Research Goals

What factors are the biggest predictors of breast cancer? Is age a bigger predictor than BMI? How much of a role do glucose, insulin, leptin, and other proteins play in predicting if a person will have breast cancer? Breast cancer is the most common cancer among American women, besides from skin cancers. By studying the predictors of breast cancer, in larger studies than the one used in this project, we could potentially avoid these risk factors and lower the chances of anyone getting breast cancer.

Data

Source: [Patricio, 2018] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer, 18(1). [doi:10.1186/s12885-017-3877-1]

Collection: The data were collected by that Faculty of Medicine of the University of Coimbra (Miguel Patrício, José Pereira, Joana Crisóstomo, Paulo Matafome, Raquel Seiça, Francisco Caramelo) and also Manuel Gomes from the University Hospital Centre of Coimbra

Variables:

- Age (numerical)
- BMI (numerical)
- Glucose (numerical)
- Insulin (numerical)
- HOMA (numerical)
- Leptin (numerical)
- Adiponectin (numerical)
- Resistin (numerical)
- MCP-1 (numerical)

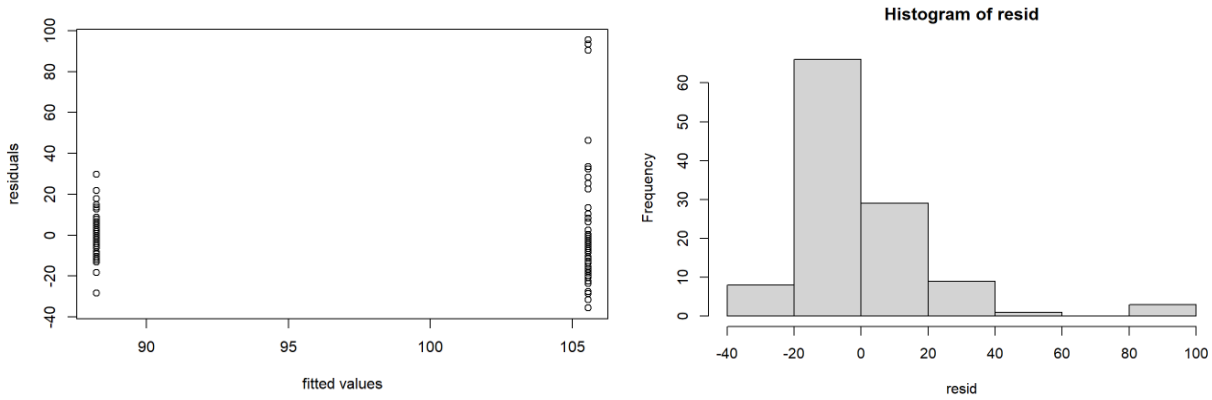
Restrictions: The sample size is relatively small and the data was all collected in Coimbra, however it is likely that the findings will hold true to a larger population.

Conditions:

116 subjects is less than 10% of the entire population of Coimbra.
64 patients with breast cancer is greater than 10.
52 healthy controls is greater than 10.

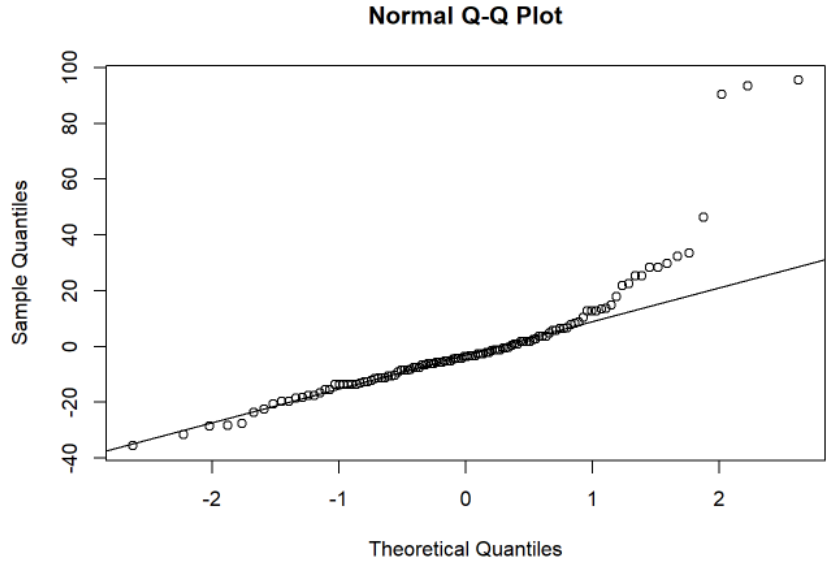
Conditions for Regression (Classification vs. Glucose):

Linearity/Normal Residuals



Linearity does not seem to be violated but is strange for the fact that classification is a boolean value, it is either 1 or 0. Data are a little right skewed but it's okay.

Constant Variability



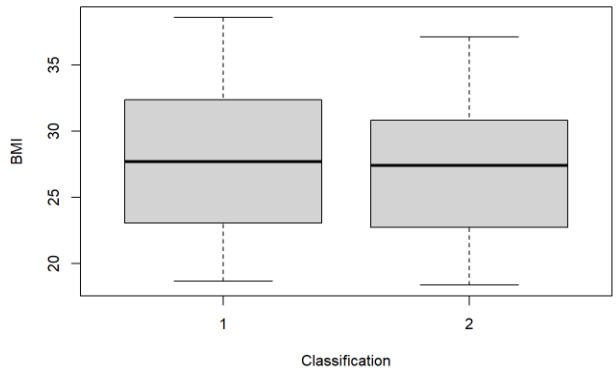
Fits the line well except for a few outliers.

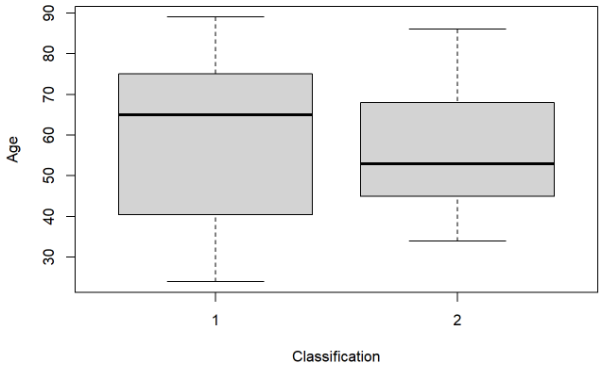
Exploratory Data Analysis:

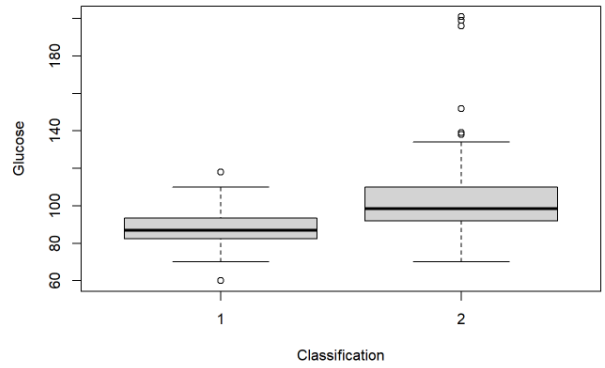
Key: Category 1 = Healthy Control, Category 2 = Breast Cancer Patient

```
boxplot(BMI~Classification)
```

```
boxplot(Age~Classification)
```







Exploratory data analysis shows that BMI is probably not a very good indicator of breast cancer, but age might be. We also see that glucose may be a strong predictor of breast cancer.

Inference:

```
#glucose level over 90:
##classification1: 20     total: 52
##classification2: 51     total: 64
prop.test(x=c(20,51),n=c(52,64))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(20, 51) out of c(52, 64)
## X-squared = 18.836, df = 1, p-value = 1.424e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.5946136 -0.2299056
## sample estimates:
##    prop 1    prop 2
## 0.3846154 0.7968750
```

A two-proportion chi-squared test was conducted to test if there was a significant difference between the glucose levels of healthy patients and the glucose levels of cancer patients. A base

number of 90 mg/dL was used for the glucose number, as the mean glucose level for healthy patients was 88.3 mg/dL and the mean glucose level for cancer patients was 105.6 mg/dL. The counts are of patients with glucose levels higher than the base.

H0: There is no difference between the glucose levels of cancer patients and the glucose level of healthy patients.

HA: There is a difference between the glucose levels of cancer patients and the glucose level of healthy patients.

p-value = 1.424*10^-5

Since the p-value is very small, we reject the H0. We have convincing evidence that there is a difference between the glucose levels of cancer patients and the glucose level of healthy patients. We can confidently claim that the glucose levels are significantly higher in cancer patients than in healthy patients.

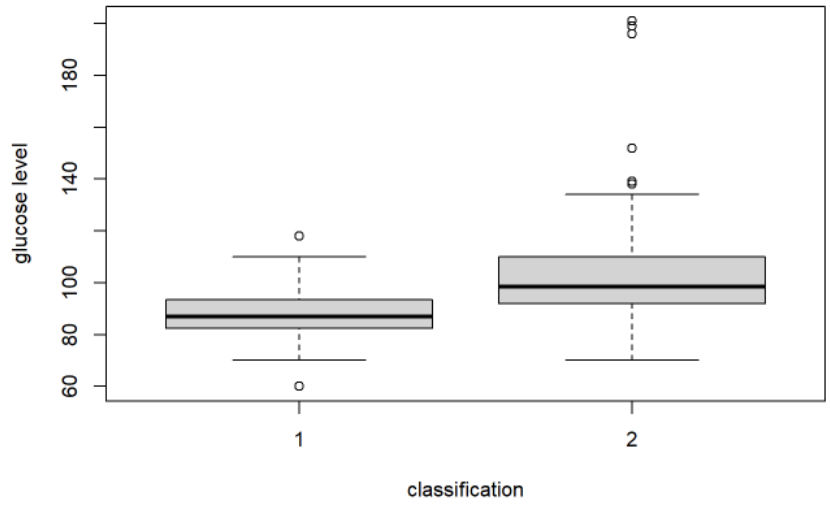
Regression:

```
##
## Call:
## lm(formula = Classification ~ Age + BMI + Glucose + Insulin +
##     HOMA + Leptin + Adiponectin + Resistin + MCP.1, data = cancer,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86576 -0.36426  0.03153  0.35611  0.81974
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.279e-01  4.316e-01   2.150  0.03383 *
## Age         -2.815e-03  2.740e-03  -1.027  0.30668
## BMI         -2.651e-02  1.118e-02  -2.371  0.01953 *
## Glucose      1.378e-02  3.102e-03   4.443 2.18e-05 ***
## Insulin      4.492e-02  1.430e-02   3.142  0.00218 **
## HOMA        -1.357e-01  4.779e-02  -2.840  0.00541 **
## Leptin      -7.904e-04  2.933e-03  -0.270  0.78805
## Adiponectin -1.730e-03  6.686e-03  -0.259  0.79630
## Resistin     7.730e-03  3.825e-03   2.021  0.04578 *
## MCP.1        1.733e-05  1.396e-04   0.124  0.90144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4375 on 106 degrees of freedom
## Multiple R-squared:  0.2929, Adjusted R-squared:  0.2329
## F-statistic: 4.879 on 9 and 106 DF,  p-value: 1.811e-05
```

According to the p-values, the largest predictor of breast cancer is the glucose level. Aside from that, the insulin level and HOMA both are large predictors of breast cancer. Resistin still is still a significant predictor, but to less of an extent than the previously mentioned factors. Some factors may be partially dependent on each other.

```
#giving simple names to the two variables of interest
x=cancer$Classification
y=cancer$Glucose

#plotting the data
boxplot(y~x,xlab="classification", ylab="glucose level")
```




```
#find the correlation **order does not matter **cor does not change if
#x or y are changed by some constant(addition,multiplication,etc)
cor(x,y)
```

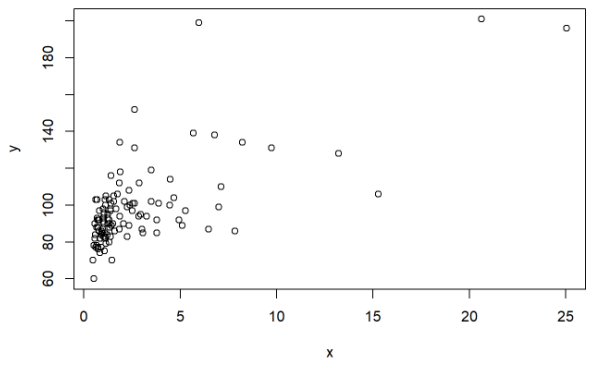
```
## [1] 0.3843154
```

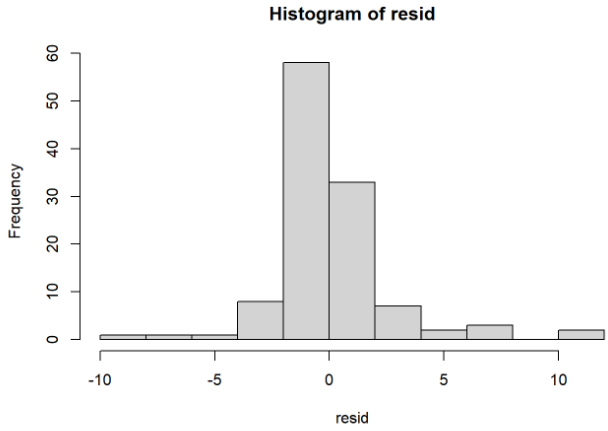
The glucose level and classification had the highest correlation with a correlation coefficient of 0.384 (age = -0.044; BMI = -0.133; insulin = 0.277; HOMA = 0.284; leptin = -0.001; adiponectin = -0.019; resistin = 0.227; MCP.1 = 0.091). The correlation coefficient of 0.384, is a positive correlation, which means that as glucose levels increase, the chances of developing breast cancer increase. While it is the highest correlation, it is still a very weak correlation.

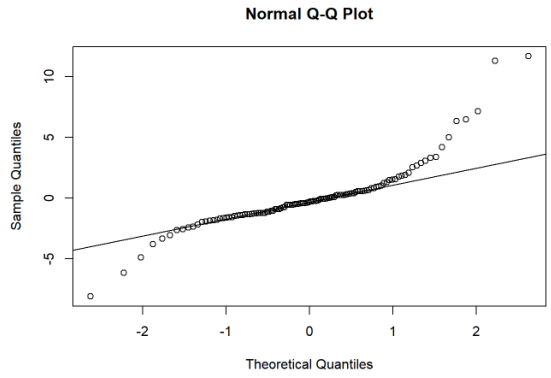
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.563 -11.231  -3.563   5.019  95.438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   70.899      6.354  11.158  < 2e-16 ***
## x             17.332      3.899   4.445 2.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.89 on 114 degrees of freedom
## Multiple R-squared:  0.1477, Adjusted R-squared:  0.1402
## F-statistic: 19.76 on 1 and 114 DF,  p-value: 2.052e-05
```

Simple linear regression test: The population for the test included healthy and cancer patients in Coimbra. The x-variable is whether or not the patient has cancer and the y-variable is the glucose level. The parameter slope is equal to 17.3 mg/dL, while the intercept is 70 mg/dL. The intercept does not necessarily give an accurate representation of anything, since it is what the glucose would be if the classification were 0, and the classification is never 0 since 1 is healthy and 2 is with cancer. However, the large slope does suggest that the jump from healthy to cancer (1 to 2) increases the glucose levels by 17 mg/dL. The regression line is essentially ineffective and nonfunctional in predicting whether a patient will have cancer from a given glucose level, and predicting the glucose levels of healthy and cancer patients. The model predicts that every healthy patient would theoretically have a glucose level of 88.2 mg/dL (the mean glucose level of all healthy patients) and that every cancer patient would theoretically have a glucose level of 105.6 mg/dL (the mean glucose level of all cancer patients). Because there is no variation in the classification variable, there is no variation in the prediction model.

There is, however, a moderate correlation between the glucose levels and HOMA (homeostasis model assessment of insulin resistance), with a correlation coefficient of 0.70. Since we know that higher glucose levels are correlated with a higher risk of developing breast cancer, we can test the regression to see how the HOMA and glucose levels affect each other to understand if there is a correlation between HOMA and cancer risk as a result of glucose levels.








```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.451 -10.862  -3.283   4.725 104.199
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.3602     7.5821  10.467   <2e-16 ***
## y             0.3217     0.1274   2.525    0.013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.02 on 114 degrees of freedom
## Multiple R-squared:  0.05295,    Adjusted R-squared:  0.04464
## F-statistic: 6.374 on 1 and 114 DF,  p-value: 0.01296
```

The conditions are satisfied for the regression analysis of glucose vs insulin. The regression line was found to be y = 0.113x - 8.313. This slope indicates that for every increase in the HOMA, their glucose level increases by 0.113 mg/dL, which is relatively small. The intercept indicates that if the HOMA was zero (unrealistic), the glucose level would theoretically be -8.313 mg/dL (also unrealistic). The added benefit of this regression line is that the theoretical glucose value could be computed given the HOMA value, or vice versa. From this data, we could predict that the chances of developing breast cancer increase with the increase in HOMA, but we do not have enough information to prove the result of this claim.

Conclusion:

The data suggest that there are many factors that can predict whether a person will develop breast cancer or not. The best predictor that the data finds is the glucose level of the individual. As the hypothesis concludes, we have convincing evidence that there is a difference between the glucose levels of cancer patients and the glucose level of healthy patients. We can confidently claim that the glucose levels are significantly higher in cancer patients than in healthy patients. We can further claim that if an individual's glucose levels are high, they are more likely to develop breast cancer. The chi-squared test is essentially the best analysis that we have of the data because a confidence interval was also given, which allows for analysis of exactly how much glucose levels differ in cancer patients than in healthy patients. The regression is a less reliable analysis due to the fact that the classification simply has two possible responses (1 or 2), but essentially shows the same result, as does the regression line for insulin and HOMA. The latter indicates that both the glucose and HOMA together play a role in predicting the chances of developing breast cancer. The possible limitations of this data is that the sample size is relatively small and all of the data was collected in Coimbra, however it is likely that the findings will hold true to a larger population.