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

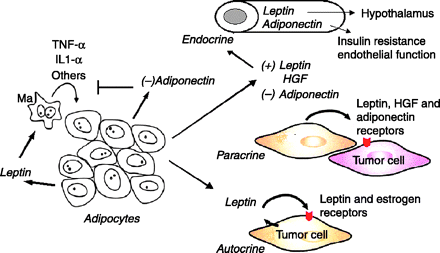
Women who don’t have insulin resistance tend to have better progression-free survival. Progression-free survival was lower in women with a HOMA score of 2.5 or higher. This essentially means that women who were insulin-resistant were more likely to have the cancer grow than women who were not insulin-resistant. 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**

**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.

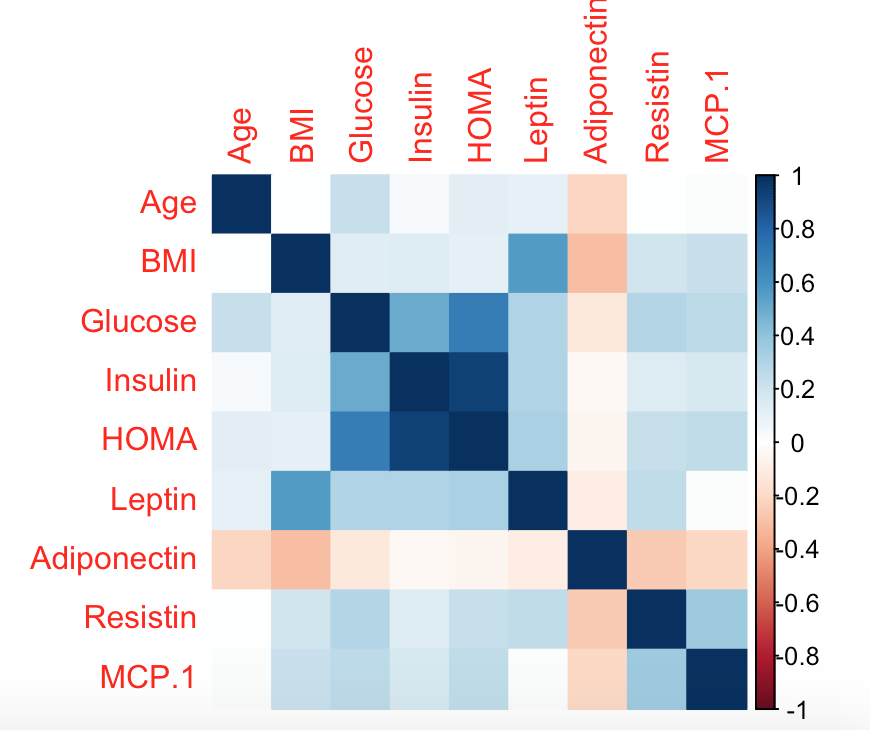


Fig. 1

**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 |

An increase in glucose indicates a 10% chance increase of cancer.

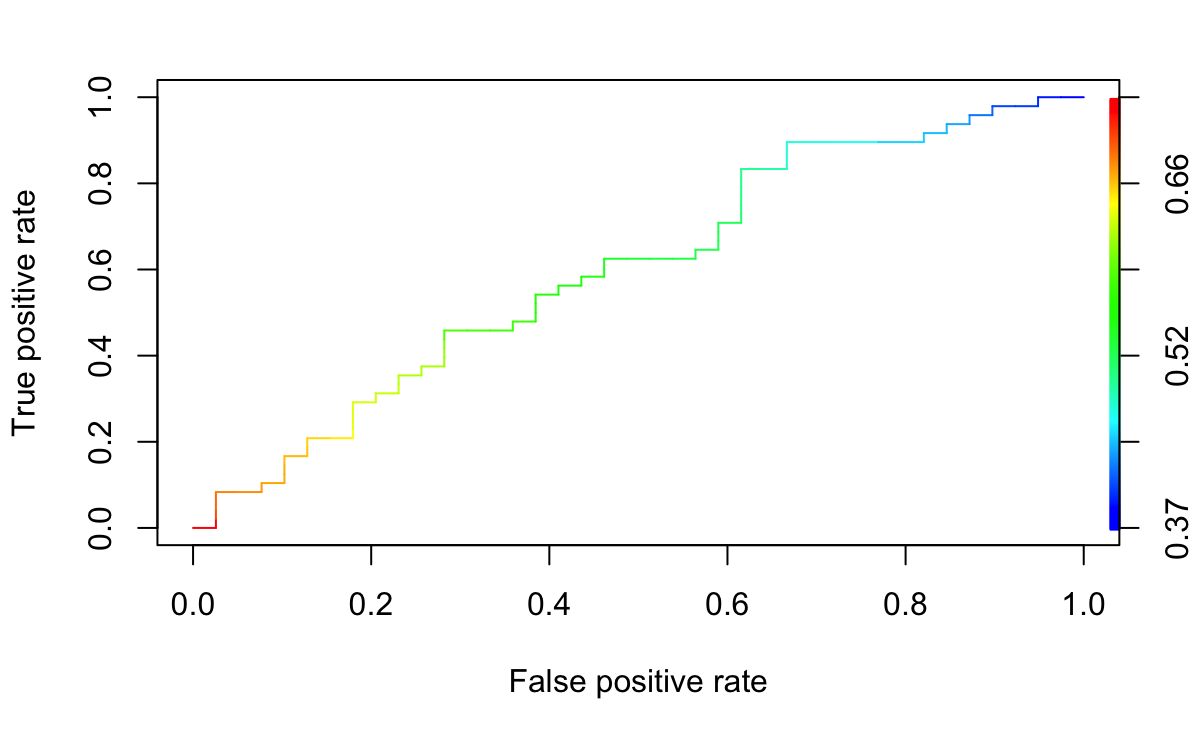
The confusion matrix depicts threshold above 0.5 on sub portion of cancer data.

FALSE TRUE

0 15 24

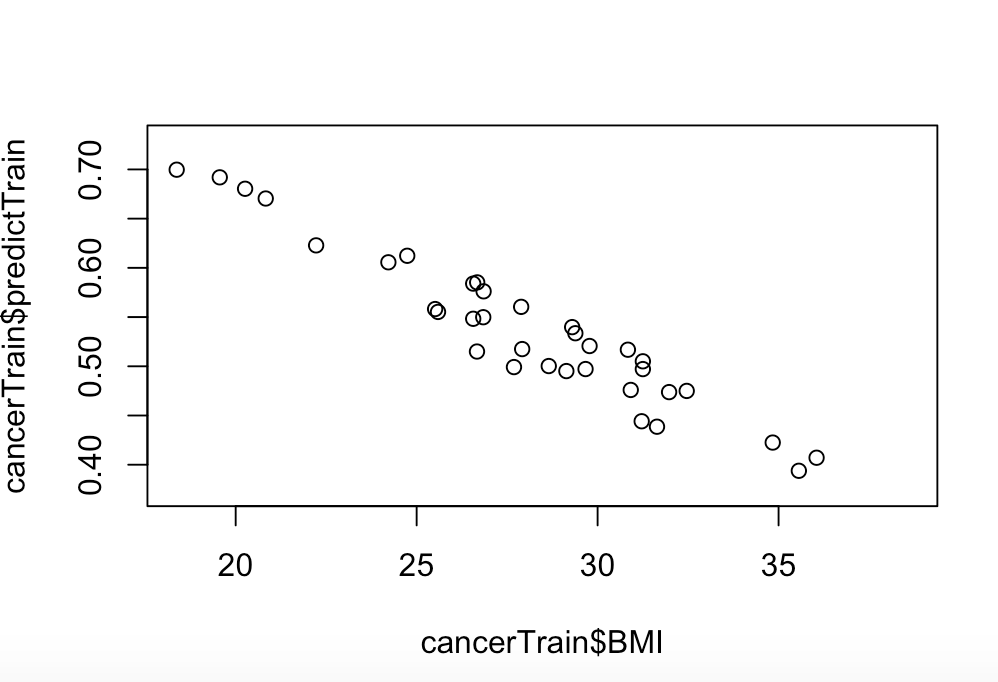
1 12 36

The sensitivity is calculated to be 36/48 = 0.75 and specificity was calculated to be 15/39 = 0.38.

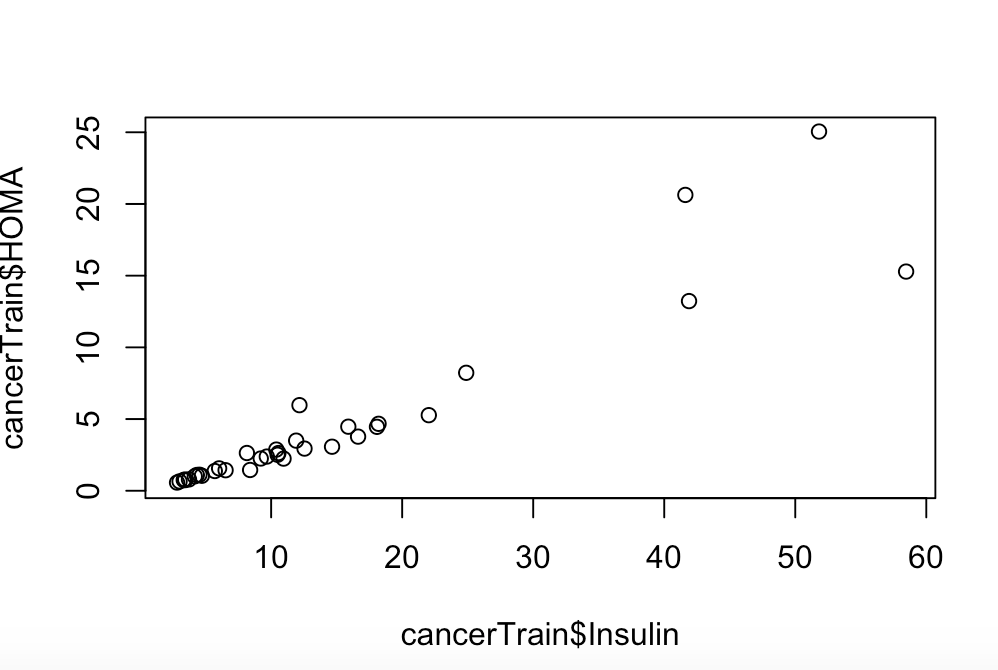


The ROC curve demonstrates the tradeoff between sensitivity and specificity. Based on the curve, a threshold of 0.5 was selected.

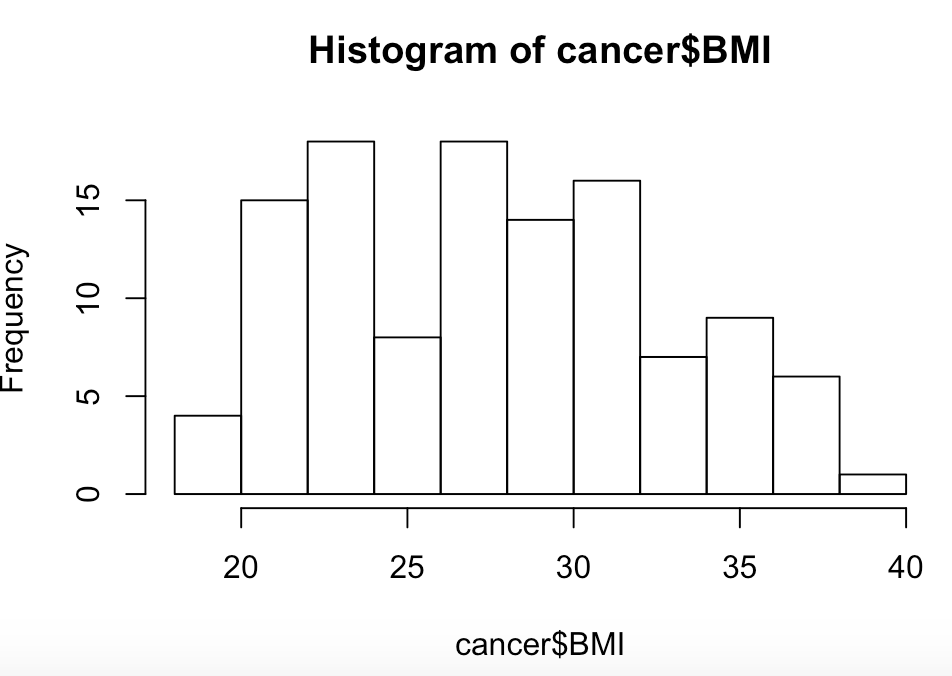
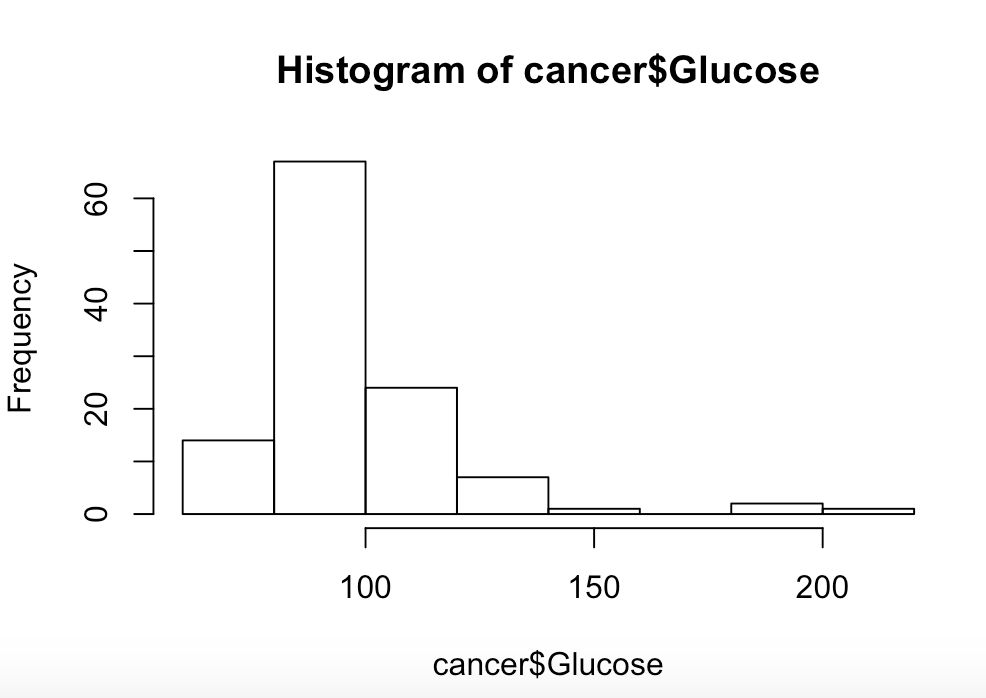
**4.3.2 Plots**

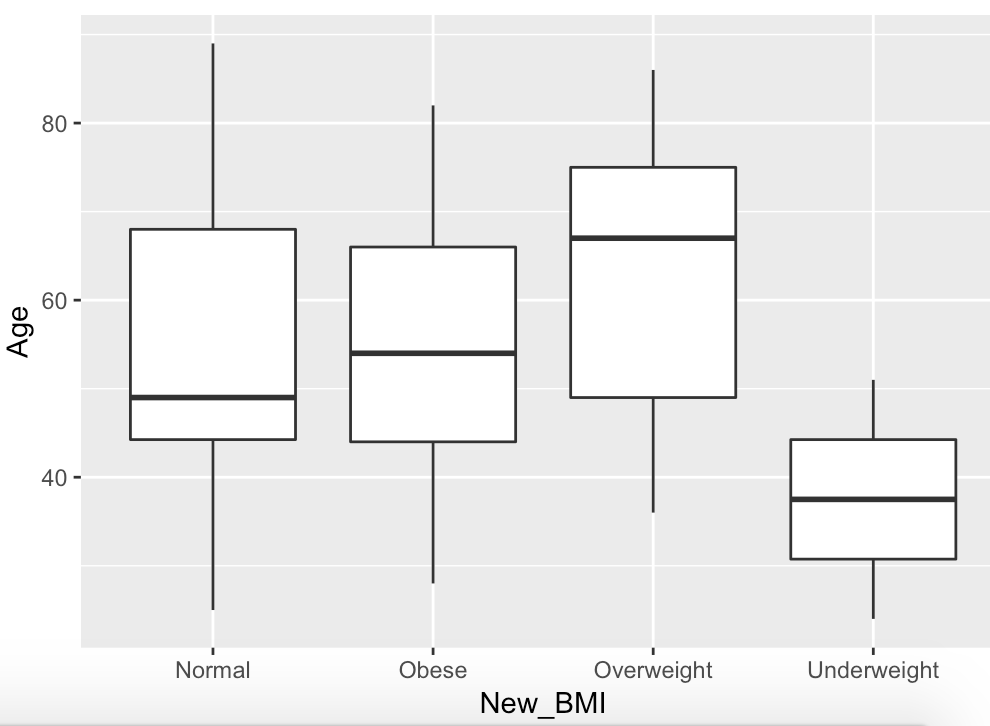


There exists a downward linear trend with BMI.



There exists a positive linear correlation between Insulin and HOMA.





**4.3.2 Residuals**

