Regressions, Short and Long

Philip J. Cross
Department of Economics
University of Wisconsin - Madison
pcross@ssc.wisc.edu

Charles F. Manski
Department of Economics and
Institute for Policy Research
Northwestern University
cfmanski@nwu.edu

November 1999

Abstract

We study the problem of identification of the long regression $E(y \mid x, z)$ when the short conditional distributions $P(y \mid x)$ and $P(z \mid x)$ are known but the long conditional distribution $P(y \mid x, z)$ is not known. This problem often arises when a researcher utilizes data from two separate data sets. (A leading example is the ecological inference problem of political science, where voting behavior across electoral districts is observed from administrative records, the demographic composition of voters within a district is observed from census data, and the researcher wants to infer voting behavior conditional on district and demographic attributes.) We isolate an identification region containing feasible values of the long regression, and show that this region forms a sharp bound on the long regression. The identification region can be calculated precisely when y has finite support. When y has infinite support we characterize two sets, one that contains the identification region, and one that is contained by it. Following this completely nonparametric analysis, we examine the identifying power yielded by exclusion restrictions across distinct covariate values. Such restrictions cause the identification region to shrink, in many cases to a single point. To illustrate the theory, we pose and address this hypothetical question: What would be the outcome if the 1996 U.S. presidential election were re-enacted in a population of different demographic composition, ceteris paribus?

We have benefitted from the opportunity to present this research in seminars at Northwestern University and the University of Wisconsin - Madison. This research was supported in part by National Science Foundation Grant SBR-9722846.

1 Introduction

Suppose that each member of a population is characterized by a triple (y, x, z). Here y is real-valued, x takes values in a finite dimensional real space X, and z takes values in a J-element finite set Z. Let P denote the population distribution of (y, x, z).

This paper studies the problem of identification of the *long* regression $E(y \mid x, z)$ when the *short* conditional distributions $P(y \mid x)$ and $P(z \mid x)$ are known but the *long* conditional distribution $P(y \mid x, z)$ is not known. The nature of the problem is revealed by the Law of Total Probability,

$$P(y \mid x) = \sum_{j \in Z} \Pr(z = j \mid x) P(y \mid x, z = j). \tag{1}$$

Knowledge of $P(y \mid x)$ and $P(z \mid x)$ restricts $[P(y \mid x, z = j), j \in Z]$ to J-vectors of distributions that satisfy (1). Our objective is to determine the implied restrictions on $E(y \mid x, z)$.

Aspects of the problem of inference on $E(y \mid x, z)$ have been studied in several literatures with varying concerns and terminology. The classical literature on linear regression compares the parameter estimates obtained in a least squares fit of y to x with those obtained in a least squares fit of y to (x, z). The expected difference between the estimated coefficients on x in the former and the latter fits is sometimes called "omitted variable bias". The findings are specific to least squares estimation of linear regressions and so do not directly inform the present nonparametric analysis. We do, however, borrow the terms *short regression* and *long regression* from Goldberger (1991), Sec. 17.2.

Stimulated by Simpson (1951), statisticians have been intrigued by the fact that $E(y \mid x)$ may be increasing in a scalar x and yet all J components of $[E(y \mid x, z = j), j \in Z]$ may be decreasing in x. Studies of *Simpson's Paradox* have sought to characterize the circumstances in which this phenomenon occurs. See, for example, Lindley and Novick (1981) and Zidek (1984).

Following Huber (1964), research on robust estimation under contaminated sampling has taken the object of interest to be $P(y \mid x, z = j)$ for a specified value of j. Values of (y, x, z) with

$z = j$ are said to be error-free, whereas those with $z \neq j$ are said to be erroneous. The researcher only observes (y, x) pairs, not (y, x, z) triples, and so does not know which observations are error free. The researcher is, however, assumed to know the conditional probability $\Pr(z = j \mid x)$ that an observation is error-free, or at least to know a lower bound on this probability. Recently, Horowitz and Manski (1995) showed that equation (1) implies a sharp bound on $E(y \mid x, z = j)$. The lower and upper bounds on $E(y \mid x, z = j)$ are the expectations of certain right-truncated and left-truncated versions of $P(y \mid x)$. This finding forms the starting point for the present analysis.

Our basic findings are developed in Section 2. We prove that the set of feasible values of the J-vector $[E(y \mid x, z = j), j \in Z]$, its *identification region*, is a bounded convex set whose extreme points are the expectations of certain J-vectors of *stacked distributions*. When $P(y \mid x)$ has finite support or $J = 2$, we are able to characterize the identification region fully as the convex hull of these extreme points. When $P(y \mid x)$ has infinite support and $J \geq 3$, we show that the identification region contains this convex hull and is contained in another convex polytope.

Whereas the analysis in Section 2 assumes no information is available beyond knowledge of $P(y \mid x)$ and $P(z \mid x)$, we entertain additional information in Section 3. Here we study *exclusion restrictions* asserting that y is either mean-independent or statistically independent of some component of x, conditional on z and the other components of x. We first characterize abstractly the identifying power of such exclusion restrictions and then present a simple *rank condition* that suffices for point identification of long regressions.

Section 4 applies our findings to the *ecological inference problem* that has long drawn the attention of sociologists and political scientists, especially since Robinson (1950). Social scientists have described ecological inference substantively as inference on individual behavior from aggregate data (e.g., King, 1997). Formally, however, the problem is inference on $P(y \mid x, z)$ given knowledge of $P(y \mid x)$ and $P(z \mid x)$. Focusing on settings in which y and z are both binary variables, Duncan and Davis (1953) and Goodman (1953) performed simple partial analyses

of the identification problem that we address in generality in Sections 2 and 3. We connect our analysis to this early literature and then show how our findings may be applied to election forecasting problems. In particular, we pose and address this hypothetical question: What would be the outcome if the 1996 U.S. presidential election were re-enacted in a population of different demographic composition, ceteris paribus? Assuming only that the long regression $E(y \mid x, z)$ would remain invariant under the hypothesized change in population composition, we obtain informative bounds on the Electoral College vote and, in some cases, are able to predict a winner.

2 Identifying $E(y \mid x, z)$ given knowledge of $P(y \mid x)$ and $P(z \mid x)$

We proceed in three steps. Section 2.1 reviews the sharp bound on the scalar $E(y \mid x, z = j)$ reported in Horowitz and Manski (1995). Section 2.2 uses this bound to characterize the identification region for the J-vector $E(y \mid x, \cdot) \equiv [E(y \mid x, z = j), j \in Z]$. Section 2.3 extends the analysis to $E(y \mid \cdot, \cdot) \equiv [E(y \mid x, \cdot), x \in X]$.

2.1 Identification of $E(y \mid x, z = j)$

Fix x. For $p \in (0, 1)$, let $q_x(p)$ denote the p-quantile of $P(y \mid x)$. Let $L_x(p)$ and $U_x(p)$ be, respectively, the right-truncated and left-truncated distributions defined by

$$L_x(p)(-\infty, t] \equiv \begin{cases} \frac{\Pr[y \leq t \mid x]}{p} & \text{if } t < q_x(p), \\ 1 & \text{if } t \geq q_x(p), \end{cases} \tag{2}$$

$$U_x(p)(-\infty, t] \equiv \begin{cases} 0 & \text{if } t < q_x(1-p), \\ \frac{\Pr[y \leq t \mid x] - (1-p)}{p} & \text{if } t \geq q_x(1-p). \end{cases}$$

Let $P(y \mid x)$ and $P(z \mid x)$ be known. Suppose that $E(y \mid x)$ exists and that $\pi_{xj} \equiv \Pr(z = j \mid x) > 0$ for all $j \in Z$.

Horowitz and Manski (1995), Proposition 4 proves that the smallest and largest feasible values of $E(y \mid x, z = j)$ are the expected values of y under $L_x(\pi_{xj})$ and $U_x(\pi_{xj})$, respectively. Thus

$$E(y \mid x, z = j) \in \left[\int y \, dL_x(\pi_{xj}), \int y \, dU_x(\pi_{xj}) \right] \equiv [\underline{E}_{xj}, \overline{E}_{xj}]. \tag{3}$$

Simple reasoning underlies this result. Consider the sub-population with covariates x. The smallest feasible value of $E(y \mid x, z = j)$ occurs if, within this sub-population, the persons with $z = j$ have the smallest values of y. Then $P(y \mid x, z = j) = L_x(\pi_{xj})$. The largest feasible value occurs if the persons with $z = j$ have the largest values of y. Then $P(y \mid x, z = j) = U_x(\pi_{xj})$.

The bound (3) has a particularly simple form when y is a binary outcome variable, taking the values 0 and 1. Then $q_x(p) = 0$ if $\Pr(y = 1 \mid x) < 1 - p$ and $q_x(p) = 1$ otherwise. It follows that

$$\underline{E}_{xj} = \max \left\{ 0, \frac{\Pr(y = 1 \mid x) - (1 - \pi_{xj})}{\pi_{xj}} \right\}, \qquad \overline{E}_{xj} = \min \left\{ 1, \frac{\Pr(y = 1 \mid x)}{\pi_{xj}} \right\}.$$

A simple direct proof of this result is given in Horowitz and Manski (1995), Corollary 1.2.

The univariate bound (3) immediately implies a bound on the J-vector $E(y \mid x, \cdot)$. That is, $E(y \mid x, \cdot)$ must lie in the J-dimensional rectangle $C_x \equiv \times_{j \in Z} [\underline{E}_{xj}, \overline{E}_{xj}]$. The set C_x, however, is not the sharp bound on $E(y \mid x, \cdot)$. The Law of Total Probability (1) implies further restrictions, including the Law of Iterated Expectations,

$$E(y \mid x) = \sum_{j \in Z} \pi_{xj} E(y \mid x, z = j). \tag{4}$$

Hence $E(y \mid x, \cdot)$ must lie in the intersection of C_x with the hyperplane satisfying (4). In what follows, we characterize further the identification region for $E(y \mid x, \cdot)$.

2.2 Identification of $E(y \mid x, \cdot)$

For each value of x, the feasible values of $E(y \mid x, \cdot)$ follow immediately, albeit abstractly, from the Law of Total Probability (1). Let Ψ denote the space of all probability distributions on

R. Let Γ_x denote the set of all J-vectors of distributions on **R** that satisfy (1). That is, $(\psi_j, j \in Z) \in \Gamma_x$ if, and only if,

$$P(y \mid x) = \sum_{j \in Z} \pi_{xj} \psi_j. \tag{5}$$

Then the identification region for $E(y \mid x, \cdot)$ is

$$D_x = \left\{ \left(\int y \, d\psi_j, j \in Z \right) : (\psi_j, j \in Z) \in \Gamma_x \right\}. \tag{6}$$

Some properties of D_x are immediate. The set Γ_x is convex and the expectation operator is linear, so D_x is convex. Moreover, D_x is contained within the J-dimensional rectangle C_x. Hence D_x is a bounded convex set.

Our objective is to characterize D_x more precisely. Proposition 1 shows that D_x has at most $J!$ distinct extreme points, these being the expectations of the stacked distributions defined below. Following Proposition 1, we develop some immediate implications through two Corollaries.

The stacked distributions J-vectors of *stacked distributions* are sequences of J distributions such that the entire probability mass of the jth distribution lies weakly to the left of that of the $(j+1)$-st distribution. To describe these distribution sequences, we now let Z be the ordered set of integers $(1, \ldots, J)$. This set has $J!$ permutations, each of which generates a distinct J-vector of stacked distributions. We label these permutations of Z as Z^m, $m = 1, \ldots, J!$, and the corresponding J-vectors of stacked distributions as $(P^m_{xj}, j = 1, \ldots, J), m = 1, \ldots, J!$.

For each value of m, the elements of $(P^m_{xj}, j = 1, \ldots, J)$ solve a recursive set of minimization problems. In what follows, we show the construction of $(P^1_{xj}, j = 1, \ldots, J)$, which is based on Z^1, the original ordering of Z. The other $(J! - 1)$ J-vectors are generated from Z^m, $m = 2, 3, \ldots, J!$, which alters the order in which the recursion is performed.

For each $j = 1, \ldots, J, P^1_{xj}$ is chosen to minimize its expectation subject to the distributions earlier chosen for $(P^1_{xi}, i < j)$, and subject to the global condition that equation (5) must hold.

The recursion is as follows: For $j = 1, \ldots, J$, P^1_{xj} solves the problem

$$\min_{\psi \in \Psi} \int y \, d\psi \tag{7}$$

subject to

$$P(y \,|\, x) = \sum_{i=1}^{j-1} \pi_{xi} P^1_{xi} + \pi_{xj} \psi + \sum_{k=j+1}^{J} \pi_{xk} \psi_k, \tag{8}$$

where $\psi_k \in \Psi$, $k = j+1, \ldots, J$ are unrestricted probability distributions.

This recursion yields a sequence of stacked distributions. For $j = 1$, equation (8) reduces to

$$P(y \,|\, x) = \pi_{x1} \psi + \sum_{k=2}^{J} \pi_{xk} \psi_k. \tag{9}$$

Horowitz and Manski (1995), Proposition 4 shows that the distribution solving (7) subject to (9) is $L_x(\pi_{x1})$, the right-truncated version of $P(y \,|\, x)$ defined in (2); thus $P^1_{x1} = L_x(\pi_{x1})$. For $j = 2$, equation (8) has the form

$$P(y \,|\, x) = \pi_{x1} L_x(\pi_{x1}) + \pi_{x2} \psi + \sum_{k=3}^{J} \pi_{xk} \psi_k. \tag{10}$$

The proof of Horowitz and Manski's (1995) Proposition 4 shows that

$$P(y \,|\, x) = \pi_{x1} L_x(\pi_{x1}) + (1 - \pi_{x1}) U_x(1 - \pi_{x1}), \tag{11}$$

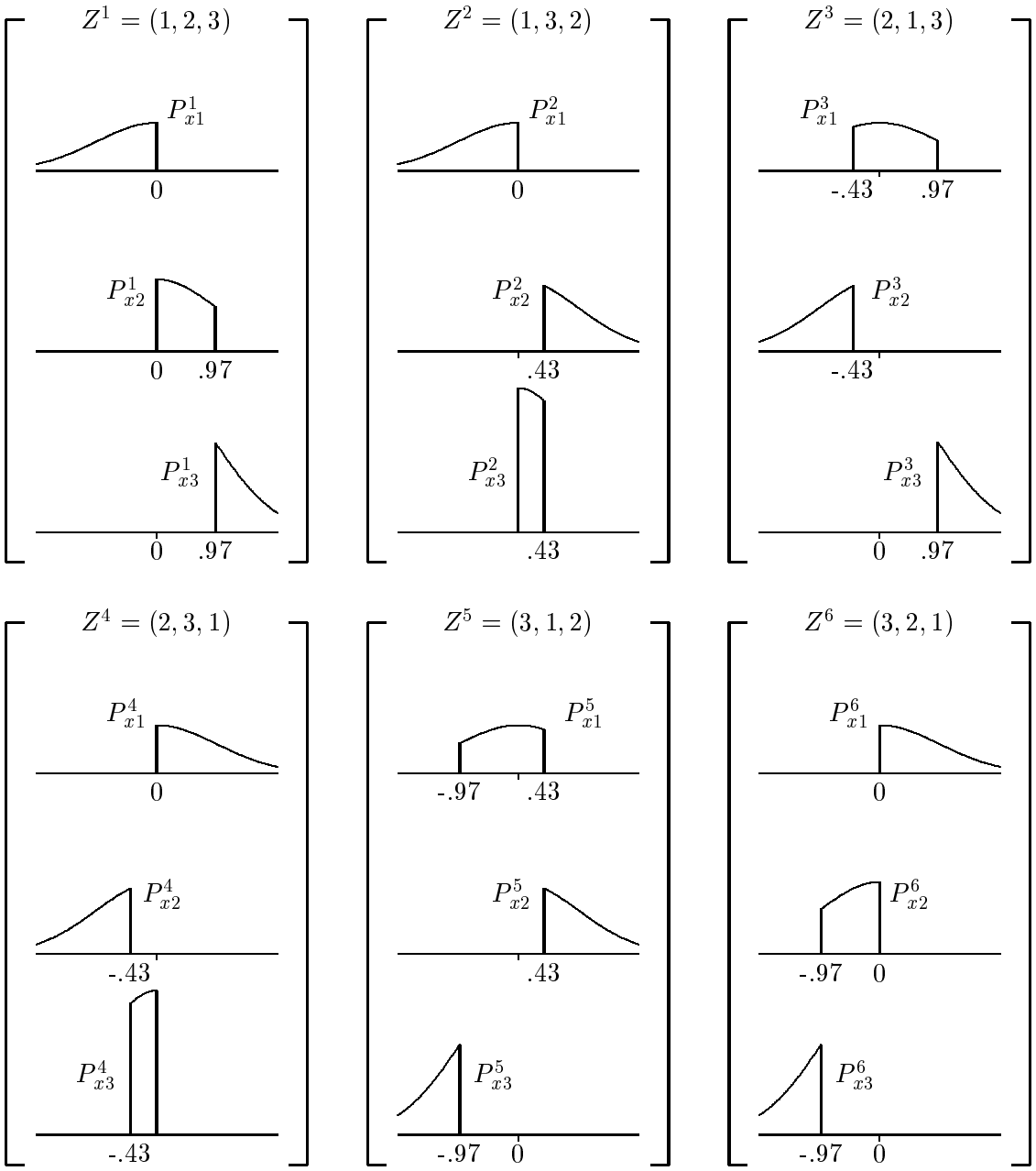
where $U_x(1 - \pi_{x1})$ is the left-truncated version of $P(y \,|\, x)$ that maximizes $E(y \,|\, x, z > 1)$. Hence (8) becomes

$$U_x(1 - \pi_{x1}) = \frac{\pi_{x2}}{1 - \pi_{x1}} \psi + \sum_{k=3}^{J} \frac{\pi_{xk}}{1 - \pi_{x1}} \psi_k. \tag{12}$$

Equation (12) has the same form as (9), with $U_x(1-\pi_{x1})$ replacing $P(y \,|\, x)$ and $\pi_{x,k+1}/(1-\pi_{x1})$ replacing π_{xk}. Hence P^1_{x2}, the solution to (7) subject to (12), is a right-truncated version of $U_x(1-\pi_{x1})$. By definition, $L_x(\pi_{x1})$ has no mass to the right of the point $q_x(\pi_{x1})$ and $U_x(1-\pi_{x1})$ has no mass to the left of this point. Hence P^1_{x1} and P^1_{x2} are stacked side-by-side, with all of the mass of the former distribution lying weakly to the left of the mass of the latter distribution. The distributions $(P^1_{xj}, j = 3, \ldots, J)$ are similarly stacked. For each j, the mass of P^1_{xj} lies to weakly to the left of the mass of $P^1_{x,j+1}$. The supremum of the support of P^1_{xj} may equal the

infimum of the support of $P^1_{x,j+1}$, but otherwise the distributions are concentrated on disjoint intervals.

Figure 1. Densities of stacked distributions for $P(y \mid x)$ standard normal, for each of the six permutations of $Z = (1, 2, 3)$.



Example: $P(y \mid x)$ standard normal, $\pi_{x1} = 1/2, \pi_{x2} = 1/3$ and $\pi_{x3} = 1/6$.

Since $J = 3$ in this example, there are $3! = 6$ 3-vectors of stacked distributions, based on Z^1 through Z^6. These are illustrated by their densities in Figure 1. Notice that the first vector of stacked distributions in the figure is $(P^1_{xj}, j = 1, 2, 3)$. P^1_{x1} is $L_x(1/2)$, the standard normal right-truncated at 0. P^1_{x2} is constructed by right-truncation at 0.97 of the distribution resulting from $L_x(1/2)$ being removed from the standard normal. And the remaining mass, which is $U_x(1/6)$, constitutes P^1_{x3}. The second vector of stacked distributions in Figure 1 is $(P^2_{xj}, j = 1, 2, 3)$, where we define $Z^2 = (1, 3, 2)$. And the remaining vectors of stacked distributions in the figure are derived from the remaining permutations, Z^3 through Z^6.

The extreme points of the identification region With the above as preliminary, Proposition 1 proves that the expectations of the stacked distributions are the extreme points of D_x.

PROPOSITION 1: Let $P(y \mid x)$ and $P(z \mid x)$ be known. Let $E(y \mid x)$ exist. Let $E^m_x \equiv (\int y\, dP^m_{xj}, j = 1, \ldots, J)$. Then the extreme points of D_x are $\{E^m_x, m = 1, \ldots, J!\}$.

Proof: By construction, each of the J-vectors in $\{E^m_x, m = 1, \ldots, J!\}$ is a feasible value of $E(y \mid x, \cdot)$. Step (i) of the proof shows that these vectors are extreme points of D_x. Step (ii) shows that D_x has no other extreme points. In what follows, we simplify the notation by suppressing the subscript x.

Step (i). It suffices to consider E^1. Permuting Z does not alter the argument below.

Suppose that E^1 is not an extreme point of D. Then there exist an $\alpha \in (0, 1)$ and distinct J-vectors $(\xi', \xi'') \in D$ such that $E^1 = \alpha\xi' + (1 - \alpha)\xi''$. Suppose that E^1, ξ', and ξ'' differ in their first component. Then either $\xi'_1 < E^1_1 < \xi''_1$ or $\xi''_1 < E^1_1 < \xi'_1$. By construction, however, $E^1_1 = \underline{E}^1$, the global minimum of $E(y \mid z = 1)$. So $\xi'_1 \geq E^1_1$ and $\xi''_1 \geq E^1_1$. Hence it must be the case that $\xi''_1 = \xi'_1 = E^1_1$.

Now suppose that E^1, ξ', and ξ'' differ in their second component. Then $\xi_2' < E_2^1 < \xi_2''$ or $\xi_2'' < E_2^1 < \xi_2'$. But E_2^1 minimizes $E(y \,|\, z = 2)$ subject to the previous minimization of $E(y \,|\, z = 1)$. So $\xi_2' \geq E_2^1$ and $\xi_2'' \geq E_2^1$. Hence $\xi_2'' = \xi_2' = E_2^1$. Recursive application of this reasoning shows that $\xi'' = \xi' = E^1$, contrary to supposition. Hence E^1 is an extreme point of D.

Step (ii). Let $\xi \in D$, with $\xi \notin \{E^m, m = 1, \ldots, J!\}$. Then ξ is the expectation of some feasible J-vector of non-stacked distributions. We want to show that ξ is not an extreme point of D. Thus, we must show that there exists an $\alpha \in (0, 1)$ and distinct J-vectors $(\xi', \xi'') \in D$ such that $\xi = \alpha\xi' + (1 - \alpha)\xi''$.

Let the set-valued function $S(\psi)$ denote the support of any probability distribution ψ on the real line. Let $(\psi_j, j \in Z) \in \Gamma$ be any feasible J-vector of distributions with expectation ξ. This J-vector is not stacked, so there exist components ψ_i and ψ_k such that $[\inf S(\psi_i), \sup S(\psi_i)] \cap [\inf S(\psi_k), \sup S(\psi_k)]$ has positive length. Thus $\sup S(\psi_i) > \inf S(\psi_k)$ and $\sup S(\psi_k) > \inf S(\psi_i)$. For ease of exposition, henceforth let $a_j \equiv \inf S(\psi_j)$ and $b_j \equiv \sup S(\psi_j)$, for $j = i, k$.

We now construct a feasible J-vector of distributions that shifts mass, in a particular balanced manner, between distributions ψ_i and ψ_k, while leaving the other components of $(\psi_j, j \in Z)$ unchanged. Let $0 < \varepsilon < \frac{1}{2}(b_i - a_k)$. Then $\psi_k[a_k, a_k + \varepsilon] > 0$, $\psi_i[b_i - \varepsilon, b_i] > 0$, and $[a_k, a_k + \varepsilon] \cap [b_i - \varepsilon, b_i] = \emptyset$. Let

$$\lambda \equiv \frac{\pi_k \psi_k[a_k, a_k + \varepsilon]}{\pi_i \psi_i[b_i - \varepsilon, b_i]}.$$

Now define the new J-vector $(\psi_j', j \in Z)$ as follows: Let $\psi_j' = \psi_j$ for $j \neq i, k$. If $\lambda \leq 1$, let

$$[\psi_i'(y), \psi_k'(y)] = \begin{cases} [\ \psi_i(y) + \frac{\pi_k}{\pi_i}\psi_k(y) \ , \ 0 \ ] & \text{if } y \in [a_k, a_k + \varepsilon], \\ [\ (1 - \lambda)\psi_i(y) \ , \ \psi_k(y) + \lambda\frac{\pi_i}{\pi_k}\psi_i(y) \ ] & \text{if } y \in [b_i - \varepsilon, b_i], \\ [\ \psi_i(y) \ , \ \psi_k(y) \ ] & \text{otherwise.} \end{cases}$$

Alternatively, if $\lambda > 1$, let

$$[\psi_i'(y), \psi_k'(y)] = \begin{cases} [\quad \psi_i(y) + \frac{1}{\lambda}\frac{\pi_k}{\pi_i}\psi_k(y) \quad , \quad (1-\frac{1}{\lambda})\psi_k(y) \quad] & \text{if } y \in [a_k, a_k+\varepsilon], \\ [\quad 0 \quad , \quad \psi_k(y) + \frac{\pi_i}{\pi_k}\psi_i(y) \quad] & \text{if } y \in [b_i-\varepsilon, b_i], \\ [\quad \psi_i(y) \quad , \quad \psi_k(y) \quad] & \text{otherwise.} \end{cases}$$

Thus, the new J-vector shifts ψ_i mass leftward from the $[b_i - \varepsilon, b_i]$ interval to the $[a_k, a_k + \varepsilon]$ interval and compensates by shifting ψ_k mass rightward to the $[b_i - \varepsilon, b_i]$ interval from the $[a_k, a_k + \varepsilon]$ interval. The λ parameter ensures that we shift equal amounts of mass and that $\pi_i\psi_i' + \pi_k\psi_k' = \pi_i\psi_i + \pi_k\psi_k$. Hence $(\psi_j', j \in Z)$ is a feasible J-vector of distributions; that is, an element of Γ . The mean of $(\psi_j', j \in Z)$ is related to the mean of $(\psi_j, j \in Z)$ as follows: $\xi_i' < \xi_i$, $\xi_k' > \xi_k$, and $\xi_j' = \xi_j$ for $j \neq i, k$.

An analogous operation switching the roles of i and k produces another new J-vector $(\psi_j'', j \in Z)$. Now let $0 < \varepsilon < \frac{1}{2}(b_k - a_i)$ and redefine λ accordingly. This construction shifts ψ_k mass leftward from the $[b_k - \varepsilon, b_k]$ interval to the $[a_i, a_i + \varepsilon]$ interval and shifts an equal amount of ψ_i mass rightward to the $[b_k - \varepsilon, b_k]$ interval from the $[a_i, a_i + \varepsilon]$ interval, while ensuring that $\pi_i\psi_i'' + \pi_k\psi_k'' = \pi_i\psi_i + \pi_k\psi_k$. The mean of this J-vector is related to the mean of $(\psi_j, j \in Z)$ as follows: $\xi_i'' > \xi_i$, $\xi_k'' < \xi_k$, and $\xi_j'' = \xi_j$ for $j \neq i, k$.

It follows from the above that $\pi_i\xi_i'' + \pi_k\xi_k'' = \pi_i\xi_i' + \pi_k\xi_k' = \pi_i\xi_i + \pi_k\xi_k$. Thus, (ξ_i, ξ_k) lies on the line connecting (ξ_i', ξ_k') and (ξ_i'', ξ_k''). Moreover, $\xi_i'' > \xi_i > \xi_i'$ and $\xi_k' > \xi_k > \xi_k''$. Hence (ξ_i, ξ_k) is a strictly convex combination of (ξ_i', ξ_k') and (ξ_i'', ξ_k''). Finally recall that $\xi_j'' = \xi_j' = \xi_j$ for $j \neq i, k$. Hence ξ is a strictly convex combination of ξ' and ξ''. Thus ξ is not an extreme point of D. *Q.E.D.*

Proposition 1 has two immediate implications that further characterize the identification region. Let conv $\{E_x^m, m = 1, \ldots, J!\}$ denote the convex hull of $\{E_x^m, m = 1, \ldots, J!\}$. Then we have

PROPOSITION 1, COROLLARY 1: $\text{conv}\,\{E_x^m, m = 1, \ldots, J!\} \subset D_x$.

Proof: D_x is a convex set containing $\{E_x^m, m = 1, \ldots, J!\}$. Hence D_x contains the convex hull of these points. *Q.E.D.*

PROPOSITION 1, COROLLARY 2: If $P(y \mid x)$ has finite support, then $D_x = \text{conv}\,\{E_x^m, m = 1, \ldots, J!\}$.

Proof: Minkowski's Theorem (e.g., Brøndsted, 1983, Theorem 5.10) shows that a compact convex set in $\mathbf{R}^J$ is the convex hull of its extreme points. We already know that D_x is a bounded convex set, so we need only show that D_x is closed. Let Y denote the support of $P(y \mid x)$ and suppose that Y has finite cardinality H. For $j \in Z$ and $\eta \in Y$, let $\varphi_{j\eta}$ be a feasible value for $\Pr(y = \eta \mid x, z = j)$. Then equation (5) becomes the following system of H linear equations in the $J \times H$ unknowns $\varphi_{j\eta}$:

$$\Pr(y = \eta \mid x) = \sum_{j \in Z} \pi_{xj} \varphi_{j\eta}, \qquad \eta \in Y.$$

Let Φ_x denote the solutions to this system of equations. Φ_x forms a closed set in $\mathbf{R}^{J \times H}$. The identification region for $E(y \mid x, \cdot)$ is $D_x = \{\, (\sum_{\eta \in Y} \eta \cdot \varphi_{j\eta}, j \in Z) \,:\, \varphi \in \Phi_x \,\}$, a linear map from Φ_x to $\mathbf{R}^J$. Hence D_x is closed. *Q.E.D.*

The identification region when $P(y \mid x)$ has infinite support Proposition 1 and its Corollaries fully characterize the identification region when $P(y \mid x)$ has finite support, but only partially so when $P(y \mid x)$ has infinite support. If D_x can be shown to be closed, then the reasoning of Corollary 2 may be applied. Unfortunately, it appears difficult to characterize D_x topologically when $P(y \mid x)$ has infinite support.

Although we currently are not able to characterize fully the identification region when $P(y \mid x)$ has infinite support, we can add to the characterization given thus far. We have already

shown that D_x contains the convex polytope $\text{conv}\,\{E_x^m, m = 1, \ldots, J!\}$. Proposition 2 uses $\{E_x^m, m = 1, \ldots, J!\}$ to construct another convex polytope that contains D_x. When $J = 2$, this yields a full characterization of D_x.

PROPOSITION 2: For each $m = 1, \ldots, J!$, let Z^m denote the mth permutation of Z. Let $j(m,k)$ be the position in Z of the kth element of Z^m. Define the following subsets of $\mathbf{R}^J$:

$$G_x^0 \equiv \left\{ \xi \in \mathbf{R}^J \,:\, \sum_{j=1}^{J} \pi_{xj}\xi_j = E(y \,|\, x) \right\},$$

$$G_x^m \equiv \left\{ \xi \in \mathbf{R}^J \,:\, \sum_{k=1}^{n} \pi_{xj(m,k)}\xi_k \geq \sum_{k=1}^{n} \pi_{xj(m,k)} E_{xj(m,k)}^m,\ n = 1, \ldots, J-1 \right\}, \qquad m = 1, \ldots, J!,$$

and

$$G_x \equiv \bigcap_{m=0}^{J!} G_x^m.$$

Then G_x is a convex polytope and $\text{conv}\,\{E_x^m, m = 1, \ldots, J!\} \subset D_x \subset G_x$. When $J = 2$, $\text{conv}\,\{E_x^m, m = 1, \ldots, J!\} = D_x = G_x$.

Proof: Proposition 1, Corollary 1 showed that $\text{conv}\,\{E_x^m, m = 1, \ldots, J!\} \subset D_x$. It is easy to see that $D_x \subset G_x$. The Law of Iterated Expectations (4) requires that every point in D_x satisfy the equality defining G_x^0. For each $m \geq 1$, the construction of E_x^m by recursive minimization implies that every point in D_x must satisfy each of the $(J-1)$ inequalities defining G_x^m. Hence $D_x \subset G_x$.

To show that G_x is a convex polytope, observe first that G_x^0 is a hyperplane and each G_x^m is the intersection of $(J-1)$ closed half-spaces. Hence G_x is a polyhedral set. Next observe that G_x is bounded from below. In particular, the first inequality used to define each set G_x^m shows that $\xi \in G_x \Longrightarrow \xi_j \geq \underline{E}_{xj}, j \in Z$. Finally, observe that this lower bound and the equality defining G_x^0 imply that G_x is bounded from above; in particular, $\xi \in G_x \Longrightarrow \xi_i \leq E(y \,|\, x) - \sum_{k \neq i} \pi_{xj(m,k)} \underline{E}_{xk}, i \in Z$. Thus G_x is a bounded polyhedral set, and hence a convex polytope. See Brøndsted (1983), Corollary 8.7.

When $J = 2$, G_x is the line segment connecting the points $(\underline{E}_{x1}, \overline{E}_{x2})$ and $(\overline{E}_{x1}, \underline{E}_{x2})$, which are the extreme points of D_x. So $\text{conv}\{E_x^m, m = 1, \ldots, J!\} = D_x = G_x$ in this special case. *Q.E.D.*

2.3 Identification of $E(y \,|\, \cdot, \cdot)$

It remains only to extend the analysis from identification of $E(y \,|\, x, \cdot)$ to identification of $E(y \,|\, \cdot, \cdot)$. This is straightforward. Knowledge of $P(y \,|\, x)$ and $P(z \,|\, x)$ implies no cross-x restrictions on $E(y \,|\, x, \cdot)$. Hence the identification region for $E(y \,|\, \cdot, \cdot)$ is the Cartesian product $\times_{x \in X} D_x$.

3 The identifying power of exclusion restrictions

Propositions 1 and 2 have characterized the restrictions on $E(y \,|\, x, z)$ implied by knowledge of $P(y \,|\, x)$ and $P(z \,|\, x)$. Tighter inferences may be feasible if additional information is available. Among the many forms that such information may take, we focus on exclusion restrictions of the type that have been found useful in resolving other identification problems.

Let us dispose first of one form of exclusion restriction whose implications are so immediate as barely to require comment. Suppose it is known that y is mean-independent of z, conditional on x; that is, $E(y \,|\, x, z) = E(y \,|\, x)$. Then knowledge of $P(y \,|\, x)$ identifies $E(y \,|\, x, z)$.

More interesting are exclusion restrictions connecting $E(y \,|\, x, z)$ across different values of x. Let $x = (v, w)$ and $X = V \times W$. One familiar form of exclusion restriction asserts that y is mean-independent of v, conditional on (w, z). Thus

$$E(y \,|\, v, w, z) = E(y \,|\, w, z). \tag{13}$$

A stronger form of exclusion asserts that y is statistically independent of v, conditional on (w, z). Thus

$$P(y \,|\, v, w, z) = P(y \,|\, w, z). \tag{14}$$

Restrictions of these forms are often called *instrumental variable* assumptions, v being the instrumental variable.

Proposition 3 below characterizes fully, albeit abstractly, the identifying power of assumptions (13) and (14). We then present a weaker, but much simpler, finding that yields a straightforward *rank condition* for point identification of $E(y \mid w, \cdot) \equiv [E(y \mid w, z = j), j \in Z]$. This rank condition indicates that, in applications, exclusion restrictions of the form (13) and (14) often suffice to identify $E(y \mid w, \cdot)$. We also call attention to the fact that these exclusion restrictions are testable assumptions.

PROPOSITION 3: Let $P(y \mid v, w)$ and $P(z \mid v, w)$ be known. Let $E(y \mid v, w)$ exist. Let D_w^* and D_w^{**} denote the identification regions for $E(y \mid w, \cdot)$ under assumptions (13) and (14) respectively. Then

$$D_w^* \equiv \bigcap_{v \in V} D_{(v,w)}, \tag{15}$$

and

$$D_w^{**} \equiv \left\{ \left(\int y \, d\psi_i, j \in Z \right) : (\psi_j, j \in Z) \in \bigcap_{v \in V} \Gamma_{(v,w)} \right\} \subset D_w^*. \tag{16}$$

The corresponding identification regions for $E(y \mid \cdot, \cdot)$ are $\times_{w \in W} D_w^*$ and $\times_{w \in W} D_w^{**}$.

Proof: Consider assumption (13). Recall that, for each value of (v, w), we have $(\psi_j, j \in Z) \in \Gamma_{(v,w)}$ if, and only if,

$$P(y \mid v, w) = \sum_{j \in Z} \pi_{(v,w)j} \psi_j.$$

Let $\xi \in \mathbf{R}^J$. Under (13), ξ is a feasible value for $E(y \mid w, \cdot)$ if, and only if, for every $v \in V$ there exists an element of $\Gamma_{(v,w)}$ whose expectation is ξ. The set D_w^* comprises these feasible values of ξ.

Consider assumption (14). Under (14), $(\psi_j, j \in Z)$ is a feasible value for $[P(y \mid w, j), j \in Z]$

if, and only if, $(\psi_j, j \in Z)$ satisfies the system of equations

$$P(y \,|\, v, w) = \sum_{j \in Z} \pi_{(v,w)j} \psi_j, \quad \text{for all } v \in V.$$

Thus the set of feasible values for $[\, P(y \,|\, w, j), j \in Z \,]$ is $\bigcap_{v \in V} \Gamma_{(v,w)}$. The set D_w^{**} comprises the expectations of these feasible J-vectors of distributions. That $D_w^{**} \subset D_w^*$ follows from the fact that assumption (14) is stronger than (13). It can also be seen directly by comparing (15) and (16).

Now consider $E(y \,|\, \cdot, \cdot)$. Neither (13) nor (14) imposes a cross-w restriction. Hence the identification regions for $E(y \,|\, \cdot, \cdot)$ are the Cartesian products of the regions for $E(y \,|\, w, \cdot)$ under these assumptions. *Q.E.D.*

A rank condition for point identification Proposition 3 is general, but it is too abstract to convey a sense of the identifying power of exclusion restrictions. A much simpler, readily applicable finding emerges if we exploit only the Law of Iterated Expectations rather than the full force of the Law of Total Probability.

Let $C_w^* \subset \mathbf{R}^J$ denote the set of solutions $\xi \in \mathbf{R}^J$ to the system of linear equations

$$E(y \,|\, v, w) = \sum_{j \in Z} \pi_{(v,w)j} \xi_j, \qquad \text{for all } v \in V. \tag{17}$$

Let $|V|$ denote the cardinality of the set V. Let Π denote the $|V| \times J$ matrix whose jth column is $(\pi_{(v,w)j}, v \in V)$. Then we have

Proposition 3, Corollary 1: $D_w^* \subset C_w^*$. If Π has rank J, then C_w^* is a singleton and $D_w^* = C_w^*$.

Proof: The Law of Iterated Expectations and assumption (13) require that feasible values of $E(y \,|\, w, \cdot)$ solve equations (17). Hence $D_w^* \subset C_w^*$. D_w^* is non-empty, so (17) must have at least one solution. If Π has rank J, then (17) has a unique solution and $D_w^* = C_w^*$. *Q.E.D.*

Testing exclusion restrictions We have thus far supposed that the specified exclusion restriction is correct. Suppose that an attempt to solve the system of equations (17) reveals that the solution set C_w^* is empty. Or, if C_w^* is non-empty, suppose that evaluation of the identification region D_w^* or D_w^{**}, as the case may be, shows the region to be empty. Any such finding implies that the specified exclusion restriction cannot be correct. Thus, exclusion restrictions of the form (13) and (14) are testable assumptions.

4 Application to ecological inference

The ecological inference problem provides a rich setting within which to demonstrate the use of Propositions 1 through 3. Section 4.1 connects our analysis to the literature on ecological inference. Section 4.2 poses a forecasting task to which the analysis may be applied. Section 4.3 uses available data to carry out the application.

4.1 Background

An application in political science serves well to illustrate the ecological inference problem. Political scientists have long been interested in the empirical variation in voting behavior across the population. Sample surveys yielding information on individual attributes and voting behavior are not always available and, when they are, the credibility of self-reports of voting behavior may be open to question. Hence political scientists have often sought to infer voting patterns from two data sources that are readily available and credible: (a) administrative records on voting by electoral district, and (b) census data on the attributes of persons in each district.

To formalize this, let y denote the voting outcome of interest. Let x denote an electoral district. Let z denote voter attributes thought to be related to voting behavior. Political scientists want to learn features of $P(y \mid x, z)$, the distribution of voting outcomes among persons

in district x with attributes z. Voting records may reveal $P(y \mid x)$ and census data may reveal $P(z \mid x)$. Ecological inference is inference on $P(y \mid x, z)$ from this information on $P(y \mid x)$ and $P(z \mid x)$.

The early major contributions to analysis of the ecological inference problem appeared in the sociology literature in the 1950s. Robinson (1950) criticized the common practice of interpreting the *ecological correlation*, the cross-x correlation of $P(y \mid x)$ and $P(z \mid x)$, as the correlation of y and z. Soon afterwards, two influential short papers were published in the same issue of the *American Sociological Review*. These papers, Duncan and Davis (1953) and Goodman (1953), foreshadowed the analysis we have presented in Sections 2 and 3, respectively.

Duncan and Davis, considering problems in which both y and z are binary, used numerical illustrations to demonstrate that knowledge of $P(y \mid x)$ and $P(z \mid x)$ implies a bound on $P(y \mid x, z)$. Duncan and Davis did not formalize the bound, but it is clear from their illustrations that they had in mind the sharp bound given in Horowitz and Manski (1995, Corollary 1.2) and, independently, in King (1997, Section 5.2). Goodman (1953), also considering problems in which y and z are binary, essentially showed that knowledge of $E(y \mid x)$ and $P(z \mid x)$ combined with an exclusion restriction yields the rank condition for point identification developed in our Proposition 3, Corollary 1.

Recent contributions to the literature on ecological inference have developed alternative routes to point identification of $P(y \mid x, z)$. In particular, see Freedman et al. (1991), King (1997), and the ensuing dispute played out in the *Journal of the American Statistical Association* (Freedman et al., 1998, 1999; King, 1999). Research has continued to focus on settings in which y is a binary outcome. There appears to be no precedent for the Horowitz and Manski (1995, Proposition 4) finding of a sharp bound on the expected value of a real-valued outcome. Nor do there appear to be precedents for our Propositions 1 through 3.

4.2 An illustrative application: forecasting the electoral effects of demographic changes

To illustrate the uses of Propositions 1 through 3, we now pose an instance of the ecological inference problem. In dynamic societies, the composition of the population changes over time as the net result of migration flows, variation in fertility and mortality rates, economic growth, and so on. We shall apply Propositions 1 through 3 to the problem of forecasting the electoral effects of these demographic changes. To make the application concrete, we pose a specific hypothetical question:

> *What would be the outcome if the 1996 U.S. presidential election were re-enacted in a population of different composition, ceteris paribus?*

Here *ceteris paribus* means that we assume the same candidates would be nominated, that these candidates would use the same election strategies, and so on. Of course, the political parties might nominate different candidates and alter their strategies if the composition of the population were to differ. Nevertheless, the ceteris paribus scenario poses an interesting baseline forecasting task.

To formalize the question, let x denote a state of the U.S., or the District of Columbia. Let z denote attributes of individual voters thought to be related to voting behavior; for concreteness we shall later let z indicate the age and ethnicity of a voter in state x. Let $Y = \{-1, 0, 1\}$ be the set of voting outcomes; $y = 1$ if a person votes Democratic, $y = -1$ if a person votes Republican, and $y = 0$ otherwise. The 1996 election did not have significant minor party candidates. Hence, for simplicity, we use $y = 0$ to aggregate persons who vote for minor party candidates and those who do not vote, either by choice or because they are ineligible.

In this setting, $P(y \mid x)$ is the distribution of voting outcomes in state x. $P(z \mid x)$ is the distribution of voter attributes in this state. $E(y \mid x, z)$ is the *Democratic plurality* among voters in state x who have attributes z. Let S_x denote the number of Electoral College seats held by state x. Then the Electoral College vote for the Democratic candidate, assuming away ties,

is $T \equiv \sum_{x \in X} S_x \cdot \mathbf{1}[E(y \mid x) > 0]$, where $\mathbf{1}[\cdot]$ is the indicator function. This candidate wins the election if $T > 269 = 538/2$, as 538 is the total number of Electoral College seats.

Now suppose that the composition of the population were different in 1996. Suppose that the distribution of attributes in state x were $P^*(z \mid x)$ and that the number of its Electoral College seats were S_x^*. What would be the election outcome under this scenario?

To address the question, we maintain the key assumption that $E(y \mid \cdot, \cdot)$ is *invariant*, in the sense that these conditional expectations remain unchanged under the hypothesized demographic change. This is a non-trivial assumption, but one that seems reasonable to entertain. To interpret the assumption, it may help to consider a behavioral model of the form $y = f(x, z, u)$, wherein a voter's behavior is some function f of his state x, personal attributes z, and other factors u. Then $E(y \mid \cdot, \cdot)$ is invariant if u is statistically independent of (x, z) and if the distribution of u remains unchanged under the hypothesized demographic change. Clearly, the reasonableness of this assumption depends on the specification chosen for the covariates z.

Under the assumption that $E(y \mid \cdot, \cdot)$ is invariant, the predicted Democratic plurality in state x is

$$E^*(y \mid x) \equiv \sum_{j \in Z} \Pr^*(z = j \mid x) E(y \mid x, z = j). \tag{18}$$

The predicted number of Electoral College votes for the Democratic candidate is $T^* \equiv \sum_{x \in X} S_x^* \times \mathbf{1}[E^*(y \mid x) > 0]$. This candidate would win the election if $T^* > 538/2$.

The essential point is that the quantities to be predicted, first $[E^*(y \mid x), x \in X]$ and then T^*, are functions of $E(y \mid \cdot, \cdot)$. The identification region for $E(y \mid \cdot, \cdot)$ determines the region for T^*. Thus, under the assumption that $E(y \mid \cdot, \cdot)$ is invariant, Propositions 1 through 3 provide the basis for forecasting the hypothetical election outcome.

4.3 Some forecasts

As a concrete application we forecast Democratic plurality, $E^*(y \mid x)$, and the number of Democratic Electoral College votes, T^*, under the ceteris paribus assumption for the estimated U.S. population composition in the next seven presidential election years, from 2000 to 2024. The z covariates are (age, ethnicity), with two age categories (18 to 54 years, 55 years and over) and three ethnicity categories (white, black, Hispanic). So $J = 2 \times 3 = 6$.

Data issues We use forecasts by the U.S. Bureau of the Census (Campbell, 1996) of each state's population and its demographic composition. The 1996 distribution of voting outcomes is based on data from the Federal Election Commission's web page, http://www.fec.gov.

The Census forecasts divide the population into four race categories; white, black, Asian/Pacific Islander, and American Indian/Eskimo/Aleut. Hispanic and non-Hispanic origin are also indicated. From these eight distinct race/Hispanic origin categories we obtain our ethnicity breakup as follows; *Hispanics* are all people of Hispanic origin regardless of race, *blacks* are non-Hispanic blacks, and *whites* are non-Hispanics from the remaining races.

We classify as voters all members of the voting age population (18 years and over), even though this includes legal and illegal aliens, persons in institutions, and others who do not possess voting rights. All such persons have $y = 0$ as their recorded voting outcome, the same outcome recorded by eligible voters who vote for minor party candidates or who choose not to vote. We are restricted to this classification because the Census Bureau does not publish forecasts of the population of eligible voters. This same classification is employed by the Federal Election Commission (Kimberling, 1988).

Electoral College seats are allocated to states according to each state's total population, as reported in the most recent decennial census. For example S_x, the seats in the 1996 presidential election, were allocated according to the population in each state from the 1990 Census,

not according to forecasts of 1996 state populations. Consequently, we estimate S_x^* in 2004 and 2008 from the Census Bureau's forecasts of each state's population in 2000, S_x^* in 2012, 2016 and 2020 from the 2010 population forecasts, and S_x^* in 2024 from the 2020 population forecasts. S_x^* in 2000 is equal to S_x, each state's number of Electoral College seats in 1996.

Bounds on $E^*(y \mid x)$ and T^* To obtain the bounds on $E^*(y \mid x)$, we first apply Proposition 1, Corollary 2 to determine the identification region D_x. We then use the right-hand side of equation (18) to determine the feasible values of $E^*(y \mid x)$. Proposition 4 below shows that the upper bound on $E^*(y \mid x)$ must occur at an extreme point of D_x, and the same argument applies to the lower bound. Hence, to compute the lower and upper bounds on $E^*(y \mid x)$, we do not have to evaluate the right-hand side of equation (18) at all points in D_x. It suffices to evaluate (18) at the $J!$ extreme points of D_x, which are the expectations of the vectors of stacked distributions. In our application, $J! = 6! = 120$, so this is quite tractable.

PROPOSITION 4: For each $j = 1, \ldots, J$, let π_{xj}^* denote the counterfactual value of π_{xj}. If y has finite support, then

$$\max_{\xi \in D_x} \sum_{j \in Z} \pi_{xj}^* \, \xi_j \;=\; \max_{\xi \in \{E_x^m,\, m=1,\ldots,J!\}} \sum_{j \in Z} \pi_{xj}^* \, \xi_j.$$

Proof: Since y has finite support, we know from Propostion 1, Corollary 2 that $D_x = \text{conv}\,\{E_x^m, m = 1, \ldots, J!\}$. So any $\xi \in D_x$ can be represented as $\sum_{m=1}^{J!} \alpha_m \, E_x^m$ for appropriately chosen convex combination weights α_m. Relabel the m's, if necessary, to ensure that

$$E_x^1 \in \arg \max_{m \in \{1,\ldots,J!\}} \sum_{j \in Z} \pi_{xj}^* \, E_{xj}^m.$$

We have to show that

$$E_x^1 \in \arg \max_{\xi \in D_x} \sum_{j \in Z} \pi_{xj}^* \, \xi_j.$$

We have

$$\sum_{j \in Z} \pi_{xj}^* \, E_{xj}^1 \geq \sum_{j \in Z} \pi_{xj}^* \, E_{xj}^m \quad \text{for all} \quad m \in \{1, \ldots, J!\}.$$

Since each α_m is non-negative, we have

$$\alpha_m \sum_{j \in Z} \pi_{xj}^* \, E_{xj}^1 \geq \alpha_m \sum_{j \in Z} \pi_{xj}^* \, E_{xj}^m \quad \text{for all} \quad m \in \{1, \ldots, J!\}.$$

Now since the α_m's sum to one, summing across the m's gives

$$\sum_{j \in Z} \pi^*_{xj} E^1_{xj} = \sum_{m=1}^{J!} \alpha_m \sum_{j \in Z} \pi^*_{xj} E^1_{xj} \geq \sum_{m=1}^{J!} \alpha_m \sum_{j \in Z} \pi^*_{xj} E^m_{xj} = \sum_{j \in Z} \pi^*_{xj} \sum_{m=1}^{J!} \alpha_m E^m_{xj} = \sum_{j \in Z} \pi^*_{xj} \xi_j.$$

Since ξ is an arbitrary element of D_x the proposition is proved. *Q.E.D.*

Table 1 reports the bounds on $E^*(y \,|\, x)$ in 2004 and 2020. The table shows that the bounds on $E^*(y \,|\, x)$ in 2020 are wider than those in 2004 for all states. In 2004 there are 25 states where the bound on Democratic plurality is entirely a positive interval, and 11 states where the bound is entirely a negative interval. In 2020 the corresponding number of states is five and zero, respectively.

The reason the bounds are wider in 2020 is simple. The forecast change in the distribution of demographic characteristics, $P(z \,|\, x)$, for each $x \in X$ is more pronounced between 1996 and 2020 than between 1996 and 2004. The more $P(z \,|\, x)$ varies, the less information $P(y \,|\, x)$ conveys about $E^*(y \,|\, x)$.

From the bounds on $E^*(y \,|\, x)$ in a particular state in 2004, we can predict the number of Electoral College seats the Democratic candidate will win in that state. For the 25 states where the bound on $E^*(y \,|\, x)$ is entirely a positive interval, we get the point prediction S^*_x as the number of seats won. And for the 11 states where the bound on $E^*(y \,|\, x)$ is entirely a negative interval, our point prediction of the number of seats won is zero. In the remaining 15 states the bound on $E^*(y \,|\, x)$ straddles zero, and so we obtain no prediction for the number of Electoral College seats won by the Democratic candidate. In the absence of any cross-x restrictions, we simply add these bounds, some of which reduce to a point, across all states to obtain the bound on T^*.

Table 1. Bounds on $E^*(y \mid x)$ and T^* in 2004 and 2020.

		$E(y \mid x)$ in 1996	Bound on $E^*(y \mid x)$ in 2004	Bound on $E^*(y \mid x)$ in 2020
Northeast	*New England*			
	Connecticut	0.102	[0.055 , 0.146]	[-0.053 , 0.252]
	Maine	0.134	[0.103 , 0.168]	[-0.028 , 0.309]
	Massachusetts	0.184	[0.141 , 0.223]	[0.025 , 0.346]
	New Hampshire	0.057	[0.023 , 0.093]	[-0.116 , 0.237]
	Rhode Island	0.171	[0.131 , 0.206]	[0.019 , 0.319]
	Vermont	0.130	[0.091 , 0.172]	[-0.035 , 0.308]
	Middle Atlantic			
	New Jersey	0.091	[0.048 , 0.130]	[-0.056 , 0.231]
	New York	0.134	[0.098 , 0.167]	[0.014 , 0.249]
	Pennsylvania	0.045	[0.024 , 0.066]	[-0.068 , 0.162]
Midwest	*East North Central*			
	Illinois	0.086	[0.051 , 0.120]	[-0.050 , 0.222]
	Indiana	-0.027	[-0.055 , -0.001]	[-0.158 , 0.098]
	Michigan	0.072	[0.042 , 0.104]	[-0.064 , 0.218]
	Ohio	0.035	[0.007 , 0.063]	[-0.097 , 0.171]
	Wisconsin	0.060	[0.028 , 0.091]	[-0.093 , 0.221]
	West North Central			
	Iowa	0.060	[0.034 , 0.086]	[-0.078 , 0.206]
	Kansas	-0.103	[-0.134 , -0.072]	[-0.262 , 0.046]
	Minnesota	0.104	[0.068 , 0.140]	[-0.073 , 0.290]
	Missouri	0.034	[0.013 , 0.056]	[-0.099 , 0.172]
	Nebraska	-0.105	[-0.134 , -0.077]	[-0.256 , 0.032]
	North Dakota	-0.038	[-0.065 , -0.013]	[-0.183 , 0.100]
	South Dakota	-0.021	[-0.034 , -0.008]	[-0.171 , 0.125]
South	*South Atlantic*			
	Delaware	0.076	[0.041 , 0.110]	[-0.079 , 0.237]
	District of Columbia	0.331	[0.299 , 0.364]	[0.253 , 0.401]
	Florida	0.028	[-0.025 , 0.078]	[-0.157 , 0.209]
	Georgia	-0.005	[-0.044 , 0.033]	[-0.166 , 0.155]
	Maryland	0.075	[0.030 , 0.117]	[-0.085 , 0.233]
	North Carolina	-0.022	[-0.060 , 0.016]	[-0.183 , 0.135]
	South Carolina	-0.024	[-0.066 , 0.016]	[-0.175 , 0.120]
	Virginia	-0.009	[-0.058 , 0.039]	[-0.172 , 0.152]
	West Virginia	0.066	[0.035 , 0.101]	[-0.057 , 0.203]
	East South Central			
	Alabama	-0.033	[-0.067 , -0.001]	[-0.176 , 0.102]
	Kentucky	0.005	[-0.028 , 0.038]	[-0.143 , 0.153]
	Mississippi	-0.023	[-0.055 , 0.008]	[-0.164 , 0.113]
	Tennessee	0.011	[-0.020 , 0.043]	[-0.135 , 0.160]
	West South Central			
	Arkansas	0.080	[0.048 , 0.117]	[-0.061 , 0.239]
	Louisiana	0.069	[0.022 , 0.117]	[-0.104 , 0.247]
	Oklahoma	-0.039	[-0.079 , 0.000]	[-0.191 , 0.109]
	Texas	-0.020	[-0.058 , 0.018]	[-0.168 , 0.128]
West	*Mountain*			
	Arizona	0.010	[-0.039 , 0.059]	[-0.166 , 0.186]
	Colorado	-0.007	[-0.068 , 0.053]	[-0.224 , 0.208]
	Idaho	-0.108	[-0.159 , -0.060]	[-0.320 , 0.080]
	Montana	-0.018	[-0.070 , 0.033]	[-0.231 , 0.190]
	Nevada	0.004	[-0.053 , 0.061]	[-0.186 , 0.194]
	New Mexico	0.034	[-0.001 , 0.068]	[-0.116 , 0.184]
	Utah	-0.106	[-0.145 , -0.074]	[-0.284 , 0.046]
	Wyoming	-0.078	[-0.133 , -0.028]	[-0.284 , 0.109]
	Pacific			
	Alaska	-0.100	[-0.155 , -0.039]	[-0.232 , 0.050]
	California	0.057	[0.003 , 0.114]	[-0.085 , 0.200]
	Hawaii	0.103	[0.079 , 0.130]	[0.028 , 0.189]
	Oregon	0.047	[-0.014 , 0.110]	[-0.156 , 0.260]
	Washington	0.069	[0.017 , 0.125]	[-0.116 , 0.272]
Democratic Electoral College votes, T^*		379	[302 , 477]	[51 , 538]

The first column of Table 2 shows the bounds on T^* for all seven election years. Observe that the bounds continually widen as we forecast further into the future. In 2000 and 2004, the bounds are tight enough to predict a Democratic election winner, since they are intervals lying entirely above 270. In contrast, the bound in 2024 conveys little information.

Table 2. Bounds on T^* in 2000 through 2024.

	No exclusion restriction	Exclusion restriction
2000	[359 , 413]	402
2004	[302 , 477]	399
2008	[193 , 514]	407
2012	[85 , 521]	361
2016	[55 , 538]	356
2020	[51 , 538]	351
2024	[18 , 538]	348

D_x versus C_x There are substantial gains from employing the sharp identification region D_x to bound $E^*(y \mid x)$ rather than the non-sharp rectangular region C_x discussed in Section 2.1. Consider, for example, the state of California. For this state, C_x is the cross product of the six j-specific bounds shown in Table 3.

Table 3. Bounds on $E(y \mid x = \text{California}, z = j)$ for each $j \in Z$.

j	$\Pr(z = j \mid x)$	Bounds on $E(y \mid x, z = j)$
White, 18 to 54 years	0.476	[-0.352 , 0.471]
White, 55 years and over	0.196	[-0.853 , 1]
Black, 18 to 54 years	0.052	[-1 , 1]
Black, 55 years and over	0.014	[-1 , 1]
Hispanic, 18 to 54 years	0.223	[-0.750 , 1]
Hispanic, 55 years and over	0.039	[-1 , 1]

Using this set C_x to bound Democratic plurality in California in 2004 yields $[-0.626, 0.771]$, considerably wider than $[0.003, 0.114]$, the bound based on D_x reported in Table 1.

The bound calculated from C_x lacks informativeness not only because of its width, but also because it straddles zero. In fact, in each of the seven election years, the bound on $E^*(y \mid x)$ obtained from C_x straddles zero in every state. If the objective is to estimate the number of Electoral College seats won by the Democratic candidate, then the bound on T^* obtained from C_x for all seven election years is $[0, 538]$, completely uninformative. This stands in sharp constrast to the bounds on T^* obtained from D_x, reported in Table 2.

Exclusion restrictions Consider x as a pair (v, w), with w indicating region of the U.S. , and v indicating the state within that region. To illustrate the identifying power of exclusion restrictions, let us now suppose that Democratic plurality, conditional on z, does not vary between states in the same region, but may vary across regions. That is, assume

$$E(y \mid v, w, z) = E(y \mid w, z) \quad \text{for all } v \in V, w \in W, \text{ and } z \in Z. \tag{19}$$

Each of the four regions of the U.S. contains more than six states. Hence the rank condition of Proposition 3, Corollary 1 implies point identification of $[E(y \mid w, z = j), j \in Z]$ if equation (17) has a unique solution, and implies that the exclusion restriction (19) is incorrect if equation (17) has no solution. We find that the exclusion restriction is rejected for all four regions.

Goodman (1953) was aware that an exclusion restriction may be rejected but, wishing to retain the restriction in an approximate form, suggested a least squares fit of equation (17). In our application, such a fit yields a point estimate of Democratic plurality in each state in a particular region. This yields a point estimate of the Electoral College seats won by the Democratic candididate in each election year. These estimates are reported in the second column of Table 2.

Notice that these point estimates for T^* lie within the previously calculated bounds in each

election year. However, this is not the case for the point estimates of $E^*(y \mid x)$. In 25 states the predicted Democratic plurality in 2020 under the least squares approximation to (19) lies outside the bound reported in Table 1. For three of these states, the prediction lies outside $[-1, 1]$, which is nonsensical. Further, in every region w, the estimates of $E(y \mid w, z = j)$ lie outside $[-1, 1]$ for several values of j. Such problems are common in applications of this *Goodman regression* approach to ecological inference (see King, 1997).

Clearly, the data reject assumption (19) when w indicates one of the four regions of the U.S. We have also considered a weaker version of this assumption, in which w indicates one of the nine sub-regions shown in Table 1. The data also reject this weaker exclusion restriction, in which the Democratic plurality conditional on z is assumed constant only across states within a sub-region.

Of course, exclusion restrictions are not the only form of assumption that an empirical researcher may wish to bring to bear. One may, for example, wish to impose upon the long regression a *monotone instrumental variable* assumption, in which the equalities defining exclusion restrictions (13) and (14) are replaced with weak inequalities (see Manski and Pepper, 2000). Regardless of what assumptions one may wish to entertain, we believe that determination of the identification region for $E(y \mid \cdot, \cdot)$ using the data alone forms a natural starting point for empirical analysis.

References

Brøndsted, A. (1983), *An Introduction to Convex Polytopes*, New York: Springer-Verlag.

Campbell, P. (1996), *Population Projections for States by Age, Sex, Race, and Hispanic Origin: 1995 to 2025*, U.S. Bureau of the Census, Population Division, PPL-47.

Duncan, O. and B. Davis (1953), "An Alternative to Ecological Correlation," *American Sociological Review*, 18, 665-666.

Freedman, D., S. Klein, J. Sacks, C. Smyth, and C. Everett (1991), "Ecological Regression and Voting Rights," *Evaluation Review*, 15, 673-711.

Freedman, D., S. Klein, M. Ostland, and M. Roberts (1998), "Review of A Solution to the Ecological Inference Problem, by G. King," *Journal of the American Statistical Association*, 93, 1518-1522.

Freedman, D., S. Klein, M. Ostland, and M. Roberts (1998), "Response to King's Comment," *Journal of the American Statistical Association*, 94, 355-357.

Goldberger, A. (1991), *A Course in Econometrics*, Cambridge, Mass.: Harvard University Press.

Goodman, L. (1953), "Ecological Regressions and Behavior of Individuals," *American Sociological Review*, 18, 663-664.

Horowitz, J. and C. Manski (1995), "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281-302.

Huber, P. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73-101.

Kimberling, W. (1988), "Voting for President: Participation in America," *The FEC Journal of Election Administration*, 15, 21-28.

King, G. (1997), *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*, Princeton: Princeton University Press.

King, G. (1999), "The Future of Ecological Inference Research: A Comment on Freedman et.al.," *Journal of the American Statistical Association*, 94, 352-355.

Lindley, D. and M. Novick (1981), "The Role of Exchangeability in Inference," *Annals of Statistics*, 9, 45-58.

Manski, C. and J. Pepper (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, forthcoming.

Robinson, W. (1950), "Ecological Correlation and the Behavior of Individuals," *American Sociological Review*, 15, 351-357.

Simpson, E. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society*, 13, 238-241.

Zidek, J. (1984), "Maximal Simpson-disaggregations of 2×2 Tables," *Biometrika*, 71, 187-190.