Mammographic Masses
# of Tumors
MATH-2200-01

Ryan Maresca

April 18, 2018

Instructor: Mark Mixer

1 Objective

The goal of this paper is to interpret the data provided by the UCI Machine Learning website and determine whether or not a tumor is malignant. Mammographies are X-ray examinations of breasts used to detect early signs of cancer, it is to be the most powerful method for breast cancer screenings today. However, this method can cause unnecessary biopsies that have benign outcomes. This obviously should be avoided to conserve the time of the Surgeons, as well as ease the cost for patients. Using a combination of numerical, ordinal, binomial and categorical variables we can create various logistic models to determine the severity of each tumor.

1.1 Definitions of Variables

BI-RADS Assessment

The acronym ”BI-RADS” stands for ”Breast Imaging Reporting and Data System” and is essentially a categorical-ordinal system for determining the findings from a mammogram screening. The system can vary but, for this data set the BI-RADS values range from $[1, 5]$, 1 being definitely benign and 5 being extremely likely of being malignant. This can be a very good determining factor of how well a CAD system compares to the radiologists.

Age

This variable is simply a numeric for the age of the patient.

Shape

For this data set the shape of the tumor will be defined categorical-nominally as the following:

round = 1, oval = 2, lobular = 3, irregular = 4

Margin

The Margin variable represents the mass margin of the tumor or distribution of the mass of the tumor, this variable is categorical-nominal with the following categories:

circumscribed = 1, microlobulated = 2, obscured = 3, ill-defined = 4, spiculated = 5

Density

Density variable determines the density of the tumor, this variable is categorical-ordinal.

high = 1, iso = 2, low = 3, fat-containing = 4

Severity

This variable is binomial with the following defined values:

benign = 0, malignant = 1

2 Analysis of Data

This data was obtained through the UCI Machine Learning repository and it has 961 observations or surveyed patients. Some of the variables contain a ”?” for the value so for the sake of this paper they will be ignored wherever found. For the analysis of this data the best model to use would be a Binomial Logistic regression. This is because the output of the Severity variable will be either a 0 or a 1. To predict goodness of fit for each model, we will use the Hosmer-Lemeshow Goodness of Fit test; The null hypothesis for this test is that there is no significant difference between the model created and the original data.

```
> library("ResourceSelection", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resourc
> mammographic_masses_data <- read.csv("~/Dropbox/Stats_2018_Maresca/Final_Paper/mammographi
> tumor_data <- na.omit(mammographic_masses_data)
> attach(tumor_data)
> M1 <- glm(Severity~as.numeric(BI.RADS_Assessment)+as.numeric(Age)+as.numeric(Shape)+as.num
> summary(M1)

Call:
glm(formula = Severity ~ as.numeric(BI.RADS_Assessment) + as.numeric(Age) +
    as.numeric(Shape) + as.numeric(Margin) + as.numeric(Density),
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6200  -0.5767  -0.2407   0.5937   3.8424

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -11.558190   0.931030 -12.414  < 2e-16 ***
as.numeric(BI.RADS_Assessment)   1.281357   0.163064   7.858 3.90e-15 ***
as.numeric(Age)                  0.044329   0.007064   6.276 3.48e-10 ***
as.numeric(Shape)                0.390401   0.084095   4.642 3.44e-06 ***
as.numeric(Margin)               0.431366   0.068081   6.336 2.36e-10 ***
as.numeric(Density)             -0.071318   0.103307  -0.690     0.49
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1326.98  on 960  degrees of freedom
Residual deviance:  798.62  on 955  degrees of freedom
AIC: 810.62

Number of Fisher Scoring iterations: 5

> hoslem.test(tumor_data$Severity,fitted(M1))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  tumor_data$Severity, fitted(M1)
X-squared = 5.6981, df = 8, p-value = 0.681

>
```

According to the summary of this model, almost all of the variables are significant in determining if a tumor is malignant or benign when they are treated as numerical variables. All except density, it appears density is not significant at all in this model so we will remove it to make a better model. Our Hosmer-Lemeshow test gives us a p-value of 0.681 which is a good sign, telling us that we can't reject the null that our model is significantly different than our data.

```
> library("ResourceSelection", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resourc
> M2 <- glm(Severity~as.numeric(BI.RADS_Assessment)+as.numeric(Age)+as.numeric(Shape)+as.num
> summary(M2)

Call:
glm(formula = Severity ~ as.numeric(BI.RADS_Assessment) + as.numeric(Age) +
    as.numeric(Shape) + as.numeric(Margin), family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6205  -0.5868  -0.2464   0.5897   3.8657

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -11.768714   0.883913 -13.314  < 2e-16 ***
as.numeric(BI.RADS_Assessment)   1.279586   0.163290   7.836 4.64e-15 ***
as.numeric(Age)                  0.044090   0.007048   6.256 3.95e-10 ***
as.numeric(Shape)                0.386605   0.083873   4.609 4.04e-06 ***
as.numeric(Margin)               0.426777   0.067776   6.297 3.04e-10 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1327.0  on 960  degrees of freedom
Residual deviance:  799.1  on 956  degrees of freedom
AIC: 809.1

Number of Fisher Scoring iterations: 5

> hoslem.test(tumor_data$Severity,fitted(M2))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  tumor_data$Severity, fitted(M2)
X-squared = 7.5963, df = 8, p-value = 0.4739
```

Because this is a Binomial Logistic model, it will take the form $ln(odds) = \beta_0 + \beta_1 x + \beta_2 x^2 + ...$

We can now solve for the odds that a tumor is benign based on the BI-RADS Assessment, Age, Shape, and Margin of the tumor.

$odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4}$
$odds = e^{-11.768714+1.279586*BI.RADS_{A}ssessment+0.044090*Age+0.386605*Shape+0.426777*Margin}$

$P(Y = Severity = 1|X = x_1, x_2, x_3, x_4) = \frac{e^{-11.768714+1.279586*BI.RADS_{A}ssessment+0.044090*Age+0.386605*Shape+0.426777*M}}{1+e^{-11.768714+1.279586*BI.RADS_{A}ssessment+0.044090*Age+0.386605*Shape+0.426777*}}$

The probability that a tumor is malignant makes sense because the slope for the BI-RADS Assessment is the largest slope for all the variables. This is because it is an ordinal variable which judges just how dangerous the tumor is which is exactly what this model is trying to do, but in a binomial fashion not a categorical fashion.

This model is a little bit better without the density playing a role in determining the severity of the tumor. This can be concluded by looking at the AIC value which goes from 810.92 to 809.1. Our model still does not reject the null given by the Hosmer-Lemeshow test.

Since we have created the best model for determining if a tumor is malignant or benign, let us see if we can predict the shape of the tumor based on the other variables.

```
> library("nnet", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/library")
> M3 <- multinom(Shape~as.numeric(BI.RADS_Assessment)+as.numeric(Age)+as.numeric(Margin)+as.

# weights:  35 (24 variable)
initial  value 1546.669834
iter  10 value 1043.067897
iter  20 value 985.223946
iter  30 value 970.742042
final  value 970.740581
converged

> summary(M3)

Call:
multinom(formula = Shape ~ as.numeric(BI.RADS_Assessment) + as.numeric(Age) +
    as.numeric(Margin) + as.numeric(Severity) + as.numeric(Density))

Coefficients:
  (Intercept) as.numeric(BI.RADS_Assessment) as.numeric(Age) as.numeric(Margin)
1    1.050640                      0.6216803    -0.021983853         -0.7633923
2    0.364269                      0.6710469    -0.019752880         -0.4117219
3   -4.842328                      1.0490475    -0.003676547         -0.1051291
4   -6.524150                      0.8614453    -0.008874169          0.5813881
  as.numeric(Severity) as.numeric(Density)
1           -0.5266550           0.2881659
2           -0.8924127           0.1484778
```

```
3           -0.1506901           0.3268852
4            0.8573925           0.5102535

Std. Errors:
  (Intercept) as.numeric(BI.RADS_Assessment) as.numeric(Age) as.numeric(Margin)
1    1.315260                      0.2166504      0.01602109          0.1484920
2    1.308818                      0.2133077      0.01595828          0.1409853
3    1.540759                      0.2529085      0.01703240          0.1500817
4    1.377535                      0.2051032      0.01608833          0.1443528
  as.numeric(Severity) as.numeric(Density)
1            0.4731833           0.1868111
2            0.4720365           0.1830840
3            0.4949076           0.2119912
4            0.4529484           0.1974924

Residual Deviance: 1941.481
AIC: 1989.481

> M4 <- multinom(Shape~as.numeric(BI.RADS_Assessment)+as.numeric(Age)+as.numeric(Margin)+as.

# weights:  30 (20 variable)
initial  value 1546.669834
iter  10 value 1035.258441
iter  20 value 976.090689
final  value 975.688105
converged

> summary(M4)

Call:
multinom(formula = Shape ~ as.numeric(BI.RADS_Assessment) + as.numeric(Age) +
    as.numeric(Margin) + as.numeric(Severity))

Coefficients:
  (Intercept) as.numeric(BI.RADS_Assessment) as.numeric(Age) as.numeric(Margin)
1   1.8487941                      0.6402731    -0.021357714        -0.72958135
2   0.7708478                      0.6750271    -0.018921108        -0.39318298
3  -3.8753184                      1.0607693    -0.002657267        -0.07413133
4  -4.9388814                      0.8786731    -0.006490288         0.61153098
  as.numeric(Severity)
1           -0.5693481
2           -0.9242965
3           -0.1898856
4            0.8084745

Std. Errors:
  (Intercept) as.numeric(BI.RADS_Assessment) as.numeric(Age) as.numeric(Margin)
```

```
1    1.220242                         0.2161833      0.01598834            0.1466075
2    1.213654                         0.2129479      0.01594581            0.1396341
3    1.413531                         0.2521459      0.01698962            0.1485503
4    1.233420                         0.2045796      0.01600518            0.1424681
  as.numeric(Severity)
1            0.4733921
2            0.4733720
3            0.4951505
4            0.4523189

Residual Deviance: 1951.376
AIC: 1991.376
```

According to our multinominal model of Shape as a function of the other variables, this is not a viable method to determine the shape of the tumor in question. This makes sense because one of the variables already involved some form of radiation scan, so using statistics wouldn't make much of a difference for the patient.

3 Conclusion

The objective of this paper was to research the relationship between Malignancy of a tumor and a variety of variables which some are obtained through mammographic screenings. For the first Binomial Logistic Regression Model used, all of the variables except density of the tumor were significant; This makes sense because comparitevly speaking the other variables are much more important, for example a BI-RADS Assessment of 5 would essentially guarentee a malignant tumor due to its large slope, this is the same for the Margin or distribution of the mass of the tumor considering it has the second largest slope in the model. Our "second" model made an attempt to predict the categorical shape of a tumor based on the Severity, BI-RADS Assessment, Margin, Age of the patient and Density. This model did not go as expected, the AIC values calculated using a combination of those variables were extremely high (lows of 1900 and highs of 2500). So we can make an inference that the Shape of the tumor is not that important or significant considering we cannot full-heartedly determine the Shape based on the other variables.

This work can be extremely important to those who are financially challenegd and have suspicion of a malignant tumor. This is because biopsies are pretty expensive and around 70% of biopsies are considered not neccessary, which can be a big problem for some people who can barely afford chemo therapy. This also allows Surgeons to spend their time on patients who are in dire need of a biopsy. The models created are by no means the "best" if we were to apply it to the work of others, but one if another data set is found with more variables and observances it will prove useful to compare.