STA 223 Project 1: Breast Cancer Diagnosis Modeling

Ana Boeriu
MS Biostatistics

Introduction

Breast cancer is a type of cancer that causes uncontrollable growth of abnormal breast cells that form a mass called a tumor. Because of genetic mutations, cancer cells have different sizes and shapes compared to healthy cells, thus are unable to function properly

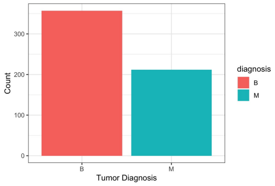
- The main goal of this project is to build a model and find the most significant variables when determining tumor diagnosis.

Data Description and Preprocessing

Data: Breast Cancer Wisconsin (Diagnostic) Data Set

This dataset consists of the tumor diagnosis and the characteristics of cell nuclei, computed from 569 digitized images of a fine needle aspirate (FNA) breast mass

- Assume a breast mass can have more than one nucleus
- No missing values
- Response variable: tumor diagnosis (M, B)



- Total of 357 benign tumors and 212 malignant tumors.

- 10 numeric predictors that describe the size and shape of the cell nucleus: area, radius, perimeter, fractal dimension, concavity, , smoothness, symmetry, compactness, concave points, texture

- Mean, standard error, worst(mean of largest three values) also computed for reach predictor resulting in 30 total predictors.
- Issue of collinearity between predictors.
  - Removed all columns of "standard error" and "worst"
  - Removed predictors of texture, area, perimeter, concave points, compactness
- Remaining variables:
  - diagnosis, radius mean, smoothness mean, symmetry mean, fractal dimension mean, concavity mean

Methods

Since tumor diagnosis is categorical, we will use logistic regression.

- Linear predictor: $\eta = X\beta$
- Random component: $nY \sim Bin(n, \pi)$
- Link function: $\eta = log\left(\frac{\mu}{1-\mu}\right)$

Before proceeding to model selection, we performed an overall regression test where we test the following:

- $H_o: \beta_p = 0$ versus $H_A: \beta_p \neq 0$ where p denotes the number of coefficients in the model

We proceed to model selection where we choose BIC as our criterion and forwards backwards as our stepwise selection

- BIC(p) = D(p) + p*log(n)

Next, we check for leverage and influential points.

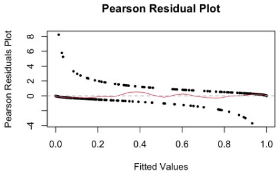
- If $h_{ii} > \frac{2p}{n}$, where n is the sample size, then the observation is a suspected leverage point.
- observations with high Cook's distance are influential points and may need to be deleted.

Overall Regression Test & Model Diagnostics

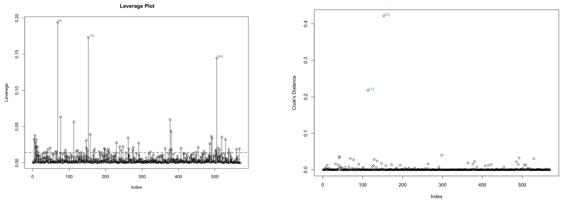
Overall Regression Test:

- $H_o: \beta_{radius\ mean} = \beta_{smoothness\ mean} = \beta_{symmetry\ mean} = \beta_{fractal\ dimention\ mean} = \beta_{concavity\ mean} = 0$
- $H_a: at\ least\ one\ \beta_i \neq 0$
- With a p-value of less than 2.2×10^{-16} and at any significant alpha, we conclude that some predictors have an overall significant effect with tumor diagnosis

Model fits the data well and there is no obvious sign of lack of fit



We found that even after removing these points and performing the same type of analysis as in part A there was no significant difference in our p-values. Thus, there is no need to remove these points



Results

Our final model with main effects is

$$\eta = -23.783 + 1.068X_{radius\ mean} + 20.284X_{concavity\ mean} + 64.89X_{smoothness\ mean}$$

β_i	Estimate	P-value
Intercept	−23.7834	$< 2.2 \times 10^{-16}$
Radius mean	1.0676	$< 2.2 \times 10^{-16}$
Concavity mean	20.2841	3.24×10^{-8}
Smoothness mean	64.8905	4.07×10^{-5}

- The predictor radius mean is the most significant predictor
- All predictors have a positive effect in determining tumor diagnosis
- As the average radius, average concavity, average smoothness increase in the cell nuclei the tumor is more likely to be malignant.
- As will all shape variables a high smoothness values corresponds to less regular contour and thus higher probability of malignancy.

Acknowledgements

I would like to thank Professor Müller, Poorbita Kundu, and Han Chen for their support and guidance throughout this project

References

Breast Cancer Wisconsin (Diagnostic) Data Set. (2016, September 25). Kaggle.
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1992). Nuclear Feature Extraction for Breast Tumor Diagnosis. *International Symposium on Electronic Imaging: Science and Technology*, *1905*, 861–870.
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.707&rep=rep1&type=pdf