PE Pset 2

Nan

2025-05-12

1.Ideal Experimental Design: RCT

In a fully unconstrained world, the ideal experimental design would be a randomized controlled trial (RCT) at the city level. A large, representative sample of cities would be randomly assigned into two groups: a treatment group, which implements a specific air quality regulation (e.g., restrictions on industrial emissions or vehicle usage), and a control group, which continues business as usual without new regulations.

To ensure balance, stratified randomization could be used based on baseline pollution levels, population size, and industrial activity. The regulation would be implemented in the treatment group simultaneously and uniformly, with strict compliance monitoring.

The primary outcome would be PM2.5 concentrations, measured daily via standardized air quality sensors across all cities. The evaluation period would span at least one year to account for seasonal variations.

To estimate the Average Treatment Effect (ATE), we would use the difference-in-means estimator:

$$ATE = \mathbb{E}[Y_1 - Y_0]$$

where Y_1 is the average PM2.5 level in treated cities and Y_0 in control cities post-intervention. With randomization ensuring exogeneity, the difference directly captures the causal effect of regulation.

This RCT eliminates confounding, allowing for a clean interpretation of the regulation's impact on air pollution. In practice, such a design faces ethical, political, and logistical barriers, but it offers the clearest causal estimate in theory.

2.Interrupted Time Series (ITS) Design

To estimate the causal effect of PPHA adoption in a single treated city, we can use an interrupted time series (ITS) design. This approach leverages the longitudinal nature of the data, comparing pollution levels before and after the policy change in 1997.

Model Specification (Regression Form)

Let t index time in years from 1990 to 2005, and let Y_t be the average PM2.5 level in the flagship city in year t. Define a treatment indicator $D_t = 1$ if $t \geq 1997$, and 0 otherwise.

We estimate:

$$Y_t = \alpha + \beta \cdot D_t + \gamma \cdot t + \epsilon_t$$

Where: - α: baseline pollution level in 1990 - γ: linear time trend - β: TS estimator for the causal effect of PPHA on PM2.5 - ϵ_t: error term

Interpretation

The coefficient β captures the discrete change in pollution levels after the implementation of PPHA in 1997, controlling for any linear trend over time.

Identifying Assumption

The key assumption is that, absent the policy, the pollution trend in the flagship city would have continued as a smooth function of time (i.e. counterfactual trend is typically linear). This is the parallel trends assumption in time.

Concern with This Approach

One concern is confounding from concurrent shocks. For example, if a nationwide environmental campaign or economic slowdown began around 1997, it could also reduce pollution and be wrongly attributed to PPHA, biasing β upward.

3.Difference-in-Differences (DiD)

Having data on multiple cities, including both cities that adopted PPHA in 1997 and those that never did, enables a difference-in-differences (DiD) approach. This improves causal inference by introducing a control group that helps account for time trends and shocks affecting all cities, not just the treated one.

Model Specification (Regression Form)

Let i index cities and t index years. Define: - Y_{it}: pollution level in city i at time t - Treat$_i = 1$ if city i adopted PPHA n 1997, 0 otherwise - Post$_t = 1$ if year $t \geq 1997$, 0 otherwise

We estimate:

$$Y_{it} = \alpha + \beta \cdot (\text{Treat}_i \times \text{Post}_t) + \lambda_i + \delta_t + \epsilon_{it}$$

Where: - λ_i: city fixed effects (controls for time-invariant city characteristics) - δ_t: year fixed effects (controls for common shocks) - β: DiD estimator for the causal effect of PPHA on PM2.5

Identifying Assumption

The parallel trends assumption: in the absence of PPHA, treated and control cities would have followed similar pollution trends over time.

Remaining Concern

A potential concern is differential trends—if treated cities were already on a steeper pollution decline before 1997 (e.g., due to prior environmental investments), the DiD estimate β may overstate the effect of PPHA.

4.Event Study

With full panel data covering all cities from 1990–2005, and variation in the timing of PPHA adoption, we can implement an event study or generalized difference-in-differences (DiD) design to flexibly estimate the dynamic causal effect of PPHA on pollution over time.

Model Specification (Regression Form)

Let i index cities and t index years. Define: - Y_{it}: pollution level in city i at time t - $D_{it}^k = 1$ if year t is k years relative to PPHA adoption in city i(e.g., $k = -1$ is one year before adoption) - Omit $k = -1$ to serve as the reference period

We estimate:

$$Y_{it} = \alpha + \sum_{k \neq -1} \beta_k \cdot D_{it}^k + \lambda_i + \delta_t + \epsilon_{it}$$

Where: - λ_i: city fixed effects(same as DiD) - δ_t: year fixed effects(same as DiD) - β_k: event-time coefficients estimating the effect of PPHA k years after (or before) adoption

Identifying Assumption

The parallel trends assumption must hold: in the absence of treatment, treated cities would have followed similar pollution trends as untreated cities. The pre-treatment coefficients (β_k for $k < 0$) also help partially test this assumption.

Remaining Concern

A key concern is selection bias caused by treatment effect heterogeneity over time—if cities adopting early (e.g., in 1997) are systematically different from those adopting later (e.g., 2001), the estimated dynamic effects may reflect both timing and selection differences, biasing interpretation.

5.Simple Comparison

```
df <- read_csv("~/Desktop/Program Evaluation/PS2/ps2_data_25.csv")
df <- df %>%
  mutate(ever_ppha = ifelse(is.na(ppha_year), 0, 1))
df_summary <- df %>%
  group_by(ever_ppha) %>%
  summarise(avg_pollution = mean(pollution_pm, na.rm = TRUE)) %>%
  mutate(group = ifelse(ever_ppha == 1, "Eventually Passed PPHA", "Never Passed PPHA"))

print(df_summary)
```

```
## # A tibble: 2 x 3
##   ever_ppha avg_pollution group
##       <dbl>         <dbl> <chr>
## 1         0         100.  Never Passed PPHA
## 2         1          99.3 Eventually Passed PPHA
```

Using the provided dataset, we compute average PM2.5 pollution levels across two groups: cities that never passed the PPHA and those that eventually did. The results indicate that cities which eventually adopted the PPHA had lower average pollution levels (99.29 ug/m³) compared to cities that never adopted the policy (99.99 ug/m³). While this suggests a potential relationship between PPHA adoption and improved air quality, this simple comparison does not account for pre-existing differences, time trends, or confounding factors. More rigorous causal inference methods are required to draw reliable conclusions.

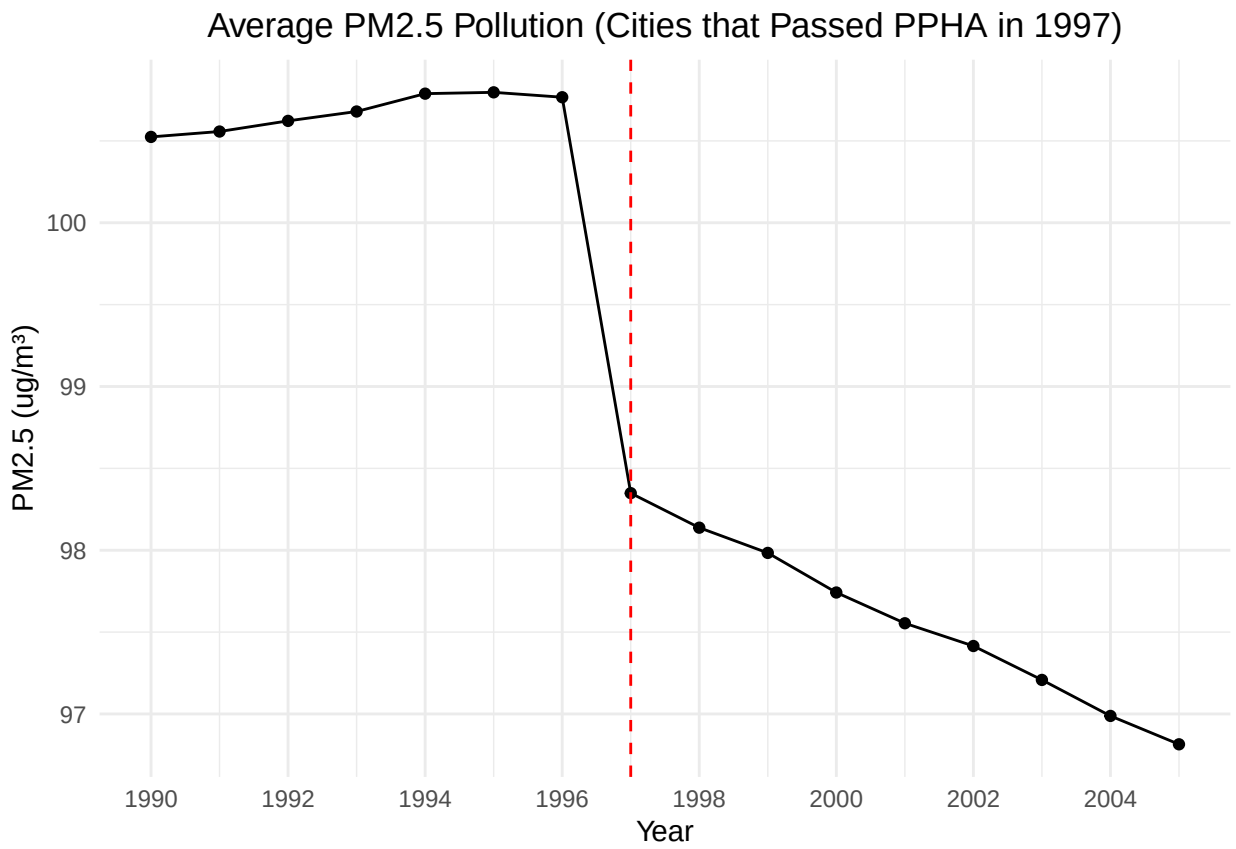
6.Time-Series Analysis

```
df_1997 <- df %>%
  filter(ppha_year == 1997) %>%
  mutate(post = ifelse(year >= 1997, 1, 0))

model1 <- lm(pollution_pm ~ post + year, data = df_1997)
summary(model1)
```

```
##
## Call:
## lm(formula = pollution_pm ~ post + year, data = df_1997)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3615 -0.2156 -0.0094  0.2253  3.3633
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 329.202635   3.068184  107.30   <2e-16 ***
## post         -2.182319   0.014305 -152.55   <2e-16 ***
## year         -0.114664   0.001539  -74.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2888 on 6397 degrees of freedom
## Multiple R-squared:  0.9669, Adjusted R-squared:  0.9669
## F-statistic: 9.347e+04 on 2 and 6397 DF,  p-value: < 2.2e-16
```

Using only cities that passed the PPHA in 1997, we estimate a time-series regression of pollution levels on a post-policy indicator and year trend. The regression shows a small but statistically significant, negative coefficient on the post-1997 indicator, suggesting a modest reduction in pollution following policy implementation. However, the time trend is also downward, indicating general improvement over time.

```
avg_by_year <- df_1997 %>%
  group_by(year) %>%
  summarise(avg_pm25 = mean(pollution_pm, na.rm = TRUE))

ggplot(avg_by_year, aes(x = year, y = avg_pm25)) +
  geom_line() +
  geom_point() +
  geom_vline(xintercept = 1997, linetype = "dashed", color = "red") +
  scale_x_continuous(breaks = seq(min(avg_by_year$year), max(avg_by_year$year), by = 2)) +
  labs(
    title = "Average PM2.5 Pollution (Cities that Passed PPHA in 1997)",
    x = "Year",
    y = "PM2.5 (ug/m³)"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

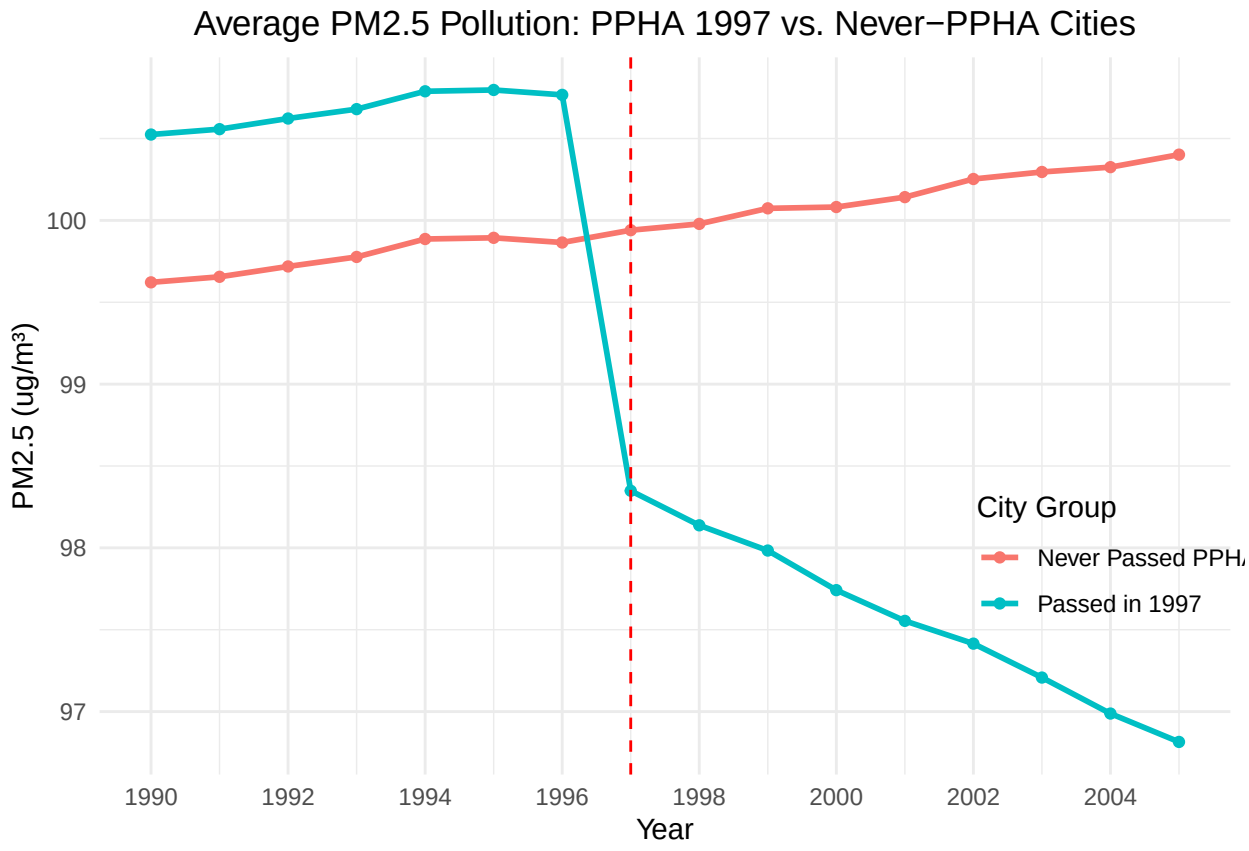


The plotted average pollution levels reveal a sharp drop in 1997, when cities adopted PPHA. Pollution levels had been relatively flat before the policy, but declined steadily thereafter. This challenges the plausibility of the parallel trends assumption in time-series analysis: absent the policy, the pollution trend would have continued as a smooth function of time. As a result, our regression estimate may plausibly capture a real policy effect. However, since we lack a control group, we cannot rule out the possibility of other confounding factors driving this change. Additional designs (e.g., DiD with untreated cities) are needed for stronger causal claims.

7.Viability of Using the Never-PPHA Cities as a Control Group

```
# Classify cities
df <- df %>%
  mutate(
    ppha_year = as.numeric(ppha_year),
    group = case_when(
      is.na(ppha_year) ~ "Never Passed PPHA",
      ppha_year == 1997 ~ "Passed in 1997",
      TRUE ~ NA_character_
    )
  ) %>%
  filter(group %in% c("Never Passed PPHA", "Passed in 1997"))

# Calculate Average Pollution by Year and Group
avg_pollution <- df %>%
  group_by(year, group) %>%
```

```r
  summarise(avg_pm25 = mean(pollution_pm, na.rm = TRUE), .groups = "drop")

# Plot PM2.5 Over Time
ggplot(avg_pollution, aes(x = year, y = avg_pm25, color = group)) +
  geom_line(size = 1) +
  geom_point() +
  geom_vline(xintercept = 1997, linetype = "dashed", color = "red") +
  scale_x_continuous(breaks = seq(min(avg_pollution$year), max(avg_pollution$year), by = 2)) +
  labs(
    title = "Average PM2.5 Pollution: PPHA 1997 vs. Never-PPHA Cities",
    x = "Year",
    y = "PM2.5 (ug/m³)",
    color = "City Group"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = c(0.90, 0.30)
  )
```



The figure plots average PM2.5 levels over time for two groups: cities that never passed PPHA and those that passed it in 1997. Prior to 1997, the trends in pollution levels for both groups are broadly similar—both increase slightly over time—suggesting that the parallel trends assumption may be reasonable in the pre-treatment period. This strengthens the case for using never-PPHA cities as a control group in a DiD design.

Still, one concern is that treated cities may differ systematically from untreated ones in ways that also affect

pollution (e.g., industrial profile, policy ambition). To ensure credible causal inference, we would ideally include fixed effects.

8.DiD

Difference-in-Means

```
df_post <- df %>%
  filter(group %in% c("Never Passed PPHA", "Passed in 1997")) %>%
  mutate(
    treated = ifelse(group == "Passed in 1997" & year >= 1997, 1, 0)
  ) %>%
  filter(year >= 1997)

mean_diff <- df_post %>%
  group_by(treated) %>%
  summarise(avg_pollution = mean(pollution_pm, na.rm = TRUE))

diff_in_means <- diff(mean_diff$avg_pollution)
print(paste("Difference in Means =", round(diff_in_means, 4)))
```

```
## [1] "Difference in Means = -2.5889"
```

Regression without Fixed Effects

```
df <- df %>%
filter(group %in% c("Never Passed PPHA", "Passed in 1997")) %>%
mutate(
treated = ifelse(group == "Passed in 1997" & year >= 1997, 1, 0)
)

model_without_fe <- lm(pollution_pm ~ treated, data = df)
summary(model_without_fe)
```

```
##
## Call:
## lm(formula = pollution_pm ~ treated, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91946 -0.24922 -0.04746  0.23462  2.90465
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 100.075309   0.002287 43749.0   <2e-16 ***
## treated      -2.498455   0.006286  -397.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3513 on 27182 degrees of freedom
## Multiple R-squared:  0.8532, Adjusted R-squared:  0.8532
## F-statistic: 1.58e+05 on 1 and 27182 DF,  p-value: < 2.2e-16
```

Regression with Fixed Effects

```
library(fixest)

model_fe <- feols(pollution_pm ~ treated | city_id + year, data = df)
summary(model_fe)
```

```
## OLS estimation, Dep. Var.: pollution_pm
## Observations: 27,184
## Fixed-effects: city_id: 1,699,  year: 16
## Standard-errors: Clustered (city_id)
##         Estimate Std. Error  t value  Pr(>|t|)
## treated -3.49141   0.008774 -397.944 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.209849     Adj. R2: 0.944087
##                  Within R2: 0.924598
```

```
results <- tibble(
  Method = c("Difference in Means", "Regression (no FE)", "Regression (with FE)"),
  Estimate = c(round(diff_in_means, 4),
               round(coef(model_without_fe)["treated"], 4),
               round(coef(model_fe)["treated"], 4))
)
print(results)
```

```
## # A tibble: 3 x 2
##   Method               Estimate
##   <chr>                   <dbl>
## 1 Difference in Means     -2.59
## 2 Regression (no FE)      -2.50
## 3 Regression (with FE)    -3.49
```

We estimate the effect of PPHA on pollution using three approaches. The difference in means estimate shows that post-1997, pollution levels were 2.59 ug/m³ lower in treated cities compared to never-treated cities. The regression without fixed effects produces a similar estimate of −2.50 ug/m³.

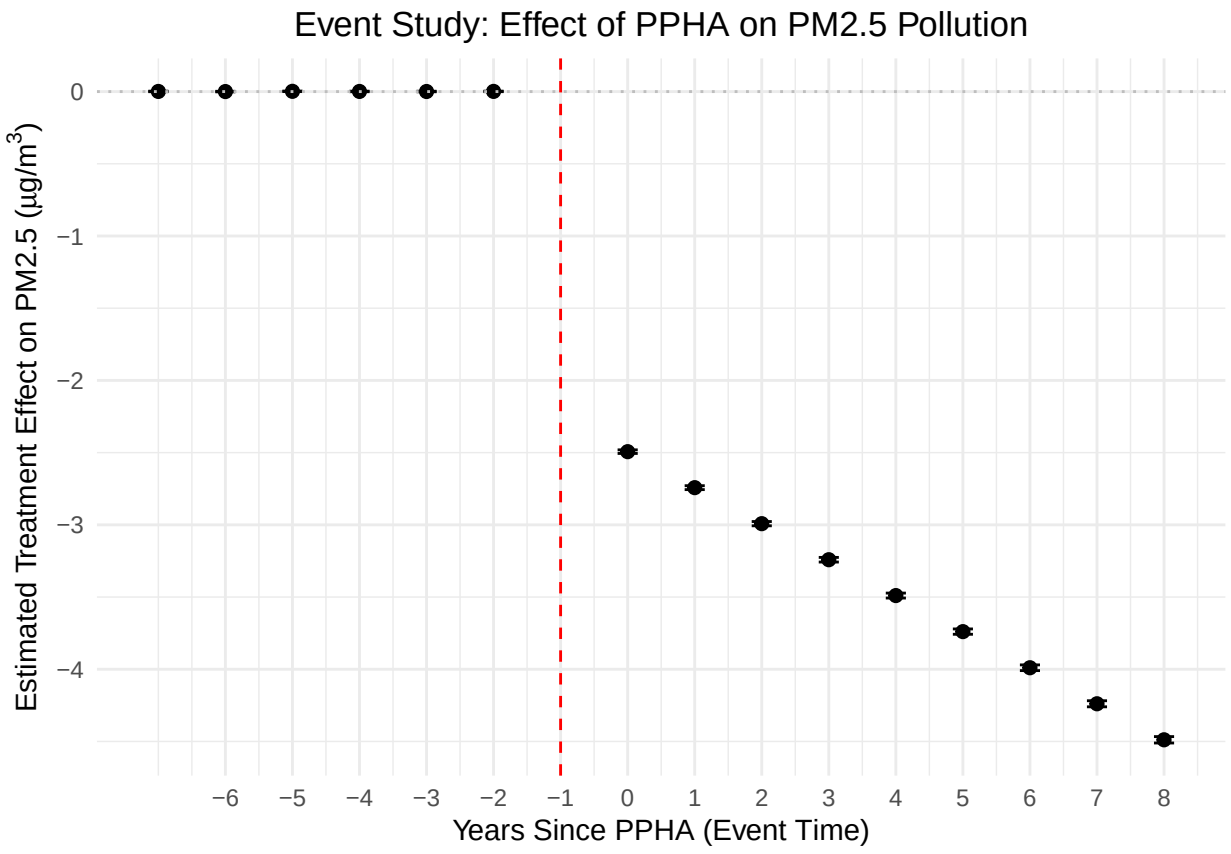
However, the regression with fixed effects (controlling for city and year) yields a larger estimate of −3.49 ug/m³, indicating a stronger treatment effect after accounting for time-invariant city characteristics and common shocks across years. This suggests that the raw comparisons may have understated the effect due to unobserved confounders that biased estimates toward zero.

The fixed effects model is the most credible specification, as it controls for both baseline differences across cities and nationwide trends. The coefficient of −3.49 implies that the adoption of PPHA in 1997 led to an average reduction of 3.49 ug/m³ in PM2.5 levels. This represents a meaningful causal effect of regulation on air quality.

9.Event Study

```
df_event <- df %>%
  mutate(
    ever_treated = ifelse(!is.na(ppha_year),1,0),
    event_time = ifelse(!is.na(ppha_year),year-ppha_year, 0))

model_event <- feols(pollution_pm ~ i(event_time, ever_treated, ref=-1) | city_id + year, cluster = ~ci
```

```r
event_effects <- broom::tidy(model_event) %>%
  filter(grepl("^event_time::", term)) %>%
  mutate(
    event_time = as.numeric(gsub("event_time::", "", gsub(":ever_treated", "", term))),
    ci_lower = estimate - 1.96 * std.error,
    ci_upper = estimate + 1.96 * std.error
  )

ggplot(event_effects, aes(x = event_time, y = estimate)) +
  geom_point(size = 2) +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), width = 0.3) +
  geom_vline(xintercept = -1, linetype = "dashed", color = "red") +
  geom_hline(yintercept = 0, linetype = "dotted", color = "gray") +
  scale_x_continuous(breaks = seq(-6, 8, 1)) +
  labs(
    title = "Event Study: Effect of PPHA on PM2.5 Pollution",
    x = "Years Since PPHA (Event Time)",
    y = expression("Estimated Treatment Effect on PM2.5 (" * mu * "g/m"^3 * ")")
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



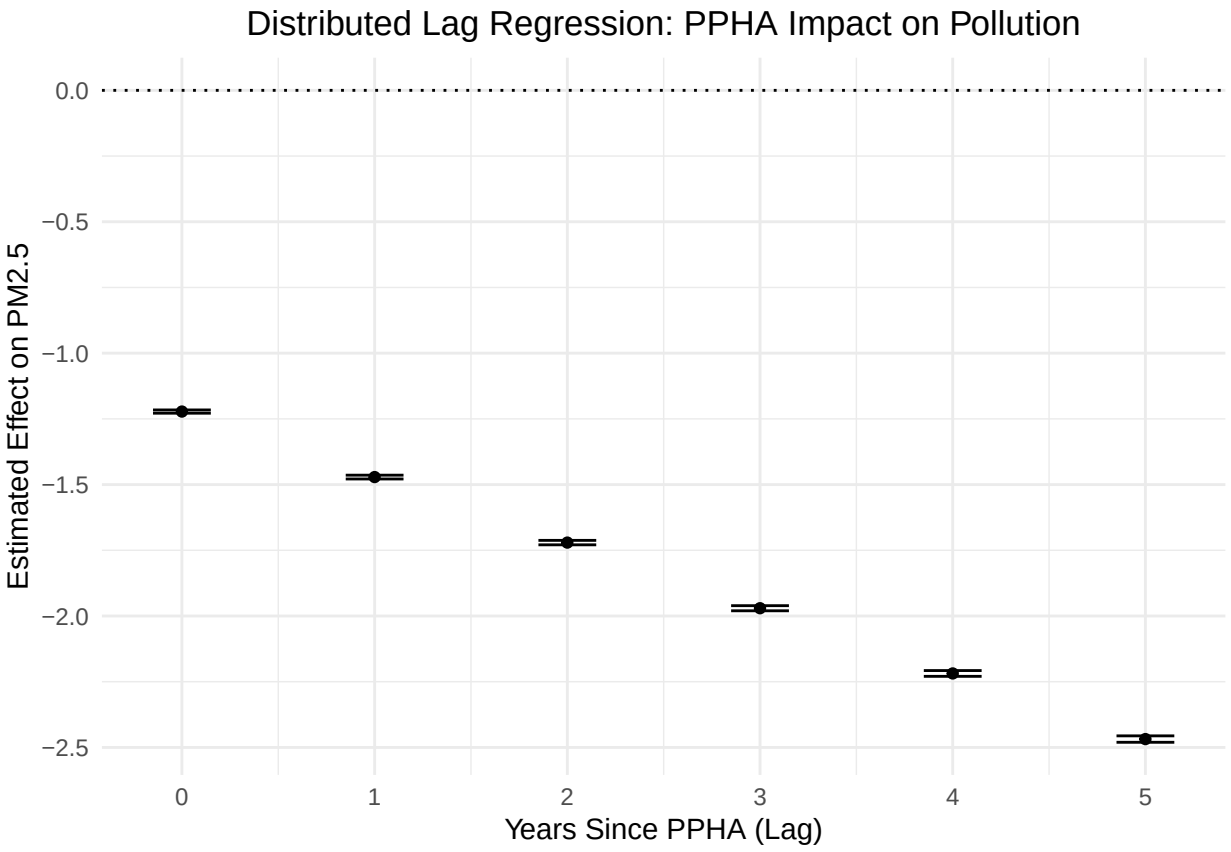
In the pre-treatment years (T = –6 to T = –2), the estimated treatment effects are all close to zero and statistically insignificant, with tight confidence intervals that include zero. This pattern supports the parallel trends assumption, suggesting that treated cities followed similar pollution trends to other cities prior to the policy implementation.

Starting at T = 0 (the year of PPHA adoption), pollution levels in treated cities drop sharply and continue to decline steadily over the following years. The estimated effects grow more negative over time, reaching a cumulative reduction of over 4 μg/m³ by year 8. All post-treatment estimates are statistically significant, with confidence intervals that exclude zero.

This pattern indicates a strong and persistent treatment effect, suggesting that the PPHA policy had a sustained impact on reducing air pollution. The gradual nature of the decline further implies that the policy's effects may have accumulated over time through enforcement, behavioral changes, or structural adjustments.

10.Distributed Lag Regression

```
lag_range <- 0:5
for (k in lag_range) {
  df[[paste0("lag_", k)]] <- ifelse(df$treated == 1 & df$year - df$ppha_year == k, 1, 0)
}

lag_vars <- paste0("lag_", lag_range)
rhs <- paste(lag_vars, collapse = " + ")
formula <- as.formula(paste0("pollution_pm ~ ", rhs, " | city_id + year"))

model_lag <- feols(formula, data = df, cluster = ~city_id)

event_effects <- broom::tidy(model_lag) %>%
  filter(term %in% lag_vars) %>%
  mutate(
    lag = as.numeric(gsub("lag_", "", term)),
    ci_lower = estimate - 1.96 * std.error,
    ci_upper = estimate + 1.96 * std.error
  )

ggplot(event_effects, aes(x = lag, y = estimate)) +
  geom_point() +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), width = 0.3) +
  geom_hline(yintercept = 0, linetype = "dotted") +
  scale_x_continuous(breaks = lag_range) +
  labs(
    title = "Distributed Lag Regression: PPHA Impact on Pollution",
    x = "Years Since PPHA (Lag)",
    y = "Estimated Effect on PM2.5"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
cumulative_effect <- sum(event_effects$estimate)
print(paste("Cumulative effect over 5 years:", round(cumulative_effect, 3)))
```

```
## [1] "Cumulative effect over 5 years: -11.072"
```