[[1]](#footnote-1)

Breast Cancer Prediction – Benign or Malignant(March 2019)

Martin Garcia, marting@smu.edu, Andrew L. Wilkins, awilkins@mail.smu.edu

*Abstract*—Data was collected from the UCI Machine Learning Respository for Breast Cancer. The data includes multiple explanatory variables of the cell being observed and a binary response variable of malignant or benign.

*Index Terms*—Cancer, Breast Cancer, malignant, benign, classification model, healthcare, supervised classification, discriminant analysis, principal component analysis, binary response

# Introduction

T

HE goal of the project is to accurately predict whether a cancer is malignant or benign based on 30 different explanatory variables. Malignancy is a general term for a cell that divides uncontrollably and spreads. These rogue cells have various names that are dependent on where they form, various examples include Carcinoma for malignancy that starts in the skin. Sarcoma begins in bones, cartilage, fat, blood vessels. Our breast cancer is known mainly as carcinoma. In contrast benign cells don’t have the ability to spread to their neighbors.

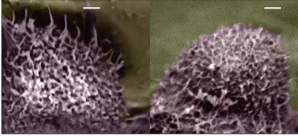


Figure - Can you guess which cell is malignant? The images above were captured with an atomic force microscope capable of capture images a fraction of a nanometer in size.

# Data Description

# Our dataset was pulled from the UCI Machine Learning Repository. This data was collected from 1989 to 1991 at University of Wisconsin Hospitals, Madison. Our explanatory variables are measurements taken from an image of cell nuclei of breast mass samples. This includes 569 observations\* paired with continuous variables that will help us predict a binary response variable, benign or malignant as 1 or 0 respectively. This includes physical measurements such as area and perimeter. Observations are divided into 357 benign and 212 malignant cases.

# Exploratory data analysis

The individual explanatory variables are observed to help meet assumptions for testing. This translates to looking at distributions, variance, independence assumptions, outliers, and linearity assuming linear model. The required assumptions are based on our test we use for our analysis; this includes assumptions required for PCA and logistic regression model addressed respectively.

## Distribution

Data sets were not normally distributed and showed skewed results for several explanatory variables. This was addressed with taking the log of these variables. Pairs of our variables also show elliptical shapes, however our observations is robust against violations as it is greater than 25.

However this only addresses normally assumptions for our PCA. Logistic regressions require univariate normal distribution for our response and multivariate normal distribution for our explanatory variables.

## Variance

Variances for 2 variables were exponentially larger for the area mean and area worst.

## Independence

????

## Linearity

????

Retrospective or observational? If retrospective must us odds with logistic regression. If observational, safe to use probability.

# Statistical Analysis

## Principal Component Analysis

With 30 different variables we reduce the count for those that account for the most variability using principal component analysis. The approach will depend on the variance of all explanatory variables which will tell us if we will approach this using a covariance matrix or correlation matrix. The covariance matrix includes our variables as is, and correlation will standardize our values if variances are substantially different from each other. Doing an analysis of the variables independently will also help us see if standardization will be required. Standardization will help us if variance among several variables is higher, by standardizing we can value all our variables proportionally. This is important because PCA will emphasize variables with high variance and we want to captures variables that account for the overall large proportion of variance. To see variance we start with summary statistics of all variables and check if any show large cumulative variation compared to the rest or stand out as outliers in variance. Knowing if there are variables that are have exponentially larger variables will help us determine if we will run our PCA on a correlation or covariance matrix. Two outliers in variance are present in area mean and area worst variables and suggest a correlation matrix which will require standardization for all variables as they are exponentially larger.

We tried 3 different approaches, using a log transformation on both correlation with scaling, covariance matrix and correlation matrix without log transformation with variance transformed with scaling. Transforming our data not only normalized our variables but also helps us compare variance differences in absolute terms and we take our data and scale further for more accurate comparison of variance for our explanatory variables. Based on this we believe doubling of our scaling through log and scaling method provided by correlation matrix was our best approach.

The first six principle components help account for 90% of the variance. During each iteration of the principle component a combination of our variables that accounts for the most amount of variance is consolidated as a new variable or principle component.

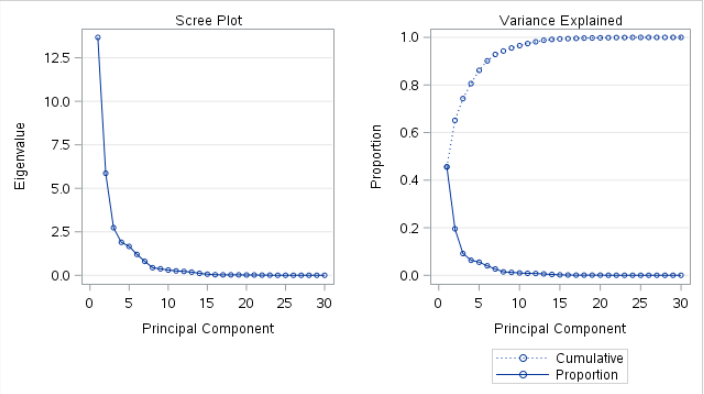


Figure 2 - Screen Plot suggest most variability is accounted for before the 10th principical component

Total variance is reflected in our eigen values and eigen vectors function as coefficients for our linear combinations of explanatory variables. New variables are created from the original minus the original variance to be measured again to find variables that account for most variance. The highest loadings are around .2 per variable.

## Logistic Regression

The response for our case study is categorical or binary and has two finite choices of malignant or benign. To accommodate for multiple explanatory we use a correlation matrix to find what effects lead to malignant cancer.

* Interpret coefficients
* Hosmer Lemeshow Test

Our classification table produced the highest percentage correctly guessed at 97.1% at probability level 42%. The 97% is translated should we should this probability level if exceeded is dumped into the correct bucket, or malignant in this example. Sensitivity at 96.7% translates into our model predicting malignancy correctly when our cell is really malignant. Opposite of sensitivity or model accurately predicts if a cell is really benign with 97.4% accuracy.

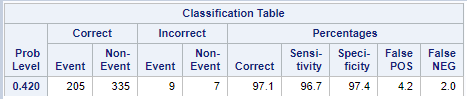
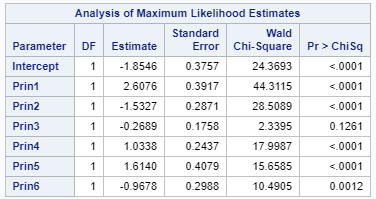
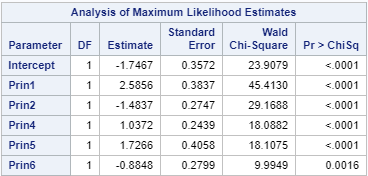


Figure - Classification table at 42% probability level





# RESOURCES NEEDED

In order to complete this project we will utilize SAS for statistics measures and python for organizing our data to create visualizations reflective of our findings.

# REFERENCES

[1] <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/activity>

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] <https://www.cancer.org/>

[4] <https://www.nature.com/articles/nnano.2009.77>

See Explanatory\_Variables.txt for more detail regarding explanatory variables\*

1. [↑](#footnote-ref-1)