PARTIAL IDENTIFICATION OF PROBABILITY DISTRIBUTIONS

Charles F. Manski

Springer-Verlag, 2003

Contents

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.

The usefulness of separating the identification and statistical components of inference has long been recognized.

It has been commonplace to think of identification as a binary event—a parameter is either identified or it is not.

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.

I specify a sampling process generating the available data and first 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.

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.

1. Missing Outcomes

Suppose that each member j of a population J has an outcome y_j in a space Y. The population is a probability space (J, Ω, P) and y: J → Y is a random variable with distribution P(y). A sampling process draws persons at random from J. A realization of y may or may not be observable, as indicated by the realization of a binary random variable z. Thus y is observable if $z = 1$ and not observable if $z = 0$.

By the Law of Total Probability

$$P(y) = P(y|z = 1)P(z = 1) + P(y|z = 0)P(z = 0).$$

The sampling process reveals $P(y|z = 1)$ and $P(z)$, but is uninformative regarding $P(y|z = 0)$. Hence, the empirical evidence reveals that $P(y)$ lies in the *identification region*

$$H[P(y)] \equiv [P(y|z = 1)P(z = 1) + \gamma P(z = 0), \gamma \in \Gamma_Y],$$

where Γ_Y is the space of all probability measures on Y.

Distributional Assumptions

Distributional assumptions may have identifying power. One may assert that the distribution $P(y|z = 0)$ of missing outcomes lies in some set $\Gamma_{0Y} \subset \Gamma_Y$. Then the identification region shrinks from $H[P(y)]$ to

$$H_1[P(y)] \equiv [P(y|z = 1)P(z = 1) + \gamma P(z = 0),\ \gamma \in \Gamma_{0Y}].$$

Or one may assert that the distribution of interest, $P(y)$, lies in some set $H_0[P(y)] \subset \Gamma_Y$. Then the identification region shrinks from $H[P(y)]$ to

$$H_1[P(y)] \equiv H_0[P(y)] \cap H[P(y)].$$

Assumptions of the former type are non-refutable but ones of the latter type may be refutable.

Partial Identification of Parameters

A common objective of empirical research is to infer a parameter of P(y). Let $\tau(\cdot)$: $\Gamma_Y \rightarrow T$ map probability distributions on Y into a space T and consider the problem of inference on the parameter $\tau[P(y)]$. The identification region for $\tau[P(y)]$ is

$$H\{\tau[P(y)]\} = \{\tau(\eta), \eta \in H[P(y)]\}$$

if only the empirical evidence is available and is

$$H_1\{\tau[P(y)]\} = \{\tau(\eta), \eta \in H_1[P(y)]\}$$

given distributional assumptions.

Statistical Inference

An empirical researcher observing a sample of finite size N must contend with issues of statistical inference as well as identification. The empirical distributions $P_N(y|z = 1)$ and $P_N(z)$ almost surely converge to $P(y|z = 1)$ and $P(z)$ respectively. Hence, a natural nonparametric estimate of the identification region $H[P(y)]$ is the sample analog

$$H_N[P(y)] \equiv [P_N(y|z = 1)P_N(z = 1) + \gamma P_N(z = 0),\ \gamma \in \Gamma_Y]$$

and a natural nonparametric estimate of $\{\tau(\eta),\ \eta \in H[P(y)]\}$ is $\{\tau(\eta),\ \eta \in H_N[P(y)]\}$.

Identification Region for the Population Mean

Let $R \equiv [-\infty, \infty]$. Let G be the space of measurable functions that map Y into R and that attain their lower and upper bounds $g_0 \equiv \inf_{y \in Y} g(y)$ and $g_1 \equiv \sup_{y \in Y} g(y)$.

Let the problem of interest be to infer the expectation $E[g(y)]$ using only the empirical evidence. The Law of Iterated Expectations gives

$$E[g(y)] = E[g(y)|z = 1]P(z = 1) + E[g(y)|z = 0]P(z = 0).$$

The sampling process reveals $E[g(y)|z = 1]$ and $P(z)$, but is uninformative regarding $E[g(y)|z = 0]$, which can take any value in the interval $[g_0, g_1]$. Hence

Proposition 1.1: Let $g \in G$. Given the empirical evidence alone, the identification region for $E[g(y)]$ is the closed interval

$$H\{E[g(y)]\} = [E[g(y)|z = 1]P(z = 1) + g_0P(z = 0),$$

$$E[g(y)|z = 1]P(z = 1) + g_1P(z = 0)]. \quad \square$$

Probabilities of Events

Let $g_B(\cdot)$ be the indicator function $g_B(y) \equiv 1[y \in B]$. Then Proposition 1.1 has this corollary.

Corollary 1.1.1: Let B be a non-empty, proper, and measurable subset of Y. Given the empirical evidence alone, the identification region for $P(y \in B)$ is the closed interval

$$H[P(y \in B)] = [P(y \in B|z = 1)P(z = 1),$$

$$P(y \in B|z = 1)P(z = 1) + P(z = 0)]. \quad \square$$

Parameters that Respect Stochastic Dominance

Let Γ_R be the space of probability distributions on the extended real line R. Distribution $F \in \Gamma_R$ stochastically dominates distribution $F' \in \Gamma_R$ if $F[-\infty, t] \leq F'[-\infty, t]$ for all $t \in R$. An extended real-valued function $D(\cdot)$: $\Gamma_R \rightarrow R$ *respects stochastic dominance* (is a *D-parameter*) if $D(F) \geq D(F')$ whenever F stochastically dominates F'.

Proposition 1.2: Let $D(\cdot)$ respect stochastic dominance. Let $g \in G$. Let $R_g \equiv [g(y), y \in Y]$ be the range set of g. Let Γ_g be the space of probability distributions on R_g. Let $\gamma_{0g} \in \Gamma_g$ and $\gamma_{1g} \in \Gamma_g$ be the degenerate distributions that place all mass on g_0 and g_1 respectively. Given the empirical evidence alone, the smallest and largest points in the identification region for $D\{P[g(y)]\}$ are

$D\{P[g(y)|z = 1]P(z = 1) + \gamma_{0g}P(z = 0)\}$

and

$D\{P[g(y)|z = 1]P(z = 1) + \gamma_{1g}P(z = 0)\}$. □

Quantiles

The α–quantile of $P[g(y)]$ is

$$Q_{\alpha}[g(y)] \equiv \min t: \{P[g(y) \leq t] \geq \alpha\}.$$

The smallest feasible value of $Q_{\alpha}[g(y)]$ is the α–quantile of $P[g(y)|z = 1]P(z = 1) + \gamma_{0g}P(z = 0)$ and the largest is the α–quantile of $P[g(y)|z = 1]P(z = 1) + \gamma_{1g}P(z = 0)$.

Outer Bounds on Differences between D-Parameters

Sometimes the parameter of interest is the difference between two D-parameters; that is, a parameter of the form $\tau_{21}\{P[g(y)]\} \equiv D_2\{P[g(y)]\} - D_1\{P[g(y)]\}$.

For example, the interquartile range $Q_{0.75}[g(y)] - Q_{0.25}[g(y)]$ is a familiar measure of the spread of a distribution.

In general, differences between D-parameters are not themselves D-parameters. Nevertheless, Proposition 1.2 may be used to obtain informative *outer bounds* on such differences. A lower bound on $\tau_{21}\{P[g(y)]\}$ is the proposition's lower bound on $D_2\{P[g(y)]\}$ minus its upper bound on $D_1\{P[g(y)]\}$; similarly, an upper bound on $\tau_{21}\{P[g(y)]\}$ is the proposition's upper bound on $D_2\{P[g(y)]\}$ minus its lower bound on $D_1\{P[g(y)]\}$.

The bound on $\tau_{21}\{P[g(y)]\}$ obtained in this manner generally is non-sharp; hence the term *outer bound*.

2. Instrumental Variables

Distributional assumptions may enable one to shrink identification regions obtained using empirical evidence alone. It has been particularly common to assert that

$$P(y) = P(y|z = 0) = P(y|z = 1).$$

$P(y|z = 1)$ is revealed by the sampling process, so $P(y)$ is point-identified.

Researchers almost inevitably find this or other point-identifying assumptions difficult to justify. This should not be surprising. The empirical evidence reveals nothing about the distribution of missing data. An assumption must be strong to pick out one among all possible distributions.

There is a fundamental tension between the credibility and strength of conclusions, which I have called the *Law of Decreasing Credibility*.

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

Inference using the empirical evidence alone sacrifices strength of conclusions in order to maximize credibility. Inference invoking point-identifying assumptions sacrifices credibility in order to achieve strong conclusions. Between these poles, there is a vast middle ground of possible modes of inference asserting assumptions that may shrink the identification region $H[P(y)]$ but not reduce it to a point.

Some Assumptions Using Instrumental Variables

Suppose that each person j is characterized by a covariate v_j in a space V. Let v: J → V be the random variable mapping persons into covariates and let P(y, z, v) denote the joint distribution of (y, z, v). Suppose that all realizations of v are observable. Observability of v provides an instrument or tool that may help to identify the outcome distribution P(y). Thus v is said to be an *instrumental variable*.

The sampling process reveals P(z), P(y, $v|z$ = 1), and P($v|z$ = 0), but is uninformative about the distributions [P($y|v$ = v, z = 0), v ∈ V]. The presence of an instrumental variable does not, per se, help to identify P(y). However, observability of v may be useful when combined with distributional assumptions.

Outcomes Missing-at-Random (Assumption MAR):

$$\mathrm{P}(y|v) \;=\; \mathrm{P}(y|v, z=0) \;=\; \mathrm{P}(y|v, z=1).$$

Statistical Independence of Outcomes and Instruments (Assumption SI):

$$\mathrm{P}(y|v) \;=\; \mathrm{P}(y).$$

Means Missing-at-Random (Assumption MMAR):

$$\mathrm{E}[\mathrm{g}(y)|v] \;=\; \mathrm{E}[\mathrm{g}(y)|v, z=0] \;=\; \mathrm{E}[\mathrm{g}(y)|v, z=1]$$

Mean Independence of Outcomes and Instruments (Assumption MI):

$$\mathrm{E}[\mathrm{g}(y)|v] \;= \mathrm{E}[\mathrm{g}(y)].$$

Proposition 2.1: Let assumption MAR hold. Then P(y) is point-identified with

$$\mathrm{P}(y) \;=\; \textstyle\sum_{\mathrm{v} \in \mathrm{V}} \mathrm{P}(y \,|\, v = \mathrm{v}, z = 1)\mathrm{P}(v = \mathrm{v}).$$

Assumption MAR is non-refutable. □

Proposition 2.2: (a) Let assumption SI hold. Then the identification region for P(y) is

$$\mathrm{H}_{\mathrm{SI}}[\mathrm{P}(y)] \;=$$

$$\bigcap_{\mathrm{v} \in \mathrm{V}} \{\mathrm{P}(y \,|\, v = \mathrm{v}, z = 1)\mathrm{P}(z = 1 \,|\, v = \mathrm{v}) + \gamma_{\mathrm{v}} \cdot \mathrm{P}(z = 0 \,|\, v = \mathrm{v}),$$

$$\gamma_{\mathrm{v}} \in \Gamma_{\mathrm{Y}}\}.$$

(b) Let the set $\mathrm{H}_{\mathrm{SI}}[\mathrm{P}(y)]$ be empty. Then assumption SI does not hold. □

3.Conditional Prediction with Missing Data

A large part of statistical practice aims to predict outcomes conditional on covariates. Suppose that each member j of population J has an outcome y_j in a space Y and a covariate x_j in a space X. Let the random variable (y, x): J → Y × X have distribution P(y, x). In general terms, the objective is to learn the conditional distributions P$(y|x = x)$, x ∈X. A particular objective may be to learn the conditional expectation E$(y|x = x)$, conditional median M$(y|x = x)$, or another point predictor of y conditional on an event $\{x = x\}$.

Suppose that a sampling process draws persons at random from J and realizations of (y, x) may be observable in whole, in part, or not at all. Two binary random variables (z_y, z_x) now indicate observability. A realization of y is observable if $z_y = 1$ but not if $z_y = 0$; a realization of x is observable if $z_x = 1$ but not if $z_x = 0$.

The sampling process reveals P(z_y, z_x), P$(y, x|z_y = 1, z_x = 1)$, P$(y|z_y = 1, z_x = 0)$, and P$(x|z_y = 0, z_x = 1)$. The problem is to use this empirical evidence to infer P$(y|x = x)$, x ∈ X.

Missing Outcomes

Recall identification of P(y) when some realizations of y are missing. The results obtained there apply immediately to P($y|x$ = x) if realizations of x are always observable. One simply needs to redefine the population of interest to be the sub-population of J for which $\{x$ = x$\}$. Then the identification region using the empirical evidence alone is

$$H[P(y|x = x)] =$$

$$[P(y|x = x, z_y = 1)P(z_y = 1|x = x) + \gamma P(z_y = 0|x = x), \quad \gamma \in \Gamma_Y].$$

Jointly Missing Outcomes and Covariates

Proposition 3.1: Let $P(z_y = z_x = 1) + P(z_y = z_x = 0) = 1$. Then

$$H[P(y \mid x = x)] = \{P(y \mid x = x, z_{yx} = 1)r(x) + \gamma[1 - r(x)], \quad \gamma \in \Gamma_Y\},$$

where

$$r(x) \equiv \frac{P(x = x \mid z_{yx} = 1)P(z_{yx} = 1)}{P(x = x \mid z_{yx} = 1)P(z_{yx} = 1) + P(z_{yx} = 0)}. \quad \square$$

Proposition 3.2: Let D respect stochastic dominance. Let $g \in G$. Let $P(z_y = z_x = 1) + P(z_y = z_x = 0) = 1$. Then the smallest and largest points in the identification region for $D\{P[g(y)]\}$ are $D\{P[g(y) \mid z_{yx} = 1]r(x) + \gamma_{0g}[1 - r(x)]\}$ and $D\{P[g(y) \mid z_{yx} = 1]r(x) + \gamma_{1g}[1 - r(x)]\}$. $\square$

Missing Covariates

Proposition 3.5: Let $P(z_y = 1) = 1$. Then

$$H[P(y \mid x = x)] = \bigcup_{p \in [0,\,1]}$$

$$\{\, P(y \mid x = x,\, z_x = 1)\ \frac{P(x = x \mid z_x = 1)P(z_x = 1)}{P(x = x \mid z_x = 1)P(z_x = 1) + pP(z_x = 0)}$$

$$+ \eta\ \frac{pP(z_x = 0)}{P(x = x \mid z_x = 1)P(z_x = 1) + pP(z_x = 0)},\ \eta \in \Gamma_Y(p)\},$$

where

$$\Gamma_Y(p) \equiv \Gamma_Y \cap \{[P(y \mid z_x = 0) - \gamma(1 - p)]/p,\ \ \gamma \in \Gamma_Y\}. \quad \square$$

Joint Inference on Conditional Distributions

Thus far, the object of interest was $P(y \mid x = x)$ for one specified value of x. Researchers often want to predict outcomes when covariates take multiple values. Then the object of interest is the set of conditional distributions $[P(y \mid x = x), x \in X]$ or some functional thereof.

The identification region for $[P(y \mid x = x), x \in X]$ necessarily is a subset of the Cartesian product of the identification regions for each component distribution. Using the empirical evidence alone, that is

$$H[P(y \mid x = x), x \in X] \subset \times_{x \in X} H[P(y \mid x = x)].$$

To go beyond this, one must specify the missing-data problem. The structure of the joint identification region is complex for sampling processes with general patterns of missing data, but simple results hold if only outcomes are missing or if (y, x) are jointly missing.

Parametric Prediction with Missing Data

Researchers often specify a parametric family of predictor functions and seek to infer a member of this family that minimizes expected loss with respect to some loss function. Let the outcome y be real-valued. Let Θ be the parameter space and $f(\cdot, \cdot)$: $X \times \Theta \rightarrow R$ be the family of predictor functions. Let $L(\cdot)$: $R \rightarrow [0, \infty]$ be the loss function. The objective is to find a $\theta^* \in \Theta$ such that

$$\theta^* \in \operatorname{argmin}_{\theta \in \Theta} \; E\{L[y - f(x, \theta)]\}.$$

Then $f(\cdot, \theta^*)$ is called a best $f(\cdot, \cdot)$-predictor of y given x under loss function L.

For example, in best linear prediction under square loss, $f(x, \theta) = x'\theta$, $L[y - f(x, \theta)] = (y - x'\theta)^2$, and $\theta^* = E(xx')^{-1}E(xy)$ if $E(xx')$ is non-singular.

Prediction Using the Empirical Evidence Alone

Using the empirical evidence alone, the identification region for θ^* is the set of parameter values that minimize expected loss under some feasible distribution for the missing data.

$$
\begin{aligned}
H(\theta^*) \;=\; & \bigcup_{(\eta_{10},\, \eta_{00},\, \eta_{01}) \;\in\; \Gamma_{10} \times\; \Gamma_{00} \times \Gamma_{01}} \\
& \{\operatorname{argmin}_{\theta \in \Theta} \;\; P(z_{yx} = 1)\, E\{L[y - f(x, \theta)] \,|\, z_{yx} = 1\} \\
& \quad + P(z_x = 1, z_y = 0) \cdot \int L[y - f(x, \theta)] d\eta_{10} \\
& \quad + P(z_x = 0, z_y = 0) \cdot \int L[y - f(x, \theta)] d\eta_{00} \\
& \quad + P(z_x = 0, z_y = 1) \cdot \int L[y - f(x, \theta)] d\eta_{01}\}.
\end{aligned}
$$

Here Γ_{10} is the set of all distributions on Y × X with x-marginal $P(x \,|\, z_x = 1, z_y = 0)$, Γ_{00} is the set of all distributions on Y × X, and Γ_{01} is the set of all distributions on Y × X with y-marginal $P(y \,|\, z_x = 0, z_y = 1)$.

The natural estimate of $H(\theta^*)$ is its sample analog, which uses the empirical distribution of the data to estimate $P(z_{yx})$, $P[(y, x)|z_{yx} = 1]$, $P(x|z_x = 1, z_y = 0)$, and $P(y|z_x = 0, z_y = 1)$. However, computation of this estimate can pose a considerable challenge. This is so even in the relatively benign setting of best linear prediction under square loss, where the sample analog of $H(\theta^*)$ is the set of least squares estimates produced by conjecturing all possible values for missing outcome and covariate data.