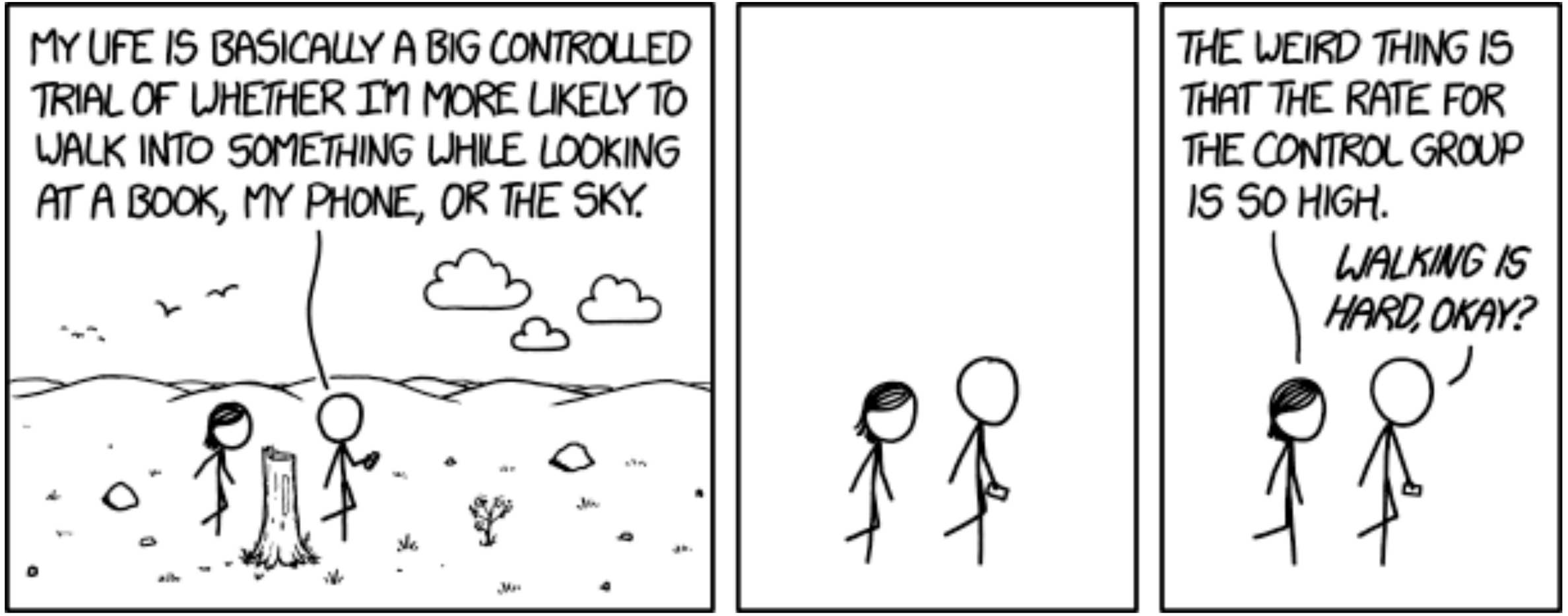
Observational Studies: DAY TWO

Lecture 2a

PP 249 (Fall 2025)


MY LIFE IS BASICALLY A BIG CONTROLLED TRIAL OF WHETHER I'M MORE LIKELY TO WALK INTO SOMETHING WHILE LOOKING AT A BOOK, MY PHONE, OR THE SKY.
THE WEIRD THING IS THAT THE RATE FOR THE CONTROL GROUP IS SO HIGH.
WALKING IS HARD, OKAY?

Plan for today

- LaLonde: Subclassification
- Regression for obs studies
- Matching
  - Exact matching
  - Approximate matching
  - Matching ***and*** regression (not matching ***or*** regression)
- LaLonde: Matching

Cold open: the assumption of "no unmeasured confounding given X" means

- Conditional on observed X, treatment assignment is as good as random
- Conditional on observed X, there are no unmeasured covariates that predict both…
- Conditional on observed X, every unit has some chance of receiving either treatm…
- B and C only
- A and B only

Cold open: the assumption of "no unmeasured confounding given X" means

Conditional on observed X, treatment assignment is as good as random

0%

Conditional on observed X, there are no unmeasured covariates that predict both the treatment and outcome

0%

Conditional on observed X, every unit has some chance of receiving either treatment or control

0%

B and C only

0%

A and B only

0%

Cold open: the assumption of "no unmeasured confounding given X" means

Conditional on observed X, treatment assignment is as good as random

0%

Conditional on observed X, there are no unmeasured covariates that predict both the treatment and outcome

0%

Conditional on observed X, every unit has some chance of receiving either treatment or control

0%

B and C only

0%

A and B only

0%

Reminder:
Smoking and Mortality

THE EFFECTIVENESS OF ADJUSTMENT BY SUBCLASSIFICATION IN REMOVING BIAS IN OBSERVATIONAL STUDIES

W. G. COCHRAN

Harvard University, Cambridge, Mass., U. S. A.

Mortality Rate

Table 5.1: Death rates per 1,000 person-years (Cochran 1968)

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Source: Cunningham *Mixtape*

Age

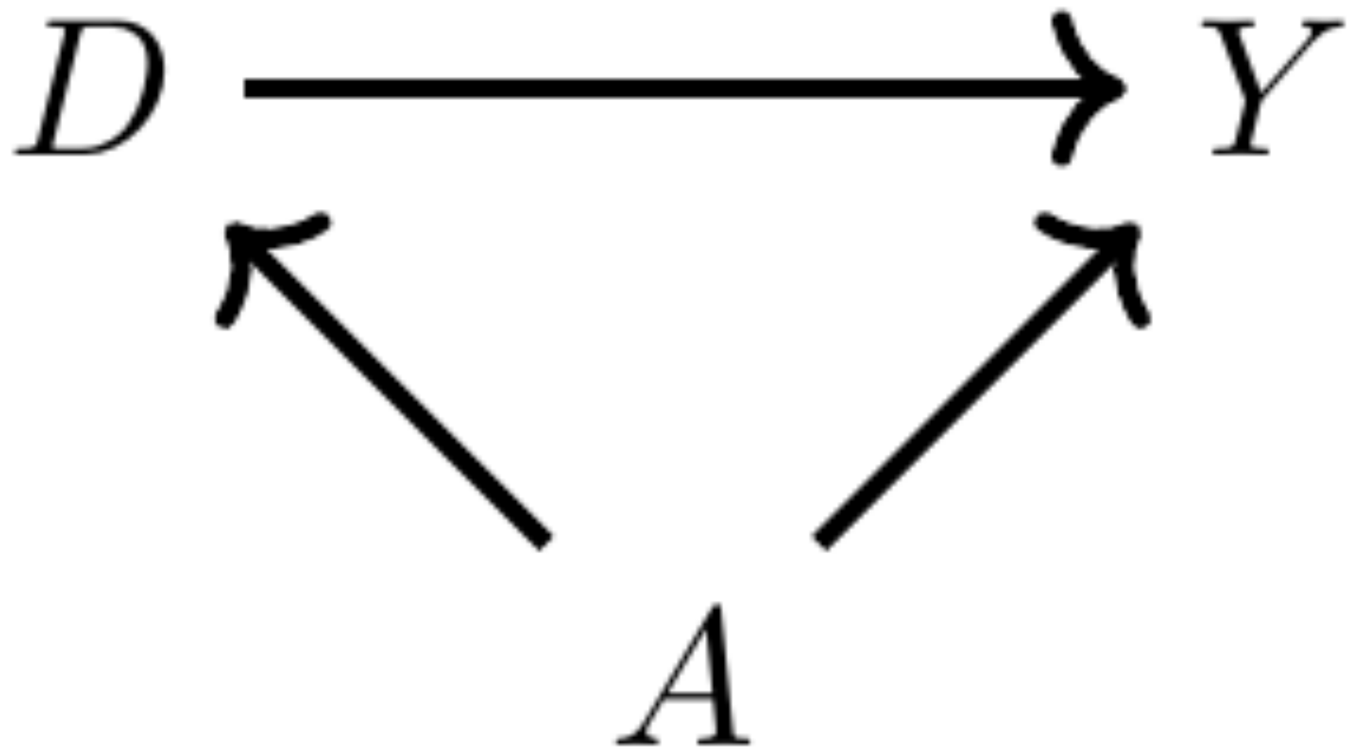
Table 5.2: Mean ages, years (Cochran 1968).

Smoking group	Canada	British	US
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Source: Cunningham *Mixtape*

Smoking

Death



Age

Distribution of smoking by age

Table 5.3: Subclassification example.

	Death rates	# of Cigarette smokers	# of Pipe or cigar smokers
Age 20–40	20	65	10
Age 41–70	40	25	25
Age $\geq$ 71	60	10	65
Total		100	100

Source: Cunningham *Mixtape*

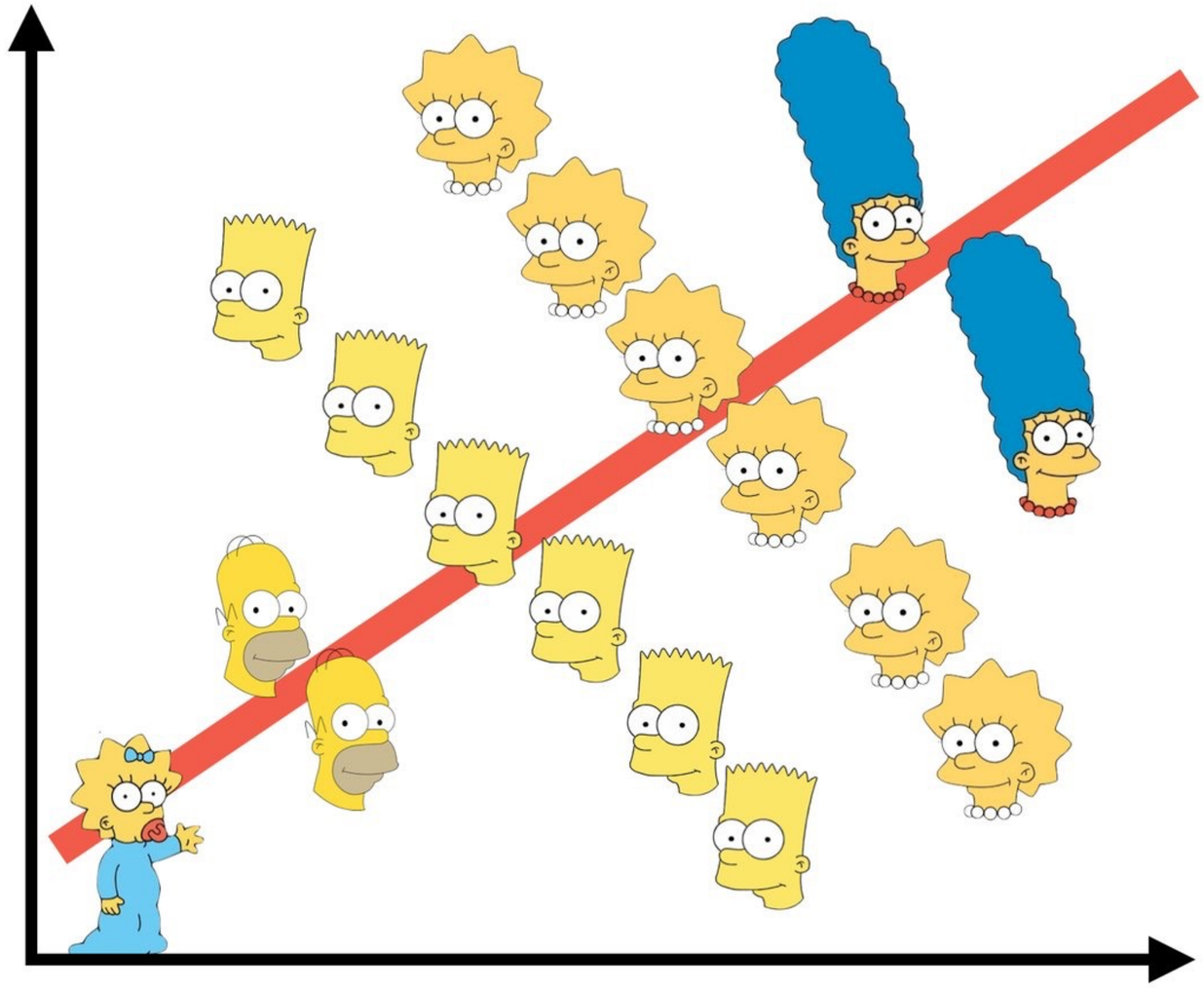
Age-adjusted mortality

Table 5.4: Adjusted mortality rates using 3 age groups (Cochran 1968).

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	29.5	14.8	21.2
Cigars/pipes	19.8	11.0	13.7

Source: Cunningham *Mixtape*

Simpson's Paradox



THERE IS NO MAGIC

We’re assuming away the problem!

(once we know **educ**)

The Obligatory LaLonde Study

Evaluating the Econometric Evaluations of Training Programs with Experimental Data

By ROBERT J. LALONDE*

This paper compares the effect on trainee earnings of an employment program that was run as a field experiment where participants were randomly assigned to treatment and control groups with the estimates that would have been produced by an econometrician. This comparison shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other nonexperimental evaluations.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Adjusted[c] (3)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978 Unadjusted (4)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons: Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons: With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78: Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78: Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

LaLonde: Covariate Imbalance

Variable	NSW		Full Samples	
	Treated	Control	CPS-1	CPS-3
	(1)	(2)	(3)	(4)
Age	25.82	25.05	33.23	28.03
Years of schooling	10.35	10.09	12.03	10.24
Black	0.84	0.83	0.07	0.20
Hispanic	0.06	0.11	0.07	0.14
Dropout	0.71	0.83	0.30	0.60
Married	0.19	0.15	0.71	0.51
1974 earninigs	2,096	2,107	14,017	5,619
1975 earnings	1,532	1,267	13,651	2,466
Number of Obs.	185	260	15,992	429

(From MHE Table 3.3.2)

Source: Borusyak, MHE

Covariate balance for race/ethnicity

Race/Eth. Category	N	Prop. Treated
White	299	0.06
Hispanic	72	0.15
Black	243	0.64

	Prop. Black	Prop. Hispanic
Treated	0.84	0.06
Control	0.20	0.14

Subclassify by race/ethnicity

Naive Difference in Means: -$635

[i.e., program has an *adverse* effect]

Subclassify by race/ethnicity

Naive Difference in Means: -$635

[i.e., program has an *adverse* effect]

Race/Eth. Category	Diff-in-Means	N	Prop. Treated
White	+$103	299	0.06
Hispanic	+$19	72	0.15
Black	+$1,283	243	0.64

Subclassify by race/ethnicity

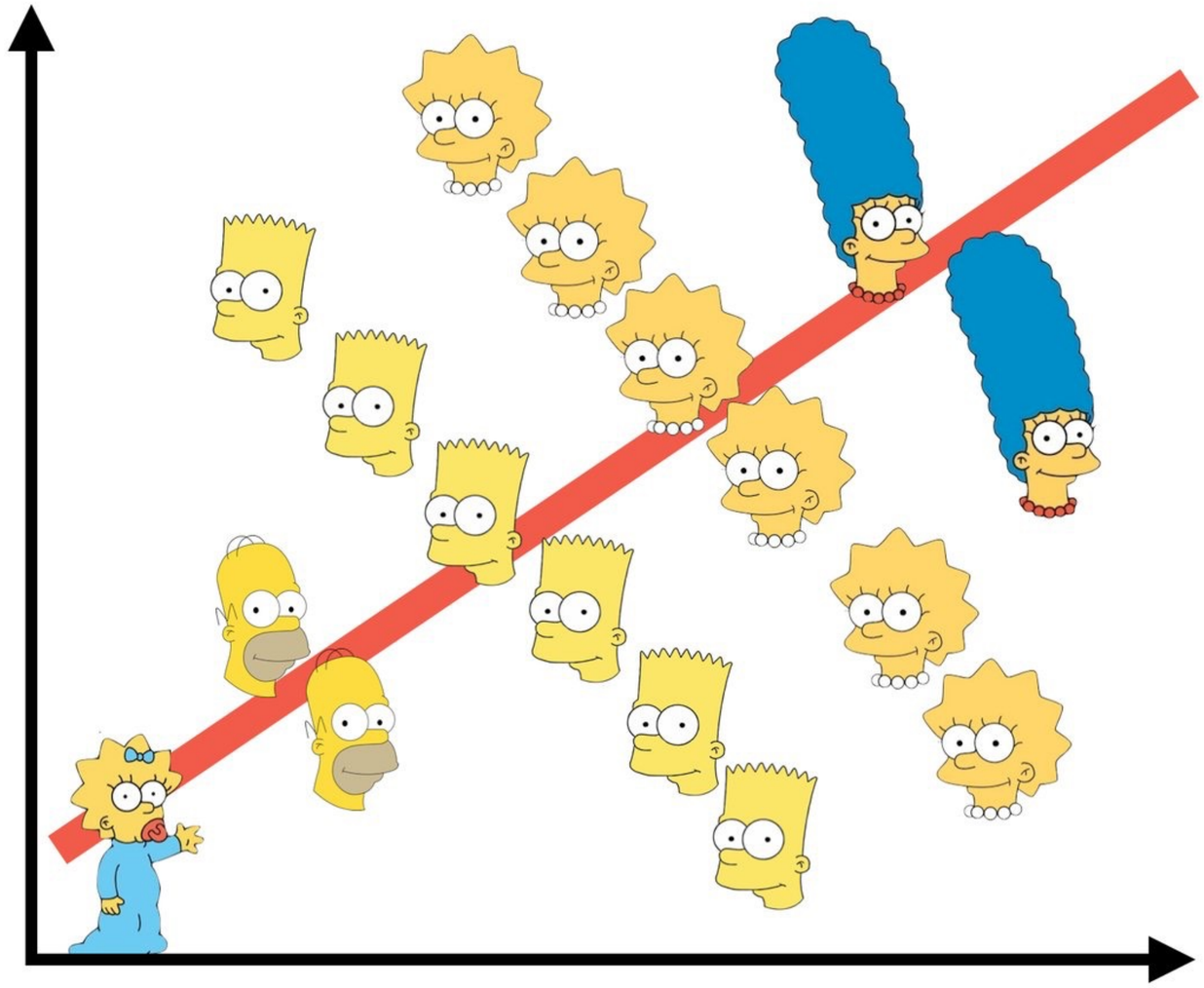
Naive Difference in Means: -$635

[i.e., program has an *adverse* effect]

Race/Eth. Category	Diff-in-Means	N	Prop. Treated
White	+$103	299	0.06
Hispanic	+$19	72	0.15
Black	+$1,283	243	0.64

Re-Weighted Difference in Means: +$560

Simpson's Paradox



Regression

Regression with discrete X

`Y ~ treat + race_eth`

```
Call:
lm_robust(formula = re78 ~ treat + black + hispan, data = lalonde)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)   7570.1      435.5 17.3824 2.840e-55   6714.9     8425 610
treat          896.2      742.1  1.2077 2.276e-01   -561.1     2354 610
black        -2468.5      700.9 -3.5217 4.608e-04  -3845.0    -1092 610
hispan        -600.4      935.9 -0.6415 5.214e-01  -2438.3     1238 610

Multiple R-squared:  0.01702 ,  Adjusted R-squared:  0.01218
F-statistic: 4.243 on 3 and 610 DF,  p-value: 0.005568
```

[Differs from stratified est of +$560. Why?]

Regression

Regression with discrete X

`Y ~ treat + race_eth`

```
Call:
lm_robust(formula = re78 ~ treat + black + hispan, data = lalonde)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)  7570.2      435.5 17.3824 2.840e-55   6714.9     8425 610
treat         896.2      742.1  1.2077 2.276e-01   -561.1     2354 610
black       -2468.4      700.9 -3.5217 4.608e-04  -3845.0    -1092 610
hispan       -600.4      935.9 -0.6415 5.214e-01  -2438.3     1238 610

Multiple R-squared:  0.01702 ,  Adjusted R-squared:  0.01218
F-statistic: 4.243 on 3 and 610 DF,  p-value: 0.005568
```

[Differs from stratified est of +$560. Why?]

Subclassify by .... more?

Race/Eth. Category	HS Degree	ATE	N	Prop. Treated
White	**No**	+$1,868	163	0.06
Hispanic	**No**	+$385	55	0.16
Black	**No**	+$885	169	0.67
White	**Yes**	-$1,903	136	0.07
Hispanic	**Yes**	-$1,677	17	0.12
Black	**Yes**	+$2,700	74	0.58

Subclassify by …. more?

Race/Eth. Category	HS Degree	ATE	N	Prop. Treated
White	No	+$1,868	163	0.06
Hispanic	No	+$385	55	0.16
Black	No	+$885	169	0.67
White	Yes	-$1,903	136	0.07
Hispanic	Yes	-$1,677	17	0.12
Black	Yes	+$2,700	74	0.58

Re-Weighted Difference in Means: +$632

Subclassify by …. more?

Race/Eth. Category	HS Degree	ATE	N	Prop. Treated
White	No	+$1,868	163	0.06
Hispanic	No	+$385	55	0.16
Black	No	+$885	169	0.67
White	Yes	-$1,903	136	0.07
Hispanic	Yes	-$1,677	17	0.12
Black	Yes	+$2,700	74	0.58

Re-Weighted Difference in Means: +$632

Regression (take 2)

Regression with discrete X?

`Y ~ treat + race_eth + hs_degree`

```
Call:
lm_robust(formula = re78 ~ treat + black + hispan + hs_degree,
    data = lalonde)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)   6633.0      491.2 13.5047 1.527e-36   5668.4   7597.6 609
treat         1003.2      745.2  1.3461 1.788e-01   -460.4   2466.8 609
black        -2223.1      705.1 -3.1531 1.695e-03  -3607.8   -838.5 609
hispan        -162.7      947.0 -0.1718 8.637e-01  -2022.5   1697.2 609
hs_degree     2046.2      650.8  3.1442 1.746e-03    768.1   3324.2 609

Multiple R-squared:  0.03395 ,  Adjusted R-squared:  0.0276
F-statistic: 5.625 on 4 and 609 DF,  p-value: 0.0001887
```

Is this the "right" model? Interactions?

[Differs from stratified est of +$632. Why?]

Regression (take 2)

Regression with discrete X?

`Y ~ treat + race_eth + hs_degree`

```
Call:
lm_robust(formula = re78 ~ treat + black + hispan + hs_degree,
    data = lalonde)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)  [illegible]   491.2 13.5047 1.527e-36   5668.4   7597.6 609
treat         1003.2       745.2  1.3461 1.788e-01   -460.4   2466.8 609
black        [illegible]   705.1 -3.1531 1.695e-03  -3607.8   -838.5 609
hispan        -162.7       947.0 -0.1718 8.637e-01  -2022.5   1697.2 609
hs_degree     2046.2       650.8  3.1442 1.746e-03    768.1   3324.2 609

Multiple R-squared:  0.03395 ,  Adjusted R-squared:  0.0276
F-statistic: 5.625 on 4 and 609 DF,  p-value: 0.0001887
```

Is this the "right" model? Interactions?

[Differs from stratified est of +$632. Why?]

OLS? OLS!

Two perspectives on regression

Regression models make it all too easy to substitute technique for work.... Regression models often seem to be used to compensate for problems in measurement, data collection, and study design. By the time the models are deployed, the scientific position is nearly hopeless.

- Freedman (1991)

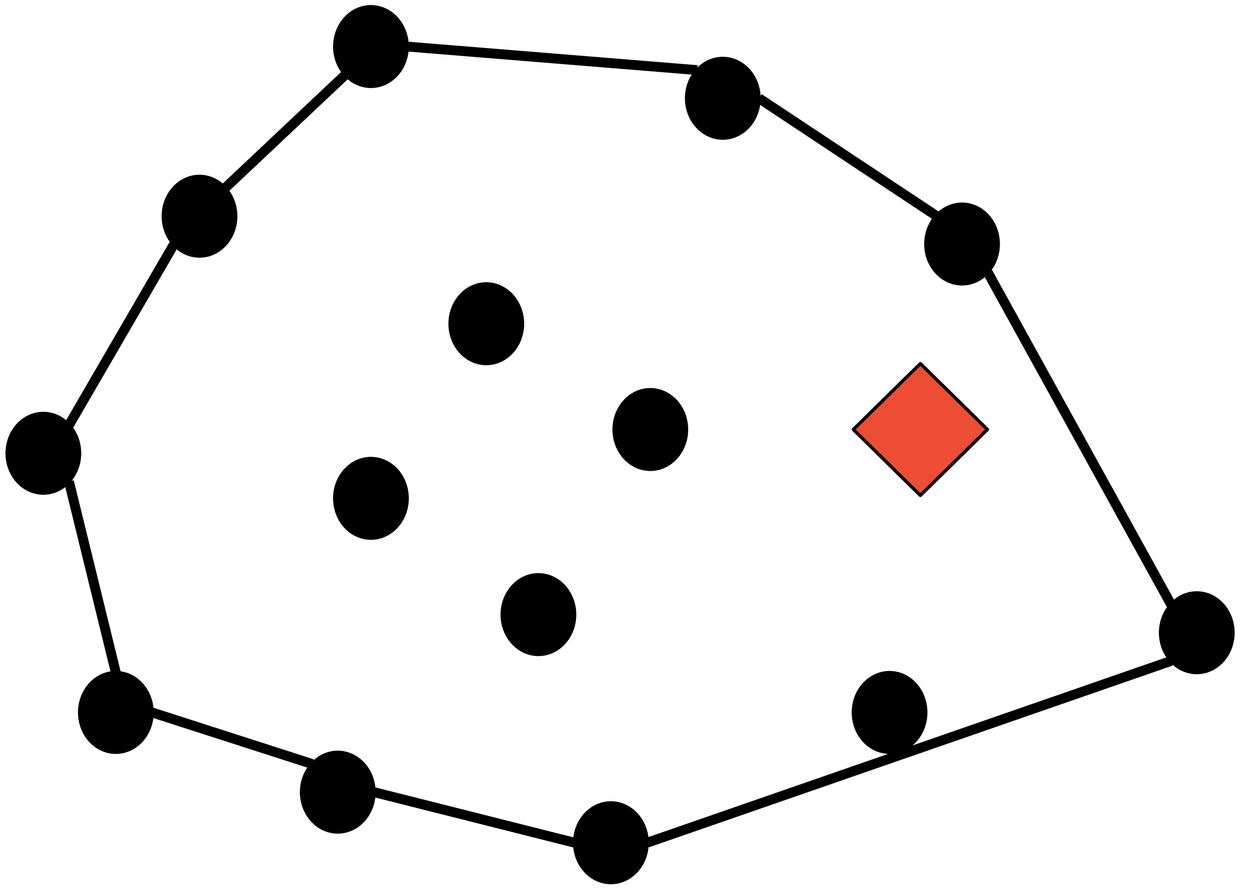
Nevertheless, we believe regression should be the starting point for most empirical projects...

- Angrist & Pischke (2012)

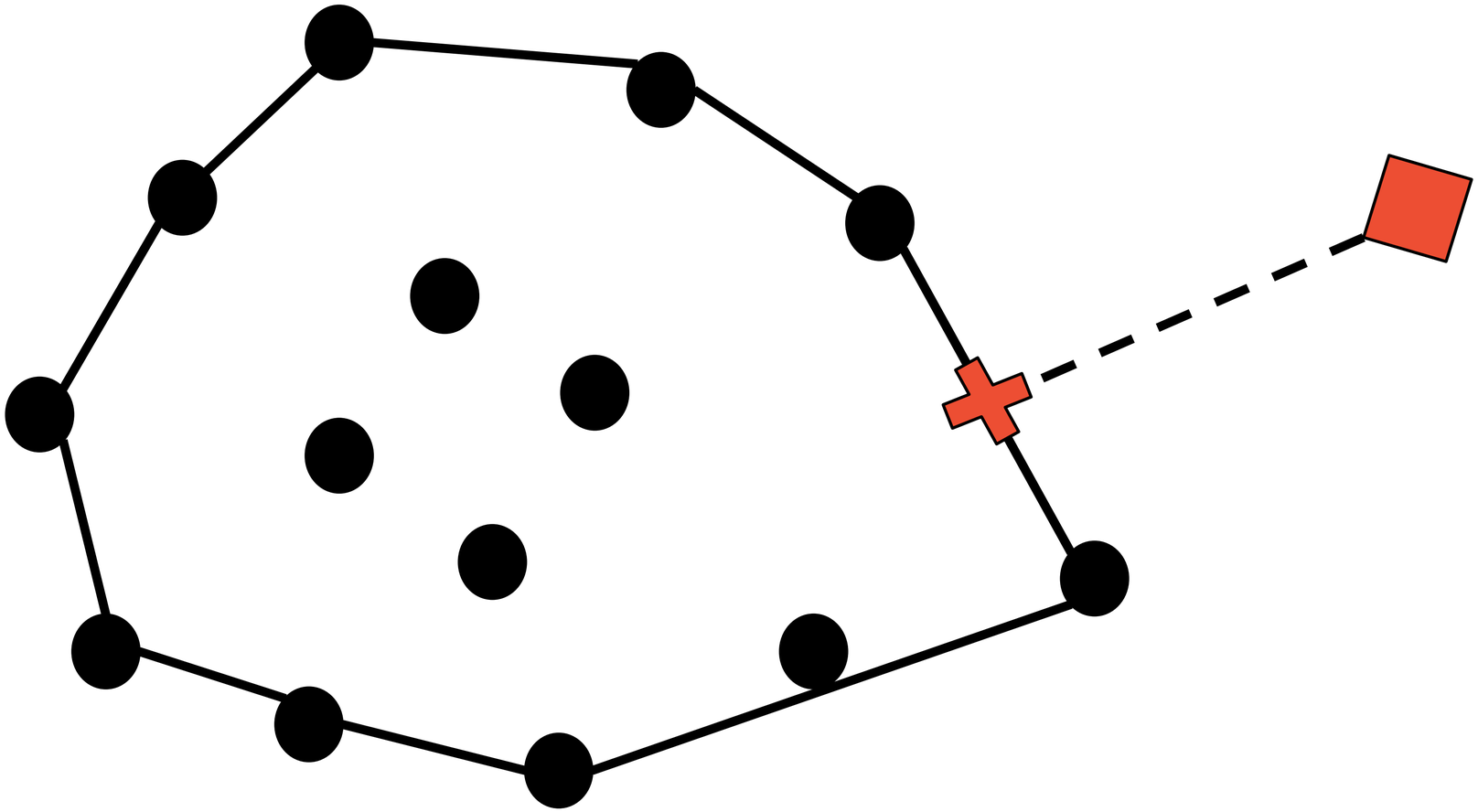
Issue #1: Mis-specification

- Estimate: $Y \sim 1 + D + X_1 + X_2$ but truth is $Y \sim 1 + D + X_1 + X_2 + X_1 \times X_2$
- Classic **omitted variable bias**
  - Remember: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i$ doesn't specify functional form
- With binary/discrete X, mis-specification alone usually isn't a huge problem
  - [Maybe ML can help? We'll come back to this]

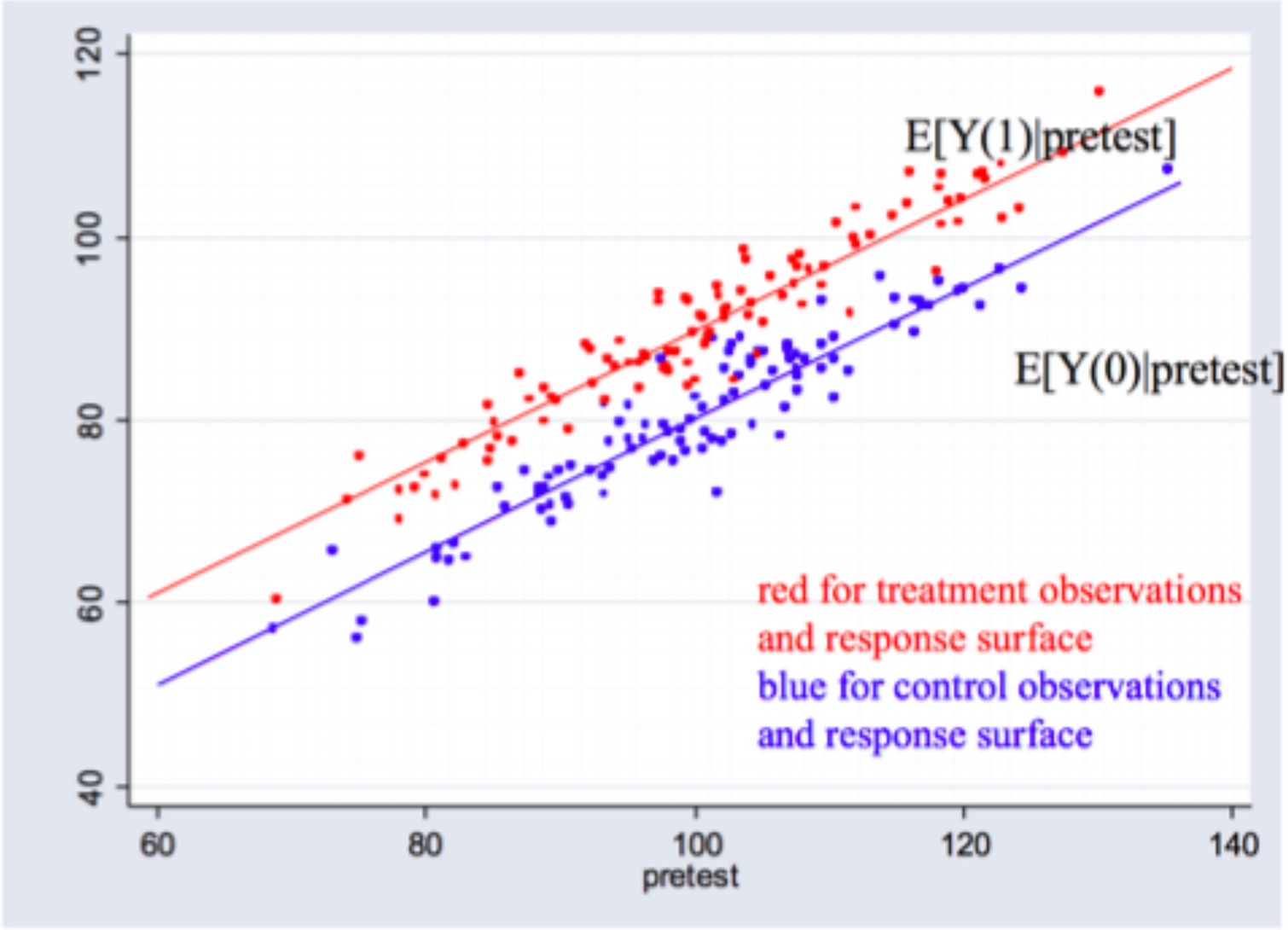
Issue #2: Extrapolation



Issue #2: Extrapolation

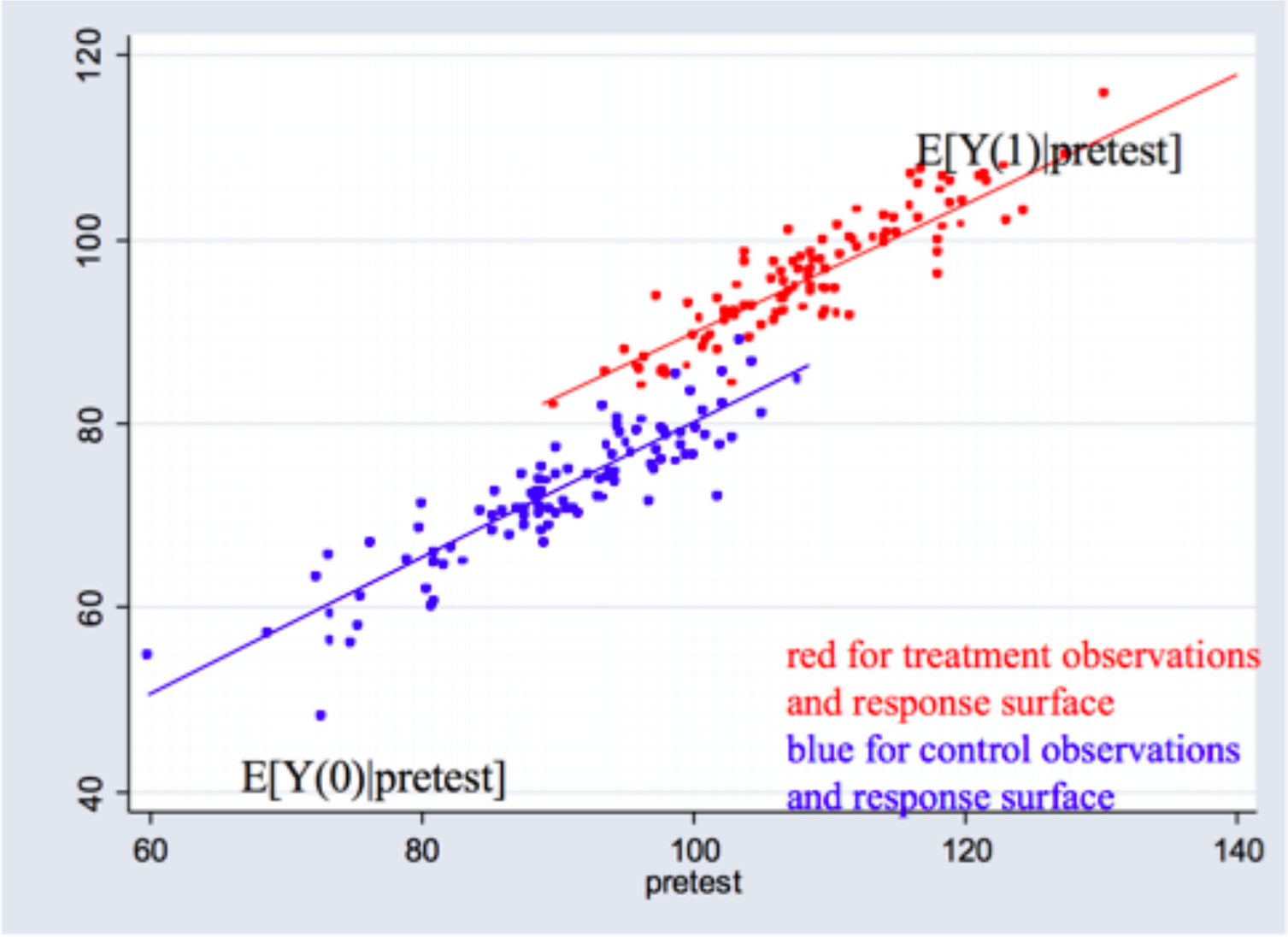


RCT: Regression only for precision



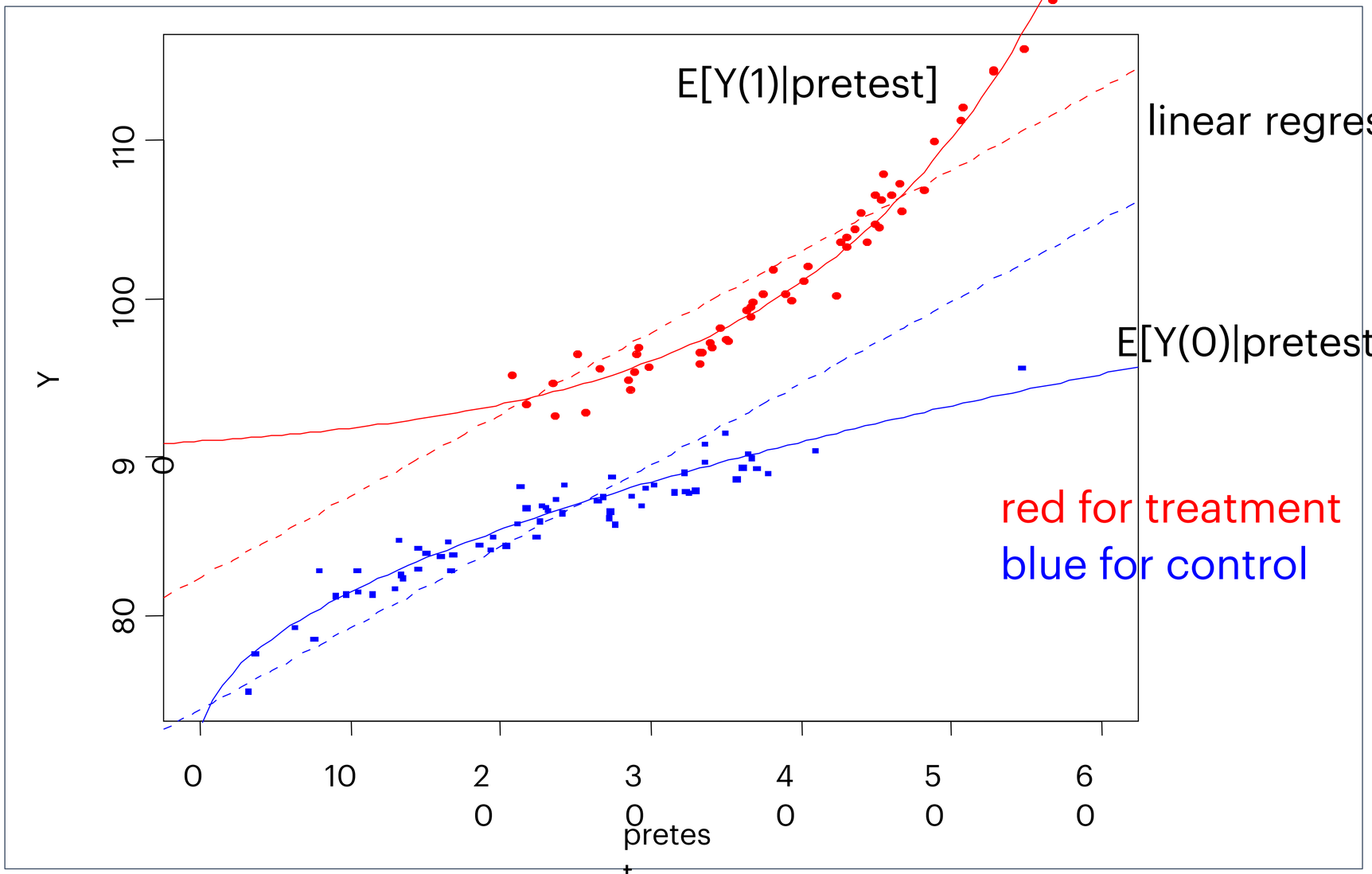
Source: Hill

Obs study: One confounder + linearity



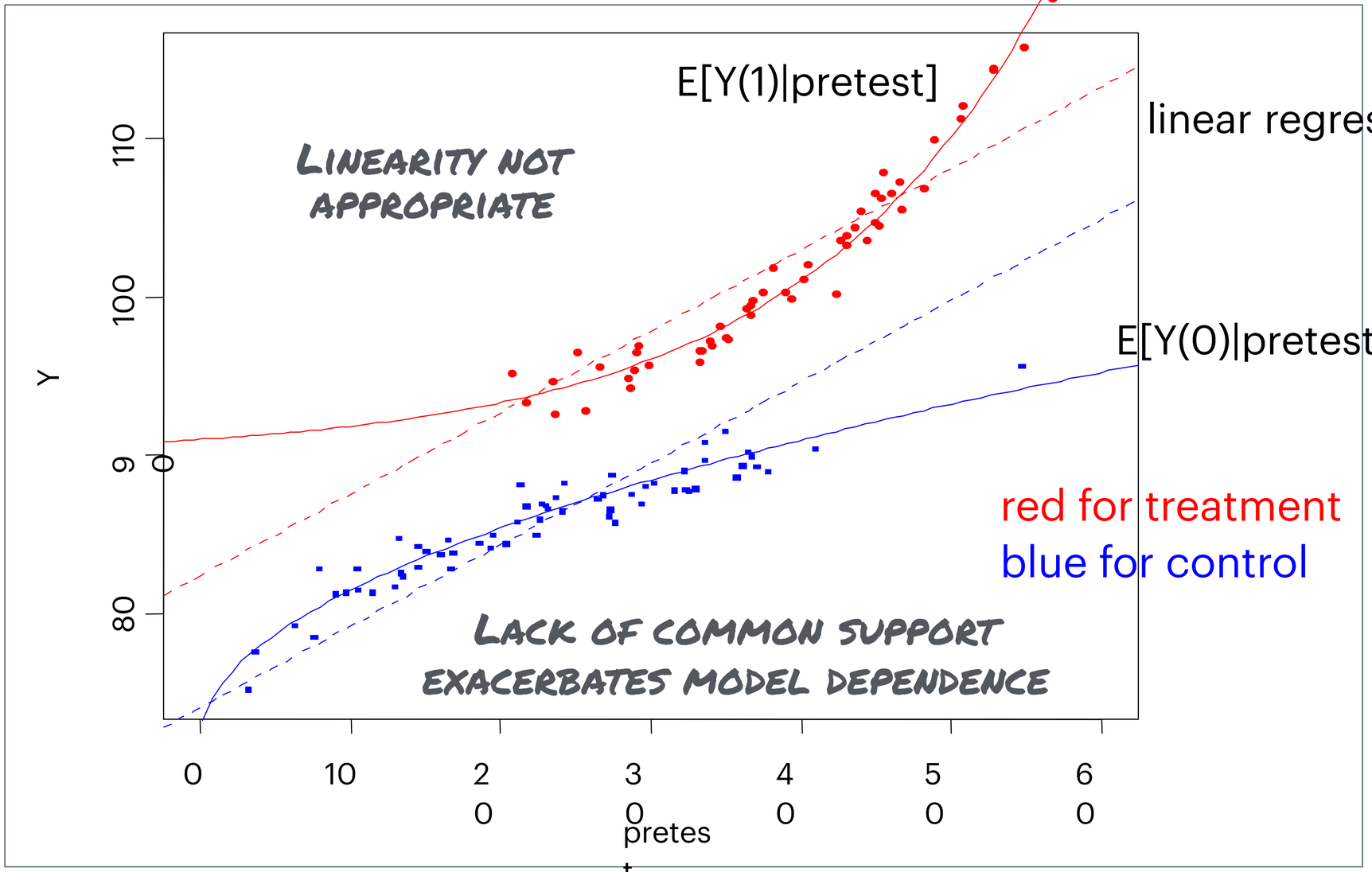
Source: Hill

Obs study: One confounder, no linearity



Source: Hill

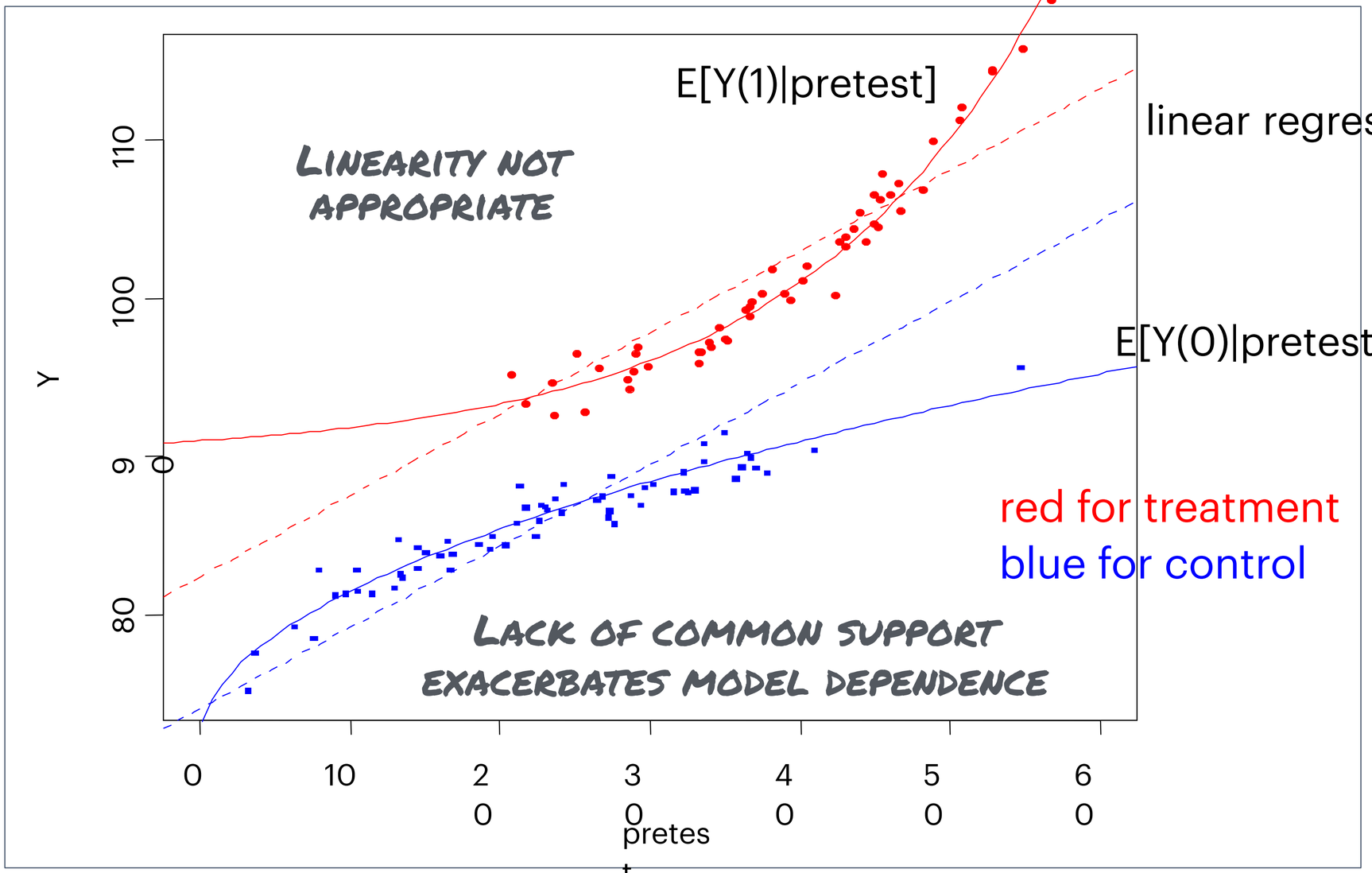
Obs study: One confounder, no linearity



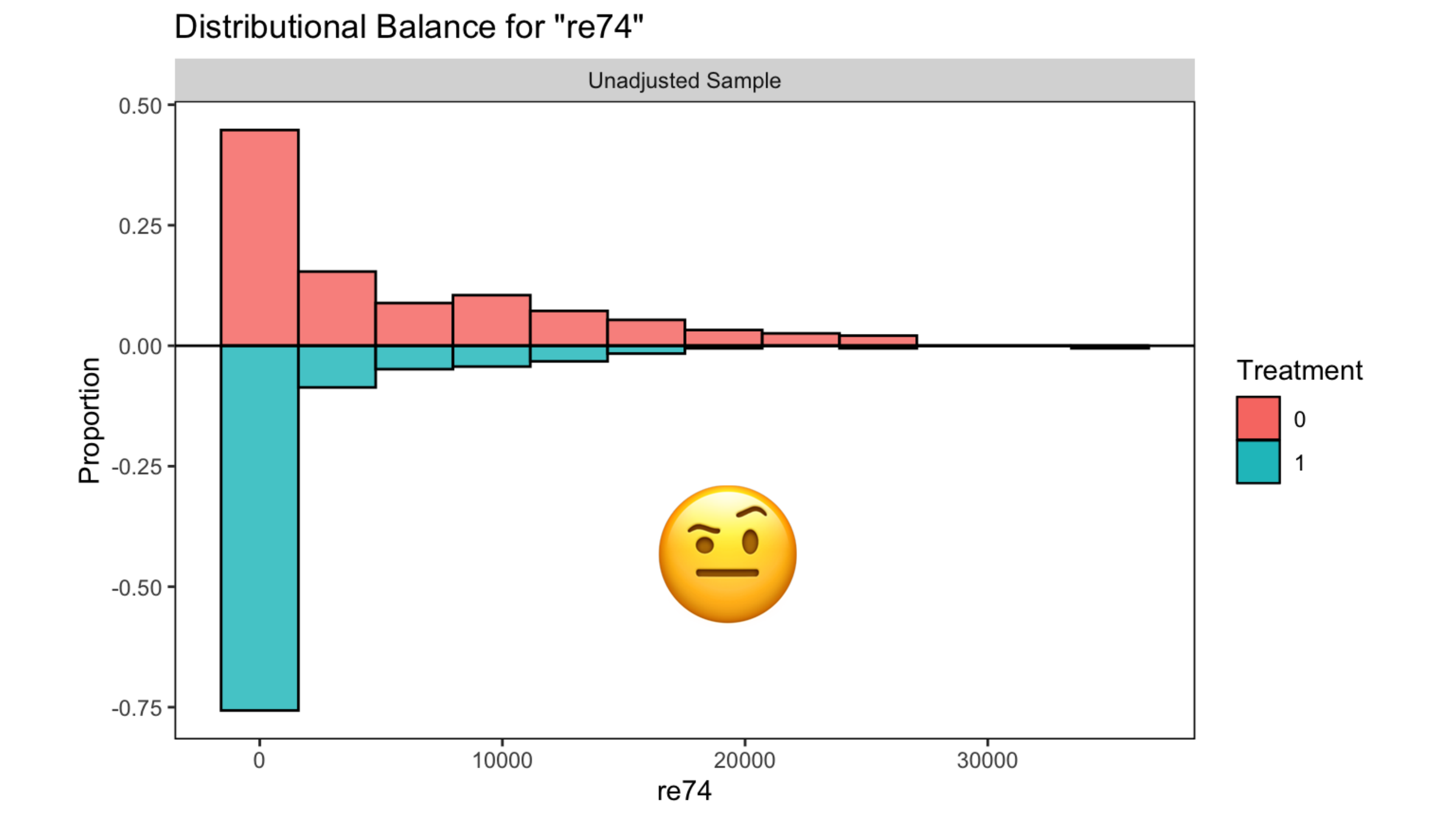
Source: Hill

Obs study: One confounder, no linearity

NO RED FLAGS!



Source: Hill


Distributional Balance for "re74"
Unadjusted Sample
0.50
0.25
0.00
-0.25
-0.50
-0.75
Proportion
0
10000
20000
30000
re74
Treatment
0
1

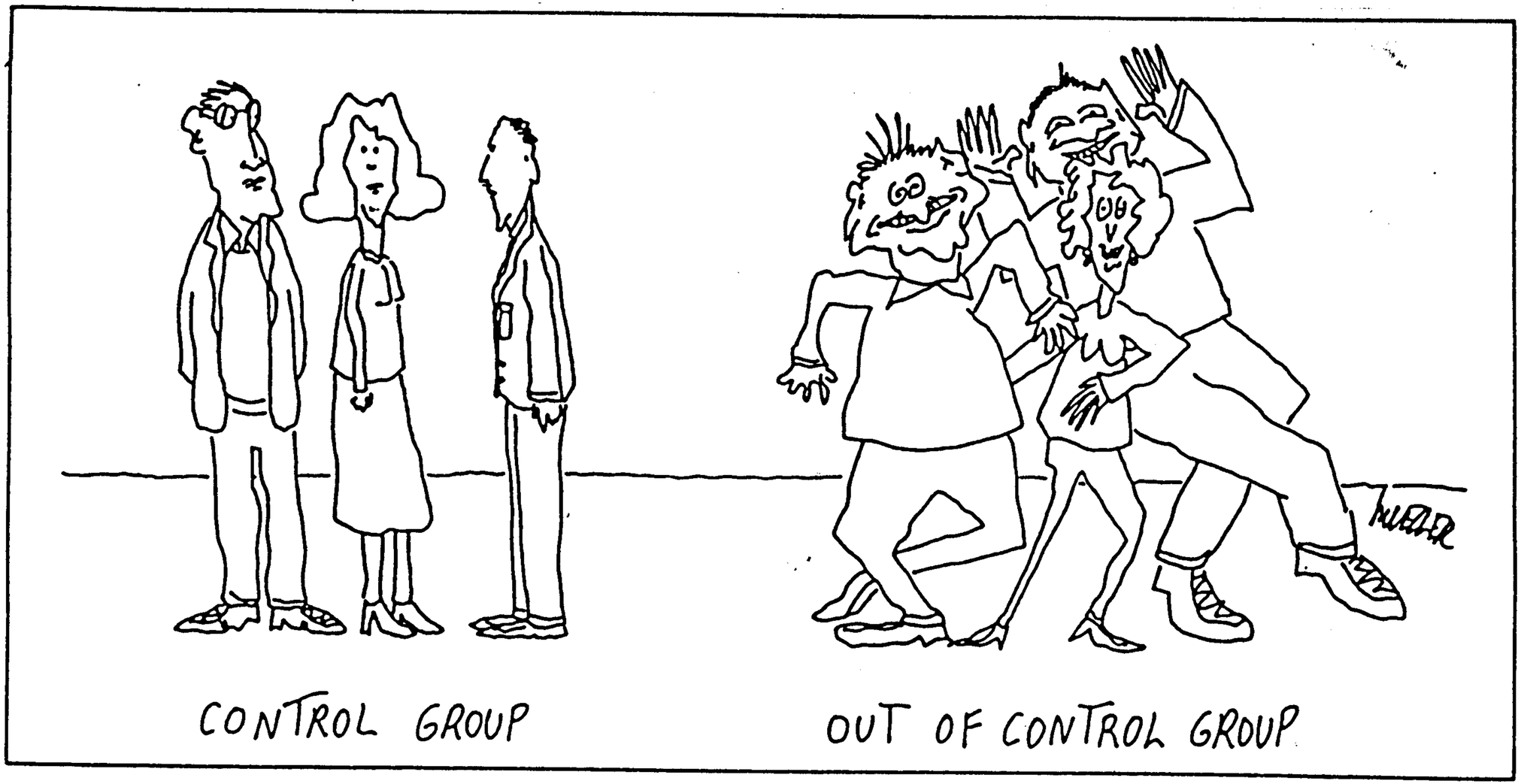
LaLonde: Regression adjustment 😞

Specification	Full Samples		
	NSW	CPS-1	CPS-3
	(1)	(2)	(3)
Raw Difference	1,794 (633)	-8,498 (712)	-635 (657)
Demographic controls	1,670 (639)	-3,437 (710)	771 (837)
1975 Earnings	1,750 (632)	-78 (537)	-91 (641)
Demographics, 1975 Earnings	1,636 (638)	623 (558)	1,010 (822)
Demographics, 1974 and 1975 Earnings	1,676 (639)	794 (548)	1,369 (809)

(From MHE Table 3.3.3. Demographics are age, years of schooling, dummies for Black, Hispanic, high school dropout, and married)

Source: Borusyak, MHE

Matching


CONTROL GROUP
OUT OF CONTROL GROUP

"Planners of observational studies should always ask themselves: How would the study be conducted if it were possible to do it by controlled experimentation?"

- Cochran (1965)

design observational study to **approximate experiments**
[i.e., to make them less terrible; aka. "target trials" or "trial emulation"]

Matching ⇝ Blocking with $n = 2$

- **Blocking:** find (possibly many) units with same X
  - What if X takes many values? (e.g., age)
  - What if we have many Xs?

- **Matching:** for each treated unit, find (at least one) control unit with similar X
  - Approximate randomized trial with many blocks of size $n = 2$
  - Start with **exact matching** if possible; usually have **inexact matching** in practice
  - Don't (necessarily) care about the matches themselves ⇝ want good balance overall

- Many choices, no right way to do matching (can be frustrating!)

Warmup: 1-1 matching with replacement

Warmup: 1-1 matching with replacement

- *Repeat*: for every treated unit
  - Find the control unit with the closest value of X
    [if there are 2+ good matches, choose one at random]

Warmup: 1-1 matching with replacement

- *Repeat*: for every treated unit
  - Find the control unit with the closest value of X
    [if there are 2+ good matches, choose one at random]

- That's it... 🤷

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
Average:		

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
Average:		

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
Average:		

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
17	29	6200
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
17	29	6200
16	24	9700
15	25	23300
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
17	29	6200
16	24	9700
15	25	23300
1	43	20900
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
17	29	6200
16	24	9700
15	25	23300
1	43	20900
8	28	8800
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
17	29	6200
16	24	9700
15	25	23300
1	43	20900
8	28	8800
4	27	9300
Average:		

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
17	29	6200
16	24	9700
15	25	23300
1	43	20900
8	28	8800
4	27	9300
8	28	8800
Average:		

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

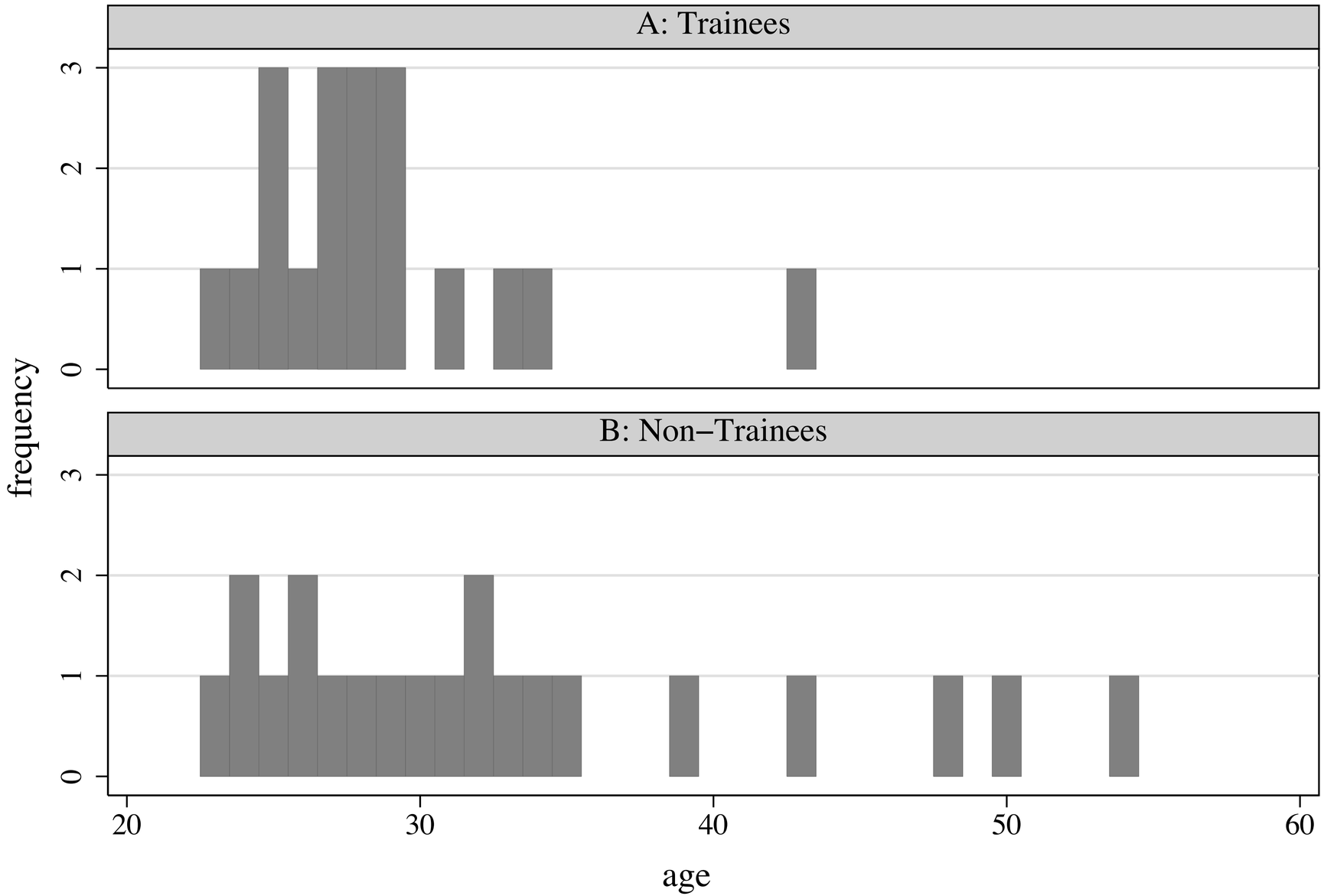
Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

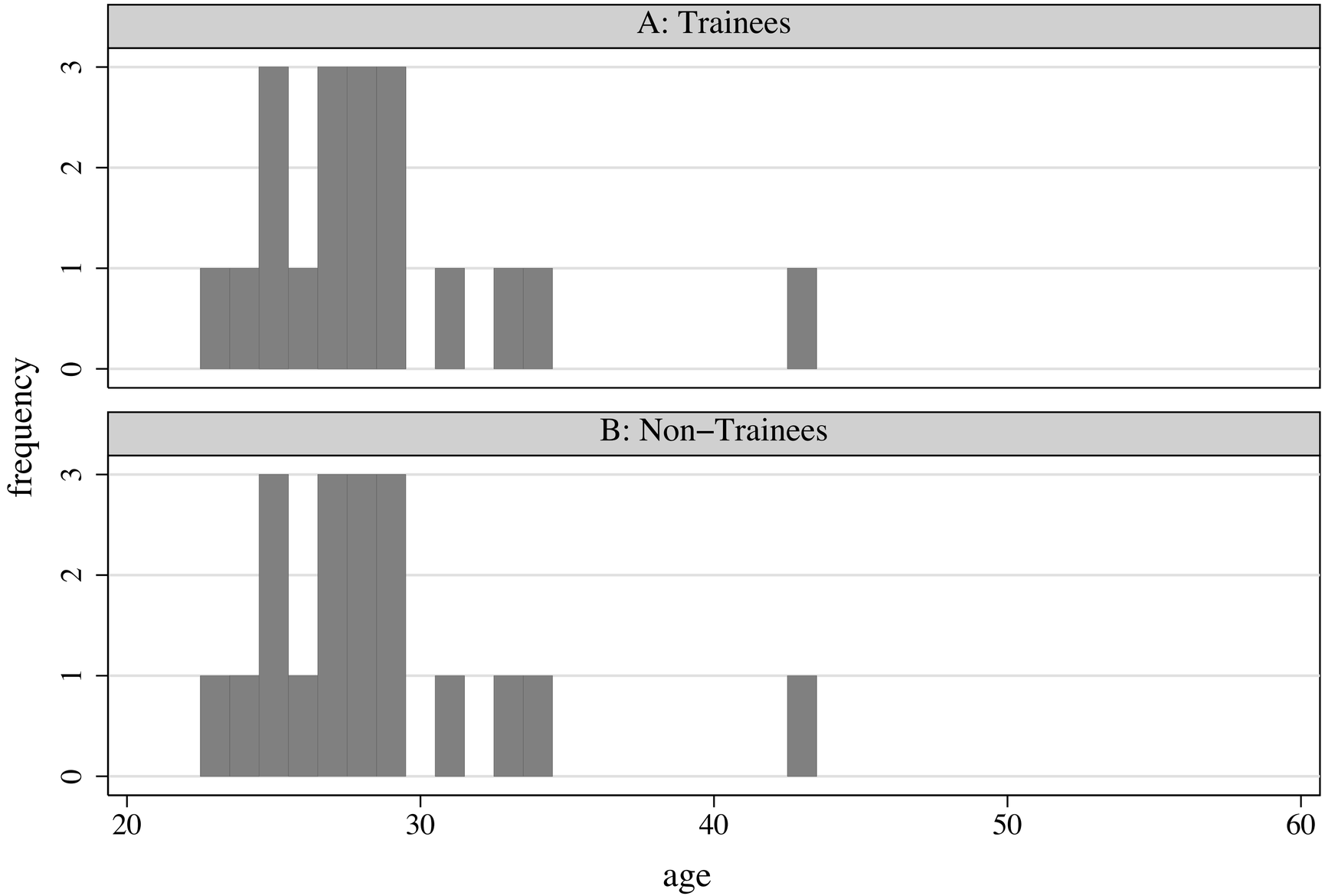
Matched Sample

unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
13	26	16500
15	25	23300
4	27	9300
17	29	6200
16	24	9700
15	25	23300
1	43	20900
8	28	8800
4	27	9300
8	28	8800
Average:	28.5	13982

Age Distribution: Before Matching



Age Distribution: After Matching



Treatment effect estimates

- Difference in average earnings between trainees and non-trainees:

- *Before matching*: $16,426 – $20,724 = **–$4, 298**
- *After matching*: $16,426 – $13, 982 = **–$2, 444**

- [use diff-in-means here; could also use regression after matching]

What's the estimand?

- Find matches for each **treated unit**
  - i.e., trying to get X to look like distribution for treated group
- ⇝ **Average Treatment Effect on the Treated (ATT)**

$$\begin{aligned} \text{ATT} &= \mathbb{E}[Y(1) - Y(0) \mid D = 1] \\ &= \mathbb{E}[Y(1) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 1] \end{aligned}$$

$\bar{Y}_{\text{treated}}$ $\bar{Y}_{\text{matched controls}}$

- *Still need:* no unmeasured confounding given X, overlap

Matching Assumptions

Selection on Observables Assumption:

Prince Charles

Male
Born in 1948
Raised in the UK
Married Twice
Lives in a castle
Wealthy and Famous

Ozzy Osbourne

Male
Born in 1948
Raised in the UK
Married Twice
Lives in a castle
Wealthy and Famous

THANKS, JAKE!

Matching complexities

- Example above **1:1 matching with replacement**
- Even here, many decisions to make!
  - What if there are multiple possible matches? [avg them? pick one?]
  - What if there are multiple treated units with the same X?
- *Biggest question: what if we can't find a match??*

Matching when we can't find an exact match

- What to do when a treated unit doesn't have an exact match on X in the control group?
- **Option 1:** Discard unit
  - Then perfectly matched sample — but **we've changed the estimand**!
  - [More on this when we get to propensity scores]
- **Option 2:** Accept a "close" match
  - Have to decide what is "close enough"? Introduce imbalance in X $\rightsquigarrow$ possible imbalance in $Y(0)$
- With more Xs, harder to find an exact match [*curse of dimensionality*]

Distance measures for matching ☣

How do we measure "closeness" between units?

- **Euclidean distance:**

$$\text{dist}(i,j) = \sqrt{\sum_k (X_{ki} - X_{kj})^2}$$

- **Normalized Euclidean distance:**

$$\text{dist}(i,j) = \sqrt{\sum_k \frac{(X_{ki} - X_{kj})^2}{\text{var}[X_k]}}$$

- **Mahalanobis distance:**

$$\text{dist}(i,j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)' \widehat{\Sigma}_X^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

(Propensity scores coming up...)

Diagnostics: Covariate Balance

- **Key idea:** A “good match” is a match that yields good covariate balance
- Important to assess covariate balance for every match
  - Differences in means are most obvious
  - But also interested in variances, overall shape
  - People often use p-values, F-stat, etc.
- If poor balance, then match again!

(Stylized) Matching Process

A little convoluted...

- Check overall balance and common support.
- Match.
- Check overall balance.
- Re-match with adjustment.
- Finalize balance.
- Estimate treatment effect.

Typically via regression ("bias correction")

Back to LaLonde

We construct a matched sample in which we exactly match on *years of education.* What can we say about covariate imbalance on *age*?

- The matched sample will be exactly balanced on age.
- The matched sample will be approximately balanced on age.
- The matched sample will be heavily imbalanced on age.
- We can’t say anything about covariate imbalance on age before looking at the data.

We construct a matched sample in which we exactly match on *years of education.* What can we say about covariate imbalance on *age*?

The matched sample will be exactly balanced on age.

0%

The matched sample will be approximately balanced on age.

0%

The matched sample will be heavily imbalanced on age.

0%

We can’t say anything about covariate imbalance on age before looking at the data.

0%

We construct a matched sample in which we exactly match on *years of education.* What can we say about covariate imbalance on *age*?

The matched sample will be exactly balanced on age.

0%

The matched sample will be approximately balanced on age.

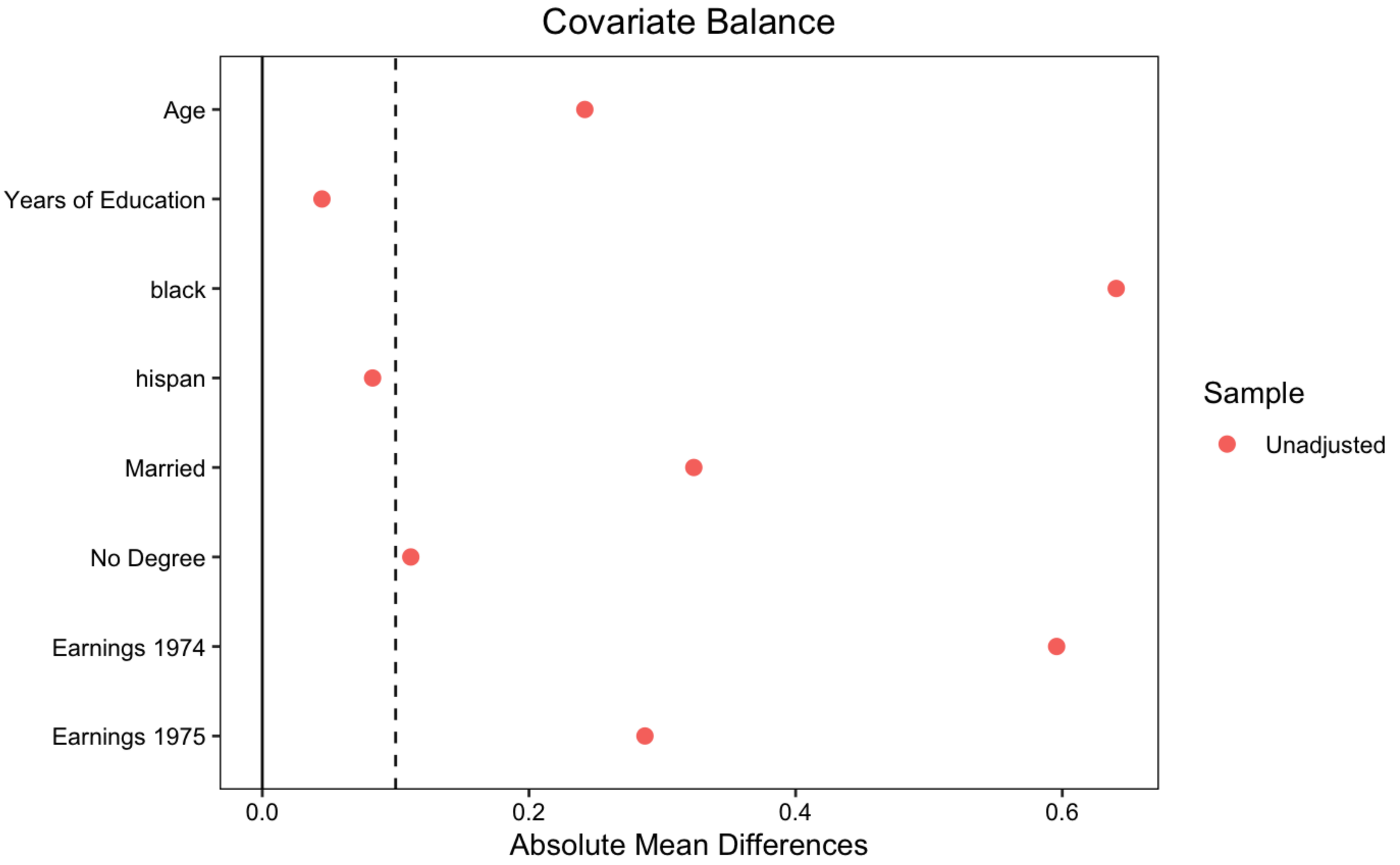
0%

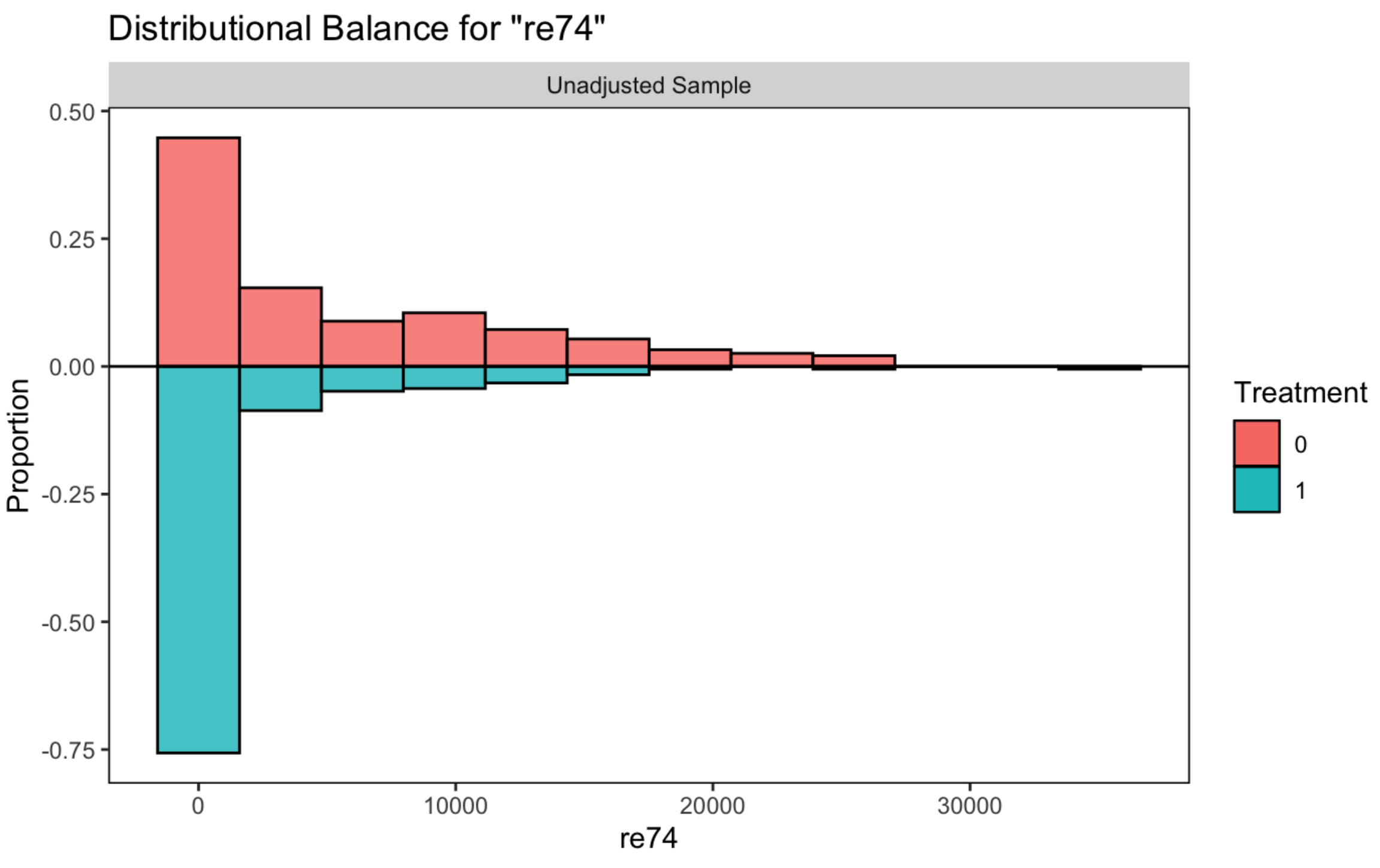
The matched sample will be heavily imbalanced on age.

0%

We can’t say anything about covariate imbalance on age before looking at the data.

0%


Covariate Balance
Age
Years of Education
black
hispan
Married
No Degree
Earnings 1974
Earnings 1975
0.0
0.2
0.4
0.6
Absolute Mean Differences
Sample
Unadjusted

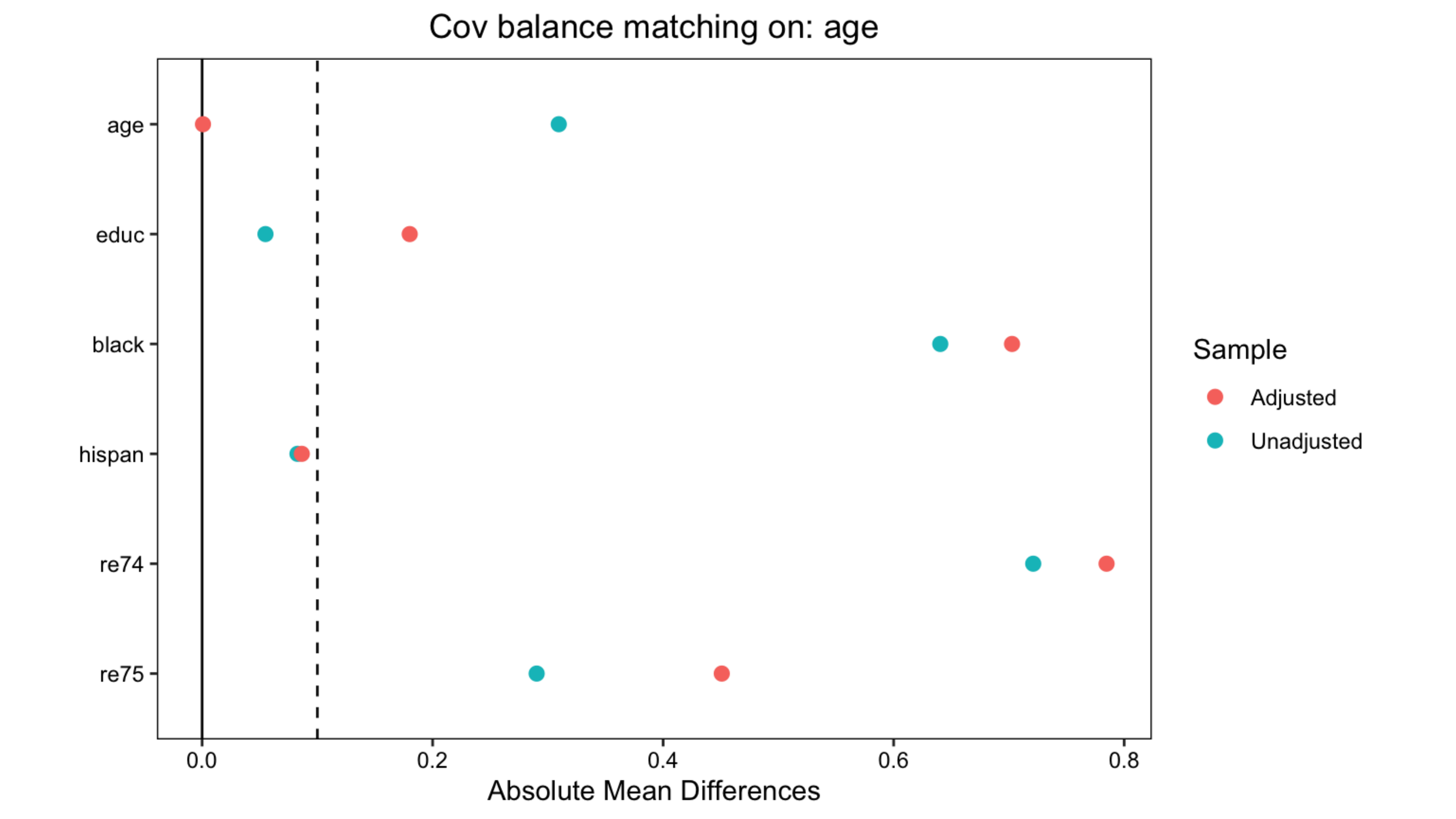

Distributional Balance for "re74"
Unadjusted Sample
0.50
0.25
0.00
-0.25
-0.50
-0.75
Proportion
0
10000
20000
30000
re74
Treatment
0
1

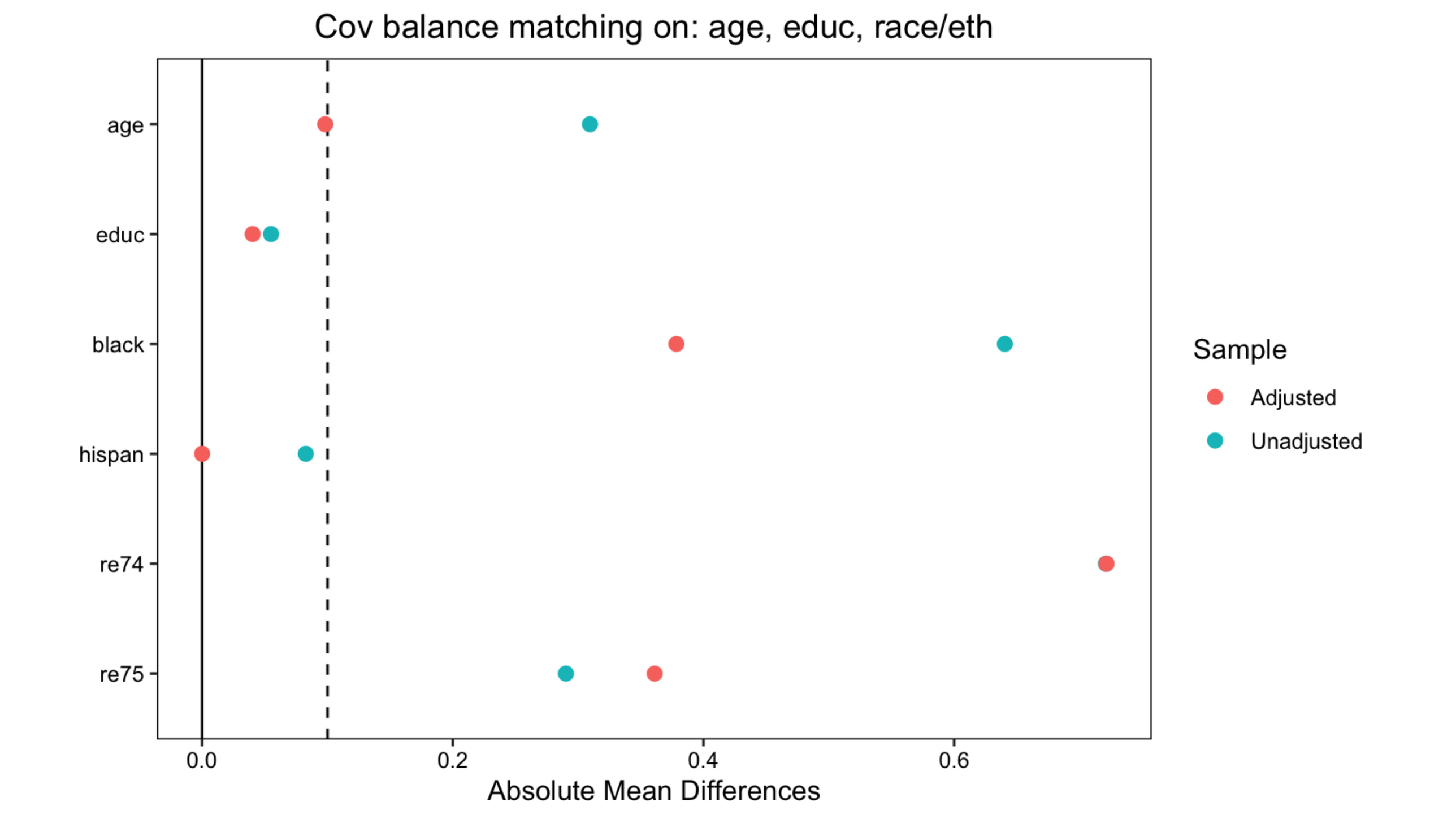
Let's Match!

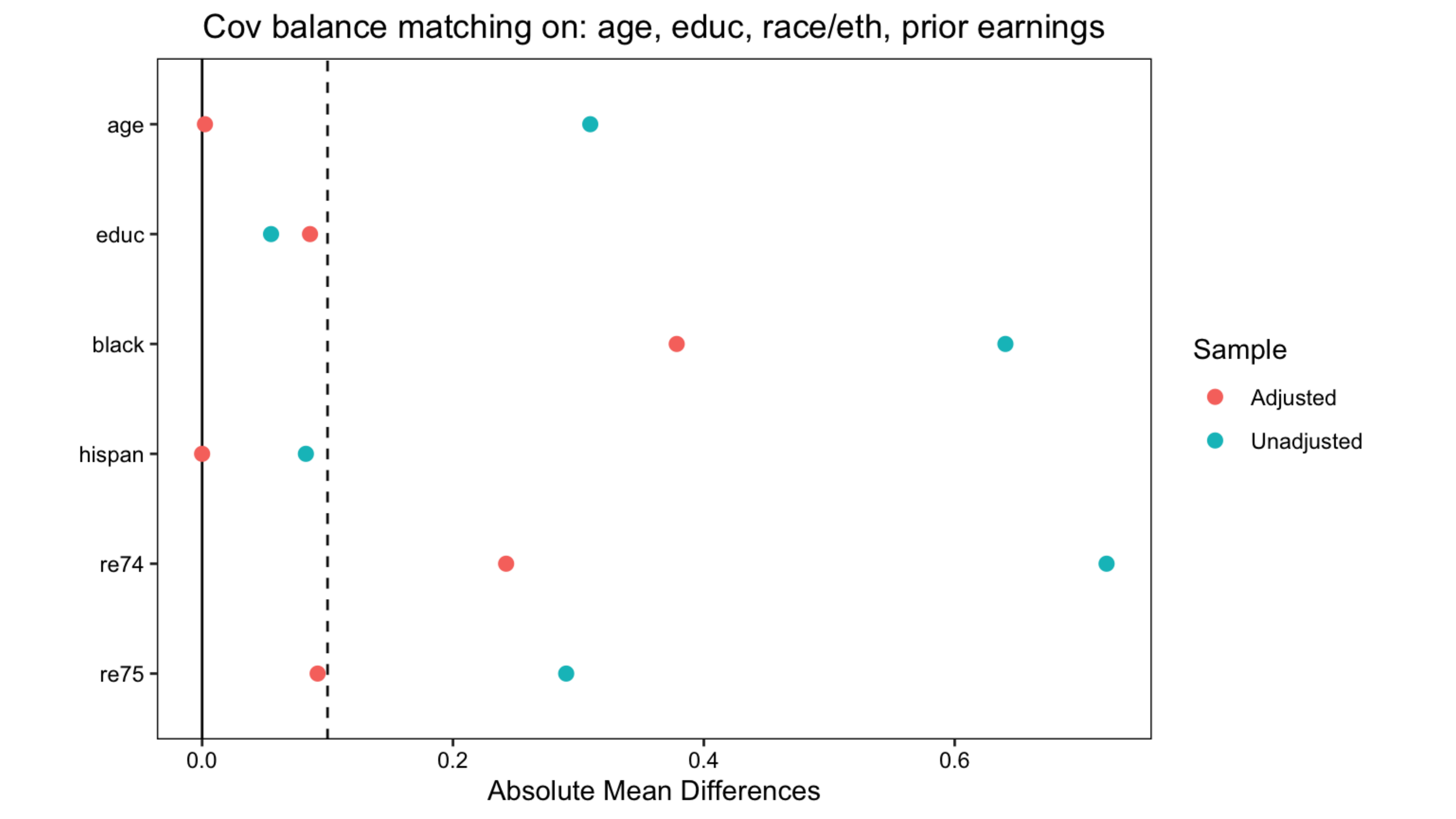
To start: **1:1 nearest neighbor matching without replacement**

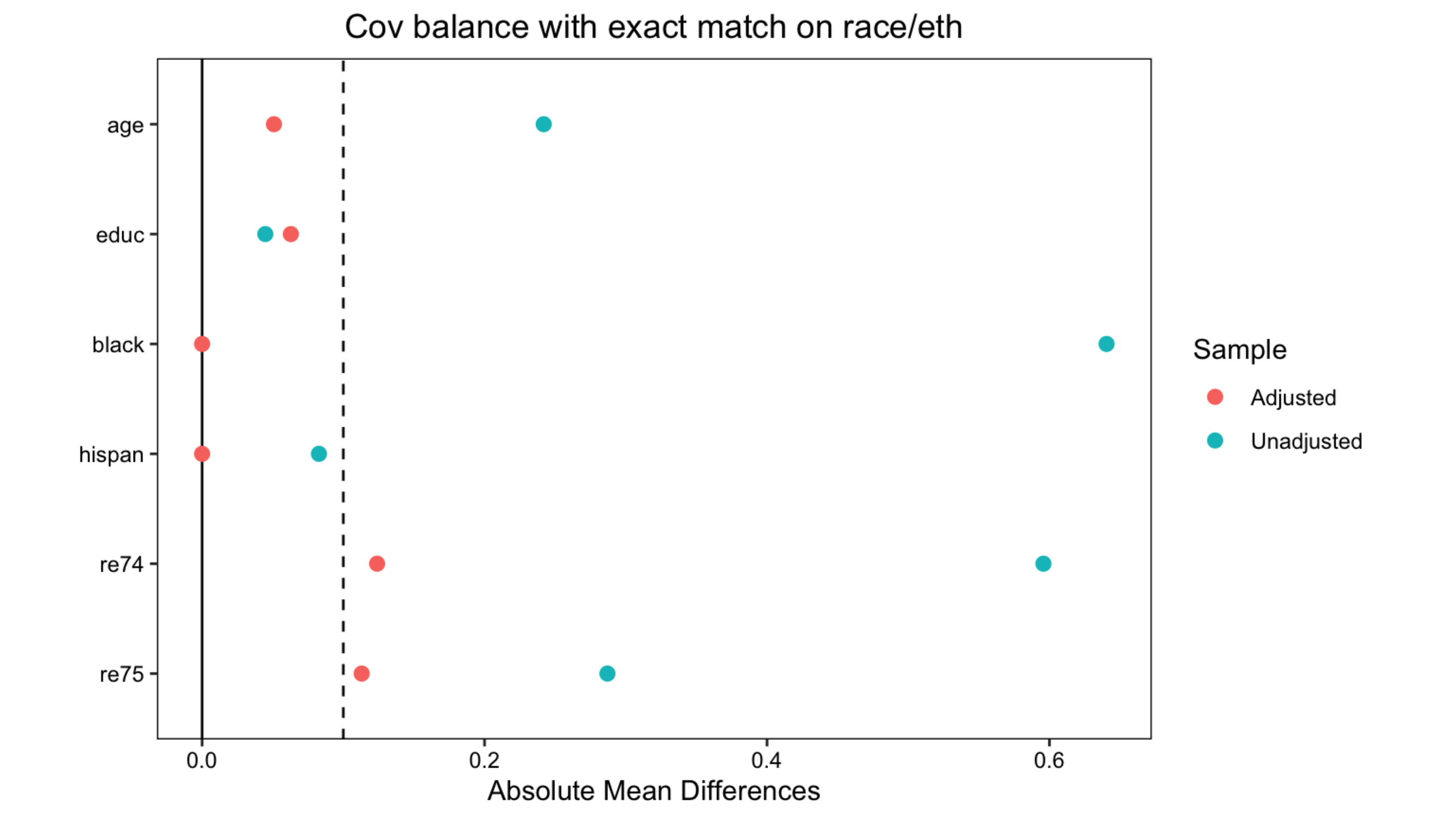
Practical questions:

- What should we use to compute "distance"?
- Should we "exact" match on anything?
  (i.e., only match treated units with a HS degree to control units with a HS degree)


Cov balance matching on: age
age
educ
black
hispan
re74
re75
0.0
0.2
0.4
0.6
0.8
Absolute Mean Differences
Sample
Adjusted
Unadjusted


Cov balance matching on: age, educ, race/eth
age
educ
black
hispan
re74
re75
0.0
0.2
0.4
0.6
Absolute Mean Differences
Sample
Adjusted
Unadjusted


Cov balance matching on: age, educ, race/eth, prior earnings
age
educ
black
hispan
re74
re75
0.0
0.2
0.4
0.6
Absolute Mean Differences
Sample
Adjusted
Unadjusted


Cov balance with exact match on race/eth
age
educ
black
hispan
re74
re75
0.0
0.2
0.4
0.6
Absolute Mean Differences
Sample
Adjusted
Unadjusted

Dropping units

When we only do 1:1 nearest neighbor matching:

- Treated units: 185 (out of 185)
- Control units: 185 (out of 429)

When we also exact match on race/ethnicity:

- Treated units: 116 (out of 185)
- Control units: 116 (out of 429)

Why?

	Control	Treat
Non-Black	342	29
Black	87	156

Target trial

Our target trial is a matched pair RCT where we match on:

- Race/ethnicity
- Age
- Education
- Prior earnings

Many other possible target trials

We like our match. Now what?

Option 1: simple difference-in-means on matched sample

→ ***+$997*** *(robust se = $867)*

Option 2: regression-adjusted estimate on matched sample
(aka. “bias correction”)

→ ***+$998*** *(robust se = $874)*

Next up, the propensity score...