Problem Set 4

Ghina Al Shdaifat - gha2009 - 003

Due Dec 13, 2023

This homework must be turned in on Brightspace by Dec. 13 2023. It must be your own work, and your own work only – you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone.

Your homework submission must be written and submitted using Rmarkdown. No handwritten solutions will be accepted. **No zip files will be accepted. Make sure we can read each line of code in the pdf document.** You should submit the following:

1. A compiled PDF file named yourNetID_solutions.pdf containing your solutions to the problems.
2. A .Rmd file containing the code and text used to produce your compiled pdf named yourNetID_solutions.Rmd.

Note that math can be typeset in Rmarkdown in the same way as Latex. Please make sure your answers are clearly structured in the Rmarkdown file:

1. Label each question part
2. Do not include written answers as code comments.
3. The code used to obtain the answer for each question part should accompany the written answer. Comment your code!

Problem 1 (100 points)

Despite the heated political and media rhetoric, there are a few causal estimates of the effect of expanded health insurance on healthcare outcomes. One landmark study, the Oregon Health Insurance Experiment, covered new ground by utilizing a randomized control trial implemented by the state of Oregon. To allocate a limited number of eligible coverage slots for the state's Medicaid expansion, about 30,000 low-income, uninsured adults (out of about 90,000 wait-list applicants) were randomly selected by lottery to be allowed to apply for Medicaid coverage. Researchers collected observable measure of health (blood pressure, cholesterol, blood sugar levels, and depression), as well as hospital visitations and healthcare expenses for 6,387 selected adults and 5,842 not selected adults.

For this problem, we will use the OHIE.dta file.

- treatment - selected in the lottery to sign up for Medicaid (instrument)
- ohp_all_ever_admin - Ever enrolled in Medicaid after notification of lottery results (compliance)
- tab2bp_hyper - Outcome: Binary indicator for elevated blood pressure (1 indicates a high blood pressure)
- tab2phqtot_high - Outcome: Binary indicator for depression
- tab4_catastrophic_exp_inp - Outcome: Indicator for catastrophic medical expenditure (1 if their total out-of-pocket medical expenses are larger than 30% of their household income)
- tab5_needmet_med_inp - Outcome: Binary indicator of whether the participant feels that they received all needed medical care in past 12 months

```r
# Load in the data
data <- haven::read_dta("OHIE.dta")
head(data)
```

```
## # A tibble: 6 x 59
##   weight_total_inp tab1_gender_inp tab2dia_dx_post_lott~1 tab2hbp_dx_post_lott~2
##              <dbl> <dbl+lbl>       <dbl+lbl>              <dbl+lbl>
## 1            1.15   1 [Female]      0 [No]                 0 [No]
## 2            0.897  0 [Male]        0 [No]                 1 [Yes]
## 3            0     NA              NA                     NA
## 4            1      1 [Female]      0 [No]                 0 [No]
## 5            1.21   0 [Male]        0 [No]                 0 [No]
## 6            1      0 [Male]        0 [No]                 0 [No]
## # i abbreviated names: 1: tab2dia_dx_post_lottery, 2: tab2hbp_dx_post_lottery
## # i 55 more variables: tab2chl_dx_post_lottery <dbl+lbl>,
## #   tab2dep_dx_post_lottery <dbl+lbl>, tab3_pcs8_score <dbl>,
## #   tab3_mcs8_score <dbl>, tab5_usual_clinic_inp <dbl+lbl>,
## #   tab5_needmet_med_inp <dbl+lbl>, tab5_chl_chk_inp <dbl+lbl>,
## #   tab5_pap_chk_inp <dbl+lbl>, tab5_fobt_chk_inp <dbl+lbl>,
## #   tab5_col_chk_inp <dbl+lbl>, tab5_psa_chk_inp <dbl+lbl>, ...
```

Hint: This was an experiment with imperfect compliance. Instead of creating a "participated" or "complied" variable, simply use "treatment" as the instrument and "ohp_all_ever_admin" (enrollment in Medicaid) as the main independent variable of interest.

Question A (25 points)

Estimate the intent-to-treat effects of being selected to sign up for Medicaid on each of the four outcomes (elevated blood pressure, depression, catastrophic medical expenditure, and whether respondents had their health care needs met). Provide 95% confidence intervals for each estimate and interpret your results. (Use lm_robust)

```r
# Estimate the ITT on elevated blood pressure
itt_bp <- lm_robust(tab2bp_hyper ~ treatment, data = data)
summary(itt_bp)
```

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

```r
# Estimate the ITT on depression
itt_depression <- lm_robust(tab2phqtot_high ~ treatment, data = data)
summary(itt_depression)
```

```
##
## Call:
## lm_robust(formula = tab2phqtot_high ~ treatment, data = data)
##
## 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
```

```r
# Estimate the ITT on catastrophic expenditures
itt_catastrophic <- lm_robust(tab4_catastrophic_exp_inp ~ treatment,
                              data = data)
summary(itt_catastrophic)
```

```
##
## Call:
## lm_robust(formula = tab4_catastrophic_exp_inp ~ treatment, data = data)
##
## 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
```

```r
# Estimate the ITT on "needs met"
itt_needs_met <- lm_robust(tab5_needmet_med_inp ~ treatment, data = data)
summary(itt_needs_met)
```

```
##
## Call:
## lm_robust(formula = tab5_needmet_med_inp ~ treatment, data = data)
##
## 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 of results:

The impact on elevated blood pressure reveals an estimate of -0.0016, where the confidence interval ranges from -0.01458 to 0.01138. The data suggests that the likelihood or severity of elevated blood pressure is not significantly altered by enrollment in Medicaid. This is evident from the confidence interval that crosses zero, implying no statistically significant change.

In regards to the effect of depression, the study indicates a significant reduction in depression among participants, with an estimate of -0.03493, with a confidence interval that extends from -0.05102 to -0.01885. This negative estimate, coupled with a confidence interval that does not cross zero, strongly suggests that Medicaid enrollment may play a role in lessening depression symptoms or lowering the risk of depression.

There is a significant decrease in catastrophic medical expenditures, as shown by an estimate of -0.01527. The confidence interval lies between -0.02287 and -0.007665. This outcome implies that Medicaid enrollment potentially alleviates the financial burden associated with major medical expenses, as indicated by the negative estimate and a confidence interval that excludes zero.

The impact on meeting health care needs shows a positive effect, with an estimate of 0.03446 and a confidence interval that ranges from 0.01732 to 0.0516. This positive estimate and its corresponding confidence interval suggest a significant improvement in the fulfillment of health care needs for those selected for Medicaid, indicating that Medicaid enrollees are more likely to have their health care requirements adequately addressed.

Overall, the findings in this specific scenario highlights that selection for Medicaid enrollment has a beneficial effect in reducing depression and catastrophic health expenditures, as well as in enhancing the satisfaction of health care needs. However, it does not appear to significantly affect elevated blood pressure levels. This analysis underscores the potential positive implications of Medicaid in specific health and financial domains.

Question B (25 points)

Suppose that researchers actually wanted to estimate the effect of Medicaid enrollment (ohp_all_ever_admin) on each of the four outcomes. Suppose they first used a naive regression of each of the the outcomes on the indicator of Medicaid enrollment. Report a 95% confidence interval for each of your estimates and interpret your results. Why might these be biased estimates for the causal effect of Medicaid enrollment?

```r
# Create a function to estimate Naive OLS effect and print the results
estimate_naive_regression <- function(outcome) {
model <- lm(as.formula(paste(outcome, "~ ohp_all_ever_admin")), data = data)
summary_result <- summary(model)
ci <- confint(model, level = 0.95)
return(list(summary = summary_result, ci = ci))
}

# Estimate the Naive OLS effect on elevated blood pressure
naive_bp <- estimate_naive_regression("tab2bp_hyper")
naive_bp
```

```
## $summary
##
## Call:
## lm(formula = as.formula(paste(outcome, "~ ohp_all_ever_admin")),
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1635 -0.1635 -0.1635 -0.1454  0.8546
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.163446   0.003914   41.76   <2e-16 ***
## ohp_all_ever_admin -0.018054   0.007310   -2.47   0.0135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3649 on 12186 degrees of freedom
##   (8557 observations deleted due to missingness)
## Multiple R-squared:  0.0005003,  Adjusted R-squared:  0.0004183
## F-statistic:   6.1 on 1 and 12186 DF,  p-value: 0.01353
##
##
## $ci
##                          2.5 %     97.5 %
## (Intercept)         0.15577430  0.1711178
## ohp_all_ever_admin -0.03238241 -0.0037255
```

```r
# Estimate the Naive OLS effect on depression
naive_depression <- estimate_naive_regression("tab2phqtot_high")
naive_depression
```

```
## $summary
##
## Call:
## lm(formula = as.formula(paste(outcome, "~ ohp_all_ever_admin")),
##     data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3206 -0.2713 -0.2713  0.6794  0.7287
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.271292   0.004843  56.021  < 2e-16 ***
## ohp_all_ever_admin 0.049317   0.009048   5.451 5.11e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4511 on 12159 degrees of freedom
##   (8584 observations deleted due to missingness)
## Multiple R-squared:  0.002438,   Adjusted R-squared:  0.002356
## F-statistic: 29.71 on 1 and 12159 DF,  p-value: 5.113e-08
##
##
## $ci
##                        2.5 %     97.5 %
## (Intercept)        0.2617994 0.28078441
## ohp_all_ever_admin 0.0315818 0.06705135
```

```
# Estimate the Naive OLS effect on catastrophic expenditures
naive_expenses <- estimate_naive_regression("tab4_catastrophic_exp_inp")
naive_expenses
```

```
## $summary
##
## Call:
## lm(formula = as.formula(paste(outcome, "~ ohp_all_ever_admin")),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04894 -0.04894 -0.04894 -0.03821  0.96179
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.048937   0.002280  21.468   <2e-16 ***
## ohp_all_ever_admin -0.010726   0.004261  -2.517   0.0118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2092 on 11793 degrees of freedom
##   (8950 observations deleted due to missingness)
## Multiple R-squared:  0.0005371,  Adjusted R-squared:  0.0004523
## F-statistic: 6.337 on 1 and 11793 DF,  p-value: 0.01184
##
##
## $ci
##                         2.5 %       97.5 %
## (Intercept)         0.04446869  0.053405171
## ohp_all_ever_admin -0.01907791 -0.002374143
```

```r
# Estimate the Naive OLS on needs met
naive_care_needs_met <- estimate_naive_regression("tab5_needmet_med_inp")
naive_care_needs_met
```

```
## $summary
##
## Call:
## lm(formula = as.formula(paste(outcome, "~ ohp_all_ever_admin")),
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6741 -0.6128  0.3259  0.3872  0.3872
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.612814   0.005164 118.665  < 2e-16 ***
## ohp_all_ever_admin 0.061266   0.009638   6.356 2.14e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4819 on 12214 degrees of freedom
##   (8529 observations deleted due to missingness)
## Multiple R-squared:  0.003297,   Adjusted R-squared:  0.003216
## F-statistic:  40.4 on 1 and 12214 DF,  p-value: 2.138e-10
##
##
## $ci
##                         2.5 %     97.5 %
## (Intercept)        0.60269157 0.62293709
## ohp_all_ever_admin 0.04237332 0.08015884
```

Interpretation of results:

The confidence interval for the influence on elevated blood pressure ranges from -0.0324 to -0.00373. This outcome implies a marginal decrease in elevated blood pressure among those enrolled in Medicaid. The effect size is minor, and the range of the confidence interval is quite tight.

On the other hand, the impact on depression has a confidence interval that extends from 0.0316 to 0.0671. This suggests a possible rise in depression cases for Medicaid participants. The statistical significance of this effect is underlined by the confidence interval not covering zero.

The effect on catastrophic medical expenditures has a confidence interval that lies between -0.0191 and -0.00237. This negative estimate hints at a decrease in catastrophic medical costs due to Medicaid enrollment and is additionally statistically significant.

With a confidence interval ranging from 0.0424 to 0.0802, the result points to an enhanced fulfillment of healthcare needs in Medicaid enrollees, a statistically significant finding.

Potential Bias:

The population enrolling in Medicaid may have distinct characteristics compared to non-enrollees. These characteristics, whether health-related, economic, or demographic, might influence both enrollment decisions and health outcomes, leading to skewed estimates. Key variables that influence both Medicaid enrollment and health outcomes, yet not accounted for in the regression, can additionally lead to misleading results. It is challenging to pinpoint if Medicaid enrollment leads to specific health outcomes or if individuals with certain health conditions are more inclined to enroll in Medicaid. It is also important to note that any errors in reporting or recording data related to Medicaid enrollment or health outcomes can also skew the findings.

In conclusion, while the initial regression approach sheds light on the association between Medicaid enrollment and various health outcomes, these findings could be skewed due to selection bias, missing important variables, the potential for a two-way relationship, and data inaccuracies. For a more accurate depiction of causality, methodologies like randomized trials or instrumental variable analysis would be more reliable.

Question C (25 points)

Suppose we were to use assignment to treatment as an instrument for actually receiving Medicaid coverage.

Consider that not everyone who was selected to apply for Medicaid actually ended up applying and receiving coverage. Likewise, some applicants who were not selected to receive the treatment nevertheless were eventually covered. What were the compliance rates (the level of Medicaid enrollment) for subjects who were selected and subjects who were not selected? Use a "first stage" regression to estimate the effect of being selected on Medicaid enrollment to estimate the compliance rates. Is the instrument of assignment-to-treatment a strong instrument for actual Medicaid enrollment?

```r
# First Stage OLS
first_stage_mod <- lm_robust(ohp_all_ever_admin ~ treatment, data = data)

# null model (compliance given an intercept only model)
null_mod<-lm_robust(ohp_all_ever_admin ~ 1, data=data)

# F - Stat for Instrument Strength (use waldtest)
instrument_strength <- waldtest(first_stage_mod, "treatment")
summary(first_stage_mod)
```

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

```r
instrument_strength
```

```
## Wald test
##
## Model 1: ohp_all_ever_admin ~ treatment
## Model 2: ohp_all_ever_admin ~ 1
##   Res.Df Df  Chisq Pr(>Chisq)
## 1  20743
## 2  20744 -1 1609.8  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation of results:

The coefficient for the treatment variable (being selected for Medicaid) is 0.236, with a 95% confidence interval ranging from 0.225 to 0.248. The statistical significance of this coefficient is high with a p-value of less than 2e-16 (p value $<$ 2e-16). The positive and statistically significant coefficient indicates that being selected to apply for Medicaid substantially increases the likelihood of actual enrollment. This is evident from the size of the coefficient and its confidence interval. A rather key measure of an instrument's strength is the F-statistic from the first stage regression. Here, the F-statistic is 1.61e+03, which is substantially larger than 10. Additionally, the relatively large R-squared value (0.0719) suggests that a significant portion of the variation in Medicaid enrollment is explained by the treatment assignment.

Based on these results, we can conclude that the instrument of assignment-to-treatment is indeed a strong instrument for actual Medicaid enrollment. The high F-statistic and the significant coefficient for the treatment variable in the first stage regression support this conclusion. In this scenario, the assignment to treatment serves as a strong instrument for Medicaid enrollment. This is confirmed by the substantial impact of the treatment assignment on enrollment, as evidenced by the significant coefficient and the robust F-statistic.

Question D (25 points)

Now estimate the effect of Medicaid enrollment on each of the four outcomes using an instrumental variables strategy. Report a 95% confidence interval for your estimates and interpret your results. Compare the estimates to those you obtained in Question B.

```r
run_iv_estimation <- function(outcome) { iv_model <- iv_robust(
formula = as.formula(paste(outcome, "~ ohp_all_ever_admin | treatment")),
data = data )
ci <- confint(iv_model, level = 0.95)
return(list(estimation = iv_model, ci = ci)) }

# Estimate the IV effect on elevated blood pressure (use iv_robust())
iv_bp <- run_iv_estimation("tab2bp_hyper")
iv_bp
```

```
## $estimation
##                        Estimate  Std. Error    t value     Pr(>|t|)    CI Lower
## (Intercept)         0.160076358 0.008180946 19.5669739 5.713710e-84  0.14404041
## ohp_all_ever_admin -0.006299556 0.026059156 -0.2417406 8.089852e-01 -0.05737964
##                      CI Upper    DF
## (Intercept)        0.17611231 12186
## ohp_all_ever_admin 0.04478052 12186
##
## $ci
##                          2.5 %     97.5 %
## (Intercept)         0.14404041 0.17611231
## ohp_all_ever_admin -0.05737964 0.04478052
```

```r
# Estimate the IV effect on depression
iv_depression <- run_iv_estimation("tab2phqtot_high")
iv_depression
```

```
## $estimation
##                      Estimate Std. Error   t value      Pr(>|t|)   CI Lower
## (Intercept)         0.3248452 0.01039309 31.255885 2.383439e-206  0.3044731
```

```
## ohp_all_ever_admin -0.1376126 0.03286133 -4.187675  2.838127e-05 -0.2020260
##                       CI Upper    DF
## (Intercept)         0.34521729 12159
## ohp_all_ever_admin -0.07319914 12159
##
## $ci
##                         2.5 %      97.5 %
## (Intercept)         0.3044731  0.34521729
## ohp_all_ever_admin -0.2020260 -0.07319914
```

```r
# Estimate the IV effect on catastrophic expenditures
iv_expenses <- run_iv_estimation("tab4_catastrophic_exp_inp")
iv_expenses
```

```
## $estimation
##                       Estimate  Std. Error   t value     Pr(>|t|)    CI Lower
## (Intercept)         0.06314350 0.005080412 12.428816 3.029925e-35  0.05318506
## ohp_all_ever_admin -0.06036067 0.015425260 -3.913105 9.162856e-05 -0.09059672
##                       CI Upper    DF
## (Intercept)         0.07310195 11793
## ohp_all_ever_admin -0.03012461 11793
##
## $ci
##                          2.5 %      97.5 %
## (Intercept)         0.05318506  0.07310195
## ohp_all_ever_admin -0.09059672 -0.03012461
```

```r
# IV estimate on needs met
iv_care_needs_met <- run_iv_estimation("tab5_needmet_med_inp")
iv_care_needs_met
```

```
## $estimation
##                     Estimate Std. Error   t value    Pr(>|t|)   CI Lower
## (Intercept)        0.5915172 0.01086399 54.447502 0.00000e+00 0.57022203
## ohp_all_ever_admin 0.1354509 0.03441917  3.935333 8.35417e-05 0.06798387
##                     CI Upper    DF
## (Intercept)        0.6128123 12214
## ohp_all_ever_admin 0.2029179 12214
##
## $ci
##                         2.5 %    97.5 %
## (Intercept)        0.57022203 0.6128123
## ohp_all_ever_admin 0.06798387 0.2029179
```

Interpretation of results:

The IV estimate (-0.0063) of elevated blood pressure shows a negligible and statistically insignificant effect of Medicaid enrollment on elevated blood pressure, as indicated by the confidence interval including zero (-0.0574 to 0.0448). Compared to the results from Question B, this finding is consistent in suggesting no significant impact on elevated blood pressure. On the other hand, the negative IV depression estimate (-0.138) has a confidence interval that does not include zero (-0.202 to -0.0732), indicating a significant reduction in depression due to Medicaid enrollment. This effect is opposite in comparison to Question B, suggesting potential bias in the naive OLS model. The IV results show a significant decrease in catastrophic medical expenditures, with an estimate value of -0.0604 and a confidence interval that ranges from -0.0906 to -0.0301 being entirely below zero. This finding aligns with the results from Question B but shows a more substantial effect, indicating that Medicaid enrollment might have a stronger protective effect against catastrophic medical expenses than previously estimated. The positive IV estimate (0.135), with a confidence interval (from 0.068 to 0.203) that does not include zero, suggests a significant improvement in meeting health care needs due to Medicaid enrollment. This result is in line with the findings from Question B but indicates a more substantial positive impact of Medicaid on meeting health care needs.

Overall, the IV estimates generally align with the direction of the effects observed in Question B but tend to show more pronounced effects. This could indicate that the naive estimates from Question B may have been biased, potentially underestimating the true effects of Medicaid enrollment on these outcomes. The instrumental variables approach likely provides a more accurate estimation of the causal effects of Medicaid enrollment by addressing potential endogeneity and unobserved confounding factors.