[[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 case study 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 the location they form; various examples include Carcinoma for malignancy that starts in the skin and Sarcoma which begins in bones, cartilage, fat, blood vessels. Breast cancer is usually categorized as a carcinoma. In contrast benign cells don’t have the ability to spread to their neighbors.

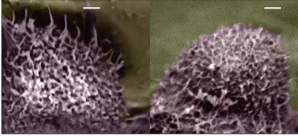


Figure - Can you guess which cell is malignant? The images above were captured with an atomic force microscope capable of capture images a fraction of a nanometer in size.

# 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.

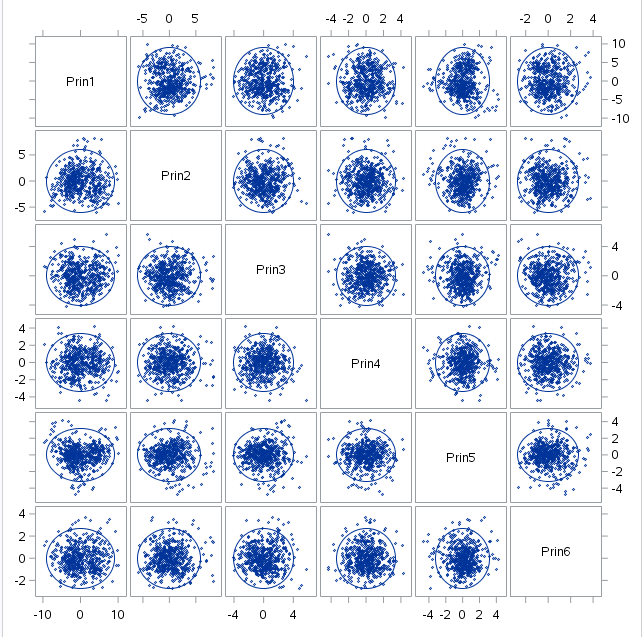
# Exploratory data analysis

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 principal component analysis and logistic regression model addressed due to a binary response variable.

## Distribution

Data sets were not individually 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. Our observations are robust against violations of normality due to the number of observations 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. Our output was circles, and not the ellipses we expected. Our response variable in univariate normal distribution would result in an ‘S’ shaped distribution. This was addresses with using log through our logistic regression.



## Variance

Variances for 2 variables were exponentially larger for the area mean and area worst. However after taking the log and adjusting with scaling through PCA our variances no longer show large absolute differences.

Variance for logistic regression, requires uniform error variance for response across all values of x.

## Independence

Independence for PCA

Independence for logistic regression

## Linearity

Linearity for PCA – relationships between all observed variables should be linear

Linearity of logistic regression – Linear relationship between scores on Y and scores on X for all variables, otherwise express relationships in terms of odds ratios

The study is retrospective and requires us to use odds ratios. This benefits ours study as it helps address assumptions not originally met.

# 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 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.

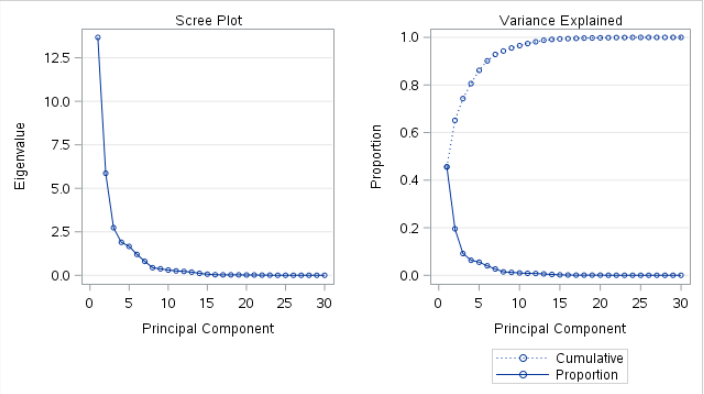


Figure 2 - Screen Plot suggest most variability is accounted for before the 10th principical component

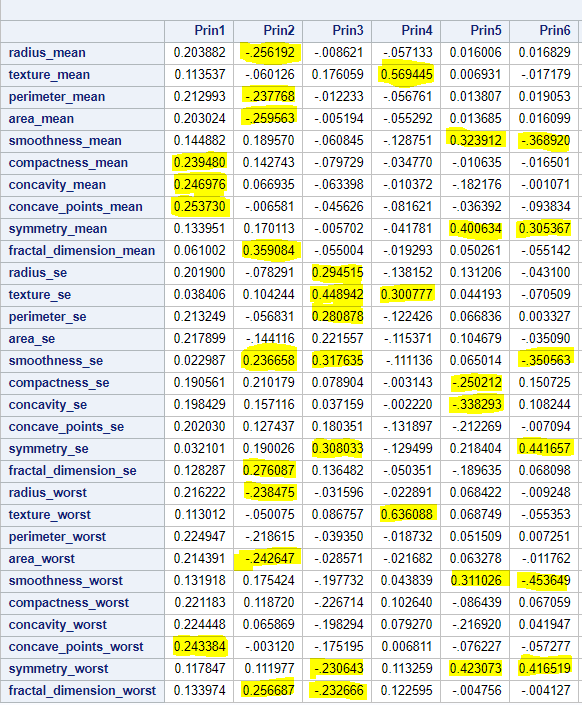
Total variance is reflected in our eigen values and eigen vectors function as coefficients for our linear combinations of explanatory variables. New variables are created from the original minus the original variance to be measured again to find variables that account for most variance. The highest loadings are around .2 per variable. Loadings help us the determine the level of impact per a particular principle component. Loadings ranged from .2 to .6 from principle component 1 to 6.

*Translate the principal components*

Translating loadings in relation to principle components in relation to a linear regression are better demonstrated using the principle components and their respective loadings.

As an example principle component 1, or think of as variable 1 had no one explanatory variable account for the variability. There was a measure an average measure of those whom did stand out at .2 for the variables below. Meaning these explanatory variables for our first principle component help explain the variability in our response. However because we have a binary response variable we change the variability in response to variability in probability of our even happening, or the falling under the category of malignant cell.

Variables that help explain the variance in probability of our model include compactness mean, concavity mean, concave points mean, and concave points worst. If translated for only 1 principle component this would mean our positive loadings increase the probability of a cell being malignant. We would then look at those 4 variables with high loadings as those mainly responsible for the variability in our response, or probability. We would place our attention at compactness of the cell, mean of concave portions of our cell, mean number of concave points, and most average number of concave points.



## 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.

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. However this applies to only an observational model and must be applied in terms of odds ratios. As a retrospective study the subjects were identified first and looked thereafter at the characteristics of cells. Find uniform error variances for response Y across all value of X, if not normally distributed, and have nonconstant variance, transform.

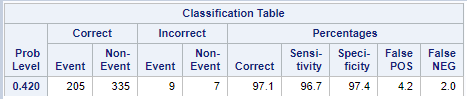
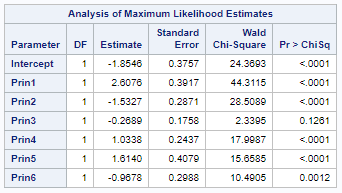


Figure - Classification table at 42% probability level

We use method of maximum likelihood in place of ordinary least squares. Instead of determining a linear relationship of our data points based in a minimum of our residuals we determine our coefficients for logistic regression based on those that contribute to the largest probability of malignant cancer. Results are all are statistically significant with the exception of principle component 3. We use the Wald Chi-Square as our test statistic.



Our odds ratio estimates show the increase odds of malignant cancer cells for each principle component. Significant increase in odds is noticed for the 1st principle component, indicating a one unit change in Prin1 increases the odds of malignant cells by 13.

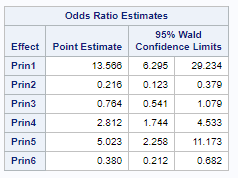


Figure - Odds ratio for malignant cell

In contrast, we can take the odds ratio estimates of the cell being benign with a reversal of our event. This translates similarly to the previous odds ratio estimate.

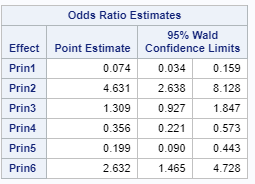


Figure - Odds ratio for benign cells

Using both these odds ratios we can compare the odds between both malignant and benign cells between principle components holding all other components constant. We use principle component 1 between both responses to calculate the probability of malignancy while holding the rest. This incorporates the ratio of odds below. B1 is our estimated coefficient for principle component 1.

Wa / Wb = exp(b1 (A-B))

13=exp(2.6)

This translates to the odds of our cell being malignant will to be 13 times the odds of the cell being benign under principle component one. Remembering that the loadings for this prin1 are compactness mean, concavity mean, concave points mean, and concave points worst. This does not discount other variables only emphasizes those that are most impactful.

Figure - Odds ratio for benign cell

Add effect plots

# Conclusion

Inferences are limited to the data set as this is not a randomized experiment but retrospective study.

# 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)