Part B

Introduction

In this section we obtain a predictive model for mammographic mass severity, a measure of the status of mammographic mass legions, 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 legions, 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.

This introduction should probably be reworked but I this hope is a good starting point

Data Entry and Cleaning

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

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

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

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

We can now check the variable types for the data:

```
str(mammo)
```

```
## '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 varaibles, and as such convert them:

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

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

```
str(mammo)
```

```
## '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:

```
summary(mammo$Age)
```

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

```
summary(mammo$Shape)
```

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

```
summary(mammo$Margin)
```

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

```
summary(mammo$Density)
```

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

```
summary(mammo$Severity)
```

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

We also create a pairwise scatterplot to see the relationships between indivisual varaibles:

```
pairs(mammo)
```

Model Fitting and Model Selection

We now fit a logistic linear model (M1) to the data, with Severity as the response variable, and Age, Shape, Margina and Density as the predictor variables:

```
M1 <- glm(Severity ~ Age+Shape+Margin+Density, data = mammo, family = "binomial")
summary(M1)
```

```
##
## 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 ***
```

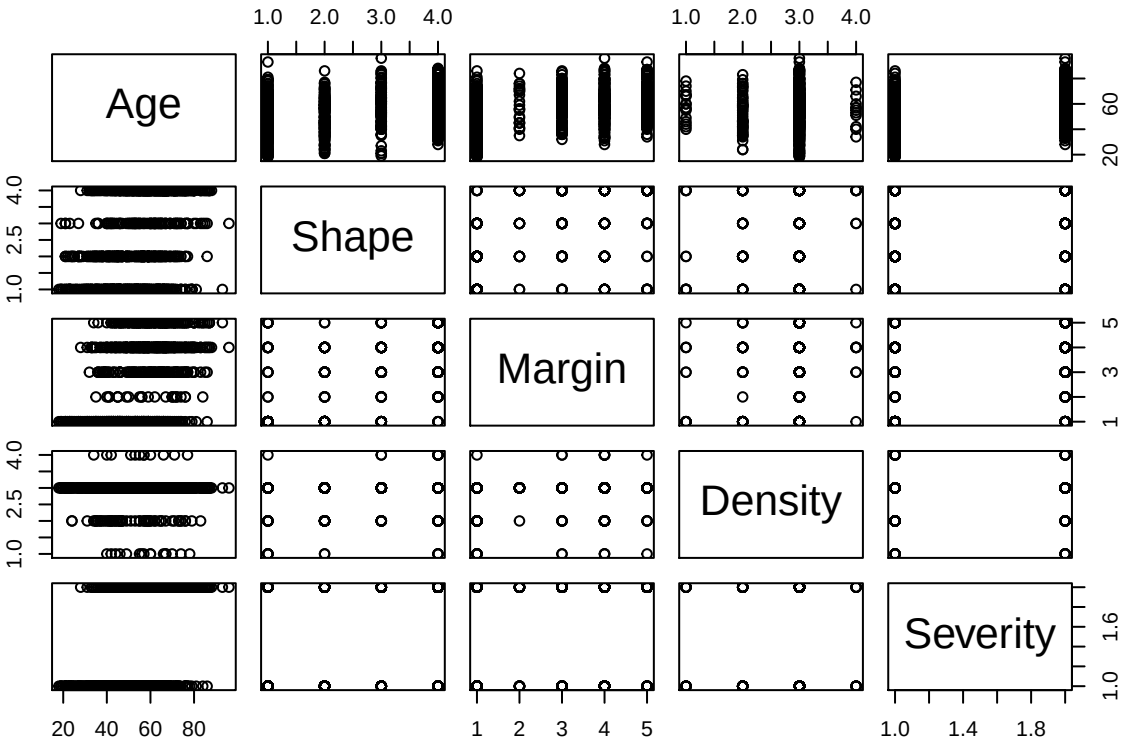

Figure 1: Pairwise scatterplot of Mammographic Mass Severity Data

```
## 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
```

In this summary, we note that the only statistically significant predictors are Age, Shape and Margin... **This is where we get a bit iffy**