Assignment 4

May 19, 2022

Contents

Background

Local air pollution is among the greatest threats to human health today. In recent years, China has embarked on a "war on pollution," using a variety of approaches to try to reduce pollution exposure. A new blue-ribbon commission, the Pollution Regulation Organization for Greater Regional Air Monitoring, Evaluation, Valuation, And Life (PROGRAMEVAL), has come to the Harris School to find a team of experts to help them understand the effectiveness of China's pollution regulations on air quality, but all the faculty are busy – so you've been asked to step in.

Question 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?)

Answer

To estimate the effect of air quality regulations on PM 2.5, ideally, we would have an experiment where we could randomly treat certain regions of the country with air quality regulations and leave the remaining regions unregulated. Ideally, we would want to have this on the level of cities, or even neighborhoods. In an unconstrained world, each neighborhood would be randomly assigned to these treatment and control groups. So, our unit of analysis would be a single neighborhood. For each neighborhood, we would want data on its treatment status and PM 2.5 levels before and after treatment. Assuming random assignment of treatment and full compliance within the experiment (all regulations would be fully implemented and enforced), we would be able to estimate the PM 2.5 difference between regulated and unregulated neighborhoods as:

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

Under the potential outcomes framework, we would want to estimate:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

But, we observe neighborhood i only once, and do not know its outcomes under both statuses simultaneously. In this experiment, randomization would allow us to estimate τ^{ATE} ensuring that:

$$\mathbb{E}[Y_i|D_i = 1] = \mathbb{E}[Y_i(1)]$$

$$\mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[Y_i(0)]$$

Question 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.

Answer

There are some problems with just looking at the difference between regulated and unregulated municipalities at a single point in time and try to infer the effect of these regulations on air quality. There might be some underlying reasons for why the regulated municiplaties were regulated in the first place. For instance, it might be that these were the provinces most affected by air pollution and this was why they were regulated. In this case, on average, regulated municipalities could still be in a worse condition than unregulated ones, but this would not indicate that regulations were harmful. Furthermore, it might be that the municipalities who implemented such regulations were also environmentally cautious in other ways. This kind of selection bias could mean that regulated municipalities had better air quality before the specific regulation was implemented. This could lead us to overestimate the impact of said regulations when we look at a single temporal snapshot of air quality.

Question 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.

Answer

If we have data on regulated municipalities at multiple points in time, we could include this time-series estimation and look whether, and to what extent, there is a significant change at 2004 to estimate the impact of air quality regulations. In other words, such time series data would allow us to compare the municipality i to itself before and after treatment. Now we could estimate the following by looking at the averages before and after 2004 (treatment):

$$\hat{\tau}^{TS} = \overline{Y}_{t \in post} - \overline{Y}_{t \in pre}$$

For this to recover the true τ, we need to assume that there are no time-varying characteristics. In other words, our pre-treatment ourcome must be a good guess of what our potential outcome would be absent treatment in the post-treatment period. So we need to assume that the counterfactual trend is zero.

We can come up with some examples of why this assumption might not hold. For instance, if the Chinese government implemented country-level regulations to mitigate air pollution around the time that municipal regulations took effect, this time-varying characteristic would also affect PM 2.5 levels in regulated municipalities, possibly leading to a decline even at the absence of the municipal level regulations. This would make it difficult to isolate the effect of the regulations of interest. Secondly, perhaps the regulated municipalities had agreements with industrial manufacturers to expand industrial output in these municipalities in 2004. In this case, absent regulations, the additional industrial activity could increase the PM 2.5 levels despite regulations. Finally, the regulated municipalities might have adopted more environmental regulations in 2004 than just the one of interest, which would make it difficult to isolate the effect of the regulation we want to estimate the effect of.

Question 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.

Answer

If we have data on many municipalities at multiple points in time including periods both before and after imposing of air quality regulations, we can come closer to estimating effect of these regulations. We can measure the difference between regulated and unregulated municipalities before regulations. This could give us the already existing difference between the two groups of municipalities. Later, by comparing this (pre) difference to the (post) difference in PM 2.5 after 2004, we could estimate the effect of regulations of PM 2.5. In other words, we would do both across-unit, within time comparisons, *and* within-unit, across time comparisons. This would be equivalent to the following expression:

$$\begin{aligned}\hat{\tau}^{DD} &= (Y(D_i = 1, post) - Y(D_i = 1, pre)) - (Y(D_i = 0, post) - Y(D_i = 0, pre)) \\ &= (\overline{Y}(treat, post) - \overline{Y}(treat, pre)) - (\overline{Y}(untreat, post) - \overline{Y}(untreat, pre))\end{aligned}$$

We could estimate this via the following regression:

$$Y_{it} = \alpha + \tau Treat \times Post_{it} + \beta Treat_i + \delta Post_t + \epsilon_{it}$$

Where $Treat_i = 1$ if the municipality i is ever treated and $Post = 1$ if period t is after treatment. This regression would give us $\hat{\tau} = \hat{\tau}^{DD}$.

For this approach to recover the causal effect of air quality regulations on PM 2.5, we would need to make the identifying assumption of *parallel trends*. In other words, the PM 2.5 trends before 2004 would need to be parallel between the regulated and unregulated municipalities. If this assumption is invalid, it means that there probably are some fundamental differences between the two groups that cannot be captured with our model. We can think of several reasons why this might be the case. For instance, it might be that treated municipalities were treated due to rapidly increasing PM 2.5 levels because of the increasing number of manufacturing plants in these regions. If such an increase is not happening in unregulated municipalities, we cannot assume parallel trends. Secondly, it might be that, besides these 2004 regulations, regulated municipalities have started imposing other environmental regulations that can impact air quality because they are more environmentally conscious. If this is the case, we might see declining pre-trends in the treated group as opposed to untreated group, which would be problematic for our difference-in-difference estimation.

Question 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.

Answer

Having data on regulated municipalities from 2003 to 2007, we would want to get the difference between average PM 2.5 levels in 2003 and average PM 2.5 levels from 2004 to 2007. We could estimate this with the following regression:

$$Y_{it} = \alpha_{it} + \tau \times \mathbb{I}(t \geq 2004) + \epsilon_{it}$$

Assuming there are no time-varying characteristics, this would give us the average treatment effect $\hat{\tau}^{TS}$.

Question 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?

Answer

```
air_df <- read_csv("ps_4_data.csv", col_types = cols(year = col_integer(),
    municipality_id = col_double(), air_quality_regulation_year = col_integer(),
    particulate_matter = col_double())) %>%
    mutate(treat = as.integer(!is.na(air_quality_regulation_year)))
```

```
air_df %>%
    group_by(treat) %>%
    summarise(`Mean PM 2.5` = mean(particulate_matter)) %>%
    kable()
```

treat	Mean PM 2.5
0	66.28139
1	41.83282

```
stargazer(lm(particulate_matter ~ treat, data = air_df), type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, May 17, 2022 - 19:02:04

Here we see that the average level of PM 2.5 for the treated group is almost 24.5 units less than the untreated group. Although, this simple comparison does not look at differentiate between pre- and post-treatment figures.

Table 2:

	Dependent variable:
	particulate_matter
treat	−24.449***
	(0.699)
Constant	66.281***
	(0.494)
Observations	22,000
R^2	0.053
Adjusted R^2	0.053
Residual Std. Error	51.815 (df = 21998)
F Statistic	1,224.499*** (df = 1; 21998)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
air_df %>%
    filter(air_quality_regulation_year == 2004) %>%
    mutate(treat = as.integer(year >= 2004)) %>%
    lm(particulate_matter ~ treat, data = .) %>%
    stargazer(type = "latex")
```

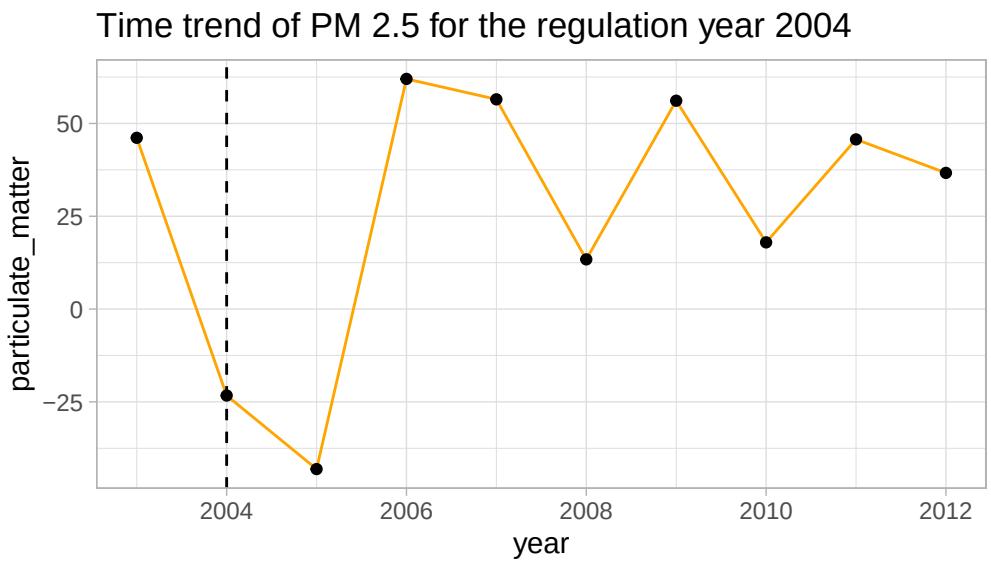
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, May 17, 2022 - 19:02:04

Table 3:

	Dependent variable:
	particulate_matter
treat	−21.470***
	(3.587)
Constant	46.132***
	(3.403)
Observations	1,000
R^2	0.035
Adjusted R^2	0.034
Residual Std. Error	34.030 (df = 998)
F Statistic	35.826*** (df = 1; 998)
Note:	*p<0.1; **p<0.05; ***p<0.01

Unlike the previous estimate, here we estimate the average difference before and after the treatment year 2004 for all municipalities who got treated in 2004. In other words, while the previous estimate was the difference across treatment groups, this is the difference within the treated group across the years.

```r
air_df %>%
    filter(air_quality_regulation_year == 2004) %>%
    group_by(year) %>%
    summarise(particulate_matter = mean(particulate_matter)) %>%
    ggplot(aes(x = year, y = particulate_matter)) + geom_line(color = "orange") +
    scale_x_continuous(breaks = pretty_breaks()) + geom_vline(aes(xintercept = 2004),
    linetype = "dashed") + geom_point() + labs(title = "Time trend of PM 2.5 for the regulation year 20
    theme_light()
```



We see a rapid decline in PM 2.5 in the two years following the regulations (2004 and 2005), later on, it seems to go back to pre-treatment levels on average. Our estimate above included all years available to us after the treatment, therefore we get an average effect which is lower than the immediate short term change we see right after treatment.

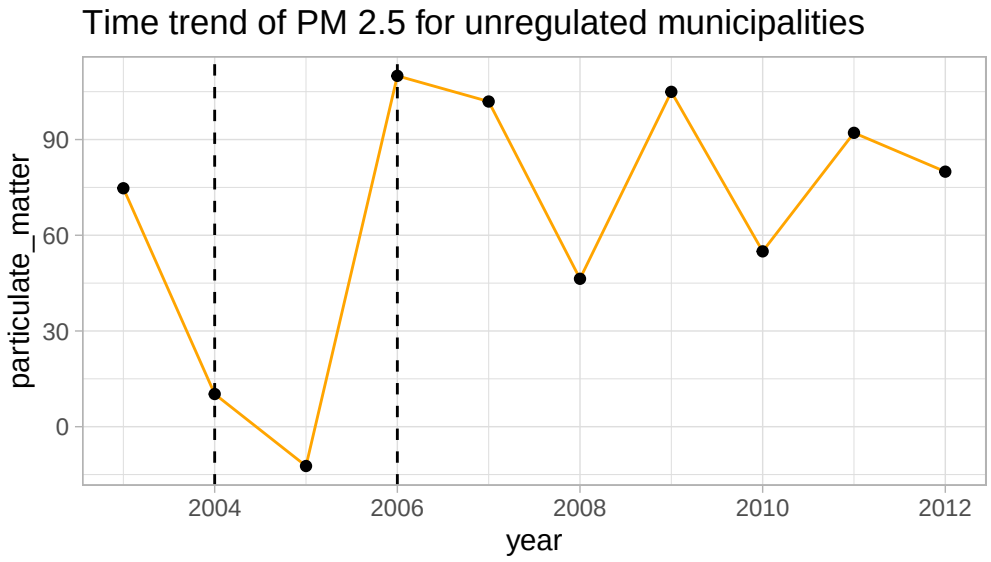
Question 7

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.

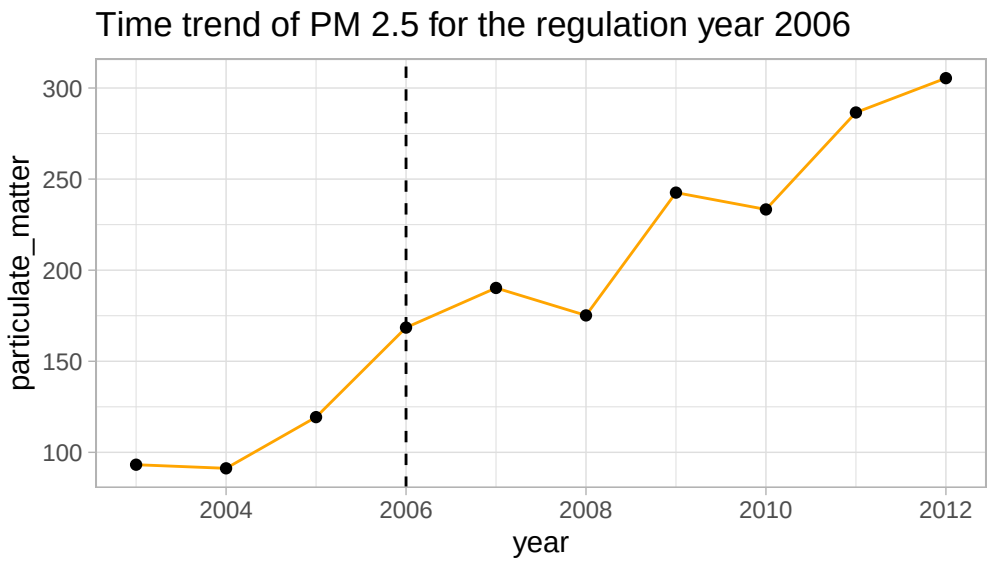
Answer

```r
air_df %>%
    filter(!treat) %>%
    group_by(year) %>%
    summarise(particulate_matter = mean(particulate_matter)) %>%
    ggplot(aes(x = year, y = particulate_matter)) + geom_line(color = "orange") +
    scale_x_continuous(breaks = pretty_breaks()) + geom_vline(aes(xintercept = 2004),
    linetype = "dashed") + geom_vline(aes(xintercept = 2006),
```

```
    linetype = "dashed") + geom_point() + labs(title = "Time trend of PM 2.5 for unregulated municipali
    theme_light()
```




```
air_df %>%
    filter(air_quality_regulation_year == 2006) %>%
    group_by(year) %>%
    summarise(particulate_matter = mean(particulate_matter)) %>%
    ggplot(aes(x = year, y = particulate_matter)) + geom_line(color = "orange") +
    scale_x_continuous(breaks = pretty_breaks()) + geom_vline(aes(xintercept = 2006),
    linetype = "dashed") + geom_point() + labs(title = "Time trend of PM 2.5 for the regulation year 20
    theme_light()
```



Since we only have one period before treatment in the case of municipalities who were regulated in 2004, the time series plot is not very convincing in terms of parallel trends assumption. Yet, we do not see any

evidence against parallel trends either. There is parallel trends after the treatment period, but this would not be indicative of this group's viability to act as controls, since it is post-treatment outcome. The municipalities that were regulated in 2006, on the other hand, clearly displays nonparallel trends with the potential control group. Therefore it is not fit to be taken as a control group in a diff-in-diff design with the 2006 group.

Question 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).

Answer

```
diff_tab <- air_df %>%
    filter(air_quality_regulation_year %in% c(2006, NA)) %>%
    mutate(post = ifelse(year >= 2006, "Post", "Pre"), Status = ifelse(treat,
        "Treated", "Untreated")) %>%
    group_by(Status, post) %>%
    summarise(mean_pm = mean(particulate_matter)) %>%
    pivot_wider(names_from = post, values_from = mean_pm)

kable(diff_tab, digits = 2)
```

Status	Post	Pre
Treated	228.83	101.27
Untreated	84.31	24.22

```
tau <- as.double((diff_tab[1, 2] - diff_tab[1, 3]) - (diff_tab[2,
    2] - diff_tab[2, 3]))
cat("Estimated tau =", tau)
```

```
Estimated tau = 67.46895
```

```
ols_did <- air_df %>%
    filter(air_quality_regulation_year %in% c(2006, NA)) %>%
    mutate(post = ifelse(year >= 2006, 1, 0)) %>%
    lm(particulate_matter ~ treat * post, data = .)

fe_did <- air_df %>%
    filter(air_quality_regulation_year %in% c(2006, NA)) %>%
    mutate(post = ifelse(year >= 2006, 1, 0)) %>%
    felm(particulate_matter ~ treat:post | municipality_id +
        year | 0 | municipality_id, data = .)

stargazer(ols_did, fe_did, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, May 17, 2022 - 19:02:04

Table 5:

	Dependent variable:	
	particulate_matter	
	OLS	*felm*
	(1)	(2)
treat	77.049***	
	(1.872)	
post	60.093***	
	(0.646)	
treat:post	67.469***	67.469***
	(2.238)	(0.545)
Constant	24.216***	
	(0.541)	
Observations	11,990	11,990
R^2	0.692	0.906
Adjusted R^2	0.692	0.895
Residual Std. Error	31.046 (df = 11986)	18.088 (df = 10781)
F Statistic	8,959.740*** (df = 3; 11986)	
Note:		*p<0.1; **p<0.05; ***p<0.01

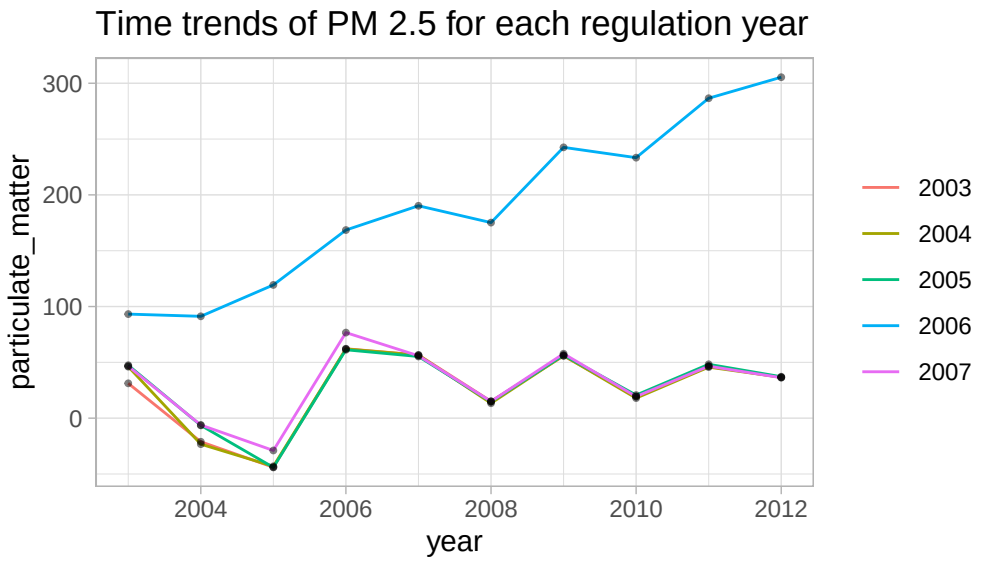
Our estimates for τ, do not change across the different methods we tried. Based on diff-in-diff findings, the causal effect of regulations on PM 2.5 emissions is a positive 67.469. This indicates that regulations have a *negative* impact on air quality, which is opposite of what we would expect. In (6) and (7), we compared the same treated group across years to estimate the causal effect of regulations, here we take the difference between the differences treatment and control groups had after treatment was administered (2006 regulations). However, the previous plots showed us that this untreated group and the group treated in 2006 do not share parallel trends, which would render these results invalid as this goes against our identifying assumption.

Question 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 (9)? 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.

Answer

```
air_df %>%
    filter(air_quality_regulation_year %in% c(2003:2007)) %>%
    group_by(year, air_quality_regulation_year) %>%
    mutate(air_quality_regulation_year = as.factor(air_quality_regulation_year)) %>%
    summarise(particulate_matter = mean(particulate_matter)) %>%
    ggplot(aes(x = year, y = particulate_matter)) + geom_line(aes(col = air_quality_regulation_year)) +
    scale_x_continuous(breaks = pretty_breaks()) + geom_point(size = 0.7,
    alpha = 0.5) + theme_light() + labs(title = "Time trends of PM 2.5 for each regulation year") +
    theme(legend.title = element_blank())
```



We cannot estimate a credible causal effect for municipalities that were regulated in 2006 because they do not display trends similar to those of unregulated municipalities, which violates our identifying assumption of parallel trends. There might be other characteristic differences that differentiate these municipalities from the unregulated ones, which do not make them a suitable counterfactual while estimating cuasal effects.

```
multi_fe_did_all <- air_df %>%
    filter(air_quality_regulation_year != 2006) %>%
    mutate(treat_post = ifelse(year >= air_quality_regulation_year,
        1, 0), treat_post = ifelse(is.na(treat_post), 0, treat_post)) %>%
    felm(particulate_matter ~ treat_post | municipality_id +
        year | 0 | municipality_id, data = .)

multi_fe_did_2003_2007 <- air_df %>%
    filter(air_quality_regulation_year %in% c(2003, 2004, 2005,
        2007, NA)) %>%
    mutate(treat_post = ifelse(year >= air_quality_regulation_year,
        1, 0), treat_post = ifelse(is.na(treat_post), 0, treat_post)) %>%
    felm(particulate_matter ~ treat_post | municipality_id +
        year | 0 | municipality_id, data = .)

stargazer(multi_fe_did_all, multi_fe_did_2003_2007, type = "latex",
    column.labels = c("With 2002", "Without 2002"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, May 17, 2022 - 19:02:05

Table 6:

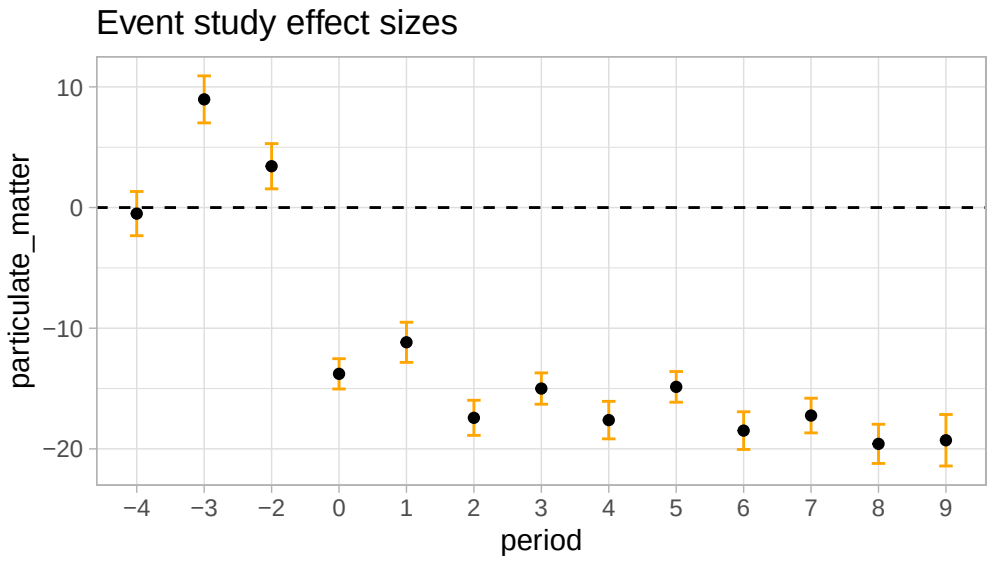
	Dependent variable:	
	particulate_matter	
	With 2002	Without 2002
	(1)	(2)
treat_post	−14.835***	−16.910***
	(0.360)	(0.376)
Observations	10,010	15,000
R^2	0.958	0.952
Adjusted R^2	0.953	0.946
Residual Std. Error	7.308 (df = 8999)	9.775 (df = 13490)
Note:		*p<0.1; **p<0.05; ***p<0.01

These results are very different from what we saw in (8). Now that we eliminated the year 2006 which displayed nonparallel trends, our estimation's sign became negative as we would expect. Based on whether we include the municipalities that got regulated in 2002 or not, the estimated causal effect changes from -14.835 to -16.910. In (8) we had estimated a negative effect which was unreliable due to nonparallel trends between the control and treatment groups.

```
event_df <- air_df %>%
    filter(air_quality_regulation_year %in% c(2003, 2004, 2005,
        2007, NA)) %>%
    mutate(year_index = ifelse(is.na(air_quality_regulation_year),
        0, year - air_quality_regulation_year), year_index = factor(year_index,
        levels = c(-1, -5:-2, 0:9)))

event_reg <- felm(particulate_matter ~ year_index | year + municipality_id |
    0 | municipality_id, data = event_df)
event_reg$coefficients
```

```
             particulate_matter
year_index-4         -0.5020447
year_index-3          8.9659943
year_index-2          3.4257054
year_index0         -13.7812545
year_index1         -11.1675313
year_index2         -17.4265682
year_index3         -15.0011507
year_index4         -17.6166307
year_index5         -14.8615917
year_index6         -18.4873463
year_index7         -17.2377823
year_index8         -19.5848909
year_index9         -19.2859685
```

```r
event_study_results <- rownames_to_column(data.frame(event_reg$coefficients),
    "period") %>%
    mutate(period = factor(substr(period, 11, 12), levels = -4:9))

event_study_results["lower"] = confint(event_reg)[, 1]
event_study_results["upper"] = confint(event_reg)[, 2]

ggplot(event_study_results, aes(x = period, y = particulate_matter)) +
    geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.2,
        color = "orange") + geom_point() + geom_hline(aes(yintercept = 0),
    linetype = "dashed") + labs(title = "Event study effect sizes") +
    theme_light()
```



Treatment effect is supposed to be centered around 0 for periods before treatment. However, for periods before treatment we see a positive treatment effect which raises questions about our identifying assumption. For periods after treatment, we see consistent negative treatment effects for all 9 years we have available in the dataset. The causal effect of regulations seem to be between -20 and -10 PM 2.5 units after the second year of regulations.

Question 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?

Answer

When we include only years with parallel trends into the study, we see a negative causal relationship between regulations and PM 2.5 levels. My preferred figure would come from the fixed effects model without the

year 2006. This model estimated that we can attribute -14.835 units decrease in PM 2.5 levels to municipality regulations. One shortcoming of such an approach is that we do not observe for sure that whether unregulated municipalities would respond exactly the same way to regulations as the regulated industries, or whether something else changed together with the imposing of regulations that could confound our analysis. Nevertheless, due to consistent results with and significance of the estimated magnitude, I believe we have enough reason to believe the effectiveness of air quality regulations and I would advise PROGRAMEVAL that they should be promoting them.