Characterizing Selection Bias Using Experimental Data

James Heckman
University of Chicago

Hidehiko Ichimura
University of College London

Jeffrey Smith
University of Michigan

Petra Todd
University of Pennsylvania

Based on Econometrica 1998 Article

Econ 312, Spring 2023

2.0 The Evaluation Problem, the Parameter of Interest in this Paper and How Randomization Estimates It

(1) The Model:
Two possible outcomes: Y_0 and Y_1.
$D = 1$ treatment, $D = 0$ its absence.

$$Y = DY_1 + (1 - D)Y_0.$$

$$\Pr(D = 1 \mid X) = P(X).$$

Parameters of Interest Considered Today

$$\begin{aligned} \Delta(X) &= E(\Delta \mid X, D = 1) && (1) \\ &= E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 1) \end{aligned}$$

or

$$\bar{\Delta}(K) = \int_K \Delta(X) dF(X \mid D = 1) \Big/ \int_K dF(X \mid D = 1). \quad (2)$$

Method of Comparison Groups:

Assumes

$$E(Y_0 \mid X, D=1) \doteq E(Y_0 \mid X, D=0)$$

Selection bias $B(X)$ for $E(\Delta \mid X, D=1)$:

$$B(X) = E(Y_0 \mid X, D=1) - E(Y_0 \mid X, D=0). \qquad (3)$$

3.0 Characterizing Selection Bias

3.1 The Method of Matching

(A-1)

$$Y_0 \perp\!\!\!\perp D \mid X, \qquad X \in \chi_c,$$

$$E(Y_0 \mid X, D=1) = E(Y_0 \mid X, D=0) \tag{4}$$

$$Y_{1i} - \sum_{j \in \{D=0\}} W_{N_0 N_1}(i,j) Y_{0j} \tag{5}$$

$$\sum_{j\in\{D=0\}} W_{N_0N_1}(i,j) = 1 \text{ for all } i.$$

Persons matched to i are in A_i

$$A_i = \{j \in \{D=0\} \mid X_j \in C(X_i)\}.$$

Nearest neighbor matching

$$C(X_i) = \min_j \|X_i - X_j\|,\ j \in \{D = 0\},$$

$$W_{N_0 N_1}(i,j) = 1,\ j \in A_i$$

and $W_{N_0 N_1}(i,j) = 0$ otherwise.

Caliper matching:

$$C(X_i) = \{X_j \mid \|X_i - X_j\| < \varepsilon\}$$

Kernel matching:

$$W_{N_0 N_1}(i,j) = \frac{G_{ij}}{\sum\limits_{k\in\{D=0\}} G_{ik}}$$

$$G_{ik} = G((X_i - X_k)/a_{N_0}), \quad \lim_{N_0\to\infty} a_{N_0} = 0$$

$$\hat{M}(K) \;=\; \sum_{i\in\{D=1\}} \omega_{N_0 N_1}(i)[Y_{1i} - \sum_{j\in\{D=0\}} W_{N_0 N_1}(i,j) Y_{0j}] \text{ for } X_i \in K$$

Rosenbaum and Rubin (1983)

(A-1) $$Y_0 \perp\!\!\!\perp D \mid P(X) \text{ for } X \in \chi_c,$$

(A-2) $$0 < P(X) < 1 \text{ for } X \in \chi_c,$$

$$E(Y_0 \mid P(X), D = 1) - E(Y_0 \mid P(X), D = 0) = B(P(X)) = 0. \tag{7}$$

3.3 Difference-in-Differences

$$B_t(X) - B_{t'}(X) = 0 \text{ for some } t, t' \tag{8}$$

4.0 Re-examining the Conventional Measure of Selection Bias

$S_{1X} = \{X \mid f(X \mid D = 1) > 0\}$ support of X for $D = 1$.

$S_{0X} = \{X \mid f(X \mid D = 0) > 0\}$ the support of X for $D = 0$

$S_X = S_{0X} \cap S_{1X}$ region of overlap.

$$\overline{B}_{S_X} = \frac{\displaystyle\int_{S_X} B(X) dF(X \mid D = 1)}{\displaystyle\int_{S_X} dF(X \mid D = 1)}.$$

Conventional measure of selection bias:

$$B = E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0).$$

Least squares regression of Y_0 on D with

$$Y_0 = \pi_0 + \pi_1 D + \tau,$$

$$E(\tau) = 0$$

$$\text{plim}\hat{\pi}_1 = B.$$

$$B = \int_{S_{1,X}} E(Y_0 \mid X, D = 1) dF(X \mid D = 1) - \int_{S_{0,X}} E(Y_0 \mid X, D = 0) dF(X \mid D = 0). \tag{9}$$

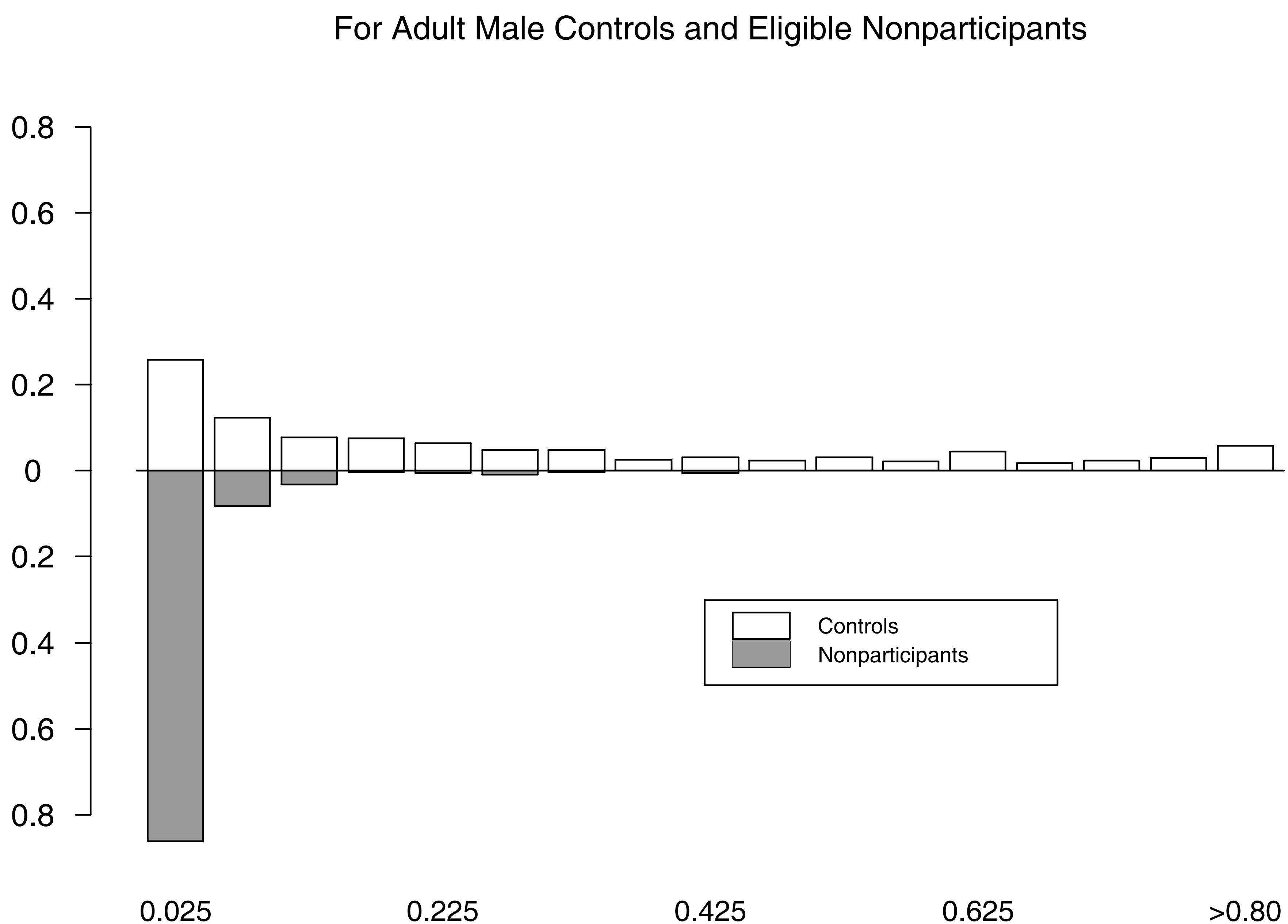
$$\textbf{Decompose B: } B = B_1 + B_2 + B_3, \tag{10}$$

$$\begin{aligned}
B_1 &= \int_{S_{1X} \setminus S_X} E(Y_0 \mid X, D=1) dF(X \mid D=1) \\
&\quad - \int_{S_{0X} \setminus S_X} E\left(Y_0 \mid X, D=0\right) dF\left(X \mid D=0\right) \\
B_2 &= \int_{S_X} E(Y_0 \mid X, D=0)[dF(X \mid D=1) - dF(X \mid D=0)] \\
B_3 &= P_X \bar{B}_{S_X}
\end{aligned}$$

$$P_X = \int_{S_X} dF(X \mid D = 1)$$

$\bar{B}_{S_X}$ is the selection bias

Figure 2: Density of Estimated Probability of Program Participation



4.2 Our Data

- We use comparison group (nonexperimental) and experimental control group
- Neither sample receives treatment

TABLE 1

DEFINITION OF VARIABLES

Variable Name	Description
Training Center: Corpus Christi, Fort Wayne, Jersey City, Providence.	Indicator variables for the geographic location of the individual.
Race and Ethnicity: black, white, Hispanic.	Indicator variables for the race/ethnicity of the individual. Individuals who reported Asian or ‘‘other’’ were included in the Hispanic category in R but not in Z.
Age: age 22-29, age 30-39, age 40-49, age 50-54.	Indicator variables for the age of the individual calculated using the average age in years of the individual within the quarter of the observation.
Education: less than 10th grade, 10-11th grade, 12th grade, 1-3 years college, 4 or more years of college.	Indicator variables for the educational attainment of the individual at the time of random assignment or eligibility determination. Missing values are imputed.*
Marital Status: currently married, last married 1-12 months before RA/EL, last married >12 months before RA/EL, single, never married at RA/EL.	Indicator variables for marital status at the time of random assignment or eligibility determination (RA/EL). Missing values are imputed.*
Children less than 6 years of age	Indicator variable for the presence of young children in the household at the time of the baseline interview. Missing values are imputed*.
Calendar Quarter: quarter 1, quarter 2, quarter 3, quarter 4.	Indicator variables for the calendar quarter for the observations. Quarter 1 refers to January, Febuary, and March etc. If an observation overlaps two quarters, then the variable takes on fractional values.
Calendar Year: year 1987, year 1988, year 1989, year 1990.	Indicator variables for the calendar year of the observation. If the observation overlaps two years, then the year indicators take on fractional values.
Local Unemployment Rate (sources: U.S. Department of Labor’s publication ‘‘Labor Force, Employment, and Unemployment Estimates for States, Labor Market Areas, Counties, and Selected Cities’’ for the years 1986-1991 provide the unemployment rates. Population weights are obtained from annual total population data available in the U.S. Department of Commerce’s Regional Economic Information System (REIS)).	This variable gives the monthly unemployment rate. The data is published at the county and metropolitan levels. We calculate the unemployment rate as a population-weighted average of the unemployment rates of the counties and metropolitan areas served by each of the four training centers in the JTPA data.

TABLE 1 (continued)

DEFINITION OF VARIABLES

Variable Name	Description
Labor Force Status Transition: employed -> employed, unemployed -> employed, OLF -> employed, employed -> unemployed, unemployed -> unemployed, OLF -> unemployed, employed -> OLF, unemployed -> OLF, OLF -> OLF.	The two most recent labor force statuses during the period composed of the month of random assignment or eligibility determination and the six preceding months define a set of nine labor force status patterns. In each case, the second status is that in the month of random assignment or eligibility determination and the first status (if different) is the most recent preceding status. Repeated patterns such as "employed -> employed" indicate persons in the same labor force status for all seven months. Missing values are imputed.*
Number of Persons in the Household	Continuous variable indicating the number of persons in the individual's household as of the baseline interview. Missing values are imputed.*
Earnings in the Month of Random Assignment or Eligibility Determination	Self-reported monthly earnings in the month of random assignment or eligibility determination from the baseline survey. Persons for whom the survey covers only a part of the month have their responses scaled up to a full month.
Ever had Vocational Training	Indicator variable for whether the respondent ever had vocational or technical training as of the baseline interview date, excluding courses taken while in high school. Missing values are imputed. *
Currently Receiving Vocational Training	Indicator variable for current receipt of vocational or technical training as of the baseline interview. Excludes courses taken in high school. Missing values are imputed. *
Number of Job Spells in the 18 Months Prior to Random Assignment or Eligibility Determination: zero, one, two, more than two.	Categories for the number of full or partial job (not employment) spells experienced during the 18 months prior to random assignment or eligibility determination. Missing values are imputed.*
Work Experience	Continuous variable indicating months of work experience prior to random assignment or eligibility determination. It is calculated using the Mincer method, (age-education-6)*12, for the period prior to our data, adding in actual experience in months for the five years prior to RA/EL.

* An appendix available upon request from the authors describes the imputation procedure for these variables.

4.3 No Good Way of Determining the Probability of Program Participation P

A. Minimization of classification error when $\widehat{P}(X) > P_C$ is used to predict $D = 1$ and $\widehat{P}(X) \leq P_C$ is used to predict $D = 0$, where $P_C = E(D)$; and

B. Statistical significance: For adult males, the two criteria produce the same model.

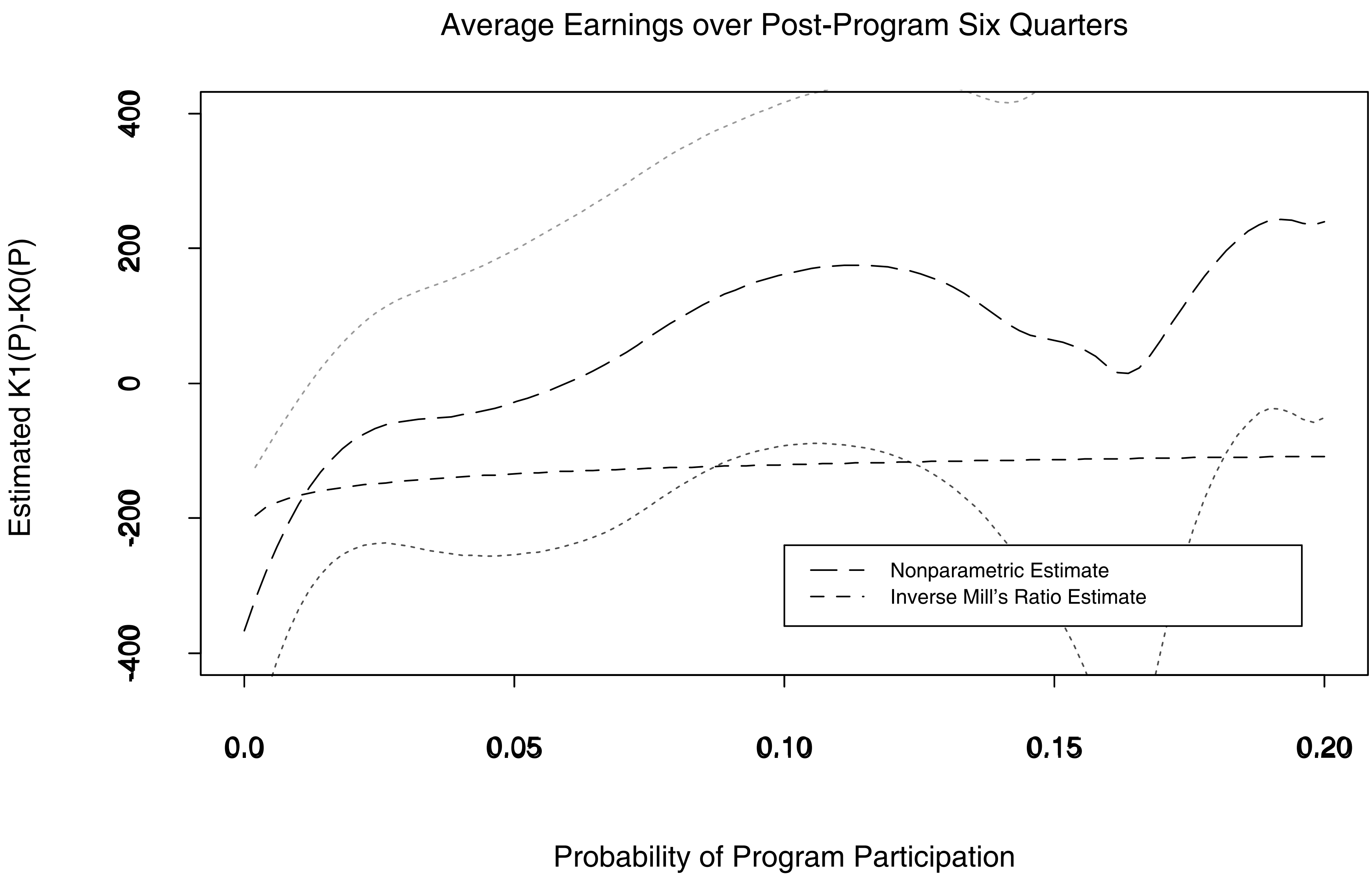
C. But as noted by Heckman and Navarro (2004) these methods are not guaranteed to pick the right model except under exogeneity conditions.

D. In general, no guide to determine the choice of X, which variables to use? Kitchen sink is usually recommended by statisticians.

E. Danger in this approach: A good predictor of D could also be correlated with U_1 and U_0 creating endogeneity problems

Consider bias in estimating models using comparison groups (compare controls with a nonexperimental comparison group).

Figure 3: Local Linear Regression Estimates of Pointwise Bias (B(P))
Adult Males, Best Predictor P Model for The Probability of Program Participation



4.4 Estimating the Components of Our Decomposition of B

$$\widehat{B} = \hat{E}(Y_0 \mid D = 1) - \hat{E}(Y_0 \mid D = 0) = \hat{B}_1 + \hat{B}_2 + \hat{B}_3 \quad (11)$$

$$\hat{B}_1 = \frac{1}{N_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_{1P} \setminus S_P}} Y_0(P_i) - \frac{1}{N_0} \sum_{\substack{i \in \{D=0\} \\ P_i \in S_{0P} \setminus S_P}} Y_0(P_i)$$

$$\hat{B}_2 = \frac{1}{N_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_P}} \hat{E}(Y_{0i} \mid D_i = 0, P_i) - \frac{1}{N_0} \sum_{\substack{i \in \{D=0\} \\ P_i \in S_P}} Y_0(P_i)$$

$$\hat{B}_3 = \frac{1}{N_1} \sum_{\substack{i \in \{D=1\} \\ P_i \in S_P}} [Y_0(P_i) - \hat{E}(Y_{0i} \mid D_i = 0, P_i)]$$

$Y_0(P_i)$ is value of Y_{0i} for person i with probability P_i, S_P.
Evaluate over regions $S_{1P} \backslash S_P$, $S_{0P} \backslash S_P$.

4.5 Estimates of the Components of B

TABLE 3

DECOMPOSITION OF MEAN SELECTION BIAS FOR THE BEST PREDICTOR MODEL FOR THE PROBABILITY OF PROGRAM PARTICIPATION

Experimental Control and Elig. Nonparticipant (ENP) Samples
Adult Males, 508 Controls and 388 ENPs

Quarter	(1) Mean Difference ($\hat{B}$)	(2) Non-overlap Support ($\hat{B}_1$)	(3) Density Weighting ($\hat{B}_2$)	(4) Selection Bias ($\hat{B}_3$)	(5) Average Bias ($\hat{\bar{B}}_{S_P}$)	(6) Experimental Treatment Impact	(7) Average Bias ($\hat{\bar{B}}_{S_P}$) as a % of Treatment Impact
Qtr1	-420	190[-45%]	-627[149%]	17[-4%]	29	5	566%
	(38)	(31)	(32)	(34)	(63)	(30)	
Qtr2	-352	209[-59%]	-581[165%]	19[-6%]	32	37	88%
	(47)	(41)	(45)	(35)	(65)	(33)	
Qtr3	-343	221[-65%]	-576[168%]	12[-3%]	20	57	35%
	(55)	(39)	(50)	(43)	(79)	(34)	
Qtr4	-294	234[-80%]	-568[194%]	41[-14%]	68	60	114%
	(57)	(40)	(46)	(42)	(79)	(34)	
Qtr5	-311	232[-75%]	-576[185%]	33[-10%]	54	44	121%
	(57)	(40)	(51)	(41)	(77)	(35)	
Qtr6	-334	223[-67%]	-573[172%]	16[-5%]	27	61	44%
	(63)	(45)	(51)	(44)	(81)	(34)	
Average of 1 to 6	-342	218[-64%]	-584[170%]	23[-7%]	38	44	87%
	(47)	(38)	(41)	(33)	(63)	(14)	

Appendix

5.0 Estimating the Form of the Selection Bias B(X)

$$Y_0 = X\beta + U_0,$$

$$E(Y_0 \mid X, D = 1) = X\beta + E(U_0 \mid X, D = 1)$$

$$E(Y_0 \mid X, D = 0) = X\beta + E(U_0 \mid X, D = 0)$$

$$Y_0 = X\beta + E(U_0 \mid X, D = 0) + B(X)D + \varepsilon \qquad (12)$$

$$\begin{aligned} B(X) &= E(U_0 \mid X, D = 1) - E(U_0 \mid X, D = 0) \\ E(\varepsilon &\mid X, D) = 0. \end{aligned}$$

Bias functions

$$K_{1t}(P_i) = E(U_{0it} \mid D = 1, P_i)$$

$$K_{0t}(P_i) = E(U_{0it} \mid D = 0, P_i),$$

define

$$\varepsilon_{it} = U_{0it} - D_i K_{1t}(P_i) - (1 - D_i) K_{0t}(P_i)$$

where $E(U_{0it}) = 0$.

Define

$$Y_i = (Y_{i1}, ..., Y_{iT}), X_i = (X_{i1}, ..., X_{iT})', K_j(P_i) = (K_{j1}(P_i), ..., K_{jT}(P_i))',$$

$$Y_i = X_i\beta + D_i K_1(P_i) + (1 - D)K_0(P_i) + \varepsilon_i. \tag{13}$$

$$Y_i - E(Y_i \mid P_i, D_i) = [X_i - E(X_i \mid P_i, D_i)]'\beta + \varepsilon_i. \qquad (14)$$

$$\arg\min_{K_j,\gamma_j} \sum_{i\in\{D=d\}} [c_i - K_j(P_0) - \gamma_j(P_0)(\hat{P}_i - P_0)]^2 G\left(\frac{\hat{P}_i - P_0}{a_N}\right), \qquad (15)$$

$d \in \{0,1\}$ $\{a_N\}$ is a sequence of smoothing parameters

$$W_{N_0N_1}(i,j) = \frac{G_{ij}^2 \sum_{k\in I_0} G_{ik}(P_k - P_i)^2 - [G_{ij}(P_j - P_k)][\sum_{k\in I_0} G_{ik}(P_k - P_i)]}{\sum_{j\in I_0} G_{ij}^2 \sum_{k\in I_0} G_{ij}(P_k - P_i)^2 - \left(\sum_{k\in I_0} G_{ik}(P_k - P_i)^2\right)} \tag{16}$$

$$G_{ik} = G\left(\frac{P_k - P_i}{a_N}\right)$$

Comparisons Using Alternative Estimators

TABLE 18A

COMPARISON OF ESTIMATED MEAN BIAS
UNDER ALTERNATIVE ESTIMATORS OF MEAN PROGRAM IMPACTS †

Quarterly Earnings Expressed in Monthly Dollars
Adult Male, 508 Experimental Controls and 388 Elig. Non-participants

Quarter	Difference in Means	Nearest Neighbor w/o Common Support	Nearest Neighbor w/ Common Support	Local Linear Matching	Regression Adjusted Local Linear Matching
	(1)††	(2)	(3)	(4)	(5)
Qtr1	-418 (38)	221 (56)	123 (67)	33 (59)	39 (60)
Qtr2	-349 (47)	-166 (151)	77 (83)	37 (61)	39 (64)
Qtr3	-337 (55)	-58 (206)	53 (96)	29 (78)	21 (80)
Qtr4	-286 (57)	161 (178)	86 (96)	80 (77)	65 (82)
Qtr5	-305 (57)	167 (196)	87 (100)	64 (77)	50 (83)
Qtr6	-328 (63)	45 (191)	34 (113)	37 (82)	17 (90)
Average of 1 to 6	-337 (47)	62 (127)	77 (80)	47 (60)	39 (64)
As a % of impact	775%	142%	176%	107%	88%

TABLE 18B

COMPARISON OF ESTIMATED MEAN BIAS UNDER ALTERNATIVE ESTIMATORS OF MEAN PROGRAM IMPACTS†

Quarterly Earnings Expressed in Monthly Dollars
Adult Male, 508 Experimental Controls and 388 Elig. Non-participants

Quarter	Difference-in-Differences w/o Common Support	Conditional on *P* Difference-in-Differences w/ Common Support	Regression-Adjusted Conditional on *P* Difference-in-Differences w/ Common Support
	(1) ††	(2)	(3)
Qtr1	172(42)	97(62)	104(63)
Qtr2	142(47)	77(89)	77(92)
Qtr3	41(56)	90(114)	74(114)
Qtr4	43(61)	112(90)	98(91)
Qtr5	-54(63)	19(95)	-5(99)
Qtr6	-111(64)	4(105)	-35(111)
Average of 1 to 6	39 (47)	67 (71)	52 (74)
As a % of impact	89%	153%	120%

TABLE 18C

COMPARISON OF ESTIMATED MEAN BIAS UNDER ALTERNATIVE ESTIMATORS OF MEAN PROGRAM IMPACTS †

Quarterly Earnings Expressed in Monthly Dollars
Adult Males, 508 Experimental Controls and 388 Elig. Nonparticipants

Quarter	Inverse Mills' Ratio w/o Common Support w/o Density Weighting (1)††	Inverse Mills' Ratio w/ Common Support w/o Density Weighting (2)	Inverse Mills' Ratio w/ Common Support w/ Density Weighting (3)
Qtr1	-610 (86)	-619 (161)	-147 (176)
Qtr2	-514 (95)	-403 (194)	3 (220)
Qtr3	-497 (96)	-365 (190)	30 (215)
Qtr4	-494 (97)	-421 (191)	-80 (215)
Qtr5	-510 (98)	-441 (190)	-69 (215)
Qtr6	-498 (102)	-323 (196)	48 (222)
Average of 1 to 6	-521 (86)	-553 (161)	-36 (37)
As a % of impact	1198%	985%	83%

Evidence from a Geographic Mismatch

TABLE 9

EFFECT OF GEOGRAPHY ON ESTIMATED BIAS
COMPARING CONTROLS AT TWO SITES TO ELIGIBLE NON-PARTICIPANTS AT TWO SITES

Earnings in the 18 Months After Random Assignment
Quarterly Earnings Expressed in Monthly Dollars

Elig. Nonparticipant (ENP) Sample at Corpus Christi and Fort Wayne
Experimental Control Sample at Jersey City and Providence
Adult Males, 149 Controls and 276 ENPs

Quarter	Difference in Means B	Local Linear Matching $\bar{B}_{S_P}$	Regression Adjusted Local Linear Matching $\bar{B}_{S_P}(adj)$	Difference-in differences for Local Linear Matching	Difference-in-differences for Regression Adjusted Local Linear Matching
Qtr1	-534(53)	-203(85)	-184(110)	-143(111)	-135(126)
Qtr2	-504(73)	-166(107)	-154(120)	-125(118)	-72(130)
Qtr3	-515(78)	-177(120)	-147(127)	-73(131)	-9(141)
Qtr4	-485(78)	-200(121)	-164(132)	-87(141)	19(151)
Qtr5	-527(72)	-272(127)	-211(132)	-254(160)	-136(167)
Qtr6	-524(75)	-281(110)	-189(112)	-257(162)	-82(165)
Average of 1 to 6	-515(63)	-216(95)	-175(108)	-157(110)	-69(123)
As a % of impact	1183%	497%	402%	360%	159%