School of Mathematical Sciences

Assessment Cover Sheet

Student Names	Thomas Carey, Nikky Gleeson, Andrew Martin, James Schoff, and Joshua Yates
Student IDs	1704570, 1707382, 1704466, 1705565, 1706458
Assessment Title	**Group Project**
Due Date	**Monday 4th June by noon**
Course	**STATS 3001 Statistical Modelling**
Tutorial Group Number	**Thurs 1pm**
Date Submitted	**4th June 2018**
Lecturer	**Professor Patty Solomon**

KEEP A COPY
It is a good idea to keep a copy of your work for your own records

PLAGIARISM AND COLLUSION
Plagiarism: using another person's ideas, designs, words or works without appropriate acknowledgement.
Collusion: another person assisting in the production of an assessment submission without the express requirement, or consent or knowledge of the assessor.

NB: In this course you are encouraged to work with other students but the work you submit must be your own. This means you must understand it and be able to explain it if required.

CONSEQUENCES OF PLAGIARISM AND COLLUSION
The penalties associated with plagiarism and collusion are designed to impose sanctions on offenders that reflect the seriousness of the University's commitment to academic integrity. Penalties may include: the requirement to revise and resubmit assessment work, receiving a result of zero for the assessment work, failing the course, expulsion and/or receiving a financial pen

I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others. I have read the University Policy Statement on Plagiarism, Collusion and Related Forms of Cheating (http://www.adelaide.edu.au/policies/?230).
I give permission for my assessment work to be reproduced and submitted to other academic staff for the purposes of assessment and to be copied, submitted and retained in a form suitable for electronic checking of plagiarism.

Signed…… Date: 4th June 2018

Signed…… Date: 4th June 2018

Signed…… Date: 4th June 2018

Signed…… Date: 4th June 2018

Signed……Joshua Yates…… Date: 4th June 2018

Statistical Modelling III Group Project

Thomas Carey, Nikky Gleeson, Andrew Martin,
James Schoff & Joshua Yates

**1704570, 1707382, 1704466,
1705565, 1706458**

June 4, 2018

Report submitted for **STATS 3001** at the School of Mathematical Sciences, University of Adelaide

Project Area: **Multiple Linear Regression and Logistic Regression**
Project Supervisor: **Patty Solomon**

In submitting this work I am indicating that I have read the University's Academic Honesty Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

1 Part A: Multiple Linear Regression

1.1 Introduction

The first section of this project investigates whether it is possible to predict the required length of a catheter for children. Catheters are inserted into a major vein or artery and pushed into the heart in order to obtain information about heart physiology and function. This heart catheterisation is performed on children with congenital heart defects. To predict the length of a catheter, the height, weight, and catheter length of 12 children was analysed in order to formulate a suitable model. This data can be seen below in Figure 1.

Child	Height (cm)	Weight (kg)	Length (cm)
1	108.7	18.14	37.0
2	161.29	42.41	49.5
3	95.25	16.10	34.5
4	100.33	13.61	36.0
5	115.57	23.59	43.0
6	97.79	7.71	28.0
7	109.22	17.46	37.0
8	57.15	3.86	20.0
9	93.98	14.97	33.5
10	59.69	4.31	30.5
11	83.82	9.53	38.5
12	147.32	35.83	47.0

Figure 1: The Height, Weight, and catheter Length of 12 children.

1.2 Relationship between Variables

The data given in Figure 1 was converted into an excel spreadsheet and subsequently read into R.

The relationship between the three recorded variables `Length`, `Height`, and `Weight` was then investigated. This was done through observation of the pairwise scatter-plot matrix and correlation matrix. The pairwise relationships and the correlation between each variable can be seen in Figure 2.

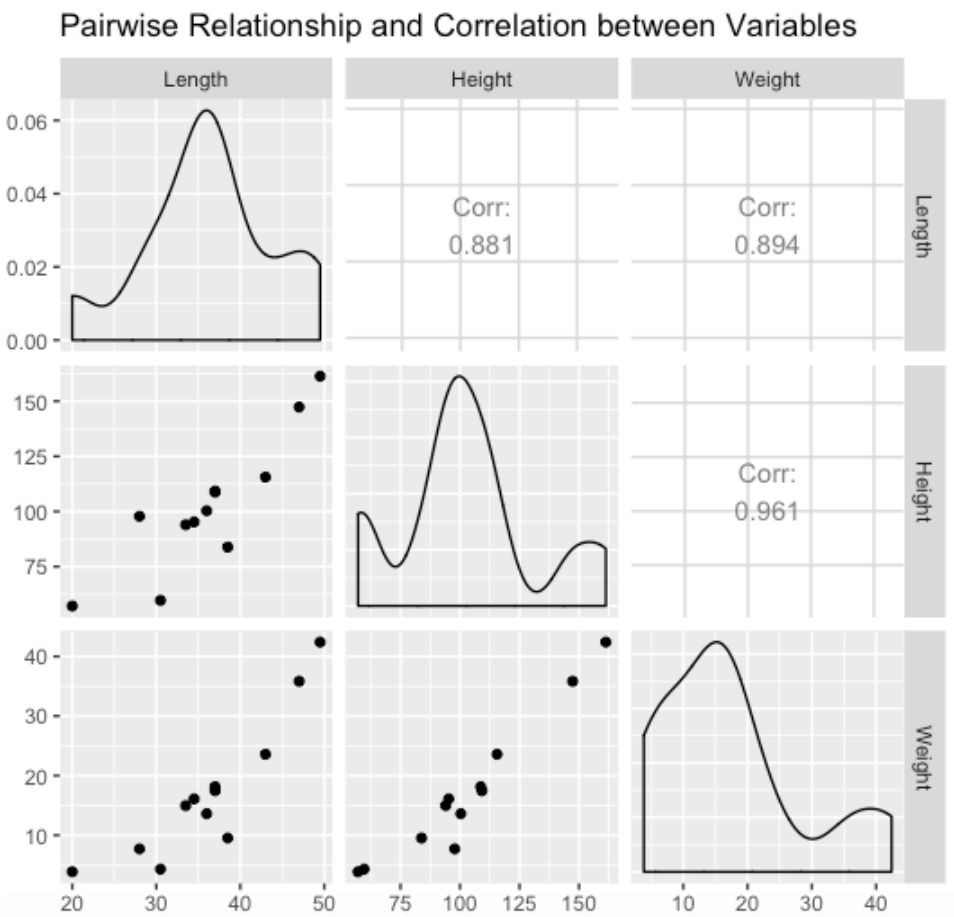


Figure 2: Pairwise Scatterplots and Correlation between Variables

Observation of the pairwise plots shows that there is a moderate-to-strong positive linear relationship between the predictor variable `Height` and the response variable `Length`.
Similarly, there is also a moderate-to-strong positive linear relationship between the predictor variable `Weight` and the response variable `Length`.
There is a strong positive linear relationship between the predictor variables `Weight` and `Height`.

These relationships can be further evidenced through observation of the correlation values, in which all variables demonstrate a relatively high correlation. The highest correlation is between the predictor variables `Weight` and `Height`, with a correlation of 0.961. The correlation between `Height` and `Length` is 0.881 and the correlation between `Weight` and `Length` is 0.894.

1.3 Model Fitting and Assumption Checking

In order to determine the most suitable model for predicting catheter length, three different linear models were fitted. These models are:

`Length` $\sim$ `Height` + `Weight`
`Length` $\sim$ `Height`
`Length` $\sim$ `Weight`

1.3.1 Assumption Checking for the Multiple Regression Model

For multiple regression models, there are four main assumptions that must be checked, three of which are checked using the residual plots. These three are *Linearity, Homoscedasticity* and, *Normality.*

The fourth is *independence*, however it cannot be checked using the residuals and must be checked through analysis of the experimental design.

Another assumption that must be checked is that of *points of high leverage and high residual.*

In order to check these assumptions, the appropriate diagnostic plots need to be analysed. These plots can be seen in Figures 3 and 4 below:

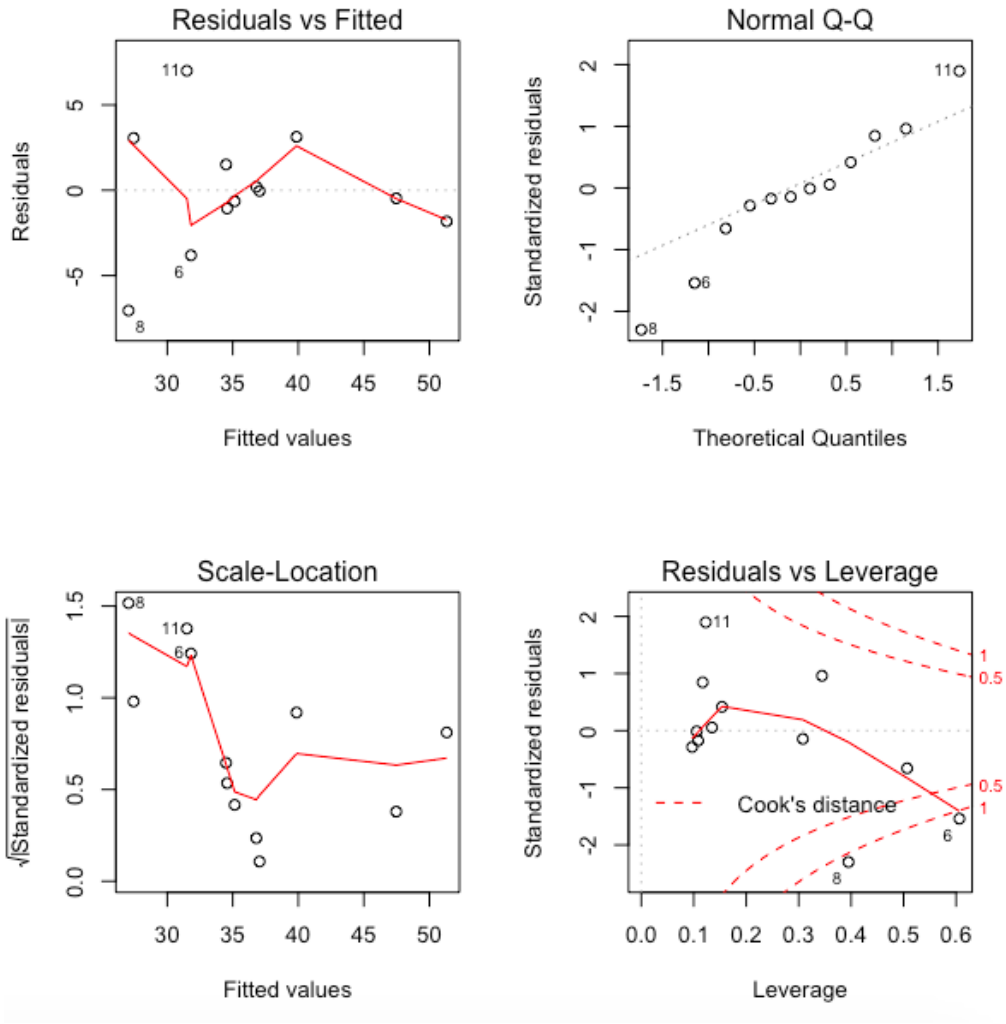
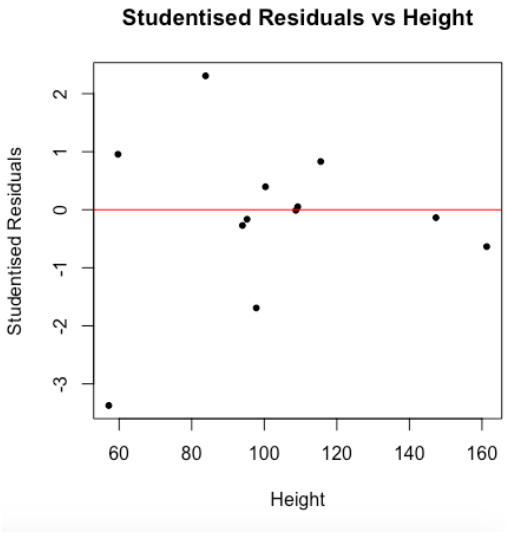
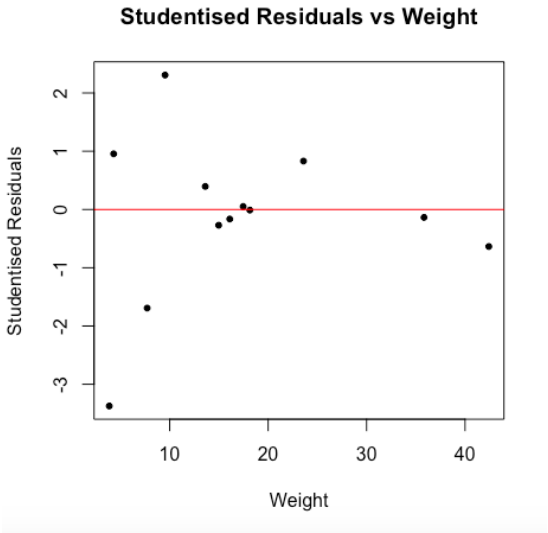


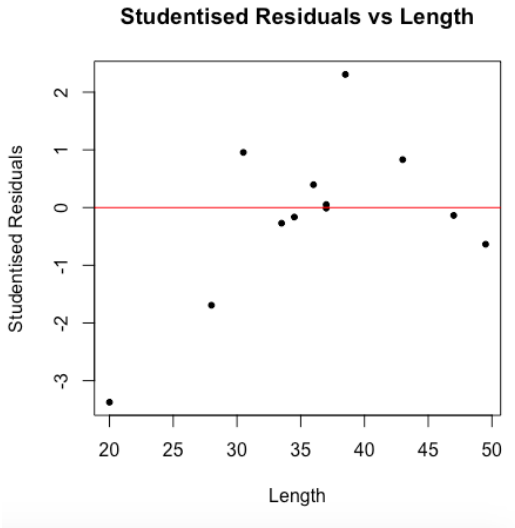
Figure 3: Diagnostic Plots



(a) Studentised Residuals vs Height



(b) Studentised Residuals vs Weight



(c) Studentised Residuals vs Length

Figure 4: Diagnostic plots of Studentised Residuals

Linearity: To check the assumption of linearity, it is necessary to look at the residuals vs fitted values plot and the individual studentised residuals vs each predictor variable plots. If the assumption holds, we expect to see random scatter about the zero-line in each of these plots. The residuals vs fitted values plot shows roughly random scatter with no significant curvature, similarly, the individual residuals vs each predictor variable plots all display no obvious curvature. So the assumption of linearity is reasonable.

Homoscedasticity: To check the assumption of homoscedasticity, it is necessary to look at the scale-location plot and the individual studentised residuals vs each predictor variable plots. If the assumption holds, we expect to see data that is symmetrically distributed about the zero-line in each of these plots. The scale-location plot suggests mild heteroscedasticity, and the plot of the studentised residuals vs length also shows mild heteroscedasticity, whilst the studentised residuals vs height plot and the studentised residuals vs weight plots show scatter which looks roughly constant. So the assumption of homoscedasticity is not quite reasonable.

Normality: To check the assumption of normality, it is necessary to look at the normal quantile plot. If the assumption holds, we expect to see a linear relationship. The normal quantile plot shows no significant departures from normality. So the assumption of normality is reasonable.

Independence: As the catheter length recorded for one child does not affect the catheter length recorded for any other child, the assumption of independence is considered reasonable in this case.

Points of High Leverage and High Influence: To check if there are any points of high leverage or high influence, it is necessary to look at the residuals vs leverage plot. As points 6 and 8 lie outside of the Cook's distance contours of 1, these are considered influential points.

Overall: Apart from mild heteroscedasticity and two points of high leverage, the assumptions of the multiple regression model appear reasonable.

1.4 Comparing Linear Regression Models

1.4.1 Comparing Regression Coefficients

The regression was performed using the code in the appendices.
The estimated coefficient for `Height` in the simple linear regression is 0.235 and is highly significant, while the estimated coefficient for `Height` in the multiple linear regression is 0.077 and is not significant.

The estimated coefficient for `Weight` in the simple linear regression is 0.611 and is highly significant, whilst the estimated coefficient for `Weight` in the multiple linear regression is 0.421 and is not significant.

1.4.2 Interpretation of the Coefficient Weight

In the context of the simple linear regression, the mean catheter `Length` increases by 0.611 for each one unit increase in `Weight` (where catheter `Length` is measured in cm and `Weight` is measured in kg).

In the context of the multiple linear regression, for fixed `Height`, the mean catheter `Length` increases by 0.421 for each one unit increase in `Weight` (where catheter `Length` is measured in cm, `Height` is measured in cm, and `Weight` is measured in kg).

The difference between these two coefficients is due to the high correlation (0.961) between the two predictor variables `Weight` and `Height`.

When interpreting the coefficient for `Weight` (0.611) in the simple linear regression model (which excludes `Height`), it is important to keep in mind that when we increase `Weight`, that `Height` also increases and both factors are associated with increased catheter `Length`. However, when interpreting the coefficient for `Weight` (0.421) in the multiple linear regression model (which includes `Height`) we are keeping `Height` fixed, and so the resulting increase in catheter `Length` is smaller.

1.5 Subspaces

The multiple regression model can be described in terms of two subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$, relating to the two simple linear regressions, one with Height as the predictor variable, and one with Weight as the predictor variable. Let x_1 denote the vector of Height values and let x_2 denote the vector of Weight values.

(a) Specify the two subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$, and also $\mathcal{L}_1 \cap \mathcal{L}_2$

$$\mathcal{L}_1 = Span\{\vec{1}, x_1\}$$
$$\mathcal{L}_2 = Span\{\vec{1}, x_2\}$$
$$\mathcal{L}_1 \cap \mathcal{L}_2 = Span\{\vec{1}\}$$

(b) Specify the two subspaces $\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp$ and $\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp$

Using a modified Gram-Schmidt process to convert $\mathcal{L}_1$ and $\mathcal{L}_2$ orthogonal spaces.

let

$$\mathcal{L}_1 = Span\{\vec{1}, {x_1}^*\}$$

such that

$$x_1^* = x_1 - proj_{(x_1)}$$

similarly, let

$$\mathcal{L}_2 = Span\{\vec{1}, x_2^*\}$$

such that

$$x_2^* = x_2 - proj_{\vec{1}}(x_2)$$

hence

$$\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp = Span\{x_1^*\}$$
$$\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp = Span\{x_2^*\}$$

(c) Noting that the subspaces in (b) are one-dimensional, calculate the angle between the two spaces and comment in light of the preceding statistical analyses.

Using the appended code, the angle was calculated to be equal to 16.0391, and $cos(16.0391) = 0.961$ which is the correlation between `Weight` and `Height`.

Clearly the two spaces are not orthogonal, meaning that there is correlation between them. An angle closer to 0° corresponds to a greater positive correlation. Similarly an angle close to 180° would correspond to a greater negative correlation. And 90° would have zero correlation.

1.6 Final Model

Select a regression model for the data, carefully justifying your choice using diagnostics and appropriate statistics.

The simple linear regression model involving the variable `Weight` is chosen to be the most suitable model in predicting catheter `Length`. This produces a final model of:

$$\text{Length} = 25.636 + 0.611 \times \text{Weight}$$

As neither predictor variables `Height` nor `Weight` are significant in the multiple regression model, this model is not chosen.

As `Weight` is more highly correlated to `Length` than `Height` is to `Length`, and the regression assumptions of the simple linear model with `Weight` are better than that with `Height`, a final model which only includes the predictor variable `Weight` is chosen. Refer to the Appendix for regression diagnostics and model summaries of the final model.

2 Part B: Logistic Regression

2.1 Introduction

Mammography is the most effective method available for breast cancer screening. However, the low predictive value of breast biopsy resulting from mammograms leads to approximately 70% of unnecessary biopsies with benign (non-malignant) outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in recent years. These systems help physicians in their decision about whether to perform a breast biopsy on a suspicious lesion seen in a mammogram, or to perform a follow-up examination instead.

The purpose of this analysis is therefore to obtain the best predictive model for mammographic mass severity, which can then be used to obtain predicted probabilities. These can then be used by clinicians to further assist their decision-making.

The dataset `mammo.txt` contains data on the true status of 961 mammographic mass lesions given by the variable `Severity`, where 0 = benign (is not cancer) and 1 = malignant (is cancer), the patients age in years (the variable `Age`), and three BI-RADS attributes which are described below:

- `Shape`: round = 1, oval = 2, lobular = 3, irregular = 4;
- `Margin`: circumscribed = 1, microlobulated = 2, obscured = 3, ill-defined = 4, spiculated = 5;
- `Density`: high = 1, iso = 2, low = 3, fat-containing = 4.

The variable `Severity` is a categorical response variable. `Age` is a continuous predictor variable, and `Shape`, `Margin`, and `Density` are all categorical predictor variables.

BI-RADS stands for Breast Imaging Reporting and Data System and was established by the American College of Radiology. The dataset also contains an assessment of the masses by physicians that have been identified on full field digital mammograms (the variable `BI.RADS`). This variable is not a predictor variable however, and is excluded from the present analysis.

2.2 Data Entry and Data Cleaning

The code for Section 2.2 is listed under the Appendices.

The first step of the cleaning process was to remove the variable `BI-RADS` as it was excluded from the analysis.

It was then necessary to check if there were any missing values from any of the five variables, `Severity`, `Age`, `Shape`, `Margin`, and `Density`. As there were missing values in all of the predictor variables, `Age`, `Shape`, `Margin`, and `Density`, they were set to `NA`. These `NA` values were not removed from the dataset, as this would remove a majority of the data.

The class of each variable was then checked, ensuring that `Age` is classed as 'numeric', and `Shape`, `Margin`, `Density` and `Severity` are all classed as 'factor'. It was necessary to change `Age` from the class 'factor' to 'numeric', and also to change `Severity` from the class 'integer' to 'factor'.

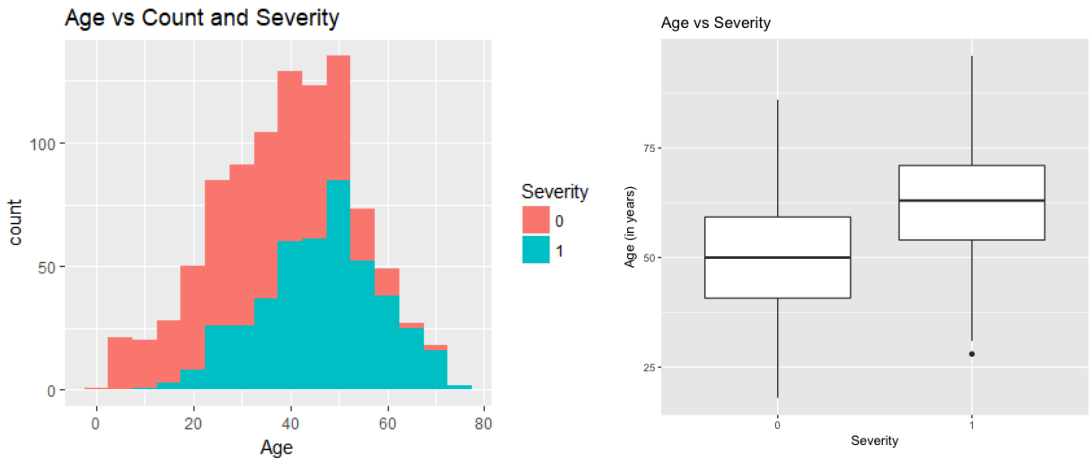
2.3 Data Visualisation and Data Summaries

The code for Section 2.3 is listed under the Appendices.

	Age	**Shape**	**Margin**	**Density**	**Severity**
Age	1.000	0.380	0.421	0.052	0.455
Shape	0.380	1.000	0.738	0.074	0.563
Margin	0.421	0.738	1.000	0.125	0.573
Density	0.052	0.074	0.125	1.000	0.068
Severity	0.455	0.563	0.573	0.068	1.000

Table 1: Correlation Matrix for the Predictors

Table 1 displays the correlation between individual variables. Margin has the strongest correlation of with the response variable Severity of 0.573, whilst Shape has a slightly weaker correlation with Severity of 0.563. Density, however, has a much weaker correlation to Severity with a value of 0.068, suggesting a weaker relationship between Density and Severity. It is also important to note that there is a high correlation between the predictor variables Shape and Margin, of 0.738.



(a) Age vs Count and Severity plot

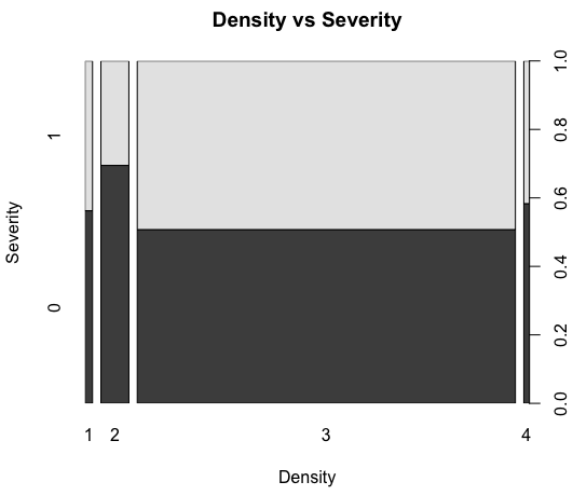
(b) Age vs Severity Boxplot

Figure 5: Plots for Age vs Severity

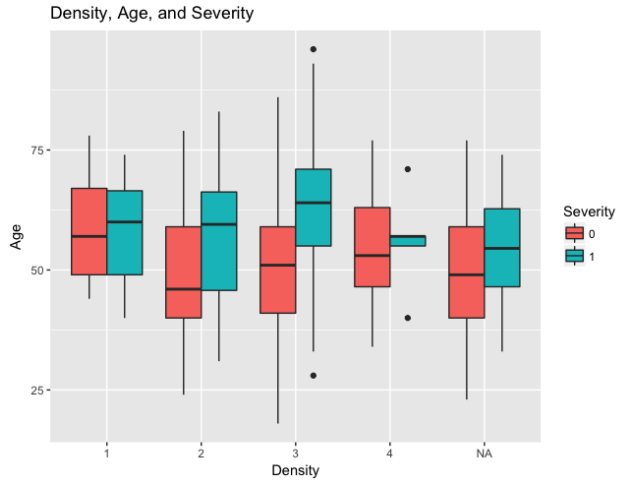
Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
18.00	45.00	57.00	55.49	66.00	96.00	5

Table 2: Summary Statistics for Age

Figures 5a and 5b show that generally a higher age indicates a higher proportion of the malignant cancer.



(a) Density vs Severity Plot



(b) Age, Density and Severity Boxplot

Figure 6: Plots of Density vs Severity

1	**2**	**3**	**4**	**NA's**
16	59	798	12	76

Table 3: Summary Statistics for Density

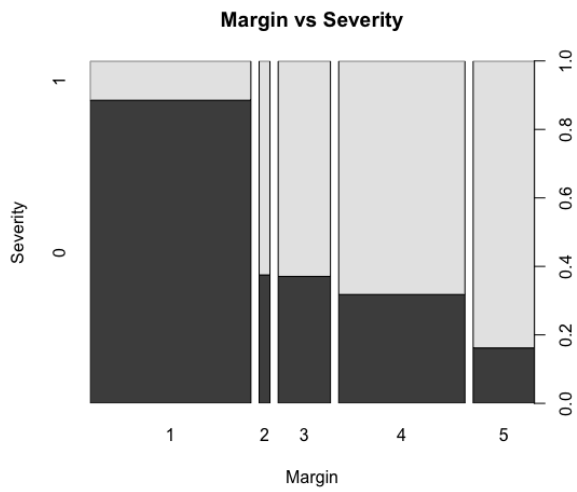
	Density			
Severity	**1**	**2**	**3**	**4**
0	9	41	405	7
1	7	18	393	5

Table 4: Density vs Severity Summary Statistics

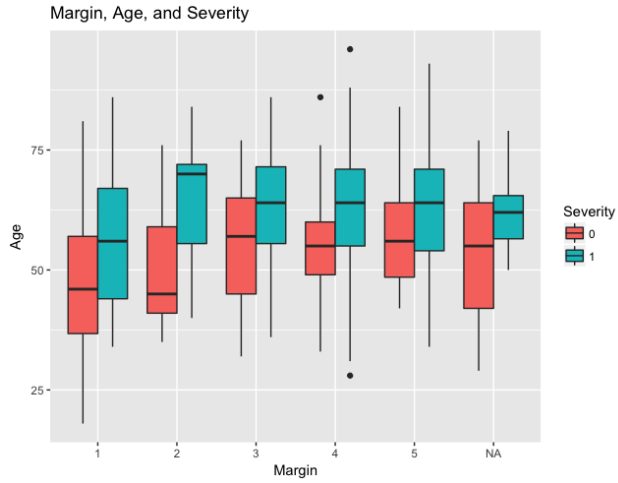
Figure 6a shows that regardless of the factor `Density`, there is a relatively even spread of cancerous and benign masses. From observation of this figure it appears that `Density` does not play a signficant role in effecting the `Severity` of the mammographic mass.

Table 4 shows that there is a relatively even spread of data for a `Density` level of 1, 3 and 4. The level 2 has a slightly higher ratio, indicating that iso tumours tend to be more malignant. It should be noted that approximately 90% of the data for `Density` is represented by level 3 (low density tumours), implying that the other 3 levels suffer from a lack of a sufficiently large sample size, making it more difficult to make an accurate assessment.

Overall, it appears that the `Density` of the mammographic mass has little significance when determining whether the mass is benign or is cancerous.



(a) Margin vs Severity plot



(b) Margin vs Severity Boxplot

Figure 7: Plots of Margin vs Severity

1	2	3	4	5	NA's
357	24	116	280	136	48

Table 5: Summary Statistics for Margin

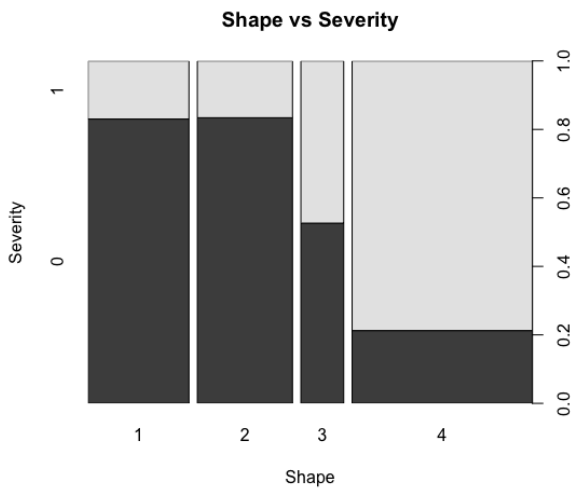
	Margin				
Severity	**1**	**2**	**3**	**4**	**5**
0	316	9	3	89	22
1	41	15	73	191	114

Table 6: Margin vs Severity Summary Statistics

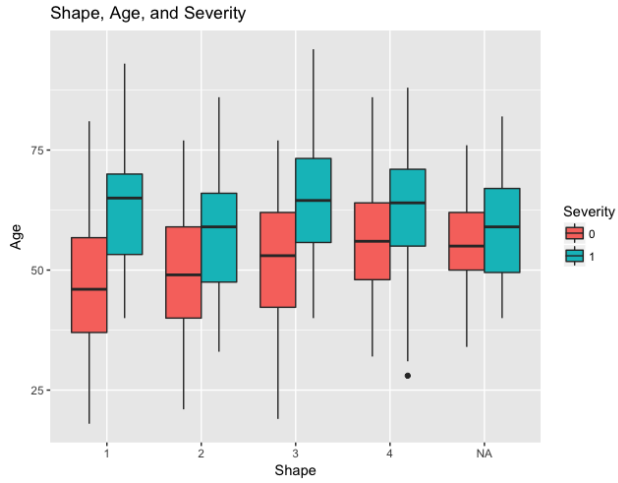
Figure 7a shows that if the mammographic mass lesion has a circumscribed margin, it is predominately a benign mass. In comparison, if the margin is either microlobulated, obscured, or ill-defined, the chance that the mass is malignant is slightly higher than it being benign. If the margin is spiculated, there was is a high chance of the mass being cancerous on average. Table 6 further shows this, with the data matching the figure 7a. It should be again noted that margin levels 2 and 3 have a very small sample number, indicating there are only a small number of microlobulated tumours in the dataset.

Figure 7b again shows that for each margin level, the age of the patient significantly impacts the chance of a severity rating of 1. Overall, a margin factor of 3, 4 and 5 have relatively similar spread of age for a severity level of 1 , with also an approximately equal mean for a severity factor of 0. A margin factor of 1 is has a considerably lower spread and mean on average in comparison with the other factors.

Overall, it appears that a margin factor of 3, 4 or 5, corresponding to an obscured, ill-defined, or spiculated margin, increase the chance of the mass being cancerous.



(a) Shape vs Severity plot



(b) Shape vs Severity Boxplot

Figure 8: Plots of Margin vs Severity

1	2	3	4	NA's
224	211	95	400	31

Table 7: Summary Statistics for Shape

	Shape			
Severity	**1**	**2**	**3**	**4**
0	186	176	50	85
1	38	35	45	315

Table 8: Shape vs Severity Summary Statistics

Figure 8a shows that the mammographic mass lesions have a low severity on average if the shape is either round or oval. This indicates that a mass of this shape is usually benign. Lobular-shaped masses have an average rating of 0.5 which indicates that the mass is cancerous half of the time, and benign the other half. The irregular shape has a high severity rating on average which indicates that a mass of that shape is likely to be cancerous. Table 7 and 8 show that there is a relatively even spread of data for each shape factor, with there being slightly less samples for level 3 and slightly more for level 4.

Overall, it appears that a higher an irregular-shaped mass implies that there is a greater probability the mass is cancerous.

2.4 Model Fitting and Model Selection

The code for Section 2.4 is listed under the Appendices.

The model was fitted as a generalised linear model (GLM).

To reduce and fit the model appropriately, all variables with NA values were initially removed from the model.

A method of backwards selection was then used with a significance level of $\alpha = 0.5$ to iteratively remove the variables from the model which had the highest p-value. All insignificant interaction terms were removed before single-order terms in order to uphold the principle of marginality.

The model was re-evaluated after each interaction term was successively removed. Once all interaction terms were removed, the Density variable was also non-significant and so was removed from the final model.

As the variables Age, Shape, and Margin are now all significant, a suitable model has been determined.

2.5 Final Model

$$\begin{aligned}\text{Severity} = -4.720 + 0.054 \times \text{Age} \\ - 0.448 \times \text{Shape2} + 0.499 \times \text{Shape3} + 1.243 \times \text{Shape4} \\ + 1.583 \times \text{Margin2} + 1.263 \times \text{Margin3} + 1.543 \times \text{Margin4} + 2.032 \times \text{Margin5}\end{aligned}$$

Where Margin# and Shape# are 0/1 (true/false) values corresponding to their respective state.

2.6 Interpretation of Final Model

Due to the model being a binary `GLM` the coefficients are in terms of log odds.

- Age 0.054 goes to $e^{0.054} = 1.055$ hence older patients have a higher chance of being diagnosed with breast cancer. That is; ageing 1 year increases the change of cancer (by 1.055%).
- Shape2 -0.448 goes to $e^{-0.448} = 0.638$ hence being diagnosed with an oval shaped tumour very slightly increases the chance of the patient having breast cancer (by 0.638%).
- Shape3 +0.499 goes to $e^{0.499} = 1.647$ hence being diagnosed with an lobular shaped tumour slightly increases the chance of the patient having breast cancer (by 1.647%).
- Shape4 +1.243 goes to $e^{1.243} = 3.465$ hence being diagnosed with an irregular shaped tumour greatly increases the patients chance of having breast cancer (by 3.465%).
- Margin2 +1.583 goes to $e^{1.583} = 4.866$ hence having an microlobulated Margin Greatly increases the patients chance of having breast cancer (by 4.866%).
- Margin3 +1.263 goes to $e^{1.263} = 3.536$ hence having obscured margins greatly increases the patients chance of having breast cancer (by 3.536%).

- Margin4 +1.543 goes to $e^{1.543} = 4.678$ hence having ill-defined margins greatly increases the patients chance of having breast cancer (by 4.678%).
- Margin5 +2.032 goes to $e^{2.032} = 7.629$ hence having spiculated margins drastically increases the patients chance of having breast cancer (by 7.629%)

2.7 Predicting Probabilities and Interpretation

Below analyses the probability of a `Severity` level of 1 for varying levels of the predictor variables. This corresponds to the probability that the mammography mass is cancerous. Each case keeps either two of variables `Age`, `Shape`, and `Margin` the same, whilst changing the other. This gives us an idea at how changing certain variables impacts the probability that the mass is malignant.

Changing Age

Firstly, 3 different ages are observed with the other variables kept constant. For this case, we look at the age at 18, 55, and 96 to gain an understanding of the effect that age has on the probability of a cancerous mass. Shape is set to level 1 and Margin is set to level 4.

	Age		
	18	**55**	**96**
Probability	0.0992	0.447	0.880

Table 9: Changing the variable of Age whilst keeping Margin and Shape Constant

This table indicates that the greater the age of the patient, the greater the probability of the mass being cancerous. This is to be expected from both the data visualisation data as well as the interpretation of the final model.

Changing Shape

Next, each different level of shape is analysed, with age being set to 55 and Margin to level 1.

Shape			
Round (1)	**Oval (2)**	**Lobular (3)**	**Irregular (4)**
0.147	0.099	0.221	0.374

Table 10: Changing the factor level of Shape whilst keeping the Age and Margin variable constant

These probabilities imply that levels 3 and 4 rating for shape significantly increase the probability that the mass is cancerous. In comparison, levels 1 and 2 decrease the probability that the mass is cancerous.

Changing Shape

Finally, each different level of Margin is analysed, with age being set to 55 and shape to 1.

Margin				
Circumscribed (1)	**Microlobulated (2)**	**Obscured (3)**	**Ill-Defined (4)**	**Spiculated (5)**
0.147	0.457	0.379	0.447	**0.568**

Table 11: Changing the factor level of Margin whilst keeping the variables Age and Shape constant

The above table indicates that a margin factor of 2 or 4 yields a fairly similar probability. A factor of 5 increases the probability of the mass being cancerous by approximately 0.1, whilst a factor of 3 has a slightly less probability.

3 Part C: Group Work

The group commenced work for the project on the 20th of April, in which an initial face-to-face meeting was held and the spokesperson for the group was elected. This meeting was then also used to introduce ourselves to the requirements of the project and to delegate tasks to individual members of the group.

We made use of various social media platforms to assist us in communicating outside of university, such as *Facebook Messenger* and *Discord*. These two applications allowed us to easily discuss the project and any progression we had made or concerns we had, without the need for an in-person meeting.

The majority of the work completed for the project was accomplished via these forms of social media as well as through face-to-face meetings, in which we met at university approximately once per fortnight and communicated via social media approximately twice per week. The use of *Overleaf* was also essential to collaborating effectively, as it allowed us to simultaneously work on the project.

After the initial meeting, the subsequent group meeting was held on the 4th of May to discuss any progress we had made, in which the majority of Part A had been completed, and Part B had been commenced. In the following fortnight, further progress was completed on Part B, and by the 18th of May the majority of the work for the project had been accomplished, however, the written report and formatting remained to be finalised. Any additional work was then completed via social media and a final meeting was held on the morning of the 4th of June in order to sign the required coversheet and to finalise and submit the project.

Statement from each member of the group outlining their contribution to the work:

Thomas: Largely contributed to Part A and B of the written report, including any typesetting and formatting.

Nikky: Primarily contributed to the coding in R, and completed a large portion of the write-up for Part A and B.

Andrew: Was the spokesperson for the group, completed Section 1.5 of the coding (question 6 of Part A), as well as mostly contributed to any final editing, typesetting, and formatting.

James: Finalised a large portion of Section B, and contributed to Part C of the written report, as well as final editing and formatting.

Josh: Contributed principally to a majority of the coding in R, as well as completed a portion of the write-up for Section B.

Appendices

Part 1

Catheter Analysis

```
# Load useful packages
library(tidyverse)
library(ggplot2)
library(readxl)
library(MASS)
library(GGally)

# PART A - MULTIPLE LINEAR REGRESSION

# (1)
# read the data into r
catheter <- read_excel("catheter.xlsx")

# remove the variable Child as it does not contain any relevant information
catheter$Child <- NULL

# create a pairwise scatterplot matrix
pairs(catheter[,c("Length","Height","Weight")])

# create a correlation matrix
round(cor(catheter[,c("Length","Height","Weight")]),3)

ggpairs(catheter[,c("Length","Height","Weight")])

# (2)
#linear regression models
catheter.lm1 <- lm(Length~Height+Weight, data=catheter)
catheter.lm2 <- lm(Length~Weight+Height, data=catheter)
catheter.lm3 <- lm(Length~Height, data=catheter)
catheter.lm4 <- lm(Length~Weight, data=catheter)

# (3) - not done in R

# (4)

par(mfrow=c(2,2))
plot(catheter.lm1)
par(mfrow=c(1,1))
catheter$studres <- studres(catheter.lm1)

with(catheter,plot(Height,studres,pch=20,pty="s",
ylab="Studentised Residuals", main = "Studentised Residuals vs Height"))
abline(0,0,col="red")

with(catheter,plot(Weight,studres,pch=20,
```

```
ylab="Studentised Residuals", main = "Studentised Residuals vs Weight"))
abline(0,0,col="red")

with(catheter,plot(Length,studres,pch=20,
ylab="Studentised Residuals", main = "Studentised Residuals vs Severity"))
abline(0,0,col="red")

summary(catheter.lm1)
summary(catheter.lm2)
summary(catheter.lm3)
summary(catheter.lm4)

# (5) - not done in R
# (6) - not done in R

# (7)

# Final model includes Weight only
# Clearly the multiple linear regression shouldn't be used
# Weight is more highly correlated to Length than Height is to Length
# and the assumptions for model checking are much better

# Assumption checking for final model
par(mfrow=c(2,2))
plot(catheter.lm4)
catheter$studres <- studres(catheter.lm4)

with(catheter,plot(Height,studres,pch=20,pty="s",
ylab="Studentised Residuals", main = "Studentised Residuals vs Height"))
abline(0,0,col="red")

with(catheter,plot(Weight,studres,pch=20,
ylab="Studentised Residuals", main = "Studentised Residuals vs Weight"))
abline(0,0,col="red")

with(catheter,plot(Length,studres,pch=20,
ylab="Studentised Residuals", main = "Studentised Residuals vs Severity"))
abline(0,0,col="red")
```

Linear Space Code

```
library(readxl)
library("dplyr")
###Q6----
Num = read_xlsx('catheter.xlsx');
Height = pull(Num[,2])
Weight = pull(Num[,3])


#a is the first column of L1 and L2
#calculate span of L1 & L2
a  = array(1,12)
b1 = Height
b2 = Weight

#Using span to calculate projection vectors
projection1 = (a%*%b1/(norm(a,"2")^2))%*%a
projection2 = (a%*%b2/(norm(a,"2")^2))%*%a

u = as.vector(b1 - projection1)
v = as.vector(b2 - projection2)

#Calculating the angle between spaces
costheta = (u%*%v)/(norm(u,"2")*norm(v,"2"))
thetarad = acos(costheta)
thetadeg = 180*thetarad/pi
```

Part 2

Data Entry and Cleaning

```
mammo <- read.csv("mammo.txt") # reads the data into R
mammo$BI.RADS <- NULL # removes the variable BI.rads

# Cleaning the Age variable
table(mammo$Age) # there are 5 missing entries
mammo$Age[mammo$Age == "?"] <- NA # sets the missing entries to NA
class(mammo$Age) # the class is "factor" and so needs to be changed
mammo$Age <- as.numeric(as.character(mammo$Age)) # sets class to be "numeric"

# Cleaning the Shape variable
table(mammo$Shape) # there are 31 missing entries
mammo$Shape[mammo$Shape == "?"] <- NA # sets the missing entries to NA
class(mammo$Shape) # the class is "factor" and so doesn't need to be changed
mammo$Shape <- droplevels(mammo$Shape) # removes the empty "?" level

# Cleaning the Margin variable
table(mammo$Margin) # there are 48 missing entries
mammo$Margin[mammo$Margin == "?"] <- NA # sets the missing entries to NA
class(mammo$Margin) # the class is "factor" and so doesn't need to be changed
mammo$Margin <- droplevels(mammo$Margin) # removes the empty "?" level

# Cleaning the Density variable
table(mammo$Density) # there are 76 missing entries
mammo$Density[mammo$Density == "?"] <- NA # sets the missing entries to NA
class(mammo$Density) # the class is "factor" and so doesn't need to be changed
mammo$Density <- droplevels(mammo$Density) # removes the empty "?" level

# Cleaning the Severity variable
table(mammo$Severity) # there are no missing entries
class(mammo$Severity) # the class is "integer" so needs to be changed
mammo$Severity <- as.factor(mammo$Severity) # sets class to be "factor"
```

Data Visualisation and Summaries

```
# to create a correlation matrix all variables must be classed as numeric
mammo$Shape <- as.numeric(mammo$Shape)
mammo$Margin <- as.numeric(mammo$Margin)
mammo$Density <- as.numeric(mammo$Density)
mammo$Severity <- as.numeric(as.character(mammo$Severity))
table(mammo$Severity)
cor(mammo, use = "complete.obs") # creates a correlation matrix accounting for NA values


% Correlation matrix here

# reclassify the necessary variables
mammo$Shape <- as.factor(mammo$Shape)
mammo$Margin <- as.factor(mammo$Margin)
mammo$Density <- as.factor(mammo$Density)
mammo$Severity <- as.factor(mammo$Severity)

# summary statistics for each individual variable
summary(mammo$Age)
summary(mammo$Shape)
summary(mammo$Margin)
summary(mammo$Density)
summary(mammo$Severity)

% Summary statistics here

# summary tables for each predictor variable against Severity
with(mammo, table(Severity, Age))
with(mammo, table(Severity, Shape))
with(mammo, table(Severity, Density))
with(mammo, table(Severity, Margin))

plot(mammo[,c("Severity","Age","Shape","Margin","Density")]) # create a pariwise scatterplot ma

# Box-plots
ggplot(data=mammo, aes(Severity, Age)) + geom_boxplot() +
labs(x= "Severity", y="Age (in years)")
ggplot(data=mammo, aes(Shape, Age, fill = Severity)) + geom_boxplot(na.rm = TRUE)
ggplot(data=mammo, aes(Density, Age, fill = Severity)) + geom_boxplot()
ggplot(data=mammo, aes(Margin, Age, fill = Severity)) + geom_boxplot()

# Other plots
plot(mammo$Age, mammo$Severity, xlab = "Age (in years)", ylab = "Severity")
plot(mammo$Shape, mammo$Severity, data=mammo, xlab = "Shape", ylab = "Severity")
plot(mammo$Margin, mammo$Severity, data=mammo, xlab = "Margin", ylab = "Severity")
plot(mammo$Density, mammo$Severity, data=mammo, xlab = "Density", ylab = "Severity")
```

Model Fitting and Model Selection

```
mammo.glm <- glm(Severity ~ (Age + Shape + Margin + Density)^2, data=mammo, family = "binomial")
summary(mammo.glm)

# Get rid of this interactions as there are NA values
mammo.glm <- update(mammo.glm, .~. -Shape:Density)
summary(mammo.glm)

# Get rid of this interaction as there are NA values
mammo.glm <- update(mammo.glm, .~. -Margin:Density)
summary(mammo.glm)

# Highest P value is 0.9898
mammo.glm <- update(mammo.glm, .~. -Age:Margin)
summary(mammo.glm)

# Highest P value is 0.98932
mammo.glm <- update(mammo.glm, .~. -Shape:Margin)
summary(mammo.glm)

# Highest P value is 0.7958
mammo.glm <- update(mammo.glm, .~. -Age:Density)
summary(mammo.glm)

# Highest P value is 0.5688
mammo.glm <- update(mammo.glm, .~. -Age:Shape)
summary(mammo.glm)

mammo.glm <- update(mammo.glm, .~. -Density)
summary(mammo.glm)
# Final model includes Age, Shape, and Margin
```

Predicting Probabilities

```
# The below give the probabilities of Severity = 1, i.e. probability of cancer.
# Have only chosen a few probabilities.

# Age
newdata1 <- with(mammo, data.frame(Age = 55, Shape = "1", Margin = "4")) # mean age
predict(mammo.glm, newdata1, type = "response")

newdata2 <- with(mammo, data.frame(Age = 18, Shape = "1", Margin = "4")) # min age
predict(mammo.glm, newdata2, type = "response")

newdata3 <- with(mammo, data.frame(Age = 96, Shape = "1", Margin = "4")) # max age
predict(mammo.glm, newdata3, type = "response")

# Shape
newdata4 <- with(mammo, data.frame(Age = 55, Shape = "1", Margin = "1")) # round
predict(mammo.glm, newdata4, type = "response")

newdata5 <- with(mammo, data.frame(Age = 55, Shape = "2", Margin = "1")) # oval
predict(mammo.glm, newdata5, type = "response")

newdata6 <- with(mammo, data.frame(Age = 55, Shape = "3", Margin = "1")) # lobular
predict(mammo.glm, newdata6, type = "response")

newdata7 <- with(mammo, data.frame(Age = 55, Shape = "4", Margin = "1")) # irregular
predict(mammo.glm, newdata7, type = "response")

# Margin
newdata8 <- with(mammo, data.frame(Age = 55, Shape = "1", Margin = "1")) # circumscribed
predict(mammo.glm, newdata8, type = "response")

newdata9 <- with(mammo, data.frame(Age = 55, Shape = "1", Margin = "2")) # microlobulated
predict(mammo.glm, newdata9, type = "response")

newdata10 <- with(mammo, data.frame(Age = 55, Shape = "1", Margin = "3")) # obscured
predict(mammo.glm, newdata10, type = "response")

newdata11 <- with(mammo, data.frame(Age = 55, Shape = "1", Margin = "4")) # ill-defined
predict(mammo.glm, newdata11, type = "response")

newdata12 <- with(mammo, data.frame(Age = 55, Shape = "1", Margin = "5")) # spiculated
predict(mammo.glm, newdata12, type = "response")
```

Model Summaries

Catheter multiple regression

```
> summary(catheter.lm1)

Call:
lm(formula = Length ~ Height + Weight, data = catheter)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0497 -1.2588 -0.2576  1.8987  7.0030 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 21.00828    8.74782   2.402   0.0398 *
Height       0.07729    0.14192   0.545   0.5993  
Weight       0.42081    0.36405   1.156   0.2775  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 3.943 on 9 degrees of freedom
Multiple R-squared:  0.8054,Adjusted R-squared:  0.7621 
F-statistic: 18.62 on 2 and 9 DF,  p-value: 0.0006332 
```

Catheter simple regression (height)

```
> summary(catheter.lm3)

Call:
lm(formula = Length ~ Height, data = catheter)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0996 -0.7246 -0.2608  1.1585  6.6826 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.12402    4.24711   2.855 0.017113 *  
Height       0.23495    0.03986   5.894 0.000152 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 4.008 on 10 degrees of freedom
Multiple R-squared:  0.7765,Adjusted R-squared:  0.7541 
F-statistic: 34.74 on 1 and 10 DF,  p-value: 0.0001523 
```

Catheter simple regression (weight)

```
> summary(catheter.lm4)

Call:
```

```
lm(formula = Length ~ Weight, data = catheter)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9958 -1.4818 -0.1334  2.0899  7.0378 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.63596    2.00425  12.791 1.60e-07 ***
Weight       0.61136    0.09698   6.304 8.86e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 3.801 on 10 degrees of freedom
Multiple R-squared:  0.7989,Adjusted R-squared:  0.7788 
F-statistic: 39.74 on 1 and 10 DF,  p-value: 8.865e-05
```

Catheter simple regression (weight)

```
> summary(mammo.glm)

Call:
glm(formula = Severity ~ Age + Shape + Margin, family = "binomial", 
    data = mammo)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5004  -0.5514  -0.2399   0.6651   2.5963  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.719544   0.465771 -10.133  < 2e-16 ***
Age          0.053879   0.007499   7.185 6.72e-13 ***
Shape2      -0.447844   0.306327  -1.462 0.143747    
Shape3       0.499251   0.364446   1.370 0.170721    
Shape4       1.242837   0.324256   3.833 0.000127 ***
Margin2      1.582943   0.539614   2.933 0.003352 ** 
Margin3      1.263073   0.342531   3.687 0.000226 ***
Margin4      1.543226   0.294045   5.248 1.54e-07 ***
Margin5      2.032105   0.362892   5.600 2.15e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1226.93  on 886  degrees of freedom
Residual deviance:  773.89  on 878  degrees of freedom
  (74 observations deleted due to missingness)
AIC: 791.89
```

```
Number of Fisher Scoring iterations: 5
```

Catheter Diagnostics

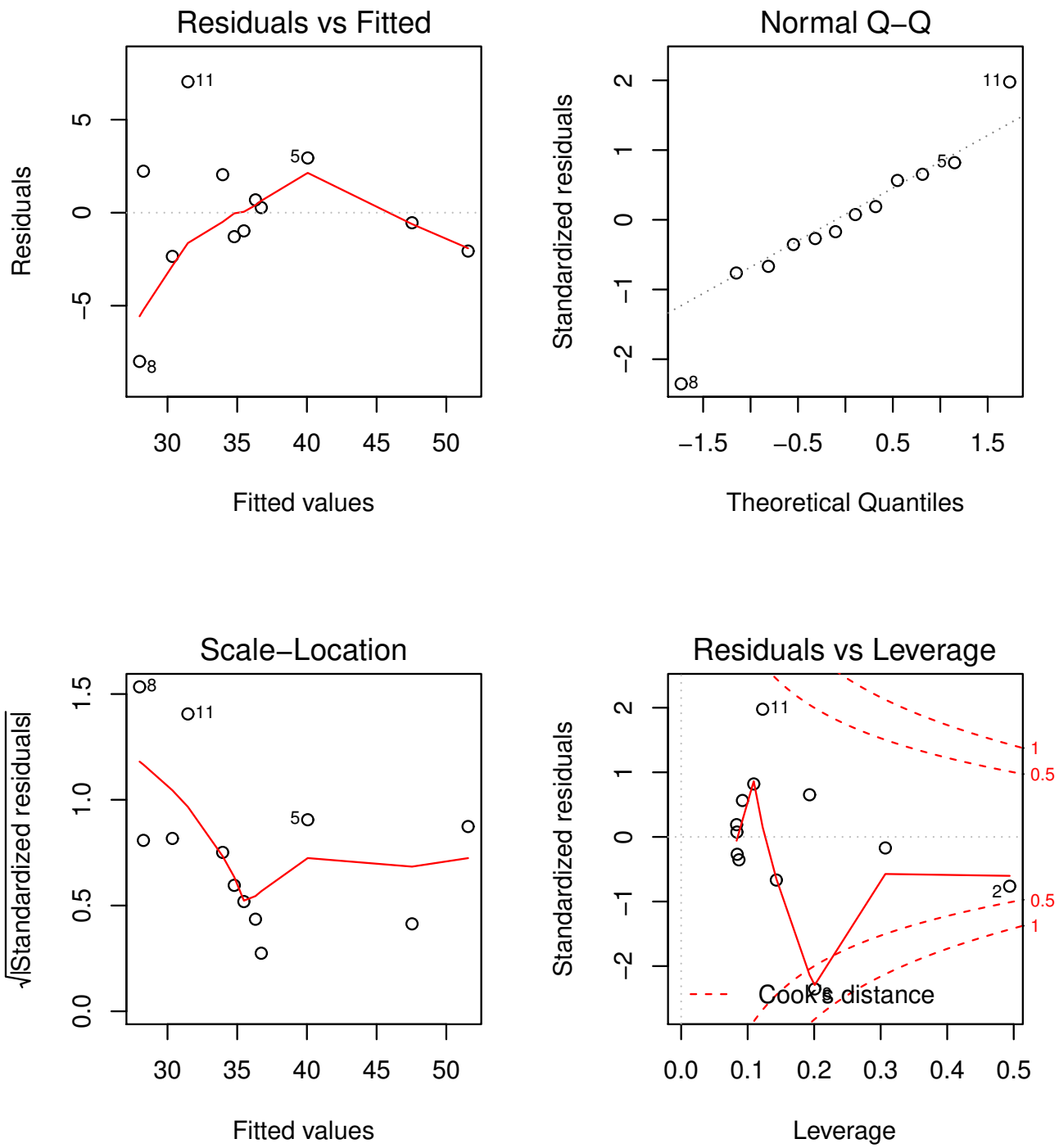


Figure 9: Plots for the final Catheter model diagnostics

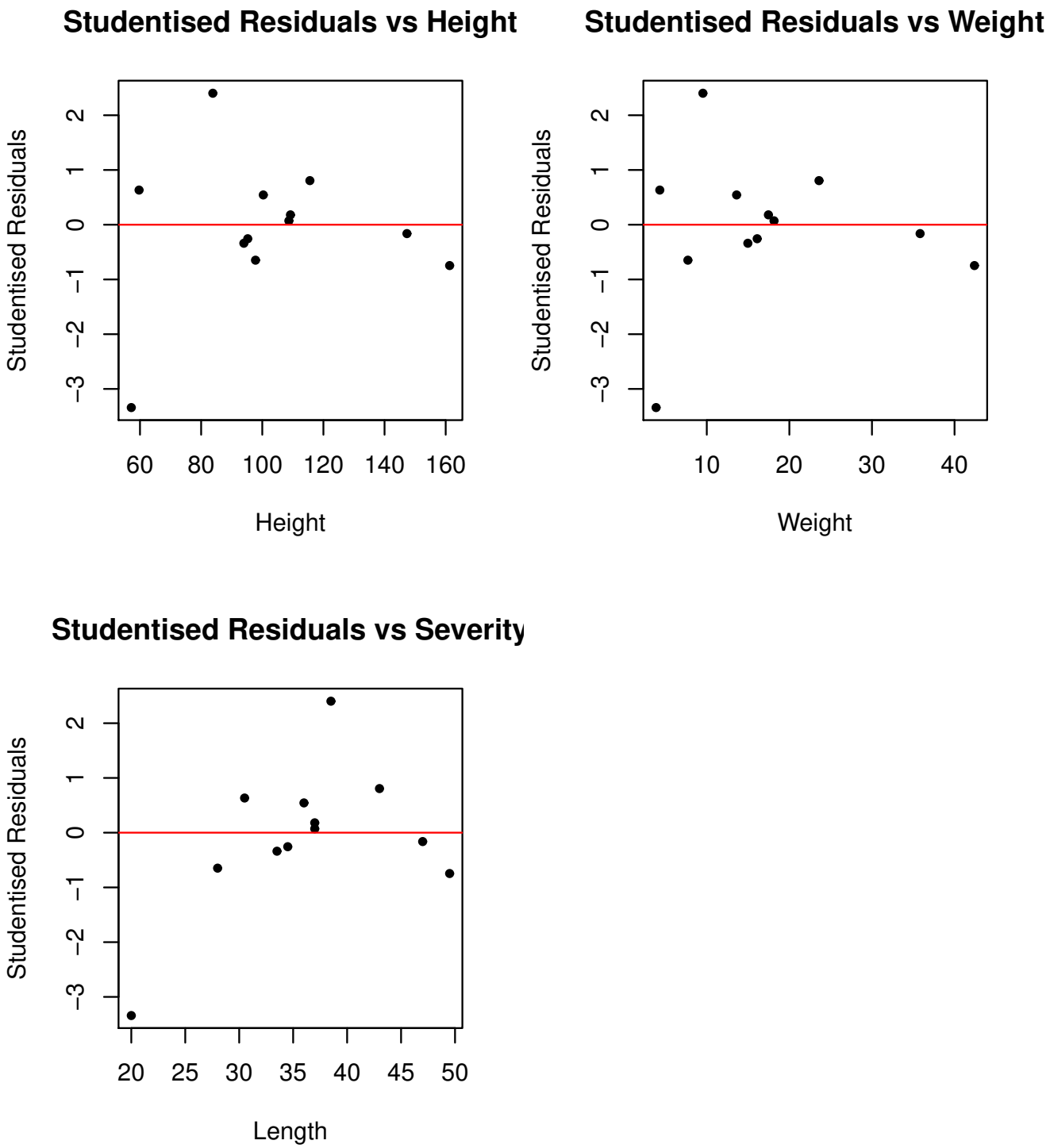


Figure 10: Plots for the final Catheter model Residuals