Statistical Modelling III Project

Shingyan Kwong, Mohammad Rezai, Xiaomeng Gu, Wenjun Liu and Oliver Lountain

June 3, 2018

Part A.

1.

Here, we produce pairwise scatterplots for height, weight and length to see the individual relationships between these varaibles (Figure 1):

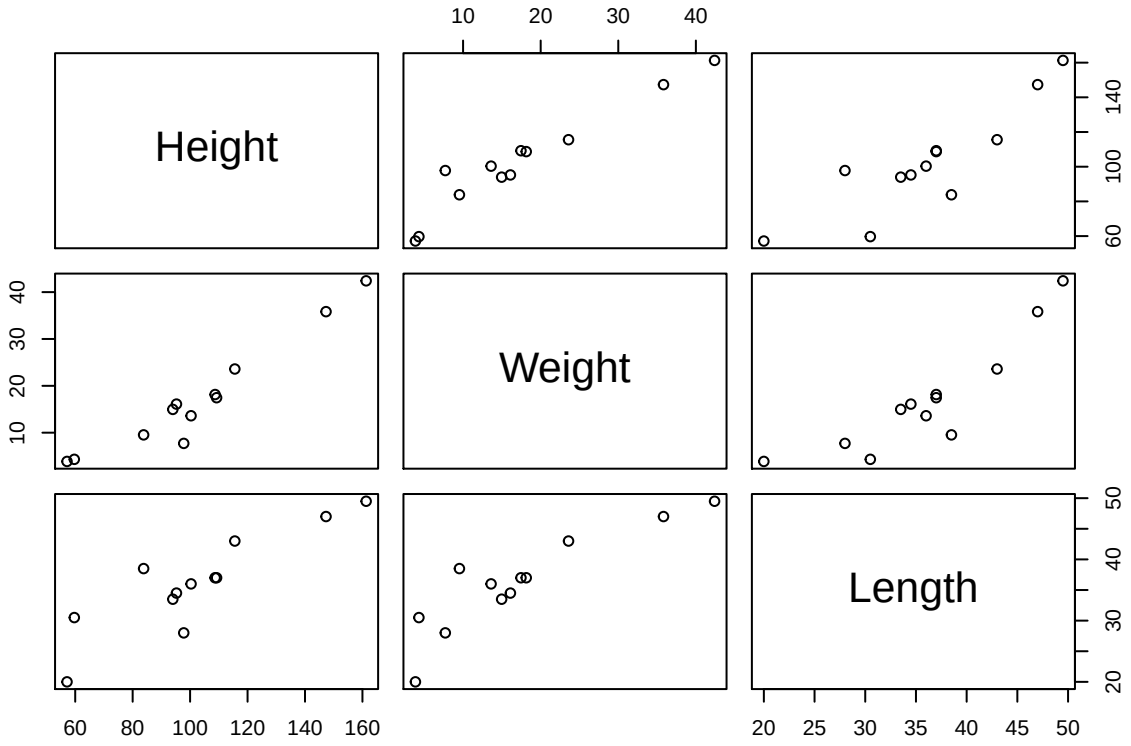

Figure 1: Pairwise Scatterplots for Height, Weight and Length

In figure 1 we see that there is evidence of a strong, positive linear relationship between length and the two predictor variables, height and weight. The associated correlation coefficients are 0.881 and 0.894 respectively. There is also a strong, positive, linear relationship between height and weight. This suggests that the two predictors may depend on one another.

2.

Here we fit three linear models: lm1 which fits Length against Height and Weight; lm2 which fits Length against Height; and lm3 which fits Length against Weight.

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

3.

The model assumptions which may be checked via diagnostic plots are as follows.

Linearity: Check the residuals vs. fitted and the residuals vs. predictor plots. The assumption of linearity is reasonable if random scatter above and below the 0 line is observed.

Constant Variance: Check scale-location plot. The assumption of constant variance is reasonable if constant variance of residuals is observed across the scale location plot.

Normality: Check normal quantile-quantile plot. The assumption of normality is reasonable if most points between -2 and 2 are on or close to the diagonal line.

Leverage: Check the residuals vs. leverage plot. The assumption that there are no points of high leverage is reasonable if there are no points beyond Cook's distance, that is, no highly influential points in the dataset.

As well as these assumptions, it is important to check that the data are independent, but this cannot be done with diagnostic plots and is based entirely on the way in which the data was collected.

4.

Here we check the regression assumptions for each of the three models.

Full model (lm1)

The assumption checking plots for the full model are given in figures 2 and 3:

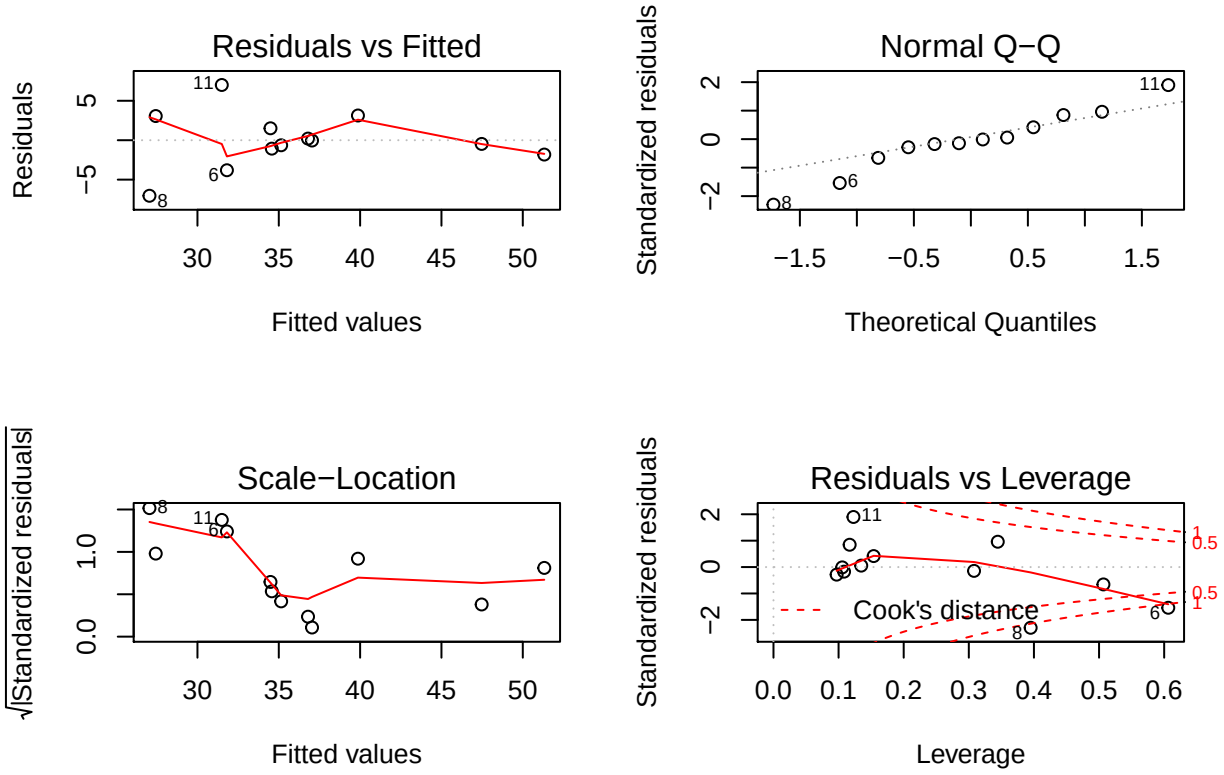

Figure 2: Residuals plots for Full Model

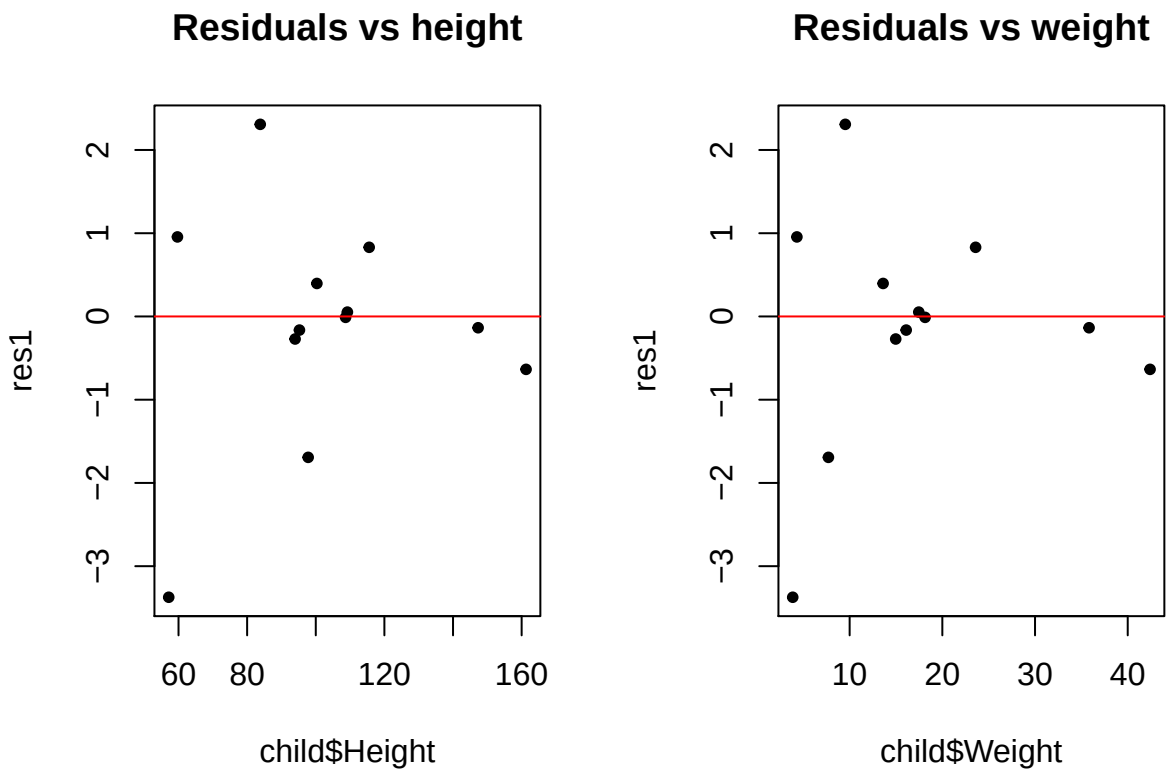

Figure 3: Residuals plots for Height and Weight Individually (Full Model)

From figures 2 and 3 we can make the following assessments:

Linearity: Given the small number of data points available, roughly random scatter is observed in the residual vs. fitted and residual vs predictor plots. There are a couple of high residual points but it is not too bad. Linearity is reasonable.

Constant variance: Scale location plots appear to show heteroscedasticity. Constant variance is not reasonable.

Normality: There is some minor departure from normality in the beginning and the tails of the standardized residuals. Overall the points are fairly close to the diagonal line. Normality is reasonable.

Leverage: There are 2 data points beyond Cook's Distance. These high leverage points are having a disproportionate effect on the model.

Height Model (lm2)

The assumption checking plots for the height model are given in figures 2 and 3:

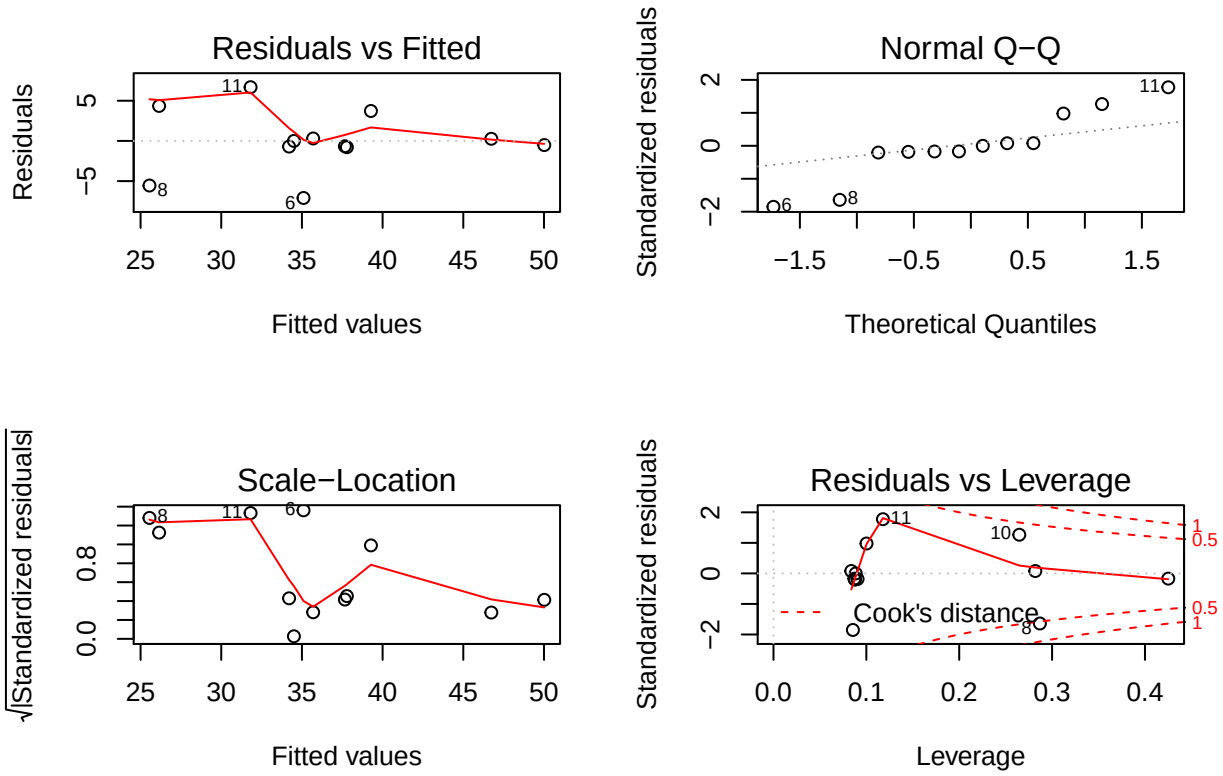

Figure 4: Residuals plots for Height Model

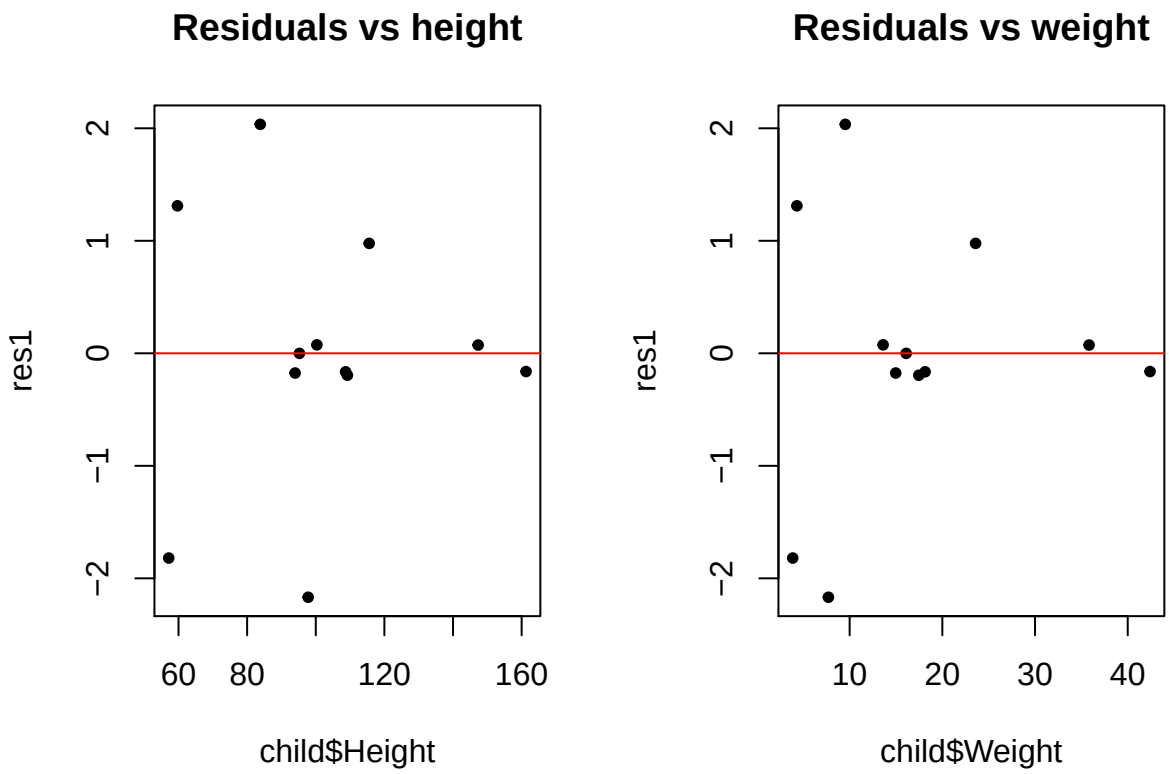

Figure 5: Residuals plots for Height and Weight Individually (Height Model)

From figures 4 and 5 we can make the following assessments:

Linearity: Non-random scatter observed in residual vs. fitted and residual vs. predictor plots. Linearity is not reasonable.

Constant Variance: Variance appears to increase for the middle fitted values and then decrease again. Constant variance is not reasonable.

Normality: There are several points deviating from the diagonal line on the normal quantile-quantile plot. Normality is not reasonable.

Leverage: There is one point with high leverage.

Weight Model (lm3)

The assumption checking plots for the weight model are given in figures 2 and 3:

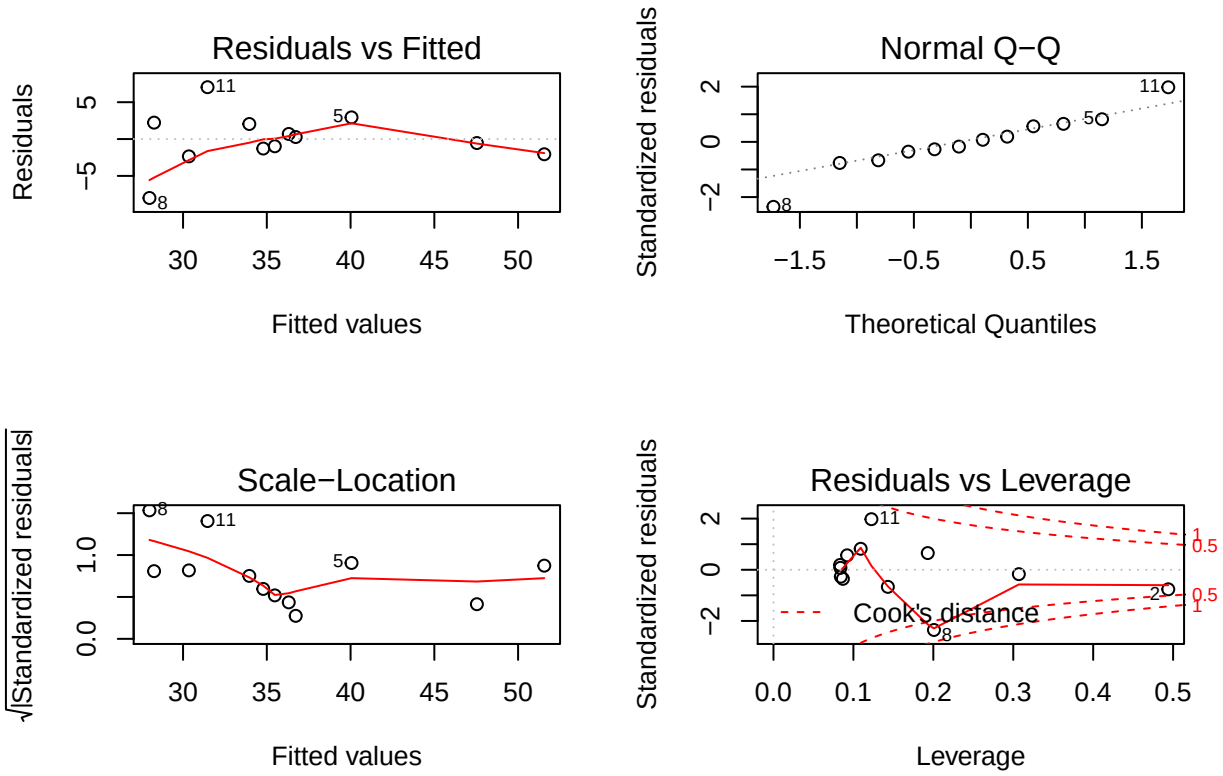

Figure 6: Residuals plots for Weight Model

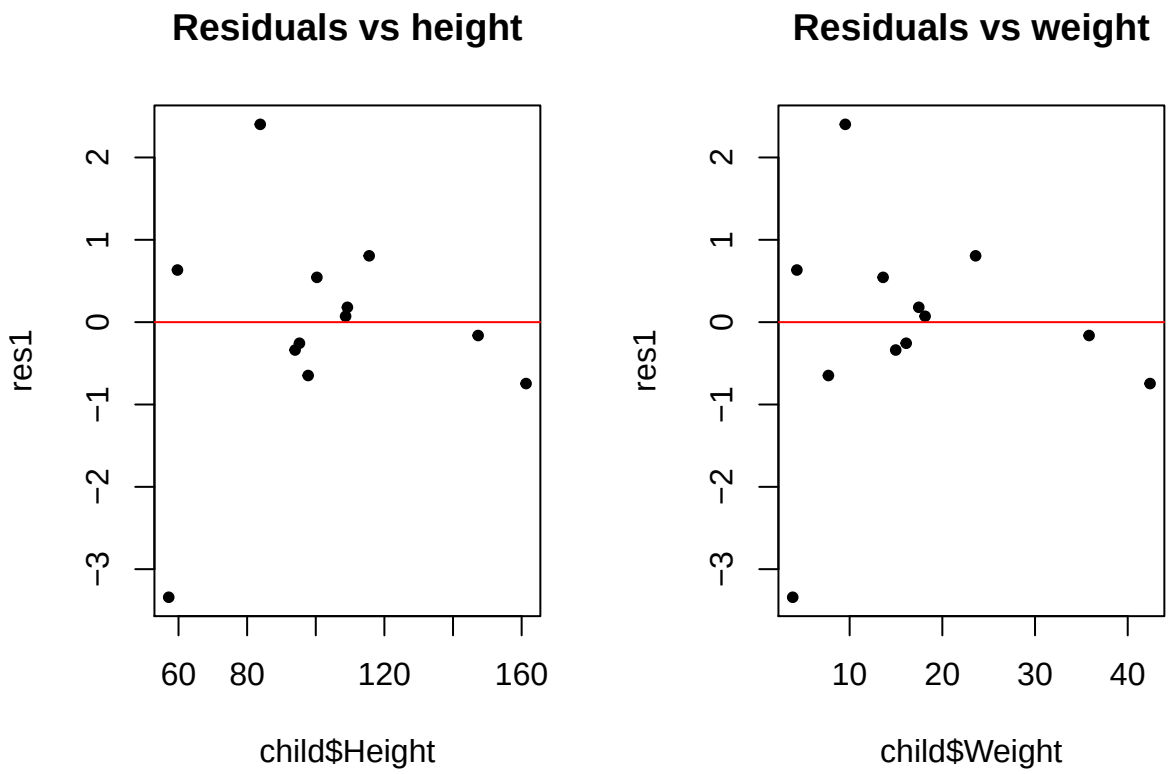

Figure 7: Residuals plots for Height and Weight Individually (Weight Model)

From figures 6 and 7 we can make the following assessments:

Linearity: Rough random scatter is observed in residual vs. fitted and residual vs. predictor plots. The fitted plot shows some evidence of curvature but overall it is acceptable. Linearity is reasonable.

Constant Variance: Variance is roughly constant across the scale-location plot. Constant variance is reasonable.

Normality: Most points are close to the diagonal line except 2. Normality is reasonable.

Leverage: There is one data point with high leverage.

5.

Here we make a comparison of the three models.

We first obtain summaries for each of the three models:

Full Model:

```
##
## Call:
## lm(formula = Length ~ Height + Weight, data = child)
##
## 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
```

Height Model:

```
##
## Call:
## lm(formula = Length ~ Height, data = child)
##
## 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
```

Weight Model:

```
##
## Call:
## lm(formula = Length ~ Weight, data = child)
##
## 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
```

(a)

In the full model, neither of the predictor variables are statistically significant (at the 5% significance level), and the numerical values of the two coefficients are both smaller than those of the single predictor models.

(b)

Full model:

Holding height constant, the full model predicts that an increase of of 1kg will on average increase the length of the catheter by 0.42081cm.

Weight only model:

Without regard for height, this model predicts that an increase of 1kg will on average increase the catheter length by 0.61136cm.

6

(a)

First, we construct the model matrices for the height only and weight only models.

Model Matrix for lm2:

```
##    (Intercept) Height
## 1            1 108.70
## 2            1 161.29
## 3            1  95.25
## 4            1 100.33
## 5            1 115.57
## 6            1  97.79
## 7            1 109.22
## 8            1  57.15
## 9            1  93.98
## 10           1  59.69
## 11           1  83.82
## 12           1 147.32
## attr(,"assign")
## [1] 0 1
```

Model matrix for lm3:

```
##    (Intercept) Weight
## 1            1  18.14
## 2            1  42.41
## 3            1  16.10
## 4            1  13.61
## 5            1  23.59
## 6            1   7.71
## 7            1  17.46
## 8            1   3.86
## 9            1  14.97
## 10           1   4.31
## 11           1   9.53
## 12           1  35.83
## attr(,"assign")
## [1] 0 1
```

Here, we define $\mathbf{1} := (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$, the vector of intercepts for both models. We also denote the vector of height values by $\mathbf{x}_1$ and the vector of weight values by $\mathbf{x}_2$. Then we find that:

$\mathcal{L}_1$ is the space spanned by the columns of M2, that is $\mathcal{L}_1 = span\{\mathbf{1}, \mathbf{x}_1\}$

$\mathcal{L}_2$ is the space spanned by the columns of M3, that is $\mathcal{L}_2 = span\{\mathbf{1}, \mathbf{x}_2\}$

Then, the intersection of the two subspaces is the intercept column, that is $\mathcal{L}_1 \cap \mathcal{L}_2 = span\{\mathbf{1}\}$.

(b)

We note that $(\mathcal{L}_1 \cap \mathcal{L}_2)^\perp$ is the subspace of all vectors orthogonal to $\mathbf{1}$. Then, in order to find the intersections of $\mathcal{L}_1$ and $\mathcal{L}_2$ with $(\mathcal{L}_1 \cap \mathcal{L}_2)^\perp$, we first find orthonormal bases for $\mathcal{L}_1$ and $\mathcal{L}_2$.

We achive this by applying the Gram-Schmidt process. First, we define the basis vectors for both subspaces, and a function norm_vec to find the norm of a vector:

```
one <- c(1,1,1,1,1,1,1,1,1,1,1,1) # intercept vector
x1 <- M2[,2] # vector of height values
x2 <- M3[,2] # vector of weight values
norm_vec <- function(x) sqrt(as.numeric(t(x) %*% x))
```

Next, we find an orthonormal basis for $\mathcal{L}_1$:

```
v1 <- one / norm_vec(one)
v2_ <- x1 - as.numeric((t(x1) %*% v1)) * v1
v2 <- v2_ / norm_vec(v2_)
```

Now

$$\mathbf{v}_1 = \frac{1}{\sqrt{12}}(1,1,1,1,1,1,1,1,1,1,1,1);$$

$$\mathbf{v}_2 = (0.0616, 0.5846, -0.0722, -0.0217, 0.1299, -0.0469, 0.0667, -0.4511, -0.0848, -0.4259, -0.1859, 0.4457)$$

Then $\mathcal{L}_1 = span\{\mathbf{v}1, \mathbf{v}2\}$.

Now, an orthonormal basis for $\mathcal{L}_2$:

```
w1 <- one / norm_vec(one)
w2_ <- x2 - as.numeric((t(x2) %*% w1)) * w1
w2 <- w2_ / norm_vec(w2_)
```

Now

$$\mathbf{w}_1 = \frac{1}{\sqrt{12}}(1,1,1,1,1,1,1,1,1,1,1,1);$$

$$\mathbf{w}_2 = (0.0216, 0.6408, -0.0304, -0.0940, 0.1606, -0.2445, 0.0043, -0.3427, -0.0593, -0.3312, -0.1981, 0.4729)$$

Then $\mathcal{L}_2 = span\{\mathbf{w}1, \mathbf{w}2\}$.

Now, we have that $\mathbf{v}_1 = \mathbf{w}_1$ is parallel to $\mathbf{1}$, and that $\mathbf{v}_2$ and $\mathbf{w}_2$ are orthogonal to $\mathbf{1}$, that is, $\mathbf{v}_2, \mathbf{w}_2 \in (\mathcal{L}_1 \cap \mathcal{L}_2)^\perp$.

As a result, we find that:

$$\mathcal{L}_1 \cap (\mathcal{L}_1 \cap \mathcal{L}_2)^\perp = span\{\mathbf{v}_2\};$$

$$\mathcal{L}_2 \cap (\mathcal{L}_1 \cap \mathcal{L}_2)^\perp = span\{\mathbf{w}_2\}.$$

(c)

Given that both $\mathcal{L}_1 \cap (\mathcal{L}_1 \cap \mathcal{L}_2)^\perp$ and $\mathcal{L}_2 \cap (\mathcal{L}_1 \cap \mathcal{L}_2)^\perp$ are one-dimensional subspace, we can compute the angle between them using the relation:

$$\cos\theta = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{u}||\ ||\mathbf{v}||}$$

Where $\mathbf{u}$ and $\mathbf{v}$ are two vectors and θ is the angle between them.

We compute the angle between the two spaces in **(b)** as follows:

```
inner <- t(v2) %*% (w2) #Defining the inner product of v2 and w2
norm_v2 <- norm_vec(v2)
norm_w2 <- norm_vec(w2)
theta <- acos(inner/(norm_v2 * norm_w2)) #Finding the angle between v2 and w2
```

Now we have that the angle between $\mathbf{v}_2$ and$\mathbf{w}_2$ is given by:

```
##           [,1]
## [1,] 0.2799352
```

The angle is not π indicating that the two spaces are not orthogonal. This suggests that height and weight are not independent. In fact they are fairly correlated.

7.

In this section we decide upon a final model for the data. We note that since the subspaces $\mathcal{L}_1 \cap (\mathcal{L}_1 \cap \mathcal{L}_2)^{\perp}$ and $\mathcal{L}_2 \cap (\mathcal{L}_1 \cap \mathcal{L}_2)^{\perp}$ are not orthogonal, the weight and height predictors are not independent; more specifially, they have a significant correlation of 0.961. As such, the full model in which Length was fitted against both Height and Weight should not be used. Now, what remains is to decide between the individual height and weight models. In figure 4, we see that the assumptions of linearity, constant variance and normality fail in the height model, and in figure 6 it is clear that these assumptions are much more reasonanble in the weight model. As a result, we can conclude that the most appropriate model for the data is lm3, the model in which engths is fit against height alone.

Part B

Introduction

In this section we obtain a predictive model for mammographic mass severity, a measure of the status of mammographic mass lesions, on a scale from 0 to 1, where 0 is assigned to a benign tumor, and 1 is assigned to a malignant tumor. Interest in this analysis arises from there being a low predicitve value of breast biopsy from mammograms. This low predictive value has been found to lead to approximately 70% of unnessessary biopsies of benign tumors. Analysis is performed on the dataset "mammo", containing the true status of 961 mammographic mass lesions, with the response variable severity as described. Four response variables are considered:

Age - the patient's age in years;

Shape - a factor variable with four levels: 1 for round, 2 for oval, 3 for lobular, and 4 for irregular;

Margin - a factor varaible with five levels: 1 for circumscribed, 2 for microlobulated, 3 for obscured, 4 for ill-defined, and 5 for spiculated;

Density - a factor with four levels: 1 for high, 2 for iso, 3 for low, and 4 for fat-containing.

The aim of this study is to produce the best predictive model for mammographic mass severity. The model must be parsimonious whilst having minimun residual error. Predicted probabilities of severity may also be generated from the model. This information may be used by clinicians to determine whether a biopsy is appropriate for the patient.

Data Entry and Cleaning

First, we enter the data and define any values which are assigned question marks to be missing values.

We then note that BI.RADS is not a predictor variable, and remove it from our analysis.

We generate a correlation matrix and a pairwise scatterplot to observe any relationships between the predictor variables and the response. It was necessary to omit observations with NA values to produce these plots. Shape and Margin were included as increases in their values are associated with a greater risk of cancer. Density was omitted as the value is not associated with a greater risk. Thus correlation between Density and Severity would not be statistically meaningful.

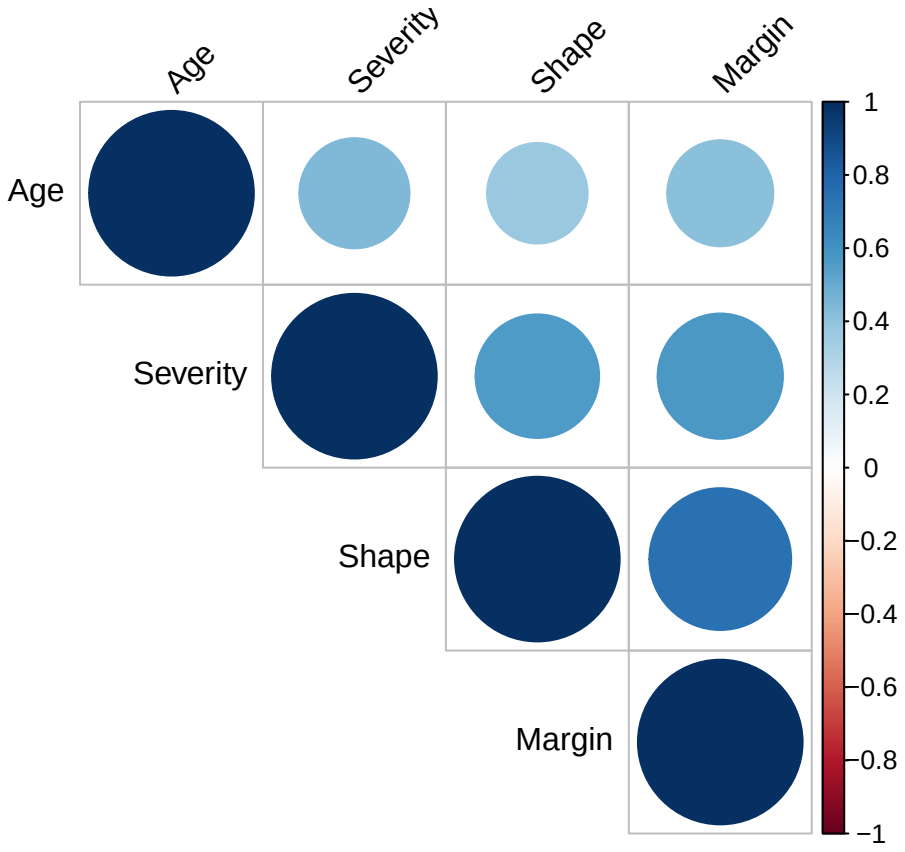

Figure 8: Correlation matrix depicting the relationship between Margin, Shape, Severity and Age.

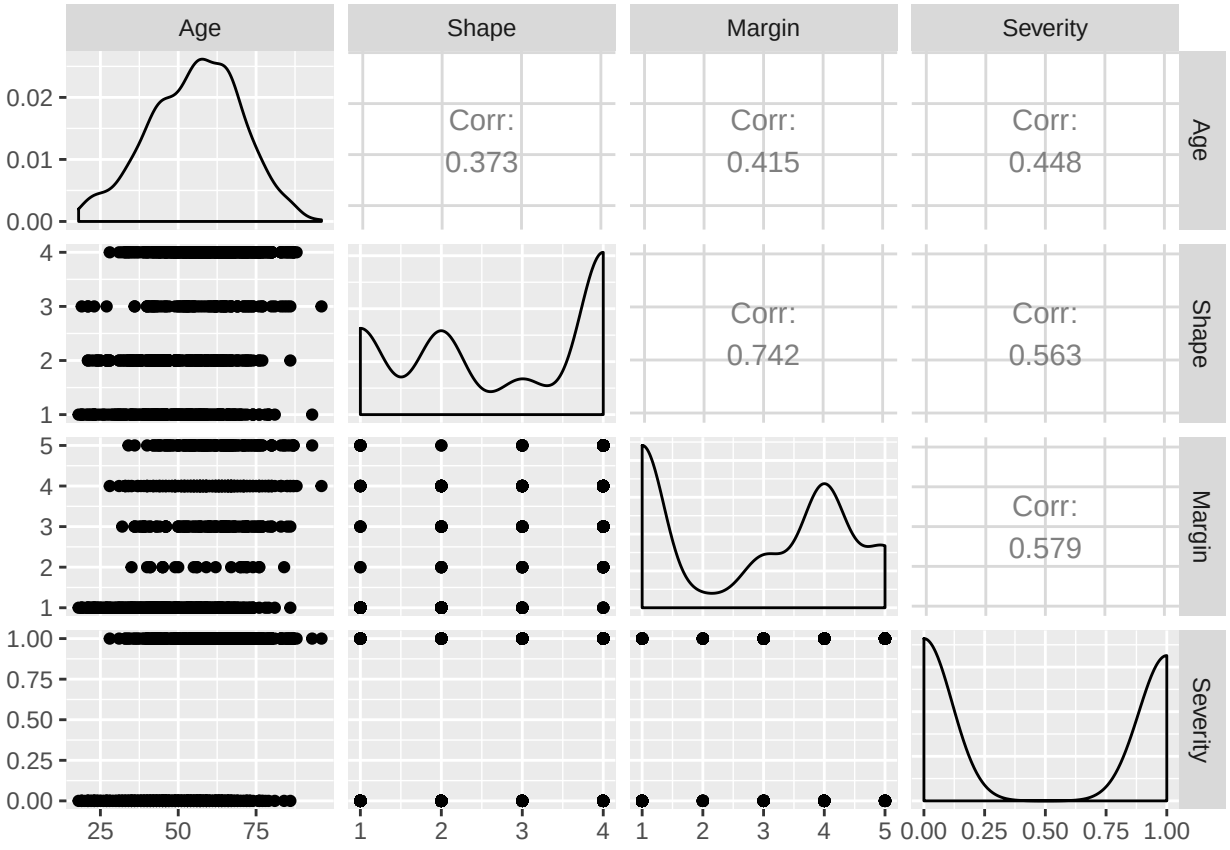

Figure 9: Comparison Matrix depicting the relationship between Margin, Shape, Severity and Age.

From figures 8 and 9 we can make the following observations:

Mild positive correlation is observed between Age and the other variables. There correlations are 0.364 and 0.411 for Shape and Age and Margin and Age respectively.

Moderate positive correlation is observed between Severity and the other 3 predictors.

Strong positive correlation between Shape and Margin with a correlation of 0.742.

We can now check the variable types for the data:

```
## 'data.frame':    961 obs. of  5 variables:
##  $ Age     : int  67 43 58 28 74 65 70 42 57 60 ...
##  $ Shape   : int  3 1 4 1 1 1 NA 1 1 NA ...
##  $ Margin  : int  5 1 5 1 5 NA NA NA 5 5 ...
##  $ Density : int  3 NA 3 3 NA 3 3 3 3 1 ...
##  $ Severity: int  1 1 1 0 1 0 0 0 1 1 ...
```

We note that Shape, Margin, Density and Severity should all be factor variables, and as such convert them.

We now see that all of the data types are correct:

```
## 'data.frame':    961 obs. of  5 variables:
##  $ Age     : int  67 43 58 28 74 65 70 42 57 60 ...
##  $ Shape   : Factor w/ 4 levels "1","2","3","4": 3 1 4 1 1 1 NA 1 1 NA ...
##  $ Margin  : Factor w/ 5 levels "1","2","3","4",..: 5 1 5 1 5 NA NA NA 5 5 ...
##  $ Density : Factor w/ 4 levels "1","2","3","4": 3 NA 3 3 NA 3 3 3 3 1 ...
##  $ Severity: Factor w/ 2 levels "0","1": 2 2 2 1 2 1 1 1 2 2 ...
```

Data Visualisations and Data Summaries

To visualise the data, we first produce summary statisitics for the dataset as a whole, and for each individual variable:

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

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

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

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

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

We can then generate boxplots of Age against the other four predictors to see how the different levels of the factor variables are affected by a patients age.

First, for Severity (Figure 10):

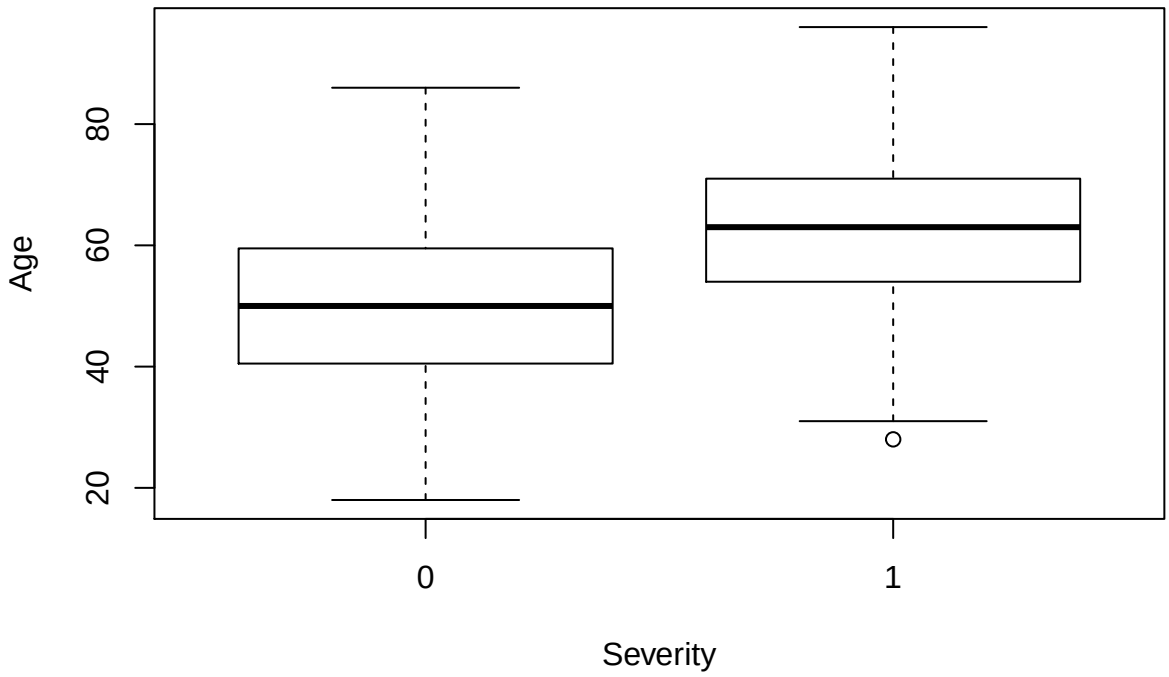

Figure 10: Boxplot of Age Against Severity

In figure 10 we note that in general, a severity value of 1 is correlated with higher ages than that of 0. In other words, patients with malignant tumors tend to be older than patients with benign tumors Also, the interquartile range for severity = 0 tends to be larger than severity = 1. We also note that there is an outlier at a low age (approximately 30) when severity = 1.

Then, for Shape (Figure 11):

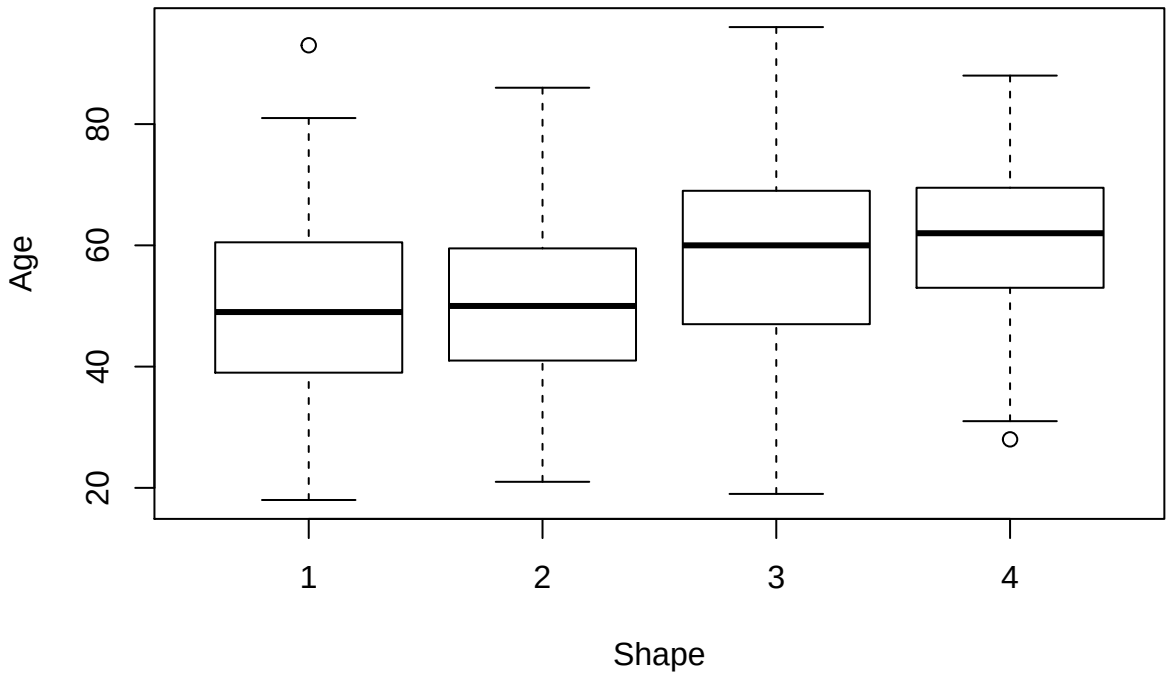

Figure 11: Boxplot of Age Against Shape

In figure 11 we see that the median ages for patient with round and oval shapes are approximately equal. Also, the boxplot for patients with lobular shape is left skewed. There is a outlier in the boxplot with round shape, and one in the boxplot with irregular shape.

Next, for Margin (Figure 12):

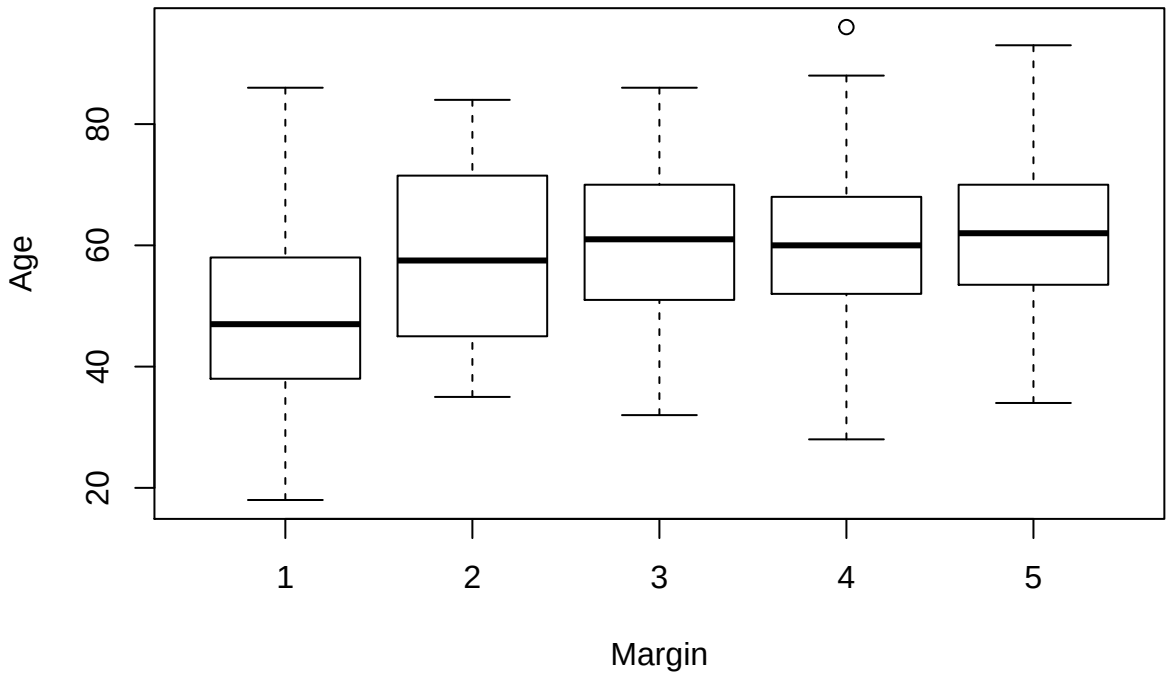

Figure 12: Boxplot of Age Against Margin

In figure 12 it can be seen that the median age for patients who have obscured, ill-defined and spiculated margins is approximately equal. Moreover, the interquartile range for microlobulated margin is larger than for others. There is a outlier in ill-defined margin.

Finally, for Density (Figure 13):

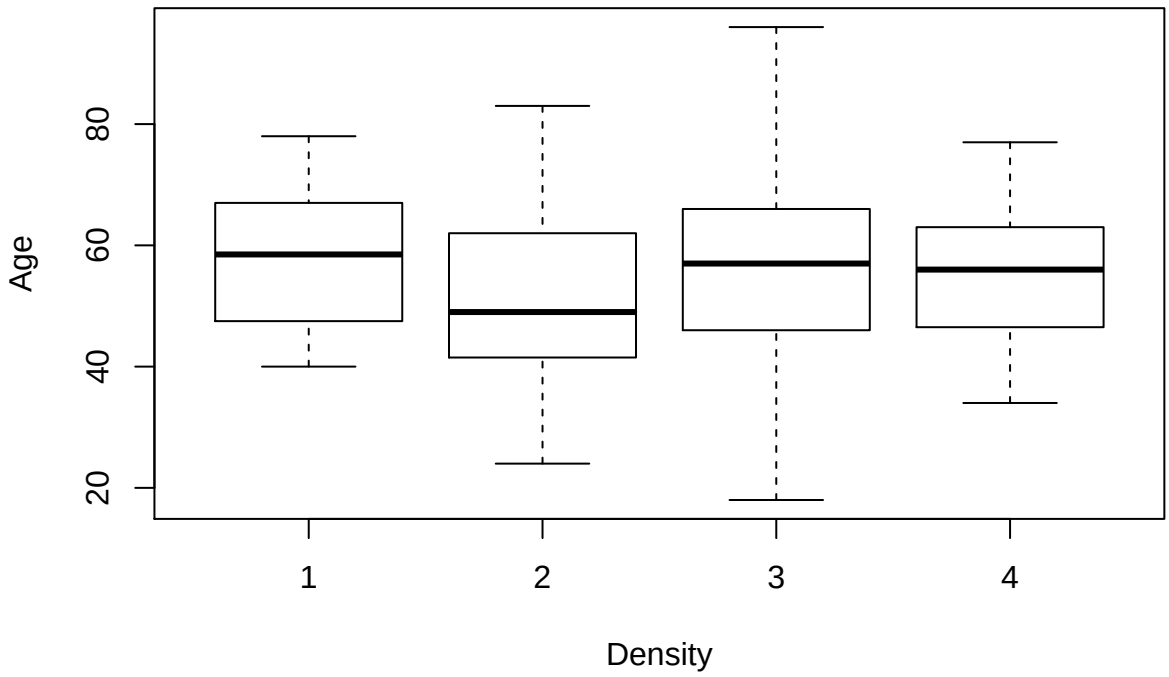

Figure 13: Boxplot of Age Against Density

In figure 13 we note that the boxplots for high and low density are left-skewed, and the boxplot for iso density is right-skewed. Although the median age for low and fat-containing density is similar, the age of low density group is more variable than the other. Moreover, the range from the lowest age to the highest age for low density is larger than others.

We can also create bar graphs to visualise how the number of patients with different values for each of the three factor predictor variables is distributed with respect to their Severity value.

First, we consider Shape (Figure 14):

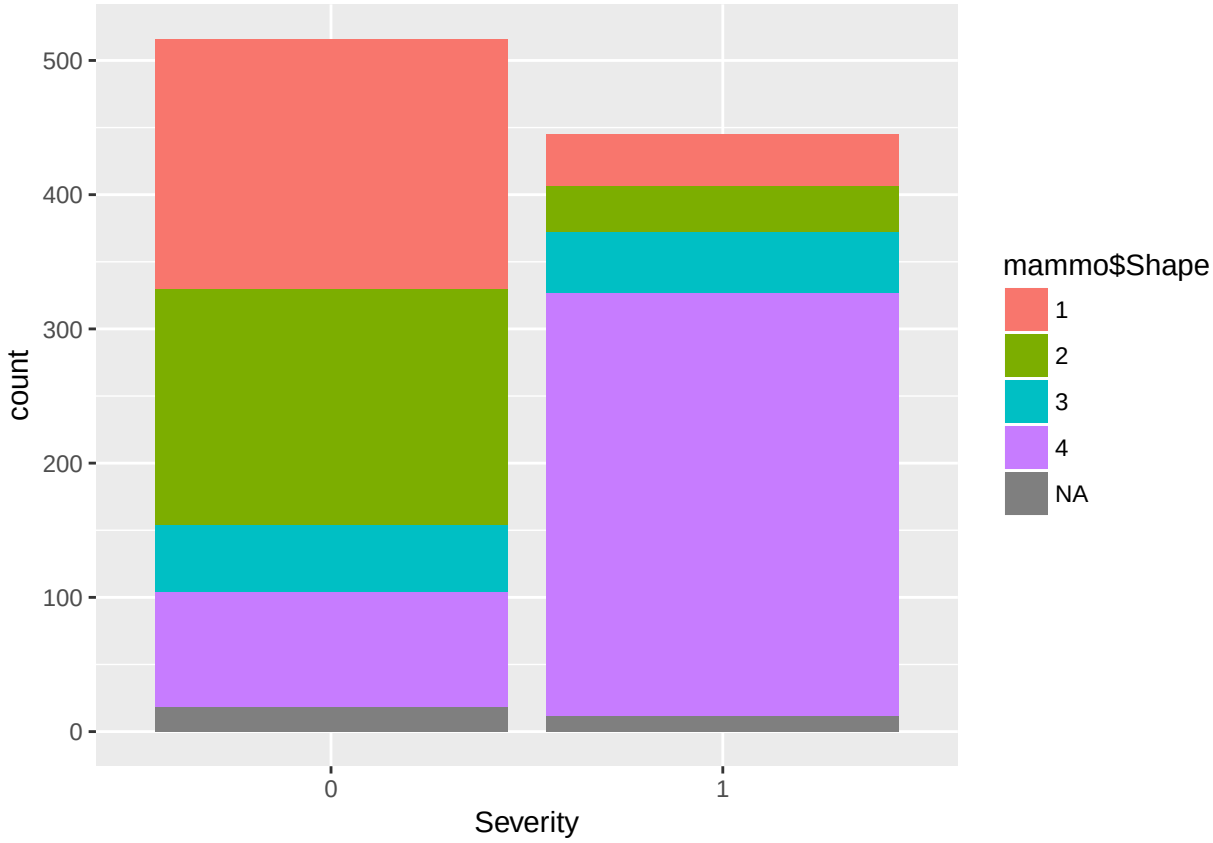

Figure 14: Bar Graph of Severity Against Shape

In figure 14 it can be seen that patients who have benign tumors tend to have more round shapes and oval shapes than patients who have malignant ones.Furthermore, patients who have benign tumors tend to have more regular shape than patients who have malignant tumors. There is an approximately equal number of patients that have lobular shape between patients who have malignant tumors and those who do not.

Next, we consider Margin (Figure 15):

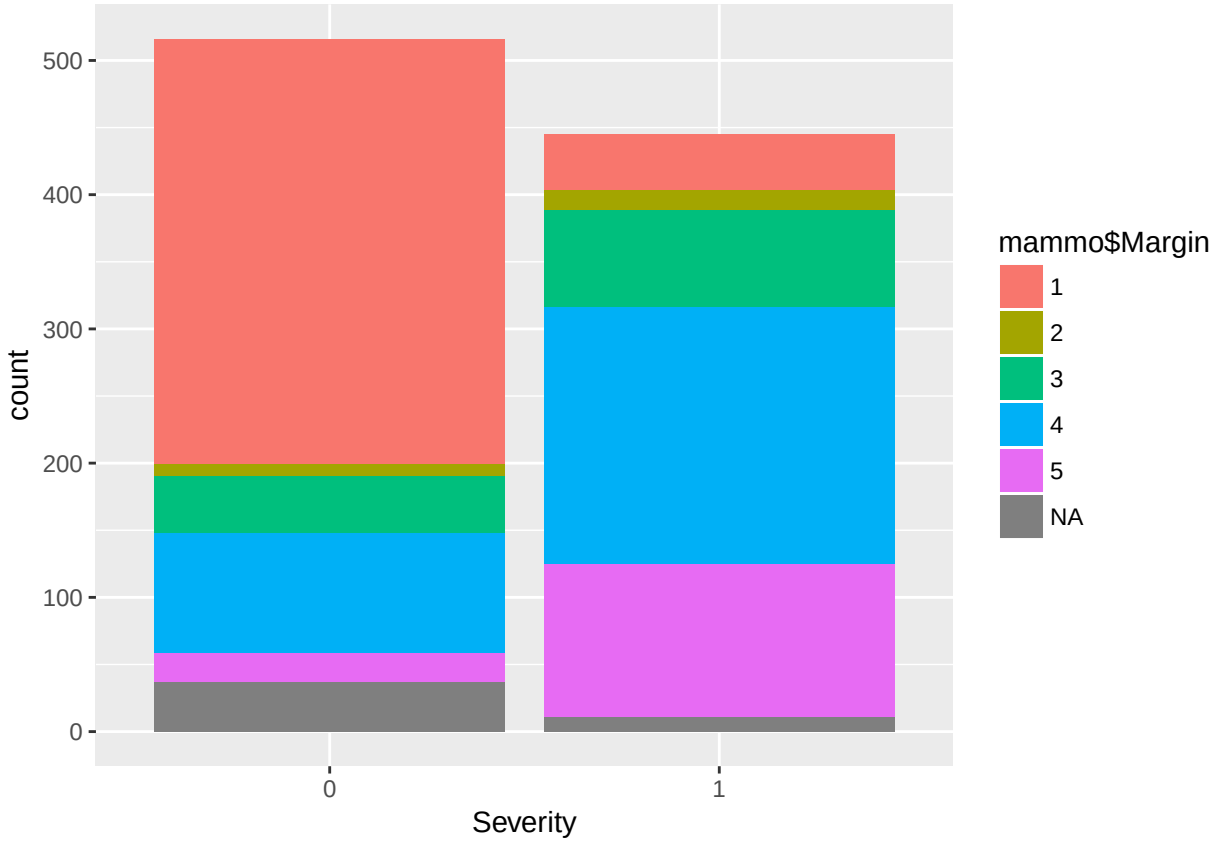

Figure 15: Bar Graph of Severity Against Margin

We see in figure 15 that patients who have benign tumors tend to have more circumscribed margin than those whose tumors are malignant. Moreover, patients who have non-malignant tumors tend to have less microlobulated, obscured, ill-defined, and spiculated margin than patients who have malignant ones.

Finally, we consider Density (Figure 16):

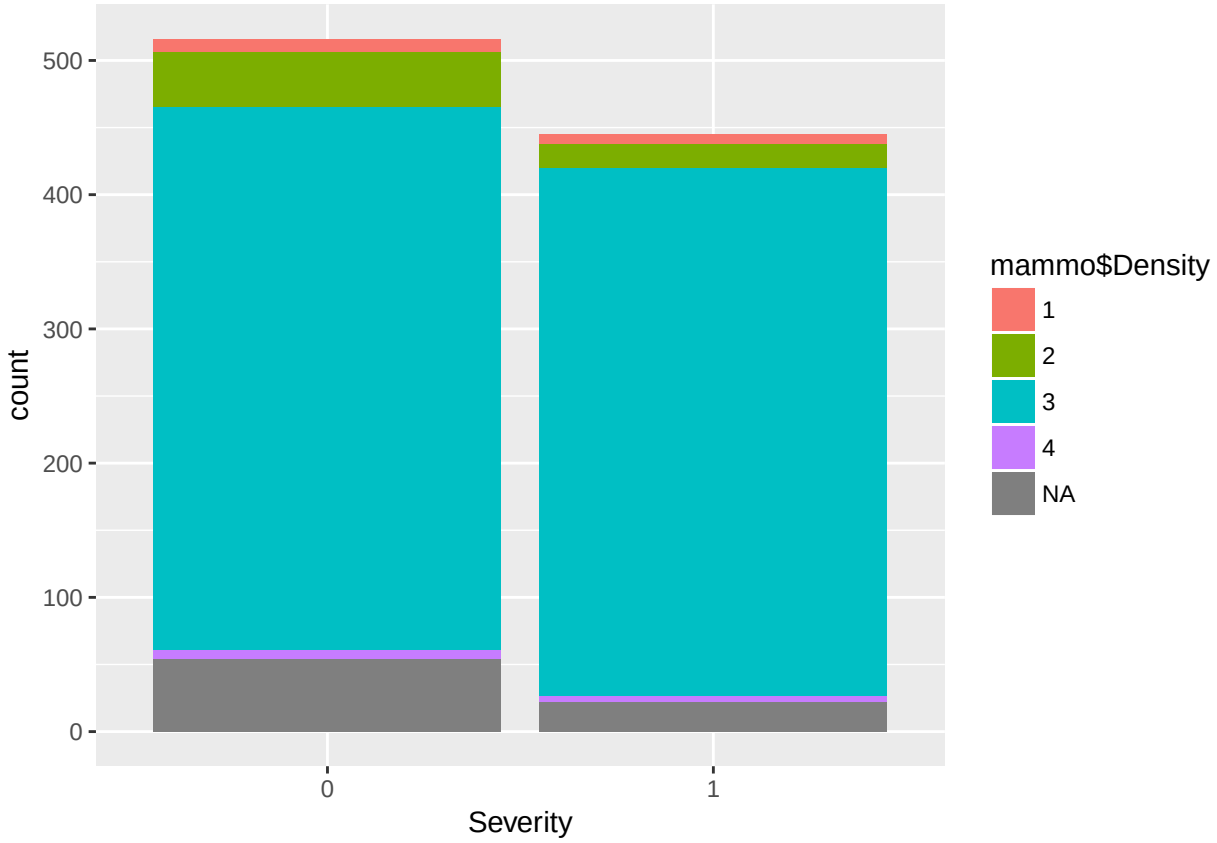

Figure 16: Bar Graph of Severity Against Density

In figure 16 we see that patients who have non-malignant tumors tend to have more iso and low density. In addition, these two groups (severity = 1 and severity = 0) tend to have equal number of patients with high and fat-containing density.

Model Fitting and Model Selection

We now fit a logistic linear model (full.glm) to the data, with Severity as the response variable, and Age, Shape, Margin and Density as the predictor variables with interaction terms up to second order. To achieve a parsimony, iInsignificant terms (p-value>0.05) were removed by the backwards selection algorithmn. Our analysis showed that all the interactions terms and Density are not statistically significant at the 0.05. Details of each step of the model selection process are available in the Part B Appendix under "Model Fitting and Model Selection."

Justification of the final model

The following is the proposed final model.

```
##          term    estimate   std.error  statistic      p.value
## 1 (Intercept) -4.71954438 0.465771163 -10.132753 3.953566e-24
## 2         Age  0.05387869 0.007498812   7.184963 6.722520e-13
## 3      Shape2 -0.44784427 0.306327242  -1.461980 1.437467e-01
## 4      Shape3  0.49925125 0.364446283   1.369890 1.707213e-01
## 5      Shape4  1.24283749 0.324255926   3.832891 1.266463e-04
## 6     Margin2  1.58294301 0.539613781   2.933474 3.351917e-03
## 7     Margin3  1.26307280 0.342530673   3.687474 2.264916e-04
## 8     Margin4  1.54322611 0.294044551   5.248273 1.535316e-07
## 9     Margin5  2.03210483 0.362892142   5.599749 2.146627e-08
```

This model was obtained by first starting with a saturated model with all the two way interaction terms. Statistically insiginificant terms were removed via the backwards selection algorithmn at the 5% significance level. Hence the final model is the most parsimonious model with all the statistically significant predictors.

Interpretation of Parameters

Intercept: A woman of Age = 0 with a mammogram of Shape = 1 (round), Margin=1 (circumscribed) has log-odds = -4.7195 of having a malignant tumour.

Age: Holding all other variables constant, a one year increase in age increases the log-odds of having a malignant tumour by 0.05388.

Shape: Holding all other variables constant, having a lesion of Shape = 2 (oval) increases the log-odds of having a malignant tumour by -.4478 compared to Shape = 1 (round). For Shape = 3 (lobular), th increase is 0.4992 and for Shape = 4 (irregular), the increase is 1.2428, all relative to Shape = 1.

Margin: Holding all other variables constant, having a lesion of Margin = 2 (microlobulated) increases the log-odds of a malignant tumour by 1.5829 comapred to Margin = 1 (circumscribed). For Margin = 3 (obscured) the change in log-odds is 1.2631, Margin = 4 (ill-defined) 1.5432 and Margin = 5 (spiculated) 2.0321, all relative to Margin = 1.

Predicting Probabilities and Interpretation

In this section, we use our final model (final.glm) to predict the probability of a specific patient, that is, a patient with given values for each of the predictor varaibles. Given that the response variable is defined to be 0 for benign (not cancerous) and 1 for malignant (cancerous), the fitted values lie between 0 and 1 and hence predict the probability for a given patient to have a malignant tumor.

We first fit the probabilties of each datapoint in the dataset based on the final model:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01736 0.12041 0.51500 0.47238 0.80361 0.96242
```

We can produce a histogram to visualise the overall distribution of probabilities (Figure 17):

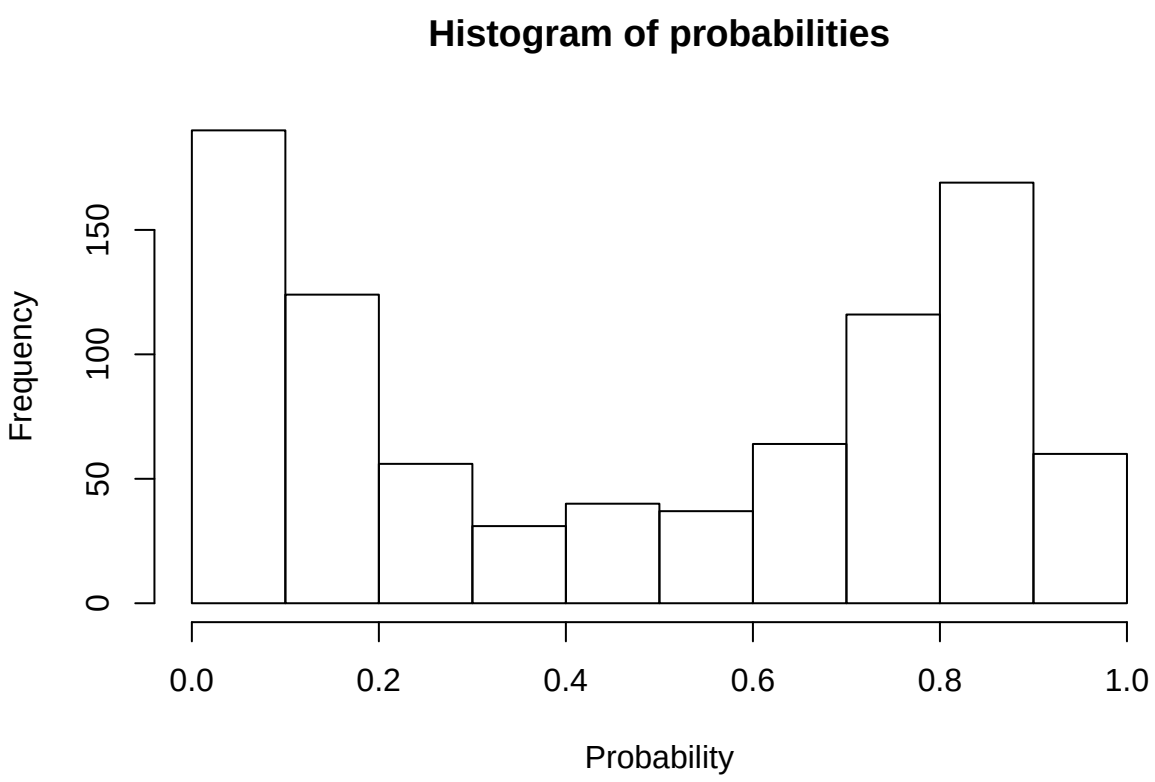

Figure 17: Histogram of Fitted Probabilities

Here we note that in general, it appears that most patients are either very likely, or very unlikely to have a malignant tumor. As a result, we might expect when predicting probabilities, that in most cases the predictions will be either very high or very low.

We can produce plots to visualise the probabilities for different levels of the predictor variables.

We first define a modified version of the mamo data, including only Age, Shape, Margin and Severity, and ignoring the missing values in order to be able to create valid plots:

We can now create plots of probabilities against Age, Shape and Margin (Figures 18, 19 and 20 respectively):

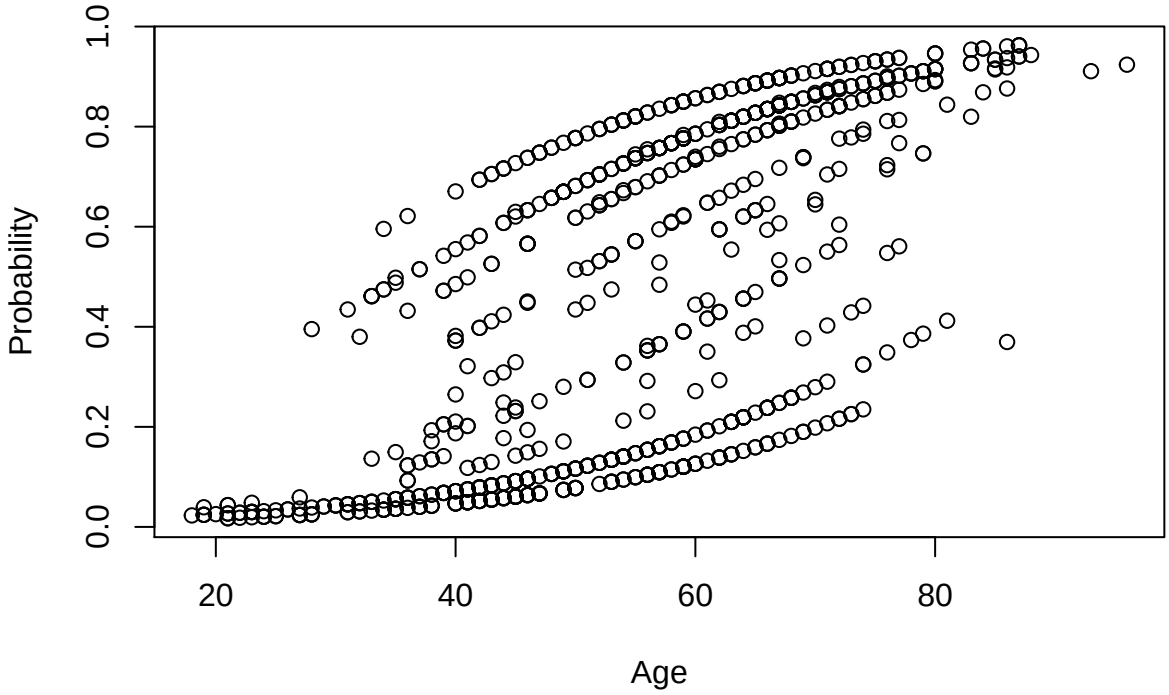

Figure 18: Probabilities against Age

Here we see that there appears to be a weak, positive relationship between age and the probability of having a malignant tumor, and it is difficult to say whether the relationship is linear or not.

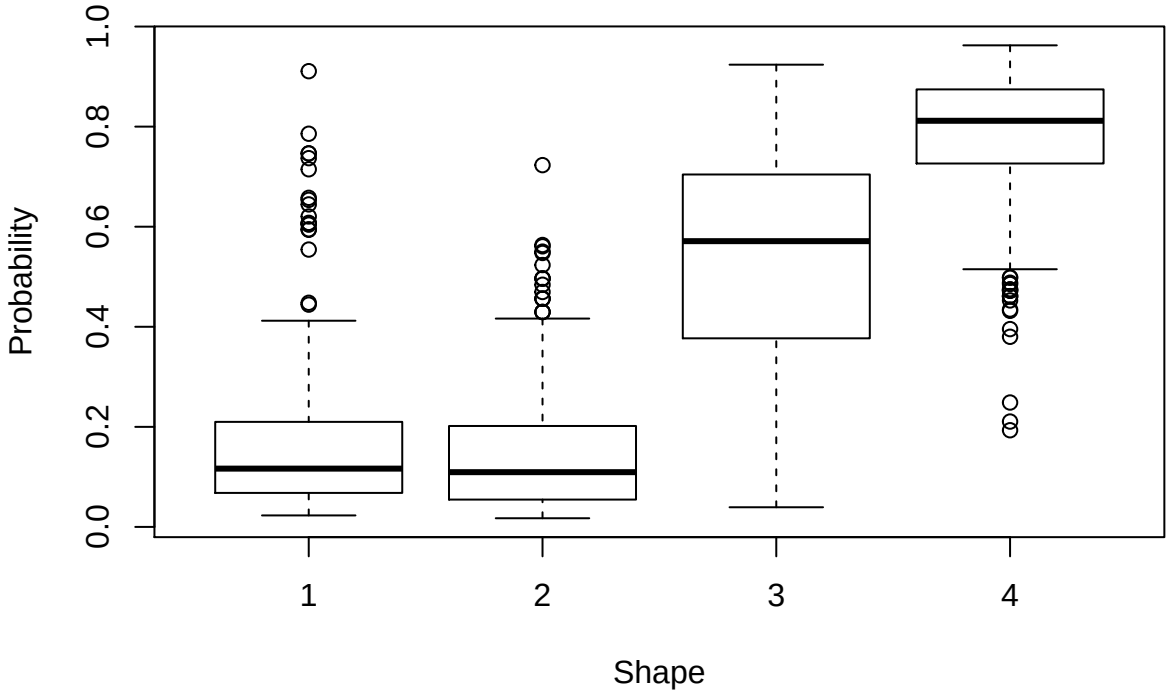

Figure 19: Probabilities against Shape

Here we see that as shape tends from the round, regular shape to a more irregular one, the predicted probabilities appear to increase in general. There appears to be a non-constant variance as the interquartile range of Shape = 3 is much larger than that of the other levels.

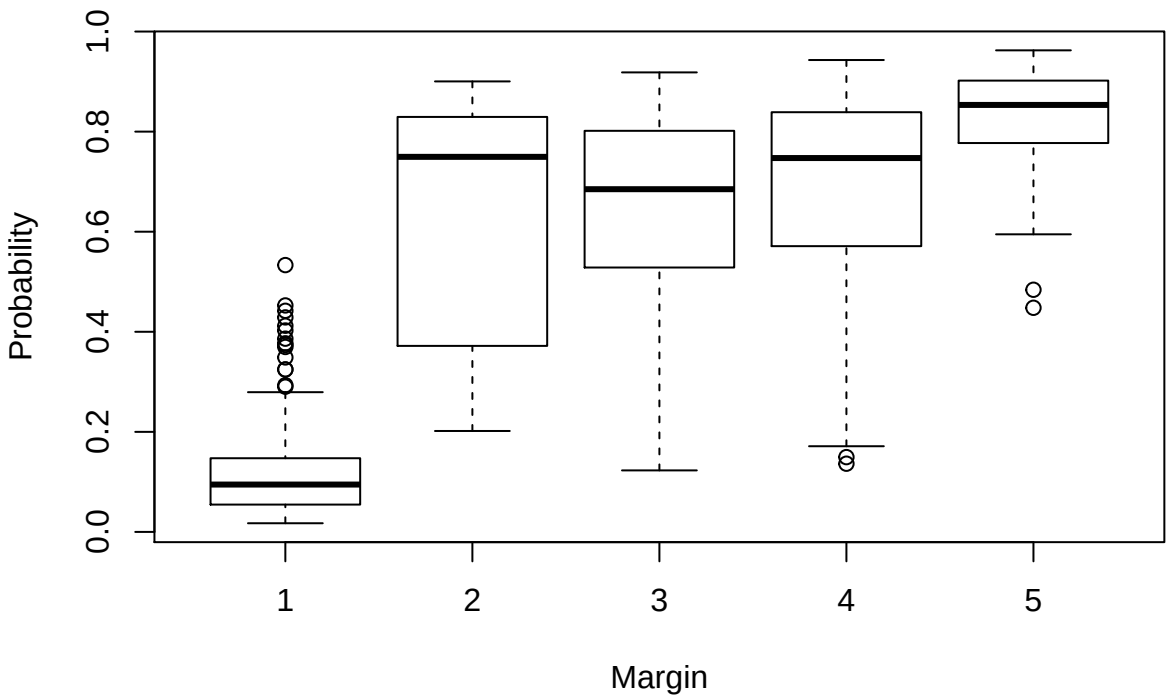

Figure 20: Probabilities against Margin

Here we see that as the margin tends from being well-defined to ill-defined, in general, the probability of the tumor being malignant seems to increase. Here we also note a non-constant variance, the interquartile range for the second level is very large, and that for the first level is very small.

Having observed these relationships, we can now predict the probability of the tumor being malignant for a few specific patients. We do so for a patient at an age of 40, with Shape = 1 (round) and Margin = 1 (circumscribed). This is a patient which we would expect to have a relatively low probability, as they are quite young, and their tumor is quite regular in shape and margin.

The predicted probability is given by:

```
##          1
## 0.07146523
```

This probability of 0.07147 alligns well with what we would expect. We can interpret this to mean that from a large group of patients, those with an age of 40, round tumors and circumscribed margins, approximately 7% would have malignant tumors.

On the other end of the spectrum, we can predict the probability for a patient with an age of 80, Shape = 4 (irregular) and Margin = 5 (speculated), that is an older patient with an irregular and very much ill-defined tumor.

The predicted probability is given by:

```
##         1
## 0.9461242
```

This probability of 0.9461 also alligns well with what we would expect. This means, that for a large group of patients, we would expect that for patients of age 80, with irregularly shaped tumors and a spiculated margins, that approximately 95% would have malignant tumors.

To find what we would expect to be a more intermediate probability, we can then predict the probability for a patient with an age of 60, Shape = 2 (oval), Margin = 3 (obscured):

```
##         1
## 0.3381399
```

This probability of 0.3381 also makes sense intuitively, as the patient's age and margin values were much more intermediate, and the shape of their tumor is closer to regular end of the spectrum than the irregular end. This means, that in a large group of patients, we would expect that for patients of age 60, with oval shaped tumors and obscured margins, that approximately 34% would have malignant tumors.

Part C

Our group comprised of five members and we commenced the project after our first face-to-face meeting on the fourth of May. During the first meeting, we decided that we would all attempt the project questions on our own first because we value the project as a good opportunity to learn and practice the course material. We planned that we would then come together with our answers to confirm that we were on the same page and identify any mistakes that we may have made.

Our initial plan was to meet one week after the first meeting, during which time we would try our best to finish part A individually. Then in the second week (13/05-19/05) we would assign one or two group members to write up part A, based on the answers we all agree upon in the second meeting. At this point we started to consider how to approach part B. Adopting a similar timeline as for part A, we decided to meet on the Friday of the second week to discuss the part B. In doing this, we would ensure that the bulk of the project was completed quite early, so that we would have ample time to format our write-up and ask questions about aspects about which we were unsure. Moreover, throughout the group project, if anyone had any problem, it was sent to the group chat and whoever was able to answer it provided some help.

However, when we started, we soon realised that some of the lecture content required for the project had not yet been covered, but we still tried to adhere to the planned timeline as much as possible. Part A was finished during the first week, other than question 6. In the second face-to-face meeting on 11th of May we had mutually agreed answers to all the other questions and a draft of the formal write-up of part A was allocated to Xiaomeng and Wenjun, who were to then send it to Anthony and Oliver to format it in Rmarkdown. The common questions we had were solved by going to both of the consulting times each week. Not all of our group members were available during those two time slots, but at least one member would go with questions prepared by the group and sent the feedback to the group chat afterwards. After the meeting on the first of June, all of us were confident with our answers for part B, and the write-up of part B was undertaken by Oliver and Anthony. During the entire period, we discussed the project in a Facebook group chat and in person after the lectures and tutorials. Wenjun was responsible for writing the reflection for part C, based on minutes kept in the meetings by Mohammad. We then met on the third of June to ensure everyone was happy to submit the project and to see if there were any possible further improvements.

Individual Contributions:

Shingyan: I have attended several group meetings to discuss the project. I assisted in writing up Part A in R Markdown. I worked on the code for Part B and wrote it up together with Oliver. Together we also edited the project script and made the final product presentable.

Mohammad: I tried to focus more on the codding side of the project and tried to make the best of all the consulting times. I was going through all the coding sections and was making sure that I understand the concept. There were times that others were unable to make to the consulting times due to their schedules to ask what they were struggling to understand. Therefore, I was asking their questions, if I was unable to provide them with the solution and was passing them on to others.

Xiaomeng: I attempted the majority of the project alone, and I did not go to the consulting due to a timetable clash. Therefore, I sought advice from the group members who went. Also, I did draft the part A with Wenjun Liu and I was also working on the data visualisation for part b.

Wenjun: Other than attempting the project on my own, participating in all the meetings and discussion, I drafted the write-up of part A with Xiaomeng and was in charge of writing Part C. I was delegated to go to several consultings and give the feedback to the group members could not make it too.

Oliver: As was the case for all group members, I attempted the majority of the project alone before we discussed answers as a group to determine what to include in the report. In the writing of the final report, my major roles were explaining the subspaces question, as well as collating and formatting the work into single polished document with Anthony.

Appendix

Here, we display the code used to generate ouput and plots following the order of the report. In part B, we display the model selection process in full.

Part A

1.

Reading data into R:

```
child<-read.table("Child_Height.txt",header=T)
```

Creating pairwise scatterplot:

```
pairs(~Height+Weight+Length,data=child)
```

2.

Fitting linear models:

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

4.

Residuals plots for lm1:

```
par(mfrow=c(2,2))
plot(lm1)
```

Individual residuals vs. predictors plots for lm1:

```
par(mfrow=c(1,2))
res1<-rstudent(lm1)
fit<-fitted(lm1)
plot(child$Height,res1,main="Residuals vs height",pch=20)
abline(0,0,col="red")
plot(child$Weight,res1,main="Residuals vs weight",pch=20)
abline(0,0,col="red")
```

Residuals plots for lm2:

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

Individual residuals vs. predictors plots for lm2:

```
par(mfrow=c(1,2))
res1<-rstudent(lm2)
fit<-fitted(lm2)
plot(child$Height,res1,main="Residuals vs height",pch=20)
```

```
abline(0,0,col="red")
plot(child$Weight,res1,main="Residuals vs weight",pch=20)
abline(0,0,col="red")
```

Residuals plots for lm3:

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

Individual residuals vs. predictors plots for lm3:

```
par(mfrow=c(1,2))
res1<-rstudent(lm3)
fit<-fitted(lm3)
plot(child$Height,res1,main="Residuals vs height",pch=20)
abline(0,0,col="red")
plot(child$Weight,res1,main="Residuals vs weight",pch=20)
abline(0,0,col="red")
```

5.

Summary for lm1:

```
summary(lm1)
```

Summary for lm2:

```
summary(lm2)
```

Summary for lm3:

```
summary(lm3)
```

6.

(a)

Model matrices for lm2 and lm3:

```
lm2 <- lm(Length ~ Height, data = child)
lm3 <- lm(Length ~ Weight, data = child)
M2 <- model.matrix(lm2)
M3 <-  model.matrix(lm3)
M2
M3
```

(b)

Defining vector of ones, $\mathbf{x}_1$ and $\mathbf{x}_2$ for the Gram Schmidt Process, as well as the function for the norm of a vector:

```
one <- c(1,1,1,1,1,1,1,1,1,1,1,1,1) # intercept vector
x1 <- M2[,2] # vector of height values
```

```
x2 <- M3[,2] # vector of weight values
norm_vec <- function(x) sqrt(as.numeric(t(x) %*% x))
```

Finding an orthonormal basis for $\mathcal{L}_1$:

```
v1 <- one / norm_vec(one)
v2_ <- x1 - as.numeric((t(x1) %*% v1)) * v1
v2 <- v2_ / norm_vec(v2_)
v2
```

Finding an orthonormal basis for $\mathcal{L}_2$:

```
w1 <- one / norm_vec(one)
w2_ <- x2 - as.numeric((t(x2) %*% w1)) * w1
w2 <- w2_ / norm_vec(w2_)
w2
```

(c)

Computing the angle between the two spaces in **(b)**:

```
inner <- t(v2) %*% (w2)
norm_v2 <- norm_vec(v2)
norm_w2 <- norm_vec(w2)
theta <- acos(inner/(norm_v2 * norm_w2))
theta
```

Part B

Data Entry and Cleaning

Entering the data and defining any values which are assigned question marks to be missing values:

```
mammo <- read.csv("mammo.txt", header=TRUE, na.strings = "?")
```

Removing BI.RADS from our analysis:

```
mammo <- dplyr::select(mammo, Age, Shape, Margin, Density, Severity)
```

Correlation matrix using corrplot package:

```
corMat<-cor(na.omit(dplyr::select(mammo, Age, Shape, Margin,Severity)))
corrplot(corMat, type="upper", order="hclust",tl.col="black",tl.srt = 45 )
```

Pairwise scatterplot for mammo data:

```
ggpairs(mammo)
```

Checking variable class:

```
str(mammo)
```

Assigning varaibles as factors:

```
mammo$Shape <- as.factor(mammo$Shape)
mammo$Margin <- as.factor(mammo$Margin)
mammo$Density <- as.factor(mammo$Density)
mammo$Severity <- as.factor(mammo$Severity)
```

Checking variable classes again:

```
str(mammo)
```

Data Visualisations and Data Summaries

Producing summary statistics:

```
summary(mammo$Age)
summary(mammo$Shape)
summary(mammo$Margin)
summary(mammo$Density)
summary(mammo$Severity)
```

Boxplot of Age Against Severity:

```
plot(mammo$Age~mammo$Severity, xlab = "Severity", ylab = "Age")
```

Boxplot of Age Against Shape:

```
plot(mammo$Age~mammo$Shape, xlab = "Shape", ylab = "Age")
```

Boxplot of Age Against Margin:

```
plot(mammo$Age~mammo$Margin, xlab = "Margin", ylab = "Age")
```

Boxplot of Age Against Density:

```
plot(mammo$Age~mammo$Density, xlab = "Density", ylab = "Age")
```

Bar Graph of Severity Against Shape:

```
ggplot(mammo, aes(x=mammo$Severity))+geom_bar(aes(fill=mammo$Shape)) + labs(x = "Severity")
```

Bar Graph of Severity Against Margin:

```
ggplot(mammo, aes(x=mammo$Severity))+geom_bar(aes(fill=mammo$Margin)) + labs(x = "Severity")
```

Bar Graph of Severity Against Density:

```
ggplot(mammo, aes(x=mammo$Severity))+geom_bar(aes(fill=mammo$Density)) + labs(x = "Severity")
```

Model Fitting and Model Selection

Fitting full model:

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

```
##
## Call:
## glm(formula = Severity ~ (Age + Shape + Margin + Density)^2,
##     family = "binomial", data = mammo)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -8.49    0.00    0.00    0.00    8.49
##
## Coefficients: (6 not defined because of singularities)
##                    Estimate Std. Error    z value Pr(>|z|)
## (Intercept)       1.145e+15  1.268e+08    9031733   <2e-16 ***
## Age              -3.945e+12  1.987e+06   -1985264   <2e-16 ***
## Shape2           -3.499e+15  7.618e+07  -45938639   <2e-16 ***
## Shape3           -1.910e+15  1.286e+08  -14850537   <2e-16 ***
## Shape4            2.176e+15  7.279e+07   29900258   <2e-16 ***
## Margin2           3.100e+13  9.231e+07     335770   <2e-16 ***
## Margin3           1.819e+15  1.024e+08   17759349   <2e-16 ***
## Margin4           7.519e+15  9.137e+07   82290971   <2e-16 ***
## Margin5           1.002e+16  1.293e+08   77489074   <2e-16 ***
## Density2         -8.211e+15  1.351e+08  -60799827   <2e-16 ***
## Density3         -8.190e+15  1.281e+08  -63946926   <2e-16 ***
## Density4         -1.330e+15  3.197e+08   -4159962   <2e-16 ***
## Age:Shape2       -7.824e+13  5.325e+05 -146931580   <2e-16 ***
## Age:Shape3       -4.877e+13  7.073e+05  -68956576   <2e-16 ***
## Age:Shape4       -5.279e+13  6.999e+05  -75428494   <2e-16 ***
## Age:Margin2       3.233e+13  1.379e+06   23446774   <2e-16 ***
## Age:Margin3       2.700e+13  7.572e+05   35662676   <2e-16 ***
## Age:Margin4      -1.031e+13  6.498e+05  -15861366   <2e-16 ***
## Age:Margin5      -2.169e+13  7.926e+05  -27372259   <2e-16 ***
## Age:Density2      9.305e+13  2.121e+06   43870897   <2e-16 ***
## Age:Density3      9.359e+13  2.021e+06   46316759   <2e-16 ***
## Age:Density4     -8.074e+13  5.944e+06  -13584315   <2e-16 ***
## Shape2:Margin2   -2.222e+15  5.882e+07  -37778481   <2e-16 ***
```

```
## Shape3:Margin2   -3.932e+14  5.897e+07   -6667074   <2e-16 ***
## Shape4:Margin2   -2.194e+15  6.504e+07  -33732011   <2e-16 ***
## Shape2:Margin3    2.698e+15  3.949e+07   68316316   <2e-16 ***
## Shape3:Margin3    3.802e+15  4.197e+07   90574690   <2e-16 ***
## Shape4:Margin3    1.029e+15  5.127e+07   20073221   <2e-16 ***
## Shape2:Margin4   -1.594e+15  2.668e+07  -59742302   <2e-16 ***
## Shape3:Margin4   -2.308e+15  3.148e+07  -73317195   <2e-16 ***
## Shape4:Margin4   -3.722e+15  4.398e+07  -84632323   <2e-16 ***
## Shape2:Margin5   -1.904e+15  5.555e+07  -34278477   <2e-16 ***
## Shape3:Margin5    2.211e+15  4.020e+07   54992752   <2e-16 ***
## Shape4:Margin5   -3.362e+15  4.687e+07  -71723706   <2e-16 ***
## Shape2:Density2   5.995e+15  7.668e+07   78187250   <2e-16 ***
## Shape3:Density2   5.592e+15  1.298e+08   43092063   <2e-16 ***
## Shape4:Density2   4.339e+15  8.224e+07   52752807   <2e-16 ***
## Shape2:Density3   8.905e+15  7.192e+07  123807786   <2e-16 ***
## Shape3:Density3   5.987e+15  1.250e+08   47880305   <2e-16 ***
## Shape4:Density3   3.704e+15  7.251e+07   51086925   <2e-16 ***
## Shape2:Density4          NA         NA         NA       NA
## Shape3:Density4          NA         NA         NA       NA
## Shape4:Density4   4.611e+15  1.195e+08   38584598   <2e-16 ***
## Margin2:Density2  2.413e+15  8.081e+07   29860071   <2e-16 ***
## Margin3:Density2 -4.886e+15  9.840e+07  -49651932   <2e-16 ***
## Margin4:Density2 -4.052e+15  8.545e+07  -47417967   <2e-16 ***
## Margin5:Density2 -7.182e+15  1.246e+08  -57636482   <2e-16 ***
## Margin2:Density3         NA         NA         NA       NA
## Margin3:Density3 -5.403e+15  9.042e+07  -59747177   <2e-16 ***
## Margin4:Density3 -3.656e+15  7.927e+07  -46116711   <2e-16 ***
## Margin5:Density3 -5.789e+15  1.177e+08  -49181149   <2e-16 ***
## Margin2:Density4         NA         NA         NA       NA
## Margin3:Density4 -9.448e+14  1.425e+08   -6632161   <2e-16 ***
## Margin4:Density4         NA         NA         NA       NA
## Margin5:Density4         NA         NA         NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  1151.3  on 830  degrees of freedom
## Residual deviance: 12543.2  on 782  degrees of freedom
##   (130 observations deleted due to missingness)
## AIC: 12641
##
## Number of Fisher Scoring iterations: 25
```

Here we note that p-values are non-existent for several of the interaction terms, as such, we begin by removing the interaction between Margin and Density, then view the summary for the updated model:

```
back.glm <- update(full.glm, .~. - Margin:Density)
summary(back.glm)
```

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

```
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -8.49    0.00    0.00    0.00    8.49
##
## Coefficients: (2 not defined because of singularities)
##                   Estimate Std. Error   z value Pr(>|z|)
## (Intercept)     -3.515e+14  1.076e+08  -3267381   <2e-16 ***
## Age              2.020e+13  1.666e+06  12121913   <2e-16 ***
## Shape2          -5.219e+15  6.423e+07 -81255963   <2e-16 ***
## Shape3           8.691e+14  1.020e+08   8522274   <2e-16 ***
## Shape4          -1.242e+15  6.253e+07 -19857241   <2e-16 ***
## Margin2          1.627e+15  9.174e+07  17732775   <2e-16 ***
## Margin3         -9.737e+14  5.191e+07 -18758010   <2e-16 ***
## Margin4          1.624e+15  4.517e+07  35945464   <2e-16 ***
## Margin5          5.254e+15  5.537e+07  94893949   <2e-16 ***
## Density2        -3.345e+15  1.160e+08 -28824738   <2e-16 ***
## Density3        -2.853e+15  1.081e+08 -26404209   <2e-16 ***
## Density4         1.083e+16  2.142e+08  50569686   <2e-16 ***
## Age:Shape2      -8.757e+12  5.297e+05 -16530193   <2e-16 ***
## Age:Shape3       3.817e+12  7.040e+05   5422779   <2e-16 ***
## Age:Shape4       5.279e+13  6.913e+05  76356261   <2e-16 ***
## Age:Margin2     -1.986e+13  1.367e+06 -14528615   <2e-16 ***
## Age:Margin3     -1.061e+13  7.535e+05 -14076728   <2e-16 ***
## Age:Margin4     -1.490e+13  6.427e+05 -23189075   <2e-16 ***
## Age:Margin5     -3.357e+13  7.808e+05 -42993465   <2e-16 ***
## Age:Density2    -1.504e+13  1.802e+06  -8345866   <2e-16 ***
## Age:Density3     2.710e+13  1.684e+06  16098110   <2e-16 ***
## Age:Density4    -3.140e+14  3.776e+06 -83157487   <2e-16 ***
## Shape2:Margin2   1.019e+15  5.879e+07  17338681   <2e-16 ***
## Shape3:Margin2   7.445e+14  5.886e+07  12647565   <2e-16 ***
## Shape4:Margin2  -7.393e+14  6.348e+07 -11647488   <2e-16 ***
## Shape2:Margin3   1.111e+15  3.937e+07  28235116   <2e-16 ***
## Shape3:Margin3   1.642e+15  4.167e+07  39412027   <2e-16 ***
## Shape4:Margin3   1.461e+15  4.929e+07  29646400   <2e-16 ***
## Shape2:Margin4   1.823e+14  2.661e+07   6853146   <2e-16 ***
## Shape3:Margin4  -1.318e+15  3.113e+07 -42322934   <2e-16 ***
## Shape4:Margin4  -2.470e+14  4.204e+07  -5875640   <2e-16 ***
## Shape2:Margin5  -1.171e+15  5.528e+07 -21190674   <2e-16 ***
## Shape3:Margin5  -5.642e+14  4.017e+07 -14044007   <2e-16 ***
## Shape4:Margin5  -2.211e+15  4.463e+07 -49541264   <2e-16 ***
## Shape2:Density2  4.499e+15  6.319e+07  71201451   <2e-16 ***
## Shape3:Density2  1.720e+15  1.010e+08  17033204   <2e-16 ***
## Shape4:Density2  2.403e+15  5.714e+07  42063762   <2e-16 ***
## Shape2:Density3  4.010e+15  5.753e+07  69708915   <2e-16 ***
## Shape3:Density3 -5.976e+14  9.634e+07  -6203280   <2e-16 ***
## Shape4:Density3 -5.724e+14  5.044e+07 -11347657   <2e-16 ***
## Shape2:Density4         NA         NA        NA       NA
## Shape3:Density4         NA         NA        NA       NA
## Shape4:Density4  7.209e+15  9.815e+07  73446873   <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:  1151.3  on 830  degrees of freedom
## Residual deviance: 11966.5  on 790  degrees of freedom
##   (130 observations deleted due to missingness)
## AIC: 12048
##
## Number of Fisher Scoring iterations: 25
```

Here we see that p-values are still non-existent for some levels of the interaction between Shape and Density, as such we remove this interaction from the model:

```
back.glm <- update(back.glm, .~. - Shape:Density)
summary(back.glm)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density + Age:Shape +
##     Age:Margin + Age:Density + Shape:Margin, family = "binomial",
##     data = mammo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3975  -0.5406  -0.1222   0.6376   2.5452
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.728e-01  3.301e+00  -0.113   0.9101
## Age            -1.195e-02  5.533e-02  -0.216   0.8290
## Shape2          2.698e+00  1.662e+00   1.623   0.1045
## Shape3          2.631e+00  2.019e+00   1.303   0.1925
## Shape4          3.344e+00  2.108e+00   1.586   0.1128
## Margin2         2.551e+00  3.420e+00   0.746   0.4557
## Margin3        -1.546e+01  1.065e+03  -0.015   0.9884
## Margin4         1.514e+00  1.821e+00   0.832   0.4055
## Margin5         3.421e+00  2.336e+00   1.464   0.1431
## Density2       -5.843e+00  3.650e+00  -1.601   0.1094
## Density3       -6.481e+00  3.400e+00  -1.906   0.0566 .
## Density4        9.525e-01  6.606e+00   0.144   0.8853
## Age:Shape2     -4.793e-02  2.894e-02  -1.656   0.0977 .
## Age:Shape3     -3.439e-02  3.430e-02  -1.003   0.3161
## Age:Shape4     -4.674e-02  3.310e-02  -1.412   0.1580
## Age:Margin2    -1.432e-02  5.261e-02  -0.272   0.7855
## Age:Margin3     8.662e-03  3.373e-02   0.257   0.7974
## Age:Margin4     1.521e-02  2.772e-02   0.549   0.5832
## Age:Margin5     4.603e-04  3.584e-02   0.013   0.9898
## Age:Density2    8.236e-02  6.185e-02   1.332   0.1830
## Age:Density3    1.011e-01  5.737e-02   1.762   0.0780 .
## Age:Density4   -3.571e-02  1.077e-01  -0.332   0.7402
## Shape2:Margin2 -1.037e+00  2.046e+00  -0.507   0.6123
## Shape3:Margin2  7.454e-01  2.045e+00   0.364   0.7155
## Shape4:Margin2  6.588e-01  2.125e+00   0.310   0.7565
## Shape2:Margin3  1.455e+01  1.065e+03   0.014   0.9891
## Shape3:Margin3  1.612e+01  1.065e+03   0.015   0.9879
## Shape4:Margin3  1.724e+01  1.065e+03   0.016   0.9871
## Shape2:Margin4 -9.464e-01  1.022e+00  -0.926   0.3545
```

```
## Shape3:Margin4 -1.239e+00  1.158e+00  -1.070   0.2848
## Shape4:Margin4 -1.434e-01  1.454e+00  -0.099   0.9214
## Shape2:Margin5 -2.036e+00  1.923e+00  -1.059   0.2896
## Shape3:Margin5  1.390e+01  8.169e+02   0.017   0.9864
## Shape4:Margin5 -9.068e-01  1.644e+00  -0.552   0.5813
## ---
## 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:  700.13  on 797  degrees of freedom
##   (130 observations deleted due to missingness)
## AIC: 768.13
##
## Number of Fisher Scoring iterations: 15
```

Now we can see that all of the terms have a valid p-value, and continue our selection process by removing the least statistically signicant terms. We see that Age:Margin5 has the highest p-value of 0.9898, and no other level of Age:Margin are significant, so it is removed from the model:

```
back.glm <- update(back.glm, .~. - Age:Margin)
summary(back.glm)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density + Age:Shape +
##     Age:Density + Shape:Margin, family = "binomial", data = mammo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4091  -0.5434  -0.1183   0.6462   2.5569
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.30269    3.27016  -0.093  0.92625
## Age              -0.01250    0.05488  -0.228  0.81976
## Shape2            2.60522    1.66156   1.568  0.11690
## Shape3            2.41408    1.92617   1.253  0.21010
## Shape4            2.92053    1.75411   1.665  0.09592 .
## Margin2           1.72914    1.67114   1.035  0.30081
## Margin3         -14.92583 1082.87384  -0.014  0.98900
## Margin4           2.38301    0.86735   2.747  0.00601 **
## Margin5           3.45360    1.18009   2.927  0.00343 **
## Density2         -5.99849    3.59952  -1.666  0.09562 .
## Density3         -6.65705    3.33506  -1.996  0.04592 *
## Density4          0.67474    6.53932   0.103  0.91782
## Age:Shape2       -0.04603    0.02875  -1.601  0.10935
## Age:Shape3       -0.03030    0.03204  -0.946  0.34427
## Age:Shape4       -0.03891    0.02489  -1.563  0.11797
## Age:Density2      0.08406    0.06118   1.374  0.16947
## Age:Density3      0.10347    0.05643   1.834  0.06669 .
## Age:Density4     -0.03151    0.10656  -0.296  0.76747
## Shape2:Margin2   -0.86503    2.03886  -0.424  0.67137
## Shape3:Margin2    0.71906    2.12769   0.338  0.73540
```

```
## Shape4:Margin2    0.53585    2.16819   0.247  0.80480
## Shape2:Margin3   14.49377 1082.87440   0.013  0.98932
## Shape3:Margin3   16.08392 1082.87413   0.015  0.98815
## Shape4:Margin3   17.18163 1082.87448   0.016  0.98734
## Shape2:Margin4   -0.94777    0.98470  -0.962  0.33580
## Shape3:Margin4   -1.24290    1.12697  -1.103  0.27008
## Shape4:Margin4   -0.16325    1.42650  -0.114  0.90889
## Shape2:Margin5   -2.08708    1.90358  -1.096  0.27290
## Shape3:Margin5   13.85038  810.22692   0.017  0.98636
## Shape4:Margin5   -0.94216    1.64296  -0.573  0.56634
## ---
## 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:  700.87  on 801  degrees of freedom
##   (130 observations deleted due to missingness)
## AIC: 760.87
##
## Number of Fisher Scoring iterations: 15
```

Now we see that the highest p-value is for Shape2:Margin3, and no other levels of the interaction between shape and margin are significant, so the interaction is removed from the model:

```
back.glm <- update(back.glm, .~. - Shape:Margin)
summary(back.glm)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density + Age:Shape +
##     Age:Density, family = "binomial", data = mammo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5002  -0.5626  -0.1418   0.6615   2.5040
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.899176   3.090399  -0.291 0.771083
## Age          -0.002679   0.052234  -0.051 0.959101
## Shape2        2.054544   1.572794   1.306 0.191450
## Shape3        1.969599   1.804794   1.091 0.275134
## Shape4        3.359258   1.394334   2.409 0.015987 *
## Margin2       1.635165   0.566231   2.888 0.003879 **
## Margin3       1.201836   0.358062   3.357 0.000789 ***
## Margin4       1.486203   0.310014   4.794 1.64e-06 ***
## Margin5       1.991353   0.381001   5.227 1.73e-07 ***
## Density2     -4.854676   3.406741  -1.425 0.154151
## Density3     -5.818577   3.116658  -1.867 0.061912 .
## Density4      0.449932   6.468724   0.070 0.944548
## Age:Shape2   -0.041274   0.027078  -1.524 0.127449
## Age:Shape3   -0.024154   0.030316  -0.797 0.425600
## Age:Shape4   -0.035522   0.023272  -1.526 0.126920
## Age:Density2  0.067510   0.058312   1.158 0.246971
```

```
## Age:Density3  0.091150   0.053070   1.718 0.085878 .
## Age:Density4 -0.027289   0.105477  -0.259 0.795851
## ---
## 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:  719.29  on 813  degrees of freedom
##   (130 observations deleted due to missingness)
## AIC: 755.29
##
## Number of Fisher Scoring iterations: 5
```

Here we see that of the interaction terms, Age:Density4 has the highest p-value of 0.7959, and no other levels of this interaction are significant, so the model is updated with the removal of this interaction:

```
back.glm <- update(back.glm, .~. - Age:Density)
summary(back.glm)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density + Age:Shape,
##     family = "binomial", data = mammo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4612  -0.5599  -0.1680   0.6685   2.4570
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.59670    1.39139  -4.022 5.76e-05 ***
## Age          0.07751    0.01899   4.081 4.49e-05 ***
## Shape2       1.65276    1.51035   1.094  0.27383
## Shape3       1.61155    1.73973   0.926  0.35428
## Shape4       3.06469    1.33593   2.294  0.02179 *
## Margin2      1.59268    0.56180   2.835  0.00458 **
## Margin3      1.15373    0.35333   3.265  0.00109 **
## Margin4      1.45682    0.30574   4.765 1.89e-06 ***
## Margin5      1.97620    0.37751   5.235 1.65e-07 ***
## Density2    -0.86281    0.80468  -1.072  0.28361
## Density3    -0.51761    0.73127  -0.708  0.47906
## Density4    -1.57940    1.07472  -1.470  0.14167
## Age:Shape2  -0.03359    0.02597  -1.293  0.19585
## Age:Shape3  -0.01658    0.02910  -0.570  0.56880
## Age:Shape4  -0.02927    0.02211  -1.324  0.18559
## ---
## 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:  724.63  on 816  degrees of freedom
##   (130 observations deleted due to missingness)
## AIC: 754.63
```

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

Here we see that Age:Shape3 has the highest p-value of 0.5688, and the other levels of the interaction between Age and Shape are also non-significant, so the interaction is removed from the model:

```
back.glm <- update(back.glm, .~. - Age:Shape)
summary(back.glm)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density, family = "binomial",
##     data = mammo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5286  -0.5632  -0.2190   0.6645   2.5553
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.175195   0.847993  -4.924 8.50e-07 ***
## Age          0.054783   0.007807   7.017 2.27e-12 ***
## Shape2      -0.259327   0.319292  -0.812 0.416681
## Shape3       0.658338   0.375749   1.752 0.079762 .
## Shape4       1.370209   0.333060   4.114 3.89e-05 ***
## Margin2      1.640629   0.559287   2.933 0.003352 **
## Margin3      1.182762   0.351871   3.361 0.000776 ***
## Margin4      1.483740   0.302603   4.903 9.43e-07 ***
## Margin5      2.012203   0.374699   5.370 7.87e-08 ***
## Density2    -0.959989   0.797154  -1.204 0.228485
## Density3    -0.653906   0.718090  -0.911 0.362497
## Density4    -1.751671   1.062852  -1.648 0.099335 .
## ---
## 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:  726.96  on 819  degrees of freedom
##   (130 observations deleted due to missingness)
## AIC: 750.96
##
## Number of Fisher Scoring iterations: 5
```

Now we note that the model has been reduced to the additive model with no interaction terms. In this model we see that the fourth level of density has the highest p-value, and no other levels are significant, so Density is removed from the model:

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

In this model, we note that all of the terms have at least on level which is statistically significant, so no terms should be removed. We assign this model the name final.glm:

```
final.glm <- back.glm
```

Justification of the final model

Printing model summary:

```
summary(final.glm)
```

Predicting Probabilities and Interpretation

Fitting probabilities from data:

```
probabilities <- fitted(final.glm)
summary(probabilities)
```

Producing histogram of probabilties:

```
hist(probabilities, xlab = "Probability")
```

Assigning new version of mammo data with only Age, Margin, Shape and Density and no missing values:

```
newMammo <- mammo %>% select(Age, Margin, Shape, Severity)
newMammo <- na.omit(newMammo)
```

Plotting probabilties against Age:

```r
plot(probabilities ~ newMammo$Age, ylab = "Probability", xlab = "Age")
```

Plotting probabilties against Shape:

```r
boxplot(probabilities ~ newMammo$Shape, ylab = "Probability", xlab = "Shape")
```

Plotting probabilties against Margin:

```r
boxplot(probabilities ~ newMammo$Margin, ylab = "Probability", xlab = "Margin")
```

Predicting probability of malignant tumor for patient of Age = 40, Shape = 1, Margin = 1:

```r
predict(final.glm,data.frame(Age=40,Shape="1",Margin="1"), type = "response")
```

Predicting probability of malignant tumor for patient of Age = 80, Shape = 4, Margin = 5:

```r
predict(final.glm,data.frame(Age=80,Shape="4",Margin="5"), type = "response")
```

Predicting probability of malignant tumor for patient of Age = 60, Shape = 2, Margin = 3:

```
##         1
## 0.3381399
```