FLS 6415: Class 10 Homework

November, 2017

Remember to answer all the questions in R markdown and produce a PDF. Email your completed homework (R markdown file and PDF) to jonnyphillips@gmail.com by midnight the night before class. Remember to refer to the example code from this week and the last couple of weeks for coding guidance.

Load the Boas and Hidalgo dataset `combined_data.RDS` using `readRDS()`. There are many tools and settings that can be adjusted when using matching - we will try and repicate the Boas and Hidalgo results using more user-friendly tools than in their own code, so we will not recover identical values.

1. The treatment variable in Boas and Hidalgo is whether a councillor that applied for a media licence received approval before the 2004 election (`treat`). The outcome variable is the councillor's vote share in the 2004 elections (`pctVV`). Conduct and interpret a basic linear regression of the outcome on treatment with no controls.

Table 1: Q1

	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

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.

2. 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?

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.

3. One 'control for confouding' approach is to use a regression. Conduct the regression of the oucome (`pctVV`) on treatment (`treat`), controlling for `male` and compare the results to your answer to Q1.

The results slightly reduce the coefficient estimate on the treatment variable, but it remains significant. Being male is positively associated with the outcome, but the difference is not statistically significant.

4. An alternative approach to overcome confounding is matching: removing units from the dataset until we have balance on gender. Let's try to implement one-to-one exact matching MANUALLY to illustrate the process:

Table 2: Q3

	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

a. Create separate datasets for treated and control units. b. Create a new column in the treated dataset (eg. matched_control_unit_pctVV) where we are going to save the outcome for the matched control unit. b. Create a `for` loop so we can analyze each *treated* unit (each row of your treated dataset) separately.
c. Within your for loop, for each treated unit, identify its gender, and filter the control units dataset to only include that same gender.
d. There are more control units than treated units, so we can pick any same-gender control unit as our matched unit. *Hint:* Use `sample_n(1)` to randomly pick one row of your control unit data.
e. Now select only the outcome variable for this matched control unit and save that value to the new column you made in the *treated* dataset (eg. matched_control_unit_pctVV). Make sure you save it only to the row of the current treated unit you are analysing in your `for` loop.

5. With this matched dataset, we now have balance on gender by definition. To estimate the treatment effect using a difference-in-means calculate the difference in outcomes between your treated and matched control units in the treated units dataset (i.e. between the `pctVV` and `matched_control_unit_pctVV` variables). Then average these differences to get an average treatment effect and interpret the result. Also conduct a t-test to get a p-value.

The estimated average treatment effect for receiving an approved media licence before the election is a 0.536% points increase in vote share, which is statistically significant with a p-value of 0.002.

6. We can also use post-matching regression to do the analysis stage. To do so, we need to reorganize our dataset to include rows for both the treated and matched control units, with their corresponding outcomes.
a. Select only the 'pctVV' and 'matched_control_unit_pctVV' columns. b. Transform the data from wide to long format using `gather`. Your dataset should now have 622 rows. c. Use `mutate` with `ifelse` to rename the treatment column values from 'pctVV' and 'matched_control_unit_pctVV' to 1 and 0 respectively. d. Conduct the simple linear regression of the outcome on treatment for this dataset.
e. Compare the results to your answer in Q5.

The estimate treatment effect is 0.536, with a p-value of 0.002. This is exactly the same as in Q5.

7. To match on continuous variables, we can't match 'exactly' so we do 'nearest' neighbour matching. Another potential confounder is the size of the electorate (`log.valid.votes`). Use

Table 3: Q6

	Dependent variable:
	pctVV_all
Treatment	0.536***
	(0.174)
Constant	2.213***
	(0.123)
Observations	622
R^2	0.015
Adjusted R^2	0.014
Residual Std. Error	2.165 (df = 620)
F Statistic	9.521*** (df = 1; 620)
Note:	*p<0.1; **p<0.05; ***p<0.01

`matchit` to construct a datset of treated and control units matched on (only) the size of the electorate with nearest neighbour matching. In your matched data, how many treated and control units are there? *Hint:* Access the units summary using `output$nn`, where `output` is the result of `matchit`.

Table 4: Q7

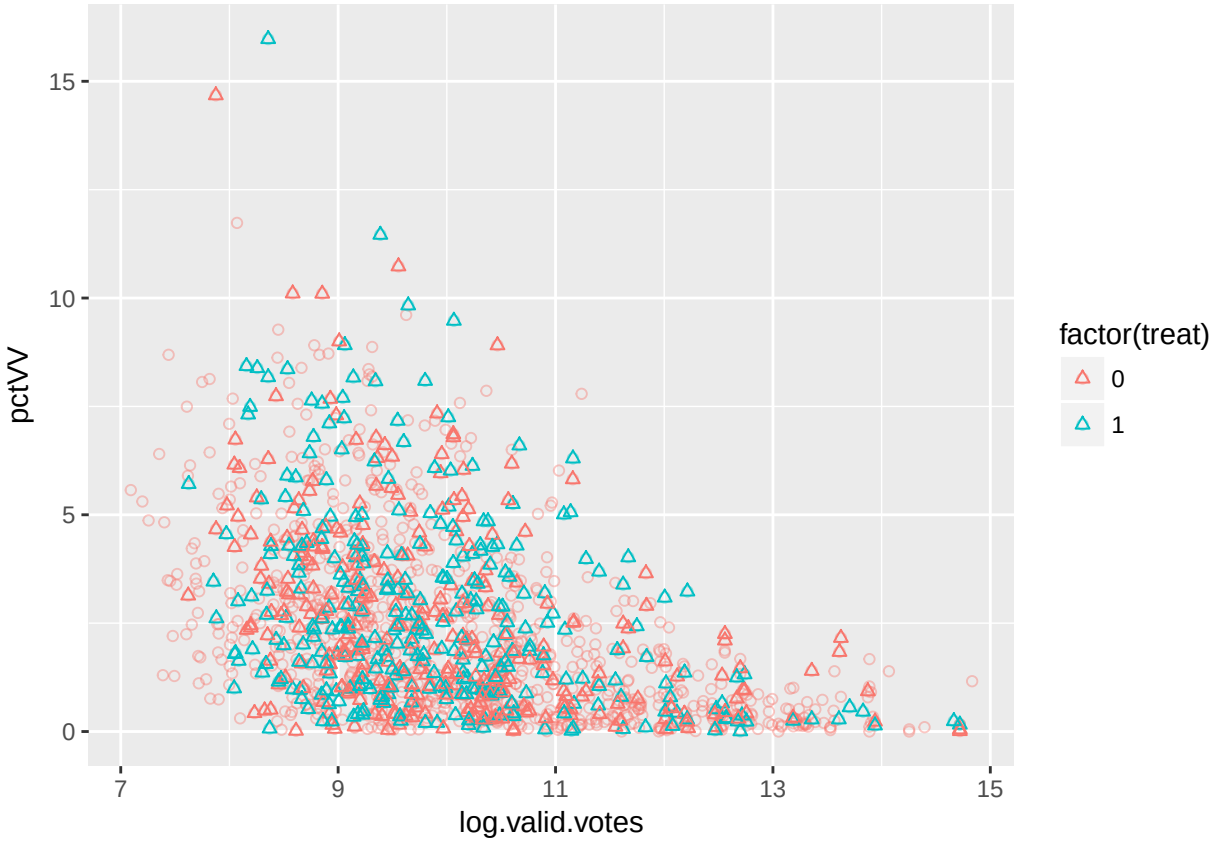
	Control	Treated
All	1144	311
Matched	311	311
Unmatched	833	0
Discarded	0	0

8. Check that this matching has improved balance by comparing the average difference in `log.valid.votes` in the original dataset and in the matched dataset. *Hint:* Remember you can use `match.data(output_of_matchit)` to get the matched dataset.

	treat	Before_Matching	After_Matching
1	0.00	10.12	9.93
2	1.00	9.93	9.93

Balance improves considerably after matching.

9. To see which units have been matched, plot two layers on a graph of the confounder `log.valid.votes` against the outcome `pctVV`. One layer should be the full dataset - make this a specific size, shape and alpha (transparency). The second layer should be the matched dataset from Q8 - make this a different size, shape and alpha so that the matched units stand out. Group and colour each layer by treatment status.



10. Use a simple least squares regression to estimate the effect of the treatment ('treat') on the outcome ('pctVV') in the matched dataset. Interpret the result.

Receiving an approved media licence application before the election is associated with a 0.199 % points increase in vote share, but this is not statistically significant.

11. Conduct the same regression as in Q8 but add a control for `log.valid.votes`, the variable we already matched on. Does this change the estimate of the treatment effect compared to your answer in Q8? Why/why not?

The coefficient on treatment is essentially unchanged because we have already gone a long way towards achieving balance through the matching procedure, and the regression only controls for any residual imbalance. The p-value is slightly smaller because we have absorbed some of the variation with the control, allowing greater precision.

12. Now lets include all of the matching variables that Boas and Hidalgo use, and use nearest neighbour matching to construct a matched dataset. Use the list of matching variables provided below for matching.

```
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",
```

Table 5: Q10

	Dependent variable:
	pctVV
treat	0.199
	(0.181)
Constant	2.549***
	(0.128)
Observations	622
R^2	0.002
Adjusted R^2	0.0003
Residual Std. Error	2.253 (df = 620)
F Statistic	1.217 (df = 1; 620)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Q11

	Dependent variable:
	pctVV
treat	0.199
	(0.167)
log.valid.votes	−0.674***
	(0.065)
Constant	9.242***
	(0.658)
Observations	622
R^2	0.149
Adjusted R^2	0.146
Residual Std. Error	2.082 (df = 619)
F Statistic	54.137*** (df = 2; 619)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
    "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 = " + ")
```

13. For your matched dataset from Q12, report a before-matching and after-matching balance table similar to the first four columns of Boas and Hidalgo's Table 3. The standardized difference in means they use allows us to compare everything on the same scale and requires calculating the difference in means and then dividing by the standard deviation of the treated group:

a. Calculate the standard deviation of the treated variable for the covariates. *Hint:* `summarise_at` is useful to apply the same summary to the list of covariates provided above. b. Summarise the average of each covariate separately for the treated and control groups, and then calculate the difference.

c. Divide this difference in means by the standard deviation of the treated group for each variable, and multiply by 100.

d. Calulate the p-value of the difference in means for each covariate between treated and control groups. *Hint:* Try something like `d %>% summarise_at(covars,funs(t.test(.[treat==0], .[treat==1])$p.value))`.

e. Now combine the columns for standardized differences in means and p-values for the full and matched datasets into a single table. *Hint:* Use `gather()` to turn wide rows into long columns, and `left_join` to combine the columns into a single table.

14. Implement the regression of the outcome on treatment (excluding controls) on the matched dataset and compare your answer to that of Boas and Hidalgo in Table 4.

The estimate is considerably smaller than Boas and Hidalgo's at 0.154, and is statistically insignificant with a p-value of 0.386

15. One risk with the nearest-neighbour method is that the control unit can still be far away from the treated unit if there are no good matches, so imbalance can still be a problem. We can put a 'caliper' on the matching method so that it only includes 'close' matches within a specific number of standard deviations. Re-run the matching process from Q12 but with a caliper of 0.01 standard deviations, and then conduct the regression from Q14. Note how the number of units changes and interpret if the result is any different.

The coefficient estimate is higher at 0.261, and has a smaller p-value of 0.207. The numer of units reduces from 622 to 492 as treated units with no matching control units within the caliper distance are discarded.

16. Boas and Hidalgo use genetic matching, which is a complex automated process. We can implement genetic matching easily by loading the package `rgenoud` and choosing `method="genetic"` in the matching process in `matchit`. Then run the same regression again and compare the results to those in Q14 and in Boas and Hidalgo. *Hint:* Genetic matching might take 10-20 minutes, so in your R chunk options set `cache=TRUE` so that it doesn't re-run every time you knit to PDF.

The coefficient estimate is higher at 0.329, and has a smaller p-value of 0.084. However, the esimate and p-value are still a little distant from the Boas and Hidalgo results.

17. As with regression, these estimates only have a causal interpretation if matching is successful at producing balance on ALL of the confounding covariates. Provide an example of a confounding variable which would bias Boas and Hidalgo's estimates despite their matching procedure.

A confounding variable could be anything that is not included in the matching procedure (or anything in the

matching procedure which remains imbalanced). For example, they do not have a perfect measure of the councillors' personal wealth; if bribes or status affect the chance of getting an application approved and also boost electoral success, the results are not causal estimates.

	key	Before_Mean	After_Mean	Before_p	After_p
1	occBlue.collar	-11.94	-1.25	0.07	0.88
2	occEducation	-2.00	-1.16	0.76	0.89
3	occGovernment	-1.27	4.79	0.84	0.54
4	occMedia	16.84	3.42	0.00	0.66
5	occNone	-6.87	-5.21	0.30	0.54
6	occOther	1.00	-4.30	0.88	0.60
7	occPolitician	-2.54	-2.75	0.69	0.74
8	occWhite.collar	9.85	-1.39	0.12	0.86
9	lat	-3.58	2.93	0.59	0.71
10	long	-0.41	-0.05	0.95	0.99
11	ran.prior	-1.70	0.00	0.79	1.00
12	incumbent	4.96	-1.58	0.43	0.84
13	log.valid.votes	-14.47	9.49	0.03	0.22
14	party.prior.pctVV	2.99	2.64	0.64	0.74
15	prior.pctVV	10.92	0.67	0.08	0.93
16	elec.year	17.35	9.31	0.01	0.25
17	match.partyPCB	-Inf		0.16	
18	match.partyPC.do.B	0.68	2.33	0.92	0.76
19	match.partyPDT	7.45	5.80	0.23	0.45
20	match.partyPFL	-4.00	-7.49	0.54	0.38
21	match.partyPL	-3.59	-9.57	0.58	0.29
22	match.partyPMDB	-5.50	4.42	0.40	0.57
23	match.partyPMN	1.33	0.00	0.83	1.00
24	match.partyPP	5.25	-1.16	0.40	0.89
25	match.partyPPS	-4.56	-1.82	0.49	0.82
26	match.partyPSB	-10.93	0.00	0.10	1.00
27	match.partyPSC	11.88	7.49	0.05	0.31
28	match.partyPSDB	4.87	6.79	0.44	0.38
29	match.partyPSDC	1.33	2.85	0.83	0.70
30	match.partyPSL	-1.00	5.70	0.88	0.41
31	match.partyPT	-8.55	-3.68	0.19	0.65
32	match.partyPTB	3.69	-7.22	0.56	0.39
33	match.partyPV	-5.44	4.33	0.41	0.56
34	uf.rs	-27.21	6.49	0.00	0.36
35	uf.sp	-1.66	-1.96	0.80	0.81
36	yob	-3.09	2.27	0.63	0.77
37	eduMore.than.Primary..Less.than.Superior	-8.59	5.14	0.18	0.52
38	eduSome.Superior.or.More	13.53	1.35	0.03	0.87
39	log.total.assets	16.72	11.95	0.01	0.13
40	pt_pres_1998	-22.75	4.34	0.00	0.60
41	psdb_2000	-10.93	0.24	0.09	0.98
42	hdi_2000	-6.13	2.14	0.34	0.79
43	income_2000	-8.21	4.93	0.21	0.53
44	log.num.apps	-111.38	-4.21	0.00	0.60

Table 7: Q13

Table 8: Q14

	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

Table 9: Q15

	Dependent variable:
	pctVV
treat	0.261
	(0.207)
Constant	2.653***
	(0.146)
Observations	492
R^2	0.003
Adjusted R^2	0.001
Residual Std. Error	2.294 (df = 490)
F Statistic	1.593 (df = 1; 490)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 10: Q16

	Dependent variable:
	pctVV
treat	0.329*
	(0.190)
Constant	2.420***
	(0.143)
Observations	546
R^2	0.005
Adjusted R^2	0.004
Residual Std. Error	2.195 (df = 544)
F Statistic	3.003* (df = 1; 544)
Note:	*p<0.1; **p<0.05; ***p<0.01