BREAST CANCER MODELS

Jessenia Morales, Benny Hernandez, Diana Sanchez, & Sonjhalyns Augustin.

Minor in Data Science
219:220 Fundamentals of Data Visualization in R: Final project

Prof Bruno Richard, PhD

Introduction: Background Info

Breast cancer is one of the most common malignancies found in women. One in three women are diagnosed with breast cancer each year. It occurs as the result of abnormal growth of cells in breast tissue. A breast tumor has the possibility of being benign, malignant, or premalignant. A malignant tumor is cancerous and requires immediate treatment, normally in the form of chemotherapy. A benign tumor is not cancerous and does not pose an immediate threat to a patient's health. A pre-malignant tumor is precancerous, and requires observation from medical professionals. This dataset was collected by the University of California-Irvine and contains information on tumor diagnoses and their characteristics. It only focuses on the benign and malignant diagnoses, and does not include any pre-malignant diagnoses. Prior to being made available to the public in 1995, a procedure called fine needle aspiration is performed. When collecting samples, fluid from a cyst in the breast is removed using a fine needle. After collection, the mean, standard error, and extreme values were calculated based on various physical characteristics of the tumor samples. The features of each sample include radius, texture, perimeter, area, smoothness , compactness , concavity, concave points, symmetry, and fractal dimension.

These features measure various characteristics that change significantly with a malignant breast tumor. For example, compactness measures how densely packed cells are within the tissue. Concave points detail how many inward curvatures and indentations are present in cell nuclei, and concavity measures how severe these concave points are. Malignant tumors are normally characterized by high levels of concavity and concave points. Many characteristics are used to record and classify these tumors. However, it is important to determine which are the most significant when diagnosing these tumors.

Introduction: Question & Hypothesis

How well do size characteristics serve as a predictor variable for a malignant tumor diagnosis in comparison to the concave measurements of a tumor ? As a medical professional, which feature should be considered the most important when testing a patient?

The null hypothesis being tested is concave points and size characteristics have no relationship or association with a malignant tumor diagnosis. The alternative hypothesis is concave points have a stronger relationship or association with a malignant tumor diagnosis.

Methods: Dataset Description

The target variable being tested is the diagnosis of each breast tumor sample, which is classified with “M” for malignant or “B” for benign. This is the only categorical variable being tested. The rest of the predictor variables include the different characteristics of a breast tumor, which are all numerical variables. The mean values of each of the characterisitics will be tested. Therefore, the worst and standard error columns of each characteristic will be excluded.

Methods: Exploratory Data Analysis

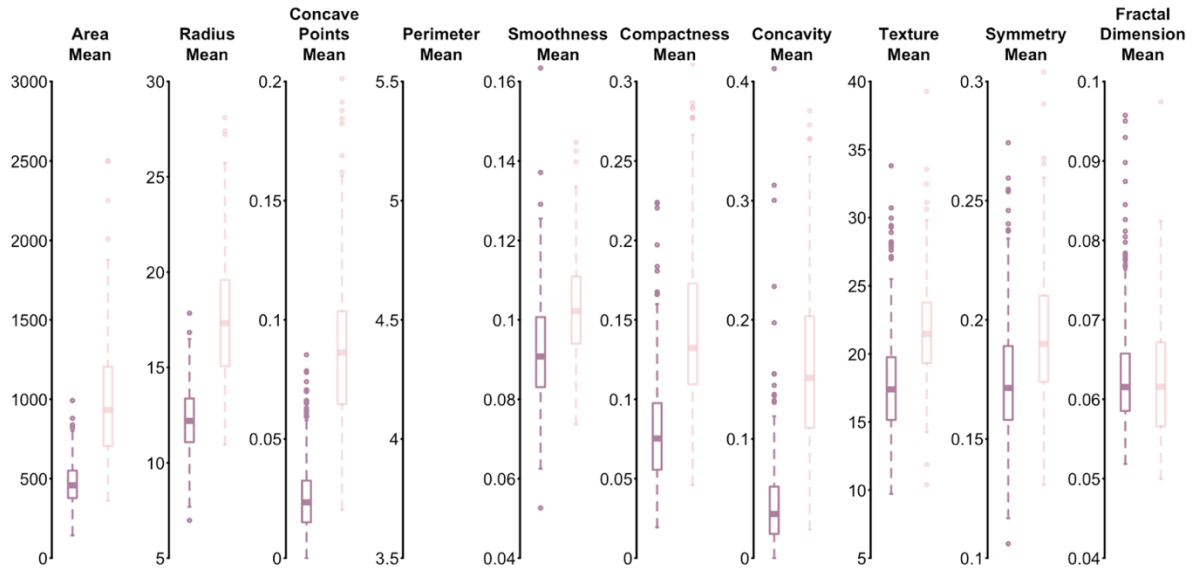


Figure 1a: The boxplots shown above

Boxplot charts are typically used to give a quick and intuitive overview of the distribution of data. In this instance, we are displaying a series of boxplots that allows us to view how the distribution of each predictor variable differs between the two categories of malignancy and benign in the diagnosis of breast cancer. The boxplot presents benign cases as light purple and malignancies as light pink. All predictor variables, with the exception of the fractal dimension mean, show how the malignancy distribution range is higher. For the fractal dimension mean, the malignancy upper and lower quartile are bigger than that of the benign case , but the mean appears to be the same.

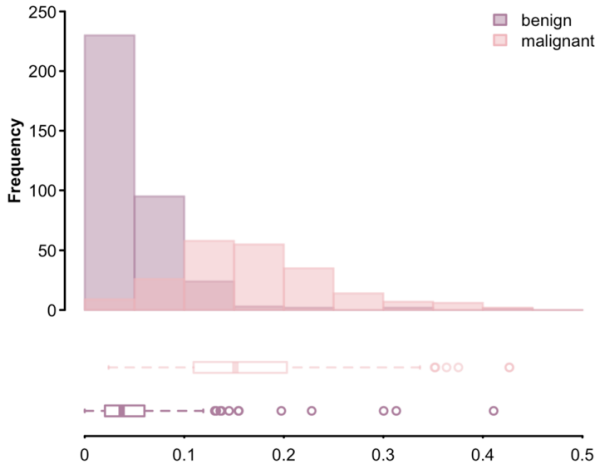


Figure 2a

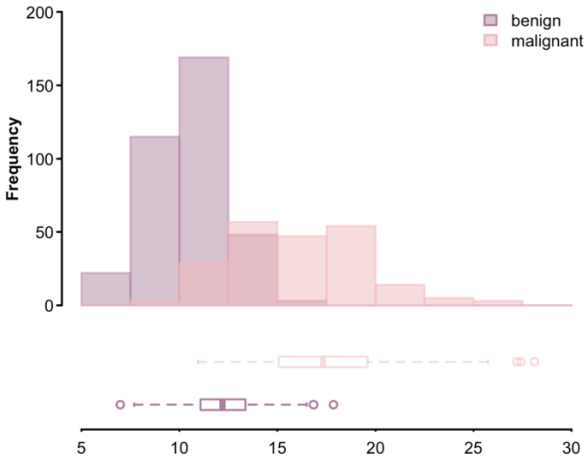


Figure 3a

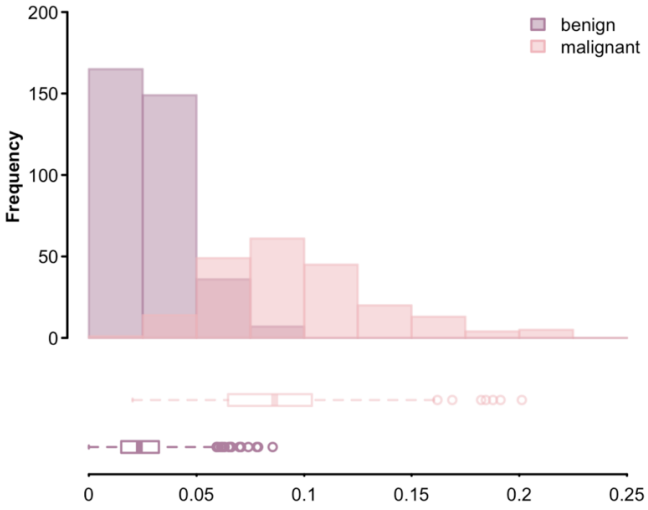


Figure 4a

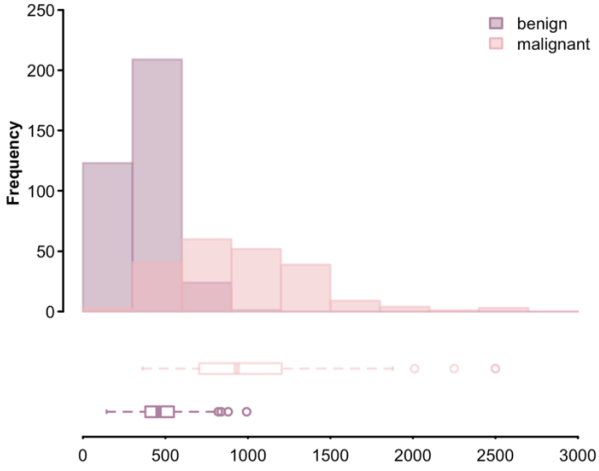


Figure 5a

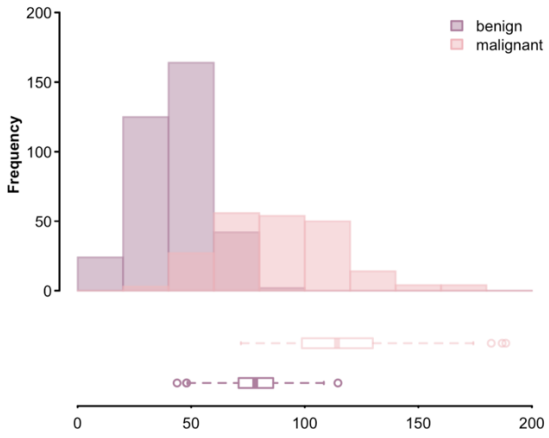


Figure 6a.

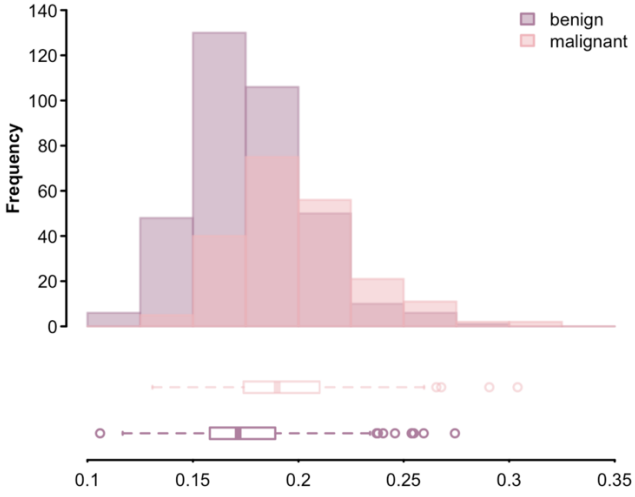


Figure 7a..

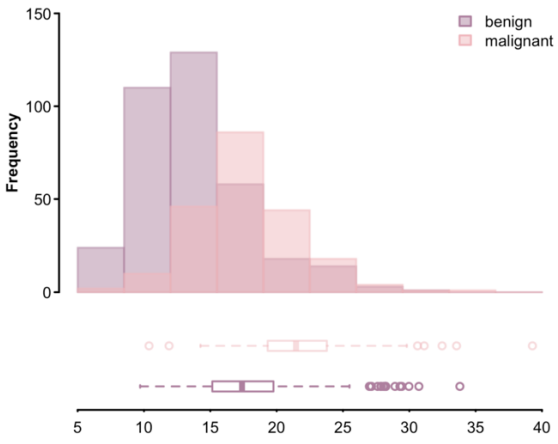


Figure 8a.

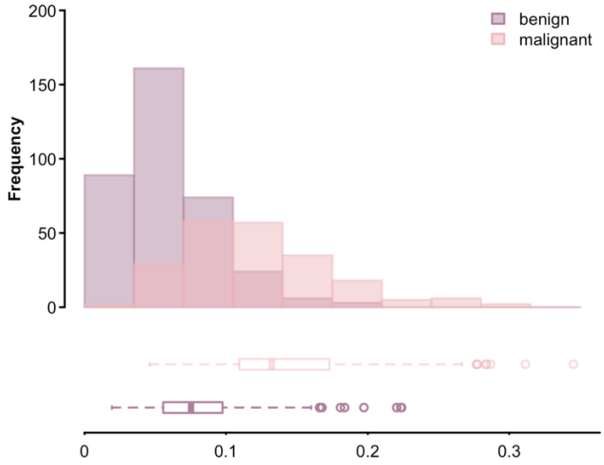


Figure 9a.

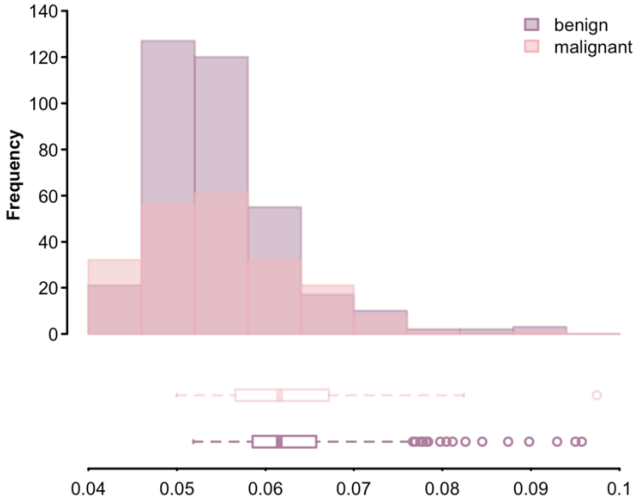


Figure 10a.

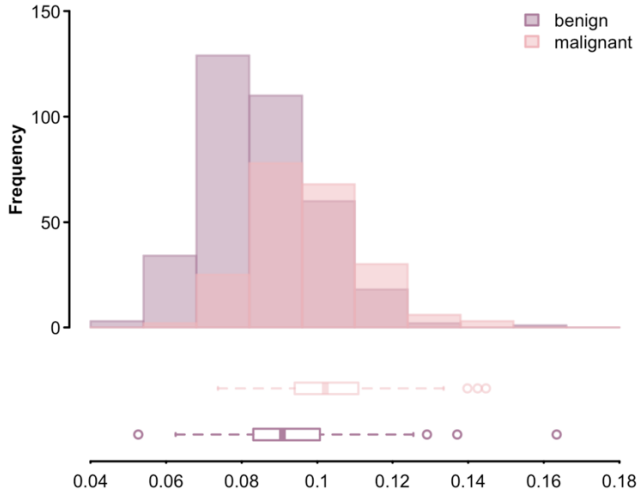


Figure 11a..

The histogram and boxplots for the predictor variables are plotted in the images above. Area Mean, radius Mean, concave points Mean, perimeter Mean, smoothness Mean, compactness Mean, concavity Mean, texture Mean, symmetry Mean, and fractal dimension Mean. The histograms are used to show the frequency distribution of values for their predictor variable. To distinguish between the various diagnoses, the histograms in the figures above are overlaid with two different colors. The histogram reveals a significant discrepancy between the two diagnoses.

Instances of malignancy are often pushed to the right, demonstrating how a frequency of higher values typically favors cases of malignancy diagnosis. Additional details on the distribution range of values are provided by the boxplots used in the figures above. We are able to observe the predictor variables that contain outliers from there and how they affect the frequency.

Correlation

In order to create our models we want to decide what variables to choose. We could try trial and error however there exists a massive amount of possible combinations then we could possibly try; One possibility could be making a function to create all possible models but we don't have anywhere near the computing power necessary to create and test all those models. Instead we will try to see what variables correlate the strongest with our target variable. However, our predictor variables are numeric and our target is a dichotomous categorical variable, so we can't use any of the coefficient variables in the R's cor function (Pearson's, Kendall's, Spearman's) and as those all assume you have a numeric target variable. We will instead opt for a point biserial correlation which means we must make our own function; The formula we are putting into code is as follows:

$$r_{pb} = \frac{g}{\sqrt{g^2 + dfw(\frac{1}{n_1} + \frac{1}{n_2})}},$$

where *g* is Hughe's G, *dfw* is degrees of freedom within, & *n* refers to the two variables After running every variable against each other we get a matrix that we then use to make the following corrplot.

Figure 2b..

From the corrplot we can see which variables correlate with diagnosis but to avoid multicollinearity we also want to make sure that the predictor variables we choose don't correlate with each other; Our cutoff to avoid multicollinearity will be >±0.5. An example would be if we want to make a model with radius as a predictor variable we can make a model with texture as it is -0.5 correlation with radius but we can't do the same with perimeter as it has a -0.9 correlation with radius. Using this methodology we have 15 models, plus one model which includes all variables. The models are as follows:

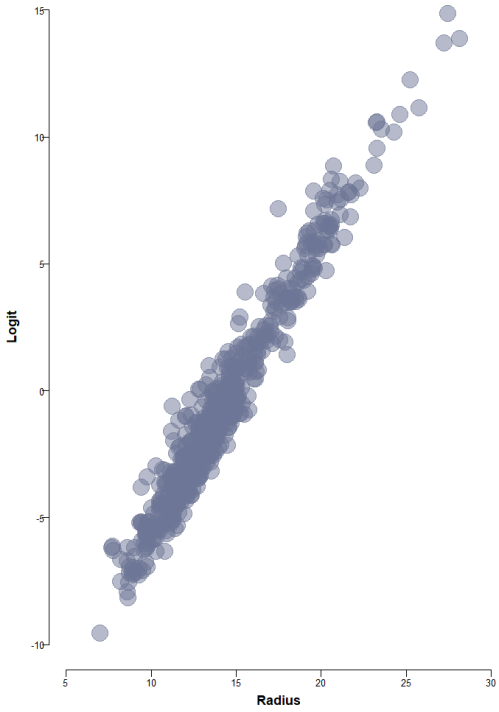
- Model 1: All variables
- Model 2: Radius & Texture
- Model 3: Smoothness, Compactness, & Concavity
- Model 4: Smoothness & Concavity
- Model 9: Compactness, Concavity, Concave Points, & Fractal Dimension
- Model 10: Compactness & Concavity
- Model 11: Compactness & Concave Points

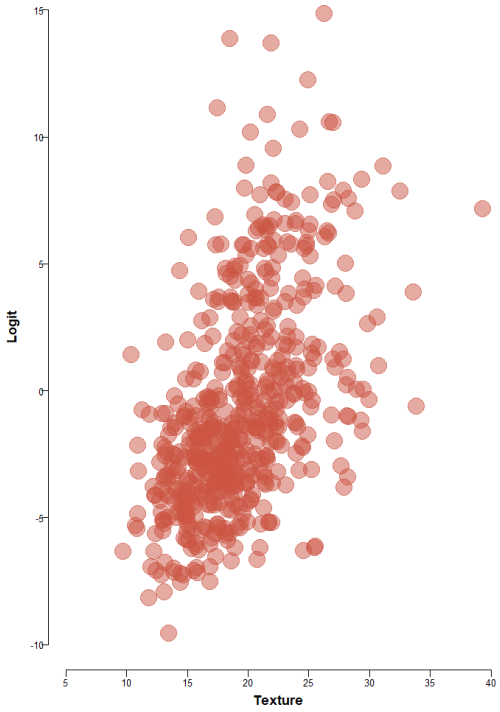
- Model 5: Smoothness & Compactness
- Model 6: Compactness, Concavity, Concave Points, & Fractal Dimension
- Model 7: Compactness, Concavity, & Concave Points
- Model 8: Compactness, Concavity, & Fractal Dimension
- Model 12: Compactness & Fractal Dimension
- Model 13: Compactness & Concavity
- Model 14: Concavity, Concave Points, & Fractal Dimension
- Model 15: Concavity & Fractal Dimension
- Model 16: Concave Points & Fractal Dimension

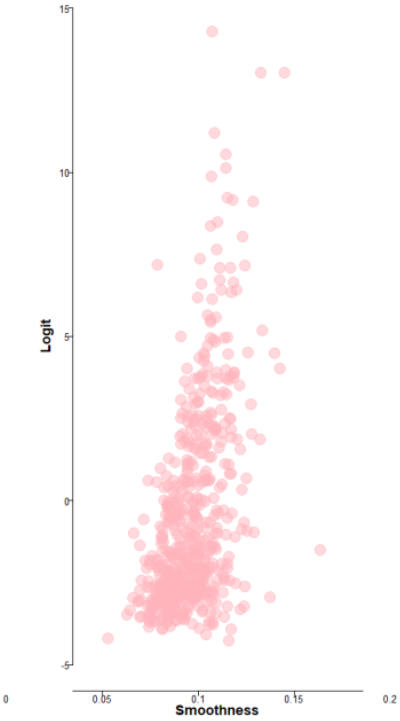
Assumptions of Logistic Regression:

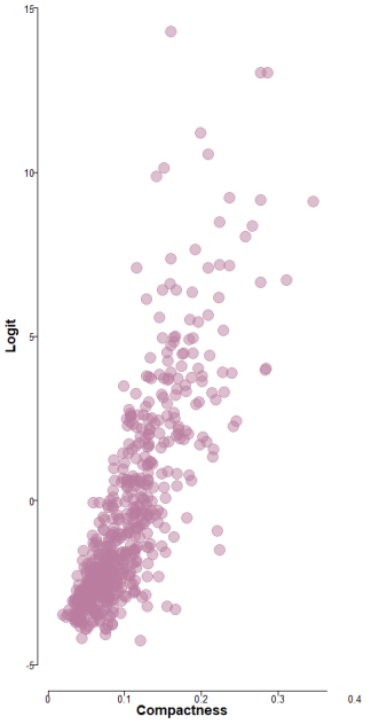
First, we test the assumptions that don't require models. The first assumption is that our target is binary, which it is because the diagnosis can either be malignant or benign. The second assumption is that our inputs are independent, which they end up being. The third assumption is that our sample is large enough to make a model with, which we decide it is so (The total size is 430).

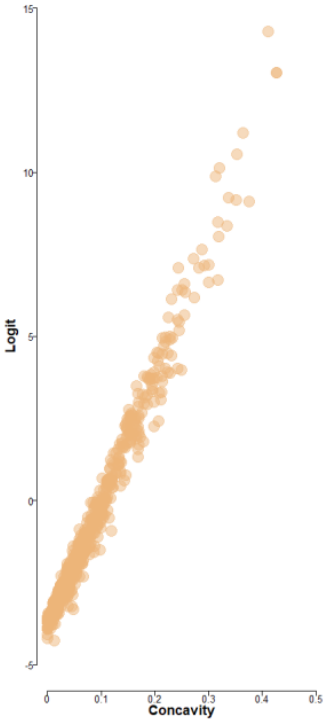
Next we test the assumptions that require our models. We test for multicollinearity and find that some of our models (Models 1, 6, & 9) violate this clause and must drop them. The next assumption is that there's a linear association between our variable and the logit (logit = p/1-p). To test this we make scatter plots with our predictor variables along the x-axis and the logit for each model along the y-axis. If any model contains a graph that doesn't appear linear we drop it.

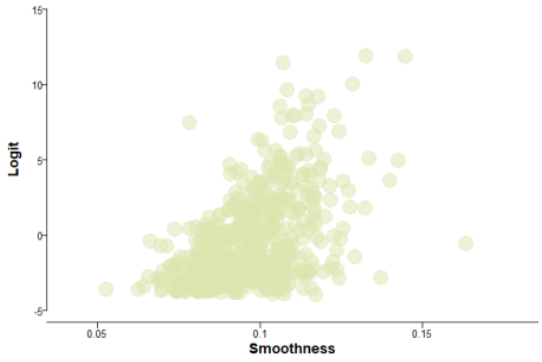

Logit
15
10
5
0
-5
-10
5
10
15
20
25
30
Radius

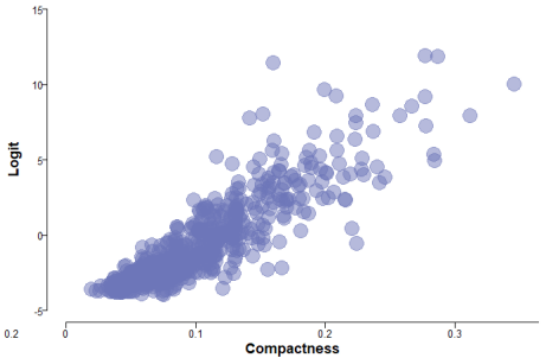

Logit
15
10
5
0
-5
-10
5
10
15
20
25
30
35
40
Texture

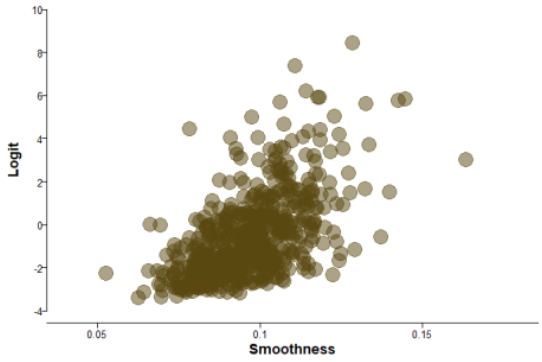

Logit
15
10
5
0
-5
0
0.05
0.1
0.15
0.2
Smoothness

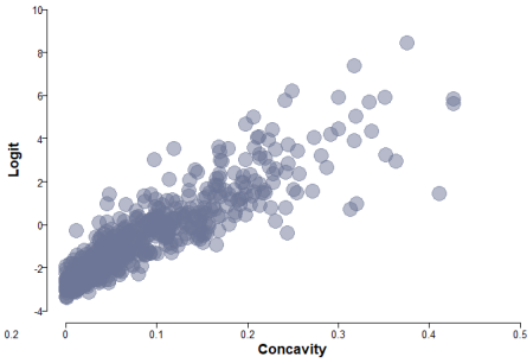

Logit
15
10
5
0
-5
0
0.1
0.2
0.3
0.4
Compactness

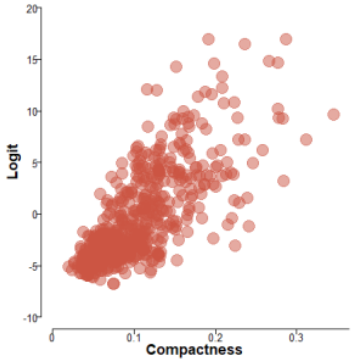

Logit
15
10
5
0
-5
0
0.1
0.2
0.3
0.4
0.5
Concavity

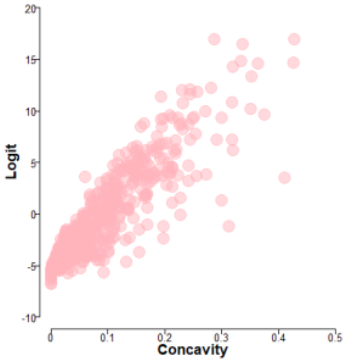

Logit
Smoothness

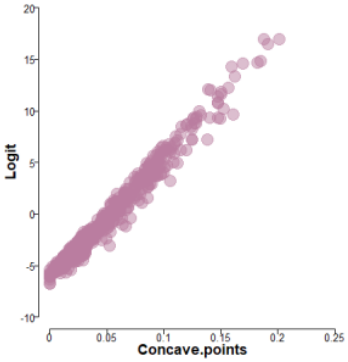

Logit
Compactness

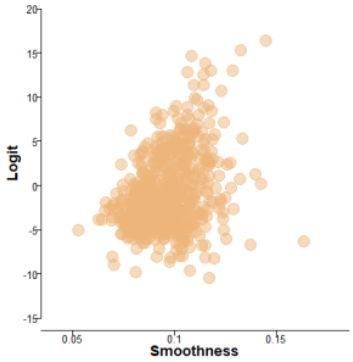

Logit
Smoothness

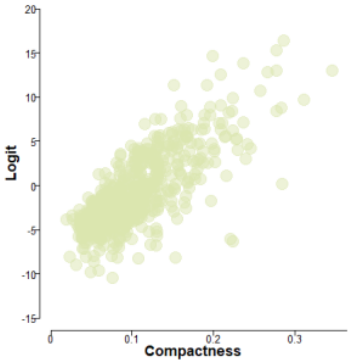

Logit
Concavity

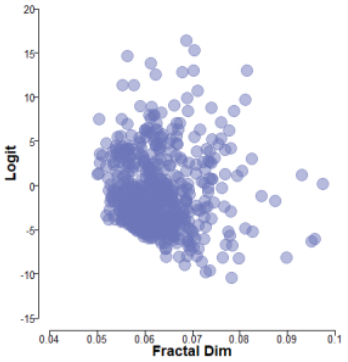

Logit
Compactness

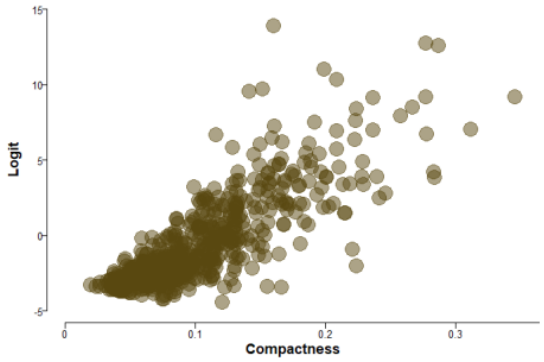

Logit
Concavity

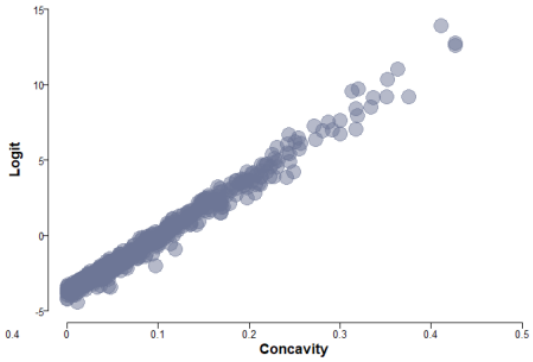

Logit
Concave.points

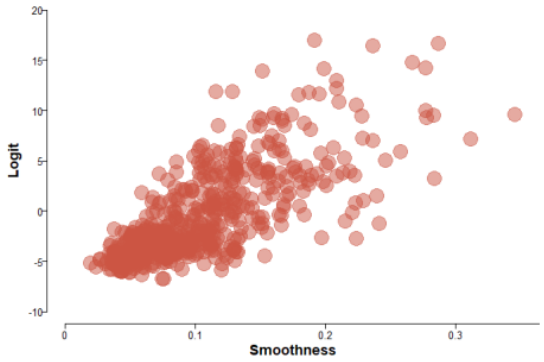

Logit
Smoothness

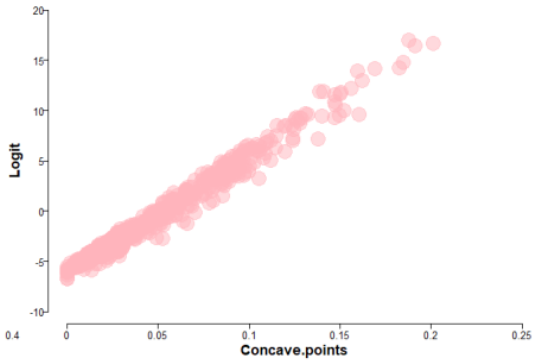

Logit
Compactness

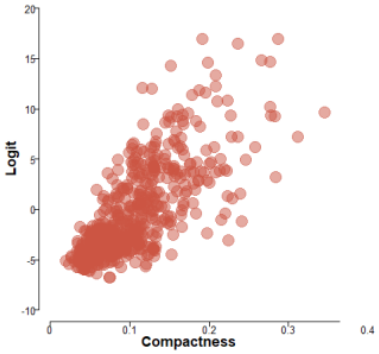

Logit
Fractal Dim

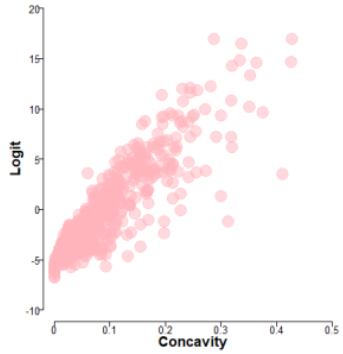

Logit
Compactness

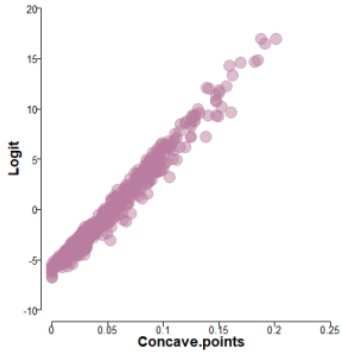

Logit
Concavity

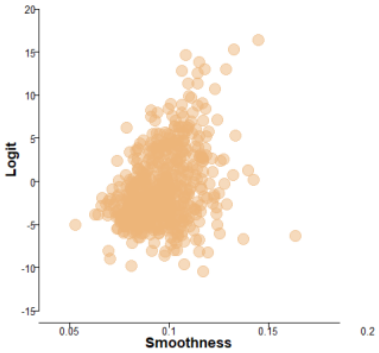

Logit
Smoothness

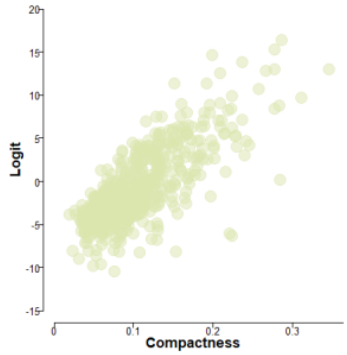

Logit
Concave.points

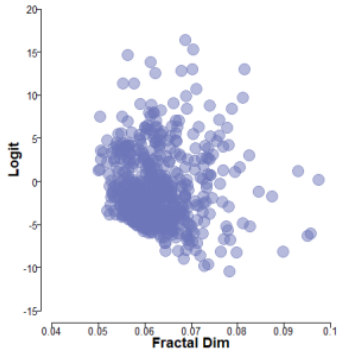

Logit
Compactness

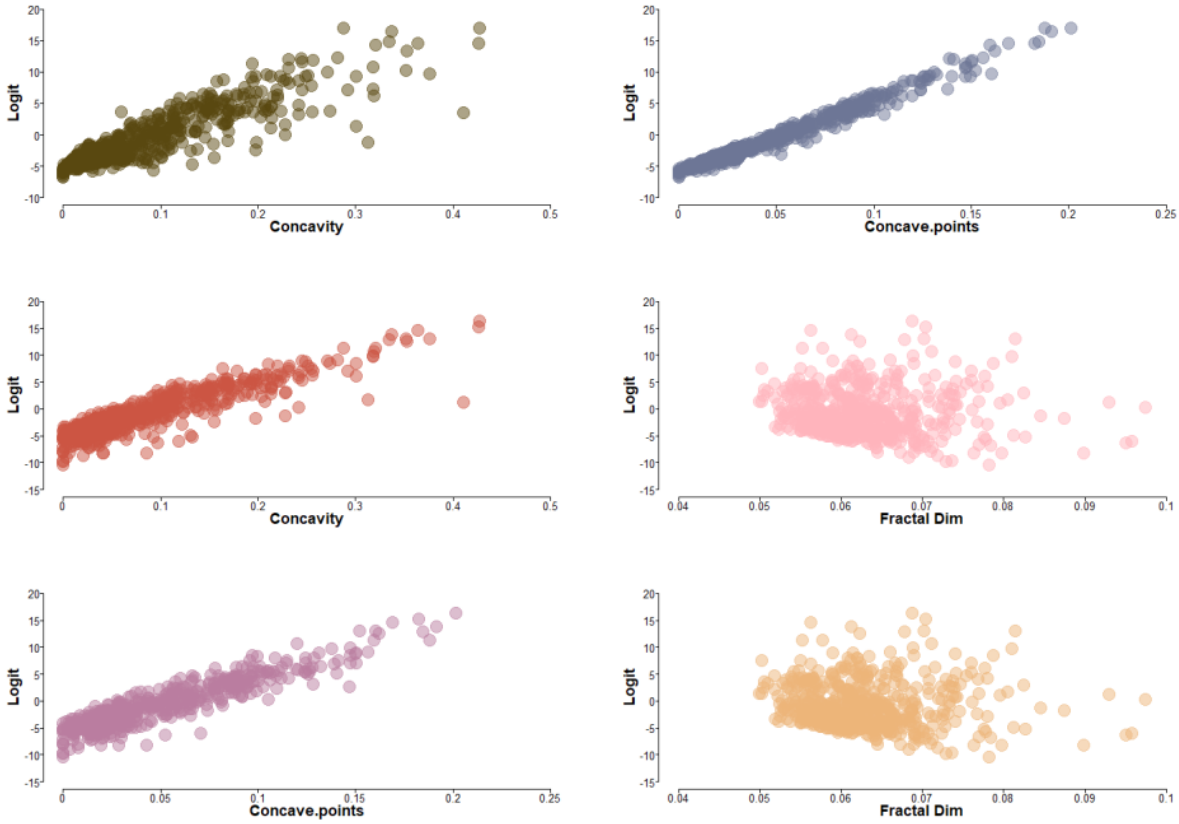

Logit
Concavity


Logit
Concave.points


Logit
Smoothness


Logit
Compactness


Logit
Fractal Dim



From the above graphs we can see that not all the graphs are linear and therefore we can drop some of the models. Models 4 & 5 are dropped due to the smoothness not being linear associated with the logit (Figure c). Model 8 is dropped due to the smoothness & fractal dimension not being linear associated with the logit (Figure d). Model 13 is dropped due to the smoothness & fractal dimension not being linear associated with the logit (Figure f). Models 15 & 16 are dropped due to fractal dimension not being linear associated with the logit (Figure g).

This leaves models 2, 3, 7, 10, 11, 12, & 14 which are the models that we will test as they pass all the assumptions.

Results

Results: Logistic Regression

Models	Log Odds	Z Value	P Value
Model 2: Radius & Texture	Radius: 1.06 Texture: 0.22	Radius: 10.42 Texture: 5.89	Radius: < 0.01 Texture: < 0.01
Model 3: Smoothness, Compactness, & Concavity	Smoothness: 14.6 Compactness: -20.3 Concavity: 48.8	Smoothness: 1.13 Compactness: -2.83 Concavity: 8.93	Smoothness: 0.259 Compactness: < 0.01 Concavity: < 0.01
Model 7: Compactness, Concavity, & Concave Points	Compactness: -25.7 Concavity: 4.36 Concave Points: 135.8	Compactness: -3.56 Concavity: 0.733 Concave Points: 8.96	Compactness: < 0.01 Concavity: 0.465 Concave Points: < 0.01
Model 10: Compactness & Concavity	Compactness: -16.17 Concavity: 47.44	Compactness: -2.61 Concavity: 8.96	Compactness: < 0.01 Concavity: < 0.01
Model 11: Compactness & Concave Points	Compactness: -22.93 Concave Points: 140.51	Compactness: -3.84 Concave Points: 10.1	Compactness: < 0.01 Concave Points: < 0.01
Model 12: Compactness & Fractal Dimension	Compactness: 85.28 Fractal Dimension: -384.2	Compactness: 7.34 Fractal Dimension: 40.1	Compactness: < 0.01 Fractal Dimension: < 0.01
Model 14: Concavity & Concave Points	Concavity: -6.99 Concave Points: 120.88	Concavity: -1.56 Concave Points: 9.13	Concavity: 0.119 Concave Points: <0.01

Table 1. The table above shows the results of the logistic regression done on the 7 listed models.

In Models 7, 11, and 14, concave points demonstrate the strongest relationship with a malignant tumor diagnosis. As the mean of concave points increases, the log odds of a malignant tumor diagnosis increases by more than 100. In comparison to other characteristics, concave points prove to be the best predictor variable of a malignant tumor. Within each model, it proves to be a

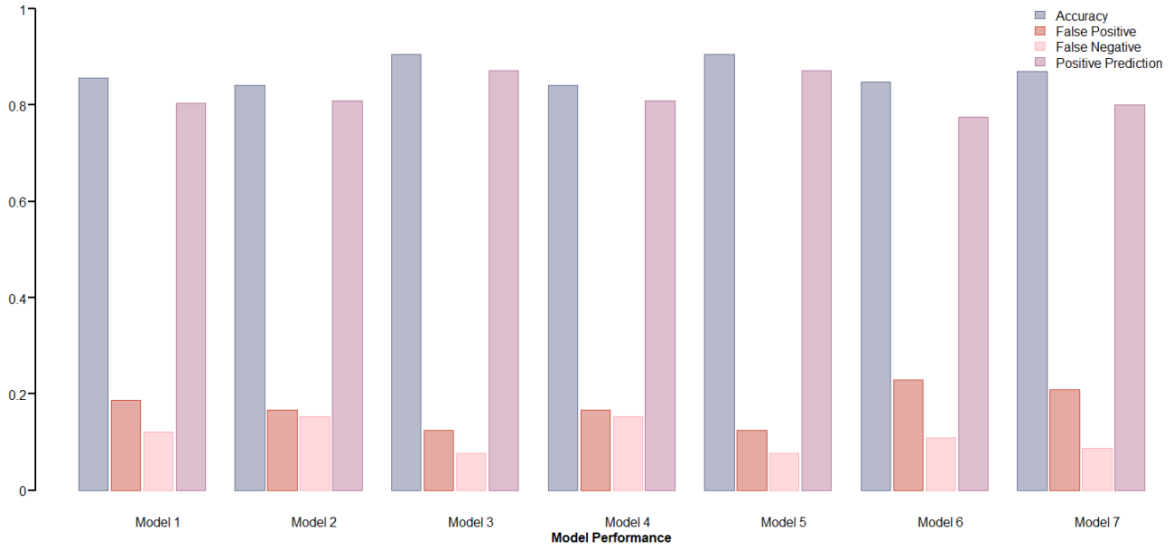
statistically significant variable. Although radius and texture prove to have a positive association with a malignant tumor diagnosis, it is not the key predictor variable when assessing a sample. For Models 3,7,10, 11, there was a strong negative association with a malignant tumor diagnosis, with the exception of Model 12. Therefore, as the compactness of a tumor sample increases, the likelihood of malignant diagnosis may decrease. In each of the models, compactness has a p-value of less than 0.01 and is statistically significant.

Results: Model Performance

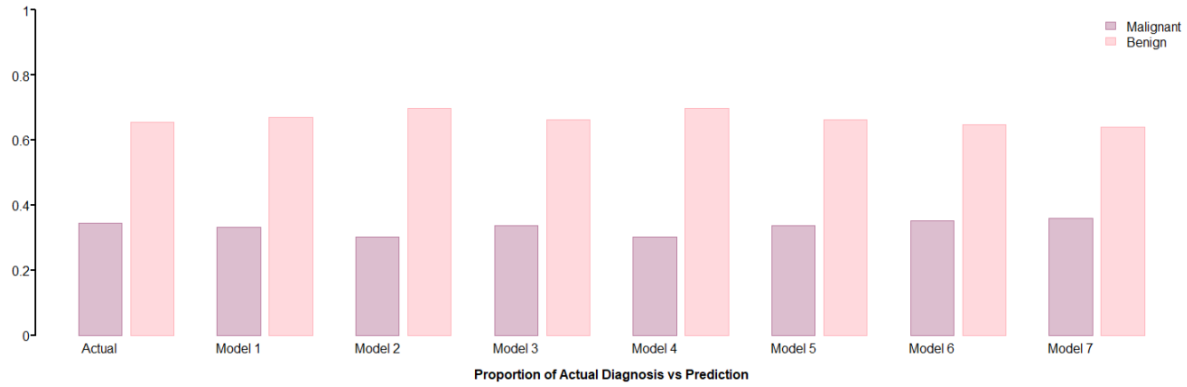
In the table below are the accuracy, false positive rate, false negative rate, and the positive predict rate of the remaining models and of which the results are visualized in figure f.

Models	Accuracy	False Positive Rate	False Negative Rate	Positive Predict Rate
Model 2: Radius & Texture	85.6%	18.8%	12.1%	80.4%
Model 3: Smoothness, Compactness, & Concavity	84.2%	16.7%	15.4%	81%
Model 7: Compactness, Concavity, & Concave Points	90.6%	12.5%	7.7%	87.2%
Model 10: Compactness & Concavity	84.2%	16.7%	15.4%	81%
Model 11: Compactness & Concave Points	90.6%	12.5%	7.7%	87.2%
Model 12: Compactness & Fractal Dimension	84.9%	22.9%	11%	77.6%
Model 14: Concavity & Concave Points	87.1%	20.8%	8.8%	80%

Table 2. The table above shows the accuracy rates calculated from the 7 logistic regression models.



We can see that the most accurate model is model 7 which uses compactness, concavity, & concavity points. This model also has the lowest false positive rate (Matching Model 11), the lowest false negative rate (Matching model 11), as well as the second highest positive predict rate (Matching model 3). As a result it has the predicted diagnoses closest to the actual diagnosis which is illustrated in the bar plot below.



Discussion of Results

Based on the findings of the research study, the null hypothesis must be rejected. Based on the numerous constructed logistic regression, it has been proven that concave points serve as the best predictor variable of a malignant tumor diagnosis, and demonstrates the strongest association of the different variables tested. These results and findings are significant to medical professionals treating breast cancer patients. Radiologists can directly implement these results when testing breast tumor samples. Although breast tumors of a specific size may be alarming, a small breast tumor with numerous concave points can be potentially more dangerous.

Discussion of Limitations

More than 5 logistic regression models could not be done, since these tests did not meet all assumptions of logistic regression. These variables may require a different form of analytical testing to be further studied.

Discussion: Future Directions

Since this dataset was collected in 1995, a similar dataset collected during previous years would likely have other variables to study and test. Also, measurements of these variables may have become slightly more accurate. Additional predictor variables may provide more insight on diagnosing breast tumor samples. As this continues to be a pressing issue, additional research and data collection can provide medical professionals with more insight to improve medical treatments.