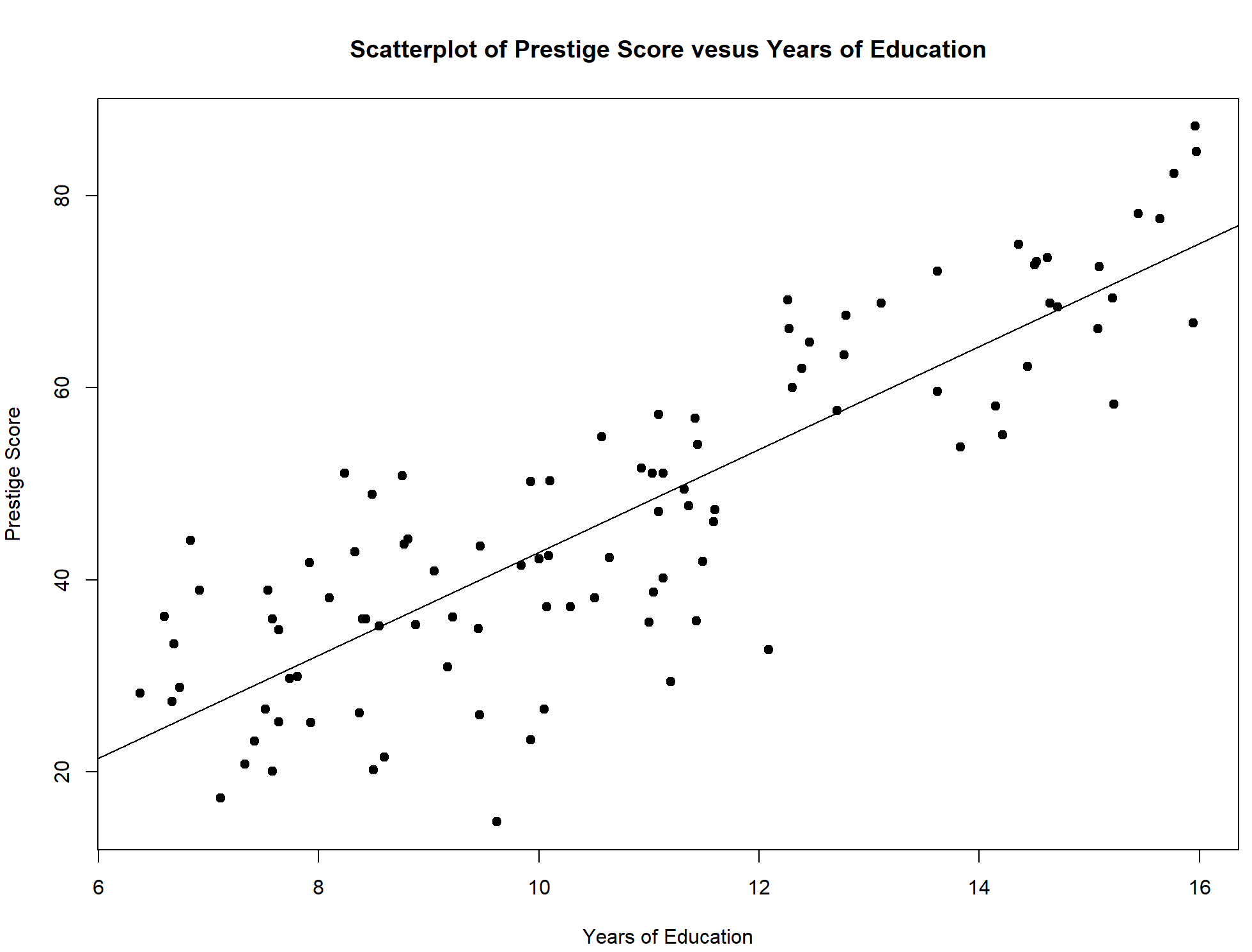
Yiduo Feng

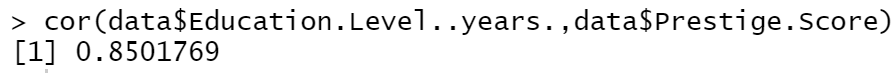
CS 555

Homework 4

08/02/2022

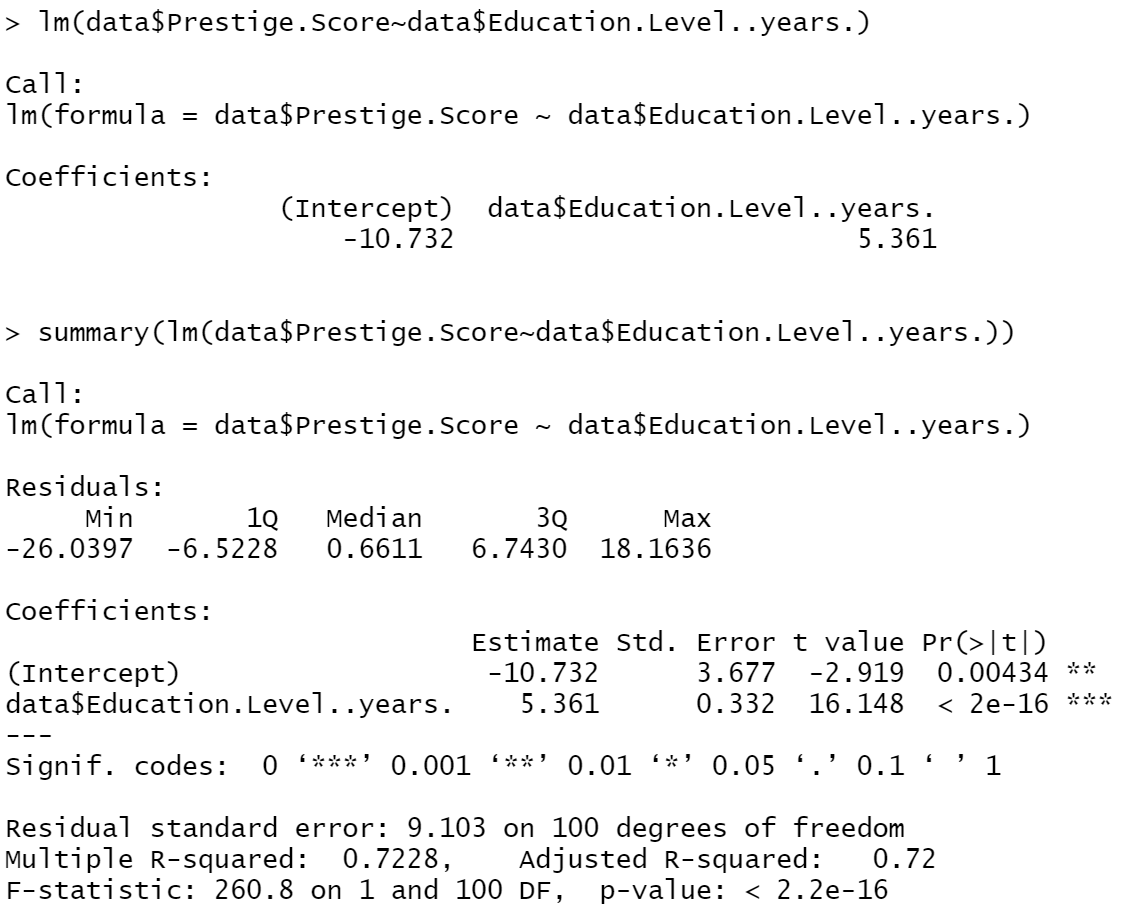
**(1) To get a sense of the data, generate a scatterplot to examine the association between** **prestige score and** **years of education. Briefly describe the form, direction, and strength of the association between the variables. Calculate the correlation coefficient.**

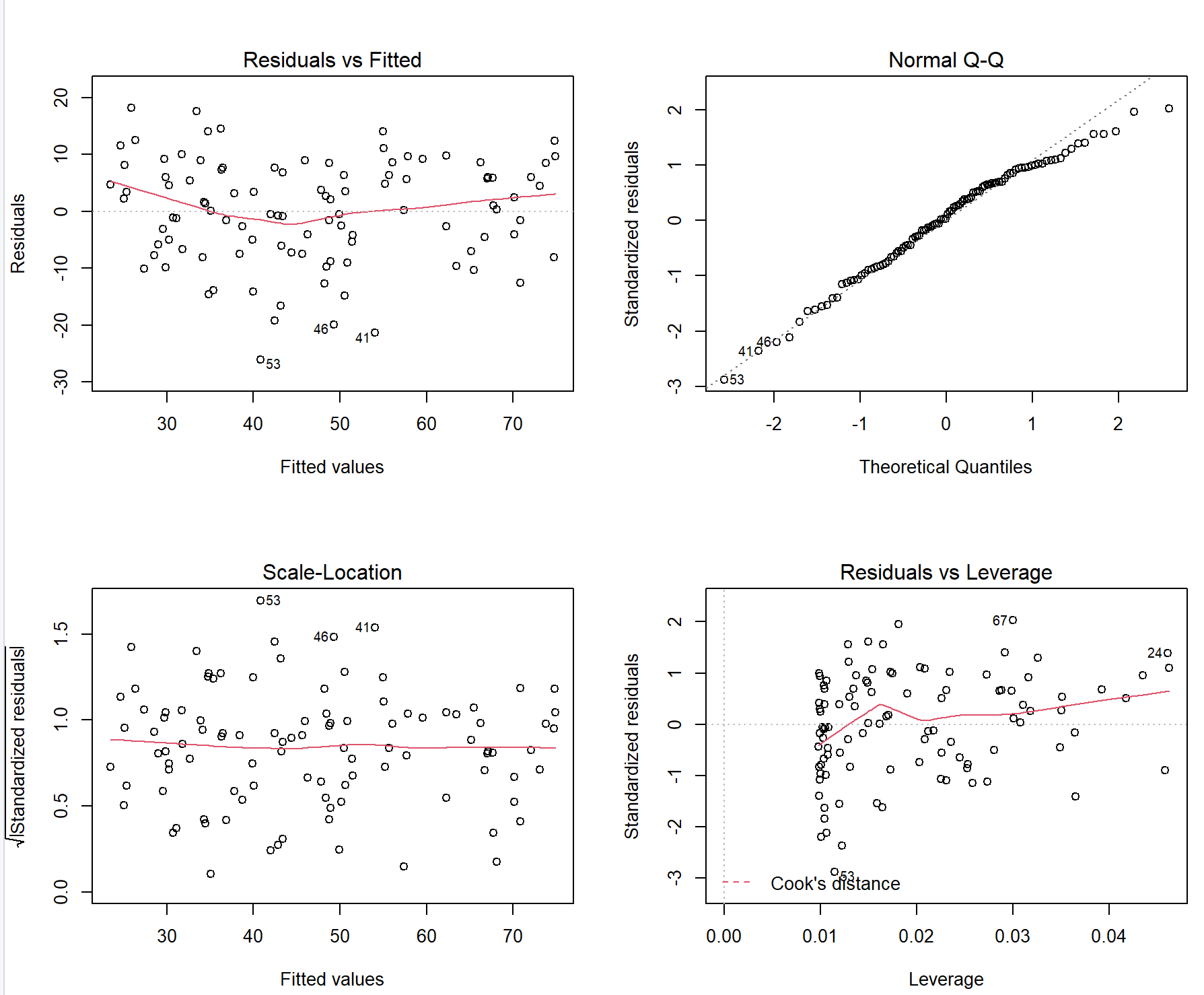




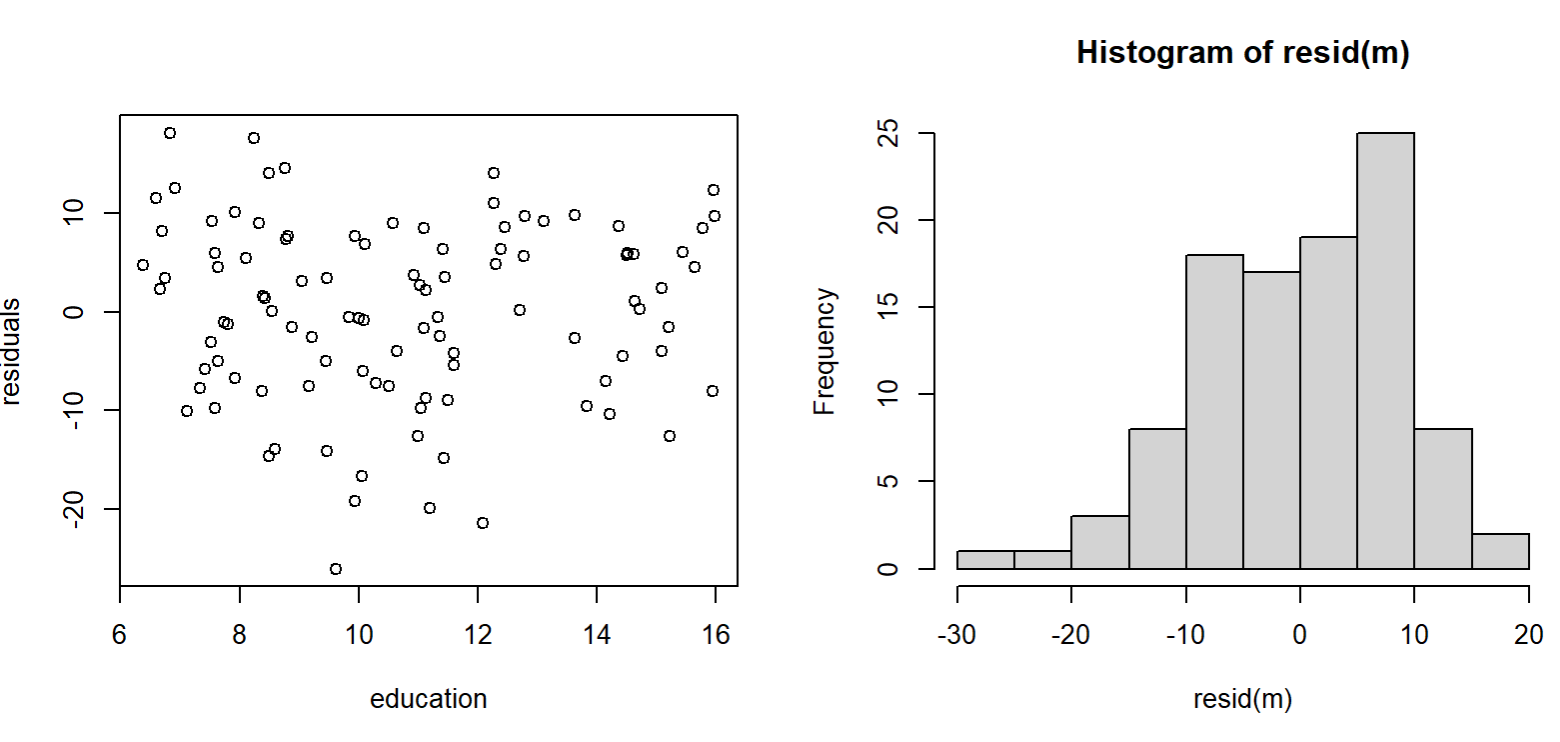
We can see a upward trend in the graph. The years of education increases the prestige score, so the direction is positive. The form is linear, and it is strong relationship, and we can see that there is few outlier. By using R, the correlation coefficient is 0.8501769. The greater the absolute value of the correlation coefficient, the stronger the correlation, the closer the correlation coefficient is to 1 or -1, the stronger the correlation, the closer the correlation coefficient is to 0, the weaker the correlation. 0.8501769 shows a large correlation coefficient and close to 1, so this correlation is strong.

**(2) Perform a simple linear regression with prestige score and years of education, and briefly summarize your conclusions (no need to do the 5-step procedure here). Generate a residual plot. Assess whether the model assumptions are met. Are there any outliers or influence points? If so, identify them by ID and comment on the effect of each on the regression.**

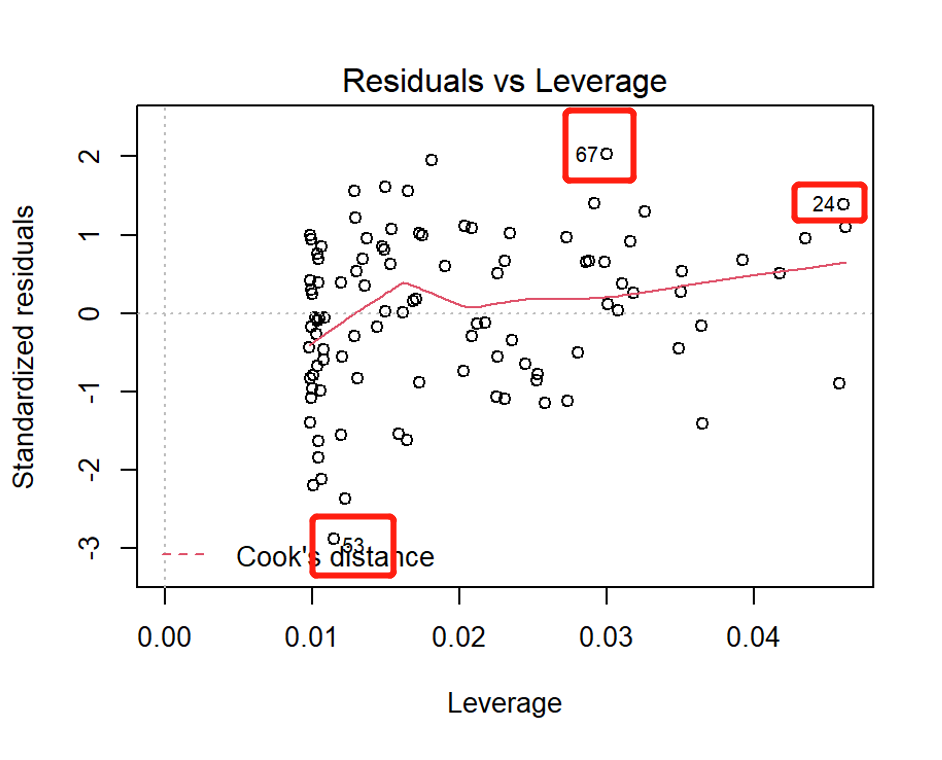




There is a violation of assumption of constant variance for residuals.

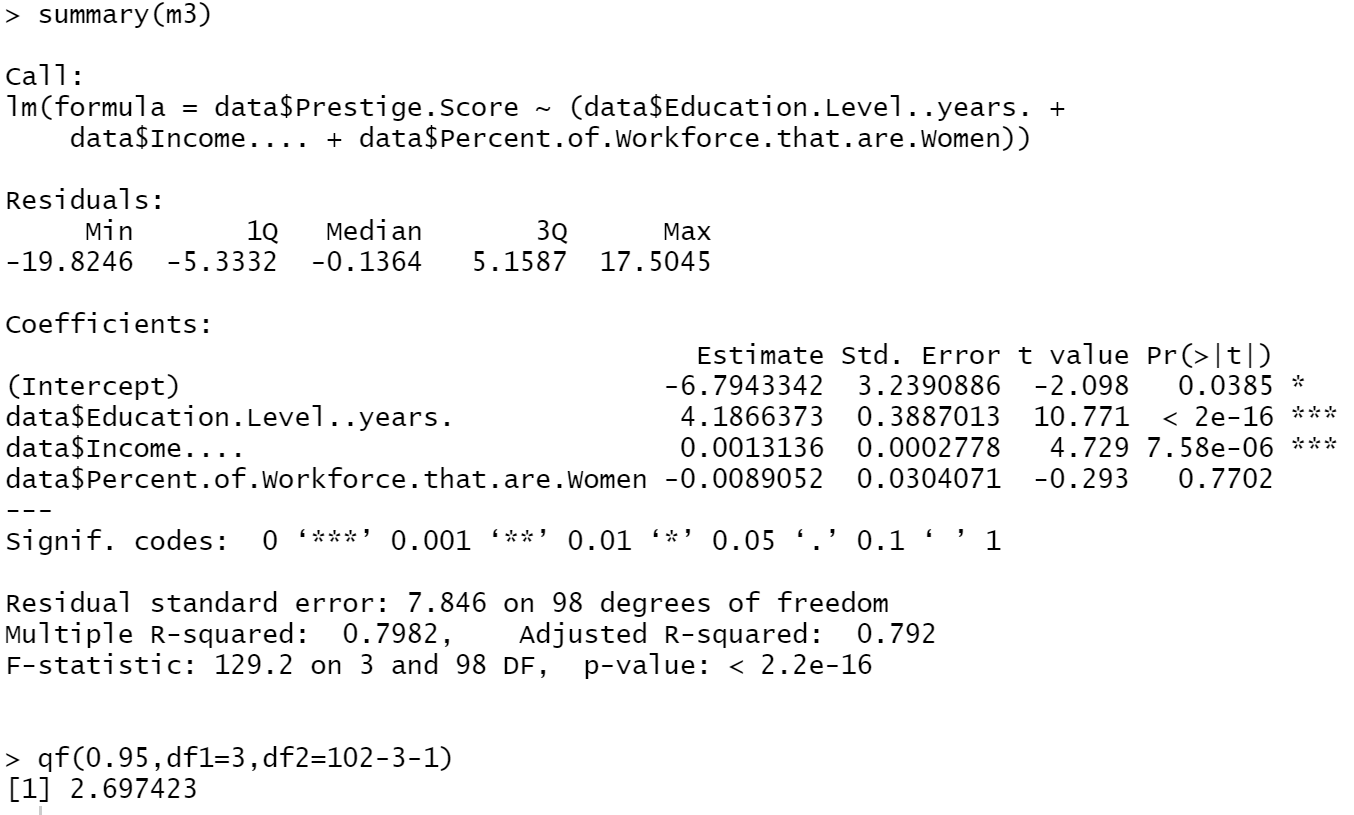


It is kind of normal and a little left skewed.



There is 3 outlier in the graph which are 24,53,67. Only 53 is lower outlier.

**(3) Calculate the least squares regression equation that predicts prestige score from education, income, and percentage of women. Formally test (using the 5-step procedure) whether the set of these predictors are associated with** **prestige score at the α = 0.05 level (Hint: You should be performing the global test).**



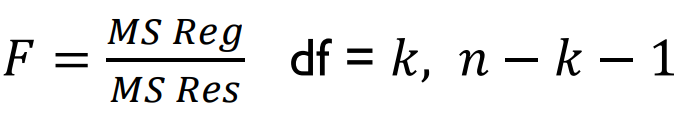
1. Set up the hypotheses and select the alpha level

H0 : 𝛽i= 0 (the set of these predictors are not associated with prestige score)

H1 : 𝛽i≠ 0 (the set of these predictors are associated with prestige score)

𝛼 = 0.01

1. Select the appropriate test-statistic



1. State the decision rule

Decision Rule: Reject H0 if 𝑝 ≤ 𝛼. Otherwise, do not reject H

1. Compute the test statistic

According to the code above,



P < 0.05

1. Conclusion

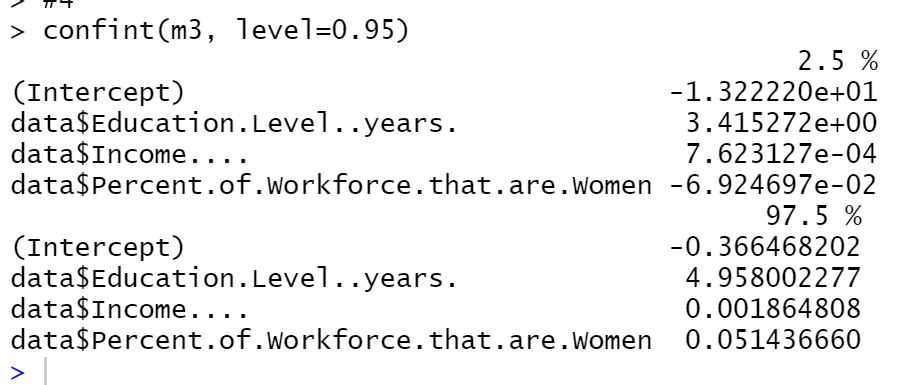
Reject H0

since 𝑝 ≤ 𝛼. We have significant evidence at the 𝛼 = 0.01 level

that these predictors when taken together are predictive of prestige score.

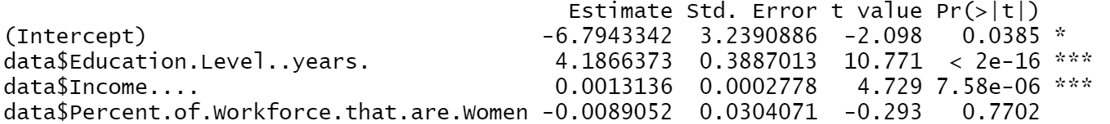
That is, there is evidence of a linear association between prestige score and Education Level , Income, Percent of Workforce that are Women. (here, p < 0.001).

**(4) If the overall model was significant, summarize the information about the contribution of each variable separately at the same significance level as used for the overall model *(no need to do a formal 5-step procedure for each one, just comment on the results of the tests)*. Provide interpretations for any estimates (of the slopes) that are significant. Calculate 95% confidence intervals for any estimates that are significant.**



According to the code above and the results from question 3

Except percent of workforce that are women, education and income are significant predictors due to the p value of percent of women is negative < 0.05.

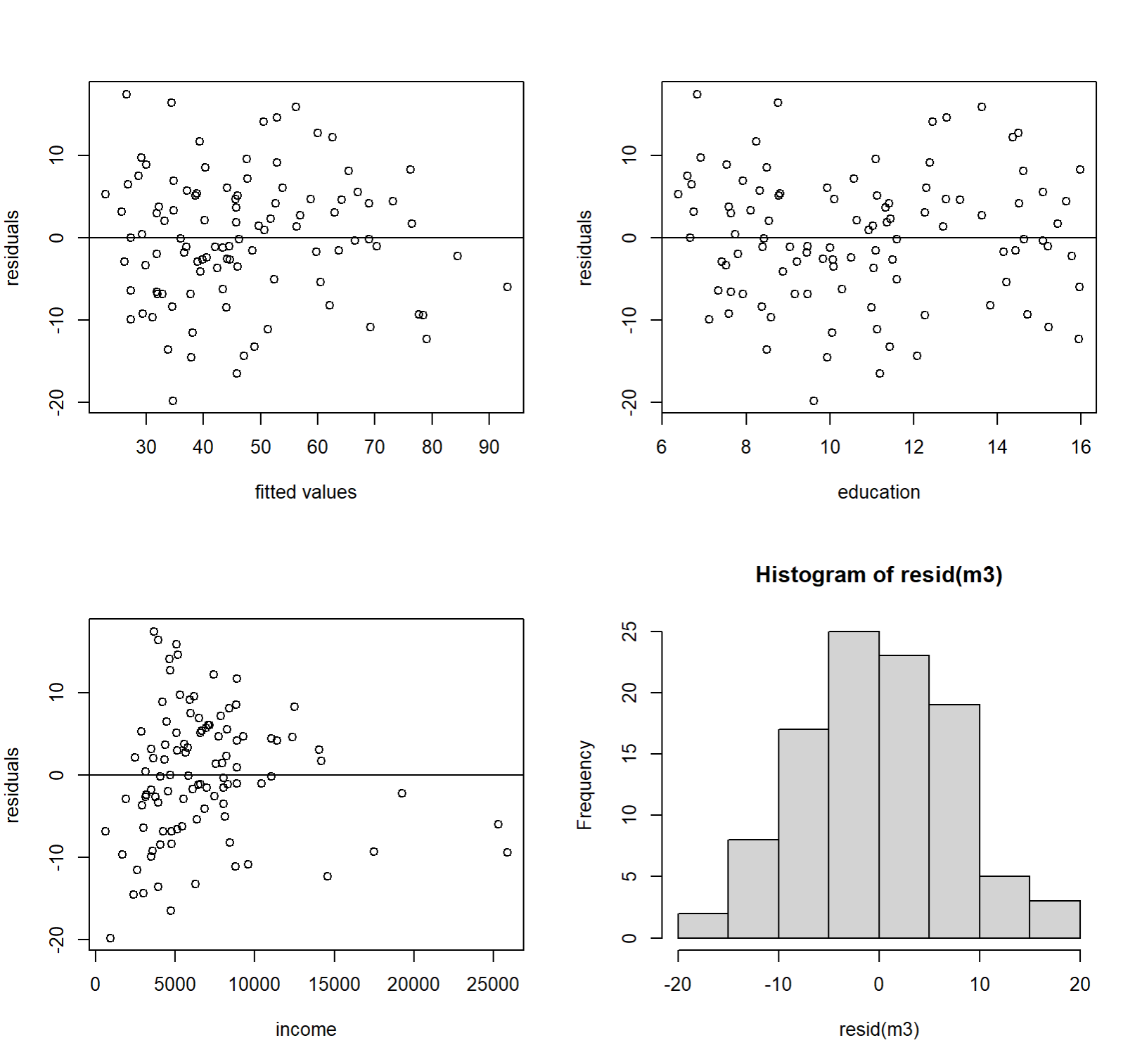


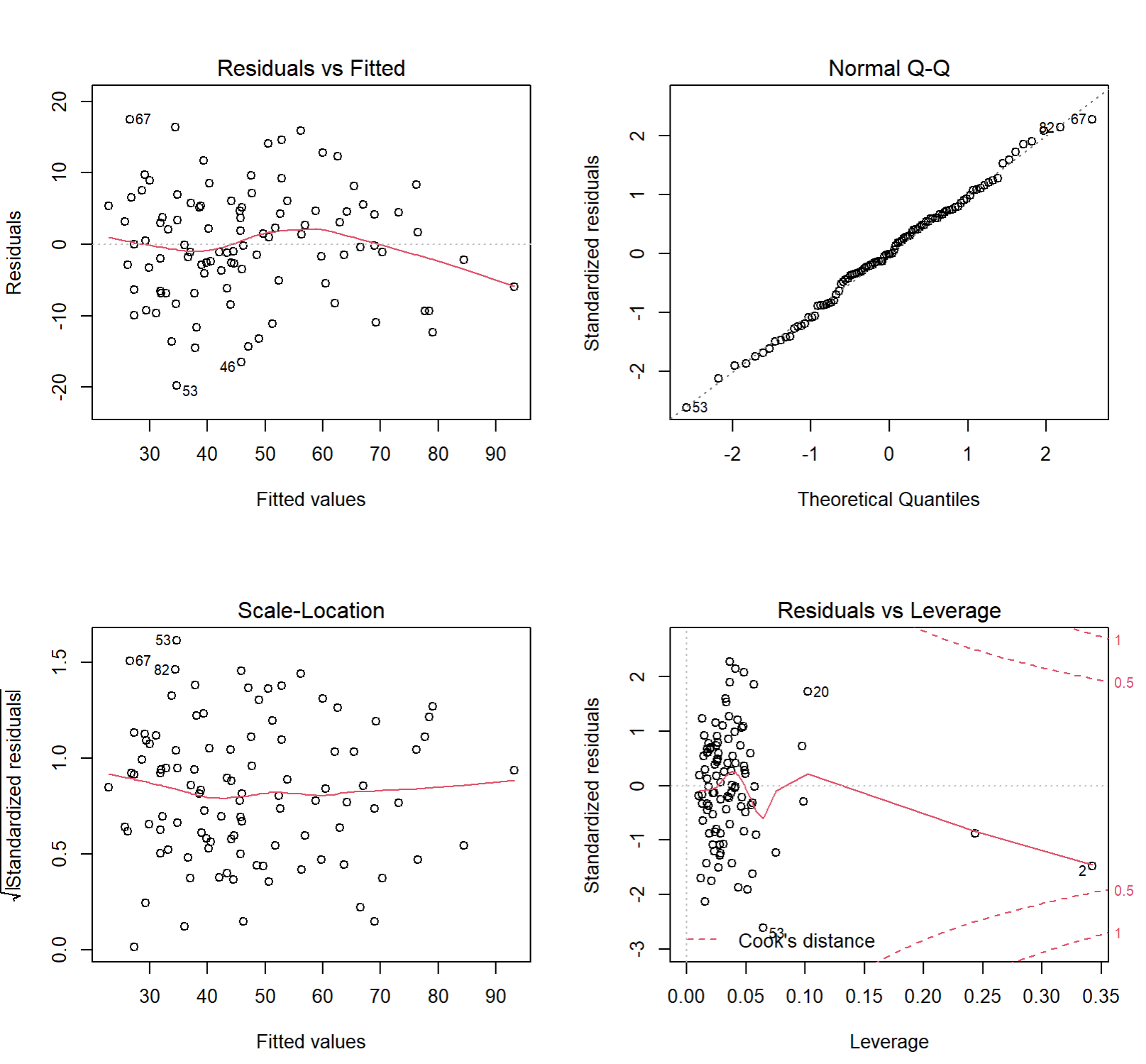
The estimate of education is 4.1866373, so that means when the year of education increase 1, the prestige score will increase about 4.1866373. The estimate of income is 0.0013136, so that means when the year of education increase 0.0013136, the prestige score will increase about 0.0013136.

Thus the 95% confidence interval for education is (3.415272, 4.958002277)

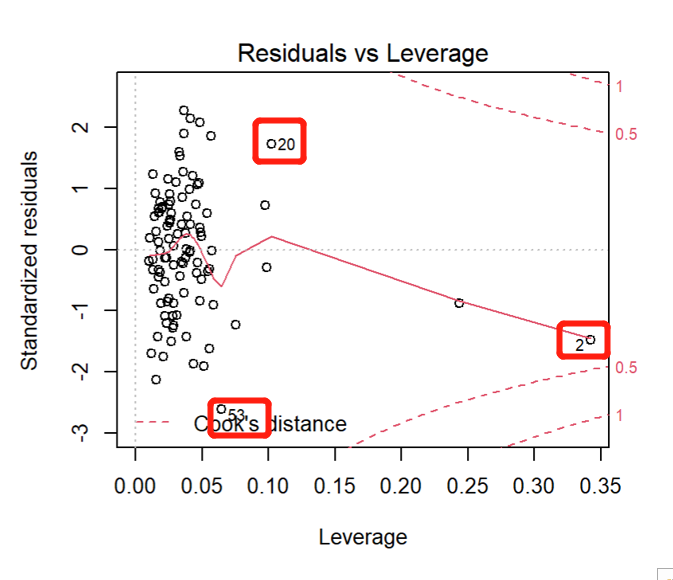
Thus the 95% confidence interval for income is (7.623127e-04, 0.001864808)

**(5) Generate a residual plot showing the fitted values from the regression against the residuals. Is the fit of the model reasonable? Are there any outliers or influence points?**





The fit of the model is reasonable.



There are 3 outliers which are 2, 20,53.

Code:

data <- read.csv(file = 'C:/Users/Yidow/Desktop/canadian.csv', fileEncoding="UTF-8-BOM")

plot(data$Education.Level..years.,data$Prestige.Score,

main="Scatterplot of Prestige Score vesus Years of Education",

xlab = "Years of Education", ylab="Prestige Score", pch=19)

abline(lm(data$Prestige.Score~data$Education.Level..years.))

cor(data$Education.Level..years.,data$Prestige.Score)

m <- lm(data$Prestige.Score~data$Education.Level..years.)

summary(m)

anova(m)

confint(m, level=0.95)

par(mfrow=c(2,2))

plot(m)

plot(data$Education.Level..years.,resid(m), axes=TRUE, frame.plot=TRUE, xlab = "education",

ylab="residuals")

hist(resid(m))

#3

m3 <- lm(data$Prestige.Score~(data$Education.Level..years.+data$Income....+data$Percent.of.Workforce.that.are.Women))

summary(m3)

qf(0.95,df1=3,df2=102-3-1)

#4

confint(m3, level=0.95)

#5

par(mfrow=c(2,2))

plot(fitted(m3),resid(m3), axes=TRUE, frame.plot=TRUE, xlab = "fitted values", ylab="residuals")

abline(h=0)

plot(data$Education.Level..years.,resid(m3), axes=TRUE, frame.plot=TRUE, xlab = "education", ylab="residuals")

abline(h=0)

plot(data$Income....,resid(m3), axes=TRUE, frame.plot=TRUE, xlab = "income", ylab="residuals")

abline(h=0)

#Checking Normality of residuals

hist(resid(m3))

plot(m3)