

4 EXPERIMENTAL AND NONEXPERIMENTAL APPROACHES TO STATISTICAL RESEARCH

Esfandiar Maasoumi

Statistical issues that concern researchers in economic education cut across all areas of scientific inquiry that lend themselves to a probabilistic approach. Broadly speaking these issues fall under two closely related headings of sampling and inference.

A first set of issues considered in this chapter concerns the information content of the data in a sample and how this is affected by how the data were observed. Such questions as experimental versus nonexperimental data, randomization across units and/or treatments, nonrandom sampling, missing observations, and selection problems fall in this set.

A second set of issues concerns the closely related questions of inference. Here I will discuss different types of statistical inference, particularly randomization inferences and model-based inferences and how these relate to data observation mechanisms.

In the assertions made here I hope to merely reflect the consensus views among those whom I believe represent the more serious and pragmatic researchers. The following partial listing of these general views will be elaborated upon in the chapter:

1. A greater degree of randomization in selection of observational units, assignment to treatment groups, levels of treatment, etc. is generally desirable whenever possible.

My thanks to Mary Welsh for her skillful typing.

2. The larger the sample of units, treatment levels, etc. the more reliable is the information in the data likely to be.
 3. There is, in general, no such thing as a pure or true experiment. Which is to say, experimental control and randomization are matters of degree and never absolute in any context of practical relevance (at least) to social studies such as economics, education, and psychology.
 4. In general, the desideratum of statistical analysis is external validity or generalizability.
 5. *Internal validity* can be regarded as an intermediate and legitimate concern of any experiment, controlled or otherwise, but such experiment is typically valuable only as a data point (information) in a progressive accumulation of knowledge which brings about *external validity*.
 6. There are often conflicting requirements for internal and external validity, but the latter are themselves complementary and not contradictory.
 7. There are no social studies which in effect can or have entirely escaped the requirements of one type of validity or the other.
 8. A randomization approach to inference, generally made possible with random sampling and/or experimental data, is generally more suited for descriptive inferences and not indispensable for analytic inferences.
 9. A model-based approach to inference, generally favored in economic studies, is convenient as it permits both descriptive inferences and analytic (parametric) inferences. This approach requires strong assumptions which are, however, quite explicit.
 10. An inferential approach that is more formal and explicit with respect to its limitations is generally preferable to one that is not. Studies that over emphasize internal validity at the expense of external validity (generalizability) are less immune to speculative inferences and hidden assumptions.
11. Nonrandom samples do permit valid inferences. The conditions under which this can be done should be more widely known and diligently investigated in every case.
 12. Honest and fuller reporting of the mapping between the underlying assumptions, simplifications, and validity characteristics of the data analysis, on the one hand, and inferences, on the other, requires greater emphasis and deserves greater encouragement.

The following discussion of points 1 to 12 bears upon an ongoing debate in (economic) education research concerning the propriety of using econometric methods in this area. A careful study of the literature in both

this area and econometric theory or, perhaps, even a casual glance at the partial listing of conclusions given above suggests that it is unwise to view the existing methodologies as competitors. There are only good inferences and poor inferences. Educational research is fortunate in having resources that provide lucid discussions of foundations with direct examples from educational research. Econometrics, on the other hand, is now a vast field that has rapidly advanced on the basis of an enormous literature in mathematical statistics and economics. It would be a folly to either characterize econometric methods on the basis of a survey of its elementary textbooks, or to regard it as a completely distinct discipline. Econometric techniques form a large subset of more generally applicable statistical techniques which have been developed by statisticians who were quite familiar with the needs of experimental and nonexperimental research in agriculture, education, economics, biology, psychology, and many other social and natural sciences. It is hoped that some of the references in this chapter alert the reader to some of the literature on foundational issues, and also to what constitutes econometric methodology and its concerns. Affirmation of the benefits of cross-fertilization among social disciplines is in the interest of learning. This is evident in the ever-expanding use of econometric techniques in educational research, and in the increasing and fruitful role of experimentation and experimental methods in econometrics (as may be surmised from, for example, Agner and Morris 1979; and Hausman and Wise 1985). As noted by Hausman and Wise, the U.S. government alone has spent at least half a billion dollars in the last 10 years on socioeconomic experiments. The corresponding analyses by econometricians represent an instructive blend of econometric techniques and what Shapiro (1984a) calls "ed-psychometric techniques."¹ Some authoritative surveys of topics, techniques, and issues that constitute "econometrics" may be found in Griliches and Intriligator (1983, 1984, and 1985).

Section 1 briefly discusses various types of experiments but not questions of optimal experimental design. I am concerned here with true experiments, semi or quasi experiments, and pseudo or hypothetical experiments. Section 2 gives an account of the elements that influence the internal and external validity of all experiments and help to determine the type of an experiment better than any physical-nonphysical allusions. Some examples are given here and a few extensions and interpretations distinguish this section from the main body of Campbell and Stanley (1966). Section 3 deals with randomization and its value to inference. Different types of inferences are distinguished and it is argued that, like "true" experiments, randomization is ideal but not always crucial to

statistical inference. Conditions under which valid inference with non-random samples is possible are exemplified.

Section 4 discusses further distinctions between the experimental and nonexperimental methods within the model-based approach to inference. The importance of clear and full reporting, and of sensitivity analysis, is brought out by example. General conclusions were stated in points 1 to 12.

1. Various Types of Experiments

Real distinctions between various types of experiments are best brought out by noting the factors that bear on their validity for statistical analysis rather than by emphasizing the physical or hypothetical circumstances that surround them. All data, be they made available by laboratory experiments or ex post measurements of natural and social phenomena, are better analyzed systematically than by casual or speculative empiricism. What is crucial is to understand those factors that limit the range and the form of questions that may be asked and inferences that may be safely drawn. Building upon the work of numerous statisticians and their own extensive studies, researchers such as D. Campbell, J. Stanley, H. Simon, H. Wold, H. Blalock, D. R. Cox, R. A. Fisher, Kempthorne, and others have provided the foundations that support statistical inference from experimental as well as nonexperimental data. Our brief discussion below does not deal with methods of optimal experimental design. Relatively recent surveys are given by Aigner (1979) and Herzberg and Cox (1969). Also discussed are issues bearing on the validity and types of experiments that may underlie, or be conceived as underlying, the sample observations. This is generally based on Campbell and Stanley (1966).

We note that in a *true experimental design* the investigator has full control over both the scheduling of observations and the timing and the levels of treatments or stimuli. In particular, the experimenter can in this setting control when to observe and the units he/she wishes to observe, and also when to give the treatments as well as which units will receive which treatments. This full level of control allows full randomization over the units, timing, and treatments. It also allows levels of isolation that are practically impossible in many other contexts.

Often, however, this level of control is not possible on one or more of the items mentioned above. This gives rise to *quasi-experiments*. For example, an educational researcher may be able to choose which courses (students) to observe and choose which semester these are to be observed, but may not be able to fully decide when a treatment (e.g., a test) is to be given and/or who will or will not be given the test.

Quasi-experiments may arise in nature and in social settings. As such they may not be willfully conducted by an investigator who, nevertheless, is able to conceive of them in an opportunistic way. This is the case, for instance, when social or economic accounting measurements are used for statistical analysis. The degree of hypothetical control varies from one set of such data to another, affecting the internal and/or external validity of the "mind experiments" that underlie such observational studies. What is important in each case is an awareness and fuller discussion (at the reporting stage) of those factors that have not been controlled, but which can or do affect the types of questions that may be validly posed (asked of the data), and the types of inferences that may be validly sought.

2. Internal Validity and External Validity

To proceed further we require serviceable notions and definitions for internal validity and generalizability (external validity).

Internal validity is a question of valid attribution. Have the stimuli (treatments) made a difference (resulted in observed effects) which can be unambiguously attributed to them? This is a minimal requirement that should be satisfied by any observational study, be it experimental, quasi-experimental, or nonexperimental (hypothetical experiments).

To the extent that the effects of other, uncontrollable, factors are thought to be present in a study, internal validity is jeopardized. Since this is inevitable to various degrees in any study, a pragmatic position is to find ways of living with it by, for instance, being modest in reporting the results and being explicit about potentially invalidating factors beyond cursory caveats and footnotes. The alternatives to this pragmatic position are evidently either too narrow (pessimistic) or too speculative. I say this since, faced with potentially invalidating circumstances, we may choose either to be so self-righteous as to stop all statistical investigations, or to engage in purely speculative and nonsystematic analysis of available information.

To proceed with the pragmatic-systematic approach it is crucial to be cognizant of the internally invalidating (confounding) effects enumerated below. For more extensive discussion the reader may consult treatises on experimental design, Campbell and Stanley (1966), and Cook and Campbell (1979) for several examples in education research.

1. *External events* are by definition uncontrolled and may have effects that are difficult to disentangle from the treatment effects. A teachers' strike or some other disturbance between observation points within the experimental period is a historical event that may partially explain

the observed differences. Attribution of the latter, either in magnitude or direction, to treatments such as a new teaching method or curriculum, can be and often is problematic.

Self-selection or withdrawal by subjects (units) from an experiment has several generally unknown causes (e.g., mortality, inconvenience, not liking the treatment, etc.) and may seriously affect attribution and thus internal validity.

Differential assignment to or *selection* of treatment groups and control groups (when such exist). An example of this may be useful. Consider providing two teaching methods (or books), one to each of two existing sections of an introductory economics course which are taught by different instructors. The instructor for the first section has a reputation as being serious, relatively dry, but very knowledgeable; the other has a reputation as being a "good entertainer." It is not unreasonable to suspect that more of the studious students will tend to attend the first section. Consequently an external element may operate here which precludes the equation of uncontrollable factors among the different treatment groups. Attribution of observed differences to treatments is proportionately suspect. Randomization in assignment of students to different groups is one way out of this problem. Repetition of the experiment over many (and possibly randomly assigned) teachers is another way of "averaging out" the externalities. This problem must be distinguished from that of sample selection, which can infringe upon generalizability. In other words, a nonrandom collection of course sections (the sample frame) does not affect internal validity, it is the differential assignment of sample units to different treatment groups that is of immediate concern here.

Instrumentation differentials and errors can impinge upon internal validity, as when students participating in an experimental test are scored by different scorers with different standards. Again randomization or equating (averaging) repetitions of the experiment may resolve this problem. Otherwise the questions that are asked may be finer than the data information allows. Instrumentation problems are a much-neglected source of bias that may be quite serious in large cross-section studies conducted by large groups of investigators, or in time series data over any appreciable length of time where, as we know, data collection methods improve gradually. Nonexperimental data are often particularly suspect in this respect. Instrumentation bias, however, is probably most serious in less systematic analyses that attempt to informally put together the results of many previous studies conducted by different investigators and/or at different points in time.

5. *Prior test exposure* biases may be serious, particularly in quasi-experiments with reduced control over the experimental units. This is an allusion to the potential for learning from previous test(s) that may confound the effects of a second test. Panel studies, for instance, generally encounter this problem when the same subjects are regularly observed in order to measure the incremental effect(s) of a treatment.
6. *Cumulative time effects* are sometimes referred to as maturation or aging effects. Passage of time itself can bring about changes in the experimental set-up, for instance in the subjects (units), that may bring about changes or an evolution in responses and the observed effects. Experiments conducted over periods of time are seldom immune to this confounding effect. As we shall see, model-based approaches are better positioned to control for this external effect than descriptive inferences. This is generally the case for many measurable external variables whose effects cannot be equated or controlled for in the primary experimental design.
7. *Regression toward the mean*. There is often a tempting but poor sampling rule that selects units according to their extreme properties. This is often based on the need for treatment, and the greater expectation of observed effects. Suppose extremely poor or very high-scoring individuals are selected after prior test(s). The scores being "unusual" outcomes, their tendency would be to move toward the center of distribution for such scores (in subsequent tests) whatever the treatment given. Once again, the attribution of all or some of the observed effects to the experimental treatment is not justified a priori.
8. *Interaction effects* as, for instance, between time effects and differential assignment or test exposure, are harder to equate across groups since they are often either unknown or are present to an unknown degree. These effects are confounded with the treatment effects, particularly in quasi- and pseudo-experiments.

We have said that the desideratum or the final objective of all sample studies and statistical analysis is generalization. This is so whether or not a particular investigator in a particular study also proceeds to generalize beyond his/her sample. His/her results, if internally valid, will generally be used for inferences about some or several populations that contain the investigated sample. The factors that may destroy or, in any case, limit the generalizability of the sample results should be carefully considered at this stage. These factors are said to affect the external validity of observational studies, experimental or otherwise. As we shall see the following factors impinge on the representativeness of the sample:

treatment levels, policies, educational tests, etc.) can not be erased and thus interferes with the effects of the "current" treatment. In a vague and rather nonexplicit sort of way, dynamic models of econometrics permit us to at least partially disentangle these different effects. Long time series or panel data are generally essential for tackling this problem.

5. *Sample selection*: The generalizability of the inferences may be reduced and serious selectivity biases may be introduced into estimates by sample selection. This is one of the most serious problems that we face in external validation of our studies. For cost reasons, and because of inevitable physical limitations, we are often unable to collect a fully representative sample from the population of interest. For convenience an investigator may randomly select from an available sampling frame. But since this rarely guarantees full representativeness, we should more often limit the range of generalizations to boundaries that are imposed by sample selectivity.
6. *Interaction effects*: Interaction between, for example, the selection mechanism, and its limitations, or among the experimental variables, will effect the representativeness of the sample. As examples, subjects who are not particularly fond of reading may be drawn to a more "active" participation in an experiment on teaching methods that emphasizes class interaction and not any text. Individuals who in any case do not use much electricity at peak-load hours are more willing participants in a price-incentive mechanism experiment that penalizes consumption during peak periods.

It is fair to say that one or several of the factors impinging on internal (eight) and external (six) validity is generally present in any observational study. Applied scientists have been generally more optimistic in their assessment of the severity of such factors than the more-purist theoreticians would like. A priori, and sometimes strong, assumptions are needed (even if implicitly) if the problems are to be substantially ignored. I believe that general and abstract discussions of the propriety of such assumptions is useful only up to a point. In each study we must be explicit about the following issues in the context of that study:

1. *Assessment of the severity* of invalidating factors that we identify in the observed sample;
2. *The importance* of these invalidating factors for the specific inferential and practical goals of our study; and
3. *The limitations* to internal validity or generalizability of our statistical

1. *Pretest or reactive effects*: These effects limit the generalizability of results measured on a pretested (experimental) population to one that was not previously tested.

2. *Experimental behavior or reactive behavior problem*. This is a common and difficult problem that arises when the responses or observed effects in an experimental situation are used to predict responses and effects of even the same subjects in nonexperimental situations. The arrangements and the environment of an experiment may result in environment-specific behavior (responses) that are not fully generalizable to nonexperimental situations. A group of economic agents (or students) when faced with the same stimuli "in nature" may behave or respond differently than what was observed within the experimental arrangements. Social experimentation, related policy studies, and experiments conducted in the field of experimental economics may be seriously affected by this phenomenon. For instance, individuals participating in a peak-load or time-of-day electricity pricing experiment may have measured price elasticities that differ from the elasticities to be observed if the same "optimal" pricing mechanism were imposed after the experiment.

3. *Novel treatment or extreme treatment-level problem*: An example from an important nonexperimental setting is more helpful here, but the same problem can occur with experimental data. In analyzing or predicting the likely consequences of a particular policy (e.g., a tax), investigators typically first analyze the observed responses to various treatments and treatment levels. For example, an econometric or sociological model is estimated on the basis of a representative sample of individuals. If this model is used to predict the response of even the same participants (sample) to an entirely new policy (e.g., flat tax rate, or a new test scoring technique), there is a danger that the response of the respondents may now be structurally or fundamentally different from that measured earlier. This may often result from new treatment levels heretofore not encountered by the participants; but it is probably more severe if a new (novel) treatment is being considered. This problem is the theme of a rather well-known critique of macroeconomic policy analysis by R. Lucas (1976), but economists have been aware of it at least since Jacob Marchak's comments in the 1950s. One way around the problem is to estimate models on the basis of responses to sufficiently varied treatments and treatment levels.

4. *Multiple-treatment interference*: In one-group experimental designs or in macrolevel socioeconomic accounting data, and in almost all non-experimental data, the effects of exposure to previous treatments (or

statements (inferences) that are the direct results of the presence of (at least) the fourteen problems mentioned above and which are not resolved satisfactorily.

Clearly, reasonable investigators can agree to disagree on how severe is "severe," and how important is "important." But while there is little room for disagreement concerning the necessity of being explicit about the "limitations" of our inferences, there is a wide spectrum of standards in reporting.

A final comment is called for. Often, requirements of internal validity may be in conflict with those of generalizability. For instance, for reasons of external validity an educational researcher would prefer to have observations on a randomly selected sample of universities in the United States. In order to have an internally valid experiment (to be conducted, or opportunistic/nonexperimental), however, the investigator will typically need to randomly assign universities to treatment groups (e.g., adopting a policy, a particular text, etc.) and to control or comparison groups (receiving a different treatment or none at all). Since the universities were randomly selected they may not agree to participate (at least in the case of those that are to receive a treatment). External validity will be affected if the units that refuse to participate are dropped. These conflicting requirements of internal and external validity have been confusing to some commentators. It has been argued that one set of methods developed in one area of observational studies is unsuited for use in another area if, as they assert, the two areas have different needs for internal vs. external validity. As have Campbell and Stanley (1966), we have argued that generalization is a desideratum of observational studies (otherwise sampling will make little sense), and that it can only be based on measurements obtained in internally valid experiments that allow clarity in attribution. A single sample (one experiment) itself is never a durable inferential target. Indeed it has value only if it is considered (albeit often informally) as one of several (many) experiments that ultimately permit generalization to some population of interest. I am excluding from this discussion those relatively trivial studies in which the population of interest is not much larger than (or identical to) the sample. Experimenting with teaching aids in an "Economics 101" class in one semester, in order to obtain results that pertain (are generalizable) only to that particular "Economics 101" class that semester, is not, in general, what "ed-psychometrics" or other observational studies are about. Different investigators, whatever areas they are in, may wish to focus on internal or external validity. The balance of their emphasis on the respective requirements will, however, determine the range of inferences they may draw from their studies. To say that an

entire field, such as econometrics, is primarily concerned with external validity and not internal validity not only reveals a dearth of information, it is practically a contradiction in terms.

3. Randomization and Inference

A careful consideration of the fourteen effects mentioned in the previous section suggests that representativeness is a basic requirement in observational studies. Representativeness is required not only with respect to the population to which generalization is to be made, but also by the need for equating factors that may affect attribution (internal validity). Randomization is a powerful method of selection and assignment that may obtain this representativeness. Randomization (equation of conditions) may be achieved with randomization distributions. In social studies, this is frequently either too expensive or, in its strictest sense, physically nearly impossible. A large number of experiments on widely varying units, treatment levels, and situations may be a way of averaging out the non-random but uncontrollable variable effects in some cases. But, it is not true that valid or reliable inferences are not possible with nonrandom samples. What is needed is a knowledge of conditions under which valid inferences are possible. To fill this need, this section will cover the main points of a relatively recent paper by T. M. F. Smith (1983) on the subject. I hope the unfamiliar reader will become interested in further readings in the area (for example, Smith 1976; Little 1982; Godambe and Spratt 1971; Rubin 1976; and Sarndal 1978.)

Types of Inferences

As in Smith (1983), we focus on finite populations of N units which are labelled $i = 1, \dots, N$. Let S denote a subset of $[1, N]$ as a sample, and $A_i = (A_{1i}, A_{2i}, \dots, A_{ki})'$ to be the selector of the units in S according to some sampling scheme such that $A_i = 1$ if $i \in S$ and $A_i = 0$ otherwise. The sampling scheme may be formally represented as

$$f(A_i | z, y, \phi) \quad (4.1)$$

This formalism is a useful and general way of representing the frequent dependence of the scheme on a priori or external information about the units (z), the matrix of unknown values (y) we wish to measure, and possibly unknown parameters (ϕ). We will see that any selection mechanism that does not depend on y can be safely ignored in certain inferences.

Randomization Inference. If A_i is random and dependent only on the known z , then the *randomization distribution*, $f(A_i|z)$, is a probability mass function and is the medium for randomization inferences. In this type of inference y 's are treated as unknown parameters about which inferences are drawn. This set-up is particularly suited for descriptive inferences about known functions of y , such as the totals and means.

Model-Based Inferences. In this case, treat y as having been generated from a super-population $f(y, \theta)$. Again, descriptive inferences (e.g., predicting y 's) are possible. But the greatest advantage of the model-based approach is that it also allows analytic inferences, usually with strong but explicit assumptions. Analytic inferences concern θ and thus the relationship that is postulated between y and θ . This is the main appeal of model-based method in econometrics where the identification of behavioral laws and empirical regularities is of much interest. It is this aspect of economic methodology that represents its concern with external validity. But a careful look at the econometric literature (even in elementary books) reveals that among the problems analyzed in econometrics there are as many problems of internal validity (specification analysis, errors in variables, latent variables models, simultaneous equations models, serial correlation, exogeneity-causality issues, etc.) as there are problems of external validity (tests of structural change, heteroskedasticity, serial correlation, discrete-choice models, error-components and error-covariance models, etc.). In model-based inferences the issue of robustness to the underlying assumptions is important. Thus, many of the above-mentioned topics may be seen to dominate the development of econometrics because of concern with robustness issues, but they also reflect concerns for internal validity as well as generalizability. Although these concerns and requirements are not mutually exclusive, an example may be useful. The problem of nonresponse and missing observations (possibly arising from self-selection or regarded as sample-selection, see section 2) impinge on internal and external validity. It is well known by now that the model-based approach favored in econometrics is particularly suited to the analysis and, sometimes, resolution of these problems (e.g., see Sargan and Drettakis 1973, Little 1982, and their references).

Randomization inference, on the other hand, is rarely possible, almost always more expensive, and subject to several inferential problems relating to such concepts as sufficiency, minimum variance, unbiased estimation (efficiency is not addressed by this approach to inference), and likelihood. (See Smith 1976, for a review.) It may suffice to say that, contrary to a widely held view, randomization inference is not free of assumptions both

because of the restrictions on the range of values for y , and because analytic inference to wider populations are inevitably based on familiar modeling-type assumptions, which are not always explicit. As Smith states:

Practical experience shows that data from social surveys are always subject to non-response and so the analysis of social survey data always requires the statistician to make assumptions beyond those of randomization. A dogmatic statistician who wished to adhere strictly to randomization inference would have to reject social survey data for analysis and retire behind a veil of statistical self-righteousness. (Smith 1983, p. 398)

And,

Most statisticians would agree that they should help analyze non-random samples, but that in so doing they should make quite clear the limitations to their conclusions. A model-based approach to inference allows the statistician to analyze nonrandom samples in a formal way while at the same time making explicit the underlying assumptions. (Smith 1983, p. 398)

To be sure, there are bad model-based inferences and good ones. This has to be judged in the light of the validity requirements that have been discussed. But it will also help to understand the effect of *selection* on model-based inferences, and inference from nonrandom samples.

Selection and Inference

In the model-based approach a first and often difficult task is to specify the probability density (or mass) function that generates y and z (and prior distributions in a Bayesian approach). One usually works with the conditional p.d.f., $f(y|z, \theta)$. This also explains the appeal of linear regression models. Given the selection distribution defined earlier, we have:

$$f(y, A_s | z; \theta, \phi) = f(y_s | z; \theta) f(A_s | y, z; \phi) \quad (4.2)$$

If y_s can only be observed when A_s has been, we have a partition of $y = (y_s, y_r)$ where \bar{y} is the complement of the sample (S). The distribution of interest is thus given by

$$f(y_s, A_s | z; \theta, \phi) = \int f(y_s | z; \theta) f(A_s | y, z; \phi) dy_r \quad (4.3)$$

We shall now see that if selection (random or otherwise) is to be ignored in our inferences about θ we must have selection that does not depend on the measurement variables (y). That is, we must have

$$f(A_s | y, z; \phi) = f(A_s | z; \phi) \quad (4.4)$$

only prior variables, z , determine the selection of units and not y (see Rubin 1976). Under this condition it can be verified that

$$f(y, A_s | z; 0, \phi) = f(A_s | z; \phi) f(y | z; \theta) \quad (4.5)$$

This last result holds for all possible A_s , we have a conditional independence factorization and, provided that ϕ and θ are distinct (or a priori independent in a Bayesian model), inferences on θ may be based on

$$f(y_s | z; \theta) \quad (4.6)$$

This is because for given A_s , sampling distributions generated by the last d.f. are identical to those generated by $f(y_s, A_s | \cdot)$ factorized as above. Random sampling guarantees this condition and hence is ideal even for model-based inference. However, the important condition (equation 4.4) is also satisfied by such other sampling schemes as "balanced sampling" and "purposive sampling" of high- z units. Thus any sampling that is solely based on the a priori z -values may be ignored without threat to external validity of inferences since the selection itself has no additional "information" (for known z). Interestingly, if z is unknown, as it might be to subsequent users of the same body of data, the design contains additional useful information not ignorable for those who do not know z . Since most investigators wish to use data that they themselves do not collect, there is a preference for simple random sampling as the only design that has no useful additional information.

In educational and other observational studies we often observe a convenient sample, for instance, those taking an introductory economics class. Here z is an indicator function and we need to specify $f(y | z; \theta)$ in order for inferences about "other groups" to be possible.

When poststratification is also employed, we may ignore the selection criteria (e.g., Introductory Economics enrollment) if it contains no information other than the stratification variables (criteria). For instance, as stratification variables, social class, age, and sex must contain all the information that a unit belongs to an introductory economics class. This is often unlikely.

Another nonrandom sampling technique is quota sampling. Here one needs to fill a quota of units on the basis of classifying variables such as sex, age, average grades in other courses, etc. These classifying variables, however, are generally known only after prior selections based on z . The decision to measure unit i now depends on both the quota requirements and on whether i is initially selected. This means that we now need both selections to be independent of the measurement variables (not the quota variables) for inferences to ignore them. (See Smith 1983, for a formal statement of these conditions.)

Another type of nonrandom sampling arises because of nonresponse, and missing values. Generally this type of sampling is nonrandom to an unknown degree. But the model-based approach does offer certain, often data-based solutions which, however, require certain modeling (or regularity) assumptions. (See the papers in the U.S. National Academy of Sciences 1980.)

4. Further Distinctions Between Design Approaches

We have argued that the distinction between the experimental and non experimental approaches is a matter of degree and not in kind. This is because there is no such thing as a "perfect" experiment, most real-life experiments lying rather far from the valid region defined in section 2. Thus statistical methods developed largely in connection with nonexperimental and quasi-experimental studies (such as in econometrics) are also useful, indeed indispensable, for the study of real-life experimental data. The further are the data from a perfect experimental setting, such techniques face increasing challenges to their successful operation and philosophical meanings.

In the modeling approach to inferences with real-life data, linear multiple regression has played a powerful role. While I am happy to note a diminishing trend in the exclusive use of this technique in applied research (particularly in econometrics), I will follow the example of Leamer (1983) in using this simple setting to discuss several issues pertaining to "good" inferences and "good" reporting of specification searches. This will also bring out some of the points we raised earlier, particularly with respect to randomization.

Consider an example in which the effect of a certain instructional method is under scrutiny. A "pure" experiment for this might be to have the same instructor, using the same textbook, teach two randomly selected sections of a given course. One section receives a new instructional technique, the other does not. Test scores of both sections are obtained by the same scorer in identical tests. There are no drop-outs or adds in either section. Assignment to different sections is random and (hypothetically) enforceable. Student backgrounds are "identical" or equated, maturation is equated, etc. The inferential question may be whether the new instructional method had an impact or increased the scores of students in the treatment group compared to the control (or comparison) group. Note that randomization in this ideal experiment does not have to mean that all the relevant external variables are equated in *this* sample. All that is needed is an equation of such factors on average, the "average" having its meaning

in a repeated-sampling sense. Thus randomization makes such equation of external variables merely more probable but not certain in every sample.

In these ideal conditions, we can analyze the scores by

$$S_i = \beta_1 + \beta_2 T_i + U_i \quad i = 1, \dots, N \quad (4.7)$$

where N is the number of students in the sample and T_i measures the treatment level given to the i th student (e.g., the number of hours of one-to-one interaction with the instructor, or the number of hours of reading required, etc.).²

In real life, however, the equation of the effects of all the a priori pertinent external variables is almost always impossible, particularly in economics and educational research. For instance, it is hard (if not impossible) to measure student ability and control for its effect (interaction with the treatments). Background variables are hard to measure and rather hard to list in an exhaustive manner. Let one of these other measurable variables be denoted by A_i , $i = 1, \dots, N$; and for the sake of argument we assume that there are no other "important" variables to consider. If A_i and T_i are neither perfectly correlated nor orthogonal, a multiple regression can separate their respective influences on S_i (collinearity causes a problem only for the identification of the separate effects). Hence the correct model is

$$S_i = \beta_1 + \beta_2 T_i + \beta_3 A_i + \epsilon_i \quad (4.8)$$

If it is assumed that $E(A_i | T_i) = \alpha_1 + \alpha_2 T_i$, and $E(\epsilon_i | T_i) = 0$, we have

$$\begin{aligned} E(S | T) &= \beta_1 + \beta_2 T + \beta_3 E(A | T) \\ &= \beta_1 + \beta_2 T + \beta_3 (\alpha_1 + \alpha_2 T) \\ &= (\beta_1 + \alpha_1 \beta_3) + (\beta_2 + \alpha_2 \beta_3) T \\ &= \gamma_1 + \gamma_2 T \end{aligned} \quad (4.9)$$

Consequently, for example, least-squares estimation on the basis of equation 4.7 provides biased inferences about β_1 and β_2 . There is no problem in drawing inferences about (γ_1, γ_2) , but these are not the parameters that are meaningful inferential targets for us. Working with equation 4.7, the data provide no information about α_1 , α_2 , and β_3 that can be used in order to disentangle the information on (β_1, β_2) and (γ_1, γ_2) . There are three ways to proceed (apart from abandoning all statistical work) at this stage.

1. Improve the experimental conditions so as to remove the influence of A . There is no ambiguity about this option. If it can be done and can be afforded relative to the benefits of the results, then it must be done.

2. Measure and control for A , in equation 4.8. There is no ambiguity about this option either. If it can be done and can be afforded then it must be done. An additional problem (aside from measurement error issues and costs), however, is that in reality there are many more candidate variables than just one A . Indeed, the a priori set of potentially influential variables grows with the decline in the degree of experimental control and the decline in the representativeness of the sample (random sampling being only one method of achieving representativeness).
3. Bring such a priori or extraneous statistical information as might be available to bear on the analysis of short models (e.g., equations 4.7 or 4.9) when long models (such as 4.8) cannot be dismissed with certainty. When are we certain about such dismissal? In our example we may assume that ability is not a factor here ($\beta_3 = 0$) and so solve the problem ($\beta_1 = \gamma_1$ and $\beta_2 = \gamma_2$).³ But we cannot be sure of this in nonexperiments and in quasi or opportunistic experiments (sometimes referred to as *natural* experiments). Thus we are typically uncertain and relatively uninformed about such biasing quantities as $\alpha_1 \beta_3$ and $\alpha_2 \beta_3$; in some problems we are more uncertain than in others (and some people are more uncertain than others in any context).

Unlike options 1 and 2, there is a good deal of ambiguity, vagueness, and specificity in pursuing option 3. Investigators and analysts can be surprisingly speculative and informal in making qualitative (corrective) statements about their inferences. This is very poor empiricism and has no place in scientific induction. Consequently, one must seek formal and systematic ways of representing a priori information about such parameters as $\alpha_1 \beta_3$ and $\alpha_2 \beta_3$, and of investigating and reporting the sensitivity or "fragility" of our inferences to variations in such prior information. It is accurate to say that most investigators in almost every field of science needing statistical tools practice option 3 in one form or other. Because the underlying problem is generally so very difficult, most practice this option rather poorly, which explains why there is a new and thriving industry in econometrics pontification (and possibly in other fields as well). One approach is to continue with classical methods of statistical analysis by being "more careful," by "building down" from larger models to smaller ones (within the bounds of possibility), by testing and testing more vigorously and at more stages, etc. Although there are undoubtedly better modes of doing classical empirical analysis than we frequently encounter, physical bounds on profligacy in model building and measurement problems (e.g., A_1 as ability), at least, define sometimes severe bounds on how well and reliable the outcomes will be. Another systematic approach

within option 3 is the Bayesian technique of bounded influence (variance) priors advocated by Leamer (1978, 1983). Before discussing the merits and/or limitations of option 3, it would be useful to see how it may be applied in the above example.

As we have said, option 3 requires a priori information on $\alpha_1\beta_3$ and $\alpha_2\beta_3$. Noting that

$$\begin{aligned} S &= E(S|T) + \epsilon \\ &= \gamma_1 + \gamma_2 T + \epsilon \\ &= \beta_1 + \beta_2 T + (\alpha_1\beta_3 + \alpha_2\beta_3 T) \end{aligned} \tag{4.10}$$

It is straightforward to show that the ordinary least-squares (OLS) estimates of (β_1, β_2) in equation 4.10 are identical with the generalized least-squares (GLS) estimates $(\hat{\beta}_1, \hat{\beta}_2)$ which use the composition of ϵ . But, from Leamer (1983),

$$\text{var}(\hat{\beta}_1, \hat{\beta}_2) = \text{var(OLS)} + \Omega \tag{4.11}$$

where Ω is the covariance matrix of a normal prior p.d.f. from which $\alpha_1\beta_3$ and $\alpha_2\beta_3$ are drawn. This prior distribution can be centered at zero with smaller covariance matrices Ω to reflect greater a priori belief in the nonimportance of the variable A . Larger Ω variances represent greater uncertainty about $\alpha_1\beta_3 = 0 = \alpha_2\beta_3$ (and the model given in equation 4.7). Ω is entirely a prioristic here and it may be regarded as a measure of the formal difference between a randomized (ideal) experiment and a "natural" experiment. Full randomization, either with respect to internal or external validity, may be represented by specification $(\alpha_1\beta_3 = 0 = \alpha_2\beta_3$ and/or $\Omega = 0)$. Least-squares estimates will be unbiased (or consistent) in this situation and valid, unbiased, inferences are possible. In other situations fixed levels of uncertainty remain with respect to external effects that influence the internal and/or external validity of statements made about the effect of T on S . If we find that inferential statements (for example about β_2) are robust with respect to relevant changes in Ω we have "nonfragile" inferences. In any case we must report the mapping between Ω (prior p.d.f.s in general) and the inferences (e.g., estimates of β_2).

There are a number of practical problems with this approach. A particular choice of prior p.d.f. is not generally one of these problems (aside from philosophical objections that some may have) because it is the mapping that is important here, and changing Ω is a useful means of achieving an "indeterminate prior probability" robustness for our inferences (see Leamer 1978 and Maasoumi 1976, 1986) Changing Ω , however, is not an exhaustive means of sensitivity analysis here.

One problem that remains is what to do when A cannot be measured (either cheaply, or accurately, or at all) and one believes that $\alpha_2\beta_3$ is properly drawn from a distribution not centered at zero. Another problem is quite serious in view of the fact that statistical advice is both demanded and supplied and is likely to remain so in the future. This problem arises from the fact that nonfragile inferences, so defined, are not to be expected in real-life where, as Leamer (1983) himself argues, the number of potential variables (A) in nonexperimental settings is without known bounds. Indeed, we can see this problem very clearly by noting the existence of an indecisive classical sensitivity analysis which is practically (if not philosophically) equivalent to this Bayesian bounded-variation fragility analysis (see McAleer, Pagan, and Volker 1985).⁴

The discussion above illustrates one notable feature of nonexperimental inference: the inevitability of specification searches by analyzing the same data set with many different models. This feature reflects the legitimate concern about lack of control in "nonexperimental" settings. But we have also discussed the possibility of inference with nonrandom samples and the role played by the model-based approach in that context. Indeed, it can be argued that the concern for lack of randomization is another feature of statistical analysis of nonexperimental data that has led to further development and use of the model-based approach to inference. I am referring here to the development of simultaneous equations models (SEM) in econometrics and the instrumental variables (IV) techniques. A brief (but abstract) example may be useful (see Leamer, 1985a): Suppose the effects, y , are to be measured of the (pseudo) administration of the treatments, z , which however cannot be reasonably randomized. Thus there will be considerable doubt about inferences from, for example, the following regression model:

$$y = z\beta + u$$

In such situations it has become popular to explain the generation of the (nonexperimental) treatments z by further relations such as

$$z = X\delta + \epsilon$$

where, within the context, X variables are reasonably thought to influence z . If $X\delta$ is approximately all that there is to the systematic part of z , and ϵ is a white-noise error term, then we have a two-equation model that solves the nonrandomization of z . This requires a belief in the randomization (or exogeneity) of the X variables and they must also be measurable. When this is so, X 's may be used as instrumental variables. Otherwise one must continue to explain some or all of the X variables by further equations in a

larger SEM. If errors of measurement are of concern, an errors-in-variables model must be analyzed within or without a SEM. The reader would have noted in this process the loose end that is associated with the assumption of randomness (exogeneity) of some variables at some stage. Thus, here as in the specification search described earlier, there are no perfect solutions because there are no perfect experiments.

It would seem that there are no fully practicable, sure-fire methods that can resolve all of the real-life problems we face, with or without experimental data (short of perfect ones). Thus I feel justified in concluding that honest reporting, be it of the data collection method or of data analysis and models, is at least as important at this juncture as improved methods of data collection and analysis. I do not see much distinction in this respect between various fields, especially within the social sciences. It is perhaps redundant to emphasize that fields of inquiry in which option 1—experimentation and better experimentation—is a realistic alternative should justifiably consider the relative merits and the net cost-benefit of conducting more and better experiments versus better handling of the much cheaper nonexperimental and quasi-experimental data.

Notes

- 1 Also see comment on Shapiro article by Becker and Walstad (1984) and his response (Shapiro 1984b).
- 2 For the discussion below, the control or comparison groups are not necessary. These were included earlier since it is generally held that an ideal experimental design should typically have a control or comparison group.
- 3 Also, if we know or find for certain that $\alpha_1 = 0 = \alpha_2$, then we do not need to control for A and no additional work (e.g., experimentation) is called for.
- 4 Also see comments on McAleer, Pagan, and Vaulker (1985) by Leamer (1985b) and Cooley and Leroy (1985).

References

- Aigner, D.J. 1979. A brief introduction to the methodology of optimal experimental design. In D.J. Aigner and C.N. Morris (eds.), *Experimental Design in Econometrics*. Amsterdam: North-Holland.
- Aigner, D.J., and C.N. Morris, eds. 1979. *Experimental Design in Econometrics*. In *Annals of Applied Econometrics*. Amsterdam: North-Holland.
- Becker, W.E., and W.B. Walstad, 1984. A misrepresentation of econometrics and educational research. *Educational Researcher* 13(9):23-24.
- Blalock, H.M. 1964. *Causal Inference in Non-Experimental Research*. Chapel Hill:

- University of North Carolina Press.
- Campbell, D.T., and J.C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Cook, T.D., and D.T. Campbell 1979. *Quasi-Experimentation: Design of Field Settings*. Boston: Houghton-Mifflin.
- Cooley, T.F. and S.F. Leroy. 1981. Identification and estimation of money demand. *American Economic Review* 71(5):825-844.
- Cooley, T.F. and S.F. Leroy. 1985. What will take the con out of econometrics? A reply to McAleer, Pagan and Volker. University of California, Santa Barbara. Mimeographed.
- Cox, D.R., and E.J. Snell. 1974. The choice of variables in observational studies. *Applied Statistics* 23:51-59.
- Fisher, R.A. 1971. *The Design of Experiments*. 8th edition. New York: Hafner Press.
- Goldambe, V.P., and D.A. Sprout, eds. 1971. *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.
- Griliches, Z., and M.D. Intriligator, eds. 1983, 1984, 1985. *Handbook of Econometrics*. Vols. 1, 2, and 3 Amsterdam: North-Holland.
- Hausman, J.A., and D. Wise. 1985. *Social Experimentation*. Chicago: University of Chicago Press.
- Herzberg, A.M., and D.R. Cox. 1969. Recent work on the design of experiments: A bibliography and a review. *Journal of the Royal Statistical Society A* 132:29-67.
- John, J.A., and M.G. Quenouille. 1977. *Experiments: Design and Analysis*. 2nd edition. London: Griffin.
- Leamer, E. 1978. *Specification Searches: Ad hoc Inferences with Non-Experimental Data*. New York: Wiley.
- Leamer, E. 1983. Let's take the con out of econometrics. *American Economic Review* 73(1):32-43.
- Leamer, E. 1985a. Non-experimental inference. In S. Kotz and N.I. Johnson (eds.), *Encyclopedia of Statistical Terms*. New York: Wiley.
- Leamer, E. 1985b. Sensitivity analyses would help. *American Economic Review* 75(3):308-313.
- Little, R.J.A. 1982. Models for non-response in sample surveys. *Journal of the American Statistical Association* 77:237-250.
- Lucas, R.E., Jr. 1976. Econometric policy evaluation: A critique. In K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labor Markets*. Amsterdam: North-Holland. Pp. 19-46.
- Maasoumi, E. 1976. A quasi-ML method for the estimation of the coefficients of the reduced forms of simultaneous equations. Unpublished paper, University of Birmingham (England).
- Maasoumi, E. 1986. Reduced form estimation and prediction from uncertain structural models: A generic approach. *Journal of Econometrics*, 31:3-29.
- McAleer, M., A. Pagan, and P. Volker. 1985. What will take the con out of econometrics. *American Economic Review*. 75(3):293-307.

- Popper, K. R. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.
- Rubin, D. B. 1976. Inferences and missing data. *Biometrika* 63:581-592.
- Sargan, J. D., and E. G. Dretakis. 1974. Missing data in an autoregressive model. *International Economic Review* 15:39-58.
- Sarndal, C. E. 1978. Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, 5:27-52.
- Shapiro, J. 1984a. On the application of econometric methodology to educational research: A meta-theoretical analysis. *Educational Researcher* 13(2):12-19.
- Shapiro, J. 1984b. A reply to Becker and Walstad. *Educational Researcher* 13(9):25.
- Smith, T. M. F. 1976. The foundations of survey sampling: A review. *Journal of the Royal Statistical Society A* 139:183-204.
- Smith, T. M. F. 1983. On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society A* 146:394-403.
- Stanley, J. C. 1961. Studying status vs. manipulating variables. In R. O. Collier and S. M. Elan (eds.), *Research Design and Analysis: The Second Phi Delta Kappa Symposium on Educational Research*. Bloomington, IN: Phi Delta Kappa. Pp. 173-208.
- Stanley, J. C. 1966a. A common class of pseudo-experiments. *American Educational Research Journal*, 3:79-87.
- Stanley, J. C. 1966b. The influence of Fisher's "The Design of Experiments" on educational research thirty years later. *American Educational Research Journal* 3:223-229.
- U.S. National Academy of Sciences. 1980. *Panel on Incomplete Data*. Washington, DC: National Academy of Sciences.
- Wold, H. 1980. Causal inference from observational data: A review of ends and means. *Journal of Royal Statistical Society A* 119:28-61.

5 MEASUREMENT INSTRUMENTS

William B. Walstad

An understanding of measurement is essential for work in all areas of education. For example, a noted measurement authority has stated:

In today's educational milieu just about 50 percent of the problems we encounter do, in fact, involve test use, test construction, or test interpretation. Consequently, just about any kind of specialist who, lacking knowledge about measurement, goes out to do battle with today's educational problems is almost certain to come back a loser. For the present and foreseeable future, educators who wish to be effective in their work simply must master the major tenets of educational measurement. (Popham 1981, p. 4)

This recommendation for educators also applies to researchers in economic education. Knowledge of econometric techniques is not sufficient to do the work; a sound understanding of measurement principles is also required.

In essence, empirical work in economic education begins with measurement. We can identify research problems, specify hypotheses, and construct an elaborate research design; but we must start our empirical studies with measurement. Many worthwhile research ideas have probably been abandoned for lack of available instruments to measure important inputs or outputs. The continuing work also depends on measurement. Statistical tests are based on comparisons among measures, and conclu-

Thanks are expressed to Steven Wise for helpful comments on an earlier draft.