Chapter 2
Regression

2.1 Introduction

This chapter reviews the essentials of regression analysis. For most readers it will be a refresher that can be skimmed quickly; it provides a concise, self-contained coverage of topics that are the staple of any good course on econometrics.

The emphasis here is on the issues that commonly arise with household survey data. Most survey data are cross-sectional, with relatively large numbers of observations. Occasionally one encounters panel data, but they are always short panels, with few time periods and many households or individuals; we address the special issues related to panel data in Chap. 9.

Survey data almost never come from simple random sampling. As noted in Chap. 3, there is typically stratification – which calls for the use of sampling weights in computing most sample statistics – and clustering, which reduces the precision of estimation, and needs to be taken into account explicitly, as explained more fully below.

Data collected in household surveys are prone to error. This is true of any data, but some of the variables that are widely used based on surveys, such as household income or expenditure, are particularly difficult to measure with precision (Haughton and Khandker 2009, Chap. 2).

The other salient feature of survey data is that they are almost always incomplete. We rarely have all the variables that we would desire; compromises are inevitable, given the cost of asking questions and the need to choose what to ask. Moreover, some of the most important variables of interest, such as an individual's "ability," may be unobservable, and this has implications for how to proceed with regression.

In the next section we outline the basics of regression, and then consider the problems that arise in the context of household survey data, beginning with measurement error, and then turning to omitted variable bias, multicolinearity,

D. Haughton and J. Haughton, *Living Standards Analytics*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4614-0385-2_2,

heteroscedasticity, outliers, clustering, and simultaneity. For a more formal treatment we recommend the classic textbook by Greene (2011); an excellent blend of theory and practice, with numerous examples using the Stata software, may be found in the book on microeconometrics by Cameron and Trivedi (2009).

2.2 Basics

The most fundamental use of regression is to summarize and describe patterns in data. To illustrate, consider the scatterplot in Fig. 2.1 (similar to the one in Haughton and Khandker 2009, p. 275), which graphs food consumption per capita against expenditure per capita for 9,189 Vietnamese households, based on data from the 2006 Vietnam Household Living Standards Survey. The problem with the graph is that the data points are so numerous and crowded that it is difficult for the eye to discern any essential underlying relationships.

A regression line goes some way toward solving this problem. The straight line in Fig. 2.1 represents the results of regressing food spending per person on total spending per person. For this example we used Stata – although any statistical software would do the job just as well – with the commands

```
regress food1 exp1
predict food1hat
```

The second command here puts the predicted values into a variable called `food1hat`, and this is what is actually graphed in Fig. 2.1. The estimated equation is

$$(\text{food spending per person}) = 1.188 + 0.22\ (\text{total spending per person}).$$

The data in Fig. 2.1 are measured in millions of dong per year; in 2006 the exchange rate was about VND15,000 per US$. This relationship shows that an extra thousand dong in total spending is associated with an extra 220 dong of spending on food. This is simply an observed association; we are not trying to make a statement about causality – a topic that we address more fully in Chap. 5.

Going back at least to the late nineteenth century, when Ernst Engel first noted the phenomenon based on household budget data collected in Belgium, it is widely accepted that higher per capita expenditure levels are associated with a falling share of spending devoted to food. This idea is captured in Fig. 2.1 by the quadratic curve, which flattens as one moves from left to right. The regression estimate that summarizes this may be written as

$$\begin{aligned}(\text{food spending per person}) = {} & 0.853 + 0.30\ (\text{total spending per person}) \\ & - 0.0021\ (\text{total spending per person})^2\end{aligned}$$

Fig. 2.1 Spending on food graphed against total spending, Vietnam, 2006. *Source*: Vietnam Living Standard Survey of 2006. Variables are in per capita terms. 9,122 observations are shown here; a further 67 are not shown (for readability)

Based on this equation, we find that for someone who is very poor, almost a third of incremental spending is devoted to food, but for those in the top percentile of the distribution the proportion of additional spending associated with food spending is just 0.14.

2.2.1 Inference

While the use of regression to summarize voluminous data is helpful, most researchers are interested in going further, using it for statistical inference. We want to know whether the link between variables is statistically significant, and we often hope to attempt to infer causality, so that we may conduct policy experiments that allow us to answer questions such as "if we were to change variable x, what would happen to variable y?" Indeed Deaton (1997, p. 65) makes the point that the essential thrust of most econometrics is to try to make causal inferences from nonexperimental data.

In order to use regression for inference, a number of additional assumptions are required. In setting these out, we broadly follow the approach taken by Cameron and Trivedi (2009).

Based on theory or intuition or prior practice, we believe that the conditional mean of some "dependent" variable y is given by a linear model of the form

$$E(y|\mathbf{X}) = \mathbf{X}\beta \equiv \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K. \tag{2.1}$$

Given data on N observations of variables y and $X_1, \ldots, X_K$, we want to estimate the vector of coefficients β.[1] Since the model cannot be expected to fit the data perfectly, we include an (additive) error to give, for the ith observation, the regression model

$$y_i = \mathbf{x}_i{}'\beta + e_i, \quad i = 1, \ldots, N. \tag{2.2}$$

The true errors (e_i) are unobservable; if we knew what the errors were, they would not be errors! In the classical regression model we begin by making three assumptions about the nature of these errors, and then spend most of our time figuring out how to proceed when we believe these assumptions have been violated. The assumptions are

Classical assumption 1
Exogeneity of Regressors: $E(e_i|\mathbf{X}_i) = 0$.

This says that the errors have zero mean; in addition, we assume that the errors are independent of the $\boldsymbol{x}_i$ regressors, which may thus be considered to be exogenous. This assumption is essential for consistency in the estimation of the parameters.

Classical assumption 2
Conditional Homoscedasticity: $E(e_i^2|\mathbf{X}_i) = \sigma^2$.

In other words, the errors are homoscedastic – that is to say, they are distributed with a constant variance that does not vary with the regressors.

Classical assumption 3
Conditionally Uncorrelated Observations: $E(e_i e_j|\mathbf{X}_i, \mathbf{X}_j) = 0, \quad i \neq j..$

This supposes that the observations are conditionally independent.

The ordinary least squares (OLS) estimates $\hat{\beta}$ of the coefficients are generated by $\hat{\beta} = (X'X)^{-1}X'y$, and when assumptions 1–3 hold, these OLS estimates are efficient; formally, they are the best linear unbiased estimates (BLUE). In large samples the vector $\hat{\beta}$, which is itself a random variable, is asymptotically distributed as (multivariate) normal, which gives

$$\hat{\beta} \dot{\sim} N\Big(\beta, \mathrm{Var}(\hat{\beta})\Big). \tag{2.3}$$

[1] Here $\mathbf{X}$ is an $N \times K$ matrix in which each column represents the observations on one of the right-hand variables.

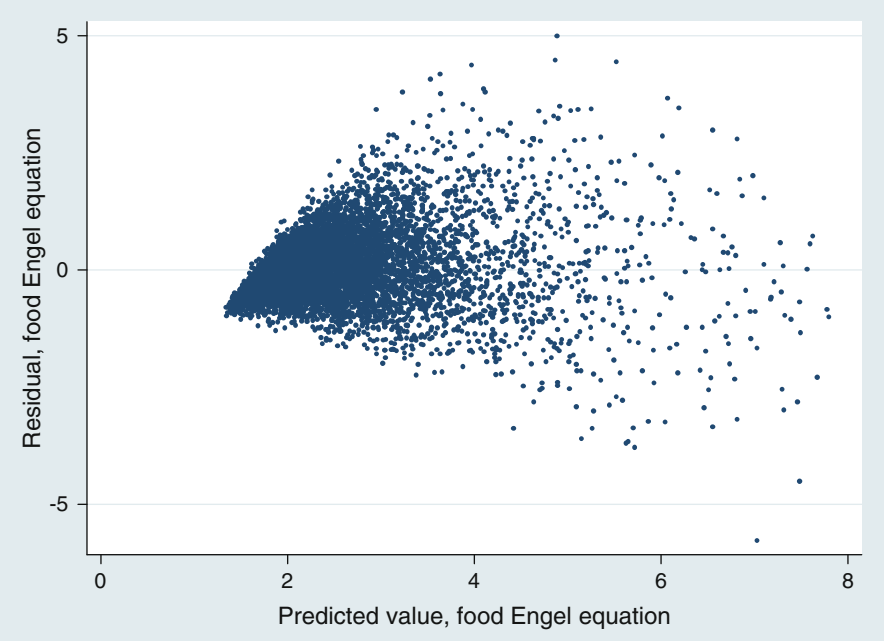


Fig. 2.2 Plot of residuals (*vertical axis*) vs. predicted values of spending on food (*horizontal axis*), Vietnam, 2006. (*Source*: As for Fig. 2.1)

Here $\text{Var}(\hat{\beta})$is the asymptotic variance–covariance matrix of the estimator (VCE), which itself needs to be estimated by $\hat{V}(\hat{\beta})$. The standard error of $\hat{\beta}_i$ is given by the square root of the *i*th diagonal element of $\hat{V}(\hat{\beta})$. Most survey samples are large enough for the asymptotic conditions to apply (approximately), allowing one to base inference tests on the normal distribution (for $\hat{\beta}$) or χ^2 distribution (for $\hat{V}(\hat{\beta})$). However, even in these cases, for data on subgroups – such as observations for a province within a country – small-sample tests based on the *t* and *F* distributions are typically needed.

Having estimated a basic equation, this is often a good point at which to look at some basic diagnostic plots, which can be very helpful in indicating whether the classical assumptions are likely to apply. A relatively standard graph puts the regression residuals on the vertical axis and the predicted values of the dependent variable on the horizontal axis; if the classical assumptions apply, the observations should appear to form a random cloud. On the other hand, if there is heteroskedasticity, or if the residuals are not independent of one another, it will usually be quite evident at this point because they will form a pattern; for instance, in Fig. 2.2, which is based on a linear regression of the data shown in Fig. 2.1, the scatter of the residuals "opens up" as we move toward the right, a sure sign of the presence of heteroskedasticity (which we discuss further below).

Another useful graph is the *Q–Q plot*, sometimes referred to as a quantile plot or a probability plot. In the version shown in Fig. 2.3 – generated using the `qnorm`

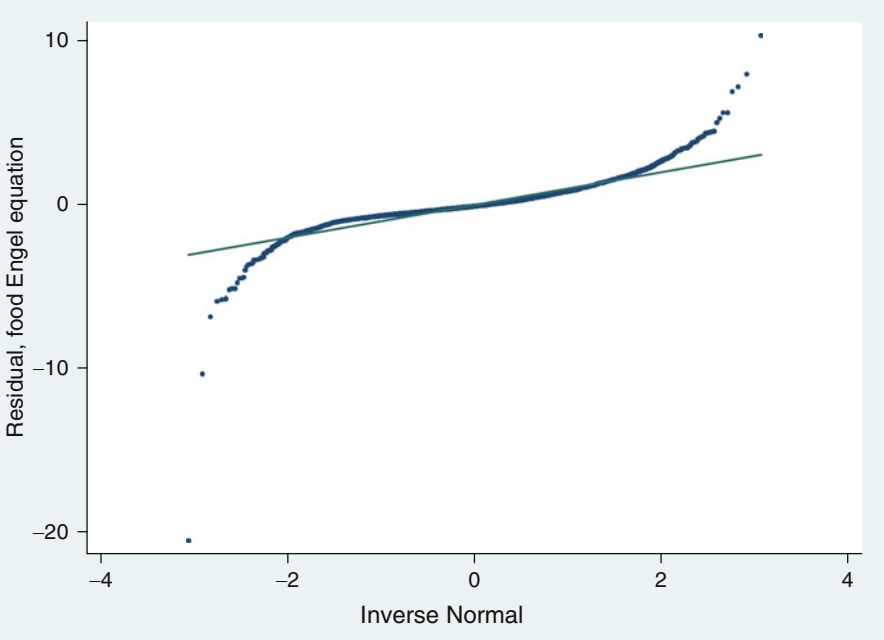


Fig. 2.3 *Q–Q* plot of residuals from food Engel curve, Vietnam, 2006. (*Source*: As for Fig. 2.1)

postestimation command in Stata – the residuals from the linear regression in Fig. 2.1 are sorted from smallest to largest, and the actual quantiles are graphed against the values that would be expected if the residuals were distributed normally. The actual values deviate from the 45° line, which means that the residuals are not quite normally distributed.

2.3 Addressing Regression Problems

Mechanically, it is straightforward to generate regression estimates. In practice, the classical assumptions are frequently violated, and we usually need to adjust our models in order to minimize the harm done by these violations.

2.3.1 Measurement Error

It is all but impossible to measure a variable with complete accuracy – for reasons we discuss more fully in Chap. 3 – and so measurement error is pervasive. For some commonly measured variables, such as household expenditure or income, the error

can be large (see Chap. 10). It even applies, in many cases, to variables that we would expect to be able to measure precisely, such as a person's age; in many surveys these values cluster at higher ages, around numbers such as 70, 75, and 80, while avoiding the ages in between.

If the measurement error is in the dependent variable, we observe y^* rather than the true y. If the error is linear, we have $y^* = y+w$, where w is a vector of (presumably random) errors. So instead of

$$y = \mathbf{X}\beta + e \tag{2.4}$$

we observe

$$y* = \mathbf{X}\beta + (e + w). \tag{2.5}$$

A regression based on (2.5) will still yield unbiased estimates of the coefficients – the classical assumptions have not been violated – but the fit will be poorer, so the value of R^2 will be lower. This is because the signal-to-noise ratio has fallen, making inference less precise.

Measurement error in the independent variables is more serious. If we observe $\mathbf{X^*} = \mathbf{X+w}$ rather than $\mathbf{X}$, then we would base our estimates on

$$y = \mathbf{X^*}\beta + (e - \mathbf{w}\beta). \tag{2.6}$$

Here the coefficient estimates $\hat{\beta}$ will be biased toward 0, because the noise in the $\mathbf{X}$ variables masks the true nature of the dependence of y on $\mathbf{X}$. More precisely, for any single right-hand-side variable,

$$E(\hat{\beta}) = \beta \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} = \beta\left(1 - \frac{\sigma_w^2}{\sigma_X^2 + \sigma_w^2}\right), \tag{2.7}$$

where the σ^2 terms are the true (not estimated) variances. Note that

$$\lim_{\sigma_w \to \infty} E(\hat{\beta}) = 0. \tag{2.8}$$

This means that with greater variance of the measurement error in the x, the estimated coefficient is pushed closer and closer to zero, and we are faced with attenuation bias. Deaton (1997, p. 99) shows how the inclusion of additional correctly measured regressors will generally make the bias worse.

It is not clear how serious these effects are in practice, but here is a simple experiment. Using the observations on 9,189 households from the 2006 Vietnam Living Standards Survey, we regressed the log of food consumption per capita on the log of expenditure per capita. The results are shown in the top panel of Table 2.1, and show a food elasticity of 0.652. Now we introduce an additive random normal

Table 2.1 Illustrating the effects of measurement error

	Coefficient	*t*-statistic	Adjusted R^2
Panel 1: Dependent variable ln(food/cap)			
Constant	−0.230	−32.7	0.732
Ln(expenditure/cap)	0.652	159.3	
Panel 2: Dependent variable ln(food/cap) + error			
Constant	−0.228	−25.1	0.623
Ln(expenditure/cap)	0.652	122.2	
Panel 3: Dependent variable ln(food/cap)			
Constant	−0.129	−16.9	0.661
Ln(expenditure/cap) + error	0.589	133.9	

Notes: 9,189 observations on households, from the Vietnam Living Standards Survey of 2006. Panel 2 adds an error distributed $N(0,0.2)$ to the dependent variable; panel 3 adds an error distributed $N(0,0.02)$ to the independent variable. The mean value of the dependent variable is 1.606, with a standard deviation of 0.614; for the independent variable these are 0.817 and 0.468, respectively

error with mean 0 and standard deviation 0.2. In the second panel we add the shock only to the dependent variable; the equation fits less well, but the food elasticity remains the same (to three decimal places). In the third panel we add the shock only to the independent variable; again, the equation fits somewhat less well, but now the estimate of the food elasticity falls to 0.589, with an apparently tight 95% confidence interval of 0.580–0.598. Errors of the magnitude introduced here are entirely plausible for measures of income and expenditure; by ignoring them we may be lulled into a false sense of confidence in the precision of our estimates.

2.3.2 Omitted Variable Bias

It is rare that the fit of an equation based on cross-sectional survey data – as measured by R^2 – comes close to 1. In other words, variation in y is not completely "explained" by variation in the $\mathbf{X}$ variables. Measurement error aside, this suggests that there may be other explanatory variables, but we have not included them in the model for one reason or another – perhaps they are unobservable (like "ability"), or unobserved (the survey did not ask enough questions), or overlooked. The key point is that some variables, denoted by $\mathbf{Z}$, have been omitted from the regression model. So instead of estimating

$$y = \mathbf{X}\beta + \mathbf{Z}\gamma + \tilde{e} \tag{2.9}$$

we actually estimate

$$y = \mathbf{X}\beta + e. \tag{2.10}$$

In effect we have a compound error term here, because $e = \mathbf{Z}\gamma + \tilde{e}$. This need not interfere with the estimates $\hat{\beta}$ unless the values of the included variables ($\mathbf{X}$) are correlated with the omitted variables ($\mathbf{Z}$), in which case

$$E(\mathbf{Z}\gamma + \tilde{e}|\mathbf{X}) \neq 0,$$

violating the first classical assumption, and making $\hat{\beta}$ biased and inconsistent.

Suppose we are trying to measure the effect of fertilizer on farm rice yields, and estimate

$$y = \beta_0 + \beta_1 X_1$$
$$\text{rice output (kg/ha)} = \beta_0 + \beta_1 \text{ fertilizer input (kg urea/ha)}$$

It is reasonable to think that more-capable and more-dynamic farmers make greater use of fertilizer (so $\text{corr}(X_1 A_1) > 0$, where A_1 refers to ability), and also raise yields in other ways. Thus a better model would be

$$y = \beta_0 + \beta_1 X_1 + \beta_2 A_1 + \tilde{e}.$$

It can be shown that

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \sigma_{X_1 A_1},$$

where $\sigma_{X_1 A_1}$ is the covariance between fertilizer use and (unobserved) ability. In this case, our estimated $\hat{\beta}$ is too large; it is in effect picking up not only the influence of fertilizer use on rice output, but also some of the effect of ability on rice output.

Since almost every model leaves out some relevant variables, omitted-variable bias is pervasive. But is it destructive? There is no simple answer to this, and no substitute for judgment, necessarily on a case-by-case basis. Unfortunately, many omitted variables are unobservable, so one can only speculate about their possible influence on the coefficient estimates.

With panel data it is sometimes possible to attenuate the effects of time-invariant unobservables. Suppose we have information on the rice crop (y) for farmer i for 2 years ($t = 1,2$), and that as before, the crop depends on fertilizer input (X) and the farmer's ability (A). Then we have

$$y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 A_{1it} + \tilde{e}_{it}. \tag{2.11}$$

If A_{1it} ("ability") does not vary over time, then using the time difference operator Δ we have

$$\Delta y_i = \beta_1 \Delta X_{1i} + \Delta \tilde{e}_i, \tag{2.12}$$

and we have purged the data of the unobservable. This is essentially the same as using fixed effects – for instance, a different intercept for each farmer – to sweep away the influence of unobservables. In practice, household-level panel data are relatively uncommon, and the effects of unobservables cannot always be washed

out so easily; we return to these issues in Chap. 9, which deals more thoroughly with panel data.

One can test whether more variables should be included in a regression model by adding the variables to the right-hand side and applying an F test of the null hypothesis that the coefficients are zero. An attractive alternative approach to model building is to use an information criterion: models with more variables have a higher likelihood, but some penalty is appropriate in order to ensure parsimony. In Stata, the Akaike Information Criterion (AIC) is given by

$$\text{AIC} : -2 \ln L + 2K,$$

where L is the likelihood and K is the number of parameters (including the constant), and the Bayesian Information Criterion (BIC) is defined as

$$\text{BIC} : -2 \ln L + K \ln N,$$

where N is the total number of observations. Defined thus, smaller values of the AIC or BIC reflect a "better" model. The AIC tends to overfit, so the BIC is usually the preferred measure (Haughton 1988).

Some scientists, especially in biostatistics, commonly include in their models only variables that are statistically significant at some level (e.g., $p < 0.05$). Stepwise commands, that either trim nonsignificant variables, or add variables from a predetermined list, do this automatically. Most economists prefer to leave, in their models, all the variables that they believe, from theory or experience, to be relevant. This reduces the risk of inconsistent parameter estimates, but also reduces the observed precision of the estimates.

2.3.3 Multicolinearity

Multicolinearity is present when two or more of the right-hand variables in a multiple regression are highly linearly related. In the case where the right-hand variables are completely uncorrelated with one another – i.e., they are mutually orthogonal – then a set of separate simple regressions would yield the same coefficients (except for the constant terms) as a multiple regression that includes all the independent variables. However, this situation almost never arises in practice, which implies that there is nearly always some degree of colinearity among the regressors.

For instance, the nutritional status of a child – measured perhaps by standardized height-for-age, which reflects stunting – may be influenced by the educational achievements of the parents (which we may define as A_{M} for the mother and A_{F} for the father). A basic model would then look like

$$y_i = \beta_0 + \beta_1 A_{\text{M}} + \beta_2 A_{\text{F}} + e. \tag{2.13}$$

However, well-educated men tend to marry well-educated women, so A_{M} and A_{F} are likely to be correlated (Haughton and Haughton 1997).

Multicolinearity does not hurt the fit of the overall equation, but it makes the coefficient estimates imprecise and inaccurate.

A tight-fitting equation where all the coefficient estimates are barely significantly different from zero is a sure sign that multicolinearity is a problem, as is the case where we cannot reject the null hypotheses that the individual coefficients are zero, but we do reject the null hypothesis that the coefficients are jointly zero. Some researchers find it helpful to look at the *variance inflation factor* for each coefficient; this is defined as $1/(1 - R^2_{k.})$, where $R^2_{k.}$ is the R^2 of a regression of variable x_k on all the other right-hand-side variables. High variance inflation factors – typically of 5 to 10 or above – indicate that multicolinearity is a potential problem. Before estimating a regression equation, it is good practice to look at a matrix of correlations among the independent variables. High (absolute) correlations indicate the potential for multicolinearity.

While multicolinearity is not a serious problem in time-series data if the goal is forecasting, it is not always easy to know what to do when it appears in other contexts, including with cross-section data. If X_1 and X_2 are highly correlated, it is tempting to leave out X_1 (or X_2), but this is likely to generate omitted variable bias, which would then attribute the wrong amount of influence to X_1 (or X_2). The textbook solution is to try to find more data, but this is not usually realistic, especially with survey data, where one is presumably using all the data that are available. Greater precision may be possible with Bayesian techniques (see Chap. 8), which bring prior information to bear on the problem, or if theory can lead to a clearer specification of the relevant variables. As data become more available, and as variables multiply and data exploration proliferates, the problems posed by multicolinearity are likely to become more common.

2.3.4 Heteroscedasticity

The second classical regression assumption supposes that the errors in the regression model are homoscedastic; in other words, the errors come from a single distribution that is unrelated to any of the independent variables. When the error term does not have a constant variance we have heteroscedasticity.

In household survey data, the assumption of homoscedasticity is unrealistic. It is far more likely that the observations are distributed as in Fig. 2.1; at low levels of the independent variable X the observations are close to the line, but at higher levels of X the errors are more scattered. This is, of course, only one of the many possible forms that heteroscedasticity may take.

The estimated coefficients are not affected by the presence of heteroscedasticity, but the OLS standard errors will be understated, making the coefficient estimates appear to be more precise than they really are.

Sometimes a straightforward transformation of the data solves the problem. For instance, we know that the distribution of income per capita is highly skewed to the right; if income per capita is used directly as a dependent or independent variable,

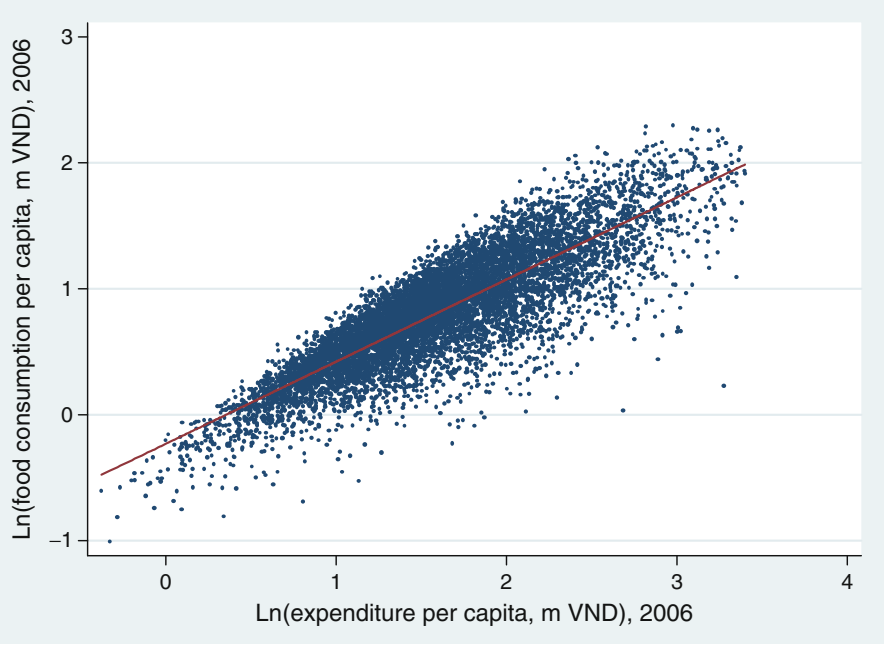


Fig. 2.4 Log of spending on food graphed against log of total spending, Vietnam, 2006. (*Source*: As for Fig. 2.1)

the errors also tend to be skewed (i.e., heteroscedastic). A log transformation of income per capita typically removes most of the skewness, although perhaps not all, since there is some evidence that at the upper end, income follows a Pareto distribution. Figure 2.4 log-transforms the variables from Fig. 2.1; the heteroscedasticity is a bit less apparent, but has by no means all disappeared.

2.3.4.1 Aside: Log Transformations

In passing, it is worth noting that if the dependent variable is in logarithmic form ($\ln y$), care must be taken when predicting values of y. Suppose the model we estimate is

$$\ln y = \mathbf{X}\beta + e. \tag{2.14}$$

From this we can get the predicted value $\widehat{\ln y} = E(\ln y|\mathbf{X}) = \mathbf{X}\hat{\beta}$, but we are typically interested in recovering $E(y|\mathbf{X})$. It is not correct to take $\exp(\widehat{\ln y})$, because $\exp\{E(\ln y)\} \neq E(y)$.

Cameron and Trivedi (2009, p. 103) point out that (2.14) implies that $y = \exp(\mathbf{X}\beta)\exp(e)$, so

$$E(y_i|\mathbf{X}_i) = \exp(\mathbf{X}_i\beta)E\{\exp(e_i)\}. \tag{2.15}$$

If $e_i \sim N(0, \sigma^2)$, then $E\{\exp(e_i)\} = \exp(0.5\sigma^2)$, where we will need to estimate $\exp(0.5\sigma^2)$. Alternatively, if we assume that the e_i are independently and identically distributed (iid), $E\{\exp(e_i)\}$ may be estimated by $\sum_{j=1}^{N} \exp(\widehat{e}_j)/N$. The relevant computations can be done easily enough in Stata, but are not generated automatically.

To illustrate, we return to our earlier example using 2006 Vietnamese household survey data, where we regressed the log of food consumption per capita on the log of expenditure per capita. We have the following (in millions of dong per capita per year):

Mean value of food spending (i.e., y):	2.540
$\exp(\widehat{\ln y})$:	2.467 (incorrect)
$\exp(\widehat{\ln y}) \times \exp(0.5\sigma^2)$:	2.539
$\exp(\widehat{\ln y}) \times \left(\sum_{j=1}^{N} \exp(\widehat{e}_j)/N\right)$:	2.536

By failing to transform correctly from logs to levels, our estimate of food spending would be off by about 3% in this case.

2.3.4.2 Testing for Heteroscedasticity

It is possible to test for heteroscedasticity in a number of ways. White (1980) proposed regressing the squared residuals on the regressors, their squares, and their cross-products; a Lagrange Multiplier test for significance then tests for the presence of any statistically significant patterns. In Stata, just type `whitetst` after running a regression in order to perform this test. It rejects the hypothesis of homoscedasticity for the regressions that are graphed in Figs. 2.1 and 2.4. Breusch and Pagan (1979) allow the researcher to control the variables that are considered likely to cause any heteroscedasticity, but their test is otherwise simple to execute (see the downloadable `bpagan` command in Stata). With large sample sizes, these tests will often reject homoscedasticity, even if it is not a major problem in practice.

It has now become almost standard practice, when using cross-sectional data, to (at a minimum) adjust for heteroscedasticity using White's estimator. Table 2.2 shows the formula for the variance–covariance matrix in the homoscedastic case (the usual default), and for White's robust estimator. The table also shows the formula for the variance–covariance matrix when the observations may be considered to be clustered, a subject we return to in the next section.

In Stata, the robust estimate of the variance–covariance matrix is obtained by adding `vce(robust)` to the `regress` command, as in

```
regress y X1 X2 X3, vce(robust)
```

The adjustment for heteroscedasticity can make a difference, especially when the sample size is not particularly large. Consider the numbers shown in Table 2.3.

Table 2.2 Standard and robust measurement of the variance–covariance matrix

	Variance–covariance matrix – i.e., $\hat{V}(\hat{\beta})$	Stata command
Standard	$s^2(\mathbf{X}'\mathbf{X})^{-1} = \left(\frac{1}{N-k}\sum_i \hat{u}_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}$	[default]
Robust (White)	$(\mathbf{X}'\mathbf{X})^{-1}\left(\frac{N}{N-k}\sum_i \hat{u}_i^2\mathbf{x}_i\mathbf{x}'_i\right)(\mathbf{X}'\mathbf{X})^{-1}$	vce(robust)
Clustering	$(\mathbf{X}'\mathbf{X})^{-1}\left(\frac{G}{G-1}\frac{N}{N-k}\sum_g \mathbf{x}_g\hat{\mathbf{u}}_g\hat{\mathbf{u}}'_g\mathbf{x}'_g\right)(\mathbf{X}'\mathbf{X})^{-1}$	vce(cluster var1)

Notes: Based on a linear regression of the form $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ and residuals $\hat{u}_i = y_i - \mathbf{X}_i\hat{\beta}$. In the case of clustering, there are G clusters denoted by g, $\hat{\mathbf{u}}_g$is the vector of residuals for the gth cluster, and the matrix of regressors for the observations in the gth cluster is given by $\mathbf{x}_g$. "var1" refers to the variable that is used for clustering, often the primary sampling unit (PSU) such as a village. N is the number of observations, and k the number of variables in the model, including the constant

Table 2.3 Comparing estimators

Estimator used	Spending per capita	Spending squared	Motorbikes per capita
OLS			
Coefficient	0.380	−0.005	0.421
Standard error	0.020	0.001	0.177
p-value	0.000	0.000	0.018
White's robust estimator: `vce(robust)`			
Coefficient	0.380	−0.005	0.421
Standard error	0.035	0.002	0.321
p-value	0.000	0.031	0.195
Clustering by commune: `vce(cluster comm)`			
Coefficient	0.380	−0.005	0.412
Standard error	0.022	0.001	0.180
p-value	0.000	0.000	0.023
Sample weights and clustering: `svy: reg`			
Coefficient	0.390	−0.005	0.412
Standard error	0.022	0.001	0.180
p-value	0.000	0.000	0.023
Robust, bootstrapped s.e.: `vce(bootstrap, rep(1000)`			
Coefficient	0.380	−0.005	0.421
Standard error	0.042	0.003	0.330
p-value	0.000	0.095	0.202
Robust, bootstrapped s.e., clustering: `vce(bootstrap, rep(1000) strata(tinh) cluster(comm))`			
Coefficient	0.380	−0.005	0.421
Standard error	0.033	0.003	0.278
p-value	0.000	0.068	0.129

Notes: Dependent variable is food spending (in millions of dong per capita per year). The same variables and data are used in each of these estimates. Based on 582 observations from the Central Highlands region of Vietnam, from the Vietnam Living Standards Survey of 2006. $R^2 = 0.755$ for the OLS regression. Estimates of the constant terms are not shown here

Each triplet of rows shows the coefficients, standard errors, and p-values, for a model of food consumption based on a sample of 582 households in the Central Highlands region of Vietnam in 2006. This is a subsample of the larger Vietnam Living Standards Survey. The model shows food expenditure per capita as a function of total household expenditure per capita, expenditure per capita squared, and a proxy for household wealth given by the number of motorbikes owned per household member. In shorthand, the model is:

$$\text{Food/cap} = a + b\ \text{Exp/cap} + c(\text{Exp/cap})^2 + d\ \text{Motorbikes/cap} \qquad (2.16)$$

The model is likely incomplete, but the purpose here is to illustrate the importance of using the right estimation method.

The first set of results are estimated using plain vanilla OLS. All the coefficients are statistically significant at the 5% level or better, and the fit is solid. But White's test shows heteroscedasticity. The second group of results have been estimated using a robust Huber/White/sandwich estimator. The coefficients are unchanged, but the standard errors are larger, so the t-statistics are smaller. Now the p-values show that the motorbikes/capita variable is not statistically significant. The correction for heteroscedasticity mattered. We will return to this example below in the context of clustering, and weighted regression.

2.3.5 Clustering

Almost every household survey uses a design that involves clustering. The primary sampling unit (PSU) is typically a village or ward, and 10–20 households are chosen for sampling from within each PSU. The problem here is that there are likely to be features of a village, possibly unobserved, that influence all the households in the village together; for instance, if there is a local flood, all villagers may suffer together, or if the local school is good, all the children may achieve excellent test scores. The existence of village-specific effects means that the errors (e_i) of households in a given village will tend to be similar – perhaps they are all too high, or all too low, at the same time.

This violates the third classical assumption, of conditionally uncorrelated observations. While the OLS estimates will not be biased, the OLS standard errors will be too small and the estimates will seem to be more precise than they really are.

The solution is to use a cluster-robust estimator of the variance–covariance matrix of the estimator. In Stata (version 10 or higher), this involves appending `vce(cluster clustervar)` to the `regress` command, where `clustervar` defines the PSU. This estimator also corrects for heteroscedasticity, and should almost always be used with household survey data.

To illustrate the importance of taking clustering into account we return to the model of food consumption in the Central Region of Vietnam, developed above in Sect. 2.3.4. Households were surveyed in clusters at the level of individual communes, and so we allowed for the errors to be correlated within, but

not across, these communes. The results are shown in the third group of numbers in Table 2.3; again, the coefficients are unchanged, but the *t*-statistics are lower than with OLS, as we would expect. With this adjustment, the motorbikes/cap variable is still statistically significantly different from zero, as in the OLS estimation.

An alternative approach to dealing with inference in the presence of heteroscedasticity and clustering is to bootstrap the standard errors. The results of doing this for the food consumption data are also shown in Table 2.3, in panels 5 (for a simple bootstrap) and 6 (for a bootstrap that takes the clustering into account). In both cases the motorbikes/cap variable is not statistically significant. A fuller discussion of the bootstrap is given in Chap. 12.

While clustering at the PSU level is pervasive in household survey data, it may also need to be taken into account at other levels. For instance, a study of child nutrition may have information on siblings; it is likely that they share features that are not fully captured by the model, which implies that the errors may be correlated within households. In this context the household itself is a form of cluster.

2.3.6 *Outliers*

When using survey data, especially data that may not have been thoroughly cleaned, it is not uncommon to encounter outliers. These are observations with an unusual value for either y and/or x. Sometimes they are simply typos – perhaps someone entered 34 instead of 3.4, or an enumerator entered values as percentages (29, 56, etc.) while another entered them as proportions (0.29, 0.56, etc.).

Box-and-whisker plots – discussed more fully in Chap. 1 – can be helpful in identifying outliers. Figure 2.5 shows such plots for data on gross domestic product (GDP) per capita for each of the (then) 53 provinces of Vietnam for 1993. The left-hand panel shows all of the observations, and it is clear that one observation lies far above any of the others; the right-hand panel excludes this outlier, and the result is a more standard-looking plot.

Outliers can have a large impact on regression estimates. Using the same Vietnamese provincial data as in Fig. 2.5, an OLS regression of tax collections per capita on GDP per capita, employing the full data set, yields

$$\begin{aligned} \text{Taxcap} &= -0.806 + 0.647\ \text{gdpcap} \\ &\quad t = -11.4 \quad t = 21.4 \qquad\qquad \bar{R}^2 = 0.898 \end{aligned}$$

where "*t*" refers to the *t*-statistic. On the other hand, if the outlier is excluded, we get a very different equation:

$$\begin{aligned} \text{Taxcap} &= -0.278 + 0.283\ \text{gdpcap} \\ &\quad t = -4.73 \quad t = 8.02 \qquad\qquad \bar{R}^2 = 0.554. \end{aligned}$$

In this case, the lines corresponding to these two estimated equations are shown in Fig. 2.6 along with the underlying data.

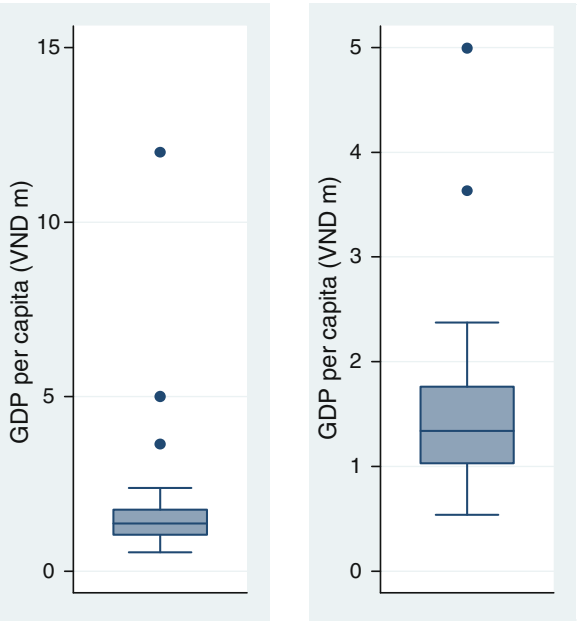


Fig. 2.5 Box-and-Whisker Plots, provincial GDP/capita, Vietnam 1993

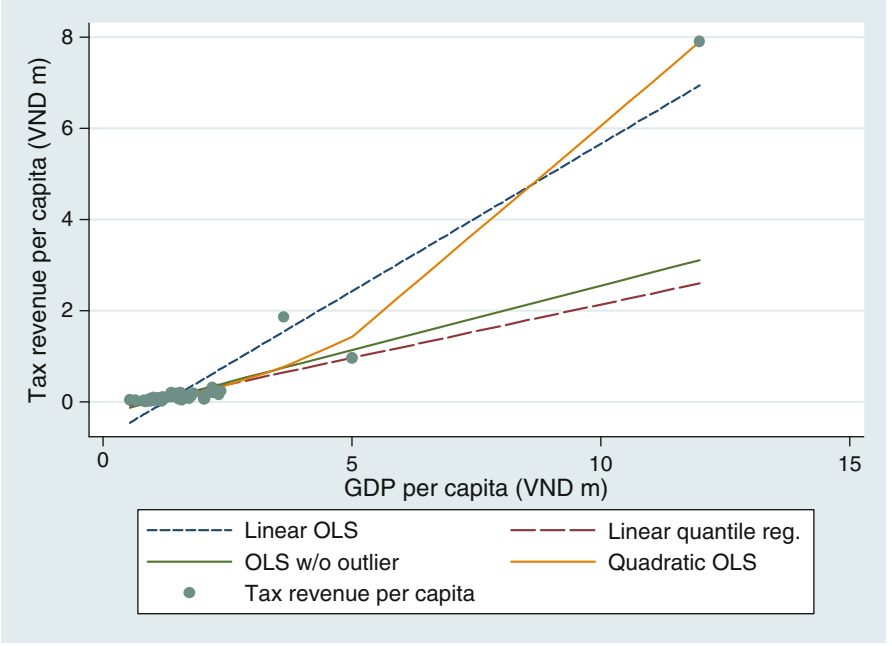


Fig. 2.6 Tax revenue vs. GDP per capita, Vietnamese Provinces, 1993

After the obvious errors have been corrected, there may still be some observations that seem somewhat extreme, and one is left with the problem of what to do. A good next step is to determine whether the outliers are influential enough to matter. There are two distinct concepts here: An *influential observation* is one that, if omitted, leads to an appreciable change in one or more of the coefficient estimates. An *observation with high leverage* is far away from the mean value of the X variable, and so potentially weighs heavily in the estimation of the coefficients.

We may illustrate this with the help of Fig. 2.7, where the top left panel presents the observations from a hypothetical dataset with homoscedastic errors. In Case 1, shown in the top right panel, there is an outlier that has a high value for both X and Y; this outlier has high leverage, but it does not affect the estimated line much, and so has low influence. Case 2, in the southwest panel, shows an outlier that has low leverage, but is somewhat influential, as it pulls the best-fit line down quite noticeably. And in Case 3, in the fourth panel, the outlier has a substantial influence on the estimated line, which makes it influential, even though it does not have a lot of leverage. Figure 2.7 also shows the p-value the Breusch–Pagan test of heteroscedasticity – a low value means that there is clear evidence of heteroskedasticity – as well as a measure of influence (the dfits number), and of leverage.

The leverage h_i of observation i is the ith diagonal entry of the hat matrix

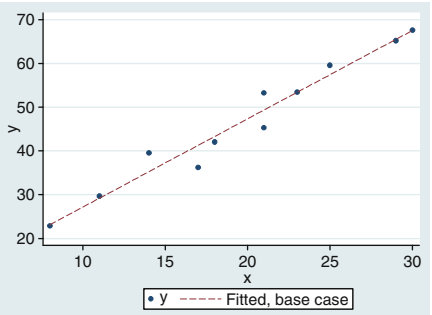
$$H = X(X'X)^{-1}X'. \tag{2.16}$$

The predicted value of y is $\widehat{y} = Hy$, so a large value for h_i implies that observation y_i has a large influence on the predicted value. In Stata it is straightforward to get the values of leverage by typing `predict leverage` after the `regress` command.

A popular measure of influence is dfits_i ("difference in fits"): obtain the predicted value $\widehat{y}_i$ using an OLS regression with, and then without, the ith observation, and take the difference. Cameron and Trivedi (2009, p. 92) suggest that if $|\text{dfits}| > 2\sqrt{K/N}$, then the observation is influential and merits closer examination. Strictly, the dfits measure is only applicable if we can assume homoscedastic errors, and such an assumption is rarely plausible with cross-section data. Nonetheless, the dfits measure can be helpful in practice.

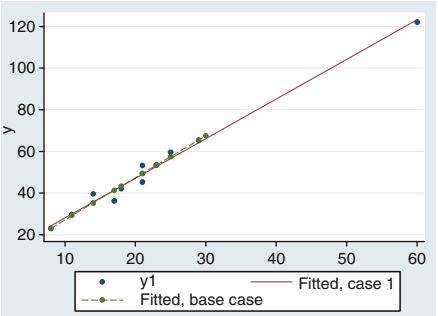
Now we may return to our real-world example, based on data for the provinces of Vietnam in 1993. The median value of leverage 0.08, the standard deviation 1.04, the minimum -0.46, and the maximum 6.95. The relatively high maximum suggests that there is a potential problem with that observation. For this same example, the dfits threshold is 0.39; the actual values range from -1.82 to 23.10, but these are the only two values outside the threshold. Once again we appear to have at least one observation with large influence.

One commonly used solution to the problem of outliers is to truncate the dataset – by excluding the top and bottom 2% of y values, for instance – and estimating the model on the remaining observations. This is only reasonable if we believe that



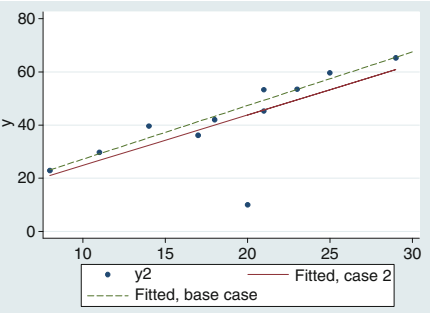
Base case

p-value (HSK)	Max |dfits|	Max leverage
0.62	0.738	0.371



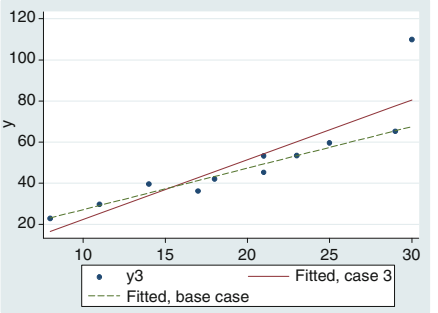
Case 1: High x, Y value

p-value (HSK)	Max |dfits|	Max leverage
0.57	1.63	0.823 (high)



Case 2: Low Y near middle of X

p-value (HSK)	Max |dfits|	Max leverage
0.63	3.56 (high)	0.40



Case 3: High Y near upper tail of X

p-value (HSK)	Max |dfits|	Max leverage
0.01 (significant)	7.28 (very high)	0.37

Fig. 2.7 Influence and leverage illustrated. (*Note*: HSK is heteroskedasticity)

the observations at the extremes are uninformative, for instance because they are considered to be noisy. On the other hand, the extreme observations may often be the most informative ones, which argues against truncating the data too often.

This interpretive problem arises with the Vietnamese data. The outlier is not the result of a typo; it shows the GDP and tax revenue (per capita) for the province of Ba Ria/Vung Tau. This is where most of Vietnam's offshore oil wells are located, and their production, and the tax revenue collected from the oil producers, are included in the statistics for the province. Certainly, one can argue that the situation is anomalous, which would justify ignoring the outlier. On the other hand, the GDP and tax revenues are real, and may have policy lessons that would be missed if the outlier were excluded.

2.3.6.1 Quantile Regression

Another popular regression technique is quantile regression, which has the advantage of being robust to outliers. In a model such as

$$y = \mathbf{X}\beta + e$$

we may think of ordinary least squares as estimating the conditional mean, so $E(Y|\mathbf{X}) = \mathbf{X}\beta$. The OLS estimate, $\widehat{\beta}$, of the coefficients are obtained by minimizing the sum of the squared residuals $\sum_{i=1}^{N} \widehat{e}_i^2 = \sum_{i=1}^{N} \left(y_i - \mathbf{X}_i\widehat{\beta}\right)^2$. Outliers typically generate large values of $\widehat{e}_i$, which are exaggerated by squaring, and so have considerable influence on the $\widehat{\beta}$.

In the same spirit, we may think of quantile regression as estimating the pth percentile $(Y|\mathbf{X}) = \mathbf{X}\beta$. An important special case is that of the 50th percentile (or 50th quantile); this gives us median regression, which finds the $\widehat{\beta}$ by minimizing the sum of the absolute deviations $\sum_{i=1}^{N} |y_i - \mathbf{X}_i\widehat{\beta}|$. The estimates β in this case are less prone than in OLS estimates to being affected by outliers.

Other quantiles are possible: the qth quantile ($q \in (0, 1)$) is the value of y that divides the data so that a proportion q of values are below it and $1 - q$ values above it. When $q = 0.5$, we put the same weight on observations above the median as below it. But if $q = 0.75$, for instance, we will put more weight on observations where $y \geq \mathbf{X}\beta$ than on those for which $y < \mathbf{X}\beta$.

Quantile regression may be implemented straightforwardly in Stata using the `qreg` command; the `bsqreg` version generates bootstrap standard errors (see Chap. 11 for further details about bootstrapping) that allow the standard errors to vary across observations. This is analogous to the use of robust estimation with OLS. The use of quantile regression with panel data is more difficult; we return to this in Chap. 9.

In Fig. 2.8 we show the same data as in Fig. 2.1, where each point represents a household surveyed by the Vietnam Household Living Standards Survey in 2006; the curves show the results of regressing food expenditure per capita on expenditure per capita, and its square. The top curve is the 75th quantile regression line, the bottom curve is the 25th quantile regression line, and the median regression is shown by the solid curve in the middle. The mean regression, based on OLS, is shown by the dotted line. It is clear that all these curves have somewhat different intercepts and curvature.

If the errors are homoscedastic, then the slopes of different quantile regressions should not vary, although the intercepts will. On the other hand, if the slopes do differ, this reflects the presence of heteroskedasticity, as illustrated in Fig. 2.8. It can be helpful to estimate the coefficients of a particular model using each quantile regression – for instance, for percentiles at 10, 20, . . . , 90% – and to graph the estimates and associated standard errors, as done in Fig. 2.9. This example comes from Thailand, where Boonperm et al. (2011) use data from the 2004 Socioeconomic Survey to estimate the effect of borrowing from the Village Fund – a major, decentralized, microcredit scheme – on household incomes. The dependent variable is the log of

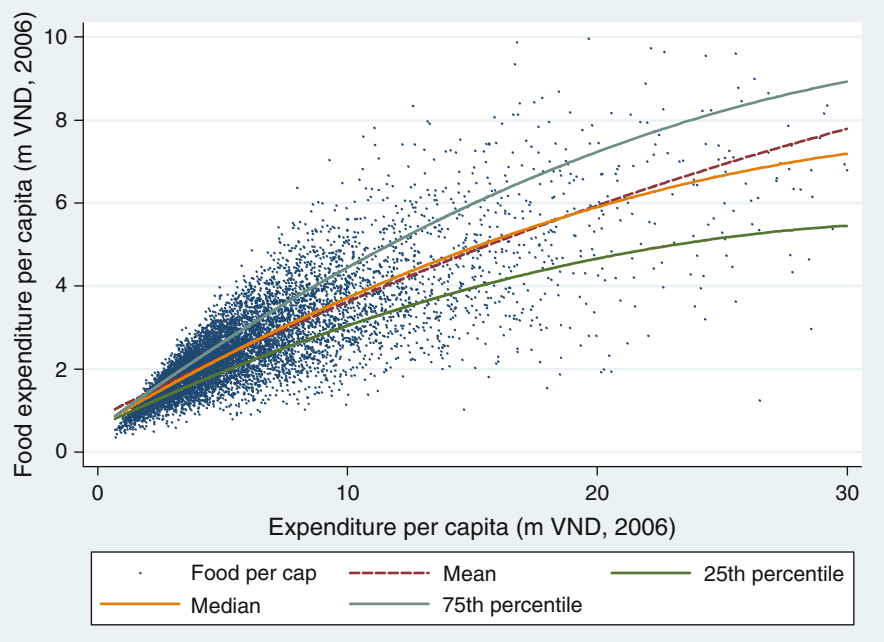


Fig. 2.8 Mean and percentile regressions of food per capita on expenditure per capita, Vietnam, 2006. (*Source*: As for Fig. 2.1)

income per capita, and the independent variable of interest is a dummy variable that equals 1 if the household borrowed from the Village Fund. The coefficient on this treatment variable should measure the impact of borrowing on income, and we are interested in knowing whether it varies across income groups.

Figure 2.9 shows that the average impact of Village Fund borrowing on incomes was a bit above 4%, and varied from 6% at the 10th percentile to just over 3% at the 90th percentile. In other words, this particular microcredit scheme appears to be relatively more influential when the borrower is poor rather than affluent.

An important use of quantile regression is to measure the differences in the marginal impact of policies across income or expenditure groups. A recent example of a distributional analysis of this nature appears in Gamper-Rabindran et al. (2010), who find that, in Brazil, the provision of piped water is particularly effective at reducing infant mortality in areas where the infant mortality rate is especially high (and, typically, incomes relatively low).

2.3.7 Simultaneity

When using regression for inference, we typically assume a direction of causality. For instance, in

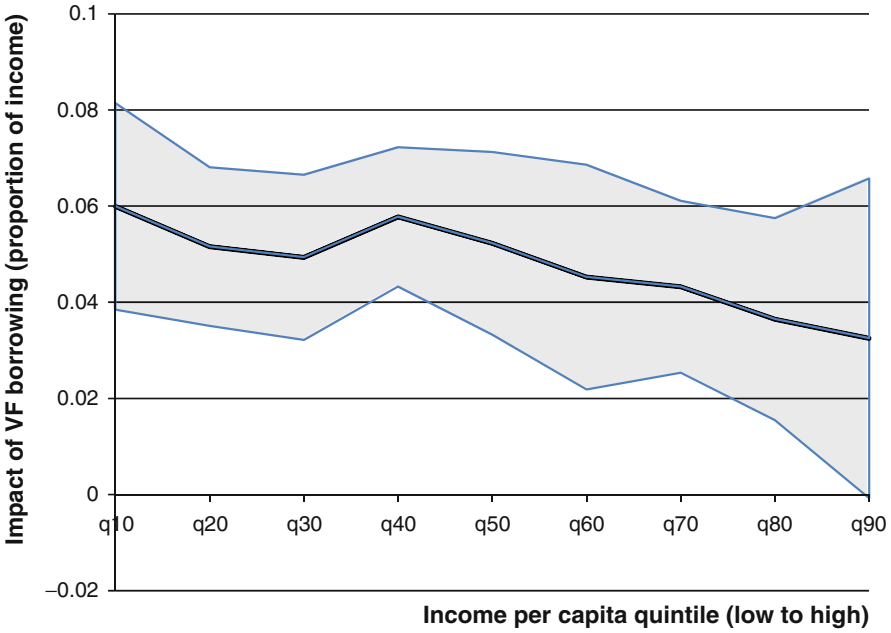


Fig. 2.9 Estimated coefficients from quantile regressions of income per capita on village fund borrowing, Thailand 2004. (*Source*: Thailand Socioeconomic Survey of 2004)

$$\text{Income} = a + b\ (\text{years of education}) + c\ (\text{years of experience}) + e \qquad (2.17)$$

we suppose that more education or experience causes higher income. We defer a fuller discussion of causality to Chap. 5, but two points are worth noting here. First, income, education, and experience may themselves be driven by an outside influence, such as a person's ability; in this case the model in (2.17) has been misspecified. The second point is that to some extent income may lead to more education (such as adult training) or experience (because high-income individuals are more likely to stay in the labor force), rather than be the outcome of these inputs.

Formally, the problem is that simultaneity such as observed here violates classical assumption 2, which states that $E(e_i^2|\mathbf{X}_i) = \sigma^2$. A positive shock ($e_i > 0$) implies a higher value of y_i, but if y_i in turn influences $\mathbf{X}_i$, then e_i and $\mathbf{X}_i$ will be correlated. The $\mathbf{X}_i$ are no longer exogenous, the OLS estimator is inconsistent, and the estimated coefficients (the $\hat{\beta}$) cannot be interpreted as measuring the effect of an exogenous change in $\mathbf{X}$ on a dependent variable y.

Where simultaneity is a concern, one may want to re-think the model. The addition of plenty of other predetermined variables can also attenuate the effects of simultaneity. But probably the most common line of attack is to use instrumental variables, which can also reduce bias related to omitted variables and measurement error.

2.3.7.1 Instrumental Variables

The intuition behind instrumental variables (IV) estimation may be seen as follows. For the ith observation, let y_{1i} be the dependent variable of interest, $\mathbf{X}_i$ be a vector of K_1 exogenous regressors, and $\mathbf{Y}_{2i}$ be a vector of M other endogenous regressors – sometimes referred to as the "troublesome regressors" (Murray 2005). Following Cameron and Trivedi (2009, p. 173), we have the following structural equation:

$$y_{1i} = \mathbf{Y}_{2i}\beta_1 + \mathbf{X}_{1i}\beta_2 + e_i, \quad i = 1, \ldots, N. \tag{2.18}$$

The problem is that while the e_i are uncorrelated with the $\mathbf{X}_{1i}$, they are correlated with the $\mathbf{Y}_{2i}$ (by construction, since this is the problem we are addressing).

Now suppose that we can find some additional variables $\mathbf{X}_2$ that are correlated with the $\mathbf{Y}_2$ but not with the error term in (2.18). Then we could regress each $\mathbf{Y}_2$ variable on the $\mathbf{X}_2$ (and $\mathbf{X}_1$), and use the predicted values from this first stage (i.e., the $\hat{\mathbf{Y}}_2$) in place of the $\mathbf{Y}_2$ when estimating (2.18). These $\hat{\mathbf{Y}}_2$ should be uncorrelated with the e_i – the components that were correlated with the e_i were purged in the first stage – so the second classical regression assumption now holds, and the $\hat{\beta}$ are no longer inconsistent. This is an asymptotic result, however, and IV estimation may not yield good results in small samples.

Formally, the probability limit of the OLS estimator in the case of simple regression is given by (Larcker and Rusticus 2010):

$$\text{plim} b_{\text{OLS}} = \beta + \frac{\text{cov}(x, e)}{\text{var}(x)} = \beta + \frac{\sigma_e}{\sigma_x}\text{corr}(x, e), \tag{2.19}$$

where σ_e and σ_x are the standard deviations of e and x respectively. When the correlation between x and the true error is zero, OLS generates a consistent estimate of β. Now if we use an instrument z for x, we will only obtain a consistent estimate of β if

$$\text{plim} b_{\text{IV}} = \beta + \frac{\text{cov}(z, e)}{\text{cov}(z, x)} = \beta + \frac{\sigma_e}{\sigma_x}\frac{\text{corr}(z, e)}{\text{corr}(z, x)}. \tag{2.20}$$

From (2.19) and (2.20), it follows that IV will only be less biased asymptotically than OLS if $R^2_{ze} < R^2_{xz}R^2_{xe}$, where the R^2 terms refer to squared population correlations; this is not a trivial requirement in practice. It is worth noting too that there is a loss in efficiency when one uses instrumental variables. The asymptotic mean square error of the IV estimate of β is given by

$$\text{asy mse}_{\text{IV}} = \left(\sigma^2_e / n\sigma^2_x\right)\left(1/R^2_{xz}\right)\left(1 + nR^2_{ze}\right) \tag{2.21}$$

(from Bartels, as in Larcker and Rusticus 2010), where n is the sample size. The first term gives the mean square error (MSE) for ordinary least squares; if there is no bias, corr(z,e) = 0; but if $R^2_{xz} = 0.25$, which is entirely plausible in practice, the estimate of the IV standard error will be twice as large as the one generated by ordinary least squares.

Larcker and Rusticus examine the use of IV estimation in the context of models that seek to explain the cost of capital. The problem here is that corporate disclosure of this information is voluntary, so OLS estimates risk being biased. However, they find, using relatively standard data, that the IV estimates may be just as seriously biased as the OLS ones.

The Larker and Rusticus results point to something that is increasingly recognized: Instrumental variables estimation, while mechanically straightforward – for instance, on can use the `ivregress` command in Stata – is hard to pull off successfully in practice. It can be very difficult to find variables that are instrumentally relevant – i.e., influence the $\mathbf{Y}_2$ enough to be "strong" instruments – while also being valid (i.e., uncorrelated with the e_i or, put another way, have no direct influence on y_1). Instrumental relevance is testable, from the correlations between the $\mathbf{Y}_2$ and $\mathbf{X}_2$, but validity is not, because the true regression errors (the e_i) are unobservable.

Ultimately, IV estimation requires creativity, and the persuasiveness of IV estimates is only as strong as the case that can be made for the appropriateness of the choice of instrument.

For IV estimation to work, one needs at least as many instruments as there are endogenous right-hand variables. When there are just enough instruments, we have standard IV estimation; with extra instruments, the model is overidentified, and we get more efficiency by using two-stage least squares, or a variant of the generalized method of moments (GMM).

It is common to test whether the $\mathbf{Y}_2$ variables are indeed endogenous. The underlying principle, due to Hausman, is simple: if the OLS and IV estimators vary by little, then the instrument is not needed. The Hausman test statistic for a single potentially endogenous regressor is distributed $\chi^2(1)$, and is given by

$$T_{\mathrm{H}} = \frac{\left(\hat{\beta}_{\mathrm{IV}} - \hat{\beta}_{\mathrm{OLS}}\right)^2}{\hat{V}(\hat{\beta}_{\mathrm{IV}} - \hat{\beta}_{\mathrm{OLS}})}. \tag{2.22}$$

This test is contingent on the choice of a particular set of instruments, and in that sense does not provide an absolute measure of whether IV estimation is needed.

An interesting example of the use of instrumental variables estimation comes from an article by Narayan and Pritchett (1999), who ask whether "social capital" raises incomes or spending in rural Tanzania. The data come from a Social Capital and Poverty Survey that was undertaken in 1995 as part of a wider poverty assessment; it was merged with data from a 1993 Human Resource Development Survey that covered the same villages (but not the same individuals). Households were asked

about their membership in groups and associations; the nature of these bodies; and about values and attitudes, including trust toward strangers. Narayan and Pritchett constructed an index of social capital by combining the number of groups to which a typical household belonged with characteristics of the groups (and in particularly how broad or inclusive the groups were). There was usable information on 1,376 households from 87 clusters (i.e., villages).

An OLS regression of household spending (HHE) on the index of social capital (SK), using 53 village-level averages, gave the following:

$$\begin{aligned} \text{HHE} &= 0.119.\text{SK} + \text{other terms for exogenous variables} \\ t &= 1.80; \qquad\qquad \text{Adjusted } R^2 = 0.272 \end{aligned}$$

There does seem to be an association between social capital and per capita spending, but which comes first? Can we reasonably say that social capital causes higher spending (and incomes), or does causality run in the other direction?

Narayan and Pritchett argue that the variable that measures "trust in strangers" (TIS) is unlikely to be caused by affluence, and on these grounds they use it as an instrument for social capital. In other words, they regress SK on TIS (and the other exogenous variables) and use the predicted value of SK (SKhat) instead of actual SK in the spending equation. The result is as follows:

$$\begin{aligned} \text{HHE} &= 0.496\ \text{SKhat}\ +\ \text{other terms for exogenous variables} \\ t &= 2.75 \end{aligned}$$

This is a stronger result than that of the OLS regression. They interpret it as showing that social capital causes more household spending, rather than the reverse; and argue that the smaller coefficient in the OLS regression reflects large measurement error in social capital, which biased that coefficient toward zero.

Whether the instrumental variables result is plausible is a matter of judgment. Perhaps more affluent villages are less wary of strangers, in which case the instrument is not exogenous. It was also less successful in the household-level (as opposed to village-level) estimates relating social capital to expenditure.

2.4 Conclusion

The literature on regression is vast, so in this chapter we have chosen to emphasize those aspects of the subject that, in our experience, are important, yet sometimes overlooked.

An appreciation of the power and limitations of regression comes with practice. This chapter gets one going; the rest of the book goes beyond basic regression applied to cross-sectional data, and introduces a wide variety of other techniques that are helpful in the study of living standards survey data.

References

Boonperm, Jirawan, Jonathan Haughton, and Shahidur R. Khandker. 2011. Does the village fund matter in Thailand? Suffolk University, Boston.

Breusch, T., and A. Pagan. 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47: 1287–1294.

Cameron, Colin, and Pravin Trivedi. 2009. *Microeconometrics using Stata*. College Station: Stata Press.

Deaton, Angus. 1997. *The analysis of household surveys: A microeconometric approach to development policy*. Baltimore: Johns Hopkins University Press.

Gamper-Rabindran, Shanti, Shakeep Khan, and Christopher Timmins. 2010. The impact of piped water provision on infant mortality in Brazil: A quantile panel data approach. *Journal of Development Economics* 92: 188–200.

Greene, William. 2011. *Econometric analysis*, 7th ed. Upper Saddle River: Prentice Hall.

Haughton, Dominique. 1988. On the choice of a model to fit data from an exponential family. *Annals of Statistics* 16: 342–355.

Haughton, Dominique, and Jonathan Haughton. 1997. Explaining child nutrition in Vietnam. *Economic Development and Cultural Change* 45(3): 541–556.

Haughton, Jonathan, and Shahidur Khandker. 2009. *Handbook on poverty and inequality*. Washington, DC: World Bank.

Larcker, David F., and Tjomme O. Rusticus. 2010. On the use of instrumental variables in accounting research. *Journal of Accounting and Economics* 49(3): 186–205.

Murray, Michael P. 2005. *The bad, the weak, and the ugly: Avoiding the pitfalls of instrumental variables estimation*. Lewiston: Bates College.

Narayan, Deepa, and Lant Pritchett. 1999. Cents and sociability: household income and social capital in rural Tanzania. *Economic Development and Cultural Change* 47(4): 871–897.

White, H. 1980. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* 48: 817–838.