STAT 425 - CASE STUDY REPORT

1. Introduction

In this case study, we looked at a dataset containing information about the cancer volume, age, benign prostatic hyperplasia, seminal vesicle invasion (svi), capsular penetration, and Gleason score of 97 men with prostate cancer. We will analyze the data to find any trends within the dataset and build a linear regression model to predict the Serum prostate-specific antigen (PSA) from a selection of some features stated above. This report includes three sections: data analysis, model building, and model diagnostics. In data analysis, we will plot various density graphs to investigate trends and make predictions for each explanatory variable. In model building, we will explore potential collinearity between each pair of predictors and which variables are insignificant to the model. Finally, we ran diagnostics on the final model to discover whether our model conformed to various assumptions and unveiled observations that may heavily influence the data.

2. Data Analysis

Firstly, we compute a summary of statistics pertaining to columns of the dataset, then plot a density graph for each numerical variable. From the statistical summary table,PSA level has a mean of 23.7 with a standard deviation of 40.8. Furthermore, 75 percentile is 21.3 but 100 percentile is 265, which means that there are some outliers far away from the mean. We should be cautious about these data points because they may yield to big residuals when fitting a regression model. Furthermore, the density graph of PSA level is right-skewed, which explains why there is a significant difference between 75 percentile and 100 percentile in the summary table.

The density age graph is somewhat unimodal with the middle 50 percentile being between 60-68 and the minimum of 40. Therefore, older patients are more likely to experience prostate cancer.

The hyperplasia and capsular density graphs are right skewed. The most striking feature of both graphs is that more than 50 percent of the data points are zero. We notice

that there is a decrease trend in the capsular graph but there is no trend in the hyperplasia graph.

The cancer volume and the capsular graphs are right-skewed like the PSA Level density graph. Furthermore, the correlation coefficients between PSA Level and cancer volume is 0.624 and between PSA Level and capsular is 0.55079. Therefore, it is possible that at least one of these two features significantly contribute to the output of the linear regression model in the next part. We will also investigate the correlation of each pair of the features to remove the collinearity between the explanatory variables.

3. Model Building

The full model shows that every predictor has a nonzero coefficient, but most predictors have statistically insignificant p-values. This could be an indication that these predictors have very little impact on the full model.

To check for collinearity, we compute the condition number and the variance inflation factors (VIFs), as well as the correlation matrix. The condition number outputs as 3.15, which is well below the threshold of 30, indicating a lack of collinearity. As confirmation, the VIFs are less than 5 for all predictors. After taking the square root of the VIFs, we can say that the standard error is less than 2 times larger than it would have been without collinearity for all variables. We can safely say that collinearity is not present in the dataset.

According to the correlation matrix, there are no predictors that are very highly correlated with each other. However, there still are still moderately high correlation values of .69 and .68 between *capsular* & *cancervolume* and *capsular* & *svi*. We can also see from the full linear model that *capsular* has a particularly high p-value of .41. Hence, it is a prime candidate for removal from the model. We can see from the partial F-test anova table that using a significance level of 0.05, *capsular* is indeed insignificant, so it is the first variable to be removed. Yet, from the summary of the resulting reduced model, there are still more insignificant variables. The new most insignificant variable is now *age*. The same process of a partial F-test using the previous model and a new one without *age* is used, again yielding

insignificance. These steps are repeated for *gleason* and *hyperplasia,* and we are left with *cancervolume* and *svi* as the only significant variables.

The final model is : *psalevel* = 1.06 + 2.477(*cancervolume)* + 24.647(*svi)*

4. Model Diagnostics

After choosing the best model, we check for diagnostics. First, we check for high leverage points and there is only one high leverage point, observation #94. To see whether that one point is “bad” or not, we designate the dependent variable and define it with both lower and upper limit to extract from the original data frame and select points outside the range. As a result, it comes out to be also a “bad” high leverage point. Second, we check for outliers and discover that there are 2 observations, #96 and #97. After using the Bonferroni correction, we see that both points have higher than the Bonferroni critical value. Third, we check for influential points using the Cook’s Distance. We can discover that we have one influential point, observation #97.

After checking diagnostics, we also move onto the Assumptions for our model and check that none of the assumptions are violated. First, we check for constant variance. Based on the Breush-Pagan test performed, we discover that the assumption of constant variance is violated, suggesting that there is heteroscedasticity. Then, we check for normality in errors. Based on the Kolmogorov-Smirnov test, we also notice that the normality of errors assumption is violated. Since both of the results have a p-value under 0.05, we use a Box-Cox log transformation to pass both the constancy of variance and normality in error.

5. Conclusion

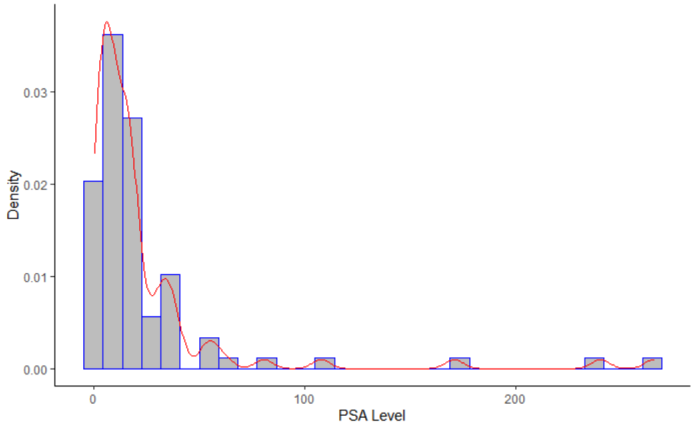
In conclusion, the results of our analysis were that only two columns of our original dataset were meaningful to the final model, seminal vesicle invasion and cancer volume. After transforming the model, we were able to satisfy assumptions that our model was linear.

Appendix

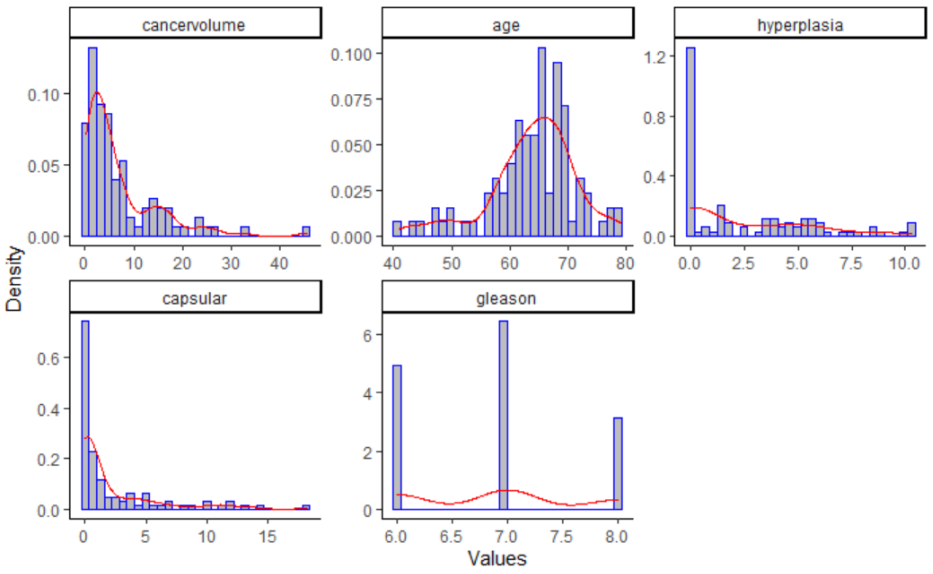
1. Statistical summary of each predictor

```
-- Data Summary ------------------------
                           Values
Name                       psa
Number of rows             97
Number of columns          7
_________________________
Column type frequency:
  numeric                  7
_________________________
Group variables            None

-- Variable type: numeric --------------------------------------------------------------
# A tibble: 7 x 10
  skim_variable n_missing complete_rate   mean     sd     p0   p25    p50   p75  p100
* <chr>             <int>         <dbl>  <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl>
1 psalevel              0             1 23.7   40.8    0.651  5.64 13.3   21.3  265.
2 cancervolume          0             1  7.00   7.88   0.259  1.67  4.26   8.41  45.6
3 age                   0             1 63.9    7.45  41     60    65     68     79
4 hyperplasia           0             1  2.53   3.03   0      0     1.35   4.76  10.3
5 svi                   0             1  0.216  0.414  0      0     0      0      1
6 capsular              0             1  2.25   3.78   0      0     0.449  3.25  18.2
7 gleason               0             1  6.88   0.740  6      6     7      7      8
```

2. Density Graph of PSA Level



3. Density graph of each predictor



4. Correlation between PSA Level and each predictor

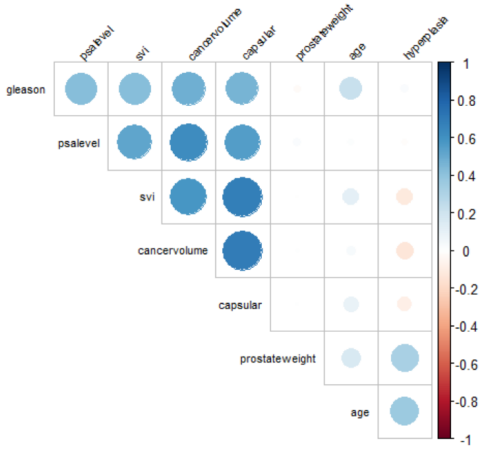
```
  psalevel cancervolume       capsular            svi       gleason prostateweight            age    hyperplasia
1.00000000   0.62415059     0.55079252     0.52861878    0.42957975    0.02621343    0.01719938   -0.01648649
```

5. Full model summary

```
##
## Call:
## lm(formula = psalevel ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.491  -8.199  -0.080   5.923 167.267
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -14.7460    40.1894  -0.367 0.714545
## cancervolume   2.0375     0.5894   3.457 0.000836 ***
## age           -0.5327     0.4724  -1.128 0.262448
## hyperplasia    1.3518     1.1434   1.182 0.240209
## svi           19.6441    10.8303   1.814 0.073038 .
## capsular       1.0974     1.3265   0.827 0.410273
## gleason        6.9942     5.1489   1.358 0.177741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31 on 90 degrees of freedom
## Multiple R-squared:  0.4584, Adjusted R-squared:  0.4223
## F-statistic:  12.7 on 6 and 90 DF,  p-value: 2.481e-10
```

6. Two visualizations of the correlation matrix

	psalevel	cancervolume	age	hyperplasia	svi	capsular	gleason
psalevel	1.00	0.62	0.02	-0.02	0.53	0.55	0.43
cancervolume	0.62	1.00	0.04	-0.13	0.58	0.69	0.48
age	0.02	0.04	1.00	0.37	0.12	0.10	0.23
hyperplasia	-0.02	-0.13	0.37	1.00	-0.12	-0.08	0.03
svi	0.53	0.58	0.12	-0.12	1.00	0.68	0.43
capsular	0.55	0.69	0.10	-0.08	0.68	1.00	0.46
gleason	0.43	0.48	0.23	0.03	0.43	0.46	1.00



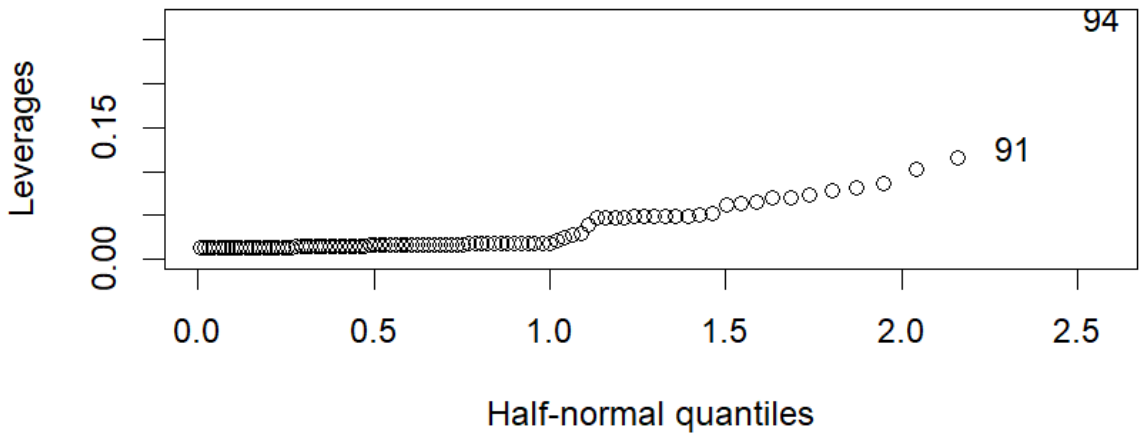
7. Summary of first reduced model

```
Call:
lm(formula = psalevel ~ cancervolume + age + hyperplasia + svi +
    gleason, data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max
-54.839  -8.758   0.206   5.181 163.883

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -18.4353    39.8719  -0.462   0.6449
cancervolume   2.2595     0.5238   4.313 4.07e-05 ***
age           -0.5261     0.4715  -1.116   0.2674
hyperplasia    1.3714     1.1412   1.202   0.2326
svi           23.6477     9.6720   2.445   0.0164 *
gleason        7.4688     5.1080   1.462   0.1471
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.94 on 91 degrees of freedom
Multiple R-squared:  0.4543, Adjusted R-squared:  0.4243
F-statistic: 15.15 on 5 and 91 DF,  p-value: 8.245e-11
```

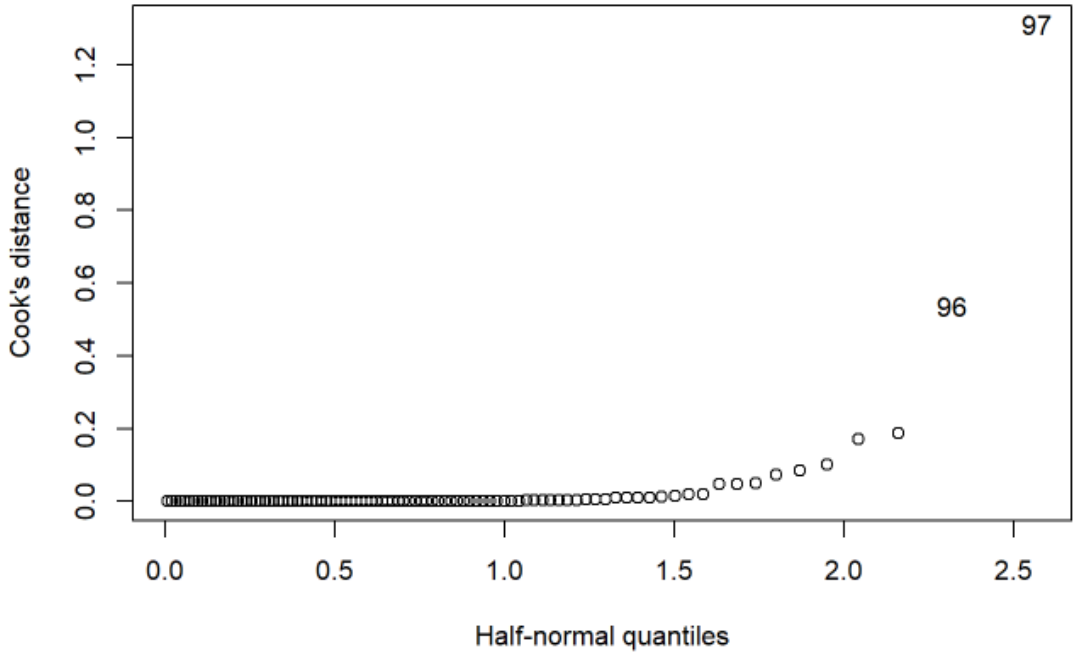
8. Summary of fourth reduced (final) model

```
Call:
lm(formula = psalevel ~ cancervolume + svi, data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max
-55.145  -7.535  -1.129   4.256 170.018

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.060      4.231   0.251   0.8027
cancervolume    2.477      0.495   5.003 2.62e-06 ***
svi            24.647      9.423   2.616   0.0104 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.09 on 94 degrees of freedom
Multiple R-squared:  0.431,  Adjusted R-squared:  0.4189
F-statistic:  35.6 on 2 and 94 DF,  p-value: 3.098e-12
```

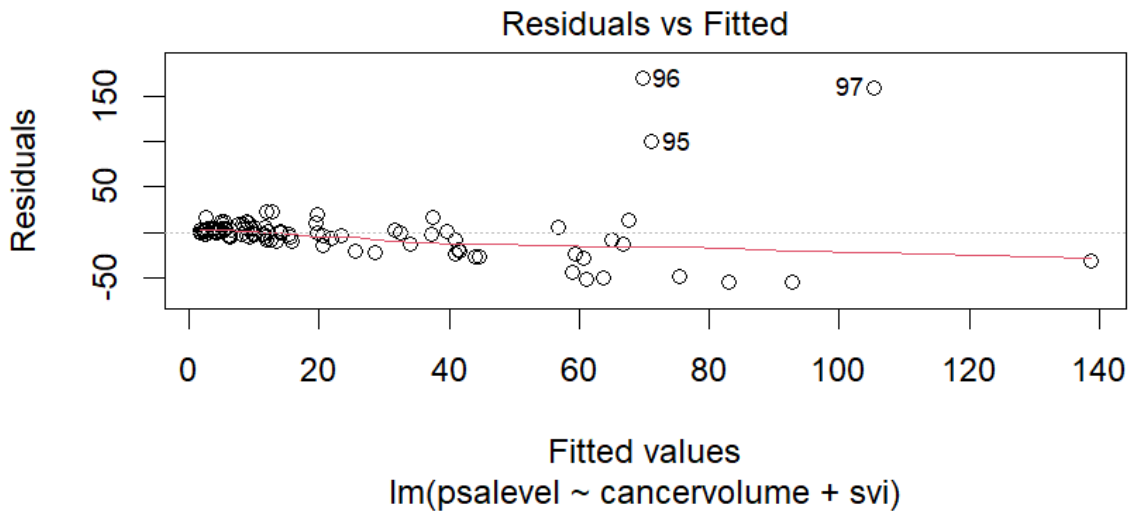
9. High leverage points plot of final model



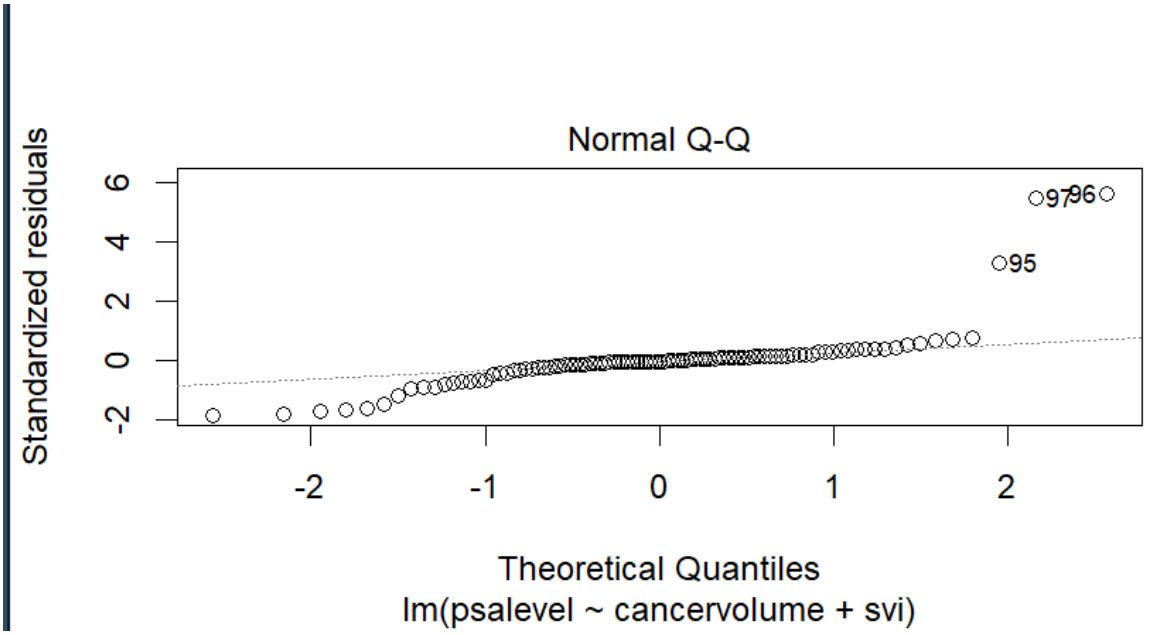
10. Cook’s distance plot of final model



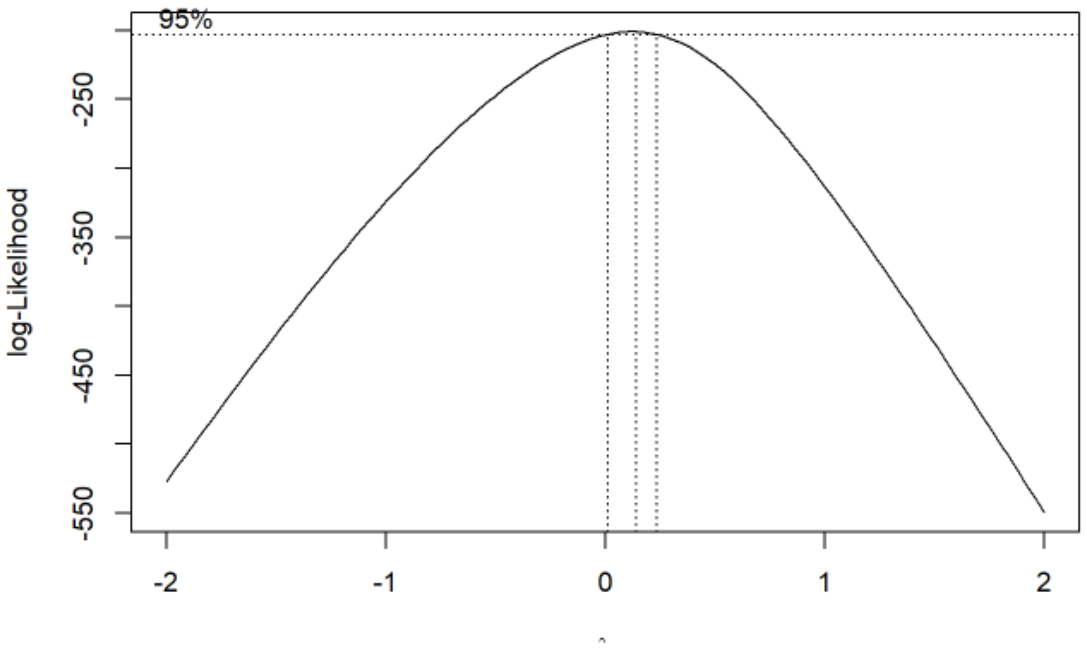
11. residual plot of final model



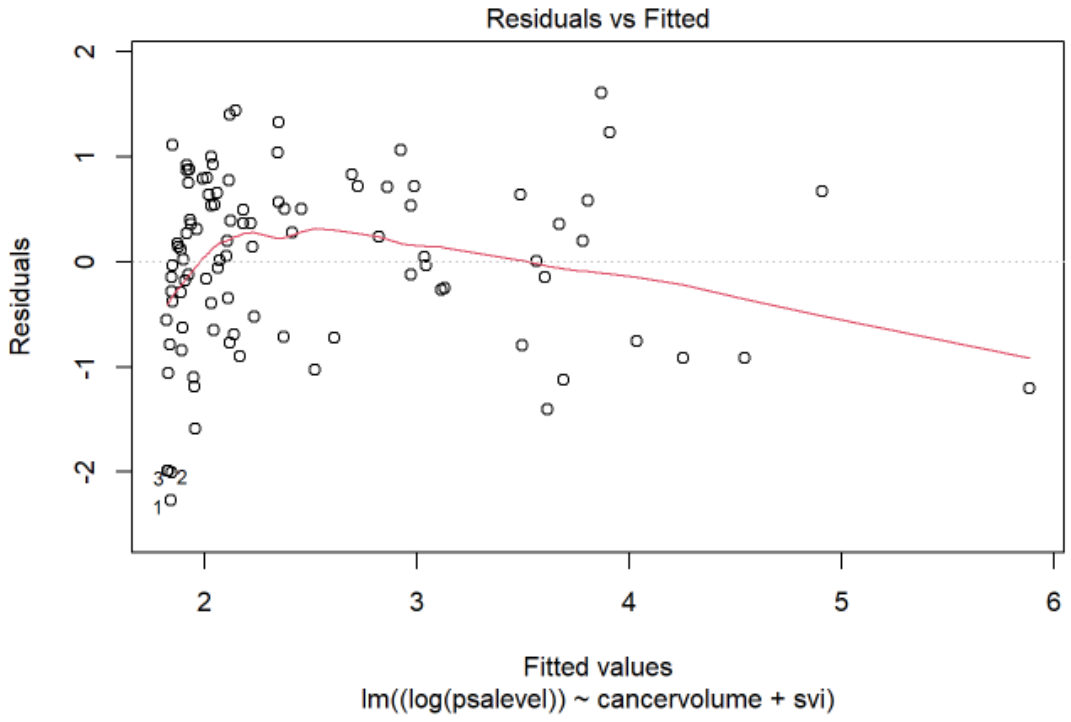
12. Normal Q-Q plot of final model



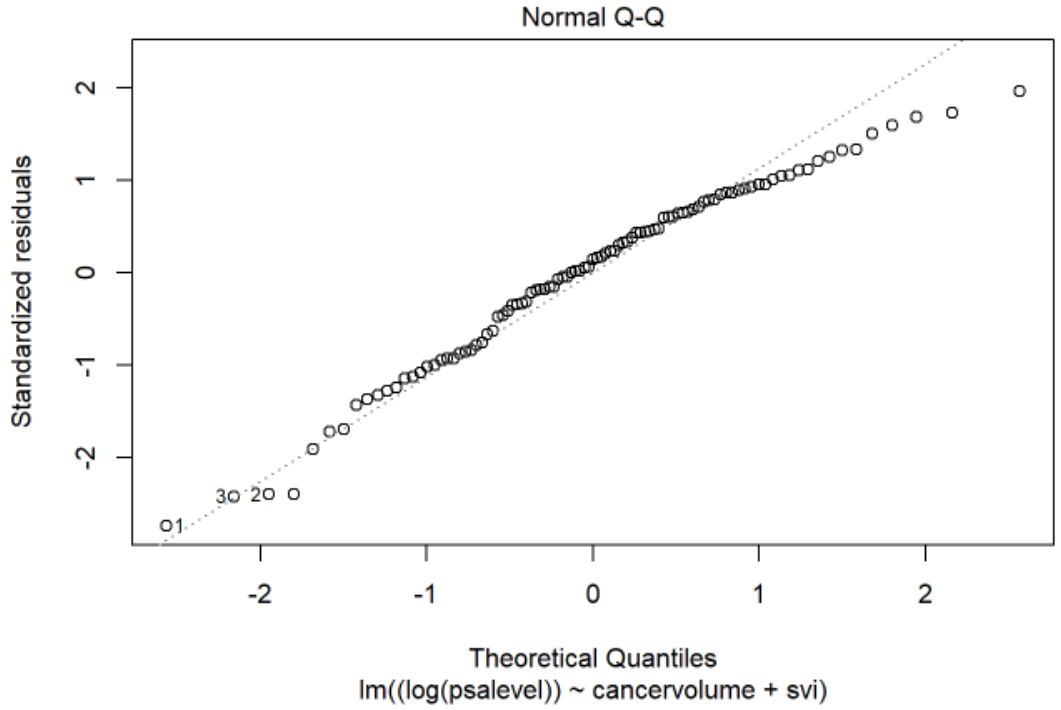
13. Log-likelihood graph of Box-Cox transformation



14. Residual plot of transformed model



15. Normal Q-Q plot of transformed model


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
3
2
1
-2
-1
0
1
2
Theoretical Quantiles
lm((log(psalevel)) ~ cancervolume + svi)