# Does a relationship exist?

Prof. Jonathan Davis

Due by Noon on Friday, October 15th

1. Explain whether each of the statements below is a fact or a correlation? Justify your answer. (BDM & Fowler Question)

a. Most top performing schools have small student bodies.

This is a fact because it only discusses the student body for top schools and does not describe the student body size for other non performing schools so we can not tell if the size of the student body actually impacts the performance level.

b. Married people are typically happier than unmarried people.

This is a correlation because it describes both married and unmarried people and their varying happiness levels.

c. Among professionals, taller basketball players tend to have lower free-throw percentages than shorter players.

This is a correlation because it describes taller and shorter players along with their varying free throw percents.

d. The locations in the U.S. with the highest cancer rates are typically small towns.

This is a fact because it describes cancer rates in small towns but does not mention larger towns.

e. Older houses are more likely to have lead paint than newer ones.

This is a correlation because it describes lead paint levels in older and newer houses so shows the variation between the two.

f. Most colds caught in Lane County are caught on cold days.

This is fact because it only describes colds caught on cold days and does not mention other kinds of days.

2. Comment on each of the graphs below. What does it tell us? What kind of graph is it? What is the x-axis? What is the y-axis? Is the graph misleading? Why? (Based on BW book)

a.

This graph is showing the rate of unemployment through the last 20 years, over the graph the times of U.S. recessions are shaded to show how the recessions impact unemployment. The graph is a line graph showing changes in unemployment over time. The x axis is the year and the y axis is the unemployment rate. The graph is not really misleading though the giant increase in unemployment in 2020 is the main viewing point and can be jarring due to the increase at the end but there is no data further than that so there is nothing they could've done.

b.

This graph is a line graph showing changes over time, the y axis is percent increase in cumulative pediatric covid cases and the x axis is the date. The graph is trying to show that pediatric covid cases have gone down over time but it is clearly misleading. Since the y axis is weekly increase in cumulative percent of cases, the amount of covid cases are still increasing even when the points reach under 10% at the end. It also shows cumulative totals so the percents will naturally be higher at the start when there is fewer initial cases. This means at the end when there are over 500,000 cases, a 10% increase will be more total new cases of covid than a 70% increase when there were only a few thousand to start with.

c.

This graph is a bar graph with amount of taxable income on the y axis and gross income level in 2008 on the x axis. This graph is showing that the middle class has a much larger total taxable income than the poorer or richer areas. The graph is misleading because the x axis is not constant. It starts out with each bar representing 5000 dollars and at the end the bar represents 5 million dollars. This results in the middle of the graph where they are trying to show the middle class has the most taxable income, the bar represents 100,000 dollars. This means that bar will represent many more people as it is a bigger range that they could fit in as compared to the 5000 dollar range at the start, so more people in that range will lead to a bigger bar.

d.

This graph is a horizontal bar graph with weekly work hours on the x axis and country on the y axis. It shows how many hours a week some countries work and also compares to the EU-28 average. The graph shows that countries near the bottom have significantly fewer hours worked per week, with bars less than half the size of the countries on the top. This is misleading because the graph does not start at 0 hours a week like it should being a bar graph. It starts at 36 hours and only shows up to 42 hours so even a country at one extreme compared to one at the other extreme would only have a difference of 6 hours while the graph tries to make it seem much larger than that.

e.

This graph is a line graph showing differences over time. It shows number of murders committed with firearms on the y axis and the year on the x axis. What the graph is trying to get across is after the implementation of the stand your ground law in 2005, the number of murders using firearms increased. The graph is misleading first because the 0 value of murders is at the top of the y axis which makes it a little harder to just glance at it since most graphs have the 0 value on the bottom. Also the number of murders at the very start of the graph are much larger than even after the implementation of stand your ground so I would be interested in what it would show if it went even earlier. By coloring the graph red they are trying to evoke feelings of violence to show that the law has increased murders drastically.

3. In class, we used 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 |

```
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(ggplot2)
library(dplyr)
library(readr)
library(gapminder)

setwd('C:/Users/campb/OneDrive/econometrics/HW2')
getwd()
```

```
## [1] "C:/Users/campb/OneDrive/econometrics/HW2"
```

```
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.
```

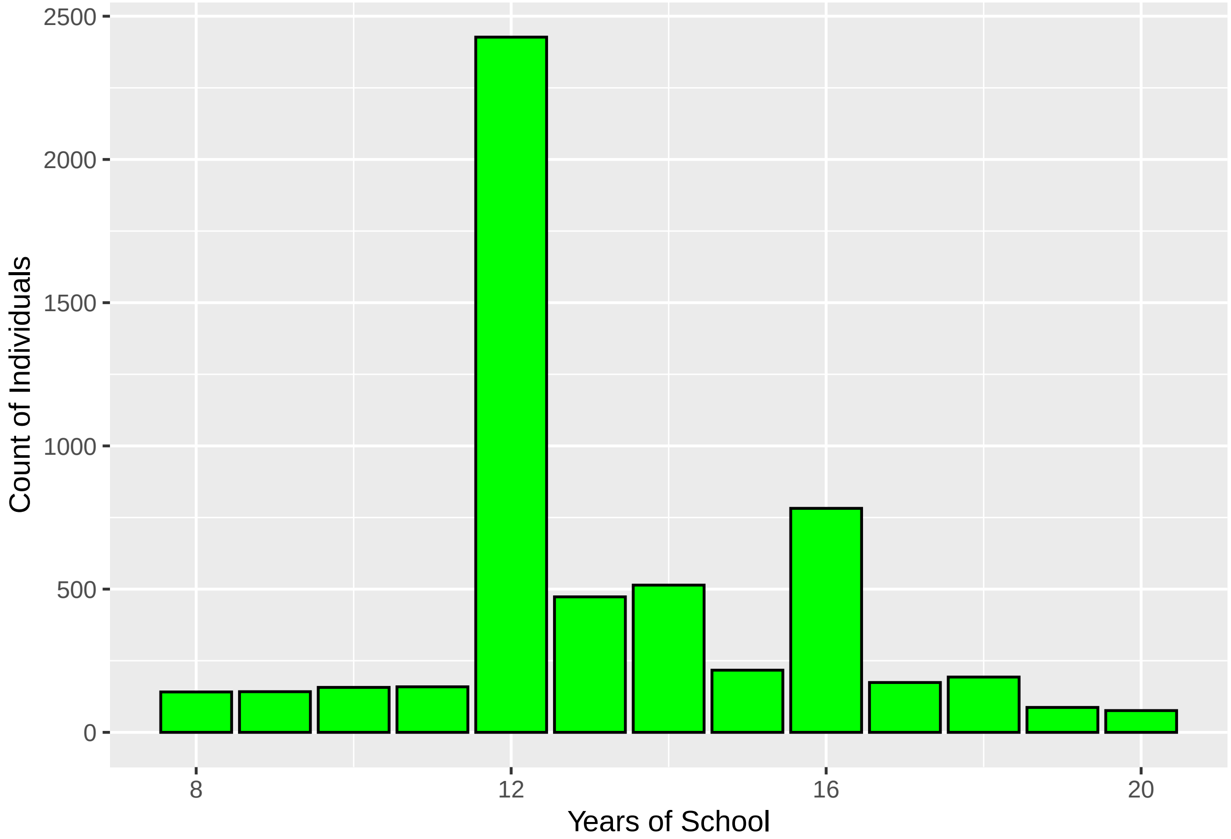
a. Re-code everyone who has 8 or fewer years of education as 8 in the data. We will interpret this category as “8 years or less”.

```
 nlsy79 <- nlsy79 %>%
mutate(hgc = ifelse(hgc<8, 8, hgc))
```

b. Use ggplot to make a bar chart showing the share of the sample with each level of education. What percentage of the population has 17 years of education or more?

```
 ggplot(nlsy79, aes(x=hgc)) +
geom_bar(color = "black", stat="count", fill = "green") +
xlab("Years of School") + ylab("Count of Individuals")
```

```
## Warning: Removed 568 rows containing non-finite values (stat_count).
```



#There appears to be around 9% of the pop #with 17 years or more of education.

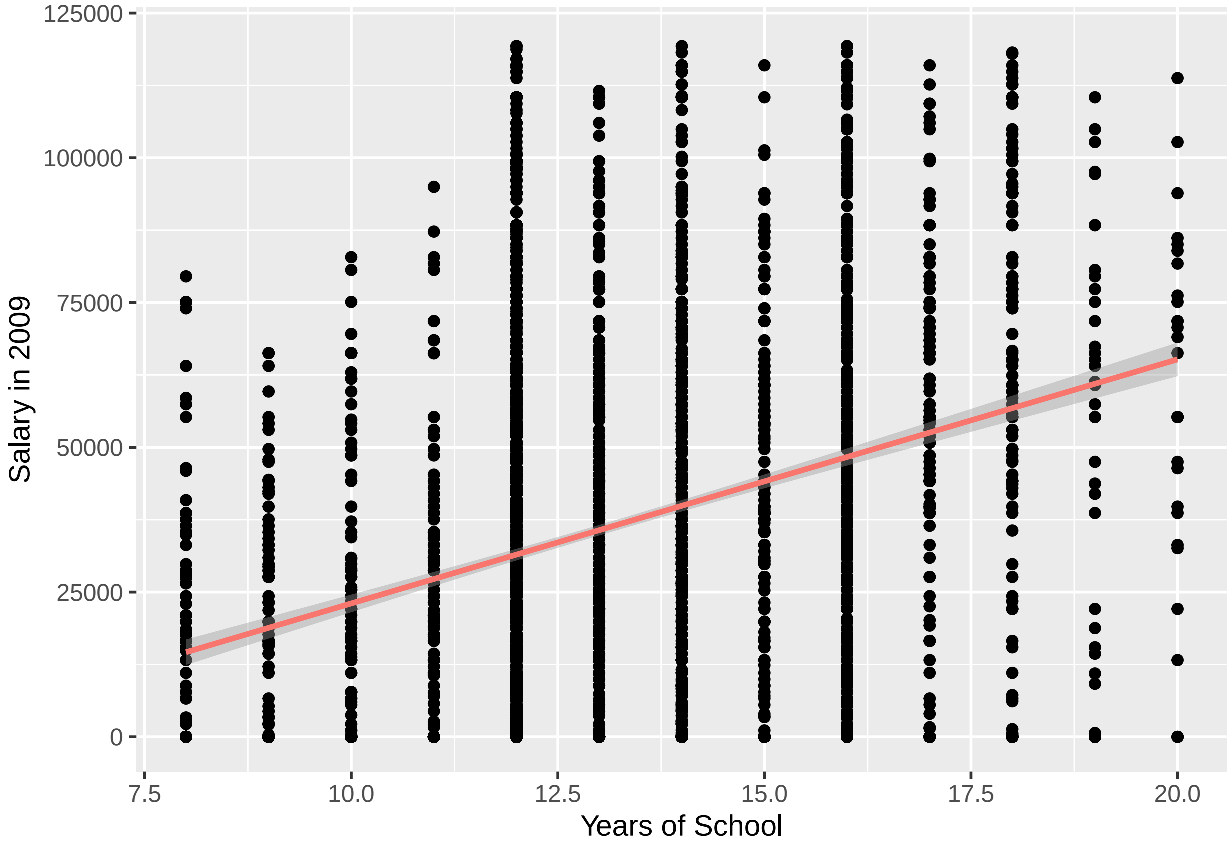
c. Use ggplot to make a scatterplot showing the relationship between earnings and education (recoded as in part a), include the line of best fit. What is the relationship between earnings and educcation?

```
 ggplot(nlsy79, aes(x = hgc, y = earn2009)) +
geom_point() +
geom_smooth(method = 'lm', aes(color = "red")) +
theme(legend.position = "None") +
ylim(0,120000) +
xlab("Years of School") + ylab("Salary in 2009")
```

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

```
## Warning: Removed 2091 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2091 rows containing missing values (geom_point).
```



#The relationship between earnings and #education appears to associated positively

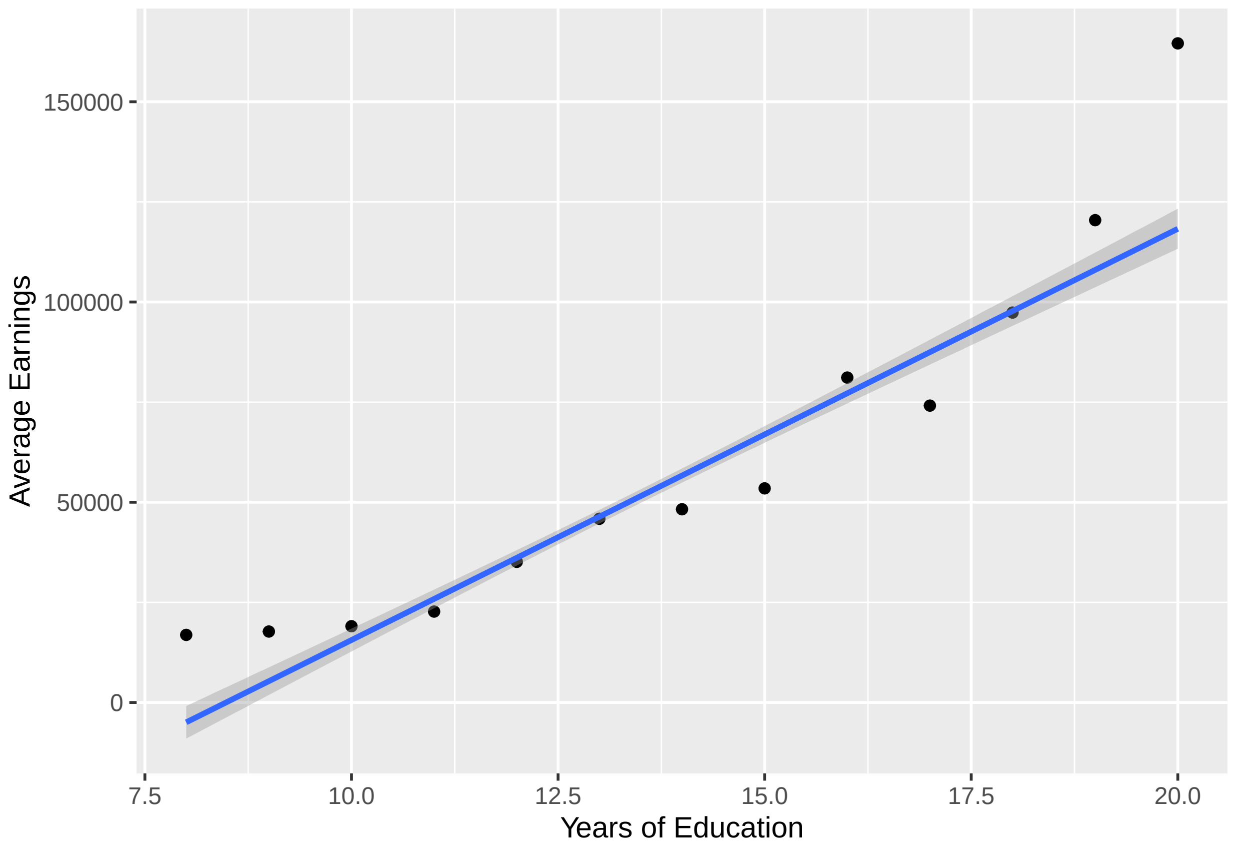
d. Use ggplot to make a figure showing the average earnings by years of education with the line of best fit. Comment on how well the line of best fit matches the average earnings data.

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

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

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

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



#The line of best fit fits the figure fairly #well due to the strong positive association