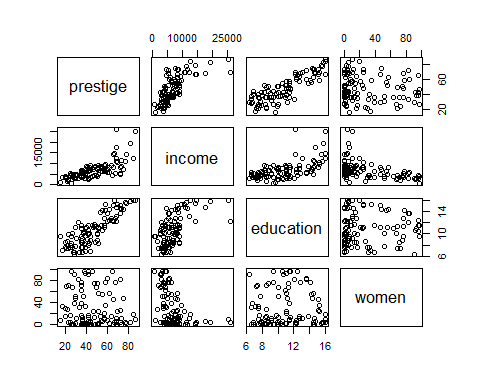
STAR511: HW 8

Megan Sears

# Q1



# Q2

## prestige income education women  
## prestige 1.0000000 0.7149057 0.85017689 -0.11833419  
## income 0.7149057 1.0000000 0.57758023 -0.44105927  
## education 0.8501769 0.5775802 1.00000000 0.06185286  
## women -0.1183342 -0.4410593 0.06185286 1.00000000

# Q3

##   
## Pearson's product-moment correlation  
##   
## data: prestige$income and prestige$prestige  
## t = 10.224, df = 100, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6044711 0.7983807  
## sample estimates:  
## cor   
## 0.7149057

# Q4

##   
## Call:  
## lm(formula = prestige$prestige ~ prestige$income)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33.007 -8.378 -2.378 8.432 32.084   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.714e+01 2.268e+00 11.97 <2e-16 \*\*\*  
## prestige$income 2.897e-03 2.833e-04 10.22 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.09 on 100 degrees of freedom  
## Multiple R-squared: 0.5111, Adjusted R-squared: 0.5062   
## F-statistic: 104.5 on 1 and 100 DF, p-value: < 2.2e-16

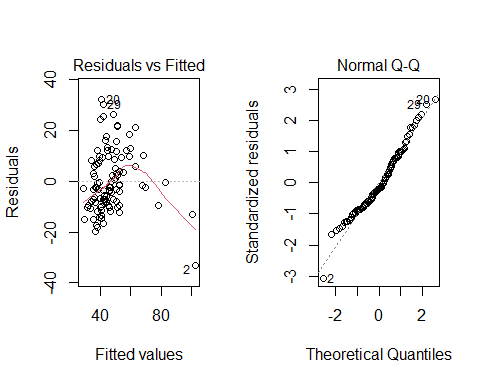
# Q5

Based on the p-value (less than alpha) from Pearson’s correlation and the correlation coefficient (0.71), there is a linear association between the two variables. Further, since there is a positive correlation (0.71), as income rises so does prestige. The null hypothesis, that there is no linear association or that the slope is 0, should be rejected based on the p-value from the correlation test and linear model (slope p-value).

# Q6

The slope from the linear model is 0.0029 with an associated p-value less than 0.05. Therefore, for one dollar increase in income, estimate average prestige changes by 0.0029.

# Q7



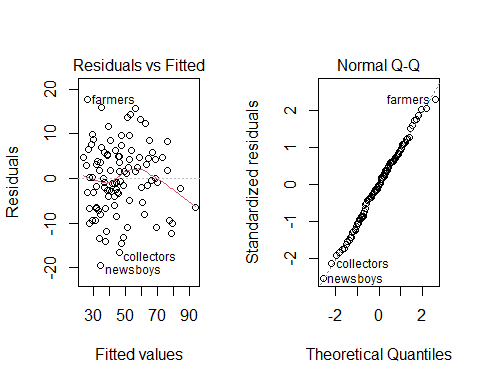
# Q8

##   
## Call:  
## lm(formula = prestige ~ income + education, data = prestige)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.4040 -5.3308 0.0154 4.9803 17.6889   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.8477787 3.2189771 -2.127 0.0359 \*   
## income 0.0013612 0.0002242 6.071 2.36e-08 \*\*\*  
## education 4.1374444 0.3489120 11.858 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.81 on 99 degrees of freedom  
## Multiple R-squared: 0.798, Adjusted R-squared: 0.7939   
## F-statistic: 195.6 on 2 and 99 DF, p-value: < 2.2e-16

# Q9

The slope from the linear model is 0.0014 with an associated p-value less than 0.05. Therefore, for every dollar increase in income, prestige changes by 0.0014 assuming the value of education is held constant.

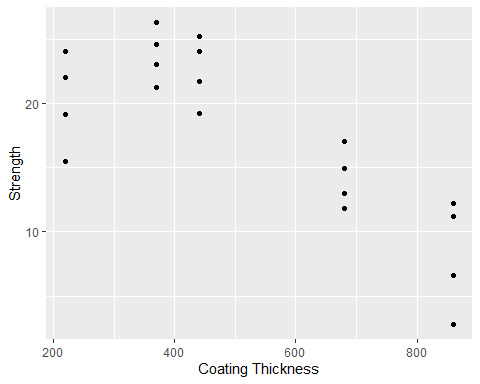
# Q10



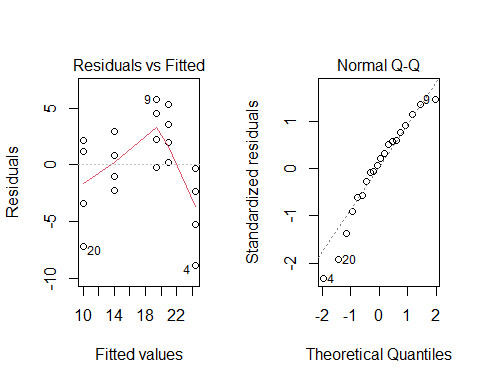
# Q11

The R-squared value is 0.798. Therefore, approximately 80% of the variability found in prestige can be explained by the linear regression on income and education.

# Q12



# Q13



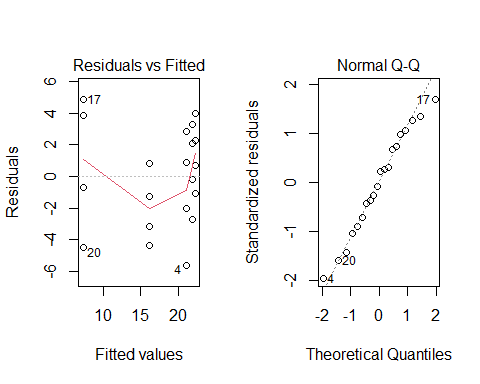
# Q14

No, the regression assumptions have not been met for equal variance and normality. The lower and higher fitted values on the residuals versus fitted plot have smaller residual values than the rest causing an almost upside u-shape in the residual data. Additionally, based on the Q-Q plot, the data are left skewed, or lower values are not near the line.

# Q15

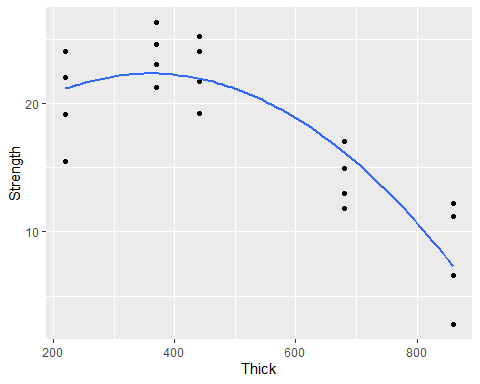
##   
## Call:  
## lm(formula = Strength ~ Thick + I(Thick^2), data = steel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.6222 -2.1960 0.2443 2.4491 4.8763   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.452e+01 4.752e+00 3.057 0.00713 \*\*  
## Thick 4.318e-02 1.980e-02 2.181 0.04354 \*   
## I(Thick^2) -5.994e-05 1.786e-05 -3.357 0.00374 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.268 on 17 degrees of freedom  
## Multiple R-squared: 0.7796, Adjusted R-squared: 0.7537   
## F-statistic: 30.07 on 2 and 17 DF, p-value: 2.609e-06

# Q16



Based on the Q-Q plot from Q15, the data appear to be normal. This is an improvement from the left skew seen in the Q-Q plot from Q13. The residuals plot has larger spread for tail ends of the data, but appears to be an improvement from the the residuals versus fitted plot in Q13.

# Q17



# Appendix

#Retain (and do not edit) this code chunk!!!  
library(knitr)  
knitr::opts\_chunk$set(echo = FALSE)  
knitr::opts\_chunk$set(message = FALSE)  
  
library(tidyverse)  
library(here)  
library(ggplot2)  
  
#Q1  
prestige <- read.csv('C:/Users/sears/Documents/Repos/STAR511/HW8/Prestige.csv', row.names=1)  
  
pairs(prestige)  
  
#Q2  
cor(prestige)  
  
#Q3  
cor.test(x=prestige$income, y=prestige$prestige)  
  
#Q4  
summary(lm(prestige$prestige ~ prestige$income))  
  
#Q7  
model <- lm(prestige$prestige ~ prestige$income)  
  
par(mfrow=c(1,2))  
  
plot(model, which = c(1:2))  
  
#Q8  
model2 <- lm(prestige ~ income + education, data = prestige)  
summary(model2)  
  
#Q10   
par(mfrow=c(1,2))  
  
plot(model2, which = c(1:2))  
  
#Q12  
steel <- read\_csv('C:/Users/sears/Documents/Repos/STAR511/HW8/Steel.csv')  
  
ggplot(steel, aes(x=Thick, y=Strength)) +  
 geom\_point() +   
 labs(x='Coating Thickness') #+  
 #geom\_smooth(method = 'lm')  
  
#Q13  
steelmod <- lm(steel$Strength ~ steel$Thick)  
  
par(mfrow=c(1,2))  
  
plot(steelmod, which = c(1:2))  
  
#Q15  
steelmod\_quad <- lm(Strength ~ Thick + I(Thick^2), data = steel)  
  
summary(steelmod\_quad)  
  
#Q16  
par(mfrow=c(1,2))  
  
plot(steelmod\_quad, which = c(1:2))  
  
qplot(x = Thick, y = Strength, data = steel) +  
 geom\_smooth(method = 'lm', formula = y ~ poly(x, 2), se = FALSE)