Assignment 1 - SS9155 - 250620601

Ravin Lathigra

2019-01-14

R Packages & Libraries

```
library(corrplot)    #Visualize Correlation between variables
library(kableExtra)  #Style tables
library(tidyverse)   #contains ggplot2,dplyr,tidyr, readr,purr,tibble,stringr,forcats
library(formatR)     #Improve readability of code
library(e1071)       #Functions for latent class analysis, Fourier transform ect.
library(VIM)         #Knn
library(ggfortify)   #Add on to ggplot2 to allow for more plot types
library(Rtsne)       #Dimension reduction classification
library(caret)       #streamlined model development
library(RColorBrewer)#Control colours of visualizations
library(GGally)      #Contains ggpairs plots
library(lmtest)      #Test for linear assumptions
library(MASS)
library(faraway)
library(lasso2)
```

Understanding the Data

The data used throughout the following analysis was gathered from the `Faraway` package, though sourced from **Andrews DF and Herzberg AM (1985)**. The data contains patient information for 97 men diagnosed with prostate cancer who were due to undergo a prostatectomy. Patient details include 9 variables namely `log(cancer volume)`, `log(prostate weight)`, `age`, `log(benign prostate hyperplasia amaunt)`, `seminal vesicle invasion`, `log(capsular penetration)`, `Gleason score`, `presentage Gleason scores 4 or 5`, & `log(prostate specific antigen)` which for the purposes of the investigation were encoded as `lcavol`, `lweight`, `age`, `lbph`, `svi`, `lcp`,`gleason`, `pgg45`, & `lpsa` respectively.

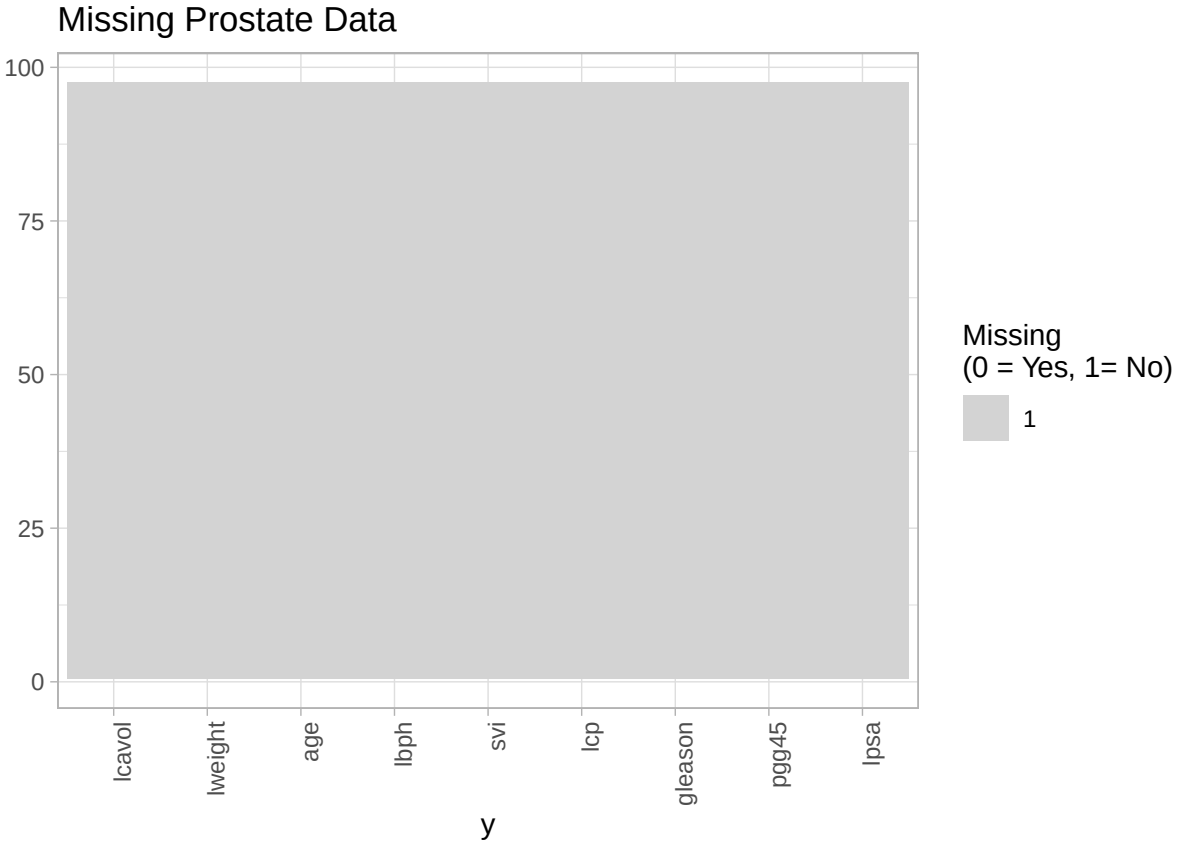
The goal of the investigation is to explore the relationship between `lpsa` and the other predictors included in the `prostate` dataset. After developing an understanding of the relationships among predictors, a regression model can be developed to predict `lpsa`. It is important that before the analysis begins that the structure of the data is understood. The following r output displays the structure of the prostate data. It should be noted that `svi` and `gleason` were origionally classified as numeric, though it is more appropriate to treat them as factor variables. `svi` is a binary indicator of whether or not the cancer has spread to the seminal vesicles, `gleason` is a discrete risk measure assigned to a biopsy of affected tissue.

```
## 'data.frame':    97 obs. of  9 variables:
##  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
##  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
##  $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
##  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ svi    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ gleason: Factor w/ 4 levels "6","7","8","9": 1 1 2 1 1 1 1 1 1 1 ...
##  $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
##  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

Missing Data

Once the structure of the data is understood, the completness of the data can be assessed. *Missing Prostate Data* shows whether data is missing or available for each observation across all predictors. Missing data is highlighted in purple, though in this case, there is no missing data across any predictors.



Data Exploration

Summary of missing and existing observations by variable

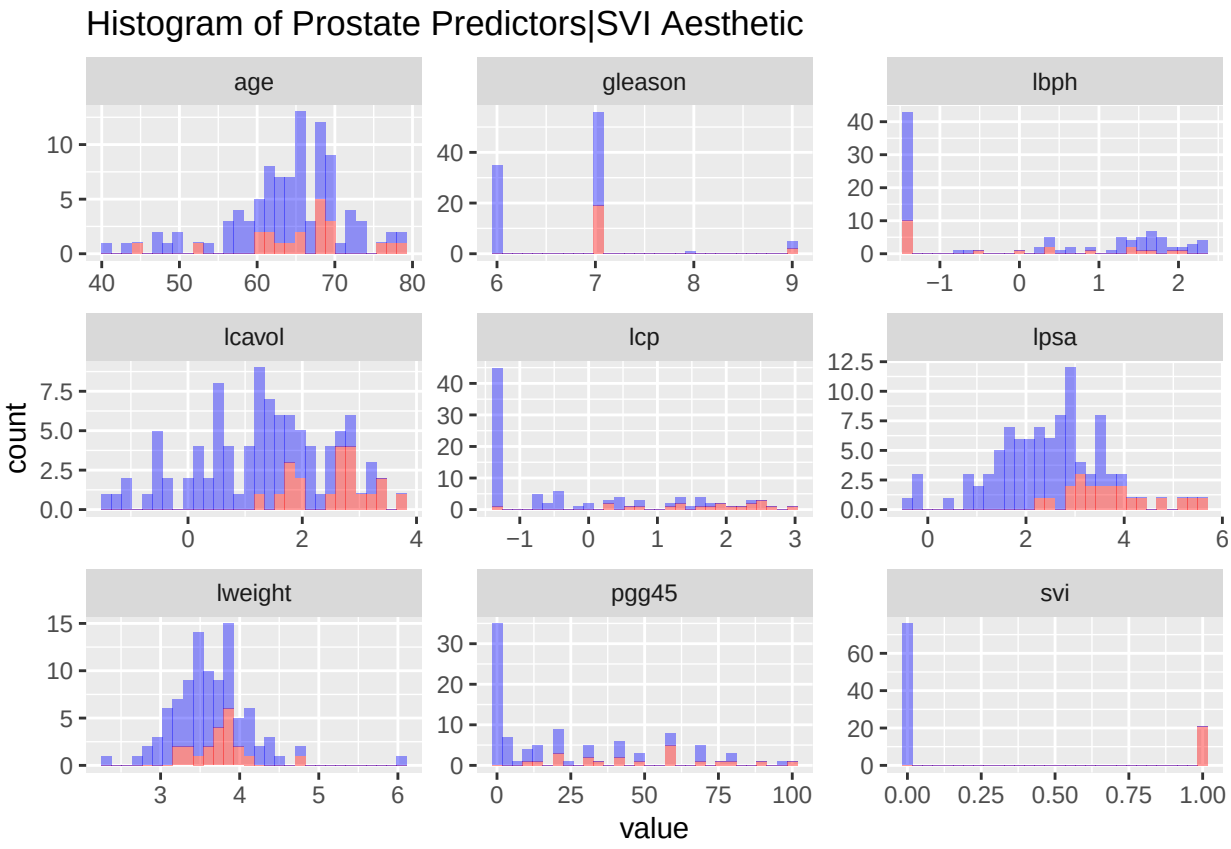
The following plots compare and contrast the distributions of each predictor using histograms. To help explore underlying groupings within the data, aesthetics can be added to the plots. Adding aesthetics allows for groups to be directly compared. There are 2 factor variables present in the data `svi` and `gleason`.

When `svi` is used as an aesthetic, blue represents a lack of invasion while red indicates invasion.

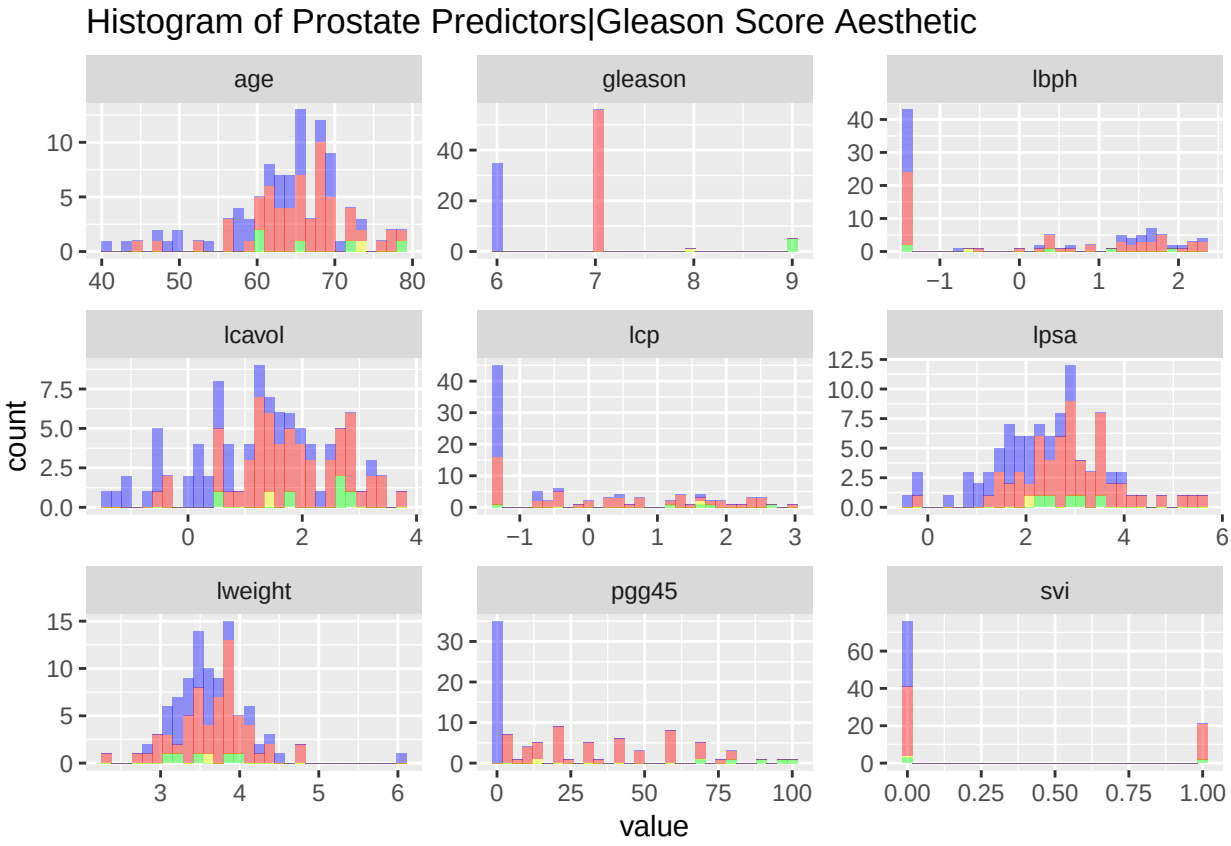
Similarly, when `gleason` is an aesthetic, blue, red,yellow and green represents gleason scores of 6,7,8 and 9 respectively.

The benefit to these visualization are they capture both discrete and continuous predictors however, since the proportion of data between the groups is not equivalent it is difficult to directly compare the distrubtions. It is worth noting, that there is an imbalance of groups, which may lead to sparse representation.

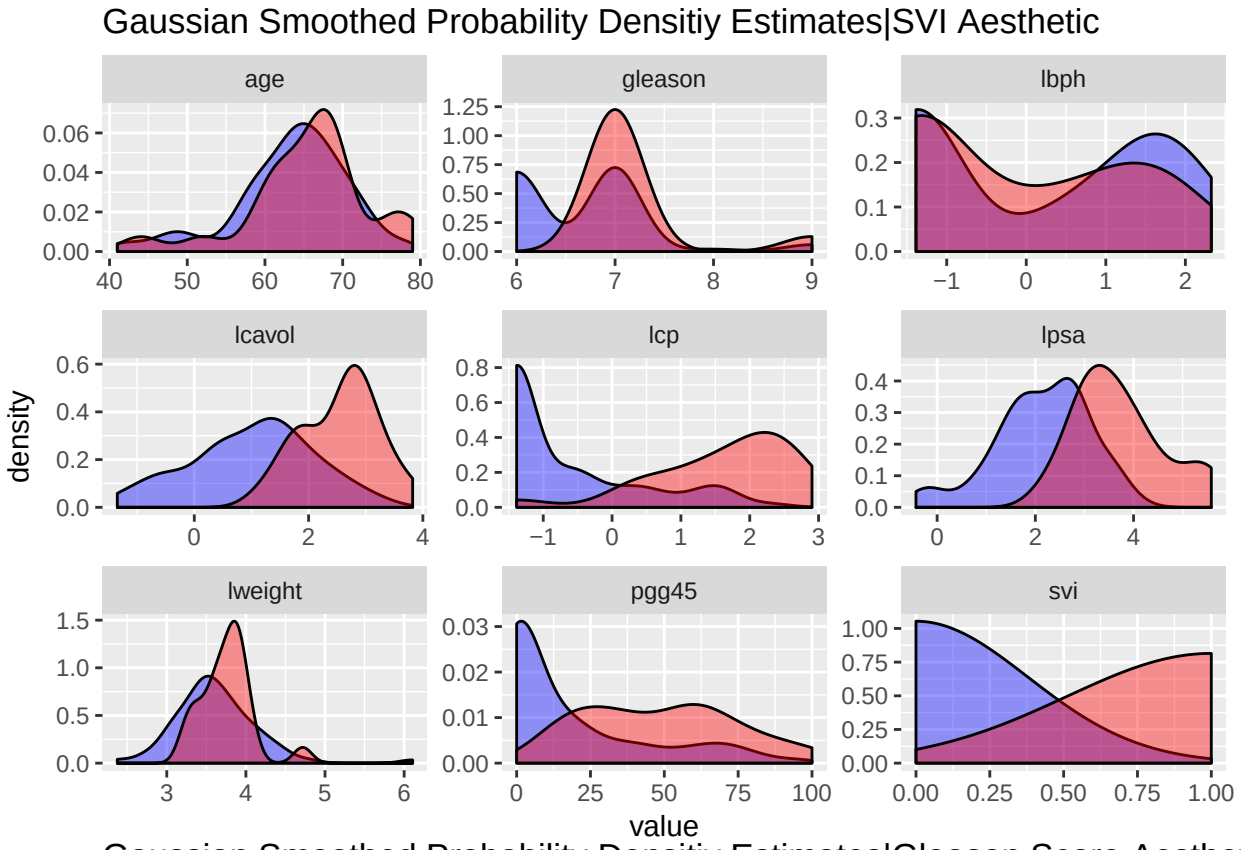
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

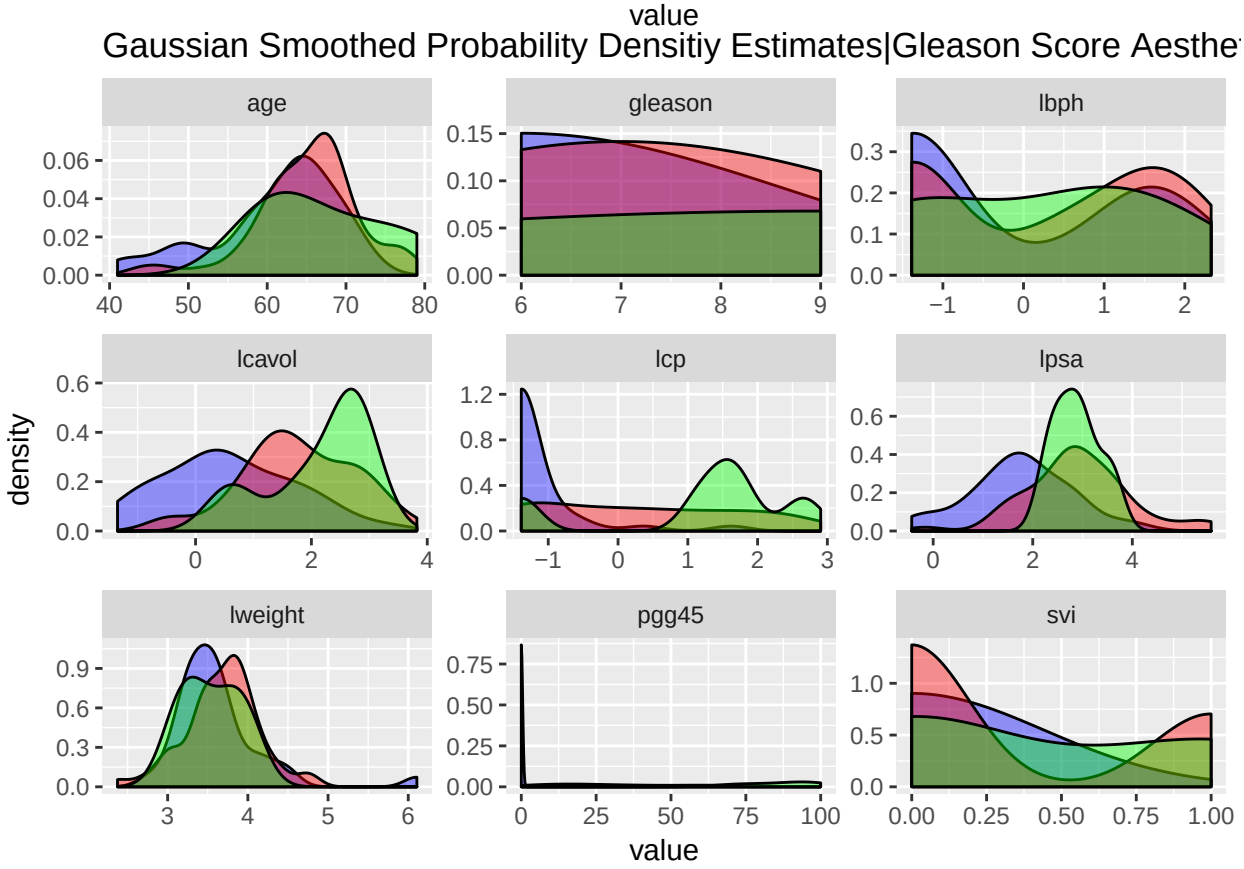



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can improve the comparisons that can be made from the data by estimating probability density functions. To promote smooth PDFs, a gaussian kernal is used. The plot *Gaussian Smoothed Probability Densitiy Estimates/SVI Aesthetic* gives insight that there is a significant degree of differentiation within the data when comparing seminal vesical invation. If instead, the data is grouped by `gleason` there is a lesser degree of seperation between the groups though there appears to be a relationship between `gleason` and `lpsa`. In particular, as gleason increases, we expect to see an increase in `lcavol` and a decrease in the variance of `lpsa`. This is inline with medical intuition though it is a good way to confirm that our data is reasonable.



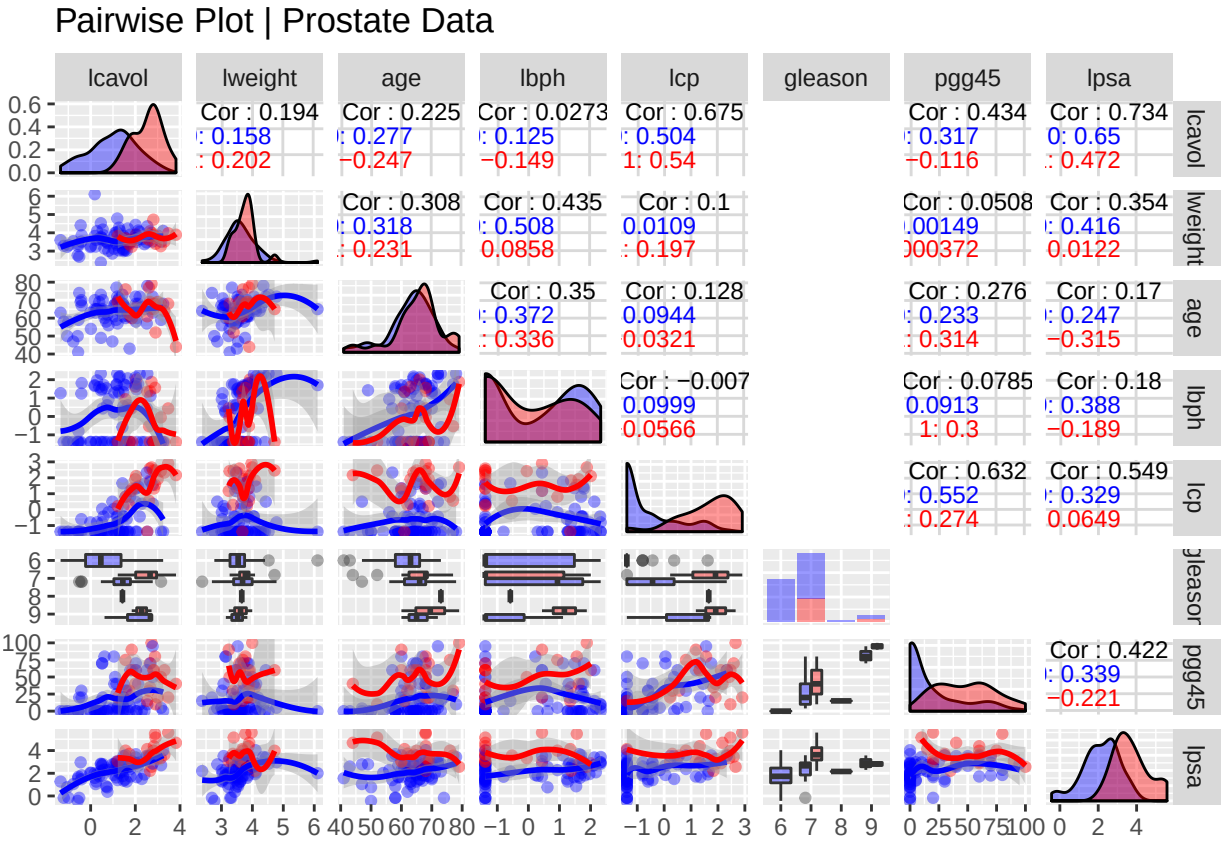


For further exploration into the data, pairwise comparisons of the data can be visualized. The plot *Pairwise*

Plot | Prostate Data has 3 important comparisons displayed. As the estimated PDFs sugggested, data is split by `svi` to allow for more detailed comparisons to be made.

- Lower: between predictor scatter plot for continuous data and boxplots for factor predictors. Further enhancements include a loess smoother applied to the scatter plot to capture relationships among predictors.
- Diagonal: Gaussian Smoothed Estimated PDFs for continuous variables and frequency plot for factor variables.
- Upper: Between predictor correlation by svi group as well as ungrouped correlation.

Perhaps the most significant inferences that can be made from is a moderate positive correlation between `lcavol` and `lpsa` which provides additional support for previous observations regarding the data. Additionally, it suggests that as gleason scores increase the `lpsa` may rise and variance decreases.



To better visualize the correlations among the predictors a corrplot can be used. The following corrplot, simplifies the correlations presented in the pairwise plot. If correlations among predictors- particularly predictors that are not `lpsa`- are large collinearity may present in the model.

From a preProcessing perspective, the feature space of the dataset can be reduced using a correlation reduction algorithm. The following algorithm is used to reduce the dimensions of the data by removing variables with the most correlated relationships with the other variables.

1. Calculate Correlation Matrix of the variables
2. Determine the 2 predictors with the greatest absolute pariwise correlation namely A & B.
3. Determine the average correlation between A and the other variables. Repeat for B.
4. If avg correlation for A > B then remove A. Otherwise remove B.
5. Repeat until theire are no pairwise correlations greater than 0.75.

The corrplot suggests that there are predictors that have a moderate degree of correlation, though they did not meet the threshold required to be removed from the model.

Further preProcessing that can be applied to the data can be easily carried out using the `caret` package.

The **caret** package offers the function `preProcess` which allows for the following:

- centering data
- scaling data
- Remove predictors with near zero or zero variance
- Remove predictors with large parwise correlation

The only preProcessing that was applicable to the prostate data set was 7 predictors centered and scaled, more specifically the non-factor predictors.

High Dimension Visualization

Principle Component Analysis

Principle component analysis is a dimenstion reduction technique that attempts to capture the maximum amount of variance within the predictors, i.e ignoring the response variable, using orthagonal linear combinations of the predictors Using the first 2 principle components (linear combinations) only 65% of the variation is captured therefore, there is not enough preservation of data to see if there is high dimension seperation within the data.

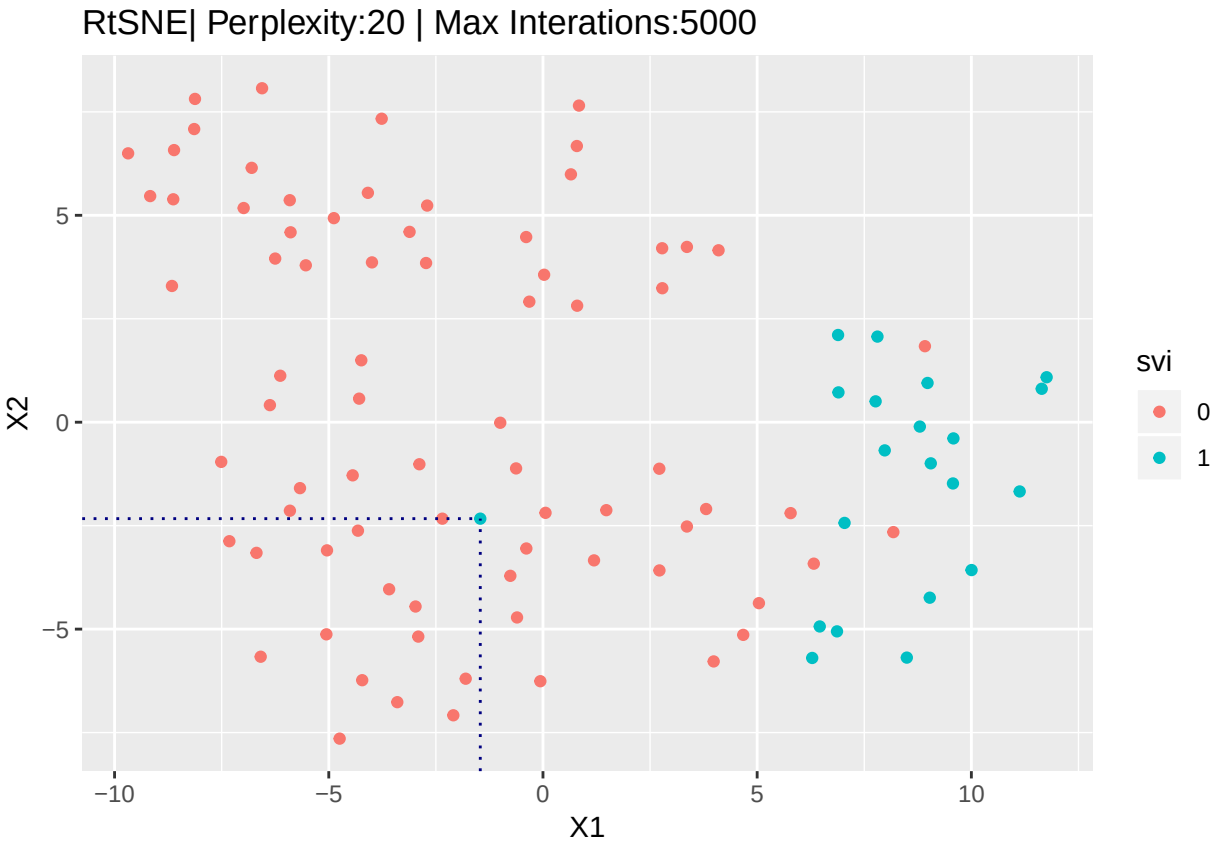
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7204 1.2520 0.9300 0.77578 0.69753 0.57637 0.43245
## Proportion of Variance 0.4228 0.2239 0.1236 0.08598 0.06951 0.04746 0.02672
## Cumulative Proportion  0.4228 0.6468 0.7703 0.85632 0.92583 0.97328 1.00000
```

t-Distributed Stochastic Neighbor Embedding

PCA can be an effective dimension reduction technique which can subsequently be used to identify clusters, however this method aims to capture variation among the feature space. As noted, there was not enough variance captured in the first 2 principle components to appropriately visualize the data in 2 dimension. Instead we can use a highly regarded method known as t-Distributed Stochastic Neighbor Embedding (tSNE). This approach says that if there exists relationships amongst the varaiables in high dimensions then we can display them in lower dimensions. This method offers advantages over pca particularly if there is only a small amount of variation explained by the first two PC's, as it looks to preserve relationships **not** maximize variance. With this, we can take our data and identify the relationships in high dimensions and display them in 2. If there is evidence of seperation, it would be interesting to determine what factors are most responsible for creating distinct grouping in the data.

tSNE does not produce a stable output i.e each iteration through thte data, the transformation of the observations can change. The parameter in tSNE that can be considered a **tuning** parameter is called perplexity. There is no way to empirically estimate this parameter, but since tSNE is simply a techniqe used to express the relationships in high dimension in lower dimensions, we can iterate through the data numerous times using tSNE considering serveral values for the tuning parameter, perplexity, such that we minimize the Kullback-Leibler divergence. For our purposes our parameter space for perplexity is 5,10,15, and 20 and we conduct tSNE 5 times, considering a maximum of 5000 iterations for each tSNE transformation to ensure that our KL divergence converges.

The following plot illstrate the *best* tSNE representation i.e lowest KL-Divergence for each of the considered perplexities:



For the purposes of regression analysis, the ability to seperate high dimensional data in lower dimension is not essential though it is interesting. It also highlights a few interesting obvservations. In particular, the dotted lines highlight a `svi` of group 1 i.e seminal vesicle invasion. It would not be surprising that there is a degree of error when dealing with biological information or human gathered data. For the purposes of the

investigation, all datapoints are considered though if the data could be reviewed with the initial publishers some points -like the highlighted one- could be further validated.

Data Splitting

Given the limited data, conservation of data is essential which suggests rather than using training, validation and test sets we should use only a training and test set. To ensure that we are appropriately estimating the generalization ("test") error we use 5 times 10 fold cross-validation. Furthermore, the data partitioned into 70/30 splits for the training and test respectively.

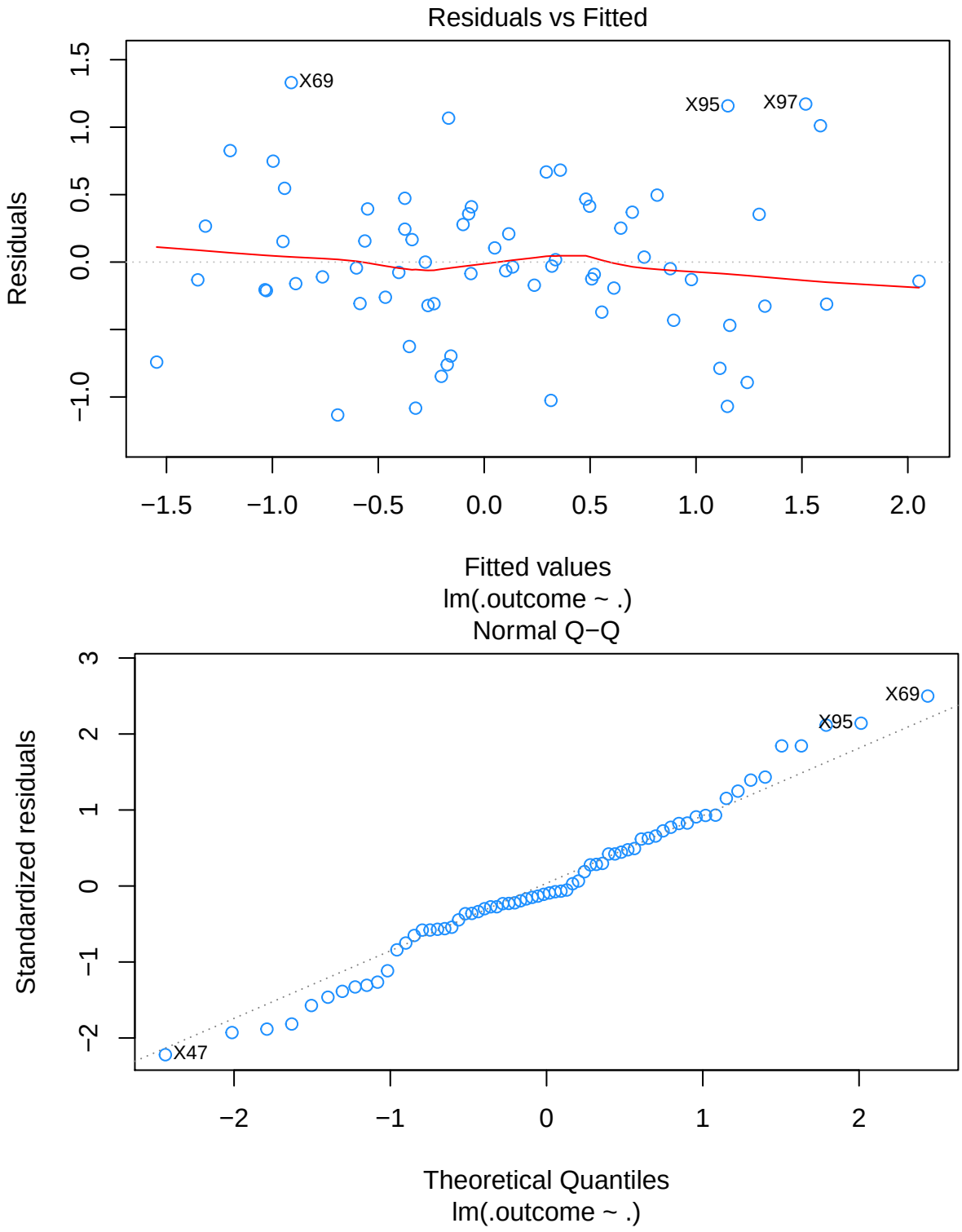
Model Development

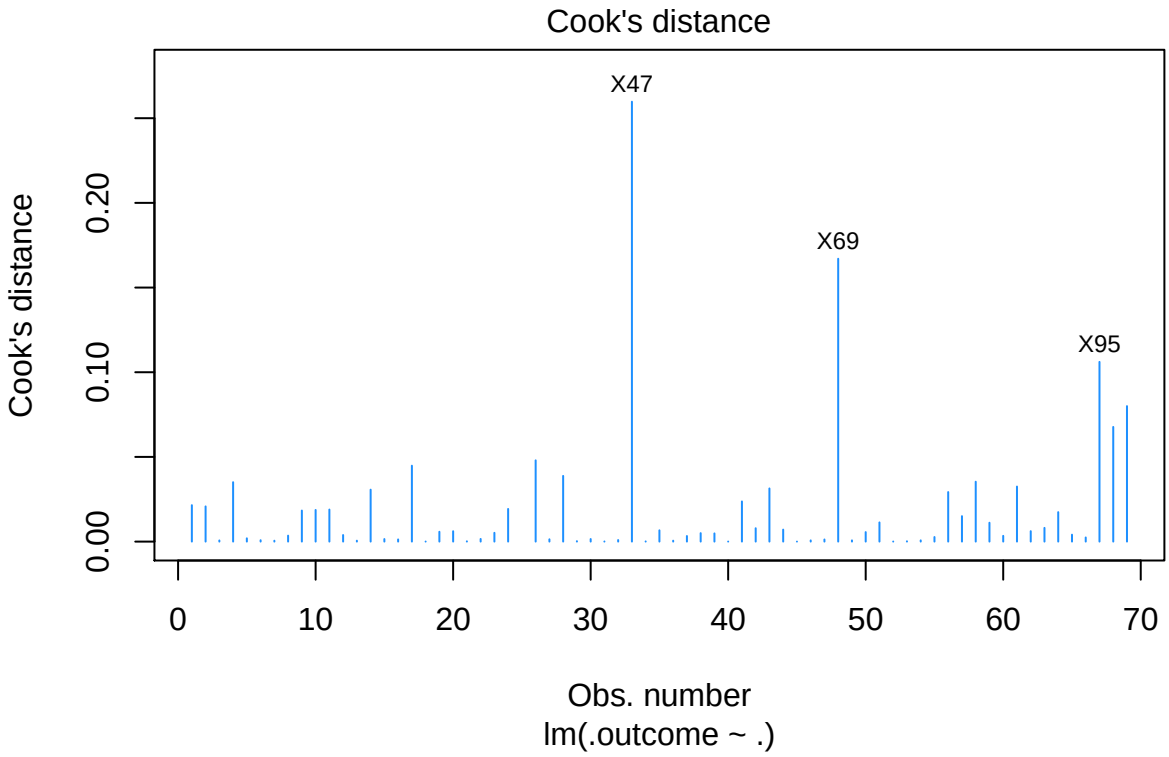
For model development, we consider the following candidate models:

- linear model
- Stepwise Linear Model

Linear Model

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1335 -0.3085 -0.0492  0.3575  1.3304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26325    0.18491  -1.424   0.1599
## lcavol       0.51982    0.11972   4.342 5.74e-05 ***
## lweight      0.22428    0.10281   2.182   0.0332 *
## age         -0.14017    0.09032  -1.552   0.1261
## lbph         0.13391    0.08901   1.504   0.1379
## svi1         0.75923    0.28656   2.649   0.0104 *
## lcp         -0.13227    0.15616  -0.847   0.4005
## gleason7     0.29272    0.23867   1.226   0.2250
## gleason8     0.38930    0.70495   0.552   0.5829
## gleason9     0.03972    0.49997   0.079   0.9370
## pgg45        0.08908    0.12591   0.708   0.4821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6055 on 58 degrees of freedom
## Multiple R-squared:  0.6839, Adjusted R-squared:  0.6294
## F-statistic: 12.55 on 10 and 58 DF,  p-value: 2.987e-11
```


Residuals vs Fitted
Residuals
X69
X95
X97
1.5
1.0
0.5
0.0
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Fitted values
lm(.outcome ~ .)
Normal Q−Q
Standardized residuals
X69
X95
X47
3
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(.outcome ~ .)




```
##
##  studentized Breusch-Pagan test
##
## data:  caret_lm$finalModel
## BP = 12.718, df = 10, p-value = 0.2399
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(caret_lm$finalModel)
## W = 0.98048, p-value = 0.3542
```

Linearity - Inspecting the plot "Residuals Vs Fitted" has a trend line that helps illustrate that there is a no clear tend between the fitted values and residuals. Furthermore, the residuals generally exhibit zero mean suggesting that a linear model may be appropriate.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see there is generally constant variance and no obvious trends suggesting a linear model may be appropriate. A BP test can confirm if this assumption is not violated.

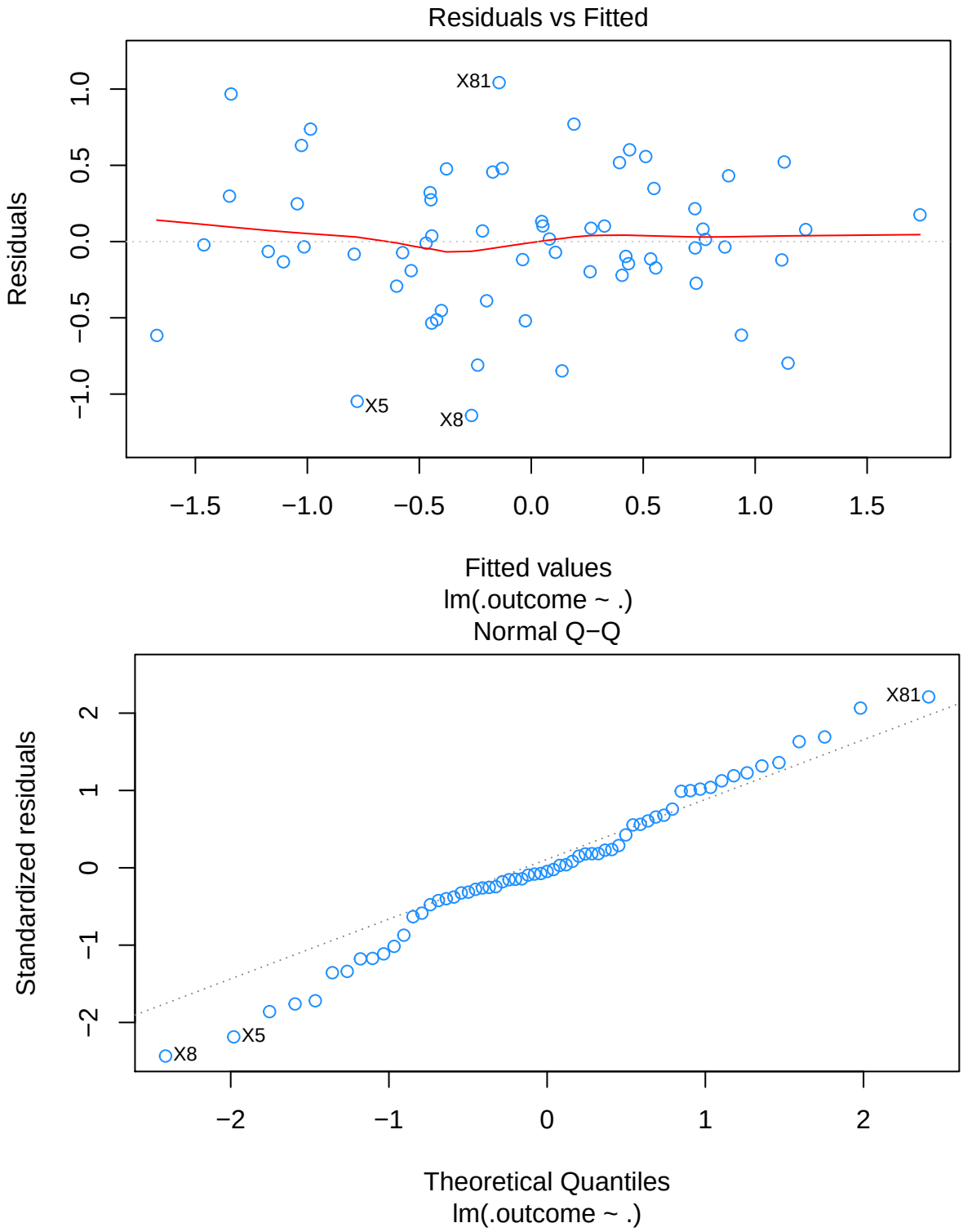
Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals moderately correspond to the theoretical quantiles of a normal distribution. At the extremes of the plot, the observed and theoretical quantiles deviate though additional testing via Shapiro test can indicate if the normality assumption is violated.

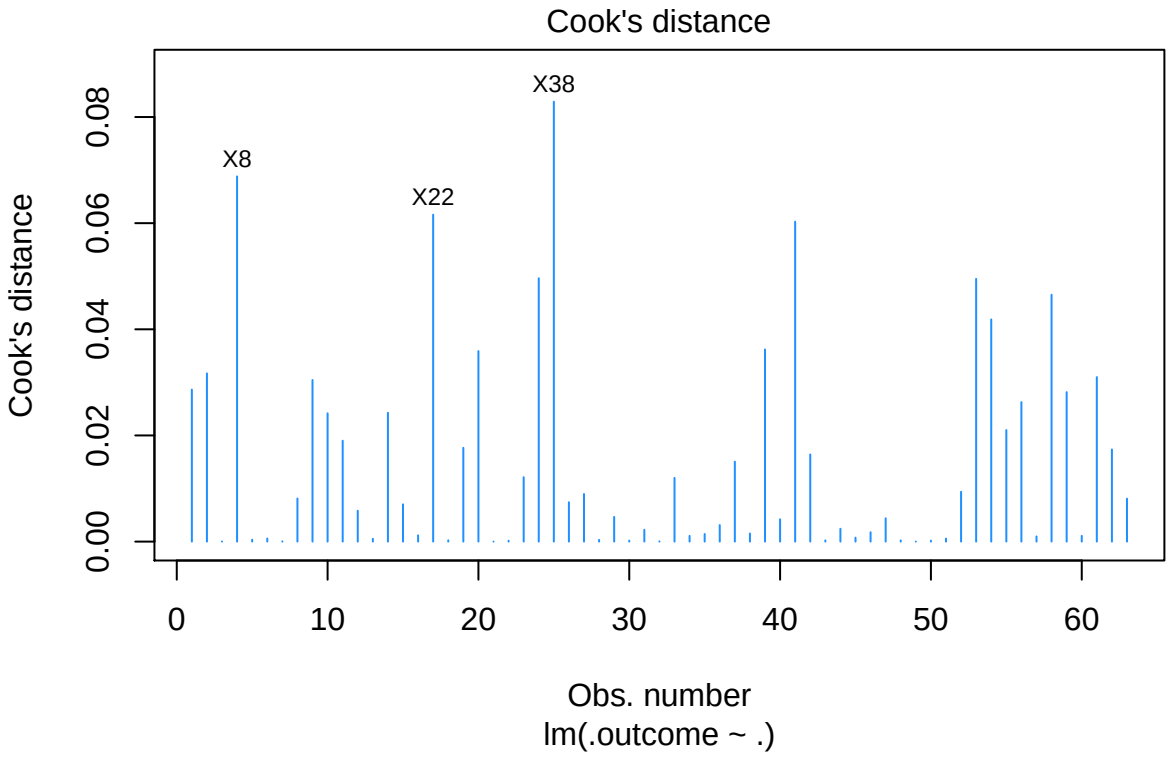
Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation. When dealing with biological experiments there can be anomolies in the observations which arise for a number of reasons including, the health of a cell and human error. Using a cooks distance threshold of 4 divided by the number of observations in the data, we can next look to remove observations and reassess linear assumptions if any are violated.

BP Test: p-value >> 5% significance level this suggests that there is no evidence against the **equal variance assumption** for this model.

Shapiro Test: p-value >> 5% significance level this suggests that there is no evidence against the **normality assumption** for this model.

Linear Model | Remove Influential Observations






```
##
##  studentized Breusch-Pagan test
##
## data:  caret_lm_inf$finalModel
## BP = 10.311, df = 9, p-value = 0.3259
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(caret_lm_inf$finalModel)
## W = 0.98328, p-value = 0.5487
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14022 -0.19428 -0.02185  0.28605  1.04239
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.27407    0.15479  -1.771  0.08237 .
## lcavol       0.55693    0.10256   5.430 1.43e-06 ***
## lweight      0.16022    0.09034   1.774  0.08189 .
## age         -0.14399    0.07697  -1.871  0.06691 .
## lbph         0.21159    0.07972   2.654  0.01047 *
## svi1         0.73891    0.25300   2.921  0.00512 **
## lcp         -0.27279    0.13678  -1.994  0.05127 .
## gleason7     0.18483    0.20330   0.909  0.36737
## gleason8          NA         NA      NA       NA
```

```
## gleason9     0.18610    0.44714   0.416  0.67895
## pgg45        0.23607    0.11764   2.007  0.04990 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4949 on 53 degrees of freedom
## Multiple R-squared:  0.7304, Adjusted R-squared:  0.6846
## F-statistic: 15.95 on 9 and 53 DF,  p-value: 3.13e-12
```

Linearity - Inspecting the plot "Residuals Vs Fitted" has a trend line that helps illustrate that there is a no clear tend between the fitted values and residuals. Furthermore, the residuals generally exhibit zero mean suggesting that a linear model may be appropriate.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see there is generally constant variance and no obvious trends suggesting a linear model may be appropriate. A BP test can confirm if this assumption is not violated.

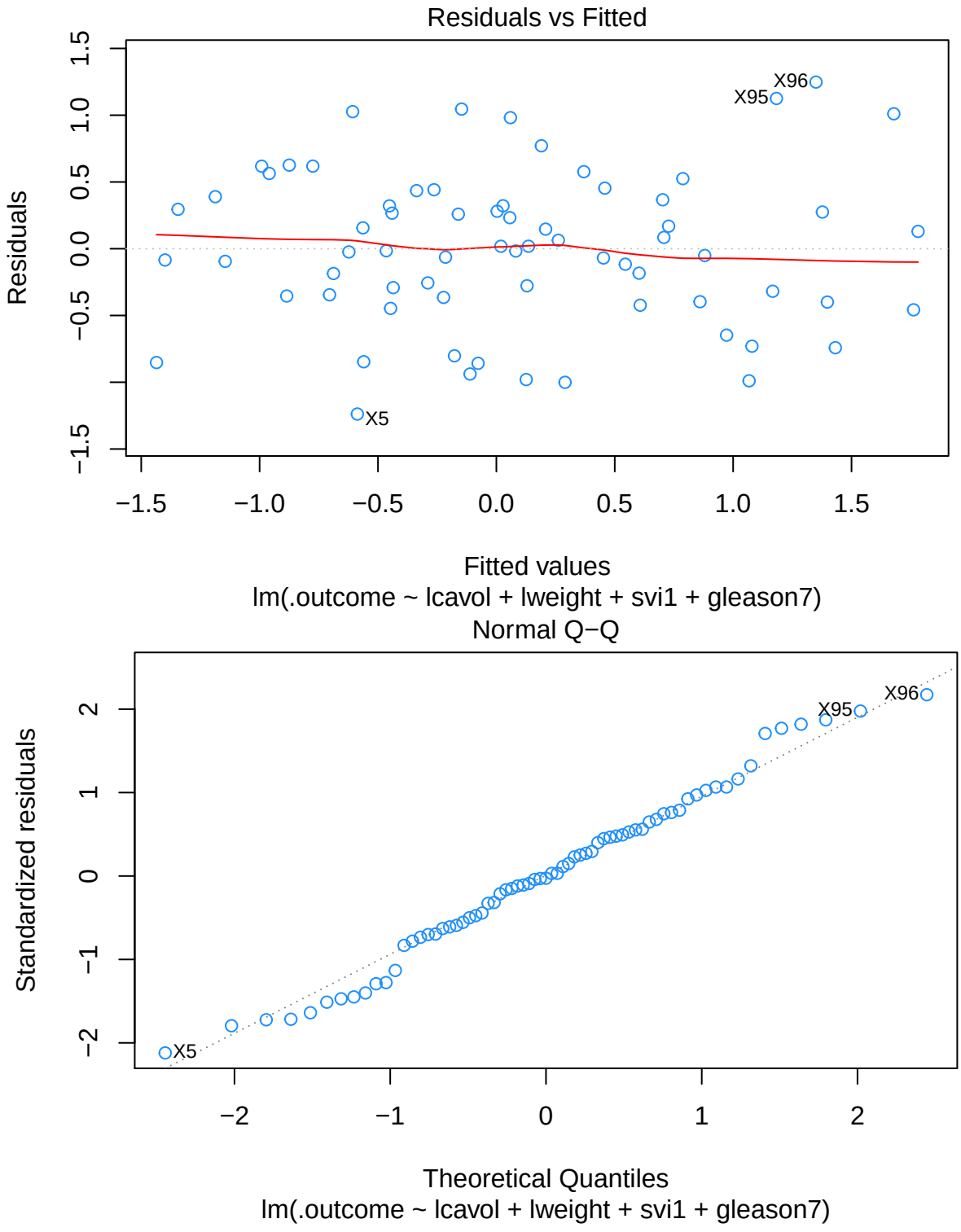
Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals moderately correspond to the theoretical quantiles of a normal distribution. At the extremes of the plot, the observed and theoretical quantiles deviate though additional testing via Shapiro test can indicate if the normality assumption is violated.

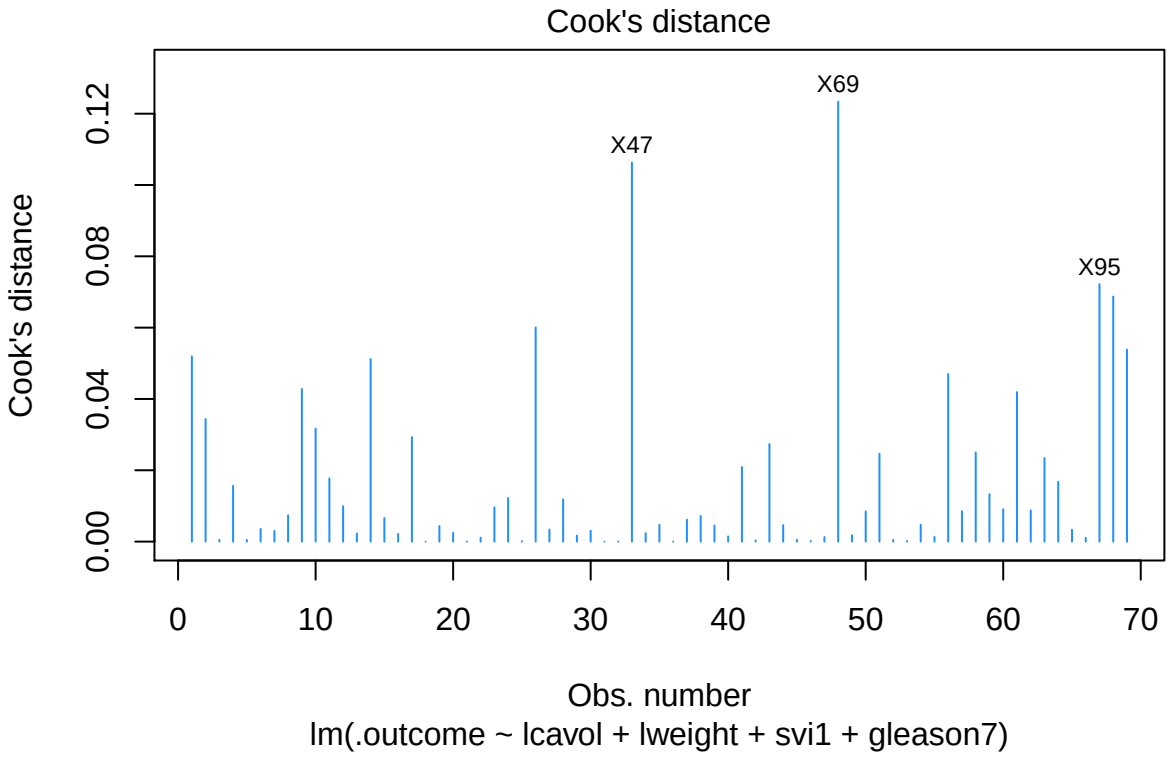
Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation. When dealing with biological experiments there can be anomolies in the observations which arise for a number of reasons including, the health of a cell and human error.

BP Test: p-value $>$ 5% significance level this suggests that there is marginal evidence against the **equal variance assumption** for this model. Note that at a 5% significance level we fail to reject the null hypothesis, if it were to increase to 10% there would be evidence against the assumption.

Shapiro Test: p-value $>>$ 5% significance level this suggests that there is no evidence against the **normality assumption* for this model.

Stepwise Linear Model






```
##
##  studentized Breusch-Pagan test
##
## data:  caret_steplm$finalModel
## BP = 6.7702, df = 4, p-value = 0.1485
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(caret_steplm$finalModel)
## W = 0.98446, p-value = 0.5488
```

```
##
## Call:
## lm(formula = .outcome ~ lcavol + lweight + svi1 + gleason7, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23798 -0.36467 -0.01523  0.36653  1.24811
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.24583    0.12398  -1.983  0.05170 .
## lcavol       0.46908    0.09534   4.920 6.36e-06 ***
## lweight      0.23247    0.09051   2.568  0.01256 *
## svi1         0.60451    0.21285   2.840  0.00604 **
## gleason7     0.32468    0.16584   1.958  0.05462 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5949 on 64 degrees of freedom
```

```
## Multiple R-squared:  0.6633, Adjusted R-squared:  0.6422
## F-statistic: 31.51 on 4 and 64 DF,  p-value: 1.662e-14
```

Linearity - Inspecting the plot "Residuals Vs Fitted" has a trend line that helps illustrate that there is a no clear tend between the fitted values and residuals. Furthermore, the residuals generally exhibit zero mean suggesting that a linear model may be appropriate.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see there is generally constant variance and no obvious trends suggesting a linear model may be appropriate. A BP test can confirm if this assumption is not violated.

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals moderately correspond to the theoretical quantiles of a normal distribution. At the extremes of the plot, the observed and theoretical quantiles deviate though additional testing via Shapiro test can indicate if the normality assumption is violated.

Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation. When dealing with biological experiments there can be anomolies in the observations which arise for a number of reasons including, the health of a cell and human error.

BP Test: p-value >> 5% significance level this suggests that there is no evidence against the **equal variance assumption** for this model.

Shapiro Test: p-value >> 5% significance level this suggests that there is no evidence against the **normality assumption* for this model.

Significance of Model Predictors

```
## Analysis of Variance Table
##
## Model 1: .outcome ~ lcavol + lweight + svi1 + gleason7
## Model 2: .outcome ~ lcavol + lweight + age + lbph + svi1 + lcp + gleason7 +
##     gleason8 + gleason9 + pgg45
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     64 22.652
## 2     58 21.265  6    1.3865 0.6303 0.7054
```

Considering the following null and alternative hypothesis and a 5% significance level

Null : age = lbph = lcp = gleason8 = gleason9 = ppg45

Alt : At least one of age,lbph,lcp,gleason8,gleason9,ppg45 is non-zero.

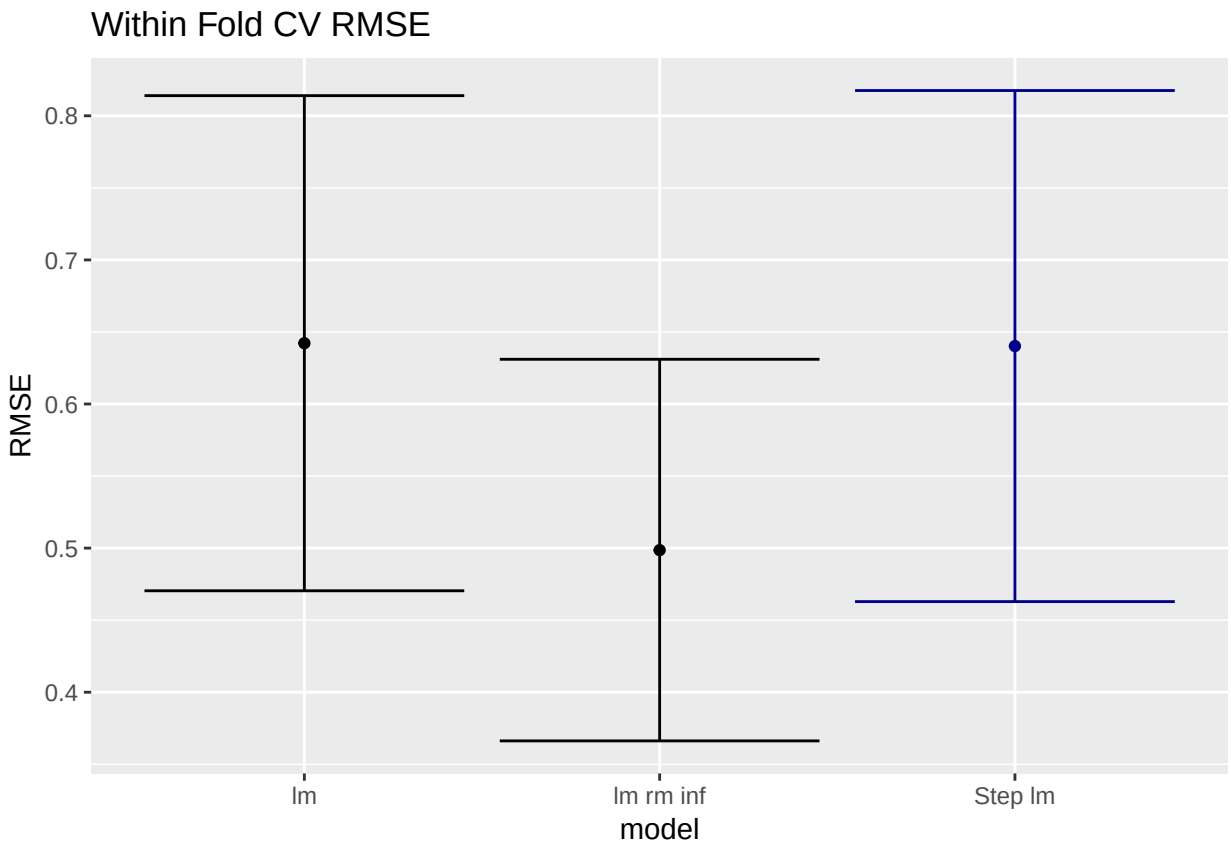
Using anova to assess the importance of the terms we fail to reject the null hyptothesis.

Rather than just considering the reduced stepwise model as a candidate model, all 3 models i.e lm, lm with influential values removed, and stepwise lm are considered and will be used to determine an estimate of the generalization error using repeated cross validation. From there, the model that optimizes the trade-off between predictive power and low variance will be the selected model.

Table 1: Model Selection

Model	RMSE
Lm Remove Inf	0.4986295
Stepwise Lm	0.6402152
Lm	**0.6421949**

Model Selection




```
##
## Call:
## lm(formula = .outcome ~ lcavol + lweight + svi1 + gleason7, data = dat)
##
## Coefficients:
## (Intercept)       lcavol      lweight         svi1     gleason7
##     -0.2458       0.4691       0.2325       0.6045       0.3247
```

On the training/validation set, the estimate of the generalization error can be seen in the table **Model Selection**. While the first model i.e lm with influential points removed had the lowest RMSE, it was close to violating the equal variance assumption and therefore the most appropriate model is either the full linear model or Stepwise model.

To assist in choosing between these two remaining models, the plot **Within Fold CV RMSE** illustrates a plot with error bars (1 standard deviation) for candidate models with respect to the RMSE across all folds. Notice that the standard deviation for the Stepwise Linear model is lower and it is less complex, therefore the most appropriate model is the stepwise linear model.

$$Selected : \hat{lpsa} = \beta_0 + \hat{\beta}_{lcavol}x_{i,1} + \hat{\beta}_{lweight}x_{i,2} + \hat{\beta}_{svi1}x_{i,3} + \hat{\beta}_{gleason7}x_{i,4}$$

The predictors relied upon in the model are rather intuitive i.e it makes perfect sense that predictors for `lpsa` are all measures of cancer size, activity of cancer and risk measure from biopsy. It seems as though the presence of seminal vesicle invasion contributes signifcantly to the 'lpsa' as does the cancer volume.

Performance Evaluation

To generate an unbiased estimate of the generalization error, the selected Stepwise function can be applied to the held out test set. The table **Generalization Error** illustrates the unbiased estimate of the generalization error.

Table 2: Generalization Error

Metric	Performance
RMSE	0.6692367