DATA 303 Assignment 1

Izzy Southon, 300597453

Assignment Questions

Q1. (28 marks) We will carry out a regression analysis to investigate the relationship between the response variable `decibels` and the other variables in the dataset as predictors.

```
# Read the csv file and save as "airfoil"
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
airfoil <- read.csv("airfoil_self_noise.csv")
head(airfoil)
```

```
##   frequency angle chord.length speed   displace decibels
## 1       800     0       0.3048  71.3 0.00266337  126.201
## 2      1000     0       0.3048  71.3 0.00266337  125.201
## 3      1250     0       0.3048  71.3 0.00266337  125.951
## 4      1600     0       0.3048  71.3 0.00266337  127.591
## 5      2000     0       0.3048  71.3 0.00266337  127.461
## 6      2500     0       0.3048  71.3 0.00266337  125.571
```

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.

```
## EDA
## Check the structure of the dataset
str(airfoil)
```

```
## '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 ...
```

```
## Convert `chord.length` and `speed' to categorical variables as they are currently numerical
airfoil$chord.length<-as.factor(airfoil$chord.length)
airfoil$speed<-as.factor(airfoil$speed)

## Get the summary statistics
summary(airfoil)
```

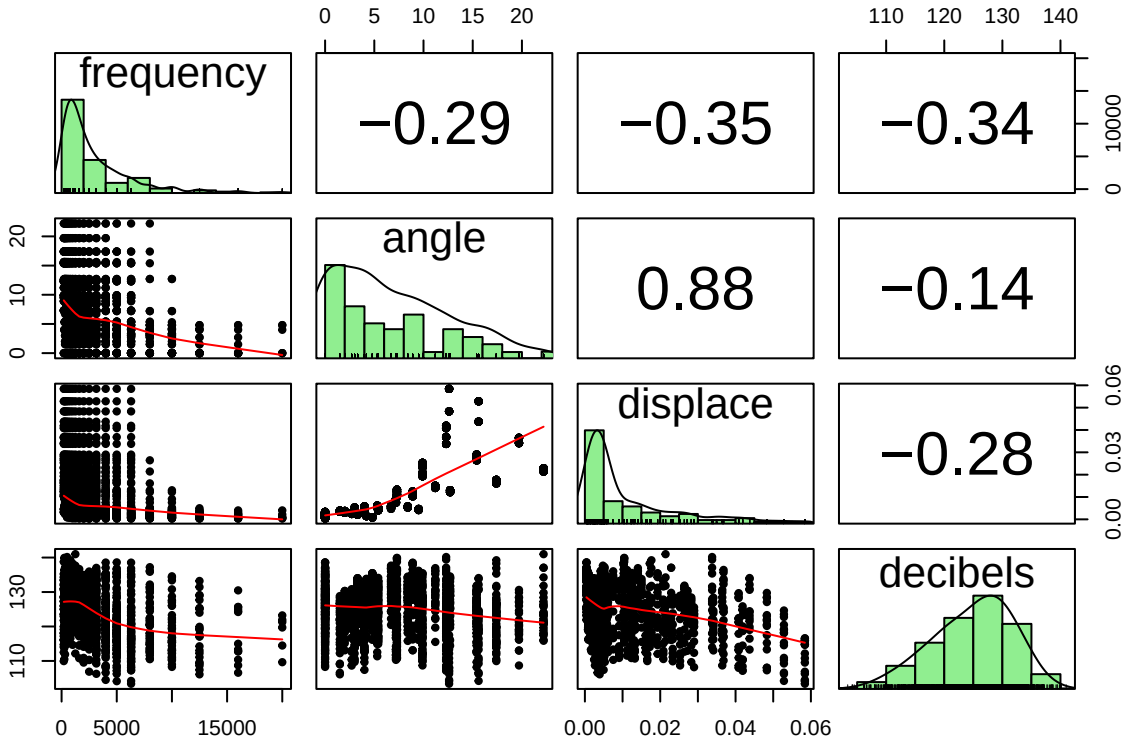
```
##    frequency         angle        chord.length  speed        displace
##  Min.   :  200   Min.   : 0.000   0.0254:278   31.7:281   Min.   :0.0004007
##  1st Qu.:  800   1st Qu.: 2.000   0.0508:237   39.6:480   1st Qu.:0.0025351
##  Median : 1600   Median : 5.400   0.1016:263   55.5:277   Median :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
```

```
## Double checking there are no missing values
any(is.na(airfoil))
```

```
## [1] FALSE
```

```
## Graphical summaries
## Scatterplot matrix for numerical variables
library(dplyr)
library(psych)

airfoil%>%
  dplyr::select(where(is.numeric))%>%
  pairs.panels(method = "spearman",
             hist.col = "lightgreen",
             density = TRUE,
             ellipses = FALSE
)
```

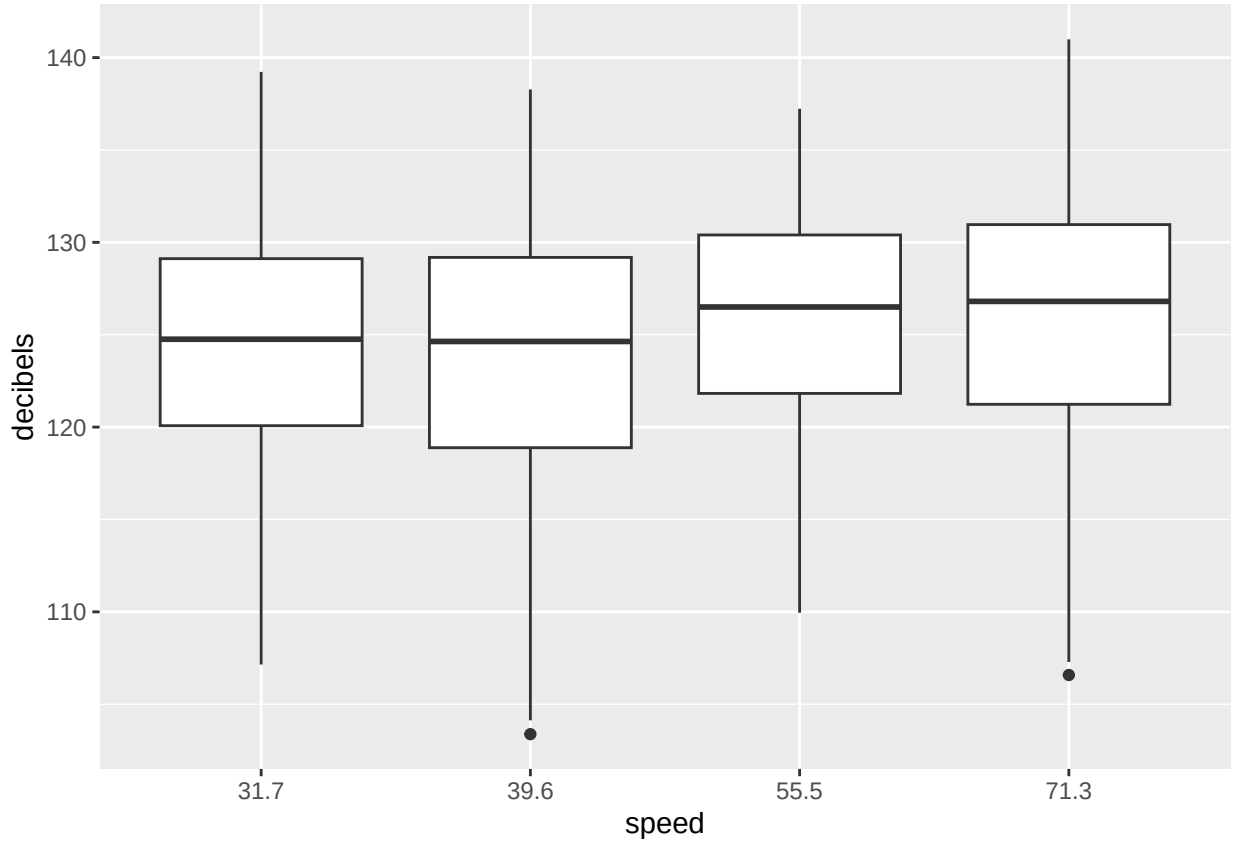



```
## Boxplots of response variable (decibels) by category for categorical predictors (speed and chord.len
library(ggplot2)
```

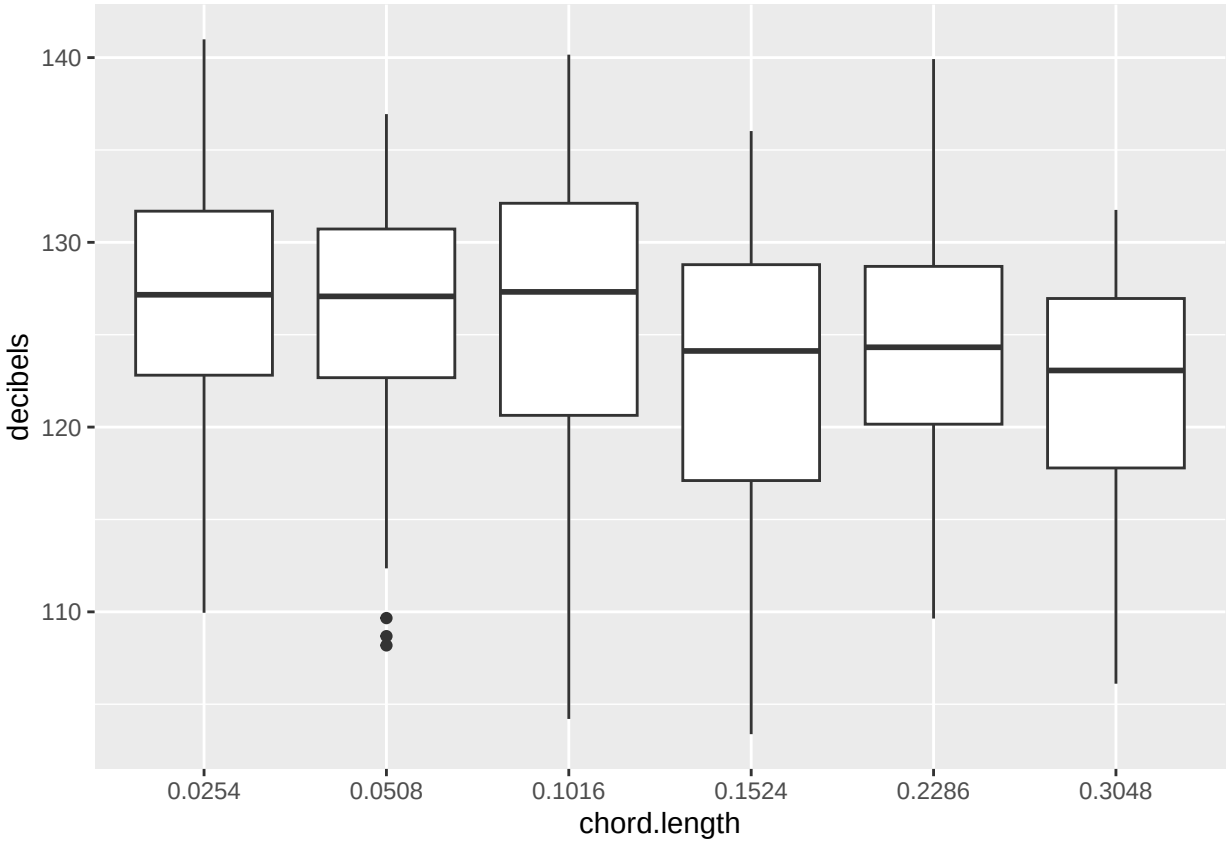
```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
airfoil%>%
  dplyr::select(decibels, where(is.factor))%>%
  ggplot(aes(x=speed,y=decibels))+
  geom_boxplot()
```




```
airfoil%>%
  dplyr::select(decibels, where(is.factor))%>%
  ggplot(aes(x=chord.length,y=decibels))+
  geom_boxplot()
```



Key points or notes from my EDA:

- There are 1503 observations (rows) in the dataset and 6 variables (so 5 predictors). Originally all of the variables were numbers or integers, but chord.length and speed have been converted to categorical variables using as.factor.
- There are no missing values in columns.
- The minimum and maximum for numerical variables seem sensible.
- As expected there are 281+480+277+465=1503 observations for speed which is as expected.
- All variables will be kept in the dataset for regression analysis.
- The median decibels seem to increase as speed increases.
- Decibels decreases the more frequency in hertz decreases.
- Decibels decreases when the angle of attack decreases.
- There are outliers for 0.0508 chord.length, where points are at lower decibels
- The distribution of the response variable decibels looks relatively symmetric
- There is some potential for non-linearity between decibels and some numerical predictors, so another kind of regression such as polynomial could be considered.
- Multicollinearity shouldn't be an issue because there are low correlations between some pairs of predictors, but it should be checked out

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 σ^2, the error variance.

```
## Fit the model: decibels = E(decibels) + E = B0 + B1frequency + B2angle + B3chord.length + B4speed + B
fit1<-lm(decibels ~ frequency + angle + chord.length + speed + displace, data = airfoil)
library(pander)
pander(summary(fit1), caption="")
```

	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

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

The fitted model equation is: $Y = decibels$

- $decibes = E(decibels) + \epsilon = \beta_0 + \beta_1 frequency + \beta_2 angle + \beta_3 chord.length + \beta_4 speed + \beta_5 displace + \epsilon$
- $E(\widehat{decibels}) = \widehat{decibels} = \widehat{\beta_0} + \widehat{\beta_1} frequency + \widehat{\beta_2} angle + \widehat{\beta_3} chord.length + \widehat{\beta_4} speed + \widehat{\beta_5} displace$
- $\widehat{decibels} = 135.8 - 0.0013 frequency - 0.47 angle - 1.6 chord.length_{0.0508} - 2.95 chord.length_{0.1016} - 6.55 chord.length_{0.1524} - 7.7 chord.length_{0.2286} - 10.2 chord.length_{0.3048} + 0.79 speed_{39.6} + 2.25 speed_{55.5} + 4.08 speed_{71.3} - 125.7 displace$
- $\sigma^2 = 4.773$

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

- `angle`: The estimate β_2 = -0.47 represents the change in expected decibels associated with a 1-unit increase in angle when all other predictors are held constant. When angle increases by 1 degree there is an associated reduction in expected decibels of -0.47, when all other predictors are kept constant.
- `speed71.3`: The table of coefficient estimates gives an estimate for speed39.6. This indicates that the speed31.7 is used as a reference level and $\widehat{\beta_8}$ = 0.7997 is the estimated difference, when keeping all other predictors constant. We interpret this to mean the expected decibels for speed39.6 was higher than the expected decibels for speed31.7 by an estimated 0.7997 decibels when comparing aircrafts of equal frequency, angle, chord.length, and displace. The coefficient estimate for speed71.3 is 4.077, meaning that the expected decibels for speed71.3 are estimated to be higher than the expected decibels for speed31.7 by an estimated 4.077 decibels, holding all other predictors constant.

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

It may make practical sense to estimate decibels when all predictor values are zero, we don't have any sample data close to zero for chord.length0.2286, chord.length0.3048, and displace. Therefore it is not appropriate to interpret B0 = 135.8 in practical terms.

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.

```
## Select the last 3 rows and excludes decibels
xdata <- tail(subset(airfoil, select = -decibels), 3) # decibels is not a predictor (it is the response

##Confidence intervals for estimating the mean response
pander(predict(fit1, newdata=xdata, 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

```
##Prediction intervals for predicting the response for new predictor values
pander(predict(fit1, newdata=xdata, 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

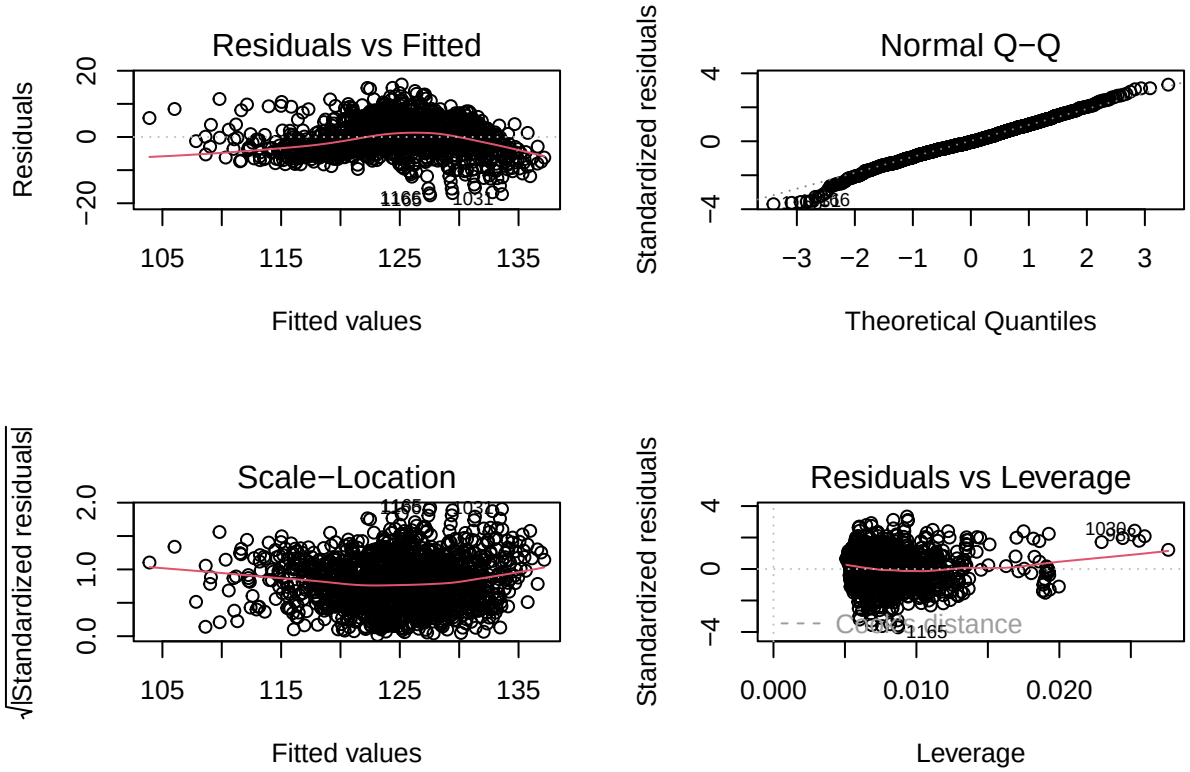
The 95% confidence interval for quantifying uncertainty around estimating $E[Y|X]$ for aircrafts with the same characteristics as the last aircraft in the data set is (110.5, 112.6). This means that 95% of intervals of this form will contain the true average aircraft decibels for such aircrafts.

For one such aircraft, the 95% prediction interval for quantifying uncertainty around prediction of the true value Y is (102.1, 121). This means that 95% of intervals of this form will contain the true aircraft decibel for this particular aircraft.

Both intervals are centred at 111.5, but the prediction interval is substantially wider than the confidence interval, reflecting greater uncertainty about the individual decibels of a given aircraft compared to uncertainty about the average decibels over many aircrafts.

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.

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



Residuals vs Fitted plot: There looks to be a linear relationship between the predictor variables and the response varaible, as the residuals are relatively equally spread around the horizontal line without a distinct pattern. This is good indication that there are no uncaptured signficant non-linear relationships.

Normal Q-Q plot: Residuals seem to be normally distributed as residuals do not stray too far from the Q-Q line. Residuals follow a straight line well. A few residuals deviate from the Q-Q line at the beginning and end, however this can happen even if the normality assumption is met.

Scale-Location plot: The residuals appear roughly equally spread along the range of fitted values for the most part, however at smaller fitted values the residuals show a slightly wider spread. The red line decreases then increases again, but by a slight slope.

Residuals vs Leverage plot: There are no highly influential observations as Cook's distance lines (a red dashed line) isn't visible because all cases are well inside of the Cook's distance thresholds. This means that there are no influential observations in the airfoil dataset.

Since there could be some non-linearity, non-normality, and non-constant variance, in further analyses transformations of the response and predictor variables should be considered. The observations of 1031 and 1166 have been labelled as being potentially problematic in 3 out of 4 plots. Observation 1165 has been labelled as potentially problematic in all plots. It may be worth looking at the aircrafts corresponding to these observation numbers to see if there is anything special about them.

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.

Normality assumption

Hypotheses:

- Null hypothesis (H_0) : The sample comes from a normal distribution.
- Alternative hypothesis (H_1) : The sample does not come from a normal distribution.

Since there are 1503 observations, the Shapiro-Wilk test can be done as the sample size is less than 5000.

```
shapiro.test(fit1$res) ## Shapiro-Wilk test
```

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

```
ks.test(fit1$res, "pnorm") ## Kolmogorov-Smirnov test
```

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

The test statistic of the Shapiro-Wilk test is 0.995 (3dp), and the p-value of 2.19 x 10^5 is less than any significance level, therefore in this test we reject the null hypothesis and conclude there is no evidence that the sample comes from a normal distribution.

The test statistic of the K-S test is 0.308 (3dp), and the p-value of 2.2 x 10^-16 is less than any significance level, therefore in this test we reject the null hypothesis and conclude there is no evidence that the sample comes from a normal distribution.

Both tests show that there is no evidence that the sample data comes from a normal distribution.

Equal (constant) variance assumption

Hypotheses:

- Null hypothesis (H_0) : Homoscedasticity is present (the residuals are distributed with equal variance).
- Alternative hypothesis (H_1) : Heteroscedasticity is present (the residuals are not distributed with equal variance)

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(fit1) ##Breusch-Pagan test
```

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

For the Breusch-Pagan test, the test statistic is 177.87 with 11 degrees of freedom, the p-value of 2.2 x 10^-16 is less than any significance level, therefore we reject the null hypothesis and conclude that heteroscedasticity is present in the regression model. This means that the residuals are not distributed with constant variance.

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.

Variance Inflation Factor (VIF)

$VIF(\widehat{\beta}_j) = \frac{1}{1-R^2_x j|x-_j}$,

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(knitr)
library(pander)
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 coeffcient is inflated by a factor of 4.6 because angle is highly correlated with at least one of the other predictors in the model.

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.

Hypotheses:

- $(H_0) : \beta_1 = \beta_2 ... = \beta_p = 0$
- (H_1) : at least one β_j is non-zero, for $j = 1, ..., p$.

```
summary(fit1)
```

```
##
## Call:
## lm(formula = decibels ~ frequency + angle + chord.length + speed +
##     displace, data = airfoil)
##
## 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
```

We find $F = 149.7$ with 11 and 1491 degrees of freedom, and a p-value of 2.2 x 10^-16. We therefore have very strong evidence to reject H_0 and conclude that there is no evidence that all regression coefficients are zero in the population. This means we can go on to further analyse and interpret the model of decibels against the eleven predictors. It is worth going on further to analyse and interpret a model of `decibels` against each of the predictors because if the coefficients were all zero, the model would contain little useful information about the association between the response variable (decibels) and the predictors.

Q2. (12 marks) Francis Galton's 1866 dataset (cleaned) lists individual observations on height for 899 children. Galton coined the term "regression" following his study of how children's heights related to heights of their parents. The data are available in the file `galton.csv` and contain the following variables:

- `familyID`: Family ID

- `father`: Height of father
- `mother`: Height of mother
- `gender`: gender of child
- `height`: Height of child
- `kids`: Number of childre in family
- `midparent`: Mid-parent height calculated as (‘father + 1.08*mother)/2
- `adltchld`: `height` if `gender`=M, otherwise 1.08*`height` if `gender`= F

All heights are measured in inches.

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?

```
galton <- read.csv("galton.csv")
head(galton)
```

```
##   familyID father mother gender height kids midparent adltchld
## 1        1   78.5   67.0      M   73.2    4     75.43   73.200
## 2        1   78.5   67.0      F   69.2    4     75.43   74.736
## 3        1   78.5   67.0      F   69.0    4     75.43   74.520
## 4        1   78.5   67.0      F   69.0    4     75.43   74.520
## 5        2   75.5   66.5      M   73.5    4     73.66   73.500
## 6        2   75.5   66.5      M   72.5    4     73.66   72.500
```

```
fit2<-lm(height ~ father + mother + gender + kids + midparent, data = galton)
library(pander)
pander(summary(fit2), caption="")
```

	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

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

The estimates for ‘midparent’ may be listed as ‘NA’ because it is using the predictor variables of ‘mother’ and ‘father’ to make its calculation for the ‘midparent’ variable. The ‘midparent’ predictor may have a strong linear correlation to one or more other predictors (multicollinearity). This is because the variable ‘midparent’ is calculated using the heights of the ‘mother’ and ‘father’ variables, and is then being used as another predictor being ‘midparent’. Multicollinearity is a problem in regression because it produces less reliable regression model results due to wider confidence intervals, that can lower the significance of regression coefficients.

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

To resolve the problem of multicollinearity identified in part a. We could remove one of the variables (either mother or father), however that wouldn't make much sense, so it would probably work best to remove both and combine them into a single variable (being 'midparent' in this case). Additionally, a dimensionality reduction technique such as PCA could be conducted to reduce the number of predictors while retaining most of the information needed.

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

The coefficient estimate for $\widehat{\beta}_3 = 5.21$ represents the expected difference in height between males and females. When all other predictors are held constant, males are associated with an expected increase in height of 5.21 inches compared to females in the galton dataset.

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

```
length(unique(galton$familyID))
```

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

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.

Residuals vs Fitted: There is no evidence for a curved pattern, therefore indicating that there looks to be a linear relationship between the predictors and the response variable (height). The residuals have a relatively even spread around the horizontal line, therefore indicating that the regression assumption of linearity has not been violated.

Q-Q Residuals: Residuals do not stray far from the Q-Q line, therefore indicating that the residuals seem to be normally distributed. A few residuals deviate from the Q-Q line at the beginning and end, although this can happen even if normality has not be violated.

Scale-Location: The residuals are relatively equally spread along the range of fitted values, therefore indicating that the assumption of equal variance has not been violated and seems to have been met.

Residuals vs Leverage: There are no highly influential points as all observations are well inside of the Cook's distance thresholds in the galton dataset. This is evident as the Cook's distance lines are barely visible on the plot.

Additionally, row 60 has shown up as a potential problem observation on all four plots. Therefore a closer look of it should be taken to see if there is anything special about it. Overall, the data has not violated the regression assumptions of linearity, normality, equal variance, and having any influential observations.