***Predicting Prostate-specific antigen from prognostic clinical measurements***

***Dawit Wolde***

**l’’**

**Department of Biostatistics**

**University of Kansas, USA**

**November 15, 2022**

**Title**

To predict prostate-specific antigen using eight prognostic clinical measurements of 97 men who are going to be undergoing radical proctectomies.

**Abstract**

It is important to study and predict prostate cancer as cancer rates are getting higher. According to PubMed Central, prostate cancer is in the top five cancer associated fatality among men and is the most common cancer following skin cancer. Prostate cancer is more likely to be found in man who are 65 or older. The most common way to screen for prostate cancer is to test prostate-specific antigen (PSA) level. While Prostate-specific antigen level above 4ng/ml is considered abnormal, it is considered alarming if the level is above 10ng/ml.

The prepose of this study is to find out the association between PCA level and predictor variables which are prognostic clinical measurements on 97 men with prostate cancer.

The seven predictor variables are cancer volume, weight, age, benign prostatic hyperplasia, seminal vesicle invasion, capsular penetration and Gleason score. The response variable for this study is prostate-specific antigen level.

**Introduction**

Age and body weight are believed to have the biggest association with the risk of prostate cancer. It has been studied that genetics, age(age) and weight(weight) are linked to risk of getting prostate cancer.

In this study, several prognostic clinical measurements including age and weight are the predictor variables. Estimate of prostate cancer volume(cancerv), presence or absence of seminal vesicle invasion(seminal), and amount of benign prostatic hyperplasia are also important predictors in this study. Other important predictor variables such as degree of capsular penetration (capsular) and scores where high scores indicating worse prognosis(score) are tested in this study.

Accordingly, the main purpose of this study is to understand the association between the response variable Serum Prostate-Specific antigen level (PSA level) and the predictor variables. By finding out the association, the appropriate predictors can be selected which can help to understand prostate cancer.

**Primary Analysis Objectives**

**Data source**

To observe linear association between the response variable prostate-specific antigen level and the predictor variables prognostic clinical measurements and find out the best predictors.

**Materials and Methods**

The data set was taken from online provided by a professor from the university of Kansas. The original dataset was collected by a university medical center urology group from 97 men who underwent radical proctectomies. The dataset includes the response variable **PSA level (PSA)** ranges from 0.651-265.072mg/ml, **Identification number(idnum)** ranges from 1-97, **cancer volume(cancerv)** in the range 0.2592 to 45.6042 cc, **weight(weight)** ranges from 10.697- 450.339 gm, **age(age)** ranges from 41-79, **benign prostatic hyperplasia(hyperplasia)** ranges from 0000-10.2779 cm^2,**seminal vesicle invasion(seminal)** ranges from 0-1:1 if yes; 0 otherwise, capsular penetration (capsular) ranges from0.0000-18.1741 cm, Gleason score(score) either 6,7 or 8 with 8 indicating worse prognosis.

**Statistical analysis**

The data used in the study is in .csv(csv) format. The data is analyzed by a statistical software R version 4.2.2(2022-03-10) and multiple regression is the statistical technique used to analyze the data. The importance of each predictor variable is individually studied. For the variable selection in regression, Automated variable selection method is used, and the best subset of predictors is identified.

**Model Assumption**

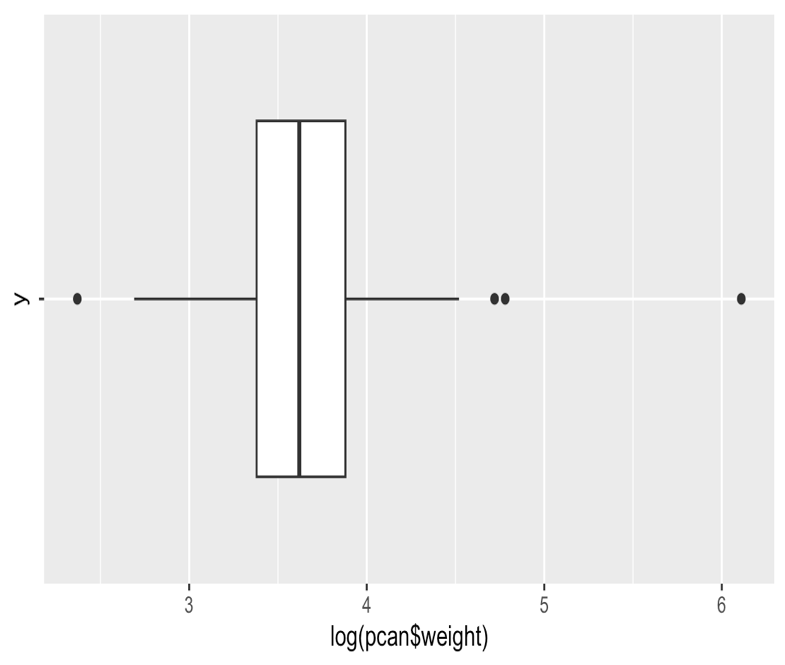
The significance alpha value used in this study is 0.05. By using mean, median, variance, maximum, minimum and standard deviation, the types of variables are identified. Most of the predictor variables are continuous variables and the two categorical variables in the data are seminal vesicle invasion and Gleason score.

**Primary Objective Analysis**

It is necessary to study each of the predictor variable before starting the actual analysis of the data. In order to have a model that is properly fit, a few important steps are needed which can be helpful by suggesting if data transformation required for a better fit. Exploring each variable in the data set and observing the degree of asymmetry by testing for skewness along with outliers is important to obtain a better fit model.

(b) Boxplot of weight

1. Boxplot of cancer volume



Chart, box and whisker chart

Description automatically generated

(c) Boxplot of age

Chart, box and whisker chart

Description automatically generated

(d)Boxplot of hyperplasia

Chart, box and whisker chart

Description automatically generated

(f) Boxplot of capsular

Chart, box and whisker chart

Description automatically generated

**Figure 1:** Analysis of predictors

**Analysis of the potential predictors** Figure 1 illustrates individual predictors distribution. Figure 1 (a) and (b) seem to be a roughly symmetrical and normally distributed. On the other hand, (d), (f) and (g) are right (Positively) skewed while (c) is left(negatively) skewed. Extreme values (outliers) are shown in (b)(c)(d) and (f).

**Analysis of the response variable prostate-specific antigen level**

1. (b)

Chart, box and whisker chart

Description automatically generated Chart, histogram

Description automatically generated  
 **Figure 2:** distribution of prostate-specific antigen level

The boxplot and Histogram for the response variable show a normal distributed data. The boxplot also shows there are few outliers.

**Scatter plot for predictors variables before model selection**

**Effect of cancer volume(cancerv) on Prostate-specific antigen level (PSA)**

The scatter plot of cancer volume and prostate-specific antigen is shown in figure 3. The regression equation of cancer volume on prostate-specific antigen can be shown as:

Yi=-11.16+1.11Xi

Yi signifies prostate -specific antigen and Xi signifies cancer volume. The model implies for every one cc increase in cancer volume, PSA increases by 1.11 mg per ml.

Chart, scatter chart

Description automatically generated

**Figure. 3**: scatter plot of cancer volume with Prostate-specific antigen level

**Effect of weight on Prostate-specific antigen level(psa)**

The scatter plot of weight and prostate-specific antigen is shown in figure 4. The regression equation of weight on prostate-specific antigen is written as:

Yi=-11.16+0.13Xi

Yi shows prostate-specific antigen and Xi shows weight. The model implies for every-one gm increase in weight, psa increases by 0.13 mg per ml.

Chart, scatter chart

Description automatically generated

**Figure 4:** scatter plot of weight with prostate-specific antigen level

**Effect of hyperplasia on Prostate-specific antigen level(psa)**

The scatter plot of hyperplasia and prostate-specific antigen is shown in figure 5. The regression equation of hyperplasia on prostate-specific antigen is shown below as:

Yi=-11.16+0.277Xi

Yi indicates prostate-specific antigen and Xi shows hyperplasia. The model shows for every-one cm2  increase in hyperplasia, psa increases by 0.277 mg per ml.

Chart, scatter chart

Description automatically generated

**Figure 5:** scatter plot of hyperplasia with prostate-specific antigen level

**Effect of age on Prostate-specific antigen level(psa)**

The scatter plot of age and prostate-specific antigen is shown in figure 6. The regression equation of age on prostate-specific antigen is written as:

Yi=-11.16-0.04Xi

Yi indicates prostate-specific antigen and Xi indicates age. The model shows for every one-yeardecrease in age, psa decreases by 0.04 mg per ml.

Chart, scatter chart

Description automatically generated

**Figure 6:** scatter plot of age with prostate-specific antigen level

**Effect of capsular on Prostate-specific antigen level(psa)**

The scatter plot of capsular and prostate-specific antigen is shown blow in figure 7. The regression equation of capsular on prostate-specific antigen is shown as:

Yi=-11.16-0.92Xi

Yi indicates prostate-specific antigen and Xi indicates capsular. The model shows for every-one cmdecrease in capsular, psa decreases by 0.92 mg per ml.

Chart, scatter chart

Description automatically generated

**Figure 7:** scatter plot of capsular with prostate-specific antigen level

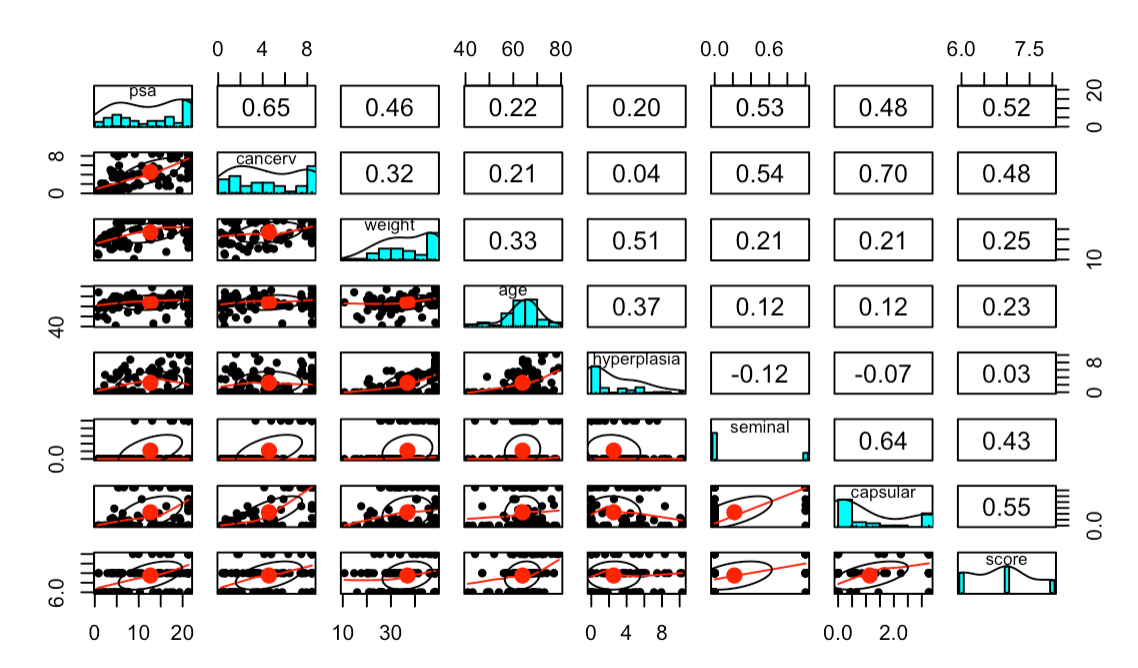
**Correlation matrix and scatterplot matrix between predictors**

Table 1 and figure 8 show the correlation between each of potential predictors. In table 1, the correlation coefficient r values are shown for cancerv, weight, age, hyperplasia, and capsular respectively. The response variable appears to have correlation between cancerv, weight and score. According to values of r, the predictor variables also show multicollinearity between each other. The closer r value is to 1(-/+), the higher the correlation. The r value between cancerv and capsular is 0.70 indicating there is a higher correlation between the two predictors compared to the correlation between other predictors.

A picture containing table

Description automatically generated

**Table 1: Correlation coefficient matrix**



**Figure 8: scatterplot matrix between predictors**

**Model selection**

**Automatic variable selection method**

By using the automatic variable selection method, the best for the data can be selected allowing the unnecessary variables to be eliminated. In order to perform the variable selection effectively, the leaps package in R is used to obtain R2, Mallow’s Cp and BIC. While the largest R2 andadj\_r2 is considered the best, the smallest Mallow’s Cp and BIC is considered the best for variable selection.

The final model is written as:

Yi = β0 + β1X1 + β2X2 + β3X3 + β4X4 +Ei

Where:

Yi is the Prostate-specific antigen level

X1 is cancer volume

X2 is weight

X3 is seminal

X4 is score

**Outliers and influential points**

1. Studentized plot (b) Cook’s distance

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Figure 9:** detecting outliers

The values that are distant from most of the values in cook’s distance showed in figure 9 are outliers. Another way of detecting outliers is by visualizing boxplots and checking the quantile percentage. In order to perform a better analysis, adjustments to the outliers are needed by changing the outlier values to the adjacent value (75% quantile) in the data pool.

**Residual diagnostics**

**Chart, histogram

Description automatically generated**

**Figure 3:** Residual diagnostics

Model assumptions for the data are shown in figure 3. Figure 3 (a) shows homoskedasticity where the model has equal variance. Figure 3 (b) and (c) show the normal distribution for error terms. The sequence plot in figure 3 (d) also shows the independence of error terms.

**Goodness of Fit Test**

linear regression models, Boxplot, Histogram, Scatterplot, correlation matrix, R2, adjusted R2a, and BIC along with t-test and p-value were used in order to support the purpose of this analysis.

All of data is analyzed by a statistical software R version 4.2.2(2022-03-10).

**Results**

The linear association between the response variable(psa) and each of the continuous predictor variable, two tailed-tailed t-test is used. There are two hypotheses for the t-test:

**Null hypothesis**: Ho: β1 = 0

**Alternative hypothesis**: Ho: β1 = 0

-t\*=b1-B1/SE(b1) is used for the decision.

-t\* is test statistics

-b1 is observed slope

-B1 is the regression model’s expected slope

-SE(b1) sampling variability of b1

To test the t\* it is compared the calculated t (1- 1 −α/2, df),

* The level of significance(α) =0.05
* Degree of freedom(df) = 1-n where n is sample size

If t\*>t (- α/2, df) and p-value ≤ α, the null hypothesis H0 is rejected.

R2 is also considered in making the decision. The proportion of variance for a response variable that is explained by the predictor variable in the model is represented by R2. R2 is in a range between 0-1. R2 that is closer to 1 considered better as the proportion of variance is explained by the predictor variables.

Text

Description automatically generated

**Table 2: coefficients for selected variables**

**Effect of cancer volume on Prostate-specific antigen level**

The significance level (α) = 0.05 is used to calculate t test. The t test suggests rejecting the null hypothesis Ho. A smaller p value (5.56e-05) for cancer volume also supports the rejection of Ho showing that there is a linear association between prostate-specific antigen level and cancer volume. The t test also supports the R2 explaining 42% of the unexplained variation in prostate-specific antigen level.

**Effect of weight on prostate-specific antigen level**

Using (α) = 0.05, t critical value (1.98) is calculated and compared to the t statistic value, and it suggests the rejection of the null hypothesis. The p value also supports the decision because it is smaller than the significance alpha level. There for the conclusion is that there is an association between weight and prostate-specific antigen. The R2 value shows that only 21% of the variation in prostate-specific antigen is explained leaving 79% of the variation unexplained.

**Effect of seminal on prostate-specific antigen level**

The significance alpha level (0.05) is used to calculate t critical value (1.98) in this study and compared to the t statistic value which was bigger. The alpha level is also bigger when compared with the p value. The decision is to reject the null hypothesis and thus giving and evidence of linear association between prostate-specific antigen level and seminal. R2 value shows 28% of the variation is explained in prostate-specific antigen leaving the other 72% unexplained.

**Effect of score on prostate-specific antigen level**

The alpha level (0.05) is greater than the p value and the t statistic value is greater than the calculated t critical value (1.98). The conclusion is to reject the null hypothesis showing that there is an association between score and prostate-specific antigen level. The R2 value is 0.27 suggesting that 27% of the variation in prostate-specific antigen level is explained.

**Primary Objective Results**

The p value and the t value indicate that there is a linear association between prostate-specific antigen level and the four selected variables (cancer volume, weight, seminal and score. R2 value is used to see choose the best predicator value. The predictor value with a higher R2 value is considered the vest and the lowest R2 value is considered the predictor that does not explain much of the variation in prostate-specific antigen level.

The using 95% prediction interval, the predicted values are compared with the observed values and the result shows that the selected variables in the model appropriately estimates prostate-specific antigen.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| observation | Cancerv | weight | seminal | score | Observed PSA level | Predicted  PSA level | Predicted range |
| 1 | 0.5599 | 15.959 | 0 | 6 | 2.961 | 2.968 | -7.04≤ Y ≤12.98 |
| 2 | 0.3716 | 27.660 | 0 | 7 | 6.745 | 6.760 | -3.16≤ Y ≤16.68 |
| 3 | 0.6005 | 14.732 | 0 | 7 | 4.702 | 4.708 | -5.36≤ Y ≤14.78 |
| 4 | 0.3012 | 26.576 | 0 | 6 | 4.574 | 4.588 | -5.317≤ Y ≤14.49 |

**Tables 3: validating model prediction**

**Discussion and conclusions**

The estimated regression equation is written as:

Yi = -11.835 + 0.905X1 +0.174X2 + 3.40X3 +1.918X4

Where:

Yi is the Prostate-specific antigen level

X1 is cancer volume

X2 is weight

X3 is seminal

X4 is score

Ei is the error term; Ei ∼iidN(0,σ2)

i =1,2, 3,…,97

The σ2 = MSE = 23.94

The selected predictor variables effectively predict prostate-specific antigen level. All the predictor variables cancer volume, weight, seminal and score have a positive linear association with the response variable prostate-specific antigen level. Each of the predictor variables are analyzed separately and tested for their linear association with the response variable. For this study, (α)=0.05 significance level and a 95% confidence interval are used to analyze and find the linear association between the response and predictor variables. Table 4 shows the min, median, and maz for the data. It also shows important findings such as coefficients, R2, R2a, t value, p value, MSE, and F-statistic. Also table 5 shows the ANOVA table for the model and which includes SSR, MSE, MSR and F values. All selected predictable variables are linearly associated with prostate-specific antigen level. The predictable variable cancer volume explains 42% of the unexplaned variation in the prostate-specific antigen level. 95% prediction interval also was used to compare predicted model with observed variables and showed that the the model can succesfuly estimate the the prostate-specific antigen level. Finally,the study suggest that there is a linear association between the response variable and the predictor variables.

|  |
| --- |
| Table  Description automatically generated |

**Table 4: statistics table for the selected regression model**

**A picture containing table

Description automatically generated**

**Table 5: ANOVA table for the regression model**

**Appendix: R-code**

**Graphical user interface, text, application, chat or text message

Description automatically generated**

**Graphical user interface, text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

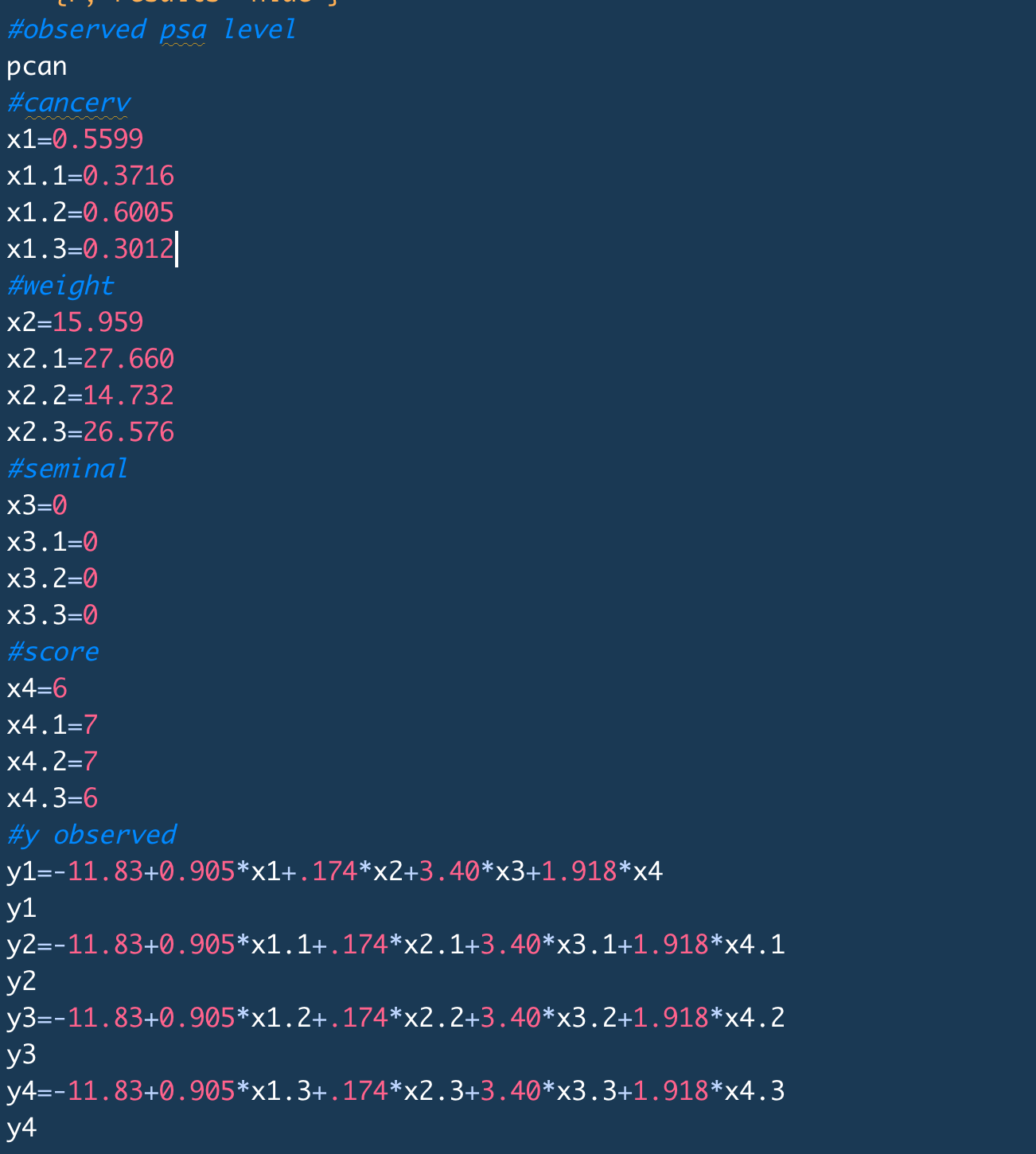
Description automatically generated**

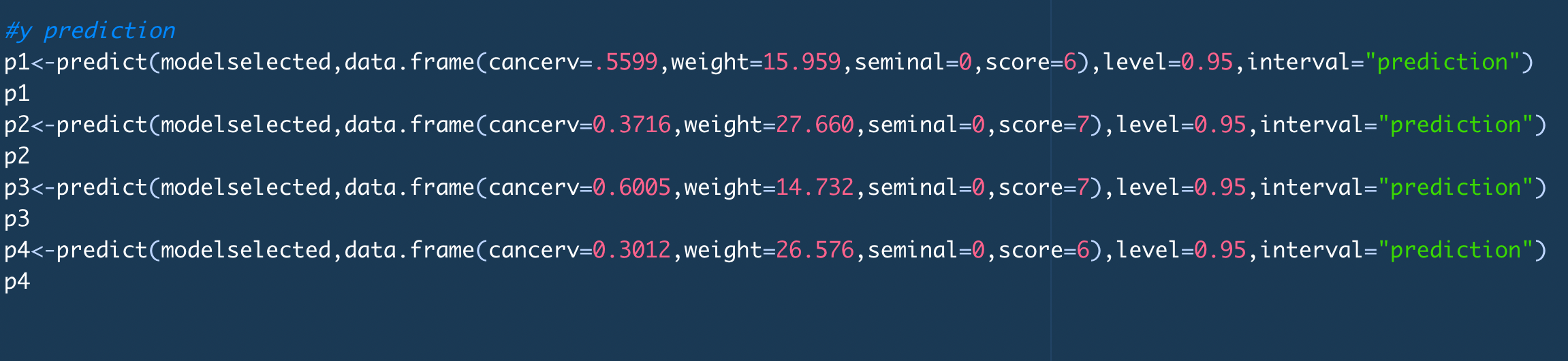
**Text

Description automatically generated**

**Text

Description automatically generated**

****

****

**References**

# [1] Key Statistics for Prostate Cancer. Available online

# <https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html>

# [2] Risk factors for prostate cancer: Available online

# <https://www.mayoclinic.org/diseases-conditions/prostate-cancer/symptoms-causes/syc-20353087>

# [3] Prostate cancer screening. Available online

# <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6283370/>