# Final Examination

Lewis White

2023-03-20

### Read in the Data

```
housing <- read.csv("/Users/lewiswhite/MEDS/policy_eval/KM_EDS241.csv")
```

(a) Using the data for 1981, estimate a simple OLS regression of 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.

```
#filter to 1981 houses
housing_1981 <- housing %>%
  filter(year == 1981)

#simple linear regression with the near incinerator column
housing_1981_slr <- lm(formula = rprice ~ nearinc, data = housing_1981)

#including heteroscedasticy robust standard erros
se_models <- starprep(housing_1981_slr, stat = c("std.error"), se_type = "HC2", alpha = 0.05)

#displaying the results
stargazer(housing_1981_slr, se = se_models, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                               rprice
## -----------------------------------------------
## nearinc                   -30,688.270***
##                             (6,243.167)
##
## Constant                  101,307.500***
##                             (2,944.810)
##
## -----------------------------------------------
## Observations                    142
## R2                             0.165
## Adjusted R2                    0.159
## Residual Std. Error    31,238.040 (df = 140)
## F Statistic           27.730*** (df = 1; 140)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

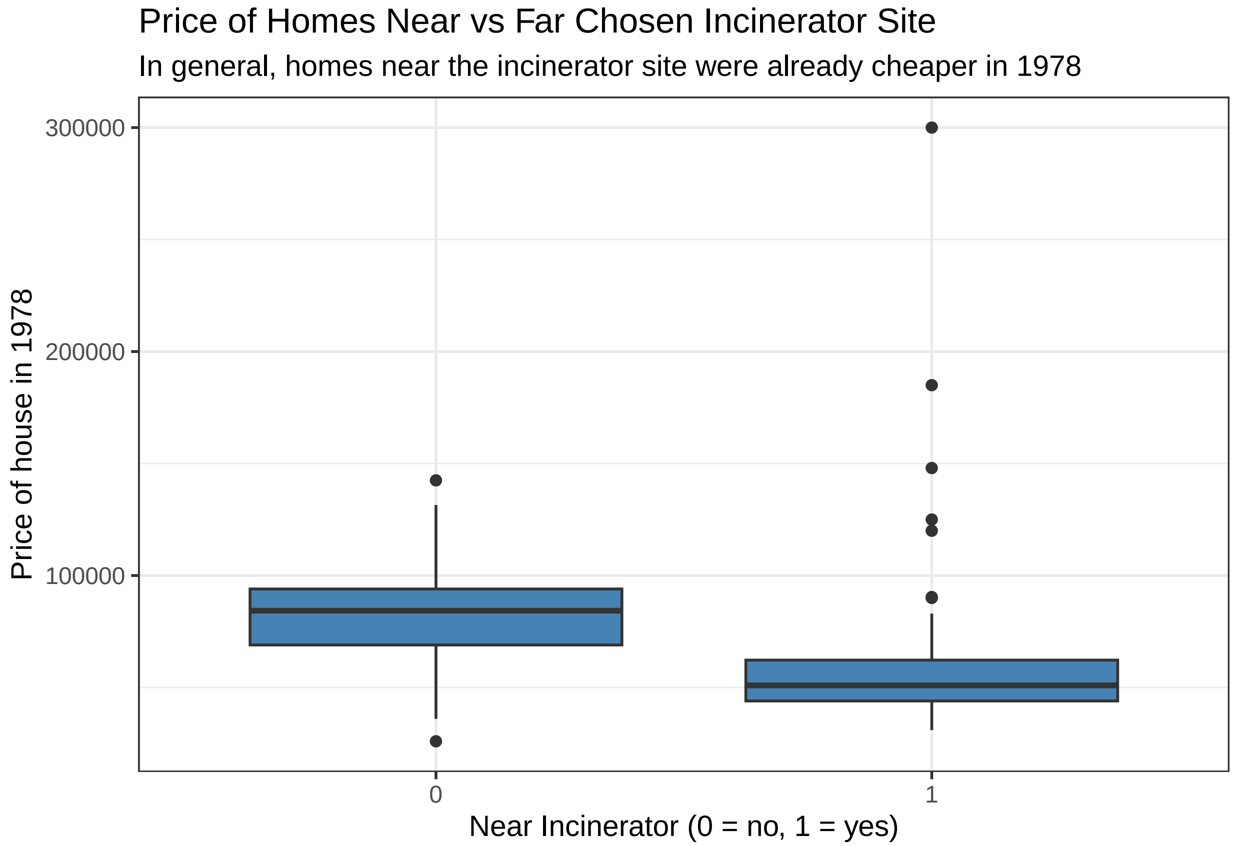
**According to this model, the house value "penalty" for houses located near the incinerator is 30,688.27 dollars. This means that houses near the incinerator in 1981 are on average 30,688.27 dollars cheaper than houses deemed far from the incinerator.**

**This coefficient does not correspond to the 'causal' effect of the incinerator. This model only takes into account the difference of home price between houses near the incinerator and far in 1981 without accounting for the price in 1978 or any of the house characteristics. In order to estimate the causal effect of the incinerator, these other variables must be taken into consideration.**

(b) Using the data for 1978, provide some evidence the location choice of the incinerator was not "random" using the data on house values and characteristics.

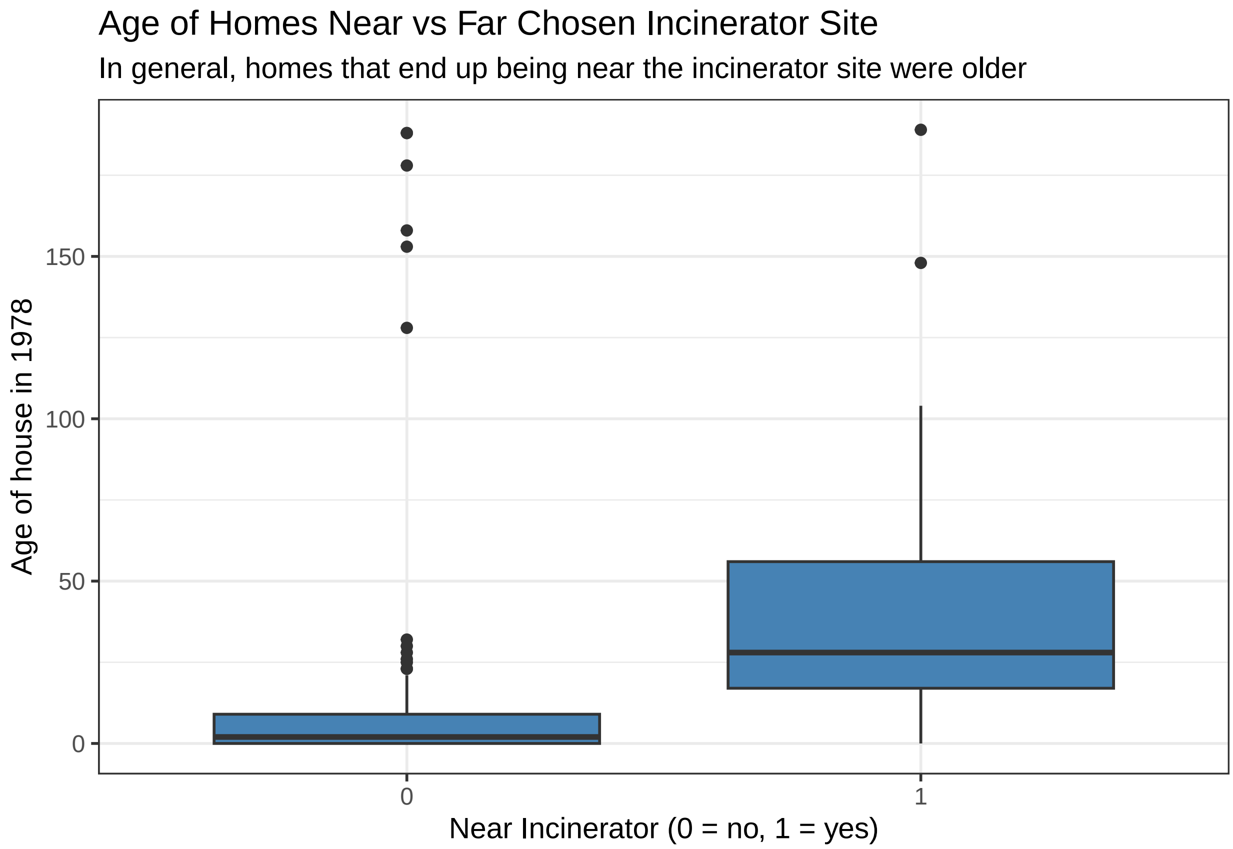
```
#filter to 1978 houses
housing_1978 <- housing %>%
  filter(year == 1978)

#comparing 1978 PRICE of homes near vs far decided incinerator location
ggplot(data = housing_1978, mapping = aes(x = as.factor(nearinc), y = rprice)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "Near Incinerator (0 = no, 1 = yes)",
       y = "Price of house in 1978",
       title = "Price of Homes Near vs Far Chosen Incinerator Site",
       subtitle = "In general, homes near the incinerator site were already cheaper in 1978") +
  theme_bw()
```

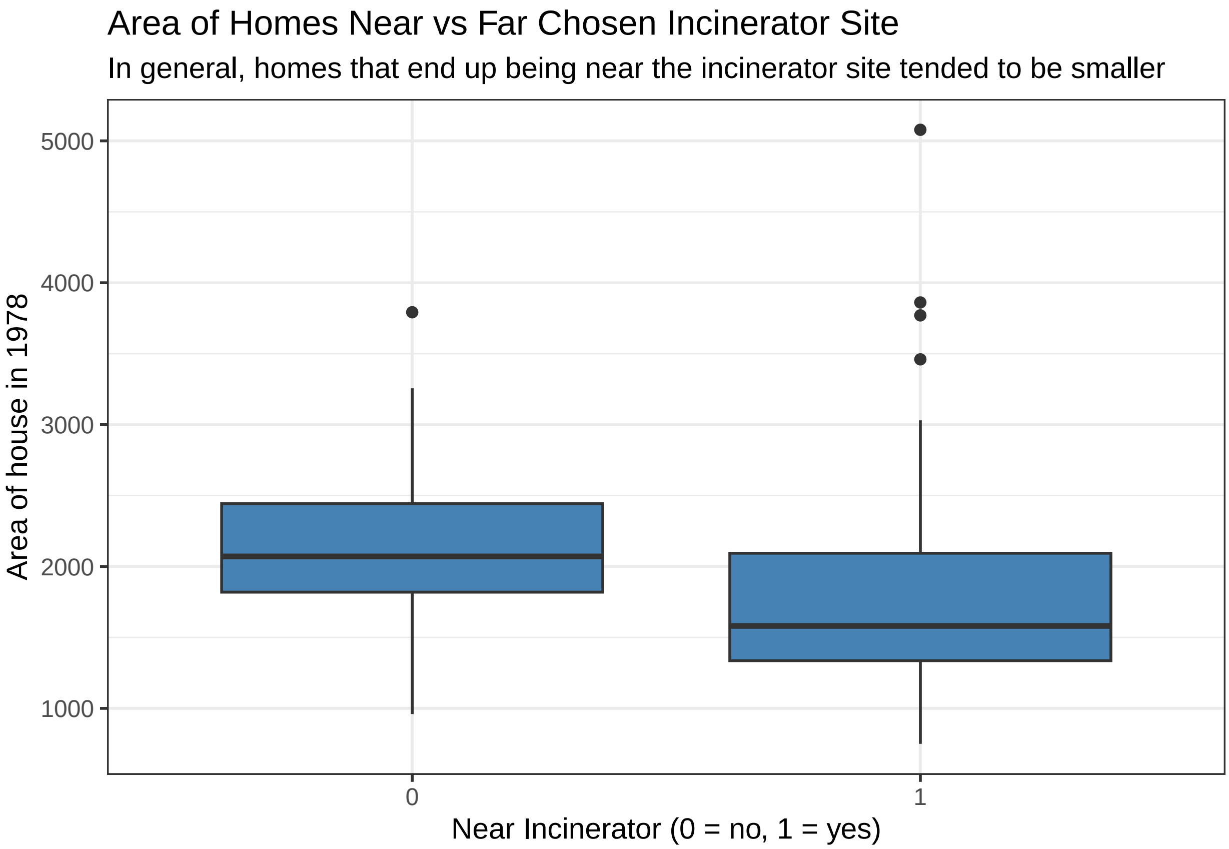


```
#comparing 1978 AGE of homes near vs far decided incinerator location
ggplot(data = housing_1978, mapping = aes(x = as.factor(nearinc), y = age)) +
  geom_boxplot(fill = "steelblue") +
```

```r
  labs(x = "Near Incinerator (0 = no, 1 = yes)",
       y = "Age of house in 1978",
       title = "Age of Homes Near vs Far Chosen Incinerator Site",
       subtitle = "In general, homes that end up being near the incinerator site were older") +
  theme_bw()
```



```r
#comparing 1978 AREA (sqft house) of homes near vs far decided incinerator location
ggplot(data = housing_1978, mapping = aes(x = as.factor(nearinc), y = area)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "Near Incinerator (0 = no, 1 = yes)",
       y = "Area of house in 1978",
       title = "Area of Homes Near vs Far Chosen Incinerator Site",
       subtitle = "In general, homes that end up being near the incinerator site tended to be smaller")
  theme_bw()
```



**For 1978 houses, the distributions for price of home, the age of homes, and the area of the homes are not very similar between houses that ended up being near the incinerator and houses that end up being far from the incinerator. Before the incinerator location was determined, houses near the eventual incinerator location tended to be cheaper, smaller, and older than houses far from it. This suggests that the location was not entirely random. This is speculation, but perhaps the incinerator location was intentionally placed in an area of less wealth because the community might have less time, money, and resources to fight against it.**

(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).

```
#showing that houses in 1978 that end up being near the incinerator
#were already cheaper than houses that end up being farther from the incinerator
housing_1978 %>%
  group_by(nearinc) %>%
  summarize(mean_value = mean(rprice))
```

```
## # A tibble: 2 x 2
##   nearinc mean_value
##     <int>      <dbl>
## 1       0     82517.
## 2       1     63693.
```

**In 1978, before the incinerator location was chosen, houses near the eventual location of the incinerator were on average roughly $20,000 cheaper than houses far from the eventual incinerator location. The estimate in a) didn't account for the previous home values and thus might attribute the pre-existing difference in home values as resulting from the erection 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.

```
#setting up the difference-in-differences data frame
housing_dd <- housing %>%
  mutate(treatment_time = ifelse(year > 1978, 1, 0)) %>%
  mutate(dd = nearinc * treatment_time)

#creating difference-in-differences model
DD_no_chars_hc2 <- lm(formula = rprice ~ dd + as.factor(nearinc) +
                        as.factor(treatment_time), data=housing_dd)

se_DD_no_chars_hc2 <- starprep(DD_no_chars_hc2, stat = c("std.error"),
                               se_type = "HC2", alpha = 0.05)

#viewing the results
stargazer(DD_no_chars_hc2, se = se_DD_no_chars_hc2, type="text")
```

```
##
## ======================================================
##                                 Dependent variable:
##                             ---------------------------
##                                       rprice
## ------------------------------------------------------
## dd                                 -11,863.900
##                                    (8,665.876)
##
## as.factor(nearinc)1              -18,824.370***
##                                    (6,010.014)
##
## as.factor(treatment_time)1        18,790.290***
##                                    (3,492.825)
##
## Constant                          82,517.230***
##                                    (1,878.277)
##
## ------------------------------------------------------
## Observations                           321
## R2                                    0.174
## Adjusted R2                           0.166
## Residual Std. Error           30,242.900 (df = 317)
## F Statistic                  22.251*** (df = 3; 317)
## ======================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

**The estimated difference in difference coefficient is -$11,863.9, indicating that the introduction of the incinerator was associated with an $11,863 reduction in the value of homes near the incinerator while holding location/time constant.**

Note: I tried using the CR2 cluster robust standard errors (because heteroskedasticity-robust standard errors assume uit are serially uncorrelated), but the values I got for the SEs were very close to 0. Given that cluster robust errors should be larger, I decided to stick with the HC2 SEs. Based on some online research, it seems like HC2 is actually commonly used when the number of clusters is relatively small, as the statistical validity of cluster-robust inference relies on large number of clusters (>30).

```
# #CODE THAT WOULD COMPARE CLUSTER ROBUST SE TO HETEROSCEDASTICITY ROBUST SE
# DD_no_chars_cluster <-
#   lm(
#     formula = rprice ~ dd + as.factor(nearinc) + as.factor(treatment_time),
#     data = housing_dd
#   )
#
# se_DD_no_chars_cluster <-
#   starprep(
#     DD_no_chars,
#     stat = c("std.error"),
#     se_type = "CR2",
#     clusters = housing$nearinc,
#     alpha = 0.05
#   )
#
# se_DD_no_chars_cluster
#
# se_models <- list(se_DD_no_chars_hc2[[1]],
#                   se_DD_no_chars_cluster[[1]])
#
# stargazer(
#   DD_no_chars_cluster,
#   DD_no_chars_cluster,
#   se = se_models,
#   keep = c("dd"),
#   type = "text"
# )
```

```
#CALCULATING MANUALLY
#(After treatment - before treatment) - (After control - before control)

housing %>%
  filter(year == 1978) %>%
  group_by(nearinc) %>%
  summarise(mean_rprice = mean(rprice))

housing %>%
  filter(year == 1981) %>%
  group_by(nearinc) %>%
  summarise(mean_rprice = mean(rprice))

#calculating the DD manually
(70619 - 63693) - (101308 - 82517) # -11865

#Calculating the DD estimator in this manner obtained a similar result
#(just a dollar added to the reduction in the home price for homes near
#the incinerator compared to the method used above).
```

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

```
#creating a confidence interval for the causal effect of the incinerator
confint(DD_no_chars_hc2, level = 0.95, vcov = vcovHC, type = "HC2")
```

```
##                                 2.5 %    97.5 %
## (Intercept)                    77152.10 87882.357
## dd                            -26534.67  2806.867
## as.factor(nearinc)1           -28416.45 -9232.293
## as.factor(treatment_time)1     10821.88 26758.691
```

**I am 95% confident that the introduction of the incinerator was associated with a change in home value for homes near the between -26534.67 and 2806.867 dollars. While the majority of this interval is well into the negatives, the confidence interval contains 0, so I am not able to claim that the incinerator caused a decrease in house prices for homes near the incinerator.**

(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.

```
#full model adding in the house/lot variables
DD_full_hc2 <- lm(formula = rprice ~ dd + as.factor(nearinc) + as.factor(treatment_time)
                  + age + rooms + area + land, data = housing_dd)

#calculating the HC2 standard erros
se_DD_full_hc2 <- starprep(DD_full_hc2, stat = c("std.error"),
                           se_type = "HC2", alpha = 0.05)

#creating a table for the results
stargazer(DD_full_hc2, se = se_DD_full_hc2, type="text")
```

```
##
## ======================================================
##                                Dependent variable:
##                            ---------------------------
##                                      rprice
## ------------------------------------------------------
## dd                                -13,320.150**
##                                    (6,785.662)
##
## as.factor(nearinc)1                 3,514.141
##                                    (7,149.521)
##
## as.factor(treatment_time)1        13,093.930***
##                                    (2,795.311)
##
## age                                -266.338***
##                                      (50.716)
##
## rooms                             6,969.002***
##                                    (1,542.265)
##
## area                                23.782***
##                                      (3.901)
##
## land                                  0.127
##                                      (0.137)
##
## Constant                          -17,688.850
##                                   (11,070.580)
##
## ------------------------------------------------------
## Observations                          321
## R2                                      0.612
## Adjusted R2                             0.603
## Residual Std. Error             20,857.870 (df = 313)
## F Statistic                   70.541*** (df = 7; 313)
## ======================================================
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

**When the house and lot characteristics are introudced to the model, the estimated difference in difference coefficient decreases to -$13,320.15. This indicates that the introduction of the incinerator was associated with an $13,320 reduction in the value of homes near the incinerator while holding location, time, and house/lot characteristics constant.**

**This change also resulting in the effect now being statistically significant at the p<0.05 significant level.**

```
#Testing the hypothesis that the coefficients on the house and lot characteristics are all jointly equa
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
linearHypothesis(DD_full_hc2, c("age=0", "rooms=0", "area=0", "land=0"), white.adjust = "hc2")
```

```
## Linear hypothesis test
##
## Hypothesis:
## age = 0
## rooms = 0
## area = 0
## land = 0
##
## Model 1: restricted model
## Model 2: rprice ~ dd + as.factor(nearinc) + as.factor(treatment_time) +
##     age + rooms + area + land
##
## Note: Coefficient covariance matrix supplied.
##
##   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
```

**To test that the coefficients of the house/lot characteristics are all jointly equal to 0, I ran a linear hypotheis test with heteroscedasicity robust SEs. The F score (34.512) resulted in a p-value near 0, indicating that at least one of these variables or a combination of them is significant in the model.**

(g) 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 that the control group (homes far from the incinerator) provides a valid counterfactual for the change in home price seen by homes near the incinerator if the incinerator had never been developed. This is the parellel trend assumption, which essentially states that the groups being compared (in this example, the homes near/far from the incinerator) would have followed similar trends in the absence of treatment. If this is true, we can attribute changes in the home price as a result of the treatment rather than pre-existing differences or differences that developed during the treatment.**

**It is possible to test this assumption by using event-study regression; however, more data (to better understand the previous / long term trend) may be necessary to complete this analysis.**