Stats2hmwk5

2024-02-27

```
#1
setwd("/Users/sarahkim/Documents/Coding/")
ohie <- read.csv("OHIE.csv")
model1 <- lm(count_visit_dr ~ I(female == 1) + treated + I(female == 1):treated
             + numhh_list, data = ohie)
summary(model1)
```

```
##
## Call:
## lm(formula = count_visit_dr ~ I(female == 1) + treated + I(female ==
##     1):treated + numhh_list, data = ohie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -7.989  -5.092  -3.092   0.281 139.281
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.7281     0.3822  17.605  < 2e-16 ***
## I(female == 1)TRUE           2.5011     0.3125   8.004 1.32e-15 ***
## treated                      0.3735     0.3257   1.147    0.251
## numhh_list                  -2.0095     0.2482  -8.097 6.16e-16 ***
## I(female == 1)TRUE:treated   0.3959     0.4317   0.917    0.359
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.78 on 12153 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.01906,    Adjusted R-squared:  0.01874
## F-statistic: 59.03 on 4 and 12153 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(count_visit_dr ~ I(age >= 50) + treated + I(age >= 50):treated
             + numhh_list, data = ohie)
summary(model2)
```

```
##
## Call:
## lm(formula = count_visit_dr ~ I(age >= 50) + treated + I(age >=
##     50):treated + numhh_list, data = ohie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -7.171  -5.743  -3.743   0.140 138.062
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                8.0153     0.3548  22.591   <2e-16 ***
## I(age >= 50)TRUE           0.8048     0.3511   2.293   0.0219 *  
## treated                    0.6368     0.2542   2.505   0.0122 *  
## numhh_list                -2.0775     0.2504  -8.295   <2e-16 ***
## I(age >= 50)TRUE:treated  -0.2083     0.4821  -0.432   0.6657    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.86 on 12152 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.006969,   Adjusted R-squared:  0.006642 
## F-statistic: 21.32 on 4 and 12152 DF,  p-value: < 2.2e-16
```

```
model3 <- lm(count_visit_dr ~ I(race_white == 1) + treated +
               I(race_white == 1):treated + numhh_list, data = ohie)
summary(model3)
```

```
##
## Call:
## lm(formula = count_visit_dr ~ I(race_white == 1) + treated + 
##     I(race_white == 1):treated + numhh_list, data = ohie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
##  -7.300  -5.427  -3.427   0.521 138.573 
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      7.2399     0.4231  17.112  < 2e-16 ***
## I(race_white == 1)TRUE           1.1835     0.3368   3.514 0.000443 ***
## treated                          0.1348     0.3854   0.350 0.726621    
## numhh_list                      -1.9481     0.2505  -7.776 8.06e-15 ***
## I(race_white == 1)TRUE:treated   0.6895     0.4637   1.487 0.137087    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.81 on 12121 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.0101, Adjusted R-squared:  0.009775 
## F-statistic: 30.92 on 4 and 12121 DF,  p-value: < 2.2e-16
```

```
model4 <- lm(count_visit_dr ~ I(health_baseline == 1) + treated +
               I(health_baseline == 1):treated + numhh_list, data = ohie)
summary(model4)
```

```
##
## Call:
## lm(formula = count_visit_dr ~ I(health_baseline == 1) + treated + 
##     I(health_baseline == 1):treated + numhh_list, data = ohie)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
##  -8.397  -5.458  -3.449   0.025 138.542 
## 
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                           7.4669     0.3514  21.246  < 2e-16 ***
## I(health_baseline == 1)TRUE           2.5166     0.3491   7.210 5.94e-13 ***
## treated                               0.6436     0.2523   2.551   0.0108 *  
## numhh_list                           -2.0087     0.2489  -8.070 7.66e-16 ***
## I(health_baseline == 1)TRUE:treated  -0.2218     0.4827  -0.460   0.6459    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 11.81 on 12153 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.01431,    Adjusted R-squared:  0.01399 
## F-statistic: 44.12 on 4 and 12153 DF,  p-value: < 2.2e-16
```

For the first regression where female is treated, the coefficient is 0.3959.

These can be interpreted to mean that holding all else constant, for those who are female, we expect that those who won the lottery to apply to medicaid to mean that an increase in the number of doctor visits by 0.7694 (.3735 + .3959). So the difference in the number of doctor visits between females and males who got the lottery and apply for medicaid on average is .3959.

Additionally, it is not statistically significant since .359 > .05. We fail to reject the null that Beta3 = 0.

Compared to the problem set 3, question 2: the coefficient of female == 1 is statistically significant so this is different than what we found above.

For the second regression where age>=50 is treated, the coefficient is -0.2083

These can be interpreted to mean that holding all else constant, for those who are over 50, we expect that those who won the lottery to apply to medicaid to mean that an increase in the number of doctor visits by 0.4285 (0.6368 + -0.2083). So the difference in the number of doctor visits between those who are younger than age 50 and those who are older than 50 who got the lottery and apply for medicaid on average is -0.2083.

Additionally, it is not statistically significant since 0.6657 > .05. We fail to reject the null that Beta3 = 0.

Compared to the problem set 3, question 2: the coefficient of age>=50 is not statistically significant similar to what we have above. The standard error for both are different but not too far different.

For the third regression where a white person is treated, the coefficient is 0.6895

These can be interpreted to mean that holding all else constant, for those who are white, we expect that those who won the lottery to apply to medicaid to mean that an increase in the number of doctor visits by 0.8243 (0.1348 + 0.6895). So the difference in the number of doctor visits between those who are white and not white who got the lottery and apply for medicaid on average is 0.6895.

Additionally, it is not statistically significant since 0.137087 > .05. We fail to reject the null that Beta3 = 0.

Compared to the problem set 3, question 2: the coefficient of race_white == 1 is statistically significant different to what we have above.

For the fourth regression where those with a health baseline are treated, the coefficient is -0.2218

These can be interpreted to mean that holding all else constant, for those who have a health baseline, we expect that those who won the lottery to apply to medicaid to mean that an increase in the number of doctor visits by 0.4218 (0.6436 + -0.2218).So the difference in the number of doctor visits between those who have a health baseline and those who do not have a health baseline who got the lottery and apply for medicaid on average is -0.2218.

Additionally, it is not statistically significant since 0.6459 > .05. We fail to reject the null that Beta3 = 0.

Compared to the problem set 3, question 2: the coefficient of health_baseline == 1 is not statistically significant similar to what we have above.

Comparing these results with Problem Set #3, Q2, we are essentially investigating whether the treatment effects vary across different demographic or health-related groups. If the beta3 coefficients are significantly different from zero, it suggests that the impact of treatment (Medicaid coverage) on the outcome (visiting a doctor) is different for the specified subgroups. This analysis allows us to explore potential heterogeneity in treatment effects based on these characteristics.

The results I got for Problem set 3, Q2: For female: 0.7926 Age >= 50: 0.4097 white person: 0.8067 health baseline: 0.4180

The numbers are drastically different. In fact the age >= 50 and health baseline are negative.

```
#2
mean_ever_medicaid_treated <- mean(ohie$ever_medicaid[ohie$treated == 1],
                                   na.rm = TRUE)
mean_ever_medicaid_control <- mean(ohie$ever_medicaid[ohie$treated == 0],
                                   na.rm = TRUE)
mean_ever_medicaid_treated
```

```
## [1] 0.4255519
```

```
mean_ever_medicaid_control
```

```
## [1] 0.1857241
```

The mean for those who won the lottery and took up medicaid is 0.4255519. Out of all of those who won the lottery, approximately 42.56% enrolled in Medicaid.

The mean for those who lost the lottery who take up medicaid is 0.1857241. Out of all of those who lost the lottery, approximately 18.57% enrolled in Medicaid.

However, the variable ever_medicaid reports whether someone actually enrolled in Medicaid coverage. While winning the lottery may increase the likelihood of enrolling in Medicaid, it doesn't guarantee that everyone who won the lottery actually enrolled, as demonstrated by the mean of ever_medicaid for those who won the lottery.

Therefore, the causal effect of winning the Medicaid lottery (beta1 in the regression) might not be the same as the causal effect of actually receiving Medicaid, as not everyone who won the lottery took up Medicaid, and some people who lost the lottery might still enroll in Medicaid through other means. The interpretation of the coefficient on treated in the regression model is specific to the random assignment process and may not fully capture the complexities of Medicaid enrollment.

```
#3
first_stage <- lm(ever_medicaid ~ treated + numhh_list, data = ohie)
summary(first_stage)
```

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

```
## lm(formula = ever_medicaid ~ treated + numhh_list, data = ohie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4394 -0.3914 -0.1951  0.5606  0.8529
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.243181   0.012650  19.224  < 2e-16 ***
## treated      0.244275   0.008133  30.033  < 2e-16 ***
## numhh_list  -0.048041   0.009380  -5.121 3.08e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4467 on 12226 degrees of freedom
## Multiple R-squared:  0.06897,    Adjusted R-squared:  0.06882
## F-statistic: 452.9 on 2 and 12226 DF,  p-value: < 2.2e-16
```

3.The coefficient for treated is 0.244275 which indicates that individuals who won the Medicaid lottery are estimated to have a .244275 higher probability enrolling in Medicaid versus those who did not win the lottery.

alpha1 is the coefficient associated with the variable, treated. The interpretation of alpha1 is the estimated effect of being treated, winning the medicaid lottery, on the likelihood of enrolling in medicaid. Since alpha1 is positive and statistically significant, it would suggest that winning the lottery increases the likelihood of enrolling in Medicaid. Alpha1 is statistically significant,it implies that winning the lottery does not have an significant effect on Medicaid enrollment. In summary, the coefficient alpha1 represents the estimated casual effect of being treated on the likelihood of enrolling in Medicaid, controlling for the number of household members

4. For instrument relevance, the assumption for this test is that it is relevant to ever_medicaid. To test this, we can check whether there is a statistically significant relationship between treated and ever_medicaid in the first-stage regression. In this case, we look at the significance of the coefficient. If it is statistically significant, it suggests that treated is relevant for predicting ever_medicaid. Since it is statistically significant, we check that that treated does fit this.

For independence assumption, treatment is not correlated with the error term in the main regression equation. Unfortunately, there is no direct statistical test for the independence assumption. So it is often rely on the plausibility of the random assignment or use additional covariates to address potential confounding factors. If there is a strong theoretical support for the randomness of the instrument assignment, and there are no omitted variables caused correlation between the instrument and the error term. There is random assignment and omitted variable bias is taken into account through control variables. However, it does fail the test since independence is not met for household size.

For exclusion restriction, the assumption is that the treatment affects ever_medicaid only through its impact on the treated treatment variable and not through any other channels. Again, there is not direct statistical test for this assumption. It relies on theoretical justification and understanding of the context. If there are plausible pathways through which the instrument affects the ever_medicaid, the exclusion restriction may be violated. Since individuals who didn’t win the lottery also got medicaid, it showed that treated has no direct effect on count_visit_dr on ever_medicaid. Treated should not have an impact on doctor visits but impact doctor visits through medicaid enrollment.

In practice, researchers often present these assumption in an interesting narrative, drawing on theoretical and empirical evidence. Diagnostic test for heteroscedasticity and other issues in the residuals of the first-stage regression may be considered. It is important to note that instrumental variable analysis relies on assumptions that cannot be definitely proven; rather, researchers must provide evidence supporting the validity of these assumptions based on context and available data.

```r
#5
ohie_no_missing <- ohie[!is.na(ohie$count_visit_dr), ]
first_stage <- lm(ever_medicaid ~ treated + numhh_list, data = ohie_no_missing)
first_stage_coeff_treated <- coef(first_stage)["treated"]
reduced_form <- lm(count_visit_dr ~ treated + numhh_list, data = ohie_no_missing)
reduced_form_coeff_treated <- coef(reduced_form)["treated"]
ratio <- reduced_form_coeff_treated / first_stage_coeff_treated
ratio
```

```
##  treated
## 2.437223
```

The ratio of the reduce form to the first stage is 2.437. Since the ratio is not close to 1, it suggest that the treated did not have a similar impact on the outcome variable (count_visit_dr) in both the reduced form and the first stage. This means that it could indicate potential issues with the validity of the instrument or the presence of confounding factors.

```r
#6
ohie_no_missing <- ohie[!is.na(ohie$count_visit_dr), ]
first_stage <- lm(ever_medicaid ~ treated + numhh_list, data = ohie_no_missing)
ohie_no_missing$predicted_ever_medicaid <- predict(first_stage)

second_stage <- lm(count_visit_dr ~ predicted_ever_medicaid + numhh_list,
                   data = ohie_no_missing)
ratio <- reduced_form_coeff_treated / first_stage_coeff_treated
iv_estimate <- coef(second_stage)["predicted_ever_medicaid"]
cat("Previous Estimate (Ratio of Reduced Form to First Stage):", ratio, "\n")
```

```
## Previous Estimate (Ratio of Reduced Form to First Stage): 2.437223
```

```r
cat("2SLS Estimate:", iv_estimate, "\n")
```

```
## 2SLS Estimate: 2.437223
```

```r
summary(second_stage)
```

```
##
## Call:
## lm(formula = count_visit_dr ~ predicted_ever_medicaid + numhh_list,
##     data = ohie_no_missing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -6.760  -5.760  -3.760  -0.028 137.834
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               7.7098     0.4417  17.453  < 2e-16 ***
## predicted_ever_medicaid   2.4372     0.8893   2.740  0.00614 **
## numhh_list               -2.0189     0.2487  -8.118 5.19e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.86 on 12155 degrees of freedom
## Multiple R-squared:  0.006276,   Adjusted R-squared:  0.006113
## F-statistic: 38.38 on 2 and 12155 DF,  p-value: < 2.2e-16
```

Previous Estimate (Ratio of Reduced Form to First Stage): 2.437223 2SLS Estimate: 2.437223 It is the same as the previous estimate.

With the estimated coefficient is 2.4372 indicates that on average for a one-unit increase in predicted_ever_medicaid, count_visit_dr is expected to increase by 2.4372 units, holding other variables constant with the impact of enrollment is conditional on winning lottery.

The p-value of ‘predicted_ever_medicaid’ is .00614 since it is below .05, it suggests that 2SLS estimate is statistically significant. the 2SLS estimate provides a consistent and unbiased estimate of the causal effect under the assumptions of the instrumental variable method. Since the instrument is valid, the 2SLS estimate can be more reliable than the OLS estimate.

```r
#7
ohie_no_missing <- ohie[!is.na(ohie$count_visit_dr), ]
first_stage <- lm(ever_medicaid ~ treated + numhh_list + female + age + I(age^2)
                  + race_white + hs_degree + college_degree + health_baseline,
                  data = ohie_no_missing)
ohie_no_missing$predicted_ever_medicaid <- predict(first_stage,
                                                   newdata = ohie_no_missing)
second_stage <- lm(count_visit_dr ~ predicted_ever_medicaid + numhh_list + female
                   + age + I(age^2) + race_white + hs_degree + college_degree +
                     health_baseline, data = ohie_no_missing)
summary(second_stage)
```

```
##
## Call:
## lm(formula = count_visit_dr ~ predicted_ever_medicaid + numhh_list +
##     female + age + I(age^2) + race_white + hs_degree + college_degree +
##     health_baseline, data = ohie_no_missing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.446  -5.264  -3.140   0.394 140.101
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.9461222  1.4527176  -1.340 0.180387
## predicted_ever_medicaid  2.4246521  0.8715028   2.782 0.005408 **
## numhh_list              -1.5734505  0.2491363  -6.316 2.78e-10 ***
## female                   2.4931789  0.2313475  10.777  < 2e-16 ***
## age                      0.2633161  0.0695231   3.787 0.000153 ***
## I(age^2)                -0.0030468  0.0008351  -3.649 0.000265 ***
## race_white               1.2373348  0.2376836   5.206 1.96e-07 ***
## hs_degree                1.0377172  0.2774203   3.741 0.000184 ***
## college_degree           2.2541510  0.4070297   5.538 3.12e-08 ***
## health_baseline          2.2871899  0.2584075   8.851  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 11.67 on 12115 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.03367,    Adjusted R-squared:  0.03295
## F-statistic:  46.9 on 9 and 12115 DF,  p-value: < 2.2e-16
```

The estimated intercept is -1.946 which represents the expected count of doctor visits when all other variables are zero impact of enrollment is conditional on winning lottery. For predicted-ever-medicaid, the coefficient is 2.425 which indicates the estimated casual effect of ever having Medicaid on the count of doctor visits. This is a slightly lower coefficient estimate from Q6.

This effect is statistically significant because the p-value is less than .05. For the model in 6, the intercept is 7.71 while the coefficient of predicted-ever-medicaid is 2.4372. This coefficient is statistically significant since its p-value is 0.006 which is smaller than .05.

The standard error for this model is 0.8715 which is lower than the standard error in Q6. This hints that this model is more precise than the model in Q6.

```r
#8
install.packages("AER", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/5p/t64709x958n6ldzknn2t_tzr0000gn/T//Rtmpd92STR/downloaded_packages
```

```r
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.3.1

## Loading required package: car

## Loading required package: carData

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 4.3.1

## Loading required package: survival

## Warning: package 'survival' was built under R version 4.3.1
```

```
endline_outcomes <- c("count_visit_dr", "count_visit_er", "out_of_pocket_spend",
                      "health_score", "happy")

results_table <- data.frame(Endline_Outcome = character(),
                             IV_Estimate = numeric(),
                             Standard_Error = numeric(),
                             stringsAsFactors = FALSE)

for (outcome in endline_outcomes) {
  print(sum(!complete.cases(ohie[c("ever_medicaid", "treated", "numhh_list",
                                   "female", "age", "race_white", "hs_degree",
                                   "college_degree", "health_baseline", outcome)])))

  ohie_subset <- ohie[complete.cases(ohie[c("ever_medicaid", "treated",
                                            "numhh_list", "female", "age",
                                            "race_white", "hs_degree",
                                            "college_degree", "health_baseline"
                                            , outcome)]), ]

  first_stage <- lm(ever_medicaid ~ treated + numhh_list + female + age +
                      I(age^2) + race_white + hs_degree + college_degree +
                      health_baseline, data = ohie_subset)

  ohie_subset$predicted_ever_medicaid <- predict(first_stage,
                                                 newdata = ohie_subset)

  iv_model <- ivreg(as.formula(paste(outcome, "~ predicted_ever_medicaid +
                                     numhh_list + female + age + I(age^2) +
                                     race_white + hs_degree + college_degree +
                                     health_baseline")), data = ohie_subset)

  iv_estimate <- coef(iv_model)[2]
  se <- summary(iv_model)$coef[2, "Std. Error"]

  results_table <- rbind(results_table, data.frame(Endline_Outcome = outcome,
                                                   IV_Estimate = iv_estimate,
                                                   Standard_Error = se,
                                                   stringsAsFactors = FALSE))
}
```

```
## [1] 104
## [1] 87
## [1] 117
## [1] 2837
## [1] 59
```

```
print(results_table)
```

```
##                              Endline_Outcome   IV_Estimate Standard_Error
## predicted_ever_medicaid       count_visit_dr    2.42465212    0.871502798
## predicted_ever_medicaid1      count_visit_er    0.04318939    0.143205790
## predicted_ever_medicaid2 out_of_pocket_spend -283.03827202   88.242983581
## predicted_ever_medicaid3        health_score   -0.00230438    0.003957147
## predicted_ever_medicaid4                happy    0.02651425    0.031317980
```

For the precision of the the estimates, it is indicate by the standard errors. Smaller standard errors suggest more precised estimates. For count_visit_dr and count_visit_er, the standard errors are relatively small, meaning a higher precision in the estimating the effects. However, out_of_pocket_spend has a relatively large standard error which means less precision in its estimating the effect.

A positive IV estimate for count_visit_dr suggests an increase in the number of doctor visits for those receiving health insurance. Negative IV estimate for out_of_pocket_spend implies a decrease in out-of-pocket spending for those with health insurance.