DATA 303/473 Assignment 1

Junn Pham, 300538618

2024-03-21

Contents

Q1.

a. **(4 marks)** Carry out an exploratory data analysis (EDA). NOTE: The predictors chord.length and speed are numerical, but only have a few different values each (6 for chord.length and 4 for speed). Such variables are best treated as categorical variables during the analysis. List any key points of note from your EDA, including any considerations you might make during a regression analysis.

```r
#Variable types
data <- read.csv("airfoil_self_noise.csv", header = TRUE)
str(data)
```

```
## 'data.frame':    1503 obs. of  6 variables:
##  $ frequency   : int  800 1000 1250 1600 2000 2500 3150 4000 5000 6300 ...
##  $ angle       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ chord.length: num  0.305 0.305 0.305 0.305 0.305 ...
##  $ speed       : num  71.3 71.3 71.3 71.3 71.3 71.3 71.3 71.3 71.3 71.3 ...
##  $ displace    : num  0.00266 0.00266 0.00266 0.00266 0.00266 ...
##  $ decibels    : num  126 125 126 128 127 ...
```

```r
#convert chord.length and speed into categorical variables
data$chord.length <- factor(data$chord.length)
data$speed <- factor(data$speed)
str(data)
```

```
## 'data.frame':    1503 obs. of  6 variables:
##  $ frequency   : int  800 1000 1250 1600 2000 2500 3150 4000 5000 6300 ...
##  $ angle       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ chord.length: Factor w/ 6 levels "0.0254","0.0508",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ speed       : Factor w/ 4 levels "31.7","39.6",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ displace    : num  0.00266 0.00266 0.00266 0.00266 0.00266 ...
##  $ decibels    : num  126 125 126 128 127 ...
```

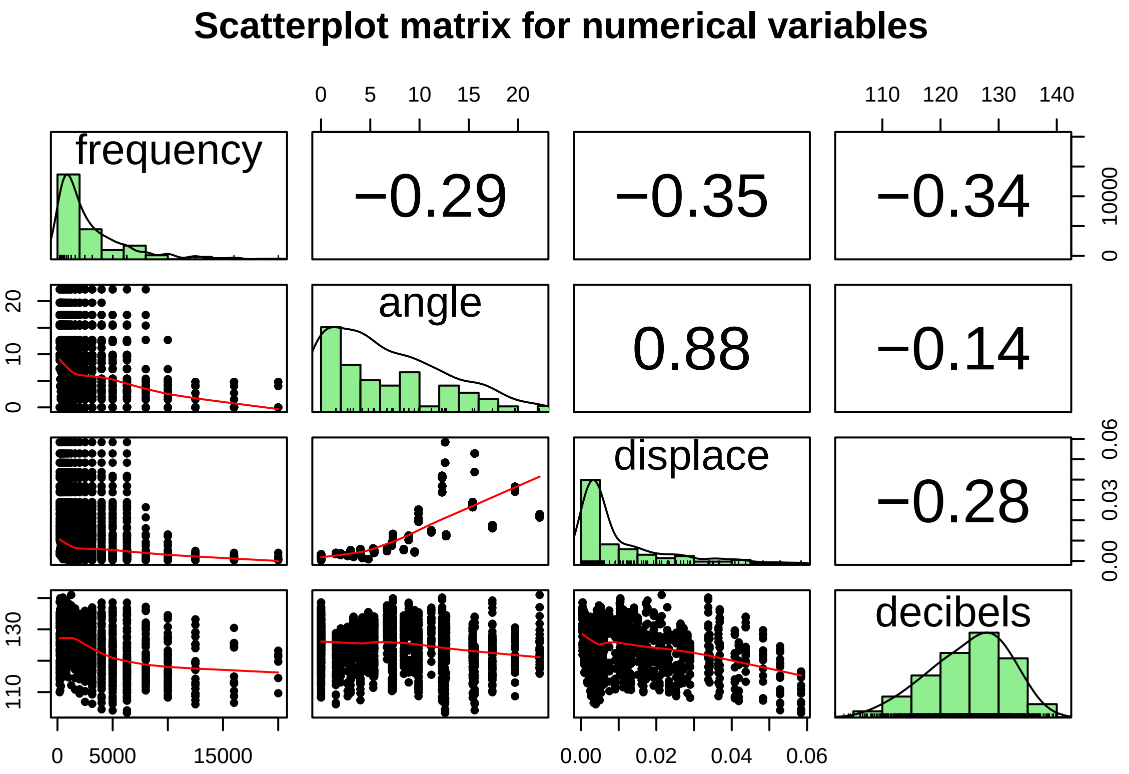
There are 1503 observations and six variables, four of which are numeric and two are converted into categorical data.

```r
#Summary statistics
summary(data)
```

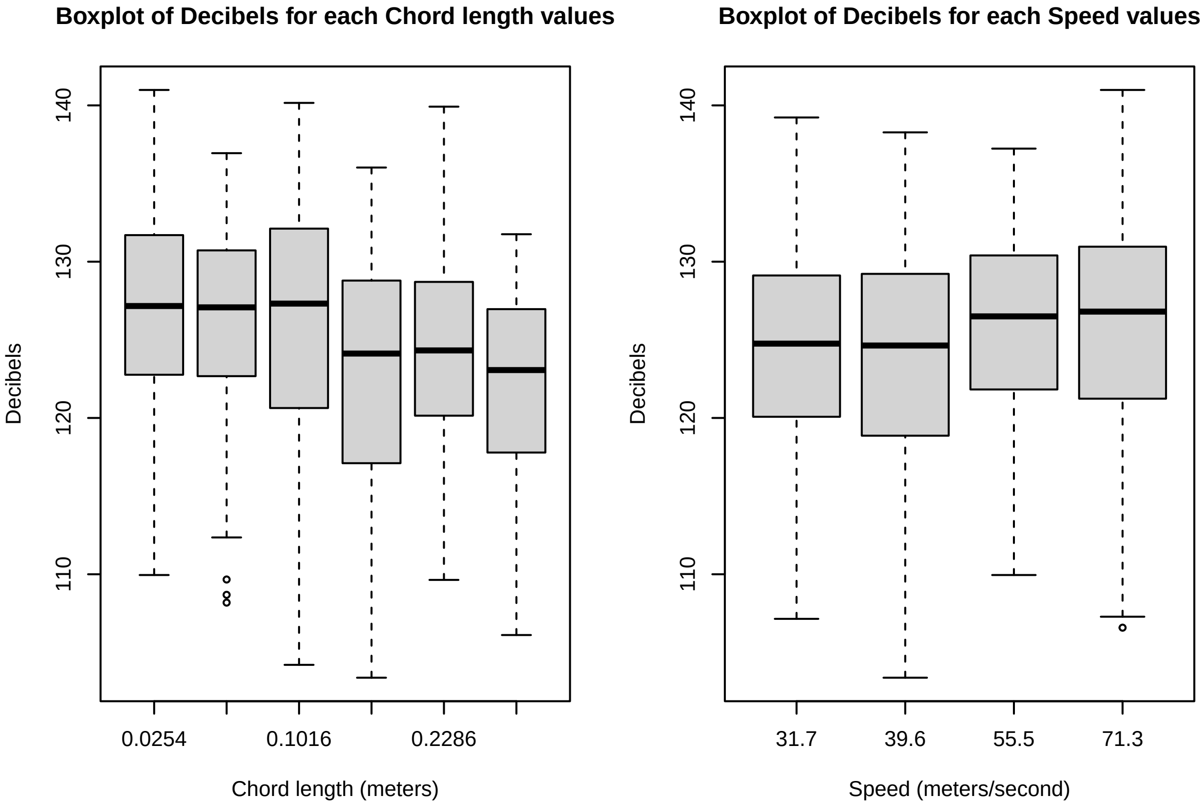
```
##    frequency         angle        chord.length  speed        displace
##  Min.   :  200   Min.   : 0.000   0.0254:278   31.7:281   Min.   :0.0004007
##  1st Qu.:  800   1st Qu.: 2.000   0.0508:237   39.6:480   1st Qu.:0.0025351
##  Median : 1600   Median : 5.400   0.1016:263   55.5:277   Median :0.0049574
##  Mean   : 2886   Mean   : 6.782   0.1524:271   71.3:465   Mean   :0.0111399
##  3rd Qu.: 4000   3rd Qu.: 9.900   0.2286:266              3rd Qu.:0.0155759
##  Max.   :20000   Max.   :22.200   0.3048:188              Max.   :0.0584113
##     decibels
##  Min.   :103.4
##  1st Qu.:120.2
##  Median :125.7
##  Mean   :124.8
##  3rd Qu.:130.0
##  Max.   :141.0
```

The minimum and maximum values seem sensible and there are no variables with missing values. the total frequency of eight distinct chord.length categorical values is 1503 as expected. And the total frequency of four distinct speed values also has the same result.

```
#Data visualisations: Scatterplot matrix for numerical variables
library(dplyr)
library(psych)
data%>%
dplyr::select(where(is.numeric))%>% #select numerical variables (includes integers)
pairs.panels(method = "spearman", # correlation method
hist.col = "lightgreen", # histogram color
density = TRUE, # show density plots
ellipses = FALSE,# do not show correlation ellipses
main = "Scatterplot matrix for numerical variables"
)
```



```r
#Data visualisations: Boxplots for categorical variables
par(mfrow=c(1,2), cex = 0.65)
boxplot(split(data$decibels, data$chord.length), xlab = "Chord length (meters)",
        ylab = "Decibels", main = "Boxplot of Decibels for each Chord length values")
boxplot(split(data$decibels, data$speed), xlab = "Speed (meters/second)",
        ylab = "Decibels", main = "Boxplot of Decibels for each Speed values")
```



The visualisations tell us:

- There appears to be weak negative correlations between decibels and the other numeric variables. This means that as these predictors decrease, decibels tend to increase.
- The highest correlation coefficient is around -0.34, which is the correlation between decibels and frequency.
- The scatterplots between decibels versus numeric variables all show weak correlations and data points are quite scattered.So, it is difficult to say definitively from this scatterplot matrix alone whether there are truly linear relationships between decibels and the other variables.

- The boxplot shows that there is a lower range of decibel values for longer chord lengths, which means that with increasing chord lengths, the sound produced can have a lower range of loudness and these variables have a negative correlation.
- On the other hand, there is a wider range of decibel values for higher speeds, which means that these variables might have a positive correlation.
- The distribution of the decibel variable is fairly symmetrical.Since the median and mean, which is 125.7 decibels and 124.8 decibels respectively, are similar and the IQR [120.2,130.0] is relatively small, the distribution is likely symmetrical with most of the values concentrated around the center (125 decibels).
- There are a very strong correlation between some angle and displace predictors - multicollinearity is likely to be an issue, so we should investigate it.

b. **(3 marks)** Fit a linear model to the data, including all predictors with no transformations or interactions. Present a summary of the model in a table and write the fitted model equation. Give an estimate of 2, the error variance.

```
fit1<-lm(decibels ~ frequency + angle + chord.length +  speed + displace, data=data)
library(pander)
pander(summary(fit1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	135.8	0.5447	249.4	0
frequency	-0.001295	4.252e-05	-30.46	8.174e-159
angle	-0.47	0.04446	-10.57	3.09e-25
chord.length0.0508	-1.603	0.449	-3.57	0.0003686
chord.length0.1016	-2.951	0.5103	-5.783	8.935e-09
chord.length0.1524	-6.551	0.5039	-13	1.117e-36
chord.length0.2286	-7.71	0.4954	-15.56	1.013e-50
chord.length0.3048	-10.22	0.5596	-18.26	2.079e-67
speed39.6	0.7997	0.3621	2.208	0.02737
speed55.5	2.252	0.4062	5.545	3.472e-08
speed71.3	4.077	0.3711	10.99	4.667e-27
displace	-125.7	17.78	-7.07	2.375e-12

Table 2: Fitting linear model: decibels ~ frequency + angle + chord.length + speed + displace

Observations	Residual Std. Error	R^2	Adjusted R^2
1503	4.773	0.5249	0.5214

The fitted model equation:

```
> Decibels = 135.8 - 0.001295\*Frequency - 0.47\*Angle - 1.603\*chord.length0.0508 - 2.951\*chord.lengt
```

The estimate of the error variance:

```
> The estimate of the error variance = 4.773\^2 = 22.7815
```

c. **(3 marks)** Based on the fitted model results in part (b), give an interpretation of the coefficients for angle and speed71.3.

- The coefficient for angle is -0.47 with a standard error of 0.04446. This indicates that for each unit increase in the angle (degree), there is a corresponding decrease of 0.47 in Decibels, holding all other variables constant. The p-value of 3.09e-25 suggests that this coefficient is statistically significant (p-value < 0.05), indicating that the relationship between angle and Decibels is unlikely to be due to random chance alone.
- The coefficient for speed71.3 is 4.077 with a standard error of 0.3711. This suggests that for each unit increase in the speed at 71.3 (meters per second), there is a corresponding increase of 4.077 in Decibels, holding all other variables constant. The p-value of 4.667e-27 suggests that this coefficient is statistically significant (p-value < 0.05), indicating that the relationship between speed at 71.3m/s and Decibels is unlikely to be due to random chance alone.

d. **(2 marks)** Does it make practical sense to interpret the intercept in this case? Justify your answer.

Interpreting the intercept is practical only if predictor values of zero are meaningful for the specific scenario and if observations near zero are present for all predictors. In this scenario, while angle, chord length, and displacement have data points near zero, the remaining variables do not share the same characteristic. Thus, interpreting Beta0 = 135.8 might not be practically meaningful in this context.

e. **(3 marks)** Obtain 95% confidence and prediction intervals for the last three observations in the dataset. Explain briefly why the prediction intervals are wider than the confidence intervals.

```r
#Get predictor values to predict for as a dataframe.
data1 <- data[(nrow(data)-2):nrow(data), -ncol(data)]

#Confidence intervals for estimating the mean response
pander(predict(fit1, newdata=data1, interval="confidence"), caption="Confidence intervals", round=2)
```

Table 3: Confidence intervals

	fit	lwr	upr
1501	114.5	113.5	115.5
1502	113.2	112.2	114.3
1503	111.5	110.5	112.6

```r
#Prediction intervals for predicting the response for new predictor values
pander(predict(fit1, newdata=data1, interval="prediction"), round=2, caption="Prediction intervals")
```

Table 4: Prediction intervals

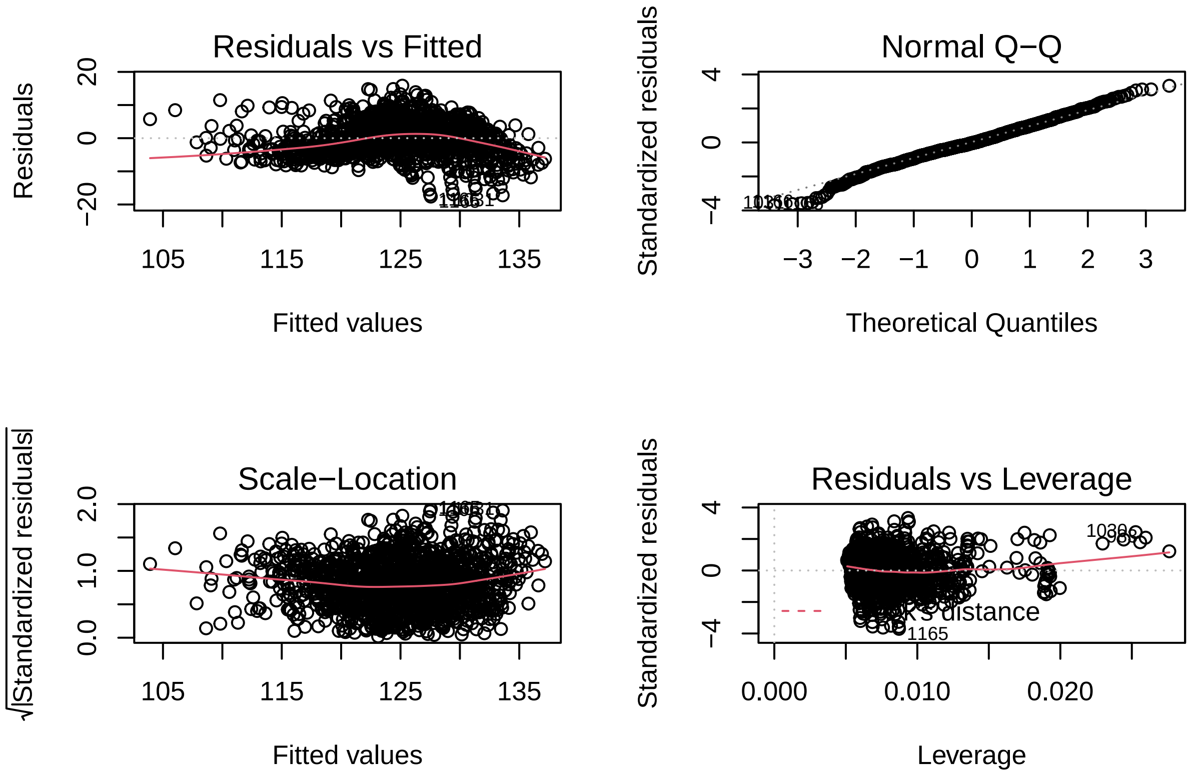
	fit	lwr	upr
1501	114.5	105.1	123.9
1502	113.2	103.8	122.6
1503	111.5	102.1	121

To explain why the prediction intervals are wider than the confidence intervals, we should interpret the terms "prediction interval" and "confidence interval". Prediction intervals consider both the spread of data around the trend line and uncertainty about where the line truly lies. Confidence intervals, on the other hand, only focus on the uncertainty about the line.

For instance, the 95% confidence interval of the last observation in the dataset is (110.2,112.6), which means that 95% of this interval contains the true average decibel value for this observation; While the 95% prediction interval of this data point, which is (102.1,121), is interpreted that 95% of the prediction interval contains the true decibel value for the particular observation. The prediction interval is wider than the confidence interval, reflecting greater uncertainty about the specific decibel value of a given observation compared to uncertainty about the average decibel values of various data points.

f. **(4 marks)** Use the plot function to carry out residual diagnostics for the model you fitted in part (b). Comment on what the residual plots indicate about regression assumptions or the existence of influential observations.

```r
par(mfrow=c(2,2))
plot(fit1)
```



- **Linearity: Residuals vs fitted plot**

The concave curve in the plot suggests that there are non-linear relationships between the response and some of the predictors that is not captured by the model. In other words, we are underfitting the model and violating the first Multiple Linear Regression (MLR) assumption about linearity in parameters.

In this case, transformations of the response and predictor variables should be considered in further analyses.

- **Normality: Normal Q-Q plot**

Due to the presence of a few data points with theoretical quantile values below -2 that deviate from the straight line, we say that there is some evidence of non-normality. However, this can occur even if the normality assumption (MLR6) holds. Hence, it's advisable to conduct additional analyses to thoroughly assess the normality of the residuals.

- **Equal variance (homoscedasticity): Scale-Location plot**

Although the red line of the S-L plot is approximately horizontal, the spread around the red line varies with the fitted values. The spread of magnitudes appears to be lowest for fitted values near 105 and highest for fitted values around 125. Therefore, There is evidence of non-constant variance (heteroscedasticity), which violates the MLR5 assumption.

- **Influential observations: Residuals vs leverage plot**

The residuals vs leverage plot suggests that there are no highly influential observations.

g. **(4 marks)** Test the assumptions of normality and constant variance in the errors. Do the results confirm the conclusions you reached in part (f) about these assumptions? In your response, include the hypotheses being tested in each test.

We test the null hypothesis : H0: The residuals come from a normal distribution.
Alternative hypothesis: H1: The residuals do not come from a normal distribution.

Shapiro-Wilk test

```
shapiro.test(fit1$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$res
## W = 0.99445, p-value = 2.191e-05
```

As the p-value < 0.05, we reject the null hypothesis. There is no evidence that the residuals come from a normal distribution. The result confirm the conclusion reached in part (f) about the assumption MLR6 of Normality.

Kolmogorov-Smirnov test

```
ks.test(fit1$res, "pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fit1$res
## D = 0.30779, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

The Kolmogorov-Smirnov test also has the same result with a p-value < 0.05. So we reject the null hypothesis. There is no evidence that the residuals come from a normal distribution.The result confirm the conclusion reached in part (f) about the assumption MLR6 of Normality.

Breusch-Pagan test

We test the null Hypothesis (H0): Homoscedasticity is present.
Alternative Hypothesis (H1): Heteroscedasticity is present.

```
library(lmtest)
bptest(fit1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit1
## BP = 177.87, df = 11, p-value < 2.2e-16
```

As the −value of the test is less than the significance level (= 0.05), we reject the null hypothesis. We conclude that heteroscedasticity is present in the regression model.The result confirm the conclusion reached in part (f) about the assumption MLR5 of Homoscedasticity.

h. **(2 marks)** Use the VIF statistic to check whether or not there is evidence of severe multicollinearity among the predictors. Comment on the results.

```
library(car)
library(knitr)
pander(vif(fit1), digits=2, caption="VIF values")
```

Table 5: VIF values

	GVIF	Df	GVIF^(1/(2*Df))
frequency	1.2	1	1.1
angle	4.6	1	2.1
chord.length	2.2	5	1.1
speed	1.1	3	1
displace	3.6	1	1.9

There is no evidence of severe multicollinearity, since all VIF values are less than 10. The largest VIF value tells us that the variance of the angle coefficient is inflated by a factor of 4.6 because angle is highly correlated with at least one of the other predictors in the model. However,this potential multicollinearity problem is not worth worrying about as the coefficient of angle variable is statistically significant at = 5% (result in part (c)).

i. **(3 marks)** Based on a global usefulness test, is it worth going on to further analyse and interpret a model of decibels against each of the predictors? Carry out the test, give the conclusion and justify your answer.

The global usefulness test:

> We test the global hypothesis : H0: Beta0 =Beta1 =…=Beta11 = 0
>
> Against the alternative: H1: at least one Betaj is non-zero, for j = 1, … , 11.

```
summary(fit1) #Conduct the F-statistic and the p-value
```

```
##
## Call:
## lm(formula = decibels ~ frequency + angle + chord.length + speed +
##     displace, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6037  -2.8999  -0.2024   3.1307  15.8329
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.358e+02  5.447e-01 249.368  < 2e-16 ***
## frequency            -1.295e-03  4.252e-05 -30.455  < 2e-16 ***
## angle              -4.700e-01  4.446e-02 -10.571  < 2e-16 ***
## chord.length0.0508 -1.603e+00  4.490e-01  -3.570 0.000369 ***
## chord.length0.1016 -2.951e+00  5.103e-01  -5.783 8.93e-09 ***
## chord.length0.1524 -6.551e+00  5.039e-01 -13.000  < 2e-16 ***
## chord.length0.2286 -7.710e+00  4.954e-01 -15.563  < 2e-16 ***
## chord.length0.3048 -1.022e+01  5.596e-01 -18.264  < 2e-16 ***
## speed39.6           7.997e-01  3.621e-01   2.208 0.027374 *
## speed55.5           2.252e+00  4.062e-01   5.545 3.47e-08 ***
## speed71.3           4.077e+00  3.711e-01  10.987  < 2e-16 ***
## displace           -1.257e+02  1.778e+01  -7.070 2.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.773 on 1491 degrees of freedom
## Multiple R-squared:  0.5249, Adjusted R-squared:  0.5214
## F-statistic: 149.7 on 11 and 1491 DF,  p-value: < 2.2e-16
```

```
qf(0.05, 11, 1491, lower.tail = TRUE) #Find the critical F-value
```

```
## [1] 0.4152767
```

In this case, the F-statistic is 149.7 on 11 and 1491 degrees of freedom, and p-value is <2.2*10^-16. The F-statistic is much higher than the critical F-value at = 5% (149.7>0.415) and the p-value is lower than (2.2*10^-16 < 0.05).

Conclusion: we can reject the null hypothesis and we have statistically significant evidence at =0.05 to show that there is at least one of the predictors is related to the dependent variable (decibels). This also means we can confidently proceed with additional analysis to explore the relationship between Decibels and the predictors, given the assurance that there is a meaningful association worth investigating.

Q2.

a. **(3 marks)** Read the data into R and fit a linear model for height with the variables father, mother, gender, kids and midparent as predictors. Provide a summary of the fitted model. You will notice that estimates for midparent are listed as NA. Why might this be the case and what regression problem does this point to?

```
#Read the data
dataq2 <- read.csv("galton.csv", header = TRUE)
str(dataq2)
```

```
## 'data.frame':    898 obs. of  8 variables:
##  $ familyID : chr  "1" "1" "1" "1" ...
##  $ father   : num  78.5 78.5 78.5 78.5 75.5 75.5 75.5 75.5 75 75 ...
##  $ mother   : num  67 67 67 67 66.5 66.5 66.5 66.5 64 64 ...
##  $ gender   : chr  "M" "F" "F" "F" ...
##  $ height   : num  73.2 69.2 69 69 73.5 72.5 65.5 65.5 71 68 ...
##  $ kids     : int  4 4 4 4 4 4 4 4 2 2 ...
##  $ midparent: num  75.4 75.4 75.4 75.4 73.7 ...
##  $ adltchld : num  73.2 74.7 74.5 74.5 73.5 ...
```

We can see the predictor gender only has two values,“F” for female and “M” for male. So, This predictor should be treated as categorical variable.

```
dataq2$gender <- factor(dataq2$gender) #convert gender into categorical variables

fit2<-lm(height ~ father + mother + gender +    kids +  midparent, data=dataq2) ##Fit the model
pander(summary(fit2))
```

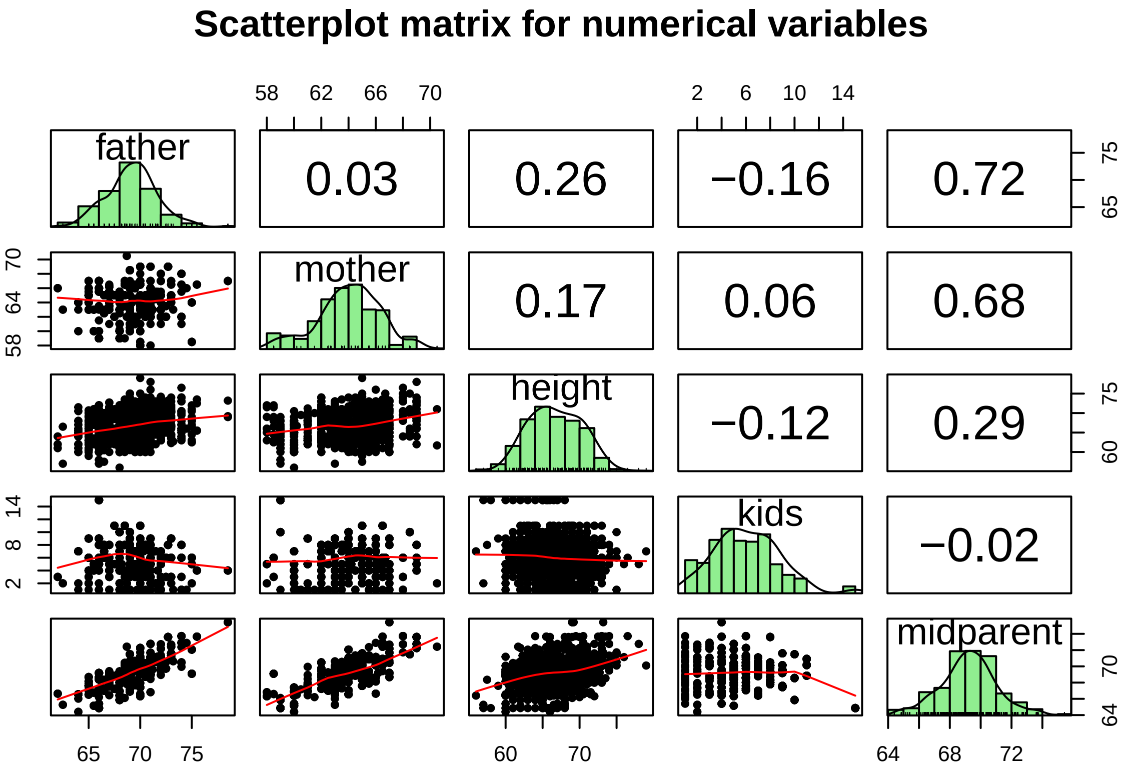
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.19	2.794	5.794	9.522e-09
father	0.3983	0.02957	13.47	8.608e-38
mother	0.321	0.03126	10.27	1.85e-23
genderM	5.21	0.1442	36.12	7.584e-177
kids	-0.04382	0.02718	-1.612	0.1073

Table 7: Fitting linear model: height ~ father + mother + gender + kids + midparent

Observations	Residual Std. Error	R^2	Adjusted R^2
898	2.152	0.6407	0.6391

We observe that the estimates for midparent are indicated as NA. This issue occurs maybe because of multicollinearity - when two predictors are strongly related. To confirm this hypothesis, we will substantiate it through the scatterplot matrix for numerical variables.

```
dataq2%>%
dplyr::select(father, mother, height,   kids,   midparent)%>% #select numerical variables (includes int
pairs.panels(method = "spearman", # correlation method
hist.col = "lightgreen", # histogram color
density = TRUE, # show density plots
ellipses = FALSE,# do not show correlation ellipses
main = "Scatterplot matrix for numerical variables"
)
```



Based on the scatterplot matrix, it is evident that the midparent variable has moderate positive correlation with both the father and mother predictor variables, with correlation coefficients of 0.72 and 0.68, respectively.This means that when the father or mother variable increases, midparent also tends to increase. Consequently, it becomes challenging to discern the individual associations of these three predictors with height because when one changes the other predictor also changes.

b. **(2 marks)** What action might you take to resolve the problem identified in part (a)?

We have two possible actions to sovle the multicolliearity issue:

- Combine the collinear predictors together into a single predictor.
- Remove one of the collinear predictors from the model.

The midparent variable is calculated as (father + 1.08*mother)/2, which means that midparent is the combination of father and mother predictors. So, if we want to keep the combination midparent, we have to remove both father and mother variables. However, the removed predictor may contain unique information that is relevant to the response variable, leading to a loss of predictive power. So, I suggest that we should only remove the midparent predictors.

c. **(2 marks)** Based on the model fitted in part (a) give an interpretation of the coefficient for genderM.

The coefficient for the variable genderM represents the difference in the mean height between males (genderM = 1) compared to females (genderM = 0), while holding all other predictors constant. The coefficient for genderM is 5.21 indicating that, on average, males are predicted to be approximately 5.21 units taller than females.The p-value of 7.584e-177 suggests that this coefficient is statistically significant (p-value < 0.05), indicating that gender is a significant predictor of height in this model.

d. **(2 marks)** Determine the number of families in the dataset.

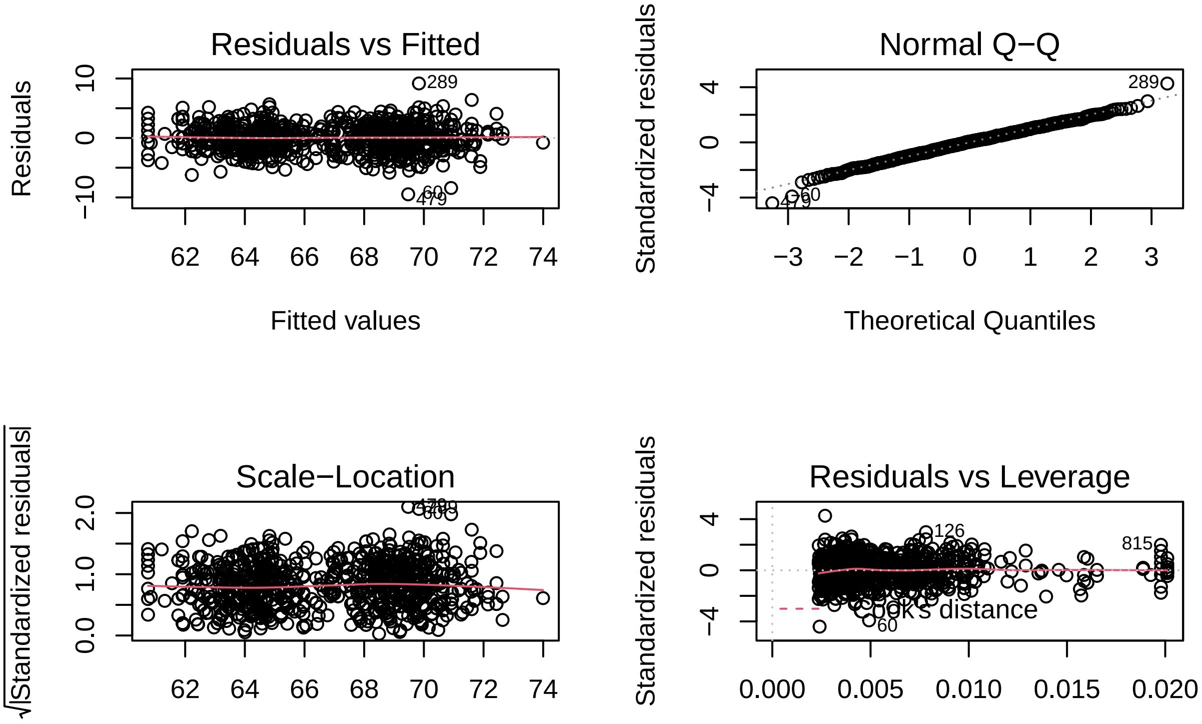
```r
length(unique(dataq2$familyID))
```

```
## [1] 197
```

There are 197 families in the dataset.

e. **(3 marks)** The problem in part (a) is resolved and a new linear model is fitted.No observations are excluded. The plots below are obtained to investigate regression assumptions for this new model. Based on your answer in part (d) and the plots below, do the data meet all the regression assumptions? Explain your answer briefly.

```r
par(mfrow=c(2,2))
plot(fit2)
```



- **Linearity: Residuals vs fitted plot**

The plot shows that the residuals are roughly equally spread around a horizontal line without a distinct pattern. This means that there are no uncaptured significant non-linear relationships and it satisfies the first MLR assumption about linearity in parameters.

- **Normality: Normal Q-Q plot**

There are only three out of 898 data points do not lie on the straight line. Due to the presence of a few data points that deviate from the straight line, we say that there is some evidence of non-normality. However, this can occur even if the normality assumption (MLR6) holds. Hence, it's advisable to conduct additional analyses to thoroughly assess the normality of the residuals.

- **Equal variance (homoscedasticity): Scale-Location plot**

The red line of the S-L plot is approximately horizontal and the spread of data points around the red line has insignificant change as the fitted values change. Therefore, There is no evidence of non-constant variance (heteroscedasticity), which does not violate the MLR5 assumption.

- **Influential observations: Residuals vs leverage plot**

The residuals vs leverage plot suggests that there are no highly influential observations.

- **There are 197 families in the dataset**

The result of part (d) indicates that observations might be related to each other due to children in the same family being siblings, highlights a violation of the assumption of random sampling in Multiple Linear Regression (MLR).

- **Conclusion** All in all, the dataset does not meet all the regression assumptions. It violates the assumption of normality of residuals (MLR6) and random sampling (MLR2).

Assignment total: 40 marks