Journal of Official Statistics
Vol. 7. No. 1, 1991, pp. 7–24

Sources of Uncertainty in Modeling Social Statistics: An Inventory[1]

Clifford C. Clogg[2] and Aref N. Dajani[3]

Abstract: We describe the major sources of uncertainty in modeling social statistics. A more comprehensive definition of a *statistical model* is developed for this setting. The components of the modeling process enumerated here alert both the analyst and the consumer to sources of uncertainty that might not be measured well by conventional measures of precision (standard errors) or goodness-of-fit summaries. We also try to assess how well some of our *supermodels* convey the true uncertainty in our conclusions.

Key words: Uncertainty; supermodels; social survey; complex samples; computer revolution; sampling frame.

1. Scope and Objectives

Our main goal is to enumerate the sources of uncertainty that affect conclusions reached from statistical models applied to social data (or social statistics). A secondary goal is to examine implications for the modeling enterprise. Although measuring uncertainty has always been an integral part of statistical science, we do not think it is necessary to say very much about the history of statistical modeling, even though aspects of that history have direct bearing on our topic. Stigler (1986) provides a definitive history of statistical modeling prior to the 20th century that can be consulted for those details. This source is important for many reasons. For example, Stigler demonstrates how basic tools such as regression, least squares, and correlation or association measures were developed, to a considerable extent at least, in the context of social statistics and empirical social science throughout the 19th century. Duncan (1984) provides a history of models used to produce social measurements that is also relevant for background. Duncan's work directs attention to models used to *combine multiple measurements* which rely on latent variable concepts. The problem of combining multiple measurements ("multiple indicators") is a major issue in most research settings involving social statistics. In addition

[1] A previous version of this paper was presented at the 1989 Annual Meetings of the American Statistical Association and appears as "Modelling Social Statistics: Current Issues," pp. 214–225 in Sesquicentennial Invited Paper Sessions: Proceedings of the American Statistical Association, Washington, D.C.: The American Statistical Association, 1989. This research was supported in part by Grant No. SES-8709254 from the National Science Foundation and by a grant from the Russell Sage Foundation. We thank Glenn Firebaugh, Richard Rockwell, the editor, three referees, and several colleagues for helpful comments. Address correspondence to the first author.
[2] Distinguished Professor, Departments of Sociology and Statistics, and Senior Scientist, Population Issues Research Center, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.
[3] Department of Statistics and Population Issues Research Center, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.

to these two fine books, some insights on the history of modeling can be inferred from contemporary textbooks in econometrics, psychometrics, and sociological methodology, to name just a few areas where models for social statistics are emphasized. In short, adequate histories of the modeling enterprise are available or are at least implicit in standard sources, so we can leap across the historical landscape and take the history of statistical ideas for granted.

Another major reason for downplaying the importance of the historical details is that the modeling enterprise has changed radically since the 1960s. The computer revolution is the main factor responsible for this dramatic change. Computer technology has fundamentally altered the ways in which we collect, organize, and distribute data. Inexpensive, versatile, and efficient computing tools have allowed the development of the main statistical models used in modern social research, most of which involve iterative estimation of models for highly multivariate data. For example, the tools of modern factor analysis (or covariance structure analysis) could not have developed as they have without cheap computing. The same can be said for most of the econometric techniques that have been so influential in recent years. Models for large contingency tables and models for complex event histories would be little more than abstract theory without the availability of modern computer technology, which now includes powerful desktop hardware equipped with flexible software. To be practical, it seems necessary to limit the subject by concentrating on the modeling enterprise as it now exists.

The topics we address are as follows. A general definition of a *statistical model* which is suited for the social-statistics setting is presented first and contrasted with a conventional definition. The parts of the modeling process that are laid out in Section 3 are appropriate for the applications we have in mind. These parts of a statistical model call attention to sources of uncertainty that are bound to have a major effect on inferences. The point is that we cannot afford to ignore sources of uncertainty brought about by a variety of statistical decisions that precede the phase of the process where some regression-type equations are estimated. Finally, we briefly consider how some of the supermodels so popular in the analysis of social statistics attempt to deal with the problem of measuring uncertainty beyond that attributed to "sampling error."

2. A Statistical Model Defined

Perhaps the most natural way to define a statistical model is to borrow from the terminology of *generalized linear models* (McCullagh and Nelder 1989). This is the approach that serves as a standard in mathematical statistics, particularly in those areas devoted to solving problems that have bearing on various applied areas.

A generalized linear model begins with two sets of measurements, both assumed to be available for a set of N units. (Unit "i" is the typical case.) One set is usually called *covariates* $(x_1, \ldots, x_k)$ and typically regarded as *fixed*; "predictors," "explanatory variables," "exogenous variables," or even "independent variables" are synonymous with covariates, in spite of the fact that each ostensibly synonymous term involves different philosophies of modeling objectives. The other set is usually called the response (Y), with expectation $E(Y) = \mu$. "Endogenous variable" and "dependent variable" are the two most common alternative names for this variable. Measurements are taken on a sample of N units, and it is almost always assumed that the measure-

ments are independent across sample units. (A more exact statement would be that the N measurements on Y are independent, or at least uncorrelated, given the X values observed.)

The goal is usually to summarize the data consisting of N measurements on x's and $Y = y$ so that Y can be predicted or explained. It is typically assumed that the set of covariates has already been selected (usually by someone else), although given a relatively small set of x's (k less than 15 or so), the modeling task might often be posed as a problem in selecting a still smaller set of covariates from a set that has already been trimmed down to a considerable extent. A *univariate* Y is taken for granted. This setup is still general enough to allow for different Y's predicted with different sets of x's, as with a simultaneous equation model. (The simultaneous equation model is a set of individual "model equations," each linking a specified response to a subset of the x's or to other response variables taken as x's.)

We next formulate an appropriate *link function* g so that $\eta_i = g(\mu_i) = \Sigma_j \beta_j x_{ij}$, with i denoting the unit of observation and j denoting the particular covariate (x_j). Here, β is the parameter vector of interest. It summarizes the relationship between the covariates and the response, determines the degree to which the response can be predicted from the x's, and, in some cases at least, permits the attribution of casual effects. We are ordinarily interested in the *size* of β, the precision of an estimator of β (such as standard errors or functions of them), and various summaries of model adequacy like goodness-of-fit (which is a function of the size of β and its precision). The linear predictor represents the *systematic* (or expectation) part of the model; departures from expectation are conceived as random and constitute the "error" component. Various assumptions for the error are possible, and these assumptions influence how the size of β or its precision are determined.

All the ingredients of a statistical model are present in the above formulation: a response variable or a transformation thereof, covariates (possibly including experimentally defined factors), systematic and random inputs assumed to determine the response, and linearity in the parameters, made possible by judicious choice of the link function. A rich body of theory and a battery of computational tools for applying the model to data (N measurements on Y and the x's) are available. These include a host of possible estimation methods (e.g., maximum likelihood, generalized least squares, robust methods, quasi-likelihood, etc.). (See McCullagh and Nelder (1989) for details and references; see Thisted (1988) for computational algorithms.) They also include formal and informal (e.g., graphical) tools for assessing goodness-of-fit, and both formal and informal methods for detecting bias created by excluding (or including) certain variables when they should have been included (or excluded). With the advent of inexpensive computing since the early 1980s, computer-intensive tools involving resampling ideas have been added to the toolbox. Bootstrap, jackknife, multiple-imputation, and cross-validation techniques belong in this category, and they give us greater flexibility in assessing precision and goodness-of-fit.

For a great many settings where statistical methodology is currently used, the generalized linear model provides a convenient summary of the modeling process. It also fairly represents the kind of statistical thinking that is inherent in that process. After the modeling problem has been framed in this fashion, the measurement of uncertainty proceeds by selecting a link function, picking the x's to be included, choosing transformations of x's, estimating β, and evaluat-

ing precision and goodness-of-fit. Even though the definition is limited to a univariate response, fixed predictors, and a linear decomposition of effects (through use of the link function g), there is no doubt that it covers a wide range of applications and has been used to organize a substantial body of statistical theory. We hasten to add that this definition of a statistical model and the framework that has developed along with it have led to some of the great success stories in modern statistics.

We think that a more general definition of a statistical model is required to appreciate the unique features of modeling efforts geared toward the analysis of social data, survey data in particular. In the definition put forth next, the apparatus above will be recognized as only a part of the overall statistical model. That is, we define the statistical model as something quite different from the *prediction equation* to be estimated.

3. A Statistical Model Defined for the Social-Statistics Setting

It is tempting to think of a statistical model as the equation that links outputs to both fixed and random inputs. If we narrow the notion of a model this far, then the definition in the previous section is as good a starting place as any. But the main purpose in constructing a *statistical* model is to make explicit the sources of uncertainty so that the degree of uncertainty in our conclusions can be quantified. Our main point is that the conventional definition in Section 2 is not sufficient for doing this whenever survey data or other social statistics are analyzed. A much broader definition of the modeling process is necessary for social statistics. What follows is a list of parts of a statistical model with deliberate (and sometimes artificial) distinctions drawn among them. We distinguish six parts.

1. We first define a universe U^* to which inferences should pertain ideally. We try next to collect a sample S^* that would give unbiased estimators (or unbiased inferences) for the multivariate relationships in U^*. This means that we try to define first what a "representative sample" should look like. We can call S^* the "target sample"; see Kish (1987). This process can be called the *universe-sample problem.*

2. We define a set of measurements to be taken on S^*. This will normally consist of multiple measurements of *specified* response variables, say $\{Y_1, \ldots, Y_p\}$, where p is the number of such variables. And it will consist of multiple measurements of *specified* predictor variables (covariates), say $\{X_1, \ldots, X_k\}$. Judging from the complexity and size of most surveys that are widely used, both p and k can be quite large, which reflects vague prior knowledge concerning what should be measured or what can be measured well. It is not uncommon to have scores of X's and Y's. An additional distinction is that both the Y's and the X's are random. Even in cases where social experiments produce *assignments* of fixed x values (for one or a few x's), the nature of such assignments is usually different for human subjects compared to nonhuman subjects, so it is best to allow for some randomness in even the predictors that might have resulted from assignment of "treatments" (see Heckman and Hotz 1989). Specifying the X's and Y's and working out an operational strategy (e.g., types of questions and question wording) can be called the *measurement problem.* We realize that many social researchers would reserve this term for what is done after the X's and Y's are available, but we shall use a different term for that operation below.

3. We next collect the sample (or samples) of interest. Problems of nonresponse as well as incomplete lists used to develop

the sampling frame often imply that the sample observed or observable is quite different from the target sample S^*. Call the sample actually obtained S; call the actual "universe" to which S refers U. Because of nonresponse, missing data on some or many of the variables, attrition or mortality in panel studies, etc., S differs from S^*. Of course, samples are collected under budget constraints and other exigencies, but even without these burdens we usually find that we cannot obtain the true target sample (S^*) for a multitude of reasons relating to the fact that we are "surveying" human subjects. The target sample might have size N (judged optimal *a priori*) whereas the sample observed has size n, different from (and usually smaller than) N. Most important, the characteristics of those that wind up in S can be very different from the characteristics of those imagined for S^*. A statistic T calculated from S might differ substantially in expected value from a statistic T^* that might have been calculated from S^*. It follows that inferences based on T might be very different from inferences that would have been based on T^*. To complicate matters further, the ideal of a simple random sample or even a judiciously stratified sample can seldom be met, due to cost or efficiency considerations. The norm is truly complex sampling involving clustering as well as stratification, and even quota sampling (more often than we would like to admit). Clustering destroys the independence among sampled units, with perverse consequences that are seldom studied carefully by researchers outside the survey houses where the data is collected. This general problem we shall call the *sampling problem*. Good references include Cochran (1977), Kish (1987), and Groves (1989).

4. We next *organize* the data and *augment* it so that it can be analyzed in the right way. Modern tools for data base management are very important for organizing data so that *multiple levels* (individual, household, local-area) can be addressed simultaneously, or so that temporal features of the data collection (panel format, true event history) can be utilized in analysis. Weights might be assigned to "cases" in order to adjust for stratification features of the design, attrition in panel studies, or prior notions about the differences between S (the actual sample) and S^* (the target sample). Finally, special *subsamples*, involving selection of particular cases from S, might be created (and usually are created) for the analyses contemplated. This problem is truly complex, and perhaps it would be better to think of organization of data (database management tools), data augmentation (addition of weights, imputation of missing values, etc.), and subsample selection (including adjustments for the selection) as distinct problems. We shall refer to this general issue in an inclusive way, as the *data organization, augmentation, and selection* problem, just to keep our list as small as possible.

5. Now the data are available and the more usual aspects of statistical modeling come to play. Before we can apply standard "statistical models," however, we have to consider ways of combining or weighting multiple measurements of both Y's and X's (y's and x's in the sample). Indexes of various kinds are commonplace in most substantive research areas, but they are ubiquitous in the analysis of social statistics. Measurements might be combined or weighted by some type of factor analysis or latent structure analysis. Of course, in some of our most advanced technology, we think we can combine or "average" our measurements (indicators) and estimate the structural model of interest at the same time. The LISREL framework (Bollen 1989) is an example of this; also see Fuller (1987). How multiple measurements

of the "true" variable Y^* are actually obtained from the multiple measures (i.e., from $\{Y_1, \ldots, Y_p\}$) shall not concern us. Even decisions to discard some Y's (or some X's) are decisions that reflect the weighting of the multiple measurements (some Y's or X's receive zero weight in the composite index). So-called summated scales calculated from 0-1 items or Likert "scales" reflect equal weighting. Principal components reflect different weighting systems. Whether to reduce the several measurements to just one measure is a related concern that goes under the heading of "dimensionality of constructs," which can have many different meanings. The point is that there is uncertainty produced by all of the steps in the process that lead to combining multiple measurements. The wrong Y's or X's might have been discarded or they may have been combined using techniques based on faulty assumptions. But we have to combine measurements somehow because it is just too unwieldy to work with all of the available measurements simultaneously. We refer to this issue as the *problem of combining multiple measurements.* Usually social researchers think of latent variables (Y^*, say) in doing these operations, but the idea of a latent variable is not necessary to appreciate the problem, however important it might be to develop a rationale for the operations.

6. At this point we are ready to become "data analysts" or "modelers" in the usual senses of these terms. We now construct a model in the formal sense (equation linking systematic and random inputs to outputs), for the subsamples selected, on the set of measurements or derived indexes picked. We suppose that this model applies to the universe U^* (or is it U?) for which the original sampling frame (i.e., S^*, but is it really only suited for S?) was devised. This model-as-equation might take the form of a generalized linear model, and usually belongs to this class in fact. At this stage of the modeling process, we usually employ likelihood methods of some kind to pick covariates (a subset of them anyway), assess goodness-of-fit, estimate the β, test hypotheses (e.g., about subsets of β), and calculate interval estimates. These methods are often sensitive to the data actually provided as an input to the estimation. We cannot have data spread too thinly across the grid of variable values possible or colinearity problems that are too severe. The apparatus summarized in the previous section comes to play at this stage; indeed, most researchers think of statistical modeling as the activity involved with this component of the analysis. We hesitate to call this set of operations "statistical modeling" because there are so many other important statistical decisions, and so many other sources of uncertainty, that have preceded this stage. For want of a better term, we call this the *estimation* problem. The point is that at this stage we are mainly seeking "good" estimates as well as the means to say how good those estimates are.

The seasoned social researcher obviously has many more things in mind than whether the "final" equation estimated represents state-of-the-art econometrics. Social researchers usually will not worry so much whether the particular estimation procedure is 99% optimal according to the criteria provided by mathematical statistics, which more often than not ignore all sources of uncertainty created by the first five phases of the modeling process. Unfortunately, optimality is usually defined in terms of "pure sampling error" for the *model equation* that is estimated, which takes the N measurements of x's and y's at face value. The methodology of modeling summarized in Section 2 deals with this sort of uncertainty very well. Our main point is that we must seriously con-

sider the uncertainty in inferences created along the way by the data collection processes so common in social research. We next examine illustrations, examples, and implications of this broader definition of a statistical model.

4. Illustrations and Implications

Our research ultimately produces estimates from one or more equations which are summarized with estimates of β and standard errors. Our claim is that $s(\hat{\beta})$ does not represent the uncertainty in our estimate of β in a proper way. Next we try to see why this is the case with illustrations drawn from an ongoing project in which we are involved (Clogg, Hogan, and Lichter 1989). The primary goal in this project is to link labor force behavior and poverty status in the careers of young adults, using information from the former to predict and explain variation in the latter. The main source of data is the National Longitudinal Study of Labor Market Experiences, Youth Survey (NLSY; see Center for Human Resources Research 1987). This sample is presumably a nationally representative sample of men and women aged 14–21 in 1979. Approximately 12 000 young people have been followed for a decade, giving both annual measurements and information on transitions among states of various kinds.

1. The universe-sample problem most definitely makes inferences more uncertain than they first appear. Perhaps both U^* and S^* have been incorrectly defined, and perhaps it is impossible to specify either in a valid way. For example, to study the development of careers in the initial stages of the life course, the universe might be thought of in terms of post-school activities of youth (all youth out of school) or it might be thought of as a combination of in-school and post-school activities, recognizing flow to and from these two states in the early career. The target sample would differ between the two cases.

Illustration. While the universe in our study of NLSY data is ostensibly the U.S. population aged 14–21 in 1979, have the military experiences of those youth been accounted for adequately? Should the universe have contained those who already dropped out of school? How can we assess the "representativeness" of the sample obtained? One standard method of answering the latter question is to compare sample results to census tabulations (for the 1980 decennial census). But the census undercounts nonwhite youths, youths in general, and poor youths, and the sample frames were derived without the benefit of the 1980 census anyway (see Clogg, Massagli, and Eliason 1989). What makes certain groups hard to count in a census also makes them hard to catch in a survey; most sample surveys are in fact samples of the census-enumerable population.

2. The measurement problem is always with us. Much of the effort in questionnaire construction (Schuman and Presser 1981) and validation of items (see Suchman and Jordan 1990) is directed to this general issue. Perhaps the wrong measurements have been taken or they have been defined in the wrong way. Perhaps self-reports are invalid, but proxy reports are also questionable. Most of the econometric techniques purporting to resolve problems of left-out variables and specification error apply to this problem in some sense, but of course the techniques associated with these topics are attempts to deal with the problem of having the wrong measurements.

Illustration. The types of variables of possible interest in a study of labor force careers and poverty or earnings are legion. There are nearly 1,000 variables in NLSY, most of which pertain to these things. How often were they measured or how

often should they have been measured? To measure poverty, several indexes have been proposed; which should be used? To measure labor force behavior, nearly one hundred existing variables might be used. Measures that are faithful to the panel format can be constructed, but most of these would not be consistent with official labor force measures calculated for cross-sectional surveys. Does that present a problem? There is no random assignment of fixed "treatment" levels at all; all variables are responses in some sense. How should the variables be treated across the nine waves of the survey? Should change measures, time-specific period measures, event-history measures, or other formats be used? How should earnings be adjusted for hours or weeks worked at possibly different jobs?

3. The sampling problem also has to be reckoned with. Often such issues are relegated to the footnotes of social research, but they are probably just as important as controlling sampling error at a later stage of the analysis. How much does the sample observed differ from the true target sample (S compared with S^*)? Have dependencies among sampled units arisen that were not taken into account? Assuming that the sample is truly a complex one (involving clustering), will it be possible to utilize the information in the sample to adjust for these facts? How well has nonresponse been managed? Were the survey instruments properly understood by the respondent?

Illustration. NLSY oversamples economically disadvantaged groups. (There are 946 Hispanic males, 978 Hispanic females, 1,451 black males, 1,472 black females, 945 poor white males, and 1,099 poor white females in the sample.) Do the weights in the file adequately reflect the oversampling? What is the universe to which the weights apply? Most variables have complicated missing data patterns. "Skip patterns" that all users of panel data have to worry about guarantee that most variables will be available only for selected subsamples for some of the periods. There is the usual level of nonresponse for many key indicators (i.e., between 10% and 20% missing values). Attrition from the sample is a major problem, and the weights are supposed to adjust for this to some extent. The response rate at the initial wave was 87% (what were the characteristics of the 13% who did not respond?), which is normally thought of as a very good rate. A full 91% of the initial sample was still in the panel by 1986, but this means that less than 80% (.91 × .87) of the original "target" sample is still present. Effects of interest are likely to be small ones, and nonresponse compounded by attrition could be a major source of bias. (We hasten to add that NLSY is usually regarded as a superior data set in terms of nonresponse and the like.) Finally, NLSY is definitely not a simple random sample or even a simple stratified random sample.

4. The data organization, augmentation, and selection problem is, after all, three separate problems. Errors creep in when primary sampling units are reorganized for analysis, particularly in panel studies where "skip patterns" make it difficult to use rectangular data files that might be made so that existing software can be used. Weights defined appropriately for cases might not be appropriate once the data is reorganized; perhaps an augmented system of weights should be used. Imputed values for missing data are randomly assigned to some extent. How do we take account of this kind of uncertainty? Sample selection is another source of uncertainty. Should we select white males in central cities, or nonblack males in metropolitan areas between ages 25 and 44 who have completed high school?

Illustration. NLSY data can be organized in different ways; a rectangular file is avail-

able. Can we assume that the rectangular file is accurate or that a reorganization based on it to emphasize event histories would be accurate? Should we do "weighted analyses" available with standard software packages (what do these packages actually do?), or should we enter the weights as covariates? Should we reweight once we select subsamples of most interest? In the interests of simplicity, should we just focus on the cases not lost to followup? (We have found that all cases lost to followup have been given zero weights.) To examine earnings functions, we have to make an allowance for the fact that only those who work get earnings. Do we want to do this through selecting special subsamples, by using missing data adjustments (Little and Rubin 1987), or by using formal selection adjustments of the econometric variety (Heckman 1976)? Should Hispanics and blacks be kept separate?

5. We must come to grips with the problem of combining multiple measurements. This creates special problems because we will usually have discarded lots of variables that have in fact been measured. We usually rely on some combination of prior experience, past research on the same topic that tells us to pick certain X's and Y's and not others, and even preliminary or exploratory analysis of sets of indicators available. Assuming that this narrows the range considerably, how do we then combine the measurements that we have? This might be done by using some type of factor analysis, principal components analysis, or latent structure technique (see Bollen 1989; Jöreskog and Sörbom 1979; Lazarsfeld and Henry 1968; Langeheine and Rost 1988). "Scales" and "indexes" abound in social research, many of which have no meaning apart from the specific indicators chosen. Summing the responses on a set of dichotomous items is often thought to be the simplest thing, but this procedure is more subtle than it first appears as the large literature on scaling test items indicates (see Clogg 1987). Producing "factor scores" or "latent class assignments" and reducing dimensionality through clustering or components analysis are standard tools in social research. Choosing the variables to combine is one problem; choosing the *method* to combine them is another. Assumptions used with such methods, such as "local independence," "random measurement error," and "simple structure" call for urgent examination (see Becker and Clogg 1988). Most methods for combining multiple measurements assume random measurement error, which in a linear model with continuous measurements means that the size of the measurement error is independent of the size of the true score that we could not measure directly. How realistic are these assumptions?

Illustration. In the analysis of NLSY data, we have boiled this down to four separate questions. First, should the three main poverty measures be used separately or in combination? If used jointly, how should the essentially categorical measures be combined? Latent class methods of some kind seem natural, but there are alternatives that are easier to implement. Second, the categories of labor force behavior, measured over a decade of experience, could run into the hundreds. Which "types" should be selected? Can an *index* or two be constructed? Association models (Goodman 1984) might be useful, but principal components methods, clustering methods, and even some kinds of factor analysis ought to be tried. Third, the covariates related to the response (poverty status) and "correlated" with the main predictor of interest (labor force behavior) must somehow be controlled. But there are scores of these, including both categorical and continuous measurements. Picking just a few of

these would be consistent with prior research but we ought to find out how much information is lost if information from many other measures is discarded. Fourth, because of the panel format (nine waves will be utilized), all of the measurements change over time, which creates even more complexity in combining the multiple measurements.

6. Finally we arrive at the problem of summarizing uncertainty, conventionally at least, in the estimation phase of the modeling process. We have to make inferences from the equations to be estimated, the so-called structural or behavior model. Proper selection of covariates, simultaneous inference problems, colinearities, the validity of assumptions (only some of which can be tested empirically), and goodness-of-fit are just some of the issues involved. The validity of sampling error properties of the estimators is another key component of uncertainty. Almost all of our statistical thinking leads to arguments based on sampling from infinite populations under iid (independent and identically distributed) assumptions. At the same time, almost all of our surveys are justified by *randomization inferences* (see Cochran 1977; Little and Rubin 1987, ch. 4) which rely on finite populations and known probabilities of selecting cases from that population for the sample. Usually estimators of β are assumed to follow Gaussian distributions, which implies that the log-likelihood for the model has a quadratic (parabolic) shape at the maximum. What is the true shape of the log-likelihood? Is the $\hat{\beta} \pm 2[s(\hat{\beta})]$ formula for interval estimates valid? Of course, modifications of likelihood methods such as conditional, partial, or quasi-likelihood can be used where necessary, but the primary justification for these latecomers is that they have almost the same logic (and almost the same large-sample rationale) as maximum likelihood. By the time we start estimating "the model" and picking from the available software (or preparing new software), this problem looks easy in comparison to all of the other problems that have preceded it.

Illustration. For categorical responses (e.g., poverty status), models of the log-linear or logit variety seem appropriate. For continuous responses (e.g., earnings), linear models are natural. We could analyze the data in terms of event history (or hazards) formulations (but the event histories are really fairly discrete), or we could take the "history" available from several waves as a covariate vector that helps predict poverty or earnings in the final waves, when sample members are in their late 20s. In fact, the response variables are mixtures of categorical and continuous (or almost continuous) variables: do we "scale" the categorical variables or group the continuous ones? After these issues are solved, we still have to think about the proper way to use weights, what subsamples to use for estimation, and rules of thumb for adjusting the variance-covariance matrix for complex sampling. Will the sample sizes permit use of conventional large-sample approximations?

Our goal in belaboring these points is to give a richer appreciation of the *statistical* decision making associated with the entire process of modeling social statistics. Uncertainty crops up all along the way. All six sources of uncertainty given above must play a role in the inferences, predictions, or hypothesis tests that are ostensibly based on the operations applied to the model equations estimated at the last stage. We should ask whether the estimates of β and $s(\hat{\beta})$, which almost all studies (ours included) produce as an end product, measure the relationships of interest or describe the uncertainty about those relationships in a manner that reflects the modeling process involved.

Although there is bound to be disagreement over what factors in our list should be emphasized the most, we are convinced that social statisticians would be in essential agreement with our claim that the modeling process is at least this complex. The simple fact is that uncertainty has to be understood in a far broader sense when social data are analyzed. And uncertainty attributable to all these sources ought to be taken into account when our conclusions from data analysis are put forth. We next try to give some reasons why this complexity has been overlooked in building the impressive modeling apparatus of modern econometrics or statistics.

5. Modeling Contexts Compared

It seems apparent that there are three main scientific contexts in which statistical modeling tools have developed. Each context defines a standard frame of reference through which sources of uncertainty can be enumerated, and in each case the context involved has had major effects of procedures for measuring uncertainty and for partitioning uncertainty into sources.

The first context is the analysis of *physical or biological* systems. Legendre and later Gauss concerned themselves most with physical measurements, such as astronomical measurements, where measurement error was the sole source of randomness in the systems analyzed (Stigler 1986). In the analysis of physical systems, it appears fairly typical to have only a few measurements, and some physical or biological theory tells the analyst how the measurements would be related if there were no measurement error (often synonymous with "experimental error"). Good science normally implies the necessity of just a few sample units, and sciences in the physical or biological areas are surely good ones. Michelson's experiments a century ago, which gave conclusive evidence that light did not travel through "ether," were based on just a few measurements in carefully contrived circumstances. (These experiments had much to do with Einstein's later formulation of relativity theory; see Clark 1971, ch. 3.) Variations of regression were thought to produce the *functional* properties of the system, or in modern terminology, functional regression relationships. The modern attention to "structural representations" or structural models comes from a different context than this one.

The second context is that of *experimental sciences* (not, of course, totally separate from the context just considered), which led to the elevation of *experimental data* as the ideal basis for casual inference. Fisher and many statisticians who followed him were concerned mostly with experimental data and rules for controlling bias and maximizing precision (reducing "sampling error" or experimental error). In the classical experimental model developed by Fisher and Yates (see Fisher 1935), nonhuman subjects or nonsocial (i.e., biological or psychophysical) aspects of human behavior were the main focus of attention. The random assignment of a few treatments to randomly chosen subjects, using balanced designs developed in the methodology of analysis-of-variance, largely avoids the problem of dealing with a multitude of variables simultaneously. Randomization allows consideration of only a few key predictors (treatments) at a time but nevertheless permits valid casual inferences even when other variables that are obviously important are unavailable. This context has led to the emphasis on analysis of variance and experimental design, two of the mainstays in any modern curriculum in statistics. Much of present-day statistical methodology is built on this model. This is indicated, for example,

by the habit of calling Y's *response variables* and X's *factors*.

The third context that we also think provides another standard for evaluating our methodology is the one created by *economic time series* data. Units of observation are often obvious in this case; so is the "universe" (even though it is an hypothetical one) and the "sample" (which modelers take for granted). The goal of time series analysis, until recently anyway, was to make predictions, not explanations of the process, which narrows the objectives considerably. This classical context for the development of statistical reasoning is the closest to that of social statistics, and it is not an accident that modern econometrics, which began with the analysis of economic time series, provides the impetus for so much that is now used in modeling social statistics. Of course, in the analysis of economic time series a great deal of attention goes toward "whitening" the error structure (removing correlations among units by conditioning on x's or past history).

Models for social statistics, including survey data, are different from models for physical, experimental, or time series data in a great many respects. The most important difference has to do with the data collection process, but of course the theory behind most efforts to collect social data is vague at this point in time. The lack of precise theory in social research stands in sharp contrast to the situation in the other areas, with the possible exception of economics. Cross-sectional surveys, repeated cross-sectional surveys, panel surveys, event-history surveys, network or "interaction" surveys, and censuses of various kinds (and samples from them) provide the data base of modern social statistics. Indeed, the term "social statistics" is practically synonymous with survey or census data of all kinds. The first four components of a statistical model enumerated in the previous two sections all refer to data collection or sampling, and their status can be appreciated immediately, and concretely, in terms of any given sample survey. The fifth component – combining multiple measurements – is a special problem because surveys typically collect hundreds of variables because it is not known exactly which ones should be collected or which ones could be measured well. This fact is the main reason why statistical methods and models for combining multiple measurements, such as factor analysis, clustering methods, and latent structure analysis, have received so much attention in social statistics. (It is important to add that these methods have now become a part of modern statistical methodology; see Bollen (1989) or Dillon and Goldstein (1984).)

The *data structures* in social statistics are also quite different from those in experimental, physical, or time series settings. The system under study, if it can be called by that name, is not known in advance, and measurement error cannot be regarded as the sole, or even the primary, source of variability. Regardless of how much is measured, most researchers now appreciate the fact that there is still quite a lot of unexplained inter-individual heterogeneity that it is hopeless to model in terms of a hypothetically closed system subject only to "experimental error." This stands in sharp contrast to the analysis of physical measurements. Experimentation of the classical variety is usually impossible, inconceivable, or difficult to implement. Experimentation is also costly. The term *observational studies* (Cochran 1977; Kish 1987) has become a catch-all phrase for survey or census data used in social statistics, and this designation certainly reinforces the view that surveys are not experiments. But it also calls attention to the fact that relationships out in the world rather than in the laboratory are being analyzed. The goal in

analyzing social statistics in most cases is to explain how a system "actually out there" works. Predicting future values of system outputs is not the ultimate objective, at least in most areas. (Some demographic modeling would, of course, take prediction or projection or forecasting as the main goal; see Keyfitz 1985.) It would be fairly easy to continue listing areas where the social statistics setting does not overlap with that of the classical settings. In fact, the only real place where the several contexts overlap is that they each lead to the production, eventually, of a data set with N measurements on some X's and some Y's. Where the N cases and the X's and Y's come from and what they actually signify are the crucial contrasts. But everything else prior to this differentiates the contexts from each other, implying that the conclusions reached from the equation estimates should be modified somehow to reflect the differing contexts.

The main point is that the estimation phase of the modeling effort should not ignore the uncertainties that abound in all phases of the modeling process. The uncertainties associated with the first five phases of the modeling process probably produce at least as much uncertainty (unknown bias, unknown precision) as "sampling error" as conceived under ideal conditions for the equation(s) actually estimated. Perhaps with physical data, experimental data, or time series data it is adequate to concentrate on the uncertainty associated with estimation. (We believe, however, that even with economic time series many of the same problems are present.) The complexity in the process by which social data (survey data) are produced calls for going substantially beyond this.

6. Supermodels

A distinction should be made between the conventional statistical model (Section 2) and the truly complex models so popular currently in the analysis of social data. The term *supermodel* shall be used for the latter. A supermodel is a framework, including a set of equations to be estimated, that tries to deal with the complexity of social data in an integrated fashion. A supermodel will normally consist of a "structural" or behavioral model imbedded in an estimation scheme that takes account of uncertainties associated with one or more of the first five phases of the modeling process. A supermodel tries to represent the complexity of causal relations along with the likely consequences of sample selection, left out variables, possible misspecification of functional form, etc. A supermodel tries to overcome deficiencies in sampling or measurement by including special parameters that are often regarded as covariates which ostensibly adjust for other uncertainties besides sampling error. We hasten to add that there does not appear to be much concern with some of the most important sources of uncertainty in our list of Section 3. For example, our most serious supermodels (or the software for implementing them) do not deal well with complex samples, with uncertainties associated with a mismatch between the observed sample and the target sample, and so on.

There are many modeling frameworks that qualify for designation as supermodels. We illustrate with three examples.

The Contingency Table Model. Traditionally, a contingency table has been thought of as a cross-classification of categorical variables where the distinction between responses and factors plays no special role. Statistical models for cell frequencies in contingency tables, usually but not always of the log-linear variety, are prominent in many areas of social statistics. Multinomial-response models (e.g., logit models) arise from this class of models by conditioning

on the variables presumed to operate as covariates or factors in the classical sense. This methodological tradition has been developed to a great extent in the social-statistics setting, by methodologists with close ties to the social sciences (see Goodman 1978, 1984; Haberman 1978, 1979; Bishop, Fienberg, and Holland 1975; Agresti 1990). Contingency table models are attractive for social data because of two factors: (a) most variables collected through survey formats are categorical in nature, and (b) effects of variables or interactions are modeled non-parametrically (i.e., no Gaussian assumptions are invoked).

Some of the uncertainties that arise before the estimation phase can be taken into account in contingency table models. To take just one example, sample cases not selected can be included as additional categories in the table actually modeled, and this strategy has been pursued by a number of researchers. Procedures for adjusting inferences for complex sampling have been developed (Rao and Thomas 1988; also see Clogg and Eliason 1987). The main obstacle preventing more widespread use of such models is the fact that they are difficult to implement in highly multivariate situations, say cases involving scores of variables where the number of cells in the table would be very large. We do not as yet have good methods for coping with *sparse data* created by creating large contingency tables. Likelihood methods fail us here because MLE's may not even exist and goodness-of-fit tests will often be invalid. Perhaps some Bayesian strategies can be used to overcome these limitations (Clogg et al. 1990), but even here the computational burden can be excessive. In addition, much more needs to be done to adapt these models for the analysis of change, for example, in panel studies. To date, there are relatively few serious studies in the social sciences where the contingency table model has been employed to study change involving a realistic number of covariates (see Clogg, Eliason, and Grego 1990). And it is difficult to simultaneously estimate a realistic model combining multiple measurements of categorical variables (latent structure or latent class ideas) and a structural model depicting relations among the latent variables.

The Covariance Structure Model. The general covariance structure model pioneered by Karl Jöreskog and his co-workers (Jöreskog and Sörbom 1979) definitely belongs in the category of supermodels. Bollen (1989) codifies this methodology very well. The basic idea is that a *measurement model* and a *structural model* are combined into one supermodel. This framework thus addresses the problem of combining multiple measurements with the problem of estimation all in one operation. Multiple measurements of either covariates or responses or both are combined implicitly into indexes ("latent variables"); causal relations are then represented in a set of linear equations connecting the indexes together. There can be no doubt that advances in this area have been important. Current software allows simultaneous analysis of many variables. Multiple-group methods add flexibility. Recent extensions (see Bollen 1989) for categorical variables allow at least some flexibility in dealing with sample selection or truncation issues. The general covariance structure model qualifies as a supermodel, and that is why it has received such sustained attention in social statistics, as well as in other areas. It can be noted that this methodology has now made its way into mainstream statistics (see Dillon and Goldstein 1984).

The chief difficulty with the covariance structure model is that it deals with only the first two moments (or cross-moments) of variables that comprise the data base. Tests

of covariance structure models examine the fit of model to variances and covariances observed, not to the real observations or the real "data". How realistic is the assumption that the data can be reduced to means, variances, and covariances? Standard techniques in this area rely strongly on parametric assumptions; typically, normality is assumed for everything in the system, both observables and unobservables. To combine multiple measurements, an assumption of *random measurement error* is often used. How sensitive are inferences to this assumption? Uncertainty in inferences should take account of uncertainty about the validity of this assumption, but seldom do we see this done. The regression relationships ultimately estimated should also be examined for omitted variable bias (or its logical complement, included variable bias) or for other types of misspecification. Uncertainty in causal inferences due to misspecification is seldom carried out beyond a ritualistic look at model-modification indexes of various kinds. This framework does not appear to be generally appropriate for longitudinal analyses, nor does it pretend to deal automatically with other uncertainties besides that associated with the last two parts of the modeling process.

The Event-History Model. To study the hazard rates governing the length of time until a single nonrepeatable event occurs, failure time models (parametric) and semiparametric hazards models (Cox 1972) have proved extremely useful. In recent years, this model has been utilized to a great extent in the analysis of social and especially demographic data. The generalization of this model for multiple events (e.g., types of labor force states) that are repeatable leads to consideration of the general event-history model. A major statement of this supermodel appears in Tuma and Hannan (1984). Probably the most complex and the most general outgrowth of this modeling framework is due to Heckman (e.g., Heckman and Walker 1987). This supermodel allows for elaborate sample selection (or left-censoring) adjustments. Misspecification by omitting key variables is dealt with with both parametric and nonparametric adjustments for unobserved heterogeneity, including latent classes of the mixing distribution. A variety of special forms of *time dependence* in the hazard functions can be considered using a kind of Box-Cox transformation. The list continues. Of all the supermodels with which we are familiar, the Heckman framework for event-history analysis qualifies the best as a supermodel. Of course, the approach constitutes a supermodel for only a certain class of problems, namely, discrete variables regarded as responses with time of the events recorded continuously. One of the most important aspects of these models is the manner in which adjustments for *right-censoring* can be made. For example, in analyzing the duration of first marriages, it is not necessary to follow a cohort of individuals until all of them have experienced failure. But these models assume *random* right censoring, an assumption that is as difficult to test as the assumption of random measurement error, and probably just as suspect as a factor adding uncertainty to our inferences.

We think the effort to build more realistic supermodels has been worthwhile. But there is a danger of becoming so enamored with the complexity of the output that we ignore the uncertainty in the inputs. Our best supermodels tell us little about the probable effects of statistical decisions that have preceded the estimation phase of the modeling effort. The "automatic" adjustments for measurement error, for sample selection, or for other factors that are built into these methodologies require further examination. In many cases, we

think it is difficult to test key assumptions or even to know what assumptions really make the model behave as it does. We need to work much harder, with the benefit of the computer, to understand how our results are sensitive to assumptions buried in the black box.

7. Concluding Remarks

The idea of statistics is not to maximize functions but rather to measure uncertainty. A truly adequate statistical model should organize the way that we summarize uncertainty and give us a means to partition it into sources and ultimately to quantify it. It is not enough, at least when analyzing social data, to "model" uncertainty using only the last-stage inputs to our latest maximum likelihood routine for the most complex model equation that our machines and software can estimate. Our purpose has been to encourage a broader view of the inputs that are oftentimes taken for granted and to note that each particular input involves statistical decision making and therefore ought to be assessed statistically *when the final outputs of our prediction equations are assessed.*

Part of the difficulty is that parts of the modeling process enumerated here have been separated from each other and have thus been dealt with by specialists having a division of labor that prevents a holistic view of the process. Survey houses (private and governmental) are primarily responsible for assessing uncertainties in the universe-sample problem. Questionnaire construction ("instrumentation") is a separate specialty. Sampling experts are called in to recommend efficient collection designs and to then build in case weights and other information necessary to exploit the data appropriately. The survey industry has made great progress in assessing "total survey error" (Groves 1989), but this does not seem to be reflected very much in the model outputs that we consume. Data base management, imputation, and related problems are also taken up separately. Latent structure analysis as well as related techniques used for combining multiple measurements constitute a major industry in itself. Finally, the consideration of formal model equations and estimation procedures for them is the domain of statistics, biostatistics, psychometrics and econometrics, as well as parts of other areas. All of these separate activities are commendable. Indeed, progress in each area has been substantial and will likely continue. At the same time, we cannot ignore or downplay the importance of all of the other sources of uncertainty that give us the set of measurements we estimate at the last stage. Perhaps the same comments apply to the modeling enterprise in other areas such as in the analysis of clinical trials in biomedical areas, but that is another story.

Although attempts to build more comprehensive supermodels will no doubt continue, it is possible that other strategies ought to be pursued in place of or in addition to these efforts. Sensitivity analysis ought to be taken seriously (Leamer 1978). We are not aware of a single empirical paper on a *substantive* topic where this has been done, perhaps because journal editors and referees alike want simple answers. Perhaps multiple-imputation tools (Rubin 1987) or some other sample reuse methods which use the computer more and analytical derivations less ought to be tried. We have not yet realized the potential of modern computing in our modeling efforts beyond the point of having the capacity to maximize estimating functions for even more complicated prediction equations. It is possible that some Bayesian strategy for the overall process ought to be attempted. (Let $U_1, U_2, \ldots, U_6$ denote uncertainty associated with stage i in

the modeling process; use empirical or other distributions to model each U_i; integrate over all sources to arrive at an overall measure for uncertainty of β; etc.) Whatever the case, we hope that social statisticians will accept the challenge to produce a more realistic appraisal of uncertainty in conclusions reached from the analysis of social data.

8. References

Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley.

Becker, M.P. and Clogg, C.C. (1988). A Note on Approximating Correlations from Odds Ratios. Sociological Methods and Research, 16, 407–424.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

Bollen, K.A. (1989). Structural Equations With Latent Variables. New York: John Wiley.

Center for Human Resources Research (1987). NLS Handbook 1987. Columbus, Ohio: Ohio State University.

Clark, R.W. (1971). Einstein: The Life and Times. New York: World Publishers.

Clogg, C.C. (1987). Latent Class Models for Measuring. In R. Langeheine and J. Rost, eds., Latent Trait and Latent Class Models. New York: Plenum, 173–205.

Clogg, C.C. and Eliason, S.R. (1987). Some Common Problems in Log-Linear Analysis. Sociological Methods and Research, 16, 8–44.

Clogg, C.C., Eliason, S.R., and Grego, J.M. (1990). Models for the Analysis of Change in Discrete Variables. In A. von Eye, ed., New Statistical Methods in Developmental Research, vol. 2. New York: Academic Press.

Clogg, C.C., Hogan, D.P., and Lichter, D.L. (1989). Underemployment and the Persistence of Poverty Among Young Adults. Proposal to the Russell Sage Foundation.

Clogg, C.C., Massagli, M.P., and Eliason, S.R. (1989). Population Undercount and Social Science Research. Social Indicators Research, 21, 559–598.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1990). Multiple Imputation of Industry and Occupation Codes in Census Public Use Samples Using Bayesian Logistic Regression. Journal of the American Statistical Association, (forthcoming).

Cochran, W.G. (1977). Sampling Techniques, 3rd edition. New York: John Wiley.

Cox, D.R. (1972). Regression Models and Life Tables. Journal of the Royal Statistical Society, ser. B, 74, 187–220.

Dillon, W.R. and Goldstein, M. (1984). Multivariate Analysis: Methods and Applications. New York: John Wiley.

Duncan, O.D. (1984). Notes on Social Measurement, Historical and Critical. New York: Russell Sage Foundation.

Fisher, R.A. (1935). The Design of Experiments. Edinburgh: Oliver and Boyd.

Fuller, W.A. (1987). Measurement Error Models. New York: John Wiley.

Goodman, L.A. (1978). Analyzing Qualitative/Categorical Data. Cambridge, Mass.: Abt Books.

Goodman, L.A. (1984). The Analysis of Cross-Classifications Having Ordered Categories. Cambridge, Mass.: Harvard University Press.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley.

Haberman, S.J. (1978). Analysis of Qualitative Data. vol. I. Introductory Topics. New York: Academic Press.

Haberman, S.J. (1979). Analysis of Qualitative Data. vol. II. New Developments. New York: Academic Press.

Heckman, J.J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for Such Models. Annals of Economic and Social Measurement, 5, 475–492.

Heckman, J.J. and Hotz, V.J. (1989). Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training Programs. Journal of the American Statistical Association, 84, 862–874.

Heckman, J.J. and Walker, J.R. (1987). Using Goodness-of-Fit and Other Criteria to Choose Among Competing Duration Models: A Case Study of Hutterite Data. In C.C. Clogg, ed., Sociological Methodology 1987. Washington, D.C.: The American Sociological Association, pp. 247–308.

Jöreskog, K.G. and Sörbom, D. (1979). Advances in Factor Analysis and Structural Equation Models. Cambridge, Mass.: Abt Books.

Keyfitz, N. (1985). Applied Mathematical Demography, 2nd ed. New York: Springer-Verlag.

Kish, L. (1987). Statistical Design for Research. New York: John Wiley.

Langeheine, R. and Rost, J., eds. (1988): Latent Trait and Latent Class Models. New York: Plenum.

Lazarsfeld, P.F. and Henry, N.W. (1968). Latent Structure Analysis. Boston: Houghton Mifflin.

Leamer, E.E. (1978). Specification Searches. New York: John Wiley.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd ed. London: Chapman and Hall.

Rao, J.N.K. and Thomas, D.R. (1988). The Analysis of Cross-Classified Categorical Data from Complex Surveys. In C.C. Clogg, ed., Sociological Methodology 1988. Washington, D.C.: The American Sociological Association, 213–270.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Schuman, H. and Presser, S. (1981). Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Content. New York: Academic Press.

Stigler, S.M. (1986). The History of Statistics: The Measurement of Uncertainty Before 1900. Cambridge, Mass.: Harvard University Press.

Suchman, L. and Jordan, B. (1990). Interactional Troubles in Face-to-Face Survey Interviews (with discussion). Journal of the American Statistical Association, 85, 232–253.

Thisted, R.A. (1988). Elements of Statistical Computing. New York: Chapman and Hall.

Tuma, N.B. and Hannan, M. (1984). Social Dynamics: Models and Methods. Orlando, Florida: Academic Press.

Received October 1989
Revised July 1990