T

Target Estimation: A New Approach to Parametric Estimation

Luisa Turrin Fernholz
Professor Emerita of Statistics
Temple University, Philadelphia, PA, USA

Introduction and Definition

Target estimation is a computer intensive procedure introduced by Cabrera and Fernholz (1999) that has proved to be effective in reducing the bias as well as the L_1 and L_2 errors of statistics in parametric settings.

For a statistical functional T, let the statistic $T(F_n)$ estimate the parameter $T(F_\theta)$, where F_n is the empirical d.f. corresponding to the sample $X_1, \ldots, X_n$ of i.i.d. random variables. Suppose that all the X_i's have common d.f. F_θ where $\theta \in \Theta$, an open subset of real numbers. If the expectation of $T(F_n)$, $g(\theta) = E_\theta(T(F_n))$, exists for all $\theta \in \Theta$ and is one-to-one and differentiable, then the functional $\widetilde{T}$ induced by T from the relation

$$g^{-1}(T) = \widetilde{T}$$

will be called the *target functional* of T. The statistic $\widetilde{T}(F_n)$ will be called the *target estimator*.

Remarks

a. Note that the target estimate of θ corresponds to choosing the value $\widetilde{\theta} = \widetilde{T}(\widehat{F}_n)$, which solves the equation

$$g(\widetilde{\theta}) = E_{\widetilde{\theta}}(T(F_n)) = T(\widehat{F}_n)$$

where $\widehat{F}_n$ is the observed value of F_n. That is, we set the expectation of a statistic equal to its observed value and we solve for θ. Also, note that g depends on the sample size n which will remain fixed.

b. It is a direct consequence of the definition that if T is a statistical functional with $g(\theta) = a\theta + b$ for $a \neq 0$, then the corresponding target estimator will be unbiased. The variance of $\widetilde{T}$ will satisfy

$$Var(\widetilde{T}) = (1/a^2)Var(T)$$

and the variance of the target estimator will be reduced if and only if $a^2 > 1$.

Properties of Target Estimators

For general estimators, Cabrera and Fernholz (1999) give some results regarding bias and variance reduction after targeting. These results can be summarized as follows:
If $g(\theta) > \theta$ and g is increasing, then:

1. If $1 < g'(\theta) < b$ then $|B_{\widetilde{T}}(\theta)| < |B_T(\theta)|$,
2. If $1 < |g'(\theta)|$ then $MSE(\widetilde{T}) < Var(T)$ and $E|\widetilde{T} - \theta| < E|T - \theta| + |Med(T) - E(T)|$,

where $B_T(\theta)$ and $B_{\widetilde{T}}$ denote the bias of T and $\widetilde{T}$ respectively, $Med(T)$ is the median of T, and MSE is the mean square error.

von Mises Expansions of Target Functionals

The von Mises expansions for the target functional $\widetilde{T}$ can be obtained using the Hadamard or Fréchet derivatives of the functional T. These expansions are useful to analyze the bias of $\widetilde{T}$ as well as the asymptotics and robustness properties of $\widetilde{T}$. For $T(F_n)$ the first order von Mises expansion is: $T(F_n) = \theta + \frac{1}{n}\sum_1^n \varphi(X_i) + Rem$. Then, under some regularity conditions, the remainder satisfies $\sqrt{n}Rem = o_P(1)$, and the statistic $T(F_n)$ is asymptotically normal (see Fernholz 1983). Moreover, when φ is properly normalized, the expectation of $T(F_n)$ gives: $g(\theta) = \theta + E_\theta(Rem)$ so that $T = \widetilde{T} + E_{\widetilde{T}}(Rem)$, and the bias of the target estimator is $B_{\widetilde{T}}(\theta) = E_\theta(Rem_1 - E_{\widetilde{T}}(Rem))$, which under certain conditions satisfies,

$$|B_{\widetilde{T}}(\theta)| = |E_\theta(Rem - E_{\widetilde{T}}(Rem))| < |E_\theta(Rem)| = |B_\theta(T)|.$$

Using the von Mises expansions of T, it can be shown that the ▶asymptotic normality of $\widetilde{T}$ is inherited from the asymptotic normality of T, with some gain in asymptotic efficiency when $|g'(\theta)| > 1$. The robustness aspects of target functionals are also analyzed using the von Mises approach and the influence functions of $\widetilde{T}$ and T are related by:

$$\mathrm{IF}_{\widetilde{T}}(x) = (1/g'(\theta))\mathrm{IF}_T(x).$$

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

This shows that the gross-error sensitivity of the target functional is lower when $|g'(\theta)| > 1$. See Fernholz (1997) and Cabrera and Fernholz (1999).

Target Estimation in Multidimensional Settings

Multivariate target estimation was treated in Cabrera and Fernholz (2004) where p-dimensional statistical functionals $T = (T_1, \ldots, T_p)$ estimate a p-dimensional parameter vector $\theta = (\theta_1, \ldots, \theta_p)$. In this case the expectation function $g(\theta)$, as defined in section "►Introduction and Definition", is p-dimensional, and for the simple case where g is an affine function of the parameter vector, the bias can be removed entirely and, under certain conditions, the variability of the bias corrected functional is reduced in the sense of smaller trace and smaller determinant. Examples of multivariate targeting for location-scale equivariant estimators and the location-scale exponential model are given in Cabrera and Fernholz (2004).

In practice, we seldom have linearity of the p-dimensional expectation function g. Quite often, the p-dimensional estimator T is defined implicitly and the corresponding target estimator must be found by solving multidimensional implicit equations in θ, of the form $g(\theta) = T(F_n)$ where $T(F_n)$ has been observed and $g(\theta) = E_\theta(T(F_n))$ is multidimensional. This amounts to inverting the function g which, if unknown, must first be estimated. The method of *stochastic approximation* introduced by Robbins and Munro (1951) and modified by Cabrera and Hu (2001) was successfully used to find the target estimates in many situations. For details and description of this methods see Cabrera and Fernholz (2004) and Cabrera et al. (2005).

Applications and Examples

Target estimation has been successfully used for bias and variance reduction in many cases. The following are just some of the more important cases developed:

1. *Ellipse estimation*. The case of ellipse estimation when only an arc of data points is available is of particular importance in computer vision since many real life problems encounter this difficulty. A study regarding the least squares estimators of five parameters identifying an ellipse can be found in Cabrera and Fernholz (2004) where a comparison of the target estimators with both the bootstrap (see ►Bootstrap Methods) and the jackknife estimators (see ►Jackknife) shows the advantages of the target estimation method in terms of reducing bias and lowering the variability of the estimators.
2. *Autoregressive Models*. Simulations were performed for autoregressive models AR(1) of the form $X_{t+1} = \theta X_t + \epsilon_t$, where the error term ϵ_t is Gaussian. The maximum likelihood estimator (MLE) of the parameter θ was compared to the corresponding target estimator for different sample sizes and different values of θ. These simulations showed a substantial reduction in the bias of the target estimator as compared to the bias of the MLE for every case considered, and they also showed that the MSE of the target estimator was reduced in most of the cases. See Cabrera and Fernholz (1999).
3. *Errors-in-variables Models*. General errors-in-variables models of the form $Y = a + bU + \epsilon$ when the observable variables are $X = U + \delta$, where ϵ and δ are independent Gaussian errors. In all the simulations performed for different sample sizes and different values of b the bias of the target estimator was substantially reduced when compared to the bias of the MLE, and in all cases the MSE of the target estimator was smaller than that of the MLE. See Cabrera and Fernholz (1999).
4. *Logistic Regression Models*. A treatment of logistic regression models (see ►Logistic Regression) of one and two parameters was given in Cabrera et al. (2005) where it is shown that the transformed MLE, i.e., the target estimator, has lower bias and MSE that the original MLE. It was also shown that another benefit of targeting is that it corrects the asymmetry of the statistic thus producing target statistics with more symmetric distributions.

Final Remarks

1. *Comparison to the Bootstrap*. Target estimation has been compared to other methods of reducing bias and variability such as the jackknife and the bootstrap. This comparison is treated in Cabrera and Fernholz (1999, 2004), where for different situations it was shown that targeting can provide considerable improvement over both the jackknife and the bootstrap in lowering the bias and the MSE.
2. *Median Target*. When the sampling distribution of the statistic is skewed or has heavy tails, the mean of the statistic may not be the proper measure of location to be considered, or may not even be defined. In such cases the mean target defined above may not be the proper approach. However, in these situations we can consider the median of the statistic as a function of θ by taking $g(\theta) = \text{med}_\theta T(F_n)$ and defining the *median target estimate* in an analogous way. The resulting median target estimate will always be *median unbiased* when the g function is monotone; this is a drastic difference with the mean target situation where

some additional regularity conditions for g are needed. Results in this direction can be found in Cabrera and Watson (1996) and Cabrera et al. (2005), but many open questions about median target estimates and their variability are still awaiting their answers.

About the Author

Biography of Fernholz is in ►Functional Derivatives in Statistics: Asymptotics and Robustness.

Cross References

►Bias Correction
►Bootstrap Methods
►Estimation
►Estimation: An Overview
►Functional Derivatives in Statistics: Asymptotics and Robustness
►Jackknife
►Logistic Regression

References and Further Reading

Cabrera J, Fernholz LT (1999) Target estimation for bias and mean square error reduction. Ann Stat 27 3:1080–1104
Cabrera J, Hu I (2001) Algorithms for target estimation using stochastic approximation. InterStat 02–04:1–18
Cabrera J, Watson GS (1997) Simulation methods for mean and median bias reduction in parametric estimation. J Stat Plann Inference 57(1):143–152
Cabrera J, Devas V, Fernholz LT (2005) Target estimation for the logistic regression model. J Stat Comput Simul 75: 121–140
Fernholz LT (1983) Von Mises calculus for statistical functionals. Lecture notes in statistics, vol 19. Springer, New York
Fernholz LT (1997) Target estimation and implications to robustness. In: L_1-statistical procedures and related topics. IMS Lecture notes, monograph series, vol 31, pp 363–372
Robbins H, Munro S (1951) A stochastic approximation method. Ann Math Stat 22:400–407

Telephone Sampling: Frames and Selection Techniques

James M. Lepkowski
Director, Program in Survey Methodology, Research Professor, Survey Research Center
University of Michigan, Ann Arbor, MI, USA

Telephone sampling is a set of techniques used to generate samples in telephone survey data collection. Telephone surveys have lower cost and time of data collection than face-to-face survey methods. (Telephone surveys are also conducted for other types of units, such as business establishments. This discussion is limited to telephone household surveys.). Cost and timeliness advantages outweigh potential loss in accuracy due to failure to cover households without telephones. However, since households without telephones vary in character over time and across countries and key subgroups, researchers must decide in any particular application whether non-coverage bias is a potentially serious source of error before choosing to use a telephone survey.

Telephone sampling methods use traditional sampling techniques or modifications of those techniques designed to address the nature of the materials available for sample selection. The materials, or frames, are of two basic types: lists of telephone household numbers and lists of groups of potential telephone household numbers.

Telephone household number lists come from commercial or government sources. Some cover virtually all, or a high percentage of all, telephone households in a target population, such as those obtained from a government agency providing telephone service. Alternatively, list frame numbers may be from published telephone directories that include a majority but not all telephone households. Telephone directories do not cover recent subscribers or subscribers who do not want to have a number appear in the directory, and substantial telephone household non-coverage arising from out-of-date or absent entries has led alternative frames with more complete coverage.

Telephone sampling for list frames uses traditional element sampling techniques such as systematic selection and stratified random sampling. The lists and samples contain numbers that are not telephone household numbers, which requires screening during data collection to eliminate non-household numbers. Some telephone households have more than one telephone number in the list, which in turn have higher chances of selection. Weights are used to compensate for the duplicate numbers. If persons within households are to be sub-selected, within household selection methods choose one or more sample persons within a household, yielding additional adjustment weights.

Alternative frames or sets of materials provide more complete, if not virtually complete, coverage than directory list frames. The alternative frames are used to complete through random generation of some portion of a telephone number telephone numbers where only an area code and local prefix combination are available, and are often referred to as random digit dialing (RDD; see Groves and Kahn 1979). The frame consists of all area code and local area prefixes for a country or region obtained from government or commercial sources. These combinations

are not complete telephone numbers, but randomly generated 'suffixes' added to a selected combination yield a valid and complete telephone number. The combination plus random digits cover, in principle, all telephone households provided all combinations in the region are available.

Simple RDD telephone number generation is typically very inefficient due to a large percentage of randomly generated numbers (sometimes in excess of 80 percent) that are not telephone households, increasing costs through screening to find telephone households among randomly generated numbers. Specialized techniques reduce the percentage of non-household numbers obtained, and improve efficiency. For example, Mitofsky–Waksberg RDD sampling (Waksberg 1978) is a two-stage sampling technique devised to randomly generate numbers that have a much lower percentage of non-telephone households, below 35 percent in early applications in the United States. Practical deficiencies led to variations to improve efficiency of the two-stage methods (see, for example, Potthoff 1987).

List-assisted methods seek efficiency gains as well, but start from a directory frame to extend coverage to all telephone households (Tucker et al. 2001). Many have a slight loss of coverage, though, compared to RDD methods. Numbers selected from a directory are selected and digits in the number altered to cover numbers that are not in the directory. Plus-one dialing, for example, replaces the last digit of a directory number with a number one larger – 8 instead of 7, for instance, replaces the last digit of a phone number ending in 7. While in principle this method should cover all telephone numbers, in practice the coverage is incomplete, and difficult to determine. Variations include changing the last digit or the last two digits randomly (Sudman 1973).

Commercial sources compile lists of all directory numbers in a country, metropolitan area, or region that are used to generate telephone numbers with higher levels of coverage. Phone numbers can be divided into sets of 100 consecutive numbers defined by all but the last two digits of a telephone number. For example, directory entry 7345551212 defines 100 consecutive numbers 7345551200 to 7345551299. Commercial sources use directory entries to find all 100 "banks" where at least one directory number is present. Telephone numbers are selected at random from all numbers occurring in the set of 100 'banks' that contain one or more directory numbers. These methods provide today higher efficiency than even the two-stage RDD methods (Casady and Lepkowski 1993).

Finally, dual frame sampling designs have been used to select samples separately from directory and RDD frames and combine the results in estimation (Lepkowski 1988). These methods are also currently being used to include telephone households that have only mobile or cell telephones and are not covered by list assisted sampling frames.

About the Author

Dr. James M. Lepkowski is Professor, Research Professor, and Director, Program in Survey Methodology, at the University of Michigan, USA. He directed the Michigan Summer Institute in Survey Research Techniques (1997–2004) and the Sampling Program for Survey Statisticians (since 1992). He is a Fellow of the American Statistical Association and an Elected Member of the International Statistical Institute. He has authored or co-authored more than 80 peer-reviewed papers and books and monographs, including the textbook *Survey Methodology* (2nd Edition, Wiley, 2009) and the edited volume *Advances in Telephone Survey Methodology* (Wiley, 2007).

Cross References

►Federal Statistics in the United States, Some Challenges
►Nonsampling Errors in Surveys
►Public Opinion Polls
►Representative Samples
►Sample Survey Methods
►Sampling From Finite Populations
►Statistical Fallacies

References and Further Reading

Casady RJ, Lepkowski JM (1993) Stratified telephone sampling designs. Surv Meth 19:103–113

Groves RM, Kahn R (1979) Surveys by telephone: A national comparison with personal interviews. Academic Press, New York

Lepkowski JM (1988) Telephone sampling methods in the United States. In: Groves RM et al (eds) Telephone survey methodology, Chap. 3, Wiley, New York

Potthoff RF (1987) Some generalizations of the Mitofsky–Waksberg techniques for random digit dialing. J Am Stat Assoc 82: 409–418

Sudman W (1973) The uses of telephone directories for survey sampling. J Market Res 10:204–207

Tucker C, Lepkowski JM, Piekarski L (2001) The current efficiency of list-assisted telephone sample designs. Public Opin Quart 66:321–338

Waksberg J (1978) Sampling methods for random digit dialing. J Am Stat Assoc 73:40–46

Testing Exponentiality of Distribution

John Haywood[1], Estate V. Khmaladze[2]
[1]Senior Lecturer
Victoria University of Wellington, Wellington, New Zealand
[2]Professor
Victoria University of Wellington, Wellington, New Zealand

The exponential distribution, defined on the positive half-line $\mathbb{R}^+$ with scale parameter $\lambda > 0$, has distribution function and density

$$F_\lambda(x) = 1 - e^{-\lambda x}, \qquad f_\lambda(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

It plays a very prominent role in probability theory and statistics, especially as a model for random times until some event, like emission of radioactive particles (Rutherford et al. 1910), or an earthquake (Gardner and Knopoff 1974), or failure of equipment (Pham 2003), or occurrence of abnormally high levels of a random process (Cramér and Leadbetter 1967), like unusually high prices (Shiryaev 1999), etc.

The characteristic "memoryless" property of the exponential distribution says that, if X is an exponential random variable, then

$$P\{X > y + x | X > y\} = P\{X > x\}, \quad \text{or}$$
$$1 - F_\lambda(x+y) = [1 - F_\lambda(x)][1 - F_\lambda(y)], \qquad (1)$$

which means that the chances to wait for longer than some time x do not change, if you have been waiting already for some time y: X "does not remember" if waiting has occurred already or not. Connected to this is another characteristic property of the exponential distribution, which states that its failure rate is constant:

$$\frac{f_\lambda(x)}{1 - F_\lambda(x)} = \lambda. \qquad (2)$$

Given a sample $X_1, \ldots, X_n$, denote by F_n and ν_n the empirical distribution function and the empirical process, respectively:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\}, \quad \nu_n(x) = \sqrt{n}[F_n(x) - F_\lambda(x)],$$

where $\mathbb{I}\{A\}$ denotes the indicator function of the event A. As is well known, after time transformation $t = F_\lambda(x)$, the process $\nu_n \circ F_\lambda^{-1}(t) = \nu_n\left(F_\lambda^{-1}(t)\right)$ converges in distribution to a standard Brownian bridge $u(t), t \in [0,1]$. Since in the majority of problems the value of the parameter λ is unknown, inference can not be based on ν_n but must use the parametric (or estimated) empirical process $\hat{\nu}_n$,

$$\hat{\nu}_n(x) = \nu_n(x, \hat{\lambda}_n) = \sqrt{n}[F_n(x) - F_{\hat{\lambda}_n}(x)],$$

where $\hat{\lambda}_n$ is an estimator of λ, based on the sample.

In any testing procedure one can use either of two types of statistics from $\hat{\nu}_n$, or a combination of the two: linear, or asymptotically linear, statistics and nonlinear omnibus statistics. Asymptotically linear statistics of the form

$$l_n(X_1, \ldots, X_n; F_{\hat{\lambda}_n}) = \int_0^\infty g(x) d\hat{\nu}_n(x) + o_P(1) \qquad (3)$$

typically lead to asymptotically optimal tests against specific "local" (or contiguous) alternatives, but have very poor power against the huge majority of other alternatives. In contrast, nonlinear statistics like

$$\sup_x |\hat{\nu}_n(x)| \quad \text{or} \quad \int_0^\infty \hat{\nu}_n^2(x) dF_{\hat{\lambda}_n}(x),$$

which may not have best power against any given alternative, have reasonable power against more or less all alternatives. These are used in goodness of fit testing problems.

It is for these omnibus tests that the asymptotic behavior of the empirical process $\hat{\nu}_n$ is somewhat unpleasant: after time transformation $t = F_{\hat{\lambda}}(x)$ it does not converge to a standard Brownian bridge, but to a different Gaussian process with more complicated distribution. While it is true that the distribution of each omnibus statistic can in principle be calculated and tables prepared, this would involve a considerable amount of computational work. Below we show versions of ▶empirical processes that are distribution free and, moreover, the distribution of many statistics from these processes are already known.

Asymptotically Linear Statistics

There are several asymptotically linear statistics, which are widely used for testing exponentiality. Papers (Deshpande 1983) and (Bandyopadhyay and Basu 1989) are based on testing whether $1 - F_\lambda(bx) = [1 - F_\lambda(x)]^b$, and the test statistic is

$$D_n = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}\{X_j > bX_i\}.$$

A statistic known as the Gini index (or coefficient),

$$G_n = \frac{\sum_{i \neq j} |X_i - X_j|}{2n(n-1)\overline{X}}, \quad \text{with } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

was originally designed as a measure of spread and is commonly used as a measure of inequality, e.g., see Deaton (1997). In Gail and Gastwirth (1978), and later Nikitin and Tchirina (1996), it was considered and recommended as a test of exponentiality.

The so-called Moran statistic was introduced in Moran (1951) as the score statistic for testing exponentiality against the alternative of a Gamma distribution and has the form

$$M_n = \frac{1}{n}\sum_{i=1}^{n} \log \frac{X_i}{\overline{X}}.$$

One more test of exponentiality, known as the Cox-Oakes statistic, was suggested in Cox and Oakes (1984) as the score test statistic against the alternative of a ▶Weibull distribution:

$$C_n = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{X_i}{\overline{X}}\right)\log \frac{X_i}{\overline{X}}.$$

One can show that all four statistics are asymptotically linear, e.g., see Haywood and Khmaladze (2008), and hence are asymptotically Gaussian. Somewhat surprisingly, although the kernels g of representation (3) in all four statistics look different, their correlation is extremely high, which means that all four statistics lead to the same test in practice; see Haywood and Khmaladze (2008).

Distribution Free Versions of Empirical Processes

As we noted above, unlike the empirical process ν_n, the time transformed parametric empirical process $\hat{\nu}_n \circ F_\lambda^{-1}$ does not converge to a standard Brownian bridge u. However, a beautiful observation, see Barlow and Campo (1975; Barlow and Proschan 1975), leads to another version of empirical process, which does. It is based on the "total time on test" (or TTT) notion of Epstein and Sobel (1953). Consider

$$\eta_n(x) = \frac{\int_0^x [1 - F_n(y)]dy}{\int_0^\infty [1 - F_n(y)]dy},$$

$$\text{where } \int_0^\infty [1 - F_n(y)]dy = \overline{X},\ x \geq 0.$$

If one interprets random variable X_i as a survival time (or time until failure) of the ith item on test, then $\eta_n(x)$ measures the time all items spent on test before the moment x, relative to the total time spent on test by all n items until they all failed. The process

$$\xi_n(x) = \sqrt{n}[F_n(x) - \eta_n(x)]$$

will converge in distribution to a Brownian bridge in time F_λ, and hence the time transformed empirical process $\xi_n \circ F_\lambda^{-1}$ converges in distribution to a standard Brownian bridge. To explain why this is true, cf. (Gill 1986; Khmaladze 1981), note that the process

$$B_n(x) = \sqrt{n}\left[F_n(x) - \int_0^x \frac{1 - F_n(y)}{1 - F(y)}dF(y)\right]$$

is a martingale (see ▶Martingales) with respect to the natural filtration $\{\mathcal{F}_x\}$ generated by F_n, for any i.i.d. observations. Using (2), in the case of the exponential distribution it reduces to

$$B_n(x, \lambda) = \sqrt{n}\left[F_n(x) - \lambda \int_0^x 1 - F_n(y)dy\right].$$

If we estimate the parameter λ through the equation $B_n(\infty, \lambda) = 0$, we get the usual estimator $\hat{\lambda}_n = 1/\int_0^\infty [1 - F_n(y)]dy = 1/\overline{X}_n$ and $B_n(x, \hat{\lambda}_n) = \xi_n(x)$. The process $B_n\left(F_\lambda^{-1}(t), \lambda\right)$ converges in distribution to standard Brownian motion on $[0,1]$ and hence $\xi_n \circ F_\lambda^{-1}$ converges to "tied up" Brownian motion, i.e., a standard Brownian bridge.

Another version of empirical process was investigated in Haywood and Khmaladze (2008). It has the form

$$w_n(x) = \sqrt{n}[F_n(x) - K(x, F_n)],$$

where

$$K(x, F_n) = \frac{\hat{\lambda}}{n}\sum_{i:X_i \leq x}\left(2X_i - \frac{\hat{\lambda}}{2}X_i^2\right) + \hat{\lambda}\left(2 + \frac{\hat{\lambda}}{2}x\right)x[1 - F_n(x)] - x\frac{\hat{\lambda}^2}{n}\sum_{i:X_i > x} X_i.$$

Asymptotically, the process w_n is also a martingale, but with respect to the "enriched" filtration $\{\hat{\mathcal{F}}_x\}$, where each σ-field is generated by the past of F_n and also the estimator $\hat{\lambda}_n$; $\hat{\mathcal{F}}_x = \sigma\{F_n(y), y \leq x, \overline{X}_n\}$. The idea behind this process follows from the general suggestion in Khmaladze (1981), but the form of compensator $K(x, F_n)$ for the exponential distribution is computationally particularly simple. Haywood and Khmaladze (2008) demonstrated quick convergence of the time transformed process $w_n \circ F_\lambda^{-1}$ to a standard Brownian motion (see ▶Brownian Motion and Diffusions).

Although not proved formally, the relationship between processes ξ_n and w_n is clear: the latter is asymptotically the innovation martingale for the former and therefore the two stay in one-to-one correspondence. The limit distribution of many statistics based on both $\xi_n \circ F_\lambda^{-1}$ and $w_n \circ F_\lambda^{-1}$ are well known.

Koul (1978) considered an empirical version of the memoryless property (1) of the exponential distribution and studied the empirical process

$$\alpha_n(x, y) = -\sqrt{n}\{1 - F_n(x + y) - [1 - F_n(x)][1 - F_n(y)]\}.$$

The asymptotic form of Koul's process is

$$\alpha_n(x, y) = \nu_n(x + y) - [1 - F_\lambda(x)]\nu_n(y) - [1 - F_\lambda(y)]\nu_n(x) + o_p(1)$$

and therefore, after the usual time transformation, it converges in distribution to β,

$$\beta(t,s) = u(ts) - tu(s) - su(t),$$

which is again a distribution free process in t and s. A particular form of this process,

$$\alpha_n(x) = -\sqrt{n}\left\{1 - F_n(bx) - [1 - F_n(x)]^b\right\}$$

with $b = 2$ was studied in Angus (1982) and Nikitin (1996). Note that the limit distributions of omnibus statistics from these α_n processes are not easy to obtain.

P-P Plots

It is easy and quick to calculate random variables $\hat{U}_i = 1 - F_{\hat{\lambda}_n}(X_i) = e^{-\hat{\lambda}_n X_i}$ and plot their ►order statistics $\hat{U}_{(i:n)}$ against expected values $i/(n+1), i = 1,\ldots,n$, of the uniform order statistics. Under exponentiality the graph should be approximately linear, as $\hat{U}_i, i = 1,\ldots,n$ are almost independent and almost uniformly distributed on $[0,1]$: they would exactly have these properties if λ was known and used instead, but with $\hat{\lambda}_n$ they are not. Visual inspection of the graph is a useful preliminary tool. However, the normalized differences

$$\sqrt{n}\left[\hat{U}_{(i:n)} - \frac{i}{n+1}\right]$$

as a process in $t = i/(n+1)$, has the same drawback as the time transformed parametric empirical process $\hat{v}_n \circ F_{\hat{\lambda}}^{-1}$: distributions of many statistics from it are not known and would require extra computational effort.

Uniform Spacings

If $0 = V_{(0:n-1)} < V_{(1:n-1)} < \cdots < V_{(n-1:n-1)} < V_{(n:n-1)} = 1$ denote the uniform order statistics from a sample of size $n-1$, the differences $\Delta V_{(i-1:n-1)} = V_{(i:n-1)} - V_{(i-1:n-1)}$, form uniform spacings. Random variables

$$\frac{X_i}{\sum_{j=1}^n X_j} = \frac{X_i}{n\overline{X}},\ i = 1,\ldots,n,$$

have the same distribution as $\Delta V_{(i-1:n-1)}, i = 1,\ldots,n$, if and only if $X_1,\ldots,X_n$ are i.i.d. exponential random variables. This characteristic property was systematically used in testing problems pertaining to uniform spacings, (Pyke 1965).

Although the normalized spacings $n\Delta V_{(i-1:n-1)}$ form a distribution free statistic, they are dependent, and the empirical process based on them does not converge to a Brownian bridge. It can be shown that this empirical process is asymptotically equivalent to the process $\hat{v}_n \circ F_{\hat{\lambda}}^{-1}$.

Other approaches for testing exponentiality include tests based on functionals from the empirical characteristic function and Laplace transform, studied, e.g., in Baringhaus and Henze (1991), Epps and Pulley (1986) and Henze (1993), and on the empirical likelihood principle, e.g., Einmahl and McKeague (2003). Surveys on tests for exponentiality, including numerical studies of their relative power against fixed alternatives, can be found in Ascher (1990) and Henze and Meintanis (2005).

About the Authors

John Haywood is Senior Lecturer, School of Mathematics, Statistics and Operations Research, Victoria University.

Estate Khmaladze completed his Ph.D. in 1971 at V.A. Steklov Mathematical Institute, Moscow, under supervision of L. N. Bolshev, who was head of department of mathematical statistics at Steklov after N.V. Smirnov. He was awarded the title Professor in Probability Theory and Mathematical Statistics in 1992. Returning to Tbilisi permanently in 1990, he was appointed Head of Department of Probability Theory and Mathematical Statistics of A. Razmadze Mathematical Institute, 1990–1999, where he is still Honorary Member. He played key role in formation of Georgian Statistical Association and served as its Vice-President from 1991 to 1998. From 1991 to 1998 he served as Vice-President of Georgian Statistical Association. In 1996, Khmaladze moved to Sydney, Australia, and from there to New Zealand, where in 2002 he was appointed a Professor in Statistics at Victoria University of Wellington. School of Mathematics, Statistics and Computer Sciences. Professor Khmaladze was the first Soviet statistician to become an Associate Editor of *The Annals of Statistics* (1989–1991). Currently, he is Associate Editor for several international journals. *Mathematical Methods of Statistics*, *Statistics and Probability Letters*, *Annals of the Institute of Statistical Mathematics* and *Sankhya*. He is widely known for broadening applications of martingale methods in statistics in general and for *Khmaladze transformation*, in particular. His research interests include recently established connections of set-valued analysis and differential geometry to statistics and the statistical theory of diversity.

"This week sees Wellington inherit its very own 'beautiful mind' with the arrival at Victoria University of outstanding mathematician and statistician, Professor Estate Khmaladze. Victoria staff and students are extremely lucky to have someone of the calibre and experience of Professor Khmaladze among us" (Victoria University of Wellington, Media Release, March 25, 2002).

Cross References

►Accelerated Lifetime Testing
►Brownian Motion and Diffusions
►Empirical Processes
►Exponential Family Models

►Lorenz Curve
►Relationships Among Univariate Statistical Distributions
►Stochastic Processes: Applications in Finance and Insurance
►Survival Data

References and Further Reading

Angus JE (1982) Goodness-of-fit tests for exponentiality based on a loss-of-memory type functional equation. J Stat Plann Inference 6:241–251
Ascher S (1990) A survey of tests for exponentiality. Commun Stat 19:1811–1825
Bandyopadhyay D, Basu AP (1989) A note on tests for exponentiality by Deshpande. Biometrika 76:403–405
Baringhaus L, Henze N (1991) A class of consistent tests for exponentiality based on the empirical Laplace transform. Ann Inst Stat Math 43:179–192
Barlow RE, Campo R (1975) Total time on test processes and applications to failure data analysis. In: Barlow RE, Fussell J, Singpurwalla ND (eds) Reliability and fault tree analysis. SIAM, Philadelphia, pp 451–481
Barlow RE, Proschan F (1975) Statistical theory of reliability and life testing: probability models. Holt, Rinehart and Winston, New York
Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, London
Cramér H, Leadbetter MR (1967) Stationary and related stochastic processes: sample function properties and their applications. Wiley, New York
Deaton A (1997) The analysis of household surveys: a microeconometric approach to development policy. Johns Hopkins University Press, Baltimore
Deshpande JV (1983) A class of tests for exponentiality against increasing failure rate average alternatives. Biometrika 70:514–518
Einmahl JHJ, McKeague IW (2003) Empirical likelihood based hypothesis testing. Bernoulli 9:267–290
Epps TW, Pulley LB (1986) A test for exponentiality vs. monotone hazard alternatives derived from the empirical characteristic function. J R Stat Soc B 48:206–213
Epstein B, Sobel M (1953) Life testing. J Am Stat Assoc 48:486–502
Gail MH, Gastwirth JL (1978) A scale-free goodness-of-fit test for exponentiality based on the Gini statistic. J R Stat Soc B 40: 350–357
Gardner JK, Knopoff L (1974) Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? Bull Seismol Soc Am 64:1363–1367
Gill RD (1986) The total time on test plot and the cumulative total time on test statistic for a counting process. Ann Stat 14: 1234–1239
Haywood J, Khmaladze EV (2008) On distribution-free goodness-of-fit testing of exponentiality. J Econom 143:5–18
Henze N (1993) A new flexible class of omnibus tests for exponentiality. Commun Stat 22:115–133
Henze N, Meintanis SG (2005) Recent and classical tests for exponentiality: a partial review with comparisons. Metrika 61:29–45
Khmaladze EV (1981) Martingale approach in the theory of goodness of fit tests. Theory Probab Appl 26:240–257
Koul HL (1978) A class of tests for testing "new is better than used." Can J Stat 6:249–271
Moran PAP (1951) The random division of an interval – part II. J R Stat Soc B 13:147–150
Nikitin Y (1996) Bahadur efficiency of a test of exponentiality based on a loss of memory type functional equation. J Nonparametric Stat 6:13–26
Nikitin Y, Tchirina AV (1996) Bahadur efficiency and local optimality of a test for the exponential distribution based on the Gini statistic. J Ital Stat Soc 5:163–175
Pham H (ed) (2003) Handbook of reliability engineering. Springer, London
Pyke R (1965) Spacings (with discussion). J R Stat Soc B 27:395–449
Rutherford E, Geiger H, Bateman H (1910) The probability variations in the distribution of α particles. Philos Mag 20:698–707
Shiryaev AN (1999) Essentials of stochastic finance. Facts, models, theory. World Scientific, Singapore

Testing Variance Components in Mixed Linear Models

Mohamed Y. El-Bassiouni
Professor and Head
United Arab Emirates University, Al-Ain, UAE

Introduction

Consider the mixed model

$$Y = X\beta + Z\gamma + \varepsilon, \tag{1}$$

where Y is an $n \times 1$ observable random vector, X is an $n \times p$ known matrix, β is a $p \times 1$ vector of unknown parameters, Z is another known $n \times m$ matrix, γ is an $m \times 1$ unobservable random vector such that $\gamma \sim N_m(0, \theta_1 I)$, $\theta_1 \geq 0$, and ε is another unobservable $n \times 1$ random vector such that $\varepsilon \sim N_n(0, \theta_0 I)$, $\theta_0 > 0$. It is also assumed that γ and ε are independent and that $n > \text{rank}(X, Z) > \text{rank}(X)$. Therefore, we have

$$Y \sim N_n(X\beta, \theta_0 I + \theta_1 ZZ'). \tag{2}$$

Model (1) can be generalized to more than two variance components and has proven useful to practitioners in a variety of fields such as genetics, biology, psychology, and agriculture, where it is usually of interest to test the null hypothesis $\theta_1 = 0$ against the alternative $\theta_1 > 0$, or equivalently,

$$H_0 : \rho = 0, \qquad \text{vs} \qquad H_1 : \rho > 0, \tag{3}$$

where $\rho = \theta_1/\theta_0$.

Wald Test

Wald (1947) proposed an exact procedure to construct confidence intervals for ρ, which can be used to test the hypotheses in (3). This test is based on the usual ANOVA F-statistic and uses the readily available F tables to determine the critical region and that is why it has been widely used in applications. The Wald test was shown by Spjotvoll (1967) to be optimal against large alternatives. Further, unless the design is strongly unbalanced and the alternative is fairly small, El-Bassiouni and Seely (1988) showed that the Wald test has reasonable efficiency relative to the corresponding MP tests.

Seely and El-Bassiouni (1983) obtained the Wald test via reduction sums of squares. This circumvents the necessity of transforming to independent variables and/or modifying Wald's method as discussed by Spjotvoll (1968). They also give necessary and sufficient conditions under which the Wald test can be used in mixed models as well as a uniqueness property that allows one to immediately determine whether or not a proposed variance component test in a mixed model is the Wald test.

Likelihood Ratio (LR) Tests

Likelihood (LR) tests were developed by Hartley and Rao (1967) who showed that such tests are consistent and unbiased and recommended that the LR tests be carried out by comparing the observed values of the test statistics with the (approximate) cutoff points obtained from the standard χ^2 tables. However, such cutoff points can yield sizes quite different from the nominal sizes (Garbade 1977). Since the computation of the maximum likelihood estimates requires the numerical solution of a constrained nonlinear optimization problem, the LR tests have not been used much in practice. Nevertheless, Harville (1977) gives some results to facilitate the computation of LR tests. It should also be noted that even for balanced models, when a UMPU (uniformly most powerful unbiased) F-test is available, the LR approach does not necessarily yield the UMPU F-test (Herbach 1959). Using the likelihood induced by maximal location-invariant statistics leads to the restricted LR (RLR) tests. For balanced models, these RLR tests are for all practical purposes equivalent to the F-tests (El-Bassiouni 1981, 1982).

For a discussion of LR and RLR tests and their comparison with the Wald and LMPI (locally most powerful invariant) tests, the reader is referred to Li et al. (1996) and the references therein.

Uniformly Most Powerful Unbiased (UMPU) Tests

Optimal tests for certain functions of the parameters of the covariance matrix were developed by El-Bassiouni and Seely (1980), where the theory in Chap. 4 of Lehmann (1959) for determining UMPU tests in exponential families is applied to a zero mean multivariate normal family that admits a complete sufficient statistic. The special case when the matrices in the covariance structure commute was emphasized. It appears that while completeness buys similarity, it is the additional assumption of commutativity that buys simple test procedures. The case of a nonzero mean family was also discussed as were some results on the completeness of families of product measures.

In balanced models, Mathew and Sinha (1988) showed that the usual ANOVA F-test is UMPU and UMPIU (uniformly most powerful invariant unbiased), but in unbalanced models, no such UMP test exists (Spjotvoll 1967).

Similar and Location-Invariant Tests

In the context of unbalanced mixed models, if one has a specific alternative $\rho^* > 0$ in mind, a most powerful test among similar location-invariant tests, which is also MPI (most powerful among location- and scale-invariant tests), was developed by Spjotvoll (1967).

As $\rho^* \to \infty$, Spjotvoll (1967) showed that the MPI test reduces to the exact F-test of Wald (1947). On the other hand, to guard against small alternatives ($\rho^* \to 0$), the LMPI test was considered by Westfall (1988, 1989) who compared the Wald and LMPI tests in classification designs and concluded that the LMPI test is better in large designs whereas the Wald test may be preferable in small designs. Further, Westfall (1989) found that the Wald test is inferior to the LMPI test whenever there is a small proportion of relatively large group sizes. The ▶harmonic mean method was used by Thomas and Hultquist (1978) to construct confidence intervals for ρ in unbalanced random one-way models. The method was generalized to unbalanced mixed models by Harville and Fenech (1985). A modified harmonic mean procedure that compares favorably with the Wald and LMPI tests was proposed by El-Bassiouni and Seely (1996) for testing the hypotheses in (3).

For $\rho_\ell < \rho_0 < \rho_u$, Lin and Harville (1991) combined the two MPI tests against ρ_ℓ and ρ_u to obtain a two-sided test of $H_0 : \rho = \rho_0$ vs $H_1 : \rho \neq \rho_0$ and showed that their NP (Neyman–Pearson) test, although computationally intensive, can be better than the Wald test in some designs. Motivated by this idea, El-Bassiouni and Halawa (2003) proposed a test that combines the LMPI test ($\rho_\ell \to 0$) and the Wald test ($\rho_u \to \infty$) to obtain a test of $H_0 : \rho_0 = 0$ vs $H_1 : \rho > 0$. The combined test statistic is easily computed

and its null distribution may be approximated by a central F distribution with the degrees of freedom of the numerator adjusted in accordance with the degree of imbalance of the design. It is also shown to be a member of the complete class of tests of El-Bassiouni and Seely (1996). Numerical methods were used to show that the approximation is accurate over a wide range of conditions and that the efficiency of the combined test, relative to the power envelope, is satisfactorily high overall.

The combined test was also adapted to the case where $n = \text{rank}(X, Z)$ (El-Bassiouni and Charif 2004). Such models with zero degrees of freedom for error occur in many applications including plant and animal breeding and time-varying regression coefficients.

About the Author

Dr. Mohamed Yahia El-Bassiouni is a Professor and Head, Department of Statistics, UAE University. He obtained his Ph.D. degree in Statistics from Oregon State University in 1977 and has received the Legion of Science, First Degree Honor, from the Egyptian President in 1984, in honor of his research. He is also the Editor of the *Journal of Economic and Administrative Sciences*, UAE University, and the Regional Editor of the I*nternational Journal of Management and Sustainable Development*, United Kingdom. Professor El-Bassiouni has authored and coauthored more than 70 papers, 15 monographs, and 30 research reports, many of which are on testing and interval estimation of variance components.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►Cross Classified and Multiple Membership Multilevel Models
►General Linear Models
►Linear Mixed Models
►Multilevel Analysis
►Panel Data

References and Further Reading

El-Bassiouni MY (1981) Likelihood ratio tests for covariance hypotheses generating commutative quadratic subspaces. Commun Stat A10(23):2461–2468
El-Bassiouni MY (1982) On a theorem of Graybill. Commun Stat A11(13):1519–1522
El-Bassiouni MY, Charif HA (2004) Testing a null variance ratio in mixed models with zero degrees of freedom for error. Comput Stat Data Anal 46:707–719
El-Bassiouni MY, Halawa AM (2003) A combined invariant test for a null variance ratio. Biom J 45:249–260
El-Bassiouni MY, Seely J (1980) Optimal tests for certain functions of the parameters in a covariance matrix with linear structure. Sankhya 42:64–77
El-Bassiouni MY, Seely JF (1988) On the power of Wald's variance component test in the unbalanced random one-way model. In: Dodge Y, Federov VV, Wynn HP (eds) Optimum design and analysis of experiments. Elsevier, North-Holland, pp 157–165
El-Bassiouni MY, Seely JF (1996) A modified harmonic mean test procedure for variance components. J Stat Plan Infer 49:319–326
Garbade K (1977) Two methods for examining the stability of regression coefficients. J Am Stat Assoc 72:54–63
Hartley HO, Rao JNK (1967) Maximum-Likelihood estimation for the mixed analysis of variance model. Biometrika 54:93–108
Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72:320–338
Harville DA, Fenech AP (1985) Confidence intervals for a variance ratio, or for heritability, in an unbalanced mixed linear model. Biometrics 41:137–152
Herbach LH (1959) Properties of model II-type analysis of variance tests, A: Optimum nature of the F-test for model II in the balanced case. Ann Math Stat 30:939–959
Lehmann EL (1959) Testing statistical hypotheses. Wiley, New York
Li Y, Birkes D, Thomas DR (1996) The residual likelihood ratio test for the variance ratio in a linear model with two variance components. Biom J 38:961–972
Lin TH, Harville DA (1991) Some alternatives to Wald's confidence interval and test. J Am Stat Assoc 86:179–187
Mathew T, Sinha BK (1988) Optimum tests for fixed effects and variance components in balanced models. J Am Stat Assoc 83:133–135
Seely JF, El-Bassiouni MY (1983) Applying Wald's variance component test. Ann Stat 11:197–201
Spjotvoll E (1967) Optimum invariant tests in unbalanced variance component models. Ann Math Stat 38:422–429
Spjotvoll E (1968) Confidence intervals and tests for variance ratios in unbalanced variance components models. Rev Int Stat Inst 36:37–42
Thomas JD, Hultquist RA (1978) Interval estimation for the unbalanced case of the one-way random effects model. Ann Stat 6:582–587
Wald A (1947) A note on regression analysis. Ann Math Stat 18: 586–589
Westfall PH (1988) Robustness and power of tests for a null variance ratio. Biometrika 75:207–214
Westfall PH (1989) Power comparisons for invariant variance ratio tests in mixed ANOVA models. Ann Stat 17:318–326

Tests for Discriminating Separate or Non-Nested Models

Basilio de Bragança Pereira
Professor of Biostatistics at the school of Medicine
Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Introduction

The Neyman–Pearson theory of hypothesis testing applies if the models belong to the same family of distributions. Alternatively, special procedures are needed if the models

belong to families that are separate or non-nested, in the sense that an arbitrary member of one family cannot be obtained as a limit of members of the other.

Let $y = (y_1, \ldots, y_n)$ be independent observations from some unknown distribution. Suppose that there are null and alternative hypotheses H_f and H_g specifying parametric densities $f(y, \alpha)$ and $g(y, \beta)$ for the random vector y. Hence α and β are unknown vector parameters and it is assumed that the families are separate.

The asymptotic tests developed by Cox (1961, 1962) were based on a modification of the Neyman–Pearson maximum likelihood ratio. If H_f is the null hypothesis and H_g the alternative hypothesis, the test statistics considered was

$$T_{fg} = T_{fg}(\hat{\alpha}, \hat{\beta}) - E_{\hat{\alpha}}\{T_{fg}(\alpha, \beta_\alpha)\}$$

where for a random sample of size n, $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimators of α and β respectively, $T_{fg}(\alpha, \beta) = 1_f(\alpha) - 1_g(\beta)$ is the log likelihood ratio, β_α is the probability limit, as $n \to \infty$, of $\hat{\beta}$ under H_f and the subscript α means that expectations, etc. are calculated under H_f.

Cox showed that, asymptotically, under the alternative hypothesis T_{fg} has a negative mean and that under the null hypothesis T_{fg} is normally distributed with mean zero and variance

$$V_\alpha(T_{fg}) = V_\alpha\{IT_{fg}(\alpha, \beta_\alpha)\} - C_\alpha^{-1} I^{-1} C_\alpha$$

where $C_\alpha = \partial E_\alpha\{T_{fg}(\alpha, \beta_\alpha)\}/\partial\alpha$, and I_α the information matrix of α. When H_g is the null hypothesis and H_f is the alternative hypothesis analogous results are obtained for a statistics T_{gf}. Therefore $T^*_{fg} = T_{fg}\{V_\alpha(T_{fg})\}^{-1/2}$ and $T^*_{gf} = T_{gf}\{V_\beta(T_{gf})\}^{-1/2}$ under H_f and H_g respectively can be considered as approximately standard normal variables and two-tailed tests can be performed. The outcomes of application of both tests are shown in the Table 1.

As an alternative to his test, Cox (1961) suggested combining the two models in a general model of which they would be both special cases. The density could be proportional to the exponential mixture

$$\{f(y, \alpha)\}^\lambda \{g(y, \beta)\}^{1-\lambda}$$

and inferences made about λ. It should be notice that these mixtures can be generalized for testing more than two models. In particular, the exponential mixture is the base of the tests developed in econometrics.

Cox also suggested a Bayesian approach and gives a general expression when losses are associated and a large sample approximation.

The posterior odds for H_f versus H_g is

$$\frac{\pi_f \int f(y; \alpha)\pi_f(\alpha)d\alpha}{\pi_g \int g(y; \beta)\pi_g(\beta)d\beta} = \frac{\pi_f}{\pi_g} B_{fg}(y)$$

where π_f and π_g are the prior probabilities of H_f and H_g respectively, $\pi_f(\alpha)$ and $\pi_g(\beta)$ are the prior probabilities for the parameters conditionally on H_f and H_g. $B_{fg}(y)$ is the Bayes Factor and represents the weight of evidence in the data for H_f over H_g.

One difficulty with this approach lies in the fact that the prior knowledge expressed by π_f and $\pi_f(\alpha)$ must be coherent with that of π_g and $\pi_g(\beta)$. If the parameter spaces have different dimensions and there is no simple relation between the parameters, the problem is not simple. When prior information is weak and improper prior is used there are also difficulties and paradox with the use of Bayes factors which is unspecified.

Alternative Approaches

Alternative approaches present in Cox (1961) were further developed under Cox supervision in unpublished Ph.D. thesis at Imperial College : O.A.Y. Jackson in 1968 and B. de B. Pereira in 1976 obtained further results on the modified likelihood ratio, A.C. Atkinson in 1970 developed the exponential compound model approach, J. K. Lindsey in 1972 used a direct relative likelihood approach. Later in 1980 A. C. Atkinson supervised L. R. Pericchi on the Bayesian approach. Published references from this work can be traced in Pereira (1977a, b). Further contributions of Cox in this area are Cox (1974), Cox and Brandwood (1959), Atkinson and Cox (1974), Chambers and Cox (1967).

Further alternative approaches such as: linear mixtures, relative likelihoods, tests based on information and divergence measures, ▶moment generating functions, multiple combinations, methods based on invariant statistics and method of moments and bootstrap are reviewed in Pereira (1998, 2005).

A huge amount of research on separate families of hypothesis was developed since the fundamental work of Cox (1961, 1962). In the 1980s econometricians, using the exponential compound model took a great interest in the subject. Bayesian statisticians in the 1990s developed alternative Bayes factors (see Araújo and Pereira 2007). For reviews and references see McAller et al. (1990), Gourieroux and Monfort (1994), McAller (1995), Pereira (1977b, 1981a, 1998, 2005), and Pesaran and Weeks (2001).

A test based on descriptive statistics for the mean and the variance of the log-likelihood ratio has been proposed by Vuong (1989) but this has not been compared with Cox test that has been shown to be consistent (Pereira 1977a)

Tests for Discriminating Separate or Non-Nested Models. Table 1 Possible results of Cox test

	T_{fg}		
T_{gf}	Significant negative	Not significant	Significant positive
Significant negative	Reject both	Accept H_f	Reject both
Not significant	Accept H_g	Accept both	Possible acceptance H_g
Significant positive	Reject both	Possible acceptance H_f	Reject both

and the only that can be extend to multivariate problems (Araújo et al. 2005; Timm and Al-Subaihi 2001) and that approaches the normal asymptotic result faster (Pereira 1978).

About the Author

Dr. Basilio de Bragança Pereira obtained his Ph.D. and D.I.C. from the Imperial College of Science, Technology and Medicine (1976), supervised by Sir David Cox. In 2003 he spent a year working on a Project on Neural Networks in Statistics with Professor C.R. Rao at Penn State University on a postdoctoral grant from the Brazilian Government (CAPES). He was Associate Professor at the Institute of Mathematics (1970–1989 and 1994–1997), and Professor Titular of Applied Statistics at COPPE (1989–1994, retired). Currently, he is Professor Titular of Biostatistics at the School of Medicine (since 1998) and the coordinator of the Statistical research consulting group at The University Hospital of UFRJ. He has supervised 19 PhD and 38 MSc students. He is an Elected member of the International Statistical Institute. Professor Pereira has coauthored over 70 refereed papers, and a monograph *Data Mining with Neural Networks: A Guide for Statisticians* (with C.R. Rao, available for download in TextBook Revolution).

Cross References

►Bayesian Statistics
►Econometrics
►Mixture Models
►Neyman-Pearson Lemma
►Significance Testing: An Overview

References and Further Reading

Araújo MI, Pereira BB (2007) Comparison among Bayes factors for separate models: some simulation results. Commun Stat – Simul Comput 36:297–309

Araújo MI, Fernandes M, de B Pereira B (2005) Alternative procedures to discriminate non nested multivariate linear regression models. Commun Stat – Theory Methods 34:2047–2062

Atkinson AC, Cox DR (1974) Planning experiments for discriminating between models (with discussion). J R Stat Soc B 36:321–348

Chambers EA, Cox DR (1967) Discriminating between alternative binary response models. Biometrica 54:573–578

Cox DR (1961) Tests of separate families of hypotheses. In: Proceedings of fourth Berkeley symposium, vol 1, pp 105–123

Cox DR (1962) Further results on tests of separate families of hypotheses. J R Stat Soc B 24:406–423

Cox DR (1974) Discussion of "Dempster, A.P. pg 353 – The direct use of likelihood for significance testing". In: Proceedings of conference on foundational questions in statistical inference, Aarhus. Memoirs no. 1. University of Aarhus, Aarhus

Cox DR, Brandwood L (1959) On a discriminatory problem connected with the works of Plato. J R Stat Soc B 21:195–200

Davidson R, MacKinnon JD (1983) Testing the specification of multivariate models in the presence of alternative hypotheses. J Econom 23:301–313

de B Pereira B (1977a) A note on the consistency and on finite sample comparisons of some tests of separate families of hypotheses. Biometrika 64:109–113 (correction in volume 64, page 655)

de B Pereira B (1977b) Discriminating among separate models: a bibliography. Int Stat Rev 45:163–172

de B Pereira B (1978) Empirical comparisons of some tests of separate families of hypotheses. Metrika 25:219–234

de B Pereira B (1981) Discriminating among separate models: an additionalbibliography. Int Stat Inf 62(2):3 (repr Katti SK (1982) On the preliminary test for the CEAS model versus the Thompson model for predicting soybean production. Technical Report 125, Department of Statistics, University of Missouri – Columbia)

de B Pereira B (1998) Separate families of hypotheses. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 5. Wiley, pp 4069–4074

de B Pereira B (2005) Separate families of hypotheses. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 7, 2nd edn. Wiley, pp 4881–4886

Gourieroux C, Monfort A (1994) Testing non-nested hypotheses. In: Engle R, McFadden DL (eds) Handbook of econometrics, vol IV (Chapter 44). Elsevier, London, pp 2585–2637

McAller M (1995) The significance of testing empirical non-nested models. J Econom 65:149–171

McAller M, Pesaran MH, Bera AK (1990) Alternative approaches to testing non-nested models with autocorrelated disturbances. Commun Stat – Theory Methods 19:3619–3644

Pesaran MH (1982) On the comprehensive method for testing non-nested regression models. J Econom 18:263–274

Pesaran MH, Deaton AS (1978) Testing non-nested regression models. Econometrica 46:677–694

Pesaran MH, Weeks M (2001) Non-nested hypothesis testing: an overview. In: Baltagi BH (ed) Companion to theoretical econometrics. Basil Blackwell, Oxford

Timm NH, Al-Subaihi AA (2001) Testing model specification in seemingly unrelated regression models. Commun Stat – Theory Methods 30:577–590

Uloa RD, Pesaran MH (2008) Non-nested hypotheses. In: Durlauf SN, Blume LE (eds) The new Palgrave dictionary of economics, 2nd edn. Palgrave Macmillan

Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica 57:307–333

Tests for Homogeneity of Variance

Nataša Erjavec
Professor, Faculty of Economics
University of Zagreb, Zagreb, Croatia

Introduction

Homogeneity of variance (*homoscedasticity*) is an important assumption shared by many parametric statistical methods. This assumption requires that the variance within each population be equal for all populations (two or more, depending on the method). For example, this assumption is used in the two-sample *t*-test and ANOVA. If the variances are not homogeneous, they are said to be *heterogeneous*. If this is the case, we say that the underlying populations, or random variables, are *heteroscedastic* (sometimes spelled as heteroskedastic).

In this entry we will initially discuss the case when we compare variances of two populations, and subsequently will extend to k populations.

Comparison of Two Population Variances

The standard F-test is used to test whether two populations have the same variance. The test statistic for testing the hypothesis if ${\sigma_1}^2 = {\sigma_2}^2$ where ${\sigma_1}^2$ and ${\sigma_2}^2$ are the variances of two populations, is

$$F = \frac{{s_1}^2}{{s_2}^2}, \tag{1}$$

where ${s_1}^2$ and s_2 2 are the sample variances for two independent random samples of n_1 and n_2 observations from normally distributed populations with variances ${\sigma_1}^2$ and ${\sigma_2}^2$, respectively. If the null hypothesis is true (i.e., $H_0 : {\sigma_1}^2 = {\sigma_2}^2$), the test statistic has the F-distribution with $(n_1 - 1)$ degrees of freedom for the numerator and $(n_2 - 1)$ degrees of freedom for the denominator. The F-test is extremely sensitive to non-normality and should not be used unless there is strong evidence that the data do not depart from normality.

In practical applications, the F ratio in (1) is usually calculated so that the larger sample variance is in the numerator, that is, ${s_1}^2 > {s_2}^2$. Thus, F statistic is always greater than one and only the upper critical values of the F-distribution are used. At the significance level α, the test rejects the hypothesis that the variances are equal if $F > F_{(\alpha;n_1-1;n_2-1)}$, where $F_{(\alpha;n_1-1;n_2-1)}$ is the upper critical value of the F distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom.

Tests for Equality of Variances of *k* Populations

The **Bartlett's test** (Bartlett 1937) is used to test if k-groups (populations) have equal variances. Hypotheses are stated as follows:

$$H_0 : {\sigma_1}^2 = {\sigma_1}^2 = \ldots = {\sigma_k}^2$$

$$H_1 : {\sigma_i}^2 \neq {\sigma_j}^2 \quad \text{for at least one pair } (i, j).$$

To test for equality of variance against the alternative that variances are not equal for at least two groups, the test statistic is defined as

$$\chi^2 = \frac{(N-k)\ln\left(\dfrac{\sum_{i=1}^{k}(n_i-1){s_i}^2}{N-k}\right) - \sum_{i=1}^{k}(n_i-1)\ln\left({s_i}^2\right)}{1 + \dfrac{1}{3(k-1)}\left[\left(\sum_{i=1}^{k}\dfrac{1}{n_i-1}\right) - \dfrac{1}{N-k}\right]} \tag{2}$$

where k is the number of samples (groups), n_i is the size of the ith sample with sample variance ${s_i}^2$, and N is the sum of all samples sizes.

The test statistic follows a ▶chi-square distribution with $(k-1)$ degrees of freedom and the standard *chi-squared test* with $(k-1)$ degrees of freedom is applied.

The Bartlett's test rejects the null hypothesis that the variances are equal if $\chi^2 > \chi^2_{(\alpha,k-1)}$, where $\chi^2_{(\alpha,k-1)}$ is the upper critical value of the chi-square distribution with $(k-1)$ degrees of freedom and a significance level of α.

The test is very sensitive to departures from normality and/or to differences in group sizes and is not recommended for routine use. However, if there is strong evidence that the underlying distribution is normal (or nearly normal), the Bartlett's test has good performance.

The Levene's test (Levene 1960) is another test used to test if k groups have equal variances, as an alternative to

the Bartlett's test. It is less sensitive to departures from normality and/or to differences in group sizes and is considered to be the standard test for homogeneity of variances. The idea of this test is to transform the original values of the dependent variable Y and obtain a new variable known as the "dispersion variable." A standard ►analysis of variance based on these transformed values will test the assumption of homogeneity of variances.

The test has two options. Given a variable Y with sample of size N divided into k-subgroups, Y_{ij} will be the jth individual score belonging to the ith subgroup. The first option of the test is to define the transformed variable as the absolute deviation of the individual's score from the mean of the subgroup to which the individual belongs, that is, as $Z_{ij} = \left|Y_{ij} - \overline{Y}_{i.}\right|$ where $\overline{Y}_{i.}$ is the mean of the ith subgroup. The transformed variable is known as the dispersion variable, since it "measures" how far the individual is displaced from its subgroup mean.

The Levene's test statistic is defined as

$$F^L = \frac{(N-k)\sum_{i=1}^{k} n_i\left(\overline{Z}_{i.} - \overline{Z}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \overline{Z}_{i.}\right)^2} \tag{3}$$

where n_i is the sample size of the ith subgroup, $Z_{ij} = \left|Y_{ij} - \overline{Y}_{i.}\right|$ is the dispersion variable, $\overline{Z}_{i.}$ are the subgroup means of Z_{ij} and $\overline{Z}$ is the overall mean of Z_{ij}.

The test statistic follows the F-distribution with $(k-1)$ and $(N-k)$ degrees of freedom and the standard F-test is applied.

The Levene's test will reject the hypothesis that the variances are equal if $F^L > F^{\alpha}_{(k-1,N-k)}$ where $F^{\alpha}_{(k-1,N-k)}$ is the upper critical value of the F distribution with $(k-1)$ and $(N-k)$ degrees of freedom at the significance level α.

The second option is to define the dispersion variable as the square of the absolute deviation from the subgroup mean, that is, as $Z_{ij}^2 = \left|Y_{ij} - \overline{Y}_{i.}\right|^2$.

The Brown–Forsythe test (Brown and Forsythe 1974) is a modification of the Levene's test, based on the same logic, except that the dispersion variable Z_{ij} is defined as the absolute deviation from the subgroup median rather than the subgroup mean, that is, $Z_{ij} = |Y_{ij} - M_{i.}|$, where $M_{i.}$ is the median of the ith subgroup. Such a definition, based on medians instead of means, provides good robustness against many types of non-normal data while retaining good power, and is therefore recommended in practical applications.

The O'Brien test (O'Brien 1979) is a modification of the Levene's Z_{ij}^2 test. In the O'Brien test, the dispersion variable Z_{ij}^2 is modified in a way to include an additional scalar W (weight) to account for the suspected kurtosis of the underlying distribution. The dispersion variable in the O'Brien test is defined as

$$Z_{ij}^B = \frac{(W + n_i - 2)\, n_i Z_{ij}^2 - W\,(n_i - 1)\, s_i^2}{(n_i - 1)\,(n_i - 2)} \tag{4}$$

where Z_{ij}^2 is the square of the absolute deviation from the subgroup mean and n_i is the size of the ith subgroup with its sample variance s_i^2. W is a constant with values between 0 and 1 and is used to adjust the transformation. The most commonly used weight is $W = 0.5$, as suggested by O'Brien (1979).

The previously discussed tests are the tests that are mostly used in empirical research and easily available in most statistical software packages. However, there are also other homogeneity of variance tests, both parametric and nonparametric. Among them are Hartley's Fmax test, David's multiple test, and Cochran's G test (also known as Cochran's C test). The Bartlett–Kendall test (like Bartlett's test) uses log transformation of the variance to approximate the normal distribution. An example of a nonparametric test is the Sidney–Tukey test that uses ranks and the chi-square approximation. A good discussion on the topic can be found in Zhang (1998).

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Bartlett's Test
►Heteroscedasticity
►Variance

References and Further Reading

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Soc Lond A 160:268–282

Brown MB, Forsythe AB (1974) Robust test for equality of variances. J Am Stat Assoc 69:364–367

Levene H (1960) Robust tests for the equality of variance. In: Olkin I (ed) Contributions to probability and statistics. Stanford University Press, Paolo Alto, pp 278–292

O'Brien RG (1979) A general ANOVA method for robust tests of additive models for variances. J Am Stat Assoc 74:877–880

Zhang S (1998) Fourteen homogeneity of variance tests: when and how to use them. Paper presented at the annual meeting of the american educational research association, San Diego, California

Tests of Fit Based on The Empirical Distribution Function

Michael A. Stephens
Professor Emeritus
Simon Fraser University, Burnaby, BC, Canada

Introduction: Tests for Continuous Distributions

Suppose a random sample $x_1, x_2, \cdots, x_n$ is given and we wish to test H_0: the parent population is the (continuous) distribution $F(x;\theta)$, where θ is a vector of parameters. The empirical distribution function (EDF) of the sample is defined by

$$F_n(x) = n(x)/n,$$

where $n(x)$ is the number of x_i which are less than or equal to x. The goodness-of-fit tests to be discussed are EDF tests, that is, based on the discrepancy

$$Y(x) = F_n(x) - F(x;\theta)$$

The most well known are the Kolmogorov-Smirnov family:

$$D_n^+ = sup\ Y(x)$$
$$D_n^- = sup\{-Y(x)\}$$
$$D_n = sup|Y(x)|$$

and the Cramér-von Mises family:

$$W_n^2 = n\int_{-\infty}^{\infty} Y^2(x)dF(x;\theta)$$
$$U_n^2 = n\int_{-\infty}^{\infty}\left\{Y(x) - \int_{-\infty}^{\infty} Y(x)dF(x;\theta)\right\}^2 dF(x;\theta)$$

and

$$A_n^2 = n\int_{-\infty}^{\infty} Y^2(x)\psi(x)dF(x;\theta)$$
$$\text{where } \psi(x) = [F(x;\theta)(1-F(x;\theta))]^{-1}$$

Statistic W_n^2 is the original Cramér-von Mises statistic, originally called $n\omega^2$. Statistic U_n^2 was introduced by Watson (1961) for testing distributions around a circle; it has the merit that its value does not depend on the origin used for measuring the observations. Statistic A_n^2 is the Anderson-Darling (1952) statistic: it emphasises the tails of the tested distribution.

Statistic D_n was introduced by Kolmogorov (1933). Distribution theory for the Kolmogorov–Smirnov family is known for the case when parameters are known; but when parameters are unknown and must be estimated from the sample, even asymptotic theory is not available and significance points must be obtained by Monte Carlo. Tables of significance points for testing a number of distributions are in Stephens (1986).

The statistics D_n^+ and D_n^- have good power when the sample EDF lies mostly on one side of the tested distribution, but the D_n statistic, in general, is less powerful as an omnibus test than the Cramér-von Mises statistics. More information on this statistic is given by Lopes (2010) in an article in this Encyclopedia and here it will not be considered further.

The Probability Integral Transformation

In practice, it is easier to work with the EDF of the transformed set $z_{(i)} = F(x_{(i)};\theta), i = 1,\cdots,n$; where $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ is the ordered sample. This transformation is called the probability integral transformation (PIT). If θ is known, the $z_{(i)}$ are ordered uniform variates. If θ is not known, an efficient estimate (for example, the MLE) should be used for the transformation. The EDF statistics are easier to calculate from the z-values, as follows.

Let $F_n(z)$ be the empirical distribution function of the z-values, and define

$$y_n(z) = \sqrt{n}\{F_n(z) - z\}.$$

The Cramér-von Mises statistics now become, in terms of $y_n(z)$:

$$W_n^2 = \int_0^1 \{y_n(z)\}^2 dz, \quad (1)$$
$$U_n^2 = \int_0^1 \{y_n(z) - \bar{y}\}^2 dz, \quad (2)$$
$$A_n^2 = \int_0^1 \{y_n(z)\}^2 w(z)dz, \quad (3)$$

where

$$\bar{y} = \int_0^1 y_n(z)dz \qquad \text{and} \qquad w(z) = 1/(z-z^2).$$

The computing formulas are

$$W_n^2 = \sum\left\{z_{(i)} - (2i-1)/2n\right\}^2 + 1/(12n) \quad (4)$$
$$U_n^2 = W^2 - n(\bar{z} - 0.5)^2 \quad (5)$$

and

$$A_n^2 = -n - (1/n)\sum(2i-1)\left\{\ln(z_{(i)}) + \ln(1 - z_{(n+1-i)})\right\}. \quad (6)$$

The distributions of these statistics, when estimated parameters are location or scale, will depend on the tested distribution, but not on the true values of the parameters. However, when an unknown parameter is a shape parameter, the distribution will depend on the shape.

Asymptotic theory of these statistics was first given by Anderson and Darling (1952), and Darling (1955); see also Anderson (2010), an entry in this Encyclopedia. Stephens (1976) used the theory to give significance points for tests of normality and exponentiality; points for other distributions are in Stephens (1986) and in Lockhart and Stephens (1985, 1994).

In general, W^2 and A^2 have been shown to be powerful in testing many distributions; A^2 has comparable power to the Shapiro-Wilk statistic for testing normality.

Tests for Discrete Distributions

EDF tests may be adapted for testing discrete distributions, by comparing the cumulated histogram of observed numbers in the cells with the cumulated histogram of the expected numbers. Choulakian et al. (1994) have given distribution theory for the Cramér-von Mises family when parameters are known, and Lockhart et al. (2007) have discussed the case when parameters must be estimated from the sample; see Stephens (2010), an entry in this Encyclopedia. These statistics are generally more powerful than Pearson's χ^2 statistic.

About the Author

For biography *see* the entry ►Cramér-Von Mises Statistics for Discrete Distributions.

Cross References

►Anderson-Darling Tests of Goodness-of-Fit
►Cramér-Von Mises Statistics for Discrete Distributions
►Exact Goodness-of-Fit Tests Based on Sufficiency
►Kolmogorov-Smirnov Test
►Normality Tests
►Parametric and Nonparametric Reliability Analysis

References and Further Reading

Anderson TW (2010) Anderson–Darling tests of goodness-of-fit. In: International encyclopedia of statistical science, Springer-Verlag, Berlin

Anderson TW, Darling DA (1952) Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann Math Stat 23:193–212

Choulakian V, Lockhart RA, Stephens MA (1994) Cramer-von Mises tests for discrete distributions. Can J Stat 22:125–137

Darling DA (1955) The Cramér-Smirnov test in the parametric case. Ann Math Stat 26:1–20

Kolmogorov AN (1933) Sulla determinazione empirica di una legge di dis-tribuzione. Giorna Ist Attuari 4:83–91

Lockhart RA, Spinelli JJ, Stephens MA (2007) Cramér-von Mises statistics for discrete distributions with unknown parameters. Can J Stat 35:125–133(9)

Lockhart RA, Stephens MA (1985) Tests of fit for the Von Mises distribution. Biometrika 72:647–652

Lockhart RA, Stephens MA (1994) Estimation and tests of fit for the three-parameter Weibull distribution. J Roy Stat Soc B 56: 491–500

Lopes RHC (2010) Kolmogorov-Smirnov test. International Encyclopedia of Statistical Science, Springer-Verlag, Berlin

Stephens MA (1976) Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann Stat 4:357–369

Stephens MA (1986) Tests based on EDF statistics, Chap 4. In: D'Agostino R, Stephens MA (eds) Goodness-of-fit techniques. Marcel Dekker, New York

Stephens MA (2010) EDF tests of fit. International Encyclopedia of Statistical Science, Springer-Verlag, Berlin

Watson GS (1961) Goodness-of-fit tests on a circle, 1. Biometrika 48:109–114

Tests of Independence

Bruno Rémillard
Professor
HEC Montréal, Montréal, QC, Canada

Testing for Interdependence

Testing independence between two of more components of a random vector is an important problem in statistics. For sake of simplicity, suppose that the law of each component is continuous. In the bivariate case, for testing independence between random variables X_1 and X_2, most of the tests proposed initially were based on some dependence measure ρ, taking usually value 0 under the null hypothesis of independence. Once a random sample $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ is collected, that is, the pairs (X_{i1}, X_{i2}), $i = 1, \ldots, n$, are independent observations of (X_1, X_2), an estimator $\hat{\rho}_n$ of ρ is obtained and it is compared with the value of ρ under the null hypothesis. In general, $\hat{\rho}_n$ must be a "good" estimator of ρ in the sense that as $n \to \infty$, $n^{1/2}(\hat{\rho}_n - \rho) \rightsquigarrow N(0, \sigma_0^2)$, where "$\rightsquigarrow$" denotes convergence in law, and σ_0 is the limiting variance of $n^{1/2}\hat{\rho}_n$. The most known example is the one based on the Pearson correlation coefficient, defined by

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X, Y)\mathrm{Var}(X, Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\{E(X^2) - E^2(X)\}\{E(Y^2) - E^2(Y)\}}},$$

provided $E(X^2)$ and $E(Y^2)$ are finite. In that case,

$$\hat{\rho}_n = r_n = \frac{\sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sqrt{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2}\sqrt{\sum_{i=1}^n (X_{i2} - \bar{X}_2)^2}}.$$

Under the null hypothesis of independence, $\rho = 0$ and $n^{1/2} r_n \rightsquigarrow N(0,1)$, as $n \to \infty$. If in addition the joint distribution of (X_1, X_2) is Gaussian, then $\frac{r_n}{\sqrt{(1-r_n^2)/(n-2)}}$ has a Student distribution with $n-2$ degrees of freedom.

Many other popular empirical measures of dependence are based on ranks. Recall that the ranks R_{ij}, $i = 1, \ldots, n$, $j = 1, 2$, are defined as follows: R_{i1} is the rank of X_{i1} amongst $X_{11}, \ldots, X_{n1}$, while R_{i2} is the rank of X_{i2} amongst $X_{12}, \ldots, X_{n2}$, and so on, where the smallest observation has rank 1. In particular, these measures do not depend on the margins, only on the so-called copula (see ▶Copulas). That notion will be defined later. The most known rank-based measures of dependence are ▶Kendall's tau and Spearman's rho. Kendall's tau is defined by

$$\tau_n = \frac{2}{n(n-1)}(C_n - D_n),$$

where C_n is the number of concordant pairs of ranks, and D_n is the number of discordant pairs, the pairs (R_{i1}, R_{i2}) and (R_{j1}, R_{j2}) being concordant if $(R_{i1} - R_{j1})(R_{i2} - R_{j2}) > 0$ and discordant otherwise. Recall that τ_n is an estimation of $\tau = 2P\{(X_1 - Y_1)(X_2 - Y_2) > 0\} - 1$, where (Y_1, Y_2) is an independent copy of (X_1, X_2). Under the null hypothesis of independence, $\tau = 0$ and it can be shown that $n^{1/2}\tau_n \rightsquigarrow N(0, 4/9)$, as $n \to \infty$.

Spearman's rho, denoted by ρ_n^S, is simply defined as the correlation between the ranks $(R_{11}, R_{12}), \ldots, (R_{n1}, R_{n2})$. Then ρ_n^S is an estimator of ρ^S, the correlation between $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$, where F_j is the distribution function of X_j, $j = 1, 2$. Under the null hypothesis of independence, $\rho^S = 0$ and $n^{1/2}\rho_n^S \rightsquigarrow N(0,1)$, as $n \to \infty$.

All tests based on a single measure of dependence usually have the same weakness: They are not consistent for testing independence in the sense that under some alternatives, the power of the test does not tend to 1 as the sample size tends to infinity. One such example of alternative is the following: Let X_1 be uniformly distributed over $(0,1)$, denoted by $X_1 \sim \text{Unif}(0,1)$ and set $X_2 = T(X_1)$, where T is the tent map, i.e., $T(u) = 2\min(u, 1-u)$. Then $X_2 \sim \text{Unif}(0,1)$ and X_1 and X_2 are strongly dependent. However, for any of the three measures of dependence ρ stated previously, the value of ρ is 0, the same value as for independence, and it can be shown than $n^{1/2}\hat{\rho}_n \rightsquigarrow N(0, \sigma^2)$, for some $\sigma > 0$ depending on ρ. As a result, the power of the associated test of level 5% tends to $2\Phi\left(-1.96\frac{\sigma_0}{\sigma}\right)$, where σ_0^2 is the asymptotic variance under the null hypothesis of independence and Φ is the distribution function of the standard Gaussian. For example, in the case of the Pearson correlation, $\sigma_0^2 = 1$ and $\sigma^2 = 6/5$, so the power tends to 0.1024, as $n \to \infty$.

To overcome the inconsistency problem, it was suggested in Blum et al. (1961) to use statistics based on the empirical distribution function. More precisely, in the bivariate case, one can compare the joint empirical distribution function H_n, given by

$$H_n(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_{i1} \le x_1, X_{i2} \le x_2)$$

with the product of its margins, i.e., $F_{n1}(x_1) = H_n(x_1, \infty)$ and $F_{n2}(x_2) = H_n(\infty, x_2)$. It can then be shown that $\mathbb{H}_n(x_1, x_2) = n^{1/2}\{H_n(x_1, x_2) - F_{n1}(x_1)F_{n2}(x_2)\} \rightsquigarrow \mathbb{H}(x_1, x_2)$, where the convergence is in the Skorohod space $\mathcal{D}\left([-\infty, +\infty]^2\right)$ and $\mathbb{H}(x_1, x_2) = \mathbb{B}\{F_1(x_1), F_2(x_2)\}$, where F_1 and F_2 are the margins of the joint distribution function H of (X_1, X_2), and $\mathbb{B}$ is a continuous centered Gaussian process with covariance function

$$\begin{aligned}\Gamma(u_1, u_2, v_1, v_2) &= \text{Cov}\{\mathbb{B}(u_1, u_2), \mathbb{B}(v_1, v_2)\} \\ &= \{\min(u_1, v_1) - u_1 v_1\} \\ &\quad \times \{\min(u_2, v_2) - u_2 v_2\}.\end{aligned}$$

Recall that by Sklar (1959), when the marginal distributions are continuous, there exists a unique distribution function C with uniform margins, called a copula, so that

$$\begin{aligned}H(x_1, x_2) &= P(X_1 \le x_1, X_2 \le x_2) \\ &= C\{F_1(x_1), F_2(x_2)\}, \qquad x_1, x_2 \in \mathbb{R}.\end{aligned}$$

Thus X_1 and X_2 are independent if and only if the copula is the independence copula $C_\perp$ defined by

$$C_\perp(u_1, u_2) = u_1 u_2, \qquad u_1, u_2 \in [0,1].$$

That relationship lead Deheuvels (1981) to proposed tests of interdependence based on the empirical copula C_n, where

$$C_n(u_1, u_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\frac{R_{i1}}{n} \le u_1, \frac{R_{i2}}{n} \le u_2\right), \quad u_1, u_2 \in [0,1].$$

The empirical copula seems to have been studied first by Rüschendorf (1976).

To tackle the d-dimensional case, $d > 2$, where the covariance of the limiting process $\mathbb{H}$ under independence is much more intricate than when $d = 2$, Blum et al. (1961) proposed a decomposition of $\mathbb{H}_n$ based on Möbius formula, leading to processes $\mathbb{H}_{n,A}$, for all $A \subset \{1, \ldots, d\}$, so that each process $\mathbb{H}_{n,A}$ is asymptotically independent of the

others and where the covariance is similar to one obtained in the bivariate case. More precisely, the covariance of the continuous centered limiting processes $\mathbb{H}_A$ is given by

$$\mathrm{Cov}\{\mathbb{H}_A(x), \mathbb{H}_A(y)\} = \prod_{j \in A} [\min\{F_j(x_j), F_j(y_j)\} - F_j(x_j)F_j(y_j)], \quad x, y \in \mathbb{R}^d.$$

That decomposition then appeared in Deheuvels (1981) for copulas, but the author came short of proposing tests of independence. That decomposition was then rediscovered by Ghoudi et al. (2001), who where also able to test independence between non-observable error terms in regression models, using the residuals. With the notable exception of the regression case, testing independence using residuals or more generally pseudo-observations can be quite difficult. See, e.g., Ghoudi and Rémillard (2004). Building on the previous work, Genest and Rémillard (2004) applied the Möbius decomposition method to empirical copulas to test interdependence and serial dependence. That lead them to define the so-called "dependogram." The work of Genest and Rémillard (2004) has been extended recently by Beran et al. (2007) and Kojadinovic and Holmes (2009) for testing independence between random vectors. In addition to test statistics constructed from empirical distribution functions, some researchers considered empirical ▶characteristic functions. See, e.g., Feuerverger (1993), Bilodeau and Lafaye de Micheaux (2005), and more recently Székely and Rizzo (2010). Because independence can be characterized in terms of characteristic functions, the associated tests are consistent in general.

Finally it is worth mentioning Genest and Rémillard (2004), Genest et al. (2006) and Genest et al. (2007) where power comparisons were made for tests of interdependence, the last two for Cramér-von Mises type test statistics.

Testing for Serial Independence

The treatment of serial dependence in (stationary) time series is almost the same as in the previous case, few modifications being necessary for taking into account their particular nature. In fact, if $Y_1, \ldots, Y_n$ represent the time series values for n consecutive periods, then in the bivariate case, one just have to define $X_{i1} = Y_i$ and $X_{i2} = Y_{i+\ell}$, for some lag $\ell \geq 1$. Then the correlation is called autocorrelation of lag ℓ, etc. The so-called correlogram of order k, introduced by Wold in his 1938 Ph.D. thesis, is the graph of the autocorrelations for lags $\ell = 1, \ldots, k$. Under the null hypothesis of serial independence, $n^{1/2} r_n(1), \ldots, n^{1/2} r_n(k)$ converge jointly to independent standard Gaussian variables. One can also adapt the rank-based measures to time series context. More precisely, if $R_1, \ldots, R_n$ are the ranks of $Y_1, \ldots, Y_n$, then one can measure dependence between the pairs $(R_i, R_{i+\ell})$, $i = 1, \ldots, n - \ell$. For more details on rank-based measures of dependence and their properties, see e.g., Hallin et al. (1985) and Ferguson et al. (2000). As before, the tests based on autocorrelations or rank-based measures are not consistent in general, so Skaug and Tjøstheim (1993) proposed to adapt the empirical distribution function methodology to time series context. More precisely, they considered the joint distribution function

$$\tilde{H}_n(x_1, x_2) = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbb{I}(Y_i \leq x_1, Y_{i+\ell} \leq x_2)$$

which was compared to $\tilde{F}_n(x_1)\tilde{F}_n(x_2)$, where $\tilde{F}_n(x) = \tilde{H}_n(x, \infty)$. It is remarkable that the limiting distribution of $n^{1/2}\{\tilde{\mathbb{H}}_n(x_1, x_2) - \tilde{F}_n(x_1)\tilde{F}_n(x_2)\}$ is the same as the limiting distribution of $\mathbb{H}_n$, defined in the previous section. That property was extended by Genest and Rémillard (2004) to the multivariate case, using the associated empirical copula and Möbius decomposition. Other work using ▶empirical processes in a serial context includes Genest et al. (2002) and Kojadinovic and Yan (2010).

Finally, one important problem in time series is checking the serial independence of the non-observable innovations, which is often considered as a test of adequacy for the underlying model. Unfortunately, in most applications, replacing the innovations by residuals changes completely the limiting distribution. See, e.g., Ghoudi and Rémillard (2004). However, using an idea of Brock et al. (1996), Genest et al. (2007) were able to propose tests of independence so that their limiting distribution was not affected by using residuals instead of innovations. However the type of models covered by their methodology is limited to additive models, so it does not include GARCH models.

About the Author

Bruno Rémillard is a Full Professor in Financial Engineering at HEC Montréal since 2001. After completing a Ph.D. in Probability at Carleton University, he was a postdoctoral fellow at Cornell University, before being a professor of Statistics at Université du Québec à Trois-Rivières. He is the author or co-author of more than fifty research articles in Probability, Statistics and Financial Engineering. In 1987, he received the Pierre-Robillard award for the best Ph.D. thesis in Probability and Statistics in Canada and in 2003, he received the prize for the best paper of the year published in the *Canadian Journal of Statistics*. He is also a consultant in the Research and Development group at Innocap since 2007, an alternative investment firm located in Montreal, owned in part by BNP-Paribas and National

Bank of Canada, where he mainly helps developing and implanting new quantitative methods for alternative and traditional portfolios.

Cross References

►Asymptotic Relative Efficiency in Testing
►Autocorrelation in Regression
►Bivariate Distributions
►Categorical Data Analysis
►Copulas
►Copulas: Distribution Functions and Simulation
►Correlation Coefficient
►Durbin–Watson Test
►Kendall's Tau
►Measures of Dependence

References and Further Reading

Beran R, Bilodeau M, Lafaye de Micheaux P (2007) Nonparametric tests of independence between random vectors. J Multivar Anal 98(9):1805–1824

Bilodeau M, Lafaye de Micheaux P (2005) A multivariate empirical characteristic function test of independence with normal marginals. J Multivar Anal 95:345–369

Blum JR, Kiefer J, Rosenblatt M (1961) Distribution free test of independence based on the sample distribution function. Ann Math Stat 32:485–498

Brock WA, Dechert WD, LeBaron B, Scheinkman JA (1996) A test for independence based on the correlation dimension. Econom Rev 15:197–235

Deheuvels P (1981) An asymptotic decomposition for multivariate distribution-free tests of independence. J Multivar Anal 11:102–113

Ferguson TS, Genest C, Hallin M (2000) Kendall's tau for serial dependence. Can J Stat 28:587–604

Feuerverger A (1993) A consistent test for bivariate dependence. Int Stat Rev 61:419–433

Genest C, Ghoudi K, Rémillard B (2007) Rank-based extensions of the Brock Dechert Scheinkman test for serial dependence. J Am Stat Assoc 102:1363–1376

Genest C, Quessy J-F, Rémillard B (2002) Tests of serial independence based on Kendall's process. Can J Stat 30:441–461

Genest C, Quessy J-F, Rémillard B (2006) Local efficiency of a Cramér-von Mises test of independence. J Multivar Anal 97:274–294

Genest C, Quessy J-F, Rémillard B (2007) Asymptotic local efficiency of Cramér-von Mises tests for multivariate independence. Ann Stat 35:166–191

Genest C, Rémillard B (2004) Tests of independence or randomness based on the empirical copula process. Test 13:335–369

Ghoudi K, Kulperger RJ, Rémillard B (2001) A nonparametric test of serial independence for time series and residuals. J Multivar Anal 79:191–218

Ghoudi K, Rémillard B (2004) Empirical processes based on pseudoobservations. II. The multivariate case. In Asymptotic methods in stochastics, Vol 44 of fields institute communications. American Mathematical Society, Providence, RI, pp 381–406

Hallin M, Ingenbleek J-F, Puri ML (1985) Linear serial rank tests for randomness against ARMA alternatives. Ann Stat 13:1156–1181

Kojadinovic I, Holmes M (2009) Tests of independence among continuous random vectors based on cramér-von mises functionals of the empirical copula process. J Multivar Anal 100(6): 1137–1154

Kojadinovic I, Yan J (2010) Tests of serial independence for continuous multivariate time series based on a Möbius decomposition of the independence empirical copula process. Ann Inst Stat Math

Rüschendorf L (1976) Asymptotic distributions of multivariate rank order statistics. Ann Stat 4(5):912–923

Skaug HJ, Tjøstheim D (1993) A nonparametric test of serial independence based on the empirical distribution function. Biometrika 80:591–602

Sklar M (1959) Fonctions de répartition á n dimensions et leurs marges. Publ Inst Stat Univ Paris 8:229–231

Székely GJ, Rizzo ML (2010) Brownian distance covariance. Ann Appl Stat 3(4):1236–1265

Time Series

Peter J. Brockwell
Professor Emeritus
Colorado State University, Fort Collins, CO, USA

A central goal of science, and indeed of a great number of human activities, is to make use of current information in order to obtain useful forecasts of what may happen in the future. If the future is completely independent of the currently available information then this information is of no help. However if there is dependence then we would like to use it to make forecasts which are as accurate as possible in some specified sense. This is one of the key goals of time series analysis (although there are others as we shall see).

A *time series* is a set of observations $\{x_t\}$, each one associated with a particular time t and usually displayed in a *time series plot*, i.e., a graph of x_t as a function of t. An example is the following graph of the natural logarithm of the daily closing value in US dollars of the Dow-Jones Industrial Average, plotted for successive trading days from August 1st, 1997 until August 5th, 2003.

In general the set of times T at which observations are recorded may be a discrete set, as is the case when observations are made at uniformly spaced times (e.g., daily rainfall, annual income etc.) or it may be a continuous interval. For reasons of space we shall restrict attention here to observations at uniformly spaced times, in which case we can label the times $1, 2, \ldots$. In order to account for randomness, we suppose that for each t the observation x_t is just one of many possible values of a random variable X_t that we *might* have observed at time t. The term *time*

series is frequently used to denote both the sequence of random variables $\{X_1, X_2, \ldots\}$ and the particular sequence of observed values $\{x_1, x_2, \ldots\}$.

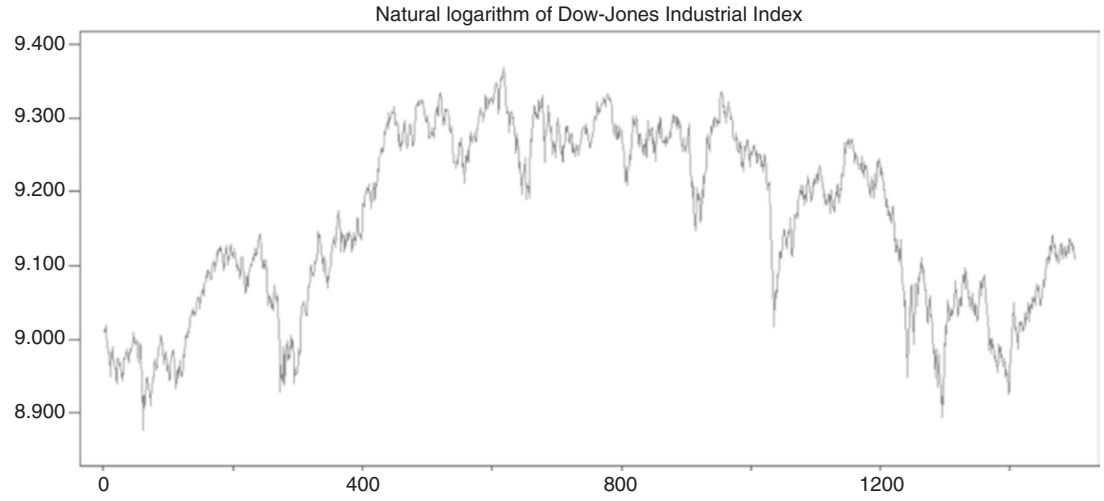
To illustrate the general problem of forecasting in concrete terms, suppose we have a sequence of jointly distributed random variables $\{X_1, X_2, \ldots\}$. Such a sequence is known as a *time series indexed by the positive integers*. Suppose also that our 'information' at time n consists of the observed values of $X_1, \ldots, X_n$. Our problem then is to predict X_{n+h}, the value of the random sequence at the future time $n + h$ using some suitably chosen function $\hat{X}_{n+h}$ of $(X_1, \ldots, X_n)$. In order to assess the performance of our forecast we need some measure of the error of $\hat{X}_{n+h}$. An especially convenient measure, if $EX_n^2 < \infty$ for all n, is the expected squared error, namely $E(X_{n+h} - \hat{X}_{n+h})^2$. Then a rather simple calculation shows that the *best* forecast, i.e., the function of $(X_1, \ldots, X_n)$ which minimizes the expected squared error is the conditional expectation, $E(X_{n+h}|X_1, \ldots, X_n)$. Unfortunately the calculation of this conditional expectation requires knowledge of the conditional distribution of X_{n+h} given $(X_1, \ldots, X_n)$ which is generally unknown and also difficult to estimate from data. (If $\{X_1, X_2, \ldots\}$ is an independent sequence then the conditional expectation is independent of $\{X_j, j \le n\}$, showing that the current information at time n is of no help in predicting X_{n+h} in this case. Time series is therefore primarily concerned with *dependent* random variables and the analysis and utilization of this dependence.) A simpler approach to forecasting X_{n+h} is to look for the *linear combination*, $\hat{X}_{n+h} = a_0 + a_1 X_n + \cdots + a_n X_1$ which minimizes the expected squared error $E(X_{n+h} - \hat{X}_{n+h})^2$. This is a much simpler problem, the solution of which depends only on the expected values EX_i and EX_iX_j, $i, j = 1, 2, \ldots$ Moreover if the joint distribution of $(X_1, X_2, \ldots, X_k)$ is multivariate normal for every positive integer k then this *best linear forecast* is the same as the best forecast.

Forecasting is just one of the many objectives of time series analysis. These depend on the particular field of application. For example, from observed values $x_1, x_2, \ldots$ of the random variables $X_1, X_2, \ldots$ we may wish to understand the mechanism generating the data or perhaps to extract a deterministic 'signal' in the data which is masked by the presence of random noise. We may simply wish to find a compact representation of the available observations or to find a mathematical model which appears to represent the observations well and to use it to simulate further realizations of the series.

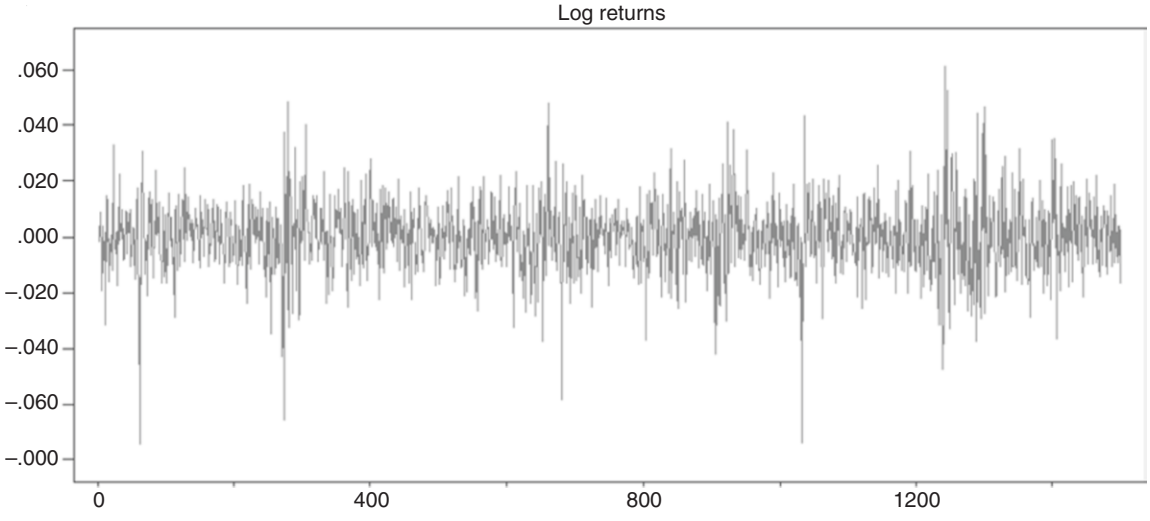
For these applications we need to find a mathematical model which gives a good representation of the data. Typically we select the best-fitting member of a specified family of models by estimating parameters from the observed data and then testing the goodness of fit of the model to the data. Once we are satisfied that the selected model is satisfactory we use it to address the questions of interest. Complete specification of a model for the time series $\{X_1, X_2, \ldots\}$ would consist of a specification of the joint distribution of $(X_1, \ldots, X_k)$ for every positive integer k. However if we are concerned with issues (such as best linear prediction) which depend only on of first and second order moments of the time series, then a model which specifies only first and second-order moments will suffice.

Much of time series analysis is concerned with *stationary* time series. It is clear that if we wish to make predictions, we must assume that *something* does not vary with time. In extrapolating deterministic functions it is common practice to assume that either the function itself or one of its derivatives is constant. The assumption of a constant first derivative leads to linear extrapolation as a means of prediction. In time series we need to predict a series that is typically not deterministic but which contains a random component. The concept of stationarity is used to extend the notion of constancy in time to incorporate randomness. Strict stationarity of the series $\{X_n\}$ means that $(X_1, \ldots, X_k)$ has the same joint distribution as $(X_{h+1}, \ldots, X_{h+k})$ for all positive integers h and k. Weak stationarity means that EX_j and $E(X_{j+h}X_j)$ exist and are both independent of j. Thus stationarity requires the probabilistic properties (or, in the case of weak stationarity, the first and second moment properties) of the series to be invariant to shifts along the time axis. Information concerning the properties of stationary processes and estimation of their parameters can be found in the many books dealing with time series analysis. Without the assumption of stationarity the formulation of appropriate models and estimation of their parameters becomes much more difficult, although in recent years progress has been made in this direction.

The practical importance of stationary processes lies in the fact that many empirically observed series, which themselves cannot be well fitted by a stationary time series model, can be simply transformed into a new series which can. If a stationary model is fitted to the transformed series, it can be used to generate forecasts of the transformed series which can then be transformed back to generate corresponding forecasts for the original series. For example if we denote by X_n the natural logarithm of the closing value of the Dow-Jones Index on day n and consider the differenced series $Y_n := X_n - X_{n-1}$ (known as the *log return* for day n) then Y_n can be rather well represented as a stationary time series. The realization of the series $\{Y_n\}$ corresponding to the realization of $\{X_n\}$ in Fig. 1 is shown in Fig. 2.



Time Series. Fig. 1 The natural logarithm of the daily closing Dow-Jones Industrial Average for successive trading days from August 1st, 1997 until August 5th, 2003



Time Series. Fig. 2 The daily log returns for the Dow-Jones Industrial Average for successive trading days from August 1st, 1997 until August 5th, 2003

The dependence between observations of a stationary time series $\{X_n\}$ is frequently measured by the *autocovariance function*,

$$\gamma(h) := E[(X_{t+h} - \mu)(X_t - \mu)],$$

where $\mu := EX_t$, or the *autocorrelation function*,

$$\rho(h) := \gamma(h)/\gamma(0),$$

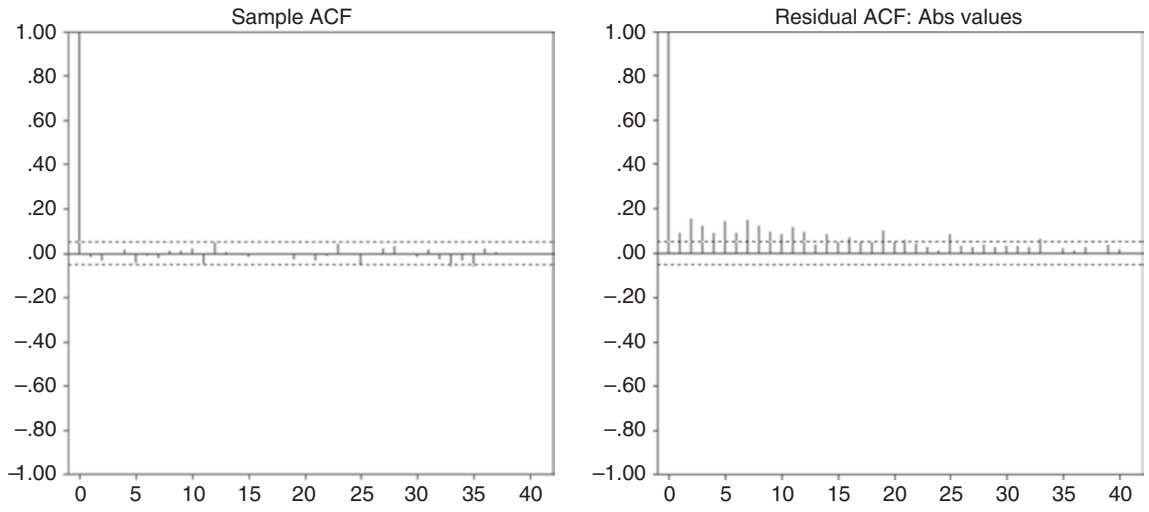
which specifies the correlation between any two observations separated by a time interval of length h. These quantities can be estimated by the *sample autocovariance function*,

$$\hat{\gamma}(h)) = n^{-1} \sum_{j=1}^{n-h} (x_{j+h} - \overline{x})(x_j - \overline{x}),$$

and *sample autocorrelation function*,

$$\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0),$$

respectively, where $\overline{x}$ denotes the sample mean, $n^{-1} \sum_{j=1}^{n} x_j$.



Time Series. Fig. 3 The sample autocorrelation function of the log returns in Fig. 2 (left) and the absolute values of the log returns (right)

The graph on the left of Fig. 3 shows the sample autocorrelation function of the differenced series in Fig. 2 with 95% significance bounds for testing the deviation of each sample autocorrelation value from zero. As there is no autocorrelation significantly different from zero from lags 1 through 40, it appears that the differenced series is uncorrelated. The best *linear* forecast of any future difference is therefore equal to the estimated mean of the differences (which is actually 0.0007). The best linear forecast of the natural logarithm of the Dow-Jones Industrial Average h trading days after August 5th, 2003 is therefore the value on August 5th (9.1090) plus $0.0007h$.

The autocorrelations in this example however do not tell the whole story. If the series of differences, instead of being merely uncorrelated with mean 0.0007, had been *independent*, then the mean value would have been the *best* rather than just the best *linear* forecast of future differences. However the graph on the right of Fig. 3, the sample autocorrelation function of the *absolute values* of the differences is clearly significantly different from zero at a number of lags. Since this implies that the absolute differences are not independent, it implies also that the differences themselves are not independent. This phenomenon of dependence with negligible correlation is a striking feature of many financial time series and has led to the development of a variety of intriguing models such as ARCH and GARCH models to account for this and related phenomena.

Probably the most widely used models for stationary time series have been the so-called ARMA (or autoregressive moving average) models. The series $\{X_n, n = 0, \pm 1, \pm 2, \ldots\}$ is said to be an ARMA(p, q) process if it is a stationary solution of the linear difference equations,

$$X_n - \phi_1 X_{n-1} - \cdots - \phi_p X_{n-p} = Z_1 + \theta_1 + \cdots \theta_q Z_{t-q},$$

where $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ are real valued coefficients, $\phi_p \neq 0$, $\theta_q \neq 0$, and $\{Z_n\}$ is a sequence of independent (or sometimes just uncorrelated) random variables, each with mean 0 and variance σ^2. Depending on the values of p and q and the coefficients ϕ_j and θ_j, an enormous range of sample autocorrelation functions can be replicated by members of the ARMA family. There is a vast literature dealing with problems of ▶model selection, estimation and forecasting for these processes. A standard technique (developed and popularized by Box and Jenkins) for dealing with observed time series which appear to be non-stationary is to apply differencing until the data appears to be representable by a stationary model and then to fit an ARMA model to the resulting series. The original data is then said to be represented by an ARIMA (or integrated ARMA) model.

In the last thirty years there has been an explosion of interest in more elaborate non-linear models to account for phenomena which cannot be accounted for in the classical linear framework provided by ARMA models. These include threshold, bilinear, ARCH, GARCH, Markov switching models and many others too numerous to be discussed here in any detail. Details can be found in some of the following references.

Acknowledgments

Work supported by NSF Grant DMS-0744058.

About the Author

Professor Brockwell is Associate Editor of the *Journal of the Japanese Statistical Society* and of the *Annals of the Institute of Statistical Mathematics*. He is a Fellow of the Institute of Mathematical Statistics, the American Statistical Association, and member of the International Statistical Institute. He was Von Neumann Guest Professor, Technical University of Munich (2001–2002). He is coauthor, with Richard Davis, of two widely used texts on Time Series Analysis: *Introduction to Time Series and Forecasting* (2nd edition, Springer, 2002), and *Time Series: Theory and Methods* (2nd edition, Springer, 1991).

Cross References

►Bayesian Approach of the Unit Root Test
►Box–Jenkins Time Series Models
►Business Forecasting Methods
►Data Mining Time Series Data
►Detecting Outliers in Time Series Using Simulation
►Detection of Turning Points in Business Cycles
►Dickey-Fuller Tests
►Exponential and Holt-Winters Smoothing
►Forecasting Principles
►Forecasting with ARIMA Processes
►Forecasting: An Overview
►Heteroscedastic Time Series
►Intervention Analysis in Time Series
►Median Filters and Extensions
►Models for Z_+-Valued Time Series Based on Thinning
►Nonlinear Time Series Analysis
►Optimality and Robustness in Statistical Forecasting
►Seasonal Integration and Cointegration in Economic Time Series
►Seasonality
►Singular Spectrum Analysis for Time Series
►Statistical Aspects of Hurricane Modeling and Forecasting
►Structural Time Series Models
►Time Series Models to Determine the Death Rate of a Given Disease
►Time Series Regression
►Trend Estimation
►Vector Autoregressive Models

References and Further Reading

Anderson TW (1971) The statistical analysis of time series. Wiley, New York
Box GEP, Jenkins GM, Reinsel GC (2008) Time series analysis: forecasting and control, 4th edn. Wiley, New York
Brockwell PJ, Davis RA (2002) Introduction to time series and forecasting, 2nd edn. Springer-Verlag, New York
Brockwell PJ, Davis RA (1991) Time series: theory and methods, 2nd edn. Springer-Verlag, New York
Fuller WA (1995) Introduction to statistical time series, 2nd edn. Wiley, New York
Hannan EJ (1970) Multiple time series. Wiley, New York
Lütkepohl H (1993) Introduction to multiple time series analysis, 2nd edn. Springer-Verlag, Berlin
Priestley MB (1981) Spectral analysis and time series. Academic Press, London
Shumway RH, Stoffer DS (2006) Time series analysis and its applications with R examples, 2nd edn. Springer-Verlag, New York
Tong H (1990) Non-linear time series: a dynamical systems approach. Oxford University Press, Oxford
Tsay RS (1990) Analysis of financial time series. Wiley, New York

Time Series Models to Determine the Death Rate of a Given Disease

Dahud K. Shangodoyin
Associate Professor
University of Botswana, Gaborone, Botswana

Statistics as a scientific subject of decision making under uncertainty is critical to the evaluation of health indicators that are of paramount importance to public health. Health issues, in most cases, are nondeterministic, which leaves their study to use the most suitable probabilistic approaches. Statistical research in health can be conducted in the following areas:

(a) ►Meta-analysis: Meta-analysis mathematically combine the results of numerous studies in order to improve the reliability of the results. Studies chosen for inclusion in a meta-analysis must be sufficiently similar in a number of characteristics in order to accurately combine their results; for instance, issues surrounding meta-analyses of individual patient data could be analyzed, and missing data can be dealt with at the patient level.

(b) Statistical Epidemiology: This aspect is broad and includes the following: (i) Clustered observational studies in which sample clusters of people are utilized for health research. This is becoming increasingly common, especially with patients in various health practices, people within health districts, and children within schools. The hierarchical nature of the data then takes on a multi-level structure that needs to be accounted for in the analysis. (ii) Ecological studies are carried out at an aggregate level, for example, the ward or district level, and can be used to investigate the relationship between socio-economic risk factors and ill-health. (iii) Longitudinal studies are useful

because following people over time is costly and time consuming and may have problems of missing data and consistency of measurement over time. Research of interest in this area could include a matched cohort study of coping and depression in parents of children newly diagnosed with terminal diseases.

(c) Survival Analysis: This is the analysis of time-to-event data, and is relevant to many clinical studies where the outcome of interest relates to the time taken for some event to occur, for instance, time to first seizure or time to death following ▶randomization.

The most important aspect of survival analysis is the measures of health indicator, especially the study of death rate for emerging and re-emerging diseases. Deliberation is continuing on how best to estimate the death rate of an emerging contagious disease, which is of paramount importance to global public health. The 2009 outbreak of influenza caused by a novel influenza A (H1N1) virus has given the World Health Organization (WHO) concern on how best to estimate the death rate arising from H1N1 throughout the world. As a matter of fact, the case of estimating the global death rate arising from the outbreak of severe acute respiratory syndrome (SARS) in 2003 also generated much public controversy (Altman (2003)). The WHO's convectional formula for computation of death rate is simply the ratio of the number of known deaths to the total number of confirmed cases (Mathers and Loncar (2006)), however, this formula is likely to underestimate the true death rate because the outcomes of many cases were still unknown or uncertain at the time these figures might have being compiled. In other words, the WHO approach has a problem of "selection" bias because the conditional probability of death among cases of known outcomes need not be equal to the unconditional probability of death. Another notable model of estimating the death rate of an emerging disease is the cohort approach. In this model cases from the same day constitute a cohort and the binomial analysis is restricted to the cases from a complete cohort, that is, cases with a known outcome at the end of the study period. The restrictions in this model lead to loss of a substantial volume of data and require some data that may not be accessible to researchers. The generalized mixed effect model of estimating death rate discussed by Chan and Tong (2006) is less biased and converges quickly to the death rate computed from the complete data, but the model specified leads to a singular precision matrix for the unknown parameters. In addition, the choice of the singular value decomposition presented may restrain this approach for practical use. Chang and Tong concluded that further research was needed on how to carry out the estimation of the conditional mean death rate with the constraint that the estimated death rate should be greater than or equal to zero. Shangodoyin (2009) proffers another method by using a novel time series model to estimate the mean death rate of an emerging or re-emerging disease with bilinear induced parameters; from the applied point of view, both the Tong and Chan (2006) and Shangodoyin (2009) models could be used by experts in monitoring and evaluating the death rate of a disease over time. For a general linear model (see ▶General Linear Model) the mean death rate could be specified as:

$$\mu_t = \sum_{a_1}^{a_2} p_j C_{t-j}$$

where a_1, $a_2 \geq 0$ are lower and upper bounds of time to death. The model is bilinear for estimating the mean deaths at time t as:

$$\mu_t = \alpha\mu_{t-1} + \beta\mu_{t-1}e_{t-1} + \sum_{1}^{u} p_j C_{t-j} + e_t.$$

By making all the necessary mathematical assumptions, the overall death rate for one-step time to death is given by

$$\dot{p}_1 = \frac{\sum_{1}^{n} \dot{\mu}_1 C_{t-1}}{\sum_{1}^{n} C_{t-1}^2}$$

where C_{t-j} is *the* number of confirmed cases at time $t - j$, $\dot{\mu}_t = \frac{\sum_{1}^{t} D_t}{t}$; $\forall t = 1, 2, \ldots$ and D_t is the *number* of deaths at time t. Readers should refer to the paper by Shangodoyin (2009) for details of the derivations.

In conclusion, statistical models play significant roles in the evaluation and monitoring of death rates from both emerging and re-emerging disease; and the use of most suitable time series models will provide the best insight to the future mortality rate for the given disease.

About the Author

Professor D. K. Shangodoyin was born in August, 1961 and started is academic career in 1986 and had taught Statistics in six African Universities. He is currently an Associate professor of Statistics at the University of Botswana, Southern Africa. He was the first alumni of the University of Ibadan in Nigeria to head the Department of Statistics at the Nigerian Premier University. He is currently the Statistics Pan African Society (SPAS) coordinator for SADC region in Africa. His broad area of research is the Theory and Application of Time Series, Bayesian inference and Econometrics modeling.

Cross References

►Meta-Analysis
►Modeling Survival Data
►Statistical Methods in Epidemiology
►Survival Data
►Time Series

References and Further Reading

Altman KL (2003, May 7) The SARS epidemic: the front-line research. New York Times

Mathers CD, Loncar D (2006) Projections of global mortality and burden of disease from 2002 to 2030. PLos Med 3(11):2011–2030

Shangodoyin DK (2009) Time series model for estimating the death rate of an emerging and re-emerging disease. In: 57th ISI Session, Durban, South Africa. www.stats.gov.za/isi2009

Tong H, Chan K (2006) Estimating the death rate of an emerging disease by Time Series Analysis. Technical Report, Department of Statistics & Actuarial Science, University of Iowa, Iowa, USA

Time Series Regression

William W. S. Wei
Professor
Temple University,
Philadelphia, PA, USA

Introduction

A regression model is used to study the relationship of a dependent variable with one or several independent variables. The standard regression model is represented by the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,$$

where Y is the dependent variable, $X_1, \ldots, X_k$ are the independent variables, $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients, and ε is the error term. When time series data are used in the model, it becomes time series regression, and the model is often written as

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \varepsilon_t,$$

or equivalently

$$Y_t = \mathbf{X}_t'\boldsymbol{\beta} + \varepsilon_t, \tag{1}$$

where $\mathbf{X}_t' = [1, X_{1,t}, \ldots, X_{k,t}]$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]'$. The standard regression assumptions on the error variable are that the ε_t are i.i.d. $N\left(0, \sigma_\varepsilon^2\right)$. Under these standard assumptions, it is well known that the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is a minimum variance unbiased estimator, distributed as multivariate normal, $N\left(\boldsymbol{\beta}, \sigma_\varepsilon^2 \mathbf{I}\right)$. When $\mathbf{X}_t'$ is stochastic in Model (1), conditional on $\mathbf{X}_t'$, the results about the OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ also hold as long as ε_s and $\mathbf{X}_t'$ are independent for all s and t. However, the standard assumptions associated with these models are often violated when time series data are used.

Regression with Autocorrelated Errors

When $\mathbf{X}_t'$ is a vector of a constant 1 and k lagged values of Y_t, i.e., $\mathbf{X}_t' = [1, Y_{t-1}, \ldots, Y_{t-k}]$, and ε_t is white noise, the model in (1) states that the variable Y_t is regressed on its own past k lagged values and hence is known as autoregressive model of order k, i.e., $AR(k)$ model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_k Y_{t-k} + \varepsilon_t. \tag{2}$$

The OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is still a minimum variance unbiased estimator. However, this result no longer holds when the ε_t are autocorrelated. In fact, when this is the case, the estimator is not even consistent and the usual tests of significance are invalid. This is an important caveat. When time series are used in a model, it is the norm rather than the exception that the error terms are autocorrelated. Even in univariate time series analysis when the underlying process is known to be an AR model as in (2), the error terms ε_t could still be autocorrelated unless the correct order of k is chosen. Thus, a residual analysis is an important step in regression analysis when time series variables are involved in the study.

There are many methods that can be used to test for autocorrelation of the error term. For example, one can use the test based on the Durbin–Watson statistic,

$$d = \frac{\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2} \approx 2(1 - \hat{\rho}_1), \tag{3}$$

where $\hat{\varepsilon}_t$ is residual series from the OLS procedure. Clearly, d lies between 0 and 4. A value close to 2 indicates no first-order autocorrelation, a value much less than 2 and close to 0 indicates a positive first-order autocorrelation and a value much greater than 2 and close to 4 indicates a negative first-order autocorrelation. To help make decision, in terms of the null hypothesis of no first-order autocorrelation against the alternative hypothesis of positive first-order autocorrelation, the critical values of Durbin-Watson, d_L and d_U can be constructed, which are functions of the number independent variables, the number of observations, and the significance level. The null hypothesis is

rejected if $0 < d < d_L$, is not rejected if $d_U < d < 2$, and inconclusive if $d_L < d < d_U$. For the null hypothesis of no first-order autocorrelation against the alternative hypothesis of negative first-order autocorrelation, the same table can be used since it is simply the mirror image of the former case when we look at the case from the endpoint of 4 instead of the endpoint of 0. Thus, the null hypothesis is rejected if $4 - d_L < d < 4$, is not rejected if $2 < d < 4 - d_U$, and inconclusive if $4 - d_U < d < 4 - d_L$.

More generally, to study the autocorrelation structure of the error term, we can perform the residual analysis with time series model identification statistics like the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF). Through these identification statistics, one can detect not only whether the residuals are autocorrelated but also identify its possible underlying model. A final analysis can then be performed on a model with autocorrelated errors as follows:

$$Y_t = \mathbf{X}_t'\boldsymbol{\beta} + \varepsilon_t \tag{4}$$

for $t = 1, 2, \ldots, n$, where

$$\varepsilon_t = \varphi_1\varepsilon_{t-1} + \cdots + \varphi_p\varepsilon_{t-p} + a_t \tag{5}$$

and the a_t are i.i.d. $N(0, \sigma^2)$.

Let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix}, \text{ and } \boldsymbol{\xi} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The matrix form of the model in (4) is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \tag{6}$$

where $\boldsymbol{\xi}$ follows a multivariate normal distribution (see ►Multivariate Normal Distributions) $N(0, \boldsymbol{\Sigma})$ When $\varphi_1, \ldots, \varphi_p$, and σ^2 are known in (5), $\boldsymbol{\Sigma}$ can be easily calculated. The diagonal element of $\boldsymbol{\Sigma}$ is the variance of ε_t, the jth off-diagonal element corresponds to the jth autocovariance of ε_t, and they can be easily computed from (5). Given $\boldsymbol{\Sigma}$, the generalized least squares (GLS) estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \tag{7}$$

is known to be a minimum variance unbiased estimator.

Normally, we will not know the variance-covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{\xi}$ because even if ε_t follows an $AR(p)$ model given in (5), the σ^2 and AR parameters φ_j are usually unknown. As a remedy, the following iterative GLS is often used:

(1) Calculate OLS residuals $\hat{\varepsilon}_t$ from OLS fitting of Model (4).
(2) Estimate φ_j and σ^2 for the $AR(p)$ model in (5) based on the OLS residuals, $\hat{\varepsilon}_t$, using any time series estimation method. For example, a simple conditional OLS estimation can be used.
(3) Compute $\boldsymbol{\Sigma}$ from the model (5) using the values of φ_j and σ^2 obtained in step (2).
(4) Compute GLS estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, using the $\boldsymbol{\Sigma}$ obtained in step (3). Compute the residuals $\hat{\varepsilon}_t$ from the GLS model fitting in step (4), and repeat the above steps (1) through (4) until some convergence criterion (such as the maximum absolute value change in the estimates between iterations becoming less than some specified quantity) is reached.

More generally, the error structure can be modified to include an ARMA model. The above GLS iterative estimation can still be used except that a nonlinear least squares estimation instead of OLS is needed to estimate the parameters in the error model. Alternatively, by substituting the error model in the regression equation (4), we can also use the nonlinear estimation or maximum likelihood estimation to jointly estimate the regression and error model parameters $\boldsymbol{\beta}$ and $\varphi_j's$, which is available in standard software.

It should be pointed out that although the error term, ε_t, can be autocorrelated in the regression model, it should be stationary. A nonstationary error structure could produce a spurious regression where a significant regression can be achieved for totally unrelated series.

Regression with Heteroscedasticity

One of the main assumptions of the standard regression model in Eq. 1 or the regression model with autocorrelated errors in Eq. 4 is that the variance, σ_ε^2, is constant. In many applications, this assumption may not be realistic. For example, in financial investment, it is generally agreed that stock markets' volatility is rarely constant.

Such a model having a non-constant error variance is called a heteroscedasticity model. There are many approaches which can be used to deal with heterocedasticity. For example, the weighted regression is often used if the error variances at different times are known or if the variance of the error term varies proportionally to the value of an independent variable. In time series regression we often have the situation where the variance of the error term is related to the magnitude of the past errors. This leads to the

conditional heteroscedasticity model, introduced by Engle (1982), where in terms of Eq. 1 we assume that

$$\varepsilon_t = \sigma_t e_t, \tag{8}$$

the e_t are i.i.d. random variable with mean 0 and variance 1, and

$$\sigma_t^2 = \theta_0 + \theta_1 \varepsilon_{t-1}^2 + \theta_2 \varepsilon_{t-2}^2 + \cdots + \theta_s \varepsilon_{t-s}^2. \tag{9}$$

Given all the information up to time $(t-1)$ the conditional variance of the ε_t becomes

$$\begin{aligned} Var_{t-1}(\varepsilon_t) &= E_{t-1}\left(\varepsilon_t^2\right) = E\left(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\right) = \sigma_t^2 \\ &= \theta_0 + \theta_1 \varepsilon_{t-1}^2 + \theta_2 \varepsilon_{t-2}^2 + \cdots + \theta_s \varepsilon_{t-s}^2. \end{aligned} \tag{10}$$

which is related to the squares of past errors, and it changes over time. A large error through ε_{t-j}^2 gives rise to the variance which tends to be followed by another large error. This is a common phenomenon of volatility clustering in many financial time series.

From the forecasting results, we see that Eq. 10 is simply the optimal forecast of ε_t^2 from the following $AR(s)$ model:

$$\varepsilon_t^2 = \theta_0 + \theta_1 \varepsilon_{t-1}^2 + \theta_2 \varepsilon_{t-2}^2 + \cdots + \theta_s \varepsilon_{t-s}^2 + a_t, \tag{11}$$

where the a_t is a $N\left(0, \sigma_a^2\right)$ white noise process. Thus, Engle (1982) called the model of the error term ε_t with the variance specification given in (8) and (9) or equivalently in (10) as autoregressive conditional heteroscedasticity model of order s ($ARCH(s)$).

Bollerslev (1986) extends the $ARCH(s)$ model to the $GARCH(r,s)$ model (generalized autoregressive conditional heteroscedasticity model of order (r,s)) so that the conditional variance of the error process is related not only to the squares of past errors but also to the past conditional variances. Thus, we have the following more general case

$$\varepsilon_t = \sigma_t e_t, \tag{12}$$

where the e_t are i.i.d. random variable with mean 0 and variance 1,

$$\sigma_t^2 = \theta_0 + \phi_1 \sigma_{t-1}^2 + \cdots + \phi_r \sigma_{t-r}^2 + \theta_1 \varepsilon_{t-1}^2 + \cdots + \theta_s \varepsilon_{t-s}^2, \tag{13}$$

and the roots of $(1 - \phi_1 B - \cdots - \phi_r B^r) = 0$ are outside the unit circle. To guarantee $\sigma_t^2 > 0$ we assume that $\theta_0 > 0$ and ϕ_i and θ_j are nonnegative.

More generally, the regression model with autocorrelated error can be combined with the conditional heteroscedasticity model, i.e.,

$$Y_t = \mathbf{X}_t' \beta + \varepsilon_t, \tag{14}$$

where

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \cdots + \phi_p \varepsilon_{t-p} + a_t, \tag{15}$$

$$a_t = \sigma_t e_t, \tag{16}$$

$$\sigma_t^2 = \theta_0 + \phi_1 \sigma_{t-1}^2 + \cdots + \phi_r \sigma_{t-r}^2 + \theta_1 a_{t-1}^2 + \cdots + \theta_s a_{t-s}^2, \tag{17}$$

and the e_t are i.i.d. $N(0,1)$. To test for the heteroscedasticity in this model, we follow:

(1) Calculate OLS residuals $\hat{\varepsilon}_t$ from the OLS fitting of (14).
(2) Fit an AR(p) model (15) to the $\hat{\varepsilon}_t$.
(3) Obtain the residuals $\hat{a}_t$ from the AR fitting in (15).
(4) Form the series $\hat{a}_t^2$, compute its sample ACF, PACF, and check whether these ACF and PACF follow any pattern. A pattern of these ACF and PACF not only indicates ARCH or GARCH errors, it also forms a good basis for their order specification.

For more detailed discussions and examples, we refer readers to Wei (2006).

About the Author

Professor Wei is a Past President, International Chinese Statistical Association (2001–2002). He was the Chair of the Department of Statistics at Temple University (1982–1987). He is a Fellow of the ASA, a Fellow of the RSS, and Elected Member of the ISI. He is currently an Associate Editor of the *Journal of Forecasting* and the *Journal of Applied Statistical Science*.

Cross References

►Autocorrelation in Regression
►Durbin–Watson Test
►Heteroscedastic Time Series
►Heteroscedasticity
►Least Squares
►Linear Regression Models
►Minimum Variance Unbiased
►Multivariate Normal Distributions
►Time Series

References and Further Reading

Bollerslev T (1986) Generalized autoregressive conditional heteroscedasticity. J Econom 31:307–327
Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
Wei WWS (2006) Time series analysis – Univariate and multivariate methods, 2nd edn. Pearson Addison-Wesley, Boston

Total Survey Error

Paul P. Biemer
Professor
RTI International and the University of North Carolina, Chapel Hill, NC, USA

Total survey error refers to the totality of error that can arise in the design, collection, processing and analysis of survey data. The concept dates back to the early 1940's although it has been revised and refined by a many authors over the years. Deming (1944), in one of the earliest works, describes "13 factors that affect the usefulness of surveys." These factors include sampling errors as well as nonsampling errors; i.e., the other factors that will cause an estimate to differ from the population parameter it is intended to estimate. Prior to Deming's work, not much attention was being paid to nonsampling errors and, in fact, textbooks on survey sampling made little mention of them. Indeed, classical sampling theory (Neyman 1934) assumes survey data are error free except for sampling error. The term "total survey error" originated with an edited volume of the same name (Andersen et al. (1977)).

A number of authors have provided a listing of the general sources of nonsampling error. For example, Biemer and Lyberg (2003) list five sources: specification, frame, nonresponse, measurement and data processing (including post-survey adjustment). A *specification error* arises when the concept implied by the survey question and the concept that should be measured in the survey differ. *Frame error* arises in the process for constructing, maintaining, and using the sampling frame(s) for selecting the survey sample. It includes the inclusion of nonpopulation members, exclusions of population members, and frame duplications. *Nonresponse error* encompasses both unit and item nonresponse. *Unit nonresponse* occurs when a sampled unit does not respond to any part of a ▶questionnaire. *Item nonresponse* error occurs when the questionnaire is only partially completed because an interview was prematurely terminated or some items that should have been answered were skipped or left blank. *Measurement error* includes errors arising from respondents, interviewers, survey questions and factors which affect survey responses. Finally, *data processing error* includes errors in editing, data entry, coding, computation of weights, and tabulation of the survey data.

The total survey error in a survey estimator, $\hat{\theta}$, for a population parameter, θ, can be summarized by the mean squared error of the estimator defined as

$$\begin{aligned} MSE(\theta) &= E(\hat{\theta} - \theta)^2 \\ &= B^2(\hat{\theta}) + Var(\hat{\theta}) \end{aligned} \tag{1}$$

where $B^2(\hat{\theta}) = E(\hat{\theta} - \theta)$ is the bias in the estimator and $Var(\hat{\theta}) = E(\hat{\theta} - \theta)^2$ is the variance of the estimator. For estimating the population mean, biases arise from systematic errors in the survey process; i.e., errors that are either predominately positive or predominately negative. As an example, sensitive items such as drug use tend to be underreported in surveys causing a negative bias in the estimated proportion of drug users. Nonresponse can also create a bias by systematically excluding from the survey data, individuals who differ on the survey characteristics from respondents.

The variance component of the MSE arises as a result of sampling error as well as variable nonsampling errors. Variable nonsampling error can be described roughly as the error remaining after accounting for the systematic errors. Variable errors tend to fluctuate randomly from unit to unit and have little or no effect on bias. As an example, interviewer estimates of housing values or neighborhood income levels may vary randomly from their true values.

To illustrate the effects of systematic and variable error, consider a ▶simple random sample of size n to estimate the mean, μ, of a large population. An elementary model for an observation, y_i, for characteristic y on sample unit i is

$$y_i = \mu_i + \varepsilon_i \tag{2}$$

where μ_i is the true value of the characteristic (i.e., the value that would have been observed without error), and ε_i is the error in the observation. Here ε_i represents the cumulative effect of all systematic and variable error sources for the ith unit. If the net error is 0, i.e., if $E(\varepsilon_i) = 0$, then there is no bias in the estimator $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. In that case, the errors are variable; i.e., no systematic errors. When systematic errors arise in the observations, $E(\varepsilon_i) = \beta \neq 0$ where β is the bias in $\bar{y}$. Under this model, $MSE(\bar{y})$ can be written as

$$MSE(\bar{y}) = \beta^2 + \frac{\sigma_\mu^2 + \sigma_\varepsilon^2}{n} \tag{3}$$

where $\sigma_\mu^2 = Var(\mu_i)$ and $\sigma_\varepsilon^2 = Var(\varepsilon_i)$. In this expression, β is the nonsampling bias, $n^{-1}\sigma_\mu^2$ is the sampling variance and $n^{-1}\sigma_\varepsilon^2$ is the nonsampling variance.

It is often useful to decompose both the nonsampling bias and variance components further by terms representing for the various sources of error in the survey process. As an example, suppose the major sources of bias include

the sampling frame, nonresponse and measurement bias. Then bias squared component can be expanded to include bias components for these sources as follows:

$$\beta = B_{FR} + B_{NR} + B_{MEAS} \tag{4}$$

where B_{FR} denotes frame bias, B_{NR}, nonresponse bias and B_{MEAS}, measurement bias. The variance component can also be expanded to include terms for all the major contributors of variable error such as sampling error, interviewers, respondents and other variable errors. Now the MSE can be rewritten as

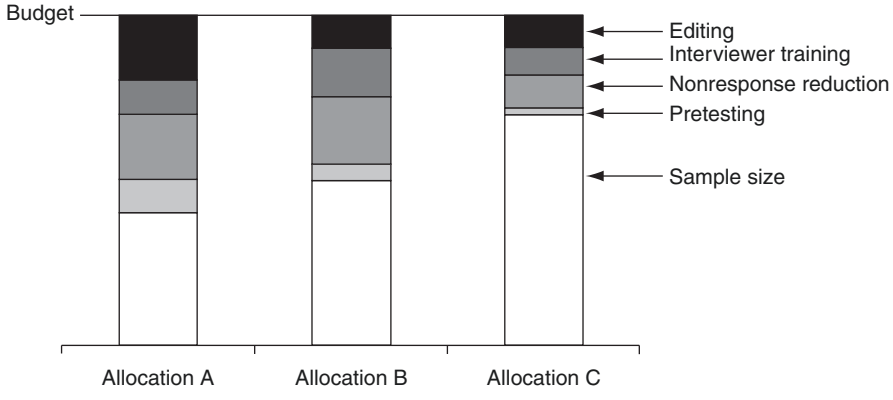
$$MSE(\bar{y}) = \left(B_{FR} + B_{NR} + B_{MEAS} + B_{DP}\right)^2 + \frac{\sigma_\mu^2 + \sigma_{int}^2 + \sigma_{res}^2 + \sigma_e^2}{n} \tag{5}$$

where σ_{int}^2 is the interviewer variance component, σ_{res}^2 is the respondent variance component, and σ_e^2 is the variance associated with all other sources. Thus, σ_ε^2 in (3) can be decomposed as $\sigma_\varepsilon^2 = \sigma_\mu^2 + \sigma_{int}^2 + \sigma_{res} + \sigma_e^2$. This form of the MSE assumes uncorrelated errors; however, the MSE can be also expanded to include correlations among the error from the same or difference error sources (see, for example, Biemer (2010)). The estimation of the components of the MSE can be quite challenging (see Mulry and Spencer 1991, for an application of the total survey error concept to the 1990 Decennial Census). Biemer (2010) provides a simplified estimator of the total MSE when multiple error sources are considered.

Finally, a critical part of the total survey error concept is error reduction and control. It is seldom possible to conduct every stage of the survey process at maximum accuracy since that would likely entail exceeding the survey budget and schedule by a considerable margin. Even under the best circumstances, some errors will necessarily remain in the data so that other, more serious errors can be avoided or reduced. For example, training interviewers adequately may require eliminating or limiting some quality control activities during data processing; but that might increase the data processing error. Efforts to reduce nonresponse bias may require substantial reductions during the survey pretesting phase to stay within budget. How should these resource allocation decisions be made? Making wise trade-offs requires an understanding of the sources of nonsampling error, their relative importance to data quality, and how they can be controlled. One answer is *optimal survey design*.

Optimal survey design aims to minimize the MSE (expressed in terms of the major error sources in the survey) subject to constraints on the survey process imposed by the budget, timeliness and other design considerations. Provide a design that is truly optimal (i.e., the best possible) may be an unattainable goal though it can be approximated. Doing so requires knowledge of the major error sources, their relative magnitudes and the most efficient and effective methods for nonsampling error reduction. Careful planning is then required to allocate survey resources to the various stages of the survey process so that the major sources of error are controlled to optimal, or near optimal levels.

To illustrate, Figure 1 depicts three possible resource allocation strategies satisfying the same budget constraint. Allocation *A* sacrifices sampling precision (i.e., sample size) for the sake of nonsampling error minimization by allocating more resources to editing, interviewer training, nonresponse reduction and questionnaire pretesting.



Total Survey Error. Fig. 1 Three potential cost allocations for the same fixed budget, each with very different implications for total survey error

Allocation *C* reduces these nonsampling error control strategies in order to boost the sample size thereby achieving greater sampling precision. Allocation *B* is a compromise between these two designs. Many other allocation schemes are possible. The challenge for the survey designer is to choose a single allocation strategy that provides the optimal balance between sampling error reduction and nonsampling error control while staying within budget. This is made even more difficult if there is insufficient information on the magnitudes of the total error components and scant knowledge regarding nonsampling error control strategies that are most effective at reducing the components of total survey error.

About the Author

Professor Biemer is RTI International Distinguished Fellow of Statistics, Associate Director of Survey Research and Development at the Odum Institute and Founding Director of the Certificate Program in Survey Methodology at the University of North Carolina, Chapel Hill. Prior to joining RTI, Professor Biemer headed the department of statistics at New Mexico State University and held a number of positions at the Bureau of the Census. Professor Biemer's book, *Introduction to Survey Quality* (with Lars Lyberg) is a widely used course text. He also co-edited following books: *Measurement Errors in Surveys*, *Survey Measurement and Process Quality*, and *Telephone Survey Methodology* which were published by John Wiley & Sons. His newest book, *Latent Class Analysis of Survey Error*, also published by John Wiley & Sons, is currently in press. He is a Fellow of the ASA and the AAAS, an Elected Member of the ISI and Associate Editor of the *Journal of Official Statistics*.

Crosss References

►Bias Analysis
►Business Surveys
►Nonresponse in Surveys
►Nonsampling Errors in Surveys
►Sample Survey Methods
►Sampling From Finite Populations

References and Further Reading

Andersen R, Kasper J, Frankel M, and Associates (1979) Total survey error. Jossey-Bass Publishers, San Francisco

Biemer PP (2010) Chapter 2 – Overview of design issues: total survey error. In: Marsden P, Wright J (eds) Handbook of survey research, 2nd edn. Bingley, United Kingdom: Emerald Group Publishing, LTD

Biemer P, Lyberg L (2003) Introduction to survey quality. Wiley, Hoboken

Deming WE (1944) On errors in surveys. Am Sociol Rev 9(4): 359–369

Mulry M, Spencer B (1991) Total error in PES estimates of population. J Am Stat Assoc 86(416):839–863

Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J Roy Stat Soc 97:558–606

Tourism Statistics

Stephen L. J. Smith
Professor
University of Waterloo, Waterloo, ON, Canada

The development of consistent measures of tourism has challenged tourism statisticians and economists since the 1930s (Smith 2004). The challenges arise, in part, from the nature of tourism as an economic activity. Although tourism is often referred to as an industry, it is fundamentally different than conventional industries; it is these differences that complicate the measurement of tourism (the definition of "tourism" and the nature of a "tourism industry" are discussed below). Further, the development of tourism statistics consistent among nations has required extensive negotiations among national statistical agencies as well as other international organizations to reach a consensus on the definition of tourism and associated concepts.

These concepts have been operationalized through new analytical tools, particularly the Tourism Satellite Account (UNWTO 1999). International agreement on core definitions and measurement techniques has now been achieved in principle. The tasks facing tourism statisticians are to refine, apply, and extend the concepts and tools that have been developed.

Fundamental to tourism statistics is, of course, the definition of "tourism." The World Tourism Organization defines tourism as the set of activities engaged in by persons temporarily away from their usual environment for a period of not more than one year, and for a broad range of leisure, business, religious, health, and personal reasons, excluding the pursuit of remuneration from within the place visited or long-term change of residence (UNWTO 1994). Thus, tourism fundamentally is something people do in certain circumstances (particularly travel outside their usual environment), not a commodity businesses produce.

There are several related concepts that are important for tourism policy, planning, marketing, and measurement purposes. One of these is *tourism commodity* – a good or service that would be produced only in a substantially reduced volume in the absence of tourism (such as passenger air services). A *tourism industry* is an industry characterized by the production of a tourism commodity (such as an airline offering scheduled passenger service). Thus, while tourism, *per se*, is not an industry, there are tourism industries such as accommodation, passenger transportation, food service, and recreation and entertainment.

Core tourism statistics include measures of the number of visitor arrivals in a destination (annually, seasonally, and/or monthly), their spending levels (often by category of commodity purchased), numbers of businesses serving visitors (by tourism industry), numbers of tourism employees, tourism's contribution to GDP, and government revenues attributable to tourism. Many specialized statistics related to persons engaged in tourism trips are also collected such as mode(s) of travel on a trip, mode(s) of accommodation used on a trip, activities engaged in during a trip, information sources used in planning a trip, routes taken, specific destinations visited, levels of satisfaction with services consumer, and so on.

Statistics related to activities not directly associated with individual behavior on specific trips are normally not considered to be tourism statistics, even though such information may be important for other purposes. Thus, government spending on infrastructure or tourism marketing, and investment in real estate or equipment (hotels, casinos, aircraft) are not considered to be within the scope of tourism statistics because they related to forms of production, and are more properly viewed as data relating to construction, manufacturing, marketing, real estate, and other forms of economic activity.

Sources of tourism statistics are numerous and diverse. They include surveys of border-crossing counts, visitors (during a trip or afterwards), business surveys, general social surveys (especially those covering household expenditures), and administrative records such as attraction ticket sales or hotel reservation records.

About the Author

Dr. Stephen L.J. Smith is Professor in the Department of Recreation and Leisure Studies and Director of the Tourism Policy and Planning Program at the University of Waterloo, Canada. He is an Elected Fellow of the International Statistical Institute (1994) and of the International Academy for the Study of Tourism (1991). He has authored more than 200 papers and eight books. His two most recent books are *Practical Tourism Research* (published by CABI, 2010) and *The Discovery of Tourism* (published by Emerald Publishing Group, 2010). He was involved in the creation of the Canadian Tourism Satellite Account through his leadership in the Canadian National Task Force on Tourism Data. Dr. Smith is Associate Editor for *Tourism Recreation Research* and the book review editor for *Annals of Tourism Research*.

Cross References

►Economic Statistics
►Seasonality
►Statistical Fallacies

References and Further Reading

Smith SLJ (2004) The measurement of global tourism: Old debates, new consensus, and continuing challenges. In: Lew AA, Hall CM, Williams AM (eds) A companion to tourism. Blackwell, Oxford, UK, pp 25–35

UNWTO (1994) Guidelines for tourism statistics. UNWTO, Madrid, Spain

UNWTO (1999) Tourism satellite account (TS): The conceptual framework. UNWTO, Madrid, Spain

Trend Estimation

Tommaso Proietti
Professor of Economic Statistics
University of Rome "Tor Vergata", Rome, Italy

Trend estimation deals with the characterization of the underlying, or long–run, evolution of a time series. Despite being a very pervasive theme in time series analysis since its inception, it still raises a lot of controversies. The difficulties, or better, the challenges, lie in the identification of the sources of the trend dynamics, and in the definition of the time horizon which defines the long run. The prevalent view in the literature considers the trend as a genuinely latent component, i.e., as the component of the evolution of a series that is persistent and cannot be ascribed to observable factors. As a matter of fact, the univariate approaches reviewed here assume that the trend is either a deterministic or random function of time.

A variety of approaches is available, which can be classified as nonparametric (kernel methods, local polynomial regression, band-pass filters, and wavelet multiresolution analysis), semiparametric (splines and Gaussian random fields) and parametric, when the trend is modeled as a

stochastic process. They will be discussed with respect to the additive decomposition of a time series $y(t)$:

$$y(t) = \mu(t) + \epsilon(t), \quad t = 1, \ldots, n, \tag{1}$$

where $\mu(t)$ is the trend component, and $\epsilon(t)$ is the noise, or irregular, component. We assume throughout that $\epsilon(t) = 0$ is a zero mean stationary process, whereas $\mu(t)$ can be a random or deterministic function of time. The above decomposition bears different meanings in different fields. In experimental sciences $\epsilon(t)$ is usually interpreted as a pure measurement error, so that a signal is observed with superimposed random noise. However, in behavioral sciences such has economics, quite often $\epsilon(t)$ is interpreted as a stationary stochastic cycle or as the transitory component of $y(t)$. The underlying idea is that trends and cycles can be ascribed to different economic mechanisms. Moreover, according to some approaches $\mu(t)$ is an underlying deterministic function of time, whereas for other it is a random function (e.g., a random walk, or a Gaussian process), although this distinction becomes more blurred in the case of splines. For some methods, like band pass filtering, the underlying true value $\mu(t)$ is defined by the analyst via the choice of a cutoff frequency which determines the time horizon for the trend.

The simplest and historically oldest approach to trend estimation adopted a global polynomial model for μ_t: $\mu(t) = \sum_{j=0}^{p} \beta_j t^j$. The statistical treatment, based on least squares, is provided in Anderson (1971). It turns out that global polynomials are amenable to mathematical treatment, but are not very flexible: they can provide bad local approximations and behave rather weirdly at the beginning and at the end of the sample period, which is inconvenient for forecasting purposes. More up to date methodologies make the representation more flexible either assuming that certain features, like the coefficients or the derivatives, evolve over time, or that a low order polynomial representation is adequate only as a local approximation.

Local polynomial regression (LPR) is a nonparametric approach that assumes that $\mu(t)$ is a smooth but unknown deterministic function of time, which can be approximated in a neighborhood of time t by a polynomial of degree p of the time distance with time t. The polynomial is fitted by locally weighted least squares, and the weighting function is known as the kernel. LPR generates linear signal extraction filters (also known as moving average filters) whose properties depend on three key ingredients: the order of the approximating polynomial, the size of the neighborhood, also known as the bandwidth, and the choice of the kernel function. The simplest example is the arithmetic moving average $m_t = \frac{1}{2h+1} \sum_{j=-h}^{h} y_{t+j}$, which is the LPR estimator of a local linear trend ($p = 1$) in discrete time using a bandwidth of $2h + 1$ consecutive observations and the uniform kernel.

Trend filters that arise from fitting a locally weighted polynomial to a time series have a well established tradition in time series analysis and signal extraction; see Kendall et al. (1983) and Loader (1999). For instance, the Maculay's moving average filters and the Henderson (1916) filters are integral part of the X-12 seasonal adjustment procedure adopted by the US Census Bureau.
The methodology further encompasses the Nayadara-Watson kernel smoother.

An important class of nonparametric filters arises from the frequency domain notion of a band-pass filter, that is popular in engineering. An ideal low-pass filter retains only the low frequency fluctuations in the series and reduces the amplitude of fluctuations with frequencies higher than a cutoff frequency ω_c. Such a filter is available analytically, but unfeasible, since it requires a doubly infinite sequence of observations; however, it can be approximated using various strategies (see Percival and Walden 1993). Wavelet multiresolution analysis provides a systematic way of performing band-pass filtering.

An alternative way of overcoming the limitations of the global polynomial model is to add polynomial pieces at given points, called knots, so that the polynomial sections are joined together ensuring that certain continuity properties are fulfilled. Given the set of points $t_1 < \ldots < t_i < \ldots t_k$, a polynomial spline function of degree p with k knots $t_1, \ldots, t_k$ is a polynomial of degree p in each of the $k + 1$ intervals $[t_i, t_{i+1})$, with $p - 2$ continuous derivatives, whereas the $p-1$-st derivative has jumps at the knots. It can be represented as follows:

$$\mu(t) = \beta_0 + \beta_1(t - t_1) + \cdots + \beta_p(t - t_1)^p + \sum_{i=1}^{k} \eta_i (t - t_i)_+^p, \tag{2}$$

where the set of functions

$$(t - t_i)_+^p = \begin{cases} (t - t_i)^p, & t - t_i \geq 0, \\ 0, & t - t_i < 0 \end{cases}$$

defines what is usually called the truncated power basis of degree p.

According to (2) the spline is a linear combination of polynomial pieces; at each knot a new polynomial piece, starting off at zero, is added so that the derivatives at that point are continuous up to the order $p-2$. The most popular special case arises for $p = 3$ (cubic spline); the additional *natural boundary conditions*, which constrain the

spline to be linear outside the boundary knots, is imposed. See Green and Silverman (1994) and Ruppert et al. (2003).

An important class of semiparametric and parametric time series models are encompassed by (2). The piecewise nature of the spline "reflects the occurrence of structural change" (Poirier 1973). The knot t_i is the timing of a structural break. The change is "smooth," since certain continuity conditions are ensured. The coefficients η_i, which regulate the size of the break, may be considered as fixed or random. In the latter case $\mu(t)$ is a stochastic process, η_i is interpreted as a *random shock* that drives the evolution of $\mu(t)$, whereas the truncated power function $(t-t_i)_+^p$ describes its *impulse response function*, that is the impact on the future values of the trend.

If the η_i's are considered as random, the spline model can be formulated as a ▶linear mixed model, which is a traditional regression model extended so as to incorporate random effects. Denoting $\mathbf{y} = [y(t_1), \ldots, y(t_n)]'$, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]'$, $\boldsymbol{\epsilon} = [\epsilon(t_1), \ldots, \epsilon(t_n)]'$, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta}$,

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{3}$$

where the t-th row of $\mathbf{X}$ is $[1, (t-1), \ldots, (t-1)^p]$, and $\mathbf{Z}$ is a known matrix whose i-th column contains the impulse response signature of the shock η_i, $(t-t_i)_+^p$.

The trend is usually fitted by penalized least squares (PLS), which chooses $\boldsymbol{\mu}$ so as to minimize

$$(\mathbf{y}-\boldsymbol{\mu})'(\mathbf{y}-\boldsymbol{\mu}) + \lambda \int \left[\frac{d^{p-1}\mu(t)}{dt^{p-1}}\right]^2 dt, \tag{4}$$

where $\lambda \geq 0$ is the smoothness parameter.

PLS is among the most popular criteria for designing filters that has a long and well established tradition in actuarial sciences and economics (see Whittaker 1923, Leser 1961, and, more recently, Hodrick and Prescott 1997). Under Gaussian independent measurement noise minimizing the PLS criterion amounts to finding the conditional mode of μ given $\mathbf{y}$, This is a solution to the smoothing problem. If $\mu(t)$ is random, the minimum mean square estimator of the signal is $\mathrm{E}(\mu(t)|\mathbf{y})$. If the model (1) is Gaussian, these inferences are linear in the observations. The computations are carried out efficiently by the Kalman filter and the associated smoother (see Wecker and Ansley 1983).

The linear mixed model representation (3) encompasses other approaches, according to which the component $\mathbf{Z}\boldsymbol{\eta}$ is a Gaussian random process (Rasmussen and Williams 2006), or a (possibly nonstationary) time series process with a Markovian representation, such as in the structural time series approach see Harvey (1989), and in the canonical decomposition of time series (see Hillmer and Tiao 1982). The Markovian nature of the opens the way to the statistical treatment by the state space methodology and signal extraction is carried out efficiently by the Kalman filter and smoother. Popular predictors, such as exponential smoothing and Holt and Winters, arise as special cases (see Harvey 1989). The representation theory for the estimator of the trend component, Wiener-Kolmogorov filter, is established in Whittle (1983).

The analysis of economic time series has contributed to trend estimation in several ways. The first contribution is the attempt to relate the trend to a particular economic mechanism. The issue at stake is whether $\mu(t)$ is better characterized as a deterministic or stochastic trends. This problem was addressed in a very influential paper by Nelson and Plosser (1982), who adopted the (augmented) Dickey Fuller test for testing the hypothesis that the series is integrated of order 1, I(1), implying that $y(t) - y(t-1)$ is a stationary process versus the alternative that it is trend-stationary, e.g., $m(t) = \beta_0 + \beta_1 t$. Using a set of annual U.S. macroeconomic time series they are unable to reject the null for most series and discuss the implications for economic interpretation. The trend in economic aggregate is the cumulative effect of supply shocks, i.e., shocks to technology that occur randomly and propagate through the economic system via a persistent transmission mechanism.

A fundamental contribution is the notion of cointegration (Engle and Granger 1987), according to which two or more series are cointegrated if they are themselves nonstationary (e.g., integrated of order 1), but a linear combination of them is stationary. Cointegration results from the presence of a long run equilibrium relationship among the series, so that the same random trends drive the nonstationary dynamics of the series; also, part of the short run dynamics are also due to the adjustment to the equilibrium.

A third contribution, related to trend estimation, is the notion of spurious cycles that may result from inappropriate detrending of a nonstationary time series. This effect is known as the Slutzky–Yule effect, and concerned with the fact that an ad hoc filter to a purely random series can introduce artificial cycles.

Finally, large dimensional dynamic factor models have become increasingly popular in empirical macroeconomics. The essential idea is that the precision by which the common components are estimated can be increased by bringing in more information from related series: suppose for simplicity that $y_i(t) = \theta_i \mu(t) + \epsilon_i(t)$, where the i-th series, $i = 1, \ldots, N$, depends on the same stationary common factor, which is responsible for the observed comovements of economic time series, plus an idiosyncratic component, which includes measurement error and local

shocks. Generally, multivariate methods provide more reliable measurements provided that a set of related series can be viewed as repeated measures of the same underlying latent variable. Stock and Watson (2002) and Forni et al. (2000) discuss the conditions on μ_t and ϵ_{it} under which dynamic or static principal components yield consistent estimates of the underlying factor μ_t as both N and the number of time series observations tend to infinity.

About the Author

Professor Proietti is Associate Editor of *Computational Statistics and Data Analysis* (Elsevier), and Co-Editor of *Statistical Methods and Applications* (Springer) He is Editor (with A.C. Harvey) of the text: *Readings in Unobserved Components Models* (Advanced Texts in Econometrics, Oxford University Press, 2005).

Cross References

►Business Forecasting Methods
►Detection of Turning Points in Business Cycles
►Dickey-Fuller Tests
►Exponential and Holt-Winters Smoothing
►Forecasting Principles
►Linear Mixed Models
►Moving Averages
►Nonparametric Estimation
►Nonparametric Regression Using Kernel and Spline Methods
►Seasonal Integration and Cointegration in Economic Time Series
►Structural Time Series Models
►Time Series

References and Further Reading

Anderson TW (1971) The statistical analysis of time series. Wiley, New York
Engle RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation, and testing. Econometrica 55: 251–276
Forni M, Hallin M, Lippi F, Reichlin L (2000) The generalized dynamic factor model: identification and estimation. Rev Econ Stat 82:540–554
Green PJ, Silverman BV (1994) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman & Hall, London
Harvey AC (1989) Forecasting, structural time series and the Kalman filter. Cambridge University Press, Cambridge
Henderson R (1916) Note on graduation by adjusted average. Trans Actuarial Soc America 17:43–48
Hillmer SC, Tiao GC (1982) An ARIMA-model-based approach to seasonal adjustment. J Am Stat Assoc 77:63–70
Hodrick R, Prescott EC (1997) Postwar U.S. business cycle: an empirical investigation. J Money Credit Bank 29(1):1–16
Kendall M, Stuart A, Ord JK (1983) The advanced theory of statistics, vol 3. Charles Griffin, London
Leser CEV (1961) A simple method of trend construction. J Roy Stat Soc B 23:91–107
Loader C (1999) Local regression and likelihood. Springer-Verlag, New York
Nelson CR, Plosser CI (1982) Trends and random walks in macroeconomic time series: some evidence and implications. J Monet Econ 10:139–162
Percival D, Walden A (1993) Spectral analysis for physical applications. Cambridge University Press, Cambridge
Poirier DJ (1973) Piecewise regression using cubic splines. J Am Stat Assoc 68:515–524
Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, Cambridge
Ruppert D, Wand MJ, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge
Stock JH, Watson MW (2002b) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97:1167–1179
Watson GS (1964) Smooth regression analysis. Shankya Series A, 26:359–372
Wecker WE, Ansley CF (1983) The signal extraction approach to nonlinear regression and spline smoothing. J Am Stat Assoc 78:81–89
Whittaker E (1923) On new method of graduation. Proc Edinburgh Math Soc 41:63–75
Whittle P (1983) Prediction and regulation by linear least squares methods, 2nd edn. Basil Blackwell, Oxford

Two-Stage Least Squares

Roberto S. Mariano
Dean and Professor of Economics and Statistics
Singapore Management University, Singapore, Singapore

In the linear regression model, $y = X_1\beta_1 + Y_1\beta_2 + u = Z\beta + u$, there are real-life situations when some of the regressors, denoted by Y_1 in the model, are correlated with the disturbance term. The vector and matrices y, X_1, and Y_1 are $N \times 1, N \times K_1$, and $N \times (G-1)$ data matrices from a sample of size N. u is the $N \times 1$ vector of disturbances, assumed to have mean zero and variance-covariance matrix $\sigma^2 I$. In this model, X_1 is assumed to be statistically independent of the disturbance term and the analysis is done conditional on X_1.

In such situations where correlation between error and regressor exists, ordinary least squares (*OLS*) estimates of regression coefficients become not only biased but also inconsistent (as sample size increases indefinitely). One of the earlier efforts to correct for this inconsistency is a

two step procedure called two-stage least squares in the econometric literature. The procedure first regresses the "disturbance-correlated" variables, Y_1, on a selected set of first-stage regressors (X) and obtains the calculated regression values $P_X(Y_1) = X(X'X)^{-1}X'Y_1$, the projection of Y_1 on the column space of X. For the second stage of the procedure, y is then regressed on X_1 and $P_X(Y_1)$ to obtain the $2SLS$ estimate b_{2SLS}. Typically $X = (X_1, X_2)$ where X_2 is $N \times K_2$, and is independent of u, and X has full column rank equal to at least $K_1 + G - 1$. Intuitively, the first-stage regression serves to "purge" Y_1 of its component that is correlated with u and this leads to consistency in the regression at the second stage where Y_1 is replaced by $P_X(Y_1)$.

$2SLS$ was developed in the econometric literature in dealing with the estimation of the linear regression model (see ▶Linear Regression Models) as part of a simultaneous equations system. In this context, the joint probability distribution of y and Y_1 is specified and X is determined from the model.

$2SLS$ appeared in an earlier form as an intermediate step in the iteration towards the calculation of the limited-information-maximum-likelihood $(LIML)$ estimator in simultaneous equation models. The $2SLS$ estimator in the linear regression model also can be interpreted as an instrumental variable (IV) estimator, using the instrument matrix $W_{2SLS} = P_X Z$ for Z; that is,

$$b_{IV} = \left(W'_{2SLS}Z\right)^{-1} W'_{2SLS}y = (Z'P_XZ)^{-1}Z'P_Xy = b_{2SLS}.$$

The $2SLS$ estimator also can be interpreted as a generalized least squares (GLS) estimator in the derived linear model $X'y = X'Z\beta + X'u$.

When Z has a large dimension, modified two-stage least squares has been suggested as an alternative approach. This is also a two-step regression procedure where the first stage of $2SLS$ is modified by regressing Y_1 on H, a $N \times h$ submatrix spanning a column subspace of X. H is chosen to be of full column rank and rank $[(I - P_1)H] \geq G - 1$, where P_1 is the projection matrix on the column space of X_1. One suggested manner of constructing H is to start with X_1 and then add at least $G-1$ of the remaining columns of X or the first K_2 principal components of $(I - P_1)X_2$. In this case, the modified $2SLS$ is exactly equivalent to the IV estimator using $(X_1, P_H Y_1)$ as the instrument matrix.

Ordinary least squares also can be interpreted as an IV estimator with Z as the instrument for itself. Another variation of an IV estimator that has been suggested is Theil's k-class estimator. This uses as its instrument matrix a linear combination of the instrument matrices for OLS and $2SLS$ and k is chosen by the investigator and can be stochastic or non-stochastic. Thus, with $W_{(k)} = kW_{2SLS} + (1-k)W_{(OLS)} = kP_XZ + (1-k)Z$, the k-class estimator is

$$b_{(k)} = \left(W'_{(k)}Z\right)^{-1} W'_{(k)}y = \beta + \left(W'_{(k)}Z\right)^{-1} W'_{(k)}u,$$

Assuming that plim(k) is finite, a necessary and sufficient condition for consistency of the k-class estimator is plim$(1-k) = 0$ – that is, the contribution of the OLS instrument matrix dies out in the limit.

The limited information maximum likelihood $(LIML)$ estimator is closely related to the $2SLS$ and other estimators introduced here and is a member of the k-class of estimators. Think of the linear regression equation introduced above as part of a complete simultaneous-equations model for the joint stochastic behavior of y, Y_1, and other dependent variables showing up in other equations of the model. The $LIML$ estimator of β maximizes the likelihood of (y, Y_1) subject to any identifiability restrictions, and is called limited in the sense that it ignores the dependent variables that do not show up in the regression equation. The constrained maximization process in $LIML$ reduces to minimizing the following variance ratio with respect to $\beta^* = (1, \beta')'$

$$v = (\beta^{*\prime}A\beta^*)/(\beta^{*\prime}S\beta^*) = 1 + (\beta^{*\prime}W\beta^*)/(\beta^{*\prime}S\beta^*),$$

where $Y = (y, Y_1)$; $S = Y'(I - P_X)Y$; $W = Y'(P_X - P_1)$; and $A = S + W = Y'(I - P_1)Y$.

This minimization problem yields the solution b_{LIML} as a characteristic vector of A with respect to S corresponding to the smallest root h of $\det(A - vS) = 0$, and

$$h = \left(b^*_{LIML}{}'Ab^*_{LIML}\right) / \left(b^*_{LIML}{}'Sb^*_{LIML}\right).$$

Note that $(\beta^{*\prime}W\beta^*)$ is the marginal regression sum of squares due to X_2 given X_1 in the regression of $Y\beta^*$ on X, while $\beta^{*\prime}S\beta^*/(N-K)$ provides an unbiased estimator of the error variance in the regression equation. Thus $LIML$ minimizes the marginal contribution of X_2 given X_1 relative to an estimate of the error variance. $2SLS$ simply minimizes this marginal contribution in absolute terms.

The $LIML$ estimator b^*_{LIML} needs to be normalized to have a unit value in its first element, to be comparable with the other estimators we have discussed so far. With such a normalization, the $LIML$ estimator of β_1 and β_2 turns out to be a k-class estimator as well, where the value of k is h, the smallest root of $\det(A - vS) = 0$. Note that $h = 1 + f$, where f is the smallest root of $\det(W - vS) = 0$. Thus, $LIML$ is an IV estimator also, whose instrument matrix is a linear combination of the OLS and $2SLS$ instrument matrices, with k stochastic and k at least equal to unity.

Two-stage least squares and the other estimators discussed above have been analyzed for statistical properties in small samples, under the standard large-sample asymptotics, and in alternative nonstandard asymptotic settings such as error variances going to zero, number of instruments going to infinity at the same rate as sample size, and so-called weak instrument asymptotics.

About the Author

Roberto S. Mariano received his PhD in Statistics in 1970, Stanford University. Currently, he is Professor of Economics and Statistics and Dean, School of Economics, Singapore Management University. He is also Director, Sim Kee Boon Institute for Financial Economics, and Co-Director, Center for Financial Econometrics. Before joining the Singapore Management University he was Professor of Economics and Statistics, Department of Economics, University of Pennsylvania (1980–2004). Dr Mariano is a Fellow, Econometric Society (2009–present), and a Fellow, Wharton Financial Institutions Center, Wharton School (2004–present). He has authored numerous research papers and books on econometric methodology and applications and has served on the editorial board of several international professional journals in economics and statistics. He was the principal investigator in research projects funded by the United Nations, US National Science Foundation, the Rockefeller Foundation, the US Department of Commerce and the US Department of Agriculture. In Singapore, he worked on a research project for the Ministry of Manpower entitled "Macroeconometric Sectoral Modeling for Manpower Planning in Singapore" from March 2000–March 2002 where he was the principal investigator with Nobel Laureate Lawrence R. Klein.

Cross References

►Econometrics
►Instrumental Variables
►Least Squares
►Linear Regression Models
►Method Comparison Studies
►Properties of Estimators

References and Further Reading

Anderson TW, Rubin H (1949) Estimation of the parameters of a single equation in a complete system of stochastic equations. Ann Math Stat 20:46–63

Basman RL (1957) A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica 25:77–83

Bound J, Jaeger DA, Baker RM (1995) Problems with instrumental variable estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J Am Stat Assoc 90:443–450

Mariano RS (1977) Finite-sample properties of instrumental variable estimators of structural coefficients. Econometrica 45:487–496

Mariano RS (2001) Chapter 6: Simultaneous equation model estimators: statistical properties and practical implications. In: Baltagi B (ed) Companion to theoretical econometrics. Blackwell Publishers, Oxford, pp 122–143

Phillips PCB (1983) Exact small sample theory in the simultaneous equations model. In: The handbook of econometrics, Volume II, Elsevier Science, North Holland, Amsterdam, pp 881–935

Theil H (1953) Repeated least squares applied to complete equation systems. Central Planning Bureau mimeograph, The Hague