37252 Regression and Linear Models

Assessment Task 2: Assignment

This assessment task is marked out of 60.

It is worth 30% of the marks for this subject.

Due: 12 noon Thursday 12st May 2022

R1: Q1 (Tegh)
R2: Q2 (Chris A --> D) (Aimann E-->G)
R3: Q3 (Katy)

QUESTION 1. Simple linear regression [20 marks]

Find a dataset suitable for demonstrating simple linear regression. It should contain two numerical variables, one that can be a response variable (y) and one an explanatory variable (x). There should be at least 50 observations.

Q1 Dataset: https://www.kaggle.com/datasets/devchauhan1/salary-datacsv

(a) [3 marks] Use a scatter plot to explore the direction, type and strength of the relationship between the two variables you have identified to be y and x (include the scatterplot with your answer).

Q1 (a)

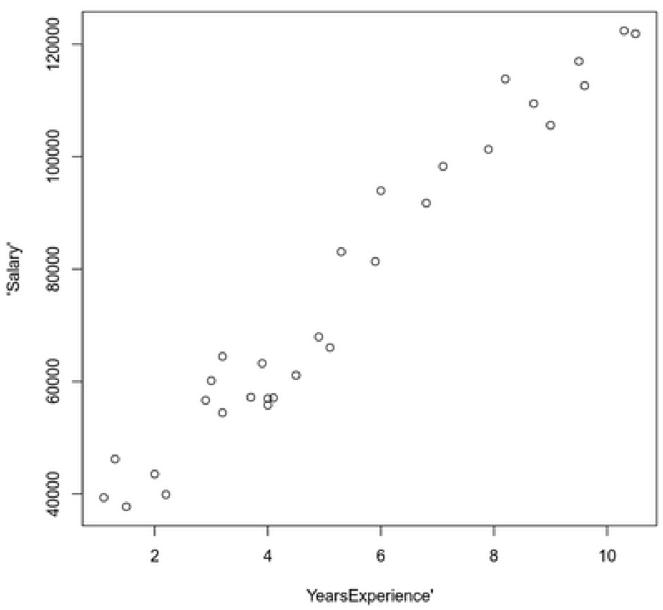
There is fairly strong, positive, linear relationship.

There is linear relationship since the dots seem to follow a line.

The relationship is quite strong as the dots seem quite close to this line.

Positive since as the x-variable increases the y-variable increases

```
df1 = read.csv("Salary_Data.csv")
plot(df1$YearsExperience, df1$Salary, xlab = "YearsExperience'", ylab ="'Salary'")
```



(b) [5 marks] Obtain and write-down the fitted regression line and comment on whether x is a useful predictor using a T-test. Make sure you clearly state the hypotheses, test statistic, test result and your conclusion in plain English.

Q1 (B)

H0: Beta1 = 0 Ha: Beta1 != 0

The regression equation is

Salary-hat = 25792.2 + 9450*YearsExperience

Since the p-value is extremely small (< 2e-16 i.e. test statistic) < 5% hence we reject the null hypothesis.

We conclude the x-variable (YearsExperience) is significant and there by the x-variable is a useful predictor of salary.

```
mod1<-lm(Salary ~ YearsExperience, data = df1)
summary(mod1)
```

```
Call:
lm(formula = Salary ~ YearsExperience, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max 
-7958.0 -4088.5  -459.9  3372.6 11448.0 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      25792.2     2273.1   11.35 5.51e-12 ***
YearsExperience   9450.0      378.8   24.95  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5788 on 28 degrees of freedom
Multiple R-squared:  0.957,	Adjusted R-squared:  0.9554 
F-statistic: 622.5 on 1 and 28 DF,  p-value: < 2.2e-16
```

(c) [2 marks] Interpret the estimated values of β_0 (the intercept) and β_1 (the slope) of the regression line in the context of the dataset.

Q1 (c)

The intercept of the equation, 25792.2 (Beta0) represents what would be the salary if the x-variable = 0.

Similarly, Beta1 represent what would be the increase of salary(9450) if YearsExperience is increase by 1

(d) [2 marks] Find the value of the coefficient of determination R^2 and interpret its value.

Q1 (d)

The coefficent of determination(R-squared) is 0.957.

This means that 95.7% of the variance can be encountered for by the model.

This suggests the model is quite useful.

"""We are presuming coefficient of determination refers to Multiple R-squared"""

(e) [2 marks] Choose a new value for x that is not in your current data. Find the 95% confidence interval for the predicted mean value of y at your chosen value of x.

If we use **df['YearsExperience'].unique()** in Python we will find the 1.2 does not appear as a value

We will take YearsExperience = 1.2 where YearsExperience is the x variable.

Now let construct the confidence interval,

The residual standard error is 5788 on 28 degrees of freedom.

The t critical value is approximately 2.048.

Note y-hat = 25792.2 + 9450*(1.2) = 37132.2

```
# We are using lecture 2, slide 32, equation 21 to construct the confidence interval
```

```
x_bar = mean(df1$YearsExperience)

numerator = (1.2 - x_bar)^2
```

```
# This is Sxx   (https://math.stackexchange.com/questions/1499752/sxx-in-linear-regression)
denominator = sum(df1$YearsExperience^2) - sum(df1$YearsExperience)^2 /30
```

```
std_yhat = 5788*sqrt(1/30 + numerator/denominator)
```

```
#calculate t-score
alpha = 0.05
t_score = qt(p=alpha/2, df=28, lower.tail=F)
```

The confidence interval is approximately (33276.12, 40988.18).

Note there is minor difference with the other answer below due to numerical issues

```
37132.15 - std_yhat*t_score
```

33276.1233361368

```
37132.15 + std_yhat*t_score
```

40988.1766638632

Other answer

```
data <- data.frame(
  YearsExperience = c(1.2)
)
```

```
predict(mod1, data, interval = "confidence")
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	37132.15	33275.92	40988.39

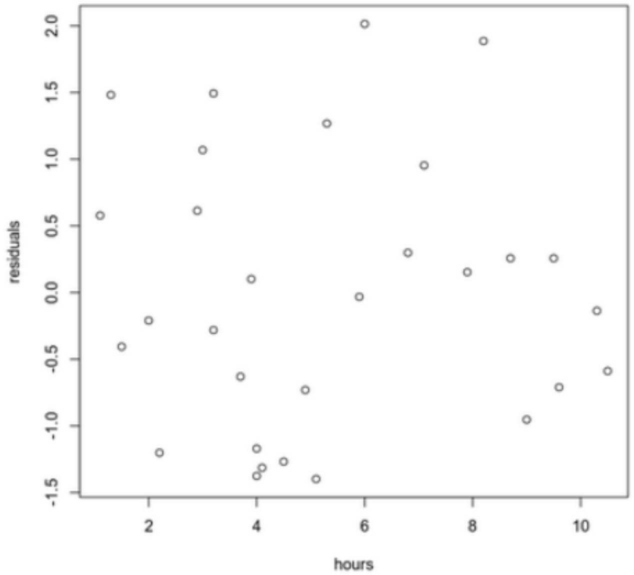
(f) [3 marks] Using appropriate plots, perform a visual analysis of the standardised residuals. Assess the assumptions made about the error terms in the model.

1(f)

Variance seems constant.

Independence seems true.

```
mod1.st.resid<-rstandard(mod1)
plot(df1$YearsExperience, mod1.st.resid, xlab = "hours", ylab = "residuals")
```



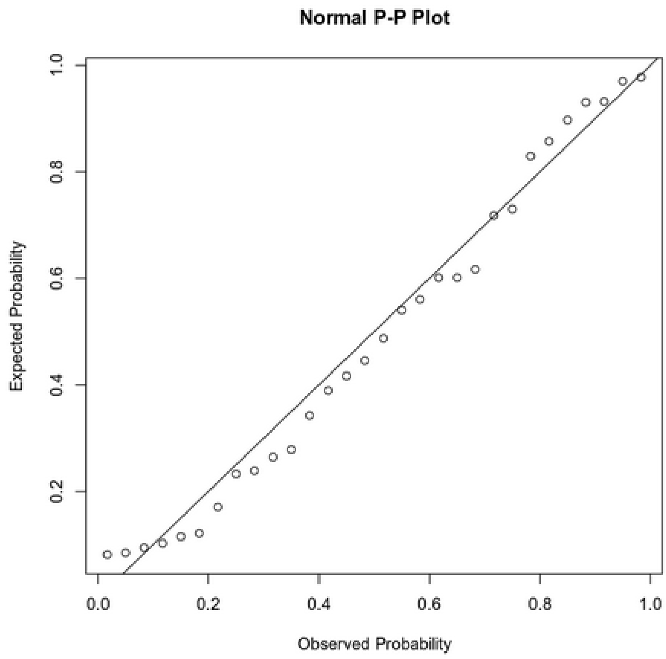
Most of the points seem to be under the line. This could potentially suggest the data is slightly skewwed.

However as such the pp-plot shows a good fit with the line. Therefore the normality assumption is reasonably valid.

```
probDist <- pnorm(mod1.st.resid)

plot(ppoints(length(mod1.st.resid)),
     sort(probDist),
     main = "Normal P-P Plot",
     xlab = "Observed Probability",
     ylab = "Expected Probability")

abline(0,1)
```



(g) [3 marks] Use Cook's D to identify the most influential observation in your data. State the observation number and remove it from the regression. Discuss any impact this has in terms of regression coefficients and R^2.

Q1 (g)

As we can see from the boxplot there are two potentially influential points from the boxplot.

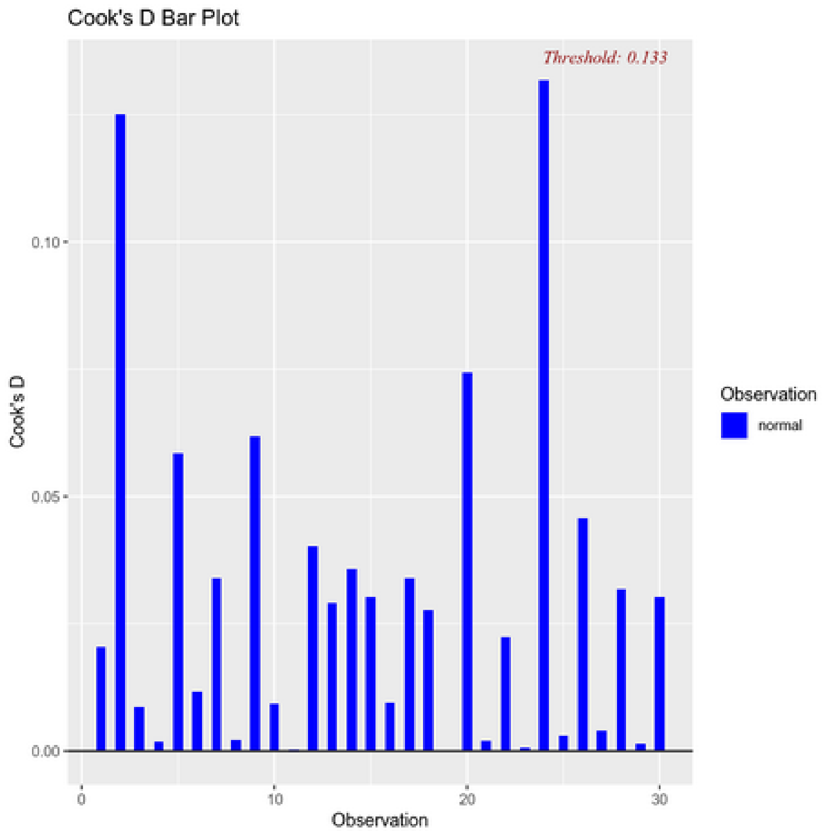
The critical value for Cook's D is 4/(30 - 2 -1) = 4/27

**Note that cooksD is calculate using 4/30 in this example because of the particular environment.

In either case no points will be removed as none of them surpass the threshold

As no points are removed the R^2 is unaffected

```
ols_plot_cooksd_bar(mod1)
```



QUESTION 2. Multiple linear regression [20 marks]

In this question we model fuel consumption. The data are observations from the forty-eight contiguous US states taken in 1980. The variables we consider are summarised in the table below.

Name	Type	Description
consumption	response	state fuel consumption
miles	predictor (continuous)	miles of paved highway
proportion	predictor (continuous)	proportion of population with driver's license

The data are available in "37252_AssessmentTask2_Autumn2022.csv*.

(a) [2 marks] Construct a linear regression model with *consumption* as response and *miles* and *proportion* as predictors. Write down the estimated regression equation and provide interpretations of the estimated coefficients.

$$\widehat{consumption} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$
$$\widehat{consumption} = -243.5 + (2.055) * 10^{-3} * miles + 1418 * proportion$$

The interpretation of the intercept (Beta0) and x coefficients (Beta1, Beta2) are as follows:

B0^Hat:
When all x-variables are equal to zero, *consumption* is equal to approximately –243.5.

B1^Hat (Miles):
For each additional unit of paved highways, with all other x-variables held constant, the estimated fuel consumption is predicted to increase by roughly (2.055) *10^-3 units.

B2^Hat (Proportion):
When all x-variables except proportion are held constant, the fuel consumption for the 48 states is predicted to increase by approximately 1418/100 if proportion increases by 1%.

```
> model <- lm(consumption ~ miles + proportion, data = scoredat)
> summary(model)

Call:
lm(formula = consumption ~ miles + proportion, data = scoredat)

Residuals:
    Min      1Q  Median      3Q     Max 
-118.88  -63.98  -16.62   49.86  250.46 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.435e+02  1.256e+02  -1.938   0.0589 .  
miles        2.055e-03  3.410e-03   0.603   0.5497    
proportion   1.418e+03  2.146e+02   6.608  3.9e-08 ***
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.45 on 45 degrees of freedom
Multiple R-squared:  0.4926,	Adjusted R-squared:  0.4701 
F-statistic: 21.85 on 2 and 45 DF,  p-value: 2.341e-07
```

(b) [4 marks] Test if the regression model is significant at the 0.05 significance level. Write down the hypotheses, the test statistic and p-value, the result of the test and conclusion in plain English.

$$H_0: \beta_0 = \beta_1 = \beta_2 = 0$$
$$H_A: At\ least\ one\ does\ not\ equal\ 0$$
$$Test\ stat = F = \frac{MSR}{MSE} = 21.85$$
$$P - value = 0.0000002341$$

As the calculated p-value has a value that is statistically significant at the 0.05 level, as 0.05 > 2.341e-07, we reject the alternate hypothesis and conclude that the model is statistically significant as all of the variables contribute to the models overall significance.

(c) [2 marks] What is the percentage of the total variation in *consumption* that can be explained by using this multiple linear regression model?

The total percentage of variation explained by this model is 47.01%.

(d) [2 marks] Which predictor variable is more important for explaining *consumption*? Explain your answer (Hint: check p-values).

The proportion of the population with their drivers' license is the stronger predictor variable since its p-value is equal to 3.9e-08 compared to miles with a p-value in this model of 0.5497. Since only the proportions p-value is significant at all levels (1%, 5% and 10%) compared with the miles p-value which is not significant at any level, proportion of population with drivers' license is a stronger predictor within this model to estimate average consumption within the US.

Below is a table of some quantiles from the relevant Student's T distribution.

$t_{0.005}$	$t_{0.01}$	$t_{0.025}$	$t_{0.05}$	$t_{0.1}$	$t_{0.9}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
-2.69	-2.41	-2.01	-1.68	-1.30	1.30	1.68	2.01	2.41	2.69

(e) **[4 marks]** Is there enough evidence to conclude that the coefficient for *proportion* is less than 1750 at the 0.05 significance level? Write down the hypotheses, calculate the test statistic, report the test result and write a conclusion in plain English.

Question 2 - e

Hypotheses

$H_o: \beta_1 = 1750$

$H_A : \beta_1 < 1750$

Test Statistic

T - value = $\dfrac{1418 - 1750}{214.6}$

= -1.547

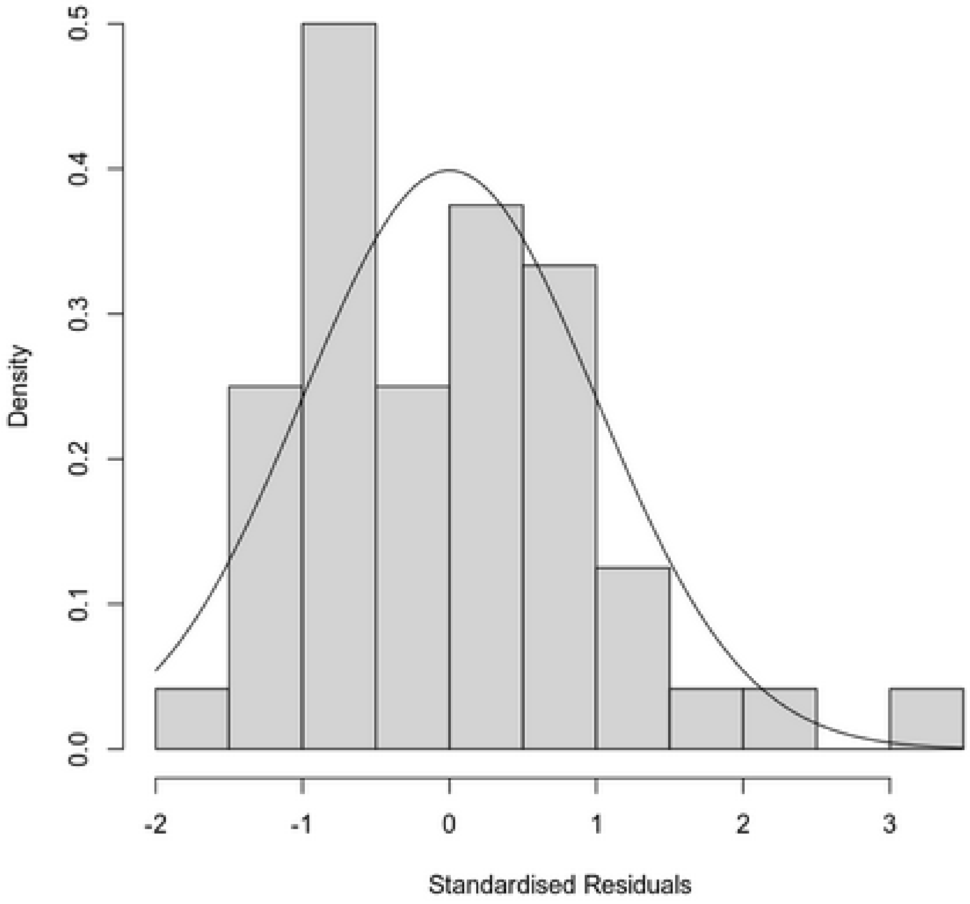
Test decision

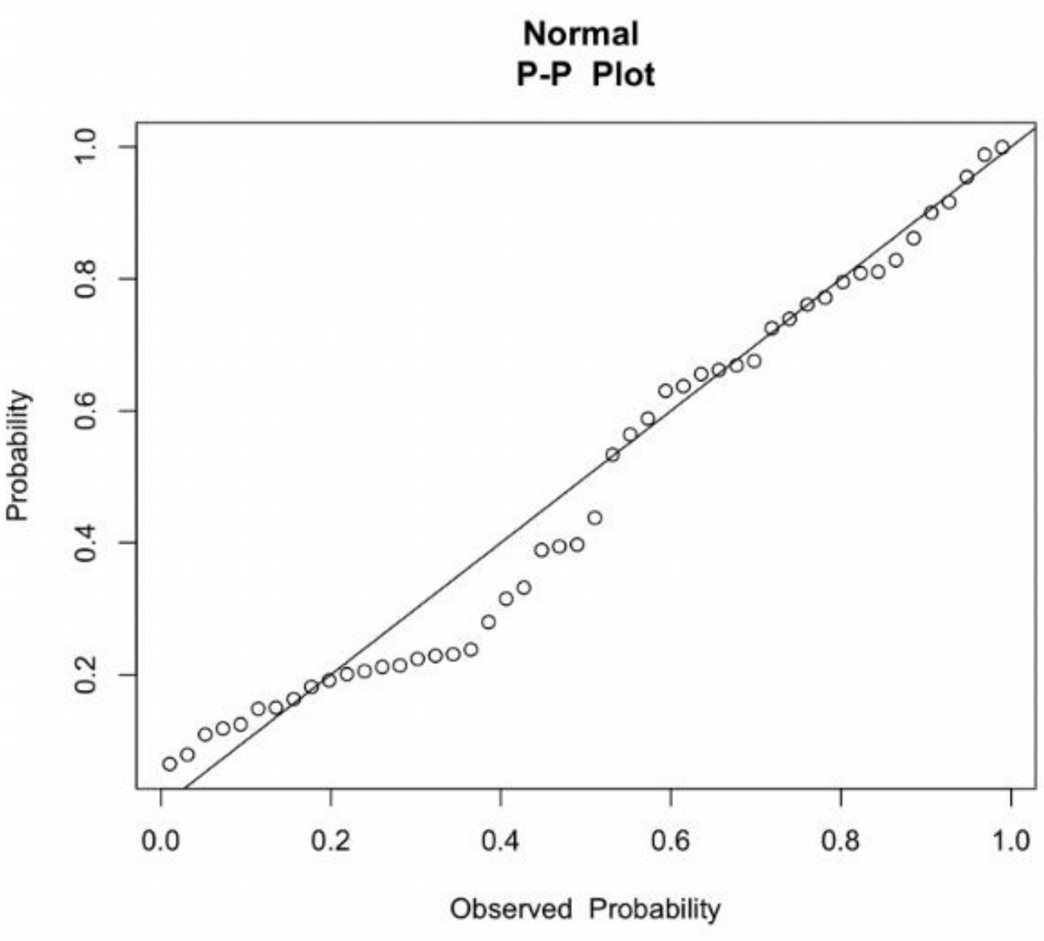
Retain the null hypothesis as $t > t_{0.05}$

Where $t_{0.05}$ equals –1.547

(f) **[3 marks]** State the assumptions made about the error terms in the model. Using appropriate plots, perform a visual analysis of the standardised residuals.

```
> mod1<-lm(consumption ~ miles + proportion, data = scoredat)
> mod1.st.resid<-rstandard(mod1)
> hist(mod1.st.resid, xlab = "Standardised residuals", freq = F, main
+ = "")
> curve(dnorm, add = T)
> probDist <- pnorm(mod1.st.resid)
> plot(ppoints(length(mod1.st.resid)), sort(probDist), main = "Normal
+ P-P  Plot",  xlab  =  "Observed  Probability",  ylab  =  "Expected
+ Probability")
> abline(0,1)
```


0.5
0.4
0.3
0.2
0.1
0.0
Density
-2
-1
0
1
2
3
Standardised Residuals

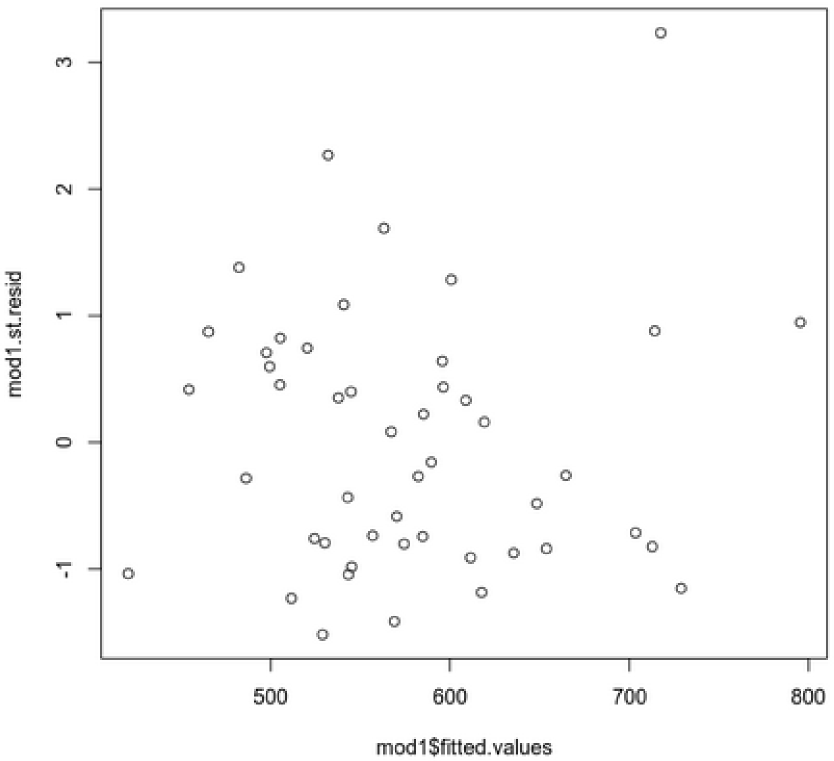


Due to the histogram being centred around –1 on the x–axis we reject normality of the “error terms” i.e. residuals.

The PP plot shows the residuals do not closely follow the line around 0.2-0.5 on the x-axis. This contradicts normality, hence the “error terms” i.e. residuals violate normality.

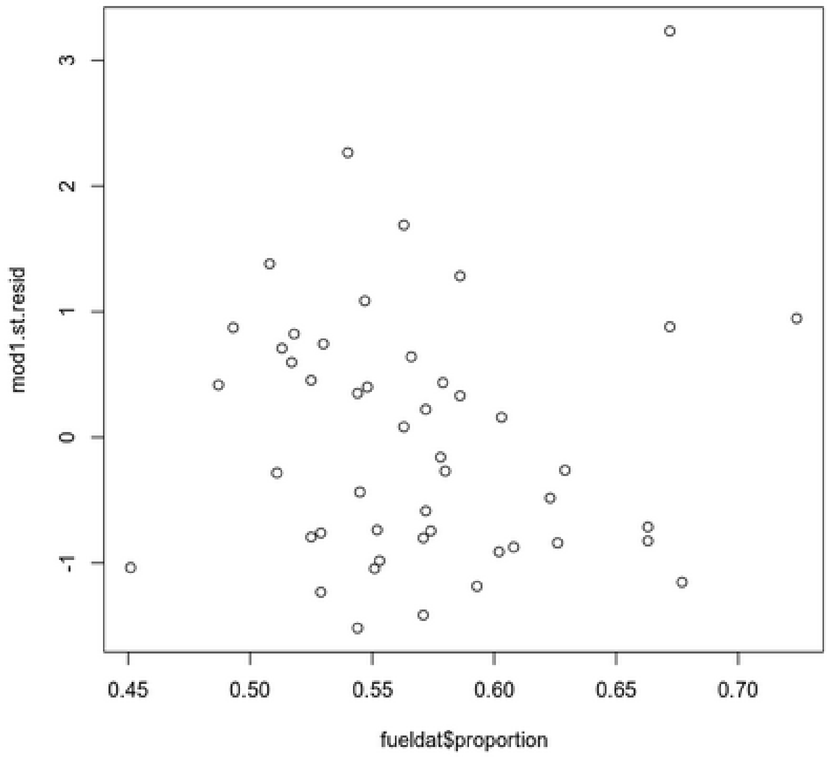
The residuals show no sign of dependence or increasing variance in the scatterplot.

```
plot(mod1$fitted.values, mod1.st.resid)
```



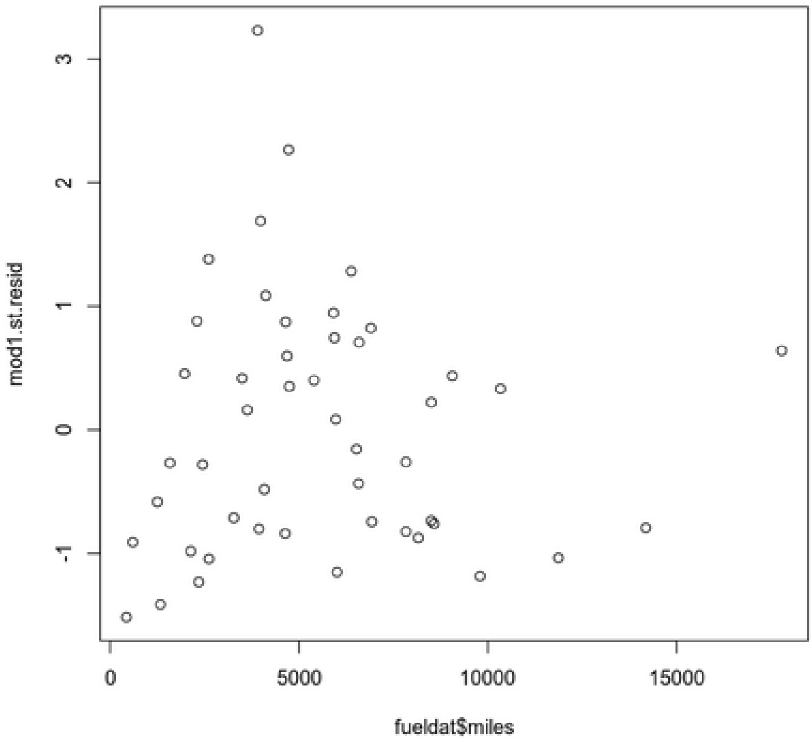
No issues was found with proportion

```
plot(fueldat$proportion, mod1.st.resid)
```



The miles plot show that there might be decreasing variance in the residuals as the miles increases.

```
plot(fueldat$miles, mod1.st.resid)
```



(g) [3 marks] Determine if the residuals are normally distributed at the 0.05 significance level. Write down the hypotheses, the test statistic and p-value, the result of the test and a conclusion in plain English.

```
shapiro.test(mod1.st.resid)
```

```
        Shapiro-Wilk normality test

data:  mod1.st.resid
W = 0.93777, p-value = 0.01334
```

Hypotheses

H_O: the residuals ϵi are normally distributed

H_A: the residuals ϵi are not normally distributed

Test Statistic and P – Value

The test statistic is sw = 0.938 with p-value reported as p = 0.013

Test decision

Reject null hypothesis as p < 0.05.

Conclusion

The evidence is strong enough to conclude that residuals in the dataset model are not normally distributed.

QUESTION 3. Regression with categorical predictor [20 marks]

In this question we extend the model built in Question 2. The variables we now consider are summarised in the table below.

Name	**Type**	**Description**
$consumption$	response	state fuel consumption
$miles$	predictor (continuous)	miles of paved highway
$proportion$	predictor (continuous)	proportion of population with driver's license
$income$	predictor (continuous)	per capita income
$taxBracket$	predictor (categorical)	petrol tax bracket: low (1), medium (2), high (3)

(a) **[5 marks]** Construct a linear regression model with $consumption$ as response and $miles$, $proportion$, $income$ and $taxBracket$ as predictors, also include interaction between $taxBracket$ and $income$. Hint: create dummy variables for $taxBracket$ with $taxBracket = 3$ as reference category. Write down the estimated regression equation and interpret coefficients of the two binary dummy variables and two interaction terms.

```
> fueldat<-read.csv("~/OneDrive/UTS 2022/Regression and Linear Models/37252_AssessmentTask2_Autummn2022_data.csv")
```

```
> fueldat$taxBracket <- as.factor(fueldat$taxBracket)
```

```
> fueldat$taxBracket <- relevel(fueldat$taxBracket, ref = "3")
```

```
> mod1 <- lm(consumption ~ miles + proportion + income + taxBracket + income*taxBracket, data = fueldat)
```

```
> summary(mod1)
```

```
Call:
lm(formula = consumption ~ miles + proportion + income + taxBracket +
    income * taxBracket, data = fueldat)

Residuals:
    Min      1Q  Median      3Q     Max
-103.62  -47.50  -11.51   37.92  227.50

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.995e+01  2.051e+02   0.341   0.7348
miles             -6.561e-04  3.189e-03  -0.206   0.8380
proportion         1.355e+03  2.276e+02   5.955 5.47e-07 ***
income            -7.036e-02  3.668e-02  -1.918   0.0622 .
taxBracket1        9.587e+01  2.149e+02   0.446   0.6580
taxBracket2       -2.426e+01  1.907e+02  -0.127   0.8994
income:taxBracket1 -6.399e-03  5.136e-02  -0.125   0.9015
income:taxBracket2  9.195e-03  4.483e-02   0.205   0.8385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.06 on 40 degrees of freedom
Multiple R-squared:  0.6851,    Adjusted R-squared:  0.6299
F-statistic: 12.43 on 7 and 40 DF,  p-value: 2.508e-08
```

General Equation

$$\widehat{consumption} = 69.95 + (-6.561 \times 10^{-4} \times miles) + (1.355 \times 10^{3} \times proportion) + (-7.036 \times 10^{-2} \times income) + (9.587 \times 10^{1} \times taxBracket_1) + (-2.426 \times 10^{1} \times taxBracket_2) + (-6.399 \times 10^{-3} \times income \times taxBracket_1) + (9.195 \times 10^{-3} \times income \times taxBracket_2)$$

Tax bracket 1 equation:

$$\widehat{consumption} = (69.95 + 9.587 \times 10^{1}) + (-6.561 \times 10^{-4} \times miles) + (1.355 \times 10^{3} \times proportion) + ((-7.036 \times 10^{-2} - 6.399 \times 10^{-3}) \times income)$$

Tax bracket 2 equation:

$$\widehat{consumption} = (69.95 - 2.426 \times 10^{1}) + (-6.561 \times 10^{-4} \times miles) + (1.355 \times 10^{3} \times proportion) + ((-7.036 \times 10^{-2} + 9.195 \times 10^{-3}) \times income)$$

Tax bracket 3 equation:

$$\widehat{consumption} = 69.95 + (-6.561 \times 10^{-4} \times miles) + (1.355 \times 10^{3} \times proportion) + (-7.036 \times 10^{-2} \times income)$$

The coefficient $\beta_{taxBracket_1} = 9.587 \times 10^{1}$ is the predicted difference in *consumption* for *taxBracket₁* compared to *taxBracket₃* with the same *miles* and *proportion* and when *income* = 0.

The coefficient $\beta_{income:taxBracket_1} = -6.399 \times 10^{-3}$ is the predicted difference in the change in *consumption* for one unit increase in *income* holding *miles* and *proportion* constant for $taxBracket_1$ compared to $taxBracket_3$.

The coefficient $\beta_{taxBracket_2} = -2.426 \times 10^{1}$ is the predicted difference in *consumption* for $taxBracket_2$ compared to $taxBracket_3$ with the same *miles* and *proportion* and when *income* = 0.

The coefficient $\beta_{income:taxBracket_2} = 9.195 \times 10^{-3}$ is the predicted difference in the change in *consumption* for one unit increase in *income* holding *miles* and *proportion* constant for $taxBracket_2$ compared to $taxBracket_3$.

(b) [2 marks] Using R to make the calculations (i.e. without using the regression equation directly), find predicted fuel consumption and 95% individual confidence interval associated with this prediction when $miles = 697$, $proportion = 0.56$ and $income = 4568$ for low petrol tax bracket states.

(b)

```
> newdata <- data.frame(taxBracket = c("1"), miles = 697, proportion = 0.56, income = 4568)
> predict(mod1, newdata)
       1
573.6924
```

$when\ miles = 697, proportion = 0.56\ and\ income = 4568:$

$$\widehat{consumption}\ for\ tax\ bracket\ 1 = 573.6924$$

```
> newdata = data.frame(taxBracket = c("1"), miles = 697, proportion = 0.56, income = 4568)
> predict(mod1, newdata,interval="confidence", level = 0.95)
       fit      lwr      upr
1 573.6924 511.332 636.0528
```

(c) [2 marks] Comment on the statistical significance of the interaction terms. What does the result imply?

The terms involving income*taxBracket(1/2) are most likely not statistically significant. As we can see the estimate values of the coefficients are extremely small and the p-value is >0.05.

(d) [2 marks] Write down the adjusted R^2, compare with the one in Question 2 and comment.

(d)

```
Residual standard error: 68.06 on 40 degrees of freedom
Multiple R-squared:  0.6851,    Adjusted R-squared:  0.6299
F-statistic: 12.43 on 7 and 40 DF,  p-value: 2.508e-08
```

Adjusted $R^2 = 0.6299$

The adjusted R^2 suggests that the model in Q3 is performing worse in Q2. Since the adjusted R^2 is lower in Q3 this means less variance is being accounted for given the number of variables.

This suggest that model 3 might be overfitting the data.

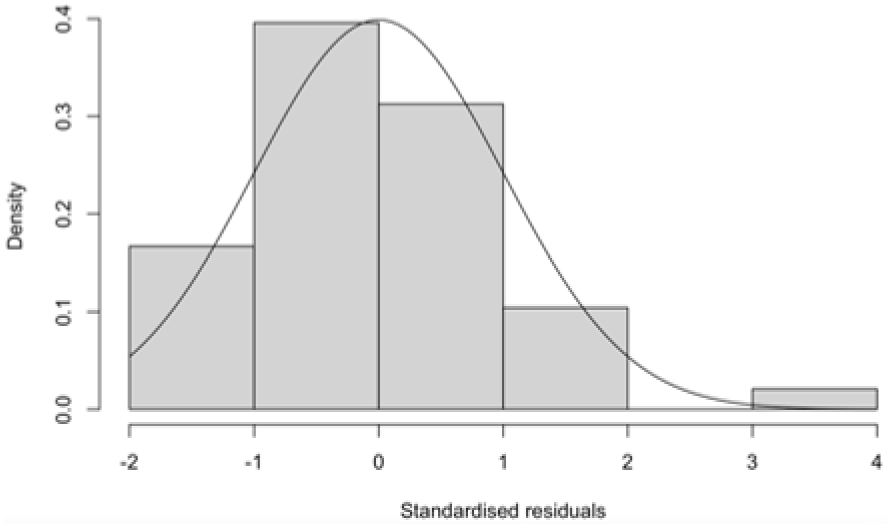
(e) [2 marks] By performing an appropriate regression, calculate the VIF for the predictor *miles* in the model in part (a). Make sure you show all working.

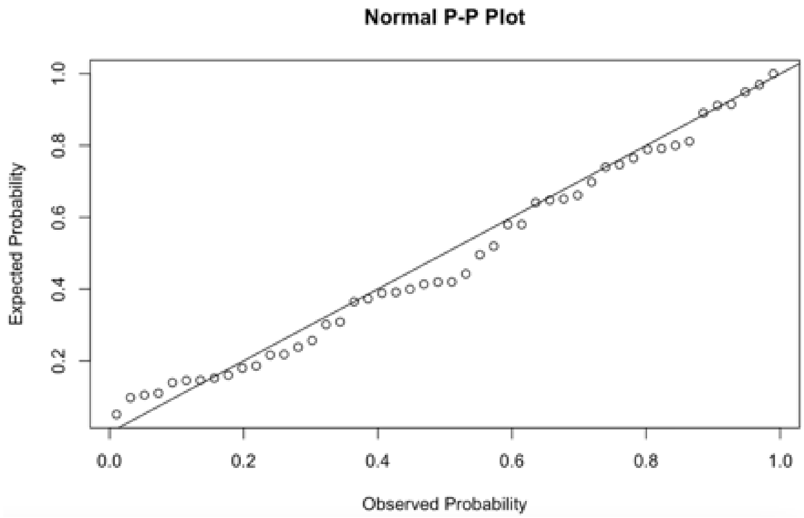
```
> vif(mod1)
           miles
        1.257954
```

As the VIF is less than 5, there is no evidence that *miles* is potentially collinear.

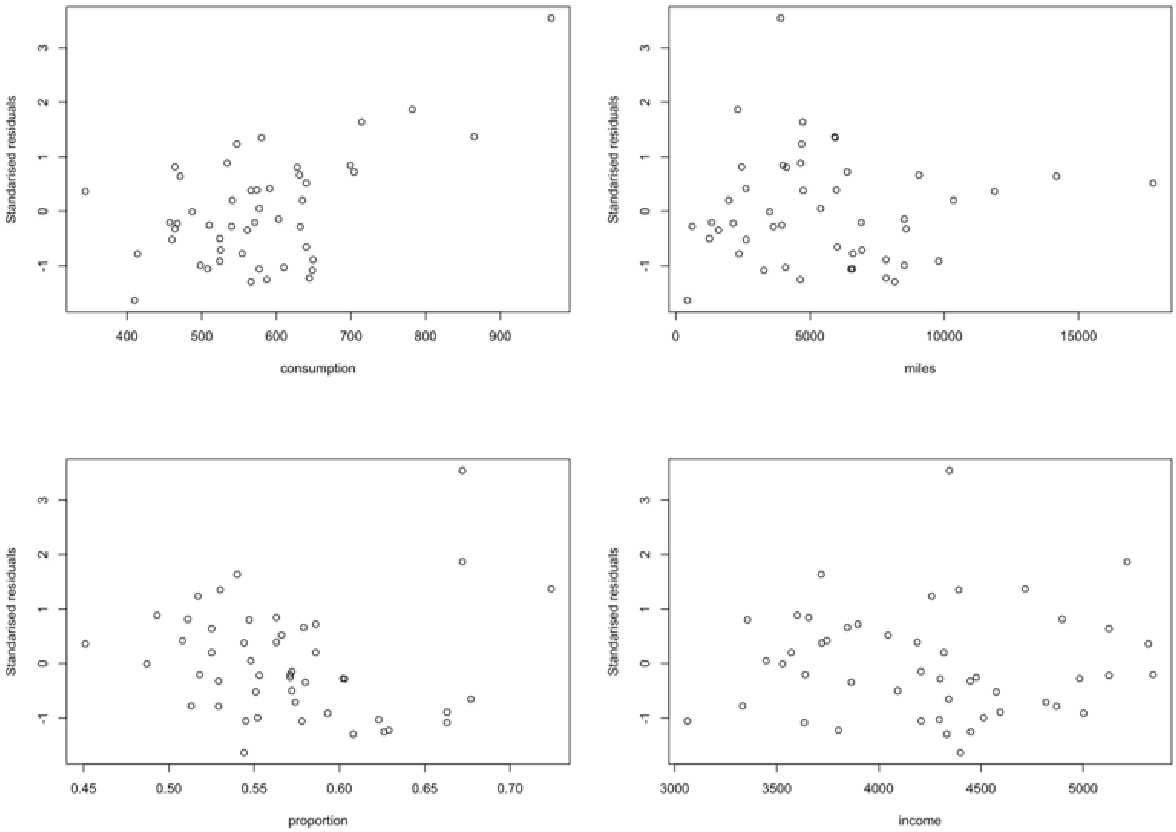
(f) [3 marks] Using appropriate plots, perform a visual analysis of the standardised residuals. Assess the assumptions made about the error terms in the model.

```
mod1.st.resid<-rstandard(mod1)
hist(mod1.st.resid, xlab = "Standardised residuals", freq = F, main = "")
curve(dnorm, add = T)
probDist <- pnorm(mod1.st.resid)
plot(ppoints(length(mod1.st.resid)), sort(probDist), main = "Normal P-P Plot", xlab = "Observed Probability", ylab = "Expected Probability")
abline(0,1)
```





Normality: the histogram and p-p plot show some deviation from normality.



Independence: there are no obvious patterns in the residuals, therefore the independence assumption is satisfied.
Constant variance: some possible fluctuation in variance for *proportion*, which may suggest a problem with constant variance.

(g) [2 marks] Determine if there is any statistical evidence against the assumption of independence of the error terms.

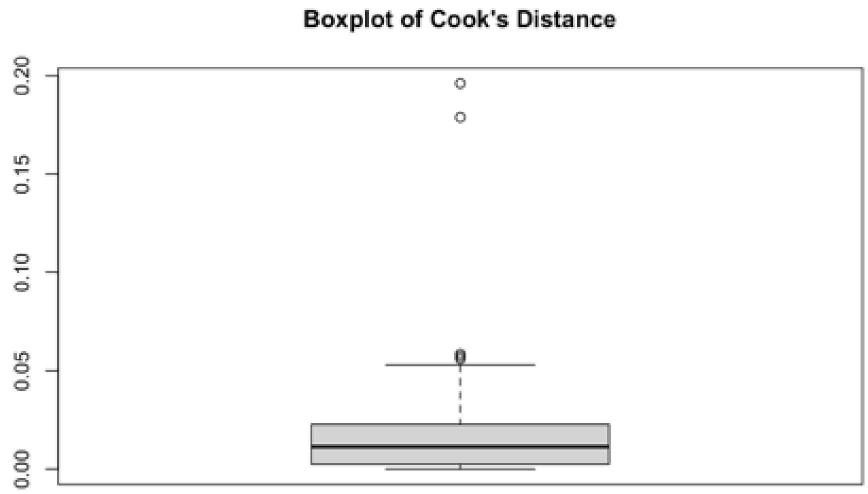
```
> durbinWatsonTest(mod1)
 lag Autocorrelation D-W Statistic p-value
   1     -0.07982098      2.142451   0.918
 Alternative hypothesis: rho != 0
```

The DW statistic is between 1 and 3, indicating no problem with serial correlation and the p-value is greater than 0.05.
There is no statistical evidence against the assumption of independence.

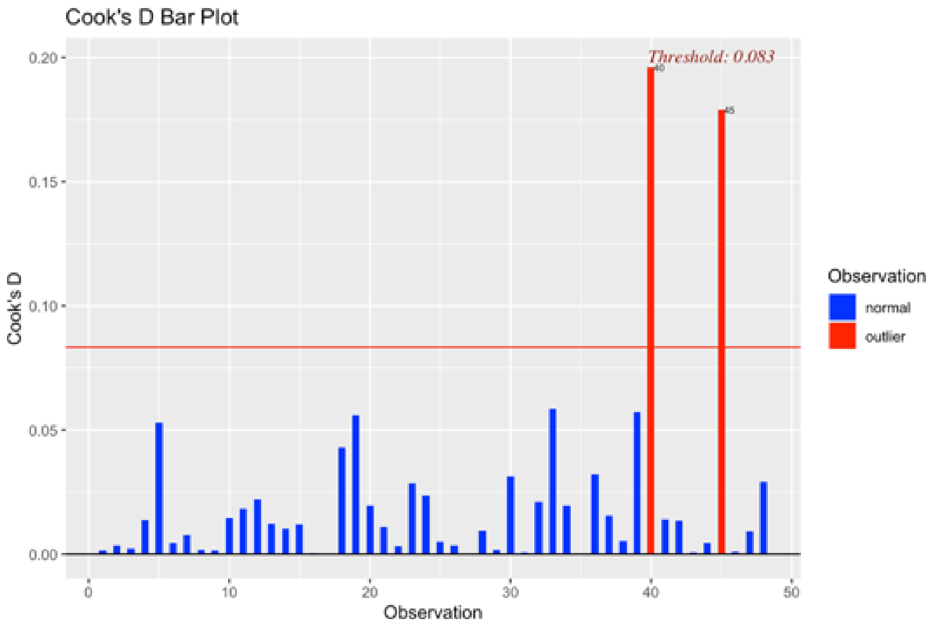
(h) [2 marks] Identify any potentially influential points by calculating the appropriate statistic.

$$\text{Critical Cook's D} = \frac{4}{n} = \frac{4}{48} = 0.083$$

```
> cooksD<-cooks.distance(mod1)
> boxplot(cooksD, main = "Boxplot of Cook's Distance")
> abline(h = 0.22)
```




```
> library('olsrr')
> ols_plot_cooksd_bar(mod1)
```



Observations 40 and 45 are two potentially influential points. They both have a Cook's D value above the critical value of 0.083.