FLS 6415: Replication 8 - Matching

May 2020

To be submitted (code + answers) by midnight, Wednesday 13th May.

First read the paper by Boas and Hidalgo (2011) on the course website. For this replication we will focus on the *second half* of their paper, not the initial RDD but the matching analysis of how possession of a radio licence affects the mayor's vote share in the next election.

The replication data is in the file *Boas_Hidalgo.csv*. A list of available variables is also provided below.

Variable	Description
pctVV	The councillor's vote share in the 2004 elections
treat	Whether a councillor that applied for a media licence received approval before the 2004 election
male	Councillor is male
log.valid.votes	Log of the size of the electorate (proxied by valid votes)

1. What is treatment? What is control? What is the outcome?

Treatment is a Councillor that applied for a media licence and received approval before the free media period of the 2004 election. Control is a Councillor who applied but was not approved by that time. The outcome variable is the Councillor's vote share in the 2004 election.

2. Why do Boas and Hidalgo not use an experiment or natural experiment to estimate the effect of possessing a radio licence?

A pure experiment is impossible since the researchers were (ethically, practically and financially) unable to randomly distribute radio licences. A natural experiment depends on some random or 'as-if' random variation in receipt of a radio licence, but the authors argue that no such variation exists; there is no discontinuity or instrument. Therefore, it is not feasible to use an alternative methodology.

3. Conduct and interpret a basic linear regression of the outcome on treatment with no controls.

```
d <- read_csv("Boas_Hidalgo.csv")

d %>% lm(pctVV ~ treat, data = .) %>% stargazer(header = F, title = "Q3")
```

Councillors that receive approved media licences before the 2004 elections are associated with a 0.453 % points increase in vote share in the 2004 election. However, this effect is not causal.

4. One potential confounding variable is gender (this could affect the chances of an application being approved if there is bias in the Ministry, and the candidate's vote share if there is bias among voters). Is there balance across control and treatment groups on the `male` variable?

```
d %>% t.test(male ~ treat, data = .)
```

```
##
##  Welch Two Sample t-test
##
```

Table 2: Q3

	Dependent variable:
	pctVV
treat	0.453***
	(0.137)
Constant	2.296***
	(0.063)
Observations	1,455
R^2	0.007
Adjusted R^2	0.007
Residual Std. Error	2.139 (df = 1453)
F Statistic	10.964*** (df = 1; 1453)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
## data:  male by treat
## t = -2.3996, df = 587.22, p-value = 0.01673
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.076928905 -0.007678609
## sample estimates:
## mean in group 0 mean in group 1
##       0.8837413       0.9260450
```

No. The treatment group has 4.23% points more men than the control group, which is a statistically significant difference with a p-value of 0.017.

5. One way of controlling for gender is to add it as a control variable to your regression in Q3. Interpret the result.

```
d %>% lm(pctVV ~ treat + male, data = .) %>% stargazer(header = F, single.row = T,
    title = "Q5")
```

Table 3: Q5

	Dependent variable:
	pctVV
treat	0.446*** (0.137)
male	0.175 (0.182)
Constant	2.141*** (0.172)
Observations	1,455
R^2	0.008
Adjusted R^2	0.007
Residual Std. Error	2.139 (df = 1452)
F Statistic	5.945*** (df = 2; 1452)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimated treatment effect reduces a little once we control for gender.

6. An alternative approach is to use matching. Let's try to do one-to-one exact matching on gender *manually*. There are 311 treated units but 1144 control units in your data, so one-to-one matching means *throwing away* 833 control units.
(a) Split your data into four different datasets: treated males, treated females, control males and control females;
(b) How many treated males do you have? Reduce your dataset of control males so you have only the same number as the number of treated males - since they are exactly matched on gender it doesn't matter which you pick so choose which ones to keep/drop randomly;
(c) Do the same for control females - reduce the number of control females to the same as the number of treated females;
(d) Join your four datasets back together to make one dataset (this will be smaller than the original dataset as we threw some data away);
(e) Check for balance in gender on the new dataset - it should be perfectly balanced, right?

```
d_treated_male <- d %>% filter(treat == 1 & male == 1)
d_treated_female <- d %>% filter(treat == 1 & male == 0)
set.seed(123)
d_control_male <- d %>% filter(treat == 0 & male == 1) %>% sample_n(dim(d_treated_male)[1])
set.seed(123)
d_control_female <- d %>% filter(treat == 0 & male == 0) %>% sample_n(dim(d_treated_female)[1])

d_matched_gender <- bind_rows(d_treated_male, d_treated_female, d_control_male,
    d_control_female)

d_matched_gender %>% t.test(male ~ treat, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  male by treat
## t = 0, df = 620, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04127914  0.04127914
## sample estimates:
## mean in group 0 mean in group 1
##        0.926045        0.926045
```

There is now perfect balance on gender in the matched dataset.

7. Using the matched dataset from Q6, conduct two analyses of the difference in outcomes between treated and control groups. One using a difference-in-means t-test and one using a simple linear regression. Interpret the results.

```
d_matched_gender %>% t.test(pctVV ~ treat, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  pctVV by treat
## t = -2.5774, df = 604.25, p-value = 0.01019
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7835116 -0.1058510
## sample estimates:
## mean in group 0 mean in group 1
```

```
##        2.304123        2.748805
```

```
d_matched_gender %>% lm(pctVV ~ treat, data = .) %>% stargazer(header = F, single.row = T,
    title = "Q7")
```

Table 4: Q7

	Dependent variable:
	pctVV
treat	0.445** (0.173)
Constant	2.304*** (0.122)
Observations	622
R^2	0.011
Adjusted R^2	0.009
Residual Std. Error	2.151 (df = 620)
F Statistic	6.643** (df = 1; 620)
Note:	*p<0.1; **p<0.05; ***p<0.01

The two methods give identical results. The value of the treatment variable has changed a little, but not by a large amount.

8. To match on continuous or multiple variables it's easier to use `matchit`.

(a) Return to your original full dataset and, using nearest neighbour matching, match only on the size of the electorate (*log.valid.votes*).

(b) How many units are matched? Why this number?

```
matched_data_Q8 <- matchit(treat ~ log.valid.votes, data = d, method = "nearest")
```

622 units are matched - one control unit for each treated unit, and the rest of the control units are dropped.

(c) Conduct a simple balance t-test on the size of the electorate for the full dataset and for your matched dataset (you can recover it with `match.data(output_of_matchit)`). How does balance change after matching?

```
d %>% t.test(log.valid.votes ~ treat, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  log.valid.votes by treat
## t = 2.1829, df = 558.4, p-value = 0.02946
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01860591 0.35279203
## sample estimates:
## mean in group 0 mean in group 1
##        10.11921         9.93351
```

```
matched_data_Q8 %>% match.data() %>% t.test(log.valid.votes ~ treat, data = .)
```

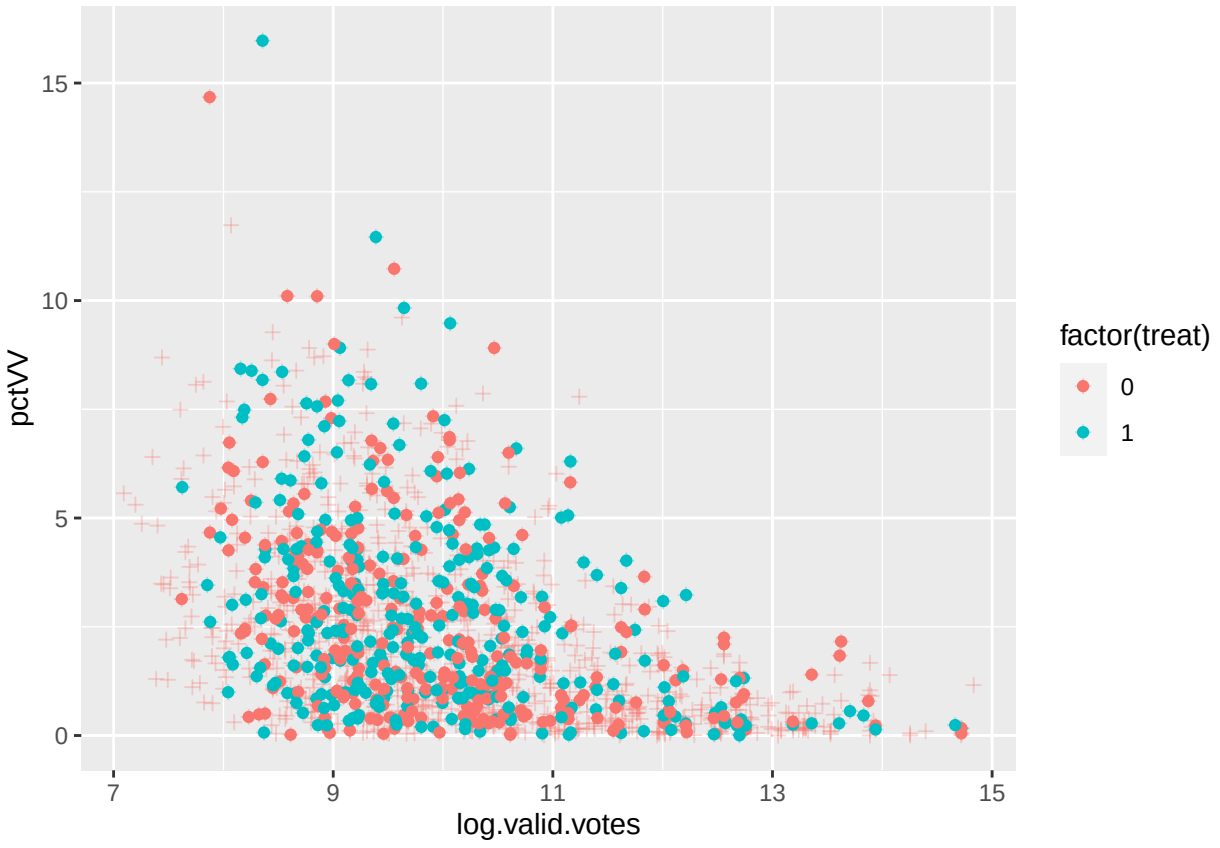
```
##
##  Welch Two Sample t-test
##
## data:  log.valid.votes by treat
```

```
## t = 0.0072001, df = 620, p-value = 0.9943
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2013492  0.2028311
## sample estimates:
## mean in group 0 mean in group 1
##        9.934251        9.933510
```

The size of the electorate is imbalanced in the full dataset but almost perfectly balanced in the matched dataset.

9. Let's see which units were dropped by our matching method in Q8. For the full (unmatched) dataset, create a graph of the size of the electorate against the outcome variable. Colour the points according to treatment status. Make this layer semi-transparent (adjust the 'alpha' of your graph in R) if you can so we can see all the points. Finally, add another layer to your graph showing the same variables for the *matched* data but with a different shape so we can distinguish them. What does this graph tell you about which units were matched?

```
d %>% ggplot() + geom_point(aes(x = log.valid.votes, y = pctVV, colour = factor(treat)),
    alpha = 0.3, shape = 3) + geom_point(data = matched_data_Q8 %>% match.data(),
    aes(x = log.valid.votes, y = pctVV, colour = factor(treat)))
```



The control units furthest from the treated units, especially those with low values of `log.valid.votes`, are dropped.

10. Using the matched dataset from Q8, conduct two analyses of the difference in outcomes between treated and control groups. One using a difference-in-means t-test and one using a simple linear regression. Interpret the results.

```
matched_data_Q8 %>% match.data() %>% t.test(pctVV ~ treat, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  pctVV by treat
## t = -1.1859, df = 617.3, p-value = 0.2361
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5671819  0.1400802
## sample estimates:
## mean in group 0 mean in group 1
##        2.535254        2.748805
```

```
matched_data_Q8 %>% match.data() %>% lm(pctVV ~ treat, data = .) %>% stargazer(header = F,
    title = "Q10, Nearest Neighbour Matching on Size of the Electorate", single.row = T)
```

Table 5: Q10, Nearest Neighbour Matching on Size of the Electorate

	Dependent variable:
	pctVV
treat	0.214 (0.180)
Constant	2.535*** (0.127)
Observations	622
R^2	0.002
Adjusted R^2	0.001
Residual Std. Error	2.246 (df = 620)
F Statistic	1.406 (df = 1; 620)
Note:	*p<0.1; **p<0.05; ***p<0.01

The results are identical - the treated group has a higher vote share than the control group but the difference is not statistically significant.

11. Now let's include all of the matching variables that Boas and Hidalgo use, and use nearest neighbour matching in `matchit` to construct a matched dataset. Use the list of matching variables provided below to conduct nearest neighbour matching.

"occBlue.collar", "occEducation", "occGovernment", "occMedia", "occNone", "occOther", "occPolitician", "occWhite.collar", "lat", "long", "ran.prior", "incumbent", "log.valid.votes", "party.prior.pctVV", "prior.pctVV", "elec.year", "match.partyPCB", "match.partyPC.do.B", "match.partyPDT", "match.partyPFL", "match.partyPL", "match.partyPMDB", "match.partyPMN", "match.partyPP", "match.partyPPS", "match.partyPSB", "match.partyPSC", "match.partyPSDB", "match.partyPSDC", "match.partyPSL", "match.partyPT", "match.partyPTB", "match.partyPV", "uf.rs", "uf.sp", "yob", "eduMore.than.Primary..Less.than.Superior", "eduSome.Superior.or.More", "log.total.assets", "pt_pres_1998", "psdb_2000", "hdi_2000", "income_2000", "log.num.apps"

```
covars <- c("occBlue.collar", "occEducation", "occGovernment", "occMedia", "occNone",
    "occOther", "occPolitician", "occWhite.collar", "lat", "long", "ran.prior",
    "incumbent", "log.valid.votes", "party.prior.pctVV", "prior.pctVV", "elec.year",
    "match.partyPCB", "match.partyPC.do.B", "match.partyPDT", "match.partyPFL",
    "match.partyPL", "match.partyPMDB", "match.partyPMN", "match.partyPP", "match.partyPPS",
    "match.partyPSB", "match.partyPSC", "match.partyPSDB", "match.partyPSDC",
    "match.partyPSL", "match.partyPT", "match.partyPTB", "match.partyPV", "uf.rs",
```

```
    "uf.sp", "yob", "eduMore.than.Primary..Less.than.Superior", "eduSome.Superior.or.More",
    "log.total.assets", "pt_pres_1998", "psdb_2000", "hdi_2000", "income_2000",
    "log.num.apps")

covars_formula <- paste0(covars, collapse = " + ")

matched_data_Q11 <- matchit(as.formula(paste0("treat~", covars_formula)), data = d,
    method = "nearest")
```

12. Using your matched dataset from Q11, conduct a simple linear regression of the outcome on treatment. Interpret the results and compare them to the result in the first column of Table 4 in Boas and Hidalgo (2011) (it probably won't be the same, see the next questions).

```
matched_data_Q11 %>% match.data() %>% lm(pctVV ~ treat, data = .) %>% stargazer(header = F,
    title = "Q12, Nearest Neighbour Matching, All Variables, No Controls", single.row = T)
```

Table 6: Q12, Nearest Neighbour Matching, All Variables, No Controls

	Dependent variable:
	pctVV
treat	0.154 (0.178)
Constant	2.594*** (0.126)
Observations	622
R^2	0.001
Adjusted R^2	−0.0004
Residual Std. Error	2.220 (df = 620)
F Statistic	0.753 (df = 1; 620)
Note:	*p<0.1; **p<0.05; ***p<0.01

The results suggest that possessing a radio licence is associated with 0.15% points higher vote share in the 2004 election, but the difference is not statistically significant and is very different from the result in Boas and Hidalgo.

13. With lots of variables it's impossible to get perfect balance on all variables, there are just too many dimensions and too few units. One option to control for 'residual confounding' is to include the matching variables as control variables in our analysis regression. How does this change your estimated treatment effect from Q12?

```
matched_data_Q11 %>% match.data() %>% lm(as.formula(paste0("pctVV ~ treat + ",
    covars_formula)), data = .) %>% stargazer(header = F, title = "Q13, Nearest Neighbour Matching, All
    single.row = T)
```

The estimated treatment effect is now marginally larger but remains insignificant.

14. One risk with nearest-neighbour matching is that the control unit can still be far away from the treated unit if there are no good matches. Re-run the matching process from Q11 but with a caliper of 0.01 standard deviations, and then re-run the regression from Q12 (no controls). How does the number of units and the result change?

```
set.seed(123)
matched_data_Q14 <- matchit(as.formula(paste0("treat~", covars_formula)), data = d,
    method = "nearest", caliper = 0.01)

matched_data_Q14 %>% match.data() %>% lm(pctVV ~ treat, data = .) %>% stargazer(header = F,
```

Table 7: Q13, Nearest Neighbour Matching, All Variables, with Controls

	Dependent variable:
	paste0("pctVV ~treat + ", covars_formula)
treat	0.181 (0.142)
occBlue.collar	0.276 (0.417)
occEducation	0.403 (0.416)
occGovernment	0.133 (0.383)
occMedia	0.011 (0.420)
occNone	0.219 (0.486)
occOther	0.373 (0.370)
occPolitician	0.570 (0.453)
occWhite.collar	0.543 (0.353)
lat	0.031** (0.015)
long	0.006 (0.014)
ran.prior	−0.873*** (0.207)
incumbent	−0.487* (0.286)
log.valid.votes	−0.533*** (0.088)
party.prior.pctVV	0.012* (0.007)
prior.pctVV	0.511*** (0.059)
elec.year	0.007 (0.062)
match.partyPCB	
match.partyPC.do.B	0.852 (0.575)
match.partyPDT	−0.144 (0.335)
match.partyPFL	0.650* (0.382)
match.partyPL	0.385 (0.444)
match.partyPMDB	0.857** (0.348)
match.partyPMN	−0.907 (0.661)
match.partyPP	0.582* (0.345)
match.partyPPS	0.445 (0.446)
match.partyPSB	0.639* (0.387)
match.partyPSC	0.369 (0.412)
match.partyPSDB	0.119 (0.309)
match.partyPSDC	0.327 (0.704)
match.partyPSL	0.079 (0.752)
match.partyPT	−0.051 (0.288)
match.partyPTB	0.490 (0.330)
match.partyPV	0.580 (0.556)
uf.rs	−0.643 (0.612)
uf.sp	−0.093 (0.253)
yob	0.019** (0.008)
eduMore.than.Primary..Less.than.Superior	0.228 (0.207)
eduSome.Superior.or.More	0.839*** (0.236)
log.total.assets	0.030 (0.020)
pt_pres_1998	0.633 (1.133)
psdb_2000	−0.620 (0.421)
hdi_2000	2.059 (2.391)
income_2000	−0.001 (0.002)
log.num.apps	−0.130 (0.172)
Constant	−45.121 (122.583)
Observations	622
R^2	0.426
Adjusted R^2	0.382
Residual Std. Error	1.745 (df = 577)
F Statistic	9.730*** (df = 44; 577)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
    title = "Q14, Nearest Neighbour Matching, All Variables, with Caliper of 0.01",
    single.row = T)
```

Table 8: Q14, Nearest Neighbour Matching, All Variables, with Caliper of 0.01

	Dependent variable:
	pctVV
treat	0.435** (0.200)
Constant	2.440*** (0.141)
Observations	488
R^2	0.010
Adjusted R^2	0.008
Residual Std. Error	2.210 (df = 486)
F Statistic	4.734** (df = 1; 486)
Note:	*p<0.1; **p<0.05; ***p<0.01

Adding a caliper greatly affects the estimated treatment effect, which rises to 0.44% points (the exact value will vary as the algorithm includes a random process) and is statistically significant.

15. Another problem with nearest neighbour matching is that it is 'greedy' - the first matches might make it harder to match well later. Boas and Hidalgo use genetic matching, which is a complex automated process to try and get the best 'overall' matches for the full dataset. Run genetic matching process with the same variables and then run your regression (with no controls) again. *Note:* Genetic matching might take 10-20 minutes.

```
matched_data_Q15 %>% match.data() %>% lm(pctVV ~ treat, data = .) %>% stargazer(header = F,
    title = "Q15, Genetic Matching", single.row = T)
```

Table 9: Q15, Genetic Matching

	Dependent variable:
	pctVV
treat	0.305 (0.195)
Constant	2.444*** (0.148)
Observations	537
R^2	0.005
Adjusted R^2	0.003
Residual Std. Error	2.228 (df = 535)
F Statistic	2.453 (df = 1; 535)
Note:	*p<0.1; **p<0.05; ***p<0.01

The result is quite different from that found in Boas and Hidalgo - smaller and not statistically significant. This probably reflects the lack of stability in the results of genetic matching, but needs further investigation.