PHP 2610 HW 2

Blain Morin

December 21, 2018

1. Let Y denote real income in 1978 and let T denote treatment group. Fit the model:

$$E(Y|T) = \beta_0 + \beta_1 Treated$$

(a) Report the estimates of beta 0 and beta 1.

Table 1: Observational Regression

	Dependent variable:
	re78
treat	−635.026
	(657.137)
Constant	6,984.170***
	(360.710)
Observations	614
R^2	0.002
Adjusted R^2	−0.0001
Residual Std. Error	7,471.134 (df = 612)
F Statistic	0.934 (df = 1; 612)
Note:	*p<0.1; **p<0.05; ***p<0.01

(b) What does the coefficient beta 1 represent?

β_1 is the average difference in real earnings between the people who had job training and those who did not. On average, the people who had job training earned $635.03 less than the people who did not have job training.

(c) Can it be interpreted as a causal effect? Why or why not?

Because our data is essentially observational, we cannot make a claim of causality here in this basic regression. We have not accounted for any confounding. There could be important differences between the control and treatment group that affect income.

2. Use propensity score matching to estimate the causal effect of job training. Select and justify the propensity score model and the method of matching that you ultimately decide to use. Please hand in the following:

(a) A description of the propensity score method, matching method, and analysis method that you use to estimate the causal effect. A few sentences is fine here.

In the following sections, I use propensity score matching in order to balance the distribution of possible confounding variables in the treatment and control arms. I use the nearest neighbor method from the "MatchIt" package. After matching one observation from the control arm to one observation from the treatment arm, I run two linear models to estimate the causal effect of treatment.

(b) The chunk of R code (or other code) that performs the matching and carries out the analysis. Not the output here, just the code.

Here is the R code that I use to perform the matching and carry out the analysis:

```r
### Nearest neighbor matching
match.nn = matchit(treat ~ age +
                 educ +
                 as.factor(black) +
                 as.factor(hispan) +
                 as.factor(married) +
                 as.factor(nodegree) +
                 re74 +
                 re75,
               data = lalonde,
               method = "nearest",
               distance = "logit")


### Get matched data
match.nn.data = match.data(match.nn)


### Run the model
match.nn.model = lm(re78 ~ treat, data = match.nn.data)

### Add other covariates
match.nn.model2 = lm(re78 ~ treat +
                       age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                     data = match.nn.data)
```

(c) A table that shows the numbers matched and not matched, and a summary of covariate distributions in the treated and control groups. For continuous variables, the summary could be (n, mean, standard deviation), or it could be (n, median, quantiles). For binary variables it should just be n and proportion. Please no graphs for this one.

This table displays the numbers matched and not matched:

Table 2: Number of Observations Matched and Unmatched

	Control	Treated
All	429	185
Matched	185	185
Unmatched	244	0
Discarded	0	0

This next table shows the balance of the matched data. For the continuous variables, I present the mean and standard deviation. For the binary variables, I only present the mean (which is interpretted as the proportion):

Table 3: Covariate Distributions: Treatment vs Control

	Control	Treated
Mean Age	25.303	25.816
Mean Education	10.605	10.346
Mean Black	0.470	0.843
Mean Hispanic	0.216	0.059
Mean Married	0.211	0.189
Mean NoDegree	0.638	0.708
Mean 74 Income	2,342.108	2,095.574
Mean 75 Income	1,614.745	1,532.055
sd Age	10.586	7.155
sd Education	2.658	2.011
sd 74 Income	4,238.976	4,886.620
sd 75 Income	2,632.353	3,219.251
n	185	185

(d) Output from the regression model or analysis method that you use to estimate the causal effect.

For model (1), I run the same simple regression as in Question 1 with the matched data. For model (2), I include additional covariates to increase the efficiency of the model:

Table 4: Regression using Matched Data

	Dependent variable:	
	re78	
	(1)	(2)
treat	908.202	1,351.960*
	(730.724)	(790.065)
age		8.797
		(42.931)
educ		608.310***
		(224.151)
as.factor(black)1		−431.414
		(1,010.292)
as.factor(hispan)1		1,108.572
		(1,300.666)
as.factor(married)1		−160.210
		(986.663)
as.factor(nodegree)1		931.682
		(1,110.315)
re74		0.026
		(0.103)
re75		0.222
		(0.160)
Constant	5,440.941***	−2,248.762
	(516.700)	(3,513.036)
Observations	370	370
R^2	0.004	0.049
Adjusted R^2	0.001	0.026
Residual Std. Error	7,027.876 (df = 368)	6,942.358 (df = 360)
F Statistic	1.545 (df = 1; 368)	2.078** (df = 9; 360)
Note:		*p<0.1; **p<0.05; ***p<0.01

(e) Report of the estimated causal effect, its standard error, and an interpretation of what the causal effect represents.

Using the regression results from model (2), the average causal effect of worker training is an $1,351.96 increase in real income (all else equal). The standard error is 790.07. The effect is significant at the .1 level.

3 G-Estimation

(a)

Because the variables other than treatment and re78 are potential confounders, I use all of them in my regressions. First, I regress re78 on all the possible confounding variables using the treatment arm data. Then, I regress re78 on all the possible confounding variables using the control arm data. I then generate predictions using both models on the treatment arm data. The difference in the two predictions is the estimated causal effect.

Here is the R code for the regressions:

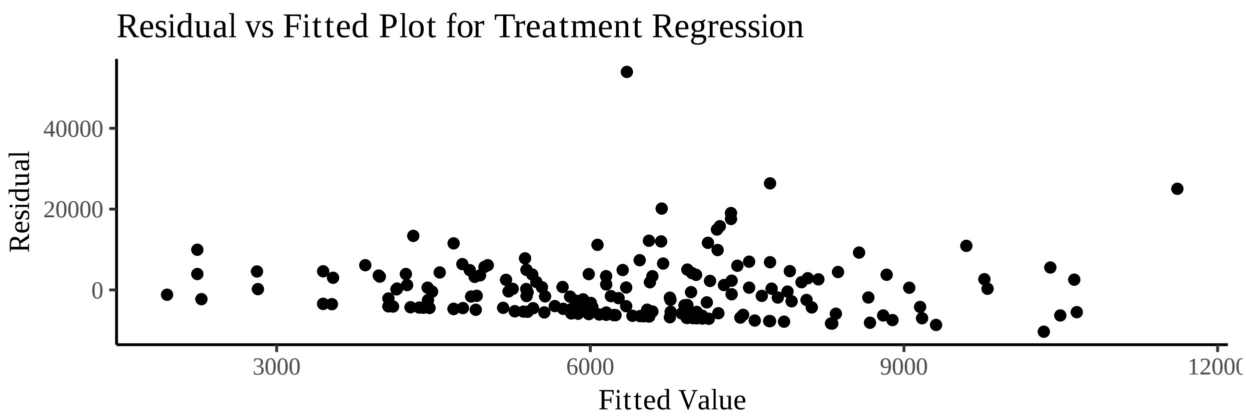
```
### Regression on the treated group
treated.model = lm(re78 ~ age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                   data = lalonde,
                   subset = (treat == 1))

### Regression on the control group
control.model = lm(re78 ~ age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                   data = lalonde,
                   subset = (treat == 0))
```

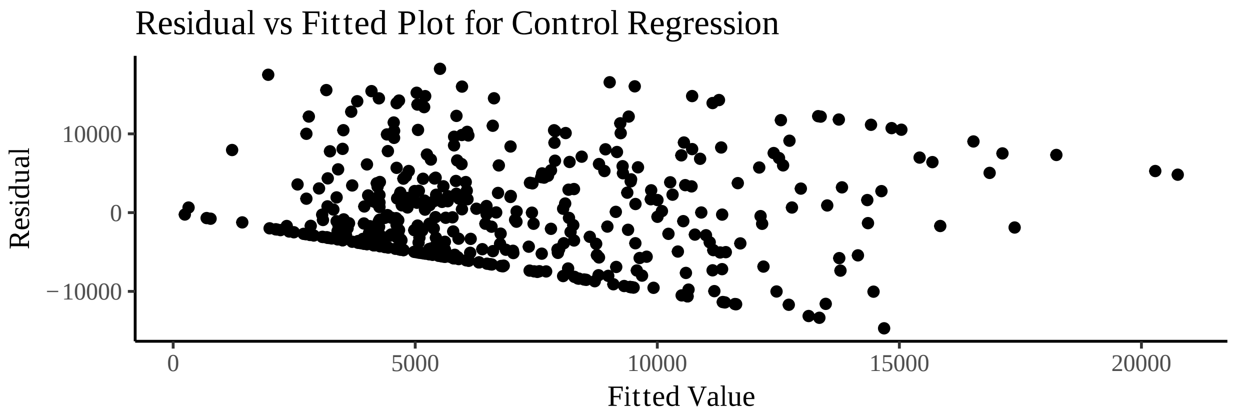
Here are the regression results for each model:

Table 5: G Estimation Regressions

	Dependent variable:	
	re78	
	Treated Model	Control Model
	(1)	(2)
age	83.562	−20.266
	(85.387)	(34.163)
educ	623.961	322.458*
	(395.287)	(170.291)
as.factor(black)1	−1,140.014	−1,167.542
	(1,981.176)	(826.335)
as.factor(hispan)1	304.327	496.088
	(3,057.645)	(944.890)
as.factor(married)1	1,032.440	−52.686
	(1,589.880)	(753.142)
as.factor(nodegree)1	−319.009	480.373
	(1,807.132)	(949.268)
re74	0.039	0.377***
	(0.160)	(0.061)
re75	0.089	0.340***
	(0.244)	(0.114)
Constant	−1,508.424	1,203.509
	(6,006.589)	(2,651.368)
Observations	185	429
R^2	0.050	0.228
Adjusted R^2	0.007	0.213
Residual Std. Error	7,840.566 (df = 176)	6,470.717 (df = 420)
F Statistic	1.158 (df = 8; 176)	15.483*** (df = 8; 420)
Note:		*p<0.1; **p<0.05; ***p<0.01

Here are residual vs fitted plots for both models:





There appears to be a pattern in the residuals for both the treatment regression and the control regression. This should be investigate in more detail, as we are violating some of the linear model assumptions.

(b)

Here are 10 random rows from the data with their corresponding Y1 and Y0 predictions:

Table 6: 10 Random Observations with Y1 and Y0 Predictions

	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78	y1	y0
94	1	21	12	1	0	0	0	0	0	9, 983.784	6, 593.891	3, 479.867
57	1	37	9	1	0	0	1	0	0	1, 067.506	5, 739.988	2, 668.605
79	1	40	11	1	0	0	1	0	0	23, 005.600	7, 238.595	3, 252.721
127	1	22	12	1	0	0	0	5, 605.852	936.177	0	6, 981.749	5, 889.450
16	1	19	10	1	0	0	1	0	0	3, 228.503	4, 859.836	3, 355.857
41	1	21	12	1	0	0	0	0	0	1, 254.582	6, 593.891	3, 479.867
50	1	23	12	1	0	0	0	0	0	4, 843.176	6, 761.015	3, 439.335
49	1	25	5	1	0	0	1	0	0	12, 187.410	2, 241.403	1, 621.969
110	1	26	10	1	0	1	1	2, 027.999	0	0	6, 557.244	3, 925.270
76	1	25	11	0	0	0	1	0	0	18, 783.350	7, 125.182	4, 724.259

(c)

Here is the R code used to calculate the causal effect:

```r
### Filter to get treatment arm
treatment.arm = lalonde %>%
  filter(treat == 1)

### Use treatment model to get y1
```

```r
y1 = predict(treated.model, newdata = treatment.arm)

### Use control model to get y0
y0 = predict(control.model, newdata = treatment.arm)

### Calculate causal Effect
mean(y1 - y0)
```

```
## [1] 1647.583
```

The estimated causal effect of worker training is a $1,647.58 average increase in real wages (all else equal).

(d) Bonus 1: Bootstrap for standard error

```r
set.seed(10)

treatment.arm = lalonde %>%
  filter(treat == 1)

control.arm = lalonde %>%
  filter(treat == 0)

### Number of bootstrap replications
sims = 1000

### Initialize a vector to store results
causal.effects = rep(NA, sims)

### Bootstrap

for (i in 1:sims) {

  bootstrap.sample.treat = treatment.arm %>%
    sample_n(nrow(treatment.arm), replace = TRUE)

  bootstrap.sample.control = control.arm %>%
    sample_n(nrow(control.arm), replace = TRUE)

  treated.model = lm(re78 ~ age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                     data = bootstrap.sample.treat)

  control.model = lm(re78 ~ age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                   data = bootstrap.sample.control)

  y1 = predict(treated.model, newdata = bootstrap.sample.treat)
  y0 = predict(control.model, newdata = bootstrap.sample.treat)

  causal.effects[i] = mean(y1 - y0)





}



sd(causal.effects)
```

```
## [1] 807.0761
```

We find the standard error of the G estimation estimate to be 807.08.

Appendix: R code

```
### Set knitr options
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)

### Load Required Libraries

library(dplyr)
library(MatchIt)
library(stargazer)
library(tidyr)
library(purrr)
library(ggplot2)
library(gridExtra)
library(extrafont)

### Load data
data("lalonde")


### Run simple model

model1 = lm(re78 ~ treat, data = lalonde)

### Create Table
stargazer(model1, header = FALSE,
          title = "Observational Regression")



### Nearest neighbor matching
match.nn = matchit(treat ~ age +
                 educ +
                 as.factor(black) +
                 as.factor(hispan) +
                 as.factor(married) +
                 as.factor(nodegree) +
                 re74 +
                 re75,
               data = lalonde,
               method = "nearest",
               distance = "logit")


### Get matched data
match.nn.data = match.data(match.nn)


### Run the model
match.nn.model = lm(re78 ~ treat, data = match.nn.data)

### Add other covariates
match.nn.model2 = lm(re78 ~ treat +
                       age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                     data = match.nn.data)


matched.table = match.nn$nn

stargazer(matched.table, header = FALSE,
          title = "Number of Observations Matched and Unmatched",
          table.placement = 'H')


covariate.dist = match.nn.data %>%
  select(-re78, -distance, -weights) %>%
  group_by(treat) %>%
  summarise_all(funs(mean, sd))

covariate.dist = t(round(covariate.dist,3))

covariate.dist = covariate.dist[-1,]

covariate.dist = as.data.frame(covariate.dist) %>%
  rename(Control = V1, Treated = V2)

covariate.dist[17,] = c(185, 185)

rownames(covariate.dist)[17] = "n"

covariate.dist = covariate.dist[-(11:14),]

row.names(covariate.dist) = c("Mean Age",
                              "Mean Education",
                              "Mean Black",
                              "Mean Hispanic",
                              "Mean Married",
                              "Mean NoDegree",
                              "Mean 74 Income",
                              "Mean 75 Income",
                              "sd Age",
                              "sd Education",
                              "sd 74 Income",
                              "sd 75 Income",
                              "n")

stargazer(covariate.dist, header = FALSE,
          table.placement = 'H',
          title = "Covariate Distributions: Treatment vs Control",
          summary = FALSE)


stargazer(match.nn.model, match.nn.model2,
          header = FALSE, table.placement = 'H',
          title = "Regression using Matched Data")



### Regression on the treated group
treated.model = lm(re78 ~ age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                   data = lalonde,
                   subset = (treat == 1))

### Regression on the control group
control.model = lm(re78 ~ age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                   data = lalonde,
                   subset = (treat == 0))



### Regression table
stargazer(treated.model, control.model, header = FALSE,
          table.placement = 'H',
          title = "G Estimation Regressions",
          column.labels = c("Treated Model",
                            "Control Model"))


### Residual vs Fitted plots
treated.plot = ggplot(treated.model) +
  geom_point(aes(x = .fitted, y = .resid)) +
  ylab("Residual") +
  xlab("Fitted Value") +
  ggtitle("Residual vs Fitted Plot for Treatment Regression") +
  theme_classic() +
  theme(text=element_text(size=11,  family="CM Sans"))

control.plot = ggplot(control.model) +
  geom_point(aes(x = .fitted, y = .resid)) +
  ylab("Residual") +
  xlab("Fitted Value") +
  ggtitle("Residual vs Fitted Plot for Control Regression") +
  theme_classic() +
  theme(text=element_text(size=11,  family="CM Sans"))

grid.arrange(treated.plot, control.plot, nrow = 2)



treatment.arm = lalonde %>%
  filter(treat == 1)

y1 = predict(treated.model, newdata = treatment.arm)

y0 = predict(control.model, newdata = treatment.arm)

treatment.arm = cbind(treatment.arm, y1, y0)

set.seed(10)
for.table = sample_n(treatment.arm, size = 10)

stargazer(for.table, header = FALSE,
          summary = FALSE,
          table.placement = 'H',
          title = "10 Random Observations with Y1 and Y0 Predictions",
          column.sep.width = "3pt",
          font.size = "tiny")


### Filter to get treatment arm
treatment.arm = lalonde %>%
  filter(treat == 1)

### Use treatment model to get y1
y1 = predict(treated.model, newdata = treatment.arm)

### Use control model to get y0
y0 = predict(control.model, newdata = treatment.arm)

### Calculate causal Effect
mean(y1 - y0)



set.seed(10)

treatment.arm = lalonde %>%
  filter(treat == 1)

control.arm = lalonde %>%
  filter(treat == 0)
```

```
### Number of bootstrap replications
sims = 1000

### Initialize a vector to store results
causal.effects = rep(NA, sims)

### Bootstrap

for (i in 1:sims) {

  bootstrap.sample.treat = treatment.arm %>%
    sample_n(nrow(treatment.arm), replace = TRUE)

  bootstrap.sample.control = control.arm %>%
    sample_n(nrow(control.arm), replace = TRUE)

  treated.model = lm(re78 ~ age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                   data = bootstrap.sample.treat)

  control.model = lm(re78 ~ age +
                       educ +
                       as.factor(black) +
                       as.factor(hispan) +
                       as.factor(married) +
                       as.factor(nodegree) +
                       re74 +
                       re75,
                   data = bootstrap.sample.control)

  y1 = predict(treated.model, newdata = bootstrap.sample.treat)
  y0 = predict(control.model, newdata = bootstrap.sample.treat)

  causal.effects[i] = mean(y1 - y0)





}



sd(causal.effects)
```