Lab 5

Taylor Blair

Math 141, Week 5

Due: Before your Week 6 lab meeting

Goals of this lab

1. Conduct exploratory data analysis.
2. Build linear regression models.
3. Interpret the estimated coefficients of a linear regression model.
4. Explore the impacts of potentially influential points.

Problems

Problem 1

In Fall of 2014, the following article was on the front page of nytimes.com: Top Colleges That Enroll Rich, Middle Class and Poor. (A copy of the article can also be found in the Articles folder in our shared class folder.) Looking at all colleges with a four-year-graduation rate of at least 75%, the authors created a measure called the "College Access Index" which tries to determine the accessibility of college for low and middle-income students. Another variable they collected was the endowment per student. Let's build a model for the college access index using endowment.

I pulled the data from the website and put it into a csv file ("collegeaccess.csv"“) for us to analyze.

```r
# Load data and omit the four schools with missing data for our two key variables
collegeaccess <-  read_csv("/home/courses/math141f20/Data/collegeaccess.csv") %>%
  drop_na(college.access.index, endowment.per.student)
```

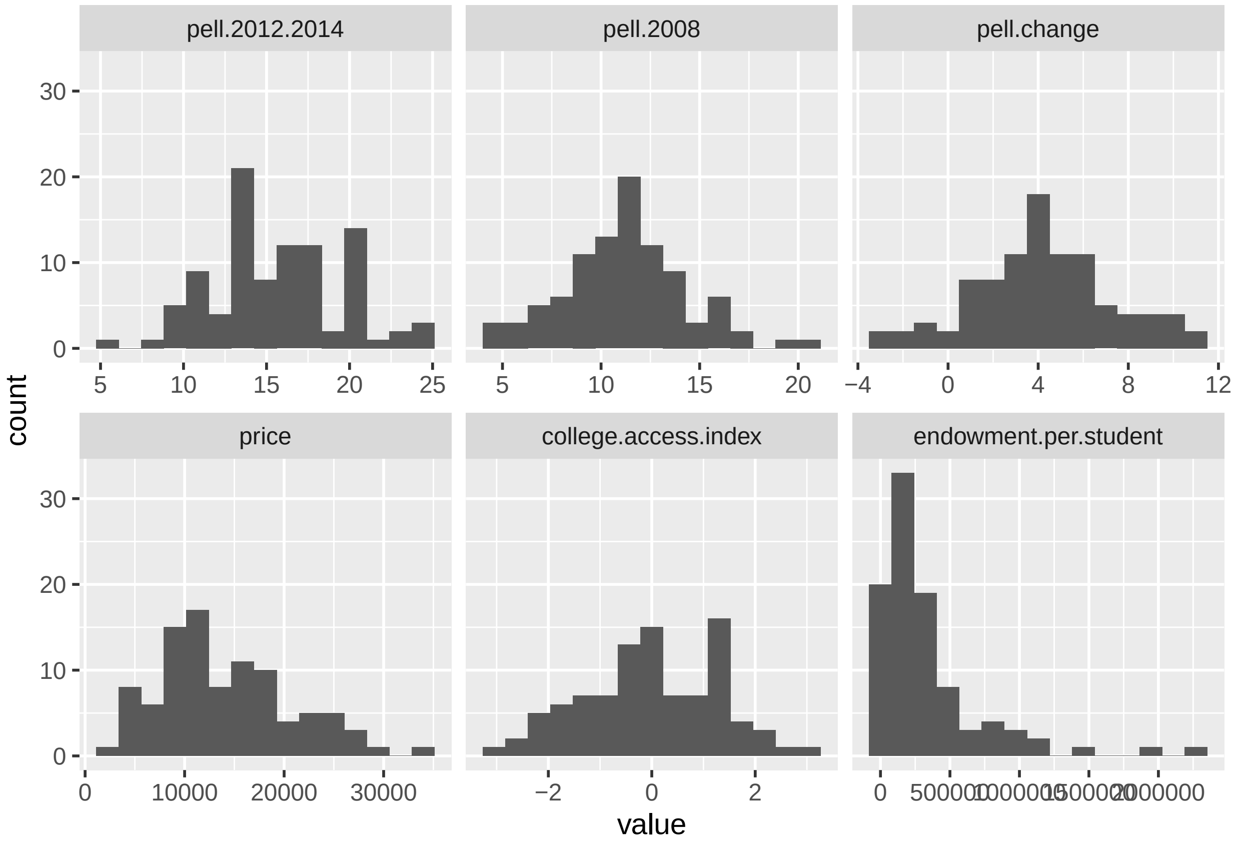
a. Identify the response variable and the explanatory variable. Also provide their code names.

- **Response Variable**
  - Change in Pell Grant recievers
  - `pell.change`
- **Explanatory Variable**
  - Endowment Per Student
    - More endowment per student typically means more resources
    - `endowment.per.student`

---

b. Explore the data by producing summary statistics and single variable plots (i.e. no scatterplots) of the college access index variable and the endowment variable.

```r
witch <- melt(collegeaccess[2:7])
ggplot(witch,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram(bins=15)
```



```r
as_tibble(cor(as.matrix(collegeaccess[2:7]))) %>%
  add_column(compared_to = colnames(collegeaccess[2:7]), .before ="pell.2012.2014") %>%
  rename(EPS = endowment.per.student, CAI = college.access.index)
```

```
## # A tibble: 6 x 7
##   compared_to        pell.2012.2014 pell.2008 pell.change   price    CAI     EPS
##   <chr>                       <dbl>     <dbl>       <dbl>   <dbl>  <dbl>   <dbl>
## 1 pell.2012.2014             1         0.653       0.613   0.140   0.652 -0.0315
## 2 pell.2008                  0.653     1          -0.197   0.0190  0.487  0.0533
## 3 pell.change                0.613    -0.197       1       0.166   0.333 -0.0971
## 4 price                      0.140     0.0190      0.166   1      -0.658 -0.531
## 5 college.access.in~         0.652     0.487       0.333  -0.658   1      0.375
## 6 endowment.per.stu~        -0.0315    0.0533     -0.0971 -0.531   0.375  1
```

c. Draw some conclusions from the statistics and plots produced in b.

Note: If you want to include the dollar sign in your narrative, use the following syntax: $

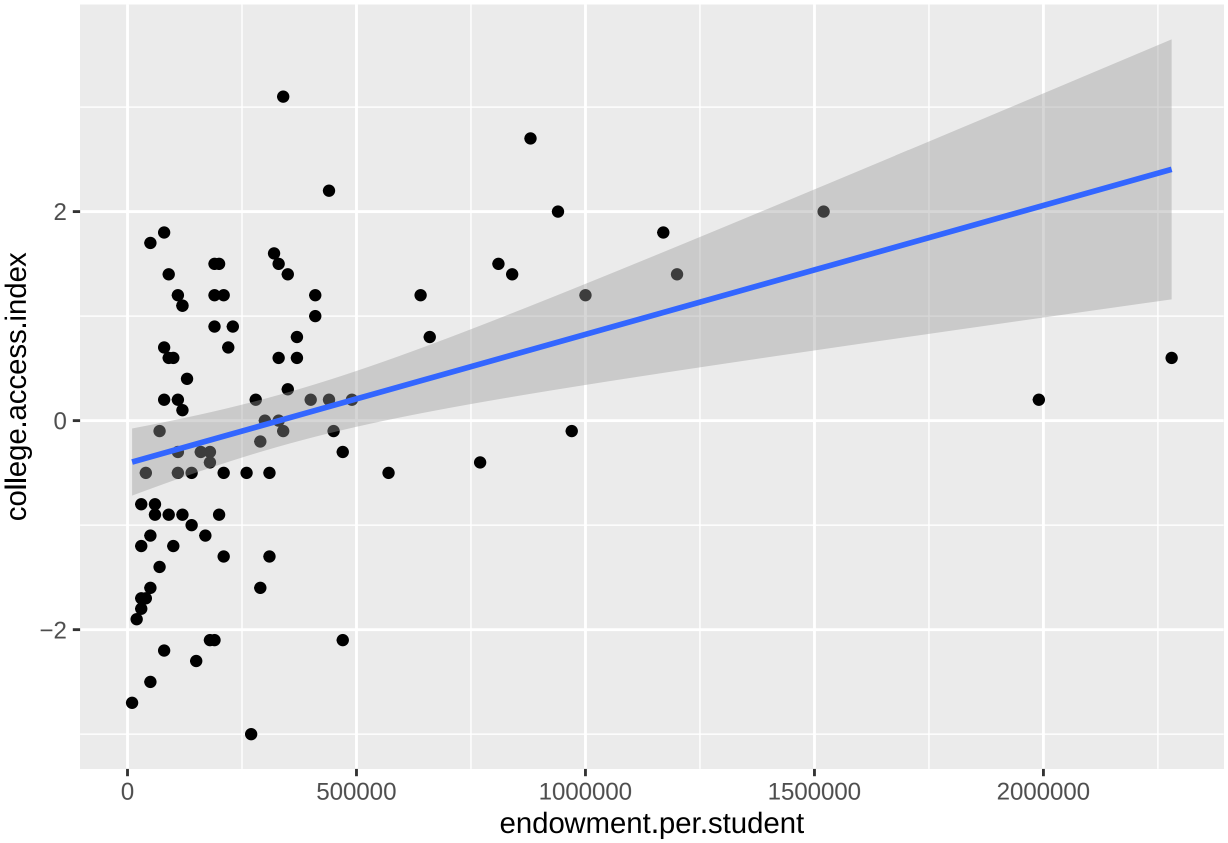
- There are weak linear coorelations for the response variable. The strongest linear cor is `pell.2012.2014`
- Data is evenely distirbuted for most variables with a few notable exceptions
  - `endowment.per,student` skews right and has a smaller spread
  - `pell.2012.2014` is somewhat chaotic while also being a partially normal distribution

- `college.access.index` is a flat normal distribution

---

d. Construct a scatterplot of the endowment variable and the college access index variable. Make sure to map the response variable to the `y` location and the explanatory variable to the `x` location.

```
ggplot(collegeaccess, aes(endowment.per.student, college.access.index)) +
  geom_point() + geom_smooth(method='lm', formula= y~x)
```



e. Compute the correlation coefficient for the endowment variable and the college access index variable.

```
cor(collegeaccess$endowment.per.student, collegeaccess$college.access.index)
```

```
## [1] 0.3752265
```

f. Based on your scatterplot and correlation coefficient, is a linear curve a good approximation for the relationship between these variables? Justify your answer.

No, the fit is rather weak as the majority of the points are within 0 to $500,000 endowment per student, with a wide spread. The points beyond half a million have a weak fit as a result

---

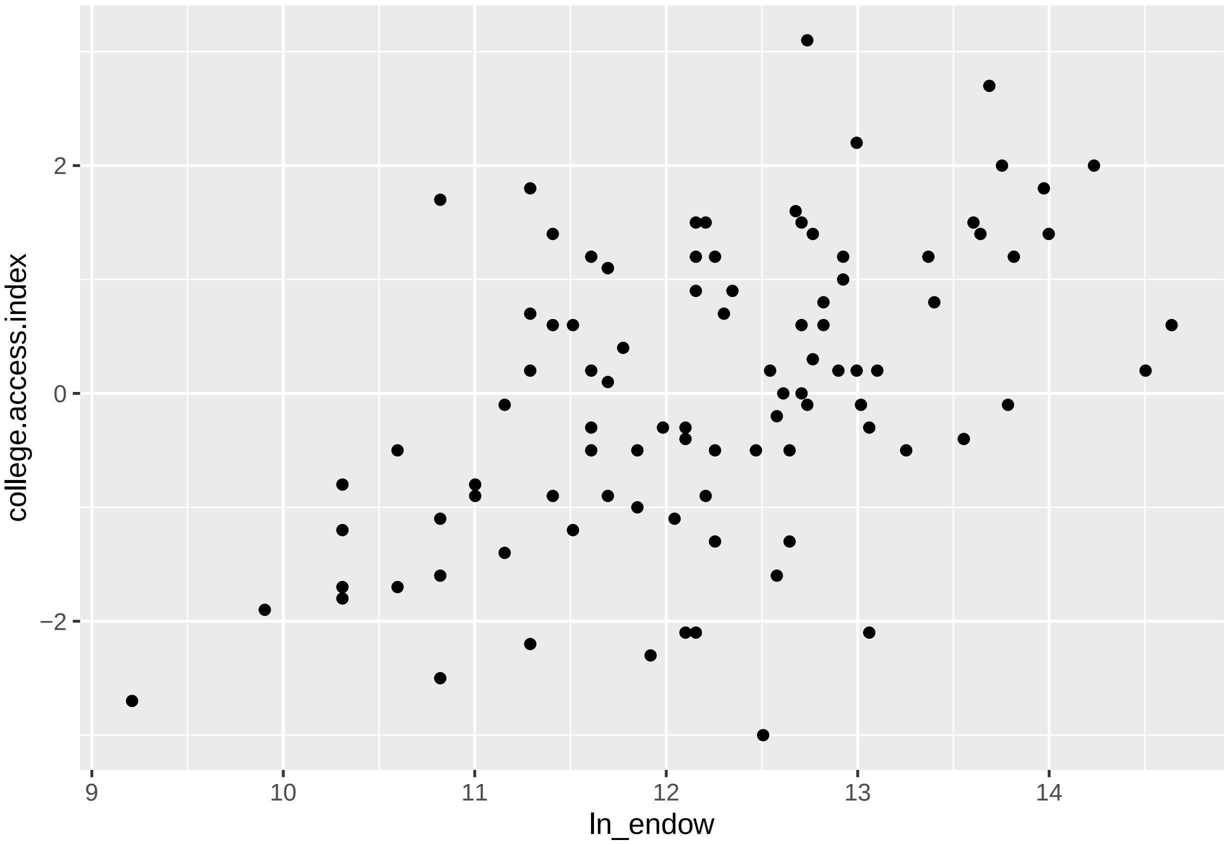
g. Add a column to the dataset that contains the log of the endowment variable. (Note: I want you to use natural log. The R function is `log()`.)

```
collegeaccess$ln_endow = log(collegeaccess$endowment.per.student)
collegeaccess
```

```
## # A tibble: 95 x 8
##    College pell.2012.2014 pell.2008 pell.change price college.access.~
##    <chr>            <dbl>     <dbl>       <dbl> <dbl>            <dbl>
##  1 Vassar              23        12          11  5600              3.1
##  2 Grinne~             24        14          10 10400              2.7
##  3 Smith               23        16           7 11600              2.2
##  4 Amherst             20        16           4  8400              2
##  5 Harvard             17        13           4  3000              2
##  6 Pomona              18        12           6  5200              1.8
##  7 St. Ma~             24        14          10 15900              1.8
##  8 Susque~             25        17           8 18000              1.7
##  9 Columb~             16        12           4  3500              1.6
## 10 Rice                18        15           3  8100              1.5
## # ... with 85 more rows, and 2 more variables: endowment.per.student <dbl>,
## #   ln_endow <dbl>
```

h. Now recreate the scatterplot and correlation coefficient but this time with the logged endowment variable.

```r
ggplot(collegeaccess, aes(ln_endow, college.access.index)) +
  geom_point()
```



```r
cor(collegeaccess$ln_endow, collegeaccess$college.access.index)
```

```
## [1] 0.4983396
```

i. Based on your updated scatterplot and correlation coefficient, is a linear curve a good approximation for the relationship between the access index and logged endowment? Justify your answer.

- It is a better fit than before. ~0.5 is a moderative positive strength slope.

---

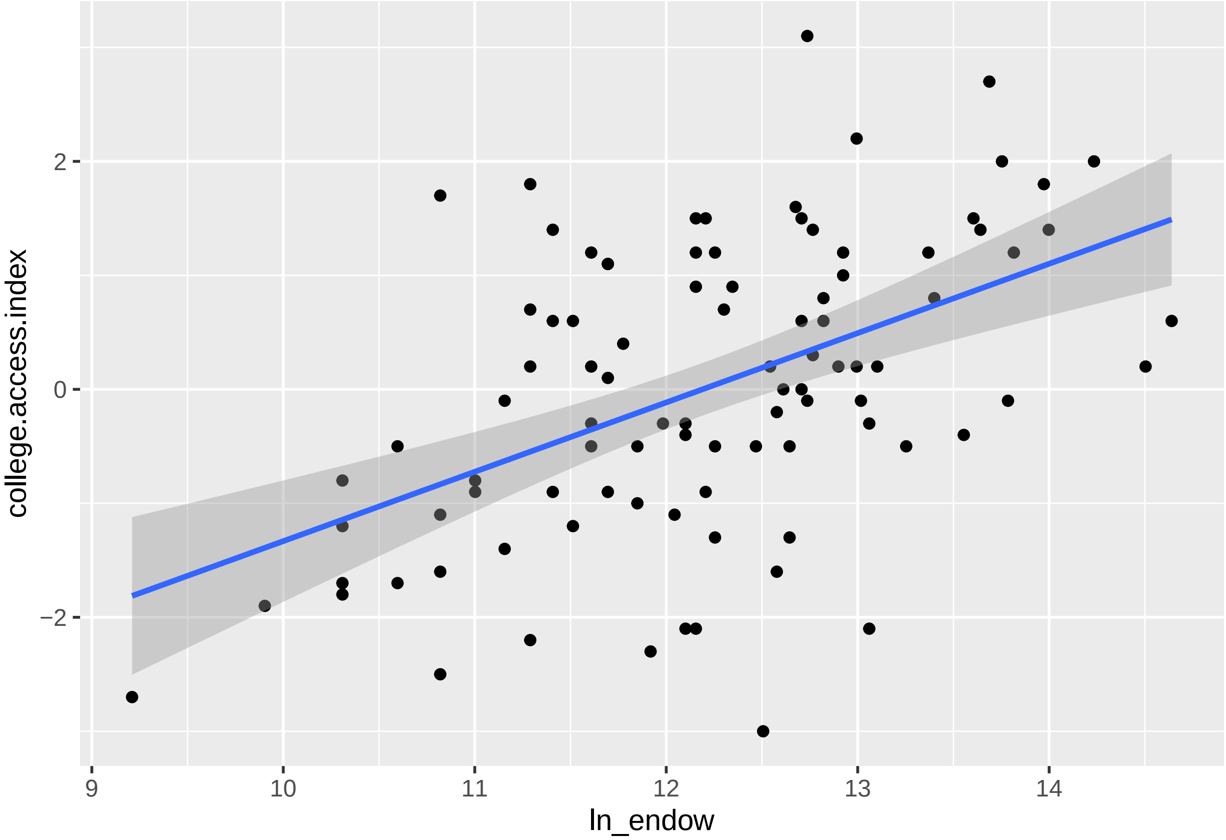
j. Build a simple linear regression model for the college access index using the logged endowment and print out the coefficients table.

```
lm(collegeaccess$college.access.index~collegeaccess$ln_endow)
```

```
##
## Call:
## lm(formula = collegeaccess$college.access.index ~ collegeaccess$ln_endow)
##
## Coefficients:
##            (Intercept)  collegeaccess$ln_endow
##                -7.4160                  0.6084
```

k. Plot the least squares line on a scatterplot of the data.

```
ggplot(collegeaccess, aes(ln_endow, college.access.index)) +
  geom_point() + geom_smooth(method='lm', formula= y~x)
```



l. Unfortunately, we can't explore how Reed College fits into this story since it is not in this dataset. Why is Reed not in the dataset? (Hint: Go back and read the problem's prompt.)

- Our 4 year graduation rate is too low (not at 75% at time of reporting)

---

m. Since I was previously teaching at Swarthmore College, I want you to predict their college access index based on your model. Compare your prediction to the true value. What does that imply about the access to college for low to middle income students at Swarthmore?

```
print("PREDICTED VALUE")
```

```
## [1] "PREDICTED VALUE"
```

```
print(-7.4160+(0.6084*collegeaccess["ln_endow"][collegeaccess["College"]=="Swarthmore"]))
```

```
## [1] 0.9708252
```

```
print("ACTUAL VALUE")
```

```
## [1] "ACTUAL VALUE"
```

```
print(collegeaccess["college.access.index"][collegeaccess["College"]=="Swarthmore"])
```

```
## [1] -0.1
```

```
print("DIFFERENCE")
```

```
## [1] "DIFFERENCE"
```

```
print(0.9708252-(0.1))
```

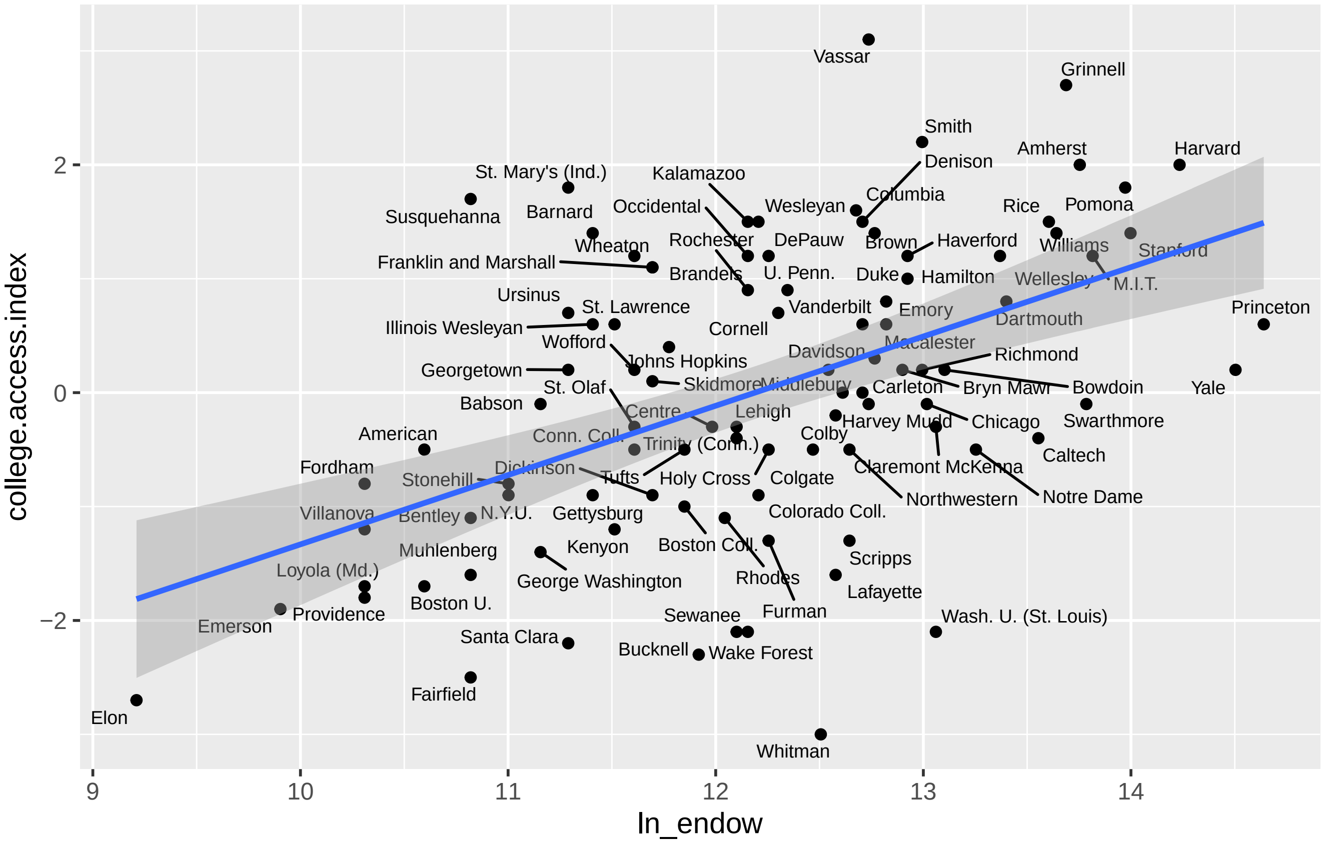
```
## [1] 0.8708252
```

- Swarthmore is less acssesible to lower income students than is predicted.
- It's predicted value would put it in the top 50 percentile as opposed to the bottom

---

n. Produce a scatterplot of log endowment and college access where the colleges are labeled. (Consider setting the `size` of the labels within `geom_text_repel()`.)

```
ggplot(collegeaccess, aes(ln_endow, college.access.index, label=College))+
  geom_point() + geom_text_repel(size=2.5) +
  geom_smooth(method='lm', formula= y~x)
```



o. Discuss how Vassar and Whitman differ from the general trend of the data.

- Vassar and Whitman have the maximum `college.access.index` score. They are both around 12.5 on `ln_endow` yet provide drastically different acsessibilties to students.

---

p. Fit the model with and without Vassar. Then fit the model with and without Whitman. Discuss changes to the slope and intercept. Make sure to address why the coefficients moved in the direction they did.

```
#Hint: Create a dataset (think dplyr) that doesn't contain Vassar
no_vass <- collegeaccess[-c(collegeaccess["College"]=="Vassar"), ]


# and a dataset that doesn't contain Whitman
no_whit <- collegeaccess[-c(collegeaccess["College"]=="Whitman"), ]


# VASSAR STARTS HERE
no_vass_fit <- lm(no_vass$college.access.index~no_vass$ln_endow)
print(no_vass_fit)
```

```
##
## Call:
## lm(formula = no_vass$college.access.index ~ no_vass$ln_endow)
##
## Coefficients:
```

```
##      (Intercept)  no_vass$ln_endow
##          -7.2740            0.5943
```

```r
print(cor(no_vass$college.access.index, no_vass$ln_endow))
```

```
## [1] 0.5016373
```

```r
#WHITMAN STARTS HERE
no_whit_fit <- lm(no_whit$college.access.index~no_whit$ln_endow)
print(no_whit_fit)# I cannot explain why they are the same
```

```
##
## Call:
## lm(formula = no_whit$college.access.index ~ no_whit$ln_endow)
##
## Coefficients:
##      (Intercept)  no_whit$ln_endow
##          -7.2740            0.5943
```

```r
cor(no_whit$college.access.index, no_whit$ln_endow)
```

```
## [1] 0.5016373
```

- `collegeaccess`
  - $y \approx 0.6084x - 7.4160$
  - 0.4983396 R
- `no_whit`
  - $y \approx 0.5943x - 7.2740$
  - 0.5016373 R
- `no_vass`
  - $y \approx 0.5943x - 7.2740$
  - 0.5016373 R
  - No different from whitman fit, which seems odd.

---

q. Do Vassar and Whitman appear to be influential points? Justify your answer.

- While they both have influence over the Cor coeffl, their impact is less than 0.01. So they are not impactful enough to be considered influential.

---

r. Read the nytimes.com article (which can be found in the Articles folder in our shared class folder). Do you think the author's college access index is a good measure of accessibility of college for low to middle income students? Justify your answer.

- **TL;DR**: Yes and no
- *Yes*
  - The statistic creates a snapshot of economic diversity that allows for an individual to comprehend a given colleges student economic backgrounds are.
- *No*
  - The stat leaves out several other variables, race, gender, and other minorty populations that a college might be recruiting for instead
  - Does not account for scholarships outside of the pell grant

  - Does not account for school region (North east schools seem to dominate the top right)
  - Leaves otu schools with lower 4 year graduation rates

Problem 2

On Wednesday we collected data to see if our class exhibits an “anchor effect”. I randomly gave half of you a survey with $X = 10$ and the other half a survey with $X = 65$. We want to answer the following: Does the X value given serve as an anchor to how you answered question two? In other words, does the previously supplied number serve as a starting point when we estimate the percentage of UN nations that are African?

```
# Load the data (also includes last two fall's Math 141 classes)
anchor <-  read_csv("/home/courses/math141f20/Data/anchor.csv")
```

a. What are the cases for our study?

- Individiuals given the survey in Math 141 F

---

b. Specify the response variable and the explanatory variable. Note: We will treat the explanatory variable as a categorical variable here instead of numeric since we have only two groups.

- **Explanatory**
  - Given 10 or 65
- **Response***
  - estimated prop in UN

---

c. Is this study an experiment or observational study? Justify your answer.

- Expirement: Individuals are given a treatment.

---

d. Was blindness used in the study? Justify your answer.

- Yes, information is withheld from the individuals that answered the survey

---

e. Compute useful summary statistics of the response variable by the groups of the explanatory variable.

```
anchor["x"][anchor["x"]=="sixty-five"]  <- "1"
anchor["x"][anchor["x"]=="ten"]  <- "0"

print(mean(anchor$Value))
```

```
## [1] 26.72898
```

```
print(mean(anchor["Value"][anchor["x"]=="1"]))
```

```
## [1] 30.95402
```

```
print(sd(anchor["Value"][anchor["x"]=="1"]))
```

```
## [1] 16.83152
```

```
print(mean(anchor["Value"][anchor["x"]=="0"]))
```

```
## [1] 22.59888
```

```
print(sd(anchor["Value"][anchor["x"]=="0"]))
```

```
## [1] 14.7201
```

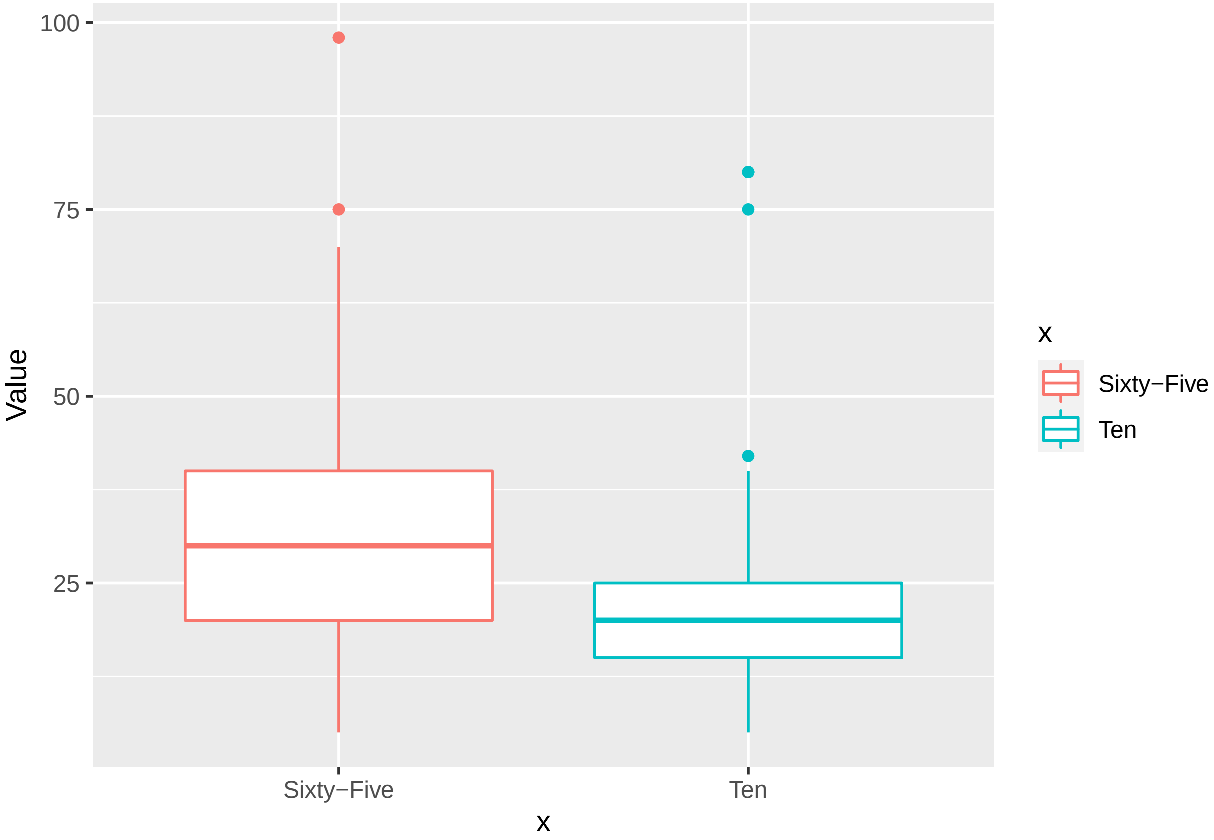
```
print(cor(as.numeric(anchor$x), anchor$Value))
```

```
## [1] 0.2569875
```

f. Construct a graphic that showcases the relationship between the two variables.

```
anchor["x"][anchor["x"]=="1"]  <- "Sixty-Five"
anchor["x"][anchor["x"]=="0"]  <- "Ten"

ggplot(anchor, mapping = aes(x, Value, color=x)) +
  geom_boxplot()
```



g. Draw some conclusions from your summary statistics and graphic.

- **Particpants given 65 treatment**
  - Larger standard deviation/IQR
  - Greater mean

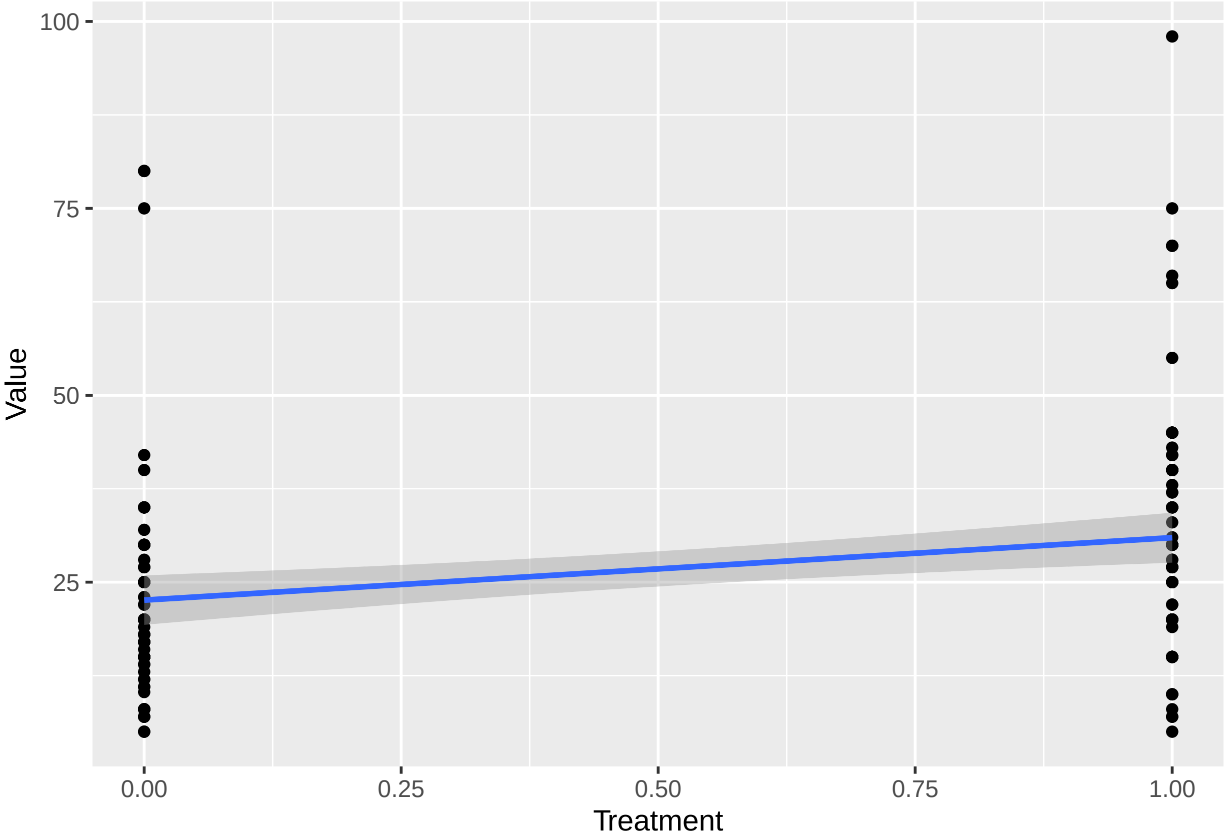
    - More outliers
- **Particpants given 10 treatment**
    - Tighter IQR/STD
    - Smaller mean

---

h. Fit the linear model.

```
anchor["x"][anchor["x"]=="Sixty-Five"]  <- "1"
anchor["x"][anchor["x"]=="Ten"]  <- "0"

ggplot(data = anchor, mapping = aes(as.numeric(x), Value)) +
  geom_point()+geom_smooth(method='lm', formula= y~x) +
  labs(x="Treatment")
```



```
lm(anchor$Value~as.numeric(anchor$x))
```

```
##
## Call:
## lm(formula = anchor$Value ~ as.numeric(anchor$x))
##
## Coefficients:
##          (Intercept)  as.numeric(anchor$x)
##               22.599                 8.355
```

i. Interpret the estimated coefficients ($\hat{\beta}_o$ and $\hat{\beta}_1$).

```
lm(anchor$Value~as.numeric(anchor$x))
```

```
##
```

```
## Call:
## lm(formula = anchor$Value ~ as.numeric(anchor$x))
##
## Coefficients:
##          (Intercept)  as.numeric(anchor$x)
##               22.599                 8.355
```

- $\hat{\beta}_o = 22.599$
  - Y-int is at 22.599. Because there are only two x values being fitted, this can be assumed to be the center of participants given the lower number
- $\hat{\beta}_1 = 8.355$
  - As group two is set to one, this is assumed to be the difference in average between the two groups.

---

j. Does there seem to be an anchoring effect? Does there appear to be a causal link between the anchor number and how the second question was answered? Justify your answer.

- Yes, the participants given the value `sixty-five` gave on average a higher value. and those that were given the number 10 had a far smaller standard deviation.

---

k. Who would you generalize these results to? Justify your answer.

- This study has been repeated numerous times, so although we only studied Reed stats students–*who had no idea what the actual value of African countries in the UN were*– it is safe to make a generalization and apply it to a larger population

---