

Intro

Angrist et al. 2002 - Background

Angrist et al. 2002 - Data, result and replication

ML Extension

Conclusions

ML4Econ - Replication Assignment

Ariel Karlinsky

28/6/2019

Intro

This markdown contains both code and text for the replication assignment in the ML4Econ course in HUJI. The paper I chose is *Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment* by: Angrist, Bettinger, Bloom, King and Kremer. From now on: Angrist et al. 2002.

This markdown has two objectives

1. Replicate some result from Angrist et al. 2002.
2. Extend the analysis in Angrist et al. 2002. with methods learned in the course, for example - using Lasso to choose variables, causal trees, etc.

We begin by loading relevant packages, setting a nice ggplot theme and loading the dataset using the `experimentdatar` package.

```
knitr::opts_chunk$set(echo = TRUE)
if (!require("pacman"))
  install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(tidyverse,
               readr,
               tree,
               randomForest,
               gbm,
               glmnet,
               pls,
               broom,
               ggpubr,
               alphahull,
               flexclust,
               experimentdatar,
               summarytools,
               janitor,
               causalTree)

theme_set(theme_classic2())

data(vouchers)
```

Angrist et al. 2002 - Background

We start by familiarizing ourselves with the paper, data and results. The paper's abstract reads:

Colombia used lotteries to distribute vouchers which partially covered the cost of private secondary school for students who maintained satisfactory academic progress. Three years after the lotteries, winners were about 10 percentage points more likely to have finished 8th grade, primarily because they were less likely to repeat grades, and scored 0.2 standard deviations higher on achievement tests. There is some evidence that winners worked less than losers and were less likely to marry or cohabit as teenagers. Benefits to participants likely exceeded the \$24 per winner additional cost to the government of supplying vouchers instead of public school places.

Unfortunately, the `experimentdatar` package doesn't contain much details about this dataset, so we will have to read the paper (!) and explore the data manually.

The paper's abstract summarizes it nicely of course, but here is some more information, summarized in built points:

- The country of Colombia has one of the largest (at least at the time) school vouchers programs in the world: the Programa de Ampliación de Cobertura de la Educación Secundaria (PACES).
- PACES provides over 125,000 students with vouchers covering a bit more than half of the cost of private secondary school.
- The vouchers are renewed each year given adequate academic performance.
- Many of the vouchers are awarded by lottery, allowing for a quasi-experimental research design - effects are estimated by comparing the outcomes between lottery winners and losers.

- Both lottery winners and losers **applied** to private schools at similar rates, yet lottery winners were 15 percentage points *more likely* to actually attend private, rather than public, schools.
- Lottery winners (vs. lottery losers):
 - Complete an additional 0.1 years of schooling.
 - Are 10 percentage points more likely to have completed 8th grade.
 - Score 0.2 SDs higher than losers on achievement tests.
 - Less likely to be married or cohabiting (in what range of years?)
 - Worked about 1.2 fewer hours per week.
- The effect on girls is larger and more precisely estimated than the effect of boys (This is a rather standard and robust results in many of these papers. I call this “Boys interrupting girls learning”).

There are some important caveats to these results:

- Only about 90% of lottery **winners** had ever used the voucher or any other type of school scholarship.
- About 24% of lottery **losers** received scholarship from other sources.
- The authors thus use an IV strategy, where the first stage is scholarship receipt instrumented by lottery win, and the second stage is the effect of scholarship on outcomes.
- The 2SLS estimates are roughly 50% higher than those in the previous bullets.

Additionally, the paper presents a cost-benefit analysis of the PACES program, from the point of view of both households and the government. We will not elaborate on this as is this is not the main focus of the paper or something that can be readily extended using modern machine learning methods.

Angrist et al. 2002 - Data, result and replication

Let's get to know the data, before that we will rename all variables to lower case:

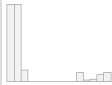
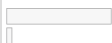



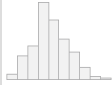
```
vouchers <- vouchers %>% rename_all(tolower)
print(dfSummary(vouchers), method = "render")
```


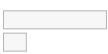




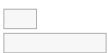
Data Frame Summary

vouchers

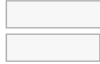

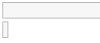

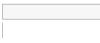
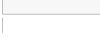
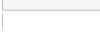
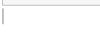
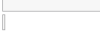

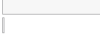
Dimensions: 25330 x 89

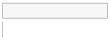
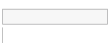

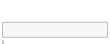
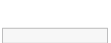







Duplicates: 0

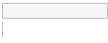
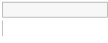
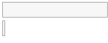
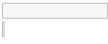
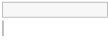
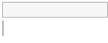
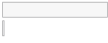
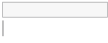
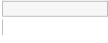
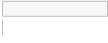
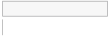
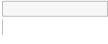
No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	id [numeric]	ID number assigned by Gabriel	Mean (sd) : 27669.7 (40938.6) min < med < max: 1 < 12665 < 141226 IQR (CV) : 12664 (1.5)	25329 distinct values		25329 (100%)	1 (0%)
2	bog95smp [numeric]		Min : 0 Mean : 0 Max : 1	0 : 24154 (95.4%) 1 : 1176 (4.6%)		25330 (100%)	0 (0%)
3	bog97smp [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25053 (98.9%) 1 : 277 (1.1%)		25330 (100%)	0 (0%)
4	jam93smp [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25165 (99.4%) 1 : 165 (0.6%)		25330 (100%)	0 (0%)
5	sex [numeric]	Gender: Equals 1 if Applicant is Male	Min : 0 Mean : 0.5 Max : 1	0 : 9447 (47.4%) 1 : 10480 (52.6%)		19927 (78.67%)	5403 (21.33%)
6	age [numeric]		Mean (sd) : 14.7 (1.7) min < med < max: 10 < 15 < 20 IQR (CV) : 2 (0.1)	11 distinct values		1938 (7.65%)	23392 (92.35%)

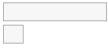
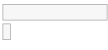
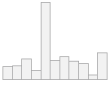
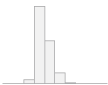
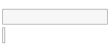
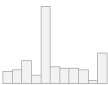
No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
7	age2 [numeric]		Mean (sd) : 13.1 (4) min < med < max: -2 < 13 < 84 IQR (CV) : 2 (0.3)	73 distinct values		19811 (78.21%)	5519 (21.79%)
8	hsvisit [numeric]	Survey was conducted in person	Min : 0 Mean : 0.2 Max : 1	0 : 1588 (81.6%) 1 : 358 (18.4%)		1946 (7.68%)	23384 (92.32%)
9	scyfnsh [numeric]	Maximum grade finished	Mean (sd) : 5.2 (0.7) min < med < max: 4 < 5 < 11 IQR (CV) : 0 (0.1)	4 : 1 (0.0%) 5 : 23498 (92.8%) 6 : 534 (2.1%) 7 : 271 (1.1%) 8 : 841 (3.3%) 9 : 157 (0.6%) 10 : 21 (0.1%) 11 : 7 (0.0%)		25330 (100%)	0 (0%)
10	inschl [numeric]	Applicant is still in school	Min : 0 Mean : 0.9 Max : 1	0 : 268 (13.9%) 1 : 1663 (86.1%)		1931 (7.62%)	23399 (92.38%)
11	prsch_c [numeric]		Min : 0 Mean : 0.7 Max : 1	0 : 658 (34.2%) 1 : 1265 (65.8%)		1923 (7.59%)	23407 (92.41%)
12	prscha_1 [numeric]		Min : 0 Mean : 0.9 Max : 1	0 : 211 (11.0%) 1 : 1710 (89.0%)		1921 (7.58%)	23409 (92.42%)
13	prscha_2 [numeric]		Min : 0 Mean : 0.8 Max : 1	0 : 469 (24.4%) 1 : 1452 (75.6%)		1921 (7.58%)	23409 (92.42%)

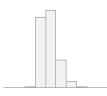


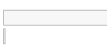
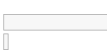


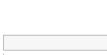

No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
14	vouch0 [numeric]	Student won voucher	Min : 0 Mean : 0.7 Max : 1	0 : 7149 (28.2%) 1 : 18180 (71.8%)		25329 (100%)	1 (0%)
15	bog95asd [numeric]		Min : 0 Mean : 0.1 Max : 1	0 : 23081 (91.1%) 1 : 2249 (8.9%)		25330 (100%)	0 (0%)
16	bog97asd [numeric]		Min : 0 Mean : 0 Max : 1	0 : 24873 (98.2%) 1 : 457 (1.8%)		25330 (100%)	0 (0%)
17	jam93asd [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25051 (98.9%) 1 : 279 (1.1%)		25330 (100%)	0 (0%)
18	dbogota [numeric]	Dummy indicating Bogota	Min : 0 Mean : 0.2 Max : 1	0 : 19515 (77.0%) 1 : 5815 (23.0%)		25330 (100%)	0 (0%)
19	djamundi [numeric]	Dummy indicating Jamundi	Min : 0 Mean : 0 Max : 1	0 : 24987 (98.7%) 1 : 343 (1.4%)		25330 (100%)	0 (0%)
20	d1995 [numeric]	Dummy indicating year=1995	Min : 0 Mean : 0.2 Max : 1	0 : 20784 (82.0%) 1 : 4546 (17.9%)		25330 (100%)	0 (0%)
21	d1997 [numeric]	Dummy indicating year=1997	Min : 0 Mean : 0.1 Max : 1	0 : 23560 (93.0%) 1 : 1770 (7.0%)		25330 (100%)	0 (0%)
22	response [numeric]		Min : 0 Mean : 0.5 Max : 1	0 : 1410 (45.9%) 1 : 1662 (54.1%)		3072 (12.13%)	22258 (87.87%)
23	test_tak [numeric]		Min : 0 Mean : 0 Max : 1	0 : 5874 (95.4%) 1 : 283 (4.6%)		6157 (24.31%)	19173 (75.69%)

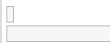
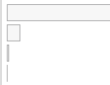

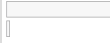
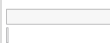


No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
24	sex_name [numeric]	Gender based on name	Min : 0 Mean : 0.5 Max : 1	0 : 2472 (51.0%) 1 : 2371 (49.0%)		4843 (19.12%)	20487 (80.88%)
25	svy [numeric]	Survey completed using new survey	Min : 0 Mean : 0.5 Max : 1	0 : 954 (49.0%) 1 : 992 (51.0%)		1946 (7.68%)	23384 (92.32%)
26	d1993 [numeric]	Dummy indicating year=1993	Min : 0 Mean : 0 Max : 1	0 : 24107 (95.2%) 1 : 1223 (4.8%)		25330 (100%)	0 (0%)
27	phone [numeric]	Applicant has phone	Min : 0 Mean : 0.6 Max : 1	0 : 10996 (43.4%) 1 : 14334 (56.6%)		25330 (100%)	0 (0%)
28	darea1 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25311 (99.9%) 1 : 19 (0.1%)		25330 (100%)	0 (0%)
29	darea2 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25326 (100.0%) 1 : 4 (0.0%)		25330 (100%)	0 (0%)
30	darea3 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25329 (100.0%) 1 : 1 (0.0%)		25330 (100%)	0 (0%)
31	darea4 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25192 (99.5%) 1 : 138 (0.5%)		25330 (100%)	0 (0%)
32	darea5 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 24834 (98.0%) 1 : 496 (2.0%)		25330 (100%)	0 (0%)
33	darea6 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25170 (99.4%) 1 : 160 (0.6%)		25330 (100%)	0 (0%)
34	darea7 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25032 (98.8%) 1 : 298 (1.2%)		25330 (100%)	0 (0%)

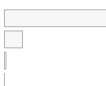



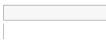



No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
35	darea8 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25234 (99.6%) 1 : 96 (0.4%)		25330 (100%)	0 (0%)
36	darea9 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25304 (99.9%) 1 : 26 (0.1%)		25330 (100%)	0 (0%)
37	darea10 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25154 (99.3%) 1 : 176 (0.7%)		25330 (100%)	0 (0%)
38	darea11 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25110 (99.1%) 1 : 220 (0.9%)		25330 (100%)	0 (0%)
39	darea12 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25258 (99.7%) 1 : 72 (0.3%)		25330 (100%)	0 (0%)
40	darea13 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25327 (100.0%) 1 : 3 (0.0%)		25330 (100%)	0 (0%)
41	darea14 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25323 (100.0%) 1 : 7 (0.0%)		25330 (100%)	0 (0%)
42	darea15 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25247 (99.7%) 1 : 83 (0.3%)		25330 (100%)	0 (0%)
43	darea16 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25282 (99.8%) 1 : 48 (0.2%)		25330 (100%)	0 (0%)
44	darea17 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25024 (98.8%) 1 : 306 (1.2%)		25330 (100%)	0 (0%)
45	darea18 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25238 (99.6%) 1 : 92 (0.4%)		25330 (100%)	0 (0%)
46	darea19 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 24546 (96.9%) 1 : 784 (3.1%)		25330 (100%)	0 (0%)

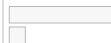
No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
47	dmonth1 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25304 (99.9%) 1 : 26 (0.1%)		25330 (100%)	0 (0%)
48	dmonth2 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25288 (99.8%) 1 : 42 (0.2%)		25330 (100%)	0 (0%)
49	dmonth3 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 24802 (97.9%) 1 : 528 (2.1%)		25330 (100%)	0 (0%)
50	dmonth4 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 24998 (98.7%) 1 : 332 (1.3%)		25330 (100%)	0 (0%)
51	dmonth5 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25110 (99.1%) 1 : 220 (0.9%)		25330 (100%)	0 (0%)
52	dmonth6 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25165 (99.4%) 1 : 165 (0.6%)		25330 (100%)	0 (0%)
53	dmonth7 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 24903 (98.3%) 1 : 427 (1.7%)		25330 (100%)	0 (0%)
54	dmonth8 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25170 (99.4%) 1 : 160 (0.6%)		25330 (100%)	0 (0%)
55	dmonth9 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25314 (99.9%) 1 : 16 (0.1%)		25330 (100%)	0 (0%)
56	dmonth10 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25321 (100.0%) 1 : 9 (0.0%)		25330 (100%)	0 (0%)
57	dmonth11 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25319 (100.0%) 1 : 11 (0.0%)		25330 (100%)	0 (0%)
58	dmonth12 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25320 (100.0%) 1 : 10 (0.0%)		25330 (100%)	0 (0%)

No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
59	bog95 [numeric]		Min : 0 Mean : 0.2 Max : 1	0 : 21285 (84.0%) 1 : 4045 (16.0%)		25330 (100%)	0 (0%)
60	bog97 [numeric]		Min : 0 Mean : 0.1 Max : 1	0 : 23560 (93.0%) 1 : 1770 (7.0%)		25330 (100%)	0 (0%)
61	mom_sch [numeric]		Mean (sd) : 5.9 (2.8) min < med < max: 0 < 5 < 11 IQR (CV) : 3 (0.5)	12 distinct values		1723 (6.8%)	23607 (93.2%)
62	mom_age [numeric]		Mean (sd) : 40.4 (7) min < med < max: 8 < 39 < 97 IQR (CV) : 10 (0.2)	46 distinct values		1813 (7.16%)	23517 (92.84%)
63	mom_mw [numeric]		Min : 0 Mean : 0 Max : 1	0 : 1749 (98.0%) 1 : 35 (2.0%)		1784 (7.04%)	23546 (92.96%)
64	dad_sch [numeric]		Mean (sd) : 5.9 (2.9) min < med < max: 0 < 5 < 11 IQR (CV) : 4 (0.5)	12 distinct values		1384 (5.46%)	23946 (94.54%)

No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
65	dad_age [numeric]		Mean (sd) : 44.3 (8) min < med < max: 1 < 43 < 91 IQR (CV) : 11 (0.2)	50 distinct values		1524 (6.02%)	23806 (93.98%)
66	dad_mw [numeric]		Min : 0 Mean : 0.1 Max : 1	0 : 1260 (90.2%) 1 : 137 (9.8%)		1397 (5.52%)	23933 (94.48%)
67	sex2 [numeric]		Min : 0 Mean : 0.5 Max : 1	0 : 2648 (51.4%) 1 : 2502 (48.6%)		5150 (20.33%)	20180 (79.67%)
68	strata1 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25032 (98.8%) 1 : 298 (1.2%)		25330 (100%)	0 (0%)
69	strata2 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 24272 (95.8%) 1 : 1058 (4.2%)		25330 (100%)	0 (0%)
70	strata3 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25065 (99.0%) 1 : 265 (1.0%)		25330 (100%)	0 (0%)
71	strata4 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25319 (100.0%) 1 : 11 (0.0%)		25330 (100%)	0 (0%)
72	strata5 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25328 (100.0%) 1 : 2 (0.0%)		25330 (100%)	0 (0%)
73	strata6 [numeric]		Min : 0 Mean : 0 Max : 1	0 : 25329 (100.0%) 1 : 1 (0.0%)		25330 (100%)	0 (0%)

No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
74	stratams [numeric]		Min : 0 Mean : 0.9 Max : 1	0 : 1635 (6.4%) 1 : 23695 (93.5%)		25330 (100%)	0 (0%)
75	rept6 [numeric]		Mean (sd) : 0.1 (0.4) min < med < max: 0 < 0 < 3 IQR (CV) : 0 (2.8)	0 : 1679 (87.5%) 1 : 214 (11.2%) 2 : 25 (1.3%) 3 : 1 (0.0%)		1919 (7.58%)	23411 (92.42%)
76	totscys [numeric]		Mean (sd) : 3.4 (1.3) min < med < max: 0 < 4 < 6 IQR (CV) : 2 (0.4)	0 : 10 (0.5%) 1 : 106 (5.5%) 2 : 507 (26.2%) 3 : 160 (8.3%) 4 : 972 (50.2%) 5 : 50 (2.6%) 6 : 130 (6.7%)		1935 (7.64%)	23395 (92.36%)
77	haschild [numeric]		Min : 0 Mean : 0 Max : 1	0 : 1882 (97.5%) 1 : 48 (2.5%)		1930 (7.62%)	23400 (92.38%)
78	married [numeric]		Min : 0 Mean : 0 Max : 1	0 : 1907 (98.8%) 1 : 24 (1.2%)		1931 (7.62%)	23399 (92.38%)
79	working [numeric]		Min : 0 Mean : 0.1 Max : 1	0 : 1647 (85.1%) 1 : 288 (14.9%)		1935 (7.64%)	23395 (92.36%)
80	rept [numeric]		Min : 0 Mean : 0.2 Max : 1	0 : 1618 (83.6%) 1 : 317 (16.4%)		1935 (7.64%)	23395 (92.36%)

No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing												
81	nrept [numeric]		Mean (sd) : 0.2 (0.4) min < med < max: 0 < 0 < 3 IQR (CV) : 0 (2.4)	<table><tr><td>0 :</td><td>1618</td><td>(83.6%)</td></tr><tr><td>1 :</td><td>282</td><td>(14.6%)</td></tr><tr><td>2 :</td><td>32</td><td>(1.6%)</td></tr><tr><td>3 :</td><td>3</td><td>(0.2%)</td></tr></table>	0 :	1618	(83.6%)	1 :	282	(14.6%)	2 :	32	(1.6%)	3 :	3	(0.2%)		1935 (7.64%)	23395 (92.36%)
0 :	1618	(83.6%)																	
1 :	282	(14.6%)																	
2 :	32	(1.6%)																	
3 :	3	(0.2%)																	
82	finish6 [numeric]		Min : 0 Mean : 0.9 Max : 1	<table><tr><td>0 :</td><td>124</td><td>(6.4%)</td></tr><tr><td>1 :</td><td>1811</td><td>(93.6%)</td></tr></table>	0 :	124	(6.4%)	1 :	1811	(93.6%)		1935 (7.64%)	23395 (92.36%)						
0 :	124	(6.4%)																	
1 :	1811	(93.6%)																	
83	finish7 [numeric]		Min : 0 Mean : 0.7 Max : 1	<table><tr><td>0 :</td><td>656</td><td>(33.9%)</td></tr><tr><td>1 :</td><td>1279</td><td>(66.1%)</td></tr></table>	0 :	656	(33.9%)	1 :	1279	(66.1%)		1935 (7.64%)	23395 (92.36%)						
0 :	656	(33.9%)																	
1 :	1279	(66.1%)																	
84	finish8 [numeric]		Min : 0 Mean : 0.5 Max : 1	<table><tr><td>0 :</td><td>924</td><td>(47.8%)</td></tr><tr><td>1 :</td><td>1011</td><td>(52.2%)</td></tr></table>	0 :	924	(47.8%)	1 :	1011	(52.2%)		1935 (7.64%)	23395 (92.36%)						
0 :	924	(47.8%)																	
1 :	1011	(52.2%)																	
85	sex_miss [numeric]		Min : 0 Mean : 0 Max : 1	<table><tr><td>0 :</td><td>1934</td><td>(100.0%)</td></tr><tr><td>1 :</td><td>1</td><td>(0.0%)</td></tr></table>	0 :	1934	(100.0%)	1 :	1	(0.0%)		1935 (7.64%)	23395 (92.36%)						
0 :	1934	(100.0%)																	
1 :	1	(0.0%)																	
86	usngsch [numeric]		Min : 0 Mean : 0.4 Max : 1	<table><tr><td>0 :</td><td>1255</td><td>(64.9%)</td></tr><tr><td>1 :</td><td>680</td><td>(35.1%)</td></tr></table>	0 :	1255	(64.9%)	1 :	680	(35.1%)		1935 (7.64%)	23395 (92.36%)						
0 :	1255	(64.9%)																	
1 :	680	(35.1%)																	
87	hoursum [numeric]		Mean (sd) : 3.6 (10.6) min < med < max: 0 < 0 < 40 IQR (CV) : 0 (2.9)	24 distinct values 	1935 (7.64%)	23395 (92.36%)													
88	tab3smpl [numeric]		1 distinct value	<table><tr><td>1 :</td><td>1577</td><td>(100.0%)</td></tr></table>	1 :	1577	(100.0%)		1577 (6.23%)	23753 (93.77%)									
1 :	1577	(100.0%)																	

No	Variable	Label	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing						
89	working3 [numeric]		Min : 0 Mean : 0.1 Max : 1	<table><tr><td>0 :</td><td>1675</td><td>(86.6%)</td></tr><tr><td>1 :</td><td>260</td><td>(13.4%)</td></tr></table>	0 :	1675	(86.6%)	1 :	260	(13.4%)		1935 (7.64%)	23395 (92.36%)
0 :	1675	(86.6%)											
1 :	260	(13.4%)											

Generated by summarytools (<https://github.com/dcomtois/summarytools>) 0.9.3 (R (<https://www.r-project.org/>) version 3.6.0)
2019-06-28

The data has 25,330 observations and 89 variables. A large portion of these variables are dummies: classifying observations into samples/cities (Bogota, Jamundi) or dummies created from some factor/categorical variable (d1993 for year == 1993, dmonth7 for month == 7, etc.).

Descriptive Replication

The paper itself is unclear about this data, as most reports are done with some stratifying. As a sanity check, let's replicate a descriptive result - the Population size (N) and Percentage awarded vouchers in each city and sample year, from Table 1 Panel A. We start by removing the NA value in id:

```
vouchers <- vouchers %>%
  filter(!is.na(id))

bog95 <- vouchers %>%
  filter(dbogota == 1 & d1995 == 1) %>%
  summarise(N = n(), pct_awarded_vouchers = sum(vouch0)/N) %>%
  mutate(place = "Bogota 1995")

bog97 <- vouchers %>%
  filter(dbogota == 1 & d1997 == 1) %>%
  summarise(N = n(), pct_awarded_vouchers = sum(vouch0)/N) %>%
  mutate(place = "Bogota 1997")

jam93 <- vouchers %>%
  filter(djamundi == 1 & d1993 == 1) %>%
  summarise(N = n(), pct_awarded_vouchers = sum(vouch0)/N) %>%
  mutate(place = "Jamundi 1993")

table1_panelA <- bind_rows(bog95, bog97, jam93) %>%
  select(place, N, pct_awarded_vouchers)

table1_panelA
```

place <chr>	N <int>	pct_awarded_vouchers <dbl>
Bogota 1995	4044	0.5880317

place <chr>	N <int>	pct_awarded_vouchers <dbl>
Bogota 1997	1770	0.8474576
Jamundi 1993	342	0.5000000
3 rows		

And thankfully we manage to replicate these basic descriptive results.

Main Result Replication

Angrist et al. 2002. estimate the lottery effects using the following regression model:

$$y_{ic} = X_i' \beta_0 + \alpha_0 Z_i + \delta_c + \epsilon_{ic}$$

Where y_{ic} is the dependent variable for child i in application cohort c (where c is defined by city and year), X_i is a vector of individual and survey characteristics (age, sex, whether the survey was telephone or in person), Z_i is an indicator of winning the lottery and δ_c is an applicant cohort effect. The main coefficient of interest is α_0 which represents the effect on y_{ic} of winning the PACES lottery.

In this section we will replicate one of the main results of the paper, the effect of lottery win on the probability to finish the 8th grade. This is shown on Table 3, column 4, row 10. The estimated α_0 is 0.094 with a standard error of 0.027, indicating a p-value for rejecting the null hypothesis of less than one percent. This model is estimated on a restricted version of the original data, where Bogota 1997 and some additional observations are exclude.

```

filtered_vouchers <- vouchers %>%
  filter(bog95smp == 1 | bog97smp == 1 | jam93smp == 1 | tab3smp1 == 1) %>%
  filter(bog95smp == 1)

fit <- estimatr::lm_robust(data = filtered_vouchers,
  formula = finish8 ~ vouch0 + svy + hsvisit + age + sex2 + strata2 +
    strata3 + strata4 + strata5 +stratams + dmonth2 + dmonth3 + dmonth4 + dmonth5 +
    dmonth6 + dmonth7 + dmonth8 + dmonth9 + dmonth10 + dmonth11 + dmonth12,
  se_type = "HC1")

huxtable::huxreg(fit, coefs=c("Win Lottery" ="vouch0"))

```

	(1)
Win Lottery	0.099 ***
	(0.027)
N	1171
R2	0.081

*** p < 0.001; ** p < 0.01; * p < 0.05.

Our estimate for α_0 is 0.099, very similar to that published in the paper (0.094). Note that the model controls for all the variables in X_i . The coefficient shown is just that for winning the lottery for brevity. Our number of observations is a bit higher (1, 171 in contrast to 1, 147), which might also account for the difference in the estimation of α_0

ML Extension

After we have replicated an important result from the paper, it's time to extend the analysis with some modern machine learning tools. In table 4, the authors estimate the same model as above but separately for boys and girls, obtaining that $\alpha_0^{boys} = 0.095$ and $\alpha_0^{girls} = 0.105$. In this extension, we will investigate heterogeneous effects with `causalTree`.

In short, `causalTree` (more precisely, Honest Causal Trees) constructs a regression/classification tree on one part of the data using a modified splitting rule optimized for treatment effects heterogeneity and ensuring that the propensity of treatment/control is similar in each leaf. The tree is grown on one split, which identifies partitions. The partitions are then used on the second split to predict outcomes. This can be done with a simple linear regression which produces valid standard errors.

The procedure requires setting some parameters such as minimum number of observations in each terminal node (leaf). To avoid complications, I use the *BreifIntro.pdf* example parameters in the *CausalTree* github page for these values.

```
set.seed(123)
filtered_vouchers2 <- filtered_vouchers %>%
  filter(!is.na(age)) %>%
  filter(!is.na(finish8)) #removing NAs, honest.causalTree doesn't deal with them properly
n <- nrow(filtered_vouchers2)
trIdx <- which(filtered_vouchers2$vouch0 == 1)
conIdx <- which(filtered_vouchers2$vouch0 == 0)

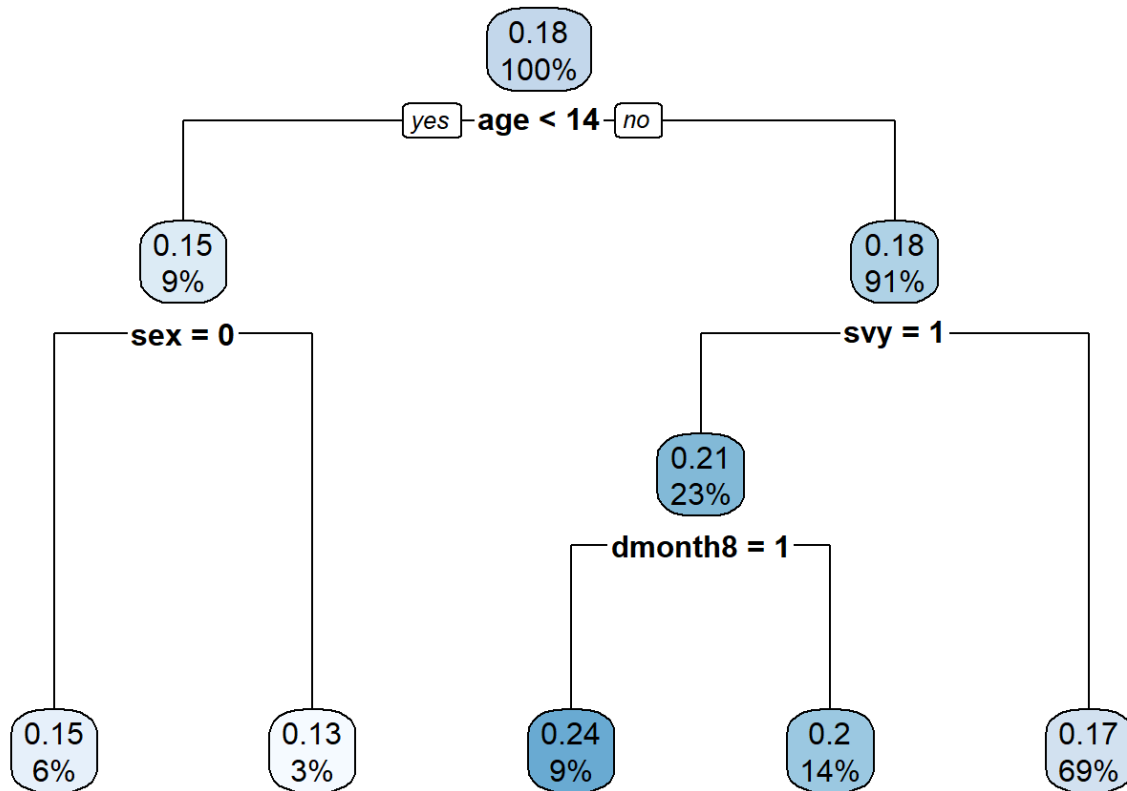
train_idx <- c(sample(trIdx, length(trIdx) / 2),
               sample(conIdx, (length(conIdx) / 2) + 1))

train_data <- filtered_vouchers2[train_idx, ]
est_data <- filtered_vouchers2[-train_idx, ]
honestTree <- honest.causalTree(
  formula = finish8 ~ svy + hsvisit + age + sex + strata2 + strat
a3 + strata4 + strata5 + stratams + dmonth2 + dmonth3 + dmonth4 + dmonth5 + dmon
th6 + dmonth7 + dmonth8 + dmonth9 + dmonth10 + dmonth11 + dmonth12,
  data = train_data,
  treatment = train_data$vouch0,
  est_data = est_data,
  est_treatment = est_data$vouch0,
  split.Rule = "CT", split.Honest = T,
  HonestSampleSize = nrow(est_data),
  split.Bucket = T, cv.option = "fit",
  cv.Honest = F, minsize = 5, na.action=na.omit
)
```



```
## [1] 6
## [1] "CTD"
```

```
opcp <- honestTree$cptable[,1][which.min(honestTree$cptable[,4])]
opTree <- prune(honestTree, opcp)
rpart.plot(opTree)
```



we see that the optimal tree is pretty shallow, with only 5 terminal nodes. All causal effects are positive. On the leftmost part of the tree we see that the causal estimate for girls ($\text{sex} == 0$) is larger than boys, similar to the original paper's table 4. Other allegedly heterogeneous groups are: interviewed in August ($\text{month} = 8$) and if the survey was completed with an updated questionnaire ($\text{svy} == 1$).

We now construct dummy variables for each leaf and estimate a simple linear regression of *finis8* (our dependent variable) on the leaf dummies, interacted with the treatment. The coefficients on $\text{treatment} \times \text{leaf}$ are the causal effects seen in the tree itself, but the standard errors are valid standard errors for the treatment effects and thus allow us to do basic statistical inference. Are these heterogeneous effects statistically significant or null?

```

est_data$leaf <- predict(opTree, est_data, type="vector")
est_data$leaf_fct <- as.factor(round(est_data$leaf, 3))

reg_model <- lm(formula = finish8 ~ -1 + leaf_fct + leaf_fct*vouch0 - vouch0,
               data = est_data)

huxtable::huxreg(reg_model, coefs=c("1st leaf" ="leaf_fct0.147:vouch0",
                                   "2nd leaf" ="leaf_fct0.13:vouch0",
                                   "3rd leaf" ="leaf_fct0.236:vouch0",
                                   "4th leaf" ="leaf_fct0.196:vouch0",
                                   "5th leaf" ="leaf_fct0.168:vouch0"))

```

	(1)
1st leaf	0.147 (0.159)
2nd leaf	0.130 (0.222)
3rd leaf	0.236 (0.129)
4th leaf	0.196 (0.102)
5th leaf	0.168 *** (0.046)
N	585
R2	0.693
logLik	-370.130
AIC	762.260
*** p < 0.001; ** p < 0.01; * p < 0.05.	

I have renamed the coefficients to better reflect the tree, and they are numbered from the left to right. The standard errors indicate that while all effects are positive, none except for the rightmost leaf are statically significant! This is not too surprising, as that leaf has 69% of the observations. The other leaves are much smaller, inducing large standard errors.

Conclusions

In this document I have replicated (part of) Angrist et al. 2002. and extended the analysis in it using a modern machine learning method for estimating heterogeneous treatment effects. My results indicate that the heterogeneous treatment effects are statically insignificant. However, this probably stems from

the low number of observations in the original paper, which is cut in half in order to conduct honest estimation of effects and standard errors.