EDS241: Take Home Final

Allie Cole

03/18/2022

1 Load and clean data

```
data <- read.csv(here::here("KM_EDS241.csv"))
```

2 Question A

(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.

```
data_1981 <- data %>%
  filter(year == 1981)

mod_1 <- lm_robust(data = data_1981,formula = rprice ~ nearinc)
huxreg(mod_1)
```

	(1)
(Intercept)	101307.515 ***
	(2944.810)
nearinc	-30688.274 ***
	(6243.167)
N	142
R2	0.165

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

The house value "penalty" for houses located near the incinerator is a decrease of 30688$ in the sales price. The estimated coefficient corresponds to the 'causal' effect of the incinerator a little bit, as we know that the housing price is most likely going to be lower in areas where the incinerator is present, however there could be some omitted variables bias as the incinerator could be placed there because of other variables that lower the price of the housing values.

3 Question B

(b) Using the data for 1978, provide some evidence the location 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?]

```
#creating a datatable for near the inc
data_1978_1 <- data %>%
  filter(year == 1978) %>%
  filter(nearinc == 1)

summary(data_1978_1) %>%
  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

```
#creating a datatable for away from the inc
data_1978_0 <- data %>%
  filter(year == 1978) %>%
  filter(nearinc == 0)

summary(data_1978_0) %>%
  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

The differences shown by the summary tables above show that on average, the houses near the incinerator are statistically different from houses far from the incinerator. Based on this we can see that the placement of the incinerator was not "random."

4 Question C

(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 because there are variables, other than the incinerator, that might play a role in bringing down the value of houses where the incinerator was placed, such as the size of the house and the amount of land it has.

5 Question D

(d) Use a difference-in-difference (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.

```
#changed after looking at Olivier's comments on Slack

#you need to make a the years into a binary = binary_time
#Then we make the dummy variable = dummy_var
 DD_data <- data %>%
  mutate(binary_time = case_when(year == 1981 ~ 1,
                                 year == 1978 ~ 0),
         dummy_var = nearinc*binary_time)

# Now we make a DD model
mod_DD <- lm_robust(data = DD_data, formula = rprice ~ binary_time + nearinc +
                      dummy_var)
summary(mod_DD)
```

```
##
## Call:
## lm_robust(formula = rprice ~ binary_time + nearinc + dummy_var,
##     data = DD_data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)    82517       1878  43.932
## binary_time    18790       3493   5.380
## nearinc       -18824       6010  -3.132
## dummy_var     -11864       8666  -1.369
##
## (Intercept) 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## binary_time 0.000001452358438347241330331025433580949268730364565271884202957153320312500000000000000000
## nearinc     0.001897088299070841775628704084510900429449975490570068359375000000000000000000000000000000
## dummy_var   0.171957017026234726397859731150674633681774139404296875000000000000000000000000000000000000
##             CI Lower CI Upper  DF
## (Intercept)    78822    86213 317
## binary_time    11918    25662 317
## nearinc       -30649    -7000 317
## dummy_var     -28914     5186 317
##
## Multiple R-squared:  0.1739 ,    Adjusted R-squared:  0.1661
## F-statistic: 17.72 on 3 and 317 DF,  p-value: 0.0000000001169
```

The estimated coefficient is -11863.9, meaning that on average houses that are near the incinerator decrease in value by 11863.9.

6 Question E

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

```
DD_ci_95 <- confint(mod_DD)
DD_ci_95
```

```
##                 2.5 %    97.5 %
## (Intercept)  78821.76 86212.692
## binary_time  11918.24 25662.335
## nearinc     -30648.93 -6999.813
## dummy_var   -28913.80  5185.997
```

```
#a way to get the variables by them selves
nearinc <- DD_ci_95[2,]
low <- round(DD_ci_95[[1]], 3)
high <- round(DD_ci_95[[2]], 3)
```

The 95% confidence interval for the estimate of the causal effect on the incinerator is [78821.763, 11918.238].

7 Question F

(f) How does your answer in (d) change 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_test <- lm_robust(data = DD_data, formula = rprice ~ binary_time +
                       nearinc +
                       age +
                       rooms +
                       area +
                       land)
summary(DD_test)
```

```
##
## Call:
## lm_robust(formula = rprice ~ binary_time + nearinc + age + rooms +
##     area + land, data = DD_data)
##
## 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
## binary_time   9019.2767  2291.2664  3.9364 0.000101916484   4511.1007
## nearinc      -2604.8161  5819.3055 -0.4476 0.654738768621 -14054.5772
## 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
## binary_time 13527.4528 314
## nearinc      8844.9450 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
```

```
nearinc_test <- DD_test$coefficients[[3]]
```

When you control for all the other variables the coefficient for `nearinc` decreases to -2604.8160767. Not only is it much lower but it is not statistically significant, so it is no longer an indicator you should use for the price of homes. Almost all the other variables are statistically significant, showing that all those are more likely to be better indicators compared to the proximity of incinerators.

```
hypo <- linearHypothesis(DD_test, c("age=0", "rooms=0", "area=0", "land=0"))
summary(hypo)
```

7.0.0.1 Now for the second part of the question

```
##      Res.Df          Df        Chisq         Pr(>Chisq)
##  Min.   :314   Min.   :4   Min.   :134.7   Min.   :0
##  1st Qu.:315   1st Qu.:4   1st Qu.:134.7   1st Qu.:0
##  Median :316   Median :4   Median :134.7   Median :0
##  Mean   :316   Mean   :4   Mean   :134.7   Mean   :0
##  3rd Qu.:317   3rd Qu.:4   3rd Qu.:134.7   3rd Qu.:0
##  Max.   :318   Max.   :4   Max.   :134.7   Max.   :0
##                NA's   :1   NA's   :1       NA's   :1
```

```
hypo$`Pr(>Chisq)`[2]
```

```
## [1] 0.00000000000000000000000000003851232
```

Because our p-value is 0, we can reject the null hypothesis that all coefficients on housing and lot characteristics are jointly equal to 0.

8 Question G

(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.

```
change <- DD_test$coefficients[[3]]
```

When using the results from the DD regression, the housing values increase by -2604.82 from 1978 to 1981.

9 Question H

(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.

The key assumption of the DD estimator in the context of the incinerator construction in North Andover is that homes that are not near an incinerator and homes that are near an incinerator will have the same trends in value if there was no incinerator built, this is known as a parallel trend assumption.