

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 predictive value of breast biopsy from mammograms. This low predictive value has been found to lead to approximately 70% of unnecessary 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 variable 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 variables, 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 statistics 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
```

```
print(" ")
```

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

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

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

```
print(" ")
```

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

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

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

```
print(" ")
```

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

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

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

```
print(" ")
```

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

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

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

We also create a pairwise scatterplot to observe the relationships between individual variables:

```
pairs(mammo)
```

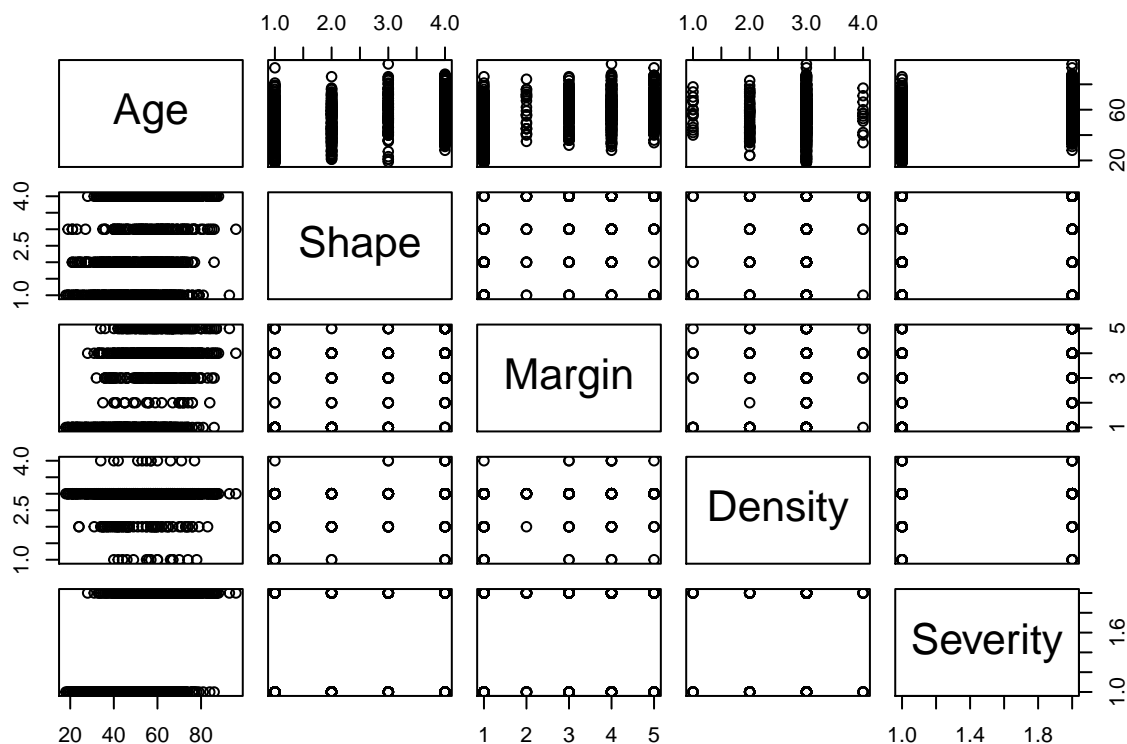


Figure 1: Pairwise scatterplot of Mammographic Mass Severity Data

There appears to be a weak, possibly linear, positive relationship between Age and Severity. There are no observable relationships between Severity and the other predictors.

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:

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

##	term	estimate	std.error	statistic	p.value
## 1	(Intercept)	1.145321e+15	126810730.3	9031732.7	0
## 2	Age	-3.945173e+12	1987227.8	-1985264.4	0
## 3	Shape2	-3.499485e+15	76177375.6	-45938639.3	0
## 4	Shape3	-1.910292e+15	128634518.7	-14850537.4	0
## 5	Shape4	2.176361e+15	72787369.0	29900257.6	0
## 6	Margin2	3.099532e+13	92311019.4	335770.5	0
## 7	Margin3	1.819304e+15	102442051.4	17759348.6	0
## 8	Margin4	7.518742e+15	91367764.2	82290970.9	0
## 9	Margin5	1.001668e+16	129265718.6	77489074.1	0
## 10	Density2	-8.211311e+15	135054845.3	-60799826.6	0

```
## 11      Density3 -8.189589e+15 128068538.9 -63946925.9      0
## 12      Density4 -1.329877e+15 319684980.9 -4159961.7      0
## 13      Age:Shape2 -7.823928e+13  532487.8 -146931580.0      0
## 14      Age:Shape3 -4.877312e+13  707302.0 -68956575.7      0
## 15      Age:Shape4 -5.279231e+13  699898.7 -75428494.0      0
## 16      Age:Margin2  3.232632e+13 1378710.6  23446773.8      0
## 17      Age:Margin3  2.700317e+13  757182.9  35662675.7      0
## 18      Age:Margin4 -1.030709e+13  649823.7 -15861365.8      0
## 19      Age:Margin5 -2.169454e+13  792573.8 -27372258.9      0
## 20      Age:Density2  9.304787e+13 2120947.6  43870897.3      0
## 21      Age:Density3  9.358883e+13 2020625.6  46316758.5      0
## 22      Age:Density4 -8.074194e+13  5943762.6 -13584315.1      0
## 23      Shape2:Margin2 -2.222298e+15 58824444.3 -37778481.4      0
## 24      Shape3:Margin2 -3.931754e+14 58972702.1 -6667074.1      0
## 25      Shape4:Margin2 -2.193930e+15 65039987.1 -33732010.6      0
## 26      Shape2:Margin3  2.697516e+15 39485681.0  68316316.2      0
## 27      Shape3:Margin3  3.801713e+15 41973234.7  90574689.7      0
## 28      Shape4:Margin3  1.029106e+15 51267626.8  20073220.5      0
## 29      Shape2:Margin4 -1.594016e+15 26681536.4 -59742302.4      0
## 30      Shape3:Margin4 -2.307825e+15 31477264.7 -73317194.6      0
## 31      Shape4:Margin4 -3.722253e+15 43981453.6 -84632323.2      0
## 32      Shape2:Margin5 -1.904170e+15 55550004.0 -34278477.0      0
## 33      Shape3:Margin5  2.210729e+15 40200379.2  54992751.8      0
## 34      Shape4:Margin5 -3.361528e+15 46867735.9 -71723706.1      0
## 35      Shape2:Density2  5.995266e+15 76678304.4  78187250.3      0
## 36      Shape3:Density2  5.591895e+15 129766250.1  43092063.0      0
## 37      Shape4:Density2  4.338634e+15 82244616.0  52752807.4      0
## 38      Shape2:Density3  8.904513e+15 71922072.5 123807786.3      0
## 39      Shape3:Density3  5.986948e+15 125039893.9  47880305.4      0
## 40      Shape4:Density3  3.704195e+15 72507682.5  51086925.3      0
## 41      Shape4:Density4  4.611210e+15 119509082.3  38584598.3      0
## 42      Margin2:Density2  2.412959e+15 80808895.7  29860070.7      0
## 43      Margin3:Density2 -4.885525e+15 98395475.4 -49651931.8      0
## 44      Margin4:Density2 -4.051699e+15 85446484.4 -47417966.6      0
## 45      Margin5:Density2 -7.182269e+15 124613243.0 -57636481.6      0
## 46      Margin3:Density3 -5.402593e+15 90424241.9 -59747177.4      0
## 47      Margin4:Density3 -3.655562e+15 79267630.6 -46116710.7      0
## 48      Margin5:Density3 -5.788975e+15 117707184.0 -49181149.3      0
## 49      Margin3:Density4 -9.447530e+14 142450245.5 -6632161.2      0
```

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 significant 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
```

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 variables. 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 probabilities of each datapoint in the dataset based on the final model:

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

```
##    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:

```
hist(probabilities)
```

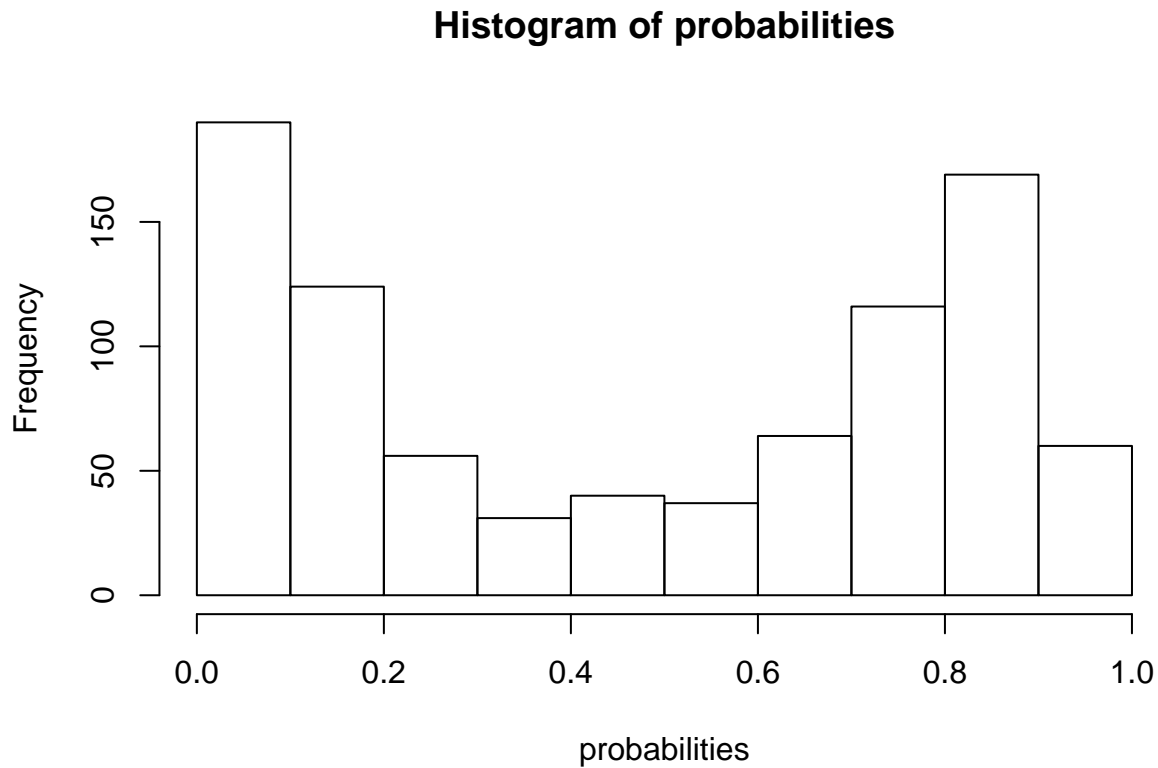


Figure 2: Histogram of Fitted Probabilities

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 mammo data, including only Age, Shape, Margin and Severity, and ignoring the missing values in order to be able to create valid plots:

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

We can now create plots of probabilities against Age, Shape and Margin:

```
plot(probabilities ~ newMammo$Age)
```

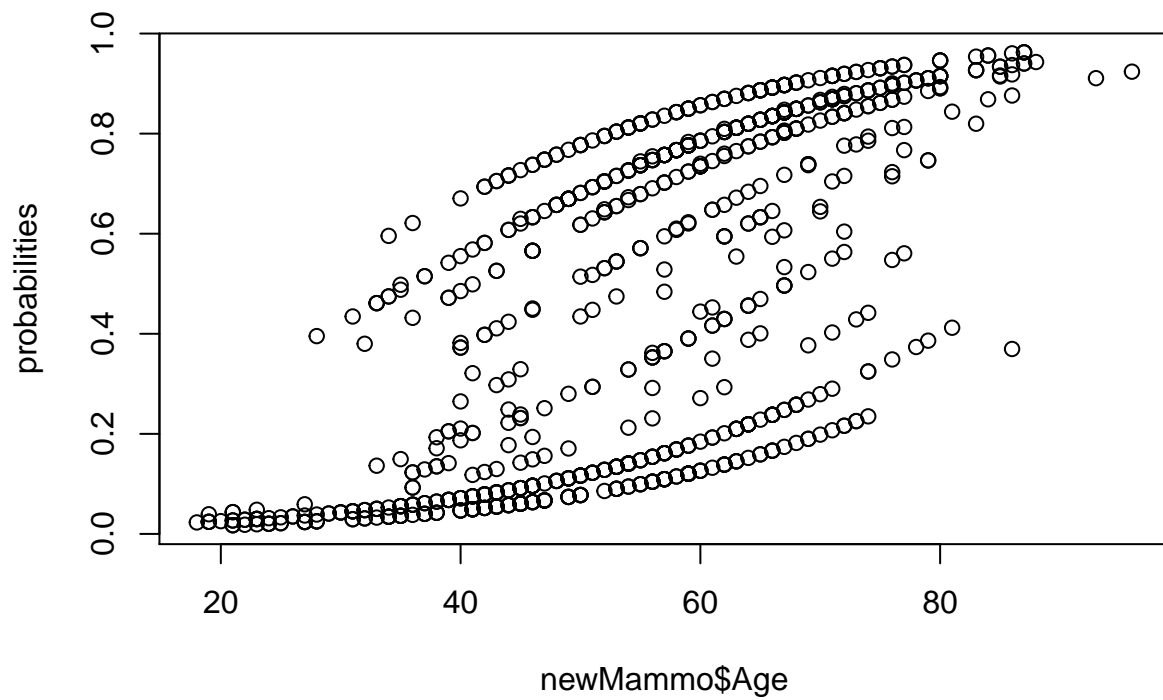


Figure 3: Probabilities against Age

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.

```
boxplot(probabilities ~ newMammo$Shape)
```

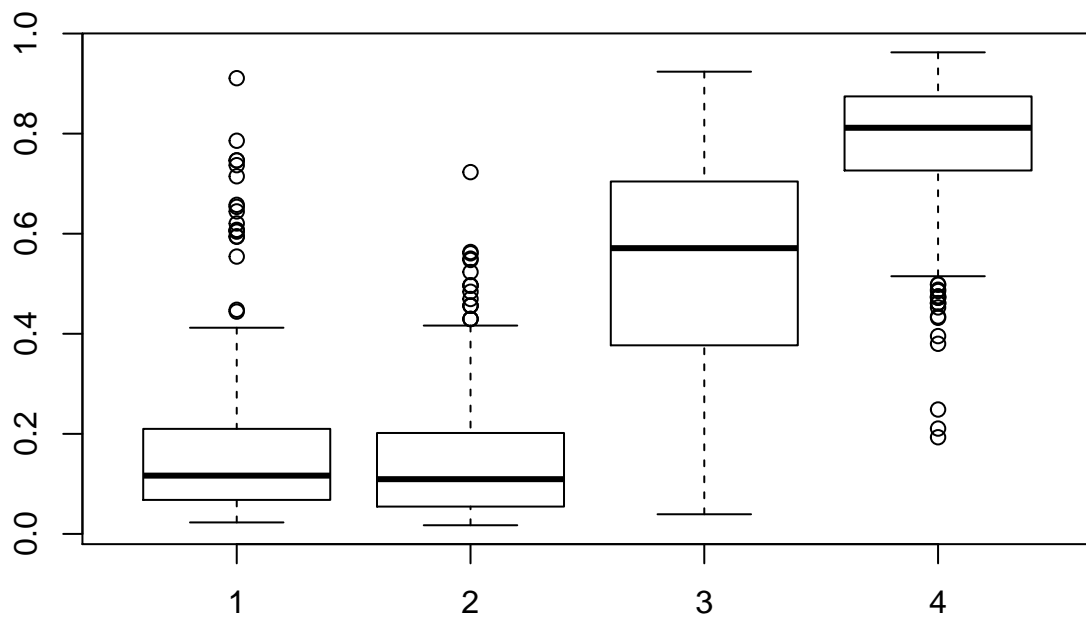


Figure 4: Probabilities against Shape

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.

```
boxplot(probabilities ~ newMammo$Margin)
```

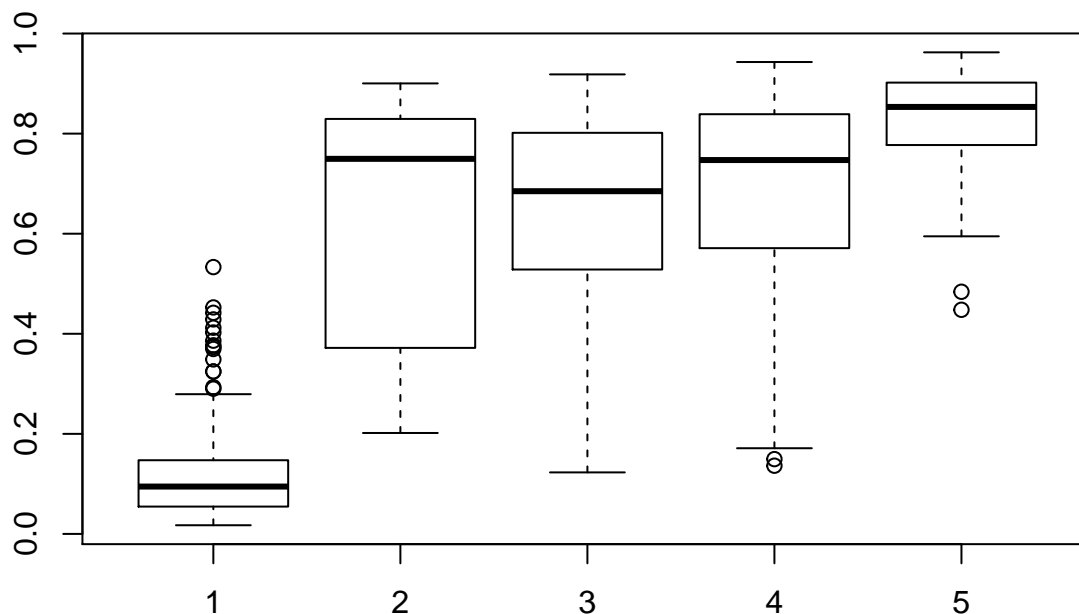


Figure 5: Probabilities against Margin

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.

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:

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

```
##          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:

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

```
##           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):

```
predict(final.glm,data.frame(Age=60,Shape="2",Margin="3"), type = "response")
```

```
##           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.