**DS 710**

**Homework 9**

**R assignment**

1. In this problem, you will do further cleaning and analysis of the data from the 1995 US News and World Report on colleges and universities in the US.
2. In the Python portion of homework 9, you created a modified version of the data set usnews.csv.  Read the modified data into R and attach it.  Check the first few values of each vector to ensure that they were read accurately.

#1a)

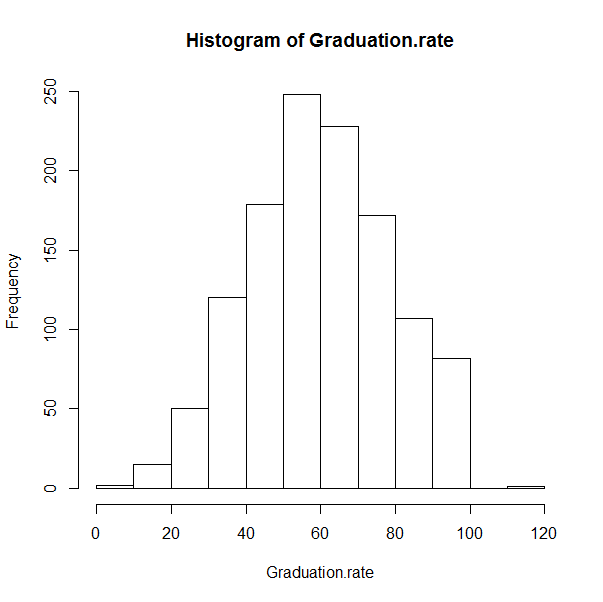
#read in the csv

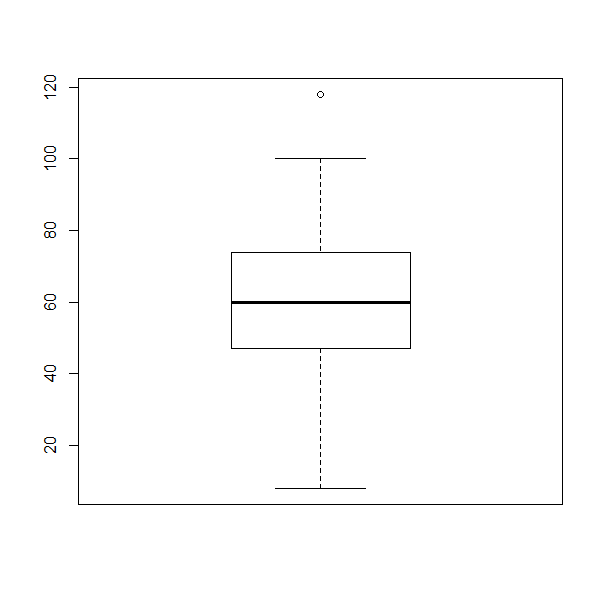
news\_CSV **=** read.csv**(**"C:/Users/pedbv9699/Documents/GitHub/ds710assignment9/new\_usnews.csv"**)**

#attach the csv in R

attach**(**new\_usnews**)**

1. Examine the summary, histogram, and boxplot of Graduation.rate.  Identify any unrealistic values and set them to missing.  Write a sentence describing what you did, naming the colleges or universities affected.  (For example, “Listed ages less than zero (ABC University, XYZ College) were converted to missing data.”)





hist**(**Graduation.rate**)**

boxplot**(**Graduation.rate**)**

**>** summary**(**Graduation.rate**)**

Min. 1st Qu. Median Mean 3rd Qu. Max. **NA**'s

8.00 47.00 60.00 60.41 74.00 118.00 98

#obviously this is something wrong with having a graduation rate as 118%

College.Name[which(Graduation.rate > 100)]

[1] Cazenovia College

1274 Levels: Abilene Christian University Adams State College Adelphi University Adrian College Agnes Scott College ... Youngstown State University

#get rid of the > 100 plus graduation rate

Graduation.Name**[**which**(**Graduation.rate **>** 100**)]** **=NA**

**>** summary**(**Graduation.rate**)**

Min. 1st Qu. Median Mean 3rd Qu. Max. **NA**'s

8.00 47.00 60.00 60.36 74.00 100.00 99

#again look at the percent of Professors or faculty with PhDs, you can't have 105%

summary**(**Pct.of.faculty.with.PhDs**)**

**>** summary**(**Pct.of.faculty.with.PhDs**)**

Min. 1st Qu. Median Mean 3rd Qu. Max. **NA**'s

8.00 57.00 71.00 68.65 82.00 105.00 32

# everything above 100 percent is changed to NA(non applicable) because it makes no sense

Pct.of.faculty.with.PhDs[which(Pct.of.faculty.with.PhDs > 100)] = 100.

> summary(Pct.of.faculty.with.PhDs)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

8.00 57.00 71.00 68.65 82.00 105.00 32

**>** Pct.of.faculty.with.PhDs**[**which**(**Pct.of.faculty.with.PhDs **>** 100**)]** **=** 100

**>** summary**(**Pct.of.faculty.with.PhDs**)**

Min. 1st Qu. Median Mean 3rd Qu. Max. **NA**'s

8.00 57.00 71.00 68.64 82.00 100.00 32

1. Find the mean percentage of alumni who donate, for private and public schools.

#percent who are public universities that are donated to

> public\_Donations **=** mean**(**Pct.alumni.who.donate**[**which**(**Private.or.Public **==** 'Public'**)]**, na.rm **=** T**)**

> public\_Donations

**[**1**]** 13.44944

#percent who are private universities that are donated to

**>** private\_Donations **=** mean**(**Pct.alumni.who.donate**[**which**(**Private.or.Public **==** 'Private'**)]**, na.rm **=** T**)**

**>** private\_Donations

**[**1**]** 24.58287

1. Test whether there is evidence that a higher percentage of alumni from private schools donate to their schools, compared to alumni from public schools.  State your conclusion in context.

#H0: mu[Public\_Donor] = mu[Private\_Donor]

#H1: mu[public\_donor != muprivate\_Donor]

**>** t.test**(**Pct.alumni.who.donate **~** Private.or.Public, alternative **=** "two.sided"**)**

Welch Two Sample t**-**test

data**:** Pct.alumni.who.donate by Private.or.Public

t **=** 17.3168, df **=** 1018.898, p**-**value **<** 2.2e**-**16

alternative hypothesis**:** true difference **in** means is not equal to 0

95 percent confidence interval**:**

9.871826 12.395044

sample estimates**:**

mean **in** group Private mean **in** group Public

24.58287 13.44944

P- value as low as it is, a significance level of .05 wouldn’t stand a chance. We can reject the null hypothese and see that there is enough evidence to claim there is infact difference in alumni donations in public vs private schools

1. Use write.csv() to save your updated data set.  Consult the R documentation to set the arguments for the write.csv function.  Your output file should not have row names or row numbers, and it should not have quotation marks around the entries
2. **>** write.csv**(**news\_CSV, file **=** 'C:/Users/pedbv9699/Documents/GitHub/ds710assignment9/US\_news.csv', quote **=** F, row.names **=** F**)**

Submit a .doc, .docx, or .pdf file to GitHub, containing your R code, R output, and written interpretations and explanations. (You may include your responses for problems 1 and 2 in the same file.)

2.  The data set cps.csv contains data from the 1985 Current Population Survey.

Dataset:  “Wages from the Current Population Survey,” <http://www.macalester.edu/~kaplan/ism/>, from Daniel Kaplan, *Statistical Modeling:  A Fresh Approach*.  Original source:  Berndt, ER.  *The Practice of Econometrics* 1991.  Addison-Wesley.

Metadata:  cps\_metadata.pdf, from p. 418 of *Statistical Modeling:  A Fresh Approach* by Daniel Kaplan.

1. Read the data into R and plot wages versus education.  Comment on the appropriateness of linear regression.

**>** cps\_CSV **=** read.csv**(**"C:/Users/pedbv9699/Documents/GitHub/ds710assignment9/cps.csv"**)**

**>**

**>** attach**(**cps\_CSV**)**

**>** head**(**cps\_CSV**)**

wage educ race sex hispanic south married exper union age sector

1 9.0 10 W M NH NS Married 27 Not 43 const

2 5.5 12 W M NH NS Married 20 Not 38 sales

3 3.8 12 W F NH NS Single 4 Not 22 sales

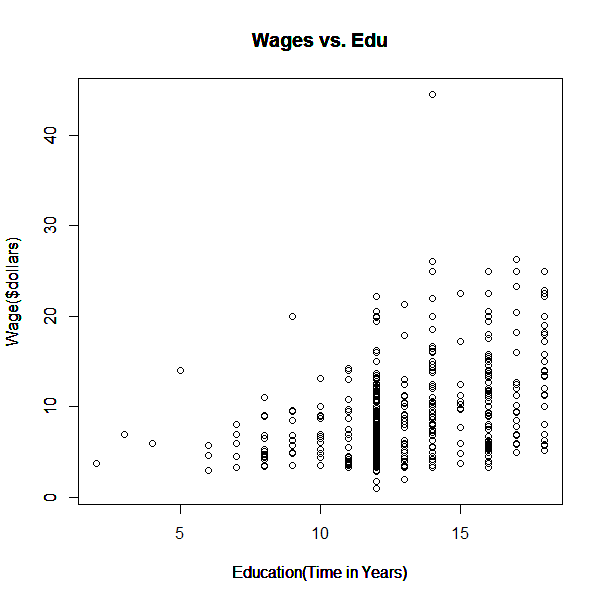
4 10.5 12 W F NH NS Married 29 Not 47 clerical

5 15.0 12 W M NH NS Married 40 Union 58 const

6 9.0 16 W F NH NS Married 27 Not 49 clerical

It is quite obvious taking look at the plot, an increase in wage and years of education seem to have a positive relationship.

plot**(**educ, wage, main **=** "Wages vs. Edu", xlab **=** "Education(Time in Years)" , ylab **=** "Wage($dollars)"**)**



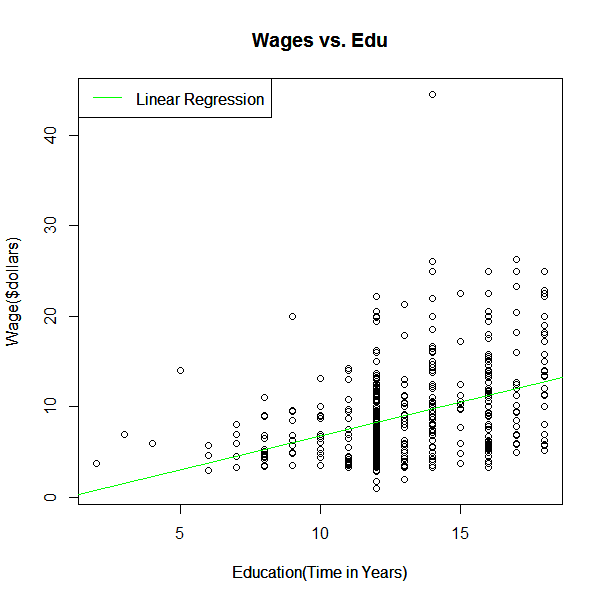
1. Perform the linear regression and examine the diagnostic plots.  Explain why transforming the wages variable is a good idea in this case.

**>** model **=** lm**(**wage**~**educ**)**

**>** plot**(**educ, wage, main **=** "Wages vs. Edu", xlab **=** "Education(Time in Years)" , ylab **=** "Wage($dollars)"**)**

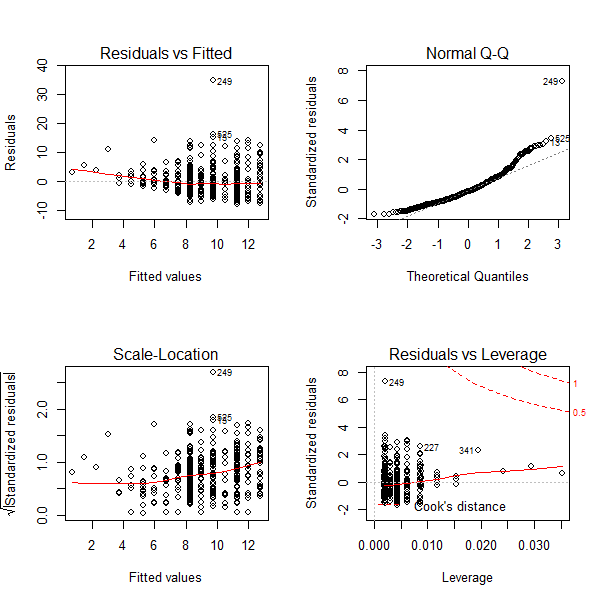
**>** abline**(**model,col **=** "green"**)**

**>** legend**(**"topleft",legend **=** "Linear Regression", col **=** "green", lwd **=** 1**)**

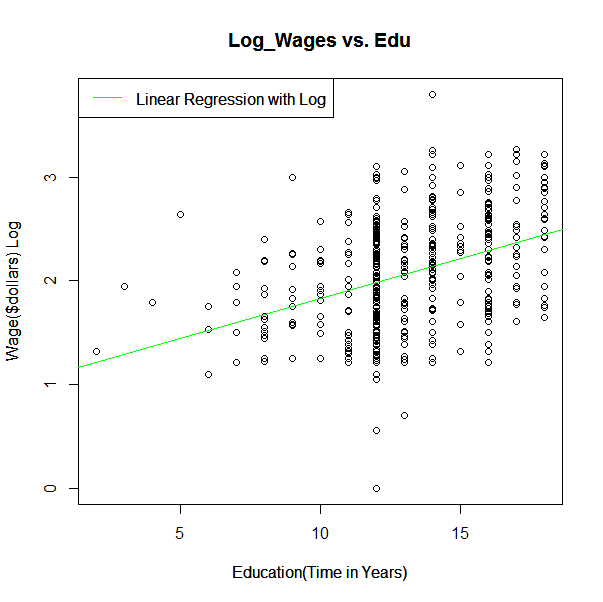


**>** par**(**mfrow **=** c**(**2,2**))**

**>** plot**(**model**)**



Q-Q plot seems to go off track starting the first theoretical quantile. There seems to be a quite a few outliers that may skew the data. The upper 13 and 14 year area are heavy and might throw off the overall or underlying data. A log transformation wouldn’t be a bad idea to consider.



**>** log\_wage **=** log**(**wage**)**

**>** plot**(**log\_wage **~** educ, main **=** "Log\_Wages vs. Edu", xlab **=** "Education(Time in Years)" , ylab **=** "Wage($dollars) Log"**)**

**>** model **=** lm**(**log\_wage**~**educ**)**

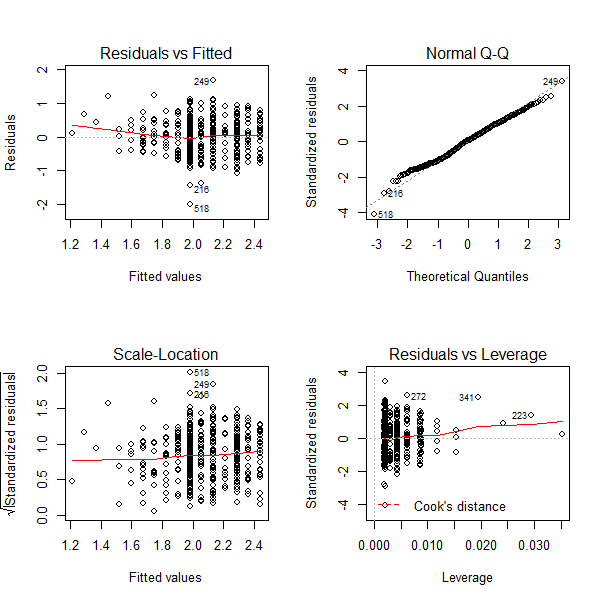
**>** legend**(**"topleft",legend **=** "Linear Regression with Log", col **=** "green", lwd **=** 1**)**

**>** abline**(**model, col **=** "green"**)**

Quite obvious the new Log model seems to be more fitting, the the abline splits directly through the plot almost dead split in the middle. I like the division and seems like the spread makes much more sense.

**(**mfrow **=** c**(**2,2**))**

**>** plot**(**model**)**



Q-Q line is very much aligned with the plots, and it seems to overlay almost perfectly. Linear regression here is much better fit and spread. Log transform here is much better fit and warranted for this particular instance.

1. The variable **wage** has units of dollars/hour.  Create a new variable, **time**, equal to 1/wage.  (So **time** has units of hours/dollar, or the length of time a person must work to earn $1.00.)

**>** head**(**cps\_CSV**)**

wage educ race sex hispanic south married exper union age sector

1 9.0 10 W M NH NS Married 27 Not 43 const

2 5.5 12 W M NH NS Married 20 Not 38 sales

3 3.8 12 W F NH NS Single 4 Not 22 sales

4 10.5 12 W F NH NS Married 29 Not 47 clerical

5 15.0 12 W M NH NS Married 40 Union 58 const

6 9.0 16 W F NH NS Married 27 Not 49 clerical

**>** time **=** 1 **/** wage

**>** head**(**time**)**

**[**1**]** 0.11111111 0.18181818 0.26315789 0.09523810 0.06666667 0.11111111

1. Plot time versus education.  Comment on the appropriateness of linear regression.

plot**(**educ, time, main **=** "Time vs Edu", xlab **=** "Yeard of Edu", ylab **=** "Time(hr/$dollar)"**)**



I see decrease in time per dollar in hours as years of education, and of course it is a good thing. From looking at the plot, the data points are very heavy on the right side than on the left side. I’d like to see more wider scale. Also The outliers at the top along 12 years of education, noticeable.

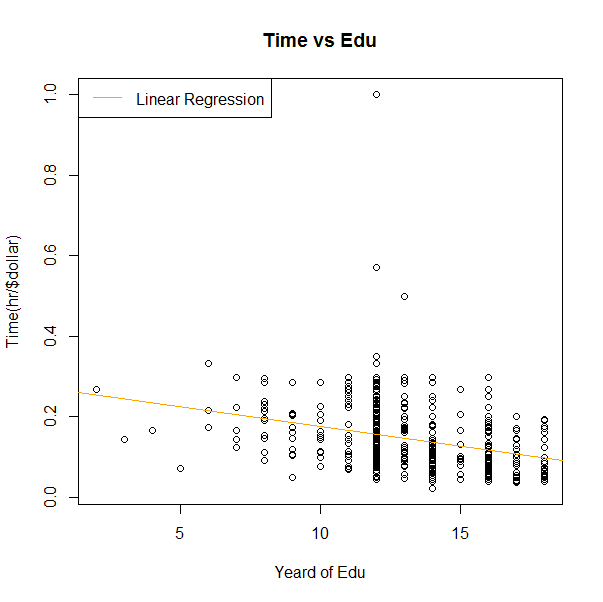
1. Perform the linear regression.  Based on these results, are you happy with your decision to pursue a master’s degree?  Explain.

**>** mod **=** lm**(**time **~** educ**)**

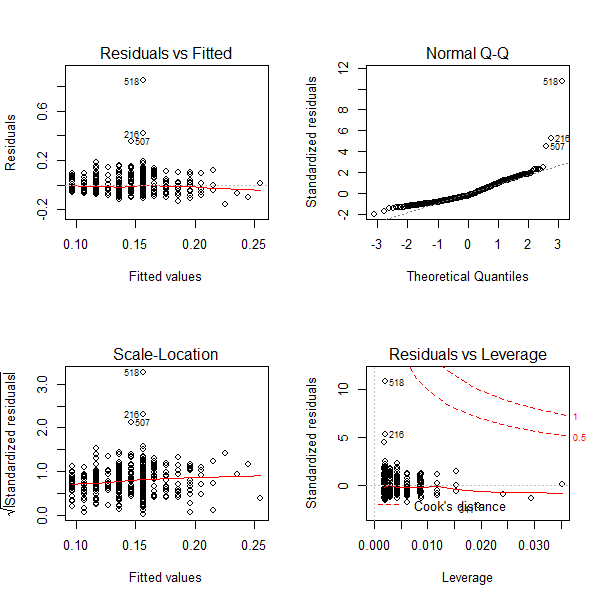
**>** plot**(**educ, time, main **=** "Time vs Edu", xlab **=** "Yeard of Edu", ylab **=** "Time(hr/$dollar)"**)**

**>** abline**(**mod, col **=** "orange"**)**

**>** legend**(**"topleft",legend **=** "Linear Regression", col **=** "orange", lwd **=** 1**)**



Yes absolutely, I am only 24 and I am already going through my Masters program in the field of Data Science. Arguably one of the hottest fields in the market with tremendous upside as far requirements and job demand goes. I am glad spending these last 2 years learning in-depth stats and programming. This char is some solid evidence showing a linear downward trend of time spent for earning dollar. Essentially indicating those who finished just a 4 year degree will likely work longer hours to make same or even lesser money than those with 16 years of education. The residual vs fitted shows a strong spread at y = 0 , x-axis. Q-Q plot seems very fitting but I do see few outliers. Those random dots hovering over R V F plot, Q-Q plot and the other two as well.



1. Examine the diagnostic plots.  Which individuals appear to be outliers on the residual vs. predicted plot?  Re-do the regression without these individuals.  Does your conclusion change?

**>** time\_outliers **=** time**[-**c**(**518,216,507**)]**

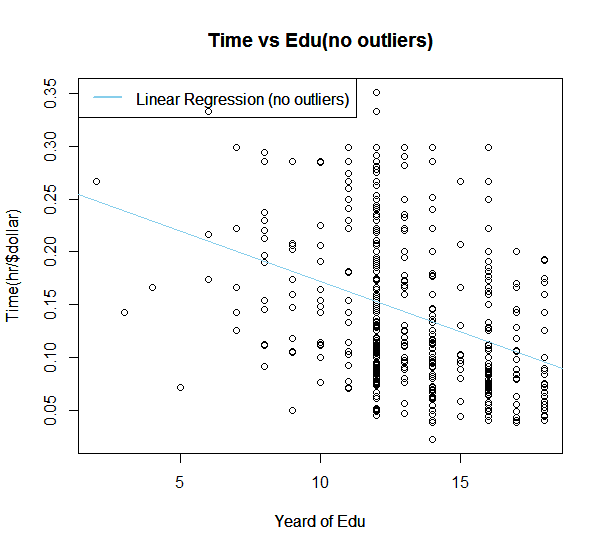
**>** education\_outliers **=** educ**[-**c**(**518,216,507**)]**

**>** final\_mod **=** lm**(**time\_outliers **~** education\_outliers**)**

**>** plot**(**education\_outliers, time\_outliers, main **=** "Time vs Edu(no outliers)", xlab **=** "Yeard of Edu", ylab **=** "Time(hr/$dollar)"**)**

**>** abline**(**final\_mod, col **=** "skyblue"**)**

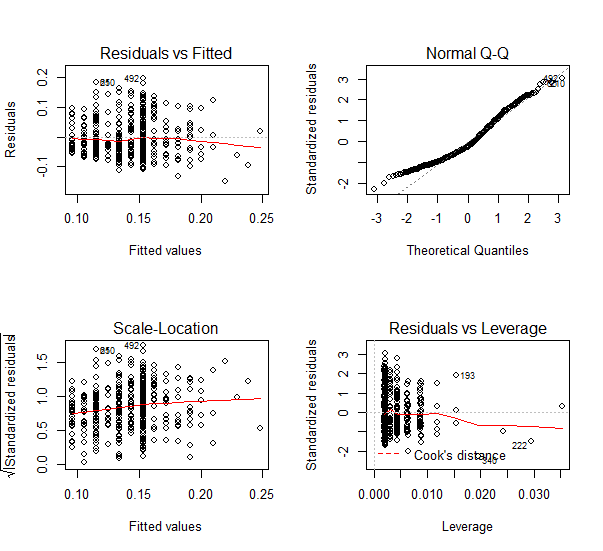
**>** legend**(**"topleft",legend **=** "Linear Regression (no outliers)", col **=** "skyblue", lwd **=** 2



> par(mfrow = c(2,2))

> plot(final\_mod)

The outliers seem to have not caused much significant change, yes they are neglible and my overall view and feel that education and experience increase will decrease amount of time needed to earn or make the same money. It makes sense because of the extra time one would spend could indeed mean senior level knowledge and expertise.



Submit a .doc, .docx, or .pdf file to GitHub, containing your R code, R output, and written interpretations and explanations.