EASTERN MICHIGAN UNIVERSITY

MASTER'S THESIS

DEPARTMENT OF MATHEMATICS AND STATISTICS

Prostate Cancer: Multiple Logistic Regression

Author
JEFFREY OSIWALA

Supervisor
Prof. KHAIRUL ISLAM

October 23, 2020

1 Proposal

In a research study, a university medical center urology group was interested in the association between prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies. The data given has identifications numbers, and provides information on 8 other variables on each person. The 8 variables being: PSA Level, Cancer Volume, Weight, Age, Benign Prostatic Hyperplasia, Seminal Vesicle Invasion, Capsular Penetration, and Gleason Score.

With this available data set, I will carry out a complete logistic regression analysis by first creating a binary response variable Y, called high-grade-cancer, by letting Y=1 if Gleason Score equals 8, and Y=0 otherwise (i.e., if Gleason Score equals 6 or 7). Thus, the response of interest is high-grade-cancer (Y), and the pool of predictors include those previously mentioned.

My analysis will consider transformations of predictors, the inclusion of second-order predictors, analysis of residuals and influential observations, model selection, goodness of fit evaluation, and the development of an ROC curve. Additionally, I will discuss the determination of a prediction rule for determining whether the grade of disease is predicted to be high grade or not, model validation, and finally asses the strengths and weaknesses of my final model.

2 Rationale

Prostate Cancer is the most common cancer in American men. The American Cancer Society (ACS), a nationwide voluntary health organization, estimates 191,930 new cases of prostate cancer and over 33,000 deaths in year 2020 alone. Additionally, the typical cost of therapy to a prostate cancer patient is $2,800/month after diagnosis (primarily from surgery and subsequently from office visits). A reliable and well understood testing/screening procedure needs to be in place support early detection, and to minimize these current and unforgiving metrics.

Research suggests that prostate cancer typically begins as a pre-cancerous condition, and these conditions are sometimes found when a man has an invasive prostate biopsy (the removal of small pieces of the prostate to look for cancer.) If prostate cancer is found early as a result of *screening*, it will probably be at an earlier and more treatable stage than if no screening were done. While this might seem like prostate cancer screening would always be a good things, there are still issues surrounding screening procedures that make it unclear if the benefits outweigh the risks for most men.

For example, the popular PSA screening test is not 100% accurate. This test can sometimes have abnormal results even when a man does not have cancer (false-positive result), or normal results when a man does have cancer (false-negative result). Consequently, false-positive results can lead to some men to get prostate biopsies (with risks of pain, infection, and bleeding) when they do not have cancer, and false-negative results can give men a false sense of security even though they may actually have cancer.

Another important issue is that even if screening does detect prostate cancer, doctors often cannot tell if the cancer is truly dangerous and needs to be treated. Prostate cancer can grow so slowly that it may never cause a man problems in his lifetime, and some men who seek screening may be diagnosed with a prostate cancer that they would have never known about otherwise. It would never have led to their death, or even cause any symptoms. Finding a "disease" like this that would never cause problems is known as **overdiagnosis**.

The problem with overdiagnosis in prostate cancer is that many of the men might still be treated with either surgery or radiation, either because the doctor cannot be sure how quickly the cancer might grow or spread, or the man is uncomfortable knowing he has cancer and is not receiving any treatment. The treatment of a cancer that would never have caused any problems is known as **overtreatment**, and the major downsides after surgery or radiation may include urinary, bowl, and/or sexual side effects that can seriously affect a man's quality of life. Thus, men and their doctors often struggle to decide if treatment is needed, or if the cancer can just be closely watched without being treated right away. Even when men are not treated right away, they still need regular blood PSA test and prostate biopsies to determine if their need for treatment in the future.

For now, the ACS recommends that men thinking about getting tested for prostate cancer learn as much as they can so they can make informed decisions based on available information, discussions with their doctors, and their own views on the possible benefits, risks, and limits of prostate cancer screening. To combat and better navigate these difficulties, research needs to continue growing the understanding of prostate cancer, and to build stronger predictive models which can improve the outlook of male lives, and also alleviate undo strain on the health care system.

3 Literature Review

3.1 Predictor Variables

An understanding of the predictor variables in this particular study can be seen as follows:

- **PSA Level**: Serum prostate-specific antigen level [mg/ml].

  -Prostate cancer can often be found early by testing for prostate-specific antigen (PSA) levels in a man's blood. However, the PSA test is not 100% accurate. (CANCER.ORG)
  - The chance of having prostate cancer increases as PSA level increases, but there is no set cutoff point that can tell for sure if a man does or does not have prostate cancer.

- **Cancer Volume**: Estimate of prostate cancer volume [cc].

  -Studies have suggested that inflammation of the prostate gland (prostatitis) may be linked to an increased risk of prostate cancer, but other studies have not found such a link.
  -Inflammation is often seen in samples of prostate tissue that also contain cancer. The link between the two it not clear, and it remains an active area of research. (CANCER.ORG)

- **Weight**: Prostate weight [gm].

  -As related to cancer volume, studies have suggested that inflammation (and an increase is prostate weight) may be linked to an increased risk of prostate cancer. This relationship remains an active area of research. (CANCER.ORG)

- **Age**: Age of patient [years].

  -Prostate cancer is rare in men younger than 40, but the chance of having prostate cancer rises rapidly after age 50. About 6 in 10 cases of prostate cancer are found in men older than 65. (CANCER.ORG)

- **Benign Prostatic Hyperplasia**: Amount of benign prostatic hyperplasia [cm^2]

  -BPH is a term used to describe common, benign type of prostate enlargement caused by an increased number of normal prostate cells. This condition is more common as men get older and is not currently known to be linked to cancer. (CANCER.ORG)

- **Seminal Vesicle Invasion**: Presence of absence of seminal vesicle invasion: 1 if yes; 0 otherwise.

  -SVI is the presence of prostate cancer in the areolar connective tissue around the seminal vesicles and outside the prostate.(NCBI.NLM.NIG.GOV)

- **Capsular Penetration**: Degree of capsular penetration [cm].

  -Cancer that has reached the outer wall of an organ (i.e. the prostate) is referred to as capsular penetration. Conversely, if cancer is strictly confined to the organ itself it is called organ-confined cancer. (PFC.ORG)

- **Gleason Score**: Pathologically determined grade of disease using total score of two patterns (summed scores were either 6, 7, or 8 with higher scores indicating worse prognosis).

  -A measure of how likely the cancer is to grow and spread quickly. This is typically determined by the results of the prostate biopsy, or surgery. (CANCER.ORG)

3.2 Related Research

Doctors are still studying if screening tests will lower the risk of death from prostate cancer. The most recent results from two large studies show conflicting evidence, and unfortunately did not offer clear answers.

The outcomes of both studies can be summarized as follows:

- Early results from a large study done in the United States found that annual screening with PSA and DRE (digital rectal exam - for a DRE, the doctor puts a gloved, lubricated finger into the rectum to feel the prostate gland) did detect more prostate cancers than in men not screened, but this screening did not lower the death rate from prostate cancer. However, questions have been raised about this study, because some men in the non-screening group actually were screened during the study, which may have affected the results.

- A European study did find a lower risk of death from prostate cancer with PSA screening (done about every 4 years), but the researchers estimated that roughly 781 men would need to be screened (and 27 cancers detected) to prevent one death from prostate cancer.
- Neither of these studies has shown that PSA screening helps men live longer overall (i.e. lowers the overall death rate).

Prostate cancer is often slow-growing, so the effects of screening in these studies might become more clear in coming years. Also, both of these studies are being continued to see if a longer follow-up will give clearer results.

4 Design and Analysis

To best model the dichotomous response variable, Y_HighGradeCancer, in the Prostate Cancer case study, I will employ a multiple logistic regression model, where 1 indicates high grade cancer and 0 indicates not high grade cancer.

In statistics, if $\pi = f(x)$ is a probability then $\frac{\pi}{1-\pi}$ is the corresponding *odds*, and the **logit** of the probability is the logarithm of the odds:

$$logit(\pi) = log(\frac{\pi}{1-\pi}) \tag{1}$$

Now, simple logistic regression means assuming that $\pi(x)$ is related to $\beta_0 + \beta_1 x$ (the *logit response function*) by the logit function. By equating $logit(\pi)$ to the logit response function (Eqn. X), we understand that the logarithm of the odds is a linear function of the predictor. In particular, the slope parameter β_1 is the change in the log odds associated with a one-unit increase in x. This implies that the odds itself changes by the multiplicative factor e^{β_1} when x increases by 1 unit.

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x \tag{2}$$

From here, straightforward algebra will then show the Simple Linear Regression Model:

$$E[Y] = \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{3}$$

Next, this simple logistic regression model is easily extended to more than one predictor variable by inclusion of the following two vectors, in matrix notation:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_{p-1} \end{bmatrix}$$

With this notation, the simple logistic response function (1) extends to the multiple logistic response function as follows:

$$E[Y] = \pi(\mathbf{X}) = \frac{exp(\mathbf{X}'\boldsymbol{\beta})}{1 + exp(\mathbf{X}'\boldsymbol{\beta})} \tag{4}$$

Fitting the logistic regression to the sample data requires that the parameters β_0, $\beta_1, \cdots, \beta_{p-1}$ be estimated. This will be done using the maximum likelihood technique provided within the statistical packages of both **R** and Python.

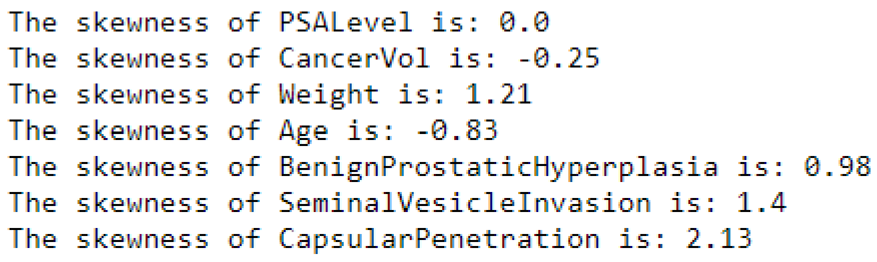
4.1 Data Transformations and Standardization

In modeling using logistic regression, the appropriate transformations on continuous variables are necessary to optimize the model predictiveness.

Variable transformation is an important technique to create robust models using logistic regression. Because the predictors are linear in the log of the odds, it is often helpful to transform the continuous variables to create a more linear relationship.

The raw data collected contained several predictors with high skewness values. A few concerning features were determined to be PSA Level (skewness = 4.39), Cancer Volume (skewness = 2.18), and Weight (skewness = 7.46). As a prepossessing step to reduce skewness, I elected to transform these continuous predictor variables using the log-transformation, and standardize *all* the data on top of that. The standardization step was used to normalize the data, did not affect any underlying distributions, and was performed by using the following design:

The finalized data skewness is summarized directly below. Following, I've included the histogram of PSA Level vs. Cancer Volume vs. Age, a helpful visual for the three predictors which carried the most significance through much of my analysis.



```
The skewness of PSALevel is: 0.0
The skewness of CancerVol is: -0.25
The skewness of Weight is: 1.21
The skewness of Age is: -0.83
The skewness of BenignProstaticHyperplasia is: 0.98
The skewness of SeminalVesicleInvasion is: 1.4
The skewness of CapsularPenetration is: 2.13
```

Figure 1: insert caption here.

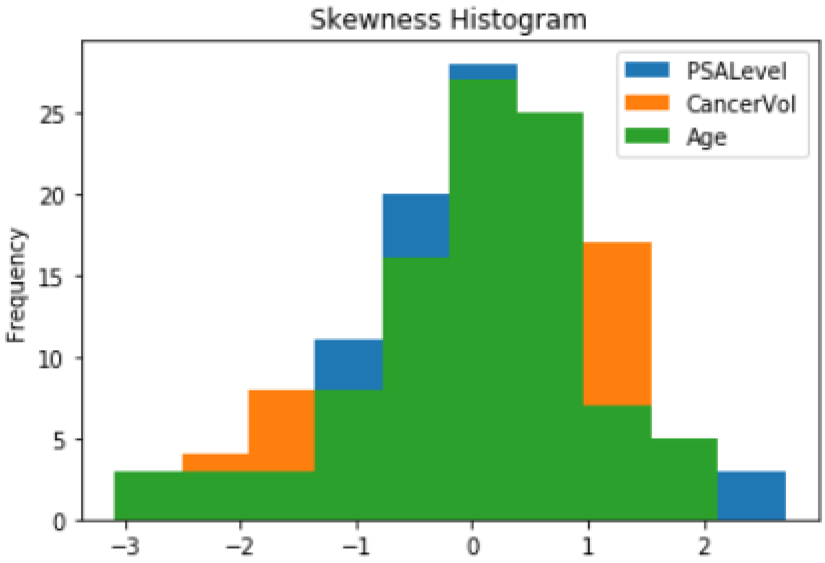


Figure 2: insert caption here.

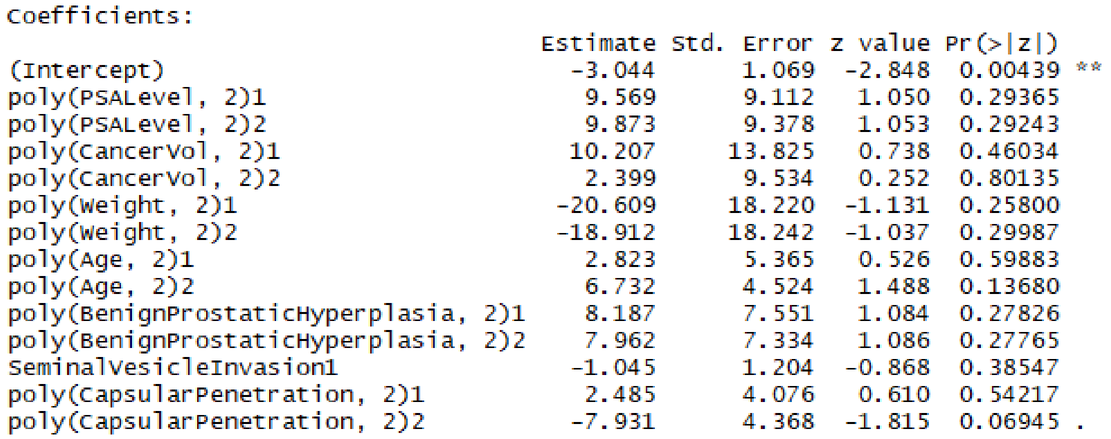
4.2 Second-Order Predictors

Occasionally, the first-order logistic model may not provide an adequate fit to the data, so I began my analysis by first attempting to fit the Prostate Cancer data to a *polynomial logistic* regression model. For simplicity, a 2nd-order polynomial model in two predictors has a logit response function as

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \tag{5}$$

and can be extended to more predictors by the inclusion of additional variables, their coefficients, and accompanying cross terms. Please recall, the Prostate Cancer data set considers 7 predictors.

In many situations the true regression function XXX has one or more peaks or valleys, and in such cases a polynomial function can provide a satisfactory approximation to XXX. However, a polynomial fit was not successful here, as indicated by *non-significant* p-values across all predictors, at 5% significance. Additionally, my preliminary scatter plot analysis did not indicate any reason to believe a polynomial fit would be suitable in this study. Without hesitation, I will move forward with the development of a multiple logistic *linear* regression model.



```
Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                               -3.044      1.069  -2.848  0.00439 **
poly(PSALevel, 2)1                         9.569      9.112   1.050  0.29365
poly(PSALevel, 2)2                         9.873      9.378   1.053  0.29243
poly(CancerVol, 2)1                       10.207     13.825   0.738  0.46034
poly(CancerVol, 2)2                        2.399      9.534   0.252  0.80135
poly(Weight, 2)1                         -20.609     18.220  -1.131  0.25800
poly(Weight, 2)2                         -18.912     18.242  -1.037  0.29987
poly(Age, 2)1                              2.823      5.365   0.526  0.59883
poly(Age, 2)2                              6.732      4.524   1.488  0.13680
poly(BenignProstaticHyperplasia, 2)1       8.187      7.551   1.084  0.27826
poly(BenignProstaticHyperplasia, 2)2       7.962      7.334   1.086  0.27765
SeminalVesicleInvasion1                   -1.045      1.204  -0.868  0.38547
poly(CapsularPenetration, 2)1              2.485      4.076   0.610  0.54217
poly(CapsularPenetration, 2)2             -7.931      4.368  -1.815  0.06945 .
```

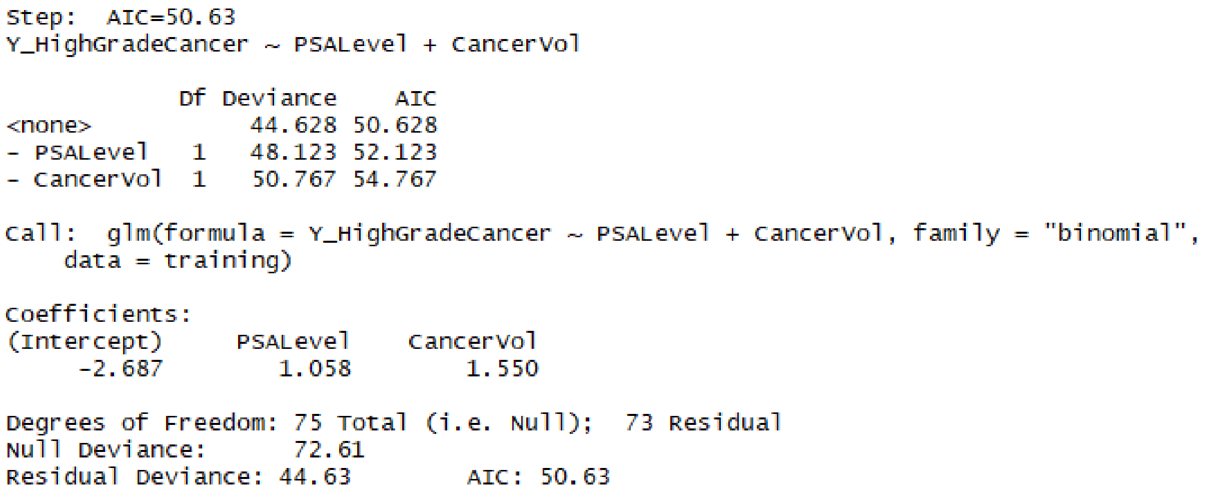
Figure 3: insert caption here

4.3 Model Selection

4.3.1 Best Subsets Procedure

The procedure outlined here will help identify a group of subset models that give the best values of a specified criterion. This technique has been developed by time-saving algorithms which can find the most promising models, without having to evaluate all 2^{p-1} candidates. The use of the best subset procedure is based on the AIC_p criteria, where promising models will yield a relatively small value.

The minimized AIC_p stepwise output given by **R** is provided in Figure XXX below.



```
Step:  AIC=50.63
Y_HighGradeCancer ~ PSALevel + CancerVol

            Df Deviance    AIC
<none>           44.628 50.628
- PSALevel   1   48.123 52.123
- CancerVol  1   50.767 54.767

Call:  glm(formula = Y_HighGradeCancer ~ PSALevel + CancerVol, family = "binomial",
    data = training)

Coefficients:
(Intercept)     PSALevel    CancerVol
     -2.687        1.058        1.550

Degrees of Freedom: 75 Total (i.e. Null);  73 Residual
Null Deviance:      72.61
Residual Deviance: 44.63        AIC: 50.63
```

Figure 4: insert my caption here

In this procedure, I instructed **R** to iterate "backwards" through all 7 predictor variables and it was determined AIC_p was minimized for $p = 3$. In particular, the results reveal that the best two-predictor model for this criteria is based on PSA Level and Cancer Volume.

4.3.2 Final Model

The data of 97 individual men in the Prostate Cancer sample was split at 80% for train and test sets. The training set is a random 76 observations and was used for fitting the model, and the remaining 21 cases were saved to serve as a validation data set. Table XXX in columns 1-X contains the variables... The primary purpose of the study was to asses the strength of the association between each of the predictor variables, the predictable nature of PSA Level, and the probability of a man having been diagnosed with high grade prostate cancer over low grade.

A first-order multiple logistic regression model with two predictor variables was considered to be reasonable by §4.3:

$$\pi(\mathbf{X}) = \frac{exp('\boldsymbol{\beta})}{1 + exp(\mathbf{X}'\boldsymbol{\beta})} = [1 + exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \tag{6}$$

where:

$$\mathbf{X}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{7}$$

This model was fitted by the method of maximum likelihood to the data from the 76 train cases. The results are summarized in Table XXX below. The estimated logistic response function is:

$$\hat{\pi} = [1 + exp(-2.6867 + 1.0577X_1 + 1.5502X_2))]^{-1} \tag{8}$$

Now, the interpretation for multiple logistic regression is that the estimated odds ratio for the predictor variable X_k assumes that all other predictor variables are held constant. We can see, for instance, that the odds of a man being diagnosed with high grade prostate cancer increase by about 105% for each additional score of PSA Level, for a given Cancer Volume. This means each unit increase of PSA Level approximately doubles the odds of said diagnosis.

4.3.3 Geometric Interpretation

When fitting a standard multiple logistic regression model with two predictors, the estimated regression shape is an S-shaped surface in three-dimensional space. Figure XXX displays a three-dimensional plot of a logistic response function that depicts the relationship between the diagnosis of high grade prostate cancer (Y, the binary outcome) and two continuous predictors, PSA Level (X_1) and Cancer Volume (X_2). This surface increases in an approximately linear fashion for larger values of PSA Level and Cancer Volume, but levels off and is nearly horizontal for small values of these predictors.

With the estimated logistic regression equation now developed, it is left to analyze the residuals, test goodness of fit, and finally apply the model to the test data and discuss the results.

4.4 Analysis of Residuals

4.4.1 Logistic Regression Residuals

4.4.2 Influential Observations

4.5 Goodness Of Fit Evaluation

The appropriateness of the fitted logistic regression model needs to be examined before it is accepted for use. In particular, we need to examine whether the estimated response function for the data is monotonic and sigmoidal in shape, as are logistic response functions. Here I will employ the Hosmer-Lemeshow test, which is useful for unreplicated data sets, as is the Prostate Cancer data. The test can detect major departures from a logistic response function, and the alternatives of interest are as follows:

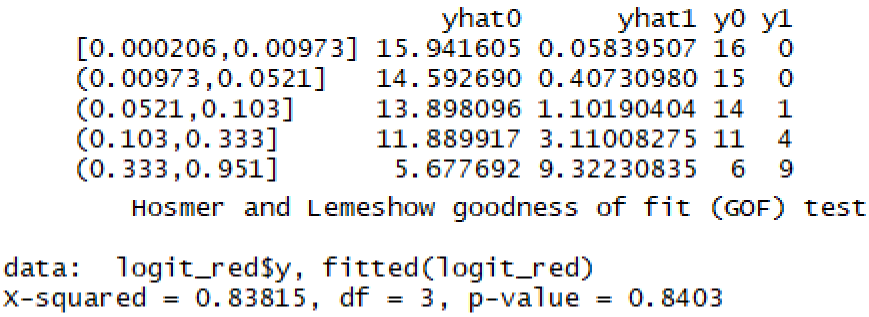
$$\begin{aligned} H_0 &: E[Y] = [1 + exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \\ H_0 &: E[Y] \neq [1 + exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \end{aligned} \tag{9}$$

4.5.1 Hosmer-Lemeshow

The Hosmer-Lemeshow Goodness of Fit procedure consists of grouping that data into classes with similar fitted values $\hat{\pi}_i$, with approximately the same number of cases in each class. Once the groups are formed, the Hosmer-Lemeshow goodness of fit statistic is calculated by using the Pearson chi-square test statistic of observed and expected frequencies. The test statistic is known to be well approximated by the chi-square distribution with $c - 2$ degrees of freedom.

$$\chi^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \tag{10}$$

The output from **R** using 5 groups is shown in Figure XXX.



```
                          yhat0      yhat1 y0 y1
[0.000206,0.00973] 15.941605 0.05839507 16  0
(0.00973,0.0521]   14.592690 0.40730980 15  0
(0.0521,0.103]     13.898096 1.10190404 14  1
(0.103,0.333]      11.889917 3.11008275 11  4
(0.333,0.951]       5.677692 9.32230835  6  9
       Hosmer and Lemeshow goodness of fit (GOF) test

data:  logit_red$y, fitted(logit_red)
X-squared = 0.83815, df = 3, p-value = 0.8403
```

Figure 5: caption here

Large values of the test statistic X^2 indicate that the logistic response function is not appropriate. The decision rule for testing the alternatives in (Eqn. XXX), when controlling the level of significance at α, therefore is:

$$\begin{aligned} &\text{If } X^2 \leq \chi^2(1-\alpha; c-p), \text{ conclude } H_0 \\ &\text{If } X^2 > \chi^2(1-\alpha; c-p), \text{ conclude } H_a \end{aligned} \tag{11}$$

Thus, for $\alpha = 0.5$ and $c - 2 = 3$, we require $\chi^2(.95;3) = 7.81$. Since $X^2 = 0.838 \leq 7.81$, we conclude H_0, that the logistic response function is appropriate. The p-value of the test is 0.8403.

4.6 Development of ROC Curve

Multiple logistic regression is often employed for making predictions for new observations. The *receiver operating characteristic* (ROC) *curve* plots $P(\hat{Y} = 1|Y = 1)$ as a function of $1 - P(\hat{Y} = 0|Y = 0)$ is an effective way to graphically display prediction rule information, and possible cutoff points.

The "True Positive" y-axis on an ROC curve is also known as *sensitivity*, and the "False Positive" x-axis is 1-*specificity*. Figure XXX below exhibits the ROC curve for model XXX for all possible cut points between 0 and 1.

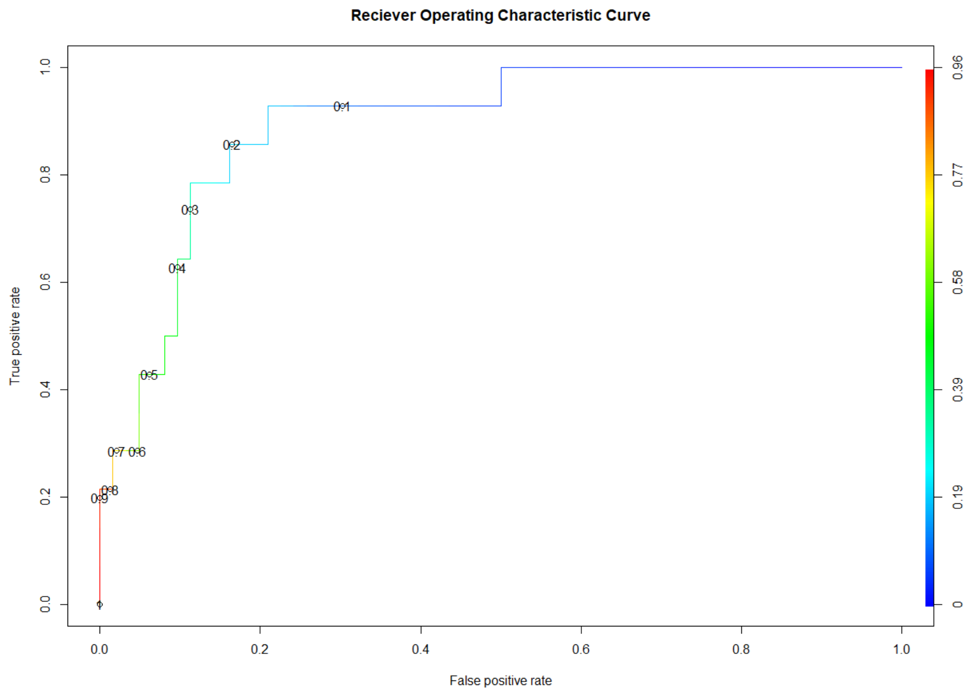


Figure 6: insert caption here

4.6.1 Prediction Rule

In the training data set (which represented a random 80% of the 97 provided observations), there were 14 men who were observed as high grade cancer patients; hence the estimated proportion of persons who had high grade cancer is $14/76 = 0.184$. This proportion can be used as the starting point in the search for the best cutoff in the prediction rule.

Thus, if $\hat{\pi}_h$ represents a newly fitted observation, my first prediction rule investigated was:

$$\text{Predict 1 if } \hat{\pi}_h \geq 0.184; \text{ predict 0 if } \hat{\pi}_h < 0.184 \tag{12}$$

Table XXX below provides a summary of the number of correct and incorrect classifications based on prediction rule XXX. Of the 62 men without high grade cancer, 13 would be incorrectly predicted to have high grade cancer, or an error rate of 21.0%. Furthermore, of the 14 persons with high grade cancer, 1 would be incorrectly predicted with rule XXX to not have it, or 7.1%. Altogether, $13 + 1 = 14$ of the 76 predictions would be incorrect, so that the prediction error rate for rule XXX is $14/76 = 0.184$ or 18.4%. Coincidentally, the model exactly matches our training set proportions with the current prediction rule.

Prediction Rule XXX			
True Classification	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 0$	49	13	62
$Y = 1$	1	13	14
Total	50	26	76

Table 1: Classification based on Logistic Response Function XXX and Prediction Rules.

With this baseline understood, it is straightforward to choose a stronger cutoff point in utilizing the ROC curve of Figure XXX. As detailed above, the false-positive rate is not ideal at 21.0%; there are too many cases where a man may opt for additional screening and treatment, even invasive actions, because he believes he has prostate cancer. It will be wise to now reference the ROC curve to better choose a prediction cutoff, while also not significantly disturbing the false-negative accuracy.

Looking at Figure XXX I see a step occurring at 0.20, and use this value for my cutoff candidate. The effects of this change can be summarized in the Confusion Matrix table below.

The model accuracy has now increased with a significantly better false-negative rate, thus intended to reduce the footprint across the health care economy.

Thus, my updated and final prediction rule is stated as follows:

$$\text{Predict 1 if } \hat{\pi}_h \geq 0.2\text{; predict 0 if } \hat{\pi}_h < 0.2 \tag{13}$$

4.7 Model: Strengths and Weaknesses

4.7.1 Strengths

4.7.2 Weaknesses

-gof has less than 5 expected frequencies -psa level and cancer volume correlations

5 Conlcusion

-Talk about "Factors that might affect PSA Levels".
-Study doesn’t indicate which type of prostate cancer we’re investigating. (eh)
-Concerns about early detection/testing. (already somewhat touched on already)

6 References

Sample text.
Test.