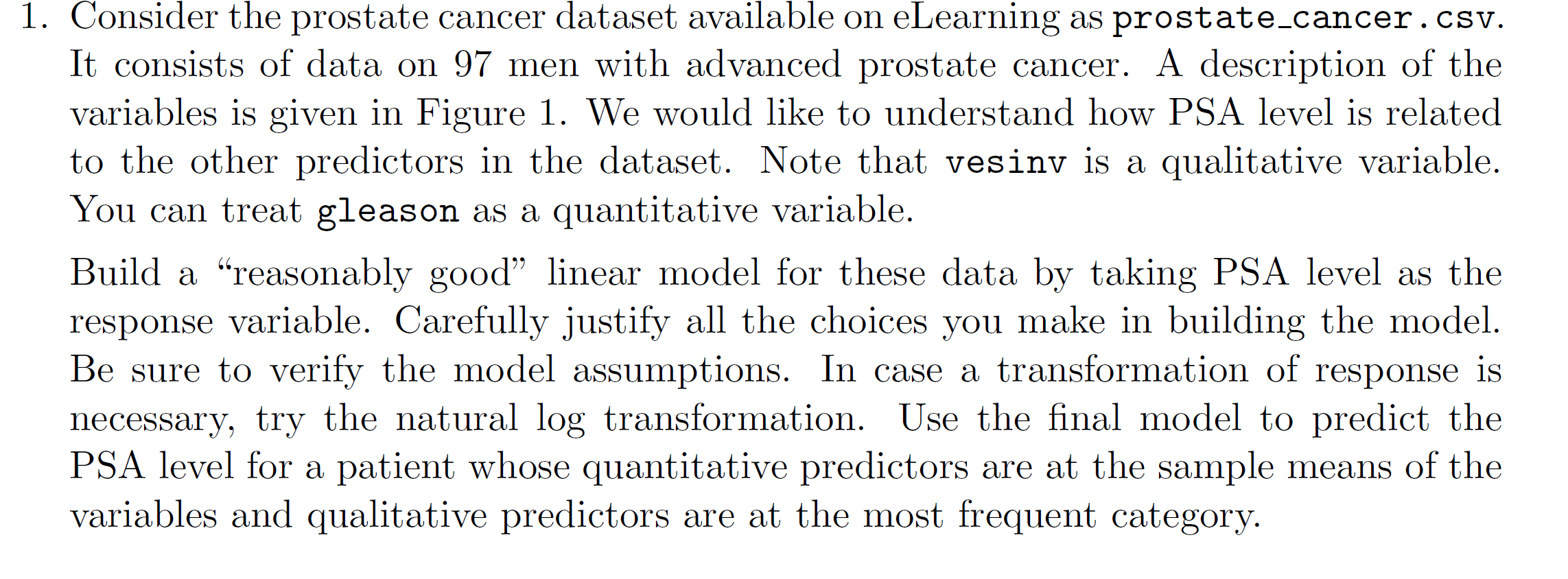
**STATISTICAL METHODS FOR DATA SCIENCE CS 6313.001 FALL 2019**

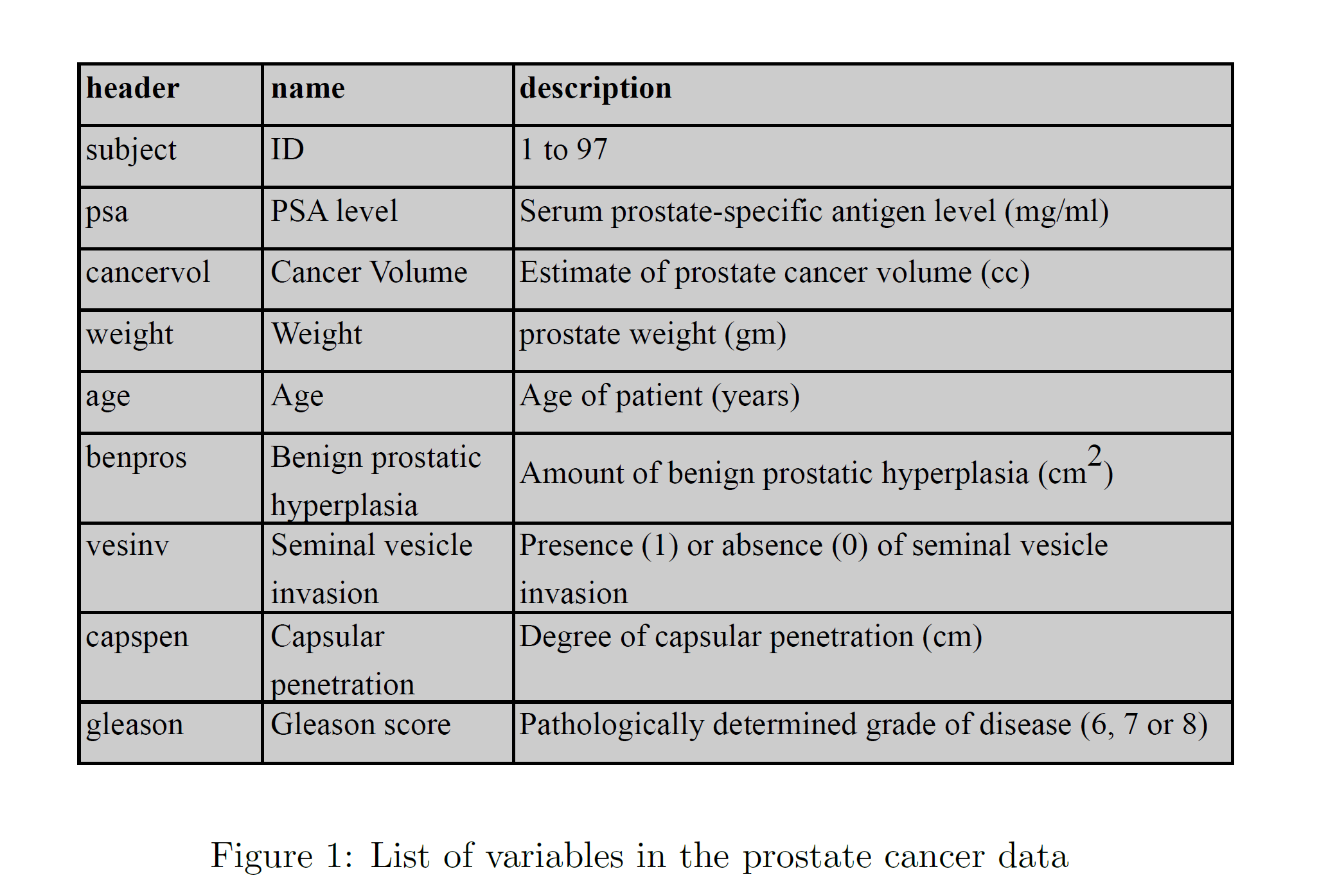
**Mini Project #6**

**Participants :**

**Reetish Chand Guntakal Patil (RXG190006)**

Question 1:

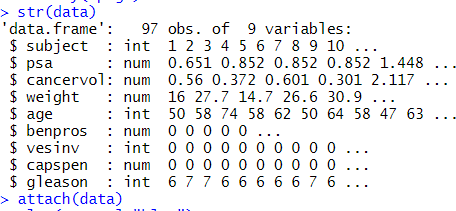




Reading the data:

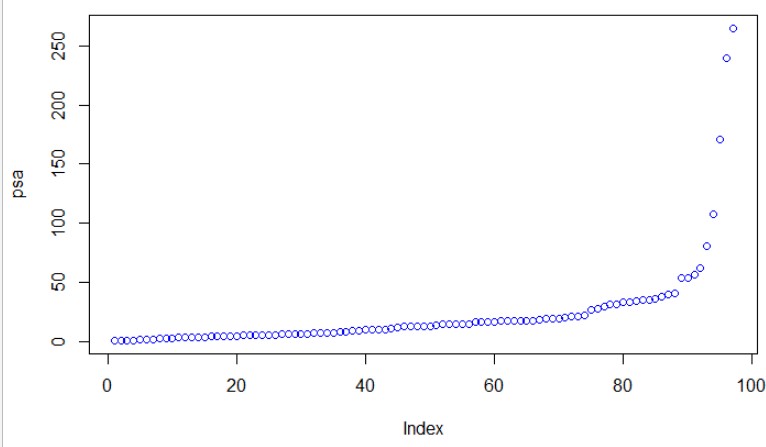
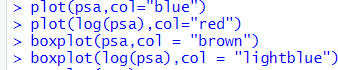
data <- read.csv("prostate\_cancer.csv")

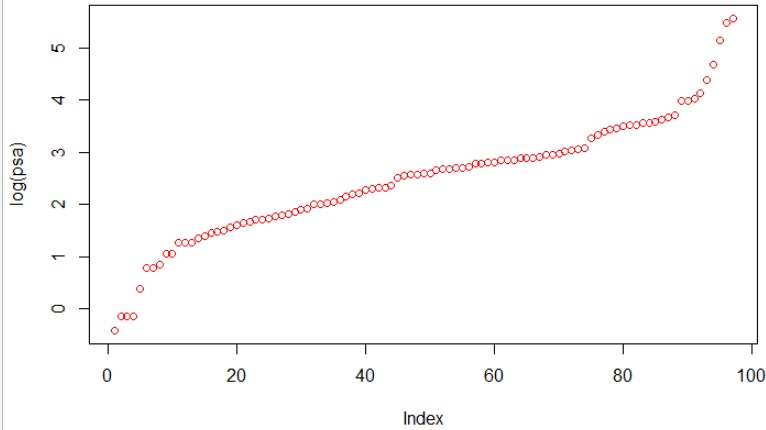
Attaching the data so that we can use the variables directly in the linear model

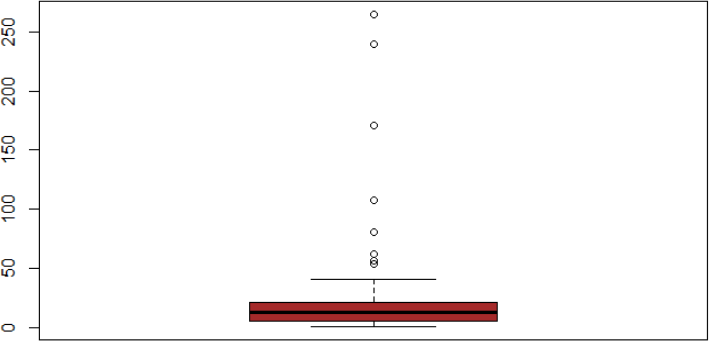


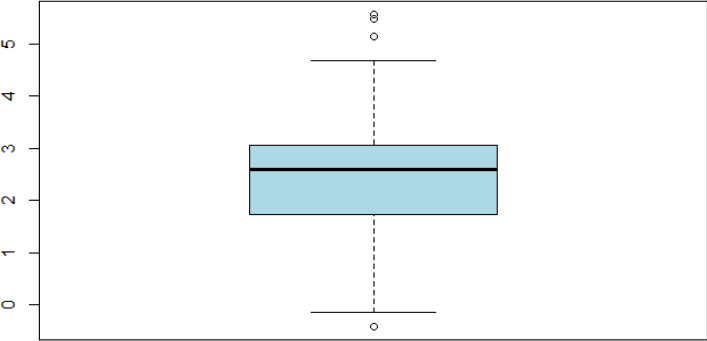
To ensure the distribution must be skewed to make it better fit our linear model condition, log transformation of the psa variable is done.

Performing the log transformation & plotting the boxplot of the psa variable, a lot of outliers can be observed which indicates transformation is needed to fit the linear model.

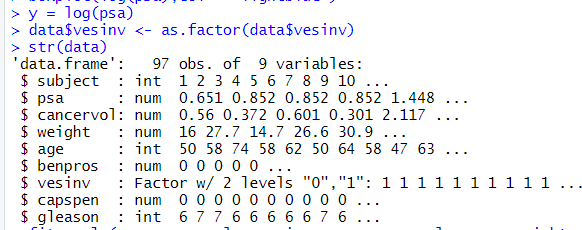








Converting the int type vesinv to factor type before giving it to the linear model in order to make vesniv a nominal variable

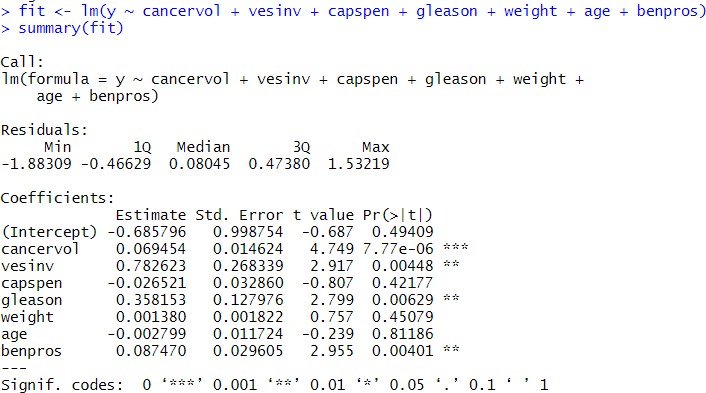


Fit the linear model 1:

Full model

Null Hypothesis Ho: None of the predictors are useful for predicting response.

Alternate Hypothesis HA: At least one of the predictors is useful for predicting the response.



Observation:

* cancervol , vesinv, gleason, benpros are the significant predictors.
* Hence, we reject the null hypothesis.

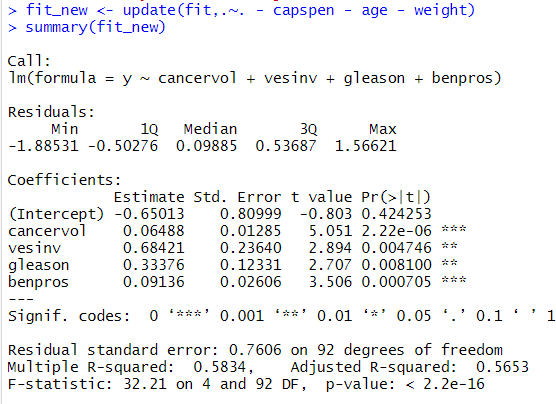
Fit the linear model 2:

Reduced Model

Null Hypothesis Ho: None of the predictors are useful for predicting response.

Alternate Hypothesis HA: At least one of the predictors is useful for predicting the response.

Note: reduced model is equivalent to the full model except that the extra predictors of the full model are unnecessary.



Observation:

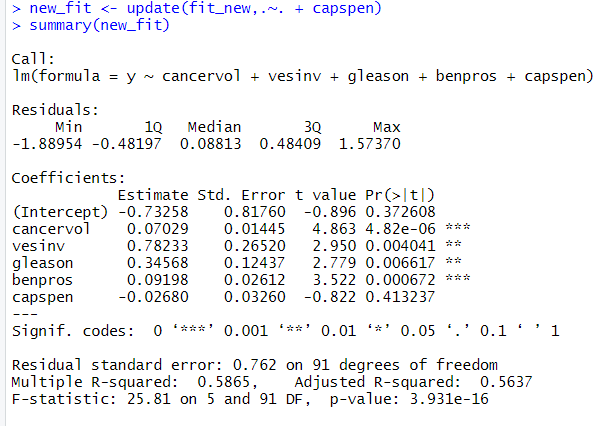
* All the predictors are the significant predictors. Hence, we reject the null hypothesis.
* The value of the adjusted R-square also increases validating the better correctness of the linear model.

Fit the linear model 3:

Null Hypothesis Ho: None of the predictors are useful for predicting response.

Alternate Hypothesis HA: At least one of the predictors is useful for predicting the response.

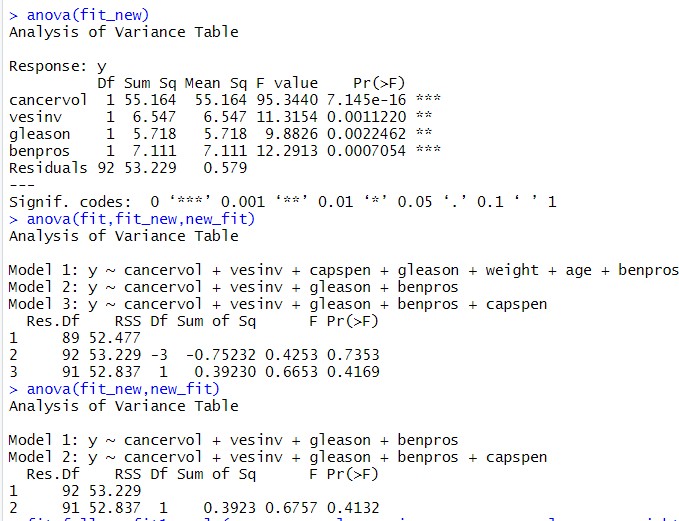
Note: Capspen is added to the linear model as it was coming out to be an important variable in the correlation plot.



Observation:

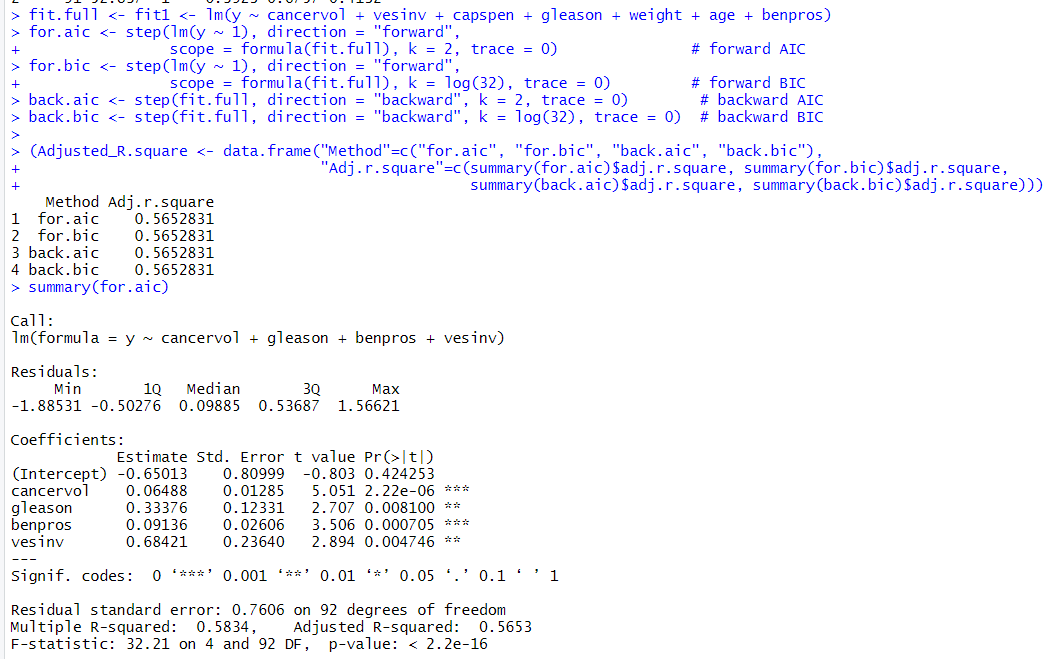
* cancervol , vesinv, gleason, benpros are the significant predictors. Hence, we reject the null hypothesis.
* The adjusted R-squared value decreases telling that capspen is not an optimal predictor for predicting the response variable.

Performing the ANOVA test:

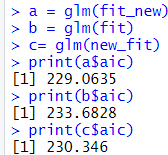


As, it’s not a good practise to infer any results from 3 model anova comparison we are

calculating the AIC score.



Comparing the AIC scores of the three fitted models



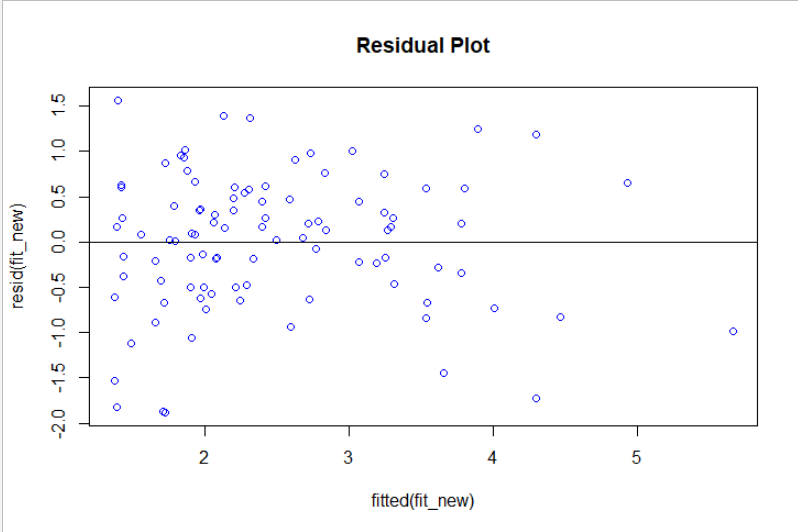
Observation:

* a or fit\_new linear model has the lowest AIC score suggesting that it’s the best

model among all the models Conclusion:

* fit\_new model is the best model

Model evaluation to find if the fitted model 2 is the good representation of the linear model



Observation:

* The mean is zero
* Not much change in the vertical spread i.e the standard deviation is constant

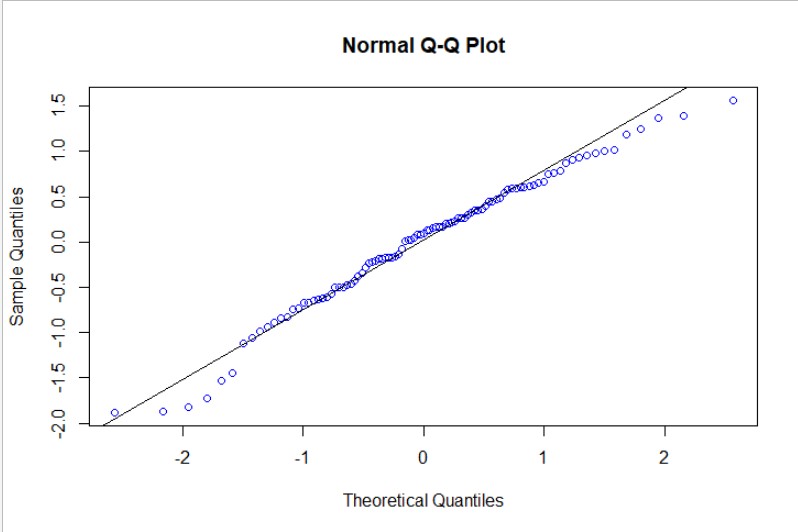
Conclusion:

* The linear model is a good estimate QQ-plot:

Assumption: the Residual error is independent and identically distributed coming from a

normal distribution with mean of 0 and standard deviation of sigma-squared.



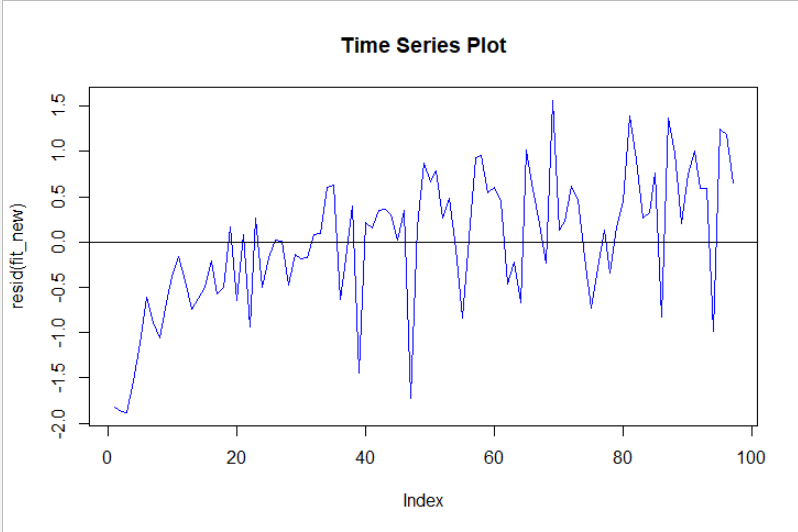


Observation:

* that the data is almost normally distributed. Conclusion:
* The assumption is valid

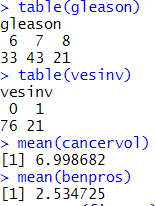
Time-series plot:





Using 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 are at the most frequent category.

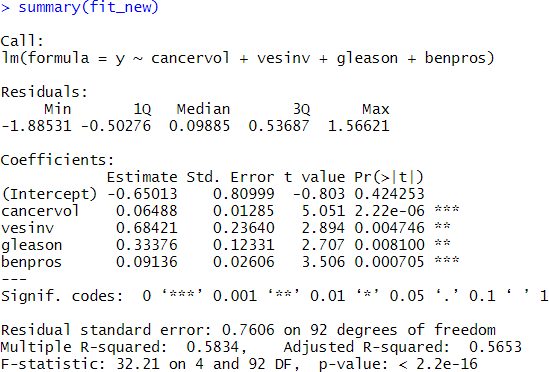
lm(formula = y ~ cancervol + vesinv + gleason + benpros)



Observation:

* gleason value 7 is being dominated in the data
* vesinv value 0 is being dominated in the data
* the mean of cancervol and benpros are 6.998 and 2.534 respectively

Predicting values with the best linear model:



predicted value is equal to:

-0.65013 + 6.998682\*(0.06488) + 7\*(0.33376) + 0.09136\*(2.534725) = 2.371837

Thus, the actual value of PSA is exp(2.371837) which is equal to 10.71706