Assignment 4

Vincent Zhang

Due on 2/8/22 @ 11:59pm

Overview

You will be working with Angrist & Krueger's data from "Does compulsory school attendance affect school and earnings?" (1991).

You can download the file from Canvas/Files/Datasets: The dataset is entitled `angrist_krueger_1991.csv`.

Once you have downloaded the data, load it.

(1)

Examine Panel B of Table III on p. 996. We have data for the 1980 census (i.e., for men born 1930-39). We do not have the data for men born 1920-29, so we will ignore Panel A.

(1a)

Calculate the mean log weekly wage for men born in the first quarter of the year. Then calculate the mean log weekly wage for men born in any other quarter of the year (i.e., 2-4). Calculate the difference. Store the difference as a new R object.

```
mean_wage_first_q <- mean((subset(angrist_krueger_dat,
                                  quarter_of_birth == 1))$log_weekly_wage)
mean_wage_other_q <- mean((subset(angrist_krueger_dat,
                                  quarter_of_birth != 1))$log_weekly_wage)

mean_wage_diff <- mean_wage_first_q - mean_wage_other_q
mean_wage_diff
```

```
## [1] -0.01109435
```

(1b)

Calculate the mean years of education for men born in the first quarter of the year. Then calculate the mean years of education for men born in any other quarter of the year (i.e., 2-4). Calculate the difference. Store the difference as a new R object.

```
mean_edu_first_q <- mean((subset(angrist_krueger_dat,
                                 quarter_of_birth == 1))$education)
mean_edu_other_q <- mean((subset(angrist_krueger_dat,
                                 quarter_of_birth != 1))$education)

mean_edu_diff <- mean_edu_first_q - mean_edu_other_q
mean_edu_diff
```

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

(1c)

Calculate the Wald estimate of the returns to education. The estimate is described by the authors on p. 995. It is also broken down on Tuesday's lecture slides. Compare your results to Table III Panel B. Are they the same?

```
wald_estimate <- mean_wage_diff / mean_edu_diff
```

Interpret the estimate in words. *HINT: Recall from discussion that here, the authors are estimating the rate of return to a year of education. You may wish to re-read pp. 994-997.*

Consider that the wald estimate is the ratio between the difference in wage to the difference in education. Then, the calculation is giving some sort of estimation of how much on average the wage increases *more* in men born in q1 *compared* to men born in q2 PER each yearly increase in educational year.

2.

(2a)

Add a new variable to the dataset named `year_of_birth_adj`. The variable should add one quarter to the year of birth for each quarter in quarter of birth. For example: If a person was born in 1930 Q2, their `year_of_birth_adj` value would be 1930 + 0.25 * 2 = 1930.5.

```
angrist_krueger_dat$year_of_birth_adj <- 0 + angrist_krueger_dat$year_of_birth +
                                         0.25*(angrist_krueger_dat$quarter_of_birth-1)
```

Then create a new variable named `states_above_16`. Add it to the dataset. The variable equals 1 when the the age for compulsory schooling is greater than 16, and 0 otherwise. See Appendix 2 at the back of the article for a list of ages for compulsory school attendance *in 1980*. Compare this to the values of the place of birth variable in the dataset.

```
# note this is for 1980, also note its exclusively GREATER than 16:
states_above_16 <- c(15, 23, 32, 35, 39, 40, 41, 42, 48, 49, 51, 53)
angrist_krueger_dat$states_above_16 <-
  as.numeric(angrist_krueger_dat$place_of_birth %in% states_above_16)
```

(2b)

Use the `aggregate()` function to group the dataset by adjusted year of birth, quarter of birth, AND whether the state has a compulsory schooling age greater than 16. At the same time, `aggregate()` should calculate the mean log weekly wage and mean level of education within these subgroups. Save the output as a new data.frame object.

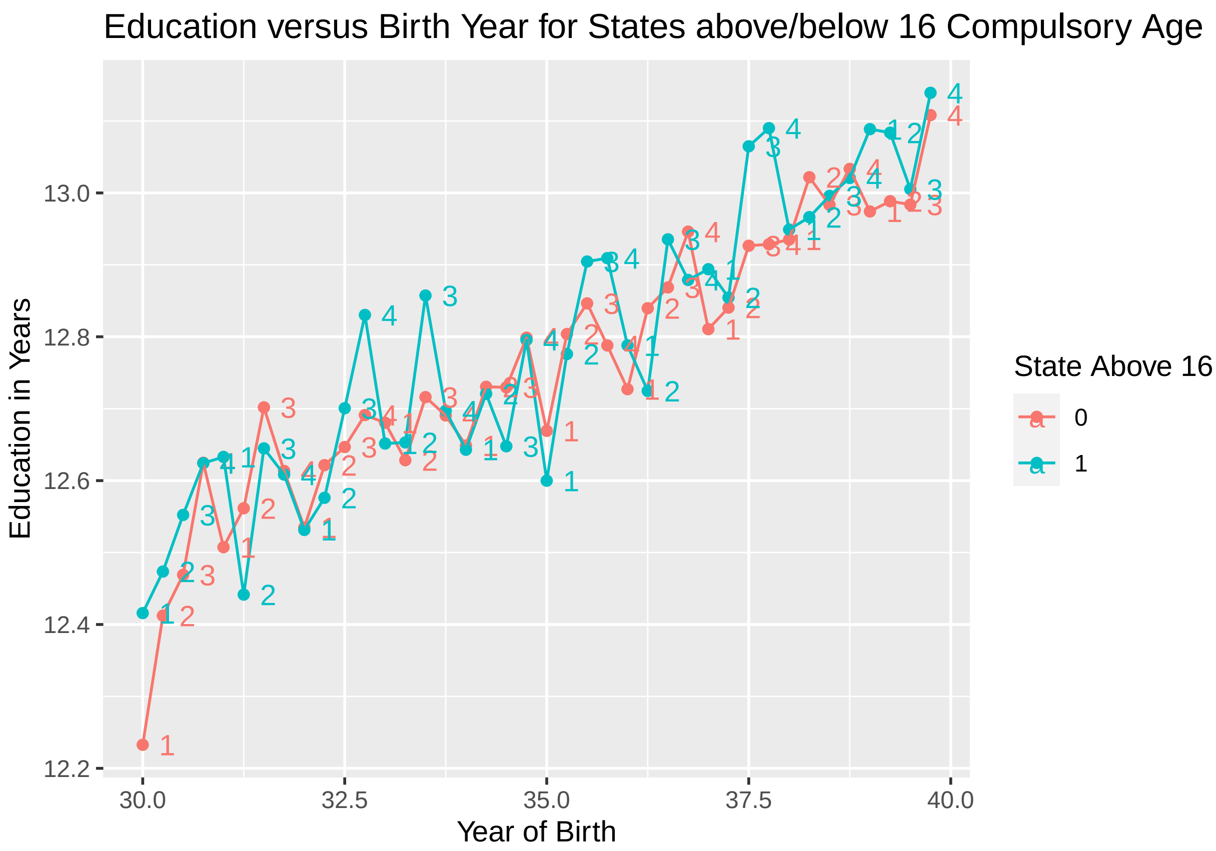
(NOTE: You could aggregate just by adjusted year of birth, as this uniquely describes quarters, but I would like you to also have quarter of birth as a variable in your new dataset.)

```
#agg_dat <- angrist_krueger_dat %>%
#            group_by(year_of_birth_adj, quarter_of_birth, above_16) %>%
#             summarize(log_weekly_wage = mean(log_weekly_wage),
#                      education = mean(education))

# I used the professors method to check that my result is what it should look like:
#
agg_dat <- aggregate(x = angrist_krueger_dat[, c("log_weekly_wage", "education")],
                    by = list(`year_of_birth_adj` = angrist_krueger_dat$year_of_birth_adj,
                              `quarter_of_birth` = angrist_krueger_dat$quarter_of_birth,
                              `above_16` = as.factor(angrist_krueger_dat$states_above_16)),
                    FUN = mean)
```

(2c)

Create a plot of your aggregated data. Use both points and lines. Adjusted year of birth should be on the x-axis, and education should be on the y-axis. Separately plot data for states with compulsory schooling ages greater than 16 and less than or equal to 16. To do this, set the color in the plot aesthetic.

```
agg_plot_1 <- ggplot(data = agg_dat, aes(x = year_of_birth_adj,
                                       y = education,
                                       label = quarter_of_birth,
                                       color = as.factor(above_16))) +
        labs(title = "Education versus Birth Year for States above/below 16 Compulsory Age",
              x = "Year of Birth",
              y = "Education in Years",
              color = "State Above 16") +
        geom_point() +
        geom_line() +
        geom_text(hjust = -1)

agg_plot_1
```

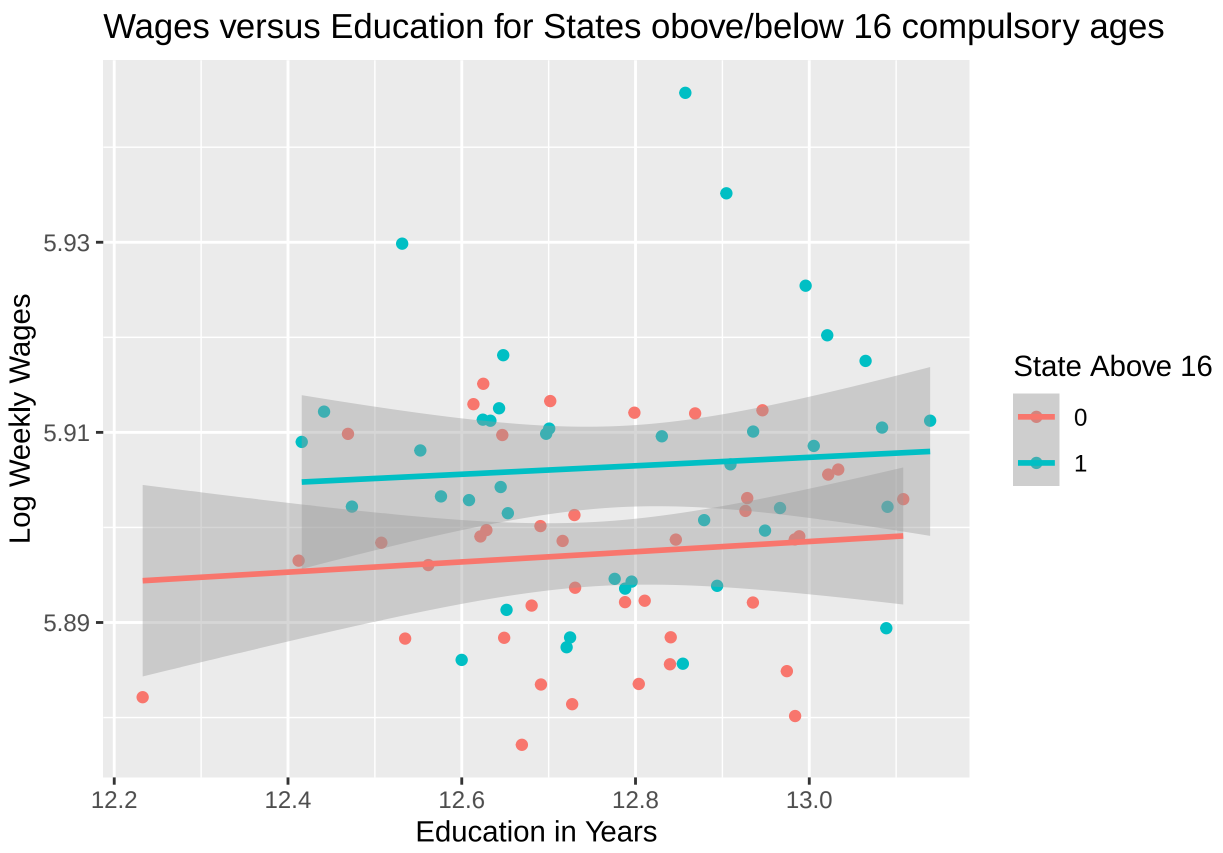


(2d)

Create a plot of your aggregated data. Education should be on your x-axis, and log weekly wages should be on your y-axis. Add a layer for points, and then plot a smooth line that demonstrates the trend across points. To do this, use `geom_smooth(method = "lm")`. Separately plot data for states with compulsory schooling ages greater than 16 and less than or equal to 16 by setting the color in the plot aesthetic.

```
agg_plot_2 <- ggplot(data = agg_dat, aes(x = education,
                                        y = log_weekly_wage,
                                        label = quarter_of_birth,
                                        color = as.factor(above_16))) +
        labs(title = "Wages versus Education for States obove/below 16 compulsory ages",
             x = "Education in Years",
             y = "Log Weekly Wages",
             color = "State Above 16") +
        geom_point() +
        geom_smooth(method = "lm")

agg_plot_2
```

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



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

The general trend, is that on average, states with age of compulsory schooling greater than 16 tend to have a greater weekly wage (amarginally small however) per educational year.

3.

(3a)

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

```
## above calculation
mean_wage_above_diff <- mean((subset(angrist_krueger_dat,
                                     (quarter_of_birth == 1 &
                                        states_above_16 == 1)))$log_weekly_wage) -
                        mean((subset (angrist_krueger_dat,
                                      (quarter_of_birth != 1 &
                                         states_above_16 == 1)))$log_weekly_wage)
mean_edu_above_diff <- mean((subset(angrist_krueger_dat,
                                    (quarter_of_birth == 1 &
                                       states_above_16 == 1)))$education) -
                       mean((subset(angrist_krueger_dat,
                                    (quarter_of_birth != 1 &
                                       states_above_16 == 1)))$education)

wald_above_estimate <- mean_wage_above_diff / mean_edu_above_diff


## below calculation
mean_wage_below_diff <- mean((subset(angrist_krueger_dat,
                                     (quarter_of_birth == 1 &
                                        states_above_16 == 0)))$log_weekly_wage) -
                        mean((subset(angrist_krueger_dat,
                                     (quarter_of_birth != 1 &
                                        states_above_16 == 0)))$log_weekly_wage)
mean_edu_below_diff <- mean((subset(angrist_krueger_dat,
                                    (quarter_of_birth == 1 &
                                       states_above_16 == 0)))$education) -
                       mean((subset(angrist_krueger_dat,
                                    (quarter_of_birth != 1 &
                                       states_above_16 == 0)))$education)

wald_below_estimate <- mean_wage_below_diff / mean_edu_below_diff

### above and below wages, educations, and wald estimates:
cat("Above Calculations:",
    "\nWage Difference:",
    mean_wage_above_diff,
    "\nEducation Difference:",
    mean_edu_above_diff,
    "\nWald Estimate:",
    wald_above_estimate)
```

```
## Above Calculations:
## Wage Difference: -0.006590999
## Education Difference: -0.09208688
## Wald Estimate: 0.07157371
```

```
cat("\n\nBelow Calculations:",
    "\nWage Difference:",
    mean_wage_below_diff,
    "\nEducation Difference:",
    mean_edu_below_diff,
    "\nWald Estimate:",
    wald_below_estimate)
```

```
##
##
## Below Calculations:
## Wage Difference: -0.01267215
## Education Difference: -0.114683
## Wald Estimate: 0.1104972
```

(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.

There is a difference in the wald estimate. Specifically, the wald estimate for states with compulsory schooling age above_16 is less than the wald estimate for schools that have compulsory schooling age below_16. I think that there is significance in the differences we see in the wald estimates. To see this, lets remember that the wald estimate, as we previously explained, provides some sort of metric to on how much the wage increases MORE for q1_birthdate people than q234_birthdate people, per yearly increase in education. As a result, the difference in wald estimate is suggesting that the difference in wage increase for people born in q1 compared to q234 per year in education is less in schools with compulsory schooling ages above_16 than in compulsory schooling ages below_16. At first, I thought that this meant that this was contradictory as in the aggregate plot, it seemed like the slopes in the aggregate plot was the same. But the i realized the wald estimate is not measuring that; instead, its measuring the DIFFERENCE between the ratio for people born in quarter_1 versus the ratio for people born in other quarters. So, it could be the case that the OVERALL trend of q1234 is the same for both above_16 and below_16, but that for above_16, the difference between q1 and q234 is much less than the difference between q1 and q234 in below_16.