The Analysis of Household Surveys

A Microeconometric Approach to Development Policy

Angus Deaton

Published for the World Bank
The Johns Hopkins University Press
Baltimore and London

2 *Econometric issues for survey data*

This chapter, like the previous one, lays groundwork for the analysis to follow. The approach is that of a standard econometric text, emphasizing regression analysis and regression "diseases" but with a specific focus on the use of survey data. The techniques that I discuss are familiar, but I focus on the methods and variants that recognize that the data come from surveys, not experimental data nor time series of macroeconomic aggregates, that they are collected according to specific designs, and that they are typically subject to measurement error. The topics are the familiar ones; dependency and heterogeneity in regression residuals, and possible dependence between regressors and residuals. But the reasons for these problems and the contexts in which they arise are often specific to survey data. For example, the weighting and clustering issues with which I begin do not occur except in survey data, although the methodology has straightforward parallels elsewhere in econometrics.

What might be referred to as the "econometric" approach is not the only way of thinking about regressions. In Chapter 3 and at several other points in this book, I shall emphasize a more statistical and descriptive methodology. Since the distinction is an important one in general, and since it separates the material in this chapter from that in the next, I start with an explanation. The statistical approach comes first, followed by the econometric approach. The latter is developed in this chapter, the former in Chapter 3 in the context of substantive applications.

From the statistical perspective, a regression or "regression function" is defined as an expectation of one variable, conventionally written y, conditional on another variable, or vector of variables, conventionally written x. I write this in the standard form

$$m(x) = E(y|x) = \int_{-\infty}^{\infty} y \, dF_c(y|x) \tag{2.1}$$

where F_c is the distribution function of y conditional on x. This definition of a regression is descriptive and carries no behavioral connotation. Given a set of variables (y,x) that are jointly distributed, we can pick out one that is of interest, in this case y, compute its distribution conditional on the others, and calculate the associated regression function. From a household survey, we might examine the

regression of per capita expenditure (y) on household size (x), which would be equivalent to a tabulation of mean per capita expenditure for each household size. But we might just as well examine the reverse regression, of household size on per capita expenditure, which would tell us the average household size at different levels of resources per capita. In such a context, the estimation of a regression is precisely analogous to the estimation of a mean, albeit with the complication that the mean is conditioned on the prespecified values of the x-variables. When we think of the regression this way, it is natural to consider not only the conditional mean, but other conditional measures, such as the median or other percentiles, and these different kinds of regression are also useful, as we shall see below. Thinking of a regression as a set of means also makes it clear how to incorporate into regressions the survey design issues that I discussed at the end of Chapter 1.

When the conditioning variables in the regression are continuous, or when there is a large number of discrete variables, the calculations are simplified if we are prepared to make assumptions about the functional form of $m(x)$. The most obvious and most widely used assumption is that the regression function is linear in x,

$$m(x) = \beta' x \tag{2.2}$$

where β is a scalar or vector as x is a scalar or vector, and where, by defining one of the elements of x to be a constant, we can allow for an intercept term. In this case, the β-parameters can be estimated by ordinary least squares (OLS), and the estimates used to estimate the regression function according to (2.2).

The econometric approach to regression is different, in rhetoric if not in reality. The starting point is usually the linear regression model

$$y = \beta' x + u \tag{2.3}$$

where u is a "residual," "disturbance," or "error" term representing omitted determinants of y, including measurement error, and satisfying

$$E(u|x) = 0. \tag{2.4}$$

The combination of (2.3) and (2.4) implies that $\beta' x$ is the expectation of y conditional on x, so that (2.3) and (2.4) imply the combination of (2.1) and (2.2). Similarly, because a variable can always be written as its expectation plus a residual with zero expectation, the combination of (2.1) and (2.2) imply the combination of (2.3) and (2.4). As a result, the statistical and econometric approaches are formally identical. The difference lies in the rhetoric, and particularly in the contrast between "model" and "description." The linear regression as written in (2.3) and (2.4) is often thought of as a model of determination, of how the "independent" variables x determine the "dependent" variable y. By contrast, the regression function (2.1) is more akin to a cross-tabulation, devoid of causal significance, a descriptive device that is (at best) a preliminary to more "serious," or model-based, analysis.

A good example of the difference comes from the analysis of poverty, where regression methods have been applied for a very long time (see Yule 1899). Suppose that the variable y_i is 1 if household i is in poverty and is 0 if not. Suppose that the conditioning variables x are a set of dummy variables representing regions of a country. The coefficients of a linear regression of y on x are then a "poverty profile," the fractions of households in poverty in each of the regions. These results could also have been represented by a table of means by region, or a regression function. A poverty profile can incorporate more than regional information, and might include local variables, such as whether or not the community has a sealed road or an irrigation system, or household variables, such as the education of the household head. Such regressions answer questions about differences in poverty rates between irrigated and unirrigated villages, or the extent to which poverty is predicted by low education. They are also useful for targeting antipoverty policies, as when transfers are conditioned on geography or on landholding (see, for example, Grosh 1994 or Lipton and Ravallion 1995.) Of course, such descriptions are not informative about the *determinants* of poverty. Households in communities with sealed roads may be well-off because of the trade brought by the road, or the road may be there because the inhabitants have the economic wherewithal to pay for it, or the political power to have someone else do so. Correlation is not causation, and while poverty regressions are excellent tools for constructing poverty profiles, they do not measure up to the more rigorous demands of project evaluation.

Much of the theory and practice of econometrics consists of the development and use of tools that permit causal inference in nonexperimental data. Although the regression of individual poverty on roads cannot tell us whether or by how much the construction of roads will reduce poverty, there exist techniques that hold out the promise of being able to do so, if not from an OLS regression, at least from an appropriate modification. Econometric theorists have constructed a catalog of regression "diseases," the presence of any of which can prevent or distort correct inference of causality. For each disease or combination of diseases, there exist techniques that, at least under ideal conditions, can repair the situation. Econometrics texts are largely concerned with these techniques, and their application to survey data is the main topic of this chapter.

Nevertheless, it pays to be skeptical and, in recent years, many economists and statisticians have become increasingly dissatisfied with technical fixes, and in particular, with the strong assumptions that are required for them to work. In at least some cases, the conditions under which a procedure will deliver the right answer are almost as implausible, and as difficult to validate, as those required for the original regression. Readers are referred to the fine skeptical review by Freedman (1991), who concludes "that statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings." One of my aims in this chapter is to clarify the often rather limited conditions under which the various econometric techniques work, and to indicate some more realistic alternatives, even if they promise less. A good starting point for all econometric work is the (obvious) realization that it is not always

possible to make the desired inferences with the data to hand. Nevertheless, even if we must sometimes give up on causal inference, much can be learned from careful inspection and description of data, and in the next chapter, I shall discuss techniques that are useful and informative for this more modest endeavor.

This chapter is organized as follows. There are nine sections, the last of which is a guide to further reading. The first two pick up from the material at the end of Chapter 1 and look at the role of survey weights (Section 2.1) and clustering (Section 2.2) in regression analysis. Section 2.3 deals with the fact that regression functions estimated from survey data are rarely homoskedastic, and I present briefly the standard methods for dealing with the fact. Quantile regressions are useful for exploring heteroskedasticity (as well as for many other purposes), and this section contains a brief presentation. Although the consequences of heteroskedasticity are readily dealt with in the context of regression analysis, the same is not true when we attempt to use the various econometric methods designed to deal with limited dependent variables. Section 2.4 recognizes that survey data are very different from the controlled experimental data that would ideally be required to answer many of the questions in which we are interested. I review the various econometric problems associated with nonexperimental data, including the effects of omitted variables, measurement error, simultaneity, and selectivity. Sections 2.5 and 2.6 review the uses of panel data and of instrumental variables (IV), respectively, as a means to recover structure from nonexperimental data. Section 2.7 shows how a time series of cross-sectional surveys can be used to explore changes over time, not only for national aggregates, but also for socioeconomic groups, especially age cohorts of people. Indeed, such data can be used in ways that are similar to panel data, but without some of the disadvantages—particularly attrition and measurement error. I present some examples, and discuss some of the associated econometric issues. Finally, section 2.8 discusses two topics in statistical inference that will arise in the empirical work in later chapters.

2.1 Survey design and regressions

As we have already seen in Section 1.1, there are both statistical and practical reasons for household surveys to use complex designs in which different households have different probabilities of being selected into the sample. We have also seen that such designs have to be taken into account when calculating means and other statistics, usually by weighting, and that the calculation of standard errors for the estimates should depend on the sample design. We also saw that, standard errors can be seriously misleading if the sample design is not taken into account in their calculation, particularly in the case of clustered samples. In this section, I take up the same questions in the context of regressions. I start with the use of weights, and with the old and still controversial issue of whether or not the survey weights should be used in regression. As we shall see, the answer depends on what one thinks about and expects from a regression, and on whether one takes an econometric or statistical view. I then consider the effects of clustering, and show that there is no ambiguity about what to do in this case; standard errors should be cor-

rected for the design. I conclude the section with a brief overview of regression standard errors and sample design, going beyond clustering to the effects of stratification and probability weighting.

Weighting in regressions

Consider a sample in which households belong to one of S "sectors," and where the probability of selection into the sample varies from sector to sector. In the simplest possible case, there are two sectors, for example, rural and urban, the sample consists of rural and urban households, and the probability of selection is higher in the urban sector. The sectors will often be sample strata, but my concern here is with variation in weights across sectors—however defined—and not directly with stratification. If the means are different by sector, we know that the unweighted sample mean is a biased and inconsistent estimator of the population mean, and that a consistent estimator can be constructed by weighting the individual observations by inflation factors, or equivalently, by computing the means for each sector, and weighting them by the fractions of the population in each. The question is whether and how this procedure extends from the estimation of means to the estimation of regressions.

Suppose that there are N_s population households and n_s sample households in sector s. With simple random sampling within sectors, the inflation factor for a household i in s is $w_{is} = N_s/n_s$, so that the weighted mean (1.25) is

$$\bar{x}_w = \frac{\sum_{s=1}^{S}\sum_{i=1}^{n_s} w_{is}x_{is}}{\sum_{s=1}^{S}\sum_{i=1}^{n_s} w_{is}} = \frac{\sum_{s=1}^{S} N_s\bar{x}_s}{\sum_{s=1}^{S} N_s} = \sum_{s=1}^{S}\frac{N_s}{N}\bar{x}_s = \bar{x}. \tag{2.5}$$

Hence, provided that the sample means for each sector are unbiased for the corresponding population means, so is the weighted mean for the overall population mean. Equation (2.5) also shows that it makes no difference whether we take a weighted mean of individual observations with inflation factors as weights, or whether we compute the sector means first, and then weight by population shares.

Let us now move to the case where the parameters of interest are no longer population totals or means, but the parameters of a linear regression model. Within each sector $s = 1, \ldots, S$,

$$y_s = X_s\beta_s + u_s \tag{2.6}$$

and, in general, the parameter vectors β_s differ across sectors. In such a case, we might decide, by analogy with the estimation of means, that the parameter of interest is the population-weighted average

$$\beta = N^{-1}\sum_{s=1}^{S} N_s\beta_s. \tag{2.7}$$

Consider the only slightly artificial example where the regressions are Engel curves for a subsidized food, such as rice, and we are interested in the effects of a general increase in income on the aggregate demand for rice, and thus on the total cost of the subsidy. If the marginal propensity to spend on rice varies from one sectors to another, then (2.7) gives the population average, which is the quantity that we need to know.

Again by analogy with the estimation of means, we might proceed by estimating a separate regression for each sector, and weighting them together using the population weights. Hence,

$$\hat{\beta} = \sum_{s=1}^{S} \frac{N_s}{N} \hat{\beta}_s, \quad \hat{\beta}_s = (X_s'X_s)^{-1}X_s'y_s. \tag{2.8}$$

Such regressions are routinely calculated when the sectors are broad, such as in the urban versus rural example, and where there are good prior reasons for supposing that the parameters differ across sectors. Such a procedure is perhaps less attractive when there is little interest in the individual sectoral parameter estimates, or when there are many sectors with few households in each, so that the parameters for each are estimated imprecisely. But such cases arise in practice; some sample designs have hundreds of strata, chosen for statistical or administrative rather than substantive reasons, and we may not be sure that the parameters are the same in each stratum. If so, the estimator (2.8) is worth consideration, and should not be rejected simply because there are few observations per stratum. If the strata are independent, the variance of $\hat{\beta}$ is

$$V(\hat{\beta}) = \sum_{s=1}^{S} \left(\frac{N_s}{N}\right)^2 V(\hat{\beta}_s) = \sum_{s=1}^{S} \left(\frac{N_s}{N}\right)^2 \sigma_s^2 (X_s'X_s)^{-1} \tag{2.9}$$

where σ_s^2 is the residual variance in stratum s. Because the population fractions in (2.9) are squared, $\hat{\beta}$ will be more precisely estimated than are the individual $\hat{\beta}_s$.

Instead of estimating parameters sector by sector, it is more common to estimate a regression from all the observations at once, either using the inflation factors to calculate a weighted least squares estimate, or ignoring them, and estimating by unweighted OLS. The latter can be written

$$\hat{\beta} = \left(\sum_{s=1}^{S} X_s'X_s\right)^{-1} \left(\sum_{s=1}^{S} X_s'y_s\right). \tag{2.10}$$

In general, the OLS estimator will not yield any parameters of interest. Suppose that, as the sample size grows, the moment matrices in each stratum tend to finite limits, so that we can write

$$\operatorname*{plim}_{n_s \to \infty} n_s^{-1} X_s'X_s = M_s; \quad \operatorname*{plim}_{n_s \to \infty} n_s^{-1} X_s'y_s = c_s = M_s\beta_s \tag{2.11}$$

where M_s and c_s are nonrandom and the former is positive definite. (Note that, as in Chapter 1, I am assuming sampling with replacement, so that it is possible to sample an infinite number from a finite population.) By (2.11), the probability limit of the OLS estimator (2.10) is

$$\text{(2.12)} \qquad \text{plim}\hat{\beta} = \left(\sum_{s=1}^{S} (n_s/n) M_s \right)^{-1} \sum_{s=1}^{S} (n_s/n) c_s$$

where I have assumed that, as the sample size grows, the proportions in each sector are held fixed. If all the β_s are the same, so that $c_s = M_s\beta$ for all s, then the OLS estimator will be consistent for the common β. However, even if the structure of the explanatory variables is the same in each of the sectors, so that $M_s = M$ for all s and $c_s = M\beta_s$, equation (2.12) gives the *sample-weighted* average of the β_s, which is inconsistent unless the sample is a simple random sample with equal probabilities of selection in all sectors.

The inconsistency of the OLS estimator for the population parameters mirrors the inconsistency of the unweighted mean for the population mean. Consider then the regression counterpart of the weighted mean, in which each household's contribution to the moment matrices is inflated using the weights,

$$\text{(2.13)} \qquad \hat{\beta}_w = \left(\sum_{s=1}^{S} \sum_{i=1}^{n_s} w_{is} x_{is} x_{is}' \right)^{-1} \left(\sum_{s=1}^{S} \sum_{i=1}^{n_s} w_{is} x_{is} y_{is} \right)$$

where x_{is} is the vector of explanatory variables for household i in sector s, and y_{is} is the corresponding value of the dependent variable. In this case, the weights are N_s/n_s and vary only across sectors, so that the estimator can also be written as

$$\text{(2.14)} \qquad \hat{\beta}_w = \left(\sum_{s=1}^{S} \frac{N_s}{n_s} X_s' X_s \right)^{-1} \left(\sum_{s=1}^{S} \frac{N_s}{n_s} X_s' y_s \right) = (X'WX)^{-1} X'Wy$$

where X and y have their usual regression connotations—the X_s and y_s matrices from each sector stacked vertically—and W is an $n \times n$ matrix with the weights N_s/n_s on the diagonal and zeros elsewhere. This is the weighted regression that is calculated by regression packages, including STATA.

If we calculate the probability limits as before, we get instead of (2.12)

$$\text{(2.15)} \qquad \text{plim}\hat{\beta}_w = \left(\sum_{s=1}^{S} \frac{N_s}{N} M_s \right)^{-1} \sum_{s=1}^{S} \frac{N_s}{N} M_s \beta_s$$

so that, where we previously had *sample* shares as weights, we now have *population* shares. The weighted estimator thus has the (perhaps limited) advantage over the OLS estimator of being independent of sample design; the right-hand side of (2.15) contains only population magnitudes. Like the OLS estimator it is consistent if all the β_s are identical, and unlike it, will also be consistent if the M_s matrices are identical across sectors. We have already seen one such case; when there is only a constant in the regression, $M_s = 1$ for all s, and we are estimating the population mean, where weighting gives the right answer. But it is hard to think of other realistic examples in which the M_s are common and the c_s differ. In general, the weighted estimator will not be consistent for the weighted sum of the parameter vectors because

$$(2.16)\quad \left(\sum_{s=1}^{S}(N_s/N)M_s\right)^{-1}\sum_{s=1}^{S}(N_s/N)c_s \neq \sum_{s=1}^{S}(N_s/N)M_s^{-1}c_s = \sum_{s=1}^{S}(N_s/N)\beta_s = \beta.$$

In this case, which is probably the typical one, there is no straightforward analogy between the estimation of means and the estimation of regression parameters. The weighted estimator, like the OLS estimator, is inconsistent.

As emphasized by Dumouchel and Duncan (1983), the weighted OLS estimator will be consistent for the parameters that would have been estimated using census data; as usual, the weighting makes the sample look like the population and removes the dependence of the estimates on the sample design, at least when samples are large enough. However, the difference in parameter values across strata is a feature of the population, not of the sample design, so that running a regression on census data is no less problematic than running it on sample data. In neither case can we expect to recover parameters of interest. The issue is not sample design, but population heterogeneity. Of course, if the population is homogeneous, so that the regression coefficients are identical in each stratum, both weighted and unweighted estimators will be consistent. In such a case, and in the absence of other problems, the unweighted OLS estimator is to be preferred since, by the Gauss-Markov theorem, least squares is more efficient than the weighted estimator. This is the classic econometric argument against the weighted estimator: when the sectors are homogeneous, OLS is more efficient, and when they are not, both estimators are inconsistent. In neither case is there an argument for weighting.

Even so, it is possible to defend the weighted estimator. I present one argument that is consistent with the modeling point of view, and one that is not. Suppose that there are many sectors, that we suspect heterogeneity, but the heterogeneity is not systematically linked to the other variables. Consider again the probability limit of the weighted estimator, (2.15), substitute $c_c = M_s\beta_s$, and write $\beta_s = \beta + (\beta_s - \beta)$ to reach

$$(2.17)\quad \text{plim}\hat{\beta}_w = \beta + \left(\sum_{s=1}^{S}(N_s/N)M_s\right)^{-1}\sum_{s=1}^{S}(N_s/N)M_s(\beta_s - \beta).$$

The weighted estimate will therefore be consistent for β if

$$(2.18)\quad \sum_{s=1}^{S}(N_s/N)M_s(\beta_s - \beta) = 0.$$

This will be the case if the variation in the parameters across sectors is random and is unrelated to the moment matrices M_s in each, and if the number of sectors is large enough for the weighted mean to be zero. The same kind of argument is much harder to make for the unweighted (OLS) estimator. The orthogonality condition (2.18) is a condition on the *population*, while the corresponding condition for the OLS estimator would have to hold for the *sample*, so that the estimator would (at best) be consistent for only some sampling schemes. Even then, its probability limit would not be β but the sample-weighted mean of the sector-specific β_s, a quantity that is unlikely to be of interest.

Perhaps the strongest argument for weighted regression comes from those who regard regression as descriptive, not structural. The case has been put forcefully by Kish and Frankel (1974), who argue that regression should be thought of as a device for summarizing characteristics of the population, heterogeneity and all, so that samples ought to be weighted and regressions calculated according to (2.13) or (2.14). A weighted regression provides a consistent estimate of the population regression function—provided of course that the assumption about functional form (in this case that it is linear) is correct. The argument is effectively that the regression function itself is the object of interest. I shall argue in the next chapter that this is frequently the case, both for the light that the regression function sometimes sheds on policy, and when not, as a preliminary description of the data. Of course, if we are trying to estimate behavioral models, and if those models are different in different parts of the population, the classic econometric argument is correct, and weighting is at best useless.

Recommendations for practice

How then should we proceed? Should the weights be ignored, or should we use them in the regressions? What about standard errors? If regressions are primarily descriptive, exploring association by looking at the mean of one variable conditional on others, the answer is straightforward: use the weights and correct the standard errors for the design. For modelers who are concerned about heterogeneity and its interaction with sample design, matters are somewhat more complicated.

For descriptive purposes, the only issue that I have not dealt with is the computation of standard errors. In principle, the techniques of Section 1.4 can be used to give explicit formulas that take into account the effect of survey design on the variance-covariance matrices of parameter estimates. At the time of writing, such formulas are being incorporated into STATA. Alternatively, the bootstrap provides a computationally intensive but essentially mechanical way of calculating standard errors, or at least for checking that the standard errors given by the conventional formulas are not misleading. As in Section 1.4, the bootstrap should be programmed so as to reflect the sample design: different strata should be bootstrapped separately and, for two-stage samples, bootstrap draws should be made of clusters or primary sampling units (PSUs), not of the households within them. Because hypothetical replications of the survey throw up new households at each replication, with new values of x's as well as y's, the bootstrap should do the same. In this context, it makes no sense to condition on the original x's, holding them fixed in repeated samples. Instead, each bootstrap sample will contain a resampling of households, with their associated x's, y's, and weights w's, and these are used to compute each bootstrap regression.

In practice, the design feature that usually has the largest effect on standard errors is clustering, and the most serious problem with the conventional formulas is that they overstate precision by ignoring the dependence of observations within the same PSU. We have already seen this phenomenon for estimation of the mean

in Section 1.4, and it is sufficiently important that I shall return to it in Section 2.2 below. It is as much an issue for structural estimation as it is for the use of regressions as descriptive tools.

The regression modeler has a number of different strategies for dealing with heterogeneity and design. At one extreme is what might be called the standard approach. Behavior is assumed to be homogeneous across (statistical or substantive) subunits, the data are pooled, and the weights ignored. The other extreme is to break up the sample into cells whenever behavior is thought likely to differ or where the sampling weights differ across groups. Separate regressions are then estimated for each cell and the results combined using population weights according to (2.8). When the distinctions between groups are of substantive interest—as will often be the case, since regions, sectors, or ethnic characteristics are often used for stratification—it makes sense to test for differences between them using covariance analysis, as described, for example, by Johnston (1972, pp. 192–207).

When adopting the standard approach, it is also wise to adopt Dumouchel and Duncan's suggestion of calculating both weighted and unweighted estimators and comparing them. Under the null that the regressions are homogeneous across strata, both estimators are unbiased, so that the difference between them has an expectation of zero. By contrast, when heterogeneity and design effects are important, the two expectations will differ. The difference between the weighted estimator (2.13) and the OLS estimator can be written as

$$\begin{aligned} \hat{\beta}_w - \hat{\beta}_{OLS} &= (X'WX)^{-1}X'Wy - (X'X)^{-1}X'y \\ &= (X'WX)^{-1}X'W(I - X(X'X)^{-1}X')y \\ &= (X'WX)^{-1}X'WM_X y \end{aligned} \tag{2.19}$$

where M_X is the matrix $I - X(X'X)^{-1}X'$. By (2.19) the difference between the two estimators is the vector of parameter estimates from a weighted regression of the unweighted OLS residuals on the x's. Its variance-covariance matrix can readily be calculated in order to form a test statistic, but the easiest way to test whether (2.19) is zero is to run the "auxiliary" regression

$$y = Xb + WXg + v \tag{2.20}$$

and to use an F-statistic to test $g = 0$ (see also Davidson and MacKinnon 1993. pp. 237–42, who discuss Hausman (1978) tests, of which this is a special case).

In the case of many sectors, when we rely on the interpretation that the intersectoral heterogeneity is random variation in the parameters as in (2.17) above, note that the residuals of the regressions, whether weighted or unweighted, will be both heteroskedastic and dependent. Rewrite the regressions (2.6) as

$$y_s = X_s\beta + X_s(\beta_s - \beta) + u_s = X_s\beta + \xi_s \tag{2.21}$$

where β is defined in (2.7) and where the compound residual ξ_s is defined by the second equality. If the intrasectoral variance-covariance matrix of the β_s is Ω_β,

the variances and covariances of the new residuals are zero between residuals in different sectors, while within each sector we have

$$E(\xi_s \xi_s') = X_s \Omega_\beta X_s' + \sigma^2 I_{n_s} \tag{2.22}$$

where I_{n_s} is the $n_s \times n_s$ identity matrix. Hence, if the different sectors in (2.21) are combined, or "stacked," into a single regression, the variance-covariance matrix of the residuals will have a block diagonal structure, displaying both heteroskedasticity and intercorrelation. In such circumstances, neither the weighted nor unweighted regressions will be efficient, and perhaps more seriously, the standard formulas for the estimated standard errors will be incorrect. In the next two sections, we shall see how to detect and deal with these problems in a slightly different but mathematically identical context.

2.2 The econometrics of clustered samples

In Chapter 1, we saw that most household surveys in developing countries use a two-stage design, in which clusters or PSUs are drawn first, followed by a selection of households from within each PSU. In Section 1.4, I explored the consequences of clustered designs for the estimation of means and their standard errors. Here I discuss the use of clusters in empirical work more broadly. When the survey data are gathered from rural areas in developing countries, the clustering is often of substantive interest in its own right. I begin with some of these positive aspects of clustered sampling, and then discuss its effects on inference in regression analysis.

The economics of clusters in developing countries

In surveys of rural areas in developing countries, clusters are often villages, so that households in a single cluster live near one another, and are interviewed at much the same time during the period that the survey team is in the village. In many countries, these arrangements will produce household data where observations from the same cluster are much more like one another than are observations from different clusters. At the simplest, there may be neighborhood effects, so that local eccentricities are copied by those who live near one another and become more or less uniform within a village. Sample villages are often widely separated geographically, their inhabitants may belong to different ethnic and religious groups, they may have distinct occupational structures as well as different crops and cropping patterns. Where agriculture is important—as it is in most poor countries—there will usually be more homogeneity within villages than between them. This applies not only to the types of crops and livestock, but also to the effects of weather, pests, and natural hazards. If the rains fail for a particular village, everyone engaged in rainfed agriculture will suffer, as will those in occupations that depend on rainfed agriculture. If the harvest is good, prices will be low for everyone in the village, and although the effects will spread out to other

villages through the market, poor transport networks and high transport costs may limit the spread of low prices to other survey villages. Indeed, there is often only one market in each village, so that everyone in the village will be paying the same prices for what they buy, and will be facing the same prices for their wage labor, their produce, and their livestock. This fact alone is likely to induce a good deal of similarity between households within a given sample cluster.

Cluster similarity has both costs and benefits. The cost is that inference is simplest when all the observations in the sample are independent, and that a positive correlation between observations not only makes calculations more complex, but also inflates variance above what it would have been in the independent case. In the extreme case, when all villagers are clones of one another, we need sample only one of them, and if the sample contains more than one person from each village, the effective sample size is the number of *villages* not the number of *villagers*. This argument applies just as much to regressions, and to other types of inference, as it does to the estimation of means.

The benefit of cluster sampling comes from the fact that the clusters are villages, and as such are often economically interesting in their own right. For many purposes it makes sense to examine what happens within each village in a different way from what happens between villages. In addition, cluster sampling gives us multiple observations from the same environment, so that we can sometimes control for unobservables in ways that would not otherwise be possible. One important example is the effects of prices, a topic to which I shall return in Chapter 5. Often, we do not observe prices directly, and since prices in each village will typically be correlated with other village variables such as incomes or agricultural production, it is impossible to estimate the effects of these observables uncontaminated by the effects of the unobservable prices. However, if we are prepared to maintain that prices have additive effects on the variable in which we are interested, differences between households within a village are unaffected by prices, and can be used to make inferences that are robust to the lack of price data. In this way the village structure of samples can be turned to advantage.

Estimating regressions from clustered samples

If the cluster design of the data is ignored, standard formulas for variances of estimated means are too small, a result which applies in essentially the same way to the formulas for the variance-covariance matrices of regression parameters estimated by OLS. At the very least then, we require some procedure for correcting the estimated standard errors of the least squares regression. There is also an efficiency issue; because the error terms in the regressions are correlated across observations, OLS regression is not efficient even within the class of linear estimators and it might be possible to do better with some other linear estimator. (Efficiency is also a potential issue for the sample mean, though I did not discuss it in Section 1.4.)

The simplest example with which to begin is where the cluster design is balanced, so that there are m households in each cluster, and where the explanatory

variables vary only between clusters, and not within them. This will be the case, for example, when we are studying the effects of prices on behavior and there is only one market in each village, or when the explanatory variables are government services, like schools or clinics, where access is the same for everyone in the same village. I follow the discussion in Section 1.4 on the superpopulation approach to clustering and write the regression equation for household i in cluster c [compare (1.64)],

$$y_{ic} = x_c'\beta + \alpha_c + \epsilon_{ic} = x_c'\beta + u_{ic} \tag{2.23}$$

so that the x's are common to all households in the cluster, and the regression error term u_{ic} is the sum of a cluster component α_c and an individual component ϵ_{ic}. Both components have mean 0, and their covariance structure can be derived from the assumption that the α's are uncorrelated across clusters, and the ϵ's both within and across clusters. Hence,

$$\begin{aligned} E(u_{ic}^2) &= \sigma^2 = \sigma_\alpha^2 + \sigma_\epsilon^2 \\ E(u_{ic}u_{jc}) &= \sigma_\alpha^2 = \left(\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}\right)\sigma^2 = \rho\sigma^2, \quad i \neq j \\ E(u_{ic}u_{jc'}) &= 0, \quad c \neq c'. \end{aligned} \tag{2.24}$$

Within the cluster, the errors are equicorrelated with intracluster correlation coefficient ρ, but between clusters, they are uncorrelated.

This case has been analyzed by Kloek (1981), who shows that the special structure implies that the OLS estimator and the generalized least squares estimator are identical, so that OLS is fully efficient. Further, the true variance-covariance matrix of the OLS estimator—as well as of the generalized least squares (GLS) estimator—is given by

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}[1 + (m-1)\rho] \tag{2.25}$$

so that, just as in estimating the variance of the mean, the variance has to be scaled up by the design effect, a factor that varies from 1 to m, depending on the size of ρ.

As before, ignoring the cluster design will lead to standard errors that are too small, and t-values that are too large. There is also a (lesser) problem with estimating the regression standard error σ^2. If N is the sample size—the number of clusters n multiplied by m, the number of observations in each—and k is the number of regressors, the standard formula $(N-k)^{-1}e'e$ is no longer unbiased for σ^2, although it remains consistent provided the cluster size remains fixed as the sample size expands. Kloek shows that an unbiased estimator can be calculated from the design effect $d = 1 + (m-1)\rho$ using the formula

$$\tilde{\sigma}^2 = e'e(N - kd)^{-1}. \tag{2.26}$$

Moulton (1986, 1990) provides a number of examples of potential underestimation of standard errors in this case, some of which are dramatic. For example, in an individual wage equation for the U.S. with only state-level explanatory variables, the design effect is more than 10; here a small but significant intrastate correlation coefficient, 0.028, is combined with very large cluster sizes, nearly 400 observations per state. In this case, ignoring the correction to (2.25) would understate standard errors by a factor of more than three.

That this is likely to be the worst case is shown in papers by Scott and Holt (1982) and Pfefferman and Smith (1985). They show that when the explanatory variables differ within clusters, (2.25)—or when there are unequal numbers of observations in each cluster, (2.25) with the size of the largest cluster replacing m—provides an *upper bound* for the true variance-covariance matrix, and that in most cases, the bound is not tight. They also show that, although the OLS estimator is inefficient when the explanatory variables are not constant within clusters, the efficiency losses are typically small. These results are comforting because they provide a justification for using OLS, and a means of assessing the maximal extent to which the design effects are biasing standard errors. Even so, the biases might still be large enough to worry about, and to warrant correction.

One obvious possibility is to estimate by OLS, use the residuals to estimate $\tilde{\sigma}^2$ from (2.26)—or even from the standard formula—as well as an estimate of the intracluster correlation coefficient

$$\hat{\rho} = \frac{\sum_{c=1}^{n}\sum_{j=1}^{m}\sum_{k\neq j}^{m} e_{ic}e_{jc}}{nm(m-1)\hat{\sigma}^2} \tag{2.27}$$

and then to estimate the variance-covariance matrix using

$$\tilde{V}(\hat{\beta}) = \tilde{\sigma}^2(X'X)^{-1}X'\tilde{\Lambda}X(X'X)^{-1} \tag{2.28}$$

where $\tilde{\Lambda}$ is a block-diagonal matrix with one block for each cluster, and where each block has a unit diagonal and a $\tilde{\rho}$ in each off-diagonal position. An alternative and more robust procedure is to use the OLS residuals from each cluster e_c to form the cluster matrices $\tilde{\Sigma}_c$ according to

$$\tilde{\Sigma}_c = e_c e_c' \tag{2.29}$$

and then to place these matrices on the diagonal of $\tilde{\Lambda}$ in (2.28). This is equivalent to calculating the variance-covariance matrix using

$$\tilde{V}(\hat{\beta}) = (X'X)^{-1}\left(\sum_{c=1}^{n} X_c'e_c e_c'X_c\right)(X'X)^{-1}. \tag{2.30}$$

Provided that the cluster size remains fixed as the sample size becomes large—which is usually the case in practice—(2.30) will provide a consistent estimate of the variance-covariance matrix of the OLS estimator, and will do so even if the error variances differ across clusters, and even in the face of arbitrary correlation

patterns within clusters (see White 1984, pp. 134–42.) In consequence, it can also be applied to the case of heterogeneity within strata discussed in the previous section; the strata are simply thought of as clusters, and the same analysis applied. As we shall see in Section 2.4 below, the same procedures can also be applied to the analysis of panel data where there are repeat observations on the same individuals—the individuals play the role of the village, and successive observations play the role of the villagers (see also Arellano 1987).

Note that the consistency of (2.30) does not suppose (or require) that the $\tilde{\Sigma}_c$ matrices in (2.29) are consistent estimates of the cluster variance-covariance matrices; indeed it is clearly impossible to estimate these matrices consistently from a single realization of the cluster residuals. Nevertheless, (2.30) is consistent for the variance-covariance matrix of the parameters, and will presumably be more accurate in finite samples the more clusters there are, and the smaller is the cluster size relative to the number of clusters. Although (2.30) will typically require special coding or software, it is implemented in STATA as the option "group" in the "huber" or "hreg" command.

Table 2.1 shows the effects of correcting the standard errors of "quality choice" regressions using data on the unit values—expenditures divided by quantities bought—of consumer purchases from the Pakistan Household Income and Expenditure Survey of 1984–85. The substantive issue here is that, because different households buy different qualities of goods, even within categories such as rice and wheat, unit values vary systematically over households, with richer households reporting higher values.

The OLS estimates of the expenditure elasticity of the unit values—what Prais and Houthakker (1955) christened "quality" elasticities—are given in the first column, and we see that there are quality elasticities of 0.13 for wheat and rice, while for the other two goods, which are relatively homogeneous and whose prices are supposedly controlled, the elasticities are small or even negative. Household size elasticities are the opposite sign to total expenditure elasticities, as would be the case (for example) if quality depended on household expenditure per head. Except for sugar, the size elasticities are all smaller in absolute value than the expenditure elasticities, so that, at constant per capita expenditure, unit values rise with household size, an effect that Prais and Houthakker attributed to economies of scale to household size. At the same level of per capita total ex-

Table 2.1. Effects of cluster design on regression *t*-values, rural Pakistan, 1984–85

Good	*Expenditure elasticity*	*t-value*		*Size elasticity*	*t-value*	
		Raw	*Robust*		*Raw*	*Robust*
Wheat	0.128	20.2	18.4	-0.070	-10.5	-9.0
Rice	0.129	12.2	8.7	-0.074	-6.9	-5.4
Sugar	0.005	3.1	1.5	-0.009	-5.2	-3.7
Edible oils	-0.004	-3.0	-1.9	0.002	1.6	1.2

Note: Underlying regression has the logarithm of unit value as the dependent variable, and the logarithms of household total expenditure and of household size as independent variables.
Source: Author's calculations using the Household Income and Expenditure Survey.

penditure, larger households are better-off than smaller households and, in consequence, buy better-quality foods. The robust t-values are smaller than the uncorrected values, although as suggested by the theoretical results, the ratios of the adjusted to unadjusted values are a good deal smaller than the (square roots of the) design effects. Even so, the reductions in the t-values for the estimated quality elasticities for sugar and edible oils are substantial. Without correction, we would almost certainly (mistakenly) reject the hypothesis that the quality elasticities for these two goods are zero; after correction, the t-values come within the range of acceptance.

2.3 Heteroskedasticity and quantile regressions

As we shall see in the next chapter, when we come to look at the distributions over households of the various components of living standards—income, consumption of various goods and their aggregate—it is rare to find variables that are normally distributed, even after standard transformations like taking logarithms or forming ratios. The large numbers of observations in many surveys permit us to look at the distributional assumptions that go into standard regression analysis, and even after transformation it is rarely possible to justify the textbook assumptions that, conditional on the independent variables, the dependent variables are independently, identically, and normally distributed. The previous section discussed how a cluster survey design is likely to lead to a violation of conditional independence. In this section, I turn to the "identically distributed" assumption, and consider the consequences of heteroskedasticity. Just as lack of independence appears to be the rule rather than the exception, so does heteroskedasticity seem to be almost always present in survey data.

The first subsection looks at linear regression models, at the reasons for heteroskedasticity, and at its consequences. I suggest that the computation of quantile regressions is useful, both in its own right, because quantile regression estimates will often have better properties than OLS, as a way of assessing the heteroskedasticity in the conditional distribution of the variable of interest, and as a stepping stone to the nonparametric methods discussed in the next two chapters. As was the case for clustering, a consequence of heteroskedasticity in regression analysis is to invalidate the usual formulas for the calculation of standard errors, and as with clustering, there exists a straightforward correction procedure.

Matters are much less simple when we move from regressions to models with limited dependent variables. In regression analysis, the estimation of scale parameters can be separated from the estimation of location parameters, but the separation breaks down in probits, logits, Tobits, and in sample selectivity models. I illustrate some of the difficulties using the Tobit model, and provide a simple but realistic example of censoring at zero where the application of maximum-likelihood Tobit techniques—something that is nowadays quite routine in the development literature—can lead to estimates that are no better than OLS. There are currently no straightforward solutions to these difficulties, but I review some of the options and make some suggestions for practice.

Heteroskedasticity in regression analysis

It is a fact that regression functions estimated from survey data are typically not homoskedastic. Why this should be is of secondary importance; indeed it is just as reasonable to ask why it should be supposed that conditional expectations should be homoskedastic. Nevertheless, we have already seen in Section 2.1 above that even when individual behavior generates homoskedastic regression functions within strata or villages, but there is heterogeneity between villages, there will be heteroskedasticity in the overall regression function. Similar results apply to heterogeneity at the individual level. If the response coefficients β differ by household, and we treat them as random, we may write

$$E(y_i|x_i,\beta_i) = \beta_i'x_i; \quad V(y_i|x_i,\beta_i) = \sigma^2. \tag{2.31}$$

Suppose that the β_i have mean β and variance-covariance matrix Ω, then (2.31) generates the heteroskedastic regression model

$$E(y_i|x_i) = \beta'x_i; \quad V(y_i|x_i) = \sigma^2 + x_i'\Omega x_i. \tag{2.32}$$

Models like (2.32) motivate the standard test procedures for heteroskedasticity such as the Breusch-Pagan (1979) test, or White's (1980) information matrix test (see also Chesher 1984 for the link with individual heterogeneity.) The Breusch-Pagan test is particularly straightforward to implement. The OLS residuals from the regression with suspected heteroskedasticity are first normalized by division by the estimated standard error of the equation. Their squares are then regressed on the variables thought to be generating the heteroskedasticity—if (2.32) is correct, these should include the original x-variables, their squares, and cross-products—and half the explained sum of squares tested against the χ^2 distribution with degrees of freedom equal to the number of variables in this supplementary regression.

In the presence of heteroskedasticity, OLS is inefficient and the usual formulas for standard errors are incorrect. In cases where efficiency is not a prime concern, we may nevertheless want to use the OLS estimates, but to correct the standard errors. This can be done exactly as in (2.30) above, a formula that is robust to the presence of *both* heteroskedasticity and cluster effects. If there are no clusters, (2.30) can be applied by treating each household as its own cluster so that there are no cross-effects within clusters and the formula can be written

$$\tilde{V}(\hat{\beta}) = (X'X)^{-1}(\sum_i e_i^2 x_i x_i')(X'X)^{-1} \tag{2.33}$$

where x_i is the column vector of explanatory variables for household i and e_i^2 is the squared residual from the OLS regression. This formula, which comes originally from Eicker (1967) and Huber (1967), was introduced into econometrics by White (1980). Its performance in finite samples can be improved by a number of possible corrections; the simplest requires that e_i^2 in (2.33) be multiplied by

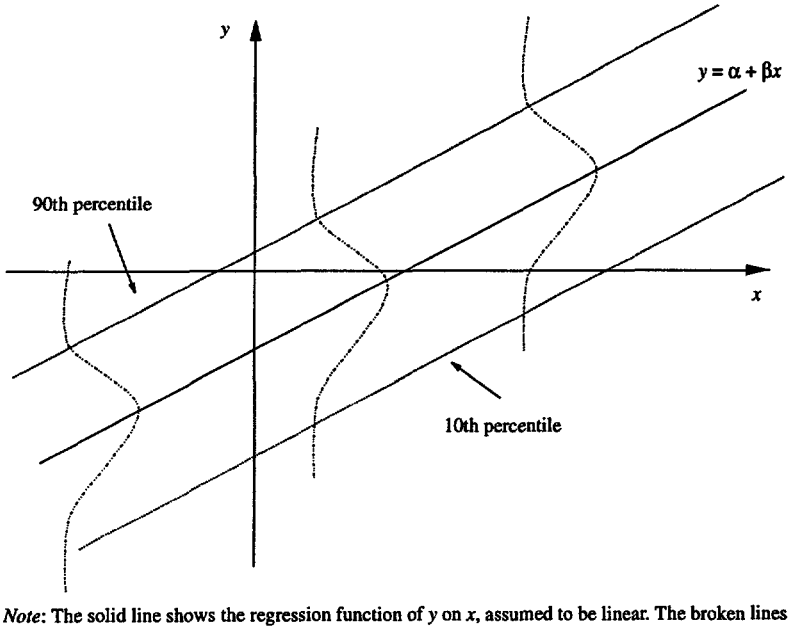
$(n-k)^{-1}n$, where k is the number of regressors and n the sample size, see Davidson and MacKinnon (1993, 552–56.) In practice, the heteroskedasticity correction to the variance-covariance matrix (2.33) is usually quantitatively less important than the correction for intracluster correlations, (2.30).

Quantile regressions

The presence of heteroskedasticity can be conveniently analyzed and displayed by estimating *quantile regressions* following the original proposals by Koenker and Bassett (1978, 1982). To see how these work, it is convenient to start from the standard *homoskedastic* regression model.

Figure 2.1 illustrates quantiles in the (standard) case where heteroskedasticity is absent. The regression line $\alpha + \beta x$ is the expectation of y conditional on x, and the three "humped" curves schematically illustrate the conditional densities of the errors given x; in principle, these densities should rise perpendicularly from the page. For each value of x, consider a process whereby we mark the percentiles of the conditional distribution, and then connect up the same percentiles for different values of x. If the distribution of errors is symmetrical, as shown in Figure 2.1, the conditional mean, or regression function, will be at the 50th percentile or median, so that joining up the conditional medians simply reproduces the regression. When the distribution of errors is also homoskedastic, the percentiles will always be at the same distance from the median, no matter what the value of x. Figure 2.1 shows the lines formed by joining the points corresponding to the 10th and 90th

Figure 2.1. Schematic figure of a homoskedastic linear regression function



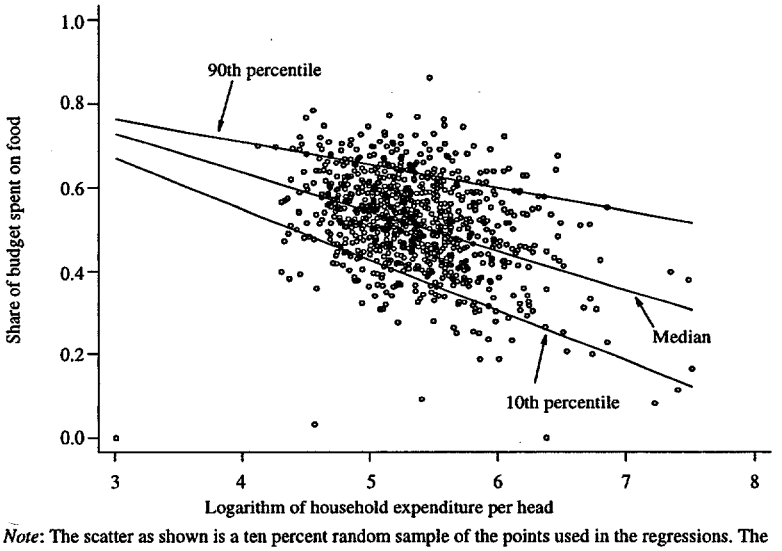
Note: The solid line shows the regression function of y on x, assumed to be linear. The broken lines show the 10th and 90th percentiles of the distribution of y conditional on x.

percentiles of the conditional distributions. Because the regression is homoskedastic, these are straight lines that are parallel to, and equidistant from, the regression line.

When regressions are heteroskedastic, or when the errors are asymmetric, marking and joining up percentiles will give quite different results. If the residuals are symmetric but heteroskedastic, the distance of each percentile from the regression line will be different at different values of x. Joining up the percentiles for different values of x will not necessarily lead to straight lines or to any other simple curve. However, we can still *fit* straight lines to the percentiles, and it is this that is accomplished by quantile regression. If the heteroskedasticity is linked to the value of x, with the distribution of residuals becoming more or less dispersed as x becomes larger, then the quantile regressions for percentiles other than the median will no longer be parallel to the regression line, but will diverge from it (or converge to it) for larger values of x.

Figure 2.2 illustrates using a food Engel curve for the rural data from the 1984–85 Household Income and Expenditure Survey of Pakistan. Previous experience has shown that the budget share devoted to food can often be well approximated as a linear function of the logarithm of household expenditure per capita, as first proposed by Working (1943). The points in the figure are a 10 percent random sample of the 9,119 households in the survey whose logarithm of per capita expenditure lies between 3 and 8; a small number of households at the extremes of the distribution are thereby excluded from the figure, but not from the calcula-

Figure 2.2. Scatter diagram and quantile regressions for food share and total expenditure, Pakistan, 1984–85



Note: The scatter as shown is a ten percent random sample of the points used in the regressions. The regression lines shown were obtained using the "qreg" command in STATA and correspond to the 10th, 50th, and 90th percentiles.

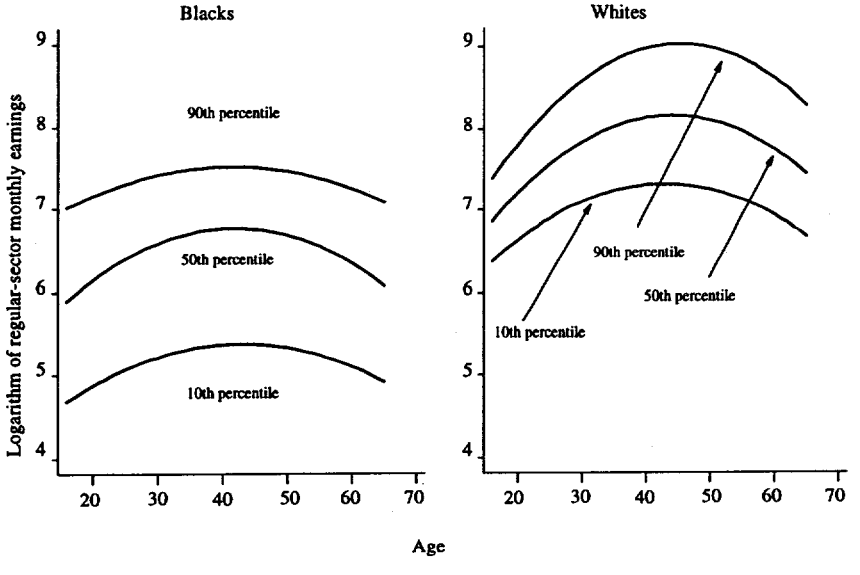
Source: Author's calculations based on Household Income and Expenditure Survey.

tions. The three lines in the figure are the quantile regressions corresponding to the 10th, 50th, and 90th percentiles of the distribution of the food share conditional on the logarithm of household expenditure per head; these were calculated using all 9,119 households. The procedures for estimating these regressions, calculated using the "qreg" command in STATA, are discussed in the technical note that follows, but the principle should be clear from the foregoing discussion.

The slopes of the three lines differ; the median regression (50th percentile) has a slope of –0.094 (the OLS slope is –0.091), while the lower line has slope –0.121, and the upper –0.054. These differences and the widening spread between the lines as we move to the right show the increase in the conditional variance of the regression among better-off households; the 10th and 90th percentiles of the conditional distribution are much further apart among richer than poorer households. Those with more to spend in total devote a good deal less of their budgets to food, but there is also more dispersion of tastes among them.

Quantile regressions are not only useful for discovering heteroskedasticity. By calculating regressions for different quantiles, it is possible to explore the shape of the conditional distribution, something that is often of interest in its own right, even when heteroskedasticity is not the immediate cause for concern. A very simple example is shown in Figure 2.3, which illustrates age profiles of earnings for black and white workers from the 1993 South African Living Standards Survey. Earnings are monthly earnings in the "regular" sector, and the graphs use

Figure 2.3. Quantile regressions of the logarithm of earnings on age by race, South Africa, 1993
(log of rand per month)



Source: Author's calculations using the South African Living Standards Survey, 1993.

only data for those workers who report such earnings. The two panels show the quantile regressions of the logarithm of earnings on age and age squared for the 10th, 50th, and 90th percentiles for Black and White workers separately. The use of a quadratic in age restricts the shapes of the profiles, but allows them to differ by race and by percentile, as in fact they do. The curves show not only the vast differences in earnings between Blacks and Whites—a difference in logarithms of 1 is a ratio difference of more than 2.7—but also that the shapes of the age profiles are different. Those whose earnings are at the top within their age group are the more highly-educated workers in more highly-skilled jobs, and because the human capital required for these jobs takes time to accumulate, the profile at the 90th percentile for whites has a stronger hump-shape than do the profiles for the 50th and 10th percentiles. There is no corresponding phenomenon for Blacks, presumably because, in South Africa in 1993, even the most able Blacks are restricted in their access to education and to high-skill jobs. These graphs and the underlying regressions do not tell us anything about the causal processes that generate the differences, but they present the data in an interesting way that can be suggestive of ideas for a deeper investigation (see Mwabu and Schultz 1995 for more formal analysis of earnings in South Africa, Mwabu and Schultz 1996 for a use of quantile regression in the same context, and Buchinsky 1994 for the use of quantile regressions to describe the wage structure in the U.S.)

There are also arguments for preferring the parameters of the median regression to those from the OLS regression. Even given the Gauss-Markov assumptions of homoskedasticity and independence, least squares is only efficient within the (restrictive) class of linear, unbiased estimators, although if the conditional distribution is normal, OLS will be minimum variance among the broader class of all unbiased estimators. When the distribution of residuals is not normal, there will usually exist nonlinear (and/or biased) estimators that are more efficient than OLS, and quantile regressions will sometimes be among them. In particular, the median regression is more resistant to outliers than is OLS, a major advantage in working with large-scale survey data.

Technical note: calculating quantile regressions

In the past, the applicability of quantile regression techniques has been limited, not because they are inherently unattractive, but by computational difficulties. These have now been resolved. Just as in calculating the median itself, median regression can be defined by minimizing the absolute sum of the errors rather than, as in least squares, by minimizing the sum of their squares. It is thus also known as the LAD estimator, for Least Absolute Deviations. Hence, the median regression coefficients can be obtained by minimizing ϕ given by

$$\phi = \sum_{i=1}^{n} |y_i - x_i'\beta| = \sum_{i=1}^{n} (y_i - x_i'\beta)\,\mathrm{sgn}(y_i - x_i'\beta) \tag{2.34}$$

where sgn(a) is the sign of a, 1 if a is positive, and –1 if a is negative or zero. (I have reverted to the standard use of n for the sample size, since there is no longer

a need to separate the number of clusters from the total number of observations.) The intuition for why (2.34) works comes from thinking about the first-order condition that is satisfied by the parameters that minimize (2.34), which is, for $j = 1, \ldots, k$,

$$\sum_{i=1}^{n} x_{ij}\,\mathrm{sgn}(y_i - x_i'\beta) = 0. \tag{2.35}$$

Note first that if there is only a constant in the regression, (2.35) says that the constant should be chosen so that there are an equal number of points on either side of it, which defines the median. Second, note the similarity between (2.35) and the OLS first-order conditions, which are identical except for the "sgn" function; in median regression, it is only the sign of each residual that counts, whereas in OLS it is its magnitude.

Quantile regressions other than median can be defined by minimizing, not (2.34), but

$$\begin{aligned}\phi_q &= -(1-q)\sum_{y \le x'\beta}(y_i - x_i'\beta) + q\sum_{y > x'\beta}(y_i - x_i'\beta) \\ &= \sum_{i=1}^{n}[q - 1(y_i \le x_i'\beta)](y_i - x_i'\beta)\end{aligned} \tag{2.36}$$

where $0<q<1$ is the quantile of interest, and the value of the function $1(z)$ signals the truth (1) or otherwise (0) of the statement z. The minimization condition corresponding to (2.35) is now

$$\sum_{i} x_{ij}[q - 1(y_i \le x_i'\beta)] = 0 \tag{2.37}$$

which is clearly equivalent to (2.35) when q is a half. Once again, note that if the regression contains only a constant term, the constant is set so that $100q$ percent of the sample points are below it, and $100(1-q)$ percent above.

The computation of quantile estimators is eased by the recognition that the minimization of (2.36) can be accomplished by linear programming, so that even for large data sets, the calculations are not burdensome. The same cannot be said for the estimation of the variance-covariance matrix of the parameter estimates. When the residuals are homoskedastic, there is an asymptotic formula provided by Koenker and Bassett (1982) that gives the variance-covariance matrix of the parameters as the usual $(X'X)^{-1}$ matrix scaled by a quantity that depends (inversely) on the density of the errors at the quantiles of interest. Estimation of this density is not straightforward, but more seriously, the formula appears to give very poor results—typically gross underestimation of standard errors—in the presence of heteroskedasticity, which is often the reason for using quantile regression in the first place!

It is therefore important to use some other method for estimating standard errors, such as the bootstrap, a version of which is implemented in the "bsqreg" command in STATA, whose manual, Stata Corporation (1993, Vol. 3, 96–106) provides a useful description of quantile regressions in general. (Note that the STATA version does not allow for clustering but it is straightforward, if time-con-

suming, to bootstrap quantile regressions using the clustered bootstrap illustrated in Example 1.3 in the Code Appendix.)

Heteroskedasticity and limited dependent variable models

In regression analysis, the presence of heteroskedasticity and nonnormality is problematic because of potential efficiency losses, and because of the need to correct the usual formulas for standard errors. However, regression analysis is somewhat of a special case because the estimation of parameters of location—the conditional mean or conditional median—is independent of the estimation of scale—the dispersion around the conditional location. In limited dependent variable models, scale and location are intimately bound together, and as a result, misspecification of scale can lead to inconsistency in estimates of location.

Probit and logit models provide perhaps the clearest examples of the difficulties that arise. There is a dependent variable y_i which is either 1 or 0 according to whether or not an unobserved or latent variable y_i^* is positive or nonpositive. The latent variable is defined by analogy to a regression model,

$$y_i^* = f(x_i) + u_i, \quad E(u_i) = 0, \quad E(u_i^2) = \sigma^2[\,g(z_i)]^2 \tag{2.38}$$

where x and z are vectors of variables controlling the "regression" and the "heteroskedasticity" respectively, and $f(.)$ and $g(.)$ are functions, the former usually assumed to be known, the latter unknown. Suppose that $F(.)$ is the cumulative distribution function (CDF) of the standardized residual $u_i/\sigma g(z_i)$, and that $F(.)$ is symmetric around 0 so that $F(a) = 1 - F(-a)$, then

$$p_i = Prob(y_i = 1) = F[\,f(x_i)/\sigma g(z_i)]. \tag{2.39}$$

If we know the function $F(.)$, which is in itself assuming a great deal, then given data on y, x, and z, the model gives no more information on which to base estimation than is contained in the probabilities (2.39). But by inspection of (2.39) it is immediately clear that it is not possible to separate the "heteroskedasticity" function $g(z)$ from the "regression" function $f(x)$. For example, suppose that $f(x)$ has the standard linear specification $x'\beta$, that the elements of z are the same as those of x, and that it so happens that $g(z) = g(x) = x'\beta/x'\gamma$. Then the application of maximum-likelihood estimation (MLE) will yield estimates that are consistent, not for β, but for γ! The latent-variable or regression approach to dichotomous models can be misleading if we treat it too seriously; we observe 1's or 0's, and we can use them to model the probabilities, but that is all.

The point of the previous paragraph is so obvious and so well understood that it is hardly of practical importance; the confounding of heteroskedasticity and "structure" is unlikely to lead to problems of interpretation. It is standard procedure in estimating dichotomous models to set the variance in (2.38) to be unity, and since it is clear that all that can be estimated is the effects of the covariates on the probability, it will usually be of no importance whether the mechanism works

through the mean or the variance of the latent "regression" (2.38). While it is correct to say that probit or logit is inconsistent under heteroskedasticity, the inconsistency would only be a problem if the parameters of the function f were the parameters of interest. These parameters are identified only by the homoskedasticity assumption, so that the inconsistency result is both trivial and obvious. (It is perhaps worth noting that STATA has "hlogit" and "hprobit" commands for logit and probit that match the "hreg" command for regression. But these should not be used to correct standard errors in logit and probit; rather they should be used to correct standard errors for clustering, so that the analogy is with (2.30), not (2.33).)

Related but more serious difficulties occur with heteroskedasticity when analyzing censored regression models, truncated regressions, or regressions with selectivity, where the inconsistencies are a good deal more troublesome. I illustrate using the censored regression model, or Tobit—after Tobin's (1958) probit—because the model is directly relevant to the analysis in later chapters, and because the technique is widely used in the development literature. Consider in particular the demand for a good, which can be purchased only in positive quantities. If there were no such restriction, we might postulate a linear regression of the form

$$y_i^* = x_i'\beta + u_i. \tag{2.40}$$

When y_i^* is positive, everything is as usual and actual demand y_i is equal to y_i^*. But negative values of y_i^* are "censored" and replaced by zero, the minimum allowed. The model for the observed y_i can thus be written as

$$y_i = \max(0, x_i'\beta + u_i). \tag{2.41}$$

I note in passing that the model can be derived more elegantly as in Heckman (1974), who considers a labor supply example, and shows that (2.41) is consistent with choice theory when hours worked cannot be negative.

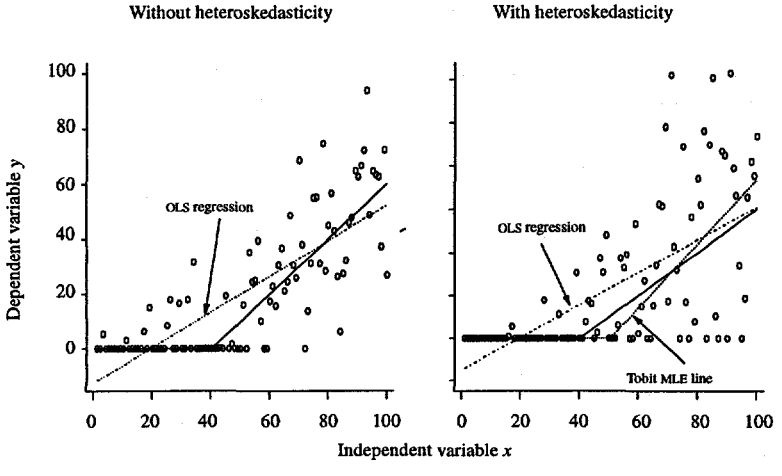
The left-hand panel of Figure 2.4 shows a simulation of an example of a standard Tobit model. The latent variable is given by $x_i - 40 + u_i$ with the x's taking the 100 values from 1 to 100, and the u's randomly and independently drawn from a normal distribution with mean zero and standard deviation 20. The small circles on the graph show the resulting scatter of y_i against x_i. Because of the censoring, which is more severe for low values of the explanatory variable, the OLS regression line has a slope less than one; in 100 replications the OLS estimator had a mean of 0.637 with a standard deviation of 0.055, so that the bias shown in the figure for one particular realization is typical of this situation. A better method is to follow Tobin, and maximize the log-likelihood function

$$\begin{aligned} \ln L = & -\frac{n_+}{2}(\ln\sigma + \ln 2\pi) - \frac{1}{2\sigma^2}\sum_{i_+}(y_i - x_i'\beta)^2 \\ & + \sum_{i_0}\ln\left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \end{aligned} \tag{2.42}$$

where n_+ is the number of strictly positive observations, i_+ and i_0 indicate that the respective sums are taken over positive and zero observations, respectively, and Φ is the c.d.f. of the standard normal distribution. The first two terms on the right-hand side of (2.42) are exactly those that would appear in the likelihood function of a standard normal regression, and would be the only terms to appear in the absence of censoring. The final term comes from the observations that are censored to zero; for each such observation we do not observe the exact value of the latent variable, only that it is zero or less, so that the contribution to the log likelihood is the logarithm of the probability of that event. Estimates of β and σ are obtained by maximizing (2.42), a nonlinear problem whose solution is guaranteed by the fact the log-likelihood function is convex in the parameters, and so has a unique maximum. This maximum-likelihood technique works well for the left-hand panel of the figure; in the 100 replications, the Tobit estimates of the slope averaged 1.009 with a standard deviation of 0.100. In this case, where the normality assumption is correct, and the disturbances homoskedastic, maximum likelihood overcomes the inconsistency of OLS.

That all will not be as well in the presence of heteroskedasticity can be seen from the likelihood function (2.42) where the last term, which is the contribution to the likelihood of the censored observations, contains both the scale and location parameters. The standard noncensored likelihood function, which is (2.42) without the last term, has the property that the derivatives of the log-likelihood function with respect to the β's are independent of σ, at least in expectation, and *vice versa*, something that is not true for (2.42). This gives a precise meaning to the notion that scale and location are independent in the regression model, but

Figure 2.4. Tobit models with and without heteroskedasticity



Note: See text for model definition and estimation procedures.
Source: Author's calculations.

dependent in these models with limited dependent variables. As a result of the dependence, misspecification of scale will cause the β's that maximize (2.42) to be inconsistent for the true parameters, a result first noted by Hurd (1979), Nelson (1981), and Arabmazar and Schmidt (1981).

The right-hand side of Figure 2.4 gives an illustration of the kind of problems that can occur with heteroskedasticity. Instead of being homoskedastic as in the left-hand panel, the u_i are drawn from independent normal distributions with zero means and standard deviations σ_i given by

$$\sigma_i = 20\left(1 + 0.2\sqrt{\max(0, x_i - 40)}\right). \tag{2.43}$$

According to this specification there is homoskedasticity to the left of the cutoff point (40), but heteroskedasicity to its right, and the conditional variance grows with the mean of the dependent variable beyond the cutoff. Although (2.43) does not pretend to be based on any actual data, it mimics reasonable models of behavior. Richer households have more scope for idiosyncracies of behavior than do the poor, and as we see in the right-hand panel, we now get zero observations among the rich as well as the poor, something that cannot occur in the homoskedastic model. This is what happens in practice; if we look at the demand for tobacco, alcohol, fee-paying schools or clinics, there are more nonconsumers among the poor, but there are also many better-off households who choose not to purchase. Not purchasing is partly a matter of income, and partly a matter of taste.

The figure shows three lines. The dots-and-dashes line to the left is the OLS regression which is still biased downward; although the heteroskedasticity has generated more very high y's at high levels of x, the censoring at low values of x keeps the OLS slope down. In the replications the OLS slope averaged 0.699 with a standard deviation of 0.100; there is more variability than before, but the bias is much the same. The second, middle (solid) line is the kinked line $\max(0, x - 40)$ which is (2.41) when all the u_i are zero. (Note that this line is not the regression function, which is *defined* as the expectation of y conditional on x.) The third line, on the right of the picture, comes from maximizing the likelihood (2.42) under the (false) assumption that the u's are homoskedastic. Because the Tobit procedure allows it to deal with censoring at low values of x, but provides it with no explanation for censoring at high values of x, the line is biased upward in order to pass through the center of the distribution on the right of the picture. The average MLE (Tobit) estimate of the slope in the replications was 1.345 with a standard error of 0.175, so that in the face of the heteroskedasticity, the Tobit procedure yields estimates that are as biased up as OLS is biased down. It is certainly possible to construct examples where the Tobit estimators are better than least squares, even in the presence of heteroskedasticity. But there is nothing odd about the current example; heteroskedasticity will usually be present in practical applications, and there is no general guarantee that the attempt to deal with censoring by replacing OLS with the Tobit MLE will give estimates that reduce the bias. This is *not* a defense of OLS, but a warning against the supposition that Tobit guarantees any improvement.

In practice, the situation is worse than in this example. Even when there is no heteroskedasticity, the consistency of the Tobit estimates requires that the distribution of errors be normal, and biases can occur when it is not (see Goldberger 1983 and Arabmazar and Schmidt 1982). And since the distribution of the u's is almost always unknown, it is unclear how one might respecify the likelihood function in order to do better. Even so, censored data occur frequently in practice, and we need some method for estimating sensible models. There are two very different approaches; the first is to look for estimation strategies that are robust against heteroskedasticity of the u's in (2.41) and that require only weak assumptions about their distribution, while the second is more radical, and essentially abandons the Tobit approach altogether. I begin with the former.

*Robust estimation of censored regression models

There are a number of different estimators that implement the first approach, yielding nonparametric Tobit estimators—nonparametric referring to the distribution of the u's, not to the functional form of the latent variable which remains linear. None of these has yet passed into standard usage, and I review only one, Powell's (1984) censored LAD estimator. It is relatively easily implemented and appears to work in practice. (An alternative is Powell's (1986) symmetrically trimmed least squares estimator.)

One of the most useful properties of quantiles is that they are preserved under monotone transformations; for example, if we have a set of positive observations, and we take logarithms, the median of the logarithms will be the logarithm of the median of the untransformed data. Since $\max(0, z)$ is monotone nondecreasing in z, we can take medians of (2.41) conditional on x_i to get

$$q_{50}(y_i|x_i) = \max[0, q_{50}(x_i'\beta + u_i|x_i)] = \max(0, x_i'\beta) \tag{2.44}$$

where $q_{50}(.|x)$ denotes the median of the distribution conditional on x and the median of u_i is assumed to be 0. But as we have already seen, LAD regression estimates the conditional median regression, so that β can be consistently estimated by the parameter vector that minimizes

$$\sum_{i=1}^{n} |y_i - \max(0, x_i'\beta)| \tag{2.45}$$

which is what Powell suggests. The consistency of this estimator does not require knowledge of the distribution of the u's, nor is it assumed that the distribution is homoskedastic, only that it has median 0.

Although Powell's estimator is not available in standard software, it can be calculated from repeated application of median regression following a suggestion of Buchinsky (1994, p. 412). The first regression is run on all the observations, and the predicted values $x_i'\beta$ calculated; these are used to discard sample observations where the predicted values are negative. The median regression is then repeated on the truncated sample, the parameter estimates used to recalculate $x_i'\beta$

for the whole sample, the negative values discarded, and so on until convergence. In (occasional) cases where the procedure does not converge, but cycles through a finite set of parameters, the parameters with the highest value of the criterion function should be chosen. Standard errors can be taken from the final iteration though, as before, bootstrapped estimates should be used.

Such a procedure is easily coded in STATA, and was applied to the heteroskedastic example given above and shown in Figure 2.4 (see Example 2.1 in the Code Appendix). To simplify the coding, the procedure was terminated after 10 median regressions, so that to the extent that convergence had not been attained, the results will be biased against Powell's estimator. On average, the method does well, and the mean of the censored LAD estimator over the 100 replications was 0.946. However, there is a price to be paid in variance, and the standard deviation of 0.305 is three times that of the OLS estimator and more than one and a half times larger than that of the Tobit. As a result, and although both Tobit and OLS are inconsistent, in only 55 out of 100 of the replications is the censored LAD closer to the truth than both OLS and Tobit. Of course, the bias to variance tradeoff turns in favor of Powell's estimator as the sample size becomes larger. With 1,000 observations instead of 100, and with the new x values again equally spaced but 10 times closer, the censored LAD estimator is closer to the truth than either OLS or Tobit in 96 percent of the cases. Since most household surveys will have sample sizes at least this large, Powell's estimator is worth serious consideration. At the very least, comparing it with the Tobit estimates will provide a useful guide to failure of homoskedasticity or normality (see Newey 1987 for an exercise of this kind).

Even so, the censored LAD estimator is designed for the censored regression model, and does not apply to other cases, such as truncation, where the observations that would have been negative do not appear in the sample instead of being replaced by zeros, nor to more general models of sample selectivity. In these, the censoring or truncation of one variable is determined by the behavior of a second latent variable regression, so that

$$\begin{aligned} y_{1i}^{*} &= x_i'\beta + u_{1i} \\ y_{2i}^{*} &= z_i'\gamma + u_{2i} \end{aligned} \tag{2.46}$$

where u_1 and u_2 are typically allowed to be correlated, y_{2i} is observed as a dichotomous variable indicating whether or not y_{2i}^{*} is positive, and y_{1i} is observed as y_{1i}^{*} when y_{2i} is 1, and is zero otherwise. Equations (2.46) are a generalization of Tobit, whereby the censoring is controlled by variables that are different from the variables that control the magnitude of the variable of interest. If the two sets of u's are assumed to be jointly normally distributed, (2.46) can be estimated by maximum likelihood, or by Heckman's (1976) two-step estimator—the "Heckit" procedure (see the next section for further discussion). As with Tobit, which is a special case, these methods do not yield consistent estimates in the presence of heteroskedasticity or nonnormality, and as with Tobit, the provision of nonparametric estimators is a lively topic of current research in econometrics. I shall return to these topics in more detail in the next section.

Radical approaches to censored regressions

Serious attention must also be given to a second, more radical, approach that questions the usefulness of these models in general. There are conceptual issues as well as practical ones. In the first place, these models are typically presented as elaborations of linear regression, in which a standard regression equation is extended to deal with censoring, truncation, selectivity, or whatever is the issue at hand. However, in so doing they make a major break from the standard situation presented in the introduction where the regression *function*, the expectation of the dependent variable conditional on the covariates, coincides with the deterministic part of the regression *model*. In the Tobit and its generalizations, the regression functions are no longer simple linear functions of the x's, but are more complex expressions that involve the distribution of the u's. For example, in the censored regression model (2.41), the regression function is given by

$$\begin{aligned} E(y_i \mid x_i) &= x_i'\beta + E(u \mid x_i'\beta + u \geq 0) \\ &= x_i'\beta + [1 - F(-x_i'\beta)]^{-1} \int_{-x_i'\beta}^{\infty} u\, dF(u) \end{aligned} \tag{2.47}$$

where $F(u)$ is the CDF of the u's. Absent knowledge of F, this regression function does not even identify the β's—see Powell (1989)—but more fundamentally, we should ask how it has come about that we have to deal with such an awkward, difficult, and nonrobust object.

Regressions are routinely assumed to be linear, not because linearity is thought to be exactly true, but because it is convenient. A linear model is often a sensible first approximation, and linear regressions are easy to estimate, to replicate, and to interpret. But once we move into models with censoring or selection, it is much less convenient to start with linearity, since it buys us no simplification. It is therefore worth considering alternative possibilities, such as starting by specifying a suitable functional form for the regression function itself, rather than for the part of the model that would have been the regression function had we been dealing with a linear model. Linearity will often not be appropriate for the regression function, but there are many other possibilities, and as we shall see in the next chapter, it is often possible to finesse the functional form issue altogether. Such an approach goes beyond partially nonparametric treatments that allow arbitrary distributional assumptions for the disturbances while maintaining linearity for the functional form of the model itself, and recognizes that the functional form is as much an unknown as is the error distribution. It also explicitly abandons the attempt to estimate the structure of selectivity or censoring, and focusses on features of the data—such as regression functions—that are clearly and uncontroversially observable. There will sometimes be a cost to abandoning the structure, but there are many policy problems for which the structure is irrelevant, and which can be addressed through the regression function.

A good example is the effect of a change in tax rates on tax revenue. A government is considering a reduction in the subsidy on wheat (say), and needs to

know the extent to which demand will be reduced at the higher price. The quantity of interest is the effect of price on average demand. Suppose that we have survey data on wheat purchases, together with regional or temporal price variation as well as other relevant explanatory variables. Some households buy positive quantities of wheat, and some buy none, a situation that would seem to call for a Tobit Estimation of the model yields an estimate of the response of quantity to price for those who buy wheat. But the policymaker is interested not only in this effect, but also in the loss of demand from those who previously purchased, but who will drop out of the market at the higher price. These effects will have to be modeled separately and added into the calculation. But this is an artificial and unnecessarily elaborate approach to the problem. The policy question is about the effect of price on average demand, averaged over consumers and nonconsumers alike. But this is exactly what we would estimate if we simply regressed quantity on price, with zeros and nonzeros included in the regression. In this case, not only is the regression function more convenient to deal with from an econometric perspective, it is also what we need to know for policy.

2.4 Structure and regression in nonexperimental data

The regression model is the standard workhorse for the analysis of survey data, and the parameters estimated by regression analysis frequently provide useful summaries of the data. Even so, they do not always give us what we want. This is particularly so when the survey data are a poor substitute for unobtainable experimental data. For example, if we want to know the effect of constructing health clinics, or of expanding schools, or what will happen if a minimum wage or health coverage is mandated, we should ideally like to conduct an experiment, in which some randomly chosen group is given the "treatment," and the results compared with a randomly selected group of controls from whom the treatment is withheld. The randomization guarantees that there are no differences—observable or unobservable—between the two groups. In consequence, if there is a significant difference in outcomes, it can only be the effect of the treatment. Although the role of policy experiments has been greatly expanded in recent years (see Grossman 1994 and Newman, Rawlings, and Gertler 1994), there are many cases where experiments are difficult or even impossible, sometimes because of the cost, and sometimes because of the moral and political implications. Instead, we have to use nonexperimental survey data to look at the differences in behavior between different people, and to try to relate the degree of exposure to the treatment to variation in the outcomes in which we are interested. Only under ideal conditions will regression analysis give the right answers. In this section, I explore the various difficulties; in the next two sections, I look at two of the most important of the econometric solutions, panel data and the technique of instrumental variables.

The starting point for a nonexperimental study is often a regression model, in which the outcome variable y is related to a set of explanatory variables x. At least one of the x-variables is the treatment variable, while others are "control" vari-

ables, included so as to allow for differences in outcomes that are not caused by the treatment and to allow the treatment effect to be isolated. These variables play the same role as the control group in an experiment. The error term in the regression captures omitted controls, as well as measurement error in the outcome y, and is assumed to satisfy (2.4), that its expectation conditional on the x's is 0. In this setup, the expectation of y conditional on x is $\beta'x$, and the effects of the treatment and controls can be recovered by estimating β. The most common problem with this procedure is the failure—or at least implausibility—of the assumption that the conditional mean of the error term is zero. If a relevant variable is omitted, perhaps because it is unobservable or because data are unavailable, and if that variable is correlated with any of the included x's, the error will not be orthogonal to the x's, and the conditional expectation of y will not be $\beta'x$. The regression function no longer coincides with the structure that we are trying to recover, and estimation of the regression function will not yield the parameters of interest. The failure of the structure and the regression function to coincide happens for many different reasons, some more obvious than others. In this section, I consider a number of cases that are important in the analysis of household survey data from developing countries.

Simultaneity, feedback, and unobserved heterogeneity

Simultaneity is a classic argument for a correlation between error terms and explanatory variables. If we supplement the regression model (2.3) with another equation or equations by which some of the explanatory variables are determined by factors that include y, then the error term in (2.3) will be correlated with one or more of the x's and OLS estimates will be biased and inconsistent. The classic textbook examples of simultaneity, the interdependence of supply or demand and price, and the feedbacks through national income from expenditures to income, are usually thought not to be important for microeconomic data, where the purchases of individuals are too small to affect price or to influence their own incomes through macroeconomic feedbacks. As we shall see, this is not necessarily the case, especially when there are local village markets. Other forms of simultaneity are also important in micro data, although the underlying causes often have more to do with omitted or unobservable variables than with feedbacks through time. Four examples illustrate.

Example 1. Price and quantities in local markets

In the analysis of demand using microeconomic data, it is usually assumed that individual purchases are too small to affect price, so that the simultaneity between price and aggregate demand can be ignored in the analysis of the microeconomic data. Examples where this is not the case have been provided by Kennan (1989), and local markets in developing countries provide a related case. Suppose that the demand function for each individual in each village contains an unobservable village-level component, and that, because of poor transportation and lack of an

integrated market, supply and demand are equilibriated at the village level. Although the village-level component in individual demands may contribute little to the total variance of demand, the other components will cancel out over the village as a whole, so that the variation in price across villages is correlated with village-level taste for the good. Villages that have a relatively high taste for wheat will tend to have a relatively high price for wheat, and the correlation can be important even when regressions are run using data from individuals or households.

To illustrate, write the demand function in the form

$$y_{ic} = \alpha_0 + \beta x_{ic} - \gamma p_c + u_{ic} = \alpha_0 + \beta x_{ic} - \gamma p_c + \alpha_c + \epsilon_{ic} \tag{2.48}$$

where y_{ic} is demand by household i in cluster c, x_{ic} is income or some other individual variable, p_c is the common village price, and u_{ic} is the error term. As in previous modeling of clusters, I assume that u_{ic} is the sum of a village term α_c and an idiosyncratic term ϵ_{ic}, both of which are mean-zero random variables. Suppose that aggregate supply for the village is z_c per household, which comes from a weather-affected harvest but is unresponsive to price (or to income). Market clearing implies that

$$z_c = \bar{y}_c = \alpha_0 + \beta \bar{x}_c - \gamma p_c + \alpha_c \tag{2.49}$$

which determines price in terms of the village taste effect, supply, and average village income. Because markets have to clear at the village level, the price is higher in villages with a higher taste for the commodity. In consequence, the price on the right-hand side of (2.48) is correlated with the α_c component of the error term, and OLS estimates will be inconsistent. The inconsistency arises even if the village contains many households, each of which has a negligible effect on price.

The bias can be large in this case. To make things simple, assume that $\beta = 0$, so that income does not appear in (2.48) nor average income in (2.49). According to the latter, price in village c is

$$p_c = \gamma^{-1}(\alpha_0 + \alpha_c - z_c). \tag{2.50}$$

Write $\hat{\gamma}$ for the OLS estimate of γ obtained by regressing individual household demands on the price in the village in which the household lives. Provided that tastes are uncorrelated with harvests, it is straightforward to show that

$$\operatorname{plim} \hat{\gamma} = \frac{\gamma \sigma_z^2}{\sigma_\alpha^2 + \sigma_z^2}. \tag{2.51}$$

The price response is biased downwards; in addition to the negative effect of price on demand, there is a *positive* effect from demand to price that comes from the effect on both of village-level tastes. The bias will only vanish when the village taste effects α_c are absent, and will be large if the variance of tastes is large relative to the variance of the harvest.

Example 2. Farm size and farm productivity

Consider a model of the determinants of agricultural productivity, and in particular the old question of whether larger or smaller farms are more productive; the observation of an inverse relationship between farm size and productivity goes back to Chayanov (1925), and has acquired the status of a stylized fact; see Sen (1962) for India and Berry and Cline (1979) for reviews.

To examine the proposition, we might use survey data to regress output per hectare on farm size and on other variables not shown, *viz.*

$$\ln(Q_i / A_i) = \alpha + \beta \ln A_i + u_i \tag{2.52}$$

where Q_i is farm output, A_i is farm size, and the common finding is that $\beta < 0$, so that small farms are "more productive" than large farms. This might be interpreted to mean that, compared with hired labor, family labor is of better quality, more safely entrusted with valuable animals or machinery, and needs less monitoring (see Feder 1985; Otsuka, Chuma, and Hayami 1992; and Johnson and Ruttan 1994), or as an optimal response by small farmers to uncertainty (see Srinivasan 1972). It has also sometimes been interpreted as a sign of *inefficiency*, and of dualistic labor markets, because in the absence of smoothly operating labor markets farmers may be forced to work too much on their own farms, pushing their marginal productivity below the market wage (see particularly Sen 1966, 1975). However, if a relationship like (2.52) is estimated on a cross section of farms, and even if the amount of land is outside the control of the farmer, (2.52) is likely to suffer from what are effectively simultaneity problems. Such issues have the distinction of being among the very first topics studied in the early days of econometrics (see Marschak and Andrews 1944).

Although it may be reasonable to suppose that the farmer treats his farm size as fixed when deciding what to plant and how hard to work, this does not mean that A_i is uncorrelated with u_i in (2.52). Farm size may not be controlled by the farmer, but farms do not get to be the size they are at random. The mechanism determining farm size will differ from place to place and time to time, but it is unlikely to be independent of the *quality* of the land. "Desert" farms that are used for low-intensity animal grazing are typically larger than "garden" farms, where the land is rich and output per hectare is high. Such a correlation will be present whether farms are allocated by the market—low-quality land is cheaper per hectare so that it is easier for an owner-occupier to buy a large farm—or by state-mandated land schemes—each farmer is given a plot large enough to make a living. In consequence, the right-hand side of (2.52) is at least partly determined by the left-hand side, and regression estimates of β will be biased downward.

We can also give this simultaneity an omitted variable interpretation where land quality is the missing variable; if quality could be included in the regression instead of in the residual, the new residual could more plausibly be treated as orthogonal to farm size. At the same time, the coefficient β would more nearly measure the effect of land size, and not as in (2.52) the effect of land size contam-

inated by the (negative) projection of land quality on farm size. Indeed, when data are available on land quality—Bhalla and Roy (1988)—or when quality is controlled by IV methods—Benjamin (1993)—there is little or no evidence of a negative relationship between farm size and productivity.

The effect of an omitted variable is worth recording explicitly, since the formula is one of the most useful in the econometrician's toolbox, and is routinely used to assess results and to calculate the direction of bias caused by the omission. Suppose that the correct model is

$$y_i = \alpha + \beta x_i + \gamma z_i + u_i \tag{2.53}$$

and that we have data on y and x, but not on z. In the current example, y is yield, and z is land quality. If we run the regression of y on x, the probability limit of the OLS estimate of β is

$$\text{plim}\hat{\beta} = \beta + \gamma \frac{\text{cov}(x,z)}{\text{var}\, x}. \tag{2.54}$$

In the example, it might be the case that $\beta = 0$, so that farm size has no effect on yields conditional on land quality. But $\gamma > 0$, because better land has higher yields, and the probability limit of $\hat{\beta}$ will be negative because farm size and land quality are negatively correlated.

The land quality problem arises in a similar form if we attempt to use equations like (2.52) to measure the effects on output of extension services or "modern" inputs such as chemical fertilizer. Several studies, Bevan, Collier and Gunning (1989) for Kenya and Tanzania, and Deaton and Benjamin (1988) for Côte d'Ivoire, find that a regression of output on fertilizer input shows extremely high returns, estimates that, if correct, imply considerable inefficiency and scope for government intervention.

Deaton and Benjamin use the 1985 Living Standards Survey of Côte d'Ivoire to estimate the following regression between cocoa output, the structure of the orchard, and the use of fertilizer and insecticide,

$$\begin{aligned} \ln(Q/LM) = {} & 5.621 + 0.526\,(LO/LM) + 0.054\,Insect \\ & (68.5) \quad (4.3) \qquad\qquad (2.5) \\ & + 0.158\,Fert \\ & (2.8) \end{aligned} \tag{2.55}$$

where Q is kilos of cocoa produced on the farm, LM and LO are the numbers of hectares of "mature" and "old" trees, respectively, and *Insect* and *Fert* are expenditures in thousands of Central African francs per cocoa hectare on insecticide and fertilizer, respectively. According to (2.55), an additional 1,000 francs spent on fertilizer will increase the logarithm of output per hectare by 0.158, which at a sample mean log yield of 5.64 implies an additional 48 kilos of cocoa at 400 francs per kilo, or an additional 19,200 francs. However, only slightly more than a half of the cocoa stands are fully mature, and the farmers pay the *mettayeurs* who harvest the crop between a half and a third of the total. But even after these

adjustments, the farmer will be left with a return of 5,400 for an outlay of 1,000 francs. Insecticide is estimated to be somewhat less profitable, and the same calculation gives a return of only 1,800 for each 1,000 francs outlay. Yet only 1 in 14 farmers uses fertilizer, and 1 in 5 uses insecticide.

On the surface, these results seem to indicate very large inefficiencies. However, there are other interpretations. It is likely that highly productive farms are more likely to adopt fertilizer, particularly if the use of fertilizer is an indicator of farmer quality and the general willingness to adopt modern methods, high-yielding varieties, and so on. Credit for fertilizer purchases may only be available to better, or to better-off farmers. Suppose also that some farmers cannot use fertilizer because of local climatic or soil conditions or because the type of trees in their stand, while others have the conditions to make good use of it. When we compare these different farms, we shall find what we have found, that farmers that use fertilizer are more productive, but there is no implication that more fertilizer should be used. Expenditure on fertilizer in (2.55) may do no more than indicate that the orchard contains new hybrid varieties of cocoa trees, something on which the survey did not collect data.

Example 3. The evaluation of projects

Analysis of the effectiveness of government programs and projects has always been a central topic in development economics. Regression analysis seems like a helpful tool in this endeavor, because it enables us to link outcomes—incomes, consumption, employment, health, fertility—to the presence or extent of programs designed to influence them. The econometric problems of such analyses are similar to those we encountered when linking farm outputs to farm inputs. In particular, it is usually impossible to maintain that the explanatory variables—in this case the programs—are uncorrelated with the regression residuals. Government programs are not typically run as experiments, in which some randomly selected groups are treated and others are left alone.

A regression analysis may show that health outcomes are better in areas where the government has put clinics, but such an analysis takes no account of the process whereby sites are chosen. Clinics may be put where health outcomes were previously very poor, so that the cross-section regression will tend to underestimate their effects, or they may be allocated to relatively wealthy districts that are politically powerful, in which case regression analysis will tend to overstate their true impact. Rosenzweig and Wolpin (1986) found evidence of underestimation in the Philippines, where the positive effect of clinics on children's health did not show up in a cross section of children because clinics were allocated first to the areas where they were most needed. The clinics were being allocated in a desirable way, and that fact caused regression analysis to fail to detect the benefits. In the next section, I shall follow Rosenzweig and Wolpin and show how panel data can sometimes be used to circumvent these difficulties. I shall return to the issue of project evaluation later in this section when I come to discuss selection bias, and again in Section 2.6 on IV estimation.

Example 4. Simultaneity and lags: nutrition and productivity

It is important to realize that in cross-section data, simultaneity cannot usually be avoided by using lags to ensure that the right-hand side variables are prior in time to the left-hand side variables. If x precedes y, then it is reasonable to suppose that y cannot affect x directly. However, there is often a third variable that affects y today as well as x yesterday, and if this variable is omitted from the regression, today's y will contain information that is correlated with yesterday's x. The land quality issue in the previous example can be thought about this way; although farm size is determined before the farmer's input and effort decisions, and before they and the weather determine farm output, both output and inputs are affected by land quality, so that there remains a correlation between output and the predetermined variables. As a final example, consider one of the more intractable cases of simultaneity, between nourishment and productivity. If poor people cannot work because they are malnourished, and they cannot eat because they do not earn enough, poor people are excluded from the labor market and there is persistent unemployment and destitution. The theory of this interaction was developed by Mirrlees (1975) and Stiglitz (1976), and it has been argued that such a mechanism helps account for destitution in India (Dasgupta 1993) and for the slow pace of premodern development in Europe (Fogel 1994).

People who eat better may be more productive, because they have more energy and work more efficiently, but people who work more efficiently also earn more, out of which they will spend more on food. Disentangling the effect of nutrition on wages from the Engel curve for food is difficult, and as emphasized by Bliss and Stern (1981), it is far from clear that the two effects can ever be disentangled. One possibility, given suitable data, is to suppose that productivity depends on nutrition with a lag—sustained nutrition is needed for work—while consumption depends on current income. Hence, if y_{it} is the productivity of individual i at time t, and c_{it} is consumption of calories, we might write

$$\begin{aligned} y_{it} &= \alpha_1 + \beta_1 c_{it-1} + \gamma_1 z_{1it} + u_{1it} \\ c_{it} &= \alpha_2 + \beta_2 y_{it} + \gamma_2 z_{2it} + u_{2it} \end{aligned} \tag{2.56}$$

where z_1 and z_2 are other variables needed to identify the system. Provided equation (2.56) is correct and the two error terms are serially independent, both equations can consistently be estimated by least squares in a cross section with information on lagged consumption. However, any form of serial dependence in the residuals u_{1it} will make OLS estimates of the first equation inconsistent. But there is a good reason to suppose that these residuals will be serially correlated, since permanent productivity differences across people that are not attributable to nutrition or the other variables will add a constant "individual" component to the error. Individuals who are more productive in one period are likely to be more productive in the next, even when we have controlled for their nutrition and other observable covariates. More productive individuals will have higher incomes and

higher levels of nutrition, not only today but also yesterday, so that the lag in the equation no longer removes the correlation between the error term and the-right-hand-side variable. In a cross section, predetermined variables can rarely be legitimately treated as exogenous.

Measurement error

Measurement error in survey data is a fact of life, and while it is not always possible to counter its effects, it is always important to realize what those effects are likely to be, and to beware of inferences that are possibly attributable to, or contaminated by, measurement error.

The textbook case is the univariate regression model where both the explanatory and dependent variables are subject to mean-zero errors of measurement. Hence, for the correctly measured variables y and x, we have the linear relationship

$$y_i = \alpha + \beta x_i + u_i \tag{2.57}$$

together with the measurement equations

$$\tilde{x}_i = x_i + \epsilon_{1i} \quad \tilde{y}_i = y_i + \epsilon_{2i} \tag{2.58}$$

where the measurement error is assumed to be orthogonal to the true variables. *Faute de mieux*, $\tilde{y}$ is regressed on $\tilde{x}$, and the OLS parameter estimate of β has the probability limit

$$\text{plim}\hat{\beta} = \frac{\beta m_{xx}}{m_{xx} + \sigma_1^2} = \beta\lambda_0 \tag{2.59}$$

where m_{xx} is the variance of the unobservable, correctly measured x, and σ_1^2 is the variance of the measurement error in x. Equation (2.59) is the "iron law of econometrics," that the OLS estimate of β is biased towards zero, or "attenuated." The degree of attenuation is the ratio of signal to combined signal and noise, λ_0, the *reliability ratio*. The presence of measurement error in the *dependent* variable does not bias the regression coefficients, because it simply adds to the variance of the equation as a whole. Of course, this measurement error, like the measurement error in x, will decrease the precision with which the parameters are estimated.

Attenuation bias is amplified by the addition of correctly measured explanatory variables to the bivariate regression (2.57). Suppose we add a vector z to the right-hand side of (2.57), and assume that z is uncorrelated with the measurement error in $\tilde{x}$ and with the original residuals. Then the probability limit of the OLS estimate of β, the coefficient of $\tilde{x}$, is now $\beta\lambda_1$, where the new reliability ratio λ_1 is

$$\lambda_1 = \frac{\lambda_0 - R_{xz}^2}{1 - R_{xz}^2} \leq \lambda_0 \tag{2.60}$$

and R^2_{xz} is the R^2 from the regression of $\tilde{x}$ on z. The new explanatory variables z "soak up" some of the signal from the noisy regressor $\tilde{x}$, so that the reliability ratio for β is reduced, and the "iron law" more severely enforced.

More generally, consider a multivariate regression where all regressors may be noisy and where the measurement error in the independent variables may be correlated with the measurement error in the dependent variable. Suppose that the correctly measured variables satisfy

$$y = X\beta + u. \tag{2.61}$$

Then the OLS parameter estimates have probability limits given by

$$\text{plim}\hat{\beta} = (M+\Omega)^{-1}M\beta + (M+\Omega)^{-1}\gamma \tag{2.62}$$

where M is the moment matrix of the true x's, Ω is the variance-covariance matrix of the measurement error in the $\tilde{x}$'s, and γ is the vector of covariances between the measurement errors in the $\tilde{x}$'s and the measurement error in $\tilde{y}$. The first term in (2.62) is the matrix generalization of the attenuation effect in the univariate regression—the vector of parameters is subject to a matrix rather than scalar shrinkage factor—while the second term captures any additional bias from a correlation between the measurement errors in dependent and independent variables. The latter effects can be important; for example, if consumption is being regressed on income, and if there is a common and noisily measured imputation term in both—home-produced food, or the imputed value of owner-occupied housing—then there will be an additional source of bias beyond attenuation effects. Even in the absence of this second term on the right-hand side of (2.62) and, in spite of the obvious generalization from scalar to matrix attenuation, the result does not yield any simple result on the direction of bias in any one coefficient (unless, of course, Ω is diagonal).

One useful general lesson is to be specific about the structure of measurement error, and to use a richer and more appropriate specification than the standard one of mean-zero, independent noise. The analysis is rarely complex, is frequently worthwhile, and will not always lead to the standard attenuation result. One specific example is worth a brief discussion. It arises frequently and is simple, but is nevertheless sometimes misunderstood. Consider the model

$$y_{ic} = \alpha + \beta x_{ic} + \gamma z_c + u_{ic} \tag{2.63}$$

where i is an individual who lives in village c, y_{ic} is an outcome variable, x_{ic} and z_c are individual and village-level explanatory variables. In a typical example, y might be a measure of educational attainment, x a set of family background variables, and z a measure of educational provision or school quality in the village. The effect of health provision on health status might be another example. What often happens in practice is that the z-variables are obtained from administrative, not survey data, so that we do not have village-level data on z, but only broader

measures, perhaps at a district or provincial level. These measures are error-ridden proxies for the ideal measures, and it might seem that the iron law would apply. But this is not so.

To see why, write z_p for the broad measure—p is for province—so that

$$z_p = n_p^{-1} \sum_{c \in p} z_c \tag{2.64}$$

where n_p is the number of villages in the province. Hence, instead of the measurement equation (2.58) where the observable is the unobservable plus an unrelated measurement error, we have

$$z_c = z_p + \epsilon_c \tag{2.65}$$

and it is now the observable z_p that is orthogonal to the measurement error. Because the measurement error in (2.65) is the deviation of the village-level z from its provincial mean, it is orthogonal to the observed z_p by construction. As a result, when we run the regression (2.63) with provincial data replacing village data, there is no correlation between the explanatory variables and the error term, and the OLS estimates are unbiased and consistent. Of course, the loss of the village-level information is not without cost. By (2.65), the averages are less variable than the individuals, so that the precision of the estimates will be reduced. And we must always be careful in these cases to correct standard errors for group effects as discussed in Section 2.2 above. But there is no errors-in-variables attenuation bias.

In Section 2.6 below, I review how, in favorable circumstances, IV techniques can be used to obtain consistent estimates of the parameters even in the presence of measurement error. Note, however, that if it is possible to obtain estimates of measurement error variances and covariances, σ_1^2 in (2.59) or Ω and γ in (2.62), then the biases can be corrected and consistent estimates obtained by substituting the OLS estimate on the left-hand side of (2.62), replacing Ω, γ, and M on the right-hand side by their estimates, and solving for β. For (2.62), this leads to the estimator

$$b = (\tilde{X}'\tilde{X} - n\Omega)^{-1}(\tilde{X}'\tilde{y} - n\gamma) \tag{2.66}$$

where n is the sample size, and the tildes denote variables measured with error. The estimator (2.66) is consistent if Ω and γ are known or are replaced by consistent estimates. This option will not always be available, but is sometimes possible, for example, when there are several mismeasured estimates of the same quantity, and we shall see practical examples in Section 5.3 and 5.4 below.

Selectivity issues

In Chapter 1 and the first sections of this chapter, I discussed the construction of samples, and the fact that the sample design frequently needs to be taken into account when estimating characteristics of the underlying population. This is

particularly important when the selection of the sample is related to the quantity under study; average travel time in a sample of travelers is likely to be quite unrepresentative of average travel time among the population as a whole: if wages influence the decision to work, average wages among workers—which are often the only wages observed—will be an upward-biased estimator of actual and potential wages. Sample selection also affects behavioral relationships. In one of the first and most famous examples, Gronau (1973) found that women's wages were higher when they had small children, a result whose inherent implausibility prompted the search for an alternative explanation, and which led to the selection story. Women with children have higher reservation wages, fewer of them work, and the wages of those who do are higher. As with the other cases in this section, the econometric problem is the induced correlation between the error terms and the regressors. In the Gronau example, the more valuable is a woman's time at home, the larger will have to be the unobserved component in her wages in order to induce her to work, so that among working women, there is a positive correlation between the number of children and the error term in the wage equation.

A useful and quite general model of selectivity is given in Heckman (1990); according to this there are two different regressions or regimes, and the model switches between them according to a dichotomous "switch" that is itself explained. The model is written:

$$y_{0i} = x_{0i}'\beta_0 + u_{0i}, \quad y_{1i} = x_{1i}'\beta_1 + u_{1i} \tag{2.67}$$

together with the {1,0} variable d_i which satisfies

$$d_i = 1(z_i'\gamma + u_{2i} > 0) \tag{2.68}$$

where the indicator function 1(.) takes the value 1 when the statement it contains is true, and is zero otherwise. The observed variable y_i is determined according to

$$y_i = d_i y_{0i} + (1 - d_i) y_{1i}. \tag{2.69}$$

The model is sometimes used in almost exactly this form; for example, the two equations in (2.67) could be wage equations in the formal and informal sectors respectively, while (2.68) models the decision about which sector to join (see, for example, van der Gaag, Stelcner and Vijverberg 1989 for a model of this sort applied to LSMS data from Peru and Côte d'Ivoire). However, it also covers several special cases, many of them useful in their own right.

If the right-hand side of the second equation in (2.67) were zero, as it would be if $\beta_1 = 0$ and the variance of u_1 were zero, we would have the censored regression model or generalized Tobit. This further specializes to the Tobit model if the argument of (2.68) and the right-hand side of the first equation coincide, so that the switching behavior and the size of the response are controlled by the same factors. However, the generalized Tobit model is also useful; for example, it is often argued that the factors that determine whether or not people smoke tobacco

are different from the factors that determine how much smokers smoke. In this case, (2.69) implies that for those values of y that are positive, the regression function is

$$E(y_i|x_i, z_i, y_i>0) = x_i'\beta + \lambda(z_i'\gamma) \tag{2.70}$$

where, since there is only one x and one β, I have dropped the zero suffix, and where the last term is defined by

$$\lambda(z_i'\gamma) = E(u_{0i}|u_{2i} \geq -z_i'\gamma). \tag{2.71}$$

(Compare this with the Tobit in (2.47) above.) This version of the model can also be used to think about the case where the data are truncated, rather than censored as in the Tobit and generalized Tobit. Censoring refers to the case where observations that fall outside limits—in this case below zero—are replaced by the limit points, hence the term "censoring." With truncation, observations beyond the limit are discarded and do not appear in our data. Censoring is easier to deal with because, although we do not observe the underlying latent variable, individual observations are either censored or not censored, and for both we observe the covariates x and z, so that it is possible to estimate the switching equation (2.68) as well as (2.70). With truncation, we know nothing about the truncated observations, so that we cannot estimate the switching process, and we are restricted to (2.70). The missing information in the truncated regression makes it difficult to handle convincingly, and it should be avoided when possible.

A second important special case of the general model is the "treatment" or "policy evaluation" case. In the standard version, the right-hand sides of the two switching regressions in (2.67) are taken to be identical apart from their constant terms, so that (2.69) takes the special form

$$y_i = \alpha + \theta d_i + x_i'\beta + u_i \tag{2.72}$$

so that the parameter θ is the effect on the outcome variable of whether or not the "treatment" is applied. If this were a controlled and randomized experiment, the randomization would guarantee that d_i would be orthogonal to u_i. However, since u_2 in (2.68) is correlated with the error terms in the regressions in (2.67), least squares will not yield consistent estimates of (2.72) because d_i is correlated with u_i. This model is the standard one for examining union wage differentials, for example, but it also applies to many important applications in development where d_i indicates the presence of some policy or project. The siting of health clinics and schools are the perhaps the most obvious examples. As we have already seen above, this version of the model can also be thought of in terms of simultaneity bias.

There are various methods of estimating the general model and its variants. One possibility is to specify some distribution for the three sets of disturbances in (2.67) and (2.68), typically joint normality, and then to estimate by maximum likelihood. Given normality, the γ-parameters in (2.68) can be estimated (up to

scale) by probit, and again given normality, the λ-function in (2.71) has a specific form—the (inverse) Mills' ratio—and as Heckman (1976) showed in a famous paper, the results from the probit can be substituted into (2.70) in such a way that the remaining unknown parameters can be estimated by least squares. Since I shall refer to this again, it is worth briefly reviewing the mechanics.

When u_0 and u_2 are jointly normally distributed, the expectation of each conditional on the other is linear, so that we can write

$$u_{0i} = \sigma_0 \rho(u_{2i}/\sigma_2) + \epsilon_i \tag{2.73}$$

where ϵ_i is orthogonal to u_{2i}, σ_0 and σ_2 are the two standard deviations, and ρ is the correlation coefficient. (Note that $\rho\sigma_0/\sigma_2 = \sigma_{02}/\sigma_2^2$ is the large-sample regression coefficient of u_0 on u_2, the ratio of the covariance to variance.) Given (2.73), we can rewrite (2.71) as

$$\lambda(z_i'\gamma) = \rho\sigma_0 E\left(\frac{u_{2i}}{\sigma_2} \Big| \frac{u_{2i}}{\sigma_2} > -\frac{z_i'\gamma}{\sigma_2}\right) = \rho\sigma_0 \frac{\phi(z_i'\gamma/\sigma_2}{\Phi(z_i'\gamma/\sigma_2)} \tag{2.74}$$

where $\phi(.)$ and $\Phi(.)$ are the density and distribution functions of the standard normal distribution, and where the final formula relies on the special properties of the normal distribution. The regression function (2.70) can then be written as

$$y_i = x_i'\beta + \rho\sigma_0 \frac{\phi(z_i'\gamma/\sigma_2)}{\Phi(z_i'\gamma/\sigma_2)}. \tag{2.75}$$

The vector of ratios γ/σ_2 can be estimated by running a probit on the dichotomous d_i from (2.68), the estimates used to compute the inverse Mills' ratio on the right-hand side of (2.75), and consistent estimates of β and $\rho\sigma_0$ obtained by OLS regression.

This "Heckit" (Heckman's probit) procedure is widely used in the empirical development literature, to the extent that it is almost routinely applied as a method of dealing with selectivity bias. In recent years, however, it has been increasingly realized that the normality assumptions in these and similar procedures are far from incidental, and that the results—and even the identification of the models—may be compromised if we are not prepared to maintain normality. Even when normality holds, there will be the difficulties with heteroskedasticity that we have already seen. Recent work has been concerned with the logically prior question as to whether and under what conditions the parameters of these models are identified without further parametric distributional assumptions, and with how identified models can be estimated in a way that is consistent and at least reasonably efficient under the sort of assumptions that make sense in practice.

The identification of the general model turns out to be a delicate matter, and is discussed in Chamberlain (1986), Manski (1988), and Heckman (1990). Given data on which observations are in which regime, the switching equation (2.68) is identified without further distributional assumptions; at least if we make the (essentially normalizing) assumption that the variance of u_2 is unity. The identification of the other equations requires that there be at least one variable in the

switching equation that does not appear in the substantive equations, and even then there can be difficulties; for example, identification requires that the variables unique to the switching equation be continuous. In many practical applications, these conditions will not be met, or at best be controversial. In particular, it is often difficult to exclude any of the selection variables from the substantive equations. Gronau's example, in which children clearly do not belong in the wage equation, seems to be the exception rather than the rule, and unless it is clear how the selection mechanism is working, there seems little point in pursuing these sorts of models, as opposed to a standard investigation of appropriate conditioning variables and how they enter the regression function.

The robust estimation of the parameters of selection models is a live research topic, although the methods are still experimental, and there is far from general agreement on which are best. In the censoring model (2.70), there exist distribution-free methods that generalize Heckman's two stage procedure (see, for example, Newey, Powell, and Walker 1990, who make use of the kernel estimation methods that are discussed in Chapters 3 and 4 below.

One possible move in this direction is to retain a probit—or even linear probability model, regressing d_i on z_i—for the first-stage estimation of (2.68), and to use the estimates to form the index $z'\gamma$, which is entered in the second-stage regression (2.70), not through the Mills' ratio as in (2.75), but in polynomial form, with the polynomial regarded as an approximation to whatever the true λ-function should be. This is perhaps an unusual mixture of parametric and nonparametric techniques, but the probit model or linear probability model (if the probabilities are typically far from either zero or one) are typically acceptable as functional forms, and it makes most sense to focus on removing the normality assumptions.

The "policy evaluation" or "treatment" model (2.72) is most obviously estimated using IV techniques as described in Section 2.6 below. Note that the classic experimental case corresponds to the case where treatment is randomly assigned, or is randomly assigned to certain groups, so that in either case the u_{2i} in (2.68) is uncorrelated with the errors in the outcome equations (2.67). In most economic applications, the "treatment" has at least some element of self-selection, so that d_i in (2.72) will be correlated with the errors, and instrumentation is required. The obvious instruments are the z-variables, although in practice there will often be difficulties in finding instruments that can be plausibly excluded from the substantive equation. Good instruments in this case can sometimes be provided by "natural experiments," where some feature of the policy design allows the construction of "treatments" and "controls" that are not self-selected. I shall discuss these in more detail below.

2.5 Panel data

When our data contain repeated observations on each individual, the resulting panel data open up a number of possibilities that are not available in the single cross section. In particular, the opportunity to compare the same individual under different circumstances permits the possibility of using that individual as his or

her own control, so that we can come closer to the ideal experimental situation. In the farm example of the previous section, the quality of the farm—or indeed of the farmer—can be controlled for, and indeed, the first use of panel data in econometrics was by Mundlak (1961)—see also Hoch (1955)—who estimated farm production functions controlling for the quality of farm management. Similarly, we have seen that the use of regression for project evaluation is often invalidated by the purposeful allocation of projects to regions or villages, so that the explanatory variable—the presence or absence of the project—is correlated with unobserved characteristics of the village. Rosenzweig and Wolpin (1986) and Pitt, Rosenzweig, and Gibbons (1993) have made good use of panel data to test for such effects in educational, health, and family planning programs in the Philippines and Indonesia.

Several different kinds of panel data are sometimes available in developing countries (see also the discussion in Section 1.1 above). A very few surveys—most notably the ICRISAT survey in India—have followed the same households over a substantial period of time. In some of the LSMS surveys, households were visited twice, a year apart, and there are several cases of opportunistic surveys returning to households for repeat interviews, often with a gap of several years. Since many important changes take time to occur, and projects and policies take time to have their effect, the longer gap often produces more useful data. It is also possible to "create" panel data from cross-sectional data, usually by aggregation. For example, while it is not usually possible to match individuals from one census to another, it is frequently possible to match locations, so as to create a panel at the location level. A good example is Pitt, Rosenzweig, and Gibbons (1993), who use several different cross-sectional surveys to construct data on facilities for 1980 and 1985 for 3,302 *kecamatan* (subdistricts) in Indonesia. In Section 2.7 below, I discuss another important example in some detail, the use of repeated but independent cross sections to construct panel data on birth cohorts of individuals. For all of these kinds of data, there are opportunities that are not available with a single cross-sectional survey.

Dealing with heterogeneity: difference- and within-estimation

To see the main advantage of panel data, start from the linear regression model

$$y_{it} = \beta' x_{it} + \theta_i + \mu_t + u_{it} \tag{2.76}$$

where the index i runs from 1 to n, the sample size, and t from 1 to T, where T is usually small, often just two. The quantity μ_t is a time (or macro) effect, that applies to all individuals in the sample at time t. The parameter θ_i is a fixed effect for observation i; in the farm size example above it would be unobservable land quality, in the nutritional wage example, it would be the unobservable personal productivity characteristic of the individual, and in the project evaluation case, it would be some unmeasured characteristic of the individual (or of the individual's region) that affects program allocation. These fixed effects are designed to cap-

ture the heterogeneity that causes the inconsistency in the OLS cross-sectional regression, and are set up in such a way as to allow their control using panel data. Note that there is nothing to prevent us from thinking of the θ's as randomly distributed over the population—so that in this sense the term "fixed effects" is an unfortunate one—but we are not prepared to assume that they are uncorrelated with the observed x's in the regression. Indeed, it is precisely this correlation that is the source of the difficulty in the farm, project evaluation, and nutrition examples.

The fact that we have more than one observation on each of the sample points allows us to remove the θ's by taking differences, or when there are more than two observations, by subtracting (or "sweeping out") the individual means. Suppose that $T = 2$, so that from (2.76), we can write

$$y_{i2} - y_{i1} = (\mu_2 - \mu_1) + \beta'(x_{i2} - x_{i1}) + u_{i2} - u_{i1} \tag{2.77}$$

an equation that can be consistently and efficiently estimated by OLS. When T is greater than two, use (2.76) to give

$$y_{it} - \bar{y}_{i.} = (\mu_t - \bar{\mu}) + \beta'(x_{i2} - \bar{x}_{i.}) + u_{it} - \bar{u}_{i.} \tag{2.78}$$

where the notation $\bar{y}_{i.}$ denotes the time mean for individual i. Equation (2.78) can be estimated as a pooled regression by OLS, although it should be noted (a) that there are $n(T-1)$ independent observations, not nT. Neither (2.77) nor (2.78) contains the individual fixed effects θ_i, so that these regressions are free of any correlation between the explanatory variables and the unobserved fixed effects, and the parameters can be estimated consistently by OLS. Of course, the fixed effect must indeed be fixed over time—which there is often little reason to suppose—and it must enter the equation additively and linearly. But given these assumptions, OLS estimation of the suitably transformed regression will yield consistent estimates in the presence of unobserved heterogeneity—or omitted variables—even when that heterogeneity is correlated with one or more of the included right-hand side variables.

In the example from the Philippines studied by Rosenzweig and Wolpin (1986), there are data on 274 children from 85 households in 20 *barrios*. The cross-section regression of child nutritional status (age-standardized height) on exposure to rural health units and family planning programs gives negative (and insignificant) coefficients on both. Because the children were observed in two years, 1975 and 1979, it is also possible to run (2.77), where changes in height are regressed on changes in exposure, in which regression both coefficients become positive. Such a result is plausible if the programs were indeed effective, but were allocated first to those who needed them the most.

The benefit of eliminating unobserved heterogeneity does not come without cost, and a number of points should be noted. Note first that the regression (2.77) has exactly half as many observations as the regression (2.76), so that, in order to remove the inconsistency, precision has been sacrificed. More generally, with T periods, one is sacrificed to control for the fixed effects, so that the proportional

loss of efficiency is greatest when there are only two observations. Of course, it can be argued that there are limited attractions to the precise estimation of something that we do not wish to know, but a consistent but imprecise estimate can be further from the truth than an inconsistent estimator. The tradeoff between bias and efficiency has to be made on a case-by-case basis. We must also beware of misinterpreting a decrease in efficiency as a change in parameter estimates between the differenced and undifferenced equations. If the cross-section estimate shows that β is positive and significant, and if the differenced data yield an estimate that is insignificantly different from both zero and the cross-section estimate, it is not persuasive to claim that the cross-section result is an artifact of not "treating" the heterogeneity. Second, the differencing will not only sweep out *the* fixed effects, it will sweep out *all* fixed effects, including any regressor that does not change over the period of observation. In some cases, this removes the attraction of the procedure, and will limit it in short panels. In the Ivorian cocoa farming example in the previous section, most of the farmers who used fertilizer reported the same amount in both periods, so that, although the panel data allows us to control for farm fixed effects, it still does not allow us to estimate how much additional production comes from the application of additional fertilizer.

Panel data and measurement error

Perhaps the greatest difficulties for difference- and within-estimators occur in the presence of measurement error. Indeed, when regressors are measured with error, within- or difference-estimators will no longer be consistent in the presence of unobserved individual fixed effects, nor need their biases be less than that of the uncorrected OLS estimator.

Consider the univariate versions of the regressions (2.76) and (2.77), and compare the probability limits of the OLS estimators in the two cases when, in addition to the fixed effects, there is white noise measurement error in x. Again, for simplicity, I compare the results from estimation on a single cross section with those from a two-period panel. The probability limit of the OLS estimator in the cross section (2.76) is given by

$$\text{plim}\hat{\beta} = \frac{\beta m_{xx} + c_{x\theta}}{m_{xx} + \sigma_1^2} \tag{2.79}$$

where $c_{x\theta}$ is the covariance of the fixed effect and the true x, σ_1^2 is the variance of the measurement error, and I have assumed that the measurement errors and fixed effects are uncorrelated. The formula (2.79) is a combination of omitted variable bias, (2.54), and measurement error bias, (2.59). The probability limit of the difference-estimator in (2.77) is

$$\text{plim}\tilde{\beta} = \frac{\beta m_{\Delta}}{m_{\Delta} + \sigma_{\Delta}^2} \tag{2.80}$$

where m_{Δ} is the variance of the difference of the true x, and σ_{Δ}^2 is the variance of the difference of measurement error in x.

That the estimate in the levels suffers from *two* biases—attenuation bias and omitted variable bias—while the difference-estimate suffers from only attenuation bias is clearly no basis for preferring the latter! The relevant question is not the number of biases but whether the differencing reduces the variance in the signal relative to the variance of the noise so that the attenuation bias in the difference-estimator is more severe than the combined attenuation and omitted variable biases in the cross-section regression. We have seen one extreme case already; when the true x does not change between the two periods, the estimator will be dominated by the measurement error and will converge to zero. Although the extreme case would often be apparent in advance, there are many cases where the cross-section variance is much larger than the variance in the changes over time, especially when the panel observations are not very far apart in time. Although measurement error may also be serially correlated, with the same individual misreporting in the same way at different times, there will be other cases where errors are uncorrelated over time, in which case the error difference will have twice the variance of the errors in levels.

Consider again the two examples of farm productivity and nutritional wages, where individual fixed effects are arguably important. In the first case, m_{xx} is the cross-sectional variance of farm size, while m_{Δ} is the cross-sectional variance of the change in farm size from one period to another, something that will usually be small or even zero. In the nutritional wage example, there is probably much greater variation in eating habits between people than there is for the same person over time, so that once again, the potential for measurement error to do harm is much enhanced. One rather different case is worth recording since it is a rare example of direct evidence on measurement error. Bound and Krueger (1991) matched earnings data from the U.S. Current Population Survey with Social Security records, and were thus able to calculate the measurement error in the former. They found that measurement error was serially correlated and negatively related to actual earnings. The reliability ratios—the ratios of signal variance to total variance—which are also the multipliers of β in (2.79) and (2.80), fall from 0.82 in levels to 0.65 in differences for men, and from 0.92 to 0.81 for women.

Since measurement error is omnipresent, and because of the relative inefficiency of difference- and within-estimators, we must be careful never to assume that the use of panel data will automatically improve our inference, or to treat the estimate from panel data as a gold standard for judging other estimates. Nevertheless, it is clear that there is more information in a panel than in a single cross section, and that this information can be used to improve inference. Much can be learned from comparing different estimates. If the difference-estimate has a different sign from the cross-sectional estimate, inspection of (2.79) and (2.80) shows that the covariance between x and the heterogeneity must be nonzero; measurement error alone cannot change the signs. When there are several periods of panel data, the difference-estimator (2.77) and the within-estimator (2.78) are mathematically distinct, and in the presence of measurement error will have different probability limits. Griliches and Hausman (1986) show how the comparison of these two estimators can identify the variance of the measurement error

when the errors are independent over time—so that consistent estimators can be constructed using (2.66). When errors are correlated over time—as will be the case if households persistently make errors in the same direction—information on measurement error can be obtained by comparing parameters from regressions computed using alternative differences, one period apart, two periods apart, and so on.

Lagged dependent variables and exogeneity in panel data

Although it will not be of great concern for this book, I should also note that there are a number of specific difficulties that arise when panel data are used to estimate regressions containing lagged dependent variables. In ordinary linear regressions, serial correlation in the residuals makes OLS inconsistent in the presence of a lagged dependent variable. In panel data, the presence of unobserved individual heterogeneity will have the same effect; if farm output is affected by unobserved farm quality, so must be last period's output on the same farm, so that this period's residual will be correlated with the lagged dependent variable. Nor can the heterogeneity be dealt with by using the standard within- or difference-estimators. When there is a lagged dependent variable together with unobserved fixed effects, and we difference, the right-hand side of the equation will have the lagged difference $y_{it-1} - y_{it-12}$, and although the fixed effects have been removed by the differencing, there is a differenced error term $u_{it} - u_{it-1}$, which is correlated with the lagged difference because u_{it-1} is correlated with y_{it-1}. Similarly, the within-estimator is inconsistent because the deviation of lagged y_{it-1} from its mean over time is correlated, with the deviation of u_{it} from its mean, not because u_{it} is correlated with y_{it-1}, but because the two means are correlated. These inconsistencies vanish as the number of time periods in the panel increases but, in practice, most panels are short.

Nor are the problems confined to lagged-dependent variables. Even if all the right-hand side variables are uncorrelated with the contemporaneous regression error u_{it}, the deviations from their means can be correlated with the average over time, $u_{i.}$. For this not to be the case, we require that explanatory variables be uncorrelated with the errors at all lags and leads, a requirement that is much more stringent than the usual assumption in time-series work that a variable is predetermined. It is also a requirement that is unlikely to be met in several of the examples I have been discussing. For example, farm yields may depend on farm size, on the weather, on farm inputs such as fertilizer and insecticide, and on (unobserved) quality. The inputs are chosen before the farmer knows output, but a good output in one year may make the farmer more willing, or more able, to use more inputs in a subsequent year. In such circumstances, the within-regression will eliminate the unobservable quality variable, but it will induce a correlation between inputs and the error term, so that the within-estimator will be inconsistent.

These problems are extremely difficult to deal with in a convincing and robust way, although there exist a number of techniques (see in particular Nickell 1981; Chamberlain 1984; Holtz-Eakin, Newey, and Rosen 1988; and particularly the

series of papers, Arellano and Bond 1991, Arellano and Bover 1993, and Alonso-Borrego and Arellano 1996). But too much should not be expected from these methods; attempts to disentangle heterogeneity, on the one hand, and dynamics, on the other, have a long and difficult history in various branches of statistics and econometrics.

2.6 Instrumental variables

In all of the cases discussed in Section 2.4, the regression function differs from the structural model because of correlation between the error terms and the explanatory variables. The reasons differ from case to case, but it is the correlation that produces the inconsistency in OLS estimation. The technique of IV is the standard prescription for correcting such cases, and for recovering the structural parameters. Provided it is possible to find instrumental variables that are correlated with the explanatory variables but uncorrelated with the error terms, then IV regression will yield consistent estimates.

For reference, it is useful to record the formulas. If X is the $n \times k$ matrix of explanatory variables, and if W is an $n \times k$ matrix of instruments, then the IV estimator of β is given by

$$\beta_{IV} = (W'X)^{-1}W'y. \tag{2.81}$$

Since $y = X\beta + u$ and W is orthogonal to u by assumption, (2.81) yields consistent estimators if the premultiplying matrix $W'X$ is of full rank. If there are fewer instruments than explanatory variables—and some explanatory variables will often be suitable to serve as their own instruments—the IV estimate does not exist, and the model is underidentified. When there are exactly as many instruments as explanatory variables, the model is said to be exactly identified. In practice, it is desirable to have more instruments than strictly needed, because the additional instruments can be used either to increase precision or to construct tests. In this overidentified case, suppose that Z is an $n \times k'$ matrix of potential instruments, with $k' > k$. Then all the instruments are used in the construction of the set W by using two-stage least squares, so that at the first stage, each X is regressed on all the instruments Z, with the predicted values used to construct W. If we define the "projection" matrix $P_Z = Z(Z'Z)^{-1}Z'$, the IV estimator is written

$$\beta_{IV} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y = (X'P_ZX)^{-1}X'P_Zy. \tag{2.82}$$

Under standard assumptions, β_{IV} is asymptotically normally distributed with mean β and a variance-covariance matrix that can be estimated by

$$V = (XP_ZX)^{-1}(X'P_ZDP_ZX)(XP_ZX)^{-1}. \tag{2.83}$$

The choice of D depends on the treatment of the variance-covariance matrix of the residuals, and is handled as with OLS, replaced by $\hat{\sigma}^2 I$ under homoskedas-

ticity, or by a diagonal matrix of squared residuals if heteroskedasticity is suspected, or by the appropriate matrix of cluster residuals if the survey is clustered (see (2.30) above). (Note that the residuals must be calculated as $y - X\beta_{IV}$, which is not the vector of residuals from the second stage of two-stage least squares. However, this is hardly ever an issue in practice, since econometric packages make the correction automatically.)

When the model is overidentified, and $k' > k$, the (partial) validity of the instruments is usually assessed by computing an *overidentification* (OID) test statistic. The simplest—and most intuitive—way to calculate the statistic is to regress the IV residuals $y - X\beta_{IV}$ on the matrix of instruments Z and to multiply the resulting (uncentered) R^2 statistic by the sample size n (see Davidson and MacKinnon 1993, pp. 232–37). (The uncentered R^2 is 1 minus the ratio of the sum of squared residuals to the sum of squared dependent variables.) Under the null hypothesis that the instruments are valid, this test statistic is distributed as a χ^2 statistic with $k'-k$ degrees of freedom. This procedure tests whether, contrary to the hypothesis, the instruments play a direct role in determining y, not just an indirect role, through predicting the x's. If the test fails, one or more of the instruments are invalid, and ought to be included in the explanation of y. Put differently, the OID test tells us whether we would get (significantly) different answers if we used different instruments or different combinations of instruments in the regression. This interpretation also clarifies the limitations of the test. It is a test of *over*identification, not of *all* the instruments. If we have only k instruments and k regressors, the model is exactly identified, the residuals of the IV regression are orthogonal to the instruments by construction, so that the OID test is mechanically equal to zero, there is only one way of using the instruments, and no alternative estimates to compare. So the OID test, useful though it is, is only informative when there are more instruments than strictly necessary.

Although estimation by IV is one of the most useful and most used tools of modern econometrics, it does not offer a routine solution for the problems diagnosed in Section 2.4. Just as it is almost always possible to find reasons—measurement error, omitted heterogeneity, selection, or omitted variables—why the structural variables are correlated with the error terms, so is it almost always difficult to find instruments that do not have these problems, while at the same time being related to the structural variables. It is easy to generate estimates that are different from the OLS estimates. What is much harder is to make the case that these estimates are necessarily to be preferred. Credible identification and estimation of structural equations almost always requires real creativity, and creativity cannot be produced to a formula.

Policy evaluation and natural experiments

One promising approach to the selection of instruments, especially for the treatment model, is to look for "natural experiments," cases where different sets of individuals are treated differently in a way that, if not random by design, was effectively so in practice.

One of the best, and certainly earliest, examples is Snow's (1855) analysis of deaths in the London cholera epidemic of 1853–54, work that is cited by Freedman (1991) as a leading example of convincing statistical work in the social sciences. The following is based on Freedman's account. Snow's hypothesis—which was not widely accepted at the time—was that cholera was waterborne. He discovered that households were supplied with water by two different water companies, the Lambeth water company, which in 1849 had moved its water intake to a point in the Thames above the main sewage discharge, and the Southwark and Vauxhall company, whose intake remained below the discharge. There was no sharp separation between houses supplied by the two companies, instead "the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. . . . The experiment, too, is on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge; one group supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity." Snow collected data on the addresses of cholera victims, and found that there were 8.5 times as many deaths per thousand among households supplied by the Southwark and Vauxhall company.

Snow's analysis can be thought of in terms of instrumental variables. Cholera is not directly caused by the position of a water intake, but by contamination of drinking water. Had it then been possible to do so, an alternative analysis might have linked the probability of contracting cholera to a measure of water purity. But even if such an analysis had shown significant results, it would not have been very convincing. The people who drank impure water were also more likely to be poor, and to live in an environment contaminated in many ways, not least by the "poison miasmas" that were then thought to be the cause of cholera. In terms of the discussion of Section 2.4, the explanatory variable, water purity, is correlated with omitted variables or with omitted individual heterogeneity. The identity of the water supplier is an ideal IV for this analysis. It is correlated with the explanatory variable (water purity) for well-understood reasons, and it is uncorrelated with other explanatory variables because of the "intimate" mixing of supplies and the fact that most people did not even know the identity of their supplier.

There are a number of good examples of natural experiments in the economics literature. Card (1989) shows that the Mariel boatlift, where political events in Cuba led to the arrival of 125,000 Cubans in Miami between May and September 1980, had little apparent effect on wages in Miami, for either Cubans or non-Cubans. Card and Krueger (1994) study fast-food outlets on either side of the border between New Jersey and Pennsylvania around the time of an increase in New Jersey's minimum wage, and find that employment rose in New Jersey relative to Pennsylvania. Another example comes from the studies by Angrist (1990) and Angrist and Krueger (1994) into earnings differences of American males by veteran status. The "treatment" variable is spending time in the military, and the out-

come is the effect on wages. The data present somewhat of a puzzle because veterans of World War II appear to enjoy a substantial wage premium over other workers, while veterans of the Vietnam War are typically paid less than other similar workers. The suspicion is that selectivity is important, the argument being that the majority of those who served in Vietnam had relatively low unobservable labor market skills, while in World War II, where the majority served, only those with relatively low skills were excluded from service.

Angrist and Krueger (1994) point out that in the late years of World War II, the selection mechanism acted in such a way that those born early in the year had a (very slightly) higher chance of being selected than those born later in the year. They can then use birth dates as instruments, effectively averaging over all individuals born in the same quarter, so that to preserve variation in the averages, Angrist and Krueger require a very large sample, in this case 300,000 individuals from the 1980 census. (Large sample sizes will often be required by "natural experiments" since instruments that are convincingly uncorrelated with the residuals will often be only weakly correlated with the selection process.) In the IV estimates, the World War II premium is reversed, and earnings are lower for those cohorts who had a larger fraction of veterans. By contrast, Angrist (1990) finds that instrumenting earnings equations for Vietnam veterans using the draft lottery makes little difference to the negative earnings premium experienced by these workers, so that the two studies together suggest that time spent in the military lowers earnings compared with the earnings of those who did not serve.

Impressive as these studies are, natural experiments are not always available when we need them, and some cases yield better instruments than others. Because "natural" experiments are not genuine, randomized experiments, the fact that the experiment is effectively (or quasi-) randomized has to be argued on a case-by-case basis, and the argument is not always as persuasive as in Snow's case. For example, government policies only rarely generate convincing experiments (see Besley and Case 1994). Although two otherwise similar countries (towns, or provinces) may experience different policies, comparison of outcomes is always bedeviled by the concern that the differences are not random, but linked to some characteristic of the country (town or province) that caused the government to draw the distinction in the first place.

However, it may be possible to follow Angrist and Krueger's lead in looking, not at programs themselves, but at the details of their administration. The argument is that in any program with limited resources or limited reach, where some units are treated and some not, the administration of the project is likely to lead, at some level, to choices that are close to random. In the World War II example, it is not the draft that is random, but the fact that local draft boards had to fill quotas, and that the bureaucrats who selected draftees did so partially by order of birth. In other cases, one could imagine people being selected because they are higher in the alphabet than others, or because an administrator used a list constructed for other purposes. While the broad design of the program is likely to be politically and economically motivated, and so cannot be treated as an experiment, natural or otherwise, the details are handled by bureaucrats who are simply trying to get the

job done, and who make selections that are effectively random. This is a recipe for project evaluation that calls for intimate knowledge and examination of detail, but it is one that has some prospect of yielding convincing results.

One feature of good natural experiments is their simplicity. Snow's study is a model in this regard. The argument is straightforward, and is easily explained to nonstatisticians or noneconometricians, to whom the concept of instrumental variables could not be readily communicated. Simplicity not only aids communication, but greatly adds to the persuasiveness of the results and increases the likelihood that the results will affect the policy debate. A case in point is the recent political firestorm in the United States over Card and Krueger's (1994) findings on the minimum wage.

Econometric issues for instrumental variables

IV estimators are invaluable tools for handling nonexperimental data. Even so, there are a number of difficulties of which it is necessary to be aware. As with other techniques for controlling for nonexperimental inconsistencies, there is a cost in terms of precision. The variance-covariance matrix (2.83) exceeds the corresponding OLS matrix by a positive definite matrix, so that, even when there is no inconsistency, the IV estimators—and all linear combinations of the IV estimates—will have larger standard errors than their OLS counterparts. Even when OLS is inconsistent, there is no guarantee that in individual cases, the IV estimates will be closer to the truth, and the larger the variance, the less likely it is that they will be so.

It must also be emphasized that the distributional theory for IV estimates is asymptotic, and that asymptotic approximations may be a poor guide to finite sample performance. Formulas exist for the finite sample distributions of IV estimators (see, for example, Anderson and Sawa 1979) but these are typically not sufficiently transparent to provide practical guidance. Nevertheless, a certain amount is known, and this knowledge provides some warnings for practice.

Finite sample distributions of IV estimators will typically be more dispersed with more mass in the tails than either OLS estimators or their own asymptotic distributions. Indeed, IV estimates possess moments only up to the degree of overidentification, so that when there is one instrument for one suspect structural variable, the IV estimate will be so dispersed that its mean does not exist (see Davidson and MacKinnon 1993, 220–4) for further discussion and references. As a result, there will always be the possibility of obtaining extreme estimates, whose presence is not taken into account in the calculation of the asymptotic standard errors. Given sufficient overidentification so that the requisite moments exist—and note that this rules out some of the most difficult cases—Nagar (1959) and Buse (1992) show that in finite samples, IV estimates are biased towards the OLS estimators. This gives support to many students' intuition when first confronted with IV estimation, that it is a clever trick designed to reproduce the OLS estimate as closely as possible while guaranteeing consistency in a (conveniently hypothetical) large sample. In the extreme case, where there are as many instruments

as observations so that the first stage of two-stage least squares fits the data perfectly, the IV and OLS estimates are identical. More generally, there is a tradeoff between having too many instruments, overfitting at the first stage, and being biased towards OLS, or having too few instruments, and risking dispersion and extreme estimates. Either way, the asymptotic standard errors on which we routinely rely will not properly indicate the degree of bias or the dispersion.

Nelson and Startz (1990a, 1990b) and Maddala and Jeong (1992) have analyzed the case of a univariate regression where the options are OLS or IV estimation with a single instrument. Their results show that the central tendency of the finite-sample distribution of the IV estimator is biased away from the true value and towards the OLS value. Perhaps most seriously, the asymptotic distribution is a very poor approximation to the finite-sample distribution when the instrument is a poor one, in the sense that it is close to orthogonal to the explanatory variable. Additional evidence of poor performance comes from Bound, Jaeger, and Baker (1993), who show that the empirical results in Angrist and Krueger (1991), who used up to 180 instruments with 30,000 observations, can be closely reproduced with randomly generated instruments. Both sets of results show that poor instruments do not necessarily reveal themselves as large standard errors for the IV estimates. Instead it is easy to produce situations in which y is unrelated to x, and where z is a poor instrument for x, but where the IV estimate of the regression of y on x with z as instrument generates a parameter estimate whose "asymptotic t-value" shows an apparently significant effect. As a result, if IV results are to be credible, it is important to establish first that the instruments do indeed have predictive power for the contaminated right-hand-side variables. This means displaying the first-stage regressions—a practice that is far from routine—or at the least examining and presenting evidence on the explanatory power of the instruments. (Note that when calculating two-stage least squares, the exogenous x variables are also included on the right-hand-side with the instruments, and that it is the predictive power of the latter that must be established, for example, by using an F-test for those variables rather than the R^2 for the regression as a whole.)

In recent work, Staiger and Stock (1993) have proposed a new asymptotic theory for IV when the instruments are only weakly correlated with the regressors, and have produced evidence that their asymptotics provides a good approximation to the finite-sample distribution of IV estimators, even in difficult cases such as those examined by Nelson and Startz. These results may provide a better basis for IV inference in future work.

2.7 Using a time series of cross sections

Although long-running panels are rare in both developed and developing countries, independent cross-sectional household surveys are frequently conducted on a regular basis, sometimes annually, and sometimes less frequently. In Chapter 1, I have already referred to and illustrated from the Surveys of Personal Income Distribution in Taiwan (China), which have been running annually since 1976, and I shall use these data further in this section. Although such surveys select

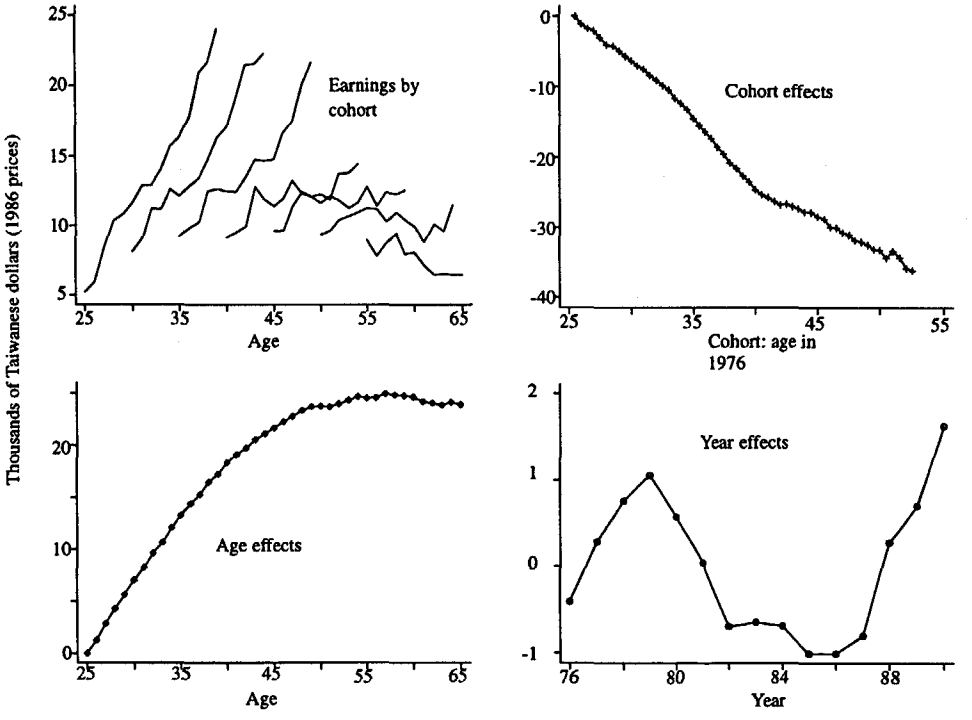
different households in each survey, so that there is no possibility of following *individuals* over time, it is still possible to follow *groups* of people from one survey to another. Obvious examples are the group of the whole population, where we use the surveys to track aggregate data over time, or regional, sectoral, or occupational groups, where we might track the differing fortunes over time of farmers versus government servants, or where we might ask whether poverty is diminishing more rapidly in one region than in another.

Perhaps somewhat less obvious is the use of survey data to follow *cohorts* of individuals over time, where cohorts are defined by date of birth. Provided the population is not much affected by immigration and emigration, and provided the cohort is not so old that its members are dying in significant numbers, we can use successive surveys to follow each cohort over time by looking at the members of the cohort who are randomly selected into each survey. For example, we can look at the average consumption of 30-year-olds in the 1976 survey, of 31-year-olds in the 1977 survey, and so on. These averages, because they relate to the same group of people, have many of the properties of panel data. Cohorts are frequently interesting in their own right, and questions about the gainers and losers from economic development are often conveniently addressed by following such groups over time. Because there are many cohorts alive at one time, cohort data are more diverse and richer than are aggregate data, but their semiaggregated structure provides a link between the microeconomic household-level data and the macroeconomic data from national accounts. The most important measures of living standards, income and consumption, have strong life-cycle age-related components, but the profiles themselves will move upward over time with economic growth as each generation becomes better-off than its predecessors. Tracking different cohorts through successive surveys allows us to disentangle the generational from life-cycle components in income and consumption profiles.

Cohort data: an example

The left-hand top panel of Figure 2.5 shows the averages of real earnings for various cohorts in Taiwan (China) observed from 1976 through to 1990. The data were constructed according to the principles outlined above. For example, for the cohort born in 1941, who were 35 years old in 1976, I used the 1976 survey to calculate the average earnings of all those aged 35, and the result is plotted as the first point in the third line from the left in the figure. The average earnings of 36-year-olds in the 1977 survey is calculated and forms the second point on the same segment. The rest of the line comes from the other surveys, tracking the cohort born in 1941 through the 15 surveys until they are last observed at age 49 in 1990. Table 2.2 shows that there were 699 members of the cohort in the 1976 survey, 624 in the 1977 survey, 879 in the 1978 survey (in which the sample size was increased), and so on until 691 in 1990. The figure illustrates the same process for seven cohorts, born in 1951, 1946, and so on backward at five-year intervals until the oldest, which was born in 1921, and the members of which were 69 years old when last seen in 1990. Although it is possible to make graphs for all

Figure 2.5. Earnings by cohort and their decomposition, Taiwan (China), 1976–90



Note: Author's calculations based on Surveys of Personal Income Distribution.

Table 2.2. Number of persons in selected cohorts by survey year, Taiwan (China), 1976–90

	Cohort: age in 1976						
Year	*25*	*30*	*35*	*40*	*45*	*50*	*55*
1976	863	521	699	609	552	461	333
1977	902	604	624	535	585	427	308
1978	1,389	854	879	738	714	629	477
1979	1,351	796	846	708	714	574	462
1980	1,402	834	845	723	746	625	460
1981	1,460	794	807	720	750	624	426
1982	1,461	771	838	695	689	655	496
1983	1,426	737	846	718	702	597	463
1984	1,477	825	820	711	695	541	454
1985	1,396	766	775	651	617	596	442
1986	1,381	725	713	659	664	549	428
1987	1,309	634	775	632	675	513	0
1988	1,275	674	700	617	595	548	0
1989	1,225	672	652	600	609	519	0
1990	1,121	601	691	575	564	508	0

Note: The year is the year of the survey, and the numbers are the numbers of individuals in each cohort sampled in each survey year. 65 is used as an age cutoff, so the oldest cohort is not observed after 1986.
Source: Author's calculations from the Surveys of Personal Income Distribution.

birth years, I have shown only every fifth cohort so as to keep the diagram clear. Note that only members of the same cohort are joined up by connecting lines, and this construction makes it clear when we are following different groups of people or jumping from one cohort to another. (See also Figures 6.3 and 6.4 below for the corresponding graphs for consumption and for a comparison of cross-sectional and cohort plots.)

The top left-hand panel of the figure shows clear age and cohort effects in earnings; it is also possible to detect common macroeconomic patterns for all cohorts. With a very few exceptions at older ages, the lines for the younger cohorts are always above the lines for the older cohorts, even when they are observed at the same age. This is because rapid economic growth in Taiwan (China) is making younger generations better-off, so that, for example, those born in 1951—the youngest, left-most cohort in the figure—have average earnings at age 38 that are approximately twice as much as the earnings at age 38 of the cohort born 10 years earlier—the third cohort in the figure. There is also a pronounced life-cycle profile to earnings, and although the age profile is "broken up" by the cohort effects, it is clear that earnings tend to grow much more rapidly in the early years of the working life than they do after age 50. As a result, not only are the younger cohorts of workers in Taiwan (China) better-off than their predecessors, but they have also experienced much more rapid growth in earnings. The macroeconomic effects in the first panel of Figure 2.5 are perhaps the hardest to see, but note that each connected line segment corresponds to the same contemporaneous span of 15 years in "real" time, 1976–90. Each segment shows the impact

of the slowdown in Taiwanese economic growth after the 1979 oil shock. Each cohort has very rapid growth from the second to third year observed, which is 1977–78, somewhat slower growth for the next two years, 1978–80, and then two years of slow or negative growth after the shock. This decomposition into cohort, age, and year effects can be formalized in a way that will work even when the data are not annual and not necessarily evenly spaced, a topic to which I return in the final subsection below. Before that, however, it is useful to use this example to highlight the advantages and disadvantages of cohort data more generally.

Cohort data versus panel data

A useful comparison is between the semiaggregated cohort data and genuine panel data in which individual households are tracked over time. In both cases, we have a time series of observations on a number of units, with units defined as either cohorts or individuals. The cohort data cannot tell us anything about dynamics within the cohorts; each survey tells us about the distribution of the characteristic in the cohort in each period, but two adjacent surveys tell us nothing about the joint distribution of the characteristic in the two periods. In the earnings example, the time series of cross sections can tell us about average earnings for the cohort over time, and it can tell us about inequality of earnings within the cohort and how it is changing over time, but it cannot tell us how long individuals are poor, or whether the people who are rich now were rich or poor at some earlier date. But apart from dynamics, the cohort data can do most of what would be expected of panel data. In particular, and as we shall see in the next subsection, the cohort data can be used to control for unobservable fixed effects just as with panel data, a feature that is often thought to be the main econometric attraction of the latter.

Cohort data also have a number of advantages over most panels. As we have seen in Chapter 1, many panels suffer from attrition, especially in the early years, and so run the risk of becoming increasingly unrepresentative over time. Because the cohort data are constructed from fresh samples every year, there is no attrition. There will be (related) problems with the cohort data if the sampling design changes over time, or if the probabilities of selection into the sample depend on age as, for example, for young men undergoing military training. The way in which the cohort data are used will often be less susceptible to measurement error than is the case with panels. The quantity that is being tracked over time is typically an average (or some other statistic such as the median or other percentile) and the averaging will nearly always reduce the effects of measurement error and enhance the signal-to-noise ratio. In this sense, cohort methods can be regarded as IV methods, where the instruments are grouping variables, whose application averages away the measurement error. Working with aggregated data at a level that is intermediate between micro and macro also brings out the relationship between household behavior and the national aggregates and helps bridge the gap between them; in Figure 2.5, for example, the behavior of the aggregate economy is clearly apparent in the averages of the household data.

It should be emphasized that cohort data can be constructed for any characteristic of the distribution of interest; we are not confined to means. As we shall see in Chapter 6, it can be interesting and useful to study how inequality changes within cohorts over time, and since we have the micro data for each cohort in each year, it is as straightforward to work with measures of dispersion as it is to work with measures of central tendency. Medians can be used instead of means—a technique that is often useful in the presence of outliers—and if the theory suggests working with some transform of the data, the transform can be made prior to averaging. When working with aggregate data, theoretical considerations often suggest working with the mean of a logarithm, for example, rather than with the logarithm of the mean. The former is not available from aggregate data, but can be routinely computed from the micro data when calculating the semiaggregated cohort averages.

A final advantage of cohort methods is that they allow the combination of data from different surveys on different households. The means of cohort consumption from an expenditure survey can be combined with the means of cohort income from a labor force survey, and the hybrid data set used to study saving. It is not necessary that all variables are collected from the same households in one survey.

Against the use of cohort data, it should be noted that there are sometimes problems with the assumption that the cohort population is constant, an assumption that is needed if the successive surveys are to generate random samples from the same underlying population. I have already noted potential problems with military service, migration, aging, and death. But the more serious difficulties come when we are forced to work, not with individuals, but with households, and to define cohorts of households by the age of the head. If households once formed are indissoluble, there would be no difficulty, but divorce and remarriage reorganize households, as does the process whereby older people go to live with their children, so that previously "old" households become "young" households in subsequent years. It is usually clear when these problems are serious, and they affect some segments of the population more than others, so that we know which data to trust and which to suspect.

Panel data from successive cross sections

It is useful to consider briefly the issues that arise when using cohort data as if they were repeated observations on individual units. I show first how fixed effects at the individual level carry through to the cohort data, and what steps have to be taken if they are to be eliminated. Consider the simplest univariate model with fixed effects, so that at the level of the individual household, we have (2.76) with a single variable

$$y_{it} = \alpha + \beta x_{it} + \mu_t + \theta_i + u_{it} \tag{2.84}$$

where the μ_t are year dummies and θ_i is an individual-specific fixed effect. If there were no fixed effects, it would be possible to average (2.84) over all the

households in each cohort in each year to give a corresponding equation for the cohort averages. When there are fixed effects, (2.84) still holds for the cohort population means, with cohort fixed effects replacing the individual fixed effects. However, if we average (2.84) over the members of the cohorts who appear in the survey, and who will be different from year to year, the "fixed effect" will not be fixed, because it is the average of the fixed effects of different households in each year. Because of this sampling effect, we cannot remove the cohort fixed effects by differencing or using within-estimators.

Consider an alternative approach based on the unobservable population means for each cohort. Start from the cohort version of (2.84), and denote population means in cohorts by the subscripts c, so that, simply changing the subscript i to c, we have

$$y_{ct} = \alpha + \beta x_{ct} + \mu_t + \theta_c + u_{ct} \tag{2.85}$$

and take first differences—the comparable analysis for the within-estimator is left as an exercise—to eliminate the fixed effects so that

$$\Delta y_{ct} = \Delta \mu_t + \beta \Delta x_{ct} + \Delta u_{ct} \tag{2.86}$$

where the first term is a constant in any given year. This procedure has eliminated the fixed effects, but we are left with the unobservable changes in the population cohort means in place of the sample cohort means, which is what we observe. If we replace Δy and Δx in (2.84) by the observed changes in the sample means, we generate an error-in-variables problem, and the estimates will be attenuated.

There are at least two ways of dealing with this problem. The first is to note that, just as the sample was used to provide an estimate of the cohort mean, it can also be used to provide an estimate of the standard error of the estimate, which in this context is the variance of the measurement error. For the example (2.86), we can use overbars to denote sample means and write

$$\begin{aligned} \Delta \bar{y}_{ct} &= \Delta y_{ct} + \epsilon_{1ct} - \epsilon_{1ct-1} \\ \Delta \bar{x}_{ct} &= \Delta x_{ct} + \epsilon_{2ct} - \epsilon_{2ct-1} \end{aligned} \tag{2.87}$$

where ϵ_{1ct} and ϵ_{2ct} are sampling errors in the cohort means. Because they come from different surveys with independently selected samples, they are independent over time, and their variances and covariance, σ_1^2, σ_2^2, and σ_{12} are calculated in the usual way, from the variances and covariance in the sample divided by the cohort size (with correction for cluster effects as necessary.) From (2.87), we see that the variances and covariances of the sample cohort means are inflated by the variances and covariances of the sampling errors, but that, if these are subtracted out, we can obtain consistent estimates of β in (2.86) from—cf. also (2.62) above,

$$\tilde{\beta} = \frac{\text{cov}(\Delta \bar{x}_{ct} \Delta \bar{y}_{ct}) - \sigma_{12t} - \sigma_{12t-1}}{\text{var}(\Delta \bar{x}_{ct}) - \sigma_{2t}^2 - \sigma_{2t-1}^2} \tag{2.88}$$

and where, for illustrative purposes, I have assumed that there are only two time periods t and $t-1$. The standard error for (2.88) can be calculated using the bootstrap or the delta method—discussed in the next section—which can also take

into account the fact that the variance and covariances of the sampling errors are estimated (see Deaton 1985, who also discusses the multivariate case, and Fuller 1987, who gives a general treatment for a range of similar models).

Another possible estimation strategy is to use IV, with changes from earlier years used as instruments. Since the successive samples are independently drawn, changes in cohort means from $t-2$ to $t-1$ are measured independently of the change from t to $t+1$. In some cases, the cohort samples may be large enough and the means precisely enough estimated so that these corrections are small enough to ignore. In any case, it is a good idea to check the standard errors of the cohort means, to make sure that regression results are not being dominated by sampling effects, and if so, to increase the cohort sizes, for example, by working with five-year age bands instead of single years. In some applications, this might be desirable on other grounds; in some countries, people do not know their dates of birth well enough to be able to report age accurately, and reported ages "heap" at numbers ending in 5 and 0.

Decompositions by age, cohort, and year

A number of the quantities most closely associated with welfare, including family size, earnings, income, and consumption, have distinct and characteristic life-cycle profiles. Wage rates, earnings, and saving usually have hump-shaped age profiles, rising to their maximum in the middle years of life, and declining somewhat thereafter. The natural process of bearing and raising children induces a similar profile in average family size. Moreover, all of these quantities are subject to secular variation; consumption, earnings, and incomes rise over time with economic development, and family size decreases as countries pass through the demographic transition. In consequence, even if the shape of the age profiles remains the same for successive generations, their position will shift from one to the next. The age profile from a single cross section confounds the age profile with the generational or cohort effects. For example, a cross-sectional earnings profile will tend to exaggerate the downturn in earnings at the highest age because, as we look at older and older individuals, we are not just moving along a given age-earnings profile, but we are also moving to ever lower lifetime profiles. The cohort data described in this section allow us to track the same cohort over several years and thus to avoid the difficulty; indeed, the Taiwanese earnings example in Figure 2.5 provides a clear example of the differences between the age profiles of different cohorts. In many cases, diagrams like Figure 2.5 will tell us all that we need to know. However, since each cohort is only observed for a limited period of time, it is useful to have a technique for linking together the age profiles from different cohorts to generate a single complete life-cycle age profile. This is particularly true when there is only a limited number of surveys, and the intervals between them are more than one year. In such cases, diagrams like Figure 2.5 are harder to draw, and a good deal less informative.

In this subsection, I discuss how the cohort data can be decomposed into age effects, cohort effects, and year effects, the first to give the typical age profile, the

second the secular trends that lead to differences in the positions of age profiles for different cohorts, and the third the aggregate effects that synchronously but temporarily move all cohorts off their profiles. These decompositions are based on models and are certainly not free of structural assumptions; they assume away interaction effects between age, cohort, and years, so that, for example, the shape of age profiles is unaffected by changes in their position, and the appropriateness and usefulness of the assumption has to be judged on a case-by-case basis.

To make the analysis concrete, consider the case of the lifetime consumption profile. If the growth in living standards acts so as to move up the consumption-age profiles proportionately, it makes sense to work in logarithms, and to write the logarithm of consumption as

$$\ln c_{ct} = \beta + \alpha_a + \gamma_c + \psi_t + u_{ct} \tag{2.89}$$

where the superscripts c and t (as usual) refer to cohort and time (year), and a refers to age, defined here as the age of cohort c in year t. In this particular case, (2.89) can be given a theoretical interpretation, since according to life-cycle theory under certainty, consumption is the product of lifetime wealth, the cohort aggregate of which is constant over time, and an age effect, which is determined by preferences (see Section 6.1 below). In other contexts where there is no such theory, the decomposition is often a useful descriptive device, as for earnings in Taiwan (China), where it is hard to look at the top left-hand panel of Figure 2.5 without thinking about an age and cohort decomposition.

In order to implement a model like (2.89), we need to decide how to label cohorts. A convenient way to do so is to choose c as the age in year $t=0$. By this, c is just a number like a and t. We can then choose to restrict the age, cohort, and year effects in (2.89) in various different ways. In particular, we can choose polynomials or dummies. For the year effects, where there is no obvious pattern *a priori*, dummy variables would seem to be necessary, but age effects could reasonably be modeled as a cubic, quartic, or quintic polynomial in age, and cohort effects, which are likely to be trend-like, might even be adequately handled as linear in c. Given the way in which we have defined cohorts, with bigger values of c corresponding to older cohorts, we would expect γ_c to be declining with c. When data are plentiful, as in the Taiwanese case, there is no reason not to use dummy variables for all three sets of effects, and thus to allow the data to choose any pattern.

Suppose that A is a matrix of age dummies, C a matrix of cohort dummies, and Y a matrix of year dummies. The cohort data are arranged as cohort-year pairs, with each "observation" corresponding to a single cohort in a specific year. If there are m such cohort-year pairs, the three matrices will each have m rows; the number of columns will be the number of ages (or age groups), the number of cohorts, and the number of years, respectively. The model (2.89) can then be written in the form

$$y = \beta + A\alpha + C\gamma + Y\psi + u \tag{2.90}$$

where y is the stacked vector of cohort-year observations—each row corresponds to a single observation on a cohort—on the cohort means of the logarithm of consumption. As usual, we must drop one column from each of the three matrices, since for the full matrices, the sum of the columns is a column of ones, which is already included as the constant term.

However, even having dropped these columns, it is still impossible to estimate (2.90) because there is an additional linear relationship across the three matrices. The problem lies in the fact that if we know the date, and we know when a cohort was born, then we can infer the cohort's age. Indeed, since c is the age of the cohort in year 0, we have

$$a_{ct} = c + t \tag{2.91}$$

which implies that the matrices of dummies satisfy

$$As_a = Cs_c + Ys_y \tag{2.92}$$

where the s vectors are arithmetic sequences $\{0,1,2,3, \ldots, \}$ of the length given by the number of columns of the matrix that premultiplies them. Equation (2.92) is a single identity, so that to estimate the model it is necessary to drop one more column from any one of the three matrices.

The normalization of age, cohort, and year effects has been discussed in different contexts by a number of authors, particularly Hall (1971), who provides an admirably clear account in the context of embodied and disembodied technical progress for different vintages of pickup trucks, and by Weiss and Lillard (1978), who are concerned with age, vintage, and time effects in the earnings of scientists. The treatment here is similar to Hall's, but is based on that given in Deaton and Paxson (1994a). Note first that in (2.90), we can replace the parameter vectors α, γ, and ψ by

$$\tilde{\alpha} = \alpha + \kappa s_a, \quad \tilde{\gamma} = \gamma - \kappa s_c, \quad \tilde{\psi} = \psi - \kappa s_y \tag{2.93}$$

for any scalar constant κ, and by (2.92) there will be no change in the predicted value of y in (2.87). According to (2.90), a time-trend can be added to the age dummies, and the effects offset by subtracting time-trends from the cohort dummies and the year dummies.

Since these transformations are a little hard to visualize, and a good deal more complicated than more familiar dummy-variable normalizations, it is worth considering examples. Suppose first that consumption is constant over cohorts, ages, and years, so that the curves in Figure 2.5 degenerate to a single straight line with slope 0. Then we could "decompose" this into a positive age effect, with consumption growing at (say) five percent for each year of age, and offset this by a negative year effect of five percent a year. According to this, each cohort would get a five percent age bonus each year, but would lose it to a macroeconomic effect whereby everyone gets five percent less than in the previous year. If this

were all, younger cohorts would get less than older cohorts at the same age, because they come along later in time. To offset this, we need to give each cohort five percent more than the cohort born the year previously which, since the older cohorts have higher cohort numbers, means a negative trend in the cohort effects. More realistically, suppose that when we draw Figure 2.5, we find that the consumption of each cohort is growing at three percent a year, and that each successive cohort's profile is three percent higher than that of its predecessor. Everyone gets three percent more a year as they age, and starting consumption rises by three percent a year. This situation can be represented (exactly) by age effects that rise linearly with age added to cohort effects that fall linearly with age by the same amount each year; note that cohorts are labeled by age at a fixed date, so that older cohorts (larger c) are poorer, not richer. But the same data can be represented by a time-trend of three percent a year in the age effects, without either cohort or year effects.

In practice, we choose a normalization that is most suitable for the problem at hand, attributing time-trends to year effects, or to matching age and cohort effects. In the example here, where consumption or earnings is the variable to be decomposed, a simple method of presentation is to attribute growth to age and cohort effects, and to use the year effects to capture cyclical fluctuations or business-cycle effects that average to zero over the long run. A normalization that accomplishes this makes the year effects orthogonal to a time-trend, so that, using the same notation as above,

$$s_y'\psi = 0. \tag{2.94}$$

The simplest way to estimate (2.90) subject to the normalization (2.94) is to regress y on (a) dummies for each cohort excluding (say) the first, (b) dummies for each age excluding the first, and (c) a set of $T-2$ year dummies defined as follows, from $t = 3,..,T$

$$d_t^* = d_t - [(t-1)d_2 - (t-2)d_1] \tag{2.95}$$

where d_t is the usual year dummy, equal to 1 if the year is t and 0 otherwise. This procedure enforces the restriction (2.94) as well as the restriction that the year dummies add to zero. The coefficients of the d_i^* give the third through final year coefficients; the first and second can be recovered from the fact that all year effects add to zero and satisfy (2.94).

This procedure is dangerous when there are few surveys, where it is difficult to separate trends from transitory shocks. In the extreme case where there are only two years, the method would attribute any increase in consumption between the first and second years to an increasing age profile combined with growth from older to younger cohorts. Only when there are sufficient years for trend and cycle to be separated can we make the decomposition with any confidence.

The three remaining panels of Figure 2.5 show the decomposition of the earnings averages into age, cohort, and year dummies. The cohort effects in the top

right-hand panel are declining with age; the earlier you are born, the older you are in 1976, and age in 1976 is the cohort measure. Although the picture is one that is close to steady growth from cohort to cohort, there has been a perceptible acceleration in the rate of growth for the younger cohorts. The bottom left-hand panel shows the estimated age effects; according to this, wages are a concave function of age, and although there is little wage increase after age 50, there is no clear turning down of the profile. Although the top left panel creates an impression of a hump-shaped age profile of earnings, much of the impression comes from the cohort effects, not the age effects, and although the oldest cohort shown has declining wages from ages 58 through 65, other cohorts observed at the same ages do not display the same pattern. (Note that only every fifth cohort is included in the top left panel, but all cohorts are included in the regressions, subject only to age lying being between 25 and 65 inclusive.) The final panel shows the year effects, which are estimated to be much smaller in magnitude than either the cohort or age effects; nevertheless they show a distinctive pattern, with the economy growing much faster than trend at the beginning and end of the period, and much more slowly in the middle after the 1979 oil shock.

Age and cohort profiles such as those in Figure 2.5 provide the material for examining the structural consequences of changes in the rates of growth of population and real income. For example, if the age profiles of consumption and income are determined by tastes and technology, and are invariant to changes in the rate of economic growth, we can change the cohort effects holding the age effects constant and thus derive the effects of growth on aggregates of consumption, saving, and income. Changes in population growth rates redistribute the population over the various ages, so that, once again, we can use the age profiles as the basis for aggregating over different age distributions of the population. Much previous work has been forced to rely on single cross sections to estimate age profiles, and while this is sometimes the best that can be done, cross-sectional age profiles confuse the cohort and age effects, and will typically give much less reliable estimates than the methods discussed in this section. I return to these techniques in the final chapter when I come to examine household saving behavior.

2.8 Two issues in statistical inference

This final section deals briefly with two topics that will be required at various points in the rest of the book, but which do not fit easily into the rest of this chapter. The first deals with a situation that often arises in practice, when the parameters of interest are not the parameters that are estimated, but functions of them. I briefly explain the "delta" method which allows us to transform the variance-covariance matrix of the estimated parameters into the variance-covariance matrix of the parameters of interest, so that we can construct hypothesis tests for the latter. Even when we want to use the bootstrap to generate confidence intervals, asymptotic approximation to variances are useful starting points that can be improved using the bootstrap (see Section 1.4). The second topic is concerned with sample size, and its effects on statistical inference. Applied econometricians often

express the view that rejecting a hypothesis using 100 observations does not have the same meaning as rejecting a hypothesis using 10,000 observations, and that null hypotheses are more often rejected the larger is the sample size. Household surveys vary in size from a few hundred to tens or even hundreds of thousands of observations, so that if inference is indeed the hostage of sample size, it is important to be aware of exactly what is going on, and how to deal with it in practice.

**Parameter transformations: the delta method*

Suppose that we have estimates of a parameter vector β, but that the parameters of interest are not β, but some possibly nonlinear transformation α, where

$$\alpha = h(\beta) \tag{2.96}$$

for some known vector of differentiable functions h. In general, this function will also depend on the data, or on some characteristics of the data such as sample means. It will also usually be the case that α and β will have different numbers of elements, k for β and q for α, with $q \leq k$. Our estimation method has yielded an estimate $\hat{\beta}$ for β and an associated variance-covariance matrix V_β for which an estimate is also available. The delta method is a means of transforming V_β into V_α; a good formal account is contained in Fuller (1987, pp. 85–88). Here I confine myself to a simple intuitive outline.

Start by substituting the estimate of β to obtain the obvious estimate of α, $\hat{\alpha} = h(\hat{\beta})$. If we then take a Taylor series approximation of $\hat{\alpha} = h(\hat{\beta})$ around the true value of β, we have for $i = 1 \ldots q$,

$$\hat{\alpha}_i \simeq \alpha_i + \sum_{j=1}^{k} \frac{\partial h_i}{\partial \beta_j} (\hat{\beta}_j - \beta_j) \tag{2.97}$$

or in an obvious matrix notation

$$\hat{\alpha} - \alpha \simeq H(\hat{\beta} - \beta). \tag{2.98}$$

The matrix H is the $q \times k$ Jacobian matrix of the transformation. If we then post-multiply (2.98) by its transpose and take expectations, we have

$$V_\alpha \simeq H V_\beta H'. \tag{2.99}$$

In practice (2.99) is evaluated by replacing the three terms on the right-hand side by their estimates calculated from the estimated parameters. The estimate of the matrix H can either be programmed directly once the differentiation has been done analytically, or the computer can be left to do it, either using the analytical differentiation software that is increasingly incorporated into some econometric packages, or by numerical differentiation around the estimates of β.

Variance-covariance matrices from the delta method are often employed to calculate Wald test statistics for hypotheses that place nonlinear restrictions on the

parameters. The procedure follows immediately from the analysis above by writing the null hypothesis in the form:

$$H_0: \alpha = h(\beta) = 0 \tag{2.100}$$

for which we can compute the Wald statistic

$$W = \hat{\alpha}' V_\alpha^{-1} \hat{\alpha}. \tag{2.101}$$

Under the null hypothesis, W is asymptotically distributed as χ^2 with q degrees of freedom. For this to work, the matrix V_α has to be nonsingular, for which a necessary condition is that q be no larger than k; clearly we must not try to test the same restriction more than once.

As usual, some warnings are in order. These results are valid only as large-sample approximations, and may be seriously misleading in finite samples. For example, the ratio of two normally distributed variables has a Cauchy distribution which does not possess any moments, yet the delta method will routinely provide a "variance" for this case. In the context of the Wald tests of nonlinear restrictions, there are typically many different ways of writing the restrictions, and unless the sample size is large and the hypothesis correct, these will all lead to different values of the Wald test (see Gregory and Veall 1985 and Davidson and MacKinnon 1993, pp. 463–71, for further discussion).

Sample size and hypothesis tests

Consider the frequently encountered situation where we wish to test a simple null hypothesis against a compound alternative, that $\beta = \beta_0$ for some known β_0 against the alternative $\beta \neq \beta_0$. A typical method for conducting such a test would be to calculate some statistic from the data and to see how far it is from the value that it would assume under the null, with the size of the discrepancy acting as evidence against the null hypothesis. Most obviously, we might estimate β itself without imposing the restriction, and compare its value with β_0. Likelihood-ratio tests—or other measures based on fit—compare how well the model fits the data at unrestricted and restricted estimates of β. Score—or Lagrange multiplier—tests calculate the derivative of the criterion function at β_0, on the grounds that non-zero values indicate that there are better-fitting alternatives nearby, so casting doubt on the null. All of these supply a measure of the failure of the null, and our acceptance and rejection of the hypothesis can be based on how big is the measure.

The real differences between different methods of hypothesis testing come, not in the selection of the measure, but in the setting of a critical value, above which we reject the hypothesis on the grounds that there is too much evidence against it, and below which we accept it, on the grounds that the evidence is not strong enough to reject. Classical statistical procedures—which dominate econometric practice—set the critical value in such a way that the probability of reject-

ing the null when it is correct, the probability of Type I error, or the *size* of the test, is fixed at some preassigned level, for example, five or one percent. In the ideal situation, it is possible under the null hypothesis to derive the sampling distribution of the quantity that is being used as evidence against the null, so that critical values can be calculated that will lead to exactly five (one) percent of rejections when the null is true. Even when this cannot be done, the asymptotic distribution of the test statistic can usually be derived, and if this is used to select critical values, the null will be rejected five percent of the time when the sample size is sufficiently large. These procedures take no explicit account of the *power* of the test, the probability that the null hypothesis will be rejected when it is false, or its complement, the Type II error, the probability of *not* rejecting the null when it is false. Indeed, it is hard to see how these errors can be controlled because the power depends on the unknown true values of the parameter, and tests will typically be more powerful the further is the truth from the null.

That classical procedures can generate uncomfortable results as the sample size increases is something that is often expressed informally by practitioners, and the phenomenon has been given an excellent treatment by Leamer (1978, pp. 100– 120), and it is on his discussion that the following is based.

The effect most noted by empirical researchers is that the null hypothesis seems to be more frequently rejected in large samples than in small. Since it is hard to believe that the truth depends on the sample size, something else must be going on. If the critical values are exact, and if the null hypothesis is exactly true, then by construction the null hypothesis will be rejected the same fraction of times in all sample sizes; there is nothing wrong with the logic of the classical tests. But consider what happens when the null is not exactly true, or alternatively, that what we mean when we say that the null is true is that the parameters are "close" to the null, "close" referring to some economic or substantive meaning that is not formally incorporated into the statistical procedure. As the sample size increases, and provided we are using a consistent estimation procedure, our estimates will be closer and closer to the truth, and less dispersed around it, so that discrepancies that were undetectable with small sample sizes will lead to rejections in large samples. Larger sample sizes are like greater resolving power on a telescope; features that are not visible from a distance become more and more sharply delineated as the magnification is turned up.

Over-rejection in large samples can also be thought about in terms of Type I and Type II errors. When we hold Type I error fixed and increase the sample size, all the benefits of increased precision are implicitly devoted to the reduction of Type II error. If there are equal probabilities of rejecting the null when it is true and not rejecting it when it is false at a sample size of 100, say, then at 10,000, we will have essentially no chance of accepting it when it is false, even though we are still rejecting it five percent of the time when it is true. For economists, who are used to making tradeoffs and allocating resources efficiently, this is a very strange thing to do. As Leamer points out, the standard defense of the fixed size for classical tests is to protect the null, controlling the probability of rejecting it when it is true. But such a defense is clearly inconsistent with a procedure that

devotes none of the benefit of increased sample size to lowering the probability that it will be so rejected.

Repairing these difficulties requires that the critical values of test statistics be raised with the sample size, so that the benefits of increased precision are more equally allocated between reduction in Type I and Type II errors. That said, it is a good deal more difficult to decide exactly how to do so, and to derive the rule from basic principles. Since classical procedures cannot provide such a basis, Bayesian alternatives are the obvious place to look. Bayesian hypothesis testing is based on the comparison of posterior probabilities, and so does not suffer from the fundamental asymmetry between null and alternative that is the source of the difficulty in classical tests. Nevertheless, there are difficulties with the Bayesian methods too, perhaps most seriously the fact that the ratio of posterior probabilities of two hypotheses is affected by their prior probabilities, no matter what the sample size. Nevertheless, the Bayesian approach has produced a number of procedures that seem attractive in practice, several of which are reviewed by Leamer.

It is beyond the scope of this section to discuss the Bayesian testing procedures in any detail. However, one of Leamer's suggestions, independently proposed by Schwarz (1978) in a slightly different form, and whose derivation is also insightfully discussed by Chow (1983, pp. 300–2), is to adjust the critical values for F and χ^2 tests. Instead of using the standard tabulated values, the null is rejected when the calculated F-value exceeds the logarithm of the sample size, $\ln n$, or when a χ^2 statistic for q restrictions exceeds $q \ln n$. To illustrate, when the sample size is 100, the null hypothesis would be rejected only if calculated F-statistics are larger than 4.6, a value that would be doubled to 9.2 when working with sample sizes of 10,000.

In my own work, some of which is discussed in the subsequent chapters of this book, I have often found these Leamer-Schwarz critical values to be useful. This is especially true in those cases where the theory applies most closely, when we are trying to choose between a restricted and unrestricted model, and when we have no particular predisposition either way except perhaps simplicity, and we want to know whether it is safe to work with the simpler restricted model. If the Leamer- Schwarz criterion is too large, experience suggests that such simplifications are indeed dangerous, something that is not true for classical tests, where large-sample rejections can often be ignored with impunity.

2.9 Guide to further reading

The aim of this chapter has been to extract from the recent theoretical and applied econometric literature material that is useful for the analysis of household-level data. The source of the material was referenced as it was introduced, and in most cases, there is little to consult apart from these original papers. I have assumed that the reader has a good working knowledge of econometrics at the level of an advanced undergraduate, masters', or first-year graduate course in econometrics covering material such as that presented in Pindyck and Rubinfeld (1991). At the same level, the text by Johnston and DiNardo (1996) is also an excellent starting

point and, on many topics, adopts an approach that is sympathetic to that taken here. A more advanced text that covers a good deal of the modern theoretical material is Davidson and MacKinnon (1993), but like other texts it is not written from an applied perspective. Cramer (1969), although now dated, is one of the few genuine applied econometrics texts, and contains a great deal that is still worth reading, much of it concerned with the analysis of survey data. Some of the material on clustering is discussed in Chapter 2 of Skinner, Holt, and Smith (1989). Groves (1989, ch. 6) contains an excellent discussion of weighting in the context of modeling versus description. The STATA manuals, Stata Corporation (1993), are in many cases well ahead of the textbooks, and provide brief discussions and references on each of the topics with which they deal.