EDS241: Take Home Final

Alex Clippinger

03/17/2022

1 Data

The following code loads and cleans the data.

```
# Load data
km_data <- read_csv("KM_EDS241.csv") %>%
  mutate(nearinc = factor(nearinc))

# Create data frames for each year
km_81 <- km_data %>% filter(year==1981)
km_78 <- km_data %>% filter(year==1978)
```

2 Question 1

(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.

```
model1 <- lm_robust(formula = rprice ~ nearinc, data = km_81)

model1 %>%
  tidy() %>%
  dplyr::select(term, estimate, std.error, p.value) %>%
  knitr::kable()
```

term	estimate	std.error	p.value
(Intercept)	101307.51	2944.810	0.0000000
nearinc1	-30688.27	6243.167	0.0000024

The house value "penalty" for houses located near the incinerator (nearinc=1) is $-30688.27. This means that, based on this simple OLS regression, houses near the incinerator are, on average, worth $30,688 less than houses away from the incinerator. The estimated coefficient does not correspond to the causal effect of the incinerator because other confounding variables, such as age of the home, square footage, and number of rooms, are not taken into account.

(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?]

```
price_diff = mean(km_78[km_78$nearinc==0,]$rprice) - mean(km_78[km_78$nearinc==1,]$rprice)
area_diff = mean(km_78[km_78$nearinc==0,]$area) - mean(km_78[km_78$nearinc==1,]$area)
rooms_diff = mean(km_78[km_78$nearinc==0,]$rooms) - mean(km_78[km_78$nearinc==1,]$rooms)
```

Prior to "intervention" (i.e., the construction of the incinerator), the mean average value of a home was $18824.37 higher for the houses that would be away from the incinerator (control group) than for the houses that would be close to the incinerator in 1981 (treatment group). This positive difference indicates that homes further from the incinerator were valued higher (on average) prior to construction, which could mean that the location of construction was not random, but instead selected based on existing home value. Additionally, homes away from construction had 240.11 greater square footage and 0.79 more rooms, on average, supporting the claim that the location of construction was based on home characteristics. These relationships can be examined using simple OLS regression.

The first regression shows that the the average home value is statistically significantly lower for homes near the incinerator prior to construction.

term	estimate	std.error	p.value	outcome
(Intercept)	82517.23	1878.277	0.0000000	rprice
nearinc1	-18824.37	6010.014	0.0020309	rprice

formula: rprice ~ nearinc

The second regression shows that the average home square footage is statistically significantly ($p<0.05$) lower for homes near the incinerator prior to construction.

term	estimate	std.error	p.value	outcome
(Intercept)	2074.7561	45.82799	0.0000000	area
nearinc1	-240.1132	120.21379	0.0473153	area

formula: area ~ nearinc

The third regression shows that the average number of rooms was statistically significantly lower for homes near the incinerator prior to construction.

term	estimate	std.error	p.value	outcome
(Intercept)	6.829268	0.0718256	0.0000000	rooms
nearinc1	-0.793554	0.1589515	0.0000014	rooms

formula: rooms ~ nearinc

Lastly, the fourth regression shows that these two variables, rooms and area, have a statistically significant relationship with the price of a home.

term	estimate	std.error	p.value	outcome
(Intercept)	-18896.760	10197.416990	0.0655432	rprice
rooms	6360.812	2372.709690	0.0080423	rprice
area	26.837	6.396528	0.0000431	rprice

formula: rprice ~ rooms + area

(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).

The estimate in (a) is likely to be biased downward because of the evidence provided in (b), which shows that there is an existing average difference in housing values between the two groups prior to construction. In 1978, homes near the eventual incinerator location were, on average, lower value and possessed characteristics that correlate with decreased home value, such as fewer number of rooms and less square footage. Therefore, the coefficient in the estimate in (a) is partially capturing the negative effect on home value from these characteristics, causing it to overstate the negative effect of the incinerator.

(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.

```
km_dd <- km_data %>%
  # Create variable for post treatment period (1981) and for interaction
  mutate(post_treatment = factor(ifelse(year==1981, 1, 0)),
         D = factor(ifelse(post_treatment==1 & nearinc==1, 1, 0)))

model_dd <- lm_robust(rprice ~ D + nearinc + post_treatment, km_dd)

model_dd %>%
  tidy() %>%
  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
D1	-11863.90	8665.876	0.1719570	-28913.80	5185.997
nearinc1	-18824.37	6010.014	0.0018971	-30648.93	-6999.813
post_treatment1	18790.29	3492.825	0.0000001	11918.24	25662.335

The estimated DD coefficient indicates that home values are $-11863.9 less for houses near the incinerator in 1981 than in 1978. To explain further, from previous sections we know the mean difference pre and post-construction for homes near the incinerator is $-30688.27 and the mean difference pre and post-construction for homes away from the incinerator is $-18824.37. The DD coefficient is the difference between these two values.

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

As indicated in the DD regression results table in (d), the 95% confidence interval for the estimate of the causal effect of the incinerator is approximately (-28914, 5186).

(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.

```
model_dd2 <- lm_robust(rprice ~ D + nearinc + post_treatment + age + rooms + area + land,
                       data = km_dd)

model_dd2 %>%
  tidy() %>%
  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
D1	-13320.1539955	6785.6622663	0.0505332	-26671.4332043	31.1252134
nearinc1	3514.1411650	7149.5211107	0.6234024	-10553.0565252	17581.3388552
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 house and lot characteristics are added to the regression, the estimated DD coefficient indicates that home values are $-13320.15 less for houses near the incinerator in 1981 than in 1978. This estimate is a more pronounced difference in price than the estimate for the uncontrolled regression. Additionally, the p-value is lower ($p<0.051$) and confidence interval includes zero closer to the upper bound.

```
linearHypothesis(model = model_dd2, c("age=0", "rooms=0", "area=0", "land=0"),
                 test="F",
                 white.adjust="hc2")
```

```
## Linear hypothesis test
##
## Hypothesis:
## age = 0
## rooms = 0
## area = 0
## land = 0
##
## Model 1: restricted model
## Model 2: rprice ~ D + nearinc + post_treatment + age + rooms + area +
##     land
##
##   Res.Df Df      F                Pr(>F)
## 1    317
## 2    313  4 34.512 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

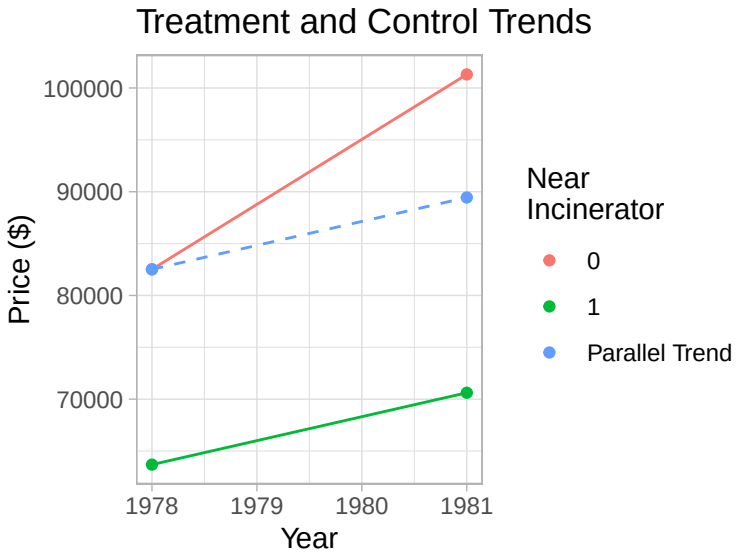
From the linear hypothesis above, the F-statistic is 34.512 (>10) with a corresponding p-value of less than 0.0001. Thus, the null hypothesis, the joint effect of the house and lot characteristics on price is equal to zero, can be rejected. This indicates that it is appropriate to control for these variables in the regression.

(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.

The change in real housing value for the control group can be derived from the intercept and `post_treatment` coefficient. The intercept coefficient is the home value when `nearinc`=0 and the year is 1978, with all other house and lot characteristics also equal to zero. The `post_treatment` coefficient is the home value when `nearinc`=0 and the year is 1981, with all other house and lot characteristics also equal to zero. Therefore, the change in real housing values for the control group, on average, is
13093.93 - (-17688.85) = $30782.79.

(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 underlying the causal interpretation of the DD estimator is that, in absence of the construction of the incinerator in North Andover, the control group (homes away from the construction location) and treatment group (homes near the construction location) would have the same trends in their outcomes in terms of home value.



The graph above displays the trends in home value for the treatment and control groups, not controlling for house and lot characteristics. It is clear that the trend for the control group is different than the trend in treatment, as shown by the disparity between the control group and the parallel trend. Thus, the difference between the parallel trend and control group after treatment is the DD estimate (in the plot above, this is equal to $11,863, which is the result from the uncontrolled regression in part d). Note that this graphic is meant to illustrate and support the explanation of the parallel trend assumption and is not intended to be interpreted as a thorough test of the assumption.