Section 2

Lily, James, Tobin

30/05/2018

Part B: Logistic Regression

Introduction

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 Mammography. 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 (round= 1, oval= 2, lobular= 3, irregular= 4), the margin (circumscribed= 1, microlobulated= 2, obscured= 3, ill-defined= 4, spiculated= 5) and the density (high= 1, iso= 2, low= 3, fat-containing= 4).

This approach 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

```
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
```

After taking the data it is clear there are a few key issues to deal with in cleaning the data. The first 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", for appropriate use within the models. These data points could be removed as they are missing some data, however this should not be done for two reasons. Firstly, incomplete data should not be thrown away as it may still contain valuable information. Furthermore, the final model may not include some of the predictor attributes, and some of the currently incomplete data may not be missing information in any of the include attributes.

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

[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].

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

Severity has no missing data.

```
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.

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

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

Shape has 31 missing data points.

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

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

Margin has 48 missing data points.

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

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

Density has 76 missing data points.

```
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.

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

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

```
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.

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

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

```
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

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

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

```
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

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

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

```
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

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

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

```
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
```

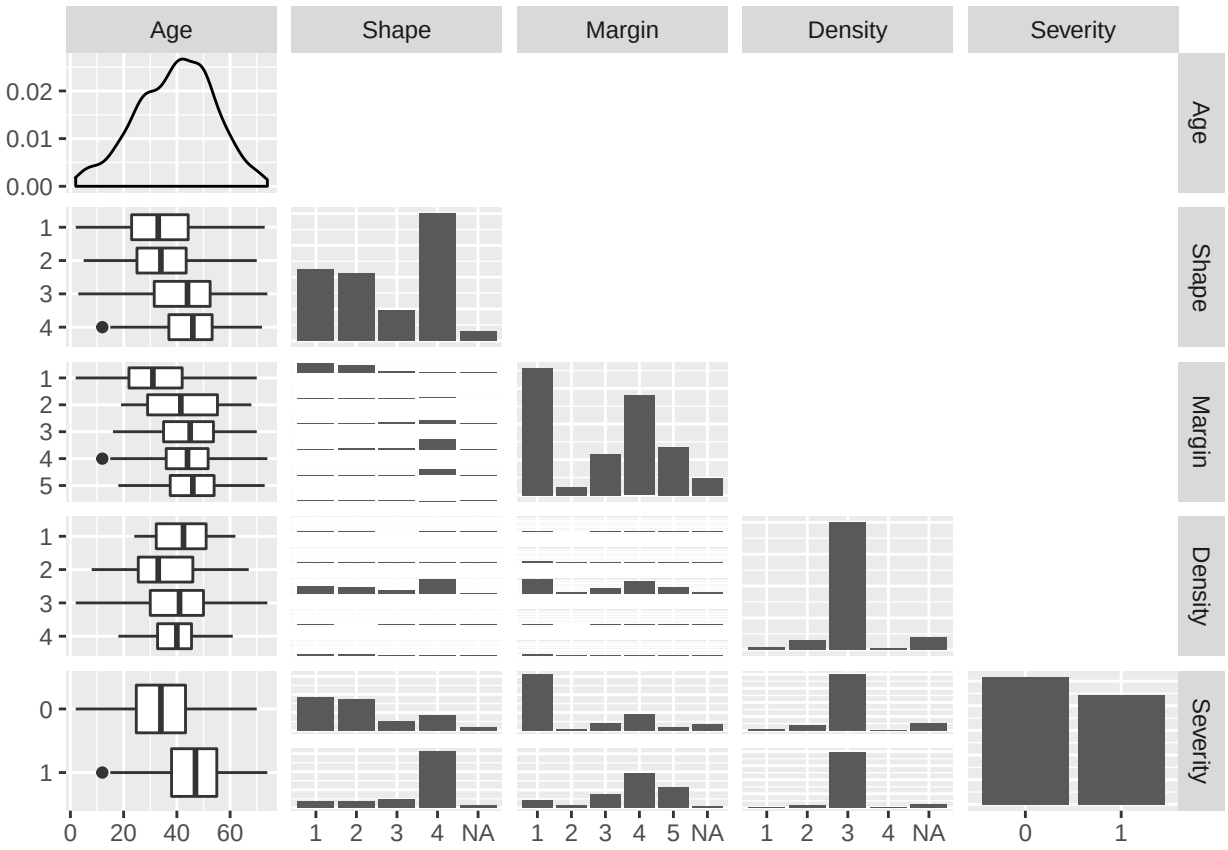
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   4.000   4.312   5.000   7.000       3
```

Can these varibles be interpolated (no)

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.

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) # Use complet
```

The Full Model

As we begin the model creation process, our first step is to create a model that uses all of our 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 because 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 malignant mass cases, while the other 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 the oddball. 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.

```
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
## -2.5293  -0.5633  -0.2188   0.6644   2.5558
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.299495   0.787279  -4.191 2.78e-05 ***
## Age          0.054818   0.007812   7.017 2.26e-12 ***
## Shape2      -0.260623   0.319239  -0.816 0.414277
## Shape3       0.658256   0.375563   1.753 0.079650 .
## Shape4       1.368033   0.332866   4.110 3.96e-05 ***
## Margin2      1.641176   0.559288   2.934 0.003342 **
## Margin3      1.183720   0.351830   3.364 0.000767 ***
## Margin4      1.485340   0.302544   4.910 9.13e-07 ***
## Margin5      2.014630   0.374494   5.380 7.46e-08 ***
## Density2    -0.960317   0.797210  -1.205 0.228359
```

```
## Density3    -0.653884   0.718144  -0.911 0.362549
## Density4    -1.752164   1.062859  -1.649 0.099242 .
## ---
## 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:  727.06  on 819  degrees of freedom
## AIC: 751.06
##
## Number of Fisher Scoring iterations: 5
```

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

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density, family = binomial,
##     data = mammo[complete, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5293  -0.5633  -0.2188   0.6644   2.5558
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.814154   0.861865  -2.105   0.0353 *
## Age          0.054818   0.007812   7.017 2.26e-12 ***
## Shape2      -0.260623   0.319239  -0.816   0.4143
## Shape3       0.658256   0.375563   1.753   0.0797 .
## Shape4       1.368033   0.332866   4.110 3.96e-05 ***
## Margin1     -1.485340   0.302544  -4.910 9.13e-07 ***
## Margin2      0.155835   0.533590   0.292   0.7702
## Margin3     -0.301620   0.270561  -1.115   0.2649
## Margin5      0.529290   0.293930   1.801   0.0717 .
## Density2    -0.960317   0.797210  -1.205   0.2284
## Density3    -0.653884   0.718144  -0.911   0.3625
## Density4    -1.752164   1.062859  -1.649   0.0992 .
## ---
## 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:  727.06  on 819  degrees of freedom
## AIC: 751.06
##
## Number of Fisher Scoring iterations: 5
```

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

Model by stepwise selection

Pretty self explanatory: find the scale, r fits the model based on AIC values.

```r
s2 <- sum((mod.full$residuals)^2)/mod.full$df.residual
s2
```

```
## [1] 10.01254
```

```r
# Stepwise Selection

mod.step <- step(mod.full, scale = s2)
```

```
## Start:  AIC=751.06
## Severity ~ Age + Shape + Margin + Density
##
##           Df Deviance    AIC
## - Density  3   730.37 745.39
## - Margin   4   762.40 746.59
## - Shape    3   762.11 748.56
## <none>         727.06 751.06
## - Age      1   782.17 754.57
##
## Step:  AIC=748.37
## Severity ~ Age + Shape + Margin
##
##          Df Deviance    AIC
## - Margin  4   765.67 743.90
## - Shape   3   765.22 745.85
## <none>        730.37 748.37
## - Age     1   786.88 752.02
##
## Step:  AIC=775.67
## Severity ~ Age + Shape
```

```r
summary(mod.step)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape, family = binomial, data = mammo[complete,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4502  -0.6555  -0.2487   0.6981   2.3990
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.892224   0.366037 -10.633  < 2e-16 ***
## Age          0.061596   0.007446   8.273  < 2e-16 ***
## Shape2       0.025425   0.290599   0.087     0.93
## Shape3       1.399304   0.317689   4.405 1.06e-05 ***
## Shape4       2.531223   0.242834  10.424  < 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:  765.67  on 826  degrees of freedom
## AIC: 775.67
##
## Number of Fisher Scoring iterations: 5
```

```
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
## -2.5293  -0.5633  -0.2188   0.6644   2.5558
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.814154   0.861865  -2.105   0.0353 *
## Age          0.054818   0.007812   7.017 2.26e-12 ***
## Shape2      -0.260623   0.319239  -0.816   0.4143
## Shape3       0.658256   0.375563   1.753   0.0797 .
## Shape4       1.368033   0.332866   4.110 3.96e-05 ***
## Margin1     -1.485340   0.302544  -4.910 9.13e-07 ***
## Margin2      0.155835   0.533590   0.292   0.7702
## Margin3     -0.301620   0.270561  -1.115   0.2649
## Margin5      0.529290   0.293930   1.801   0.0717 .
## Density2    -0.960317   0.797210  -1.205   0.2284
## Density3    -0.653884   0.718144  -0.911   0.3625
## Density4    -1.752164   1.062859  -1.649   0.0992 .
## ---
## 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:  727.06  on 819  degrees of freedom
## AIC: 751.06
##
## Number of Fisher Scoring iterations: 5
```

```
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
## -2.5221  -0.5543  -0.2172   0.6692   2.5605
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.500638   0.465642  -5.370 7.86e-08 ***
## Age          0.055165   0.007772   7.098 1.26e-12 ***
## Shape2      -0.276329   0.317199  -0.871    0.384
## Shape3       0.610999   0.372984   1.638    0.101
## Shape4       1.345583   0.332329   4.049 5.14e-05 ***
## Margin1     -1.455603   0.299929  -4.853 1.22e-06 ***
## Margin2      0.189788   0.534100   0.355    0.722
## Margin3     -0.277907   0.269810  -1.030    0.303
## Margin5      0.541925   0.292604   1.852    0.064 .
## ---
## 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:  730.37  on 822  degrees of freedom
## AIC: 748.37
##
## Number of Fisher Scoring iterations: 5
```

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

```
##
## Call:
## glm(formula = Severity ~ Age + Margin, family = binomial, data = mammo[complete,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4735  -0.5926  -0.2196   0.7219   2.4790
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.652190   0.347059  -4.761 1.93e-06 ***
## Age          0.057279   0.007557   7.580 3.46e-14 ***
## Margin1     -2.404041   0.229427 -10.478  < 2e-16 ***
## Margin2     -0.144246   0.486316  -0.297  0.76676
## Margin3     -0.251294   0.256682  -0.979  0.32758
## Margin5      0.768274   0.284493   2.701  0.00692 **
## ---
## 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:  765.22  on 825  degrees of freedom
## AIC: 777.22
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(mod.full)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density, family = binomial,
##     data = mammo[complete, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5293  -0.5633  -0.2188   0.6644   2.5558
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.814154   0.861865  -2.105   0.0353 *
## Age          0.054818   0.007812   7.017 2.26e-12 ***
## Shape2      -0.260623   0.319239  -0.816   0.4143
## Shape3       0.658256   0.375563   1.753   0.0797 .
## Shape4       1.368033   0.332866   4.110 3.96e-05 ***
## Margin1     -1.485340   0.302544  -4.910 9.13e-07 ***
## Margin2      0.155835   0.533590   0.292   0.7702
## Margin3     -0.301620   0.270561  -1.115   0.2649
## Margin5      0.529290   0.293930   1.801   0.0717 .
## Density2    -0.960317   0.797210  -1.205   0.2284
## Density3    -0.653884   0.718144  -0.911   0.3625
## Density4    -1.752164   1.062859  -1.649   0.0992 .
## ---
## 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:  727.06  on 819  degrees of freedom
## AIC: 751.06
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(mod.back)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Margin, family = binomial, data = mammo[complete,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4735  -0.5926  -0.2196   0.7219   2.4790
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.652190   0.347059  -4.761 1.93e-06 ***
## Age          0.057279   0.007557   7.580 3.46e-14 ***
## Margin1     -2.404041   0.229427 -10.478  < 2e-16 ***
## Margin2     -0.144246   0.486316  -0.297  0.76676
## Margin3     -0.251294   0.256682  -0.979  0.32758
```

```
## Margin5      0.768274   0.284493   2.701  0.00692 **
## ---
## 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:  765.22  on 825  degrees of freedom
## AIC: 777.22
##
## Number of Fisher Scoring iterations: 5
```

```
summary(mod.full)
```

```
##
## Call:
## glm(formula = Severity ~ Age + Shape + Margin + Density, family = binomial,
##     data = mammo[complete, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5293  -0.5633  -0.2188   0.6644   2.5558
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.814154   0.861865  -2.105   0.0353 *
## Age          0.054818   0.007812   7.017 2.26e-12 ***
## Shape2      -0.260623   0.319239  -0.816   0.4143
## Shape3       0.658256   0.375563   1.753   0.0797 .
## Shape4       1.368033   0.332866   4.110 3.96e-05 ***
## Margin1     -1.485340   0.302544  -4.910 9.13e-07 ***
## Margin2      0.155835   0.533590   0.292   0.7702
## Margin3     -0.301620   0.270561  -1.115   0.2649
## Margin5      0.529290   0.293930   1.801   0.0717 .
## Density2    -0.960317   0.797210  -1.205   0.2284
## Density3    -0.653884   0.718144  -0.911   0.3625
## Density4    -1.752164   1.062859  -1.649   0.0992 .
## ---
## 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:  727.06  on 819  degrees of freedom
## AIC: 751.06
##
## Number of Fisher Scoring iterations: 5
```

```
qchisq(1-0.05, df = 7)
```

```
## [1] 14.06714
```

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

```
## [1] 38.60816
```

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

```
## Analysis of Deviance Table
##
## Model 1: Severity ~ Age + Shape
## Model 2: Severity ~ Age + Shape + Margin + Density
##   Resid. Df Resid. Dev Df Deviance
## 1       826     765.67
## 2       819     727.06  7   38.608
```

```r
qchisq(1-0.05, df = 2)
```

```
## [1] 5.991465
```

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

```
## [1] 0.4523015
```

```r
anova(mod.step, mod.back) # These literally say the same thing but in the prac we use the difference
```

```
## Analysis of Deviance Table
##
## Model 1: Severity ~ Age + Shape
## Model 2: Severity ~ Age + Margin
##   Resid. Df Resid. Dev Df Deviance
## 1       826     765.67
## 2       825     765.22  1   0.4523
```

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

```
##          df      AIC
## mod.full 12 751.0626
## mod.step  5 775.6708
## mod.back  6 777.2185
```

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 79.6% success rate.

```r
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] 0.796
```

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 78.7% which is barely less 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, 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] 0.787
```

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

Above each row represents a different age and each column a different shape.