STAT-HW-5

AUTHOR
Damini Vadrevu

1 (a)

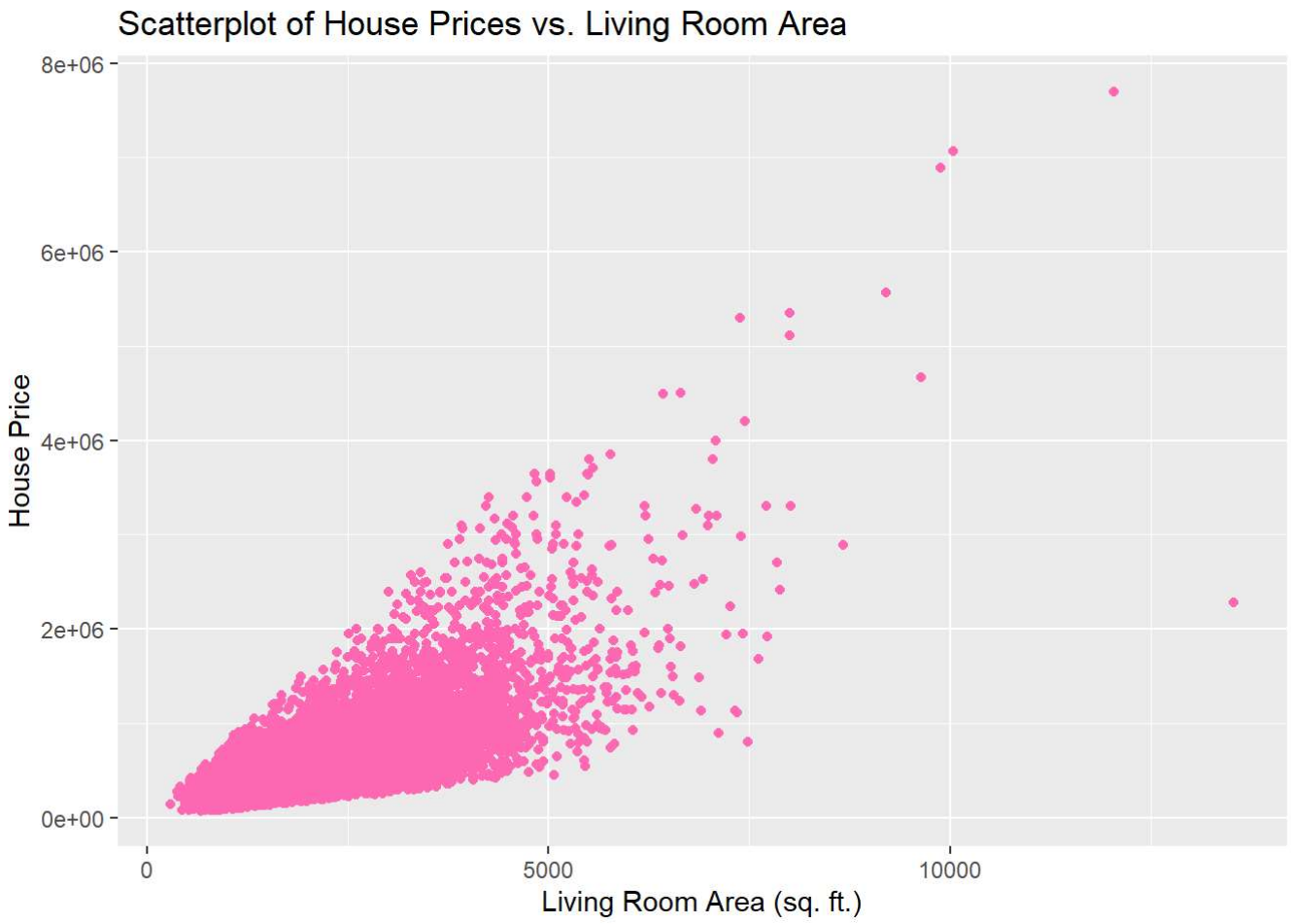
```
library(ggplot2)
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
library(car)
```

```
Loading required package: carData
```

```
data <- read.csv('C:/Users/vadre/Downloads/kchouse.csv')

# Scatterplot
ggplot(data, aes(x=sqft_living, y=price)) +
  geom_point(color="hotpink") +
  ggtitle("Scatterplot of House Prices vs. Living Room Area") +
  xlab("Living Room Area (sq. ft.)") +
  ylab("House Price")
```

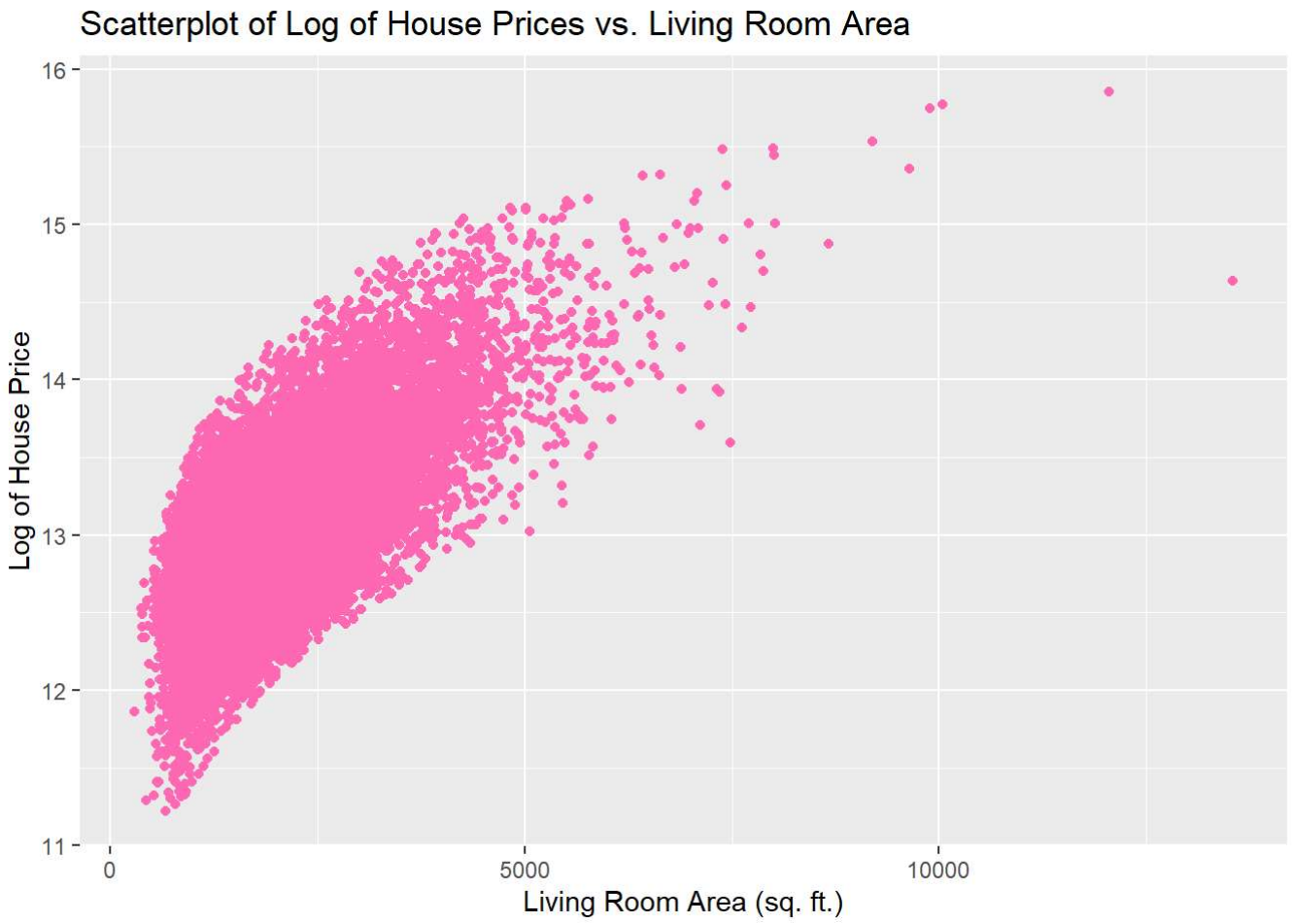


> Observations:

- There seems to be less variability in price range of the houses between 0 to 5000 square feet indicating that living area might not be the only influencing factor.
- A positive slope can be observed from the scatter plot, so as the house prices increase the living area in square feet seems to increase along.
- The dense cluster indicates that the most houses sold fall have living areas within 500 to 4000 sq.ft

1 (b)

```
data$log_price <- log(data$price)

# Scatterplot for the log transform
ggplot(data, aes(x=sqft_living, y=log_price)) +
  geom_point(color='hotpink') +
  ggtitle("Scatterplot of Log of House Prices vs. Living Room Area") +
  xlab("Living Room Area (sq. ft.)") +
  ylab("Log of House Price")
```



The data points appear to form a cloud that flows from the bottom left to the top right, especially for living room areas smaller than 5000 sq. ft. This suggests that the living room area and the log of home values have a positive linear relationship.

The size of the living room appears to correlate with the price of the home.In conclusion, the scatterplot shows a broad linear trend between the log of housing prices and living room size, notably for houses with living rooms less than 5000 square feet.

1 (c)

```
cor_coeff <- cor(data$log_price, data$sqft_living)
print(cor_coeff)
```

```
[1] 0.6953406
```

The correlation coefficient shows a significantly positive relationship between the log transformed house price and living area in sq. ft.

1 (d)

```
model <- lm(log_price ~ sqft_living, data=data)
summary(model)
```

```
Call:
lm(formula = log_price ~ sqft_living, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.97781 -0.28543  0.01472  0.26070  1.27628

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.222e+01  6.374e-03  1916.9   <2e-16 ***
sqft_living 3.987e-04  2.803e-06   142.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3785 on 21611 degrees of freedom
Multiple R-squared:  0.4835,    Adjusted R-squared:  0.4835
F-statistic: 2.023e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

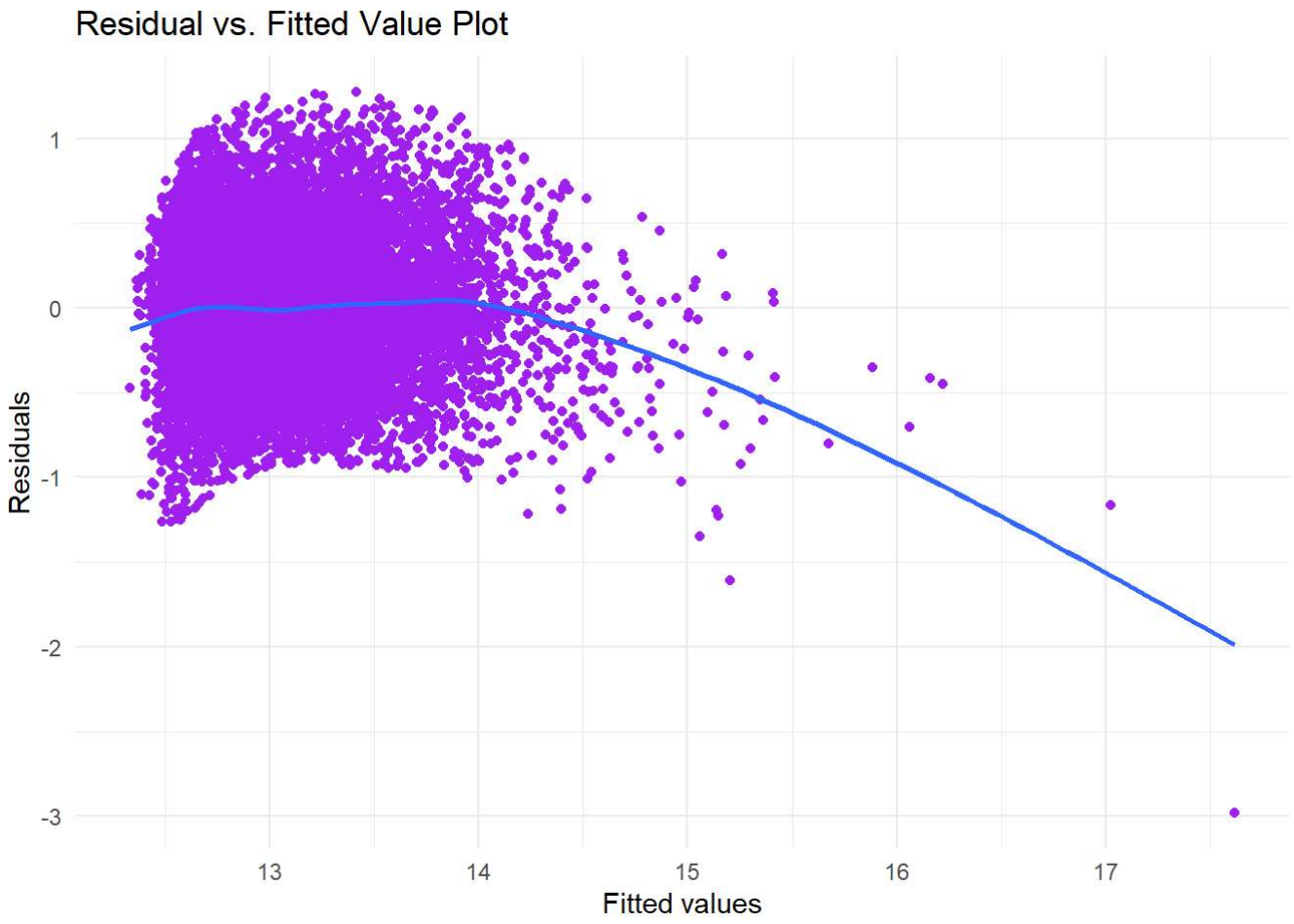
The logarithm of the house price is anticipated to rise by roughly 0.0003987 for every additional square foot in the living room area. The size of the living room accounts for about 48.35% of the variation in house price logarithms. The p-values of the model coefficients show that they are statistically significant.

The p-values for the sqft_living intercept and coefficient are less than 0.05 indicating that there is substantial evidence that the size of the living room has an impact on the dataset's logarithm of the price of homes.

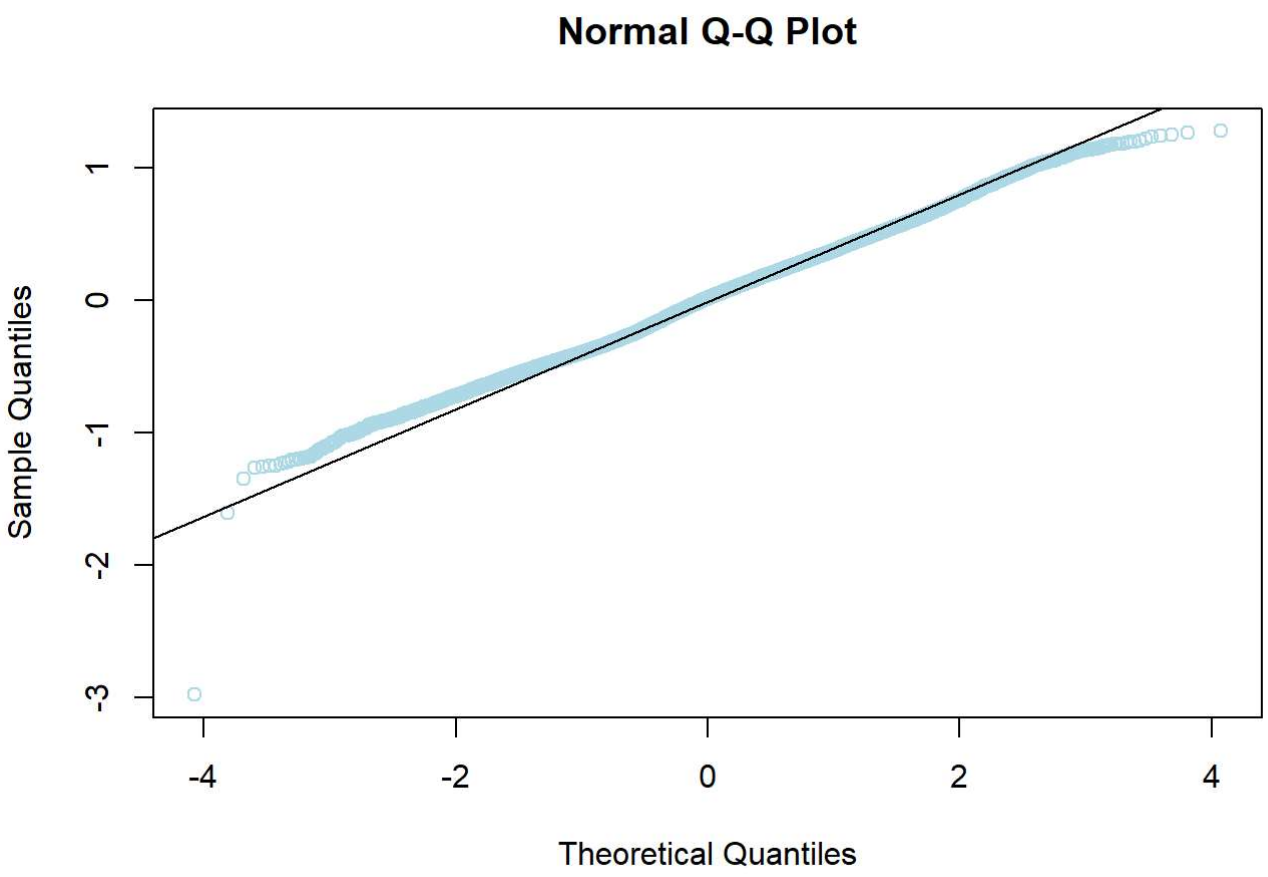
1 (e)

```
# Residual vs. Fitted Value Plot for checking homoscedasticity
ggplot(data, aes(x=fitted(model), y=residuals(model))) +
  geom_point(color='purple') +
  geom_smooth(se=FALSE) +
  theme_minimal() +
  ggtitle("Residual vs. Fitted Value Plot") +
  xlab("Fitted values") + ylab("Residuals")
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
# Q-Q Plot for checking normality
qqnorm(residuals(model), col = "lightblue")
qqline(residuals(model))
```

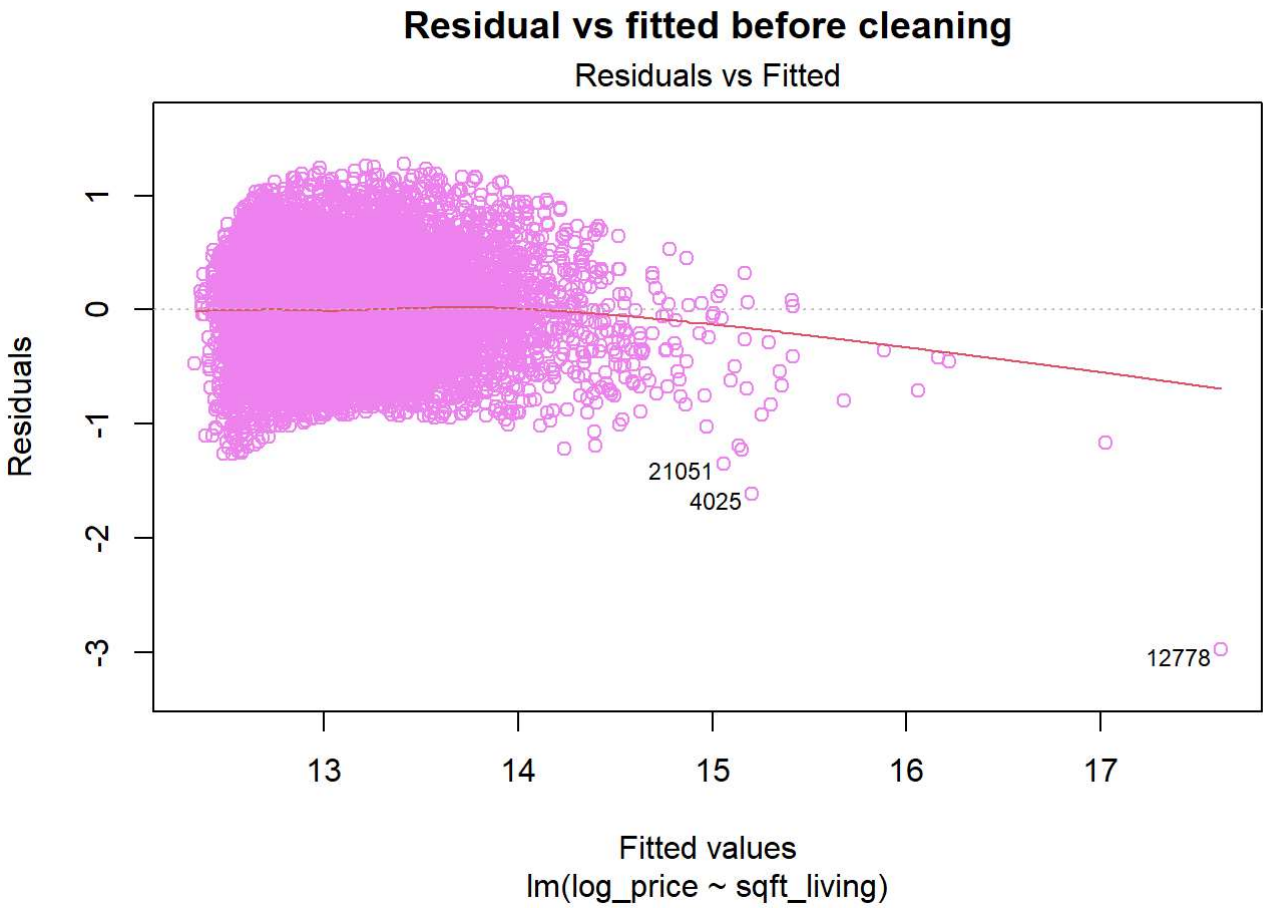


The qq plot shows that the residuals follow well along the reference line indicating normality.

Residual vs fitted plot : A funnel-like pattern may be seen, especially on the right side. This means that not all levels of fitted values will have the same residual variance. The antithesis of homoscedasticity, this pattern suggests probable heteroscedasticity.

1 (f)

```
# Diagnostic plot to identify outliers
plot(model, which=1, col="violet", main = "Residual vs fitted before cleaning")
```

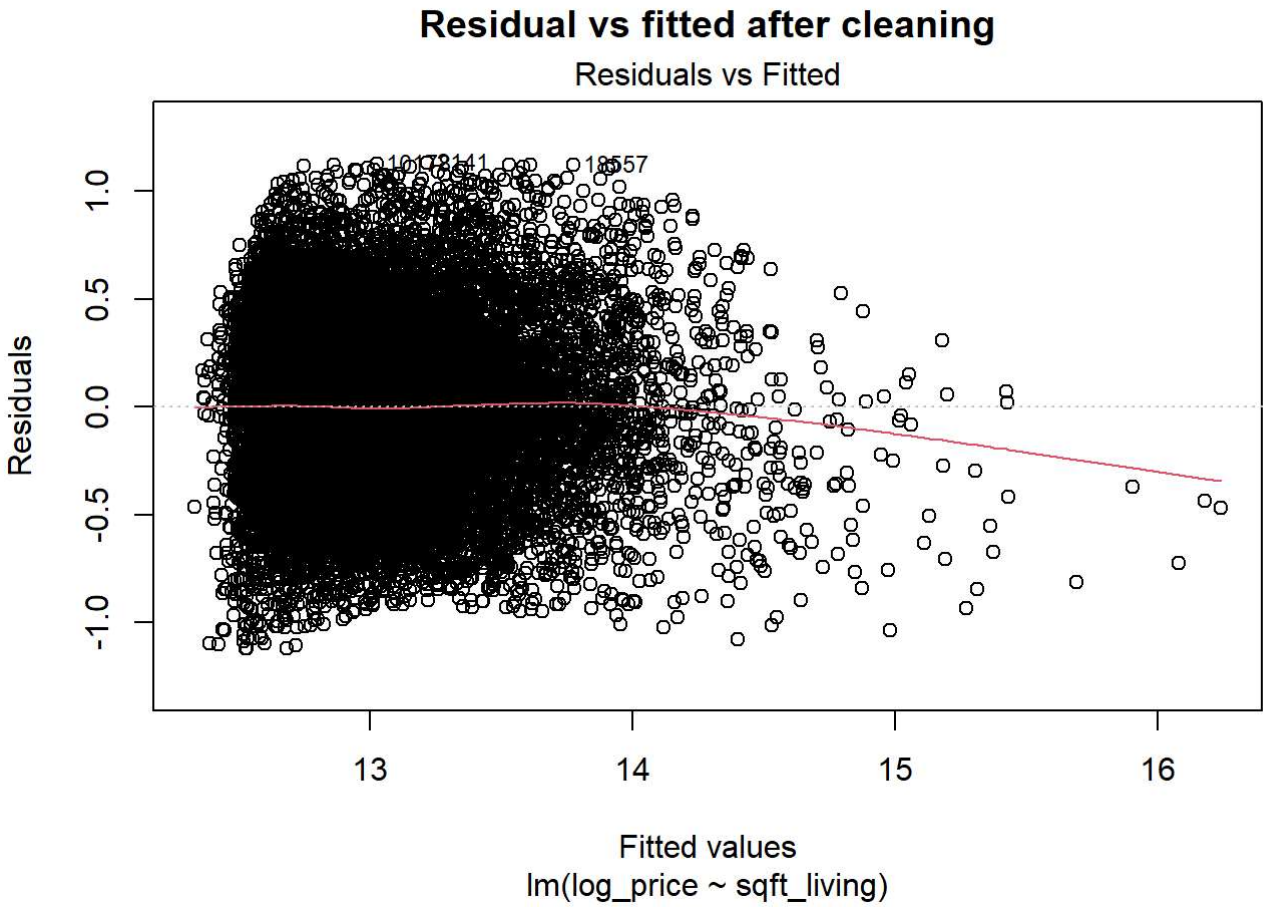



```r
std_residuals <- rstandard(model)
outliers <- which(abs(std_residuals) > 3)

data_clean <- data[-outliers, ]
model_refit <- lm(log_price ~ sqft_living, data=data_clean)
summary(model_refit)
```

```
Call:
lm(formula = log_price ~ sqft_living, data = data_clean)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.12266 -0.28462  0.01431  0.26038  1.13015 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.221e+01  6.351e-03  1923.0   <2e-16 ***
sqft_living 4.010e-04  2.800e-06   143.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3739 on 21559 degrees of freedom
```

```
Multiple R-squared:  0.4875,    Adjusted R-squared:  0.4875
F-statistic: 2.051e+04 on 1 and 21559 DF,  p-value: < 2.2e-16
```

```
# Diagnostic plot to check residuals
plot(model_refit, which=1, main = "Residual vs fitted after cleaning")
```




```
# Compare coefficients of both models
summary(model)$coefficients
```

```
               Estimate   Std. Error   t value Pr(>|t|)
(Intercept) 1.221846e+01 6.374131e-03 1916.8830        0
sqft_living 3.987465e-04 2.803481e-06  142.2326        0
```

```
summary(model_refit)$coefficients
```

```
               Estimate   Std. Error  t value Pr(>|t|)
(Intercept) 1.221341e+01 6.351128e-03 1923.030        0
sqft_living 4.010336e-04 2.800495e-06  143.201        0
```

After cleaning the data by removing outliers and refitting the model, there is still some slight pattern, but it's less pronounced than in the initial plot. indicating that the outliers do not influence the model to some degree.

1 (g)

> ***Intercept:***

Null Hypothesis $H_0 : \beta_0 = 0$ The true intercept is 0.

Alternative Hypothesis $H_1 : \beta_0 \neq 0$ The true intercept is not 0

The t-value for the intercept is 1916.9, which is significantly high. The p-value is given as <2e-16, which is practically zero. This provides strong evidence against the null hypothesis. Thus, we reject the null hypothesis and conclude that the intercept is significantly different from zero..

> *Slope:*

Null Hypothesis: $H_0 : \beta_1 = 0$ The log price does not change with the living area in sq ft.

Alternative Hypothesis $H_1 : \beta_1 \neq 0$ The log price changes with a change in living area in sq. ft.

The t-value for sqft_living is 142.2, and the p-value is again <2e-16, practically zero. This indicates that there's strong evidence against the null hypothesis for the slope as well. We reject the null hypothesis for the slope and conclude that the true slope of sqft_living in predicting log_price is significantly different from zero.

```
model <- lm(log_price ~ sqft_living, data = data)
confint(model)
```

```
                  2.5 %       97.5 %
(Intercept) 1.220597e+01 1.223096e+01
sqft_living 3.932515e-04 4.042416e-04
```

Interpretation:

- This interval suggests that we are 95% confident that the true intercept (when sqft_living is zero) lies between approximately 12.20597 and 12.23096 for the log_price. Given that this interval does not contain zero, it reaffirms our earlier conclusion that the intercept is significantly different from zero.
- The confidence interval for the slope doesn't include 0, depicting that there is a significant relationship between the living room area and the log price. Specifically, for each additional square foot of living area, the log price is expected to increase by an amount between 0.0003932515 and 0.0004042416, holding all else constant.

1 (h)

```
new_data <- data.frame(sqft_living = 1500)
predicted_log_price <- predict(model, newdata = new_data, interval = "prediction", level = 0.95)
```

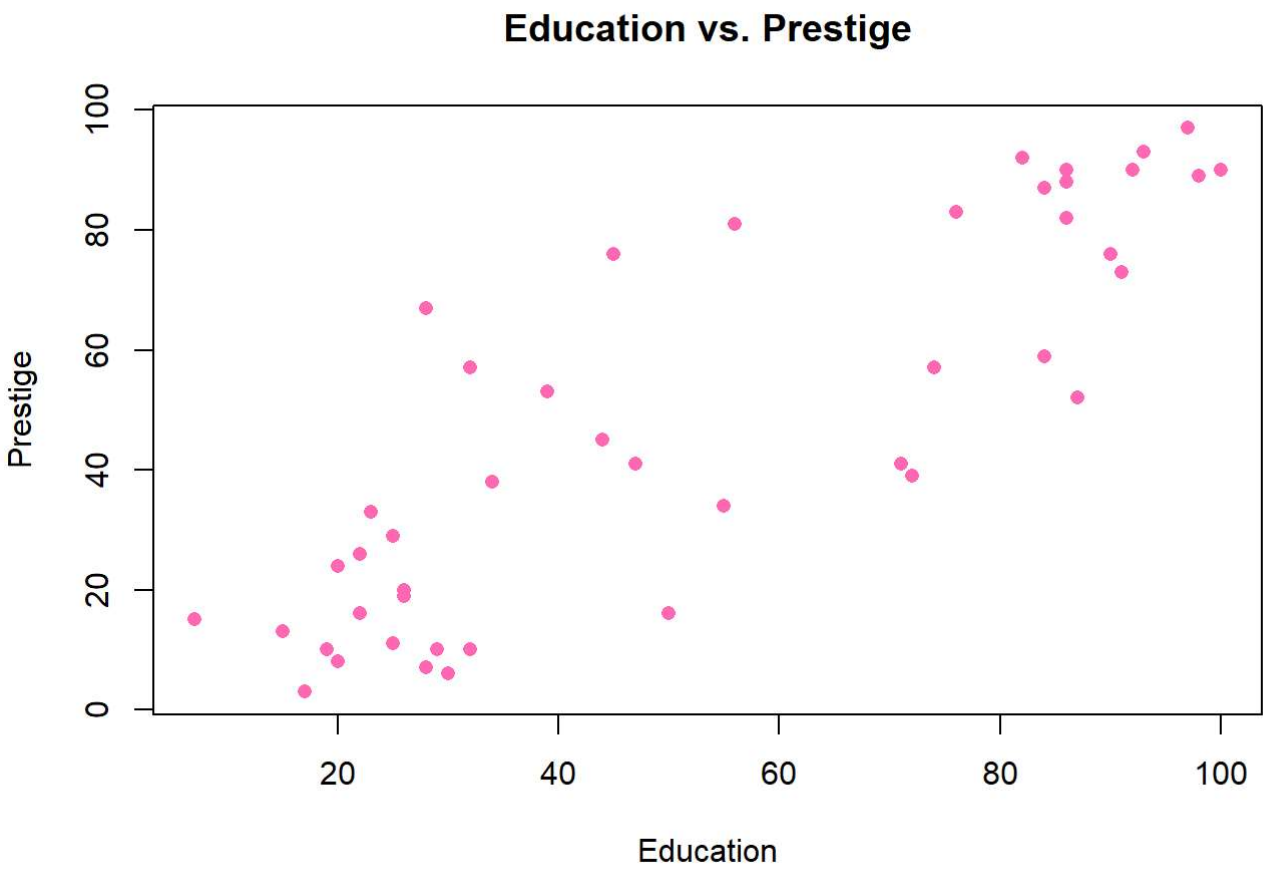
```
expected_price <- exp(predicted_log_price)
expected_price
```

```
      fit      lwr      upr
1 368274.5 175364.1 773397.1
```

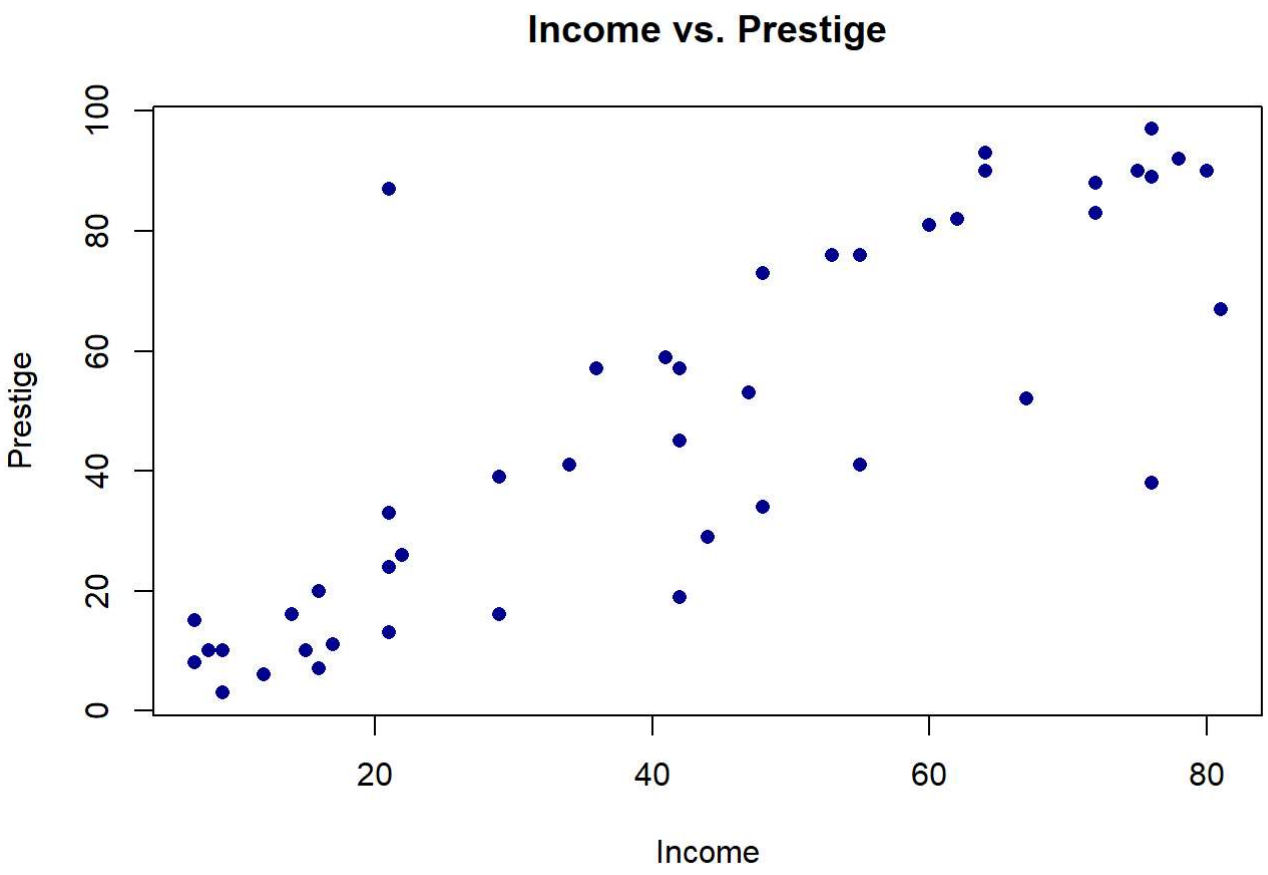
- The predicted price for a house with 1500 square feet of living area is approximately $368,274.5.
- You can be 95% confident that the price of a house with 1500 square feet of living area will fall between $175,364.1 and $773,397.1.

2 (a)

```
library(carData)
data(Duncan)
plot(Duncan$education, Duncan$prestige, main="Education vs. Prestige", xlab="Education", ylab="Pr
```




```
plot(Duncan$income, Duncan$prestige, main="Income vs. Prestige", xlab="Income", ylab="Prestige",
```

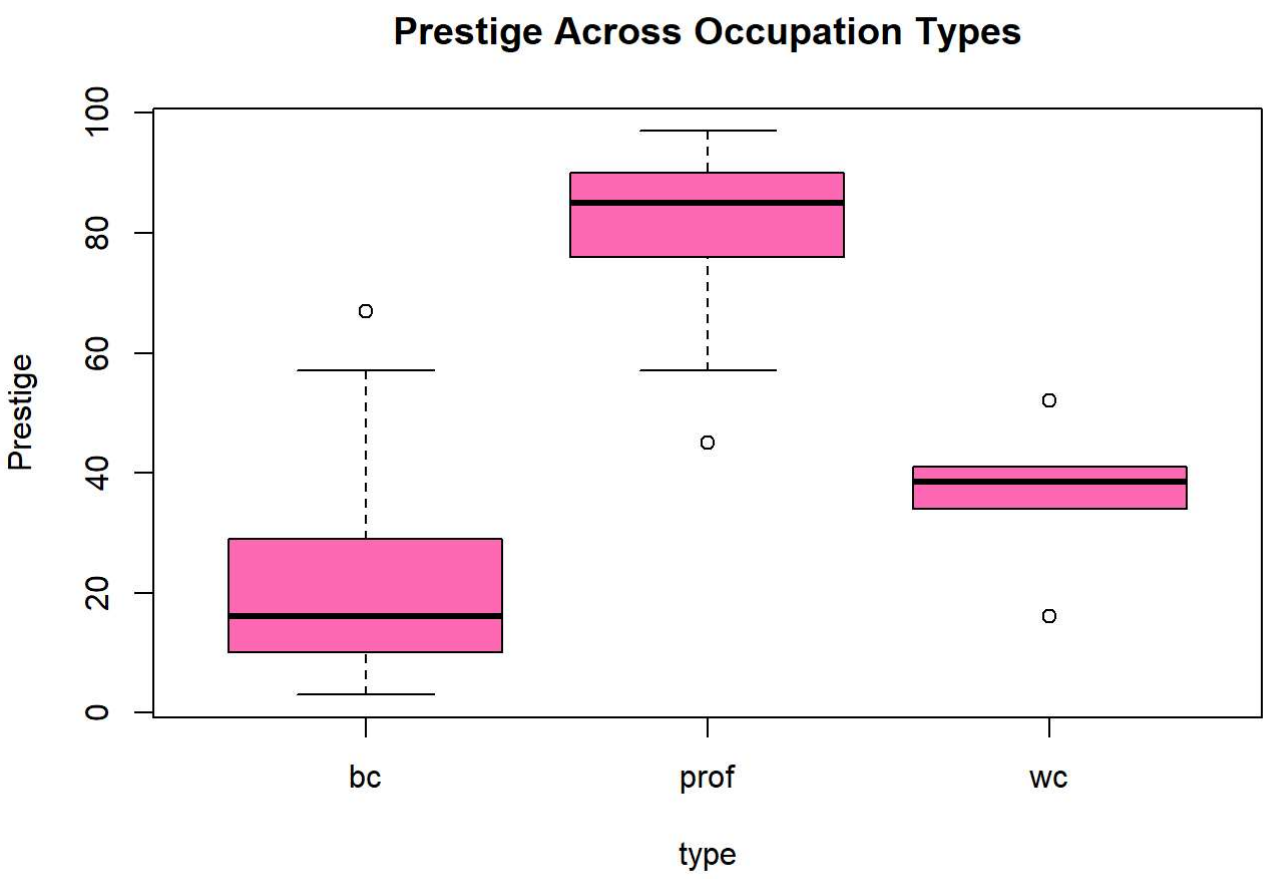


Education vs. Prestige: The scatterplot shows a generally upward trend indicating that as education increases, prestige tends to increase as well. However, the relationship doesn't seem to be perfectly linear. For lower values of education, prestige seems to be more clustered in the lower region. As education values increase, the spread of prestige values also increases. The relationship appears to be somewhat curvilinear, perhaps logarithmic or quadratic.

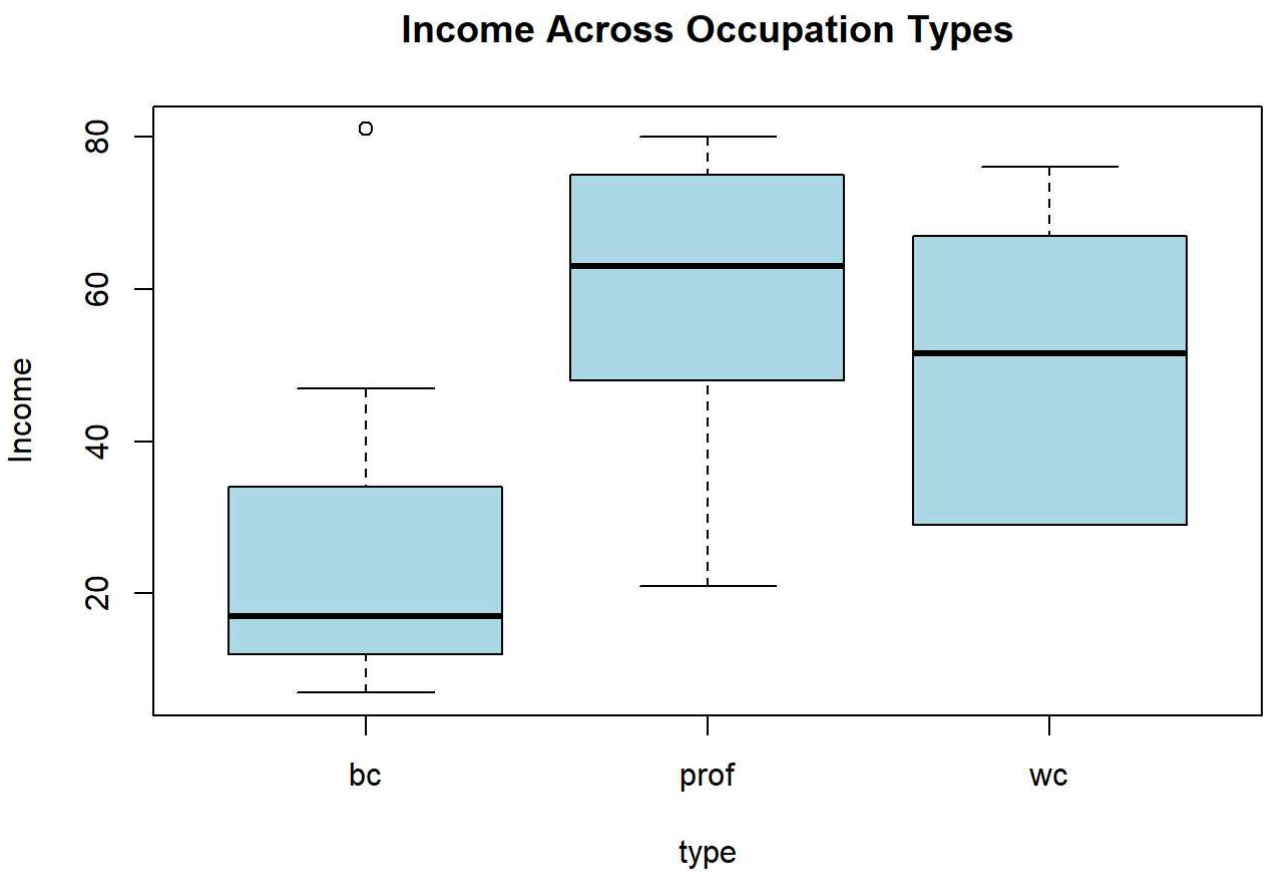
Income vs. Prestige: The relationship between income and prestige seems more linear than the previous plot. There's a clear upward trend: as income increases, prestige tends to increase. However, there's some variability, especially in the mid-range of income values, where the prestige values spread out more. But overall, this relationship is more linear than the one between education and prestige.

(b)

```
boxplot(prestige ~ type, data=Duncan, main="Prestige Across Occupation Types", ylab="Prestige", c
```




```
boxplot(income ~ type, data=Duncan, main="Income Across Occupation Types", ylab="Income",col ="li
```




```
library(carData)
aggregate(Duncan$prestige, by=list(Type=Duncan$type), FUN=mean)
```

```
  Type        x
1   bc 22.76190
2 prof 80.44444
3   wc 36.66667
```

```
aggregate(Duncan$income, by=list(Type=Duncan$type), FUN=mean)
```

```
  Type        x
1   bc 23.76190
2 prof 60.05556
3   wc 50.66667
```

- More Prestige: prof or Professional occupations have the highest prestige.
- More Income: Again, prof or Professional occupations also command the highest income on average.
- Both the prestige and income boxplots show that professionals enjoy the most benefits, followed by white-collar jobs. Blue-collar jobs, on average, have the lowest prestige and income. However, it's also worth noting that there are outliers in both prestige and income across occupation types, indicating that there are exceptions to these general trends.

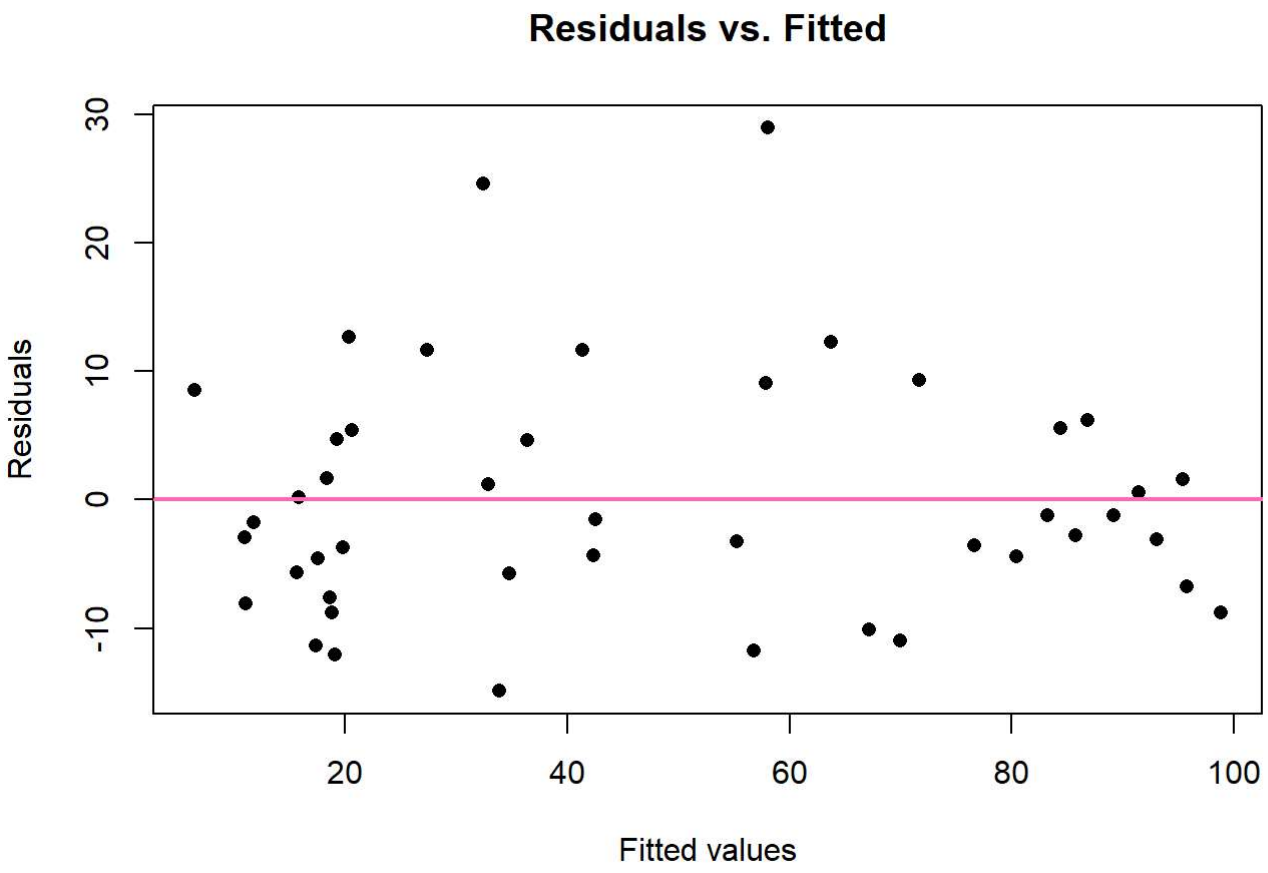
2 (c)

Assumptions to be met before fitting a regression model:

- Linearity
- Homoscedasticity
- Normality
- Multicollinearity

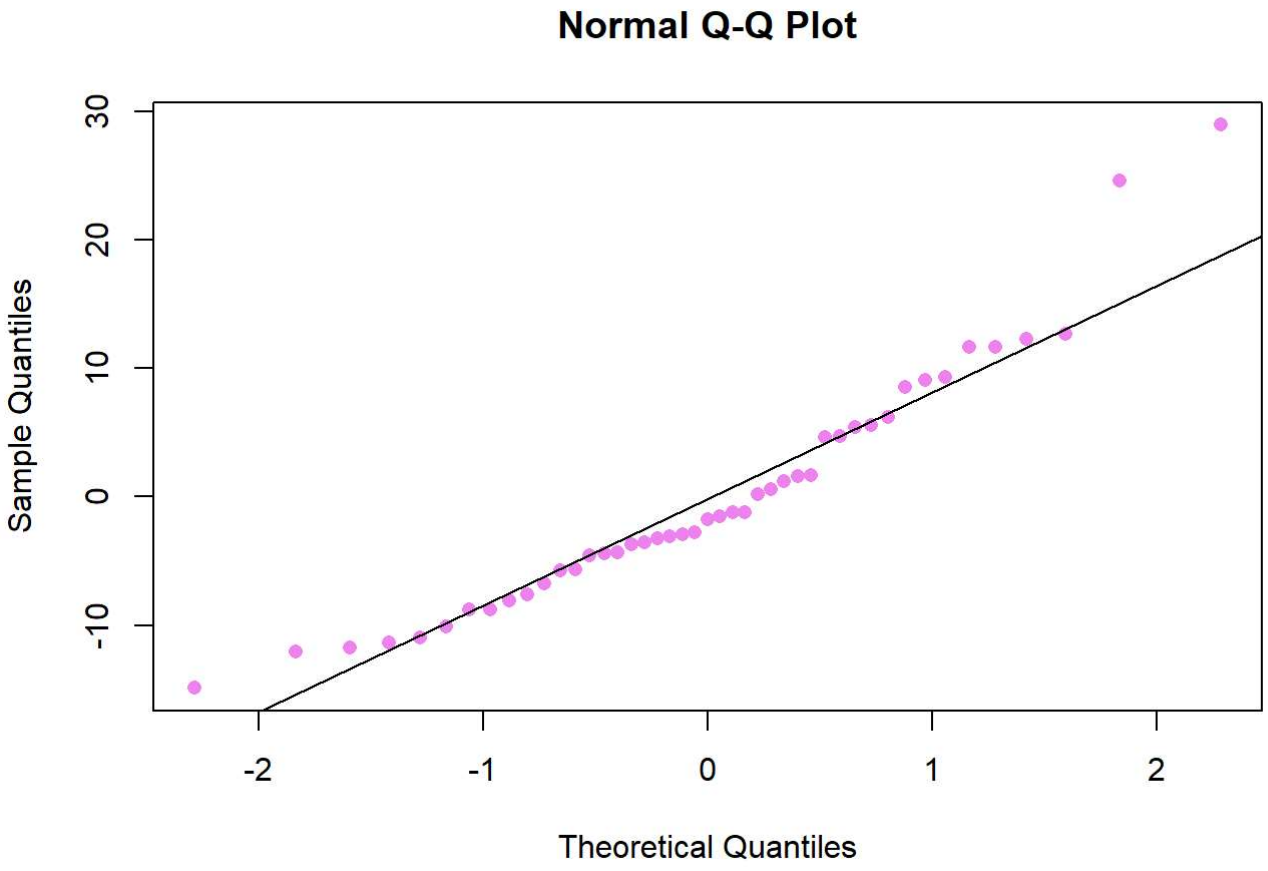
Linearity has been satisfied from the above scatterplots in 2 a

```
mod <- lm(prestige ~ education + income + type, data=Duncan)
plot(mod$fitted.values, mod$residuals, main="Residuals vs. Fitted", xlab="Fitted values", ylab="R
abline(h=0, col="hotpink", lwd=2)
```




```
#normality
```

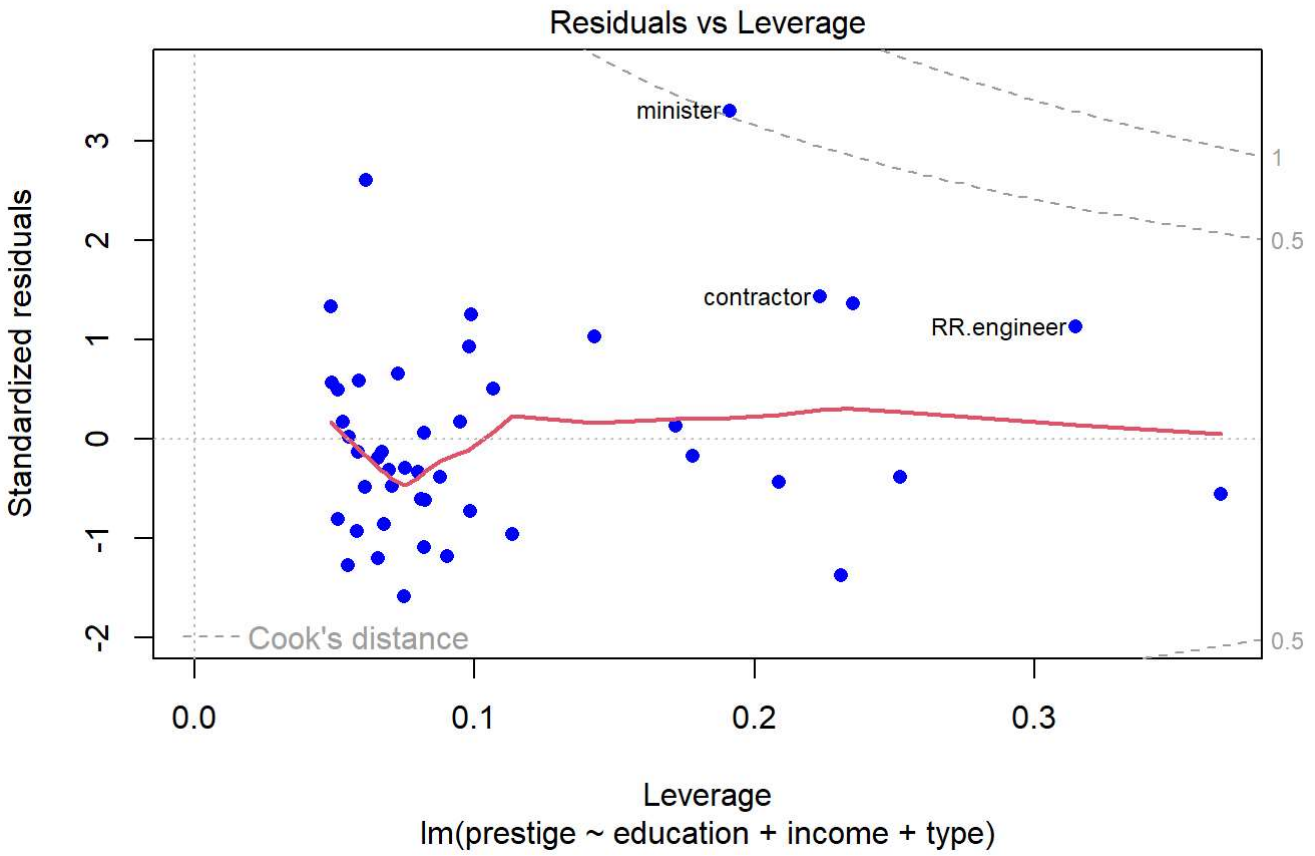
```
qqnorm(residuals(mod), col = "violet", pch=16)
qqline(residuals(mod))
```



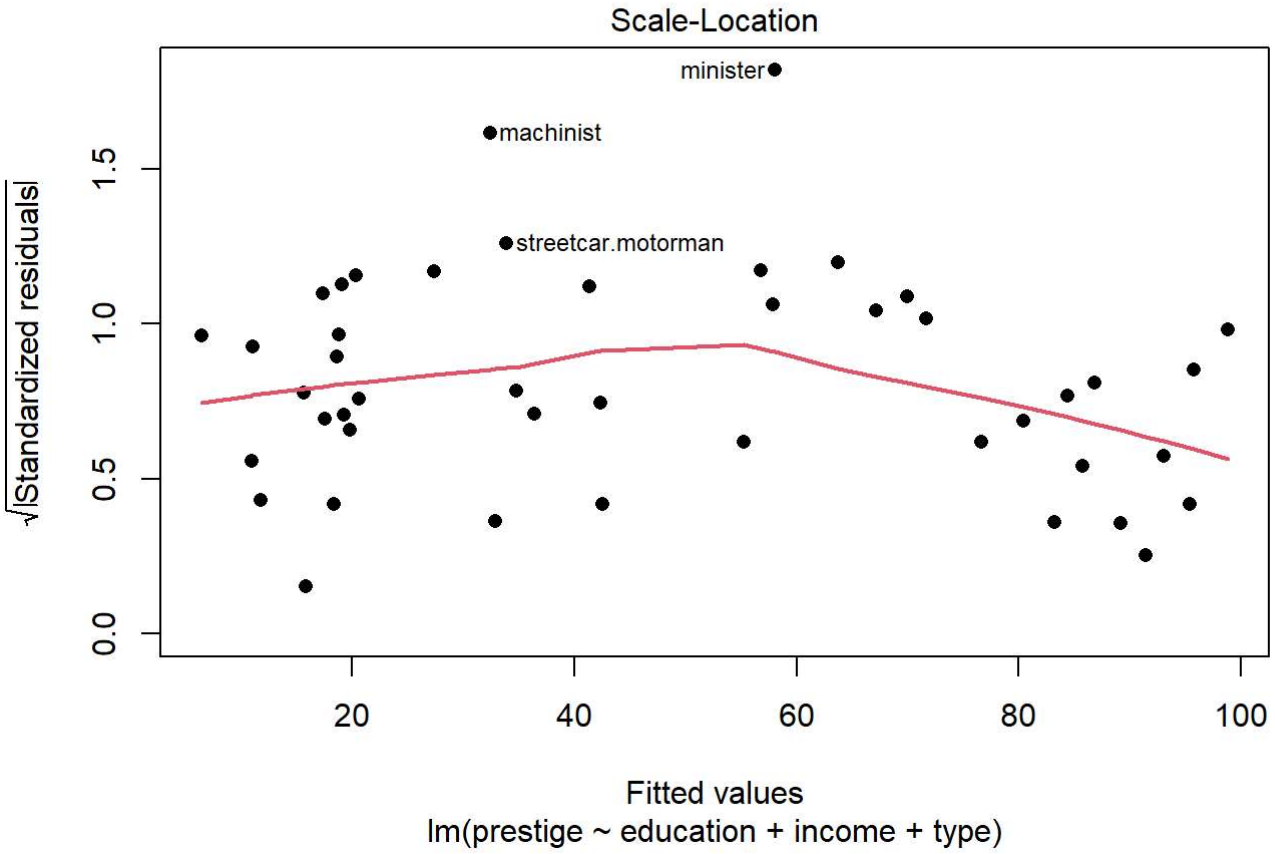

```
#Independence
library(car)
durbinWatsonTest(mod)
```

```
lag Autocorrelation D-W Statistic p-value
  1       0.2468701      1.497477    0.03
Alternative hypothesis: rho != 0
```

```
#residuals vs leverage plot
plot(mod, which=5, lwd=2, pch =16, col="blue")
```




```
#scale location plot
plot(mod, which=3, lwd=2, pch=16, col = "black")
```



Although the p-value of 0.282 implies that there is no statistically significant autocorrelation, the Durbin-Watson statistic value of 1.750404 is somewhat close to 2, suggesting that there may be a very little positive autocorrelation. Accordingly, based on the results of this test, there isn't much proof that autocorrelation exists in the residuals at lag 1.

"coal.miner", "minister", and "store.manager" are labeled and seem to be of interest. Among them, **"coal.miner" has high leverage** and is outside the Cook's distance, making it particularly influential. **"minister" and "store.manager" have a substantial residual but less leverage,** suggesting they might not be as influential as "coal.miner", but they still need consideration.

Normality : The assumption of normality has been met as the points follow along the reference line.

```
#fitting regression model
mod <- lm(prestige ~ education + income + type, data=Duncan)
summary(mod)
```

```
Call:
lm(formula = prestige ~ education + income + type, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-14.890  -5.740  -1.754   5.442  28.972
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.18503    3.71377  -0.050  0.96051
education     0.34532    0.11361   3.040  0.00416 **
income        0.59755    0.08936   6.687 5.12e-08 ***
typeprof     16.65751    6.99301   2.382  0.02206 *
typewc      -14.66113    6.10877  -2.400  0.02114 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.744 on 40 degrees of freedom
Multiple R-squared:  0.9131,    Adjusted R-squared:  0.9044
F-statistic:   105 on 4 and 40 DF,  p-value: < 2.2e-16
```

Null Hypothesis (H0): All regression coefficients (excluding the intercept) are equal to zero. This implies that none of the predictors have any effect on the response variable.

Alternative Hypothesis (H1): At least one regression coefficient is different from zero. This means that at least one of the predictors has a significant effect on the response variable.

2 (d)

```r
# Fit a model with interactions between education and income, education and type, and income and
model_interaction <- lm(prestige ~ education * income * type, data = Duncan)
summary(model_interaction)
```

```
Call:
lm(formula = prestige ~ education * income * type, data = Duncan)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6819  -4.6186  -0.4928   3.1740  22.3316

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               17.89437   12.36638   1.447   0.1573
education                 -0.55965    0.50082  -1.117   0.2719
income                    -0.29036    0.53576  -0.542   0.5915
typeprof                 -85.48040   58.97604  -1.449   0.1567
typewc                   -86.67731   58.71481  -1.476   0.1494
education:income           0.03918    0.01903   2.058   0.0475 *
education:typeprof         2.03495    0.84976   2.395   0.0225 *
education:typewc           1.97504    1.07925   1.830   0.0763 .
income:typeprof            2.49297    1.19344   2.089   0.0445 *
income:typewc              1.59260    1.00829   1.580   0.1238
education:income:typeprof -0.06008    0.02269  -2.648   0.0123 *
education:income:typewc   -0.05461    0.02395  -2.280   0.0292 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.015 on 33 degrees of freedom
Multiple R-squared:  0.9386,    Adjusted R-squared:  0.9182
F-statistic: 45.87 on 11 and 33 DF,  p-value: < 2.2e-16
```

Two-way Interactions

- The statistically significant positive coefficient for the education:income relationship shows that the combined impact of income and education on prestige exceeds the sum of each factor's individual effects.
- The relationship between education and prestige appears to be stronger for professionals and white-collar jobs than for other job types, according to the positive coefficients for education:typeprof and education:typewc.
- The positive coefficient for income:typeprof implies that, in comparison to other job kinds with the same income level, professionals have a bigger increase in prestige.

Three-way Interactions:

Although the combined effect of education and income on prestige is generally favorable, it appears to be less pronounced for professionals and white-collar employees, as indicated by the negative coefficients for all three-way interactions.

2 (e)

- Linearity: The near-horizontal line in this plot suggests that the residuals do not exhibit any systematic patterns against the fitted values. This is supportive of the assumption that the relationship is linear.
- Normality: In the provided qq plot, the data points closely follow the reference line without any pronounced deviations. This indicates that the residuals are approximately normally distributed, fulfilling the normality assumption.
- Homoscedasticity: The "Residuals vs. Fitted" plot does not show any discernible pattern, and the points are dispersed randomly, which supports the homoscedasticity assumption. However, the non-constant spread observed in the "Scale-Location" plot suggests the presence of some heteroscedasticity. It is imperative to consider this when making inferences based on the model.
- Influence of Points: While there are a few high-leverage observations in the "Residuals vs. Leverage" plot, they do not appear to have a significant influence on the regression outcome. High-leverage points can unduly influence the model's results, but in this case, they seem not to exert a substantial impact.

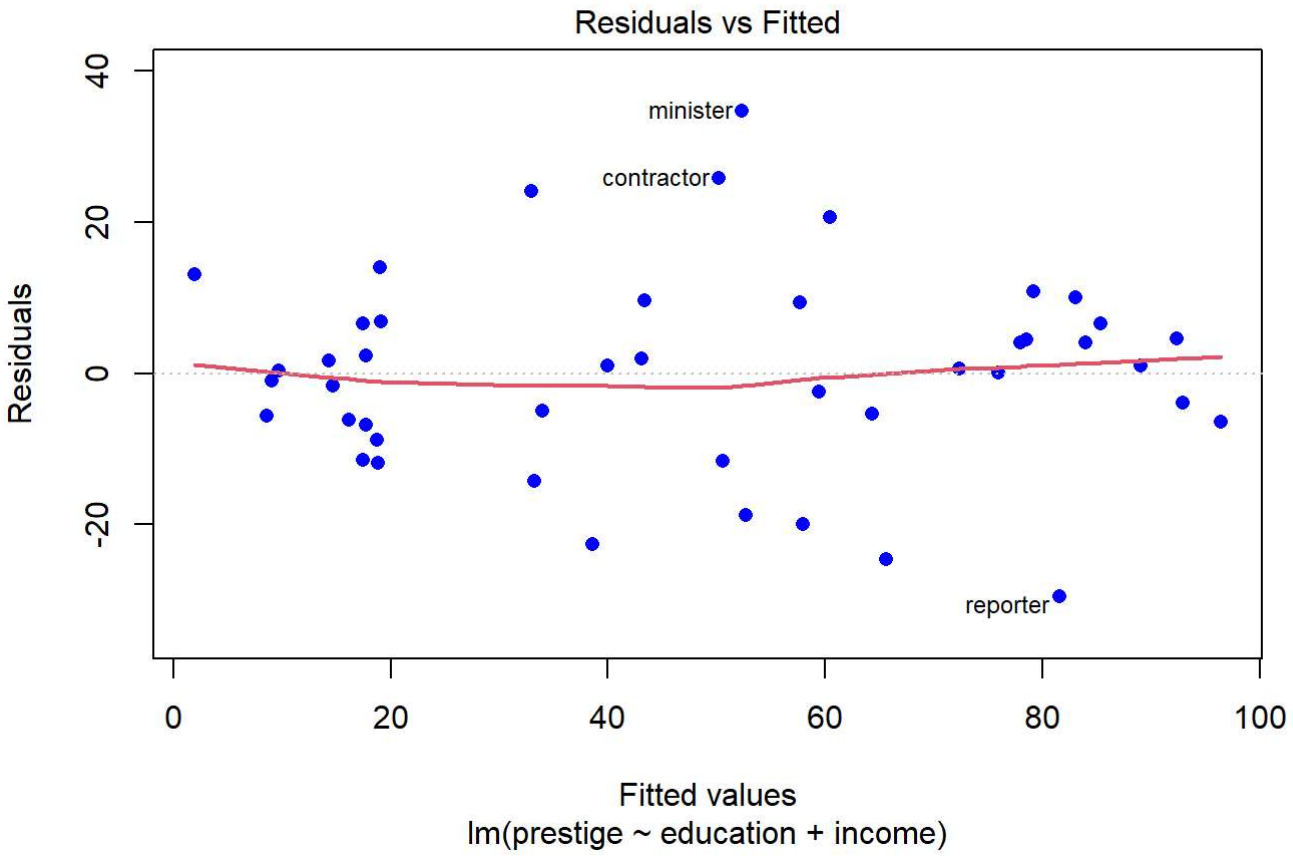
Conclusion: Taking all these diagnostics into account, the model demonstrates a commendable fit for the data. It adequately captures the linear relationship and satisfies key regression assumptions, making it a robust and reliable tool for interpreting linear relationships in the dataset.

2 (f)

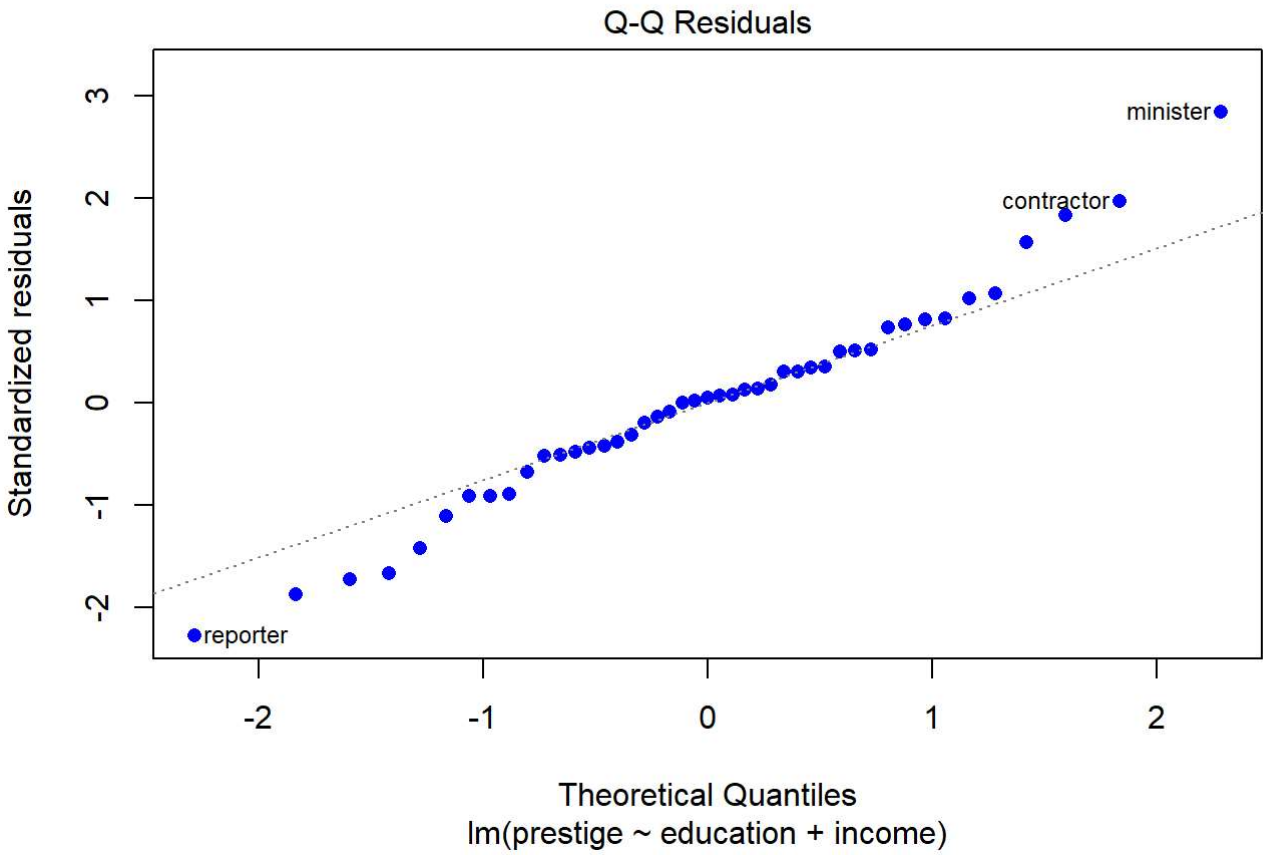
```r
mod_new <- lm(prestige ~ education + income, data = Duncan)
summary(mod_new)
```

```
Call:
lm(formula = prestige ~ education + income, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.538  -6.417   0.655   6.605  34.641 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.06466    4.27194  -1.420    0.163    
education    0.54583    0.09825   5.555 1.73e-06 ***
income       0.59873    0.11967   5.003 1.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared:  0.8282,    Adjusted R-squared:   0.82 
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

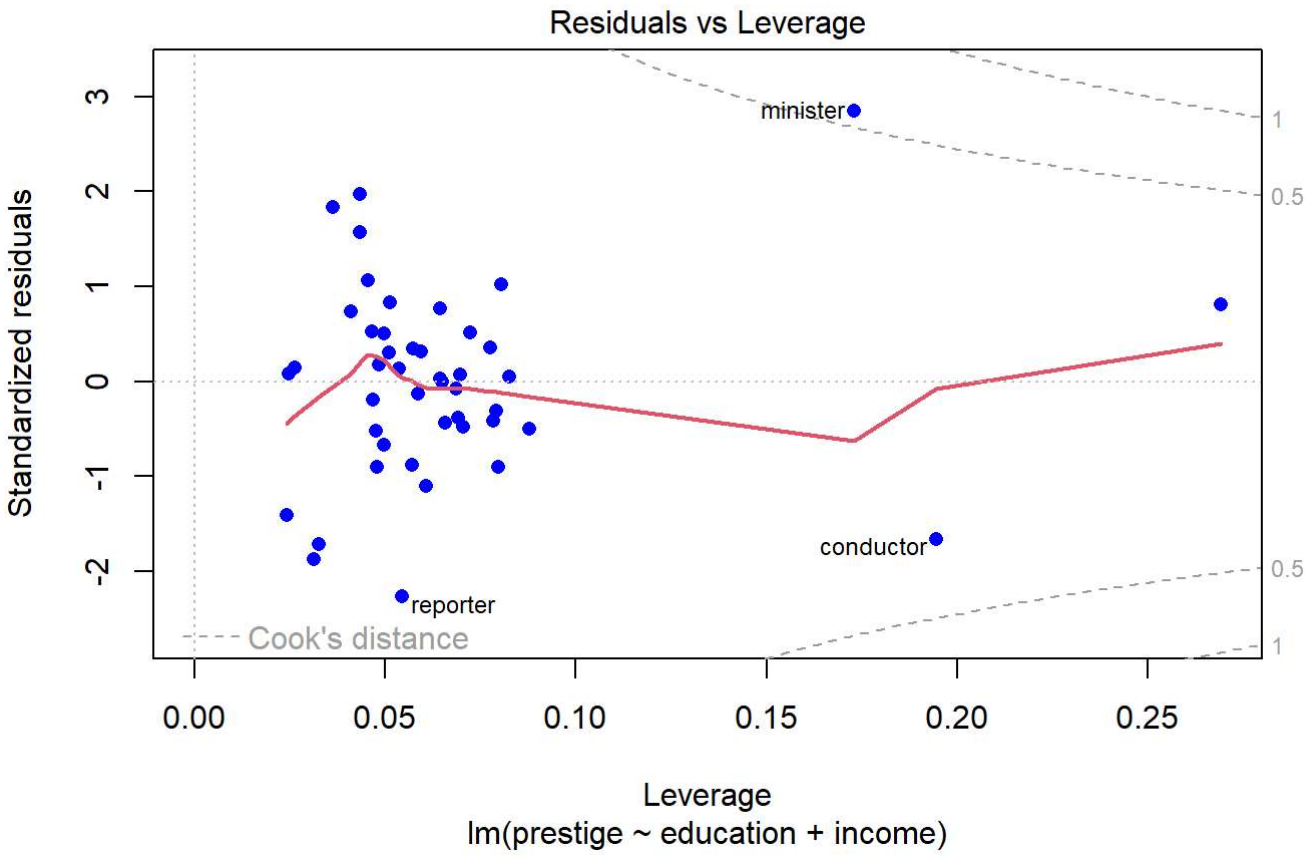
```r
#homoscedasticity and linearity
plot(mod_new, which = 1,pch =16, lwd=2, col="blue")
```



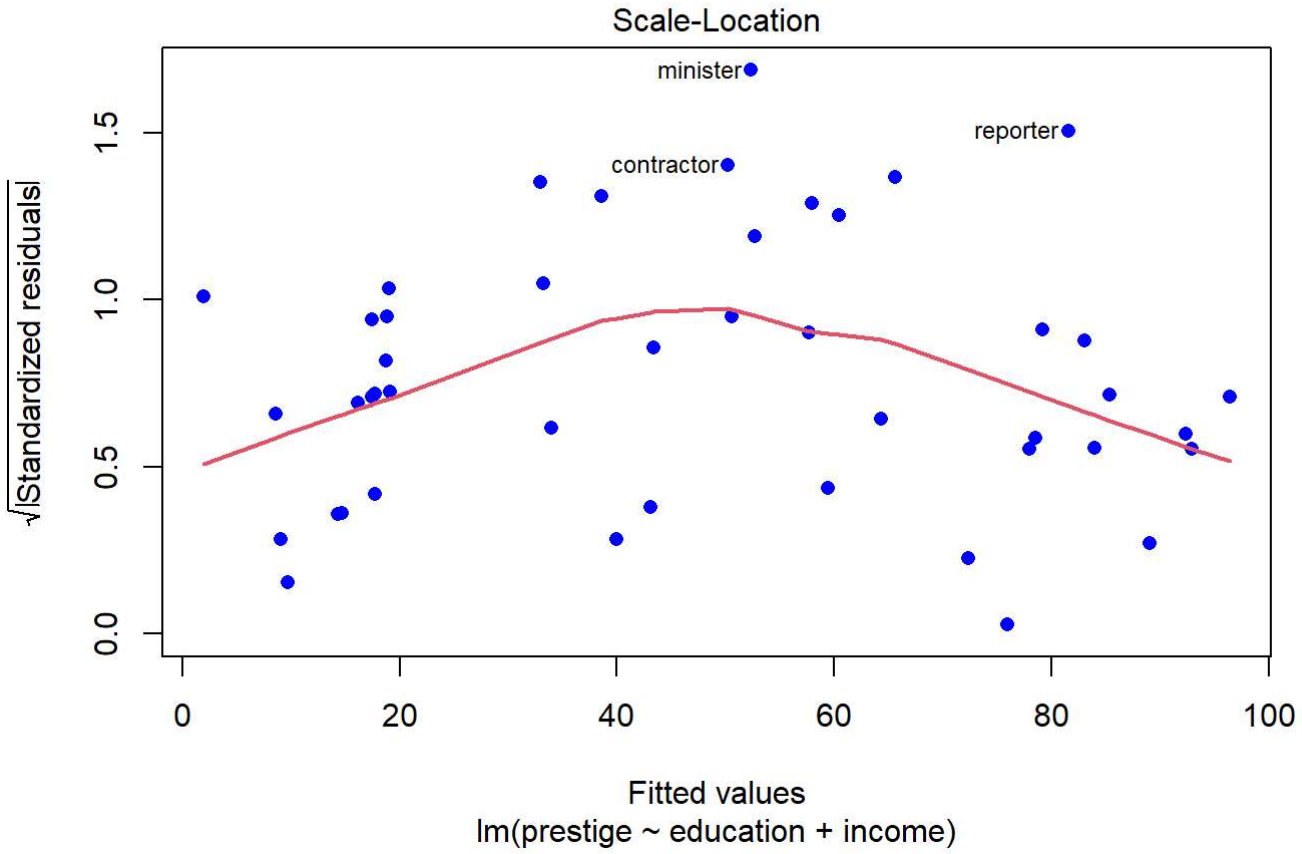

```
#normality of residuals
plot(mod_new, which = 2, pch = 16, lwd=2, col="blue")
```




```
#influence points
plot(mod_new, which = 5, pch=16, lwd=2, col="blue")
```




```
#scale location
plot(mod_new, which=3, pch=16, lwd = 2, col="blue")
```



SUMMARY INTERPRETATION

- With a Multiple R-squared value of 0.8282, the model appears to account for about 82.82% of the variation in prestige.
- The F-statistic evaluates the model's overall significance. Because the p-value is so small (2.2e-16), the model is statistically significant and that education and income are significant predictors.
- Both the education and income p-values for the t-tests are extremely significant (much less than 0.05). This implies that the null hypothesis, according to which these coefficients are equal to zero in the population, is strongly refuted. In other words, there is a statistically significant correlation between status and both income and education.
- *Null Hypothesis* H_0 : The coefficient of education is equal to zero. This means that education has no effect on prestige when holding other variables constant.

  $\beta_{education}$ = 0
- *Alternative Hypothesis* H_1 : The coefficient of education is not equal to zero. This suggests that there is some effect of education on prestige when holding other variables constant.

  $\beta_{education} \neq 0$

Given that the p-value for the coefficient of education is 1.73e-06 (which is much smaller than the commonly used significance level of 0.05), we reject the null hypothesis

- *Null Hypothesis* H_0 : The coefficient of income is equal to zero.

  β_{income} = 0

- *Alternative Hypothesis* H_1 : The coefficient of income is not equal to zero.

  $\beta_{income} \neq 0$

Given that the p-value for the coefficient of income is 1.05e-05 (which is also much smaller than the 0.05 significance level), we reject the null hypothesis. This indicates that income also has a significant effect on prestige when holding other variables constant.

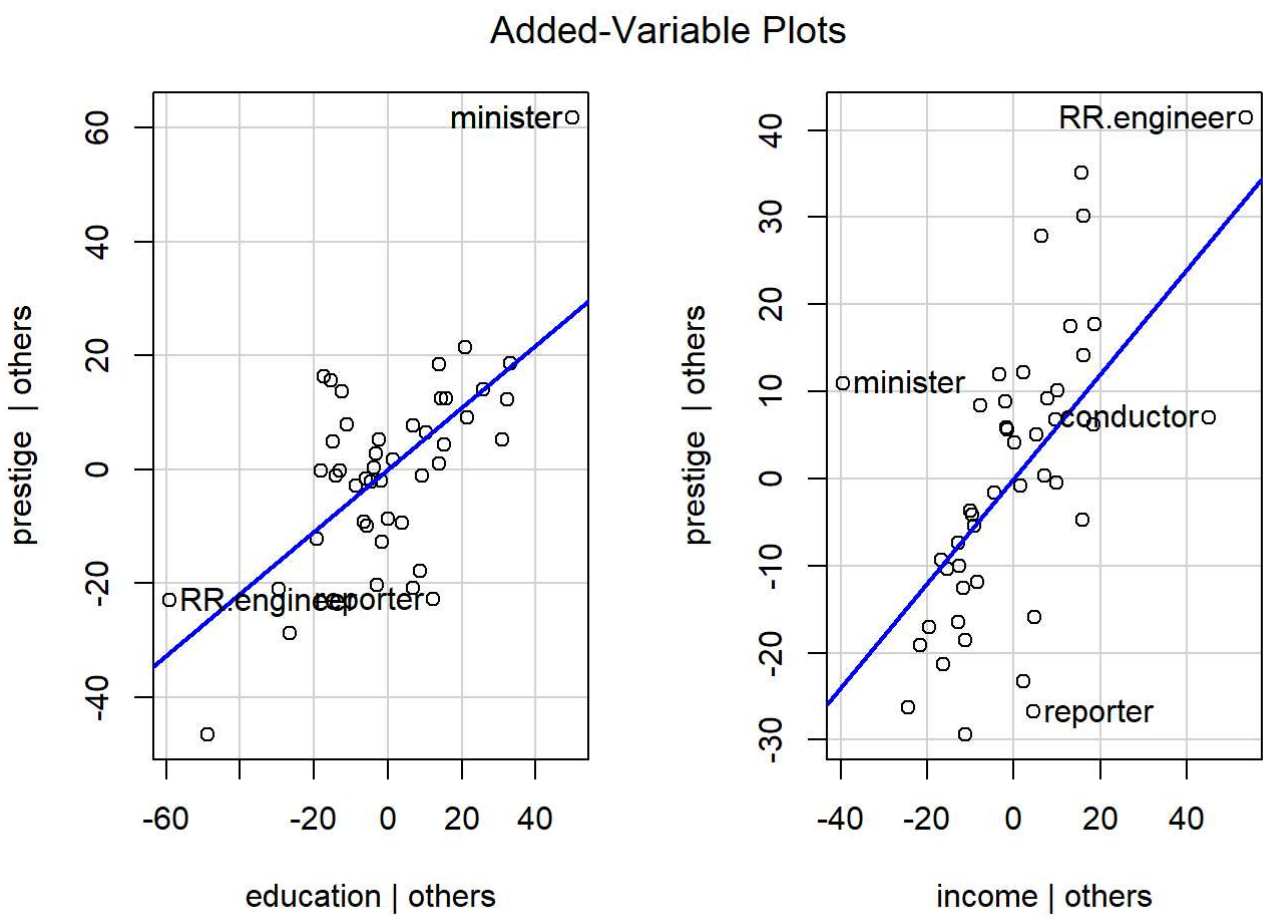
In conclusion, the model provides evidence to reject the null hypotheses for both education and income, suggesting that both predictors have significant effects on prestige.

> ***ASSUMPTIONS***

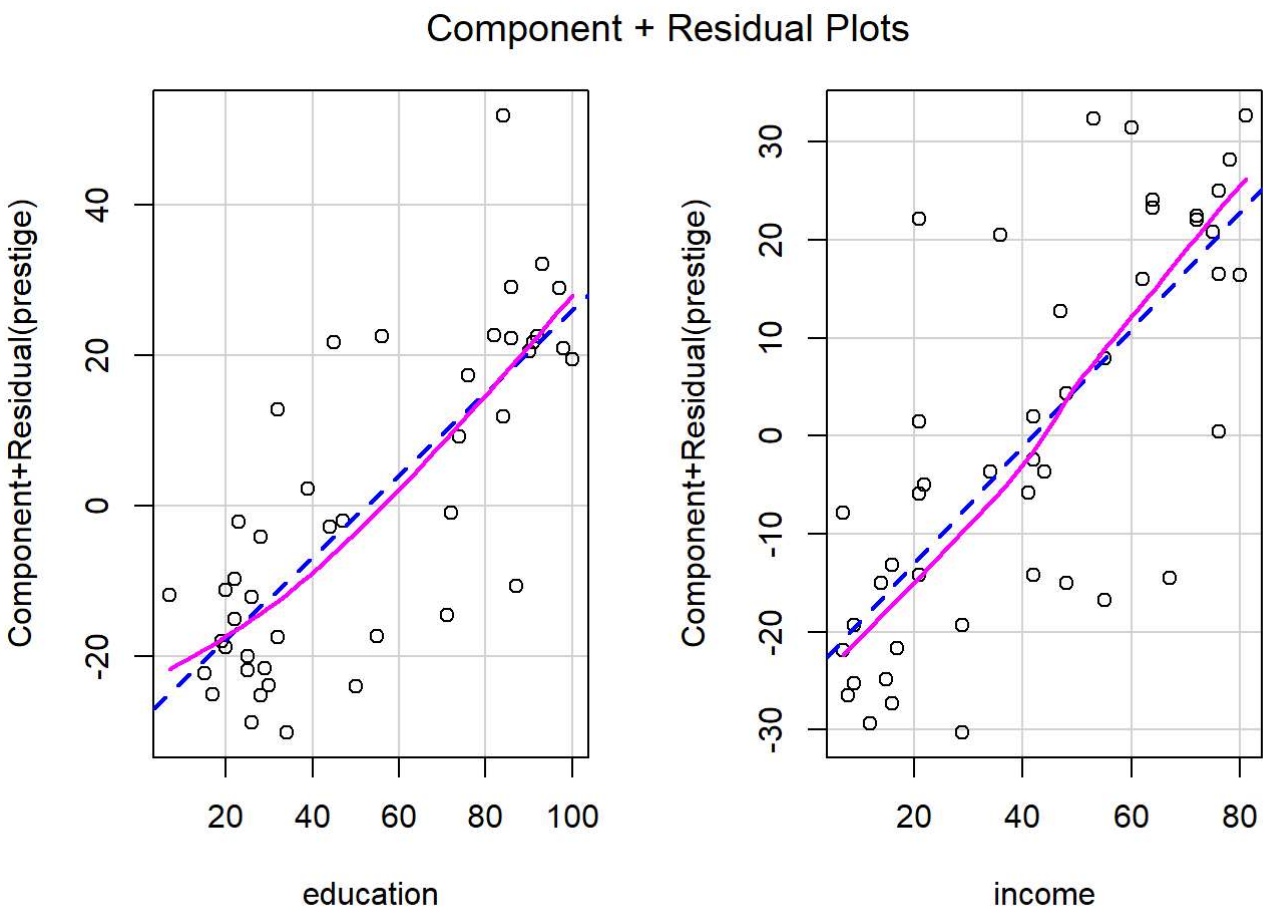
- A horizontal band around the center of the plot indicates homoscedasticity. The presented plot seems to suggest a roughly constant spread, but there's a noticeable trend or curvature, indicating potential heteroscedasticity.
- the residual vs. fitted plot suggests that the relationship is quite linear and due to no distinguishing pattern present, it depicts equal variances.

2 (g)

```
library(car)

#partial regression plots
avPlots(mod_new)
```




```
#partial residual plots
crPlots(mod_new)
```



Partial Residual Plots:

- Education: The plot for education indicates a positive tendency, implying that despite maintaining a constant income, as education rises, status also tends to rise.
- Income: After taking into account the impact of education, the plot for income likewise shows a positive link between income and prestige.

Plots with an Added Variable (Partial Regression)

- Education: Despite taking into account other factors (in this case, income), the plot clearly shows a positive trend, indicating that higher levels of education are linked to greater prestige. Marked words like “minister” and “RR.engineer” are prominent points, presumably signifying their importance.
- Income : Once the impact of education is taken into consideration, there is also a positive correlation between income and prestige. Considerations like “RR.engineer”, “minister”, and “conductor” may have an impact

Differences:

- Nature: The component-plus-residual graphic combines the residuals from the multiple regression with the fitted line from a simple regression of the response on the predictor. The link between the answer and one predictor is depicted in the added-variable plot, on the other hand, after the linear effects of the other predictors have been subtracted from both.

- Analysis: The slope of the line in the partial residual plot offers the same predicted coefficient as in the multiple regression. The slope of the line in the added-variable plot, which accounts for the linear effect of other predictors, depicts the link between the answer and the predictor.

- Residual Representation: The residuals from the whole model are added to the component line in the component-plus-residual graphic. The added-variable plot directly illustrates the particular influence of the predictor on the response by removing the effects of other variables from both the response and the predictor.

- After accounting for additional variables, both charts aid in determining if the connection between the predictor and the response is linear. However, because they would appear further from the line in the added-variable plot, influential points can be found more easily.

In conclusion, despite taking into consideration the interactions between them, both plots show that status and money have positive correlations. The two types of plots differ from one another in their construction, interpretation, and the specifics they reveal about the model.

2 (h)

```
mod_new_summary <- summary(mod_new)
r_squared <- mod_new_summary$r.squared
print(r_squared)
```

```
[1] 0.8281734
```

The value of 0.8281734 means that approximately 82.82% of the variability in the dependent variable (prestige) is explained by the predictors (education and income) in the model.

In other words, education and income together account for about 82.82% of the variation in prestige. This is a high value and suggests that the model fits the data well. The predictors, education and income, appear to be strong indicators of prestige in the dataset.