Pooja Vasudevan

**Biomarkers for Breast Cancer**

**Predicting Presence of Breast Cancer**

**1.Abstract**

Breast Cancer is a common disease affecting women worldwide. Studies predict that one in eight women will be diagnosed with breast cancer in her lifetime. This disease is the result of malignant cells forming in the tissue of the breast. There are many variables that can contribute to the presence of breast cancer: Insulin, Gucose, HOMA, Leptin, Adiponectin, and Resistin to name a few. Many epidemiological studies indicate the relationship between obesity and prevalence of breast cancer. The medical diagnosis and analysis of breast cancer through MRI scans can be relatively time consuming and expensive. Hence, statistical methods to determine the most likely biomarkers for breast cancer would be valuable. This paper predicts what would most important biomarkers for breast cancer through logistic regression modeling and analysis using the R statistical programming language.

**2. Background**

We will begin the analysis with a brief explanation of the factors that contribute to breast cancer diagnosis and how these molecules interact eachother to play a critical role in cancer development. Then, we will analyze a real world, large breast cancer data set, originally posted to UCI Machine Learning (<https://archive.ics.uci.edu/ml/datasets>)

**2.1 Factors Affecting Breast Cancer**

**2.1.1 Glucose, Insulin, and HOMA**

A number of studies have revealed a link between diabetes, diabetes medicine, and breast cancer risk. Research suggests that women diagnosed with diabetes are more likely to be diagnosed with breast cancer than women who aren’t diabetic. Early studies have also shown that women diagnosed with breast cancer have higher levels of insulin and have a worse prognosis than normal levels of insulin. The hormone insulin helps our bodies regulate blood sugar (glucose). Many diabetic and obese patients tend to have higher levels of insulin, which facilitates the growth of breast cancer cells. To measure insulin resistance, blood sugar levels were measured using a HOMA index, a math formula to assess insulin sensitivity. A normal range is around 2. People with a HOMA score of 2.5 or higher are likely to have insulin resistance. The HOMA index is a method for assessing β-cell function and insulin resistance (IR) from basal (fasting) glucose and insulin or C-peptide concentrations. The following table below depicts the approximating equation for insulin resistance, using a fasting plasma sample, derived by the use of glucose-insulin product.

|  |  |
| --- | --- |
| HOMA-IR = (Glucose X Insulin)/22.5 | HOMA-IR = (Glucose X Insulin)/405 |
| HOMA- β = (20 X Insulin)/(Glucose-3.5) % | HOMA- β = (20 X Insulin)/(Glucose-63) % |
| Glucose in Molar Units mmol/L | Glucose in mass units mg/dL |

**Fig. 1** Equation for calculating HOMA based on Insulin and Glucose levels

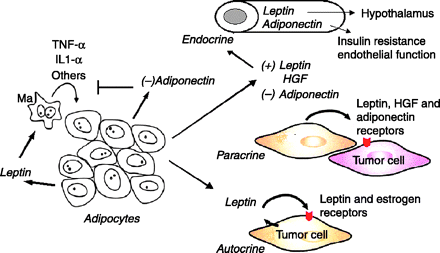
Women who don’t have insulin resistance tend to have better progression-free survival. Progression-free survival was lower in women with a HOMA score of 2.5 or higher. This essentially means that women who were insulin-resistant were more likely to have the cancer grow than women who were not insulin-resistant. Insulin is an important growth factor for all body tissues, even if it’s not clear how it affects the development of cancer cells.

While genetics play a role in your insulin levels, many people have higher insulin levels because of an unhealthy diet and lifestyle: too much sugar and too many simple carbohydrates

combined with not enough exercise. The best way to keep insulin at appropriate levels would be to have a healthy diet and lifestyle, which includes performing the following: eat a diet low in added sugar, exercise every day, maintain a healthy weight, don’t smoke, and limit alcohol use.

**2.1.2 Resistin, Adiponectin, and Leptin**

As studies have indicated that obesity as reflected by increased body mass index (BMI) is associated with increased risk of more aggressive breast cancer, adipose tissue, an endocrine organ producing and secreting a large range of factors, may interfere with cancer development. These factors called adipokines are involved in the mediation of inflammatory diseases and obesity. Adipokines, such as leptin, adiponectin, and resistin are produced by different fat depots. They act on breast cancer tissue in an endocrine manner, in a paracrine pathway, and in an autocrine action. The structure of the mammary gland may be in favor of close interaction between mammary adipose tissue and breast tissue, which suggests that adipokines produced by mammary adipose tissue and the tumor cell microenvironment may be the major link between obesity and breast cancer progression and metastasis.



**Fig. 2** In obesity and breast cancer, adipokines (leptin, adiponectin, and HGF) circulate in the plasma to interact with preneoplastic or cancerous breast epithelium.

Endocrine-, paracrine-, and autocrine-mediated relationships exist between leptin and the cellular microenvironment to support the growth of tumor cells via leptin and estrogen receptor activation. A paracrine relationship exists between HGF-synthesizing adipocytes and nearby mammary tumor cells to stimulate growth. Adiponectin exerts a direct growth-inhibitory effect on the tumor cells, blocks leptin secretion from surrounding breast adipose tissue, and prevents macrophages from producing inflammatory cytokines (TNF-α and IL-1β). HGF, hepatocyte growth factor; TNF-α, tumor necrosis factor-α; IL-1β, interleukin-1β; Ma, macrophage.

**3. Data Set**

This data was extracted from UCI Machine Learning Repository in “Center for Machine Learning and Intelligent Systems” (<http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>). The data was obtained by observing and measuring clinical features for 64 patients with breast cancer and 52 healthy controls. To indicate the presence or absence of breast cancer, there are 10 predictors, all quantitative, and a binary dependent variable. This data included anthroprometric parameters, which can be gathered from routine blood analysis. The following table depicts all the important independent variables in prediction of breast cancer.

|  |
| --- |
| Quantitative Attributes |
| Age (years) |
| BMI (kg/m2) |
| Glucose (mg/dL) |
| Insulin (µU/mL) |
| HOMA |
| Leptin (ng/mL) |
| Adiponectin (µg/mL) |
| Resistin (ng/mL) |
| MCP-1(pg/dL) |

**Fig. 3** A table of quantitative attributes, which have been used as

independent variables for this study

**4. Analysis**

The following description of the analysis is given without the complete R code, which can be found on Github and as an addendum to the report.

**4.1 Read and Examine Data Set**

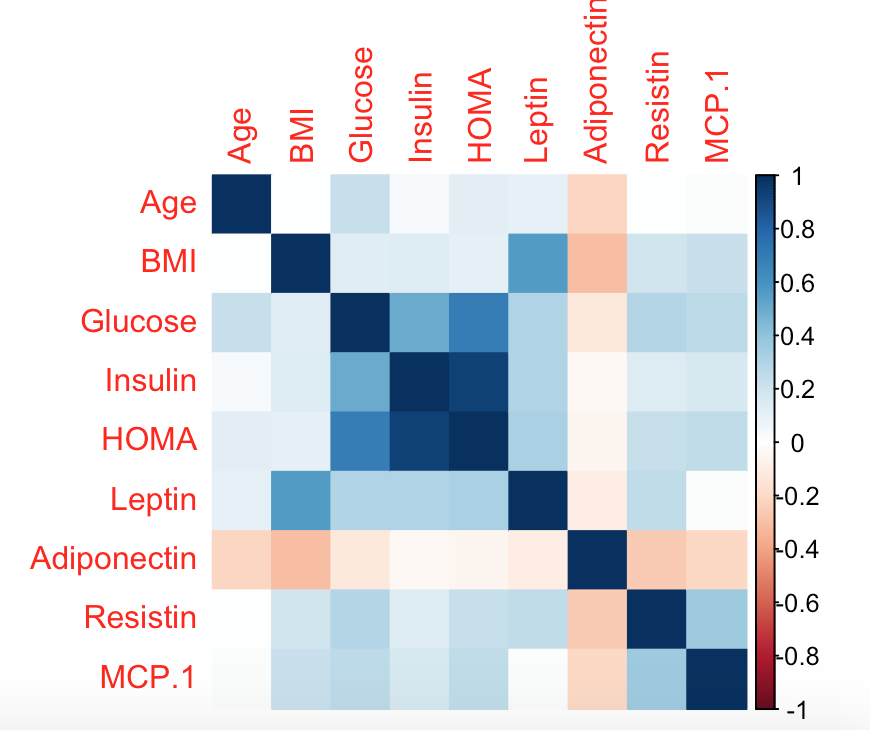
The data was read into Rstudio using read.csv(), then checked for missing values. Classification binary variable values were converted by creating a target column, that indicated 1 for presence of breast cancer and 0 for absence of breast cancer. Data was clean and tidy without any missing values.

**4.2 Renaming Columns**

The classification column was renamed to ‘target’ column as the binary dependent variable.

**4.3 Exploratory Data Analysis**

The color plot demonstrates which variables have a stronger correlation. From examining the color plot, Insulin and HOMA have the strongest correlation as well as insulin and Glucose, with a value close to 0.8.



**Fig. 4** A color plot to determine which variables have the

strongest relationship.

The correlation of BMI and Leptin was determined to be 0.6. The variables that have a weaker correlation, around 0.2-0.4, include BMI and Glucose, HOMA and Adiponectin, and Insulin and Adiponectin.

**4.3.1 Regression Analysis**

A logistic regression model was created to plot the independent variables against the binary dependent variable for predicting breast cancer, where 1 = presence and 0 = absence. The following table below provides an estimate for the probabilistic increase of breast cancer for each independent variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -5.438684 | 3.338638 | -1.629 | 0.10331 |
| Age | -0.022373 | 0.015563 | -1.438 | 0.15055 |
| BMI | -0.132158 | 0.063186 | -2.092 | 0.03648 |
| Glucose | 0.101003 | 0.033868 | 2.982 | 0.00286 |
| HOMA | -0.568062 | 1.067928 | -0.532 | 0.59478 |
| Insulin | 0.203698 | 0.257784 | 0.790 | 0.42942 |
| Leptin | -0.013537 | 0.016748 | -0.808 | 0.41893 |
| Adiponectin | -0.005895 | 0.037323 | -0.158 | 0.87451 |
| Resistin | 0.065705 | 0.030615 | 2.146 | 0.03186 |

**Fig. 5** Summary of logistic regression model on quantitative attributes

Out of the independent variables, insulin seems to have the highest probabilistic increase of cancer of about 20% for every unit increase in insulin, with every unit of glucose having a 10% increase chance of cancer. However, HOMA and BMI have a negative correlation with cancer. About every unit increase in BMI, there’s 13% decrease in chance of cancer., and every unit of HOMA, there’s a 6% chance of decrease in breast cancer. Adiponectin has the lowest correlation to cancer, with 0.5% decrease chance of breast cancer for every unit increase of Adiponectin.

Cancer data was split with a ratio of 0.75. A logistic regression model was developed for the splitted cancer data. The confusion matrix depicts threshold above 0.6 on sub portion of training cancer data and 0.8 for cancer test data for the full logistic regression model.

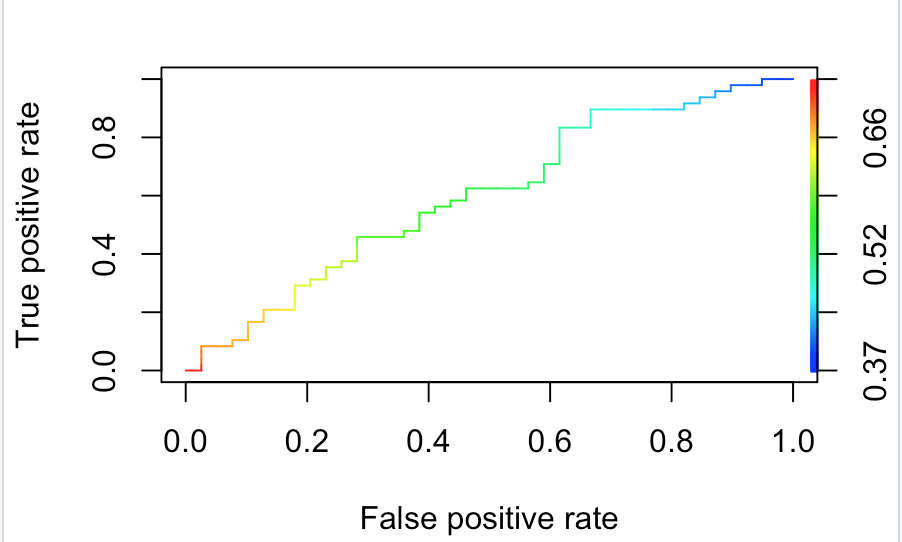
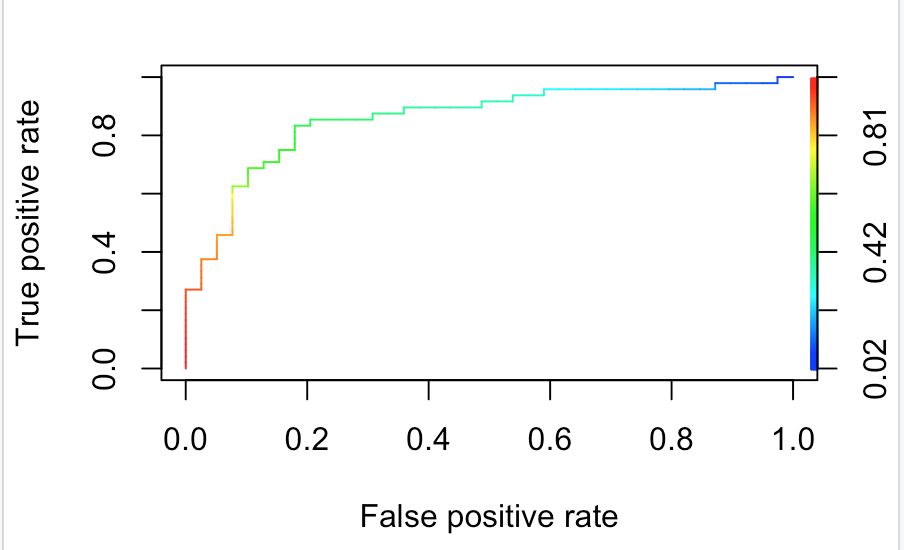
|  |  |  |
| --- | --- | --- |
|  | Predictor(P) | Predictor (N) |
| Actual  (P) | 28  (TP) | 11  (FN) |
| Actual  (N) | 28  (FP) | 20  (TN) |

|  |  |  |
| --- | --- | --- |
|  | Predictor(P) | Predictor (N) |
| Actual  (P) | 35  (TP) | 4  (FN) |
| Actual  (N) | 17  (FP) | 31  (TN) |

A. B.

**Fig. 6** Confusion Matrix for A. training data on full logistic model with 0.6 threshold B. test data on logistic regression model with 0.8 threshold

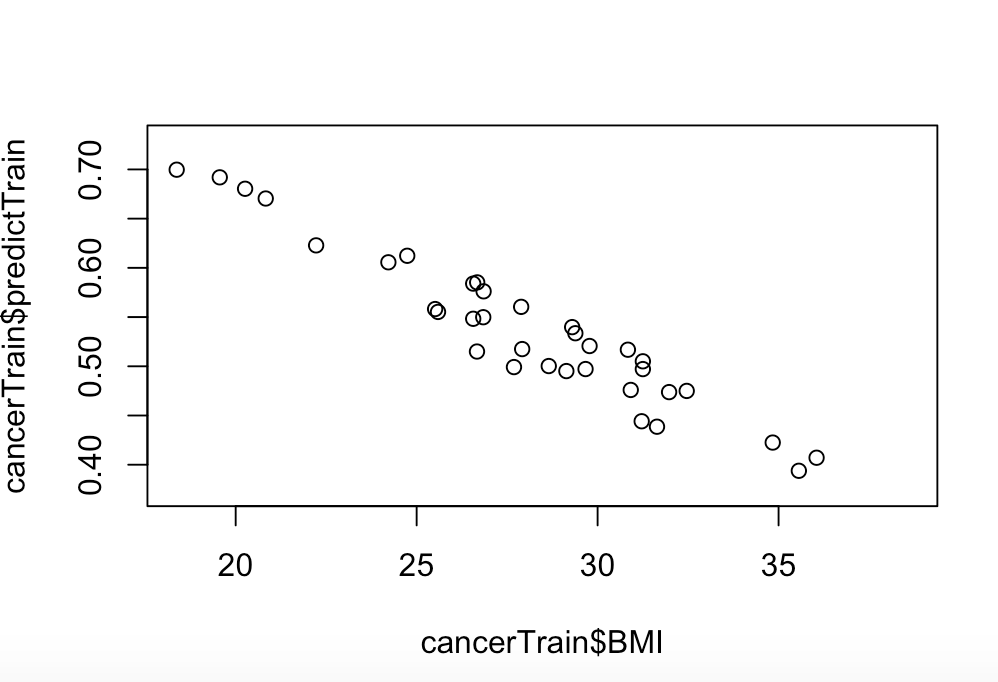
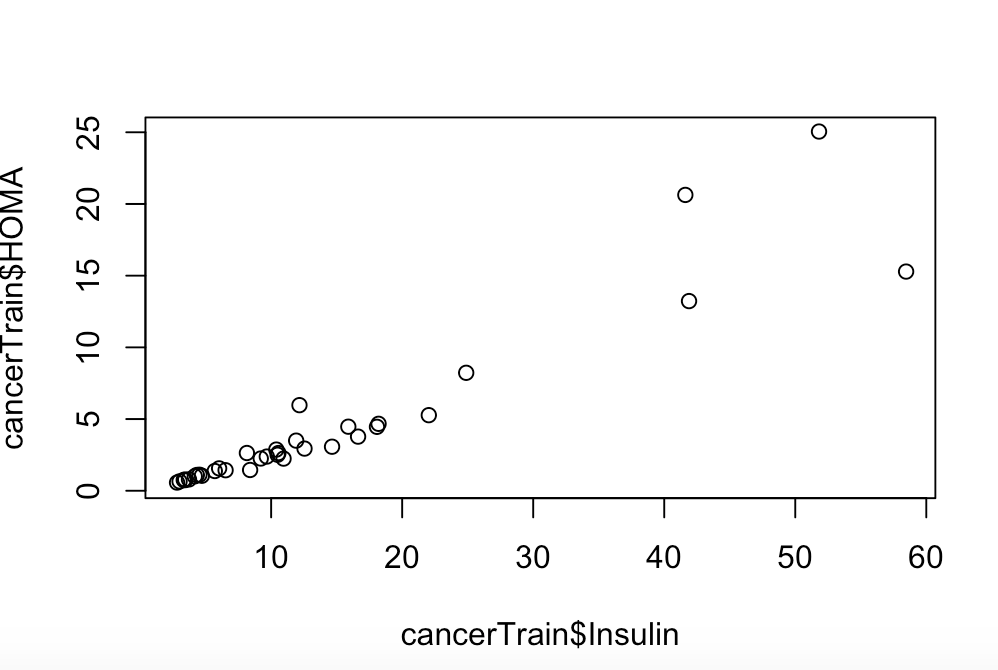
The sensitivity, specificity, and accuracy were calculated for both the cancer training and test data. For training data, the specificity is calculated to be 20/48 = 0.42, sensitivity was calculated to be 28/39 = 0.72, and accuracy was calculated to be 0.55. For test data, the specificity is calculated to be 31/48 = 0.42, sensitivity was calculated to be 35/39 = 0.92, and accuracy was calculated to be 0.76. The high sensitivity appropriately identifies patients with the actual disease.

A.B.

**Fig. 7** ROC curve of false positive rate against true positive rate for A. training data B. test data

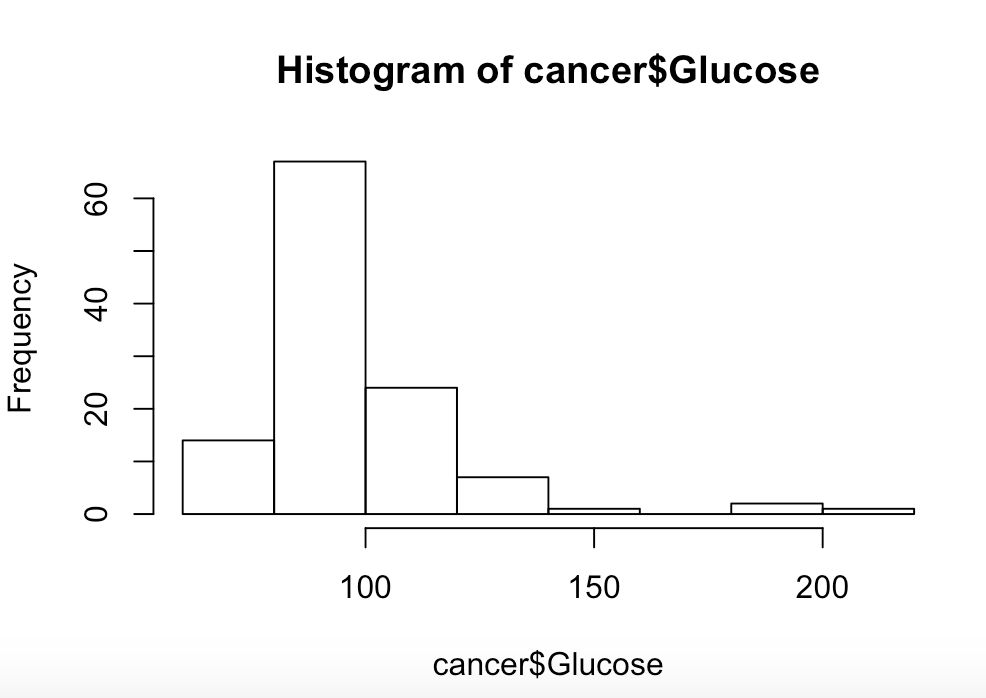
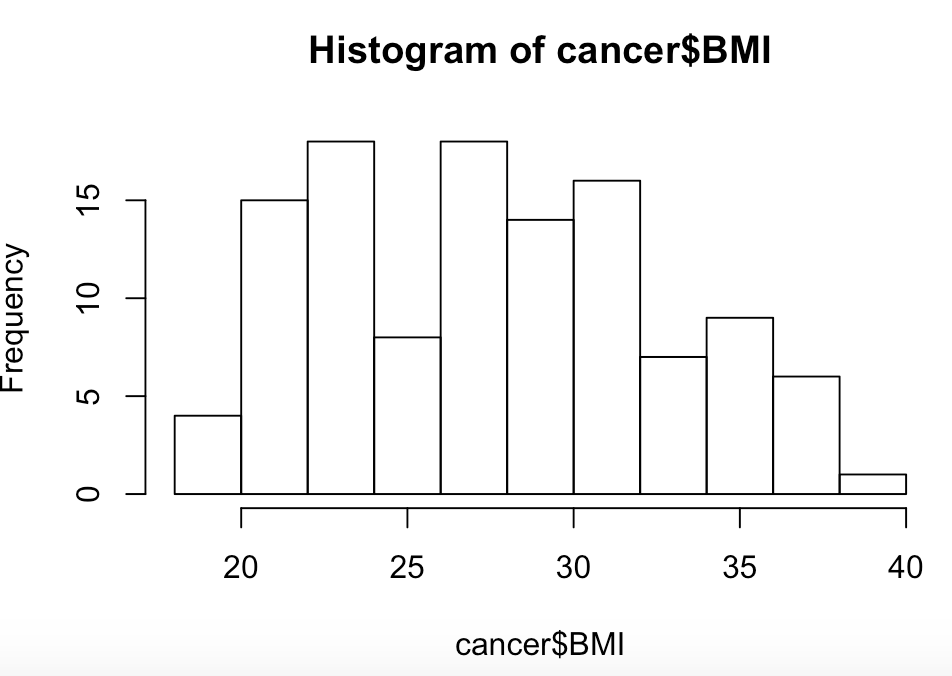
The ROC curve demonstrates the tradeoff between sensitivity and specificity. Based on the curve, a threshold of 0.6 was selected for the training data, and 0.8 for the test data.

**4.3.2 Plots**

**A****B**

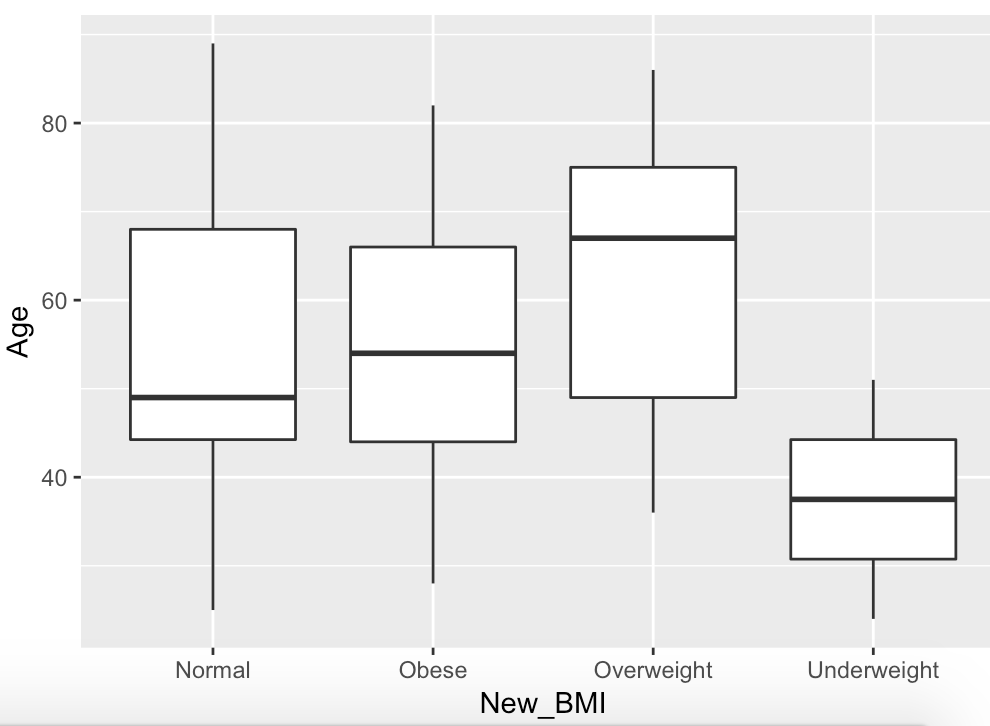
**Fig. 8** Graphical plots depicting the correlation of the following: A. BMI and splitted cancer data B. Relationship between Insulin and HOMA

There exists a downward linear trend with BMI and an upward positive linear trend with Insulin and HOMA. There exists a positive linear correlation between Insulin and HOMA with a Pearson’s correlation coefficient around 0.7.

A.B. 

**Fig. 9** Histogram displaying prevalence of Glucose ad BMI levels

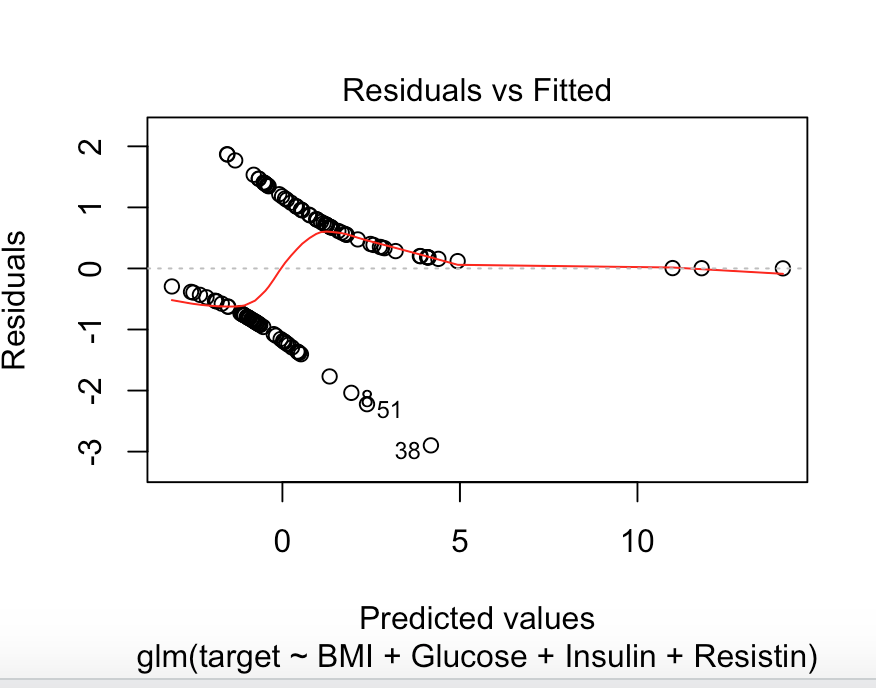
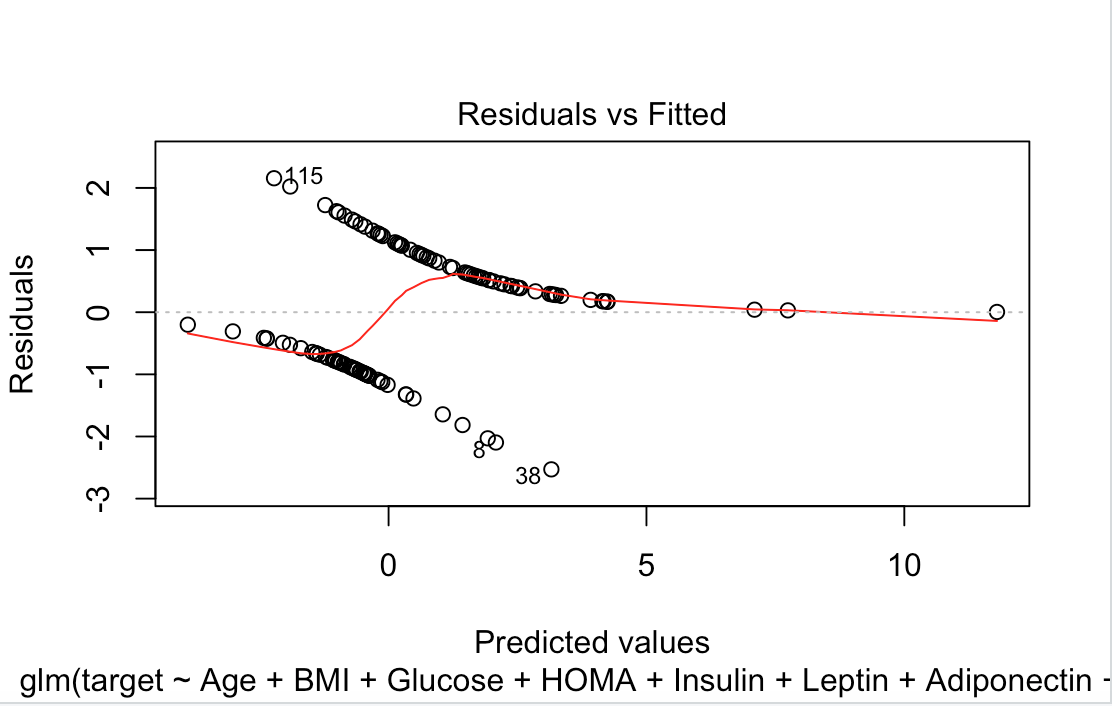
The Histogram displayed above displays the relative frequency of Glucose and BMI levels in patients. Most patients seem to have a glucose level at about 90-100 mg/dL, which is considered normal. The BMI level seems to be highest around 23 and 28 kg/m2. BMI distribution is fairly well spread.



**Fig. 10** Box and whisker plot depicting age range across different BMI levels.

A box and whisker plot was constructed to measure BMI levels across different ages. BMI levels were categorized into the following levels: Underweight (less than 19 kg/m2), Normal (19 to 25 kg/m2), Overweight (25 to 30 kg/m2), and Obese (over 30 kg/m2). Age distribution is more well distributed for individuals with a normal BMI. The median age for obesity and overweight BMI levels tends to be individuals who are older in age, approximately 58-63 years. The median age of underweight BMI levels are found more in younger adults, who are approximately 36-38 years old.

**4.3.3 Residuals**

A.B. 

**Fig. 11** Residual plots of A. Step-wise model B. Total model

Residuals measure the difference between the actual and expected value, and how a close a data point is to its real value. Residual plots help to explains the appropriateness of model in describing a relationship between variables. It shows how well the data fits a sigmoidal line. The metrics of the stepwise function is almost similar to the total model, with metrics slightly improved in terms of performance. The data points more closely follow the sigmoidal line.

**5. Results and Discussion**

The results can be summarized as follows:

* Insulin seems to be the most significant factor in contributing to the presence of breast cancer with 20% increase chance of cancer with every unit increase. Insulin is linked to diabetes and diabetic patients have a higher risk of breast cancer.
* Higher BMI levels are characterized in senior citizens who are about 58-65 years old, which poses a higher risk to cancer development
* Insulin and HOMA have a positive linear correlation. Thus, patients with high levels of insulin have a higher HOMA index.
* Resistin, an important protein contained in adipose tissue, contributes to obesity.

**6. Future Work**

As the focus of this research is understanding the quantitative attributes to predicting breast cancer, more research in understanding the chemical and physical properties of these molecules would be useful in understanding how these factors promote cancer development. It would be interesting to further investigate which important molecular and domain components interact to stimulate cancer development.

**7. Acknowledgements**

I would like to thank my mentor, Chris Esposo, for his input regarding this study, especially with statistical analysis.

**8. Addendum – R code**

#library

library(caTools)

library(dplyr)

library(ggplot2)

library(ISLR)

library(boot)

install.packages("corrplot")

library(corrplot)

cancer = read.csv("/Users/pooja.vasudevan/Downloads/cancerdata.csv")

summary(cancer)

str(cancer)

print(head(cancer, n=4))

View(cancer)

#Data Wrangling

cancer$BMI\_scale <- (cancer$BMI-mean(cancer$BMI))/sd(cancer$BMI)

cancer$target[cancer$Classification==2]=1

cancer$target[cancer$Classification==1]=0

View(cancer)

mod <- glm(formula=target ~ Age+BMI+Glucose+HOMA+Insulin+Leptin+Adiponectin+Resistin, family=binomial, data=cancer)

step.mod <- step(mod) #Step-wise model

step.mod

summary(step.mod)

#Cancer Train and Test Data

set.seed(88)

split = sample.split(cancer$target, SplitRatio = 0.75)

cancerTrain = subset(cancer, split==TRUE)

View(cancerTrain)

cancerTest = subset(cancer, split==FALSE)

head(cancerTrain)

CancerLog = glm(target ~ Age+BMI+Glucose+HOMA+Insulin+Leptin+Adiponectin+Resistin, family=binomial, data=cancerTrain)

final.mod2 <- step(CancerLog) #Step-wise model

plot(final.mod2)

summary(CancerLog)

View(cancerTest)

cancerTrain$predictTrain = predict(CancerLog, type="response")

pred = predict(CancerLog, newData=cancerTest, type="response")

cancerTrain$predictTrain2 = predict(final.mod2, type="response")

pred2 = predict(final.mod2, newData=cancerTest, type="response")

#Confusion Matrix - Total Model - Cancer Train

summary(predictTrain)

tapply(predictTrain, cancerTrain$target, mean)

table(cancerTrain$target, predictTrain > 0.6)

#Confusion Matrix - Total Model - Cancer Test

summary(pred)

tapply(pred, cancerTrain$target, mean)

table(cancerTrain$target, pred > 0.8)

#Confusion Matrix - Stepwise Model - Cancer Train

summary(predictTrain2)

tapply(predictTrain2, cancerTrain$target, mean)

table(cancerTrain$target, predictTrain2 > 0.6)

#Confusion Matrix - Step-wise Model - Cancer Test

summary(pred2)

tapply(pred2, cancerTrain$target, mean)

table(cancerTrain$target, pred2 > 0.85)

#Residuals

mod.res=resid(mod)

plot(mod)

plot(step.mod)

#ROC Curve - Cancer Train

ROCRpred = prediction(predictTrain, cancerTrain$target)

ROCRpref = performance(ROCRpred, "tpr", "fpr")

plot(ROCRpref, colorize=TRUE)

#ROC Curve - Cancer Test

ROCRpred = prediction(pred, cancerTrain$target)

ROCRpref = performance(ROCRpred, "tpr", "fpr")

plot(ROCRpref, colorize=TRUE)

#ROC Curve - Stepwise - Cancer Test

ROCRpred2 = prediction(pred2, cancerTrain$target)

ROCRpref2 = performance(ROCRpred2, "tpr", "fpr")

plot(ROCRpref2, colorize=TRUE)

#ROC Curve - Stepwise - Cancer Train

ROCRpred3 = prediction(predictTrain2, cancerTrain$target)

ROCRpref3 = performance(ROCRpred3, "tpr", "fpr")

plot(ROCRpref2, colorize=TRUE)

#Plots

plot(cancer$Age, cancer$BMI, col = cancer$target)

plot(cancer$BMI, cancer$Glucose, col = cancer$target)

plot(cancer$Glucose, cancer$HOMA, col = cancer$target)

plot(cancer$Leptin, cancer$BMI, col = cancer$target)

plot(cancerTrain$BMI, cancerTrain$predictTrain, col = cancer$target)

plot(cancerTrain$Insulin, cancerTrain$HOMA, col = cancer$target)

View(cancer)

cancer$New\_BMI[which(cancer$BMI <= 19)] <-'Underweight'

cancer$New\_BMI[which(cancer$BMI > 19 & cancer$BMI <= 25)] <-'Normal'

cancer$New\_BMI[which(cancer$BMI > 25 & cancer$BMI <= 30)] <-'Overweight'

cancer$New\_BMI[which(cancer$BMI > 30)] <- 'Obese'

View(cancer)

box\_plot <- ggplot(cancer, aes(x = New\_BMI, y = Age))

# Add the geometric object box plot

box\_plot +

geom\_boxplot()

hist(cancer$Glucose, bin=50)

hist(cancer$BMI)

## LOOCV approach

set.seed(1)

#Fit a linear model

m = glm(BMI ~ Age, data = cancerTrain)

MSE\_LOOCV = cv.glm(cancerTrain, m) #test model on data that hasb't been trained on

MSE\_LOOCV$delta[1]

MSE\_10\_fold\_cv = NULL

for(i in 1:10){

m = glm(BMI~poly(Age, i), data=cancerTrain)

MSE\_10\_fold\_cv[i] = cv.glm(cancerTrain, m, K=10)$delta[1] ##divide data set in 10 parts

}

MSE\_10\_fold\_cv

MSE\_LOOCV

#Color plot

cancer\_feature\_set <- cancer[1:9]

View(cancer\_feature\_set)

cor1 <- cor(cancer\_feature\_set)

corrplot(cor1, method="color")

cancer2 = unique(cancer)

str(cancer2)

distances = dist(cancer2, method="euclidean")

clusterCancer2 = hclust(distances, method="ward")

clusterGroups = cutree(clusterCancer2, k=10)

tapply(cancer2$BMI, clusterGroups, mean)

cluster2 = subset(cancer2, clusterGroups==2)

cluster2$BMI

plot(clusterCancer2)