marshad_HW 5

2023-03-01

```
setwd("/Users/misbaharshad/Downloads/")
data <- read.csv("MTO (2).csv")
library(AER)
```

```
## 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
```

```
## Loading required package: survival
```

```
#1/a
reg <- lm(working_end ~ moved_using_voucher_end, data = data)
summary(reg)
```

```
##
## Call:
## lm(formula = working_end ~ moved_using_voucher_end, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6121 -0.4610  0.3878  0.5390  0.5390
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.46100    0.01066  43.239   <2e-16 ***
## moved_using_voucher_end   0.15115    0.01823   8.289   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4948 on 3271 degrees of freedom
## Multiple R-squared:  0.02058,    Adjusted R-squared:  0.02028
## F-statistic: 68.72 on 1 and 3271 DF,  p-value: < 2.2e-16
```

```r
#The control group (i.e. didn't have a voucher) had an average likelihood of being
#employed at 46%, and the voucher increased that likelihood by 15 percentage points
#on average. The coefficient is statistically significant at the 5% level and hence
#we reject the null that the voucher has no effect on employment at the end of the
#study. However, even though we previously determined the characteristics are
#balanced, OLS doesn't guarantee a causal effect.

#2/b
reg_HS <- lm(working_end ~ moved_using_voucher_end + high_school_grad_start,
             data = data)
summary(reg_HS)
```

```
##
## Call:
## lm(formula = working_end ~ moved_using_voucher_end + high_school_grad_start,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00272 -0.22168 -0.00272  0.07792  0.77832
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.22168    0.00899  24.659  < 2e-16 ***
## moved_using_voucher_end    0.08064    0.01339   6.021 1.93e-09 ***
## high_school_grad_start     0.70040    0.01312  53.402  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3617 on 3270 degrees of freedom
## Multiple R-squared:  0.4768, Adjusted R-squared:  0.4765
## F-statistic:  1490 on 2 and 3270 DF,  p-value: < 2.2e-16
```

```r
#Once we control for high school, the control group has an average likelihood of
#being employed at 22% and the voucher increased that likelihood by 8 percentage
#points on average at the end of the study. The coefficient is still statistically
#significant at the 5% level (i.e. we still reject the null that the voucher has no
#effect on employment at the end of the study). However, the fact that the estimate
#decreased from 46% to 22% and 15pp to 8 pp suggests omitted variable bias in the
#previous regression.

#Without controlling for the confound, the outcome variable in the first regression
#captured both the estimate of the treatment (the voucher) and the effect of the
#omitted variable (high school). The fact the our second regression significantly
#decreased suggests we overestimated the treatment effect in the analysis of (a).

#3/c
#The independence assumption requires an instrument to be "as good as randomly
#assigned" meaning that it is not related to the omitted variables. Using treated
#satisfies the independence assumption because receiving a voucher is random, and
#thus is not influenced by potential confounding variables that may lead to omitted
#variable bias. Another way of explaining this is the instrument is not related to
#anything in the error term.
```

```r
#The exclusion restriction means there is a single channel through which the
#instrument affects the outcome. This means the instrument ("treated") can't
#directly affect the outcome (i.e. everyone who receives the voucher automatically moved), otherwise we
#of events. By ensuring the instrument passes through the explanatory to the outcome
#is a single channel, we ensure we are observing the treatment effect and not an external variable. The
#moving (Y)  through the voucher (X). Z -> X -> Y

#4/d
reg4 <- lm(moved_using_voucher_end ~ treated, data = data)
summary(reg4)
```

```
##
## Call:
## lm(formula = moved_using_voucher_end ~ treated, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5244 -0.5244  0.0000  0.4756  0.4756
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.341e-15  1.195e-02    0.00        1
## treated      5.244e-01  1.480e-02   35.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4034 on 3271 degrees of freedom
## Multiple R-squared:  0.2773, Adjusted R-squared:  0.2771
## F-statistic:  1255 on 1 and 3271 DF,  p-value: < 2.2e-16
```

```r
#Yes, the instrument satisfies the first stage assumption because the F-stat is
#1255 with a very small p-value, meaning that the instrument has strong
#joint-significance. It is relevant for the endogenous variable, because "treated"
#is highly correlated with "moving". The estimate being less than 1 can be
#explained by the fact that the instrument does not perfectly predict the
#treatment variable (i.e. receiving a voucher doesn't necessarily mean the person
#will move).

#5/e
reg5 <- lm(working_end ~ treated, data = data)
summary(reg5)
```

```
##
## Call:
## lm(formula = working_end ~ treated, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5162 -0.5108  0.4838  0.4892  0.4892
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.516242   0.014815  34.846   <2e-16 ***
## treated     -0.005464   0.018347  -0.298    0.766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5 on 3271 degrees of freedom
## Multiple R-squared:  2.712e-05,  Adjusted R-squared:  -0.0002786
## F-statistic: 0.08871 on 1 and 3271 DF,  p-value: 0.7658
```

```
#The ratio is the "reduced form" over the "first stage"
ratio <- -0.005464 / 0.524367385192118229
print(ratio)
```

```
## [1] -0.01042018
```

```
#6/f
iv_reg_6 <- ivreg(working_end ~ moved_using_voucher_end | treated, data = data)
summary(iv_reg_6)
```

```
##
## Call:
## ivreg(formula = working_end ~ moved_using_voucher_end | treated,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5162 -0.5162  0.4838  0.4838  0.4942
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.51624    0.01484  34.795   <2e-16 ***
## moved_using_voucher_end -0.01042    0.03504  -0.297    0.766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5007 on 3271 degrees of freedom
## Multiple R-Squared: -0.002935,   Adjusted R-squared: -0.003242
## Wald test: 0.08844 on 1 and 3271 DF,  p-value: 0.7662
```

```
#When we calculate the ratio and verify with the instrument regression in R, we
#get -0.01042 for both. We can interpret the local average treatment effect as
#testing the subpopulation of compliers (those who receive the treatment only
#because of the instrument and moved or didn't receive a voucher and didn't move)
#being not statistically different from the control group. Although the estimate
#is -0.01042 for both, it is not statistically significant. We fail to reject the
#null that the number employed after the survey is different between the treated
#(received a voucher) and control (didn't receive a voucher) groups.

#7/g
attach(data)
x2 <- cbind(male_respondant, black_respondant, high_school_grad_start, never_married_start,
            parent_before_18, verydissat_neighborhood_start, on_welfare_start,
```

```
            victim_of_crime_start, want_move_gangs_drugs_start)
Z <- treated

#first-stage
first_stage <- lm(moved_using_voucher_end ~ Z + x2)
summary(first_stage)
```

```
##
## Call:
## lm(formula = moved_using_voucher_end ~ Z + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67978 -0.42274 -0.00345  0.37827  0.63086
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.133358   0.019468  -6.850 8.78e-12 ***
## Z                                0.522178   0.014528  35.942  < 2e-16 ***
## x2male_respondant               -0.009931   0.058705  -0.169  0.86568
## x2black_respondant               0.047393   0.016023   2.958  0.00312 **
## x2high_school_grad_start        -0.053600   0.022204  -2.414  0.01583 *
## x2never_married_start            0.028797   0.022346   1.289  0.19759
## x2parent_before_18               0.043900   0.022594   1.943  0.05210 .
## x2verydissat_neighborhood_start  0.110258   0.022246   4.956 7.55e-07 ***
## x2on_welfare_start               0.026696   0.023630   1.130  0.25867
## x2victim_of_crime_start         -0.004449   0.022269  -0.200  0.84168
## x2want_move_gangs_drugs_start    0.033919   0.023832   1.423  0.15475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3952 on 3262 degrees of freedom
## Multiple R-squared:  0.3083, Adjusted R-squared:  0.3062
## F-statistic: 145.4 on 10 and 3262 DF,  p-value: < 2.2e-16
```

```
#The estimate in (d) was 0.5243 and now when we include the baseline characteristics
#it goes down slightly to 0.5221. This affirms that our instrument satisfies the
#exclusion restriction because the instrument is only affecting our outcome through
#1 channel. If the adding the controls changed the estimate significantly then we
#violate the exclusion restriction because the instrument would be impacting the
#controls.

#8/h
full_reg_iv <- ivreg(working_end ~ moved_using_voucher_end + x2 | Z + x2)
summary(full_reg_iv, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = working_end ~ moved_using_voucher_end + x2 |
##     Z + x2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.02151 -0.09245 -0.01103  0.05277  0.98983
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       0.01103    0.01353   0.815  0.41505
## moved_using_voucher_end          -0.01939    0.02097  -0.924  0.35537
## x2male_respondant                -0.04090    0.04426  -0.924  0.35550
## x2black_respondant               -0.06380    0.01215  -5.251 1.61e-07 ***
## x2high_school_grad_start          0.27538    0.01677  16.425  < 2e-16 ***
## x2never_married_start             0.24700    0.01684  14.671  < 2e-16 ***
## x2parent_before_18                0.04510    0.01705   2.645  0.00821 **
## x2verydissat_neighborhood_start   0.18561    0.01699  10.926  < 2e-16 ***
## x2on_welfare_start                0.08233    0.01782   4.621 3.97e-06 ***
## x2victim_of_crime_start           0.15743    0.01679   9.378  < 2e-16 ***
## x2want_move_gangs_drugs_start     0.08142    0.01798   4.529 6.13e-06 ***
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 3262  1291.862 < 2e-16 ***
## Wu-Hausman          1 3261     7.029 0.00806 **
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2979 on 3262 degrees of freedom
## Multiple R-Squared: 0.646,   Adjusted R-squared: 0.6449
## Wald test: 596.5 on 10 and 3262 DF,  p-value: < 2.2e-16
```

```
#The coefficient changed slightly, form -0.0104 to -0.01939, both are negative
#numbers suggesting that the voucher isn't correlated with better employment outcomes
#at the end of the survey but neither are statistically significant. The standard
#error decreased slightly from 0.03 to 0.02 as a results of greater precision from
#adding the baseline characteristics. The regression with the baselines is slightly
#better but doesn't alter our conclusion in any significant way from (f).

##############################################
#Question 2

#a
#False, median income doesn't satisfy the exclusion restriction in order to be a valid good instrument.

#b
#False, precision falls compared to OLS. The IV estimator uses only a subset if
#the variation in the explanatory variable to estimate the relationship between
#the explanatory variable and the outcome variable.  By reducing the sample size
#we lose precision.

#c
#False IF nobody at Harris has low undergraduate grades to begin with. If nobody
#has low undergrad grades to begin with, it is not selection bias to not include
#students with low undergrad grades. However, it is true if she is singling out
#only students with non-low undergrad grades. This will skew the results, and lead
#to sample selection bias.
```

R Markdown