## **Introduction**:

This paper will look at examples of Prostate Cancer diagnoses. The diagnosis of Prostate Cancer will be modeled as a binomial distribution with y = 1 if a patient has a diagnosis of Prostate Cancer and 0 otherwise. This is being modeled based on the assumption that the number of patients observed is fixed, each patient observed is independent of each other, and that there are only two outcomes for a diagnosis of Prostate Cancer.

Since the Binomial distribution is an exponential family function, a generalized linear model will be used. A logistic regression on the odds of having a diagnosis of Prostate Cancer can be done since exponential families satisfy the assumptions for logistic regression. The variables:

- Prostate Specific Antigen Level as psa (mg/ml)
- Prostate Cancer Volume Estimate as c.vol (cc)
- Prostate Weight as weight (grams)
- Age as age (years)
- Amount of Benign Prostatic Hyperplasia as benign ($cm^2$)
- Presence or Absence of Seminal Vesicle Invasion modeled as inv with
    - 0 for absence
    - 1 for presence
- Degree of Capsular Penetration (cm)

are going to be used to try to predict the value of y for the said model.

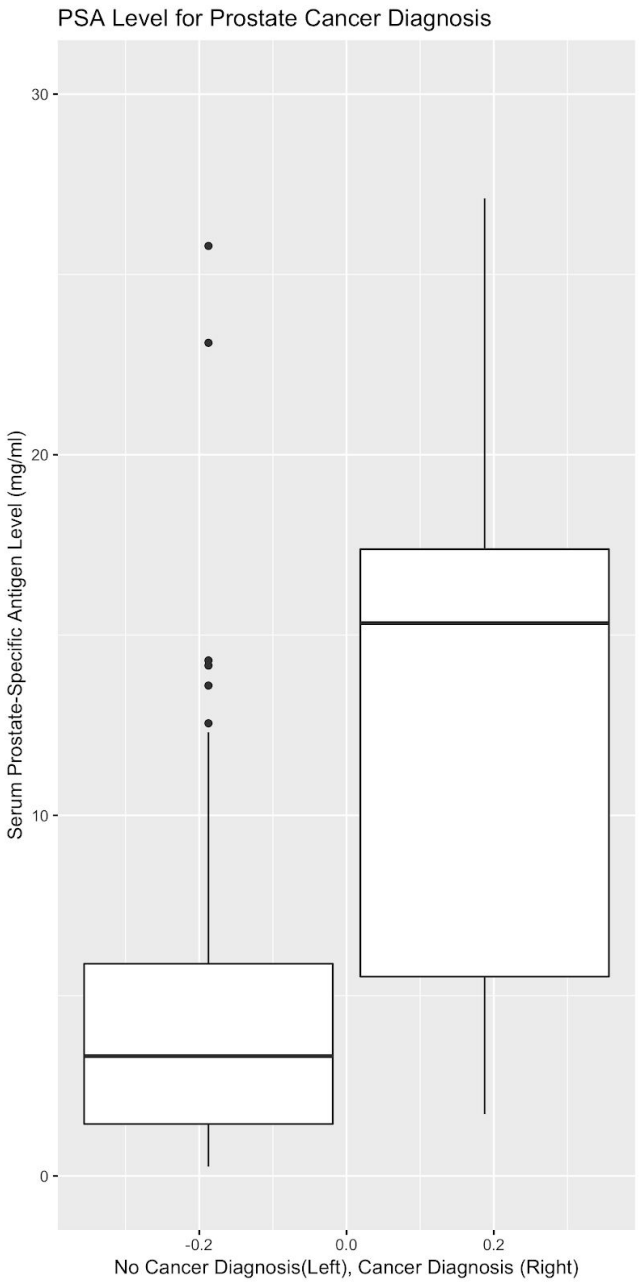
## **Summary**:

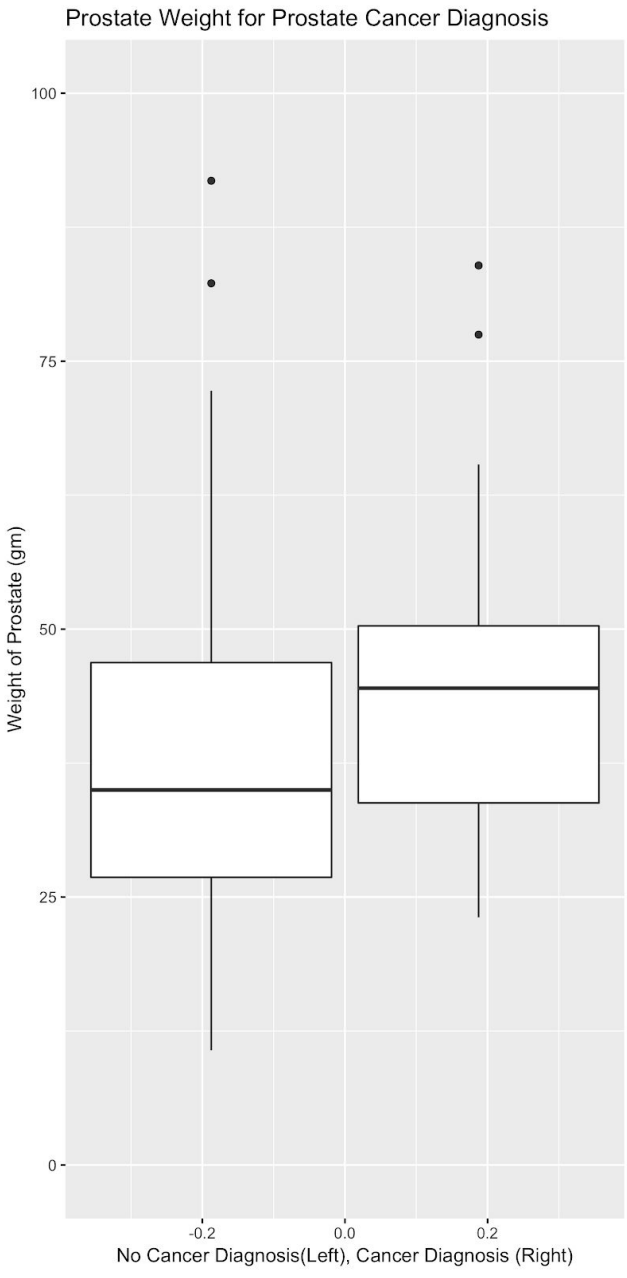
**5-Number Summary and Mean for Continuous Variables Table**

| | y | psa | c.vol | weight | age | benign | cap |
|---|---|---|---|---|---|---|---|
| min: | 0 | 0.651 | 0.2592 | 10.70 | 41.00 | 0 | 0 |
| 1st Quartile: | 0 | 5.641 | 1.6653 | 29.37 | 60.00 | 0 | 0 |
| Median: | 0 | 13.330 | 4.2631 | 37.34 | 50.00 | 1.350 | 0.4493 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Mean: | 0.2165 | 23.730 | 6.9987 | 45.49 | 63.87 | 2.535 | 2.2454 |
| 3rd Quartile: | 0 | 21.328 | 8.4149 | 48.42 | 68.00 | 4.759 | 3.2544 |
| Max: | 1 | 265.072 | 45.6042 | 450.34 | 79.00 | 10.278 | 18.1741 |

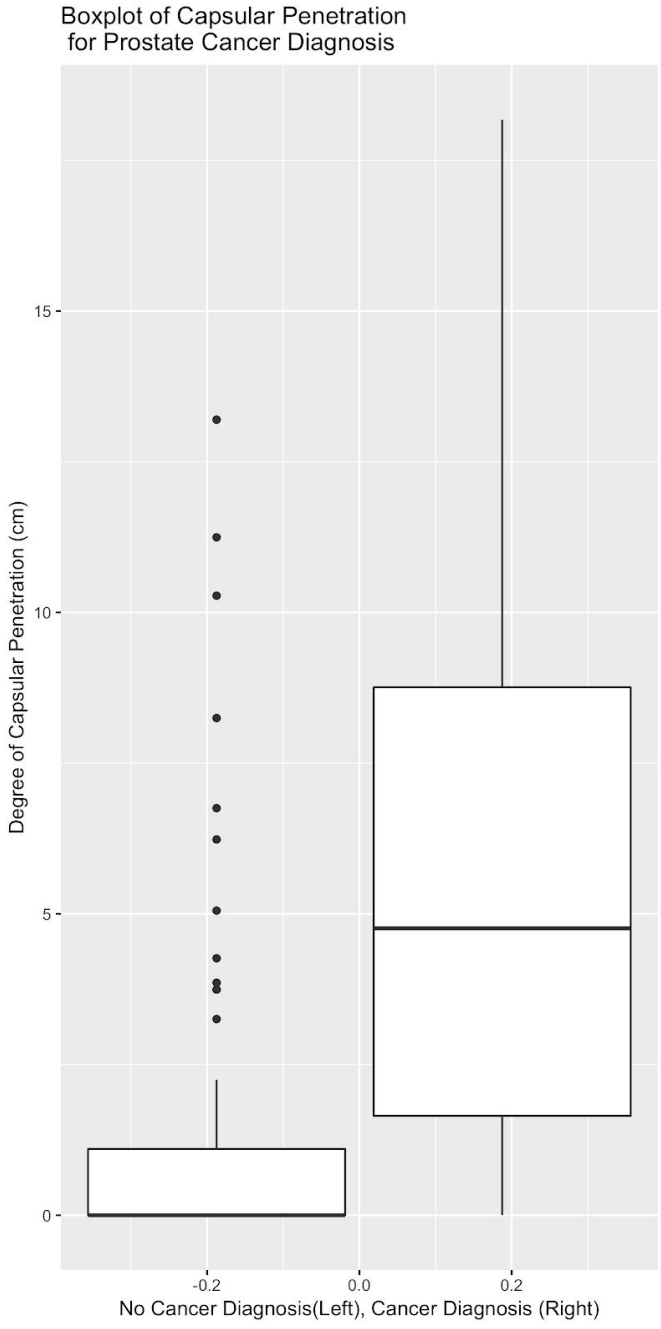
From the five-number summary we can see that there are possible influential points and outliers for the weight, psa, and c.vol columns. However, the effect of these

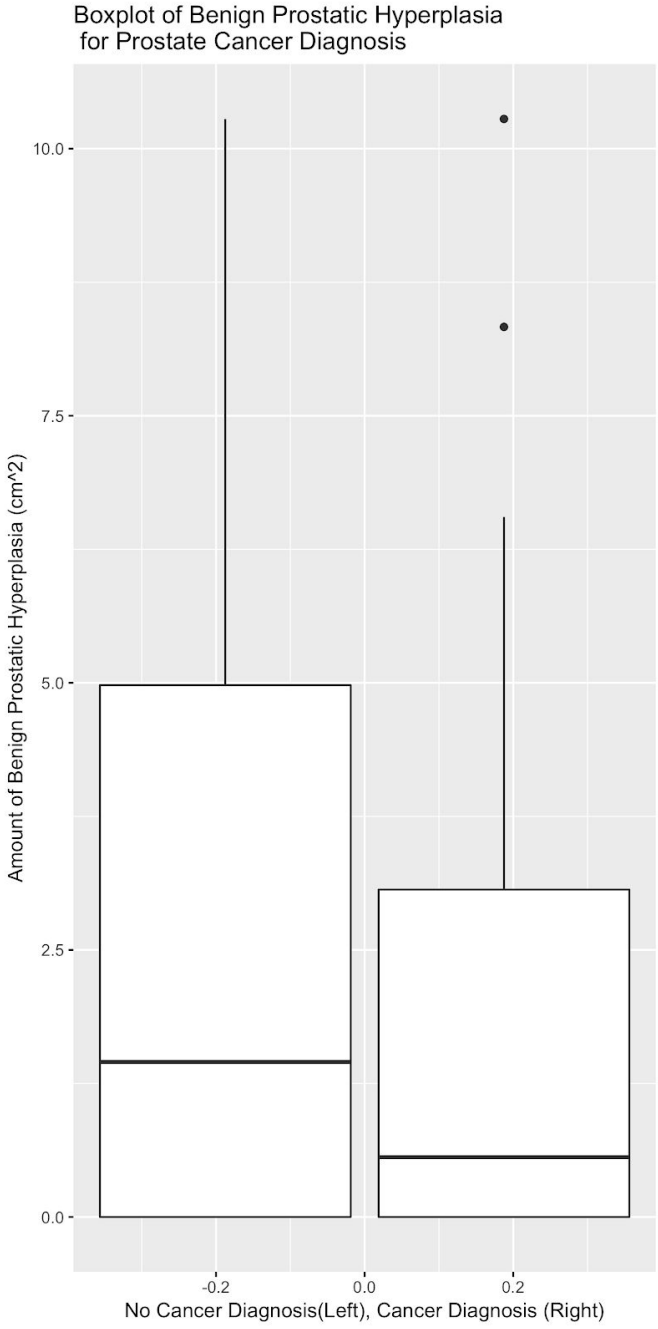


potential influential points will be talked about more in the Data Preparation section of this paper.

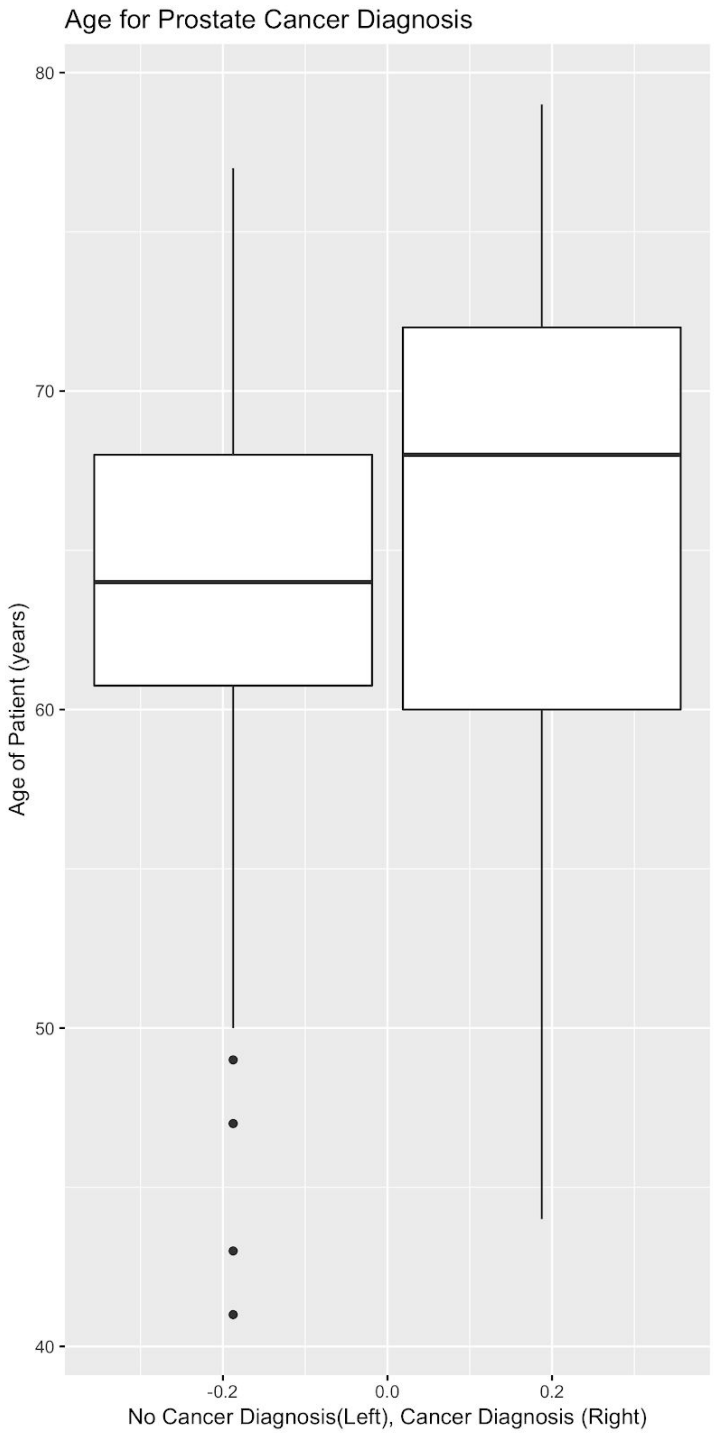
The above graphs had the y - axis range limited to allow us to better see the plots. This was done because the presence of influential points greatly increased the plot size. The left graph shows the Prostate Specific Antigen Level  level for different diagnoses. We can see that the median psa is quite higher for those with a diagnosis of



Prostate Cancer. On the right is the prostate weight for the different cancer diagnoses. We see that the median is higher for those with a cancer diagnosis as well.

The above graph on the left shows the Degree of Capsular Penetration for different Prostate Cancer diagnoses. We can see that for no cancer diagnoses a majority of the points are 0 since the median is 0. For cancer diagnoses, the median is once again higher than that of the no cancer diagnosis group. On the right is the graph of the amount of Benign Prostatic Hyperplasia for each cancer diagnosis. Contrary to the past plots the median for the no cancer group is lower than that of the cancer group.

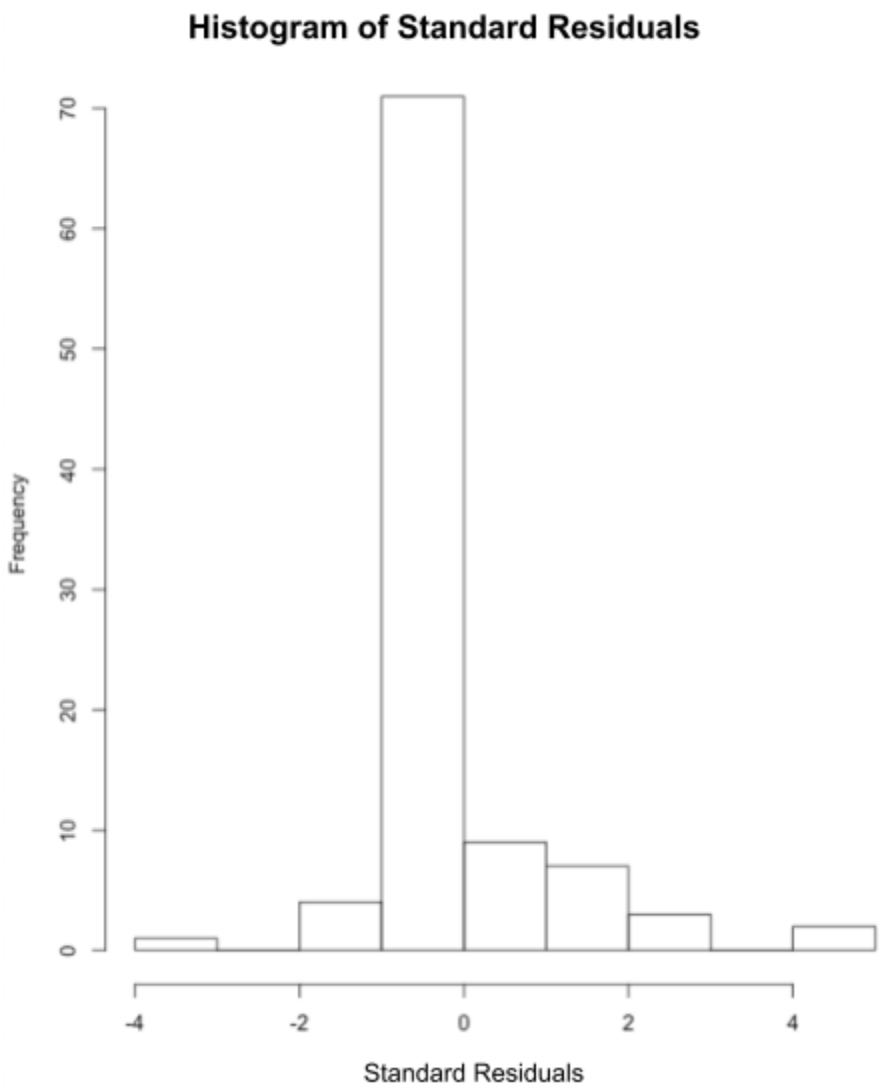



On the left is the boxplot of the age of patients and their cancer diagnosis. Once again, the cancer group has a higher median than the the no cancer group.

## Data Preparation:

The full log-odds model was used below for investigating influential points and outliers.

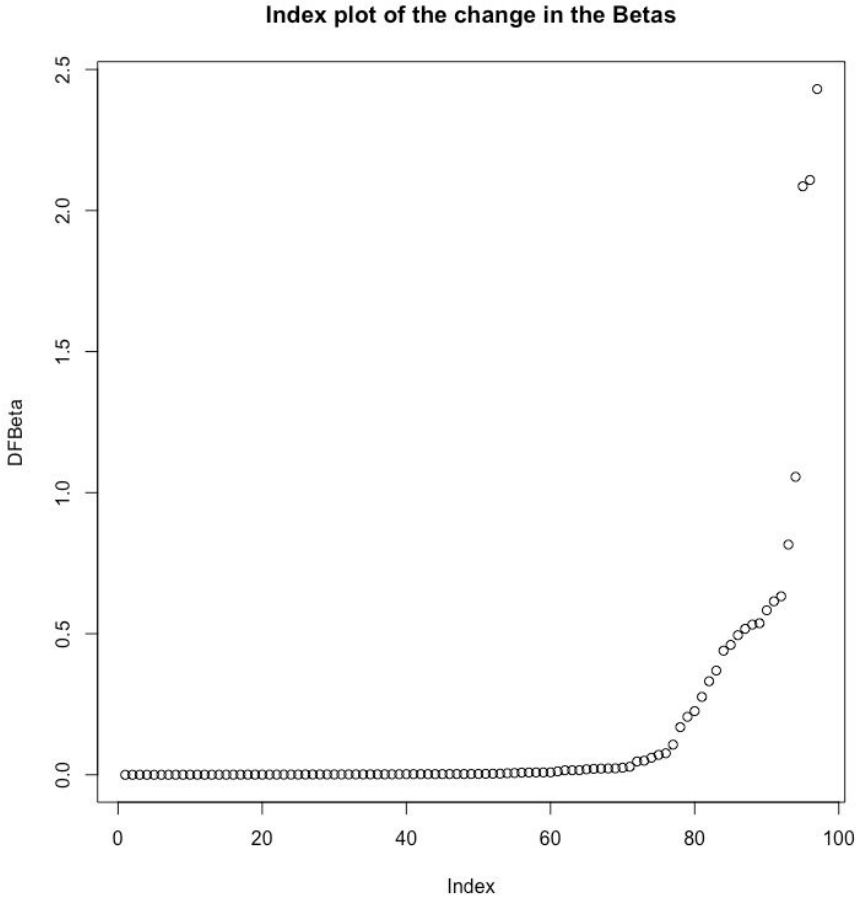
The right is a histogram of the standard residuals for the observations in the dataset. From this histogram we can see that there are a few outliers on both sides of the histogram. Investigating further, the observations with standard residuals with a magnitude greater than 3 are:




**Significant Standardized Residuals Table**

| psa | c.vol | weight | age | benign | inv | cap | Standard Residual |
|---|---|---|---|---|---|---|---|
| 9.974 | 1.8589 | 23.104 | 60 | 0.0000 | 1 | 0.0000 | 4.997933 |
| 21.758 | 6.2965 | 25.534 | 60 | 1.5527 | 0 | 3.2544 | 4.703014 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 56.261 | 25.7903 | 60.340 | 68 | 0.0000 | 1 | 0.0000 | -3.880665 |




On the left is the index plot of the change in beta hats for each observation. This was measured and recorded as df.beta and the cut off for significant df.betas was if they were greater than 1. The significant observation were:

**Significant df.beta Table**

| psa | c.vol | weight | age | benign | inv | cap | Change in Beta Hat |
|---|---|---|---|---|---|---|---|
| 21.758 | 6.2965 | 25.534 | 60 | 1.5527 | 0 | 3.2544 | 1.055984 |
| 14.880 | 23.3361 | 33.784 | 59 | 0.0000 | 1 | 0.0000 | 2.085914 |
| 7.463 | 1.1972 | 450.339 | 65 | 5.4739 | 1 | 0.0000 | 2.108261 |
| 56.261 | 25.7903 | 60.340 | 68 | 0.0000 | 1 | 0.0000 | 2.430568 |

The significant df.beta table does not include all of the observations in the standard residual outlier table. This may be because of the cutoff for the df.beta was too high and did not capture all of the outlier points. Or that the outliers observed did not have a significant effect on beta hats.

In addition, the df.beta table contains points that were not in the standard residual outlier table. This is due to the fact that influential points are not always outliers. It is also not always a bad for an observation to be an influential point. However, the outliers observed in the significant df.beta table will be removed before model selection.

## Model Selection/Analysis:

**Subset Selection Table**

| | Criterion | Value | Model |
|---|---|---|---|
| Forward Backward Subset | BIC | 63.06939 | y ~ c.vol + psa |
| Backward Forward Subset | BIC | 57.27822 | y ~ c.vol + psa + age + inv |
| All Subset | BIC | 57.27822 | y ~ c.vol + psa + age + inv |

The above table shows the model outputs and criterion values used for a particular subset selection method. The BIC criterion was mainly used since it puts a higher penalty on bigger models. The models for BIC did not converge for forward backward subsetting and backward forward subsetting. Because of this the all subsetting method was used and the best model was found to be y ~ c.vol + psa + age + inv.

The no interaction model is:

$$\ln(\pi /(1 - \pi)) = -19.09174266 + 0.08372397 * X_1 + 0.22971340 * X_2 + 0.17706457 * X_3 + 2.76134535 * X_4$$

Where $X_1$ is the psa, $X_2$ is the c.vol, $X_3$ is the age, and $X_4$ is inv.

Wald 99% confidence intervals were done for the beta hat values and the results were:

| | Lower Bound | Upper Bound |
|---|---|---|
| Intercept | -35.477689863 | -2.7057955 |
| c.vol | 0.044304847 | 0.1652061 |
| psa | 0.002241841 | 0.09892968 |
| age | -0.018944606 | 0.3730737 |
| invno-invasion | -0.588835610 | 6.1115263 |

Below is the log test for interactions in this model. A log-likelihood test to drop the interaction term from the model was used with:

**Log-Likelihood Tests for Interactions Table**

| | $G^2$ | d.f. | p-value |
|---|---|---|---|
| $H_0$: No interaction between c.vol*psa<br>$H_1$: Interaction between c.vol*psa is present | 0.1737480 | 1 | 0.6768021 |
| $H_0$: No interaction between psa*age<br>$H_1$: Interaction between psa*age is present | 0.1667525 | 1 | 0.6830142 |
| $H_0$: No interaction between psa*inv<br>$H_1$: Interaction between psa*inv is present | 0.1422010 | 1 | 0.7061026 |
| $H_0$: No interaction between c.vol*age<br>$H_1$: Interaction between c.vol*age is present | 0.08620778 | 1 | 0.76905470 |
| $H_0$: No interaction between c.vol*inv | 0.09543415 | 1 | 0.75737920 |

$H_1$: Interaction between c.vol*inv is present

$H_0$: No interaction between age*inv
$H_1$: Interaction between age*inv is present 1.0122075 1 0.3143746

Based on these tests we cannot say there are interaction effects between any of the variables in the model.

## Interpretation:

For the best model, as the prostate specific antigen level increases by 1 mg/ml the odds of having a diagnosis of Prostate Cancer vs not are multiplied by 1.087329, holding all other variables constant. As the prostate cancer volume estimate increases by 1 cc the odds of having a diagnosis of Prostate Cancer vs not are multiplied by 1.258239, holding all other variables constant. As the age increases by 1 year the odds of having a diagnosis of Prostate Cancer vs not are multiplied by 1.193708, holding all other variables constant. The odds of having a diagnosis of Prostate Cancer for someone who had seminal vesicle invasion is 15.82111 times that of not having a diagnosis of Prostate Cancer, holding all other variables constant. The odds of someone having a Prostate Cancer without having a seminal vesicle invasion is $5.111655 * 10^{-09}$ times that of not having a diagnosis of Prostate Cancer, holding all other variables constant.

The Wald CIs tell us that the true change in odds for having Prostate Cancer are 99% likely to be between [1.002244, 1.17963620] for prostate specific antigen level, [1.045301, 1.51455542] for prostate cancer volume estimate, [.9812337, 1.45219143]

for age, [.549731, 451.02659687] for those who had seminal vesicle invasion, and [3.910525* $10^{-16}$, 0.06681715] for those who did not have seminal vesicle invasion. One thing to note is that the CIs for age and seminal vesicle invasion both contains the value 1 which suggests that we may drop the corresponding variables from the model.

The interaction tests tell us that at any reasonable alpha level we cannot say there are any interactions for our model. If there were no interaction between

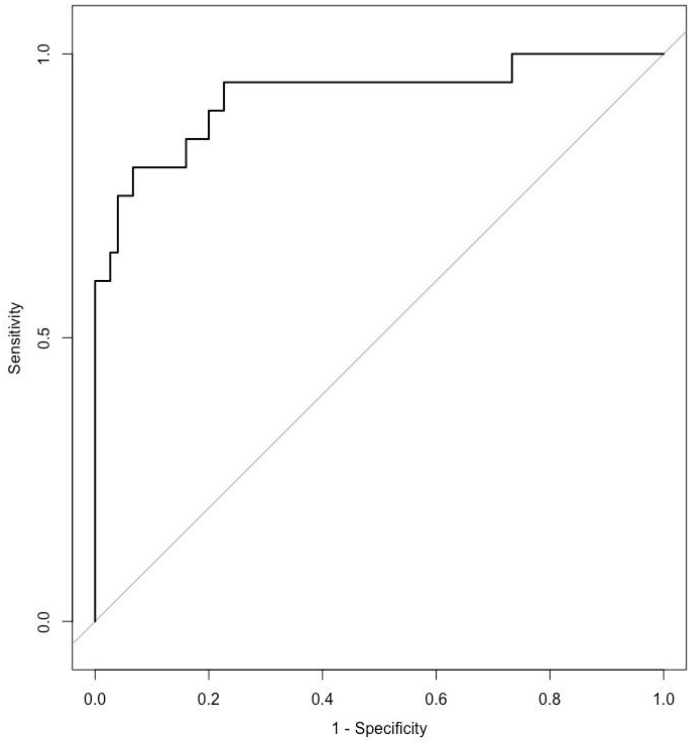
- prostate specific antigen level and prostate cancer volume estimate we would observe this data 67.6802% of the time.
- prostate specific antigen level and age we would observe this data 68.30142% of the time.
- prostate specific antigen level and seminal vesicle invasion we would observe this data 70.61026% of the time.
- age and prostate cancer volume estimate we would observe this data 76.905470% of the time.
- seminal vesicle invasion and prostate cancer volume estimate we would observe this data 75.737920% of the time.
- age and seminal vesicle invasion we would observe this data 31.43746 % of the time.

## Prediction:

Based on the model above, the probability that a person who has 10 psa, 5 c.vol, 40g for weight, age 67, with 2.5 benign, with no seminal vesicle invasion, and with 0.5 cm cap has a diagnosis of Prostate Cancer is 1.08502%.

The error matrix for the model used is:

| | |
|---|---|
| Sensitivity: | 0.6500000 |
| Specificity: | 0.9600000 |
| Error-Rate: | 0.1052632 |




Based on the error matrix, the sensitivity, correctly predicting a diagnosis of Prostate Cancer is 65%, the specificity, correctly predicting a no diagnosis of Prostate Cancer is 96%, and the error-rate, overall proportion of wrong predictions is 10.52%.

Further, the ROC was plotted to the left and the area under the curve was found to be 0.9253. The model used has an extremely good fit since the AUC is 0.9253. A 99% Delong confidence interval was done on the AUC and the result was that the true AUC is 99% likely to be between [0.8479, 1].

Next, the proportion of reduction in error for the best model was calculated to be 0.4110331. The proportion of reduction in error represents how much of the error is reduced when using our model to predict the diagnosis of Prostate Cancer as opposed to using the mean of Prostate Cancer diagnoses.

## Conclusion:

In conclusion, the most significant variables found for predicting a diagnosis of Prostate Cancer are Prostate Specific Antigen Level, Prostate Cancer Volume Estimate, Age, and the presence of Seminal Vesicle Invasion. Due to the low number of predictor variables used, the proportion of reduction in error was somewhat low. However, since we were interested in model correctness this was okay since we did not want to include variables that are not statistically significant.