Data and Econometrics

Jeff Parker, Econ 311, Fall 2017

Data are the raw material from which econometric analysis is constructed. Just as a building is no stronger than the wood or steel used in its framework, an econometric study is only as reliable as the data used in its analysis.

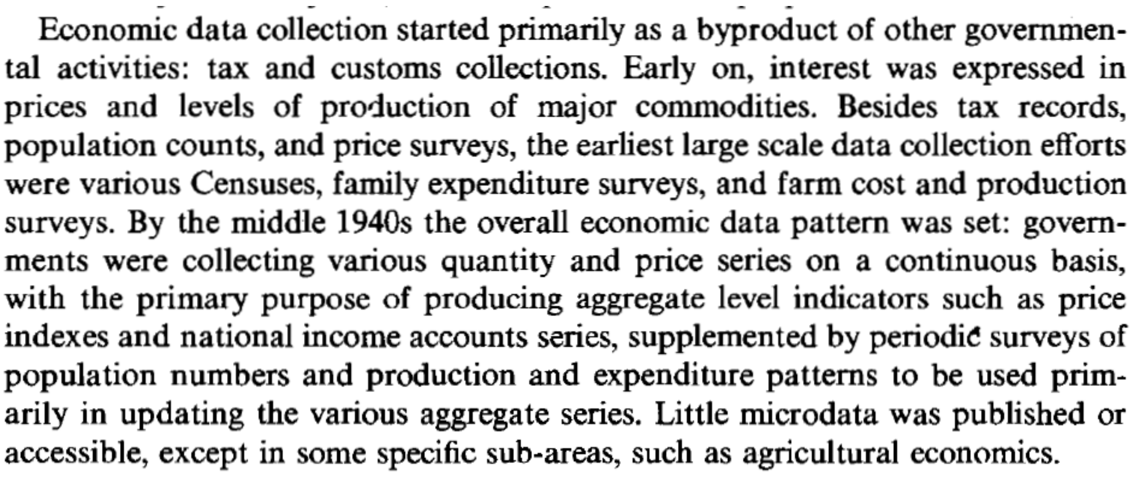
Many econometricians over the years have written about problems with data. One of the most comprehensive and comprehensible is a chapter that noted econometrician Zvi Griliches wrote for the third volume of Elsevier's *Handbook of Econometrics* back in 1986. Obviously much has changed in the world of data and econometrics in the last 30 years, but many of the points that Griliches made are still relevant today, and some are even more important.

This document uses extensive quotes from Griliches's chapter to highlight some important issues that every practitioner of econometrics should consider. You are encouraged to read the full chapter if you want more details; I've tried to give his summaries and introductions to the most important issues while leaving out the formal mathematical analysis and examples. Apologies for the imperfections in the scans and screen shots!

1. Who collects our data and why?

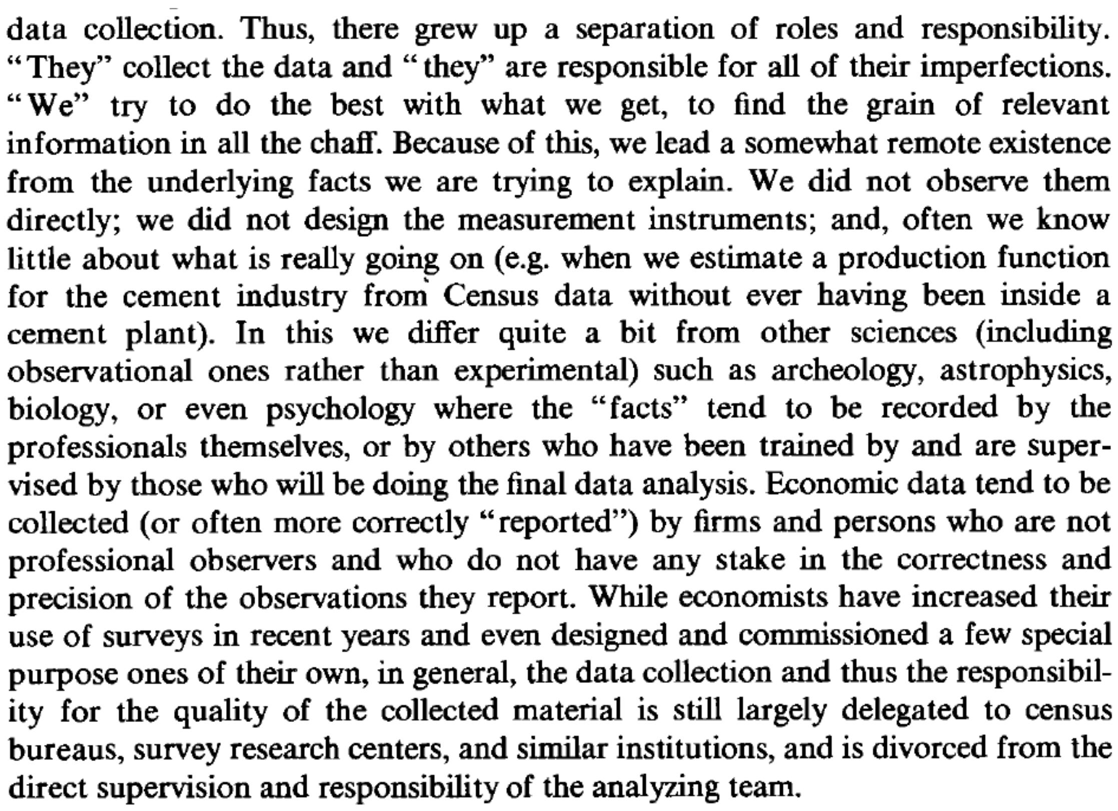
Economists sometimes collect their own data from experiments or surveys, but most econometric analysis relies on "found data," often from government sources. Were those data collectors trying to measure the same concept that economists want to measure?

From page 1466:



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.

From page 1467:

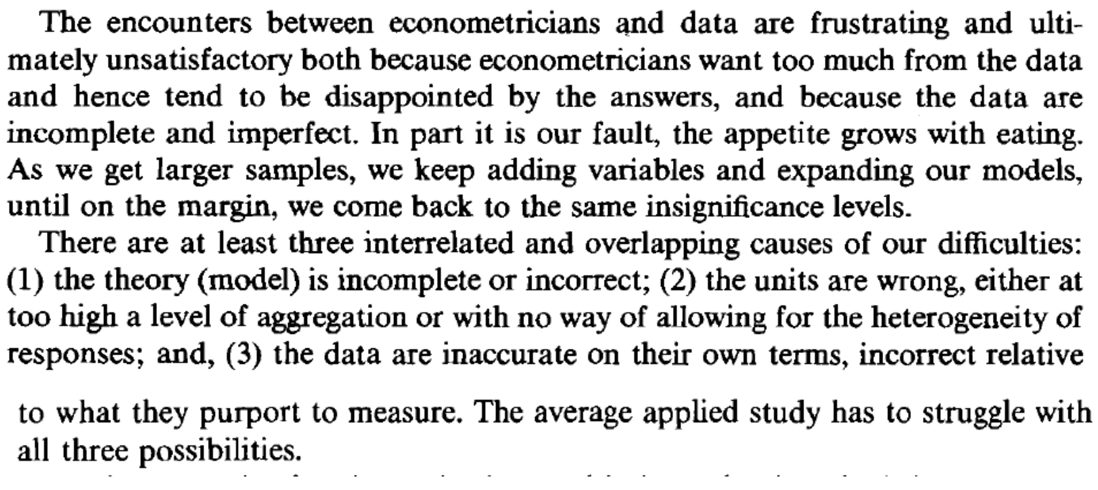


data collection. Thus, there grew up a separation of roles and responsibility. "They" collect the data and "they" are responsible for all of their imperfections. "We" try to do the best with what we get, to find the grain of relevant information in all the chaff. Because of this, we lead a somewhat remote 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 institutions, and is divorced from the direct supervision and responsibility of the analyzing team.

2. Matching the data to the model

We develop a theoretical model using variables that express a very specific theoretical concept. Can we even imagine how we would ask survey questions that would measure that concept? Should we tailor the theoretical/econometric model to the available data, try to get the ideal data for the best theoretical model, or simply make do with what's available and pretend that the data match the theory?

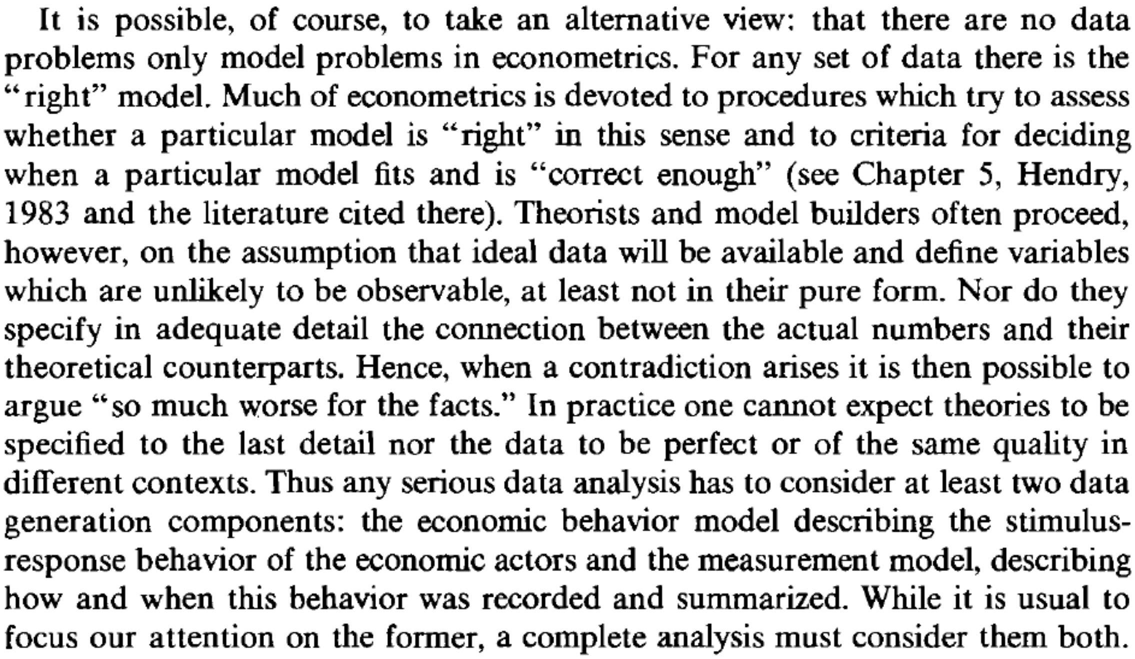
From pp. 1468–69



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 come back to 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.

From page 1469

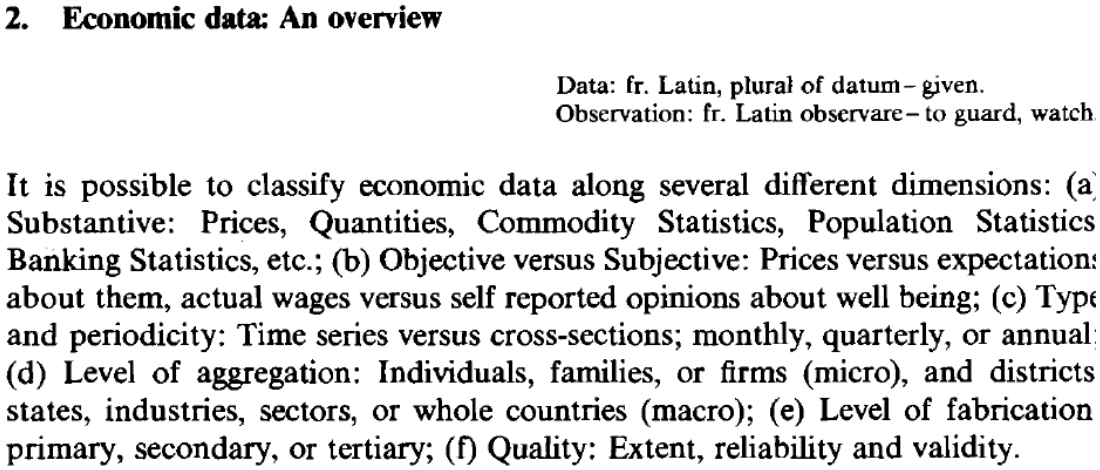


It is possible, of course, to take an alternative view: that there are no data problems only model problems in econometrics. For any set of data there is the "right" model. Much of econometrics is devoted to procedures which try to assess whether a particular model is "right" in this sense and to criteria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

3. How should we begin to consider the characteristics and "quality" of economic data?

What do we mean by "good data"? Can we make such an assessment without reference to the specific use that is being contemplated?

From page 1470:

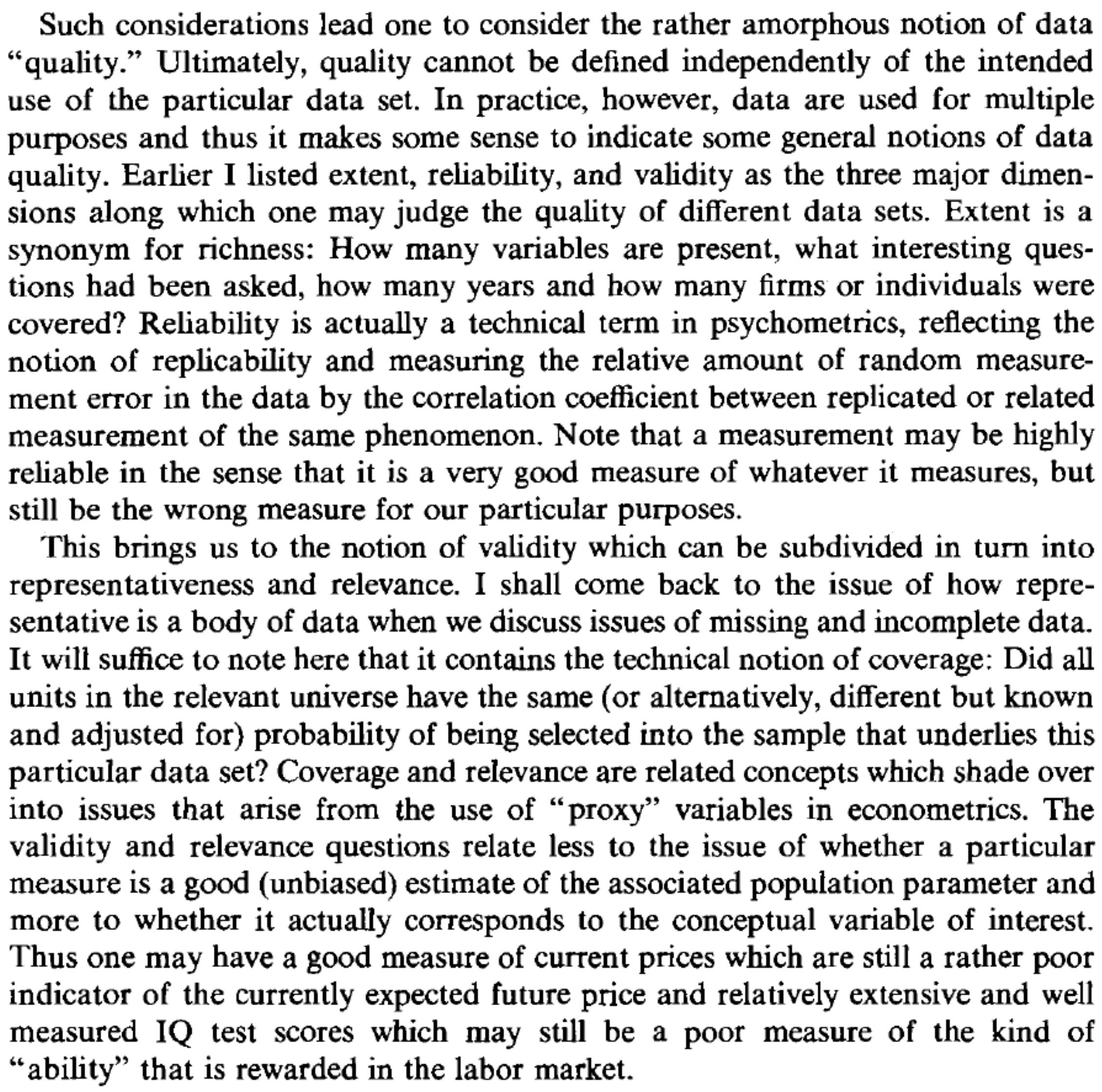


2. Economic data: An overview

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.

From page 1472 (talking about some data problems in the previous pages):



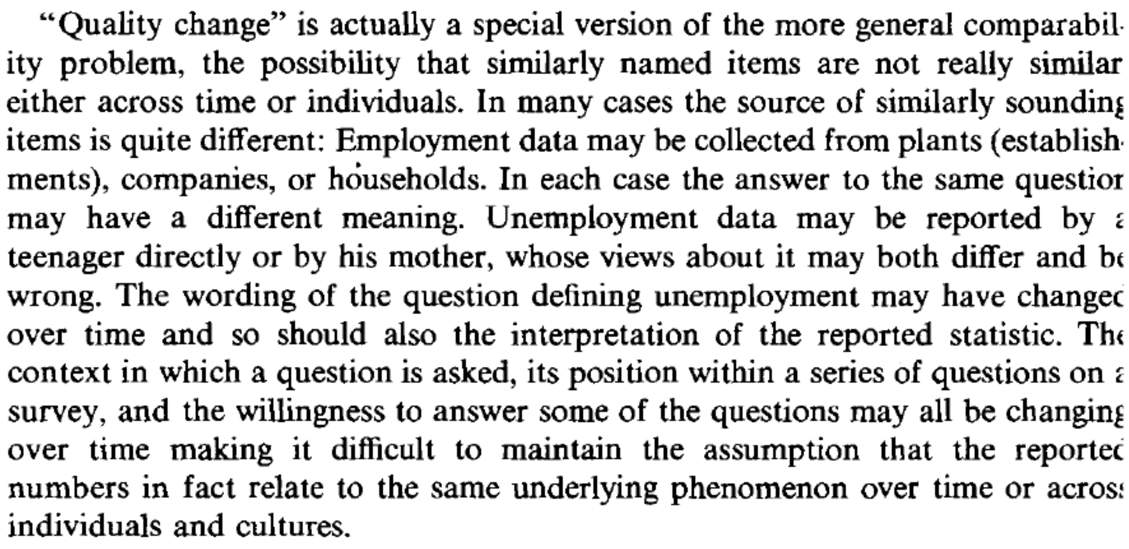
Such considerations lead one to consider the rather amorphous notion of data "quality." Ultimately, quality cannot be defined independently of the intended use of the particular data set. In practice, however, data are used for multiple purposes and thus it makes some sense to indicate some general notions of data quality. Earlier I listed extent, reliability, and validity as the three major dimensions along which one may judge 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 measurement 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 relate 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.

4. Are variables measured in the same way for all observations in the sample?

If the data span a considerable time period, have there been changes in the thing being measured or the method used to measure it during that period? If the data span a large geographic region, are the definitions and collection methods reliably the same for all observations?

From page 1474:

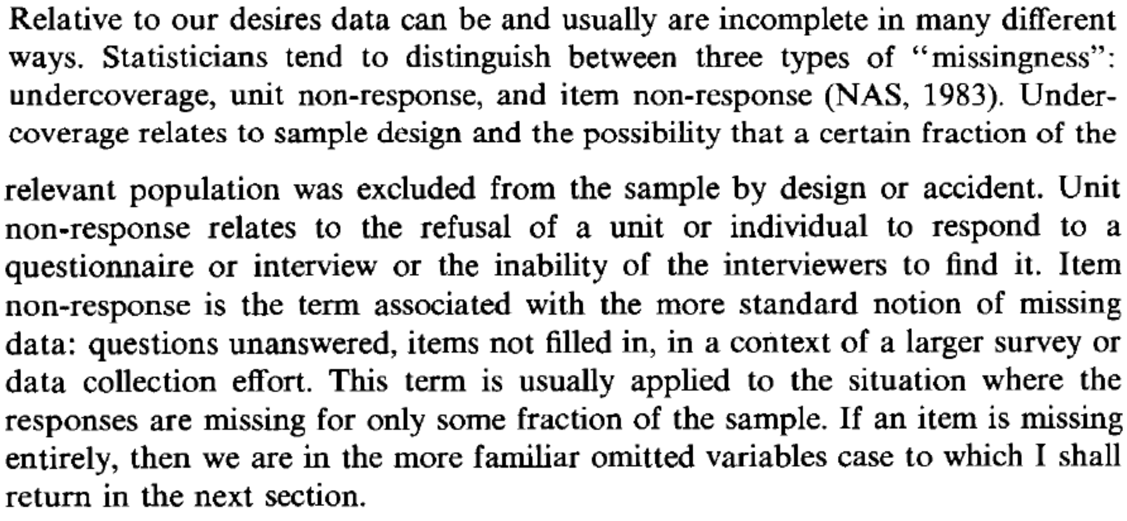


"Quality change" is actually a special version of the more general comparability problem, the possibility that similarly named items are not really similar either across time or individuals. In many cases the source of similarly sounding items is quite different: Employment data may be collected from 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.

5. Missing observations and incomplete data

Almost all data sets have observations for which some or all variables are missing. Not everyone responds to surveys and those who do may not answer every question. How can econometricians cope with these issues?

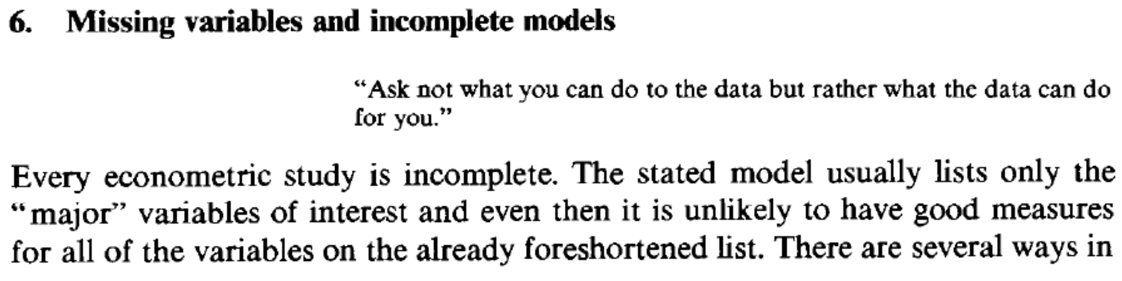
From pp. 1485–86:



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.

The related problem of missing/omitted variables is one that econometricians have studied a lot, but given the limitations of the data have rarely been able satisfactorily to solve.

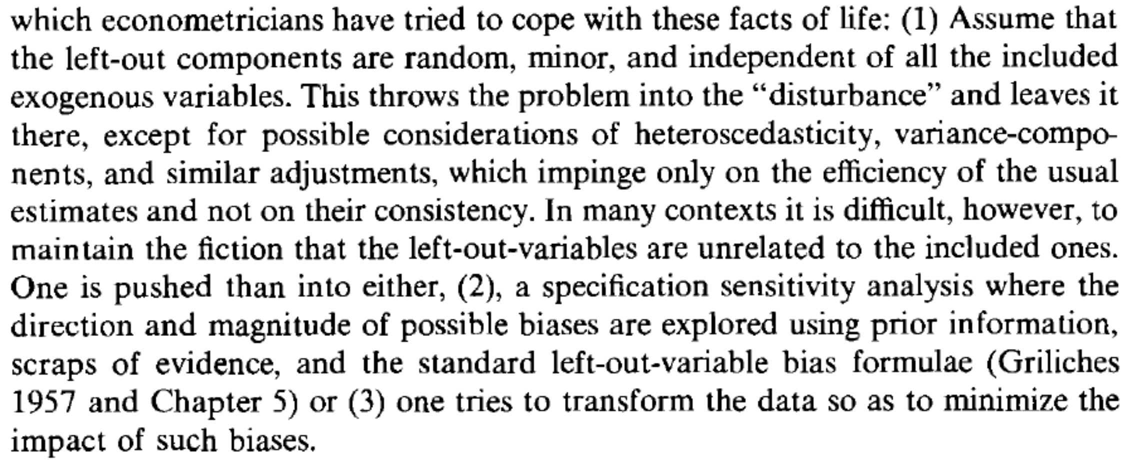
From pp. 1495–96:



6. Missing variables and incomplete models

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

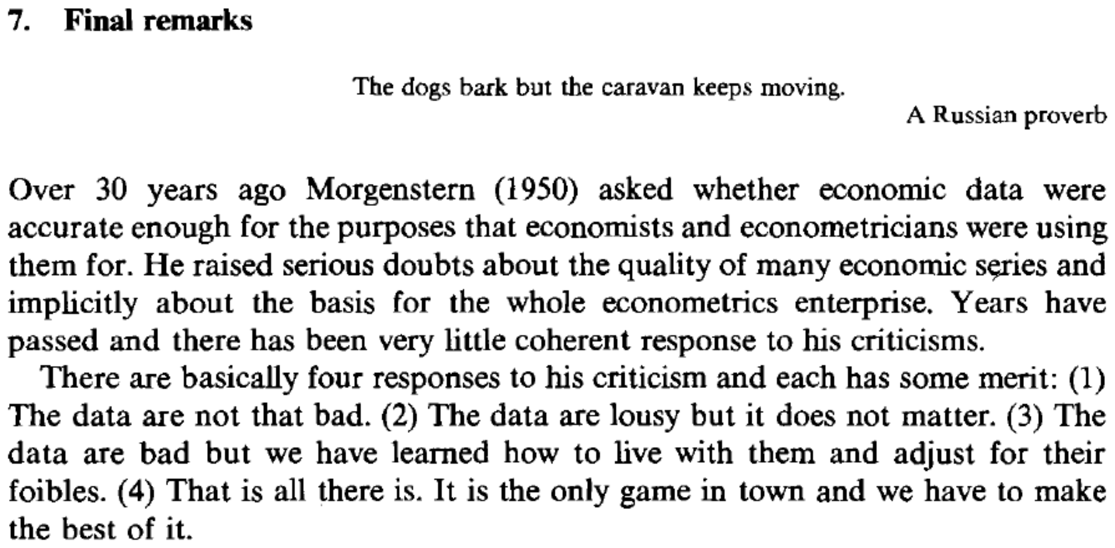
Every econometric study is incomplete. The stated model usually lists only the "major" variables of interest and even then it is unlikely to have good measures for all of the variables on the already foreshortened list. There are several ways in



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 exogenous 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.

7. Where does this leave the econometrician?

From page 1507:



7. 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.