PROBLEM SET 2

Jacob Jameson

Due on Tuesday March 7, 2023

IDENTIFICATION

(1) Your information

Jacob Jameson

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

Bohan Li, Jenna Rogers

(3) Compliance with Harvard Kennedy School Academic Code

I certify that my work in this problem set complies with the Harvard Kennedy School Academic Code

Conceptual Questions

1. Read the paper. In a couple of sentences, explain the primary research question that the authors are trying to answer. What is the key logic behind the authors' answer? In a separate sentence, explain why policymakers should care about that question. Use non-technical jargon so that someone without statistical training could understand.

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

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

4. Generally, what characteristics must an instrument have for it to be considered valid?

   a. Explain these characteristics in words, and specific terms of the instrument(s) used in the paper.

   b. Explain these characteristics using random variables and potential outcomes.

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 for how the exogeneity assumption may be violated.

6. To assess whether the instrument is relevant, we can examine whether the instrument (quarter of birth) predicts the instrumentalized variable (compulsory schooling). In the following parts a - c, provide interpretations with concrete units, in a way a policymaker can understand.

   a. Explain how Table 1 is constructed, and give some intuition for the authors' choices.

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

   c. Why do the authors estimate coefficients for the bottom part of Table 1 ("College graduate", "Completed master's degree", "Completed doctoral degree")? How do the results support the validity of their instrument? What assumption of the IV model are they addressing here?

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?

8. How would you construct a reduced form table? What would be the purpose thereof? What figure in the paper fulfills this purpose?

9. Subsequent papers have found the instrument to be weak for some specifications of the paper.

   a. Provide an intuition for the reason why weak instruments are problematic.

   b. Read the following, explain the intuition, and explain the implications for weak instruments.

   c. Optional: if you know what the bootstrap does, why does bootstrapping not solve the weak identification issue?

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?

   b. How are Anderson-Rubin confidence sets constructed?

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 does it means to be a defier in this study.

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?).

13. With 3-5 sentences, discuss the how you believe these results inform policy outside of the specified context. In your discussion, be sure to touch on the problems specific to generalizing from instrumental variable findings and the particular scope conditions of this paper's findings.

Data Analysis Questions (22 points + 2 extra points)

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 1940. Replicate Figure I. You can highlight the mean of years of education of those born in the first quarter (for each year between 1930 and 1940).

Hint: you may want to create year of birth and quarter of birth dummies. They will also be useful for the following questions.

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

```
data <- read_dta("data/AK91_1930_39.dta") %>%
  filter(YOB <= 39)

data$QOB <- as.factor(data$QOB)

data %>%
  group_by(YOB, QOB) %>%
  summarise(mean.EDUC = mean(EDUC),
            quarter.year = YOB + (as.numeric(QOB)/4) - 0.25) %>%
  ungroup() %>%
  unique() %>%
ggplot(aes(x = quarter.year, y = mean.EDUC)) +
  geom_path() +
  geom_point(aes(color=QOB), size = 3) +
  geom_text(aes(label = paste(QOB)),
          vjust = 2.5, size = 3) +
  scale_x_continuous(breaks = seq(30, 40, by = 2)) +
  scale_y_continuous(breaks = seq(12.2, 13.2, by = .2)) +
  scale_color_manual(values = c("red", "black", "black", "black")) +
  labs(x = "Year of Birth", y = "Years of Completed Education",
       title = "Relationship between birth year and years of education",
       caption = paste("Figure I\n",
       "Years of Education and Season of Birth 1980 Census \n",
       "Note: Quarter of birth is listed below each observation")) +
  theme_classic() +
  theme(plot.caption = element_text(
    color = "black",
    size = 8,
    lineheight = 1.2,
    hjust = 0.5,
    margin = margin(t = 20)),
    legend.position = "none")
```

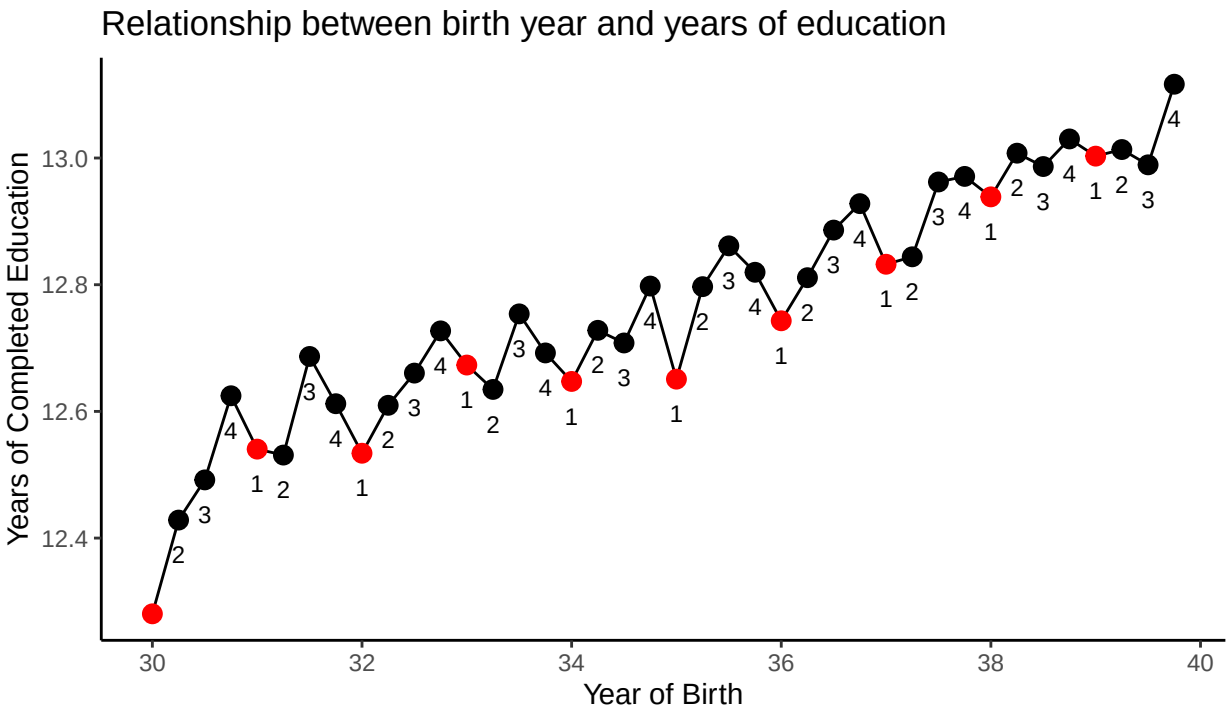

Figure I
Years of Education and Season of Birth 1980 Census
Note: Quarter of birth is listed below each observation

Table 1: Quarter-of-birth effect

Outcome Variable	Birth Cohort	Mean	I	II	III	F-test
Total years of Education	1930-1939	12.77	-0.151 (0.016)	-0.095 (0.016)	-0.034 (0.016)	34.00945

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. Your estimates do not need to be exactly the same as in the paper, but roughly of the same magnitude.

```
data <- within(data, QOB <- relevel(QOB, ref = 4))

ols <- tidy(lm(formula = paste0('EDUC', " ~ QOB"), data = data))

table <- data.frame(
    `Outcome variable` = character(),
    `Birth cohort` = character(),
    Mean = double(),
    I = double(),
    II = double(),
    III = double(),
    `F-test` = character())

table <- rbind(table, data.frame(
    `Outcome variable` = 'Total years of Education',
    `Birth cohort` = '1930-1939',
    Mean = round(mean(data$EDUC, na.rm = TRUE), 3),
    I = paste(round(ols$estimate[2], 3),
              paste0('(',round(ols$std.error[2], 3)),')'),
    II = paste(round(ols$estimate[3], 3),
               paste0('(',round(ols$std.error[2], 3)),')'),
    III = paste(round(ols$estimate[4], 3),
                paste0('(',round(ols$std.error[2], 3)), ')'),
    `F-test` = summary(lm(formula = paste0('EDUC', " ~ QOB"),
                          data = data))$fstat[[1]]))

kable(
    table,
    caption = "Quarter-of-birth effect",
    align = "c",
    col.names = c("Outcome Variable", "Birth Cohort",
                  "Mean", "I", "II", "III", "F-test")) %>%
  kable_styling(full_width = FALSE)
```

16. Create a reduced form table. In other words, you want to regress the log of weekly earnings on the quarter of birth dummies (our instrument). You may want to include year fixed effect.

```
reg.reduced.form <- felm(QOB ~ LWKLYWGE|0|0|0,data)

reg.reduced.form.fe <- felm(QOB ~ LWKLYWGE|YOB|0|0,data)
```

17. Table III reports OLS and Wald estimates of returns of education. Replicate both estimates for men born 1930-1939 (Panel B).

```
data$QOB_dummy <- ifelse(data$QOB == 1, 1, 0)
wage_1st <- data %>% filter(QOB == 1) %>% summarise(mean(LWKLYWGE))
wage_other <- data %>% filter(QOB != 1) %>% summarise(mean(LWKLYWGE))
wage_diff <- wage_1st - wage_other

wage_err <- sqrt(
sum(data %>% filter(QOB == 1) %>% summarise(sd(LWKLYWGE)^2),
    data %>% filter(QOB != 1) %>% summarise(sd(LWKLYWGE)^2))
)

educ_1st <- data %>% filter(QOB == 1) %>%
  summarise(mean(EDUC))

educ_other <- data %>% filter(QOB != 1) %>%
  summarise(mean(EDUC))

educ_diff <- educ_1st - educ_other

educ_err <- sqrt(
sum(data %>% filter(QOB == 1) %>% summarise(sd(EDUC)^2),
    data %>% filter(QOB != 1) %>% summarise(sd(EDUC)^2))
)

# wald return to education
data$EDUC_pred <- lm(EDUC ~ QOB_dummy, data = data) %>% predict()
wald_rslt <- lm(LWKLYWGE ~ EDUC_pred, data = data)

# ols return to education
ols_rslt <- lm(LWKLYWGE ~ EDUC, data = data)

wald_est = coef(wald_rslt)["EDUC_pred"]
wald_err = sqrt(diag(vcovHC(wald_rslt, type = "HC1"))["EDUC_pred"])
ols_est = coef(ols_rslt)["EDUC"]
ols_err = sqrt(diag(vcovHC(ols_rslt, type = "HC1"))["EDUC"])

# create table
table <- data.frame(
    Variable = c("ln (wkly. wage)", "",
                 "Education", "",
                 "Wald est. of return to education", "",
                 "OLS return to education", ""),
    `(1) Born in 1st quarter of year` =
      c(round(wage_1st$`mean(LWKLYWGE)`,5), "",
        round(educ_1st$`mean(EDUC)`,5), "",
        "", "",
        "", ""),
    `(2) Born in 2nd, 3rd, or 4th quarter of year` =
      c(round(wage_other$`mean(LWKLYWGE)`,5), "",
        round(educ_other$`mean(EDUC)`,5), "",
        "", "",
        "", ""),
```

Table 2: Panel B: Wald Estimates for 1980 Census - Men Born 1930-1939

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Diff
ln (wkly. wage)	5.8916	5.90269	
Education	12.68807	12.79688	
Wald est. of return to education			
OLS return to education			

```
    `(3) Difference (std.error) (1) - (2)` =
      c(paste(round(wage_diff$`mean(LWKLYWGE)`,5),
        paste0("(", round(wage_err ,5), ")")), "",
        paste(round(educ_diff$`mean(EDUC)`,5),
        paste0("(", round(educ_err, 5), ")")), "",
        paste(round(wald_est[[1]], 5),
        paste0("(", round(wald_err[[1]], 5), ")")), "",
        paste(round(ols_est[[1]], 5),
        paste0("(", round(ols_err[[1]], 5), ")")), "")
  )

kable(
    table,
    caption =  "Panel B: Wald Estimates for 1980 Census - Men Born 1930-1939",
    align = "c",
    col.names = c("", "(1) Born in 1st quarter of year",
                  "(2) Born in 2nd, 3rd, or 4th quarter of year",
                  "(3) Difference (std.error) (1) - (2)")) %>%
  kable_styling(full_width = FALSE)
```

18. Table V reports different specifications of the TSLS for men born 1930-1939. Run TSLS similar as in Column 2 and Column 6. First, instrument education with a set of quarter-of-birth x year-of-birth dummies, and add year fix effects as control. Why do we want to include year fix effects?

Second, similarly to Column 6, instrument education with the same interaction dummies, and add regional fix effects, year fix effects, race, married status, and an urban dummy.

```
data$YOB <- as.factor(data$YOB)

ivreg.2 <- ivreg(LWKLYWGE ~ EDUC + YOB | QOB*YOB, data = data)

ivreg.6 <- ivreg(LWKLYWGE ~ EDUC + YOB + NEWENG +
                   MIDATL + ENOCENT + WNOCENT +
                   SOATL + ESOCENT + WSOCENT + MT +
                   SMSA + MARRIED + RACE | QOB*YOB, data = data)


varnames <- c("EDUC", "RACE", "SMSA", "MARRIED")
labels <- c("Years of Education", "Race (1 == black)",
```

```
            "SMSA (1 == city center)", "Married (1 == married)")

birth_dummy <- c("Yes", "Yes")
region_dummy <- c("No", "Yes")

stargazer(ivreg.2, ivreg.6, type = "latex",
          coef.names = labels,
          covariate.labels = labels,
          covariate.include = varnames,
          column.labels = c("Model 2 (TSLS)", "Model 6 (TSLS)"),
          add.lines = list(c("9 Year of Birth dummies", birth_dummy[1], birth_dummy[2]),
                           c("8 Region of Residence dummies",
                             region_dummy[1], region_dummy[2])),
                    omit = c("QOB", "YOB", "NEWENG", "MIDATL",
                             "ENOCENT", "WNOCENT", "SOATL", "ESOCENT",
                             "WSOCENT", "MT"))
```

Table 3:

	Dependent variable:	
	LWKLYWGE	
	Model 2 (TSLS)	Model 6 (TSLS)
	(1)	(2)
Years of Education	0.089***	0.068***
	(0.016)	(0.024)
Race (1 == black)		−0.783*
		(0.415)
SMSA (1 == city center)		0.486
		(0.522)
Married (1 == married)		−0.411
		(0.514)
Constant	4.793***	4.872***
	(0.201)	(0.821)
9 Year of Birth dummies	Yes	Yes
8 Region of Residence dummies	No	Yes
Observations	329,509	329,509
R^2	0.110	−0.082
Adjusted R^2	0.110	−0.082
Residual Std. Error	0.640 (df = 329498)	0.706 (df = 329487)
Note:		*p<0.1; **p<0.05; ***p<0.01

19. Now repeat the same exercise as the first 2LS2 in Question 18 (without additional controls, only year fix effects), but without using a built-in IV function – use only lm instead. Are your results the same as in the previous question?

20. For the purpose of this question, define treatment as completing high school (12 or more years of education), and the instrument as binary, with Zi=1 if one is born in the fourth quarter, and 0 otherwise.

    a. What is the share of the complier population?

    b. What is the average untreated outcome among the never-takers?

    c. What is the average treated outcome among the always-takers?

    d. Is there selection into treatment? State the assumptions necessary for your conclusion.