CAE Working Paper #05-09

Generalization of a Result on “Regression, Short and Long”

by

Francesca Molinari
and
Marcin Peski

March 2005

.

Generalization of a Result on "Regressions, Short and Long"[*]

Francesca Molinari
Department of Economics
Cornell University[†]

Marcin Peski
Department of Economics
Northwestern University

March 2005

Abstract

This paper is concerned with the problem of combining a data set that identifies the conditional distribution $P(y|x)$ with one that identifies the conditional distribution $P(z|x)$, in order to identify the regressions $E(y|x,\cdot) \equiv [E(y|x,z=j), j \in Z]$ when the conditional distribution $P(y|x,z)$ is unknown. Cross and Manski (2002) studied this problem and showed that the identification region of $E(y|x,\cdot)$ can be precisely calculated, when y has finite support. Here we generalize Cross and Manski's result showing that the identification region can be precisely calculated also in the case in which y has infinite support.

Motivation and Results

Applied economists often face the problem that no single data set contains all the variables that are necessary to conduct inference on a population of interest. When this is the case, they need to integrate the information contained in different samples; for example, they might need to combine survey data with administrative data (see Ridder and Moffitt (2003) for a survey of the econometrics of data combination). From a methodological perspective, the problem is that while the samples being combined might contain some common variables, other variables belong only to one of the samples. When the data is collected at the same aggregation level (e.g., individual level, household level, etc.), if the common variables include a unique (and correctly recorded) identifier of the units constituting each sample, and there is a substantial overlap of units across all samples, then exact matching of the data sets is relatively straightforward, and the combined data set provides

[*] We are grateful to the coeditor Paolo Paruolo, an anonymous referee, Maria Goltsman, Nick Kiefer, Tymon Tatur and Tim Vogelsang for useful comments. Any remaining errors are our own responsibility.

[†] Corresponding author. 492 Uris Hall, Ithaca NY 14853-7601, USA, fm72@cornell.edu. Financial support from Northwestern University's Dissertation Year Fellowship is gratefully acknowledged.

all the relevant information to identify features of the population of interest. However, it is rather common that there is a limited overlap in the units constituting each sample, or that variables that allow identification of units are not available in one or more of the input files, or that one sample provides information at the individual or household level (e.g., survey data) while the second sample provides information at a more aggregate level (e.g., administrative data providing information at the precinct or district level).

This empirical problem can be formalized in the following way. Suppose that each member l of a population of interest L has an outcome y_l in $\Re$ and covariates $(\mathrm{x}_l, \mathrm{z}_l)$ in a space $X \times Z$. Let the random variables $(y, x, z) : L \rightarrow \Re \times X \times Z$ have distribution $Q(y, x, z)$. Here X is a finite dimensional real space, and Z is a $J-$element finite set. The identification problem arises from the fact that the joint realizations of (y, x, z) are not observable. All that the researcher can observe are realizations from two separate sampling processes, one which draws persons at random from L and collects realizations of (x, z) but not y, and the other that draws persons at random from L and collects realizations of (y, x) but not z. Abstracting from statistical considerations, these sampling processes reveal the conditional probabilities $\Pr(z = j| x)$, $j \in Z$, and the conditional distributions $P(y| x)$, but not the conditional distributions $P(y| x, z = j)$, $j \in Z$. Features of these conditional distributions, e.g., conditional mean, conditional quantiles, etc., are the objects that the researcher would like to identify and estimate; the empirical evidence alone, however, does not allow for point identification and point estimation.

The literature on *statistical matching* has aimed at using the common variable(s) x as a bridge to create synthetic records containing (y, x, z). As Sims (1972) pointed out,[1] the inherent assumption at the base of statistical matching is that conditional on x, y and z are independent. This conditional independence assumption is very strong and untestable. While it does guarantee point identification of features of the conditional distributions $P(y| x, z = j)$, $j \in Z$, it often finds very little justification in practice. Cross and Manski (2002) (CM hereafter) have recently proposed an alternative approach to the problem of data combination when exact matching is not possible. They suggested a method to learn features of the conditional expectations $E(y| x, \cdot) \equiv [E(y| x, z = j),\ j \in Z]$, using the empirical evidence alone and without maintaining any untestable assumption.[2]

CM showed that the vector of conditional expectations $E(y| x, \cdot)$ belongs to an *identification region* (that is, a set of values of $E(y| x, \cdot)$ which are feasible given the maintained assumptions and the empirical evidence), given by a bounded and convex set. They derived the extreme points of such an identification region, showing that they are the expectations of certain $J-$vectors of distributions. When the support of $P(y| x)$ is finite or $J = 2$, CM were able to characterize the identification region fully, as the convex hull of its extreme points. However, for the case in

[1] See also Okner (1972).

[2] See Cross (2000) for an investigation of the identifying power of additional assumptions.

which $P\left(y|\,x\right)$ has infinite support and $J \geq 3$, CM could not fully characterize the topology of the identification region (they did show, however, that the identification region contains the convex hull of its extreme points, and is contained in another convex polytope.). This implies that for empirical problems in which the variable y is continuous (e.g., y is given by capital gains, income, etc.), CM's results provide the extreme points of the identification region, but are not conclusive as to whether the identification region itself is given by the convex hull of these extreme points.

The purpose of this paper is to fully characterize the topology of the identification region in CM. A full characterization allows empirical researchers to precisely calculate the identification region of interest, and to estimate it when only a finite sample is available by replacing population parameters with sample analogs. We show that the identification region is given exactly by the convex hull of its extreme points irrespective of the support of $P\left(y|\,x\right)$. The paper is organized as follows. Section 2 formally states the problem; Proposition 1 presents the main result of the paper, showing that the identification region of $E\left(y|\,x,\cdot\right)$ is closed even when $P\left(y|\,x\right)$ has infinite support and $J \geq 3$. Corollary 2 uses well known results of convex analysis to argue that compactness of the identification region (established through Proposition 1) implies that the identification region is, in fact, given by the convex hull of its extreme points.

Proofs and Discussion

By the Law of Total Probability:

$$P\left(y|\,x\right) = \sum_{j\in Z} P\left(y|\,x, z=j\right)\Pr\left(z=j|\,x\right) \tag{1}$$

and by the Law of Iterated Expectations:

$$E\left(y|\,x\right) = \sum_{j\in Z} E\left(y|\,x, z=j\right)\Pr\left(z=j|\,x\right)$$

Assume that $E\left(y|\,x\right)$ exists and that $\Pr\left(z=j|\,x\right) > 0\ \forall j \in Z$. Following CM's notation, let Ψ denote the set of all probability distributions on $\Re$, and let Γ_x denote the set of all J-vectors of distributions on $\Re$ that satisfy (1). That is, $\left(\psi_{jx}, j\in Z\right) \in \Gamma_x$ if and only if

$$P\left(y|\,x\right) = \sum_{j\in Z} \psi_{jx}\pi_{xj} \tag{2}$$

where $\pi_{xj} \equiv \Pr\left(z=j|\,x\right) > 0\ \forall j \in Z$. Let $\psi_{jx}\left(A\right) = P\left(y\in A|\,x, z=j\right)$, $A\subset\Re$. The identification region of $E\left(y|\,x,\cdot\right)$ is

$$D_x = \left\{\left(\int y d\psi_{jx}, j\in Z\right) : \left(\psi_{jx}, j\in Z\right) \in \Gamma_x\right\}$$

CM show that D_x is a bounded and convex set. Below we show that D_x is a closed set.

Proposition 1 *Suppose that $E\left(y|\, x\right)$ exists and that $\pi_{xj} > 0\ \forall j \in Z$. Then D_x is a closed set.*

Proof. The proof proceeds in two steps. In all that follows we will suppress the subscript x and the conditioning on x for ease of notation, and we will show the following: Suppose that there is a sequence $\xi^n \in D$ such that $\xi^n \to \xi$; then $\xi \in D$. As a starting point, notice that there is a sequence of elements of Γ, $\psi^n = (\psi_1^n, \psi_2^n, \ldots, \psi_J^n) \in \Gamma$, such that $\left(E_{\psi_1^n}y, E_{\psi_2^n}y, \ldots, E_{\psi_J^n}y\right) = \xi^n \in D$, where $E_{\psi_j}y$ will denote $\int y d\psi_j$ in all that follows.

Step 1: The sequence $\{\psi^n\}$ contains a convergent subsequence $\{\psi^{n_m}\}$ with $\psi^{n_m} \Longrightarrow \psi^* \in \Gamma$. As a consequence of (2), for any set $A \subset Y$,

$$\pi_j \psi_j\left(A\right) \leq P\left(y \in A\right)\ \forall j \in Z \tag{3}$$

Since $P\left(y\right)$ is tight, for any $\varepsilon > 0$ we can find a compact set K_ε s.t. $P\left(y \in K_\varepsilon^c\right) < \varepsilon$, where K_ε^c is the complement of K_ε. This implies that for any $\varepsilon > 0$ we can find a compact set K_ε s.t. $\psi_j\left(K_\varepsilon^c\right) < \frac{\varepsilon}{\pi_j}$ $\forall \psi_j \in \Gamma_j$, and hence the set Γ_j is *tight*. Therefore, by Prohorov's theorem, $\left\{\psi_j^n\right\}$ contains a subsequence which converges weakly to a probability measure. Denote $\psi_j^n \supset \left\{\psi_j^{n_m}\right\} \Longrightarrow \psi_j^*$. Since the linear relation in (2) is continuous in the ψ_j's, we have that $\sum \psi_j^* \pi_j = P\left(y|x\right)$ and $\psi^* = \left(\psi_j^*, j \in Z\right) \in \Gamma$.

Step 2. $\left(E_{\psi_1^*}y, E_{\psi_2^*}y, \ldots, E_{\psi_J^*}y\right) = \xi$, hence $\xi \in D$.
Step 2 implies that $\xi^* \equiv \left(\int y \cdot d\psi_j^*(y)\right)_{j=1,..,J} \in D$. We will complete the proof by showing that $\xi = \xi^*$. To this end, we only need to show that for any $j \in \{1, .., J\}$, the distributions $\left\{\psi_j^{n_m}\right\}$ are uniformly integrable. By (3), for any $\alpha \in \Re$

$$\int_{|y| \geq \alpha} |y|\, d\psi_j^n \leq \frac{\int_{|y|\geq\alpha} |y|\, dP}{\pi_j},\ \forall j \in Z,\ \forall n.$$

Therefore

$$0 \leq \lim_{\alpha\to\infty} \sup_n \int_{|y|\geq\alpha} |y|\, d\psi_j^n \leq \frac{1}{\pi_j} \lim_{\alpha\to\infty} \int_{|y|\geq\alpha} |y|\, dP = 0,$$

where the second limit converges since $E\left(y\right)$ exists. This establishes uniform integrability of the distributions $\left\{\psi_j^{n_m}\right\}$, $j \in Z$, and completes the proof.

■

Corollary 2 *Suppose that $E\left(y|\, x\right)$ exists and that $\pi_{xj} > 0\ \forall j \in Z$. Then D_x is given by the convex hull of its extreme points.*

Proof. CM showed that D_x is a bounded convex set. We showed that D_x is closed. Hence D_x is a compact convex set. Standard results of convex analysis (e.g., Rockafellar (1970), Theorem 18.5 and Corollary 18.5.1) insure that a compact convex set in $\Re^J$ is the convex hull of its extreme points, hence the result follows.

■

References

Billingsley, P. (1999): *Convergence of Probability Measures.* John Wiley and Sons, New York.

Cross, P. J. (2000): "Three Essays in Nonparametric Identification," Ph.D. thesis, University of Wisconsin–Madison.

Cross, P. J., and C. F. Manski (2002): "Regressions, Short and Long," *Econometrica*, 70(1), 357–368.

Okner, B. A. (1972): "Constructing a New Microdata Base from Existing Microdata Sets: The 1966 Merge File," *Annals of Economic and Social Measurement*, 1, 325–362.

Ridder, G., and R. Moffitt (2003): "The Econometrics of Data Combination," Chapter for the Handbook of Econometrics, Vol. 6.

Rockafellar, R. T. (1970): *Convex Analysis.* Princeton University Press, Princeton, New Jersey.

Sims, C. A. (1972): "Comments and Rejoinder (On Okner (1972))," *Annals of Economic and Social Measurement*, 1, 343–345, 355–357.