II. MATCHMAKER, MATCHMAKER

Josh Angrist
MIT 14.387

Fall 2014

Agenda

- Matching. What could be simpler? We look for causal effects by comparing treatment and control within subgroups where everything . . . or most things . . . or the things that matter most . . .

  are held fixed

- *There are many ways to hold fast and hold fixed: Full covariate matching, propensity-score methods, and, my favorite: regression, the ultimate matchmaker*
- *What matters? The OVB formula suggests an answer*
- *It's usually the controls that matter, not the nitty gritty 'metrics of how you use 'em*

- I'll illustrate this through two examples: military service and job training

Background: *Volunteers of America*!

- The volunteer military is the largest single employer of young men and women in the United States
- Between 1989 and 1992, enlistments by men and women without prior military service fell by 27 percent
- Enlistments by white men declined by 25 percent while enlistments by black men, the group hardest hit by military downsizing, declined by 47 percent (Angrist, 1993a)
- The major avenue used to effect these declines was an increase in applicant test-score cutoffs and other changes in entry standards
- I asked: *What were the consequences of military service for recruits?*
  - Answering this, we learn whether military downsizing constitutes a lost economic opportunity (as many believed at the time)
  - Selection bias makes comparisons by veteran status misleading (Seltzer and Jablon, 1974)

Angrist (1998) Matching Strategy

1. Compare veteran and nonveteran applicants (only half of qualified applicants serve)
2. Control for the characteristics the military uses to screen soldiers

- The matching estimand is an average of contrasts or comparisons across cells defined by covariates
- Focus on effects of treatment on the treated (TOT):

$$E[Y_{1i} - Y_{0i} | D_i = 1]$$

  This tells us the difference between the average observed earnings of soldiers, $E[Y_{1i}|D_i = 1]$, and the counterfactual average if they had not served, $E[Y_{0i}|D_i = 1]$
- The earnings differential by veteran status is a biased measure of TOT, unless D_i is independent of Y_{0i}:

$$\begin{aligned} & E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ = \; & E[Y_{1i} - Y_{0i}|D_i = 1] + \{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\} \end{aligned}$$

The Conditional Independence Assumption (CIA)

Conditional on observed characteristics, X_i, treatment is as good as randomly assigned:

$$\{\mathrm{Y}_{0i}, \mathrm{Y}_{1i}\} \amalg \mathrm{D}_i | \mathrm{X}_i$$

- Given the CIA, causal effects can be constructed by iterating expectations over X_i:

$$\begin{aligned} \delta_{TOT} &\equiv E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{D}_i = 1] \\ &= E\{E[\mathrm{Y}_{1i} | \mathrm{X}_i, \mathrm{D}_i = 1] - E[\mathrm{Y}_{0i} | \mathrm{X}_i, \mathrm{D}_i = 1] | \mathrm{D}_i = 1\} \end{aligned}$$

- $E[\mathrm{Y}_{0i} | \mathrm{X}_i, \mathrm{D}_i = 1]$ is counterfactual, but, *by virtue of the CIA*:

$$\begin{aligned} \delta_{TOT} &= E\left\{E\left[\mathrm{Y}_{1i} | \mathrm{X}_i, \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i} | \mathrm{X}_i, \mathrm{D}_i = 0\right] | \mathrm{D}_i = 1\right\} \quad (1) \\ &= E[\delta_X | \mathrm{D}_i = 1], \end{aligned}$$

where

$$\delta_X \equiv E[\mathrm{Y}_i | \mathrm{X}_i, \mathrm{D}_i = 1] - E[\mathrm{Y}_i | \mathrm{X}_i, \mathrm{D}_i = 0],$$

is the (random) X-specific difference in mean earnings by veteran status at each value of X_i

Angrist (1998) Details and Results

- Angrist (1998) constructs the sample analog of the right-hand-side of (1) for discrete covs:

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{D}_i = 1] = \sum_x \delta_x P(\mathrm{X}_i = x | \mathrm{D}_i = 1), \quad (2)$$

  where $P(\mathrm{X}_i = x | \mathrm{D}_i = 1)$ is the dsn of X_i for vets
- Hats are donned by replacing δ_X with the sample veteran-nonveteran earnings difference in each cell, and weighting by the empirical $P(\mathrm{X}_i = x | \mathrm{D}_i = 1)$
  - White veterans earn more than nonveterans, but this effect becomes negative once covariates are matched away
  - Non-white veterans earn much more than nonveterans, but controlling for covariates reduces this considerably
- Angrist (1998) **tables and figures**

TABLE I

APPLICANT POPULATION AND SAMPLE

Race	Application Year 1976	1977	1978	1979	1980	1981	1982
A. *Population*[a]							
White	339.5	286.9	235.9	253.1	348.6	387.3	309.8
Percent veteran[b]	53	52	54	55	53	49	52
Nonwhite	128.6	114.8	103.6	119.5	134.3	149.3	112.5
Percent veteran	44	46	50	46	41	36	43
B. *Sample*[c]							
White	49.2	46.5	40.0	39.4	52.9	57.9	47.3
Percent veteran	56	53	55	57	54	50	53
Nonwhite	50.9	48.1	44.6	51.9	57.0	63.7	48.7
Percent veteran	49	49	52	49	44	38	45

[a] The population is as in Angrist (1993a, Table 4), excluding those with less than a 9th grade education at the time of application. Numbers reported are thousands.

[b] Veterans are applicants identified as entrants to the military within two years following application.

[c] Approximately 90 percent of the sample is self-weighting, conditional on race.

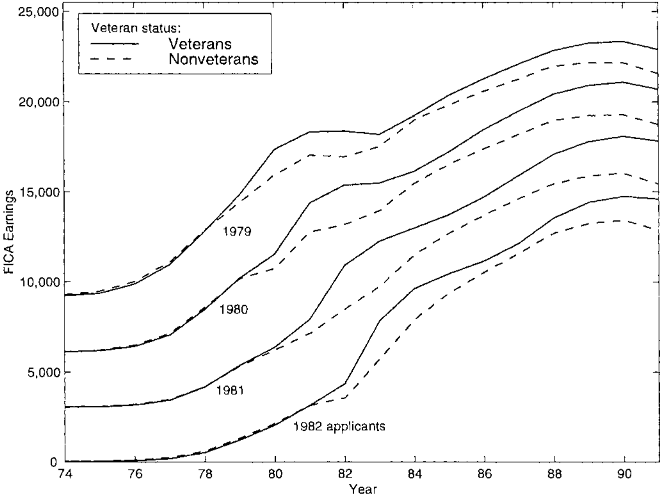

FIGURE 2.—Earnings profiles by veteran status and application year for men who applied 1979–82, with AFQT scores in categories III and IV. The plot shows the actual earnings of men who applied in 1982, earnings + \$3,000 for men who applied in 1981, earnings + \$6,000 for men who applied in 1980, and earnings + \$9,000 for men who applied in 1979.

TABLE II

Alternative Estimates of the Effects of Military Service

	Whites				Nonwhites			
Year	Mean (1)	Difference in Means[c] (2)	Controlled Contrast (3)	Regression Estimates (4)	Mean (5)	Difference in Means (6)	Controlled Contrast (7)	Regression Estimates (8)
A. *Earnings*[a]								
74	182.7	−26.1	−14.0	−13.0	157.2	−4.9	−2.0	−3.9
		(7.0)	(9.2)	(9.4)		(4.4)	(6.0)	(5.8)
75	237.9	−41.4	−14.2	−12.0	216.9	−.6	−17.1	−15.2
		(6.3)	(7.6)	(7.8)		(4.5)	(6.0)	(5.5)
76	473.4	−47.9	−14.8	−12.7	413.6	−14.5	−33.3	−30.2
		(8.1)	(9.0)	(9.3)		(6.4)	(8.0)	(7.4)
77	1012.9	−7.1	−8.6	−9.4	820.9	−13.0	−56.0	−51.3
		(11.3)	(12.3)	(12.2)		(9.1)	(11.1)	(10.0)
78	2147.1	40.3	−23.5	−22.4	1677.9	58.1	−53.6	−42.5
		(16.7)	(18.1)	(17.2)		(13.4)	(16.1)	(14.1)
79	3560.7	188.0	−8.4	−11.2	2797.0	340.3	119.1	122.3
		(21.0)	(23.2)	(21.6)		(16.2)	(20.1)	(17.2)
80	4709.0	572.9	178.0	175.9	3932.2	1154.3	741.6	738.5
		(23.4)	(27.2)	(24.6)		(18.0)	(23.4)	(19.5)
81	6226.0	855.5	249.5	249.9	5218.8	1920.0	1299.9	1318.5
		(27.2)	(32.4)	(29.1)		(20.7)	(28.2)	(23.1)
82	7200.6	1508.5	783.3	782.4	6150.2	2917.1	2186.0	2210.1
		(30.3)	(36.4)	(32.5)		(23.4)	(32.0)	(26.0)
83	8398.1	1390.5	588.8	601.5	7221.1	2889.9	2103.8	2142.3
		(34.4)	(41.1)	(36.6)		(27.0)	(36.7)	(29.8)
84	9874.2	652.8	−235.7	−198.5	8377.2	2202.9	1333.0	1428.9
		(39.5)	(46.9)	(41.7)		(30.5)	(41.4)	(33.4)
85	10972.7	469.8	−521.3	−459.6	9306.8	1955.5	932.3	1059.2
		(44.6)	(52.6)	(46.8)		(34.4)	(46.2)	(37.3)
86	12004.5	543.7	−557.3	−491.7	10106.2	1881.3	720.9	872.3
		(50.4)	(59.0)	(52.5)		(38.7)	(51.2)	(41.6)
87	13045.7	663.9	−548.0	−464.3	10833.0	2050.1	751.0	925.0
		(54.6)	(63.9)	(56.8)		(41.8)	(55.2)	(44.8)
88	14136.1	904.3	−415.5	−311.7	11480.1	2175.0	708.2	923.7
		(58.3)	(68.2)	(60.6)		(44.9)	(59.5)	(48.1)
89	14716.1	1169.1	−248.6	−136.3	11751.4	2379.1	799.7	1031.9
		(61.0)	(71.2)	(63.2)		(47.6)	(62.7)	(50.9)
90	14886.1	1300.8	−154.5	−53.2	11904.3	2483.6	824.9	1064.0
		(63.0)	(73.6)	(65.2)		(49.4)	(65.4)	(52.7)
91	14407.9	1559.6	29.8	146.2	11518.7	2758.8	1026.1	1277.9
		(64.6)	(75.6)	(66.9)		(50.8)	(67.2)	(54.3)

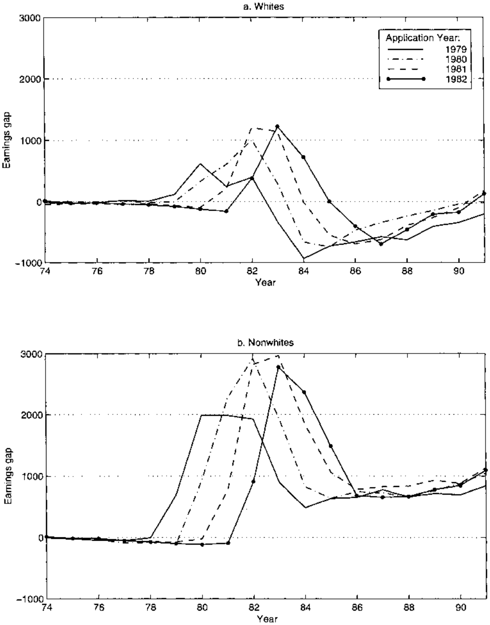

FIGURE 3.—Controlled contrasts by application year and calendar year for whites (a) and nonwhites (b).

TABLE 3.3.1
Uncontrolled, matching, and regression estimates of the effects of voluntary military service on earnings

Race	Average Earnings in 1988–1991 (1)	Differences in Means by Veteran Status (2)	Matching Estimates (3)	Regression Estimates (4)	Regression Minus Matching (5)
Whites	14,537	1,233.4 (60.3)	−197.2 (70.5)	−88.8 (62.5)	108.4 (28.5)
Non-whites	11,664	2,449.1 (47.4)	839.7 (62.7)	1,074.4 (50.7)	234.7 (32.5)

Notes: Adapted from Angrist (1998, tables II and V). Standard errors are reported in parentheses. The table shows estimates of the effect of voluntary military service on the 1988–91 Social Security–taxable earnings of men who applied to enter the armed forces between 1979 and 1982. The matching and regression estimates control for applicants' year of birth, education at the time of application, and AFQT score. There are 128,968 whites and 175,262 nonwhites in the sample.

Regression Meets Matching

Angrist (1998) reports estimates of δ_R in

$$\mathrm{Y}_i = \sum_x d_{ix}\beta_x + \delta_R \mathrm{D}_i + \varepsilon_i, \tag{3}$$

where d_{ix} indicates $\mathrm{X}_i = x$, β_x is a regression-effect for $\mathrm{X}_i = x$, and δ_R is the regression treatment effect

- Simplifying,

$$\begin{aligned}
\delta_R &= \frac{Cov(\mathrm{Y}_i, \tilde{\mathrm{D}}_i)}{V(\tilde{\mathrm{D}}_i)} = \frac{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])\mathrm{Y}_i]}{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2]} &(4)\\
&= \frac{E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])E[\mathrm{Y}_i|\mathrm{D}_i, \mathrm{X}_i]\}}{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2]}. &(5)
\end{aligned}$$

- Saturating in X_i means $E[\mathrm{D}_i|\mathrm{X}_i]$ is linear. Hence, $\tilde{\mathrm{D}}_i$, the residual from regressing D_i on X_i, is $\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i]$
- The regression of Y_i on D_i and X_i is the same as the regression of Y_i on $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i]$

Regression Meets Matching (cont.)

- Using
$$E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i] = E[\mathrm{Y}_i|\mathrm{D}_i = 0,\mathrm{X}_i] + \delta_X \mathrm{D}_i$$
to substitute for $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i]$ in the numerator of δ_R:
$$\begin{aligned} & E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i]\} \\ = \; & E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])E[\mathrm{Y}_i|\mathrm{D}_i = 0,\mathrm{X}_i]\} + E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])\mathrm{D}_i\delta_X\} \\ = \; & E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])\mathrm{D}_i\delta_X\} \end{aligned}$$
because $E[\mathrm{Y}_i|\mathrm{D}_i = 0,\mathrm{X}_i]$ and $(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])$ are uncorr. Similarly,
$$E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])\mathrm{D}_i\delta_X\} = E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2\delta_X\}.$$

- Iterating over X, we've shown
$$\delta_R = \frac{E\{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]\delta_X\}}{E\{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]\}} = \frac{E[\sigma_D^2(\mathrm{X}_i)\delta_X]}{E[\sigma_D^2(\mathrm{X}_i)]}, \qquad (6)$$
where $\sigma_D^2(\mathrm{X}_i)$ is var D cond on X:
$$\sigma_D^2(\mathrm{X}_i) = E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]$$

Regression Meets Matching (cont.)

- Regression produces a variance-weighted average of δ_X. Since D_i is a dummy, $\sigma_D^2(\mathrm{X}_i) = P(\mathrm{D}_i = 1|\mathrm{X}_i)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i))$, so

$$\delta_R = \frac{\sum_x \delta_x \left[P(\mathrm{D}_i = 1|\mathrm{X}_i = x)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x))\right] P\left(\mathrm{X}_i = x\right)}{\sum_x \left[P(\mathrm{D}_i = 1|\mathrm{X}_i = x)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x))\right] P\left(\mathrm{X}_i = x\right)}$$

- In contrast, TOT is

$$\begin{aligned} E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i &= 1] = \sum_x \delta_x P(\mathrm{X}_i = x|\mathrm{D}_i = 1) \\ &= \frac{\sum_x \delta_x P(\mathrm{D}_i = 1|\mathrm{X}_i = x) P(\mathrm{X}_i = x)}{\sum_x P(\mathrm{D}_i = 1|\mathrm{X}_i = x) P(\mathrm{X}_i = x)} \end{aligned}$$

because

$$P(\mathrm{X}_i = x|\mathrm{D}_i = 1) = \frac{P(\mathrm{D}_i = 1|\mathrm{X}_i = x) \cdot P(\mathrm{X}_i = x)}{P(\mathrm{D}_i = 1)}$$

Regression *vs* Matching

- TOT weights covariate cells in proportion to the *probability* of treatment
- Regression weights in proportion to the conditional *variance* of treatment.
  - This is maximized when $P(\mathrm{D}_i = 1|\mathrm{X}_i = x) = \frac{1}{2}$
- Angrist (1998) Figure 4 shows why this matters (some): treatment varies with the score
- Macartan Humphreys (2009) shows that if δ_x is monotone in the score (as is roughly true in Angrist 1998) then the regression estimand lies between TOT and TNT . . . pretty neat!

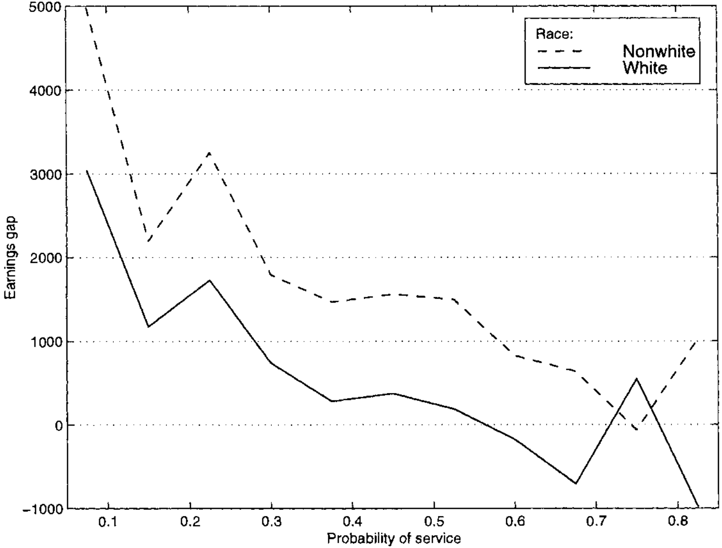

FIGURE 4.—Controlled contrasts by race and probability of service. These estimates are for pooled 1988–91 earnings.

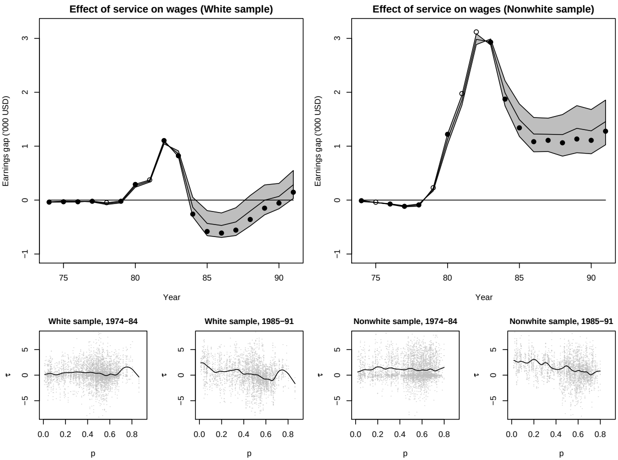

Figure 3 The top panel shows effects of participation in the military on earnings of white and nonwhite veterans for years 1974-91. The shaded band marks the region between ATT and ATC, the center line gives the ATE. The OLS estimate is marked with circles, which are filled whenever b_{OLS} lies between ATT and ATC. Bottom panels show the relation between p and τ for each group for two time periods. In early periods these relations are approximately flat and so $ATC \approx ATT$; in later periods there is a negative (near) monotonic relation and so $ATC > ATT$. In these later periods b_{OLS} always lies between ATE and ATT.