EDS241: Take Home Final

Scout Leonard

03/17/2022

Read in and clean data

```
#read in data
housing <- read.csv(here("data", "KM_EDS241.csv")) %>%
  mutate(nearinc = as.factor(nearinc),
         year = as.factor(year))
```

Question A

Using the data for 1981, estimate a simple OLS regression of real house values on the indicator for being located near the incinerator in 1981. What is the house value "penalty" for houses located near the incinerator? Does this estimated coefficient correspond to the 'causal' effect of the incinerator (and the negative amenities that come with it) on housing values? Explain why or why not.

The code chunk below estimates a simple OLS regression of real house values on the indicator for being located near the incinerator in 1981.

```
#filter housing for only observations from 1981
housing_1981 <- housing %>%
  filter(year == 1981)

#model simple ols regression
ols_mod_1981 <- lm_robust(rprice ~ nearinc,
                          data = housing_1981)
```

The code chuck below inputs the outputs of the simple OLS regression `ols_mod_1981` into a table using the function `kable()`.

```
#create table with regression results
ols_mod_1981_table <- tidy(ols_mod_1981)

#print table
ols_mod_1981_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	101307.51	2944.810	0.0000000	95485.47	107129.6
nearinc1	-30688.27	6243.167	0.0000024	-43031.35	-18345.2

The house value "penalty" for houses located near the incinerator is 30688.27 dollars.

This estimated coefficient does respond to the 'causal' effect of the incinerator, although there can be other variables which effect the penalty for houses located near the incinerator, due to omitted variables bias.

Question B

Using the data for 1978, provide some evidence the location choice of the incinerator was not “random”, but rather selected on the basis of house values and characteristics. [Hint: in the 1978 sample, are house values and characteristics balanced by nearinc status?]

In the code chunk below, I filter the housing data for North Andover to only observations from 1978.

```
#filter housing for only observations from 1981
housing_1978 <- housing %>%
  filter(year == 1978)
```

In the next two sections, I filter first the 1978 data for houses near the incinerator and return summary statistics for each variable, then the summary statistics for houses far from the incinerator. Comparing these helped me to select variables to test for non-randomness using simple OLS regressions.

Houses Near the Incinerator: 1978 Summary Exploration

The code chunk below filters the data for observations from the housing in 1978 subset for only the houses near the incinerator. It returns the summary statistics for each of the seven variables for this subset of a subset.

```
#filter data for houses near incinerator
housing_1978_nearinc <- housing_1978 %>%
  filter(nearinc == 1)

#print summary table
summary(housing_1978_nearinc) %>%
  kable()
```

	year	age	rooms	area	land	nearinc	rprice
	1978:56	Min. : 0.00	Min. :4.000	Min. : 750	Min. : 1710	0: 0	Min. : 31000
	1981: 0	1st Qu.: 17.00	1st Qu.:5.000	1st Qu.:1336	1st Qu.: 8143	1:56	1st Qu.: 44000
	NA	Median : 28.00	Median :6.000	Median :1581	Median : 10684	NA	Median : 50950
	NA	Mean : 39.79	Mean :6.036	Mean :1835	Mean : 21840	NA	Mean : 63693
	NA	3rd Qu.: 56.00	3rd Qu.:6.250	3rd Qu.:2093	3rd Qu.: 17724	NA	3rd Qu.: 62250
	NA	Max. :189.00	Max. :9.000	Max. :5078	Max. :282704	NA	Max. :300000

Houses Far from the Incinerator: 1978 Summary Exploration

The code chunk below filters the data for observations from the housing in 1978 subset for only the houses far from the incinerator. It returns the summary statistics for each of the seven variables for this subset of a subset.

```
#filter data for houses far from incinerator
housing_1978_farinc <- housing_1978 %>%
  filter(nearinc == 0)

#print summary table
summary(housing_1978_farinc) %>%
  kable()
```

	year	age	rooms	area	land	nearinc	rprice
	1978:123	Min. : 0.00	Min. : 4.000	Min. : 960	Min. : 7858	0:123	Min. : 26000
	1981: 0	1st Qu.: 0.00	1st Qu.: 6.000	1st Qu.:1819	1st Qu.: 43560	1: 0	1st Qu.: 69000
	NA	Median : 2.00	Median : 7.000	Median :2071	Median : 44431	NA	Median : 84300
	NA	Mean : 12.75	Mean : 6.829	Mean :2075	Mean : 52569	NA	Mean : 82517
	NA	3rd Qu.: 9.00	3rd Qu.: 7.000	3rd Qu.:2443	3rd Qu.: 48593	NA	3rd Qu.: 94000
	NA	Max. :188.00	Max. :10.000	Max. :3792	Max. :544500	NA	Max. :142500

In comparing summary statistics for variables from 1978 Andover houses from near and far from the incinerator, we can see that houses far from the incinerator have higher mean and median area, land, and price, and lower mean and median age. This suggests location choice of the incinerator *may not* be random.

In the code chunks below, linear regressions return the average difference in outcomes for area, lot area, and age between houses near and far from the incinerator.

Simple OLS: Average Differences

Average Difference in Area of the House

```
mod1_1978 <- lm_robust(area ~ nearinc, data  = housing_1978)

#create table with regression results
mod1_1978_table <- tidy(mod1_1978)

#print table
mod1_1978_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	2074.7561	45.82799	0.0000000	1984.317	2165.195671
nearinc1	-240.1132	120.21379	0.0473153	-477.350	-2.876464

The results indicate that on average, houses near the incinerator have an area 240.11 square feet smaller than houses far from the incinerator.

Average Difference in Area of the Lot

```r
mod2_1978 <- lm_robust(land ~ nearinc, data = housing_1978)

#create table with regression results
mod2_1978_table <- tidy(mod2_1978)

#print table
mod2_1978_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	52569.06	4635.138	0.0000000	43421.81	61716.30
nearinc1	-30729.13	7140.626	0.0000278	-44820.85	-16637.41

On average, houses near the incinerator have 30729.13 fewer square feet of lot compared to houses far from the incinerator.

Average Difference in Age of the House

```r
mod3_1978 <- lm_robust(age ~ nearinc, data = housing_1978)

#create table with regression results
mod3_1978_table <- tidy(mod3_1978)

#print table
mod3_1978_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	12.74797	3.226605	0.0001123	6.38040	19.11553
nearinc1	27.03775	5.759048	0.0000053	15.67251	38.40298

On average, houses near the incinerator are 27.04 year older than houses far from the incinerator.

These differences show that on average, houses near and far from the incinerator have statistically different characteristics. This serves as evidence that the siting of the generator was not “random.”

Question C

Based on the observed differences in (b), explain why the estimate in (a) is likely to be biased downward (i.e., overstate the negative effect of the incinerator on housing values).

Based on the observed differences in (b), the estimate in (a) is likely to be biased downward (overstate the negative effect of the incinerator on housing values), because variables such as size and area probably already brought down the value of houses where the incinerator was placed, but the effect of the incinerator on housing values emphasized the difference in values between houses near and far from the incinerator in North Andover.

Question D

Use a difference-in-differences (DD) estimator to estimate the causal effect of the incinerator on housing values without controlling for house and lot characteristics. Interpret the magnitude and sign of the estimated DD coefficient.

The code chunk below runs a difference in difference regression to estimate the causal effect of the incinerator on housing values without controls for house and lot characteristics.

```
#add post treatment indicator and interaction effect (slide 12)
dd_df <- housing %>%
  mutate(post_treatment = factor(ifelse(year == 1981, 1, 0)),
         d = factor(ifelse(post_treatment == 1 & nearinc == 1, 1, 0))) # 1 if in treatment group and po

#difference in difference regression without controls
dd <- lm_robust(formula = rprice ~ nearinc +
           d +
             post_treatment,
         data = dd_df)

#create table with regression results
dd_table <- tidy(dd)

#print table
dd_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	82517.23	1878.277	0.0000000	78821.76	86212.692
nearinc1	-18824.37	6010.014	0.0018971	-30648.93	-6999.813
d1	-11863.90	8665.876	0.1719570	-28913.80	5185.997
post_treatment1	18790.29	3492.825	0.0000001	11918.24	25662.335

The DD value tells us that the houses near the incinerator in 1981 cost 11863.9 less than houses near the incinerator in 1978.

Question E

Report the 95% confidence interval for the estimate of the causal effect on the incinerator in (d).

The code chuck below calculates the 95% confidence interval for the causal effect on the incinerator.

```
dd_ci_95 <- confint(dd)

dd_ci_95
```

```
##                      2.5 %    97.5 %
## (Intercept)      78821.76 86212.692
## nearinc1        -30648.93 -6999.813
## d1              -28913.80  5185.997
## post_treatment1  11918.24 25662.335
```

The 95% confidence interval for the estimate of the causal effect on the incinerator is (d) is -28913.8 to 5186

Question F

How does your answer in (d) changes when you control for house and lot characteristics? Test the hypothesis that the coefficients on the house and lot characteristics are all jointly equal to 0.

The code chunk below runs a difference in difference regression to estimate the causal effect of the incinerator on housing values with controls for house and lot characteristics.

```
#difference in difference regression with controls
dd_2 <- lm_robust(formula = rprice ~ nearinc +
                    d +
                    post_treatment +
                    age +
                    rooms +
                    area +
                    land,
                  data = dd_df)

#create table with regression results
dd_2_table <- tidy(dd_2)

#print table
dd_2_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	-17688.8531406	11070.5839684	0.1110910	-39471.0243962	4093.3181151
nearinc1	3514.1411650	7149.5211107	0.6234024	-10553.0565252	17581.3388552
d1	-13320.1539955	6785.6622663	0.0505332	-26671.4332043	31.1252134
post_treatment1	13093.9318727	2795.3113134	0.0000042	7593.9555468	18593.9081987
age	-266.3382888	50.7157180	0.0000003	-366.1251166	-166.5514611
rooms	6969.0019675	1542.2646814	0.0000088	3934.4851337	10003.5188012
area	23.7821135	3.9011610	0.0000000	16.1062983	31.4579286
land	0.1268062	0.1370292	0.3554731	-0.1428086	0.3964211

When you control for house and lot characteristics, the difference in price between houses near the incinerator from 1978 to 1981 is greater than in the model in which we do not control for these characteristics. This smaller coefficient value, however, is no longer statistically significant at the 0.05 p value in the new model.

The code chunk below uses a linear hypothesis test to test the hypothesis that the coefficients on the house and lot characteristics are all jointly equal to 0.

```
#linear hypothesis
linearHypothesis(model = dd_2,
                 c("area = 0",
                   "land = 0",
                   "rooms = 0",
                   "age = 0"),
                   white.adjust  = "hc2")
```

```
## Linear hypothesis test
##
## Hypothesis:
## area = 0
## land = 0
## rooms = 0
## age = 0
##
## Model 1: restricted model
## Model 2: rprice ~ nearinc + d + post_treatment + age + rooms + area +
##     land
##
##   Res.Df Df  Chisq            Pr(>Chisq)
## 1    317
## 2    313  4 138.05 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With this P value, I would reject the null hypothesis that the coefficients on the house and lot characteristics are all jointly equal to 0.

Question G

Using the results from the DD regression in (f), calculate by how much did real housing values change (for the control group) on average between 1978 and 1981.

```r
#far house price in 1978
control_1978 <- dd_2$coefficients[1]

#far house price in 1981
control_1981 <- dd_2$coefficients[4]

#difference
real_housing_value_change <- control_1981 - control_1978
```

The real housing value change between 1978 and 1981 was 30782.79 dollars.

Question H

Explain (in words) what is the key assumption underlying the causal interpretation of the DD estimator in the context of the incinerator construction in North Andover.

A key assumption of the DD estimator corresponds to the parallel trend assumption, where the control group, the houses far from the incinerator, provide a valid counterfactual for the temporal evolution of mean outcomes (housing prices) in the treatment group, houses near the incinerator, in the absence of a change in treatment (placement of the incinerator).