PS4

5/19/2022

(1) PROGRAMEVAL are interested in answering the following question: What is the impact of provincial air quality regulations on local particulate matter (PM 2.5)? In order to get started, they'd like you to present your ideal experiment. Explain what you would do to answer this question in a completely unconstrained world, and describe the dataset that you'd like to have to perform the analysis. Use math, words, and the potential outcomes framework to explain what you would estimate and how you would do so. Make sure to be clear about the unit of analysis (ie, what is "i" here?)

In an unconstrained world we would be able to measure the causal impact of provincial air quality regulation on local PM 2.5 by running a randomized experiment, RCT, in which some municipalities are randomly assigned to treatment, that is to say, some municipalities would have air quality regulations emitted and other municipalities will not. The municipalities that are randomized into not having air quality regulations emitted would act as the control group to compare the group that gets treated (ie. regulated).

In the ideal experiment, we would have a dataset of Chinese municipalities with baseline observable characteristics that might be correlated with pollution (for example number and types of factories, quantity of trees, economic growth, etc.). Then, I could randomize municipalities into receiving air quality regulation and to check that randomization was effective, I would construct a balance table by conducting t-tests so as to ensure treatment and control municipalities are balanced in all baseline characteristics, and subsequently run the experiment.

Using the **potential outcomes framework**, we could estimate the causal effect of provincial air quality regulation on local PM 2.5 of the **average municipality** by estimating the **Average Treatment Effect (ATE):**

$$\tau^{ATE} = E[Y_i(1)] - E[Y_i(0)]$$

Where:

- Y is the outcome variable of interest, ie. local PM 2.5
- i is each individual municipality

Under randomization, the **expected** potential PM 2.5 for regulated municipalities (ie. $PM2.5_{municipality}(1)$) is equivalent to the **mean** PM 2.5 of municipalities whose air quality is regulated, and vice versa for municipalities in the control group.

Therefore, we can estimate the ATE by taking the difference in means between both groups:

$$\tau^{ATE} = \overline{Y_i}(D = 1) - \overline{Y_i}(D = 0)$$

Where:

- D refers to treatment status: $D = 1$ for municipalities whose air quality is regulated and $D = 0$ for unregulated municipalities.
- Y is local particulate matter PM 2.5
- i is each individual municipality

Through an effective process of randomization, Chinese municipalities assigned to being regulated would be statistically similar to municiplaities assigned to the control group, such that the difference we observe between both can be attributed to the air quality regulation and not to other external factors. This way, we address the **selection bias problem** and can effectively recover the ATE.

Further, in the ideal experiment there would be **perfect compliance**. That is, municipalities assigned to air quality regulation would actually get regulated and municipalities that are assigned to the control group will not get regulated.

(2) PROGRAMEVAL, being a commission and not a Chinese regulator, can't impose pollution regulations themselves. But they do have some data that they'd be willing to let you work with. They have a single temporal snapshot of air quality across many municipalities. They'd like you to look at average differences in air quality between municipalities with and without air quality regulations to get a sense of what these regulations do to air quality. You have a sense that this is not a great idea. Describe three concrete examples of why this comparison might not provide the answer that PROGRAMEVAL want.

Their single temporal snapshot of air quality across many municipalities is cross sectional data, and in the absence of randomizing air quality regulation, if we simply take the average difference in air quality between municipalities with and without air quality regulation, then we would be calculating the **naive estimator:**

$$\tau^N = \overline{Y}(1) - \overline{Y}(0)$$

Where:

- $N = municipalities$ since the unit of analysis regulated and unregulated municipalities
- Y is local particulate matter PM 2.5

This naive estimator is simply based on **observed outcomes** and it can be very problematic since the treated and untreated municipalities are very likely to be **systematically different in many characteristics**, both observable and unobservable, in ways that are correlated with municipalities selection into treatment (emitting air quality regulations or not).

If we use this estimator, we would have **selection bias** and would erroneously attribute the impact on local particulate matter PM 2.5 to air quality regulation, when it can well be the result of many other factors, without the regulation necessarily having an effect at all.

To give three concrete examples:

1. Some municipalities might have a concentration of pollution-emitting factories while others might not. If, for example, it so happens that municipalities without factories are the ones that have air quality regulations in place, due to environmentally-conscious local regulators and citizens that hold them accountable, then we might incorrectly conclude that the low levels of pollution are a causal effect of the regulation when in fact these municipalities were pollution-free to begin with.

2. It's possible that owners of the biggest and most polluting factories have strong lobbying power and can influence local regulators to act in their favor by not emitting air quality regulations. Then, we might systematically observe that those municipalities without regulations are those that are most polluted. We could erroneously assume that air quality regulations have no effect but this might be due because these regulations are not emitted in localities with actual pollution.

3. There could be an enforcement problem. For example, it could be the case that polluted municipalities have air quality regulations in place but due to lack of resources, the local government is unable to enforce these regulations. Then, polluting factories would continue polluting without facing any fines or restrictions. As such, we could conclude that air quality regulations are ineffective when they might actually work if these regulations were enforced.

(3) Explain to the PROGRAMEVAL what the benefit would be of being able to observe municipalities at multiple points in time. Their database goes from 2001 to 2019, but they only want to share what's absolutely necessary, for confidentiality reasons. First, describe, in words and math, what you would do with data on many municipalities, all of which imposed air quality regulations in 2004. Be sure to discuss the identifying assumption that would be required for this approach to recover the causal effect of air quality regulations on particulate matter. Provide three concrete examples of concerns with this approach.

Observing municipalities at different points in time, aka time-series data, can be a better way to address the selection bias problem. When we have cross-sectional data, observations of municipalities in a *single point in time*, we are limited to comparing *different municipalities*, (ie. municipality i to municipality j) as a way of addressing the **fundamental problem of causal inference** and have municipality j act as a counterfactual of municipality i. The problem with this, as explained in part (2) is that municipalities can be systematically different in many aspects, both observable and unobservable.

Instead, with **time-series data**, we can address selection problems by comparing each municipality to itself in different points in time (ie. municipality i itself in time t vs. municipality i in time t-1). Essentially, the municipality acts like a control for itself, before being regulated (aka before treatment). The notion is that any given municipality is inherently more similar to itself in different points in time than two different municipalities are.

Given time-series data from 2001 to 2019 for municipalities that imposed air quality regulations in 2004, we would like to estimate the **Average Treatment Effect (ATE)**:

$$\tau^{ATE} = E[Y_t(D_t = 1) - Y_t(D_t = 0)]$$

Where:

- Y is the local PM 2.5 air quality regulations for each municipality with regulations in place $Y_t(D_t = 1)$ and without regulations $Y_t(D_t = 1)$ in 2004
- $t = 2004$ since that is the year air quality regulations were imposed

Due to the **fundamental problem of causal inference** it is **impossible** to observe a given municipality at the same time (ie. 2004) with air quality regulation in place and without it *at the same time.*

Instead, what we can do, is compare those same municipalities **before and after** their air quality was eventually regulated. Here, we use the **time-series estimator**:

$$\hat{\tau}^{TS} = \overline{Y}_{t \in POST} - \overline{Y}_{t \in PRE}$$

Where:

- $\overline{Y}$ is the average local PM 2.5, such that
  $overline Y_{t \in POST}$ during the POST air quality regulation years (ie. 2005-2019)and $\overline{Y}_{t \in PRE}$ during the PRE regulation years (ie. 2001 to 2003)

IDENTIFYING ASSUMPTION

Supposing we have the following regression to estimate ATE with the time-series estimator:

$$Y_{it} = \tau D_{it} + \beta X_i + \gamma U_i + \delta V_i$$

Where:

- Y_{it} is the local PM 2.5 that varies by municipality over time
- X_i are municipality observable characteristics that DO NOT vary over time (ie. geographic location of a municipality)
- U_i are municipality unobservable characteristics that DO NOT vary over time

- V_i are municipality observable and unobservable characteristics that DO change over time (ie. number of factories in a municipality, quantity of pollution)

With this model that includes *time-varying characteristics* we would recover the following:

$$\hat{\tau}^{TS} = Y_{i,t=1} - Y_{i,t=0}$$
$$\hat{\tau}^{TS} = \tau + \delta(V_{i,t=1} - V_{i,t=0})$$

Where:

- Y_{it} is the local PM 2.5 that varies by municipality over time, before and after air quality regulations in 2004
- V_i are time-varying characteristics

When we have time-varying characteristics, observable and unobservable, the **problem** is we are not estimating only τ, but rather $\tau + \delta(V_{i,t=1} - V_{i,t=0})$ which is a **biased estimate**.

$\rightarrow$ As such, for $\hat{\tau}^{TS} = \tau$, the identifying assumption is that there are NO time-varying characteristics, observable or unobservable. Mathematically:

$$\delta = 0 \text{ OR } V_{i,t=1} = V_{i,t=0} = V_i$$

This means that the local PM 2.5 Y_i would be **unchanged in the absence of treatment**. That is, we assume that the *counterfactual trend = 0*.

EXAMPLES OF CONCERNS

The time-series estimator identifying assumption is untestable and there is reason to believe it is unlikely to hold. That is, we can imagine that we usually are faced with time varying characteristics. For example:

1. There could be an economic boom in some Chinese municipalities that bring in extra investment opportunities for local factories, allowing them to increase their production and therefore pollution as well. Then, the effect of air quality regulations may be estimated to be ineffective because coincidentally at the same time as municipalities had these regulations imposed they also had an economic boom that led to an increase in pollution. This would affect our estimated ATE and lead us to reach a wrong conclusion.

2. Due to the pandemic, there has been a global shortage of certain commodities. If air quality regulations were placed at the same time as this shortage, then we could observe a fall in factory production in China due to limited production inputs. The accompanying reduction in pollution would then be a consequence of the global shortage and not of the air quality regulations.

3. In China, there have been very strong lockdown measures during the pandemic as part of zero-tolerance for Covid-19 policy. Having citizens in lockdown forces a decrease in production without workers to operate factories. Globally, there was a clear reduction in pollution during COVID-19 lockdowns. If the air quality regulations happened to be imposed at the same time as the lockdowns, then we could incorrectly conclude that the regulations were the reason behind the drastic drop in pollution.

(4) Next, explain why it would be even better to have data on multiple municipalities, divided into two groups: municipalities that never imposed air quality regulations, and municipalities that imposed air quality regulations in 2004. Explain, in words and math, the estimator that you would use with this dataset. You should include an estimating equation in the form of a regression. Describe how this larger dataset would allow you to resolve the concerns you had above. Be sure to discuss the identifying assumption that would be required for this approach to recover the causal effect of air quality regulations on particulate matter. Provide two examples of remaining concerns, even in this larger dataset.

Unlike having data only on municipalities that eventually get air quality regulations as in part (3), having data on municipalities that never impose air quality regulations AND municipalities that do, can help us

address the short-comings previously addressed.

Essentially, by comparing treated and untreated units over time, we can address the selection bias problems we face if we compare different municipalities at the same time (here we have to deal with municipalities beins systematically different from one another, in characteristics different from the air-quality regulations) and, the problems we face when comparing the same municipalities over-time (here we have to deal with a strong assumption that there are no time-invariant characteristics, which is unlikely to hold, particularly with an outcome such as pollution which can be affected by macroeconomic shocks).

Moving from time-series data, we would now have **panel data** of *multiple municipalities each observed over time*. With this dataset, we can now use the **Difference-in-differences estimator (DD)** which is a combination of the naive estimator in part (2) and time-series estimator in part (3).

The DD estimator, essentially takes two differences:

1. Across municipalities, within 2004, comparisons (ie. naive estimator)
2. Within municipalities, across time 2001-2019, comparisons (ie. time-series estimator)

Mathematically, we can express the DD estimator as:

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

Where:

- $\hat{\tau}^{TS}_{D_i=1}$ are regulated municipalities over time (2001-2019)
- $\hat{\tau}^{TS}_{D_i=0}$ are unregulated municipalities over time (2001-2019)

In particular, to estimate ATE with a DD approach, we could run the following regression:

$$PM2.5_{it} = \alpha + \tau ACRegulation \times Post_i + \beta ACRegulation_i + \delta Post_t + \gamma X_i + \varepsilon_i t$$

Where:

- i is each individual municipality
- t is each year, from 2001-2019
- X_i are covariates that can be added to soak up variation and increase precision (ex: number and size of factories)
- $\varepsilon_i t$ is the error time of anything not captured by the model that can affect local PM 2.5

IDENTIFYING ASSUMPTION

In order to recover an unbiased estimate of ATE with DD, the main assumption is **common parallel trends** between municipalities that have air-quality regulations and municipalities that do not. In other words, we assume that the trend overtime between unregulated municipalities is the same as the trend of regulated municipalities over time.

This assumption only requires for treated and untreated municipalities to have common trends, but it doesn't require for the trends to be at the same levels. If we assume that both groups of municipalities are trending the same, then we can use the trend in the unregulated municipalities as a counterfactual for the regulated municipalities to measure the effect of air-quality regulations on local particulate matter.

EXAMPLES OF REMAINING CONCERNS

If the parallel trends assumption is violated, we would have a problem as we would be recovering a **biased ATE**. Two examples of how this could happen:

1. For example, we could be comparing municipalities whose local particulate matter is *trending downward* because of active environmental policies different than air-quality regulations, to municipalities that have a *flat trend* because pollution has not changed by much over time. It could be the case that air quality regulations are effective in decreasing local 2.5 matter but since the counterfactual we are using

was trending downwards, we would be led to erroneuously conclude that the decline in pollution was bigger than it actually was.

2. Likewise, if regulated municipalities were trending upwards and unregulated municipalities were trending downwards, and there is an effect in lowering pollution from air-quality control regulations, then we would be drastically *overestimating* this effect because the POST difference would be much higher.

(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.

In this situation, unlike question (3), municipalities imposed air-quality regulations in different years instead of all at the same time.

In order to compare them, we can implement an **Even Study design**, which is a general form of a Fixed Effects design that allows for differential effects on local particulate matter overtime. The equation for this regression would be the following:

$$Y_{it} = \sum_{r=-s}^{R} \tau_r D_i * 1[\text{periods post treatment} = r] + \alpha_i + \delta_t + \beta X_{it} + \epsilon_{it}$$

Where:

- Y_{it} is local particulate matter 2.5, which varies by municipality over time
- $1[\text{periods post treatment} = r]$ is a dummy variable that equals 1 when we are r periods after air-quality regulation, and 0 otherwise
- $X_i t$ are covariates that vary by municipality over time
- τ_r recover the ATE r periods after air-quality regulations are imposed
- R is the number of post-regulation periods
- S is the number of pre-regulation periods

What this design achieves is "lining-up" air-quality regulation (ie. treatment) at the same time for all municipalities so we can compare them and recover ATE. In other words, by estimating this model, we are comparing municipalities on a given number number of years before and after they were regulated, indifferent of whether they were regulated in different points in time. It's a way of putting all municipalities on the same playing field so we can compare them.

For this Event Study design, we require pre-treatment $\tau_s' s$ to be centered around 0 instead of trending. It's very important to show these type of designs graphically to visualize the effects of the time to treatment and from treatment.

An additional benefit, is that this is a way to get a partial test of the identifying assumption of **common parallel trends** because we can visually observe if unregulated municipalities were trending the same before being regulated.

(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?

1. We implement a simple comparison of average local 2.4 between municipalities with and without air quality regulations.

```
# Create dummies for regulated and unregulated municipalities
data$treat_municipalities = ifelse(is.na(data$air_quality_regulation_year), 0, 1)
```

From this simple regression we find that, the estimated difference between a municipality that has air-quality regulations in place and a municipality that does not, is predicted to be approximately -24.45 micrograms per cubic meter of PM 2.5. This difference in means, is highly statistically significant at all conventional levels of significance, such that we reject the null in favor of the alternative that this difference is significantly different from 0. The sign of the coefficient is negative, as we would expect, meaning that air quality regulations is associated with a decline in particulate matter.

2. Next, we use time-series of the effect of air quality regulations on particulate matter, using only municipalities who introduced regulations in 2004.

```
# Create filter of 2004 treated municipalities
data_2004 <- data %>%
  filter(air_quality_regulation_year == 2004)

data_2004$post <- ifelse(data_2004$year >= 2004, 1, 0)
reg2 <- lm(particulate_matter ~ post, data = data_2004)
summary(reg2)
```

```
##
## Call:
## lm(formula = particulate_matter ~ post, data = data_2004)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.937 -14.320   5.748  28.010  56.389
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   46.132      3.403  13.556 < 0.0000000000000002 ***
## post         -21.470      3.587  -5.985        0.00000000301 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.03 on 998 degrees of freedom
## Multiple R-squared:  0.03465,    Adjusted R-squared:  0.03369
## F-statistic: 35.83 on 1 and 998 DF,  p-value: 0.000000003006
```

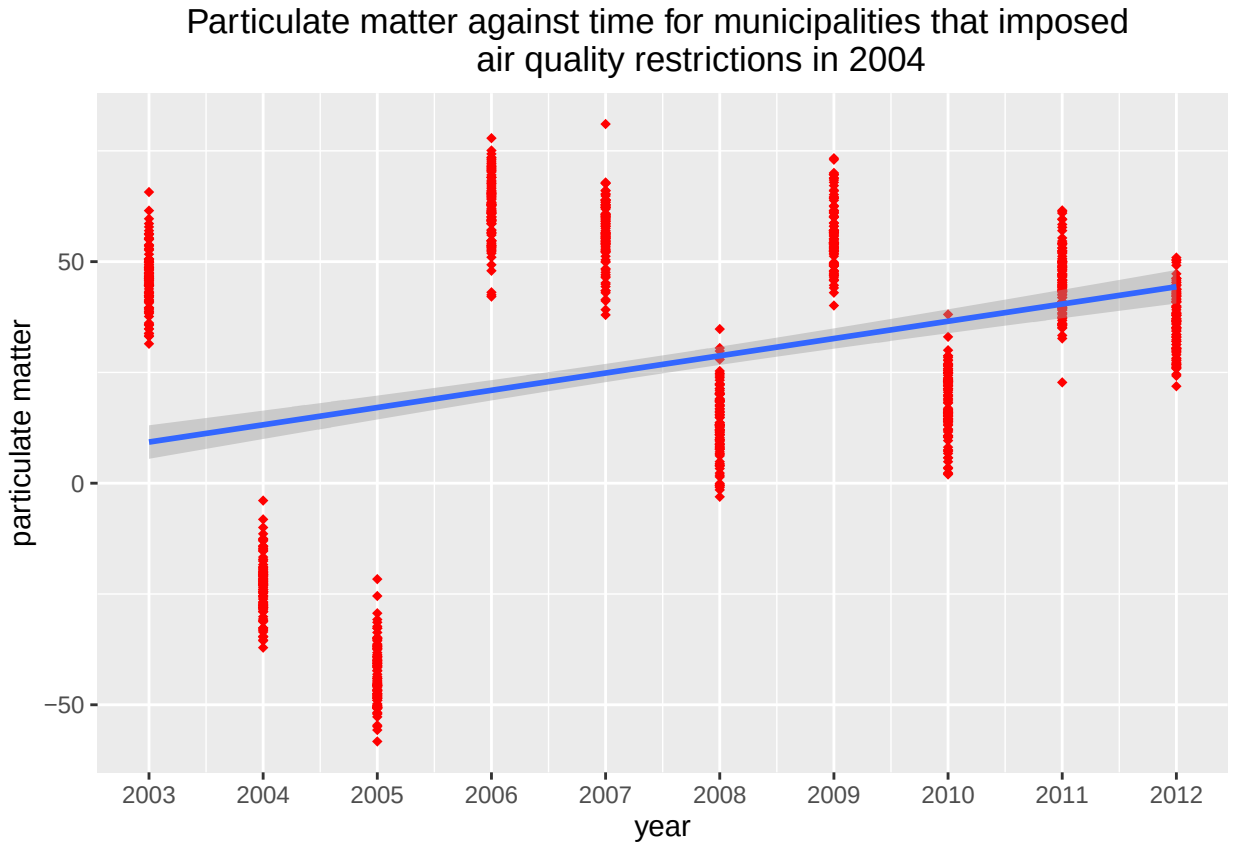
Next, for the time-series estimator, we find that *overtime*, the estimated difference between a municipality that has air-quality regulations in place and a municipality that does not, is approximately -21.47 micrograms per cubic meter of PM 2.5. This ATE estimate is also highly statistically significant at all conventional levels of significance and like before, retains the negative sign on the coefficient.

Compared to the previous naive estimate, the size of the estimated ATE is somewhat smaller (-21.47 vs. the -24.45 previously obtained). In other words, there is not much of a difference and this may be due to the problems previously described in part (2) and part (3) of selection bias and inexistence of time-varying characteristics.

3. We now plot particulate matter against time for municipalities that imposed air quality restrictions in 2004.

```
ggplot(data_2004, aes(x=year, y=particulate_matter)) +
geom_point(shape=18, color="red") + geom_smooth(method=lm) +
  scale_x_continuous(breaks = seq(2003, 2012, 1)) +
ggtitle("Particulate matter against time for municipalities that imposed
        air quality restrictions in 2004") + theme(plot.title = element_text
        (hjust = 0.5)) + xlab("year") +
        ylab("particulate matter")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

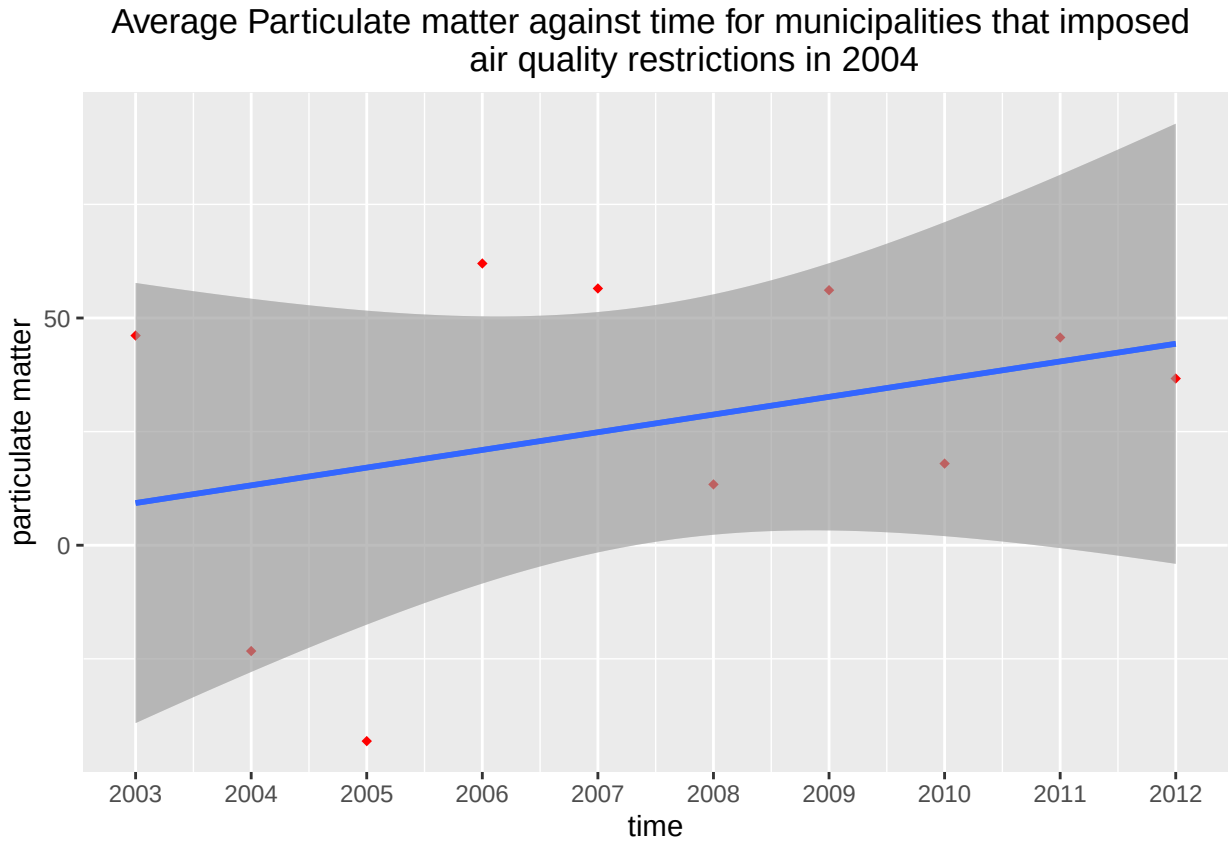


What we can see is that in the given years, for municipalities that imposed air quality regulations in 2004, there is an upwards trend particulate matter over time. In other words, there is a positive trend in pollution over time. It looks like right after 2004, the levels of particulate matter are much lower than the levels in particulate matter before 2004. The levels of particulate matter decrease even more in 2005. However, this effect is dissipated, as moving forward from 2005, there is an overall increase in particulate matter, some years lower some years higher, but overall much higher than in 2004 and 2005.

```
grouped <- data_2004 %>%
  group_by(year) %>%
  summarise(particulate_matter = mean(particulate_matter))
```

```
ggplot(grouped, aes(x=year, y=particulate_matter)) +
geom_point(shape=18, color="red") + geom_smooth(method=lm) +
  scale_x_continuous(breaks = seq(2003, 2012, 1)) +
ggtitle("Average Particulate matter against time for municipalities that imposed
        air quality restrictions in 2004") + theme(plot.title = element_text
        (hjust = 0.5)) + xlab("time") +
        ylab("particulate matter") +
         geom_smooth(formula = y ~ x, method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



If instead, we plot the mean particulaate matter consumption across municipalities, we can make similar conclusions as the previous graph. In 2004 there is a sharp decrease in mean PM 2.5 consumption that decreases further in 2005. However, after 2005, there is an upwards trend. In some years there is much higher consumption like 2006, 2007 and 2009, whereas other years like 2008 and 2010 have a comparatively lower PM 2.5 but not as low as in 2004-2005 right when and right after the air-quality regulations were imposed. Only looking at these graphs, we might believe the regulations were effective right after being imposed with effects dissipating overtime.

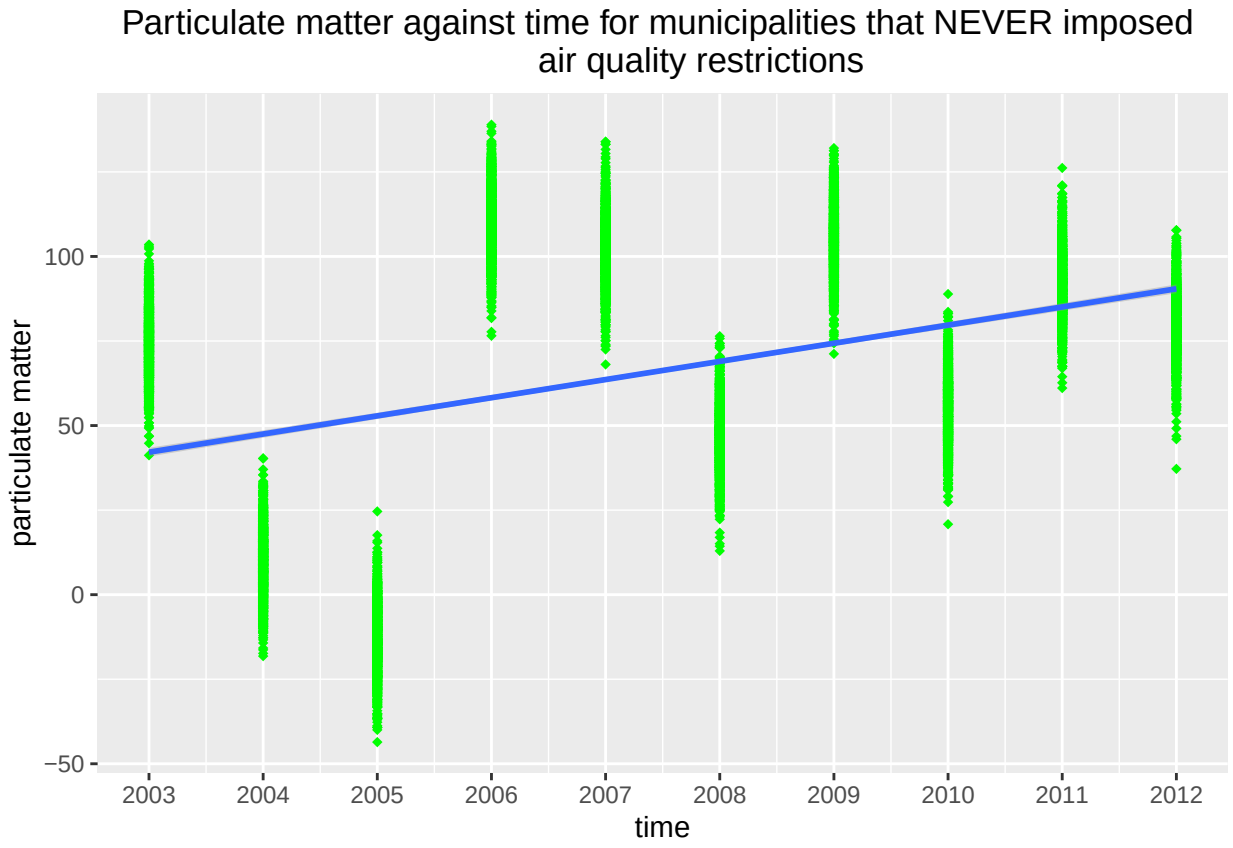
(7) Plot (average) particulate matter against time for municipalities who never imposed air quality regulations. Assess the viability of using these municipalities as a control group for the 2004 regulators. Plot (average) particulate matter against time for municipalities who passed air quality regulation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

1 Particulate matter against time for municipalities who never imposed air quality regulations

```r
# Filter of municipalities that never imposed regulations
no_regulation <- data %>%
  filter(treat_municipalities == 0)

# Plot
ggplot(no_regulation, aes(x=year, y=particulate_matter)) +
geom_point(shape=18, color="green") + geom_smooth(method=lm) +
  scale_x_continuous(breaks = seq(2003, 2012, 1)) +
ggtitle("Particulate matter against time for municipalities that NEVER imposed
        air quality restrictions") + theme(plot.title = element_text
        (hjust = 0.5)) + xlab("time") +
        ylab("particulate matter")
```

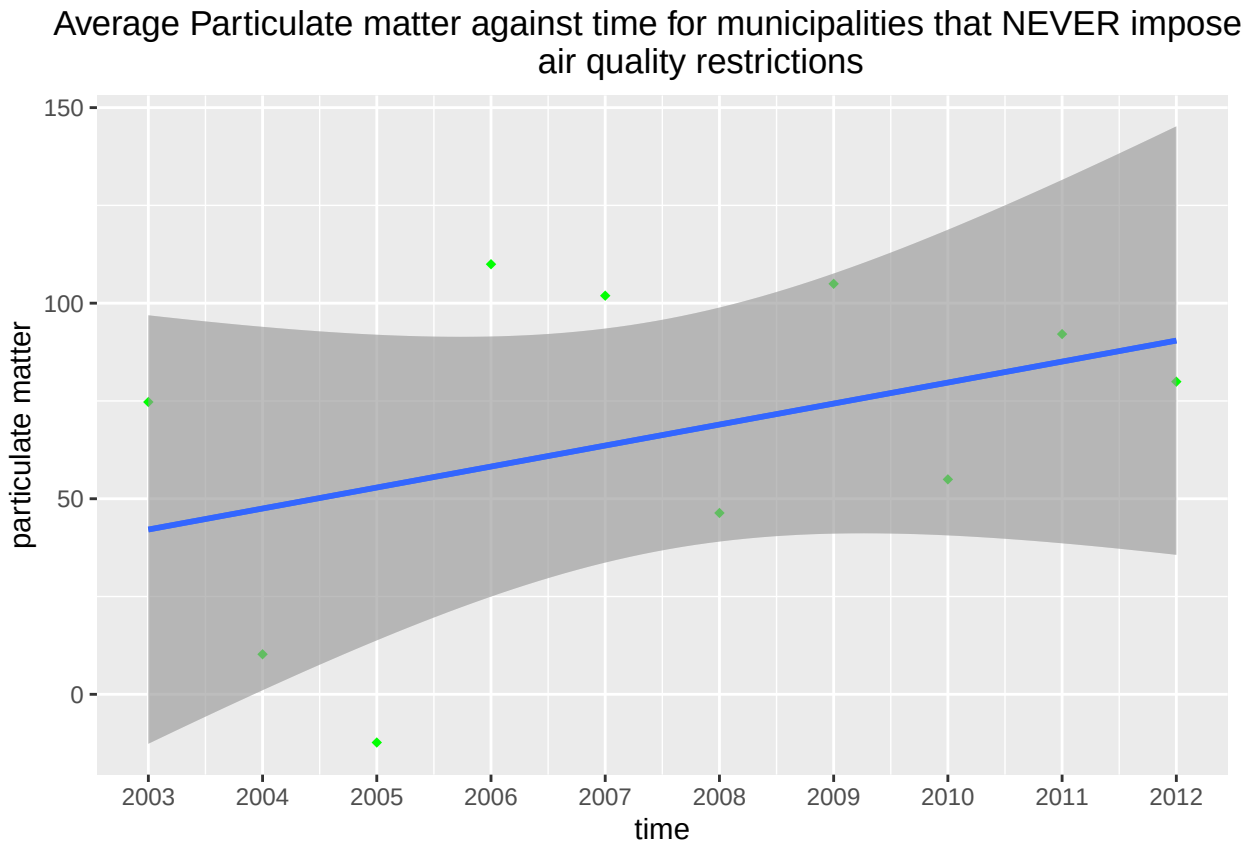
```
## `geom_smooth()` using formula 'y ~ x'
```




```r
grouped <- no_regulation %>%
  group_by(year) %>%
  summarise(particulate_matter = mean(particulate_matter))

ggplot(grouped, aes(x=year, y=particulate_matter)) +
geom_point(shape=18, color="green") + geom_smooth(method=lm) +
  scale_x_continuous(breaks = seq(2003, 2012, 1)) +
ggtitle("Average Particulate matter against time for municipalities that NEVER imposed
        air quality restrictions") + theme(plot.title = element_text
        (hjust = 0.5)) + xlab("time") +
        ylab("particulate matter") +
         geom_smooth(formula = y ~ x, method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Now, when we look at particulate matter consumption for municipalities that never imposed air quality restrictions, we find that the trend is actually pretty similar as that for for municipalities that impose air quality restrictions. The decline in 2004 and 2005 is also present for unregulated municipalities and the overall positive trend overtime, particularly after 2005 remains the same. This provides evidence in favor that there are common parallel trends between unregulated and regulated municipalities and motivates the use of a DD estimator as its identifying assumption is satisfied.

We might notice that the levels of the trends are different. For example, the range of the scale for unregulated municipalities is lower than the range in scale for regulated municipalities. Nonetheless, they are trending the same, which is the key DD identifying assumption to retrieve an unbiased ATE estimate. Therefore, using municipalities that never imposed air-quality regulations is a good control group for the 2004 regulators.

2 Plot average particulate matter against time for municipalities who passed air quality regulation in 2006

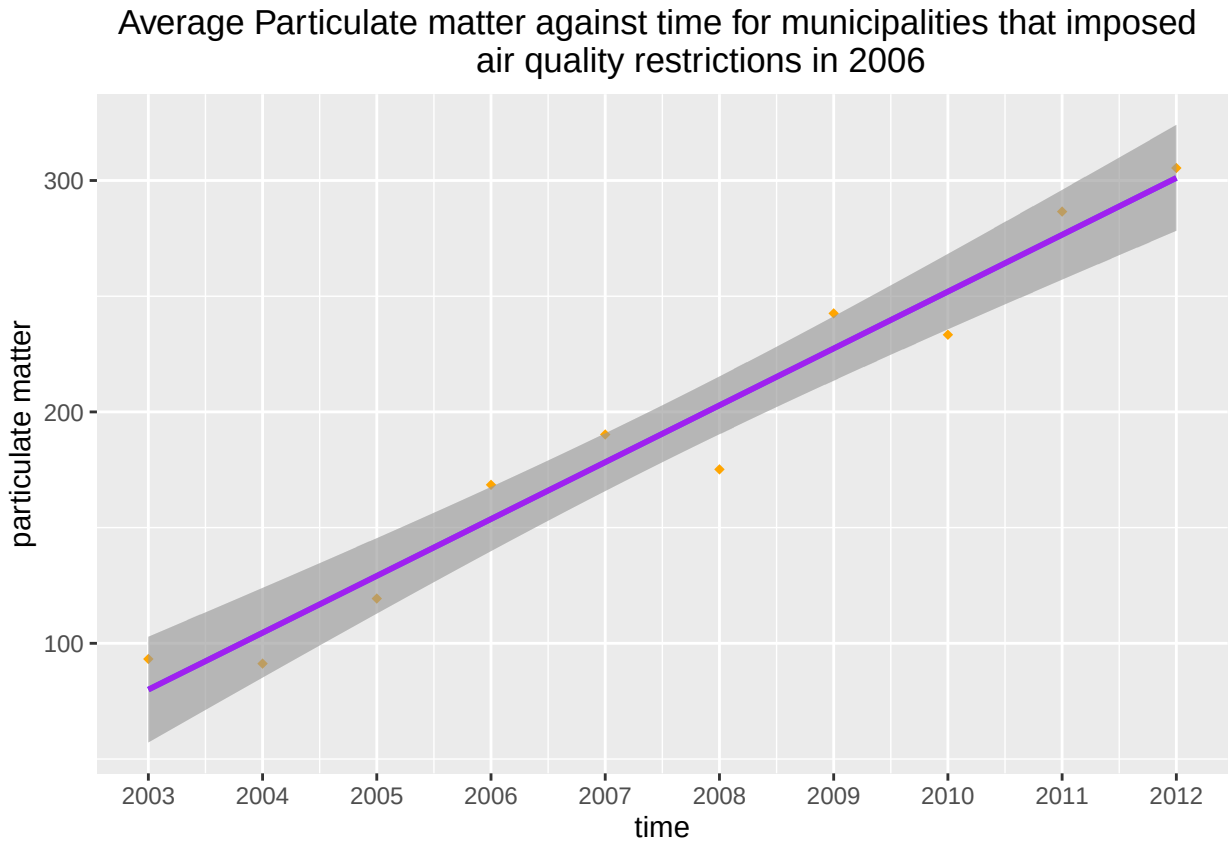
```
# Create filter of 2006 treated municipalities
data_2006 <- data %>%
  filter(air_quality_regulation_year == 2006) %>%
  mutate(post = ifelse(year  >= 2006, 1, 0))

grouped <- data_2006 %>%
  group_by(year) %>%
  summarise(particulate_matter = mean(particulate_matter))


ggplot(grouped, aes(x=year, y=particulate_matter)) +
geom_point(shape=18, color="orange") + geom_smooth(method=lm) +
  scale_x_continuous(breaks = seq(2003, 2012, 1)) +
```

```
ggtitle("Average Particulate matter against time for municipalities that imposed
        air quality restrictions in 2006") + theme(plot.title = element_text
        (hjust = 0.5)) + xlab("time") +
        ylab("particulate matter") +
         geom_smooth(formula = y ~ x, method = "lm", color="purple")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Looking at this plot, we can observe a *very different* trend than the previous two graphs. The slope of the trend is much steeper compared to the previous graphs where it was relatively more flat, albeit positive trending. This means that using non-regulators as a control group for municipalities that imposed air-quality regulations in 2006 violates the common parallel trends assumption and would not be valid.

(8) Using just the non-regulators and the 2006 regulators, estimate the causal impact of imposing air quality regulation on particulate matter. To do this, begin with a simple difference in means (rather than regression). Next, use a simple regression (no fixed effects). Finally, use fixed effects to control for common time shocks and time-invariant municipality characteristics (you can do this either via dummy variables or de-meaning). Report what you find. Be sure to adjust your standard errors appropriately in the regression-based estimates. Describe how this compares to what you estimated in (6) and (7).

1. Simple difference in means

```
dd<-data %>%
 filter(air_quality_regulation_year==2006|is.na(air_quality_regulation_year))
dd[is.na(dd)]<-0
dd<-dd %>%
 mutate(treated=ifelse(air_quality_regulation_year==2006,1,0))
dd<-dd %>%
```

```
 mutate(post=ifelse(year>=2006,1,0))
dd<-dd %>%
 mutate(treat_post=treated*post)

naive_estimator<-mean(dd$particulate_matter[dd$air_quality_regulation_year==
2006])-mean(dd$particulate_matter[dd$air_quality_regulation_year==0])
naive_estimator
```

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

If we only do a difference in means berween regulated and non-regulated municipalities in 2006, we obtain a difference of 124.27 micrograms per cubic meter of PM 2.5. This is can be argued to be large in size but also, it has a positive sign, which is concerning. It suggests air quality control regulations is correlated with an increase in particulate matter. This is the naive estimator and the problems with selection bias previously identified hold, which is why it is unreliable

2. Simple regression with no FE

```
reg3<-felm(particulate_matter ~ treated + post + treat_post|0|0|
                municipality_id,
                data = dd)
summary(reg3)
```

```
##
## Call:
##    felm(formula = particulate_matter ~ treated + post + treat_post |      0 | 0 | municipality_id, da
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.681 -26.395   1.227  21.816  92.576
##
## Coefficients:
##             Estimate Cluster s.e. t value            Pr(>|t|)
## (Intercept)  24.2163       0.1690   143.3 <0.0000000000000002 ***
## treated      77.0488       0.4666   165.1 <0.0000000000000002 ***
## post         60.0929       0.2022   297.1 <0.0000000000000002 ***
## treat_post   67.4689       0.5447   123.9 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.05 on 11986 degrees of freedom
## Multiple R-squared(full model): 0.6916   Adjusted R-squared: 0.6915
## Multiple R-squared(proj model): 0.6916   Adjusted R-squared: 0.6915
## F-statistic(full model, *iid*): 8960 on 3 and 11986 DF, p-value: < 0.00000000000000022
## F-statistic(proj model): 1.322e+05 on 3 and 1198 DF, p-value: < 0.00000000000000022
```

Here we find, that the estimated ATE for municipalities after they were regulated, compared to unregulated municipalities, is is 67.47 micrograms per cubic meter of PM 2.5. This coefficient is statistically significant at all conventional levels. It has decreased in magnitude from the naive estimator but it also has a positive sign, which is counter intuitive.

3.

```
reg4<-plm(particulate_matter~ treated + post + treated*post,data =dd,
 model = "within",
 effect = "twoway",
 index = c("municipality_id", "year"))
```

```
summary(reg4)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = particulate_matter ~ treated + post + treated *
##     post, data = dd, effect = "twoway", model = "within", index = c("municipality_id",
##     "year"))
##
## Balanced Panel: n = 1199, T = 10, N = 11990
##
## Residuals:
##         Min.      1st Qu.       Median      3rd Qu.         Max.
## -100.5959207   -7.5620690   -0.0039282    7.5231201   90.4826066
##
## Coefficients:
##              Estimate Std. Error t-value              Pr(>|t|)
## treated:post  67.4689     1.3037  51.751 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    4403400
## Residual Sum of Squares: 3527200
## R-Squared:      0.19898
## Adj. R-Squared: 0.10923
## F-statistic: 2678.16 on 1 and 10781 DF, p-value: < 0.000000000000000222
```

Here, with the DD estimator, we find that the estimated ATE is is 67.47 micrograms per cubic meter of PM 2.5. This coefficient is statistically significant at all conventional levels and is exactly the same as the one obtained previously. The positive sign remains, suggesting there is a positive relation between air-quality regulations and particulate matter, which is contrary to what we would expect. While it is likely to imagine that air-quality regulations may have no effect on pollution (for example if they are not enforced), it seems strange to image a scenario where they might actually lead to an increase in pollution, rather than a decrease. This may be due to some noise in the data that is inducing bias in our estimates.

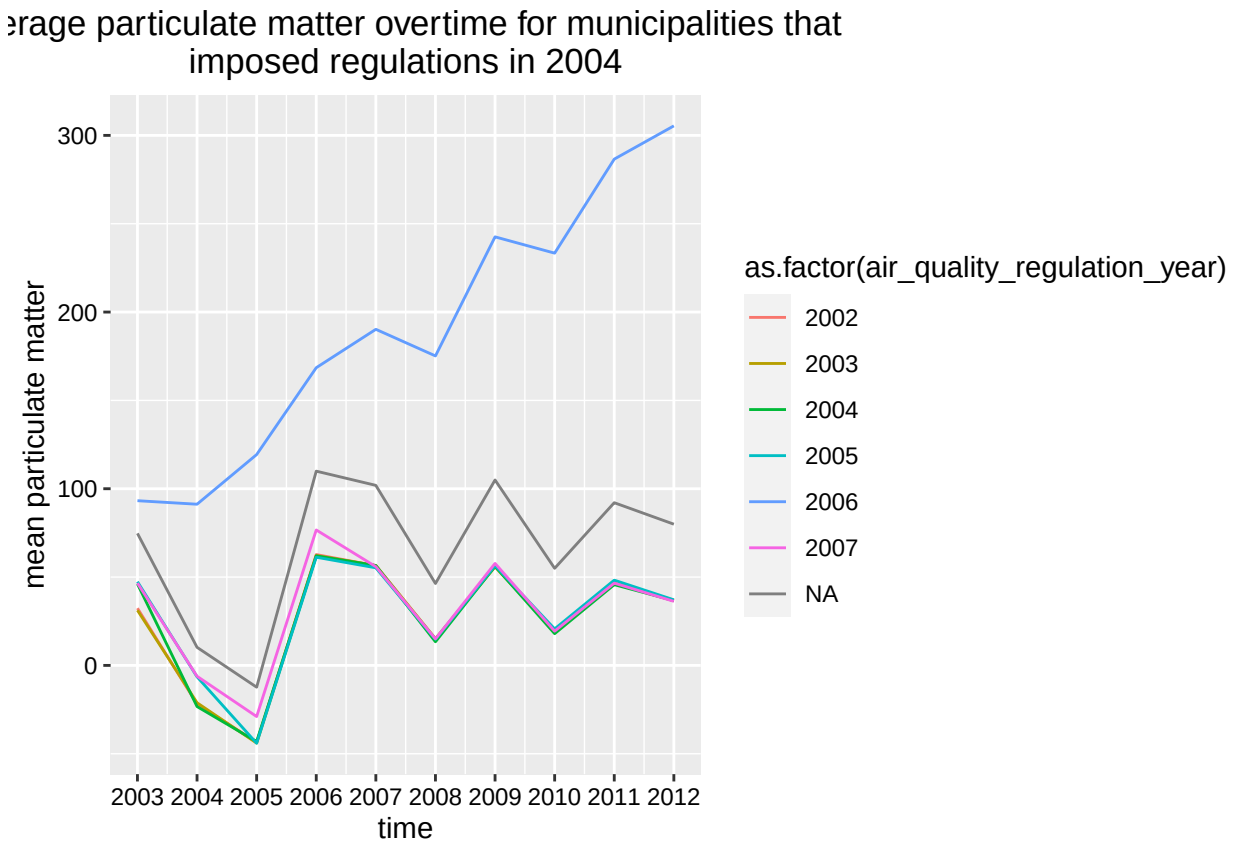
(9) Plot average particulate matter over time, separately (but on the same graph) for municipalities that imposed air quality regulations in each of the years from 2003 to 2007. Drop the municipalities that regulated air quality in the year that looks different from the rest of the years, and explain why you can't estimate a credible causal effect for these municipalities. Using the remaining municipalities to estimate a panel fixed effects regression to identify the causal effect of air quality regulation on particulate matter. Describe what you find. How does this compare to what you estimated in (8)? Use an event study regression to estimate how this treatment effect varies over time. Note that you will have to omit one of the event study treatment dummies (otherwise everything will be collinear). Standard practice is to leave out the T-1 dummy. Plot the resulting event study point estimates and 95 percent confidence intervals. Describe how the treatment effect varies over time, if at all.

1. Plot average particulate matter over time for municipalities that imposed air-quality regulations

```
regulators <- data %>%
  group_by(air_quality_regulation_year, year) %>%
  summarize(mean_particulate = mean(particulate_matter),
            .groups = 'keep') %>% ungroup()

regulators %>%
```

```
ggplot(aes(y = mean_particulate, x = year, color =
             as.factor(air_quality_regulation_year))) +
geom_line() + scale_x_continuous(breaks = seq(2003, 2012, 1)) +
scale_y_continuous(labels = scales::comma) +
  theme(axis.text=element_text(colour="black")) +
ggtitle("Average particulate matter overtime for municipalities that
imposed regulations in 2004") + theme(plot.title = element_text
        (hjust = 0.5)) + xlab("time") +
        ylab("mean particulate matter")
```



From this graph, we can observe that the blue line that represents the year 2006 is trending *very differently* from all other years. All other years follow the same trend and, even at the same level from 2007 onwards. Overall, all other years show a sharp decline up to 2005, followed by an general increase with ups and downs. However, 2006 has a very steep increase overtime that is on a different trend entirely.

We need to omit the year 2006 from the dataset because it violated the **parallel trends** assumption we need for a DD estimator when using panel data. As explained in part (4), we cannot recover a credible unbiased ATE estimate for 2006 municipalities because, given that the identifying assumption is violated, our counterfactual is no longer valid. Essentially, after imposing regulations, we would be incorrectly *under-estimating* an effect that cannot be attributed solely to air-quality regulations but rather that effect PLUS noise from other factors that are correlated with particulate matter levels. It would result in a **biased estimate**.

2. Estimate panel fixed effects regression of air quality regulation on particulate matter

```
# Remove municipalities from the year 2006, include treat and post dummy
# variables:

data_clean <- data %>%
filter(air_quality_regulation_year != 2006 | is.na(air_quality_regulation_year)) %>%
```

```
mutate(treat = ifelse(is.na(air_quality_regulation_year), 0, 1),
      post = ifelse(year >= air_quality_regulation_year &
                      air_quality_regulation_year!=0, 1, 0),
      dd = treat*post)

# Run fixed effects regression
reg5 <- felm (particulate_matter ~ dd|municipality_id +
                year|0|municipality_id, data = data_clean)
summary(reg5)
```

```
##
## Call:
##    felm(formula = particulate_matter ~ dd | municipality_id + year |      0 | municipality_id, data =
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.109  -4.636   0.012   4.580  29.445
##
## Coefficients:
##    Estimate Cluster s.e. t value            Pr(>|t|)
## dd   -14.84         0.36  -41.21 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.308 on 8999 degrees of freedom
##   (10990 observations deleted due to missingness)
## Multiple R-squared(full model): 0.9579   Adjusted R-squared: 0.9532
## Multiple R-squared(proj model): 0.1517   Adjusted R-squared: 0.05654
## F-statistic(full model, *iid*):202.9 on 1010 and 8999 DF, p-value: < 0.00000000000000022
## F-statistic(proj model):  1698 on 1 and 1000 DF, p-value: < 0.00000000000000022
```

From the panel fixed effects regression, we find our DD coefficient is -14.84, which means that the predicted local particulate matter for municipalities that impose air-quality regulations, in comparison with unregulated municipalities, is predicted to be a decrease in -14.84 micrograms per cubic meter of PM 2.5. This estimate is statistically significant at all conventional levels. Contrary to the previous estimates in part 8, this time our estimate has a negative sign, suggesting a negative correlation between air quality regulations and pollution, which is more in line with what we would expect.

In comparison to part (8), this is a more reliable estimate because we dropped municipalities that imposed regulations in 2006 from the data, which was inducing noise in our estimate by violating the parallel trends assumption. Those observations were pushing the coefficient to have a positive sign, which was an unusual result. Having gotten rid of that noise, has improved the reliability of our estimate.

3. Event study regression

```
event_study <- data_clean %>%
  mutate(air_quality_regulation_year = ifelse(air_quality_regulation_year == 0, NA, air_quality_regulati
t_min_1 = year - air_quality_regulation_year)

# Exclude periods t minus to prevent collinearity
tmin1 <- event_study%>%
  filter(t_min_1 != -1)

# Run Event Study regression with time and municipality fixed effects
reg6<- plm(particulate_matter~ as.factor(t_min_1)*post,
          data = tmin1,
```

```
          model = "within",
          effect = "twoway",
          index = c("municipality_id", "year"))
summary(reg6)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = particulate_matter ~ as.factor(t_min_1) * post,
##     data = tmin1, effect = "twoway", model = "within", index = c("municipality_id",
##         "year"))
##
## Unbalanced Panel: n = 1001, T = 9-10, N = 9709
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -34.187331  -4.605916   0.061209   4.558532  29.406855
##
## Coefficients: (13 dropped because of singularities)
##                        Estimate Std. Error  t-value              Pr(>|t|)
## as.factor(t_min_1)-3   2.25155    1.09674   2.0530              0.040107 *
## as.factor(t_min_1)-2   2.54125    0.84962   2.9910              0.002788 **
## as.factor(t_min_1)0  -10.97258    0.70613 -15.5389 < 0.00000000000000022 ***
## as.factor(t_min_1)1   -9.05262    0.57274 -15.8059 < 0.00000000000000022 ***
## as.factor(t_min_1)2   -8.32792    0.61046 -13.6421 < 0.00000000000000022 ***
## as.factor(t_min_1)3   -7.11559    0.56690 -12.5517 < 0.00000000000000022 ***
## as.factor(t_min_1)4   -6.04677    0.54655 -11.0635 < 0.00000000000000022 ***
## as.factor(t_min_1)5   -4.92683    0.50695  -9.7185 < 0.00000000000000022 ***
## as.factor(t_min_1)6   -4.07089    0.53752  -7.5735   0.00000000000004006 ***
## as.factor(t_min_1)7   -3.14738    0.50452  -6.2383   0.00000000046301721 ***
## as.factor(t_min_1)8   -2.14075    0.52073  -4.1111   0.00003974110184566 ***
## as.factor(t_min_1)9   -1.56909    0.55373  -2.8337              0.004612 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    518330
## Residual Sum of Squares: 463490
## R-Squared:      0.1058
## Adj. R-Squared: 0.00070008
## F-statistic: 85.6501 on 12 and 8687 DF, p-value: < 0.00000000000000222
```

```
results <- coeftest(reg6, vcov = vcovHC(reg6, type = "HC0", cluster = "group"))
results
```

```
##
## t test of coefficients:
##
##                        Estimate Std. Error  t value              Pr(>|t|)
## as.factor(t_min_1)-3   2.25155    1.14934   1.9590              0.050146 .
## as.factor(t_min_1)-2   2.54125    0.86510   2.9375              0.003317 **
## as.factor(t_min_1)0  -10.97258    0.73225 -14.9848 < 0.00000000000000022 ***
## as.factor(t_min_1)1   -9.05262    0.59324 -15.2597 < 0.00000000000000022 ***
## as.factor(t_min_1)2   -8.32792    0.61963 -13.4401 < 0.00000000000000022 ***
## as.factor(t_min_1)3   -7.11559    0.57104 -12.4607 < 0.00000000000000022 ***
## as.factor(t_min_1)4   -6.04677    0.56864 -10.6338 < 0.00000000000000022 ***
```

```
## as.factor(t_min_1)5   -4.92683    0.51290  -9.6059 < 0.00000000000000022 ***
## as.factor(t_min_1)6   -4.07089    0.54538  -7.4643   0.00000000000009185 ***
## as.factor(t_min_1)7   -3.14738    0.49074  -6.4136   0.00000000014958295 ***
## as.factor(t_min_1)8   -2.14075    0.50044  -4.2777   0.00001908373100343 ***
## as.factor(t_min_1)9   -1.56909    0.56009  -2.8015              0.005098 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coef <- results[1:12]
se <- results[13:24]
es <- data.frame(t_min_1 = c(c(-3,-2), c(0:9)), coef,se,
                 lb = coef - 1.96 * se, ub = coef + 1.96 * se)
```

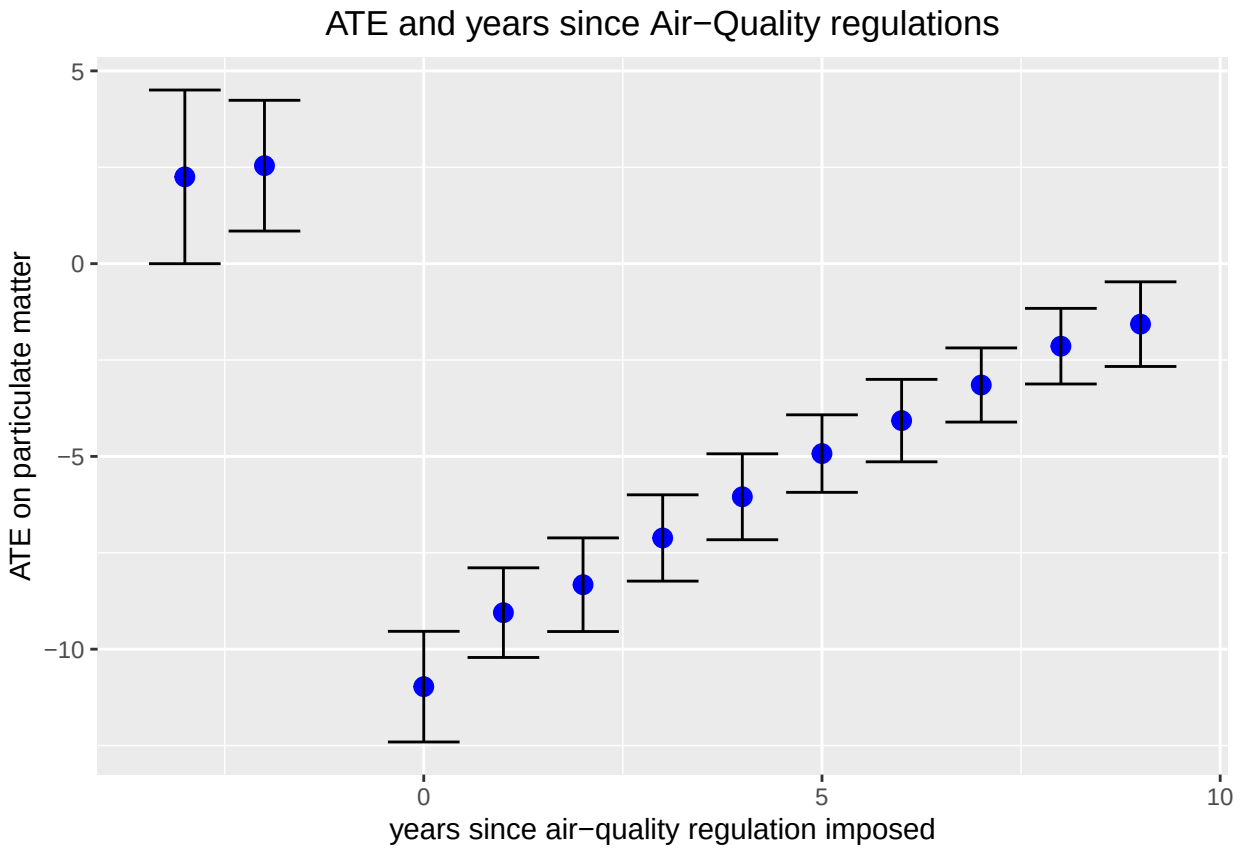
As described in part (5), an event study allows us to *line-up* air-quality regulations for all municipalities, so that we can compare and estimate the ATE of regulations on particulate matter, regardless of municipalities imposing regulations at different points in time.

This design lets us compare muncipalities PRE and POST treatment (ie. imposing air-quality regulations). What we can observe from the estimated coefficients is that the effects are statistically significant for all years for which we have data, that is in range from -3 years from regulations being imposed up to 9 years after they were imposed. In particular, the effects are highly statistically significant for years 0-8.

We can also observe that before treatment, at years -2 and -3, the coefficients were positive but in the year the regulations are imposed, on average, there is an estimated -10.97 decrease in per cubic meter of PM 2.5. This effect can be argued to be sizable, considering the change from 2.54 to -10.97 per cubic meter of PM 2.5, on average. However, the effects dissipate over time by approximately 1 full unit per year, where the magnitude of the ATE decreases all the way to -1.57 per cubic meter of PM 2.5. This result suggests air-quality regulations are effective right after they are imposed but overtime, their effect on pollution decreases substantially.

4. Plot event study estimates with 95% confidence intervals

```
ggplot(es, aes(x = t_min_1, y = coef)) +
  geom_point(size = 3, color = "blue") +
  geom_errorbar(aes(ymin = lb, ymax = ub)) +
  xlab('years since air-quality regulation imposed') +
  ylab('ATE on particulate matter') +
  ggtitle(label = 'ATE and years since Air-Quality regulations') +
  theme(plot.title = element_text
        (hjust = 0.5))
```



The plot shows the results previously discussed in a much more visual way. The ATE before air-quality regulations are imposed has a positive sign 2 and 3 years before these regulations are implemented. The year the regulations are implemented is where we see the most marked decline in local particulate matter. Each year after the regulation is imposed, on average, there is an increase in PM 2.5, showing the waning effects of this policy.

(10) PROGRAMEVAL would like a summary of your findings. Explain which of the results you've come up with is your preferred estimate, and why. Make sure to describe at least one remaining potential shortcoming with these results. Finally, interpret the magnitude of your estimated effects: do your results suggest that the PROGRAMEVAL should be strongly promoting air quality regulations?

From the results so far, my preferred estimate is the Event Study Design in part (9) because, after dropping municipalities that imposed regulations in 2006, which was violating the parallel trends assumption, we were able to recover a more reliable ATE that has a negative sign.

We can reasonably expect the sign to be negative or for the coefficient to be close to 0, in which case it would mean air-control regulations have no effect on particulate matter (for example if the policy is not enforced or has such high pollution caps that it is ineffective). However, it seems difficult to come up with reasoning that would sustain a positive correlation. It would be a situation in which, there are defying companies that would counter regulations with purposefully more polluting just to spite the government, which seems highly unlikely, particularly in the political context of China.

Unlike the panel fixed effects, the Event Design estimates are much more powerful in analyzing the effect of air-quality control regulations on PM 2.5 not only on average, but rather, overtime. The accompanying plot provides a more substantial and complete analysis that, even though the effect is negative as would be desirable, this effect dissipates over time, to the point it has barely any effect after 9 years of these regulations being implemented. This would allow PROGRAMEVAL to make more informed decisions about

implementing air-quality control policies.

Based on these results, I would recommend to PROGRAMEVAL to promote air-quality regulations and show that they should expect the biggest effects to happen near the time they are imposed. However, I would also advise they put measures in place that allow them to maintain these effects over time, like tightening air-quality regulations overtime, or increasing/maintaining enforcement so that they effects are sustained.

One potential short-coming of these results is we have very limited data before air-quality regulations were imposed. We are comparing up to 9 years after regulations vs only 3 years before these regulations. It's possible that going back there are other considerations in the way these municipalities were trending that could be related with the effectiveness of regulations on pollution. For instance, if it just so happens that those two years happened to be years with relatively low pollution and in the last decade the levels of pollution have in reality been much higher, then we might be over stating the economic significance of these results.