EDS241: Take Home Final

Alexandra Yousefivand

03/17/2022

```r
# Load data
data_raw <- read_csv(here("KM_EDS241.csv"))

data <- data_raw %>%
  mutate(year = as.factor(year),
         nearinc = as.factor(nearinc))
```

(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.

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

```r
model <- lm_robust(formula = rprice ~ nearinc, data = data_1981)
summary(model)
```

```
##
## Call:
## lm_robust(formula = rprice ~ nearinc, data = data_1981)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   101308       2945  34.402
## nearinc1      -30688       6243  -4.915
##                                                                           Pr(>|t|)
## (Intercept) 0.00000000000000000000000000000000000000000000000000000000000000000000003633
## nearinc1    0.00000244235036239296636972027909617821705978712998330593109130859375000000
##             CI Lower CI Upper  DF
## (Intercept)    95485   107130 140
## nearinc1      -43031   -18345 140
##
## Multiple R-squared:  0.1653 ,    Adjusted R-squared:  0.1594
## F-statistic: 24.16 on 1 and 140 DF,  p-value: 0.000002442
```

```r
penalty <- abs(round(model$coefficients[2]))
```

The house value "penalty" for houses located near the incinerator is 30688; in other words, on average, houses near the incinerator cost $30688 less than houses not near the incinerator. The very low p-value indicates that this is a statistically significant result and this estimated coefficient correlates with price. This might correspond to the 'causal' effect of the incinerator; however, there are other variables that may contribute to the difference in housing prices as well, which implies the possibility of omitted variables bias.

(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?]

```
# subset data
data_1978 <- data %>% filter(year == 1978)
```

```
data_nearinc <- data_1978 %>% filter(nearinc == 1)
data_not_nearinc <- data_1978 %>% filter(nearinc == 0)
```

```
# unadjusted mean difference
nearinc_mean_price <- mean(data_nearinc$rprice)

not_nearinc_mean_price <- mean(data_not_nearinc$rprice)

difference_price <- not_nearinc_mean_price - nearinc_mean_price
difference_price
```

```
## [1] 18824.37
```

Houses near the incinerator cost, on average, $18824 less than houses not near the incinerator.

```
# unadjusted mean difference
nearinc_mean_age <- mean(data_nearinc$age)

not_nearinc_mean_age <- mean(data_not_nearinc$age)

difference_age <- not_nearinc_mean_age - nearinc_mean_age
difference_age
```

```
## [1] -27.03775
```

Houses near the incinerator are, on average, 27 years older than houses not near the incinerator.

```
# unadjusted mean difference
nearinc_mean_rooms <- mean(data_nearinc$rooms)

not_nearinc_mean_rooms <- mean(data_not_nearinc$rooms)

difference_rooms <- not_nearinc_mean_rooms - nearinc_mean_rooms
difference_rooms
```

```
## [1] 0.793554
```

Houses near the incinerator have, on average, 0.79 fewer rooms than houses not near the incinerator.

```
# unadjusted mean difference
nearinc_mean_area <- mean(data_nearinc$area)

not_nearinc_mean_area <- mean(data_not_nearinc$area)

difference_area <- not_nearinc_mean_area - nearinc_mean_area
difference_area
```

```
## [1] 240.1132
```

Houses near the incinerator have, on average, 240 less square footage (of the house) than houses not near the incinerator.

```
# unadjusted mean difference
nearinc_mean_land <- mean(data_nearinc$land)

not_nearinc_mean_land <- mean(data_not_nearinc$land)

difference_land <- not_nearinc_mean_land - nearinc_mean_land
difference_land
```

```
## [1] 30729.13
```

Houses near the incinerator have, on average, 30729 less square footage (of the lot) than houses not near the incinerator.

```
# unadjusted mean difference using linear regression
model_age   <- lm_robust(formula = age ~ nearinc, data = data_1978)
summary(model_age)
```

```
##
## Call:
## lm_robust(formula = age ~ nearinc, data = data_1978)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    12.75      3.227   3.951 0.000112327     6.38    19.12 177
## nearinc1       27.04      5.759   4.695 0.000005329    15.67    38.40 177
##
## Multiple R-squared:  0.1106 ,    Adjusted R-squared:  0.1055
## F-statistic: 22.04 on 1 and 177 DF,  p-value: 0.000005329
```

```
model_rooms <- lm_robust(rooms ~ nearinc, data = data_1978)
summary(model_rooms)
```

```
##
## Call:
## lm_robust(formula = rooms ~ nearinc, data = data_1978)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   6.8293    0.07183  95.081
## nearinc1     -0.7936    0.15895  -4.992
##
## (Intercept) 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## nearinc1    0.000001419911976227534418978221131268924182222690433263778686523437500000000000000000000
##             CI Lower CI Upper  DF
## (Intercept)    6.688   6.9710 177
## nearinc1      -1.107  -0.4799 177
##
## Multiple R-squared:  0.1481 ,    Adjusted R-squared:  0.1433
## F-statistic: 24.92 on 1 and 177 DF,  p-value: 0.00000142
```

```
model_area  <- lm_robust(area ~ nearinc, data = data_1978)
summary(model_area)
```

```
##
## Call:
## lm_robust(formula = area ~ nearinc, data = data_1978)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   2074.8      45.83  45.273
## nearinc1      -240.1     120.21  -1.997
##
## (Intercept) 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## nearinc1    0.047315283494353183035840970660501625388860702514648437500000000000000000000000000000000
##             CI Lower CI Upper  DF
## (Intercept)   1984.3 2165.196 177
## nearinc1      -477.4   -2.876 177
##
## Multiple R-squared:  0.03091 ,   Adjusted R-squared:  0.02543
## F-statistic:  3.99 on 1 and 177 DF,  p-value: 0.04732
```

```
model_land  <- lm_robust(land ~ nearinc, data = data_1978)
summary(model_land)
```

```
##
## Call:
## lm_robust(formula = land ~ nearinc, data = data_1978)
##
## Standard error type:  HC2
##
## Coefficients:
```

```
##             Estimate Std. Error t value                    Pr(>|t|) CI Lower
## (Intercept)    52569       4635  11.341 0.00000000000000000000009291    43422
## nearinc1      -30729       7141  -4.303 0.00002777959403285577551050   -44821
##             CI Upper  DF
## (Intercept)    61716 177
## nearinc1      -16637 177
##
## Multiple R-squared:  0.08082 ,   Adjusted R-squared:  0.07563
## F-statistic: 18.52 on 1 and 177 DF,  p-value: 0.00002778
```

Additionally, each of these coefficients (or mean difference values) are statistically significant (p = 0.05). The above evidence implies the location choice of the incinerator was not "random", but rather selected on the basis of housing prices and characteristics.

(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), which is based on the observed differences in (b), is likely to be biased downward because this value captures the impact of other characteristics related to housing price (such as the age and size of the home) other than location relative to the incinerator. Before construction of the incinerator in 1978, homes near the incinerator site were older, smaller, and cost less, on average. Because the previous estimate absorbs the affect of these housing characteristics, it is likely to overstate the negative effect of the incinerator on housing values.

(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.

```r
diff_diff <- lm_robust(formula = rprice ~ nearinc, data = data)
summary(diff_diff)
```

```
##
## Call:
## lm_robust(formula = rprice ~ nearinc, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)    91035       1793  50.783
## nearinc1      -24457       4419  -5.534
##
## (Intercept) 0.0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## nearinc1    0.0000000652024090332346133343233152146467546117492020130157470703125000000000000000000000
##             CI Lower CI Upper  DF
## (Intercept)    87509    94562 319
## nearinc1      -33151   -15763 319
##
## Multiple R-squared:  0.1147 ,    Adjusted R-squared:  0.1119
## F-statistic: 30.63 on 1 and 319 DF,  p-value: 0.0000000652
```

The DD estimator is -24457, which implies houses near the incinerator are worth, on average, $24457 less than houses not near the incinerator.

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

```r
conf_low <- diff_diff$conf.low[[2]]
conf_high <- diff_diff$conf.high[[2]]
```

There is a 95% probability that the estimate of the causal effect on the incinerator is between $-33151 and $-15763.