Mini Project 2

Jose Alfaro

Due: February 13, 2019

1. Consider the prostate cancer dataset available on eLearning as prostate cancer.csv. It consists of data on 97 men with advanced prostate cancer. A description of the variables is given in Figure 1. We would like to understand how PSA level is related to the other predictors in the dataset. Note that vesinv is a qualitative variable. You can treat gleason as a quantitative variable.

(a) Perform an exploratory analysis of data.

To explore as much of the data as possible, I created a scatterplot matrix (*Figure 1*) in which all of the variables (except ID) were plotted against each other. This allowed me to observe all trends between each combination of variables.

(b) Is psa appropriate as a response variable or a transformation is necessary? In case a transformation of response is necessary, try the natural log transformation or some other transformation and use it for the rest of this problem.

To assess whether or not PSA Level was a good response variable, I observed the scatterplot matrix produced in Part A (*Figure 1*) and focused on the row comparing PSA Level to all other predictors. This led me to conclude that PSA Level could be a decent response variable, but some transformation would be needed to fix the number of outliers in a lot of the scatterplots. To try to remedy this, I enforced a natural log transformation on the PSA Level variable and produced a second scatterplot matrix (*Figure 2*). In the updated scatter plot matrix, there is an extra variable titled "log_psa" which is simply the natural log transformation of the original psa_level variable. Again, I compared all of the scatterplots between all other variables and log_psa in order to determine if log_psa is a good response variable. As a result, the transformation produced a smaller number of outliers which confirmed it as a better response variable.

(c) Do part (a) of Exercise 15 in Chapter 3 for these data.

Part (a) of the textbook required us to fit simple linear regression models for each predictor in the dataset. This resulted in the creation of seven simple linear regression models, each of which were consisted of log(PSA Level) as the response variable. As a result, only four of these models produced a statistically significant association between the predictor and the response variable. The significant models are displayed below.

- Significant Models:
  - $log(PSA\ Level) = \beta_0 + \beta_1 * Cancer\ Volume$ *(Figure 3)*
    - Note: This model produced a p-value of 2.668e-13
  - $log(PSA\ Level) = \beta_0 + \beta_1 * Vesicular\ Invasion$ *(Figure 4)*
    - Note: This model produced a p-value of 1.481e-09
  - $log(PSA\ Level) = \beta_0 + \beta_1 * Capsular\ Penetration$ *(Figure 5)*
    - Note: This model produced a p-value of 5.503e-08
  - $log(PSA\ Level) = \beta_0 + \beta_1 * Gleason\ Score$ *(Figure 6)*
    - Note: This model produced a p-value of 1.228e-08

For the most part, these simple linear regression models follow the basic assumptions of constant variance and normally distributed with minor deviations. This serves as an indicator that one or more of these variables should be included in the final multiple regression model.

(d) Do part (b) of Exercise 15 in Chapter 3 for these data.

Part (b) of the textbook required us to create a multiple regression model to predict the response. For this step, two models were fit. The first multiple regression model was the "full" model and consisted of all the predictor variables. The second model fitted was a "reduced" model which consisted of only those variables that proved to be significant in part C. The models were fitted are demonstrated below.

- Multiple Regression Models Fitted:
  - $Full\ Model: log(PSA\ Level) = \beta_0 + \beta_1 * Cancer\ Volume + \beta_2 * Weight + \beta_3 * Age + \beta_4 * Benign\ Prostatic\ Hyperplasia + *\beta_5 * Seminal\ Vesicle\ Invasion + \beta_6 * Capsular\ Penetration + \beta_7 * Gleason\ Score$
  - $Reduced\ Model: log(PSA\ Level) = \beta_0 + \beta_1 * Cancer\ Volume + \beta_2 * Seminal\ Vesicle\ Invasion + \beta_3 * Capsular\ Penetration + \beta_4 * Gleason\ Score$

However, when I conducted a partial F-test to test whether the reduced model explained as much information as the full model, the test resulted in a small p-value of 0.00722. Since this p-value is small, this indicates that we reject the null hypothesis and conclude that the full and reduced models do not contain the same information. In other words, we cannot conclude that the reduced model is more useful than the full model, thus we must use the full model in this case.

(e) Build a "reasonably good" multiple regression model for these data. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions.

Since our reduced model in the previous step was not as useful as the full model, we must find a simpler model. To do this, we used model selection procedures such as forward, backward, and stepwise selection based on AIC values. As a result, all three of the model selection procedures selected the same model shown below.

- Forward, Backward, and Stepwise Model Selection:
  - $Reduced\ Model: log(PSA\ Level) = \beta_0 + \beta_1 * Cancer\ Volume + \beta_2 * Benign\ Prostatic\ Hyperplasia + \beta_3 * Seminal\ Vesicle\ Invasion + \beta_4 * Gleason\ Score$

F-Test Hypothesis:
$H_0 : \beta_{p-q+1} = ... = \beta_p = 0$
$H_1 : At\ least\ one\ of\ these\ q\ slopes \neq 0$

Similar as before, to test whether this new reduced model obtained via model selection procedures explains the same amount of information as the full model, we conducted a partial F-test. However, this time the partial F-test produced a p-value of 0.7353 which leads us to fail to reject the null hypothesis and conclude that the reduced model explains the same amount of information as the full model. In other words, it is acceptable for us to use the reduced multiple regression model as opposed to the full multiple regression model.

In order to validate that this multiple regression model is valid, we must varify that the following model assumptions are met: constant variance, variance centered around 0, normally distributed data.

By observing *Figure 7*, we can easily tell that the residuals of this multiple regression model are centered around 0. Furthermore, one could say that the variance is non-constant since the residual plot in *Figure 7* shows a slight curvature. This would then indicate that one of our key regression assumptions is violated. However, this astute observation could easily be due to the small sample size provided in the data. If we were provided with more data points, then this issue of slight non-constant variance could potentially be resolved. Lastly, we checked our QQ plot in search of a linear trend that closely follows the 45 degree line. As a result, *Figure 7* presents a QQ plot for the model which shows a a strong linear pattern indicating that the data are normally distributed.

In conclusion, since there are no strong violations of the regression assumptions, then this model is deemed valid and is thus selected as our final multiple regression model.

(f) Write the final model in equation form, being careful to handle qualitative predictors (if any) properly.

Since the Seminal Vesicle Invasion predictor is a categorical variable, we must be careful in writing the model. Therefore, we must account for situations where the categorical predictor takes on both of its values (0 and 1) by fitting two distinct models. The models are shown below.

- Final Model:
  - $Model\ when\ SVI = 1 : log(PSA\ Level) = \beta_0 + \beta_1 * Cancer\ Volume + \beta_2 * Benign\ Prostatic\ Hyperplasia + \beta_3 * Seminal\ Vesical\ Invasion + \beta_4 * Gleason\ Score$
  - $Model\ when\ SVI = 0 : log(PSA\ Level) = \beta_0 + \beta_1 * Cancer\ Volume + \beta_2 * Benign\ Prostatic\ Hyperplasia + \beta_4 * Gleason\ Score$

(g) Use the final model to predict the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors (if any) are at the most frequent category.

The sample means for the qualitative predictors as well as the mode of the quantitative predictor variable in the model are listed below.

- Summary Statistics:
  - Mean of Cancer Volume: 6.998682
  - Mean of Benign Prostatic Hyperplasia: 2.534725
  - Mean of Gleason Score: 6.876289
  - Mode of Seminal Vesicular Invasion: 0

We used these statistics to predict the PSA Level for a patient who consists of values similar to the sample means of our predictor variable. Since we cannot take the sample mean of a categorical variable, we simply took the mode. Therefore the predicted PSA level of a patient whose quantitative predictors are the sample means of the variables is 10.28357.

$exp(log(PSALevel)) = exp(-0.65013 + (0.06488 * 6.998682) + (0.09136 * 2.534725) + (0.33376 * 6.876289)) = 10.28357$

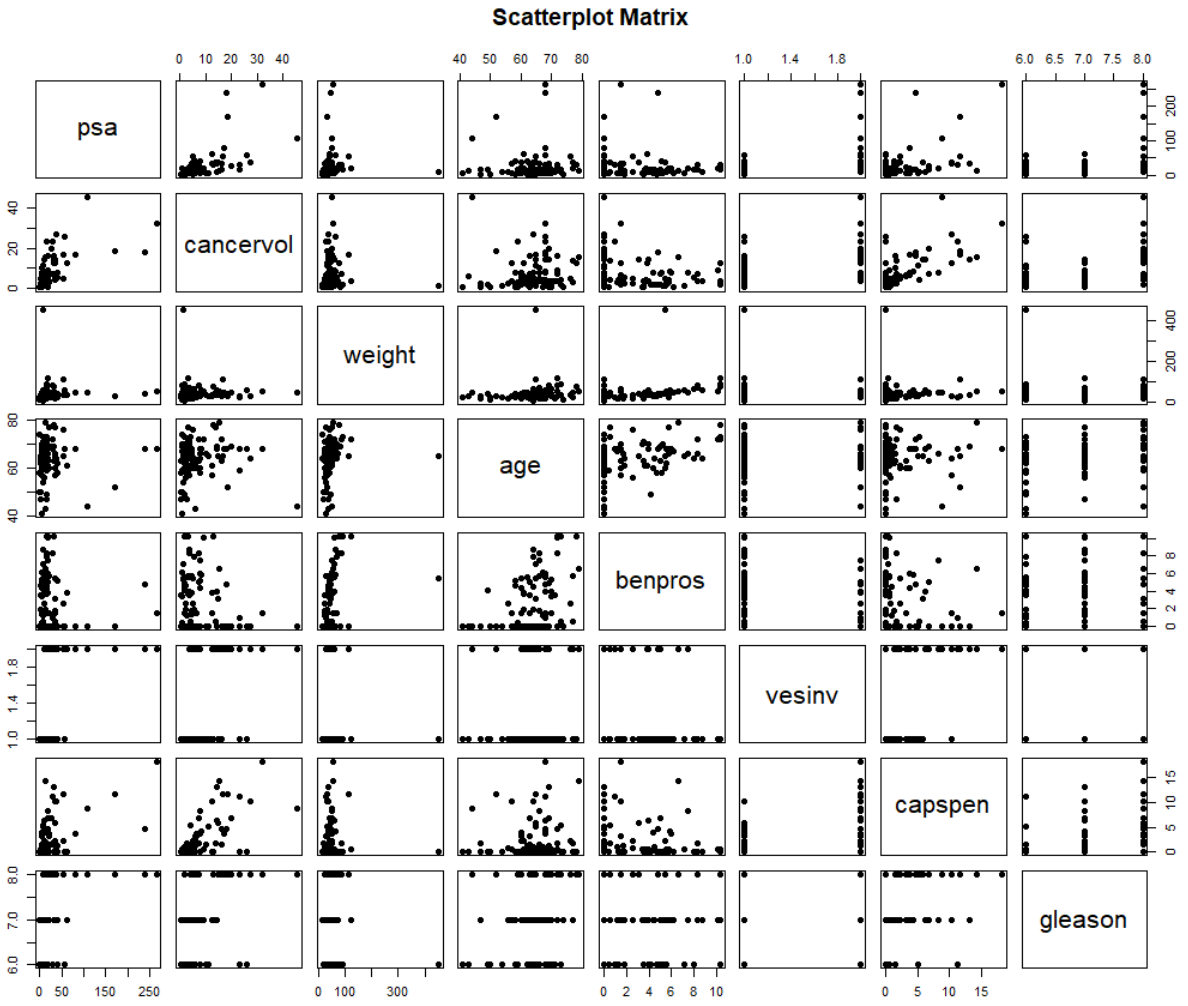

Figure 1: Scatterplot Matrix for Prostate Cancer Dataset

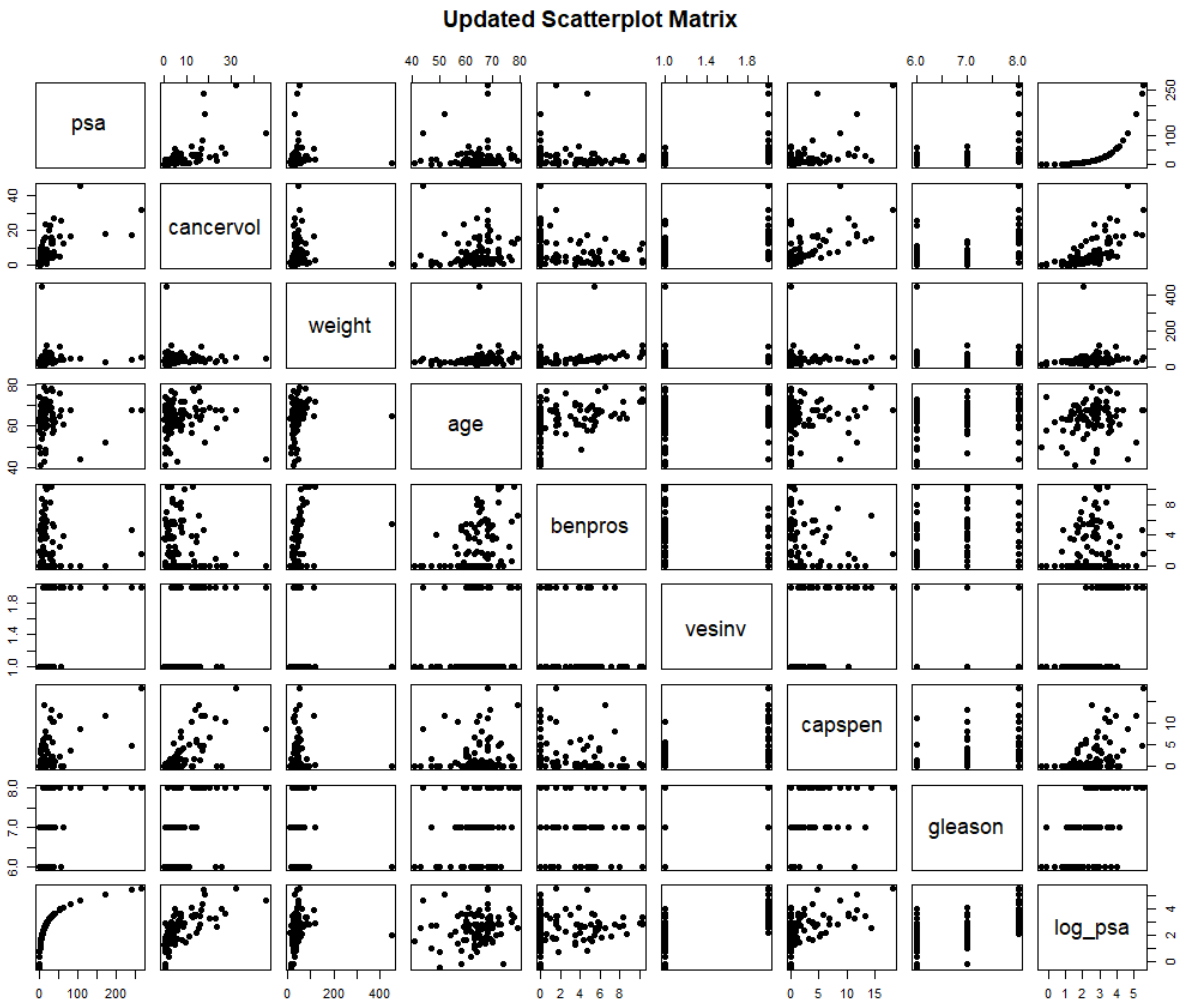

Figure 2: Updated Scatterplot Matrix

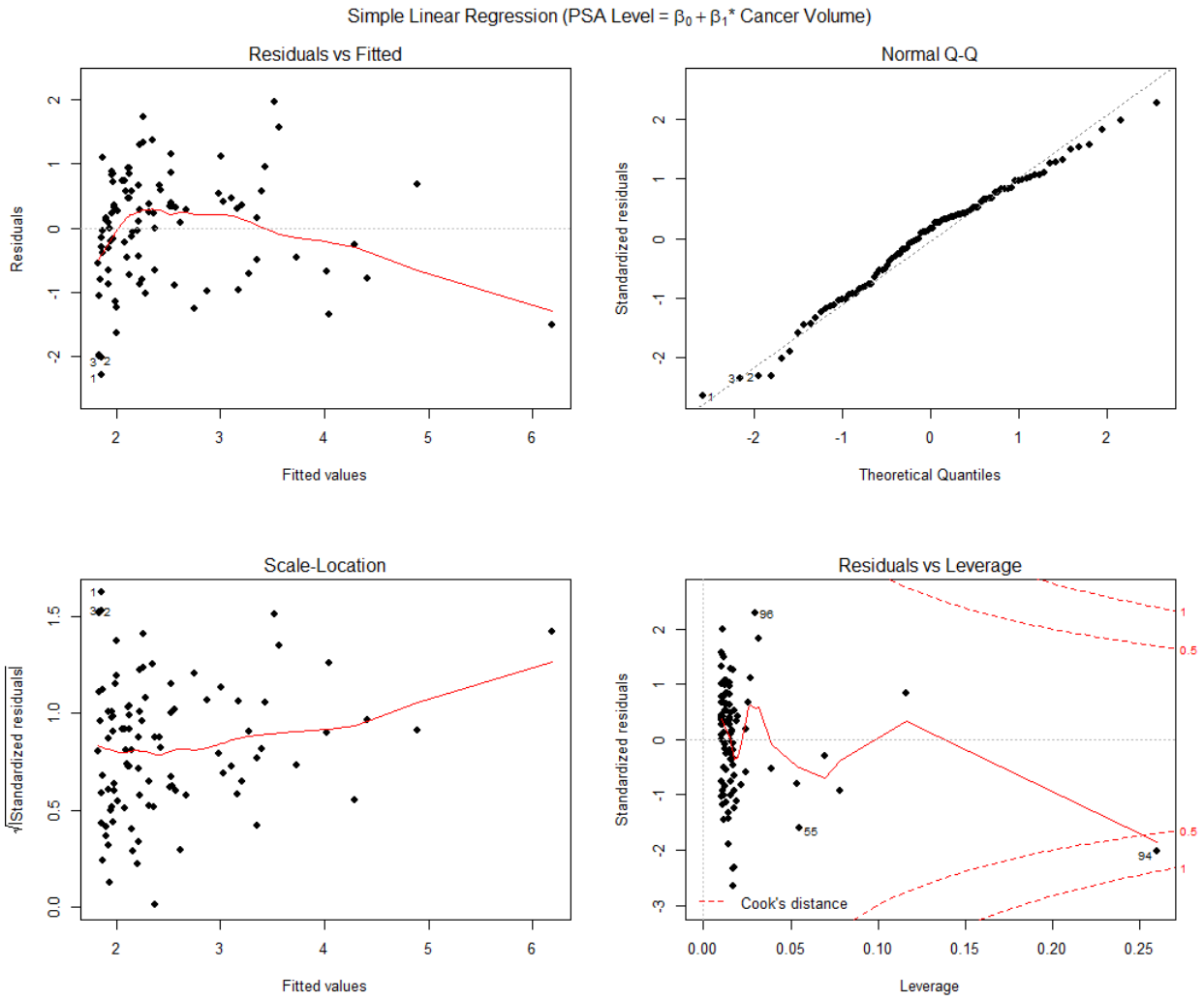

Figure 3: Simple Linear Regression (Cancer Volume) Diagnostics

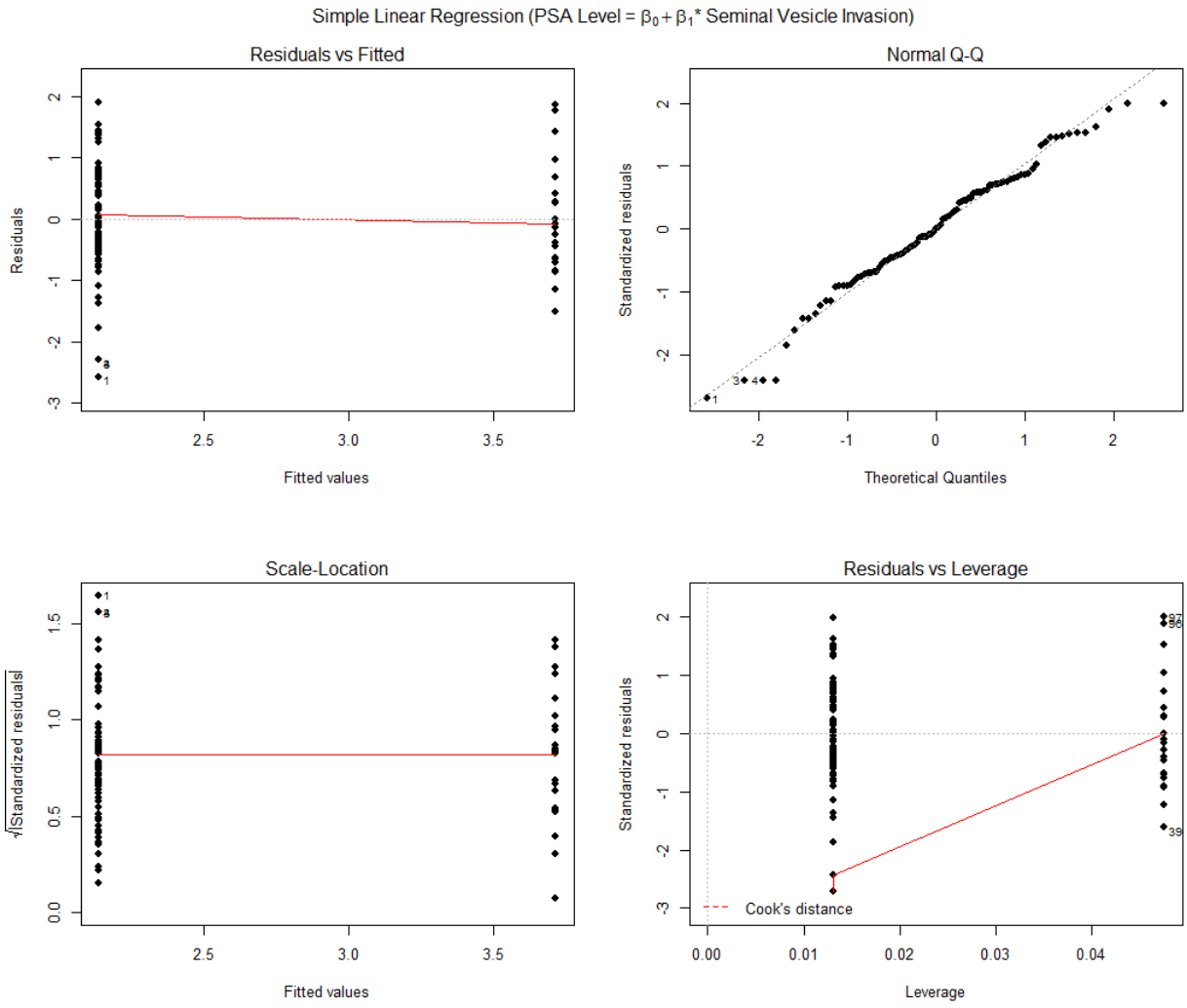

Figure 4: Simple Linear Regression (Seminal Vesicle Invasion) Diagnostics

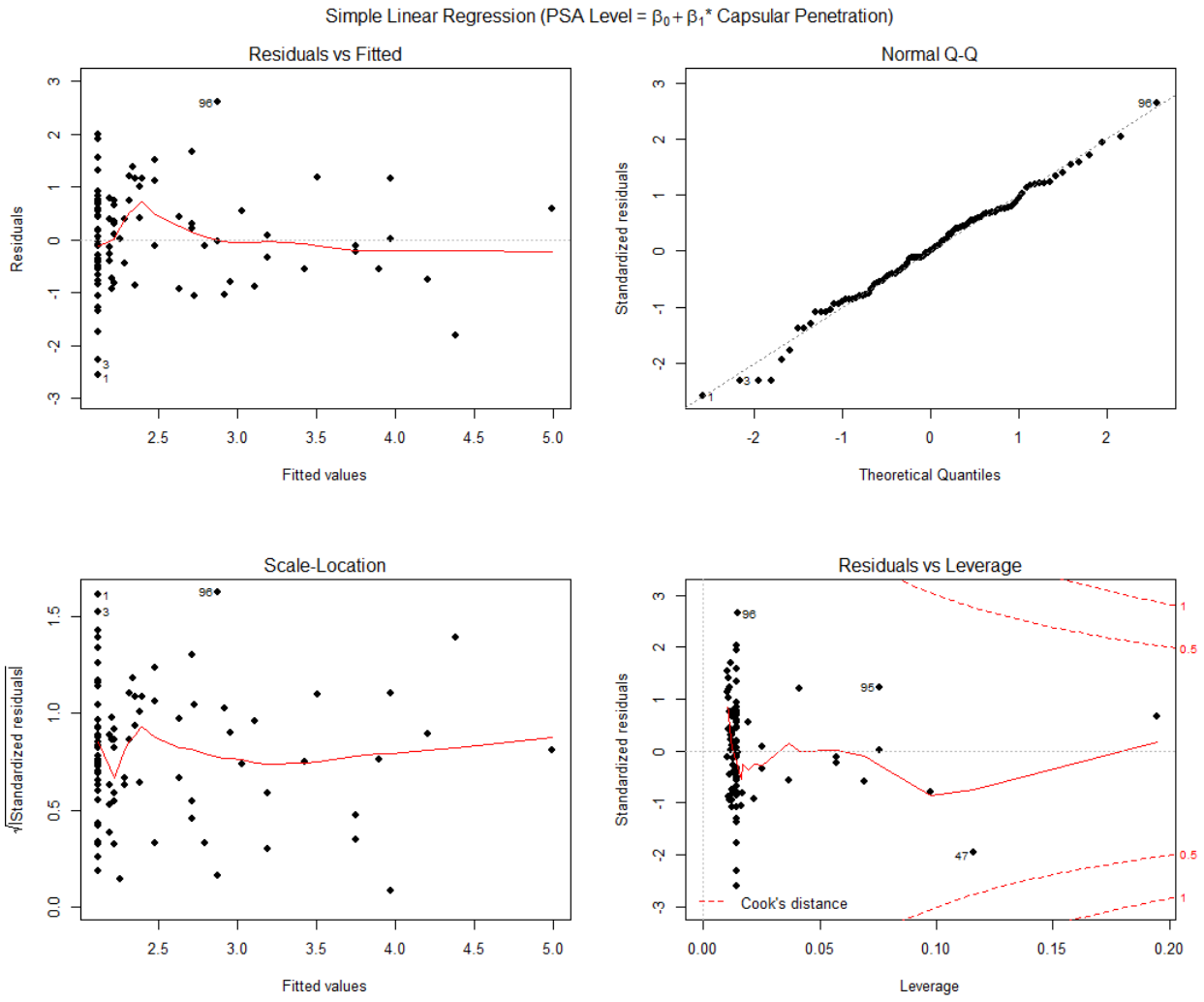

Figure 5: Simple Linear Regression (Capsular Penetration) Diagnostics

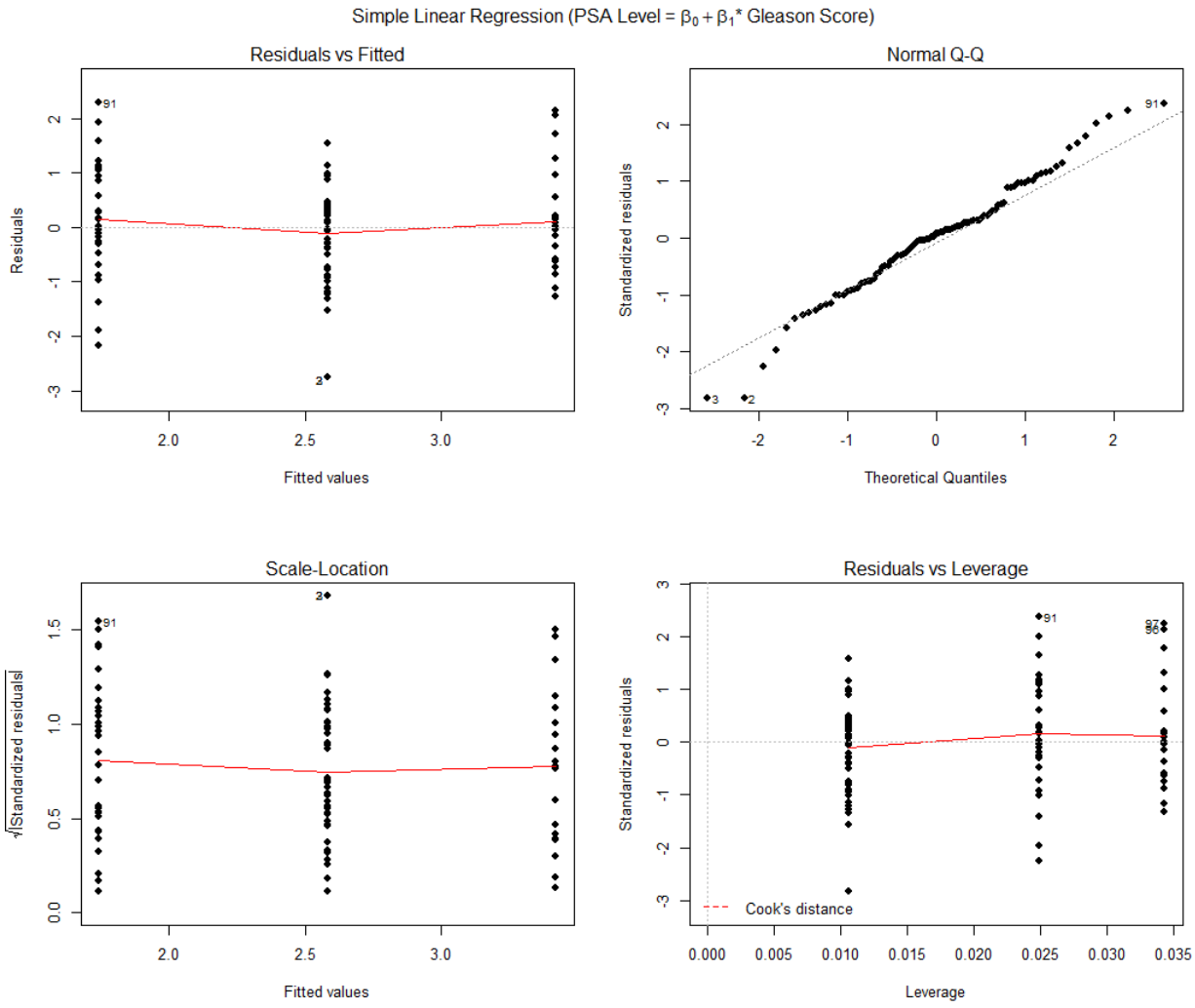

Figure 6: Simple Linear Regression (Gleason Score) Diagnostics

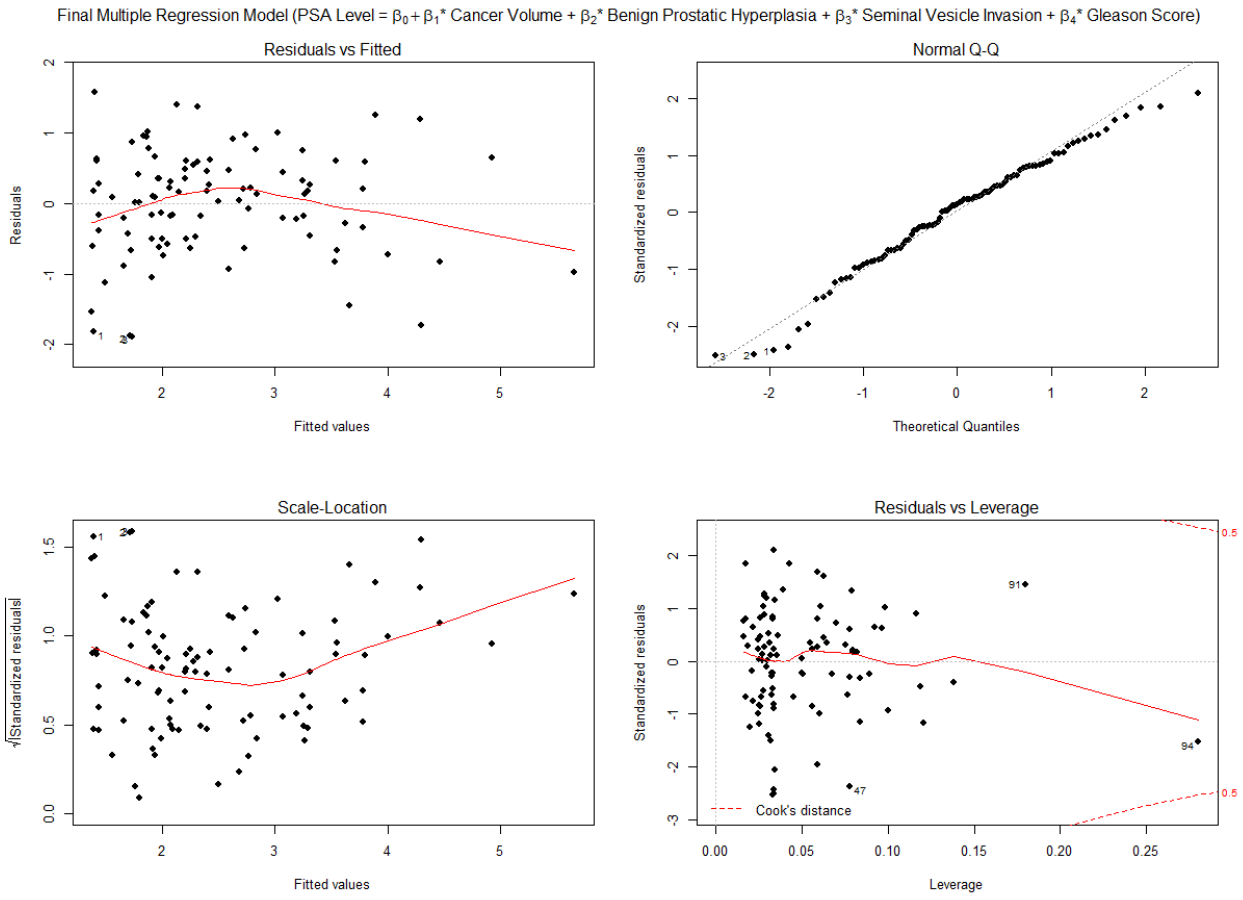

Figure 7: Final Model Diagnostics