Social Science Inquiry II

Week 4: Joint relationships, part I

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2024

Loading packages for this class

```
> library(ggplot2)
```

Homework

- Solution sets will be posted *at the same time* as problem sets.
- Do as much as you can on the problem set before checking the solutions.
- Check your work, and then fill out a form on how you did, what you understood and didn't.
- You get marked both on completion of the problem set, **AND** filling out the form.
- (If you find errors in the solution set, post them on the class StackOverflow and you will get extra credit)
- For homework assignments, always submit *both* your .R file showing your work, and and a compiled .pdf file on Canvas.

Homework grading

check(+/-)

- Check: You fully completed the assignment, and submitted all components. (A)
- Check plus: You went above and beyond, your solutions were clear and detailed. (A+)
- Check minus: You made an attempt, but it wasn't complete. Maybe you didn't submit all components, or didn't fully answer some of the questions. (B or C)
- Unmarked: You did not submit enough of an assignment for credit.

Angrist, Joshua D., and Alan B. Krueger. (1991) "Does compulsory school attendance affect schooling and earnings?"

An aside on Nobels

Figure: Joshua Angrist, Guido Imbens, David Card

An aside on Nobels

Figure: Alan Krueger

Krueger and Card were economists at Princeton when they started collaborating (Card started his career at the Booth School). Josh Angrist started collaborating on the returns to schooling project as a PhD student in the department.

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question? For empirical papers:
  - Data (Where does it come from/how is it generated? What is the sample population? What is being measured?)
  - Research design/strategy
  - Statistical tools
- What is the answer that the authors get to the main question?

How would you answer these questions with the Angrist and Keueger (1991) paper?

Establishing evidence: relationship between birth quarter and education

Loading the data

```
> dat <- read.csv('../data/angrist-krueger.csv', as.is = TRUE)
> head(dat)
  log_weekly_wage education year_of_birth quarter_of_birth place_of_birth
1        5.790019        12            30                1             45
2        5.952494        11            30                1             45
3        5.315949        12            30                1             45
4        5.595926        12            30                1             45
5        6.068915        12            30                1             37
6        5.793871        11            30                1             45
```

Examining the data

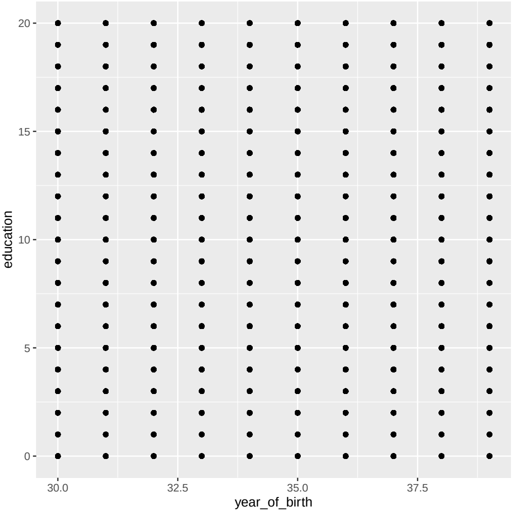
```
> str(dat)
'data.frame':	329509 obs. of  5 variables:
 $ log_weekly_wage : num  5.79 5.95 5.32 5.6 6.07 ...
 $ education       : int  12 11 12 12 12 11 11 12 11 7 ...
 $ year_of_birth   : int  30 30 30 30 30 30 30 30 30 30 ...
 $ quarter_of_birth: int  1 1 1 1 1 1 1 1 1 1 ...
 $ place_of_birth  : int  45 45 45 45 37 45 36 51 45 45 ...
> summary(dat)
 log_weekly_wage    education     year_of_birth  quarter_of_birth
 Min.   :-2.342   Min.   : 0.00   Min.   :30.0   Min.   :1.000   
 1st Qu.: 5.637   1st Qu.:12.00   1st Qu.:32.0   1st Qu.:2.000   
 Median : 5.952   Median :12.00   Median :35.0   Median :3.000   
 Mean   : 5.900   Mean   :12.77   Mean   :34.6   Mean   :2.506   
 3rd Qu.: 6.257   3rd Qu.:15.00   3rd Qu.:37.0   3rd Qu.:3.000   
 Max.   :10.532   Max.   :20.00   Max.   :39.0   Max.   :4.000   
 place_of_birth 
 Min.   : 1.00  
 1st Qu.:19.00  
 Median :34.00  
 Mean   :30.69  
 3rd Qu.:42.00  
 Max.   :56.00  
```

Data

- Where does it come from/how is it generated?
- What is the sample population?
- What is being measured?

Data exploration: Education on birth year

```
> ggplot(dat, aes(x = year_of_birth, y = education)) +
+   geom_point()
```



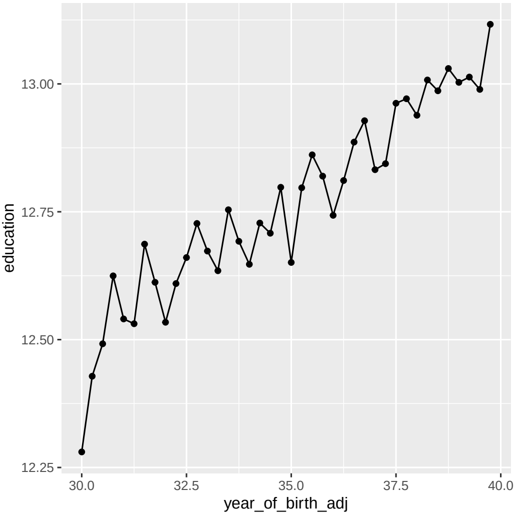
Data exploration: Education on birth year

```
> dat_agg <- aggregate(x = dat[, c('log_weekly_wage', 'education')],
+                      by = list(`year_of_birth` = dat$year_of_birth,
+                                `quarter_of_birth` = dat$quarter_of_birth),
+                      FUN = mean)
> dat_agg$year_of_birth_adj <- dat_agg$year_of_birth +
+   0.25 * (dat_agg$quarter_of_birth-1)
> head(dat_agg)
```

```
  year_of_birth quarter_of_birth log_weekly_wage education year_of_birth_adj
1            30                1        5.889133  12.28041                30
2            31                1        5.902136  12.54043                31
3            32                1        5.899809  12.53393                32
4            33                1        5.891946  12.67319                33
5            34                1        5.895157  12.64726                34
6            35                1        5.879843  12.65091                35
```

Data exploration: Education on birth year

```
> ggplot(dat_agg, aes(x = year_of_birth_adj, y = education)) +
+   geom_point() + # points
+   geom_line() # lines
```



Data exploration: Education on birth year

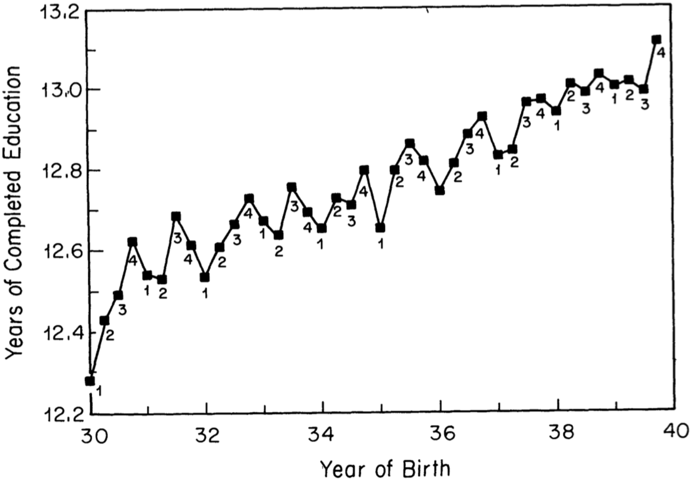


FIGURE I

Years of Education and Season of Birth

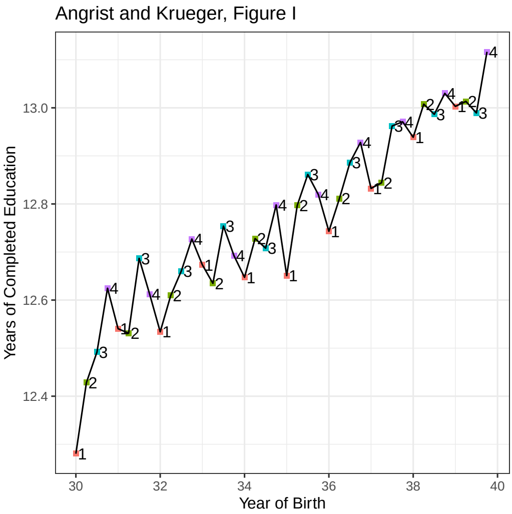
1980 Census

Note. Quarter of birth is listed below each observation.

Data exploration: Education on birth year

```
> ggplot(dat_agg, aes(x = year_of_birth_adj,
+                     y = education,
+                     label = quarter_of_birth)) +
+   geom_point(pch = 15,
+              aes(color = as.factor(quarter_of_birth) )) + # points with color
+   geom_line() + # lines
+   geom_text(hjust = 0, nudge_x = 0.05) + # text labels on points
+   theme_bw() + # plot style
+   theme(legend.position = '') + # remove legend from colored text labels
+   ylab('Years of Completed Education') +  # y-axis label
+   xlab('Year of Birth') + # x-axis label
+   ggtitle('Angrist and Krueger, Figure I') + # title
+   scale_x_continuous(breaks = seq(30, 40, 2)) + # x-axis ticks
+   scale_y_continuous(breaks = seq(12.2, 13.2, .2)) # y-axis ticks
>
```

Data exploration: Education on birth year



Data exploration: Education on birth year

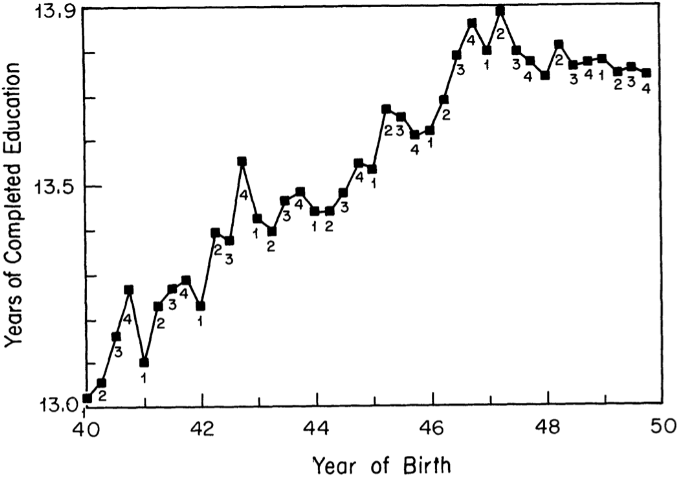


FIGURE II
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

Data exploration: Education on birth year

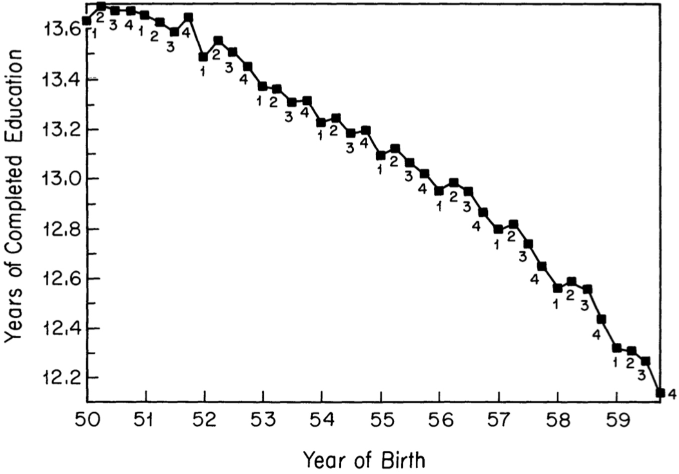


FIGURE III

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

Data exploration: Education on birth year

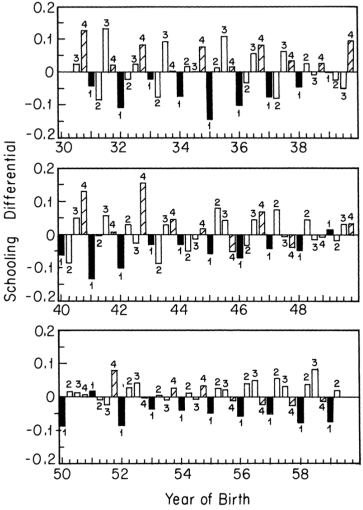


FIGURE IV
Season of Birth and Years of Schooling
Deviations from $MA(+2,-2)$

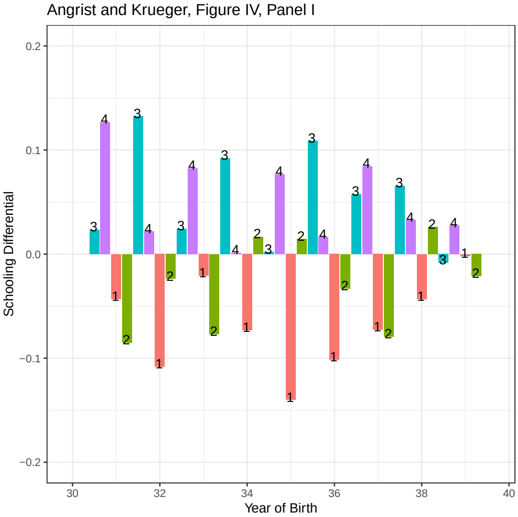
Data exploration: Education on birth year

```
> # function for moving average
> ma <- function(x, n = 5){
+   ma_x <- as.numeric(filter(x, rep(1 / n, n), sides = 2))
+   ma_x2 <- (ma_x - x/5)*5/4
+   return(ma_x2)
+ }
> # get dat_agg in right order
> dat_agg <- dat_agg[order(dat_agg$year_of_birth_adj),]
> # calculate moving average
> dat_agg$moving_average <- ma(dat_agg$education)
> # update adjusted birth year in main dataset
> dat$year_of_birth_adj <- dat$year_of_birth + 0.25 * (dat$quarter_of_birth-1)
> # and match aggregated moving average to main data
> dat$moving_average <- dat_agg$moving_average[match(dat$year_of_birth_adj,
+                                                   dat_agg$year_of_birth_adj)]
> # calculate deviation from moving average
> dat$deviation <- dat$education-dat$moving_average
> # get aggregate deviation
> dat_agg$deviation <- aggregate(x = dat$deviation,
+                               by = list(dat$year_of_birth_adj), mean)$x
>
```

Data exploration: Education on birth year

```
> ggplot(dat_agg, aes(x = year_of_birth_adj,
+                    y = deviation,
+                    fill = as.factor(quarter_of_birth),
+                    label = quarter_of_birth)) +
+   geom_col(na.rm = TRUE) +
+   geom_text(hjust = 0, nudge_y = 0.003, nudge_x = -0.1, na.rm = TRUE) + # text la
+   coord_cartesian(ylim = c(-0.2, 0.2)) +
+   theme_bw() + # plot style
+   theme(legend.position = '') + # remove legend from colored text labels
+   ylab('Schooling Differential') +  # y-axis label
+   xlab('Year of Birth') + # x-axis label
+   ggtitle('Angrist and Krueger, Figure IV, Panel I') + # title
+   scale_x_continuous(breaks = seq(30, 40, 2)) # x-axis ticks
>
```

Data exploration: Education on birth year



Data exploration: Education on birth year

What is the case that the difference in education across quarters is due to compulsory schooling?

Inference

- Over what population do these effects apply?
    - Time frame
    - Geography
    - Policy

Policy implications

What should we do with this evidence?

- Should we change compulsory school attendance laws in the US?
- If you were hired as a consultant for another country, would you recommend to change compulsory school attendance laws? Under what conditions?

Estimating causal effects: returns to education

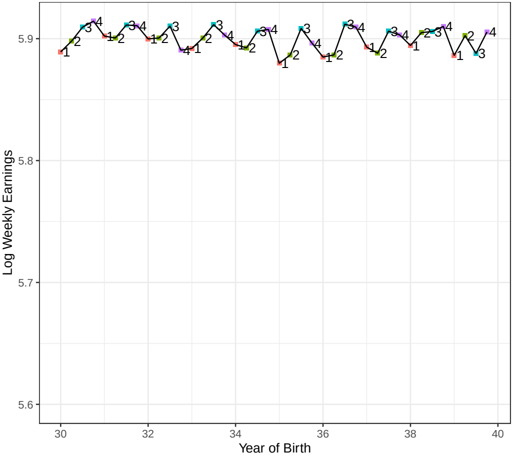
Data analysis: Returns to education

```
> ggplot(dat_agg, aes(x = year_of_birth_adj, y = log_weekly_wage,
+                     label = quarter_of_birth)) +
+   geom_point(pch = 15,
+              aes(color = as.factor(quarter_of_birth) )) + # points with color
+   geom_line() + # lines
+   geom_text(hjust = 0, nudge_x = 0.05) + # text labels on points
+   theme_bw() + # plot style
+   theme(legend.position = '') + # remove legend from colored text labels
+   scale_x_continuous(breaks = seq(30, 40, 2)) + # x-axis ticks
+   scale_y_continuous(breaks = seq(5.6, 6.1, .1)) + # y-axis ticks
+   coord_cartesian(ylim = c(5.6, NA)) +
+   ylab('Log Weekly Earnings') +  # y-axis label
+   xlab('Year of Birth') + # x-axis label
+   ggtitle('Angrist and Krueger, Figure V',
+           subtitle = 'Mean Log Weekly Wage, by Quarter of Birth\nAll Men Born 193
```

Data analysis: Returns to education

Angrist and Krueger, Figure V

Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930−1949; 1980 Census



Data analysis: Returns to education

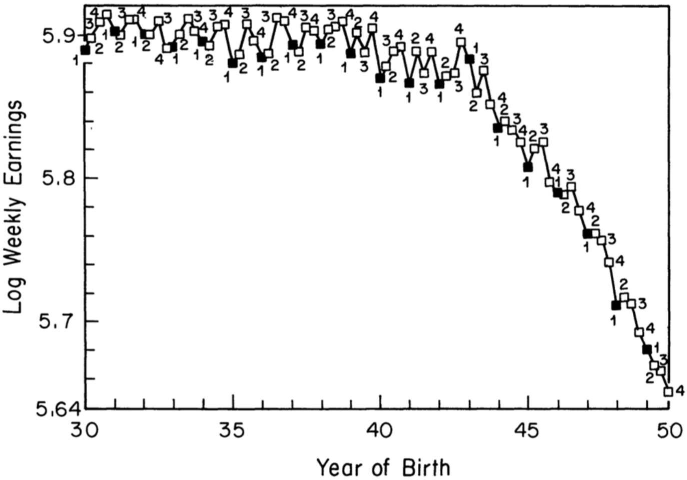


Figure V
Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930–1949; 1980 Census

Wald estimator

Computes returns to education as ratio:

- numerator: the difference in earning by quarter of birth
- denominator: the difference in education by quarter of birth
- comparison: men born in first quarter vs. men born in last three quarters

Wald estimator

The Wald estimator is a simple example of **Instrumental Variables** analysis, where you *instrument* for changes in X with changes in some instrument, Z.

$$\beta_{IV} = \frac{\delta y/\delta z}{\delta x/\delta z}$$

With the Wald estimator, Z is binary.

$$\hat{\beta}_{Wald} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

Data analysis: Returns to education

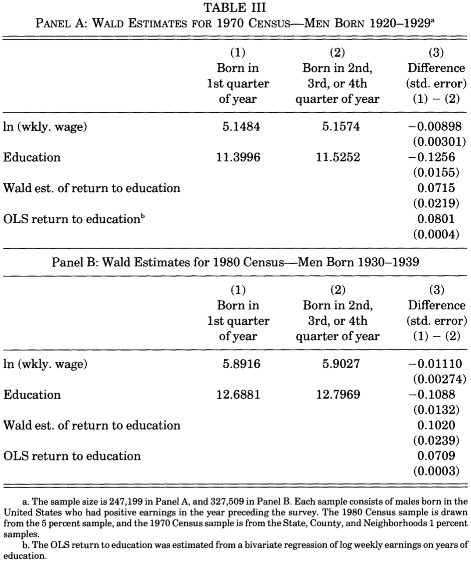


TABLE III

Panel A: Wald Estimates for 1970 Census—Men Born 1920–1929[a]

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) − (2)
ln (wkly. wage)	5.1484	5.1574	−0.00898 (0.00301)
Education	11.3996	11.5252	−0.1256 (0.0155)
Wald est. of return to education			0.0715 (0.0219)
OLS return to education[b]			0.0801 (0.0004)

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) Difference (std. error) (1) − (2)
ln (wkly. wage)	5.8916	5.9027	−0.01110 (0.00274)
Education	12.6881	12.7969	−0.1088 (0.0132)
Wald est. of return to education			0.1020 (0.0239)
OLS return to education			0.0709 (0.0003)

a. The sample size is 247,199 in Panel A, and 327,509 in Panel B. Each sample consists of males born in the United States who had positive earnings in the year preceding the survey. The 1980 Census sample is drawn from the 5 percent sample, and the 1970 Census sample is from the State, County, and Neighborhoods 1 percent samples.

b. The OLS return to education was estimated from a bivariate regression of log weekly earnings on years of education.

What makes this paper so compelling? What is its contribution to research methods in the social sciences?

References I

Angrist, J. D. and Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics, 106(4):979–1014.