V

Valeriani, Luigi Molinari (1758–1828)

Ugo Rabbeno

Born at Imola, near Bologna, Valeriani was a learned man, and well acquainted with the classical languages; he studied poetry, physics, law, and economics. He was appointed in 1797 a member of the legislative body in Milan, and in 1801 professor of public economy at the university of Bologna where Pellegrino Rossi was his pupil.

In his day Valeriani was widely known; he wrote many works, some of which were never published. Though diffuse and obscure in style, his writings deserve attention for the learning they display and a certain originality of conception. Trained both as a lawyer and an economist, his writings bear especially on the relation between economics and law. He devoted himself with assiduity to the theory of value, and wrote a book on the subject. He maintains that the law of value depends rigidly on supply and demand, supporting this theory with a geometrical illustration from the relative quantities of both; he combats the theory of cost of production and engaged in a controversy on this question with Melchiorre Gioja. In illustrating the theory of value he employs mathematical formulae. These are, however, not employed as a means of investigating the phenomena of prices, but are only symbols employed to express in mathematical language economic laws already known – as Montanari justly said.

Selected Works

1806. *Del prezzo cose tutte mercantili*.
1807. *Trattato sulle misure*.
1823. *Trattato dei cambi*.
1827. *Saggio di erotemi di quella parte del gius delle genti e pubblico che dicesi pubblica economia*.

Bibliography

Cavazzoni-Pederzini, A. 1859. *Intorno alla vita, opere e dottrine di L. Molinari Valeriani.*

Cossa, L. 1891–1901. Saggi bibliografica di economia politica. *Giornale degli economisti*.

Cossa, L. 1893. *Introduction to the study of political economy*. Trans. L. Dyer, London: Macmillan.

Lampertico, F. 1904. Della vita e degli scritti di Luigi Valeriani Molinari, economista. *Atti della Reale Academia dei Lincei* 11(1), Rome.

Montanari, G. 1804. Della moneta trattato mercantile. *Economia politica* 44, Milan.

Montanari, G. 1876. *Notizie e lettere inedite de G. Montanari*. Modena: G. Campori.

Montanari, G. 1892. *La matematica applicata all' economia politica da Cesare Beccaria, Guglielmo Silio, Luigi Molinari Valeriani*. Ed. A. Scialoja.

Rossi, P.L.E. 1863–7. *Oeuvres complètes*. Paris.

Macmillan Publishers Ltd (ed.), *The New Palgrave Dictionary of Economics*,
https://doi.org/10.1057/978-1-349-95189-5

Value and Price

Meghnad Desai

The problem of the relationship between value and price – the so called Transformation Problem – is a central issue in Marxian economics. In one sense it can be posed as a technical or mathematical problem of deriving a set of prices from a given set of value equations. But if it were only a technical problem then it should have a definite answer – either a solution exists or it does not. It is surprising therefore that this problem has continued to attract succeeding generations of economists since the date of publication of volume 3 of *Capital* in 1894 (Marx 1894).

The debate shows no signs of abating and seems a rare example of a problem which continues to invite new solutions or versions in new mathematical language of the old solution. There can rarely have been a question in economic theory which has been solved so many times in so many different mathematical languages but yet not resolved finally. This continuing fascination of the Transformation Problem leads one to suspect that there is more than a technical issue at stake.

The *locus classicus* of the debate is chapter IX of *Capital* Vol. 3 (3/IX), which was published posthumously by Engels from notes left by Marx. There is evidence however that the material contained in this volume was written some time in the 1860s before the publication of *Capital* Vol. 1 (Marx 1867). This is of more than biographical interest in the debate. In Vol. 1, Marx developed his theory on the explicit assumption that values and prices were proportional to each other. This was done in awareness of two qualifying conditions; *first* that this was a special case and generally value and prices were related systematically but not proportionally, but *second* that values and value relations were unobservable, latent or structural whereas prices were observable, actual and phenomenal. The hidden nature of value relations – commodity fetishism – is crucial to Marx's argument and hence it would have been totally uncharacteristic of Marx's approach not to have foreseen that values and prices diverge from each other.

This divergence of prices from values emerged as a central result of 3/IX and was seized upon by Böhm-Bawerk in his *Karl Marx and the Close of His System* (1896; Sweezy 1949) as a basic deficiency and disproof of Marx's theory of profits. He took it to be a complication that may have arisen in Marx's work after he had written the first volume and an impression was conveyed that the price value divergence, being contrary to the proportionality assumed in Vol. 1, invalidated the conclusions in that volume.

If Böhm-Bawerk was able to gain and convey this impression it was because Marx's attempt at solving the Transformation Problem *looks* unfinished. Having derived a numerical solution for prices from a set of value equations, as we will see below, Marx confronts the divergence as a puzzle and then spends some pages tacking around the problem but in no way presenting it as a systematic outcome. Thus it could be thought from reading 3/IX that the Transformation Problem was left unsolved.

The Problem

Marx's theory of profit was that profits were the money form of surplus value produced by labour during the production process. The conversion of surplus value into profits was accomplished not at the level of the firm but of the whole economy. This conversion had to be effected in the context of a contractual purchase of labour by employers (i.e. no extraeconomic coercion) and secondly, the rate of profit had to be equal in all activities. The first consideration meant that the wage rate – the exchange value of the commodity sold by the labourer and bought by the employer – was determined on the same principles as any other commodity. Thus the existence of surplus value had to be reconciled with an economic determination of the exchange value of the commodity labour power.

To drive a wedge between the product of labour and its price, Marx used the accepted distinction between use value and exchange value of a commodity. The commodity in question, labour power, is the labourer's potential for production. The use-value of labour power to the purchaser of the commodity – the capitalist employer – was measured in terms of the total labour time contracted to be spent by the labourer in production – the length of the working day in hours. The exchange value of labour power, like that of any other commodity, was the amount of labour time required for its reproduction, measured by the labour time equivalent of the basket of wage goods purchasable by the given wage. Having thus obtained two commensurable measures of the use value and the exchange value of labour power, the wedge between them was identified as surplus value, produced by the labourer but retained by the purchaser of labour power, the capitalist employer.

Now the total value of a commodity comprised the value contained in the materials used up in the production process – raw materials and energy used as well as the wear and tear of the fixed means of production – which Marx labelled *constant capital* (c) and the total value contributed by labourers. The latter consists of the exchange value of the wage, i.e. of paid labour, labelled variable capital (v) by Marx, and surplus labour (value) (s) which was the remainder. Given this framework the proportion of surplus value to value paid for (constant capital plus variable capital) is defined as the (value) rate of profit. This quantity can be expressed as a product of the rate of surplus value (s/v) and the organic composition of capital ($c/c + v$). Thus, the (value) rate of profit ρ in the ith economic activity

$$p_i = \frac{s_i}{\upsilon_i}\left[1 - \frac{c_i}{(c_i + \upsilon_i)}\right] = r_i(1 - g_i) \qquad (1)$$

where r_{i} is the rate of surplus value and g_{i} is the organic composition of capital. But if this were the basis of actual profits, activities with higher proportion of living labour would earn a higher rate of profit (given identical rates of exploitation) relative to one with the lesser labour intensive activity. But since we have to provide for equal rates of profit in all activities, a further step has to be taken to reconcile the theory of unequal value rates of profit with equal actual (or price) rates of profit.

Marx envisaged a pooling of surplus value from all activities at the level of economy and then its redistribution in a transformed form as profits equiproportional to the amount of capital (fixed and variable) invested in each activity. This was done by the price of a product departing from its unit value. The ratio would be above one for activities with organic composition of capital above average and below one for those below average. This condition will reconcile the unequal value rates of profit, given equal rates of surplus value with equal (price) rates of profit. Indeed for Marx this gives a usable rule to predict transfer of surplus value from one sector to another as he did use in his chapter on Absolute Rent (3/XLV).

The problem is however that the numerical example used in 3/IX contained a conceptual error (though this is disputed as we shall see below) which gave the calculations a tentative, halffinished, unsolved appearance. This can be best explained by setting out Marx's numerical example but in a more general notation. He took five activities labeled $i = 1,\ldots,5$, each using as inputs constant capital c_i and variable capital v_i with the g_i being different in each activity from the other. The output of the activities were not specifically identified nor was it clear whether they were of the constant capital or the variable capital category. To keep the inputs and outputs separate therefore let input prices be labelled p_c, p_v and output prices p_i.

The value of output can be expressed as

$$\begin{aligned} y_i = c_i + \upsilon_i + s_i &= \{[1 + r(1 - g_i)]/(1 - g_i)\}\upsilon_i \\ &= [(1 + \rho_i)/(1 - g_i)]\upsilon_i \end{aligned} \qquad (2)$$

In Eq. (2), we have used Eq. (1) and assumed as Marx did that the rate of exploitation is identical in all activities. (All the variables total value y_i as well as c_i, v_i could be interpreted as being per unit of physical output if thought convenient.)

Corresponding to Eq. (2), the price (total revenue) of output was written by Marx as

$$\begin{aligned} p_i &= (1+\pi)(c_i+v_i) \\ &= ((1+\pi)/(1-g_i))\upsilon_i \end{aligned} \tag{3}$$

Again but especially in this case, variables could be thought of in terms of per unit of output.

To determine π, the actual (price) rate of profit, Marx imposed the condition that the sum of surplus values in all activities was equal to the total of profits over all activities i.e.

$$\sum_i s_i = r\sum_i \upsilon_i = \pi \sum_i (c_i+\upsilon_i). \tag{4a}$$

Since however his five units were taken to be of the same size in terms of total value, he also trivially obtained an alternative normalization condition that the total value produced equalled total revenue, i.e.

$$\sum y_i = \sum p_i \tag{4b}$$

Using the normalization conditions notice that Eqs. (2) and (3) together yield

$$\begin{aligned} p_i/y_i &= (1+\pi)/(1+\rho_i) \\ &= (1+r(1-g))/(1+r(1-g_i)). \end{aligned} \tag{5}$$

Thus strict proportionality of prices and values can only hold if either the rate of exploitation is zero i.e. no exploitation or for the case of identical organic compositions of capital $g_i = g$. Given Eq. (4b) it was not difficult to see that the price value differences cancel out in the aggregate. While Marx found some positive and some negative deviations of p_i from y_i, he had no precise explanation to offer at this stage. It is obvious however as he saw that Eq. (5) implies

$$\begin{aligned} p_i/y_i \gtrless 1 \quad &\text{as } g_i \gtrless \overline{g} \quad \text{where} \quad \overline{g} \\ &= \sum \overline{c}_i / \sum (c_i+\upsilon_i). \end{aligned}$$

The problem with Marx's calculation is not that prices diverge from values; that they must, but that the specification of Eq. (3) is mistaken if Eq. (5) holds. The correct way to write the price equation is to weight the inputs by their respective prices, i.e.

$$p_i = (1+\pi)(p_c c_i + p_\upsilon \upsilon_i). \tag{3a}$$

At one level, we can see that Marx made a mistake in considering the cost of inputs in value terms rather than in price terms. It has been argued however (Shaikh 1977; Morishima and Catephores 1975) that Eqs. (2), (3), (4a), (4b), and (5) can be thought of as the first stage of an ergodic process. By substituting the values obtained by Eq. (5) into Eq. (3) to modify the input prices, the calculations will converge so that the prices in Eqs. (5) and (3) would be consistent with each other.

But this can only be done if the physical specification of c_i and v_i is matched to one or more of the commodities produced. If this is not done then we have two more prices than we can solve for. It was Bortkiewicz's merit to have reformulated Marx's problem using Marx's Reproduction Schemes outlined in *Capital* Vol. 2 to allow for matching specification of physical outputs and inputs with constant and variable capital. This allowed him to reduce the size of the problem (the number of unknowns) and allow for aggregate availability constraints on inputs and outputs. He took a model with three commodities (industries or departments) with Department 1 'capital' good (constant capital), Department 2 'wage' good (variable capital) and Department 3 capitalists' consumption (luxury) good. Thus, two of his three commodities were inputs as well as outputs in the production process i.e. they are basic in the sense of Sraffa but the third one is an output to be consumed but not an input.

Let the three departments (commodities) be denoted as $j = 1,2,3$. The value equations are the same as in Marx but Bortkiewicz's treatment allows a clearer input–output demarcation. Thus, the value equations can be written

$$y_j = y_{1j} + (1+r)y_{2j} \tag{6}$$

where y_{ij} is the input of good i in the output of good j etc. The price equations are

$$p_j = (1 + \pi)\sum_i p_i y_{ij}. \tag{7}$$

Bortkiewicz preserved Eq. (4a) as the normalization condition. But in addition he took care to ensure that the conditions of simple reproduction were satisfied. Thus, he imposed for the two inputs

$$y_i = \sum_i y_{ij}, \qquad i = 1, 2. \tag{8}$$

But having implicitly chosen his magnitudes to satisfy Eq. (4b) as well, he imposed a condition

$$\sum_j s_j = y_3. \tag{9}$$

While Eq. (8) are conditions on total availability of inputs to sustain the required level of output, Eq. (9) is a 'consumption function' for the recipients of surplus value. As there is no accumulation by assumption, we require that all surplus value is spent on the 'luxury good' produced by Department 3.

Thus Bortkiewicz correctly formulated the problem and even put it in the appropriate general equilibrium framework lacking in Marx's formulation in 3/IX. The solution is straightforward and need not be given here (see Sweezy 1942, 1949; Desai 1979). This should have settled any debate about the problem. It emerges that prices are systematic functions of values but are not proportional to them. But the solution was published in German in 1907 and did not become generally known until Sweezy described it in his *Theory of Capitalist Development*, nor did it become available until Sweezy's translation of it in 1949. Within this forty-year interval, economists' knowledge of the linear model had advanced as a result of the works of Leontieff and von Neumann. It was obvious therefore that the problem could be reformulated in these terms. Winternitz proposed such a formulation in 1949 and full general solution in terms of *n* goods was given by Morishima and Seton (1961). Roemer (1980) has shown that the linearity assumption can be dropped and a solution in the 'Arrow–Debreu language' can be obtained.

Two areas of controversy arose during the 1970s. First was whether it was necessary to go through the transformation problem at all to solve for prices from physical input–output data. This was raised by Samuelson (1971). Second is a more serious question about the conditions required for solution when there is joint production in the von Neumann–Sraffa sense.

Samuelson's point can be simply made. In order to arrive at value equations such as Eqs. (2) or (6), we have to translate the data which are in terms of physical output flows and labour inputs into the direct and indirect labour content of inputs. After such a translation, we proceed with the transformation. But as we know from input–output analysis, from the physical input data, one can directly solve for prices from the dual of the Leontieff matrix. If one thought of the purpose of the exercise to provide merely a set of prices consistent with a set of values, he is entirely right. What the criticism misses, however, is that if we were to follow Marx's purpose in providing a theory of profits, the separation of labour input into paid and unpaid components (which assumes a political economic background) and the use of the concept of the rate of exploitation are required. If one is to reject Marx's theory of profits, it can be done quite independently of the Transformation Problem, as Wicksteed was able to do even before the publication of *Capital* Vol. 3 since he rejected the labour theory of value, classical or Marxian, as such (Wicksteed 1884; see Desai 1979, for details).

The second line of criticism is much more serious. This is because it claims that positive surplus value is neither necessary nor sufficient for positive profits i.e. it denies the existence of any mapping from values to prices that can satisfy certain general conditions. The problem is with Marx's treatment of fixed capital. In his formulation of the value equations, Marx takes a flow measure of non-labour inputs. This suffices if all capital equipment has only one period life since then the stock and flow measures are equivalent. But if the capital equipment lives beyond the production period some account has to be taken of this in writing the value and prices equations. Bortkiewicz was also able to formulate this problem with different rates of turnover of capital i.e. different lengths of life in another, even lesser known, paper of his (Bortkiewicz 1906–7). But he took the rates of turnover to be fixed and known in advance. this is less general than one wishes (see Desai (1979) for a description). Marx can be said to have used implicitly a neoclassical accounting whereby the rental on capital correctly measures its productive contribution. But as Morishima (1973) points out a von Neumann accounting scheme in a 'joint production' model is more appropriate.

It was Steedman (1977) who first constructed a numerical example in which there is negative surplus value but positive profit. This is an example of the generic case of nonconvexities which are known to arise in activity analysis (Koopmans 1951). Steedman made it however an argument for abandoning Marxian value theory in favour of a Ricardo–Sraffa formulation. This suggestion has parallels with Samuelson's suggestion since the detour via labour values can be shown to be misleading in some cases. It has also been pointed out that the non-convexity problem can arise in the Ricardo–Sraffa scheme just as much as in the Marx scheme. Morishima (1973, 1975) has taken the view that all that is necessary is to reformulate the value price problem under joint production with appropriate inequality constraints so that non-negativity of (surplus) values and prices are assured. This would seem the more rigorous formulation. The question does remain however of the behavioural foundations of the mechanism that will ensure that in a capitalist economy, only activities with positive surplus values are chosen.

The transformation problem thus continues to fascinate economists even as they debate its relevance. It formed the basis in Bortkiewicz's case for an early formulation of a general equilibrium problem in linear terms. It has been argued that it is more appropriate for planning calculations in a socialist economy than in a capitalist economy whose workings it was supposed to illuminate (Samuelson and Weiszacker 1971; Morishima 1973). To Marxists as to their opponents, more important issues such as the moral justification for capitalism seem to be at stake in the solution or non-solution of this seemingly arid technical problem. This is one reason why it will no doubt go on attracting new solutions and new attacks.

See Also

- Bortkiewicz, Ladislaus von (1868–1931)
- Sweezy, Paul Marlor (1910–2004)

Bibliography

von Böhm-Bawerk, E.R. 1896. Zum Abschluss des Marxschen System. In *Staatswissenschaftliche Arbeiten: festgaben für Karl Knies*, ed. O.V. Boenig, Berlin. Trans. as 'Karl Marx and the close of his system' in Sweezy (1949).

von Bortkiewicz, L. 1906–7. Wertrechnung und Preisrechnung im Marxschen System. *Archiv fur Sozialwissenschaft und Sozialpolitik*, July 1906, July and September 1907. Trans. as: Value and price in the Marxian system. In *International economic papers* No. 2, ed. Alan T. Peacock et al., London/New York: Macmillan, 1952.

von Bortkiewicz, L. 1907. Zur Berichtigung der grundlegenden theoretischen konstruktion von Marx im dritten Band des 'Kapital'. *Jahrbucher für Nationalökonomie und Statistik*, July. Trans. as: 'On the correction of Marx's fundamental theoretical construction in the third volume of *Capital*' as Appendix in Sweezy (1949).

Desai, M. 1979. *Marxian economics*. Oxford: Basil Blackwell.

Koopmans, T.C. 1951. *Activity analysis of production and allocation*, Cowles Commission Monograph, vol. 13. New York: John Wiley.

Marx, K. 1894. *Das Kapital*, Vol. III, ed. F. Engels. Hamburg: Otto Meissner.

Morishima, M. 1973. *Marx's economics: A dual theory of value and growth*. Cambridge: Cambridge University Press.

Morishima, M., and G. Catephores. 1975. The transformation problem: A Markov process. In *Value exploitation and growth – Marx in the light of modern economic theory*, ed. M. Morishima. New York: McGraw–Hill.

Morishima, M., and F. Seton. 1961. Aggregation in Leontief matrices and the labour theory of value. *Econometrica* 29: 203–220.

Roemer, J. 1980. A general equilibrium approach to marxian economics. *Econometrica* 48: 505–530.

Samuelson, P.A. 1971. Understanding the marxian notion of exploitation: A summary of the so-called transformation problem between marxian values and competitive prices. *Journal of Economic Literature* 9(2): 399–431.

Samuelson, P.A. 1972. *The collected scientific papers of Paul A. Samuelson*, Vol. 3, ed. Robert C. Merton. Cambridge, MA: MIT Press.

Samuelson, P.A. and, C. Weiszacker. 1971. A new labour theory of value for rational planning through the use of the bourgeois profit rate. *Proceedings of the national academy of sciences*, June. Also in Samuelson (1972).

Schwartz, J. 1977. *The subtle anatomy of capitalism*. California: Santa Monica.

Shaikh, A. 1977. Marx's theory of value and the transformation problem. In *Schwartz* (1977).

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

Sweezy, P.M. 1942. *The theory of capitalist development*. New York: Monthly Review Press.

Sweezy, P.M., ed. 1949. *Karl Marx and the close of his system* by E. von Böhm-Bawerk and Böhmbawerk's criticism of Marx by Hilferding. New York: Augustus Kelly.

Wicksteed, P.H. 1884. Das Kapital: A criticism. First published in *Today*, October 1884, reprinted in P.H. Wicksteed, *The Commonsense of Political Economy,* Vol. II, 1933.

Winternitz, J. 1948. Values and prices: A solution of the so-called transformation problem. *Economic Journal* 58: 276–280.

Value Elicitation

Glenn W. Harrison

Abstract

Economists are interested in eliciting values at the level of the individual because market values do not provide the information needed to measure consumer surplus, value new products, or value goods that have no market. Direct and indirect procedures have been developed to elicit values, and each has some strengths and weaknesses. The evidence points to several recommendations for best practice in the reliable elicitation of values, trading off transparency and rigour.

Keywords

Auctions; Bias correction; Cheap talk; Conjoint choice; Contingent valuation; Cost–benefit analysis; English auction; Latent choice models; Maximum likelihood; Multiple price lists; Revealed preference theory; Second-price auction; Value elicitation; Vickrey auction

JEL Classifications

C9

Why elicit values? The prices observed on a market reflect, on a good competitive day, the equilibrium of marginal valuations and costs. They do not quantitatively reflect the infra-marginal or extra-marginal values, other than in a severely censored sense. We know that infra-marginal values are weakly higher, and extra-marginal values are weakly lower, but beyond that one must rely on functional forms to extrapolate. For policy purposes this is generally insufficient to undertake cost–benefit calculations.

When producers are contemplating a new product or innovation they have to make some judgement about the value that will be placed on it. New drugs, and the R&D underlying them, provide an important example. Unless one can heroically tie the new product to existing products in terms of shared characteristics, and somehow elicit values on those characteristics, there is no way to know what price the market will bear. Value elicitation experiments can help fill that void, complementing traditional marketing techniques (see Hoffman et al. 1993).

Many goods and services effectively have no market, either because they exhibit characteristics of public goods or it is impossible to credibly deliver them on an individual basis. These non-market goods have traditionally been valued using surveys, where people are asked to state a valuation 'contingent on a market existing for the good'. The problem is that these surveys are hypothetical in terms of the deliverability of the good and the economic consequences of the response, and this understandably generates controversy about their reliability (Harrison, 2006).

Procedures

Direct methods for value elicitation include auctions, auction-like procedures and 'multiple price lists'.

Sealed-bid auctions require the individual to state a valuation for the product in a private manner, and then award the product following certain rules. For single-object auctions, the second-price (or Vickrey) auction awards the product to the highest bidder but sets the price equal to the highest rejected bid. It is easy to show, to students of economics at least, that the bidder has a dominant strategy to bid his true value: any bid higher or lower can only end up hurting the bidder in expectation. But these incentives are not obvious to inexperienced subjects. A real-time counterpart of the second-price auction is the English (or ascending bid) auction, in which an auctioneer starts the price out low and then bidders increase the price to become the winner of the product. Bidders seem to realize the dominant strategy property of the English auction more quickly than in comparable second-price sealed-bid auctions, no doubt due to the real-time feedback on the opportunity costs of deviations from that strategy (see Rutström, 1998; Harstad, 2000). Familiarity with the institution is also surely a factor in the superior performance of the English auction: first encounters with the second-price auction rules lead many non-economists to assume that there must be some 'trick'.

Related schemes collapse the logic of the second-price auction into an auction-like procedure due to Becker et al. (1964). The basic idea is to endow the subject with the product, and to ask for a 'selling price'. The subject is told that a 'buying price' will be picked at random, and that, if the buying price that is picked exceeds the stated selling price, the product will be sold at that price and the subject will receive that buying price. If the buying price equals or is lower than the selling price, the subject keeps the lottery and plays it out. Again, it is relatively transparent to *economists* that this auction procedure provides a formal incentive for the subject to truthfully reveal the certainty-equivalent of the lottery. One must ensure that the buyout range exceeds the highest price that the subject would reasonably state, but this is not normally a major problem. One must also ensure that the subject realizes that the choice of a buying price does not depend on the stated selling price; a surprising number of respondents appear not to understand this independence, even if they are told that a physical randomizing device is being used.

Multiple price lists present individuals with an ordered menu of prices at which they may choose to buy the product or not. In this manner the list resembles a menu, akin to the price comparison websites available online for many products. For any given price, the choice is a simple 'take it or leave it' posted offer, familiar from retail markets. The set of responses for the entire list is incentivized by picking one at random for implementation, so the subject can readily see that misrepresentation can only hurt for the usual revealed preference reasons. Refinements to the intervals of prices can be implemented, to improve the accuracy of the values elicited (see Andersen et al. 2006). These methods have been widely used to elicit risk preferences and discount rates, as well as values for products (see Holt and Laury, 2002; Harrison et al. 2002; Andersen et al., 2007).

Indirect methods work by presenting individuals with simple choices and using a latent structural model to infer valuations. The canonical example comes from the theory of revealed preference, and confronts the decision-maker with a series of purchase opportunities from a budget line and asks him to pick one. By varying the budget lines one can 'trap' latent indifference curves and place nonparametric or parametric bounds on valuations. The same methods extend naturally to variations in the non-price characteristics of products, and merge with the marketing literature on 'conjoint choice' (for example, Louviere et al. 2000; Lusk and Schroeder, 2004). Access to scanner data from the massive volume of retail transactions made every day promises rich characterizations of underlying utility functions, particularly when merged with experimental methods that introduce exogenous variation in characteristics in order to statistically condition and 'enrich' the data (Hensher et al. 1999). One of the attractions of indirect methods is that one can employ choice tasks which are familiar to the subject, such as binary 'take it or leave it' choices or rank orderings. The lack of precision in that type of qualitative data requires some latent structure before one can infer values, but behavioural responses are much easier to explain and motivate for respondents.

One major advantage of undertaking structural estimation of a latent choice model is that valuations can be elicited in a more fundamental manner, explicitly recognizing the decision process underlying a stated valuation. A structural model can control for risk attitudes when choices are being made in a stochastic setting, which is almost always the case in practical settings. Thus one can hope to tease apart the underlying deterministic valuation from the assessment of risk. Likewise, non-standard models of choice posit a myriad of alternative factors that might confound inference about valuation: respondents might distort preferences from their true values, they might exhibit loss aversion in certain frames, and they might bring their own home-grown reference points or aspiration levels to the valuation task. Only with a structural model can one hope to identify these potential confounds to the valuation process. Quite apart from wanting to identify the primitives of the underlying valuation free of confounds, normative applications will often require that some of these distortions be corrected for. That is only possible if one has a complete structural model of the valuation process.

A structural model also provides an antidote to those that claim that valuations are so contextual as to be an unreliable will-o'-the-wisp. If someone is concerned about framing, endowment effects, loss aversion, preference distortions, social preferences, and any number of related behavioural notions, it is impossible to generate a scientific dialogue without being able to write out a structural model and jointly estimate it.

Lessons and Concerns

The most important lesson that has been learned from decades of experimental research into the behavioural properties of these procedures to elicit values is: keep it simple. This refers primarily to the nature of the task given to respondents. It can be dangerous to rely on fancy rules that ensure incentives to truthfully reveal valuations only if everyone sees a complete chain of logic, even if that logic is apparent to trained economists. Of course, one can use 'cheap talk' and just tell people to reveal the truth since it is in their best interests, but one cannot be sure that such admonitions work reliably. Cultural familiarity with institutions counts for a lot when subjects are otherwise placed in an artefactual valuation task.

The desire to keep it simple has a corollary: the use of more rigorous statistical techniques to infer valuations. This implication follows from the need to make inferences about valuations on a cardinal scale when responses are often between subject and qualitative. Progress has been made in the use of numerical simulation methods for the maximum likelihood estimation of random utility models that allow extraordinary flexibility (for example, Train, 2003).

We also have a better understanding now of the manner in which valuations may be biased by being hypothetical, due to procedural devices in the institution being employed, and because of field context (for example, Harrison et al. 2004). More constructively, methods have been developed to undertake *ex ante* 'instrument calibration' to remove biases using controlled experiments, and to implement *ex post* 'statistical calibration' to filter out any remaining systematic biases (see Harrison, 2006).

Finally, the manner in which valuations change with states of nature is starting to be understood. Insights here again come from thinking about valuation as a latent, structural decision process. If we observe the same person state a different value for the same product at two different times, is it because he has a shift in his utility function, a change in some argument of his utility function, a change in his perceived opportunity set, or something else? If valuation is viewed as a process we can begin to design procedures that can help us identify answers to these questions, and better understand the valuations that are observed.

See Also

- ▶ Auctions (experiments)
- ▶ Demand theory
- ▶ Experimental economics
- ▶ Experimental methods in economics
- ▶ Revealed preference theory
- ▶ Utility

Bibliography

Andersen, S., G.W. Harrison, M.I. Lau, and E.E. Rutström. 2006. Elicitation using multiple price lists. *Experimental Economics* 4: 383–405.

Andersen, S., G.W. Harrison, M.I. Lau, and E.E. Rutström. 2007. Valuation using multiple price list formats. *Applied Economics* 39: 675–682.

Becker, G.M., M.H. DeGroot, and J. Marschak. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9: 226–232.

Harrison, G.W. 2006. Experimental evidence on alternative environmental valuation methods. *Environmental and Resource Economics* 34: 125–162.

Harrison, G.W., M.I. Lau, and M.B. Williams. 2002. Estimating individual discount rates for Denmark: A field experiment. *American Economic Review* 5: 1606–1617.

Harrison, G.W., R.M. Harstad, and E.E. Rutström. 2004. Experimental methods and elicitation of values. *Experimental Economics* 2: 123–140.

Harstad, R.M. 2000. Dominant strategy adoption and Bidders' experience with pricing rules. *Experimental Economics* 3: 261–280.

Hensher, D., J. Louviere, and J.D. Swait. 1999. Combining sources of preference data. *Journal of Econometrics* 89: 197–221.

Hoffman, E., D.J. Menkhaus, D. Chakravarti, R.A. Field, and G.D. Whipple. 1993. Using laboratory experimental auctions in marketing research: A case study of new packaging for fresh beef. *Marketing Science* 3: 318–338.

Holt, C.A., and S.K. Laury. 2002. Risk aversion and incentive effects. *American Economic Review* 5: 1644–1655.

Louviere, J.J., D.A. Hensher, and J.D. Swait. 2000. *Stated choice methods: Analysis and application*. New York: Cambridge University Press.

Lusk, J.L., and T.C. Schroeder. 2004. Are choice experiments incentive compatible? A test with quality differentiated beef steaks. *American Journal of Agricultural Economics* 2: 467–482.

Rutström, E.E. 1998. Home-grown values and the design of incentive compatible auctions. *International Journal of Game Theory* 3: 427–441.

Train, K.E. 2003. *Discrete choice methods with simulation*. New York: Cambridge University Press.

Value Judgements

John C. Harsanyi

Keywords

Harsanyi, J. C.; Preferences; Public choice; Value judgements

JEL Classifications

B0

The Claim of Objective Validity

One may define value judgements as judgements of approval or disapproval claiming objective validity. Many of our judgements of approval and disapproval do not involve such claims. When I say that I like a particular dish, I do not mean to imply that other people ought to like it too or that those disliking it are making a mistake. All I am doing is expressing my personal preference and my personal taste. (But an expert chef or an expert food critic may very well claim that his judgements about food have some degree of objective validity – in the sense that other gastronomic experts would tend to agree with his judgements. Of course, it is an empirical question whether his claim would be justified and, more generally, how much agreement there is in fact among expert judges of food.) Yet when I say that Hitler's murder of many millions of innocent people was a moral outrage, I do mean to do more than express my personal moral attitudes and do mean to imply that anybody who tried to defend Hitler's actions would be morally wrong.

In claiming objective validity, value judgements resemble factual judgements (both those dealing with empirical facts and those dealing with logical–mathematical facts). But they resemble judgements of personal preference in expressing human attitudes (those of approval or disapproval) rather than expressing beliefs about matters of fact, as factual judgements do. But this immediately poses a difficult philosophical problem: We can understand what it means for factual judgements to be objectively valid, that is, to be true, or to be objectively invalid, that is, to be false. They will be true if they describe the relevant facts as these facts actually are, and will be false if they fail to do so. But in what sense can judgements expressing human attitudes be objectively valid or invalid?

It seems to me that this can happen in at least two different ways. Such judgements can be objectively invalid either because they are contrary to the facts or because they are based on the wrong value perspective. Value judgements can be contrary to the facts in the following sense: When we form our attitudes, we do so on the basis of some specific factual assumptions so that our attitudes and our judgements expressing these attitudes will be contrary to the facts if they are based on false factual assumptions. Mistaken factual assumptions may vitiate both our value judgements about instrumental values and those about intrinsic values. Thus, if I approve of using A as a means to achieve some end B, I will do this on the assumption that A is causally effective in achieving B. Hence, my approval will be mistaken if this assumption is incorrect. Likewise, if I approve of A as an intrinsically desirable goal, I will do this on the assumption that A has some qualities I find intrinsically attractive. My approval will be mistaken if in fact A does not possess these qualities.

Another way of value judgement may be objectively invalid is by being based on a value perspective different from the one it claims to have. For example, I may claim that my support for some government policy is based on its being in the public interest, even though actually it is based on its being in my own personal interest. Or, I may praise a very undistinguished novel as a great work of art merely because it supports my own political point of view. When a person claims to base his value judgement on one value perspective though actually he bases it on another, he may be simply lying, being fully aware of not telling the truth. Another possibility is that he is unaware, or only half aware, of using a value perspective different from the one he claims to use. (Likewise, when a person is making a value judgement based on false factual assumptions, he may or may not be fully aware of the falsity of these assumptions.)

Disagreements in Value Judgements

As we all know, disagreements in value judgements are extremely common and in many cases are very hard, or even impossible, to resolve. It seems to me that in most cases careful analysis would show that these disagreements about values are based on disagreements about the facts. Yet they may be very hard to resolve because these factual disagreements may be about very subtle facts about which reliable information is very hard, or even impossible, to obtain. For instance, our value judgements about a person's behaviour will often crucially depend on what we think his motives are. Some observers may attribute very noble motives to him, while others may do the opposite. Yet the available evidence might be consistent with either assumption. Other value judgements we make may hinge on our predictions about future facts. Thus, different economists may advocate very different economic policies because they have very different expectations on the likely effects of specific policies – even if their ultimate policy objectives are much the same. Yet, at the present stage of our knowledge about the economic system, we may be unable to tell with any degree of confidence which predictions are right and which are wrong.

Of course, we could avoid most of these disagreements if we refrained from making value judgements until we could ascertain with some assurance that the factual assumptions underlying the value judgements we want to make are correct. But this would require more intellectual self-discipline than most of us can muster. We have to act one way or another; and it is psychologically much easier for us to act if we can manage to entertain value judgements justifying our actions – even if the factual assumptions underlying these value judgements go far beyond, or are even clearly inconsistent with, the available evidence.

Let me add that most disagreements in value judgements are not disagreements about what the basic values of human life actually are. Rather, most disagreements are about the relative weights and the relative priorities to be assigned to different basic values. Some individuals and some societies will learn from their experience – possibly based on a very idiosyncratic personal or national history – that things tend to work out best if value A is given far greater weight than value B is. Other individuals and other societies will reach very much the opposite conclusion on the basis of their experience. Once a given ranking of these two values has been adopted, it may be retained for a long time even when conditions change and make this ranking utterly inappropriate. For instance, an individual or a society that suffered a good deal from lack of individual freedom may be so preoccupied with political liberty as to neglect the need for social discipline – even under conditions that would make the need for social discipline paramount.

Besides disagreements about the facts, another source of value conflicts is philosophical disagreements about the correct value perspectives to be used in making various classes of value judgements. For instance, even if two people agree about all the relevant facts, they may still make conflicting moral value judgements if they disagree about the nature of morality and, therefore, disagree about the nature of the moral perspective to be used in making moral value judgements. (For instance, one individual may favour a utilitarian interpretation of morality – see, for example, Harsanyi 1977 – while the other may favour an entitlement interpretation – see Nozick 1974.) In the same way, disagreements about the nature of the aesthetic perspective to be used in making aesthetic value judgements may lead to disagreements about the artistic quality of various works of art.

Value Judgements in Economics

There was a time when many economists wanted to ensure the objectivity of economic analysis by excluding value judgements, and even the study of value judgements, from economics. (A very influential advocate of this position has been Robbins 1932.) Luckily, they have not succeeded; and we now know that economics would have been that much poorer if they had.

After some important preliminary work in the 1930s and the 1940s, mainly in welfare economics, a new era in the study of economically relevant value judgements, has started with Arrow's *Social Choice and Individual Values* (1951). This book has shown how to express alternative value judgements in the form of precisely stated formal axioms, how to investigate their logical implications in a rigorous manner, and how to examine their mutual consistency or inconsistency. Arrow's book and the research inspired by it have greatly enriched economic theory not only in welfare economics but also in several other fields, including the theory of competitive equilibrium. It has given rise to a new sub-discipline called *public choice theory*, which is a rigorous study of voting and of alternative voting systems and which has made important contributions to the study of alternative political systems and of alternative moral codes and, more indirectly, to the study of alternative economic systems as mechanisms of social choice.

Of course, value judgements often play an important role in economics even when they are not the main subjects of investigation. They influence the policy recommendations made by economists and their judgements about the merits of alternative systems of economic organization. But this need not impair the social utility of the work done by economists as long as it is work of high intellectual quality and as long as the economists concerned *know* what they are doing, *know* the qualifications their conclusions are subject to, and *tell* their readers what these qualifications are. In particular, intellectual honesty requires economists to *state* their political and moral value judgements and to make clear how their conclusions differ from those that economists of different points of view would tend to reach on the problems under discussion. What is no less important, they should make clear how *uncertain* many of their empirical claims and their predictions actually are. This is particularly important in publications addressed mainly to people outside the economist profession.

See Also

- ► Interpersonal Utility Comparisons
- ► Philosophy and Economics
- ► Positivism

Bibliography

Arrow, K.J. 1951. *Social choice and individual values*. New York: Wiley.

Harsanyi, J.C. 1977. Morality and the theory of rational behavior. *Social Research* 44: 623–656.

Nozick, R. 1974. *Anarchy, state and Utopia*. Oxford: Blackwell.

Putnam, H. 2004. *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.

Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.

Value of Life

W. Kip Viscusi

Abstract

The economic approach to valuing risks to life focuses on risk–money trade-offs for very small risks of death, or the value of statistical life (VSL). These VSL levels will generally exceed the optimal insurance amounts. A substantial literature has estimated the wage–fatality risk trade-offs, implying a median VSL of $7 million for US workers. International evidence often indicates a lower VSL, which is consistent with the lower income levels in less developed countries. Preference heterogeneity also generates different trade-off rates across the population as people who are more willing to bear risk will exhibit lower wage–risk trade-offs

Keywords

Compensating differentials; Contingent valuation; Deterrence; Hedonic models; Hedonic prices; Human capital; India; Life insurance; Risk to life; Schelling, T.; Smith, A.; South Korea; Taiwan; Value of life; Value of statistical life

JEL Classifications

J17

Issues pertaining to the value of life and risks to life are among the most sensitive and controversial in economics. Much of the controversy stems from a misunderstanding of what is meant by this terminology. There are two principal value-of-life concepts – the amount that is optimal from the standpoint of insurance, and the value needed for deterrence. These concepts address quite different questions that are pertinent to promoting different economic objectives.

The insurance value received the greater attention in the literature until recent decades. The basic principle for optimal insurance purchases is that it is desirable to continue to transfer income to the post-accident state until the marginal utility of income in that state equals the marginal utility of income when healthy. In the case of property damage, it is desirable to have the same level of utility and marginal utility of income after the accident as before. In contrast, fatalities and serious injuries affect one's utility function, decreasing both the level of utility and the marginal utility for any given level of income, making a lower income level after a fatality desirable from an insurance standpoint. Thus, the value of life and limb from the standpoint of insurance may be relatively modest.

The second approach to valuing life is the optimal deterrence amount. What value for a fatality sets the appropriate incentives for those avoiding the accident? In the case of financial losses, the optimal insurance amount, the optimal deterrence amount, and the 'make whole' amount are identical; however, for severe health outcomes, such as fatalities, the optimal deterrence amount will exceed the optimal level of compensation.

The economic measure for the optimal deterrence amount is the risk–money trade-off for very small risks of death. Since the concern is with small probabilities, not the certainty of death, these values are referred to as the value of statistical life (VSL). Economic estimates of the VSL amounts have included evidence from market decisions that reveal the implicit values reflected in behaviour as well as the use of survey approaches to elicit these money–risk trade-offs directly. Government regulators in turn have used these VSL estimates to value the benefits associated with risk reduction policies. Because of the central role of VSL estimates in the economics literature, those analyses will be the focus here rather than income replacement for accident victims.

Valuing Risks to Life

Although economics has devoted substantial attention to issues generally termed the 'value of life', this designation is in many respects a misnomer. What is at issue is usually not the value of life itself but rather the value of small risks to life. As Schelling (1968) observed, the key question is how much people are willing to pay to prevent a small risk of death. For small changes in risk, this amount will be approximately the same as the amount of money that they should be compensated to incur such a small risk. This risk–money trade-off provides an appropriate measure of deterrence in that it indicates the individual's private valuation of small changes in the risk. It thus serves as a measure of the deterrence amount for the value to the individual at risk of preventing accidents and as a reference point for the amount the government should spend to prevent small statistical risks. Because the concern is with statistical lives, not identified lives, analyses of government regulations now use these VSL levels to monetize risk reduction benefits.

Suppose that the amount people are willing to pay to eliminate a risk of death of 1/10,000 is $700. This amount can be converted into a value of statistical life estimate in one of two ways. First, consider a group of 10,000 individuals facing that risk level. If each of them were willing to contribute $700 to eliminate the risk, then one could raise a total amount to prevent the statistical death equal to 10,000 people multiplied by $700 per person, or $7 million. An alternative approach to conceptualizing the risk is to think of the amount that is being paid per unit risk. If we divide the willingness to pay amount of $700 by the risk probability of one in 10,000, then one obtains the value per unit risk. The value per statistical life is $7million using this approach as well.

Posing hypothetical interview questions to ascertain the willingness-to-pay amount has been a frequent survey technique in the literature on the value of life. Such studies are often classified as 'contingent valuation surveys' or 'stated preference surveys', in that they seek information regarding respondents' decisions given hypothetical scenarios (see Jones-Lee, 1989; Viscusi, 1992). Survey evidence is most useful in addressing issues that cannot be assessed using market data. How, for example, do people value death from cancer compared with acute accidental fatalities? Would people be interested in purchasing pain-and-suffering compensation, and does such an interest vary with the nature of the accident? Potentially, survey methods can yield insights into these issues.

Evidence from actual decisions that people make is potentially more informative than trade-offs based on hypothetical situations if suitable market data exists. Actual decision-makers are either paying money to reduce a risk or receiving actual compensation to face a risk, which may be a quite different enterprise from dealing with hypothetical interview money. In addition, the risks to them are real so that they do not have to engage in the thought experiment of imagining that they face a risk. It is also important, however, that individuals accurately perceive the risks they face. Surveys can present respondents with information that is accurate. Biased risk perceptions may bias estimates of the money–risk trade-off in the market. Random errors in perceptions will bias estimates of the trade-off downward. The reason for this result can be traced to the standard errors-in-variables problem. A regression of the wage rate on the risk level, which is measured with error, will generate a risk variable coefficient that will be biased downward if the error is random. The estimated wage–risk trade-off will consequently understate its true value.

Empirical Evidence on the Value of Statistical ife

A large literature has documented significant trade-offs between income received and fatality risks. Most of these studies have examined wage–risk trade-offs but many studies have focused on product and housing risks as well. The wage–risk studies have utilized data from the United States as well as many other countries throughout the world. The primary implication of these results is that estimates of the value of life in the United States are clustered in the \$4 million to \$10 million range, with an average value of life in the vicinity of \$7 million.

Since the time of Adam Smith (1776), economists have observed that workers will require a ‘compensating differential’ to work on jobs that pose extra risk. These wage premiums in turn can be used to assess risk–money trade-offs and the value of life. The underlying methodology used for this analysis derives from the hedonic price and wage literature, which focuses on ‘hedonic’ or ‘quality-adjusted’ prices and wages. Rosen (1986) and Smith (1979), among others, review this methodology.

To see how the hedonic model works, let us begin with the supply side of the market. The worker’s risk decision is to choose the job with the fatality risk p that provides the highest level of expected utility (*EU*). The worker faces a market offer curve $w(p)$ that is the outer envelope of the individual firms’ market offer curves. Let there be two states of the world: good health with utility $U(w)$ and death with utility $V(w)$, where this term is the worker’s bequest function. The utility function has the property that good health is preferable to ill health, and workers are risk-averse or risk-neutral, or $U(w) > V(w)$; U'; $V' > 0$; and U'', $V'' \leq 0$. The job choice is to

$$\underset{p}{MAX} \quad EU = (1-p)U(w(p)) + pV(w(p)),$$

leading to the result

$$\frac{dw}{dp} = \frac{U(w) - V(w)}{(1-p)\,U'(w) + pV'(w)}.$$

The wage-risk trade-off *dw/dp* based on the worker’s choice of a wage–risk combination for a job is the value of statistical life, which equals the difference in utility between the two health states divided by the expected marginal utility of consumption.

What trade-off rate *dw/dp* the worker will select will depend not only on worker preferences but also on the shape of the market offer curve. The best available market opportunities will be those that offer the highest wage for any given level of risk, or the outer envelope of the offer curves for the individual firms. Each individual firm will offer a wage that is a decreasing function of the level of safety. The cost function for producing safety increases with the level of safety, so the wage decline associated with incremental improvements in safety must be increasingly great to keep the firm on its isoprofit curve.

Figure 1 illustrates the nature of the hedonic labour market equilibrium. The curves OC_1 and OC_2 represent two possible market offer curves from firms with risky jobs. As the risk level is reduced, firms will offer lower wages. EU_1 and EU_2 are expected utility loci of two workers, both of whom have selected their optimal job risk from available market opportunities. The curve $w(p)$ represents the locus of market equilibria, which consists of the points at which worker indifference curves are tangent to the market offers. Thus, the empirical estimation of the hedonic labour market equilibrium focuses on the joint influence of demand and supply.

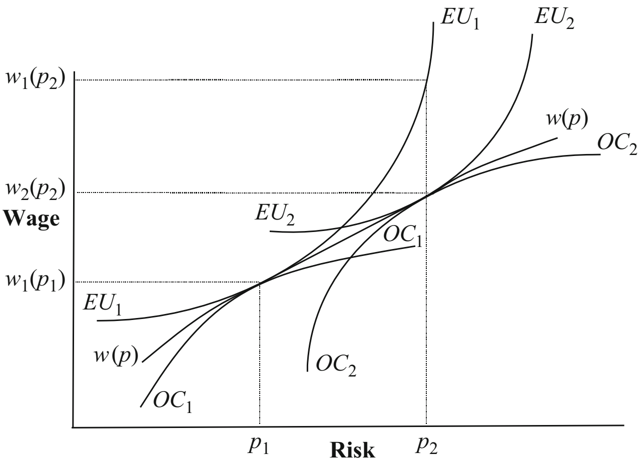
The trade-offs reflected in market equilibria do not represent a schedule of individual VSL trade-off values at different risks, but rather different VSLs for different workers. Worker 1 chooses risk p_1 with associated wage $w(p_1)$, and worker 2 chooses risk p_2 for wage $w(p_2)$. However, worker 1 would not accept risk p_2 for $w(p_2)$ even when that is the point on the hedonic equilibrium curve. Rather, worker 1 will require wage $w_1(p_2) > w_2(p_2)$ to accept this risk.

The canonical hedonic wage equation is

$$\ln w_i = \alpha + X_i'\beta + \gamma_1 p_i + \gamma_2 q_i + \gamma_3 WC_i + \varepsilon_i,$$

where w_i is worker *i*’s wage, X_i is a vector of personal characteristics and job characteristics, p_i is the worker’s fatality risk, q_i is the nonfatal injury and illness risk, and WC_i is a measure of the worker’s compensation benefits. Not all labour market studies of VSL include the q_i and

Value of Life, Fig. 1 Market process for determining compensating differentials



WC_i terms. Moreover, there are some differences in the form of the workers' compensation benefit term that is included. The most common is the expected workers' compensation replacement rate, which is the product of the injury risk and the benefit level divided by the wage rate. These differences in the empirical specification account for some of the differences across studies in the estimated VSL.

As a practical matter, there are many systematic differences that have becomes apparent in these studies. Workers at very high-risk jobs tend to have lower values of life on average since they have self-selected themselves into the very risky occupation. Through their job choices these individuals have revealed their greater willingness to endanger their lives. Workers at lower-risk jobs typically have greater reluctance to risk their lives, which accounts for their selection into these safer pursuits. Such differences are apparent in practice, as the estimated values of life for workers in the average risk jobs tend to be several times greater than those for workers in very risky jobs.

Other differences correlated with worker affluence are also evident. Health status is a normal economic good, and individuals' willingness to pay to preserve their health increases with income. Blue-collar workers, for example, have a lower value of life than do white-collar workers. In addition, there is a positive income elasticity of the estimated values of risks to life and health. Based on a sample of 50 wage–risk studies from ten countries, Viscusi and Aldy (2003) estimate that VSL has an income elasticity of 0.5 to 0.6.

Value of Life, Table 1 Labour market estimates of value of statistical life throughout the world

Study/Country	Value of statistical life ($ millions)
Median value from 30 US studies	7.0
Australia	4.2
Austria	3.9–6.5
Canada	3.9–4.7
Hong Kong	1.7
India	1.2–1.5
Japan	9.7
South Korea	0.8
Switzerland	6.3–8.6
Taiwan	0.2–0.9
United Kingdom	4.2

Note: All estimates are in year 2000 US dollars.
Source: Viscusi and Aldy (2003). For concreteness, single representative studies are drawn from their Table 4.

These differences by income level in the VSL amounts are also borne out in the international evidence on wage-risk trade-offs, such as the study of Australia and Japan by Kniesner and Leeth (1991). Table 1 summarizes representative VSL studies from throughout the world. More affluent countries such as Japan and Canada tend to have higher revealed VSL levels than countries such as South Korea, India and Taiwan. The major international anomaly is the United Kingdom, for which labour market estimates have been very unstable across studies and sometimes quite high. Deficiencies of the UK fatality risk data or correlation of these values with other unobservables may account for this pattern. Because of these limitations, the benefit assessments for risk reductions in the UK are based on stated preference values rather than labour market values, which is the approach taken by US regulatory agencies.

Because of individual heterogeneity in preferences and resources, it is not surprising that estimated values of life often differ considerably across empirical studies. These differences are not a sign that such studies are necessarily in error. These samples often consist of workers with quite different risk levels and who are situated differently. International comparisons, for example, consistently reveal differences across countries, not only because of the aforementioned aspects of heterogeneity, but because of the differences in the social insurance and workers' compensation arrangements that may be present in these countries.

The role of heterogeneity is evidenced in estimates for the implicit value for non-fatal job injuries for different worker groups. This analysis follows the same general methodological approach as does the literature on the implicit value of life. The difference is that the focus is on non-fatal job risks rather than fatalities. On average, workers value non-fatal loss injuries on the job at values ranging from \$20,000 to \$70,000 per expected job injury. Thus, for example, a worker at the high end of this range would require \$2,000 to face a one chance in 25 of being injured that year.

The estimates of the implicit values of injuries for other labour market groups who have different attitudes towards risk vary substantially from this amount. Interestingly, women often work at hazardous jobs and appear to have wage–risk tradeoffs similar to those of men. Other personal characteristics generate more evidence of heterogeneity in preferences. Cigarette smokers and people who don't use seat belts in their automobiles work on risky jobs for less per expected injury than do people who don't smoke and who use seat belts in their automobiles. What is noteworthy is that these results are not hypothetical willingness-to-pay values that these groups have expressed with respect to risks. Rather, they represent actual differences in compensation based on observed patterns of decisions in the marketplace. Markets work as expected in that they match workers to the jobs that are most appropriate for their preferences. This is a constructive role of market sorting that promotes a more efficient match-up than if, for example, all individuals were constrained to have the same job riskiness.

Preference heterogeneity has additional implications. Recall from Fig. 1 that workers may settle along different points of the available market opportunities. However, if workers face the same opportunities locus, then the worker choosing the higher risk $p2$ must always be paid a wage $w(p_2) > w(p_1)$ if $p_2 > p_1$. Interestingly, that pattern does not always hold. As shown by Viscusi and Hersch (2001), smokers choose jobs that are riskier than non-smokers' jobs but offer less additional wage compensation for incurring the risks. Smokers and non-smokers face different market offer curves and, most important, these offer curves provide for a flatter wage–risk gradient for smokers. There may be an efficiency-based rationale for these differences, as smokers are more prone to job accidents, so that there safety-related productivity is less.

Studies of the money–risk trade-offs are not restricted to the labour market. There have been a number of efforts to assess price–risk trade-offs for a variety of commodities. The contexts analysed by economists include the choice of highway speed, seat belt use, installation of smoke detectors, property values in polluted areas, and prices of automobiles. The most reliable of these studies outside the labour market are those pertaining to automobile prices in that they follow the same kind of approach as is used in the wage–risk literature. In particular, the analysts obtain price information on a wide variety of automobile models. Using regression analysis, they assess the incremental contribution of the safety characteristics per se to the product price, controlling for other product attributes. The results of these studies suggest a value of life around \$5 million.

The Duration and Quality of Life

The value-of-life terminology is misleading to the extent that risk reduction efforts do not confer immortality but simply extend life. Because of that, the major concern should not be with the value of life but with the value of extending life for different periods. In the case of preventing the risk of death to a young person, the increase in life expectancy that will be generated will exceed that for preventing a risk of death to older people. Some kind of age adjustment may be appropriate. The quantity of life matters, but which years of life matter most? Is a year of life at age 45 more valuable than a year of life at age 5 or age 70? How do various health impairments correlated with age affect the value one should attach to such years of life, and should the fact that very young children have not yet received the value of the education and rearing by their parents matter? The total ‘human capital’, which is the set of personal attributes such as education and training that affect one’s income, will be greater for older children who are further along in their development. Resolving such questions remains highly problematic.

Considerable attention has been devoted to economic analysis of age effects, including studies by Shepard and Zeckhauser (1984) and Johansson (2002). If capital markets were perfect, then VSL would steadily decline with age, reflecting the shortening of life expectancy. If, however, there are capital market imperfections, then VSL will display an inverted U-shaped relationship with age. A similar pattern is exhibited empirically by lifetime consumption patterns, which some theoretical models have linked to VSL levels over the life cycle. Although empirical estimates of the age effects are still being refined, the available evidence from survey data and market-based studies suggests that there is an inverted-U-shaped relation. The main empirical controversies concern the tails of the age distribution. To what extent is there a flattening of the VSL–age relation for the very old age groups, and how should VSL levels be assigned to children?

The quality of the life of the years saved clearly matters as well. Life years in deteriorating health may be less valuable to the individual than years in good health. Some analysts have suggested that the measure should focus on quality-adjusted life years. Making these quality adjustments has yet to receive widespread empirical implementation and is often controversial. There may be quite legitimate fears of government efforts to target expenditures by denying health care to those whose life quality is deemed to be low. People often adapt to changes in health status so that external observers may overstate the decline in well-being that occurs with serious illnesses.

Conclusion

Economic estimates of the trade-offs people make between risk and either prices or wages serve a variety of functions. First, they provide evidence on how people make decisions involving risk in labour market and product market contexts. The fact that there are probabilistic health effects does not imply that markets cease to function. Second, these estimates have proved useful in providing a reference point for how the government should value the benefits associated with regulations and other policies that reduce risk. Third, the existence of these estimates and economists’ continuing efforts to refine the values has served to highlight many of the fundamental ethical issues involved, such as how society should value reducing risks to people in different age groups.

See Also

- ▶ Compensating Differentials
- ▶ Hedonic Prices

Bibliography

Hersch, J. 1998. Compensating differentials for gender-specific job injury risks. *American Economic Review* 88: 598–627.

Johansson, P.-O. 2002. On the definition and age-dependency of the vale of statistical life. *Journal of Risk and Uncertainty* 25: 251–263.

Jones-Lee, M. 1989. *The economics of safety and physical risk*. Oxford: Basil Blackwell.

Kniesner, T., and J. Leeth. 1991. Compensating wage differentials for fatal injury risk in Australia, Japan, and the United States. *Journal of Risk and Uncertainty* 4: 75–90.

Rosen, S. 1986. The theory of equalizing differences. In *Handbook of labor economics*, ed. O. Ashenfelter and R. Layard. Amsterdam: North-Holland.

Schelling, T. 1968. The life you save may be your own. In *Problems in public expenditure analysis*, ed. S. Chase. Washington, DC: Brookings Institution.

Shepard, D., and R. Zeckhauser. 1984. Survival versus consumption. *Management Science* 30: 423–439.

Smith, A. 1776. *The wealth of nations*, ed. E. Cannan. New York: Modern Library, 1937.

Smith, R. 1979. Compensating differentials and public policy: A review. *Industrial and Labor Relations Review* 32: 339–352.

Viscusi, W. 1992. *Fatal tradeoffs: Public and private responsibilities for risk*. NewYork: Oxford University Press.

Viscusi, W., and J. Aldy. 2003. The value of a statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27: 5–76.

Viscusi, W., and J. Hersch. 2001. Cigarette smokers as job risk takers. *The Review of Economics and Statistics* 83: 269–280.

Zeckhauser, R., and D. Shepard. 1976. Where now for saving lives? *Law and Contemporary Problems* 39: 5–45.

Value of Time

Reuben Gronau

JEL Classifications

J2

Time is a scarce resource or, to use a popular adage – 'time is money'. The value of time depends on its usage and the complementary resources used with it. Firms pay for their workers' time according to the workers' value or marginal product. Households naturally place a value on their time when they sell it in the market, but they also assign a value to the time they use in the home sector. This value determines (and is sometimes determined by) the optimum combination of activities a person engages in, and the optimum combination of goods and time used in each activity. It affects the supply of labour and the demand for goods.

The recognition of the importance of time for many economic decisions related and unrelated to the labour market (e.g., schooling; transportation) is not new. The generalization of the model is associated with Becker's (1965) theory of home production. Becker (following Mincer 1963) reformulates traditional consumption theory by shifting the focus of analysis from goods to activities ('commodities', in his terms). By this approach the source of the household's welfare is its activities, which in turn, are a combination of goods and time. Welfare is maximized subject to home technology, the budget constraint, and the time constraint. Formally, the welfare function depends on the activity levels (Z_i)

$$U = U(Z_1, \ldots, Z_n)$$

where each activity is 'produced' through a combination of goods (X_i) and time (T_i)

$$Z_i = d_i(X_i, T_i).$$

The consumer's welfare is maximized subject to the budget constraint

$$\sum P_i X_i = W(Z_n) + V$$

and the time constraint

$$\sum T_i = T,$$

where P_i denote prices, $W(Z_n)$ is labour income (Z_n denoting the activity work in the market), V is non-labour income, and T is the total time available.

The maximization of welfare subject to the home production technology and the time and budget constraints yields the optimum allocation of activities:

$$\partial U/\partial Z_i = \lambda \hat{\Pi}_i,$$

and the optimum combination of inputs in the production of each activity

$$(\partial Z_i/\partial T_i)/(\partial Z_i/\partial X_i) = \hat{W}/P_i,$$

where λ denotes the marginal value of income, $\hat{\Pi}_i$ is the shadow price of activity *i*, and $\hat{W}$ is the shadow price of time. The shadow price of the activity equals its marginal cost of production

$$\hat{\Pi}_i = P_i(\partial X_i/\partial Z_i) + \hat{W}(\partial T_i/\partial Z_i).$$

Thus an increase in the shadow price of time leads to substitution of time in favour of goods and a substitution from time-intensive to goods-intensive activities.

When there are no external constraints on hours of market work the value people place on their time depends on their marginal wage rate

$$\hat{W} = w + (u_n/\lambda),$$

where w is the marginal wage rate (the change in earnings as a result of a change in market work net of taxes and any expenditures associated with work) and u_n denotes the marginal utility of labour. However, even when one is not free to change one's working hours the shadow price of time increases with wages and with income because of the increase in time scarcity.

The importance of the value of time to allocative decisions has been shown in a wide range of contexts: fertility (Becker 1960; Willis 1973; Schultz 1975), health (Grossman 1972) and most notably, labour supply and transportation. Thus, women with higher wages have higher opportunity costs of raising children and therefore tend to reduce fertility, substituting 'quality' for 'quantity'. Travellers who place a high value on their time prefer faster but more expensive modes of transport to slower and cheaper modes. Married women with young children or with high earning husbands place higher value on their time and are, therefore, more reluctant to participate in the labour force.

Theory predicts that the shadow price of time changes with the person's wage rate. It does not imply that the two are equal; they differ if the marginal net wage differs from the average wage, when labour involves direct utility (or disutility), or when it is assumed that the utility generated by an activity depends on the time inputs involved (Bruzelius 1979).

The value of time saving is a major component of the benefits of the investment in many transportation projects (Beesley 1965; Tipping 1968). To evaluate the shadow price of time transportation economists studied the trade-off between time and money implicit in the choice of modes of transport, choice of route, location decision, and demand for travel. Studying commuter choices it is found that the value placed by commuters on their time is only 1/5 to 1/2 of their wage rate. The value of walking and waiting time is found to be 2.5–3.0 times greater than the value of in-vehicle time. Differences in convenience, comfort, effort etc. are reflected in estimates of time value in bus travel that are higher than travel by car, and values that tend to increase with the length of the trip. Finally, differences between the gross and the net wage and constrained working hours result in estimates that are higher for business travel than for personal travel. (For a recent discussion of the estimating methods and results see Bruzelius 1979).

A second source for the study of the value of time at home is labour-force-participation behaviour. A person is supposed to participate in the labour force if the wage he is offered exceeds the value of his marginal productivity at home – that is, his value of home time. Studying the labour force participation patterns of US married women Gronau (1973) found that the value of time of these women increases with their schooling (most noticeably with college attendance). It is little affected by the husband's schooling and income and by age, and increase sharply when the family has children. Having a child under 3 years of age increases the value of its mother's time at home by over 25 per cent (in particular if she has a college education), but this effect diminishes as the child grows older.

See Also

- ▶ Family Economics
- ▶ Gender Roles and Division of Labour
- ▶ Household Production and Public Goods
- ▶ Leisure
- ▶ Women's Work and Wages

Bibliography

Becker, G.S. 1960. An economic analysis of fertility. In *Demographic and economic change in developed countries*, University-National Bureau Conference Series, No. 11. Princeton: Princeton University Press.

Becker, G.S. 1965. A theory of the allocation of time. *Economic Journal* 75 (September): 493–517.

Beesley, M.E. 1965. The value of time spent in travelling: Some new evidence. *Economica* 45 (May): 174–185.

Bruzelius, N. 1979. *The value of travel time*. London: Croom Helm.

Gronau, R. 1973. The effect of children on the housewife's value of time. *Journal of Political Economy* 81 (2): 168–199. March–April, Supplement.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80 (2): 223–255. March–April.

Mincer, J. 1963. Market prices, opportunity costs, and income effects. In *Measurement in economics: Studies in mathematical economics and econometrics in memory of Yehuda Grunfeld*, ed. C. Christ et al. Stanford: Stanford University Press.

Schultz, T.W., ed. 1975. *Economics of the family: Marriage, children, and human capital*. London/Chicago: Chicago University Press for the National Bureau of Economic Research.

Tipping, D.G. 1968. Time savings in transport studies. *Economic Journal* 78 (December): 843–854.

Willis, R.J. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81 (2): 14–64. March–April, Supplement.

Value-Added Tax

Gilbert E. Metcalf

Abstract

A value-added tax (VAT) is a tax on the value created in goods or services during production, distribution, and sales. VATs are generally constructed as consumption taxes. In principle, a consumption VAT is neutral in its treatment of savings and consumption. In practice, VATs are often designed with exemptions and zero-ratings that generate consumption distortions. An important issue for VATs is their implementation in a federal system. A number of modifications of the basic VAT structure have been proposed to strike a balance between the efficiency of a common rate across political units and the desire for country-specific VAT rates.

Keywords

Compensating VAT (CVAT); Consumption taxation; Depreciation; Tax evasion; Tax neutrality; Value-added tax; Vertical integration; Viable integrated VAT (VIVAT)

JEL Classifications

H2

A value-added tax (VAT) is a tax on the value created in a good or service by a business at any stage of production, distribution or sales.

Definitions and Equivalencies

Value-added is simply the difference between the value of the goods and services sold and the value of goods and services purchased as intermediate inputs. Consider the general cash flow equation for a firm.

$$S + K^{+} = L + M + K^{-} \qquad (1)$$

The equation states that cash comes into a firm from capital inflows (K^{+}) – new equity and borrowing – and proceeds from sales. Cash is used for payments for labour (L) and intermediate goods (M). Capital purchases are generally included in M. This makes a VAT a consumption tax. If capital depreciation is included in M, then this would be an income-type VAT. If no capital deduction of any form is allowed, then this would be a gross output VAT. For the remainder of the article, I focus on a consumption-type VAT. In addition, cash is retained or used for dividend and interest payments as well as any retirement of debt and equity (K^{-}).

Value added was defined above as the difference between revenue from sales and the cost of inputs:

$$VA = S - M = L + (K^{-} - K^{+}) \qquad (2)$$

Equation 2 demonstrates that there are several ways to impose a VAT. We could tax gross sales net of intermediate input purchases at each stage of production. This forms the basis for a 'subtraction method' VAT. Alternatively, we could tax gross sales and allow a credit for taxes paid by registered suppliers of intermediate inputs to the firm. The 'credit method' VAT works in this fashion. A third method is to tax the factor payments to capital and labour. This forms the basis for an 'addition method' VAT.

Value-added taxes are common throughout the world with the notable exception of the United States. Most countries use the credit method, arguing that this method is self-enforcing since the ability to take a credit for VAT paid at an earlier stage of production requires suppliers to provide an invoice detailing their VAT payments.

Tax Neutrality

As described at this most general level, a VAT shares all the attributes of a broad- based consumption tax. If comprehensively applied, it is neutral across all forms of purchased consumption. And since capital purchases are expensed (immediately deducted from the tax base), the rate of return on capital is unaffected by the tax. As with all consumption taxes, the efficiency gains from a switch from income to consumption taxation depend significantly on the tax treatment of old capital; on this point, see Auerbach and Kotlikoff (1987). In practice, VATs are not neutral for a number of reasons. First, if capital is not expensed, returns to capital are affected by the tax. Most VATs are consumption-type VATs so this is not a significant problem. Second, as noted by Cnossen (1998), some countries extend the VAT up through the manufacturing or wholesale stage but not through the retail stage. This creates distortions across consumption given the different ensuing tax rates on different commodities.

Finally, many VATs exempt certain sectors from the tax, 'zero-rate' sectors or commodities, or apply a reduced rate to certain commodities (for example, food for home preparation). Zero-rating in a credit method VAT means that firms apply a zero rate to their tax base but receive a credit for all VAT paid by suppliers. Zero-rating has no impact if applied at an intermediate stage of production since any taxes forgone at one stage are made up at the next stage. Zero-rating at the retail stage means the commodity is untaxed by a VAT. Exempting sectors from the VAT process may create peculiar outcomes. If an intermediate sector is exempted from taxation, downstream stages of production will pay a VAT not only on their value added but on the value added created in sectors upstream from the exempt sector for which no credit was received. The result is that exemptions at an intermediate stage of production can lead to the effective VAT rate being *higher* than the nominal rate. For this reason, many countries that exempt certain sectors (generally small businesses) allow voluntary participation in the VAT system. Note too that exemptions at the retail stage create incentives for vertical integration to increase the proportion of value added exempt from taxation.

Design Issues

A VAT can be levied on an 'origin' or 'destination' basis. An origin VAT taxes value added in the country in which the value added is produced, while a destination VAT taxes value added where it is consumed. Most countries employ a destination VAT and use a border tax adjustment whereby a VAT is applied to the value of imports and a rebate provided for the value of exports. While it is popularly believed that border tax adjustments favour export industries, a flexible exchange rate in general leads to the same trade balance whether the VAT is applied on an origin or destination basis. Grossman (1980) demonstrates that this proposition fails in a world with trade in intermediate goods.

Border tax adjustments are commonly applied by customs authorities, and this has given rise to special problems for the European Union with its abolition of border controls in 1992. Keen and Smith (1996) note conflicts between two important goals: maximum autonomy for individual countries to set their own tax rates and a system of country VAT structures that does not impede the creation of a single European market. Keen and

Smith propose a ‘viable integrated VAT’ (VIVAT) to address this problem. The VIVAT applies a harmonized VAT rate to intermediate producers in all European countries and a different rate for final consumption sales. The rate on final sales would vary across countries based on individual country preferences. The VIVAT can be thought of as a harmonized EU-wide VAT and a system of individual country retail sales taxes, a point that reminds us of the close connection between a VAT and a retail sales tax.

McLure (2000) notes that the VIVAT requires firms to charge different rates to different classes of customers, a non-trivial burden. He also notes that a destination-based system of VATs in a subnational system can give rise to tax evasion by households or unregistered firms importing goods (which are zero-rated at the exporting country’s border). McLure proposes a compensating VAT (CVAT), essentially an additional federal-level tax to guarantee the tax revenues that might otherwise be lost to cross-border tax evasion. The key point here is that considerable administrative complexity comes into play when a VAT is implemented by a group of countries (or states) within a common trading system (or federal government).

As with any other consumption tax framework, taxing housing and financial services is problematic with a VAT. One approach for treating housing services follows from an arbitrage argument that the present value of the stream of future consumption services from housing is equal to its purchase price. With this assumption, a tax-prepayment approach levies the VAT on the first sale of a house (but not subsequent sales) as well as additions or maintenance. With constant tax rates, this tax payment is equal to the present value of the taxes that would be paid on the housing services enjoyed by occupants of the house. If tax rates rise (fall) in the future, the tax prepayment approach raises less (more) revenue than if the housing services were taxed directly. Alternatively, the sale of all residential housing and rental income are subject to tax while the purchase of residential housing is deductible. This approach treats housing like any other capital asset which produces services (housing). Measuring and taxing the imputed rental income on owner occupied housing is a significant problem for this approach. For this reason, most consumption taxes favour the tax prepayment approach.

Financial services are even more difficult to handle under consumption taxes. One approach is to tax the net cash flow from financial services. In the terminology of Meade (1978), this would be an R + F (real plus financial) consumption tax. As Auerbach and Gordon (2002) point out, this creates considerable administrative problems since other transactions are treated on an R basis, thus giving rise to arbitrage opportunities to avoid the tax. In the European Union financial services are exempt from VAT, though Huizanga (2002) has argued that it is increasingly feasible to bring this sector into the VAT system. This sanguine perspective is not shared by all economists.

Tax Incidence and Impacts on Saving and Labour Supply

Because a VAT in its purest form is a consumption tax, its distributional impact as well as behavioural impacts are the same as those of any broad-based consumption tax. To the extent that the VAT is implemented in non-neutral ways (exemptions and zero-rating of sectors, multiple tax rates, and so forth) consumption distortions will arise similar to those of any differential rate commodity tax system.

Many countries apply a VAT tax structure with lower rates on perceived necessities (food, for example) on distributional grounds. Most economic analyses of VAT proposals recommend a uniform tax rate on all commodities to avoid consumption distortions, and recommend using an income tax to effect desired income redistribution. Cnossen (1998), however, recommends a dual rate system for developing countries on the grounds that income taxes are administratively unfeasible in these countries. While reducing tax rates on food and other necessities provides benefits to low-income households, this is a blunt instrument for redistribution given the resulting reduction in taxes to wealthy people’s purchase of food (and other low or zero-rated commodities).

See Also

- ▶ Consumption Taxation
- ▶ Fiscal Federalism
- ▶ Optimal Taxation
- ▶ Tax Competition

Bibliography

Auerbach, A.J., and R.H. Gordon. 2002. Taxation of financial services under a VAT. *American Economic Review* 92: 411–416.

Auerbach, A.J., and L.J. Kotlikoff. 1987. *Dynamic fiscal policy*. New York: Cambridge University Press.

Cnossen, S. 1998. Global trends and issues in value added taxation. *International Tax and Public Finance* 5: 399–428.

Grossman, G. 1980. Border tax adjustments: Do they distort trade? *Journal of International Economics* 10: 117–128.

Huizanga, H. 2002. A European VAT on financial services? *Economic Policy* 35: 499–534.

Keen, M., and S. Smith. 1996. The future of value-added tax in the European Union. *Economic Policy* 23: 375–420.

McLure, C. 2000. Implementing subnational value added taxes on internal trade: The compensating VAT (CVAT). *International Tax and Public Finance* 7: 723–740.

Meade, J.E. 1978. *The structure and reform of direct taxation*. London: Allen & Unwin.

Vanderlint, Jacob (Died 1740)

Peter Groenewegen

A timber merchant at Blackfriars, London, about whose life little is known except that in 1734 he published *Money Answers All Things, or an Essay to make Money Plentiful among all Ranks of People and increase our Foreign and Domestick Trade.* This work appears to have received little attention during the 18th century until Dugald Stewart referred to it as anticipating the Physiocrats on the single tax of land rent and on free trade. Stewart compared him also with David Hume 'in point of good sense and liberality' (Stewart 1794, pp. 342, 343, 346). McCulloch used Stewart's opinions on several occasions (e.g. 1845, p. 162) and may have provided the basis for Marx's charge (1878, p. 327, cf. 1867, p. 124, n.1) that 'Hume follows step by step, and often even in his personal idiosyncrasies' Vanderlint's work.

The essay itself presents a complex argument supporting a proposal for alleviating the distress from a diagnosed trade depression (pp. 134–48). This was designed to ensure prosperity for all including the labouring poor to whose plight Vanderlint was most sympathetic (pp. 72–7, 83, 88, 100). As Vanderlint explains in the opening remarks of his preface, reducing labour costs is the best way to stimulate domestic and foreign trade; the problem is how to achieve this end without the reduction in domestic demand following a cut in money wages. Vanderlint's solution rests on his proposal to extend agriculture by making more labour and land available for cultivation (pp. 117–19, 163–8). Assuming constant returns (pp. 81–2) this policy leads to increased agricultural produce, the starting point for his causal analysis. As Vickers (1960, p. 180) demonstrates, Vanderlint argues that increased agricultural produce lowers the price of wage goods, hence the money wage level, hence cost of production, hence favourably affects the balance of trade by increasing export competitiveness, increasing money supply, which increases demand for output in general and brings about full employment and prosperity. Vanderlint combines real and monetary factors in this analysis as Hume was to do two decades later. Aware of the specie mechanism (see Viner 1937, pp. 83–4), Vanderlint suggests ways of neutralizing monetary effects on the prices of provisions. He also provides interesting reflections on war and peace, marriage, luxury, and more equal distribution of income and taxation. His analysis is enriched by empirical material drawn from contemporary political arithmetick sources.

References

Marx, K.H. 1867. *Capital*. Moscow: Foreign Languages Publishing House, 1959.

Marx, K.H. 1878. From the critical history. In *Anti-Dühring*, ed. F. Engels. Moscow: Foreign Languages Publishing House, 1954.

McCulloch, J.R. 1845. *The literature of political economy.* London: LSE. Reprint, 1938.

Stewart, D. 1794. *Account of the life and writings of Adam Smith, L.L.D.* In *Adam Smith, Essays on philosophical subjects,* ed. W.P.D. Wightman, J.C. Bryce, and I.S. Ross. Oxford: Clarendon Press, 1980.

Vanderlint, J. 1734. *Money answers all things or, an essay to make money sufficiently plentiful*. New York: Johnson Reprint Corporation, n.d.

Vickers, D. 1960. *Studies in the theory of money 1690–1776*. London: Peter Owen.

Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper & Brothers.

Vansittart, Nicholas, Lord Bexley (1766–1851)

H. R. Tedder

Son of Henry Vansittart, sometime governor of Bengal, Vansittart took his MA degree at Oxford in 1791, and was called to the bar at Lincoln's Inn, where he became a bencher in 1812. He was MP for Hastings in 1796, and in 1801 was sent as minister plenipotentiary with Parker and Nelson to Copenhagen to endeavour to detach Denmark from the Northern Alliance. In April 1801 he was appointed joint-secretary to the treasury by Addington.

Between 1802 and 1812 he sat for Old Sarum, and afterwards for Harwich. In 1804 he was a lord of the treasury in Ireland and in the following year secretary to the lord lieutenant. He was reappointed joint secretary to the treasury, 1806–7, under Grenville's administration; and in 1812 became a cabinet minister, succeeding Perceval as chancellor of the exchequer. He held this office during Lord Liverpool's administration until January 1823, when he retired, and was raised to the peerage. He remained in the cabinet as chancellor of the duchy of Lancaster until 1828. He died 8 February 1851, in his 85th year.

Vansittart was a poor debater, with feeble voice and indistinct utterance, but he at one time had a certain financial reputation, and his gentle manners and benevolent character secured the attention which his natural abilities were unable to command. The eleven years during which he was chancellor of the exchequer were from a financial point of view perhaps the most critical England ever saw, but Vansittart never showed dexterity either in imposing or in remitting taxation. He introduced no measure of first importance. He was not responsible for the repeal of the income tax in 1816, the surrendering of the war malt tax, nor the return to cash payments. His resolutions on the report of the Bullion Committee have not added to his fame, and a praiseworthy scheme for converting the navy five per cents to four per cents in 1822 was coupled with an objectionable proposal to farm the pensions known as the 'dead weight annuity'. He introduced alterations into the sinking fund far from successful. He was simply an honest and industrious clerk, finally dismissed from his office with little ceremony.

Selected Works

1793. *Reflections on the propriety of an immediate conclusion of the peace*. London.

1794. *A reply to the addressed to Mr. Pitt by Jasper Wilson*. London.

1796a. *Letter to Mr. Pitt on the conduct of the Bank directors*. London.

1796b. *An inquiry into the state of the finances of Great Britain, in answer to Mr. Morgan's facts*. London.

1811. *Substance of two speeches on the bullion question*. London.

1813. Outline of a plan of finance proposed to be submitted to parliament. *Pamphleteer* 1: 255.

1815a. The budget of 1815. *Pamphleteer* 6: 27.

1815b. Speech . . . February 20 1815 in the Committee of ways and means. *Pamphleteer* 6: 1.

1818. Speech 16 March 1818 on a Grant of £1 M for (Churches). *Pamphleteer* 12: 3.

1819. Speech of the chancellor of the exchequer on the budget of 1819. *Pamphleteer* 15: 1.

Bibliography

Attwood, T. 1817. Letter to N. Vansittart on the creation of money and . . . upon the national prosperity. Birmingham.

Colchester, Lord. 1861. Diary and correspondence of Charles, Lord Colchester by his son, 3 vols. London.
Dunn, W. 1820. The Vansittart plan of finance. *Pamphleteer* 16: 263.
Walpole, S. 1878–86. *History of England*, 5 vols. London.

Varga, Evgeny (Jenö) (1879–1964)

Rudolf Nötel

Soviet economist, political activist and analyst, Varga was born in Nagytétény, Hungary, on 6 November 1879, and died in Moscow on 7 November 1964. He was a college teacher, economic journalist, Professor of Political Economy (1918) and People's Commissar of the Hungarian Soviet Republic in 1919. He was forced to leave Hungary in the first days of August 1919 for Austria (where he was detained for several months). While in exile, he worked for the Secretariat of the Communist International in Moscow and the Soviet Trade Mission in Berlin. From 1927 to 1947 he held the position of Director of the Institute of World Economy and World Politics, and from 1929 to 1964 he was a Full Member of the USSR Academy of Sciences.

In feudal-capitalist Hungary of the declining Habsburg Monarchy and two subsequent short-lived revolutions, Varga systematically covered all vital economic policy issues, including industrialization (1912), land reform, and inflation (1918). His experience as Commissar he summed up in 'The economic policy problems facing the proletarian dictatorship' (1920).

In the following period of capitalism, imperialism, colonialism, fascism and war, he found confirmation for many basic tenets of Marxism: in *The Great Crisis and its Political Consequences. Economics and Politics, 1928–1924* (1935) he empirically demonstrated the validity of the theories of exploitation, imperialism, class warfare and crises, and correctly foresaw the inescapable drift towards war and revolution.

After decades of exceptionally intensive research and varied experience (which permitted him to become one of the chief architects of Hungary's spectacularly successful Forint stabilization), he published 'Changes in the Capitalist Economy following the Second World War' (1946). Now he attributed lasting importance to reinforced state control, rising consumption shares, decolonization and the increased role of international credit in the capitalist economy and, accordingly, doubted the fatality of world crises and world wars.

These conclusions were officially rejected in the Soviet Union and he was demoted from his leading Institute position (1947). But after some interruption he resumed scientific work and restated and extended his theses (1953, 1964). His Selected Works were posthumously published in three volumes (1974).

Selected Works

1912. Az ipartelepülés és Magyarország iparosodásának problémája (The location of industry and the problem of Hungary's industrialization). *Közgazdasági Szemle* 303–313; 393–411.

1918. *A pénz: uralma a békében, bukása a háboruban* (*On money: Its peace-time power and war-time collapse*). Budapest: Népszava.

1920. *Die wirtschaftspolitishen Probleme der proletarischen Diktatur* (*The economic policy problems facing the proletarian dictatorship*), 2nd ed. Vienna: Verlag der Arbeiterbuchhandlung.

1935. *The great crisis and its political consequences. Economics and politics, 1928–1934*. London: Modern Books.

1946. *Izmeneniia v ekonomike kapitalizma v itoge vtoroi mirovoi voiny* (*Changes in the capitalist economy as a result of the Second World War*). Moscow: Gospolitizdat.

1953. *Osnovnye voprosy ekonomiki i politiki imperializma – posle vtoroi mirovoi voiny* (*Basic problems of imperialist economics and politics – After the Second World War*). Moscow: Gospolitizdat.

1964. *Politico-economic problems of capitalism*. First published in Russian. Moscow: Progress Publishers, 1968.

1974. *Kapitalizm posle vtoroi mirovoi voiny. Izbrannye proizvedeniia* (*Capitalism after the Second World War*). In *Selected works*. Moscow: Nauka; includes a bibliography listing 749 titles.

Bibliography

Sociological Institute of the CC of the HSWP. 1979. *Varga Jenömüveinek bibliográfiája*. A bibliography listing 1158 titles and editorial contribution to 30 volumes. Budapest: MSZMP KB Társadalomtudományi Intézete. variable capital. *See* Constant and Variable Capital.

Variance Decomposition

Helmut Lütkepohl

Abstract

Variance decomposition is a classical statistical method in multivariate analysis for uncovering simplifying structures in a large set of variables (for example, Anderson 2003). For example, *factor analysis or principal components* are tools that are in widespread use. Factor analytic methods have, for instance, been used extensively in economic forecasting (see for example, Forni et al. 2000; Stock and Watson 2002). In macroeconomic analysis the term 'variance decomposition' or, more precisely, 'forecast error variance decomposition' is used more narrowly for a specific tool for interpreting the relations between variables described by vector autoregressive (VAR) models. These models were advocated by Sims (1980) and used since then by many economists and econometricians as alternatives to classical simultaneous equations models. Sims criticized the way the latter models were specified, and questioned in particular the exogeneity assumptions common in simultaneous equations modelling.

Keywords

Bayesian methods; Bootstrap; Choleski decompositions; Cointegrated variables; Cointegration; Factor analysis; Forecasting; Least squares; Maximum likelihood; Multivariate analysis; Principal components; Simultaneous equations models; Structural vector autoregressions; Variance decomposition; Vector autoregressions

JEL Classifications

C32

Variance decomposition is a classical statistical method in multivariate analysis for uncovering simplifying structures in a large set of variables (for example, Anderson 2003). For example, *factor analysis or principal components* are tools that are in widespread use. Factor analytic methods have, for instance, been used extensively in economic forecasting (see for example, Forni et al. 2000; Stock and Watson 2002). In macroeconomic analysis the term 'variance decomposition' or, more precisely, 'forecast error variance decomposition' is used more narrowly for a specific tool for interpreting the relations between variables described by vector autoregressive (VAR) models. These models were advocated by Sims (1980) and used since then by many economists and econometricians as alternatives to classical simultaneous equations models. Sims criticized the way the latter models were specified, and questioned in particular the exogeneity assumptions common in simultaneous equations modelling.

VAR models have the form

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \qquad (1)$$

where $y_t = (y_{1t}, \ \ldots, y_{Kt})'$ (the prime denotes the transpose) is a vector of K observed variables of interest, the A_i's are $(K \times K)$ parameter matrices, p is the lag order and u_t is a zero mean error process which is assumed to be white noise, that is, $E(u_t) = 0$, the covariance matrix, $E\left(u_t u_t'\right) = \Sigma_u$, is time invariant and the u_t's are serially uncorrelated or independent. Here deterministic terms such as constants, seasonal dummies or polynomial trends are neglected because they are of no interest in the following. In the VAR model (1) all the variables are a priori endogenous. It is usually difficult to disentangle the relations between the variables directly from the coefficient matrices. Therefore it is useful to have special tools which help with the interpretation of VAR models. Forecast error variance decompositions are such tools. They are presented in the following.

An h steps ahead forecast or briefly h-step forecast at origin t can be obtained from (1) recursively for $h = 1\ ,\ 2\ ,\ \ldots$, as

$$y_{t+h|t} = A_1 y_{t+h-1|t} + \cdots + A_p y_{t+h-p|t}. \qquad (2)$$

Here $y_{t\,+\,j|t} = y_{t\,+\,j}$ for $j \leq 0$. The forecast error turns out to be

$$\begin{aligned} y_{t+h} - y_{t+h|t} = \quad & u_{t+h} \\ & + \sum_{i=1}^{h-1} \Phi_i u_{t+h-i} {}^{\sim} \left(0, \Sigma_h = \Sigma_u + \sum_{i=1}^{h-1} \Phi_i \Sigma_u \Phi_i' \right), \end{aligned}$$

that is, the forecast errors have mean zero and covariance matrices Σ_h. Here the Φ_i's are the coefficient matrices of the power series expansion $\left(I_K - A_1 z - \cdots - A_p z^p\right)^{-1} = I_K + \sum_{i=1}^{\infty} \Phi_i z^i$. Note that the inverse exists in a neighbourhood of $z = 0$ even if the VAR process contains integrated and cointegrated variables. (For an introductory exposition of forecasting VARs, see Lütkepohl 2005.)

If the residual vector u_t can be decomposed in instantaneously uncorrelated innovations with economically meaningful interpretation, say, $u_t = B\varepsilon_t$ with $\varepsilon_t \sim (0,\, I_K)$, then $\Sigma_u = BB'$ and the forecast error variance can be written as $\Sigma_h = \sum_{i=0}^{h-1} \Theta_i \Theta_i'$, where $\Theta_0 = B$ and $\Theta_i = \Phi_i B$; $i = 1\ ,\ 2\ ,\ \ldots$. Denoting the (n,m)th element of Θ_j by $\theta_{nm\ ,\ j}$, the forecast error variance of the kth element of the forecast error vector is seen to be

$$\begin{aligned} \sigma_k^2(h) &= \sum_{j=0}^{h-1} \left(\theta_{k1,j}^2 + \cdots + \theta_{kK,j}^2 \right) \\ &= \sum_{j=1}^{K} \left(\theta_{kj,0}^2 + \cdots + \theta_{kj,h-1}^2 \right). \end{aligned}$$

The term $\left(\theta_{kj,0}^2 + \cdots + \theta_{kj,h-1}^2 \right)$ may be interpreted as the contribution of the jth innovation to the h-step forecast error variance of variable k. Dividing the term by $\sigma_k^2(h)$ gives the percentage contribution of innovation j to the h-step forecast error variance of variable k. This quantity is denoted by $\omega_{kj\ ,\ h}$ in the following. The $\omega_{kj\ ,\ h}\ , j = 1\ ,\ \ldots\ , K$, decompose the h-step ahead forecast error variance of variable k in the contributions of the ε_t innovations. They were proposed by Sims (1980) and are often reported and interpreted for various forecast horizons.

For such an interpretation to make sense it is important to have economically meaningful innovations. In other words, a suitable transformation matrix B for the reduced form residuals has to be found. Clearly, B has to satisfy $\Sigma_u = BB'$. These relations do not uniquely determine B, however. Thus, restrictions from subject matter theory are needed to obtain a unique B matrix and, hence, unique innovations ε_t. A number of different possible sets of restrictions and approaches for specifying restrictions have been proposed in the literature in the framework of *structural VAR models*. A popular example is the choice of a lower-triangular matrix B obtained by a Choleski decomposition of Σ_u (for example, Sims 1980). Such a choice amounts to setting up a system in recursive form where shocks in the first variable have potentially instantaneous effects also on all the other variables, shocks to the second variable can also affect the third to last variable instantaneously, and so on. In recursive systems it may be possible to associate the innovations with variables, that is, the jth component of $\varepsilon_t$is primarily viewed as a shock to the jth observed variable. Generally, the innovations can also be associated with unobserved variables, factors or forces and they may be named accordingly. For example, Blanchard and Quah (1989) consider a bivariate model for output and the unemployment rate, and they investigate effects of supply and demand shocks. Generally, if economically meaningful innovations can be found, forecast error variance decompositions provide information about the relative importance of different shocks for the variables described by the VAR model.

Estimation of reduced form and structural form parameters of VAR processes is usually done by least squares, maximum likelihood or Bayesian methods. Estimates of the forecast error variance components, $\omega_{kj\ ,\ h}$, are then obtained from the VAR parameter estimates. Suppose the VAR coefficients are contained in a vector α, then $\omega_{kj\ ,\ h}$ is a function of α, $\omega_{kj\ ,\ h} = \omega_{kj\ ,\ h}(a)$. Denoting the estimator of α by $\hat{\alpha}$, $\omega_{kj\ ,\ h}$ may be estimated as $\hat{\omega}_{kj,h} = \sigma_{kj,h}(\hat{\alpha})$. If $\hat{\alpha}$ is asymptotically normal, that is, $\sqrt{T}(\hat{\alpha} - \alpha) \rightarrow d\mathcal{N}(0, \Sigma_{\hat{\alpha}})$, then, under general conditions, $\hat{\omega}_{kj,h}$ is also asymptotically normally distributed, $\sqrt{T}\left(\hat{\omega}_{kj,h} - \omega_{kj,h}\right) \rightarrow d\ \mathcal{N}\left(0, \sigma^2_{kj,h} = \frac{\partial\ \omega_{kj,h}}{\partial\ \alpha'} \Sigma_{\hat{\alpha}} \frac{\partial\ \omega_{kj,h}}{\partial\ \alpha}\right)$, provided the variance of the asymptotic distribution is non-zero. Here $\partial\omega_{kj\ ,\ h}/\partial a$ denotes the vector of first-order partial derivatives of $\omega_{kj\ ,\ h}$ with respect to the elements of α (see Lütkepohl 1990, for the specific form of the partial derivatives). Unfortunately, $\sigma^2_{kj,h}$ is zero even for cases of particular interest, for example, if $\omega_{kj\ ,\ h} = 0$ and, hence, the jth innovation does not contribute to the h-step forecast error variance of variable k (see Lütkepohl 2005, Sect. 3.7.1, for a more detailed discussion). The problem can also not easily be solved by using bootstrap techniques (cf. Benkwitz et al. 2000). Thus, standard statistical techniques such as setting up confidence intervals are problematic for the forecast error variance components. They can at best give rough indications of sampling uncertainty. The estimated $\omega_{kj\ ,\ h}$'s are perhaps best viewed as descriptive statistics.

See Also

- ▶ Impulse Response Function
- ▶ Structural Vector Autoregressions
- ▶ Vector Autoregressions

Bibliography

Anderson, T. 2003. *An introduction to multivariate statistical analysis*. 3rd ed. New York: John Wiley.

Benkwitz, A., H. Lütkepohl, and M. Neumann. 2000. Problems related to bootstrapping impulse responses of autoregressive processes. *Econometric Reviews* 19: 69–103.

Blanchard, O., and D. Quah. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79: 655–673.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. 2000. The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics* 82: 540–552.

Lütkepohl, H. 1990. Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *Review of Economics and Statistics* 72: 116–125.

Lütkepohl, H. 2005. *New introduction to multiple time series analysis*. Berlin: Springer-Verlag.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.

Stock, J., and M. Watson. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167–1179.

Variance, Analysis Of

Andrew Gelman

Abstract

Analysis of variance (ANOVA) is a statistical procedure for summarizing a classical linear model – a decomposition of sum of squares into a component for each source of variation in the model – along with an associated test (the *F*-test) of the hypothesis that any given source of variation in the model is zero. More generally, the variance decomposition in ANOVA can be extended to obtain inference for the variances of batches of parameters (sources of variation) in multilevel regressions. ANOVA is a useful addition to regression in that it structures inferences about batches of parameters.

Keywords

Analysis of variance (ANOVA); Balanced and unbalanced data; Bayesian inference; Classical linear models; Classical method of moments; Contrast analysis; Experimental economics; Finite-population standard deviation; Fixed effects and random effects; Generalized linear models; Linear models; Linear regression; Multilevel models; Nonexchangeable models; Probability; Super-population standard deviation; Variance decomposition

JEL classifications

C1; C3

Introduction

Analysis of variance (ANOVA) represents a set of models that can be fit to data, and also a set of methods for summarizing an existing fitted model. We first consider ANOVA as it applies to classical linear models (the context for which it was originally devised; Fisher 1925) and then discuss how ANOVA has been extended to generalized linear models and multilevel models. Analysis of variance is particularly effective for analysing highly structured experimental data (in agriculture, multiple treatments applied to different batches of animals or crops; in psychology, multi-factorial experiments manipulating several independent experimental conditions and applied to groups of people; industrial experiments in which multiple factors can be altered at different times and in different locations).

At the end of this article, we compare ANOVA with simple linear regression.

Analysis of Variance for Classical Linear Models

ANOVA as a Family of Statistical Methods

When formulated as a statistical model, analysis of variance refers to an additive decomposition of data into a grand mean, main effects, possible interactions and an error term. For example, Gawron et al. (2003) describe a flight-simulator experiment that we summarize as a 5 × 8 array of measurements under five treatment conditions and eight different airports. The corresponding two-way ANOVA model is $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij.}$ The data as described here have no replication, and so the two-way interaction becomes part of the error term. (If, for example, each treatment x airport condition were replicated three times, then the 120 data points could be modelled as $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, with two sets of main effects, a two-way interaction, and an error term.)

This is a linear model with 1 + 4 + 7 coefficients, which is typically identified by constraining the $\sum_{i=1}^{5} \alpha_i = 0$ and $\sum_{j=1}^{8} \beta_j = 0$. The corresponding ANOVA display is shown in Table 1:

1. For each source of variation, the degrees of freedom represent the number of effects at that level, minus the number of constraints (the five treatment effects sum to zero, the eight airport effects sum to zero, and each row and column of the 40 residuals sums to zero).
2. The total sum of squares – that is, $\sum_{i=1}^{5} \sum_{j=1}^{8} \left(y_{ij} - \bar{y}_{..}\right)^2$ – is 0.078 + 3.944 + 1.417, which can be decomposed into these three terms corresponding to variance described by treatment, variance described by airport, and residuals.
3. The mean square for each row is the sum of squares divided by degrees of freedom. Under the null hypothesis of zero row and column effects, their mean squares would, in expectation, simply equal the mean square of the residuals.
4. The F-ratio for each row (except for the last) is the mean square, divided by the residual mean square. This ratio should be approximately 1 (in expectation) if the corresponding effects are zero; otherwise we would generally expect the F-ratio to exceed 1. We would expect the F-ratio to be less than 1 only in unusual models with negative within-group correlations (for example, if the data y have been renormalized

Variance, Analysis Of, Table 1 Classical two-way analysis of variance for data on five treatments and eight airports with no replication

Source	Degrees of freedom	Sum of squares	Mean square	F-ratio	p-value
Treatment	4	0.078	0.020	0.39	0.816
Airport	7	3.944	0.563	11.13	<0.001
Residual	28	1.417	0.051		

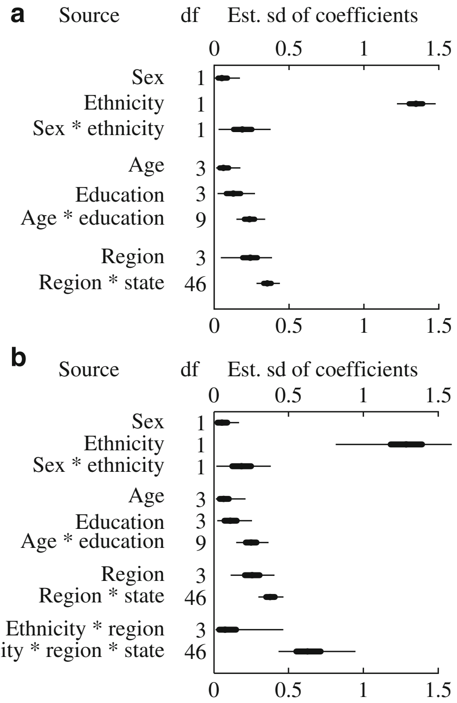
Note: The treatment-level variation is not statistically distinguishable from noise, but the airport effects are statistically significant

Sources for all examples in this article: Gelman (2005) and Gelman and Hill (2006)

in some way, and this had not been accounted for in the data analysis).

5. The p-value gives the statistical significance of the F-ratio with reference to the F_{v_1,v_2}, where v_1 and v_2 are the numerator and denominator degrees of freedom, respectively. (Thus, the two F-ratios in Fig. 1 are being compared to $F_{4,28}$ and $F_{7,28}$ distributions, respectively.) In this example, the treatment mean square is lower than expected (an F-ratio of less than 1), but the difference from 1 is not statistically significant (a p-value of 82%), hence it is reasonable to judge this difference as explainable by chance, and consistent with zero treatment effects. The airport mean square is much higher than would be expected by chance, with an F-ratio that is highly statistically significantly larger than 1; hence we can confidently reject the hypothesis of zero airport effects.

More complicated designs have correspondingly complicated ANOVA models, and complexities arise with multiple error terms. We do not intend to explain such hierarchical designs and analyses here, but we wish to alert the reader to such complications. Textbooks such as Snedecor and Cochran (1989) and Kirk (1995) provide examples of analysis of variance for a wide range of designs.



Variance, Analysis Of, Fig. 1 ANOVA display for two logistic regression models of the probability that a survey respondent prefers the Republican candidate for the 1988 US presidential election. *Notes*: Point estimates and error bars show median estimates, 50% intervals and 95% intervals of the standard deviation of each batch of coefficients. The large coefficients for ethnicity, region and state suggest that it might make sense to include interactions, hence the inclusion of ethnicity × region and ethnicity × state interactions in the second model (Source: data from seven CBS News polls)

ANOVA to Summarize a Model That Has Already Been Fitted

We have just demonstrated ANOVA as a method of analysing highly structured data by decomposing variance into different sources, and comparing the explained variance at each level with what would be expected by chance alone.

Any classical analysis of variance corresponds to a linear model (that is, a regression model, possibly with multiple error terms); conversely, ANOVA tools can be used to summarize an existing linear model.

The key is the idea of 'sources of variation', each of which corresponds to a batch of coefficients in a regression. Thus, with the model $y = X\beta + \varepsilon$, the columns of X can often be batched in a reasonable way (for example, in Table 1, a constant term, four treatment indicators, and seven airport indicators) and the mean squares and F-tests then provide information about the amount of variance explained by each batch.

Such models could be fitted without any reference to ANOVA, but ANOVA tools could then be used to make some sense of the fitted models, and to test hypotheses about batches of coefficients.

Balanced and Unbalanced Data

In general, the amount of variance explained by a batch of predictors in a regression depends on which other variables have already been included in the model. With *balanced data*, however, in which all groups have the same number of observations (for example, each treatment applied exactly eight times, and each airport used for exactly five observations), the variance decomposition does not depend on the order in which the variables are entered. ANOVA is thus particularly easy to interpret with balanced data. The analysis of variance can also be applied to unbalanced data, but then the sums of squares, mean squares, and F-ratios will depend on the order in which the sources of variation are considered.

ANOVA for More General Models

Analysis of variance represents a way of summarizing regressions with large numbers of predictors that can be arranged in batches, and a way of testing hypotheses about batches of coefficients. Both these ideas can be applied in settings more general than linear models with balanced data.

F-tests

In a classical balanced design (as in the example in Table 1), each F-ratio compares a particular batch of effects to zero, testing the hypothesis that this particular source of variation is not necessary to fit the data.

More generally, the F-test can compare two nested models, testing the hypothesis that the smaller model fits the data adequately (so that the larger model is unnecessary). In a linear model, the F-ratio is $\frac{(\mathrm{SS}_2 - \mathrm{SS}_1)/(\mathrm{df}_2 - \mathrm{df}_1)}{\mathrm{SS}_1/\mathrm{df}_1}$, where SS_1, df_1 and SS_2, df_2 are the residual sums of squares and degrees of freedom from fitting the larger and smaller models, respectively.

For generalized linear models, formulas exist using the *deviance* (the log-likelihood multiplied by -2) that are asymptotically equivalent to F-ratios. In general, such models are not balanced, and the test for including another batch of coefficients depends on which other sources of variation have already been included in the model.

Inference for Variance Parameters

A different sort of generalization interprets the ANOVA display as inference about the variance of each batch of coefficients, which we can think of as the relative importance of each source of variation in predicting the data. Even in a classical balanced ANOVA, the sums of squares and mean squares do not exactly do this, but the information contained therein can be used to estimate the variance components (Cornfield and Tukey 1956; Searle et al. 1992). Bayesian simulation can then be used to obtain confidence intervals for the variance parameters. As illustrated in this article we display inferences for standard deviations (rather than variances) because these are more directly interpretable. Compared with the classical ANOVA display, our plots emphasize the estimated variance parameters rather than testing the hypothesis that they are zero.

Generalized Linear Models

The idea of estimating variance parameters applies directly to generalized linear models as well as unbalanced data-sets. All that is needed is that the parameters of a regression model are batched into 'sources of variation'. Figure 1 illustrates with a multilevel logistic regression model, predicting vote preference given a set of demographic and geographic variables.

Multilevel Models and Bayesian Inference

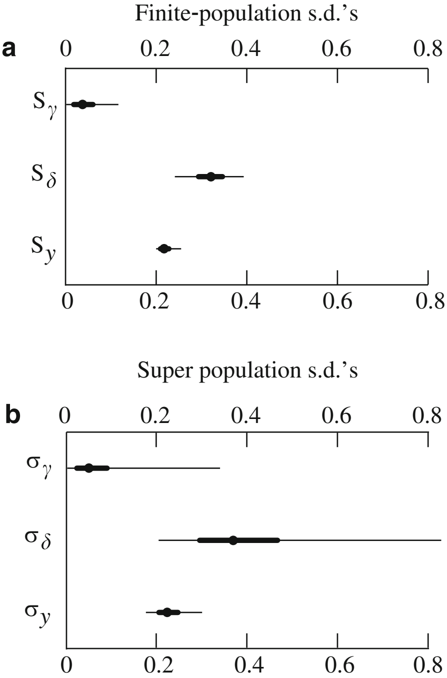
Analysis of variance is closely tied to multilevel (hierarchical) modelling, with each source of variation in the ANOVA table corresponding to a variance component in a multilevel model (see Gelman 2005). In practice, this can mean that we perform ANOVA by fitting a multilevel model, or that we use ANOVA ideas to summarize multilevel inferences. Multilevel modelling is inherently Bayesian in that it involves a potentially large number of parameters that are modelled with probability distributions (see, for example, Goldstein 1995; Kreft and De Leeuw 1998; Snijders and Bosker 1999). The differences between Bayesian and non-Bayesian multilevel models are typically minor except in settings with many sources of variation and little information on each, in which case some benefit can be gained from a fully Bayesian approach which models the variance parameters.

Related Topics

Finite Population and Super-Population Variances

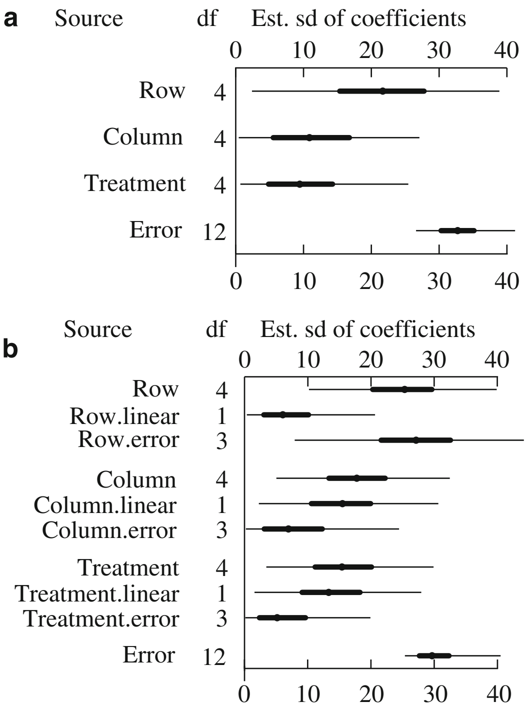
So far in this article we have considered, at each level (that is, each source of variation) of a model, the standard deviation of the corresponding set of coefficients. We call this the *finite-population* standard deviation. Another quantity of potential interest is the standard deviation of the hypothetical *super-population* from which these particular coefficients were drawn. The point estimates of these two variance parameters are similar – with the classical method of moments, the estimates are identical, because the super-population variance is the expected value of the finite-population variance – but they will have different uncertainties. The inferences for the finite-population standard deviations are more precise, as they correspond to effects for which we actually have data.

Figure 2 illustrates the finite-population and super-population inferences at each level of the model for the flight-simulator example. We know much more about the five treatments and eight airports in our data-set than for the general populations of treatments and airports. (We similarly know more about the standard deviation of the 40 particular errors in out data-set than about their hypothetical super-population, but the differences here are not so large because the super-population distribution is fairly well estimated from the 28 degrees of freedom available from these data.)

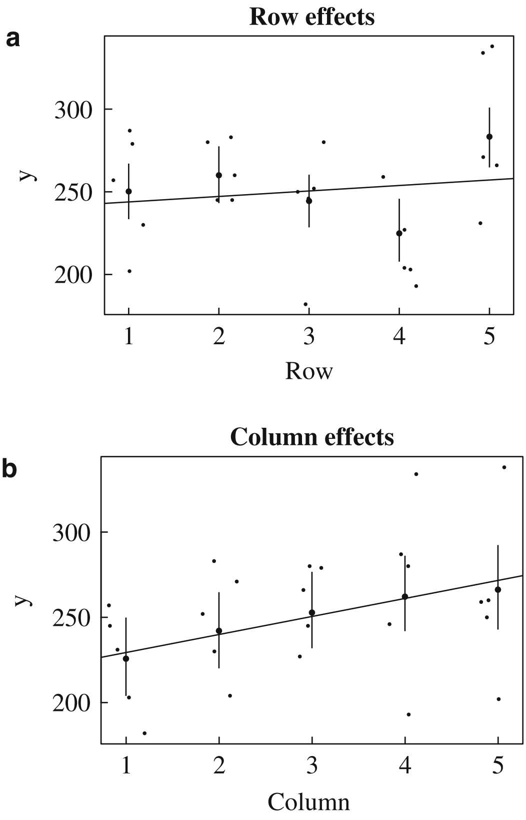


Variance, Analysis Of, Fig. 2 Median estimates, 50% intervals and 95% intervals for (a) finite population and (b) super-population standard deviations of the treatment-level, airport-level and data-level errors in the flight-simulator example from Table 1. *Note*: The two sorts of standard deviation parameters have essentially the same estimates, but the finite-population quantities are estimated much more precisely. (We follow the general practice in statistical notation, using Greek and Roman letters for population and sample quantities, respectively)

There has been much discussion about fixed and random effects in the statistical literature (see Eisenhart 1947; Green and Tukey 1960; Plackett 1960; Yates 1967; LaMotte 1983; and Nelder 1977, 1994, for a range of viewpoints), and unfortunately the terminology used in these discussions is incoherent (see Gelman 2005, sec. 6). Our resolution to some of these difficulties is to always fit a multilevel model but to summarize it with the appropriate class of estimand – super-population



Variance, Analysis Of, Fig. 3 ANOVA displays for a 5 × 5 latin square experiment (an example of a crossed three-way structure): (**a**) with no group-level predictors, (**b**) contrast analysis including linear trends for rows, columns and treatments. *Note*: See also the plots of coefficient estimates and trends in Fig. 4



Variance, Analysis Of, Fig. 4 Estimates ± 1 standard error for the row, column, and treatment effects for the latin square experiment summarized in Fig. 3. *Note*: The five levels of each factor are ordered, and the lines display the estimated linear trends

or finite population – depending on the context of the problem. Sometimes we are interested in the particular groups at hand; at other times they are a sample from a larger population of interest. A change of focus should not require a change in the model, only a change in the inferential summaries.

Contrast Analysis

Contrasts are a way to structuring the effects within a source of variation. In a multilevel modelling context, a contrast is simply a group-level coefficient. Introducing contrasts into an ANOVA allows a further decomposition of variance. Figure 3 illustrates for a 5 × 5 latin square experiment (this time, not a split plot): the left plot in the figure shows the standard ANOVA, and the right plot shows a contrast analysis including linear trends for the row, column and treatment effects. The linear trends for the columns and treatments are large, explaining most of the variation at each of these levels, but there is no evidence for a linear trend in the row effects.

Figure 4 shows the estimated effects and linear trends at each level (along with the raw data from the study), as estimated from a multilevel model. This plot shows in a different way that the variation among columns and treatments, but not among rows, is well explained by linear trends.

Non-exchangeable Models

In all the ANOVA models we have discussed so far, the effects within any batch (source of variation) are modelled exchangeably, as a set of coefficients with mean 0 and some variance. An important direction of generalization is to non-exchangeable models, such as in time series, spatial structures (Besag and Higdon 1999), correlations that arise in particular application areas such as genetics (McCullagh 2005), and dependence in multi-way structures (Aldous 1981; Hodges et al. 2005). In these settings, both the hypothesis-testing and variance-estimating extensions of ANOVA become more elaborate. The central idea of clustering effects into batches remains, however. In this sense, 'analysis of variance' represents all efforts to summarize the relative importance of different components of a complex model.

ANOVA Compared with Linear Regression

The analysis of variance is often understood by economists in relation to linear regression (for example, Goldberger 1964). From the perspective of linear (or generalized linear) models, we identify ANOVA with the structuring of coefficients into batches, with each batch corresponding to a 'source of variation' (in ANOVA terminology).

As discussed by Gelman (2005), the relevant inferences from ANOVA can be reproduced by using regression – but not always least-squares regression. Multilevel models are needed for analysing hierarchical data structures such as 'split-plot designs', where between-group effects are compared with group-level errors, and within-group effects are compared with data-level errors.

Given that we can already fit regression models, what do we gain by thinking about ANOVA? To start with, the display of the importance of different sources of variation is a helpful exploratory summary. For example, the two plots in Fig. 1 allow us to quickly understand and compare two multilevel logistic regressions, without getting overwhelmed with dozens of coefficient estimates.

More generally, we think of the analysis of variance as a way of understanding and structuring multilevel models – not as an alternative to regression but as a tool for summarizing complex high-dimensional inferences, as can be seen, for example, in Fig. 2 (finite-population and superpopulation standard deviations) and Figs. 3 and 4 (group-level coefficients and trends).

See Also

- ▶ Bayesian Statistics
- ▶ Fisher, Ronald Aylmer (1890–1962)
- ▶ Linear Models
- ▶ Two-Stage Least Squares and the k-Class Estimator

We thank Jack Needleman, Matthew Rafferty, David Pattison, Marc Shivers, Gregor Gorjanc, and several anonymous commenters for helpful suggestions and the National Science Foundation for financial support.

Bibliography

Aldous, D. 1981. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*: 581–598.

Besag, J., and D. Higdon. 1999. Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society B*: 691–746.

Cochran, W., and G. Cox. 1957. *Experimental designs*. 2nd ed. New York: Wiley.

Cornfield, J., and J. Tukey. 1956. Average values of mean squares in factorials. *Annals of Mathematical Statistics*: 907–949.

Eisenhart, C. 1947. The assumptions underlying the analysis of variance. *Biometrics* 3: 1–21.

Fisher, R.A. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Gawron, V., B. Berman, R. Dismukes, and J. Peer. 2003. New airline pilots may not receive sufficient training to cope with airplane upsets. *Flight Safety Digest* (July–August): 19–32.

Gelman, A. 2005. Analysis of variance: Why it is more important than ever (with discussion). *Annals of Statistics* 33: 1–53.

Gelman, A., and J. Hill. 2006. *Data analysis using regression and multilevel/ hierarchical models*. New York: Cambridge University Press.

Gelman, A., C. Pasarica, and R. Dodhia. 2002. Let's practice what we preach: Using graphs instead of tables. *American Statistician* 56: 121–130.

Goldberger, A. 1964. *Econometric theory*. New York: Wiley.

Goldstein, H. 1995. *Multilevel statistical models*. 2nd ed. London: Edward Arnold.

Green, B., and J. Tukey. 1960. Complex analyses of variance: General problems. *Psychometrika* 25: 127–152.

Hodges, J., Y. Cui, D. Sargent, and B. Carlin. 2005. *Smoothed ANOVA*. Technical report: Department of Biostatistics, University of Minnesota.

Kirk, R. 1995. *Experimental design: Procedures for the behavioral sciences*. 3rd ed. Pacific Grove: Brooks/Cole.

Kreft, I., and J. De Leeuw. 1998. *Introducing multilevel modeling*. London: Sage.

LaMotte, L. 1983. Fixed-, random-, and mixed-effects models. In *Encyclopedia of statistical sciences*, ed. S. Kotz, N. Johnson, and C. Read. New York: Wiley.

McCullagh, P. 2005. Discussion of Gelman (2005). *Annals of Statistics* 33: 33–38.

Nelder, J. 1977. A reformulation of linear models (with discussion). *Journal of the Royal Statistical Society A* 140: 48–76.

Nelder, J. 1994. The statistics of linear models: Back to basics. *Statistics and Computing* 4: 221–234.

Plackett, R. 1960. Models in the analysis of variance (with discussion). *Journal of the Royal Statistical Society B* 22: 195–217.

Searle, S., G. Casella, and C. McCulloch. 1992. *Variance components*. New York: Wiley.

Snedecor, G., and W. Cochran. 1989. *Statistical methods*. 8th ed. Ames: Iowa State University Press.

Snijders, T., and R. Bosker. 1999. *Multilevel analysis*. London: Sage.

Yates, F. 1967. A fresh look at the basic principles of the design and analysis of experiments. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 4: 777–790.

Varying Coefficient Models

Andros Kourtellos and Thanasis Stengos

Abstract

Varying coefficient models offer a compromise between fully nonparametric and parametric models by allowing for the desired flexibility of the response coefficients of standard regression models to uncover hidden structures in the data without running into the serious curse of the dimensionality issue.

Keywords

Functional coefficient models; Heteroskedasticity; Least squares; Linear regression models; Maximum likelihood; Nonparametric estimation; Parameter heterogeneity; Random coefficients model; Smooth coefficient models; Tuning variables; Varying coefficient models

JEL Classifications

C1

One of the most interesting forms of nonlinear regression models is the varying coefficient model (VCM). Unlike the linear regression model, VCMs were introduced by Hastie and Tibshirani (1993) to allow the regression coefficients to vary systematically and smoothly in more than one dimension. It is worth noting the distinction between the VCM and the so-called random coefficients model, which assumes that the coefficients vary non-systematically (randomly). Versions of the VCM are encountered in the literature as functional coefficient models (see Cai et al. 2000b) and smooth coefficient models (see Li et al. 2002).

VCMs are very useful tools in applied work in economics as they can be used to model parameter heterogeneity in a general way. For example, Durlauf et al. (2001) estimate a version of the Solow model that allows the parameters for each country to vary as functions of initial income. This work is extended in Kourtellos (2005), who finds parameter dependence on initial literacy, initial life expectancy, expropriation risk and ethnolinguistic fractionalization. Li et al. (2002) use the above smooth coefficient model to estimate the production function of the non-metal mineral industry in China. Stengos and Zacharias (2006) use the same model to examine an intertemporal hedonic model of the personal computer market, where the coefficients of the hedonic regression were unknown functions of time. Hong and Lee (2003) forecast the nonlinearity in the conditional mean of exchange rate changes using a VCM that allows the autoregressive coefficients to vary with investment positions. Ahmad et al. (2005) apply the VCM in the estimation of a production function in China's manufacturing industry to show that the marginal productivity of labour and capital depends on the firm's R&D values. Mamuneas et al. (2006) study the effect of human capital on total factor productivity in an empirical growth framework. In what follows we present the basic structure of the standard VCM specification as it appears in the literature and then proceed to discuss certain aspects of estimation and some of its recent generalizations.

Basic Specification

Consider the following VCM

$$y_i = \beta(z_i)'X_i + u_i \tag{1}$$

with $E(u_i|X_i)= 0$, where $X_i = (1, x_{i2}, \ldots , x_{ip})'$ is a p dimensional vector of slope regressors and $\beta(z_i)' = (\beta_1(z_{i1}), \beta_2(z_{i2}), \ldots , \beta_p(z_{ip}))$ is a p dimensional vector of varying coefficients, which take the form of unknown smooth functions of $z_{i1}, z_{i2}, \ldots , z_{ip}$, respectively. Notice that $\beta_1(z_i)$ is a varying intercept that measures the direct relationship between the tuning variable z_i and the dependent variable in a nonparametric way. We refer to the variables z_i's as tuning variables, and they can be one-dimensional or multidimensional. These functions map the tuning variables into a set of local regression coefficient estimates that imply that the effect of X_i on y_i will not be constant but rather it will vary smoothly with the tuning variables. These tuning variables could take the form of a scalar like time or it could be a vector of dimension q. A common situation in the literature arises when the z_j is the same for all j.

It is worth is noting that the VCM (1) is a very flexible and rich family of models. One of the reasons is that the general additive separable structure of (1) offers also a very useful compromise to the general high-dimensional nonparametric regression that is known to suffer from the curse of dimensionality. This allows for nonparametric estimation even when the conditioning regressor space is in high dimensional. Another is that it nests many well-known models as a special case. For instance, consider the following cases. If $\beta_j(z_{ij})= \beta_j$, for all j then we are dealing with the usual linear model. If $\beta_j(z_{ij})= \beta_j z_{ij}$ for some variable j, we simply have the interaction term $\beta_j x_{ij} z_{ij}$ entering the regression function. If $x_i = c$ (a constant) or if $z_{ij} = x_{ij}$ for all $j = 1, \ldots p$, then the model takes the generalized additive form where the additive components are unknown functions (see Hastie and Tibshirani 1990; Linton and Nielsen 1995).

We now set out some estimation issues. A popular estimation approach is based on local polynomial regression, as illustrated by Fan (1992), Fan and Gijbels (1996), and Fan and Zhang (1999), which we present in the context of a random sample design. Given a random sample $\{(z_i, X_i, y_i)\}_{i=1}^n$, the estimation procedure solves a simple local least squares problem. To be precise, for each given point z_0 the functions $\beta_j(z), j = 1 \ldots p$ are approximated by local linear polynomials $\beta_j(z) \approx c_{j0} + c_{j1}(z - z_0)$ for z in a neighborhood of z_0. This leads to the following weighted local least squares problem:

$$\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{p}\{c_{j0} + c_{j1}(z - z_0)\}X_{ij}\right]^2 K_h(z_i - z_0) \tag{2}$$

for a given kernel function K and bandwidth h, where $K_h(\cdot) = K(\cdot/h)/h$. While this method is simple, it is implicitly assumed that the functions $\beta_j(z)$ possess the same degrees of smoothness and hence can be approximated equally well in the same interval. Fan and Zhang (1999) allow for different degrees of smoothness for different coefficient functions by proposing a two-stage method. This is similar in spirit to what Huang and Shen (2004) do for global smoothers using regression splines but allowing each coefficient function to have different (global) smoothing parameters.

An attractive alternative to local polynomial estimation is a global smoothing method based on general series methods such as polynomial splines and trigonometric approximation (see Ahmad et al. 2005; Huang et al. 2004; Huang and Shen 2004; Xue and Yang 2006a). All these papers emphasize the computational savings from having to solve only one minimization problem. Ahmad, Leelahanon and Li stress the efficiency gains of the series approach over a kernel-based approach when one allows for conditional heteroskedasticity. We should note that the inference for the estimated coefficients will differ for different choices of approximation, and the asymptotic properties of such estimators are generally not easy to obtain.

Although the model was initially developed for i.i.d. data, it has been extended for time series data by Chen and Tsay (1993), Cai et al. (2000b), Huang and Shen (2004), and Cai (2007) for strictly stationary processes with different mixing conditions. The coefficient functions typically now become functions of time and/or lagged values of the dependent variable. It is worth noting that estimation issues such as bandwidth selection are similar, as in the i.i.d. data case (see Cai 2007). The varying coefficient model has also been employed to analyse longitudinal data (see Brumback and Rice (1998), Hoover et al. (1998), and Huanget al. (2004).

The Partially Linear Varying Coefficient Model

An interesting special case of eq. (1), where the unknown coefficient functions depend on a common z_i, is the partially linear VCM. Here some of the coefficients are constants (independent of z_i). In that case, eq. (1) can be rewritten as

$$y_i = \alpha' W_i + \beta(z_i)' X_i + u_i \qquad (3)$$

where W_i is the ith observation on a $(1 \times q)$ vector of additional regressors that enter the regression function linearly. The estimation of this model requires some special treatment as the partially linear structure may allow for efficiency gains, since the linear part can be estimated at a much faster rate, namely, $\sqrt{n}$.

The partially linear VCM has been studied by Zhang et al. (2002), Xia et al. (2004), Ahmad et al. (2005), and Fan and Huang (2005). Zhang et al. (2002) suggest a two-step procedure where the coefficients of the linear part are estimated in the first step using polynomial fitting with an initial small bandwidth using cross validation (see Hoover et al. 1998). In other words, the approach is based on under-smoothing in the first stage. Then these estimates are averaged to yield the final first-step linear part estimates which are then used to redefine the dependent variable and return to the environment of eq. (1), where local smoothers can be applied as described above. Alternatively, Xia et al. (2004) separate the estimation of γ from that of $\beta(z_i)$ by noting that the former can be estimated globally, but the latter locally. This is what they call a 'semi-local least squares procedure', and they achieve a more efficient estimate of γ without under-smoothing using standard bandwidth selection methods. Once γ has been estimated, then again the linear part can be used to redefine the dependent variable and return to the environment of eq. (3).

More recently, Fan and Huang (2005) use a profile least squares estimation approach to provide a simple and useful method for (3). More precisely, they construct a Wald test and a profile likelihood ratio test for the parametric component that share similar sampling properties. More importantly, they show that the asymptotic distribution of the profile likelihood ratio test under the null is independent of nuisance parameters, and follows an asymptotic χ^2 distribution. They also propose a generalized likelihood ratio test statistic to test whether certain parametric functions fit the nonparametric varying coefficients. This hypothesis test includes testing for the significance of the slope variables X (zero coefficients) and the homogeneity of the model (constant coefficients). Other work on specification testing includes Li et al. (2002), Cai et al. (2000b), Cai (2007), Yang et al. (2006) that mainly rely on bootstrapping in their implementation.

Generalizations and Extensions

A useful generalization of (1) is to allow the dependent variable to be related to the regression function nonlinearly $m(X_i, Z_i)= \beta(z_i)'X_i$ via some given link function $g(\cdots)$

$$yi = g(\beta(z_i)'X_i) + u_i \qquad (4)$$

This generalization is known as the generalized varying coefficient model and was originally proposed by Hastie and Tibshirani (1993). Cai et al. (2000a) study this model using local polynomial techniques and propose an efficient one-step local maximum likelihood estimator. Notice that if $g(\cdots)$ is the normal CDF then (4) generalizes the standard tool of the discrete choice literature, namely the probit model.

Another strand of the literature allowed for a multivariate tuning variable z_l, $l = 1, 2, \ldots, q$. Although Hastie and Tibshirani (1993) proposed a back-fitting algorithm to estimate the varying coefficient functions, they did not provide any asymptotic justification. The most notable advance in this context has been by Xue and Yang (2006a), who propose a generalization of the VCM as in (1) that allows the varying coefficients to have an additive coefficient structure on regression coefficients to avoid the curse of dimensionality

$$\beta_j(z) = \gamma_{j0} + \gamma_{j1}(z_1) + \cdots + \gamma_{jq}(z_q) \quad \text{for all } j.$$

Under mixing conditions, Xue and Yang (2006a) propose local polynomial marginal integration estimators, while Xue and Yang (2006b) study this model using polynomial splines.

Finally, Cai et al. (2006) have shifted the discussion to consider a structural VCM. They examine the case of endogenous slope regressors, and propose a two-stage IV procedure based on local linear estimation procedures in both stages. We believe that this line of research is fruitful for economic applications.

Conclusion

VCMs have increasingly been employed as useful tools that allow for a compromise between fully nonparametric and parametric models. This compromise allows for the desired flexibility to uncover hidden structures that underlie the response coefficients of standard regression models without running into the serious curse of the dimensionality issue. More importantly, the structure of the VCM that allows the regression coefficients to vary with a tuning variable is very appealing in many economic applications, for it has a natural interpretation of non-constant marginal effects.

See Also

- ▶ Economic Growth Non-linearities
- ▶ Non-parametric Structural Models

Bibliography

Ahmad, I., S. Leelahanon, and Q. Li. 2005. Efficient estimation of a semiparametric partially varying linear model. *Annals of Statistics* 33: 258–283.

Brumback, B., and J. Rice. 1998. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 93: 961–976.

Cai, Z. 2007. Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics* 136: 163–188.

Cai, Z., J. Fan, and R. Li. 2000a. Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* 95: 888–902.

Cai, Z., J. Fan, and Q. Yao. 2000b. Functional coefficient regression models for nonlinear time series models. *Journal of the American Statistical Association* 95: 941–956.

Cai, Z., M. Das, H. Xiong, and Z. Wu. 2006. Functional coefficient instrumental variables models. *Journal of Econometrics* 133: 207–241.

Chen, R., and R. Tsay. 1993. Functional coefficient autoregressive models. *Journal of the American Statistical Association* 88: 298–308.

Durlauf, S., A. Kourtellos, and A. Minkin. 2001. The local Solow growth model. *European Economic Review* 45: 928–940.

Fan, J. 1992. Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87: 998–1004.

Fan, J., and I. Gijbels. 1996. *Local polynomial modelling and its applications*. London: Chapman and Hall.

Fan, J., and T. Huang. 2005. Profile likelihood inferences on semiparametric varying- partially linear models. *Bernoulli* 11: 1031–1057.

Fan, J., and W. Zhang. 1999. Statistical estimation in varying-coefficient models. *Annals of Statistics* 27: 1491–1518.

Hastie, T., and R. Tibshirani. 1990. *Generalized additive models*. New York: Chapman and Hall.

Hastie, T., and R. Tibshirani. 1993. Varying coefficient models. *Journal of the Royal Statistical Society, Series B* 55: 757–796.

Hong, Y., and T.-H. Lee. 2003. Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *The Review of Economics and Statistics* 85: 1048–1062.

Hoover, D., C. Rice, C. Wu, and L. Yang. 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85: 809–822.

Huang, J., and H. Shen. 2004. Functional coefficient regression models for nonlinear time series: A polynomial spline approach. Scandinavian Journal of Statistics 31, 515–534.

Huang, J., C. Wu, and L. Zhou. 2004. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 14: 763–788.

Kourtellos, A. 2005. Modeling parameter heterogeneity in cross-country growth regression models. Mimeo, Department of Economics, University of Cyprus.

Li, Q., C. Huang, D. Li, and T. Fu. 2002. Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics* 20: 412–422.

Linton, O., and J. Nielsen. 1995. A kernel method of estimating structural nonparametric regression based on marginal integration. *Biometrika* 82: 93–100.

Mamuneas, T., A. Savvides, and T. Stengos. 2006. Economic development and the return to human capital: A smooth coefficient semiparametric approach. *Journal of Applied Econometrics* 21: 111–132.

Stengos, T., and E. Zacharias. 2006. Intertemporal pricing and price discrimination: A semiparametric hedonic analysis of the personal computer market. *Journal of Applied Econometrics* 21: 371–386.

Stone, C. 1977. Consistent nonparametric regression. *Annals of Statistics* 5: 595–620.

Xia, Y., W. Zhang, and H. Tong. 2004. Efficient estimation for semivarying-coefficient models. *Biometrika* 91: 661–681.

Xue, L., and L. Yang. 2006a. Estimation of semiparametric additive coefficient model. *Journal of Statistical Planning and Inference* 136: 2506–2534.

Xue, L., and L. Yang. 2006b. Additive coefficient modeling via polynomial spline. *Statistica Sinica* 16: 1423–1446.

Yang, L., B. Park, L. Xue, and W. Härdle. 2006. Estimation and testing for varying coefficients in additive models with marginal integration. *Journal of the American Statistical Association* 101: 1212–1227.

Zhang, W., S.-Y. Lee, and X. Song. 2002. Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis* 82: 166–188.

Veblen Goods

B. Curtis Eaton

Keywords

Conspicuous consumption; Leisure; Leisure class; Relative consumption trap; Status and economics; Upward-sloping demand curves; Veblen goods; Veblen, T.; Well-being

JEL Classifications

B2

The utility that an individual derives from a Veblen good is an increasing function of the individual's consumption of the good *relative* to the consumption of others.

In *The Theory of the Leisure Class*, Thorstein Veblen observed that people value status, and further that in modern societies one's status is determined primarily by one's relative consumption of highly visible goods. 'In order to gain and hold the esteem of men it is not sufficient merely to posses wealth . . . The wealth . . . must be put in evidence, for esteem is awarded only on evidence' (1899, p. 36). The evidence consists of the conspicuous consumption of certain costly goods as prescribed by 'the accredited cannons of [conspicuous] consumption, the effect of which is to hold the consumer up to a standard of expensiveness and wastefulness in his consumption of goods' (1899, p. 116). Veblen was certainly not the first person to articulate the view that esteem can be achieved by conspicuous displays of wealth, but he saw more clearly than others the futility and wastefulness of this form of status seeking.

Following Leibenstein (1950), much of the literature on Veblen goods has focused on the possibility that the demand curve might be upward sloping. The inefficiency or wastefulness associated with Veblen goods is perhaps a more serious matter – see Hopkins and Kornienko (2004) for a theoretical analysis. Veblen seems to have thought that beyond some modest level of affluence societies get caught in what might be called the *relative consumption trap* in which all added productivity is soaked up by the wasteful consumption of Veblen goods with no effect on well-being: 'The need of conspicuous waste . . . stands ready to absorb any increase in the community's industrial efficiency or output of goods, after the most elementary physical wants have been provided for' (1899, p. 110).

The recent literature on perceived well-being suggests that affluent societies may in fact be caught in this trap. A number of studies have shown that the correlation of average well-being and per capita income in affluent societies is very weak, in some cases non-existent. Much of the evidence is surveyed in Robert Frank's (1999) provocative book, *Luxury Fever*. Others have shown that an individual's well-being is negatively associated with the incomes of one's neighbours, and further that the effects on one's well-being of an increase in one's own income and an increase of the same magnitude in the average income of one's neighbours are approximately offsetting (see Luttmer 2005, for example).

With the aid of a simple representative agent model, we can readily see how affluent societies can get stuck in the relative consumption trap. There is a continuum of agents, all of whom have identical preferences and budgets. Preferences of a representative person are captured in the following utility function:

$$U_r(x_r, y_r, v_r) = D(v_r - v) + F(x_r) + G(y_r),$$

where v_r, x_r and y_r are, respectively, quantities of a *pure Veblen good*, leisure, and a standard consumption good, and v is average consumption of the Veblen good. The Veblen good is *pure* in the sense that the utility derived from it, $D(v_r - v)$, is dependent only on relative consumption, $v_r - v$. The functions D, F and G are strictly increasing and concave. Leisure and the standard good are essential, but the Veblen good is not ($D'(0)$ is finite). Each individual is endowed with 1 unit of time to be allocated to leisure and work, and with asset income a. The wage rate is w, and the prices of the Veblen and standard goods are both 1.

For an interior solution to the individual's choice problem, the following marginal conditions must hold:

$$\frac{F'(x_r)}{w} = G'(y_r) = D'(v_r - v).$$

In addition the budget constraint, $wx_r + y_r + v_r = w + a$, will be satisfied.

Since everyone is identical, in equilibrium $v_r = v$, so the conditions that characterize an interior equilibrium are

$$\frac{F'(x^*)}{w} = G'(y^*) = D'(0),$$

and

$$wx^* + y^* + v^* = w + a.$$

Notice that, in equilibrium, the marginal value of the Veblen good, $D'(0)$, is independent of w and a, and since $G'(y^*) = D'(0)$, so too is the equilibrium quantity of the standard good.

What happens as a increases? Clearly, y^* doesn't change, and neither does x^*, since $F'(x^*)/w = G'(y^*)$ and w hasn't changed. So all of the added purchasing power is devoted to the Veblen good, and, since no one's relative consumption of the Veblen good has changed, there will be no change in equilibrium utility.

What happens as w increases? As in the first scenario, y^* doesn't change, but w having increased, x^* must decrease to satisfy $F'(x^*)/w = G'(y^*)$. But this implies that the increase in expenditure on the Veblen good (dv^*) exceeds the increase in full income (dw^*), so in this case *more* than all of the added purchasing power is soaked up by the Veblen good. In addition, since neither y^* nor equilibrium relative consumption of the Veblen good changes, and x^* decreases, equilibrium utility decreases.

So, in this model, if the equilibrium is interior, then

$$\frac{dy^*}{da} = 0, \quad \frac{dx^*}{da} = 0, \quad \frac{dv^*}{da} = 1, \quad \frac{du^*}{da} = 0, \quad \frac{dy^*}{dw} = 0, \quad \frac{dx^*}{dw} < 0, \quad \frac{dv^*}{dw} > 1, \quad \frac{du^*}{dw} < 0.$$

Of course, the equilibrium is not necessarily interior. In particular, since $D'(0)$ is finite, unless the society is sufficiently affluent, in equilibrium nothing is spent on the Veblen good ($v^* = 0$). But once the society is affluent enough so that it begins to squander its resources on the wasteful Veblen good, it is stuck in the relative consumption trap.

See Also

- ▶ Conspicuous Consumption
- ▶ Consumption Externalities
- ▶ Happiness, Economics of

- ► Leisure Class
- ► Social Status, Economics and
- ► Veblen, Thorstein Bunde (1857–1929)

Bibliography

Frank, R. 1999. *Luxury fever: Why money fails to satisfy in an era of excess*. New York: Free Press.

Hopkins, E., and T. Kornienko. 2004. Running to keep in the same place: Consumer choice as a game of status. *American Economic Review* 94: 1085–1107.

Leibenstein, H. 1950. Bandwagon, snob, and Veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics* 64: 183–207.

Luttmer, E.F.P. 2005. Neighbors as negatives: Relative earnings and well-being. *Quarterly Journal of Economics* 120: 963–1002.

Veblen, T. 1899. *The theory of the leisure class: An economic study of institutions*, 1934. New York: The Modern Library.

Veblen, Thorstein Bunde (1857–1929)

Geoffrey M. Hodgson

Abstract

This article outlines the work of the American institutional economist Thorstein Veblen (1857–1929), stressing his critique of neoclassical economics and his development of an alternative, evolutionary approach to the analysis of social, economic and technological change. Veblen's analytical approach to both technology and institutions is discussed, as well as his explicit application of the evolutionary ideas from Darwinian biology to economics.

Keywords

Ayres, C. E.; Clark, J. M.; Commons, J. R.; Conspicuous consumption; Consumer sovereignty; Customs; Darwin, C.; Economic man; Evolutionary economics; Explanation; Habit; Hedonism; Innovation; Institutional economics; James, W.; Mark, K. H.; McDougall, W.; Mitchell, W. C.; Natural selection; Neoclassical economics; Old institutionalism; Peirce, C. S.; Pragmatism; Preferences; Spencer, H.; Sumner, W. G.; Technical change; Utililitarianism; Veblen, T. B.

JEL Classifications

B31

Thorstein Veblen was one of the most influential economists of the early twentieth century and one of the founders of the American school of institutional economics.

Veblen was the fourth son and sixth child of Norwegian immigrants who settled in eastern Minnesota in United States. Educated at Carleton College, Johns Hopkins University, Yale University and Cornell University, he took various university posts at Chicago, Stanford, Missouri and New York. At Johns Hopkins he came in contact with the pragmatist philosopher Charles Sanders Peirce, and at Yale he was influenced by William Graham Sumner. Veblen read widely in biology, psychology and philosophy, as well as the social sciences. The works of Immanuel Kant, Charles Darwin, William James, Karl Marx, William McDougall and Herbert Spencer also made an enduring mark. Despite the popularity of his ideas, Veblen's career was marred by scandal and he never held a senior academic post (Jorgensen and Jorgensen 1999). He died in meagre circumstances in California.

Several of his most important theoretical works date from the 1890s, when he was at the University of Chicago. In 1898 he published his classic article 'Why is Economics Not an Evolutionary Science?' in the *Quarterly Journal of Economics.* The following year saw the appearance of his first book *The Theory of the Leisure Class.* Although this is an original and sophisticated theoretical work, its mockery of the wasteful rich turned it into a bestseller. Other academic articles followed in the *Quarterly Journal of Economics,* the *Journal of Political Economy* and elsewhere, the most important of which have been collected in *The Place of Science in Modern Civilization and Other Essays* (1919). These articles provided a critique of ‘neoclassical’ economics (a term he coined to refer to equilibrium-oriented approaches involving individual utility maximization) and suggestions of a new approach to economics on ‘evolutionary’ and ‘Darwinian’ lines.

He is remembered today as the founder of the school of ‘institutional economics’ that prospered in the United States between the first and second world wars. This school involved leading American economists such as John Maurice Clark, John Rogers Commons and Veblen’s student, Wesley Clair Mitchell.

However, Veblen and his followers did not construct an integrated system of economic theory. This is partly because the original foundations of Veblenian institutionalism were challenged. Pragmatist philosophy, instinct-habit psychology and evolutionary ideas had been foundational for Veblen’s thought. However, by the 1920s they had lost much of their former popularity. Thus, at the high point of its influence, American institutionalism faced fundamental philosophical and theoretical difficulties. After 1940, the ‘old’ institutional economics lost ground to the rising generation of formal and mathematically inclined theorists. By the 1960s the American institutional school was confined to a small minority of adherents. However, in economics in recent years there has been a revival of interest in both evolutionary ideas and the legacy of the ‘old’ institutional school.

Veblen (1919, p. 73) argued that neoclassical economics adopted a faulty and ‘hedonistic’ psychology, involving ‘a passive and substantially inert and immutably given human nature’. He criticized the idea of the individual as a given ‘globule of desire’, lambasting the neoclassical picture of the optimizing and omniscient economic agent as ‘a lightning calculator of pleasures and pains’. He saw this ‘economic man’ as having ‘neither antecedent nor consequent’, lacking an account of how human wants are formed and portraying humans as utility-maximizing automata. Veblen (1914) proposed an alternative theory of human agency, in which ‘instincts’ such as ‘workmanship’, ‘emulation’, ‘predatoriness’ and ‘idle curiosity’ play a major role. Habit and instinct replaced the utilitarian pleasure-pain principle.

Following the pragmatist philosophy of Peirce and James, Veblen rejected the Cartesian notion of the supremely rational and calculating individual, instead seeing agents as propelled in the main by habits and customs. Habits of thought provide the point of view from which facts and events are interpreted. When they are shared and reinforced within a society or group, individual habits assume the form of socioeconomic institutions. In turn, institutions create and reinforce habits of action and thought: ‘The situation of today shapes the institutions of tomorrow through a selective, coercive process, by acting upon men’s habitual view of things, and so altering or fortifying a point of view or a mental attitude handed down from the past’ (Veblen 1899, pp. 190–1).

In *The Theory of the Leisure Class* and elsewhere, he argued that consumption is a ‘conspicuous’ and social process. Through consumption, humans signal status and social position, and thereby stimulate the desires of others. Accordingly, individual tastes are malleable and the idea of unalloyed ‘consumer sovereignty’ is a myth.

Veblen saw conventions, customs and institutions as repositories of social knowledge. Institutional adaptations and behavioural norms were stored in individual habits and could be passed on by education or imitation to succeeding generations. His explanations of economic growth privileged knowledge and institutions, rather than the accumulation of physical assets.

Veblen addressed the ‘evolutionary’ processes of innovation and transformation in a modern economy. Neoclassical theory is defective in this respect because it indicated ‘the conditions of survival to which any innovation is subject, supposing the innovation to have taken place, not the conditions of variational growth’ (Veblen 1919, pp. 176–7). He saw it as important to consider why innovations take place, and not merely to dwell over equilibrium conditions with given technological possibilities. The question for him was not how things stabilize themselves in a ‘static state’, but how they endlessly grow and change.

Veblen saw Darwinian evolutionary principles as crucial to the understanding of the processes of institutional and technological development in a capitalist economy. He was the first economist to argue at length that Darwinian evolutionary principles should be applied to economics. He upheld that economics should become an 'evolutionary' and 'post-Darwinian' science. There is a current revival in 'evolutionary' approaches in economics but the Veblenian precedent for this type of approach is not always acknowledged.

Darwinian evolution involves three essential features. First, there must be sustained variation among the members of a species or population. Variations may be random or purposive in their origin, but without them, as Darwin insisted, natural selection cannot operate. Second, there must be some principle of heredity or continuity involving some mechanism through which individual characteristics are passed on to succeeding generations. Third, natural selection operates because better-adapted organisms leave increased numbers of offspring, or because the variations that are preserved bestow advantage in the struggle for survival.

Veblen applied the same three Darwinian principles to economic evolution. He recognized the role of creativity and novelty with his concept of 'idle curiosity'. Habits and institutions were regarded as relatively durable heritable traits. Concerning selection, Veblen (1899, p. 188) wrote: 'The life of man in society, just as the life of other species, is a struggle for existence, and therefore it is a process of selective adaptation. The evolution of social structure has been a process of natural selection of institutions.' This did not mean that social phenomena were to be explained wholly or largely in biological terms, but that Darwinian principles could be applied to social and economic units and processes.

Veblen saw Darwinian evolutionary processes as open-ended and suboptimal. Unlike advocates of laissez faire, he did not use Darwinian principles to justify market competition. He was critical of apologetic tendencies in social science which regard existing institutions as necessarily efficient or optimal. He described particularly regressive or disserviceable institutions as 'archaic', 'ceremonial' or even 'imbecile'. Furthermore, he used Darwinian ideas to rebut of Marx's teleological suggestions that history was leading inevitably to a communist future.

Like Darwin, Veblen emphasized the importance of processual, causal explanation. Although he did not use the word, he had an appreciation of Darwinian evolution as an 'algorithmic' process. Veblen used phrases such as 'cumulative causation', 'theory of a process, of an unfolding sequence' and 'impersonal sequence of cause and effect' to connote the same idea. This focus on algorithmic processes is revolutionary and modern; it directs attention to ongoing processes rather than static equilibria alone.

Consequently, rather than taking individual reasons or preferences as themselves sufficient to understand motivations, Veblen pointed to the need for causal explanations of reasons or preferences themselves. He did not underestimate the importance of human intentionality – but it had to be explained rather than assumed. Such explanations involved the evolution of social institutions and their interplay with biological and psychological characteristics. He thus acknowledged processes of dual inheritance or coevolution (again to use modern terms) where there was evolution and transmission at both the instinctive and the cultural levels.

Along with the assumption of fixed preference functions, Veblen also criticized the widespread assumption in economic theory of a fixed set of technological possibilities. Technological change can challenge established institutions and vested interests. In *The Theory of Business Enterprise* and elsewhere Veblen distinguished between industry (making goods) and business (making money). This dichotomy parallels the earlier suggestion in *The Theory of the Leisure Class* that there is a distinction between serviceable consumption to satisfy human need and conspicuous consumption for status and display. Subsequently, institutionalists such as Clarence E. Ayres elevated the different conflict between technology and institutions into a universal principle, and dubbed it the 'Veblenian dichotomy'. This is misleading, because Veblen never saw such a conflict as universal, and he saw institutions as the indispensable fabric of economic life (Hodgson 2004).

In the last two decades of the twentieth century, evolutionary and institutional ideas again become prominent in economics. Pragmatism has again become fashionable in philosophy and the concept of habit has returned to psychology. Many of Veblen's ideas, including those on institutional evolution and the role of knowledge in economic growth, now seem strikingly modern. The conditions exist for a deeper appreciation of his contribution to economics and social science.

See Also

- ▶ Ayres, Clarence Edwin (1891–1972)
- ▶ Clark, John Maurice (1884–1963)
- ▶ Commons, John Rogers (1862–1945)
- ▶ Evolutionary Economics
- ▶ Institutionalism, Old
- ▶ Mitchell, Wesley Clair (1874–1948)
- ▶ 'Neoclassical'

Selected Works

1898. Why is economics not an evolutionary science? *Quarterly Journal of Economics* 12: 373–397. Reprinted in Veblen (1919).

1899. *The theory of the leisure class: An economic study of institutions*. New York: Macmillan.

1904. *The theory of business enterprise.* New York: Charles Scribners.

1914. *The instinct of workmanship, and the state of the industrial arts*. New York: Macmillan.

1915. *Imperial Germany and the industrial revolution*. New York: Macmillan.

1918. *The higher learning in America: A memorandum on the conduct of universities by business men*. New York: Huebsch.

1919. *The place of science in modern civilization and other essays*. New York: Huebsch.

1921. *The engineers and the price system.* New York: Harcourt Brace and World.

1923. *Absentee ownership and business enterprise in recent times*. New York: Huebsch.

1934. *Essays on our changing order.* New York: Viking Press.

Bibliography

Daugert, S. 1950. *The philosophy of Thorstein Veblen*. New York: Columbia University Press.

Dorfman, J. 1934. *Thorstein Veblen and his America*. New York: Viking Press.

Hodgson, G. 2004. *The evolution of institutional economics: Agency, structure and Darwinism in American institutionalism*. London: Routledge.

Jorgensen, E., and H. Jorgensen. 1999. *Thorstein Veblen: Victorian Firebrand*. Armonk: M. E. Sharpe.

Rutherford, M. 1994. *Institutions in economics: The old and the new institutionalism*. Cambridge: Cambridge University Press.

Tilman, R. 1996. *The intellectual legacy of Thorstein Veblen: Unresolved issues*. Westport: Greenwood Press.

Vecchio, Gustavo del (1883–1972)

F. Caffè

Del Vecchio was born at Lugo in Romagna on 22 June 1883, and died in Rome on 6 September 1972. He initially attended the university in Rome, where he followed the history of philosophy course under Antonio Labriola. He continued his studies in Bologna, where he was greatly influenced by the teaching of Tullio Martello, follower of Francesco Ferrara's work. His postgraduate studies, which were completed in Berlin, gave Del Vecchio's entire work a wide cultural outlook, influenced by historical, philosophical and sociological factors, as well as purely economic considerations. He became Professor of Political Economy at the Universities of Trieste and Bologna, and Professor of Public Finance at the University of Rome. He lectured at the Bocconi University of Milan, where he was Chancellor from 1934 to 1938. During this last year he was forced to give up his teaching because of the anti-semitic measures adopted by the Fascist government. He went into exile in Switzerland in the latter years of World War II, and on his return to Italy started teaching once again. He was Minister of the Treasury from 1947 to 1950, but these public duties represented only a brief intermission in his life as a dedicated academic.

Del Vecchio's scientific work shows that he constantly tried to unify the tradition of Italian economic thought, whose main personality was Francesco Ferrara, with the theories of equilibrium, whether of the approach suggested by Marshall, or by Walras and Pareto. He was, moreover, profoundly influenced by the work of Maffeo Pantaleoni in the task of constructing an economic dynamics, to be understood not merely as a modification of static analysis, but as a building of a new economic framework. On the one hand, in a series of books which lasted from 1922 to 1950, Del Vecchio realized a unified exposition of political economy, public finance and economic policy, which he believed to be 'successive stages in the passage from a major to a minor level of abstraction in a unique theoretical framework'. On the other hand, on the academic plane, he carried out pioneering analyses which have received wide recognition in the literature (by Schumpeter, Ohlin, Knight and Stigler, among others). In particular, his research into the application of the marginal principle to money can be traced back to 1909; this research was started with the previously scarcely recognized Walrasian analysis of money, but criticizing some aspects of it and carrying out original developments. Among his important early works are his analyses of the process of the formation of savings which (in 1915) he linked not to the interest rate (the generally held view) but to the quality of income, that is, to its sources. He also carried out important research into the process of accumulation which he felt could not be explained purely in terms of economic factors; he believed that in order to obtain a realistic understanding of the whole accumulation process, it was necessary that non-economic factors should also be taken into account. His contributions to the pure theory of international trade, to the concept of risk as an uncertainty related to the passing of time, and to the empirical investigation into consumer behaviour by means of the investigations of relations between income and consumption have all been recognized. From all his contributions emerges the need to analyse the economy from a broader perspective, avoiding the aridity of abstraction and formalism.

Selected Works

1909. I principi della teoria economica della moneta. *Giornale degli Economisti*.

1912. Relazioni tra entrata e consumo. *Giornale degli Economisti*.

1915. *Lineamenti generali della teoria dell'interesse*. Rome: Athenaeum.

1928. Teoria economica dell'assicurazione. *Annali di Economia dell'Universita Bocconi*, Milan.

1936. *Progressi della teoria economica*. Padua: Cedam.

1930. *Grundlinien der Geldtheorie*. Tübingen: J.C.B. Mohr (Paul Siebeck).

1932. *Ricerce sopra le teoria generale della moneta*. Milan: Università Bocconi.

1956. *Vecchie e nuovo teorie economiche*. Turin: Utet.

1961. *Economia generale*. Turin: Utet.

1983. *Anthology of the Writings of Gustavo del Vecchio on the Centenary of His Birth*. Milan: F. Angeli.

Vector Autoregressions

Tao Zha

Abstract

Vector autoregressions are a class of dynamic multivariate models introduced by Sims (1980) to macroeconomics. These models have been primarily used to bring empirical regularities out of the time series data, to provide forecasting and policy analysis, and to serve as a benchmark for model comparison. Economic applications often impose more restrictions on vector autoregressions than originally thought necessary. Recent econometric developments have made it feasible to handle vector autoregressions with a wide class of restrictions and have narrowed the gap between these models and dynamic stochastic general equilibrium models.

Keywords

Bayesian econometrics; Bayesian priors; Cowles commission; Dynamic multivariate models; Dynamic stochastic general equilibrium models; Gibbs samplers; Identification; Impulse responses; Likelihood function; Marginal data density; Markov chain Monte Carlo method; Markov-switching vector autoregressions; Probability density function; Recursive identification; Structural shocks; Structural vector autoregressions; Vector autoregressions

JEL Classifications

C3

Vector autoregressions (VARs) are a class of dynamic multivariate models introduced by Sims (1980) to macroeconomics. These models arise mainly as a response to the 'incredible' identifying assumptions embedded in traditional large-scale econometric models of the Cowles Commission. The traditional approach uses predetermined or exogenous variables, coupled with many strong exclusion restrictions, to identify each structural equation. VARs, by contrast, explicitly recognize that all economic variables are interdependent and thus should be treated endogenously. The philosophy of VAR modelling begins with a multivariate time series model that has minimal restrictions, and gradually introduces identifying information, with emphasis always placed on the model's fit to data.

While the traditional econometric approach allows disturbances or shocks to structural equations to be correlated, the VAR methodology insists that structural shocks ought to be independent of one another. The independence assumption plays an essential role in achieving unambiguous economic interpretations about structural shocks such as technology and policy shocks; it can be tested using recently developed econometric tools (Leeper and Zha 2003). The bulk of VAR work has focused on identifying structural shocks as a way to specify the contemporaneous relationships among economic variables. With most dynamic relationships unrestricted, the intent of such an identifying strategy is to construct models that have both economic interpretability and superior fit to data. Dynamic responses to a particular shock, called impulse responses, are often used as economic interpretations to the model. They summarize the properties of all systematic components of the system and have become a major tool in modern economic analysis.

Modelling policy shocks explicitly is important in addressing the practical importance of the Lucas critique. If policy switches regime, such a change may be viewed as a sequence of random shocks from the public's viewpoint (Sims 1982). If this sequence displays a persistent pattern, the public will adjust its expectations formation accordingly and the Lucas critique may be consequential. For the practice of monetary policy, however, it is an empirical question how significant this adjustment is. Leeper and Zha (2003) construct an econometric measure from the sequence of policy shocks implied by regime switches to gauge whether the public's behaviour could be well approximated by a linear model. This measure is particularly useful if counterfactual exercises regarding the effects of policy changes are conducted with respect to the Lucas critique.

VARs have also been used for other tasks. Armed with a Bayesian prior, VARs have been known to produce out-of-sample forecasts of economic variables as well as, or even better than, those from commercial forecasting firms (Litterman 1986; Geweke and Whiteman 2006). Because of their ability to forecast, VARs have given researchers a convenient diagnostic tool to assess the feasibility or plausibility of real-time policy projections of other economic models (Sims 1982). VARs have been increasingly used for policy analysis and as a benchmark for comparing different dynamic stochastic general equilibrium (DSGE) models. Restrictions on lagged coefficients have been gradually introduced to give more economic interpretations to individual equations. All these developments are positive and help narrow the gap between statistical and economic models.

This article discusses these and other aspects of VARs, summarizes some key theoretical results for the reader to consult without searching for different sources, and provides a perspective on where future research in this area will be headed.

General Framework

Structural Form

VARs are generally represented in a structural form of which the reduced form is simply a byproduct. The general form is

$$\mathbf{y}_t'\mathbf{A} = \sum_{l=1}^{p} \mathbf{y}_{t-l}'\mathbf{A}_l + \mathbf{Z}_{ti}'\mathbf{D} + \varepsilon_t', \qquad (1)$$

where $\mathbf{y}_t$ is an $n \times 1$ column vector of endogenous variables, $\mathbf{A}$ and $\mathbf{A}_l$ are $n \times n$ parameter matrices, $\mathbf{z}_t$ is an $h \times 1$ column vector of exogenous variables, $\mathbf{D}$ is an $h \times n$ parameter matrix, p is the lag length, and ε_t is an $n \times 1$ column vector of structural shocks. The parameters of individual equations in (1) correspond to the columns of $\mathbf{A}$, $\mathbf{A}_l$, and $\mathbf{D}$. The structural shocks are assumed to be i.i.d. and independent of one another:

$$E(\varepsilon_t | \mathbf{y}_{t-s}, s > 0) = \underset{n\times 1}{0}, \quad E(\varepsilon_t {\varepsilon'}_t | \mathbf{y}_{t-s}, s > 0) = \underset{n\times n}{\mathbf{I}},$$

where $\mathbf{0}_{n\times n}$ is the $n \times n$ matrix of zeros and $\mathbf{I}_{n\times n}$ is the $n \times n$ identity matrix. It follows that the reduced form of (1) is

$$\mathbf{y}_t' = \sum_{l=1}^{p} \mathbf{y}_{t-l}'\mathbf{B}_l + \mathbf{z}_t'\mathbf{C} + u_t', \qquad (2)$$

where $\mathbf{B}_l = \mathbf{A}_l\mathbf{A}^{-1}$, $\mathbf{C} = \mathbf{D}\mathbf{A}^{-1}$, and $u_t' = \varepsilon_t'\mathbf{A}^{-1}$. The covariance matrix of u_t is $\Sigma = (\mathbf{A}\mathbf{A}')^{-1}$

In contrast to the traditional econometric approach, the VAR approach puts emphasis almost exclusively on the dynamic properties of endogenous variables $\mathbf{y}_t$ rather than exogenous variables $\mathbf{z}_{t.}$ In most VAR applications, $\mathbf{z}_t$ simply contains the constant terms.

Identification

One main objective in the VAR literature is to obtain economically meaningful impulse responses to structural shocks ε_t. To achieve this objective, it is necessary to impose at least $n(n - 1)/2$ identifying restrictions, often on the contemporaneous coefficients represented by $\mathbf{A}$ in the structural system (1). In his original work, Sims (1980) makes the contemporaneous coefficient matrix $\mathbf{A}$ triangular for identification. The triangular system, often called the recursive identification, has a 'Wold chain causal' interpretation which is based on the timing of how shocks affect variables contemporaneously. It assumes that some shocks may influence only a subset of variables within the current period. This identification is still popular because it is straightforward to use and can yield some results that match widely held views. Christiano et al. (1999) discuss extensively how recursive identification can be used in policy analysis.

There are fundamental economic applications that require identification under alternative assumptions rather than the recursive system. One familiar example is the determination of price and quantity as discussed in Sims (1986) and Gordon and Leeper (1994). Both variables are often determined *simultaneously* by the supply and demand equations in equilibrium; this simultaneity is inconsistent with recursive identification. Bernanke (1986) and Blanchard and Watson (1986) pioneered other applications of non-recursive identified VARs. Estimation of non-recursive VARs presents technical difficulties that are absent in recursive systems. These difficulties help explain the use of recursive VARs even if this maintained assumption is implausible. Recent developments in Bayesian econometrics, however, have made it feasible to estimate non-recursive VARs.

All of these works focus on the contemporaneous coefficient matrix. There are other ways to achieve identification. Blanchard and Quah (1993) and Gali (1992) propose using identifying restrictions directly on short-run and long-run impulse responses, which have been used in quantifying the effects of technology shocks and various nominal shocks, although the unreliable statistical properties of long-run restrictions are documented by Faust and Leeper (1997).

Many VAR applications rely on exact identification: the number of identifying restrictions equals$n(n-1)/2$. This counting condition is necessary but *not* sufficient for identification. To see this point, consider a three-variable VAR with the following restrictions

$$\mathbf{A} = \begin{bmatrix} * & * & 0 \\ 0 & * & * \\ * & 0 & * \end{bmatrix}$$

where *'s indicate unrestricted coefficients and 0's indicate exclusion restrictions. This VAR is *not* identified because in general there exist two distinct sets of structural parameters that deliver the same dynamics of $\mathbf{y}_t$. For larger and more complicated systems with both short-run and long-run restrictions, there has been, until recently, no practical guidance as to whether the model is identified. The paper by Rubio-Ramirez et al. (2005) develops a theorem for a necessary and sufficient condition for a VAR to be exactly identified. This theorem applies to a wide range of identified VARs, including those used in the literature. The basic idea is to transform the original structural parameters to the $(np + h) \times n$ matrix F (which is a function of $\mathbf{A}, \mathbf{A}_1, \ldots \mathbf{A}_p, \mathbf{D}$) so that linear restrictions can be applied to each column of F. The linear restrictions for the ith column of F can be summarized by the matrix Q_i of rank q_i, where q_i is the number of restrictions. According to their theorem, the VAR model is exactly identified if and only if$q_i = n - i$ for $1 \leq i \leq n$. This result gives the researcher a practical way to determine whether a VAR model is identified.

When the number of identifying restrictions is greater than $n(n-1)/2$, a VAR is over-identified. Allowing for over-identification is important since economic theory often implies more than $n(n-1)/2$ restrictions. Moreover, many economic applications call for restrictions on the model's parameters beyond the contemporaneous coefficients (Cushman and Zha 1997). Restrictions on the lag structure, such as block recursions, offer an effective way to handle over-parameterization when the lag length is long (Zha 1999). Classical or Bayesian econometric procedures can be used to test over-identifying restrictions. A review of theoretical results for Bayesian estimation and inference for both exactly identified and over-identified VARs is discussed below.

Impulse Responses

Impulse responses are most commonly used in the VAR literature and are defined as $\partial_{\mathbf{y}_{t+s}}/\partial\varepsilon'_t$ for$s \geq 0$. Let Φ_s be the $n \times n$ impulse response matrix at step s and the ith row of Φ_s be responses of the n endogenous variables to the ith one-standard-deviation structural shock. One can show that the impulse responses can be recursively updated as

$$\Phi_s = \Phi_{s-1}\mathbf{B}_1 + \ldots + \Phi_{s-p}\mathbf{B}_p \qquad (3)$$

with the convention that $\Phi_0 = \mathbf{A}^{-1}$ and $\Phi_\upsilon = \mathbf{0}_{n\times n}$ for $\upsilon < 0$.

The concept of impulse response is economically appealing and is used in strands of literature other than VAR work. For example, impulse responses to technology shocks or monetary policy shocks in a DSGE have been often compared to those in a VAR model. In empirical monetary economics, impulse responses of various macroeconomic variables to policy shocks have been a focal point in the recent debate on the effectiveness of monetary policy. These shocks can be thought of as shifts (deviations) from the systematic part of monetary policy that are hard to predict from the viewpoint of the public.

It is sometimes argued that identified VARs are unreliable because certain conclusions are sensitive to the specific identifying assumptions. This argument is a sophism. All economic models, DSGE model and VARs alike, are founded on 'controversial' assumptions, and the results can be sensitive to these assumptions. What researchers should do is to select a class of models based on how well they fit to the data, analyse how reasonable the underlying assumptions are, and examine whether there are robust conclusions across models.

Christiano et al. (1999) and Rubio-Ramirez et al. (2005) show some important robust results across different VAR models that have reasonable assumptions and fit to the data equally well. One prominent example is the robust conclusion that a large fraction of the variation in policy instruments, such as the short-term interest rate, can be attributed to the systematic response of policy to shocks originating from the private economy. Such a conclusion is expected of good monetary policy, but it also explains the subtle and difficult task of identifying monetary policy shocks separately from the other shocks affecting the economy.

Estimation and Inference

Bayesian Prior

When one estimates a VAR model for macroeconomic time series data, there is a trade-off between using short and long lags. A VAR with a short lag is prone to misspecification, and a VAR with a long lag length is likely to suffer from the over-fitting problem. The Bayesian prior proposed by Sims and Zha (1998) is designed to eliminate the over-fitting problem without reducing the dimension of the model. It applies to not only reduced-form but also identified VARs.

To describe this prior simply, let $\mathbf{z}_t$ contain only a constant term and thus $\mathbf{D}$ is a 1 × n vector of parameters. Rewrite the structural system (1) in the compact form of$\mathbf{y}_t'\mathbf{A} = \mathbf{X}_t'\mathbf{F} + \varepsilon_t'$, where

$$\underset{1\times k}{\mathbf{x}_t'} = \left[\mathbf{y}_{t-1}'\Lambda\mathbf{y}_{t-p}'\mathbf{z}_t'\right], \quad \underset{n\times k}{\mathbf{F}'} = \left[\mathbf{A}'\Lambda\mathbf{A}_p'\mathbf{D}'\right],$$

and $k = np + h$. For $1 \leq j \leq n$, let $\mathbf{a}_i$ be the jth column of $\mathbf{A}$ and $\mathbf{f}_{\mathrm{i}}$ be the jth column of $\mathbf{F}$. The first component of the prior is that $\mathbf{a}_j$ and $\mathbf{f}_i$ have Gaussian distribution

$$\mathbf{a}_j \sim N(0, \mathbf{S}) \quad \text{and} \quad \mathbf{f}_j|\mathbf{a}_j \sim N\left(\mathbf{P}\mathbf{a}_j, \mathbf{H}\right), \qquad (4)$$

where $\mathbf{P}_{n\times k}' = [\mathbf{I}_{n\times n} \quad \mathbf{0}_{n\times n} \quad \ldots \quad \mathbf{0}_{n\times n} \quad \mathbf{0}_{n\times 1}]$, which is consistent with the reduced-form random walk prior of Litterman (1986). The covariance matrices $\mathbf{S}$ and $\mathbf{H}$ are assumed to be diagonal matrices and are treated as hyperparameters. In principle, one could estimate these hyperparameters or integrate them out in a hierarchical framework. In practice, the values of these hyperparameters are specified before estimation. The ith diagonal element of $\mathbf{S}$ is λ_0/σ_i. The diagonal element of $\mathbf{H}$ that corresponds to the coefficient on lag 1 of variable i in equation j is $\left(\lambda_0\lambda_1{\lambda_2}^{\delta(i,j)}\right)/\left(\sigma_i l^{\lambda_3}\right)$ where $\delta(i, j)$ equals 1 if $i = j$ and 0 otherwise. The diagonal element of $\mathbf{H}$ corresponding to the constant term is the square of $\lambda_0\lambda_4$. The hyperparameter λ_0 controls the overall tightness of belief about the random walk feature, as well as tightness on the prior of $\mathbf{A}$ itself; λ_1 further controls the tightness of belief on random walk and the relative tightness on the prior of lagged coefficients; λ_2 controls the influence of variable i in equation j; λ_3 controls the rate at which the influence of lag decreases as its length increases; and λ_4 controls the relative tightness on the zero value of the constant term. The hyperparameters σ_i are scale factors to make the units uniform across variables, and are chosen at the sample standard deviations of residuals from univariate autoregressive models fitted to the individual time series in the sample (Litterman 1986).

A VAR with many variables and a long lag is likely to produce relatively large coefficient estimates on distant lags and thus volatile sampling errors. The prior described here is designed to reduce the influence of distant lags and the unreasonable degree of explosiveness embedded in the system. It is essential for ensuring reasonable small-sample properties of the model, especially when there are relatively few degrees of freedom in a large VAR.

The aforementioned prior, however, does not take into account the features of unit roots and cointegration relationships embedded in many time series. For this reason, Sims and Zha (1998) add another component to their prior. This component uses Litterman's idea of dummy observations to express beliefs on unit roots and cointegration. Specifically, there are $n + 1$ dummy observations added to the original system, which can be written as

$$\mathbf{Y}_d\mathbf{A}{=}\mathbf{X}_d\mathbf{F}{+}\mathbf{E}, \qquad (5)$$

where $\mathbf{E}$ is a matrix of random shocks,

$$\underset{(n+1)\times n}{\mathbf{Y}_d} = \begin{bmatrix} \mu_5\bar{y}_1^0 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mu_5\bar{y}_n^0 \\ \mu_6\bar{y}_1^0 & \dots & \mu_6\bar{y}_n^0 \end{bmatrix}, \quad \underset{(n(n+1)\times 1}{\mathbf{c}_d} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mu_6 \end{bmatrix},$$

$$\underset{(n+1)\times(np+1)}{\mathbf{X}_d} = [\,\mathbf{Y}_d \quad \dots \quad \mathbf{Y}_d \quad \mathbf{c}_d\,],$$

and $\bar{y}_i^0$ is the sample average of the p initial conditions for the ith variable of $\mathbf{y}_t$ and μ_5 and μ_6 are hyperparameters. The first $n + 1$ dummy-observation equations in (5) express beliefs that all variables are stationary with means equal to $\bar{y}_i^0$'s or cointegration is present. The larger the values of μ_5 and μ_6, the stronger these beliefs.

Since the values of λ's and μ's move in opposite directions to increase or loosen the tightness of the prior, the two symbols λ and μ are kept distinct. In applied work, the values of the hyperparameters for quarterly data are typically set to $\lambda_0 = 1, \lambda_1 = 0.2$, and $\lambda_2 = \lambda_3 = \lambda_4 = \mu_5 = \mu_6 = 1.0$. For monthly data, $\lambda_0 = 0.6$, $\lambda_1 = 0.1$, $\lambda_2 = 1.0$, $\lambda_4 = 0.1$, and $\mu_5 = \mu_6 = 5.0$, while the choice of the lag decay weight λ_3 is somewhat complicated and is elaborated in Robertson and Tallman (1999).

By taking into account the cointegration relationships among macroeconomic variables, this additional component of the prior helps improve out-of-sample forecasting, reduces the difference in forecasting accuracy between using the vintage and final data, and produces robust impulse responses to monetary policy shocks across VARs with different identification assumptions (Robertson and Tallman 1999, 2001). Furthermore, Leeper et al. (1996) demonstrate that with this prior it is feasible to estimate VAR models with as many as 18 variables – far more than the current DSGE models can handle. Because the prior proposed by Sims and Zha (1998) reflects *widely held* beliefs in the behaviour of macroeconomic time series, it has been often used as a base line prior in the Bayesian estimation and inference of VAR models.

Marginal Data Density

If a model is used as a candidate for the 'true' data-generating mechanism, it is imperative that the model's fit to the data is superior to those of alternative models. Recent developments in Bayesian econometrics have made it feasible to compare nested and non-nested models for their fits to the data (Geweke 1999). With a proper Bayesian prior, one can numerically compute the marginal data density (MDD) defined as

$$\int_\Theta L(\mathbf{Y}_T|\,\varphi)p(\varphi)d\varphi, \tag{6}$$

where φ is a collection of all the model's parameters, Θ is the domain of φ, $\mathbf{Y}_T$ is all the data up to T, and $L(\mathbf{Y}_T|\,\varphi)$ is the proper likelihood function. To determine the goodness of fit of a DSGE model, for example, one can compare its MDD with that of a VAR model (Smets and Wouters 2003; Del Negro and Schorfheide 2004).

As a VAR is often used as a benchmark for comparing different models, it is important that one compute its MDD efficiently and accurately. For an unrestricted reduced-form VAR as specified in (2), there is a standard closed-form expression for (6) so that no Markov chain Monte Carlo (MCMC) method is needed to obtain the MDD. For restricted (tightly parameterized) VARs implied by a growing number of economic applications, there is in general no closed-form solution to (6), and a numerical approximation to (6) is needed. Because of a high dimension in the VAR parameter space and possible simultaneity in an identified model, popular MCMC approaches such as importance sampling and modified harmonic mean methods require a long sequence of posterior draws to achieve numerical reliability in approximating (6), and thus are computationally very demanding.

Chib (1995) offers a procedure for accurate evaluations of the MDD that requires the existence of a Gibbs sampler by partitioning φ into a few blocks. One can sample alternately from the conditional posterior distribution of one block of parameters given other blocks. While sampling between blocks entails additional simulations, the Chib algorithm can be far more efficient than other methods because each conditional posterior probability density function (PDF) can be evaluated in closed form. The objects needed to complete this algorithm are the closed-form prior PDF and the conditional posterior PDF for each block.

Because the prior discussed so far includes the dummy observations component, there is a question as to whether this overall prior has a standard PDF. To answer this question, it can be shown from (4) and (5) that the overall prior PDF is

$$\mathbf{a}_j \sim N\left(0\overline{\mathbf{S}}\right) \quad \text{and} \quad \mathbf{f}_j|\mathbf{a}_j \sim N\left(\mathbf{P}\mathbf{a}_j,\overline{\mathbf{H}}\right), \qquad (7)$$

where $\overline{\mathbf{S}}=\mathbf{S}$ and $\overline{\mathbf{H}}=\left(\mathbf{X}_d'\mathbf{X}_d+\mathbf{H}^{-1}\right)^{-1}$. The result (7) follows from the two claims:

$$\left(\mathbf{X}_d'\mathbf{X}_d+\mathbf{H}^{-1}\right)^{-1}\left(\mathbf{X}_d'\mathbf{Y}+\mathbf{H}^{-1}\mathbf{P}\right)=\mathbf{P};$$

$$\mathbf{Y}_d'\mathbf{Y}_d+\mathbf{P}'\mathbf{H}^{-1}\mathbf{P}=\left(\mathbf{Y}_d'\mathbf{X}_d+\mathbf{P}'\mathbf{H}^{-1}\right)\mathbf{P}.$$

Given the prior (7), Waggoner and Zha (2003a) develop a Gibbs sampler for identified VARs with the linear restrictions studied in the VAR literature. These restrictions can be summarized as

$$\underset{n\times n}{\mathbf{Q}_j}\,\mathbf{a}_j = \underset{n\times 1}{0}, \quad \underset{n\times k}{\mathbf{R}_j}\,\mathbf{f}_j = \underset{n\times 1}{0}; \quad j=1,\ldots n. \qquad (8)$$

If there are q_j restrictions on $\mathbf{a}_j$ and r_j restrictions on $\mathbf{f}_j$, the ranks of $\mathbf{Q}_j$ and $\mathbf{R}_j$ are q_j and r_j respectively. Let $\mathbf{U}_j$ ($\mathbf{R}_j$) be an $n \times q_j$ ($n \times r_j$) matrix whose columns form an orthonormal basis for the null space of $\mathbf{Q}_j$ ($\mathbf{R}_j$). The conditions in (8) are satisfied if and only if there exist a $q_j \times 1$ vector $\mathbf{b}_j$ and an $r_j \times 1$ vector $\mathbf{g}_j$ such that $\mathbf{a}_j = \mathbf{U}_j\mathbf{b}_j$ and $\mathbf{f}_j = \mathbf{V}_j\mathbf{g}_j$. The vectors $\mathbf{b}_j$ and $\mathbf{g}_j$ are the free parameters of $\mathbf{a}_j$ and $\mathbf{f}_j$ dictated by the conditions in (8). It follows from (7) that the prior distribution of $\mathbf{b}_j$ and $\mathbf{g}_j$ is jointly normal.

As for the conditional posterior PDFs, it can be shown that the posterior distribution of $\mathbf{g}_j$ conditional on $\mathbf{b}_j$ is normal and that the posterior distribution of $\mathbf{b}_j$ conditional on $\mathbf{b}_i$'s for $i \neq j$ has a closed-form PDF and can be simulated from it exactly. These results enable one to use the efficient method of Chib (1995). The MDD calculated this way is reliable and requires little computing time. For example, it takes less than one minute to obtain a very reliable estimate of the MDD for a large VAR with 13 lags and 10 variables. Such accuracy and speed make it feasible to compare a large number of identified VARs with different degrees of restriction.

Error Bands

Because impulse responses are of central interest in interpreting dynamic multivariate models and helping guide the directions for new economic theory to be developed (Christiano et al. 2005), it is essential that measures of the statistical reliability of estimated impulse responses be presented as part of the process of evaluating models. The Bayesian methods reviewed so far in this essay make it feasible to construct the error bands around impulse responses. The error bands can contain any probability and are typically expressed in both .68 and .90 probability bands to characterize the shapes of the likelihood implied by the model.

The error bands of impulse responses reported in most VAR works are constructed as follows. One begins with the Gibbs sampler to draw $\mathbf{b}_j$ and $\mathbf{g}_j$ for $j = 1, \ldots n$. For each posterior draw, the free parameters $\mathbf{b}_j$'s and $\mathbf{g}_j$'s are transformed to the original structural parameters $\mathbf{A}$, $\mathbf{A}_l$ $(1 = 1, \ldots p)$, and $\mathbf{D}$; then the impulse responses are computed according to (3). The empirical distribution for each element of the impulse responses is formed and the equal-tail .68 and .90 probability intervals around each element are computed. The probability intervals have exact small-sample properties from a Bayesian point of view; and .90 or .95 probability intervals have been used in the empirical literature to approximate classical small-sample confidence intervals when the high dimensional parameter space and a large number of nuisance parameters make it difficult or impossible to obtain exact classical inferences.

One issue related to the error bands around impulse responses, whose importance is beginning to be recognized, is normalization. A normalization rule selects the sign of each draw of impulse responses from the posterior distribution. If there is no restriction imposed on the sign of each column of the contemporaneous coefficient matrix $\mathbf{A}$, then the likelihood or the posterior function remains the same when the sign of a column of $\mathbf{A}$ is reversed. Without any sign restriction, the error bands for impulse responses would be symmetric around zero and thus the estimated responses would be determined to be imprecise.

The conventional normalization is to keep the diagonal of $\mathbf{A}$ always positive, based on the notion that a choice of normalization cannot have substantive effects on the results. But this notion is mistaken. If an identified VAR is non-recursive, normalization can generate ill-determined or unreasonably wide error bands around some impulse responses because some coefficients on the diagonal may be insignificantly different from zero.

Waggoner and Zha (2003b) show that normalized likelihoods can be different across normalization rules and that inappropriate normalization tends to produce a multi-modal likelihood. They propose a normalization rule designed to prevent the normalized likelihood from being spuriously multi-modal and thus avoid unreasonably wide error bands caused by the multi-modal likelihood. The algorithm for their normalization is straightforward to implement: for each posterior draw of $\mathbf{a}_j$, keep $\mathbf{a}_j$ if $\mathbf{e}_j'\mathbf{A}^{-1}\hat{\mathbf{a}}_j > 0$ and replace $\mathbf{a}_j$ with $\mathbf{a}_j$ if $\mathbf{e}_j'\mathbf{A}^{-1}\hat{\mathbf{a}}_j < 0$, where $\mathbf{e}_j$ is the jth column of the $n \times n$ identity matrix. This algorithm works for not only short-run but also long-run restrictions (Evans and Marshall 2002).

Another important issue related to error bands, not addressed until recently, is the characterization of the uncertainty around estimated impulse responses not only at one particular point but also around the shape of the responses as a whole. Let $\Phi_s(i, j)$ be the s-step impulse response of the jth variable to the ith structural shock. The associated error band is only pointwise. It is very unlikely in economic applications, however, that uncertainty about $\Phi_s(i, j)$ is independent across j or s. For example, the response of output to a policy shock is likely to be negatively correlated with the response of unemployment, and the response of inflation this period is likely to be positively correlated with the previous and next responses.

The procedure proposed by Sims and Zha (1999) takes into account these possible correlations across variables and across time. To use this procedure, one can simply stack all the relevant impulse responses into a column vector denoted by $\tilde{\mathbf{c}}$, where the tilde refers to a posterior draw. From a large number of posterior draws, the mean $\mathbf{c}$ and covariance matrix $\overline{\Omega}$ of $\tilde{\mathbf{c}}$ are computed. For each posterior draw $\tilde{\mathbf{c}}$ the kth component $\widetilde{\gamma}_k = (\tilde{\mathbf{c}} - \overline{\mathbf{c}})'\overline{\mathbf{w}}_k$ is calculated, where $\overline{\mathbf{w}}_k$ is the eigenvector corresponding to the kth largest eigenvalue of $\overline{\Omega}$. From the empirical distribution of $\widetilde{\gamma}_k$, one can tabulate different quantiles such as $\gamma_{k,.16}$ and $\gamma_{k,.84}$. Thus, the .68 probability error bands explained by the kth component of variation in the group of impulse responses can be computed as $\mathbf{c}_{.16} = \overline{\mathbf{c}} + \gamma_{k,.16}\overline{\mathbf{w}}_k$ and $\mathbf{c}_{.84} = \overline{\mathbf{c}} + \gamma_{k,.84}\overline{\mathbf{w}}_k$. For a particular economic application, if it turns out that only one to three eigenvalues dominate the covariance matrix of $\tilde{\mathbf{c}}$, these kinds of connecting-dots error bands can be useful in understanding the magnitudes and directions of uncertainty among a group of interrelated impulse responses. This method has proven to be particularly useful in economic applications that characterize the uncertainty around the entire paths, not just points one at a time (Cogley and Sargent 2005; Nason and Rogers 2006).

Markov-Switching VARs

The class of VARs discussed thus far assumes that the parameters are constant over time. This assumption is made mainly for the technical constraint on estimation and inference, however. Many macroeconomic time series display patterns that seem impossible to capture by constant-parameter VARs. One prominent example is changes in volatility over time. In the VAR framework, volatility changes mean that the reduced-form covariance matrix Σ is not constant. In policy analysis, there is a serious debate on whether the coefficients in the policy rule have changed over time, or whether the variances of shocks in the private sector have changed over time, or both. Time-varying VARs are designed to answer these kinds of questions. Stock and Watson (2003) use the reduced-form VAR framework to show that fluctuations in US business cycles can be largely explained by changes in Σ. Sims and Zha (2006b) identify the behaviour of monetary policy from the rest of the VAR system and show that changes in the coefficients in monetary policy are, at most, modest and the variance changes in shocks originating from the private sector dominate aggregate fluctuations.

There have been a number of studies on time-varying VARs that allow the coefficients or the covariance matrix of residuals or both to change over time. These models typically let all the coefficients drift as a random walk or persistent process. To the extent that this kind of modelling tries to capture possible changes in the model's parameters, the model tends to over-fit because the dimension of time variation embedded in the data is much lower than the model's specification. Conceptually, there is a problem of distinguishing shocks to the residuals from shocks to the coefficients. The inability to distinguish among these shocks makes it difficult to interpret the effects of, say, monetary policy shocks.

The Markov-switching VAR introduced by Sims and Zha (2006a) is designed to overcome the over-fitting problems present in the other time-varying VARs and, at the same time, maintain clear interpretation of structural shocks. It builds on the Markov-switching model of Hamilton (1989), but emphasizes ways to restrict the degree of time variation allowed in the VAR. It has a capability to approximate parameter drifts arbitrarily well with the growing number of states, while restricting the transition matrix to be concentrated on the diagonal. This feature also allows discontinuous jumps from one state to another, which appears to matter for aggregate fluctuations.

To see how this method works, suppose that the parameter z_t drifts according to the process $z_t = \rho z_{t-1} + v_t$ where $v_t \sim N(0, \sigma^2)$. By discretizing this autoregressive process, one can let the probability of the transition from state j to i be proportional to

$$\Pr\left[z_t \in \left(\frac{\tau_i \sigma}{\sqrt{1-\rho^2}}, \frac{\tau_{i+1}\sigma}{\sqrt{1-\rho^2}}\right) | z_{t-1} = \frac{\tau_j + \tau_{j+1}}{2} \frac{\rho\sigma}{\sqrt{1-\rho^2}}\right]$$
$$= \Psi\left(\frac{\tau_{i+1}}{\sqrt{1-\rho^2}} - \frac{\tau_j + \tau_{j+1}}{2} \frac{\rho}{\sqrt{1-\rho^2}}\right)$$
$$-\Psi\left(\frac{\tau_i}{\sqrt{1-\rho^2}} - \frac{\tau_j + \tau_{j+1}}{2} \frac{\rho}{\sqrt{1-\rho^2}}\right),$$

where Ψ( ) is the standard normal cumulative probability function. The values of τ divide up the interval between − 2 and 2 (two standard deviations). For nine states, for example, one has $\tau_1 = 2$, $\tau_2 = 1.5$, $\tau_3 = 1, \ldots, \tau_8 = 1.5$, and $\tau_9 = 2$. Careful restrictions on the degree of time variation, as well as on the constant parameters themselves, will put VARs a step closer to DSGE modelling. Recent work by Davig and Leeper (2005) shows an example of how to use a DSGE model to restrict a VAR on monetary and fiscal policy.

Conclusion

There is a tension between models that have clear economic interpretations but offer a poor fit to data and models that fit well but have few a priori assumptions and are therefore less interpretable (Ingram and Whiteman 1994; Del Negro and Schorfheide 2004). The original philosophy motivating VARs assumes that the economy is sufficiently complex and that simplified theoretical models, while useful in organizing thought about how the economy works, generally abstract from important aspects of the economy. VAR modelling begins with the minimal restrictions on dynamic time-series models, explores empirical regularities that have been ignored by simple models, and insists on the model's fit to data. The emphasis on fit has begun to bear fruit, as an increasing array of dynamic stochastic general equilibrium models have been tested and compared with VARs (Christiano et al. 2005; Smets and Wouters 2003). Markov-switching VARs go a step further in bringing VARs even closer to the data and thus provide a new benchmark for model comparison.

At the same time, considerable progress has been made to narrow the gap between VARs and DSGE models. Some results from VARs have provided empirical support to the key assumption made by real business cycle (RBC) models that monetary policy shocks play insignificant roles in generating business fluctuations. Nason and Cogley (1994) and Cogley and Nason (1995) discuss similar results from both VAR and RBC approaches. Fernandez-Villaverde et al. (2005) provide conditions and examples under which there exists the VAR representation of a DSGE model. Sims and Zha (2006a) display a close connection between an identified VAR and a DSGE model, and provide a measure for determining whether the 'invertibility problem' is a serious issue.

Undoubtedly there are payoffs in moving beyond the original VAR philosophy by imposing more restrictions on both contemporaneous relationships and lag structure while the restrictions are guided carefully by economic theory. Although moving in this direction is desirable, it is essential to maintain the spirit of VAR analysis as originally proposed by Sims (1980). This requires that heavily restricted VARs be subject to careful evaluation in terms of fit. Recent advances in Bayesian estimation and inference methods of restricted VARs make it feasible to compute the MDD accurately and efficiently and, therefore, to determine whether the restrictions have compromised the fit. These methods, however, still fall short of handling VARs with cross-equation restrictions implied by DSGE models. Thus, the challenge ahead of us is to develop new tools for VARs with possible cross-equation restrictions.

See Also

- ▶ Bayesian Econometrics
- ▶ Bayesian Methods in Macroeconometrics
- ▶ Markov Chain Monte Carlo Methods
- ▶ Structural Vector Autoregressions

Bibliography

Bernanke, B. 1986. Alternative exploration of the money-income correlation. *Carnegie-Rochester Conference Series on Public Policy* 25: 49–99.

Blanchard, O., and D. Quah. 1993. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 83: 655–673.

Blanchard, O., and M. Watson. 1986. Are business cycles all alike? In *The American business cycle: Continuity and change*, ed. R. Gordon. Chicago: University of Chicago Press.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90: 1313–1321.

Christiano, L., M. Eichenbaum, and C. Evans. 1999. Monetary policy shocks: What have we learned and to what end? In *Handbook of Macroeconomics*, ed. J. Taylor and M. Woodford, Vol. 1A. Amsterdam: North-Holland.

Christiano, L., M. Eichenbaum, and C. Evans. 2005. Nominal rigidities and the dynamics effects of a shock to monetary policy. *Journal of Political Economy* 113: 1–45.

Cogley, T., and J. Nason. 1995. Output dynamics in real business cycle models. *American Economic Review* 85: 492–511.

Cogley, T., and T. Sargent. 2005. Drifts and volatilities: Monetary policies and outcomes in the post WWII U.S. *Review of Economic Dynamics* 8: 262–302.

Cushman, D., and T. Zha. 1997. Identifying monetary policy in a small open economy under flexible exchange rates. *Journal of Monetary Economics* 39: 433–448.

Davig, T., and E. Leeper. 2005. Fluctuating macro policies and the fiscal theory. Working Paper No. 11212. Cambridge, MA: NBER.

Del Negro, M., and F. Schorfheide. 2004. Priors from general equilibrium models for VARs. *International Economic Review* 45: 643–673.

Evans, C., and D. Marshall. 2002. Economic determinants of the nominal treasury yield curve. Working paper. Federal Reserve Bank of Chicago.

Faust, J., and E. Leeper. 1997. When do long-run identifying restrictions give reliable results? *Journal of Business and Economic Statistics* 15: 345–353.

Fernandez-Villaverde, J., J. Rubio-Ramirez, and T. Sargent. 2005. A, B, C's (and D's) for understanding VARs. Working Paper No. 2005–9. Federal Reserve Bank of Atlanta.

Gali, J. 1992. How well does the IS-LM model fit postwar U.S. data? *Quarterly Journal of Economics* 107: 709–738.

Geweke, J. 1999. Using simulation methods for Bayesian econometric models: Inference, development, and communication. *Econometric Reviews* 18: 1–73.

Geweke, J., and C. Whiteman. 2006. Bayesian forecasting. In *The handbook of economic forecasting*, ed. G. Elliott, C. Granger, and A. Timmermann. Amsterdam: North-Holland.

Gordon, D., and E. Leeper. 1994. The dynamic impacts of monetary policy: An exercise in tentative identification. *Journal of Political Economy* 102: 1228–1247.

Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384.

Ingram, B., and C. Whiteman. 1994. Supplanting the Minnesota prior: Forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics* 34: 497–510.

Leeper, E., and T. Zha. 2003. Modest policy interventions. *Journal of Monetary Economics* 50: 1673–1700.

Leeper, E., C. Sims, and T. Zha. 1996. What does monetary policy do? *Brookings Papers on Economic Activity* 2: 1–78.

Litterman, R. 1986. Forecasting with Bayesian vector autoregressions – Five years of experience. *Journal of Business and Economic Statistics* 4: 25–38.

Nason, J., and T. Cogley. 1994. Testing the implications of long-run neutrality for monetary business cycle models. *Journal of Applied Econometrics* 9: S37–S70.

Nason, J., and J. Rogers. 2006. The present-value model of the current account has been rejected: Round up the usual suspects. *Journal of International Economics* 68: 159–187.

Robertson, J., and E. Tallman. 1999. Vector autoregressions: Forecasting and reality. *Federal Reserve Bank of Atlanta Economic Review* 84(1): 4–18.

Robertson, J., and E. Tallman. 2001. Improving federal-funds rate forecasts in VAR models used for policy analysis. *Journal of Business and Economic Statistics* 19: 324–330.

Rubio-Ramirez, J., D. Waggoner, and T. Zha. 2005. Markov-switching structural vector autoregressions: Theory and applications. Working Paper No. 2005–27. Federal Reserve Bank of Atlanta.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–47.

Sims, C. 1982. Policy analysis with econometric models. *Brookings Papers on Economic Activity* 1: 107–152.

Sims, C. 1986. Are forecasting models usable for policy analysis. *Federal Reserve Bank of Minneapolis Quarterly Review* 10(1): 2–16.

Sims, C., and T. Zha. 1998. Bayesian methods for dynamic multivariate models. *International Economic Review* 39: 949–968.

Sims, C., and T. Zha. 1999. Error bands for impulse responses. *Econometrica* 67: 1113–1155.

Sims, C., and T. Zha. 2006a. Does monetary policy generate recessions? *Macroeconomic Dynamics* 10(2): 231–272.

Sims, C., and T. Zha. 2006b. Were there regime switches in US monetary policy? *American Economic Review* 96: 54–81.

Smets, F., and R. Wouters. 2003. An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association* 1: 1123–1175.

Stock, J., and M. Watson. 2003. Has the business cycle changed? Evidence and explanations. Prepared for the federal reserve bank of Kansas city symposium 'Monetary policy and uncertainty: Adapting to a changing economy', Jackson Hole, Wyoming, 28–30 August.

Waggoner, D., and T. Zha. 2003a. Likelihood-preserving normalization in multiple equation models. *Journal of Econometrics* 114: 329–347.

Waggoner, D., and T. Zha. 2003b. A Gibbs simulator for structural vector autoregressions. *Journal of Economic Dynamics & Control* 28: 349–366.

Zha, T. 1999. Block recursion and structural vector autoregressions. *Journal of Econometrics* 90: 291–316.

Velocity of Circulation

J. S. Cramer

The *velocity of circulation* of money is V in the *identity of exchange*

$$MV \equiv PT \tag{1}$$

which is due to Irving Fisher (1911). On the left-hand side, M is the stock of money capable of ready payment, i.e. currency and demand deposits, or, in modern parlance, M_1, on the right, P is the price level and T stands for the volume of trade. PT is usually identified with total transactions at current value, which must be identically equal to total payments. All these variables are aggregates. The identity defines V as PT/M, that is the ratio of a flow of payments to the stock of money that performs them; its dimension is time^{-1}.

Apart from defining V, the identity (1) also serves for rudimentary quantity theories of money. If V is assumed constant, we have a theory of money demand, with PT determining M. Again, with both V and T constant, changes in M imply changes in P; this is still a popular explanation of inflation, with 'too much money chasing too little goods'. The above quantity theory of money demand has however long been replaced by a more sophisticated argument, whereby money demand is determined along with demand for other assets by yield and liquidity differentials and by net wealth or income Y. This has led, by analogy, to the unfortunate term *income velocity* for the ratio Y/M. It should not be thought that Y here acts as a proxy for PT of the earlier theory: the underlying argument is quite different, and if Y is a proxy at all it represents net wealth. The term velocity is inappropriate in this context. We shall here reserve it for the *transactions velocity* V as defined above, and for its constituents parts.

This V has no place in modern economic analysis; it attracted some interest in the decades before 1940. When we divide M into currency M_c and demand deposits M_d, and acknowledge that there are several different types of transaction, (1) becomes

$$M_c V_c + M_d V_d = \sum_j P_j T_j. \qquad (2)$$

Among the variables in this expression, M_d and V_d are in principle observable at short notice, and in the absence of production indices and of national income estimates $M_d V_d$ (or $M_d V_d/P$) is a useful indicator of economic activity. It was used as such by authors like Angell (1936), Edie and Weaver (1930), Keynes (1930) and Snyder (1934). As for the data, M_d is demand deposit balances, available from banking returns, and V_d is the ratio of debits to balances, which can also be obtained from banks. The US Federal Reserve Board has long published monthly statistics of this *debits ratio* or *deposit turnover rate*, and still does so; there have been some drastic changes in definition and coverage over the years. The Bank of England provided a similar series from 1930 to 1938. Comparable statistics are available for several other countries.

The main trouble with this approach is that there is more than one type of transaction, and that (bank) payments are not limited to transactions in connection with current production. Some debits even have no economic meaning at all, as when a depositor has several accounts, and shifts funds between them, or when currency is withdrawn. Moreover bank debits can also reflect the sale of capital assets, income transfers, and money market dealings. The latter are by far the largest single category of turnover. These elements hinder the interpretation of V_d, and various attempts have been made to identify and remove them. We refer to Keynes' distinction between *industrial* and *financial* circulation, and to the Federal Reserve's practice of separately recording turnover in major financial centres. Failing a detailed classification of debits by the banks, however, all corrections are limited to approximate adjustments.

The observed value of V_d thus varies considerably with the definition of the relevant payments. For the US we quote the overall annual V_d, inclusive of financial transactions and the money market. This gross V_d, inclusive V_d rose from just under 30 in 1919 to about 35 in 1929, and then declined until 1945 when it was under 15. After the war it started on a long rise. It was about 50 by 1965, and from then onwards it soared to over 400 in 1984 (Garvy and Blyn 1970; Federal Reserve Bulletin). In Britain, *net* velocity, exclusive of the money market, was roughly stable at values between 15 and 20 from 1920 to 1940; later it rose from 20 in 1968 to 40 in 1977 (Cramer 1981). In the Netherlands, similarly defined net debits series show a V_d of between about 40 in 1965 and 45 in 1982 (Boeschoten and Fase 1984).

It is hard to find a single common interpretation of these movements. The development in the US until the 1960s suggests strong business cycle effects, but the enormous later increase of gross V_d must in large part be due to new techniques like overnight lending and repurchase agreements. These generate a huge amount of debits on the basis of quite small average balances. New banking techniques that go hand in hand with improved cash management explain increases in V_d outside the money market, too. The process is induced by the pressure of rising interest rates. Increased speed and precision of bank transfers permit a reduction of working balances at a given turnover level, and the reduction of demand moreover calls forth additional debits, as when idle funds are shifted to time deposits. Debits may thus increase *because* balances are reduced, and the rise of V_d is accentuated.

As regards currency payments, the currency stock M_c is well documented, but the estimation of velocity V_c or payments M_cV_c presents intractable problems. There are two solutions, but both use major assumptions that defy verification.

The first method is based on the redemption rates of worn-out banknotes of different denominations. Under stationary conditions these rates are the reciprocal of average lifetime, and this turns out to be positively related to face value. While this may well be due to more careful handling of the larger notes, it is usually inferred from this that larger denominations circulate less rapidly and are hoarded more often, and for longer periods, than small notes. Laurent (1970) uses these specific redemption rates to estimate currency payments. He assumes that a banknote is redeemed if and only if it has completed G transfers. Assigning G transfers to notes that are redeemed, and ½G to notes still in circulation, he builds up cumulative estimates of the transfers performed by each US denomination from 1861 onwards. This yields annual transfers by denomination, and hence total currency payments per year, ignoring coins. All estimates are of course a multiple of the unknown G, which is regarded as a physical constant like the number of times a note can be handled. Laurent assumes implicitly that it equals the number of payments a note can perform in its lifetime. He constructs currency payments series for various G, adds bank debits, and examines the correlation of this sum with GNP over the period 1875 to 1967. The maximum correlation occurs at $G = 129$, and this value is adopted. Since currency in circulation, bank debits, and GNP all share the same real growth and price movements, the constructed payment series will be closely correlated with GNP for *any* G, and the maximum correlation is not a good criterion for determining this constant. It is moreover uncertain that G *is*, constant. Laurent's estimates of currency payments imply that V_c is about 30 from 1875 to 1890; it then rises to a peak of 120 in 1928, and thereafter declines steeply to 32 in 1945, remaining at that level since. We shall argue that this level is too high.

The second method of estimating currency payments is due to Fisher (1909). He observes that most people obtain the currency they spend from banks, and that most recipients return their takings to banks. The currency circulation thus consists of *loops* of payments connecting withdrawals with deposits, and currency payments can be established by multiplying aggregate withdrawals (or deposits) by the average number of intervening payments, or the *loop length*. Withdrawals and deposits are of course recorded at the banks, and should be readily available statistics (although in fact they are not); as for the loop length, there is no way of measuring it, and it must be inferred from common sense considerations. In consumer spending the loop consists of a single payment, as households draw cash from the banks and spend it at retail shops that deposit all their takings. This is of course a minimum: some agents do not deposit their currency receipts, but spend them; some agencies, like post offices or stores that cash customers' cheques, act in a double capacity, paying out currency they have received and thus doubling the number of payments it performs before returning to the banks. Such considerations together suggest an average loop length of about two for present-day industrialized countries.

In recent years, V_c has been estimated for two countries for which series or estimates of cash withdrawals could be established. Fisher's method gives a constant V_c of about 18.5 for Britain over the period 1960–78 (Cramer 1981). For the Netherlands, a combination of Laurent's and Fisher's methods gives a constant value of about 15.3 for the years 1965–82 (Boeschoten and Fase 1984). These results suggest that currency velocity is a constant, as if it were set by physical limitations to the speed of currency circulation, and that it lies between 15 and 20.

This estimate often arouses strong feelings, as casual observation suggests that currency performs far more than 15 or 20 payments a year. A higher value of V_c does however mean higher currency payments M_cV_c, and it is not at all clear where these take place. Even with a velocity of 15 this is a problem, for at this value currency payments in most countries far exceed consumer spending, let alone retail sales. Yet consumer spending is commonly believed to be the major repository of cash. A fair proportion must by our estimate take place elsewhere, and it appears that crime or the informal economy cannot account for this vast amount. Over and again the currency stock is much larger than common sense would suggest. Where are these payments made? Where is all the currency used or hoarded? The plain answer is that no one knows, and that very few people care. Attempts to find the answer by a sample survey have failed (Cramer and Reekers 1976).

The above results suggest that even for current transactions (excluding the money market) bank velocity is larger than currency velocity, so that the steady and continuing shift from currency to demand deposits must mean a gradual increase in the overall velocity *V*.

See Also

▸ Demand for Money: Empirical Studies

Bibliography

Angell, J.W. 1936. *The behaviour of money*. New York: McGraw-Hill.

Boeschoten, W.J., and M.M.G. Fase. 1984. *The volume of payments and the informal economy in the Netherlands 1965–1982*, Monetary Monographs no. 1. Amsterdam/Dordrecht: de Nederlandsche Bank/Nijhoff.

Cramer, J.S. 1981. The volume of transactions and of payments in the United Kingdom, 1968–1977. *Oxford Economic Papers* 33(2): 234–255.

Cramer, J.S., and G.M. Reekers. 1976. Money demand by sector. *Journal of Monetary Economics* 2(1): 99–112.

Edie, L.D., and D. Weaver. 1930. Velocity of bank deposits in England. *Journal of Political Economy* 38: 373–403.

Fisher, I. 1909. A practical method for estimating the velocity of circulation of money. *Journal of the Royal Statistical Society* 72: 604–611.

Fisher, I. 1911. *The purchasing power of money*, 2nd ed, 1922. Reprinted New York: Kelley, 1963.

Garvy, G., and M.R. Blyn. 1970. *The velocity of money*. New York: Federal Reserve Bank, available from Microfilm International, Ann Arbor and London.

Keynes, J.M. 1930. *A treatise on money*. London: Macmillan.

Laurent, R.D. 1970. *Currency transfers by denomination*. PhD Dissertation, University of Chicago.

Snyder, C. 1934. On the statistical relation of trade, credit, and prices. *Revue de l'Institut international de Statistique* 2: 278–291.

Vent for Surplus

H. Myint

Conventionally, international trade theory focuses attention on the pattern of comparative costs existing at a point of time on the basis of the given resources and technology of the trading countries. Adam Smith, writing before the theory of comparative costs became formalized as a cross-section type of analysis, was concerned with the process of interaction between trade and development over a period of time. Thus his writings provide a more promising starting point for the study of the historical process of export expansion and economic development in the underdeveloped countries (Williams 1929; Myint 1958; and Myint 1977).

Actually, there were two strands in Adam Smith's analysis: the first, which may be called the 'productivity' theory, emphasized the role in international trade in widening the extent of the market and the scope for division of labour and specialization, thereby raising the productivity of labour by encouraging technical progress and enabling the trading country to enjoy increasing returns by overcoming technical indivisibilities imposed by the narrowness of the home market; the second, which may be called the 'vent for surplus' theory, emphasized the role of international trade in providing a wider market outlet or the 'vent' for the surplus productive capacity which would have remained underutilized in the absence of international trade.

When applied to the historical experience of the expansion of primary exports from the underdeveloped countries, Smith's 'productivity' theory of trade suggested too optimistic a picture of the rise in labour productivity through specialization and the possibility of reaping increasing returns through export expansion. It is true that the introduction of foreign investment and technology raised labour productivity in the mining and plantation exports. But this was usually of a one-off character and the subsequent expansion of output relied heavily on an abundant supply of unskilled labour at low wages. When the local labour supply was exhausted, the typical reaction was to recruit immigrant foreign labour from countries such as India and China with their vast reservoir of cheap labour, rather than to economize local labour and raise its productivity. This fell short of Adam Smith's optimistic vision of division of labour, with specialization continually raising labour's productivity. Smith's 'productivity' theory also did not accord with the typical process of expansion of peasant exports. Here, apart from the improvements in transport and communications and law and order, there was no significant improvement in agricultural techniques and the productivity of resources. Peasant exports simply expanded by bringing more land under cultivation and drawing upon the underemployed labour from the subsistence economy (Myint 1954).

This left unanswered the question of why the primary exports from the underdeveloped countries expanded so rapidly and in a sustained manner when these countries were opened up to multinational trade in the latter half of the 19th century or the early 20th century. Smith's vent for surplus theory serves to fill this gap. The typical process of expansion of primary exports may be looked upon as a long 'transition process' during which the expected tendency to diminishing returns was held in check by drawing upon the underutilized or the surplus natural resources and labour into export production; that is to say, exports expanded approximately under conditions of constant returns during the vent for surplus phase in many peasant export economies of South-East Asia and Africa seems to have continued rather longer than expected, lasting well into the recent postwar decades.

The significance of the vent for surplus theory for the study of the underdeveloped countries may be elaborated as follows. Under normal conditions (i.e. in the absence of short-run economic fluctuations), there is generally a gap in any country between the actual level of production attained and the theoretically attainable level of production with the 'given' resources and technology idealized in international trade theory as the production possibility frontier. This gap between the actual and the attainable level of output may be expected to be wider for the underdeveloped countries than for the developed countries, even if both were pursuing similar economic policies. An important reason for this may be traced to the fact that the domestic economic organization of the poorer countries is less well developed. Specifically, it is characterized by a poor internal system of transport and communications, by an incomplete development of the markets, particularly for the factors of production, and by an inadequate development of the administrative and fiscal machinery of the government. According to the vent for surplus theory, a substantial reserve of 'surplus' resources is likely to exist in a traditional economy not yet fully opened up to external economic relations, reflecting the underdeveloped nature of the domestic economic framework. In such a setting, international trade would provide a major force for economic development. It would bring about not only 'direct gains' from trade in the form of cheaper imports raising the economic welfare of the country, but also important 'indirect gains' transforming the organization of the domestic economy: through the extension and development of the exchange economy in the traditional agricultural sector, through the improvements in transport and communications and through a better provision of public services financed by increasing government revenue from the expanding exports (Myint 1958).

Further, the vent for surplus theory suggests that the 'direct gains' from trade would also be much larger than those envisaged in the conventional theory of multinational trade. In the conventional trade theory, the resources are assumed to be fully employed before a country enters into international trade and export production can be expanded only at the cost of contracting output for the domestic market. The gains from trade are therefore confined to the gains in allocative efficiency obtained by reallocating the given and fully employed resources according to the comparative advantage offered by international trade. In contrast, according to the vent for surplus theory, there is a considerable scope for expanding the exports of an underdeveloped country *without* contracting output for domestic consumption-by drawing upon the surplus land and labour. Thus the gains from trade would be larger because imports can be obtained with little or no resource cost. This hypothesis is supported by the experiences of the peasant export economies in South-East Asia and Africa. In Burma and Thailand, where rice happened to be both the main food crop and the export crop, rice exports expanded very rapidly for many decades without any contraction in the domestic food supply. If anything, it is possible to argue that the domestic food supplies of these countries were made more secure through the development of a large exportable surplus of rice brought about by the extension of cultivation to unused land. Similarly, African peasant economies such as Ghana, Nigeria or Uganda were able to expand their peasant exports in the prewar decades without any appreciable reduction in their domestic food production. Indeed, in the initial phase of export expansion export crops such as cocoa or cotton were usually interplanted with the food crops, such as yam, on the newly cleared pieces of land so that export production and domestic food production tended to increase together (Myint 1963, chs. 3 and 4).

The vent for surplus phase of peasant export expansion has continued somewhat longer than one would have expected at first sight. This is so because the existence of the 'unused' land is not given once for all in a physical sense by the geographical area but depends importantly on the improvements in transport and communications and the growth of the market system (the 'unused' labour being replenished by population growth). Thus, it is noteworthy that Thailand, which has been expanding her rice exports on a vent for surplus basis since the early 1900s, still managed to go through a rapid phase of expansion of new peasant exports, such as maize and tapioca in the 1960s and 1970s-mainly through an improvement in internal transport (Myint 1972, chs. 1 and 4). Similarly, in the 1950s and the 1960s, many African countries experienced a rapid expansion of new peasant exports, notably the tropical beverages, by bringing more land under cultivation. In particular, Ivory Coast continued with its rapid expansion of exports during the 1960s and 1970s. It is true that in recent times the expansion of peasant exports from many South-East Asian and African countries has slackened. In some countries, such as the Philippines, this is due to a genuine exhaustion of the supply of exploited land, which seems to have occurred by the end of the 1950s (Hayami and Ruttan 1985, ch. 10). In other countries, particularly those in Africa, the slackening in peasant export production may be attributed not to the end of the vent for surplus phase, but to the very unfavourable prices fixed for the peasant producers by the State Agricultural Marketing Boards (World Bank 1981, ch. 5) and, in some countries, to political instability. Sooner or later, of course, the vent for surplus phase of agricultural expansion will come to an end with the growing population pressure on limited land. But as suggested by the more recent phases of peasant export expansion in countries such as Thailand and the Ivory Coast, the possibility for the vent for surplus mechanism may not as yet be completely exhausted-given the policies of providing adequate incentives to the peasant farmers and political stability.

The vent for surplus theory may be extended on a somewhat different basis to the agricultural surpluses of the advanced countries such as the United States and the EEC countries. The reason for this type of surplus productive capacity is of course not the underdevelopment of the domestic economic organization, but the various farm support programmes induced by powerful political pressure (Hayami and Ruttan 1985, ch. 8). Despite this, however, it is instructive to study the international trade and aid policies of the advanced countries in terms of the vent for surplus theory and the desire to find an international outlet from the existing surplus productive capacity, rather than in terms of adapting their productive capacity to the world market demand.

See Also

▶ British Classical Economics

Bibliography

Caves, R.E. 1965. 'Vent for surplus' models of trade and growth. In *Trade, growth and the balance of payments*, ed. R.E. Baldwin et al. Chicago: Rand McNally.

Hayami, Y., and V.W. Ruttan. 1985. *Agricultural development: An international perspective*. Baltimore: Johns Hopkins Press.

Myint, H. 1954. An interpretation of economic backwardness. *Oxford Economic Papers* 6: 132–162.

Myint, H. 1958. The 'classical theory' of international trade and the underdeveloped countries. *Economic Journal* 68: 317–337.

Myint, H. 1963. *The economics of the developing countries*. London: Hutchinson.

Myint, H. 1972. *Southeast Asia's economy: Development policies in the 1970s*. Harmondsworth: Penguin Books.

Myint, H. 1977. Adam Smith's theory of international trade in the perspective of economic development. *Economica* 44: 231–248.

Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. London: Methuen. 1950.

Williams, J.H. 1929. The theory of international trade reconsidered. *Economic Journal* 39: 195–209.

World Bank. 1981. *Accelerated development in Sub-Saharan Africa: Agenda for action*. Washington, DC: World Bank.

Venture Capital

Josh Lerner

Abstract

Venture capital is independently managed, dedicated capital focusing on equity or equity-linked investments in privately held, high-growth companies. Research into venture capital has focused on the structure and financing of venture partnerships, the financial and operational interactions of venture capitalists with portfolio firms, and the exiting of venture capital investments. Major areas needed further research include the internationalization of venture capital, the impact of public policy, and the real economic effects of these funds.

Keywords

Agency conflicts; Asymmetric information; Capital gains taxation; Corporate governance; Covenants; Entrepreneurship; Financial intermediaries; High-risk assets; Initial public offerings; Investment; Limited liability; Limited partnerships; Liquidity constraints; Maximum likelihood; Monitoring; Patents; Pay-for-performance incentives; Reputation; Research and development; Venture capital

JEL Classifications

G3

Venture capital is independently managed, dedicated capital focusing on equity or equity-linked investments in privately held, high-growth companies. The first venture firm, American Research and Development, was formed in 1946 and invested in companies commercializing technology developed during the Second World War. Because institutions were reluctant to invest, it was structured as a publicly traded closed-end fund and marketed mostly to individuals, a structure emulated by its successors.

By 1978 limited partnerships had become the dominant investment structure. Limited partnerships have an important advantage: capital gains taxes are not paid by the limited partnership. Instead, only the taxable investors in the fund pay taxes. Venture partnerships have predetermined, finite lifetimes. To maintain limited liability, investors must not become involved in the management of the fund.

Activity in the venture industry increased dramatically in early 1980s. Much of the growth stemmed from the US Department of Labor's clarification of Employee Retirement Income Security Act's 'prudent man' rule in 1979, which had prohibited pension funds from investing substantial amounts of money into venture capital or high-risk asset classes. The rule clarification explicitly allowed pension managers to invest in high-risk assets, including venture capital.

The subsequent years saw both very good and trying times for venture capitalists. Venture capitalists backed many successful companies, including Apple Computer, Cisco, Genentech, Google, Netscape, Starbucks, and Yahoo! But commitments to the venture capital industry were very uneven, creating a great deal of instability. The annual flow of money into venture funds increased by a factor of ten during the early 1980s. From 1987 through 1991, however, fund-raising steadily declined as returns fell. Between 1996 and 2003, this pattern was repeated.

Venture capital investing can be viewed as a cycle. In this article, I follow the cycle of venture capital activity. I begin with the formation of venture funds. I then consider the process by which such capital is invested in portfolio firms, and the exiting of such investments. I end with a discussion of open research questions, including those relating to internationalization and the real effects of venture activity.

Fund-Raising

Research into the formation of venture funds has focused on two topics. First, the commitments to the venture capital industry have been highly variable since the mid-1970s. Understanding the determinants of this variability has been a topic of continuing interest to researchers. Second, the structure of venture partnerships has attracted increasing attention.

First, Poterba (1987, 1989) notes that the fluctuations could arise from changes in either the supply of or the demand for venture capital. It is very likely, he argues, that decreases in capital gains tax rates increase commitments to venture funds, even though the bulk of the funds are from tax-exempt investors. The drop in the tax rate may spur corporate employees to become entrepreneurs, thereby increasing the need for venture capital. The increase in demand due to greater entrepreneurial activity leads to more venture fund-raising.

Gompers and Lerner (1998b) find empirical support for Poterba's claim: lower capital gains taxes have particularly strong effects on venture capital supplied by tax- exempt investors. This suggests that the primary mechanism by which capital gains tax cuts affect venture fund-raising is the higher demand of entrepreneurs for capital. The authors also find that a number of other factors influence venture fund-raising, such as regulatory changes and the returns of venture funds.

A second line of research has examined the contracts that govern the relationship between investors (limited partners) and the venture capitalist (general partner). Gompers and Lerner (1999) find that compensation for older and larger venture capital organizations is more sensitive to performance than that of other venture groups. Also, the cross-sectional variation in compensation terms for younger, smaller venture organizations is considerably lower. The fixed component of compensation is higher for smaller, younger funds and funds focusing on high-technology or early stage investments. Finally, Gompers and Lerner do not find any relationship between the incentive compensation and performance.

The authors argue that these results are consistent with a learning model in which neither the venture capitalist nor the investor knows the venture capitalist's ability. With his early funds, the venture capitalist will work hard even without explicit pay-for-performance incentives: if he can establish a good reputation, he can raise subsequent funds. These reputation concerns lead to lower pay for performance for smaller and younger venture organizations. Once a reputation has been established, explicit incentive compensation is needed to induce the proper effort.

Covenants also play an important role in limiting conflicts in venture partnerships. Their use may be explained by two hypotheses. First, because negotiating and monitoring covenants are costly, they will be employed when monitoring is easier and the potential for opportunistic behaviour is greater. Second, in the short run the supply of venture capital services may be fixed, with a modest number of funds of carefully limited size raised each year. Increases in demand may lead to higher prices when contracts are written. Higher prices may include not only increases in monetary compensation, but also greater consumption of private benefits through fewer covenants.

Gompers and Lerner (1996) show that both supply and demand conditions and costly contracting are important in determining contractual provisions. Fewer restrictions are found in funds established during years with greater capital inflows and funds, when general partners enjoy higher compensation. The evidence illustrates the importance of general market conditions on the restrictiveness of venture partnerships. In periods when venture capitalists have relatively more bargaining power, the venture capitalists are able to raise money with fewer stings attached.

Lerner and Schoar (2004) examine rationales for constraints on liquidity. Venture groups often impose severe restrictions on transfers of partnership interests beyond what is required by securities law. They argue that these curbs allow general partners to screen for long-term investors. A limited partner who expects many liquidity shocks would find these restrictions especially onerous. Thus, the limited partners investing will be highly liquid, facilitating fund-raising in follow-on funds. The authors show that restrictions on liquidity are less common in later funds organized by the same venture group, when information problems are presumably less severe.

Investing

A second broad area of research has focused on the ties between venture capitalists and the firms in which they invest.

This literature emphasizes the informational asymmetries that characterize young firms, particularly in high-technology industries. These problems make it difficult for investors to assess firms, and permit opportunistic behaviour by entrepreneurs after finance is received. Specialized financial intermediaries, such as venture capitalists, address these problems by intensively scrutinizing firms before providing capital and monitoring them afterwards.

Economic theory examines the role that venture capitalists play in mitigating agency conflicts between entrepreneurs and investors. The improvement in efficiency might be due to the active monitoring and advice that is provided (Cornelli and Yosha 2003; Hellmann 1998; Marx 1994), the screening mechanisms employed (Chan 1983), the incentives to exit (Berglöf 1994), the proper syndication of the investment (Admati and Pfleiderer 1994), or investment staging (Bergemann and Hege 1998; Sahlman 1990).

Staged capital infusion is the most potent control mechanism a venture capitalist can employ. The shorter the duration of an individual round of financing, the more frequently the venture capitalist monitors the entrepreneur's progress. The duration of funding should decline and the frequency of re-evaluation increase when the venture capitalist believes that conflicts with the entrepreneur are likely.

If monitoring and information gathering are important – as models such as those of Amit et al. (1990) and Chan (1983) suggest – venture capitalists should invest in firms where asymmetric problems are likely, such as early stage and high-technology firms with intangible assets. The capital constraints faced by these companies will be large and these investors will address them.

Gompers (1995) shows that venture capitalists concentrate investments in early stage companies and high-technology industries where informational asymmetries are significant and monitoring is valuable. He finds that early stage firms receive significantly less money per round. Increases in asset tangibility are associated with longer financing duration and reduce monitoring intensity.

In a related paper, Kaplan and Strömberg (2003) document how venture capitalists allocate control and ownership rights contingent on financial and non- financial performance. If a portfolio company performs poorly, venture capitalists obtain full control. As performance improves, the entrepreneur obtains more control. If the firm does well, the venture capitalists relinquish most of their control rights but retain their equity stake.

Related evidence comes from Hsu (2004), who studies the price entrepreneurs pay to be associated with reputable venture capitalists. He analyses firms which received financing offers from multiple venture capitalists. Hsu shows that high investor experience is associated with a substantial discount in firm valuation.

Venture capitalists usually make investments with peers. The lead venture firm involves other venture firms. One critical rationale for syndication in the venture industry is that peers provide a second opinion on the investment opportunity and limit the danger of funding bad deals.

Lerner (1994a) finds that in the early investment rounds experienced venture capitalists tend to syndicate only with venture firms that have similar experience. He argues that, if a venture capitalist were looking for a second opinion, then he would want to get one from someone of similar or greater ability, certainly not from someone of lesser ability.

The advice and support provided by venture capitalists is often embodied in their role on the firm's board of directors. Lerner (1995) examines whether venture capitalists' representation on the boards of the private firms in their portfolios is greater when the need for oversight is larger, looking at changes in board membership around the replacement of CEOs. He finds that an average of 1.75 venture capitalists are added to the board between financing rounds when a firm's CEO is replaced in the interval; between other rounds 0.24 venture directors are added. No differences are found in the addition of other outside directors.

Hochberg (2005) studies the influence of venture capitalists on the governance of a firm following its initial public offering (IPO). Venture-backed firms manage earnings less in the IPO year, as measured by discretionary accounting accruals. Venture-backed firms also experience a stronger wealth effect when they adopt a poison pill, which implies that investors are less worried that the poison pill will entrench management at the expense of shareholders. Finally, venture-backed firms more frequently have independent boards and audit and compensation committees, as well as separate CEOs and chairmen.

It is natural to ask why other financial intermediaries (such as banks) cannot duplicate these features of the venture capitalists, and undertake the same sort of monitoring. Economists have suggested several explanations for the apparent superiority of venture funds in this regard. First, because regulations limit banks' ability to hold shares, they cannot freely use equity. Second, banks may not have the necessary skills to evaluate projects with few collateralizable assets and significant uncertainty. Finally, venture funds' high-powered compensation schemes give venture capitalists incentives to monitor firms closely. Banks sponsoring venture funds without high-powered incentives have found it difficult to retain personnel.

So far, this section has highlighted the ways in which venture capitalists can successfully address agency problems in portfolio firms. During periods when the amount of money flowing into the industry grows dramatically, however, competition between venture groups can introduce distortions.

Gompers and Lerner (2000) examine the relation between the valuation of venture deals and inflows into venture funds. Doubling inflows leads to a 7–21 per cent increase in valuation levels. But success rates don't differ significantly between investments made during periods of low inflows and valuations on the one hand and those made in booms on the other. The results indicate that the price increases reflect increasing competition for investment.

Exiting

A third major area of research has been the process whereby venture funds exit investments. This topic is important because, in order to make money on their investments, venture capitalists must sell their equity stakes.

Initial research into the exiting of venture investments focused on IPOs. This reflects the fact that typically the most profitable exit opportunity is an IPO. Barry et al. (1990) and Megginson and Weiss (1991) document that venture capitalists hold significant equity stakes and board positions in the firms they take public, which they continue to hold a year after the IP-O. They argue that this pattern reflects the certification they provide to investors that the firms they bring to market are not overvalued. Moreover, they show that venture-backed IPOs have less of a positive return on their first trading day, a finding that has been subsequently challenged (Lee and

Wahal 2004; Kraus 2002). The authors suggest that investors need a smaller discount because the venture capitalist has certified the offering's quality.

Subsequent research has examined the timing of the exit decision. Several potential factors affect when venture capitalists choose to bring firms public. Lerner (1994b) examines how the valuation of public securities affects when venture capitalists choose to finance companies in another private round in preference to taking the firm public. He shows that investors take firms public when market values are high, relying on private financings when valuations are lower. Seasoned venture capitalists appear more proficient at timing IPOs.

Another consideration may be the venture capitalist's reputation. Gompers (1996) argues that young venture firms have incentives to 'grandstand', or take actions that signal their ability to potential investors. Specifically, young venture firms bring companies public earlier than older one to establish a reputation and successfully raise new funds. Gompers shows that the effect of recent IPOs on the amount of capital raised is stronger for young venture firms, providing them with greater incentives to bring companies public earlier.

Lee and Wahal (2004) propose a variant of the 'grandstanding' hypothesis: they posit that venture firms have an incentive to underprice IPOs. The publicity surrounding a successful offering will enable the venture group to raise more capital than it could otherwise. Lee and Wahal confirm this hypothesis by showing a positive relationship between first-day returns and subsequent fund-raising by venture firms.

The typical venture firm, however, does not sell its equity at the time of the IPO. After some time, venture capitalists usually return money to their limited partners by distributing their shares. Gompers and Lerner (1998a) examine distributions. After significant increases in stock prices prior to distribution, abnormal returns around the distribution are negative. Cumulative excess returns for the 12 months following the distribution also appear to be negative. While the overall level of venture capital returns does not exhibit abnormal returns relative to the market (Brav and Gompers 1997), there is a distinct rise and fall around the time of the stock distribution. The results are consistent with venture capitalists possessing inside information and with the (partial) adjustment of the market to that information.

A related research area is venture-fund performance. Kaplan and Schoar (2005) show substantial persistence across consecutive venture funds. General partners that outperform the industry in one fund are likely to outperform in the next fund, while those who underperform in one fund are likely to underperform with the next fund. These results contrast with those of mutual funds, where persistence is difficult to identify.

Cochrane (2005) estimates the returns of venture capital investments. He notes that many analyses of returns focus only on investments that go public, get acquired, or go out of business. Such calculations may produce biased returns by concentrating only on the portfolio's 'winners' and outright failures. Cochrane develops a maximum likelihood estimate that uses existing data, but adjusts for these selection biases. While these papers – as well as Gompers and Lerner (1997) and Jones and Rhodes-Kropf (2003) – represent a first step towards understanding these issues, much more work remains to be done.

Future Research

While financial economists know much more about venture capital than they did a decade ago, there are many unresolved issues. I highlight here three promising areas.

The rapid growth in the US venture capital market has led institutional investors to look abroad. In a pioneering study, Jeng and Wells (2000) examine the factors that influence venture fund-raising internationally. They find that the strength of the IPO market is an important determinant of venture commitments, supporting Black and Gilson's (1998) hypothesis that the key to a successful venture industry is the existence of robust IPO markets. Jeng and Wells find, however, that the IPO market does not influence commitments to early-stage funds as much as those to later-stage ones. Much more remains to be explored regarding the internationalization of venture capital.

One provocative finding from Jeng and Wells's analysis is that government policy can dramatically affect the health of the venture sector. Researchers have only begun to examine the ways in which policymakers can catalyse the growth of venture capital and the companies in which they invest (Irwin and Klenow 1996; Lerner 1999; Wallsten 2000). Clearly, much more needs to be done in this arena.

A final area is the thorniest: the impact of venture capital on the economy. Demonstrating a causal relationship between innovation, job growth and venture activity is a challenging empirical problem. Kortum and Lerner (2000) examine the influence of venture capital on patented inventions in the United States over three decades, finding that increases in venture capital activity in an industry are associated with significantly higher patenting rates. One dollar of venture capital, they suggest, is three times more likely than one dollar of corporate R&D to stimulate patenting. (Hellmann and Puri 2000, also explore the impact of venture capital on innovation.) Many research opportunities remain in this arena.

See Also

- ▶ Entrepreneurship

Bibliography

Admati, A., and P. Pfleiderer. 1994. Robust financial contracting and the role for venture capitalists. *Journal of Finance* 49: 371–402.

Amit, R., L. Glosten, and E. Muller. 1990. Entrepreneurial ability, venture investments, and risk sharing. *Management Science* 36: 1232–1245.

Barry, C., C. Muscarella, J. Peavy III, and M. Vetsuypens. 1990. The role of venture capital in the creation of public companies: Evidence from the going public process. *Journal of Financial Economics* 27: 447–471.

Bergemann, D., and U. Hege. 1998. Venture capital financing, moral hazard, and learning. *Journal of Banking and Finance* 22: 703–735.

Berglöf, E. 1994. A control theory of venture capital finance. *Journal of Law, Economics, and Organizations* 10: 247–267.

Black, B., and R. Gilson. 1998. Venture capital and the structure of capital markets: Banks versus stock markets. *Journal of Financial Economics* 47: 243–277.

Brav, A., and P. Gompers. 1997. Myth or reality? Long-run underperformance of initial public offerings; Evidence from venture capital and nonventure capital-backed IPOs. *Journal of Finance* 52: 1791–1821.

Chan, Y. 1983. On the positive role of financial intermediation in allocation of venture capital in a market with imperfect information. *Journal of Finance* 38: 1543–1568.

Cochrane, J. 2005. The risk and return of venture capital. *Journal of Financial Economics* 75: 3–52.

Cornelli, F., and O. Yosha. 2003. Stage financing and the role of convertible debt. *Review of Economic Studies* 70: 1–32.

Gompers, P. 1995. Optimal investment, monitoring, and the staging of venture capital. *Journal of Finance* 50: 1461–1489.

Gompers, P. 1996. Grandstanding in the venture capital industry. *Journal of Financial Economics* 42: 133–156.

Gompers, P., and J. Lerner. 1996. The use of covenants: An empirical analysis of venture partnership agreements. *Journal of Law and Economics* 39: 463–498.

Gompers, P., and J. Lerner. 1997. Risk and reward in private equity investments: The challenge of performance assessment. *Journal of Private Equity* 1: 5–12.

Gompers, P., and J. Lerner. 1998a. Venture capital distributions: Short- and long-run reactions. *Journal of Finance* 53: 2161–2183.

Gompers, P., and J. Lerner. 1998b. What drives venture fundraising? In *Brookings papers on economic activity – Microeconomics*, 149–192. Washington, DC: Brookings Institution Publishing.

Gompers, P., and J. Lerner. 1999. An analysis of compensation in the U.S. venture capital partnership. *Journal of Financial Economics* 51: 3–44.

Gompers, P., and J. Lerner. 2000. Money chasing deals? The impact of fund inflows on private equity valuations. *Journal of Financial Economics* 55: 281–325.

Hellmann, T. 1998. The allocation of control rights in venture capital contracts. *RAND Journal of Economics* 29: 57–76.

Hellmann, T., and M. Puri. 2000. The interaction between product market and financing strategy: The role of venture capital. *Review of Financial Studies* 13: 959–984.

Hochberg, Y. 2005. Venture capital and corporate governance in the newly public firm. Working paper, Northwestern University.

Hsu, D. 2004. What do entrepreneurs pay for venture capital affiliation? *Journal of Finance* 59: 1805–1844.

Irwin, D., and P. Klenow. 1996. High tech R&D subsidies: Estimating the effects of Sematech. *Journal of International Economics* 40: 323–344.

Jeng, L., and P. Wells. 2000. The determinants of venture funding: Evidence across countries. *Journal of Corporate Finance* 6: 241–289.

Jones, C., and M. Rhodes-Kropf. 2003. The price of diversifiable risk in VC and private equity. Working paper, Columbia University.

Kaplan, S., and A. Schoar. 2005. Private equity performance: Returns, persistence, and capital. *Journal of Finance* 60: 1791–1823.

Kaplan, S., and P. Strömberg. 2003. Financial contract theory meets the real world: An empirical analysis of venture capital contracts. *Review of Economic Studies* 70: 281–315.

Kortum, S., and J. Lerner. 2000. Assessing the contribution of venture capital to innovation. *RAND Journal of Economics* 31: 674–692.

Kraus, T. 2002. Underpricing of IPOs and the certification role of venture capitalists: Evidence from Germany's Neuer Markt. Working paper, University of Munich.

Lee, P., and S. Wahal. 2004. Grandstanding, certification and the underpricing of venture capital backed IPOs. *Journal of Financial Economics* 73: 375–407.

Lerner, J. 1994a. The syndication of venture capital investments. *Financial Management* 23: 16–27.

Lerner, J. 1994b. Venture capitalists and the decision to go public. *Journal of Financial Economics* 35: 293–316.

Lerner, J. 1995. Venture capitalists and the oversight of private firms. *Journal of Finance* 50: 301–318.

Lerner, J. 1997. An empirical exploration of a technology race. *RAND Journal of Economics* 28: 228–247.

Lerner, J. 1999. The government as venture capitalist: The long-run effects of the SBIR program. *Journal of Business* 72: 285–318.

Lerner, J., and A. Schoar. 2004. The illiquidity puzzle: Theory and evidence from private equity. *Journal of Financial Economics* 72: 3–40.

Marx, L. 1994. Negotiation and renegotiation of venture capital contracts. Working paper, University of Rochester.

Megginson, W., and K. Weiss. 1991. Venture capital certification in initial public offerings. *Journal of Finance* 46: 879–893.

Poterba, J. 1987. How burdensome are capital gains taxes? Evidence from the United States. *Journal of Public Economics* 33: 157–172.

Poterba, J. 1989. Venture capital and capital gains taxation. In *Tax policy and the economy*, ed. L. Summers. Cambridge, MA: MIT Press.

Sahlman, W. 1990. The structure and governance of venture capital organizations. *Journal of Financial Economics* 27: 473–524.

Wallsten, S. 2000. The effects of government-industry R&D programs on private R&D: The case of the Small Business Innovation Research program. *RAND Journal of Economics* 31: 82–100.

Verdoorn's Law

J. S. L. McCombie

One of the most notable features of the postwar economic performance of the advanced countries has been the substantial and persistent differences between the various economies in their rates of growth of productivity and output. Yet these disparities are merely one aspect of the more general picture of economic development. Since the beginning of the Industrial Revolution, at which time there appears to have been little variation between areas in terms of per capita income, some countries have achieved a sustained growth in productivity whilst others have shown little or no improvement. The reasons for this, of course, remain a source of controversy.

Verdoorn's Law is an empirical generalization that provides the basis for one such explanation. Although originally discussed in terms of the differences in productivity growth of the advanced countries, the law is now recognized as having a wider significance for the more general process of economic growth and development.

In its simplest form, the law states that there is a close relationship between the long run growth of manufacturing productivity and that of output. (The law has also been found to hold for public utilities and the construction industries but not for any other sector of the economy.) The importance of the law is that it suggests that a substantial part of productivity growth is endogenous to the growth process, being determined by the rate of expansion of output through the effect of economies of scale.

The development of this approach to the theory of economic growth owes much to the writings of Lord Kaldor (see, in particular, Kaldor 1978a, b, and the symposium on Kaldor's growth laws published in the 1983 edition of the *Journal of Post Keynesian Economics*). Indeed, interest in the law primarily dates from Kaldor's (1966) inaugural lecture which examined why the United

Kingdom had grown so much more slowly over the postwar period than most other industrial countries. (It was P.J. Verdoorn, however, who had first discussed the relationship between productivity and output growth in an article published in 1949. The paper was written in Italian which may explain why it had largely escaped notice, with the notable exception of Colin Clark (1957), until Kaldor drew attention to it. Kaldor was also the first to discuss the broader implications of the law for economic growth.)

In the inaugural lecture, Kaldor observed that there was a close relationship for the advanced countries between the growth of manufacturing output per worker (p) and that of output (q). When the Verdoorn Law was estimated in the form $p = a + bq$ using cross-country data for twelve advanced countries over the early postwar period, it was found that the estimate of b, the 'Verdoorn coefficient', took a value of about one half. (Other studies have discovered similar results using cross-industry, time-series and regional data for both the advanced and the less developed countries.) Since the exponential growth of productivity is definitionally equal to the difference between output and employment growth (e), the law is sometimes expressed as $e = -a + (1 - b)\ q$. But the implications are the same. An increase in the growth of output will cause an increase in the growth of employment of about half a percentage point and an increase in productivity growth of a similar magnitude. Kaldor argued that this implies that manufacturing is subject to substantial increasing returns to scale.

The emphasis on the role of economies of scale as an important factor in determining the rate of economic progress has a long history. It is the basis of Adam Smith's (1776) principle enunciated in the opening sentence of Book I of *The Wealth of Nations* that '[the] greatest improvement in the productive powers of labour, and the greater part of the skill, dexterity, and judgement with which it is anywhere directed, or applied, seem to have been the effect of the division of labour'. The latter in turn is limited by the extent of the market. This is nothing more than the phenomenon of economies of scale, in the broad sense of the term. The theme was subsequently elaborated in Allyn Young's (1928) classic paper. In particular, Young argued that an important implication is that the capital–labour ratio is not to be understood as a response to relative factor prices but is primarily determined by the scale of production. He further stressed that economies of scale are primarily a macroeconomic phenomenon, the result of increased inter-industry specialization. (But it should be emphasized that the law has been found to apply to individual manufacturing industries.)

Another major tenet of the argument is that the law reflects both static and dynamic economies of scale. The former is a function of the volume of output and the gains in productivity from this source are reversible – if output contracts so the benefits of scale will be lost. Dynamic returns to scale, on the other hand, reflect such factors as 'learning by doing' and are usually ascribed to the rate of growth of output. These gains in productivity represent the acquisition of knowledge concerning more efficient methods of production and as such are irreversible. Substantial gains in productivity have been found to arise from this source even in the absence of any gross investment. A more rapid expansion of production will also lead to (as well as be the result of) a greater rate of innovation and a climate more favourable to risk taking. Investment will also be more efficiently used if it is introduced as part of a planned modernization scheme under conditions of rapidly expanding output rather than added, in an ad hoc manner, to existing capacity in stagnating industries. (Lamfalussy 1963, has termed these 'enterprise' and 'defensive' investment, respectively.)

For the law to provide evidence of the degree of returns to scale, it must be interpreted as reflecting a production relationship such as a form of the technical progress function. This being the case, the law is now usually specified as including the growth of the capital stock. This allows a separation to be made between the growth of productivity due to the greater use of machinery and that resulting from increasing returns to scale, per se. The inclusion of the growth of capital has not led to any major revision of the interpretation of the law.

The technical progress function was developed by Kaldor in an attempt to avoid the misleading dichotomy of growth into shifts of the production function and movements along the function. It is therefore all the more ironic that Verdoorn (1949, 1980) himself regards the law as being derived from the neoclassical Cobb–Douglas production function, although with the latter expressed in terms of growth rates. (The *linear* technical progress function may also be integrated to yield a conventional production function, although this is not necessarily true of the non-linear specifications.) Nevertheless, a paradox arises in that the estimation of the law using the *levels* of the various variables (the 'static Verdoorn Law') suggests either constant or small increasing returns to scale, whereas large estimates are obtained by estimating the 'dynamic law' using the same data sets. One explanation is that while the Verdoorn Law may be derived by differentiating a Cobb–Douglas production function with respect to time, it does not follow that the latter is the correct underlying structure. Integrating the law will lead to innumerable structures, depending upon the constant of integration.

The implications of Verdoorn's Law are far-reaching. It suggests that there is an inherent tendency for growth to proceed in a self-reinforcing manner and provides an economic rationale for Myrdal's (1957) notion of 'cumulative causation'. An increase in output causes a faster growth of productivity for the reasons already noted. Provided all the gains are not absorbed by increased real wages, countries (or firms) will experience an increasing cost advantage over their competitors. Improvements in the non-price aspects of competition, such as quality, are also positively related to productivity growth. Of course, growth is not observed to be explosive and formalizations of the cumulative causation model show how the growth of various countries may converge to (differing) equilibrium rates.

(However, it has been suggested that the Verdoorn Law may simply result from this reverse causation from productivity to output growth. Large differences in *exogenous* productivity growth could lead to variations in output growth through the price mechanism – the 'Salter effect'. This could generate a Verdoorn-type relationship even though constant returns to scale prevail. However, the evidence suggests that this is unlikely to be significant for total manufacturing or for an individual industry, although it may be an important factor in crossindustry studies.)

Since the Verdoorn Law shows that differences in productivity growth are caused by variations in the growth of output, the problem is to explain why disparities in the latter arise. In the inaugural lecture, Kaldor argued that the United Kingdom's economic problems stemmed from the limited supply of labour available to the manufacturing sector and it was this that prevented a faster rate of growth. If this is the case, the Verdoorn Law may be mis-specified since employment and not output growth should be the regressor (Rowthorn 1975). When this specification (sometimes confusingly known as Kaldor's Law) is estimated, most studies find that constant returns to scale prevail. However, Kaldor later retracted his earlier position. The long run growth of the advanced countries (and, equally, the less developed countries) is not determined by the exogenously given growth of factor inputs but rather by the growth of 'effective demand'. Under these circumstances, the original specification of the law is to be preferred, although the very nature of the cumulative causation mechanism suggests that both output and employment growth may be jointly determined.

The importance of the rate of growth of demand as the driving force behind the pace of economic growth extends beyond the issues concerning the correct specification of the law. Long-run growth is best understood in a Keynesian (or, more appropriately, 'Kaldorian') framework. The rate of capital accumulation cannot be seen as an independent determinant of development since it is as much a result as a cause of the growth of output. The evidence further suggests that labour supplies were not a serious factor in limiting the growth of the advanced countries even during their most rapid expansionary phase which lasted from the end of World War II until 1973 (Cornwall 1977). There was either disguised unemployment in the primary and tertiary sectors or sufficient immigration to satisfy the demand for labour emanating from the manufacturing sector. The question naturally arises as to what is it that determines the growth of exogenous demand. In the early stages of development it is the growth of the agricultural surplus and the rate of land-saving innovations. With industrialization and the decline of the importance of agriculture, the key determinant becomes the growth of exports. This provides a source of the growth of autonomous demand both directly through the Harrod foreign trade multiplier and indirectly by relaxing the balance of payments constraint. Growth can thus be regarded as being ‘export-led’.

An important result of this approach is that, given the cumulative nature of economic growth, there is no inherent tendency for free trade to be to the benefit of all countries. Trade liberalization may well lead to a further deterioration in the growth of those countries which are already lagging as they find they become increasingly less competitive internationally. This is, of course, the converse of the inference that is sometimes drawn from the neoclassical theory of trade.

See Also

- ► Cumulative Causation
- ► Increasing Returns to Scale

References

Clark, C. 1957. *The conditions of economic progress*, 3rd ed. London: Macmillan.

Cornwall, J. 1977. *Modern capitalism: Its growth and transformation*. London: Martin Robertson.

Kaldor, N. 1966. *Causes of the slow rate of economic growth of the United Kingdom: An inaugural lecture*. Cambridge: Cambridge University Press.

Kaldor, N. 1978a. *Further essays on economic theory*. London: Duckworth.

Kaldor, N. 1978b. *Further essays on applied economics*. London: Duckworth.

Lamfalussy, A. 1963. *The United Kingdom and the Six. An essay on economic growth in Western Europe*. London: Macmillan.

Myrdal, G. 1957. *Economic theory and underdeveloped regions*. London: Duckworth.

Rowthorn, R.E. 1975. What remains of Kaldor’s Law? *Economic Journal* 85: 10–19.

Salter, W.E.G. 1960. *Productivity and technical change*. Cambridge: Cambridge University Press.

Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nation*, ed. E. Cannan. London: Methuen, 1961.

Thirlwall, A.P. (ed.). 1983. Symposium: Kaldor’s growth laws. *Journal of Post Keynesian Economics:* 5(3).

Verdoorn, P.J. 1949. Fattori che regolano lo sviluppo della produttività del lavoro. *L’Industria* 1: 45–53.

Verdoorn, P.J. 1980. Verdoorn’s Law in retrospect: A comment. *Economic Journal* 90: 382–385.

Young, A.A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.

Vernon, Raymond (Born 1913)

S. Hirsch

Vernon has been a most prolific writer on international economic relations in the post World War II era. His writings reflect a multi-faceted career which includes nearly two decades in government service, a short stint with private business, three years as director of the New York Metropolitan Region Study and, since 1959, a fruitful association with Harvard University, first at the Business School, where he was the leading figure in the teaching and research of international business, and later at the John F. Kennedy School of Government, where he was incumbent of the Clarence Dillon Chair of International Affairs until his retirement.

The policy orientation of his writing and the acute awareness it reflects of the interests and point of view of foreign governments, their institutional make up and constraints, surely owe much to his years of service with the State Department. His abiding interest in the restructuring of international trade, investment and payments systems, economic development, especially of Latin America, and economic relations between East and West must be similarly attributed to his State Department experience.

One of Vernon’s early analytical contributions concerns the economics of location. In the New York Metropolitan Region Study he adapted the notion of ‘external economies’ to the specific environment of urban agglomeration. The term was used by him to characterize the cost advantage enjoyed by firms located in urban centres because of their closeness to sources of information and to a large variety of specialized services. The availability of these services and their low costs determine the characteristics of industries, such as electronics, fashion goods, printing and publishing, which tend to flourish in agglomerates despite the high costs of more conventional production factors such as labour, space and transportation.

Information and specialized services also figure prominently in Vernon's extensive writings on the multinational corporation. In this case, Vernon has shown how information and specialized services are internalized and transformed into proprietary knowledge, which is used by the firm to obtain a monopolistic position in the domestic and international markets. This position is extended from the early to the mature phase of the 'product cycle' by transferring production to subsidiaries located in countries where conventional production factors are least costly, while retaining the location of the head office in the most developed markets where the new product and process specifications originate.

Alone and in collaboration with colleagues and doctoral students at the Harvard Business School, Vernon published numerous books and articles about the multinationals. He studied their dominant role in world production and trade of technology-based industries on the one hand the resource-based ones on the other, using the 'product cycle' as well as the more traditional industrial organization models to explain their distinct competitive structure, their insoluble conflicts with both their host and home governments, conflicts which evolve through a predictable cycle of power relations which Vernon aptly termed the 'obsolescing contract'.

His books *Sovereignty at Bay* (1971) and *Storm over the Multinationals* (1977), which summarize his work on the multinational corporation, will be regarded as major contributions to our knowledge of the multinational corporations for many years to come.

Business–government relations had been dealt with by Vernon early in his career as a civil servant. He returned to the theme in his work on the multinationals. The subject figures even more prominently in his more recent work conducted at the Kennedy School of Government, which focuses on state-owned enterprises and on government relations with private sector firms against the background of the energy crisis of the mid-seventies and its aftermath.

In *Two Hungry Giants,* which compares US and Japanese responses to the threat of resource shortage, Vernon attributes Japan's superior performance to the skilful way in which the Japanese government managed to harness private sector corporations to the 'national interest'.

Marxist doctrine claims that the state is being used by capitalists to advance their class interests. Vernon's analysis offers a less dogmatic view of the role of the state: to enhance their goals, even governments of 'market economies' increasingly use both state and privately owned enterprises as instruments of national policy.

Selected Works

1971. *Sovereignty at bay: The multinational spread of U.S. Enterprises*. New York: Basic Books.

1977. *Storm over the multinationals: The real issues*. Cambridge, MA: Harvard University Press.

1983. *Two hungry giants: The United States and Japan in the quest for oil and ores*. Cambridge, MA: Harvard University Press.

Verri, Pietro (1728–1797)

Peter Groenewegen

Keywords

Balance of production and consumption; Consumption taxation; Marginalism; Mathematical economics; Tax incidence; Verri, P.

JEL Classifications
B31

Italian economist, administrator and philosopher, Verri was born in Milan in 1728, educated in Rome and Parma, served with Austria in the Seven Years War and at this time was introduced to the study of economics by General Henry Lloyd (Venturi 1978, 1979). His economic writings of the 1760s, such as *Elementi di Commercio* (1760) and the dialogues on monetary disorders in the State of Milan (1762), led to his appointment to a number of positions in the Austrian civil service in Milan. His administrative achievements include the abolition of tax farming (1770) and lowering and simplifying the tariff (1786). From 1764 to 1766 he edited with his brother Alessandro the periodical *Il Caffè*, which attracted contributions on economics from Beccaria and Frisi as well as himself (Verri 1764). His most important economic publication, *Reflections on Political Economy,* appeared in 1771, went through numerous editions and was translated into French, German and Dutch and more recently into English. Other economic works on monetary and trade questions, including his 1769 pamphlet advocating freedom of the domestic corn trade, contribute to his reputation as a most important 18th-century Italian economist (McCulloch 1845, pp. 26–7). More recently he has been noted for inspiring early developments in mathematical economics (Theocharis 1961, pp. 27–34). He died in 1797.

Verri's *Reflections* is a complete treatise on political economy, reminiscent of Turgot's *Reflections on the Production and Distribution of Wealth* (1766) with its tight, logical framework and division into fairly short sections. Although these cover a wide range of subjects, they are interconnected by the basic theme of the work, the increase in annual reproduction of the nation through trade of surplus product which Verri related to the balance of production and consumption. This ratio or balance is the key concept in Verri's economic analysis, since it not only influences economic growth but also value (it approximates the ratio of sellers to buyers at home and abroad), the rate of interest (it represents thriftiness conditions) and, via its influence on the balance of trade, it also determines national money supply. An excess of production over consumption lowers the price level and the rate of interest, expands the money supply, animates industry and facilitates the collection of taxes. Some features of this analysis may be specifically noted. Verri does not appear to have been aware of the importance of capital, as is demonstrated in his general discussion of production (sections 26–8) and his treatment of the interest rate as a monetary phenomenon (sections 14–15). Secondly, his emphasis on supply and demand (used to determine all prices including the rate of interest) combined with references to utility and scarcity in the context of value (section 4) explains why this part of his work has been linked with marginalist economics. The last 11 sections discuss taxation and public finance, including a presentation of five canons of taxation (section 30), a tax incidence analysis arguing against the Physiocratic view that all taxes fall on the landlord (sections 32–3) and a plea for indirect consumption taxation as a fair and administratively easy way to raise revenue. Anti-Physiocratic elements in his economics are not confined to tax issues, but apply to his discussion of special classes (section 24), the importance of agriculture (section 28) and are apparent in his view that free trade should be largely confined to domestic activity (section 40). Verri's *Reflections* were highly regarded when they appeared, and could be found, for example, in Smith's library. His work, though now largely ignored, may therefore have exerted greater influence than is generally believed.

Selected Works

1760. Degli elementi di commercio. In Scrittori classici italiani di economia politica, *parte moderna*, vol. 17, ed. P. Custodi. Milan, 1804.

1762. Dialogo sui disordine delle monete nello stato di Milano nel 1762. In Scrittori classici italiani di economia polìtica, *parte moderna*, vol. 16, ed. P. Custodi. Milan, 1804.

1764. Considerazioni sul lusso. In *Scrittori classici italiani di economia politica,* parte moderna, vol. 17, ed. P. Custodi. Milan, 1804.

1771. *Reflections on political economy.* Trans. B. McGilvray and ed. P. Groenewegen, Reprints of economic classics, Series 2, No. 4. Sydney: University of Sydney, 1986.

Bibliography

McCulloch, J.R. 1845. *The literature of political economy.* London: LSE reprint, 1938.

Theocharis, R.D. 1961. *Early developments in mathematical economics*. London: Macmillan.

Venturi, F. 1978. Le 'Meditazioni sulla economica politica' di Pietro Verri: edizioni, echi e discussioni. *Rivista storica italiana* 90: 530–594.

Venturi, F. 1979. Le avventure del generale Henry Lloyd. *Rivista storica italiana* 91: 369–433.

Vertical Integration

Michael H. Riordan

Abstract

Modern economics takes a two-way approach to vertical integration. The theory of the firm approach focuses on how the unified control of successive production and distribution processes changes investment incentives, while the industrial organization approach studies how vertical integration affects the exercise of market power.

Keywords

Asset specificity; Backward integration; Bargaining costs; Bilateral vertical contracts; Chicago School; Commitment; Control rights; Double markups; Enforceable contracts; Exclusive dealing; Firm, theory of; Foreclosure; Forward integration; Free-rider problem; Hold-up problem; Imperfect information; Incomplete contracts; Industrial organization; Market exchange; Market power; Quasi-rent; Raising rivals' costs; Relationship-specific assets; Transaction costs; Variable proportions distortions; Vertical integration

JEL Classifications

L1

Vertical integration is the unified ownership and operation of successive production and distribution processes by a single firm. Backward integration occurs when a manufacturer controls the production of inputs, and forward integration occurs when the manufacturer controls distribution. The alternative (market exchange) is to procure inputs and distribution services from independent suppliers. Vertical integration is a matter of degree, as firms often are only partially integrated in one direction or the other.

Vertical integration raises issues for business strategy and public policy. A major theme in the theory of the firm literature is that vertical integration remedies underinvestment in relationship-specific assets due to opportunistic bargaining when contracts are incomplete. Accordingly, vertical integration enhances operational efficiency by improving investment incentives and reducing bargaining costs. Major themes of the industrial organization literature are that vertical integration reduces a firm's procurement or distribution costs, or raises those of its rivals. Accordingly, vertical integration is a strategy for competitive advantage. Policy issues concern whether the prevention or regulation of vertical integration improves consumer and social welfare.

Theory of the Firm Approach

The neoclassical theory of the firm assumes managers choose inputs and outputs to maximize profits subject to a production function, on the assumption that the governance of transactions is costless. The modern theory, in contrast, focuses explicitly on transaction costs, including efficiency losses, arising within and between firms.

The transaction-cost approach views vertical integration as a response to difficulties negotiating and executing market contracts (Coase 1937; Klein et al. 1978; Williamson 1975, 1985, 1996). The transaction-cost advantages of vertical integration over market exchange are most pronounced when contracts are incomplete, and uncertain future transactions require prior investments in relationship-specific assets for operational efficiency. In these circumstances, market exchange runs afoul of the hold-up problem. Relationship-specific assets by definition are strictly more valuable in a particular transactional relationship than in alternative uses; the difference in use value is called a quasi-rent. Thus asset specificity locks investors into bilateral relationships, while contractual incompleteness exposes them to costly bargaining over quasi-rents. Bargaining costs include failures to adapt transactions efficiently to unfolding circumstances and the direct costs of dispute resolution. Vertical integration improves operational efficiency by replacing dysfunctional bargaining with centralized authority over transactions, but adds bureaucratic costs, including efficiency losses from low-powered managerial incentives. A key hypothesis is that bargaining costs of market exchange rise with asset specificity faster than the bureaucratic costs of vertical integration, leading to two propositions: first, vertical integration is more likely the more important asset specificity is for efficiency; second, vertical integration supports more investment in relationship-specific assets than market exchange (Riordan and Williamson 1985). Empirical research generally bears out the implied positive correlation between vertical integration and the level of asset specificity (Shelanski and Klein 1995).

The more formal property-rights approach studies how *ex post* bargaining when contracts are incomplete distorts *ex ante* relationship-specific investments (Grossman and Hart 1986; Hart and Moore 1990; Hart 1995). Ownership confers control rights over the use of non-human assets used in production. While some specific control rights may be contracted away, the residual rights are held by the owners. Furthermore, managers make non-contractable relationship-specific investments to increase the value of these assets. The hold-up problem is manifest because *ex post* bargaining distributes the returns from these investments. Owner-managers who control the non-human assets of a firm undertake relationship-specific investments to the extent that the hold-up problem does not discourage them. Employee-managers, however, have less incentive because owners of the complementary non-human assets appropriate much of the investment returns. Thus vertical integration has mixed effects on managerial incentives. By eliminating the hold-up problem of market exchange, vertical integration improves investment incentives of owner- managers, while converting owner-managers into employees diminishes their incentives. Accordingly, the direction of vertical integration matters. Backward integration enhances investment incentives of downstream managers and degrades managerial incentives upstream, while forward integration has opposite effects. Optimal vertical integration depends on the importance of relationship-specific investments by managers at each stage of production and distribution. For example, forward integration is predicted when upstream managerial effort is particularly important for operational efficiency. Predictions of the property-rights approach depend sensitively on managers' investment opportunities to improve efficiency, and are difficult to test empirically (Whinston 2003).

Vertical integration also improves efficiency by reducing information imperfections at the root of bargaining costs (Arrow 1975; Riordan 1990). At the same time, the changed information structure diminishes investment incentives of employee-managers by aggravating the hold-up problem. This perspective reconciles with the property-rights approach by interpreting business information as an asset for which the owner has control rights. An open question is how the change in information structure derives from primitive conditions (Hart 1995).

Vertical integration might be motivated by the pursuit of greater bargaining leverage, rather than just greater efficiency (Bolton and Whinston 1993). A vertically integrated supplier with scarce capacity over-invests in its downstream unit in order to negotiate better terms from independent customers. The unfortunate effect is to discourage independents' investments in relationship-specific assets. Consequently, vertical integration tends to be excessive from a social welfare perspective.

Industrial Organization Approach

While the theory of the firm deals mainly with the reasons for vertical integration, industrial organization is more concerned with its effects on competition. Building on the neoclassical theory of the firm, industrial organization studies how market power distorts transactions. Much of this literature presupposes that transactions are governed by uniform prices for inputs and outputs. In this context, the Chicago School approach identifies efficiencies of vertical integration arising from a more profitable exercise of market power, including output expansion resulting from the elimination of 'double markups' when vertically related firms each exercise market power, the correction of ' variable proportions distortions' when independent downstream firms substitute towards more competitively supplied inputs, and the prevention of free-riding on point-of-sale services (Perry 1989).

The post-Chicago approach, by contrast, studies how foreclosure resulting from vertical integration reduces competition and raises rivals' costs (Ordover et al. 1990; Riordan 1998; Salinger 1988; Salop and Scheffman 1987). Foreclosure might drive up procurement costs or deny scale economies. Accordingly, an appropriate policy analysis weighs efficiencies against possible anti-competitive effects of vertical integration (Riordan and Salop 1995). The post-Chicago approach demonstrates conditions for anti-competitive foreclosure more rigorously than the traditional foreclosure doctrine attacked by the Chicago School (Bork 1978).

Another recent approach to vertical foreclosure studies the commitment problem of a supplier with market power who deals with customers bilaterally (Hart and Tirole 1990; Rey and Tirole 2007). Multilateral contracts involving a supplier and several downstream rival customers might be prevented by antitrust policy, or be unenforceable due to monitoring problems. Allowing more sophisticated contracting than just uniform pricing, the privacy of bilateral vertical contracts nevertheless fosters opportunism. A supplier has an adverse incentive to negotiate individual contracts that disadvantage other rival customers. Consequently, equilibrium supply contracts with favourable terms result in more downstream competition than would maximize total profits. Vertical integration restores monopoly power because the vertically integrated supplier is loath to set terms that hurt its own downstream division. The vertically integrated supplier offers less favourable terms to downstream rivals, raising their variable costs and causing higher downstream prices. Enforceable contracts with multilateral elements, such as exclusive dealing, also improve profits. Moreover, a vertically integrated firm has a greater incentive to enter into exclusive supply deals that foreclose upstream competitors and effectively cartelize a downstream industry (Chen and Riordan 2007). Such non-efficiency motives for vertical integration sometimes are contrary to consumer and social welfare, but are inconsequential if market power is absent.

See Also

- ► Firm Boundaries (Empirical Studies)
- ► Hold-up Problem
- ► Incomplete Contracts

Bibliography

Arrow, K.J. 1975. Vertical integration and communication. *Bell Journal of Economics* 6: 173–183.

Bolton, P., and M.D. Whinston. 1993. Incomplete contracts, vertical integration, and supply assurance. *Review of Economic Studies* 60: 121–148.

Bork, R.H. 1978. *The antitrust paradox: A policy at war with itself*. New York: Basic Books.

Chen, Y., and M.H. Riordan. 2007. Vertical integration, exclusive dealing, and *ex post* cartelization. *RAND Journal of Economics* 38: 1–21.

Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.

Grossman, S.J., and O.D. Hart. 1986. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* 94: 691–719.

Hart, O. 1995. *Firms, contracts and financial structure*. New York: Oxford University Press.

Hart, O., and J. Moore. 1990. Property rights and the nature of the firm. *Journal of Political Economy* 98: 1119–1158.

Hart, O. and Tirole, J. 1990. Vertical integration and market foreclosure. *Brookings papers on economic activity: Microeconomics*, 205–276.

Klein, B., R.A. Crawford, and A.A. Alchian. 1978. Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics* 21: 297–326.

Ordover, J.A., G. Saloner, and S.C. Salop. 1990. Equilibrium vertical foreclosure. *American Economic Review* 80: 127–142.

Perry, M.K. 1989. Vertical integration: Determinants and effects. In *Handbook of industrial organization*, vol. 1, ed. R. Schmalensee and R. Willig. Amstersdam: North-Holland.

Rey, P. and Tirole, J. 2007. A primer on foreclosure. In *Handbook of industrial organization*, vol. 3, ed. M. Armstrong and R.H. Porter. Amsterdam: North-Holland.

Riordan, M.H. 1990. What is vertical integration? In *The firm as a nexus of treaties*, ed. M. Aoki, B. Gustafsson, and O.E. Williamson. London: Sage.

Riordan, M.H. 1998. Anticompetitive vertical integration by a dominant firm. *American Economic Review* 88: 1232–1248.

Riordan, M.H., and S.C. Salop. 1995. Evaluating vertical mergers: A post-Chicago approach. *Antitrust Law Journal* 63: 513–568.

Riordan, M.H., and O.E. Williamson. 1985. Asset specificity and economic organization. *International Journal of Industrial Organization* 3: 365–378.

Salinger, M.A. 1988. Vertical mergers and market foreclosure. *Quarterly Journal of Economics* 103: 345–356.

Salop, S.C., and D.T. Scheffman. 1987. Cost-raising strategies. *Journal of Industrial Economics* 36: 19–34.

Shelanski, H.A., and P.G. Klein. 1995. Empirical research in transaction cost economics: review and assessment. *Journal of Law, Economics, and Organization* 11: 335–361.

Whinston, M.D. 2003. Transaction cost determinants of vertical integration. *Journal of Law, Economics, and Organization* 19: 1–23.

Williamson, O.E. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Williamson, O.E. 1985. *The economic institutions of capital*. New York: Free Press.

Williamson, O.E. 1996. *The mechanisms of Governance*. New York: Oxford University Press.

Vickrey, William Spencer (1914–1996)

Richard Arnott

Abstract

William Spencer Vickrey was awarded the 1996 Nobel Prize in Economics 'for his fundamental contributions to the economic theory of incentives under asymmetric information'. While best known as the father of auction theory, he made important contributions on a broad range of topics including social choice, marginal cost pricing, the design of tax systems, transportation economics, urban economics, and macroeconomics.

Keywords

Asymmetric information; Auctions (theory); Congestion pricing; Cumulative averaging taxation; Efficient allocation; Estate taxes; Gibbard–Satterthwaite theorem; Groves–-Clark–Vickrey mechanisms; Land use; Land value taxation; Marginal cost pricing; Mirrlees, J. A.; Natural rate of unemployment; Optimal income tax; Philanthropy; Philosophy and economics; Ramsey pricing; Responsive pricing; Revelation of preferences; Second best; Shoup, C. S.; Site value taxation; Social choice; Strategic misrepresentation; Transport economics; Urban economics; User fees; Veil of ignorance; Vickrey second-price auction; Vickrey, W. S.

JEL Classifications

B31

William Spencer Vickrey was awarded the 1996 Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, jointly with James A. Mirrlees, 'for [his] fundamental contributions to the economic theory of incentives under asymmetric information' (Royal Swedish Academy of Sciences, 1996). His most influential papers, as well as a comprehensive bibliography, are contained in *Public Economics: Selected Papers by William Vickrey* (Arnott et al., 1994). Articles that pay tribute to his contributions include Arnott (1998), Drèze (1995), Harriss (2000), and Laffont (2003).

Vickrey was born in Victoria, Canada, where his maternal grandfather had built a group of department stores. Towards the end of the First World War, his father, a Congregational minister, became actively involved in the relief of

Armenian and Greek refugees from the Ottoman Empire. In conjunction with this work, the family moved to New York and later Switzerland. Vickrey attended high school in Scarsdale, NY, and Phillips Andover Academy. He received his bachelor's degree in engineering and mathematics from Yale University in 1935. He then moved to Columbia University for graduate studies in economics. After two years of course work, he went to Washington, DC, to work for the National Resources Committee, in a group pioneering studies on the structure of the US economy. During the Second World War, as a conscientious objector, he served in the Division of Tax Research of the Department of the Treasury, working on broad macroeconomic issues related to war finance and more specific issues of tax structure. In 1946 he returned to Columbia to teach and to complete his doctoral dissertation, published as *An Agenda for Progressive Taxation* (1947). Apart from sabbaticals and missions abroad, he remained at Columbia until his death.

Throughout his career, Vickrey had considerable practical policy experience. As part of Carl Shoup's team, he helped design several countries' tax systems, including Japan's after the Second World War. He also advised many public utilities, and even introduced skip-stop scheduling to the Indian rail service. But he was never a major player on the policy front. His legacy is his body of publications. He published eight longer works, including graduate textbooks in microeconomics and macroeconomics, three technical monographs, and two co-authored country tax system studies, as well as his thesis. Apart from his thesis, however, he is best known for his some 200 papers.

Though covering an unusually broad range of topics, his papers are consistent in theme and style. While his choice of topics stemmed from social and moral concerns, his treatments of them stressed improvements in resource allocation. 'Greater efficiency for the common good' would be an appropriate slogan for his work. His style of writing and reasoning is idiosyncratic and paradoxical. Most of his papers advocate specific policy innovations. To reach a broader audience, he developed his ideas primarily verbally, and with literary flair, but the precision and sophistication of his economic reasoning largely defeated this purpose. He also tended to emphasize practical issues of policy implementation, while presenting in an offhand manner the novel theoretical and conceptual insights for which the papers have been remembered. Many of his policy recommendations, though derived with impeccable logic, were impractical at the time he proposed them. (Technological advances have since rendered some, such as auto congestion pricing and land value taxation, more practical.) These smaller paradoxes can be resolved by understanding Vickrey as a social crusader with a theorist's cast of mind. The larger paradox is that the tension between crusader and theorist was the source first of his creativity, and then of the neglect for many years of much of his work, and ultimately of the distinction of his intellectual legacy. Many of his ideas were overlooked until they were independently discovered many years later, while others lay dormant until their time had come.

Vickrey's major contributions lie in four areas: social choice and resource allocation mechanisms, taxation, marginal cost pricing, and urban transportation.

Vickrey is best known as the father of auction theory, due to his seminal paper, 'Counterspeculation, Auctions, and Competitive Sealed Tenders' (1961). The question posed by Vickrey in that paper is how to achieve efficiency in resource allocation with a small number of buyers or sellers under asymmetric information. He presented and analysed two classes of mechanisms that circumvent the strategic misrepresentation of costs and preferences. The first is auctions. Consider the simplest auction in which there is a single item for sale, for which bidders have different private valuations. Efficiency entails the item being sold to the bidder with the highest valuation. If the item is sold to the highest bidder at his bid, then each bidder has an incentive to bid less than his valuation, since if he bids his valuation he gains no surplus from purchase of the item. In deciding on his bid, each bidder must guess others' valuations, and there is no guarantee that the item will be sold to the bidder with the highest valuation. Suppose instead that the item is sold to the highest bidder but at the *second* highest bid (the Vickrey second-price auction). Whatever other bidders do, if a particular bidder bids more than his valuation he will win more often but only when the second-highest bid, and therefore the price he pays, exceeds his valuation, while if he bids less than his valuation he will win less often and only when the winning bid falls short of his valuation. Since bidding one's private valuation is the dominant strategy, the item should go to the bidder with the highest valuation, achieving efficiency. Auction theory has developed from this insight. Auctions are now extensively used in the allocation of goods with a small number of buyers; timber and drilling rights, bandwidth, Treasury bills and sealed bid tenders are well-known examples. Vickrey's paper also investigated a class of demand-revealing mechanisms – now known as Groves–Clark–Vickrey mechanisms – that elicit truthful revelation of preferences, for pure public goods for example.

Vickrey made several other contributions to the literature on social choice and resource allocation mechanisms. In 'Measuring Marginal Utility by Reactions to Risk' (1945), he provided the first statement of social choice based on the maximization of expected utility behind the veil of ignorance, which was independently stated by Harsanyi (1955) a decade later, and also the first formulation of the optimal income tax problem, which was not solved until a quarter-century later by James Mirrlees (1971). 'Utility, Strategy, and Social Decision Rules' (1960) provides a masterful survey of social choice theory as of that date and conjectures what is now known as the Gibbard–Satterthwaite theorem.

The efficiency of marginal cost pricing in general and of short-run marginal-cost pricing in particular were well understood when Vickrey was a graduate student. Vickrey's contributions were in communicating the breadth of application of the principles and in conceiving ingenious technological schemes for their implementation. From the early 1950s he was a strong advocate of responsive marginal cost pricing, whereby the current price reflects the current realization of stochastic demand and supply; for instance, he proposed varying the parking meter rate on a city block according to the meters' realized occupancy rate (1959) and dealing with airline overbooking via responsive pricing (1972). His crusading for congestion pricing in transportation, for site value taxation (1970), and for the extended application of user fees to finance local public services (1963) are examples of his advocacy of marginal cost pricing in novel contexts.

His major contributions to the theory of taxation derived from his experience in the Department of the Treasury during the Second World War. All are contained in his thesis, which was a tour de force. The goal of the thesis was the comprehensive design of a practical, coherent, fair, and efficient tax system. At the time, a steeply progressive income tax was the primary source of federal tax revenue. Rates had been increased sharply to generate the revenue needed to finance the war effort. The combination of steep progression and high rates encouraged taxpayers to devote considerable effort to altering the timing of expenditures and receipts in order to average income. To eliminate this waste, Vickrey proposed cumulative averaging, a method of taxing individuals on their discounted lifetime incomes, with a minimum of accounting. A steeply progressive estate tax was also in place, which wealthy individuals avoided through generation-skipping trusts. Vickrey proposed an integrated successions tax, which retained the progressivity of the tax while reducing the incentives for complex tax avoidance schemes. Since then, income and estate tax rates have been lowered and made less progressive, mitigating the problems that Vickrey's proposed reforms addressed. His contributions lie not so much in his specific proposals, however, as in his conception of redesigning the tax system from basic principles. How much influence his book had on the Carter Commission in Canada, the Meade Committee in the UK, or the Reagan tax reforms is hard to say, but they are in the same spirit.

While best known for his auctions paper, Vickrey was also the pre-eminent transport economic theorist of his generation. As a transport economist, he is famous for his 50-year-long advocacy of auto congestion pricing (of charging drivers for the externality cost each imposes on other drivers by slowing them down), and in North America at least is known as the father of congestion pricing. After many years of political resistance, urban auto congestion pricing is slowly being adopted, first in Singapore and more recently in London and Stockholm. His first work on the subject (1955) was a proposed fare structure for New York City's subway system, based on marginal cost pricing principles. His empirical research on the project entailed travelling the subway system, stopwatch in hand, while his discussion of deviations from first-best marginal cost pricing to take into account equity concerns and the transit authority's budgetary constraints anticipates the theory of the second best. His second work (1959) detailed an automobile congestion-pricing scheme for downtown Washington, DC. A schedule gives the prices of traversing major intersections by time of day. Each car is equipped with a transponder. When the car enters an intersection, its transponder sends a signal to a roadside receptor and is conveyed to a central computer, which adds the appropriate charge to the car's bill. The theoretical work on the project included an independent derivation of Ramsey pricing with a marginal cost of public funds. His later work in urban transport economics is noteworthy in two respects. He, more than any other urban transport economist, grappled with the complex physics of auto congestion. He also introduced the bottleneck model of traffic congestion (1969), the first analytically tractable model of equilibrium rush-hour traffic dynamics. Each commuter decides when to leave home in the morning, trading off schedule inconvenience against congestion delay. Congestion takes the form of a queue behind a bottleneck of fixed flow capacity, with the queue length (and hence the departure time distribution) evolving to achieve equilibrium.

His other work spans a diversity of topics. Viewing unemployment as a tragic waste of human resources, he wrote many macroeconomic papers arguing against a natural rate of unemployment and for Keynesian macroeconomic stabilization. He also made important contributions to urban economics, most noteworthy of which are pioneering papers on traffic congestion and land use (Solow and Vickrey, 1971) and on the Henry George theorem (1977) – which states that efficient spatial economies can be decentralized via marginal cost pricing, with land rents being used to cover the deficits deriving from the economies of scale underlying agglomeration. His miscellaneous papers cover such topics as game theory, student loans, gerrymandering, international dispute resolution, cost-of-living indices, equivalence scales and sorting theory. One paper on the economics of traffic accidents (1968), another on the economics of philanthropy (1962), and another on economics and philosophy (1950) have been influential. The last of these papers provides insight into the moral purpose underlying Vickrey's work.

Prior to his receipt of the Nobel Prize, Vickrey's work, apart from his justly celebrated auctions paper, was not well known to most economists. But Vickrey is much more than just a 'one-paper economist'. The same intellectual qualities that spawned the auctions paper are evident in the rest of his work. Perhaps the whole of the rest of his work is greater than the sum of the individual papers, demonstrating the wealth of ideas that are generated when a brilliant economic theorist applies his creativity to devising solutions to practical public policy problems.

See Also

- ▶ Auctions (Theory)
- ▶ Congestion
- ▶ Land Tax
- ▶ Marginal and Average Cost Pricing
- ▶ Progressive and Regressive Taxation
- ▶ Utilitarianism and Economic Theory

Selected Works

1945. Measuring marginal utility by reactions to risk. *Econometrica* 13: 319–333.

1947. *An agenda for progressive taxation*. New York: Ronald Press.

1950. Ethics and economics: An exchange of questions between economics and philosophy. In *Goals of economic life*, ed. A.D. Ward. New York: National Council of Churches.

1955. A proposal for revising New York's subway fare structure. *Journal of Operations Research Society of America* 3: 38–68.

1959. Statement on the pricing of urban street use. *Hearings: US Congress, Joint Committee on Metropolitan Washington Problems* 11(November): 466–477.

1960. Utility, strategy, and social decision rules. *Quarterly Journal of Economics* 74: 507–535.

1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16: 8–37.

1962. One economist's view of philanthropy. In *Philanthropy and public policy*, ed. F.G. Dickinson. New York: NBER.

1963. General and specific financing of urban services. In *Public expenditure decisions in the Urban community*, ed. H.G. Schaller. Washington, DC: Resources for the Future.

1968. Automobile accidents, tort law, externalities, and insurance: an economist's critique. *Safety: Law and Contemporary Problems* 33: 464–487.

1969. Congestion theory and transport investment. *American Economic Review* 59: 251–260.

1970. Defining land value for taxation purposes. In *The assessment of land value*, ed. D.-H. Holland. Madison: University of Wisconsin Press.

1971. (With R.M. Solow.) Land use in a long, narrow city. *Journal of Economic Theory* 3: 403–447.

1972. Airline overbooking: some further solutions. *Journal of Transport Economics and Policy* 6: 257–270.

1977. The city as a firm. In *The economics of public services*, ed. M.S. Feldstein and R.F. Inman. New York: Wiley.

Bibliography

Arnott, R. 1998. William Vickrey: Contributions to public policy. *International Tax and Public Finance* 5: 93–113.

Arnott, R., K. Arrow, A. Atkinson, and J. Drèze, eds. 1994. *Public economics: Selected papers by William Vickrey*. Cambridge: Cambridge University Press.

Drèze, J.H. 1995. Forty years of public economics: A personal perspective. *Journal of Economic Perspectives* 9 (2): 111–130.

Harriss, C.L. 2000. William Spencer Vickrey, 1914–1996: Nobel Laureate in Economics. *Economic Journal* 110: 708–719.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Laffont, J.-J. 2003. William Vickrey: A pioneer in the economics of incentives. In *Nobel lectures, economics, 1996–2000*, ed. T. Persson. Singapore: World Scientific Publishing Co.

Mirrlees, J.A. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.

Royal Swedish Academy of Sciences. 1996. Press release, 8 Oct Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1996/press.html. Accessed 18 Jan 2007.

Vind, Karl (1933–2004)

Hans Keiding

Keywords

Control theory; Cores; Exchange equilibrium; Expected utility hypothesis; Fairness; General equilibrium; Lyapunov theorem; Mathematical economics; Optimal control; Statistics; Utility representations; Vind, K; Walrasian equilibria

JEL Classifications

B31

Karl Vind was born in a small provincial town in Denmark on 3 April 1933. His mother died when he was only a few years old, and his father was often absent from home, so Karl and his two brothers were taken care of by relatives. The family reports early interest and skill in economics and mathematics.

Karl Vind finished his school years in 1951. He studied economics at the University of Copenhagen, and attended lectures in mathematics by Werner Fenchel (known for his contributions to the theory of convexity). He graduated in 1958, and after finishing military service was employed at the Faculty of Social Sciences; he also had a position as scientific researcher at what was later known as the Institute of Economics. His future scientific orientation was formed in the years 1962–3, which he spent as a Rockefeller Fellow at the University of California (Berkeley), where he was inspired by the highly fertile research environment around Gerard Debreu. He returned to Berkeley as visiting associate professor from 1964 to 1966. While at Berkeley Karl Vind was married in 1962 to Anni (Mortensen); they had sons named Lars and Jacob and adopted a daughter named Dorthe.

After his return to Copenhagen in 1966 Karl Vind obtained a position as professor in economics, later changed to a chair in mathematical economics, which he held until his retirement in April 2003 at the age of 70. He spent several long periods in Berkeley as well as at the Center for Operations Research and Analysis (CORE) in Louvain. After retirement, he retained his office at the institute and participated in its everyday activities until his death in July 2004 after a short illness.

Karl Vind's published research covers many fields – international trade theory, control theory, general equilibrium, game theory, the theory of choice under uncertainty – so that his publication record is consistent with his own interpretation of mathematical economics as 'the derivative of economic theory' – what is mathematical economics today becomes economic theory tomorrow. However, most of the topics he studied engaged him over long periods; some of the results appearing in the later years were at least partially obtained in the early years of his career.

In the years after graduating, Karl Vind had been interested in mathematical statistics and control theory, something which is witnessed by his work on optimal control with jumps in the state variables (1967). He never returned to control theory, but it usefully inspired him to apply Lyapunov's theorem, which at that time was known to control theorists but not to researchers in general equilibrium. Vind demonstrated that it could be used to show the equivalence of core and Walrasian equilibria in large economies (1964). This was a major breakthrough at the time, achieved independently by Aumann (1964). The approach using Lyapunov's theorem was innovative and offered a new approach to modelling large economies.

Also from this period is the short piece on the core of an exchange economy (1965), which pioneers the extension of the results obtained for economies with infinitely many agents to economies with a finite number of agents. Vind's result does not go all the way to establishing a connection between core and equilibrium, something which was achieved only several years later. This later development might perhaps have been simpler and faster if researchers had followed Vind's early approach; he had the bad luck of being ahead of his time.

In the following years, Vind's published research dealt with extensions of the general equilibrium model in several directions. An example is the paper written with David Schmeidler (1972) on fair net trades, proposing a new approach to the concept of fairness as well as an elegant formalism. In much of his work from this period Karl Vind was concerned with the structural properties of exchanges. His concept of an exchange equilibrium was intended to capture the essential properties of trade in markets. He was, however, not quite satisfied with the initial formulations of the exchange equilibrium, which were never published. After several reformulations the concept appeared in 1983 as 'equilibrium with coordination'.

In the late 1960s Karl Vind started on another project, dealing with utility representations of preferences, which remained at the draft stage until it was finally published as a monograph in 2003. His work on the so-called mean groupoids was inspired by the need to extend the general equilibrium model to include time and uncertainty, which at that time seemed to call for specific functional forms of utility representations. Vind realized that there was a common structure behind utility representations over time and under uncertainty, related to the operation of taking mixtures of consumption programmes, and this led Vind's theory of mean groupoids. The work had already taken shape as a draft for a research monograph around 1970, but its final publication was considerably delayed, partly for practical reasons and partly due to the emergence of new results from Vind himself and others. Some of these had to do with the extension of the expected utility hypothesis to preferences that are not necessarily complete, one of Karl Vind's later research projects.

As a researcher, Karl Vind remained active after retirement as an organizer of and participant in scientific meetings and seminars. His influence on the mathematical economics profession goes beyond his published work, since he took great pleasure in following the work of other researchers, in particular young ones, who received valuable suggestions from a person genuinely interested in their work. Due to this aspect of his scientific activity, he has had a lasting influence on the development of the field.

See Also

► Mathematics and Economics

Selected Works

1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5: 165–177.

1965. A theorem on the core of an economy. *Review of Economic Studies* 32: 47–48.

1967. Control systems with jumps in the state variables. *Econometrica* 35: 273–277.

1972. (With D. Schmeidler.) Fair net trade. *Econometrica* 40: 637–642.

1983. Equilibrium with coordination. *Journal of Mathematical Economics* 12: 275–285.

2003. *Independence, additivity, uncertainty.* Berlin: Springer.

Bibliography

Aumann, R. 1964. Markets with a continuum of traders. *Econometrica* 32: 265–290.

Viner, Jacob (1892–1970)

Henry W. Spiegel

Keywords

Bentham, J.; Chicago school,; Cost theory; Dumping; Economic development; Envelope curve; History of economic thought; International trade theory; Kinked demand curve; Knight, F. H.; Marshall, A.; Mill, J. S.; Monopolistic competition; Partial equilibrium; Religion and economics; Taussig, F. W.; Trade creation; Trade diversion; Utilitarianism; Viner, J.

JEL Classifications

B31

Jacob Viner, the economic theorist and historian of economic thought, was born and raised in Montreal, the son of immigrant parents from eastern Europe. As an undergraduate he attended McGill University, where he was taught economics by Stephen Leacock, the famous humorist. Leacock used texts by Mill and Walker, Milk and Water, as the students referred to them, showing 'good judgment' according to an account that Viner gave later in life. For graduate work he went to Harvard, where he earned a Ph.D. in 1922. He was a student and eventually became a close friend of Frank W. Taussig, the well-known authority on economic theory and international economics. At that time and during the earlier part of Viner's career he and Taussig were rare specimens in what was, except for a very few others, essentially a 'wasp' establishment. But in other respects their background was quite different. Viner was a self-made man who had emancipated himself from the immigrant quarter of Montreal, while Taussig was born into a patrician family with wealth and native culture.

Taussig's specialities were the fields to which Viner himself was drawn and in which he earned great distinction, in addition to his perhaps even more distinguished work in the history of economics, where his accomplishments were almost without rival.

During the two world wars, during the Great Depression, and on and off at other times, Viner did consulting and other work in Washington, but he was foremost an academic, who taught at the University of Chicago in 1916–17 and from 1919 to 1946, when he went to Princeton and taught there until his retirement in 1960. Viner advanced rapidly at Chicago, where the department then was headed by J.M. Clark, and he became a full professor at age 32. A few decades earlier, in the same department, Veblen had risen to the rank of assistant professor only at age 43. But Veblen had defied convention both in his writings and personal life.

Viner's tenure at Chicago coincided in part with his editorship of the *Journal of Political Economy* for a period of 18 years. Most of the time the post was held jointly with Frank H. Knight, who, after having earlier spent two years at Chicago, returned to it in 1927. Both men imprinted on the journal the mark of their own great gifts.

Viner's contributions to economic theory and the history of economic thought are embodied in periodical articles that were reprinted in book form in 1958 under the title *The Long View and the Short.* His contributions to general theory consist principally of two remarkable articles, one published in 1921 and the other ten years later. Of the two, the second on 'Cost Curves and Supply Curves' (Viner 1958, pp. 50–78) made an immediate and powerful impact on the profession. Written, as it was, by a then well-established scholar, it contained virtually the whole of the modern exposition, graphic and otherwise, of the theory of cost, including the envelope curve, about which Viner had a legendary dispute with his mathematically more proficient Chinese draftsman. It also contained, perhaps for the first time in print, the words 'marginal revenue'. All this matter eventually entered into the elementary textbooks. Viner's accomplishment paralleled that of Knight, whose graphic portrayal of the theory of production in *Risk, Uncertainty and Profit* (1921) likewise entered into the mainstream of economic theory and became the basis for the textbook treatment of the matter. Among the two great scholars there was forged a substantial portion of partial equilibrium analysis as it evolved during the first half of the 20th century.

Viner's earlier article, 'Price Policies: The Determination of Market Price', published in 1921 and covering barely five pages in the reprint of 1958, was in some respects an even more dazzling achievement than the later and much better-known one. Five years ahead of Sraffa, six years ahead of the publication of Joan Robinson's and Chamberlin's books on the subject, Viner developed here, in a short paragraph, the outlines of the theory of monopolistic competition. He writes of inflexible prices, 'differentiation' of products, advertising, non-price competition and other characteristics of markets that are neither fully competitive nor completely monopolistic. In such markets producers may succeed in creating a special demand for their products. They can then to some extent determine prices independently of the prices charged by their competitors and still maintain their sales (Viner 1958, pp. 5–6). In the same context Viner also developed, in a few sentences, the theory of what became later known as the kinky demand curve, 18 years ahead of Sweezy's article on the subject.

These were indeed path-breaking contributions, but their existence was virtually ignored until Viner's article was reprinted in 1958. The place of the original contribution – L.C. Marshall, ed., *Business Administration,* University of Chicago Press, 1921 – was not exactly obscure but elusive nevertheless from the standpoint of a reader looking for innovations in economic theory. Chamberlin did not mention Viner in the bibliographies that he appended to successive editions of his book and which eventually listed around 1,500 items. As regards the kinky demand curve, there is no reference to Viner in Sweezy's article in the *Journal of Political Economy* for 1939, of which Viner then was the co-editor, nor in Stigler's critique published in the same journal in 1947. All this is an unresolved puzzle. No one knows why Viner never developed more fully the ideas sketched in his brief article of 1921 and why he, who in other contexts did not shy away from announcing his priority, remained silent about this one. The ideas surely were his own and not derived from an oral tradition at Harvard, whose only potential fount, Allyn Young, came to Harvard only in 1920, when Viner was already teaching at Chicago. He could not very well have anticipated unfriendly criticism, because the Chicago of the 1920s, where J.M. Clark had a senior position, was not the Chicago of the later so-called Chicago School, all of whose leaders, beginning in the mid-1940s, voiced disapproval of the theory of monopolistic competition.

Viner not only had an analytical mind that was stocked with original ideas, but combined with this a stupendous book learning that within the scope of the humanities and social sciences, and especially their history, was virtually universal and gave special depth to his studies in the history of economics. He was perhaps not as scintillating a writer as Schumpeter, nor did he turn out, as did Schumpeter, a comprehensive treatise on the subject, but his work, scattered in periodical articles, contains far more reliable and judicious interpretations of such matters as utilitarianism, and classical and Marshallian economics. The most important of Viner's articles on the history of economics were reprinted in Part II of the collection published in 1958. Their coverage extends all the way from the mercantilists to Marshall and Schumpeter. The essay on mercantilist thought shows the mercantilists in pursuit both of power and wealth as ultimate ends of national policy. Another on Adam Smith demonstrates, among other matters, that Smith was not a doctrinaire advocate of laissez-faire, a quality that he shares with Viner. Smith was a favoured subject of Viner's studies, and in 1965 he contributed an introduction of 145 pages to a new edition of Rae's *Life of Adam Smith,* the standard biography. An essay about the utility concept in value theory defends the concept against its critics. Writing about Bentham and J.S. Mill (1949), Viner clarifies the meaning of the former's hedonic calculus and by restricting it to comparisons between pain and pleasure contributes to the rehabilitation of this concept, for which, he believes, an idea of Benjamin Franklin's may have been the inspiration. Mill and Marshall are both viewed in their Victorian setting. The former's *Principles,* a combination of 'hard-headed rules and utopian aspirations', was 'exactly the doctrine that Victorians of goodwill yearned for'. Marshall fitted into the Victorian age that was complacent about the present and optimistic with respect to future progress.

Except for the collection of his articles published in 1958 and two posthumous publications, all of Viner's books are about international economics, with a collection of his articles in this field, titled *International Economics,* published in 1951. His work in international economics covers virtually all its phases – theory, history of thought, and policy – with occasional use of empirical material. His earliest book, *Dumping* (1923), contained the first comprehensive and systematic study of this subject. It was followed a year later by Viner's doctoral dissertation on *Canada's Balance of International Indebtedness, 1900–1913,* which was written on the suggestion of Taussig, who directed a number of related empirical studies designed to demonstrate the operation of the balance-of-payments adjustment process. In 1937 there was published Viner's masterwork, *Studies in the Theory of International Trade*, which blends in an inimitable manner theoretical analysis and erudite doctrinal history. Its aim was to trace the evolution of the modern theory of international trade. It starts out with the mercantilists and continues with the bullionist controversies, the currency school–banking school controversy, the international mechanism of adjustment, and the doctrine of the gains from trade. In Viner's view, the comparative-cost doctrine is dependent on a real-cost theory of value rather than on opportunity cost. While this view was not in tune with the time, there are many forward-looking sections in the book, including references to a lecture given by Viner in 1931 in which later models of Lerner, Leontief and Hicks were anticipated.

In 1950 Viner published *The Customs Union Issue,* which contained the distinction between trade creation and trade diversion, the starting point of later discussions of the matter. Viner's articles on *International Economics,* collected in 1951, start with one on the most-favoured-nation clause and end with an essay on the economic foundations of international organizations. Many of the articles are indispensable for the study of the policy issues of the time. In 1952 Viner made his contribution to the emerging field of economic development in a book on *International Trade and Economic Development*. In this work he took a far less favourable view of a number of public policies designed to accelerate economic development than was commonly held at that time. He refused to identify agriculture with poverty, stressed that industrialization was more often a consequence than a cause of prosperity, and placed the main burden of promoting development on the underdeveloped country itself.

Viner had for long been interested in theological ideas, especially of the more remote past, and after his retirement he started out on a project designed to explore the relationship between religious and economic thought. This great project proved open-ended. After Viner's death only two fragments were published, one on *The Role of Providence in the Social Order* (1972), and the other on *Religious Thought and Economic Society* (1978). The first of these works is an original accomplishment that traces the derivation of a number of economic ideas from theological precedents, for example, the theory of international trade that is grounded in differences in factor endowments, Smith's invisible hand, and the providential origin of social inequality that was claimed in the past. The second work is written along more conventional lines and reviews the economic doctrines of the Fathers of the Church, of the Scholastics, secularizing tendencies in later Catholic social thought, and Protestantism and the rise of capitalism. This last chapter contains a critical analysis of the Weber–Tawney thesis of the Calvinist origin of capitalism.

To place Viner's work into its proper historical setting, a word is in order about his relation to the Chicago School. A common conception takes his membership or leadership in this school for granted, but this view is mistaken. Viner himself said that much in a remarkable letter to Patinkin written shortly before his death (Patinkin 1981, p. 266; the letter is also reproduced by Reder 1982, p. 7). It must be remembered that at the time when Viner taught at Chicago, the designation 'Chicago School' was not yet a commonly used term. To be sure, Viner's views about laissez-faire, Keynesian economics and government intervention had something in common with the views held by representatives of the Chicago School, but on the whole he was a more pragmatic thinker and more aware of the need for qualification and consideration of circumstances of time and place. Moreover, Viner, from whom stems the definition 'economics is what economists do', would not have felt comfortable within the confines of a school, especially of one that at times has come close to defining economics as the study of competitive markets. The early leaders of what later became known as the Chicago School were Henry Simons and Knight, not Viner. Like no one else, Knight had a charismatic appeal that yielded conversions to libertarianism in his classroom – James Buchanan has testified to this – and that made him the more likely founder of a school. It is significant also that Viner, and, for that matter, Knight too, urged deficit spending during the Great Depression. Viner called the plea for an annual balanced budget a mouldy fallacy (Viner 1933, p. 129). He was critical of Hayek's libertarianism (Viner 1961). He denied that competition was both a norm and normal, pointing out instead that

> monopoly is so prevalent in the markets of the western world today that discussions of the merits of the free competitive market as if that were what we are living with or were at all likely to have the good fortune to live with in the future seem to me academic in the only pejorative sense of that adjective.

He also insisted that 'no modern people will have zeal for the free market unless it operates in a setting of "distributive justice" with which they are tolerably content' (Viner 1960, pp. 66, 68). (The article in which Viner developed these ideas was ostensibly an exposition on the *rhetoric* of laissez-faire, an early exercise in an approach that D.N. McCloskey was to apply on a wider scale more than a quarter century later.) Against Friedman Viner supported discretionary monetary management rather than conduct in conformity with a 'rule' (Viner 1962). And, last but not least, it was Viner who created the substance of the theory of monopolistic competition, which in a peculiar dialectic was later to become the target of the Chicago School.

See Also

- ► Chicago School

Selected Works

1923. *Dumping*: *A problem in international trade.* Chicago: University of Chicago Press.

1924. *Canada's balance of international indebtedness, 1900–1913: An inductive study in the theory of international trade*. Cambridge, MA: Harvard University Press.

1930. *Lectures in price and distribution theory.* Economics 301, University of Chicago, Summer Quarter, 1930; ed. M.D. Ketchum, mimeo, 1931.

1933. Inflation as a remedy for depression. In *Proceedings of the Institute of Public Affairs,* Seventh Annual Session. Athens: University of Georgia.

1937. *Studies in the theory of international trade. New York: Harper.*

1949. Bentham and J.S. Mill: The utilitarian background. *American Economic Review* 39, 360–82.

1950. *The customs union issue.* New York: Carnegie Endowment for International Peace; London: Stevens.

1951. *International economics*: *Studies. Glencoe: Free Press.*

1952. *International trade and economic development*: *Lectures delivered at the National University of Brazil. Glencoe: Free Press; Oxford: Clarendon Press.*

1958. *The long view and the short*: *Studies in economic theory and policy. Glencoe: Free Press. (With bibliography).*

1960. The intellectual history of laissez faire. *Journal of Law and Economics* 3: 45–69.

1961. Hayek on freedom and coercion. *Southern Economic Journal* 2(3): 230–236.

1962. The necessary and the desirable range of discretion to be allowed to a monetary authority. In *In search of a monetary constitution,* ed. L.B. Yeager. Cambridge, MA: Harvard University Press.

1963. Review article on C.B. Macpherson's *Political Theory of Possessive Individualism: Hobbes to Locke. Canadian Journal of Economics and Political Science* 29: 548–559.

1965. Guide to John Rae's 'Life of Adam Smith'. Introduction to J. Rae, *Life of Adam Smith.* New York: Kelley.

1972. *The role of providence in the social order*: *An essay in intellectual history. Philadelphia: American Philosophical Society.*

1978. *Religious thought and economic society*: *Four chapters of an unfinished work,* ed. J. Melitz and D. Winch. Durham: Duke University Press. Also published in *History of Political Economy* 10: 9–189.

Bibliography

Baumol, W.J., and E.V. Seiler 1979. Jacob Viner. In *International encyclopedia of the social sciences,* vol. 18, Biographical Supplement. New York: Free Press.

Davis, J.R. 1971. *The new economics and the old economists*. Ames: Iowa State University Press.

Machlup, F. 1972. What was left on Viner's desk. *Journal of Political Economy* 80: 353–364.

Machlup, F., P.A. Samuelson, and W.J. Baumol. 1972. In Memoriam, Jacob Viner (1892–1970). *Journal of Political Economy* 80: 1–15.

Patinkin, D. 1981. *Essays on and in the Chicago tradition*. Durham: Duke University Press.

Reder, M.W. 1982. Chicago economics: Permanence and change. *Journal of Economic Literature* 20: 1–38.

Robbins, L. 1970. *Jacob Viner: A tribute*. Princeton: Princeton University Press.

Samuels, W.J., ed. 1976. *The Chicago School of political economy.* Published jointly by the Association for Evolutionary Economics and Division of Research, Graduate School of Business Administration, Michigan State University, East Lansing.Winch, D. 1981. Jacob Viner. American Scholar 50, 519–525.

Vintage Capital

Raouf Boucekkine, David de la Croix and Omar Licandro

Abstract

This article reviews the early vintage capital literature of the 1960s, and identifies the factors behind the revival of topic from the 1990s.

The fundamental properties of the seminal vintage capital growth models are disentangled, and the origins of the associated controversy on the importance of embodied technical progress are evoked. The recent revival of this literature is analysed with special emphasis on the rising support for the Solowian view of investment following Gordon's 1990 fundamental work on the price of durable goods, and the emergence of a new vintage human capital literature devoted to some fundamental economic growth issues.

Keywords

Capital–labour substitution; Cobb–Douglas technology; Cohort size; Consumption smoothing; Demographic transition; Demography; Denison, E.; Depreciation; Diffusion of technology; Embodiment hypothesis; Fertility; Human capital; Inequality (explanations); Information technology revolution; Innovation; Investment cycles; Leontief technology; Life expectancy; Neoclassical growth theory; Neoclassical production functions; Obsolescence; Overlapping generations models; Phelps, E.; Population growth; Productivity slowdown; Quasi-rent; Replacement echoes; Solow, R.; Technical progress; Transition dynamics; Vintage capital; Vintage human capital

JEL Classifications

E21

In neoclassical growth theory capital is assumed to be homogeneous and technical progress disembodied, meaning that all capital units benefit equally from any technological improvement. The disembodied nature of technical progress looks unrealistic, as acknowledged by Solow (1960, p. 91):

> This conflicts with the casual observation that many if not most innovations need to be embodied in new kinds of durable equipment before they can be made effective. Improvements in technology affect output only to the extent that they are carried into practice either by net capital formation or by the replacement of old-fashioned equipment by the latest models...

Accounting for the age distribution of capital is a way to cope with this criticism. It actually inspired an important stream of the growth literature of the 1950s and 1960s, giving birth to vintage capital theory.

An economy is said to have a *vintage capital* structure if machines and equipment belonging to separate generations have different productivity, or face different depreciation schedules as in Benhabib and Rustichini (1991). Let us denote by $I(v)$ the number of machines of vintage v. With zero physical depreciation, vintage technology v is

$$Y(v,\ \ t) = F(I(v), L(v,t), e^{\gamma v}),$$

where $Y(v,\ t)$ is the output of vintage v at time $t \geq v$ and $L(v,\ t)$ is the amount of labour assigned to this vintage. Parameter $\gamma > 0$ designates the rate of technical progress, which is said to be *embodied* since it benefits only vintage v. $F(.)$ has the properties of a neoclassical production function. Vintages produce the same final good

$$Y(t) = \int_{t-T(t)}^{t} Y(v,\ \ t)\ \ dv,$$

where $Y(t)$ is total production and $T(t)$ is the lifetime of the oldest operative vintage.

Besides realism, vintage capital models were initially thought to be able to generate quite different long-run properties and short-term dynamics from neoclassical growth models. Because the productivity gap between new and old vintages is increasing over time, the latter need not be operated for ever, and, contrary to the neoclassical growth theory, the lifetime of capital goods might well be finite (Johansen 1959). Such a property was thought to involve non-monotonic transition dynamics governed by the replacement of scrapped goods, known as the replacement echoes principle, which again sharply departs from neoclassical growth models.

On more general ground, vintage capital models were at the heart of the embodiment controversy, which opposed Solow to some leading growth theorists and empiricists, among them Phelps (1962) and Denison (1964). While the former argued that accounting for the fraction of technological progress which is exclusively conveyed by capital accumulation (that is, embodied technical progress) is important to accounting for growth, Phelps argued that the decomposition of technical progress is irrelevant in the long run. Recent studies notably by Gordon (1990) have resuscitated this controversy, as we shall see. Before developing all these themes, it should be noted that, whereas early vintage capital theory primarily focused on physical capital accumulation, recent contributions have taken the same view of human capital accumulation (see Chari and Hopenhayn 1991). Vintage human capital is generated either by successive vintages of technologies which induce specific skills or by demographic conditions. Such contributions have brought out a new and quite appealing understanding of the mechanisms behind technology diffusion and demographic transitions, for example. We briefly review them also.

The Lifetime of Capital

In Johansen (1959), technology is 'putty-clay', meaning that capital–labour substitution is permitted *ex ante* but not once capital is installed. Technological progress is assumed to be labour-saving. Because factor proportions are fixed *ex post*,

$$Y(v,\ \ t) = F(I(v),\ \ e^{\gamma v}L(v,\ \ t)) = g(\lambda(v))\ \ I(v),$$

where the labour–capital ratio $\lambda(v)$ and the size of the capital stock $I(v)$ are both decided at the time of installation, and employment is $L(v,\ \ t) = \lambda(v)$ $e^{\gamma v}I(v)$. In Johansen, *obsolescence* determines the range of active vintages. Quasi-rents of vintage v at date t are proportional to $g(\lambda(v)) - \lambda(v)$ $e^{\gamma v}w(t)$, where $w(t)$ is the equilibrium wage. Since wages are permanently growing as a direct consequence of technical progress, quasi-rents are decreasing. Machines of vintage v are operated as long as their quasi-rents remain positive. Consequently, the scrapping age is defined by $T = t^{*} - v$ where $g(\lambda(v)) = \lambda(v)\ e^{\gamma v}w(t^{*})$. Therefore, Johansen's framework leads to an endogenous, finite lifetime of capital.

Replacement Echoes

If capital lifetime is finite, there might be a room for replacement echoes, as mentioned above. Solow et al. (1966) examine this question in the simpler case of a Leontief technology, when factor substitution is not allowed either *ex ante* or *ex post*. In such a case, $Y(v, t) = Y(v) = I(v) = e^{\gamma v}$ $L(v)$, for all $t \geq v$. One unit of vintage capital v produces one unit of output once combined with $e^{-\gamma v}$ units of labour.

Technical progress is embodied and takes the form of a decreasing labour requirement. For the same reasons as in Johansen, capital goods are scrapped in finite time. Using in addition a constant saving rate, and some technical assumptions, Solow et al. show convergence to a unique balanced growth path, delivering the same qualitative asymptotic behaviour as the neoclassical growth model. This was quite disappointing, since under finite lifetime one would have expected an investment burst from time to time, giving rise to replacement echoes.

Let us normalize the labour supply to unity. From labour market clearing, $\int_{t-T(t)}^{t} L(v)\ \ dv = 1$. Under constant lifetime, time differentiation of the equilibrium condition yields $L(t) = L(t - T)$, implying that investment is mainly driven by replacement activities. When obsolete capital is destroyed, new investments are needed to replace the scrapped machines, creating enough jobs to clear the labour market. As a direct consequence, job creation and investment have a periodic behaviour, implying that investment cycles are reproduced again and again in the future.

Solow et al. (1966) did not find echoes because of the constant saving rate assumption, which completely decouples investment from replacement. In an optimal growth model with linear utility and the same technological assumptions, Boucekkine et al. (1997) show (finite time) convergence to a constant lifetime, letting replacement echoes operate and generate everlasting fluctuations in investment, output and consumption. Under strictly concave preferences, fluctuations do arise in the short run but get dampened in the long run by consumption smoothing (see

Boucekkine et al. 1998). Therefore, the short-run dynamics of vintage capital models differ strikingly from the neoclassical growth model, provided capital and labour are to some extent complementary, consistently with the observed dynamics of investment at both the plant level (Doms and Dunne 1998) and the aggregate level (Cooper et al. 1999). Non-monotonic behaviour has also been shown by Benhabib and Rustichini (1991) for vintage models with non-geometric depreciation.

The Embodiment Hypothesis

A crucial property of vintage capital models is the embodied nature of technological progress: the incorporation of innovations into the production process cannot be achieved without the acquisition of the new vintages which are their exclusive material support. According to Solow (1960), embodiment can have crucial implications for growth accounting. To make the point, he considers a Cobb–Douglas vintage technology

$$Y(v,t) = [\mathrm{e}^{\gamma v}\ I(v)]^{1-\alpha}\ L(v,\ \ t)^{\alpha},$$

and the capital–labour ratio adjusts continuously. The embodiment hypothesis takes the form of quality adjustments, with capital's quality growing at rate γ. In sharp contrast to Johansen, capital lifetime needs not be finite, since under Cobb–Douglas technology any wage cost could be covered by assigning arbitrarily small amounts of labour.

A striking outcome of Solow's model is its aggregation properties. Denote by $L(t)$ the total labour supply, and define quality adjusted capital as

$$K(t) = \int_{-\infty}^{t} \mathrm{e}^{\gamma v} I(v)\ \ dv. \qquad (1)$$

Since marginal labour productivity equalizes across vintages, aggregate output becomes

$$Y(t) = K(t)^{1-\alpha} L(t)^{\alpha}.$$

Aggregate vintage technology in Solow (1960) degenerates into a neoclassical production function. However, by differentiating (1), the motion law for capital is slightly different

$$K'(t) = \mathrm{e}^{\gamma t} I(t),$$

reflecting embodied technical change. Since $\mathrm{e}^{-\gamma t}$ measures the relative price of investment goods at equilibrium, the value of capital is by definition $A(t) = e^{-\gamma t}K(t)$, and evolves following

$$A'(t) = I(t) - \gamma A(t).$$

Technological progress operates as a steady improvement in equipment quality, which in turn implies obsolescence of the previously installed capital. In Solow, obsolescence does not show up through finite time scrapping but through labour reallocation reflecting a declining value of capital.

This important point has been at the heart of recent literature on the *productivity slowdown* and the *information technology revolution* (see Whelan 2002). Indeed, the potential implications for growth of embodied technical progress were tremendously controversial in the 1960s. In a famous statement, Denison (1964) claimed 'the embodied question is unimportant'. His argument was merely quantitative and restricted embodiment to changes in the average age of capital in a one-sector growth accounting exercise. In particular, his reasoning omits *de facto* the relative price of capital channel. Greenwood et al. (1997), by using Gordon's (1990) estimates of the relative price of equipment, quantitatively evaluate the two-sector Solow model, claiming that around 60 per cent of US per-capita growth is due to embodied technical change. As pointed out by Hercowitz (1998), Gordon's series have been good news for the Solowian view.

Vintage Human Capital

The vintage capital growth literature typically considers labour as a homogenous good. However, just as physical capital is heterogenous, so too is the labour force. The concept of vintage human capital was explicitly used in the 1990s to treat some specific issues related to technology diffusion, inequality and economic demography.

In a world with a continuous pace of innovations, a representative individual faces the typical question of whether to stick to an established technology or to move to a new and better one. The trade-off is the following: switching to the new technique would allow him to employ a more advanced technology but he would lose the expertise, the *specific human capital*, accumulated on the old technique. In Chari and Hopenhayn (1991) and Parente (1994), individuals face exactly this dilemma. In such frameworks the generated vintage human capital distributions essentially mimic the vintage distribution of technologies, the time sequence of innovations being generally exogenously given. Chari and Hopenhayn (1991) consider a two-period overlapping generations model where different vintage technologies, operated by skilled and unskilled workers, coexist. Old workers are experts in the specific vintage technology they have run when young. The degree of complementary between skilled and unskilled labour affects negatively the velocity of technological diffusion, since young individuals have strong incentives to invest in old technologies when their unskilled labour endowment is highly complementary to the skilled labour of the old.

Jovanovic (1998) argues that vintage capital models are particularly well suited to explain income disparities across individuals and across countries. The main mechanism behind them is the following. Under the assumption that machines' quality and labour's skill are complementary, the best machines are assigned to the best-skilled individuals, exacerbating inequality. If reassignment is frictionless, then the best-skilled workers are immediately assigned to the frontier technology, the second-best go to the machines just below the frontier, and so on. Even though it is at odds with Chari and Hopenhayn, where adoption costs induce a much slower switching of technologies, frictionless reassignment has the virtue, consistent with cross-country evidence, of implying persistent inequality, in contrast to Parente (1994), which bears leapfrogging.

On the theoretical side, Jovanovic makes an important contribution to the vintage capital literature to the extent that he addresses the hard problem of combining vintage physical capital and vintage human capital in a framework where the vintage distributions of both assets are endogenous. Jovanovic uses an assignment model *à la* Sattinger (1975) to solve this difficult problem. Firms combine machines and workers in fixed proportions, say one machine for one worker, at every instant. Because labour resources are fixed, the latter fixed-proportions assumption implies that old machines become unprofitable at a finite time, as in Johansen. Vintage human capital comes from human capital accumulation *à la* Lucas (1988): the growth of the stock of human capital determines the maximal quality of human capital available: if the worker has human capital, h, and works a fraction of time u (in production), then her skill is given by $s = u\ h$. The typical assignment problem of a firm having acquired capital of a given vintage is to find the optimal vintage human capital or skill of the associated worker (via profit maximization), which makes it possible to achieve the pairing of skills and machines on the basis of the persistent inequality mechanism outlined above.

Demographics

One likely channel through which demographics affect growth is the size, quality and composition of the workforce. From this perspective, generations of workers can be understood as being vintages of human capital. In a continuous-time overlapping generations framework, Boucekkine et al. (2002) model the vintage specificity of human capital from schooling decisions. Individuals optimally decide how many years to spend at school as well as their retirement age; life expectancy has a positive effect on both because of its beneficial impact on the return to education. In such a framework, the vintage specificity of human capital depends, not on technological vintages as in Chari and Hopenhayn (1991), but on cohort-specific demographic characteristics, including education.

The observed relation between demographic variables, such as mortality, fertility and cohort sizes, and growth is anything but linear. Since a key element is between-generation differences in human capital, these nonlinearities may be modelled by the mean of a vintage structure of population. Boucekkine et al. (1998) generate nonlinear relationships between economic growth and both population growth and life expectancy. A longer life, for example, has several conflicting effects. On the one hand it increases the incentives to acquire education and reduces the depreciation rate of aggregate human capital. But on the other, an older population, which finished its schooling a long time ago, is harmful for economic growth.

Conclusion

After a relatively long stagnation, the vintage capital literature, which was a fundamental growth area in the 1960s, has been experiencing a revival since the early 1990s. This revival is due to several factors, among them the rising support for the Solowian view of investment following Gordon's fundamental work on the price of durable goods, the emergence of a new vintage capital growth theory led by Benhabib and Rustichini (1991) relying on a novel and appropriate mathematical set-up, and notably the increasingly common view that some fundamental economic growth issues (like technology diffusion, for example) do require the vintage structure to be better appraised. Of course, many tasks within this new literature remain to be addressed. In particular, much work is needed to bring the vintage models closer to the data. The work of Gilchrist and Williams (2000) is fundamental is this respect.

See Also

- ▸ Diffusion of Technology

Bibliography

Benhabib, J., and A. Rustichini. 1991. Vintage capital, investment, and growth. *Journal of Economic Theory* 55: 323–339.

Boucekkine, R., D. de la Croix, and O. Licandro. 2002. Vintage human capital, demographic trends and growth. *Journal of Economic Theory* 104: 340–375.

Boucekkine, R., M. Germain, and O. Licandro. 1997. Replacement echoes in the vintage capital growth model. *Journal of Economic Theory* 74: 333–348.

Boucekkine, R., M. Germain, O. Licandro, and A. Magnus. 1998. Creative destruction, investment volatility and the average age of capital. *Journal of Economic Growth* 3: 361–384.

Chari, V., and H. Hopenhayn. 1991. Vintage human capital, growth, and the diffusion of new technology. *Journal of Political Economy* 99: 1142–1165.

Cooper, R., J. Haltiwanger, and L. Power. 1999. Machine replacement and the business cycle: Lumps and bumps. *American Economic Review* 84: 921–946.

Denison, E. 1964. The unimportance of the embodied question. *American Economic Review: Papers and Proceedings* 54: 90–94.

Doms, M., and T. Dunne. 1998. Capital adjustment patterns in manufacturing plants. *Review of Economic Dynamics* 1: 409–429.

Gilchrist, S., and J. Williams. 2000. Putty-clay and investment: A business cycle analysis. *Journal of Political Economy* 108: 928–960.

Gordon, R.J. 1990. *The measurement of durable goods prices*. Chicago: University of Chicago Press.

Greenwood, J., Z. Hercowitz, and P. Krusell. 1997. Long-run implications of investment specific technological change. *American Economic Review* 87: 342–362.

Hercowitz, Z. 1998. The embodiment controversy: A review essay. *Journal of Monetary Economics* 41: 217–224.

Johansen, L. 1959. Substitution versus fixed production coefficients in the theory of economic growth. *Econometrica* 27: 157–176.

Jovanovic, B. 1998. Vintage capital and inequality. *Review of Economic Dynamics* 1: 497–530.

Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.

Parente, S. 1994. Technology adoption, learning-by-doing, and economic growth. *Journal of Economic Theory* 63: 346–369.

Phelps, E. 1962. The new view of investment: A neoclassical analysis. *Quarterly Journal of Economics* 76: 548–567.

Sattinger, M. 1975. Comparative advantage and the distribution of earnings and abilities. *Econometrica* 43: 455–468.

Solow, R. 1960. Investment and technological progress. In *Mathematical methods in social sciences 1959*, ed. K. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

Solow, R., J. Tobin, C. Von Weizsacker, and M. Yaari. 1966. Neoclassical growth with fixed factor proportions. *Review of Economic Studies* 33: 79–115.

Whelan, K. 2002. Computers, obsolescence and productivity. *The Review of Economics and Statistics* 84: 445–461.

Vintages

Dale W. Jorgenson

Abstract

Vintages are durable goods acquired at different points of time. The acquisition prices for capital goods of each vintage at each point of time together with investments of all vintages at each point of time constitute the basic data on quantities and prices. These data can be employed in generating the complete vintage accounting system.

Keywords

Capital measurement; Capital services; Depreciation; Durable goods model of production; Investment; Jorgenson, D. W.; Price–quantity duality; Production functions; Rate of return; Vintage accounting; Vintages

JEL Classifications

E21

Investment represents the acquisition of capital goods at a given point of time. The quantity of investment is measured in the same way as the durable goods themselves. For example, investment in equipment is the number of machines of a given specification and investment in structures is the number of buildings of a particular description. The price of acquisition of a durable good is the unit cost of acquiring a piece of equipment or a structure.

By contrast with investment, capital services are measured in terms of the use of a durable good for a stipulated period of time. For example, a building can be leased for a period of years, an automobile can be rented for a number of days or weeks, and computer time can be purchased in seconds or minutes. The prices of the services of a durable good is the unit cost of using the good for a specified period.

Aggregation over Vintages

We can refer to durable goods acquired at different points of time as different *vintages* of capital. The flow of capital services is a quantity index of capital inputs from durable goods of different vintages. Under perfect substitutability among the services of durable goods of different vintages, the flow of capital services is a weighted sum of past investments. The weights correspond to the relative efficiencies of the different vintages of capital.

The durable goods model of production is characterized by price-quantity duality. The rental price of capital input is a price index corresponding to the quantity index given by the flow of capital services. The rental prices for all vintages of capital are proportional to the price index for capital input. The constants of proportionality are given by the relative efficiencies of the different vintages of capital.

We develop notation appropriate for the intertemporal theory of production by attaching time subscripts to the variables that occur in the theory. We can denote the quantity of output at time t by y_t and the quantities of J inputs at time t by x_{jt} $(j = 1, 2, \ldots, J)$. Similarly, we can denote the price of output at time t by q_t and the prices of the J inputs at time t by p_{jt} $(j = 1, 2, \ldots, J)$.

In order to characterize capital as a factor of production, we require the following additional notation:

A_t – quantity of capital goods acquired at time t.

$K_{t,\tau}$ – quantity of capital services from capital goods of age τ at time t.

$p_{A,t}$ – price of acquisition of new capital goods at time t.

$p_{Kt,\tau}$ – rental price of capital services from capital goods of age τ at time t.

To present the durable goods model of production we first assume that the production function, say *F*, is homothetically separable in the services of different vintages of capital:

$$y_t = F\left[G\left(K_{t,0}, K_{t,1} \ldots K_{t,\tau} \ldots\right), x_{2t} \ldots x_{Jt}\right]. \tag{1}$$

Where K_t is the flow of capital services, we can represent this quantity index of capital input as follows:

$$K_t = G,$$

where the function *G* is homogeneous of degree one in the services from capital goods of different ages.

If we assume that the quantity index of capital input K_t is characterized by perfect substitutability among the services of different vintages of capital, we can write this index as the sum of these services:

$$K_t = \sum_{\tau=0}^{\infty} K_{t,\tau}.$$

Under the additional assumption that the services provided by a durable good are proportional to initial investment in this good, we can express the quantity index of capital input in the form:

$$K_t = \sum_{\tau=0}^{\infty} d_\tau A_{t-\tau}. \tag{2}$$

The flow of capital services is a weighted sum of past investments with weights given by the relative efficiencies $\{d_\tau\}$ of capital goods at different ages.

Under constant returns to scale we can express the price of output as a function, say *Q*, of the prices of all inputs. The price function *Q* is homothetically separable in the rental prices of different vintages of capital:

$$q_t = Q\left[P\left(p_{K,t,0}, p_{K,t,1} \ldots p_{K,t,\tau} \ldots\right), p_{2t} \ldots p_{Jt}\right]. \tag{3}$$

Where $p_{K,t}$ is a price index of capital services, we can represent this index as follows:

$$p_{K,t} = P,$$

where the function *P* is homogeneous of degree one in the rental prices of capital goods of different ages.

Under perfect substitutability among the services of different vintages of capital, we can write the price index of capital input *P* as the price of the services of a new capital good:

$$p_{K,t} = p_{K,t,0}.$$

Under the additional assumption that the services provided by a durable good are proportional to the initial investment, we can express the rental prices of capital goods of different ages in the form:

$$p_{K,t,\tau} = d_\tau p_{K,t}, \quad (\tau = 0, 1, \ldots). \tag{4}$$

The rental prices are proportional to the rental price of capital input with constants of proportionality given by the relative efficiencies $\{d\tau\}$ of capital goods of different ages.

Given the quantity of capital input K_t, representing the flow of capital services, and the price of capital inputs $p_{K,t}$, representing the rental price, capital input plays the same role in production as any other input. We next derive the prices and quantities of capital inputs from the prices and quantities for acquisition of durable goods $p_{A,t}$ and A_t.

Vintage Accounting

We begin our description of the measurement of capital input with the quantities estimated by the perpetual inventory method. Taking the first difference of the expression for capital stock in terms of past investments (2), we obtain:

$$\begin{aligned} K_t - K_{t-1} &= A_t + \sum_{\tau=1}^{\infty} (d_\tau - d_{\tau-1}) A_{t-\tau}, \\ &= A_t - R_t, \end{aligned}$$

where R_t is the level of replacement requirements in period t. The change in capital stock from period to period is equal to the acquisition of investment goods less replacement requirements.

We turn next to a description of the price data required for the measurement of the price of capital input. There is a one-to-one correspondence between the vintage quantities that appear in the perpetual inventory method and the prices that appear in our vintage price accounts. To bring out this correspondence we use a system of present or discounted prices. Taking the present as time zero, the discounted price of a commodity, say q_t, multiplied by a discount factor:

$$q_t = \prod_{s=1}^{t} \frac{1}{1+r_s} p_t.$$

The notational convenience of present or discounted prices results from dispensing with explicit discount factors in expressing prices for different time periods.

In the correspondence between the perpetual inventory method and its dual or price counterpart the price of acquisition of a capital good is analogous to capital stock. The price of acquisition, say $q_{A,t}$ is the sum of future rental prices of capital services, say $q_{K,t}$, weighted by the relative efficiencies of capital goods in all future periods:

$$q_{A,t} = \sum_{\tau=0}^{\infty} d_\tau q_{K,t+\tau+1} \tag{5}$$

This expression may be compared with the corresponding expression (2) giving capital stock as a weighted sum of past investments.

Taking the first difference of the expression for the acquisition price of capital goods in terms of future rentals (5), we obtain:

$$\begin{aligned} a_{A,t} - q_{A,t-1} &= -q_{K,t} - \sum_{\tau=1}^{\infty} (d_\tau - d_{\tau-1}) q_{K,t+\tau} \\ &= -q_{K,t} + q_{D,t}, \end{aligned}$$

where $q_{D,t}$ is depreciation on a capital good in period t. The period-to-period change in the price of acquisition of a capital good is equal to depreciation less the rental price of capital. Postponing the purchase of a capital good makes it necessary to forgo one period's rental and makes it possible to avoid one period's depreciation. In the correspondence between the perpetual inventory method and its price counterpart, investment corresponds to the rental price of capital and replacement corresponds to depreciation.

We can rewrite the expression for the first difference of the acquisition price of capital goods in terms of undiscounted prices and the period-to-period discount rate:

$$p_{K,t} = p_{A,t-1} r_t + p_{D,t} - \left(p_{A,t} - p_{A,t-1}\right), \tag{6}$$

where $p_{A,t}$ is the undiscounted price of acquisition of capital goods, $p_{K,t}$ the price of capital services, $p_{D,t}$ depreciation, and r_t the rate of return, all in period t. The price of capital services $p_{K,t}$ is the sum of return per unit of capital $p_{A,\ t\ -\ 1} r_t$, depreciation *pD,t*, and the negative of revaluation, $p_{A,\ t} - p_{A,\ t\ -\ 1}$. To apply this formula we require a series of undiscounted acquisition prices for capital goods $p_{A,t}$, rates of return r_t, depreciation on new capital goods, $p_{D,t}$, and revaluation of existing capital goods $p_{A,\ t} - p_{A,\ t\ -\ 1}$.

To calculate the rate of return in each period we set the formula for the rental price $p_{K,t}$ times the quantity of capital K_{t-1} equal to property compensation. All of the variables entering this equation – current and past acquisition prices for capital goods, depreciation, revaluation, capital stock and property compensation – except for the rate of return, are directly observable. Replacing these variables by the corresponding data we solve this equation for the rate of return. To obtain the capital service price itself we substitute the rate of return into the original formula along with the other data. This completes the calculation of the service price.

In the perpetual inventory method data on the quantity of investment goods of every vintage are used to estimate capital formation, replacement requirements and capital stock. In the price counterpart of the perpetual inventory method data on the acquisition prices of investment goods of every vintage is required. In the full price–quantity duality that characterizes the vintage accounts, capital stock corresponds to the acquisition price of durable goods and investment corresponds to the rental price of capital services.

Conclusion

The distinguishing feature of capital as a factor of production is that durable goods contribute capital services to production at different points of time. The services provided by a given durable good are proportional to the initial investment. In addition, the services provided by different durable goods at the same point of time are perfect substitutes. The weights correspond to the relative efficiencies of the different vintages of capital. The durable goods model of production was originated by Walras (1954) and is discussed in greater detail by Jorgenson (1973) and Diewert (1980).

The durable goods model is characterized by price–quantity duality. The rental price of capital input is a price index corresponding to the quantity index given by the flow of capital services. The rental prices for all vintages of capital are proportional to the price index for capital input. The constants of proportionality are given by the relative efficiencies of the different vintages of capital. The dual to the durable good model of production was introduced by Hotelling (1925) and Haavelmo (1960). The dual to this model has been further developed by Arrow (1964) and Hall (1968).

The acquisition prices for capital goods of each vintage at each point of time together with investments of all vintages at each point of time constitute the basic data on quantities and prices. These data can be employed in generating the complete vintage accounting system originated by Christensen and Jorgenson (1973) and described by Jorgenson (1980). Price and quantity data that we have described for a single durable good are required for each durable good in the system. These data are used to derive price and quantity indexes for capital input in the theory of production presented in the entry on production functions.

See Also

- ► Production Functions
- ► Technical Change

Bibliography

Arrow, K.J. 1964. Optimal capital policy, the cost of capital, and myopic decision rules. *Annals of the Institute of Statistical Mathematics* 16: 16–30.

Christensen, L.R., and D.W. Jorgenson. 1973. Measuring economic performance in the private sector. In *The measurement of economic and social performance*, ed. M. Mess. New York: Columbia University Press for the National Bureau of Economic Research.

Diewert, W.E. 1980. Aggregation problems in the measurement of capital. In *The measurement of capital*, ed. D. Usher. Chicago: University of Chicago Press.

Haavelmo, T. 1960. *A study in the theory of investment*. Chicago: University of Chicago Press.

Hall, R.E. 1968. Technical change and capital from the point of view of the dual. *Review of Economic Studies* 35 (1): 35–46.

Hotelling, H.S. 1925. A general mathematical theory of depreciation. *Journal of the American Statistical Association* 20: 340–353.

Jorgenson, D.W. 1973. The economic theory of replacement and depreciation. In *Econometrics and economic theory*, ed. W. Sellekaerts, 189–221. New York: Macmillan.

Jorgenson, D.W. 1980. Accounting for capital. In *Capital, efficiency, and growth*, ed. G. von Furstenberg, 251–319. Cambridge: Ballinger.

Walras, L. 1874. *Elements of pure economics*. Trans. W. Jaffé. Homewood: Irwin, 1954.

Virtual Economy

Clifford G. Gaddy

Abstract

The virtual economy was the system of informal rent distribution that arose in postSoviet Russia in the 1990s as nonviable Soviet-era manufacturing industries sought to protect themselves from the discipline of the market. Enterprise directors and their allies throughout the economy (including government officials) colluded to use nonmarket prices and various forms of nonmonetary exchange such as barter to transfer value from resource sectors to manufacturing industry. The article discusses the system's historical roots, describes some of its characteristic phenomena, and outlines a model for behaviour of enterprises.

Keywords

Arrears; Barter; Barter in transition; Command economy; Corruption; Non-market prices; Non-monetary exchange; Relational capital; Rent sharing; Soviet Union, economics in; Tax offsets; Virtual economy

JEL Classifications

P3

The virtual economy was the name given to the system of informal rent sharing or value distribution that prevailed in Russia in the 1990s. Featuring widespread use of nonmonetary exchange and nonmarket prices to conceal transfers of value, especially from resource sectors to manufacturing industry, the virtual economy reached a peak in the run-up to the country's financial crisis in August 1998.

The strategies used by enterprise directors to participate in this nonmonetary economy fundamentally changed the behaviour of hundreds and thousands of noncompetitive manufacturing enterprises in Russia during the transition process. The behavioural adaptation permitted enterprises to survive in the transition environment where they ought to have failed. The expectation had been that when the old Soviet industrial structure was shocked by the sudden collapse of central planning and the subsequent launching of radical market reforms – including mass privatization and elimination of overt subsidies – economic agents would be forced to change their behaviour to become competitive in a market economy. The transition was thus intended as a Darwinian process whereby only those enterprises that could transform themselves into competitive operations would survive. But in the case of Russia, the dinosaurs survived – without restructuring. They did change, but instead of adapting *to* the market, they changed to protect themselves *from* the market.

In essence the virtual economy was a peculiar system of rent distribution in which the primary vehicle through which agents laid claim to rents was *production.* The virtual economy was the set of informal institutions that facilitated the production of goods that were value-subtracting, that is, worth less than the value of the inputs used to produce them. Enterprises were able to engage in such production because they had recipients who were willing to accept fictitious (nonmarket) pricing of the goods at levels that masked their lack of profitability. Buyers and sellers colluded to hide the fictitious nature of the pricing. In the classic form of the virtual economy, they did so by avoiding money, instead using barter and other forms of nonmonetary exchange, as well as even more intricate subterfuges.

Since value was being destroyed as the system operated, there had to be a source of value. The ultimate 'value pump' in Russia was the fuel and energy sector, above all one single company, Gazprom – Russia's natural gas monopoly. In exchange for the rights to keep what it earned from exports, Gazprom pumped value into the system by supplying gas without being paid for it (or, more generally, at a cost that was low enough to keep enterprises operating). Gazprom subsidies, which then led to arrears to the government, were the primary way in which unprofitable activity was supported in Russia.

The virtual economy evolved and persisted because it met the needs of so many actors in the economy. Workers and managers at industrial dinosaurs benefited because the virtual economy postponed the ultimate reckoning for loss-making firms. Government, especially at the sub-national level, where much of the important action took place, benefited because the virtual economy system maintained employment and the provision of social services. Gazprom also benefited, since the value transfers it made to the virtual economy were the price it paid to appropriate the massive rents from exports.

The roots of the virtual economy mechanisms lay in the Soviet system, especially the production relationships that had developed under the Soviet command economy. These relationships represented a peculiar type of asset, ‘relational capital’, which supplemented the enterprise’s conventional physical and human capital. Thanks to relational capital, market reform policies did not necessarily compel the enterprise to restructure in order to be able to compete in the market environment. Enterprises chose between whether to become more competitive in the market, by investing in physical and human capital, or to be better protected from the market, by investing in relational capital.

The Term

The term ‘virtual economy’ was coined in 1998 by Gaddy and Ickes, building on terminology in a Russian government report from 1997. In early 1996, alarmed by the extent of tax delinquency in the country, President Boris Yeltsin appointed a special blue-ribbon panel to investigate the low rate of collection of taxes in Russia. Presenting its findings after an 18-month investigation, the panel reported that the country’s largest companies conducted 73 per cent of all their business in the form of barter and other nonmonetary forms of settlement. Especially alarming was the extent of nonmonetary payments of taxes. During the period of review, these large enterprises paid less than eight per cent of their tax bills in actual cash. They simply failed to meet 29 per cent of their obligations at all, while ‘paying’ the remaining 63 per cent in the form of offsets and barter goods. The market value of the goods delivered was far below the nominal price used in the offsets, leaving the government with substantially less in real revenues than officially accounted for. In summing up their own conclusions about the contemporary Russian economy, the investigatory commission wrote:

> An economy is emerging where prices are charged which no one pays in cash; where no one pays anything on time; where huge mutual debts are created that also can’t be paid off in reasonable periods of time; where wages are declared and not paid; and so on. . . . [This creates] illusory, or virtual earnings, which in turn lead to unpaid, or virtual fiscal obligations, [with business conducted at] non-market, or virtual prices. (Karpov 1997)

Gaddy and Ickes (1998) suggested that the entire system be called a virtual economy ‘because it [was] based on illusion, or pretense, about almost every important parameter of the economy: prices, sales, wages, taxes, and budgets’. The pretence that had become the norm was as characteristic of the virtual economy as were the colourful forms of nonmonetary exchange.

The Nonmonetary Economy

The nonmonetary means of payment that characterized the virtual economy spanned a wide range. They included direct exchanges of goods (true barter), either bilaterally or through ‘chains’ with multiple participants, offsets (where debts accrued by one party were later paid off not in money but in goods), and promissory notes called *veksels. Veksels* – the name is derived from the German *Wechsel* (‘promissory note’) – were a widespread nonmonetary payment mechanism that ranged from being a substitute for money to essentially a form of barter.

There were several key nodes in the barter chains, above all the major natural monopolies known popularly as the ‘three fat boys’ *(tri tolstyaka)* – Gazprom (the natural gas monopoly), RAO UES (the electricity monopoly), and MPS (the state railways). All three frequently complained that they collected as little as ten per cent of their revenues in cash. Almost all enterprises in Russia were consumers of the output of these three companies, rail freight transport, gas and electricity. The three monopolies also accounted for about 25 per cent of taxes due to the federal budget. The fact that everyone needed to purchase services from the ‘fat boys’ meant that there was a ready demand for the *veksels* (IOUs) of these companies. It was this special position that put them at the core of the non-payments system in Russia.

The other key player in the barter economy was the government, or rather, governments at all levels. Here again was an agent to whom nearly everyone had an obligation. The volume of accrued unpaid taxes, plus the huge fines and penalties levied for nonpayment, presented governments with an almost inexhaustible supply of debts. And, in turn, governments themselves owed many others. They were, like the natural monopolies, a key node for barter.

One particularly important phenomenon was tax offsets. An enterprise owed taxes to the government, and concluded an agreement whereby those tax obligations were settled by delivering goods or performing services for the government. Of all the forms of nonmonetary transactions observed in Russia in the 1990s, the mechanism of tax offsets was the most characteristic of the virtual economy. Russian governments at all levels grew increasingly willing to offset enterprises' tax obligations against goods or services delivered to the government. By the end of 1997, the accumulated tax debt was enormous. Industrial enterprises were particularly egregious delinquents. The sum owed by the enterprises at the end of 1997 was equal to 46 per cent of the amount they actually remitted in taxes for the whole of 1997. These enormous debts gave impetus to the practice of tax offsets.

Consider, for example, an enterprise that was able to supply the local government with services in lieu of taxes. The enterprise could have paid its tax liability in money, but that would have required selling its output for cash. Alternatively, the enterprise could negotiate with the government to supply some service as an offset for taxes. If the enterprise had resources that were not fully utilized, the latter alternative was likely to reduce the effective tax burden on the enterprise. Moreover, once the government showed itself to be willing to engage in tax offsets, the options open to enterprises expanded. The enterprise could now potentially pay its taxes not only with its own products but also with products it received in barter deals from other enterprises. This greatly reduced the cost to the seller of accepting goods rather than cash.

The motivation for governments to join in the barter economy was simple. They reasoned that if they could not get cash, it was better to reach some sort of settlement than receive nothing at all. In some cases, especially at the local level, an enterprise could offer to deliver goods or services to the city or regional government in lieu of taxes. At the federal level, it was more common for the government to cancel tax arrears or taxes due by writing off the government's own debt to the enterprise in question for state orders. Once the practice was established with respect to past arrears, there was an anticipatory factor: enterprises began to feel confident that they could henceforth ship off products to the government, knowing that later they would be allowed to offset their taxes in an equivalent amount. Less than 60 per cent of all federal taxes collected in 1997 were paid in cash; the rest were in the form of offsets.

The federal government was particularly victimized by these schemes. Enterprises frequently colluded with regional and local officials to hide income and hence keep revenues away from the federal government for taxes whose revenues were split between local and national authorities. In other cases, local governments demanded that enterprises pay their taxes in the form of goods and services that could be used only locally and not be shared with the federal government (for instance, by providing road construction or repairs of buildings). Often, if the federal government received anything at all in these schemes, it was only what the regional governments did not want.

In one notorious case reported in the Russian press in the spring of 1998, the oblast (province) government of Samara had permitted enterprises to pay their regional taxes in the form of goods. One of the items offered turned out to be ten tons of toxic chemicals from a local chemical plant. Although the plant claimed (and was given) credit for 400 million rubles (80,000 dollars) in taxes, auditors later determined that the chemicals were worthless (and indeed dangerous). The Samara government never suffered from this curious deal, however, since it had previously sought and received permission from the federal ministry of labour to fulfil its obligations to the federal unemployment compensation fund by delivering goods instead of money. Among the goods it offered were the ten tons of toxic chemicals (Gaddy and Ickes 2002, p. 176).

As a result of these practices, the Russian budget ran massive deficits. Even using the inflated prices used in the offset deals, federal revenues plummeted – from 16.2 per cent of GDP in 1995 to 12.4 per cent in 1998. To finance its deficits, the government had resorted to extensive borrowing outside and inside Russia at increasing and unsustainably high costs, thus digging itself even deeper in debt. Finally, on 17 August 1998, the government defaulted on about 40 billion dollars' worth of its own ruble-denominated debt instruments (so-called GKOs), some 17 billion dollars of which were held by foreigners.

The Soviet Roots

The roots of the virtual economy lay in the structure and institutions bequeathed to Russia by its Soviet predecessor. Some parts of the economy, notably the resource industries, were value-adding. But most of the vast manufacturing sector that Russia had inherited could not compete in a market setting. In fact, by the final years of the Soviet era, the manufacturing sector was in poor condition even on the terms of the planned economy. By official Soviet standards, more than one-third of equipment in Russian industry was physically obsolete. Soviet planning practice, which emphasized output over costs, set physical, rather than economic, obsolescence as the criterion for removing a machine from the factory. As long as the machine could produce anything at all, it was kept in production. The result was very low replacement rates for capital equipment.

The location of industry in the Soviet economy was another problem. Not only did Soviet location policy ignore transport costs but it also failed to take into account the costs associated with the cold Russian climate – in terms of energy use, health maintenance and many other factors. By being placed in some of Russia's coldest and most remote regions, the manufacturing enterprises were rendered even less competitive and less attractive for foreign investment.

Equally important as the structure of the Soviet economy and its lack of competitiveness was the fact that this reality was hidden. As the market transition began, past history and performance gave no information about which sectors, or enterprises, were value-adding and which were value-destroying. The culprit was distorted Soviet pricing.

Soviet Pricing and the 'Circus Mirror' Effect

Soviet prices were not based on opportunity cost, or value; rather, they were simply an accounting instrument to measure plan fulfilment. Although Soviet prices were set arbitrarily, they were not set randomly. They were determined by specific rules of the system, which produced some systematic biases. First, the planners underpriced raw material inputs, especially energy. They based raw materials prices only on the operating costs of extraction, while ignoring rent. In so doing, they disregarded the opportunity cost of using the resources now rather than in the future. The planners' overriding goal was to increase today's output. Scarcity pricing might have induced more conservation, but it would have militated against maximizing current production. This bias in raw materials prices fed into the system of industrial prices. Heavy consumers of energy were, in effect, subsidized. So, too, were heavy users of capital, thanks to the absence of interest charges. In short, costs of production were calculated on the basis of an incomplete enumeration of costs. This led to lower prices for inputs, especially resource inputs, than for final uses and thus an understatement of the share of gross output used in production and, hence, an overstatement of net output.

In addition to incomplete cost-based pricing, the Soviet system was explicitly biased towards certain users. The Soviet leadership assigned priority in the economy to heavy industry, especially the defence industry, and it was important that it appear that these sectors were producing value. This non-scarcity-based pricing was like a distorting mirror at the carnival. It created the illusion that many enterprises were value-producing when in fact they were value-destroying.

The 'Loot Chain'

A further factor contributing to the opaqueness of the Soviet economy and its post-Soviet successor was the way in which income from control of assets was passed down as payoffs through what Gregory Grossman (1998) referred to as the loot chain. In the USSR, wealth diverted from the official state economy into private hands was shared among networks of individuals in the form of payoffs, bribes, and other schemes. Over time an ever greater proportion of people's incomes depended on the chain of corruption and side payments.

The virtual economy perpetuated the loot chain in post-Soviet Russia. The living standards of a huge number of people depended on the chain of production and distribution of goods and services in the virtual economy system, where value redistribution, in contrast to looting pure and simple, occurred in a form that paralleled and was intertwined with actual productive economic activity. This made it especially difficult for agents to discern what their own value and the value of their assets would have been in a well-developed and transparent economy. Basic ideas of a market economy, such as the relationship between individual effort and reward, became almost impossibly obscure. One's static position in the production process – for instance, membership in the workforce of a particular enterprise – was more important for success than individual skills and abilities. The Soviet system separated 'what you get' from 'what you do'. The reality was that the effort–reward nexus was random. Instead of 'from each according to his ability, to each according to his needs (or ability)', it was 'to each according to some unknowable, random criterion'. The durability of the misperception depended on its opaqueness. There was no alternative, competing information about the real relationship. This meant that the loot chain was also a constraint on the future evolution of the economy. Individuals were dependent on the prevailing system and they could not know what an alternative system would offer. The uncertainty caused them to resist abandoning the prevailing system.

Impermissibility of True Reform

While there was no accurate information about the economic importance of the large Soviet manufacturing sector, its social and political importance was unavoidable. Many of the least competitive enterprises – the so-called dinosaurs of Soviet industry – were socially the most important. They employed millions of workers and provided for tens of millions of their family members. Entire cities depended on them. The sheer size of this sector – as shown by employment – operated to maintain its social and political importance, and the illusion of its economic performance. In a sense, then, the importance of the manufacturing sector in Russia was an illusion economically but continued to be a political and social reality.

This latter reality constrained serious market reform policies. Russia did not formally reject the policies themselves; instead, it continued with a pretence of market reform. Policymakers launched one measure after another in their attempt to transform Russia into a market economy. But very few of those measures were allowed to play themselves out to their full extent. The consequences of complete and proper implementation would have been politically intolerable. Thus, while the nation's leadership proclaimed reform policies, enterprises and other agents continued to behave in ways that rendered the policies ineffective.

The Behavioural Implications

The range of behavioural options in the virtual economy was broad. The ability to use non-monetary mechanisms to pay taxes to governments and bills to the natural monopolies fundamentally changed the range of opportunities for action available to Russian enterprise directors. By allowing enterprises to settle their obligations by delivering goods for which there was no effective demand, the governments and the monopolies offered an incentive to avoid restructuring. For many enterprises it was easier to produce such goods than to restructure and earn additional monetary income to pay bills in cash. Producing those goods allowed for the use of idle capital and labour. In short, offsets and barter permitted some enterprises to survive without restructuring. To represent the full range of choice, not only market-oriented activity but also behaviour characteristic of the virtual economy, Gaddy and Ickes (2002) employed the notion of a two-dimensional space, called *r*-*d* space. The following sections outline their model.

Market Distance, *d*

The impact of liberalization on the Soviet economy can be expressed with a spatial metaphor: liberalization revealed the distance that a Russian enterprise would have to travel to compete in the world economy. Let *d* designate the enterprise's 'distance to the market' at the start of transition. Clearly, *d* depends on the enterprise's initial endowments of the things that matter for market viability – physical and human capital, as well as the enterprise's marketing structure and organizational behaviour, but also the characteristics of the good that the enterprise produces (its quality and cost of production). Formally, define an enterprise's *d* as the amount of capital expenditure needed to enable the enterprise to produce a product that is competitive in the market. The fundamental reason for measuring *d* in terms of the investment cost is that transition causes a divergence between the value of existing (inherited) capital and that of newly installed capital.

One may begin to grasp this point by recalling what happened to traditional models of investment in market economies during the energy crisis of the 1970s. Those models predicted that investment would decline, given the tremendous increase in the price of energy. In fact, however, spending on new equipment and buildings soared. The reason for this discrepancy between model and reality was the divergence between the value of installed capital that was energy intensive and new capital that was energy saving. The conventional model ignored the sharp decline in the economic value of the existing capital stock as a result of the 1973 energy crisis. Installed capital had been the result of investment decisions based on low energy prices; hence, its value fell dramatically once energy prices quadrupled. This in turn only increased the demand for new investment in energy-saving equipment. The result was a divergence between the value of installed capital (which lost value) and that of new capital (which had full economic value). In the Russian context, measuring market distance *d* by the need for new capital investment is a way of capturing the cost of filling the gap between the value of inherited (Soviet) capital and new (market) capital.

Distribution of *d*

The level of *d* differs widely among enterprises in the economy. An enterprise that already produces a product it can sell in world markets at a price above cost will have a value of *d* equal to 0. A completely noncompetitive enterprise will have an enormously large d. Everyone else will be somewhere between. For example, an oil- producing enterprise will have a very low *d*. Its product is already right for the market. It may need only relatively small investments in marketing, and so on. A Soviet-style machine tool producer, in contrast, is likely to have a long distance to travel.

The distribution of *d*'s in transition economies differs in two respects from that in market economies. In transition economies the range of *d*'s is greater and the distribution is more skewed. Both differences stem from the dissimilarity in the process of entry and exit in market and planned economies. In a market economy, whether or not a new firm attempts to enter an industry depends on the founders' expectations about the new firm's competitiveness. They will enter if they expect the firm's potential costs to be lower (its productivity to be higher) than those of existing firms. No firm enters an industry in which it expects it will be noncompetitive. Over time the competitiveness of some firms declines, so *d* increases. But if a firm in a market economy has too high a level of *d*, it will be forced to close. Competition and hard budget constraints cause high-*d* enterprises to shut down.

In a transition economy, by contrast, some enterprises have very high *d*'s that would not be observed in a market economy. There are several reasons for these high-d enterprises. First, in socialist economies entry was not determined by expectations of profitability or competitiveness but rather by the need to fulfil plan targets. Second, insulation from the world economy meant that enterprises were created that produced goods for which the country might not have had a comparative advantage. Third, especially in the case of Russia, the priority given to defence production led to a proliferation of enterprises that produced goods whose market collapsed with the end of the Soviet Union. Fourth, since the geographic location of industry in the Soviet period was based on ignoring transport costs (as well as the costs associated with extraordinarily cold temperatures), the location of enterprises was also a factor in increasing the *d* in many cases. For all these reasons, the distribution of the *d*'s in Russia at the onset of the transition had a much higher mean and was more skewed to the right than in a mature market economy. This extra mass of high-d enterprises was the burden of the Soviet legacy. And it was this burden that was the essence of the restructuring problem: so many enterprises had to radically reduce their distance to the market at the same time.

One way to think of the purpose of economic reform is to reduce the average distance in the economy. This occurs through three means: (1) exit of high-*d* enterprises; (2) entry of new low-d enterprises; (3) and reduction of the *d* of surviving enterprises. In an ideal market world, market distance would be the only condition that characterized the state of an enterprise. If the only important difference in enterprises were their initial level of *d*, then policies that put pressure on existing high-*d* enterprises and encourage creation of new low-*d* enterprises would have the effect of pushing the distribution in the direction of the market.

Relational Capital

The conventional view of restructuring, whereby reform means reducing d, assumes that each enterprise has one set of resources – its physical and human capital – which it must use ever more efficiently in order to survive. The virtual economy view, by contrast, posits that some enterprises have another resource, relational capital, which they can draw on to enhance their chances for survival. Relational capital is the stock of goodwill that an enterprise can use to avoid the strictures of the budget constraint. An enterprise that has high relational capital can undertake transactions (bartering, using tax offsets, delaying payment) that other enterprises, with low amounts of relational capital, cannot get away with. To put it another way, relational capital is goodwill that can be translated into the ability to continue to engage in production and exchange without reducing the distance to the market. It is therefore the existence of this second dimension that can explain the persistent survival of high-*d* enterprises in the Russia of the 1990s.

At the onset of transition enterprises differed in their inherited relational capital call it *r* – just as they differed in their d. Some enterprises (or their directors) had very good relations with local and/or federal officials. Relations with other enterprises also varied.

Origins of Relational Capital

The relational capital of Russian enterprises was initially accumulated in the Soviet system. Enterprise directors relied heavily on the accumulation and use of personal connections. Relational capital was passed forward to the post-Soviet system in a deceptively simple manner: it was spontaneously privatized. And here lies an important aspect of economic transition in Russia. As Hewett (1988) described, plan fulfilment in the Soviet economy required enterprise directors to use informal skills. Their ability to accomplish this, and their position in the economic hierarchy, was critical to their incomes. While directors earned income from these positions, they did not legally own the source of these incomes. The demise of the planning system, which had already begun with Mikhail Gorbachev's reforms in the late *perestroika* period, had the effect of increasing the autonomy of enterprise directors. With the start of economic reform and privatization, the role of the enterprise director increased; other mediating actors (planners, party officials) played less and less of a formal role in economic allocation. Directors used this opportunity to appropriate the returns to the relationships they had developed and cultivated under the previous system. However, in order for directors to appropriate these returns, the enterprises had to continue to operate. Much of the relational capital was both enterprise-specific and person-specific. To the extent that it was enterprise-specific, the director could not cash out the relational capital. The primary form of these connections was relationships with directors of other enterprises, often in related lines of activity, and with ministerial officials and local government officials. The relational capital was worthless to the incumbent director unless he remained in that particular enterprise. He could not leave the enterprise and take the relational capital with him. Furthermore, because it was person-specific, he could not sell it to someone else. Instead, in order to appropriate the rents accruing to his relational capital, he had to remain in the enterprise and keep it operating. The privatization of relational capital is thus an important part of the explanation of why directors fought to keep open enterprises that had few prospects in the market economy.

r–d Space

The concept of relational capital can be used to revise the spatial representation of the Russian transition economy. There are now two state variables that describe the nature of an enterprise. In addition to the dimension of market distance, enterprises can be arrayed in terms of their level of relational capital. The initial conditions of an enterprise can thus be described by a two-dimensional space, *r–d* space, in which each enterprise has its own location.

Whether one views the enterprise sector in a single (*d*) dimension or in the two dimensions of *r–d* space is critical for how reform policy is understood. The conventional, one-dimensional view assumes that economic reform measures will have the greatest impact on those enterprises that have the highest level of *d*. According to this assumption, for example, if budget constraints are tightened, enterprises that are farthest from the market will be under greatest competitive pressure. Similarly, it is assumed that if the economy is opened to international competition, the greatest impact will be on those enterprises that are most in need of restructuring. In the two-dimensional *r–d* space environment, the effects of market- type reforms need not have this property at all. Tightening the budget constraint will not necessarily put the most pressure on the enterprise that is most inefficient (with the highest *d*). If the enterprise has been endowed with high *r*, it may be insulated against the impact of this policy; it can use relations to evade the budget constraint. And if tight budget constraints are enforced against enterprises that are lower in *r*, then the policy may, in fact, have greater impact on low-*d* enterprises than high-*d* enterprises.

It is not just the initial levels of either *r* or *d* that matter, of course. An enterprise's location in *r–d* space is not the immutable relic of its past; it depends on the path of enterprise investment decisions. If the enterprise has invested in *r*, it will improve its resistance to policies of tight budget constraints. The enterprise director's problem is to decide how much to invest in reducing distance and how much to invest in relational capital.

See Also

- ► Barter
- ► Barter in Transition
- ► Soviet Union, Economics in

Bibliography

Gaddy, C., and B. Ickes. 1998. Russia's virtual economy. *Foreign Affairs* 77: 53–67.

Gaddy, C., and B. Ickes. 2002. *Russia's virtual economy.* Washington, DC: Brookings Institution Press.

Grossman, G. 1998. Subverted sovereignty: Historic role of the Soviet underground. In *The tunnel at the end of the light: Privatization, business networks, and economic transformation in Russia*, ed. S.S. Cohen, A. Schwartz, and J. Zysman. Berkeley: University of California Press.

Hewett, E. 1988. *Reforming the Soviet economy.* Washington, DC: Brookings Institution Press.

Karpov, P. 1997. On the causes of the low rate of tax collection (nonpayments in the fiscal system), general causes of the 'payments crisis', and the possibility of restoring the solvency of Russian enterprises. Report of the Inter-Agency Balance-Sheet Commission, chaired by P.A. Karpov. Moscow (December). [In Russian].

Viti de Marco, Antonio de (1858–1943)

F. Caffe

Italian economist and politician; born in Lecce on 30 September 1858; died in Rome on 1 December

1943. He graduated in law at the University of Rome in 1881 and embarked on an academic career, first teaching political economy and then public finance at Camerino, Macerata and Pavia. In 1887–8 he took up the post of teaching public finance in the Faculty of Law in Rome, where he remained until 1931. From 1901 until 1921, with only a brief intermission, he was a member of the Italian Parliament. He attempted unsuccessfully to found a liberal democratic group whose main aim was to fight the protectionism and exploitation of Southern Italy. The volume entitled *Un trentennio di lotte politiche (1894–1922)* is a testimony to his political ideas. In keeping with his political beliefs, he avoided taking the oath of allegiance to the fascist regime by giving up his university post in 1931. De Viti de Marco's cultural interests led him, together with some other economists, to complete the purchase in 1890 of the *Giornale degli Economisti*, of which he was co-editor until 1919 with Maffeo Pantaleoni, Ugo Mazzola and, later on, with Vilfredo Pareto. It was in this way that the *Giornale degli Economisti* became the most authoritative voice of liberal Italian thinking.

De Viti de Marco was not a prolific writer – he spent much time patiently revising his own works – but he exerted a fundamental influence on the typically Italian tradition of creating a 'pure' theory of public finance. He dedicated his first essay (*Il carattere teorico dell'economia finanziaria*) in 1888 to this particular area of economic research. At the same time he studied monetary and credit problems, on which in 1898 he published the volume entitled *La funzione della Banca*, which he revised several times before the definitive edition was published in 1934. De Viti de Marco's name, however, is primarily connected with his *Principi di Economia Finanziaria*, which was the subject of various drafts and revisions in 1923, 1928, 1934 and 1939. The definitive edition of this work contains a masterly preface by Luigi Einaudi which fully upholds 'for spontaneous universal recognition' the position of supremacy held by De Viti de Marco over other researchers in the field of public finance. In addition, when the book was translated into English, it was generally judged to be 'the best book ever written on public finance'. De Viti de Marco's *Principi* has been translated into all major foreign languages, and it embodies the most complete attempt to construct an 'economic' theory of the entire financial system, whose final aim is the systematic application of the theory of marginal utility to financial problems.

The origins of De Viti de Marco's beliefs can be traced to the work of Francesco Ferrara, in as much as the latter believed public spending to be an integral part of the study of public finance, and recognized the productive aspect of the public services. The significance of the study of financial problems had already been foreseen in the writings of Maffeo Pantaleoni and Ugo Mazzola. But it was De Viti de Marco who, after forty years of methodical work, advanced the economic concept of public finance based on two abstract types of political constitution of the State: a monopolist state in which a privileged oligarchy acts in its own interests in the decisions concerning the levying of taxes and the distribution of public expenditures; and a cooperative state where the interests of tax-payers and those who are entitled to benefit from the services of the state coincide. This latter type of state was referred to most extensively by De Viti de Marco in his work in order to examine the whole fiscal problem, because in the cooperative state choices and the decisions are reduced to economic calculus on an individualistic level and the resulting finance is devoid of any coercive character. The precise reasoning of this premise and its rigorous development explain why De Viti's work was internationally acclaimed. It also explains the criticisms of those who followed a sociological approach and did not consider economic calculus at an individual level to be a valid basis for collective decisions. But De Viti's undisputed merit lies in his having created a scientific model which has remained a point of reference and a focus of discussion for alternative ideas about the nature, the causes and the effects of fiscal phenomena.

Selected Works

1885. *Moneta e prezzi*. Città de Castello: S. Lapi.

1888. *Il carattere teorico dell'economia finanziaria*. Rome: Pasqualucci.

1898a. *Saggi di economia e di finanza*. Edited from the Giornale degli Economisti. Rome.

1898b. *La funzione della Banca*. Rome: Accademia dei Lincei.

1930. *Un trentennio di lotte politiche*. Rome: Collezione Meridionale Editrice.

1932. *Grundlehren der Finanzwirtschaft*. Tübingen: J.C.B. Mohr.

1934a. *La funzione della Banca*. Revised and definitive ed. Turin: Einaudi.

1934b. *Principi di economia finanziaria*. Turin: Einaudi.

1934c. *Principios fondamentales de economia financier*. Madrid: Editorial Revista de Derecho Privado.

1936. *First principles of public finance*. Trans. E.P. Marget. New York: Harcourt, Brace & Co.

References

Buchanan, J.M. 1960. 'La scienza delle finanze': the Italian tradition in fiscal theory and political economy. In *Selected essays*, ed. J. Buchanan. Chapel Hill: University of North Carolina Press.

Cardini, A. 1985. *Antonio De Viti de Marco*. Bari: Laterza.

Ricci, U. 1946. Antonio De Viti de Marco. *Studi Economici*.

Volterra, Vito (1860–1940)

Giancarlo Gandolfo

A mathematician by vocation, Volterra graduated at the Scuola Normale in Pisa in 1882 and obtained the Chair of Rational Mechanics at the University of Pisa in 1883. Subsequently he held chairs at the Universities of Turin and Rome. He became a Senator in 1905, was President of the Consiglio Nazionale delle Ricerche, of the Academia dei Lincei, Fellow of the Royal Society, etc. In 1931 he refused to take the required oath of loyalty to the Fascist government and was deprived of his Rome chair and forced to resign from all Italian scientific academies.

Volterra is renowned for his contributions to pure and applied mathematics. He is recognized as the founder of the general theory of functionals (1887, 1927a, 1929). In biological mathematics (independently of Lotka, who had examined the two-species case earlier) he introduced the prey–predator equations generalized to n species (1926, 1927b, 1931).

In 1906, Volterra reviewed Pareto's *Manuale*. Pareto, in treating the problem of integrating the differential equation of the indifference curve to obtain the 'ophelimity' (the utility function) had stressed the case in which the 'elementary ophelimity' (the marginal utility) of each good was a function solely of the quantity of that good, giving the impression that this was the case in which the integration could be performed with certainty. Volterra reminded Pareto that in the two-variable case there always exists an integrating factor so that it is always possible to perform the integration; he also pointed out that – as there exists an infinite number of integrating factors – the utility function is, in general, indeterminate. The real integrability problem arises when one has to deal with more than two commodities, and Volterra invited Pareto to go more fully into this problem. This was the beginning of the integrability problem in the theory of consumer demand.

Although (1906) was Volterra's only contribution to economic theory, his work is of interest to economists for at least other two reasons. One is his functional analysis, now so important in problems involving infinite horizons, numbers of goods, etc. This, however, is like any other important mathematical tool whose availability enabled and continues to enable mathematical economists to solve their problems (for example, fixed point theorems or Liapunov's second method). The other and more important reason is his study of predator–prey equations, which directly inspired an economic model, Goodwin's growth cycle (1965): 'Finally, at some happy moment, I remembered Vito Volterra's formulation of the struggle for existence, and suddenly all became clear to me' (Goodwin's foreword to Vercelli (ed.), 1982, p. 72). This is a two-class model which can be reduced to a system of two differential equations of the Lotka–Volterra type (the variables are the workers' share of the product and the employment ratio). The result is a growth cycle; i.e. the economy grows, but with cycles in growth rates. Goodwin's was the first successful attempt at integrating (not merely superimposing) growth and cycles, and his seminal paper has given rise to many important developments which use predator–prey equations as the basic tool (see, e.g., Izzo 1971; Desai 1973; Vercelli (ed.), 1982; Goodwin et al. (eds) 1984).

See Also

- ▶ Functional Analysis
- ▶ Integrability of Demand
- ▶ Predator–Prey Models

Selected Works

A full bibliography is included in Whittaker's biography of Vito Volterra (originally published in 'Obituary Notices of the Royal Society' 1941) as reproduced and completed in the 1959 reprint of (1929). This biography also contains a detailed evaluation of Volterra's scientific work. Volterra's scientific papers have been collected in five volumes by the Accademia Nazionale dei Lincei as V. Volterra, *Opere matematiche: memorie e note*. Rome: Cremonese for the Accademia nazionale dei Lincei, 1954–62; the fifth volume includes the complete bibliography of Volterra's works.

1887. Sopra le funzioni che dipendono da altre funzioni. *Rendiconti della R. Accademia dei Lincei*, series IV, 3(2). Reprinted in *Opere*, vol. I.

1906. L'economia matematica ed il nuovo manuale del Prof. Pareto. *Giornale degli economisti* 32: 296–301. Reprinted in *Opere*, vol. III.

1926. Variazioni e fluttuazioni del numero di individui in specie animali conviventi. *Memorie della R. Academia dei Lincei*, series VI, 2: 31–113.

1927a. *Teoria de los funcionales y de las ecuaciones integrales e integro-diferenciales*. Conferencias explicadas en la Facultad de Ciencias de la Universidad, 1925, redactadas por L. Fantappié. Madrid: Imprenta Clasica Española.

1927b. Variazioni e fluttuazioni in specie animali conviventi. *Rendiconti del R. Comitato talassografico italiano*, Memoria CXXX-I. Reprinted in *Opere*, vol. V.

1929. *Theory of functionals and of integral and integro-differential equations*. London: Blackie & Son (revised English translation of 1927a). Reprinted (with a Preface by G.C. Evans and the Biography by E. Whittaker). New York: Dover, 1959.

1931. Leçons sur la théorie mathématique de la lutte pour la vie. Paris: Gauthier-Villars.

Bibliography

Desai, M. 1973. Growth cycles and inflation in a model of the class struggle. *Journal of Economic Theory* 6(6): 527–545.

Gandolfo, G. 1971. *Mathematical methods and models in economic dynamics*, 409–416. Amsterdam: North-Holland and 436–442.

Goodwin, R.M. 1965. A growth cycle. Paper presented at the First World Congress of the Econometric Society. Rome. Published in *Socialism, capitalism and economic growth: Essays presented to Maurice Dobb*, ed. C.-H. Feinstein. Cambridge: Cambridge University Press, 1967. Reprinted in *Essays in economic dynamics*, ed. R.-M. Goodwin. London: Macmillan, 1982.

Goodwin, R.M., M. Krüger, and A. Vercelli (eds.). 1984. *Nonlinear models of fluctuating growth*. Berlin: Springer-Verlag.

Izzo, L. 1971. La moneta in un modello di sviluppo ciclico. In *Saggi di analisi e teoria monetariaII*, ed. L. Izzo. Milano: F. Angeli.

Vercelli, A. (ed.) 1982. Non-linear theory of fluctuating growth. *Economic Notes*: 69–190.

Voluntary Contribution Model of Public Goods

Richard Cornes

Abstract
This article surveys the literature on the model of voluntary contributions to public goods that

has developed since the early 1980s. This literature draws explicitly on noncooperative game theory. We present a recent novel statement of the problem, based on ‘replacement functions’, which is both more powerful and more transparent than the traditional formulation that uses players’ best response functions. We survey existence, uniqueness and comparative static properties of the basic model, and also sketch some of the extensions of that model – impure public goods, weakest link and best shot – that have been proposed and applied to such problems as global public goods and the global commons. We also draw attention to recent attempts to dynamize the model.

Keywords

Pure public goods; Impure public goods; Noncooperative games; Neutrality; Free riding; Underprovision; Weakest and weaker links; Best and better shots; Replacement functions; Lotteries and public good provision

JEL Classification

H41

Two classic papers by Samuelson (1954, 1955) played a major role in provoking interest in the problem of public good provision. However, they did not provide an explicit model of decentralized provision. His formal analysis focused on necessary conditions for their optimal provision. Elements of a positive model of decentralized provision – hereafter the standard model – were gradually developed during the following decades, and more complete formal analyses were provided by Cornes and Sandler (1985) and by Bergstrom et al. (1986).

Introduction

Consider a community with an exogenous number, n, of members. They have preferences over a private good and a public good. Player i’s consumption of the private good is y_i, and the total provision of the public good is G. Preferences, resource constraints, and the technology that converts individual contributions into the total available public good are summarized, respectively, by the following assumptions:

Preferences Player i’s preferences are represented by a utility function, $u_i(y_i, G)$, which is strictly increasing in both arguments and quasi-concave. Both goods are normal.

Resource constraint $y_i + c_i g_i \leq m_i$, where player i’s unit cost as a contributor c_i, and money income m_i, are exogenously given. g_i is player i’s contribution to the public good.

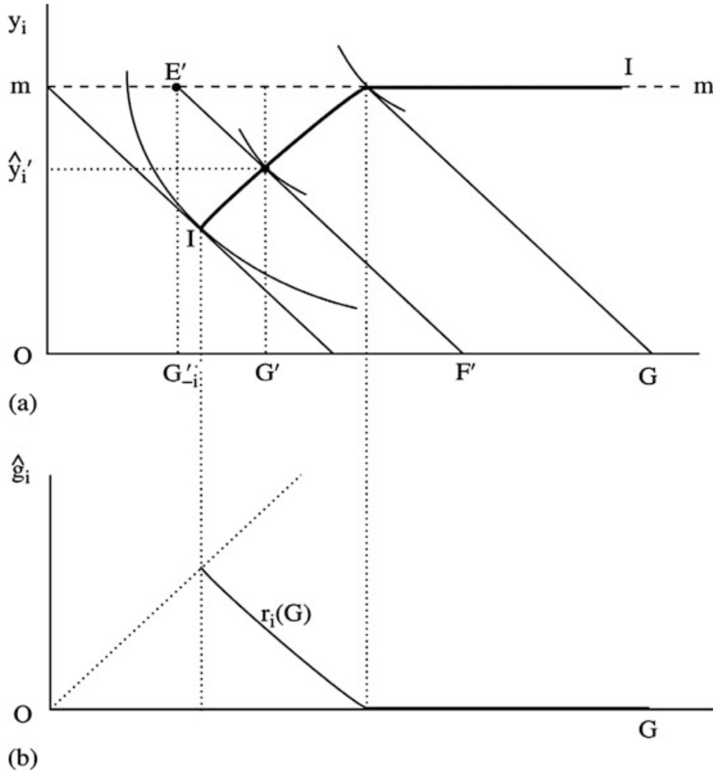
Technology of public good provision $G = \Sigma_{j=1}^{n} g_j$.

The model considers the Nash equilibrium of the static noncooperative game containing these elements when each player is choosing her best response, $\hat{g}_i$, to the choices made by all others, $G_{-i} = \Sigma_{j=1, j \neq i}^{n} g_j$.

This formulation slightly generalizes the standard model in that we allow unit costs to differ across contributors. This extension, initially explored by Ihori (1996), has interesting implications.

A Graphical Treatment

Analyses typically derive a best response function for each player. This determines the player’s most preferred choice of contribution as a function of the choices made by all other players: $\widehat{g}_i = b_i(G_{-i})$, where $\hat{g}_i$ is player $\bar{\imath}$ ‘– i’s utility-maximizing response. A Nash noncooperative equilibrium is an allocation at which every player chooses her best response. Formally, it is a solution to the n equations provided by the individual best response functions in the n unknowns, g_1, g_2,. . ., g_n. Questions about existence, uniqueness and other properties of equilibrium become questions about the existence, uniqueness and other properties of solutions to this set of equations. Such an approach, though naturally suggested by noncooperative game theory, is not the most helpful or transparent method of tackling these issues. We shall briefly sketch an alternative approach, suggested by Cornes and Hartley (2007a), which



Voluntary Contribution Model of Public Goods, Fig. 1 A player's replacement function

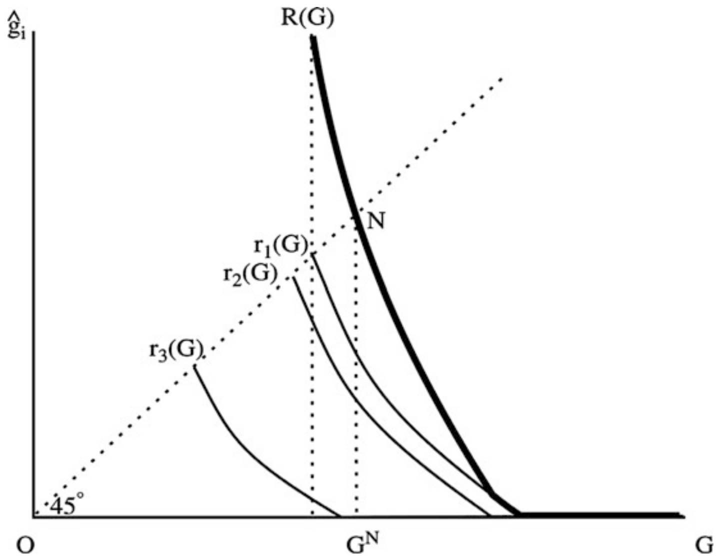
provides both a rigorous and powerful tool of analysis, and a simple and transparent geometric representation.

Individual Behaviour

Figure 1a shows player i's preferences, constraints and choices. Suppose her income is m_i. If the sum of all other players' contributions is G'_{-i}, player i can devote all her money income to private goods consumption and enjoy the public good provided by others. This allocation is the point E'. Each unit of private good consumption given up by i augments total public good provision by the amount $1/c_i$. Thus her budget constraint is the line $E'F'$. Her most preferred choice is the point of tangency, P'. By varying G_{-i} parametrically, we can trace out the income expansion path II, which summarizes the player's behaviour. The locus II is everywhere continuous, and slopes upwards at all points at which player i is at an interior point, choosing strictly positive values of both y_i and g_i. If there is a finite value of G_{-i} at which $\widehat{y}_i = m_i$, at that point the locus become horizontal.

Note that, for any given value of G, the vertical distance between the income expansion path and the locus of the income constraint mm measures the implied value of expenditure by i on the public good, $c_i\widehat{g}_i = m_i - \widehat{y}_i$. If $c_i = 1$, this same distance measures the quantity of the public good. If $c_i \neq 1$, then a simple scaling up or down of the vertical axis in panel (b) allows us to depict the quantity g_i. In any event, under our assumptions, to any given level of total public good provision G above a certain value there corresponds a unique level of contribution by player i, $\hat{g}_i$, that is consistent with that observed level, in the sense that $\hat{g}_i$ is a best response to the quantity $G_{-i} = G - \widehat{g}_i$. We write the implied functional relationship as $\widehat{g}_i = r_i(G)$ and call this player i's replacement function. The figure suggests that every player has a replacement function that is continuous, everywhere non increasing, and strictly decreasing in G wherever the replacement value itself is positive.

One further property of an individual's replacement function is significant. Suppose that, at a



Voluntary Contribution Model of Public Goods, Fig. 2 Nash equilibrium

given level of G, player i is a strictly positive contributor. Consider the consequence of an increase of Δm_i in i's money income. At that given level of G, Figure 1 panel (a) shows that her chosen allocation is unchanged. She consumes an unchanged quantity of the private good. Thus, her contribution to the public good changes by the amount

$$\Delta\widehat{g}_i = \frac{\Delta m_i}{c_i}$$

Geometrically, the graph of (the positive section of) her replacement function rises vertically by an amount that, appropriately deflated by the cost parameter, equals the income change. This property plays a crucial role in comparative static analysis.

Nash Equilibrium

Figure 2 shows the graphs of players' replacement functions in a three-player game. Equilibrium is an allocation at which the aggregate quantity of the public good is consistent with the replacement values to which it gives rise. In an n-player voluntary contribution game, it is an allocation at which

$$R(G)\equiv\sum_{j=1}^{n} r_j(G) = G$$

The 'aggregate replacement function' $R(G)$ is shown as the thick line in Fig. 2. It is simply the vertical sum of the individual graphs. A Nash equilibrium may be depicted graphically as a point where the graph of $R(G)$ intersects the 45° ray through the origin in Fig. 2. This relationship describes a Nash equilibrium in the form of a single equation in a single unknown, G, regardless of how many players there are, and how they differ with respect to preferences, unit costs and money incomes. Armed with the properties already sketched above of the individual replacement functions, scrutiny of this equation is sufficient to provide a complete positive analysis of the model. First, however, note the following simple points. First, the sum of two continuous functions is continuous. Second, the sum of two monotonic functions is itself monotonic.

Properties of the Equilibrium

We now have all the ingredients for a rigorous analysis of the equilibrium properties of the model, which we now investigate.

Existence of Equilibrium

Consider the player whose replacement graph reaches the 45° ray furthest from the origin in

(G, $R(G)$) space. It is possible that, at that level of G, all other players are choosing to contribute zero. In this case, we have found an equilibrium, at which the chosen player is the sole contributor. Alternatively, there may be other players whose replacement values are positive. In this case, we have found a value of G at which $R(G) \equiv \sum_{j=1}^{n} r_j(G) > G$. Then monotonicity implies that, as G rises, the left-hand side of this inequality falls, while the right-hand side rises. Continuity implies that there must be a finite value of G at which the equilibrium condition holds. Either way, an equilibrium certainly exists. In Fig. 2, this is the point G^N, at which the sum of all players' contribution levels that are individually consistent with G^N is also collectively consistent.

Uniqueness of Equilibrium

Monotonicity implies that $R(G)$ is everywhere nonincreasing. Clearly, G is a strictly increasing function of itself. Thus, there can only be one value of G at which $R(G) = G$.

Presumptive Inefficiency of Equilibrium

In the basic model, in which a common unit cost is assumed across players, there is a general presumption that too little of the public good is provided at equilibrium, in the sense that Pareto-superior allocations can be obtained by increasing the level of public good provision. This may be confirmed by a simple envelope argument. Suppose that, at an equilibrium, players j and k are both positive contributors. Starting from the equilibrium, a small increase in player j's contribution imposes a second-order cost on player j, but generates a first-order benefit for player k. Similarly, a small increase in player k's contribution imposes a second-order cost for player k and a first-order utility gain for player j. Thus it is possible for both to be made better off if both raise their contributions slightly above their equilibrium levels. Furthermore, such a move will not hurt other players, and will generally benefit them. Thus it is Pareto-improving.

In the current model, in which unit costs are allowed to differ across players, this remains true. There is also, however, a second source of inefficiency. This arises from the fact that the 'wrong' people contribute at equilibrium. Consider an equilibrium at which both a high-cost and a low-cost contributor are making positive contributions. An initial transfer of income from the high to the low-cost player shifts the replacement function of the high-cost player down, and that of the low-cost player up, in the neighbourhood of the equilibrium value of G. But the latter shift is quantitatively greater, so that the aggregate replacement function shifts upwards. The equilibrium provision therefore must rise, and contemplation of Fig. 1a makes it clear that all players are better off in the new equilibrium.

Note that we talk of presumptive, not necessary, underprovision. This is for two reasons. First, as Cornes and Sandler (1996, p. 160) point out, if every player prefers to consume the private and public good in fixed proportions, so that their indifference curves are L-shaped, then the equilibrium is Pareto efficient. This possibility disappears if we allow some substitutability between the private and public goods. A second possibility, which certainly needs to be taken more seriously in policy discussions of public good provision than is sometimes done, is that the equilibrium involves zero total provision and that, even when provision is zero, the sum of all player's marginal valuations is less than the minimum cost of producing an increment of the public good. In this case, the public good neither is, nor should be, provided.

Neutrality

Suppose that two players – say i and j – have the same value for the cost parameter. Consider an equilibrium, G^*, at which both are strictly positive contributors. Now transfer an amount of income, Δm, from one to the other. In the neighborhood of G^*, the recipient's replacement graph shifts upwards by the amount Δm. The donor's graph shifts downwards by the amount Δm. Thus, if both remain positive contributors at G^*, that value remains the sole equilibrium public good provision level. Nothing real has changed – equilibrium levels of private good consumption and of total public good provision, and therefore equilibrium utility levels, are unaffected by the income transfer. This is the famous neutrality property of the standard model, which assumes a common value of the cost parameter for all players. Often attributed to Warr (1983), it was foreshadowed in earlier work by Shibata (1971).

Non-neutrality

The reasoning that led to the neutrality result allows us to understand easily the circumstances under which neutrality fails to hold. First, suppose that the source of the income transfer is initially choosing to contribute zero. Then, at the initial level of G, the reduction in her income cannot shift the relevant portion of her replacement function downwards – she is already contributing zero. The recipient's replacement graph shifts upwards. Therefore the aggregate replacement graph shifts upwards, and the equilibrium provision of the public good must now be higher. Transfers between existing contributors and noncontributors will have real consequences, leading to changes in both the equilibrium total public good provision and also in individual equilibrium utility levels. It is even possible, as Cornes and Sandler (2000) point out, that transfers from each of several noncontributors to contributors leads to a Pareto-superior allocation.

Second, our discussion of the presumptive inefficiency of equilibrium has already shown that an income transfer from a high-unit-cost contributor to a low-unit-cost contributor will lead to a higher level of equilibrium provision and to a Pareto improvement.

Implications of a Cost Change

Suppose that player i is initially a positive contributor, and that she enjoys an exogenous reduction in her unit cost. Consideration of Fig. 1 shows that the level of her preferred contribution that is associated with the initial equilibrium value of G must now be higher. In the absence of any other shocks, the equilibrium level of total provision must rise. Thus, every player except for i enjoys a higher equilibrium utility. However, player i herself may be either better or worse off – on the one hand, total provision is higher, but on the other hand she is now contributing a higher share of that total, since her fellow contributors have reduced their contributions.

Limiting Behaviour as n Gets Large

The implications of adding players to the community are very easy to trace using our suggested approach. Suppose a fourth player joins the group of three depicted in Fig. 2. To identify the new equilibrium, we merely add the new player's replacement graph to the existing ones. There are two possibilities. It is possible that, at the equilibrium of the three-player community the fourth player would choose to contribute zero. This will be the case if the extra player's replacement graph hits the horizontal axis in Fig. 2 at a point to the left of $G^N_.$ The equilibrium level of total provision, and the choices and utilities of the three initial players, are unchanged. The fourth player chooses to contribute nothing, and enjoys the existing level of public good, while allocating all of his money income to private good consumption. Alternatively, the replacement value of the new player is positive at the existing equilibrium. In this case, the graph of the aggregate replacement function shifts upwards in the neighborhood of the initial equilibrium. The new equilibrium involves a higher total provision level. Existing contributors will reduce their individual contributions, and all are advantaged by the addition of the extra player.

In the presence of a large number of potential contributors who may differ in terms of incomes, preferences or unit costs, the diagram strongly suggests the conclusion reached by Andreoni (1988) – namely, that when n is large, the proportion of players who make strictly positive equilibrium contributions may be vanishingly small. Almost all players choose zero contributions.

Extensions

Early attempts to apply the voluntary contributions model – for example, to charitable giving, in which the aggregate G is the total quantity subscribed to some good cause – suggest that the very strong implications of the simple model – neutrality when unit costs are the same across contributors, and its implication that, when n is large, the number of strictly positive contributors will be a very small fraction of n – are difficult to square with empirical evidence. In addition, recent concerns with global and regional public goods have led to an interest in situations that naturally seem to involve public good technologies other than the summation one described above. We now briefly review some of the recent extensions and modifications of the model.

Technology of Public Good Provision

Hirshleifer (1983) suggested two types of public good which are not captured by the summation technology and which, he argued, may be of empirical significance. They are characterized by different public good provision technologies. Best-shot and weakest-link public goods are captured, respectively by the following technologies:

$$\text{Best}-\text{shot}: G = Max\{g_1, g_2, \ldots, g_n\}$$

$$\text{Weakest}-\text{link}: G = Min\{g_1, g_2, \ldots, g_n\}.$$

Hirshleifer's example of a best-shot public good involves defensive guns ringing a city, each trying to shoot down an approaching missile. What matters to the city's inhabitants is the accuracy of the single most accurate shot. His example of a weakest-link involves a group of farmers, each owning a pie-shaped slice of land within a circular area surrounded by sea. Each is responsible for the maintenance of his part of the perimeter dyke. In the event of a storm that threatens to breach the dyke, it is the level of maintenance of the least well-maintained stretch of wall that determines the level of security enjoyed by all. Sandler (2004) suggests a wide range of situations involving regional or global public goods that are better captured by one or other of these formulations than by the standard summation formulation.

These formulations have very distinctive equilibrium properties. For example, consider a two-player model with the weakest link technology in which there is an equilibrium at which both contribute, say, ten units to the public good. Then any allocation at which each is contributing x units, where x lies between zero and ten, is also an equilibrium. After all, if the other player is contributing x units, it does not pay you to contribute any more than x, since the total provision is defined by the smaller individual contribution. This game there can have a continuum of equilibria. Hirshleifer himself suggested that the players may be expected to choose the Paretodominant equilibrium. However, experimental evidence suggests that players find it surprisingly hard to coordinate on the Pareto dominant equilibrium.

Cornes (1993) and Cornes and Hartley (2007b) consider the class of games in which the total level of a public good is generated by individual contributions according to a constant returns to scale CES production process: $G = \left[\sum_{i=1}^{n} g_i^{v}\right]^{\frac{1}{v}}$ The summation model is obtained by putting $v = 1 : v \rightarrow +\infty$ generates the best shot, and $v \rightarrow -\infty$ generates the weakest link. They show that, if $-\infty < v < 1$, the resulting weaker link model has a unique equilibrium. It is only at the limit, when the isoquants associated with the production technology are L-shaped, that Hirshleifer's problem of multiple equilibria arises. Moreover, if player i is contributing less than player j at an equilibrium – perhaps because her income is lower, or because she has less interest in the public good – then player i has a higher marginal product as a contributor. Hence, neutrality with respect to income transfers breaks down, and a transfer from player j to player i may lead to a higher equilibrium level of public good provision and may be Pareto improving.

Situations involving $v > 1$ are better-shot games. Here, the production technology is inherently nonconvex, and again multiple equilibria may arise. For finite values of v, an equilibrium may involve positive contributions by each of a team of positive contributors, while the rest make zero contributions. However, there may be many such equilibria, each involving a different team of contributors. In Hirshleifer's best-shot case, if there are n players, there may be n equilibria, each of which involves a single 'champion', or 'dragon-slayer', who is the sole positive contributor, while all others make zero contributions. Again, achieving an equilibrium requires the players to resolve a tricky coordination problem.

Preferences

Cornes and Sandler (1984, 1994) extend the basic model by modifying the individual preferences. They included player i's own contribution as an argument of her own utility function, in addition to the aggregate quantity G:

$$u_i(.) = u_i(y_i, g_i, G).$$

They suggest this formulation as a model of charitable giving, a suggestion explored by Andreoni (1988). Donor i not only cares about the total amount given to the charitable giving, G, but also experiences a 'warm glow' of satisfaction from her own contribution, g_i. If the standard resource constraint and public good technology are retained, this modification is sufficient to produce very rich comparative static possibilities: neutrality does not generally hold, and an increase in player i's money income alone may either increase or reduce the equilibrium level of G. Finally, note that if the utility function is assumed to take the Cobb-Douglas form – $u_i(.) = y_i^\alpha g_i^\beta G^\gamma$ – then at any equilibrium every player will make a positive contribution to the public good. A proliferation of noncontributors as the number of players increases is no longer implied.

This extension significantly broadens the range of potential applications of the model. First, there is nothing to stop us from considering situations in which player i regards G as a public bad $-\frac{\partial u_i(.)}{\partial G} < 0$. Thus the model may be interpreted as one involving congestion or pollution. Each may still be a positive contributor at equilibrium, the pollution or congestion being an incidental by-product that is jointly generated alongside the private good g_i. Kotchen (2006) and Ruebbelke (2002) have explored such models.

Morgan (2000) and Duncan (2002) have used a slight modification of this model to investigate the potential role that lotteries, or raffles, may play in raising the public good level above that implied by the voluntary contribution model. The basic idea is simple. The presence of the public good by itself involves a positive externality, and will tend to be underprovided. If individuals buy lottery tickets, each of which partially contributes to the public good and also gives its purchaser a probability of winning a money prize, a negative externality is thereby added – by buying a ticket, and increasing my chance of winning the prize, I inflict a negative externality on other ticket holders. There are two externalities, one beneficial and one harmful. The resulting equilibrium, at which these externalities tend to counteract each other, may involve a higher level of public good provision than if it were provided simply by individual contributions in the absence of the lottery.

Dynamic Models

Up to this point, our discussion has remained firmly within the context of a one-shot static game. It is natural to wonder how the properties of equilibrium – in particular its presumptive inefficiency – are affected if we allow the contribution game to be played over many time periods. Schelling (1960, p. 45) suggested that such a setting may allow each player to make a small contribution, then wait to see whether others follow suit, before deciding whether to make a further small contribution. His suggestion has been analysed more formally by others, notably by Admati and Perry (1991) and Marx and Matthews (2000).

The last-named authors, whose analysis includes a useful discussion of the difference between their model and that of Admati and Perry, allow every player to choose a contribution level in each time period – any non-negative contribution, however large or small, is admissible. The properties of equilibria depend on (i) the degree of heterogeneity of players' valuations of the public good, (ii) the rate at which future costs and benefits are discounted, and (iii) whether or not there is a significant step in the benefit function – for example, a bridge generates no benefits until it is completed, thus representing an extreme example of a benefit function with a discrete step. They provide good news and bad news. The good news is that, if contributions can be made in small increments over time, equilibria can be attained that are more efficient than the equilibrium associated with the one-shot game. They argue that, if players' valuations are similar, and the rate of discount low, then nearly efficient equilibria exist. Furthermore, the presence of a significant benefit jump helps the prospects of successful completion of a project. An efficient equilibrium of the dynamic game may exist even in situations in which the only equilibrium of the static game involves zero contributions. The bad news is that, in common with many other dynamic games, there also exist other equilibria involving zero contributions.

Duffy et al. (2007) have investigated the properties of such dynamic models experimentally. They confirm that contributions do indeed tend to be higher in dynamic games of the kind proposed by Marx and Matthews, but their results cast doubt on the claimed importance of jumps in the benefit function.

See Also

- ► Non-cooperative Games (Equilibrium Existence)
- ► Public Goods

Bibliography

Admati, A., and M. Perry. 1991. Joint projects without commitment. *Review of Economic Studies* 58: 259–276.

Andreoni, J. 1988. Privately provided public goods in a large economy: The limits of altruism. *Journal of Public Economics* 35: 57–73.

Bergstrom, T.C., L. Blume, and H. Varian. 1986. On the private provision of public goods. *Journal of Public Economics* 29: 25–49.

Cornes, R.C. 1993. Dyke maintenance and other stories: Some neglected types of public good. *Quarterly Journal of Economics* 107: 259–271.

Cornes, R.C., and R. Hartley. 2007a. Aggregative public good games. *Journal of Public Economic Theory* 9: 201–219.

Cornes, R.C., and R. Hartley. 2007b. Weak links, good shots and other public good games: Building on BBV. *Journal of Public Economics* 91: 1684–1707.

Cornes, R.C., and T. Sandler. 1984. Easy riders, joint production and collective action. *Economic Journal* 94: 580–598.

Cornes, R.C., and T. Sandler. 1985. The simple analytics of pure public good provision. *Economica* 52: 103–116.

Cornes, R.C., and T. Sandler. 1994. Comparative static properties of the impure public good model. *Journal of Public Economics* 54: 403–421.

Cornes, R.C., and T. Sandler. 1996. *The theory of externalities, public goods and club goods*, 2nd ed. New York: Cambridge University Press.

Cornes, R.C., and T. Sandler. 2000. Pareto-improving redistribution in the pure public good model. *German Economic Review* 1: 169–186.

Duffy, J., J. Ochs, and L. Vesterlund. 2007. Giving little by little: Dynamic voluntary contribution games. *Journal of Public Economics* 91: 1708–1730.

Duncan, B. 2002. Pumpkin pies and public goods: The raffle fundraising strategy. *Public Choice* 111: 49–71.

Hirshleifer, J. 1983. From weakest-link to best-shot: The voluntary provision of public goods. *Public Choice* 41: 371–386.

Ihori, T. 1996. International public goods and contribution productivity differentials. *Journal of Public Economics* 61: 139–154.

Kotchen, M. 2006. Green markets and private provision of public goods. *Journal of Political Economy* 114: 816–834.

Marx, L., and S. Matthews. 2000. Dynamic voluntary contribution to a public project. *Review of Economic Studies* 67: 327–358.

Morgan, J. 2000. Financing public goods by means of lotteries. *Review of Economic Studies* 67: 761–784.

Ruebbelke, D. 2002. *International climate policy to combat global warming: An analysis of the ancillary benefits of reducing carbon emissions*. Cheltenham: Edward Elgar.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Samuelson, P.A. 1955. A diagrammatic exposition of a theory of public expenditure. *Review of Economics and Statistics* 37: 350–356.

Sandler, T. 2004. *Global collective action*. New York: Cambridge University Press.

Schelling, T. 1960. *The strategy of conflict*. Oxford: Oxford University Press.

Shibata, H. 1971. A bargaining model of the pure theory of public expenditure. *Journal of Political Economy* 79: 1–29.

Warr, P.G. 1983. The private provision of a public good is independent of the distribution of income. *Economics Letters* 13: 207–211.

Von Neumann Ray

A. Rubinov

The von Neumann ray determines the proportions of maximal balanced growth in a von Neumann technology. The economic growth trajectory, which realizes the maximum possible growth rate, that the economy could with-stand for infinite time is located on this ray. Let us give a more formal discription of the problem being discussed. The trajectory $x(0),\ldots, x(t),\ldots$ generated by this technology is called stationary if the proportions between goods in the state $x(t)$ are independent of time t. A stationary trajectory can be written in the form rm $x\ (t) = \gamma^t x$ where $x = x(0)$ is the initial state. This trajectory is generated by a technologically feasible activity (x, y) under which $\gamma x \le \gamma_1$. Usually a stationary trajectory is called the trajectory of balanced growth (although actual growth will take place only for $\gamma > 1$). The maximum number γ which enables a balanced growth is called the von Neumann (or technological) rate of growth for the technology Z. Thus the technological rate α is the solution of the following optimization problem find $\alpha = \max\ \gamma$ subject to

$$(x, y) \in Z, \qquad y \ge \gamma x.$$

If $y \ge \alpha x$, the process (x, y) is called the von Neumann activity (process), the corresponding vector x – the von Neumann vector and the ray passing through x – the von Neumann ray.

J. von Neumann in his pioneering paper (1937) established that a stationary price trajectory corresponds to the growth rate α, i.e., there exists a sequence $p(0), p(1),\ldots, p(t),\ldots$ of price vectors such that $p(t) = \alpha^{-\ \ t}p$ and $pw \le \alpha p\upsilon$ for all $(\upsilon, w) \in Z$ (this means exactly that $p(t)$ is a price trajectory). The vector p appearing in the definition of this trajectory is called the von Neumann price vector.

Thus for every von Neumann technology Z we can find the number α which is the solution of the problem (1), and a technologically feasible activity (x, y) and a price vector p satisfying the relations

$$(x, y) \in Z, \alpha x \le y, pw \le \alpha p\upsilon, ((\upsilon, w) \in Z) \qquad (2)$$

It can occur in degenerate cases that $p = 0$, i.e. all goods serving as inputs in a von Neumann process have zero prices. We shall exclude this (senseless from the economic point of view) situation and call $(\alpha, (x, y), p)$ a von Neumann equilibrium if it satisfies (2) (where α is the solution of the problem (1) and $px > 0$. The equilibrium has the following economic interpretation. If in the initial time period $t = 0$ the system is in the state $x(0) = X$ then it can develop with the maximum possible rate of growth α (the same for all goods) maintaining the initial proportions between goods. This development is implemented by the activity (x, y). It is possible to choose time-constant prices in such a manner that the interest factor pw/pv (equal to 1 + the rate of return) for any technologically admissible activity (υ, w) does not exceed α. For the activity (x, y) this interest factor is maximal and equals α.

Using the notion of characteristic prices we can say that the stationary equilibrium trajectory of the economic system moving along the von Neumann ray with the rate α admits as a characteristic a stationary price trajectory with the same price decline rate α.

Now we consider a von Neumann technology in the narrow sense Z. Recall that it is defined by an input matrix A and an output matrix B. For this technology the conditions (2) reduce to the following inequality system $\alpha\ Au \le Bu$, $pb \le \alpha\ pA$ where u is an m-vector of intensities.

Let the vector u, p satisfy this system with

$$\alpha = \max\{\gamma : \gamma Au \le Bu, u \ge 0\}$$

Then p is the vector of von Neumann prices, u is the so-called vector of von Neumann intensities; it determines the equilibrium vector $x = Au$.

In terms of equilibrium it is possible to characterize goods for which growth at a rate exceeding the von Neumann growth rate α is technologically possible. Let (x, y) be an activity such that the output of good i is greater than its input multiplied by α. Then it can be easily seen that the equilibrium price of the good i is equal to zero; in other words, this good is free. In short, this property of the equilibrium can be stated as follows: if the growth rate for some good exceeds the technological growth rate, then this good is free.

Now we point out another property of equilibrium for a von Neumann technology, in the narrow sense defined by an input matrix A and an output matrix B. The pair (a, b), where a is the ith column of A, b is the ith column of B, defines the ith basic activity of this technology. To every basic activity we can associate its interest factor pb/pa.

We can choose among the basic activities the most profitable ones, i.e. those for which the interest factor is maximal (equal to *pb/pa*). An important property of an equilibrium activity (*x, y*) is that it can be obtained by a joint use (with some intensities) only of the most profitable activities. If **u** is a von Neumann intensity vector then its components corresponding to the activities with non-maximal profitability are equal to zero.

We characterized the growth rate from a purely technological point of view. If the technology *Z* is 'indecomposable', i.e. for the production of some goods all goods are (directly or indirectly) used, then this growth rate admits an economic description. To demonstrate this consider stationary price trajectories, i.e. sequences of the form.

$$q, \beta^{-1}q, \ldots, \beta^{-t}q, \ldots \qquad (3)$$

where q is the price vector such that $qw \leq \beta\, q\upsilon$ for all technologically admissible activities (υ, w). If q is given then the minimal number β for which the sequence (3) is a price trajectory coincides with $\beta(q) = \max\}(q(w)/q(\upsilon) : (\upsilon, w) \in Z\}$ which is the maximal (at prices q) growth rate. The quantity $\beta(q) - 1$ is the maximal rate of return at prices q,

The economic growth rate for the technology Z is the minimal number β for which a stationary price trajectory exists. If this trajectory is generated by a price vector p, i.e. has the form

$$p(0), \ldots, p(t), p(t+1), \ldots$$

with $p(t) = \beta^{-1}p$ then the vector p is such that the maximal rate of return $\beta\,(p) - 1$ defined by p does not exceed the rate of return $\beta(q) - 1$ for any price vector q.

It turns out that if the technology Z is indecomposable in the aforementioned sense then the economic growth rate β co-incides with the technological growth rate α, the prices p with the minimal rate of return $\beta\,(p) - 1$ being von Neumann prices. To clarify the situation, introduce the following definition. The number α for which there exist an activity (x, y) and a vector p satisfying (2) and the inequality $px > 0$ is called a growth rate. It turns out that for the indecomposable nondegenerate case the technology admits only one growth rate which is simultaneously the technological and the economic one. Thus if some number α, for some $(x, y)) \in Z$ and p the inequalities (2) and $px > 0$ are satisfied, then α simultaneously solves the problems of maximizing the rate of reproduction and of minimizing the rate of return $\beta(p) - 1$.

In the decomposable case the situation is much more tangled: several growth rates can exist. Nevertheless their number does not exceed the number of goods.

Further we shall consider only indecomposable technologies. Let $x = x(0)$ be a vector with non-negative components representing the endowments at the moment $t = 0$. Choosing in one way or another the activities we can form various trajectories of length T begining in $x(0)$. Among those of special interest are trajectories which are optimal in terms of some price vector q. If the point $x(0)$ belongs to the von Neumann ray and q coincides with the von Neumann price vector then optimal behaviour consists in moving with the maximum technologically possible rate α along the von Neumann ray. It turns out that for a sufficiently wide class of initial states $x(0)$ and vectors q the optimal trajectories must grow with a rate which differ little from α.

Let us discuss this in more detail. Let p be the von Neumann price vector. If the trajectory $x(0)$, $\ldots$, $x(T)$ of length T is such that for a sufficiently large number of moments t the inequality $px(t+1)/px(t) \leq$ with $\gamma < \alpha$ holds then the mentioned trajectory cannot be optimal. This assertion can be elaborated in many ways. It has a very elegant and transparent geometrical interpretation.

Consider a von Neumann technology Z and choose among its activities the most profitable ones (i.e. those with the maximal rate of return according to von Neumann prices p).

These activities form a facet of the convex cone Z which is called a von Neumann facet. The further it is from the von Neumann facet the less profitable is any activity. Thus, an overwhelming majority of the activities taking part in the construction of the optimal trajectory lie near the von Neumann facet. Such assertions are usually caled turnpike theorems in the weak form. More precisely, the number of activities lying 'far' from the facet does not exceed some number independent of the length of the trajectory. Under some additional assumptions the activities essentially different from the facet can occur only at the beginning and the end of the trajectory (turnpike theorem in the strong form). Finally, some additional assumptions guarantee that the activities forming the trajectory simply belong to the facet (turnpike theorem in the strongest form).

Suppose that Z is a von Neumann technology in the narrow sense. Then the von Neumann facet has as its extreme rays the most profitable basic activities. We recall that every activity (υ, w) from Z is formed as a combination of basic activities with some intensities. The closeness of (υ, w) to the facet means that in its formation the most profitable activities are used with substantially greater intensities than the other activities. This activity belongs to the facet if only the most profitable activities are actually used.

We mention now the case when there is only one most profitable activity (x, y) (this case is typical for the technologies described by production functions). The von Neumann facet in this case coincides with the ray passing through the $2n$-dimensional vector (x, y). Instead of deviation of the activities from this ray we can speak about the deviation of the trajectory itself (more precisely, of its state $x(t)$) from the von Neumann ray which in this case is spanned by the vector x. The fact that a point has a small deviation from the von Neumann ray means simply that the proportions between its coordinates differ insignificantly from the proportions on the ray. This permits us to interpret the turnpike theorems from another point of view, for example, the theorem in the strong form means that the proportions between products for the states of the optimal trajectory can differ substantially from those on the ray only at the beginning and the end of the trajectory. (The first is caused by the difference of the initial state $x(0)$ from the von Neumann vector x, the second by the difference of the optimality criterion from the vector of von Neumann prices.)

Bibliography

Kemeny, J., O. Morgenstern, and G. Thompson. 1956. A generalization of the von Neumann model of an expanding economy. *Econometrica* 24: 115–135.

Makarov, V.L., and A.M. Rubinov. 1977. *Mathematical theory of economic dynamics and equilibria*. New York: Springer-Verlag.

Nikaido, H. 1964. Persistence of continual growth near the von Neumann ray: A strong version of the Radner turnpike theorem. *Econometrica* 32(1–2): 151–163.

Radner, R. 1961. Paths of economic growth that are optimal with regard only to final states; A turnpike theorem. *Review of Economic Studies* 28: 98–104.

von Neumann, J. 1945–6. A model of general economic equilibrium. *Review of Economic Studies*. 13: 1–9.

Von Neumann Technology

V. Makarov

The von Neumann technology is a convenient tool for the description and analysis of a wide variety of economic systems. It can be considered a special form of describing the production possibility set (i.e. the production process of the economic system, a form mostly designed for mathematical research of development dynamics).

The production process of this technology is determined by definition of input and output of goods corresponding to contiguous time intervals. An arbitrary production process can be described in this framework by introducing additional intermediate goods. We give a more formal description of the considered situation. Consider an economy with n goods, where we understand the term 'goods' in a very broad sense. Depending on the economic situation we can number among the goods not only goods in the usual sense of the world but also various types of capital, labour, natural resources as well as some conditional goods (e.g. the effect of consumption of some other goods).

A technology is a set Z consisting of technologically feasible processes (activities) Z. Every activity transforms a given set of goods (input vector) into another set (outout vector). Thus formally the activity is represented by a pair of vectors $Z = (x, y)$, where x is the input vector and y the output vector, both of them being n-dimensional vectors with non-negative components.

Considering the technology we assume that all technologically admissible activities have the same duration (a unit time interval). This hypothesis is based on the assumption that a longrun process can be decomposed into several processes of unit length. As a result of this decomposition intermediate goods (e.g. capital vintages or unfinished products) can be introduced.

Now we point out the essential features of the von Neumann technology Z.

(1) Any activity can be used at any intensity: i.e. $(x, y) \in Z, \lambda \geqslant 0$ implies $\lambda(x, y) \in Z$. This property reflects the possibility of an unlimited use of resources.

(2) Any two activities can be used jointly: $(x, y) \in Z, (u, \upsilon) \in Z$ implies $(x + u, y + u) \in Z$. Geometrically (1) and (2) mean that the von Neumann technology can be described by a convex cone.

(3) All goods can be produced. This means (together with (2) that there exists an activity (x, y) such that all coordinates of the vector y are positive.

(4) Non-zero output is impossible without input.

The von Neumann technology Z in the narrow sense (in another terminology: the von Neumann model, the model of an expanding economy) is defined through specification of m activities which are termed basic; it is the set of all input–output vectors which can be obtained by the joint use of the basic activities with arbitrary intensities. Geometrically, Z is a polyhedral cone with activities as its extreme rays. Algebraically, it is convenient to define Z by a pair of $m \times n$-matrices: the input matrix A and the output matrix B. If (a, b) is the i th basic activity, the vector a is the i th column of the matrix A, b is the i th column of the matrix B. Then

$$Z = \{(x, y) : x = Au, yBu, u \geq 0\}$$

where u is an m-vector of intensities. The condition (3) (resp. 4) is equivalent to the absence of zero columns in the matrix A (resp. to the absence of zero rows in the matrix B). These properties were formulated by Kemeny, Morgenstern and Thompson in 1956. von Neumann in his fundamental paper (1937; English translation: 1946) assumed a stronger condition: in every activity every good is either consumed or produced.

The von Neumann technology in a broad sense (in another terminology: the Neumann–Gale model) is merely a closed (in the topological sense) set for which the conditions (1)–(4) are fulfilled. It was introduced by Gale in 1956. Such technologies arise, for example, in connection with the use of production functions.

The von Neumann technology is a formal mathematical object that can be used for modelling various economic situations. One such situation was considered by J. von Neumann. He studied a closed economic system (i.e. having no connections with the outer world). The production possibilities of the system are given by the input and output matrices. There is no outflow of consumption, the process of production includes the reproduction of labour force, the workers save nothing, all capitalists' returns are invested. In other works, von Neumann abstracts from consumption and savings and concentrates solely on the process of production. A detailed analysis of the underlying economic assumptions is given in (Champernowne 1946).

Some deep generalizations of the von Neumann technology describing an open economy and explicitly taking into account consumption, labour and wages were studied by Morgenstern and Thompson and by J. Los and his pupils.

Various modifications of this model in the framework of a von Neumann technology (possibly, in a broad sense) can be given. As an example we describe a simple macroeconomic model of a firm. Let $F(K, L)$, the production function describing the performance of the firm where K is the capital and L the labour force. It is supposed that any part of the output $F(K, L)$ obtained with the capital K and the labour force L can be turned into investment I, the remaining part being used for purchasing the labour force l. The wage rate ω and the capital deterioration rate μ re given. The set of states (k, l) the firm can reach (in a unit time interval) from the state (*K, L*) is described by a system of inequalities

$$0 \leq \mathrm{k} \leq (1-\mu)k + I, I + \omega l \leq F(k,l)l \geq 0,\\ I > 0. \tag{$*$}$$

If the function *F* satisfies the traditional assumptions of concavity and homogeneity of degree 1 then the set of activities ((*K, L*), (*k, l*) satisfying (*)) is a von Neumann technology (in a broad sense).

The von Neumann technology is often used for representing the production part in various models of economic dynamics. Models with utility functions explicitly taking into account consumption, as well as dynamic Leontief models can be stated and analysed in this framework as well. We note furthermore than many other problems not connected with economic dynamics can be embedded into a von Neumann technology scheme, in particular, 'bottleneck problems'.

Thus with the help of the von Neumann technology we can study a demographic model of population movement, based on the following hypothesis: the number of marriages between men and women under certain ages is proportional to the minimum of the numbers of unmarried men and women under these ages. Men and women under certain ages, and also their newly created families which are distinguished according to the terms of their existence, play the role of 'products' here.

The technological activities describe a shift of 'products' from one age group to another, and the processes of family increase and decrease.

As a rule the von Neumann technology is analysed from two viewpoints. First, equilibrium states of the economic system can be determined in these terms. J. von Neumann introduced it specially for this purpose. Second, this technology is a convenient tool for analysing development trajectories of the economic system. Both directions are closely interconnected. The concept of von Neumann equilibrium (geometrically: the von Neumann ray) is extremely important in these problems. Here we focus our attention on the trajectory concept.

In many situations modelled with the von Neumann technology it is reasonable to guess that the development of the underlying economic system is such that the input vector at the beginning of some time period does not exceed the output vector at the end of the preceding period. First of all, it is true for the original von Neumann construction; the same holds true for the model of the firm described in (*). Thus we can give the following formal definition. The sequence $x(0), \ldots,$ $x(T)$ is called a trajectory of length *T* generated by a von Neumann technology *Z* if the relations

$$(x(t), y(t+1)) \in Z, x(t+1) \leq y(t+1),\\ t = 0, 1, \ldots, T-1$$

hold for some vectors $y(t)$. The trajectories which are optimal in the sense that, the output value $p(T)$ $x(T)$ at moment *T* is greater than or equal to the output value for any other trajectory beginning at $x(0)$ are of special interest here ($p(T) \geqslant 0$) is the given price vector at the moment *T, px* is the scalar product of the vectors *p* and *x*).

Sometimes efficient trajectories $x(0), \ldots, x(T)$ are considered. Efficiency means that from the point $x(0)$ it is impossible to reach in *T* steps the point $\lambda x(t)$ with $\lambda \geqslant 1$; in other words trajectories of the form $x(0), \ldots, \lambda x(T)$ do not exist. Under some natural assumptions on the technology the trajectory is efficient if and only if there exists a price vector $p(T)$ for which the trajectory is optimal. One can consider infinite trajectories $x(0),$ $\ldots, x(t), x(t+1), \ldots$ as well. An infinite trajectory is called efficient if each of its segments $x(0),$ $\ldots, x(t)$ is efficient for any $t > 0$. The interest is infinite efficient trajectories is not motivated solely by the desire to understand the system's behaviour in the far future. Much more concretely, the fact that $x(1)$ must belong to the infinite efficient trajectory beginning at $x(0)$ is often a very restrictive assumption, which allows us to determine uniquely the output $x(1)$ among all feasible outputs generated by the input $x(0)$.

The von Neumann technology *Z* generates not only the trajectories of goods describing the material flows in the economy but the price trajectories describing the financial flows. It is supposed that the price vector $q \geqslant 0$ at the moment $t + 1$ (given the price vector p at the moment t) is chosen in such a manner that the value of any output y (at moment $t + 1$) does not exceed the value of the input x at moment t). Thus we have the following definition: the sequence $p(0), \ldots, p(t), \ldots$ is a price trajectory if $p(t+1)y \leq p(t)x$ for all $(x, y) \in Z, t = 0, 1 \ldots,$ If $x(0), \ldots, x(t), \ldots$ is a goods trajectory, and $p(0), \ldots, p(t), \ldots$ is a price trajectory, then the inequalities

$$p(0)x(0) \geq p(1)x(1) \geq \ldots p(t)x(t) \ldots$$

are valid.

Let us consider now the case of a von Neumann technology (in the narrow sense) given by an input matrix A and an output matrix B. The (goods) trajectory $x(0), \ldots, x(t)$ generated by the technology Z is determined by the sequence of intensity vectors $u(t)$ such that $Bu(t) \leq Au(t+1)$. This sequence is called the intensity trajectory. In this case the price trajectory is a sequence $p(t)$ such that $p(t+1)B \leq p(t)A$.

The efficient trajectory $x(0), \ldots, x(t), \ldots$ generated by some von Neumann technology Z can be characterized by a system of 'shadow prices' $p(0), \ldots, p(t), \ldots$. The corresponding result (often called the characteristic theorem) is in a sense analogous to the duality theorem of linear programming and can be interpreted in a similar manner. Under some natural additional assumptions it is: the trajectory $x(0), \ldots, x(t), \ldots$ is efficient if and only if there exists a price trajectory $p(0), \ldots, p(t), \ldots$ such that $p(t) \neq 0$ for all t and

$$p(0)x(0) = p(1)x(1) = \cdots = p(t)(t) = \cdots$$

All this can be fully carried over to the case when at every moment t a new technology $Z(t)$ is used. The discussion of trajectory properties and, in particular, the characteristics theorem is contained in Makarov and Rubinov (1977).

See Also

- ▶ General Equilibrium
- ▶ Linear Models
- ▶ Von Neumann Ray

Bibliography

Champernowne, D.G. 1945–6. A note on J. von Neumann's article on 'A model of general economic equilibrium'. *Review of Economic Studies* 13: 10–18.

Gale, D. 1956. The closed linear model of production. In *Linear inequalities and related systems*, Annals of Mathematics Studies, vol. 38, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Kemeny, J., O. Morgenstern, and G. Thompson. 1956. A generalization of the von Neumann model of an expanding economy. *Econometrica* 24: 115–135.

Los, J. 1978. Mathematical theory of von Neumann economic models. Report on recent results. *Colloquia Mathematica* 40(2): 327–346.

Makarov, V.L., and A.M. Rubinov. 1977. *Mathematical theory of economic dynamics and equilibria*. New York: Springer.

Morgenstern, O., and G. Thompson. 1976. *Mathematical theory of expanding and contracting economies*. Lexington: D.C. Heath.

von Neumann, J. 1945–6. A model of general economic equilibrium. *Review of Economic Studies* 13: 1–9.

von Neumann, John (1903–1957)

Gerald L. Thompson

JEL Classifications
B31

His Life

Jansci (John) von Neumann was born to Max and Margaret Neumann on 28 December 1903 in Budapest, Hungary. He showed an early talent for mental calculation, reading and languages. In 1914, at the age of ten, he entered the Lutheran Gymnasium for boys. Although his great intellectual (especially mathematical) abilities were recognized early, he never skipped a grade and instead stayed with his peers. An early teacher, Laslo Ratz, recommended that he be given advanced mathematics tutoring, and a young mathematician Michael Fekete was employed for this purpose. One of the results of these lessons was von Neumann's first mathematical publication (joint with Fekete) when he was 18.

Besides his native Hungarian, Jansci (or Johnny, as he was universally known in his later life) spoke German with his parents and a nurse and learned Latin and Greek as well as French and English in school. In 1921 he enrolled in mathematics at the University of Budapest but promptly left for Berlin, where he studied with Erhard Schmidt. Each semester he returned to Budapest to take examinations without ever having attended classes. While in Berlin he frequently took a three-hour train trip to Göttingen, where he spent considerable time talking to David Hilbert, who was then the most outstanding mathematician of Germany. One of Hilbert's main goals at that time was the axiomatization of all of mathematics so that it could be mechanized and solved in a routine manner. This interested Johnny and led to his famous 1928 paper on the axiomatization of set theory. The goal of Hilbert was later shown to be impossible by Kurt Gödel's work, based on an axiom system similar to von Neumann's, which resulted in a theorem, published in 1930, to the effect that every axiomatic system sufficiently rich to contain the positive integers must necessarily contain undecidable propositions.

After leaving Berlin in 1923 at the age of 20, von Neumann studied at the Eidgenossische Technische Hochschüle in Zurich, Switzerland, while continuing to maintain his enrolment at the University of Budapest. In Zurich he came into contact with the famous German mathematician, Hermann Weyl, and also the equally famous Hungarian mathematician, George Polya. He obtained a degree in Chemical Engineering from the Hochschüle in Zurich in 1925, and completed his doctorate in mathematics from the University of Budapest in 1926. In 1927 he became a privatdozent at the University of Berlin and in 1929 transferred to the same position at the University of Hamburg. His first trip to America was in 1930 to visit as a lecturer at Princeton University, which turned into a visiting professorship, and in 1931 a professorship. In 1933 he was invited to join the Institute for Advanced Study in Princeton as a professor, the youngest permanent member of that institution, at which Albert Einstein was also a permanent professor. Von Neumann held this position until he took a leave of absence in 1954 to become a member of the Atomic Energy Commission.

Von Neumann was married in 1930 to Marietta Kovesi, and his daughter Marina (who became a vice-president of General Motors) was born in 1935. The marriage ended in a divorce in 1937. Johnny's second marriage in 1938 was to Klara Dan, whom he met on a trip to Hungary. They maintained a very hospitable home in Princeton and entertained, on an almost weekly basis, numerous local and visiting scientists. Klara later became one of the first programmers of mathematical problems for electronic computers, during the time that von Neumann was its principal designer.

In 1938 Oskar Morgenstern came to Princeton University. His previous work had included books and papers on economic forecasting and competition. He had heard of von Neumann's 1928 paper on the theory of games and was eager to talk to him about connections between game theory and economics. In 1940 they started work on a joint paper which grew into their monumental book, *Theory of Games and Economic Behavior* published in 1944. Their collaboration is described in Morgenstern (1976).

Von Neumann became heavily involved in defence-related consulting activities for the United States and Britain during World War II. In 1944 he became a consultant to the group developing the first electronic computer, the ENIAC, at the University of Pennsylvania. Here he was associated with John Eckert, John Mauchly, Arthur Burks and Herman Goldstine. These five were instrumental in making the logical design decisions for the computer, for example, that it be a binary machine, that it have only a limited set of instructions that are performed by the hardware, and most important of all, that it run an internally stored program. It was acknowledged by the others in the group that the most important design ideas came from von Neumann. The best account of these years is Goldstine (1972). After the war von Neumann and Goldstine worked at the Institute of Advanced study where they developed (with others) the JONIAC computer, a successor to the ENIAC, which used principles some of which are still being used in current computer designs.

In 1943 von Neumann became a consultant to the Manhattan Project which was developing the atomic bomb in Los Alamos, New Mexico. This work is still classified but it is known that Johnny performed superbly as a mathematician, an applied physicist, and an expert in computations. His work continued after the war on the hydrogen bomb, with Edward Teller and others. Because of this work he received a presidential appointment to the Atomic Energy Commission in 1955. He took leave from the Institute for Advanced Study and moved to Washington. In the summer of 1955 he fell and hurt his left shoulder. Examination of that injury led to a diagnosis of bone cancer which was already very advanced. He continued to work very hard at his AEC job, and prepared the Silliman lectures (von Neumann 1958), but was unable to deliver them. He died on 8 February 1957 at the age of 53 in the Walter Reed Hospital, Washington, DC.

The Theory of Games

Without question one of von Neumann's most original contributions was the theory of games, with which it is possible to formulate and solve complex situations involving psychological, economic, strategic and mathematical questions. Before his great paper on this subject in 1928 there had been only a handful of predecessors: a paper by Zermelo in 1912 on the solution in pure strategies of chess; and three short notes by the famous French mathematician E. Borel. Borel had formulated some simple symmetric two-person games in these notes, but was not able to provide a method of solution for the general case, and in fact conjectured that there was no solution concept applicable to the general case. For a commentary on the priorities involved in these two men's work see the notes by Maurice Frechet, translations (by L.J. Savage) of the three Borel papers, and a commentary by von Neumann, all of which appeared with von Neumann (1953a).

The three main results of von Neumann's 1928 paper were: the formulation of a restricted version of the extensive form of a game in which each player either knows nothing or everything about previous moves of other players; the proof of the minimax theorem for two-person zero-sum games; and the definition of the characteristic function for and the solution of three-person zero-sum games in normal form. Von Neumann also carried out an extensive study of simplified versions of poker during this time, but they were not published until later.

The *extensive* form of a game is the definition of a game by stating its rules, that is, listing all of the possible legal moves that a player can make for each possible situation he can find himself in during a play of the game. A *pure strategy* in a game is a much more complicated idea – a listing of a complete set of decisions for each possible situation in which the player can find himself. A complete enumeration of all possible strategies shows that the number of such strategies is equal to the product of the number of legal moves for each situation, which implies that there is an astronomical number of possible strategies for any non-trivial game such as chess. Most of these are bad, and would never be used by a skilful player, but they must be considered to find its solution. The *normalized* form of a game is obtained by replacing the definition of a game as a statement of its rules, as is done in its extensive form, by a listing of all of the possible pure strategies for each player. To complete the normalized form of the game, imagine that each player has made a choice of one of his pure strategies. When pitted against another a unique (expected) outcome of the game will result. For the moment we will imagine that the outcome of the game is monetary, and therefore each player gets a 'pay-off' at the end of the game which is actually money. (Later we will replace money by 'utility'.) If the sum of the payments to all players is zero the game is said to be *zero-sum*; otherwise it is a *non-zero-sum* game.

The normalized form of a game is also called a *matrix game*, and any real $m \times n$ matrix can be considered a two-person zero-sum game. The row player has m pure strategies, $i = 1, \ldots, m$, and the column player has n pure strategies, $j = 1, \ldots, n$. If the row player chooses i and the column player chooses j then the payoff $a(i, j)$ is exchanged between them, where $a(i, j) > 0$ means that the row player receives $a(i, j)$ from the column player, while a negative payoff means that the column player receives the absolute value of that amount from the row player.

The importance of the careful analysis of the extensive and normalized forms of a game is that it separates out the concept of strategy and psychology in any discussion of a game. As an example, in poker bidding high when having a weak hand is commonly called 'bluffing', and considered an aggressive form of play. As a result of this formulation, and the solution of simplified versions of the game von Neumann showed that in order to play poker 'optimally' it is necessary to bluff part of the time, i.e., it is a required part of the strategy of any good poker player. A similar analysis for simplified bridge shows that a required part of an optimal bridge strategy is to signal, via the way one discards low cards in a suit, whether the player holds higher cards in that suit.

The analysis of special kinds of games shows that some of them can be solved by using pure strategies. This class includes the games of 'perfect information' such as the board games of chess and checkers. However, even such a simple game as matching pennies shows that an additional strategic concept is needed, namely, that of a 'mixed strategy'. This concept appeared first in the context of symmetric two-person games in Borel's 1921 paper. Briefly, a mixed strategy for either player is a finite probability function on his set of pure strategies. For matching pennies the common strategy of flipping the penny to choose whether to play heads or tails is a mixed strategy that chooses both alternatives with equal probability (1/2), and is, in fact, an optimal strategy for that game.

We now discuss the way that von Neumann made precise the definition of a solution to a matrix game. Let A be an arbitrary $m \times n$ matrix with real number entries. Let x be an m-component row vector, and let f be an m-component column vector all of whose components are ones. Then x is a *mixed strategy* vector for the row player in the matrix game A if it satisfies: $xf = 1$ and $x \geq 0$. Similarly, let y be an n-component column vector, and let e be an n-component row vector of all whose components are ones. Then y is a mixed strategy vector for the column player in the matrix game A if it satisfies: $ey = 1$ and $y \geq 0$. Mixed strategy vectors are also called *probability* vectors because they have non-negative components that sum to one, and hence could be used to make a random choice of a pure strategy by spinning a pointer, choosing a random number, etc. To complete the definition of the solution to a game, we need a real number v, called the *value of the game*. The solution to the matrix game A is now a triple, a mixed strategy x for the row player, a mixed strategy y for the column player, and a value v for the game: these quantities must solve the following pair of (vector) inequalities:

$$xA \geq ve \, \text{and} \, Ay \leq vf.$$

Because these are linear inequalities, one might suspect (and would be correct) that the optimal x, y and v can be found by using a linear programming code and a computer.

However, in the 1920s it was not clear that such a solution existed. In fact, Borel conjectured that it did not. The most decisive result of von Neumann's 1928 paper was to establish, using an argument involving a fixed point theorem, his famous *minimax theorem* to the effect that for an arbitrary real matrix A there exists a real number v and probability vectors x and y such that

$$\underset{x}{Maximum}\ \underset{y}{Maximum}\ xAy$$
$$= \underset{y}{Maximum}\ \underset{x}{Maximum}\ xAy$$

This theorem became the keystone not only for the theory of two-person matrix games, but also for n-persons games via the characteristic function (to be discussed later).

We now discuss the major differences between von Neumann and Morgenstern (1944) and von Neumann's 1928 paper. The information available to each player was assumed, in the 1928 paper, to be the following: when required to move, each player knows either everything about the previous moves of his opponents (as in chess), or nothing (as in matching pennies). By using information trees, and partitioning the nodes of such trees into information sets, in 1944 this concept was extended to games in which players have only partial information about previous moves when they are required to make a move. This was a difficult but major extension, which has not been substantially improved upon since its exposition in the 1944 treatise.

A second major change in the basic theory of games was in the treatment of payoff functions. In the 1928 paper payoffs were treated as if they were monetary, and it was implicitly assumed that money was regarded as equally important by each of the players. In order to take into account the well-known objections, such as those of Daniel Bernoulli, to the assumption that a dollar is equally important to a poor man as a rich man, a monetary outcome to a player was replaced by the *utility* of the outcome. Although Bernoulli had suggested that the utility of x dollars should be the natural logarithm of x, so that the addition of a dollar to a rich man's fortune would be valued less than the addition of a dollar to a poor man's fortune, this specific utility concept was never universally accepted by economists. So utility remained a fuzzy, intuitive concept. Von Neumann and Morgenstern made the absolutely decisive step of axiomatizing utility theory, making it unambiguous and they can properly be said to have started the modern theory of utility, not only for game theory, but for all of economics and the social sciences.

Almost two-thirds of the 1944 treatise consists of the theory of n-person constant-sum games, of which only a small part, the three person zero-sum case, appears in the 1928 paper. When $n > 2$, there are opportunities for cooperation and collusion as well as competition among the players, so that there arises the problem of finding a way to evaluate numerically the position of each player in the game. In 1928 von Neumann handled this problem for the zero-sum case by introducing the idea of the *characteristic* function of a game defined as follows: For each *coalition*, that is, subset S of players, let $v(\mathrm{S})$ be the minimax value that S is assured in a zero-sum two-person game played between S and its complementary set of players.

To describe the possible division of the total gain available among the players the concept of an imputation, which is a vector $(x(1),\ldots, x(n))$ where $x(i)$ represents the amount the player i obtains, was introduced. For a coalition C in a constant-sum game $v(C)$ is the minimum amount that the coalition C should be willing to accept in any imputation, since by playing alone against all the other players, C can achieve that amount for itself. Except for this restriction there is no other constraint on the possible imputations that can become part of a solution. An imputation vector x is said to dominate imputation vector y if there exists a coalition C such that (1) $x(i) \geq y(i)$ for all i in C, and (2) the sum of $x(i)$ for i in C does not exceed $v(C)$. The idea is that that the coalition C can 'enforce' the imputation x by simply threatening to 'go it alone', since it can do no worse by itself.

One might think, or hope, that a single imputation could be taken as the definition of a solution to an n-person constant-sum game. However, a more complicated concept is needed. By a von Neumann–Morgenstern solution to an n-person game is meant a set S of imputations such that (1) if x and y are two imputations in S then neither dominates the other; and (2) if z is an imputation not in S, then there exists an imputation x in S that dominates z. Von Neumann and Morgenstern were unable (for good reasons, see below) to prove that every n-person game had a solution, even though they were able to solve every specific game they considered, frequently finding a huge number of solutions.

At the very end of the 1944 book there appears a chapter of about 80 pages on general non-zero-sum games. These were formally reduced to the zero-sum case by the technique of introducing a fictitious player, who was entirely neutral in terms of the game's strategic play, but who either consumed any excess, or supplied any deficiency so that the resulting $n + 1$ person game was zero-sum. This artifice helped but did not suffice for a completely adequate treatment of the non-zero-sum case. This is unfortunate because such games are the most likely to be found useful in practice.

About 25 years after the treatise appeared, William Lucas (1969) provided as a counter-example, a general sum game that did not have a von Neumann–Morgenstern solution. Other solution concepts have been considered since, such as the Shapley value, and the core of a game.

One of the most interesting non zero-sum games considered in that chapter was the so-called *market game*. The first example of a market game (though it was not called that) was the famous horse auction of Böhm-Bawerk, published in 1881. The horses had identical characteristics, each of 10 buyers had a maximum price he was willing to bid, and each of 8 sellers had a minimum price he was willing to accept. Böhm-Bawerk's solution was to find the 'marginal pairs' of prices, which turned out to be included in the von Neumann–Morgenstern solution to this kind of game. Later work on this problem was done by Shapley and Shubik (1972) and Thompson (1980, 1981).

The Expanding Economy Model

Another of von Neumann's original contribution to economics was von Neumann (1937), which contained an expanding economy model unlike any other economic model that preceded it. When von Neumann gave a seminar to the Princeton economics department in 1932 on the model, which was stated in terms of linear inequalities not equations, and whose existence proof depended upon a fixed point theorem more sophisticated than any published in the mathematics literature of the time, it is little wonder that he made no impression on that group. He repeated his talk on the subject at Karl Menger's mathematical seminar in Vienna in 1936, and published his paper in German in 1937 in the seminar proceedings. The paper became more widely known after it was translated into English and published in *The Review of Economic Studies* in 1945 together with a commentary by Champernowne.

Von Neumann's model consists of a closed production economy in which there are m processes and n goods. In order to describe it we use the vectors e and f previously defined together with the following notation:

x is the $m \times 1$ intensity vector with $xf = 1$ and $x \geq 0$.
y is the $1 \times$ price vector with $ey = 1$ and $y \geq 0$.
$\alpha = 1 + a/100$ is the expansion factor, where a is the expansion rate.
$\beta = 1 + b/100$ is the interest factor, where b is the interest rate. The model satisfies the following axioms:

Axiom 1 . $xB \geq \alpha xA$ or $x(B - \alpha A) \geq 0.$
Axiom 2 . $By \leq \beta Ay$ or $x(B - \beta A)y \leq 0.$
Axiom 3 . $x(B - \alpha A)y = 0.$
Axiom 4 . $x(B - \beta A)y = 0.$
Axiom 5 . $xBy > 0.$

Axiom 1 makes the model closed, i.e., the inputs for a given period are the outputs of the previous. Axiom 2 makes the interest rate be such that the economy is *profitless*. Axiom 3 requires that overproduced goods be *free*. Axiom 4 forces inefficient processes not to be used. And Axiom 5 requires the total value of all goods produced to be positive.

In order to demonstrate that for any pair of nonnegative matrices A and B, solutions consisting of vectors x and y and numbers α and β exist, an additional assumption was needed:

Assumption V . $A + B > 0.$

This assumption means that every process requires as an input or produces as an output some amount, no matter how small, of every good. With this assumption, and the assumption that natural resources needed for expansion were available in unlimited quantities, von Neumann showed that necessarily $\alpha = \beta$, that is, that the expansion and interest factors were equal. In his paper, von Neumann proved a sophisticated fixed point theorem and used it to prove the existence theorem for the EEM.

D.G. Champernowne (1945) provided the first acknowledgement that the economics profession had seen the article, and also provided its first criticisms. We mention three:

(1) Assumption V which requires that every process must have positive inputs or outputs of every other good was economically unrealistic.
(2) The fact that the model has no consumption, so that labour could receive only subsistence amounts of goods as necessary inputs for production processes, also seems unrealistic.
(3) The consequence of Axiom 3 that overproduced good should be free is too unrealistic.

Criticisms 1 and 2 were removed by Kemeny et al. (1956), who replace Assumption V by:

Assumption KMT-1. Every row of A has at least one positive entry.

Assumption KMT-2. Every column of B has at least one positive entry. The interpretation of KMT-1 is that every process must use at least one good as an input. And the interpretation of KMT-2 is that every good must be produced by some process. With these assumptions they were able to show that there were a finite number of possible expansion factors for which intensity and price vectors existed satisfying the axioms. They also showed how consumption could be added into the model, which responded to criticism 2.

An alternative way of handling these criticisms appears in Gale (1956).

In Morgenstern and Thompson (1969, 1976), the third criticism above was answered by generalizing the model to become an ‘open economy’. In such an economy the price of an overproduced good cannot fall below its export price, and it cannot rise above its import price. Generalizations of the open model have been made by Los (1974) and Moeschlin (1974).

Von Neumann’s Influence on Economics

Although von Neumann has only three publications that can directly be called contributions to economics, namely, his 1928 paper on the theory of games, his 1937 paper (translated in 1945) on the expanding economy model and his 1944 treatise (with Morgenstern) on the theory of games, he had enormous influence on the subject. The small *number* of contributions is deceptive because each one consists of several different topics, each being important. We discuss these separately.

The expanding economy model, von Neumann (1937) consisted of two parts: the first input–output equilibrium model that permits expansion; and second the fixed point theorem. The linear input–output model is a precursor of the Leontief model, of linear programming as developed by Kantorovich and Dantzig, and of Koopman’s activity analysis. This paper, together with A. Wald (1935) raised the level of mathematical sophistication used in economics enormously. Many current younger economists are high-powered applied mathematicians, in part, because of the stimulus of von Neumann’s work.

The theory of games, von Neumann (1928) and von Neumann and Morgenstern (1944), was an enormous contribution consisting of several different parts: (1) the axiomatic theory of utility; (2) the careful treatment of the extensive form of games; (3) the minimax theorem; (4) the concept of a solution to a constant-sum n-person game; (5) the foundations of non-zero-sum games; (6) market games. Each of these topics could have been broken into a series of papers, had von Neumann taken the time to do so. And he could have forged a brilliant career in economics by publishing them. However, he found that making an exposition of the results that he had worked out in notes or in his head was less interesting to him than investigating still other new ideas.

Von Neumann’s indirect contributions, such as the theory of duality in linear programming, computational methods for matrix games and linear programming, combinatorial solution methods for the assignment problem, the logical design of electronic computers, contributions to statistical theory, etc. are equally, important to the future of economics. Each of his contributions, direct or indirect, was monumental and decisive. We should be grateful that he was able to do so much in his short life. His influence will persist for decades and even centuries in economics.

Selected Works

1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100: 295–320.

1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. In *Ergebnisse eines mathematische Kolloquiums*, vol. 8, ed. Karl Menger. Trans. as 'A model of general equilibrium', *Review of Economic Studies* 13 (1945–6): 1–9.

1944. (With O. Morgenstern.) *Theory of games and economic behavior*. Princeton: Princeton University Press. 2nd ed, 1947; 3rd ed, 1953.

1947. Discussion of a maximum problem. Unpublished working paper, Princeton, November, 9 pp.

1948. A numerical method for determining the value and the best strategies of a zero-sum two-person game with large numbers of strategies. Mimeographed, May, 23 pp.

1953a. Communications on the Borel notes. *Econometrica* 21: 124–125.

1953b. (With G.W. Brown.) Solutions of games by differential equations. In *Contributions to the theory of games*, Annals of mathematics studies no. 28, vol. 1, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

1953c. (With D.B. Gillies and J.P. Mayberry.) Two variants of poker. In *Contributions to the theory of games*, Annals of mathematics studies no. 28, vol. 1, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

1954. A numerical method to determine optimum strategy. *Naval Research Logistics Quarterly* 1: 109–115.

1958. *The computer and the brain*. New Haven: Yale University Press.

1963. *Collected works*, vols. I–VI. New York: Macmillan.

Bibliography

Champernowne, D.G. 1945–6. A note on J. von Neumann's article. *Review of Economic Studies* 13: 10–18.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*, Cowles Foundation monograph no. 17. New York: Wiley.

Gale, D. 1956. The closed linear model of production. In *Linear inequalities and related systems*, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press. A counter-example showing that optimal prices need not exist in Gale's original model was published by J. Hulsman and V. Steinmitz in *Econometrica* 40 (1972): 387–390. Proof of the existence of optimal prices in a modified Gale model was given by A. Soyster in *Econometrica* 42 (1974): 199–205.

Goldstine, H.H. 1972. *The computer from Pascal to von Neumann*. Cambridge, MA: MIT Press.

Heims, S.J. 1980. *John von Neumann and Norbet Wiener*. Cambridge, MA: MIT Press.

Kemeny, J.G., O. Morgenstern, and G.L. Thompson. 1956. A generalization of von Neumann's model of an expanding economy. *Econometrica* 24: 115–135.

Los, J. 1974. The existence of equilibrium in an open expanding economy model (generalization of the Morgenstern–Thompson model). In *Mathematical models in economics*, ed. J. Los and M.W. Los. Amsterdam/New York: North-Holland.

Lucas, W. 1969. The proof that a game may not have a solution. *Transactions of the American Mathematics Society* 137: 219–229.

Luce, R.D., and H. Raiffa. 1957. *Games and decisions: Introduction and critical survey*. New York: Wiley.

Moeschlin, O. 1974. A generalization of the open expanding economy model. *Econometrica* 45: 1767–1776.

Morgenstern, O. 1958. Obituary, John von Neumann, 1903–57. *Economic Journal* 68: 170–174.

Morgenstern, O. 1976. The collaboration between Oskar Morgenstern and John von Neumann on the theory of games. *Journal of Economic Literature* 14: 805–816.

Morgenstern, O., and G.L. Thompson. 1969. An open expanding economy model. *Naval Research Logistics Quarterly* 16: 443–457.

Morgenstern, O., and G.L. Thompson. 1976. *Mathematical theory of expanding and contracting economies*. Boston: Health–Lexington.

Oxtoby, J.C., B.J. Pettis, and G.B. Price (eds.). 1958. John von Neumann 1903–1957. *Bulletin of the American Mathematical Society* 64(3), Part 2.

Shapley, L.S. 1953. A value for n-person games. In *Contributions to the theory of games*, vol. II, ed. H.-W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Shapley, L.S., and M. Shubik. 1972. The assignment game. I: The core. *International Journal of Game Theory* 1: 111–130.

Shubik, M. 1982. *Game theory in the social sciences: Concepts and solutions*. Cambridge, MA: MIT Press.

Shubik, M. 1985. *A game theoretic approach to political economy*. Cambridge, MA: MIT Press.

Thompson, G.L. 1956. On the solution of a game-theoretic problem. In *Linear inequalities and related systems*, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Thompson, G.L. 1980. Computing the core of a market game. In *Extremal methods and systems analysis*, ed. A.V. Fiacco and K.O. Kortanek. Berlin: Springer.

Thompson, G.J. 1981. Auctions and market games. In *Essays in game theory and mathematical economics in honor of Oskar Morgenstern*, ed. R.J. Aumann et al. Mannheim: Bibliographisches Institut Mannheim.

Wald, A. 1935. Über die eindeutige positive Losbarkeit der neuen Produktionsgleichungen. In *Ergebnisse eines Mathematischen Kolloquiums*, vol. 6, ed. K. Menger, 12–20.

Vorob'ev, Nikolai N. (1925–1995)

Leon A. Petrosyan

Keywords

Coalitional games; Game theory; Markov, A.; Morgenstern, O.; Probability theory; Von Neumann, J.; Vorob'ev, N

JEL Classifications

B31

Nikolay Vorob'ev is commonly regarded as the founder and the leader of game-theoretic school in the former Soviet Union.

Vorob'ev was born on 18 September 1925 in Leningrad (now St Petersburg). His father was a lawyer and his mother a surgeon. Beginning his education at technical institutes in Izevsk and Moscow, he returned to Leningrad in 1944 and become a student at the Leningrad Shipbuilding Institute. In 1946 he began study at the Faculty of Mathematics and Mechanics of the Leningrad State University. In 1948 he left the Shipbuilding Institute and graduated from the university. In 1947 Vorob'ev published his first paper in semigroup theory.

In 1948 Vorob'ev started a postgraduate programme at the Leningrad branch of the Steklov Mathematical Institute. His supervisor was Professor A.A. Markov, under whose influence he studied constructive mathematical logic, which was rapidly developing at that time. His Candidate of Science thesis in mathematics was devoted to logical deduction rules in systems with strong negation. He received his Candidate of Science degree in 1952. In the same year he joined the Steklov Mathematical Institute as a junior research associate. Here he once more changed his scientific interests and started research concerned with both algebra and probability theory.

Axiomatic training in algebra and logic, along with studies in probability theory, permitted Vorobe'ev to make a transition to the study of game theory. His paper 'Controlled Processes and Game Theory' (1955) was the first paper in game theory published in the former Soviet Union. His 1959 review article 'Finite Noncooperative Games' served for many years as a primary Russian language source for understanding game theory. In the next five years Vorob'ev made an attempt to develop the theory of coalitional games, that is, games in which players belonging to one coalition are acting as one player, and therefore mixed strategies have to be defined as correlated families of measures. To prove the existence of stable outcomes in such games, he solved some non-standard problems from combinatorial topology and probability theory, thus combining ideas and methods from various branches of mathematics. At that time he also made interesting generalizations of H. Kuhn's equivalence theorem about behaviour strategies in extensive games with perfect recall, proposed an algorithm on enumerating equilibrium points in bimatrix games and studied games with forbidden situations. These results constituted the basis of his Doctor of Science thesis, which he defended in 1961. In the same year he organized the Soviet Union's first laboratory for game theory and operations research at the Steklov Mathematical Institute of the Academy of Sciences. Under his supervision more than 30 students obtained candidate and doctoral degrees in game theory. In 1968 Vorob'ev organized the first all-Union game theory conference in Erevan (Armenia) and the second in 1971 in Vilnius (Lithuania). He was the main speaker at both conferences, which each attracted more than 100 participants. His addresses focused on methodological and philosophical aspects of game theory as well as areas of applications. He forecast the application of game theory in economics, military affairs, biology, law, ethics, sociology, medicine and literature.

In 1975 his laboratory moved to the Institute for Socio-Economic Problems. Unfortunately, the administration of the institute considered any application of mathematical methods in social sciences as inconsistent with prevailing Marxist-Leninist dogmas. Game theory was no exception, which was why the laboratory was forced to concentrate on mathematical problems arising in game theory. Vorob'ev wrote an interesting monograph *Foundations of Game Theory: Noncooperative Games* (published in English translation in 1994) and considered it the first volume in a planned series of books on game theory. The second volume, 'Cooperative Games' was not completed. He also wanted to write a volume titled 'Dynamic Games'.

Vorob'ev was a brilliant lecturer. He taught part-time at the Leningrad State University and many other universities in Russia and elsewhere, delivering courses in game theory, algebra, probability theory and number theory. He wrote many textbooks, the most popular among which is *Game Theory for Economists and System Scientists* (published in English translation in 1977). He edited most of the translations of the principal Western scientific monographs into Russian, including the famous *Theory of Games and Economic Behavior* by J. von Neumann and O. Morgenstern (1944). He also edited two bibliographic indices on game theory literature up to 1974. They contain about 5,000 summaries of game-theory books and papers from all over the world.

See Also

- ▶ Aumann, Robert J. (Born 1930)
- ▶ coalitions
- ▶ game theory
- ▶ Harsanyi, John C. (1920–2000)
- ▶ Morgenstern, Oskar (1902–1977)
- ▶ Nash, John Forbes (Born 1928)
- ▶ von Neumann, John (1903–1957)

Selected Works

1955. Controlled processes and game theory. *Viestnik of Leningrad University* 4(11):49.

1958. Equilibrium points in bi-matrix games. *Theory of Probability and its Applications* 3:297–309.

1959. Finite non-cooperative games. *Russian Mathematical Surveys* 14(4):21–56.

1960. About partitioned strategies. *Probability Theory and Applications* 5:457–9.

1970. The present state of the theory of games. *Russian Mathematical Surveys* 25(2):77–36.

1976. *Game theory: Bibliographic index*. Leningrad: Nauka.

1977. *Game theory for economists and system scientists*. New York: Springer.

1980. *Game theory: Bibliographic index*. Leningrad: Nauka.

1994. *Foundations of game theory: Noncooperative games*. Boston: Birkhauser.

Bibliography

von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.

Voting

Nicholas R. Miller

Virtually all economic doctrines prescribe that certain activities – for example, the provision of public goods – be undertaken by government. Accordingly, such doctrines implicitly prescribe that certain allocative decisions – for example, determining the level of supply of public goods – be made by political rather than market processes. Thus voting (and government decision making generally), though logically a part of political science, is of clear relevance to economic theory.

Historically, economists have contributed at least as much as political scientists to the pure theory of voting. The theory of voting has its origins in the work of such enlightenment philosophers and mathematicians as Borda, Condorcet and Laplace. Little further progress was made until some forty years ago when the economist Duncan Black wrote a series of articles (most notably Black 1948) on the logic of committees and elections, which were subsequently consolidated into a book (Black 1958). Since Black revived the subject, a number of economists and political scientists have made important contributions. Indeed, the theory of voting has to some extent been subsumed by the more recent and abstract theory of social choice, which was virtually invented by the economist Kenneth Arrow (1951).

Here we review the generic voting problem of selecting, on the basis of the declared preferences of several individuals, one alternative out of a set of alternatives. The voting body may be a small committee, a legislature, or a mass electorate. The alternatives may be proposed budgets, programmes, policies, or candidates for some single office – the common problem is that several alternatives are available from which exactly one must be chosen. (We exclude, therefore, the somewhat different problem of voting for representative bodies, to which several candidates may be elected simultaneously.)

The simplest voting problem is that in which there are just two alternatives, one of which is to be chosen. In this case, voting by simple majority rule strikes most people as fair and reasonable. Each voter votes for one or other alternative (or abstains), and whichever alternative receives more votes is selected. May (1952) formalized our intuition concerning majority rule: he identified four conditions that we probably want a voting rule to meet in a two-alternative contest, and he demonstrated that majority rule, and only majority rule, meets these conditions. May's conditions are: *decisiveness* – however people vote, there is always a clear outcome (even if that is 'social indifference', i.e., a tie); *anonymity* (of voters) – we do not need to know who cast which votes to determine the outcome; *neutrality* (between alternatives) – if everyone voted in the opposite fashion (or continued to abstain), the other alternative would win (or, if the outcome were initially a tie, it would remain a tie); and *positive responsiveness* – if alternative A at least ties B and someone then changes his vote to make it more favourable to A (i.e., by voting for A instead of abstaining or voting for B, or by abstaining instead of voting for B), A then wins. May demonstrated that majority rule meets these four conditions and is the only decision rule that does so. (Decision rules distinct from majority rule can meet any three of them.)

In sum, voting based on majority rule to choose between two alternatives is essentially straightforward, though objections can still be raised against it. One common objection is that, on any particular decision, the winning majority may be, in some sense, 'wrong' or misinformed. Another objection, stated in terms of political theory, is that an 'apathetic' majority (with only weak preferences for alternative A) may override an 'intense' minority (with strong preferences for alternative B); in economic terms, there is no assurance – supposing that some interpersonal accounting of costs and benefits is possible – that selection of A provides to the group as a whole greater benefits net of costs than selection of B. Finally, it may be remarked that, in some circumstances, one or more of May's conditions – and thus also majority rule itself – may not seem so fair and reasonable; an example may be provided if alternative A represents a fundamental change in constitutional arrangements and B represents maintenance of the constitutional status quo, in which case it may well seem appropriate to treat the alternatives in a non-neutral fashion (by, for example, requiring greater than majority support for the selection of A).

But more vexing problems arise when the domain of choice is expanded to three or more alternatives. Many different apparently fair and reasonable voting procedures are possible (and in actual use), all of which reduce to simple majority rule in the event there are just two alternatives, but which operate differently in the event there are three or more alternatives. It is not clear which, if any, of these procedures is the 'natural' or appropriate extension of simple majority rule. On closer inspection, they all have serious flaws – that is, they turn out not to be so fair and reasonable; indeed such flaws appear to be unavoidable in the general case.

With three or more alternatives, a procedure may require voters to declare their preferences either 'nominally' or 'ordinally' – a distinction that collapses when just two alternatives are being voted on. Under a *nominal* procedure, each voter divides the alternatives into two sets – those he votes for and (implicitly) those he votes against. Under an *ordinal* procedure, each voter ranks orders the alternatives according to his preferences. (There are other ballot forms, but they are rarely used in practice.)

For descriptive purposes, we may assign commonly used voting procedures that select one alternative out of many to three broad types (for a more extended recent discussion see Dummett 1984), which we may label *aggregation* procedures, *elimination* procedures, and *sequential binary* procedures. To simplify the following discussion, we sidestep the question of how procedures may break ties and we suppose voters are never indifferent between alternatives.

An aggregation procedure takes declared preferences and aggregates them in a single step to determine the selected alternative; thus only one vote is taken. The simplest voting procedure is *plurality* (or 'first-past-the-post') voting: on a nominal ballot, each voter votes for no more than one alternative; the aggregation rule selects the alternative with the most votes. A recently proposed variant is *approval* voting (Brams and Fishburn 1983): on a nominal ballot, each voter votes for any number of alternatives; the aggregation rule is the same as plurality. The most common aggregation procedure using an ordinal ballot is *preferential* (or *Borda count*) procedure. The aggregation rule is this: if there are m alternatives altogether, an alternative is awarded m points for each ballot on which it is ranked first, $m - 1$ points for each on which it is ranked second, and so forth; the alternative with the most points is selected.

An elimination procedure initially aggregates declared preferences in some fashion, on the basis of which weaker alternatives are eliminated. A new vote is then taken on the remaining alternatives. (If an ordinal ballot was used at the outset, the original ballots can be reaggregated with the eliminated alternatives deleted from each ranking.) Elimination and revoting (or reaggregation) continue until every alternative but one has been eliminated. *Plurality plus runoff* voting initially aggregates in the manner of plurality voting, eliminates all alternatives except those receiving the most and second most votes, and holds a simple majority vote runoff between these two. The *alternative vote* procedure also aggregates in the manner of plurality voting, but only the alternative with the fewest number of votes is eliminated at each stage; thus $m - 1$ votes are required altogether. The *exhaustive* (or *Coombs*) procedure uses an ordinal ballot and eliminates from among the remaining alternatives the one with the most last-place, rather than the fewest first-place, preferences. Still other elimination procedures aggregate in the manner of preferential voting.

A sequential binary procedure is a voting procedure of the parliamentary type, in which a sequence of binary choices (e.g., yes or no) is put to the voters. A very simple sequential binary procedure – which approximately (but not exactly) mimics Anglo-American parliamentary voting – is what Black called *ordinary committee* procedure and is now generally referred to as *standard amendment* procedure: two alternatives are paired for a simple majority vote, the winner is paired with a third alternative for a second vote, and so forth until every alternative has entered the voting. The alternative that wins the final vote is selected. Another sequential procedure is variously referred to as *sequential elimination* or *successive* procedure: each alternative in turn is voted up or down on a simple majority vote; the first alternative to receive majority support is selected; if every alternative but one has been rejected, the one remaining alternative is selected by default.

Under any sequential procedure, the alternatives must be placed in some kind of voting order; this raises the possibility that such procedures may violate the spirit of May's neutrality condition, in that whether an alternative is selected may depend on when it enters the voting.

The reader may easily check that each procedure described above reduces to simple majority rule in the event that there are just two alternatives. Moreover, at first blush, they all appear to be fair and reasonable – in any case, certainly not perverse. Thus each procedure appears to be a natural extension of simple majority rule when the domain of choice is expanded beyond two alternatives. However, the reader may also check that, for given declarations of preferences by voters, each procedure may imply a different selected alternative. By way of partial illustration, consider the following declaration of preferences over four alternatives (the number above each ordering indicates the number of voters declaring such preferences):

Example 1	4	4	2	9
First preference	A	B	B	C
Second preference	B	A	D	D
Third preference	D	D	A	A
Fourth preference	C	C	C	B

Under plurality voting, C is selected (with 9 votes, as opposed to 6 for B, 4 for A, and none for D). Under approval voting, if we suppose that each voter votes for his top two alternatives, D wins (with 11 votes, as opposed to 10 for B, 9 for C, and 8 for A). Under plurality plus runoff voting, B is selected (the runoff is between B and C and the four voters whose first preference A has been eliminated prefer B to C). The alternative vote, in this case, works in just the same way as plurality plus runoff. Exhaustive voting selects D (C, with 10 last-place preferences, is eliminated first, then B with the 9 last-place preferences, and then A). Preferential voting selects A (with 50 points, as opposed to D with 49 points, C with 46 points, and B with 45 points). With respect to sequential binary procedures, voting under both amendment and successive procedures voting can select any alternative other than C (which loses every possible pairwise vote), depending on the voting order (specifically, the alternative other than C that enters the voting last is selected).

In choosing among competing voting procedures, an appealing approach is to do what May did for simple majority rule – that is, identify a set of attractive criteria and then determine which procedure uniquely meets them. (See, for example, Young 1974.) The problem here is that different procedures meet different sets of criteria, and no procedure meets all criteria that we might regard as necessary for a fair and reasonable system to meet. (In effect, voting theory runs up against Arrow's 'general impossibility theorem' in social choice theory; cf. Arrow 1951.)

A particularly severe flaw that affects all these voting procedures is that they are subject to *agenda manipulation* – that is, individuals who can add alternatives to, or delete alternatives from, the agenda of choice can influence the outcome *even if the alternatives that may be added or deleted cannot themselves win*. (It is this property of plurality voting that makes the presence or absence of 'third-party' candidates, who cannot themselves win, so significant in British parliamentary elections or US Presidential elections.) Consider Example 1 again. If all four alternatives are on the agenda, C is selected under plurality voting, but if A is removed from the agenda (and thus deleted from each preference ordering), B is selected. (This is why B wins under plurality plus runoff voting. More generally, it is only because the elimination of alternatives can alter the relative strength of surviving alternatives under aggregation procedures that there is any reason to devise elimination versions of these procedures.) Similar illustrations could be provided for other procedures. Thus voting under such procedures violates the Weak Axiom of Revealed Preference – which economists usually take to be an aspect of rational choice – and indeed weaker consistency criteria as well.

Let us now consider one apparently attractive approach to extending simple majority rule to the multi-alternative case that none of the procedures described above exactly implements (for good reason, it turns out). Let us consider the *majority preference relation* – that is, simple majority rule between all pairs of alternatives. Consider the following declaration of preferences by five voters (we will discuss the 'social ordering' momentarily):

Example 2	2	1	2	Social ordering
First preference	A	B	C	B
Second preference	B	A	B	A
Third preference	C	C	A	C

We may note that B, though it has the fewest first preferences and would lose under many procedures, has a particular strength and perhaps a strong claim to be the alternative that should be selected. This is due to the fact that B can defeat each other alternative in a pairwise vote (or 'straight fight') under simple majority rule. An alternative that can do this is called the *Condorcet winner*, and the criterion for voting procedures which requires that the Condorcet winner be the selected alternative is called the *Condorcet criterion*. Every procedure described above, other than the standard amendment procedure, violates this criterion – that is, we can find some declaration of preferences such that the procedure selects an alternative other than the Condorcet winner.

The approach of looking at pairwise contests based on majority rule apparently has this further attraction: for the example above, we can identify not only the Condorcet winner but a 'social ordering' based on majority rule, as shown above – that is, A is majority preferred to both B and C, and B is majority preferred to C. Given such a 'social ordering', if it turned out that A was in fact not a feasible alternative, the group could simply move to B as its second 'social preference'. The majority preference relation, moreover, is quite immune to agenda manipulation, as majority preference between two alternatives depends only on individual preferences between the same two alternatives and is unaffected by the presence or absence of other alternatives or by changes in individual preferences among alternatives other than the two in question.

The majority preference relation has further descriptive significance. Most electoral and legislative voting rules are *majoritarian* in nature – that is, they empower any majority of voters acting in concert to select whatever alternative that majority agrees upon. Thus to say A is majority preferred to B is equivalent to saying, in the language of cooperative game theory, that A *dominates* B, i.e., that there is a coalition of individuals who all prefer A to B and who collectively have the power to bring about A. The Condorcet winner is, therefore, the *undominated* or *core* alternative. Thus, if voters treat voting as a game of strategy in which coalitions can form freely, the outcome will be determined by the majority preference relation, independent of the particular (majoritarian) voting procedure nominally in use.

It may appear, therefore, that we have satisfactorily solved the problem of generalizing majority rule to the multi-alternative case, but unfortunately we have not. The reason is that majority preference (like game-theoretic domination) does not in general generate a 'social ordering'. This is illustrated by Example 1, in which it may be checked that, in pairwise votes, A defeats B, B defeats D, and D defeats A. (It was for this reason that the selected alternative under amendment procedure depended on the order of voting.) This phenomenon is variously called the 'paradox of voting', the 'Condorcet effect', the 'Arrow problem' and the phenomenon of 'cyclical majorities'. It is most simply illustrated by the following three voter, three alternative example.

Example 3	1	1	1
First preference	A	B	C
Second preference	B	C	A
Third preference	C	A	B

This phenomenon evidently was first discovered by Condorcet, and it was then alternately forgotten and rediscovered until the work of Black and Arrow appeared in the late 1940s. It results from some declarations of preferences (e.g., Example 3) but not others (e.g., Example 2). The question naturally occurs of whether we can specify general conditions on preference declarations under which the paradox of voting does, and does not, occur.

The most obvious condition that excludes the paradox is *majority consensus*, i.e., a majority of voters declare the same preferences; but we may note that this does not explain the absence of paradox in Example 2. What is true in Example 2 is that the declared preferences are – to use the term introduced by Black (1948) – *single-peaked*. (See Sen 1966, for generalization of this concept.) What this means is that the declared preferences are consistent with the supposition that the alternatives are perceived by all voters as arrayed along a single dimension of evaluation. For example, three alternatives might be arrayed along an ideological dimension such that one is the (relatively) 'leftist' alternative, another is the (relatively) 'rightist' alternative, and the third is the 'centrist' alternative that represents a compromise between the other two. If all voters structure their preferences accordingly, it follows that there is some alternative – namely, the centrist one – that no voter ranks last. Then in turn it follows that either an absolute majority of voters prefers one or other extreme alternative or the centrist alternative defeats each extreme alternative in a pairwise majority vote (since the voters who most prefer one extreme alternative prefer the centrist alternative to the other extreme); in any event there is a Condorcet winner. It may be checked that the declared preferences in Example 2 meet the single-peakedness condition (with B the alternative that no one ranks last), while the preferences in Example 3 do not.

The notion of single-peaked preferences extends readily to a continuum of alternatives. Each voter has an *ideal point* of highest preference or maximum utility somewhere along the continuum and his utility declines as distance from his ideal point increases in either direction.

Whether alternatives are discrete points along a dimension or a continuum of points, if preferences are single-peaked voter ideal points can be rank ordered from left to right (or whatever is the nature of the evaluative dimension). It then follows that the alternative M corresponding to the median of voter ideal points is the Condorcet winner. This is the *median voter theorem* due originally to Black (1948, 1958). Consider any point A to the left of M. M is preferred to A by the median voter and all voters whose ideal points lie to the right of M and, by definition of a median point, this is a majority of the voters. Obviously the same argument can be made for any point B to the right of M. Thus M defeats every other point and is the Condorcet winner.

The notion of single-peaked preferences can be generalized to a multidimensional space of alternatives, where each point in the space represents a different combination of policies, programmes, appropriations, points on distinct evaluative dimensions, or whatever. Generalized to this setting, the notion requires that all voter preferences with respect to sets of alternatives lying on any straight line through the space be single-peaked. This is equivalent to the standard economic assumption that individual preferences on a space (of, for example, commodity bundles) be convex. But, in the multidimensional case, there almost never is a point that is the median ideal point in all directions, so there is almost never a Condorcet winner, and cyclical majorities almost always exist (Plott 1967). Moreover, it turns out that, in the almost certain event that there is no Condorcet winner, a massive majority cycle encompasses all points in the space (McKelvey 1979). Despite all this, recent work indicates that, even in the multidimensional case, common voting processes, in particular those of a competitive nature, lead to selection of more or less centrist alternatives.

Throughout the discussion thus far, we have consistently sidestepped one further complexity in voting. Voting procedures operate on the *declared preferences* on voters. The question arises of whether it always is expedient for voters to declare (or reveal) their 'honest' or 'sincere' preferences. In fact, it is well known to both students and practitioners of politics that, under common voting procedures, voters who cast 'honest' votes may regret doing so. For example, suppose the preferences displayed in Example 1 are actually the honest preferences of all voters. Under plurality voting, alternative C is selected, if preferences are honestly revealed. But it would be to the advantage of the four voters whose preference ordering appears in the first column to declare their preferences otherwise, specifically by ranking B first, for then B – which they all prefer to C – would be selected. For another example, suppose the preferences displayed in

Example 3 are actually honest preferences. Under standard amendment procedure with the alternatives voted on in alphabetical order, A defeats B in the initial vote and C, which defeats A in the second vote, is ultimately selected. However, if the voter whose preference ordering appears in the first column were to vote insincerely for B instead of A at the first vote, B would be ultimately selected and that voter prefers B to C. In general, if voting is treated as a game of strategy, voting in a manner that reveals true preferences may not be the best strategy.

Several questions then naturally occur. First, if voting is treated as a game of strategy, is it possible to identify 'best' strategies for all voters? If so, and if all voters use their best strategies, is the selected alternative different from what would be selected if all voters used honest strategies? (Note that, in the two examples above, we did not consider possible counter-strategies of the remaining voters.) If the outcomes are different, how do the 'strategic' and 'honest' outcomes compare? Finally, it is possible to design a voting procedure such that best and honest strategies always coincide for all voters – that is, can we devise a 'strategy proof' voting procedure?

The first question was first systematically treated by Farquharson (1969), who introduced the concept of *sophisticated* voting, i.e., voting that is strategically optimal, which is in general different from *sincere* voting, i.e., voting that honestly reveals preferences. Farquharson stated a theorem that says this: if no voters are indifferent between alternatives, sophisticated voting under any sequential binary procedure is determinate, i.e., the game of strategy has a definite solution. However, Farquharson's method for solving such voting games, based on successive elimination of dominated strategies, is cumbersome to employ in even the simplest situation and, for all practical purposes, impossible to employ if there are more than a few voters or alternatives. Fortunately, an alternative definition of sophisticated voting under sequential binary procedures, and an alternative and much easier method of solution, exist. This is the *multi-stage* or *tree* method, which has been definitively characterized by McKelvey and Niemi (1978).

Using this method, sophisticated voting outcomes under binary procedures may easily be identified and compared with sincere outcomes. First, sincere and sophisticated outcomes often diverge – that is, strategic behaviour on the part of all voters does not necessarily 'cancel out'. Second, and perhaps contrary to 'common sense' expectations, sophisticated voting outcomes are, by several criteria, superior to sincere outcomes. (For example, sophisticated voting, but not sincere, always complies with the Condorcet criterion.) Third, if voting is sincere, alternatives are favoured by being placed later in the voting order; if voting is sophisticated, the reverse is true. Finally, these differential effects are magnified to the extent that majority preference is cyclical.

With respect to the final question, voting theorists conjectured for many years that a strategy proof voting procedure could not exist, but two fundamental problems stood in the way of decisively demonstrating this. First, it is not at all clear how to define the class of objects that we might call 'voting procedures'. Thus, no matter how many procedures we can demonstrate to be vulnerable to strategy, there seems always to be the logical possibility that something else exists that we might be willing to call a 'voting procedure' and that is strategy proof. Second, especially with more exotic procedures (e.g., approval voting), it is not always clear what constitutes 'sincere' or 'honest' voting.

Gibbard (1973) neatly sidestepped both of these problems and proved the conjecture. He did this by solving a much more general problem in game theory. First, he said, however we define the set of all voting procedures, it is certainly a subset of all 'game forms', where a *game form* is a game (in the sense of game theory) minus the preferences of players over outcomes. A game form is *dictatorial* if there is some player who, for every outcome of the game, has a strategy that is *decisive* for that outcome, i.e., its selection guarantees that outcome, regardless of the strategy selections of the other players. In a game, a strategy is *dominant* for a player if he would never regret selecting it, regardless of the strategies selected by other players. A game form is *straightforward* if it gives every player, for all possible preferences over outcomes, a dominant strategy. Gibbard then proved (using Arrow's theorem) that every straightforward game form with three or more outcomes is dictatorial.

Now suppose that a voting procedure is strategy proof. Then no voter, regardless of his preferences, can ever have reason to regret voting sincerely, regardless of how other voters vote. But this means that every voter, regardless of his preferences, must always have a dominant strategy (which, moreover, must be a sincere strategy). But, even apart from the requirement that the dominant strategies be sincere, this requires that the voting procedure be a straightforward game form. Thus, once we move beyond choice between just two alternatives, and at the same time make selection depend on the declared preferences of more than one individual, we cannot avoid the possibility that individuals may have an incentive to declare other than their true preferences.

See Also

- ▶ Black, Duncan (1908–1991)
- ▶ Borda, Jean-Charles de (1733–1799)
- ▶ Carroll, Lewis (Charles Lutwidge Dodgson) (1832–1898)
- ▶ Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis De (1743–1794)
- ▶ Public Choice
- ▶ Social Choice

Bibliography

Arrow, K.J. 1951. *Social choice and individual values*, Cowles foundation monograph, vol. 17. New York: Wiley.

Brams, S., and P. Fishburn. 1983. *Approval voting*. Boston: Birkhauser.

Black, D. 1948. On the rationale of group decision-making. *Journal of Political Economy* 56(1): 23–34.

Black, D. 1958. *The theory of committees and elections*. Cambridge: Cambridge University Press.

Dummett, M. 1984. *Voting procedures*. Oxford: Clarendon Press.

Farquharson, R. 1969. *Theory of voting*. New Haven: Yale University Press.

Gibbard, A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41(4): 587–601.

May, K. 1952. A set of independent necessary and sufficient conditions for simple majority rule. *Econometrica* 20(4): 680–684.

McKelvey, R. 1979. General conditions for global intransitivities in formal voting models. *Econometrica* 47(5): 1085–1112.

McKelvey, R., and R. Niemi. 1978. A multistage game representation of sophisticated voting for binary procedures. *Journal of Economic Theory* 18(1): 1–22.

Plott, C. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57(4): 787–806.

Sen, A.K. 1966. A possibility theorem on majority decisions. *Econometrica* 34(2): 491–499.

Young, H.P. 1974. An axiomatization of Borda's rule. *Journal of Economic Theory* 9(1): 43–52.

Voting Paradoxes

Donald G. Saari

Abstract

After using an example to motivate why voting theory is so central to the social sciences, this survey describes some of the more recent (and, surprisingly, benign) interpretations of Arrow's Impossibility Theorem as well as explanations of the wide selection of voting paradoxes that drive this academic area. As described, it now is possible to explain all positional voting paradoxes while creating any number of illustrating examples.

Keywords

Anti-plurality system; Approval voting; Arrow, K; Axiomatic approach to decision rule choice; Borda Count; Condorcet, M; Cumulative voting; Elections; Gibbard–Satterthwaite theorem; Impossibility theorem; Independence of irrelevant alternatives; Kruskal–Wallis test; Luce, D; Plurality vote; Sonnenshein–Mantel–Debreu theorem; Transitivity; Voting paradoxes; Voting rules; Walras's Law

JEL Classifications

D7

Almost daily, news articles describe important elections being held somewhere in the world. The newsworthiness of these events is obvious: election outcomes can change the political, societal and economic directions of a city, a state, or even a country. Elections, in fact, are everywhere; their use ranges from legislative bodies busily determining laws to a kindergarten class selecting a recess treat 'with a show of hands'. As elections are important, we impose safeguards such as the secret ballot. But a strong message coming from voting theory is that the choice of a voting rule can do more to frustrate the 'will of the voters' than any scheming, cigar-smoking political boss.

To illustrate this comment, consider the following three-candidate example where A > B > C means a voter prefers A to B to C. Let four voters prefer A > B > C, three prefer A > C > B, two prefer C > A > B, two prefer C > B > A, and six prefer B > C > A. With the:

- *plurality vote*, or 'vote for one', *A wins* with the A > B > C ranking;
- *Borda Count*, where 2, 1, 0 points are assigned, respectively, to a voter's first, second and third ranked candidate, *B wins* with the B > C > A ranking;
- *anti-plurality*, or 'vote for two', system, which is equivalent to voting against a candidate, *C wins* where its C > B > A ranking happens to reverse the plurality ranking.

Not all candidates reflect the 'will of these voters', yet each 'wins' by selecting an appropriate voting rule. Pairwise majority votes offer no help with their A > B, B > C, C > A cycle. The message is that, rather than capturing the views of the voters, an election outcome may more accurately reflect the *choice of the voting rule*.

More general rules include n-candidate positional methods defined by n weights w_1, w_2, . . ., $w_n = 0$; $w_1 > 0$ and $w_j \geq w_{j+1}$ where w_{j} points are assigned to a voter's jth ranked candidate; candidates are ranked by the sums of assigned points. While (1, 0, 0), (2, 1, 0) and (1, 1, 0) represent the above rules, (8, 3, 0) is still another choice. Different weights, however, may generate other election outcomes. Indeed, the above example allows *seven* different positional election rankings. For instance, the (8, 3, 0) outcome is a fourth strict ranking B > A > C; the three remaining rankings involve ties.

One probable reason for the many different election rules is that inventing new ones is limited only by one's imagination; for example, positional methods define run-off rules whereby, after the bottom-ranked candidates are dropped, the remaining two are reordered. With our example, the plurality, Borda, and anti-plurality run-offs elect, respectively, A, B and B. Other approaches allow each *voter* to select a positional method to tally his ballot. With *cumulative voting*, for instance, a voter splits, say, three points in any integer manner; for example, she may use (3, 0, 0), or (2, 1, 0). *Approval voting* (AV) allows a voter to vote for (approve) any number of candidates; for example, he could select (1, 0, 0) or (1, 1, 0). But, whenever voters can determine how to tally their own ballots, we must anticipate that a single profile (that is, listing of voters' preferences) can admit many different outcomes. Indeed, while changing positional methods generates seven different rankings for our example, *all 13* ways to rank three candidates are admissible cumulative or AV outcomes. Some theorists view this flexibility as a virtue (for example, Brams et al. 1988); others treat this extreme indeterminacy as a serious failing (for example, Saari and Van Newenhizen 1988).

As our example demonstrates, selecting an inappropriate voting or decision rule could inadvertently cause inferior outcomes – with negative concomitant consequences. This is not an isolated phenomenon: with conservative assumptions, about 69 % of contested three-candidate elections allow election rankings to change with different positional methods (Saari and Tataru 1999). The percentage significantly increases with more candidates.

Further underscoring the complexity is Arrow's (1951) seminal impossibility theorem. He first requires voters to have complete (all pairs are ranked), transitive (a voter preferring A > B and B > C prefers A > C) preferences without restrictions, and the societal outcomes to be complete transitive rankings. Then Arrow characterizes all rules satisfying two basic properties. The first (Pareto) is a unanimity condition whereby, if everyone ranks a pair of candidates in the same manner, that is the societal ranking.

To motivate the second, 'independence of irrelevant alternatives' (IIA), condition with a reoccurring phenomenon in the judging of figure-skating, suppose a committee's ranking is Susan > Barb > Jeannie. Imagine Barb's anguish if, told that had more judges liked Jeannie, Barb would have ranked over Susan. Why should the judges' opinion of Jeannie affect the (Susan, Barb) ranking? Arrow's 'independence of irrelevant alternatives' (IIA) condition prohibits this difficulty. Essentially, IIA requires each pair's ranking to depend only on each voter's relative ranking of this pair.

With these minimal conditions, Arrow proves that, for three or more candidates, the only admissible rule is a *dictator* – a specified voter whereby the societal outcome *always* agrees with her preferences independent of what other voters want. Understandably, Arrow's result is often interpreted to mean 'no voting rule is fair'. An alternative, significantly more benign explanation is given below.

The overpowering message is that the choice of a decision rule is crucial. Indeed, determining which rules are 'optimal' is the primary concern of voting theory, where finding axiomatic characterizations of rules, or discovering paradoxical examples, seems to dominate. Another approach (Luce 1959) imposes structure on the outcomes; this structure determines what voting rules are admitted and what restrictions must be imposed on voter choices. A third, recent emphasis examines the data structure – voter preferences – to determine what the voters want and then which voting rules deliver the appropriate outcome (Saari 2000).

For a template, treat a voting rule as a mapping from the domain (space of individual preferences) to the range (space of societal outcomes). The axiomatic approach emphasizes properties of the mapping, Luce's approach emphasizes the structure of the range, and my recent approach emphasizes the structure of the domain. All three approaches are briefly described.

Axiomatic Approach and Paradoxes

Borrowed from mathematics, a standard justification for the 'axiomatic approach' is that 'it tells us what we are getting'. After all, axioms are intended to form the fundamental building blocks of a theory, so axiomatic characterizations should specify what to expect from different voting rules. But this expectation requires the conditions to be true axioms; most often they are not. Instead, many results *uniquely identify* a rule in terms of special, perhaps idiosyncratic, properties rather than characterizing the rule. As an analogy, it is easy to envision settings where certain properties uniquely identify 'John' as a studious, well-behaved student, while different properties uniquely identify 'John' as a street-wise juvenile delinquent. By concentrating on particular traits, both sets of properties uniquely identify John, but neither completely describes nor characterizes him.

Similarly, many so-called 'axiomatic characterizations' of voting rules are, in reality, properties that inadvertently emphasize *special profiles*, so while they uniquely *identify* certain rules, they do not characterize them. As an example, certain technical assumptions plus the condition 'a candidate top-ranked by most voters wins' uniquely *identifies* the plurality vote. Alternatively, the same technical conditions accompanied with the 'with *n*-candidates, a candidate may win even if bottom-ranked by all but one more than $1/n$ of the voters' property also uniquely *identifies* the plurality vote. Neither is an axiomatic characterization: by depending on special profiles, neither really 'tells us what we are getting'.

This literature, however, identifies valued voting rule properties. Another widely used approach with the same objective is to find 'voting paradoxes', that is, unexpected outcomes. Indeed, the origin of this field derives from a 1770 example (published in Borda 1781) that Borda constructed to question the plurality vote: with his example the C > B > A plurality outcome conflicts with the pairwise rankings that are consistent with A > B > C; his (2, 1, 0) Borda Count conclusion agrees with the pairwise rankings.

In contrast, Condorcet (1785) believed we should decide via pairwise comparisons: a Condorcet winner (loser) is the candidate who beats (loses to) all other candidates in majority pairwise votes. To distinguish his approach from Borda's, he constructed an example whereby the Condorcet winner is not top-ranked by the Borda Count – or any positional rule. The controversy over whether Borda's or Condorcet's method is superior continues: comments on this debate are given below.

With examples, Condorcet illustrated that his method can fail; for example, the *Condorcet triplet* A > B > C, B > C > A, C > A > B defines the pairwise cycle A > B, B > C, C > A where neither a Condorcet winner nor loser exists. Later I explain why Condorcet's example remains central to voting theory. Others continuing Condorcet's philosophy explored ways to handle cyclic outcomes; for example, Dodgson's (1876) (Lewis Carroll from 'Alice in Wonderland') method finds the 'closest' Condorcet winner (that is, over all possible lists of pairwise rankings, find the list with a Condorcet winner that is 'closest' to the actual election tallies), while Kemeny's Rule finds the 'closest' transitive ranking. Surprisingly, as Ratliff (2001) proved, the Dodgson winner need not be Kemeny top-ranked; it can be anywhere within the Kemeny ranking. As Ratliff (2003) also proved with examples, if Dodgson's method is extended to select the top two, or top three, candidates, the outcomes need not be consistent; that is, examples exist where the Dodgson winner is not a Dodgson top-two candidate, and none of them is in the Dodgson top three. Voting behaviour is very complex.

'Paradoxes', then, identify new properties of voting rules. Nurmi (1999, 2002), for instance, creates several examples illustrating how major voting rules disagree over a wide selection of desirable properties. His work suggests it may be futile to select voting rules based on specified properties because no rule may satisfy all of them, and most surely there are other valued properties that we have yet to recognize. Fishburn creates many fascinating examples; one (1981) has a plurality ranking of A > B > C > D, but, if D drops out, the same voters have the plurality ranking of C > B > A; Fishburn's example illustrates an unexpected reversal property of the plurality vote.

Examples disclose subtle properties of voting rules, so a way to find all such properties is to find *everything* that can happen: that is, a profile defines a list – an election ranking for each possible subset of candidates. The goal is to find *all* lists that can be created with all possible choices of positional rules and all possible profiles. Call this collection of lists a 'dictionary'. Entries in a dictionary, then, describe all possible ranking properties for all positional rules and even for methods, such as AV and run-offs, based on positional and pairwise rules. Even entries outside the dictionary describe properties; for example, lists of the (A > B > C, B > A, C > A, C > B) type, where some profile allows the pairwise rankings to reverse the positional ranking, never are in the Borda Dictionary, so, by being a missing listing, it describes a Borda consistency property.

Such dictionaries exist (for example, Saari 1989; Saari and Merlin 2000) showing, for instance, that most positional rules allow *anything* to happen. For instance, rank seven candidates in any desired manner. Next, re-rank the seven six-candidate subsets (created by dropping someone) in any desired manner; for example, if you wish, reverse the original ranking, or select them randomly. Continue doing so with each subset of five, four, three and two candidates. While the choices could be chaotic, a profile exists where the voters' plurality ranking for each subset is the selected one. (The same conclusion holds for most choices of positional rules over the different subsets.) What provides hope from these dictionaries is that the Borda Count – defined by $(n-1, n-2, \ldots, 1, 0)$ – is the unique rule (when used with every subset of candidates) that significantly minimizes the number and kinds of allowed paradoxes. Thus, the Borda Count enjoys the maximum number of positive properties; for example, only Borda always ranks a Condorcet winner over a Condorcet loser.

A related 'dictionary' result (Saari 1992a) proves that a ten-candidate profile exists where 9(9!) (recall, $9! = (9)(8)(7) \ldots (2)(1)$, so 9(9!) is over three million) different election rankings without ties result from changing the positional method; each candidate is top ranked with some rules and bottom ranked with others. (For n-candidates, up to $(n-1)[(n-1)!]$ different strict election rankings can emerge from changes in positional methods.)

Luce's Approach

Arrow (1951) proved that with three or more candidates no voting rule satisfying his conditions always has transitive outcomes. Luce (1959) adopted a different approach; he imposed constraints on admissible election outcomes. His conditions, which are described in terms of probabilities to reflect his interest in individual decisions, are stricter than Arrow's. Expressed in terms of voting, Luce requires a candidate's vote percentage to remain consistent over all subsets of candidates. For instance, if A, B, and C receive, respectively, 1/3, 1/2, and 1/6 of the vote, then in a pairwise comparison B beats A by receiving $(1/2)/[(1/3)+(1/2)=3/5$ of the vote. Luce's conditions, then, capture settings where a candidate's support is intrinsic; relative to other candidates, the support remains fixed over all sets of candidates even should new ones join.

The accompanying voting rule and admissible profiles are not specified; they are selected to be consistent with Luce's conditions. But, even with his strong conditions, the accompanying profile restriction with the plurality vote is surprisingly relaxed. Only limited extensions of this approach have been explored for voting theory, but more is possible for settings where candidates have intrinsic support.

Emphasizing the Data

So far I have sampled ways to analyse voting rules through properties of the rules and by imposing restrictions on admissible election outcomes. It remains to explore how the domain structure – the individual preferences – sheds light on these rules. The approach mimics how we might determine whether an election outcome reflects the 'will of the voters': one way is to compare the outcome with what the voters say they want. To develop methodology, reverse the order: first determine what the voters want, and then determine which voting rules respect these outcomes.

To indicate how to determine what the voters want, consider tallying an Alice > Barb '22:20' election outcome. One tallying approach combines an Alice and a Barb vote – a tie. After counting the 20 ties, Alice breaks the tie as she has two extra supporters. For more candidates, the approach is to determine configurations of preferences that arguably constitute ties. This provides a filter; if a voting rule fails to deliver a tie, expect it to introduce a bias in election outcomes. While this is the motivation, the technical objective is to find a coordinate system for the space of profiles. Different coordinates represent how portions of profiles influence different voting rules.

Such a coordinate system for profiles has been established for any number of candidates (Saari 1999, 2000). For intuition about how this is done and the kinds of available results, the three-alternative setting (Saari 1999) is outlined. The space of profiles is divided into three distinct coordinates, or subspaces, capturing:

- profiles that cause *all possible* positional method problems, but with no effect on pairwise rankings;
- profiles that cause *all* problems with pairwise majority votes, but with no effect on positional rankings; and
- profiles where no problems arise with any positional or majority vote rule.

The power of such a coordinate decomposition is apparent. As a sample:

- The coordinates allow us to explain properties of election rules. For instance, positional rules failing to have a tie for the first class of profiles can seriously disagree with pairwise majority vote outcomes.

- The second class of profiles explains problems dating to the 1780s about conflicts between pairwise and positional methods as well as agendas, tournaments and so forth.
- Conflicts associated with any profile, such as our initial one, can be explained; for example, finding the portions of a profile in each of these directions identifies why different rules have different election outcomes.
- Examples illustrating any possible paradox can be constructed. Start with a profile in the last class where there is complete agreement among all rules. To introduce a conflict with positional methods, add a profile portion from the first class; to create conflict with pairwise outcomes, add a profile portion from the second class.

To determine the first coordinate direction, we must find all profiles affecting only positional outcomes. While this is done mathematically, for an intuitive explanation combine a ranking with its reversal, for example, (A > B > C, C > B > A): it is arguable that the outcome should be a tie. It is a tie for majority votes over pairs. But with positional rules (w_1, w_2, 0), the A:B:C tallies are w_1:$2w_2$:w_1. where a tie occurs if and only if (iff) $w_1 = 2w_2$; that is, the desired tie occurs iff the Borda Count is used. If this configuration is used as a filter, then beware of a non-Borda rule. This is because, instead of a tie, rules with $w_1 > 2w_2$ (for example, the plurality vote) have an A = C > B outcome, while rules with $w_1 < 2w_2$ (for example, the anti-plurality vote) have a B > A = C outcome. Consequently, profiles exist where non-Borda positional rankings must differ from majority vote outcomes.

Surprisingly, *all possible differences* among three-candidate positional election rankings reflect how different rules handle these *reversal* profile components. Indeed, to create the initial example, I started with one voter with the B > C > A preference. To generate differences in positional outcomes, add x reversal units of (A > B > C, C > B > A) and y of (A > C > B, B > C > A). As the plurality and anti-plurality tallies for A:B:C are, respectively, $x + y$:y:x and $x + y$:$2x + y$:$x + 2y$, algebra yields my $x = 2$, $y = 3$ choices creating the desired positional outcomes – and conflicts. (Borda is not affected by reversal terms, so its ranking remains the starting B > C > A.) As all possible positional differences are generated by reversal terms, any justification for one positional rule (for example, properties that uniquely identify one rule over others) must reduce to analysing the reversal component (A > B > C, C > B > A) tally.

The second coordinate direction, capturing all conflict among pairwise majority votes, is the Condorcet triplet with its resulting cycle. This component is responsible for all pairwise voting mysteries, including the majority vote cycles, differences in Dodgson's and Kemeny's methods, problems with agendas, tournaments and so forth. This assertion holds for any number of candidates. To create a Condorcet *n*-tuple, start with an *n*-candidate ranking, say A > B > C > D > E. For the next ranking, place the top candidate on the bottom, creating B > C > D > E > A. Continue until each candidate is in first, second, . . ., last place precisely once. This configuration should define a tie, and it does for all positional methods. But the profile also creates majority vote cycles. Surprisingly, these profile coordinate components cause all possible pairwise problems.

To illustrate with our initial example, start with the B > C > A preference. Adding *z* units of (A > B > C, B > C > A, C > B > A) results in A:B, B:C, C:A pairwise votes of, respectively, $2z$:$1 + z$, $2z + 1$:z, $2z$:1:z. So $z = 2$ creates the desired cycle. Adding these reversal and Condorcet terms to the starting ranking yields the initial example.

The remaining coordinate directions, where nothing goes wrong, are called *Basic* directions. For candidate A, it consists of two preferring A > B > C, two preferring A > C > B, one preferring B > A > C, one preferring C > A > B; that is, two for each ranking where A is top-ranked, one for each where A is second-ranked. More generally with *n*-candidates, candidate X's Basic direction has (n–j) voters with each ranking where X is *j*th ranked. While not intuitive, these coordinate directions come from mathematics. The important point is that no conflict occurs in this profile space; for example, the tallies for *any* voting rule for all candidates identifies the tally for *all* voting rules over any subset of candidates. Nothing goes wrong. These three kinds of directions span the six dimensions of profile space, so they complete the three-alternative analysis. (A profile, of course, normally has only parts in each direction.)

Explaining All Differences

All possible differences among three-candidate standard voting rules, then, reflect how voting rules react to reversal and Condorcet profile components. The many desirable properties of the Borda Count, for instance, arise because it is the only rule based on positional and majority votes that always delivers a tie for these components.

I indicated how all positional differences reflect how positional rules treat reversal terms, so it remains to describe the Condorcet components. For motivation, suppose three voters must vote for one of two candidates from each of three schools. Suppose the candidates are [Anne, Bob], [Connie, Dave], [Ellen, Fred]. Does a [Bob, Dave, Fred] outcome, each by 2:1, reflect the voters' views? To answer this question without knowing the actual preferences, all supporting preferences must be listed.

Four of the five profiles have two voters selecting different candidates from each school; this causes a tie. Breaking the tie is the last voter's [Bob, Dave, Fred] preference. The fifth profile has the preferences [Anne, Dave, Fred], [Bob, Connie, Fred], [Bob, Dave, Ellen]. It is difficult to argue against the outcome for the first four profiles as a tie is broken. At least statistically, then, the outcome respects most supporting profiles. But it is difficult to justify the fifth 'outlier' profile other than pointing to the 2:1 votes.

While most profiles justify the conclusion, suppose the fifth 'outlier' profile is the actual one where each voter wanted to elect a woman and a man. The profile reflects their wishes; the outcome does not. The reason is clear: the majority vote strictly emphasizes information about specific pairs; it ignores information – even intended relationships – among pairs. Consequently, rather than recognizing the added 'balanced gender' condition, the majority vote must ignore it.

To connect this example with the Condorcet triplet, identify Anne = B > A, Bob = A > B; Connie = C > B, Dave = B > C; Ellen = A > C, Fred = C > A: the Condorcet triplet becomes the outlier 'fifth profile', and the 'balanced gender condition' is equivalent to 'transitivity'. Because any argument applied to one setting transfers to the other, it follows that the cyclic outcome for the Condorcet triplet (the 'paradox of voting') occurs because (*a*) this outcome reflects most supporting profiles (even though, by involving cyclic preferences, they are not admitted), and (*b*) the majority vote strips all connecting information, *including transitivity*, from the profile. (*c*) While majority pairwise voting may suffice if candidates have 'intrinsic support', it can distort outcomes for usual cases.

In general:

- Pairwise outcomes reflect the average over *all possible* supporting profiles; paradoxes, such as with the Condorcet triplet, indicate that the actual profile is an outlier relative to the average.
- Majority votes strip away all intended relationships, including transitivity, from the profile.
- Whenever intended relations are dropped, they come from profile portions based on Condorcet *n*-tuples.

Explaining Mysteries

The above structure explains several mysteries. The ones described here compare the Borda and Condorcet rules, briefly discuss all rules based on pairwise outcomes, and explain Arrow's Impossibility Theorem.

As indicated, for any number of candidates all possible differences between the Borda and pairwise rankings manifest the majority vote's reaction to Condorcet *n*-tuples, which introduce cyclic affects. As an illustrating example, with two preferring A > B > C, and one preferring B > A > C, both the Borda and pairwise rankings reflect A > B > C. Adding x units of the Condorcet [B > A > C, A > C > B, C > B > A] never affects the Borda ranking, but its cyclic effect changes the A:B, B:C, C:A pairwise tallies to $2 + x{:}2x + 1$, $x + 3{:}2x$, $2{:}2x + 3$ where $x = 2$ makes B the Condorcet winner, $x \geq 4$ creates a cycle.

Any difference between the Borda and Condorcet winners, then, reflects Condorcet profile components. Thus, any argument supporting Condorcet over Borda must justify something other than a tie for a Condorcet triplet or n-tuple.

Voting rules relying on majority vote pairwise rankings, such as Kemeny's and Dodgson's rules, inherit the majority vote difficulties caused by Condorcet n-tuples. As these rules are primarily intended to handle cyclic behaviour, their value presumably emerges when the Condorcet component is dominant. But the stripping action of the majority vote over these components means that, unexpectedly, the rule cannot use information about the voters' transitive preferences. Consequently, if the transitivity of voter preferences is valued, such rules should not be used. If transitivity is not valued, we must question using rules that impose transitivity on the outcomes.

A similar analysis holds for Arrow's Theorem (Saari 2001). An unexpected feature of IIA, as with the majority vote, is to strip from the decision rule all information that individuals have transitive preferences. But, if the rule cannot use the transitivity of individual preferences, then transitive societal outcomes cannot be expected unless profiles are severely restricted; that is, the societal outcome reflects the imposed data structure rather than properties of the rule. One severe restriction is to use the preferences of a single voter; this explains Arrow's dictator.

As Arrow's negative result is strictly caused by IIA unintentionally stripping away valued information about individual preferences, resolutions must modify IIA to allow the rule to use this information. To illustrate, a transitive ranking, say $A > B > C$, separates some alternatives from others. Listing these separations as $[A > B, 0]$, $[B > C, 0]$, $[A > C, 1]$ provides information about the transitive individual preferences. Let IIIA (*Intensity IIA*) be where a pair's societal ranking is determined by how each voter ranks the pair *and the number of separating alternatives*. By replacing IIA with IIIA in Arrow's conditions, Arrow's dictator is replaced with the Borda Count, and rules based on the Borda Count.

Strategic Behaviour

Beyond the above 'single-profile' problems, multiple-profile concerns catalogue interesting changes in outcomes by changing a profile. They include the seminal Gibbard (1973)–Satterthwaite (1975) theorem asserting that, with three or more alternatives, no decision rule is immune from strategic behaviour: that is, with any rule, situations exist where some voter ensures a personally better outcome by voting according to other than her true preferences. There is, in fact, a host of related behaviour; see, for example, Nurmi (1999, 2002). Some rules, for instance, can cause a winning candidate to *lose* by attracting more supporting voters. Similarly, Fishburn and Brams (1983) discovered the 'no-show' paradox where, with the plurality run-off, a voter obtains a personally better outcome by *not* voting.

These results reflect the higher dimensionality of profiles that accompanies added alternatives. With two candidates, a voter can vote for, or against, her favorite. With more alternatives, beyond her top and bottom choice, a voter can consider intermediate options. As suggested by the 'don't waste your vote' cry for strategic voting, situations exist where, by voting strategically, some voters can block personally lower-ranked candidates from winning. The Gibbard–Satterthewaite result proves this happens for all realistic rules.

A common source of problems, such as the no-show paradox, or where two subcommittees elect 'A' but the combined committee does not, and so forth, is when the rule loses monotonicity. Positional methods are monotonic; that is, with added support a candidate has higher tallies. But difficulties occur with rules involving several subsets of candidates; for example, a run-off involves {all n-candidates} and {top two}. What causes problems is that the first election determines who is advanced to the second. Consequently, added support for a winning candidate could also advance a stronger opponent to the run-off.

Implications for Economics

Voting rules are aggregation methods: voters' preference rankings are aggregated into a societal ranking. But as much of economics, and the social sciences, also involves aggregation rules, we must anticipate that the behaviour of voting rules predicts behaviour elsewhere in economics and other disciplines. This happens. As illustrations, the above result allowing 9(9!) different positional election rankings for a single ten-candidate profile, where almost any specified outcome can occur, has a parallel with the Sonnenshein (1972)–Mantel (1972)–Debreu (1974) Theorem asserting that any continuous function satisfying Walras's Laws can be (up to minor technical conditions on prices) the aggregate excess demand function for some exchange economy. As another example, recall the voting result stating that, even if the rankings for the different subsets of candidates are selected in an arbitrary manner, a supporting profile can be found. The same behaviour arises in economics. The voting result allowing a ranking to be selected for each subset of candidates, and a profile can be found so that the selected ranking is the actual election ranking also has an economic parallel: that is, the Sonnenshein–Mantel–Debreu Theorem extends to where a different function can be selected for each subset of commodities, and an economy (initial endowment and utility function for each agent) can be found so that (with the same technical condition) the aggregate excess demand for each subset is the selected one (Saari 1992b).

Voting results have parallels in non-parametric statistics, namely, select rankings for each subset of alternatives: for most non-parametric rules, a data-set can be found so that each set's actual ranking is the selected one. In voting, the positional rule most immune from the 'anything can happen' difficulty is the Borda Count. In nonparametric statistics, the Kruskal–Wallis test has similar properties (Haunsperger 1992).

See Also

- ► Democratic Paradoxes
- ► Paradoxes and Anomalies
- ► Rational Choice and Political Science

Bibliography

Arrow, K. 1951. *Social choice and individual values*, Cowles foundation monograph No. 17. New York: Wiley.

Borda, J. 1781. *Mémoire sur les élections au scrutin*. Paris: Histoire de l'Académie Royale des Sciences.

Brams, S., P. Fishburn, and S. Merrill. 1988. The responsiveness of approval voting: Comments on Saari and Van Newenhizen. *Public Choice* 59: 112–131.

Condorcet, M. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1: 15–23.

Dodgson, C. 1876. A method for taking votes on more than two issues. In *Classics of social choice*, ed. I. McLean and A. Urken. Ann Arbor: University of Michigan Press, 1995.

Fishburn, P. 1981. Inverted orders for monotone scoring rules. *Discrete Applied Mathematics* 3: 27–36.

Fishburn, P., and S. Brams. 1983. Paradoxes of preferential voting. *Mathematics Magazine* 56: 207–214.

Gibbard, A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41: 587–601.

Haunsperger, D. 1992. Dictionaries of paradoxes for statistical tests on k-samples. *Journal American Statistical Association* 87: 149–155.

Luce, D. 1959. *Individual choice behavior*. Mineloa: Dover Publications, 2005.

Mantel, R. 1972. On the characterization of aggregate excess demand. *Journal of Economic Theory* 7: 348–353.

Nurmi, H. 1999. *Voting paradoxes and how to deal with them*. New York: Springer.

Nurmi, H. 2002. *Voting procedures under uncertainty.* New York: Springer.

Ratliff, T. 2001. A comparison of Dodgson's method and Kemeny's rule. *Social Choice & Welfare* 18: 79–89.

Ratliff, T. 2003. Some startling paradoxes when electing committees. *Social Choice & Welfare* 21: 433–454.

Saari, D. 1989. A dictionary for voting paradoxes. *Journal of Economic Theory* 48: 443–475.

Saari, D. 1992a. Millions of election rankings from a single profile. *Social Choice & Welfare* 9: 277–306.

Saari, D. 1992b. The aggregate excess demand function and other aggregation procedures. *Economic Theory* 2: 359–388.

Saari, D. 1999. Explaining all three alternative voting outcomes. *Journal of Economic Theory* 8: 313–355.

Saari, D. 2000. Mathematical structure of voting paradoxes. *Economic Theory* 15: 1–101.

Saari, D. 2001. *Decisions and elections*. New York: Cambridge University Press.

Saari, D., and V. Merlin. 2000. A geometric examination of Kemeny's rule. *Social Choice & Welfare* 17: 403–438.

Saari, D., and M. Tataru. 1999. The likelihood of dubious election outcomes. *Economic Theory* 13: 345–363.

Saari, D., and J. Van Newenhizen. 1988. Is approval voting an 'unmitigated evil?'. *Public Choice* 59: 133–147.
Satterthwaite, M. 1975. Strategyproofness and Arrow's conditions. *Journal of Economic Theory* 10: 187–217.
Sonnenshein, H. 1972. Market excess demand functions. *Econometrica* 40: 649–663.

Voznesensky, Nikolai Alekseevich (1903–1950)

M. C. Kaser

Voznesensky (born the son of a timber dealer in Teploe, Russia, on 18 November 1903; executed on 30 September 1950) joined the Bolshevik Party in 1919 and studied political economy at the Institute of Red Professors, Moscow, where he stayed on as lecturer. His publications – fewer than 30, his culminating manuscript being destroyed by the police – have been analysed by Harrison (1985) and Sutela (1984). In a concept later to be termed 'unbalanced growth' by A. O. Hirschman, he saw that the national plan 'must localize bottlenecks, not for adapting them, but for doing away with them'. Ranging himself against those who argued that comprehensive planning invalidated money calculations, he had by 1935 embraced the position – which was to figure in Stalin's indictment of him in 1949 – that money would have a distributive function even when all means of production had been nationalized. His association with the Leningrad circle which eventually led to his execution also began in 1935, for A.A. Zhdanov, having replaced the assassinated S.M. Kirov as Leningrad Party Secretary, invited Voznesensky to lead that city's plan organization under an Executive Committee headed by A.N. Kosygin.

Voznesensky was promoted to the chairmanship of the USSR State Planning Committee in January 1938 and brought order into the chaos resulting from the 1937 Great Purge (Voznesensky 1938, 1940; Harrison 1985), but so inadequate were his plans for a war economy both before and after the German attack of June 1941 that Zhdanov's rivals, G.M. Malenkov and L.P. Beria (Ra'anan 1983) ran the newly created State Defence Committee, from which Voznesensky was excluded until February 1942. He regained chairmanship of the Planning Committee in December 1942, and achieved in 1943 a peak of armaments production and economic expansion in the unoccupied territory. He allowed market forces to operate in the household sector, alongside rations at controlled prices, absorbing some of the inflation in purchasing power through highly taxed off-ration prices in state shops, and intended to liquidate the inflationary overhang generated by free sales by farmers in a monetary reform as soon as the war ended (though famine caused postponement and retail price restructuring until December 1947).

At the height of Voznesensky's economic leadership (he was elected Academician in 1943) an unsigned editorial, 1943, condemned the 'voluntarism' which disregarded the 'objectivelydetermined process of development' and confirmed, as had been adumbrated in 1941 (Kaser 1965), that a law of value operated under socialism. His postwar Reconstruction Plan evoked 'economic levers in the organization of production and distribution, such as price, money, credit, profit and incentives' (*Selected Works*, 1979, p. 465): he brought in Kosygin as Minister of Finance to oversee the cut in subsidies required by his reform of wholesale prices; the measures which took effect on 1 January 1949 would have been a major contribution to rational economic management (Kaser 1950).

Political realignments led to Voznesensky's dismissal within weeks of his reform and his eventual execution without trial; the life of the dismissed Kosygin, in Khrushchev's later words, 'hung by a thread'. Stalin reversed the reform of both retail and wholesale prices and soon (Stalin 1952) limited the role of 'commodity relations' to the interface of the socialist sector with non-state entities (such as collective farmers and foreigners), vilifying Voznesensky's analysis of the war economy (Voznesensky 1948) for the very 'voluntarism' that the author rejected. The death or disgrace of those in the Leningrad circle was a triumph, albeit short-lived, for Beria and

Malenkov in a political power struggle, but the open disputations were on economic issues: on one, to stop dismantling capital in the Soviet Zone of Germany in favour of current deliveries, Voznesensky had been right; in the others – where E.S. Varga argued that east Europe should be allowed to be 'state capitalist' with market relations with the West and that Keynesian policies had halted the 'general crisis of capitalism' – he had been wrong.

Selected Works

1938. K itogam sotsialisticheskogo vosproizvodstva vo vtoroi piatiletke (On the results of socialist reproduction in the second Five-year Plan). *Bol'shevik* No. 2. In *Selected Works*, 346–362.

1940. Tri stalinskie piatiletki stroitel'stva sotsializma (Three Stalinist Five-year Plans for building socialism). *Bol'shevik* No. 1. Not reproduced in *Selected Works*.

1948. *The war economy of the USSR in the period of the Patriotic War*. Washington, DC: Public Affairs Press and the American Association of Learned Societies. Translation of *Voennaia ekonomika SSSR v period Otechestvennoi voiny*, Moscow: Gospolitizdat. In *Selected Works*, 484–604.

1979. *Izbrannye proizvedeniia 1931–1947* (Selected Works 1931–1947). Moscow: Izdatel'stvo politicheskoy literatury.

Bibliography

Harrison, M. 1985. *Soviet planning in peace and war, 1938–1945*. Cambridge: Cambridge University Press.

Kaser, M.C. 1950. Soviet planning and the price mechanism. *Economic Journal* 60: 81–91.

Kaser, M.C. 1965. Le débat sur la loi de la valeur en URSS. Etude rétrospective 1941–1953. *Annuaire de l'URSS 1965*. Paris: CNRS.

Ra'anan, G.D. 1983. *International policy formation in the USSR: Factional 'Debates' during the Zhdanovshchina*. Hamden: Archon.

Pod znamenem marxizma (Under the banner of Marxism). 1943. No. 7–8. Editorial.

Stalin, J.V. 1952. *Economic problems of socialism in the USSR*. Moscow: Foreign Languages Publishing House. (Translation of *Ekonomicheskie problemy sotsializma v SSSR*, Moscow.)

Sutela, P. 1984. *Socialism, planning and optimality. A study in soviet economic thought*, Commentationes Scientiarum Socialium No. 25. Helsinki: Societas Scientiarum Fennica.

Varga, E.S. 1946. *Izmeneniia v ekonomike kapitalizma v itoge vtoroi mirovoi voiny (Changes in the economy of capitalism as a result of the Second World War)*. Moscow: Gospolitizdat.

Vulgar Economy

Krishna Bharadwaj

Karl Marx used the epithet 'vulgar economy' to describe certain analytical positions which, beginning in classical political economy in the works of Malthus, Say, some of the post-Ricardians including John Stuart Mill, developed eventually into an 'analytical system' (as in Say) and took an 'academic form' (as in the writings of Roscher, among others) (see *Theories of Surplus Value*, Vol. III, pp. 500–502). The epithet was not simply a derogatory label but had thus a specific analytical content and significance. Marx contrasted sharply the 'vulgar' from the classical political economy, the latter comprising of 'all the economists who since the time of W. Petty have investigated the real internal framework of bourgeois relations of production' (*Capital*, Vol. I, pp. 174–5). Vulgar economy, while drawing upon the materials provided by scientific political economy – and therefore lacking in originality – ruminated instead over the 'appearances'. Marx saw, in the capitalist production, 'more than in any other', a 'reality', 'the inner physiology of the system' – which was captured in scientific political economy, in their analysis locating the generation of surplus in production, in their theory explaining the manner in which surplus is appropriated by the owners of the means of production and distributed as the tripartite revenues of rents, profits and wages, and which brought to light the inevitable and endemic conflicts of class interests and thence the contradictions incipient in the processes of generation, distribution and accumulation of surplus. Marx was himself to build his theory on the rudiments provided by political economy. However, this 'reality' hides behind 'appearances' which assume forms and emerge as esoteric concepts and categories of analysis pertaining to the sphere of exchange where 'Freedom, Equality, Property and Bentham' reign supreme; exchange appears as between 'equivalents', governed entirely by competition on the market. Also, the true social relations take fetishistic forms in 'false consciousness', forming the subjectivist perceptions of the participant agents of production. Marx attacked vulgar political economy for remaining at the level of these 'appearances'; since these often reflected perceptions of the bourgeois agents of production, vulgar economy tends to defend, rationalize and therefore to serve the interests of the bourgeois class. While Marx thus recognized, in vulgar political economy, an explicit or implicit ideological function, providing apologetics for the bourgeoisie, his critique was not confined only to the ideological; he painstakingly traced its analytical roots and development and criticized the logical inconsistencies and ambivalences of their theoretical positions.

For Marx, the significant achievement of scientific political economy was in tracing the source of surplus in production and identifying the role of labour as a cause of value and the source of surplus value. It grasped the 'internal interconnections' of capitalist production through recognizing the different roles that the 'agents' – land, capital and labour – played in the process of production and in generating value and the different principles by which their revenues were governed. It identified the constraint binding upon the wage – profit relation. In contrast, vulgar political economy adopted the 'trinity formula' concerning the form and sources of these revenues. Treated as having a symmetric coordinate status, land was seen as the source of rent and capital, of profits just as labour is of wages, it being held that the agents are all paid according to their productivity. Thus land as well as capital is as much a source of value and of surplus as labour. Thus 'we have complete mystification of the capitalist mode of production, the conversion of social relations into relations among things'; to Marx, the entitlement to surplus in the form of rents and profits, originating from the property relations, is here confounded with the creation of surplus by the material means themselves. Further, through giving a symmetric role and status to the trinity, by envisaging their revenues as determined by the same process of competition, and independently of each other, a harmonious view of classes was constructed. This view, explaining distributive revenues in 'doctrinaire language' helped their theory to conform to the bourgeois perceptions: wages appeared as the competitive return to labour and, analogously, as Senior proposed, profits as the recompense for abstinence. The rise in distributive revenues of any one class, reflecting its enhanced productive contribution could not interfere with others' revenues which were determined alike but independently.

Marx sees the roots of the later vulgar economy in certain 'vulgar representations' or 'elements' in classical political economy. While generously praising the masterly vision of Adam Smith for fathoming 'the inner connection' and, for the first time, describing and providing 'a nomenclature and corresponding mental concepts' for 'the external, apparent forms of its life', Marx criticizes, at length, an important 'vulgar' element in Smith: when Smith constructs the natural price of a commodity from adding up wages, rents and profits, determined independently of each other and separately, they become *sources of value* instead of having 'a source *in* value'. After having revealed the intrinsic connection among wages and profits, Smith leaps into 'the connection as it appears in competition'. Marx attaches a great historical significance to Ricardo, 'for science' in that he brought back 'the inner connection – the contradiction between the apparent and the actual movement of the system and brought into the open the objective basis for the inescapable antagonism of class interests'.

This apart, Marx also discusses a number of other shortcomings of classical political economy that provided scope for vulgarization, such as their inadequate recognition of the historical and transient character of the capitalist mode, of the full implications of labour-power becoming a 'commodity' and of capital as a 'social relation' apart from its 'material form'; of the processes of transforming surplus value into profits and of the intervention of money into barter and the evolution of its functions over the advancing stages of capitalist accumulation. All these inadequacies were exploited by vulgar political economy in building up a sanguine and harmonious view of the functioning and growth of the capitalist system, whereas Marx found the system ridden with internal contradictions and recurrent crises.

Marx traced the growth of vulgar political economy and its ascendancy over scientific political economy in terms of the concrete conditions of the historical stages of class struggle. He saw the period between 1820 and 1830 as the last decade of scientific activity when Ricardo's theory was popularized and extended and when 'unprejudiced polemics' was possible. By 1830, the bourgoisie had conquered political power in France and England, their ascendancy over the landed interests was firmly established while the class struggle of labour was assuming threatening proportions. 'It sounded the knell of scientific bourgeois economics. It was thenceforth no longer a question whether this or that theorem was true but whether it was useful to capital or harmful, expedient or inexpedient' (Preface to the second edition, *Capital*, Vol. I).

Vulgar political economy itself passed through analytical stages in the period. Marx notices: 'Only when political economy has reached a certain stage of development and has assumed well-established forms – that is, after Adam Smith – does . . . the vulgar element become a special kind of political economy.' Thus, Say separates the vulgar notions in Smith's work (such as the supply and demand determination of value) and puts them forward as a distinct system. Borrowing from the advancing political economy, vulgar economy also thrives: after Ricardo, particularly, the decline of his theory sets in; the erosion and obfuscation occurring in the hands of his own followers. The hostility to Ricardian theory was sharpened by the use made of labour theory by the utopian writers who, on the basis of their naive interpretation, advocated a radical change in social order. Vulgar political economy becomes increasingly apologetic, as in Bastiat, with the capital-labour confrontation emerging sharply in society, until it assumes a further 'academic form' where apologetics was concealed in an 'insipid erudition' (Marx refers to Roscher as a 'master of this form'!) (1861–3, Vol. III, pp. 500–502.)

What emerges from Marx's detailed critique, particularly in the *Theories of Surplus Value*, is that his attack was not only ideological but also analytical. While a fully-fledged alternative system to replace classical political economy had not yet emerged in Marx's time, the latter had been eroded and conditions become ripe for its subversion.

See Also

▶ Marx, Karl Heinrich (1818–1883)

Bibliography

Marx, K. 1861–3. *Theories of surplus value*, Vols. I–III. London: Lawrence & Wishart, 1972.

Marx, K. 1890. *Capital*, vol. I, 4th ed. London: Pelican Marx Library, 1976.

Marx, K. 1894. *Capital*, vol. III. Moscow: Progress Publishers, 1974.