PARTIAL IDENTIFICATION ON PROBABILITY DISTRIBUTIONS

Charles F. Manski

Springer Series in Statistics, 2003
178 pages

This book deals with statistical inference based on data and assumptions that only partially identify population parameters. Its origin is the research of the author on partial identification of probability distributions, started in the late 1980s on nonparametric regression analysis with missing outcome data, and continued by investigating more general incomplete data situations.

The chosen approach to statistical inference is nonparametric analysis, that enables us to learn from the available data without imposing additional assumptions on the population distribution (assumptions that are not often well motivated) and to know about the limitations of the data in order to support inferences about the population parameters.

This book complements an earlier book of the same author, entitled *"Identification Problems in the Social Sciences" (Manski, 1995)*, that introduces in an accessible way the principles of partial identification to students and researchers in the social sciences. The present book develops this subject in a more rigorous manner, with the aim of providing the foundations for further studies by statisticians and econometricians.

The background needed to follow the contents of the book is only elementary probability theory, especially the Law of Total Probability and Bayes Theorem. At the end of each chapter there is some complements and endnotes to place it in context and to provide historical perspective. The first endnote of each chapter cites its sources, basically research articles of the author, alone or with co-authors, until 2002.

There is a common structure in every chapter: first of all, the sampling processes are specified which generate the available data; then the question is considered of what can be said of population parameters without imposing restrictive assumptions on the population distribution, by obtaining the set-valued identification region containing the parameters. Finally, it is studied the possibility that these identification regions may shrink if certain assumptions on the population distribution are imposed, such as statistical independence and monotonicity assumptions. The complementary

approach which begins with some point-identifying assumption and then examine how identification becomes more partial as the assumption is weakened, is not considered here. This last approach is referred to as *sensitivity, perturbation* or *robustness* analysis, and has been followed by Rosenbaum (1995) and Robins (1999) among others.

Chapter 1, "Missing Outcomes", deals with the problem of identification by using only empirical evidence, when the data are generated by random sampling and some outcome realizations are not observable at all. It is also considered the generalization to cases in which data from multiple sampling processes are available, and where outcomes that are observable under some sampling processes may be missing under others; the objective is then to combine data generated by the sampling processes to learn as much as possible about the population distribution. Sometimes the real situation for empirical researchers is the intermediate one, corresponding to the partial knowledge that the realization belongs to a set-valued identification region. This case is studied at the end of the chapter.

Chapter 2, "Instrumental Variables", treats the use of instrumental variables in the formulation of distributional assumptions that help to identify the distribution of outcomes. Some of such assumptions imply point identification, whereas others have less identifying power and, possibly, more credibility. The supposition that data are missing-at-random (MAR) is one of such assumptions, that is weakened to the mean-missing-at-random assumption (MMAR). Another interesting assumption is the mean independence of outcomes of an instrumental variable (MI).

A large part of statistical practice aims to predict outcomes conditional on covariates. In practice it is common to have missing outcomes and/or covariates. While analysis of chapters 1 and 2 extends immediately to inference on conditional outcome distributions when the conditioning event is always observed, in Chapter 3, entitled "Conditional Prediction with Missing Data", the case of data on outcomes and/or conditioning events missing is considered. Therefore, Chapters 1, 2 and 3 form a unit on prediction with missing outcome and/or covariates.

Chapters 4, "Contaminated Outcomes", and 5, "Regressions, Short and Long", form a unit on decomposition of finite mixtures. Inference on the components of finite probability mixtures has application in distinct areas, as contaminated sampling, ecological inference and regression with missing covariate data, that is the problem that originally motivates the author in his research. In fact, Chapter 4 deals with the mixture model of data errors, that presents the available data as realizations of a probability mixture of an error-free realization and a data-error (that imperfectly measures the variable of interest). On the other hand, Chapter 5 studies the problem of ecological inference, that is well known by the social scientists who aim to predict outcomes conditional on (two) covariates. It uses the terminology *short regression* (respectively, *long regression*) from Goldberger (1991), that corresponds to

the conditional expectation of the outcome with respect to one covariate (respectively, with respect both covariates).

The analysis of the response-based sampling, often motivated by practical considerations as cost reduction, is treated in Chapter 6, entitled “Response-based Sampling”. The response-based sampling consists on divide the population into some sub-populations or strata, according to the values of the outcome (response) and sample at random within each stratum. It is particularly effective, for instance, in generating observations of serious diseases, as ill persons are clustered in treatment centres.

Chapter 7, “Analysis of Treatment Response”, and the next chapters form a unit on the study of the problem of missing outcomes given by the non-observability of *counterfactual outcomes* in empirical analysis of treatment response. In studies of treatment response, treatments are mutually exclusive, so it is not possible to observe the outcomes that an experimental unit would experience under other treatment that its own. This study starts with Chapter 7 and is continued in Chapter 8, “Monotone Treatment Response”, that considers the situation in which there exist consistent reasons to believe that outcomes vary monotonically with the intensity of the treatment. Chapter 9, “Monotone Instrumental Variables”, studies identification of mean treatment response under distributional assumptions weaker (and more credible) that the assumption of independence between outcomes and instrumental variables. The last chapter is Chapter 10, “The Mixing Problem”, and studies prediction of outcomes when treatment, unlike the three previous chapters, may vary within the group of experimental units who share the same value of the covariates. In this sense, it is a extrapolation from classical randomized experiments (that does not point-identify outcome distributions under rules in which treatment may vary within groups).

Rosario Delgado
Departament de Matemàtiques
Universitat Autònoma de Barcelona
Spain