PARTIAL IDENTIFICATION OF PROBABILITY DISTRIBUTIONS

Charles F. Manski

Springer-Verlag, 2003

Contents

Introduction: Partial Identification and Credible Inference

Statistical inference uses sample data to draw conclusions about a population of interest. However, data alone do not suffice. Inference always requires assumptions about the population and the sampling process. Statistical theory illuminates the logic of inference by showing how data and assumptions combine to yield conclusions.

Empirical researchers should be concerned with both the logic and the credibility of their inferences. Credibility is a subjective matter, yet I take there to be wide agreement on a principle that I shall call:

The Law of Decreasing Credibility: The credibility of inference decreases with the strength of the assumptions maintained.

This principle implies that empirical researchers face a dilemma as they decide what assumptions to maintain: Stronger assumptions yield inferences that may be more powerful but less credible. Statistical theory cannot resolve the dilemma but can clarify its nature.

It is useful to distinguish combinations of data and assumptions that point-identify a population parameter of interest from ones that place the parameter within a set-valued identification region. Point identification is the fundamental necessary condition for consistent point estimation of a parameter. Strengthening an assumption that achieves point identification may increase the attainable precision of estimates of the parameter. Statistical theory has had much to say about this matter. The classical theory of local asymptotic efficiency characterizes, through the Fisher information matrix, how attainable precision increases as more is assumed known about a population distribution. Nonparametric regression analysis shows how the attainable rate of convergence of estimates increases as more is assumed about the shape of the regression. These and other achievements provide

important guidance to empirical researchers as they weigh the credibility and precision of alternative point estimates.

Statistical theory has had much less to say about inference on population parameters that are not point-identified (see the historical note at the end of this Introduction). It has been commonplace to think of identification as a binary event—a parameter is either identified or it is not—and to view point identification as a precondition for meaningful inference. Yet there is enormous scope for fruitful inference using data and assumptions that partially identify population parameters. This book explains why and shows how.

Origin and Organization of the Book

The book has its roots in my research on nonparametric regression analysis with missing outcome data, initiated in the late 1980s. Empirical researchers estimating regressions commonly assume that missingness is random, in the sense that the observability of an outcome is statistically independent of its value. Yet this and other point-identifying assumptions have regularly been criticized as implausible. So I set out to determine what random sampling with partial observability of outcomes reveals about mean and quantile regressions if nothing is known about the missingness process or if assumptions weak enough to be widely credible are imposed. The findings were sharp bounds whose forms vary with the regression of interest and with the maintained assumptions. These bounds can readily be estimated using standard methods of nonparametric regression analysis.

Study of regression with missing outcome data stimulated investigation of more general incomplete data problems. Some sample realizations may have unobserved outcomes, some may have unobserved covariates, and others may be entirely missing. Sometimes interval data on outcomes or covariates are available, rather than point measurements. Random sampling with incomplete observation of outcomes and covariates generically yields partial identification of regressions. The challenge is to describe and estimate the identification regions produced by incomplete-data processes when alternative assumptions are maintained.

Study of regression with missing outcome data also naturally led to examination of inference on treatment response. Analysis of treatment response must contend with the fundamental problem that counterfactual outcomes are not observable; hence my findings on partial identification of regressions with missing outcome data were directly applicable. Yet analysis of treatment response poses much more than a generic missing-data problem. One reason is that observations of realized outcomes, when combined with suitable assumptions, can provide information about counterfactual ones. Another is that practical problems of treatment choice

motivate much research on treatment response and thereby determine what population parameters are of interest. So I found it productive to examine inference on treatment response as a subject in its own right.

Another subject of study has been inference on the components of finite probability mixtures. The mathematical problem of decomposition of finite mixtures arises in many substantively distinct settings, including contaminated sampling, ecological inference, and regression with missing covariate data. Findings on partial identification of mixtures have application to all of these subjects and more.

This book presents the main elements of my research on partial identification of probability distributions. Chapters 1 through 3 form a unit on prediction with missing outcome or covariate data. Chapters 4 and 5 form a unit on decomposition of finite mixtures. Chapter 6 is a stand-alone analysis of response-based sampling. Chapters 7 through 10 form a unit on the analysis of treatment response.

Whatever the particular subject under study, the presentation follows a common path. I first specify the sampling process generating the available data and ask what may be inferred about population parameters of interest in the absence of assumptions restricting the population distribution. I then ask how the (typically) set-valued identification regions for these parameters shrink if certain assumptions are imposed. There are, of course, innumerable assumptions that could be entertained. I mainly study statistical independence and monotonicity assumptions.

The approach to inference that runs throughout the book is deliberately conservative and thoroughly nonparametric. The traditional way to cope with sampling processes that partially identify population parameters has been to combine the available data with assumptions strong enough to yield point identification. Such assumptions often are not well motivated, and empirical researchers often debate their validity. Conservative nonparametric analysis enables researchers to learn from the available data without imposing untenable assumptions. It enables establishment of a domain of consensus among researchers who may hold disparate beliefs about what assumptions are appropriate. It also makes plain the limitations of the available data. When credible identification regions turn out to be uncomfortably large, researchers should face up to the fact that the available data do not support inferences as tight as they might like to achieve.

By and large, the analysis of the book rests on the most elementary probability theory. As will become evident, an enormous amount about identification can be learned from judicious application of the Law of Total Probability and Bayes Theorem. To keep the presentation simple without sacrificing rigor, I suppose throughout that conditioning events have positive probability. With appropriate attention to smoothness and support

conditions, the various propositions that involve conditioning events hold more generally.

The book maintains a consistent notation and usage of terms throughout its ten chapters, with the most basic elements set forth in Chapter 1 and elaborations introduced later as required. Random variables are always in *italics* and their realizations in normal font. The main part of each chapter is written in textbook style, without references to literature. However, each chapter has complements and endnotes that place the analysis in context and elaborate in eclectic ways. The first endnote of each chapter cites the sources on which the chapter draws. These primarily are research articles that I have written, often with co-authors, over the period 1989–2002.

This book complements my earlier book *Identification Problems in the Social Sciences* (Manski, 1995), which exposits basic themes and findings on partial identification in an elementary way intended to be broadly accessible to students and researchers in the social sciences. The present book develops the subject in a rigorous, thorough manner meant to provide the foundation for further study by statisticians and econometricians. Readers who are entirely unfamiliar with partial identification may want to scan at least the introduction and first two chapters of the earlier book before beginning this one.

Identification and Statistical Inference

This book contains only occasional discussions of problems of finite-sample statistical inference. Identification and statistical inference are sufficiently distinct for it to be fruitful to study them separately. As burdensome as identification problems may be, they at least have the analytical clarity of exercises in deductive logic. Statistical inference is a more murky matter of induction from samples to populations.

The usefulness of separating the identification and statistical components of inference has long been recognized. Koopmans (1949, p. 132) put it this way in the article that introduced the term *identification* into the literature:

> In our discussion we have used the phrase "a parameter that can be determined from a sufficient number of observations." We shall now define this concept more sharply, and give it the name *identifiability* of a parameter. Instead of reasoning, as before, from "a sufficiently large number of observations" we shall base our discussion on a hypothetical knowledge of the probability distribution of the observations, as defined more fully below. It is clear that exact knowledge of this probability distribution cannot be derived from any finite number of observations. Such knowledge is the limit approachable but not attainable by extended observation. By hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical

inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is suspect.

Historical Note

Partial identification of population parameters has a long but sparse history in statistical theory. Frisch (1934) developed sharp bounds on the slope parameter of a simple linear regression when the covariate is measured with mean-zero errors; fifty years later, his analysis was extended to multiple regression by Klepper and Leamer (1984). Frechét (1951) studied the conclusions about a joint probability distribution that may be drawn given knowledge of its marginals; see Ruschendorf (1981) for subsequent findings. Duncan and Davis (1953) used a numerical example to show that ecological inference is a problem of partial identification, but formal characterization of identification regions had to wait more than forty years (Horowitz and Manski, 1995; Cross and Manski, 2002). Cochran, Mosteller, and Tukey (1954) suggested conservative analysis of surveys with missing outcome data due to nonresponse by sample members, but Cochran (1977) subsequently downplayed the idea. Peterson (1976) initiated study of partial identification of the competing risk model of survival analysis; Crowder (1991) and Bedford and Meilijson (1997) have carried this work further.

Throughout this book, I begin with the identification region obtained using the empirical evidence alone and study how distributional assumptions may shrink this region. A mathematically complementary approach is to begin with some point-identifying assumption and examine how identification decays as this assumption is weakened in specified ways. Methodological research of the latter kind is variously referred to as *sensitivity*, *perturbation*, or *robustness* analysis. For example, studying the problem of missing outcome data, Rosenbaum (1995) and Scharfstein, Rotnitzky, and Robins (1999) investigate classes of departures from the point-identifying assumption that data are missing at random.