Breast Cancer Diagnosis Modeling

By: Ana Boeriu

UC Davis

Graduate Group Biostatistics

1. Introduction

Cells are the basic unit that make up a human body. They can grow and divide to make new cells as the body needs them. When cells die, new ones will take their place. Breast cancer is a type of cancer that causes uncontrollable growth of abnormal breast cells that form a mass called a tumor. Benign tumors tend to grow slowly and do not spread while malignant tumors grow rapidly, invade, and destroy nearby normal tissues, and spread throughout the body (Cancer, 2019). Because of genetic mutations, cancer cells have different sizes and shapes compared to healthy cells are unable to function properly. To diagnose a tumor, doctors will perform a biopsy where a small amount of tissue is removed and examined. In some cases, where the patient has a breast lesion the doctor may recommend fine needle aspiration (FNA). In this minimally invasive procedure, a very thin hollow needle attached to a syringe is used to collect tissue from the breast lesion (Biopsy, 2020). Once the biopsy is complete, a pathologist will examine the tissue under the microscope and will determine if the breast lesion is a cyst, a benign tumor of malignant tumor. Breast cancer is a complex disease that is still being researched. The continued support for breast cancer awareness has helped create advances in diagnosis and in treatments.

The main goal of this project is to build a model and find the most significant variables when determining tumor diagnosis. We will then try and interpret our predictors to see how they affect tumor diagnosis.

The dataset we will use is the Breast Cancer Wisconsin (Diagnostic) Data Set. This dataset consists of the tumor diagnosis and the characteristics of cell nucleus, computed from 569 digitized images of a fine needle aspirate (FNA) breast mass. Assume that a breast mass can have more than one cell nucleus. There are 10 numeric predictors which describe the size and shape of the cell nucleus for each image. The size of the cell nucleus described by the predictors radius, area, and perimeter while the shape of the cell nuclei is described by the predictors texture, smoothness, concavity, compactness, concave points, symmetry, and fractal dimension. The mean, worst (mean of the three largest values), and standard error of these 10 predictors and were computed for each image, resulting in a total of 30 predictors. The response variable is a binary categorical variable that describes the diagnosis of the tumor as either benign or malignant.

2. Methodology

2.1 Preprocessing

Before we begin modeling our data, we will perform some exploratory data analysis, identify any problems in the data and explain how we addressed these problems. There were no missing values thus, there was no need to address issue. In this dataset 212 women were diagnosed with a malignant tumor while 357 women were diagnosed with a benign tumor. Since the number of benign tumors is not excessively more than the number of malignant samples, our response variable is not unbalanced, and we do not need to adjust for this. We verified for any unusual low or high values however we did not find any. We then addressed the issue of collinearity between our predictor variables. This is an important step which prevents the reduced performance of a model by eliminating redundant predictors. The "worst" and "mean" columns of the data have correlation values greater than 0.87. This is somewhat inevitable since the "worst" columns are just a subset of the "mean" columns. Thus, we removed all "worst" columns from the data. We also removed all "standard error" columns as we generally do not want them as predictors. Furthermore, the predictor "radius mean" and "perimeter mean" have a correlation of 0.997 while the predictors "area mean" and "radius mean" have a correlation of 0.988. This is not surprising because area and perimeter are functions of radius. Thus, we choose radius as our predictor variable to represent the size of a cell, removing the predictors perimeter mean and area mean. We removed the predictor "concave points mean" since it had a correlation of 0.87 with "radius mean". To summarize we have kept the following predictors: radius mean, smoothness mean, fractal dimension mean, symmetry mean, compactness mean and concavity mean. We notice that compactness mean, and concavity mean are highly correlated with each other but not highly correlated with the other predictors. Rather than removing both variables, we fit two logistic regression models with the remaining four of six predictors including compactness mean in the first model and concavity mean in the second. We chose the model that has the smallest deviance which was the model with concavity.

2.2 Logistic regression

Because our response variable of tumor diagnosis has two categories, either benign or malignant, we will be using a logistic regression model. This model is part of an entire class of models called generalized linear models (GLM). The three important components that allow us to relate the response variable to our predictors are the linear predictor, link function and systematic component. In logistic regression the linear predictor is $\eta = X\beta$, the random component is $nY \sim Bin(n, \pi)$, and the link function is $\eta = log\left(\frac{\mu}{1-\mu}\right)$. We will also perform diagnostics, hypothesis test, as well as interpret our results.

2.3 Overall Regression effect

Before we begin model selection, we must first conduct the overall regression test. We want to test the null that all slopes are zero against the alternative that at least one of the slopes is not zero. The test statistic is the difference between the null deviance and the residual deviance of the model and approximately follows a χ_5^2 distribution. By rejecting the null, we indicate that the predictors indeed have an overall regression effect in determining tumor diagnosis. Once we established an overall regression effect, we may proceed to model selection where we want to obtain the most significant predictors in our model.

2.4 Model Selection

Compared to AIC, BIC tends to keep less predictors due to the larger penalty term related to model complexity. Since we are interested in inference, we will focus on using BIC criterion and forwards backwards stepwise selection. These criteria will give us a smaller model that is more interpretable.
We then perform diagnostics to make sure there are no abnormalities with our data. More specifically, we will be examining the goodness of fit and examining leverage and influential points.

2.5 Diagnostic Measures: Goodness of fit.

The Pearson residual plot is used to check if the model fits the data well and the aim is to check if there are any systematic patterns left in the residuals. We plot the Pearson residuals against fitted values, using smoothing splines as visual aids, and check for systematic patterns which signify a lack of fit. After determining there is no obvious lack of fit, we then check for leverage and influential points.

2.6 Diagnostic Measures: Leverage/Influential points

To identify influential data points, we plot the leverage h_{ii} (diagonal of the hat matrix) against the index of the points. If $h_{ii} > \frac{2p}{n}$,where p is the number of coefficients in the model and n is the sample size, then the observation is a suspected leverage point. To detect outliers or influential observations, we will use Cooks distance plot. Points that have a high Cooks Distance are suspiciously influential points and we may need to delete them if they greatly affect our model fit.

3. Results

3.1 Overall Regression Effect

We want to test the following:

$$H_o: \beta_{radius\ mean} = \beta_{smoothness\ mean} = \beta_{symmetry\ mean} = \beta_{fractal\ dimention\ mean} = \beta_{\text{concavity}\ mean} = 0$$

$$H_a: at\ least\ one\ \beta_i \neq 0$$

Since our p-value of 2.2×10^{-16} is less than any significant alpha of 0.01, 0.05, 0.1 we reject the null hypothesis and conclude that at least one β_i will have a significant effect with tumor diagnosis. This suggests that the predictors have an overall effect in determining tumor diagnosis

3.2 Final Model

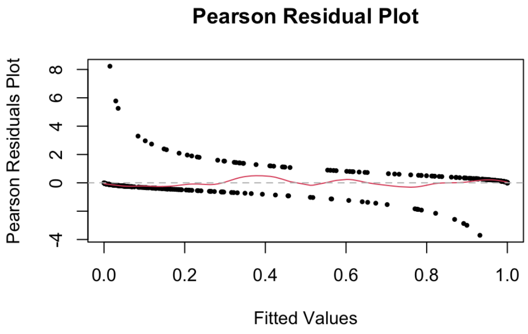
After performing model selection, our final model, with main effects only is:

$$\boxed{\eta = -23.783 + 1.068X_{radius\ mean} + 20.284X_{concavity\ \ mean} + 64.89X_{smoothness\ mean}}$$

β_i	Estimate	P-value
Intercept	-23.7834	$< 2.2 \times 10^{-16}$
Radius mean	1.0676	$< 2.2 \times 10^{-16}$
Concavity mean	20.2841	3.24×10^{-8}
Smoothness mean	64.8905	4.07×10^{-5}

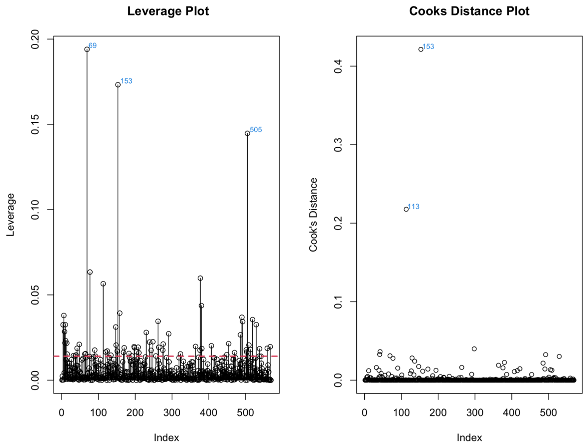
Notice that radius mean is the most significant predictor followed by concavity mean and smoothness mean when determining tumor diagnosis. Furthermore, these predictors all have a positive effect on determining if the tumor is malignant or benign. In other words, as the average radius, average concavity, and average smoothness of cell nuclei in a tumor increase, the probability of malignancy increases.

3.3 Goodness of fit



There is no systematic pattern thus we conclude that there is no obvious lack of fit.

3.4 Outliers and influential Points



From the leverage plot, we observe several influential points greatly exceeding the threshold $\frac{2p}{n}$ denoted by the dotted red line. Observations 69, 153, and 505 are highly influential, while observations 153 and 113 have a large cook's distance. We will check to see if our regression line is affected when removing these points. We found that even after removing these points and performing the same type of analysis as in part 3.1 there was no significant difference in our p-values. Thus, since the regression line is not significantly affected, we will not remove any points.

Generally, the next step would be to test for interaction terms however there is no biological reason indicating that the effect of one of these predictors depends on another predictor. Thus, we did not test for the significance of an interaction term and our final model is shown in 3.2

4. Discussion

In this project we found that the average radius, mean smoothness and mean concavity are all significant when determining the tumor diagnosis for a breast mass. It is not surprising that we see radius mean as the most significant predictor since in general cancer cells have larger nuclei than normal cells (Cancer Treatment Centers of America (CTCA) Comprehensive Cancer Care Network, 2018). But just looking at size doesn't tell the whole story. A small tumor can be aggressive while larger tumor is not – or it could be the other way around (*Size of the Breast Cancer*, 2020). Genetic mutations in breast cancer cells cause the cell nuclei to have different sizes and shapes compared to healthy cells. This can cause functional and structural irregularities which prevent the cells from properly functioning. Note that smoothness measures the nuclear contour, while concavity refers to the number of indentations a cell nucleus has. As with all the shape variables, a higher value corresponds to a less regular contour and thus to a higher probability of malignancy (Street, et al., 1992). Furthermore, the higher the average number of indentations is for cell nuclei in a tumor the more likely the tumor is malignant. To conclude, a breast mass is more likely to be malignant tumor if the average radius of cell nuclei is large and if the average smoothness and average concavity of cell nuclei have a high number of morphological irregularities.

For future studies we could expand this concept further by including the number of cell nuclei each breast mass had. This could give us an even better understanding in exploring if number of cell nuclei is a significant characteristic.

References

Biopsy. (2020, August 24). Cancer.Net. https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures/biopsy

Breast Cancer Wisconsin (Diagnostic) Data Set. (2016, September 25). Kaggle. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Cancer. (2019, December 3). Stanford Health Care. https://stanfordhealthcare.org/medical-conditions/cancer/cancer.html

Cancer Treatment Centers of America (CTCA) Comprehensive Cancer Care Network. (2018, February 20). *How does cancer do that? Sizing up cells and their shapes*. Cancer Treatment Centers of America. https://www.cancercenter.com/community/blog/2018/02/how-does-cancer-do-that-sizing-up-cells-and-their-shapes

Muller, H. G. (2022, January 12). *UC Davis Canvas Discovery* [Lecture Notes]. Canvas Discovery. https://canvas.ucdavis.edu/courses/655751/files/?preview=15312915

Size of the Breast Cancer. (2020, September 21). Breastcancer.Org. https://www.breastcancer.org/symptoms/diagnosis/size

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1992). Nuclear Feature Extraction for Breast Tumor Diagnosis. *International Symposium on Electronic Imaging: Science and Technology*, *1905*, 861–870. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.707&rep=rep1&type=pdf

R appendix

Preprocessing

```
bcancer <-
read.csv("Documents/Winter22/STA223
/data-2.csv", header=TRUE)
bcancer$X=NULL
bcancer$id = NULL
bcancer$diagnosis =
as.factor(bcancer$diagnosis)
View(bcancer)
summary(bcancer)
table(bcancer$diagnosis)
plot(bcancer[,c(2,4,5,12,14,15,
22,24,25)])
plot(bcancer[,c(2,4,5)])
plot(bcancer[,c(12,14,15)])
plot(bcancer[,c(22,24,25)])
library(ggcorrplot)
correlationData = cor(bcancer[,2:31])
ggcorrplot(correlationData,type =
"lower")
corrVals = data.frame(correlationData)
corrVals$Xname = rownames(corrVals)
corrVals = corrVals[,c(31,1:30)]
rownames(corrVals) =NULL
View(corrVals)
View(corrVals[1:11, 1:11])
corrVals = sort(corrVals, decreasing =
TRUE)
bigCorr = which(corrVals[,2:31] > 0.5 |
corrVals[,2:31] < -0.5)
View(bigCorr)
library(dplyr)
bcancer_mean_only_measure= bcancer
%>% select(-contains(c("perimeter",
"area","worst","se")))
names(bcancer_mean_only_measure)
View(cor(bcancer_mean_only_measure[
,c(2,5:8,3,4,9)]))
################ run the 2 models
##############
## run 2 logistic model with the 2
correlated variables and choose one
## with smallest deviance
m1_compact = glm(diagnosis~
radius_mean + texture_mean +
smoothness_mean +
         fractal_dimension_mean +
symmetry_mean + compactness_mean,
         data =
bcancer_mean_only_measure, family
="binomial")
m2_concav = glm(diagnosis~
radius_mean + texture_mean +
smoothness_mean +
      symmetry_mean +
fractal_dimension_mean +
concavity_mean,
   data = bcancer_mean_only_measure,
family ="binomial" )
m1_compact$deviance
m2_concav$deviance
```

Regression effect

```
init_model = m2_concav
```

```
empty_model = glm(diagnosis~1,
family = "binomial", data =
bcancer_mean_only_measure)
anova(empty_model, init_model, test =
"Chi")
```

Final Model

```
fbBIC = step(empty_model, scope =
list(lower = empty_model, upper =
init_model),
            direction= "both",
k=log(nrow(bcancer_mean_only_measu
re)))
fbBIC$coefficient
anova(empty_model, fbBIC, test =
"Chi")
summary(fbBIC)
```

Goodness of Fit

```
res.P = residuals(fbBIC,
type="pearson")
res.D = residuals(fbBIC,
type="deviance") #or residuals(fit), by
default
#boxplot(cbind(res.P, res.D), names =
c("Pearson", "Deviance"))
plot(fbBIC$fitted.values, res.P, pch=16,
cex=0.6, ylab='Deviance Residuals',
    xlab='Fitted Values', main =
"Residuals vs Fitted Values")
lines(smooth.spline(fbBIC$fitted.values,
res.D, spar=0.9), col=2)
abline(h=0, lty=2, col='grey')
```

Outliers and influential Points

```
cooks = cooks.distance(fbBIC)
leverage = hatvalues(fbBIC)
par(mfrow = c(1,2))
leverage = hatvalues(fbBIC)
plot(names(leverage), leverage,
ylab="Leverage", xlab="Index",
    type="h", main =" Leverage Plot")
points(names(leverage), leverage)
abline(h=2*length(fbBIC$coefficients)/
nrow(bcancer_mean_only_measure),col
=2,lwd=2,lty=2)
plot(cooks, ylab="Cook's Distance",
cex=0.9,main = "Cooks Distance Plot")
susPts <-
as.numeric(names(sort(cooks[infPts],
decreasing=TRUE)[1:2]))
text(susPts, cooks[susPts], susPts,
adj=c(-0.1,-0.1), cex=0.7, col=4)
```