

Answers to Problem Set 4 - Program Evaluation PPHA 34600

Diego Diaz

26-05-2020

Other group members: Piyush Tank, Matthew Mauer.

1 - An ideal experiment for this question requires an RCT experiment since we require random assignment of the treatment. Ideally we would have a group of N provinces and we would assign *provincial air quality regulations* randomly among a portion of them. Having the regulations in effect in *province_i* means the province is being *treated* (ideally we set up the regulations the exact same way to treated provinces). We are interested in estimating:

$$\tau_i = PM_i(D_i = 1) - PM_i(D_i = 0)$$

Where PM_i is particulate matter for province i and D_i indicates the treatment status. Although we can't observe τ_i since we can't observe a province in both states, having the regulations assigned randomly means that both groups, provinces with and without regulations, are balanced. Given that both groups are equal in terms of both observable and unobservable characteristics on average, we can estimate τ_{ATE} , the *average treatment effect* of air regulations, by calculating the difference in the outcome; that is, measuring the prevalence of *local particulate matter* (PM) and calculating the difference between groups.

The dataset that the experiment would produce is the treatment binary indicator for provinces ($D_i \in \{0, 1\}$) and the measurement of *local particulate matter* (PM_i) a certain amount of time after the policy is in effect, measured in the same unit for each province as a concentration of particules (*e.g.* : $[gr/m^3]$). Using the potential outcomes framework, we estimate the τ_{ATE} for *province_i* as:

$$\tau^{ATE} = E[PM_i | D_i = 1] - E[PM_i | D_i = 0]$$

Differences between average particulate matter among provinces gives us an unbiased estimate of τ_{ATE} :

$$\hat{\tau}^{ATE} = \overline{PM(D_i = 1)} - \overline{PM(D_i = 0)}$$

2- If we have a single snapshot of municipalities and we have measurements of air quality, we can compare the difference in air quality between municipalities with regulations and municipalities without. However, this comparison will not give us a causal effect of air quality regulations on air quality. Three examples are provided next:

- Municipalities with bad air quality are likely more prone to set up air quality regulations to try to fix the problem. Without knowing the impact of the policy, but having a need to address a problem demanded by the general public, these municipalities will likely be willing to try untested methods once the problem is costly enough. In the data this causes a self selection problem that generates a downward bias in the estimate of the average treatment effect.
- Municipalities that place air quality regulations may be fundamentally different from those that do not. For example, they may have more income per capita on average as highly populated urban areas tend to have higher incomes. The high urban density causes traffic and traffic jams, which increase the prevalence of particulate matter in the air. Given that the units are different, it will be impossible to

identify the *ATE* by comparing both groups. A more sophisticated analysis as a selection on observables would be an alternative to control for differences in income, but the units may also be different in unobservables, which would be impossible to correct with that approach.

- Unintended consequences of regulation may involve people moving out of regulated municipalities to less regulated municipalities. A business that produces large amounts of particulate matter may find it profitable to set up production in a less regulated municipalities when facing higher costs due to regulation. This interaction between treated and control groups will produce upward biased estimates of the treatment effect, since regulated municipalities will have less economic activity while unregulated ones have more.

3- Having multiple observations in different points in time would be beneficial since we could compare each unit with in two different times, which is a better comparison since we can be almost certain that the unit is the same in observables and unobservables. By adding a temporal dimension to the data, and supposing we have two periods and that municipality i puts air quality regulations in a given year after 2004 (let's call it $t = 1$), we can estimate τ_i as:

$$\tau_i = PM_i(D_{i,t=1} = 1) - PM_i(D_{i,t=0} = 0)$$

Since now we can observe a municipality in both states, we can get each τ_i for every municipality for which we have observations in both states. Knowing that every municipality installed regulations in 2004, $D_i = 1$, that is, being treated or having regulations on, only happens when $t = 1$. Therefore, the average treatment effect of air quality regulations becomes:

$$\tau_i^{ATE} = E[PM_i(t = 1) - PM_i(t = 0)]$$

Calling our time series estimator $\hat{\tau}^{TS}$, we can estimate it as:

$$\hat{\tau}_i^{TS} = PM_{it=1} - PM_{it=0}$$

And if we have data on multiple points in time from 2001 to 2019 we can estimate:

$$\hat{\tau}_i^{TS} = \bar{PM}_{it \in \text{post-2004}} - \bar{PM}_{it \in \text{pre-2004}}$$

The assumption we are making when implementing this approach is that we have *time invariant* characteristics, both in observables and unobservables. Another way to think about this assumption is that *the counterfactual trend must be zero**, which is untestable.

Three possible concerns are provided next:

- In 2004 or after ($t > 0$) Technologies are developed that decrease the concentration of particulate matter (PM2.5) in the air, increasing air quality for treated municipalities. This would result in an upward bias in the estimator since municipalities would get cleaner air for reasons unrelated to regulation.
- Volcanic eruptions in some municipalities during $t > 0$ would make the air quality decrease in the affected locations, causing a downward bias in $\hat{\tau}_i^{TS}$. As in the previous example, the differences observed would be contaminated by the eruptions, causing a potential total decrease of air quality even after regulations are in effect by breaking the *zero counterfactual trend* assumption.
- Another way to break the required assumption would happen if regular economic activity causes emissions of particulate matter to increase. This would also prevent our *zero counterfactual trend* to be true and would bias the estimate of $\hat{\tau}_i^{TS}$ as some of the difference of air quality would be for reasons unrelated to the policy.

4- Having data of groups with and without regulations would be beneficial to estimate the causal effect of air quality regulations on air quality because the *zero counterfactual trend* could be relaxed by a *difference-in-difference* approach. Comparing different municipalities can be a problem for estimating average treatment

effects because they may be systematically different, while comparing the same municipalities across time can also be a problem if we suffer from time-varying unobservables (non-zero trends). Both problems cause bias in our estimate of average treatment effects, but combining observations of both group of municipalities across time allows us to perform comparisons in both dimensions. Our concern with the *time-series* estimand was that the regulated municipalities might have changed for unobservables unrelated to the treatment. By estimating how the regulated municipalities would have behaved without treatment we can isolate the treatment effect, and to do this we can make a guess using the unregulated municipalities. The unregulated municipalities may have changed their air quality as well, and subtracting this change from the ones would have been the same in absence of treatment, therefore, it is called the *parallel trends assumption*. It is less restrictive than the *zero counterfactual trend* assumption because it can be tested in practice by looking at the trends before treatment. We call the estimator the difference-in-differences estimator (*DD*) and we can represent it as:

$$\hat{\tau}^{DD} = \hat{\tau}_{D_i=1}^{TS} - \hat{\tau}_{D_i=0}^{TS}$$

In words, the *DD* estimator is the difference between the *TS* estimator between regulated and unregulated municipalities, which compares *treated* and *untreated* units over time.

Assuming municipality i puts regulations, the following equation would allow us to obtain the estimator from a linear regression:

$$PM_{it} = \alpha + \tau \text{ regulations}_i \times \text{post-2004}_t + \beta \text{ regulations}_i + \delta \text{ post-2004}_t + \varepsilon_{it}$$

Next we provide two examples of possible concerns remaining:

- The parallel trends assumption can break if there is interaction between groups that affect the air quality. Higher regulations can increase costs (and prices) of goods produced in regulated municipalities after 2004. In turn this will lower demand for those goods, decreasing economic activity. However, this is a problem if the economic activity is taken by the unregulated municipalities. If they start producing the goods that saw a price increase and have increased economic activity as a result, our parallel trends assumption breaks, as the municipality would not be a suitable counterfactual for the regulated municipality.
- Fundamental differences in unobservables between groups would also likely break the parallel trends assumption. If regulated municipalities have higher starting concentrations of particulate matter, for example as a result of higher urbanization level, then it is likely that the trends of air quality would be different if the municipalities were never regulated.

5- PROGRAMEVAL, given your even-handed discussion of various approaches, is willing to put their faith in you. They will give you data on the universe of their consumers from 2003 to 2007. This includes municipalities that imposed air quality regulations across several different years. Describe, in words and math, how you would estimate the effect of air quality regulations on particulate matter using this dataset. You should include an estimating equation in the form of a regression.

We can modify the previous model to a more general fixed effects model to allow for different municipalities to be treated at different times periods.

$$PM_{it} = \alpha_i + \tau D_{it} + \delta_t + \varepsilon_{it}$$

Where D_{it} is a dummy that indicates whether the municipality i has regulations in place at time t .

6- Use the included `ps4_data.csv` dataset to implement a simple comparison of average particulate matter between municipalities with and without air quality regulations. Describe what you find. Use regression to perform a time-series analysis of the effect of air quality regulations on particulate matter, using only municipalities who introduced regulations in 2004. Describe what you find. How does this differ from what you estimated using the initial estimator. Plot particulate matter against time for municipalities that imposed

air quality restrictions in 2004. What do you see? (It may also be helpful to plot average consumption across municipalities). Does this figure affect how you interpret your estimates?

We first compare average particulate matter between municipalities with and without air quality regulations. We are forgetting at many issues when considering this estimator, including selection and time of treatment. From the potential outcomes framework, this is the naive estimator:

$$\tau_{NAIVE} = \overline{PM(D_i = 1)} - \overline{PM(D_i = 0)}$$

We estimate it in R with the following code:

```
mean_treated <- mean(ps_4_data[!is.na(ps_4_data$air_quality_regulation_year),]$particulate_matter)
mean_untreated <- mean(ps_4_data[is.na(ps_4_data$air_quality_regulation_year),]$particulate_matter)
naive <- mean_treated - mean_untreated
print(naive)
```

```
## [1] -24.44856
```

As we can see, $\tau_{NAIVE} = -24.45$, indicating that the air quality regulations in affected municipalities decreased the amount of particulate matter by 24.45. Considering the issues of selection and possible different time trends between units, the naive estimator is not believable.

Next we perform a regression using municipalities that introduced restrictions in 2004. Following the approach given in (3), since we have data on multiple points post-regulation, we average air quality after 2004 and until 2004. Our time series estimator becomes:

$$\hat{\tau}^{TS} = \bar{PM}_{t \in \text{post}-2004} - \bar{PM}_{t \in \text{pre}-2004}$$

```
units_2004 <- ps_4_data[ps_4_data$air_quality_regulation_year == 2004,]
units_2004 <- units_2004[!is.na(units_2004$air_quality_regulation_year),]
mean_post <- mean(units_2004[units_2004$year > 2003,]$particulate_matter)
mean_pre <- mean(units_2004[units_2004$year < 2004,]$particulate_matter)
ts_estimator <- mean_post - mean_pre
print(ts_estimator)
```

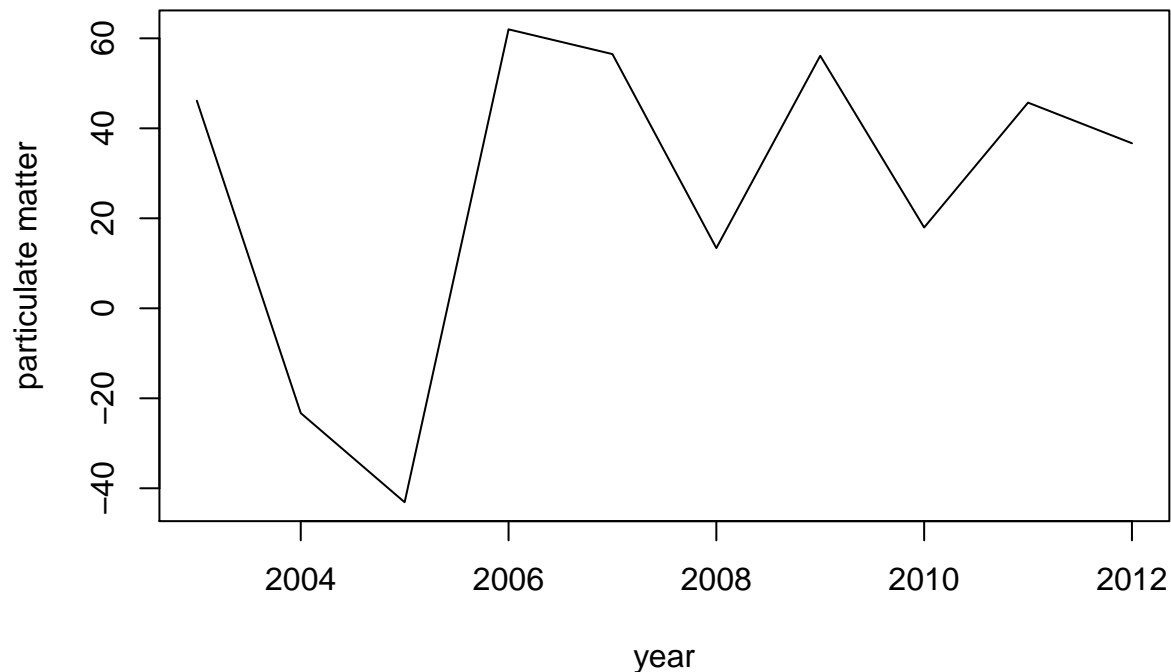
```
## [1] -21.46998
```

As we can see, $\hat{\tau}_{TS} = -21.47$, indicating that the air quality regulations in affected municipalities on average through the years decreased the amount of particulate matter by 21.47.

Next we plot average particulate matter among municipalities through the years:

```
units_2004 <- ps_4_data[ps_4_data$air_quality_regulation_year == 2004,]
units_2004 <- units_2004[!is.na(units_2004$air_quality_regulation_year),]
avg_year = aggregate(units_2004[, "particulate_matter"], list(units_2004$year), mean)
plot(avg_year$x, x = avg_year$Group.1, type="l", xlab = 'year', ylab = 'particulate matter', main = 'mu
```

municipalities that placed regulations on 2004



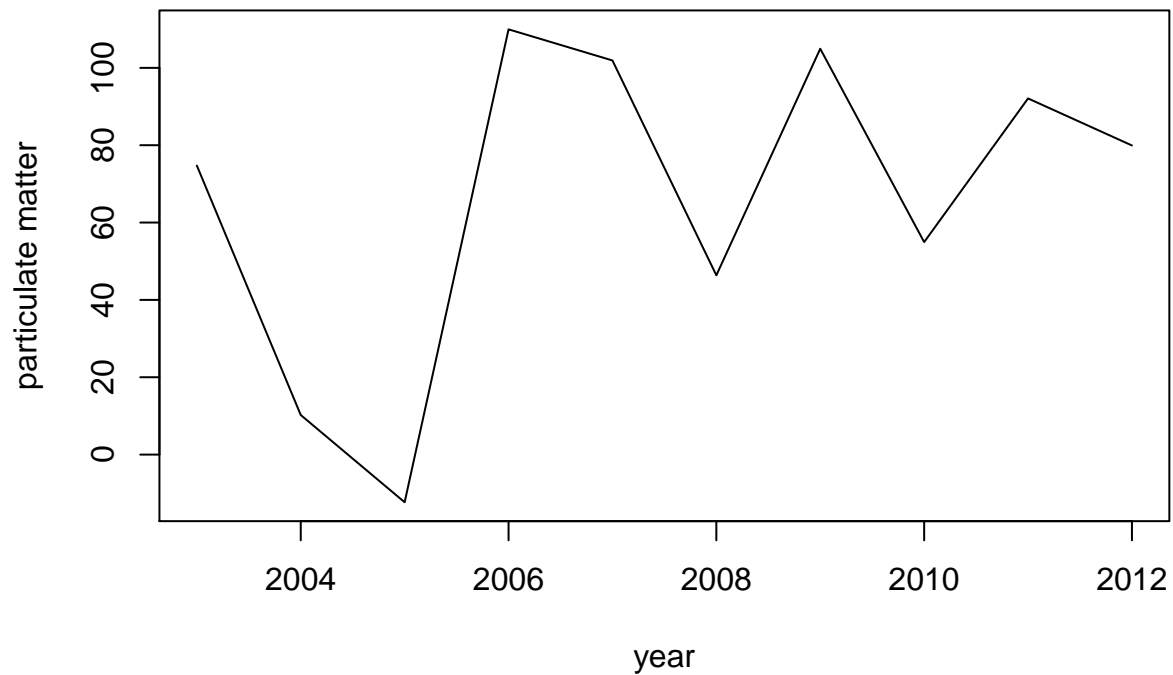
We can say that the amount of particulate matter decreased on average for municipalities that imposed regulations on 2004 in the first years after the policy was implemented, however, it went back to previous levels after that.

7- Plot (average) particulate matter against time for municipalities who never imposed air quality regulations. Assess the viability of using these municipalities as a control group for the 2004 regulators. Plot (average) particulate matter against time for municipalities who passed air quality regulation in 2006. Assess the viability of using the non-regulating municipalities as a control group for the 2006 regulators.

We plot particulate matter (municipalities average) against time for those who never imposed regulations:

```
untreated <- ps_4_data[is.na(ps_4_data$air_quality_regulation_year),]  
avg_year = aggregate(untreated[, "particulate_matter"], list(untreated$year), mean)  
plot(avg_year$x, x = avg_year$Group.1, type="l", xlab = 'year', ylab = 'particulate matter', main = 'mu
```

municipalities that did not place air quality regulations

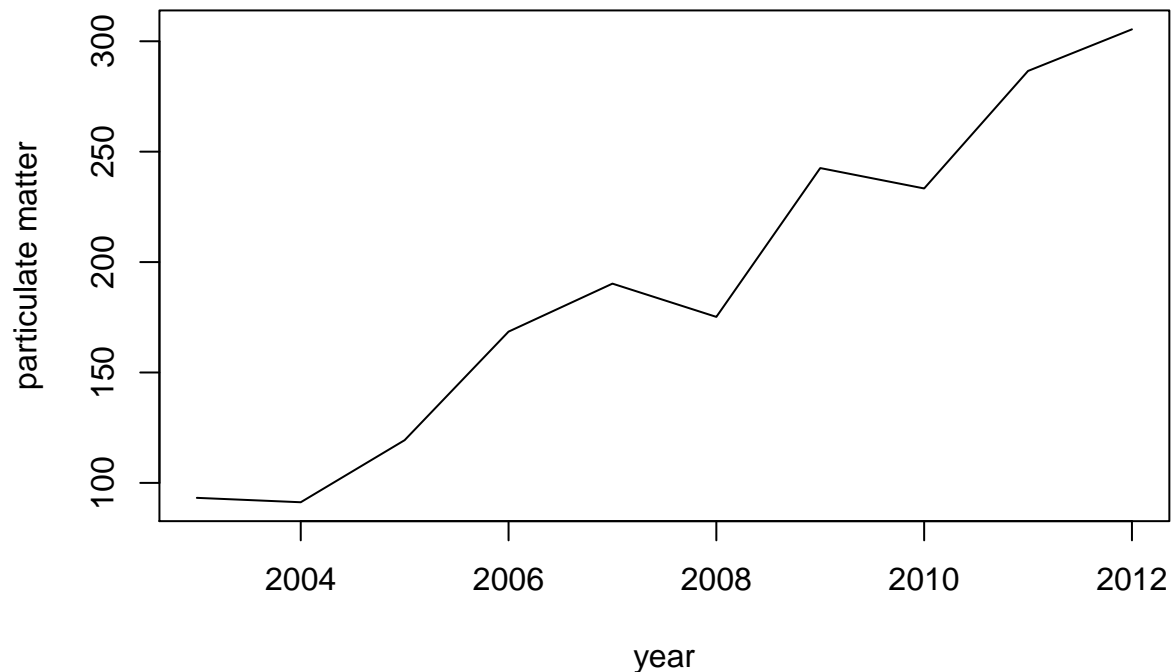


Given our **parallel trends assumption** explained in (4), we can say that the municipalities that imposed air quality regulations on 2004 on average are a viable control group for a difference-in-difference approach, because they have similar trends in particulate matter before the treatment starts.

Next we plot average particulate matter for municipalities that placed regulations on 2006:

```
units_2006 <- ps_4_data[ps_4_data$air_quality_regulation_year == 2006,]  
units_2006 <- units_2006[!is.na(units_2006$air_quality_regulation_year),]  
avg_year = aggregate(units_2006[, "particulate_matter"], list(units_2006$year), mean)  
plot(avg_year$x, x = avg_year$Group.1, type="l", xlab = 'year', ylab = 'particulate matter', main = 'municipalities that placed air quality regulations on 2006')
```

municipalities that placed regulations on 2006



Given the different trends for these municipalities that started regulating on air quality on 2006 (parallel trends assumption does not stand), they are not a suitable control.

8 - We first obtain the naive estimator comparing 2006 regulators to non-regulators. We obtain:

```
treated <- ps_4_data[!is.na(ps_4_data$air_quality_regulation_year),]  
treated_2006 <- treated[treated$air_quality_regulation_year == 2006,]  
mean_treated_2006 <- mean(treated_2006$particulate_matter)  
mean_untreated <- mean(ps_4_data[is.na(ps_4_data$air_quality_regulation_year),]$particulate_matter)  
naive <- mean_treated_2006 - mean_untreated  
print(naive)
```

```
## [1] 124.2771
```

Next we perform a simple regression of implementing regulations on particulate matter. Performing this on R we obtain:

```
treated <- ps_4_data[!is.na(ps_4_data$air_quality_regulation_year),]  
untreated <- ps_4_data[is.na(ps_4_data$air_quality_regulation_year),]  
treated_2006 <- treated[treated$air_quality_regulation_year == 2006,]  
  
df <- rbind(untreated, treated_2006)  
df$treat = ifelse(is.na(df$air_quality_regulation_year), 0, 1)  
  
ols_2006 <- lm(particulate_matter ~ treat, data = df)  
summary(ols_2006)
```

```
##
```

```
## Call:
```

```
## lm(formula = particulate_matter ~ treat, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.458  -24.523    9.488   33.522  130.845
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   66.2814     0.4206  157.60  <2e-16 ***
## treat        124.2771     1.4563   85.34  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.09 on 11988 degrees of freedom
## Multiple R-squared:  0.3779, Adjusted R-squared:  0.3779
## F-statistic: 7282 on 1 and 11988 DF, p-value: < 2.2e-16
```

The coefficient for the treatment (placing regulations) is 124.27 and is statistically significant at 1%. This would mean that implementing regulations, on average, causes particulate matter to increase by 124.27 units. However, the issues with this approach mentioned in previous exercises remain and is therefore this conclusion is not believable.

We now move to a fixed effects approach, controlling for time shocks and time-invariant municipality characteristics. Proceeding in R we obtain:

```
treated <- ps_4_data[!is.na(ps_4_data$air_quality_regulation_year),]
untreated <- ps_4_data[is.na(ps_4_data$air_quality_regulation_year),]
treated_2006 <- treated[treated$air_quality_regulation_year == 2006,]

df <- rbind(untreated, treated_2006)

df$treat = ifelse(is.na(df$air_quality_regulation_year), 0, 1)
df$post = ifelse(df$year > 2005, 1, 0)

fe_2006 <- plm(particulate_matter ~ treat:post + factor(year) +
               factor(municipality_id), model='within', effect='twoways',
               data=df)
#summary(fe_2006)
coeftest(fe_2006, vcov = vcovHC(fe_2006, type = "HC1", cluster = "group"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## treat:post    67.469     33.818   1.995  0.04606 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen on the results, our estimate for τ from our fixed effects regression is 67.47. Indicating a positive and statistically significant effect of air quality regulations. However, as was shown in the previous exercise, the parallel trends assumption does not hold. Therefore, the coefficient is biased.

We have obtained three positive estimates of the average treatment effect by using the naive estimator, a simple regression, and a fixed effects regression. Comparing our results with (6) and (7), in (6) we got a negative value for the time series estimator of -21.47, which is more in line with which we would expect since it is reasonable to assume that regulations for air quality would have, if any, a negative impact in particulate matter. The difference in the analysis here and on (6) and (7) is that here we consider municipalities that

started regulating in 2006, while before we considered those that started in 2004. As the municipalities that started in 2006, as we showed in (7), kept increasing their levels of particulate matter through the years in a trend different from the municipalities that never regulated and different from those that regulated in 2004, it is not surprising that our econometric approaches find that those municipalities that started regulating on 2006 experienced an increase of particulate matter due to the regulation. It is likely that unobservable time-variant characteristics are accounting for these differences.

9- We plot particulate matter over time for municipalities that imposed air quality regulations from 2003 to 2007:

```
library(ggplot2)
```

```
treated <- ps_4_data[!is.na(ps_4_data$air_quality_regulation_year),]
treated <- treated[(treated$air_quality_regulation_year) >= 2003,]
treated <- treated %>% group_by(year ,air_quality_regulation_year) %>%
  summarise_at(vars(-municipality_id), funs(mean(., na.rm=TRUE)))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
```

```
## Please use a list of either functions or lambdas:
```

```
##
```

```
## # Simple named list:
```

```
## list(mean = mean, median = median)
```

```
##
```

```
## # Auto named with `tibble::lst()`:
```

```
## tibble::lst(mean, median)
```

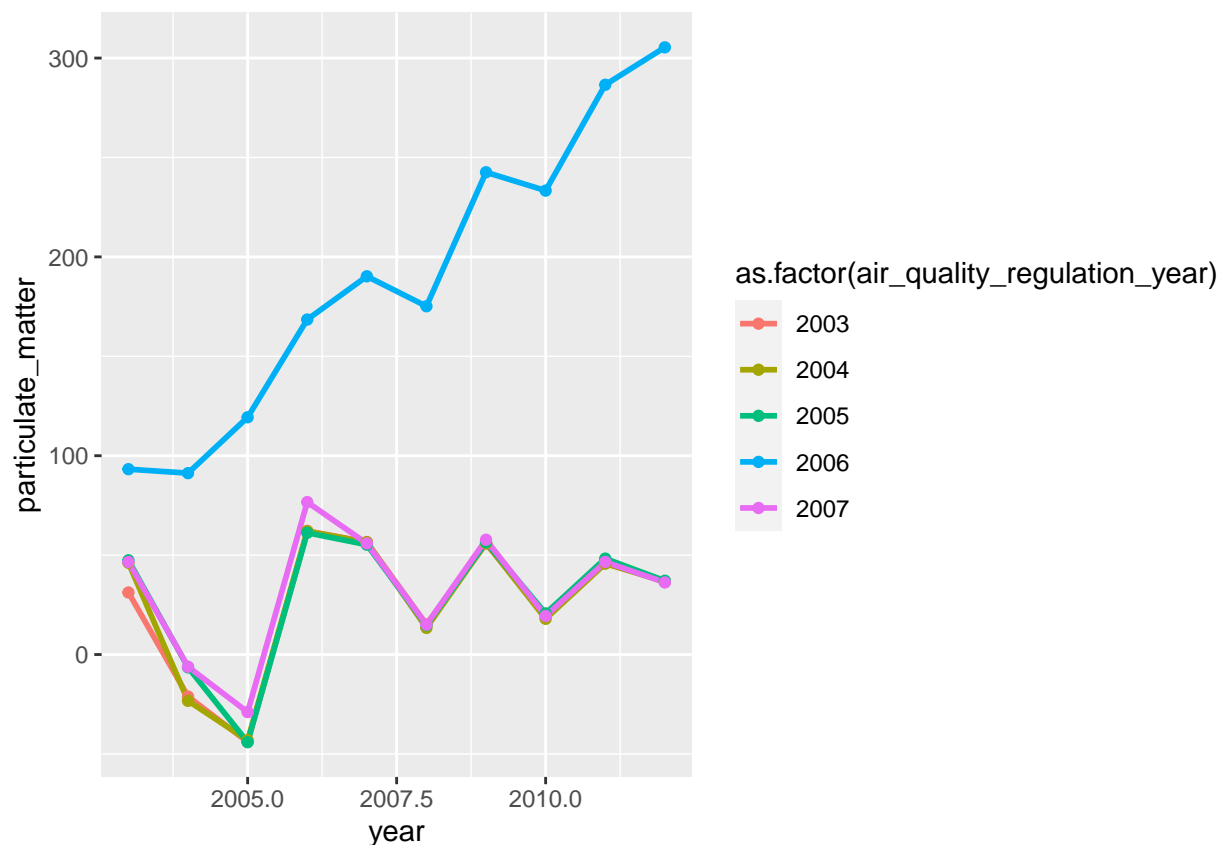
```
##
```

```
## # Using lambdas
```

```
## list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
## This warning is displayed once per session.
```

```
ggplot(treated, aes(x=year, y=particulate_matter, group=(air_quality_regulation_year), color= as.factor
  geom_line(size=1) +
  geom_point(size=1.5) +
  scale_fill_brewer()
```



We immediately noticed that the municipalities that started on 2006 have a different trend than the rest both pre and post treatment. Given our parallel trends assumption would be broken when implementing a **diff-in-diff/fixed effects** approach, we must drop these municipalities from our data to obtain credible causal effects.

With the remaining municipalities (those that implemented regulations in 2003, 2004, 2005, 2007, and those that never did) we estimate a fixed effect model. We show our results next:

```
untreated <- ps_4_data[is.na(ps_4_data$air_quality_regulation_year),]
treated <- ps_4_data[!is.na(ps_4_data$air_quality_regulation_year),]
treated_post_2002 <- treated[(treated$air_quality_regulation_year) >= 2003,]

parallel_trends_df <- treated_post_2002[(treated_post_2002$air_quality_regulation_year) != 2006,]

parallel_trends_df <- rbind(untreated, parallel_trends_df)
parallel_trends_df$treat = ifelse(is.na(parallel_trends_df$air_quality_regulation_year), 0, 1)

parallel_trends_df$post = ifelse(parallel_trends_df$year >= parallel_trends_df$air_quality_regulation_year, 1, 0)
parallel_trends_df$post = ifelse(is.na(parallel_trends_df$post), 0, parallel_trends_df$post)

fe_q9 <- plm(particulate_matter ~ treat:post, index=c('municipality_id','year'), model='within', effect='municipality')
coeftest(fe_q9, vcov = vcovHC(fe_q9, type = "HC1", cluster = "group"))

##
## t test of coefficients:
##
```

```
##           Estimate Std. Error t value Pr(>|t|)
## treat:post -16.91016    0.37572 -45.007 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We obtain an statistically significant estimate for τ of -16.91 units, indicating that the effect of regulating for air quality reduced particulate matter on average for those municipalities that regulated. This is the opposite effect to that found on (8), but given that our parallel trends assumption holds, our new estimate should be the correct one.

Next we perform an event study in R, adding a period variable to compare every regulated municipality from the same point in time. Leaving out the T-1 dummy, we obtain the following values for the estimates:

```
parallel_trends_df <- parallel_trends_df %>%
  mutate(period = year - air_quality_regulation_year)

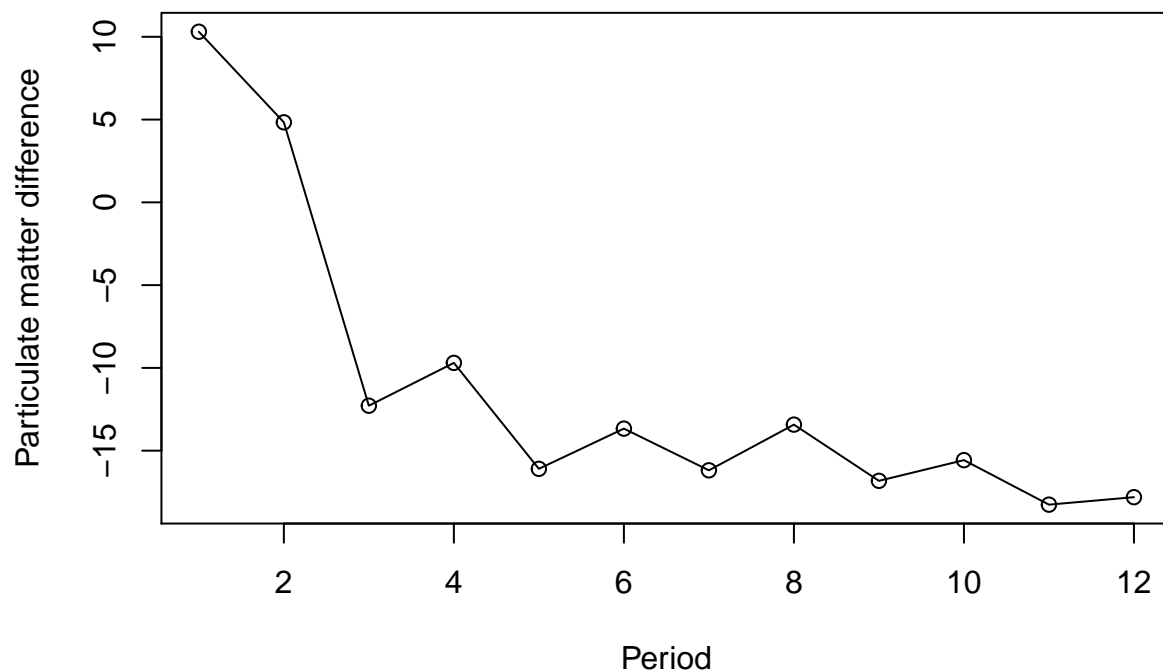
event_study <- plm(particulate_matter ~ factor(period, exclude='-1'):treat + treat,
  index = c('municipality_id', 'year'),
  model = 'within',
  effect = 'twoways',
  data = parallel_trends_df)

coeftest(event_study, vcov = vcovHC(event_study, type = "HC1", cluster = "group"))
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## factor(period, exclude = "-1")-3:treat  10.30520    1.15653   8.9105 < 2.2e-16
## factor(period, exclude = "-1")-2:treat   4.83442    1.00975   4.7878 1.705e-06
## factor(period, exclude = "-1")0:treat  -12.28301    0.97337 -12.6190 < 2.2e-16
## factor(period, exclude = "-1")1:treat   -9.69653    0.99866  -9.7096 < 2.2e-16
## factor(period, exclude = "-1")2:treat  -16.09674    0.97002 -16.5943 < 2.2e-16
## factor(period, exclude = "-1")3:treat  -13.66791    1.00325 -13.6236 < 2.2e-16
## factor(period, exclude = "-1")4:treat  -16.18616    1.01941 -15.8780 < 2.2e-16
## factor(period, exclude = "-1")5:treat  -13.42574    0.98586 -13.6183 < 2.2e-16
## factor(period, exclude = "-1")6:treat  -16.82183    1.06892 -15.7372 < 2.2e-16
## factor(period, exclude = "-1")7:treat  -15.57126    1.02676 -15.1654 < 2.2e-16
## factor(period, exclude = "-1")8:treat  -18.25909    1.08667 -16.8028 < 2.2e-16
## factor(period, exclude = "-1")9:treat  -17.80829    1.30157 -13.6822 < 2.2e-16
##
## factor(period, exclude = "-1")-3:treat ***
## factor(period, exclude = "-1")-2:treat ***
## factor(period, exclude = "-1")0:treat ***
## factor(period, exclude = "-1")1:treat ***
## factor(period, exclude = "-1")2:treat ***
## factor(period, exclude = "-1")3:treat ***
## factor(period, exclude = "-1")4:treat ***
## factor(period, exclude = "-1")5:treat ***
## factor(period, exclude = "-1")6:treat ***
## factor(period, exclude = "-1")7:treat ***
## factor(period, exclude = "-1")8:treat ***
## factor(period, exclude = "-1")9:treat ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All coefficients for different periods are statistically significant. Putting them in a plot we obtain:

```
plot(event_study$coefficients, lwd = 1, type='o', xlab = 'Period', ylab = 'Particulate matter difference')
```



The effect of regulations decreases particulate matter drastically after implementation and then slowly trends down.

10- To assess the impact of regulating air quality in different municipalities, several approaches were taken. We compared the means of treated municipalities to untreated ones ($\tau_{N AIVE} = -24.45$), we compared those that were treated before and after they started regulating, obtaining a time series estimator ($\tau = -21.47$). We then performed a simple regression only for municipalities regulated in 2006 and those without, without considering time differences, obtaining a coefficient $\tau_{OLS} = 124.3$. Then we performed a fixed effect regression on the same data, obtaining $\tau_{FE} = 67.47$. Finally we performed a fixed effects regression with the groups of municipalities that have parallel trends with the untreated ones, fulfilling the main requirement for our fixed effects estimator, obtaining a coefficient $\tau_{FE} = -16.9$. Finally we performed an event study, finding that every year the municipalities that regulated air quality experienced a decrease in particulate matter because of the policies. The last fixed effects estimation and event study are the only one that potentially can be unbiased because the parallel trends assumption is satisfied, and therefore PROGRAMEVAL should be promoting air quality regulations. A potential shortcoming is that we haven't looked at the distributions of the trends of particulate matter between municipalities. We want the trends to be similar between treated and untreated, but so far we only compared the averages. The trends between units could potentially be entirely different and that would render our parallel trends assumption invalid.