Predicting Breast Cancer II - Logistic Regression Model

Logistic regression is one of the mainstays in terms of predicting binary outcomes (in this case: benign or malignant tumor cells). Besides, the model features allows us to get a deeper understanding of which factors play more risks to patients.

Let's see this example, with data from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumours for 699 patients up to 15 July 1992. (R Documentation, 2019)

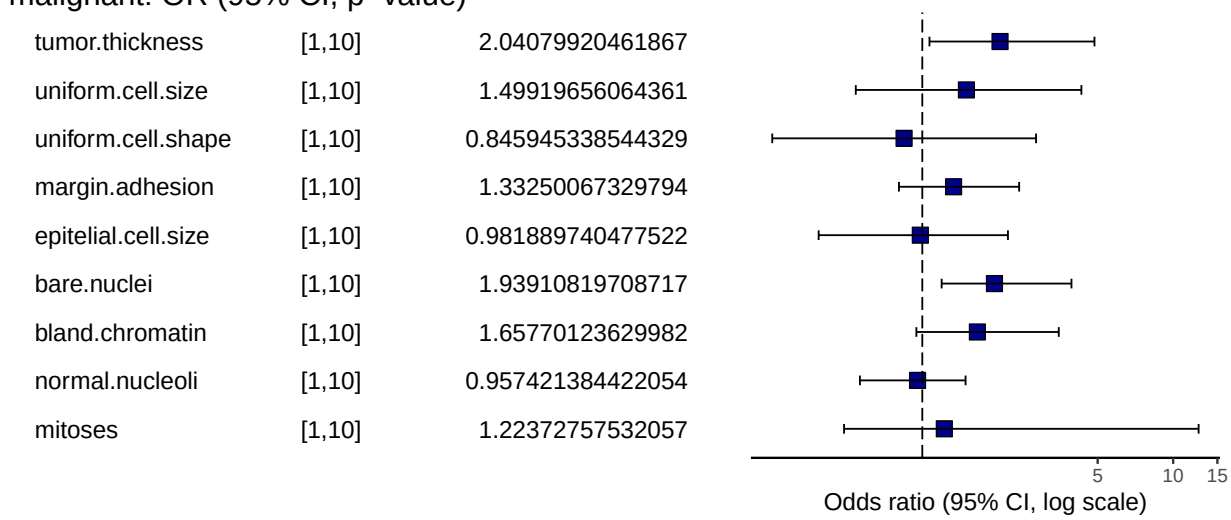
```
##
## Call:
## glm(formula = malignant ~ tumor.thickness + uniform.cell.size +
##     uniform.cell.shape + margin.adhesion + epitelial.cell.size +
##     bare.nuclei + bland.chromatin + normal.nucleoli + mitoses,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.70846  -0.11571  -0.05802   0.02165   2.61223
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -10.07583    1.29057  -7.807 5.84e-15 ***
## tumor.thickness       0.52431    0.15135   3.464 0.000532 ***
## uniform.cell.size    -0.04453    0.23464  -0.190 0.849470
## uniform.cell.shape    0.40374    0.24423   1.653 0.098302 .
## margin.adhesion       0.18574    0.14076   1.320 0.186986
## epitelial.cell.size   0.15132    0.17296   0.875 0.381623
## bare.nuclei           0.39692    0.10493   3.783 0.000155 ***
## bland.chromatin       0.33168    0.19411   1.709 0.087505 .
## normal.nucleoli       0.24692    0.12932   1.909 0.056224 .
## mitoses               0.63762    0.31679   2.013 0.044143 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 706.912  on 558  degrees of freedom
## Residual deviance:  80.291  on 549  degrees of freedom
## AIC: 100.29
##
## Number of Fisher Scoring iterations: 8
```

Interpreting results:

Multivariate Analysis shows a p-value lower than 0.05 to the features:

- tumor.thickness
- bare.nuclei
- mitoses

Predicting Breast Cancer - Odds Ratio Analysis



Testing Model Parameters:

```
# TESTING & EVALUATING MODEL:

# testing predictions on the model:
logit_pred <- as.factor(round(predict(mylogit,test,type='response'),0))

#building a confusion matrix
logit_confMat <- table(test$malignant,logit_pred)

#building a confusion matrix
print('Confusion Matrix:')
```

```
## [1] "Confusion Matrix:"
```

```
print(logit_confMat)
```

```
##    logit_pred
##      0  1
##   0 79  3
##   1  4 54
```

```
# getting model accuracy:
logit_accuracy <- sum(diag(logit_confMat))/sum(logit_confMat)
print('model accuracy (in %):')
```

```
## [1] "model accuracy (in %):"
```

```r
print(round(logit_accuracy,3))
```

```
## [1] 0.95
```

As we see in the Odds plot, the only variables who are statistically significant on increasing OR for malignancy and tumor thickness and bare nuclei.

Model performance

We have a confusion matrix which shows that our model got 3 false positives and 4 false negative results, for a total of 140 testing samples.

In terms of success metrics, it means that the model has an accuracy of 95%.

Observations:

As we ran simulations for the model, we noticed that the p-values for regression in the train dataset changed in ways of having different significative variables in each random sampling. This means that probably the sample size needs to increase, to confirm that this model can be generalized in scale.

Comparing with the results from our Decision Tree Model (https://git.io/fh7qR), we see a convergence of importance of bare nuclei to determine tumor malignancy.

Final recommendations for improvements:

- Drop variables which are statistically insignificant for the model, to reduce "noise"
- increase sample size, focusing on generalize the model fit

Find the source code for this report at https://git.io/fh7qR