Problem Set 4

Erin Choi - eyc321 - Section 2

Due Dec 9, 2022

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
- ohp_all_ever_admin - Ever enrolled in Medicaid after notification of lottery results
- 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

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

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)

```
# ITT of treatment on elevated blood pressure
lm_robust(tab2bp_hyper ~ treatment, data=data)
```

```
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   0.1591    0.00480  33.182 3.21e-231   0.1497   0.1685 12186
## treatment    -0.0016    0.00662  -0.242  8.09e-01  -0.0146   0.0114 12186
```

The ITT of treatment on elevated blood pressure was -0.0016, with a 95% confidence interval of [-0.0146, 0.0114]. The point estimate indicates that being selected to enroll in Medicaid has a very slight negative impact on having elevated blood pressure, i.e. selection for Medicaid enrollment slightly decreases blood pressure. However, the confidence interval does contain zero, so this ITT estimate does not represent a statistically significant effect.

```
# on depression
lm_robust(tab2phqtot_high ~ treatment, data=data)
```

```
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   0.3037    0.00603   50.33 0.00e+00    0.292   0.3155 12159
## treatment    -0.0349    0.00821   -4.26 2.09e-05   -0.051  -0.0188 12159
```

The ITT of treatment on depression was -0.0349, with a 95% confidence interval of [-0.051, -0.0188]. This point estimate is also negative, indicating that being selected to enroll in Medicaid decreases depression by a small margin. The confidence interval is completely lower than zero and does not contain zero, so this estimate is statistically significant.

```
# on catastrophic medical expenditure
lm_robust(tab4_catastrophic_exp_inp ~ treatment, data=data)
```

```
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   0.0538    0.00300   17.92 6.77e-71   0.0479  0.05971 11793
## treatment    -0.0153    0.00388   -3.94 8.34e-05  -0.0229 -0.00766 11793
```

The ITT of treatment on catastrophic medical expenditures is estimated at -0.0153, with a 95% confidence interval of [-0.0229, -0.00766]. Again, this point estimate is negative, so it is estimated that selection for Medicaid enrollment slightly decreases the occurrence of a catastrophic medical expenditure (having out-of-pocket medical expense that is over 30% of household income). The confidence interval does not contain zero, so this estimate represents a statistically significant effect.

```
# on meeting healthcare needs
lm_robust(tab5_needmet_med_inp ~ treatment, data=data)
```

```
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   0.6124    0.00638   96.02 0.00e+00   0.5999   0.6249 12214
## treatment     0.0345    0.00875    3.94 8.19e-05   0.0173   0.0516 12214
```

Finally, the ITT of treatment on whether the participant feels they received all needed medical care in the past year is 0.0345, with a 95% confidence interval of [0.0173, 0.0516]. This estimate is positive, signaling that selection for Medicaid enrollment increases the proportion of participants who report that all their healthcare needs have been met. The confidence interval does not contain zero, so the estimate is statistically significant.

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?

```
# Estimate effect of Medicaid enrollment on elevated blood pressure
lm_robust(tab2bp_hyper ~ ohp_all_ever_admin, data=data)
```

```
##                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)          0.1634    0.00397   41.21   0.0000   0.1557  0.17122 12186
## ohp_all_ever_admin  -0.0181    0.00716   -2.52   0.0117  -0.0321 -0.00401 12186
```

The estimated effect of Medicaid enrollment on elevated blood pressure was -0.0181, with a 95% confidence interval of [-0.0321, -0.00401]. The point estimate indicates that Medicaid enrollment slightly decreases blood pressure. The confidence interval does not contain zero, so this estimate represents a statistically significant effect.

```
# on depression
lm_robust(tab2phqtot_high ~ ohp_all_ever_admin, data=data)
```

```
##                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)          0.2713    0.00477   56.83 0.00e+00   0.2619   0.2806 12159
## ohp_all_ever_admin   0.0493    0.00924    5.34 9.52e-08   0.0312   0.0674 12159
```

The effect of Medicaid enrollment on depression was estimated to be 0.0492, with a 95% confidence interval of [0.0312, 0.0674]. This point estimate is positive, indicating that enrolling in Medicaid actually increases depression. The confidence interval is completely above than zero and does not contain zero, so this estimate is statistically significant.

```
# on catastrophic medical expenditure
lm_robust(tab4_catastrophic_exp_inp ~ ohp_all_ever_admin, data=data)
```

```
##                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)          0.0489    0.00235   20.81 1.66e-94   0.0443  0.05355 11793
## ohp_all_ever_admin  -0.0107    0.00405   -2.65 8.13e-03  -0.0187 -0.00278 11793
```

The effect of enrolling in Medicaid on catastrophic medical expenditures was estimated as -0.0107, with a 95% confidence interval of [-0.0187, -0.00278]. This point estimate is negative, so Medicaid enrollment is estimated to slightly decrease the occurrence of a catastrophic medical expenditure. The confidence interval does not contain zero, thus the estimate represents a statistically significant effect.

```
# on meeting healthcare needs
lm_robust(tab5_needmet_med_inp ~ ohp_all_ever_admin, data=data)
```

```
##                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)          0.6128    0.00522  117.40 0.00e+00   0.6026   0.6230 12214
## ohp_all_ever_admin   0.0613    0.00948    6.46 1.08e-10   0.0427   0.0799 12214
```

Lastly, the effect of Medicaid enrollment on whether the participant feels they received all needed medical care in the past year is 0.0613, with a 95% confidence interval of [0.0427, 0.0799]. The estimate is positive, so it is estimated that Medicaid enrollment increases the proportion of participants with completely met healthcare needs. This confidence interval does not contain zero, so the estimate is statistically significant.

These estimates for the causal effect of Medicaid enrollment may be biased because of a compliance issue; it is very unlikely that all participants who were selected to be allowed to enroll in Medicaid actually did enroll then actually received coverage. It is also likely that some of those who were not selected through lottery were able to enroll in Medicaid through other means. Unobserved variables affect compliance in such ways in the observational study that make it difficult to estimate an unbiased causal effect.

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?

```
# first stage regression, estimate compliance rates
# = effect of being selected in lottery on Medicaid enrollment
# Note the treatment is binary (either selected or not)
stage1 = lm_robust(ohp_all_ever_admin ~ treatment, data=data)
tidy(stage1)
```

```
##          term estimate std.error statistic p.value conf.low conf.high    df
## 1 (Intercept)    0.145   0.00347      42.0       0    0.139     0.152 20743
## 2   treatment    0.236   0.00589      40.1       0    0.225     0.248 20743
##              outcome
## 1 ohp_all_ever_admin
## 2 ohp_all_ever_admin
```

```
# find the estimate value and F-statistic
cat(paste("Estimates of compliance rate",
          "\nFor selected subjects:", stage1$coefficients[1]+stage1$coefficients[2],
          "\nFor not selected subjects:", stage1$coefficients[1],
          "\nF-statistic:", stage1$fstatistic[1]))
```

```
## Estimates of compliance rate
## For selected subjects: 0.381835655934658
## For not selected subjects: 0.145454545454567
## F-statistic: 1609.82611136488
```

```
# check estimates against computed conditional probabilities:
# nrow(data[data$treatment==1 & data$ohp_all_ever_admin==1, ])/
  # nrow(data[data$treatment==1, ]) yields 0.382
# nrow(data[data$treatment==0 & data$ohp_all_ever_admin==1, ])/
  # nrow(data[data$treatment==0, ]) yields 0.145
```

The effect of being selected in the lottery on enrolling in Medicaid is estimated as 0.236. The compliance rate for subjects who were not selected is equal to the intercept of 0.145; 14.5% of subjects who were not selected in the lottery eventually enrolled in Medicaid. The compliance rate for subjects who were selected is the sum of the coefficients, since 0.236 represents the increase in Medicaid enrollment that is attributed to selection in the lottery. Thus the compliance rate for selected subjects is $0.236 + 0.145 = 0.381$, meaning 38.1% of subjects who were selected in the lottery enrolled in Medicaid. This is actually not a very high compliance rate, if fewer than half of the "treated" participants actually chose to enroll in Medicaid.

The F-statistic for this first-stage regression is about 1609, which is much greater than the threshold value for a strong instrument of about 104. This indicates that the instrument of assignment-to-treatment (selection to actual enrollment) is a strong instrument for actual Medicaid enrollment.

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.

For the instrumental variables strategy, there are four assumptions that must be satisfied.

- The instrument ("treated" variable) is "as-if randomly assigned"; it *was* randomly assigned through lottery!
- Exclusion restriction is satisfied, as being chosen in the lottery (the instrument) can only affect each of the outcomes through compliance (actually enrolling for Medicaid after selection).
- There is a strong first-stage relationship, as verified by checking that the F-statistic (1609) is higher than about 104 in Question C.
- Monotonicity is assumed in this case by assuming that enrolling for Medicaid can only improve outcomes (change outcomes in one direction) by lowering blood pressure, decreasing depression rates, lessening proportions of catastrophic medical expenditure for subjects, and increasing the proportions of people whose healthcare needs were met.

Now the iv_robust function can be used to not only estimate the effect of Medicaid enrollment on the four outcomes, but to also estimate the standard errors accurately.

```
# Estimate of effect of Medicaid enrollment on high blood pressure
iv_robust(tab2bp_hyper ~ ohp_all_ever_admin|treatment, data = data)
```

```
##                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)          0.1601    0.00818  19.567 5.71e-84   0.1440   0.1761 12186
## ohp_all_ever_admin  -0.0063    0.02606  -0.242 8.09e-01  -0.0574   0.0448 12186
```

The estimated effect of Medicaid enrollment on higher blood pressure was -0.0063, with a 95% confidence interval of [-0.0574, 0.0448]. The point estimate is closer to 0 than that produced using a naive regression (-0.0181), but it is still negative, indicating that Medicaid enrollment slightly decreases blood pressure. However, the standard error of this estimate (0.02606) is quite large compared to that of the naive regression (0.00716). This results in the confidence interval containing zero, unlike the confidence interval constructed using naive regression, so the estimate using instrumental variables does not represent a statistically significant effect.

```
# on depression
iv_robust(tab2phqtot_high ~ ohp_all_ever_admin|treatment, data = data)
```

```
##                    Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper
## (Intercept)           0.325     0.0104   31.26 2.38e-206    0.304   0.3452
## ohp_all_ever_admin   -0.138     0.0329   -4.19  2.84e-05   -0.202  -0.0732
##                        DF
## (Intercept)         12159
## ohp_all_ever_admin  12159
```

The effect of Medicaid enrollment on depression was estimated to be -0.138, with a 95% confidence interval of [-0.202, -0.0732]. This point estimate is negative, unlike the positive naive estimate of 0.0492, so the instrumental variables approach estimates a slight decrease in depression from Medicaid enrollment. Though the standard error of 0.0329 is large than the naive regression's standard error (0.00924), this confidence interval also does not contain zero, so this estimate is statistically significant.

```r
# on catastrophic medical expenditure
iv_robust(tab4_catastrophic_exp_inp ~ ohp_all_ever_admin|treatment, data = data)
```

```
##                     Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)           0.0631    0.00508   12.43 3.03e-35   0.0532   0.0731 11793
## ohp_all_ever_admin   -0.0604    0.01543   -3.91 9.16e-05  -0.0906  -0.0301 11793
```

The effect of enrolling in Medicaid on catastrophic medical expenditures was estimated as -0.0604, with a 95% confidence interval of [-0.0906, -0.0301], The point estimate is more negative (about 6 times more so) than the naive estimate of -0.0107, so both the naive and instrumental variable methods estimate decreases in the occurrence of catastrophic medical expenditures. Like the naive confidence interval, this confidence interval does not contain zero, thus the estimate represents a statistically significant effect. This standard error (0.01543) is also much larger than the naive standard error of 0.00405.

```r
# on meeting healthcare needs
iv_robust(tab5_needmet_med_inp ~ ohp_all_ever_admin|treatment, data = data)
```

```
##                     Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)            0.592     0.0109   54.45 0.00e+00    0.570    0.613 12214
## ohp_all_ever_admin     0.135     0.0344    3.94 8.35e-05    0.068    0.203 12214
```

Lastly, the effect of Medicaid enrollment on fully met medical care in the past year is 0.135, with a 95% confidence interval of [0.068, 0.203]. This estimate is positive and over twice as large as the naive estimate of 0.0613, so this approach also estimates that Medicaid enrollment increases the proportion of participants with completely met healthcare needs. This confidence interval does not contain zero, so the estimate is statistically significant. Again, the instrumental variable approach's standard error of 0.0344 is much larger than that of the naive regression (0.00948).

Besides the adding assignment to treatment as an instrumental variable, more differences in estimates and standard errors between the two methods may come from confounding. There are definitely many confounders involved when tracking the paths from selection to enrollment, then enrollment to each of the four outcomes, but there was no controlling for covariates done in both the naive regression and the instrumental variables approach.