Predicting Breast Cancer Diagnosis

Authors: Matthew Timmerman, Elizabeth Nieto, Dayana Giunta, Ishita Mehta, Wilson Wu

Executive Summary

Current screening procedures such as biopsies of tumors can be invasive, expensive, and time-consuming. Instead, the images of the tumors with proper measurements can be used to screen for breast cancer, in conjunction with traditional methods. The data *Breast Cancer Wisconsin (Diagnostic)* was retrieved from kaggle. Using the image 10 features of each cell nucleus within the image were collected: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. From these features the mean, standard error and the worst score of each feature were collected for the image and made up the data set. For each of the original 10 features 3 features were created. There were 685 images and therefore 685 data points with 30 features describing the cell nuclei and one ID feature and one Diagnosis feature. The research goal of this project was to understand if these collected features of a cell nuclei from a breast mass image could be used to accurately classify a breast mass as malignant or benign.

The methods used to analyze the dataset are a combination of multivariate continuous and discrete models, along with cluster analysis techniques. To understand the relationship between the variables and the desired Diagnosis variable, we started with Linear Regression. Since there are often unprecedented cases when it comes to cancer, we decided to not remove any outliers and instead employ regularization techniques such as Ridge and Lasso Regression which provide a way to reduce the variance error introduced by the outliers. Furthermore, the chosen dataset consists of variables that are highly related to each other (such as nuclei radius and area). We used methods that would group redundant variables together such as Principal Component Analysis and Canonical Correlation Analysis which provide a deeper insight into the relationships between the potential variable groupings. Finally, since the goal of the research is to predict where the certain mass is either benign or malignant, we used Logistic Regression and Linear Discriminant Analysis to do the prediction and analyse the performance of our models.

Application results

The original dataset has more samples of benign cases than malignant cases, which could potentially mean that our chosen prediction model is biased towards benign cases. An implication of a biased model would mean a higher chance of diagnosing cancer when there is no cancer (i.e. false negatives.) Although this risk can be mitigated by using the same model multiple times on the same image. A better solution would be using a higher sample size to train the prediction model. By having a better balance between the provided benign and malignant cases, we can build a more specific prediction model. Future work could also include using the more significant variable groupings to improve the image collection and pre-processing. Despite the fact that this research focused on the physical characteristics of the mass, there is abundant research to indicate that the demographics of the patient also plays a role into the cancer

treatment plan required. Therefore by including demographics with the mass characteristics, we can build a better prediction model that reflects the sociological realities along with molecular level analysis.

Conclusion

Abstract

The determination of what type of tumor is one of the first steps that doctors applied after detecting it. For this examination, one sample it is taken from the cells and is analyzed through a biopsy procedure. This paper studies the detection of the type of tumor following the measurements of the features of the breast mass cell nuclei. Different methods were applied, such as linear regression, ridge and lasso regression, logistic regression, linear discriminant analysis, principal component analysis and canonical correlation analysis. The exploratory results obtained show that the type of breast cancer can be predicted by a coefficient of the measurement of the images of the cell that are represented in our data.

Introduction

The data *Breast Cancer Wisconsin(Diagnostic)* was retrieved from kaggle. Using the image 10 features of each cell nucleus within the image were collected: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. From these features the mean, standard error and the worst score of each feature were collected for the image and made up the data set. For each of the original 10 features 3 features were created. There were 685 images and therefore 685 data points with 30 features describing the cell nuclei and one ID feature and one Diagnosis feature. The focus of this project was to understand if features of a cell nuclei from a breast mass image can be used to accurately classify a breast mass as malignant or benign annd if so what featues would be most

Literature Review

It is no secret that breast cancer and breast cancer awareness is talked about and advertised everywhere today. What isn't talked about however, is the fact that breast cancer rates in America are rising among certain demographics. According to the CA : A Cancer Journal for Clinicians, "breast cancer incidence rates increased among Asian/Pacific Islander (1.7% per year), non-Hispanic black (NHB) (0.4% per year), and Hispanic (0.3% per year) women" from 2005 to 2014 and "approximately 252,710 new cases of invasive breast cancer and 40,610 breast cancer deaths are expected to occur among US women in 2017". Among different breast cancers, there is a different in molecular shape depending on the exact type of breast cancer and tumor. The Breast Cancer Research Journal claims that "In breast cancer, gene expression analyses have defined five tumor subtypes (luminal A, luminal B, HER2-enriched, basal-like and claudin-low), each of which has unique biologic and prognostic features". All five of these molecular subtypes of breast cancer tumors can be identified and then specific treatment can be given to target that individual subtype. In addition to molecular subtyping, a link has been found between the vascularity in breast cancer tumors and certain factors in the bone marrow of a patient. There is a “significant positive association between angiogenesis at the primary tumor site and micrometastasis” in the bone marrow of a patient, as outlined by the Journal of the National Cancer Institute.

Methods

To address the following research question of whether features of the breast mass cell nuclei can be used to predict whether the mass is benign or malignant, we employed the following methods. Studying the breast cancer prediction is a rigorous job. the risk of having an erroneous model can cost a person's life. For this reason it is very important to have the most accurate formula. Initially, to explore the relationship between the features and the response variable *Diagnosis*, linear regression and logistic regression were used, dividing the data by training (70%) and test (30%). Even though it is recommended to use logistic regression model when having binary variable, linear regression model helps us to explore the variables giving us interesting results. The best results were undoubtedly the logistic regression, giving us the formula to predict using whether the mass is benign or malignant using 13 different measures.

Since we noticed outliers in the dataset that could not be removed due to the highly sensitive nature of the medical domain, we explored regularization methods to help address the error in the regression models. We split the data into 70% training and 30% test and apply data into Tikhonov regularization (Ridge regression) and least absolute shrinkage and selection operator (LASSO) to evaluate the Breast Cancer Wisconsin. The Ridge is a regression method that performs L2 regularization. The Lasso is a regression method that performs L1 regularization and variable selection when the data has a huge number of features. After finding the best lambda.min and lambda.1se by using 10-fold cross-validation to fit the model, we apply 30% test into each Ridge and Lasso model to predict the mass and find out the best accuracy model.

The dataset has a smaller sample size than usually recommended for the conventional training and testing split approach for prediction. Therefore, we also used the classification technique Linear Discriminant Analysis (LDA) to predict diagnosis based on all the features of the mass. Due to the small size of the dataset, k-cross validation was used and the model’s performance was compared with the rest of the models.

Furthermore, the features are also highly multicollinear. To avoid overfitting the dataset, we explored cluster analysis technique Principal Component Analysis (PCA) for dimension reduction. To prep the data for PCA two features were removed. These were the *Diagnosis* feature and the *ID* feature. Kaiser Meyer Olkin(KMO), Bartlet’s Test of Sphericity and Chronbachs’s Alpha were used to test the data for Factorability. This was followed by the first PCA model on the data. Using this first model the scree plot was used to derive the number of components using Kaiser Meyer and the knee method. Once the number of components was selected PCA was ran again, this time with a selected cut off method to remove overlapping features among components to reduce correlation among the components.

With our 30 measurement values, we decided to look at our 10 mean measurements for a Canonical Correlation Analysis (CCA) as these 10 measurements were the stronger choice to look for correlation over standard error and worst case. These ten measurements were split into two groups, with 4 variables being in the standard measurements group, and

6 variables being in the specialized measurement group as seen in figure **Matt1.** These measurements were originally split by intuition and how the data was grouped originally, and then verified to work through testing. Canonical Correlation dimensions were judged based on their p-value after running a hypothesis test on all 4 dimensions using the Wilk’s Lambda test statistic. Canonical Dimension groupings were then formed using the standardized coefficients of the two groupings, and groupings were labeled with the amount of variability they explained from the data.

Discussion and Results

Preliminary Results

Linear Regression

For the first linear regression model, all the variables were plugged into the model, getting good R2 and adjusted R2. However, this model was discarded due the multicollinearity (Dayana figures.1).

A second try was made using the stepwise selection method, where backward and forward selection play an important role. Even though this model improves respect to the statistically significant, the multicollinearity still in the model. (Dayana figures.2)

For the last try using linear regression model, the variables with high multicollinearity were deleted, ending with a total of 14 variables. This model had a satisfactory result in reference to the R2 and adjusted R2 , also this model is statistically significant and there is no multicollinearity (Dayana figures.3).

Ridge and Lasso

For the Ridge regression result. The plot coefficients with log lambda (Wilson figures.1) shows the result that when we increase Log Lambda more and more, almost all the variables shrink into coefficients close to zero, but never drop off from the model. Next, to find the best lambda for our Ridge Model, the plot misclassification error with log lambda (Wilson figures.2) indicate that when log lambda around -2 which are the vertical dotted line interval, the model has =low misclassification error. One is Lambda.min 0.0677 which is the value that gives a minimum mean error, and the other is Lambda.1se: 0.207 which is the value that gives one standard error of the minimum. Between Lambda.min and Lambda.1se have amount difference. We apply both Lambda.min and Lambda.1se into our model. Both Lambda model keeps all the variables and shows the same accuracy. After we apply this model to our testing set, this mode indicates that prediction accuracy for the mass is 94.73%, Moreover, the false negatives for the prediction 8.49%.

For the Lasso regression result. The plot coefficients with log lambda (Wilson figures.3) shows the result that when we increase Log Lambda more and more, more variables shrink into coefficients to zero. Moreover, we found that when the log lambda is -2, the only concave.points_worst variable still stay significant. To find the best lambda for our Lasso Model, the plot misclassification error with log lambda (Wilson figures.4) indicate that when log lambda around -5.3 which are the vertical dotted line interval, the model has good prediction and acceptable numbers of variables. One is Lambda.min 0.0045 which is the value that gives a minimum mean error, and the other is Lambda.1se: 0.0054 which is the value that gives one standard error of the minimum.

We apply both Lambda.min and Lambda.1se into our model. Lambda.1se indicate less dimension and better accuracy, so we pick Lambda.1se model for our model. Lasso Lambda.1se shrink all variables to 10 significant variables. fractal dimension in se, smoothness in worst, and concave.points in worst have high coefficient values for the models.

$$\text{Lasso Model: } Log\ b\ \frac{P}{1-P} = 26.4165412 - 3.0971651 * concavity_{mean} - 6.5388170 * concave.points_{mean} - 1.2684859 * radius_{se} + 115.4896347 * fractal_{dimension_{se}} - 0.7323596 * radius_{worst} - 0.1712081 * texture_{worst} - 0.0045227 * perimeter_{worst} - 30.6757178 * smoothness_{worst} - 23.5827645 * concave.points_{worst} - 5.0950463 * symmetry_{worst}$$

After we apply this model to our testing set, this mode indicates that prediction accuracy for the mass is 97.66%, Moreover, the false negatives for the prediction 0.94%. Finally, we compare Ridge regression and Lasso regression. We maintain that Lasso model for breast cancer has higher prediction accuracy and not overfitting model. Furthermore, Lasso model has much lower false negatives prediction. We think is important for the not predict malignant to benign.

Logistic regression

Using the third model, from the linear regression analysis with 13 variables, we applied the logistic regression model. As a result, we end up with statistically significant for almost all variables, giving us the good signal to choose and make our predictions analysis (Dayana figures.4).

One advantage of using linear regression model is to be able to use the r2 to ensure the model can be useful or not. On the contrary, in logistic regression this step is difficult to calculate, due to the fact that the prediction line is not a straight line. McFadden (1973) suggested an alternative for the logistic regression, known as “likelihood ratio index”, comparing a model without any predictor to a model including all predictors. It is defined as one minus the ratio of the log likelihood with intercepts only, and the log likelihood with all predictors. For this model the McFadden score is 0.70 meaning that this model could be a strong predictor in order to answer our research question.

Now, having the coefficients we can explain the assumptions for example: If the texture_mean increase by one, the logOdds will increase by 0.42. To better understand the log of the odds is the division of the probability of getting breast cancer benign and the probability of getting breast cancer malignant. In other words the odds will be the exp(coefficient) and the logOdds will be the (odds-1)*100 this result is the percent of increase when the variable increase by 1 unit.

Previously the data was divided by training (70%) and test (30%). Now, in order to see if the model works, the probability of the prediction using the test data was calculated using the best model from the logistic regression(13 variables), creating a confusion matrix. Therefore, we end up with a really good results of 103 cases of true positive vs 3 case of false positive and 58 cases of true negative vs 7 cases of false negative (Dayana figures.5).

After that, we calculate the error rate that is 0.0994152 that its very satisfactory and subtracting it to 1 we will get the accuracy rate that is almost 90%. Additional to this, we plot a ROC curve, getting a 0.96.As we know, the upper level of the line is, the better is the model (Dayana figures.6).

Linear Discriminant Analysis

Principal Component Analysis

Factorability testing yielded the following results. A KMO Overall MSA of .83, a P value from Bartlet's Test of Sphericity of 2.22e-16 and a raw alpha from Chronbach's Alpha of .59 (Elizabeth Table 1). Three methods for component selection were implemented: Keiser Meyer, Knee Method, and Cumulative Variance(Elizabeth Figure 1, 2, Table 6.2). Moving forward with 3 components a cutoff value of .654 was selected. This resulted in three components. Component 1 named Size contained the following variables, perimeter mean, area mean, concavity mean, concave points mean, radius se, perimeter se, area se, radius worst, perimeter worst, area worst, comcavepoints worst. The variables with the highest variability contribution were perimeter mean and area mean(.971, .971). Component 2 names Spread contained the following variables, smoothness, compactness mean, fractal dimension mean, smoothness worst, compactness worst, concavity worst, symmetry worst, and fractal dimension worst. The variable contributing the most variability to this component was fractal dimension worst(.889). The third variable named Symmetry contained three variables: smoothness se, symmetry se and fractal dimension se. The variable contributing the most variability to this component was fractal dimension se(.733)(Elizabeth Table 3). These three components accounted for 73% of the variability(Elizabeth Table 4).

Canonical Correlation Analysis

With our two groups, standard and specialized measurements, a Canonical Correlation was run. We see with figure **Matt2** that the Canonical Correlation values are 0.97, 0.87, 0.44, and 0.20 for dimensions 1, 2, 3, and 4 respectively. These are all relatively high numbers for explaining the amount of variability in the data using CCA. Figure **Matt3** shows that our hypothesis test using the Wilks' Lambda statistic yielded all four dimensions being statistically significant at the 0.05 level. Next the dimensions were broken down into the standardized coefficients that can be seen in figure **Matt4**. We notice that dimension one does not yield any coefficients that are the largest among any of the variables, and the largest coefficient among each variable is bolded in the table. Based on this information, we were able to break down and group the variables for dimensions 2, 3, and 4 and these can be seen in figure **Matt5**. These groups ended up quite nicely with our first group being Outer measurements. This contains the perimeter and radius variables. Our second group, named Inner measurements contains area, compactness, concavity points, and fractal dimension. Finally, we have the Characteristics Measure group that contains texture, concavity points, and smoothness.

Common Trends

Limitations

In the Breast Cancer Wisconsin data set, the diagnosis variable is distributed as follows, 357 observations are benign and 212 observations are malignant. Since there is a higher frequency of benign cases, the unbalanced dataset might affect model building and be biased for prediction analysis. Variances are not equally distributed between the two cases, which also may lead to additional discrepancies. The dataset it also relatively small considering the minimum data point requirements for the methods implemented. Finally, because of insufficient knowledge in the healthcare industry, we weren't able to fully understand how to address the outliers in the dataset and the significance of the weights found in the regularization techniques.

Future Work

The factors discovered using PCA and CCA can be useful for breast cancer data collectors to focus on grouping these variables together in such a manner, and for future thought of adding additional measurements in their data collection. Maybe they are thinking of adding additional ways to measure the breast cancer tumor, measurements that deal with the outer, such as radius and perimeter could yield another measurement such as surface area. The same could be said with inner measurements of possibly density. Characteristic measurements are even more abundant.

Conclusion

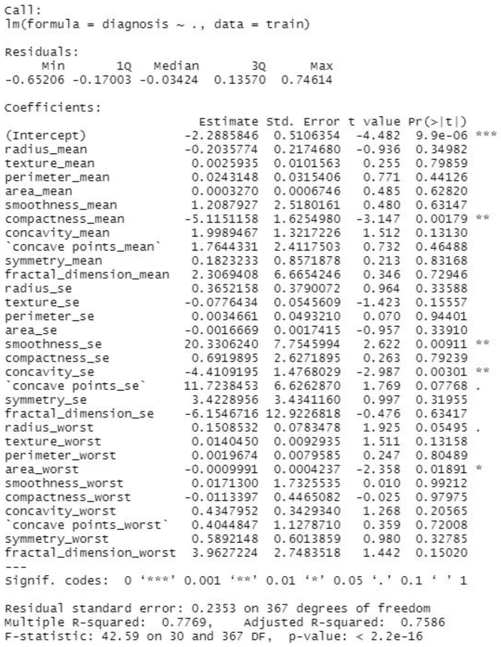
- logistic regression showed that not all variables have the same importance to the model, such as: symmetry_mean, texture_se, concavity_se and fractal_dimension_se.

- The CCA analysis has yielded us 3 canonical correlation dimensions that are all statistically significant, and explain 0.87, 0.44, and 0.2 of the variability each respectively.

Below conclusion section for PCA written after draft submission:

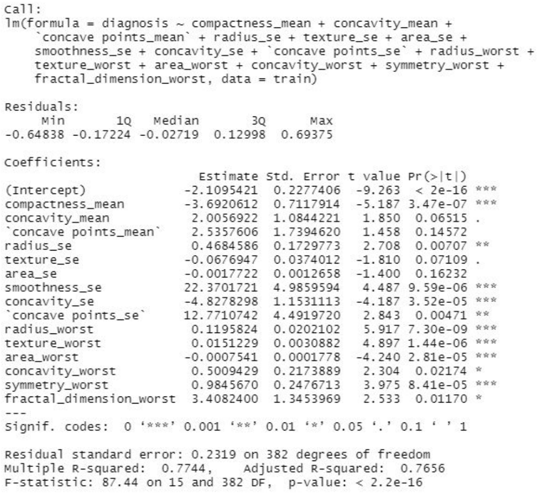
The three PCA components describing the shape, spread and symmetry of the cell nuclei contribute for 73% of the variance when it comes to classifying breast masses as benign or malignant. Theses latent features drawn from the original 30 can be used to focus future breast mass imaging on characteristics similar to or relating to the cell nuclei’s shape spread or symmetry. Present and future diagnosis may also rely more heavily on these features. In the event of conflicting feature measures for a particular mass those pertaining to shape size and symmetry can be used as a tiebreaker to classify the mass.

Appendix



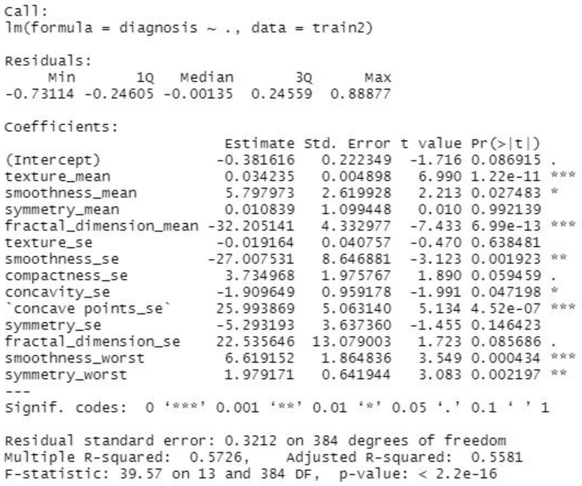
```
Call:
lm(formula = diagnosis ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65206 -0.17003 -0.03424  0.13570  0.74614 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -2.2885846  0.5106354  -4.482  9.9e-06 ***
radius_mean             -0.2035774  0.2174680  -0.936  0.34982    
texture_mean             0.0025935  0.0101563   0.255  0.79859    
perimeter_mean           0.0243148  0.0315406   0.771  0.44126    
area_mean                0.0003270  0.0006746   0.485  0.62820    
smoothness_mean          1.2087927  2.5180161   0.480  0.63147    
compactness_mean        -5.1151158  1.6254980  -3.147  0.00179 ** 
concavity_mean           1.9989467  1.3217226   1.512  0.13130    
`concave points_mean`    1.7644331  2.4117503   0.732  0.46488    
symmetry_mean            0.1823233  0.8571878   0.213  0.83168    
fractal_dimension_mean   2.3069408  6.6654246   0.346  0.72946    
radius_se                0.3652158  0.3790072   0.964  0.33588    
texture_se              -0.0776434  0.0545609  -1.423  0.15557    
perimeter_se             0.0034661  0.0493210   0.070  0.94401    
area_se                 -0.0016669  0.0017415  -0.957  0.33910    
smoothness_se           20.3306240  7.7545994   2.622  0.00911 ** 
compactness_se           0.6919895  2.6271895   0.263  0.79239    
concavity_se            -4.4109195  1.4768029  -2.987  0.00301 ** 
`concave points_se`     11.7238453  6.6262870   1.769  0.07768 .  
symmetry_se              3.4228956  3.4341160   0.997  0.31955    
fractal_dimension_se    -6.1546716 12.9226818  -0.476  0.63417    
radius_worst             0.1508532  0.0783478   1.925  0.05495 .  
texture_worst            0.0140450  0.0092935   1.511  0.13158    
perimeter_worst          0.0019674  0.0079585   0.247  0.80489    
area_worst              -0.0009991  0.0004237  -2.358  0.01891 *  
smoothness_worst         0.0171300  1.7325535   0.010  0.99212    
compactness_worst       -0.0113397  0.4465082  -0.025  0.97975    
concavity_worst          0.4347952  0.3429340   1.268  0.20565    
`concave points_worst`   0.4044847  1.1278710   0.359  0.72008    
symmetry_worst           0.5892148  0.6013859   0.980  0.32785    
fractal_dimension_worst  3.9627224  2.7483518   1.442  0.15020    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2353 on 367 degrees of freedom
Multiple R-squared:  0.7769,	Adjusted R-squared:  0.7586 
F-statistic: 42.59 on 30 and 367 DF,  p-value: < 2.2e-16
```

Dayana fig.1 Linear Regression Full model.



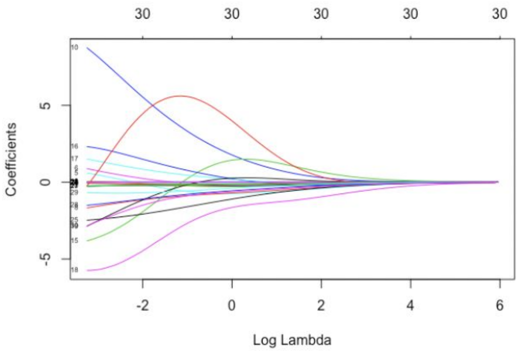
```
Call:
lm(formula = diagnosis ~ compactness_mean + concavity_mean + 
    `concave points_mean` + radius_se + texture_se + area_se + 
    smoothness_se + concavity_se + `concave points_se` + radius_worst + 
    texture_worst + area_worst + concavity_worst + symmetry_worst + 
    fractal_dimension_worst, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.64838 -0.17224 -0.02719  0.12998  0.69375 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -2.1095421  0.2277406  -9.263  < 2e-16 ***
compactness_mean        -3.6920612  0.7117914  -5.187 3.47e-07 ***
concavity_mean           2.0056922  1.0844221   1.850  0.06515 .  
`concave points_mean`    2.5357606  1.7394620   1.458  0.14572    
radius_se                0.4684586  0.1729773   2.708  0.00707 ** 
texture_se              -0.0676947  0.0374012  -1.810  0.07109 .  
area_se                 -0.0017722  0.0012658  -1.400  0.16232    
smoothness_se           22.3701721  4.9859594   4.487 9.59e-06 ***
concavity_se            -4.8278298  1.1531113  -4.187 3.52e-05 ***
`concave points_se`     12.7710742  4.4919720   2.843  0.00471 ** 
radius_worst             0.1195824  0.0202102   5.917 7.30e-09 ***
texture_worst            0.0151229  0.0030882   4.897 1.44e-06 ***
area_worst              -0.0007541  0.0001778  -4.240 2.81e-05 ***
concavity_worst          0.5009429  0.2173889   2.304  0.02174 *  
symmetry_worst           0.9845670  0.2476713   3.975 8.41e-05 ***
fractal_dimension_worst  3.4082400  1.3453969   2.533  0.01170 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2319 on 382 degrees of freedom
Multiple R-squared:  0.7744,	Adjusted R-squared:  0.7656 
F-statistic: 87.44 on 15 and 382 DF,  p-value: < 2.2e-16
```

Dayana fig.2 Linear Regression Model Stepwise selection.

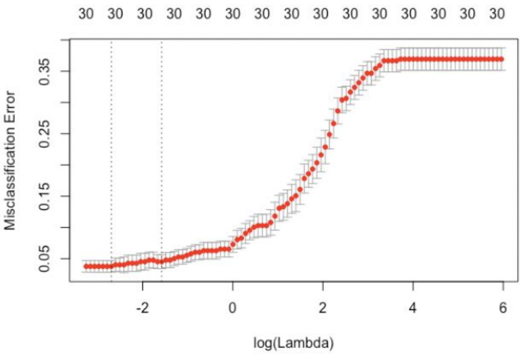


```
Call:
lm(formula = diagnosis ~ ., data = train2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73114 -0.24605 -0.00135  0.24559  0.88877 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            -0.381616   0.222349  -1.716 0.086915 .  
texture_mean            0.034235   0.004898   6.990 1.22e-11 ***
smoothness_mean         5.797973   2.619928   2.213 0.027483 *  
symmetry_mean           0.010839   1.099448   0.010 0.992139    
fractal_dimension_mean -32.205141   4.332977  -7.433 6.99e-13 ***
texture_se             -0.019164   0.040757  -0.470 0.638481    
smoothness_se         -27.007531   8.646881  -3.123 0.001923 ** 
compactness_se          3.734968   1.975767   1.890 0.059459 .  
concavity_se           -1.909649   0.959178  -1.991 0.047198 *  
`concave points_se`    25.993869   5.063140   5.134 4.52e-07 ***
symmetry_se            -5.293193   3.637360  -1.455 0.146423    
fractal_dimension_se   22.535646  13.079003   1.723 0.085686 .  
smoothness_worst        6.619152   1.864836   3.549 0.000434 ***
symmetry_worst          1.979171   0.641944   3.083 0.002197 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3212 on 384 degrees of freedom
Multiple R-squared:  0.5726,	Adjusted R-squared:  0.5581 
F-statistic: 39.57 on 13 and 384 DF,  p-value: < 2.2e-16
```

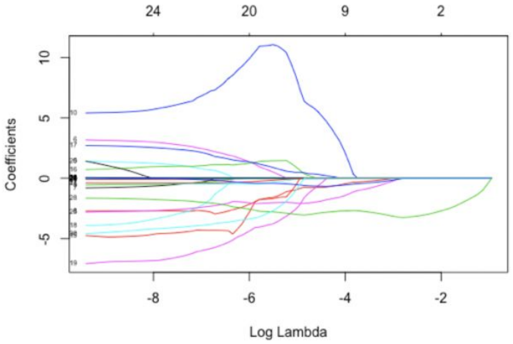
Dayana fig.3 Linear Regression 13 variables.



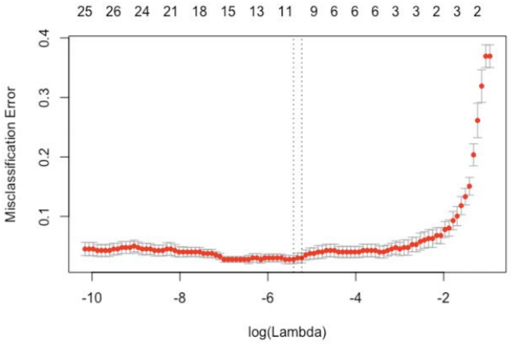
Wilson fig.1. Lambda coefficients plot



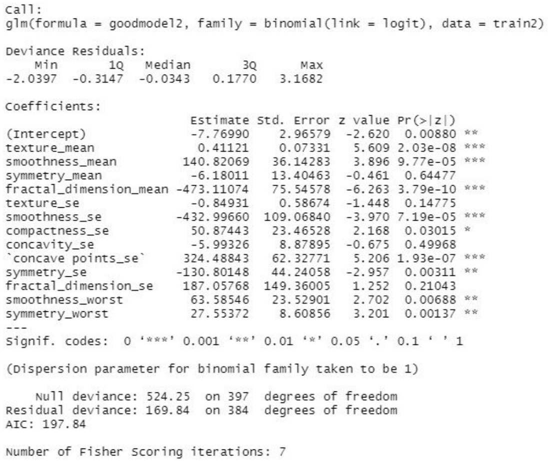
Wilson fig.2. Lambda misclassification error plot



Wilson fig.3. Lambda coefficients plot

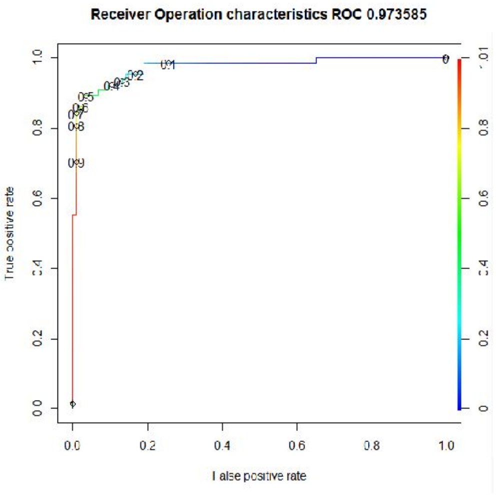


Wilson fig.4. Lambda misclassification error plot

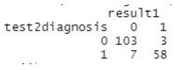


```
Call:
glm(formula = goodmodel2, family = binomial(link = logit), data = train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0397  -0.3147  -0.0343   0.1770   3.1682

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -7.76990    2.96579  -2.620  0.00880 **
texture_mean            0.41121    0.07331   5.609 2.03e-08 ***
smoothness_mean       140.82069   36.14283   3.896 9.77e-05 ***
symmetry_mean          -6.18011   13.40463  -0.461  0.64477
fractal_dimension_mean -473.11074  75.54578  -6.263 3.79e-10 ***
texture_se             -0.84931    0.58674  -1.448  0.14775
smoothness_se        -432.99660  109.06840  -3.970 7.19e-05 ***
compactness_se         50.87443   23.46528   2.168  0.03015 *
concavity_se           -5.99326    8.87895  -0.675  0.49968
`concave points_se`   324.48843   62.32771   5.206 1.93e-07 ***
symmetry_se          -130.80148   44.24058  -2.957  0.00311 **
fractal_dimension_se  187.05768  149.36005   1.252  0.21043
smoothness_worst       63.58546   23.52901   2.702  0.00688 **
symmetry_worst         27.55372    8.60856   3.201  0.00137 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 524.25  on 397  degrees of freedom
Residual deviance: 169.84  on 384  degrees of freedom
AIC: 197.84

Number of Fisher Scoring iterations: 7
```

Dayana fig.4 Logistic Regression 13 variables.



Dayana fig.6 ROC



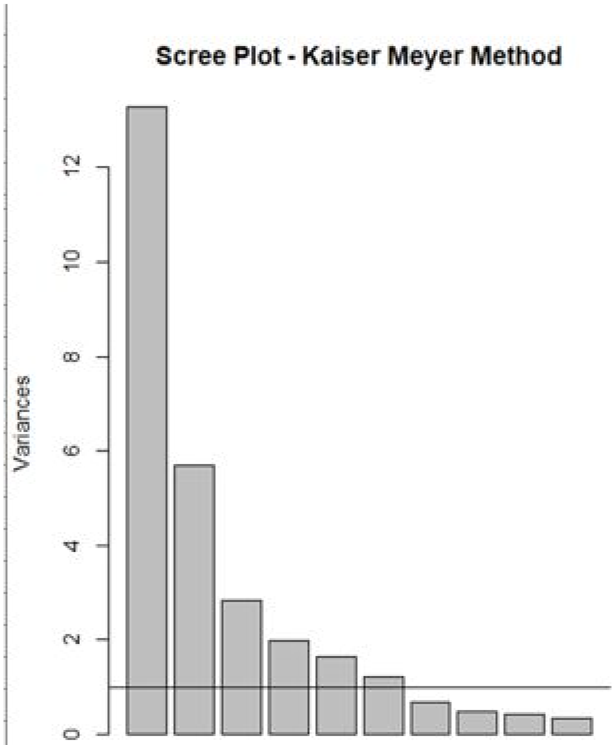
```
             result1
test2diagnosis   0   1
             0 103   3
             1   7  58
```

Dayana fig.5 Confusion Matrix.

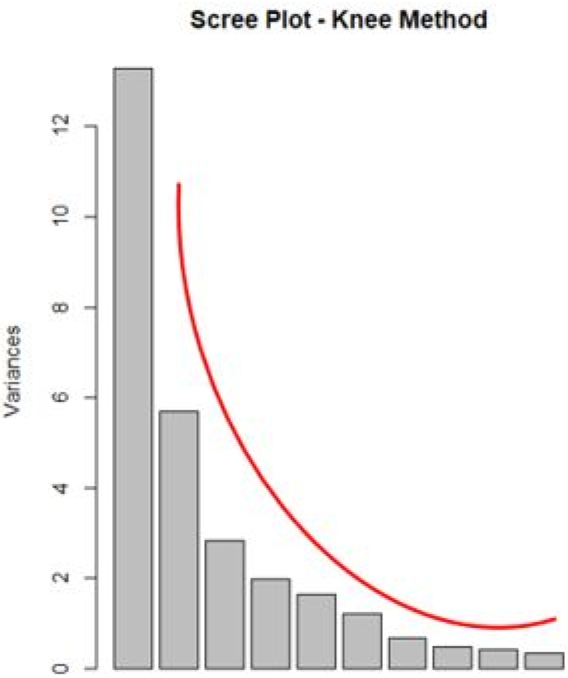
Tests for Factorability

KMO	Bartlett's Test	Chronbach's Alpha
Overall MSA	P-value	Raw Alpha
0.83	< 2.22e-16	0.59

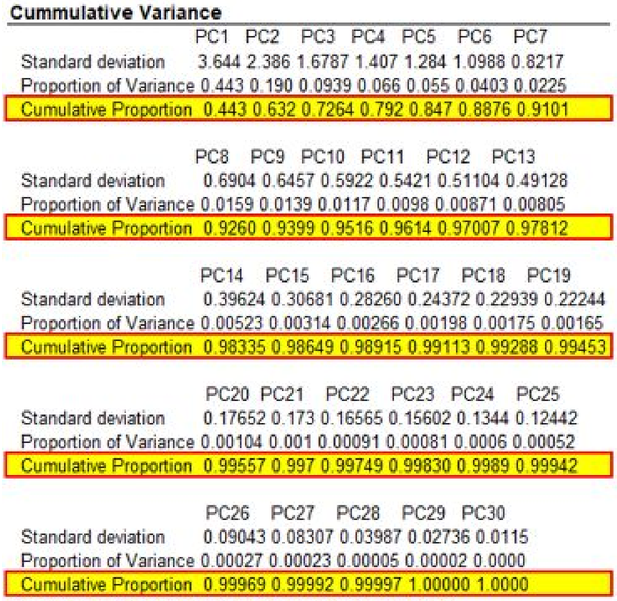
Elizabeth Table 1 Tests for Factoability



Elizabeth Figure 1. Scree Plot with Applied Keiser Meyer method suggesting 6 components for PCA



Elizabeth Figure 1. Scree Plot with Applied Knee method suggesting 3 components for PCA

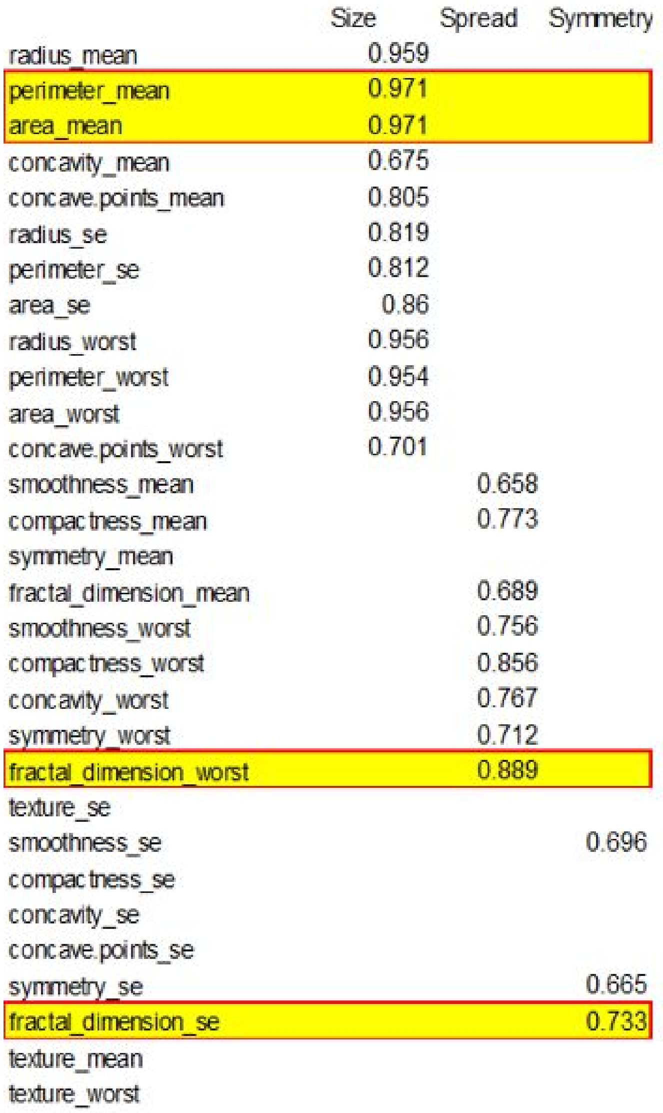


Cummulative Variance

```
                         PC1   PC2    PC3    PC4   PC5    PC6    PC7
Standard deviation       3.644 2.386 1.6787 1.407 1.284 1.0988 0.8217
Proportion of Variance 0.443 0.190 0.0939 0.066 0.055 0.0403 0.0225
Cumulative Proportion  0.443 0.632 0.7264 0.792 0.847 0.8876 0.9101

                         PC8    PC9    PC10   PC11    PC12    PC13
Standard deviation      0.6904 0.6457 0.5922 0.5421 0.51104 0.49128
Proportion of Variance 0.0159 0.0139 0.0117 0.0098 0.00871 0.00805
Cumulative Proportion  0.9260 0.9399 0.9516 0.9614 0.97007 0.97812

                          PC14    PC15    PC16    PC17    PC18    PC19
Standard deviation      0.39624 0.30681 0.28260 0.24372 0.22939 0.22244
Proportion of Variance 0.00523 0.00314 0.00266 0.00198 0.00175 0.00165
Cumulative Proportion  0.98335 0.98649 0.98915 0.99113 0.99288 0.99453

                          PC20  PC21    PC22    PC23   PC24    PC25
Standard deviation      0.17652 0.173 0.16565 0.15602 0.1344 0.12442
Proportion of Variance 0.00104 0.001 0.00091 0.00081 0.0006 0.00052
Cumulative Proportion  0.99557 0.997 0.99749 0.99830 0.9989 0.99942

                          PC26    PC27    PC28    PC29   PC30
Standard deviation      0.09043 0.08307 0.03987 0.02736 0.0115
Proportion of Variance 0.00027 0.00023 0.00005 0.00002 0.0000
Cumulative Proportion  0.99969 0.99992 0.99997 1.00000 1.0000
```

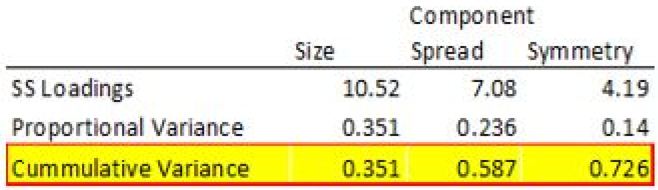
Elizabeth Table 2: Cummultive variance from initial 30 components of PCA

PCA Components by Feature Variance



	Size	Spread	Symmetry
radius_mean	0.959		
perimeter_mean	0.971		
area_mean	0.971		
concavity_mean	0.675		
concave.points_mean	0.805		
radius_se	0.819		
perimeter_se	0.812		
area_se	0.86		
radius_worst	0.956		
perimeter_worst	0.954		
area_worst	0.956		
concave.points_worst	0.701		
smoothness_mean		0.658	
compactness_mean		0.773	
symmetry_mean			
fractal_dimension_mean		0.689	
smoothness_worst		0.756	
compactness_worst		0.856	
concavity_worst		0.767	
symmetry_worst		0.712	
fractal_dimension_worst		0.889	
texture_se			
smoothness_se			0.696
compactness_se			
concavity_se			
concave.points_se			
symmetry_se			0.665
fractal_dimension_se			0.733
texture_mean			
texture_worst			

Elizabeth Table 3: this table shows all three components with the corresponding variables at a .654 cutoff point. The variables contributing the most variance to each coponent is highlighted



		Component	
	Size	Spread	Symmetry
SS Loadings	10.52	7.08	4.19
Proportional Variance	0.351	0.236	0.14
Cummulative Variance	0.351	0.587	0.726

Elizabeth Table 4: This table shows the proportion of variance for each component and the cummulative variance

References

Desantis, C. E., Ma, J., Sauer, A. G., Newman, L. A., & Jemal, A. (2017). Breast cancer statistics, 2017, racial disparity in mortality by state. *CA: A Cancer Journal for Clinicians, 67*(6), 439-448. doi:10.3322/caac.21412

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., . . . Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research, 12*(5). doi:10.1186/bcr2635

Stephen B. Fox, Kevin C. Gatter, Russel D. Leek, Adrian L. Harris, Judith Bliss, Janine L. Mansi, Barry Gusterson, Association of Tumor Angiogenesis With Bone Marrow Micrometastases in Breast Cancer Patients, *JNCI: Journal of the National Cancer Institute*, Volume 89, Issue 14, 16 July 1997, Pages 1044–1049, https://doi.org/10.1093/jnci/89.14.1044

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics (Edited by P. Zarembka), 105-42. Academic Press, New York.