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

Use this section for introducing logistic regression.
High level...show the function, explain it.

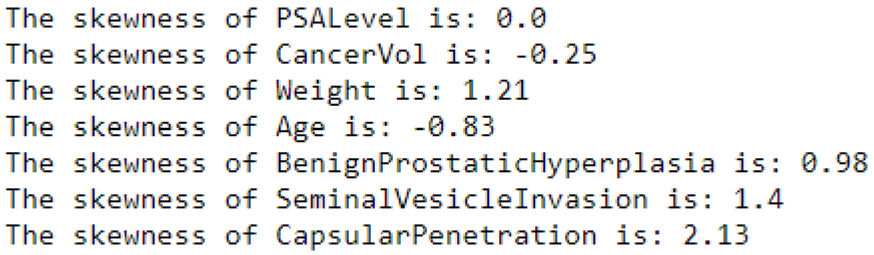
4.1 Data Transformations and Standardization

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

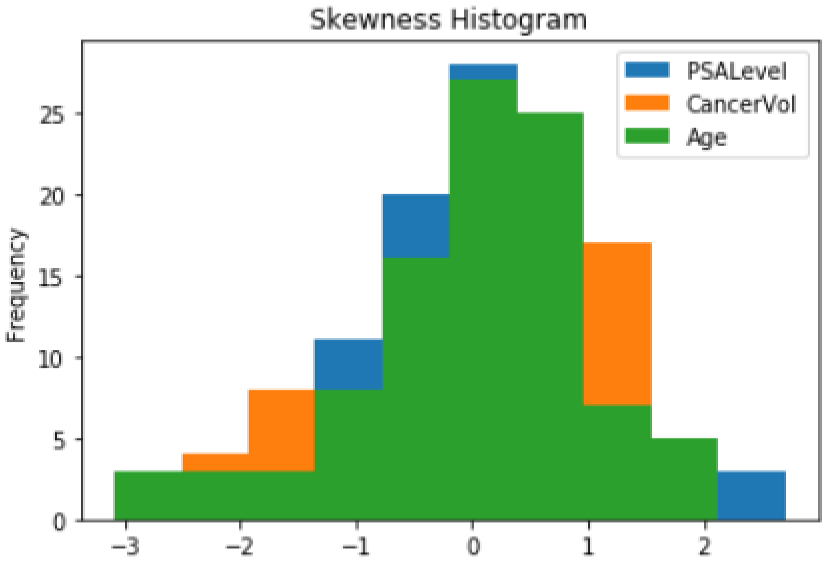
Variable transformation is an important technique to create robust models using logistic regression. Because the predictors are linear in the log of the odds, it is often helpful to transform the continuous variables to create a more linear relationship.

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
```



4.2 Second order Predictors

```
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 . 
```

4.3 Model Selection

4.4 Analysis of Residuals

4.4.1 Influential Observations

4.5 Goodness Of Fit Evaluation

4.6 Development of ROC Curve

4.6.1 Prediction Rule

4.7 Model: Strengths and Weaknesses

-discuss correlation matrix
-PSALevel and CancerVol show a mild level of correlation: 0.624151

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.