ECON 7710
Homework 5: Instrumental Variables

General instructions

The `.Rmd` source for this document will be the template for your homework submission. You must submit your completed assignment as a single html document saved as a PDF file, uploaded to eLC by **11:59p on March 27** using the filename `sectiontime_teamnumber_hw1.html` (e.g. `935_1_hw5.html`).

Notes:

- Include the name of each teammate under `author` in the `yaml` .
- Do not alter the formatting code in this template.
- For questions requiring analytical solutions, you can type them in using markdown math code. Or, you can submit handwritten solutions, embedding them in the knitted document as clearly readable images.
- For questions requiring computation, some or all of the required code is included in associated chunks. Modify chunks where and how you are directed.
- For (almost) all questions about R Markdown, consult The Definitive Guide (https://bookdown.org/yihui/rmarkdown/).
- The `setup` chunk above indicates the packages required for this assignment.
- You will find a description of the variables in the referenced dataset through the Help tab in the Plot pane of RStudio.
- **Switch `eval` to `TRUE` in the global options command to execute code chunks**.

In this homework assignment, you will use data from Angrist and Krueger 1991 to estimate the effects of years of schooling on labor market outcomes. This paper uses quarter of birth as an instrument for years of schooling since students born in the first quarter of the year will have attended fewer years of school when they hit the compulsory schooling age.

```
data <- read.csv("hw5_data.csv")
str(data)
```

```
## 'data.frame':    816435 obs. of  9 variables:
##  $ education       : int  12 12 12 16 14 12 12 12 12 7 ...
##  $ log_wage        : num  6.25 5.85 6.65 6.71 6.36 ...
##  $ married         : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ mid_atlantic    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mountain        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ new_england     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ quarter_of_birth: int  1 4 1 1 4 4 1 1 3 2 ...
##  $ year_of_birth   : int  33 33 30 33 37 35 38 30 39 36 ...
##  $ first_quarter   : int  1 0 1 1 0 0 1 1 0 0 ...
```

1. First calculate the OLS estimate of the relationship between years of schooling and log wages. Why might you not trust this estimate?

```
# insert command to run linear regression with robust standard errors here

model_ols <- lm_robust(log_wage ~ education, data = data)


summary(model_ols)
```

```
##
## Call:
## lm_robust(formula = log_wage ~ education, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper     DF
## (Intercept)  5.05223  0.0035331  1430.0        0  5.04531  5.05916 816433
## education    0.05937  0.0002578   230.3        0  0.05887  0.05988 816433
##
## Multiple R-squared:  0.07905 ,   Adjusted R-squared:  0.07905
## F-statistic: 5.303e+04 on 1 and 816433 DF,  p-value: < 2.2e-16
```

The estimate might not be trusted because of factors such as OVB, Reverse Causality, and Sample Selection. For OVB, variables like ability or family background/upbringing that influence both education and earnings are not included. Reverse Causality would be where individuals might pursue more education because they anticipate higher earnings. In this situation, Sample Selection would be where the sample may only include employed individuals. This omits those with different educational and employment characteristics

2. Now, you decide you use an instrumental variables approach and instrument for education using quarter of birth. What assumptions do you need for this to be a valid strategy?

Crucial assumptions include relevance, exclusion, and monotonicity. Relevance ensures quarter of birth significantly influences educational attainment, as it dictates school start age which then affects years of schooling. The exclusion assumption states that quarter of birth impacts log wages ONLY through education. Monotonicity means the effect of quarter of birth on education moves in a single direction, uniformly affecting all individuals. This ensures that the instrumental variable provides a valid estimate of education's causal impact on wages.

3. Write down and estimate the first stage equation. Report the main coefficient estimate in a sentence. Store your predicted values.

```
# insert command to estimate the first stage regression here
first_stage <- lm(education ~ quarter_of_birth, data = data)
educ_hat <- predict(first_stage)

summary(first_stage)
```

```
##
## Call:
## lm(formula = education ~ quarter_of_birth, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.317  -1.317  -1.177   2.683   6.823
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      13.129787   0.008579 1530.49   <2e-16 ***
## quarter_of_birth  0.046878   0.003109   15.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.13 on 816433 degrees of freedom
## Multiple R-squared:  0.0002784,  Adjusted R-squared:  0.0002772
## F-statistic: 227.4 on 1 and 816433 DF,  p-value: < 2.2e-16
```

The estimated coefficient for quarter_of_birth in the first stage regression is 0.046878. This indicates that each additional quarter of birth is associated with an increase of approximately 0.046878 years in educational attainment. According to its p-value, it is statistically significant.

4. Write down and estimate the reduced form equation. Report the main coefficient estimate in a sentence.

```
# insert command to estimate the reduced form regression here

reduced_form <- lm(log_wage ~ quarter_of_birth, data = data)


summary(reduced_form)
```

```
##
## Call:
## lm(formula = log_wage ~ quarter_of_birth, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1808 -0.2562  0.0830  0.3476  5.3864
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      5.8384932  0.0018119 3222.334   <2e-16 ***
## quarter_of_birth 0.0001305  0.0006566    0.199    0.842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.661 on 816433 degrees of freedom
## Multiple R-squared:  4.836e-08,  Adjusted R-squared:  -1.176e-06
## F-statistic: 0.03948 on 1 and 816433 DF,  p-value: 0.8425
```

This indicates that each additional quarter of birth is associated with an increase of approximately 0.0001305 in log wages, although this effect is not statistically significant because p-value = 0.842. This means the instrument may not have a direct impact on log wages independent of education.

5. Using your answers in 3) and 4), report the Wald estimate of the relationship between education and wages.

The Wald estimate is the difference in log wages between individuals born in different quarters divided by the difference in educational attainment born in different quarters, which is equal to .0001305/.046878 = .00278

6. To estimate this using 2SLS, write down the second stage equation. Estimate this using your predicted values from 3).

```
# insert command to estimate the second stage regression here
data$educ_hat <- educ_hat
second_stage <- lm_robust(log_wage ~ educ_hat, data = data)

summary(second_stage)
```

```
##
## Call:
## lm_robust(formula = log_wage ~ educ_hat, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper     DF
## (Intercept) 5.801951    0.18624  31.153 6.147e-213  5.43693  6.16698 816433
## educ_hat    0.002783    0.01406   0.198  8.431e-01 -0.02477  0.03034 816433
##
## Multiple R-squared:  4.836e-08 , Adjusted R-squared:  -1.176e-06
## F-statistic: 0.0392 on 1 and 816433 DF,  p-value: 0.8431
```

The second stage equation models the effect of predicted education on log wages. The equation looks like log_wage = 5.80195 + .00278 * educ_hat

7. Now, use the ivreg command to direct estimate this effect. Compare this estimate to your OLS estimate in 1).

```
iv = ivreg(log_wage ~ education | quarter_of_birth, data = data)
summary(iv)
```

```
##
## Call:
## ivreg(formula = log_wage ~ education | quarter_of_birth, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.19942 -0.25576  0.08127  0.34734  5.38989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.801951   0.184885  31.381   <2e-16 ***
## education   0.002783   0.013955   0.199    0.842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6586 on 816433 degrees of freedom
## Multiple R-Squared: 0.007237,    Adjusted R-squared: 0.007236
## Wald test: 0.03977 on 1 and 816433 DF,  p-value: 0.8419
```

Using ivreg we direct estimate the effect of education on log wages to be .00278 which is smaller than the OLS estimate and it is not significant in IV but is in OLS. This suggests that when we account for endogenity using quarter birth, the effect that education has on wages is weaker and less reliable.

8. Now, add control variables to your IV regression. How do these change your results?

```
# IV regression with control variables
iv2 <- ivreg(log_wage ~ education + married + mid_atlantic + mountain + new_england + first_quar
ter |
                quarter_of_birth + married + mid_atlantic + mountain + new_england + first_quart
er,
                data = data)

# Display the results
summary(iv2)
```

```
##
## Call:
## ivreg(formula = log_wage ~ education + married + mid_atlantic +
##     mountain + new_england + first_quarter | quarter_of_birth +
##     married + mid_atlantic + mountain + new_england + first_quarter,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.64082 -0.29908  0.06034  0.37916  5.30790
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.065246   0.315856  19.203   <2e-16 ***
## education     -0.034697   0.023895  -1.452   0.1465
## married        0.268464   0.002453 109.435   <2e-16 ***
## mid_atlantic   0.065872   0.006770   9.730   <2e-16 ***
## mountain       0.030040   0.015886   1.891   0.0586 .
## new_england    0.020063   0.011919   1.683   0.0923 .
## first_quarter -0.007607   0.002973  -2.559   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6916 on 816428 degrees of freedom
## Multiple R-Squared: -0.09472,    Adjusted R-squared: -0.09472
## Wald test:  3067 on 6 and 816428 DF,  p-value: < 2.2e-16
```

The addition of control variables like marital status, region, and birth quarter in the IV regression has changed the coefficients, particularly for education. The fact that education is no longer significant with these controls suggests that unobserved factors related to marriage, location, and birth timing may have been influencing the relationship between education and log wages.

9. The relationship between quarter of birth and years of schooling should only hold for those who are affected by r compulsory school laws. Estimate a regression that tests whether there is a relationship between quarter of birth and years of schooling for those who have at least a college education (16 years of schooling). What do you find?

```
data_16 = subset(data, education > 16)
# Regression of education on quarter of birth
reg_16 <- lm(education ~ quarter_of_birth, data = data_16)

# Display the results
summary(reg_16)
```

```
##
## Call:
## lm(formula = education ~ quarter_of_birth, data = data_16)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3564 -1.3335 -0.3411  0.6665  1.6665
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      18.364104   0.008295 2213.773   <2e-16 ***
## quarter_of_birth -0.007662   0.003000   -2.554   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.148 on 118432 degrees of freedom
## Multiple R-squared:  5.506e-05,  Adjusted R-squared:  4.662e-05
## F-statistic: 6.522 on 1 and 118432 DF,  p-value: 0.01066
```

The regression results show a small but statistically significant relationship between quarter of birth and years of schooling for individuals with at least a college education. The coefficient for quarter of birth is -0.007662, with a p-value of 0.0107, suggesting that individuals born later in the year tend to have slightly fewer years of schooling. However, the effect is minor, and the very low R-squared value (5.506e-05) indicates that quarter of birth explains only a tiny portion of the variation in years of schooling.

10. Someone now tells you that quarter of birth is related to parental income. Does this change how you feel about your IV analysis?

If quarter of birth is related to parental income, it raises concerns about the validity of the instrument because it violates the exclusion restriction. The exclusion assumption of IV is that the instrument affects the outcome only through its impact on education. If parental income also influences wages, then quarter of birth could affect wages through this channel, violating the exclusion restriction. This could bias the IV estimate, making it unreliable for estimating the true causal effect of education on wages.