Statistical Modelling Project

Joshua Bean, James Beck, Lily Harris, Miriam Slattery, Tobin South

2 June 2018

Contents

Introduction **3**

Multiple Linear Regression **3**
Data Visualisation . . . 3
Fitting Models . . . 5
Assumption Checking . . . 5
Comparing Models . . . 9
Interpreting Coefficients . . . 12
Model as a Projection of Data on Subspaces . . . 12
Final Model Selection . . . 14

Logistic Regression **15**
Data cleaning . . . 15
Data Visualisation . . . 18
Confidence intervals . . . 19
Making a Model . . . 24
Using the right data . . . 24
The Full Model . . . 24
Model by stepwise selection . . . 26
Model by removal of non-signficant terms . . . 26
Choosing models . . . 28
Prediction to verify model . . . 30

Teamwork Reflection **31**
Contribution of Group Members . . . 32
Miriam Slattery . . . 32
Joshua Bean . . . 32
Tobin South . . . 33

Introduction

This report will explore two different forms of regression, multiple linear regression and logistic regression, through two different data sets.

Multiple linear regression uses several explanatory (predictor) variables to predict the outcome of a response variable. It attempts to model the relationship between these variables as a linear relationship by using data,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

where y_i is the ith data point of the response variable, $x_{i1}, \ldots, x_{ip}$ are the ith points of the p predictor variables, $\beta_0, \ldots, beta_p$ are constant coefficients and ϵ_i is the error in the ith prediction. In regression, we find the least squares estimates for $\beta_0, \ldots, \beta_p$ so that the error for each prediction is minimised.

Multiple linear regression will be performed on data about heart catheters in children. The aim is to be able to predict the length of catheter required for a child with the possible predictor variables of Height and Weight. First, we look at the relationship between predictors and the response variable to confirm that it is a linear relationship and that a linear model is appropriate. Then we will fit three models, one with both predictors and model with each single predictor. We will check the assumptions of linear models for each of these and compare the models. After examining the model subspaces, we will find that Height and Weight are correlated and neither is significant in the full model. Thus, we will select a simple regression model.

Logistic regression is used when the response variable has only two outcomes (usually denoted 0 and 1), and can include multiple predictor variables. We use the log-odds as the response variable in the regression, $\eta_i = \log \frac{\pi_i}{1-\pi_i}$, where pi_i is the probability of the ith response being 1. We find the best estimates, β, such that $\eta = \beta^T X$.

We will use logistic regression for modelling the data collected about mammograms such that we can predict whether a mass lesion found is malignant or benign to avoid unnecessary biopsies. This will use the measurements found in mammograms as predictors variables: Shape, Margin and Density. First, we will clean the data by making sure variables are appropriately categorised and missing data handled and denoted accordingly. Then we will use visualisations of the data and confidence intervals to explore the relationship between the variables. Then we can fit the full logistic model. Using stepwise selection with AIC, we will then reduce the model by removing statistically insignificant predictors. Finally, we will verify our model selection using prediction.

Multiple Linear Regression

Heart cathetisation is sometimes performed on children with congenital heart defects. We want to find if it is possible to predict the required length of catheter. The height, weight and catheter length was recorded for 12 individuals. The data is summarised below.

Variable	Description
Height	Heightof child (cm)
Weight	Weight of child (kg)
Length	Length of required heart catheter (cm)

Data Visualisation

First, we enter the data into R to allow for analysis.

```
Child<- c(1:12)
Height<-c(108.7,161.29,95.25,100.33,115.57,97.79,109.22,57.15,93.98,59.69,83.82,147.32)
```

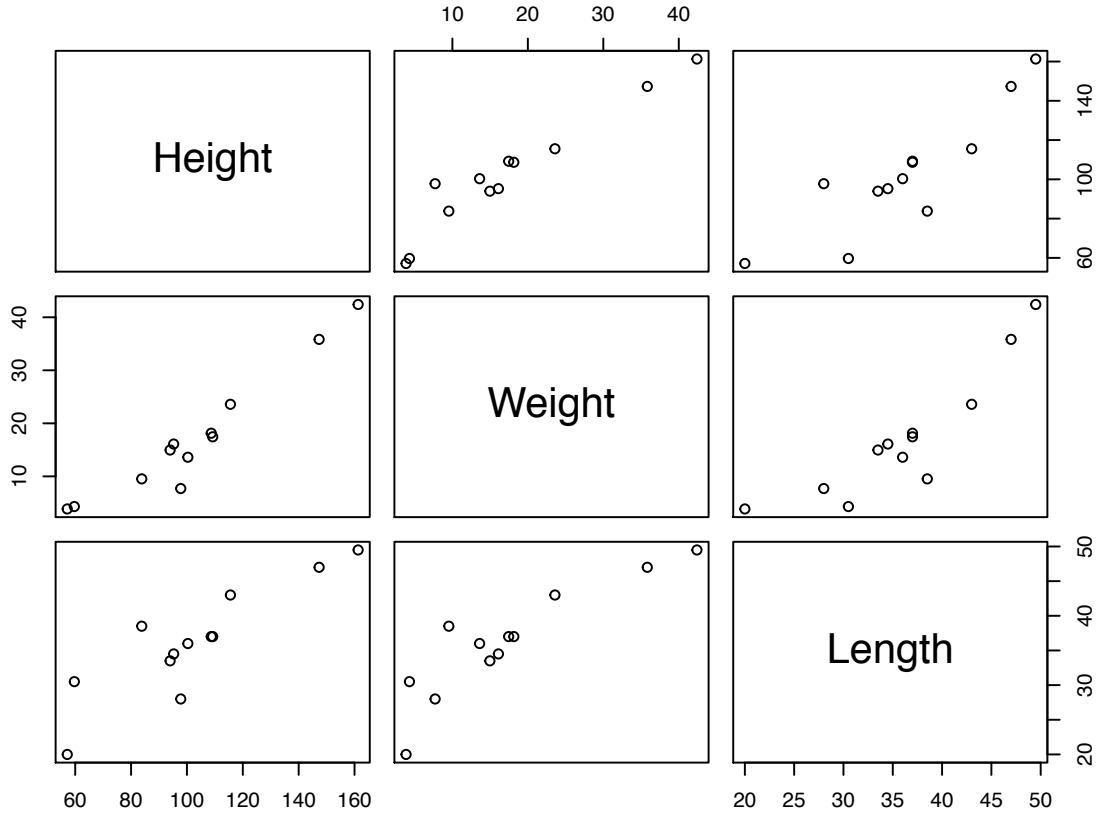


Figure 1: A pariwise scatter plot of all variables.

```
Weight<-c(18.14,42.41,16.10,13.61,23.59,7.71,17.46,3.86,14.97,4.31,9.53,35.83)
Length<-c(37,49.5,34.5,36,43,28,37,20,33.5,30.5,38.5,47.0)
catheter <-data.frame(Child, Height, Weight, Length)
```

Now we should explore the relationship between the predictor variables and the response variable. To do so, we will create a pairwise scatter plot matrix as follows,

```
pairs(subset(catheter, select=c(2:4)))
```

Before examining these plots closely, note that there are only 12 data points so there will be a lot of unexplained scatter and it is difficult to recognise trends.

To investigate the relationship between Length, the response variable, and the two predictor variables, Heigth and Weight, we need to consider the two plots in the top right of Figure 1.

The scatter plot between Length and Weight (top centre plot) shows the relationship is strong, positive and linear. Similarly, the plot between Length and Height (top right) shows the relationship is moderate, positive and mostly linear with slight positive curvature. The trend is only moderate to moderately strong because for the lower values, there is some deviation from the strong trendline. So we can conclude that the relationship between the resonse and the predictor variables is positive linear.

Now consider the relationship between the predictor variables, to look for correlation. To do so, consider the plot of Height against Weight (centre right plot), which shows the relationship is moderately strong, positive and mostly linear with slight positive curvature, and with a few outlying points below the curve. Although there is a small number of data points, we can see that the trend is predominantly linear.

Fitting Models

Since the data is linear with respect to all variables, we will fit linear models to the data. As there are only two predictor variables, we can perform an exhaustive check for all three different models. These models are as follows,

$$\begin{aligned}
\text{Length} &= \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Weight} \\
\text{Length} &= \beta_0 + \beta_1 \text{Height} \\
\text{Length} &= \beta_0 + \beta_2 \text{Weight}.
\end{aligned}$$

These linear models will be fit using the lm() function built into R as below.

```
lm1<-lm(Length~Height+Weight, data=catheter)
lm2<-lm(Length~Height, data=catheter)
lm3<-lm(Length~Weight, data=catheter)
```

Assumption Checking

For linear models of this form there are four assumptions to check, for each model, these assumptions are: linearity, homoscedasticity, normality and independence. Note that for multiple linear regression, we need to check for linearity and homoscedasticity between the residuals and the overall fitted data as well as between residuals and each predictor variable. For simple linear regression, the residuals vs fitted plot is simply the residuals against the one predictor variable.

Now we will check the assumptions for the first model, Length $= \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Weight}$, using the following diagnostic plots,

```
res1<-residuals(lm1)
par(mfrow=c(3,2))
plot(lm1,which=1)
plot(lm1,which=2)
plot(lm1,which=3)
plot(lm1,which=5)
plot(catheter$Height, res1,xlab = "Height",ylab="Residuals",main="Residuals vs Height")
plot(catheter$Weight, res1,xlab = "Weight",ylab="Residuals",main="Residuals vs Weight")
```

Linearity

To test for linearity between residuals and fitted values, consider the top left hand plot, residuals vs. fitted, in Figure 2.. If the data is linear, we would expect an equal number of points above and below the line '$y = 0$' for each 'x' value, with no curvature. For this data set, we can see that the data follows a reasonably linear trend with an equal spread of values above and below the horizontal line. Normally, the included red line may be helpful when discussing linearity. In this plot, the line clearly is not linear due to the sheer lack of data. Overall, there is insufficient evidence in the plot to invalidate the assumption of linearity.

To check for linearity between the residuals and the individual predictor variables, we need to consider the two bottom plots in Figure 2. In the residuals vs Height plot, we can see the data follows a similar trend to that described in the residuals vs fitted plot, once again due to the lack of data points, the assumption of linearity holds. In the residuals vs Weight plot, we can see that the data also follows a similar trend to the other two plots considered, thus the same justification can be applied to claim that the assumptions of linearity holds for the whole model.

Homoscedasticity

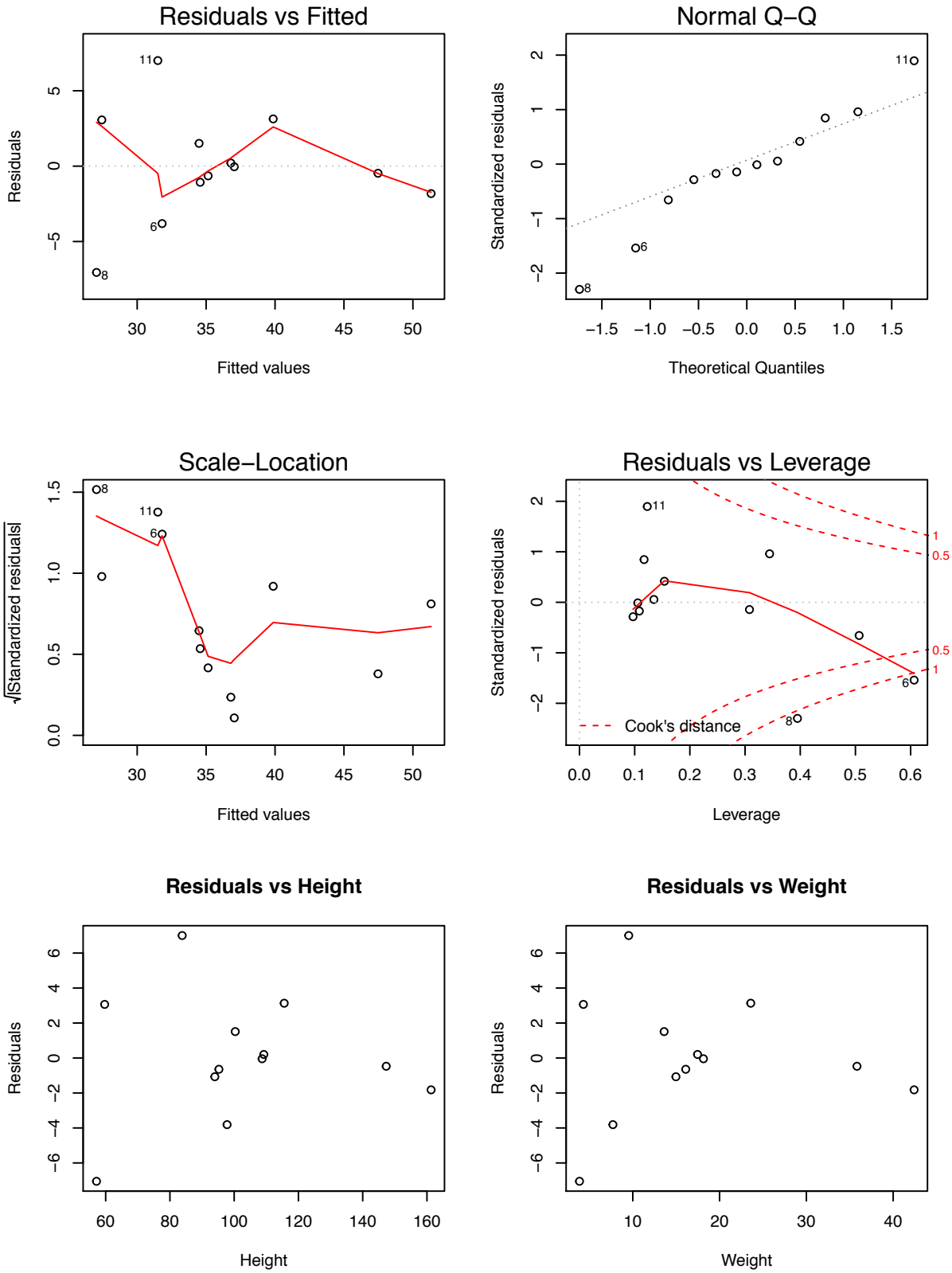


Figure 2: The six plots required to thoroughly check the assumptions for the first linear model.

To test for homoscedasticity between residuals and fitted values, consider the top left hand plot, residuals vs. fitted, in Figure 2. If the data is homoscedastic, we would expect to see an equal distance between the points and the horizontal line described above. However for this data set, we have that the plot has significant fanning, where the variance decreases as the Height and Weight increase. This is unsurprising since the data is of predominantly young children, with only two data points corresponding to large catheters (longer than 40cm). This explains why the variance is smaller for these values. Given the small number of data points, the assumption of homoscedasticity holds, albeit weakly. Another way to check for constant variance is to consider the scale location (center left) plot and look for linearity in the data points. However for this data set, we can see some curvature, which corresponds to the variance not being constant.

Now, to check homoscedasticity between residuals and height and weight independently, we need to consider the two bottom plots of 2. In both of these plots, we can see that the data follows a similar trend to that of the plot discussed above, with significant fanning. However this fanning will be due to the small data set, so once again the assumption of constant variance holds.

Normality

To test for Normality, consider the Normal-QQ (top right) plot in Figure 2. If the residuals are Normally distributed, we would expect to see a straight, linear line following the trend of $y = x$. For this data set, we can see that in general it follows a straight line, with some significant deviation from the trend for lower values. This significant deviation implies some negative skewing in the distribution. However since there are only 12 data points, randomness that is intrinsic in this plot is expected. Thus given the small data set, the residuals appear to be passably normal, thus there is insufficient evidence to invalidate this assumption.

Independence

There is no formal test for independence in the data, instead we need to consider the way the data was collected. This data set was collected from children with congenital heart defects, implying there is a slim chance that there is any relationship between the individuals. Thus it appears that there is no connection between each subject, thus the data is independent to the best of our knowledge.

Now we have checked the assumptions for the multiple regression model, we also need to check the same assumptions for the two simple regression models. However since we have discussed them in considerable detail above, we will breifly discuss them for the simple models.

For the simple regression model in just Height, we have the following diagnostic plots.

```
par(mfrow=c(2,2))
plot(lm2)
```

Using Figure 3, we will check the following assumptions,

- *Linearity*
  - In the residuals vs fitted plot, the data follows a reasonably linear trend with an almost even spread of points above and below the horizontal line. Although this lack of linearity can be explained by the small data set, hence the data is approximately linear and the assumptions holds.
- *Homoscedasticity*
  - In the residuals vs fitted plot, there is considerable fanning in the points, particularly for the smaller fitted values. Similarly in the scale location plot, the data has some curvature present in it further implying a non-constant variance. However due to the small data set, this is insufficient evidence to invalidate the assumption of homoscedasticity.
- *Normality*
  - In the normal-QQ plot, we can see that the data follows an almost linear trend with significant deviation at the ends. However, the small data set has an even bigger impact on normality as it does not satisfy the conditions of the central limit theorem, so we have insufficient evidence to invalidate the assumptiosn of normality.

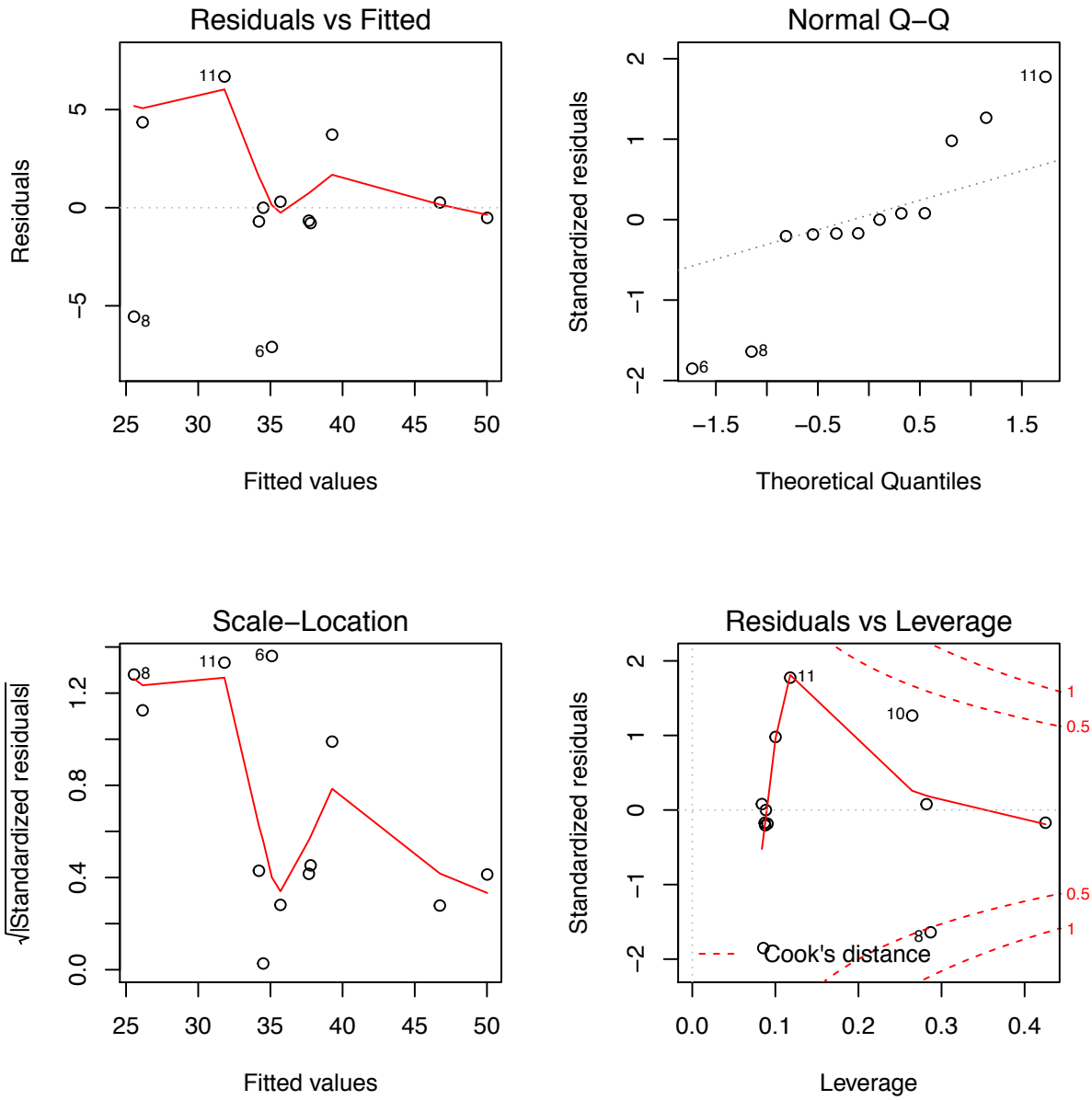


Figure 3: The four plots required to check the assumptions for the second linear model.

- *Independence*
  - The data used to create this model is the same as that used to create the multiple regression model. Since the assumption of independence is satisfied in the previous model it is also satisfied in this model.

It should be noted that there is one point in this model that has significant leverage, however once again due to the small number of data points, one outlier has a much bigger effect than for larger data sets.

For the simple regression model in just Weight, we have the following diagnostic plots.

```
par(mfrow=c(2,2))
plot(lm3)
```

Using Figure 4, we will check the following assumptions,

- *Linearity*
  - In the residuals vs fitted plot, the data follows a reasonably linear trend with an almost even spread of points above and below the horizontal line. This lack of linearity can be explained by the small data set, hence the data is approximately linear and the assumption holds.
- *Homoscedasticity*
  - In the residuals vs fitted plot, there is considerable fanning in the points, particularly for the smaller fitted values. Similarly, in the scale location plot the data has some curvature present in it further implying a non-constant variance. However, due to the small data set, this is insufficient evidence to invalidate the assumption of homoscedasticity.
- *Normality*
  - In the normal-QQ plot, we can see that the data follows an almost linear trend with some deviation at the ends. The small data set has an even bigger impact on normality as it does not satisfy the conditions of the central limit theorem. However this plot follows a significantly stronger trend of normality than the model with Height as a predictor.
- *Independence*
  - The data used to create this model is the same as that used to create the multiple regression model. Since the assumption of independence is satisfied in the previous model it is also satisfied in this model.

It should be noted that there is one point in this model that has significant leverage, however once again due to the small number of data points, one outlier has a much bigger affect than for larger data sets. Interestingly, this one point in both models corresponds to the same child, a 3.86kg baby.

Comparing Models

Now we can compare the coefficients of the predictor variables in the different linear models. Below are summaries of the three linear models.

```
summary(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
##
```

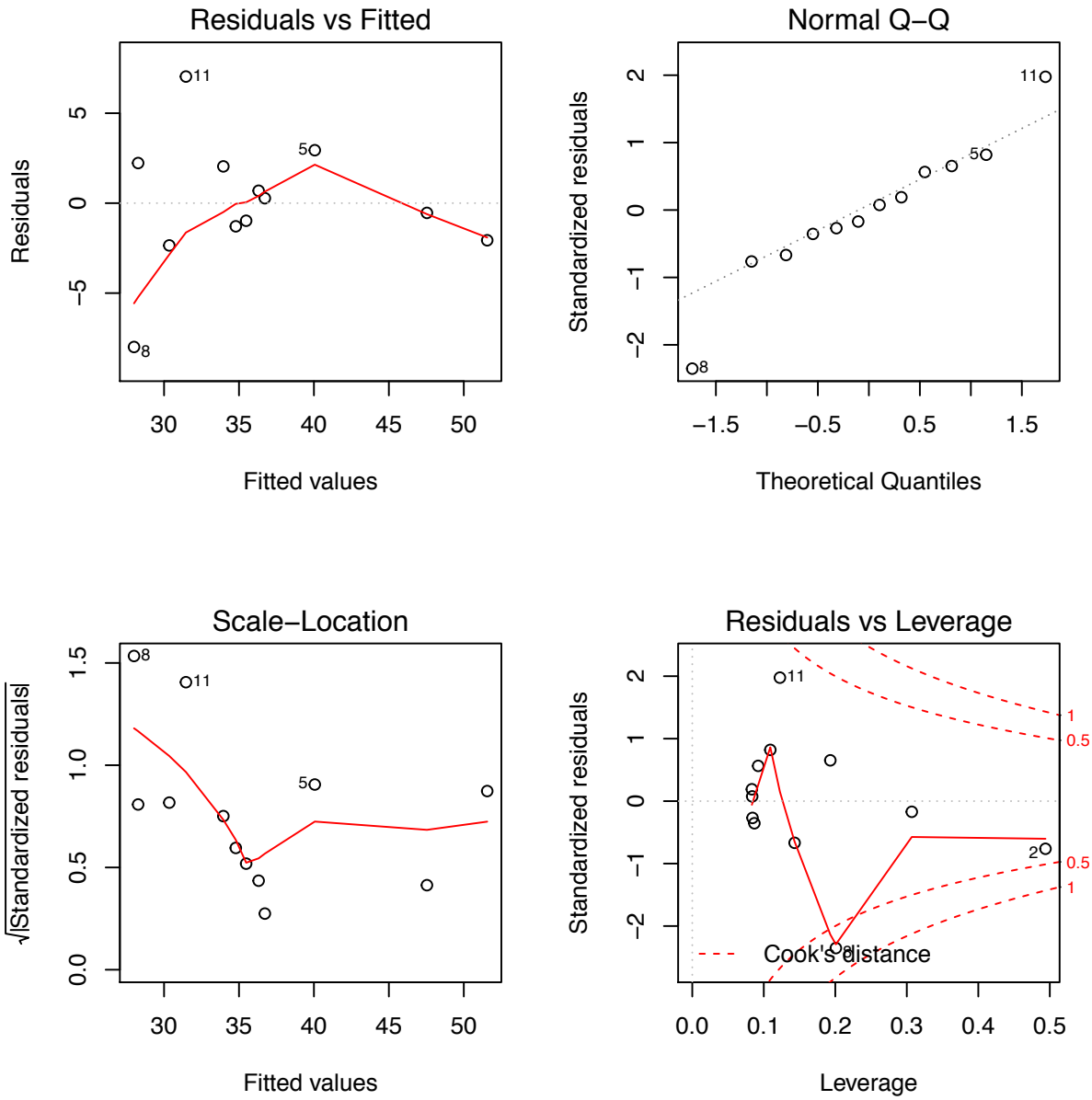


Figure 4: The four plots required to check the assumptions for the third linear model.

```
## 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
```

```r
summary(lm2)
```

```
##
## 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
```

```r
summary(lm3)
```

```
##
## 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
```

First we will consider the coefficients of Height, in the simple linear regression, we have $\beta_1 = 0.235$, where as in the multiple linear regression, we have, $\beta_1 = 0.077$. The significant difference in these values can be

explained by the inclusion of Weight in the linear model. For the simple linear regression model, the intercept coefficient is, $\beta_0 = 12.124$, where as in the multiple linear regression, the intercept coefficient is, $\beta_0 = 21.008$. If we naively ignore the inclusion of Weight in the model, the larger coefficient of Height in the simple linear regression corresponds to a smaller intercept and vice versa.

When we include Weight in the multiple regression model, it has coefficient, $\beta_2 = 0.421$, in constrast to the simple regression model where the coefficient is, $\beta_2 = 0.611$. Once again, the large coefficient corresponds to a smaller intercept coefficient and vice versa.

Now for both simple regression models, the coefficients of Height and Weight (independently) are statistically significant and should be included in the model. However, for the multiple regression model, both coefficients are statistically insignificant implying a strong correlation between these two predictors. In performing model selection, we would choose to drop one of the non-significant predictor variables in the multiple regression model and simplify it to one of the single regression models.

Interpreting Coefficients

As an example of coefficient interpretation, consider the coefficient of Weight in the two models that use it as a predictor.

For the simple regression model containing Weight, we can interpret the coefficient to mean that for any child, an increase in Weight by one (in kilograms) will result in an increase in length of the catheter tube by 0.611cm.

Similarly, for the multiple regression model the coefficient of Weight can be interpreted as, for an increase in one (kilogram) of Weight for a child of fixed Height, the length of the catheter tube will increase by 0.421cm.

Model as a Projection of Data on Subspaces

Up until now, we have considered the multiple linear regression as a linear function of the data, but it can also be thought of as projecting the data onto model subspaces. Let $\mathcal{L}_1$ be the model subspace corresponding to the simple linear regression with Height and $\mathcal{L}_2$ be the model subspace corresponding to the simple lienar regression with Weight.

$\mathcal{L}_1$ is the column space of the model matrix corresponding to the simple regression model with Height. Thus,

$$\mathcal{L}_1 = \text{span}\left\{ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} 108.7 \\ 161.29 \\ \vdots \\ 147.32 \end{pmatrix} \right\}.$$

Similarly, $\mathcal{L}_2$ is the column space of the model matrix corresponding to the simple regression model with Weight. Hence we have,

$$\mathcal{L}_2 = \text{span}\left\{ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} 18.14 \\ 42.41 \\ \vdots \\ 35.83 \end{pmatrix} \right\}.$$

Now consider the intersection of these two subspaces,

$$
\begin{aligned}
\mathcal{L}_1 \cap \mathcal{L}_2 &= \left\{ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} 108.7 \\ 161.29 \\ \vdots \\ 147.32 \end{pmatrix} \right\} \cap \left\{ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} 18.14 \\ 42.41 \\ \vdots \\ 35.83 \end{pmatrix} \right\} \\
&= \text{span} \left\{ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\} \\
&= \text{span}\{\mathbf{1}\}
\end{aligned}
$$

Thus,

$$
\{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} = \{\mathbf{1}\}^{\perp}.
$$

Now we need to consider the following subspaces,

$$
\begin{aligned}
\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} &= \text{span} \left\{ \begin{pmatrix} 108.7 \\ 161.29 \\ \vdots \\ 147.32 \end{pmatrix} \right\} \\
&= \boldsymbol{h} \\
\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} &= \text{span} \left\{ \begin{pmatrix} 18.14 \\ 42.41 \\ \vdots \\ 35.83 \end{pmatrix} \right\} \\
&= \boldsymbol{w},
\end{aligned}
$$

where $\boldsymbol{h}$ is the column vector containing the values for Height and $\boldsymbol{h}$ is the column vector containing the values for Weight.

These subspaces are the components in the models that are orthogonal to the intersection of the two models. That is to say it is the components that are not shared by either model, so if these two subspaces are orthogonal then the two models are disjoint.

Since these two subspaces are one-dimensional, we can calculate the angle between these two subspaces as follows.

$$
\begin{aligned}
\theta &= \cos^{-1} \left(\frac{\boldsymbol{h}^T \boldsymbol{w}}{||\boldsymbol{h}||||\boldsymbol{w}||} \right) \\
&= 0.322 \text{ radians.}
\end{aligned}
$$

Since the angle between these two subspaces is not $\frac{\pi}{2} \approx 1.57$, they are not perpendicular, and so the two predictor variables are correlated. This reuslt is not surprising since taller children are generally heavier and vice versa. This could also be observed in the postive linear relationship observed in the Height vs. Weight scatterplot when considering relationships between variables.

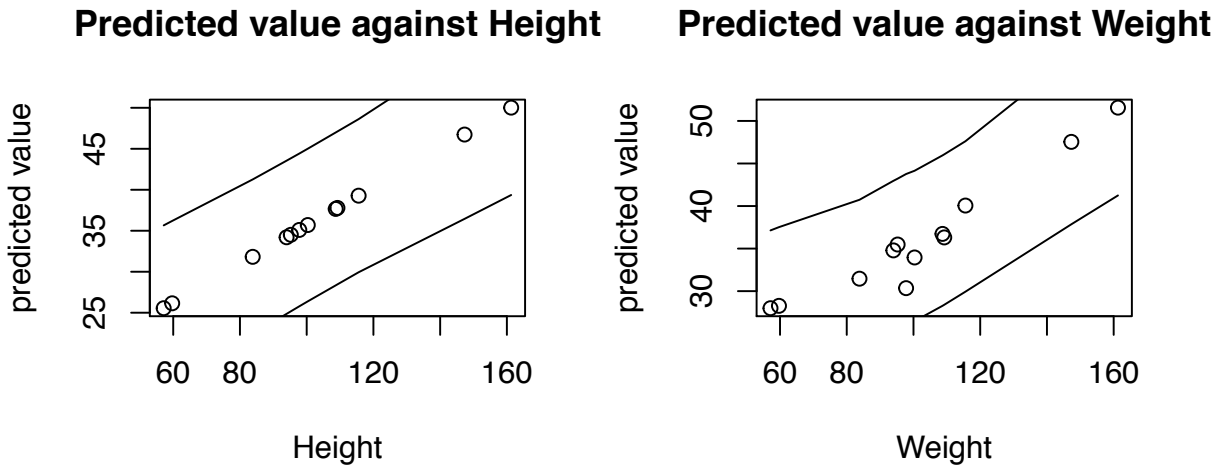


Figure 5: Prediction intervals plotted over the data for the two models under consideration.

Final Model Selection

We can now select a final model to be used in predicting the lengths of catheter required for children. In the multiple regression model, we saw that the coeffficients were statistically insignificant. This was then explained by looking into the angle between the two vector subspaces for Height and Weight, which showed that Height and Weight were moderately correlated. Therefore, it seems unnecessary to fit both Height and Weight as predictors. We will avoid overfitting by using only the one of the predictor variables.

To select between the two simple regression models, we can use the Akaike Information Criterion, AIC, which weighs up goodness of fit against the number of predictors. Since the number of predictors is the same between the two models, AIC is just a measure of goodness of fit. The two simple regression models, in terms of Height and Weight have AIC values, 71.19 and 69.92 respectively. Although the difference in these two values is very small, it implies that the simple regression model as a function of Weight is a slightly better fit than Height.

Since we want to use the model for prediction, we can comapre the prediction intervals for both models as in figure 5.

```
par(mfrow=c(1,2))

predicts.lm2 <- predict(lm2,interval="prediction")
plot(catheter$Height,predicts.lm2[,1],xlab = 'Height',ylab='predicted value',main='Predicted value agai
lines(lowess(catheter$Height,predicts.lm2[,2]))
lines(lowess(catheter$Height,predicts.lm2[,3]))

predicts.lm3 <- predict(lm3,interval="prediction")
plot(catheter$Height,predicts.lm3[,1],xlab = 'Weight',ylab='predicted value',main='Predicted value agai
lines(lowess(catheter$Height,predicts.lm3[,2]))
lines(lowess(catheter$Height,predicts.lm3[,3]))
```

Here we can see that the two plots are almost identical with a slightly tighter prediction interval for the model in terms of Weight. As a result, the prediction interval is slightly wider for more extreme values of Weight.

Another approach considered in determining which model we should choose is cross validation, however this approach requires a partitioning the data into training and testing sets. Since we have only 12 data points, this partitioning will leave us with too little training data and does not provide satisfactory comparison of the models.

All that has been discussed above highlights that the two simple regression models are nearly identical. However the model corresponding to Weight, satisfied the assumptions for linear regression slightly better than the model predicted by Height. Thus, the model we will choose to make the final predictions will be the second model, Length $= \beta_0 + \beta_2$Weight.

Logistic Regression

Breast cancer was the second most commonly diagnosed cancer in Australia in 2013 [1]. The most effective method for breast cancer screening is mammography. To confirm the diagnosis, invasive biopsies are performed. However, 70% of biopsies come back benign, indicating a high false positive rate in mammographies. To improve this process several computer-aided diagnosis (CAD) systems have been developed to help aid clinicians in making informed diagnoses.

The data in this examination contains 961 mammographic mass lesions with 445 of those lesions being malignant, given by the indicator variable Severity. Additionally for each of these lesions there are three attributes from the Breast Imaging Reporting and Data System (BI-RADS), including the lesion shape, the margin and the density. The data is summarised below.

Variable	Description
Severity	Indicator variable for malignant lesion (0=benign, 1=malignant)
Age	Age of patient in years
Shape	1=round, 2=oval, 3=lobular, 4=irregular
Margin	1=circumscribed, 2=microlobulated, 3=obscured, 4=ill-defined, 5=spiculated
Density	1=high, 2=iso, 3=low, 4=fat-containing

The dependent variable (Severity) is binary in this case. We can also assume independence between the observations as each oberservation is a different individual (assumed to not be related) and breast cancer is not contagious. Therefore, we can use logistic regression on this data.

We will attempt to develop a logistic predictive model for mammographic mass severity using the available predictor variables and to obtain predicted probabilities of mass severity that can used by clinicians to make informed diagnoses.

Data cleaning

```r
mammo <- read.csv('mammo.txt', header = TRUE)
head(mammo)
```

```
##   BI.RADS Age Shape Margin Density Severity
## 1       5  67     3      5       3        1
## 2       4  43     1      1       ?        1
## 3       5  58     4      5       3        1
## 4       4  28     1      1       3        0
## 5       5  74     1      5       ?        1
## 6       4  65     1      ?       3        0
```

A simple inspection of the data makes it clear that cleaning is required. There is the incorrect classes of several of the variables and, as can be seen below, there are a number of data points that are missing certain attributes. These are currently set to "?". but should be set to "NA", to comply with R's notation for missing data. These data points could be removed as they are missing some data, however this should not be done

[1] Australian Institute of Health and Welfare, Cancer compendium: information and trends by cancer type, https://www.aihw.gov.au/reports/cancer/cancer-compendium-information-and-trends-by-cancer-type/report-contents/breast-cancer-in-australia, [Accessed May 2018].

for two reasons. Firstly, incomplete data may still contain valuable information. Secondly, the final model may not include some of the predictor attributes, and so some incomplete data may actually be complete for the predictors used in the final model.

```r
table(mammo$Severity)
```

```
##
##   0   1
## 516 445
```

Severity has no missing data.

```r
table(mammo$Age)
```

```
##
##  ? 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
##  5  1  4  1  6  3  7  4  3  2  6  5  3  3  7  6  9  9 13 11  8  9 11 19 16
## 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66
## 19 20 18 21 28 13 11 23 21 16 20 24 25 26 23 32 23 36 25 13 25 24 27 25 31
## 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 93 96
## 32 20 12 13 22 17  9 12  7 17  9  4  8 11  2  1  5  3  4  5  5  1  1  1
```

Age has 5 missing data points.

```r
table(mammo$Shape)
```

```
##
##   ?   1   2   3   4
##  31 224 211  95 400
```

Shape has 31 missing data points.

```r
table(mammo$Margin)
```

```
##
##   ?   1   2   3   4   5
##  48 357  24 116 280 136
```

Margin has 48 missing data points.

```r
table(mammo$Density)
```

```
##
##   ?   1   2   3   4
##  76  16  59 798  12
```

Density has 76 missing data points.

```r
table(mammo$BI.RADS)
```

```
##
##   ?   0   2   3   4   5  55   6
##   2   5  14  36 547 345   1  11
```

BI-RADS has 2 missing data points, but this class will not be used in our model, and is not of high importance.

This data is now cleaned by setting "?"'s to NA's and fixing the attribute classes.

```r
class(mammo$Age)
```

```
## [1] "factor"
```

```r
mammo$Age[mammo$Age == "?"] <- NA
mammo$Age <- as.numeric(mammo$Age)
summary(mammo$Age, exclude = FALSE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00   29.00   41.00   39.48   50.00   74.00       5
```

Age has its 5 missing data points set to NA and the variable is set to numeric.

```r
class(mammo$Shape)
```

```
## [1] "factor"
```

```r
mammo$Shape <- as.character(mammo$Shape)
mammo$Shape[mammo$Shape == "?"] <- NA
mammo$Shape <- factor(mammo$Shape)
summary(mammo$Shape, exclude = FALSE)
```

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

Shape has its 31 missing data points set to NA and the variable is set to a factor

```r
class(mammo$Margin)
```

```
## [1] "factor"
```

```r
mammo$Margin <- as.character(mammo$Margin)
mammo$Margin[mammo$Margin == "?"] <- NA
mammo$Margin <- factor(mammo$Margin)
summary(mammo$Margin, exclude = FALSE)
```

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

Margin has its 48 missing data points set to NA and the variable is set to a factor

```r
class(mammo$Density)
```

```
## [1] "factor"
```

```r
mammo$Density <- as.character(mammo$Density)
mammo$Density[mammo$Density == "?"] <- NA
mammo$Density <- factor(mammo$Density)
summary(mammo$Density, exclude = FALSE)
```

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

Density has its 76 missing data points set to NA and the variable is set to a factor

```r
class(mammo$Severity)
```

```
## [1] "integer"
```

```r
mammo$Severity <- as.numeric(mammo$Severity)
summary(mammo$Severity, exclude = FALSE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4631  1.0000  1.0000
```

Severity is set to a integer.

Finally, BI-RADS has it's 2 missing data points set to NA. Additionally, the data has a clear outlier in it that is set to NA as well. Again, this is not overly important as BI-RADS will not be used as a predictor in this model.

```
class(mammo$BI.RADS)
```

```
## [1] "factor"
```

```
mammo$BI.RADS <- as.character(mammo$BI.RADS)
mammo$BI.RADS[mammo$BI.RADS == "?"] <- NA
mammo$BI.RADS <- factor(mammo$BI.RADS)
summary(mammo$BI.RADS, exclude = FALSE) # 2 NAs, 1 outlier
```

```
##    0    2    3    4    5   55    6 NA's
##    5   14   36  547  345    1   11    2
```

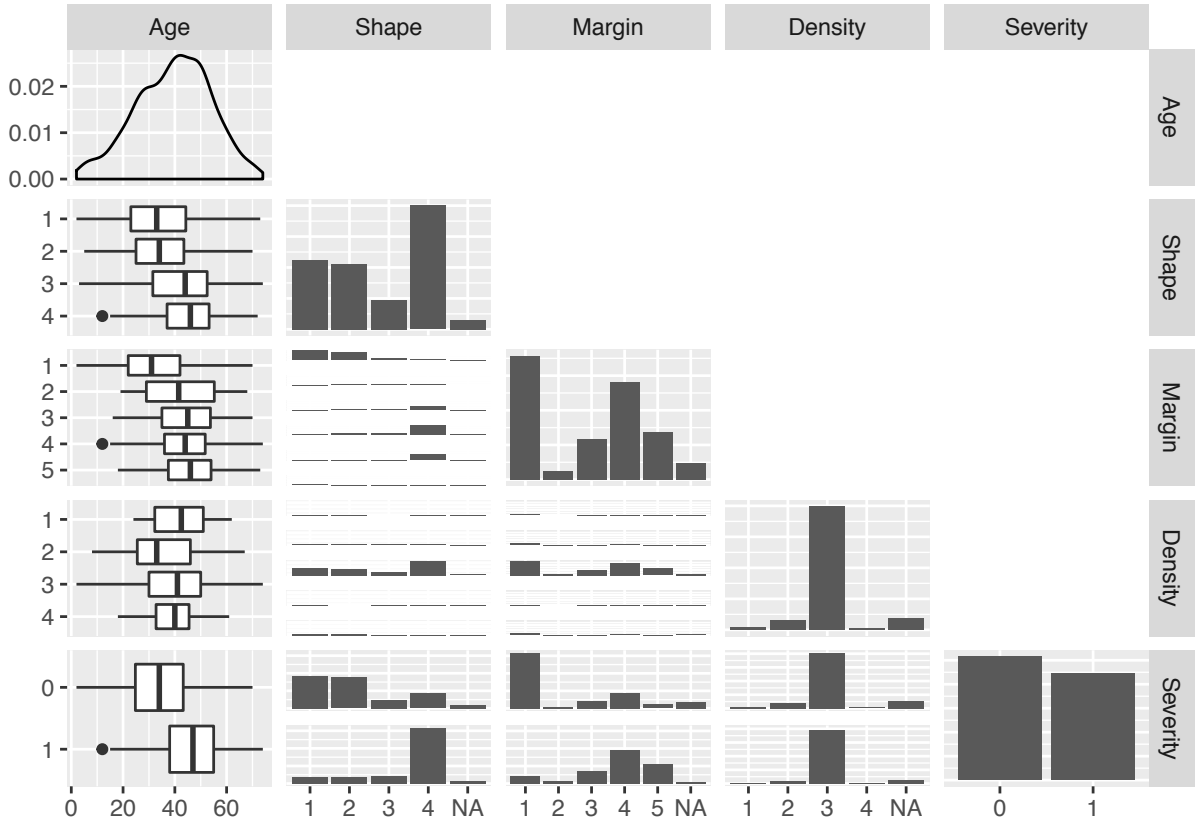
```
mammo$BI.RADS[mammo$BI.RADS == 55] <- NA # Set outlier to NA
mammo$BI.RADS <- as.numeric(mammo$BI.RADS)
summary(mammo$BI.RADS, exclude = FALSE) # 3 NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   4.000   4.312   5.000   7.000       3
```

Data Visualisation

To examine the relationship between each of the individual variables, a plot of the relationships between the variable is made.

```
mammo %>%
  mutate(Severity = as.factor(Severity)) %>%
  select(2:6) %>%
  ggpairs(upper = list(continuous = "blank",
                       combo ="blank",
                       discrete = "blank",
                       na = "blank"),
          lower = list(continuous = "cor",
                       combo = "box_no_facet",
                       discrete = "facetbar",
                       na = "na"))
```



Age is mostly normally distributed around a mean of 39.48. Shape and Margin have a slight linear increase with age but Density does not. Severity appears to have a correlation between being malignant and a higher age.

Shape has a large portion in the lobular category, this category also has a high proportion of malignancy.

Margin has a strong correlation between the first category, circumscribed, and being benign.The second category has very little data and may not be of much value to the model.

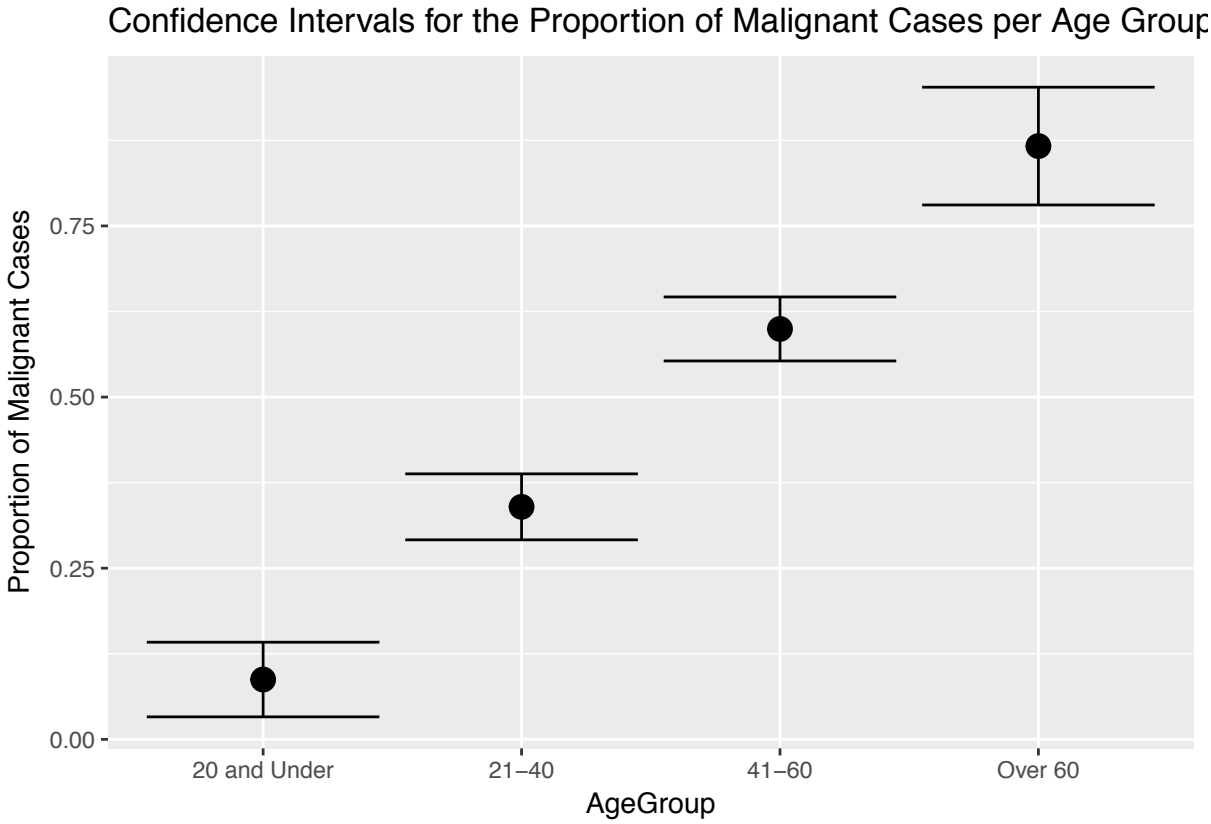
Density has a overwhelming popularity of the third class of "low". This makes it difficult to mind any significant findings with respect to how the classes effect the Severity.

Finally Severity has a mostly balanced proportion of benign to malignant which is ideal for fitting an accurate model.

Confidence intervals

```
mammo$ageGroup <- cut(mammo$Age,
                        breaks = c(0,20,40,60,80),
                        labels = c("20 and Under","21-40","41-60","Over 60"))

confidenceInt <- mammo %>%
  split(.$ageGroup) %>%
  map_df(~conf_int_prop(.$Severity), .id = "AgeGroup")

ggplot(confidenceInt, aes(x = AgeGroup, y = Proportion)) +
  geom_point(size = 4) +
  geom_errorbar(aes(ymax = Upper, ymin = Lower)) +
```

```
  labs(y = "Proportion of Malignant Cases") +
  ggtitle("Confidence Intervals for the Proportion of Malignant Cases per Age Group")
```



First, we looked at the confidence intervals for age. To do this, we had to group ages, as finding confidence intervals for each value given in age would be impractical and largely unhelpful. Each interval shows where we expect the true proportion of malignant masses for that age group. For example, we can say that we are 95% confident the true proportion of people aged 20 and under with a malignant mass lies between 3% and 14%. From the plot, we can clearly see that, as the age of patients increases, the proportion of cases with malignant masses also increase. This means that in patients over 40, more than half of the lesions are malignant. It is fair to predict that this variable will be influential in the model.

```
mammo$Shape.orig <- mammo$Shape

confidenceInt <- mammo %>%
  split(.$Shape.orig) %>%
  map_df(~conf_int_prop(.$Severity), .id = "Shape")

confidenceInt %>% knitr::kable(caption = "Table of the confidence intervals for the proportion of malig
```

Table 1: Table of the confidence intervals for the proportion of malignant cases per shape of the variable shape

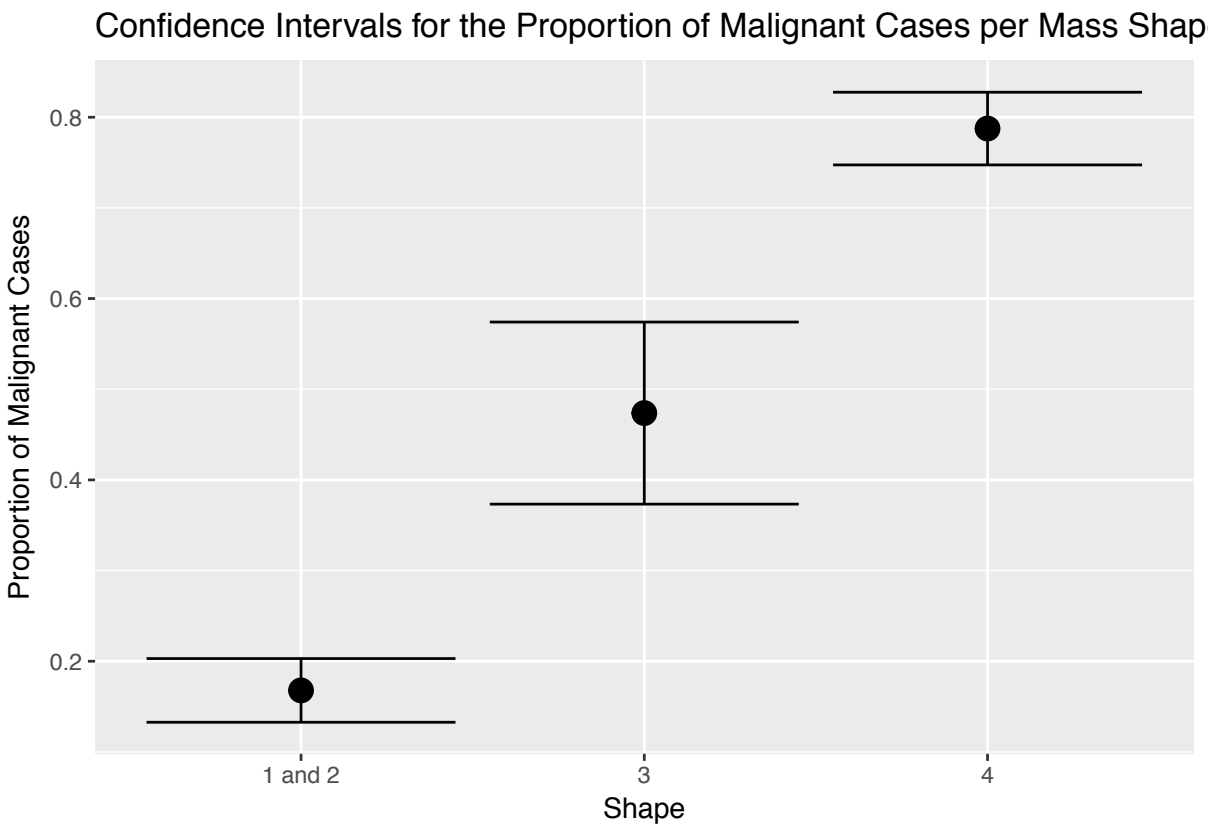
Shape	MalignantCases	TotalCases	Lower	Proportion	Upper
1	38	224	0.1204928	0.1696429	0.2187929
2	35	211	0.1156871	0.1658768	0.2160665
3	45	95	0.3732795	0.4736842	0.5740889
4	315	400	0.7474112	0.7875000	0.8275888

Our next variable, Shape provided some more challenges. Our first attempt at creating confidence intervals for the shape of the mass showed that the true proportion of cases with malignant lesions with round and oval shaped masses lies roughly between 16% and 22% for both categories. Basic geometry knowledge tells us that these two shapes are very similar, therefore we will combine these categories.

```
levels(mammo$Shape)[levels(mammo$Shape)=="1"] <- "1 and 2"
levels(mammo$Shape)[levels(mammo$Shape)=="2"] <- "1 and 2"
table(mammo$Shape)
```

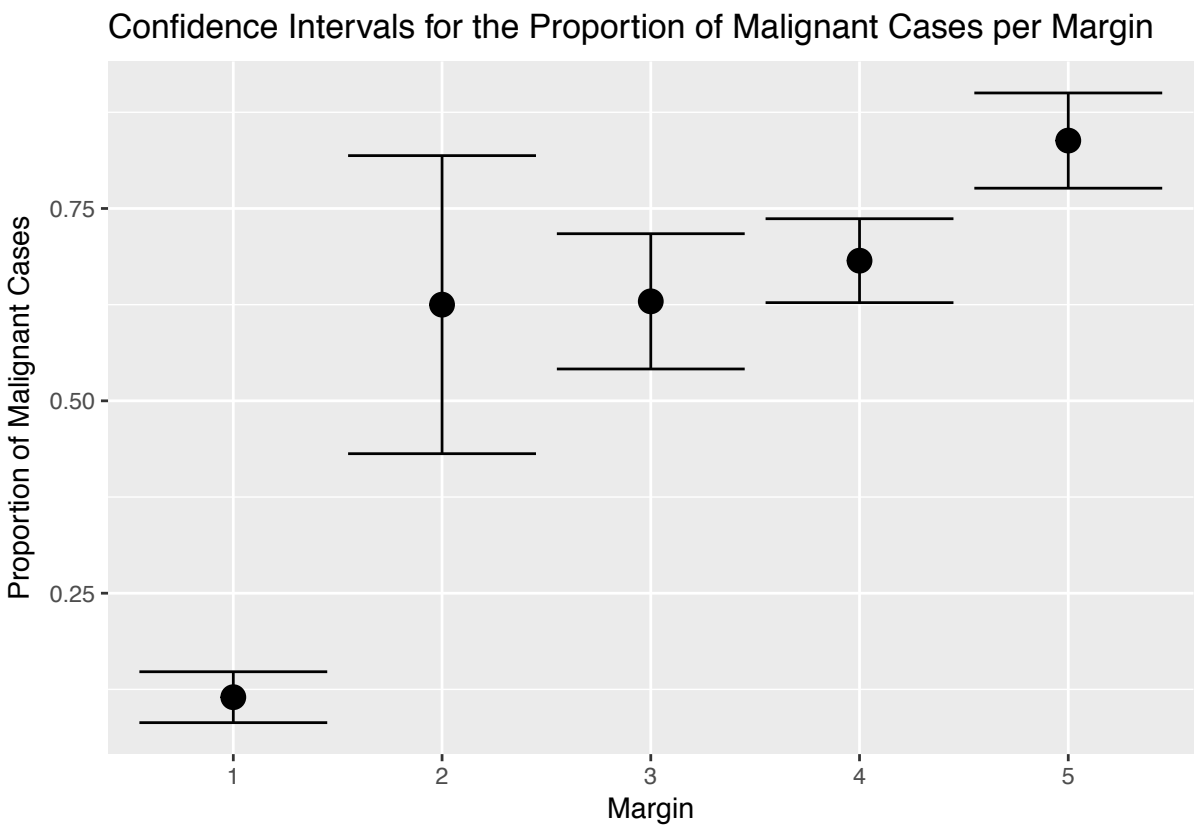
```
##
## 1 and 2       3       4
##     435      95     400
```

```
confidenceInt <- mammo %>%
  split(.$Shape) %>%
  map_df(~conf_int_prop(.$Severity), .id = "Shape")

ggplot(confidenceInt, aes(x = Shape, y = Proportion)) +
  geom_point(size = 4) +
  geom_errorbar(aes(ymax = Upper, ymin = Lower)) +
  labs(y = "Proportion of Malignant Cases") +
  ggtitle("Confidence Intervals for the Proportion of Malignant Cases per Mass Shape")
```



Now, looking at the confidence intervals above, we can see that more patients with lobular and irregular shaped masses, categories 3 and 4 respectively, had a malignant mass than those with round or ovular masses. As with age, this clear trend seems to indicate that shape may provide useful information when predicting severity of lesions, and therefore will probably be useful in any models we create.

```
confidenceInt <- mammo %>%
  split(.$Margin) %>%
  map_df(~conf_int_prop(.$Severity), .id = "Margin")

ggplot(confidenceInt, aes(x = Margin, y = Proportion)) +
  geom_point(size = 4) +
  geom_errorbar(aes(ymax = Upper, ymin = Lower)) +
  labs(y = "Proportion of Malignant Cases") +
  ggtitle("Confidence Intervals for the Proportion of Malignant Cases per Margin")
```



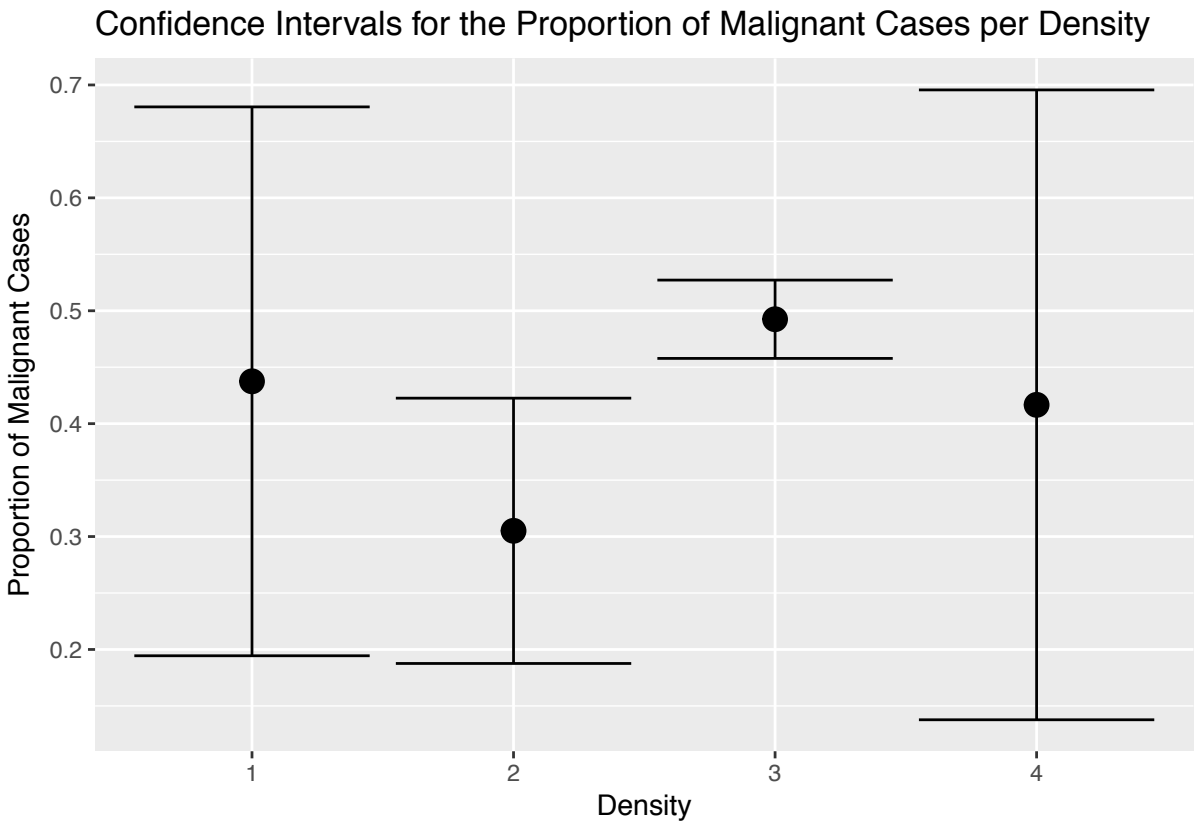
The confidence intervals for the variable Margin contain more overlap than those for Age and Shape. The large size of the confidence interval for category 2 is a result of the limited number of cases with this margin in our data set. Here we can see the true proportion of malignant cases with margin 2, 3, 4, and 5 fall within a similar range, while the proportion for those with margin 1 falls in a very different range. This makes it harder to predict whether this variable will be useful in a model as 4 of the 5 categories have similar outcomes of Severity.

```
confidenceInt <- mammo %>%
  split(.$Density) %>%
  map_df(~conf_int_prop(.$Severity), .id = "Density")

confidenceInt %>% knitr::kable(caption = "Table of the confidence intervals for the proportion of malig
```

Table 2: Table of the confidence intervals for the proportion of malignant cases per category of the variable density

Density	MalignantCases	TotalCases	Lower	Proportion	Upper
1	7	16	0.1944261	0.4375000	0.6805739
2	18	59	0.1875955	0.3050847	0.4225740
3	393	798	0.4577941	0.4924812	0.5271683
4	5	12	0.1377270	0.4166667	0.6956063

```
ggplot(confidenceInt, aes(x = Density, y = Proportion)) +
  geom_point(size = 4) +
  geom_errorbar(aes(ymax = Upper, ymin = Lower)) +
  labs(y = "Proportion of Malignant Cases") +
  ggtitle("Confidence Intervals for the Proportion of Malignant Cases per Density")
```



Finally, we created confidence intervals for the variable Density. This contains little useful information. The ranges for the confidence intervals for categories 1 and 4 pretty much span the entire range of possible proportions. This could be because there is no relationship between these densities and malignant cases, or it could be a result of the how few patients had these densities. As a result, we can predict that density probably will not be a useful variable in our model.

Making a Model

Using the right data

Firstly, we are only able to fit a model to entries in the data frame that are complete. To do this, we created a logical vector for all the data that references which data entries have information for all variables.

```
complete=!is.na(mammo$Age)&!is.na(mammo$Shape)&!is.na(mammo$Margin)&!is.na(mammo$Density)
```

The Full Model

Firstly, we fit a model that uses all the predictor variables to model our response variable, Severity. Of the 4 predictor variables, only Age is a numerical variable. Shape, Margin, and Density are all categorical variables which contain levels.

In R, when fitting a model to categorical variables, the model uses a reference category. The reference category for Shape should clearly be the combined 1 and 2 level, as each level seems to show a distinct difference in the proportion of malignant masses, and this combined level contains the greatest number of data entries.

However, for Margin, choosing a reference category is more difficult. The default reference level 1 is not the best selection. As we saw from the confidence intervals, category 1 has a relatively low proportion of malignant mass cases, while categories 2,3,4, and 5 do not deviate from one another all that much. With 1 as the reference level, all the other levels appear significantly different, when really it is 1 that is odd. Hence we must decide on another reference level. Category 4 has the largest number of data points, and the true proportion of malignant cases with a margin of 4 has a narrow range of possibilities relative to the other categories, meaning any other facets of the data will be more prevalent when fitting the model. Therefore this is the reference point we will use.

```
levels(mammo$Shape)[levels(mammo$Shape)=="1"] <- "1 and 2"
levels(mammo$Shape)[levels(mammo$Shape)=="2"] <- "1 and 2"

mammo$Margin <- relevel(mammo$Margin, ref="1")

mod.full1 <- glm(formula = Severity ~ Age + Shape + Margin + Density,
                family = binomial,
                data = mammo[complete,])

table(mammo$Margin)
```

```
##
##   1   2   3   4   5
## 357  24 116 280 136
```

```
mammo$Margin <- relevel(mammo$Margin, ref="4")

mod.full4 <- glm(formula = Severity ~ Age + Shape + Margin + Density,
                  family = binomial,
                  data = mammo[complete,])


summary(mod.full1)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density, family = binomial,
##     data = mammo[complete, ])
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5368  -0.5647  -0.2162   0.6637   2.5016
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.398139   0.780114  -4.356 1.32e-05 ***
## Age          0.055034   0.007803   7.053 1.76e-12 ***
## Shape3       0.816696   0.321534   2.540  0.01109 *
## Shape4       1.534198   0.263437   5.824 5.75e-09 ***
## Margin2      1.584004   0.551647   2.871  0.00409 **
## Margin3      1.127629   0.343066   3.287  0.00101 **
## Margin4      1.425197   0.291192   4.894 9.86e-07 ***
## Margin5      1.974963   0.370198   5.335 9.56e-08 ***
## Density2    -0.994808   0.796858  -1.248  0.21188
## Density3    -0.676635   0.719649  -0.940  0.34710
## Density4    -1.769463   1.065232  -1.661  0.09669 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1151.26  on 830  degrees of freedom
## Residual deviance:  727.73  on 820  degrees of freedom
## AIC: 749.73
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(mod.full4)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density, family = binomial,
##     data = mammo[complete, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5368  -0.5647  -0.2162   0.6637   2.5016
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.972942   0.841265  -2.345   0.0190 *
## Age          0.055034   0.007803   7.053 1.76e-12 ***
## Shape3       0.816696   0.321534   2.540   0.0111 *
## Shape4       1.534198   0.263437   5.824 5.75e-09 ***
## Margin1     -1.425197   0.291192  -4.894 9.86e-07 ***
## Margin2      0.158807   0.531245   0.299   0.7650
## Margin3     -0.297568   0.270079  -1.102   0.2706
## Margin5      0.549767   0.293322   1.874   0.0609 .
## Density2    -0.994808   0.796858  -1.248   0.2119
## Density3    -0.676635   0.719649  -0.940   0.3471
## Density4    -1.769463   1.065232  -1.661   0.0967 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1151.26  on 830  degrees of freedom
## Residual deviance:  727.73  on 820  degrees of freedom
## AIC: 749.73
##
## Number of Fisher Scoring iterations: 5
```

```r
mod.full <- mod.full4
```

Model by stepwise selection

We will use step-wise selection to choose which predictors are significant in the model, based on the AIC values.

```r
mod.step <- step(mod.full, direction = "both", trace = 0)
summary(mod.step)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin, family = binomial,
##     data = mammo[complete, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5302  -0.5637  -0.2212   0.6699   2.5034
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.695149   0.410359  -6.568 5.11e-11 ***
## Age          0.055413   0.007762   7.139 9.43e-13 ***
## Shape3       0.778804   0.319330   2.439   0.0147 *
## Shape4       1.522701   0.262822   5.794 6.89e-09 ***
## Margin1     -1.391147   0.288084  -4.829 1.37e-06 ***
## Margin2      0.193505   0.531686   0.364   0.7159
## Margin3     -0.273404   0.269301  -1.015   0.3100
## Margin5      0.563723   0.291981   1.931   0.0535 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1151.26  on 830  degrees of freedom
## Residual deviance:  731.14  on 823  degrees of freedom
## AIC: 747.14
##
## Number of Fisher Scoring iterations: 5
```

Model by removal of non-signficant terms

Here, we will be creating a model by removing the least significant variable. The first iteration was simple as none of the coefficients for Density were significant. We then also removed Margin, as only 2 of the 4 coefficients were statistically significant.

```r
mod.back <- mod.full

summary(mod.back)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density, family = binomial,
##     data = mammo[complete, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5368  -0.5647  -0.2162   0.6637   2.5016
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.972942   0.841265  -2.345   0.0190 *
## Age          0.055034   0.007803   7.053 1.76e-12 ***
## Shape3       0.816696   0.321534   2.540   0.0111 *
## Shape4       1.534198   0.263437   5.824 5.75e-09 ***
## Margin1     -1.425197   0.291192  -4.894 9.86e-07 ***
## Margin2      0.158807   0.531245   0.299   0.7650
## Margin3     -0.297568   0.270079  -1.102   0.2706
## Margin5      0.549767   0.293322   1.874   0.0609 .
## Density2    -0.994808   0.796858  -1.248   0.2119
## Density3    -0.676635   0.719649  -0.940   0.3471
## Density4    -1.769463   1.065232  -1.661   0.0967 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1151.26  on 830  degrees of freedom
## Residual deviance:  727.73  on 820  degrees of freedom
## AIC: 749.73
##
## Number of Fisher Scoring iterations: 5
```

```r
mod.back <- update(mod.back, .~. - Density)
summary(mod.back)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin, family = binomial,
##     data = mammo[complete, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5302  -0.5637  -0.2212   0.6699   2.5034
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.695149   0.410359  -6.568 5.11e-11 ***
## Age          0.055413   0.007762   7.139 9.43e-13 ***
## Shape3       0.778804   0.319330   2.439   0.0147 *
```

```
## Shape4       1.522701   0.262822   5.794 6.89e-09 ***
## Margin1     -1.391147   0.288084  -4.829 1.37e-06 ***
## Margin2      0.193505   0.531686   0.364   0.7159
## Margin3     -0.273404   0.269301  -1.015   0.3100
## Margin5      0.563723   0.291981   1.931   0.0535 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1151.26  on 830  degrees of freedom
## Residual deviance:  731.14  on 823  degrees of freedom
## AIC: 747.14
##
## Number of Fisher Scoring iterations: 5
```

```r
mod.back <- update(mod.back, .~. - Margin)
summary(mod.back)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape, family = binomial, data = mammo[complete,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4501  -0.6593  -0.2502   0.6981   2.4040
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.879346   0.334745 -11.589  < 2e-16 ***
## Age          0.061589   0.007444   8.273  < 2e-16 ***
## Shape3       1.386747   0.283167   4.897 9.72e-07 ***
## Shape4       2.518659   0.195455  12.886  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1151.26  on 830  degrees of freedom
## Residual deviance:  765.68  on 827  degrees of freedom
## AIC: 773.68
##
## Number of Fisher Scoring iterations: 5
```

Choosing models

Now we have three models to choose from. Each smaller model only required the removal of one variable. First, we wanted to ensure that removing this variable actually had a significant impact on the model. We tested if we could set the coefficients for that variable equal to 0. This could be found by comparing the difference in residual deviances to the 95th percentile of the chi squared distributions.

```r
qchisq(1-0.05, df = 3)
```

```
## [1] 7.814728
```

```r
mod.step$deviance - mod.full$deviance
```

```
## [1] 3.405636
```

```r
anova(mod.step, mod.full) # Accept hypothesis that all coefficients are equal
```

```
## Analysis of Deviance Table
##
## Model 1: Severity ~ Age + Shape + Margin
## Model 2: Severity ~ Age + Shape + Margin + Density
##   Resid. Df Resid. Dev Df Deviance
## 1       823     731.14
## 2       820     727.73  3   3.4056
```

We first compared our two largest models, the stepwise selection model and the full model. The only difference between these models the stepwise model removes the variable Density. As we can see, the difference in residual deviance, approximately 3.405, is well within the 95th percentile, therefore including Density does not significantly alter the residuals of the model. Since it is not providing any useful information, its inclusion only makes the model larger and more inefficient, meaning the stepwise selection model is the better model.

```r
qchisq(1-0.05, df = 4)
```

```
## [1] 9.487729
```

```r
mod.back$deviance - mod.step$deviance
```

```
## [1] 34.53966
```

```r
anova(mod.back, mod.step)
```

```
## Analysis of Deviance Table
##
## Model 1: Severity ~ Age + Shape
## Model 2: Severity ~ Age + Shape + Margin
##   Resid. Df Resid. Dev Df Deviance
## 1       827     765.68
## 2       823     731.14  4    34.54
```

Next we compared the stepwise selection model to the backwards selection model, which omits the variables Margin and Density. Here we can see that the 95th percentile value is 9.488, while the difference in residual deviance is 34.539, much higher than the 95th percentile value. Therefore, we reject the null hypothesis which states that the values of the coefficients for the variable Margin could be equal to 0 without significantly impacting the model. As this is not the case, we can see that including Margin in the model changes the model signficantly.

```r
AIC(mod.full, mod.step, mod.back)
```

```
##          df      AIC
## mod.full 11 749.7331
## mod.step  8 747.1388
## mod.back  4 773.6784
```

A final way to compare models is to use the Akaike Information Criterion or AIC. When we compare the AIC values for each model, we can see that the full model and the stepwise model are very close in value, with stepwise just a little lower. While this is not a significant difference, as we have already shown previously, including the extra variable Density in the full model does not significantly influence the results. Thus, the best model for predicting whether a mass is malignant or not is the stepwise model.

Prediction to verify model

Another way of assessing the goodness of fit of a model it to see how well predicts values. This works especially well for predicting binary outcomes as you can calculate a simple proportion of correct predictions. Firstly we predicted values for the data that was used to fit the model. This gives us an about 81.5% success rate.

```
predict <- predict(mod.step, newdata = mammo ,type="response")
predict.df <- data.frame(predict.prob = predict)

predict.df.indexed <- data.frame(predict.df, id = row.names(predict.df))
mammo.indexed <- data.frame(mammo, id = row.names(mammo))

mammo.predict <- left_join(mammo.indexed, predict.df.indexed, by="id")
mammo.predict <- mammo.predict[ , !names(mammo.predict) %in% c("id")]


#Calculate the percentage that our model correctly predicts Severity
mammo.predict %>%
  mutate(predict = (predict.prob >= 0.5),
         predict = as.integer(predict),
         correct = (predict == Severity)) %>%
  count(correct) %>%
  summarise(hit.rate = n[2]/(n[1] + n[2])) %>%
  first() %>%
  round(3)
```

```
## [1] 0.815
```

The second approach is to fit the model to only half of the data we have available and then attempt to predict the values of the other half of the data. Doing this we get a success rate of about 81% which is barely lower than above. This is evidence to justify our model as valid and useful for prediction.

```
train <- slice(mammo[complete,], 1:400)
test <- slice(mammo[complete,], 401:831)

mod.train <- glm(Severity ~ Age + Shape + Margin, data = train, family = "binomial")

predict <- predict(mod.train, newdata = test ,type="response")
predict.df <- data.frame(predict.prob = predict)

predict.df.indexed <- data.frame(predict.df, id = row.names(predict.df))
mammo.indexed <- data.frame(test, id = row.names(test))

mammo.predict <- left_join(mammo.indexed, predict.df.indexed, by="id")
mammo.predict <- mammo.predict[ , !names(mammo.predict) %in% c("id")]


#Calculate the percentage that our model correctly predicts Severity
mammo.predict %>%
  mutate(predict = (predict.prob >= 0.5),
         predict = as.integer(predict),
         correct = (predict == Severity)) %>%
  count(correct) %>%
  summarise(hit.rate = n[2]/(n[1] + n[2])) %>%
  first() %>%
  round(3)
```

```
## [1] 0.81
```

Prediction can also be used to inform clinical decisions. To that end we produced a table that gives predictive probabilities for different ages, shapes and margins.

```
for (age in seq(from = 20, to = 80, by = 10)) {
pred1 <- predict(mod.step, newdata = data.frame(Age = age, Margin = c("1", "2", "3", "4", "5"), Shape =
pred3 <- predict(mod.step, newdata = data.frame(Age = age, Margin = c("1", "2", "3", "4", "5"), Shape =
pred4 <- predict(mod.step, newdata = data.frame(Age = age, Margin = c("1", "2", "3", "4", "5"), Shape =

pred <- cbind(pred1, pred3, pred4)

diag.table <- data.frame(round(pred, 2), row.names = c("Circumscribed", "Microlobulated", "Obscured", "
colnames(diag.table) <- c("Round or Oval", "Lobular", "Irregular")
tab <- xtable(diag.table, caption = paste('Predictive probabilities for Age =', age))
print(tab, type="latex")
}
```

	Round or Oval	Lobular	Irregular
Circumscribed	0.05	0.10	0.19
Microlobulated	0.20	0.35	0.53
Obscured	0.13	0.25	0.42
Ill-Defined	0.17	0.31	0.48
Spiculated	0.26	0.44	0.62

Table 3: Predictive probabilities for Age = 20

	Round or Oval	Lobular	Irregular
Circumscribed	0.08	0.16	0.29
Microlobulated	0.30	0.48	0.66
Obscured	0.21	0.37	0.55
Ill-Defined	0.26	0.44	0.62
Spiculated	0.38	0.58	0.74

Table 4: Predictive probabilities for Age = 30

	Round or Oval	Lobular	Irregular
Circumscribed	0.13	0.25	0.41
Microlobulated	0.43	0.62	0.78
Obscured	0.32	0.51	0.68
Ill-Defined	0.38	0.57	0.74
Spiculated	0.52	0.70	0.83

Table 5: Predictive probabilities for Age = 40

These tables represent the probability that a growth is malignant based on Shape and Margin for different ages.

Teamwork Reflection

As this project was completed in a team of five people, collaboration and communication were key. The first meeting was undertaken at the university campus where we all met face to face to partition the workload evenly between all members. The division of labour was to have two people work on the first section (linear

	Round or Oval	Lobular	Irregular
Circumscribed	0.21	0.37	0.55
Microlobulated	0.57	0.74	0.86
Obscured	0.45	0.64	0.79
Ill-Defined	0.52	0.70	0.83
Spiculated	0.65	0.81	0.90

Table 6: Predictive probabilities for Age = 50

	Round or Oval	Lobular	Irregular
Circumscribed	0.32	0.50	0.68
Microlobulated	0.69	0.83	0.91
Obscured	0.59	0.76	0.87
Ill-Defined	0.65	0.80	0.90
Spiculated	0.77	0.88	0.94

Table 7: Predictive probabilities for Age = 60

regression) and have the remaining three work on the second section (logistic regression). The finer division will be discussed in more detail later in the personal statements of contribution.

After the initial meeting and distribution of work, we created a group chat that allowed discussion of ideas and an easy medium to send files. Similarly, we created a Git repository that contained the code and R-Markdown files which allowed each member to keep up to date with the files easily and manage version control (which did come in handy once or twice).

Because this group consists of people who had met before and are close friends, it made communicating trivial as there was already a rapport. The group working on the first section met two more times after the initial meeting to discuss plans for section, while the group working on the second section met three more times for a similar purpose.

Once each section was completed independently, one person was tasked with taking the two section and ensuring the final report flowed appropriately. Then each member was given an opportunity to read over the final report and correct any mistakes or include more detail in places where it was required.

We held our initial meeting towards the end of Week 10, where we delegated groups to Sections. We agreed to complete both sections by Wednesday Week 12. Then the two sections were combined and edited by the whole group by Friday Week 12. We then used final weekend before the submission date to do final editing.

Contribution of Group Members

Miriam Slattery

I was tasked with working on Section 1 (linear regression), with Joshua Bean. We met to analyse the data and develop the code for this section together. In particular, I wrote about the model assumptions. I also wrote the Introduction and combined the two sections into a cohesive report.

Joshua Bean

My contribution to this report was to co-write section 2 with Miriam Slattery, we met together to complete the code. Then I wrote about the model creation and analyses, culminating in the conclusion about the final model chosen.

	Round or Oval	Lobular	Irregular
Circumscribed	0.45	0.64	0.79
Microlobulated	0.80	0.90	0.95
Obscured	0.71	0.84	0.92
Ill-Defined	0.77	0.88	0.94
Spiculated	0.85	0.93	0.96

Table 8: Predictive probabilities for Age = 70

	Round or Oval	Lobular	Irregular
Circumscribed	0.59	0.76	0.87
Microlobulated	0.87	0.94	0.97
Obscured	0.81	0.90	0.95
Ill-Defined	0.85	0.93	0.96
Spiculated	0.91	0.96	0.98

Table 9: Predictive probabilities for Age = 80

Tobin South

I worked with both Lily and James on Section 2. We had an initial meeting to plan section 2 and divide up the work. We also had an additional meeting to clarify some details later in the project as well as spending time casually working on the project with my teammate. I contributed mainly to the data cleaning and data visualization sections of the logistic model and acted as the GitHub technical support for the team.