Chapter 33

DATA AND ECONOMETRIC TOOLS FOR DEVELOPMENT ANALYSIS

ANGUS DEATON

Princeton University

Contents

Introduction 1786
1. Data for development economics 1787
   1.1. Household survey data 1787
   1.2. Panel data 1801
   1.3. National income and other data 1807
2. Econometric tools for development analysis 1815
   2.1. Econometric analysis of survey data 1815
   2.2. Econometric issues in time-series 1841
   2.3. Introduction to non-parametric methods 1861
References 1874

*I am grateful to Tim Besley, Jere Behrman, Anne Case, David Card, Alan Krueger, Ron Miller, Christina Paxson, James Powell and T.N. Srinivasan for help in the preparation of this paper.

Handbook of Development Economics, Volume III, Edited by J. Behrman and T.N. Srinivasan

Introduction

Almost all the tools of econometric analysis that have been used in empirical work in economics in general have also been applied to the specific problems of economic development. My choice of topics in this review is therefore to some extent arbitrary, although I have been guided by three considerations. First, there are a number of methods that are so widely encountered in the development literature that they demand some coverage, even when their use raises no issues beyond those encountered in a good standard text. Second, I have been led by the first section of the chapter, on data questions, since discussion of data frequently leads naturally into econometric technique. Third, I have followed much of the recent general econometric literature in emphasizing robustness of inference and estimation. Much recent theoretical work has been devoted to methods that allow applied workers to dispense with unnecessary supporting assumptions, so that, for example, standard errors can be calculated even when standard textbook assumption fail, simultaneity and selection bias can be allowed for without making arbitrary and often incredible assumptions, and key effects can be measured with minimal assumptions about functional form. In keeping with the balance of the development literature, I have chosen to emphasize microeconomic more than macroeconomic applications, although I have included a section on recent results in time-series analysis and their application to problems of economic development.

The plan of the chapter is as follows. Section 1 is concerned with data, and Section 2 with tools and their application. However, since many of the data issues lead directly into the econometrics, I have sometimes found it convenient not to make the separation, and to include both in Section 1. The first and largest part of Section 1 deals with household survey data, with survey design in developing countries, with data collection, with measurement issues, and with the experience of using such data in econometric analysis. Section 1.2 is concerned with national income accounts, and with the index number and other problems that underlie international comparisons of income levels and growth rates. I also give some attention to the quality of country data, looking beyond national incomes to demographic, trade, and other measures.

Section 2 turns to econometric tools. There are three main sections. Section 2.1 is mainly concerned with tools for microeconomic analysis, emphasizing the use of survey data. I work through a range of more or less familiar econometric topics, illustrating their uses in the development literature, discussing methods for strengthening the robustness of inference, and trying to identify common pitfalls and difficulties. Section 2.2 turns to time-series techniques and their uses in the analysis of development questions. The modern time-series literature is very large and is rapidly growing, and I cover only a small selection of the many possible topics. Section 2.3 provides an introduction to non-parametric techniques for estimating density functions, regression functions, and the derivatives of regression functions. Although non-parametric analysis typically requires a great deal of data, there are a number of questions in development economics that are susceptible to a non-parametric treatment using survey data.

1. Data for development economics

1.1. Household survey data

1.1.1. Content and purpose

There are few Less Developed Countries (LDCs) that have not collected survey data of some sort at some time, and many LDCs have multiple surveys that are run on a regular and continuing basis, many of which meet the highest international standards of data collection, editing, and publication of results. Many (perhaps most) of these surveys have a specific *raison d'etre*; household expenditure surveys are used to monitor living standards or to collect weights for the consumer price index, labor force surveys are used to estimate unemployment rates, and censuses to estimate total population. Other surveys, such as surveys of firms or of farms are used to collect production or output data, and use a unit of observation other than a household. In many countries, the statutory authority establishing each survey is explicit about its purpose, and official statisticians design the surveys with these aims in mind. Of course, once the data are collected they can be used for many other purposes, to which they may be more or less suited, and to which government statistical offices may be more or less sympathetic. In the last ten to fifteen years there has been a great expansion in the use of survey data in development economics – as in other branches of the profession – much of it a consequence of better computing facilities, and much of it attributable to the increased willingness of statistical offices around the world to release their data to researchers. Ministers and civil servants are realizing that they have relatively little to fear from econometric analysis, and perhaps something to learn.

The difference between the original statutory purposes of the surveys and the uses to which the data are put in development economics poses a number of problems. In the short run, there are various statistical issues associated with using data for purposes that are different from the original intent and design, and in the longer run, there is the more fundamental (and much more difficult) question of how surveys ought to be redesigned for the broader policy and analytical purposes for which they are increasingly being used. I shall have something to say on both of these topics.

1.1.2. Survey data in policy and development

Why should development economists be interested in household survey data? If the ultimate aim of economic activity is the welfare of individuals, then the data from household surveys are the measure of its success. Although GDP and GNP per capita are often used as summary measures of welfare, in many countries they are derived with the help of household survey data, and even when this is not the case and consumption is derived as a residual, survey data provide a cross-check, and in many cases will provide higher quality data. But even at their best, national income measures can tell us only a very limited amount about distributional issues, about allocation by region, by ethnic group, by poor versus rich, or by rural versus urban. As economic development expands opportunities, we want to know who is benefiting, and who (if anyone) is losing. Indeed, as Stigler (1954) has documented, the first explorations of household budgets were carried out by social activists in the late eighteenth and early nineteenth centuries, and their object was to inform (and shock) policy makers and to lay the basis for reform. Counting the poor, documenting their living-standards (including nutritional standards), and measuring inequality remain important uses of household survey data by development economists.

Household survey data also yield direct measures of the effects of policy changes, whether these operate through price changes or through changes in the provision of public services. They can therefore provide the background information for informed discussion of possible changes in policy. In particular, quantities produced and consumed provide a local approximation to the derivative of welfare with respect to price. To see this in an example, suppose that a farm (or non-farm) household faces output prices p (labor is an output) and input prices v, and receives off-farm income y, that its technology can be represented by the (restricted) profit function $\pi(p, v; a)$ where a is a vector of fixed factors such as land, and that its preferences can be represented by the expenditure or cost function $c(u, p)$ for utility u, since without loss of generality, all consumption goods can be taken to be outputs. Then, since utility must be financed from farm profits or other income, we have

$$c(u, p) = y + \pi(p, v; z) . \tag{1}$$

Equation (1) immediately tells us by how much income y would have to be increased to compensate the household for a change in a price of one of its inputs or outputs. Since the partial derivatives of the cost function with respect to elements of p are the quantities consumed, q, and since the derivatives of the profit function with respect to output and input prices are quantities of outputs, z, and (minus) inputs r, respectively, total differentiation of (1) gives

$$\mathrm{d}y = \sum_i (q_i - z_i)\,\mathrm{d}p_i + \sum_i r_i\,\mathrm{d}v_i \,. \tag{2}$$

This is the familiar result, that those who are net producers benefit from a price change, and that those who are net consumers lose, and that, to a first approximation, the *amount* of the money-equivalent benefit (or loss) is the net amount produced (production less consumption) multiplied by the price change. Hence, the survey data not only identify the gainers and losers of a price change, but also quantify the sizes of their gains and losses. All this is obvious enough, but is nevertheless important. In many LDCs, where tax and welfare instruments are limited in number, there is a wide range of commodity taxes and subsidies. Many of these are justified on distributional grounds; imported consumer goods should be taxed because only the rich use them, or bus services should be subsidized to support the poor. By looking at (2) for different households, survey data can be used to check whether such claims are in fact correct, or whether they are simply a cover for special interests.

Provided that we accept the underlying economic assumptions of atomistic maximizing agents in competitive markets with minimal uncertainty, the evaluation of (2) requires only the raw data; no econometric model is required. Of course, there are different ways of presenting the results, and I shall give examples in Section 2.3 below of how non-parametric techniques can be used to illustrate the distributional issues in an immediately assimilable form. Note too that the basic result can be extended in various directions. In particular, (2) is a local approximation and so cannot safely be used except for small price changes. For large changes, a better approximation can be made by including substitution effects, effects that in some circumstances can also be estimated from the survey data, a topic to which I return in Section 2.1.

Note what happens when the policy involves a quantity change rather than a price change, as when additional health, education, or agricultural extension services are provided. If these publicly provided quantities are incorporated into the cost or profit functions, and a compensation is calculated as in (2), the result involves the shadow prices of the public goods, prices that can often be estimated using appropriate behavioral models, [see for example Gertler and van der Gaag (1990) and the studies reviewed in Jimenez (1987)]. However, even without such calculations, the survey data frequently tell us who uses the public goods, and by how much, something that is frequently of direct concern, even where we do not have estimates of how much the households value the services.

Beyond the direct use of survey data for policy, household surveys provide much of the raw material for modeling and trying to understand household behavior. I shall discuss a number of such studies and their results as I illustrate the various techniques.

1.1.3. Survey design and its implications for analysis

I shall use the "typical" household income and expenditure survey as my example, but many of the arguments can be applied to other types of surveys too. Such surveys typically collect data on a household basis – a household usually being defined as a group of people who share the same "cooking-pot" – and ask how much was spent over some reference period on a lengthy list of consumption items; the reference period can be anything from a day to a year, and sometimes varies by category of expenditure, with shorter recall periods for high frequency items like food, and longer periods for unusual purchases, like clothes or durable goods. In countries with near universal literacy, households can be asked to keep diaries; otherwise enumerators verbally ask respondents to recall individual purchases. Data are also collected on the respondents, at the very least covering the numbers, sexes, and ages, or people in the household. This can be extended to a range of household characteristics, such as education, occupation, and race. Data are frequently also collected on quantities consumed as well as expenditures, at least for readily measurable goods such as food. There will also be data on location, and perhaps more if the enumerators collect and retain data on the environment, for example on the size of the village, whether it has a school, and so on. Such surveys are sometimes carried out on an annual basis [Taiwan, Korea, India until 1973–1974 and since 1991], but more usually are done at intervals, often quinquennially, on the grounds that consumption patterns and levels of living and poverty do not change very quickly. The surveys are typically nationally representative, with each remaining in the field for a year, although there are also many special purpose surveys that are restricted in geographical coverage, and which last for a period shorter than a year.

Households are chosen at random, but there is a wide range of designs. The simplest is where each household in the country has an equal chance of being selected, but such simplicity is uncommon, if only because there are very different costs of obtaining data from different types of households. Rural households are more widely scattered than urban households, and in many LDCs, there are some households that live in inaccessible (sometimes even dangerous) areas. Any sample design that minimizes cost for a given degree of precision (or equivalently maximizes precision at given cost) will therefore lead to oversampling of urban and under sampling of rural households. Beyond this, interview procedures usually require more than one visit to each household. For example if a diary is kept, there will be an initial visit, a second visit to check that it is being kept correctly, and a collection visit after seven or fourteen days. In rural areas, where transport is a major cost of the survey, it therefore makes sense for the survey to group households into survey "clusters", often villages, with typically six to fifteen households in each cluster. The optimal number of households in the cluster involves a trade-off between the low marginal cost of drawing another household in a village that is already being visited, and the relatively low contribution to precision of such a household, given that it is likely to look rather similar to other households in the same village. The survey team remains in the village for a week or two, surveying all households in the cluster, and then moves on to a new cluster. Such surveys frequently attempt to give each household an equal chance of inclusion by using a two stage design, in which clusters are selected first, with a probability of inclusion proportional to size (i.e. the number of households in the cluster), while individual households are randomly selected at the second stage. The random selection of clusters and households is made from a "sampling frame", often a census. However, censuses are often badly outdated, and in some countries are not reliable, either because of political interference – census returns are typically used to make voter rolls and sometimes to allocate resources – or because of difficulties of collection. Problems with censuses can be avoided as in India, where the National Sample Survey (NSS) selects villages from a village frame, and then lists all households in the village at a preliminary stage. The final drawing selects a stratified random sample from their own list, with stratification based on a few variables collected at the list stage, [see Murthy (1977, Chapter 15)] who also describes many of the other features of the design of the NSS, or (more generally) Casley and Lury (1981) for further description.

The relatively simple – and sensible – designs of the previous paragraphs can be complicated ad infinitum. Adjustments can be made for non-response, and for the consequences of replacing non-responders by "look-alike" households, although it should be noted that, unless households are approached at obviously inappropriate times, like harvests, non-response is typically much less of a problem in LDCs than in the United States (US). Probabilities of selection can also be linked to any ancillary information in the sampling frame or listing, such as occupation, housing status, or landholdings. As a result, survey tapes usually report for each household a sampling probability, or its reciprocal, an "inflation factor", which is the number of households in the country for which the household stands proxy. In complex designs, the inflation factors will be different for every household in the survey. Although designs are often "self-weighting", whereby in spite of the many strata and levels of strata each household has an equal chance of being included, refusals or other practical problems often frustrate the intention, and inflation factors remain relevant.

Many survey statisticians, in the US as well as in LDCs, see their role as producing an optimum design that will estimate the target magnitudes – for example the weights for the price index – in a way that trades off precision against cost. From such a perspective, any variable that is observed prior to the survey and is correlated with the target magnitude is a potential candidate for stratification. However, the more complex the design, the greater are the difficulties of using the data for anything other than the original purpose. Households may be stratified by variables that are endogenous to the processes that economists want to model, and even when this is not the case, the fact that samples are not simple random samples raises questions about the extent to which econometric results can be regarded as nationally representative. If it were to be widely recognized that household surveys have a wide range of important uses, then it would also be recognized that complex designs are dysfunctional, with sometimes quite small gains in precision obtained at the price of large compromises in the usefulness of the surveys. While there exist econometric techniques to correct samples for selectivity, as in Heckman (1976) or Manski and Lerman (1977), it is much better not to have to use them, see also the discussion in Section 2.1 below.

Given that development economists only rarely have control over survey protocols, there are a number of implications of design that should be born in mind when using survey data in econometric applications. I focus on three of the most important: the definition of the household, the measurement of dispersion, and the effects of designs other than simple random surveys.

1.1.4. The definition of the household

In many societies, people do not live in households that resemble the typical nuclear families of the US or Europe. Extended families, or members of a common lineage, may live in close proximity to one another, and only sometimes share the same cooking pot. The closeness of the group may vary with economic circumstances, with subunits becoming independent in good times and reuniting in adversity, see for example Ainsworth (1992) on fostering in West Africa. Even when there are separate households undertaking separate economic activities, assets may be held at an extended family level, as with the *chia* in Taiwan, [see Liu (1982) or Greenhalgh (1982)]. In many surveys, the decision whether to count multiple units as one or many households is essentially arbitrary, and in Thailand, a change from one to the other between the 1975–1976 and 1980–1981 surveys caused average household size to drop by about one person per household, [Government of Thailand (1977, 1983)]. Even for identical populations, the survey that distinguishes more households will show higher inequality and higher poverty, since combining households and assuming that each member has the same income or consumption amounts to a mean-preserving reduction in spread, see Haddad and Kanbur (1990). Since most surveys retain the same practices over time, trends in inequality and poverty are unlikely to be misleading, at least for these reasons, but the absolute levels will be incorrect, and international comparisons will be compromised.

1.1.5. Measuring means versus measuring dispersion

Surveys are usually designed to measure *means*, not dispersion, and there is a wide variety of designs that will measure means accurately, but will give very poor estimates of inequality, of poverty, or of other quantities that depend upon higher moments. Consider the measurement of income in an agricultural society where, to take an extreme case, all agricultural income is received in the month of the harvest. A design in which one twelfth of the sample is interviewed in each month and asked to report the previous month's income will generate an estimate of average income that is unbiased. But even if every household has the same annual income, the survey will appear to show that 100 percent of income is concentrated in the hands of 7.5 percent of households, and that 92.5 percent of households are "absolutely poor". Some surveys avoid these problems, at least in part, by revisiting households on a seasonal basis, but most do not. Once again, international comparisons of inequality and poverty are rendered meaningless, and in predominately agricultural societies with variable and weather-affected harvests, there will even be spurious shifts in apparent dispersion between different surveys in the same country, so that even the time path of inequality can be obscured.

The variability of income is one reason why many analysts prefer to use consumption as a basis for measurement. But consumption is not immune to the problem. Different surveys use different reporting periods, from a day to a year. Some purchases are made infrequently, and households stock up when they shop, so that the shorter the reporting period, the larger will be the apparent dispersion. For example, suppose that everyone has consumption c but that purchases are random, with a fraction p of households buying cp^{-1} during the reporting period, and the rest buying nothing. Simple calculation gives:

$$E(x) = p.cp^{-1} + (1-p).0 = c; \quad V(x) = c^2(1-p)p^{-1} \tag{3}$$

where x is reported expenditure. As the reporting period becomes shorter, p will get smaller, and although the mean is unaffected, the variance will rise. There is no point in comparing distributions of expenditures from different surveys unless we know that the reporting periods are the same. Exactly the same point arises if we attempt to compare two countries one of which has a perishable staple that is bought frequently, while the other uses a storable staple that is bought rarely. Problems over reporting periods and over the definition of the household are two (of the many) reasons why we know so little about international comparisons of inequality and poverty; for others [see Berry (1985) and Fields (1992)].

1.1.6. Estimation of means in stratified samples

When different households have different probabilities of being included in the survey, unweighted sample means will generally be biased for the population means. Consider the simplest example where there are two sectors, sector 1, "urban" and sector 2, "rural", and where households in each are sampled with probabilities π_1 and π_2. We are interested in the random variable x, which is distributed in the populations of the two sectors with means μ_1 and μ_2. There are n observations in total, n_1 urban households and $n_2 = n - n_1$ rural households; these correspond to population figures of N, N_1 and N_2, so that $\pi_s = n_s N_s^{-1}$, $s = 1, 2$. The sample mean is

$$\bar{x} = (n_1 + n_2)^{-1} \sum_i x_i \tag{4}$$

with expectation

$$E(\bar{x}) = \frac{n_1}{n} \mu_1 + \frac{n_2}{n} \mu_2 \, . \tag{5}$$

The population mean, by contrast, is given by

$$\mu = \frac{N_1}{N} \mu_1 + \frac{N_2}{N} \mu_2 \tag{6}$$

so that the sample mean is biased unless either $\pi_1 = \pi_2$, in which case the sample is a simple random sample, or $\mu_1 = \mu_2$, so that the population is homogeneous, at least as far as the parameter of interest is concerned.

Neither of these requirements would usually be met in practice; for example, rural households are likely to be both poorer and costlier to sample. To get the right answer, we do the obvious thing, and compute a weighted mean. This can be done by defining "inflation factors" for each observation, equal to the reciprocals of the sampling probabilities, so that here

$$\theta_i = \pi_s^{-1}, \quad i \in s, s = 1, 2 \, . \tag{7}$$

Note that if we multiply each sample observation by its inflation factor and add, we obtain an unbiased estimate of the population *total*, something that is often of separate interest. However, if the inflation factors are scaled by their total to derive sampling weights $w_i = \theta_i / \Sigma\, \theta_k$, and we calculate a weighted mean, when we take expectations we get

$$E\Big(\sum w_i x_i\Big) = \frac{n_1 \pi_1^{-1} \mu_1 + n_2 \pi_2^{-1} \mu_2}{n_1 \pi_1^{-1} + n_2 \pi_2^{-1}} = \frac{N_1 \mu_1 + N_2 \mu_2}{N_1 + N_2} = \mu \ . \tag{8}$$

which is the right answer. Similar weighting schemes can be applied to the estimation of any other population statistic that can be written as an average, including variances, quantiles, measures of inequality and of poverty. The simple idea to remember is that each household should be inflated to take account of the households that it represents but were not sampled, so as to make the inflated sample "as like" the population as possible.

While the underlying population in these exercises is finite (it is the population of all households in the country at the time of the survey), and although much of the inference in the sampling literature is conducted explicitly from such a perspective, so that expectations are taken over all the possible samples that can be drawn from the finite population, this is not the *only* framework for inference. In particular, the finite population can be regarded as itself being a "sample" from a "superpopulation" of similar households, households that might have existed or might exist in the future. In this way, the parameter μ (for example) is not the mean characteristic for the current population, but a parameter that characterizes the distributional law by which that population was generated. In this way, the superpopulation approach brings survey-sampling theory much closer to the usual sampling theory in econometric analysis where we are usually making inferences about behavioral parameters, not characteristics of finite populations.

1.1.7. Econometric estimation in stratified samples

All this is so familiar and so natural that it seems hardly worth the exposition. However, the simple weighting of observations is less obviously appropriate once we move from the estimation of means to even the simplest of econometric estimates, including ordinary least squares regression. Again, consider the simplest possible case, where there exists a linear relationship between y and x, but with coefficients β_1 and β_2 that differ by sector. Assuming zero means for both variables, write this

$$y_i = x_i \beta_s + u_i, \quad s = 1, 2 \ . \tag{9}$$

Suppose that the parameter of interest is β, the population-weighted average of β_1 and β_2; given the two coefficients, this could be obtained by weighting each by the inflation factor for its sector. For example, if β_i is the marginal propensity to consume in each sector, the population-weighted average β would be the marginal propensity to consume out of a randomly allocated unit of currency, a quantity that is often of interest in discussions of tax and benefit reform.

As with the case of estimating the population mean, it is immediately clear that the (unweighted) OLS estimator using all of the data is biased and inconsistent. Instead, we might follow the principle of the previous subsection, weighting each household by the number of households that it represents in the survey, and compute the weighted estimator

$$\tilde{\beta} = \Big(\sum_i w_i x_i^2\Big)^{-1}\Big(\sum_i w_i x_i y_i\Big) \tag{10}$$

where w_i is the normalized inflation factor. This estimator converges, not to β, but to

$$\operatorname{plim} \tilde{\beta} = \frac{\beta + N_2 N^{-1}\beta_2(m_2 - m_1)m_1^{-1}}{1 + N_2 N^{-1}(m_2 - m_1)m_1^{-1}} \tag{11}$$

where m_1 and m_2 are the (population) variances of x in each of the two sectors. Unlike the unweighted estimator, this quantity at least has the (limited) virtue of being independent of sample design; indeed, as is to be expected from the general argument for inflation factors, it is what OLS would give if applied to the data from the whole population [see Dumouchel and Duncan (1983)]. However, it is not equal to the parameter of interest β unless either $\beta_1 = \beta_2$, or $m_1 = m_2$; the former is ruled out by hypothesis, and there is no reason to suppose that the latter will hold in general.

Of course, the fundamental issue here is not the sample design but the fact that the regression is not homogeneous within the population being studied. As such, the problem is not a sampling issue – exactly the same issues arise in regressions using pooled time-series for a cross-section of countries – but a heterogeneity issue, and it comes to the fore in the sampling context because it is heterogeneity that justifies the stratification in the first place. As a result, it is often plausible that behavioral parameters will differ across strata, just as they are likely to vary across countries. When this is not the case, and regression coefficients are identical, then both weighted and unweighted regressions are unbiased and consistent, and the Gauss-Markov theorem tells us that the *unweighted* regression is to be preferred. If instead the regression coefficients differ by strata, that fact has to be explicitly faced and cannot be finessed by running regressions weighted by inflation factors. Such recommendations were once standard in econometric texts – [see for example Cramer (1969, pp. 142–143)] – but even so, many regressions using survey data are run in weighted form.

In cases where heterogeneity is suspected, there are several useful strategies. When there are only a few strata – rural versus urban would be the most frequent – it clearly makes sense to run separate regressions, and to use covariance analysis where the homogeneity hypothesis is of separate interest. When the number of strata is large, with relatively few observations in each, random coefficient specifications would seem more useful, and, as a result, analysts should routinely expect heteroskedasticity in OLS regressions. Standard heteroskedasticity tests can be used, for example that given by Breusch and Pagan (1979), which in this case would involve regressing squared residuals on dummy variables for each stratum and comparing half the resulting explained sum of squares with a χ^2 with degrees of freedom equal to the number of strata. Heteroskedastic consistent variance covariance matrices should also be routinely used, see Section 2.1 below.

I should conclude by noting that there is a school of thought that does not accept the argument against weighted regressions, Kish and Frankel (1974) being perhaps the most eloquent example. They argue that the stratification in many surveys is not of substantive interest in its own right, and that the parameters of a hypothetical census regression are indeed of interest. Others, such as Pfefferman and Smith (1985) take a view similar to that here, arguing (among other things) that a complete population is of no great interest since it is only one of the many possible populations with which we might have been confronted.

1.1.8. Estimation and other design features: clustering

Even if the regression coefficients are homogeneous across strata, standard formulae for standard errors may be incorrect depending on the survey design. Two-stage sampling will induce non-independence between households in the same cluster if households who live in the same village are subject to common unobservables, such as weather, tastes, or prices. Under such circumstances, whether we are estimating means or regressions, standard formulae for variances are incorrect and can be seriously misleading.

Consider first the straightforward use of survey data to estimate a mean. Given a set of n observations x_i, standard procedures call for the estimation of the mean and variance according to

$$\hat{\mu} = n^{-1} \sum_{1}^{n} x_j; \;\; \hat{\sigma}^2 = (n-1)^{-1} \sum_{1}^{n} (x_1 - \bar{x})^2 \,. \tag{12}$$

If the observations are independent and identically distributed, the variance of $\hat{\mu}$ is given by

$$V(\hat{\mu}) = n^{-1}\sigma^2 \tag{13}$$

which can be estimated by replacing σ^2 by its estimate from (12). Consider now what happens when the x's are no longer i.i.d., but belong to clusters, and that within each cluster

$$E(x_i - \mu)(x_j - \mu) = \rho\sigma^2 \tag{14}$$

for some quantity ρ, while for two observations in different clusters, we retain the assumption of independence. Then, as shown by Kish (1965), and as may be readily confirmed, (13) must be replaced by

$$V(\hat{\mu}) = n^{-1}\sigma^2 d \tag{15}$$

where d is the Kish design effect, or "deff", defined by

$$d = 1 + (\tilde{n}_c - 1)\rho\ . \tag{16}$$

The quantity $\tilde{n}_c$ is the number of households in each cluster when the clusters are all the same size; more generally it is the weighted average of cluster sizes, where the weights are the cluster sizes themselves, i.e. $n^{-1}\Sigma n_c^2$ for individual cluster sizes n_c. An estimate of ρ can be obtained from the "intracluster correlation coefficient"

$$\hat{\rho} = \frac{\Sigma_c \Sigma_i \Sigma_{j \neq i} (x_{ci} - \hat{\mu})(x_{jc} - \hat{\mu})}{\hat{\sigma}^2 \Sigma_c n_c(n_c - 1)}\ . \tag{17}$$

A number of points should be noted. In the presence of positive intracluster correlations, the number of "effective" observations is smaller than the sample size. In the extreme case, when ρ is unity, d is the cluster size, and the effective sample size is the number of clusters, not the number of observations. Even when ρ is 0.5, a high but not unusual figure, the usual formula for the standard error of a mean is optimistic by a factor of 2.34 (the square root of 5), a correction that could make a substantial difference to the conclusions being drawn. Second, although I have illustrated using clusters, the same analysis might be useful within strata, or regions, or sectors, or any other partition of the sample for which there is reason to believe that the observations within each partition are correlated. When the partition is large, ρ is likely to be small, but the size of "deff" depends on the product, and might still be large.

Third, while a similar analysis applies to the residuals in a linear regression, and while it is still true that standard formulas are likely to underestimate the standard errors, the formulas are not exactly the same, and I postpone discussion of the regression case until Section 2.1 below. This is a major issue that has been much neglected, not only in development economics, but in other applied fields using survey data.

1.1.9. General measurement issues

I do not believe that there is any reason to suppose that survey data are always and automatically of lower quality in LDCs, as if "backwardness" were a condition that applied equally to GDP and its measurement. While statistical services are sometimes poorly funded and staffed, especially in Africa, survey data are often relatively cheap to collect in poor countries, and responses are likely to be accurate where there is a high degree of literacy, and where the respondents have time to talk to the enumerators. There are also some very poor countries (such as India) where survey practice is (or at least was) second to none. Indeed, Indian statisticians have played a leading role in the development of sample surveys and of sampling techniques; the surveys of jute production in Bengal by the Indian Statistical Institute under the direction of Mahalanobis were among the first successful large-scale sample surveys [see Mahalanobis (1944, 1946)]. It is also true that respondents tend to be much more patient in LDCs, that they rarely refuse to participate in the survey, and that they will usually tolerate instruments that take several hours to administer. The differences in quality of survey data between poor and rich countries comes, not from survey administration, but from differences in the structure of income and employment. In particular, difficulties in estimating income arise, not because of respondent unwillingness or because of fear that enumerators will pass information to the fiscal authorities, but because a large fraction of poor people in LDCs are self-employed, mostly in agriculture. Self-employment incomes are notoriously difficult to estimate in developed economies, and if income estimates in general are of lower quality in LDCs, it is because self-employment income is a larger fraction of the total.

The problems are easily seen. Self-employed traders or farmers typically have no need of any concept that corresponds to economists' definitions of income. Direct questions about income or profitability cannot therefore generate useful answers, especially for individuals whose personal and business transactions are not clearly separated. Instead, it is necessary for surveys to ask detailed questions about business or agricultural operations, about sales and purchases, about quantities and prices, about taxes and transfers, about multiple business activities, about transactions in kind, and about assets. From this detailed information, an income measure has to be built up by imposing an accounting framework on each household's activities. This is a very time-consuming and complex operation, and the value so obtained is likely to be an extremely noisy estimate of the underlying theoretical magnitude, even supposing that the theory has any behavioral relevance. For example, an appropriate accounting framework might well include some allowance for depreciation of assets, tools, buildings, trees, and animals. Yet unless farmers actually think in those terms, it is unclear that the measure will be useful in understanding the farmer's behavior, however relevant it may be for measuring welfare.

A further major issue is how to handle *autoconsommation*, that fraction of consumption that is produced (or grown, hunted or bartered) by the household without going through a market. Some societies, for example large fractions of rural West Africa, are not extensively monetized, and in extreme cases, non-monetized transactions may account for nearly a half of GDP, [Heston (1994)], and a good deal more of consumption. The standard survey procedure is for values to be imputed to such consumption, typically by surveying quantities, and then by multiplying by some suitable price. The results are added to consumption purchased in markets, as well as to the value of total income. Some mechanical and apparently sensible imputation algorithms can give absurd results. For example, in one comprehensive African survey, values were imputed for water consumption. Where no price was available for a particular transaction, imputation was done at the average of the prices reported by those households who did make monetary purchases. However, the only observed prices for water were for bottled water in the main city, so that rural households were credited with very high levels of total consumption and income, much of it "Perrier" from the local river. Such extreme cases are rare, but the problems are not.

The choice of prices for imputation is rarely obvious; selling prices differ from buying prices, and there are often quality differences (perhaps better, perhaps worse) between goods sold and those retained for home consumption. In extreme cases, where monetization is the exception rather than the rule, *autoconsommation* is the tail that wags the dog; not only is most of consumption measured by making essentially arbitrary assumptions, but there must be legitimate doubts as to the usefulness of imposing a market-based accounting framework on a household or village economy in which markets play little part. Even if all these problems are solved (or ignored), it should always be borne in mind that any errors of imputation will be common to both consumption and income – and perhaps other variables, such as landholdings, or agricultural output – and the communality must be taken into account when the effects of measurement error are being explored.

Note finally that the decision of what to impute is largely arbitrary. By convention, home produced goods are included, but most home produced services are excluded. Meat and vegetables from the home farm are added to both consumption and income, but no similar allowances are made for the value of work in the home, child-minding, or the preparation of meals. While there is a good deal of agreement on the desirability of including these services, and while it is clear that there are systematic biases from failing to do so – the failure to value leisure understates the relative poverty of single parents who have very little of it – there is little agreement on how to value time. If labor markets are sufficiently well-developed so that everyone can work as many hours as they wish at the market wage, then that wage would be the appropriate price for imputing time. But if people have limited opportunities for work, as is often the case for women in many parts of the world, the appropriate rate would be less, perhaps very much less. The mislabelling of unemployment as leisure is injury enough, without adding the insult of labelling the unemployed as wealthy on the basis of their enforced leisure.

1.2. Panel data

1.2.1. Data collection

Most household surveys, in both developed and developing countries, draw new households for each new survey, so that it is generally impossible to track any given household through successive surveys. In a few cases however, most notably the World Bank's Living Standards Surveys (LSS) in Côte d'Ivoire and Ghana, the ICRISAT data from six villages in southern India, and in data collected in Pakistan and the Philippines under the auspices of the International Food Policy Research Institute, have individual households been revisited on a systematic basis at intervals of a year or more. The Living Standards Surveys have a rotating structure, with half of the households from the previous year retained and half replaced, so that data are obtained from each household on two occasions separated by a year. There have been a few other cases where households from a previous survey have been revisited, even though the original survey was not designed to be a panel. The National Council for Applied Economic Research (NCAER) in Delhi revisited a sample of Indian households after a ten year gap, Bevan, Collier, and Gunning (1989) used follow-up surveys in Kenya and Tanzania, and Smith, Thomas and Karoly (1992) report on a 1990 follow-up survey of the households in the 1978 Malaysian Family Life Survey. In all these cases, a large fraction of households or household members was found, nearly three-quarters of the latter in the Malaysian case, which is presumably a much higher fraction of those who are still alive and still resident in the country. Since the fractions reinterviewed would presumably have been higher had the resurvey been planned from the start, these experiences do not support any general supposition that panel data are more difficult to collect in LDCs, because households are "hard to find" or because of attrition in general.

Unlike cross-sections, panel surveys yield data on *changes* for individuals or individual households. Individual changes are of interest in their own right; we want to know how individual living standards change during the development process, the "who is benefiting from development" question, and we want to know whether poverty and deprivation are transitory or long-lived, the income dynamics question. Even beyond the individual, a panel design will allow more precise measurement of *aggregate* changes if the variable being measured is positively autocorrelated in the individual data, [see for example Hansen, Hurwitz, and Madow (1953, pp. 268–272) and Ashenfelter, Deaton, and Solon (1986, pp. 44–51)] for formulae. These results suggest that, even for general purpose surveys, and even when we are interested in levels as well as changes, it will generally be undesirable to replace *all* households from one survey to the next.

Changes over time in the behavior of individuals can also reveal regularities that may be obscured by individual heterogeneity in the cross-section. For example, the cross-section relationship between age and wages usually has a humped shape, with wages rising early in the life cycle, and falling later. However, older workers may be systematically different from younger workers; they may be less educated or have less experience in working with modern techniques, and their wages may have been lower throughout their lives. If so, the cross-section age-wage profile will be quite different from the profile that would result from following an individual or a cohort of individuals through time, something that is possible with panel data. By making comparisons for individuals with their own earlier behavior, each individual is effectively acting as his or her own control. There exists an extensive econometric literature that exploits this insight using panel data, and the techniques are frequently used in work on economic development. I shall return to the topic in Sections 2.1 and 2.2 below.

1.2.2. The Living Standards Surveys

The general usefulness of panel data in LDCs is an issue that is unlikely to be decided for some time, but our knowledge has recently been much expanded by the experience of the World Bank's Living Standards Surveys. These surveys have sometimes been independent cross-sections, but rotating panel data have been collected in Côte d'Ivoire, from 1985 through 1988, and in Ghana, from 1987 on a continuing basis. The LSS was originally seen as a device for monitoring poverty and inequality, and the project was begun in the Bank in response to the then (as now) extremely unsatisfactory situation in respect of international comparisons of poverty and inequality. In one example that was key at the time, it was essentially impossible in the late 1970s to deduce what had happened to distribution in Brazil during the "economic miracle" of the 1960s, whether the poor had benefitted from the income growth, or whether the benefits had flowed to a narrow wealthy group, see the original analysis by Fields (1977) and the criticism by Ahluwalia et al. (1980). Although a set of international comparisons of inequality had earlier been produced within the World Bank by Jain (1975), these were simply compilations of survey data that happened to be available within the organization at the time, with no attempt to allow for differences in definition, or to correct for non-comparabilities between countries. (Given the difficulties, Jain's figures are not a sound basis on which to make international comparisons, and results that rely on them should be viewed with great skepticism, [see for example Anand and Kanbur's (1993) critique of Ahluwalia (1976)], although the lesson is widely ignored in the recent political economy literature, for example Persson and Tabellini (1990) and Alesina and Perotti, (1992).

However, by the time the first LSS surveys were ready to be implemented, as a cross section in Peru in 1984, and with a rotating panel in Côte d'Ivoire a year later, the emphasis within the World Bank had shifted away from poverty more towards a household modelling approach. Influenced by Beckerian models of household behavior, by their extension to integrated farm-household models as in Singh, Squire, and Strauss (1986), as well as by previous experience with RAND's Malaysian Family Life Survey, the philosophy was to collect data from a relatively small number of households, but to attempt to be comprehensive, covering consumption, all income generating activities, agriculture, labor supply, business activities, gifts and transfers, as well as education (including parents' education), migration, demographics, health, and fertility, as well as some limited measurement of anthropometrics. The Ivorian data, for example, come from 1600 households, selected as a simple random sample, 800 of whom were retained as panel members, with a new 800 added each year. The 50 percent rotation pattern comes from a desire to collect at least some panel data combined with doubts about the feasibility of running a much longer panel in Africa, and from the ever present need to produce results relatively quickly.

One of the most impressive achievements of the LSS is its demonstration that microcomputer technology can be used effectively in collecting data in LDCs. A full description of the methodology is given in Ainsworth and Muñoz (1986). Responses were quickly taken to local headquarters, and entered into PCs, and then immediately run through editing programs, so that cross-checks and corrections could be carried out on subsequent visits. The rapid data entry and editing programs also mean that data are available very quickly at the end of the survey, and the teams produced preliminary survey reports within a few months of leaving the last household. The data are thus immediately available for policy analysis, an enormous improvement over previous practice, where survey results were in most countries available only years – in some cases many years – after the end of the survey.

To the extent that it is possible to tell from internal evidence, the LSS data are typically of good quality, although the breadth of the survey clearly carries some price in terms of depth and in the ability to monitor subpopulations. For example, the agricultural modules typically do not produce the sort of reliable harvest estimates that could be obtained from sample crop-cutting in an agricultural survey. But this was by design, and in many applications is offset by knowing the farmer's other activities, his and his parents' education levels, his migration history, ethnic group, and so on. The retrospective questions appear to have worked well, so that, for example, it is possible to use the fertility questions to construct reasonable estimates of changes in infant mortality over time, [see Benefo and Schultz (1993) for estimates for Ghana and Côte d'Ivoire].

Like most surveys, the LSS surveys are designed to collect data, not to experiment with survey methodology, so that it is difficult to use their results to come to general conclusions. The surveys have certainly been expensive relative to most established surveys in LDCs, with costs per household per year ranging between \$100–\$200 at 1990 prices, although it could be argued that high costs reflect the set-up costs of a new product.

One lesson from these surveys is that in countries where economic development is slow or non-existent, as in much of Africa, and where survey measures of living standards are error prone, as in Africa and elsewhere, measures of change, at both individual and aggregate levels will be dominated by measurement error. Over short periods, living standards in agriculture are variable in any case, so that short panels of a year or two are unlikely to give useful measures either of income dynamics or of the change in living standards, except possibly in the most rapidly developing countries. This is even true for "straightforward" concepts such as household size; in West Africa there is a great deal of genuine mobility among both adults and children – see particularly Ainsworth (1992) – but even here there appears to be a good deal of measurement error.

A second lesson is that it is very difficult to maintain new surveys in the field for any length of time. In the Ivorian case, personnel changes in the World Bank led to a loss of interest, and the survey ceased after 1988 apparently without leaving any permanent enhancement of Ivorien survey capability. This was particularly unfortunate because, in the face of collapsing world prices, procurement prices of cocoa and coffee – the main cash crops in Côte d'Ivoire – were cut by a half, the first such cut since independence. Had the panel been in place, the survey could have observed the process of adaptation as smallholders reacted to the cuts, but the opportunity was lost.

Third, while the computer technology has been successfully applied to the collection of data, it has been much less used for its rapid analysis, particularly in a policy context within the countries themselves. The data are now widely analyzed in academia and in international organizations, but neither analytical capacity nor software exists to make survey data rapidly available to support domestic policy making. As a result, there is less local interest in continuing surveys than is warranted by their potential utility. There have also been difficulties over setting up proper mechanisms to allow access to scholars and to the policy community. The World Bank is an operational entity, not a research foundation, and there are also legitimate interests of countries that have to be protected. Nevertheless, there would have been great benefits to constructing adequate access agreements before any data were collected, agreements that provided for public-use versions of the data at marginal cost.

Fourth and finally, I suspect that if there is a real payoff to panel data,it is over relatively long time periods, five or ten years, or even longer. Perhaps the most interesting and important work using the PSID has come from looking at income changes over long periods of time, or of comparing incomes and consumption patterns of parents and their children [Behrman and Taubman (1990), Solon (1992), Zimmerman (1992), Altonji, Hayashi and Kotlikoff (1989) and Hayashi, Altonji, and Kotlikoff (1991)]. Even here, some of the results are identical to those obtained earlier using recall data, see Sewell and Hauser (1975), and this much cheaper alternative may not be inferior for many applications. Even at best, economic development is far from instantaneous, so that changes from one year to the next are probably too noisy and too short-term to be really useful. It is hard to imagine nationally representative panels being maintained for ten or twenty years, and international organizations and foundations do not have the attention span nor the ability to commit resources over such periods. Perhaps the most promising line of research is one in which one time surveys are designed with at least the possibility of a revisit at some unspecified future date, so that ad hoc panel data can be collected on an opportunistic basis. We also need more evidence on the reliability of recall data for different kinds of information; [again see Smith, Thomas and Karoly (1992) who compare reports of the same migration events obtained in two surveys twelve years apart]. Alternatively, national survey programs might usefully incorporate some panel element, either by deciding in advance to revisit some subsample of households quinquennially or decennially, or by adding a small component of shorter period rotating panel households to their pre-existing surveys.

While there is likely to be some payoff to further experiments with panel data, it is important not to overstate the potential benefits. The PSID in the United States has generated a great deal of important *methodological* work in econometrics, but it is hard to point to any *substantive* conclusion that depends on the existence of these data. Attrition problems, especially in the early years, and the continuing presence of measurement error have made it difficult even to describe the "facts" of dynamic household behavior. Beyond that, the use of the PSID in the more ambitious research programs on life-cycle labor supply and consumption can only be described as a disaster, [see Card (1991) for a review of the labor supply literature and Deaton (1992a, Chapter 5) for the work on the intertemporal allocation of consumption].

1.2.3. Panels from a series of cross-sections

Many countries carry out their household surveys on a regular basis, often using the same instrument over time, in which case there will exist a time-series of cross-sections. Such data can be used for many of the purposes to which panel data are put, and in some respects provide a superior database.

Consider, for example, the Surveys of Personal Income Distribution that have been carried out in Taiwan every year since 1976. While it is not possible to track individuals or households from 1976–1991, it is perfectly feasible to track *cohorts* of individuals. If for example, we are interested in how individual earnings have changed in any economy experiencing very rapid growth, we can follow the mean earnings of the same group through time by looking at the members of the group who are randomly selected into each survey. If our first cohort is those born in 1951, who were 25 years old in 1976, we use the 1976 survey to calculate average earnings – or average log earnings, if that is the variable of interest – for all 25 year-olds, the 1977 survey for the average earnings of 26 year-olds, and so on, up to the average earnings of 40 year-olds in 1991. Figure 33.1, taken from Deaton and Paxson (1994a), shows the results for every fifth cohort; the connected lines track the behavior of each cohort. The figure shows a life-cycle pattern in earnings, together with strong cohort effects, with the younger cohorts earning more. As a result, it is the youngest groups whose earnings have grown the most rapidly; the average 55 year-old in 1976 had relatively little earnings growth over the subsequent fifteen years.

Such data cannot be used to look at income dynamics; even if the membership of the cohort is constant, we can estimate only the marginal distributions of income in each year, whereas estimation of income dynamics require us to observe the joint distributions, which can only come from panel data. That case apart, time-series of cross-sections can perform many of the other functions of panel data. Linear regressions with individual fixed effects can be averaged to give cohort relationships with cohort fixed effects, and can be consistently estimated by differencing the cohort level data or by using within estimators. Note too that, since we start from the individual data, the aggregation can be done over whatever function of the data is prescribed by the theory; averages of logs or of powers are as easily calculated as averages of levels. Since the cross-sections draw new households in each survey, there is no

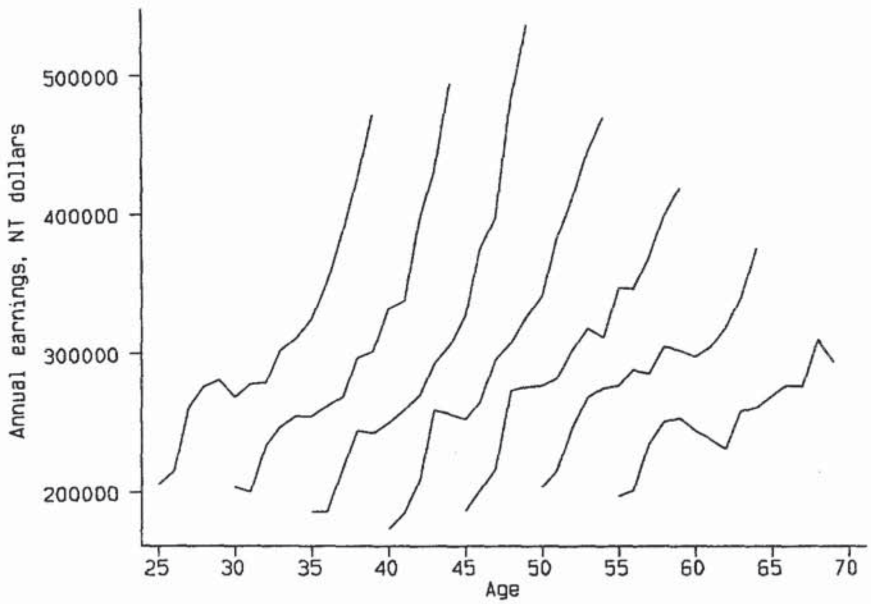

Figure 33.1. Annual earnings of seven age cohorts, Taiwan 1976–1990.

attrition bias as there is in genuine panel data, although with older cohorts, there will be (typically non-random) attrition through death, and immigration and migration will change cohort membership at all ages. The averaging will also yield less measurement error in the cohort than in the micro data provided that, as is plausible, misreporting is sufficiently uncorrelated across members of the cohort. Of course, unless the cohorts are very large, observed cohort means will estimate population cohort means with a sampling error, but the size of the error can be estimated and the appropriate corrections made using what are essentially error-in-variable estimators [see Deaton (1985) and Tan (1991) who applies these methods to life-cycle labor supply in Korea]. I shall return to the theory of these estimators when I come to the econometrics of measurement error in Section 2.1 below.

1.3. National income and other data

I have discussed household survey data at some length because, in that case, it is possible to go beyond ritual complaints about quality and quantity, and to think constructively about the effects on econometric practice of data collection, design, and measurement error. However, a great deal of econometric work in development uses non-survey data. Indeed, there has been a recent explosion of empirical work on economic growth, [see for example, Barro (1991), Barro and Sala-i-Martin (1992), Mankiw, Romer, and Weil (1992), and Levine and Renelt (1992) for four leading examples]. Most of this work is based on (and to some extent inspired by) the internationally comparable national accounts data constructed by the international price comparion project at the University of Pennsylvania, [Kravis, Heston, and Summers (1978)], and whose latest incarnation is the Penn World Table, Mark V, [Summers and Heston (1991)]. Many researchers also use the World Development Indicators, published annually by the World Bank, and which contain, in addition to a large number of social and other indicators, a competing set of national accounts – converted at official exchange rates rather than purchasing power parity exchange rates – and which, like the Summers–Heston data, are conveniently available on diskette. The Bank, the International Monetary Fund, the United Nations, and the International Labor Office all produce a wide range of other data relevant for development work, on trade, on debt, on international finance, on labor, and on social and demographic indicators.

Any sort of evaluation of this multiplicity of sources would quickly fill the whole of this Handbook. I confine myself to (*a*) a discussion of some of the index number problems that underlie international and intertemporal comparisons of income and output, and (*b*) a brief review of quality issues, the latter drawing on a recent set of conference papers on the topic.

1.3.1. Index number problems and international comparisons

Before looking at the *practical* quality issues, it is worth reviewing the *conceptual* index-number problems that underlie international comparisons of income and output. The actual Penn World Tables are a good deal more complex than the examples here, which are chosen to illustrate only the main points. Current price local currency GDP for country c at time t can be written as the sum of its component goods and services, or

$$y_{ct} = \sum_k p_{ckt} q_{ckt}\,, \tag{18}$$

where p's are prices, q's quantities, and y is income or output. Since GDP is an aggregate of value added, not of output, we must assume that there is some quantity or quantity aggregate that represents value added, something that requires suitable separability assumptions on the structure of production [see Sims (1969) and Arrow (1974)]. However, my main concern here is with different index number problems.

Suppose that there is some base country b, say, and prices are collected for each good in each country – and this is the main task of an international price comparison project – so that GDP can be repriced, using period s prices in country b as

$$y_{ct}^{bs} = \sum_k p_{bks} q_{cki} . \tag{19}$$

If country b is, for example, the US, and $s = t$, then y_{ct}^{bt} is country c's GDP at US prices, and the ratio of y_{ct} to y_{ct}^{bt} is the purchasing power parity (*PPP*) exchange rate of country c's currency in terms of US dollars. If the *PPP* exchange rate were equal to the official exchange rate – which is not usually the case – GDP at US prices could be obtained without collecting price data simply by conversion, as is done for the data reported in the *WDR*. For measuring real economic growth, we need constant price series, so that, in addition to a base country, we need a base year with a base set of relative prices. Alternatively, as in the recommended and most commonly used series in the Penn World Table, the base can be updated year by year to construct a chain index of GDP.

The problem of choosing base prices and a base country, like all index number "problems", is a conceptual and not a practical one. In principle, there is no reason other than convention to use US prices rather than Korean, Kenyan, or Chilean prices, and since they measure essentially different things, the ratio (for example) of Indian to Chinese GDP will differ depending on the choice. When making comparisons of GDP over time within a single developed country, the same conceptual difficulties arise, but because relative prices change slowly over time, the growth rate of GDP is hardly affected by the choice of base period. For those LDCs where a large share of GDP is concentrated in one or two primary commodities, this is not true, and even comparisons over time become hazardous. These difficulties are perhaps most severe for non-diversified oil exporters, although there are many other commodities (e.g. copper, cocoa, coffee) that have highly variable prices, and that make up a large fraction of GDP for some countries.

Figures 33.2 and 33.3 illustrate the time-series and cross-section implicationsof the choice of base prices. Figure 33.2 shows real GDP in Nigeria from 1965 to 1985 using two different Summers–Heston measures; according to both sets of estimates, GDP rose until the late 1970s, and has been declining since. The terms-of-trade corrected measure of GDP on the vertical axis allows for the effects on national income of changes in commodity exports and imports, while the chain measure on the horizontal axis does not. Nigeria is a major oil exporter and so has much greater growth using the terms-of-trade corrected measure. Since the Summers–Heston measures are equal by construction in 1985, Nigeria's GDP is very much lower in the earlier years when relative commodity prices are continuously adjusted; in 1965, the adjusted GDP

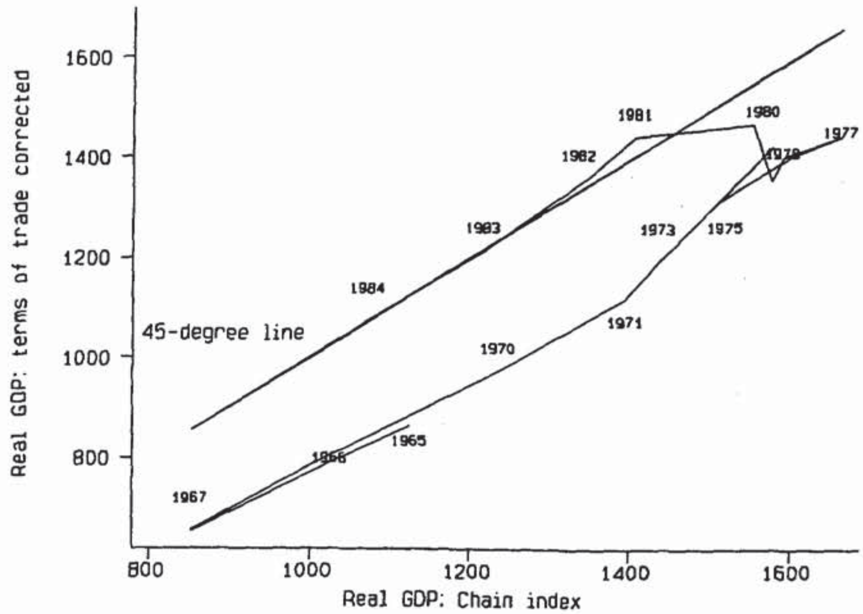

Figure 33.2. The effects of commodity prices on national income, Nigeria 1965–1985.

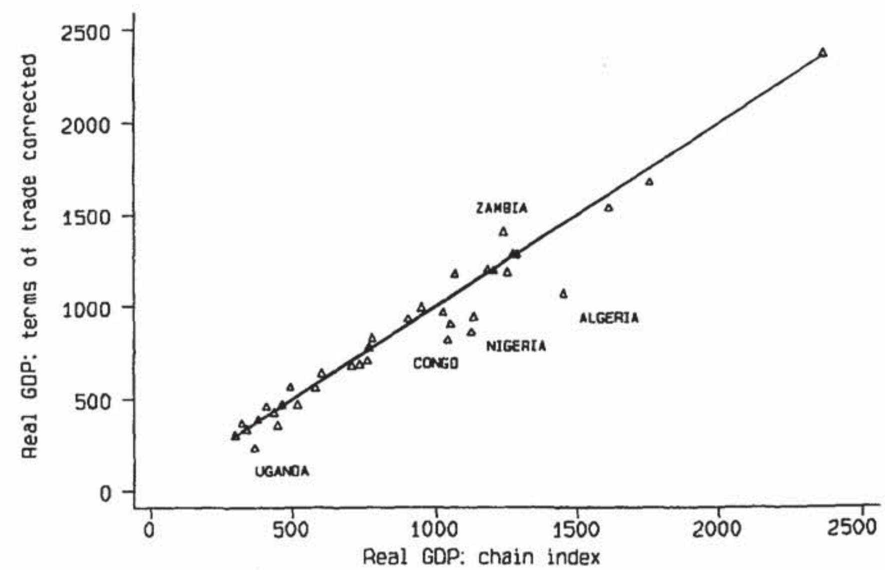

Figure 33.3. Real GDP in 1970, chain and terms of trade corrected, Africa.

estimate is only 77 percent of the chain estimate. Figure 33.3 shows both estimates for 1970 for all the African economies in the Penn World table. Most countries are close to the 45-degree line, but there are many exceptions: the ratios of adjusted to unadjusted GDP were 69 percent for Algeria, 80 percent for Lesotho, 79 percent for Nigeria, 62 percent for Uganda, and 128 percent for Zambia (the price of copper was *lower* in 1985 than in 1970.) Although some of these differences reflect the difference between output and income measures of GDP–commodity price changes have no (direct) effect on physical output although they make the country richer or poorer – these examples should illustrate the conceptual difficulties of making international comparisons in a many commodity world. (Note that I am not concerned here with measurement error, but with what is essentially an aggregation problem. The Penn World Table take their underlying data from the national accounts of the countries themselves, and these data are repriced, not corrected.)

Although these difficulties are real enough, they are minor compared with those in making comparisons across space. International differences in relative prices are both large and systematic, so that the choice of base country makes a large difference to the estimates. Because labor is relatively cheap in poor countries, the relative price of non-tradeables to tradeables rises with economic development, so that, for example, services and government tend to be relatively cheap, and investment relatively expensive in poorer countries. There are associated substitution patterns in both production and consumption which give rise to the standard biases associated with fixed weight or current weight index numbers. Using American wage rates to revalue Indian labor costs will tend to overstate Indian relative to American GDP, because Indian GDP is (or should be) more specialized in labor-intensive activities, a substitution effect that is turned into apparently high income by applying the prices of a labor-scarce economy. In India, servants – both domestic and civil – are cheap and widely used, so that, at American prices, the real size of the domestic service sector in India is exaggerated. For the same reasons, making comparisons in American prices will bias down the estimated growth rates of the poorer countries, since rising real wages will narrow the relative price differentials, and progressively reduce the exaggeration of GDP in LDCs.

Once we go beyond output measures to interpret GDP as a measure of living standards, then we also have to face the question of whether it makes sense to treat preferences as identical across countries, or at the very least, whether international differences in climate and the conditions of production do not severely compromise international welfare comparisons. The problem of calculating the comparative costs-of-living for an American diplomat in Karachi or Reykjavik is well-defined and calculable to some degree of approximation. It is more of an open question whether it makes sense to compare the living standards of a Nepalese peasant and a Taiwanese fisherman, let alone those of an American lawyer and a Namibian bushman.

1.3.2. Quality issues in development data

This brief review section is based on a set of papers from a conference on databases for development. I have benefitted particularly from the overview paper by Srinivasan (1994), as well as those by Bouis (1994) on nutrition, by Chamie (1994) on demography, by Evenson and Pray (1994) on agriculture, by Rozansky and Yeats (1994) on trade, by Heston (1994) on national income and growth rate comparisons, and by Behrman and Rosenzweig (1994) on labor force and education data. The interested reader should consult these papers; only a few highlights are summarized here.

There are a number of other important *practical* issues in international *national income data*. Heston (1994) points out that the share of non-monetized subsistence in GDP can be greater than 40 percent in the poorest countries, that its measurement is fraught with difficulties, and that the solutions are far from uniform across countries. Many LDCs estimate GDP growth using growth rates of physical indicators, with benchmark weights that are frequently seriously outdated. Given GDP, consumption is obtained as a residual by subtracting net exports from trade flows, government expenditure, and investment in plant and machinery. Over-invoicing of imports and under-invoicing of exports are common methods of transferring funds abroad in countries with exchange controls and overvalued exchange rates, and such practices compromise not only the trade data, but will lead to overstatement of consumption and understatement of saving. In largely agricultural societies, estimation of physical output is difficult, and evidence [in Srinivasan (1994) and Evenson and Pray (1994)] suggests that discrepancies between household survey and national accounts estimates of food consumption and production in India may come more from the national accounts than from the surveys. This is an important lesson with implications beyond India, since national income estimates of income and consumption are nearly always given more weight than survey estimates when there are discrepancies between the two, a practice that has little justification in general.

There have also been suggestions that estimates of GNP are manipulated for political ends. It is certainly true that one of the more widely noted ratios, the relative per capita GDP of India and China is a number about which it is hard to obtain reliable estimates. The Penn World Tables Mark 5 estimate that in 1985 international prices, China's GDP per capita at $1,883 was 2.71 times that of India in 1985. The previous version (Mark 4) of the same tables gives the ratio, again for 1985, at 3.26, now calculated in 1980 prices. Srinivasan (1994) quotes the 1992 WDR figures of $350 for India and $370 for China in 1990, and points out that the respective growth rates over 1965–1990 (in the same publication) are 1.9 percent and 5.5 percent. If these figures are correct, in 1965 GNP per capita in China was only 44 percent of that in India, a statistic that defies belief.

Counting *people, births, and deaths* is also problematic. Chamie (1994) points out that there are a number of LDCs who have yet to carry out their first census, that only a third of LDCs have had a census since 1985, and that 27 percent of countries have a latest census that was conducted prior to 1975. Recent, reliable data on life-expectancy (infant mortality) are available for only a half (a quarter) of LDCs, and two-thirds of African countries have collected no data on life-expectancy since 1970. Many of the figures published in the *World Development Report* and the UN's *Human Development Report* are estimates and projections, not measurements.

There are also puzzles and discrepancies in data on *health, education,* and *nutrition*. Self-reported health data in LDCs typically show a *positive* correlation between living standards and ill-health, something that is usually attributed to better-off people reporting a larger fraction of health problems. Recent work at RAND appears to have made real progress on this issue [see Strauss, Gertler, Rahman, and Fox (1992)]. Questions about ADLs (Activities of Daily Living, such as walking and eating) and IADLs (Instrumental Activities of Daily Living, such as shopping) ask respondents whether, for example, they would find it easy, difficult, or very difficult, to perform a set of specified tasks (climbing stairs, fetching water) that are relevant to everyday life. The results of these questions reveal more sensible, richer, and interesting patterns of health with income and age than do the previous self-reported measures. Education data often exaggerate enrolments, by reporting attendance on the first day of school, or by expressing total enrollments, including those of adult students and grade repeaters as a fraction of the normal age groups for those grades, so that enrolment fractions greater than unity are possible [see Behrman and Rosenzweig (1994)].

Nutritional data are usually obtained from survey data on household purchases of food, and less often from 24-hour food consumption recall data. The latter generate much lower income elasticities of calories and of foods than do the former, Bouis (1994). Bouis argues in favor of the lower figures, on the grounds that traditional food elasticities imply implausible weight patterns. If the food (and calorie) elasticity is 0.4, say, then people in the top decile of the income distribution, who are perhaps six times as well off as people in the bottom decile, consume more than twice as much food and calories as those in the bottom decile, and ought therefore to weigh more than twice as much, something that we do not observe. Not everyone would accept the existence of such a reliable and simple relationship between calories and weight, even in the long-term, nor is it clear that the purchase method of calculating nutrition is necessarily worse than the more invasive and detailed recall surveys. The problem cannot be attributed to imputation biases in the survey data along the lines discussed above [see Bouis and Haddad (1992) and Subramanian and Deaton (1992)]; the latter paper also rules out functional form problems. However, it is possible, as Bouis argues, that there is a very high income elasticity of food wastage and of food gifts to servants, relatives, and even animals, thus reconciling the purchase with the intake data.

Finally, there is an excellent discussion of the quality of international trade data by Rozansky and Yeats (1994) who look for inconsistencies (*a*) across different sources, particularly the U.N., the Fund, and the Bank, (*b*) between trading partners, comparing recorded imports of A from B with recorded exports from B to A, (*c*) between trade totals over commodity groups and their component sums, and (*d*) across revisions of SITCs, for those groups not affected by the reclassifications. The results are far from encouraging, and by all criteria, trade data from LDCs show more and greater discrepancies than data for OECD countries, with discrepancies apparently worsening over time. To take just one example, comparisons under (*b*) show that only 2–3 percent of US or French trade gets "lost", compared with more than 50 percent for South Africa (not surprisingly), Venezuela, Seychelles, and Bahrain. The IMF's estimate of Venezuela's 1982 exports is 20 times larger than that compiled by the UN.

1.3.3. Some implications

The news from this section is dismal. National income and growth comparisons across countries are plagued by conceptual index number problems, and by immense practical difficulties. Many frequently used data from LDCs are of poor quality, or only pretend to exist, having their only reality in the mind of bureaucrats in New York or Washington. And while the Penn World Table, which provides probably the best and certainly the most heavily used set of national income data, has provided a great step forward in producing data at a common set of prices, it cannot be better than the raw (and uncorrected) data from the individual countries on which it is based.

What then should be done? Researchers should obviously be encouraged to be critical of the data, and to take every opportunity to explore the consequences of measurement error for their analysis. However, when the data are of such low quality that it is difficult to establish any results – as with much of the official macroeconomic data for Africa – it is difficult to pinpoint specific problems, or to know where to press for improvement. It is also clearly sensible to press for more resources to be devoted to data collection, and it would be a notable improvement if international agencies were to advertise their data more precisely, so that, for example, projections and estimates were clearly separated from genuine measurements. There have been questions as to whether the international organizations have any real interest in improving data collection. Skeptics have argued that the World Bank (or at least its loan staff) is interested in the *quantity* of loans, not ultimately in their *quality*, and that without an interest in the latter, there is little chance that the necessary resources will be committed to the improvement of the data either on its own account, or by helping to improve data collection by its member countries. In defense, it must be remembered that international organizations are responsible to their members, and in many cases are limited in the extent to which they can correct, question, or criticize the data that are provided by the member countries. Unless policy makers can be persuaded that the quality of their decisions are being compromised by poor data, they are unlikely to find the resources to improve matters.

2. Econometric tools for development analysis

2.1. Econometric analysis of survey data

In this second part of the review, I discuss a series of econometric techniques that are particularly appropriate for or are widely used in the analysis of development questions. In this first of three sections, I shall be concerned mostly with techniques used in the analysis of survey data, although a good deal of the material applies more generally. Subsequent sections deal with time-series and non-parametric issues respectively. My focus is on developments in econometric practice over the last ten or fifteen years, and how they relate to practice in published work in economic development. In particular, I attempt to identify a number of topics where best practice is somewhat ahead of what is readily available in the textbooks. One topic that will occur repeatedly is *robustness*. Inferences that rest on arbitrary – sometimes even incredible – assumptions are hard to take seriously, and there has been a major effort in econometrics – as in statistics more generally – to find ways of generating conclusions that are both credible and convincing and that are not the more or less immediate consequence of arbitrary supporting assumptions.

An important role of econometrics is to substitute for experimentation, and much of the econometric literature on simultaneity, heterogeneity, selectivity, omitted variables, and measurement error can be thought of as finding procedures that can bring the non-experimental results closer to the experimental ideal. Many of these procedures rest on strong parametric assumptions, some of them necessarily so, but others do not, and in some cases it is possible to obtain results with quite unobjectionable assumptions. When this is not so, the fact is in itself important, since it implies that robust inferences are not possible, and that the assumptions of the investigator are as necessary as the data for drawing the conclusions.

Even for standard and well-understood techniques, such as linear regression, inferences can be made more robust, either by moving away from OLS to alternatives such as quantile regression, or less radically, by calculating standard errors in ways that are robust against the failures of standard assumptions that are common in survey data. I begin this section with these topics.

2.1.1. Heteroskedasticity and linear regression

As is well-known, the presence of heteroskedasticity in linear regression affects neither the unbiasedness nor the consistency of OLS estimation. However, the assumptions of the Gauss-Markov theorem are violated, so that OLS is no longer efficient, and the usual formula for the variance-covariance matrix of the parameter estimates is no longer valid. In particular, if the regression model is, for $i = 1, \ldots n$,

$$y_i = x_i\beta + u_i;\ E(u_i) = 0;\ E(u_i^2) = d_i > 0\,, \tag{20}$$

and the OLS estimator is, as usual, $(X'X)^{-1}X'y$, then the variance-covariance matrix is given by

$$V = (X'X)^{-1}X'DX(X'X)^{-1} \tag{21}$$

where D is an $n \times n$ diagonal matrix whose diagonal is the d's from (20). Although V in (21) cannot be evaluated without knowledge of the d's, it has been shown by Eicker (1967), Huber (1967), Fuller (1975) and White (1980), that it can be consistently estimated by replacing D by the diagonal matrix whose elements are the squared OLS residuals. Note that the consistency here is of the matrix V, not of D, the number of elements in which increases with the sample size, and which therefore cannot be consistently estimated. Following White and MacKinnon (1985), this relatively straightforward calculation can be modified and extended in a number of ways, some of which are likely to yield improvements in performance. These methods yield estimates of the variance covariance matrix that are asymptotically valid, and do not require the user to know or to specify the specific form of the heteroskedasticity in (20).

As I argued in Section 1, the stratification of surveys is likely to generate heteroskedasticity, and even without it, experience suggests that residuals are more often heteroskedastic than not. There are a number of tests for heteroskedasticity, of which perhaps the most convenient is that suggested by

Breusch and Pagan (1979), in which the squared OLS residuals are regressed on variables that are thought to be likely candidates for causing the heteroskedasticity, usually including the levels, squares, and interactions of the original explanatory variables. Indeed, as is easily checked, this is the correct specification if β in (20) is taken to be distributed randomly in the population. Under the assumption that the original regression errors are normally distributed, the null of homoskedasticity implies that the explained sum of squares of this supplementary regression will be distributed as χ^2 with degrees of freedom equal to the number of regressors in the supplementary regression. This test is closely related to the information matrix test proposed by White (1980).

It is clearly good practice to calculate and report standard errors and other test statistics that are robust to departures from homoskedasticity. Furthermore, my own experience suggests that it is difficult to pass the Breusch–Pagan test in practical applications, and that heteroskedasticity is usually revealed not just by this test, but by others, such as the quantile regression techniques discussed below. That said, the heteroskedasticity-consistent standard errors and tests are rarely very different from those given by the standard formulas. An upward correction of about 30 percent to standard errors appears to be common, and this correction would not normally lead to startling differences in inference.

2.1.2. Clustering and linear regression

In Section 1 above, I showed that when observations within survey clusters are correlated, survey cluster sampling requires a revision of the formula for the standard error of an estimated mean. In particular, the usual variance, which is the population variance divided by the sample size, has to be multiplied by the Kish design effect (16), which depends on the average number of observations per cluster and the size of the intracluster correlation coefficient. Similar considerations apply to the estimation of linear regressions when there are grounds for believing that the errors are correlated within clusters. The fact that the sample is clustered does not in itself imply that there must be a non-zero intracluster correlation once other explanatory variables have been taken into account. However, survey clusters in rural areas in LDCs are typically geographically dispersed villages, so that there are likely to be unobserved communalities that are shared between households in the same village, and that differentiate them from those in other villages. Note too that there may be intrahousehold correlations between households beyond the cluster levels, for example across provinces or regions, correlations that could come from ethnic factors, from the way in which markets operate, or from the way that the government allocates services across administrative areas.

To illustrate the issues, I shall suppose that the survey is clustered, that there is a regression model such as (20), and that the residuals are positively correlated across observations in the same cluster. As is the case with heteroskedasticity, OLS remains unbiased and consistent, but is inefficient, and as with heteroskedasticity, our main concern is with the invalidity of standard formulas for the variance-covariance matrix of the OLS estimator. A useful result, due to Scott and Holt (1982), is that the Kish design effect is the *maximal* correction that is required, and that, in general, the estimated variances will understate the true variances by a factor that is less than the design effect. However, the maximum is attained when all the right hand side variables in the regression are constant within clusters, as would be the case when the x's are cluster prices, wages, or variables measuring access to schools, health clinics or the like [see also Kloek (1981)]. If some x's vary across members of the cluster, and are correlated between clusters with the other variables, the design effect will overstate the correction.

As with heteroskedasticity, there are parametric and non-parametric methods for correcting the variance-covariance matrix. Among the former would be to specify a variance components model at the cluster level, the estimation of which would allow the calculation of the intracluster correlation coefficient, which can then be used to calculate standard errors. Alternatively, an intracluster correlation coefficient can be calculated from the OLS residuals using (17) and the result used to estimate the correct variance covariance matrix for the OLS estimator. More generally, it is possible to allow for cluster fixed effects, and to work with deviations from village means. This is a useful technique in some contexts, and I shall discuss it below, but note that it does not permit us to estimate coefficients for any regressors that do not vary within the clusters.

A useful procedure is based on the fact that cluster sizes are typically small relative to the total sample size, say 10 or 16 households per cluster, so that it is possible to correct the variance covariance matrix non-parametrically by using the OLS residuals to "estimate" the variance-covariance matrix of the residuals in each cluster, just as the squared OLS residuals are used to "estimate" the variances in the heteroskedasticity-robust calculations. (I use the inverted commas around "estimate" because in neither case are we trying to obtain a consistent estimate of the individual residual variance or individual cluster residual variance covariance matrix.)

Suppose then that we have estimated the regression by OLS, and that for cluster c we have obtained the OLS residuals e_c. We then calculate a robust OLS variance covariance matrix by calculating [see White (1984)],

$$\tilde{V}(\hat{\beta}) = (X'X)^{-1} \sum_{c=1}^{C} X_c' e_c e_c' X_c (X'X)^{-1} \tag{22}$$

where X_c is the submatrix of X corresponding to cluster c, and C is the total number of clusters. Note that in the case where there is only one household per cluster, (22) is the standard formula for the heteroskedasticity-consistent variance covariance matrix. Note too that (22) does not require that there be homoskedasticity, either within clusters or between them, nor that there be a common intracluster correlation coefficient. It is therefore robust against quite general forms of intracluster correlations. The equation is implemented in the software package STATA as part of the huber command, and the corresponding procedure for panel data is described by Arellano (1987).

How much does all this matter? The answer seems to be a great deal, certainly more than is the case for the more familiar heteroskedasticity correction. In many applications, the correction is not much less than the design effect, and in my own work, I have frequently found that the usual formulas give standard errors that are understated by a factor of two to three, a much more serious matter than the 30 percent that seems to be common for the heteroskedasticity correction. The problem is exacerbated by the fact that in so many development applications, the explanatory variables are constant within the clusters, the wage, price, and access variables listed above. It would be invidious to list papers that use clustered data without correction, although [see Deaton (1988, 1990a)] for two selected examples, but there are hundreds of papers in development economics looking at labor market questions, at the demand for commodities and nutrition as a function of prices, and at access to education and health services where the true significance levels for t-values should probably be closer to 6 than to 2. Many of these studies will have to be redone, and I suspect that there will have to be a good deal of revision of conclusions. Of course, these problems are not confined to studies of economic development, and similar considerations apply for example, in labor economics. Indeed, Moulton (1990) has provided a particularly dramatic example using American state level data, where a small intrastate correlation coefficient is combined with large numbers of observations in each state to yield a design effect of nearly 10.

2.1.3. Quantile regressions

The method of quantile regression is not one that has been much used in economics to date, perhaps because of computational considerations. These have now been solved – the qreg command in STATA is an example – so that this extremely useful tool is readily available without the need for special coding. The basic idea was first introduced into economics by Koenker and Bassett (1978) and can be described as follows.

Quantile regression, like linear regression, is concerned with the distribution of a scalar random variable y conditional on a vector of covariates x. In linear regression, it is assumed that one characteristic of this distribution, its mean, is a linear function of x, or at least we attempt to fit a linear function to the conditional expectation, or *regression function*. Instead of the mean, we might choose to work with the median, and to assume that the medians of y conditional on x are linear in x, or at least to fit a linear function to the medians. This would be a median regression, or 0.5 quantile regression. In principle, we can do the same for any other quantile of the distribution, thus constructing the p-quantile regression, where p is any number between 0 and 1.

Given the idea, why should we be interested, and if we are interested, how can such regressions be calculated? Start with the former. First, by looking at a number of different quantile regressions, we can explore different parts of the conditional distribution. For example, consider the relationship between wages and schooling; at any given number of years of schooling, there is a (conditional) distribution of wages, presumably reflecting unobserved abilities and other labor market skills. In general, thereis no reason to require that the rate of return to an additional year's schooling should be the same at all points in the distribution of abilities conditional on schooling, and quantile regression would pick up the differences, see Chamberlain (1991). Used in this way, quantile regression is essentially a non-parametric technique that describes the shape of the empirical distribution without imposing prior restrictions. As such, it can also provide an indication of heteroskedasticity. If the conditional distribution changes shape with one or more of the explanatory variables, quantile regressions at different quantiles will have different slopes, [see Koenker and Bassett (1982)] for a test that uses this property.

Second, just as the median is less sensitive to outliers than is the mean, so are quantile regressions more resistant to outliers than are mean (least-squares) regressions. Median regression is affected by the presence of an outlier, but not by changes in its position, provided of course that it remains above or below the median. As such, quantile regression is one of several regression techniques that have robustness properties superior to OLS [see in particular Huber (1981) and Hampel, Ronchetti, Rousseeuw and Stahel (1986)]. Standard methods of robust regression typically downweight large residuals identified from a previous regression, iterating to convergence. Such procedures require an estimate of the scale of the residuals in order to identify outliers, and thus are sensitive to patterns of heteroskedasticity that are handled naturally by quantile regressions.

Third, quantiles are not affected by monotonic transformations of the data, so that, for example, the median of the logarithm of y conditional on x is the logarithm of the median of y conditional on x. As we shall see in the next subsection, this property has useful consequences.

The estimation of quantile regressions rests on extensions of the well-known result that the median is the point closest to the data in the sense of minimizing the sum of the absolute deviations. Median linear regression parameters are given as the value of the vector β that minimizes

$$\sum_{i=1}^{n} |y_i - x_i'\beta| = \sum_{i=1}^{n} (0.5 - 1(y_i \geq x_i'\beta))(y_i - x_i'\beta) . \tag{23}$$

Koenker and Bassett (1978) show that the p-quantile estimator can be calculated by minimizing a generalization of the second expression in (23),

$$\tilde{\beta} = \operatorname{argmin} \sum_{i=1}^{n} (p - 1(y_i \leq x_i'\beta))(y_i - x_i'\beta) . \tag{24}$$

Although these expressions do not permit explicit solutions, the parameters can be obtained quickly by linear programming methods.

2.1.4. Zeros: probits and Tobits

In development applications, as elsewhere in economics, many variables of interest have limited ranges, either a set of discrete values, or are continuous but limited to some interval. The most frequent example of the latter is when a variable is restricted to positive values; a farmer can produce nothing or something, but cannot grow negative amounts, a consumer may or may not smoke, but cannot sell tobacco, and so on.

Binary discrete choices are typically modelled by using probit or logit models, and often less formally using the linear probability model, in which a dichotomous dependent variable is regressed on the covariates. Provided the standard errors of the linear probability model are corrected for the heteroskedasticity that is inevitable in such a specification, there is no good reason not to use it, especially when sample sizes are large enough so that computational costs of probit and logit are non-trivial. The fact that linearity is an inappropriate functional form for a probability is unlikely to be problematic provided the bulk of the data are in the range where predicted probabilities are far from either zero or unity.

Cases where the data are a partly discrete and partly continuous are harder to handle. The most common case is where a continuous response is censored at zero, for which the standard model is the Tobit, viz.

$$y_i = \max(0, \beta' x_i + u_i);\ E(u_i|x_i) = 0;\ E(u_i^2|x_i) = \sigma^2 . \tag{25}$$

The model is also interpreted as one in which $x_i'\beta + u_i$ is a latent variable, observed when zero or positive, but censored to zero, i.e. replaced by zero, when it would otherwise be negative. A more general version of this model, in which the censoring is controlled by a second latent variable, will be discussed in the subsection on selection below. Estimation is usually done by assuming that the (conditional) distribution of u_i is normal, and following Tobin's (1958) original procedure of estimating the parameters by maximum likelihood. The log likelihood function for this problem is globally concave, so that it is a routine problem in non-linear estimation, typically no more time consuming than the estimation of a probit. Note also that if (25) is correct, OLS will be inconsistent for the parameters β. The regression function is

$$E(y_i|x_i) = (1 - F(-x_i'\beta/\sigma))x_i'\beta + \sigma \int_{-x_i'\beta/\sigma}^{\infty} \tau \, \mathrm{d}F(\tau) \tag{26}$$

where $F(.)$ is the distribution function of $\sigma^{-1}u_i$. (26) will generally not be linear in x_i.

Tobin's maximum likelihood method works well when its assumptions are satisfied. However, the estimates will typically be inconsistent if normality fails, or perhaps more seriously, if there is heteroskedasticity [see Arabmazar and Schmidt (1981, 1982) and Goldberger (1983)]. This is more than a technical problem, and it is straightforward to construct realistic examples with heteroskedasticity where the maximum likelihood estimates are worse than OLS. Particularly in survey data, where heteroskedasticity is endemic, there is no reason to suppose that Tobit will give estimates that are any better than OLS ignoring the censoring. With heteroskedasticity and censoring, neither technique is likely to give satisfactory estimates.

There are two approaches that make sense in practical applications. The first is to abandon this way of thinking about the problem. The standard approach starts from a linear model, and then complicates it to allow for censoring, treating the linearity as a maintained structural hypothesis. In the standard linear regression, this makes sense, because the structural regression coincides with the regression function, and is readily recovered from the data. In many cases, this structural assumption of linearity is merely a convenience, and there is no particular reason to believe that the underlying relationship is genuinely linear. When this is so, the standard procedure for dealing with censoring is not an attractive one, because the original linearity assumption has nothing to support it but convenience, and the convenience is lost with the censoring. The regression function (26) is not a convenient object to handle, and a more suitable alternative would be to start, not from the structure, but from some suitable direct specification for the regression function. Given the presence of the zeros, linearity might not be plausible, but some other flexible functional form might do perfectly well. I shall discuss one particular non-parametric procedure in Section 2.3 below.

The second approach is less radical, and makes sense when there is good reason to retain the linear structure. In this case, it is desirable to use an estimation technique that will deliver consistent estimates in the absence of normality and homoskedasticity. There are a number of these in the literature, all more or less experimental. One that is relatively straightforward to compute is Powell's (1984) censored least absolute deviations estimator, which can be implemented given a program (such as STATA) that allows quantile regression.

Powell's estimator rests on the previously noted fact that medians are preserved by monotone functions. Hence, if $q_{50}(y_i|x_i)$ is the median of the conditional distribution of y_i, then from (25)

$$q_{50}(y_i|x_i) = \max[0,\, q_{50}(x_i'\beta + u_i)] = \max(0,\, x_i'\beta) \tag{27}$$

since $\max(0,\, z)$ is monotone in z, and where the last equality rests on the assumption that the median of u_i is zero. Given (27), consistent estimates of the parameters can be obtained by running a nonlinear median (50th percentile) regression, or equivalently by minimizing

$$\sum_{i=1}^{n} |y_i - \max(0,\, x_i'\beta)|\,. \tag{28}$$

Buchinsky (1994) suggests a simple – if not necessarily efficient – computational strategy is to run a median regression of y on x, to calculate predicted values and discard any that are negative before rerunning the regression. Repetition of this procedure, if it converges, will lead to the parameters that minimize (28). In my own – admittedly limited – experience, this works quite satisfactorily even if terminated after five cycles. As is to be expected from a robust procedure, the estimates are a good deal less efficient that Tobit when Tobit's assumptions are correct, and the technique is probably not suitable for a small number of observations. Nevertheless, it is certainly worth trying on survey data, and given large enough samples is likely to be safer than either OLS or Tobit.

2.1.5. Regression bias

Censoring is only one of many cases where the model of interest does not coincide with the regression function, the conditional expectation of y on x. There are a wide range of circumstances where the explanatory variables are correlated with the disturbance, so that least squares regression does not yield consistent estimates of the structural parameters. Omitted variables, simultaneity, heterogeneity, measurement error, and sample selection are all capable of rendering OLS inconsistent, and a great deal of effort in the development literature has gone towards developing techniques that will deliver consistent estimates for a range of specific problems. Some of these techniques draw on panel data when available, and many others rely on one form or another of instrumental variable estimation. In the next few subsections, I review a number of specific topics that illustrate the use of these techniques and some of the issues associated with them.

2.1.6. Agricultural production functions: heterogeneity and panel data

The estimation of farm production functions is a problem that often arises in development applications, whether we are simply attempting to relate physical outputs to physical inputs, or whether we are concerned with more elaborate models of farm-households and the associated integrated models of consumption and production [see for example Singh, Squire, and Strauss (1986)]. Production functions are one of the oldest topics in econometrics; many of the issues reviewed by Marschak and Andrews in 1943 are still relevant, and Mundlak's (1961) paper on agricultural production functions is the first – or at least one of the first – to use fixed effect estimators with panel data as a remedy for unobserved heterogeneity. The simultaneity and omitted heterogeneity problems in this case arise in many other related applications.

A good starting point is the "obvious" procedure, which is to regress outputs on inputs, as for example in

$$\ln(q_i/A_i) = \beta_0 + \beta_1 \ln A_i + \beta_2 \ln h_i + \beta_3 \ln z_i + u_i \tag{29}$$

where A_i is land, so that q_i/A_i is the yield per hectare of farm i, h_i is labor input, and z_i is some other input, such as fertilizer, or perhaps the farmer's education. The sign of β_1 is relevant to the question of whether large or small farms are more "productive", the coefficient β_2 tells us about the marginal productivity of labor on family farms, and the size of β_3 might tell us whether inputs are being efficiently used, since a very large marginal product of fertilizer relative to its costs might be used as an argument for intervention in distribution or extension services.

The problem is that OLS estimation of (29) will tell us none of these things. The finding that $\beta_1 > 0$, that smaller farms have higher yields, is the traditional one since Chayanov's (1925) findings for Russian farmers, and has been widely observed elsewhere [see for example Sen (1962) for India, and Berry and Cline (1979)] for a review of other research. There are many interpretations of the result; that higher output per head is an optimal response to uncertainty by small farmers [Srinivasan (1972)], that there are dualistic labor markets, [Sen (1966, 1975)], or that hired labor requires more monitoring than family labor, [Feder (1985)]. Perhaps the simplest explanation is that (29) omits unobserved heterogeneity, in this case land quality, and that this omitted variable is systematically correlated with the explanatory variables. Farms in low-quality marginal areas (semi-deserts) are typically large, and farms in high-quality land areas are often much smaller. That a garden adds more value-added per hectare than a sheep station does not imply that sheep-stations should be reorganized as gardens. The omitted quality variable is negatively correlated with A_i and so causes the estimated coefficient to be downward biased, from the true value of zero to the observed negative value. Indeed there is some evidence that controlling for quality either reduces or removes the effect [see Bhalla and Roy (1988) and Benjamin (1993)].

Similar arguments apply to the other variables in the production function. For example, it is sometimes found that the returns to fertilizer use, estimated from regression coefficients, are many times larger than would be consistent with productive efficiency [see for example Benjamin and Deaton (1988) for Côte d'Ivoire and Bevan, Collier, and Gunning (1989) for Kenya and Tanzania]. Should fertilizer use be encouraged, and extension services expanded? Not if what we are seeing is that the farms with the higher quality land, or with the most go-ahead farmers, are also those who adopt new technologies. Output is high, not because of the return to inputs, but because of unobservables, land and farmer quality, that are correlated both with inputs and outputs.

Omitted heterogeneity induces correlations between explanatory variables and the error term in a way that has the same consequences as simultaneity bias. Indeed, the production function is likely to suffer from genuine simultaneity bias even in the absence of heterogeneity; inputs, like outputs, are under the control of the farmer, and can have no general claim to exogeneity. The combination of genuine simultaneity and heterogeneity has the further effect of ruling out the use of lags to remove the former; while it is true that seeds have to be planted before the crop is harvested, heterogeneity across farmers will mean that seeds are not exogenous for the harvest, a problem that I shall return to in Section 2.2 in the context of using predetermined variables with panel data. The result of all these considerations is that the regression function of physical output conditional on physical inputs will rarely be informative about the underlying technology.

There are a number of possible econometric solutions to these problems. Note first that, under the standard neoclassical assumptions of the farm-household model, the appropriate exogenous variables for production are not inputs, but the prices of inputs, and the appropriate estimation technique is either instrumental variables applied to the physical production function, or the estimation of a dual specification, in which the technology is specified as a profit function whose derivatives are the demand functions for inputs and the supply functions of outputs, all functions of prices.

There are two problems here, one theoretical and one practical. First, many development economists are comfortable with physical relationships between inputs and outputs, but are unwilling to commit themselves to a neoclassical or "capitalist" view of agriculture in LDCs. Although it is easy to make fun of such "engineering" as opposed to "economic" approaches, there are many cases where misgivings have a real basis. Some markets are not well-developed, and farm inputs are sometimes allocated in ways other than through competitive markets with parametric prices. Second, and consistent with these views, it my impression that it is much more difficult to estimate satisfactory relationships in which inputs and outputs are functions of *prices*, rather than of each other, where the omitted heterogeneity will often guarantee a good if entirely spurious fit. This practical problem will be exacerbated in those cases where there is relatively little price variation across farms. Certainly, it is rare for researchers to report first-stage regressions of inputs on input prices.

When panel data are available, the heterogeneity can be addressed by assuming that it takes the form of additive fixed effects in (29). Consistent estimates of the parameters can then be obtained by OLS applied to differences across periods, or to deviations from individual means. Hence, if (29) is rewritten in standard regression form with i denoting the farm, $i = 1, \ldots, n$, and t the time period, $t = 1, \ldots, T$, we have for the differenced estimator

$$\Delta y_{it} = \beta'.\Delta x_{it} + u_{it} - u_{it-1} \tag{30}$$

for $t = 1, \ldots, T-1$, while for the within-estimator, we have

$$y_{it} - y_{i.} = \beta'(x_{it} - x_{i.}) + u_{it} - u_{i.} \tag{31}$$

where the suffix $\{i.\}$ indicates the time mean for farm i. Mundlak's (1961) original application of (31) to Israeli farms was designed to remove the effect of "management bias", the heterogeneity that arises from some farmers being better farmers than others.

The ability to deal with heterogeneity does not come without cost, and indeed many of the most important difficulties are recognized in Mundlak's paper. First, the technique depends on the specific functional form for the heterogeneity, that it take the form of an additive fixed effect. There are often good theoretical reasons why this will not be the case, and there is no straightforward way of dealing with fixed effects in nonlinear models. Second, the differencing or demeaning loses n observations, so that if T is small, as is often the case, there will be a substantial loss in precision. Third, when the x's are positively correlated over time, differencing or demeaning reduces variation, so that once again precision is lost. In the extreme case when some of the x's are constant, there is zero precision, and the parameters are not identified. In the agricultural production case, farm size will usually change little or not at all over short periods, which thus precludes any attempt to resolve the "small farms are more productive" question by using fixed effect estimators. Fourth, and perhaps most serious, in the presence of white noise measurement error in the explanatory variables, demeaning or differencing of positively autocorrelated x's will not only reduce the variability of the signal – variability in the true x's – but it will inflate the ratio of noise to signal in the regressors. In the standard case where measurement error induces attenuation bias, the attenuation will be worse using the difference or within estimator. The combination of loss of precision and increased attenuation bias often erases in the difference or within estimates effects that were significant in the cross-section, even when the model is correctly specified and there is no heterogeneity bias. Such results provide no indication that heterogeneity bias is an issue in the cross-section. It clearly makes sense to use Hausman (1978) tests to check whether the estimates from the difference or within estimates are indeed significantly different from the cross-section estimates, although when significant differences are found, further information is needed to discriminate between measurement error or heterogeneity bias as an explanation.

2.1.7. Panel data in practice

Perhaps for the reasons given in the previous paragraph, it is difficult to use panel data – especially short panel data – to generate convincing conclusions and it is particularly difficult to disentangle measurement error from omitted heterogeneity. In particular, it is clear that panel data are no panacea, and that there is no guarantee that difference or within estimates will be preferable to OLS on a cross-section. Even so, panel data have allowed investigators to consider alternatives that could not otherwise have been explored, and to relax previously maintained assumptions.

The techniques that were originally developed for agricultural production functions have been widely applied to other sorts of "production", from the production of health in terms of health inputs – where exactly the same issues of simultaneity and heterogeneity arise – as well as to wage equations, where earnings are a function of schooling and heterogeneity arises because econometricians cannot control for unobserved ability. Such studies are extensively reviewed by Behrman and Deolalikar (1987) and by Strauss and Thomas in this volume. At their best, these studies are sensitive to the difficulties, and much can be done to interpret results by using prior information about the likely size of measurement errors, so that changes between cross-section and within estimates can plausibly be explained. Investigators have also been creative in using the panel data idea, not just for differences over time, but in other applications. A number of studies, for example Behrman and Wolfe (1984, 1989) on education, and Rosenzweig and Wolpin (1988) on child health, use data on siblings to allow for family fixed effects, and to estimate within-family regressions. Fixed effects can also be associated with the villages from which survey clusters are selected, so that village means can be swept out from all the households in each cluster, thus allowing consistent estimation of the effects of quantities that vary within the village in the presence of arbitrary inter-village effects [see for example Deaton (1988) and the further discussion below].

Other studies [Rosenzweig and Wolpin (1986, 1988) and Pitt, Rosenzweig, and Gibbons (1993)] have used panel data to approach the important problem of using regression analysis to aid project evaluation. For example, Pitt et al. look at (among other things) the effects of grade-school proximity on school attendance in Indonesia combining survey with administrative data. One potential problem is that the placement of the schools is unlikely to be random – indeed the whole point of project evaluation would be to avoid random allocation – and that allocation may be influenced by unobservable local factors that themselves have a direct effect on outcomes. The simplest example would be when the government allocates schools to areas with poor attendance, so that an ultimately successful program would be one in which school attendance is the same everywhere, and where a regression analysis would show no effect of school proximity on attendance. (It is also possible that already successful areas are better at getting resources, for example through influential politicians, or by being able to turn money into votes.) Although the Indonesian data are not panels, the same administrative units (*kecamatans*) show up in successive surveys, so that it is possible to compute a difference estimator at *kecamatan* level, a procedure that is closely related to the panel data from cross-sections methodology discussed above. This difference estimator shows much larger effects of school location on school attendance than are visible in the cross-section.

2.1.8. Latent variables and measurement error

Instrumental variables and panel data are only two of the possible ways of dealing with unobserved heterogeneity. In some cases, a more direct approach is available in which the data provide enough information to identify the effects of interest even in the presence of latent variables. These cases fall into the class of multiple indicator and multiple cause, or MIMIC models, which are related both to factor analysis and to models of measurement error [see in particular Goldberger (1974) and Jöreskog (1973)]. Rather than discuss the general case, I look at two particular applications from the development literature.

The first is the model of imperfect fertility control of Rosenzweig and Schultz (1983), used again in Rosenzweig and Schultz (1987) and in Rosenzweig (1990), and in a somewhat different context by Pitt, Rosenzweig, and Hassan

(1990). A skeletal form of the model can be written as

$$\begin{aligned} y_{1i} &= \alpha_1 + \beta y_{2i} + \gamma_1 . z_{1i} + \mu_i + u_{1i} \\ y_{3i} &= \alpha_2 + \gamma_2 . z_{2i} + \theta \mu_i + u_{2i} \end{aligned} \tag{32}$$

where y_1, y_2, and y_3 are endogenous variables, z_1 and z_2 are vectors of exogenous variables, u_{1i} and u_{2i} are error terms, and μ_i is unobserved heterogeneity. In the Rosenzweig and Schultz papers, the first equation explains the number of births in terms of the endogenous contraceptive effort y_2, so that μ_i is couple-specific fecundity. The second equation is used to explain various characteristics of child health which are also influenced by latent fecundity. In the Pitt, Rosenzweig and Hassan paper, which is concerned with nutritional status and consumption, the first equation relates weight for height to calorie consumption (the two endogenous variables) and an individual "endowment" μ_i. In this case, $y_3 = y_2$, which is calorie consumption, and the parameter θ measures the extent to which the household reinforces $(\theta > 0)$ or offsets $(\theta < 0)$ natural endowments in the intrahousehold allocation of food.

As always with MIMIC models, the major issue is identification, and strong assumptions are required to be able to recover θ. Provided that the β's and γ's are identified – which poses no non-standard issues – θ is identified from the covariance matrix of the residuals provided that u_1 and u_2 are orthogonal – which requires that there be no common omitted variables in the two equations – and provided the instruments are orthogonal to the unobservable μ's, a set of conditions that would seem to be indefensible in any real application. In practice, (32) is usually estimated by applying instrumental variables to the first equation and then using the residuals as a regressor in the second equation, with some allowance for the "measurement error" that comes from the presence of u_1 in the first equation. (Note also that without correction, such a two-step procedure will not generally lead to valid standard errors.) An alternative (and more direct) procedure would be to substitute for μ in the second equation from the first, and to estimate the resulting equation by instrumental variables.

A second example comes from my own work on estimating price elasticities of demand using the spatial price variation that is revealed in cross-sectional household surveys when respondents report, not only how much they have spent, but also the physical quantity bought, so that for each household we can construct a unit value index. This unit value index reflects both local prices and the quality choices of individuals, with unit values usually higher for better-off households who purchase more expensive varieties, even of relatively homogeneous goods. A stripped-down version of the model can be written as follows, see Deaton (1988):

$$
\begin{aligned}
y_{ic} &= \alpha_1 + \beta_1 \ln x_{ic} + \theta \ln p_c + f_c + u_{1ic} \\
\ln \nu_{ic} &= \alpha_2 + \beta_2 \ln x_{ic} + \psi \ln p_c + u_{2ic}
\end{aligned}
\tag{33}
$$

where i is a household, and c is the cluster or village in which it lives. The first equation explains y_{ic}, the demand for the good – for example the logarithm of quantity or the budget share of the good – in terms of household total expenditure x, the unobservable price p, and village fixed effect f, and a random error term. The price is assumed to be the same for all households in the village and is therefore not indexed on i. The fixed (or random) effect is uncorrelated with the price, but can be correlated with x or with any other included variables that are not constant within the cluster. The unobservable price also manifests itself through the unit value ν which is the dependent variable in the second equation. The parameter β_2 is the elasticity of unit value to total expenditure, or quality elasticity – Prais and Houthakker (1955) – while ψ allows for possible quality shading in response to price changes. If price and unit value were identical, ψ would be unity and β_2 would be zero, but quality effects will make $\beta_2 > 0$ and $\psi \leq 1$.

Once again, identification is a problem, and as is intuitively obvious from using the second equation to substitute out for the unobservable log price, only the ratio θ/ψ can be estimated. The β-parameters can be estimated by a within-estimator in which village effects are swept out, a procedure that also provides estimates of the variances and covariances of u_1 and u_2. Given the β's from the within-village estimates, construct the corrected village averages

$$
z_{1c} = y_{.c} - \hat{\beta}_1 \ln x_{.c}, \; z_{2c} = \ln \nu_{.c} - \hat{\beta}_2 \ln x_{.c} \, . \tag{34}
$$

At the second stage, we calculate the estimate

$$
\hat{\phi} = \frac{cov(z_1 z_2) - n_c^{-1} \hat{\sigma}_{12}}{var(z_2) - n_c^{-1} \hat{\sigma}_{22}} \tag{35}
$$

where the covariances are taken over villages, where n_c is the number of households per cluster, and the σ's are estimated from the first stage variance covariance matrix of the residuals. Using (33), it is straightforward to show that (35) is a consistent estimate of the ratio θ/ψ.

Note that (35) is a standard errors-in-variables estimator in which the OLS estimator, which is the ratio of the covariance to the variance, is corrected for the component that is attributable to measurement error. The standard error for $\hat{\phi}$ can be obtained from first principles by application of the "delta method", or by adapting the formulas in Fuller (1987) for the effects of the first-stage estimation. The model can also be extended to allow for other covariates both within and between villages, it can be expanded into a system of demand functions with many goods and many prices – [Deaton and Grimard (1992)] – and a simple theory of quality shading can be appended to the model so as to allow θ and ψ to be separately identified, [Deaton (1988)].

2.1.9. Selection models

Selection bias occurs in many different forms; one very general formulation is due to Heckman (1990) and is useful for thinking about a number of issues that arise in development practice. Heckman's formulation has three equations, two regression equations, and a switching equation that governs which of the two determines behavior. The regressions are:

$$y_{0i} = x'_{0i}\beta_0 + u_{0i}, \;\; y_{1i} = x'_{1i}\beta_1 + u_{1i} \,. \tag{36}$$

The dichotomous switch variable d_i takes values 1 or 0 and satisfies

$$d_i = 1(z'_i\gamma + u_{2i} > 0) \tag{37}$$

where the "indicator function" 1(.) is defined to take the value 1 when the statement in brackets is true, and 0 otherwise. The dependent variable y_i is thus determined by

$$y_i = d_i y_{0i} + (1 - d_i)y_{1i} \,. \tag{38}$$

There are several cases in the development literature that use the model in essentially this form. In van der Gaag, Stelcner, and Wijverberg (1989) and Stelcner, van der Gaag and Wijverberg (1989), (36) are wage equations for the formal and informal sectors in Peru, while (37) is the equation determining choice of sector. Pitt and Sumodiningrat (1991) look at the adoption of high yielding versus traditional variety rice in Indonesia, so that the equations (36) are variety specific profit functions, and (37) is the profit maximizing choice between them. Bell, Srinivasan and Udry (1992) model credit markets in the Punjab using a demand equation, a supply equation, and a condition that enforces a ration whenever the supply is less than the demand. In the appropriate notation, all of these fall within the general framework of the previous paragraph. Various special cases of Heckman's model occur even more frequently in development practice.

Consider first setting both β_1 and the variance of u_1 to be zero in the second equation in (36). Given this, we have the Tobit model when the right hand side of the first equation in (36) coincides with the argument of the indicator function in (37). When the two quantities are different, we have the important case generalizing Tobit where the censoring is determined by different factors than determine the magnitude of the dependent variable when it is not censored. This would be the correct model (for example) for the demand for fertilizer if what determines whether a farmer uses fertilizer at all – perhaps the existence of a local extension agent – is different from what determines how much is used conditional on use – perhaps the price of fertilizer, land quality, or the anticipated price of output.

For this generalized Tobit model, (36) and (37) imply that, if we condition on y being positive, the regression function is

$$E(y_i|x_i, z_i, y_i > 0) = x_i'\beta + \lambda(z_i'\gamma) \tag{39}$$

where I have suppressed the zero suffix and where

$$\lambda(z_i'\gamma) = E(u_{0i}|u_{2i} \geq -z_i'\gamma)\,. \tag{40}$$

Equation (40) can also be applied to the case of truncation. In contrast to censoring, where we see zeros when the observation is censored, with truncation, the observation does not appear in the sample. In this case, although (40) holds, and although the switching equation (37) still explains the truncation, we cannot use it to estimate the switching parameters in the absence of the information that would have been contained in the truncated observations. We have only (40) to work with, and it is clear from inspection that identification, if it is to be achieved at all, will require strong supplementary assumptions. In cases where truncation cannot be avoided, it will rarely be possible to make a convincing separation between the truncation variables and the variables in the structural equation. With censoring, we have both (37) and (40) and, as we shall see below, identification is easier.

Heckman's general formulation can also be used to analyze the "policy evaluation" or "treatment" case that was discussed in the context of heterogeneity. In (36), set $u_1 = u_2$ and $\beta_0 = \beta_1$ except for the constant term. Equation (38) then becomes

$$y_i = \alpha + \theta\, d_i + x_i'\beta + u_i \tag{41}$$

where the parameter θ is the difference between the two constant terms and captures the effect of the policy on the outcome. Given the structure of the model, and the determination of d_i by (37), the policy indicator will generally be correlated with the error term in (41) so that the policy effect cannot be consistently estimated by least squares. This is simply another way of looking at the same problem discussed above, that when we want to estimate the effects of a policy or a project, we must take into account what determines it, and having done so, we will usually find that we cannot discover its effects by standard regressions. The basic issue here is the correlation of explanatory variables with the error term, and it matters less whether we think of that correlation as coming from simultaneity, heterogeneity, selection, or omitted variables.

The fully general model (36) through (38) can be estimated as it stands by using maximum likelihood once some joint distribution – typically joint normality – is specified for the three error terms u_0, u_1, and u_3. Given normality, the special case of generalized Tobit can be estimated using a short-cut technique that avoids the need for maximizing a custom built likelihood function. In a famous paper, Heckman (1976) proposed what has come to be known as the "Heckit" or Heckman's probit, by analogy with Tobit or Tobin's probit. At the first stage, the γ-parameters in (37) are estimated up to scale by probit applied to a dichotomous variable that is 0 when y is censored and 1 otherwise. The results are then used to calculate the λ-function in (40), which under normality takes the form of a Mill's ratio, which can then be used on the right hand side of (40) to estimate the β's. This technique is very widely used in the applied development literature, although (notably) not in the study of wage equations among Panamanian males by Heckman and Hotz (1986).

The role of the distributional assumptions in these models has come under increased scrutiny in recent years. As we have already seen, maximum likelihood estimation of the Tobit model is inconsistent when homoskedasticity fails. In the general model, even identification can hinge on the distributional assumptions on the error terms, a situation that is practically little different from lack of identification altogether. The identification of the general model under minimal distributional assumptions has been addressed in papers by Manski (1988), Chamberlain (1986) and Heckman (1990). The identification of the switching equation (38) is straightforward, provided of course that we normalize the variance to unity. The identification of the structural equations in the absence of knowledge of the joint distribution of u_0, u_1, and u_2 requires that there is at least one variable in the switching equation (37) that is absent from the structural equations, although this in itself is not sufficient; for example, at least one of the variables unique to the switching equation must be continuous.

Finding variables that affect switching but are absent from the structure is closely akin to the general problem of finding instruments, and is frequently as difficult. In the paper that introduced selection effects into applied econometrics, Gronau (1973) found that women's wages were systematically higher when they had small children. The implausibility of children directly increasing labor market productivity led to a model in which children acted as selection variables, with higher reservation wages required to bring women with children into the labor force. But this sort of clear separation appears to be rare in practice, and in cases where there are no grounds for excluding the selection variables from the structure, there is little point in pursuing the selectivity through a normality-dependent correction, as opposed to estimating the regression function without any attempt to separate structure from selection.

When the models are identified, it is still desirable to pursue estimation strategies that do not rest on normality. There exist a number of robust techniques for various special cases of the general model. For the "policy evaluation" model given by (37) and (41), the obvious technique is instrumental variables – although see the earlier discussion on heterogeneity – which is dealt with in the next subsection. Robust techniques for dealing with generalized Tobit are still in the experimental stage, and there is little practical experience upon which to draw. However, one straightforward method is given by Newey, Powell, and Walker (1990), who generalize the Heckit to make it robust against departures from normality. At the first stage, they estimate a non-parametric version of probit using the kernel techniques discussed in Section 2.3 below. Alternatively, if we are not too concerned with the role of normality in the probit, the first stage of Heckit can be retained to provide an estimate of the index $z'\gamma$. Indeed, the linear probability model is also a competitive technique for the first stage. At the second stage, Newey, Powell and Walker suggest that the index be entered into the regression, not through the Mill's ratio, but as a polynomial that will mimic the unknown and distribution dependent λ-function in (39). This procedure avoids having to specify a joint distribution for the two error terms, and will force us to confront the lack of identification where it exists. For example, the procedure will break down if the x and z variables are the same.

2.1.10. *Instrumental variables and natural experiments*

The "policy evaluation" model (37) and (40) is only one of the many regression models where the technique of instrumental variables can be useful. Indeed, whenever there is a correlation between an explanatory variable and the error term, whether induced by heterogeneity, simultaneity, measurement error, omitted variables, or selectivity, instrumentation can be used to generate consistent estimates provided that it is possible to find instruments that are (at least asymptotically) correlated with the explanatory variable and uncorrelated with the error terms. Of course, the variance of IV estimators will be larger than OLS, so that even when the latter is inconsistent, there is no guarantee that the IVE will be closer to the truth; as usual, the price of greater generality is decreased precision, and as usual, it is important not to interpret an insignificant estimate from IVE as evidence that the OLS estimate is spurious.

Once again, the Hausman test is useful for checking whether the IV estimates of the parameters of interest are significantly different from OLS.

In this subsection, I have two points to make in addition to what is contained in any good textbook treatment of instrumental variable estimation. The first concerns the policy evaluation model and the role of "natural" experiments. The second is concerned with some aspects of the finite sample distribution of instrumental variables estimates that are important for the interpretation of results in practice.

Perhaps the best environment for policy evaluation is where the "treatment" is randomly allocated, so that the effects of the policy can be assessed by post-experimental comparison of outcomes for treatment and controls. For obvious reasons, projects and policies in LDCs are typically not allocated randomly, although there is occasional scope for randomization when there are more suitable individuals or localities that would like to be treated than there are funds to support them [see Newman, Gertler, and Rawlings (1993)]. There is also scope for genuine experimentation prior to policy evaluation, something that has been carried furthest in the US [see Grossman (1993) for a review], but is also of increasing interest in LDCs [see again Newman et al.]. When experimentation is not possible, or simply was not done, the question arises as to whether and in what circumstances econometric technique is a substitute.

A good deal of recent attention has been devoted to "natural" experiments, situations where there is no experimental intent, but where the design or implementation of a program has features that allows the construction of good instruments, or of groups of experimentals and controls where it can be convincingly argued that selection into each group is effectively random. Good examples of this technique are provided by papers on veteran status and wages by Angrist (1990) and by Angrist and Krueger (1989). Angrist studies the effects of Vietnam veteran status on wages using the fact that selection for the draft was at least in part random through the allocation of lottery numbers. The comparison of wage rates between those who received high and low lottery numbers reveals the veteran effect without contamination by other omitted variables because the latter cannot be correlated with the selection. Angrist and Krueger note that in the last years of World War II the selection procedures into the military generated a (weak) correlation between the likelihood of induction and the position of an individual's birthday within the year, with those born earlier more likely to be drafted. Using a sample of 300,000 individuals from the 1980 census, Angrist and Krueger show that the positive association between World War II veteran status and wages is reversed when birth dates are used as instruments, and that the subsequent negative wage effect is consistent with the negative effect on wages of having been a veteran of the Vietnam War.

These are impressive studies, and they show how to make good use of natural experiments. However, it is important to note the features of these examples that are responsible for the credibility of the results. In one case, randomization is actually present, so that we are quite close to a genuine experiment. In the other, the birth date effect comes from an accidental feature of the program. In cases where the element of natural experiment comes from deliberate choice by an individual or an institution, the othogonality between the instrument and the error terms can be much harder to defend. In particular, differences in government policy between areas or individuals can rarely be treated as experimental, and indeed one of the achievements of the political economy literature has been to stop economists automatically treating government behavior as an exogenous explanatory variable. Differences in educational policy between two otherwise "similar" countries such as Kenya and Tanzania may provide useful insights on educational outcomes, since at least some hard to observe features are automatically controlled for [see Knight and Sabot (1990)], but the policy differences are neither random nor accidental, and the comparison can hardly be labelled a natural – or any other kind of – experiment. Technological change, such as the green revolution in India – Rosenzweig (1990) – can plausibly be taken as exogenous to Indian farmers, but that is a different matter from the adoption of the technology, which is always likely to be correlated with farm-specific features that make it appear more likely to succeed. Some components of fertility can be thought of as random, such as Rosenzweig and Wolpin's (1980) use of the birth of twins as a natural experiment, but when differences in "the intercouple variation in the biological propensity to conceive" can only be measured as the residuals from a regression, again see Rosenzweig (1990), the validity of the instruments requires that the residuals be uncontaminated by other omitted factors, something that will often not be credible. Even regional price variation, which is routinely treated as exogenous – as in my own work in spatial demand analysis that was discussed above – will not provide valid instruments in the presence of regional taste variation [see Kennan (1989) and Deaton (1994, Chapter 2)].

The natural experiment methodology works best when, as in the Angrist and Krueger examples, it focusses on some detail of the program that is plausibly random. While there is no guarantee that all programs will have this sort of feature, it is possible that a detailed examination of administrative procedures can yield a useful instrument. Although major program outlines are set by politicians or administrators who are well aware of the consequences of their actions, the microstructure of the implementation is often undertaken by bureaucrats who are allowed administrative discretion, whose motivation is no more than completing their task, and whose actions may sometimes be close to random, as when decisions are influenced by birthdates or alphabetical order.

Even in these favorable cases, there are serious econometric problems associated with instrumental variables. Even in the best cases, where samples are large enough to make asymptotic theory useful, the variance covariance matrix of the IV estimator exceeds that of OLS by a positive definite matrix, so the removal of bias – if it is present – comes at the price of precision. But the greatest practical difficulties relate to the finite sample properties of IVEs, and to the fact that practical inference is inevitably based on large-sample distributions that often give very little idea of the true behavior of the estimates. Although general results are available on the finite sample distribution of instrumental variable estimates – see Phillips (1980) – the formulae are not readily calculated and are thus of limited value for applied researchers. Even so, a good deal is known. The finite sample distributions of IVEs can be thick-tailed, so much so that IVEs possess finite sample moments only up to the degree of overidentification [see Davidson and McKinnon (1993, 220–224) for discussion and references]. Thus in the (fairly common) case where there is exact identification, the IVE does not possess a mean. Hypothesis testing in such circumstances is obviously hazardous, especially when asymptotic standard errors are used to compute t-values. Even when there is sufficient overidentification to guarantee the exitence of the moments, IVEs are biased towards OLS in finite samples [see Nagar (1959) and Buse (1992)]. In the extreme case, where the first stage regression has no degrees of freedom and fits perfectly, OLS and IVE are mechanically identical. There is therefore a tradeoff between having too many instruments, and risking the close replication of the biased OLS estimates, or of having too few, and risking dispersion and apparently extreme estimates.

Special cases of small sample distributions of IVE have recently been investigated by Nelson and Startz (1990a,b) and by Maddala and Jeong (1992). Nelson and Startz analyze the simplest case of a linear regression with a single right hand side variable and a single instrument. They show that the asymptotic distribution of the IVE will often be a bad approximation to the true distribution when the instrument is a poor one in the sense of being only weakly correlated with the explanatory variable. In particular, there is no guarantee that a "poor" instrument will necessarily result in insignificant estimates when the instrument is used. Indeed, Nelson and Startz produce examples where the opposite occurs, and where apparently significant estimates are generated spuriously by the use of a weak instrument. The moral is that it is important to present the results of the first-stage estimation in two-stage least squares – a practice that is far from routine – and that little credibility should be given to instrumental estimates where the predictive power of the first stage has not been established. Although Nelson and Startz's analysis covers only the univariate case, the natural extension would be to check the joint significance in the first-stage regression of the identifying instruments, for example by calculating an F-test for the variables not included in the structural equation.

Buse's results are also concerned with finite sample bias and with the fit of the first-stage regressions. Using Nagar approximations to the moments of the estimator – a technique that will only be valid in those cases where enough moments exist – Buse asks how the finite sample bias changes as the number of instruments is increased. In the case where there is one endogenous right hand side variable, Buse's formula implies that one additional instrument will decrease the approximate bias if

$$R_2^2 - R_1^2 > (L_1 - 2)^{-1}(R_1^2 - R_0^2) \tag{42}$$

where $L_1 > 2$ (the asymptotic approximations are not useful for smaller values) is the number of instruments before the addition, R_0^2 is the fit of the regression of the endogenous right hand side on the exogenous variables included in the structural equation, and R_1^2 and R_2^2 refer to the same regression but with the addition of L_1 and $L_1 + 1$ instruments respectively. Not surprisingly, it is possible for poor instruments to increase the bias, but according to (42) the bias can be exacerbated even by the addition of an instrument that makes a substantial contribution to the fit. These results, like those of Nelson and Startz, underline the importance of examining the first-stage regression in two-stage least squares.

2.1.11. Test statistics

The construction and interpretation of test statistics follows the same general principles in development practice as in the rest of econometrics. The "trinity" of likelihood based tests, Wald, likelihood ratio, and Lagrange Multiplier tests are widely used in the development literature as elsewhere [see Engle (1984) for a review]. Many of these tests are based on the normal distribution, so that in parallel with the increased focus on robust estimation techniques, there has been a move towards robust test statistics. In particular, and as we have already seen, Hausman (1978) tests are frequently useful, since they provide a general framework for comparing efficient estimates obtained under restrictive assumptions with more robust estimators whose consistency is more generally guaranteed but that are less efficient under the conditions that justify the original estimator. The generalized methods of moments (GMM) estimators introduced by Hansen (1982) also provide an integrated framework for estimation and inference. Although GMM estimators are perhaps most frequently used in a time-series context for the estimation of rational expectations models, they also provide a useful way of thinking about many of the techniques discussed above, since it is often the case in development practice that estimation is based on conditional moment restrictions.

For example, suppose that we generalize the instrumental variable models discussed above by writing the structural model in the form

$$f(y_i, x_i, \beta) = u_i \tag{43}$$

and appending the k conditional moment restrictions

$$E(u_i|z_{ij}) = 0, \quad j = 1, \ldots, k . \tag{44}$$

The sample analog of (44) is the condition $n^{-1}Z'u = 0$, and the GMM estimator of β is given by making the quantity as small as possible, or

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ u'ZW^{-1}Z'u \tag{45}$$

where W is a suitable positive definite weighting matrix. Hansen shows that the optimal choice for W is the variance covariance matrix of the vector $Z'u$, $Z'\Omega Z$ say, in which case the criterion function in (45) is

$$u'Z(Z'\Omega Z)^{-1}Z'u . \tag{46}$$

Under the null that all the instruments are valid, (46) will have an asymptotic χ^2 distribution with degrees of freedom equal to the number of instruments in excess of the number required to identify the parameters. Hence (46) is often referred to as an overidentification test, and it can be used to good purpose in most of the situations described above that involve the use of instrumental variables. In practice, the matrix Ω is not known but is constructed from the residuals analogously with the heteroskedastic and cluster effect models discussed above, see equations (21) and (22) above. Following Newey (1986), we can also estimate the model using a subset of instruments and calculate a Hausman test by comparing the value of the criterion function with that obtained using the full set. As we shall see in the next section, this framework also provides a useful way of addressing a number of important issues in the analysis of macroeconomic and panel data.

When standard test procedures are applied in the context of large scale surveys, the issue often arises as to whether critical values should be adjusted for sample size. In time-series work, including most applied macroeconomics, sample sizes do not vary widely from one application to another, and the issue does not arise. But in survey based econometrics, the number of data points can vary from a few hundred to a few million, and there are arguments that suggest that it is inappropriate to use the same critical values at all sample sizes. In particular, when working with very large sample sizes, investigators often find that standard statistical procedures lead to hypotheses being rejected "too frequently", and that even when the null hypothesis seems like a good approximation, t-tests and F-ratios are large relative to conventional significance levels.

The intuitive argument is that no one believes that hypotheses are literally true, so that when we say $\beta = 0$, we really mean that β is close to zero. With a small sample size, β will not be very precisely estimated, so that an estimate close to zero will not lead to a rejection of the null. However, if β is being consistently estimated, the sampling distribution will tighten as the sample size increases so that even very small values of β can be associated with very large t-statistics when we test the hypothesis that it is zero. Of course, if we literally believe that $\beta = 0$, standard test procedures give the right answer, and the t-test will reject the null 5 percent of the time whatever the sample size.

There are deep issues of statistical philosophy here, which it is not appropriate to rehearse here. Many economists are adherents of classical inference, while others believe that it is inherently nonsensical to test point nulls, and reject testing altogether in favor of estimation. The Bayesian point of view has been eloquently argued by Leamer (1978). He points out that if we hold the probability of Type I error fixed as the sample size increases – which is the classical prescription – all of the increased precision of estimation from the additional observations is being devoted to reducing the Type II error. If β is not zero, the probability of failing to detect the fact may be large with 100 observations, and infinitesimally small with 200,000. The classical procedure, by holding fixed at 5 percent (say) the probability of Type I error, is one that commits us to lexicographic preferences or loss functions over the two types of error; lower Type I error is always preferred, independently of Type II errors. Although classical statisticians emphasize the asymmetry of Type I and Type II errors, as well as the care that should be taken in formulating the null, it is still hard to see why we should subscribe to such preferences, rather than trading-off the two types of error as the sample size increases.

Recognizing the persuasiveness of Leamer's argument is a good deal easier than doing something about it. From the Bayesian perspective argued by Leamer, the solution is to choose models based on posterior probabilities. For example, if we want to test that a parameter or subset of parameters is zero, we compare the posterior probability of the restricted and unrestricted models, and select whichever is the larger. As shown by Schwarz (1978), in sufficiently large samples this rule leads to a simple adjustment to the likelihood ratio test. In particular – see Chow (1983, pp. 300–302) for an exposition – the posterior probability of each of the models is dominated for large enough sample sizes by terms of the form

$$\ln L(\hat{\beta}) - \frac{k}{2} \ln n \tag{47}$$

where $\hat{\beta}$ is the maximum likelihood estimate, k is the number of parameters, and n the sample size. Hence, standard likelihood ratio tests can be transformed into large sample Bayesian posterior probability tests by comparing twice the log likelihood ratio, not with a χ^2 distribution with q degrees of freedom, where q is the number of restrictions, but with q multiplied by the logarithm of the sample size. For a standard F-test, which in large samples is q times the χ^2, the Schwarz procedure calls for the restrictions to be rejected when the F is larger than the logarithm of the sample size.

While the use of such criteria should ultimately depend on the philosophy of inference, I have frequently found in my own work that the Schwarz criterion gives sensible answers. As far as it is possible to tell on other grounds, it seems to discriminate between what I intuitively think of as "large" and "small" violations of the null; accepting the null according to these sample dependent criteria rarely leads to untoward consequences, while its rejections are ignored at one's peril. However, it should also be emphasized that some of the large test statistics that are frequently encountered using survey data may be attributable to some of the other considerations discussed in this section, such as a failure to treat heteroskedasticity or to allow for cluster effects, and these may provide more mundane explanations for the apparent frequency with which sensible nulls are rejected.

2.2. Econometric issues in time-series

The way that econometricians think about time-series data has undergone major changes in the last ten to fifteen years. Much of the change has come from macroeconomics, and its relevance to development economics lies largely but not exclusively in that area. In this section, I provide a brief introduction to the modern language of time series analysis, and to some of the major issues that have been debated in recent years, particularly those associated with unit roots. I begin with univariate descriptions of time series, and with alternative methods of handling trends. I use illustrations from my own and others' work on modelling commodity prices, one of the leading time-series topics in economic development. Commodity prices are extremely variable, arguably trending, and their behavior generates major problems of macroeconomic stabilization and growth for the large number of LDCs that are dependent on exports of primary commodities. Adequate univariate time-series representations of commodity prices would be a considerable aid to the design of macroeconomic policy in much of the world. The second subsection turns from univariate to multivariate time-series analysis and considers the problems that arise in regression analysis when the variables are trending. These two sections can do no more than scratch the surface of what has become an immense topic. For readers interested in pursuing these questions further, Campbell and

Perron (1991) provide an accessible discussion as well as an excellent synthesis and guide to the literature.

Time-series issues are important not only in statistical description and regression analysis, but also in modelling behavior, particularly when individuals are assumed to behave as dynamic intertemporal optimizers. Such approaches have had a considerable impact on the way we think about saving, and the recent literature on saving and development has begun to make use of the tools of dynamic optimization. I discuss Euler equation approaches to the econometric analysis of these models, as well as the more elaborate structural estimation strategies that have recently made their appearance in development and elsewhere.

The final subsection returns to the topic of panel data, and to the time-series issues involved in its use. Although many of the applications here are again macroeconomic, particularly to the analysis of growth using time-series data from a cross-section of countries, the same issues arise in many microeconomic applications where there are dynamic features, such as lagged dependent or predetermined variables.

2.2.1. Univariate time-series models

For many purposes it is useful to have a simple descriptive model of a time series. Although such models rarely have a direct behavioral content, they can encapsulate the stylized facts about a time-series, whether or not it has a trend, its autocorrelations at different lags, and how long-lived are the effects of unanticipated shocks. In the case of commodity prices, the existence or otherwise of a downward trend in the terms of trade of LDCs has been a topic of debate in economic development since Prebisch and Singer in the 1950s. Similarly, sensible macroeconomic stabilization in the face of shocks to commodity prices is greatly eased if it is possible to come to some sort of understanding of what a given shock means for future prices. In principle, univariate time-series analysis can cast light on these questions.

A useful starting point is the familiar ARIMA formulation of Box and Jenkins (1970). For a univariate time series y_t this can be written

$$A(L)\Delta^d y_t = B(L)\varepsilon_t \tag{48}$$

where L is the lag operator, $A(L)$ and $B(L)$ are finite degree polynomials, Δ is the backward difference operator $(1-L)$ so that Δ^d indicates differencing d times, and ε_t is an independently and identically distributed or "white noise" process. The roots of the polynomials $A(L)$ and $B(L)$ are assumed to lie outside the unit circle, so that the series $\Delta^d y_t$ is stationary. Since the quantity ε_t is white noise, it – or at least the deviation from its mean – is not predictable and is therefore frequently referred to as a shock or innovation. The basic idea of (48) is first to difference the series to induce stationarity, and then to use low order autoregressive and moving average polynomials to capture the autocorrelation structure of the differenced process. If the series is stationary to begin with, d will be zero. However, in many macroeconomic applications, the change or the rate of growth of the series is more naturally thought of as stationary, and there are some cases, such as perhaps the level of prices, where even the rate of growth will be non-stationary, requiring twice-differencing to generate a stationary series. A non-stationary series that has to be differenced d times to induce stationarity is said to be integrated of order d, or $I(d)$, so that a quantity such as GDP would typically be $I(1)$ – or difference stationary – while the price level could be $I(2)$ and unemployment rates or interest rates either $I(0)$ or $I(1)$.

When the parameter d in (48) is greater than 0, then we say that the series y_t has a unit root; the term comes from noting that the left hand side of (48) can be written as $A(L)(1-L)^d$ which is a polynomial with d unit roots. The simplest case of a unit root model is the random walk with drift, which is written

$$\Delta y_t = \varepsilon_t = \eta + u_t \tag{49}$$

where η is the mean of ε and is the average rate at which y increases in each period. Other more general unit root models come from treating u_t in (49) as a general stationary process, although the restriction in (48) that the roots of $B(L)$ lie outside the unit circle rules out the case where u_t is itself the first difference of a stationary process; otherwise we could write any *stationary* process – including white noise – in the form (49).

Equation (49) is an example of how ARIMA models deal with a trend. With or without additional serial correlation, such models often provide a natural and straightforward method of summarizing the behavior of a series. However, integrated processes are not the only way of modelling trends. The most important competitor is the standard deterministic trend, whereby, instead of differencing the series to induce stationarity, we first remove some deterministic function of time. As we have seen, an $I(1)$ series that is stationary in first-differences is called difference stationary. When $y_t - f(t)$ is stationary for a deterministic function $f(t)$, then we say that y_t is trend stationary. Since 0 is a deterministic function of time, trend-stationary series include stationary series as a special case. For the typical upward trending macroeconomic series, $f(t)$ will usually be linear or exponential, although there is nothing to rule out other possibilities. Note that a series that is stationary about a deterministic linear trend will have a first difference that appears to satisfy (49) (or equivalently (48) with $d = 1$). That it cannot in fact be written in this unit root form follows from the fact that the associated polynomial $B(L)$ would have a unit root, which is ruled out by the definition of (48).

The distinction between trend stationarity and difference stationarity can sometimes be made in terms of econometric convenience, but it is often a good deal more. The long-term behavior of the series and our ability to forecast it depends a good deal on which one is correct. If a series is trend stationary, it is tied to its deterministic trend from which it can never stray too far. However autocorrelated it might be in the short run, and however slowly it comes back to its anchor, come back it must. As a result, once we know the parameters of the process, and once the effects of the original position of the series relative to trend have worn off, our long-term forecasts will be forecasts of the trend. And since the trend is deterministic, there is no more uncertainty about its position in the far distant future than there is in the short or medium term.

For a difference stationary series, the position is quite different. Although the average rate of change per period is fixed, just as it is in the trend stationary case, there is nothing that ties the series to a particular position or particular trend line. For example, the random walk with drift in (49) increases at η per period on average, but in any given period it will increase by more or less depending on the random shock u_t. But once the new position has been attained, the series will increase at an expected rate of η per period from that new position, whatever was the route by which it got there. In particular, there is no non-stochastic trend to which the series is tied, and the series will eventually depart as far as we like from any that we try to delineate. As a result, even when we know the parameters of a difference stationary process, the uncertainty about its future position will grow steadily as we look further into the future. Put another way, because the series is an integrated process, the effects of shocks never wear off; disturbances have permanent effects.

The distinction between trend stationarity and difference stationarity means that in practical work it is necessary not only to estimate the parameters of a given type of model, but it is also necessary to choose a modelling strategy, and to decide which type of model to use. A number of test statistics have been developed to help make the choice. To take the simplest example, suppose that y_t is a trending series and that we consider the model

$$\Delta y_t = a + bt + \pi y_{t-1} + u_t \,. \tag{50}$$

When $b = \pi = 0$ (50) is a random walk with drift; alternatively, y_t is a stationary AR(1) around the linear trend $(1-\pi)^{-1}(a + bt)$. The random walk with drift model can therefore be treated as the null hypothesis in (50). However, and although the parameters in (50) are consistently estimated by OLS, conventionally calculated F-tests of the hypothesis that b and π are jointly zero do not have the F-distribution, even asymptotically. Special distributions must therefore be used, and in the case of (50), the appropriate critical values have been calculated and tabulated by Dickey and Fuller (1981).

Although the random walk with drift is a leading special case, we more frequently need to test for a unit root, which requires us to allow for the possibility that y_t is serially correlated even after first-differencing. To do so, (50) is augmented to read

$$\Delta y_t = a + bt + \pi y_{t-1} + \sum_{j=1}^{k} \theta_j \Delta y_{t-j} + u_t \tag{51}$$

where k is selected – typically by some model selection procedure – to be large enough to account for the serial correlation in the differences under the null. Once again, the unit root null can be tested by calculating a standard F-test, and once again the results must be checked against the critical values tabulated by Dickey and Fuller.

The recent literature on unit roots contains a large number of tests that are similar in purpose to those outlined above. There are a variety of different types of trends that can be used, from no trend at all, through polynomials in time, to nonlinear or "breaking" trends. Even in the cases given here, we could consider simply the t-test on π in place of the F-statistic on b and π jointly. These and other possibilities are reviewed by Campbell and Perron (1991) and by Mills (1990). However, the Dickey–Fuller tests shown above illustrate the general principles, and are sufficient background for me to discuss a number of practical implications.

The first point to note is that these tests take the unit root hypothesis as the null, so that if the test results in an acceptance, we have only failed to reject the unit root hypothesis, and we cannot legitimately claim that we have rejected trend stationarity. It is not possible to use these unit root tests to prove that time series have unit roots. The second issue is one of power. In most cases of interest, the data can be equally well represented by both difference stationary and trend stationary formulations. If π in (51) is nonzero but small, which is what the parameter estimates frequently show, it will be difficult to reject the hypothesis that there is a unit root, even though the series is in fact trend stationary. With enough data, it will always be possible to tell difference stationary and trend stationary models apart, but the distinction hinges on the long-run behavior of the series, behavior that can be very difficult to establish in finite samples.

It has become standard practice among some time-series econometricians and macroeconomists to accept the unit root hypothesis when it cannot be rejected, and then to follow the standard Box–Jenkins prescription of estimating a parsimonious (i.e. low-order) ARMA for the differenced series. In many cases, this will be a sensible descriptive strategy, but it is important to recognize it as such, and to note that there are cases where it can be quite misleading. Commodity prices again provide some instructive examples. When deflated by some suitable price index, such as the US consumer price index or an index of imports of manufactures by developing countries, the series are quite volatile but show relatively little trend over long enough periods. (What trends there are are typically negative as suggested by the Prebisch–Singer hypothesis [see for example Grilli and Yang (1988) and Ardeni and Wright (1992)]. The series also display very high first-order autocorrelations, typically 0.9 or higher even in annual data [see for example Deaton and Laroque (1992a) and Cuddington (1992)], and these autocorrelations typically decline only very slowly at higher orders. Such autocorrelation patterns are typically regarded as symptoms of non-stationarity, and indeed it is entirely plausible that commodity prices should be non-stationary over periods long enough to see changes in the technology of material use in production and consumption.

Even so, it comes as somewhat of a surprise to discover to what simple conclusions the unit root methodology leads. Cuddington (1992) looks at 26 commodity prices using long-run annual data, and find that unit roots cannot be rejected for half of them, and that in some cases, such as beef, copper, and rubber, a simple random walk with drift cannot be rejected. Looking at more recent data but at a monthly rather than annual frequency, Deaton (1992b) also finds that low-order unit root models provide an excellent fit to the data for cotton, copper, cocoa, and coffee prices. But these models make neither statistical nor economic sense. The price of annual crops like cotton are influenced by weather shocks that are known to be stationary, so that a model that asserts that all shocks are permanent is essentially absurd. Even with tree crops, where a frost may generate a price increase that can last for several years, the frost is not permanent and the trees will grow again, so that there is no reason to suppose that there will be *any* long term effect on the price. Nor do the data suggest that there are such long term effects, and in spite of the very high levels of volatility, the real prices of many commodities are remarkably close to what they were at the beginning of the century. What seems to be the case is that commodity prices are in fact tied to slowly evolving trends, perhaps deterministic but perhaps also stochastic with low volatility, and that the mechanism that brings them back to their trends operates only very slowly. Because the trend-reversion is small enough in any given month or year, the tests are unable to reject the unit root, and the first-differences of the series are well-approximated by low order autoregressive or moving average processes. But the statistical procedure has effectively discarded all the long-term information in the data, and the fitted process has quite different long-term properties from the data. For policy purposes, such errors can be very serious. If the copper price is a random walk, the income boom that accompanies a price boom can be treated as permanent, and it is appropriate for policy makers to engineer a consumption boom by the full amount of the income boom. But if prices are ultimately trend reverting, the income is ultimately transitory, and can support only a fraction of itself in permanently increased consumption.

A debate along these lines has characterized the estimation of unit roots using American macroeconomic data, where Campbell and Mankiw (1987) found that quarterly data on GDP showed persistence of macroeconomic shocks, while Cochrane (1988) found much less persistence using longer run annual data. Out of these debates came a useful and more direct measure of persistence that can be calculated with minimal parametric assumptions. Suppose that Δy_t is the (demeaned) first difference of the series, and that we write its moving average representation as

$$\Delta y_t = C(L)\varepsilon_t = \varepsilon_t + \sum_{j=1}^{\infty} c_j \varepsilon_{t-j} . \tag{52}$$

Note that (52) does not commit us to either trend or difference stationarity, or indeed to nonstationarity at all; if y_t is white noise, for example, $C(L) = 1 - L$. Campbell and Mankiw (1987) propose that γ defined by

$$\gamma = C(1) = 1 + \sum_{j=1}^{\infty} c_j \tag{53}$$

be used as a measure of persistence. Note that when the series is stationary or trend stationary, the $C(L)$ polynomial will have a unit root and γ will be zero. If the series is a random walk, γ will be unity, if positively correlated in first differences greater than unity, and so on. Note also that if the series is slowly trend reverting, with positive low order autocorrelations that are eventually succeeded by individually small negative ones, the infinite sum $C(1)$ will be much smaller than the sum of its first two terms.

Campbell and Mankiw's persistence measure is closely related to Cochrane's (1988) variance ratio, and both are closely related to the spectral density at frequency zero of the difference Δy_t. This opens the way for the estimation of persistence based on the standard range of non-parametric (window) estimates of the spectral density [see Campbell and Mankiw, Cochrane, and Priestly (1981) for the details]. In the work on quarterly measures of GDP, Campbell and Mankiw's calculations support the persistence measures that are estimated by low-order ARMA processes. However, in the commodity price case, the non-parametric persistence estimates are very much lower than those obtained from the parametric models, and in most cases are not significantly different from zero [see again Deaton and Laroque (1992a)]. Standard unit root modelling of models that are slowly trend reverting will generate a misleading picture of long-run behavior.

2.2.2. Multivariate issues in time-series models

The time-series issues associated with the presence of unit roots arise not only in univariate modelling, but also when we construct models containing several trending regressors. The issue arises most sharply in macroeconomic work, where there is a range of estimation and inference problems associated with the use of regression analysis to estimate relationships between trending variables. In this section, I give only the barest outline of the topic; I am unaware of any issues that are specific to development questions, and fuller treatments are readily available elsewhere [see again Campbell and Perron (1991) and Stock and Watson (1988)].

A useful starting point is a standard linear regression using time-series data,

$$y_t = \beta' x_t + u_t \,. \tag{54}$$

When all the variables in (54) are stationary, there are no non-standard issues of inference or estimation, so that we suppose some or all of the variables are non-stationary as indeed will be the case in most macroeconomic applications. The knowledge that there can be difficulties with such an apparently innocuous regression goes back to Yule (1926), with later refinements by Granger and Newbold (1974) and Phillips (1986). This is the literature on spurious regression, which is what happens when two essentially unrelated variables are regressed on one another, and appear to be related because, over the period of the sample, they exhibit trend-like behavior. Yule illustrated the point with the correlation at high frequencies of a sine and cosine wave that are orthogonal over frequencies longer than a complete cycle, while Granger and Newbold work with two independent random walks. They show that when two independently generated random walks are regressed on one another, a situation in which the OLS estimate of β converges to zero, there will often be spurious significance with apparently significant values of the t-statistic and of R^2. The very low Durbin–Watson statistics that typically accompany such regressions should indicate that something has gone wrong, although it is not hard to find examples in the literature where it is simply interpreted as evidence of positive autocorrelation, rather than of more fundamental difficulties. Even when investigators are more alive to the dangers, spurious regressions provide a good example of how standard distributional theory can break down in the time-series context.

A central concept in the multivariate analysis of trending variables is *cointegration*, introduced by Engle and Granger (1987). A vector of non-stationary variables z_t is said to be cointegrated when there exists at least one linear combination, $\gamma' z_t$ say, that is stationary. Clearly, if the regression (54) is to make sense, with u_t a zero mean stationary process, the composite vector

$(y_t x_t)$ must be cointegrated with cointegrating vector (1β). By contrast, two independent random walks are clearly not cointegrated, and their failure to be so is part of the problem in spurious regressions. The idea behind cointegration is that two (or more) variables are tied together in the long run, so that while in any given period the cointegrating relationship will never exactly be satisfied, the deviation is always within the bounds defined by a stationary distribution. Indeed, Engle and Granger show that when variables are cointegrated, there exist "error correction" representations of the relationships between them. In the case of two variables, this can be written

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta x_t + \alpha_2 (y_{t-1} - \beta x_{t-1}) + u_t \tag{55}$$

where β is the cointegrating parameter, and $y_t - \beta x_t$ is stationary. In many cases, (55) can be interpreted as a causal mechanism running from x to y, whereby y adjusts both to changes in x and to the previous "disequilibrium" in the long-run relationship. However, because (55) is a general consequence of cointegration, there can be many other interpretations.

The error-correction representation is also important because it highlights the consequences of a strategy that used to be recommended by time-series analysts, which is to difference non-stationary variables prior to regression analysis, so that standard inferential procedures can be applied. Equation (55) shows that when variables are cointegrated, a regression in differences ignores the long-run information in the data, and will generate results that do not reflect the long-run behavior of the series. This is analogous to the discussion in the previous section of what happens when a slowly trend-reverting series is modelled by applying a low-order ARMA to its difference. In both cases, the long-run behavior of the series is lost. The same result also applies to the analysis of a set of trending variables among which there may be one or more cointegrating relationships. The dynamics of such systems are often investigated in an atheoretical way by estimating vector autoregressions or VARs, where each variable is regressed on its own lags and those of the other variables. The long-run relationships will be lost if the VAR is estimated using differences.

Given the concept of cointegration, we need procedures for estimation and for inference. Tests for cointegration versus spurious regressions are typically based on the residuals from estimating the cointegrating regression by OLS. In the example of equation (54), we would first estimate by least squares, and then test whether the residuals have a unit root as would be the case if there is no cointegration and the regression is spurious. The recommended way of doing so is to use the augmented Dickey–Fuller test as described above, see equation (51) but without the time trend, and using the Dickey–Fuller tables to test whether the coefficient on the lagged residual is zero. Note once again the structure of the null and alternative hypotheses. Acceptance means that we cannot reject that the residuals have a unit root, which means that we cannot reject the null that there is no cointegration. A significant value of the test leads to the conclusion that the variables are cointegrated. Although the Durbin–Watson statistic may be helpful in indicating the possibility of a spurious regression, it should not be used as a formal test in place of the procedure outlined above.

There remains the question of how inference should be done in relationships like (54) and (55) given cointegration. In general, although OLS will yield consistent estimators, the distributional theory is non-standard, so that it is not possible to rely on the standard normal and χ^2 asymptotic theory. There are a number of alternative approaches. The first was suggested by Engle and Granger and relies on first estimating the cointegrating parameters in a levels regression ignoring any serial correlation in the errors, and then estimating the error-correction equation treating the cointegrating parameters as if they were known. Standard inference applies asymptotically to this second regression in spite of the fact that the cointegrating parameters are estimated not known, a result that hinges on Stock's (1987) proof that the parameters in the first stage regression converge at a rate proportional to $1/T$ rather than the usual $1/\sqrt{T}$. Unfortunately, this procedure does not give a way to make inferences about the cointegrating parameters themselves, nor does it appear to give satisfactory results for the other parameters in the sample sizes usually encountered in macroeconomic applications even in the US, let alone in LDCs, where macro data are often only annual and of relatively short duration.

Several theoretically superior methods are discussed by Campbell and Perron, although it seems that none of these have gone beyond the experimental stage into standard econometric usage. However, it is often possible to avoid the problems altogether. In a number of cases, the cointegrating relationships are "great ratio" relationships, so that a transformation to logarithms gives a cointegrating parameter of unity. If by this or some other argument the cointegrating parameters are known in advance, estimation is straightforward, using for example the error-correction representation. There are other contexts in which non-standard distributions can be avoided, and these have been explored by West (1988) and by Sims, Stock, and Watson (1988). In particular, even in the presence of cointegrating relationships, standard distributional theory applies to OLS estimates of parameters attached to stationary variables, or to parameters that could be attached to stationary variables in an appropriately rewritten regression. This is clearer in an example than it is to state. Consider the regression

$$y_t = \alpha + \beta y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + u_t \tag{56}$$

which with y as consumption and x as income could be an old-fashioned consumption function as included in innumerable macroeconometric models. This equation can be rewritten in either of the two forms

$$y_t = \alpha + \beta y_{t-1} + \gamma_0 \Delta x_t + (\gamma_0 + \gamma_1) x_{t-1} + u_t \tag{57}$$

or

$$y_t = \alpha + \beta y_{t-1} + (\gamma_0 + \gamma_1) x_t - \gamma_1 \Delta x_t + u_t \,. \tag{58}$$

If x_t is $I(1)$ so that its first difference is stationary, then (57) and (58) show that both of the γ-parameters can be attached to stationary variables, and that hypotheses concerning either or both can be tested using standard test procedures. Note however that in this example it is not possible to treat β in the same way.

2.2.3. The econometrics of dynamic programs

One of the many recent research programs in macroeconomics has been the use of representative agent dynamic programming models to characterize the economy. Although there is clearly scope for disagreement as to whether it makes sense to model the economy as a fully optimized system – indeed such a program would seem not to leave much scope for traditional development economics – the work has generated a good deal of new econometric tools including techniques for modelling aggregate time-series data. At the same time, there has been an increased interest in using dynamic programming models to help understand microeconomic data. In this section, I briefly review both topics with a focus towards applications in the development literature.

One topic that has been studied at both the macro and micro levels is saving, something that has always been seen as central to the understanding of the development process. Recent papers in development that interpret either micro or macro data using an explicit intertemporal optimization model include Giovannini (1985), Rossi (1988), Corbo and Schmidt-Hebbel (1991), Raut and Virmani (1989), Morduch (1990), and Atkeson and Ogaki (1990). Review papers that cover at least some of this work are Gersovitz (1988) and Deaton (1990b). Deaton (1992a) covers the literature from developed and developing countries, and develops the arguments below at much greater length.

Since the macroeconomic data are often modelled in terms of a single representative agent, we can use the same model to discuss both macro and micro applications. The usual version starts from an intertemporal utility function, whose expected value is the quantity that the agent seeks to maximize. This is typically written as

$$Eu_t = E_t \sum_{k=1}^{\infty} (1+\delta)^{-k} v(c_{t+k}) \tag{59}$$

where E_t indicates the mathematical expectation conditional on information at time t, δ is the rate of time preference, and the representative agent is assumed to live for ever. Utility is maximized subject to a budget constraint, conveniently written in terms of the evolution of a single asset

$$A_{t+1} = (1 + r_{t+1})(A_t + y_t - c_t) \tag{60}$$

where y_t is labor income – income excluding income from assets – and r_{t+1} is the rate of interest on funds carried from t to $t+1$ and is assumed not to be known in period t. A necessary condition for intertemporal optimality is the Euler equation which for this problem takes the form

$$v'(c_t) = E_t\left[\frac{1 + r_{t+1}}{1+\delta} v'(c_{t+1})\right]. \tag{61}$$

A good deal of econometric expertise has been devoted to estimating the parameters of (61) including the parameters of the marginal utility functions, as well as to testing its validity on both aggregate and microeconomic data.

There are several different econometric approaches to the analysis of Euler equations. One of the most straightforward is to select a functional form and to approximate the expectation by taking the first few terms of a Taylor series expansion. The leading example for (61) is to assume isoelastic utility so that the marginal utility is $c_t^{-\rho}$, in which case we can obtain the approximation, see for example Deaton (1992a, p. 64)

$$E_t \Delta \ln c_{t+1} \approx \frac{E_t r_{t+1} - \delta}{\rho} + \frac{\rho \, var_t(\Delta \ln c_{t+1} - \rho^{-1} r_{t+1})}{2} \tag{62}$$

so that, if the last term is small, as is plausibly the case for aggregate (but not micro) data, the Euler equation can be tested by regressing the change in consumption on the real interest rate. The term $E_t r_{t+1}$ can be replaced by r_{t+1} but should then be instrumented using instruments dated period t or earlier, a technique first suggested in the rational expectations literature by McCallum (1976). Variants of (62) are tested on aggregate consumption data for various LDCs by Rossi (1988), Giovannini (1985), and Raut and Virmani (1989).

A more formal estimation and testing technique avoids the approximation by using generalized methds of moments (GMM), an approach pioneered in this context by Hansen and Singleton (1982). Once again, use the isoelastic form for utility and rewrite (61) as

$$\frac{1+r_{t+1}}{1+\delta}c_{t+1}^{-\rho}-c_t^{-\rho}=\eta_{t+1} \tag{63}$$

where η_{t+1} is the difference between the outcome and its expectation, or innovation and as such is uncorrelated with any information dated t or earlier. In particular, if we have k such variables z_{it}, we have k conditional moment restrictions $E(\eta_{t+1}|z_{it}=0)$, whose sample counterparts are the k conditions, $j=1,\ldots k$,

$$T^{-1}d_j=\frac{1}{T}\sum_{t=1}^{T-1}z_{tj}\left[\frac{1+r_{t+1}}{1+\delta}c_{t+1}^{-\rho}-c_t^{-\rho}\right]=0\,. \tag{64}$$

If $k=2$, so that there are only two instruments, the parameters δ and ρ can be estimated in order to satisfy (64); otherwise, the model is overidentified, and the parameters are estimated by minimizing the quadratic form

$$d'(ZDZ)^{-1}d \tag{65}$$

where D is a diagonal matrix with the estimated η_t^2 on the diagonal, cf. (45) and (46) above.

The theory of GMM estimation requires that both sides of (63) be stationary, something that will typically require some form of detrending prior to estimation. The method also involves nonlinear estimation, which in some cases will make it less attractive than techniques based on the approximations. However, the technique is a clean one that is closely tied to the economics underlying the model, and since the criterion (65) is asymptotically distributed as χ^2 with $k-2$ degrees of freedom (there are k instruments and two parameters), GMM offers not only estimates, but a natural way of testing the overidentifying restrictions. However, it is not difficult by approximating (64) to show that GMM is closely related to the informal procedure, enforcing very much the same restrictions, and its overidentification test can be regarded as a test of whether the rate of growth of consumption is unpredictable except by variables that predict the real rate of interest.

The final approach to the Euler equation (61) is through a different set of assumptions, that the real rate of interest is constant and equal to the rate of time preference δ, and that preferences are quadratic, so that the marginal utility functions are linear. Together with (61) these assumptions imply that consumption is a martingale, so that the change in consumption is itself an innovation, unpredictable by earlier information

$$\Delta c_{t+1}=\eta_{t+1}\,. \tag{66}$$

The interest rate question has now been removed from consideration, but (66) can still be tested by finding out whether the change in consumption is in fact predictable by information available in period t, particularly information that is capable of predicting the change in income. This is the approach adopted by many writers in the US, and by Corbo and Schmidt-Hebbel (1991) for a range of Latin American countries together with Ghana, Pakistan, the Philippines, Thailand, and Zimbabwe. However, it is important to note that (66) is not rejected by finding that Δc_{t+1} is correlated with the contemporaneous change in income Δy_{t+1} since the latter will typically contain new information, and thus be correlated with the innovation η_{t+1}. An appropriate procedure, as with the interest rate in (62), is to regress the change in consumption on the contemporaneous change in income by instrumental variables using as instruments any variables dated t or earlier that help predict the next period's income change.

I have discussed these three methods in some detail, partly because the consumption issue is of such great importance, but also because these or similar techniques can be applied to a range of similar problems that generate stochastic intertemporal optimality conditions. Real business cycle models of macroeconomics provide one example; another occurs in finance, where arbitrage conditions are often interpretable as Euler equations, something that will happen quite generally when optimal intertemporal allocations are decentralizable by speculative behavior. For example, Deaton and Laroque (1992a) use GMM to estimate (a subset of) the parameters for the arbitrage conditions for speculative storage of primary commodities.

There are also a number of difficulties with these methods, some of which are general, and some of which are specific to the consumption example. Perhaps most serious is the inherent absurdity of modelling aggregate consumption as the optimal solution to a representative agent problem. Aggregation conditions for these type of models have been discussed (for example) by Grossman and Shiller (1982) and the conditions require (*a*) that people live for ever, (*b*) linearity in functional forms, and (*c*) that information about aggregate macroeconomic variables is known to everyone. In an economy of finitely-lived agents, it is possible for each person's consumption to grow (or to decline) in each year of life, but for aggregate consumption to be constant because the high (low) consumption levels of the old are constantly being replaced by the low (high) consumption levels of the young. The dynamics of individual consumption tells us nothing about the dynamics of aggregate consumption, and vice versa. There is some evidence from developed countries that functional form issues are important in aggregation, Attanasio and Weber (1991), and it is straightforward to use the failure of (*c*) to construct examples where the Euler equation is rejected at the aggregate level even though each agent in the economy is behaving as prescribed, see Pischke (1991) and Deaton (1992a, Chapter 5).

Aggregation problems are avoided by working with individual data, and there is a literature that looks at whether Euler equations like (61) describe the behavior of individual households, or whether there are deviations from (61) in the direction that would be predicted if households were credit constrained. One of the best known (and best) studies for the US is that by Zeldes (1989), and similar work has been undertaken for the South Indian ICRISAT households by Morduch (1990). At least to this writer, these studies make a good deal more sense than do the macroeconomic representative agent models discussed above. However, they are not without technical difficulties. First, these models typically require panel data, and their dynamic structure poses a number of delicate econometric issues that complicate both estimation and inference, see the next subsection. Second, because panel data are typically of short duration, it is impossible to apply the orthogonality conditions that identify these rational expectations models and provide the stuff of hypothesis tests. For example, equation (64) relates to a time average that will not be adequately estimated by two or three observations. In consequence, the panel data studies are forced to replace the time averages in (64) by cross-sectional averages across households or individuals. But the theory has no predictions for the cross-section averages, and in the presence of macroeconomic shocks that are common to many households, the theory can be rejected even when it is correct. Third, it is possible for consumers to be liquidity constrained but for the Euler equation to hold in almost all periods. In Deaton (1990b, 1991) I construct a model of a farmer in an LDC who faces i.i.d. income shocks, and who cannot borrow, but for whom the Euler equation (61) will fail only on the rare equations when it is optimal for him to spend down all his assets. In such circumstances, tests based on the failure of the Euler equation will lack power.

While Euler equations are necessary for optimality, they are not sufficient, so that estimates and tests based on them are less efficient, powerful, or informative than tests based on a complete characterization of behavior. In a few cases, such as the third example above where consumption is a martingale, explicit solutions are available, but this is exceptional and is purchased only at the price of strong assumptions. Recent work in empirical applications of stochastic dynamic programming has taken up this challenge, and attempts to solve for the functions that characterize behavior, not analytically, but numerically. The easiest way to see how this works is to look at an example, and to generalize from it. Since we have the notation at hand, the consumption example is a convenient one.

The basic idea is to have the computer work through the solution to the stochastic dynamic program, mimicking the possibilities that the agent will face. As always with dynamic programming, this is done by starting from the last period and working backwards. Consider then the consumption problem given by (59) and (60), but with a terminal period T. In this final period, since there is no bequest motive, there is an obvious solution to the consumption problem, which is to spend everything. Hence

$$c_T = g_T(A_T + y_T) = A_T + y_T \,. \tag{67}$$

We can think of this (trivial) solution as period T's decision rule, by which the decision variable c_t is related to the current state variables, here the sum of assets and labor income. This is a clumsy way to describe this obvious result, but the language is useful later.

Given last period behavior, we can begin to work backwards using the Euler equation. In period $T-1$, (61) implies that

$$v'(c_{T-1}) = E_{T-1}\left[\frac{1+r_T}{1+\delta} v'(y_T + (1+r_T)(A_{T-1} + y_{T-1} - c_{T-1}))\right]. \tag{68}$$

In order to solve (68) for consumption, we need an explicit functional form not only for the marginal utility function but also for the joint distribution of one period ahead interest rates conditional on information at $T-1$. For example, if this conditional distribution depends only on current values of income and interest rates, the solution to (68) will be of the form

$$c_{T-1} = g_{T-1}(A_{T-1}, y_{T-1}, r_{T-1}; \beta) \tag{69}$$

where β contains parameters of both preferences and the distribution function of interest rates and income. The solution (69) could contain more variables if for example income and interest rates are not first-order Markovian, or if utility depends on variables other than consumption, or less variables, if income and/or interest rates were i.i.d. In any case, there will usually not be an analytical solution for (69), although it will be possible to find it numerically given values of the state variables and of the parameters.

As we pursue the recursion back through time, the equations and the calculations become more complex, but the general principle remains the same. For any period t, there will exist some solution

$$c_t = g_t(s_t; \beta) \tag{70}$$

where s_t is a vector of state variables, for this problem, current and lagged values of income, assets, and interest rates, plus anything else that appears in preferences. There will also be a set of "updating" equations that relate the state variables to their earlier values; the budget constraint (60) is one, and others will represent the dynamic structure of the exogenous or forcing variables, here income and interest rates. Equation (70) is the "policy function" for time t, and is the equation that forms the basis for the econometric analysis of the relationship between consumption and the state variables. In some cases, the functions g_t may converge to a period independent or stationary policy function; this will typically require stationarity of the environment together with restrictions on preferences, conditions that have to be determined on a case by case basis.

This framework, together with a discussion of empirical applications has recently been reviewed by Rust (1992). Most of the empirical work, including Rust's (1987) own path-breaking study of engine replacement at the Madison, Wisconsin bus depot, has been concerned with cases where the choice is over a finite number of alternatives. In such a case, the econometric analysis proceeds as follows. Starting from a parametric form for preferences and for the laws of motion of the exogenous state variables, and a set of trial parameters, the computer first solves the decision problem. In the stationary case, and with discrete state variables, this essentially involves tabulating the function (70) for all possible values of its arguments conditional on the trial parameters. These solutions require iterative methods, and can be (extremely) computationally expensive for large problems with many choices and many state variables. Once the tabulation has been done, the actual values of the state variables can be inserted and the predicted choices compared with actuals. To avoid point predictions, at least one unobservable state variable is introduced, the distribution of which will induce probabilities on each of the outcomes, so that a likelihood function can be constructed in the usual way, the maximization of which is the outer loop of the estimation procedure.

Examples of the use of these techniques in the development literature include Wolpin's (1984) model of fertility using the Malaysian Family Life Survey, and Rosenzweig and Wolpin's (1993) examination of bullocks and pump-sets using the ICRISAT data. The latter study displays both the strengths and weaknesses of the general approach. The strength is the model's close links to the theoretical structure, especially the fact that solutions are genuine solutions of the dynamic problem rather than ad hoc intertemporal allocation rules whose costs and benefits are largely unknown. In principle, functional forms for preferences and distributions can be chosen for their plausibility and suitability to the problem at hand, here the agroclimatic and production conditions in the Indian semi-arid tropics. However, the costs are also high. Even with the use of supercomputers, it is necessary to keep down the number of state variables and to restrict the heterogeneity of individual agents, restrictions that can compromise the realism of the model. For example, Rosenzweig and Wolpin are forced for essentially computational reasons to have only a single asset, bullocks, so that all intertemporal transactions have to pass through the bullock market. There is no money, no jewelry, and no other assets. Note also that these models, by their heavy reliance on parametric functional forms, are extreme examples of the detailed structural specifications that are being increasingly discarded in other applications.

At least in simple cases, it is possible to implement dynamic stochastic programming models where the choice and state variables are continuous. An example is provided by Deaton and Laroque (1992b), who formulate and estimate a model of speculatively determined commodity prices. Under the assumptions that the commodity is an agricultural one whose harvests are i.i.d. from year to year, that demand is a linear function of price, that speculators will store the commodity whenever its one-period ahead expected price exceeds the current price by the costs of storage, and that storage cannot be negative, it is possible to write price as a function of the amount on hand, defined as this year's harvest plus any storage from the previous year. The observed price is thus a function of the unobserved state variable, the amount on hand, which is itself updated by the amount of the new harvest less consumption and the amount taken into storage by the speculators. In consequence, price follows a stationary but nonlinear first-order Markov process. As in the discrete case, estimation is done by nesting two sets of calculations, one to calculate the functions, and one to do the estimation. The function relating prices to the amount on hand can be characterized by a functional equation, the solution of which can be obtained by contraction mapping techniques for any given set of parameter values. The practical difficulty here is that instead of filling in a table, as in the discrete case, we have to solve for a continuous function, and much of the work is concerned with finding suitable discrete approximations that allow us to tabulate the function with interpolation methods (cubic splines) used to fill in the other values. Once the function has been obtained, the Markov process for prices is determined, and likelihood techniques are used to match outcomes to the data, with the results used to modify the parameters.

While these techniques are currently at the frontiers of computational complexity, they are no more difficult than were the now standard nonlinear estimation calculations when they were first explored twenty years ago. Calculations that require supercomputers now will be undertaken on notebook computers in only a few years. I suspect that the limitations to these methods lie not in the computation, but in their reliance on strong parametric assumptions. Robustness issues for both identification and estimation remain to be explored.

2.2.4. Time-series issues in panel data

In this final subsection, I return to the topic of panel data, and to the special problems that arise in using them to examine dynamic issues. I have two sets of applications in mind. The first is those studies that pool time-series from a cross-section of countries to explore the political and economic determinants of economic growth, see for example the references at the beginning of section 1.2. The second is provided by microeconomic studies of household or farm behavior, such as those using the ICRISAT panel data. The difference between the two lies less in methodology than in sample sizes. The country data usually have 20 to 30 years of observations, combined with a cross-section size of between 20 and 120, depending on how many countries are selected into the study. However, many authors work with five or ten year averages of growth rates, so that the time dimension is often reduced to three or four observations. The micro panel studies usually have many more observations in the cross-section but with a smaller number of years. The ICRISAT data, which covered 240 households in six villages for 10 years, have fewer observations and more time periods than is typically the case.

When panel data are used to confront dynamic issues, such as whether countries with low initial levels of GDP will subsequently grow faster, whether political changes cause economic changes or vice versa, or how assets affect current consumption, the regression equations will contain lagged dependent and predetermined variables. A useful model is the following:

$$y_{it} = \alpha + \beta y_{it-1} + \gamma' x_{it} + \eta_i + \varepsilon_{it} \tag{71}$$

where the x's are covariates whose degree of exogeneity and predeterminateness will be discussed, η_i is an individual (or country) specific effect that might or might not be correlated with the x's, and ε_{it} is an error term that will be treated as uncorrelated both across individuals and over time, assumptions than can be relaxed in special cases.

There is a substantial econometric literature on the estimation of (71) [see in particular Chamberlain (1984), Holtz-Eakin, Newey and Rosen (1988), Arellano and Bond (1991), and Schmidt, Ahn, and Wyhowski (1992)], the last three offering particularly clear and accessible treatments that are useful for practical work. As usual, the basic issue is correlation between right-hand side variables and error-terms, with the additional complications involved in dealing with fixed effects by differencing or using within estimators.

Consider first the case where there are no η's. There is no special problem with estimating (71) by OLS, and the model can be estimated using all the data without differencing or sweeping out means. Now suppose that there is individual heterogeneity, that we treat η_i as random, and that there is no lagged dependent variable. Provided the x's are uncorrelated with the random effects, this is a standard random-effects model of the Balestra and Nerlove (1966) variety, and estimation should be by (feasible) GLS. The problem here is that it is often quite implausible that the x's and the η's should be uncorrelated. In particular, if the x's are endogenous but predetermined, the random effects estimator will be inconsistent. For example, if the dependent variable is almost any aspect of household behavior and one of the x's is a measure of assets in any previous period, see for example Rosenzweig and Binswanger (1993), households with high η's will have high y's, and thus high or low x's. Because the η's are present in every period, they determine not only the current value of the dependent variable, but also all previous values, so that any feedback between dependent and independent variables will generate inconsistency, no matter how far lagged are the independent variables. Consistency of OLS (IVE) – or of the random effects estimator – requires that the x's (instruments) be strictly (strongly) exogenous, that they be uncorrelated with the compound error at all leads and lags. For the lagged dependent variable, strict exogeneity is logically impossible, so that OLS or GLS are automatically inconsistent in this case.

If the individual heterogeneity parameters are treated as fixed effects with no further structure, they must either be estimated – which is equivalent to sweeping out the individual means – or the data must be differenced to remove them. The former, whether implemented by estimating individual specific constants or by sweeping out means, is also inconsistent unless the explanatory variables are strictly exogenous. To see why, suppose that there is no lagged dependent variable, and that we sweep out the means from (71):

$$y_{it} - y_{i.} = \gamma'(x_{it} - x_{it-1}) + (\varepsilon_{it} - \varepsilon_{i.})\,. \tag{72}$$

The fixed effects have been removed, but only at the price of introducing the average of the ε's. Hence, if there is any feedback from y's to x's the x's will be correlated with the new compound error term. This is a problem of small numbers of time periods, since as T becomes large the time-average of the ε's will converge to its limit of zero. However, in many panel applications, the time dimension is small relative to the cross-section, and large n will not provide consistent estimates. It certainly makes more sense to use the within-estimators on cross-country data with 25 observations per country than it does to use them on a three period cross-sectional panel or on country data that has been averaged by decade.

How then should (71) be estimated? Holtz-Eakin, Newey and Rosen suggest that the equation be differenced to give

$$y_{it} - y_{it-1} = \beta(y_{it-1} - y_{it-2}) + \gamma'\Delta x_{it} + (\varepsilon_{it} - \varepsilon_{it-1}) \tag{73}$$

and that the parameters be estimated by instrumental variables using the fact that x's and y's lagged at least two periods are orthogonal to the differenced error terms. GMM approaches can also be used and will generate efficient estimates, see the elegant treatments in Arellano and Bond and in Schmidt et al. In practice, it may well be difficult to find instruments that are well-correlated with the right hand side variables in (73); especially in micro data, changes are often much harder to predict than are levels. In consequence, the within estimator may be worth considering in spite of its inconsistency, especially if there are more than a few time-periods.

2.3. Introduction to non-parametric methods

Many policy questions can be illuminated from survey data using only very straightforward concepts and techniques. In countries where there is little statistical information, the characterization of univariate and bivariate distributions from survey data is often extremely useful, in much the same way that simple unconditional forecasts are useful for time-series. Poverty measurement typically relies on univariate distributions of income and consumption, while questions of the distributional incidence of price changes depend on the joint distribution of living standards and net purchases. Simple multivariate relationships are often also useful, as when we disaggregate Engel curves by region or by family type. Basic "facts" such as these are routinely obtained from survey data using histograms, cross-tabulations, scatter diagrams, and linear regressions. While there is nothing wrong with any of these techniques, they are frequently not the best that can be done, particularly given the richness of survey data, and the typically large sample sizes. The numerical outputs of cross-tabulations and regressions are harder to communicate to policy makers or non-specialists than is graphical evidence, yet graphs of scatter diagrams are not very informative with very large samples, and plots of linear regressions have the opposite problem of compressing too much information into one graphic. Non-parametric estimation of densities and regressions is often useful in these situations, and allows the analyst to extract and display many more features of the data in a way that is sensitive to sample size. These techniques come in many varieties and flavors, from those that are obvious cousins of cross-tabulations and scatters, through non-parametric regression, to semi-parametric estimation, where a balance is struck between sample and prior information. There is a large and rapidly growing literature on these topics, and I only pick out a few development related applications. Readers interested in following these issues further should begin with the splendid book by Silverman (1986), and then draw from Bierens (1987), Härdle (1991), Stoker (1991), Powell (1992) and Härdle and Linton (1993).

2.3.1. *Estimating densities*

A histogram is the standard and most familiar non-parametric method for representing a univariate density, for example that of incomes in a household survey. For continuous data, however, the discrete steps in histograms usually do not have any counterpart in the underlying population, but are arbitrary artifacts of the selection of the discrete bins into which the observations are separated.

Instead of using a fixed bin, we might use a "sliding" bin or band, so that at each point x, say, we count the number of observations that fall in a symmetric band around x. If the sample points are $x_1, x_2, \ldots, x_N$, the estimated density, $\hat{f}_r(x)$ is given by

$$\hat{f}_r(x) = (2hN)^{-1} \sum_1^N I(x - h \leq x_i \leq x + h) \tag{74}$$

where $I(e)$ is the indicator function, which is unity if the statement e is true, and zero otherwise, and where the positive number h controls the width of the band. According to (74), the density at x is estimated by the fraction of the sample per unit length (the band is $2h$ long) that is close to (i.e. within h of) x. The procedure here is akin to taking a moving average in time series; rather than take the fraction of the sample *at* x, which can only be zero or j/N for positive integer j and so is very variable from point to point, we average over a nearby interval.

Like the time-series moving average, (74) has the disadvantage of being discontinuous at every point where an observation comes into or falls out of the band. As a result, if (74) is applied to a finite amount of data, the resulting estimator will have a "step" for every data point. This roughness can be removed by calculating a weighted average of the number of points near x, with maximum weight when points are at x, and with the weight steadily declining for points that are further away, and becoming zero for points that are just about to drop out of the band. We write this in the form:

$$\hat{f}_h(x) = \frac{1}{hN} \sum_1^N K\left(\frac{x - x_i}{h}\right) \tag{75}$$

where $K(z)$ is a positive symmetric function, so that $K(z) = K(-z)$, which is declining in the absolute value of its single argument, and which integrates to unity. The function $K(z)$ is a *kernel*, and (75) is a kernel estimator of the density. Note that the naive estimator (74) is a special case of (75) with $K(z) = 0.5I(|z| \leq\ = 1)$; for obvious reasons it is known as the rectangular kernel estimator. The intuition of (75) is exactly the same as that of (74); we are adding up the number of points per unit of line in an interval around x, but we use weights, that sum to unity but that weight more heavily the points near x.

There are several kernels that give good results in practice; three of the most commonly used are the Epanechnikov

$$K_e(z) = 0.75(1 - z^2)I(|z| \leq 1) . \tag{76}$$

the Gaussian

$$K_n(z) = (2\pi)^{-0.5} \exp(-0.5z^2) \tag{77}$$

and the quartic

$$K_q(z) = \tfrac{15}{16}(1 - z^2)^2 I(|z| \leq 1) . \tag{78}$$

For estimating densities (and regression functions) the choice between these (and other) kernels appears not to be of great importance, see Silverman (1986, p. 43). The Gaussian kernel does not require a separate line of code to detect whether the observation is or is not in the kernel, but many computers are unhappy exponentiating large negative numbers. The Epanechnikov and quartic kernels do not have this problem; the latter has the advantage of being everywhere differentiable, a property that is necessary if we want the estimated density – and later regression function – also to be differentiable, as will be the case in several of the applications discussed below.

More important than the choice of the kernel is the choice of the bandwidth. When the bandwidth is large, the estimator averages over a wide interval around each point, the estimated density will be relatively smooth, but we risk bias by missing genuine features of the underlying density. When the bandwidth is small, sample irregularities will be transmitted to the estimate, and while we will not miss real irregularities, and so will be protected against bias, we risk having too much variability. In any case, as the sample size increases, the bandwidth should shrink, so that with infinite amounts of data, the bandwidth will be zero, and we measure the fraction of the sample – and the population – at each data point. But the bandwidth must decrease more slowly than the sample size expands, so that the number of points in each band increases fast enough to ensure that consistent estimates are obtained at each point. If the bandwidth decreased in proportion to N^{-1}, there would be the same number of points in each band no matter what the sample size, and the variance of the estimate would never decrease. It turns out that it is optimal for the bandwidth to decrease in proportion to the fifth root of the sample size. Beyond that, there are a large number of methods of determining the constant of proportionality, each with its own adherents. Silverman (1986, pp. 45–46) suggests the following as a useful bandwidth when using a Gaussian kernel

$$h = 1.06 \min(\sigma, \chi/1.34) N^{-0.2}, \tag{79}$$

where σ is the standard deviation of the sample, and χ is its interquartile range, i.e. the difference between the 75th and 25th percentiles. The same formula can be used with the Epanechnikov and quartic kernels if the 1.06 is modified to 2.34 and 2.42 respectively.

Figure 33.4 shows densities for the logarithm of household income (deflated by a single price index in each year) in Taiwan for the years 1976 through 1990 inclusive. The underlying data come from the fifteen household surveys discussed in Section 1; there are 9,000–10,000 households in 1976 and 1977, and thereafter between 14,000 and 16,000. Each graph is calculated using (75) and a Gaussian kernel with bandwidth given by (79), and so varying from survey to survey. The densities are estimated at each point on a grid of 100 evenly spaced points for each survey; in order to draw graphs, it is unnecessary (and expensive) to estimate the density for each individual sample point. Note that the underlying distribution of incomes in levels is positively skewed, with observations much more densely distributed at low values than at high values. In such circumstances, it would be better to use smaller bandwidths at low incomes than at high. The logarithmic transformation is a simple way of avoiding having to do so, and the plots of the roughly symmetric densities are easier to see. The graphs show clearly the pattern of real income growth in

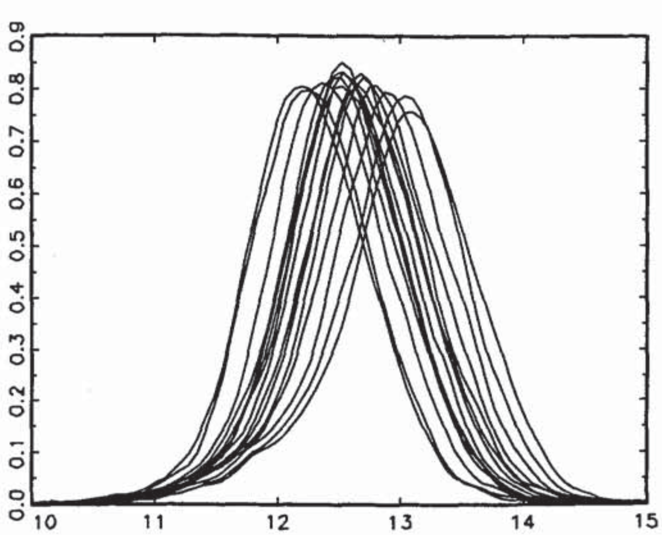

Figure 33.4. Non-parametric estimates of densities of log real income, Taiwan 1976–1990.

Taiwan, and modal real household income approximately doubles from 1976 to 1990. The graphs also show that the *shape* of the income distribution has been changing over time, with the lower tail stretching out as some households are left behind, and the density at the mode falling. Taiwan has had, and continues to have one of the world's most equal distributions of income, but inequality has been widening over the last decade [see also Government of Taiwan, (1989) and Deaton and Paxson (1994b)].

It is also possible to look at income distribution non-parametrically by plotting cumulative distribution functions or Lorenz curves. I have focussed here on density estimation, not because it is necessarily superior, but because the estimation of distribution functions or Lorenz curves using the individual data is straightforward because it is not usually necessary to smooth the estimates. The empirical estimator of the distribution function

$$\hat{F}(x) = N^{-1} \sum_{1}^{N} I(x_i \leq x) \tag{80}$$

is discontinuous for the same reasons as is the naive estimator, but for reasonable sample sizes, the discontinuities will not usually be apparent to the naked eye, and will not usually generate difficulties for the calculation of poverty or other measures. The same is true for Lorenz curves.

Non-parametric methods can also be used to estimate multivariate densities, although the "curse of dimensionality" means that very large numbers of observations are required when the dimensionality is high, see Silverman (1986, pp. 93–94). In practice, it is frequently useful to disaggregate by discrete variables, and to estimate densities by region, or by ethnic group, and these estimates present no new issues of principle. Estimates of the joint density of two continuous variables are practical and often informative, and can be constructed using the bivariate extensions of the methods given above. The standard procedure is to transform the data using its variance covariance matrix so as to make the scatter spherical, and then to make a weighted count of the number of observations in a circle around each point (x, y). The details are again in Silverman, pp. 65–80. To illustrate using the quartic kernel, the joint density estimator is

$$\hat{f}(z) = \hat{f}(x, y) = 3(\pi\sqrt{|S|}Nh^2)^{-1} \sum_{1}^{N} (1 - h^{-2}t_i^2)^2 I(|t_i| \leq h) \tag{81}$$

where S is the sample variance-covariance matrix, $|S|$ is its determinant, and t^2 is the quadratic form $(z - z_i)'S^{-1}(z - z_i)$, the squared distance of $z_i = (x_i, y_i)$ from z.

Figure 33.5 shows a contour map of the joint density of log consumption and log income using the Taiwanese data for 1979. This was calculated from 16,424

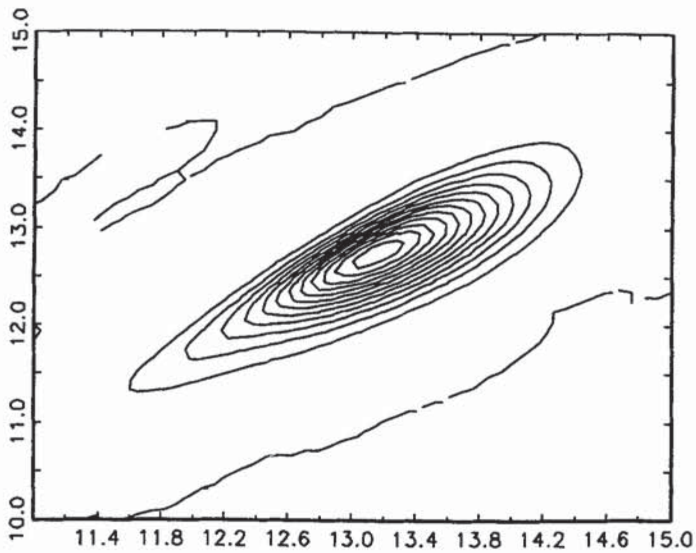

Figure 33.5. Non-parametric estimate of joint density of log income and log consumption, Taiwan 1990.

observations using (81), a 99 by 99 point grid, and a bandwidth of 1. (The calculations took a little less than an hour using GAUSS on a 66-Mhz PC.) Once again, the fact that the contours are not far from elliptical shows that the transformation to logs has induced something close to normality, not only for the marginals but also in the joint distribution of log income and log consumption. Even so, the contours are more egg-shaped than elliptical, so that although the regression function of log consumption on log income will be approximately linear, it will not be homoskedastic. The contour map is the preferred way of showing detail in the joint distribution, although for some purposes, a rapid appreciation of its general shape can be better seen from "net-maps", which provide a projection of the three dimensional object, see Deaton (1989) for net-maps of the joint distributions of living standards and net rice consumption across different regions of Thailand.

2.3.2. Non-parametric regression

Joint densities, such as that illustrated in Figure 33.5, contain the information that goes into regression functions, whether the conditional density of x on y, or the conditional density of y on x. More generally, the regression (function) of y on a vector $(x_1, x_2, \ldots, x_k)$ is defined as the expectation of y conditional on the x's, or

$$m(x_1, x_2, \ldots, x_k) = E(y/x_1, x_2, \ldots, x_k) .$$

If two variables x and y are jointly normally distributed, then both regression functions (of x on y and of y on x) will be linear and homoskedastic, but in general we cannot expect this to be the case. Indeed, there are a number of examples in development econometrics where nonlinearity has been an important issue, and perhaps the leading example is the relationship between nutrition and income. As emphasized by Ravallion (1990), it is likely that very poor people, who are close to subsistence, will spend a large fraction of any additional available resources on calories, but that this fraction will become much less as basic needs are met. As a result, the regression function of calories on income (or total expenditure) may be steeply sloped at low income levels, and flatten out as income increases. A transformation to logs may help, but there is no reason to suppose that the relationship is exactly of the form that would thereby become linear. The obvious procedure, especially when data are plentiful, is to estimate a non-parametric regression.

The procedures are closely analogous to those for density estimation, and the basic ideas are, if anything, even more transparent. With an infinite amount of data, the regression function of y on x could be calculated by averaging all the observations on y at each point x. With finite data, such a procedure is impractical, but we can use the same idea as in density estimation, and take an average over some interval around x. An early form of such non-parametric regression is Mahalanobis' (1960) fractile graphical analysis by which the x's are first sorted and then partitioned into fractile groups, such as deciles, so that averages of y can be calculated and plotted for each fractile. This method is analogous to the construction of histograms to represent densities, and like the latter can be improved by calculating averages for each x, and by weighting within the averages.

Alternatively, and more formally, we can write the regression function of y on x, $m(x)$ as

$$m(x) = \int yf(y|x)\,\mathrm{d}y = f(x)^{-1} \int yf(x, y)\,\mathrm{d}y \tag{83}$$

for which we can construct an estimate

$$\hat{m}(x) = \left[\frac{1}{Nh} \sum y_i K\left(\frac{x - x_i}{h}\right)\right] \div \left[\frac{1}{Nh} \sum K\left(\frac{x - x_i}{h}\right)\right] \tag{84}$$

which is known as the Nadaraya–Watson kernel regression estimator. Procedures for choosing the kernel and the bandwidth are similar to those in density estimation, though the kernel and bandwidth that optimally trade-off bias and variance for the density will generally not do so for the regression function. However, if the main object of the exercise is graphical representation of the regression function, a bandwidth can be selected by trial and error, with some preference for erring on the side of too small a bandwidth, since it is easier for the eye to smooth over irregularities than to uncover invisible features from an over-smoothed curve.

Note that since these methods estimate regression functions, and not the underlying structure, they require no modification for dichotomous or limited dependent variables. If the y's are 1 or 0, the regression function will be a conditional probability, and that is what the nonparametric regression will estimate. If the true model is Tobit, a linear model with censoring at zero, the nonparametric regression will deliver the Tobit regression function, equation (26).

One useful feature of (84) is that it allows straightforward estimation, not only of the regression function at each point, but of its derivatives, which are often of as much or more interest. Direct calculation gives:

$$\hat{m}'(x) = \Big(h \sum K(t_i)\Big)^{-1} \Big(\sum K'(t_i)(y_i - \hat{m}(x))\Big) \tag{85}$$

where $t_i = (x - x_i)/h$. These derivatives are readily calculated at the same time as the estimates of the regression function. Note however that, at the same bandwidth, the derivatives will be much less smooth than the regression as a function of x, so that higher bandwidths are typically desirable, see also Härdle (1991).

Figure 33.6, reproduced from Subramanian and Deaton (1992), shows the derivatives of the regression of the logarithm of calories on the logarithm of per capita household total expenditure for a sample of 5,600 households in rural Maharashtra in India. Since the regression function is in logs, the figure shows the elasticity of calories with respect to per capita total expenditure at each level of the latter. The results for two different bandwidths are shown; the smaller value is clearly too small and introduces what is almost certainly spurious fluctuations as total expenditure changes. Using the larger of the two bandwidths, there is some evidence of a slow decline as per capita expenditure rises, but not of any critical (or subsistence) point where the elasticity falls sharply. This is in contrast to Strauss and Thomas' (1990) results for Brazil, obtained using Cleveland's (1979) LOWESS estimator – an alternative nonparametric regression method – where elasticities are found to be much higher for poor than for rich households. The fact that the Indian households are much poorer may be relevant; perhaps they are all poorer than the cut-off

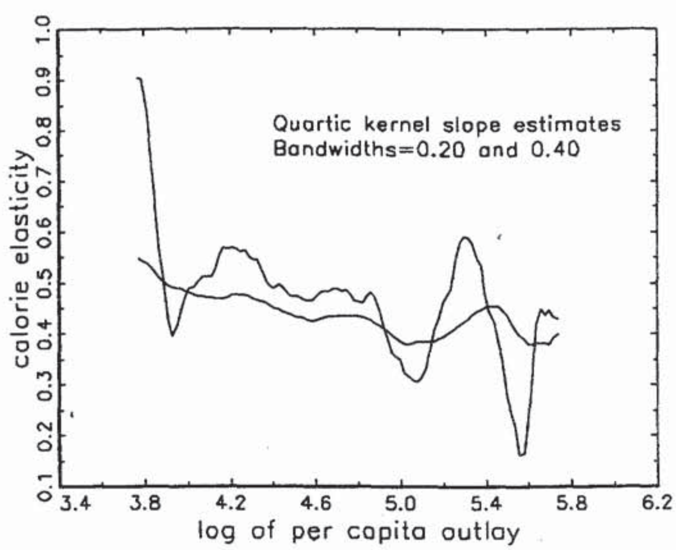

Figure 33.6. Elasticity of the calorie-expediture relationship with different bandwidths, Maharashtra, India 1983.

point in the Brazilian data. Note too that the elasticities in the figure are all a great deal higher than the figures that Bouis (1992) and Bouis and Haddad (1992) regard as sensible, perhaps because of the fact that nothing has been done to eliminate plate wastage, or food passed on to other people. These varying estimates are also fully consistent with that obtained by Behrman and Deolalikar (1987) using panel data from the ICRISAT data, from a nearby part of India. However, Behrman and Deolalikar's estimate of 0.37, based on a much smaller sample, and obtained from the first-differences of panel data, is not significantly different from zero, and the authors choose to interpret their result as showing that there is no relationship between calories and income.

2.3.3. Other non-parametric estimation methods

I have emphasized kernel regression methods in this presentation, but there are many other possibilities. I have already referred to Cleveland's (1979) LOWESS procedure, whereby the regression function is estimated by a series of local OLS regressions, appropriately weighted. Non-parametric regressions can also be estimated using splines, see Engle et al. (1986) and Härdle (1991), or by nearest-neighbor methods, which are similar to kernel methods, but the averaging is done, not over a fixed bandwidth centered on the point, but over the k nearest neighbors of the point, with k playing the role of the bandwidth. Such estimators are very similar to variable kernel estimators, with the bandwidth small in regions of high density, and large in regions of low. It is also possible to approximate regression functions using series expansions, with polynomial and Fourier series the obvious choices; for the latter see Gallant (1981), El-Badawi, Gallant, and Souza (1983), and Gallant and Souza (1991). Such estimators are convenient because they can be estimated by OLS, but can be dangerous with "dirty" survey data. A few outliers can have a major effect on fitted polynomials, while, with Fourier series, the regression function tends to "send out" tailored sine waves to pick up individual outliers. Kernel estimators have the advantage that observations beyond the bandwidth do not have any effect on the regression line at a point, so that the effect of outliers is limited. Kernel estimators of densities and of regression functions are also readily generalizable from the univariate and bivariate cases to multivariate analysis.

All of these nonparametric estimation techniques are most valuable in low dimensional situations, for one or two dimensional densities, or for regressions with at most two or three right hand side variables. In higher dimensions, data requirements are usually prohibitive, even by the standards of household survey data, and even armed with several million observations, it is unclear how one would display the results of a high dimensional density or regression. (Although programs such as *Mathematica* will draw three-dimensional netmaps with color variation representing the fourth dimension.) While there are many problems in development economics where the problems are low dimensional – calorie expenditure curves, Engel curves, and income distributions, for example – there are many other problems that require a different approach.

There has been a good deal of recent work on *semi-parametric* estimation techniques, where the idea is to mix parametric and non-parametric approaches, using prior structure where data are weak or scarce, and letting the data speak for themselves through non-parametric specifications where they are capable of doing so. One possibility is to write a regression function that is a mixture of linear and unspecified parts

$$E(y|x, z) = m(x, z) = \beta'x + f(z) \tag{86}$$

where z is a variable about which the data are likely to be quite informative, and x is a vector of variables where prior structure is required. For example, z might be income in an Engel curve, and x a set of sociodemographic variables whose effects are likely to be less well-defined in even a very large data set. Techniques for estimating (86) have been investigated by Robinson (1988).

2.3.4. Average derivatives and index models

Perhaps even more useful for many problems in empirical development is the *index* model, in which the regression function is written in the form

$$E(y|x) = m(x) = \theta(\beta' x) \tag{87}$$

so that while the standard linear structure is retained in the index $\beta' x$, the function $\theta(.)$ is left unspecified. This index model has the obvious attractions of mixing parametric and non-parametric specifications, something that is necessary if we are to bring some element of non-parametric methodology to higher dimensional problems. It also arises naturally in limited dependent variable models; for example, the regression function of Tobit – equation (26) – is of this form. Stoker (1991) shows that a range of other models – logit and probit models, truncation models, Box – Cox and Generalized Linear Models, and duration models – can all be written in the index form (87). While estimation of the parameters of (87), which are clearly identified only up to scale, will typically not provide estimates of all the parameters of these models, it can provide some information, and can do so without the distribution assumptions that are required for traditional maximum likelihood procedures.

Estimation of index models has been studied by Powell, Stock, and Stoker (1989), who show that it is possible to estimate the β's consistently and with convergence to the true values at the usual rate of root-N not at the much slower rates, typically the fifth root of N, that characterize non-parametric regressions. The basic idea comes from Stoker (1986). Write $f(x)$ for the marginal density of the x's, the conditioning variables in the regression, and define the vector of "scores" $z(x)$, by, $j = 1, \ldots, K$,

$$z_j = -\partial \ln f(x)/\partial x_j \tag{88}$$

so that, for each of the K x-variables, we have a value of its score for each observation. Consider then the population expectation of $z_j(x)y$, where the expectation is taken over both y and x. Note first that, from (88),

$$E(z_j y) = \int\int z_j(x) y f(y, x)\, \mathrm{d}y\, \mathrm{d}x = -\int\int f_j(x) y f(y|x)\, \mathrm{d}y\, \mathrm{d}x \tag{89}$$

where $f_j(x)$ is the jth partial derivative of the joint density. If the right hand side of (89) is integrated by parts, and if $f(x)$ is zero on the boundary of x, then

$$E(z_j y) = \int\int f(x) y f_j(y|x)\, \mathrm{d}y\, \mathrm{d}x = E_x\left(\frac{\partial E_y(y|x)}{\partial x_j}\right). \tag{90}$$

The last term on the right hand side of (90) is the vector of population average derivatives of the regression function, while the left hand side involves the scores z_j which can be estimated by the kernel methods discussed above; note that estimates of the derivatives of the joint densities can be obtained by differentiating (75) or its multivariate extension. Replacing population quantities by their estimates in (90) yields the *average derivative estimator*

$$\hat{\delta}_j = N^{-1} \sum_{i=1}^{N} \hat{z}_j(x_i) y_i \tag{91}$$

$\hat{\delta}_j$ will converge to the average over the population of the jth partial derivative of the regression function.

There are a number of complications and improvements to this basic idea that make it work better in practice. First, note that the scores are *logarithmic* derivatives of the joint density, and so will be very badly estimated where the joint density is small. This problem is dealt with by "trimming", which means dropping such observations from the sum in (91). Second, note from (90) that the $E(z_j x_k)$ is the derivative with respect to x_j of the expectation of x_j conditional on x, which is 1 if $j = k$, and is otherwise zero. As a result, if the scores are placed in an N by K matrix Z, the "instrumental variable estimator" $(Z'X)^{-1}(Z'y)$ also converges to the vector of average derivatives. To trim this, define the diagonal N by N matrix W by its typical element

$$w_{ii'} = \delta_{ii'} I(\hat{f}(x_i) > \alpha) \tag{92}$$

for some small positive number α. We then construct the "practical" estimator

$$\hat{\delta} = (\hat{Z}'WX)^{-1}(\hat{Z}'Wy) \tag{93}$$

where $\hat{Z}$ is the matrix of estimated scores.

Härdle and Stoker (1989) and Stoker (1991) discuss a number of other possible variants of this average derivative estimator. The remarkable thing about these estimators is not that they consistently estimate the average partial derivatives of the regression, but that they converge at the standard rate of $1/\sqrt{N}$. Although the kernel estimates may give very imprecise and slowly converging estimates of the scores, the averaging over the sample cancels out the variability in the components of the estimator, and allows the δ's to converge at the standard rate.

Consider first the application of average derivative estimators to the index models discussed above. If the expectation is given by (87), the average derivative estimator $\tilde{\delta}_j$ will converge to β_j multiplied by the average derivative of θ, a quantity that is independent of j. For such index models, Powell, Stock, and Stoker (1989) actually recommend a "density-weighted" average derivative estimator which differs from (93) in replacing the matrix W by one with the estimated densities on its diagonal.

Perhaps the most enticing potential application of average derivative estimators is to the case where the regression function is unspecified, in which case it offers a non-parametric estimate of the derivatives. A good example comes from the theory of tax reform, and its application to problems of pricing in LDCs. The basic material is discussed at length in Newbery and Stern (1987), Deaton (1987), and for India and Pakistan in Ahmad and Stern (1991).

A policy reform of increasing the price of good i is under consideration. The welfare cost of an infinitesimal price change Δp_i is the sum over each agent h of $\xi^h q_i^h \Delta p_i$, where q_i^h is *net* consumption of good i by h, and the weight ξ^h is the marginal social utility of money to individual h, independent of h for unrepentant Harbergians, see Harberger (1978, 1984), or varying with income (or ethnicity or region) if we want to apply distributional or other weights. The benefit of the price change is the value of additional revenue that it generates, so that the cost to benefit ratio of increasing the price can be written

$$\lambda_i = \frac{H^{-1} \Sigma_h \xi^h q_i^h}{H^{-1} \Sigma_h q_i^h + \Sigma_k t_k (H^{-1} \Sigma_h \partial q_k^h / \partial p_i)} \tag{94}$$

where t_k is the tax rate (or shadow tax rate, see Stern, 1987) on good k, and H is the population size. Commodities that have large λ-ratios are those through which it is (socially) harmful to raise government revenue, and whose prices should ideally be reduced, and vice versa for commodities with low λ-ratios.

The standard practice for evaluating tax reform proposals is to use household survey data to evaluate the equity effects in the numerator, weighting the consumption patterns of different groups by whatever importance policy-makers attach to their consumption at the margin. In practice this is usually done by using an Atkinson (1970) social welfare function, in which ξ^h is taken to be proportional to income (or per capita expenditure) to the power of $-\varepsilon$. Calculations are then done for a range of values of ε, with zero representing indifference to distribution, and larger values representing greater concern for the poor. Survey (or administrative) data can also be used to estimate mean consumption, the first term on the denominator, so that all but the second term in the denominator can be calculated without need of a parametric model. This last term is usually obtained by using time-series data or spatial variation in prices to estimate a demand system in which quantities are a function of prices, incomes, and household characteristics, and to calculate the responses using the estimated parameters, see for example Ahmad and Stern (1991) or Deaton and Grimard (1992) for alternative specifications applied to Pakistan. However, the term can be estimated without a parametric demand system using average derivative estimators. Not only can we avoid distributional assumptions about unobservables, but we do not even require a functional form for demand. In cases like this, and the general point applies to any average or weighted average of behavioral responses, the use of a functional form is essentially a detour, and an unnecessary one at that. Note too that in these sort of cases, there is no loss in precision from the non-parametric treatment, at least asymptotically.

The point made, a number of caveats ought to be entered. Practical experience with average derivative estimators in real situations is essentially nil, although see the experimental calculations in Deaton and Ng (1993). Although the asymptotic theory has been fully worked out, we do not know how difficult are the computational problems with data sets of several thousand observations, nor what is the best way to select the bandwidth for the estimation of the scores. Nor does the fact that these estimators are root-N consistent in itself guarantee that they will perform well in small samples. Indeed, the literature provides no guidance on what sample sizes are required to make the methods work. Nevertheless, these techniques are exceptionally promising, and provide a potential antidote to that body of empirical work that derives its results using restrictive functional forms and arbitrary distributional assumptions.

References

Ahluwalia, M.S. (1976) 'Inequality, poverty, and development', *Journal of Development Economics*, 3:307–342.

Ahluwalia, M.S., Duloy, J.H., Pyatt, G. and Srinivasan, T.N. (1980) 'Who benefits from economic development? Comment', *American Economic Review*, 70:242–245.

Ahmad, E. and Stern, N. (1991) *The theory and practice of tax reform in developing countries*, Cambridge. Cambridge University Press.

Ainsworth, M. (1992) 'Economic aspects of child fostering in Côte d'Ivoire', The World Bank, processed.

Ainsworth, M. and Muñoz, J. (1986) 'The Côte d'Ivoire living standards study', LSMS Working Paper 26, The World Bank, Washington, D.C.

Alesina, A. and Perotti, P. (1992) 'Income distribution, political instability, and investment', Washington, D.C. Institute for Policy Reform Working Paper No. 53, (September.)

Altonji, J., Hayashi, F. and Kotlikoff, L.J. (1992) 'Is the extended family altruistically linked? Direct tests using micro data', *American Economic Review*, 82:1177–1198.

Anand, S. and Ravi Kanbur, S.M. (1993) 'Inequality and development: a critique', *Journal of Development Economics*, 41:19–44.

Angrist, J.D. (1990) 'Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records', *American Economic Review*, 80:313–336.

Angrist, J.D. and Krueger, A.B. (1994) 'Why do World War II veterans earn more?' *Journal of Labor Economics*, 12:74–97.

Arabmazar, A. and Schmidt, P. (1981) 'Further evidence on the robustness of the Tobit estimator to heteroscedasticity', *Journal of Econometrics*, 17:253–258.

Arabmazar, A. and Schmidt, P. (1982) 'An investigation of the robustness of the Tobit estimator to non-normality', *Econometrica*, 50:1055–1063.

Ardeni, P.G. and Wright, B. (1992) 'The Prebisch – Singer hypothesis: a reappraisal independent of stationarity hypotheses', *Economic Journal*, 102:803–812.

Arellano, M. (1987) 'Computing robust standard errors for within-groups estimators', *Oxford Bulletin of Economics and Statistics*, 49:431–434.

Arellano, M. and Bond, S. (1991) 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies*, 58:277–297.

Arrow, K.J., 1974, 'The measurement of real value added', in P. David, and M. Reder, eds., *Nations and households in economic growth: essays in honor of Moses Abramovitz*, New York. Academic Press, 3–19.

Ashenfelter, O., Deaton, A. and Solon, G. (1986) 'Collecting panel data in developing countries: does it make sense?' LSMS Working Paper 23, The World Bank, Washington, D.C.

Atkeson, A. and Ogaki, M. (1990) 'Engel's law and saving', University of Chicago and University of Rochester (July), mimeo.

Atkinson, A.B. (1970) 'On the measurement of inequality', *Journal of Economic Theory*, 2:244–263.

Attanasio, O. and Weber, G. (1991) 'Consumption growth, the interest rate and aggregation', Stanford University and University College, London (June), mimeo. (*Review of Economic Studies*, forthcoming).

Balestra, P. and Nerlove, M. (1966) 'Pooling cross-section and time-series data in the estimation of a dynamic model: the demand for natural gas', *Econometrica*, 34:585–612.

Barro, R.J. (1991) 'Economic growth in a cross-section of countries', *Quarterly Journal of Economics*, 106:407–444.

Barro, R.J. and Sala-i-Martin, X. (1992) 'Convergence', *Journal of Political Economy*, 100:223–251.

Behrman, J.R. and Deolalikar, A.B. (1987) 'Will developing country nutrition improve with income? A case study for rural south India', *Journal of Political Economy*, 95:108–138.

Behrman, J.R. and Deolalikar, A.B. (1988) 'Health and nutrition', Chapter 14 in H. Chenery, and T.N. Srinivasan, eds., *Handbook of Development Economics*, 1:631–711.

Behrman, J.R. and Rosenzweig, M.R. (1994) 'The quality of aggregate inter-country, time-series data on educational investments and stocks, economically active populations, and employment', *Journal of Development Economics*, forthcoming.

Behrman, J.R. and Taubman, P. (1990) 'The intergenerational correlation between children's adult earnings and their parents' income: results from the Michigan panel survey of income dynamics', *The Review of Income and Wealth*, 36:115–127.

Behrman, J.R. and Wolfe, B.L. (1984) 'The socioeconomic impact of schooling in a developing country', *Review of Economics and Statistics*, 66:296–303.

Behrman, J.R. and Wolfe, B.L. (1989) 'Does more schooling make women better nourished and healthier?' *Journal of Human Resources*, 24:644–663.

Bell, C., Srinivasan, T.N. and Udry, C. (1992) 'Segmentation, rationing and spillover in credit markets: the case of rural Punjab', New Haven. Yale University, processed (February)

Benefo, K. and Schultz, T.P. (1993) 'Determinants of fertility and child mortality in Côte d'Ivoire and Ghana', New Haven. Yale University, processed (March)

Benjamin, D. (1993) 'Can unobserved land quality explain the inverse productivity relationship?' University of Toronto, processed. (February)

Benjamin, D. and Deaton, A. (1988) 'The Living Standards Survey and price policy reform: a study of cocoa and coffee production in Côte d'Ivoire', Living Standards Measurement Working Paper No. 44, Washington, D.C., The World Bank.

Berry, A.R. (1985) 'On trends in the gap between rich and poor in less developed countries: why we know so little', *Review of Income and Wealth*, 31:337–354.

Berry, A.R. and William R. Cline, 1979, *Agrarian structure and productivity in developing countries*, Baltimore. Johns Hopkins University Press.

Bevan, D., Collier, P. and Gunning, J.W. (1989) *Peasants and government: An economic analysis*, Oxford. Clarendon.

Bhalla, S.S. and Roy, P. (1988) 'Mis-specification in farm productivity analysis: the role of land quality', *Oxford Economic Papers*, 40:55–73.

Bhalla, S.S. and Roy, P. (1988) 'Mis-specification in farm productivity analysis: the role of land quality', *Oxford Economic Papers*, 40:55–73.

Bierens, H.J. (1987) 'Kernel estimators of regression functions', in: T.F. Bewley, ed., *Advances in Econometrics: Fifth World Congress*, 1:99–144.

Bouis, H.E. (1994) 'The effect of income on demand for food in poor countries: are our databases giving us reliable estimates?' *Journal of Development Economics*, forthcoming.

Bouis, H.E. and Haddad, L.J. (1992) 'Are estimates of calorie-income elasticities too high? A recalibration of the plausible range', *Journal of Development Economics*, 39:333–364.

Box, G.E.P. and Jenkins, G.M. (1970) *Time-series analysis: forecasting and control*, San Francisco: Holden–Day.

Breusch, T.S., and Pagan, A.R. (1979) 'A simple test for heteroskedasticity and random coefficient variation', *Review of Economic Studies*, 47:1287–1294.

Buchinsky, M. (1994), 'Changes in the U.S. wage structure 1963–1987: application of quantile regression; *Econometrica*, 62:405–465.

Buse, A. (1992) *'The bias of instrumental variables estimators'*, *Econometrica*, 60:173–179.

Campbell, J.Y. and Mankiw, N.G. (1987) 'Are output fluctuations transitory?' *Quarterly Journal of Economics*, 102:857–880.

Campbell, J.Y. and Perron, P. (1991) 'Pitfalls and opportunities: What macroeconomists should know about unit roots', in: O.J. Blanchard and S. Fischer, eds., *Macroeconomics Annual 1991*, Cambridge, MA: MIT Press for NBER, 141–201.

Card, D.E. (1991) 'Intertemporal labor supply: an assessment', Cambridge, MA: NBER Working Paper No. 3602, processed. (January)

Casley, D.J. and Lury, D.A. (1981) *Data collection in developing countries*, Oxford: Clarendon.

Chamberlain, G. (1984) 'Panel data', in: Z. Griliches and M.D. Intriligator, eds., *Handbook of Econometrics*, *Volume 2*, Amsterdam: North-Holland. 1247–1322.

Chamberlain, G. (1986) 'Asymptotic efficiency in semiparametric models with censoring', *Journal of Econometrics*, 32:189–192.

Chamberlain, G. (1991) 'Quantile regression, censoring, and the structure of wages', Cambridge, MA: Harvard University, processed, (June.)

Chamie, J. (1994) 'Population databases in development analysis', *Journal of Development Economics*, forthcoming.

Chayanov, A.V. (1966) *The theory of peasant economy*, Homewood, IL: Irwin. (Originally published in Russian, 1925.)

Chow, G.C. (1983) *Econometrics*, New York: McGraw-Hill.

Cleveland, W.S. (1979) 'Robust locally weighted regression and smoothing scatter plots', *Journal of the American Statistical Association*, 74:829–836.

Cochrane, J.H. (1988) 'How big is the random walk in GNP?' *Journal of Political Economy*, 96:893–920.

Corbo, V. and Schmidt-Hebbel, K. (1991) 'Public policy and saving in developing countries', *Journal of Development Economics*, 36:89–115.

Cramer, J.S., 1969, *Empirical econometrics*, Amsterdam: North-Holland.

Cuddington, J.T. (1992) 'Long-run trends in 26 primary commodity prices: a disaggregated look at the Prebisch–Singer hypothesis', *Journal of Development Economics*, 39:207–227.

Davidson, R. and MacKinnon, J.G. (1993) *Estimation and inference in econometrics*, Oxford: Oxford University Press.

Deaton, A.S. (1985) 'Panel data from time series of cross-sections', *Journal of Econometrics*, 30:109–126.

Deaton, A.S. (1987) 'Econometric issues for tax design in developing countries', in: D. Newbery and N. Stern, eds., *The theory of taxation for developing countries*, Oxford: Oxford University Press for The World Bank, 92–113.

Deaton, A.S. (1988) 'Quality, quantity and spatial variation in price', *American Economic Review*, 78:418–430.

Deaton, A.S. (1989) 'Rice prices and income distribution in Thailand: a non-parametric analysis', *Economic Journal*, 99:(Supplement) 1–37.

Deaton, A.S. (1990a) 'Price elasticities from survey data: extensions and Indonesian results', *Journal of Econometrics*, 44:281-309.

Deaton, A.S. (1990b) 'Saving in developing countries: theory and review', *World Bank Economic Review*, Special Issue, Proceedings of the First Annual World Bank Conference on Development Economics, 61–96.

Deaton, A.S. (1991) 'Saving and liquidity constraints', *Econometrica*, 59:1221–1248.

Deaton, A.S. (1992a) *Understanding consumption*, Oxford: Oxford University Press.

Deaton, A.S. (1992b) 'Commodity prices, stabilization and growth in Africa', Washington, D.C. Institute for Policy Reform Working Paper No. 40, processed. (September)

Deaton, A.S. (1994) *The analysis of household surveys: microeconomic analysis for development policy*, Research Program in Development Studies, Princeton University, processed. (423 pp.) (Apr.)

Deaton, A.S. and Grimard, F. (1992) 'Demand analysis for tax reform in Pakistan', Princeton, N.J. Princeton University, Research Program in Development Studies, processed. (January)

Deaton, A.S. and Laroque, G. (1992a) 'On the behavior of commodity prices', *Review of Economic Studies*, 59:1–23.

Deaton, A.S. and Laroque, G. (1992b) 'Estimating the commodity price model', Research Program in Development Studies, Princeton University, processed.

Deaton, A.S. and Ng, S. (1993) 'Parametric and non-parametric approaches to price and tax reform', Princeton, N.J. Princeton University, Research Program in Development Studies, processed. (March)

Deaton, A.S. and Paxson, C.H. (1994a) 'Saving, growth, and aging in Taiwan', in: D. Wise, ed., *Studies in the economics of aging*, Chicago: Chicago U. Press for National Bureau of Economic Research.

Deaton, A.S. and Paxson, C.H. (1994b) 'Intertemporal choice and inequality', *Journal of Political Economy*, 102:437–468.

Dickey, D.A. and Fuller, W.A. (1981) 'Likelihood ratio statistics for autoregressive time series with a unit root', *Econometrica*, 49:1057–1072.

Dumouchel, W.H. and Duncan, G.J. (1983) 'Using sample survey weights in multiple regression analysis of stratified samples', *Journal of the American Statistical Association*, 78:535–543.

Eicker, F. (1967) 'Limit theorems for regressions with unequal and dependent errors', in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, Berkeley, CA: University of California Press, 59–82.

El-Badawi, I., Gallant, A.R. and Souza, G. (1983) 'An elasticity can be estimated consistently without knowledge of functional form', *Econometrica*, 51:1731–1751.

Engle, R.F. (1984) 'Wald, likelihood ratio, and Lagrange multiplier tests in econometrics', in: Z. Griliches and M.D. Intriligator, eds., *Handbook of Econometrics, Volume 2*, Amsterdam: North-Holland, 775–826.

Engle, R.F. and Granger, C.W.J. (1987) 'Co-integration and error correction: representation, estimation, and testing', *Econometrica*, 55:251–276.

Engle, R.F., Rice, J. and Weiss, A. (1986) 'Semiparametric estimates of the relation between weather and electricity sales', *Journal of the American Statistical Association*, 81:310–320.

Evenson, R.E. and Pray, C.E. (1994) 'Measurement of food production and consumption', *Journal of Development Economics*, forthcoming.

Feder, G. (1985) 'The relation between farm size and farm productivity', *Journal of Development Economics*, 18:297–313.

Fields, G.S. (1977) 'Who benefits from economic development? A re-examination of Brazilian growth in the 60's', *American Economic Review*, 67:570–582.

Fields, G.S. (1992) 'Data for measuring poverty and inequality changes in the developing countries', Ithaca, N.Y.: Cornell University, processed. (June)

Fuller, W.A., 1975, 'Regression analysis for sample survey, *Sankhya*, Series C, 37:117–132.

Fuller, W.A. (1987) *Measurement error models*, New York: John Wiley.

Gallant, A.R. (1981) 'On the bias in flexible functional forms and an essentially unbiased form: the Fourier flexible form', *Journal of Econometrics*, 15:211–245.

Gallant, A.R. and Souza, G. (1991) 'On the asymptotic normality of Fourier flexible form estimation', *Journal of Econometrics*, 50:329–353.

Gersovitz, M. (1988) 'Saving and development', in: H. Chenery and T.N. Srinivasan, eds., *Handbook of Development Economics, Volume 1*, Amsterdam: North-Holland, 381–424.

Gertler, P. and van der Gaag, J. (1990) *The willingness to pay for medical care: evidence from two developing countries*, Baltimore, MD: Johns Hopkins University Press for The World Bank.

Giovannini, A. (1985) 'Saving and the real interest rate in LDCs', *Journal of Development Economics*, 18:197–217.

Goldberger, A.S. (1974) 'Unobservable variables in econometrics', in: P. Zarembka, ed., *Frontiers of Econometrics*, New York: Academic Press, 193–213.

Goldberger, A.S. 1983, 'Abnormal selection bias', in S. Karlin et al., eds., *Studies in econometrics, time series, and multivariate statistics*, New York: Academic Press.

Government of Thailand (1977) *Report of the 1975/6 Socio-Economic Survey: Whole Kingdom*, Bangkok: National Statistical Office.

Government of Thailand (1983) *Report of the 1981 Socio-Economic Survey: Whole Kingdom*, Bangkok: National Statistical Office.

Government of Taiwan (1989) *Report on the survey of personal income distribution in Taiwan area of the Republic of China*, Taipei: Directorate General of Budget, Accounting and Statistics Executive Yuan.

Granger, C.W.J. and Newbold, P. (1974) 'Spurious regressions in econometrics', *Journal of Econometrics*, 2:111–120.

Greenhalgh, S. (1982) 'Income units: the ethnographic alternative to standardization', *Population and Development Review*, 8:(Supplement), 70–91.

Grilli, E. and Yang, M.C. (1988) 'Primary commodity prices, manufactured goods prices, and the terms of trade in developing countries', *World Bank Economic Review*, 2:1–47.

Gronau, R. (1973) 'The effects of children on the housewife's value of time', *Journal of Political Economy*, 81:S168–99.

Grossman, S.J. and Shiller, R.J. (1982) 'Consumption correlatedness and risk measurement in economies with nontraded assets and heterogeneous information', *Journal of Financial Economics*, 10:195–210.

Grossman, J.B. (1994) 'Evaluating social policies: principles and U.S. experience', *World Bank Research Observer*, 9:159–180.

Haddad, L.J. and Ravi Kanbur, S.M. (1990) 'How serious is the neglect of intra-household inequality', *Economic Journal*, 100:866–881.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust statistics: the approach based on influence functions*, New York: Wiley.

Hansen, L.P. (1982) 'Large sample properties of generalized method of moments estimators', *Econometrica*, 50:1029–54.

Hansen, L.P. and Singleton, K.J. (1982) 'Generalized instrumental variables estimation of nonlinear rational expectations models', *Econometrica*, 50:1269–1286.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953) *Sample survey methods and theory*, 2, New York: Wiley.

Harberger, A.C. (1978) 'On the use of distributional weights in social cost benefit analysis', *Journal of Political Economy*, 86:S87–120.

Harberger, A.C. (1984) 'Basic needs versus distributional weights in social cost benefit analysis', *Economic Development and Cultural Change*, 32:455–474.

Härdle, W. (1991) *Applied nonparametric regression*, Cambridge: Cambridge University Press.

Härdle, W. and Stoker, T.M. (1989) 'Investigating smooth multiple regression by the method of average derivatives', *Journal of the American Statistical Association*, 84:986–995.

Härdle, W. and Linton, O. (1993) 'Applied nonparametric methods', Humbolt-Universität zu Berlin and Nuffield College Oxford, processed. (June), in: R.F. Engle and D. McFadden, eds., *Handbook of Econometrics, Volume 3* (1994), Amsterdam: North-Holland.

Hausman, J. A. (1978) 'Specification tests in econometrics', *Econometrica*, 46:1251–72.

Hayashi, F., Altonji, J. and Kotlikoff, L.J. (1991) 'Risk-sharing, altruism, and the factor structure of consumption', NBER working paper no. 3834 (September), processed.

Heckman, J.J. (1974) 'Shadow prices, market wages, and labor supply', *Econometrica*, 42:679–693.

Heckman, J.J. (1976) 'The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models', *Annals of Economic and Social Measurement*, 5:475–492.

Heckman, J.J. (1990) 'Varieties of selection bias', *American Economic Review* (paps. and procs.), 80:313–318.

Heckman, J.J. and Hotz, V.J. (1986) 'An investigation of the labor market earnings of Panamanian males: evaluating the sources of inequality', *Journal of Human Resources*, 21:507–542.

Heston, A.C. (1994) 'A brief review of some problems of using national accounts data in level comparisons and growth studies', *Journal of Development Economics*, forthcoming.

Holtz-Eakin, D., Newey, W. and Rosen, H. (1988) 'Estimating vector autoregressions with panel data', *Econometrica*, 56:1371–1395.

Huber, P.J. (1967) 'The behavior of maximum likelihood estimates under non-standard conditions', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233.

Huber, P.J. (1981) *Robust statistics*, New York: John Wiley.

Jain, S. (1975) *Size distribution of income: a compilation of data*, Washington, D.C.: The World Bank.

Jimenez, E. (1987) *Pricing policy in the social sectors: cost recovery for education and health in developing countries*, Baltimore: Johns Hopkins University Press for The World Bank.

Jöreskog, K.G. (1973) 'A general method of estimating a linear structure equation system', in: A.S. Goldberger and O.D. Duncan, eds., *Structural equation models in the social sciences*, New York: Seminar Press, 85–112.

Kennan, J. (1989) 'Simultaneous equations bias in disaggregated econometric models', *Review of Economic Studies*, 56:151–156.

Kish, L. (1965) *Survey sampling*, New York: John Wiley.

Kish, L. and Frankel, M.R. (1974) 'Inference from complex samples' (with discussion), *Journal of the Royal Statistical Society*, Series B, 36:1–37.

Kloek, T. (1981) 'OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated', *Econometrica*, 49:205–207.

Knight, J.B. and Sabot, R.H. (1990) *Education, productivity, and inequality: the East African natural experiment*, Oxford: Oxford University Press for The World Bank.

Koenker, R. and Bassett, G. (1978) 'Regression quantiles', *Econometrica*, 46:33–50.

Koenker, R. and Bassett, G. (1982) 'Robust tests for heteroscedasticity based on regression quantiles', *Econometrica*, 50:43–61.

Kravis, I.B., Heston, A. and Summers R. (1978) *International comparisons of real product and purchasing power*, Baltimore: Johns Hopkins University Press.

Leamer, E.E. (1978) *Specification searches: ad hoc inference from non-experimental data*, New York: Wiley.

Levine, R. and Renelt, D. (1992) 'A sensitivity analysis of cross-country growth regressions', *American Economic Review*, 82:942–963.

Liu, P. (1982) 'Determinants of income inequality over the family development cycle: the case of Taiwan', *Population and Development Review*, 8 (Supplement), 53–69.

Maddala, G.S. and Jeong, J. (1992) 'On the exact small sample distribution of the instrumental variable estimator', *Econometrica*, 60:181–183.

Mahalanobis, P.C. (1944) 'On large scale sample surveys', *Philosophical Transactions of the Royal Society*, Series B, 231:329–451.

Mahalanobis, P.C. (1946) 'Recent experiments in statistical sampling in the Indian Statistical Institute', *Journal of the Royal Statistical Society*, 109:325–370.

Mahalanobis, P.C. (1960) 'The method of fractile graphical analysis', *Econometrica*, 28:325–351.

Manski, C.F. (1988) 'Identification of binary response models', *Journal of the American Statistical Association*, 83:729–738.

Manski, C.F. and Lerman, S. (1977) 'The estimation of choice probabilities from choice-based samples', *Econometrica*, 45:1977–1988.

Mankiw, N.G., Romer, D. and Weil, D. (1992) 'A contribution to the empirics of economic growth', *Quarterly Journal of Economics*, 107:407–437.

Marschak, J. and Andrews, W.H. (1944) 'Random simultaneous equations and the theory of production', *Econometrica*, 12:143–205.

McCallum, B.T. (1976) 'Rational expectations and the natural rate hypothesis: some consistent estimates', *Econometrica*, 44:43–52.

Mills, T. (1990) *Time series techniques for economists,* Cambridge: Cambridge University Press.
Morduch, J. (1990) 'Risk, production, and saving: theory and evidence from Indian households', Harvard University, processed.
Moulton, B.R. (1990) 'An illustration of a pitfall in estimating the effects of aggregate variables on micro units', *Review of Economics and Statistics*, 72:334–338.
Mundlak, Y. (1961) 'Empirical production function free of management bias', *Journal of Farm Economics*, 43:44–56.
Murthy, M.N. (1977) *Sampling theory and methods*, Calcutta. Statistical Publishing Society. (2nd ed.)
Nadaraya, E.A. (1964) 'On estimating regression', *Theory of Probability and Applications*, 10:186–190.
Nagar, A.L. (1959) 'The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations', *Econometrica*, 27:575–595.
Nelson, C. and Startz, R. (1990a) 'The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one', *Journal of Business*, S125–140.
Nelson, C. and Startz, R. (1990b) 'Some further results on the exact small sample properties of the instrumental variable estimator', *Econometrica*, 58:967–976.
Newbery, D., and Stern, N. eds. (1987) *The theory of taxation for developing countries*, Oxford: Oxford University Press for The World Bank.
Newey, W.K. (1986) 'Maximum likelihood specification testing and conditional moment tests', *Econometrica*, 53:1047–1070.
Newey, W.K., Powell, J.L. and Walker, J.R. (1990) 'Semiparametric estimation of selection models: some empirical results', *American Economic Review*, 80:324–328.
Newman, J., Gertler, P. and Rawlings, L. (1994) 'Using randomized central designs in evaluating social sector programs in developing countries', *World Bank Research Observer*, 9:181–201.
Persson, T. and Tabellini, G. (1991) 'Is inequality harmful for growth?' Cambridge, MA: NBER Working Paper No. 3599, processed. (January)
Pfefferman, D. and Smith, T.M.F. (1985) 'Regression models for grouped populations in cross-section surveys', *International Statistical Review*, 53:37–59.
Phillips, P.C.B. (1980) 'The exact distribution of instrumental variables estimators in an equation containing $n+1$ endogenous variables', *Econometrica*, 48:861–878.
Phillips, P.C.B. (1986) 'Understanding spurious regression', *Journal of Econometrics*, 33:311–340.
Pischke, J.-S. (1991) 'Individual income, incomplete information, and aggregate consumption', Industrial Relations Section Working Paper no. 289, Princeton University, mimeo.
Pitt, M.M., Rosenzweig, M.R. and Gibbons, D.M. (1993) 'The determinants and consequences of the placement of government programs in Indonesia', *World Bank Economic Review*, 7:319–348.
Pitt, M.M., Rosenzweig, M.R. and Hassan, M.N. (1990) 'Productivity, health and inequality in the intrahousehold distribution of food in low-income countries', *American Economic Review*, 80:1139–1156.
Pitt, M.M. and Sumodiningrat, G. (1991) 'Risk, schooling and the choice of seed technology in developing countries: a meta-profit function approach', *International Economic Review*, 32:457–473.
Powell, J.L. (1984) 'Least absolute deviations estimation for the censored regression model', *Journal of Econometrics*, 25:303–325.
Powell, J.L. (1992) 'Estimation of semiparametric models', Princeton, N.J.: Princeton University, processed. (September)
Powell, J.L., Stock, J.H. and Stoker, T.M. (1989) 'Semiparametric estimation of index coefficients', *Econometrica*, 57:1403–1430.
Prais S. and Houthakker, H.S. (1955) *The analysis of family budgets*, Cambridge: Cambridge University Press.
Priestly, M.B. (1981) *The spectral analysis of time-series*, London: Academic Press.
Raut, L. and Virmani, A. (1989) 'Determinants of consumption and savings behavior in developing countries', *World Bank Economic Review*, 3:379–393.
Ravallion, M. (1990) 'Income effects on undernutrition', *Economic Development and Cultural Change*, 38:489–515.
Robinson, P.M. (1988) 'Root-*N*-consistent semiparametric regression', *Econometrica*, 56:931–954.

Rosenzweig, M.R. (1990) 'Population growth and human capital investments: theory and evidence', *Journal of Political Economy*, 98:S38–70.
Rosenzweig, M.R. and Binswanger, H. (1993) 'Wealth, weather risk and the composition and profitability of agricultural investments', *Economic Journal*, 103:56–78.
Rosenzweig, M.R. and Schultz, T.P. (1983) 'Estimating a household production function: heterogeneity, the demand for health inputs, and their effects on birth weight', *Journal of Political Economy*, 91:723–46.
Rosenzweig, M.R. and Schultz, T.P. (1987) 'Fertility and investments in human capital: estimates of the consequences of imperfect fertility control', *Journal of Econometrics*, 36:163–184.
Rosenzweig, M.R. and Wolpin, K.J. (1980) 'Testing the quantity-quality model of fertility: results of a natural experiment – twins', *Econometrica*, 48:227–240.
Rosenzweig, M.R. and Wolpin, K.J. (1986) 'Evaluating the effects of optimally distributed public programs', *American Economic Review*, 76:470–487.
Rosenzweig, M.R. and Wolpin, K.J. (1988) 'Migration selectivity and the effects of public programs', *Journal of Public Economics*, 37:265–289.
Rosenzweig, M.R. and Wolpin, K.J. (1993) 'Credit market constraints, consumption smoothing, and the accumulation of durable production assets in low-income countries: investments in bullocks in India', *Journal of Political Economy*, 101:223–244.
Rossi, N. (1988) 'Government spending, the real interest rate, and the behavior of liquidity-constrained consumers in developing countries', *IMF Staff Papers*, 35:104–140.
Rozansky, J. and Yeats, A. (1994) 'On the (in)accuracy of economic observations: an assessment of trends in the reliability of international trade statistics', *Journal of Development Economics*, forthcoming.
Rust, J. (1987) 'Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher', *Econometrica*, 55:999–1033.
Rust, J. (1992) 'Do people behave according to Bellman's principle of optimality?' Stanford, CA: Hoover Institution, processed. (May)
Schmidt, P., Ahn, S.C. and Wyhowski, D. (1992) 'Comment', *Journal of Business and Economic Statistics*, 10:10–14.
Schwarz, G. (1978) 'Estimating the dimensions of a model', *Annals of Statistics*, 30:461–464.
Scott, A.J. and Holt, D. (1982) 'The effect of two-stage sampling on ordinary least squares methods', *Journal of the American Statistical Association*, 77:848–854.
Sen, A.K. (1962) 'An aspect of Indian agriculture', *Economic Weekly*, 14, Annual Number.
Sen, A.K. (1966) 'Peasants and dualism with and without surplus labor', *Journal of Political Economy*, 74:425–450.
Sen, A.K. (1975) *Employment, technology, and development*, Oxford: Clarendon Press.
Sewell, W.H. and Hauser, R.M. (1975) *Education, occupation and earnings: Achievement in the early career*, New York: Academic Press.
Silverman, B.W. (1986) *Density estimation for statistics and data analysis*, London: Chapman and Hall.
Sims, C.A. (1969) 'Theoretical basis for a double deflated index of real value added', *Review of Economics and Statistics*, 51:470–471.
Sims, C.A., Stock, J.H. and Watson, M.W. (1988) 'Inference in linear time-series models with some unit roots', *Econometrica*, 58:113–144.
Singh, I., Squire, L. and Strauss, J. eds. (1986) *Agricultural households models: extensions, applications, and policy*, Baltimore, MD: Johns Hopkins University Press for The World Bank.
Smith, J.P., Thomas, D. and Karoly, L.A. (1992) 'Migration in retrospect: differences between men and women', Santa Monica, CA: Rand Corporation, processed. (April)
Solon, G.R. (1992) 'Intergenerational income mobility in the United States', *American Economic Review*, 82:393–408.
Srinivasan, T.N. (1972) 'Farm size and productivity – implications of choice under uncertainty', *Sankhya*, Series B, 34:409–420.
Srinivasan, T.N. (1994) 'Data base for development analysis: an overview', *Journal of Development Economics*, forthcoming.
Stelcner, M., van der Gaag, J. and Wijverberg, W. (1989) 'A switching regression model of public – private sector wage effects in Peru, 1985–6', *Journal of Human Resources*, 24:545–559.

Stern, N. (1987) 'Aspects of the general theory of tax reform', Chapter 3 in: D. Newbery and N. Stern, eds., *The theory of taxation for developing countries*, Oxford: Oxford University Press for the World Bank.

Stigler, G.J. (1954) 'The early history of empirical studies of consumer behavior', *Journal of Political Economy*, 42:95–113.

Stock, J.H. (1987) 'Asymptotic properties of least-squares estimators of cointegrating vectors', *Econometrica*, 55:1035–1056.

Stock, J.H. and Watson, M.W. (1988) 'Variable trends in economic time-series', *Journal of Economic Perspectives*, 2:147–174.

Stoker, T.M. (1986) 'Consistent estimation of scaled coefficients', *Econometrica*, 54:1461–1481

Stoker, T.M. (1991) *Lectures on semi-parametric econometrics*, Louvain-la-Neuve, Core Foundation.

Strauss, J. and Thomas, D. (1990) 'The shape of the calorie expenditure curve', New Haven: Yale University, processed.

Strauss, J., Gertler, P., Rahman, O. and Fox, K. (1992) 'Gender and life-cycle differentials in the patterns and determinants of adult health', RAND Corporation and Ministry of Health, Government of Jamaica, processed. (June)

Strauss, J. and Gertler, P. (1993) 'Human resources: household decisions and markets', *this volume*.

Subramanian, S. and Deaton, A. (1992) 'The demand for food and calories: further evidence from India', Princeton, N.J.: Princeton University, Research Program in Development Studies, processed. (February)

Summers, R. and Heston, A.C. (1991) 'The Penn World Table (Mark 5): an expanded set of international comparisons, 1950–1988', *Quarterly Journal of Economics*, 106:327–368.

Tan, C.C.H. (1991) 'Drawing out the life cycle of Korean households with synthetic panels of rational consumers', Yale University and University of Hong Kong, processed. (November)

Tobin, J. (1958) 'Estimation of relationships for limited dependent variables', *Econometrica*, 26:24–36.

van der Gaag, J., Stelcner, M. and Wijverberg, W. (1989) 'Wage differentials and moonlighting by civil servants: evidence from Côte d'Ivoire and Peru', *World Bank Economic Review*, 3:67–95.

Watson, G.S. (1964) 'Smooth regression analysis', *Sankhyā*, Series A, 26:359–372.

West, K.D. (1988) 'Asymptotic normality when regressors have a unit root', *Econometrica*, 56:1397–1418.

White, H., and MacKinnon, J.G. (1985) 'Some heteroskedasticity-consistent covariance matrix estimates with improved finite sample properties', *Journal of Econometrics*, 29:305–325.

White, H. (1980) 'A heteroskedasticity consistent covariance matrix estimatorand a direct test for heteroskedasticity', *Econometrica*, 48:817–838.

White, H. (1984) *Asymptotic theory for econometricians*, New York: Academic Press.

Wolpin, K.J. (1984) 'An estimable dynamic stochastic model of fertility and child mortality', *Journal of Political Economy*, 92:852–874.

Yule, G.U. (1926) 'Why do we sometimes get nonsense correlations between time-series?' *Journal of the Royal Statistical Society*, 89:1–64.

Zeldes, S.P. (1989) 'Consumption and liquidity constraints: an empirical investigation', *Journal of Political Economy*, 97:305–346.

Zimmerman, D.J. (1992) 'Regression toward mediocrity in economic stature', *American Economic Review*, 82:409–429.