What's up with regression?

Prof. Jonathan Davis

Due by Friday, October 22nd at Noon

In the first two questions of this homework, we will once again analyaze data from the National Longitudinal Survey of Youth 1979 to measure the correlation between income and years of education.

The data file `nlsy79.csv` includes this information and some additional variables we may use later in the course. A description of the variables in the data is in the table below.

Name	Description
`CASEID`	Unique identifier
`earn2009`	Earnings in 2009
`hgc`	Years of education
`race`	Race and Ethnicity
`sex`	Gender
`bmonth`	Birth Month
`byear`	Birth Year
`afqt`	Armed Forces Qualifying Test Percentile
`region_1979`	Region
`faminc1978`	Family Income in 1978
`nsibs79`	Number of Siblings

1. The Shape of the Returns to Schooling

a. Regress earnings on years of education. How much do earnings increase on average for every additional year of schooling?

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("broom")
library("stargazer")
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
setwd('C:/Users/campb/OneDrive/econometrics/hw3')
nlsy79 <- read_csv('nlsy79.csv')
```

```
## Rows: 6110 Columns: 11
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): race, sex
## dbl (9): CASEID, earn2009, bmonth, byear, afqt, region_1979, faminc1978, hgc...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
fel1 <- lm(earn2009 ~ hgc, data = nlsy79)
fel1
```

```
##
## Call:
## lm(formula = earn2009 ~ hgc, data = nlsy79)
##
## Coefficients:
## (Intercept)          hgc
##      -82664         9949
```

```
tidy(fel1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  -82664.     4720.     -17.5 2.03e- 66
## 2 hgc            9949.      350.      28.5 2.08e-163
```

```
#earnings increase by9949 for every extra year of #school
```

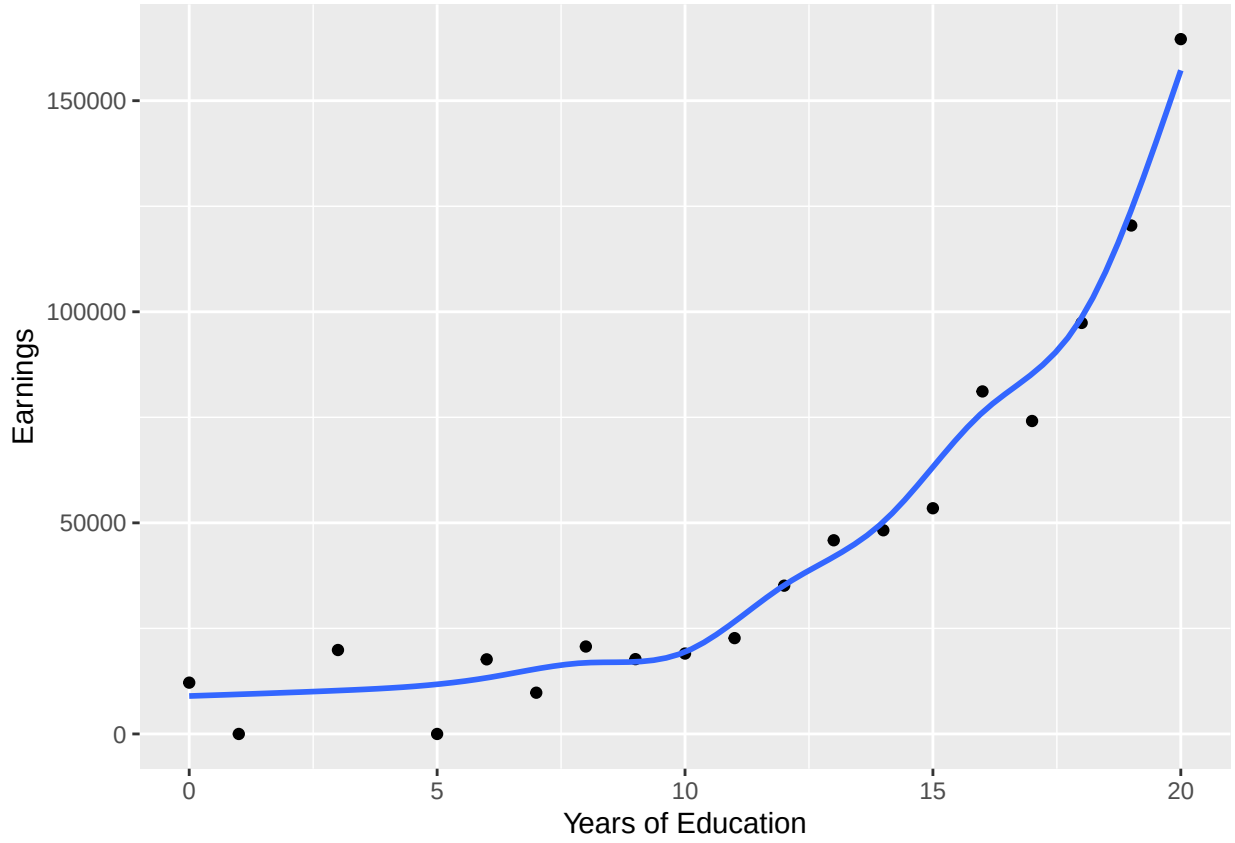
b. Use ggplot2 to plot the conditional expectation of earnings with respect to years of education? Do you think it is reasonable to assume earnings increase linearly with years of schooling?

```
ggplot(nlsy79, aes(x = hgc, y = earn2009)) +
  geom_point(stat = "summary", fun = "mean") +
  xlab("Years of Education") + ylab("Earnings") +
  geom_smooth(se = FALSE)
```

```
## Warning: Removed 1776 rows containing non-finite values (stat_summary).
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1776 rows containing non-finite values (stat_smooth).
```




```
#Assuming earnings increase with years of school #seems reasonable after 12 years of school
```

c. Generate a variable that equals years of education squared. Regress earnings on years of education and years of education squared. How much do earnings increase for someone who gets 10 instead of 9 years of schooling? What about someone who gets 17 instead of 16

```
nlsy79 <- mutate(nlsy79, hgcsq  = (hgc*hgc))

reg1 <- lm(earn2009 ~ hgc + hgcsq, data = nlsy79)
reg1
```

```
##
## Call:
## lm(formula = earn2009 ~ hgc + hgcsq, data = nlsy79)
##
## Coefficients:
## (Intercept)          hgc        hgcsq
##     47122.7      -9711.8        719.4
```

```
#for 10 yrs over 9 yrs there is a $3,950 diff
#for 17 yrs over 16 yrs there is a $14,016 diff

tidy(reg1)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   47123.   16206.       2.91 3.66e- 3
## 2 hgc           -9712.    2376.      -4.09 4.43e- 5
## 3 hgcsq           719.      86.0      8.37 7.99e-17
```

```
summary(reg1)
```

```
##
## Call:
## lm(formula = earn2009 ~ hgc + hgcsq, data = nlsy79)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -140661  -33628   -9090   15742  310794
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  47122.7    16206.1   2.908  0.00366 **
## hgc          -9711.8     2375.7  -4.088 4.43e-05 ***
## hgcsq          719.4       86.0   8.365  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56610 on 4331 degrees of freedom
##   (1776 observations deleted due to missingness)
## Multiple R-squared:  0.1709, Adjusted R-squared:  0.1705
## F-statistic: 446.2 on 2 and 4331 DF,  p-value: < 2.2e-16
```

```
stargazer(reg1, keep.stat = c("n", "rsq"), covariate.labels = "earnings", dep.var.labels = "earnings",
```

```
##
## ========================================
##                  Dependent variable:
##              ---------------------------
##                        earnings
## ----------------------------------------
## earnings            -9,711.773***
```

```
##                       (2,375.713)
##
## hgcsq                 719.434***
##                        (86.002)
##
## Constant             47,122.710***
##                      (16,206.080)
##
## ---------------------------------------
## Observations             4,334
## R2                       0.171
## =======================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

d. Code years of education as a factor. Now regress earnings on years of education. Comment on the results.

```
#hgcf the factor of hgc
nlsy79 <- nlsy79 %>% mutate(hgcf = factor(hgc))

felshy <- lm(earn2009 ~ hgcf, data = nlsy79)
tidy(felshy)
```

```
## # A tibble: 19 x 5
##    term        estimate std.error statistic  p.value
##    <chr>          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)   12150.    39907.    0.304  0.761
##  2 hgcf1        -12150.    69122.   -0.176  0.860
##  3 hgcf3          7732.    69122.    0.112  0.911
##  4 hgcf5        -12150.    56438.   -0.215  0.830
##  5 hgcf6          5523.    44119.    0.125  0.900
##  6 hgcf7         -2380.    41308.   -0.0576 0.954
##  7 hgcf8          8559.    40517.    0.211  0.833
##  8 hgcf9          5570.    40286.    0.138  0.890
##  9 hgcf10         6891.    40220.    0.171  0.864
## 10 hgcf11        10551.    40231.    0.262  0.793
## 11 hgcf12        22972.    39929.    0.575  0.565
## 12 hgcf13        33724.    40014.    0.843  0.399
## 13 hgcf14        36095.    40005.    0.902  0.367
## 14 hgcf15        41311.    40149.    1.03   0.304
## 15 hgcf16        68980.    39972.    1.73   0.0845
## 16 hgcf17        61980.    40191.    1.54   0.123
## 17 hgcf18        85211.    40166.    2.12   0.0339
## 18 hgcf19       108285.    40508.    2.67   0.00754
## 19 hgcf20       152428.    40578.    3.76   0.000175
```

```
stargazer(felshy, keep.stat = c("n", "rsq"), covariate.labels = "years of school", dep.var.labels = "ea
```

```
##
## ========================================
##                     Dependent variable:
##                 ---------------------------
```

```
##                           earnings
## ----------------------------------------
## years of school          -12,149.910
##                         (69,121.710)
##
## hgcf3                     7,731.760
##                         (69,121.710)
##
## hgcf5                    -12,149.910
##                         (56,437.640)
##
## hgcf6                     5,522.683
##                         (44,119.330)
##
## hgcf7                    -2,379.608
##                         (41,308.120)
##
## hgcf8                     8,558.512
##                         (40,516.750)
##
## hgcf9                     5,569.621
##                         (40,285.710)
##
## hgcf10                    6,891.023
##                         (40,220.440)
##
## hgcf11                   10,550.780
##                         (40,230.580)
##
## hgcf12                   22,972.380
##                         (39,928.540)
##
## hgcf13                   33,724.240
##                         (40,013.710)
##
## hgcf14                   36,095.210
##                         (40,005.130)
##
## hgcf15                   41,311.450
##                         (40,148.570)
##
## hgcf16                   68,980.450*
##                         (39,972.380)
##
## hgcf17                   61,980.310
##                         (40,191.480)
##
## hgcf18                  85,210.740**
##                         (40,165.740)
##
## hgcf19                 108,284.900***
##                         (40,507.580)
##
## hgcf20                 152,428.400***
```

```
##                         (40,578.200)
##
## Constant                 12,149.910
##                         (39,907.440)
##
## -----------------------------------------
## Observations               4,334
## R2                         0.179
## =========================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

```
#the effect on earnings based on years of school
#definitely increases as years increase.

#the 'slope' of 1 and 5 years of school is the #negative of the intercept.

#the std error starts out large and gets smaller
#near the middle and starts growing at the end
```

e. Do your answers to this question make you more or less worried about the linearity of linear regression?

The linearity of linear regression assumes the relationship between x and the mean of y is linear. The data seems to be linear and the standard error gets smaller in the middle years showing that it is a more representative sample of the overall population.

2. Indicators and Interactions

a. Regress the natural logarithm of earnings on an indicator variable for being male. How do the estimates relate to the average log earnings of men and women?

```
table(nlsy79$sex)
```

```
##
## FEMALE   MALE
##   3108   3002
```

```
#make a log earn variable
nlsy79 <- nlsy79 %>%
  mutate(learn = ifelse(earn2009>0, log(earn2009), NA))

#indicator variable for male and female
nlsy79 <- nlsy79 %>% mutate(male = sex == "MALE")

#regress log earn on male
bymale <- lm(learn ~ male, data = nlsy79)
bymale
```

```
##
## Call:
## lm(formula = learn ~ male, data = nlsy79)
##
```

```
## Coefficients:
## (Intercept)     maleTRUE
##     10.3266       0.5216
```

```
#table of results
stargazer(bymale, keep.stat = c("n", "rsq"), covariate.labels = "Male", dep.var.labels = "Log Earnings"
```

```
##
## Regression of log earnings on male
## ========================================
##                   Dependent variable:
##               ---------------------------
##                       Log Earnings
## ----------------------------------------
## Male                   0.522***
##                         (0.034)
##
## Constant               10.327***
##                         (0.024)
##
## ----------------------------------------
## Observations             3,559
## R2                       0.061
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

```
#average earnings of men
nlsy79 %>%
  filter(sex == "MALE") %>%
  drop_na(learn) %>%
  summarize(avgM = mean(learn))
```

```
## # A tibble: 1 x 1
##    avgM
##   <dbl>
## 1  10.8
```

```
#average earnings for female
nlsy79 %>%
  filter(sex == "FEMALE") %>%
  drop_na(learn) %>%
  summarize(avgF = mean(learn))
```

```
## # A tibble: 1 x 1
##    avgF
##   <dbl>
## 1  10.3
```

```
#The avg earnings of men is 10.848
#The avg earnings of women is 10.326

#the difference in these values is equal to the
```

```
#coefficient in the regression .522

#the increase in earn for being a male is .522
```

b. Regress log earnings on years of education and an indicator for being male. Next, regress log earnings on years of education and an indicator for being female. Compare the estimated returns to education from both specifications. What do you notice?

```
#regress logearn on years school ANd male
nlsy79 <- nlsy79 %>%
mutate(male = sex == "MALE")
reg2 <- lm(learn ~ hgc + male, data = nlsy79)

tidy(reg2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    8.37    0.0931       89.9 0
## 2 hgc            0.144   0.00666      21.7 9.35e-98
## 3 maleTRUE       0.534   0.0325       16.4 2.21e-58
```

```
#regress learn on hgc and female
nlsy79 <- nlsy79 %>%
  mutate(female = sex == "FEMALE")
reg3 <- lm(learn ~ hgc + female, data=nlsy79)

tidy(reg3)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    8.91    0.0926       96.1 0
## 2 hgc            0.144   0.00666      21.7 9.35e-98
## 3 femaleTRUE    -0.534   0.0325      -16.4 2.21e-58
```

```
#The estimated returns of education is .144 for each and the estimated
#returns of being male are .533 while female is the negative of that
# at -.533
```

c. Regress log earnings on years of education separately for the samples of men and women. Compare these estimates to what you found in part a of this question.

```
nlsy79m <- nlsy79 %>%
filter(male == 1)

regm <- lm(learn ~ hgc, data = nlsy79m)
tidy(regm)
```

```
## # A tibble: 2 x 5
```

```
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    8.68    0.123        70.8 0
## 2 hgc            0.161   0.00894      18.0 9.01e-67
```

```
#dataset female
nlsy79f <- nlsy79 %>%
  filter(male == 0)

regf <- lm(learn ~ hgc, data = nlsy79f)

tidy(regf)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    8.64    0.136        63.3 0
## 2 hgc            0.125   0.00993      12.6 8.03e-35
```

```
#Using a sample of only the males results in a return in another year of schooling of .161 which is muc
# For females this effect was seen even more as the returns to another
# year of schooling are .124.
# The results are smaller than a since we are looking at essential half
# of the population for each regression here.
```

d. Generate the interaction between years of education and indicators for being a man and woman. Regress log earnings on years of education, the interaction between years of education and the indicator for being a woman, and the indicator for being a woman. How do these estimates relate to what you found in part c?

```
#create interaction betwwen hcgmale and hgcfemale
nlsy79 <- nlsy79 %>%
  mutate(hgc_male = hgc * male, hgc_female = hgc * female)

#regress learn on hgc and the two interactions
schoolingBySex <- lm(learn ~ hgc + hgc_female + female, data = nlsy79)

tidy(schoolingBySex)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   8.68     0.125      69.6   0
## 2 hgc           0.161    0.00911    17.7   2.59e-67
## 3 hgc_female   -0.0364   0.0133     -2.73  6.45e- 3
## 4 femaleTRUE   -0.0425   0.183      -0.232 8.17e- 1
```

```
# These results show that the return of another year of school is
# .161, the return of schoolwhile being a woman is -.036, and the
# return of being a female is -.042.
# The return of another year of school minus the interaction
# of school and being a woman is the same as the return of schooling
# calculated in part c.
```

e. Now, also generate a variable equal to the interaction between years of education and an indicator for being a man. Regress log earnings on years of education interacted with being a man and a woman (both interactions, but not years of education alone) and an indicator for being a woman. How do these estimates relate to what you found in part c?

```
#variable of interaction for hgc and indicator for man

#trying to make the indicator
nlsy79 <- nlsy79 %>%
  mutate(hgc_male = hgc * male, hgc_female = hgc * female)


#regress learn on hgc(interacted with man), hgc(interacted with woman), & indicator for woman
reg5 <- lm(learn ~ hgc_male + hgc_female + female, data= nlsy79)
tidy(reg5)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   8.68     0.125      69.6   0
## 2 hgc_male      0.161    0.00911    17.7   2.59e-67
## 3 hgc_female    0.125    0.00976    12.8   1.14e-36
## 4 femaleTRUE   -0.0425   0.183      -0.232 8.17e- 1
```

```
#The results for the slope of male and female ran like this
# are the same as the slopes found for them in their own sample.
```

3. In a 2004 Nature article (one of the leading scientific publications), Tatem, Guerra, Atkinson, and Hay predict that women will have faster times than men in the Olympic 100 meter dash by the year 2156. This one page article is posted on Canvas.

a. How do the authors come to their conclusion?

They reached their conclusion by plotting the winning times of men and women finals over the past 100 years against the competition date.

b. What assumption do they make about the rate at which men's and women's times will improve? (Hint: Are they improving faster or slower over time?) Is this reasonable?

The womens times are improving more as they have a plus or minus of ,232 seconds while the men only have .144. So the womens time difference each year could be much lower or much higher while the mens is a bit more consistent.

c. Based on their model, about how fast will men and women run the 100 meter dash in the year 2600?

On their model womens time will decrease by 2.491 seconds in the 148 years it takes to go from 2008 to 2156. To go another 444 years to 2600, at the same rate, the model would predict women to run a time of .606. Mens expected time in 2600 is 3.202.

4. Blog post brainstorm

a. What is a topic you are interested in?

League of Legends

b. Google that topic and “data”. What do you find?

data for champion win rates, pro player salary and team win rates.