Assignment 4, Social Science Inquiry II (SOSC13200-W22-3)

Monday 2/7/22 at 5pm

Packages

```
library(ggplot2)
```

Read in the data. We will use the data from:

Angrist, Joshua D., and Alan B. Krueger. "Does compulsory school attendance affect schooling and earnings?" The Quarterly Journal of Economics 106.4 (1991): 979-1014.

```
file <- "https://raw.githubusercontent.com/UChicago-pol-methods/SOSC13200-W22/main/data/angrist-krueger
dat <- read.csv(file, as.is = TRUE)
```

1.

Consider Angrist and Kruger (1991) Table III Panel B on p. 996. We have the data for the 1980 census, for men born 1930-1939–we don't have the 1920-1929 data, so you can ignore Panel A.

(1a)

Calculate the mean log weekly wage for men born in the first quarter of the year, and men born in any other quarter of the year. Calculate the difference, and save the difference as an R object.

```
(lnwage1 <- mean(dat$log_weekly_wage[which(dat$quarter_of_birth==1)]))
```

```
## [1] 5.891596
```

```
(lnwage2 <- mean(dat$log_weekly_wage[which(dat$quarter_of_birth!=1)]))
```

```
## [1] 5.902695
```

```
(lnwagediff <- lnwage1 - lnwage2)
```

```
## [1] -0.01109888
```

(1b)

Calculate the mean years education for men born in the first quarter of the year, and men born in any other quarter of the year. Calculate the difference, and save the difference as an R object.

```
(ed1 <- mean(dat$education[which(dat$quarter_of_birth==1)]))
```

```
## [1] 12.68807
```

```
(ed2 <- mean(dat$education[which(dat$quarter_of_birth!=1)]))
```

```
## [1] 12.79688
```

```
(eddiff <- ed1 - ed2)
```

```
## [1] -0.1088179
```

(1c)

Calculate the Wald estimate of the returns to education as the ratio of the difference in mean log earnings by quarter of birth to the difference in years of mean education by quarter of birth. Compare your results to Table III Panel B. Are they the same?

```
(wald_est <- lnwagediff/eddiff)
```

```
## [1] 0.101995
```

Interpret the estimate in words.

```
# [Your explanation here].
```

2.

(2a)

In the Angrist and Krueger data, create a new variable, `year_of_birth_adj`, which adds a quarter on to year of birth for each quarter in quarter of birth *after the first quarter*. For example, if a person was born in 1930 Q2, their `year_of_birth_adj` value would be 1930 + 0.25 * (2-1) = 1930.25.

```
dat$year_of_birth_adj <- dat$year_of_birth +
  0.25 * (dat$quarter_of_birth-1)
```

Then create a new variable, `states_above_16`, which is a 1 when the the age for compulsory schooling is above 16, and 0 otherwise. Check Appendix 2 for the list of ages for compulsory school attendance *in 1980*. Compare this to the values of the place of birth variable in the data set.

```
states_above_16 <- c(15, 23, 32, 35, 39, 40, 41, 42, 48, 49, 51, 53)
dat$states_above_16 <- 1 * (dat$place_of_birth %in% states_above_16)
```

(2b)

Using the `aggregate()` function, group the data set by adjusted year of birth, quarter of birth, AND whether the state has a compulsory schooling age above 16, and calculate mean log weekly wage and mean education within each of the subgroups. Save this as a new data.frame object in R.

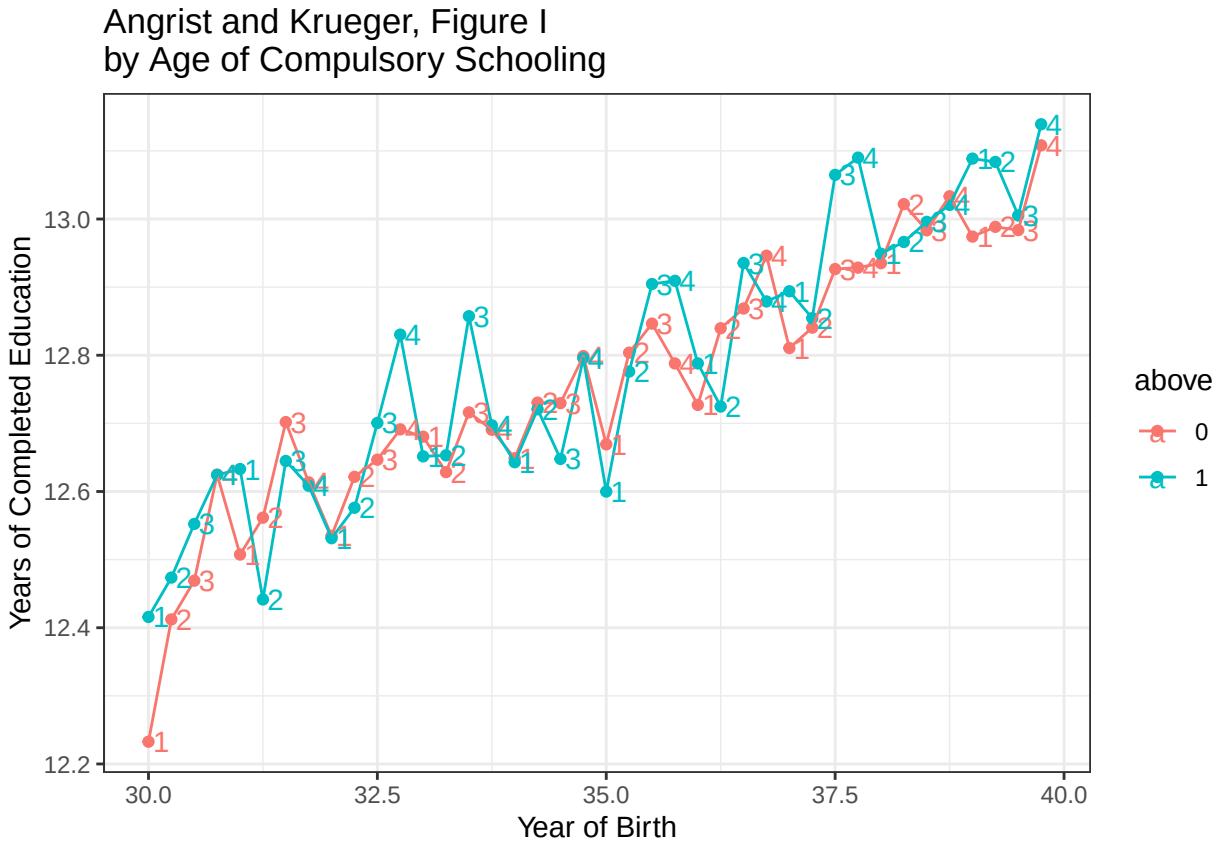
(Note: you could aggregate just by adjusted year of birth, as this uniquely describes quarters, but I would like you to also have quarter of birth as a variable in your new dataset.)

```
dat_agg <- aggregate(x = dat[, c('log_weekly_wage', 'education')],
                     by = list(`year_of_birth_adj` = dat$year_of_birth_adj,
                               `quarter_of_birth` = dat$quarter_of_birth,
                               `above` = as.factor(dat$states_above_16)),
                     FUN = mean)
```

(2c)

Create a plot of your aggregated data, using both points and lines, with adjusted year of birth on the x-axis, and education on the y-axis. Separately plot data for states with compulsory school ages above 16 and for 16 and below by setting the color in the plot aesthetic.

```
ggplot(dat_agg, aes(x = year_of_birth_adj,
                    y = education,
                    label = quarter_of_birth,
                    color = above)) +
  geom_point() + # points with color
  geom_line() + # lines
  geom_text(hjust = 0, nudge_x = 0.05) + # text labels on points
  theme_bw() + # plot style
  ylab('Years of Completed Education') +  # y-axis label
  xlab('Year of Birth') + # x-axis label
  ggtitle('Angrist and Krueger, Figure I\nby Age of Compulsory Schooling') # title
```



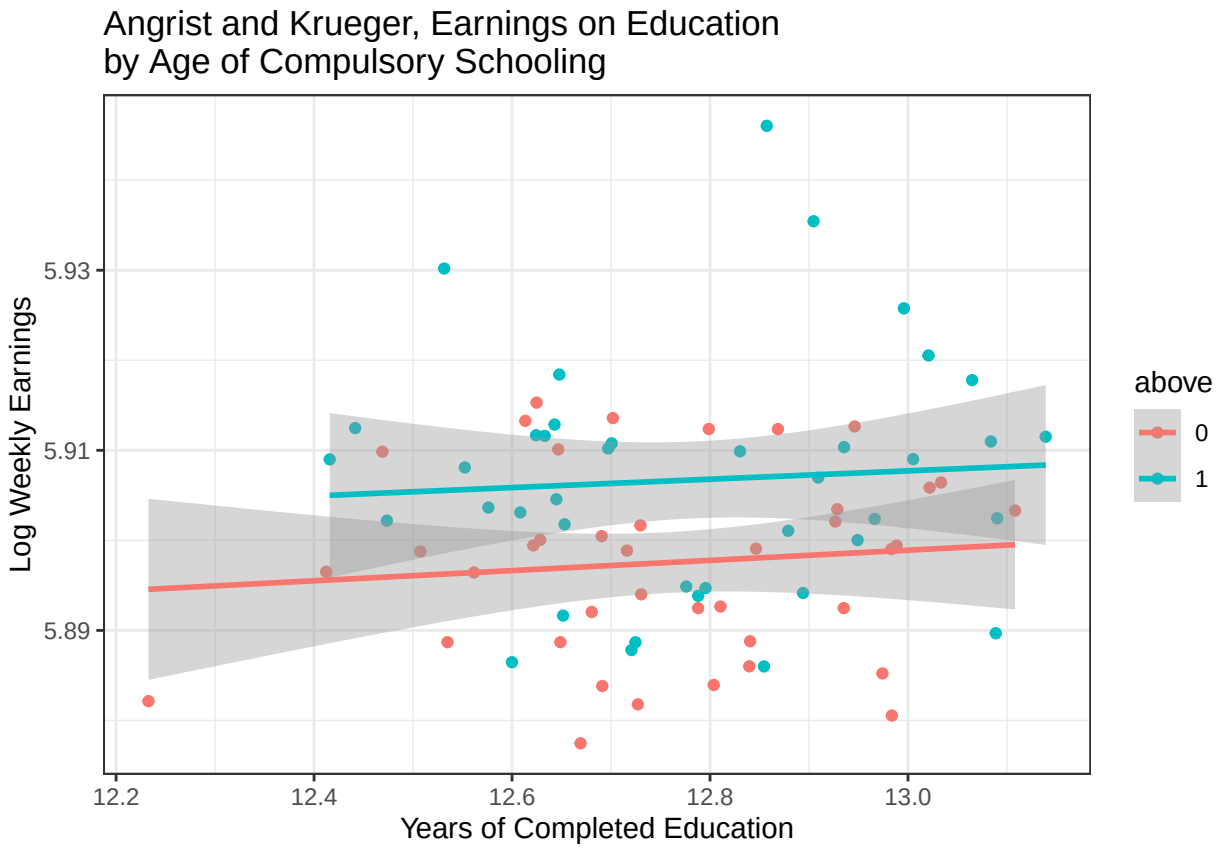
(2d)

Create a plot of your aggregated data, with education on the x-axis, and log weekly wages on the y-axis; add a layer for points, and then show a smoothed line demonstrating the trend across points with `geom_smooth(method = 'lm')`. Separately plot data for states with compulsory school ages above 16 and for 16 and below by setting the color in the plot aesthetic.

```
ggplot(dat_agg, aes(x = education,
                    y = log_weekly_wage,
```

```
                    color = above)) +
  geom_point() + # points with color
  geom_smooth(method = 'lm') + # lines
  theme_bw() + # plot style
  ylab('Log Weekly Earnings') +  # y-axis label
  xlab('Years of Completed Education') + # x-axis label
  ggtitle('Angrist and Krueger, Earnings on Education\nby Age of Compulsory Schooling') # title
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Do you see differences in trends across states with age of compulsory schooling above 16 and 16 and below?

```
# [Your explanation here]
```

3.

(3a)

Redo your calculations from question 1, but separately for states with compulsory school ages above 16 and for 16 and below.

```
# Earnings differences
(lnwage11 <- mean(dat$log_weekly_wage[which(dat$quarter_of_birth==1 & dat$states_above_16 == 1)]))
```

```
## [1] 5.902041
```

```
(lnwage21 <- mean(dat$log_weekly_wage[which(dat$quarter_of_birth!=1& dat$states_above_16 == 1)]))
```

```
## [1] 5.908638
```

```
(lnwagediff1 <- lnwage11 - lnwage21)
```

```
## [1] -0.00659704
```

```
(lnwage10 <- mean(dat$log_weekly_wage[which(dat$quarter_of_birth==1 & dat$states_above_16 == 0)]))
```

```
## [1] 5.887929
```

```
(lnwage20 <- mean(dat$log_weekly_wage[which(dat$quarter_of_birth!=1& dat$states_above_16 == 0)]))
```

```
## [1] 5.900605
```

```
(lnwagediff0 <- lnwage10 - lnwage20)
```

```
## [1] -0.01267616
```

```
# Education differences
(ed11 <- mean(dat$education[which(dat$quarter_of_birth==1& dat$states_above_16 == 1)]))
```

```
## [1] 12.72104
```

```
(ed21 <- mean(dat$education[which(dat$quarter_of_birth!=1& dat$states_above_16 == 1)]))
```

```
## [1] 12.81313
```

```
(eddiff1 <- ed11 - ed21)
```

```
## [1] -0.09208688
```

```
(ed10 <- mean(dat$education[which(dat$quarter_of_birth==1& dat$states_above_16 == 0)]))
```

```
## [1] 12.67649
```

```
(ed20 <- mean(dat$education[which(dat$quarter_of_birth!=1& dat$states_above_16 == 0)]))
```

```
## [1] 12.79117
```

```
(eddiff0 <- ed10 - ed20)
```

```
## [1] -0.114683
```

```
# Final estimates
(wald_est1 <- lnwagediff1/eddiff1)
```

```
## [1] 0.07163931
```

```
(wald_est0 <- lnwagediff0/eddiff0)
```

```
## [1] 0.1105322
```

(3b)

Do you get different estimates for the two conditions? If so, propose an explanation for why returns to education might be different in these two cases. If you think the results are not meaningfully different, make a case for why we should not see a difference.

```
# [Your explanation here]
```