Program Evaluation Problem Set 4

Hye-Min Jung

5/28/2020

1.

- **In a completely unconstrained world, where I could do anything with almighty ability, I would run randomized controlled trial, to estimate "the impact of provincial(municipal) air quality regualtions on local particulate matter(PM 2.5)". Because under the randomized controlled trial, we can easily estimate the impact of treatment. ATE is simply the difference in means between treated and control provincial's particulate matter in the air.**
- Define: $D_i \in \{0, 1\}$ as the regulation indicator for municipal i
  - when municipal i is treated, $D_i = 1$
  - when municipal i is not treated, $D_i = 0$
- Define: $Y_i\left(D_i\right)$ as the local particulate matter as a function of D_i
  - when municipal i is treated, we observe local particulate matter$Y(1)$
  - when municipal i is not treated, we observe local particulate matter $Y(0)$
- Then, the impact of provincial(municipal) air quality regualtions on local particulate matter(PM 2.5) is: $\tau_i = Y_i(1) - Y_i(0)$
- To put it simply in words, impact of treatment is a parameter that measures 'What is the outcome of air quality regualtion, compared to the outcomes that we would have observed without the the regulation?'
  - Impacts is 'changes in outcomes caused by the regulation' whereas, outcomes is 'things that we could "potentially" observe'.
  - We need to consider all possible outcomes we could have observed, spans both actual and alternative programs, to explain the impact of treatment.
- To perform the analysis, I would need dataset at the municipal level, treatment status, pre-treatment observable chracteristics, post-treatment characteristics, post-treatment outcome.

2.

- If you compare the 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 cannot avoid fundamental problem of causal inference due to selection bias.
- Selection bias ex1
  - Treated(regulated) municipals rely heavily on manufacturing industry(factory), whereas untreated municipals are rural areas that does not have any pollution.
  - If you compare those two groups and estimate the impact of the regulation, you might end up with underestimated policy impact.
- Selection bias ex2
  - Treated municipal residents care more about the air quality(highly educated), so residents forced the policy to be implemented and also follows the guideline very well. On the other hand untreated municipal residents don't care about the air quality at all and don't follow the guidelines.
  - In this scenario, you might end up with overestimated policy impact.
- Selection bias ex3
  - Treated municipal urbanized cities with 99% of oil-engine cars, whereas untreated municipals are the areas where the 80% of residents has adopted eco-friendly electric car.
  - In this case, you might end up with underestimated policy impact.

3.

- We can leverage data for identification by comparing i in t vs. i in t-1 and by doing so, it becomes possible to get rid of the selection bias.
  - In this formulation, i serves as a control for itself.
  - As ‘post-treatment i’ is much more similar to ‘i before treatment’ than comparing to unit j, it will be less likely to have changes on unit i, which can cause unit i not to select into treatment yesterday, but to select in today for unit i.
  - Therefore, we can get rid of the selection bias by comparing i of t-1 to i to t. (Where as in cross-sectional data, there are many cases that units can be selected into treatment.)
- With this within unit comparison, we can estimate:
  - $\hat{\tau}^{TS} = Y_{i,t=1} - Y_{i,t=0}$
  - In words, this is going to be the outcome difference in before 2004 regualtion and after 2004 regulation.
- In order for $\hat{\tau}^{TS} = Y_{i,t=1} - Y_{i,t=0}$ to recover the causal effect, we need follwing identifying assumption:
  - $\delta = 0$ or $V_{i,t=1} = V_{i,t=0}\,(= V_i)$ from $\hat{\tau}^{TS} = Y_{i,t=1} - Y_{i,t=0} = \tau + \delta\,(V_{i,t=1} - V_{i,t=0})$, where V_{it} is a set of observed and unobserved time-varying characteristics.
- If any time-varying observed or unobserved chraceteristics(variables) matter for the outcome, that will create bias in $\hat{\tau}^{TS}$ and make you to worry, because we will confound the treand with τ and we will not be able to eliminate the trend with one time series.
- ex 1:
  - Suppose 2004, if the global pandemic outbreak put vast majority of factories into shut down, contributing to significant amount of reduction in particulate matter, we cannot separate the regulation effect from the global pandemic effect.
- ex 2:
  - Suppose 2004, if the destructive and unstoppable wildfire occurres in China and spread throughout the country emitting significant amount of particulate matter, then we cannot separate the regulation effect from the wildfire effect.
- ex 3:
  - Suppose 2004, if the extremely hazardous volcano erupts in China, emitting significant amount of particulate matter, then we cannot separate the regulation effect from the volcanic eruption effect.

4.

- Being able to have data on municipalities that had never been imposed to regualations as a gauge of regulated municipalities's trend over time, we can reasonably guess over the missing counterfactual trend for treated unit, in which we can never observe in real world.
- Differences-in-differences, $\hat{\tau}^{DD} = \hat{\tau}^{TS}_{D_i=1} - \hat{\tau}^{TS}_{D_j=0}$ is the estimator that I would like to estimate using this dataset.
  - I will have 2 time series estimator and take the difference between two. More specifically, I will compare the time series estimator for regulated municipalities to the time series estimator for unregulated municipalities.
  - In other words, this estimator compars treated to untreated units over time.
- Let, regulated units: $Y_{it} = \tau D_{it} + \beta X_i + \delta S_t$, never regulated units: $Y_{jt} = \beta X_j + \delta S_t$
  - We have time-invariant characteristics X_i and time-varying S_t
  - $\hat{\tau}^{DD} = (Y_{i,t=1} - Y_{i,t=0}) - (Y_{j,t=1} - Y_{j,t=0}) = \tau\,(D_{i,t=1} - D_{i,t=0}) = \tau$
- We can implement DD via the following regression:
  - $Y_i = \alpha + \tau Treat \times Post_{it} + \beta Treat_i + \delta Post_t + \varepsilon_{it}$
  - Running this regression yields $\hat{\tau} = \hat{\tau}^{DD}$
- The identifying assumption of the DD is "parallel counterfactual trends".
- The larger dataset, the panel data, would allow me to resolve the concern that I had with only time-series data.
  - Because using other units that never got treated, we can make a guess on the missing counterfactual.
  - If there is no change in outcome is for never regulated municipalities in pre and post regulation, it will be reasonable to extrapolate same treand for regulated municipalities too.
- Remaining concern with panel data still remains with the paralle trends assumption.
  - If the treated and untreated units are on different trajectory to begin with, comparing treated units pre and post treatment and untreated units pre and post will mis-adjustment and resulting in bad approximation of the counterfactual.
- ex 1.
  - Suppose, never regulated municipalities were well-educated and conscious about the air quality, so had decreasing PM 2.5 trend to begin with, whereas regulated municipalities had stable PM 2.5. In this case, the average PM 2.5 would be smaller for the post period for never regulated municipalities and in response, we will mis-adjust the counterfactual for regualted municipalities.
- ex 2.
  - Suppose, regulated municipalities were well-educated and conscious about the air quality, where as never regualted municipalities are careless about the air quality. So the trends for regulated and unregualted are different to begin with. We will have the similar mis-adjustment and bad approximation for the counterfactual in this case too.

5.

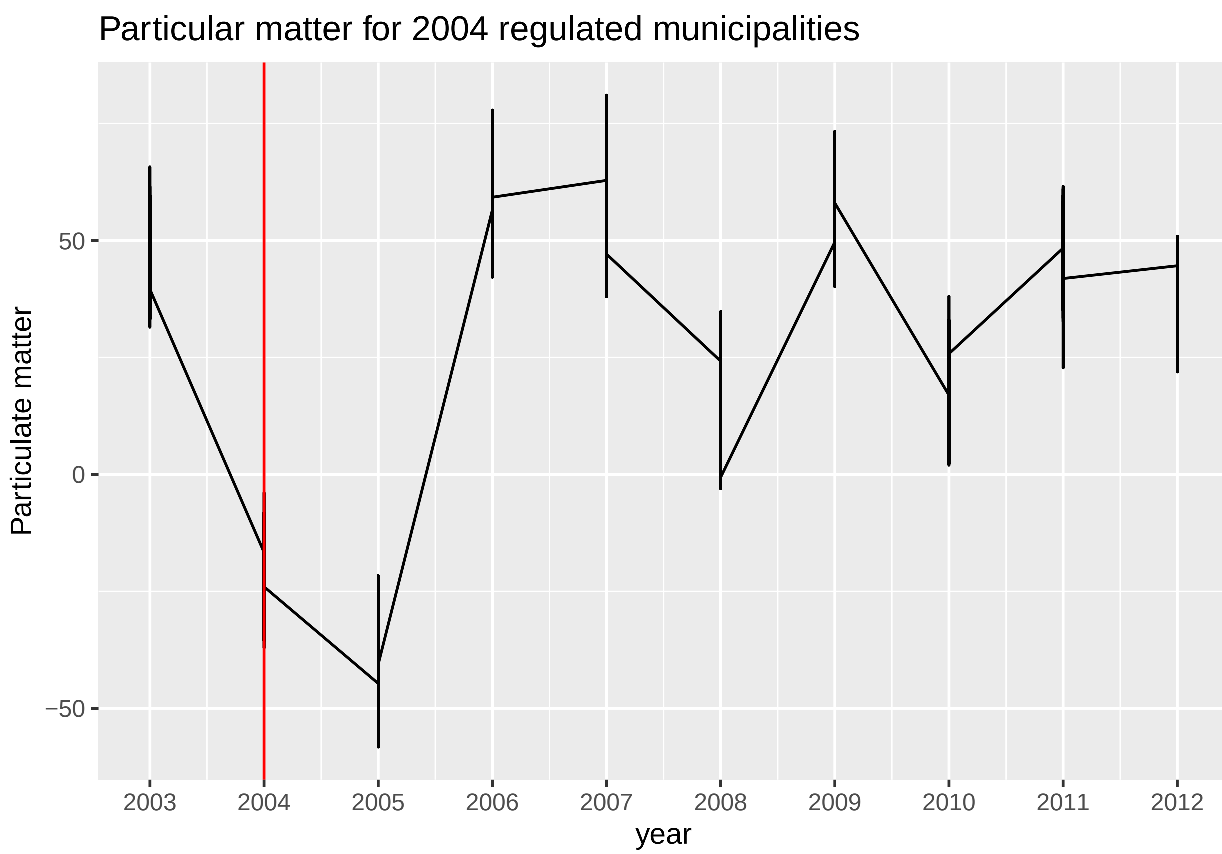
- Using data on the universe of their consumers from 2003 to 2007, I would estimate fixed effect to know the impact of regulation on air quality by:
  - Controlling for individual-specific time-invariant effects and common time-period-speicific effects that might be correlated with D_{it}.
  - In math, I would like to control for α_i and δ_t, that might be correlated with D_{it}. (In the basic panel data regression $Y_{it} = X_{it}\beta + \tau D_{it} + \varepsilon_{it}$, α_i and δ_t is in the error term.)
  - $Y_{it} = X_{it}\beta + \tau D_{it} + \alpha_i + \delta_t + \nu_{it}$

6.

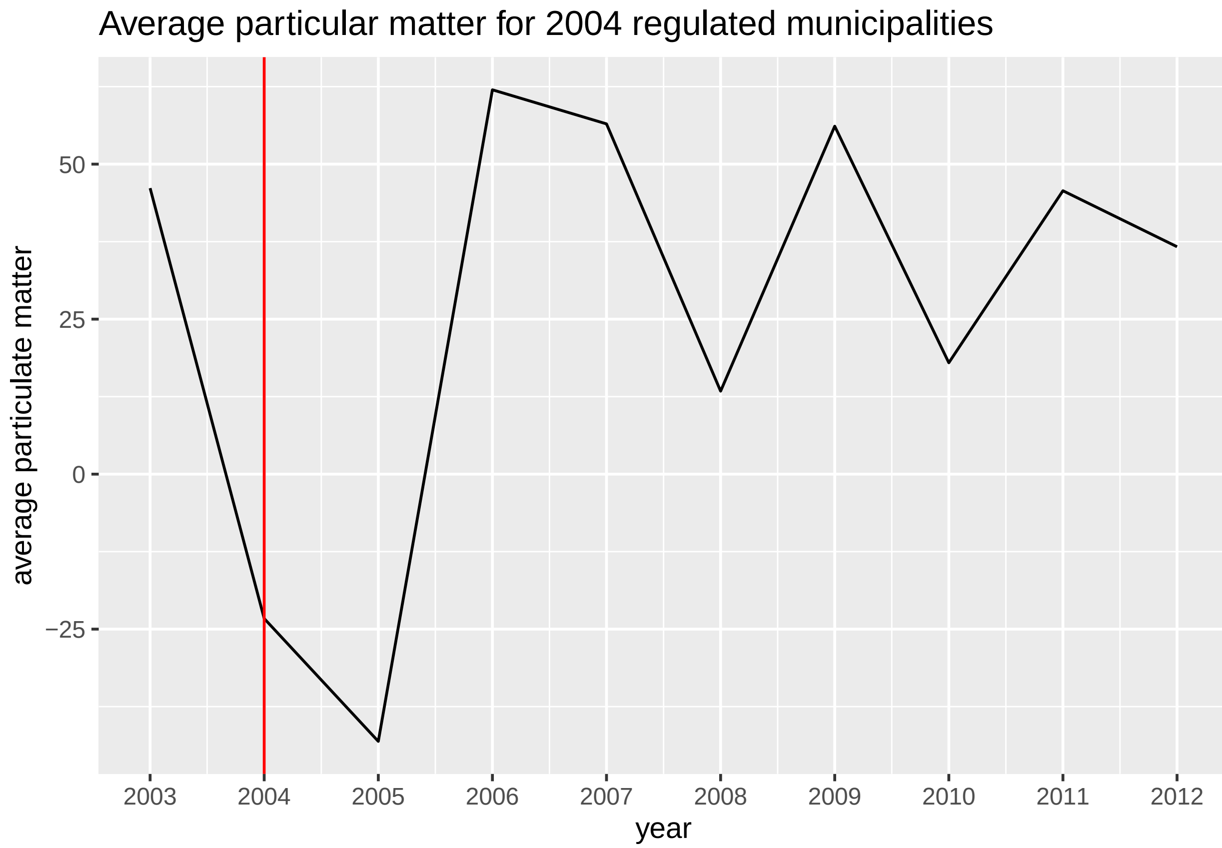
- Simple comparison result:
  - PM 2.5 is 24.44856 lower in average for 2004 regulated municipalities compare to the never regulated municipalities.
- Regression result for time-series analysis:
  - PM 2.5 is 21.470 lower in average for municipalities after the regulation in 2004 compare to PM 2.5 level before the regulation.
- Regression result is slightly different from the simple comparison. The magnitude is bigger when calculating with simple comparison.

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

```
##
## ==========================================
##                  Dependent variable:
##              ----------------------------
##                  particulate_matter
## ------------------------------------------
## post                  -21.470***
##                         (3.772)
##
## ------------------------------------------
## Observations            1,000
## R2                      0.035
## Adjusted R2            -0.073
## F Statistic    32.394*** (df = 1; 899)
## ==========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```



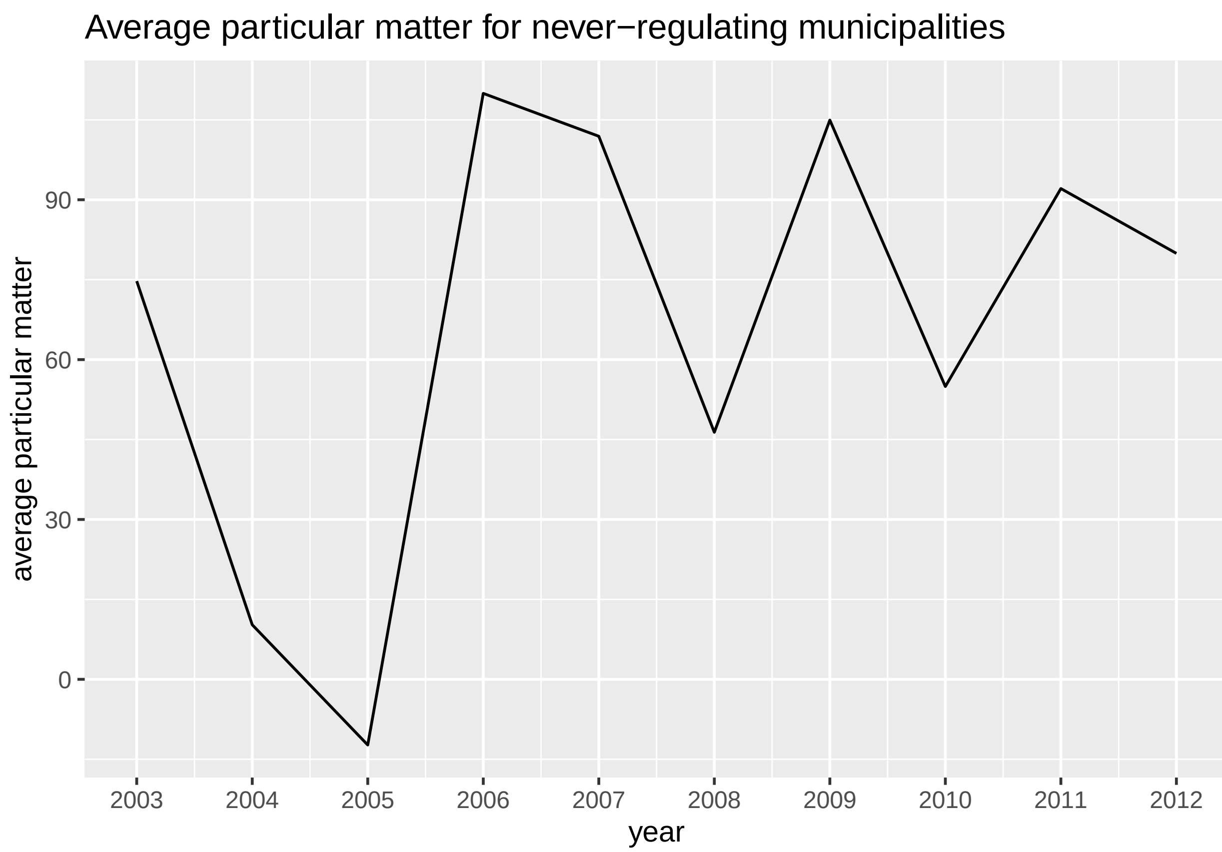

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```



- The average PM 2.5 decreases after the regulation, but it increases in 2005 and it continues to have fluctuating trajectory.
- This figure suggests that my estimate might be negative due to the big short term effect reducing PM 2.5 but does due to long lasting effect that reduces PM 2.5.

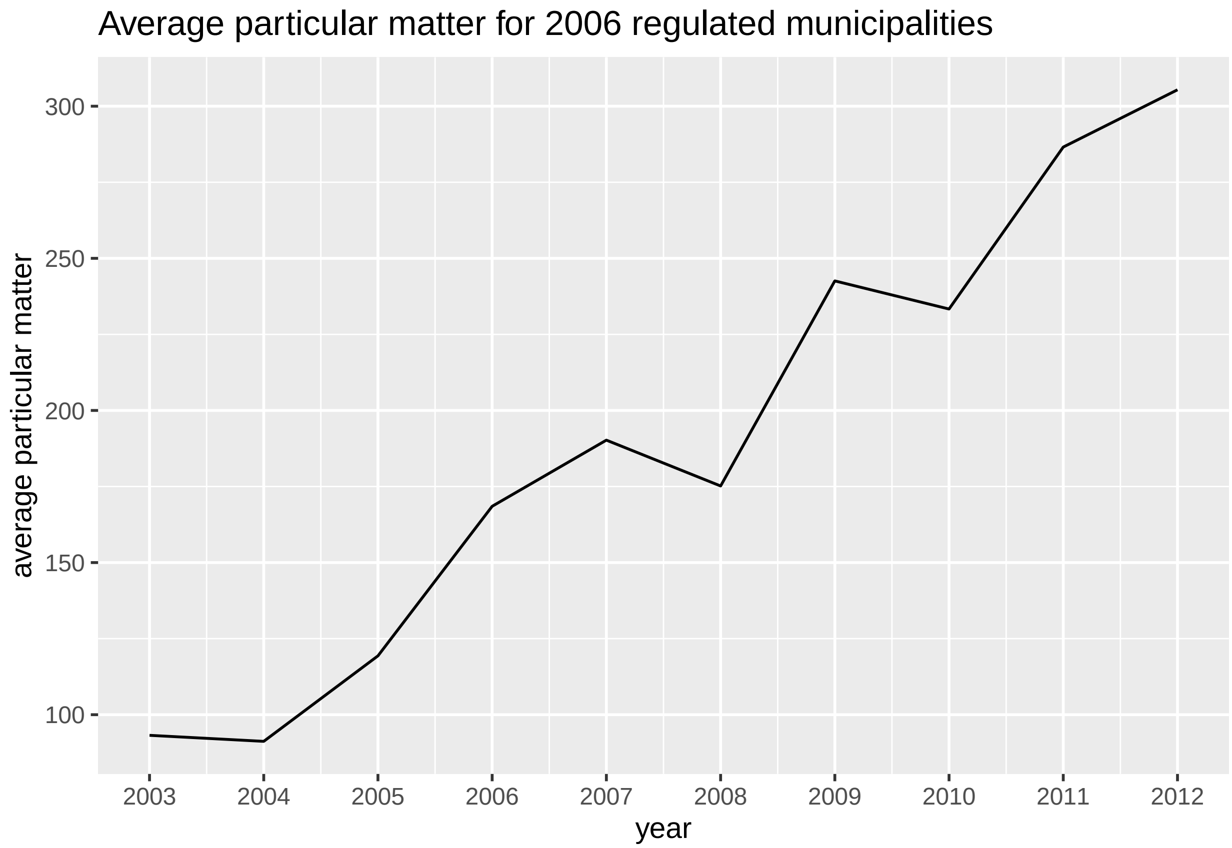
7.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



- As never-regulating municipalities before 2004 has parallel trend with 2004 regulating municipalities, it is reasonable to assume 'untreated trend=counterfactual trend', using it as a control group for the 2004 regulators.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



- However, never-regulating municipalities does not have parallel trend with 2006 regulating municipalities, before 2006, it is unreasonable to assume 'untreated trend=counterfactual trend', and it would be bad idea to use it as a control group for the 2006 regulators.

8.

- Simple difference in mean and regression both yield same result:
  - PM 2.5 is 124.277 higher in average for 2006 regulated municipalities compare to the never regulated municipalities.
- Fixed effects controling for common time shocks and time-invariant municipalitiy characteristics yield different result:
  - PM 2.5 is 67.469 higher in average for 2006 regulated municipalities compare to the never regulated municipalities.
- Q8 estimate is significantly different from Q6 estimate.
  - Q8: PM 2.5 is 124.277 or 67.469 higher in average for 2006 regulated municipalities compare to the never regulated municipalities.
  - Q6: PM 2.5 is 24.44856 lower in average for regulated municipalities compare to the never regulated municipalities.
  - As we have discovered in Q7 plots, this significant difference comes from the unparallel pre-trend of 2006 regulated municipalities compare to the control group.

  - Because 2006 regulator and control were trending differently to begin with, comparing those two groups of municipalities gives the bad approximation.

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

```
##
## ================================================
##                         Dependent variable:
##                     ----------------------------
##                          particulate_matter
## ------------------------------------------------
## treat                        124.277***
##                               (1.456)
##
## Constant                      66.281***
##                               (0.421)
##
## ------------------------------------------------
## Observations                   11,990
## R2                             0.378
## Adjusted R2                    0.378
## Residual Std. Error      44.090 (df = 11988)
## F Statistic         7,282.370*** (df = 1; 11988)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

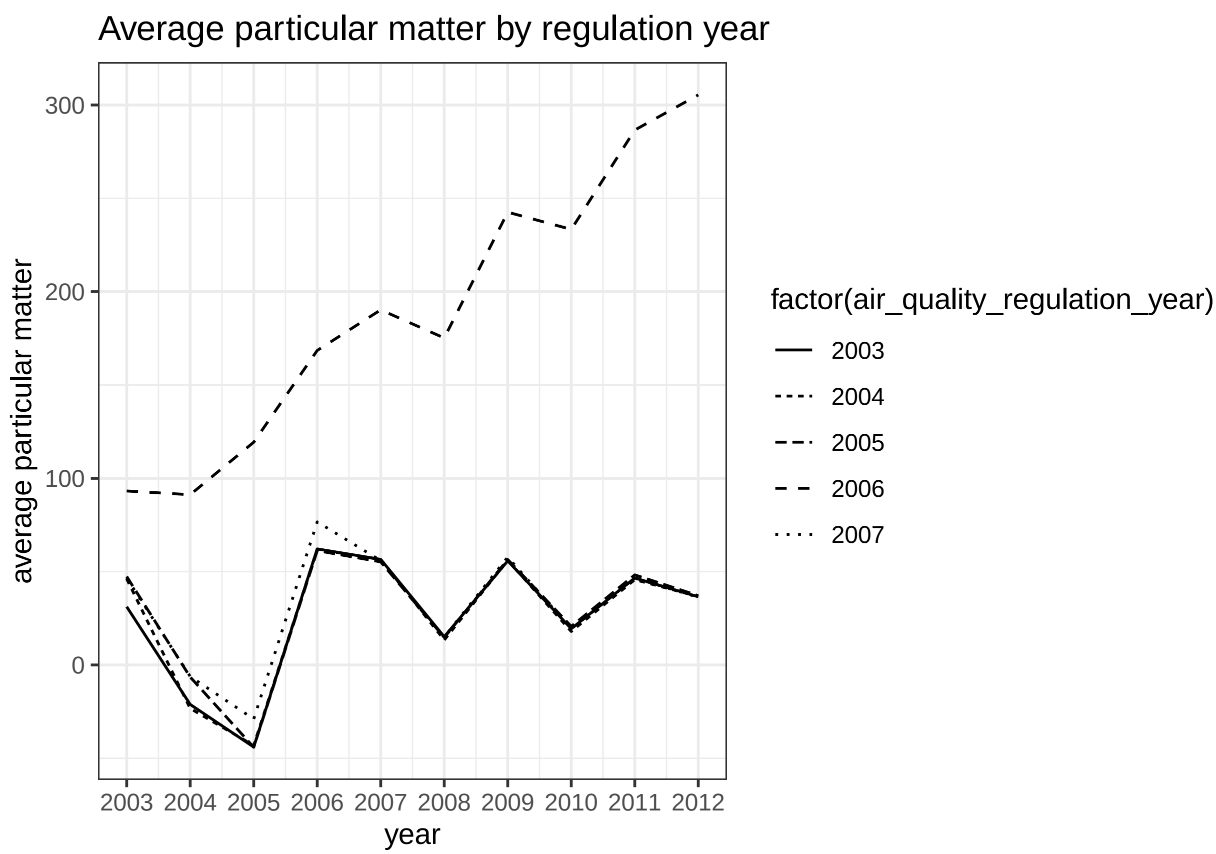
```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                         particulate_matter
```

```
## ----------------------------------------------
## treat
##                               (0.000)
##
## post
##                               (0.000)
##
## treat:post                    67.469*
##                              (35.513)
##
## ----------------------------------------------
## Observations                  11,990
## R2                             0.906
## Adjusted R2                    0.895
## Residual Std. Error     18.088 (df = 10781)
## ==============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

9.

```
## `summarise()` regrouping output by 'year' (override with `.groups` argument)
```




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

```
##
##            Estimate Std. Error t value  Pr(>|t|)
## treatment -14.83518    0.35979 -41.233 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- I dropped 2006 regulated dataset, because of the unparallel pre-trend.
  - If we have different trajectory to begin with, we would be overestimate or underestimate the treatment effect, because we have falsely assumed the counterfactual.
- After dropping 2006:
  - PM 2.5 is 14.83518 lower in average for regulated municipalities compare to the never regulated municipalities.

```
##
## t test of coefficients:
##
##                 Estimate Std. Error  t value  Pr(>|t|)
## delta_0        -11.92232    0.51755 -23.0361 < 2.2e-16 ***
## delta_before_2   0.64714    0.72800   0.8889  0.374069
## delta_before_3  -0.57481    0.98448  -0.5839  0.559323
## delta_before_4  -3.13029    0.95627  -3.2734  0.001067 **
## delta_after_1   -9.61189    0.41972 -22.9009 < 2.2e-16 ***
## delta_after_2   -8.54454    0.41235 -20.7215 < 2.2e-16 ***
## delta_after_3   -6.52717    0.36950 -17.6651 < 2.2e-16 ***
## delta_after_4   -5.10026    0.35141 -14.5137 < 2.2e-16 ***
## delta_after_5   -3.61716    0.34381 -10.5208 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Treatment effect do vary over time. The effect of regulation is the most effective on the regulated year and it diminishes in magnitude of the effect as time goes by.

10.

- Preferred estimating method for PROGRAMEVAL case is event study, to allow differential effects over time.
  - With event study we can line up treatment at the same time for everyunit, still use fixed effects to soak up confounders, get a partial test of the identifying assumption.
- Remaining shortcmoing with event study is that we still have threats to identification.
  - Selection into treatment: Treated and untreated individuals on different trends.
  - Coincident treatments: If a treated unit does two things at once, we can't cleanly estimate treatment.
  - Anticipatory effects and the Ashenfelter dip.
- I would like to recommend strongly promote air quality regulations, refering to the below event study results.
  - Although the effect of regulation is most effective on the regulated year and it diminishes as time goes by, but still, it is effective.
- Regulated year effect:

  - PM 2.5 is 11.92232 lower in average for regulated municipalities compare to the never regulated municipalities.
- Regulated year+1 effect:
  - PM 2.5 is 9.61189 lower in average for regulated municipalities compare to the never regulated municipalities.
- Regulated year+2 effect:
  - PM 2.5 is 8.54454 lower in average for regulated municipalities compare to the never regulated municipalities.
- Regulated year+3 effect:
  - PM 2.5 is 6.52717 lower in average for regulated municipalities compare to the never regulated municipalities.
- Regulated year+4 effect:
  - PM 2.5 is 5.10026 lower in average for regulated municipalities compare to the never regulated municipalities.
- Regulated year+5 effect:
  - PM 2.5 is 3.61716 lower in average for regulated municipalities compare to the never regulated municipalities.