

jh8186 solutionsHW5

Jerry Huang

2024-08-06

Problem 1

Part a. Parallel trends assumption

Parallel trends assumption states that the selection bias in time 1 is the same as the selection bias in time 0.
 $Y_{i1}(0) = \delta_t + u_i + \epsilon_{it}$, for $t = 0, 1$.

The difference for treated group:

$$E[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 1]$$

$$\begin{aligned} E[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 1] &= E[(\delta_1 + u_i + \epsilon_{i1}) - (\delta_0 + u_i + \epsilon_{i0}) \mid D_i = 1] \\ &= E[\delta_1 - \delta_0 + \epsilon_{i1} - \epsilon_{i0} \mid D_i = 1] \\ &= \delta_1 - \delta_0 + E[\epsilon_{i1} - \epsilon_{i0} \mid D_i = 1] \end{aligned}$$

The difference for control group:

$$E[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 0]$$

$$\begin{aligned} E[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 0] &= E[(\delta_1 + u_i + \epsilon_{i1}) - (\delta_0 + u_i + \epsilon_{i0}) \mid D_i = 0] \\ &= E[\delta_1 - \delta_0 + \epsilon_{i1} - \epsilon_{i0} \mid D_i = 0] \\ &= \delta_1 - \delta_0 + E[\epsilon_{i1} - \epsilon_{i0} \mid D_i = 0] \end{aligned}$$

Given that $E[\epsilon_{i1} \mid D_i = d] = E[\epsilon_{i0} \mid D_i = d] = \eta_i(d)$ for both $d = 1$ and $d = 0$, we will have $E[\epsilon_{i1} - \epsilon_{i0} \mid D_i = 1] = \eta_i(1) - \eta_i(1) = 0$ and $E[\epsilon_{i1} - \epsilon_{i0} \mid D_i = 0] = \eta_i(0) - \eta_i(0) = 0$. The expected difference in the error terms is zero for both treated and control groups. The expected difference will be $E[\epsilon_{i1} - \epsilon_{i0} \mid D_i = d] = 0$ regardless of whether $\eta_i(1)$ equals $\eta_i(0)$ or not.

Thus,

$$E[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 1] = E[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 0] = \delta_1 - \delta_0,$$

Part b. Unbiased estimator

We need to show that $E[\hat{\tau}] = \frac{1}{n} \sum_{i=1}^n \tau_i$.

For the treated group:

$$E \left[\frac{1}{n_t} \sum_{i=1}^n (Y_{i1} - Y_{i0}) D_i \right]$$

Substitute Y_{i1} and Y_{i0} :

$$Y_{i1} = \tau_i D_i + \delta_1 + u_i + \epsilon_{i1}$$

$$Y_{i0} = \delta_0 + u_i + \epsilon_{i0}$$

So,

$$Y_{i1} - Y_{i0} = (\tau_i D_i + \delta_1 + u_i + \epsilon_{i1}) - (\delta_0 + u_i + \epsilon_{i0}) = \tau_i D_i + (\delta_1 - \delta_0) + (\epsilon_{i1} - \epsilon_{i0})$$

Now,

$$E \left[\frac{1}{n_t} \sum_{i=1}^n (Y_{i1} - Y_{i0}) D_i \right] = E \left[\frac{1}{n_t} \sum_{i=1}^n (\tau_i D_i + (\delta_1 - \delta_0) + (\epsilon_{i1} - \epsilon_{i0})) D_i \right]$$

Since $E[\epsilon_{i1} - \epsilon_{i0} \mid D_i = d] = 0$, the expectation of the error term difference is zero:

$$\begin{aligned} & E \left[\frac{1}{n_t} \sum_{i=1}^n (\tau_i D_i + (\delta_1 - \delta_0)) D_i \right] \\ &= \frac{1}{n_t} \sum_{i=1}^n E[\tau_i D_i D_i + (\delta_1 - \delta_0) D_i] \\ &= \frac{1}{n_t} \sum_{i=1}^n (\tau_i E[D_i] + (\delta_1 - \delta_0) E[D_i]) \end{aligned}$$

Since $E[D_i] = \frac{n_t}{n}$:

$$\begin{aligned} &= \frac{1}{n_t} \sum_{i=1}^n (\tau_i \frac{n_t}{n} + (\delta_1 - \delta_0) \frac{n_t}{n}) \\ &= \frac{1}{n} \sum_{i=1}^n \tau_i + \frac{1}{n} \sum_{i=1}^n (\delta_1 - \delta_0) \end{aligned}$$

For the control group We will have a similar simplification process:

$$E \left[\frac{1}{n_c} \sum_{i=1}^n (Y_{i1} - Y_{i0}) (1 - D_i) \right]$$

Substitute Y_{i1} and Y_{i0} :

$$\begin{aligned} Y_{i1} &= \tau_i D_i + \delta_1 + u_i + \epsilon_{i1} \\ Y_{i0} &= \delta_0 + u_i + \epsilon_{i0} \end{aligned}$$

So,

$$Y_{i1} - Y_{i0} = (\tau_i D_i + \delta_1 + u_i + \epsilon_{i1}) - (\delta_0 + u_i + \epsilon_{i0}) = \tau_i D_i + (\delta_1 - \delta_0) + (\epsilon_{i1} - \epsilon_{i0})$$

Now,

$$E \left[\frac{1}{n_c} \sum_{i=1}^n (Y_{i1} - Y_{i0}) (1 - D_i) \right] = E \left[\frac{1}{n_c} \sum_{i=1}^n (\tau_i D_i + (\delta_1 - \delta_0) + (\epsilon_{i1} - \epsilon_{i0})) (1 - D_i) \right]$$

Since $E[\epsilon_{i1} - \epsilon_{i0} \mid D_i = d] = 0$, the expectation of the error term difference is zero:

$$\begin{aligned} & E \left[\frac{1}{n_c} \sum_{i=1}^n (\tau_i D_i + (\delta_1 - \delta_0)(1 - D_i)) \right] \\ &= \frac{1}{n_c} \sum_{i=1}^n E[\tau_i D_i(1 - D_i) + (\delta_1 - \delta_0)(1 - D_i)] \end{aligned}$$

Here, $D_i(1 - D_i) = 0$ since it is binary:

$$= \frac{1}{n_c} \sum_{i=1}^n (0) + (\delta_1 - \delta_0) E[1 - D_i]$$

Since $E[1 - D_i] = \frac{n_c}{n}$:

$$\begin{aligned} &= \frac{1}{n_c} \sum_{i=1}^n (0) + (\delta_1 - \delta_0) \frac{n_c}{n} \\ &= \frac{1}{n} \sum_{i=1}^n (0) + \frac{1}{n} \sum_{i=1}^n (\delta_1 - \delta_0) \end{aligned}$$

Combining treatment and control group

$$E[\hat{\tau}] = \frac{1}{n} \sum_{i=1}^n \tau_i + \frac{1}{n} \sum_{i=1}^n (\delta_1 - \delta_0) + (0) - \frac{1}{n} \sum_{i=1}^n (\delta_1 - \delta_0)$$

Eventually, we will have:

$$E[\hat{\tau}] = \frac{1}{n} \sum_{i=1}^n \tau_i$$

Part c.

$$E[\hat{\tau}_{t=1}] = E \left[\frac{1}{n_t} \sum_{i=1}^n Y_{i1} D_i \right] - E \left[\frac{1}{n_c} \sum_{i=1}^n Y_{i1} (1 - D_i) \right] = \frac{1}{n_t} \sum_{i=1}^n E[Y_{i1} D_i] - \frac{1}{n_c} \sum_{i=1}^n E[Y_{i1} (1 - D_i)]$$

Substitute Y_{i1} :

$$= \frac{1}{n_t} \sum_{i=1}^n E[(\tau_i D_i + \delta_1 + u_i + \epsilon_{i1}) D_i] - \frac{1}{n_c} \sum_{i=1}^n E[(\tau_i D_i + \delta_1 + u_i + \epsilon_{i1})(1 - D_i)]$$

Since D_i is binary and hence $D_i(1 - D_i) = 0$, we will further reduce the equation to:

$$= \frac{1}{n_t} \sum_{i=1}^n E[\tau_i D_i + \delta_1 D_i + u_i D_i + \epsilon_{i1} D_i] - \frac{1}{n_c} \sum_{i=1}^n E[\delta_1 (1 - D_i) + u_i (1 - D_i) + \epsilon_{i1} (1 - D_i)]$$

Given $E[\epsilon_{i1} \mid D_i = d] = \eta_i(d)$, we can split the expectation:

$$= \frac{1}{n_t} \sum_{i=1}^n E[\tau_i D_i + \delta_1 D_i + u_i D_i + \eta_i(1) D_i] - \frac{1}{n_c} \sum_{i=1}^n E[\delta_1 (1 - D_i) + u_i (1 - D_i) + \eta_i(0)(1 - D_i)]$$

Simplify the expectation terms:

$$= \frac{1}{n_t} \sum_{i=1}^n \frac{n_t}{n} (\tau_i + \delta_1 + u_i + \eta_i(1)) - \frac{1}{n_c} \sum_{i=1}^n \frac{n_c}{n} (\delta_1 + u_i + \eta_i(0))$$

Combining and Simplifying the fractions, the result of $E[\hat{\tau}_{t=1}]$ will be:

$$E[\hat{\tau}_{t=1}] = \frac{1}{n} \sum_{i=1}^n \tau_i + \frac{1}{n} \sum_{i=1}^n \eta_i(1) - \frac{1}{n} \sum_{i=1}^n \eta_i(0)$$

Thus, the bias term is:

$$\text{Bias}(\hat{\tau}_{t=1}) = E[\hat{\tau}_{t=1}] - \frac{1}{n} \sum_{i=1}^n \tau_i = \frac{1}{n} \sum_{i=1}^n (\eta_i(1) - \eta_i(0))$$

Problem 2 - Instrumental variable

We want to show that

$$E[\hat{\tau}_{Wald}^p] = E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)]$$

$$E[\hat{\tau}_{Wald}^p] = \frac{1}{p} \left(E \left[\frac{1}{n_t} \sum_{i=1}^n Y_i Z_i \right] - E \left[\frac{1}{n_c} \sum_{i=1}^n Y_i (1 - Z_i) \right] \right)$$

For the first term:

$$E \left[\frac{1}{n_t} \sum_{i=1}^n Y_i Z_i \right] = \frac{1}{n_t} \sum_{i=1}^n E[Y_i \mid Z_i = 1] \times \mathbb{P}(Z_i = 1)$$

Since $\mathbb{P}(Z_i = 1) = \frac{n_t}{n}$, this simplifies to:

$$\mathbb{E} \left[\frac{1}{n_t} \sum_{i=1}^n Y_i Z_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i \mid Z_i = 1].$$

For the second term:

$$\mathbb{E} \left[\frac{1}{n_c} \sum_{i=1}^n Y_i (1 - Z_i) \right] = \frac{1}{n_c} \sum_{i=1}^n \mathbb{E}[Y_i \mid Z_i = 0] \times \mathbb{P}(Z_i = 0).$$

Since $\mathbb{P}(Z_i = 0) = \frac{n_c}{n}$, this simplifies to:

$$\mathbb{E} \left[\frac{1}{n_c} \sum_{i=1}^n Y_i (1 - Z_i) \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i \mid Z_i = 0].$$

Combine the Terms The expected value of $\hat{\tau}_{Wald}^p$ is therefore:

$$\mathbb{E}[\hat{\tau}_{Wald}^p] = \frac{1}{p} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i \mid Z_i = 1] - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i \mid Z_i = 0] \right).$$

Relate to LATE By the definition of LATE, under monotonicity and independence assumptions, we have:

$$\mathbb{E}[Y_i | Z_i = 1] - \mathbb{E}[Y_i | Z_i = 0] = \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) > D_i(0)] \times \mathbb{P}(D_i(1) > D_i(0)).$$

Thus:

$$\mathbb{E}[\hat{\tau}_{\text{Wald}}^p] = \frac{1}{p} \times p \times \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) > D_i(0)] = \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) > D_i(0)].$$

Here we have confirmed that the estimator $\hat{\tau}_{\text{Wald}}^p$ is unbiased for the Local Average Treatment Effect (LATE).

Problem 3 -

```
# Loading Library needed
library(haven) # dta reader
library(estimatr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

diff_in_means <- function(treated, control){
  # Point Estimate
  point <- mean(treated) - mean(control)

  # Standard Error
  se <- sqrt(var(treated)/length(treated) + var(control)/length(control))

  # Asymptotic 95% CI
  ci_95 <- c(point - qnorm(.975)*se,
             point + qnorm(.975)*se)

  # P-value
  pval <- 2*pnorm(-abs(point/se))

  # Return as a data frame
  output <- data.frame(meanTreated = mean(treated), meanControl = mean(control), est = point, se = se, ci_95 = ci_95, pval = pval)

  return(as_tibble(output))
}
```

Question 1.

```
ohie <- read_dta("OHIE.dta")
```

```
itt_bp_model <- lm_robust(tab2bp_hyper ~ treatment, data = ohie)
summary(itt_bp_model)
```

```
##
## Call:
## lm_robust(formula = tab2bp_hyper ~ treatment, data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)  0.1591    0.004795 33.1816 3.215e-231  0.14971  0.16851 12186
## treatment   -0.0016    0.006621 -0.2417  8.090e-01 -0.01458  0.01138 12186
##
## Multiple R-squared:  4.796e-06 , Adjusted R-squared:  -7.727e-05
## F-statistic: 0.05842 on 1 and 12186 DF, p-value: 0.809
```

```
itt_depression_model <- lm_robust(tab2phqtot_high ~ treatment, data = ohie)
summary(itt_depression_model)
```

```
##
## Call:
## lm_robust(formula = tab2phqtot_high ~ treatment, data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)  0.30367    0.006034 50.327 0.000e+00  0.29184  0.31549 12159
## treatment   -0.03493    0.008207  -4.257 2.091e-05 -0.05102 -0.01885 12159
##
## Multiple R-squared:  0.001493 , Adjusted R-squared:  0.001411
## F-statistic: 18.12 on 1 and 12159 DF, p-value: 2.091e-05
```

```
itt_exp_model <- lm_robust(tab4_catastrophic_exp_inp ~ treatment, data = ohie)
summary(itt_exp_model)
```

```
##
## Call:
## lm_robust(formula = tab4_catastrophic_exp_inp ~ treatment, data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)  0.05382    0.003003 17.923 6.769e-71  0.04794  0.059711 11793
## treatment   -0.01527    0.003879  -3.936 8.336e-05 -0.02287 -0.007665 11793
##
## Multiple R-squared:  0.001329 , Adjusted R-squared:  0.001245
## F-statistic: 15.49 on 1 and 11793 DF, p-value: 8.336e-05
```

```
itt_care_model <- lm_robust(tab5_needmet_med_inp ~ treatment, data = ohie)
summary(itt_care_model)
```

```
##
## Call:
```

```
## lm_robust(formula = tab5_needmet_med_inp ~ treatment, data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)  0.61241    0.006378   96.02 0.00e+00  0.59990   0.6249 12214
## treatment    0.03446    0.008746    3.94 8.19e-05  0.01732   0.0516 12214
##
## Multiple R-squared:  0.001272 , Adjusted R-squared:  0.00119
## F-statistic: 15.52 on 1 and 12214 DF, p-value: 8.19e-05
```

Interpretation:

The estimated intent-to-treat (ITT) effect of being selected to apply for Medicaid on **elevated blood pressure** is -0.0016. This means that, on average, being selected in the lottery to apply for Medicaid is associated with a slight decrease of 0.0016 in the probability of having elevated blood pressure. The 95% confidence interval for this estimate ranges from **-0.0146 to 0.0114**. Since this interval includes 0, it indicates that the effect is not statistically significant at the 95% confidence level. Given that the **p-value is 0.809**, which is much greater than the typical threshold of 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to conclude that being selected to apply for Medicaid has an effect on the probability of having elevated blood pressure.

The estimated ITT effect of being selected to apply for Medicaid on **depression** is -0.0349. This means that being selected in the lottery to apply for Medicaid is associated with a decrease of 0.0349 on average in the probability of having depression. The 95% confidence interval for this estimate ranges from **-0.051 to -0.0188**. Since it does not include 0, it indicates that the effect is statistically significant at the 95% confidence level. Also, given that **p-value is 0.0000209**, which is much smaller than the threshold of 0.05, we can reject the null hypothesis. We do have sufficient evidence to conclude that being selected to apply for Medicaid has an effect on the probability of having depression.

The estimated ITT effect of being selected to apply for Medicaid on **catastrophic medical expenditure** is -0.0153. This means that being selected in the lottery to apply for Medicaid is associated with a decrease of 0.0349 on average in the probability of having catastrophic medical expenditure. The 95% confidence interval for this estimate ranges from **-0.0229 to -0.00766**. Since it does not include 0, it indicates that the effect is statistically significant at the 95% confidence level. Also, given that **p-value is 0.0000834**, which is much smaller than the threshold of 0.05, we can reject the null hypothesis. We do have sufficient evidence to conclude that being selected to apply for Medicaid has an effect on the probability of having catastrophic medical expenditure.

The estimated ITT effect of being selected to apply for Medicaid on **whether respondents had their health care needs met** is 0.0345. This means that being selected in the lottery to apply for Medicaid is associated with an increase of 0.0345 on average in the probability of having whether respondents had their health care needs met. The 95% confidence interval for this estimate ranges from **0.0173 to 0.05166**. Since it does not include 0, it indicates that the effect is statistically significant at the 95% confidence level. Also, given that **p-value is 0.0000819**, which is much smaller than the threshold of 0.05, we can reject the null hypothesis. We do have sufficient evidence to conclude that being selected to apply for Medicaid has an effect on the probability of having whether respondents had their health care needs met.

Question 2:

```
naive_bp_model <- lm_robust(tab2bp_hyper ~ ohp_all_ever_admin , data = ohie)
summary(naive_bp_model)
```

```
##
## Call:
## lm_robust(formula = tab2bp_hyper ~ ohp_all_ever_admin, data = ohie)
```

```
##
## Standard error type: HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)      0.16345   0.003966  41.212  0.00000   0.15567   0.171220
## ohp_all_ever_admin -0.01805   0.007162  -2.521  0.01173  -0.03209  -0.004014
##               DF
## (Intercept)      12186
## ohp_all_ever_admin 12186
##
## Multiple R-squared:  0.0005003 , Adjusted R-squared:  0.0004183
## F-statistic: 6.354 on 1 and 12186 DF,  p-value: 0.01173
naive_depression_model <- lm_robust(tab2phqtot_high ~ ohp_all_ever_admin, data = ohie)
summary(naive_depression_model)

##
## Call:
## lm_robust(formula = tab2phqtot_high ~ ohp_all_ever_admin, data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)      0.27129   0.004773  56.833 0.000e+00   0.26194   0.28065
## ohp_all_ever_admin 0.04932   0.009237   5.339 9.516e-08   0.03121   0.06742
##               DF
## (Intercept)      12159
## ohp_all_ever_admin 12159
##
## Multiple R-squared:  0.002438 , Adjusted R-squared:  0.002356
## F-statistic: 28.5 on 1 and 12159 DF,  p-value: 9.516e-08
naive_exp_model <- lm_robust(tab4_catastrophic_exp_inp ~ ohp_all_ever_admin, data = ohie)
summary(naive_exp_model)

##
## Call:
## lm_robust(formula = tab4_catastrophic_exp_inp ~ ohp_all_ever_admin,
##           data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)      0.04894   0.002351  20.812 1.658e-94   0.04433   0.053546
## ohp_all_ever_admin -0.01073   0.004052  -2.647 8.128e-03  -0.01867  -0.002784
##               DF
## (Intercept)      11793
## ohp_all_ever_admin 11793
##
## Multiple R-squared:  0.0005371 , Adjusted R-squared:  0.0004523
## F-statistic: 7.007 on 1 and 11793 DF,  p-value: 0.008128
```



```
naive_care_model <- lm_robust(tab5_needmet_med_inp ~ ohp_all_ever_admin, data = ohie)
summary(naive_care_model)
```

```
##
## Call:
## lm_robust(formula = tab5_needmet_med_inp ~ ohp_all_ever_admin,
##           data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)      0.61281   0.005220 117.399 0.000e+00  0.60258  0.62305
## ohp_all_ever_admin 0.06127   0.009482   6.461 1.078e-10  0.04268  0.07985
##              DF
## (Intercept)    12214
## ohp_all_ever_admin 12214
##
## Multiple R-squared:  0.003297 , Adjusted R-squared:  0.003216
## F-statistic: 41.75 on 1 and 12214 DF, p-value: 1.078e-10
```

Answer:

The result may be biased due to the following reasons:

- Selection Bias: Individuals who enrolled in Medicaid might differ systematically from those who did not. For example, those who chose to enroll might have different health behaviors or socioeconomic characteristics that also affect their health outcomes. This can lead to confounding, where the observed effect is partly due to these underlying differences rather than the Medicaid enrollment itself.

- Unobserved Confounders: We may have unobserved confounders that influence both treatment and outcome.
- Compliance Issues: Subjects who are assigned to treatment may not be compliant to the treatment.
- Measurement Error: If there is any error in measuring Medicaid enrollment, this can also lead to biased estimates.

Interpretation:

The estimated effect of Medicaid enrollment on **elevated blood pressure** is -0.0181, suggesting that enrollment is associated with a 1.81 percentage point decrease in the probability of having elevated blood pressure. The 95% confidence interval ranges from **-0.0321 to -0.00401**, indicating that this effect is **statistically significant at the 95% confidence level**. The p-value is **0.0117**, which is below the 0.05 threshold, further indicating statistical significance. In contrast, the ITT effect of treatment on elevated blood pressure is -0.0016 with a confidence interval (-0.0146, 0.0114), which shows that the point estimate is smaller and there is no statistically significant effect.

The estimated effect of Medicaid enrollment on **depression** is 0.0493, suggesting that enrollment is associated with a 4.93 percentage point increase in the probability of having depression. The 95% confidence interval ranges from **0.0312 to 0.0674**, indicating that this effect is statistically significant at the 95% confidence level. The p-value is **0.0000000952**, which is below the 0.05 threshold, further indicating statistical significance. In contrast, the ITT effect of the treatment on depression is -0.0349 with a CI (-0.051, -0.0188). Here we can see the direction of effect changes, so there might be unobserved confounders.

The estimated effect of Medicaid enrollment on **catastrophic medical expenditure** is -0.01073, suggesting that enrollment is associated with a 1.073 percentage point decrease in the probability of having catastrophic medical expenditure. The 95% confidence interval ranges from **-0.01867 to -0.002784**, indicating that this effect is statistically significant at the 95% confidence level. The p-value is **0.008128**, which is below the 0.05 threshold, further indicating statistical significance. In contrast, the ITT effect of the treatment on

catastrophic medical expenditure is -0.01527 with a CI (-0.02287, -0.007665).

The estimated effect of Medicaid enrollment on **whether respondents had their health care needs met** is 0.06127, suggesting that enrollment is associated with a 1.073 percentage point decrease in the probability of having whether respondents had their health care needs met. The 95% confidence interval ranges from **0.04268 to 0.07985**, indicating that this effect is statistically significant at the 95% confidence level. The p-value is **1.078e-10**, which is below the 0.05 threshold, further indicating statistical significance. In contrast, the ITT effect of the treatment on whether respondents had their health care needs met is 0.0345 with a CI (0.0173, 0.05166).

In conclusion, as we can see the result is quite different from ITT effect of treatment on those variables, there might be unobserved confounders between `ohp_all_ever_admin` and different outcomes. The `ohp_all_ever_admin` on blood pressure is significant, whereas the ITT effect of treatment on blood pressure is not.

Question 3:

```
compliance_selected <- sum(ohie$treatment == 1 & ohie$ohp_all_ever_admin == 1) / sum(ohie$treatment == 1)
compliance_non_selected <- sum(ohie$treatment == 0 & ohie$ohp_all_ever_admin == 0) / sum(ohie$treatment == 0)
print(c(compliance_selected, compliance_non_selected))
```

```
## [1] 0.3818357 0.8545455
```

```
first_stage <- lm_robust(ohp_all_ever_admin ~ treatment, data = ohie)
summary(first_stage)
```

```
##
## Call:
## lm_robust(formula = ohp_all_ever_admin ~ treatment, data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   0.1455    0.003467  41.95      0  0.1387   0.1523 20743
## treatment     0.2364    0.005891  40.12      0  0.2248   0.2479 20743
##
## Multiple R-squared:  0.07189 , Adjusted R-squared:  0.07184
## F-statistic: 1610 on 1 and 20743 DF, p-value: < 2.2e-16
```

Answer:

The compliance rate among those selected is around 38.2%, while among those not selected, it is around 85.5%. This indicates that a significant portion of the control group still managed to enroll, which might have implications for the analysis.

The result shows a high F-statistic (1610). A rule of thumb is that the F-statistic should be greater than 10 for the instrument to be considered strong, and here it is much higher, suggesting that the assignment to treatment is indeed a strong instrument for actual Medicaid enrollment. The p-value is extremely small and the CI does not include 0, further indicating that the assignment to treatment is a strong instrument.

Question 4:

```
# Second Stage result
iv_model_bp <- iv_robust(tab2bp_hyper ~ ohp_all_ever_admin | treatment, data = ohie)
summary(iv_model_bp)
```

```
##
```

```
## Call:
## iv_robust(formula = tab2bp_hyper ~ ohp_all_ever_admin | treatment,
##           data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)      0.1601   0.008181 19.5670 5.714e-84  0.14404  0.17611
## ohp_all_ever_admin -0.0063   0.026059 -0.2417 8.090e-01 -0.05738  0.04478
##               DF
## (Intercept)      12186
## ohp_all_ever_admin 12186
##
## Multiple R-squared:  0.0002882 , Adjusted R-squared:  0.0002062
## F-statistic: 0.05844 on 1 and 12186 DF,  p-value: 0.809

iv_model_dp <- iv_robust(tab2phqtot_high ~ ohp_all_ever_admin|treatment, data = ohie)
summary(iv_model_dp)

##
## Call:
## iv_robust(formula = tab2phqtot_high ~ ohp_all_ever_admin | treatment,
##           data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)      0.3248   0.01039 31.256 2.383e-206  0.3045  0.3452
## ohp_all_ever_admin -0.1376   0.03286 -4.188 2.838e-05 -0.2020 -0.0732
##               DF
## (Intercept)      12159
## ohp_all_ever_admin 12159
##
## Multiple R-squared: -0.03258 , Adjusted R-squared: -0.03267
## F-statistic: 17.54 on 1 and 12159 DF,  p-value: 2.838e-05

iv_model_exp <- iv_robust(tab4_catastrophic_exp_inp ~ ohp_all_ever_admin|treatment, data = ohie)
summary(iv_model_exp)

##
## Call:
## iv_robust(formula = tab4_catastrophic_exp_inp ~ ohp_all_ever_admin |
##           treatment, data = ohie)
##
## Standard error type: HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)      0.06314   0.00508 12.429 3.030e-35  0.05319  0.07310
## ohp_all_ever_admin -0.06036   0.01543 -3.913 9.163e-05 -0.09060 -0.03012
##               DF
## (Intercept)      11793
## ohp_all_ever_admin 11793
```

```
##
## Multiple R-squared:  -0.01096 , Adjusted R-squared:  -0.01105
## F-statistic: 15.31 on 1 and 11793 DF,  p-value: 9.163e-05

iv_model_need <- iv_robust(tab5_needmet_med_inp ~ ohp_all_ever_admin|treatment, data = ohie)
summary(iv_model_need)

##
## Call:
## iv_robust(formula = tab5_needmet_med_inp ~ ohp_all_ever_admin |
##           treatment, data = ohie)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper
## (Intercept)      0.5915   0.01086  54.448 0.000e+00  0.57022  0.6128
## ohp_all_ever_admin 0.1355   0.03442   3.935 8.354e-05  0.06798  0.2029
##              DF
## (Intercept)    12214
## ohp_all_ever_admin 12214
##
## Multiple R-squared:  -0.001537 , Adjusted R-squared:  -0.001619
## F-statistic: 15.49 on 1 and 12214 DF,  p-value: 8.354e-05
```

Answer:

Elevated Blood Pressure - LATE: -0.0063

- Naive estimate: -0.0181

The 95% Confidence interval is (-0.0574, 0.0448), which indicates that the result is not statistically significant since this interval includes 0. The p-value is 0.809, which is higher than the 0.05 threshold, further indicating it is no statistical significance. The LATE for the effect of Medicaid enrollment on elevated blood pressure is -0.0063. This means that for the subgroup of individuals who enrolled in Medicaid because they were selected in the lottery (the compliers), there is an estimated decrease of 0.63 percentage points in the probability of having elevated blood pressure. The Naive estimate does not account for potential biases due to selection effects and endogeneity, which might explain why it differs from the LATE.

Depression

- LATE: -0.0181

- Naive estimate: 0.0493 The 95% Confidence interval is (-0.202, -0.0732), which indicates that the result is statistically significant since this interval does not include 0. The p-value is 0.0000284, which is below the 0.05 threshold, further indicating statistical significance. The LATE for the effect of Medicaid enrollment on depression is -0.0181. This means that for the subgroup of individuals who enrolled in Medicaid because they were selected in the lottery (the compliers), there is an estimated decrease of 1.81 percentage points in the probability of having depression. The Naive estimate provides a different result (0.0493), which means that there is an estimated increase of 4.93 percentage points in the probability of having depression for all individuals enrolling in Medicaid. The fact that the naive estimate suggests an effect in the opposite direction compared to the LATE indicates that the bias is indeed strong. The result is significant here but is not in 3.2.

Catastrophic Medical Expenditure

- LATE: -0.0604 - Naive estimate: -0.0107 The 95% Confidence interval is (-0.0906, -0.0301), which indicates that the result is statistically significant since this interval does not include 0. The p-value is 0.0000916, which is below the 0.05 threshold, further indicating statistical significance. The LATE for the effect of Medicaid enrollment on Catastrophic Medical Expenditure is -0.0604. This means that for the subgroup of individuals who enrolled in Medicaid because they were selected in the lottery (the compliers), there is an estimated decrease of 6.04 percentage points in the probability of having Catastrophic Medical Expenditure. The Naive estimate provides a different result (-0.0107), which means that there is an estimated decrease of

-1.07 percentage points in the probability of having Catastrophic Medical Expenditure for all individuals enrolling in Medicaid.

Health care needs - LATE: 0.135 - Naive estimate: 0.0613 The 95% Confidence interval is (0.0000835, 0.068), which indicates that the result is statistically significant since this interval does not include 0. The p-value is 0.0000835, which is below the 0.05 threshold, further indicating statistical significance. The LATE for the effect of Medicaid enrollment on Catastrophic Medical Expenditure is 0.135. This means that for the subgroup of individuals who enrolled in Medicaid because they were selected in the lottery (the compliers), there is an estimated increase of 13.5 percentage points in the probability of having whether respondents had their health care needs met. The Naive estimate provides a smaller result (0.0613), which means that there is only an estimated increase of 6.13 percentage points in the probability of having whether respondents had their health care needs met for all individuals enrolling in Medicaid.

Problem 5.

Answer:

Randomization: Instrument is independent of both sets of potential outcomes (potential outcomes for the treatment and potential outcomes for the outcome).

Exclusion restriction: Random assignment affects the outcomes only through Medicaid enrollment.

First-stage relationship: The instrument must at least have some weak effect on treatment.

Monotonicity: There are no individuals who would enroll in Medicaid only if they were not assigned to the treatment group.

Homogeneous Treatment Effect: The treatment effect of Medicaid enrollment is the same for all individuals in the sample.

Problem 4 -

Question 1:

```
bases <- read_dta("bases_replication_final.dta")
number_control <- sum(bases$bases6 == 0)
print(number_control)

## [1] 16272

check_bases <- bases %>%
  group_by(municipality) %>%
  summarise(sd_bases = sd(bases6, na.rm = TRUE)) %>%
  summarise(change_over_time = any(sd_bases > 0))

check_military <- bases %>%
  group_by(municipality) %>%
  summarise(sd_military = sd(lrmilnar_col, na.rm = TRUE)) %>%
  summarise(change_over_time = any(sd_military > 0))

print(check_bases)

## # A tibble: 1 x 1
##   change_over_time
##   <lgl>
## 1 FALSE
print(check_military)

## # A tibble: 1 x 1
##   change_over_time
```

```
## <lg1>
## 1 TRUE
```

Answer:

There are 16272 units in the ‘control’ group. Bases do not change over time but the logged military aid variable does change across units for a given year. The author is assuming that while the presence of military bases is a fixed characteristic (time-invariant), the amount of U.S. military aid distributed to municipalities can vary both over time and across municipalities. This variation in aid might depend on factors like the severity of conflict, strategic importance, or other unobserved characteristics of the municipalities.

Question 2.

In the two-way fixed effects model used to estimate the effect of U.S. military and narcotics aid on paramilitary attacks, the authors make several key assumptions to identify the treatment effect:

1. Parallel Trends Assumption:

Explanation: The parallel trends assumption assumes that in the absence of the treatment (U.S. military aid), the trends in paramilitary attacks would have been the same across municipalities with and without military bases. If this assumption holds, the difference in trends between treated and control municipalities after the treatment can be attributed to the treatment effect (i.e., the interaction of military aid with the presence of a base). Violations of this assumption could lead to biased estimates of the treatment effect.

2. No Time-Varying Confounders: Explanation: The model assumes that there are no unobserved confounders that vary over time and are correlated with both the treatment (military aid) and the outcome (paramilitary attacks). This means that any factors influencing the level of paramilitary attacks that change over time must be either controlled for in the model or assumed to be unrelated to the treatment. If there are unobserved, time-varying confounders (e.g., changes in local governance or economic conditions) that influence both the allocation of military aid and paramilitary violence, the treatment effect estimate may be biased.

3. Homogeneity of Treatment Effect: Explanation: The two-way fixed effects model assumes that the treatment effect (the impact of military aid on paramilitary attacks) is homogeneous across all municipalities and time periods. This means the model assumes that the effect of the aid does not vary between different municipalities or over different years. If the treatment effect varies across municipalities (e.g., if aid is more effective in some areas than others), the model’s estimate of the average treatment effect may not accurately represent the true impact of military aid.

4. Exogeneity of Treatment: Explanation: The model assumes that the treatment (military aid, as interacted with the presence of a military base) is exogenous. This means that the allocation of military aid is not influenced by potential outcomes (i.e., it is not driven by the current or expected levels of paramilitary violence). If military aid is allocated based on expected levels of violence (e.g., more aid is sent to areas already experiencing high violence), this could lead to endogeneity, resulting in biased estimates of the treatment effect.

Question 3.

```
# install.packages("plm")
library(plm)
```

```
##
## Attaching package: 'plm'

## The following objects are masked from 'package:dplyr':
##
##   between, lag, lead
```

```
library(lmtest) # For coeftest function
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

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

library(sandwich) # For clustered standard errors

bases_4 <- bases %>%
  select(paratt, year, municipality, bases6xlrmlnar_col, lnnewpop, lrmilnar_col)

data <- pdata.frame(bases_4, index = c('municipality', 'year'))

did_model <- plm(paratt ~ bases6xlrmlnar_col + lnnewpop ,
  data = data,
  model = 'within',
  effect = 'twoways')
summary(did_model)

## Twoways effects Within Model
##
## Call:
## plm(formula = paratt ~ bases6xlrmlnar_col + lnnewpop, data = data,
##      effect = "twoways", model = "within")
##
## Unbalanced Panel: n = 936, T = 8-18, N = 16606
##
## Residuals:
##      Min.      1st Qu.      Median      3rd Qu.      Max.
## -3.276524 -0.129748 -0.029909  0.053485  13.828254
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## bases6xlrmlnar_col 0.150312   0.017293  8.6918 < 2.2e-16 ***
## lnnewpop           0.117848   0.034770  3.3894 0.0007022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    3412.5
## Residual Sum of Squares: 3391.9
## R-Squared:    0.0060294
## Adj. R-Squared: -0.054558
## F-statistic: 47.4696 on 2 and 15651 DF, p-value: < 2.22e-16
# Calculate clustered standard errors at the municipality level
clustered_se <- vcovHC(did_model, type = "HC1", cluster = "group")

# Display the summary with clustered standard errors
coeftest(did_model, vcov = clustered_se)

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## bases6xlrmlnar_col 0.150312   0.060090  2.5014 0.012379 *
```

```
## lnnewpop          0.117848   0.045237   2.6051 0.009193 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

coef_value <- 0.150312
se_value <- 0.060090

# Calculate the confidence intervals
CI_upper <- coef_value + 1.96 * se_value
CI_lower <- coef_value - 1.96 * se_value

# Print the confidence intervals
print(paste("95% CI: [", CI_lower, ", ", CI_upper, "]", sep = ""))

## [1] "95% CI: [0.0325356, 0.2680884]"
```

The confidence interval for the coefficient is [0.11641772, 0.18420628]. The positive and statistically significant coefficient for bases6xlrmlnar_col (0.150312) suggests that the interaction between U.S. military and narcotics aid (lrmlnar_col) and the presence of a military base (bases6) has a significant positive effect on the number of paramilitary attacks. Specifically, for each one-unit increase in the interaction term, the number of paramilitary attacks increases by approximately 0.15, holding other factors constant.

The results from this two-way fixed effects model indicate that the interaction of U.S. military and narcotics aid with the presence of a military base is associated with a significant increase in paramilitary attacks in Colombian municipalities. Additionally, municipalities with larger populations also experience more paramilitary attacks. These findings suggest that U.S. military aid, when coupled with military bases, may have unintended consequences, potentially exacerbating violence in areas with military presence.

```
{r} # # Run the two-way fixed effects model # result_model <-
lm_robust(paratt ~ bases6xlrmlnar_col + lnnewpop + factor(municipality)
+ factor(year), # data = bases, clusters =
municipality) # summary(result_model) #
```