Team Cinco

Analyzing Breast Cancer

Authors: Gautham Arra, Harika Putti, Pavan Kumar Madineni, Jeevan Reddy Rachepalli, Priyadarshini Vijjigiri

---

ABSTRACT

Diagnosis of Breast cancer is extremely expensive in United states. Since a large demographic of women are affected by it in their lifetime, its critical to find a cheaper and a reliable way of detecting Breast Cancer. We propose a model that can do this. Using information about the attributes of a tumor that is present in women's chest, we try to analyze weather that tumor is malignant or benign i.e., if it causes cancer or not. We explore different models that can predict the chances of a cancerous tumor using these attributes as our predictor variables.

INTRODUCTION

1 in 8 women in the U.S is diagnosed with Breast cancer at least once in their life time [(BCOrg)]. If a cancerous tumor is detected in the early stages, it can almost always be cured. If detected in the first stage, it can be cured 99% of the time and about 86% of the time in stage 2. Most of the breast cancer cases are due to genetic mutation that happens as women age. It's is rarely hereditary and hence cannot be easily predicted. It's important for women to get regularly examined to reduce the risk. Currently, women can only undergo screening tests which are physical examinations and mammograms. If any abnormalities are found, then they have to undergo a bunch of very expensive tests for diagnosis. Most of the time these tests have a risk of false-negatives, meaning that the results of these tests depend on the expertise of the examiner and might not always be the most reliable. If there are false-positives, then it leads to more expensive and time-consuming tests which can cause stress and anxiety in the patient [(cdc)].

We know that there are two types of tumors- Benign and Malignant. A Benign tumor is not cancerous and Malignant is cancerous. Women tend to have doubt if the lumps that are formed in their chest are cancerous or not. There are many ways to test this- some tests take longer time and some tests are costly. One of such tests is Cytology- testing the abnormality of the tumor tissues under a microscope. Once we obtain the attributes of the tumor our objective is to find the main attributes that can help identify the cancerous ones. We then build a model that predicts if the tumor cell is cancerous. This would eliminate the need for physical examination and can be more reliable.

DATASET DESCRIPTION

The data set was collected by University of Wisconsin Hospitals, Madison from women who had undergone laboratory test to find out whether the lump formed in their chest is a Benign or a Malignant tumor. There are 9 attributes and one class- whether the tumor is malignant or not. The attributes are –

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses

Each of these describes the tumor in the chest. Depending on the types and measurements of these attributes, each of these were categorized on a scale of 1-10, higher the value, greater the chance of it being malignant. This scale was determined by the University of Wisconsin Hospitals using prognostic and diagnostic data of these tumors [(O. L. Mangasarian, 1990)]. It would have been better if they have provided the raw data instead of the scaled data from 1-10. We had a total of 699 observations out of which 16 had missing values (we dropped those because they constituted < 2.5% of the data). The class variable represented 2 for benign and 4 for malignant. And the distribution was 65.5% of the cases as benign and 34.5 as malignant.

VARIABLE SELECTION

We first plotted a correlation heat map to see the dependencies of one attribute to the other (Figure 1). From the graph, it is clear that all the variables are positively correlated to each other. The most important is their relationship with the class variable and clearly all of the predictors are positively correlated with it. We wanted to find the most important attributes that can help detect malignant tumors. That way researchers can find optimal ways of extracting information about those attributes alone. Also, it pretty tenuous to perform exploratory data analysis using 9 variables since there is always a threat of multicollinearity. So, we performed a series of variable elimination techniques to help us determine these most important attributes.

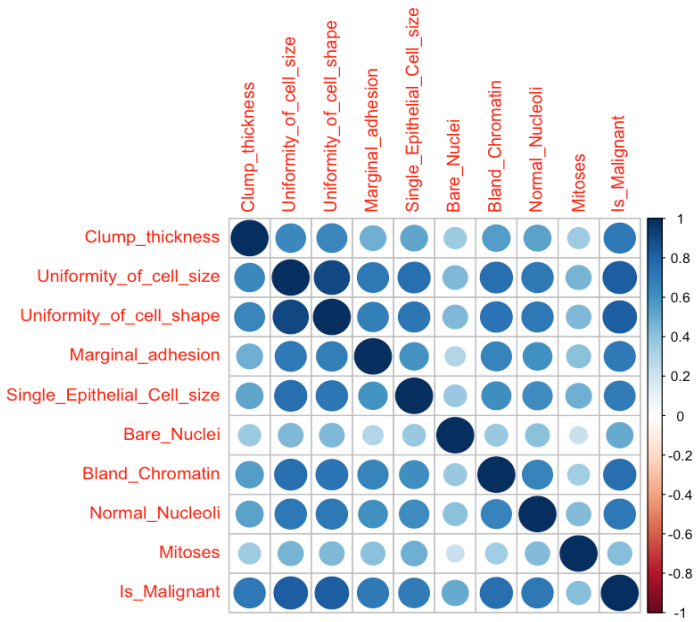

Figure 1: Correlation Heat Map

The first step of elimination was using the Tie-fighter plot.

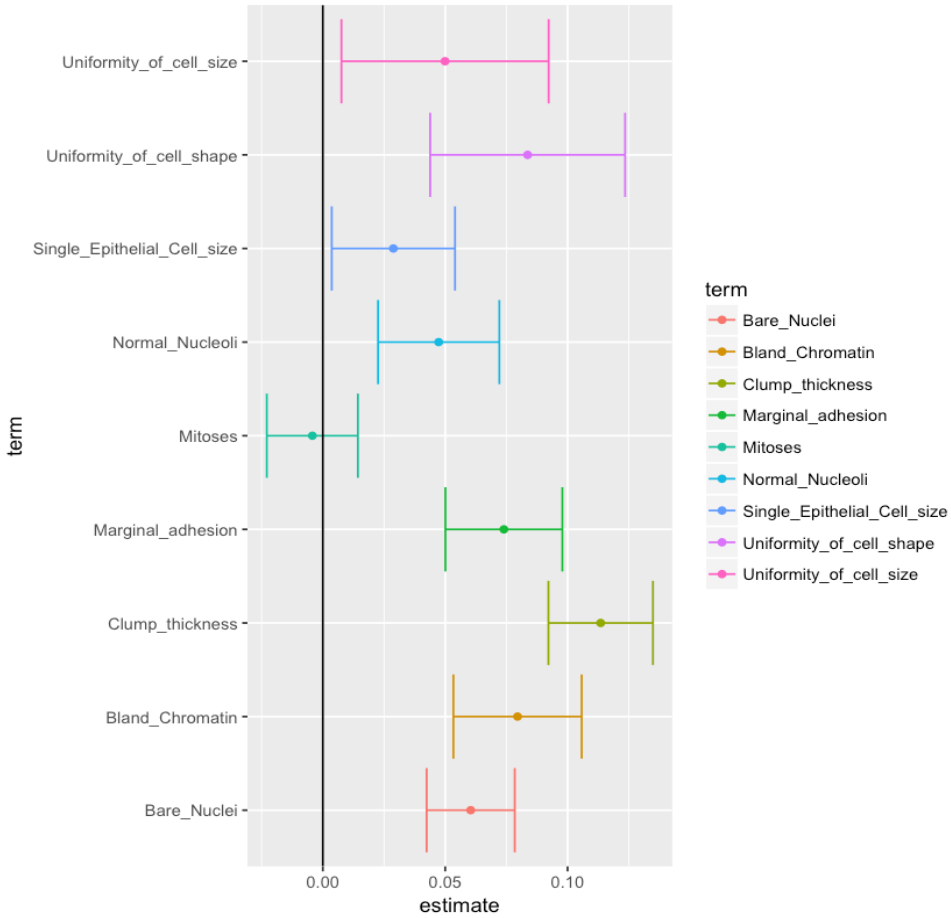

Figure 2: Tie-fighter plot

After setting the significance level to 0.01, we found that variables like Mitoses, Single Epithelial Cell Size, Normal Nucleoli and Uniformity of Cell Size might not be significant for our model. To confirm this and pick the top 4 variables we have done Backward elimination with Logistic Regression. In this process we observe the changes in residual deviance by eliminating the variables one by one. We only eliminate the ones that cause minor change the residual deviance or ones that don't increase it substantially. We finally settled on the following 4 variables –

- Clump Thickness
- Uniformity of Cell Shape
- Marginal Adhesion
- Bland Chromatin

UNDERSTANDING THE VARIABLES

Clump thickness:

Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayers.

Uniformity of cell shape:

Cancer cells tend to vary in size and shape. So, uniformity of cell size/shape points in a benign direction.

Bland Chromatin:

Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser.

Marginal adhesion:

Normal cells stick together while cancer cells lose their ability to do so. So, loss of adhesion is a sign of malignancy.

EXPLORING THE VARIABLES

We compared each of the predictor variable with the response variable to see that difference in the distributions. Since the number of benign cases were substantially more than the number of malignant cases, we expected the graphs to be skewed in one direction more than the other. We could infer the following from the graphs (Figure 3) –

- Distribution of benign cases are all right-skewed meaning when the values are low the chances of the tumor being malignant are less.
- Distribution of malignant cases depending on Marginal Adhesion for is surprisingly also right skewed (till value 9) instead of left skewed. This was particularly peculiar because

we had previously thought that the malignant cases would increase as the values for these predictors increase.

- The distribution of uniformity of cell shape for malignant cases is almost uniform meaning we need to be watchful of this variable because even small values of it might lead to cancer
- When the values of all these predictors is 8 the cells are definitely malignant/cancerous.
- When the clump thickness is high, there are higher chances of the cells being cancerous.
- When the uniformity of cell shape and the Marginal Adhesion is <3 there is higher chances of the cells being benign.

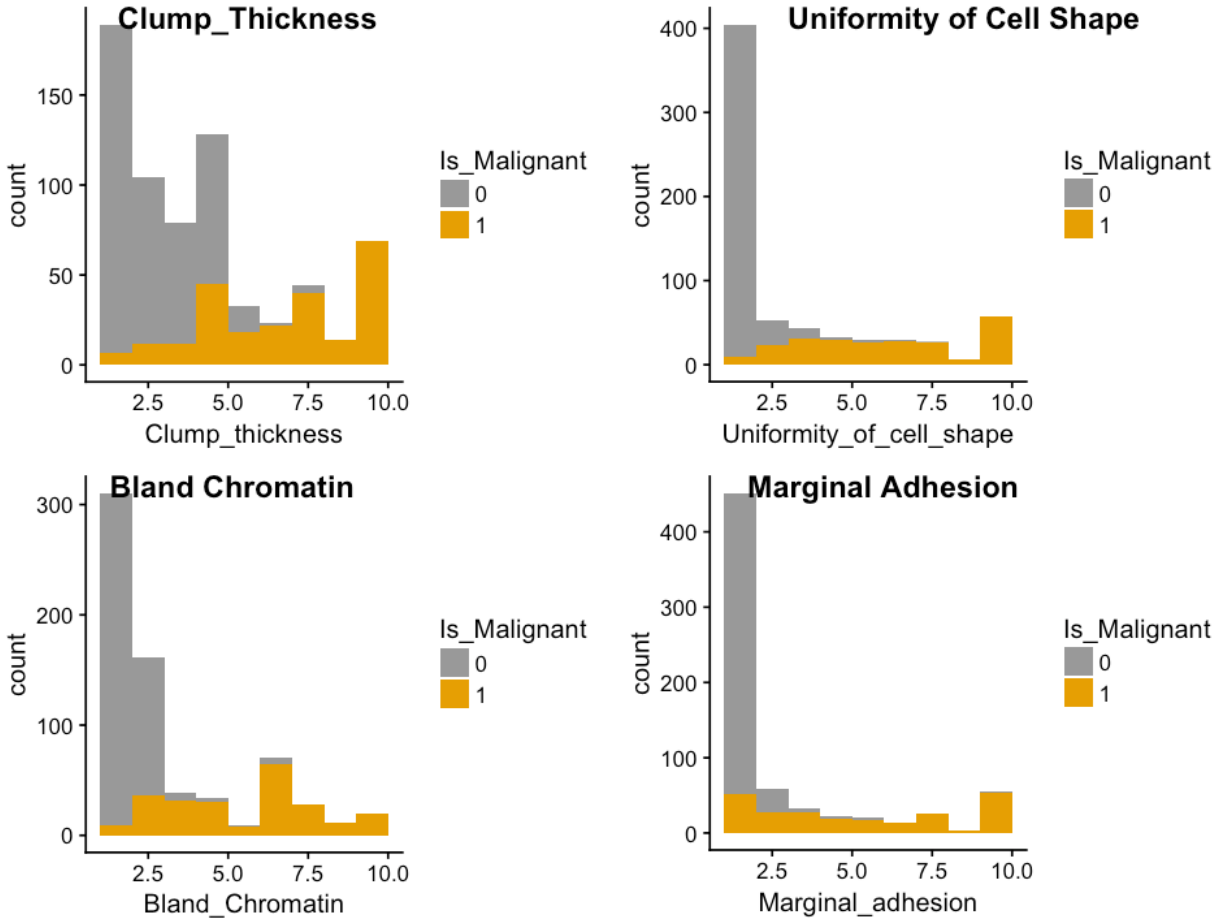

Figure 3: Predictor Variables vs Class Variable

AVERAGE CHANGE IN MALIGNANCE

Each of the explanatory variables is compressed to a scale between 1 to 10. In order to estimate the severity of each attribute on this compressed scale, the average number of malignant cases are evaluated. We already know that as the values of the predictor variables increase, the chances of a malignant tumor increase as they are positively correlated. So, we wanted to see at what pace this probability grows.

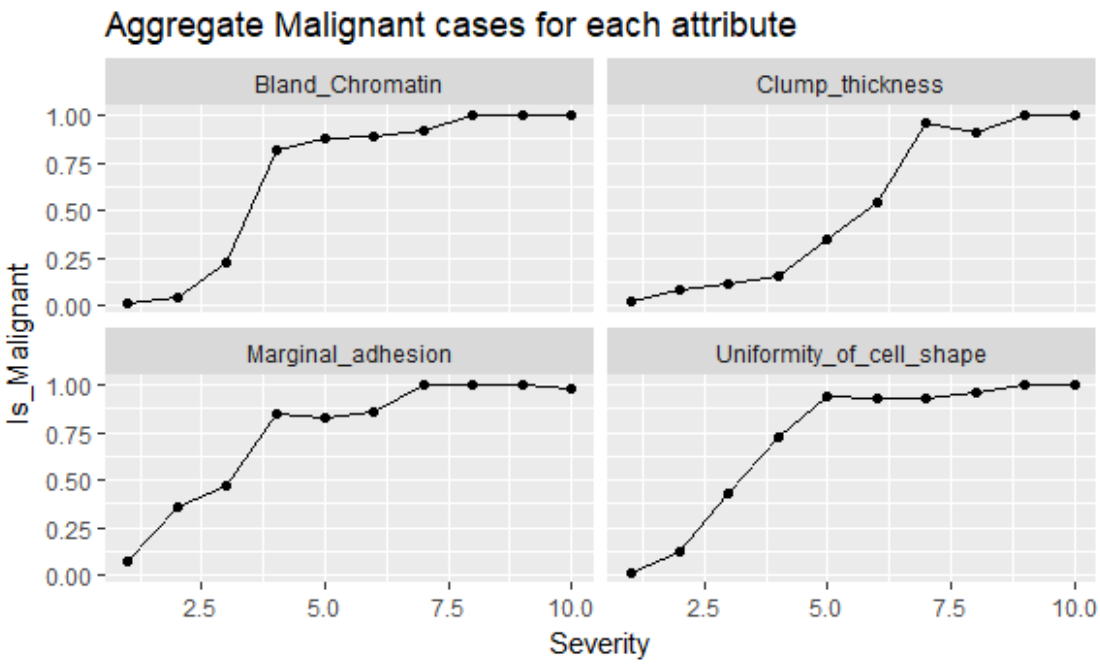

Figure 4: Average change in Malignance

Drawing inferences from this –

- It is evident that any value over 7 for any attribute is considered to be fatal since this range of value has a malignancy rate over 90%.
- It can also be observed that a rapid increase in the number of malignant cases takes place in the range 2 – 5 for Bland Chromatin, Marginal Adhesion and Uniformity of cell shape.
- So, Clump thickness is only fatal when its values are above 7 but for the rest of the variables if they're above 4-5 then there is cause for panic.

TWO WAY DISTRIBUTIONS

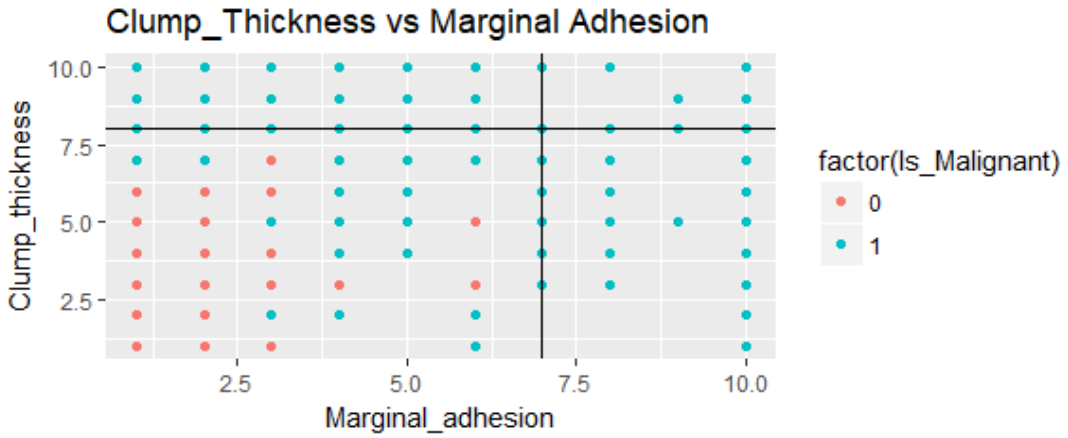

Figure 5: 2-Way distribution of classes

To check the distribution of the class variable based on 2 of the predictors, we drew these plots. Like we saw earlier, the values above 7-8 for each of the predictor variable means there is a definite chance of a cancerous tumor. But there is a lot of uncertainty for the values the lie between 3 and 6.

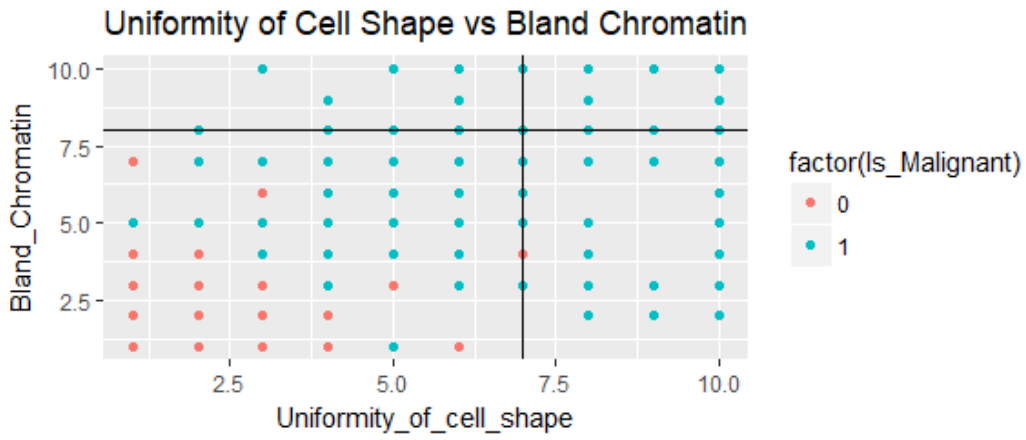

Figure 6: 2-way distribution of classes

INTERACTIONS BETWEEN PREDICTORS

Before fitting a model, we checked all possible interactions that we could incorporate in the model.

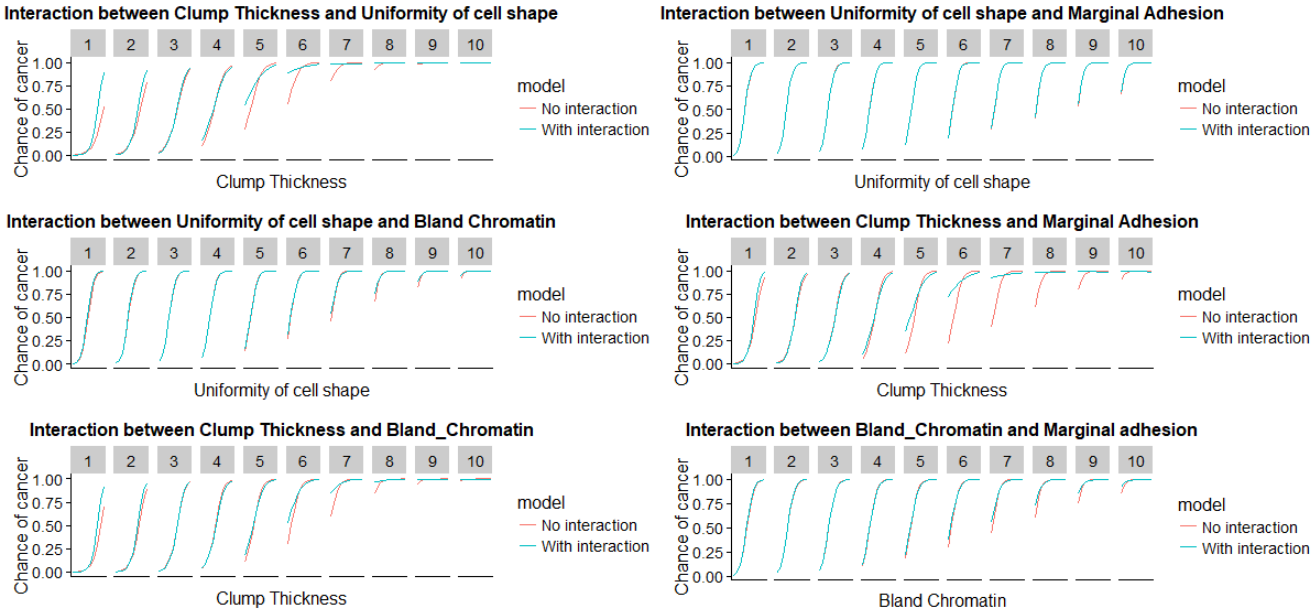

Figure 7: Interactions between predictors

We found strong interactions between Clump thickness and other predictors, and the rest of the predictors had very strained interactions. So, we decided to include only the ones with clump thickness.

One other thing we noticed is that there is much more interaction for the higher values of the predictors and very subtle interaction for the smaller values. Meaning that when all the values are high there are higher chances of cancer cells(which is obvious).

FITTING THE MODEL

We fit a Logistic regression model using glm and the survey weights and "quasibinomial" as the family. The model gave the following results –

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.81582  -0.08713  -0.02247   0.06079   3.00260

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -14.92552    2.40593  -6.204 5.52e-10 ***
Clump_thickness                          1.58569    0.38883   4.078 4.54e-05 ***
Marginal_adhesion                        0.94342    0.31700   2.976  0.00292 **
Uniformity_of_cell_shape                 1.25862    0.43419   2.899  0.00375 **
Bland_Chromatin                          1.08953    0.38048   2.864  0.00419 **
Clump_thickness:Marginal_adhesion       -0.10144    0.04981  -2.037  0.04168 *
Clump_thickness:Uniformity_of_cell_shape -0.09581    0.06822  -1.404  0.16020
Clump_thickness:Bland_Chromatin          -0.07267    0.06364  -1.142  0.25353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 121.97  on 675  degrees of freedom
AIC: 137.97

Number of Fisher Scoring iterations: 8
```

The results indicate –

- All the variables influence our response variable positively.
- They all are significant since the p values are < 0.05
- The residual deviance is pretty good since the variables decrease the deviance by 722.38 with a loss of 7 degrees of freedom.

The residual plot of the model looks pretty good even for values = 1. This means our model will give reliable results as we had hoped. We then drew plots to see how the trends vary.

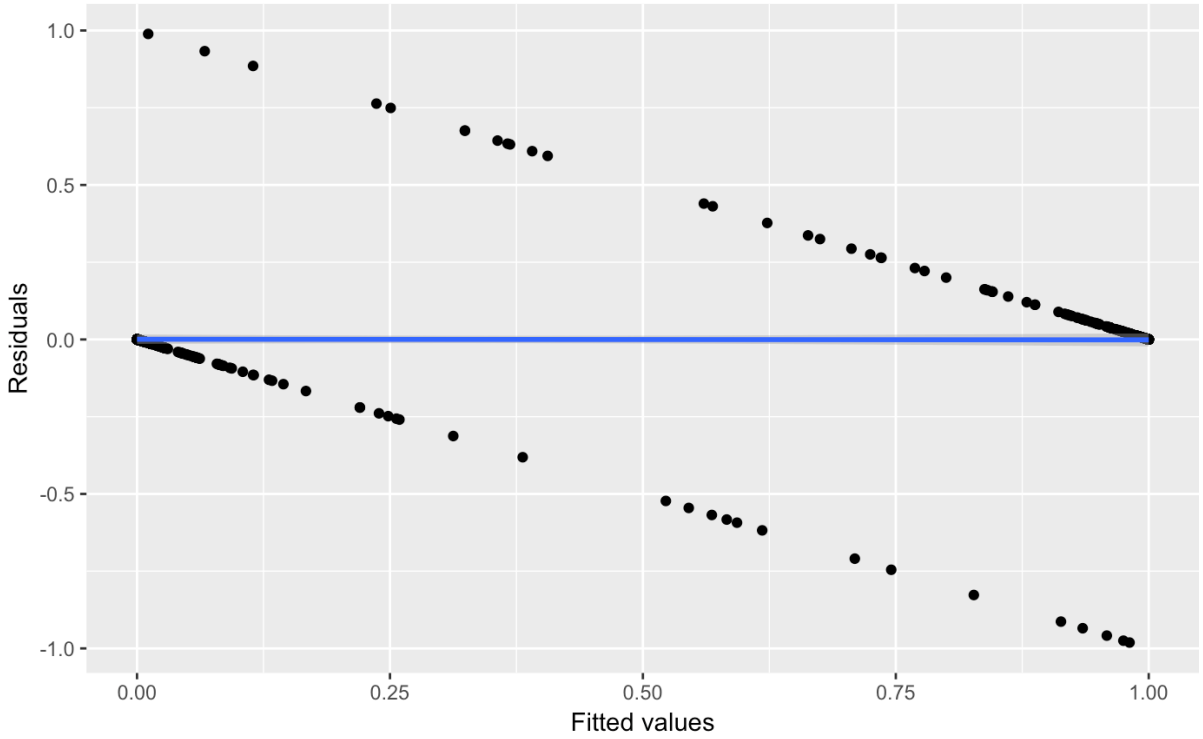

Figure 8: Residual Plot

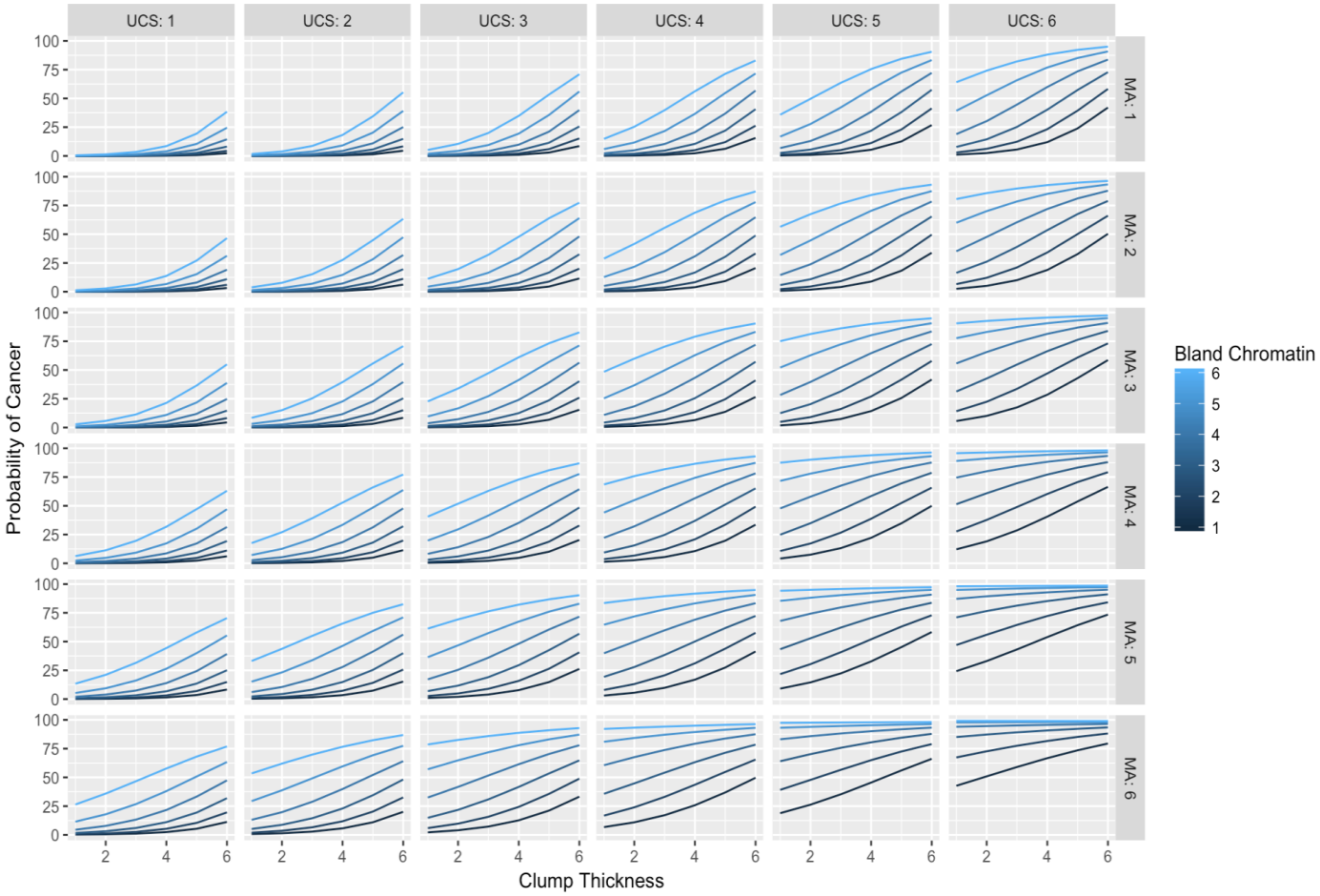

Figure 9: Prediction on changes in probability using the predictors.

Since we know higher values indicate definite malignance, we focused on exploring the smaller values. Some important take aways from this was –

- If Uniformity of Cell Shape and Bland Chromatin levels were high enough, the other variables didn't matter as much indicating these could be much more important and malign.
- This graph helps us tell things like if Uniformity is above 4 and bland chromatin is about 5-6, then there are high chances of malignance. But even if Marginal Adhesion is above 4, if the values of Uniformity and bland chromatin is less then there is comparitively less probability of malignance.
- Trend shows how as all the values increase the probability also increase.

ACCURACY

Our model can also give the prediction of malignance if we provide random values. In order to check the accuracy, we trained the model on 2/3 of the data we had and test it on the other 1/3 and that yielded us 98% accuracy.

CONCLUSION

We have done a detailed analysis of which attributes can be more helpful in determining if a tumor is cancerous. There is scope for R&D focusing on methods to measure these attributes precisely and with cheaper methods. We built a model as a preliminary test of breast cancer. We explored laboratory data and drew inferences from that. We require real-life data in-order to showcase real time accuracy or loss. This is a faster way of classifying the tumor and is probably more reliable than physical examination, but it is highly advised to go for further tests once a malignant tumor is detected. All the way while performing various analyses on this data, we used the values of the diagnostic tests which were compressed to a scale of 1 – 10. As a topic for future research, we are looking for the original data with the quantitative values from the diagnostic tests which could provide deeper insights for detailed analysis.

AKNOWLEDGEMNET

We wish to thank Prof. Brad Luen for numerous helpful discussions and comments. Also, thanking Miguel Pebes Trujillo, lead Associate Instructor, for his guidance and valuable suggestions. We also want to thank the University of California, Irvine for making the dataset available in their Machine Learning Repository. We also want to extend thanks to the R and RStudio community for their open source packages that inspired us to solve the problem effortlessly and precisely.

Works Cited

(BCOrg). Retrieved from http://www.breastcancer.org/symptoms/understand_bc/statistics

(cdc). Retrieved from https://www.cdc.gov/cancer/breast/basic_info/benefits-risks.htm

(O. L. Mangasarian, 1990). O. L. Mangasarian, R. S. (1990). Pattern recognition via linear programming: Theory and application to medical diagnosis. In *Large-scale numerical optimization* (pp. 22-30). SIAM Publications, Philadelphia.