# Author: Brian Pankau

# Class: DS710 Summer 2017

# Assignment: R 9-2

# Problem 2(a)

# Read the data from cps.csv into R and

# plot wages versus education.

# Comment on the appropriateness of linear regression.

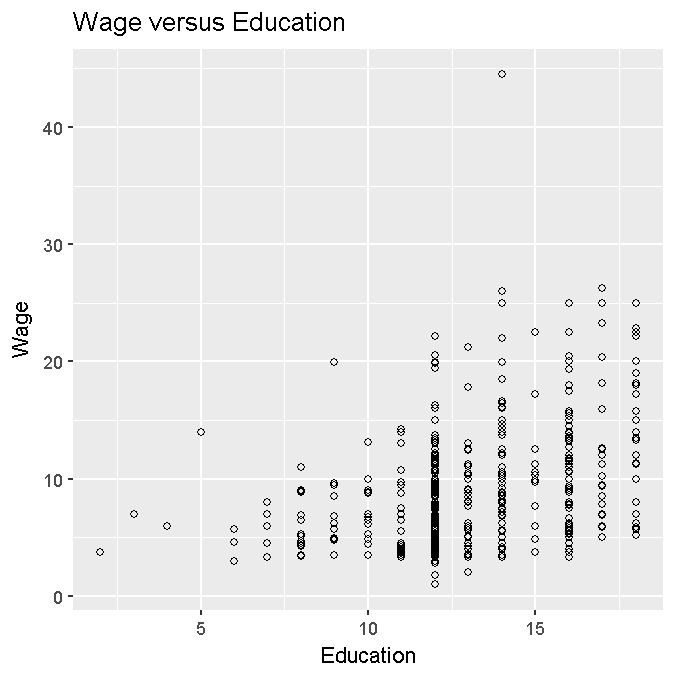
ggplot(dat, aes(x=df.educ, y=df.wage)) +

geom\_point(shape=1) +

labs(x = "Education") +

labs(y = "Wage") +

labs(title = "Wage versus Education")



Comment:

This scatter plot of Wage versus Education indicates that there is a positive and an increasing association between the number of years of education and the predicted Wage in dollars earned. These is a high degree of dispersal and range in the predicted value of Wage for a specific years of education.   
This scatter plot shows that the data is skewed left, indicating that a transformation of the Wage variable may be appropriate for application – prehaps using a square transformation.

# Problem 2(b)

# Perform the linear regression and examine the diagnostic plots.

# Explain why transforming the wages variable is a good idea in

# this case.

ggplot(dat, aes(x=df.educ, y=df.wage)) +

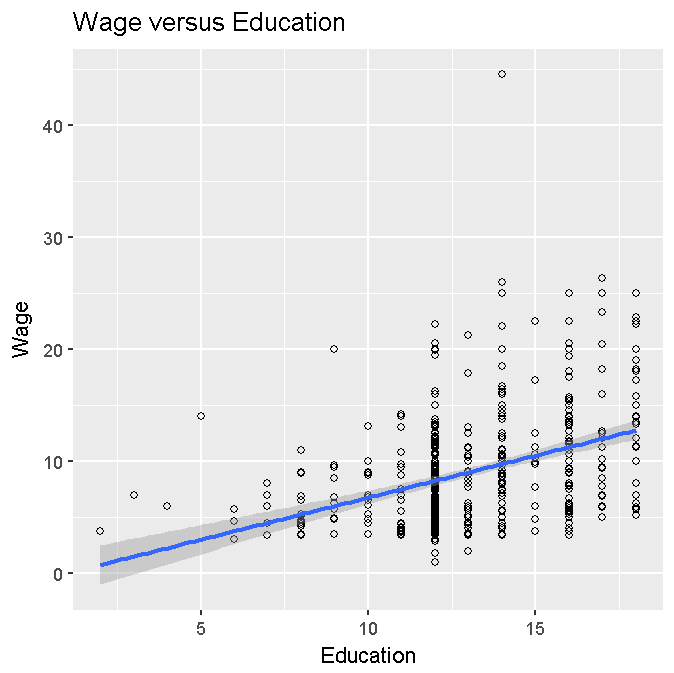
geom\_point(shape=1) +

labs(x = "Education") +

labs(y = "Wage") +

labs(title = "Wage versus Education") +

geom\_smooth(method=lm)



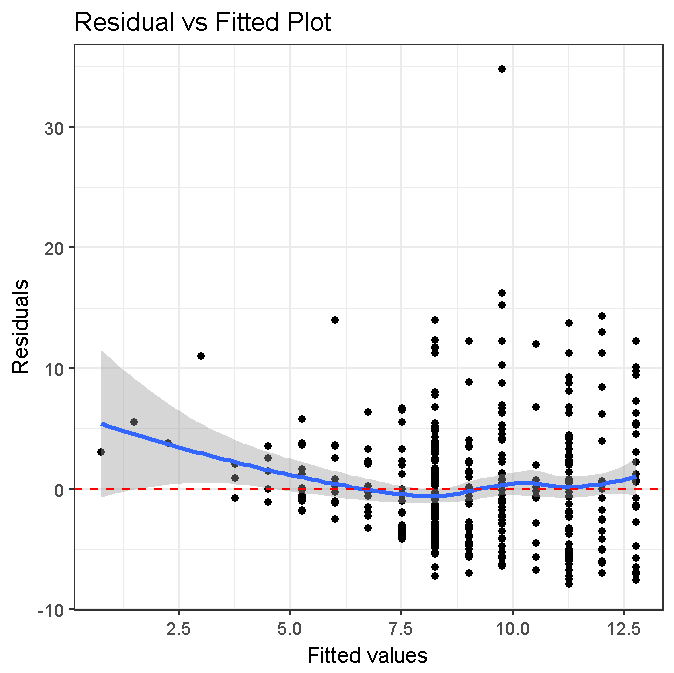
Comment:

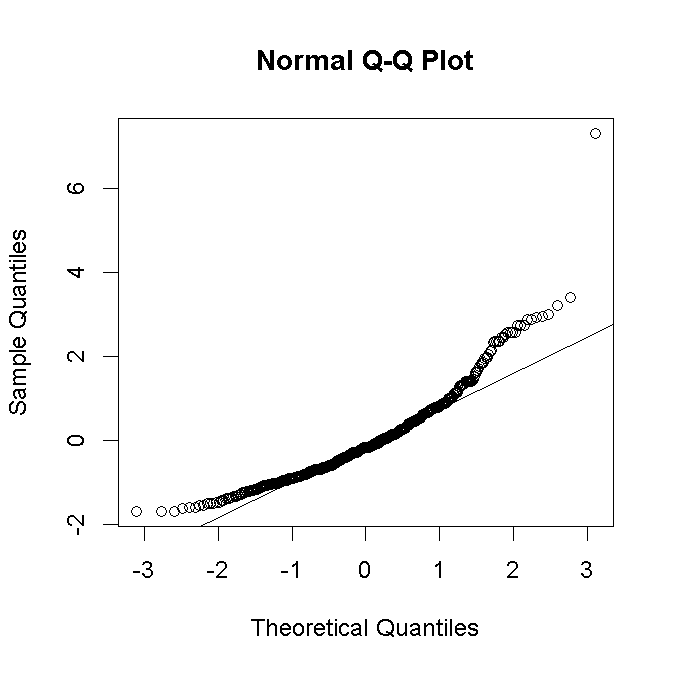
This scatter plot of Wage versus Education and the accompaning regression line indicates a positive and an increasing correlation between the number of years of education and the resulting income in wages. As indicated by the regression line, as the number of years increase, so does the size of the wage that is earned. The confidence of the association as indicated by the shaded area around the regression line. Since there is limited information at low levels of eduction, one can not say with a high degree of confidence about the predicted wage; in addition, there seems to be a high degree of variability in these few data points at the lower end of the education range (< ~7). Confidence of the prediction is higher towards the median of the education range, but again there is a dispersal of data points both above and below the regression line at these higher education ranges.

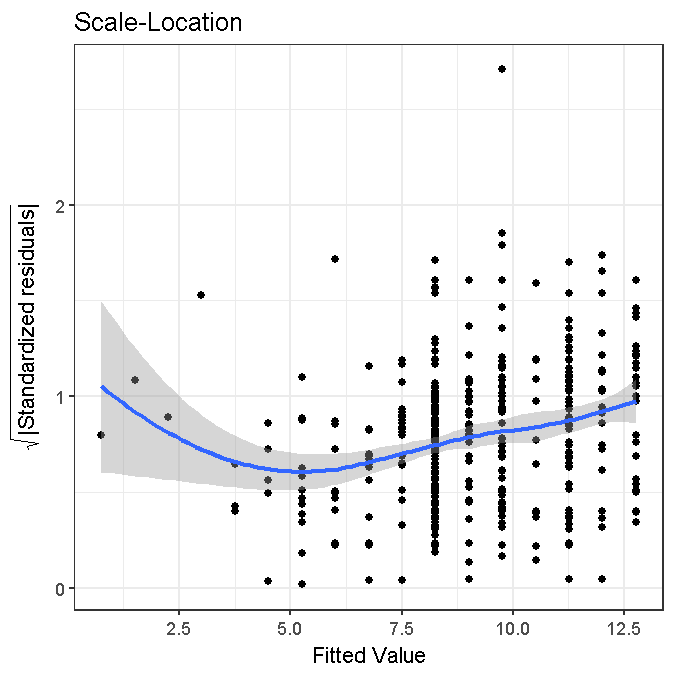
# generate diagnostic plots

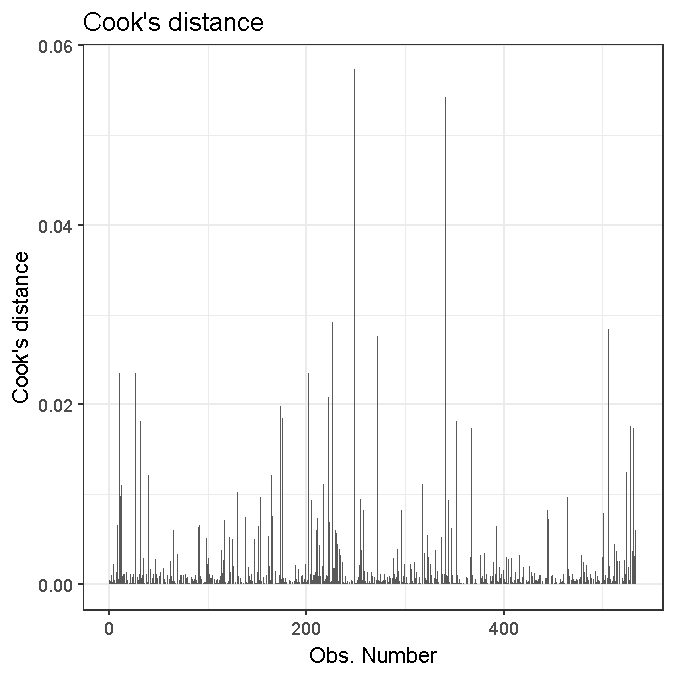
fit = lm(df.wage~df.educ, dat)

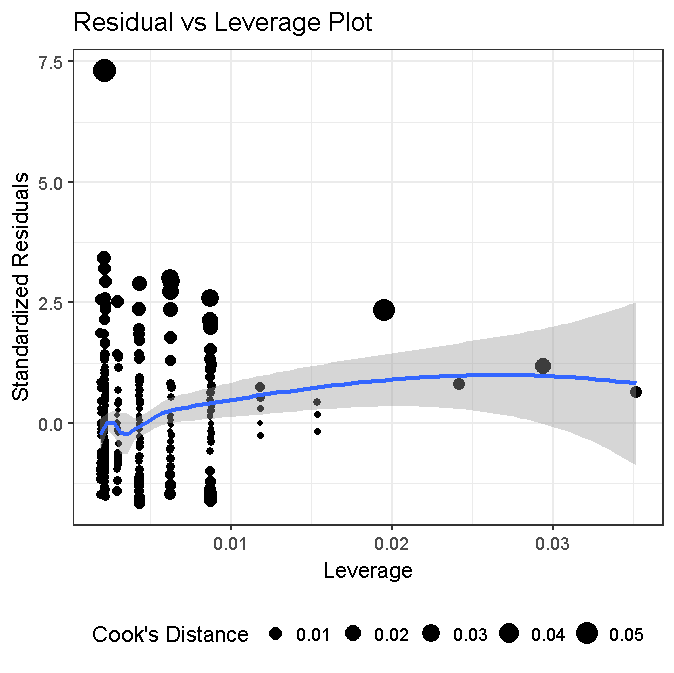
plot(fit)

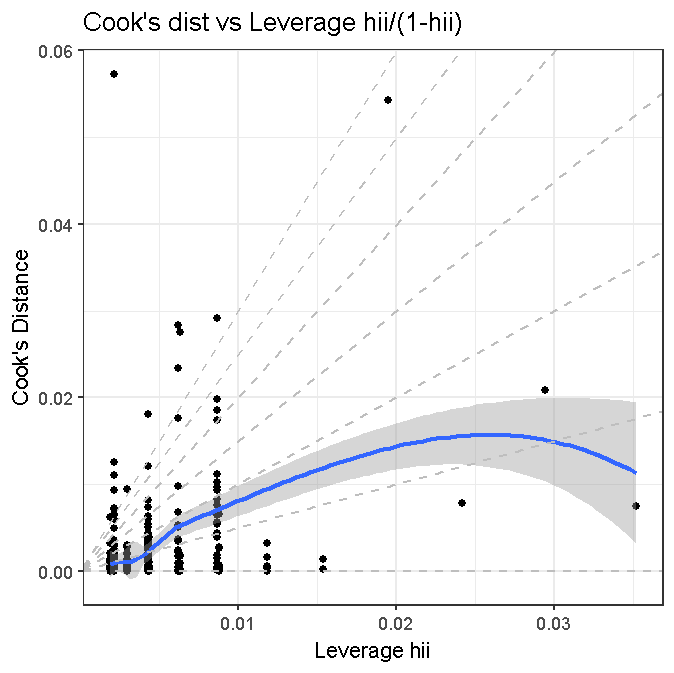












# Problem 2(c)

# The variable wage has units of dollars/hour.

# Create a new variable, time, equal to 1/wage.

dat$time <- with(

dat,

1/df.wage

)

summary(dat$time)

# Min. 1st Qu. Median Mean 3rd Qu. Max.

#0.02247 0.08889 0.12853 0.14624 0.19048 1.00000

# Problem 2(d)

# Plot time versus education.

# Comment on the appropriateness of linear regression.

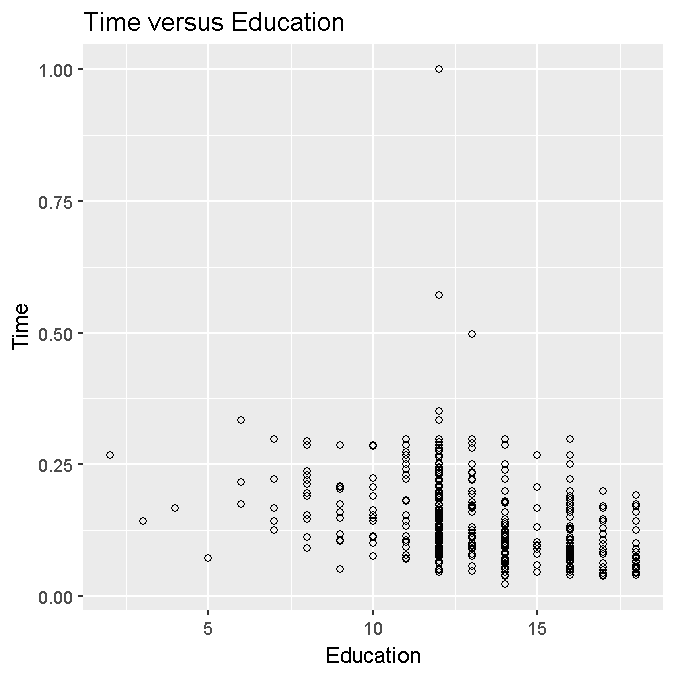
ggplot(dat, aes(x=df.educ, y=time)) +

geom\_point(shape=1) +

labs(x = "Education") +

labs(y = "Time") +

labs(title = "Time versus Education")



Comment:

By applying the inverse transformation on the Wage variable to generate Time (Time = 1/Wage), the distribution of the data points is less skewed to the left and the data points have less dispersion along the Time axis. However, there is still a grouping / concentration of the data points as the number of years of education increases, and there does not seem to be any indication of a linear assocation between the number of years of education and the predicted wage.

# Problem 2(e)

# Perform the linear regression. Based on these results, are you

# happy with your decision to pursue a master’s degree? Explain.

ggplot(dat, aes(x=df.educ, y=time)) +

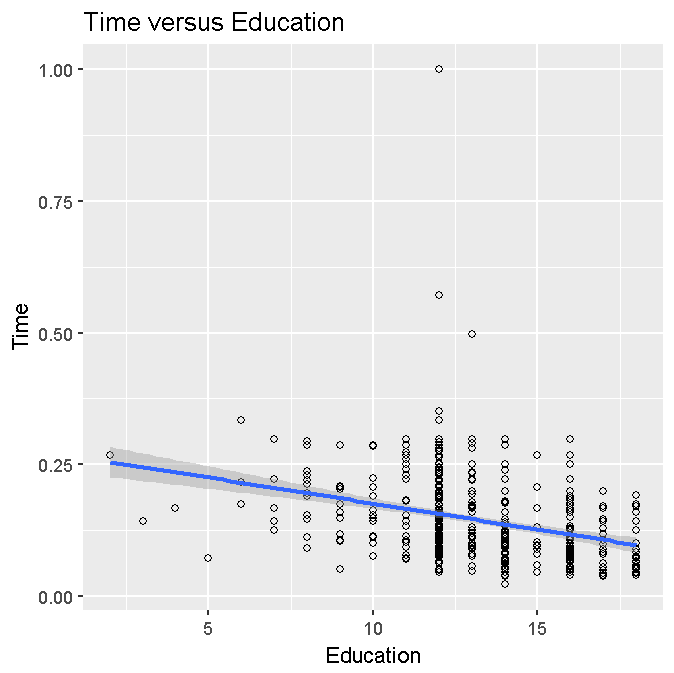
geom\_point(shape=1) +

labs(x = "Education") +

labs(y = "Time") +

labs(title = "Time versus Education") +

geom\_smooth(method=lm)



Comment:

The regression line in this transformed scatter plot indicates that as the number of years of education increases, the length of time that a person has to work to earn $1 decreases. There is a negative correlation between the years of education versus the amount of hours it takes a person to earn $1.

Based on this negative relationship, I am very happy with my decision to persue this advanced degree in data science, for this relationship indicates that it is probable that my earning power per time unit will increase once I achieve the degree.

# Problem 2(f)

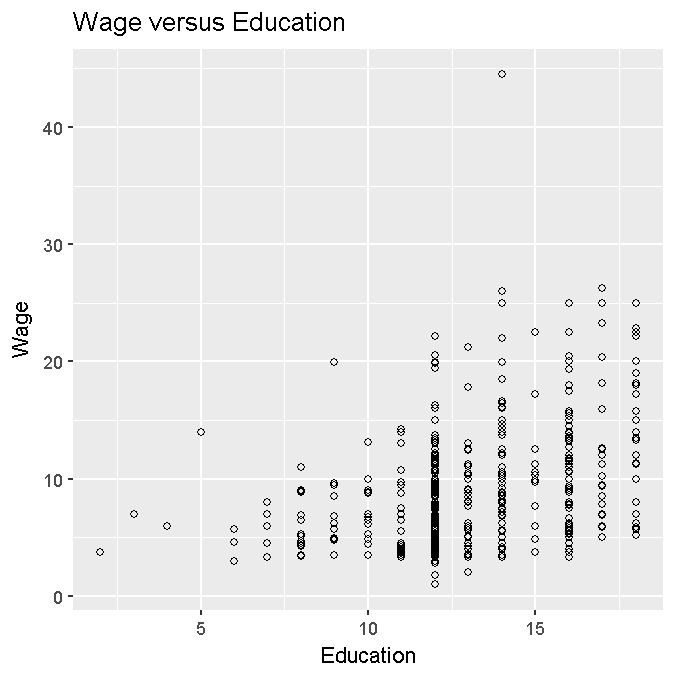
# Examine the diagnostic plots.

# Which individuals appear to be outliers on the residual vs. predicted plot?

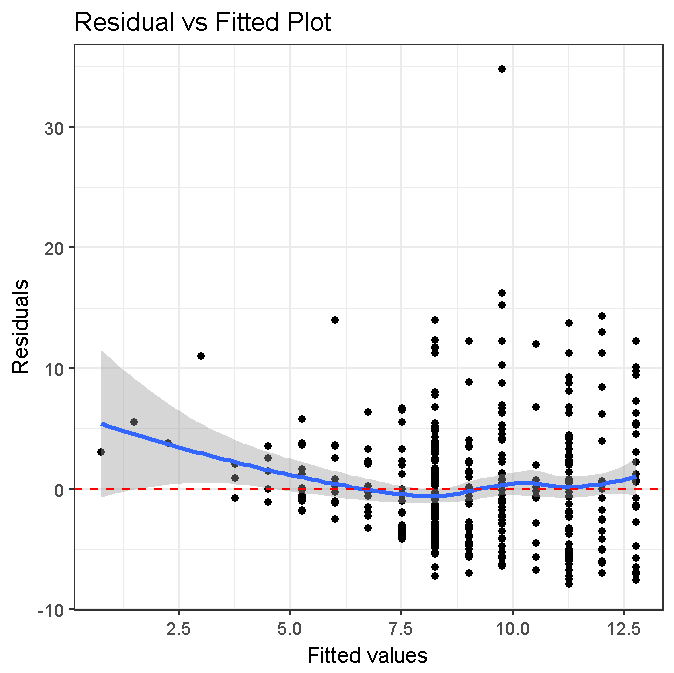
# Re-do the regression without these individuals.

# Does your conclusion change?

# With individuals



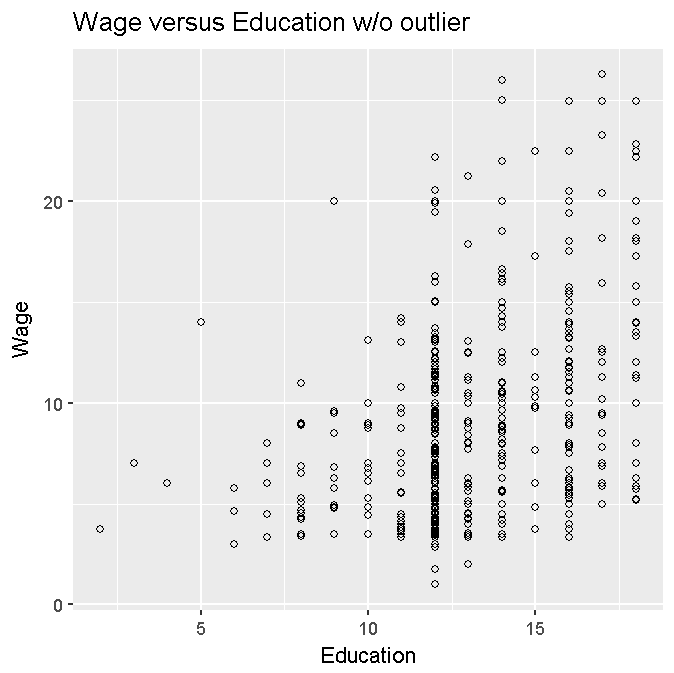
# With individuals

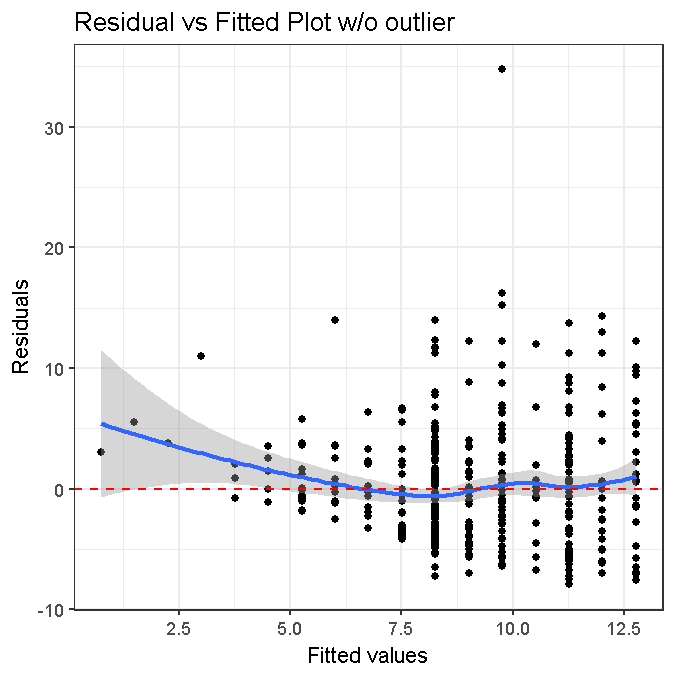


**Comment**

**By comparing the scatter plot to the Residual vs Fitted plot, we can see that there is an outlier in the upper right hand corner which can be removed (Wage = 40.50, Education = 14).**

**# with outlier removed:**





**Comment**

**By removing the outlier, the distribution in the scatter plot improves, but there still is a outlier in the Residual vs Fitted plot. My conclusion remains the same that there is an association between years of education and the predicted wage.**

**One could attempt at removing the outliers at the lower end of the education~wage ranges.**