PROBLEM SET 2

Due on Monday, February 26, 2024

I - INSTRUCTIONS

To successfully complete this problem set, please follow these steps:

1. Download this Word document file into your computer
2. Insert all your answers into this Word document. Guidance here on how to insert non-Word objects such as handwritten work or screenshot images in your answers.
3. **Once your document is complete, please save it as a PDF**. This is important to make sure all your work is preserved in the process of submission to Canvas.
4. Please submit an electronic copy of the PDF and your **replicable Stata or R script** to the Canvas assignment page.

II - IDENTIFICATION

(1) Your information

Your Last Name:	*Boochever*
Your First Name:	*Oscar*

(2) Group Members (please list the classmates you worked with on this problem set):

n/a

(3) Compliance with Harvard Kennedy School Academic Code[1] (mark with an X below)

	Yes	No
I certify that my work in this problem set complies with the Harvard Kennedy School Academic Code	X	

[1] We abide by the Harvard Kennedy School Academic code for all aspects of the course. In terms of problem sets, unless explicitly written otherwise, the norms are the following: You are free (and encouraged) to discuss problem sets with your classmates. However, you must hand in your own unique written work and code in all cases. Any copy/paste of another's work is plagiarism. In other words, you can work with your classmate(s), sitting side-by-side and going through the problem set question-by-question, but you must each type your own answers and your own code. For more details, please see syllabus.

For this problem set, we will be examining the methods used in the following paper:

> Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. *The Quarterly Journal of Economics*, *106*(4), 979-1014.

Conceptual Questions (40 points + 1 extra point)

1. Clearly state the primary research question that the author is trying to answer. Why should policymakers care about this question? (2 points)

The primary research question is revealed in the title: does compulsory school attendance affect educational attainment and earnings? Policymakers should care about this question because it has clear implications for education policies around compulsory schooling, in addition to suggesting the benefits of more schooling, which is policy relevant to early childhood education.

2. The authors used an instrumental variable approach because they believed a naïve regression specification (regressing earnings on education) would be insufficient. What are two possible confounders (omitted variables) that could bias the results from this regression? Explain the mechanism by which each omitted variable could bias the results and use the omitted variable bias formula to argue whether it would lead to an understatement or overstatement of the true effect. (3 points)

Naïve regression of earnings on education could omit certain variables that correlate both with education and earnings. For example, parental income and standardized test scores.

1. ***Income:*** *individuals with higher-income parents are likely to have higher future earnings, and likely attain more schooling.*

Earnings = a0 + a1Educationt_i + vi

Earnings = b0 + b1Education_i + b2Income_i + ei

Income_i = l + dEducation_i = n_i

*Bias: a1 = b1 + b2*d*

Educational attainment increases earnings so b1: ***positive***

Income increases earnings so b2: ***positive***

Positive correlation between educational attainment and parental income ***→ d: positive***

Since the effect of educational attainment would be positive, and bias is positive, omitting parental income in the naïve regression would lead to an ***overstatement*** *of the true effect of educational attainment.*

2. ***Standardized test scores:*** *individuals with higher test scores are likely to have higher future earnings, and likely attain more schooling.*

Earnings = a0 + a1Educationt_i + vi

Earnings = b0 + b1Education_i + b2Scores_i + ei

Scores_i = l + dEducation_i = n_i

*Bias: a1 = b1 + b2*d*

Educational attainment increases earnings so b1: ***positive***

Test scores increase earnings so b2: ***positive***

Positive correlation between educational attainment and standardized test scores ***→ d: positive***

Once again, since the effect of educational attainment would be positive, and bias is positive, omitting standardized test scores in the naïve regression would lead to an ***overstatement*** *of the true effect of educational attainment.*

**Note that income and standardized test scores are also certainly correlated*

3. What is/are the instrument(s) used by the authors in this study, and what are the authors instrumenting for? (2 points)

The instrument used by the authors in this study is the quarter of birth. They are instrumenting for educational attainment.

4. Generally, what conditions must an instrument satisfy to be considered valid?
    a. Explain these conditions in broad terms and in the specific context of the instrument(s) used in the paper. (3 points)

        1. ***Relevance****: The instrument must be correlated with the endogenous variable (the variable of interest) in the equation. In the context of the paper, the quarter of birth is relevant because it is correlated with educational attainment.*

Children born in different quarters start school at different ages, leading to differences in educational attainment.

2. ***Exclusion Restriction****: The instrument should only affect the outcome variable (in this case, earnings) through its effect on the endogenous variable (educational attainment), and not through any other pathways. In the paper, the exclusion restriction is satisfied because the quarter of birth affects earnings only through its impact on educational attainment, not directly.*
3. ***Independence****: The instrument should be as good as randomly assigned, meaning that it should be unrelated to any unobserved factors that affect the outcome variable. In the paper, the quarter of birth is assumed to be independent of other factors that affect earnings (eg., from before, parental income and standardized test scores).*

b. Explain these characteristics using random variables and potential outcomes. (2 points)

1. ***Relevance:*** *Relevance, in terms of potential outcomes, means that there is a difference in potential outcomes between treated and untreated units, given the instrument Zi. In the context of the paper, relevance implies that children born in different quarters (Zi) have different potential outcomes for educational attainment (Xi). For example, children born later in the year (Q4) might have a higher potential outcome for educational attainment compared to those born earlier in the year (Q1) if they are treated differently in terms of school entry age.*
2. ***Exclusion Restriction:*** *The exclusion restriction, when expressed using potential outcomes, asserts that the instrument affects the outcome variable (Yi) solely through its impact on the treatment variable (Xi). In this case, it means that the instrument Zi affects earnings (Yi) only through its influence on educational attainment (Xi), represented as Zi→Xi→Yi. For example, the quarter of birth affects earnings only through its impact on educational attainment, not directly.*
3. ***Independence:*** *Independence, in terms of potential outcomes, means that the instrument Zi is independent of the potential outcomes under different treatment conditions, Yi(0) and Yi(1). This implies that the instrument Zi is unrelated to any unobserved factors that influence the potential outcomes. In the paper, this means that the quarter of birth (Zi) is assumed to be unrelated to unobserved factors that affect earnings under different levels of educational attainment.*

5. Do you believe that the instrument(s) in the paper is/are truly exogenous? Why or why not? If so, provide a brief argument for this assumption. If not, provide an alternate mechanism through which the instrument(s) might affect the outcome variable, which suggests the exogeneity assumption may be violated. (2 points)

*I believe the instrument is exogenous – if the quarter of birth was **not** exogenous, that would mean it would have an effect on earnings through some other mechanism other than educational attainment. While birthdays are not uniformly distributed, it is hard to argue that when you are born might be correlated with other factors that affect future earnings. For example, I do not believe wealthier parents have children at certain points of the year relative to lower income parents.*

6. To assess whether the instrument is relevant, we can examine whether the instrument (quarter of birth) predicts the instrumentalized variable (compulsory schooling).
   a. Explain how Table I is constructed, and give some intuition for the authors' choices. (2 points)

Table I is constructed by estimating the effect of each quarter of birth (relative to Q4) on various educational outcomes, after "detrending" years of education across cohorts by subtracting off a moving average of the surrounding birth cohort's average education. The authors do this because there exists cohort-level trends in years of education that may bias the coefficients.

   b. Interpret the coefficient of the first quarter for the outcome variables "Total years of education" and "High school graduate" for the 1930-1939 cohort. (2 points)

Total years of education: Male students from the 1930-1939 cohort born in the first quarter of the year, on average, attain 0.124 fewer years of education relative to students born in the fourth quarter, statistically significant at the 5% level.

High school graduate: Male students from the 1930-1939 cohort born the first quarter of the year are, on average, 1.9 percentage points less likely to graduate from high school than men born in the last quarter of the year.

   c. Why do the authors estimate the coefficients displayed in the bottom part of Table 1 ("College graduate", "Completed master's degree", "Completed doctoral

degree")? How do these results support the validity of their instrument? Which assumption of the IV model are they addressing here? (3 points)

The authors estimate coefficients for post-secondary educational outcomes such as "College graduate", "Completed master's degree", and "Completed doctoral degree" to investigate whether the observed seasonal pattern in education persists beyond compulsory schooling (which is non-binding after high school). This analysis aims to provide further support for the validity of their instrument, the quarter of birth, by examining whether birth quarter continues to influence educational attainment even among individuals who are not constrained by compulsory schooling laws.

By assessing the effect of birth quarter on post-secondary educational outcomes, the authors address the ***exclusion*** *restriction assumption in the IV model, determining whether birth quarter affects earnings solely through its impact on educational attainment during the compulsory schooling period. If birth quarter remains a significant predictor of post-secondary educational outcomes, it suggests a potential violation of the exclusion restriction assumption. Conversely, if birth quarter has little to no effect on post-secondary educational outcomes, it strengthens the validity of the instrument and supports the assumption that the instrument affects earnings solely through its influence on educational attainment during compulsory schooling.*

The results support this claim; first quarter births are just slightly less likely to graduate college, and there is no discernible pattern for the proportion of men with master's or doctoral degrees by quarter of birth. Since the quarter a person is born is linked to when they start school, the fact that there's no clear trend in post-secondary education by birth season implies that starting school earlier or later doesn't affect educational attainment. So, without mandatory schooling, we wouldn't anticipate seeing any differences in total (or high-school) education either based on when someone is born.

7. Consider Table III and Table IV. Provide a general formula and a basic intuition for the Wald estimator. How does it compare to the OLS estimate? What is the advantage of using TSLS, instead of the Wald estimator? (4 points)

The Wald estimator computes the return to education as the ratio of difference in earnings by quarter of birth to difference in years of education by quarter of birth: $\frac{\Delta Earnings}{\Delta Education}$*. Intuitively, the Wald estimator captures how changes in education (measured by quarter of birth) impact changes in earnings, holding other factors constant.*

The results are very similar to the OLS estimates, and the Wald estimator presents consistent estimates since unobserved earnings determinants like innate ability or family background are likely to be uniformly distributed across people regardless of which day of the year they were born.

Instead of the Wald estimator, the TSLS estimation allows for incorporation of additional covariates – particularly, age-related trends in earnings. By using instrumental variables to address endogeneity, TSLS provides more consistent estimates of the return to education compared to Wald, allowing us to identify effects of education across birth quarters within each birth year. In other words, TSLS removes variation of years of education that is related to the error term.

8. How would you construct a reduced form table? Why might you want to report reduced form estimates? What figure in the paper fulfills this purpose? (3 points)

You would estimate the effect of the instrument (quarter of birth) on earnings. Each row in the table would represent a different specification of the regression model, possibly including different control variables. The coefficients and standard errors for the instrument variable (quarter of birth) would be reported, along with any other relevant statistics, such as R-squared or F-statistics.

You would want to report this to demonstrate whether the instrument has an effect on the variable of interest. Table III column III fulfills this purpose because it demonstrates the difference in log weekly wages for people born in the first quarter versus second, third, or fourth quarters. Taken together, this reduced form estimate, and the first stage estimate (which measures relevance), can be used to back out the IV estimator – which is good for fact checking your estimates!

9. Subsequent papers have found that the instrument (quarter-of-birth) is weak for some specifications in the paper.
    a. What is the intuition for why weak instruments are problematic? (1 point)

Weak instruments mean the relevance assumption is loosely, or not, met. Thus, in the equation for our IV estimator, the first stage estimator (denominator) would approach 0, thus creating a division problem, biasing TSLS towards OLS.

    b. Read the following, explain the intuition, and explain the implications for weak instruments. (2 points)

Suppose we have an IV model with T_i, Z_i scalar

$$Y_i = T_i\beta + X_i\eta + \varepsilon_i$$

$$T_i = Z_i\pi + X_i'\delta + V_i$$

and re-write as reduced-form, first stage

$$Y_i = Z_i\pi\beta + X_i'\tau + U_i$$

$$T_i = Z_i\pi + X_i'\delta + V_i$$

Usual IV estimate (e.g. 2SLS) equal to ratio of OLS coefficients on Z_i in reduced-form and first stage,

$$\hat{\beta} = \frac{\pi\beta}{\hat{\pi}}$$

Under mild conditions, the reduced-form and first stage are jointly normal:

$$\sqrt{n}\begin{pmatrix} \hat{\pi}\beta - \pi(\theta)\beta(\theta) \\ \hat{\pi} - \pi(\theta) \end{pmatrix} \rightarrow_d N(0, \Sigma)$$

Provided $\pi(\theta) \neq 0$, for $\hat{\Sigma}_\beta$ asymptotic variance estimate for $\hat{\beta}$

$$\sqrt{n}\hat{\Sigma}_\beta^{-\frac{1}{2}}(\hat{\beta} - \beta(\theta)) \rightarrow_d N(0, 1)$$

However, for $\pi(\theta) = 0$, the continuous mapping theorem implies:

$$\hat{\beta} \rightarrow_d \frac{\xi_1}{\xi_2}, \xi \sim N(0, \Sigma)$$

Can show $\sqrt{n}\hat{\Sigma}_\beta^{-\frac{1}{2}}(\hat{\beta} - \beta(\theta))$ has non-standard distribution

c. Optional: if you know what the bootstrap does, why does bootstrapping *not* solve the weak identification issue? (1 extra point)

10. If you were to write the paper today, how would you detect weak instruments, and what statistic would you use for inference?
    Hint: you may want to refer to Andrews, Stock and Sun (2019).
    a. How is the effective F-statistic constructed? (1 point)
    b. How are Anderson-Rubin confidence sets constructed? (1 point)

The Local Average Treatment Effect

11. Explain the monotonicity assumption in the context of this study. What is required regarding the relationship between variables for monotonicity to be met, and is it reasonable to assume that defiers do not exist? In your explanation, be sure to touch on what it means to be a defier in this study. (3 points)

The monotonicity assumption in this study requires that the effect of the instrumental variable (QOB) on the treatment variable (education) is consistently in one direction. This means that individuals always receive more education when exposed to a higher quarter of birth (born later in the year). In other words, QOB only affects education as expected.

It is reasonable to assume that defiers do not exist because defiers are individuals who would receive less education when required to get more (due to compulsory schooling laws – this seems fairly impossible.

12. Interpret the IV estimates in Table IV with appropriate units in the context of the study's research question, treating them as a local average treatment effect. In your interpretation, clarify the population for which this local average treatment effect is identified (i.e., who are the compliers?). (2 points)

The LATE is the effect for the compliers, who in this study are the people who get more school because they are required to. The LATE suggests compliers receive an average of 0.06-0.09 more years of schooling due to being required to (quarter of birth).

13. In 3-5 sentences, discuss how these results might inform policy outside of this setting. In your discussion, be sure to comment on the challenges of generalizing instrumental variable findings. (2 points)

The results underscore the potential impact of compulsory schooling laws on educational outcomes, suggesting that individuals compelled to stay in school longer due to their quarter of birth tend to achieve higher levels of education. This insight could guide policymakers in refining compulsory schooling policies to enhance educational attainment across different birth quarters – for example, requiring X years of education, rather than a drop out age.

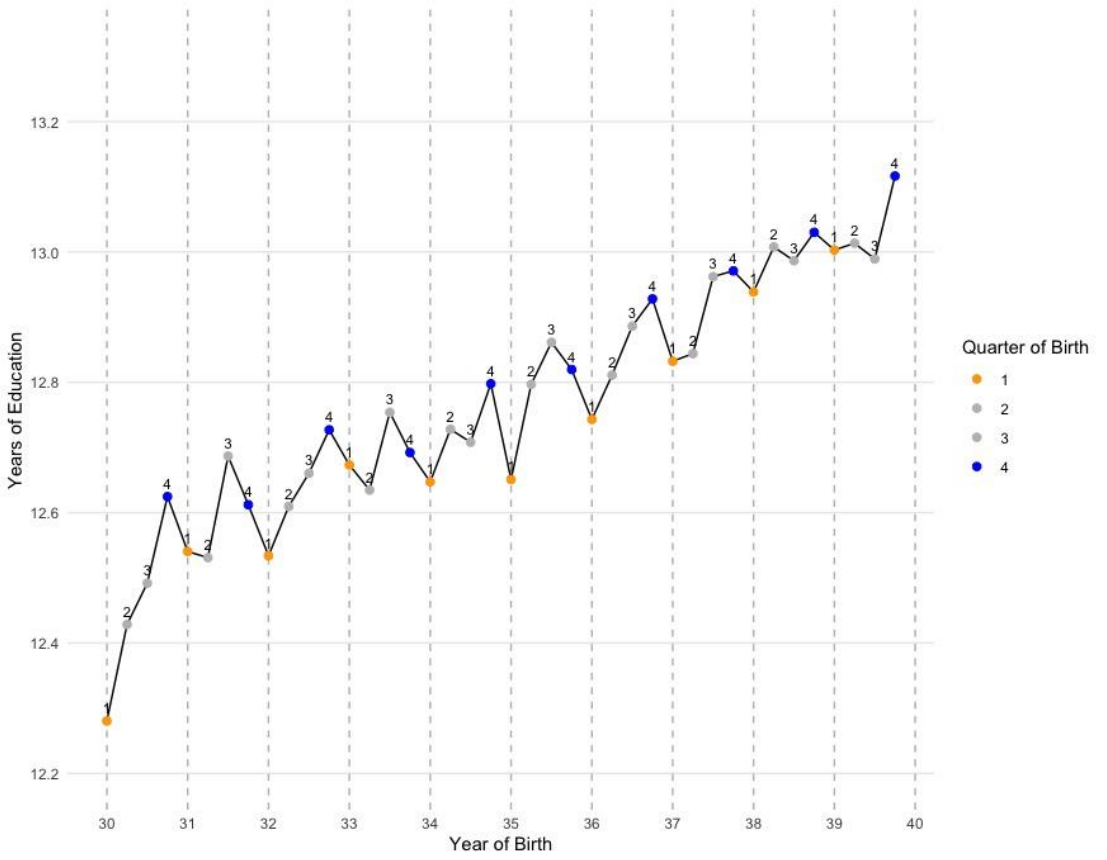
However, generalizing these findings to broader policy contexts requires caution. Specifically, these findings relate to those who received more schooling become they were compelled to. This doesn't evaluate the effectiveness of the policy generally speaking – for example, perhaps ***everyone*** *could receive more schooling under a different policy.*

Data Analysis Questions (20 points)

The enclosed is a subsample from Angrist and Krueger's dataset. Specifically, for men born between 1930 and 1939, it includes the following information from the 1980 Census:

- LWKLYWGE: log of weekly earnings
- EDUC: years of completed education
- YOB: year of birth
- QOB: quarter of birth
- Age, marriage status (1=married), race (1=black), urban dummy (SMSA, 1= center city)
- 8 region of residence dummies (NEWENG, MIDATL, ENOCENT, WNOCENT, SOATL, ESOCENT, WSOCENT, MT)

14. Figure I can be thought of as a "graphical first-stage", as it shows the mean of completed years of education by quarter-of-birth for each year of birth between 1930 and 1939. Replicate Figure I, and highlight those born in the first quarter (for each year between 1930 and 1939) in your figure. (2 points)
    Hint: you may want to create year-of-birth and quarter-of-birth dummies. They will also be useful for the following questions.



15. Table I shows the relationship between quarter-of-birth and educational outcomes. Replicate the first row of Table I, i.e., find the coefficients of the first, second, and third quarter-of-birth dummies on total years of education. (2 points)

term	estimate
(Intercept)	0.057
quarter_1	-0.124
quarter_2	-0.086
quarter_3	-0.015

My standard errors were extremely low and we couldn't figure out why after troubleshooting in office hours.

16. Create a reduced form table that illustrates the relationship between quarter-of-birth and weekly earnings. In other words, regress log weekly earnings on the quarter-of-birth dummies (our instruments). Include year fixed effects. (2 points)

```
Call:
   felm(formula = LWKLYWGE ~ quarter_2 + quarter_3 + quarter_4 |      YOB, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2536 -0.2613  0.0635  0.3612  4.6424

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
quarter_2 0.004875   0.003190   1.528    0.126
quarter_3 0.015463   0.003126   4.946 7.58e-07 ***
quarter_4 0.014366   0.003184   4.512 6.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6775 on 367713 degrees of freedom
Multiple R-squared(full model): 0.0001811   Adjusted R-squared: 0.0001457
Multiple R-squared(proj model): 9.198e-05   Adjusted R-squared: 5.663e-05
F-statistic(full model):5.123 on 13 and 367713 DF, p-value: 3.383e-09
F-statistic(proj model): 11.28 on 3 and 367713 DF, p-value: 2.159e-07
```

17. Table III reports OLS and Wald estimates of returns of education. Replicate both estimates (in the last two rows) for men born between 1930-1939 (Panel B). *Hint: See footnote 13 in Angrist and Krueger (1991) for details on how they calculate the Wald estimate. Note that if you want to use the function felm, since there are no covariates, you will need to include 1 as a covariate (i.e., y ~ 1 | 0 | x ~ z).* (2 points)

```
Call:
lm(formula = LWKLYWGE ~ EDUC, data = cohort_one)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7540 -0.2367  0.0726  0.3318  4.6357

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.9951823  0.0044644  1118.9   <2e-16 ***
EDUC        0.0708510  0.0003386   209.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6378 on 329507 degrees of freedom
Multiple R-squared:  0.1173,    Adjusted R-squared:  0.1173
F-statistic: 4.378e+04 on 1 and 329507 DF,  p-value: < 2.2e-16
```

```
Call:
   felm(formula = LWKLYWGE ~ 1 | 0 | (EDUC ~ quarter_1), data = cohort_one)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9792 -0.2558  0.0692  0.3542  4.8154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.59748    0.30583  15.033  < 2e-16 ***
`EDUC(fit)`  0.10200    0.02395   4.259 2.06e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6459 on 329507 degrees of freedom
Multiple R-squared(full model): 0.09463   Adjusted R-squared: 0.09462
Multiple R-squared(proj model): 0.09463   Adjusted R-squared: 0.09462
F-statistic(full model):18.14 on 1 and 329507 DF, p-value: 2.055e-05
F-statistic(proj model): 18.14 on 1 and 329507 DF, p-value: 2.055e-05
F-statistic(endog. vars):18.14 on 1 and 329507 DF, p-value: 2.055e-05
```

18. Table V reports different specifications of the TSLS for men born between 1930-1939. Run TSLS regressions replicating Column 2 and Column 6. For Column 2, instrument for education with a full set of quarter-of-birth times year-of-birth dummies, and include year fixed effects. For Column 6, instrument for education with the same set of quarter-of-birth times year-of-birth dummies, and include regional fixed effects, year fixed effects, race, urban, and married status dummies. (4 points)

Column 2

```
Call:
   felm(formula = LWKLYWGE ~ 1 | YOB | (EDUC ~ year_dummies * quarter_dummies),      data = cohort_one)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8781 -0.2404  0.0703  0.3417  4.7491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
`EDUC(fit)`  0.08912    0.01611   5.532 3.17e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6404 on 329498 degrees of freedom
Multiple R-squared(full model): 0.1102   Adjusted R-squared: 0.1101
Multiple R-squared(proj model): 0.1101   Adjusted R-squared: 0.1101
F-statistic(full model):4.167 on 10 and 329498 DF, p-value: 8.581e-06
F-statistic(proj model):  30.6 on 1 and 329498 DF, p-value: 3.175e-08
F-statistic(endog. vars): 30.6 on 1 and 329498 DF, p-value: 3.175e-08
```

Column 6

term	estimate	std.error
RACE	-0.2302	0.0261
SMSA	-0.1581	0.0174
MARRIED	0.2440	0.0049
`EDUC(fit)`	0.0806	0.0164

```
Call:
   felm(formula = LWKLYWGE ~ RACE + SMSA + MARRIED | YOB + NEWENG +      MIDATL + ENOCENT + WNOCENT + SOATL + ESOCENT + WSOCENT +
MT | (EDUC ~ year_dummies * quarter_dummies), data = cohort_one)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9749 -0.2305  0.0574  0.3247  4.6815

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
RACE        -0.230218   0.026126  -8.812  < 2e-16 ***
SMSA        -0.158147   0.017423  -9.077  < 2e-16 ***
MARRIED      0.243969   0.004871  50.088  < 2e-16 ***
`EDUC(fit)`  0.080552   0.016385   4.916 8.83e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6228 on 329487 degrees of freedom
Multiple R-squared(full model): 0.1584   Adjusted R-squared: 0.1584
Multiple R-squared(proj model): 0.1419   Adjusted R-squared: 0.1419
F-statistic(full model): 1436 on 21 and 329487 DF, p-value: < 2.2e-16
F-statistic(proj model):  5654 on 4 and 329487 DF, p-value: < 2.2e-16
F-statistic(endog. vars):24.17 on 1 and 329487 DF, p-value: 8.831e-07
*** Standard errors may be too high due to more than 2 groups and exactDOF=FALSE
```

19. Now repeat the first TSLS regression in Question 18 (Column 2, without additional controls and only year fixed effects), but instead of using a built-in IV function, regress education directly on the instruments and then use predicted education to estimate the wage return of education. (Use `lm` if you choose to use R, since `felm` does not support `predict`). Do your results match your results from the previous question? (3 points)

```
Call:
lm(formula = LWKLYWGE ~ predicted_educ + factor(YOB), data = cohort_one)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2496 -0.2612  0.0610  0.3609  4.6425

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.792727   0.212733  22.529  < 2e-16 ***
predicted_educ 0.089115   0.017077   5.218 1.81e-07 ***
```

They are very, very close.

20. For this question, define treatment $\mathbf{X_i}$ as completing high school (**12 or more years of education**), and set the instrument as binary, with Z_i equal to 1 if **born in the fourth quarter**, and 0 otherwise. The sample includes men born between 1930-1939.
    a. What is the share of the complier population? (1 point)

A complier is an individual who would attend high school (receive treatment) if and only if they were assigned to the treatment group based on the instrument. In other words, a complier is someone who completes high school if they were born in the fourth quarter, and would not complete high school if they were born in any other quarter.

Complier share can be calculated by taking the first stage coefficient.

```
Call:
lm(formula = treatment ~ instrument, data = cohort_one)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7819  0.2181  0.2324  0.2324  0.2324

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7676392  0.0008424 911.290   <2e-16 ***
instrument  0.0142618  0.0017006   8.386   <2e-16 ***
```

1.4%

    b. What is the average untreated outcome (log of weekly earnings) for never-takers? (1 point)

```
> mean(cohort_one$LWKLYWGE[cohort_one$treatment == FALSE & cohort_one$instrument == TRUE])
[1] 5.597343
```

Assume no defiers given monotonicity.

    c. What is the average treated outcome (log of weekly earnings) for always-takers? (1 point)

because treatment true and instrument true includes the people who switched (compliers), but treatment true and instrument false means no one could've switched because of the instrument

```
> mean(cohort_one$LWKLYWGE[cohort_one$treatment == TRUE & cohort_one$instrument == FALSE])
[1] 5.990221
```

d. Is there selection into treatment? State the assumptions necessary for your conclusion. (2 points)

There appears to be some selection into the treatment, as treatment (12 or more years of schooling) are not balanced across race and urban status.

```
> t.test(RACE ~ treatment, data = cohort_one)

        Welch Two Sample t-test

data:  RACE by treatment
t = 66.053, df = 95926, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 0.08924606 0.09470445
sample estimates:
mean in group FALSE  mean in group TRUE
         0.15260171          0.06062645

> t.test(SMSA ~ treatment, data = cohort_one)

        Welch Two Sample t-test

data:  SMSA by treatment
t = 49.602, df = 110289, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 0.08324891 0.09009866
sample estimates:
mean in group FALSE  mean in group TRUE
          0.2531693           0.1664955
```

Without selection into the treatment, we would expect characteristics of those receiving the treatment and not receiving the treatment to be equivalent. While these characteristics are equivalent among those that receive or do not receive the instrument (born in Q4)

IVs in Your Own Work (8 points)

21. Think about a social relationship that would be best studied using an IV approach. Briefly state the research question and the main variables of interest in non-technical terms. (4 points)

Research Question: Does the deployment of body-worn cameras (BWCs) by police departments reduce incidents of police use of force?

Variables of Interest

Dependent Variable: Incidents of police use of force

Independent Variable: Deployment of body-worn cameras

Instrumental Variable: Funding for body-worn cameras

We use an IV approach because there is likely correlation between deployment of BWCs and incidents of use of force that would bias a simple OLS regression analysis. Funding for BWCs is unlikely affect police use of force except through actual BWC deployment.

22. Write out the empirical specification you would use and explain the equation. (2 points)

***First Stage**:*

Equation: $BWC_{it} = \alpha 0 + \alpha 1 Funding_{it} + \gamma_i + \lambda_t + \eta_{it}$

- *BWC_{it} represents the deployment of body-worn cameras by police department i at time t.*
- *$Funding_{it}$ is the funding allocated for body-worn cameras by police department i at time t.*
- *γ_i and λ_t represent police department and time fixed effects, respectively, capturing unobserved time-invariant characteristics of police departments and common shocks affecting all departments over time.*
- *η_{it} is the error term capturing unobserved factors affecting funding for body-worn cameras.*

***Second Stage**:*

Equation: Use of Force_it = β0 + β1BWC^_it + γ_i + λ_t + ϵ_it

- *Use of Force_it represents the incidence of police use of force at time t in police department i.*
- *BWC^_it is the predicted value of BWC deployment obtained from the first-stage regression.*
- *γ_i and λ_t are the same police department and time fixed effects included in the first stage.*
- *ϵ_it is the error term capturing unobserved factors affecting police use of force incidents.*

23. If you clustered your standard errors or included fixed effects, explain why these methods reduced the likelihood of bias in your results (and if applicable, in which direction). If you did not, explain why these methods were not appropriate in your setting. (2 points)

Since outcomes (police use of force) may be correlated within the same police department over time, clustering standard errors by police department would be appropriate. This approach considers the possibility of uneven variability and patterns of correlation in use-of-force incidents within police departments. By doing so, it helps prevent the risk of underestimating the precision of our results and drawing inaccurate conclusions about the relationships between variables.

Additionally, including fixed effects for police departments would control for time-invariant characteristics of each department that may affect both the deployment of body-worn cameras and the incidence of police use of force. By controlling for these fixed effects, we mitigate the risk of omitted variable bias and obtain more reliable estimates of the causal effect of BWC deployment on police use of force.