NBER TECHNICAL PAPER SERIES

DATA PROBLEMS IN ECONOMETRICS

Zvi Griliches

Technical Working Paper No. 39

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 1984

The research reported here is part of the NBER's research program in Productivity. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Technical Paper #39
July 1984

Data Problems in Econometrics

ABSTRACT

This review of data problems in econometrics has been prepared for the Handbook of Econometrics (Vol. 3, Chap. 25, forthcoming). It starts with a review of the ambivalent relationship between data and econometricians, emphasizing the largely second-hand nature of economic data and the consequences that flow from the distance between econometricians as users of data and its producers. Section II describes the major types of economic data while Section III reviews some of the problems that arise in trying to use such data to estimate model parameters and to test economic theories. Section IV reviews the classical errors in variables model and its applicability to micro-data, especially panel data. Section V discusses missing data models and methods and illustrates them with an empirical example. Section VI focuses on the problem of estimating models in the absence of a full history, suggests a possible range of solutions, and provides again an empirical example: using a short panel to investigate the weights to be used in constructing a correct "capital" measure. The chapter closes (Section VII) with some final remarks on the existential problem of econometrics: life with imperfect data and inadequate theories.

Zvi Griliches
National Bureau of
Economic Research
1050 Massachusetts Ave.
Cambridge, MA. 02138

Zvi Griliches

Handbook of Econometrics
Volume III, Chapter 25
Revised July 1984

DATA PROBLEMS IN ECONOMETRICS

Table of Contents

Handbook of Econometrics
Volume III, Chapter 25
Revised July 1984

DATA PROBLEMS IN ECONOMETRICS*

Zvi Griliches

I. Introduction Data and Econometricians -- The Uneasy Alliance

> Then the officers of the children of Israel came and cried unto Pharaoh, saying, Wherefore dealest thou thus with thy servants?
> There is no straw given unto thy servants, and they say to us, Make brick: and behold thy servants are beaten; but the fault is in thine own people.
> But he said, Ye are idle, ye are idle: Therefore ye say, Let us go and do sacrifice to the Lord.
> Go therefore now, and work; for there shall no straw be given you, yet shall ye deliver the tale of bricks.
>
> Exodus 5, 15-18

Econometricians have an ambivalent attitude towards economic data. At one level, the "data" are the world that we want to explain, the basic facts that economists purport to elucidate. At the other level, they are the source of all our trouble. Their imperfection makes our job difficult and often impossible. Many a question remains unresolved because of "multicollinearity" or other sins of the data. We tend to forget that these imperfections are what gives us our legitimacy in the first palce. If the data were perfect, collected from well designed randomized experiments, there would be hardly room for a separate field of econometrics.

*I am indebted to the National Science Foundation (SOC78-04279 and PRA81-08635) for their support of my work on this range of topics, to John Bound, Bronwyn Hall, J.A. Hausman, and Ariel Pakes for research collaboration and many discussions, and to E. Berndt, F.M. Fisher, R.M. Hauser, M. Intriligator, S. Kuznets, J. Medoff, and R. Vernon for comments on an earlier draft.

Given that it is the "badness" of the data that provides us with our living, perhaps it is not all that surprising that we have shown little interest in improving it, in getting involved in the grubby task of designing and collecting original data sets of our own. Most of our work is on "found" data, data that have been collected by somebody else, often for quite different purposes.

Economic data collection started primarily as a byproduct of other governmental activities: tax and customs collections. Early on, interest was expressed in prices and levels of production of major commodities. Besides tax records, population counts, and price surveys, the earliest large scale data collection efforts were various Censuses, family expenditure surveys, and farm cost and production surveys. By the middle 1940s the overall economic data pattern was set: governments were collecting various quantity and price series on a continuous basis, with the primary purpose of producing aggregate level indicators such as price indexes and national income accounts series, supplemented by periodic surveys of population numbers and production and expenditure patterns to be used primarily in updating the various aggregate series. Little microdata was published or accessible, except in some specific sub-areas, such as agricultural economics.

A pattern was also set in the way the data were collected and by whom they were analyzed.[1] With a few notable exceptions, such as France and Norway, and until quite recently, econometricians were not to be found inside the various statistical agencies, and especially not in the sections that were responsible for data collection. Thus, there grew up a separation

---

[1]. See Kuznets (1971) and Morgenstern (1950) for earlier expressions of similar opinions. Morgenstern's Cassandra like voice is still very much worth listening to on this range of topics.

of roles and responsibility. "They" collect the data and "they" are responsible for all of their imperfections. "We" try to do the best with what we get, to find the grain of relevant information in all the chaff. Because of this, we lead a somewhat remote existence from the underlying facts we are trying to explain. We did not observe them directly; we did not design the measurement instruments; and, often we know little about what is really going on (e.g., when we estimate a production function for the cement industry from Census data without ever having been inside a cement plant). In this we differ quite a bit from other sciences (including observational ones rather than experimental) such as archeology, astrophysics, biology, or even psychology where the "facts" tend to be recorded by the professionals themselves, or by others who have been trained by and are supervised by those who will be doing the final data analysis. Economic data tend to be collected (or often more correctly "reported") by firms and persons who are not professional observers and who do not have any stake in the correctness and precision of the observations they report. While economists have increased their use of surveys in recent years and even designed and commissioned a few special purpose ones of their own, in general, the data collection and thus the responsibility for the quality of the collected material is still largely delegated to census bureaus, survey research centers, and similar institutitions, and is divorced from the direct supervision and responsibility of the analyzing team.

It is only relatively recently, with the initiation of the negative income tax experiments and various longitudinal surveys intended to follow up the effects of different governmental programs, that econometric professionals had actually become involved in the primary data collection process. Once attempted, the job turned out to be much more difficult that was thought

originally, and taught us some humility.[2] Even with relatively large budgets, it was not easy to figure out how to ask the right question and to collect relevant answers. In part this is because the world is much more complicated than even some of our more elaborate models allow for, and partly because economists tend to formulate their theories in non-testable terms, using variables for which it is very hard to find empirical counterparts. For example, even with a large budget, it is difficult to think of the right series of questions, answers to which would yield an unequivocal number for the level of "human capital" or "permanent income" of an individual. Thinking about such "alibi-removing" questions should make us a bit more humble, restrain our continuing attacks on the various official data producing agencies, and push us towards formulating theories with more regard to what is observable and what kind of data may be available.

Even allowing for such reservations there has been much progress over the years as a result of the enormous increase in the quantity of data available to us, in our ability to manipulate them, and in our understanding of their limitations. Especially noteworthy have been the development of various longitudinal microdata sets (such as the Michigan PSID tapes, the Ohio State NLS surveys, the Wisconsin high school class follow-up study, and others),[3] the computerization of the more standard data bases and their more easy accessibility at the micro, individual response level (I have in mind here such developments as the Public Use Samples from the U.S. Population Census and the Current Population Surveys).[4] Unfortunately, much more progress has been made with labor force and income type data, where the samples are large, than in the

---

2. See Hausman and Wise (1985).

3. See Borus (1982) for a recent survey of longitudinal data sets.

4. This survey is, perforce, centered on U.S. data and experience, which is what I am most familar with. The overall developments, however, have followed similar patterns in most other countries.

availability of firm and other market transaction data. We are still in our infancy as far as our ability to interrogate and get reasonable answers about firm behavior is concerned. Most of the available microdata at the firm level are based on legally required responses to questions from various regulatory agencies who do not have our interests exactly in mind.

We do have, however, now a number of extensive longitudinal microdata sets which have opened a host of new possibilities for analysis and also raised a whole range of new issues and concerns. After a decade or more of studies that try to use such data, the results have been somewhat disappointing. We, as econometricians, have learned a great deal from these efforts and developed whole new subfields of expertise, such as sample selection bias and panel data analysis. We know much more about these kinds of data and their limitations but it is not clear that we know much more or more precisely about the roots and modes of economic behavior that underlie them.

The encounters between econometricians and data are frustrating and ultimately unsatisfactory both because econometricians want too much from the data and hence tend to be disappointed by the answers, and because the data are incomplete and imperfect. In part it is our fault, the appetite grows with eating. As we get larger samples, we keep adding variables and expanding our models, until on the margin, we are back with the same insignificance levels.

There are at least three interrelated and overlapping causes of our difficulties: (1) the theory (model) is incomplete or incorrect; (2) the units are wrong, either at too high a level of aggregation or with no way of allowing for the heterogeneity of responses; and, (3) the data are inaccurate on

their own terms, incorrect relative to what they purport to measure. The average applied study has to struggle with all three possibilities.

At the macro level and even in the usual industry level study, it is common to assume away the underlying heterogeneity of the individual actors and analyze the data within the framework of the "representative" firm or "average" individual, ignoring the aggregation difficulties associated with such concepts. In analyzing microdata, it is much more difficult to evade this issue and hence much attention is paid to various individual "effects" and "heterogeneity" issues. This is wherein the promise of longitudinal data lies -- their ability to control and allow for additive individual effects. On the other hand, as is the case in most other aspects of economics, there is no such thing as a free lunch: going down to the individual level exacerbates both some of the left out variables problems and the importance of errors in measurement. Variables such as age, land quality, or the occupational structure of an enterprise, are much less variable in the aggregate. Ignoring them at the micro level can be quite costly, however. Similarly, measurement errors which tend to cancel out when averaged over thousands or even millions of respondents, loom much larger when the individual is the unit of analysis.

It is possible, of course, to take an alternative view: that there are no data problems only model problems in econometrics (see Hendry, 1983). For any set of data there is the "right" model. Much of econometrics is devoted to procedures which try to assess whether a particular model is "right" in this sense and to criteria for deciding when a particular model fits and is "correct enough" (see Chapter 5 and the literature cited in Hendry). Theorists and model builders often proceed, however, on the assumption that ideal data will

be available and define variables which are unlikely to be observable, at least not in their pure form. Nor do they specify in adequate detail the connection between the actual numbers and their theoretical counterparts. Hence, when a contradition arises it is then possible to argue "so much worse for the facts." In practice one cannot expect theories to be specified to the last detail nor the data to be perfect or of the same quality in different contexts. Thus any serious data analysis has to consider at least two data generation components: the economic behavior model describing the stimulus-response behavior of the economic actors and the measurement model, describing how and when this behavior was recorded and summarized. While it is usual to focus our attention on the former, a complete analysis must consider them both.

In this chapter, I discuss a number of issues which arise in the encounter between the econometrician and economic data. Since they permeate much of econometrics, there will be quite a bit of overlap with some of the other chapters in the Handbook. The emphasis here, however, will be more on the problems that are posed by the various aspects of economic data than on the specific technological solutions to them.

After a brief review of the major classes of economic data and the problems that are associated with using and interpreting them, I shall focus on issues that are associated with using erroneous or partially missing data, discuss several empirical examples, and close with a few final remarks.

II. Economic Data: An Overview

Data: fr. Latin, plural of datum -- given.

Observation: fr. Latin observare -- to guard, watch.

It is possible to classify economic data along several different dimensions: (a) Substantive: Prices, Quantities, Commodity Statistics, Population Statistics, Banking Statistics, etc. (b) Objective versus Subjective: Prices versus expectations about them, actual wages versus self reported opinions about well being; (c) Type and periodicity: Time series versus cross-sections; monthly, quarterly, or annual; (d) Level of aggregation: Individuals, families, or firms (micro), and districts, states, industries, sectors, or whole countries (macro); (e) Level of fabrication: primary, secondary, or tertiary; (f) Quality: Extent, reliability and validity.

As noted earlier, the bulk of economic data is collected and produced by various governmental bodies, often as a by-product of their other activities. Roughly speaking, there are two major types of economic data: aggregate time series on prices and quantities at the commodity, industry, or country level, and periodic surveys with much more individual detail. In recent years, as various data bases became computerized, economic analysts have gained access to the underlying microdata, especially where the governmental reports are based on periodic survey results. This has led to a great flowering of econometric work on various microdata sets including longitudinal panels.

The level of aggregation dimension and the micro-macro dichotomy are are not exactly the same. In fact, much of the "micro" data is already

aggregated. The typical U.S. firm is often an amalgam of several enterprises and some of the larger ones may exceed in size some of the smaller countries or states. Similarly, consumer surveys often report family expenditure or income data which have been aggregated over a number of individual family members. Annual income and total consumption numbers are also the result of aggregation over more detailed time periods, such as months or weeks, and over a more detailed commodity and sources of income classification. The issues that arise from the mismatch between the level of aggregation at which the theoretical model is defined and expected to be valid and the level of aggregation of the available data have not really received the attention they deserve (see Chapters 20 and 30 for more discussion and some specific examples).

The level of fabrication dimension refers to the "closeness" of the data to the actual phenomenon being measured. Even though they may be subject to various biases and errors, one may still think of reports of hours worked during last week by a particular individual in a survey or the closing price of a specific common stock on the New York Stock Exchange on December 31 as primary observations. These are the basic units of information about the behavior of economic actors and the information available to them (though individuals are also affected by the macro information that they receive). They are the units in which most of our microtheories are denominated. Most of our data are not of this sort, however. They have usually already undergone several levels of processing or fabrication. For example, the official estimate of total corn production in the State of Iowa in a particular year is not the result of direct measurement but the outcome of a rather complicated process of blending sample information on physical yields, reports on grain

shipments to and from elevators, benchmark census data from previous years, and a variety of informal Bayes-like smoothing procedures to yield the final official "estimate" for the state as a whole. The final results, in this case, are probably quite satisfactory for the uses they are put to, but the procedure for creating them is rarely described in full detail and is unlikely to be replicable. This is even more true at the aggregated level of national income accounts and other similar data bases, where the link between the original primary observations and the final aggregate numbers is quite tenuous and often mysterious.

I do not want to imply that the aggregate numbers are in some sense worse than the primary ones. Often they are better. Errors may be reduced by aggregation and the informal and formal smoothing procedures may be based on correct prior information and result in a more reliable final result. What needs to be remembered is that the final published results can be affected by the properties of the data generating mechanism, by the procedures used to collect and process the data. For example, some of the time series properties of the major published economic series may be the consequence of the smoothing techniques used in their construction rather than a reflection of the underlying economic reality. (This was brought forceable home to me many years ago while collecting unpublished data on the diffusion of hybrid corn at the USDA when I came across a circular instructing the state agricultural statisticians: "When in doubt -- use a growth curve.") Some series may fluctuate because of fluctuations in the data generating institutions themselves. For example, the total number of patents granted by the U.S. Patent Office in a particular year depends rather strongly on the total number of patent examiners available to do the

job. For budgetary and other reasons, their number has gone through several cycles, inducing concomitant cycles in the actual number of patents granted. This last example brings up the point that while particular numbers may be indeed correct as far as they go, they do not really mean what we though they did.

Such considerations lead one to consider the rather amorphous notion of data "quality." Ultimately, quality cannot be defined independently of the intended use for the particular data set. In practice, however, data are used for multiple purposes and thus it makes some sense to indicate some general notions of data quality. Earlier I listed extent, reliability, and validity as the three major dimensions along which one may just the quality of different data sets. Extent is a synonym for richness: How many variables are present, what interesting questions had been asked, how many years and how many firms or individuals were covered? Reliability is actually a technical term in psychometrics, reflecting the notion of replicability and measuring the relative amount of random measurement error in the data by the correlation coefficient between replicated or related measurements of the same phenomenon. Note that a measurement may be highly reliable in the sense that it is a very good measure of whatever it measures, but still be the wrong measure for our particular purposes.

This brings us to the notion of validity which can be subdivided in turn into representativeness and relevance. I shall come back to the issue of how representative is a body of data when we discuss issues of missing and incomplete data. It will suffice to note here that it contains the technical notion of coverage: did all units in the relevant universe have the same (or alternatively, different but known and adjusted for) probability of being selected into the sample that underlies this particular data set?

Coverage and relevance are related concepts which shade over into issues that arise from the use of "proxy" variables in econometrics. The validity and relevance questions related less to the issue of whether a particular measure is a good (unbiased) estimate of the associated population parameter and more to whether it actually corresponds to the conceptual variable of interest. Thus one may have a good measure of current prices which are still a rather poor indicator of the currently expected future price and relatively extensive and well measured IQ test scores which may still be a poor measure of the kind of "ability" that is rewarded in the labor market.

III. Data and Their Discontents.

> My father would never eat "cutlets" (minced meat patties) in the old country. He would not eat them in restaurants because he didn't know what they were made of and he wouldn't eat them at home because he did.
>
> AN OLD FAMILY STORY

I will be able to touch only a few of the many serious practical and conceptual problems that arise when one tries to use the various economic data sets. Many of these issues have been discussed at length in the national income and growth measurement literature but are not usually brought up in standard econometrics courses or included in their curriculum. Among the many official and semi-official data base reviews one should mention especially the Creamer GNP Improvement report (U.S. Department of Commerce, 1979), the Rees committee report on productivity measurement (National Academy of Sciences, 1979), the Stigler committee (National Bureau of Economic Research, 1961)

and the Ruggles (Council on Wage and Price Stability, 1977) reports on price statistics, The Gordon (President's Committee to Appraise Employment Statistics, 1962), and the Levitan (National Committee on Employment and Unemployment Statistics, 1979) committee reports on the measurement of employment and unemployment, and the many continuous and illuminating discussions reported in the proceedings volumes of the Conference on Research in Income and Wealth, especially in volumes 19, 20, 22, 25, 34, 38, 45, 47, and 48 (National Bureau of Economic Research, 1957 ... 1983). All these references deal almost exclusively with U.S. data, where the debates and reviews have been more extensive and public, but are also relevant for similar data elsewhere.

At the national income accounts level there are serious definitional problems about the borders of economic activity (e.g., home production and the investment value of children) and the distinction between final and intermediate consumption activity (e.g., what fraction of education and health expenditures can be thought of as final rather than intermediate "goods" or "bads?"). There are also difficult measurement problems associated with the existence of the underground economy and poor coverage of some of the major service sectors. The major serious problem from the econometric point of view probably occurs in the measurement of "real"output, GNP or industry output in "constant prices," and the associated growth measures. Since most of the output measures are derived by dividing ("deflating") current value totals by some price index, the quality of these measures is intimately connected to the quality of the available price data. Because of this, it is impossible to treat errors of measurement at the aggregate level as being independent across price and "quantity" measures.

The available price data, even when they are a good indicator of what

they purport to measure, may still be inadequate for the task of deflation. For productivity comparisons and for production function estimation the observed prices are supposed to reflect the relevant marginal costs and revenues in a, at least temporary, competitive equilibrium. But this is unlikely to be the case in sectors where output or prices are controlled, regulated, subsidized, and sold under various multi-part tariffs. Because the price data are usually based on the pricing of a few selected items in particular markets, they may not correspond well to the average realized price for the industry as a whole during a particular time period, both because "easily priced" items may not be representative of the average price movements in the industry as a whole and because of the fact that many transactions are made with a lag, based on long term contracts. There are also problems associated with getting accurate transactions prices (Kruskal and Telser, 1960 and Stigler and Kindahl, 1970), but the major difficulty arises from getting comparable prices over time, from the continued change in the available set of commodities, the "quality change" problem.

"Quality change" is actually a special version of the more general comparability problem, the possibility that similarly named items are not really similar, either across time or individuals. In many cases the source of similarly sounding items is quite different: Employment data may be collected from plants (establishments), companies, or households. In each case the answer to the same question may have a different meaning. Unemployment data may be reported by a teenager directly or by his mother, whose views about it may both differ and be wrong. The wording of the question defining unemployment may have changed over time and so should also the interpretation of the reported statistic. The context in which a question is asked, its position within a series of questions on a survey, and the willingness to answer some of the

questions may all be changing over time making it difficult to maintain the assumption that the reported numbers in fact relate to the same underlying phenomenon over time or across individuals and cultures.

The common notion of quality change relates to the fact that many commodities are changing over time and that often it is impossible to construct appropriate pricing comparisons because the same varieties are not available at different times and in different places. Conceptually one might be able to get around this problem by assuming that the many different varieties of a commodity differ only along a smaller number of relevant dimensions (characterisitics, specifications), estimate the price-characteristics relationship econometrically and use the resulting estimates to impute a price to the missing model or variety in the relevant comparison period. This approach, pioneered by Waugh (1928) and Court (1936) and revived by Griliches (1961) has become known as the "hedonic" approach to price measurement. The data requirements for the application of this type of an approach are quite severe and there are very few official price indexes which incorporate it into their construction procedures. Actually, it has been used much more widely in labor economics and in the analyses of real estate values than in the construction of price deflator indexes. See Griliches (1971), Gordon (1983), Rosen (1974) and Triplett (1975) for expositions, discussions, and examples of this kind of an approach to price measurement.

While the emergence of this approach has sensitized both the producers and the consumers of price data to this problem and contributed to significant improvements in data collection and processing procedures over time, it is fair to note that much still remains to be done. In the U.S. GNP deflation procedures, the price of computers has been kept constant since the early 1960s, for lack of an agreement of what to do about it, resulting in a significant

underestimate in the growth of real GNP during the last two decades. Similarly, for lack of a more appropriate price index, aircraft purchases have been deflated by an equally weighted index of gasoline engine, metal door, and telephone prices.[5] One could go on adding to this gallery of horror stories but the main point to be made here is not that a particular price index is biased in one or another direction. Rather, the point is that one cannot take a particular published price index series and interpret it as measuring adequately the underlying notion of a price change for a well specified, unchanging, commodity or service being transacted under identical conditions and terms in different time periods. The particular time series may indeed be quite a good measure of it, or at least better than the available alternatives, but each case requires a serious examination whether the actual procedures used to generate the series do lead to a variable that is close enough to the concept envisioned by the model to be estimated or by the theory under test. If not, one needs to append to the model an equation connecting the available measured variable to the desired but not actually observed correct version of this variable.

The issues discussed above affect also the construction and use of various "capital" measures in production function studies and productivity growth analyses. Besides the usual aggregation issues connected with the "existence" of an unambigous capital concept (see Diewert 1980 and Fisher 1969 on this) the available measures suffer from potential quality change problems, since they are usually based on some cumulated function of past investment expenditures deflated by some combination of available price indexes. In addition, they are also based on rather arbitrary assumptions about the

[5] For a recent review and reconstruction of the price indexes for durable producer goods see Gordon's (1985) forthcoming monograph.

pattern of survival of machines over time and the time pattern of deterioration in the flow of their services. The available information on the reasonableness of such assumptions is very sparse, ancient, and flimsy. In some contexts it is possible to estimate the appropriate patterns from the data rather than impose them a priori. We shall explore an example of this kind of an approach below.

Similar issues arise also in the measurement of labor inputs and associated variables: hours of work, unemployment, and wage rates; both at the macro and micro levels. At the macro level the questions revolve about the appropriate weighting to be given to different types of labor: young-old, male-female, black-white, educated vs. uneducated, and so forth. The direct answer here as elsewhere is that they should be weighted by their appropriate marginal prices but whether the observed prices actually reflect correctly the underlying differences in their respective marginal productivities is one of the more hotly debated topics in labor economics. (See Griliches 1970 on the education distinction and Medoff and Abraham 1980 on the age distinction). Connected to this is also the difficulty of getting relevant labor prices. Most of the usual data sources report or are based on data on average annual, weekly, or hourly earnings which do not represent adequately either the marginal cost of a particular labor hour to the employer or the marginal return to a worker from the additional hour of work. Both are affected by the existence of overtime premia, fringe benefits, training costs, and transportation costs. Only recently has an employment cost index been developed in the United States. (See Triplett 1983 on this range of issues.) From an individual workers point of view the existence of non-proportional tax schedules introduces another source of discrepancy between the observed wage

rates and the unobserved marginal after tax net returns from working (see Hausman, 1982,for a more detailed discussion).

While the conceptual discrepancy between the desired concepts and the available measures dominates at the macro level the more mundane topics of errors of measurement and missing and incomplete data come to the fore at the micro, individual survey level. This topic is the subject of the next section.

IV. Random measurement errors and the classic EVM.

To disavow an error is to invent retroactively.

Goethe

While many of the macro series may be also subject to errors, the errors in them rarely fit into the framework of the classical errors-in-variables model (EVM) as it has been developed in econometrics (see Chapter 23 for a detailed exposition). They are more likely to be systematic and correlated over time.[6] Micro data are subject to at least three types of discrepancies, "errors," and fit this framework much better:

(a) Transcription, transmission, or recording error, where a correct response is recorded incorrectly either because of clerical error (number transposition, skipping a line or a column) or because the observer misunderstood or misheard the original response.

(b) Response or sampling error, where the correct underlying value could be ascertained by a more extensive sampling, but the actual observed value is not equal to the desired underlying population parameter. For example, an IQ test is based on a sample of responses to a selected number of questions. In principle, the mean of a large number of tests over a wide range of questions would converge to some mean level of "ability" associated with the range of subjects being tested. Similarly, the simple permanent income hypothesis would assert that reported income in any particular year is a random draw from a potential population of such incomes whose mean is "permanent income." This is the case where the observed variable is a direct but fallible indicator of the underlying relevant "unobservable," "latent" "factor" or variable (see Chapter 23 and Griliches, 1974, for more discussion of such concepts).

---

[6] For an "error analysis" of national income account data based on the discrepancies between preliminary and "final" estimates see Cole (1969), Young (1974), and Haitovsky (1972). For an earlier more detailed evaluation based on subjective estimates of the differential quality of the various "ingredients" (series) of such accounts see Kuznets (1954, chapter 12).

(c) When one is lacking a direct measure of the desired concept and a "proxy" variable is used instead. For example, consider a model which requires a measure of permanent income and a sample which has no income measures at all but does have data on the estimated market value of the family residence. This housing value may be related to the underlying permanent income concept, but not clearly so. First, it may not be in the same units, second it may be affected by other variables also, such as house prices and family size, and third there may be "random" discrepancies related to unmeasured locational factors and events that occurred at purchase time. While these kinds of "indicator" variables do not fit strictly into the classical EVM framework, their variances, for example, need not exceed the variance of the true "unobservable," they can be fitted into this framework and treated with the same methods.

There are two classes of cases which do not really fit this framework: Occasionally one encounters large transcription and recording errors. Also, sometimes the data may be contaminated by a small number of cases arising from a very different behavioral model and/or stochastic process. Sometimes, these can be caught and dealt with by relatively simple data editing procedures. If this kind of problem is suspected, it is best to turn to the use of some version of the "robust estimation" methods discussed in Chapter 11 . Here we will be dealing with the more common general errors-in-measurement problem, one that is likely to affect a large fraction of our observations.

The other case that does not fit our framework is where the true concept, the unobservable is distributed randomly relative to the measure we have. For example, it is clear that the "number of years of school completed" (S) is an erroneous measure of true "education" (E), but it is more likely that the

discrepancy between the two concepts is independent of S rather than E. I.e., the "error" of ignoring differences in the quality of schooling may be independent of the measured years of schooling but is clearly a component of the true measure of E. Similarly, if we use the forecast of some model, based on past data, to predict the expectations of economic actors, we clearly commit an error, but this error is independent of the forecast level (if this forecast is optimal and the actors have had access to the same information). This type of "error" does not induce a bias in the estimated coefficients and can be incorporated into the standard disturbance framework (See Berkson 1950).

The standard EVM assumes the existence of a true relationship

$$y = \alpha + \beta z + e, \tag{4.1}$$

the absence of direct observations on z, and the availability of a fallible measure of it

$$x = z + \varepsilon \tag{4.2}$$

where ε is a purely random i.i.d. measurement error, with $E\varepsilon = 0$, and no correlation with either z or y. This is quite a restrictive set of assumptions, especially the assumption of the errors not being correlated with anything else in the model including their own past values. But it turns out to be very useful in many contexts and not too far off for a variety of micro data sets. I will discuss the evidence for the existence of such errors further on, when we turn to consider briefly various proposed solutions to the estimation problem in such models, but the required assumptions are not more difficult than those made in the standard linear regression model which requires that the "disturbance" e, the model discrepancy, be uncorrelated with all the included explanatory variables.

It may be worthwhile, at this point, to summarize the main conclusions from the EVM for the standard OLS estimates in contexts where one has ignored

the presence of such errors. Estimating

$$y = a + bx + u \tag{4.3}$$

where the true model is the one given above yields $-\beta\lambda$ as the assymptotic bias of the OLS $\hat{b}$, where $\lambda = \sigma_\varepsilon^2/\sigma_x^2$ is a measure of the relative amount of measurement error in the observed x series. The basic conclusion is that the OLS slope estimate is biased towards zero, while the constant term is biased away from zero. Since, in this model one can treat y and x symmetrically, it can be shown (Schultz 1938, Frisch 1934, Klepper and Leamer 1983) that in the "other regression," the regression of x on y, the slope coefficient is also biased towards zero, implying a "bracketing" theorem

$$\text{plim } b_{yx} < \beta < 1/\text{plim } b_{xy} \tag{4.4}$$

These results generalize also to the multivariate case. In the case of two independent variables (x_1 and x_2), where only one (x_1) is subject to error, the coefficient of the other variable (the one not subject to errors of measurement) is also biased (unless the two variables are uncorrelated). That is, if the true model is

$$y = \alpha + \beta_1 z_1 + \beta_2 x_2 + e \tag{4.5}$$

$$x_1 = z_1 + \epsilon$$

then

$$\text{plim}(b_{yx_1 \cdot x_2} - \beta_1) = -\beta_1 \lambda/(1-\rho^2) \tag{4.6}$$

where ρ is the correlation between the two observed variables x_1 and x_2, and if we scale the variables so that $\sigma_{x_1}^2 = \sigma_{x_2}^2 = 1$, then

$$\text{plim}(b_{yx_2 \cdot x_1} - \beta_2) = \rho\, \beta_1 \lambda/(1 - \rho^2) \qquad (4.7)$$

$$= -\rho\, [\text{bias } \beta_1]$$

That is, the bias in the coefficient of the erroneous variable is "transmitted" to the other coefficients, with an opposite sign (provided, as is often the case, that $\rho > 0$), (see Griliches and Ringstad, 1971, Appendix C, and Fisher 1980 for the derivation of this and related formulae).

If more than one independent variable is subject to error, the formulae become more complicated, but the basic pattern persists. If both z_1 and z_2 are unobserved and $x_1 = z_1 + \varepsilon_1$, $x_2 = z_2 + \varepsilon_2$, where the ε's are independent (of each other) errors of measurement, and we have normalized the variables so that $\sigma_{x_1}^2 = \sigma_{x_2}^2 = 1$, then

$$\text{plim}(b_{y1 \cdot 2} - \beta_1) = -\beta_1 \lambda_1/(1-\rho^2) + \beta_2\, \lambda_2\, \rho/(1-\rho^2) \qquad (4.8)$$

$$= -\frac{\beta_1 \lambda_1}{1-\rho^2} \{1 - \frac{\beta_2 \lambda_2}{\beta_1 \lambda_1} \rho\}$$

with a similar symmetric formula for plim $b_{y2 \cdot 1}$. Thus, in the multivariate case, the bias is increased by the factor $1/(1-\rho^2)$, the reduction in the independent variance of the true signal due to its intercorrelation with the other variable(s), and attenuated by the fact that the particular variable compensates somewhat for the downward bias in the other coefficients caused by the errors in the other variables. Overall, there is still a bias towards zero. For example, in this case the sum of the estimated coefficients is always biased towards zero:

$$\text{plim}[(b_{y1.2} + b_{y2.1}) - (\beta_1+\beta_2)] = [\beta_1\lambda_1 + \beta_2\lambda_2]/(1+\rho) \qquad (4.9)$$

It is a declining function of ρ, for $\rho > 0$, which is reasonable it we remember that ρ is defined as the intercorrelation between the observed x's. The higher it is, the smaller must be the role of independent measurement errors in these variables.

The impact of errors in variables on the estimated coefficients can be magnified by some transformations. For example, consider a quadratic equation in the unobserved true z:

$$y = \alpha + \beta z + \gamma z^2 + e \qquad (4.10)$$

with the observed

$$x = z + \varepsilon,$$

substituted instead.
If both z and ε are normally distributed, it can be shown (Griliches and Ringstad, 1970) that

$$\text{plim } \hat{b} = \beta(1-\lambda) \qquad (4.11)$$

while

$$\text{plim } \hat{c} = \gamma(1-\lambda)^2$$

where $\hat{b}$ and $\hat{c}$ are the estimated OLS coefficients in the $y = a + bx + cx^2 + u$ equation. That is, higher order terms of the equation are even more affected by errors in measurement than lower order ones.

The impact of errors in the levels of the variables may be reduced by aggregation and aggravated by differencing. For example, in the simple model $y = \alpha + \beta z + e$, $x = z + \varepsilon$, the asymptotic bias in the OLS b_{yx}

is equal to $-\beta\lambda$, while the bias of the first differenced estimator $[y_t - y_{t-1} = b(x_t - x_{t-1}) + v_t]$ is equal to $-\beta\lambda/(1-\rho)$ where ρ now stands for the first order serial correlation of the x's, and can be much higher than in levels (for $\rho > 0$ and not too small). Similarly, computing "within" estimates in panel data, or differencing across brothers or twins in micro data, can result in the elimination of much of the relevant variance in the observed x's, and a great magnification of the noise to signal ratio in such variables. (See Griliches, 1979, for additional exposition and examples.)

In some cases, errors in different variables cannot be assumed to be independent of each other. To the extent that the form of the dependence is known, one can derive similar formulae for these more complicated cases. The simplest and commonest example occurs when a variable is divided by another erroneous variable. For example, "wage rates" are often computed as the ratio of payroll to total man hours. To the extent that hours are measured with a multiplicative error, so will be also the resulting wage rates (but with opposite sign). In such contexts, the biases of (say) the estimated wage coefficient in a log-linear labor demand function will be towards -1 rather than zero.

The story is similar, though the algebra gets a bit more complicated, if the z's are categorical or zero-one variables. In this case the errors arise from misclassification and the variance of the erroneously observed x need not be higher than the variance of the true z. Bias formulae for such cases are presented in Aigner (1973) and Freeman (1984).

How does one deal with errors of measurement? As is well known, the standard EVM is not identified without the introduction of additional information, either in the form of additional data (replication and/or instrumental variables) or additional assumptions.

Procedures for estimation with known λ's are outlined in Chapter 23. Occasionally we have access to "replicated" data, when the same question is asked on different occasions or from different observers, allowing us to estimate the variance of the "true" variable from the covariance between the different measures of the same concept. This type of an approach has been used in economics by Bowles (1972) and Borus and Nestel (1973) in adjusting estimates of parental background by comparing the reports of different family members about the same concept, and by Freeman (1984) on a union membership variable, based on a comparison of worker and employer reports. Combined with a modelling approach it has been pursued vigorously and successfully in sociology in the works of Bielby, Hauser, and Featherman (1977), Massagli and Hauser (1983), and Mare and Mason (1980). While there are difficulties with assuming a similar error variance on different occasions or for different observers, such assumptions can be relaxed within the framework of a larger model. This is indeed the most promising approach, one that brings in additional independent evidence about the actual magnitude of such errors.

Almost all other approaches can be thought of as finding a reasonable set of instrumental variables for the problem, variables that are likely to be correlated with the true underlying z, but not with either the measurement error ε or the equation error (disturbance) e. One of the earlier and simpler applications of this approach was made by Griliches and Mason (1972) in estimating an earnings function and worrying about errors in their ability measure (AFQT test scores). In a "true" equation of the form

$$y = \alpha + \beta s + \gamma a + \delta x + e \qquad (4.12)$$

where y = log wages, s = schooling, a = ability, and x = other variables, they substituted an observed test score t for the unobserved ability variable

and assumed that it was measured with random error: $t = a + \varepsilon$. They used then a set of background variables (parental status, regions of origin) as instrumental variables, the crucial assumption being that these background variables did not belong in this equation on their own accord. Chamberlain and Griliches (1975 and 1977) used "purged" information from the siblings of the respondents as instruments to identify their models (see also Chamberlain, 1971).

Varous "grouping" methods of estimation, which use city averages (Friedman, 1957), industry averages (Pakes 1983), or size class averages (Griliches and Ringstad 1971), to "cancel out" the errors, can be all interpreted as using the classification framework as a set of instrumental dummy variables which are assumed to be correlated with differences in the underlying true values and uncorrelated with the random measurement errors or the transitory fluctuations.[7]

The more complete MIMIC type models (Multiple indicators - multiple causes models, see Hauser and Goldberger, 1971) are basically full information versions of the instrumental variables approaches, with an attempt to gain efficiency by specifying the complete system in greater detail and estimating jointly. In the Griliches-Mason example, such a model would consist of the following set of equations:

$$\begin{aligned} a &= x\delta_1 + g \\ t &= a + \varepsilon \\ s &= x\delta_2 + \gamma_1 a + v \\ y &= \beta s + \gamma_2 a + e \end{aligned} \qquad (4.13)$$

[7] Grouping methods that do not use an "outside" grouping criterion but are based on grouping on x alone (or using its ranks as instruments) are not in general consistent and need not reduce the EV induced bias. (See Pakes (1982)).

where a is an unobserved "ability" factor, and the "unique" disturbances g, e, v, and ε are assumed all to be mutually uncorrelated. With enough distinct x's and $\delta_1 \neq \delta_2$, this model is estimable either by instrumental variable methods or maximum likelihood methods. The maximum likelihood versions are equivalent to estimating the associated reduced form system:

$$\begin{aligned} t &= x\delta_1 + g + \varepsilon \\ s &= x(\delta_2 + \gamma_1\delta_1) + \gamma_1 g + v \\ y &= x[\delta_2 + (\gamma_1\beta + \gamma_2)\delta_1] + (\gamma_1\beta + \gamma_2)g + \beta v + e \end{aligned} \tag{4.14}$$

imposing the non-linear parameter restrictions across the equations and retrieving additional information about them from the variance-covariance matrix of the residuals, given the no-correlation assumption about the ε's, g's, v's, and e's. It is possible, for example, to retrieve an estimate of $\beta + \gamma_2/\gamma_1$ from the variance-covariance matrix and pool it with the estimates derived from the reduced form slope coefficients. In larger, more over-identified models, there are more binding restrictions connecting the variance-covariance matrix of the residuals with the slope parameter estimates. Chamberlain and Griliches (1975) used an expanded version of this type of model with sibling data, assuming that the unobserved ability variable has a variance-components structure. Aasness (1983) uses a similar framework and consumer expenditures survey data to estimate Engel functions and the unobserved distribution of total consumption.

All of these methods rely on two key assumptions: (1) The original model $y = \alpha + \beta z + e$ is correct for all dimensions of the data. I.e., the β parameter is stable and (2) The unobserved errors are uncorrelated in some well specified known dimension. In cross-sectional data it is common to assume that

the z's (the "true" values) and the ε's (the measurement errors) are based on mutually independent draws from a particular population. It is not possible to maintain this assumption when one moves to time series data or to panel data (which are a cross-section of time series), at least as far as the z's are concerned. Identification must hinge then on known differences in the covariance generating functions of the z's and the ε's. The simplest case is when the ε's can be taken as white (i.e., uncorrelated over time) while the z's are not. Lagged x's then can be used as valid instruments to identify β. For example, the "contrast" estimator suggested by Karnai and Weisman (1974) which combines the differentially biased level (plim $b = \beta-\beta\lambda$) and first difference estimators (plim $b_\Delta = \beta - \beta\lambda/(1-\rho)$) to derive consistent estimators for β and λ, can be shown, for stationary x and y, to be equivalent (asymptotically) to the use of lagged x's as instruments.

While it may be difficult to maintain the hypothesis that errors of measurement are entirely white, there are many different interesting cases which still allow the identification of β. Such is the case if the errors can be thought of as a combination of a "permanent" error or misperception of or by individuals and a random independent over time error component. The first part can be encompassed in the usual "correlated" or "fixed" effects framework with the "within" measurement errors being white after all. Identification can be had then from constrasting the consequences of differencing over differing lengths of time. Different ways of differencing all sweep out the individual effects (real or errors) and leave us with the following kinds of bias formulae:

$$\text{plim } b_{1\Delta} \simeq \beta(1 - 2\sigma_v^2/s_{1\Delta}^2)$$
$$\text{plim } b_{2\Delta} \simeq \beta(1 - 2\sigma_v^2/s_{2\Delta}^2) \qquad (4.15)$$

where σ_v^2 is the variance of the independent over time component of the ε's, 1Δ denotes the transformation x_2-x_1 while 2Δ indicates differences taken two periods apart: x_3-x_1 and so forth, and the s^2's are the respective variances of such differences in x. (4.15) can be solved to yield:

$$\hat{\beta} = \frac{\omega_{2\Delta} - \omega_{1\Delta}}{s_{2\Delta}^2 - s_{1\Delta}^2} \quad \text{and} \quad \hat{\sigma}_v^2 = \frac{(\hat{\beta} - b_{2\Delta})s_{2\Delta}^2}{2\hat{\beta}} \qquad (4.16)$$

where $\omega_{j\Delta}$ is the covariance of j period differences in y and x. This in turn, can be shown to be equivalent to using past and future x's as instruments for the first differences.[8]

More generally, if one were willing to assume that the true z's are non-stationary, which is not unreasonable for many evolving economic series, but the measurement errors, the ε's, are stationary, then it is possible to use panel data to identify the parameters of interest even when the measurement errors are correlated over time.[9] Consider, for example, the simplest case of $T = 2$. The probability limit of the variance-covariance matrix between y and x if given by:

---

[8] See Griliches and Hausman, 1984, for details, generalizations, and an empirical example.

[9] I am indebted to A. Pakes for this point.

	x_1	x_2
y_1	βs_1^2	βs_{12}
y_2	βs_{21}	βs_2^2
x_1	$s_1^2 + \sigma^2$	$s_{12} + \rho\sigma^2$
x_2	$s_{21} + \rho\sigma^2$	$s_2^2 + \sigma^2$

(4.17)

where now s_{th} stands for the variances and covariances of the true z's, σ^2 is the variance of the ε's, and ρ is their first order correlation coefficient. It is obvious that if the z's are non-stationary then $(\text{cov } y_1x_1 - \text{cov } y_2x_2)/(\text{var } x_1 - \text{var } x_2)$ and $(\text{cov } y_1x_2 - \text{cov } y_2x_1)/(\text{cov } x_1x_2 - \text{cov } x_2x_1)$ yield consistent estimates of β. In longer panels this approach can be extended to accomodate additional error correlations and the superimposition of "correlated effects" by using its first differences analogue.

Even if the z's were stationary, it is always possible to handle the correlated errors case provided the correlation is known. This rarely is the case, but occasionally a problem can be put into this framework. For example, capital measures are often subject to measurement error but these errors cannot be taken as uncorrelated over time, since they are cumulated over time by the construction of such measures. But if one were willing to assume that the errors occur randomly in the measurement of investment and they are uncorrelated over time, and the weighting scheme (the depreciation rate) used in the construction of the capital stock measure is known, then the correlation between the errors in the stock levels is also known.

For example, if one is interested in estimating the rate of return to some capital concept, where the true equation is

$$\pi_t = a + rK_t^* + e_t \tag{4.18}$$

π is a measure of profits and K^* is defined as a geometrically weighted average of past true investments I_t^* :

$$K_t^* = I_t^* + \lambda K_{t-1}^* = I_t^* + \lambda I_{t-1}^* + \lambda^2 I_{t-2}^* + \ldots. \tag{4.19}$$

but we do not observe I_t^* or K_t^* only

$$I_t = I_t^* + \varepsilon_t \tag{4.20}$$

where ε_t is an i.i.d. error of measurement and the observed $K_t = \Sigma\lambda^i I_{t-1}$ is constructed from the erroneous I series, then if λ is taken as known, which is implicit in most studies that use such capital measures, instead of running versions of (4.18) involving K_t and dealing with correlated measurement errors we can estimate

$$\pi_t - \lambda\pi_{t-1} = a(1-\lambda) + rI_t + u_t - \lambda u_{t-1} - r\varepsilon_t \ , \tag{4.21}$$

which is now in standard EVM form, and use lagged values of I as instruments.

Hausman and Watson (1983) use a similar approach to estimate the seasonality in the unemployment series by taking advantage of the known correlation in the measurement errors introduced by the particular structure of the sample design in their data.

One needs to reiterate, that in these kinds of models (as is also true for the rest of econometrics) the consistency of the final estimates depends both on the correctness of the assumed economic model and the correctness of the assumptions about the error structure.[10] We tend to focus here on the latter, but the former is probably more important. For example, in Friedman's (1957) classical permanent income consumption function model, the estimated elasticity of consumption with respect to income is a direct estimate of one minus the error ratio (the ratio of the variance of transitory income to the variance of measured income). But this conclusion is conditional on having assumed that the true elasticity of consumption with respect to permanent income is unity. If that is wrong, the first conclusion does not follow. Similarly in the profit-capital stock example above, we can do something because we have assumed that the true depreciation is both known and geometric. All our conclusions about the amount of error in the investment series are conditional on the correctness of these assumptions.

---

[10]The usual assumption of normality of such measurement and response errors may not be tenable in many actual situations. See Ferber (1966) and Hamilton (1981) for empirical evidence on this point.

V. Missing Observations and Incomplete Data

This could but have happened once,
And we missed it, lost it forever.
Browning

Relative to our desires data can be and usually are incomplete in many different ways. Statisticians tend to distinguish between three types of "missingness": undercoverage, unit non-response, and item non-response (NAS, 1983). Undercoverage relates to sample design and the possibility that a certain fraction of the relevant population was excluded from the sample by design or accident. Unit non-response relates to the refusal of a unit or individual to respond to a questionnaire or interview or the inability of the interviewers to find it. Item non-response is the term associated with the more standard notion of missing data: questions unanswered, items not filled in, in a context of a larger survey or data collection effort. This term is usually applied to the situation where the responses are missing for only some fraction of the sample. If an item is missing entirely, then we are in the more familiar omitted variables case to which I shall return in the next section.

In this section I will concentrate on the case of partially missing data for some of the variables of interest. This problem has a long history in statistics and somewhat more limited history in econometrics. In statistics, most of the discussion has dealt with the randomly missing, or in newer terminology, ignorable case (See Rubin 1976, and Little 1982) where, roughly speaking, the desired parameters can be estimated consistently from the complete data subsets and "missing data" methods focus on using the rest of the available data to improve the efficiency of such estimates.

The major problem in econometrics is not just missing data but the possiblity (or more accurately, probability) that they are missing for a variety of self-selection reasons. Such "behavioral missing" implies not

only a loss of efficiency but also the possibility of serious bias in the estimated coefficients of models that do not take this into account. The recent revival of interest in econometrics in limited dependent variables models, sample-selection, and sample self-selection problems has provided both the theory and computational techniques for attacking this problem. Since this range of topics is taken up in Chapter 28, I will only allude to some of these issues as we go along. It is worth noting, however, that this area has been pioneered by econometricians (especially Amemiya and Heckman) with statisticians only recently beginning to follow in their footsteps (e.g., Little 1983).

The main emphasis here will be on the no-self-selection ignorable case. It is of some interest, because these kinds of methods are widely used, and because it deals with the question of how one combines scraps of evidence and what one can learn from them. Consider a simple example where the true equation of interest is

$$y = \beta x + \gamma z + e \tag{5.1}$$

where e is a random term satisfying the usual OLS assumptions and the constant has been supressed for notational ease. β and γ could be vectors and x and z could be matrices, but I will think of them at first as scalars and vectors respectively. For some fraction λ $[n_2/(n_1+n_2)]$ of our sample we are missing observations (responses) on x. Let us rearrange the data and call the complete sample A and the incomplete sample B. Assume that it is possible to describe the data generation mechanism by the following model

$$\begin{aligned} d &= 1 \quad \text{if} \quad g(x, z, m; \theta) + \varepsilon \geq 0 \\ d &= 0 \quad \text{if} \quad g(x, z, m; \theta) + \varepsilon < 0 \end{aligned} \tag{5.2}$$

where $d = 1$ implies that the observation is in set A, it is complete; $d = 0$ implies that x is missing, m is another variable(s) determining the response or sampling mechanism, θ is a set of parameters, and ε is a random variable, distributed independently of x, z, and m. The incomplete data problem is ignorable if (a) ε (and m) are distributed independently of e and (b) there is no connection or restrictions between the parameters θ and β and γ. If these conditions hold than one can estimate β and γ from the complete data subset A and ignore B. Even if θ and β and γ are connected, if ε and e are independent, β and γ can be estimated consistently in A but now some information is lost by ignoring the data generation process. (See Rubin 1976 and Little 1982 for more rigorous versions of such statements.)

Note that this notion of ignorability of the data generation mechanism is more general than the simpler notion of randomly missing x's. It does not require that the missing x's be similar to the observed ones. Given the assumptions of the model (a constant β irrespective of the level of x), the x's can be missing "non-randomly," as long as the conditional expectation of y given x does not depend on which x's are missing. For example, there is nothing especially wrong if all "high" x's are missing, provided e and x are independent over the whole range of the data.

Even though with these assumptions β and γ can be estimated consistently in the A subsample there is still some more information about them in sample B. The following questions arise then: (1) How much additional information is there in sample B and about which parameters? (2) How should the missing values of x be estimated (if at all)? What other information can be used to improve these estimates?[11]

[11] This section borrows heavily from Griliches, Hall and Hausman (1978).

Options include using only z, using z and y, or using z and m, where m is an additional variable, related to x but not appearing itself in the y equation.

To discuss this, it is helpful to specify an "auxiliary" equation for x:

$$x = \delta z + \phi m + v \tag{5.3}$$

where $E(v)=0$ and $E(ve)=0$. Note that as far as this equation is concerned, the missing data problem is one of missing the dependent variable for sub-sample B. If the probability of being present in the sample were related to the size of v, we would be in the non-ignorable case as far as the estimation of δ and ϕ are concerned. Assume this is not the case and let us consider at first only the simplest case of $\phi = 0$, with no additional m variables present.

One way of rewriting the model is then

$$\begin{aligned} y_a &= \beta x_a + \gamma z_a + e_a \\ x_a &= \delta z_a + v_a \\ y_b &= (\beta+\gamma\delta) z_b + e_b + \beta v_b \end{aligned} \tag{5.4}$$

How one estimates β, γ, and δ depends on what one is willing to assume about the world that generated such data. There are two kinds of assumptions possible: The first is a "regression" approach, which assumes that the parameters which are constant across different subsamples are the slope coefficients β, γ, and δ but does not impose the restriction that σ_v^2 and σ_e^2 are the same across all the various subsamples. There can be heteroscedasticity across samples as long as it is independent

from the parameters of interest. The second approach, the maximum likelihood approach, would assume that conditional on z, y and x are distributed normally and the missing data are a random sample from such a distribution. This implies that $\sigma^2_{e_a} = \sigma^2_{e_b}$ and $\sigma^2_{v_a} = \sigma^2_{v_b}$.

The first approach starts by recognizing that under the general assumptions of the model Sample A yields consistent estimates of β, γ, and δ with variance covariance matrix Σ_a. Then a "first order" procedure, i.e., one that estimates missing x's by z alone and does not iterate, is equivalent to the following: Estimate $\hat{\beta}_a$, $\hat{\gamma}_a$, $\hat{\delta}_a$ from sample A, rewrite the y equation as

$$\begin{pmatrix} y_a - \hat{\beta}_a x_a \\ y_b - \hat{\beta}_a \hat{\delta}_a z_b \end{pmatrix} = \gamma z + \begin{pmatrix} e_a \\ e_b + \beta v \end{pmatrix} + \varepsilon \qquad (5.5)$$

where ε involves terms which are due to the discrepancy between the estimated $\hat{\beta}$ and $\hat{\delta}$ and their true population values. Then just estimate γ from this "completed" sample by OLS.

It is clear that this procedure results in no gain in the efficiency of β, since $\hat{\beta}_a$ is based solely on sample A. It is also clear that the resulting estimate of γ could be improved somewhat using GLS instead of OLS.[12]

How much of a gain is there in estimating γ this way? Let the size of sample A be N_1 and of B be N_2. The maximum (unattainable)

[12] See Gourieroux and Monfort (1981).

gain in efficiency would be proportional to $(N_1 + N_2)/N_1$ (when $\sigma_v^2 = 0$). Ignoring the contribution of ε's, which is unimportant in large samples, the asymptotic variance of γ from the sample as a whole would be

$$\text{Var}(\hat{\gamma}_{a+b}) \simeq [N_1\sigma^2 + N_2(\sigma^2 + \beta_1^2\sigma_v^2)]/(N_1 + N_2)^2\sigma_z^2$$

and (5.6)

$$\text{Eff}(\hat{\gamma}_{a+b}) = \frac{\text{Var } \hat{\gamma}_{a+b}}{\text{Var}(\hat{\gamma}_a)} \simeq (1 - \lambda)(1 + \lambda \frac{\beta^2\sigma_v^2}{\sigma^2})$$

where $\sigma^2 = \sigma_e^2$; and $\lambda = N_2/(N_1 + N_2)$. Hence efficiency will be improved as long as $\beta^2\sigma_v^2 / \sigma^2 < 1/(1-\lambda)$, i.e., the unpredictable part of x (unpredictable from z) is not too important relative to σ^2, the overall noise level in the y equation.[13]

Let us look now at a few illustrative calculations. In the work to be discussed below, y will be the logarithm of the wage rate x is IQ, and z is schooling. IQ scores are missing for about one-third of the sample, hence $\lambda = 1/3$. But the "importance" of IQ in explaining wage rates is relatively small. Its independent contribution $(\beta^2\sigma_v^2)$ is small relative to the large unexplained variance in y. Typical numbers are $\beta = .005$, $\sigma_v = 12$, and $\sigma = .4$, implying

$$\text{Eff}(\hat{\gamma}_{a+b}) = 2/3\ [1 + \frac{1}{3}\frac{.0036}{.16}] = .672,$$

[13] Thus, remark 2 of Gourieroux and Monfort (1981) p. 583 is in error. The first order method is not always more efficient.

which is about equal to the 2/3's one would have gotten ignoring the terms in the brackets. Is this a big gain in efficiency? First, the efficiency (squared) metric may be wrong. A more relevant question is by how much can the standard error of γ be reduced by incorporating sample B into the analysis. By about 18 percent ($\sqrt{.672} = .82$) for these numbers. Is this much? That depends how large the standard error of γ was to start out with. In Griliches, Hall and Hausman (1978) a sample consisting of about 1,500 individuals with complete information yielded an estimate of $\gamma_a = .0641$ with a standard error of .0052. Processing another 700 plus observations could reduce this standard error to .0043, an impressive but rather pointless exercise, since nothing of substance depends on knowing γ within .001.

If IQ (or some other missing variable) were more important, the gain would be even smaller. For example, if the independent contribution of x to y were on the order of σ^2, then with one-third missing, $\text{Eff}(\gamma_{a+b}) \simeq 8/9$, and the standard deviation of γ would be reduced by only 5.7 percent. There would be no gain at all, if the missing variable was one and a half times as important as the disturbance [or more generally if $\beta^2\sigma_v^2/\sigma^2 > 1/(1-\lambda)$].

The efficiency of such estimates can be improved a bit more by allowing for the implied heteroscedesticity in these estimates and by iterating further across the samples. This is seen most clearly by noting that sample B yields an estimate of $\hat{\pi} = \beta + \gamma\delta$ with an estimated standard error σ_π. This information can be blended optimally with the sample A estimates of β, γ, δ, and Σ_a using non-linear techniques and maximum likelihood is one way of doing this.

If additional variables which could be used to predict x but which do not appear on their own accord in the y equation were available, then there is also a possibility to improve the efficiency of the estimated β and not just of γ. Again, unless these variables are very good predictors of x and unless the amount of complete data available is relatively small, the gains in efficiency from such methods are unlikely to be impressive. (See Griliches, Hall and Hausman 1978, and Haitovsky, 1968, for some illustrative calculations.)

The maximum likelihood approaches differ from the "first-order" ones by using also the dependent variable y, to "predict" the missing x's, and by imposing restrictions on equality of the relevant variances across the samples. The latter assumption is not usually made or required by the first order methods, but follows from the underlying likelihood assumption that conditional on z, x and y are jointly normally (or some other known distribution) distributed, and that the missing values are missing at random. In the simple case where only one variable is missing (or several variables are missing at exactly the same places), the joint likelihood connecting y and x to z, which is based on the two equations

$$\begin{aligned} y &= \beta x + \gamma z + e \\ x &= \delta z + v \end{aligned} \tag{5.7}$$

with $Ee = \sigma^2$, $Ev^2 = \eta^2$, $Eev = 0$ can be rewritten in terms of the marginal distribution function of y given z, and the conditional distribution function of x given y and z, with the corresponding equations:

$$\begin{aligned} y &= cz + u \\ x &= dy + fz + w \end{aligned} \tag{5.8}$$

and $Eu^2 = g^2$, $Ew^2 = h^2$, $Ewu = 0$. Given the normality assumption, this is just another way of rewriting the same model, with the new parameters related to the old ones by

$$c = \gamma + \beta\delta \qquad g^2 = \beta\eta^2 + \sigma^2$$
$$d = \beta\eta^2/(\beta^2\eta^2+\sigma^2) \qquad f = \delta - cd \qquad h^2 = \eta^2\sigma^2/g^2 \tag{5.9}$$

In this simple case the likelihood factors and one can estimate c and g^2 from the complete sample; d, f, and h^2 from the incomplete sample and solve back uniquely for the original parameters β, γ, δ, σ^2, and η^2. In this way all of the information available in the data is used and computation is simple, since the two regressions (y on z in the whole sample and x on y and z in the complete data portion) can be computed separately. Note, that while x is implicitly "estimated" for the missing portion, no actual "predicted" values of x are either computed or used in this framework.[14]

Table 1 illustrates the results of such computations when estimating a wage equation for a sample of young women from the National Longitudinal Survey, 30 percent of which were missing IQ data. The first row of the table gives estimates computed solely from the complete data subsample. The second one uses the schooling variable to estimate the missing IQ values in the incomplete portion of the data and then re-computes the OLS estimates. The third row uses GLS, reweighting the incomplete portion of the data to allow for the increased imprecision due to the estimation of the missing IQ values. The last row reports the maximum likelihood

[14] Marini et al (1980) describe such computations in the context of more than one set of variables missing in a nested pattern.

estimates. All the estimates are very close to each other. Pooling the samples and "estimating" the missing IQ values increases the efficiency of the estimated schooling coefficient by 29 percent. Going to maximum likelihood adds another percentage point. While these gains are impressive, substantively not much more is learned from expanding the sample except that no special sample selectivity problem is caused by ignoring the missing data subset. The χ^2_2 test for pooling yields the insignificant value of .8. That the samples are roughly similar, can be also seen from computing the biased schooling coefficient (ignoring IQ) in both matrices: It is equal to .057 (.10) in the complete data subset and .054 in the incomplete one.

The maximum likelihood computations get more complicated when the likelihood does not factor as neatly as it does in the simple "nested" missing case. This happens in at least two important common cases: (1) If the model is overidentified then there are binding constraints between the $L(y|z, \theta_1)$ and $L(x|y, z, \theta_2)$ pieces of the overall likelihood function. For example, if we have an extra exogeneous variable which can help predict x but does not appear on its own in the "structural" y equation, than there is a constraining relationship between the θ_1 and θ_2 parameters and maximum likelihood estimation will require iterating between the two. This is also the case for multi-equation systems where, say, x is itself structurally endogeneous because it is measured with error. (2) If the pattern of "missingness" is not nested, if observations on some

variables are missing in a number of different patterns which cannot be arranged in a set of nested blocks, then one cannot factor the likelihood function conveniently and one must approach the problem of estimating it directly.

There are two related computational approaches to this problem: The first is the EM algorithm (Dempster et al, 1977). This is a general approach to maximum likelihood estimation where the problem is divided into an iterative two-step procedure. In the E-step (estimation), the missing values are estimated on the basis of the current parameter values of the model (in this case starting with all the available variances and covariances) and an M-step (maximization) in which maximum likelihood estimates of the model parameters are computed using the "completed" data set from the previous step. The new parameters are then used to solve again for the missing values which are then used in turn to reestimate the model, and this process is continued until convergence is achieved. While this procedure is easy to program, its convergence can be slow, and there are no easily available standard error estimates for the final results (though Beale and Little, 1975, indicate how they might be derived).

An alternative approach, which may be more attractive to model oriented econometricians and sociologists, given the assumption of ignorability of the process by which the data are missing, is to focus directly on pooling the available information from different portions of the sample which under the assumptions of the model are independent of each other. That is, the data are summarized by their relevant variance-covariance matrices (and means, if they are constrained by the model) and the model is expressed in terms of constraints on the elements of such matrices. What is done next is to "fit" the model to the observed matrices. This

approach is based on the idea that for multivariate normal distributed random variables the observed moment matrix is a sufficient statistic. Many models can be written in the form $\Sigma(\theta)$, where Σ is the true population covariance matrix associated with the assumed multivariate normal distribution and θ is a vector of parameters of interest. Denote the observed covariance matrix as S. Maximizing the likelihood function of the data with respect to the model parameters comes down to maximizing

$$\ell n\ L(\Sigma|S,\theta) = k - \frac{n}{2}\{\ell n\ |\Sigma(\theta)| + \text{tr}\ \Sigma(\theta)^{-1}S\} \qquad (5.10)$$

with respect to θ. If θ is exactly identified, the estimates are unique and can be solved directly from the definition of Σ and the assumption that S is a consistent estimator of it. If θ is overidentified, then the maximum likelihood procedure "fits" the model $\Sigma(\theta)$ to the data S as best as possible. If the observed variables are multivariate normal this estimator is the Full Information Maximum Likelihood estimator for this model. Even if the data are not multivariate normal but follow some other distribution with $E(S|\theta) = \Sigma(\theta)$, This is a pseudo- or quasi-maximum likelihood estimator yielding a consistent $\hat{\theta}$.[15] The correctness of the computed standard errors will depend, however, on the validity of the normality assumption. Robust standard errors for this model can be computed using the approach of White.

There is no conceptual difficulty in generalizing this to a multiple sample situation where the resulting $\Sigma_j(\theta_j)$ may depend on somewhat different parameters. As long as these matrices can be taken as arising independently,

---

[15] See Van Praag, 1983.

their respective contributions to the likelihood function can be added up. and as long as the θ_j's have parameters in common, there is a return from estimating them jointly. This can be done either utilizing the multiple samples feature of LISREL-V (see Allison, 1981 and Joreskog and Sorbom, 1981) or by extending the MOMENTS program (Hall, 1979) to the connected-multiple matrices case. The estimation procedure combines these different matrices and their associated pieces of the likelihood function, and then iterates across them until a maximum if found. (See Bound, Griliches and Hall, 1984 for more exposition and examples.)

I will outline this type of approach in a somewhat more complex, multi-equation context: the estimation of earnings functions from sibling data while allowing for an unobserved ability measure and errors of measurement in the variable of interest -- schooling. (See Griliches 1974 and 1979 for an exposition of such models.) The simplest version of such a model can be written as follows:

$$\begin{aligned} t &= a + e_1 &&= (f+g) + e_1 \\ s &= \delta a + h + e_2 &&= \delta(f+g) + (w+v) + e_2 \\ y &= \beta a + \lambda(s-e_2) + e_3 &&= \pi(f+g) + \gamma(w+v) + e_3 \end{aligned} \qquad (5.11)$$

where t is a reported IQ-type test score, s is the recorded years of school completed, and y = ln wage rate, is the logarithm of the wage rate on the current or last job, $a = (f+g)$ is an unobserved "ability" factor

with f being its "family" component. $h = (w+v)$ is the individual opportunity factor (above and beyond a and hence assumed to be orthogonal to it), with w, "wealth," as its family component. The e's are all random, uncorrelated and untransmitted measurement errors. That is

$$Eee' = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$$

and $\pi = \beta + \gamma\delta$. In addition, it is convenient to define

$$\begin{aligned} &\text{Var } a = a^2, \quad \text{Var } h = h^2 \\ &\tau = \text{Var } f/a^2, \quad \rho = \text{Var } w/h^2 \end{aligned} \tag{5.12}$$

where τ and ρ are the ratios of the variance of the family components to total variance in the a and h factors respectively.

Given these assumptions, the expected values of the variance-covariance matrix of all the observed variables across both members of a sib-pair is given by

	t_1	s_1	y_1	t_2	s_2	y_2
t_1	$a^2+\sigma_1^2$	δa^2	πa^2	τa^2	$\tau\delta a^2$	$\tau\pi a^2$
s_1		$\delta^2\alpha^2+h^2+\sigma_2^2$	$\delta\pi a^2+\gamma h^2$		$\tau\delta^2 a^2+\rho h^2$	$\tau\delta\pi a^2+\rho\gamma h^2$
y_1			$\pi^2 a^2+\gamma^2 h^2+\sigma_3^2$			$\tau\pi^2 a^2+\rho\gamma^2 h^2$

(5.13)

where only the 12 distinct terms of the overall 6 x 6 matrix are shown, since the others are derivable by symmetry and by the assumption that all the relevant variances (conditional on a set of exogeneous variables) are the same across sibs. With 10 unknown parameters this model would be under-identified without sibling data. This type of model was estimated by Bound, Griliches and Hall (1984) using sibling data from the National Longitudinal Surveys of Young Men and Young Women.[16] They had to face, however, a very serious missing data problem since much of the data, especially test scores, were missing for one or both of the siblings. Data were complete for only 164 brother pairs and 151 sister pairs but additional information subject to various patterns of "missingness" was available for 315 more male and 306 female sibling pairs and 2852 and 3398 unrelated male and female respondents respectively. Their final estimates were based on pooling the information from 15 different matrices for each sex and were used to test the hypothesis that the unobserved factors are the same for both males and females in the sense that their loadings (coefficients) are similar in the male and female versions of the model and that the implied correlation between the male and female family components of these factors was close to unity. The latter test utilized the cross-sex cross-sib covariances arising from the brother-sister pairs (N = 774) in these panels.

Such pooling of data reduced the estimated standard errors of the major coefficients of interest by about 20 to 40 percent without changing the results significantly from those found solely in the "complete data" portions of their sample. Their major substantive conclusion was that taking out the mean

16. The cited paper uses a more detailed 4 equation model based on an additional "early" wage rate.

differences in wages between young males and females, one could not detect significant differences in the impact of the unobservables or in their patterns between the male and female portions of their samples. As far as the IQ-Schooling part of the model is concerned, families and the market appeared to be treating brothers and sisters identically.

A class of similar problems occurs in the time series context: missing data at some regular time intervals, the "construction" of quarterly data from annual data and data on related time series, and other "interpolation" type issues. Most of these can be tackled using adaptations of the methods described above, except for the fact that there is usually more information available on the missing values and it makes sense to adapt these methods to the structure of the specific problem. A major reference in this area is Chow and Lin (1971). More recent references are Harvey and Pierse (1982) and Palm and Nijman (1982).

VI. Missing Variables and Incomplete Models.

"Ask not what you can do to the data
but rather what the data can do for you."

Every econometric study is incomplete. The stated model usually lists only the "major" variables of interest and even then it is unlikely to have good measures for all of the variables on the already foreshortened list. There are several ways in which econometricians have tried to cope with these facts of life: (1) Assume that the left-out components are random, minor, and independent of all the included exogeneous variables. This throws the problem into the "disturbance" and leaves it there, except for possible considerations of heteroscedasticity, variance-components, and similar adjustments, which impinge only on the efficiency of the usual estimates and not on their consistency. In many contexts it is difficult, however, to maintain the fiction that the left-out-variables are unrelated to the included ones. One is pushed than into either, (2), a specification sensitivity analysis where the direction and magnitude of possible biases are explored using prior information, scraps of evidence, and the standard left-out-variable bias formulae (Griliches 1957 and Chapter 5) or (3) one tries to transform the data so as to minimize the impact of such biases.

In this section, I will concentrate on this third way of coping which has used the increasingly available panel data set to try to get around some of these problems. Consider, then, the standard panel data set-up:

$$y_{it} = \alpha + \beta(i,t)\, x_{it} + \gamma(i,t)\, z_{it} + e_{it} \qquad (6.1)$$

where y_{it} and x_{it} are the observed dependent and "independent" variables respectively, β is the set of parameters of interest, z_{it} represents various possible misspecifications of the model in the form of left out variables, and e_{it} are the usual random shocks assumed to be well behaved and independently distributed (at this level of generality almost all possible deviations from this can be accommodated by redefining the z's). Two basic assumptions are made very early on in this type of model. The first one, that the relationship is linear, is already implicit in the way I have written (6.1). The second one is that the major parameters of interest, the β's, are both stable over time and constant across individuals. I.e.,

$$\beta(i,t) = \beta. \qquad (6.2)$$

Both of these assumptions are in principle testable, but are rarely questioned in practice. Unless there is some kind of stability in β, unless there is some interest in its central moments, it is not clear why one would engage in estimation at all. Since the longitudinal dimension of such data is usually quite short (2 - 10 years), it makes little sense to allow β to change over time, unless one has a reasonably clear idea and a parsimonious parameterization of how they change over time. (The fact that the β's are just coefficients of a first order linear approximation to a more complicated functional relationship and hence <u>should</u> change as the level of x's changes

can be allowed for by expanding the list of x's to contain high order terms).

The assumption that $\beta_i = \beta$, that all individual response are alike (up to the additive terms, the z_i, which can differ across individuals), is one of the more bothersome ones. If longer time series were available, it would be possible to estimate separate β_i's for each individual or firm. But that is not the world we find ourselves in at the moment. Right now there are basically three outs from this corner: (1) Assume that all differences in the β_i's are random and uncorrelated with everything else. Then we are in the random coefficients world (Chapter 21) and except for the issues of heteroscedasticity the problem goes away; (2) Specify a model for the differences in β_i, making them depend on additional observed variables, either own individual ones or higher-order macro ones (cf. Mundlak 1980). This results in defining a number of additional "interaction" variables within the x set. Unless there is strong prior information on how they differ, this introduces a whole additional dimension to the "specification search" (in Leamer's term) and is not very promising; (3) Ignore it, which is what I shall proceed to do for the moment, focusing instead on the heterogeneity which is implicit in the potential existence of the z_i's, the ignored or unavailable variables in the model.

Even if (6.1) is simplified to

$$y_{it} = \alpha + \beta x_{it} + \gamma_t z_{it} + e_{it} \qquad (6.3)$$

β is not identified from the data in the absence of direct

observations on z . Somehow, assumptions have to be made about the source of the z's and their distributional properties, before it is possible to derive consistent estimators of β. There are (at least) three categories of assumptions that can be made about such z's which lead to different estimation approaches in this context:

(a) The z's are random and independent of x's. This is the easy but not too likely case. The z's can be collapsed then into the e_i's with only the heteroscedasticity issue remaining for the "random effects" model to solve.

(b) The z's are correlated with the x's but are constant over time and have also constant effects on the y's. I.e.,

$$\gamma(t)z_{it} = z_i \tag{6.4}$$

where we have normalized $\gamma = 1$. This is the standard "fixed" or "correlated" effects model (see Maddala 1971, and Mundlak 1978) which has been extensively analyzed in the recent literature. This is the case for which the panel structure of the data provides a perfect solution. Letting each individual have its own mean level and expressing all the data as deviations from own means eliminates the z's and leads to the use of "within" estimators:

$$y_{it} - \bar{y}_{i\cdot} = \beta(x_{it} - \bar{x}_{i\cdot}) + e_{it} - \bar{e}_{i\cdot} \tag{6.5}$$

where $\bar{y}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, etc., and yields consistent estimates of β .

I have only two cautionary comments on this topic: As is true in many other contexts, and as was noted earlier, solving one problem may aggravate another. If there are two reasons for the z_{it}, e.g., both "fixed"effects and errors in variables, then

$$z_{it} = \alpha_i - \beta\varepsilon_{it} \tag{6.6}$$

where α_i is the fixed individual effect and ε_{it} is the random uncorrelated over time error of measurement in x_{it}. In this type of model α_i causes an upward bias in the estimated β from pooled samples while ε_{it} results in a negative one. Going "within" not only eliminates α_i but also increases the second type of bias through the reduction of the signal to noise ratio. This is seen easiest in the simplest panel model where $T = 2$ and within is equivalent to first differencing. Undifferenced, an OLS estimate of β would yield

$$\text{plim}(\hat{\beta}_T - \beta) = b_{\alpha_i x} - \beta\lambda_T \tag{6.7}$$

where $b_{\alpha_i x}$ is the auxiliary regression coefficient in the projection of the α_i's on the x's, while $\lambda_T = \sigma^2_\varepsilon/\sigma^2_x$ is the error variance ratio in x. Going "within", on the other hand would eliminate the first term and leave us with

$$\text{plim}\ (\hat{\beta}_w - \beta) = -\beta\lambda_w = -\beta\ \lambda_T/(1-\rho) \tag{6.8}$$

where ρ is the first order serial correlation coefficient of the x's. A plausible example might have $\beta = 1$, $b_{\alpha_i x} = .2$, $\lambda_T = .1$, and $\hat{\beta}_T = 1 + .2 - .1 = 1.1$. Now, as might not be unreasonable, if $\rho = .67$, then $\lambda_w = .3$ and $\hat{\beta}_w = .7$, which is more biased than was the case with the original $\hat{\beta}_T$.

This is not an idle comment. Much of the recent work on production function estimation using panel data (e.g., see Griliches-Mairesse, 1984) starts out worrying about fixed effects and simultaneity bias, goes within, and winds up with rather unsatisfactory results (implausible low coefficients). Similarly, the rather dramatic reductions in the schooling coefficient in earnings equations achieved by analyzing "within" family data for MZ twins, is also quite likely the result of originally rather minor errors of measurement in the schooling variable (see Griliches 1979 for more detail).

The other comment has to do with the unavailability of the "within" solution if the equation is intrinsically non-linear since, for example, the mean of $e^x + \varepsilon$ is not equal to $e^{\bar{x}} + \bar{\varepsilon}$. This creates problems for models in which the dependent variables are outcomes of various non-linear probability processes. In special cases, it is possible to get around this problem by conditioning arguments. Chamberlain (1980) discusses the logit case while Hausman, Hall and Griliches (1984) show how conditioning on the sum of outcomes over the period as a whole converts a Poisson problem into a conditional multinomial logit problem and allows an equivalent "within" unit analysis.

(c) Non-constant effects. The general case here is one of a left out variable(s) and nothing much can be done about it unless more explicit assumptions are made about how the unseen variables behave and/or what their effects are. Solutions are available for special cases, cases that make restrictive enough assumptions on the $\gamma(t)z_{it}$ terms and their correlations with the included x variables (see Hausman and Taylor, 1981).

For example, it is not too difficult to work out the relevant algebra for

$$\gamma(t)z_{it} = \gamma_t \cdot z_i \tag{6.9}$$

or

$$\gamma(t)z_{it} = -\beta\varepsilon_{it} \tag{6.10}$$

where ε_{it} is an i.i.d measurement error in x. The first version, equation (6.9) is one of a "fixed" common effect with a changing influence over time. Such models have been considered by Stewart (1983) in the estimation of earnings function, by Pakes and Griliches (1984) for the estimation of geometric lag structures in panel data where the unseen truncation remainders decay exponentially over time, and by Anderson and Hsiao (1982) in the context of the estimation of dynamic equations with unobserved initial conditions. The second model, equation (6.10), is the pure EVM in the panel data context and was discussed in Section IV. It is estimable by using lagged x's as instruments, provided the "true" x's are correlated over time, or by grouping methods if independent (of the errors) information is available which allows

one to group the data into groups which differ in the underlying "true" x's (Pakes, 1983). Identification may become problematic when the EVM is superimposed on the standard fixed effects model. Estimation is still possible, in principle, by first differencing to get rid of the α_i's, the fixed effects, and then using past and future x's as instruments. (See Griliches and Hausman, 1984.)

Some of these issues can be illustrated by considering the problem of trying to estimate the form of a lag structure from a relatively short panel.[17] Let us define a flexible distributed lag equation

$$y_{it} = \alpha_i + \beta_0 x_{it} + \beta_1 x_{it-1} + \beta_2 x_{it-2} + \ldots + \varepsilon_{it} \qquad (6.11)$$

$$y_{it} = \alpha_i + \sum_{\tau=0}^{?} \beta_\tau x_{it-\tau} + \varepsilon_{it}$$

where the constancy of the β's is imposed across individuals and across time. The empirical problem is how does one estimate, say, 9 β's if one only has

[17] The following discussion borrows heavily from Pakes and Griliches (1984).

four to five years history on the y's and x's. In general this is impossible. If the length of the lag structure exceeds the available data, than the data cannot be informative about the unseen tail of the lag distribution without the imposition of stronger a priori restrictions. There are at least two ways of doing this: (a) We can assume something strong about the β's. For example, that they decline geometrically after a few free terms, that $\beta_{T+1} = \lambda\beta_T$. This leads us back to the geometric lag case which we know more or less how to handle.[18] (b) We can assume something about the unseen x's, that they were constant in the past (in which case we are back to the fixed effects with changing coefficient case), or that they follow some simple low order autoregressive process (in which case their influence on the included x's dies out after a few terms).

Before proceeding along these lines, it is useful to recall the notion of the Π-matrix , introduced in Chapter 22, which summarizes all the (linear) information contained in the standard time series - cross section panel model. This approach, due to Chamberlain (1982), starts with the set of unconstrained multivariate regressions, relating each year's y_{it} to all of the available x's, past, present, and future. Consider, for example, the case where data on y are available for only three years (T = 3) and on x's for four. Then the Π matrix consists of the coefficients in the following set of regressions:

$$
\begin{aligned}
y_{1i} &= \pi_{13}x_{3i} + \pi_{12}x_{2i} + \pi_{11}x_{1i} + \pi_{10}x_{0i} + v_{1i} \\
y_{2i} &= \pi_{23}x_{3i} + \pi_{22}x_{2i} + \pi_{21}x_{1i} + \pi_{20}x_{0i} + v_{2i} \qquad (6.12) \\
y_{3i} &= \pi_{33}x_{3i} + \pi_{32}x_{3i} + \pi_{31}x_{1i} + \pi_{30}x_{0i} + v_{3i}
\end{aligned}
$$

[18] See Anderson and Hsiao (1982) and Bhargava and Sargan (1983).

where we have ignored constants to simplify matters. Now all that we know from our sample about the relationship of the y's to the x's is summarized in these π's (or equivalently in the overall correlation matrix between all the y's and the x's), and any model that we shall want to fit will impose a set of constraints on it.[19]

A series of increasingly complex possible worlds can be written as:

$$\text{a.} \quad y_{it} = \beta_0 x_{it} + \beta_1 x_{it-1} + e_{it} \tag{6.13}$$

$$\text{b.} \quad y_{it} = \beta_0 x_{it} + \beta_1 x_{it-1} + \alpha_i + e_{it}$$

$$\text{c.} \quad y_{it} = \beta_0 x_{it} + \beta_1 (x_{it-1} + \lambda x_{it-2} + \lambda^2 x_{it-3} + \ldots) + e_{it}$$

$$\text{d.} \quad y_{it} = \beta_0 x_{it} + \beta_1 (x_{it-1} + \lambda x_{it-2} + \lambda^2 x_{it-3} + \ldots) + \alpha_i + e_{it}$$

$$\text{e.} \quad y_{it} = \beta_0 x_{it} + \beta_1 x_{it-1} + \beta_2 x_{it-2} + \beta_3 x_{it-3} + \beta_4 x_{it-4} \cdots + e_{it}$$

$$x_{it} = \rho x_{it-1} + \varepsilon_{it}$$

$$\text{f.} \quad y_{it} = \beta_0 x_{it} + \beta_1 x_{it-1} + \beta_2 x_{it-2} + \beta_3 x_{it-3} + \beta_4 x_{it-4} \cdots + \alpha_i + e_{it}$$

$$x_{it} = k\alpha_i + \rho x_{it-1} + \varepsilon_{it}$$

going from the simple one lag, no fixed effects case (a) to the arbitrary lag structure with the one factor correlated effects structure (f). For each of these cases we can derive the expected value of Π. It is obvious that (a) implies

---

[19] There may be, of course, additional useful information in the separate correlation matrices between all of the y's and all of the x's respectively.

$$\Pi(a) = \begin{pmatrix} 0 & 0 & \beta_0 & \beta_1 \\ 0 & \beta_0 & \beta_1 & 0 \\ \beta_0 & \beta_1 & 0 & 0 \end{pmatrix}$$

For the b case, fixed effects with no lags, we need to define the wide sense least squares projection (E*) of the unseen effects (α_i) on all the available x's

$$E^*(\alpha_i | x_{01} \ldots x_3) = \delta_3 x_{i3} + \delta_2 x_{i2} + \delta_1 x_{i1} + \delta_0 x_{i0} \tag{6.14}$$

Then

$$\Pi(b) = \begin{pmatrix} \delta_3 & \delta_2 & \delta_1+\beta_0 & \delta_0+\beta_1 \\ \delta_3 & \delta_2+\beta_0 & \delta_1+\beta_1 & \delta_0 \\ \delta_3+\beta_0 & \delta_2+\beta_1 & \delta_1 & \delta_0 \end{pmatrix}$$

To write down the Π matrix for c, the geometric lag case, we rewrite (6.11) as

$$\begin{aligned} y_{1i} &= \beta_0 x_{1i} + \beta_1 x_{0i} + z_i + e_{1i} \\ y_{2i} &= \beta_0 x_{2i} + \beta_1 x_{1i} + \beta_1 \lambda x_{0i} + \lambda z_i + e_{2i} \\ y_{3i} &= \beta_0 x_{3i} + \beta_1 x_{2i} + \beta_1 \lambda x_{1i} + \beta_2 \lambda^2 x_{0i} + \lambda^2 z_i + e_{3i} \end{aligned} \tag{6.15}$$

and (6.14) as

$$E^*(z_i|x) = m'x \tag{6.16}$$

which gives us the Π matrix corresponding to the geometric tail case

$$\Pi(c) = \begin{pmatrix} m_3 & m_2 & m_1+\beta_0 & m_0+\beta_1 \\ \lambda m_3 & \lambda m_2+\beta_0 & \lambda m_1+\beta_1 & \lambda(m_0+\beta_1) \\ \lambda^2 m_3+\beta_0 & \lambda^2 m_2+\beta_1 & \lambda^2 m_1+\lambda\beta_1 & \lambda^2(m_0+\beta_1) \end{pmatrix}$$

This imposes a set of non-linear constraints on the Π matrix, but is estimable with standard non-linear multivariate regression software (in SAS or TSP). In this case we have seven unknown parameters to estimate (4 m's, 2 β's, and λ) from the 12 unconstrained Π coefficients.[20]

Adding fixed effects on top of this, as in d, adds another four coefficients to be estimated and strains identification to its limit. This may be feasible with longer T but the data are unlikely to distinguish well between fixed effects and slowly changing initial effects, especially in short panels.

---

20. An alternative approach would take advantage of the geometric nature of the lag structure, and use lagged values of the dependent variable to solve out the unobserved z_i's. Using the lagged dependent variables formulation would introduce both an errors-in-variables problem (since y_{t-1} proxies for z subject to the e_{t-1} error) and a potential simultaneity problem due to their correlation with the α_i-s (even if the α's are not correlated with the x's). Instruments are available, however, in the form of past y's and future x's and such a system is estimable along the lines outlined by Bhargava and Sargan (1983).

Perhaps a more interesting version is represented by (e), where we are unwilling to assume an explicit form for the lag distribution since that happens to be exactly the question we wish to investigate, but are willing instead to assume something restrictive about the behavior of the x's in the unseen past; specifically that they follow an autoregressive process of low order. In the example sketched out, we never see x_{-1}, x_{-2} and x_{-3}, and hence cannot identify β_4 (or even β_3) but may be able to learn something about β_0, β_1, and β_2. If the x's follow a first order autoregressive process, than it can be shown (see Pakes and Griliches, 1984) that in the projection of $x_{-\tau}$ on all the observed x's

$$E^*(x_{-\tau}|x_3, x_2, x_1, x_0) = g'x = 0\cdot x_{i3} + 0\cdot x_{i2} + 0\cdot x_{i1} + g_\tau \cdot x_{i0} \qquad (6.17)$$

only the last coefficient is non-zero, since the partial correlation of $x_{-\tau}$ with all the subsequent x's is zero, given its correlation with x_o. If the x's had followed a higher order autoregression, say third order, then the last three coefficients would be non-zero. In the first order case the Π matrix is

$$\Pi(e) = \begin{pmatrix} 0 & 0 & \beta_0 & \beta_1 + \beta_2 g_1 + \beta_3 g_2 + \beta_4 g_3 \\ 0 & \beta_0 & \beta_1 & \beta_2 + \beta_3 g_1 + \beta_4 g_2 \\ \beta_0 & \beta_1 & \beta_2 & \beta_3 + \beta_4 g_1 \end{pmatrix}$$

where now only β_0, β_1 and β_2 are identified from the data. Estimation

proceeds by leaving the last column of Π to be free and constraining the rest of it to yield the parameters of interest.[21] If we had assumed that the x's are AR(2), we would be able to identify only the first two β's, and would have to leave the last two columns of Π free.

The last case to be considered, represents a mixture of fixed effects and truncated lag distributions. The algebra is somewhat tedious (see Pakes and Griliches, 1984) and leads basically to a mixture of the (c) and (e) case, where the fixed effects have changing coefficients over time, since their relationship to the correlated truncation remainder is changing over time:

$$\Pi(f) = \begin{bmatrix} \delta_3 & \delta_2 & \delta_1+\beta_0 & \Pi_{10} \\ m_2\delta_3 & m_2\delta_2+\beta_0 & m_2\delta_1+\beta_1 & \Pi_{20} \\ m_3\delta_3+\beta_0 & m_3\delta_2+\beta_1 & m_3\delta_1+\beta_2 & \Pi_{30} \end{bmatrix}$$

where I have normalized $m_1 = 1$. The first three β's should be identified in this model but in practice it may be rather hard to distinguish between all these parameters, unless T is significantly larger than 3, the underlying samples are large, and the x's are not too collinear.

[21] This is not fully efficient. If we really believe that the x's follow a low order Markov process with stable coefficients over time (which is not necessary for the above), then the equations for x can be appended to this model and the g's would be estimated jointly, constraining this column of Π also.

Following Chamberlain, the basic procedure in this type of model is first to estimate the unconstrained version of the Π matrix, derive its correct variance-covariance matrix allowing for the heteroscedasticity introduced by our having thrust those parts of the α_i and z_i which are uncorrelated with the x's into the random term (using the formulae in Chamberlain 1982, or White 1980), and then impose and test the constraints implied by the specific version deemed relevant.

Note that it is quite likely (in the context of longer T) that the test will reject all the constraints at conventional significance levels. This indicates that the underlying hypothesis of stability over time of the relevant coefficient may not really hold. Nevertheless, one may still use this framework to compare among several more constrained versions of the model to see whether the data indicate, for example, that "if you believe in a distributed lag model with fixed coefficients, then two terms are better than one."

Some of these ideas are illustrated in the following empirical example which considers the ubiquitous question of "capital." What is the appropriate way to define it and measure it? This is, of course, an old and much discussed question to which the theoretical answer is that in general it cannot be done in a satisfactory fashion (Fisher, 1969) and that in practice it depends very much on the purpose at hand (Griliches, 1963). There is no intention of reopening the whole debate here (see the various papers collected in Usher 1980 for a review of the recent state of this topic); the focus is rather on the much narrower question of what is the appropriate functional form for the depreciation or deterioration function used in the construction of conventional capital stock measures. Almost all of the data used empirically

are constructed on the basis of conventional "length of life" assumptions developed for accounting and tax purposes and based on very little direct evidence on the pattern of capital services over time. These accounting estimates are then taken to imply rather sharp declines in the service flows of capital over time using either the straight line or double declining balance depreciation formulae. Whatever independent evidence there is on this topic comes largely from used assets markets and is heavily contaminated by the effects of obsolescence due to technical improvements in newer assets.

Pakes and Griliches (1984) provide some direct empirical evidence on the question. In particular they asked: What is the time pattern of the contribution of past investments to current profitability? What is the shape of the "deterioration of services with age function" (rather than the "decline in present value" patterns)? All versions of capital stock measures can be thought of as weighted sums of past investments:

$$K_t = \Sigma\, w_\tau\, I_{t-\tau} \tag{6.18}$$

with w_τ differing according to the depreciation schemes used. Since investments are made to yield profits and assuming that ex ante the expected rate of return comes close to being equalized across different investments and firms, one would expect that

$$\Pi_t = \rho K_t + e_t = \rho\, (\Sigma\, w_\tau\, I_{t-\tau}) + e_t \tag{6.19}$$

where e_t is the ex post discrepancy between expected and actual profits assumed to be uncorrelated with the ex ante optimally chosen I's. Given a

series on Π_t and I_t, in principle one could estimate all the w parameters except for the problem that one rarely has a long enough series to estimate them individually, expecially in the presence of rather high multi-collinearity in the I's. Pakes and Griliches used panel data on U.S. firms to get around this problem, which greatly increases the available degrees of freedom. But even then, the available panel data are rather short in the time dimension (at least relatively to the expected length of life of manufacturing capital) and hence some of the methods described above have to be used.

They used data on the gross profits of 258 U.S. manufacturing firms for the nine years 1964-72 and their gross investment (deflated) for 11 years 1961-71. Profits were deflated by an overall index of the average gross rate of return (1972 = 100) taken from Feldstein and Summers (1977) and all the observations were weighted inversely to the sum of investment over the whole 1961-71 period to adjust roughly for the great heteroscedasticity in this sample. Model (6.13f) of the previous section was used. That is, they tried to estimate as many unconstrained w terms as possible asking whether these coefficients in fact decline as rapidly as is assumed by the standard depreciation formulae. To identify the model, it was assumed that in the unobserved past the I's followed an autoregressive process. Preliminary calculations indicated that it was adequate to assume a third order autoregression for I. Since they had also an accounting measure of capital stock as of the beginning of 1961, it could be used as an additional indicator of the unseen past I's. The possibility that more profitable firms may also invest more was allowed for by including individual

firm effects in the model and allowing them to be correlated with the I's and the initial K level. The resulting set of multivariate regressions with non-linear constraints on coefficients and a free covariance matrix was estimated using the LISREL-V program of Joreskog and Sorbom (1981).

Before their results are examined a major reservation should be noted about this model and the approach used. It assumes a fixed and common lag structure (deterioration function) across both different time periods and different firms which is far from being realistic. This does not differ, however, from the common use of accounting or constructed capital measures to compute and compare "rates of return" across projects, firms, or industries. The way "capital" measures are commonly used in industrial organization, production function, finance, and other studies implicitly assumes that there is a stable relationship between earnings (gross or net) and past investments; that firms or industries differ only by a factor of proportionality in the yield on these investments, with the time shape of these yields being the same across firms and implicit in the assumed depreciation formula. The intent of the Pakes-Griliches study was to question only the basic shape of this formula rather than try to unravel the whole tangle at once.

Their main results are presented in Table 2 and can be summarized quickly. There is no evidence that the contribution of past investments to current profits declines rapidly as is implied by the usual straight line or declining balance depreciation formula. If anything, they rise during the first three years! Introducing the 1961 stock as an additional indicator improves the estimates of the later w's and indicates no noticeable decline in the contibution of past investments during their first seven years. Compared against a single traditional capital stock measure (column 3), this

model does a significantly better job of explaining the variance of profits across firms and time. But it does not come close to doing as well as the estimates that correspond to the free Π matrix, implying that such lag structures may not be stable across time and/or firms. Nevertheless, it is clear that the usual depreciation schemes which assume that the contribution of past investments declines rapidly and immediately with age are quite wrong. If anything, there may be an "appreciation" in the early years as investments are completed, shaken down, and adjusted to.[22]

---

[22] For a methodologically related study see Hall, Griliches and Hausman (1983) which tried to figure out whether there is a significant "tail" to the patents as a function of past R&D expenditures lag structure.

VII. Final Remarks

The dogs bark but the caravan keeps moving.

A Russian proverb

Over 30 years ago Morgenstern (1950) asked whether economic data were accurate enough for the purposes that economists and econometricians were using them for. He raised serious doubts about the quality of many economic series and implicitly about the basis for the whole econometrics enterprise. Years have passed and there has been very little coherent response to his criticisms.

There are basically four responses to his criticism and each has some merit: (1) The data are not that bad. (2) The data are lousy but it does not matter. (3) The data are bad but we have learned how to live with them and adjust for their foibles. (4) That is all there is. It is the only game in town and we have to make the best of it.

There clearly has been great progress both in the quality and quantity of the available economic data. In the U.S. much of the agricultural statistical data collection has shifted from judgment surveys to probability based survey sampling. The commodity coverage in the various official price indexes has been greatly expanded and much more attention is being paid to quality change and other comparability issues. Decades of criticisms and scrutiny of official statistics have borne some fruit. Also, some of the aggregate statistics have now much more extensive micro-data underpinnings. It is now routine, in the U.S., to collect large periodic labor force activity and related

topics surveys and release the basic micro-data for detailed analysis with relatively short lags. But both the improvements in and the expansion of our data bases have not really disposed of the questions raised by Morgenstern. As new data appear, as new data collection methods are developed, the question of accuracy persists. While quality of some of the "central" data has improved, it is easy to replicate some of Morgenstern's horror stories even today. For example, in 1982 the U.S. trade deficit with Canada was either $12.8 or $7.9 billion depending on whether this number came from U.S. or Canadian publications. It is also clear that the national income statistics for some of the LDC's are more political than economic documents (Vernon, 1983).[23]

Morgenstern did not distinguish adequately between levels and rates of change. Many large discrepancies represent definitional differences and studies that are mostly interested in the movements in such series may be able to evade much of this problem. The tradition in econometrics of allowing for "constants" in most relationships and not over-interpreting them, allows implicitly for permanent "errors" in the levels of the various series. It is also the case that in much of economic analysis one is after relatively crude first order effects and these may be rather insensitive even to significant inaccuracies in the data. While this may be an adequate response with respect to much of the standard especially macro-economic analysis, it seems inadequate when we contemplate some of the more recent elaborate non-linear multi-equational models being estimated at the frontier of the subject. They are much more likely to be sensitive to errors and inconsistencies in the data.

[23] See also Prakash (1974) for a collection of confidence shattering comparisons of measures of industrial growth and trade for various developing countries based on different sources.

In the recent decade there has been a revival of interest in "error" models in econometrics, though the progress in sociology on this topic seems more impressive. Recent studies using micro-data from labor force surveys, negative-tax experiments and similar data sources exhibit much more sensitivity to measurement error and sample selectivity problems. Even in the macro area there has been some progress (see de Leeuw and McKelvey, 1983) and the "rational expectations" wave has made researchers more aware of the discrepancy between observed data and the underlying forces that are presumably affecting behavior. All of this has yet to make a major dent on econometric textbooks and econometric teaching but there are signs that change is coming. [24] It is more visible in the areas of discrete variable analysis and sample selectivity issues, (e.g., note the publication of the Maddala (1983) and Manski-McFadden (1981) monographs) than in the errors of measurement area per se, but the increased attention that is devoted to data provenance in these contexts is likely to spill over into a more general data "aware" attitude.

---

[24] Theil (1978) devotes five pages out of 425 to this range of problems. Chow (1983) devotes only six pages out of 400 to this topic directly, but does return to it implicitly in the discussion of rational expectations models. Dhrymes (1974) does not mention it explicitly at all, though some of it is implicit in his discussion of factor analysis. Dhrymes (1978) does devote about 25 pages out of 500 to this topic. Maddala (1977) and Malinvaud (1980) devote separate chapters to the EVM, though in both cases these chapters represent a detour from the rest of the book. The most extensive textbook treatment of the EVM and related topics appears in a chapter in Judge et al (1980). The only book that has some explicit discussion of economic data is Intriligator (1978). Except for the sample selection literature there is rarely any discussion of the processes that generate economic data and the resultant implications for econometric practice.

One of the reasons why Morgenstern's accusations were brushed off was that they came from "outside" and did not seem sensitive to the real difficulties of data collection and data generation. In most contexts the data are imperfect not by design but because that is all there is. Empirical economists have over generations adopted the attitude that having bad data is better than having no data at all, that their task is to learn as much as is possible about how the world works from the unquestionably lousy data at hand. While it is useful to alert users to their various imperfections and pitfalls, the available economic statistics are our main window on economic behavior. In spite of the scratches and the persistent fogging, we can not stop peering through it and trying to understand what is happening to us and to our environment, nor should we. The problematic quality of economic data presents a continuing challenge to econometricians. It should not cause us to despair, but we should not forget it either.

In this somewhat disjointed survey, I discussed first some of the long standing problems that arise in the encounter between the practicing econometrician and the data available to him. I then turned to the consideration of three data related topics in econometrics: errors of measurement, missing observations and incomplete data sets, and missing variables. The last topic overlapped somewhat with the chapter on panel analysis (Chapter 22), since longitudinal data provide us with one way of controlling for missing but relatively constant information on individuals and firms. It is difficult to shake off the impression that here also, the progress of econometric theory and computing ability is outracing the increased availability of data and our understanding and ability to model economic behavior in increasing detail. While we tend to look at the newly available data as adding degrees of freedom grist to our computer mills, the increased detail often raises

more questions than it answers. Particularly striking is the great variety of responses and differences in behavior across firms and individuals. Specifying additional distributions of unseen parameters rarely adds substance to the analysis. What is needed is a better understanding of the behavior of individuals, better theories and more and different variables. Unfortunately, standard economic theory deals with "representative" individuals and "big" questions and does not provide much help in explaining the production or hiring behavior of a particular plant at a particular time, at least not with the help of the available variables. Given that our theories, while couched in micro-language, are not truly micro-oriented, perhaps we should not be asking such questions. Then what are we doing with micro-data? We should be using the newly available data sets to help us find out what is actually going on in the economy and in the sectors that we are analyzing without trying to force our puny models on them.[25] The real challenge is to try to stay open, to learn from the data, but also, at the same time, not drown in the individual detail. We have to keep looking for the forest among all these trees.

---

[25] An important issue not discussed in this chapter is the testing of models which is a way of staying open and allowing the data to reject our stories about them. There is a wide range of possible tests that models can and should be subjected to. See, e.g., Chapters 5, 13, 14, 15, 18, 19, and 33 and Hausman (1978) and Hendry (1983).

References

Aasness, J., 1983. "Engle Functions, Distribution of Consumption and Errors in Variables," (paper presented at the European Meeting of the Econometric Society in Pisa) Oslo: Institute of Economics.

Aigner, D.J., 1973. "Regression with a Binary Independent Variable Subject to Errors of Observation," Journal of Econometrics (17), 49-59.

Allison, P.D., 1981. "Maximum Likelihood Estimation in Linear Models When Data Are Missing," Sociological Methodology, forthcoming.

Anderson, T.W. and C. Hsiao, 1982. "Formulation and Estimation of Dynamic Models Using Panel Data," Journal of Econometrics, 18(1), 47-82.

Beale, E.M.L. and R.J.A. Little, 1975. "Missing Values in Multivariate Analysis," Journal of the Royal Statistical Society, Ser. B., 37, 129-146.

Berkson, J., 1950. "Are There Two Regressions?", Journal of the American Statistical Association, 45, 164-180.

Bhargava, A. and D. Sargan, 1983. "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods," Econometrica, 51(6), 1635-1660.

Bielby, W.T., R.M. Hauser and D.L. Featherman, 1977. "Response Errors of Non-Black Males in Models of the Stratification Process," in Aigner and Goldberger (eds.) Latent Variables in Socioeconomic Models, Amsterdam: North Holland Publishing Company, 227-251.

Borus, M.E., 1982. "An Inventory of Longitudinal Data Sets of Interest to Economists," Review of Public Data Use, 10(1-2), 113-126.

Borus, M.E. and G. Nestel, 1973. "Response Bias in Reports of Father's Education and Socioeconomic Status," Journal of the American Statistical Association, 68(344), 816-820.

Bound, J., Z. Griliches and B.H. Hall, 1984. "Brothers and Sisters in the Family and Labor Market," unpublished paper presented at the Conference on Economics of the Family, April 12-13, Philadelphia.

Bowles, S., 1972. "Schooling and Inequality from Generation to Generation," Journal of Political Economy, Part II, 80(3), S219-S251.

Center for Human Resource Research, 1979. The National Longitudinal Survey Handbook, Ohio State University, Columbus, Ohio.

Chamberlain, Gary, 1977. "An Instrumental Variable Interpretation of Identification in Variance Components and MIMIC models," Chapter 7, in P. Taubman (ed.) Kinometrics, Amsterdam: North Holland Publishing Company, 235-254.

_____, 1980. "Analysis of Covariance with Qualitative Data," Review of Economic Studies, 47(1). 225-238.

_____, 1982. "Multivariate Regression Models for Panel Data," Journal of Econometrics 18(1), 5-46.

Chamberlain, G. and Z. Griliches, 1975. "Unobservables with a Variance-Components Structure: Ability, Schooling and the Economic Success of Brothers," International Economic Review 16(2), 422-449.

_____, 1977. "More on Brothers," in P. Taubman (ed.) Kinometrics: Determinants of Socioeconomic Success Within and Between Families, New York: North Holland Publishing Company, 97-124.

Chow, G.C., 1983. Econometrics, New York: McGraw Hill.

Chow, G.C. and A. Lin, 1971. "Best Linear Unviased Interpolation, Distribution and Extrapolation of Time Series by Related Series," Review of Economics and Statistics, 53(4), 372-375.

Cole, R., 1969. Error in Provisional Estimates of Gross National Product, Studies in Business Cycles #21, New York: NBER.

Council on Wage and Price Stability, 1977. The Wholesale Price Index: Review and Evaluation, Washington, D.C.: Executive Office of the President.

Court, A.T., 1939. "Hedonic Price Indexes with Automotive Examples," in The Dynamics of Automobile Demand, New York: General Motors Corporation, 99-117.

de Leeuw, F. and M.J. McKelvey, 1983. "A 'True' Time Series and Its Indicators," Journal of the American Statistical Association, 78(381), 37-46.

Dempster, A.P., N.M. Laird, and D.B. Rubin, 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Ser. B, 39(1), 1-38.

Dhrymes, P.J., 1974. Econometrics, New York: Springer-Verlag.

_____, 1978. Introductory Econometrics, New York: Springer-Verlag.

Diewert, W.E., 1980. "Aggregation Problems in the Measurement of Capital," in D. Usher (ed.), The Measurement of Capital, Studies in Income and Wealth, Vol. 45, University of Chicago Press for NBER, 433-538.

Eicker, F., 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Berkeley: University of California

Feldstein, M. and L. Summers, 1977. "Is the Rate of Profit Falling?", Brookings Papers on Economic Activity, 211-227.

Ferber, R., 1966. "The Reliability of Consumer Surveys of Financial Holdings: Demand Deposits," Journal of the American Statistical Association 61(313), 91-103.

Fisher, F.M., 1969. "The Existence of Aggregate Production Functions," Econometrica, 37(4), 553-577.

_____, 1980. "The Effect of Sample Specification Error on the Coefficients of 'Unaffected' Variables" in L.R. Klein, M. Nerlove and S.C.

Tsiang (eds.), Quantitative Economics and Development, New York: Academic Press, 157-163.

Freeman, R.B., 1984. "Longitudinal Analyses of the Effects of Trade Unions," Journal of Labor Economics 2(1), 1-26.

Friedman, M., 1957. A Theory of the Consumption Function, NBER General Series 63, Princeton: Princeton University Press.

Frisch, R., 1934. Statistical Confluence Analysis by Means of Complete Regression Systems, Oslo: University Economics Institute, Publication No. 5.

Gordon, R.J., 1982. "Energy Efficiency, User-Cost Change, and the Measurement of Durable Goods Prices," in NBER, Studies in Income and Wealth, Vol. 47, The U.S. National Income and Products Accounts, 205-268, M. Foss (ed.), Chicago: University of Chicago Press.

_____, 1985. The Measurement of Durable Goods Prices, NBER, forthcoming.

Gourieroux, C. and A. Monfort, 1981. "On the Problem of Mising Data in Linear Models," Review of Economic Studies, XLVIII(4), 579-586.

Griliches, Z. 1957. "Specification Bias in Estimates of Production Functions," Journal of Farm Economics 39(1), 8-20.

_____, 1961. "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change," in The Price Statistics of the Federal Government, NBER, 173-196.

_____, 1963. "Capital Stock in Investment Functions: Some Problems of Concept and Measurement," in Christ et al (eds.) Measurement in Economics, Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld, Stanford: Stanford University Press, 115-137.

_____, 1970. "Notes on the Role of Education in Production Functions and Growth Accounting," in Education, Income and Human Capital, W.L. Hansen (ed.), NBER Studies in Income and Wealth, Vol. 35, 71-127.

_____, 1971. Price Indexes and Quality Change, Cambridge, Mass.:

Harvard University Press.

Griliches, Z., 1974. "Errors in Variables and Other Unobservables," Econometrica, 42(6), 971-998.

_____, 1977. "Estimating the Returns to Schooling: Some Econometric Problems," Econometrica, 45(1), 1-22.

_____, 1979. "Sibling Models and Data in Economics: Beginnings of a Survey," Journal of Political Economy, 87(5), Part 2, S37-S64.

Griliches, Z., B.H. Hall and J.A. Hausman, 1978. "Missing Data and Self-Selection in Large Panels," Annales de L'INSEE, No. 30-31, 138-176.

Griliches, Z., and J.A. Hausman, 1984. "Errors-in-Variables in Panel Data: NBER Technical Paper No. 37, May.

Griliches, Z. and J. Mairesse, 1984. "Productivity and R&D at the Firm Level, in Z. Griliches (ed.), R&D, Patents and Productivity, NBER, Chicago: University of Chicago Press, 339-374.

Griliches, Z. and W.M. Mason, 1972. "Education, Income and Ability," Journal of Political Economy, Part II, 80(3), S74-S103.

Griliches, Z. and V. Ringstad, 1970. "Error in the Variables Bias in Non-Linear Contexts," Econometrica 38(2), 368-370.

_____, 1971. Economies of Scale and the Form of the Production Function, Amsterdam: North Holland Publishing Company.

Haitovsky, Y., 1968. "Estimation of Regression Equations When a Block of Observations is Missing," ASA, Proceedings of the Business and Economic Statistics Section, 454-461.

_____, 1972. "On Errors of Measurement in Regression Analysis in Economics, International Statistical Review 40(1), 23-35.

Hall, B.H., 1979. Moments: The Moment Matrix Processor User Manual, Stanford, California.

Hall, B.H., Z. Griliches, and J.A. Hausman, 1983. "Patents and R&D: Is There A Lag Structure?", NBER Working Paper No. 1227.

Hamilton, L.C., 1981. "Self Reports of Academic Performance: Response Errors Are Not Well Behaved," Sociological Methods and Research, 10(2), 165-185.

Harvey, A.C. and R.G. Pierse, 1982. "Estimating Missing Observations in Economic Time Series," London: London School of Economics Econometrics Programme Discussion Paper No. A33.

Hauser, R.M. and A.S. Goldberger, 1971. "The Treatment of Unobservable Variables in Path Analysis," in H.L. Costner (ed.) Sociological Methodology 1971, San Francisco, Jossey-Bass, 81-117.

Hausman, J.A., 1978. "Specification Tests in Econometrics," Econometrica, 46(6), 1251-1271.

_____, 1982. "The Econometrics of Non Linear Budget Constraints," Fisher-Schultz Lecture given at the Dublin Meetings of the Econometric Society, Econometrica, forthcoming.

Hausman, J.A., B.H. Hall and Z. Griliches, 1984. "Econometric Models for Count Data with Application to the Patents-R&D Relationship," Econometrica, 52(4), 909-938.

Hausman, J.A. and W. E. Taylor, 1981. "Panel Data and Unobservable Individual Effects," Econometrica 49(6), 1377-1398.

Hausman, J.A. and M. Watson, 1983. "Seasonal Adjustment with Measurement Error Present," National Bureau of Economic Research Working Paper No. 1133.

Hausman, J.A. and D. Wise, (eds.), 1985. Social Experimentation, NBER, Chicago: University of Chicago Press, forthcoming.

Hendry, D.F., 1983. "Econometric Modelling: The 'Consumption Function' in Retrospect," Nuffield College, Oxford, unpublished.

Intriligator, M.D., 1978. Econometric Models, Techniques and Applications. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Joreskog, K. and D. Sorbom, 1981. LISRELV, Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Method (National Educational Resources, Chicago, Illinois).

Judge, G.G., W.R. Griffiths, R.C. Hill and T.C. Lee, 1980. The Theory and Practice of Econometrics, New York: Wiley.

Karni, E. and I. Weissman, 1974. "A Consistent Estimator of the Slope in a Regression Model with Errors in the Variables," Journal of the American Statistical Association 69(345), 211-213.

Klepper, S. and E.E. Leamer, 1983. "Consistent Sets of Estimates for Regressions with Errors in All Variables," Econometrica 52(1), 163-184.

Kruskal, W.H. and L.G. Telser, 1960. "Food Prices and The Bureau of Labor Statistics," Journal of Business 33(3), 258-285.

Kuznets, S. 1954. National Income and Its Composition 1919-1938, New York: NBER.

_____, 1971. "Data for Quantitative Economic Analysis: Problems of Supply and Demand," lecture delivered at the Federation of Swedish Industries. Stockholm: Kungl Boktryckeriet P.A. Norsted and Soner.

Little, R.J.A., 1979. "Maximum Likelihood Inference for Multiple Regressions with Missing Values: A Simulation Study," Journal of the Royal Statistical Society, Ser. B, 41(1), 76-87.

_____, 1983. "Superpopulation Models for Non-Response," in National Academy of Sciences, Incomplete Data in Sample Surveys, Madow, Olkin, and Rubin (eds.), Vol. II, Part VI, 337-413, New York: Academic Press.

_____, 1982. "Models for Non-Response in Sample Surveys," Journal of the American Statistical Association 77(378), 237-250.

MaCurdy, T.E., 1982. "The Use of Time Series Processes to Model the Error Structure of Earnings in Longitudinal Data Analysis," Journal of Econometrics 18(1), 83-114.

Maddala, G.S., 1971. "The Use of Variance Components Models in Pooling Cross Section and Time Series Data," Econometrica, 39(2), 341-358.

_____, 1977. Econometrics, New York: McGraw Hill.

_____, 1983. Limited-Dependent and Qualitative Variables in Econometrics, Cambridge, England: Cambridge University Press.

Malinvaud, E., 1980. Statistical Methods of Econometrics, 3rd revised ed., Amsterday: North-Holland Publishing Company.

Manski, C.F. and D. MacFadden (eds.), 1981. Structural Analysis of Discrete Data with Econometric Applications, Cambridge, Massachusetts, MIT Press.

Mare, R.D. and W.M. Mason, 1980. "Children's Report of Parental Socioeconomic Status: A Multiple Group Measurement Model," Sociological Methods and Research 9, 178-198.

Marini, M.M., A.R. Olsen, and D.B. Rubin, 1980. "Maximum-Likelihood Estimation in Panel Studies with Missing Data," Sociological Methodology 1980, 9, 315-357.

Massagli, M.P. and R.M. Hauser, 1983. "Response Variability in Self- and Proxy Reports of Paternal and Filial Socioeconomic Characteristics," American Journal of Sociology, 89(2), 420-431.

Medoff, J. and K. Abraham, 1980. "Experience, Performance, and Earnings," Quarterly Journal of Economics, XVC(4), 703-736.

Morgenstern, O., 1950. On the Accuracy of Economic Observations, Princeton: Princeton University Press.

Mundlak, Y., 1978. "On the Pooling of Time Series and Cross Section Data," Econometrica, 46(1), 69-85.

_____, 1980. "Cross Country Comparisons of Agricultural Productivity," Unpublished manuscript.

National Academy of Sciences, 1979. Measurement and Interpretation of Productivity, Washington, D.C.

_____, 1983. Incomplete Data in Sample Surveys, Vol. 1-3, Madow, Olkin and Rubin (eds.), New York: Academic Press.

National Bureau of Economic Research, 1961. The Price Statistics of the Federal Government, Report of the Price Statistic Review Committee, General Series, No. 73, New York.

National Bureau of Economic Research, 1957a. Studies in Income and Wealth, Vol. 19, Problems of Capital Formation: Concepts, Measurement, and Controlling Factors, New York: Arno Press.

_____, 1957b. Studies in Income and Wealth, Vol. 20, Problems in International Comparisons of Economic Accounts, New York: Arno Press.

_____, 1958. Studies in Income and Wealth, Vol. 22, A Critique of the United States Income and Products Accounts, New York: Arno Press.

_____, 1961. Studies in Income and Wealth, Vol. 25, Output, Input and Productivity Measurement, New York, NBER.

_____, 1969. Studies in Income and Wealth, Vol. 34. Production and Productivity in the Service Industries, V.R. Fuchs, (ed.), New York: Columbia University Press.

_____, 1973. Studies in Income and Wealth, Vol. 38, The Measurement of Economic and Social Performance, M. Moss (ed.), New York: Columbia University Press.

_____, 1983a. Studies in Income and Wealth, Vol. 47, The U.S. National Income and Product Accounts, M. Foss (ed.), Chicago: University of Chicago Press.

_____, 1983b. Studies in Income and Wealth, Vol. 48. The Measurement of Labor Cost, J. Triplett (ed.), Chicago: University of Chicago Press.

National Commission on Employment and Unemployment Statistics, 1979. Counting the Labor Force, Washington, D.C.: Government Printing Office.

Pakes, A. 1982. "On the Asymptotic Bias of Wald-Type Estimators of a Straight Line When Both Variables Are Subject to Error," International Economic Review, 23(2), 491-497.

Pakes, A. 1983. "On Group Effects and Errors in Variables in Aggregation," Review of Economics and Statistics, LXV(1), 168-172.

Pakes, A. and Z. Griliches, 1984. "Estimating Distributed Lags in Short Panels with An Application to the Specification of Depreciation Patterns and Capital Stock Constructs," Review of Economic Studies, LI(2), 243-262.

Palm, F.C. and Th. E. Nijman, 1982. "Missing Observations in the Dynamic Regression Model," Paper presented at the Dublin Meeting of the Econometric Society. Amsterdam: Vrije Universiteit.

President's Committee to Appraise Employment and Unemployment Statistics, 1962. Measuring Employment and Unemployment, Washington, D.C.: Government Printing Office.

Rosen, S. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," Journal of Political Economy 82(1), 34-55.

Rubin, D.B., 1976. "Inference and Missing Data," Biometrika 63(3), 581-592.

Schultz, H., 1938. The Theory and Measurement of Demand, Chicago: University of Chicago Press.

Stewart, M.B., 1983. "The Estimation of Union Wage Differentials from Panel Data: The Problems of Not-So-Fixed Effects," National Bureau of Economic Research Conference on the Economics of Trade Unions, Cambridge, Mass., Unpublished.

Stigler, G.J., and J.K. Kindahl, 1970. The Behavior of Industrial Prices, National Bureau of Economic Research, New York: Columbia University Press.

Theil, H., 1978. Introduction to Econometrics, Englewood Cliffs, N.J.: Prentice Hall, Inc.

Triplett, J.E., 1975. "The Measurement of Inflation: A Survey of Research on the Accuracy of Price Indexes," in P.H. Earl (ed.) Analysis of Inflation, Chapter 2, Lexington, Mass.: Lexington Books, 19-82.

_____, 1983. "An Essay on Labor Cost," in National Bureau of Economic Research, Studies in Income and Wealth, Vol. 49, The Measurement of Labor Cost, Chicago: University of Chicago Press, 1-60.

U.S. Department of Commerce, 1979. Gross National Product Improvement Report, Washington, D.C.: Government Printing Office.

Usher, D. (ed.), 1980. The Measurement of Capital, National Bureau of Economic Research: Studies in Income and Wealth, Vol. 45, Chicago: University of Chicago Press.

Van Praag, B. 1983. "The Population-Sample Decomposition in Minimum Distance Estimation," unpublished paper presented at the Harvard-MIT Econometrics seminar.

Vernon, R., 1983. "The Politics of Comparative National Statistics," Cambridge, Massachusetts, Unpublished.

Waugh, F.V., 1928. "Quality Factors Influencing Vegetable Prices," Journal of Farm Economics, 10, 185-196.

White, H., 1980. "Using Least Squares to Approximate Unknown Regression Functions," International Economic Review, 21(1), 149-170.

Young, A.H., 1974. "Reliability of the Quarterly National Income and Product Accounts in the United States, 1947-71," Review of Income and Wealth, 20(1), 1-39.

Table 1: Earnings Equations for NLS Sisters: Various Missing Data Estimators

Estimation Method	Y Dependent S	Y Dependent T	T Dependent S	σ^2	η^2
OLS on complete data sample N = 366	.0434 (.0109)	.00433 (.00148)	3.211 (.398)	.1217	152.58
Total Sample: N = 520					
OLS with predicted IQ in missing portion*	.0423 (.00916)	.00433 (.00148)		.1186	
GLS with predicted IQ *	.0432 (.00915)	.00433 (.00148)			
Maximum Likelihood	.0427 (.00912)	.00421 (.00144)	3.205 (.346)	.1177	152.48

Y = log of wage rate, S = years of schooling completed, T = IQ type test score.
*The standard errors are computed using the Gourieroux-Monfort (1982) formulae.
All variable have been conditioned on age, region, race, and year dummy variables. The conditional moment matrices are:

	Complete Data (N=366)			Incomplete (154)		
LW	.13488			.12388		
IQ	1.2936	187.71		---	---	
SC	.19749	11.0703	3.4476	.23472	---	4.3408

Data Source: The National Longitudinal Survey of Young Women (See Center for Human Resource Research, 1979).

Table 2

The Relationship of Profits to Past Investment Expenditures for U.S. Manufacturing Firms: Parameter Estimates Allowing for Heterogeneity*

Parameter (Standard Error)	Without k^g_{-2}	With k^g_{-2}	Comparison Model (System 10)	3 Years Investment + $k^n_{i,t-4}$	3 Years Investment + $k^g_{i,t-4}$
	(1)	(2)	(3)	(4)	(5)
w_1	0.067 (0.028)	0.068 (0.027)		0.073 (0.022)	0.057 (0.021)
w_2	0.115 (0.033)	0.112 (0.032)		0.104 (0.022)	0.077 (0.022)
w_3	0.224 (0.041)	0.222 (0.040)		0.141 (0.024)	0.120 (0.024)
w_4	0.172 (0.046)	0.208 (0.046)			
w_5	0.072 (0.049)	0.198 (0.050)			
w_6	0.096 (0.062)	0.277 (0.057)			
w_7	-0.122 (0.094)	0.202 (0.076)			
w_8	-0.259 (0.133)	0.087 (0.103)			
Coefficient of:					
$k^n_{i,t}$			0.095 (0.012)		
$k^n_{i,t-4}$				0.103 (0.011)	
$k^g_{i,t-4}$					0.045 (0.006)
(Trace $\hat{\Omega}$)/253.6[1]		1.18	2.04	1.35	1.37

1. $\hat{\Omega}$ = Estimated covariance matrix of the disturbances from the system of profit equations (across years).

For the free Π matrix: trace $\hat{\Omega}$ = 253.6

Table 2 (continued)

*The dependent variable is gross operating income deflated by the implicit GNP deflator and an index of the overall rate of return in manufacturing (1972 = 1.0). The w_τ refer to the coefficients of gross investment expenditures in period $t - \tau$ deflated by the implicit GNP producer durable investment deflator. k^n_{it} and k^g_{it} are deflated Compustat measures of net and gross capital at the beginning of the year. k^g_{-2} refers to undeflated gross capital in 1961 as reported by Compustat. All variables are divided by the square root of the firm's mean investment expenditures over the 1961-71 period. Dummy variables for the nine time periods are included in all equations. $N = 258$ and $T = 9$.

The overall fit, measured by $1 - (\text{trace } \hat{\Omega})/1208.4$, $1208.4 = \sum_1^T s^2_{yt}$, where s^2_{yt} is the sample variance in y_t, is .72 for the model in Column 2 as against .79 for the free Π matrix.

From: Pakes and Griliches 1984.