Cancer Prediction

Group 2 (Colin Fitzpatrick, Sam Mosiman, Brandon Winder)

10/22/2020

# Introduction

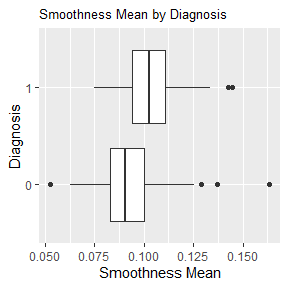
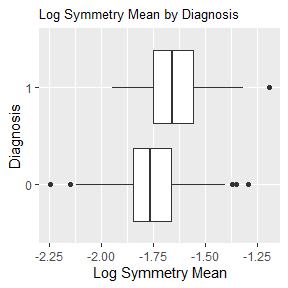
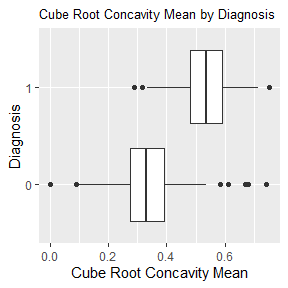
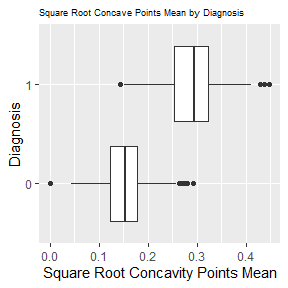
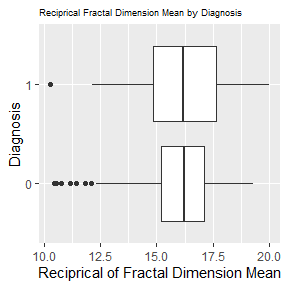
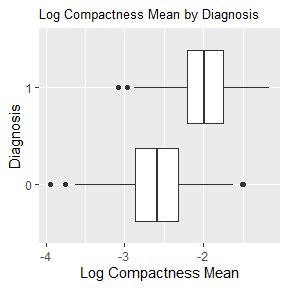
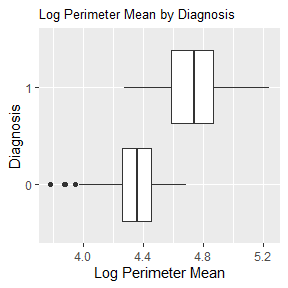
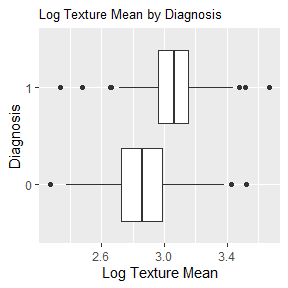
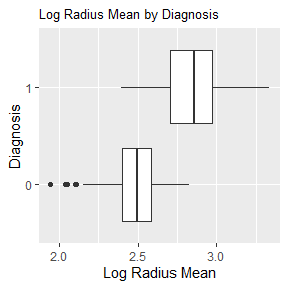
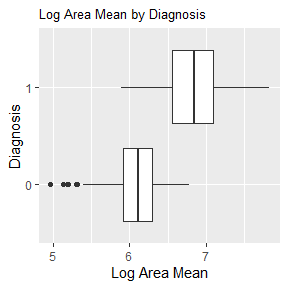
The goal of our work is to use data on tissue samples taken from cancerous breast tissue to create a logistic regression model that can be used to accurately predict if a cancer sample is malignant or benign. Our approach is to first analyze all 10 mean predictor variables to check for extreme values that could influence our models accuracy. We chose to use mean predictors exclusively because the worst predictors should be included in the mean, but we didn’t want to focus on the outliers for each specific observation. Additionally we did not include the standard error variables to predict whether cancer cells are malignant or benign because they are solely based off the mean variables and would not provide much information on the diagnosis. To do this, we will plot predictor variables and look for outliers that could skew the variable. If need be, we can then transform the skewed variables through the use of square roots, logarithms, and reciprocals. We will also look at the predictor variables to determine if any are highly correlated so that we can reduce the number of predictor variables to account for overfitting our regression model. Once we have checked our predictor variables and transformed those where we felt necessary, we will build models and test their predictive capabilities.

# Exploring and Transforming the Data

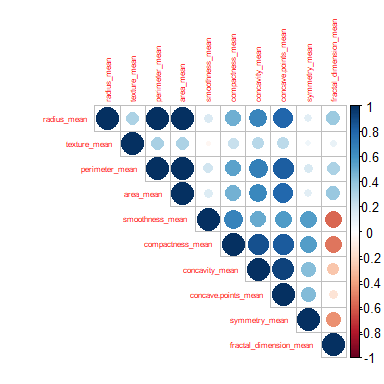
We first looked into the descriptive statistics for each variable mean and these are displayed in the table below.

## radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean  
## Min. 6.98 9.71 43.79 143.50 0.05  
## 1st Qu. 11.63 16.17 74.68 415.10 0.09  
## Median 13.24 18.82 85.48 541.60 0.10  
## Mean 14.04 19.31 91.41 648.49 0.10  
## 3rd Qu. 15.61 21.68 103.20 761.30 0.11  
## Max. 28.11 39.28 188.50 2501.00 0.16  
## compactness\_mean concavity\_mean concave.points\_mean symmetry\_mean  
## Min. 0.02 0.00 0.00 0.11  
## 1st Qu. 0.06 0.03 0.02 0.16  
## Median 0.09 0.06 0.03 0.18  
## Mean 0.10 0.09 0.05 0.18  
## 3rd Qu. 0.13 0.13 0.07 0.20  
## Max. 0.31 0.43 0.20 0.30  
## fractal\_dimension\_mean  
## Min. 0.05  
## 1st Qu. 0.06  
## Median 0.06  
## Mean 0.06  
## 3rd Qu. 0.07  
## Max. 0.10

After analyzing the summaries and the distribution of the variables, we have decided to perform many transformations to account for abnormally distributed data. The transformations on each variable can be seen in the boxplots displayed below. These transformations were performed because the normal boxplots of the predictors had outliers and/or were highly skewed. Then the data was transformed either by a log, reciprocal, square root, or cube root transformations. These transformations make the distributions less skewed and contain less outliers, making sure the data is as close to normal as possible. After performing the transformations, the box plots shown below show less outliers and are less skewed than the originals. These transformations have a more likely chance of getting better prediction models later when we conduct tests of the data. Models will also be more accurate with these changes in transformations as there are less outliers to skew the models that we will conduct later. Overall, these transformations were needed to better fit models and get the data as close to normal as possible. The boxplots shown below help us identify which variables need to be transformed. When we plot these we are looking for three things. We are looking to see how many outliers the data has, the spread of the data, and if it is consistent throughout all of plots. We are also looking to see if the data is heavily skewed in these boxplots as well. These three factors will help us determine if we need to make the data fit better to a normal curve.



Based on this correlation plot, we can see that the variable ‘concave.points\_mean’ and ‘concavity\_mean’ have a high positive correlation with most other predictor variable means. Also when looking at the predictors standard errors and worst values we see the same trends. So, we will want to pay close attention to these variables when creating the logistic regression model to avoid multicollinearity. Additionally, we will want to examine the variables ‘radius\_mean’, ‘perimeter\_mean’, and ‘area\_mean’ because they are also highly positively correlated.



## The Logistic Model

First, we separated the data into a training and validation set. This allowed us to build our model in the training set and then test against the validation set. This is needed to make sure the model is not overfitted. Testing our model on a validation set that wasn’t used to create the model allows us to check that the model is actually useful in prediction with new data.

Next, we created a model for prediction with all of the variables used in the model. We then created contingency tables and found the sensitivity, specificity, and the overall accuracy of the model.

## [1] "The senstivity for this model on the validation data is: "

## [1] 0.952381

## [1] "The specificity for this model on the validation data is: "

## [1] 0.972973

## [1] "The overall accuracy for this model on the validation data is: "

## [1] 0.9655172

## Model Selection

After making our model utilizing all predictors, we made two other models using both forward and backward stepwise regression to select the most significant predictors. We then created contingency tables and found the sensitivity, specificity, and the overall accuracy for both of the models. Overall, we ended choosing the backward stepwise regression model as it had slightly better accuracy while also using fewer predictor variables.

## [1] "The senstivity using the forward selection model on the validation data is: "

## [1] 0.952381

## [1] "The specificity using the forward selection model on the validation data is: "

## [1] 0.972973

## [1] "The overall accuracy using the forward selection model on the validation data is: "

## [1] 0.9655172

## [1] "The senstivity using the backward selection model on the validation data is: "

## [1] 0.952381

## [1] "The specificity using the backward selection model on the validation data is: "

## [1] 0.9864865

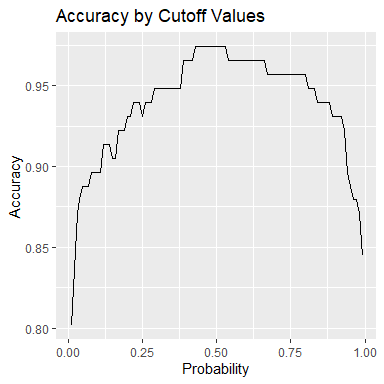
## [1] "The overall accuracy using the forward selection model on the validation data is: "

## [1] 0.9741379

Based off of this we find that the best model selected using the backward stepwise regression results in the highest overall accuracy. Its specificity has increased while its sensitivity has remained the same. This model is:

## Optimal Threshold

We created an accuracy plot for all varying classification thresholds and it showed that thresholds between .43 and .53 have the same accuracy and also the highest accuracy among all thresholds. We ended up choosing 0.5 as our threshold because it was one of the thresholds that resulted in the highest accuracy for the data. Additionally, 0.5 is the middle of all probabilities and if we were to decrease this classification threshold it would result in more samples being classified as malignant even if they are benign.



# Results Summary

In conclusion, the backwards elimination model is the best model for predicting whether a cancerous breast tissue is malignant or benign. This comes from how the backwards elimination model has the highest accuracy out of all three models tested and also has the lowest amount of predictors in any of the models. The accuracy of this model was 97.4% and the classification threshold we choose was at 0.5. This model also has a sensitivity of 95.2% and a specificity of 98.6%. Overall, the backwards elimination model does the best job in all categories of predicting cancerous breast tissue.