241 Take Home Final

Shale

3/18/2022

Data

The key variables for the analysis are: rprice (inflation-adjusted sales price of house), nearinc (=1 if house located near the incinerator, =0 otherwise), age (age of the house), land (square footage of the lot), area (square footage of the house), rooms (number of rooms in the house), and a year indicator (1978 or 1981).

```
data = read_csv(here("KM_EDS241.csv"))
```

Questions

(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.

```
da81 = data %>% filter(year == 1981)

ols1 = estimatr::lm_robust(data = da81, formula = rprice ~ nearinc)
ols1
```

```
##              Estimate Std. Error   t value     Pr(>|t|)  CI Lower CI Upper  DF
## (Intercept) 101307.51   2944.810 34.402059 3.632995e-70  95485.47 107129.6 140
## nearinc     -30688.27   6243.167 -4.915498 2.442350e-06 -43031.35 -18345.2 140
```

A simple OLS shows that the penalty for being near the incinerator in 1981 is about $30,688 (the mean house price near the incinerator is $30,688 less than the mean for houses farther away). However, this is not representative of the causal effect of adding the incinerator: the incinerator probably wouldn't have been placed in a place with expensive houses to begin with (NIMBY, city planners listening to rich people more than poor neighborhoods, etc).

(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?]

```
da78 = data %>% filter(year == 1978)

ols2 = estimatr::lm_robust(data = da78, formula = rprice ~ nearinc)
ols2
```

```
##              Estimate Std. Error   t value     Pr(>|t|)  CI Lower  CI Upper  DF
## (Intercept)  82517.23   1878.277 43.932406 3.948812e-97  78810.53 86223.927 177
## nearinc     -18824.37   6010.014 -3.132167 2.030868e-03 -30684.88 -6963.864 177
```

The above OLS regression for house prices in 1978 for the same area shows that houses near the future incinerator location are already on average worth $18,824 less than houses farther away. This supports the theory that the location for the incinerator was not random and was selected to be in a less wealthy area to begin with.

(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).

Because the characteristics that created a housing value difference in 1978 are still present in 1981, the estimate from (a) is biased downward. In other words, not all of the $30,000 difference can be reasonably attributed to the presence of the incinerator, because without the incinerator there was already a $18,000 difference. But the OLS from (a) cannot separate the effect of the incinerator and the static location effect that was present before the effect (under the parallel trends assumption): so the effect of the incinerator is added on to the $18,000 to make it look like $30,000.

(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.

```
# Using plm():
# DDmodel = plm::plm(data = data, formula = rprice ~ nearinc,
#                    index = c("year"), effect = "twoways", model = "within")
# DDmodel
```

```
# manual test

# time, treatment
m11 = data %>% filter(year == 1981 & nearinc == 1)
after_inc_m = mean(m11$rprice)
m10 = data %>% filter(year == 1981 & nearinc == 0)
after_far_m = mean(m10$rprice)
m01 = data %>% filter(year == 1978 & nearinc == 1)
before_inc_m = mean(m01$rprice)
m00 = data %>% filter(year == 1978 & nearinc == 0)
before_far_m = mean(m00$rprice)

D_control = after_far_m - before_far_m
D_inc = after_inc_m - before_inc_m
DD = D_inc - D_control
DD
```

```
## [1] -11863.9
```

```
data = data %>% mutate(year = as.factor(year))

# DD REGRESSION using lm package
DD_1 <- estimatr::lm_robust(formula = rprice ~ nearinc*year, data=data)

summary(DD_1)
```

```
##
## Call:
## estimatr::lm_robust(formula = rprice ~ nearinc * year, data = data)
##
## Standard error type:  HC2
```

```
##
## Coefficients:
##                   Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)          82517       1878  43.932 7.429e-137    78822    86213 317
## nearinc             -18824       6010  -3.132  1.897e-03   -30649    -7000 317
## year1981             18790       3493   5.380  1.452e-07    11918    25662 317
## nearinc:year1981    -11864       8666  -1.369  1.720e-01   -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 difference-in-differences model using `lm()` predicts a coefficient of -11,864 for the interaction between `nearinc` and `year`. This corresponds to a $11,864 lower price for houses near the incinerator in 1981 compared to what those same houses *would* have been worth in 1981 without the incinerator (thus, it is the causal effect of the incinerator, not biased by the systemic differences in housing values seen in 1978). This is done using the parallel trends assumption.

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

```
summary(DD_1)
```

```
##
## Call:
## estimatr::lm_robust(formula = rprice ~ nearinc * year, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##                   Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)          82517       1878  43.932 7.429e-137    78822    86213 317
## nearinc             -18824       6010  -3.132  1.897e-03   -30649    -7000 317
## year1981             18790       3493   5.380  1.452e-07    11918    25662 317
## nearinc:year1981    -11864       8666  -1.369  1.720e-01   -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 95% CI for the estimate in (d) is between -28914 and 5186. Because this range includes 0 (and because of the p-value of 0.17) this is not statistically significant at the 0.05 level.

(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_full <- estimatr::lm_robust(formula = rprice ~ nearinc*year + age + rooms + area + land,
                               data=data)

dds = summary(DD_full)
dds
```

```
##
## Call:
## estimatr::lm_robust(formula = rprice ~ nearinc * year + age +
```

```
##     rooms + area + land, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##                      Estimate Std. Error t value       Pr(>|t|)    CI Lower
## (Intercept)      -17688.8531  11070.584 -1.5978 0.111090982713 -39471.0244
## nearinc            3514.1412   7149.521  0.4915 0.623402359190 -10553.0565
## year1981          13093.9319   2795.311  4.6842 0.000004195095   7593.9555
## age                -266.3383     50.716 -5.2516 0.000000279088   -366.1251
## rooms              6969.0020   1542.265  4.5187 0.000008832216   3934.4851
## area                 23.7821      3.901  6.0962 0.000000003194     16.1063
## land                  0.1268      0.137  0.9254 0.355473122621     -0.1428
## nearinc:year1981 -13320.1540   6785.662 -1.9630 0.050533201725 -26671.4332
##                      CI Upper  DF
## (Intercept)       4093.3181 313
## nearinc          17581.3389 313
## year1981         18593.9082 313
## age               -166.5515 313
## rooms            10003.5188 313
## area                31.4579 313
## land                 0.3964 313
## nearinc:year1981    31.1252 313
##
## Multiple R-squared:  0.612 , Adjusted R-squared:  0.6034
## F-statistic: 79.94 on 7 and 313 DF,  p-value: < 2.2e-16
```

```
car::linearHypothesis(DD_full,c("age+rooms+area+land=0"), white.adjust = "hc2")
```

```
## Linear hypothesis test
##
## Hypothesis:
## age  + rooms  + area  + land = 0
##
## Model 1: restricted model
## Model 2: rprice ~ nearinc * year + age + rooms + area + land
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    314
## 2    313  1 18.765 0.00001479 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When controlling for house and lot characteristics (`age`, `rooms`, `area`, `land`) the effect of the incinerator+year interaction is stronger as seen in a reduced p-value from 0.17 in the model in (d) to 0.0505 here (this is or is not significant at the 0.05 level depending on rounding). The house and lot characteristics are jointly significant ($p < 0.0001$) and not equal to 0. NOTE: heteroskedasticity appears to be present in this dataset. The p-value for the interaction is calculated using the heteroskedasticity-robust `lm_robust()` function. If the regular `lm()` function is used, the p-value goes down to 0.011. The same is true in (e), where the reported heteroskedasticity-robust p-value of 0.17 is different than the p-value returned by `lm()`, which is 0.11.

```
# using lm() (not heteroskedasticity-robust)
DD_lm <- lm(formula = rprice ~ nearinc*year + age + rooms + area + land,
                              data=data)
```

```
summary(DD_lm)
```

```
##
## Call:
## lm(formula = rprice ~ nearinc * year + age + rooms + area + land,
##     data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -85798  -9571   -320   9004 139885
##
## Coefficients:
##                       Estimate   Std. Error t value        Pr(>|t|)
## (Intercept)      -17688.85314   9869.16141  -1.792          0.0740 .
## nearinc            3514.14117   3820.60400   0.920          0.3584
## year1981          13093.93187   2902.99282   4.510 0.0000091583350 ***
## age                -266.33829     38.48087  -6.921 0.0000000000254 ***
## rooms              6969.00197   1628.64592   4.279 0.0000249785589 ***
## area                 23.78211      2.04952  11.604         < 2e-16 ***
## land                  0.12681      0.03142   4.036 0.0000683935069 ***
## nearinc:year1981 -13320.15400   5198.13712  -2.562          0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20860 on 313 degrees of freedom
## Multiple R-squared:  0.612,  Adjusted R-squared:  0.6034
## F-statistic: 70.54 on 7 and 313 DF,  p-value: < 2.2e-16
```

(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 [for the control group].

The coefficient for the `year1981` term in the model in (f) shows the average increase in housing values for the control group (houses not near the incinerator) between 1978 and 1981 when controlling for house age, rooms, area, and land. On average, these houses increased in value by $13,094.

(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 here is the parallel trends assumption. That is, the calculation of the "base" rate of house price change between 1978-1981 for houses near the incinerator (i.e. what their prices would have been if the incinerator hadn't been built) is assumed to be the same as the rate for the non-treatment houses (in this case the houses farther away from the incinerator).