EDS241: Take Home Final

Peter Menzies

03/18/2022

```
# quick function to make formatted table
tablr <- function(obj) {

    obj %>%
    tidy() %>%
    xtable()
}
```

Reading in data

```
df <- read_csv(here("data", "KM_EDS241.csv")) %>%
  clean_names() %>%
  mutate(nearinc = as.factor(nearinc))
```

(a) OLS regression of real house values on the indicator for being located near the incinerator in 1981.

```
df_81 <- df %>%
  filter(year == "1981")
```

```
ols_a <- lm_robust(rprice ~ nearinc, df_81)
```

```
tablr(ols_a)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	101307.51	2944.81	34.40	0.00	95485.47	107129.56	140.00	rprice
2	nearinc1	-30688.27	6243.17	-4.92	0.00	-43031.35	-18345.20	140.00	rprice

The estimated "penalty" in value for houses near the incinerator based on the above OLS regression is -30688.27 dollars.

This estimate does *not* correspond to the causal effect of being near the incinerator on housing values. The regression does not include house value prior to construction of the incinerator nor the other observed determinants of housing value included in the dataset (`age`, `rooms`, `area`, `land`—which are in fact significantly correlated with both `rprice` and `nearinc`). Additionally, the regression does not control for unobserved determinants of housing value which may also be unbalanced between treated and non-treated houses. The estimator is thus subject to omitted variable bias and we cannot infer causality.

(b) Provide evidence that the location choice of the incinerator was not "random", but rather selected on the basis of house values and characteristics.

```
df_78 <- df %>%
  filter(year == 1978)
```

```
ols_b1 <- lm_robust(rprice ~ nearinc, df_78)
ols_b2 <- lm_robust(area ~ nearinc, df_78)
ols_b3 <- lm_robust(rooms ~ nearinc, df_78)
```

```
# rprice ~ nearinc
tablr(ols_b1)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	82517.23	1878.28	43.93	0.00	78810.53	86223.93	177.00	rprice
2	nearinc1	-18824.37	6010.01	-3.13	0.00	-30684.88	-6963.86	177.00	rprice

```
# area ~ nearinc
tablr(ols_b2)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	2074.76	45.83	45.27	0.00	1984.32	2165.20	177.00	area
2	nearinc1	-240.11	120.21	-2.00	0.05	-477.35	-2.88	177.00	area

```
# rooms ~ nearinc
tablr(ols_b3)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	6.83	0.07	95.08	0.00	6.69	6.97	177.00	rooms
2	nearinc1	-0.79	0.16	-4.99	0.00	-1.11	-0.48	177.00	rooms

The above regressions show that house value, house square footage, and number of rooms all had statistically significant correlations with being near the incinerator prior to its construction, thus it is highly unlikely that the location choice of the incinerator was "random".

(c) Based on the observed differences in (b), explain why the estimate in (a) is likely to be biased downward

As shown in (b), prior to incinerator construction (and rumors of such), houses near the incinerator location-to-be tended to be lower in value and were less likely to possess characteristics that are highly correlated with higher value than those not near the incinerator location. This means that our initial regression in (a) is likely to be biased downward because we did not control for those imbalances in pre-treatment characteristics, and thus the estimated effect of being near the incinerator is attempting to account for those differences to some extent.

(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

```
df <- df %>%
  # creating diff-in-diffs variable indicating treated AND after construction started
  mutate("DD" = ifelse(year == 1981 & nearinc == 1, 1, 0) %>%
           as.factor())
```

```
DD1 <- lm_robust(rprice ~ nearinc + as.factor(year) + DD, df)
```

```
tablr(DD1)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	82517.23	1878.28	43.93	0.00	78821.76	86212.69	317.00	rprice
2	nearinc1	-18824.37	6010.01	-3.13	0.00	-30648.93	-6999.81	317.00	rprice
3	as.factor(year)1981	18790.29	3492.83	5.38	0.00	11918.24	25662.34	317.00	rprice
4	DD1	-11863.90	8665.88	-1.37	0.17	-28913.80	5186.00	317.00	rprice

```
# calculating diff-in-diffs by hand (just to confirm the above coefficient on DD)
df_before <- df %>%
  filter(year == 1978)

df_after <- df %>%
  filter(year == 1981)

mean_t_before <- mean(df_before[df_before$nearinc == 1,]$rprice)
mean_t_after <- mean(df_after[df_after$nearinc == 1,]$rprice)
mean_c_before <- mean(df_before[df_before$nearinc == 0,]$rprice)
mean_c_after <- mean(df_after[df_after$nearinc == 0,]$rprice)

(mean_t_after - mean_t_before) - (mean_c_after - mean_c_before)
```

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

The estimated causal effect of the incinerator on housing values without controlling for house and lot characteristics is -11863.9 dollars. This represents an estimated *decrease* of 11863.9 dollars in housing value for houses near the incinerator after the construction began. The estimated effect is lower in magnitude than that of our initial simple OLS regression which makes sense as we're now employing difference-in-differences in order to control for imbalances in pre-treatment characteristics.

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

The 95% confidence interval for the estimate of the causal effect of the incinerator in (d) is -28913.8 to 5185.997 (dollars).

(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.

```
DD2 <- lm_robust(rprice ~ nearinc + as.factor(year) + DD + land + area + rooms + age, df)
```

```
tablr(DD2)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	-17688.85	11070.58	-1.60	0.11	-39471.02	4093.32	313.00	rprice
2	nearinc1	3514.14	7149.52	0.49	0.62	-10553.06	17581.34	313.00	rprice
3	as.factor(year)1981	13093.93	2795.31	4.68	0.00	7593.96	18593.91	313.00	rprice
4	DD1	-13320.15	6785.66	-1.96	0.05	-26671.43	31.13	313.00	rprice
5	land	0.13	0.14	0.93	0.36	-0.14	0.40	313.00	rprice
6	area	23.78	3.90	6.10	0.00	16.11	31.46	313.00	rprice
7	rooms	6969.00	1542.26	4.52	0.00	3934.49	10003.52	313.00	rprice
8	age	-266.34	50.72	-5.25	0.00	-366.13	-166.55	313.00	rprice

When we control for house and lot characteristics, the estimated causal effect of the incinerator on housing value increases in magnitude compared with the prior DD regression—representing a larger estimated decrease in housing value. The estimated causal effect of being near the incinerator in this regression is -13320.15 dollars. Additionally, the estimated effect in this regression has a lower p-value than that of (d).

```
linearHypothesis(DD2, c("land = 0", "age = 0", "rooms = 0", "area = 0"),
                 white.adjust = "hc2", test = "F") %>%
  tablr()
```

	res.df	df	statistic	p.value
1	317.00			
2	313.00	4.00	34.51	0.00

Based on the above linear hypothesis test, we can reject the null hypothesis that the coefficients on the house and lot characteristics are all jointly equal to 0 with a p-value of less than 0.001.

(g) Calculate by how much did real housing values in the control group change on average between 1978 and 1981.

```
mean_c_before <- mean(df_before[df_before$nearinc == 0,]$rprice)
mean_c_after <- mean(df_after[df_after$nearinc == 0,]$rprice)

mean_change_value <- (mean_c_after - mean_c_before)
```

Real housing values in the control group increased on average by 18790.29 dollars between 1978 and 1981.

(h) 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 in the context of the incinerator construction in North Andover is that in the absence of the incinerator, the mean difference in value between houses near the incinerator and those not near the incinerator would remain constant—i.e. the parallel trends assumption.