

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 2013¹. 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 be 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 included attributes.

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

¹Australian Institute of Health and Welfare, Cancer compendium: information and trends by cancer type, <https://www.aihw.gov.au/reports/cancer/cancer-compedium-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$Severity <- factor(mammo$Severity)
summary(mammo$Severity, exclude = FALSE)
```

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

Severity is set to a factor \ Finally, BI-RADS has its 2 missing data points set to NA. Additioally, the data has a clear outlier in it that is set to NA aswell. 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

```

TODO: Ask if Age is forced to be a parameter

Can these variables be interpolated

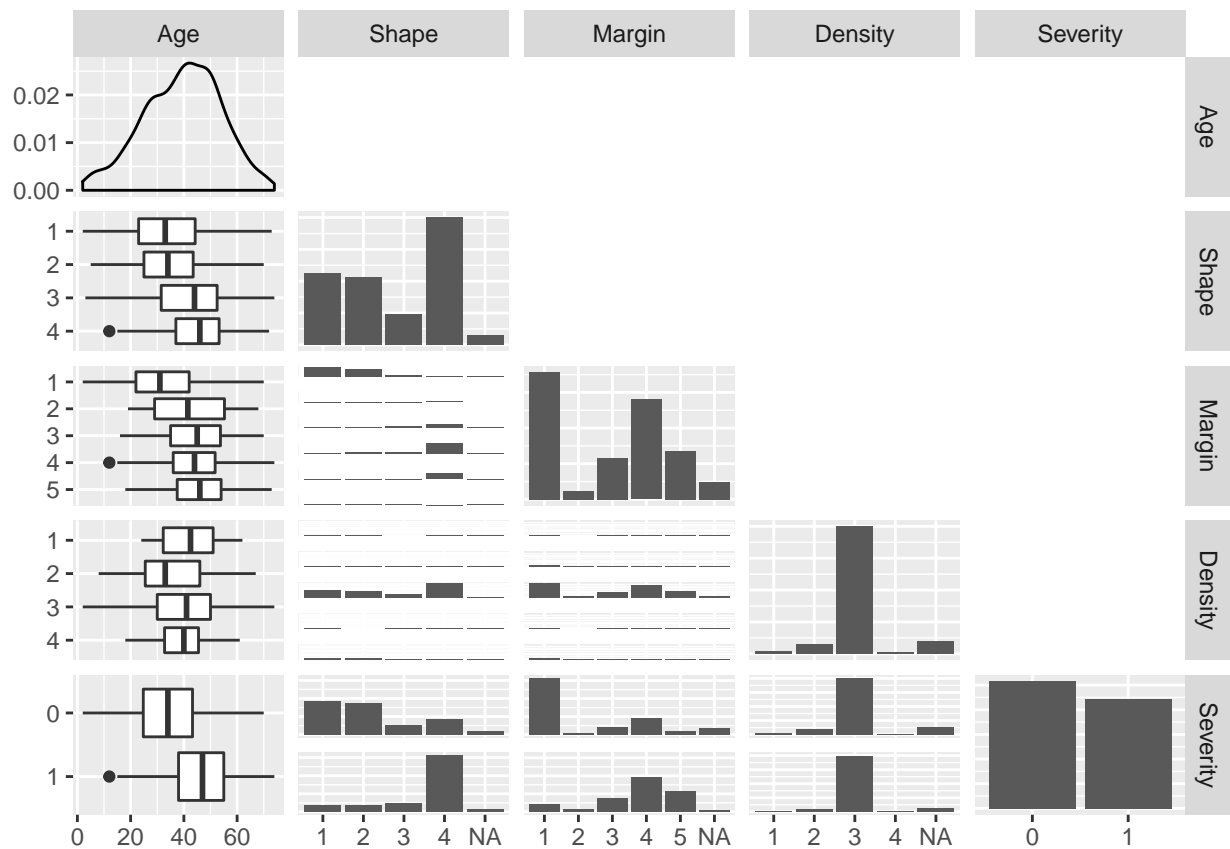
Data Visualisation

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"))

```



Now I'll talk about the relationships between variables but I'm sleepy and frankly that density is giving me the middle finger so I'll do it tomorrow.