EDS_241_Final

Marie Rivers

03/17/2022

This assignment examines the impact of the opening of a garbage incinerator on housing values in North Andover, MA. The data for the exercise are a subset of the data in the paper: K.A. Kiel and K.T. McClain (1995): "House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor Through Operation," Journal of Environmental Economics and Management 28, 241-255.

Background

Construction of a new garbage incinerator in North Andover in the early 1980s was controversial due to the increases in ambient pollution that it would create. Rumors of the incinerator began after 1978. Construction started in 1981, and the incinerator began operating in 1985. In Economics, land market theory suggests that local amenities are capitalized in housing values, and predicts that the prices of houses located near the incinerator would fall compared to the price of houses located further away from the incinerator. By 1981, you can assume that all market participants had full information on the upcoming garbage incinerator, so that housing values had capitalized the upcoming arrival of the incinerator.

Read Data

```
data <- read_csv(here("data", "KM_EDS241.csv")) %>%
  mutate(nearinc = as.factor(nearinc)) %>%
  mutate(year = as.factor(year)) %>%
  mutate(age = na_if(age, 0))
```

Data Variables

- year = year house was sold (1978 is before rumors of the incinerator; 1981 = during construction of the incinerator)
- age = age of house in years
- rooms = number of rooms
- area = living area in square feet
- land = lot size in square feet
- nearinc = dummy variable (0 = not near incinerator; 1 = near incinerator)
- rprice = real house values

```
data_summary <- data %>%
  group_by(year, nearinc) %>%
  summarise(num_houses = n(),
```

year	nearinc	num_houses	mean_age	mean_rooms	mean_area	mean_land	mean_price
1978	0	123	21.2	6.8	2075	52569	82517
1978	1	56	40.5	6.0	1835	21840	63693
1981	0	102	15.5	6.8	2351	40251	101308
1981	1	40	36.1	6.2	1962	23164	70619

year	num_houses	mean_age	mean_rooms	mean_area	mean_land	mean_price
1978	179	29.4	6.6	2000	42955	76628
1981	142	22.8	6.6	2242	35438	92663

```
            mean_age = round(mean(age, na.rm = TRUE), 1),
            mean_rooms = round(mean(rooms), 1),
            mean_area = round(mean(area), 0),
            mean_land = round(mean(land), 0),
            mean_price = round(mean(rprice), 0))
```

```
data_summary_table <- data_summary %>%
  kable(col.names = c("year", "nearinc", "num_houses", "mean_age", "mean_rooms", "mean_area", "mean_lan
  kable_paper(full_width = FALSE) %>%
  row_spec(c(0), background = "lightgray")
data_summary_table
```

```
data_summary2 <- data %>%
  group_by(year) %>%
  summarise(num_houses = n(),
            mean_age = round(mean(age, na.rm = TRUE), 1),
            mean_rooms = round(mean(rooms), 1),
            mean_area = round(mean(area), 0),
            mean_land = round(mean(land), 0),
            mean_price = round(mean(rprice), 0))
```

```
# compare just year
data_summary_table2 <- data_summary2 %>%
  kable(col.names = c("year", "num_houses", "mean_age", "mean_rooms", "mean_area", "mean_land", "mean_p
  kable_paper(full_width = FALSE) %>%
  row_spec(c(0), background = "lightgray")
data_summary_table2
```

```
data_summary3 <- data %>%
  group_by(nearinc) %>%
  summarise(num_houses = n(),
            mean_age = round(mean(age, na.rm = TRUE), 1),
            mean_rooms = round(mean(rooms), 1),
            mean_area = round(mean(area), 0),
            mean_land = round(mean(land), 0),
            mean_price = round(mean(rprice), 0))
```

```
# compare just nearinc status
data_summary_table3 <- data_summary3 %>%
  kable(col.names = c("nearinc", "num_houses", "mean_age", "mean_rooms", "mean_area", "mean_land", "mea
```

nearinc	num_houses	mean_age	mean_rooms	mean_area	mean_land	mean_price
0	225	18.7	6.8	2200	46985	91035
1	96	38.9	6.1	1888	22392	66579

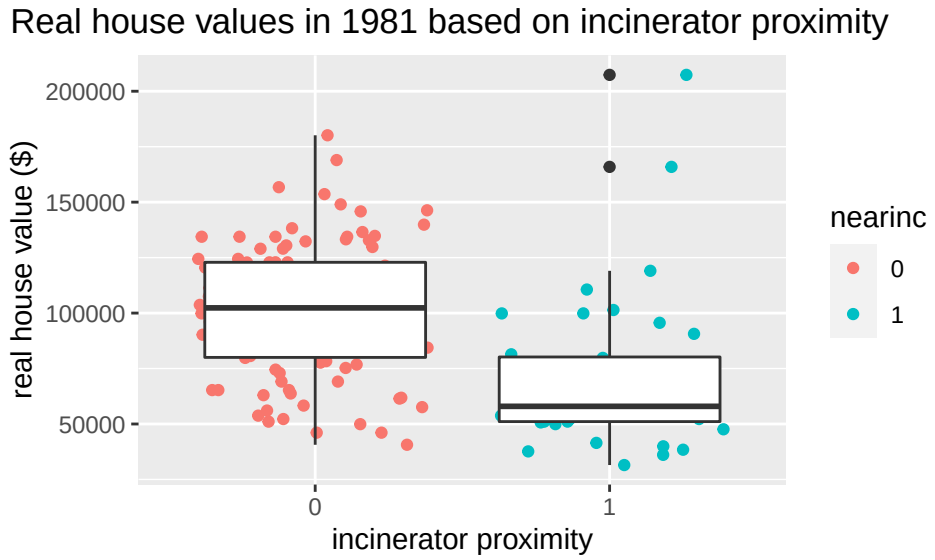
```
  kable_paper(full_width = FALSE) %>%
  row_spec(c(0), background = "lightgray")
data_summary_table3
```

Question 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 amenitites that come with it) on housing values? Explain why or why not?

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

```
ggplot(data = data_1981, aes(x = nearinc, y = rprice)) +
  geom_jitter(aes(color = nearinc)) +
  geom_boxplot() +
  labs(title = "Real house values in 1981 based on incinerator proximity",
       x = "incinerator proximity", y = "real house value ($)") +
  theme(plot.title.position = "plot")
```




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

```
huxreg("real house value" = model_a)
```

	real house value
(Intercept)	101307.515 ***
	(2944.810)
nearinc1	-30688.274 ***
	(6243.167)
N	142
R2	0.165

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

```
model_a_coef <- abs(round(model_a$coefficients[2], 2))
model_a_coef
```

```
## nearinc1
## 30688.27
```

```
model_a_se <- round(model_a[[2]][2], 2)
model_a_se
```

```
## nearinc1
##  6243.17
```

Answer A:

Based on a simple OLS regression of real house values on the indicator for being located near the incinerator in 1981, the house value "penalty" for houses located near the incinerator was $30688.27. The estimated coefficient does correspond to the negative 'causal effect' of the incinerator on housing values because the coefficient is negative and, based on the p-value, this coefficient is statistically different from 0. On average houses located near the incinerator had lower values than houses not near the incinerator. This analysis has omitted variable bias and excludes the fact that, on average, homes near the incinerator are older and smaller than homes not near the incinerator.

Question b:

Using the data for 1978, provide some evidence that 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?

nearinc	num_houses	mean_age	mean_rooms	mean_area	mean_land	mean_price
0	123	21.2	6.8	2075	52569	82517
1	56	40.5	6.0	1835	21840	63693

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

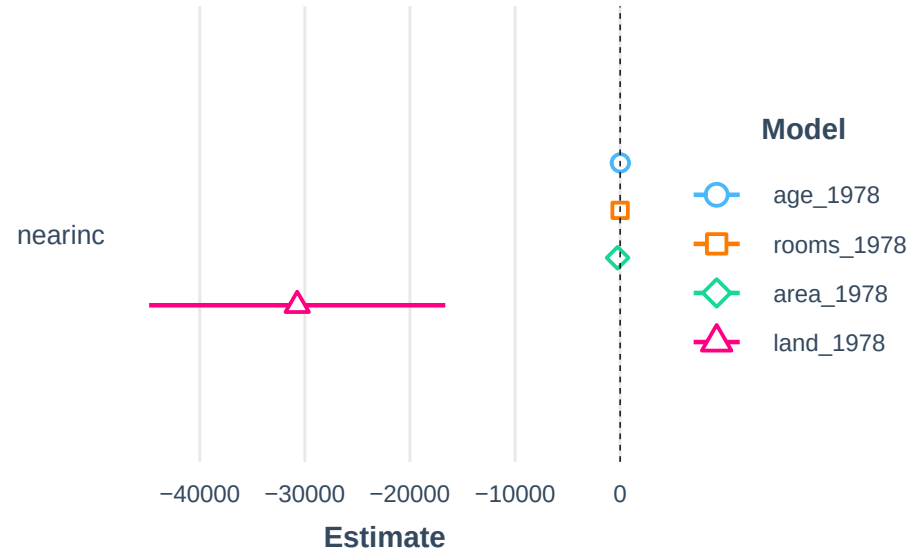
```
data_1978_summary <- data_1978 %>%
  group_by(nearinc) %>%
  summarise(num_houses = n(),
            mean_age = round(mean(age, na.rm = TRUE), 1),
            mean_rooms = round(mean(rooms), 1),
            mean_area = round(mean(area), 0),
            mean_land = round(mean(land), 0),
            mean_price = round(mean(rprice), 0))
```

```
data_1978_summary_table <- data_1978_summary %>%
  kable(col.names = c("nearinc", "num_houses", "mean_age", "mean_rooms", "mean_area", "mean_land", "mean
  kable_paper(full_width = FALSE) %>%
  row_spec(c(0), background = "lightgray")
data_1978_summary_table
```

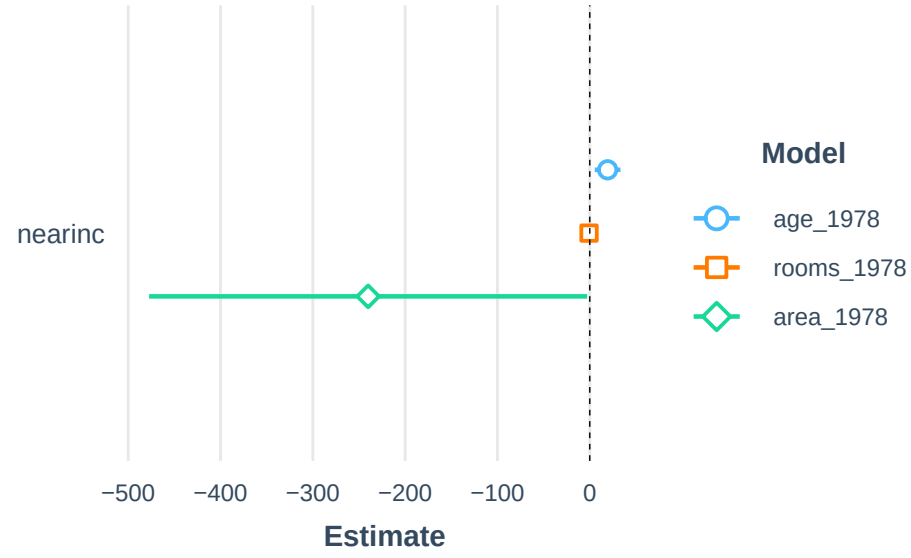
```
# models to estimate the difference in house characteristics for houses near and far from the future in
age_1978 <- lm(formula = age ~ nearinc, data = data_1978)
age_dif_1978 <- round(age_1978$coefficients[2], 2)

rooms_1978 <- lm(formula = rooms ~ nearinc, data = data_1978)
rooms_dif_1978 <- abs(round(rooms_1978$coefficients[2], 2))

area_1978 <- lm(formula = area ~ nearinc, data = data_1978)
area_dif_1978 <- abs(round(area_1978$coefficients[2], 2))

land_1978 <- lm(formula = land ~ nearinc, data = data_1978)
land_dif_1978 <- abs(round(land_1978$coefficients[2], 2))
```

Plots of the estimated difference and 95% confidence interval for age and number of rooms, living area, and lot size for houses near and far from the incinerator

```
plot_summs(age_1978, rooms_1978, area_1978, land_1978,
           scale = TRUE, ci_level = 0.95,
           robust = list("HC2", "HC2", "HC2", "HC2"),
           model.names = c("age_1978", "rooms_1978", "area_1978", "land_1978"))
```



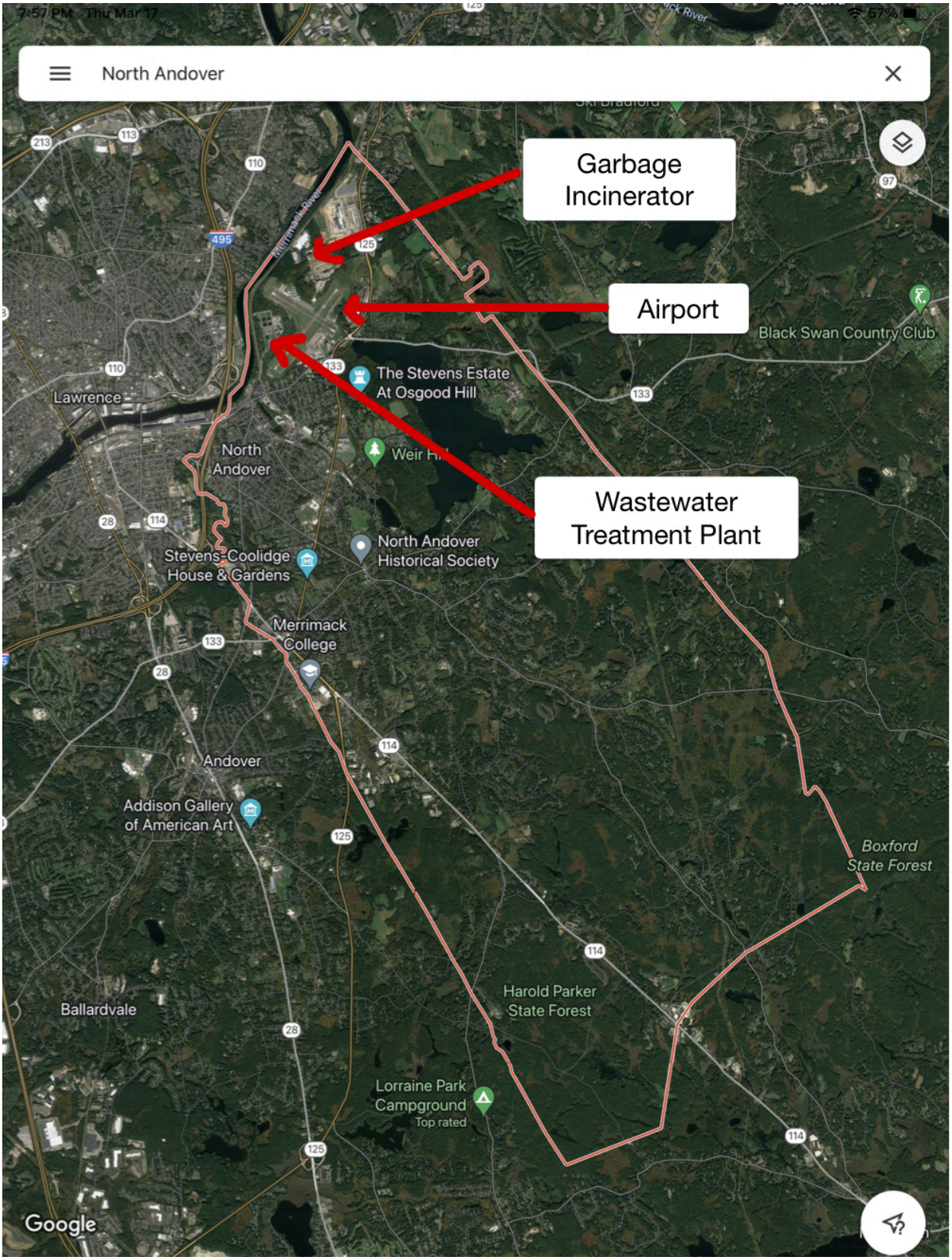

```
# separated out age, rooms, and area so they could be seen on a smaller scale
plot_summs(age_1978, rooms_1978, area_1978,
           scale = TRUE, i_level = 0.95,
           robust = list("HC2", "HC2", "HC2"),
           model.names = c("age_1978", "rooms_1978", "area_1978"))
```



Answer b:

Based on the summary table of 1978 sample, house values and characteristics are NOT balanced by `nearinc` status. On average, the houses near the future incinerator are older, smaller in living area and lot size, and have lower real prices. For example, houses near the future incinerator on average are 19.32 years older, have 0.79 fewer rooms, have living areas that are 240.11 square feet smaller, and have lot sizes that are 30729.13 square feet smaller than houses away from the future incinerator site. This suggests that the location choice of the incinerator was not random, but rather selected on the basis of house values and characteristics. If the incinerator location was randomly selected, then there would be no statistical difference in housing characteristics between houses near and far from the future incinerator site.

A consideration not reflected in the data is that the location choice of the incinerator is near the existing Lawrence Municipal Airport (established in 1934) and Greater Lawrence Sanitary District wastewater treatment plant (sanitary district established in 1968; sewage plant on-line April 1977). Both the airport and wastewater treatment facility are located in North Andover. The incinerator may have been intentional sited near the airport and wastewater treatment plant rather than intentionally sited near houses with lower values and less desirable characteristics. House values near the future incinerator could be lower due to the airport and wastewater treatment plant.


North Andover
Garbage Incinerator
Airport
Wastewater Treatment Plant
Black Swan Country Club
The Stevens Estate At Osgood Hill
Lawrence
North Andover
North Andover Historical Society
Stevens-Coolidge House & Gardens
Merrimack College
Andover
Addison Gallery of American Art
Boxford State Forest
Harold Parker State Forest
Ballardvale
Lorraine Park Campground
Top rated
Google

Question c:

Based on the observed differences in (b), explain why the estimate in (a) is likely to be biased downward (ie. overstate the negative effect of the incinerator on housing values).

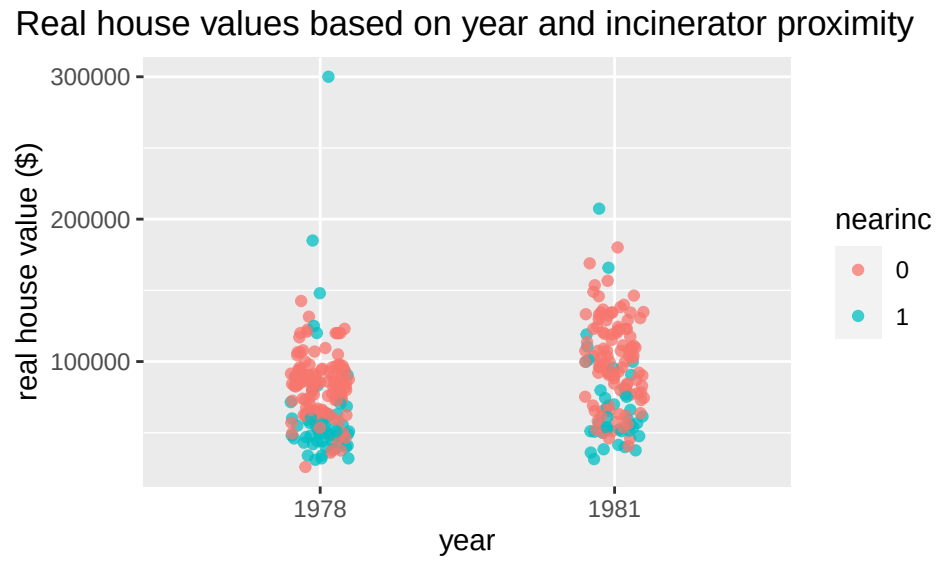
Answer c:

Based on the observed differences in (b), the estimate in (a) is likely to be biased downward (ie. overstate the negative effect of the incinerator on housing values) because characteristics such as old age, few rooms, small living area, and small lot size are generally less desirable and can all contribute to lower house values. The simple OLS regression in (a) omitted all of these variables. Since the location of the incinerator is not random, simple regression of outcomes on treatment status yields a biased estimate of the treatment effect. The treatment and control groups are not balanced.

Question 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 estiamted DD coefficient.

```
ggplot(data = data, aes(x = as.factor(year), y = rprice)) +
  geom_jitter(width = 0.1, aes(color = nearinc), alpha = 0.75) +
  labs(title = "Real house values based on year and incinerator proximity",
       x = "year", y = "real house value ($)") +
  theme(plot.title.position = "plot")
```




```
mean_inc_1978 <- data %>%
  filter(nearinc == 1) %>%
  filter(year == 1978) %>%
  summarise(mean(rprice)) %>%
  as.numeric()
```

```r
mean_inc_1981 <- data %>%
  filter(nearinc == 1) %>%
  filter(year == 1981) %>%
  summarise(mean(rprice)) %>%
  as.numeric()

mean_noinc_1978 <- data %>%
  filter(nearinc == 0) %>%
  filter(year == 1978) %>%
  summarise(mean(rprice)) %>%
  as.numeric()

mean_noinc_1981 <- data %>%
  filter(nearinc == 0) %>%
  filter(year == 1981) %>%
  summarise(mean(rprice)) %>%
  as.numeric()
```

```r
# compute the DD estimator for the treatment effect 'by hand' as a check to the linear model
DD_hand <- (mean_inc_1981 - mean_inc_1978) - (mean_noinc_1981 - mean_noinc_1978)
DD_hand
```

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

```r
# compute the DD estimator using a linear model
DD_model_d <- lm_robust(formula = rprice ~ nearinc * as.factor(year) + nearinc + as.factor(year), data =
```

```r
summary(DD_model_d)
```

```
##
## Call:
## lm_robust(formula = rprice ~ nearinc * as.factor(year) + nearinc +
##     as.factor(year), data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)                      82517       1878  43.932
## nearinc1                        -18824       6010  -3.132
## as.factor(year)1981              18790       3493   5.380
## nearinc1:as.factor(year)1981   -11864       8666  -1.369
##
## (Intercept)                   0.000000000000000000000000000000000000000000000000000000000000000000000000
## nearinc1                      0.001897088299070841775628704084510900429449975490570068359375000000000
## as.factor(year)1981           0.000000145235843834724133033102543358094926873036456527188420295715332
## nearinc1:as.factor(year)1981 0.171957017026234809664586578037415165454149246215820312500000000000000
##                               CI Lower CI Upper  DF
## (Intercept)                      78822    86213 317
## nearinc1                        -30649    -7000 317
## as.factor(year)1981              11918    25662 317
## nearinc1:as.factor(year)1981   -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
```

```
huxreg("real house value" = DD_model_d)
```

	real house value
(Intercept)	82517.228 ***
	(1878.277)
nearinc1	-18824.370 **
	(6010.014)
as.factor(year)1981	18790.287 ***
	(3492.825)
nearinc1:as.factor(year)1981	-11863.903
	(8665.876)
N	321
R2	0.174

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

```
DD_model_d_coef <- round(DD_model_d$coefficients[4], 0)
DD_model_d_coef
```

```
## nearinc1:as.factor(year)1981
##                        -11864
```

Answer d:

Based on the difference-in-difference (DD) estimator, the causal effect of the incinerator on housing values without controlling for house and lot characteristics is $-11864. The DD estimator represents how much more being near the incinerator effects the price in 1981 compared to 1978. The sign of the estimated DD coefficient is negative and that mean that houses for the treatment group (near the incinerator) have a lower price than houses away from the incinerator. The magnitude of this estimator is the difference between the two groups. Based on the p-value, the DD coefficient is not statistically significant. Between 1978 and 1981, there is no statistical evidence that house values changed more for houses near the incinerator than they did for houses not near the incinerator.

The difference-in-difference method estimates the average treatment effect as the difference between before and after averages of real house values for houses near the incinerator (treatment group) and not near the incinerator (control group). This method controls for the unobserved characteristics of treated and control groups and is useful for the North Andover incinerator problem because the treatment and control groups differ in time-invariant characteristics such as age, number of rooms, living area, and lot size. To adjust for pre-incinerator differences, the DD estimator subtracts the average house value during the pre-incinerator time period so that you can estimate how much of the difference is due to the treatment.

Question e:

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

```
ci_95_low <- round(confint(DD_model_d)[4], 0)
ci_95_low
```

```
## [1] -28914
```

```
ci_95_high <- round(confint(DD_model_d)[8], 0)
ci_95_high
```

```
## [1] 5186
```

Answer e:

The 95% confidence interval for the estimate of the causal effect on the incinerator in (d) is between -28914 and 5186. This means that there is a 95% chance that this interval includes the true causal effect of the incinerator on housing values (without controlling for house and lot characteristics). Note that the 95% confidence interval crosses 0 so the DD estimate is not statistically different from zero.

Question 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_model_f <- lm_robust(formula = rprice ~ nearinc * year + nearinc + year + age + rooms + area + land,
                        data = data)
```

```
summary(DD_model_f)
```

```
##
## Call:
## lm_robust(formula = rprice ~ nearinc * year + nearinc + year +
##     age + rooms + area + land, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##                      Estimate  Std. Error t value     Pr(>|t|)    CI Lower
## (Intercept)       -12825.81781 10422.92556 -1.2305 0.2198850582 -33373.9345
## nearinc1           -1076.96907  4395.61535 -0.2450 0.8066903538  -9742.6375
## year1981            8443.37290  2991.56692  2.8224 0.0052291574   2545.6942
## age                 -205.50467    49.01561 -4.1926 0.0000407838   -302.1357
## rooms               7319.94435  1343.19363  5.4497 0.0000001420   4671.9259
```

```
## area                  20.33832     3.41888  5.9488 0.0000000113     13.5982
## land                   0.05693     0.08835  0.6444 0.5200522718     -0.1172
## nearinc1:year1981  -4470.41217  5705.66040 -0.7835 0.4342219891 -15718.7488
##                       CI Upper  DF
## (Intercept)         7722.2989 208
## nearinc1            7588.6994 208
## year1981           14341.0516 208
## age                 -108.8736 208
## rooms               9967.9628 208
## area                  27.0784 208
## land                   0.2311 208
## nearinc1:year1981   6777.9245 208
##
## Multiple R-squared:  0.6068 ,    Adjusted R-squared:  0.5936
## F-statistic: 50.04 on 7 and 208 DF,  p-value: < 0.00000000000000022
```

```
huxreg("real house value" = DD_model_f)
```

```
DD_model_f_coef <- round(DD_model_f$coefficients[8], 0)
DD_model_f_coef <- as.numeric(DD_model_f_coef)
DD_model_f_coef
```

```
## [1] -4470
```

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

Answer f:

Based on the difference-in-difference (DD) estimator, the causal effect of the incinerator on housing values when controlling for house and lot characteristics is $-4470. The sign of the estimated DD coefficient when controlling for house/lot characteristics is negative (same as for part d) which means that houses near the incinerator have a lower price than houses away from the incinerator. The magnitude of the estimated DD coefficient is smaller when controlling for house/lot characteristics. This means that there is less of a difference between the value of houses near and away from the incinerator when controlling for house/lot characteristics. Based on the p-value, the DD coefficient is not statistically significant; between 1978 and 1981, house values may not have changed more for houses near the incinerator than they did for houses not near the incinerator.

Since the p-value of the linear hypothesis test is statistically significant, we reject the null hypothesis that the coefficients on the house and lot characteristics are all jointly equal to 0.

Question g:

Using the results from DD regression in (f), calculate by how much did real housing values change on average between 1978 and 1981 for the control group (not near the incinerator).

	real house value
(Intercept)	-12825.818
	(10422.926)
nearinc1	-1076.969
	(4395.615)
year1981	8443.373 **
	(2991.567)
age	-205.505 ***
	(49.016)
rooms	7319.944 ***
	(1343.194)
area	20.338 ***
	(3.419)
land	0.057
	(0.088)
nearinc1:year1981	-4470.412
	(5705.660)
N	216
R2	0.607

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

```
DD_model_f_coef_control <- round(DD_model_f$coefficients[3], 0)
DD_model_f_coef_control
```

```
## year1981
##     8443
```

Answer g:

On average, real housing values increased by $8443 between 1978 and 1981 for houses not near the incinerator when controlling for house and lot characteristics. This effect is statistically significant.

Res.Df	Df	Chisq	Pr(>Chisq)
212			
208	4	135	3.41e-28

Question 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.

Answer h:

The key assumption underlying the causal interpretation of the DD estimator in the context of the incinerator construction in North Andover is the parallel trend assumption. The control group (houses away from the incinerator) provides a valid counterfactual for the temporal evolution of the mean outcomes in the treatment group in absence of a change in treatment. A given house can't be near the incinerator one year and not near the incinerator another year.