**CS-6313 – Statistical Methods for Data Science**

**Mini Project #6**

**Group No - 5**

**Manan Dalal (MUD200000)**

**Lipi Patel (LDP210000)**

**Contribution of Team Members:**

We both collaborated and solved the questions together for a thorough understanding of functions in R and solved the question concurrently to check for accuracy, debugging and application.

**Question 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.**

**Build a “reasonably good” linear model for these data by taking PSA level as the response variable. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions. In case a transformation of response is necessary, try the natural log transformation. 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 are at the most frequent category.**

**Table

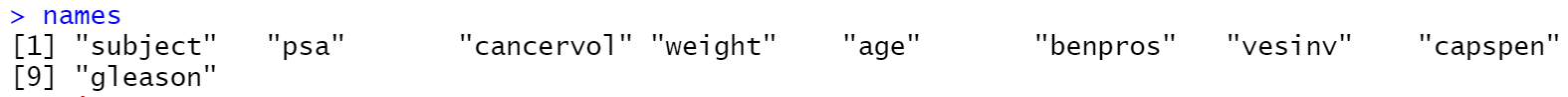
Description automatically generated**

**Solution**

* First, we read the csv file containing the prostate cancer data.
* Then, we study the different columns of the data and summarize the data.
* Then, we analyze the correlation between attributes.
* We then create a boxplot to analyze the PSA attribute.
* We also transform the PSA attribute using natural logarithm function.
* Then, we check the distribution of age and PSA column by plotting a graph.

Graphical user interface, text, application, email

Description automatically generated



A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Chart

Description automatically generated with medium confidence Chart, scatter chart

Description automatically generated

* Now, we will compare all the predictors with PSA to find out how they are related.
* To do this, we will build univariate models for all the predictors.
* Following are the models we built for all the various predictors.

Predictor: ***subject***

Text

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence Chart, histogram

Description automatically generated

Predictor:  ***cacervol***

Text

Description automatically generated

Text, letter

Description automatically generated Chart, scatter chart

Description automatically generated

Predictor: **weight**

Text

Description automatically generated with medium confidence

Text, letter

Description automatically generated Chart, scatter chart

Description automatically generated

Predictor: ***age***

Text

Description automatically generated with medium confidence

Text, letter

Description automatically generated Chart, scatter chart

Description automatically generated

Predictor: ***benpros***

Text

Description automatically generated with medium confidence

Text

Description automatically generated Chart, scatter chart

Description automatically generated

Predictor: ***vesinv***

Text

Description automatically generated

Text, letter

Description automatically generated Chart, box and whisker chart

Description automatically generated

Predictor: **capspen**

Text

Description automatically generated

Text

Description automatically generated Chart, scatter chart

Description automatically generated

Predictor: **gleason**

Text

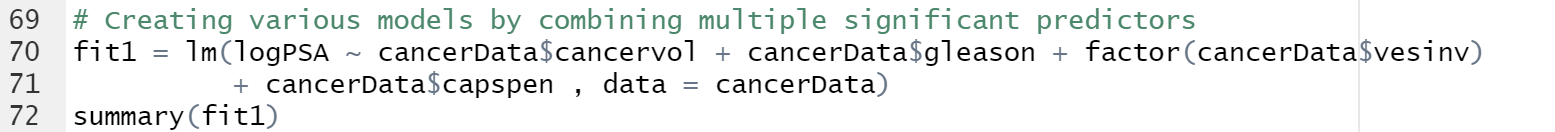
Description automatically generated

Text, letter

Description automatically generated Chart

Description automatically generated

* Upon Observing the models above, we can see that cancervol, gleason, vesinv, benepros and capspan are significant.
* These predictors show an evident linear relation with PSA.
* We will now use various combinations of these predictors to predict our PSA.



Text

Description automatically generated

* Then, we remove capspen from the model and compare the 2 models to check if capspen is significant or not.

Graphical user interface, text, application

Description automatically generated

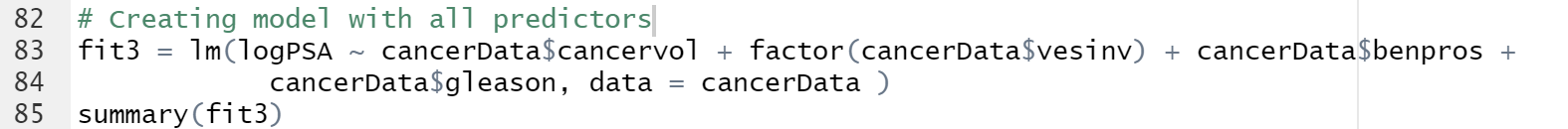
Text

Description automatically generated with low confidence

Text

Description automatically generated

* By looking at the above output we can say that capspen is not a significant predictor.
* Thus, in the next model we add all other predictors and ignore capspen.



Text

Description automatically generated

* From the above summary we can observe minimum residual standard error, thus so far this is the best model.
* We will now create a residual and a QQ plot for this model.

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated Chart, histogram

Description automatically generated

**Our Assumptions:**

• Errors are centered around zero with constant variance, this can be seen from the residual chart thus this is verified.

• Errors are also normally distributed as seen by QQ plot where QQ line fits very well. Thus, this assumption is also verified well.

* Now, we shall use this final model to predict the PSA level of a patient.
* We will consider the sample means of all the quantitative predictors and the highest frequency count for the qualtitative predictor.

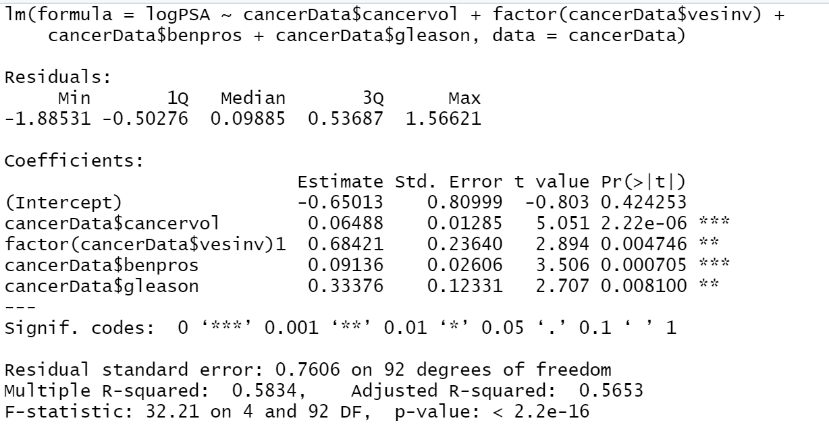
Text

Description automatically generated

Text

Description automatically generated

* From the below summary of fit3, we can see the values of beta0, beta1, beta2, beta3, beta4 are -0.65013, 0.06488, 0.68421, 0.09136, 0.33376 respectively.



* Now, we use these values to predict the PSA value.

A picture containing text

Description automatically generated



* Thus, the final value of PSA level of a patient comes out to be 10.28357 using our best fit model.