

EDS241: Take Home Final

Scout Leonard

03/16/2022

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 chunk 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?]

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

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.

```
housing_1978_nearinc <- housing_1978 %>%
  filter(nearinc == 1)

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.

```
housing_1978_farinc <- housing_1978 %>%  
  filter(nearinc == 0)  
  
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 chunk below, linear regressions return the average difference in outcomes 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

Average Difference in Area of the Lot

```
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

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

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

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.

```
summary(dd)
```

The DD value tells us that by holding time constant, the houses near the incinerator cost 23896 less than houses far from the incinerator.

Question E

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

```
dd_ci_95 <- confint(dd)
```

```
dd_ci_95
```

```
##              2.5 %    97.5 %  
## (Intercept) 80195.422 88012.34  
## nearinc1    -32480.829 -15311.16  
## year1981      8593.456  21987.17
```

The 95% confidence interval for the estimate of the causal effect on the incinerator is (d) is -32480.83 to -15311.16

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.

```
dd_2 <- lm_robust(formula = rprice ~ nearinc +
                  year +
                  age +
                  rooms +
                  area +
                  land,
                  data = housing)

summary(dd_2)
```

```
##
## Call:
## lm_robust(formula = rprice ~ nearinc + year + age + rooms + area +
##          land, data = housing)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)    CI Lower
## (Intercept) -14144.3562 10765.2862  -1.3139 0.189843745252 -35325.5703
## nearinc1    -2604.8161  5819.3055  -0.4476 0.654738768621 -14054.5772
## year1981     9019.2767  2291.2664   3.9364 0.000101916484  4511.1007
## age         -260.6588    50.5237  -5.1591 0.000000440517  -360.0667
## rooms        6593.7854  1547.5197   4.2609 0.000026950500  3548.9666
## area          24.2933     3.9928   6.0843 0.000000003402    16.4372
## land          0.1197     0.1349   0.8878 0.375327708821   -0.1456
##
##              CI Upper  DF
## (Intercept)  7036.8580 314
## nearinc1     8844.9450 314
## year1981    13527.4528 314
## age         -161.2509 314
## rooms       9638.6042 314
## area         32.1493 314
## land         0.3851 314
##
## Multiple R-squared:  0.6039 ,    Adjusted R-squared:  0.5963
## F-statistic: 89.07 on 6 and 314 DF,  p-value: < 0.00000000000000022
```

When you control for house and lot characteristics, the difference in price between houses near and far from the incinerator, holding all else equal is, is smaller than in the model in which we do not control for these characteristics. This smaller Beta 1 value, however, is no longer statistically significant in the new model. This means that being near or far from the incinerator may not be a good indicator for the price of a house in North Andover.

```
#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 + year + age + rooms + area + land
##
##   Res.Df Df Chisq          Pr(>Chisq)
## 1      318
## 2      314   4 134.7 < 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.

Between 1978 and 1981, real housing values increased by 9019.28 dollars if all other variables are held equal.

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).