CENTRO DE ESTUDIOS ECONÓMICOS

Maestría en Economía 2024–2026

Microeconometrics for Evaluation

6 Matching

Disclaimer: I AM NOT the original intellectual author of the material presented in these notes. The content is STRONGLY based on a combination of lecture notes (from Aurora Ramirez), textbook references, and personal annotations for learning purposes. Any errors or omissions are entirely my own responsibility.

Índice

Today's Agenda

- **Matching**. We search for *causal effects* by comparing **treated** and **control groups** inside subgroups where almost everything . . . or most factors . . . or the *key elements* . . . stay fixed
- **Similarities** between *regression* approaches and **matching**
- **Example** of matching from *Angrist (1998)*

Intuition

- Matching is a way to *imitate experiments* by checking "similar" groups
- It works by **holding constant** the most *relevant characteristics*
- Today's focus: learn the method and see a **practical example**

Background: Volunteers of America!

- The **army** is the largest single employer of *young men and women* in the United States
- Between **1989 and 1992**, the enlistments of men and women with *no prior service* in the military dropped by **27 %**
- Recruitments of **white men** declined by **25 %**, while enlistments of **black men**—the group most affected by the *military cutback*—fell by **47 %** (*Angrist, 1993a*)
- The main channel behind these decreases was a **raise in cutoff scores** on applicant tests and *changes in entry rules*

Intuition

- The military is a **big employer**, so changes hit many people
- During cuts, enlistments fell, especially among *black recruits*
- Rules got tougher: **higher test thresholds** and stricter entry norms
- What were the **consequences** of military service for the **recruits**?
- By answering this, we learn whether the **military cuts** were a **lost economic opportunity** (as many believed at that time)
- The problem of **selection bias** makes comparisons between **veterans** and **non-veterans** misleading (**Seltzer and Jablon, 1974**)

Intuition

- The key question: did service **help or hurt** recruits?
- Cuts may mean **missed chances** for economic gains
- We must beware of **biased comparisons** between groups

Angrist (1998). Matching Strategy

1. Compare **veterans** and **non-veterans** who applied (only half of the **qualified candidates** serve in the army)
2. Control for the **characteristics** that the military uses to **select** soldiers

- The **matching estimator** is an average of **contrasts** or comparisons across cells defined by **covariates**
- Focusing on the **treatment effect** on the treated (TOT):

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

- This shows the gap between the **average observed earnings** of soldiers, $E[Y_{1i} \mid D_i = 1]$, and the **counterfactual average** had they not served, $E[Y_{0i} \mid D_i = 1]$

Intuition

- We compare **similar applicants** to isolate service effects
- Matching builds a **fair comparison** by holding traits constant
- The goal is to measure the **earnings gap** caused by service
- The **income gap** by veteran status is a **biased estimator** of the TOT, unless D_i is independent of Y_{0i}:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \quad \textbf{Diff. means} \tag{1}$$

$$= \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0] \quad \textbf{Def. of } Y_i \tag{2}$$

$$= \Big(\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]\Big) + \Big(\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]\Big) \quad \textbf{Add \& subtract } \mathbb{E}[Y_{0i} \mid D_i = 1] \tag{3}$$

$$= \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] + \Big(\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]\Big) \quad \textbf{Group terms} \tag{4}$$

$$= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]}_{\textbf{TOT}} + \underbrace{\Big(\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]\Big)}_{\textbf{Selection Bias}} \quad \textbf{Identify components} \tag{5}$$

Intuition

- The raw gap mixes **treatment effect** and **bias**
- **TOT**: what veterans gain from service
- **Bias**: pre-existing differences in Y_0 across groups

Conditional Independence Assumption (CIA)

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

$$\{Y_{0i}, Y_{1i}\} \perp D_i \mid X_i$$

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

$$\delta_{TOT} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] \quad \textbf{Def. TOT} \tag{6}$$

$$= \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1] \quad \textbf{Linearity of } \mathbb{E} \tag{7}$$

$$= \mathbb{E}\Big[\mathbb{E}[Y_{1i} \mid X_i, D_i = 1] \mid D_i = 1\Big] - \mathbb{E}\Big[\mathbb{E}[Y_{0i} \mid X_i, D_i = 1] \mid D_i = 1\Big] \quad \textbf{Iterated Exp.} \tag{8}$$

$$= \mathbb{E}\Big[\mathbb{E}[Y_{1i} \mid X_i, D_i = 1] - \mathbb{E}[Y_{0i} \mid X_i, D_i = 1] \Big| D_i = 1\Big] \quad \textbf{Group terms} \tag{9}$$

$$= \int \Big(\mathbb{E}[Y_{1i} \mid X_i, D_i = 1] - \mathbb{E}[Y_{0i} \mid X_i, D_i = 1]\Big) dF(X_i \mid D_i = 1) \quad \textbf{Integral form: average over distribution of } X_i \tag{10}$$

$$\mathbb{E}[Y_{0i} \mid X_i, D_i = 1] \quad \textbf{Counterfactual term} \tag{11}$$

$$\stackrel{CIA}{=} \mathbb{E}[Y_{0i} \mid X_i, D_i = 0] \quad \textbf{Use CIA} \tag{12}$$

$$\delta_{TOT} = \mathbb{E}\Big[\mathbb{E}[Y_{1i} \mid X_i, D_i = 1] - \mathbb{E}[Y_{0i} \mid X_i, D_i = 0] \,\Big|\, D_i = 1\Big] \quad \textbf{Replace } Y_0 \textbf{ term} \tag{13}$$

$$= \mathbb{E}[\delta_{X_i} \mid D_i = 1] \quad \textbf{Define } \delta_{X_i} \tag{14}$$

- where $\delta_{X_i} = \mathbb{E}[Y_{1i} \mid X_i, D_i = 1] - \mathbb{E}[Y_{0i} \mid X_i, D_i = 0]$ is the (random) **X-specific causal effect**.

Intuition

- CIA says: once we **control for** X_i, treatment is **"as if random"**
- TOT is built by **averaging treatment effects** across X_i
- Counterfactuals are replaced by **non-treated with same** X_i

Angrist (1998)

- Angrist (1998) builds the **sample analog** of the right-hand side of eq. (1) for **discrete covariates**:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] = \sum_x \delta_x \, P(X_i = x \mid D_i = 1) \quad \textbf{Discrete version} \tag{15}$$

$$\delta_x = \mathbb{E}[Y_{1i} \mid X_i = x, D_i = 1] - \mathbb{E}[Y_{0i} \mid X_i = x, D_i = 0] \quad \textbf{Cell-specific effect} \tag{16}$$

- Here $P(X_i = x \mid D_i = 1)$ is the **distribution of** X_i among veterans
- The estimator is obtained by replacing δ_x with the **mean earnings difference** between veterans and non-veterans in each cell x, and weighting by the **empirical distribution** $P(X_i = x \mid D_i = 1)$

Intuition

- With discrete X, the integral turns into a **sum over cells**
- Each δ_x captures the **local treatment effect** for covariate profile x
- The final TOT is a **weighted average** of these cell effects

Angrist (1998). Results

- **White veterans** earn more than non-veterans, but this effect becomes **negative** once covariates are matched
- **Non-white veterans** earn much more than non-veterans, but covariate controls **shrink the gap** considerably

Intuition

- At first glance, veterans seem to earn **more**, but once we compare **similar people**, the advantage often **disappears or reverses**
- Think of it like comparing two runners: one starts with a head start. After correcting for that, the real performance looks different
- For non-white veterans, the raw difference looks huge, but careful matching shows it's **partly explained by background factors**

Tabla 1. Applicant Population and Sample

	1976	1977	1978	1979	1980	1981	1982
A. Population[a]							
White	393.5	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	40	38	44	48	47	38	45

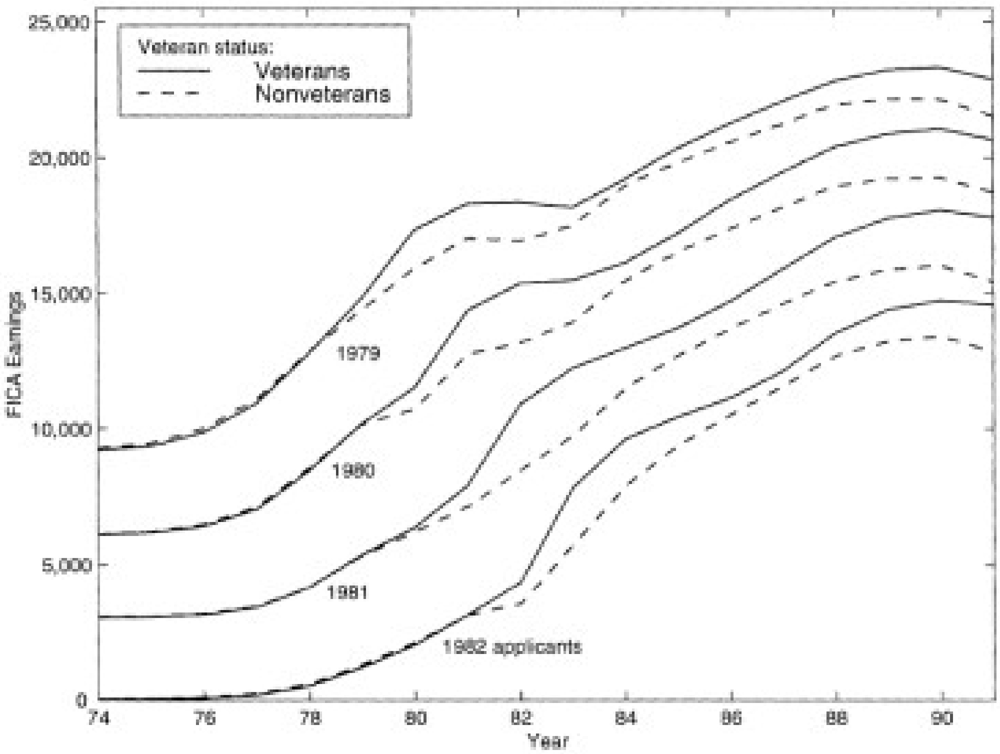
[a] As in Angrist (1993a, Table 4), excluding applicants with less than 9th grade education at application. Numbers in thousands.
[b] Veterans are applicants identified as entrants to the military within two years following application.
[c] Approximately 90 % of the sample is self-weighting, conditional on race.

Explanation

Table 1 reports the size of the applicant pool and the sample used in the analysis, split by **race** and **year of application** (1976–1982). Panel A shows the total number of applicants (in thousands) and the share who are **veterans**. Panel B shows the selected sample of applicants who actually enlisted, again by race and year, with corresponding veteran shares.

Intuition

- White applicants are the majority, but the share of **veterans** is similar across races
- The **sample** is much smaller than the population, since not all applicants enlist
- Think of it like a **funnel**: many apply (Panel A), fewer serve (Panel B). This shows why careful sampling matters in the analysis



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 actual earnings: + \$3,000 for 1982 applicants, + \$6,000 for 1981 applicants, + \$6,000 for 1980 applicants, and + \$9,000 for 1979 applicants.

Explanation

The figure plots **earnings profiles** over time for veterans and nonveterans, by year of application (1979–1982). Each curve shows the actual FICA earnings for men with AFQT scores in categories III and IV. Veterans consistently earn more than nonveterans in the early years, with the earnings advantage varying by application cohort.

Intuition

- The **gap** between the solid (veterans) and dashed (nonveterans) lines shows how much extra veterans earned

- Earlier cohorts (like 1979) had the **largest boost**, while later ones (1982) saw smaller gains

- Think of it like different **graduating classes**: those who entered earlier cashed in more, while later classes faced lower returns

Tabla 2. Alternative Estimates of the Effects of Military Service

	Whites				Nonwhites			
Year	Mean	Diff. in Means	Controlled Contrast	Regression Est.	Mean	Diff. in Means	Controlled Contrast	Regression Est.
74	182.7	-26.1	-14.0	-13.0	157.2	-4.9	-2.0	-3.9
75	237.9	-41.4	-14.2	-12.0	216.9	-6.6	-17.1	-15.2
76	473.4	-47.9	-14.8	-12.7	413.6	-14.5	-33.3	-30.2
77	1012.9	-7.1	-8.6	-9.4	820.9	-13.0	-56.0	-51.3
78	2147.1	40.3	-23.5	-22.4	1677.9	58.1	-53.6	-42.5
79	3560.7	188.0	-8.4	-11.2	2797.0	340.3	119.1	122.3
80	4790.0	572.9	178.0	175.9	3932.2	1154.3	741.6	738.5
81	6226.0	855.5	249.5	249.9	5218.8	1920.0	1299.9	1318.5
82	7200.6	1508.5	783.3	782.4	6150.2	2917.1	2186.0	2210.1
83	7957.3	1305.3	584.4	532.5	6852.1	2834.9	2320.0	2260.0
84	9874.2	1593.5	888.6	805.1	8377.2	2902.7	1330.6	1289.2
85	10972.7	2097.2	1281.3	1258.9	9306.8	3555.5	1942.3	1939.2
86	12004.5	543.7	-557.3	-491.7	10106.2	1381.3	720.9	872.3
87	13045.7	654.3	-598.0	-464.3	10833.0	2050.1	751.0	925.4
88	14006.4	593.0	-440.8	-333.5	11506.7	2158.2	858.0	924.3
89	14117.5	-61.0	-71.2	-62.9	11751.4	843.2	189.6	267.9
90	14886.1	-139.8	-166.2	-139.6	11904.3	2483.6	624.9	1064.0
91	14407.9	1559.6	29.8	16.2	11518.7	2758.8	1062.1	1277.9

Explanation

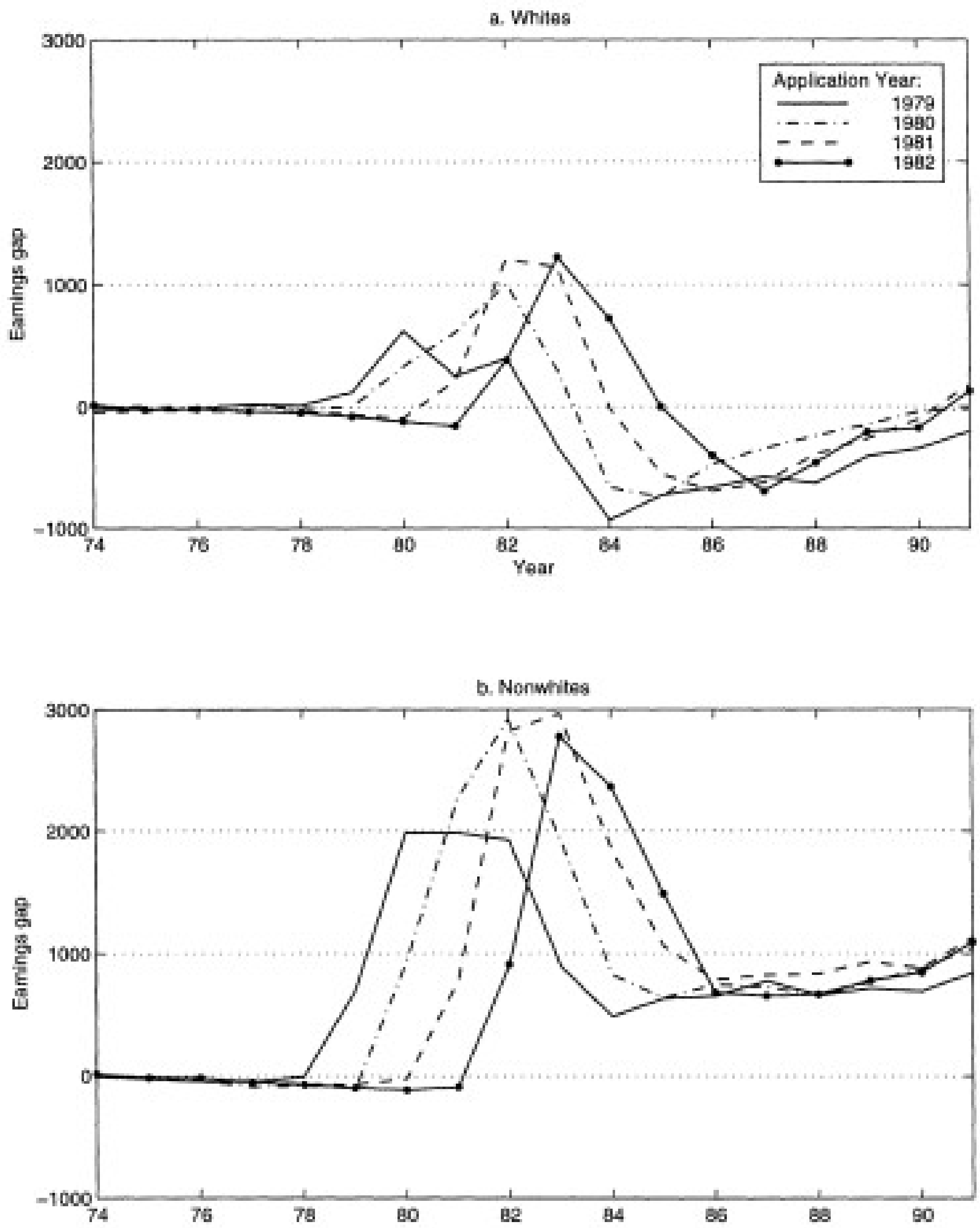
Table 2 reports alternative estimates of the effect of **military service** on earnings, shown separately for **whites** and **nonwhites**, by year of application. Each panel shows four measures: the **mean earnings**, the simple **difference in means** between veterans and nonveterans, a **controlled contrast** adjusting for covariates, and a **regression estimate**. Standard errors are reported in parentheses in the original source.

Intuition

- Raw comparisons (diff. in means) often exaggerate or mislead; once we **control for covariates**, the effects change noticeably

- For whites, the estimated benefit is sometimes small or even **negative**, while for nonwhites the veteran premium is usually **larger and more positive**

- Think of it like comparing test scores: if we just compare averages, one group looks much better — but once we compare students with the **same background**, the real difference is smaller (and sometimes flips sign)



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

Explanation

Figure 3 shows the **controlled contrasts** in earnings by year of application and calendar year, separately for **whites** (panel a) and **nonwhites** (panel b). Each line corresponds to a different application cohort (1979–1982). The vertical axis measures the earnings gap between veterans and nonveterans after controlling for covariates.

Intuition

- The veteran **premium** rises sharply in the early 1980s and then gradually declines for both groups
- For whites, the boost is smaller and sometimes disappears; for nonwhites, the peak is **larger and more sustained**
- Think of it like **waves of opportunity**: earlier cohorts (1979–1980) rode a higher wave, while later ones (1982) caught a smaller one

Tabla 3. Uncontrolled, Matching, and Regression Estimates of the Effects of Voluntary Military Service on Earnings

Race	Average earnings 1988–1991 (1)	Diff. in means by veteran status (2)	Matching estimates (3)	Regression estimates (4)	Regression minus matching (5)
Whites	14537	1233.4	-197.2	-88.8	108.4
	(60.3)	(60.3)	(70.5)	(62.5)	(28.5)
Nonwhites	11664	2449.1	839.7	1074.4	234.7
	(47.4)	(47.4)	(62.7)	(50.7)	(32.5)

Notes: Adapted from Angrist (1998, Tables II and V). Standard errors in parentheses. Estimates cover 1988–1991 Social Security-taxable earnings of men who applied to enter the armed forces between 1979 and 1982. Controls include year of birth, education at application, and AFQT score. Sample sizes: 128,968 whites and 175,262 nonwhites.

Explanation

Table 3 summarizes the estimated effects of **voluntary military service** on 1988–1991 earnings. Column (1) shows the average earnings for whites and nonwhites. Column (2) presents the **raw differences in means** between veterans and nonveterans. Columns (3) and (4) report **matching** and **regression** estimates, which adjust for birth year, education, and AFQT scores. Column (5) shows the difference between regression and matching estimates. Standard errors are in parentheses.

Intuition

- Raw differences suggest veterans earn **more**, especially among nonwhites
- Once we **control for background factors**, the advantage for whites **disappears**, but remains **positive and large** for nonwhites
- Think of it like comparing two classrooms: if we just take the average grades, veterans seem ahead. But once we compare students with the **same prior preparation**, the gap mostly closes for whites, while nonwhites still keep a real boost

Regression vs. Matching

- Angrist (1998) reports estimates of δ_R in the regression:

$$Y_i = \sum_x d_{ix}\beta_x + \delta_R D_i + \varepsilon_i \qquad \textbf{Model with covariates} \tag{17}$$

- where d_{ix} indicates $X_i = x$, β_x is the regression effect for cell $X_i = x$, and δ_R is the treatment effect from regression.
- Simplifying, the OLS estimator is:

$$\delta_R = \frac{\text{Cov}(Y_i, \tilde{D}_i)}{\text{Var}(\tilde{D}_i)} \qquad \textbf{OLS formula} \tag{18}$$

$$\tilde{D}_i = D_i - \mathbb{E}[D_i \mid X_i] \qquad \textbf{Residualized treatment} \tag{19}$$

$$\text{Cov}(Y_i, \tilde{D}_i) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(D_i - \mathbb{E}[D_i \mid X_i])] \qquad \textbf{Def. of Cov} \tag{20}$$

$$= \mathbb{E}[(D_i - \mathbb{E}[D_i \mid X_i])Y_i] \qquad \textbf{Cross-term simplification} \tag{21}$$

$$\delta_R = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i \mid X_i])Y_i]}{\mathbb{E}[(D_i - \mathbb{E}[D_i \mid X_i])^2]} \qquad \textbf{Plug into OLS ratio} \tag{22}$$

$$= \frac{\mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i]) \cdot \mathbb{E}[Y_i \mid D_i, X_i]\}}{\mathbb{E}[(D_i - \mathbb{E}[D_i \mid X_i])^2]} \qquad \textbf{Use law of iterated exp.} \tag{23}$$

- Interpretation:

$$\delta_R = \frac{\text{Weighted average of conditional treatment effects}}{\text{Variance of residualized } D_i} \qquad \textbf{Interpretation of estimator} \tag{24}$$

Intuition

- Regression with controls is like comparing veterans and nonveterans **after "residualizing" treatment** on covariates
- Matching does the same idea more directly: build **pairs of similar units** and compare
- Example: Suppose age predicts enlistment. Regression subtracts the "age effect" before comparing earnings; matching compares **same-age individuals** directly

Regression vs. Matching II

- Start from the conditional expectation of Y_i given D_i and X_i:

$$\mathbb{E}[Y_i \mid D_i, X_i] = \mathbb{E}[Y_i \mid D_i = 0, X_i] + \delta_X D_i \qquad \textbf{Treatment effect by } X \tag{25}$$

- Substitute this into the numerator of δ_R:

$$\mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i])Y_i\} = \mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i]) \cdot \mathbb{E}[Y_i \mid D_i, X_i]\} \qquad \textbf{Replace } Y_i \textbf{ with cond. exp.} \tag{26}$$

$$= \mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i])[\mathbb{E}[Y_i \mid D_i = 0, X_i] + \delta_X D_i]\} \qquad \textbf{Expand using formula above} \tag{27}$$

$$= \mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i]) \cdot \mathbb{E}[Y_i \mid D_i = 0, X_i]\} + \mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i]) \cdot \delta_X D_i\} \qquad \textbf{Distribute terms} \tag{28}$$

$$= 0 + \mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i])D_i\delta_X\} \qquad \textbf{Uncorrelated with baseline term} \tag{29}$$

$$= \mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i])D_i\delta_X\} \qquad \textbf{Keep treatment part} \tag{30}$$

$$= \mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i])^2\delta_X\} \qquad \textbf{Because factor } D_i \textbf{ multiplies residual} \tag{31}$$

- Therefore:

$$\delta_R = \frac{\mathbb{E}\{(D_i - \mathbb{E}[D_i \mid X_i])^2 \delta_X\}}{\mathbb{E}[(D_i - \mathbb{E}[D_i \mid X_i])^2]} \qquad \textbf{Plug into OLS ratio} \tag{32}$$

- Iterate expectations over X:

$$\delta_R = \frac{\mathbb{E}\Big[\mathbb{E}[(D_i - \mathbb{E}[D_i \mid X_i])^2 \mid X_i] \cdot \delta_X\Big]}{\mathbb{E}\Big[\mathbb{E}[(D_i - \mathbb{E}[D_i \mid X_i])^2 \mid X_i]\Big]} \qquad \textbf{Law of Iterated Expectations} \tag{33}$$

$$\delta_R = \frac{\mathbb{E}[\sigma_D^2(X_i) \cdot \delta_X]}{\mathbb{E}[\sigma_D^2(X_i)]} \qquad \textbf{Define conditional variance} \tag{34}$$

where $\sigma_D^2(X_i) = \mathbb{E}[(D_i - \mathbb{E}[D_i \mid X_i])^2 \mid X_i]$.
Intuition

- Regression puts more weight on cells where D_i has higher **variance within** X_i (more treated and untreated units to compare)
- Matching weights all cells more evenly, regression emphasizes those with **better overlap**
- Example: If veterans and nonveterans are balanced in education group A but rare in group B, regression leans on group A for precision, while matching treats both groups similarly

Regression vs. Matching III

Step 1. Variance of D_i conditional on X_i
Since D_i is binary:

$$\sigma_D^2(X_i) = \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2 \mid X_i]$$

But $\mathbb{E}[D_i|X_i] = P(D_i = 1 \mid X_i)$. So,

$$\sigma_D^2(X_i) = P(D_i = 1|X_i) \cdot (1 - P(D_i = 1|X_i))$$

Step 2. Regression estimand as weighted average
We had:

$$\delta_R = \frac{\mathbb{E}[\sigma_D^2(X_i)\delta_X]}{\mathbb{E}[\sigma_D^2(X_i)]}$$

Expanding over support of X:

$$\delta_R = \frac{\sum_x \delta_x \, \sigma_D^2(x) \, P(X_i = x)}{\sum_x \sigma_D^2(x) \, P(X_i = x)}$$

Explicitly:

$$\delta_R = \frac{\sum_x \delta_x \left[P(D_i = 1|X_i = x)(1 - P(D_i = 1|X_i = x))\right] P(X_i = x)}{\sum_x [P(D_i = 1|X_i = x)(1 - P(D_i = 1|X_i = x))] P(X_i = x)}$$

Regression = weighted avg.
weights $\propto$ var$(D|X)$

Step 3. Matching / TOT definition
The treatment-on-the-treated (TOT) effect:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] = \sum_x \delta_x \, P(X_i = x \mid D_i = 1)$$

Step 4. Expand conditional probability
By Bayes rule:

$$P(X_i = x \mid D_i = 1) = \frac{P(D_i = 1 \mid X_i = x) P(X_i = x)}{P(D_i = 1)}$$

So,

$$\text{TOT} = \frac{\sum_x \delta_x P(D_i = 1 \mid X_i = x) P(X_i = x)}{\sum_x P(D_i = 1 \mid X_i = x) P(X_i = x)}$$

TOT = weighted avg.
weights $\propto P(D = 1|X)P(X)$

Step 5. Comparison
- Regression estimator δ_R weights by *overlap/variance*: $P(D = 1|X)(1-P(D = 1|X))P(X)$.
- TOT weights by *treated mass*: $P(D = 1|X)P(X)$.

Intuition

- **Regression**: emphasizes strata x where treatment is more <u>balanced</u>, since variance of D is maximized when $P(D = 1|X) = 0{,}5$.
- **TOT**: emphasizes strata x where more people are treated, regardless of balance.
- If veterans are spread across education groups, regression will lean more on groups with mix of veterans and non-veterans, while TOT leans more on groups with many veterans (even if imbalance is large).

Regresión vs. Matching IV

- **TOT**: pondera cada celda de covariantes x en proporción a la *probabilidad de tratamiento*, $P(D_i = 1|X_i = x)P(X_i = x)$.
- **Regresión**: pondera cada celda en proporción a la *varianza condicional del tratamiento*,

$$P(D_i = 1|X_i = x)\Big(1 - P(D_i = 1|X_i = x)\Big)P(X_i = x).$$

- Esta varianza se maximiza cuando $P(D_i = 1|X_i = x) = \frac{1}{2}$ (máximo balance).

Ejemplo breve:

- Supongamos 2 grupos $X = A, B$ con efectos $\delta_A = 1000$ y $\delta_B = 2000$.
- En A, 90 % tratados; en B, 50 % tratados.
- TOT da más peso al grupo A (muchos tratados) $\Rightarrow$ estimador ≈ 1000.
- Regresión da más peso al grupo B (más balance) $\Rightarrow$ estimador ≈ 2000.

Intuición:

- **TOT** sigue a los tratados: mide "lo que realmente experimentaron".
- **Regresión** sigue al balance: se concentra en los grupos donde la comparación tratado/no tratado es más creíble.

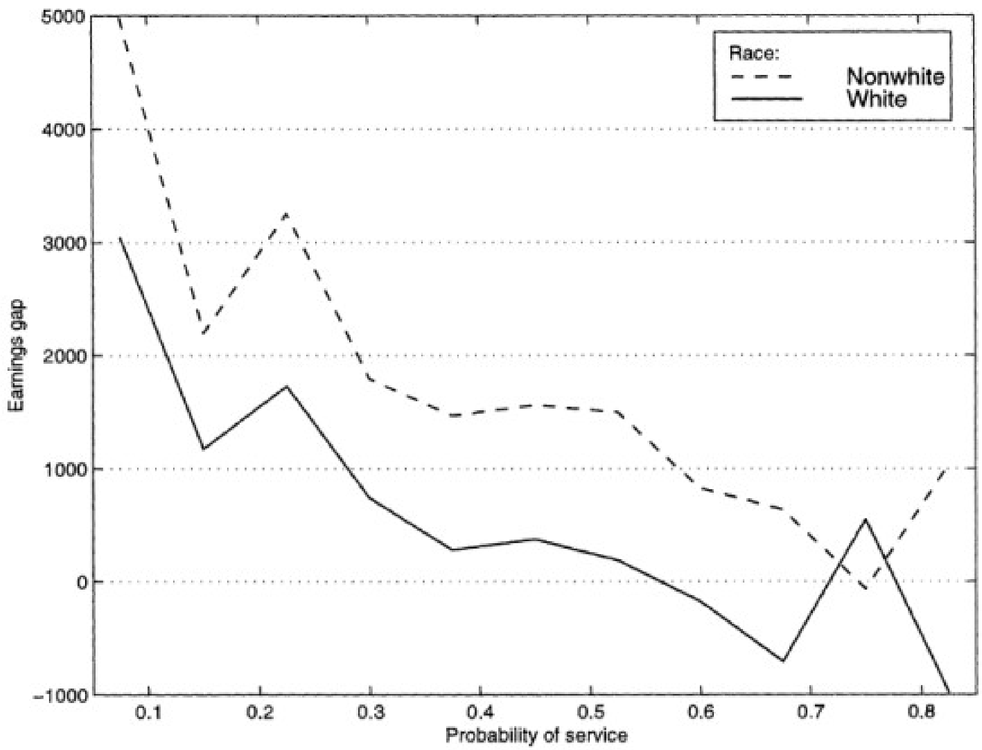


Figura 1: Controlled contrasts by race and probability of service. These estimates are for pooled 1988–91 earnings.

- **Explanation:** The earnings gap is large and positive when the probability of service is low, but it shrinks and even disappears as the probability rises. This reflects selection: early entrants differ from later ones.
- **Intuition:** At first, only the most motivated or skilled joined, so veterans look richer. As more people serve, including less advantaged groups, the average veteran advantage fades.

References

- **Angrist, Joshua D.** and **Joern-Steffen Pischke** (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press. Section 3.3.1.
- **Angrist, Joshua D.** (1998). "Estimating the Labor Market Impact of Voluntary Military Service using Social Security Data on Military Applicants." *Econometrica*, 66(2): 249–288.