Advanced Statistical Modeling Final Project

Diya Vinod

2024-12-06

Project Topic

The objective of this analysis is to assess the relationship between tumor characteristics and the diagnosis of breast cancer, specifically focusing on the potential influence of smoothness and concavity. The primary question being addressed is:

Can the characteristics of smoothness and concavity effectively predict whether a breast cancer diagnosis is malignant ('M') or benign ('B')?

Breast cancer is one of the most prevalent cancers worldwide, and early detection plays a critical role in improving patient outcomes. Accurate diagnosis, which distinguishes between malignant (cancerous) and benign (non-cancerous) tumors, is fundamental to deciding the appropriate treatment course. Tumor characteristics such as smoothness and concavity are important features that may help in determining whether a tumor is malignant or benign.

This project is important because it explores the predictive power of two specific features—smoothness and concavity—in classifying breast tumors as malignant or benign. To answer this question, a Generalized Linear Model (GLM) is used. The GLM is a statistical method that allows for modeling the relationship between a set of predictors (in this case, smoothness and concavity) and a binary outcome (malignant vs. benign). By applying GLM to this dataset, we aim to determine whether these two features alone, or in combination, are effective predictors of malignancy.

Data Description

This dataset was acquired from UC Irvine's Machine Learning Repository. The dataset is titled "Breast Cancer Wisconsin (Diagnostic)" and can be found here: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

Below are the columns in the data:
1) ID number
2) Diagnosis (M = malignant, B = benign)
3) radius (mean of distances from center to points on the perimeter)
4) texture (standard deviation of gray-scale values)
5) perimeter
6) area
7) smoothness (local variation in radius lengths)
8) compactness (perimeter^2 / area - 1.0)
9) concavity (severity of concave portions of the contour)
10) concave points (number of concave portions of the contour)
11) symmetry
12) fractal dimension ("coastline approximation" - 1)

We will be focusing on just the smoothness and concavity for this project. These features are chosen due to their potential association with the malignancy of the tumor. Tumors with irregular contours or rough surfaces tend to be malignant, while benign tumors typically exhibit smoother, more uniform contours.

Statistical and Graphical Methods

We start determining the correlation between the smoothnesss and concavity variables. We obtain a value of 0.52 from the correlation matrix which is moderate correlation. A correlation of 0.52 does not necessarily indicate multicollinearity and we can proceed with including both variables in the GLM.

```
##                 smoothness_mean Concavity_mean
## smoothness_mean       1.0000000      0.5190017
## Concavity_mean        0.5190017      1.0000000
```

Given that the response is binary, we will use a binomial family for the GLM with a logit link function. The parameter for the intercept is -3.3128, the parameter for the smoothness is -5.3719 and the parameter for the concavity is 37.3630. Below are the approximate 95% confidence intervals for smoothness and concavity:

Smoothness: (-26.24923, 15.50543)

Concavity: (31.0663, 43.6597)

Predicted Probabilities vs. Predictors

Figure 1 visualizes the relationship between the predictor variable Concavity_mean and the predicted probability of a diagnosis being malignant, as computed by the fitted GLM model. The plot shows a clear relationship between concavity mean and the predicted probability of malignancy. This is because there is a fairly clear separation between groups. For the output of malignant (1), the predicted probabilities are closer to 1 and for the output of benign (0) the predicted probabilities are closer to 0. The predicted probabilities also follow a sigmoid (S-shaped) curve and the curve has some steepness to it suggesting a strong predictive relationship between concavity_mean and the diagnosis.

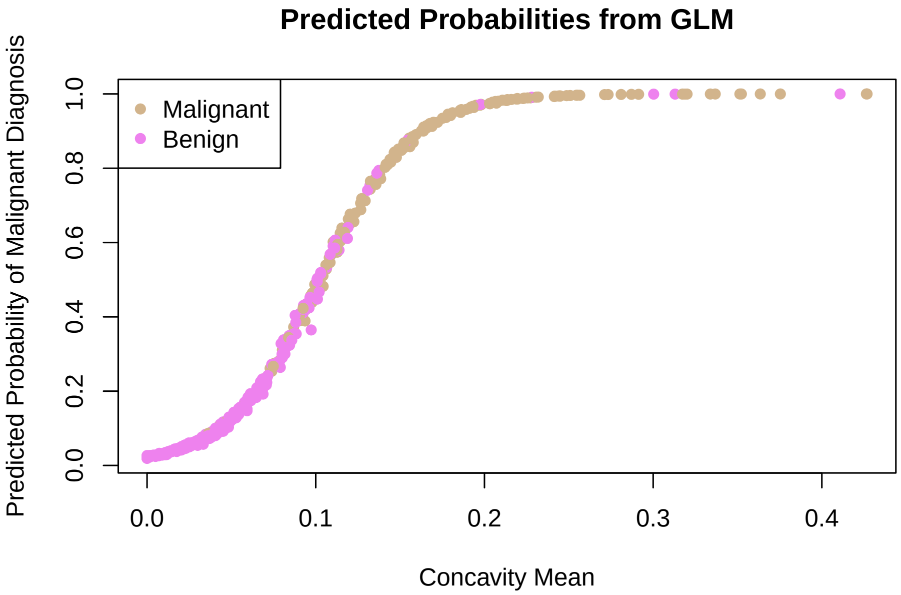

Figure 1

Figure 2 visualizes the relationship between the predictor variable smoothness_mean and the predicted probability of a diagnosis being malignant, as computed by the fitted GLM model. In contrast to figure 1 there is no clear sigmoid curve. The predicted probability of malignancy increases slightly as the smoothness mean grows. However, there is significant overlap between benign (violet) and malignant (tan) predictions, especially at lower smoothness values. There isn't much of a clear separation between groups. This suggests that smoothness alone has limited discriminative power for distinguishing malignant from benign tumors.

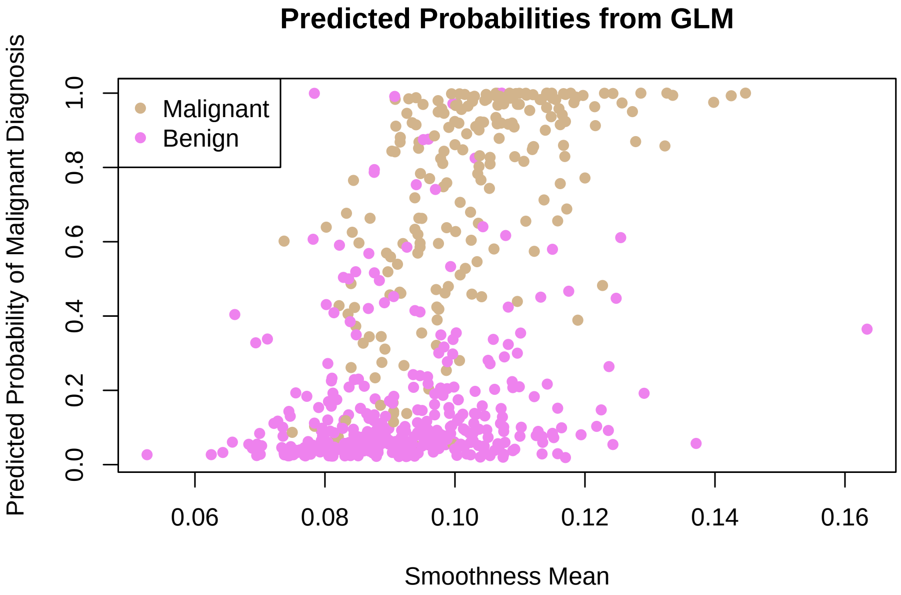

Figure 2

Next we make boxplot (*Figure 3*) of quantile residuals for the fits. It is shown that the residuals are centered approximately around 0 which indicates the model predicts the response variable well for both classes on average. The spread (IQR) of residuals for both classes is similar. Overall, the boxplot of quantile residuals indicates that the model predicts well.

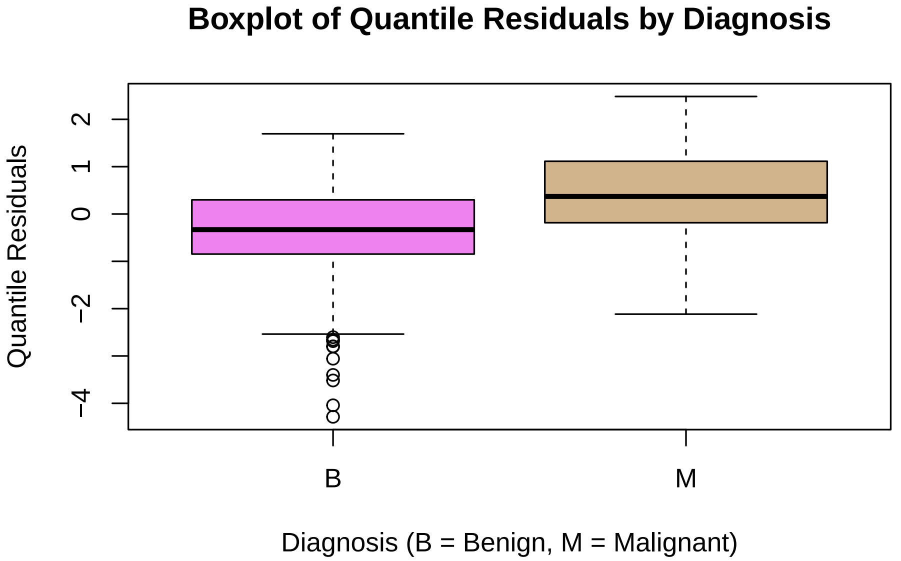

Figure 3

Results and Conclusion

In this project we aimed to answer the question **Can the characteristics of smoothness and concavity effectively predict whether a breast cancer diagnosis is malignant ('M') or benign ('B')?** From the analysis performed, we get the parameter value `-3.3128` for the intercept, the parameter value of `-5.3719` for the smoothness and the parameter value of `37.3630` for the concavity. Our p-value for smoothness_mean is `0.614035` which is not statistically significant using a significance level of 0.05. However the p-value for concavity_mean which is $p < 2e^{-16}$ is less than 0.05; this means the coefficient for concavity_mean is highly significant and has a large positive effect. This is further supported by figure 1 and figure 2 above. The strong predictive power of concavity mean is consistent with the clear trend observed in the figure 1, where malignant cases dominate higher concavity values. The statistical insignificance of smoothness mean matches the lack of a strong relationship in figure 2.

To conclude, the GLM analysis shows that concavity mean is a effective predictor of malignant breast cancer diagnoses, with a clear positive relationship and strong discriminative power. In contrast, smoothness mean is not a significant predictor and provides limited value in distinguishing between malignant and benign cases.

Future Work

One promising avenue for future work involves experimenting with additional predictor variables. While this project focused on smoothness and concavity, there are likely several other variables that could serve as effective predictors of diagnoses. Features such as radius, texture, and perimeter could be incorporated into the model to assess their predictive value. Moreover, it would be worthwhile to evaluate these features individually, excluding smoothness and concavity, to determine their standalone contributions. Exploring an interaction term between smoothness and concavity could also provide insights into whether their combined effect enhances the model's predictive performance.

Another direction for future research would be to investigate whether incorporating non-linear terms for smoothness and concavity improves the model fit. Furthermore, while this study employed a GLM, it could be beneficial to compare its results with those from other modeling approaches, such as Random Forests, Gradient Boosting Machines, or Neural Networks. These comparisons could offer a deeper understanding of how different methodologies perform with the same data.