Assignment 4

Mark Deming

SPR 2025

Overview

You will replicate some parts of Angrist & Krueger's (1991) analysis of the effect of compulsory schooling laws on educational attainment and earnings. Specifically, you will replicate:

- Figure I (p. 983)
- Table V (p. 1000)

Both of these focus on males born during 1930-1939. The basic procedure applies to other cohorts too (e.g., 1940-1949). **The main aim of the assignment is to practice regression with instrumental variables, or two-stage least squares regression (TSLS).**

The data are contained in `angrist-krueger_1991.csv`, which is available at the course GitHub repository: https://github.com/jmdeming/independent_study/tree/main/datasets. There is alos a `README.txt` file that describes the variables in the dataset.

Get started

1. Load the `angrist-krueger_1991.csv` dataset.

2. You will also require the following packages. Load them:

- `here`
- `readr`
- `tidyverse`
- `ivreg`
- `modelsummary`

Figure I (p. 983)

Examine Figure I on page 983. To replicate the Figure, you must first summarize the data. Specifically, the Figure shows the average years of schooling by year and quarter of birth. The Figure is limited to individuals born in 1930 through 1939. This means that you must:

- Filter the data to include only individuals born in 1930 through 1939.
- Group the data by year of birth (`yob`) and quarter of birth (`qob`).
- Calculate the average years of schooling (`educ`) by year and quarter of birth.

3. Summarize the data, calculating the average years of schooling by year and quarter of birth for individuals born during 1930-1939.

```
# Summarize
summarized_dat <- ak %>%
  filter(yob %in% c(1930:1939)) %>%
  group_by(yob, qob) %>%
  summarize(mean_educ = mean(educ))
```

```
## `summarise()` has grouped output by 'yob'. You can override using the `.groups`
## argument.
```

The next step is to plot the summarized data above. The challenge is to get the data along the x-axis to show the year and quarter of birth. To do this, we will combine the year and quarter of birth into a single variable and then plot that variable. For example:

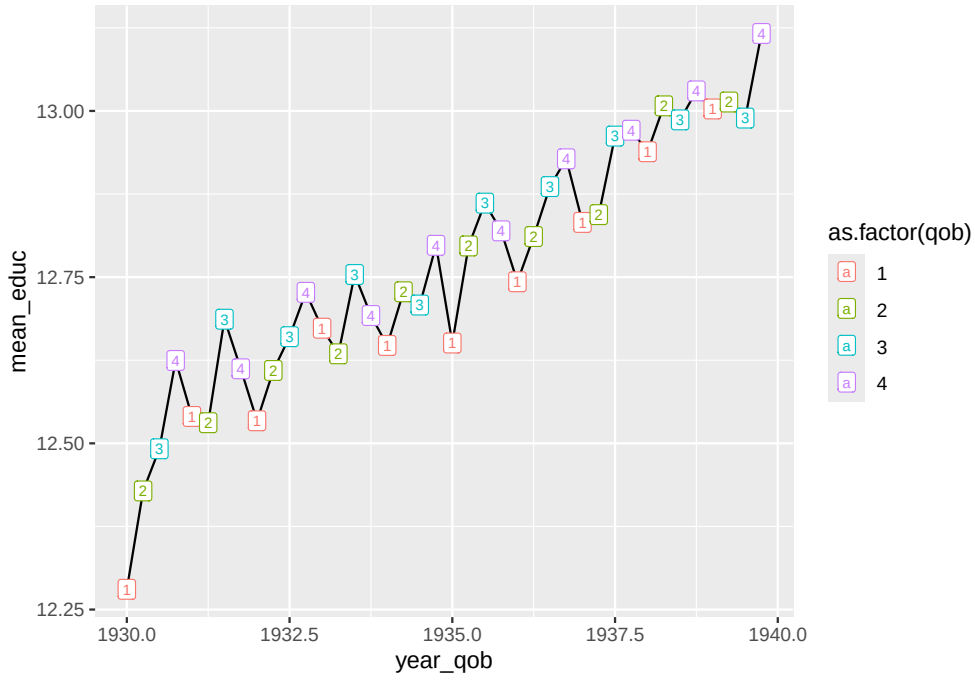
- yob = 1930 / qob = 1 becomes 1930.00
- yob = 1930 / qob = 2 becomes 1930.25
- yob = 1930 / qob = 3 becomes 1930.50
- yob = 1930 / qob = 4 becomes 1930.75
- ... and so forth

My code below does this for you. Examine my code. Be sure that you know what it is doing and why. Then run the code.

```
# Create a new variable that combines the year and quarter of birth
summarized_dat <- summarized_dat %>%
  mutate(year_qob = yob + (qob - 1) * .25)
```

4. Plot the summarized data. The x-axis should show the year and quarter of birth. The y-axis should show the average years of schooling. Label the points with the quarter of birth.

```
# Basic plot
 ggplot(summarized_dat,
         aes(x = year_qob,
             y = mean_educ,
             label = qob)) +
  geom_line() +
  geom_label(aes(color = as.factor(qob)),
             size = 2.5)
```




```
# Advanced plot
 ggplot(summarized_dat,
        aes(x = year_qob,
            y = mean_educ,
            label = qob)) +
  geom_line() +
  geom_label(aes(color = as.factor(qob)),
             size = 2.5) +
  scale_x_continuous(limits = c(1930, 1940),
                     breaks = seq(1930, 1940, 2),
                     labels = seq(1930, 1940, 2)) +
  scale_y_continuous(limits = c(12.2, 13.2),
                     breaks = seq(12.2, 13.2, .2),
                     labels = seq(12.2, 13.2, .2)) +
  scale_color_manual(values = c("steelblue", "steelblue", "steelblue", "firebrick")) +
  labs(title = "Figure I: Years of education and season of birth",
       subtitle = "1980 census\n",
       x = "\nYear of birth",
       y = "Years of completed education\n") +
  theme_bw() +
  theme(legend.position = "none",
        plot.title.position = "plot",
        axis.ticks = element_blank(),
        panel.grid = element_line(linetype = "dotted", color = "gray"))
```

Figure I: Years of education and season of birth

1980 census

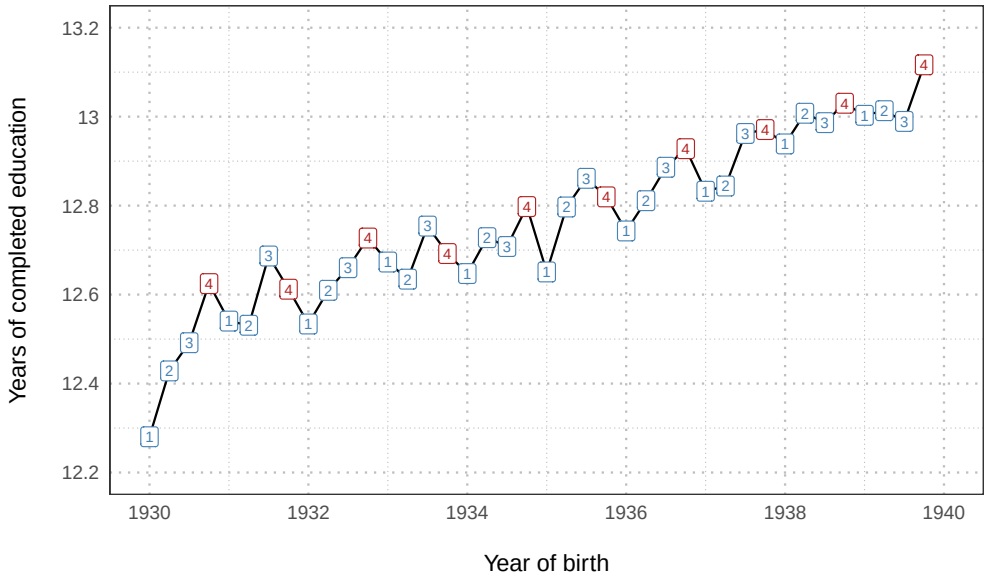


Table V (p. 1000)

You will now replicate Table V on page 1000 of the article. The Table presents results from 8 regression models. Models 1, 3, 5, and 7 are OLS models. Models 2, 4, 6, and 8 are TSLS (Two Stage Least Squares) models. The models are estimated using the `ivreg` function from the `ivreg` package.

Table V analyzes men born during 1930-1930. The first step, then, is to create a subset of the data that grabs on rows in which `yob` is within 1930:1939.

5. Subset the data to grab only rows in which `yob` is within `1930:1939`.

```
# Subset the data, grabbing men born 1930-1939
ak_sub <- ak %>%
  filter(yob %in% c(1930:1939) & !is.na(educ) & !is.na(lnw))
```

The next step is ensure that the `qob` and `yob` variables are treated as factors. This is important because the models will include dummies for these variables.

6. Convert `qob` and `yob` to factors.

```
# Convert qob and yob to factor variables
ak_sub <- ak_sub %>%
  mutate(qob_f = factor(qob),
         yob_f = factor(yob))
```

The next step is to create the instrument. The authors use quarter of birth (`qob`) as an instrument for education. But, they do so *within each year of birth (`yob`)*. This is because compulsory schooling laws change over time. So, instead of using just `qob`, we interact `qob` with `yob`. This gives us 30 instruments: `Q1_1930`,`Q2_1930`, ..., `Q4_1939`. These allow the effect of quarter of birth to vary by cohort — a key assumption in the authors' identification strategy.

The relevant R function for creating interactions with factor variables is `interaction()`.

7. Create the instrument using the `interaction()` function.

```
# Create instrument
ak_sub$qobyob <- interaction(ak_sub$qob_f, ak_sub$yob_f)
```

You are now ready to estimate the models. You should estimate all models using the `lm()` and `ivreg()` functions. The main coefficients for each model should be nearly identical to those shown in Table V. The exception is the coefficient for `ageq` and `agesq`. The coefficients for these variables will be slightly different (not sure why!). Standard errors may also differ somewhat.

8. Estimate models 1-8 using the `lm()` and `ivreg()` functions.

```
# Column 1: OLS with yob dummies
col1 <- lm(lnw ~ educ + yob_f, data = ak_sub)

# Column 2: TSLS with yob dummies
col2 <- ivreg(lnw ~ educ + yob_f |
                qobyob + yob_f, data = ak_sub)

# Column 3: OLS with yob, ageq, and agesq controls
col3 <- lm(lnw ~ educ + yob_f + ageq + agesq, data = ak_sub)

# Column 4: TSLS with yob, ageq, and agesq controls
col4 <- ivreg(lnw ~ educ + yob_f + ageq + agesq |
                qobyob + yob_f + ageq + agesq, data = ak_sub)

# Column 5: OLS with race, smsa, married, yob, and region controls
col5 <- lm(lnw ~ educ + yob_f + race + smsa + married + yob_f + neweng +
             midatl + soatl + enocent + esocent + wnocent + wsocent + mt,
           data = ak_sub)

# Column 6: OLS with race, smsa, married, yob, and region controls
col6 <- ivreg(lnw ~ educ + yob_f + race + smsa + married + yob_f + neweng +
                midatl + soatl + enocent + esocent + wnocent + wsocent + mt |
              qobyob + yob_f + race + smsa + married + yob_f + neweng + midatl +
                soatl + enocent + esocent + wnocent + wsocent + mt,
              data = ak_sub)

# Column 7: OLS with full controls
col7 <- lm(lnw ~ educ + yob_f + race + smsa + married + ageq + agesq + yob_f + neweng +
             midatl + soatl + enocent + esocent + wnocent + wsocent + mt,
           data = ak_sub)

# Column 8 : TSLS with full controls
col8 <- ivreg(lnw ~ educ + yob_f + race + smsa + married + ageq + agesq + yob_f + neweng +
                midatl + soatl + enocent + esocent + wnocent + wsocent + mt |
                qobyob + yob_f + race + smsa + married + ageq + agesq + yob_f + neweng +
                midatl + soatl + enocent + esocent + wnocent + wsocent + mt, data = ak_sub)
```

The next step is to create a table that summarizes the results of the models. You should use `modelsummary` to do so. If you are up for the challenge, you might see if you can add rows to the table that show whether or not the model includes year-of-birth dummies and region-of-residence dummies (along the lines of what the authors have done in their own Table V).

9. Create a table that summarizes the results of the models. You should use `modelsummary` to do so.

NOTE: The `modelsummary` table will exceed the width of the PDF document upon knitting. One solution is to set `output = "kableExtra"` inside `modelsummary` and then pipe the output to `kableExtra::kable_styling()`. This will allow you to use the `latex_options = "scale_down"` option in `kableExtra`, which will scale the table down to fit within the PDF document. This option requires several latex packages, which I have loaded in the `YAML` header at the top of this RMD. It also requires that we add `results = 'asis'` to the code chunk options. This allows the table to be printed as-is, without being wrapped in a `kable` function. Here is an example of how to do this:

```
# You may need to install the kableExtra package
model_summary(models, output = "kableExtra") %>%
  kableExtra::kable_styling(latex_options = "scale_down")
```

```
# It is not necessary to add the rows, as I have below. But
# it is a nice touch!

# Grab models
models <- list(col1, col2, col3, col4, col5, col6, col7, col8)

# Add rows to table
rows <- tibble::tribble(~IV, ~i, ~ii, ~iii, ~iv, ~v, ~vi, ~vii, ~viii,
                        "9 Year-of-birth dummies", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes",
                        "8 Region-of-residence dummies", "No", "No", "No", "No", "Yes", "Yes", "Yes", "Y
attr(rows, "position") <- c(9, 10)

# Create table
modelsummary(models,
             output = "kableExtra",
             stars = TRUE,
             fmt = 4,
             gof_omit = "AIC|BIC|Log.Lik|F|RMSE",
             title = "Table V: OLS and TSLS Estimates of the Return to Education for Men Born 1930-1939
             coef_map = c("educ" = "Years of education",
                          "race" = "Race (1 = black)",
                          "smsa" = "SMSA (1 = center city)",
                          "married" = "Married (1 = married)",
                          "ageq" = "Age",
                          "agesq" = "Age-squared"),
             add_rows = rows) %>%
  kableExtra::kable_styling(latex_options = "scale_down")
```

Wrap up

When you have finished:

- Knit the RMD to PDF.
- Review the PDF for completeness and accuracy.
- Submit the RMD and PDF via email.

Table 1: Table V: OLS and TSLS Estimates of the Return to Education for Men Born 1930-1939: 1980 Census

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.0711***	0.0891***	0.0711***	0.0754**	0.0632***	0.0806***	0.0632***	0.0595*
	(0.0003)	(0.0161)	(0.0003)	(0.0289)	(0.0003)	(0.0164)	(0.0003)	(0.0290)
Race (1 = black)					−0.2575***	−0.2302***	−0.2575***	−0.2633***
					(0.0040)	(0.0261)	(0.0040)	(0.0458)
SMSA (1 = center city)					−0.1763***	−0.1581***	−0.1763***	−0.1802***
					(0.0029)	(0.0174)	(0.0029)	(0.0305)
Married (1 = married)					0.2479***	0.2440***	0.2479***	0.2487***
					(0.0032)	(0.0049)	(0.0032)	(0.0073)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age			−3.0694	−3.2162			−3.0028	−2.8830
			(2.6767)	(2.8539)			(2.6040)	(2.7681)
Age-squared			0.0008	0.0008			0.0008	0.0007
			(0.0007)	(0.0007)			(0.0007)	(0.0007)
Num.Obs.	329 509	329 509	329 509	329 509	329 509	329 509	329 509	329 509
R2	0.118	0.110	0.118	0.117	0.165	0.158	0.165	0.165
R2 Adj.	0.118	0.110	0.118	0.117	0.165	0.158	0.165	0.165

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$