Breast Cancer Prediction from Cell Physiology

Daniel Miao, Rylan Keniston, Neha Bhattacharyya, Allyssa Weinbrecht

11/21/2020

# Abstract

* The purpose of this study it to determine of certain physical characteristics of cell nuclei are most significant when predicting a cancer diagnosis

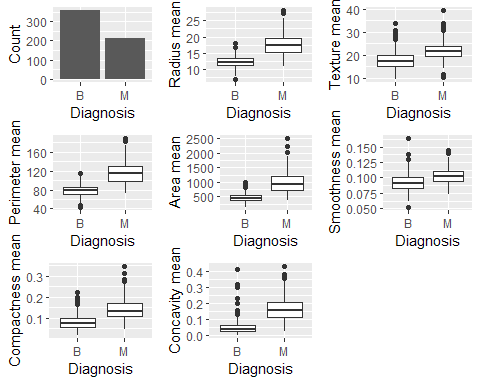
# I. Background and Introduction

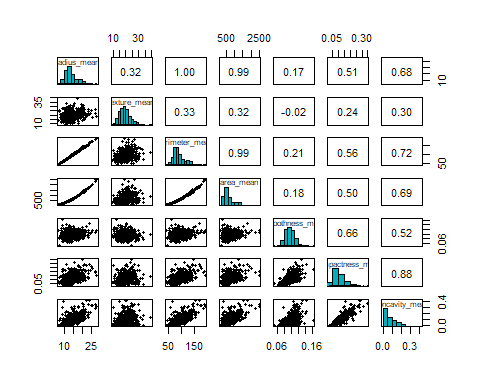
Breast cancer is the second most common form of cancer, contributing 12.3% of the total number of new cases and 25.4% of the number of new cases in women diagnosed in 2018 (excluding non-melanoma skin cancer) (Worldwide cancer data). Early suspicion of breast cancer typically derives from the findings of a clump of irregular cells in the breast, also known as a tumor. Fine-needle aspiration (FNA) and computer-based analytic techniques are used to determine various physical characteristics of the cell nuclei , which are used to diagnose the breast tumors as either benign or malignant. Although many different attributes associated with malignant cytology have been used in previous analyses, we used data available from the UCI Machine Learning Repository containing individual cells properties in our machine learning techniques to classify cells as either benign or malignant.

# II. Methods/Data and Exploratory Analysis

**Data Collection and Variables:** The Breast Cancer Wisconsin Diagnostic data set was originally created and used by individuals in the General Surgery and Computer Science departments at the University of Wisconsin. The set contains 569 observations of 32 characteristics. Our response variable, diagnosis, was determined prior to the measurement of the cell properties, but by establishing models and regressions through analysis, there is potential in the future for a measure of accuracy for prognosis of breast cancer. The seven predictors used to predict the diagnosis response variable, which was mutated to be either 0 for a benign diagnosis or a 1 for a cancer diagnosis, are explained below. The radius mean is the mean measure of the distance from the center to points on the perimeter. Texture mean is the standard deviation of the gray-scale values. Perimeter mean is the distance around the perimeter. Area mean is the area of the cell nuclei. Smoothness mean describes the local variation in radius length. Compactness mean was is by perimeter^2 / area - 1.0. LAstly, concavity mean is the severity of the concave portions of the contour.

**Exploratory Data Analysis:**

 **Fig 1.** From the bar graph in figure 1, we can see that there are around 50% more benign breast cells in the data sample than malignant. Now looking at the boxplots, we see that there are large differences in the physical properties of benign versus malignant (B v M) breast tissue cells. Malignant cells have a greater spread, especially in the top quartile, for most variables and have a greater median for each property than benign cells. Texture mean and smoothness mean show roughly the same spreads, but still an increased median for malignant diagnoses. For radius mean, perimeter mean, area mean, compactness mean, and concavity mean the upper 75% of malignant cells have greater measurements than the lower 75% of benign cells. The capacity for more extreme outliers seems to appear in malignant cells, but surprisingly texture and concavity have a large number of benign outliers as well. Taking these observations into consideration, it can be determined that malignant or cancer diagnosis is associated with an increased size and greater texture. To further examine the data, we should perform a regression analysis of an initial full model.

 **Fig 2.** Since the response variable is categorical and all of our predictors are numerical variables, the exact linearity between the response and each predictor variable is hard to determine. Comparing predictors, radius mean, perimeter mean, and area mean all have highly correlation with each other, which does not come as a surprise since each is a measure associated with size. Other variable combinations with significant collinearity are compactness mean and concavity mean, as well as radius mean, perimeter mean, and area mean each with concavity mean. Smoothness mean seems to be the least correlated with the other predictors. Because almost all the predictors are measuring mostly similar properties of the cell, it would be reasonable that most of the predictors would be moderately correlated with one another. However, this should not be an issue when performing a logistic regression.

# III. Model and Results

**Analytical Methods:**

Initially, a multiple linear regression model was used with diagnosis has the response variable being predicted by all 7 predictors. Examining the graph with our linear model, as well as the plots to check the assumptions that must be met (Appendix A and Appendix B), we quickly determined that a the multiple linear regression model is not the best for for the data and that logistic graph would be better for the categorical response variable. However, the greater the combined value of the predictors, the more likely it seems to be that the cell is malignant.

We created another model, this time using logistic regression. The full regression illustrates the comprehensive graph combining all the simple logistic regression models. The equation of this model is:

π/(1-π) = e^(-24.3696 - 2.6048(radius) + 0.3857(texture) + 0.2462(perimeter) + 0.0692(area) + 136.1192(smoothness) - 14.4426(compactness) + 21.1783 (concavity))

where the output π/(1-π) is equal to the odds that a breast tumor cell should be classified as malignant. The equation can also be written in a different form:

p^hat = (e^(-24.3696 - 2.6048(radius) + 0.3857(texture) + 0.2462(perimeter) + 0.0692(area) + 136.1192(smoothness) - 14.4426(compactness) + 21.1783 (concavity)))) / (1 + e^(-24.3696 - 2.6048(radius) + 0.3857(texture) + 0.2462(perimeter) + 0.0692(area) + 136.1192(smoothness) - 14.4426(compactness) + 21.1783 (concavity))))

where the output p^hat is the probability that a breast tumor cell is malignant.

We used the p-values from the full regression model to determine that texture, area, smoothness, and concavity means are the predictors with statistic significance to the logistic model when using an alpha level of a=0.01. This tells us that the odds of a cell being malignant is most impacted by changes in texture, area, smoothness, and concavity means. The p-values for these predictors are 0, .0692, 0, and .0004, respectively. The exponent e raised to the slope of each predictor gives us the odds, which enhances interpretability.

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

**Final Model and Results:** To improve our logistic regression model, we took a subset of 4 of the original 7 predictors that were the most statistically significant to the original logistic function. The equation of this model is:

log(π/(1-π)) = -30.8875 + 0.3817(texture) + 0.0152(area) + 119.5156(smoothness) + 19.3929(concavity)

All the predictors are very significant at the alpha level a=0.01.

Transitioning from the full logistic model to the improved one, there seems to be a significant improvement in AIC (171.9614 to 168.1136), at moderate costs to both deviance (155.9614 to 158.1136) and log likelihood (-77.9807 to -79.0568), with only minimal impact on the pseudo R^2 value (0.7924 to 0.7896). Overall, the improved model appears to be a significant improvement as it factors in the most significant variables in determining the logistic regression model, reducing the risk of overfitting to the noise in the data. In addition, many of the highly correlated predictors have also been removed as a result of the high p-values from before.

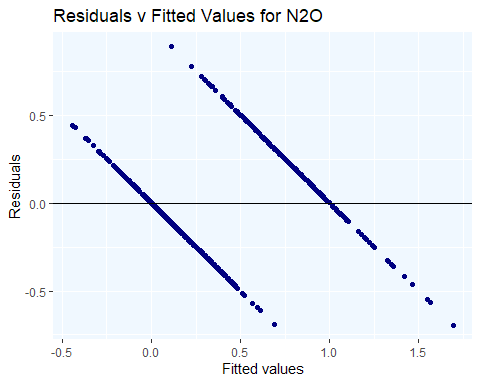
##8. VIF Values

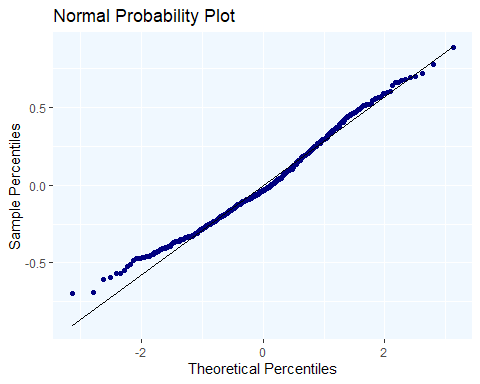
Transitioning from the full to the improved logistic model, we see that some very high VIF values (907.4, 671.2, 109.4) are removed, and other predictors such as smoothness mean have reduced VIF values as a result of eliminating the correlated predictors. In the improved model, all VIF values are 2.12 or lower, which indicates a well-performing model.

# IV. Conclusion and Discussion

# References

# Appendix

**Appendix A**  From the residuals versus fitted values above, we can see that the resiudals are not randomly scattered at all, which indicates that our multiple regression linear model does not satisfy the assumption of equal error variances. Hence, a linear model is not the best model to predict a cancer diagnosis.

**Appendix B**  Looking at the location of clusters and how they follow the normal distribution line, the linear model seems to satisfy the normality of the residuals assumption. However, we cannot say that the linear model performs accurately since in Appendix A we found that not all assumptions were met.