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

Monday 1/30/22 at 5pm

Packages

```
library(ggplot2)
library("modelsummary")
```

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-W23/main/data/angrist-krueger
dat <- read.csv(file, as.is = TRUE)
logww <- dat$log_weekly_wage
q <- dat$quarter_of_birth
edu <- dat$education
```

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.

```
#mean log weekly wage for men born in first quarter:
mean(logww[q==1]) #5.891596
```

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

```
#mean log weekly wage for men born in any other quarter:
mean(logww[q!=1]) #5.902695
```

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

```
#difference:
diff.logww <- mean(logww[q==1])-mean(logww[q!=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.

```
#mean yrs education for men born in  first quarter:
mean(edu[q==1]) #12.68807
```

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

```
#mean yrs education for men born in any other quarter:
mean(edu[q!=1]) #12.79688
```

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

```
#difference:
diff.edu <- mean(edu[q==1])-mean(edu[q!=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?

```
diff.logww/diff.edu #0.101995
```

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

Interpret the estimate in words.

```
#My estimate, 0.101995, is the same as the literature value (0.1020). This tells me the indirect "effec
```

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*(q-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.

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

(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.

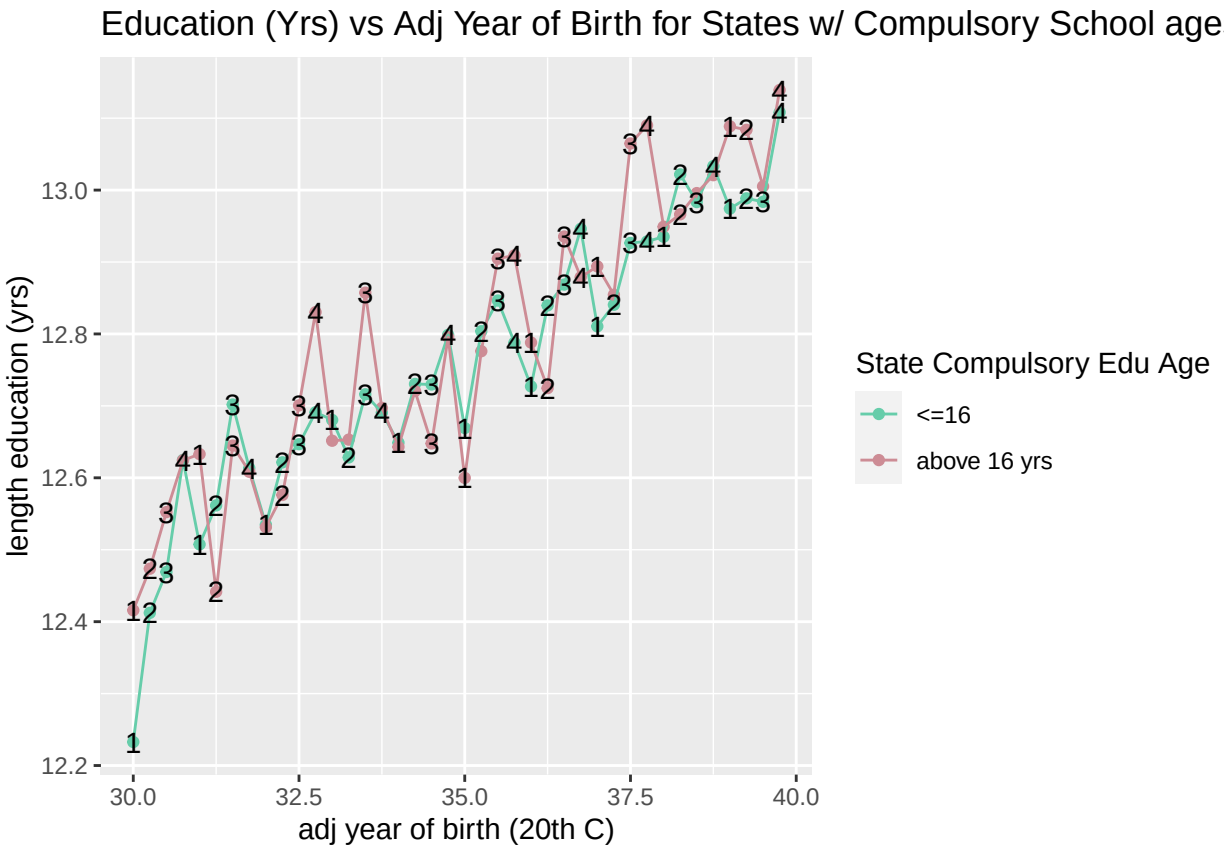
(Note: you could aggregate just by adjusted year of birth, as this uniquely describes quarters, but I would like you to also have quarter of birth as a variable in your new dataset.)

```
agg <- aggregate(dat[,1:2],list(`yr_born`=dat$year_of_birth_adj,`quarter`=q,`state`=as.factor(dat$state
```

(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(agg,aes(x=yr_born,y=education,color=state,label=quarter))+
  geom_point()+
  geom_line()+
  geom_text(check_overlap=TRUE,color="black")+
  labs(x="adj year of birth (20th C)",
       y="length education (yrs)",
       title="Education (Yrs) vs Adj Year of Birth for States w/ Compulsory School ages above/for+below
  scale_color_manual(labels=c("<=16","above 16 yrs"),values=c("aquamarine3","lightpink3"))
```



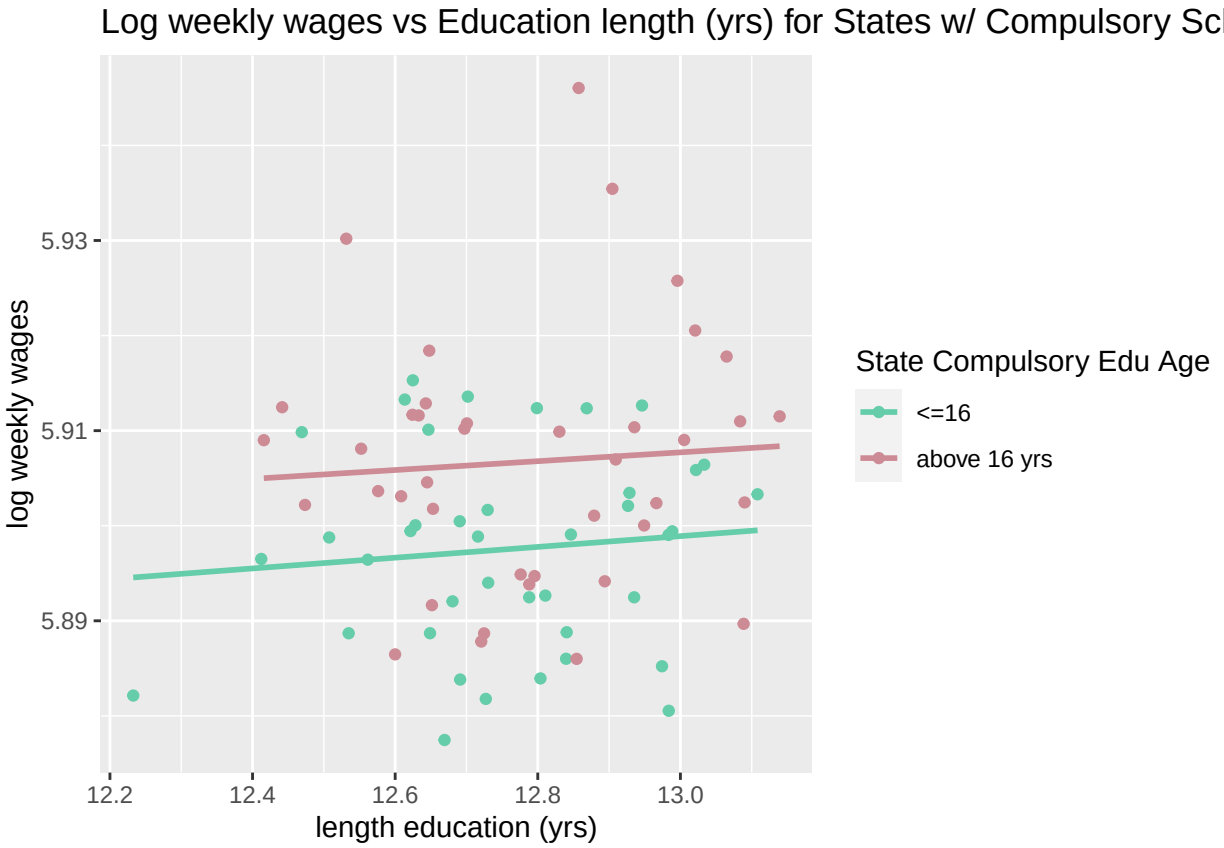

```
#state = 1 --> generally more education for any year of birth; gap is slightly smaller for those in q1
#makes sense b/c older --> get more education overall
```

(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(agg,aes(x=education,y=log_weekly_wage,color=state))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)+ #linear model
  labs(x="length education (yrs)",
       y="log weekly wages",
       title="Log weekly wages vs Education length (yrs) for States w/ Compulsory School ages above/for
  scale_color_manual(labels=c("<=16","above 16 yrs"),values=c("aquamarine3","lightpink3"))
```

```
## `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?

```
#After separating differences in compulsory schooling age being above vs =/< 16 years old and looking a

lm(log_weekly_wage~education,data=agg) #0.00684 -- adj year of birth
```

```
##
## Call:
## lm(formula = log_weekly_wage ~ education, data = agg)
##
## Coefficients:
## (Intercept)    education
##     5.81477      0.00684
```

```
lm(log_weekly_wage~education,data=agg[which(agg$state==0),]) #0.005663
```

```
##
## Call:
## lm(formula = log_weekly_wage ~ education, data = agg[which(agg$state ==
##     0), ])
##
## Coefficients:
## (Intercept)    education
##    5.825289     0.005663
```

```
lm(log_weekly_wage~education,data=agg[which(agg$state==1),]) #0.004648
```

```
##
## Call:
## lm(formula = log_weekly_wage ~ education, data = agg[which(agg$state ==
##     1), ])
##
## Coefficients:
## (Intercept)    education
##    5.847288     0.004648
```

3.

(3a)

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

```
st <- dat$states_above_16
#earnings:
#state above 16 (st=1):
  #mean log weekly wage for men born in first quarter:
  mean(logww[q==1&st==1]) #5.902041
```

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

```
  #mean log weekly wage for men born in any other quarter:
  mean(logww[q!=1&st==1]) #5.908638
```

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

```
  #difference:
  diff.logww1 <- mean(logww[q==1&st==1])-mean(logww[q!=1&st==1]) #-0.00659704
#state at or below 16 (st=0):
  #mean log weekly wage for men born in first quarter:
  mean(logww[q==1&st==0]) #5.887929
```

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

```
  #mean log weekly wage for men born in any other quarter:
  mean(logww[q!=1&st==0]) #5.900605
```

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

```
  #difference:
  diff.logww0 <- mean(logww[q==1&st==0])-mean(logww[q!=1&st==0]) #-0.01267616

#education:
#state above 16 (st=1):
  #mean yrs education for men born in first quarter:
  mean(edu[q==1&st==1]) #12.72104
```

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

```
  #mean yrs education for men born in any other quarter:
  mean(edu[q!=1&st==1]) #12.81313
```

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

```
  #difference:
  diff.edu1 <- mean(edu[q==1&st==1])-mean(edu[q!=1&st==1]) #-0.09208688
#state at or below 16 (st=0):
  #mean yrs education for men born in first quarter:
  mean(edu[q==1&st==0]) #12.67649
```

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

```
  #mean yrs education for men born in any other quarter:
  mean(edu[q!=1&st==0]) #12.79117
```

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

```
  #difference:
  diff.edu0 <- mean(edu[q==1&st==0])-mean(edu[q!=1&st==0]) #-0.114683

#Wald estimates:
  #state above 16 (st=1):
  diff.logww1/diff.edu1 #0.07163931
```

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

```
  #state at or below 16 (st=0):
  diff.logww0/diff.edu0 #0.1105322
```

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

```
#OLS for reference
mod1 <- lm(logww~edu,data=dat)
mod2 <- lm(logww~edu+states_above_16,data=dat)
msummary(list(mod1,mod2))
```

(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.

```
#While we can see from the last graph a similarly small, positive return in education for both kinds of
```

	(1)	(2)
(Intercept)	4.995	4.993
	(0.004)	(0.005)
edu	0.071	0.071
	(0.000)	(0.000)
states_above_16		0.008
		(0.003)
Num.Obs.	329 509	329 509
R2	0.117	0.117
R2 Adj.	0.117	0.117
AIC	638 703.1	638 696.2
BIC	638 735.3	638 739.0
Log.Lik.	−319 348.574	−319 344.086
RMSE	0.64	0.64