Logistic Regression in Breast Cancer Patients

Carmen Canedo

11 August 2020

Objective

We will be looking at data collected on breast cancer patients in Wisconsin by Dr. William H. Wolberg from 1989-1991. The data observes the growths in 699 individuals, measuring variables such as clump thickness, size uniformity, marginal adhesion, cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and growth classification. In this report, I will be using a backwards selection process to create a logistic model to determine what explanatory variables most affect a benign or malignant diagnosis.

Loading necessary libraries

These are the necessary libraries you will need to have loaded to follow along with the code below.

```
library(tidyverse)
library(corrplot)
library(aod)
```

Loading in data

The data provided to me by Professor Kongoli of American University was initially provided as an Excel document. You will need to convert the file to a .csv to read it into R.

```
# Reading in .csv
breast_cancer_original <- read.csv("Breast_cancer_data_Wisconsin_Stat302_2020.csv")
```

Cleaning data

Before beginning the analysis, I will make some adjustments to the data. I am renaming the columns for clarity, changing their class, and modifying the values of the growth classification.

```
# Saving renamed columns into new data frame
breast_cancer <- breast_cancer_original %>%
  rename(clump_thickness = ClumpThick) %>%
  rename(size_uniformity = SizeUniforrm) %>%
  rename(shape_uniformity = shapeUniform) %>%
  rename(marginal_adhesion = Adhesion) %>%
  rename(cell_size = cell.cize) %>%
```

```
  rename(bare_nuclei = bareNuclei) %>%
  rename(bland_cromatin = Cromatin) %>%
  rename(normal_nucleoli = normalNuclei) %>%
  rename(mitoses = Mitozis) %>%
  rename(growth_class = Btype)
```

Normally, we would need to covert growth_class to a factor because its two values, 2 and 4, represent benign and malignant growths respectively. However, later in the code the `cor()` function requires that all columns are numeric. We will make those changes now, but it is imperative to keep in mind that calculations should not be done on these values.

```
# Checking class
sapply(breast_cancer, class)
```

```
##                ID   clump_thickness   size_uniformity  shape_uniformity
##         "integer"         "integer"         "integer"         "integer"
## marginal_adhesion         cell_size       bare_nuclei    bland_cromatin
##         "integer"         "integer"       "character"         "integer"
##   normal_nucleoli           mitoses      growth_class
##         "integer"         "integer"         "integer"
```

```
# List of column names
names <- c("ID", "clump_thickness", "size_uniformity", "shape_uniformity", "marginal_adhesion", "cell_s

# Changing type to numeric using sapply and list of column names
breast_cancer[names] <- sapply(breast_cancer[names], as.numeric)

# Checking to make sure answer is correct
sapply(breast_cancer, class)
```

```
##                ID   clump_thickness   size_uniformity  shape_uniformity
##         "numeric"         "numeric"         "numeric"         "numeric"
## marginal_adhesion         cell_size       bare_nuclei    bland_cromatin
##         "numeric"         "numeric"         "numeric"         "numeric"
##   normal_nucleoli           mitoses      growth_class
##         "numeric"         "numeric"         "numeric"
```

Lastly, we need to re-code 2 and 4 to 0 and 1, so that the program will be able to run its regression on the data.

```
# Changing to 0 and 1
breast_cancer <- breast_cancer  %>%
  mutate(growth = case_when(growth_class == 2 ~ 0, growth_class == 4 ~ 1)) %>%
  select(-growth_class) %>%
  rename(growth_class = growth)
```

Analysis

Correlation

Let's take a look at the correlation matrix to get a general idea of the patterns in the data.

```r
# Computing correlation for breast_cancer data set
bc_correlations <- cor(breast_cancer)

# Calling object
bc_correlations
```

```
##                            ID clump_thickness size_uniformity shape_uniformity
## ID                 1.00000000     -0.05530844     -0.04160334      -0.04157607
## clump_thickness   -0.05530844      1.00000000      0.64491250       0.65458908
## size_uniformity   -0.04160334      0.64491250      1.00000000       0.90688191
## shape_uniformity  -0.04157607      0.65458908      0.90688191       1.00000000
## marginal_adhesion -0.06487808      0.48635624      0.70558181       0.68307920
## cell_size         -0.04552828      0.52181622      0.75179913       0.71966844
## bare_nuclei               NA              NA              NA               NA
## bland_cromatin    -0.06005053      0.55842816      0.75572098       0.73594845
## normal_nucleoli   -0.05207195      0.53583455      0.72286482       0.71944632
## mitoses           -0.03490066      0.35003386      0.45869315       0.43891093
## growth_class      -0.08022565      0.71600136      0.81790374       0.81893374
##                   marginal_adhesion   cell_size bare_nuclei bland_cromatin
## ID                      -0.06487808 -0.04552828          NA    -0.06005053
## clump_thickness          0.48635624  0.52181622          NA     0.55842816
## size_uniformity          0.70558181  0.75179913          NA     0.75572098
## shape_uniformity         0.68307920  0.71966844          NA     0.73594845
## marginal_adhesion        1.00000000  0.59959907          NA     0.66671533
## cell_size                0.59959907  1.00000000          NA     0.61610184
## bare_nuclei                      NA          NA           1             NA
## bland_cromatin           0.66671533  0.61610184          NA     1.00000000
## normal_nucleoli          0.60335241  0.62888069          NA     0.66587781
## mitoses                  0.41763278  0.47910148          NA     0.34416950
## growth_class             0.69680021  0.68278453          NA     0.75661615
##                   normal_nucleoli     mitoses growth_class
## ID                    -0.05207195 -0.03490066  -0.08022565
## clump_thickness        0.53583455  0.35003386   0.71600136
## size_uniformity        0.72286482  0.45869315   0.81790374
## shape_uniformity       0.71944632  0.43891093   0.81893374
## marginal_adhesion      0.60335241  0.41763278   0.69680021
## cell_size              0.62888069  0.47910148   0.68278453
## bare_nuclei                    NA          NA           NA
## bland_cromatin         0.66587781  0.34416950   0.75661615
## normal_nucleoli        1.00000000  0.42833575   0.71224362
## mitoses                0.42833575  1.00000000   0.42317026
## growth_class           0.71224362  0.42317026   1.00000000
```

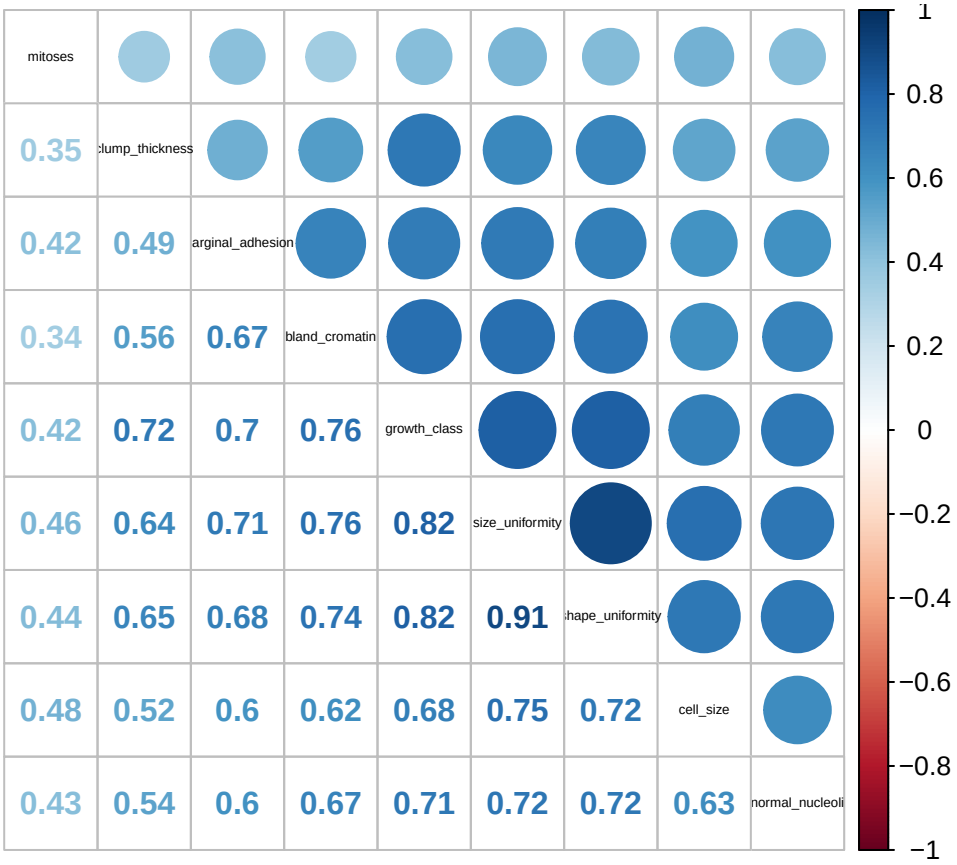
```r
# Getting rid of missing values and ID
bc_corr <- breast_cancer %>%
  select(-bare_nuclei, -ID)

# Final correlation matrix
bc_corr <- cor(bc_corr)
bc_corr
```

```
##                   clump_thickness size_uniformity shape_uniformity
## clump_thickness         1.0000000       0.6449125        0.6545891
```

```
## size_uniformity         0.6449125       1.0000000        0.9068819
## shape_uniformity        0.6545891       0.9068819        1.0000000
## marginal_adhesion       0.4863562       0.7055818        0.6830792
## cell_size               0.5218162       0.7517991        0.7196684
## bland_cromatin          0.5584282       0.7557210        0.7359485
## normal_nucleoli         0.5358345       0.7228648        0.7194463
## mitoses                 0.3500339       0.4586931        0.4389109
## growth_class            0.7160014       0.8179037        0.8189337
##                    marginal_adhesion cell_size bland_cromatin normal_nucleoli
## clump_thickness            0.4863562 0.5218162      0.5584282       0.5358345
## size_uniformity            0.7055818 0.7517991      0.7557210       0.7228648
## shape_uniformity           0.6830792 0.7196684      0.7359485       0.7194463
## marginal_adhesion          1.0000000 0.5995991      0.6667153       0.6033524
## cell_size                  0.5995991 1.0000000      0.6161018       0.6288807
## bland_cromatin             0.6667153 0.6161018      1.0000000       0.6658778
## normal_nucleoli            0.6033524 0.6288807      0.6658778       1.0000000
## mitoses                    0.4176328 0.4791015      0.3441695       0.4283357
## growth_class               0.6968002 0.6827845      0.7566161       0.7122436
##                     mitoses growth_class
## clump_thickness   0.3500339    0.7160014
## size_uniformity   0.4586931    0.8179037
## shape_uniformity  0.4389109    0.8189337
## marginal_adhesion 0.4176328    0.6968002
## cell_size         0.4791015    0.6827845
## bland_cromatin    0.3441695    0.7566161
## normal_nucleoli   0.4283357    0.7122436
## mitoses           1.0000000    0.4231703
## growth_class      0.4231703    1.0000000
```

This matrix can be seen in its visual representation below.

```
# Visualizing correlation matrix
corrplot.mixed(bc_corr,
         order = "hclust",
         tl.cex = .4,
         tl.col = "black")
```



Interpretation

From this visualization, we can easily see that all combinations of variables have positive associations. The strongest correlations (which in this instance we are determining as greater than or equal to 0.75) are between the following:

- Size uniformity and cell size: 0.75
- Bland chromatin and growth classification: 0.76
- Bland chromatin and size uniformity: 0.76
- Growth classification and size uniformity: 0.82
- Growth classification and shape uniformity: 0.82
- Size uniformity and shape uniformity: 0.91

Logistic model

We are going to use the backwards method to create our logistic model. We will first start with all potential explanatory variables and perform regression on them.

```
# Creating model using all variables to determine growth_class
logistic_1 <- glm(growth_class ~ ., data = breast_cancer, family = "binomial")

# Results of logistic model
summary(logistic_1)
```

```
##
## Call:
## glm(formula = growth_class ~ ., family = "binomial", data = breast_cancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4877  -0.1156  -0.0613   0.0223   2.4668
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.015e+01  1.454e+00  -6.983 2.89e-12 ***
## ID                 4.008e-08  7.331e-07   0.055 0.956396
## clump_thickness    5.349e-01  1.420e-01   3.767 0.000165 ***
## size_uniformity   -6.818e-03  2.091e-01  -0.033 0.973995
## shape_uniformity   3.234e-01  2.307e-01   1.402 0.161010
## marginal_adhesion  3.306e-01  1.234e-01   2.679 0.007389 **
## cell_size          9.624e-02  1.567e-01   0.614 0.539233
## bare_nuclei        3.840e-01  9.546e-02   4.022 5.77e-05 ***
## bland_cromatin     4.477e-01  1.716e-01   2.608 0.009099 **
## normal_nucleoli    2.134e-01  1.131e-01   1.887 0.059224 .
## mitoses            5.344e-01  3.294e-01   1.622 0.104740
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 102.89  on 672  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 124.89
##
## Number of Fisher Scoring iterations: 8
```

Now we will exclude the variables that are not statistically significant at a 95% significance level. We will now remove the following: ID, size_uniformity, shape_uniformity, cell_size, normal_nucleoli, and mitoses.

```r
# Removing variables that are not statistically significant
reduced_bc <- breast_cancer %>%
  select(-ID, -size_uniformity, -shape_uniformity, -cell_size, -normal_nucleoli, -mitoses)

# Creating model using remaining variables to determining growth_class
logistic_2 <- glm(growth_class ~ ., data = reduced_bc, family = "binomial")

# Results of logistic model
summary(logistic_2)
```

```
##
## Call:
## glm(formula = growth_class ~ ., family = "binomial", data = reduced_bc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6964  -0.1451  -0.0609   0.0232   2.4476
```

```
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -10.11370    1.03264  -9.794  < 2e-16 ***
## clump_thickness     0.81166    0.12585   6.450 1.12e-10 ***
## marginal_adhesion   0.43412    0.11403   3.807 0.000141 ***
## bare_nuclei         0.48136    0.08816   5.460 4.76e-08 ***
## bland_cromatin      0.70154    0.15196   4.616 3.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 125.77  on 678  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 135.77
##
## Number of Fisher Scoring iterations: 8
```

```
# Exponentiated coefficients
exp(logistic_2$coefficients)
```

```
##       (Intercept)   clump_thickness marginal_adhesion       bare_nuclei
##      4.052058e-05      2.251649e+00      1.543606e+00      1.618270e+00
##    bland_cromatin
##      2.016859e+00
```

```
# Wald Test
wald.test(b = coef(logistic_2), Sigma = vcov(logistic_2), Terms = 3:4)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 48.9, df = 2, P(> X2) = 2.4e-11
```

All of our variables are now statistically significant at 0.05. We can move forward to the interpretation of the model.

Interpretation

Using the information above, we can infer which variables have an effect on whether breast cancer patients in Wisconsin are likely to have benign versus malignant growths.

The values on the y-axis of logistic regression are limited from 0 to 1 in order to calculate the probability. We can transform this to the log odds of cancerous growth to increase the domain of the y-axis from negative infinity to positive infinity. We will later visualize the logistic model using a domain of 0 to 1, but the coefficients are determined using logg odds.

Our model for the growth classification of breast cancer patients in Wisconsin from 1989-1991 is as follows:

growthclassification = -10.1137 + 0.81166(clumpthickness) + 0.43412(marginaladhesion) + 0.48136(barenuclei) + 0.70154(blandcromatin)

Based on our Wald test with 2 degrees of freedom, the p-values indicate that the associations are statistically significant at alpha less than 0.05 for clump thickness, magrinal adhesion, bare nuclei, and bland chromatin.

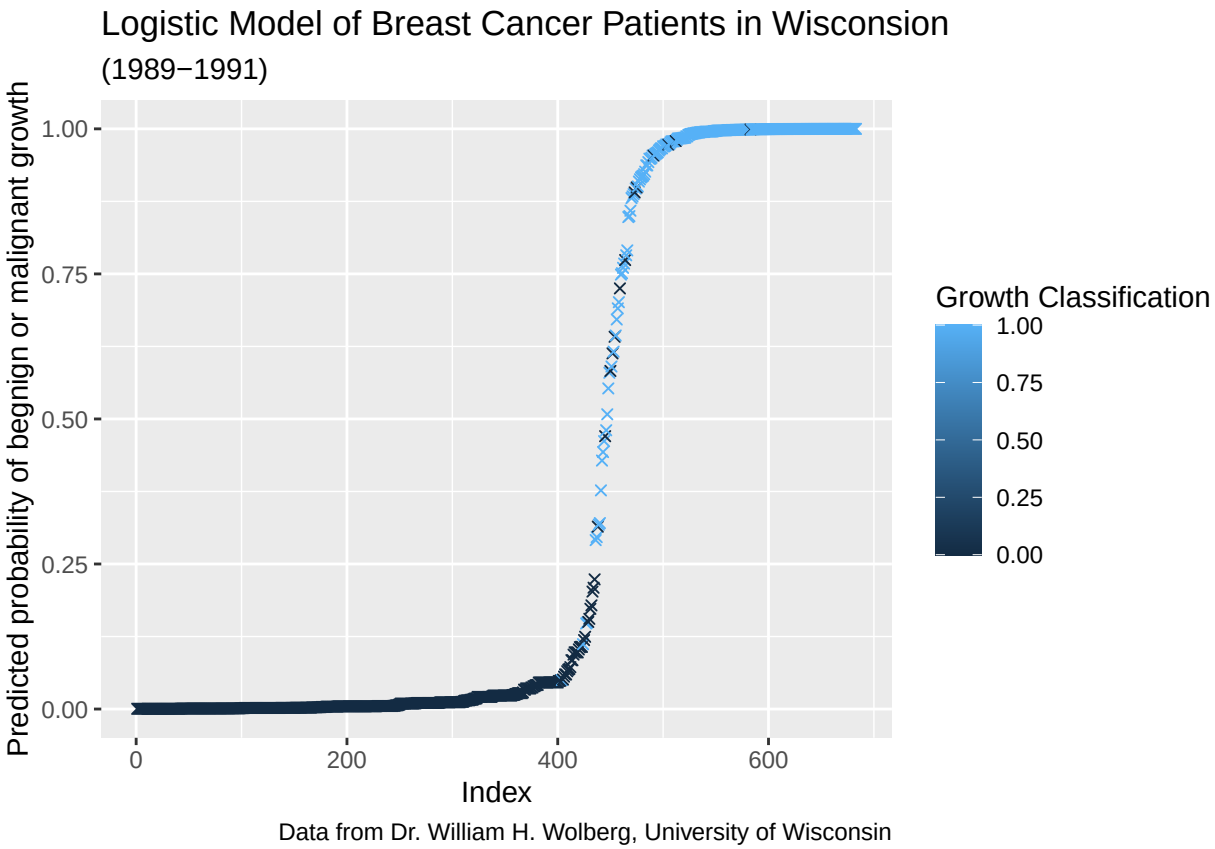
Graphing

In order to graph the probabilities of the model above, we need to create a new data frame that has the probabilities.

```
# Creating new data frame of probabilities
predicted_data <- data.frame(probability_of_growth_class = logistic_2[["fitted.values"]], growth_class =

# Sorting from low to high
predicted_data <- predicted_data[order(predicted_data$probability_of_growth_class, decreasing = FALSE),]

# Ranking probability from low to high
predicted_data$rank <- 1:nrow(predicted_data)
```

Creating graph

```
graphed_model <- ggplot(data = predicted_data, aes(x = rank, y = probability_of_growth_class)) +
  geom_point(aes(color = growth_class.growth_class), alpha = 1, shape = 4, stroke = 0.5) +
  theme_set(theme_linedraw()) +
  labs(title = "Logistic Model of Breast Cancer Patients in Wisconsion",
       subtitle = "(1989-1991)",
       caption = "Data from Dr. William H. Wolberg, University of Wisconsin",
       x = "Index",
       y = "Predicted probability of begnign or malignant growth",
       color = "Growth Classification")

graphed_model
```



Conclusion

Now that we have performed our logistic regression, we can see that clump thickness, marginal adhesion, bare nuclei, and bland chromatin have the greatest effect on the outcome of a benign or malignant diagnosis.