STAT 406 Project Final Report

Group 3

February 6, 2023

1 Introduction

Globally, the most frequently appearing cancer diagnosis is breast cancer, with 12.5 percent of all new cancer diagnoses annually being breast cancer. It is one of the leading causes of death for women in the United States today. In 2022, about 43,250 women in the United States are expected to die from breast cancer, almost 1 in 8 women will develop an invasive type of breast cancer and will need treatment, and, as of January 2022, there are more than 3.8 million women with a history of breast cancer in the U.S. This includes women currently being treated and women who have finished treatment. Despite all of this, breast cancer death rates have been decreasing since 1989. This decrease is thought to be a result of treatment advances and earlier detection through screening. [1]

One such treatment advance is fine needle aspiration, or FNA. FNA is a common biopsy technique used to aid in the diagnosing of breast cancer. During a FNA, a small amount of breast tissue or fluid is removed from a suspicious area with a thin, hollow needle and checked for cancer cells through digital scans. A computer finds the nuclei in the image, then calculates nuclear size, shape, and textural features and reports their mean, standard errors, and worst values [2]. With these features, a diagnosis is made. FNA is inexpensive, widely used, and proven to be safe and effective.

2 Data Description

The data was collected from a breast cancer study in Wisconsin between 1989 to 1991. FNA procedures were performed on 569 women. Using digital scans of the tissue gathered, images were collected, and certain cytologic features were measured.

The breast tumors were diagnosed as either benign (noncancerous) or malignant (cancerous). There were three categorical variables for each patient:

1. ID Number (Patient assigned identifying number)
2. Diagnoses ('B' is benign and 'M' is malignant)
3. Unnamed-32 (Three measurements were taken for each continuous variable)

For the continuous features, each value was recorded with four significant digits. There are no missing values in the dataset. The dataset was assembled with 10 factors that were thought to cause breast cancer:

1. radius (mean of distances from the center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness ($perimeter^2/area - 1.0$)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension ("coastline approximation" - 1)

2.1 Data Cleaning

In this dataset, three measures were taken for each of the continuous characteristics. The three measurements were the mean, standard error, and worst (mean of the three largest values) of the measurement. This resulted in 30 total measurements for each patient.

The 'ID' and 'Unnamed-32' columns in the dataset are not predictors, and were not included in the model. We chose to only use the mean measurements of the ten factors for our model.

2.2 Data Visualization

To understand the dataset better, we made a pie chart to see the proportion of benign to malignant diagnoses. About 62.7% of the diagnoses were benign, and 37.3% were malignant so we kept this in mind during our procedures (Figure 1). We also created histograms to understand the distribution of the factor means split by diagnoses. Benign diagnoses had overall lower value centers of the data for each of the measurements, but more noticeably so for perimeter, area, concavity, and concave points means. Malignant diagnoses seem to have a greater spread, but there is also less observations compared to benign diagnoses (Figure 2).

3 Research Questions

The aim of this project is to determine whether there is a significant relationship between the diagnosis status (benign or malignant) and any of the measurements gathered by FNA. Which features have the most predictive power in determining diagnosis? The answer might benefit physicians by prioritizing the measurement of the most useful features and eliminating costs and time towards procedures that have less marginal benefit in diagnosis. Our null and alternative hypotheses are as follows:

H_0: There is no relationship between all variables and the diagnosis $\beta_i = 0, i = 1, 2, 3..n$

H_A: There is at least one feature that is related to the diagnosis. At least one $\beta_i \neq 0, i = 1, 2, 3..n$

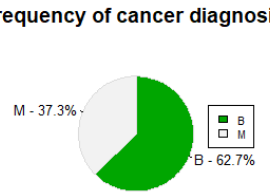

Figure 1: Frequency of Cancer Diagnosis

Figure 2: Histograms of features by diagnosis

4 Model

We chose a logistic regression model to assess the relationship between features and diagnoses since the response variable is binary (benign or malignant).

4.1 Assumption Checking

We first checked the assumptions for the logistic model.

There was no violation of the binary response or no significant outliers assumptions. Each of the 10 variables also have a fairly linear relationship with $logit(diagnosis)$. However, the independence of the variables and the multicollinearity assumption are violated. The variables radius_mean, area_mean, perimeter_mean, and concave.points_mean have correlation coefficients greatly above 0.7 with one another. The same is true for concavity_mean and compactness_mean.

4.2 Full Model

Two full models were built, one with only main effects and another that also includes their two-way interactions.

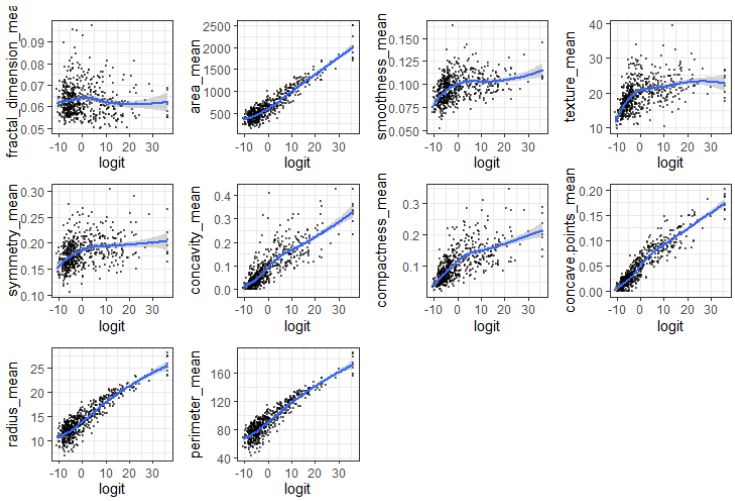

Figure 3: Checking Linearity Assumption against logit of the response

4.2.1 Full Model A - Main Effects Only

This full model we ran included all the original mean features except radius, concavity points, perimeter, and compactness. The two-dimensional measurement area_mean likely provides more information than concavity points and one-dimensional measurements such as radius and perimeter. So we decided it was appropriate to keep the variable area_mean. We also kept concavity_mean over compactness_mean in the model since when running the model with one or the other, concavity_mean proved significant while compactness_mean did not.

Our model, named **Full Model A**, is as follows:

$$\begin{aligned} Logit(\pi(x)) = \beta_0 + \beta_1(\text{texture\_mean}) + \beta_2(\text{area\_mean}) + \beta_3(\text{smoothness\_mean}) \\ +\beta_4(\text{concavity\_mean}) + \beta_5(\text{symmetry}) + \beta 7(\text{fractal.dimension\_mean}) \end{aligned}$$

4.2.2 Model B - Main Effects with Interaction

In addition to the same variables as Full Model A, we also added all two-way interactions of the six factors. The model, called **Full Model B**, is:

$$\begin{aligned} Logit(\pi(x)) = \beta_0 + \beta_1(\text{texture\_mean}) + \beta_2(\text{area\_mean}) + \beta_3(\text{smoothness\_mean}) \\ +\beta_4(\text{concavity\_mean}) + \beta_5(\text{symmetry}) + \beta_6(\text{fractal.dimension\_mean}) \\ +\beta_7(\text{texture\_mean*area\_mean}) + \beta_8(\text{texture\_mean*smoothness\_mean}) \\ +... + \beta_{21}(\text{symmetry\_mean*fractal.dimension\_mean}) \end{aligned}$$

4.3 Reduced Model

To obtain reduced models, we used AIC backwards elimination and selection of the initial model's significant factors.

4.3.1 AIC Backwards Selection

From Full Model A, the model with all the main effects was selected based on the AIC criterion. There were no differences between the AIC suggested model, called **AIC Model A**, and the full model.

$$\begin{aligned} Logit(\pi(x)) = \beta_0 + \beta_1(\text{texture\_mean}) + \beta_2(\text{area\_mean}) + \beta_3(\text{smoothness\_mean}) \\ +\beta_4(\text{symmetry}) + \beta_5(\text{concavity\_mean}) + \beta_6(\text{fractal.dimension\_mean}) \end{aligned}$$

From Full Model B, the AIC criterion included all main effects besides symmetry_mean, and five other interaction terms. The model, called **AIC Model B**, is:

$$\begin{aligned} Logit(\pi(x)) = \beta_0 + \beta_1(\text{texture\_mean}) + \beta_2(\text{area\_mean}) + \beta_3(\text{smoothness\_mean}) \\ +\beta_4(\text{concavity\_mean}) + \beta_5(\text{texture\_mean*smoothness\_mean}) \\ +\beta_6(\text{smoothness\_mean*concavity\_mean}) \end{aligned}$$

4.3.2 Significant Variables

The second pair of reduced models were built using the significant variables indicated by JMP in the original full models. The **Reduced Model A**, contained significant variables from the main effects **Full Model A**, and eliminated symmetry and fractal dimension. The **Reduced Model B**, contained significant variables from the main effects and interactions **Full Model B**, and eliminated symmetry, fractal dimension, and all but two two-way interaction effects. These two reduced models were run in JMP and evaluated.

From **Full Model A**, we obtain the following model, **Reduced Model A**:

$$\begin{aligned} Logit(\pi(x)) = \beta_0 + \beta_1(\text{texture\_mean}) + \beta_2(\text{area\_mean}) \\ +\beta_3(\text{smoothness\_mean}) + \beta_4(\text{concavity\_mean}) \end{aligned}$$

From **Full Model B**, we obtain the following mode, **Reduced Model B**:

$$\begin{aligned} Logit(\pi(x)) = \beta_0 + \beta_1(\text{texture\_mean}) + \beta_2(\text{area\_mean}) + \beta_3(\text{smoothness\_mean}) \\ +\beta_4(\text{concavity\_mean}) + \beta_5(\text{texture\_mean*smoothness\_mean}) \\ +\beta_6(\text{smoothness\_mean*concavity\_mean}) \end{aligned}$$

5 Analysis and Results

To evaluate the fit of each of the six models, we compared their negative log-likelihood values shown in Table 1.

Table 1: **Negative Log Likelihood Values of Various Models**

Model	- Log Likelihood
Full Model A	76.77600
Full Model B	52.64357
AIC Model A	76.77600
AIC Model B	57.50085
Reduced Model A	79.05678
Reduced Model B	73.03539

All models were significant according to a Chi-Square test. When picking which model to use, Full Model B has the lowest negative log-likelihood value 52.64357 and is seemingly the best-fitting model. This is not surprising since Full Model B makes use of the most terms, total of 21, to account for more variability. The next best model would be AIC Model B since it has the second smallest negative log-likelihood value 57.50085, which is still much less than the other three models. However, AIC Model B has almost half as many terms than Full Model B. For practicality purposes, AIC Model B would be most ideal.

6 Conclusion

Our logistic regression models prove there is a significant relationship between some of the factor means, their two-way interactions, and diagnoses, so we can reject the null hypothesis.

Of our models, Full Model B had the smallest negative log-likelihood value, but AIC Model B is our most ideal model to keep the model smaller and more computationally feasible. AIC Model B had a comparably small negative log-likelihood value calculated and the number of terms used to build the model.

For future exploration, we would want to try nonparametric methods to fit the data, since meeting logistic regression assumptions may have resulted in loss of information. Forward feature selection using AIC may also produce a better fitting model. We also suggest fitting a model with the other measurements for the factors, such as "worst", and including more interaction terms if computationally feasible.

7 References

1. Breastcancer.org. "Breast Cancer Facts and Statistics ." Breast Cancer Facts & Statistics 2022, 2021, https://www.breastcancer.org/facts-statistics.

2. The American Cancer Society medical and editorial content team. "Fine Needle Aspiration (FNA) of the Breast." American Cancer Society, American Cancer Society, Jan. 2022, https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast

3. Dataset: Learning, UCI Machine. "Breast Cancer Wisconsin (Diagnostic) Data Set." Kaggle, 25 Sept. 2016, https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.