Econometrics Final

Yuhuan Huang

2025-03-10

```
library(AER)
library(dplyr)
library(broom)
data <- read.csv("draft.csv", header = TRUE)
```

Q1

```
model <- lm(lwage ~ vet, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = lwage ~ vet, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7797 -0.2645  0.0987  0.4067  2.0992
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.436395   0.007867 691.047   <2e-16 ***
## vet         -0.020522   0.016717  -1.228     0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6976 on 10099 degrees of freedom
## Multiple R-squared:  0.0001492,  Adjusted R-squared:  5.02e-05
## F-statistic: 1.507 on 1 and 10099 DF,  p-value: 0.2196
```

From the regression result, we can see that the coefficient of vet is -0.0205. Therefore, we can say that the average effect of being a veteran on log wage is -0.0205, given other things the same. However, with a p-value of 0.22 (larger than 0.05), and t-value of -1.228 (whose absolute value is smaller than 1.96), we can say that this effect is not statistically significant.

```
# Check endogeneity
residuals_ols <- resid(model)
cor_test <- cor.test(residuals_ols, data$vet)
print(cor_test)
```

```
##
##  Pearson's product-moment correlation
##
## data:  residuals_ols and data$vet
## t = 7.7591e-16, df = 10099, p-value = 1
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01950183  0.01950183
## sample estimates:
##          cor
## 7.720952e-18
```

```r
# Hausman test
iv_model <- ivreg(lwage ~ vet | elig, data = data)

hausman_test <- function(ols, iv) {
  # coefs
  b_ols <- coef(ols)
  b_iv <- coef(iv)

  # covarians matrix
  V_ols <- vcov(ols)
  V_iv <- vcov(iv)

  #
  diff_beta <- b_iv - b_ols
  V_diff <- V_iv - V_ols

  #
  stat <- t(diff_beta) %*% solve(V_diff) %*% diff_beta
  p_value <- pchisq(stat, df = length(b_ols) - 1, lower.tail = FALSE)

  return(list(statistic = stat, p_value = p_value))
}

result <- hausman_test(model, iv_model)
print(result)
```

```
## $statistic
##          [,1]
## [1,] 2.882534
##
## $p_value
##            [,1]
## [1,] 0.08954498
```

When the regression has endogeneity problem, or breaking the GM_3 assumption $E[vet'\,\varepsilon_i] = 0$, it wouldn't be consistent.

However, as I applied residual check and Hausman test based on IV, we can see that there is no endogeneity problem. Therefore, the estimator is already consistent.

Q2

The compliers are those individuals: If drafted by the lottery (draft eligible), then they serve the army (are veteran); If they are not drafted by the lottery (draft not eligible) , then they don't serve the army (are not veteran).

The always-takers are those individuals: Whether or not they are drafted by the lottery (whether or not draft eligible), they always serve the army (are veteran).

The never-takers are those individuals: Whether or not they are drafted by the lottery (whether or not draft eligible), they never serve the army (are not veteran).

The defiers are those individuals: If drafted by the lottery (draft eligible), then they don't serve the army (are not veteran); If they are not drafted by the lottery (draft not eligible), then they serve the army (are veteran).

Q3

The monotonicity assumption here means there is no defiers: If drafted by the lottery, it would only make individuals more likely (willing) to serve the army (become a veteran), and would not affect their decision in an opposite way.

Q4

Since the given IV should satisfy the instrument relevance and exclusion restriction to be a valid IV (under the "first-stage" and exclusion restriction assumption), and under the monotonicity assumption (which has been talked about in Q3), and under the random assignment assumption, we can use the LATE theorem. The Wald estimator is the LATE:

```
#first stage
first_stage <- lm(vet ~ elig, data=data)
summary(first_stage)  # Check whether Z influence D
```

```
##
## Call:
## lm(formula = vet ~ elig, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3111 -0.1874 -0.1874 -0.1874  0.8126
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.187406   0.004810   38.96   <2e-16 ***
## elig        0.123745   0.009169   13.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4116 on 10099 degrees of freedom
## Multiple R-squared:  0.01772,    Adjusted R-squared:  0.01762
## F-statistic: 182.1 on 1 and 10099 DF,  p-value: < 2.2e-16
```

```r
#second stage
late_model <- ivreg(lwage ~ vet | elig, data = data)
summary(late_model)
```

```
##
## Call:
## ivreg(formula = lwage ~ vet | elig, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8269 -0.2763  0.1015  0.4081  2.2651
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4836     0.0289 189.761   <2e-16 ***
## vet          -0.2336     0.1266  -1.845   0.0651 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7032 on 10099 degrees of freedom
## Multiple R-Squared: -0.01593,    Adjusted R-squared: -0.01603
## Wald test: 3.404 on 1 and 10099 DF,  p-value: 0.06506
```

```r
late_estimate <- coef(late_model)["vet"]
cat("Estimated LATE:", late_estimate, "\n")
```

```
## Estimated LATE: -0.2335831
```

The estimated LATE is -0.2335. It is different in value from what we get in question Q1. I think the problem is that, when we directly do the OLS regression, what we get is the ATT, which is on the individuals who take the however here what we get is the LATE

Q5

From our lecture topic 7, we know that LATE can be used to deal with the IV heterogeneity problem. Therefore, I applied to a cross-term of *vet* and *yob*, and *elig* and *yob*.

$$vet * yob_i = \pi_0 + \pi_1 elig * yob_i + \nu_i$$

$$lwage_i = \alpha + \delta \cdot \hat{vet * yob_i} + \epsilon_i$$

```r
#1.Use LATE for interaction term:

data$vet_yob <- data$vet * data$yob
data$elig_yob <- data$elig * data$yob

# First stage regression: Predict D using instrument Z, allowing for birth year interaction
multi_first_stage <- lm(vet ~ elig_yob, data = data)
summary(multi_first_stage)
```

```
##
## Call:
## lm(formula = vet ~ elig_yob, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3131 -0.1880 -0.1880 -0.1880  0.8120
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.1880003  0.0048113   39.08   <2e-16 ***
## elig_yob    0.0023595  0.0001779   13.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4117 on 10099 degrees of freedom
## Multiple R-squared:  0.01711,    Adjusted R-squared:  0.01702
## F-statistic: 175.8 on 1 and 10099 DF,  p-value: < 2.2e-16
```

```
# Second stage regression: Estimate heterogeneous LATE using 2SLS
multi_late_model <- ivreg(lwage ~ vet_yob | elig_yob, data = data)
summary(multi_late_model)
```

```
##
## Call:
## ivreg(formula = lwage ~ vet_yob | elig_yob, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8298 -0.2771  0.1025  0.4074  2.2741
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.486512   0.029353 186.916   <2e-16 ***
## vet_yob     -0.004813   0.002510  -1.918   0.0552 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7038 on 10099 degrees of freedom
## Multiple R-Squared: -0.01764,    Adjusted R-squared: -0.01774
## Wald test: 3.677 on 1 and 10099 DF,  p-value: 0.05519
```

```
late_estimate0 <- coef(multi_late_model)["vet_yob"]
cat("Estimated LATE:", late_estimate0, "\n")
```

```
## Estimated LATE: -0.00481332
```

We can also regress on both vet, yob and the interaction term of vet and yob:

```
#2. Use LATE for both vet, yob and the interaction term
# First stage regression: Predict D using instrument Z, allowing for birth year interaction
first_stage1 <- lm(vet~ elig * yob, data = data)
summary(first_stage1)  # Check instrument strength
```

```
##
## Call:
## lm(formula = vet ~ elig * yob, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4070 -0.2432 -0.1697 -0.1329  0.8671
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.082570   0.222166   9.374  < 2e-16 ***
## elig         1.456318   0.421799   3.453 0.000557 ***
## yob         -0.036787   0.004311  -8.532  < 2e-16 ***
## elig:yob    -0.025852   0.008184  -3.159 0.001589 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4085 on 10097 degrees of freedom
## Multiple R-squared:  0.03246,    Adjusted R-squared:  0.03217
## F-statistic: 112.9 on 3 and 10097 DF,  p-value: < 2.2e-16
```

```r
# Second stage regression: Estimate heterogeneous LATE using 2SLS
late_model1 <- ivreg(lwage ~ vet * yob | elig * yob, data = data)
summary(late_model1)
```

```
##
## Call:
## ivreg(formula = lwage ~ vet * yob | elig * yob, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8345 -0.2745  0.1007  0.4057  2.2307
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.752513   1.933168   2.976  0.00293 **
## vet          8.137423   9.076509   0.897  0.36999
## yob         -0.005226   0.037572  -0.139  0.88938
## vet:yob     -0.163219   0.177594  -0.919  0.35809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7039 on 10097 degrees of freedom
## Multiple R-Squared: -0.01773,    Adjusted R-squared: -0.01803
## Wald test: 8.721 on 3 and 10097 DF,  p-value: 8.955e-06
```

```r
# Extract LATE estimates by birth year
late_estimates <- coef(late_model1)
print(late_estimates)
```

```
##  (Intercept)          vet          yob      vet:yob
##  5.752513013  8.137423485 -0.005226189 -0.163218669
```

Then we calculate the weighted average:

```
# Identify treatment effect coefficients (interaction terms "vet:yob")
late_coef_indices <- grep("vet:yob", names(late_estimates))  # Extract relevant coefficients
late_values <- late_estimates[late_coef_indices]  # LATE estimates by birth year
# Define weights (ensure they are positive and sum to 1)
# Example: Use sample size for each birth year as weights
weights <- table(data$yob)  # Count observations per birth year
weights <- weights[names(late_values)]  # Match with available coefficients
weights <- weights / sum(weights)  # Normalize to sum to 1
# Compute weighted LATE
weighted_LATE <- sum(weights * late_values, na.rm = TRUE)
```

Also, we can use group-by regression to deal with the heterogeneity.

$$lwage = \beta_{0,yob} + \beta_{1,yob} \cdot vet + \varepsilon$$

```
# 3.Group by regression
group_results <- data %>%
  group_by(yob) %>%
  do(tidy(lm(lwage ~ vet, data = .))) %>%
  filter(term == "vet")

print(group_results)
```

```
## # A tibble: 4 x 6
## # Groups:   yob [4]
##     yob term  estimate std.error statistic  p.value
##   <int> <chr>    <dbl>     <dbl>     <dbl>    <dbl>
## 1    50 vet    0.0349     0.0317     1.10  0.270
## 2    51 vet   -0.00762    0.0332    -0.229 0.819
## 3    52 vet   -0.0573     0.0336    -1.71  0.0880
## 4    53 vet   -0.125      0.0370    -3.38  0.000744
```

And we weighted them with the proportion of year of birth.

```
yob_weights <- data %>%
  group_by(yob) %>%
  summarise(weight = n()/nrow(data))

# weighted average
weighted_avg_effect <- sum(group_results$estimate * yob_weights$weight)
print(weighted_avg_effect)
```

```
## [1] -0.03972542
```

Q6

In the IV that allows for parameter heterogeneity, we assume that:

- Random assignment, Z is random assigned
- Exclusion restriction, Z can only influence y through D
- Monotonicity $D_1 \geq D_0$
- First stage $E[D_1 - D_0] \neq 0$

In this case, since Z is no longer completely random with i, random assignment is violated, and the validity of LATE theorem is influenced. Also, when November and December births were systematically assigned to lower draft numbers, it is a possible that, the effect of being drafted is correlated with the effect of be born in different time of the year, which has something directly to do with adulthood wages (like some careers in sports may be affeted by month to be born due to the timing of the competitions)