

[corrected]0403_Shuxi Chen_Problem Set 4

2025-04-03

Packages Used:

```
pacman::p_load(tidyverse,gt,foreign,knitr,webshot2,
               reshape2,viridis,ggthemes,stargazer,texreg,
               sandwich,modelsummary,Rmisc,neatStats,gmodels, dplyr, tidyr,
               plm, fixest, DRDID, gridExtra)
```

Difference-in-differences

```
df_Malesky <- read.dta("maleskyetal.dta")
```

Load and Inspect the Data

```
table(is.na(df_Malesky$column_name))
```

check NA

```
## < table of extent 0 >
```

Make a new indicator variable (treat) for whether observation is pre- or post-treatment period (2009)

```
df_Malesky$treat <- ifelse(df_Malesky$year == 2010, 1, 0)
```

Estimation: difference in means

Use the difference in means estimator to calculate the difference-in-differences estimate of the treatment effect for each service delivery indicator. Be sure to present your results in a nicely formatted table.

```
indices <- paste0("index", 1:6)
```

```
# Loop thru
```

```
results <- lapply(indices, function(idx) {
  treat_post <- mean(df_Malesky[[idx]][df_Malesky$treatment == 1 & df_Malesky$treat == 1], na.rm = TRUE)
  treat_pre  <- mean(df_Malesky[[idx]][df_Malesky$treatment == 1 & df_Malesky$treat == 0], na.rm = TRUE)
  control_post <- mean(df_Malesky[[idx]][df_Malesky$treatment == 0 & df_Malesky$treat == 1], na.rm = TRUE)
  control_pre  <- mean(df_Malesky[[idx]][df_Malesky$treatment == 0 & df_Malesky$treat == 0], na.rm = TRUE)

  did <- (treat_post - treat_pre) - (control_post - control_pre)

  data.frame(
    Index = idx,
    Treat_Post_Mean = treat_post,
    Treat_Pre_Mean = treat_pre,
    Control_Post_Mean = control_post,
```

```

    Control_Pre_Mean = control_pre,
    DiD_Estimate = did
  )
})

# Combine all results
results_df <- do.call(rbind, results)
print(results_df, row.names = FALSE)

```

```

##   Index Treat_Post_Mean Treat_Pre_Mean Control_Post_Mean Control_Pre_Mean
## index1      3.3882784      2.8980392      3.3273239      3.0859803
## index2      2.6659433      2.6375042      2.5118641      2.4959413
## index3      0.2825966      0.1821555      0.2264496      0.2447586
## index4      1.8126371      1.6335083      1.6290218      1.5482063
## index5      2.0439560      1.9333333      1.9650704      2.0098576
## index6      1.9830318      1.9513001      1.8348503      1.8363255
## DiD_Estimate
##   0.24889551
##   0.01251626
##   0.11875016
##   0.09831340
##   0.15540990
##   0.03320696

```

Estimation: regression

Estimate the model given above using ordinary least squares in R. Do this six times, once with each of the six service delivery indices as the outcome variable, using the data for 2008 through 2010. Be sure to estimate your standard errors appropriately. Report your results in a table, and interpret the statistical significance and direction of the causal effect of recentralization. The causal effect of recentralization appears positive and statistically significant in improving Health and Communication services, with some evidence of improvement in Infrastructure. However, there is no meaningful impact on Agriculture, Education, or Business Development, suggesting that the policy change was most effective in improving sectors requiring centralized coordination (health, communication), while having limited influence on more locally-dependent services like agriculture and education.

```
library(lmtest)
```

```

## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##   as.Date, as.Date.numeric

```

```
library(sandwich)
```

```

run_did <- function(outcome) {
  formula <- as.formula(paste(
    outcome, "~ treatment * treat + lnarea + lnpopden + city + factor(reg8)"
  ))

  # clustered SE

```

```

model <- lm(formula, data = df_Malesky)
clust_se <- vcovCL(model, cluster = ~tinh)
coeftest(model, vcov. = clust_se)
}

results <- lapply(indices, run_did)
names(results) <- indices

# Extract DiD coefficients
did_coefs <- lapply(results, function(x) {
  if (!"treatment:treat" %in% rownames(x)) {
    return(data.frame(Estimate = NA, SE = NA, p_value = NA))
  }
  coef <- x["treatment:treat", ]
  data.frame(
    Estimate = coef[1],
    SE = coef[2],
    p_value = coef[4]
  )
})

did_coefs_df <- do.call(rbind, did_coefs)
did_coefs_df$Index <- c("Infrastructure", "Agriculture", "Health",
  "Education", "Communication", "Business Dev")
results_table <- did_coefs_df[, c("Index", "Estimate", "SE", "p_value")] # Reorder columns

stargazer(
  results_table,
  title = "DiD Estimates",
  summary = FALSE,
  rownames = FALSE,
  type = "text",
  digits = 3,
  notes = "Clustered SE by district"
)

```

```

##
## DiD Estimates
## =====
## Index      Estimate  SE    p_value
## -----
## Infrastructure  0.225   0.125  0.072
## Agriculture    -0.003   0.103  0.975
## Health         0.123   0.032  0.0001
## Education      0.091   0.086  0.294
## Communication  0.152   0.071  0.032
## Business Dev   0.007   0.096  0.938
## -----
## Clustered SE by district

```

Parallel trends

```
df_placebo <- read.dta("maleskyetal_placebo.dta")
```

Load and Inspect the Data

```
table(is.na(df_placebo$column_name))
```

check NA

```
## < table of extent 0 >
```

Using the data from 2006 (available in `maleskyetal_placebo.dta`), and 2008 and 2010 (in `maleskyetal.dta`) create six parallel trend plots, one for each of the six indices. Present the six plots as a single figure on a single page of your problem set. In each panel of the figure, the x-axis should have be years (2006 through 2010), and the y-axis should take appropriate values given the particular index. For each year, for each of the control and treatment groups, plot the mean value of the particular index and the 95% confidence interval around that mean. Connect the means with lines to aid in visual interpretation of the trends. Be sure to also include a vertical line to indicate when in time treatment occurred. What do you conclude? For most indices (2, 4, 6), parallel trends hold pre-treatment, suggesting the DID design is valid. We observe positive effects in index4 (Education) where the treatment group outperforms the control post-2009 with non-overlapping CIs. However, index1, index3, index5 show pre-existing diverging trends before treatment, making causal interpretation unreliable. The remaining indices show no significant treatment effects.

```
library(foreign)
library(ggplot2)

df_main <- read.dta("maleskyetal.dta")

# Combine datasets
df_2006 <- df_placebo[df_placebo$year == 2006, ]
df_2008 <- df_placebo[df_placebo$year == 2008, ]
df_2010 <- df_main[df_main$year == 2010, ]
df_combined <- rbind(df_2006, df_2008, df_2010)

plot_list <- list()

# calculate mean and CI for each group
get_summary <- function(data, index) {
  # Control
  control <- data[data$treatment == 0, ]
  mean_control <- mean(control[, index], na.rm = TRUE)
  n_control <- sum(!is.na(control[, index]))
  se_control <- sd(control[, index], na.rm = TRUE) / sqrt(n_control)

  # Treatment
  treat <- data[data$treatment == 1, ]
  mean_treat <- mean(treat[, index], na.rm = TRUE)
  n_treat <- sum(!is.na(treat[, index]))
  se_treat <- sd(treat[, index], na.rm = TRUE) / sqrt(n_treat)

  return(data.frame(
    year = unique(data$year),
    mean_control = mean_control,
    lower_control = mean_control - 1.96 * se_control,
    upper_control = mean_control + 1.96 * se_control,
    mean_treat = mean_treat,
```

```

    lower_treat = mean_treat - 1.96 * se_treat,
    upper_treat = mean_treat + 1.96 * se_treat
  ))
}

# Loop thru to create plots
for (i in 1:6) {
  index_name <- paste0("index", i)

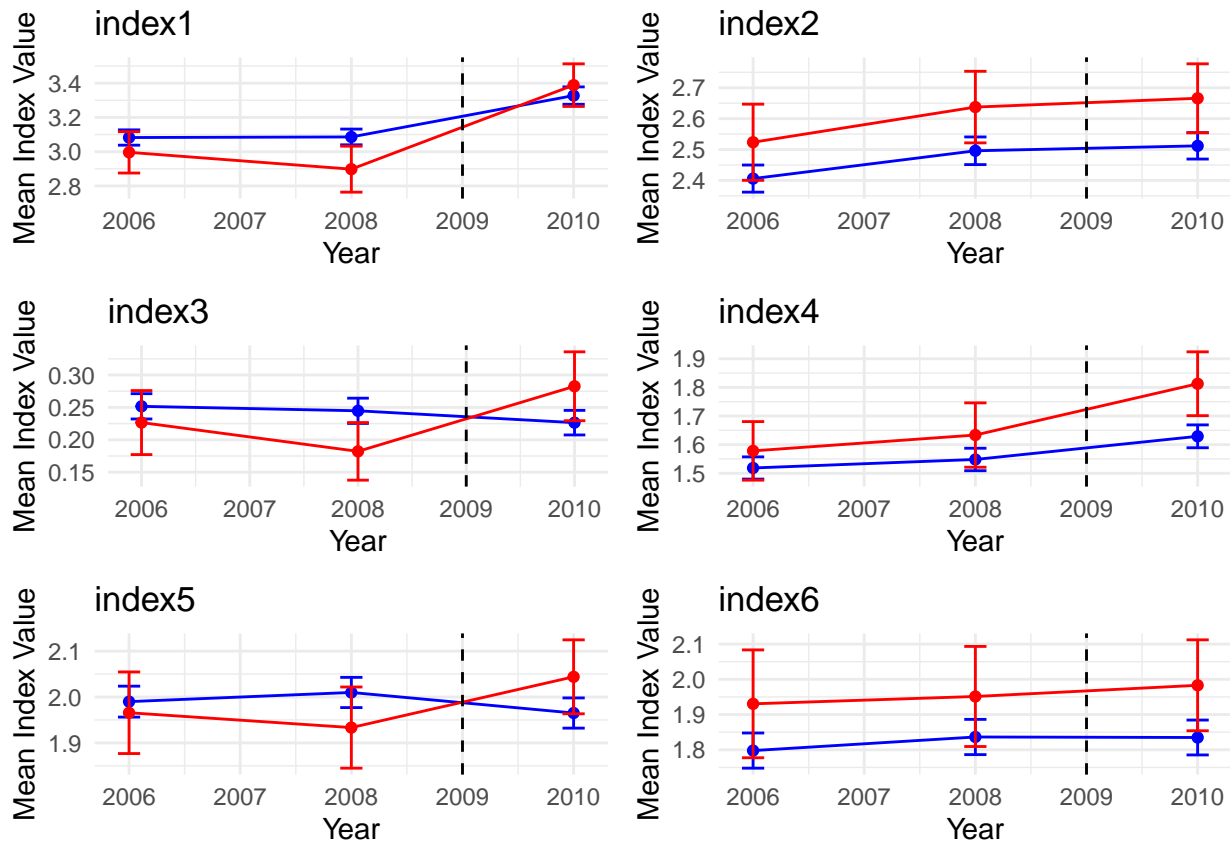
  # Get summary for each year
  summary_2006 <- get_summary(df_2006, index_name)
  summary_2008 <- get_summary(df_2008, index_name)
  summary_2010 <- get_summary(df_2010, index_name)
  plot_data <- rbind(summary_2006, summary_2008, summary_2010)

  # plot
  p <- ggplot(plot_data, aes(x = year)) +
    # Control
    geom_line(aes(y = mean_control), color = "blue") +
    geom_point(aes(y = mean_control), color = "blue") +
    geom_errorbar(aes(ymin = lower_control, ymax = upper_control),
                  width = 0.2, color = "blue") +
    # Treatment
    geom_line(aes(y = mean_treat), color = "red") +
    geom_point(aes(y = mean_treat), color = "red") +
    geom_errorbar(aes(ymin = lower_treat, ymax = upper_treat),
                  width = 0.2, color = "red") +
    # Treatment line
    geom_vline(xintercept = 2009, linetype = "dashed") +
    labs(title = index_name, x = "Year", y = "Mean Index Value") +
    theme_minimal()

  plot_list[[i]] <- p
}

grid.arrange(grobs = plot_list, ncol = 2)

```



Placebo test

Now, use the data from 2006 and 2008 (i.e., use only the data in `maleskyetal_placebo.dta`) to statistically assess whether treated and control units showed divergent trends just prior to 2009. We will do this by re-estimating the DID regression model before treatment, with a “fake” treatment occurring in 2007. (Note: You are replicating Table 3, Panel B, from the APSR paper). Present your results in a table and interpret what you find. What does it suggest about the main findings you estimated in the “regression” section? The main DiD estimates for most indices appear reliable, as they show no significant pre-treatment trends (p -values > 0.1). The exception is Infrastructure, where a marginally significant placebo effect ($p < 0.1$) suggests potential pre-existing divergence. This casts some doubt on the validity of its main positive treatment effect ($p = 0.072$), indicating that the observed gains may partially reflect pre-intervention trends rather than a true causal impact. In contrast, the improvements in health and communication services provide robust evidence of the policy’s effectiveness.

```
df_placebo$fake_treat <- ifelse(df_placebo$year >= 2007, 1, 0)

# to run placebo DID test for each index
run_placebo_test <- function(index_name) {
  formula <- as.formula(paste(index_name, "~ treatment*fake_treat + lnarea + lnpopden + city + factor(r
  model <- lm(formula, data = df_placebo)

  # Cluster SE
  vcov_cluster <- vcovCL(model, cluster = df_placebo$tinhh)
  coeffs <- coefest(model, vcov = vcov_cluster)

  return(coeffs["treatment:fake_treat", c("Estimate", "Std. Error", "Pr(>|t|)"))}
```

```

}

placebo_results <- t(sapply(paste0("index", 1:6), run_placebo_test))
colnames(placebo_results) <- c("Estimate", "Std. Error", "p-value")

print(placebo_results)

##           Estimate Std. Error    p-value
## index1 -0.11399921 0.06673895 0.08768483
## index2  0.07959141 0.13353240 0.55117681
## index3 -0.03344242 0.03163344 0.29048748
## index4  0.03266792 0.04719941 0.48889661
## index5 -0.05053366 0.05015370 0.31371612
## index6 -0.02393077 0.07143804 0.73765204

```

Panel data

Load and Inspect the Data

```
df_wide <- read.csv("wagepan_wide.csv")
```

check NA

```
table(is.na(df_wide$column_name))
```

```
## < table of extent 0 >
```

First, let's start by reshaping the data into a long format. You may use any method you prefer, just be sure to verify your data.

```

df_long <- df_wide %>%
  pivot_longer(
    cols = -nr,
    names_to = c(".value", "year"),
    names_pattern = "(.*)_(\\d+)"
  ) %>%
  mutate(year = as.numeric(year))

glimpse(df_long)

```

```

## Rows: 4,360
## Columns: 10
## $ nr      <int> 13, 13, 13, 13, 13, 13, 13, 13, 17, 17, 17, 17, 17, 17, 17, 17~
## $ year    <dbl> 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1980, 1981, 19~
## $ lwage   <dbl> 1.1975400, 1.8530600, 1.3444620, 1.4332130, 1.5681250, 1.69989~
## $ educ    <int> 14, 14, 14, 14, 14, 14, 14, 14, 13, 13, 13, 13, 13, 13, 13, 13~
## $ black   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ hisp    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ exper   <int> 1, 2, 3, 4, 5, 6, 7, 8, 4, 5, 6, 7, 8, 9, 10, 11, 4, 5, 6, 7, ~
## $ married <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, ~
## $ union   <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ hours   <int> 2672, 2320, 2940, 2960, 3071, 2864, 2994, 2640, 2484, 2804, 25~

table(df_long$year)

```

```
##
## 1980 1981 1982 1983 1984 1985 1986 1987
## 545 545 545 545 545 545 545 545
```

```
head(df_long)
```

```
## # A tibble: 6 x 10
##   nr year lwage educ black hisp exper married union hours
##   <int> <dbl> <dbl> <int> <int> <int> <int> <int> <int> <int>
## 1 13 1980 1.20 14 0 0 1 0 0 2672
## 2 13 1981 1.85 14 0 0 2 0 1 2320
## 3 13 1982 1.34 14 0 0 3 0 0 2940
## 4 13 1983 1.43 14 0 0 4 0 0 2960
## 5 13 1984 1.57 14 0 0 5 0 0 3071
## 6 13 1985 1.70 14 0 0 6 0 0 2864
```

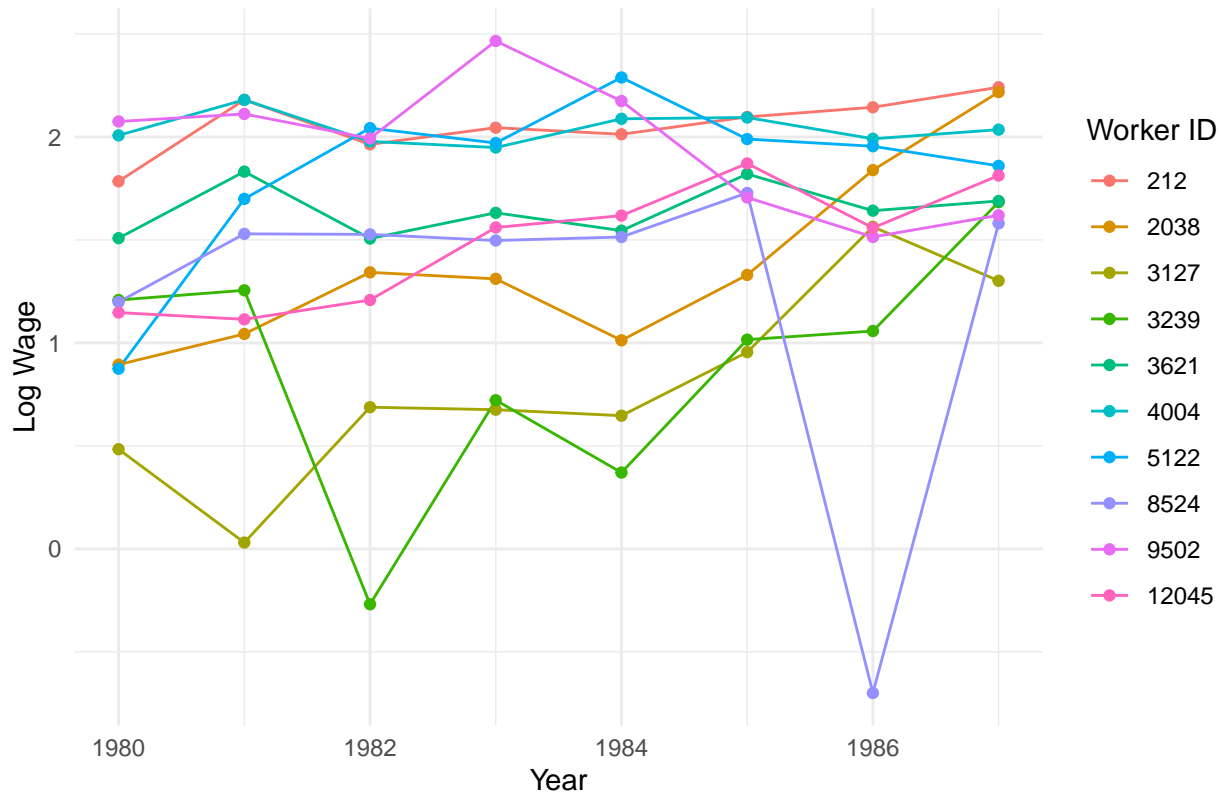
Variation between and within individuals

Choose 10 workers randomly and plot a line graph of the log of wages (lwage) for each unit over time on the same graph. Describe what you see - does the variation between individuals seem bigger than the variation within individuals? From the plot we can see for most of the sampled workers, the variation in wages between individuals is greater than within individuals over time. While workers maintain stable hierarchical positions, most show only minor fluctuations around their personal wage levels across years. This suggests that time invariant individual characteristics drive most wage differences, while short-term changes account for a smaller portion of overall wage variation.

```
set.seed(123)
sample_ids <- sample(unique(df_long$nr), 10)

ggplot(df_long[df_long$nr %in% sample_ids, ], aes(x = year, y = lwage, group = nr, color = factor(nr)))
  geom_line() +
  geom_point() +
  labs(title = "Wage Trajectories for 10 Random Workers",
       x = "Year", y = "Log Wage", color = "Worker ID") +
  theme_minimal()
```


Wage Trajectories for 10 Random Workers



Pooled OLS

Fit the following model and report the results. Describe what you are estimating and the assumptions necessary for your estimate(s) to be unbiased. In general, what does this analysis say about the effect of union membership on wages? The results show union members earn 18% higher wages after controlling for education, experience, and demographics. Key assumptions like no omitted variables, exogeneity for unbiased estimates may not hold, and in turn, the premium could partly reflect unobserved worker quality rather than pure union effects. While suggestive of union benefits, panel methods would better isolate causality.

```
pooled_model <- lm(lwage ~ educ + black + hisp + exper + I(exper^2) + married + union, data = df_long)
```

```
summary(pooled_model)
```

```
##
## Call:
## lm(formula = lwage ~ educ + black + hisp + exper + I(exper^2) +
##     married + union, data = df_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.2689 -0.2487  0.0332  0.2962  2.5608
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0347056  0.0645690  -0.537   0.591
## educ         0.0993878  0.0046776  21.248 < 2e-16 ***
## black       -0.1438417  0.0235595  -6.105 1.11e-09 ***
```

```
## hisp      0.0156980  0.0208112  0.754    0.451
## exper     0.0891791  0.0101110  8.820 < 2e-16 ***
## I(exper^2) -0.0028487  0.0007074 -4.027 5.74e-05 ***
## married   0.1076656  0.0156965  6.859 7.90e-12 ***
## union     0.1800725  0.0171205 10.518 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4807 on 4352 degrees of freedom
## Multiple R-squared:  0.1866, Adjusted R-squared:  0.1853
## F-statistic: 142.6 on 7 and 4352 DF,  p-value: < 2.2e-16
```

Within estimator

To estimate this model, use the within estimator. Do not use a canned function, instead demean your variables and run the appropriate model. For now report conventional OLS standard errors (but make sure you do the appropriate degrees of freedom correction).

```
wage_within <- df_long %>%
  group_by(nr) %>%
  mutate(
    lwage_demean = lwage - mean(lwage),
    exper_demean = exper - mean(exper),
    exper2_demean = (exper^2) - mean(exper^2),
    married_demean = married - mean(married),
    union_demean = union - mean(union)
  ) %>%
  ungroup()

within_model <- lm(lwage_demean ~ 0 + exper_demean + exper2_demean + married_demean + union_demean,
  data = wage_within)

# DoF
n <- length(unique(wage_within$nr))
k <- length(coef(within_model))
T <- nrow(wage_within)
df_correction <- sqrt((T - 1) / (T - k - n - 1))

# SE
raw_se <- summary(within_model)$coefficients[, "Std. Error"]
adjusted_se <- raw_se * df_correction

within_results <- data.frame(
  Estimate = coef(within_model),
  Std_Error = adjusted_se,
  t_value = coef(within_model) / adjusted_se,
  p_value = 2 * pt(abs(coef(within_model) / adjusted_se), df = T - k - n - 1, lower.tail = FALSE)
)

within_results

##              Estimate    Std_Error  t_value    p_value
## exper_demean    0.114022090 0.0113077924 10.083497 1.289611e-23
## exper2_demean  -0.006351987 0.0007752004 -8.193994 3.414642e-16
## married_demean  0.158460389 0.0174075451  9.102972 1.385110e-19
```

```
## union_demean      0.161206602 0.0191822981  8.403925 6.020794e-17
```

How are the assumptions different from the pooled model? What happens with time invariant variables, like race or ethnicity? Compare your estimate of the effect of union membership with the pooled OLS estimate. How has it changed? The within estimator reveals a slightly smaller but still robust union premium (16.1%), likely closer to the true causal effect after accounting for worker-level heterogeneity. The pooled OLS estimate (18.0%) was biased due to unobserved, time-invariant individual characteristics like race or ethnicity that are correlated with both union membership and wages. The within estimator accounts for these fixed traits by demeaning the data, thus offering a more credible causal estimate.

But it cannot estimate the effects of time-invariant variables, which are absorbed by the individual fixed effects. In contrast, the pooled OLS model can estimate those effects but relies on the stronger and less realistic assumption that individual-specific factors are uncorrelated with the regressors.

****Calculate the total number of workers that do not experience a change in their unionization status over time (i.e., for which the variance of union by individual, over time is zero). What happens to these units when we include unit fixed effects? How does that change our estimand? Finally, how are individuals with high variance in their unionization differ from individuals with low variance in calculating our estimate?*** When unit fixed effects are included in the model, any worker with no variation in a covariate like union status would be dropped from the estimation of that variable's effect, as the fixed effects absorb all between-person variation. In our case, since everyone's union status changes at least once, no workers are excluded, so the model can fully leverage within-person changes to estimate the effect of unionization. This allow us to isolate the causal effect of joining or leaving a union.

```
workers_constant_union <- df_long %>%
  group_by(nr) %>%
  summarize(union_var = var(union)) %>%
  filter(union_var == 0) %>%
  nrow()

total_workers <- length(unique(df_long$nr))

cat("# of workers with no change:", workers_constant_union, "\n")

## # of workers with no change: 0

cat("Total workers:", total_workers, "\n")

## Total workers: 545
```

Fixed effects

Given that we have panel data where units are observed repeatedly over time, it is unlikely that there is no serial correlation. We would expect individuals' potential outcomes to be correlated with themselves in the past. Estimate the same “within” model as before using an R package of your choice⁴. But now assume there is correlation within individuals when calculating your standard errors. Assume there is no correlation between units in any given year. Verify your point estimates from the manual within estimator. Show your results and the appropriate standard errors.

```
fe_model <- plm(lwage ~ exper + I(exper^2) + married + union,
  data = df_long,
  index = c("nr", "year"),
  model = "within",
  effect = "individual")

# cluster SE
fe_results <- coeftest(fe_model, vcov = vcovHC(fe_model, type = "HC1", cluster = "group"))
```

```
# manual vs package estimates
results_comparison <- data.frame(
  Manual_Estimate = coef(within_model),
  Manual_SE = summary(within_model)$coefficients[,2],
  Package_Estimate = fe_results[1:4, "Estimate"],
  Package_SE = fe_results[1:4, "Std. Error"]
)

rownames(results_comparison) <- c("exper", "I(exper^2)", "married", "union")
results_comparison
```

```
##           Manual_Estimate   Manual_SE Package_Estimate   Package_SE
## exper           0.114022090 0.0105717499      0.116846688 0.010703148
## I(exper^2)      -0.006351987 0.0007247413     -0.004300889 0.000685462
## married         0.158460389 0.0162744598      0.045303333 0.020984862
## union           0.161206602 0.0179336912      0.082087135 0.022805664
```

Now estimate the following model

```
twfe_model <- plm(lwage ~ educ + black + hisp + exper + I(exper^2) + married + union + factor(year),
  data = df_long,
  index = c("nr", "year"),
  model = "within",
  effect = "twoways")

# clustered SE
twfe_results <- coeftest(twfe_model,
  vcov = vcovHC(twfe_model,
    type = "HC1",
    cluster = "group"))

summary(twfe_model)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = lwage ~ educ + black + hisp + exper + I(exper^2) +
## married + union + factor(year), data = df_long, effect = "twoways",
## model = "within", index = c("nr", "year"))
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##      Min.    1st Qu.    Median    3rd Qu.     Max.
## -4.159280 -0.125273  0.011266  0.154869  1.492088
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## I(exper^2) -0.00518550 0.00070444 -7.3612 2.222e-13 ***
## married    0.04668038 0.01831044  2.5494  0.01083 *
## union      0.08000185 0.01931031  4.1430 3.503e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Total Sum of Squares:    479.09
## Residual Sum of Squares: 468.75
## R-Squared:    0.021568
## Adj. R-Squared: -0.12089
## F-statistic: 27.959 on 3 and 3805 DF, p-value: < 2.22e-16
```

```
twfe_results
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## I(exper^2) -0.00518550 0.00080884 -6.4110 1.623e-10 ***
## married    0.04668038 0.02096768  2.2263 0.0260523 *
## union      0.08000185 0.02270396  3.5237 0.0004306 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Compare with one-way FE
```

```
screenreg(list("One-way FE" = fe_model, "Two-way FE" = twfe_model),
  custom.model.names = c("Individual FE", "Individual + Year FE"),
  digits = 3)
```

```
##
## =====
##           Individual FE  Individual + Year FE
## -----
## exper          0.117 ***
##                (0.008)
## exper^2        -0.004 ***      -0.005 ***
##                (0.001)          (0.001)
## married         0.045 *          0.047 *
##                (0.018)          (0.018)
## union           0.082 ***        0.080 ***
##                (0.019)          (0.019)
## -----
## R^2             0.178           0.022
## Adj. R^2        0.060           -0.121
## Num. obs.    4360             4360
## =====
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Assume correlation within individuals. Show your results and the appropriate standard errors. What has changed? What does adding time fixed effects do? Do some variables drop out? If so, why? Adding clustered standard errors accounts for serial correlation within individuals, making our inference more reliable. The coefficients stay similar, but standard errors increase slightly. Adding year fixed effects controls for common shocks over time, further isolating the treatment effects. Time-invariant variables like race and education drop out because their effects are absorbed by the individual fixed effects.