# Assignment5

## Question1

library(tidyverse)

jobs <- read\_csv('prestige.csv')

set.seed(17053777)

my\_jobs <- jobs %>%

sample\_n(95) %>%

mutate(income = income \* 10) %>%

mutate(job\_type = factor(job\_type)) %>%

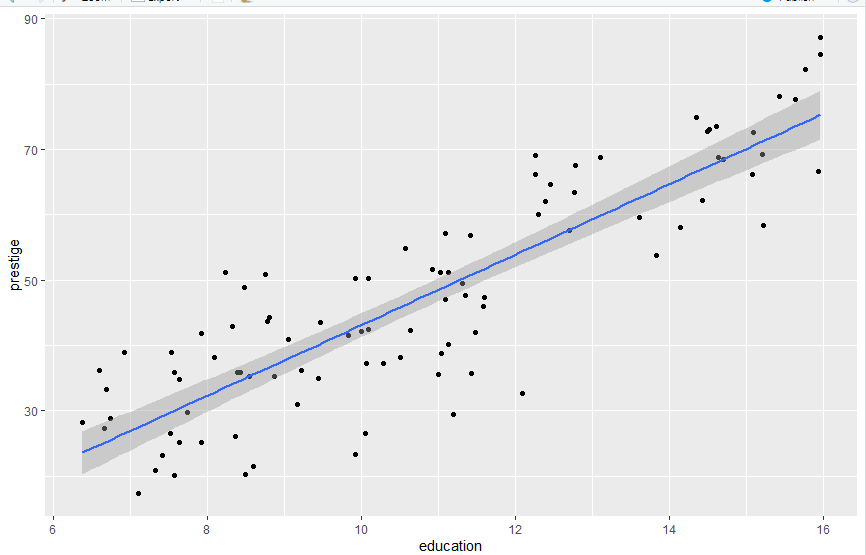
mutate(log\_income = log10(income))

## question2

library(GGally)

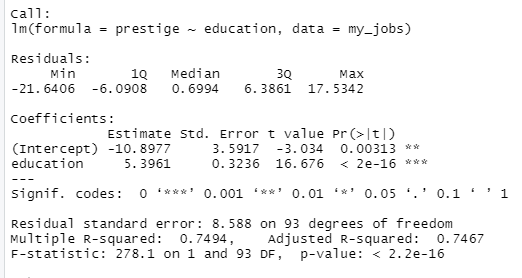
my\_jobs %>%

ggplot(aes(x=education, y=prestige)) + geom\_point() + geom\_smooth(method='lm')



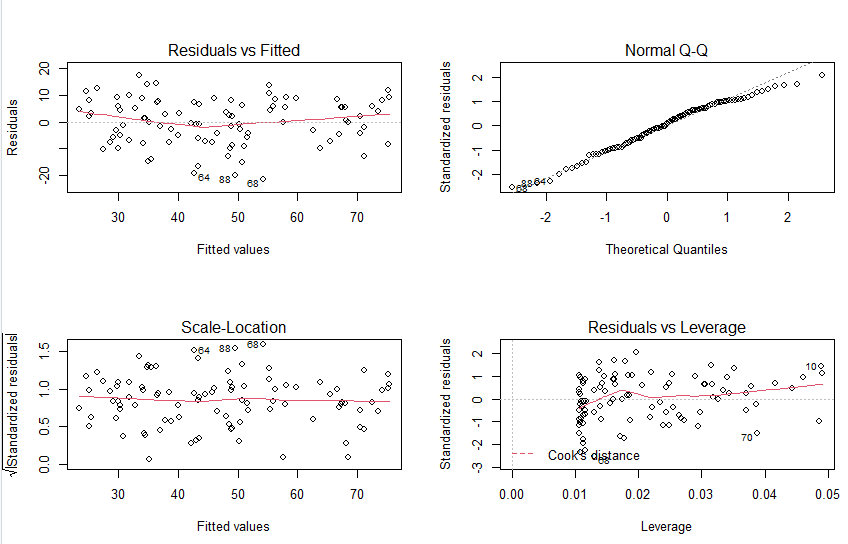
m1 <- lm(prestige ~ education, data = my\_jobs)

summary(m1)



par(mfrow = c(2, 2))

plot(m1)



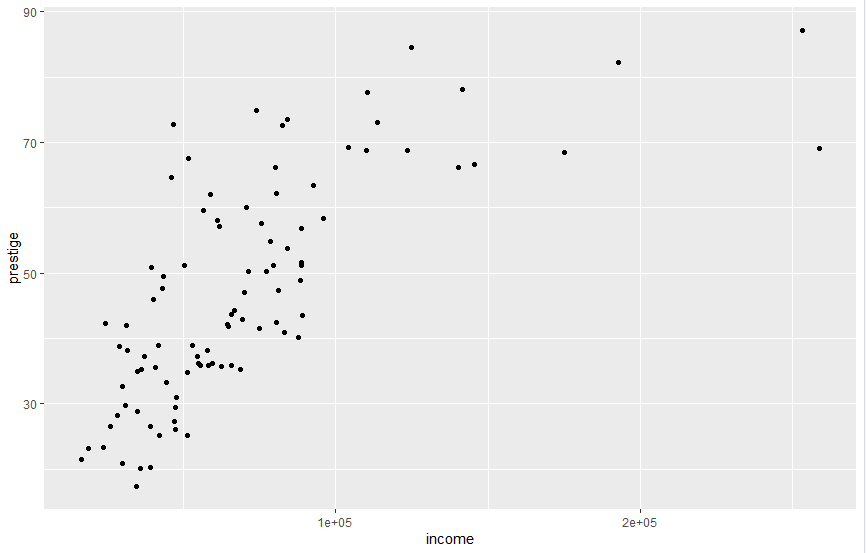
# From Residuals vs Fitted, we can see the red line is almost around the 0-dash line, so we can say there is not too much deviation from linearity.

# From Normal Q-Q. we observe that after the theoretical qualities are greater than 1, the observation points are further way from the predicted line. The residuals increase as the quality increases.

## question3

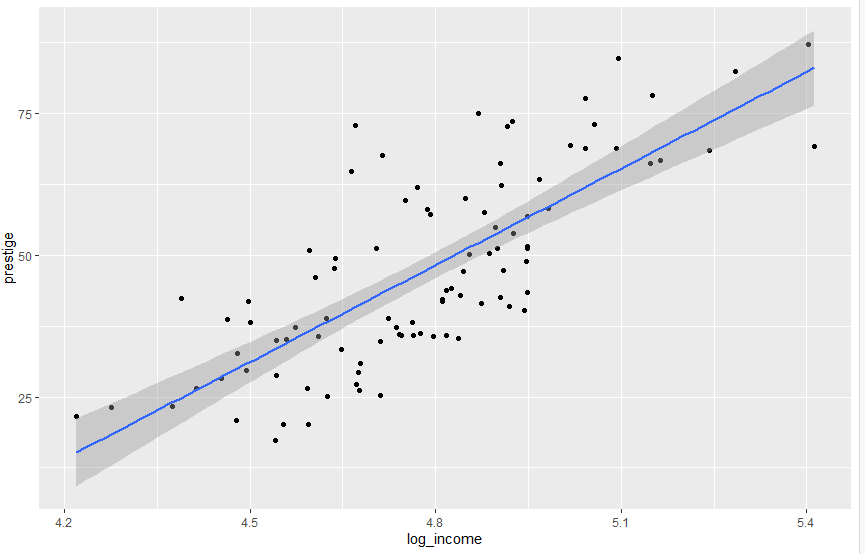
my\_jobs %>%

ggplot(aes(x = income, y = prestige)) + geom\_point()



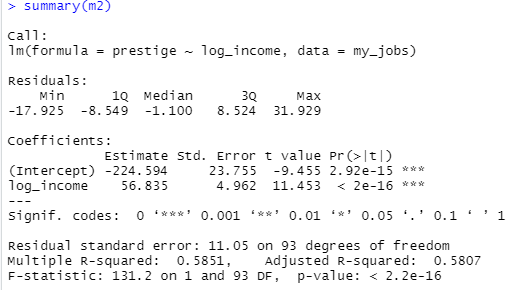
my\_jobs %>%

ggplot(aes(x = log\_income, y = prestige)) + geom\_point() + geom\_smooth(method='lm')



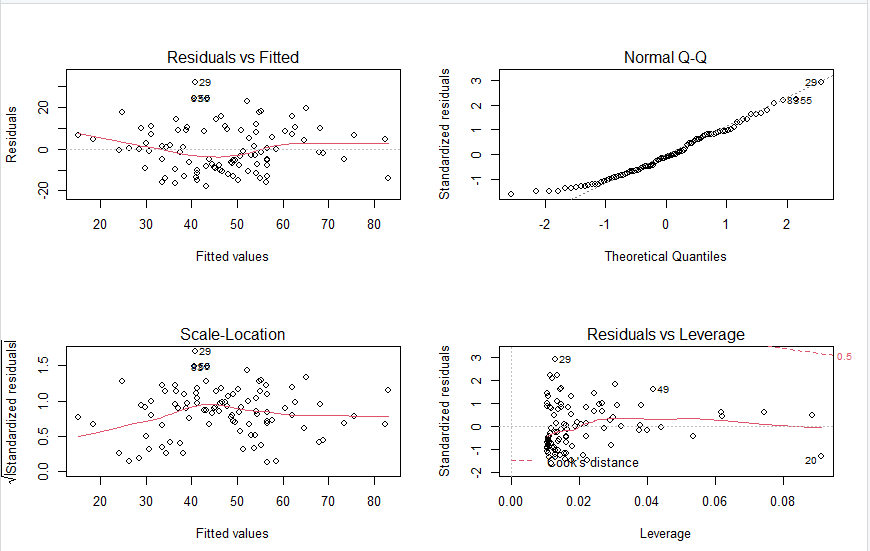
m2 <- lm(prestige ~ log\_income, data = my\_jobs)

summary(m2)



par(mfow = c(2, 2))

plot(m2)



# From Residuals vs Fitted, we can observe that the red line is almost match with the 0 das line.

# From Normal Q-Q, we can see almost all the observations match with the predicted line.

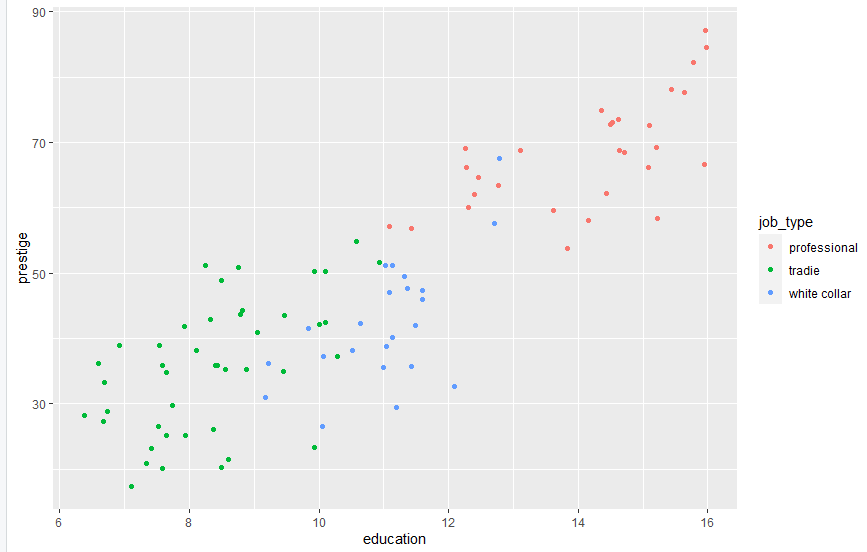
# From the scatter plot of prestige versus income, we found we cannot fit in a straight line, we can only fit the points with a curved line.

# But for the scatter plot of prestige versus log\_income, we found we can see a straight line clearly.

# so the log\_income would perform slightly better than income.

## question4

my\_jobs %>% ggplot(aes(x = education, y = prestige, colour= job\_type)) + geom\_point()

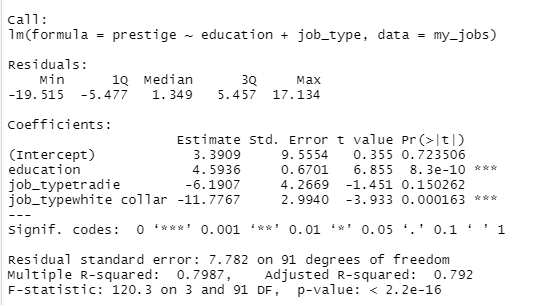


# Yes I think the regression lines will differ between seasons. For tradie, the regression line is going to be higher than the line of white collar. For professional, the regression line is going to be higher and other two and starts from 12.

## question5

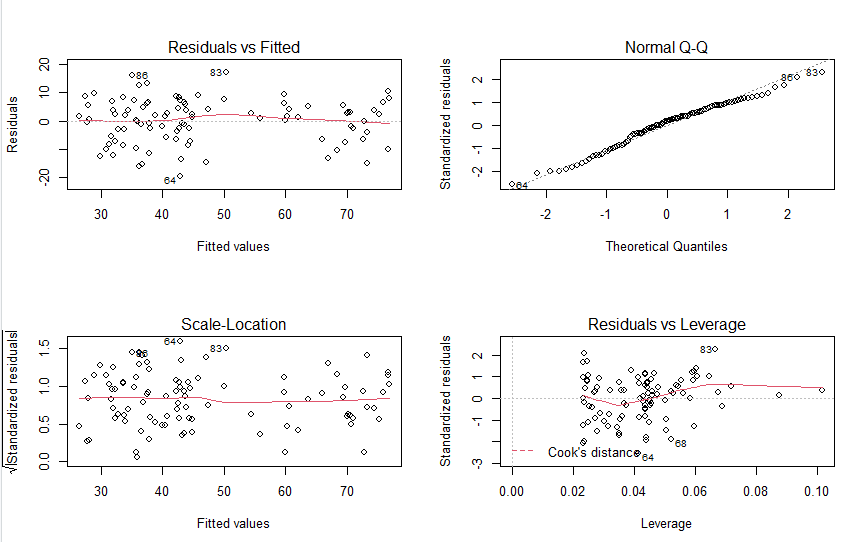
m3 <- lm(prestige~education+job\_type, data=my\_jobs)

summary(m3)



par(mfrow = c(2, 2))

plot(m3)



# The model is slightly improved from m1 to m3 since the residual standard error is decreasing from 11.05 on 93 degrees of freedom to 7.782 on 91 degrees of freedom and the multiple R^2 is increasing from 0.5851 to 0.7987 and the adjusted R^2 increased from 0.5807 to 0.792

# From Residuals vs Fitted, we can observe that the red line fits better with 0 dash line, not too much deviation. From Normal Q-Q, we can see the residuals are generally distributed. From the scale-location we can see the red line is almost a straight line comparing to the scale-location graph of m1.

## question6

tradie = (2.9859 -6.0239) + 4.6134 \* education

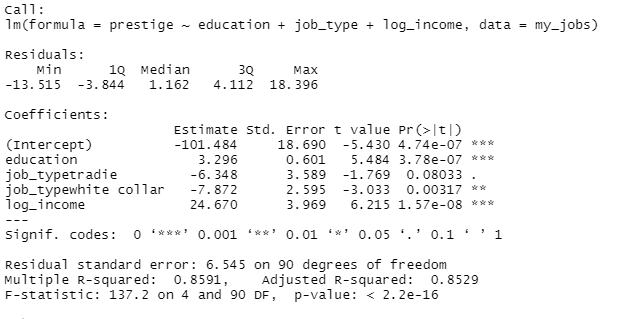
White collar = (2.9859 - 11.4750) + 4.6134 \* education

professional = 2.9859 + 4.6134 \* education

## question7

m4 <- lm(prestige~education+job\_type+log\_income, data=my\_jobs)

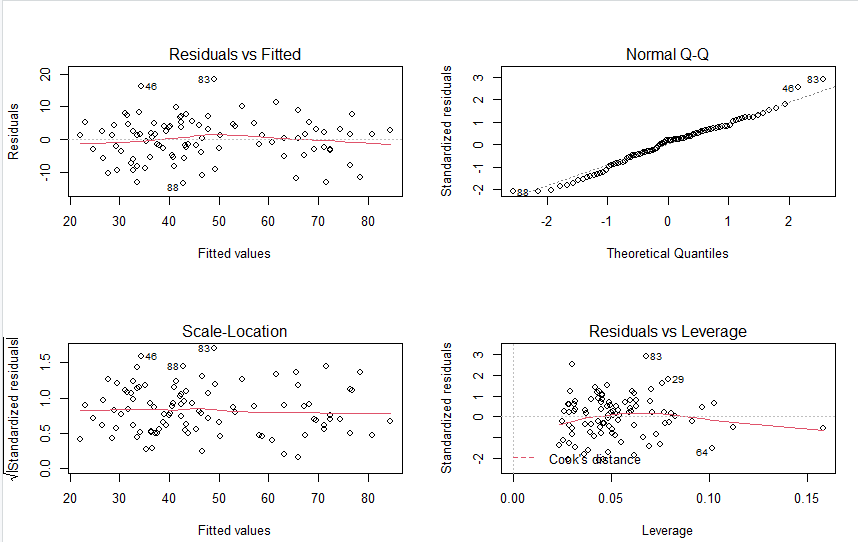
summary(m4)



# The model is slightly improved from m3 to m4 since the residual standard error is decreasing from 7.782 on 91 degrees of freedom to 6.545 on 90 degrees of freedom and the multiple R^2 is increasing from 0. 7987 to 0.8591 and the adjusted R^2 increased from 0. 792to 0.8529

par(mfrow = c(2, 2))

plot(m4)



# From the Residual v Fitted graph we can see the red line is almost a straight line the linearity is improved. From the Normal Q-Q graph we can see there is nearly no residuals since the observations are all with the predicted line