STA 206: Statistical Methods for Research I
University of California, Davis
Fall 2022

Multiple Regression Analysis of Prostate-Specific Antigens

Yemisi Obasemo and Monica Orme

December 5, 2022

Abstract

We studied 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 prostatectomies (Stameyet al., 1989). The objectives of this project were to investigate which factors significantly affect PSA levels, which regression model best explains the variability in PSA levels, and determine the reliability of the final model's results through data analysis. The project explored some exploratory data analysis to check for the relationship between the variables. Then went further to check for the model that best describes the relationship between the variables. There were a number of limitations in the process of deciding what model best describes the association between the variables like, the training and validation datasets not being large enough (both are less than 50) among others. The procedure showed that the model that best explained variability in the log-transformed PSA levels was the one that included seminal vesicle invasion, square-root transformed cancer volume, benign prostatic hyperplasia, and gleason score. Most of the model coefficients were significant in their effect on the model's response variable and the model's R_a^2 (adjusted R-squared) metric was 0.5691, which means that the 56.91% of the variation in the log-transformed response can be explained by the model.

Introduction

A university medical center urology group was interested in determining 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 prostatectomies (Stamey et al., 1989). The prognostic clinical measurements in men that comprised the dataset include 7 continuous variables, which are PSA levels, prostate cancer volume levels, weight of the prostate, age of the patients, the amount of benign prostatic hyperplasia, and the degree of capsular penetration. The dataset also includes 2 categorical variables, which are presence or absence of seminal vesicle invasion and the gleason score. The descriptions for each of these variables is contained in Figure 1.1. Hence, we performed data analysis by building a multiple regression in this project to determine if there is an association between PSA levels and these clinical measurements.

In this project, we investigated which factors significantly affected PSA levels, which regression model best explains the variability in PSA levels, and the reliability of the final model's results through data analysis. Particularly, these objectives are important to investigate since the solutions to these objectives can help scientific researchers determine what factors significantly impact PSA levels (National Cancer Institute 2022). This can ultimately help such researchers determine factors that may also influence the risk of prostate cancer.

Methods and Results

The data analysis process for investigating the questions of interest include conducting exploratory data analysis (EDA), performing preliminary model fitting on the entire dataset, conducting model selection and validation, and performing statistical inference on the final model.

During the EDA phase, one of the first steps was to check the data types of the variables in the dataset. In particular, the continuous variables in the dataset were PSA levels, cancer

volume, weight, age, benign prostatic hyperplasia (BPH), and capsular penetration whereas the categorical variables in the dataset comprised seminal vesicle invasion and gleason score. We updated the data types of the continuous variables and categorical variables to be numeric and factor respectively in R. The distributions of the variables were investigated in their raw form through boxplots and histograms (Figures 2.1 - 2.3). We found that the distributions of the PSA levels, cancer volume, weight, and capsular penetration were right skewed. The distribution of age was roughly symmetric. The distribution of benign prostatic hyperplasia was bimodal with the highest mode being 0. To address the issue of this variable being zero-inflated, the categorical variable BPH presence (entitled bph_presence in the dataset) was added to account for the presence and absence of BPH; a BPH amount above 0 corresponded to presence of BPH whereas a BPH amount equal to 0 corresponded to absence. Furthermore, many individuals in this study had an absence of seminal vesicle invasion by 78%, 34% of individuals had a gleason score of 6, 44% had a gleason score of 7, and 22% had a gleason score of 8. Most individuals had a presence of BPH by 56% whereas the remaining 44% did not. Boxplots were also plotted for each of the continuous variables on the basis of each of the categorical variables to determine if there were differences across groups. There were noticeable differences in the distributions of seminal vesicle invasion on the basis of cancer volume and seminal vesicle invasion (Figures 2.4 - 2.6).

A scatterplot matrix of the continuous variables was made to investigate the relationships between PSA levels and the remaining continuous variables as well as the relationship that the remaining continuous variables had with each other (Figure 2.7). We found that there was no noticeable linear relationship between PSA levels and the remaining continuous variables and their correlations were positive. PSA-levels were log-transformed since its histogram was symmetric (Figure 2.8) and the Box-Cox procedure suggested this transformation with the log-likelihood being maximized at around $\lambda = 0$ (Figure 2.9). Another scatterplot matrix was made of the log-transformed PSA levels and the remaining continuous variables were made to verify linearity and most of the other continuous variables had a linear relationship with the transformed PSA levels. However, cancer volume and the log-transformed PSA levels did not appear to have a linear relationship, so a square root transformation of cancer volume was applied since the data points in the plot between these variables were increasing and concaved upward (Figure 2.11). After this transformation, most of the variables generally had a linear relationship with the transformed PSA levels, which is necessary in order for the linearity assumption of multiple linear regression to hold. For the rest of this study, the log-transformed PSA levels replaced PSA levels and the square root transformed cancer variable replaced cancer volume.

After investigating the variables' distributions and linearity with the transformed PSA levels, a preliminary regression model was fitted onto the transformed PSA levels regressing onto the remaining variables. We found that there was no clear pattern between the residuals and the fitted values and the QQ plot of the residuals was left-tailed (Figure 2.12), which indicated the constant variance assumption was satisfied and the normality assumption was not satisfied. Hence, the cook's distance measurements (D_i values) were calculated to determine if there were any extreme outliers affecting the model and we deemed D_i values > 1 to be extreme. The 32nd observation in this study had a D_i value of 5.36. This observation was investigated through the cook's distance plot, a summary of the percentage change between the fitted values of the full model and the predicted values of model without the 32nd observation applied to the entire dataset, and a plot of the fitted values without using the 32nd case against the fitted values using

all cases. We found that the 32nd observation had D_i that was much higher than all of the other ones in this study, the range of the percentage change was between 0.18% and 194.6%, and not all of the values in the plot fell on a straight line (Figures 2.14 - 2.15). Hence, since the 32nd observation was very extreme relative to other observations and added variability in the fitted values, it was removed from this study. At this point, the data is ready for further analysis and model selection.

We proceeded to divide the data into two parts: training data for model building and validation data for model validation. We had 96 dataset left so it was easier to divide it into two parts. Each part had 48 dataset. We drew the boxplot to check the distribution of the variables in the validation and training data. The variables (PSA_level, sqrt_cancer_volume, weight,age, benign_prostatic_hyperplasia, capsular_penetration) had very similar distributions (Figure_).

We fit the entire dataset with the training data and went ahead to get the best subset regression. We considered 8 X variables: cancer_volume, weight, age, benign_prostatic_hyperplasia (BPH), seminal_vesicle_invasion, capsular_penetration, gleason_score, bph_presence. We got the one best model for all model sizes and named it model 1 and we used the function regsubsets in the leaps library to get a summary for the one best model.

We went ahead to calculate the model *AIC* and *BIC*. We fit the none-model with no X variable since the regsubsets function does not fit it.We can see that the SSE was decreasing as the r_squared and r-squared-adjusted was increasing.

We proceeded to use the forward or backward stepwise to select the final best model . The final best model based on the forward or backward stepwise is "log_PSA_level ~ sqrt_cancer_volume + seminal_vesicle_invasion + benign_prostatic_hyperplasia + gleason_score". This model was also chosen based on the AIC criterion.

We used the training data to rerun the first model (best model) and found out that the coefficients of the chosen X variables are significant.We did model diagnostic for model one and from the QQ-plot and fitted versus Residual plot, we can see a linear relationship between the log PSA level and the chosen X variables and we can say that model 1 is a good model.

We obtained the $Press_p$ value (22.08743) for model 1. The value seems to be reasonably small. Hence, we can conclude the model does not overfit the data.

We used the validation data and training data to rerun model 1. We obtained the estimated regression and standard errors of model 1 built on both the training and validation datasets. Most of the estimated coefficients as well as their standard errors agree quite closely on the two data sets (Figure 2.22). We also examine the SSE and adjusted R squares using both the training data and validation data.The SSE values are quite far, but the adjusted R squares are very close. We found the *SSE*/*n* under the training model, and compare it to the *MSPEv* when we apply our training model to the validation data. The MSPE is not much larger than the SSE divided by n, so it doesn't overfit the data as much.

We went ahead to choose another model with log_PSA_level regressing onto sqrt_cancer_volume, seminal_vesicle_invasion, benign_prostatic_hyperplasia, capsular_penetration, and gleason_score based on the r_adjusted criterion and named it model 2. We repeated the procedures we took above for model 1 and applied it to model 2. Then, compare the two models.

We used the training data to rerun the second model and found out that the models' coefficients except the ones corresponding to capsular_penetration and gleason_score7 were

significant. We did model diagnostics for model 2 and from the QQ-plot and fitted versus residuals plot, we can see an almost linear relationship between the log PSA level.

We obtained the $Press_p$ value (27.32679) for model 2. The value seems to be reasonably small. Hence, we can conclude the model does not overfit the data.

We used the validation data and training data to rerun model 2. We obtained the estimated regression and standard errors of model 2 built on both the training and validation datasets (Figure 2.26). Most of the estimated coefficients as well as their standard errors agree quite closely on the two data sets. We also examined the SSE and adjusted R squares using both the training data and validation data. The SSE values are quite far, but the adjusted R squares are very close. We found the *SSE*/*n* under the training model, and compared it to the *MSPEv* when we apply our training model to the validation data. The MSPE is not much larger than the SSE divided by n, so it doesn't overfit the data as much.

We compared Model 1 and Model 2 and found that the variable "capsular_penetration" was not present in model 1 but present in model 2. However, the coefficient corresponding to this variable was not significant in model 2, which makes it a justifiable reason for not being present in the "best model" (Model 1). The $Press_p$ for Model 1 (22.08743) appears better than the $Press_p$ for model 2 (27.32679), since it's smaller, this implies that model 1 fits the data better. Most of the estimated coefficients as well as their standard errors for the training and validation data are closer to each other in model 1 compared to model 2. The SSE values and adjusted R squares are farther from each other in the training and validation data in model 2 compared to model 1, which implies that there's more variability between the training and validation data in model 2. The difference between the MSPEv and SSE/n in model 2 is greater than that of model 1 and this implies that model 1 fits the data better than model 2 (Figure 2.29). Based on the comparison above, our final preferred model is Model 1, which is the final model chosen by forward or backward stepwise selection.

The final model that has been deemed to best address our questions of interest is the log-transformed PSA levels regressing onto seminal vesicle invasion, benign prostatic hyperplasia, gleason score, and square-root transformed cancer volume. Each of the slopes in the model are interpreted as the change in the average log-transformed PSA levels by the coefficient value with a unit increase in the corresponding predictor when all other predictors are held constant. The intercept is interpreted as the change in the average log-transformed PSA levels when all other predictors are equal to 0. This model is denoted as follows:
$\log(\text{PSA\_levels}) = 1.211 + 0.092*X_{\text{benign\_prostatic\_hyperplasia}} + 0.272*X_{\text{sqrt\_cancer\_volume}} + 0.837*X_{\text{seminal\_vesicle\_invasion\_1}} + 0.242*X_{\text{gleason\_score7}} + 0.69*X_{\text{gleason\_score8}}$.

The model assumptions were all reasonably satisfied since the model's standardized residuals fell on a straight line in the QQ plot, there was no distinct pattern in the residuals vs. fitted values plot, and the quantitative predictors all have a linear relationship with the response (Figure 2.17). This allowed us to conduct statistical inference on the final model. A significance level of $\alpha = 0.05$ has been used and a decision rule of a p-value less than α has been applied to the statistical tests in this analysis. In the F-test for regression relation, the p-value was $7.866 * 10^{-8}$, which means that there is sufficient evidence to conclude that not all of the coefficients are 0 in the model and there's a significant regression relation. From there, we conducted individual t-tests on each of the coefficients to determine whether they were significantly different from 0. The p-values for each of the coefficients except the one for gleason_score7 were less than α, which means that we have sufficient evidence to conclude that the each of the coefficients except the one for gleason_score7 have a significant effect on then log-transformed PSA levels (Figure

2.16). 95% confidence intervals were constructed to determine the reliability of the results for the estimated coefficients and each of the intervals mean that we're confident by 95% that the true value of the model's coefficients are contained in the intervals (Figure 2.17). Most of the confidence intervals contained positive values except for the coefficient corresponding to gleason_score7, which ultimately means that the coefficient estimate for gleason_score7 is unreliable for prediction and interpretation.

The model's goodness of fit was also assessed. The MSE was roughly 0.4136, which means that the average squared difference between the actual log-transformed PSA levels and the fitted values was low. The model's R_a^2 (adjusted R-squared) metric was 0.5691, which means that the 56.91% of the variation in the log-transformed response can be explained by the model. Overall, the model moderately explains the variation in the transformed response and has low variability, which makes it generally reliable for prediction and interpretation.

Conclusion and Discussion

Based on the data analysis conducted, the model that best explained variability in the log-transformed PSA levels was the one that included seminal vesicle invasion, square-root transformed cancer volume, benign prostatic hyperplasia, and gleason score. The factors that significantly affected the log-transformed PSA levels were seminal vesicle invasion, square-root transformed cancer volume, benign prostatic hyperplasia, and gleason score (with only two levels based on a gleason score of 8) based on the t-tests on the model's coefficients. Since most of the coefficients were significant, the MSE was low, and there was a significant regression relation, the model is generally reliable for predicting the PSA levels of a typical or average man. However, the coefficient corresponding to a gleason score of 7 was insignificant and its confidence interval was wide, which means that the predictor corresponding to this coefficient increases the sampling variability in the model and isn't reliable for interpreting the change in the response. It's important to note that this model shouldn't be used to predict the PSA levels of a man with extreme levels of cancer volume and/or benign hyperplasia since extreme observations weren't contained in building this model.

The possible limitations were the training and validation datasets aren't the largest since both are less than 50 and not all of the criteria were used for selecting the potential best models as we primarily used the stepwise AIC and the adjusted R^2 as criterions for selecting these models. Essentially, it's possible that another good model was overlooked as we didn't use all of the criteria to select the best model. Benign Prostatic Hyperplasia (BPH) and Seminal Vesicle Invasion had several 0 values, which may have skewed the results of the final model. We attempted to address the issue with BPH by adding a categorical variable to account for the presence and absence of BPH. However, we could've also added a categorical variable to account for the presence and absence of seminal vesicle invasion and determine if this predictor would've been significant to include in the model. A few transformations were used to achieve a better fit for the multiple regression model, which may make some of the results slightly more difficult to interpret. The adjusted R^2 for the final model was moderate at 56.91%. Although more than half of the variance in the response is explained by the model, this goodness of fit isn't strong. Hence, it may have been worth checking to see if other regression models such as ridge regression and if adding interaction terms could have provided a better fit.

Appendix 1: Figures and tables

Variable Name	Data Type	Description
PSA_level	Quantitative	PSA level - Serum prostate-specific antigen level (mg/ml)
log_PSA_level	Quantitative	The logarithm transformed serum prostate-specific antigen level (mg/ml). This was used as the response in the final model.
cancer_volume	Quantitative	Prostate cancer volume: Estimate of prostate cancer volume (cc)
sqrt_cancer_volume	Quantitative	The square root transformed estimate of prostate cancer volume (cc). This variable replaced cancer_volume in the final model.
weight	Quantitative	Weight of Prostate: Prostate weight (gm)
age	Quantitative	Age: Age of patient (years)
benign_prostatic_hyperplasia	Quantitative	Benign prostatic hyperplasia: Amount of benign prostatic hyperplasia (cm^2)
bph_presence	Categorical	Benign prostatic hyperplasia status: Presence or absence of benign prostatic hyperplasia: 1 if presence; 0 if absence. This variable was added during data analysis to account for the many zeros in the benign_prostatic_hyperplasia variable.
seminal_vesicle_invasion	Categorical	Seminal vesicle invasion: Presence or absence of seminal vesicle invasion: 1 if presence; 0 if absence
capsular_penetration	Quantitative	Capsular penetration: Degree of capsular penetration (cm)
gleason_score	Categorical	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)

Figure 1.1: Description of the Variables from the Prostate Dataset

Appendix 2: R codes and outputs

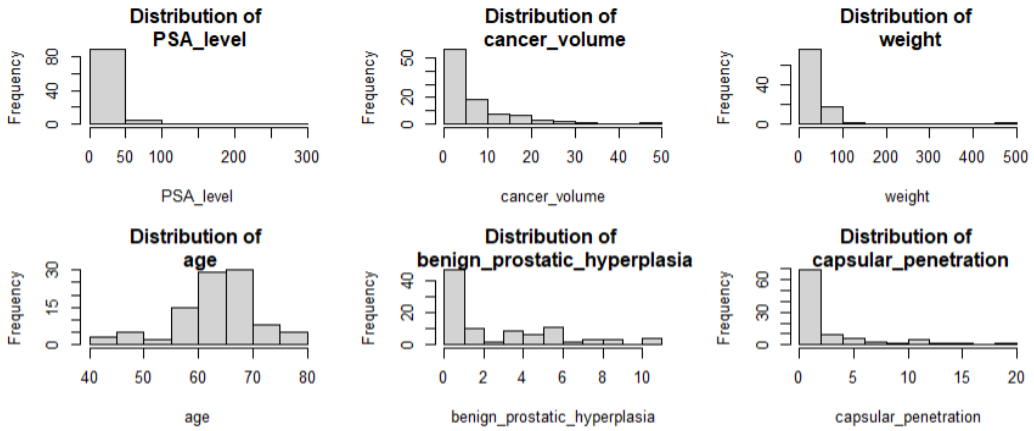


Figure 2.1: Histograms of the Initial Continuous Variables from the Prostate Dataset

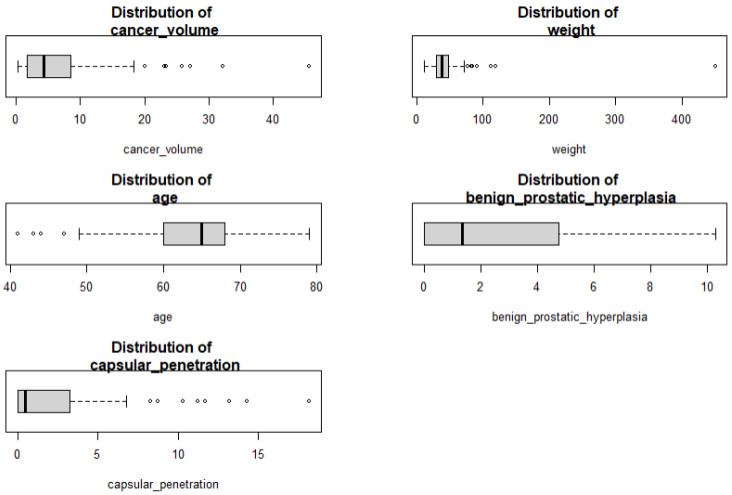


Figure 2.2: Boxplots of the Initial Continuous Variables from the Prostate Dataset

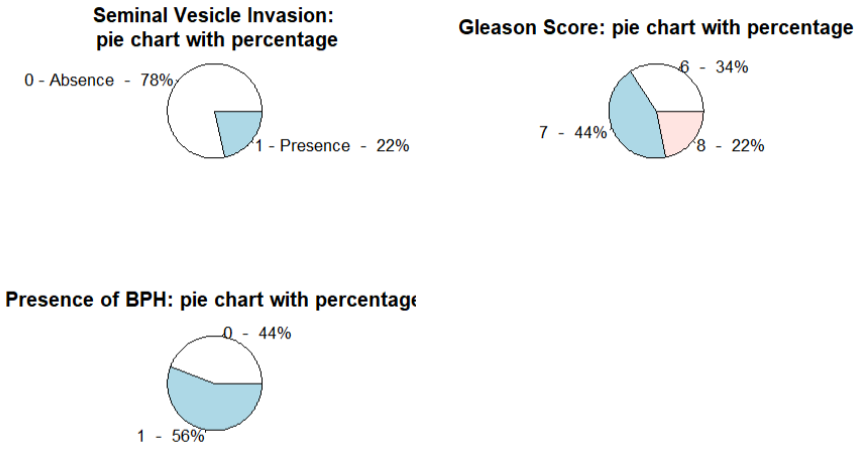


Figure 2.3: Pie charts of the Initial Categorical Variables from the Prostate Dataset

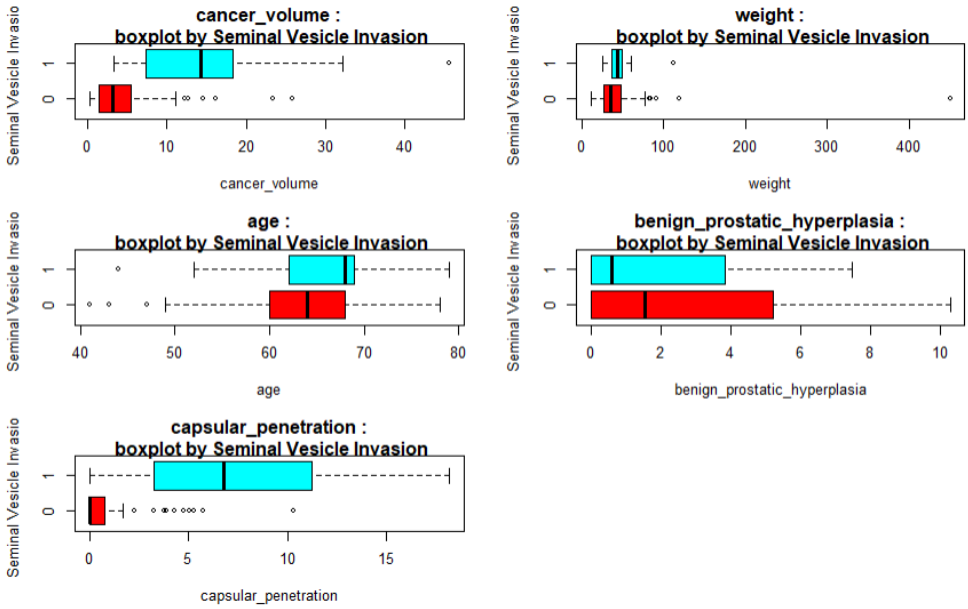


Figure 2.4: Boxplots of the Continuous Variables based on the levels of seminal_vesicle_invasion from the Prostate Dataset

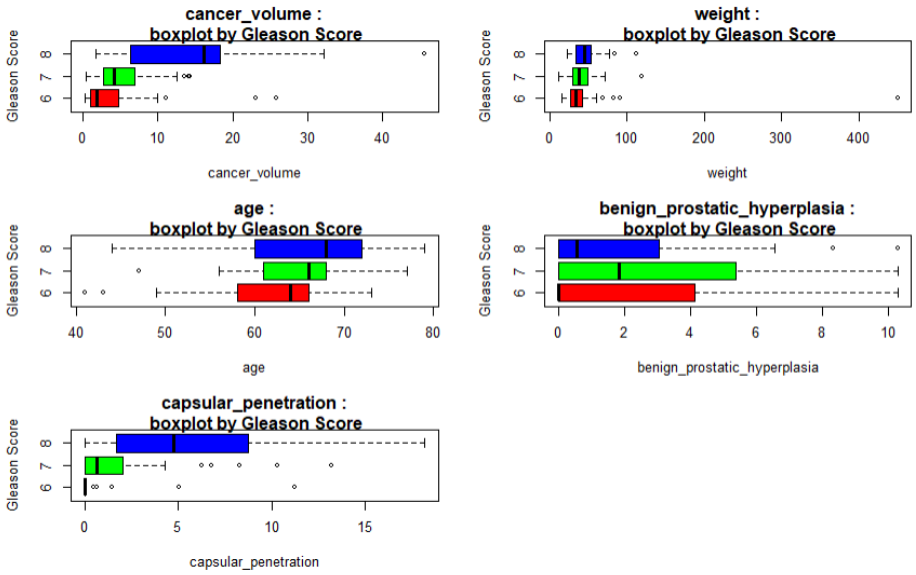


Figure 2.5: Boxplots of the Continuous Variables based on the levels of gleason_score

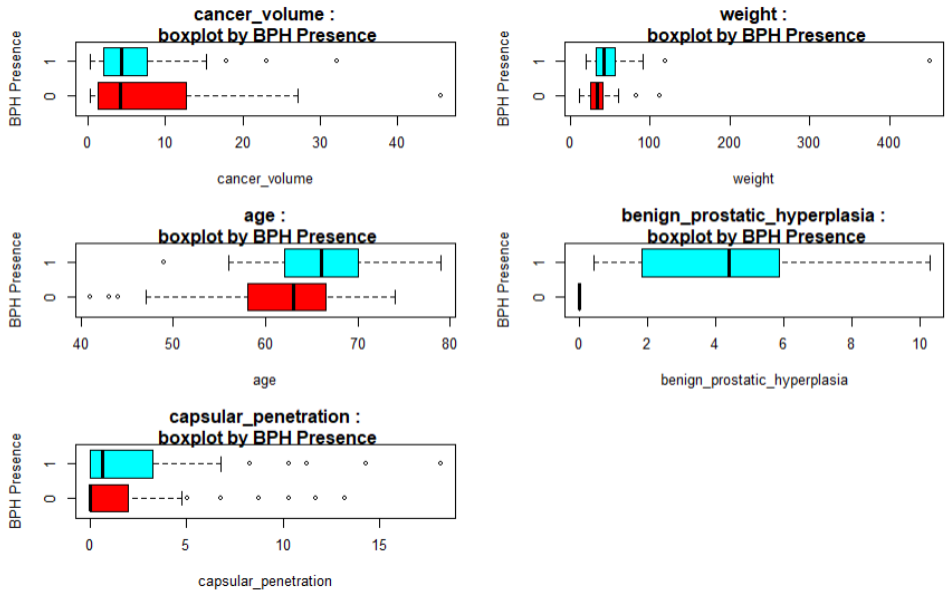


Figure 2.6: Boxplots of the Continuous Variables based on the levels of bph_presnce from the Prostate Dataset

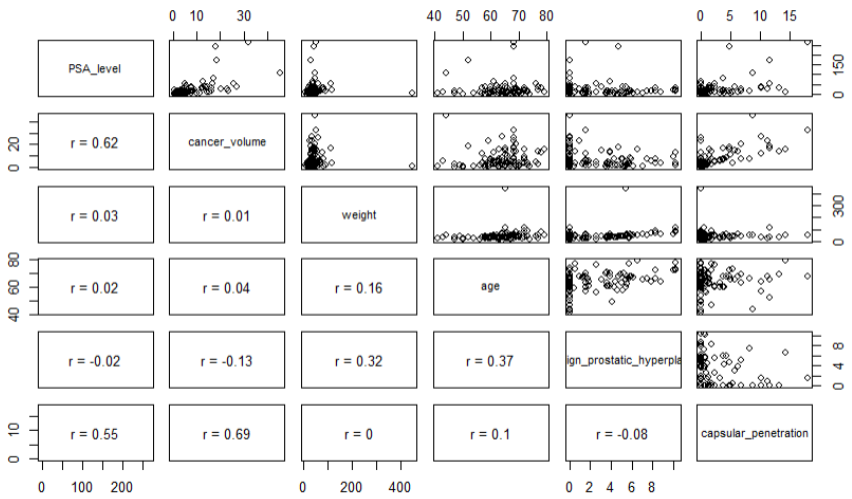


Figure 2.7: Preliminary Scatterplot Matrix of the Continuous Variables

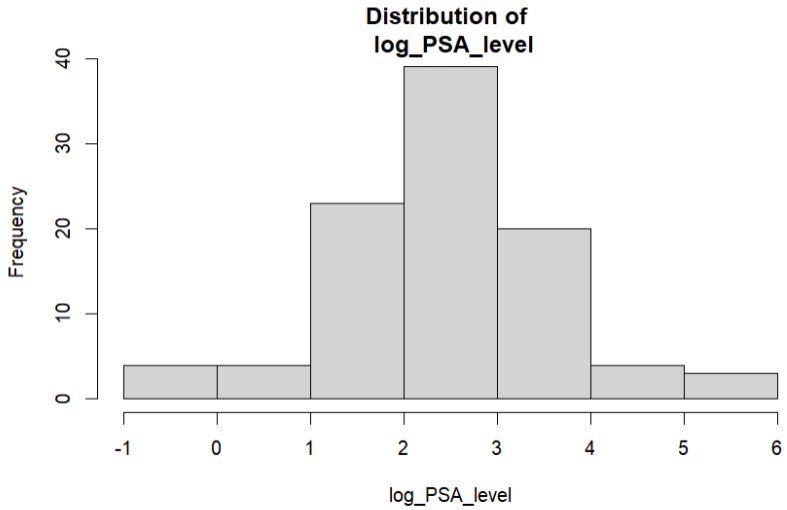


Figure 2.8: Histogram of the logarithmic transformation of PSA_values

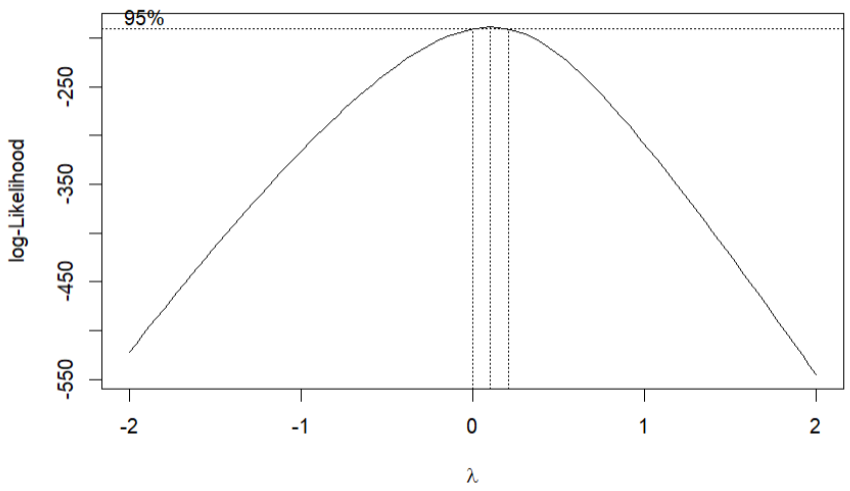


Figure 2.9: Box-Cox Procedure Plot of PSA_values regressing onto the remaining variables

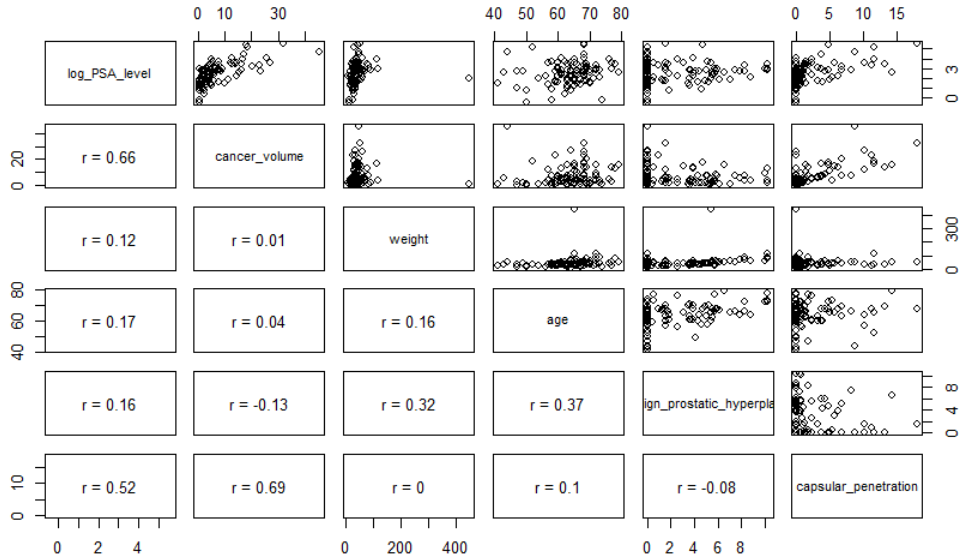


Figure 2.10: Preliminary Scatterplot Matrix of the Continuous Variables except for PSA_levels and sqrt_cancer_volume

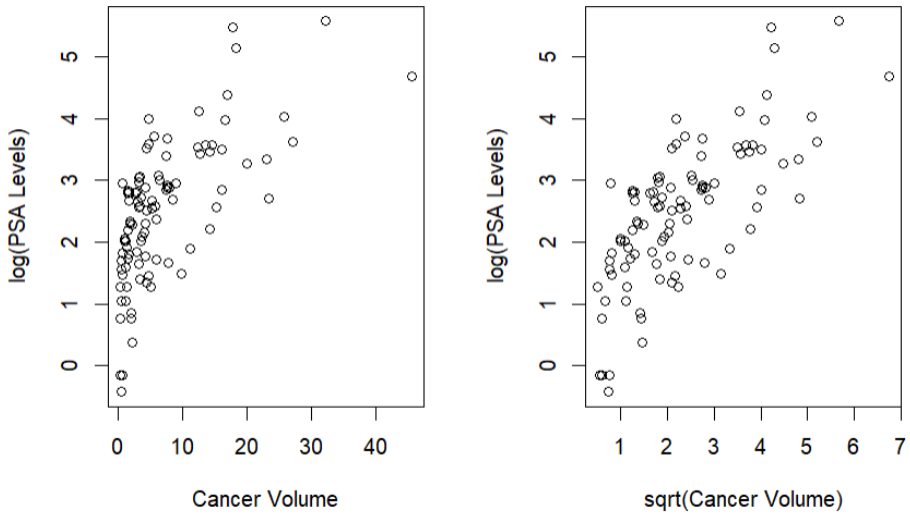


Figure 2.11: Comparison of log_PSA_levels vs. cancer_volume and log_PSA_levels vs. sqrt_cancer_volume

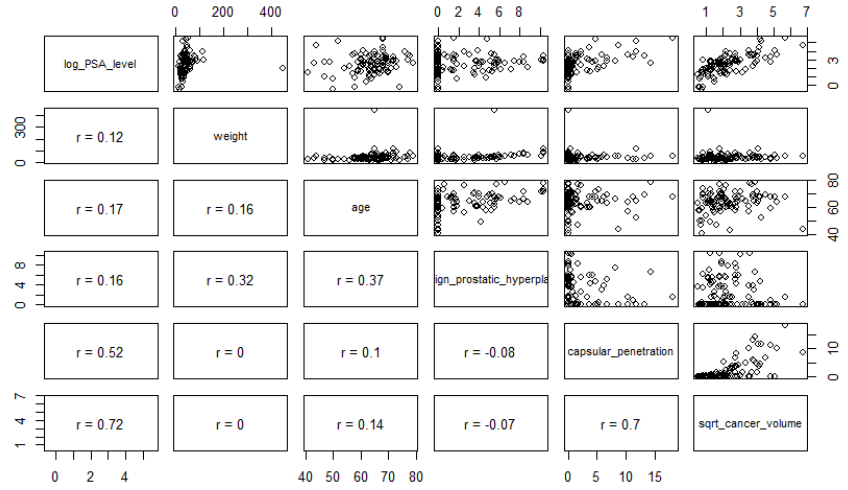


Figure 2.10: Preliminary Scatterplot Matrix of the Continuous Variables except for PSA_levels and cancer_volume

```
lm(formula = log_PSA_level ~ I(sqrt(cancer_volume)) + weight +
    age + benign_prostatic_hyperplasia + bph_presence + seminal_vesicle_invasion +
    capsular_penetration + gleason_score, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.60907 -0.39191  0.02564  0.46287  1.84616

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.336175   0.681033   1.962  0.05296 .
I(sqrt(cancer_volume))      0.509993   0.087967   5.798 1.06e-07 ***
weight                      0.001873   0.001742   1.075  0.28539
age                        -0.011357   0.011131  -1.020  0.31042
benign_prostatic_hyperplasia  0.034346   0.039819   0.863  0.39076
bph_presence1               0.394717   0.230603   1.712  0.09052 .
seminal_vesicle_invasion1   0.731401   0.254611   2.873  0.00511 **
capsular_penetration       -0.046871   0.031321  -1.496  0.13814
gleason_score7              0.229795   0.177168   1.297  0.19804
gleason_score8              0.664539   0.250809   2.650  0.00957 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7244 on 87 degrees of freedom
Multiple R-squared:  0.6427,    Adjusted R-squared:  0.6058
F-statistic: 17.39 on 9 and 87 DF,  p-value: 4.449e-16
```

Figure 2.12: Summary of the preliminary regression model of log_PSA_levels regressing onto the remaining variables except cancer_volume and PSA_levels using all of the data

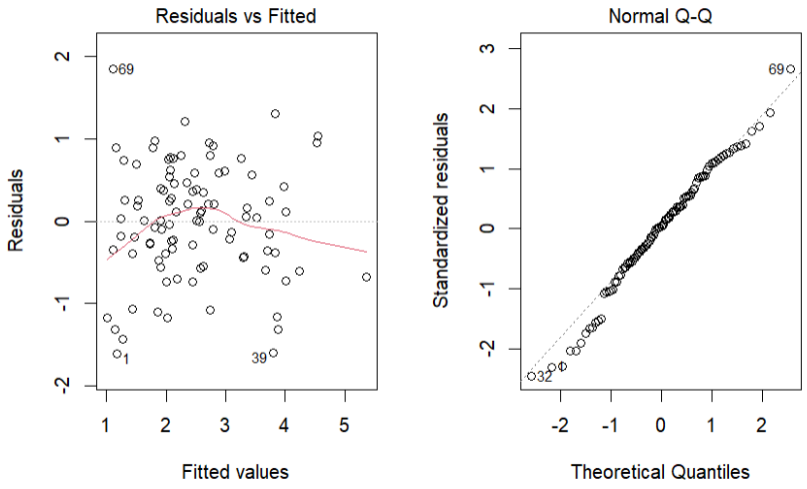


Figure 2.13: Diagnostics the preliminary regression model of log_PSA_levels regressing onto the remaining variables except cancer_volume and PSA_levels using all of the data

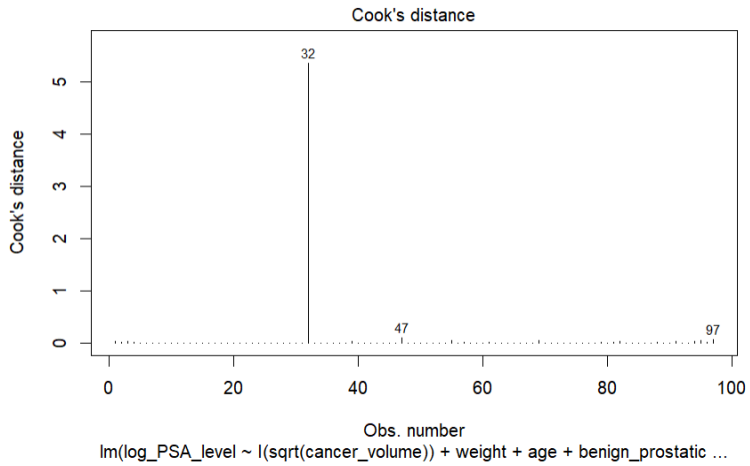


Figure 2.14: Plot of the cook's distance values for each observation in the preliminary model

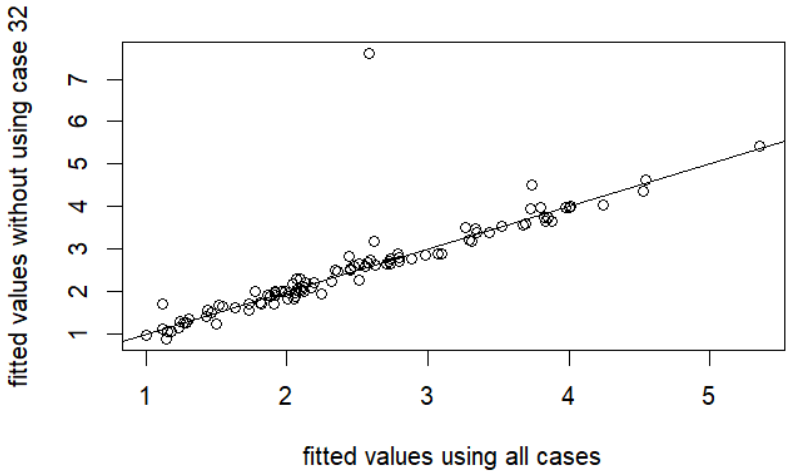


Figure 2.15: Plot of the fitted values without case 32 vs. fitted values using all cases for the preliminary model

```
lm(formula = log_PSA_level ~ sqrt_cancer_volume + seminal_vesicle_invasion +
    benign_prostatic_hyperplasia + gleason_score, data = data.c)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47054 -0.43833 -0.01212  0.43671  1.33867

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.21095    0.28813   4.203 0.000135 ***
sqrt_cancer_volume          0.27208    0.11061   2.460 0.018099 *
seminal_vesicle_invasion1   0.83690    0.27127   3.085 0.003592 **
benign_prostatic_hyperplasia  0.09201    0.03215   2.862 0.006535 **
gleason_score7              0.24185    0.22926   1.055 0.297499
gleason_score8              0.68956    0.28986   2.379 0.021986 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6431 on 42 degrees of freedom
Multiple R-squared:  0.6149,    Adjusted R-squared:  0.5691
F-statistic: 13.41 on 5 and 42 DF,  p-value: 7.866e-08
```

Figure 2.16: Summary of Model 1 (the final model) of log_PSA_levels regressing onto the remaining variables except cancer_volume and PSA_levels using the training data

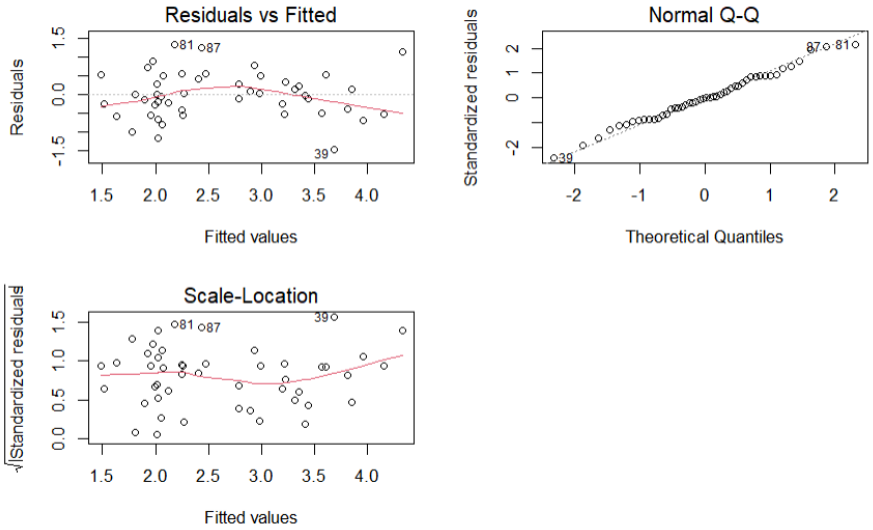


Figure 2.17: Diagnostic Plots of Model 1 (the final model)

```
                                2.5 %    97.5 %
(Intercept)                 0.62947488 1.7924283
sqrt_cancer_volume          0.04886199 0.4953076
seminal_vesicle_invasion1   0.28945524 1.3843411
benign_prostatic_hyperplasia  0.02713558 0.1568898
gleason_score7             -0.22081965 0.7045193
gleason_score8              0.10458909 1.2745209
```

Figure 2.18: 95% Confidence Intervals of the Model 1’s Coefficients

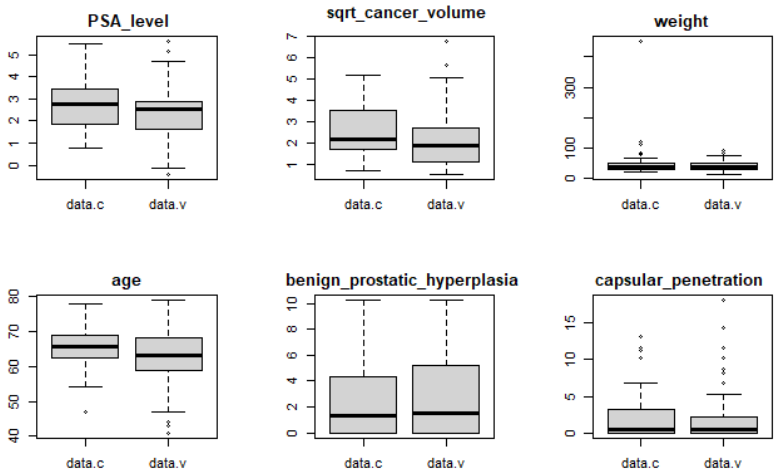


Figure 2.19: Boxplot showing distribution of variables in the validation and training data.

```
Step:  AIC=-36.79
log_PSA_level ~ sqrt_cancer_volume + seminal_vesicle_invasion +
    benign_prostatic_hyperplasia + gleason_score

                              Df Sum of Sq    RSS     AIC
<none>                                      17.371 -36.787
+ capsular_penetration          1    0.5051 16.866 -36.204
+ age                          1    0.1945 17.176 -35.328
+ bph_presence                 1    0.0331 17.338 -34.879
+ weight                       1    0.0034 17.367 -34.797
- gleason_score                2    2.3567 19.728 -34.681
- sqrt_cancer_volume           1    2.5026 19.873 -32.327
- benign_prostatic_hyperplasia 1    3.3882 20.759 -30.234
- seminal_vesicle_invasion     1    3.9366 21.308 -28.983
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
log_PSA_level ~ 1

Final Model:
log_PSA_level ~ sqrt_cancer_volume + seminal_vesicle_invasion +
    benign_prostatic_hyperplasia + gleason_score


                           Step Df  Deviance Resid. Df Resid. Dev         AIC
1                                                  47   45.10856  -0.9821951
2           + sqrt_cancer_volume  1 19.242418        46   25.86614 -25.6767818
3     + seminal_vesicle_invasion  1  2.549863        45   23.31628 -28.6583694
4 + benign_prostatic_hyperplasia  1  3.588630        44   19.72765 -34.6806414
5                 + gleason_score  2  2.356694        42   17.37095 -36.7872780
```

Figure 2.20: Results showing the stepwise procedure for choosing the final model(model 1)

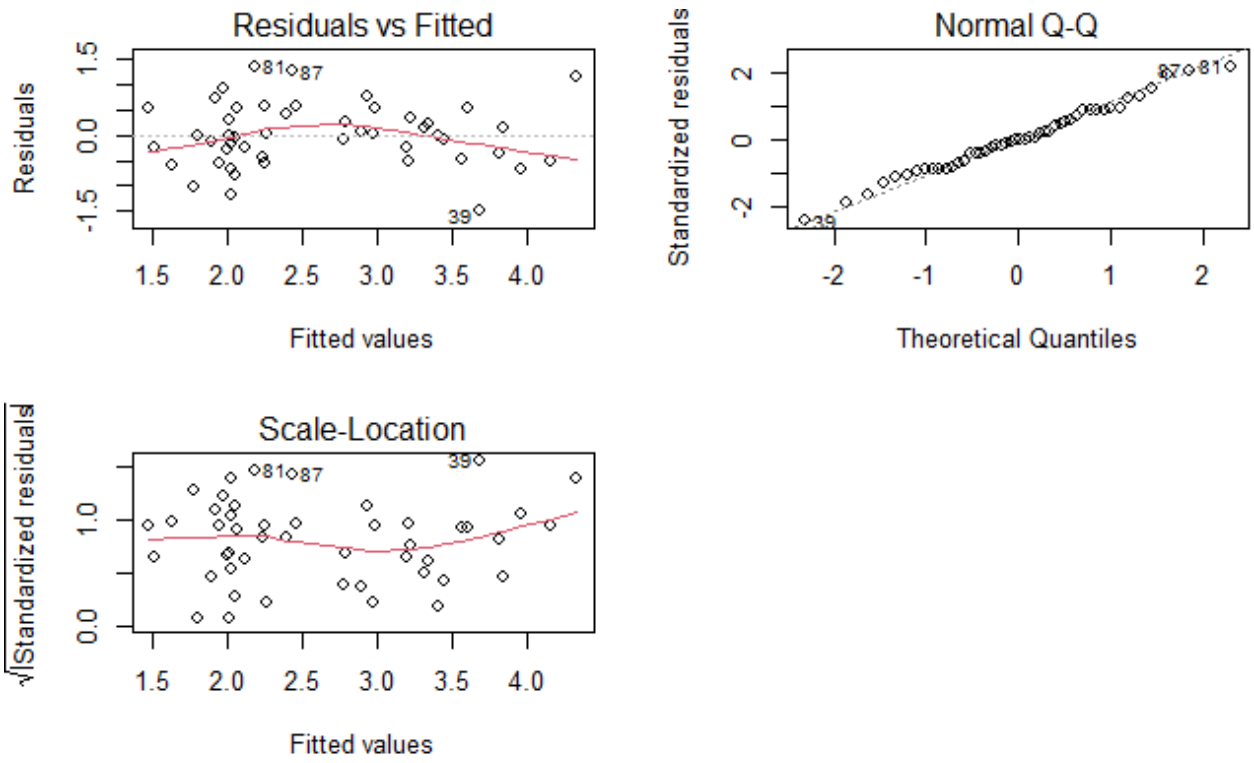


Figure 2.21: Model Diagnostics for final model(model 1) when fitted with the training data.

	Train Est1	Valid Est1	Train s.e1.	Valid s.e1.
(Intercept)	1.21095161	0.72558271	0.2881334	0.27247089
sqrt_cancer_volume	0.27208481	0.55928274	0.1106114	0.11834660
seminal_vesicle_invasion1	0.83689816	0.27523053	0.2712690	0.45596768
benign_prostatic_hyperplasia	0.09201267	0.09061299	0.0321479	0.04098196
gleason_score7	0.24184981	0.06129269	0.2292620	0.27426810
gleason_score8	0.68955502	0.62178729	0.2898624	0.46898977

Figure 2.22: Estimated regression and standard errors of model 1 built on both the training and validation datasets.

	SSE1	R2_adj1
train_sum1	17.37095	0.5690635
valid_sum1	29.13088	0.5956635

Figure 2.23: SSE and adjusted R squares using both the training data and validation data for model 1.

```
MSPE1 <- mean((data.v$log_PSA_level -log_PSA_level.hat1 )^2)
MSPE1

SSE_over_N1 = sse_t1/n
SSE_over_N1
```

```
[1] 0.7004003
[1] 0.3618948
```

Figure 2.24: MSPE and SSE / n for model 1

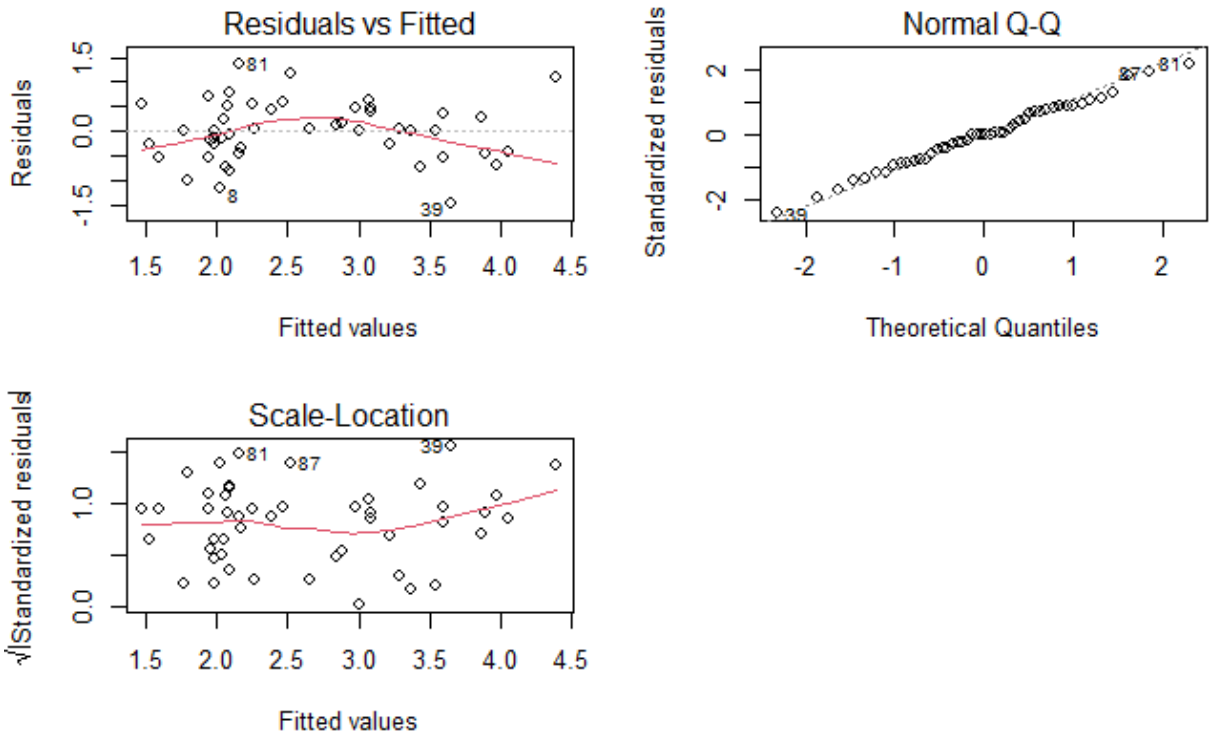
```
Call:
lm(formula = log_PSA_level ~ sqrt_cancer_volume + seminal_vesicle_invasion +
    benign_prostatic_hyperplasia + capsular_penetration + gleason_score,
    data = data.c)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4483 -0.4318  0.0259  0.4421  1.3638

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.13546    0.29532   3.845 0.000413 ***
sqrt_cancer_volume          0.33893    0.12573   2.696 0.010140 *
seminal_vesicle_invasion1   0.97124    0.29646   3.276 0.002146 **
benign_prostatic_hyperplasia 0.08699   0.03238   2.687 0.010377 *
capsular_penetration       -0.04687    0.04230  -1.108 0.274286
gleason_score7              0.21998    0.22949   0.959 0.343401
gleason_score8              0.66489    0.28993   2.293 0.027034 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6414 on 41 degrees of freedom
Multiple R-squared:  0.6261,    Adjusted R-squared:  0.5714
F-statistic: 11.44 on 6 and 41 DF,  p-value: 1.747e-07
```

Figure 2.26: Summary of model 2 (chosen based on R_squared_adjusted)



.Figure 2.25: Model Diagnostics for model 2 when fitted with the training data.

	Train Est2	Valid Est2	Train s.e.2	Valid s.e.2
(Intercept)	1.13545838	0.68231775	0.29532119	0.28624706
sqrt_cancer_volume	0.33892559	0.58966284	0.12572802	0.13199138
seminal_vesicle_invasion1	0.97124078	0.39928821	0.29646044	0.51423252
benign_prostatic_hyperplasia	0.08699184	0.09149602	0.03237966	0.04135697
capsular_penetration	-0.04686808	-0.02830124	0.04229674	0.05253867
gleason_score7	0.21998164	0.08946019	0.22949283	0.28146025
gleason_score8	0.66489101	0.68592116	0.28993499	0.48766607

Figure 2.26: Estimated regression and standard errors of model 2 built on both the training and validation datasets.

	SSE2	R2_adj2
train_sum2	16.86587	0.5713886
valid_sum2	28.93100	0.5888768

Figure 2.27: SSE and adjusted R squares using both the training data and validation data for model 2.

```
MSPE2 <- mean((data.v$log_PSA_level -log_PSA_level.hat2 )^2)
MSPE2

SSE_over_N2 = sse_t2/n
SSE_over_N2
```

```
[1] 0.6911011
[1] 0.3513722
```

Figure 2.28: MSPE and SSE / n for model 2

```
MSPE1-SSE_over_N1
MSPE2-SSE_over_N2
```

```
[1] 0.3385055
[1] 0.3397289
```

Figure 2.29: Difference between MSPE and SSE / n for model 1 and model 2

References

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients, Journal of Urology 16: 1076 - 1083.

Prostate-specific antigen (PSA) test. National Cancer Institute. (n.d.). Retrieved March 11, 2022, from https://www.cancer.gov/types/prostate/psa-fact-sheet