EDS241: Take Home Final

Scout Leonard

03/15/2022

Read in and clean data

```
#read in data
housing <- read.csv(here("data", "KM_EDS241.csv"))
```

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
nearinc	-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
Min. :1978	Min. : 0.00	Min. :4.000	Min. : 750	Min. : 1710	Min. :1	Min. : 31000
1st Qu.:1978	1st Qu.: 17.00	1st Qu.:5.000	1st Qu.:1336	1st Qu.: 8143	1st Qu.:1	1st Qu.: 44000
Median :1978	Median : 28.00	Median :6.000	Median :1581	Median : 10684	Median :1	Median : 50950
Mean :1978	Mean : 39.79	Mean :6.036	Mean :1835	Mean : 21840	Mean :1	Mean : 63693
3rd Qu.:1978	3rd Qu.: 56.00	3rd Qu.:6.250	3rd Qu.:2093	3rd Qu.: 17724	3rd Qu.:1	3rd Qu.: 62250
Max. :1978	Max. :189.00	Max. :9.000	Max. :5078	Max. :282704	Max. :1	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
Min. :1978	Min. : 0.00	Min. : 4.000	Min. : 960	Min. : 7858	Min. :0	Min. : 26000
1st Qu.:1978	1st Qu.: 0.00	1st Qu.: 6.000	1st Qu.:1819	1st Qu.: 43560	1st Qu.:0	1st Qu.: 69000
Median :1978	Median : 2.00	Median : 7.000	Median :2071	Median : 44431	Median :0	Median : 84300
Mean :1978	Mean : 12.75	Mean : 6.829	Mean :2075	Mean : 52569	Mean :0	Mean : 82517
3rd Qu.:1978	3rd Qu.: 9.00	3rd Qu.: 7.000	3rd Qu.:2443	3rd Qu.: 48593	3rd Qu.:0	3rd Qu.: 94000
Max. :1978	Max. :188.00	Max. :10.000	Max. :3792	Max. :544500	Max. :0	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
nearinc	-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
nearinc	-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.

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.

```
dd <- lm(formula = rprice ~ nearinc,
         data = housing)

se_dd <- starprep(dd, stat = c("std.error"), se_type = "HC2", alpha = 0.05)

summary(dd)
```

```
##
## Call:
## lm(formula = rprice ~ nearinc, data = housing)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65035 -18579  -5579  13965 233421
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    91036       2081  43.752 < 0.0000000000000002 ***
## nearinc       -24457       3805  -6.428       0.000000000472 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31210 on 319 degrees of freedom
## Multiple R-squared:  0.1147, Adjusted R-squared:  0.1119
## F-statistic: 41.32 on 1 and 319 DF,  p-value: 0.0000000004715
```

```
# DD REGRESSION
DD <- plm(rprice ~ nearinc,
               index = c("year"),
          model = "within",
          effect = "twoways",
          data = housing)

summary(DD)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = rprice ~ nearinc, data = housing, effect = "twoways",
##     model = "within", index = c("year"))
##
## Unbalanced Panel: n = 2, T = 142-179, N = 321
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
##  -74834  -12152       0   12152   74834
##
## Coefficients:
##         Estimate Std. Error t-value  Pr(>|t|)
## nearinc -25052.5     5900.9 -4.2455 0.0000395 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    119120000000
## Residual Sum of Squares: 105530000000
## R-Squared:      0.11406
## Adj. R-Squared: -1.025
## F-statistic: 18.0247 on 1 and 140 DF, p-value: 0.000039498
```

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 %
## nearinc -36618.02 -13486.96
```

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 <- plm(rprice ~ nearinc +
            area +
            land +
            rooms +
            age,
               index = c("year"),
          model = "within",
          effect = "twoways",
          data = housing)

summary(DD_2)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = rprice ~ nearinc + area + land + rooms + age, data = housing,
##     effect = "twoways", model = "within", index = c("year"))
##
## Unbalanced Panel: n = 2, T = 142-179, N = 321
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -49958.9  -6537.7      0.0   6537.7  49958.9
##
## Coefficients:
##             Estimate   Std. Error t-value           Pr(>|t|)
## nearinc -8878.357469  4413.997215 -2.0114          0.0462583 *
## area       23.376524     2.918406  8.0100 0.0000000000004591 ***
## land        0.020809     0.046259  0.4498          0.6535404
## rooms    8500.439559  2088.463008  4.0702 0.0000792649908328 ***
## age      -202.309147    51.541679 -3.9252          0.0001371 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    119120000000
## Residual Sum of Squares: 49347000000
## R-Squared:      0.58573
## Adj. R-Squared: 0.025251
## F-statistic: 38.4579 on 5 and 136 DF, p-value: < 0.000000000000000222
```

```
#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 + area + land + rooms + age
##
##   Res.Df Df  Chisq            Pr(>Chisq)
## 1    140
## 2    136  4 154.84 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question G

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

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.