# Difference in Differences with a Pooled Panel

**R Tutorials for Applied Statistics**

**Note on required packages:** The following code requires packages in the `tidyverse`. The `tidyverse` contains many packages that allow you to organize, summarize, and plot data. If you have not already done so, download and install the libraries (needed only once per computer), and load the libraries (need to do every time you start R) with the following code:

```
# This only needs to be executed once for your machine
install.packages("tidyverse")

# This needs to be executed every time you load R`
library("tidyverse")
```

---

# 1 Example: Living Near An Incinerator

The code below downloads and loads an R data set that includes housing prices in North Andover, MA in 1978 and 1981, and the distance to the houses to an incinerator that began operating in 1985, but whose construction was announced in 1979.

```
load(url("http://murraylax.org/datasets/kielmc.RData"))
```

It is widely believed that it is undesirable to live near a garbage incinerator. Researchers and city government officials are interested in whether the announcement of the incinerator led to lower housing prices for homes within three miles of the incinerator.

One *tempting, but incorrect* way to determine this is to estimate a regression model that predicts housing price in 1981, and use living near the incinerator as a dummy variable.

The variable `dist` is equal to the number of *feet* the home is from the incinerator. Let us create a dummy variable for whether the house is *within three miles* of the incinerator. That is, the dummy variable will equal 1 when the `dist<=15840` (There are 15,840 feet in

one mile). The following line creates a new dummy variable in the data set, `data`, called `close` to accomplish this:

```
data$close <- as.numeric( data$dist <= 15840)
```

The code `data$dist <= 15840` returns a vector where all the values are TRUE or FALSE, depending on whether the condition less than three miles was true. The call to `as.numeric()` converts these to 1s and 0s.

Let us split up the data set for housing prices in 1978 and housing prices in 1981, so that we can focus on just one particular year if necessary. The following line filters out the rows in the the data set, `data`, where the year is 1978:

```
data1978 <- filter(data, year == 1978)
```

Let us similarly construct a data set for only the year 1981:

```
data1981 <- filter(data, year == 1981)
```

# 2 Doing it wrong: Estimating the impact of the incinerator announcement

It was announced in 1979 that the incinerator would be built, so any affect the incinerator has on house prices will happen after 1979. Let us run the following regression focusing only on data from 1981 to predict the (log) house price based on whether or not the house is within 3 miles of the incinerator:

```
lmprice <- lm( log(price) ~ close, data=data1981)
summary(lmprice)
```

```
##
## Call:
## lm(formula = log(price) ~ close, data = data1981)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86626 -0.21706  0.02718  0.24035  1.16633
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.74242    0.03428 342.543  < 2e-16 ***
## close       -0.40257    0.06459  -6.233 5.06e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3462 on 140 degrees of freedom
## Multiple R-squared:  0.2172, Adjusted R-squared:  0.2116
## F-statistic: 38.85 on 1 and 140 DF,  p-value: 5.061e-09
```

We can see the impact on housing price is estimated to be negative and it is strongly statistically significant. One would be tempted to say we found statistical evidence that the incinerator announcement caused lower house prices for houses within three miles of the incinerator. The magnitude of the coefficient (`-0.402`) suggests that the incinerator caused housing prices within three miles to be 40.2% lower.

What is wrong with this?

It may be that housing prices near the incinerator were low to begin with. Let's see if that is true. Let's estimate the same regression as above, but focus on 1978, before the incinerator was ever announced. Does living within three miles within the location of a future incinerator that no one knows will exist *influence* housing prices?

```
lmprice <- lm( log(price) ~ close, data=data1978)
summary(lmprice)
```

```
##
## Call:
## lm(formula = log(price) ~ close, data = data1978)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11957 -0.18781  0.01555  0.15439  1.66604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.28542    0.02995 376.845  < 2e-16 ***
## close       -0.33992    0.05354  -6.349 1.77e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 177 degrees of freedom
## Multiple R-squared:  0.1855, Adjusted R-squared:  0.1809
## F-statistic: 40.31 on 1 and 177 DF,  p-value: 1.769e-09
```

We see here that the coefficient on `close` is again negative and *strongly* statistically significant. The idea of an incinerator does not even exist yet, so it cannot be the incinerator causing these lower housing prices. Clearly the area where the incinerator is to be built already had, on average, lower housing prices.

# 3 Doing it right: Difference in Difference

A **difference-in-difference (DinD)** estimator takes into account differences that already existed before the treatment, regardless of the causes. First is computes the differences in housing prices explained by the location of the house, both before the treatment (the incinerator announcement) and after the treatment. Then it computes the differences of these differences.

We can estimate the DinD effect with the a regression model that uses the whole data set and includes an interaction term between a dummy for the year and the dummy variable for proximity to the incinerator:

$$\log(price_i) = \beta_0 + \beta_1 yr1981_i + \beta_2 close_i + \beta_3 yr1981_i close_i + \epsilon_i$$
$$\log(pricei) = \beta 0 + \beta 1 yr1981i + \beta 2 closei + \beta 3 yr1981i\ closei + \epsilon i$$

The coefficient on the interaction term, $\beta_3$β3, measures how much *more* the effect of closeness to the incinerator has on price in 1981 versus 1978. The data set includes a dummy variable for 1981 called `y81`. Let's estimate the regression:

```r
lmprice <- lm( log(price) ~ y81 + close + y81:close, data=data)
summary(lmprice)
```

```
##
## Call:
## lm(formula = log(price) ~ y81 + close + y81:close, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11957 -0.20328  0.02226  0.18909  1.66604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.28542    0.03051 369.839  < 2e-16 ***
## y81          0.45700    0.04532  10.084  < 2e-16 ***
## close       -0.33992    0.05456  -6.231 1.48e-09 ***
## y81:close   -0.06265    0.08344  -0.751    0.453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3384 on 317 degrees of freedom
## Multiple R-squared:  0.4091, Adjusted R-squared:  0.4035
## F-statistic: 73.15 on 3 and 317 DF,  p-value: < 2.2e-16
```

We see here that while the coefficient on the interaction term is negative, it is not statistically significant. We failed to find evidence that housing prices changed more from 1978 to 1981 for houses that were within three miles of the incinerator.

The coefficient on `y81` captures on average how different prices were overall in 1981 compared to 1978. The coefficient suggests that housing prices were on average 45.7% greater in 1981 that 1978. The coefficient is strongly statistically significant.

The coefficient on `close` captures on average how different housing prices are near the location of the incinerator, but only for 1978 (when the dummy `y81` is equal to 0). Even before the incinerator was built, housing prices near the location of the future incinerator were 34% less than other housing prices in the city.