**­­­­­­­­­­**

**Statistical Methods for Data Science (Spring 2017)**

Mini Project 5

Contributing Members:

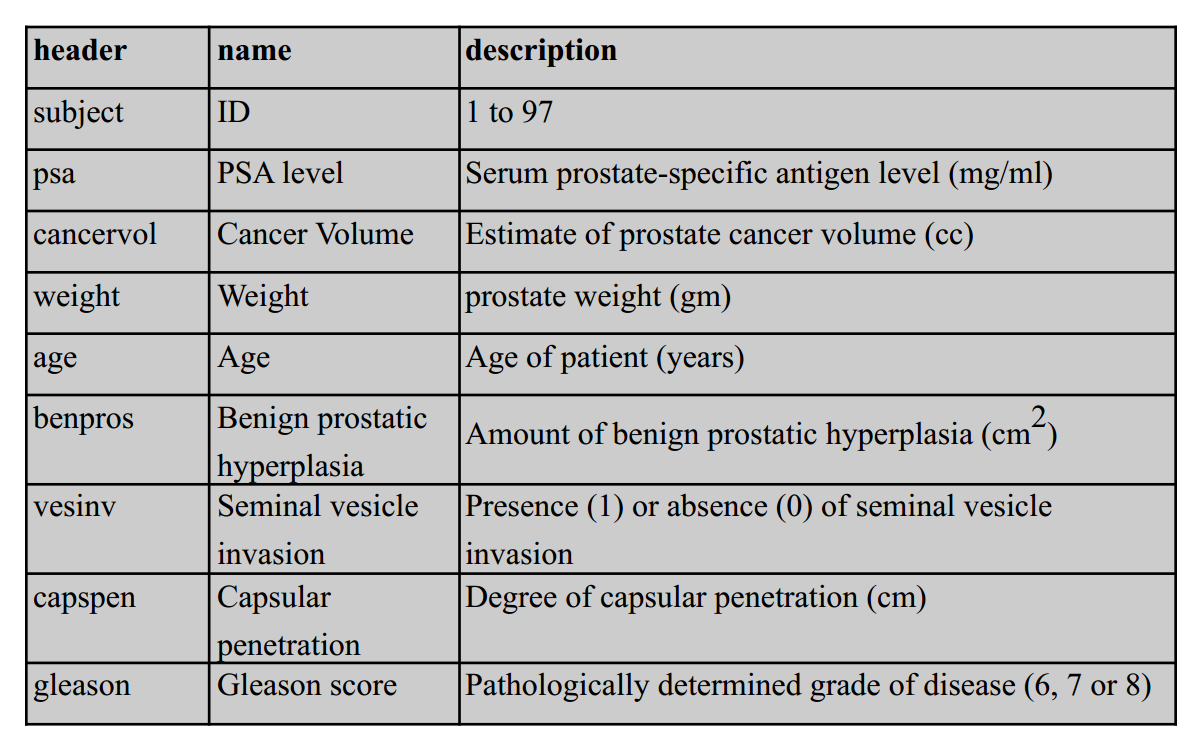
**Akhilesh Kumar Kagalvadi Chinnaswamy (axk167131)**

**Vidya Sri Mani (vxm163230)**

**Contribution of Group Members:**

We individually found out the model and later discussed different models and anova analysis and residual plots of the reports together and after reaching a consensus on our solution, we compiled the report together.

Consider the prostate cancer dataset available on eLearning. It consists of data on 97 men with advanced prostate cancer. Following is a description of the variables:

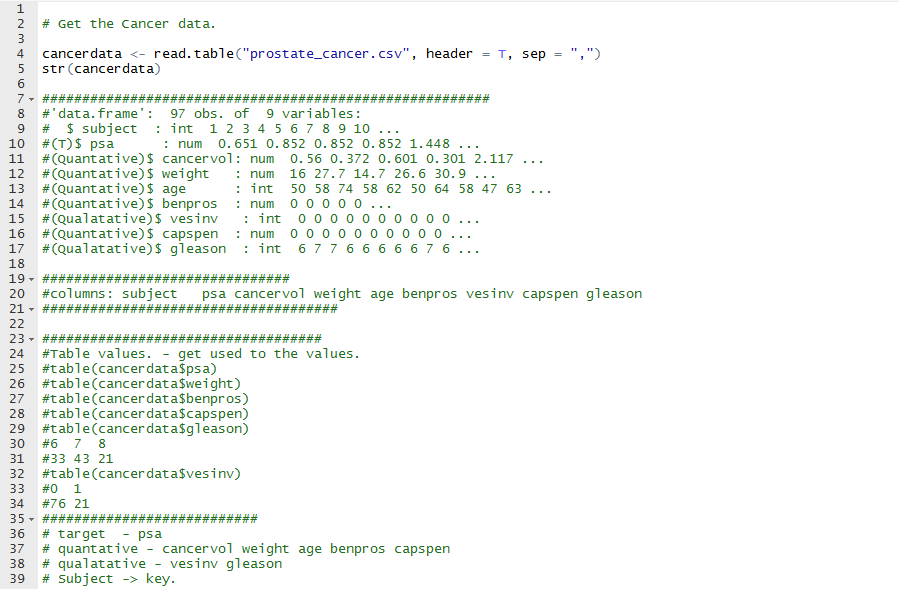
  
Build a “reasonably good” linear model for these data by taking PSA level as the response variable. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions and also to distinguish between quantitative and qualitative variables. Use the final model to predict the PSA level for a patient whose  
predictors are at the sample means of the variables.

Solution:  
The Work consists of following sections:  
1) The data (exploratory analysis)  
2) Linear Regression fitting with one parameter.  
 a) fitting quantitative data.  
 b) fitting categorical data like gleason, vesinv.

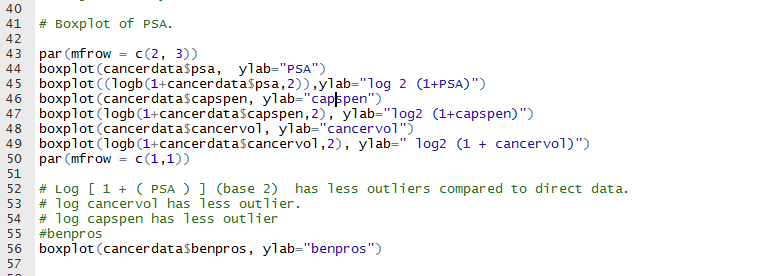
3) Multi Regression manual fitting.  
4) comparing the fitted model using residual, qq plot.  
5) Automatic fitting using AIC and comparing the results.  
6) predict the value from the final model at the mean of the predictors.­  
  
\* Note since the code is long, most of the comments/explanation are added to the R-code. R-code is 800 lines so it is not completed pasted in the appendix.

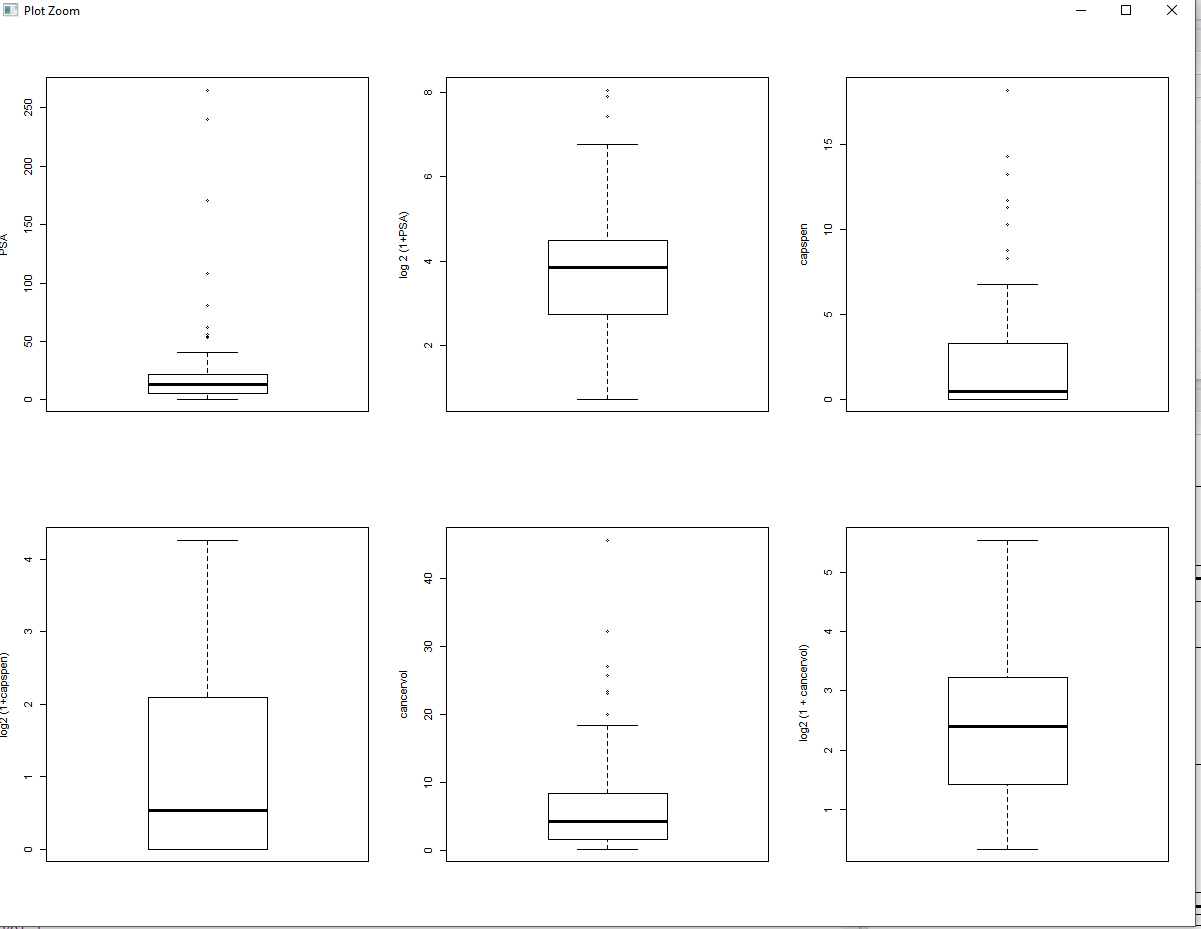
# 1) The data (exploratory analysis)

Getting used to the data.

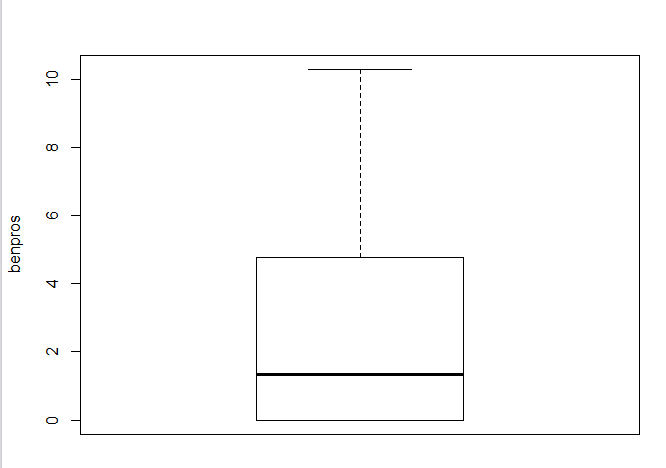


Box Plot Analysis of the Variables:  
 cancervol, PSA , Capsen and log Transformation to reduce the outliers, so that data is in the correct scale.



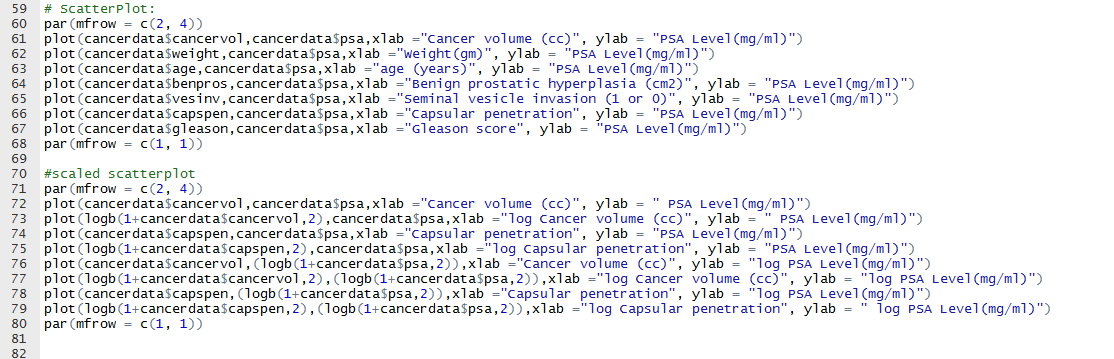


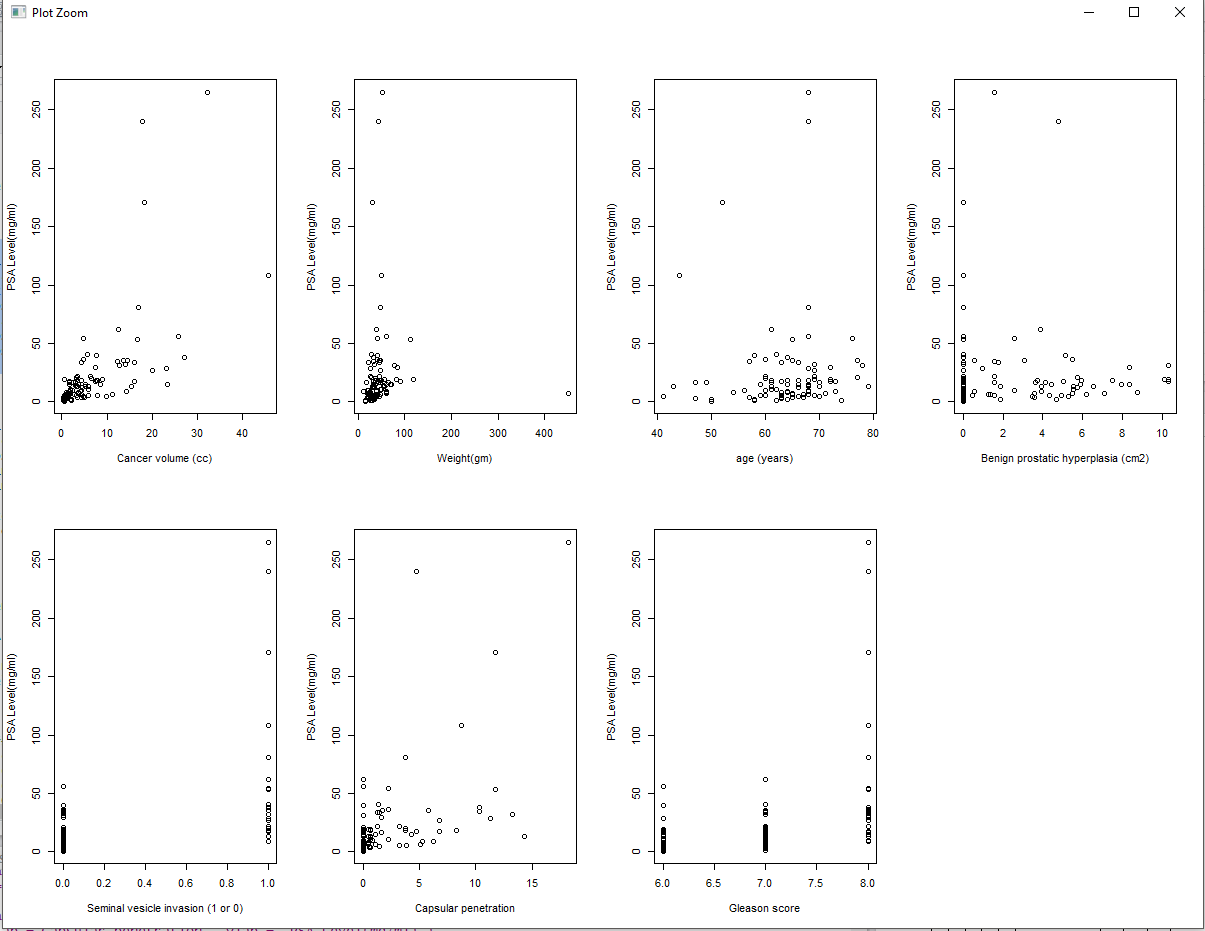
Boxplot for Cancervol, PSA, Capsen, log2(x+1) of three variables.

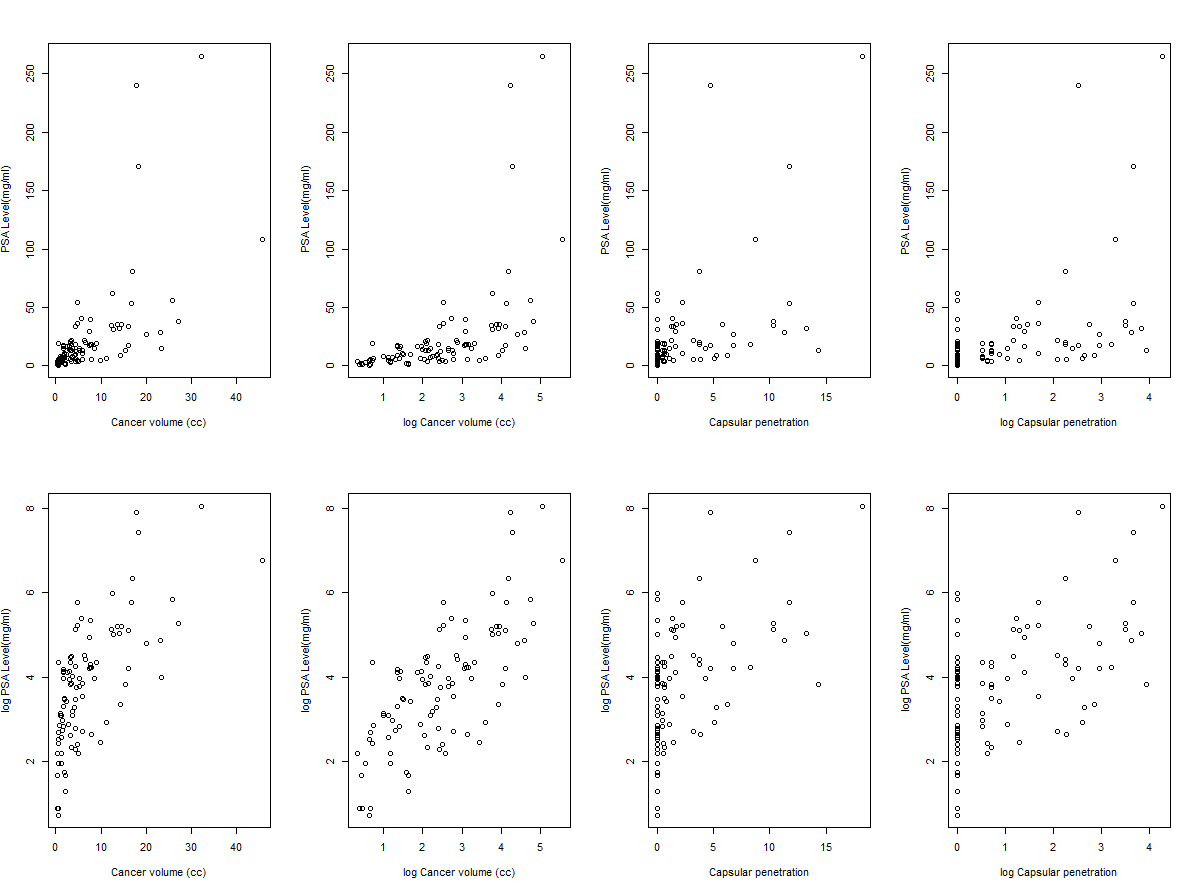


Boxplot for Benpros doesn’t need a log transform.

Scatter plot Analysis to see the data behavior:

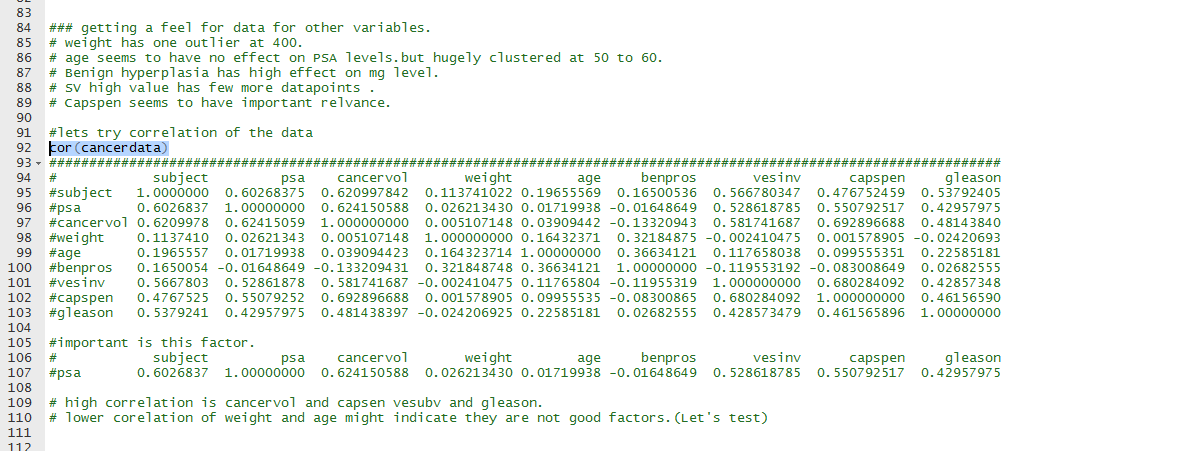




Scatter plot of all the variables.  


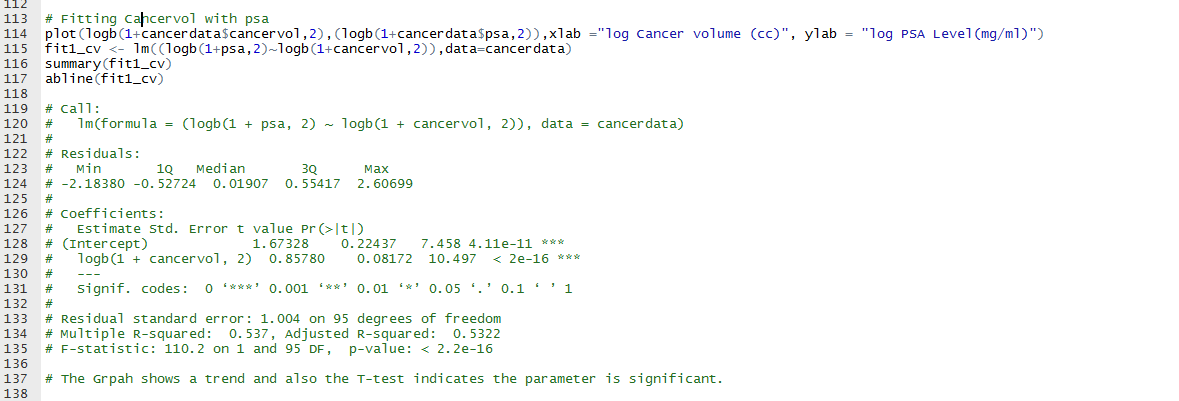
Scatter plot with transforms.

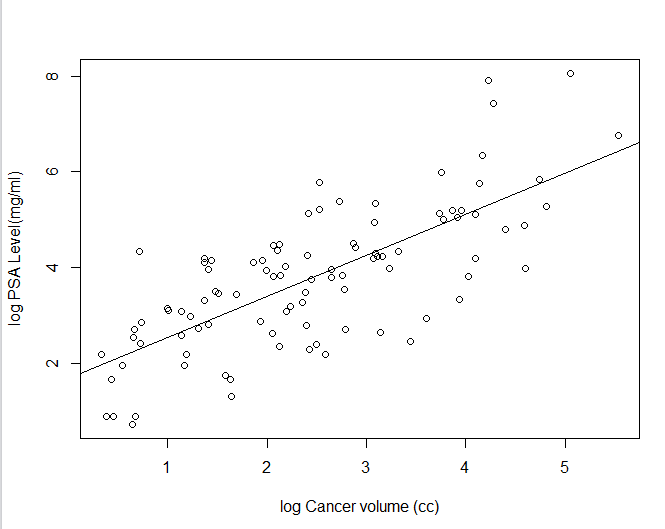
Correlation among the variables:



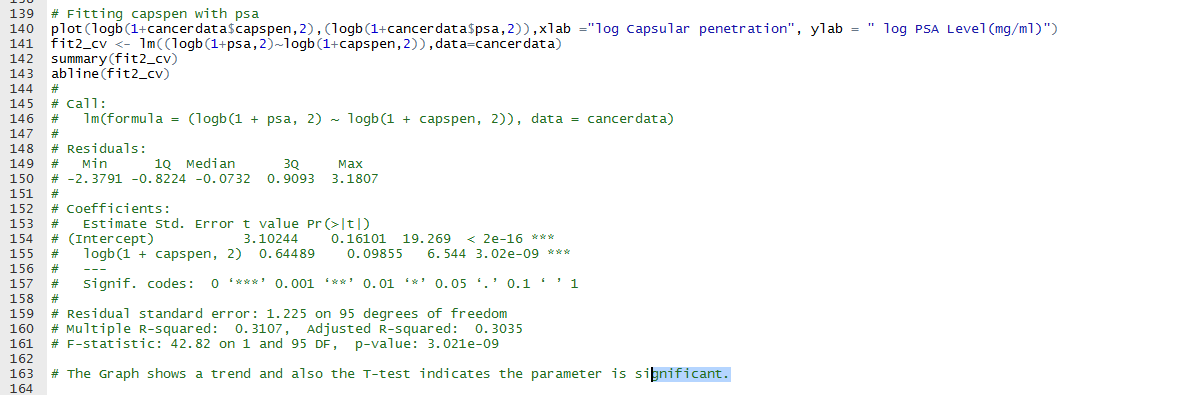
2) Linear Regression fitting with one parameter.

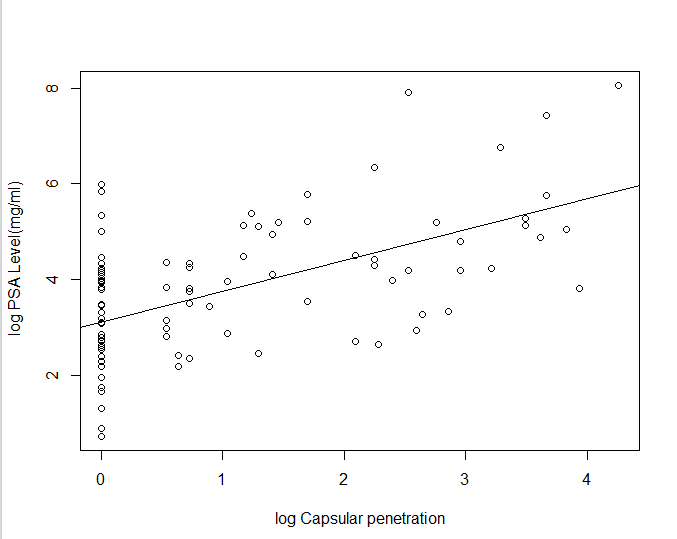
Fitting quantitative data.  
1. Linear fitting of the data: Cancer vol.



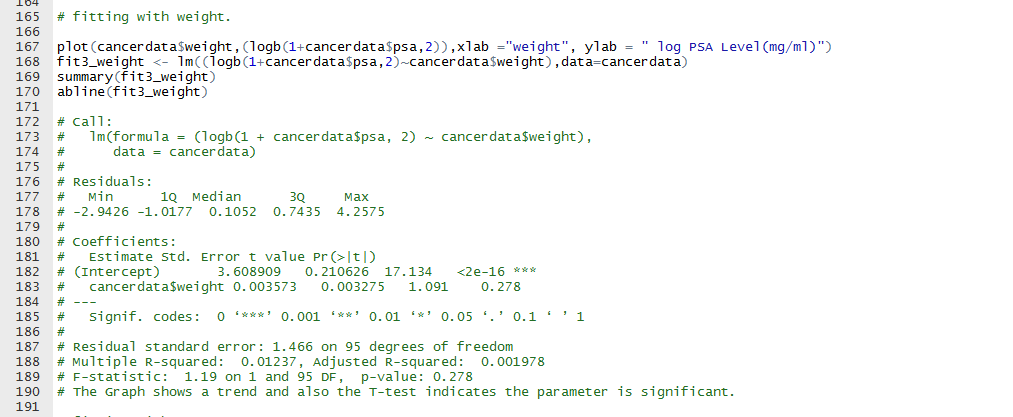


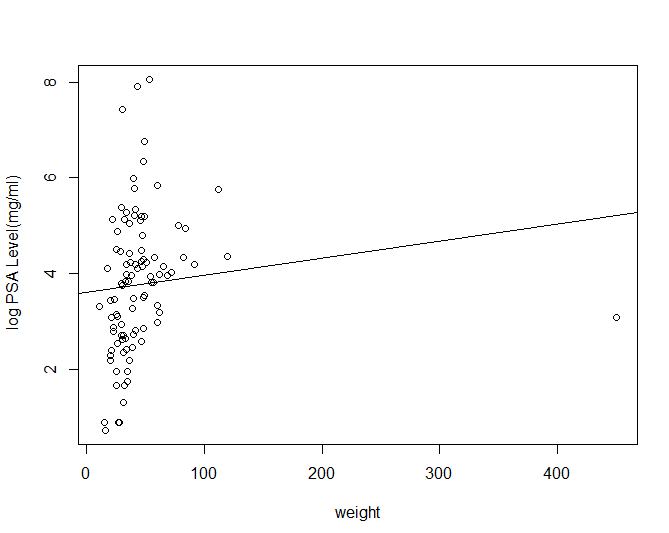
2.Linear fitting of the data: Capsular Penetration.



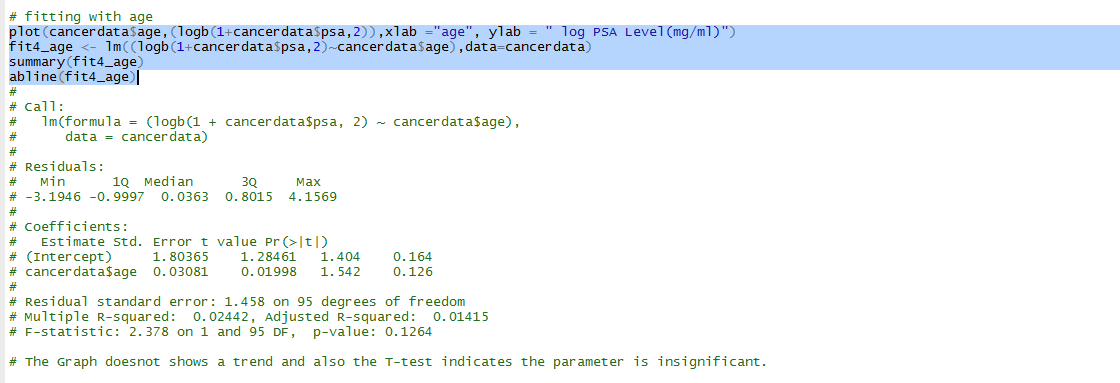


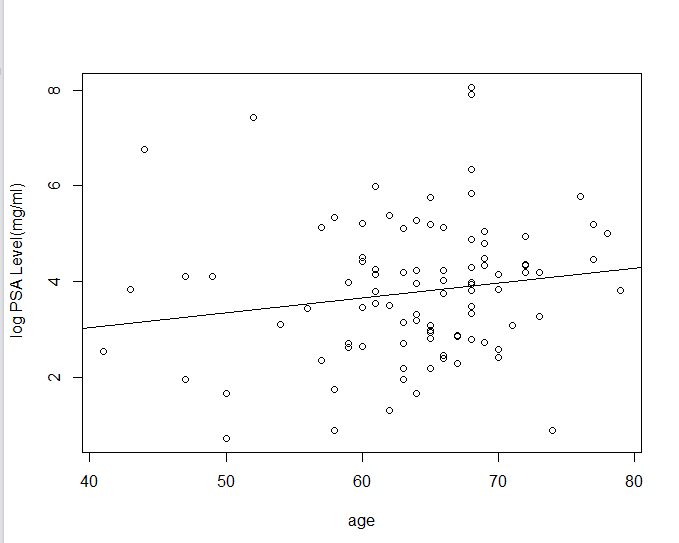
3.Linear regression for weight:



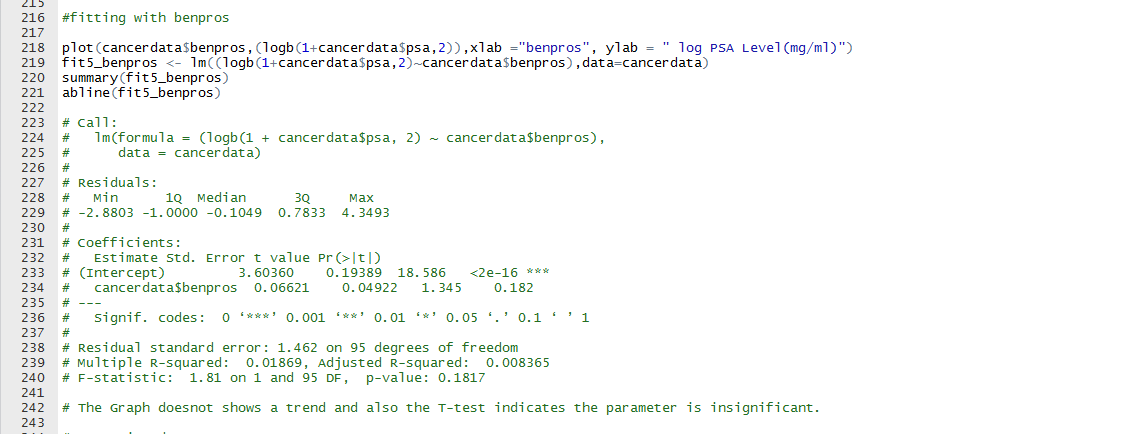


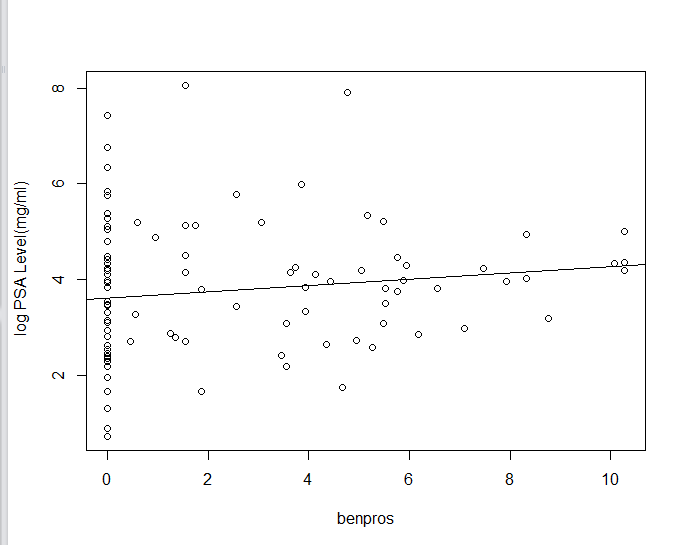
4. Linear Regression for age:



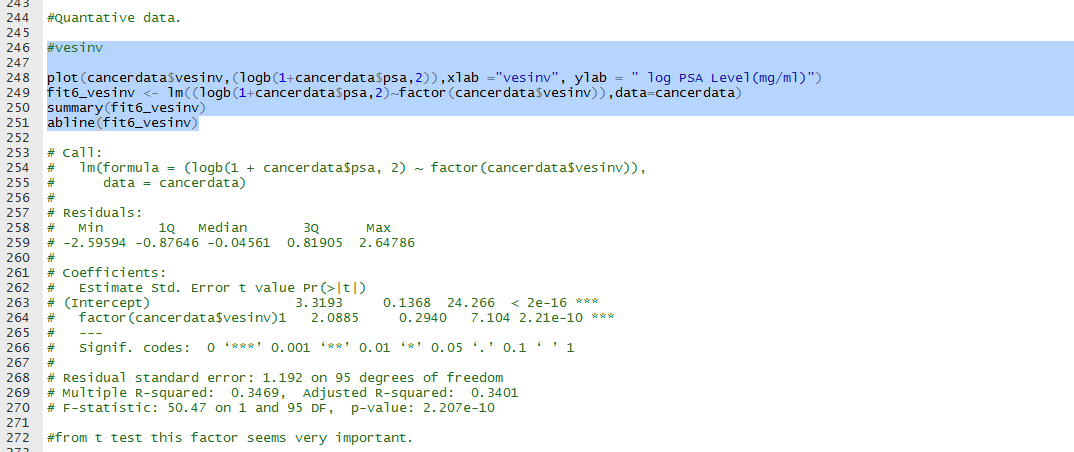


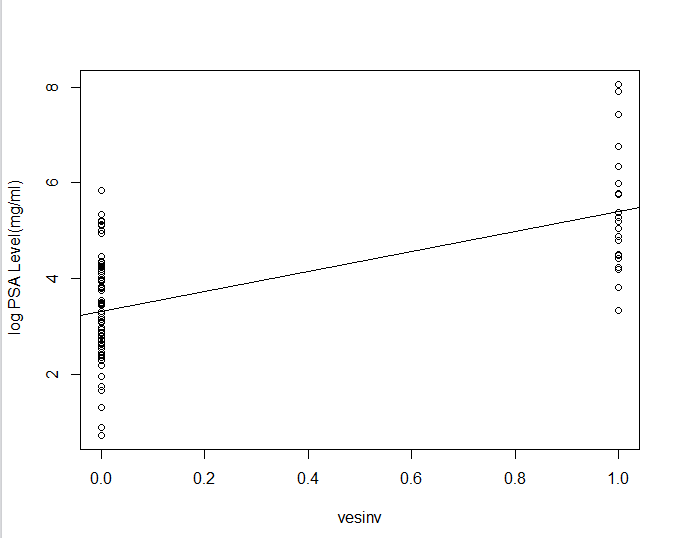
5.Linear regression with Benpros data:



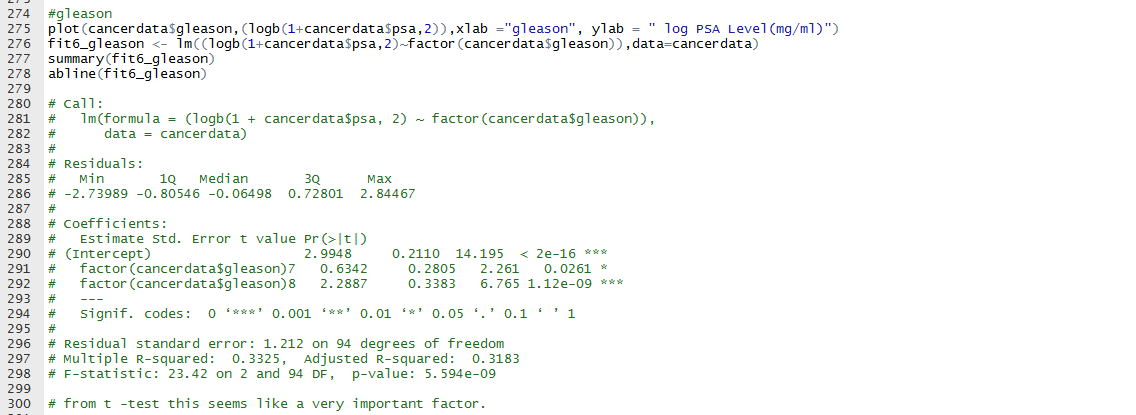


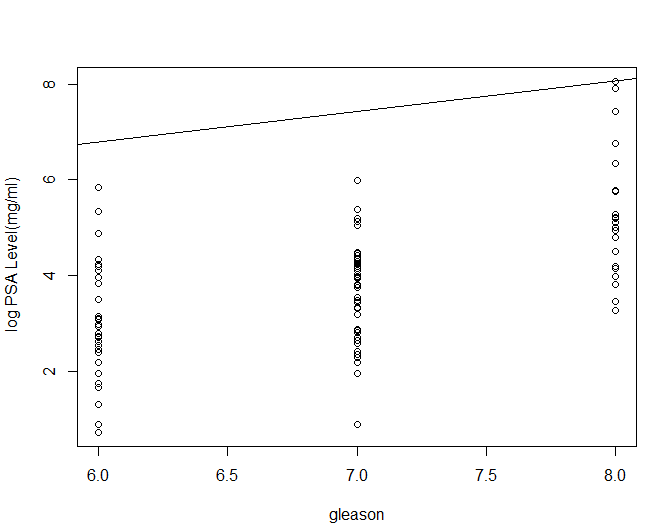
2b) Qualitative data (categorical data)  
vesinv:



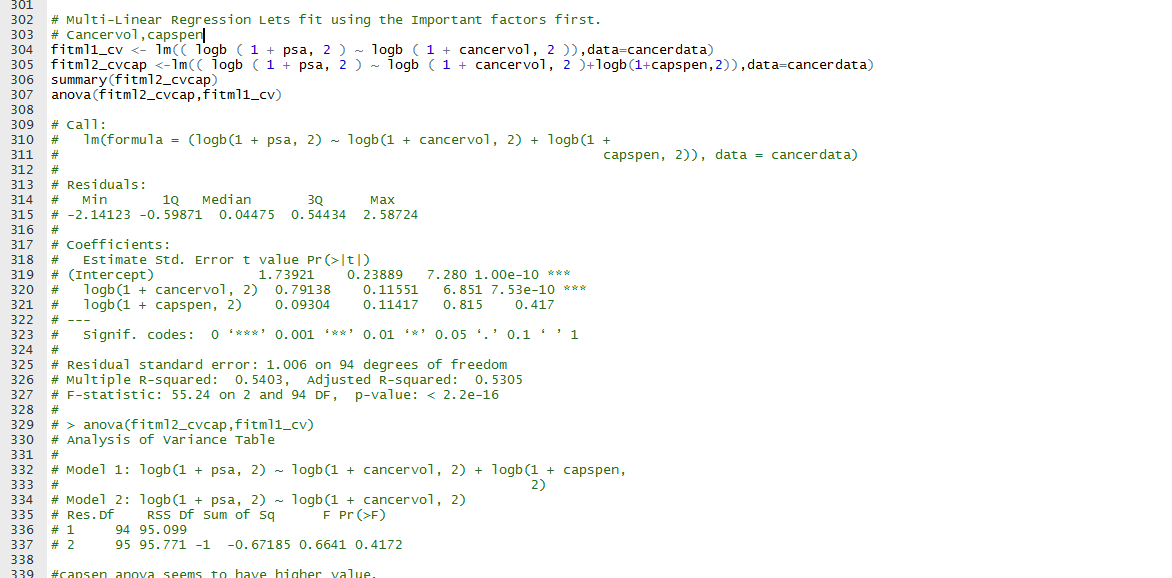


## Gleason:

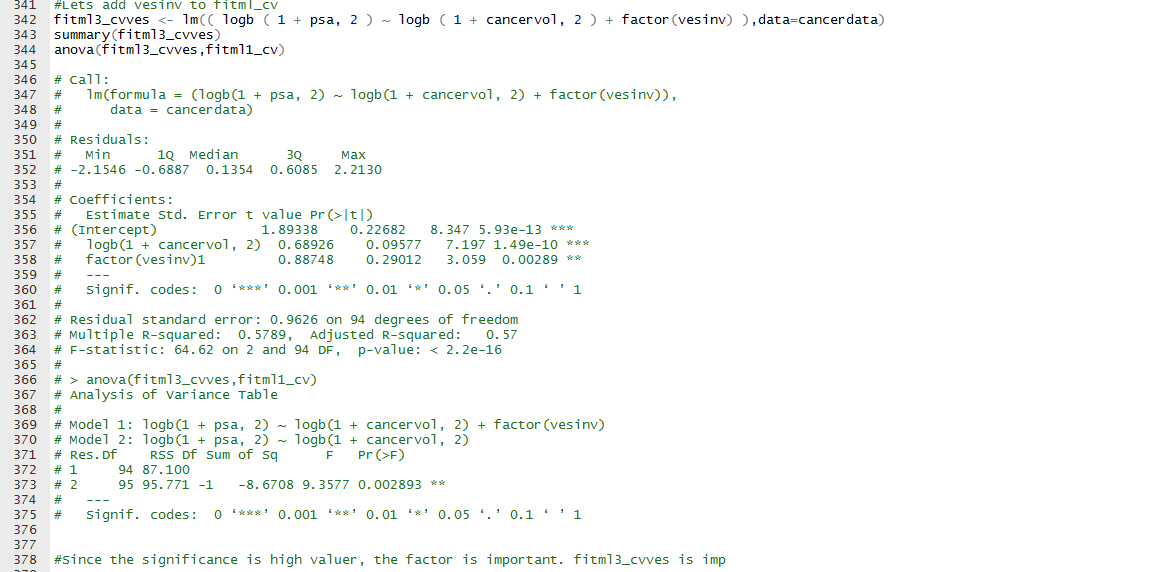




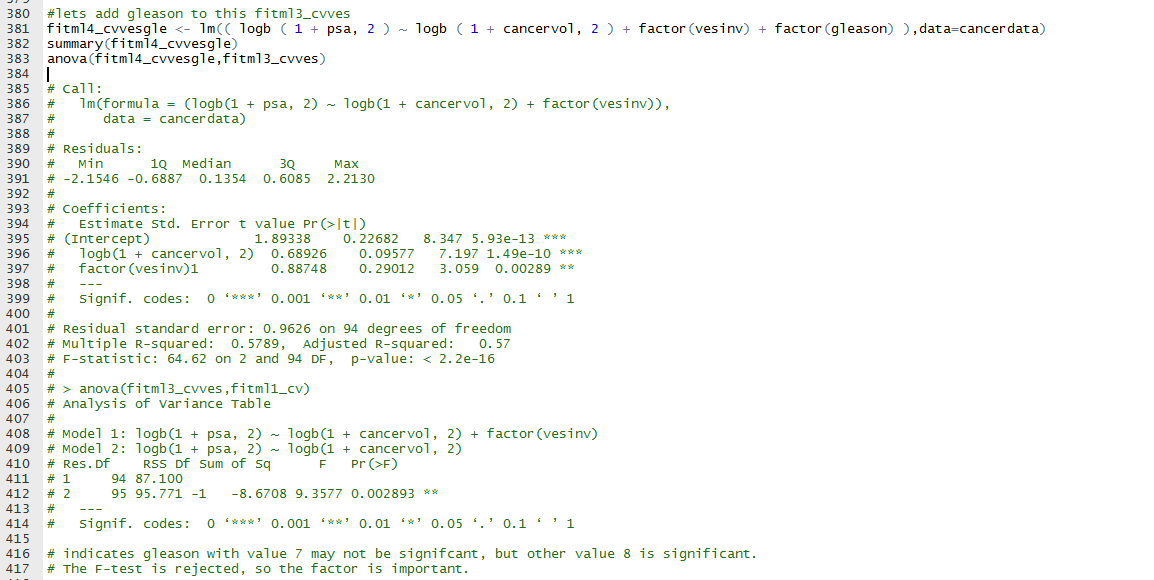
3) Multi Regression manual fitting.

3a) Cancervol + Capspen  


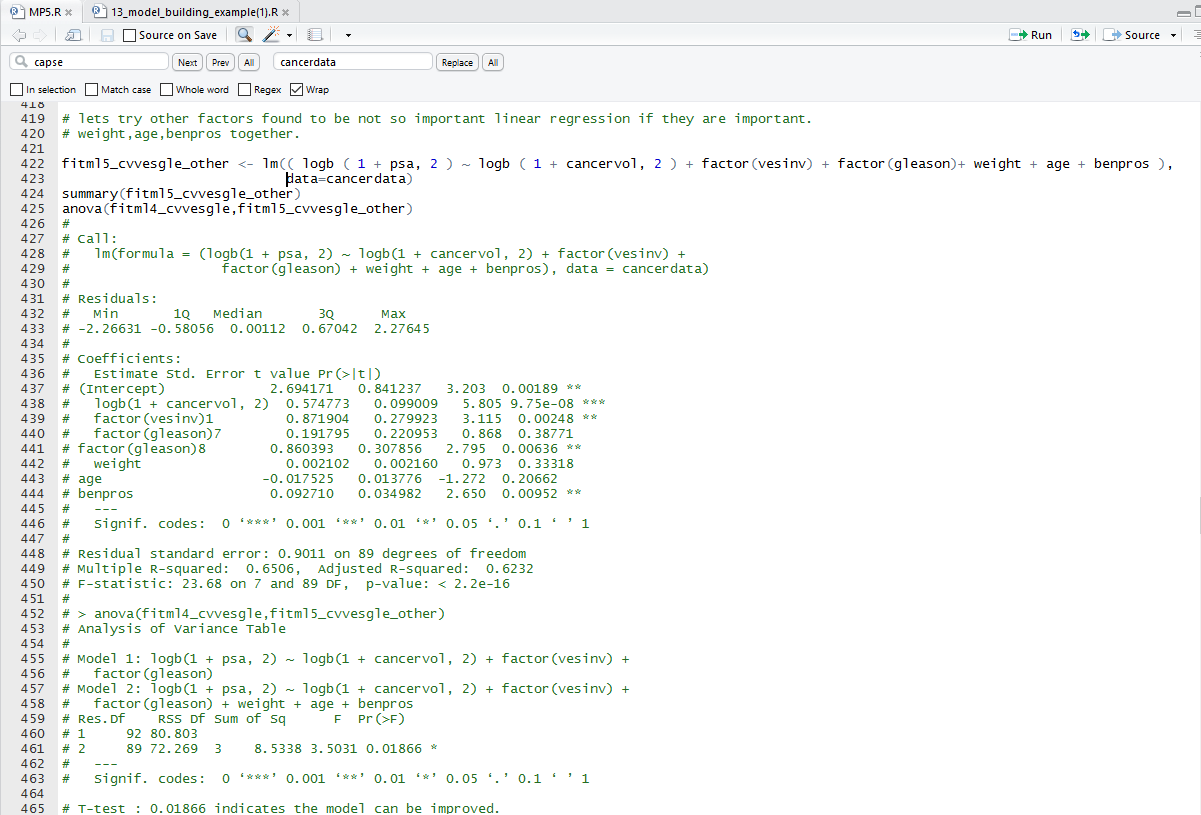
3b) Cancervol + Capspen + vesinv



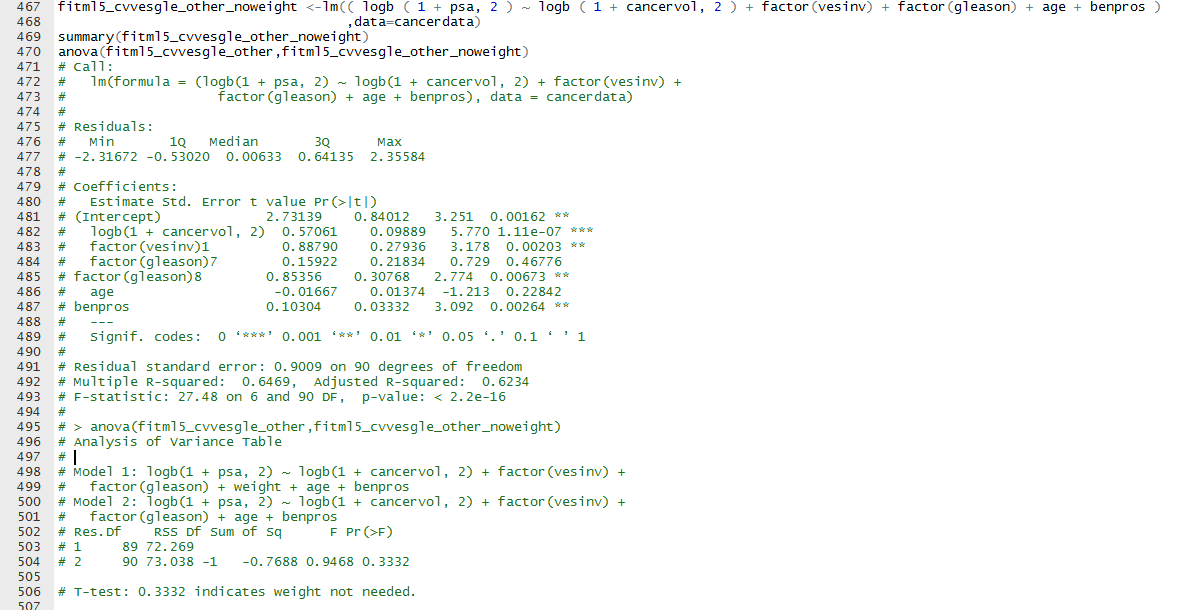
Cancervol + vesinv + gleason



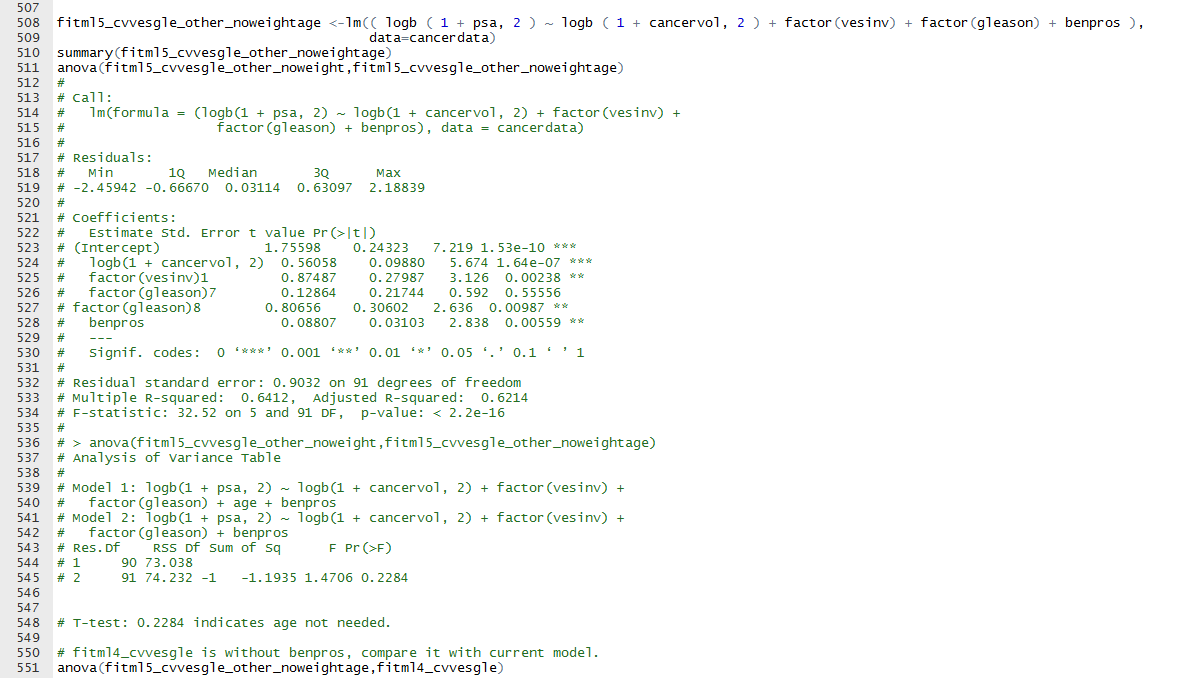
3 c) Cancervol + vesinv + gleason + weight + age + benpros

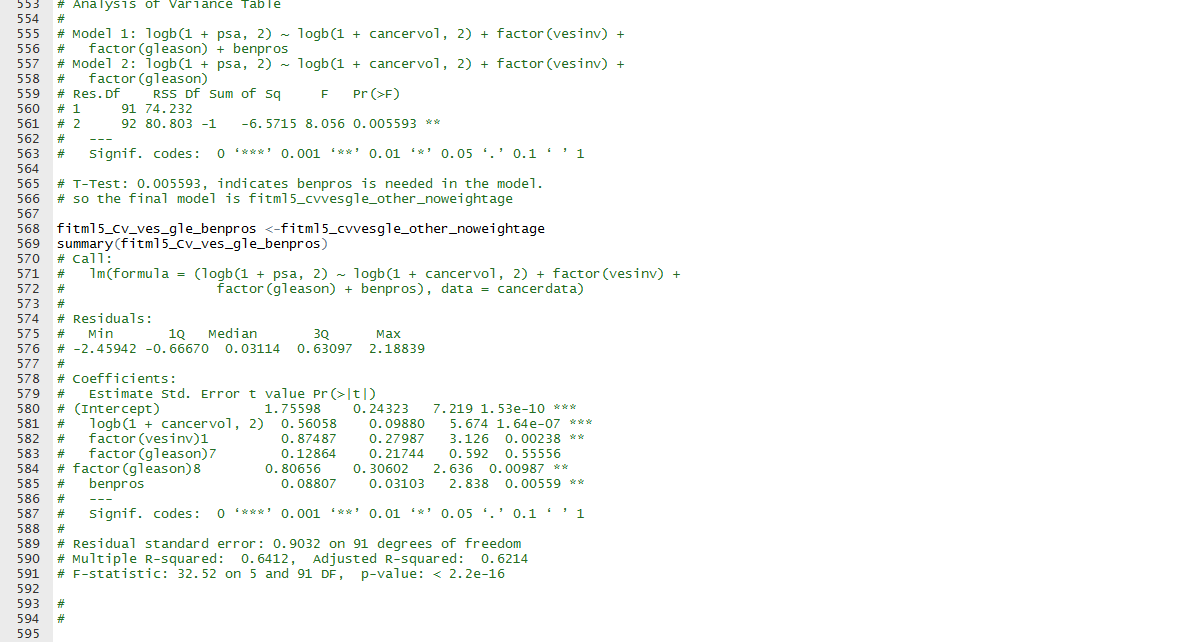


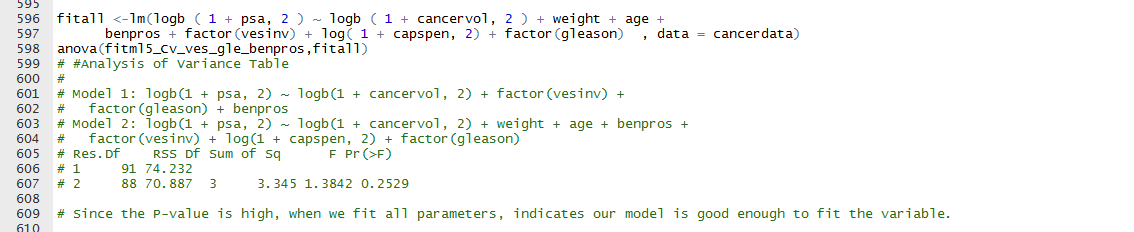
3 d) Cancervol + vesinv + gleason + age + benpros



3 e) Cancervol + vesinv+ gleason + benpros



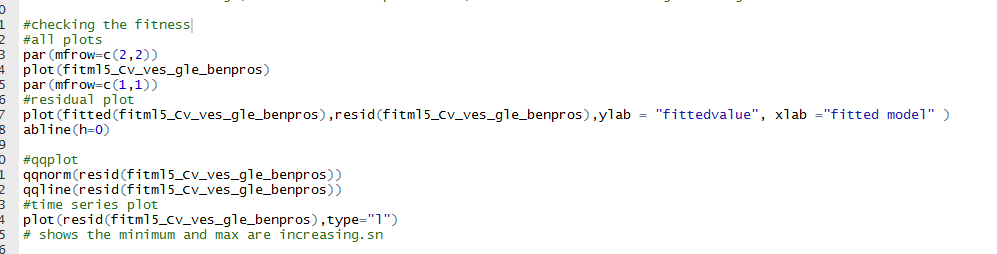




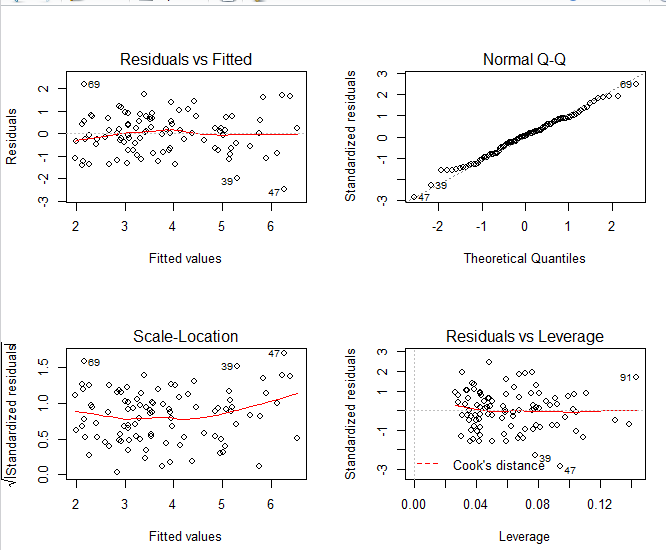
Finally, our model is fit which is

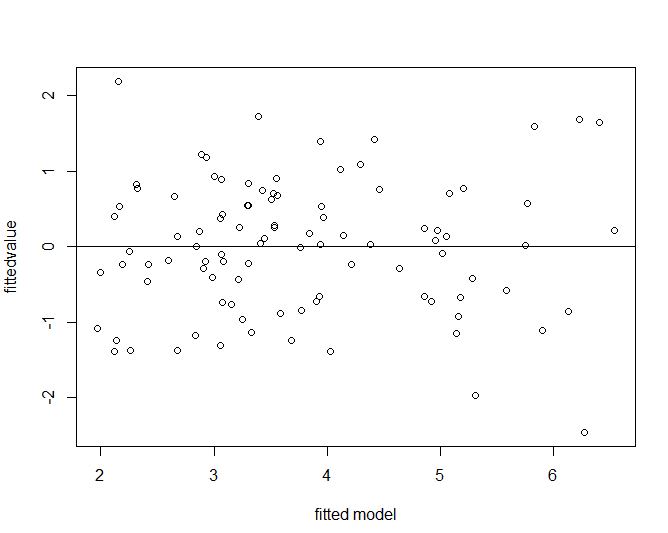
Y = fitml5\_Cv\_ves\_gle\_benpros <- <-lm(( logb ( 1 + psa, 2 ) ~ logb ( 1 + cancervol, 2 ) + factor(vesinv) + factor(gleason) + benpros ), data=cancerdata)

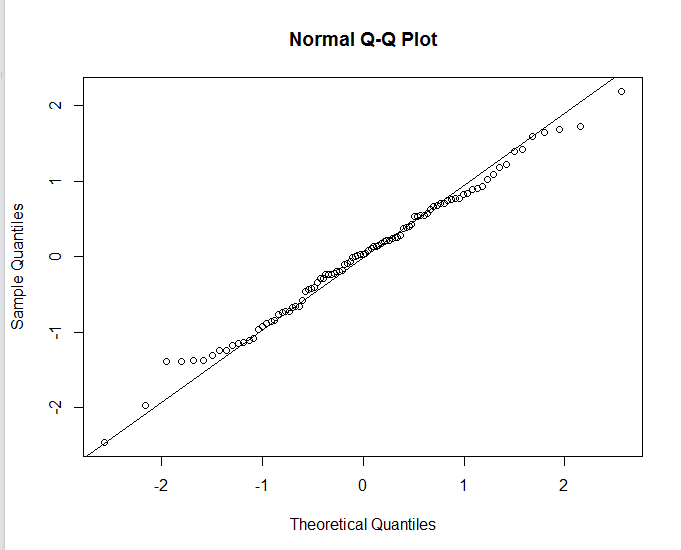
# 4) comparing the fitted model using residual, QQ plot, time series plot.

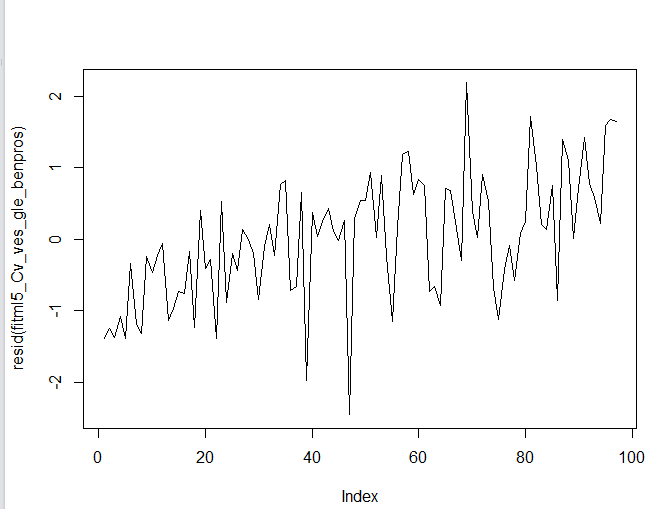


Standard library plot of the fit, shows that residual plots vary across equal across the mean. The normal QQ plot shows that the fit is normal. Hence Normality assumption for residual plot is good.





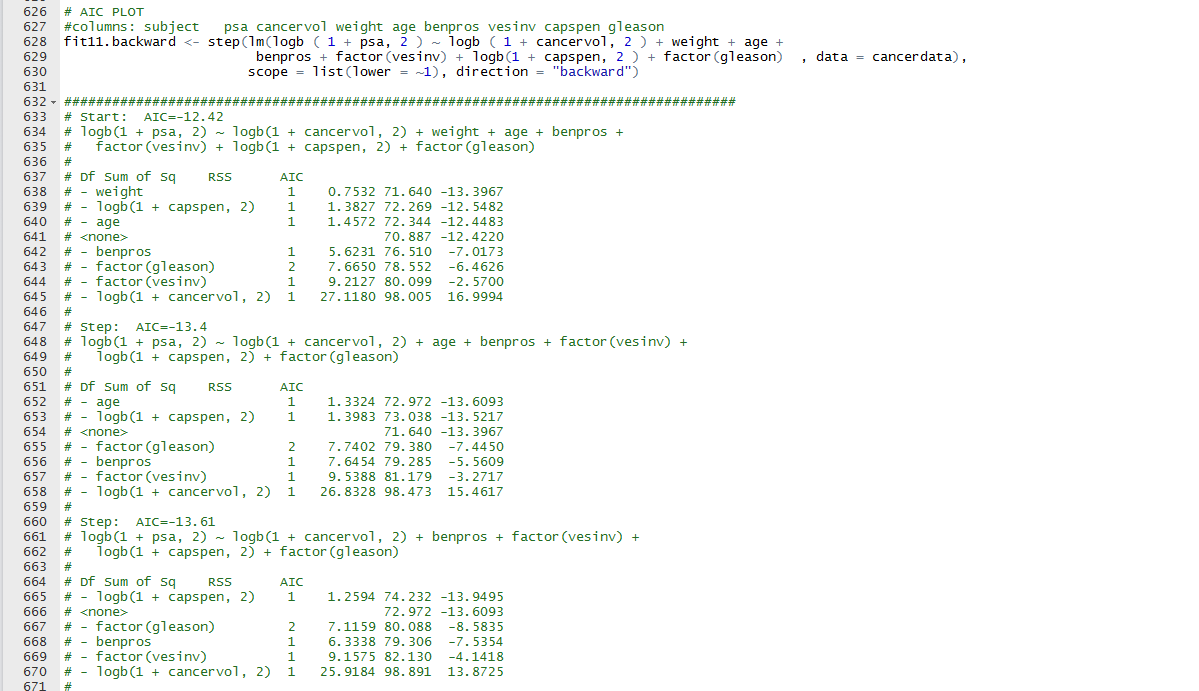


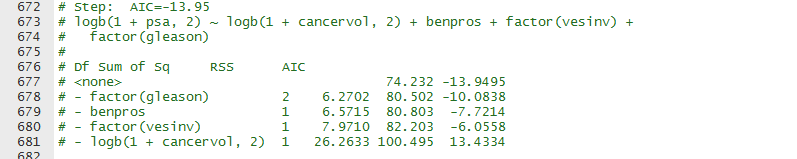


The QQ-plot shows our fit with normal assumption is true, residual plot vary equally across the mean and the time series plot indicates the values go almost from less lower to high lower values with time.

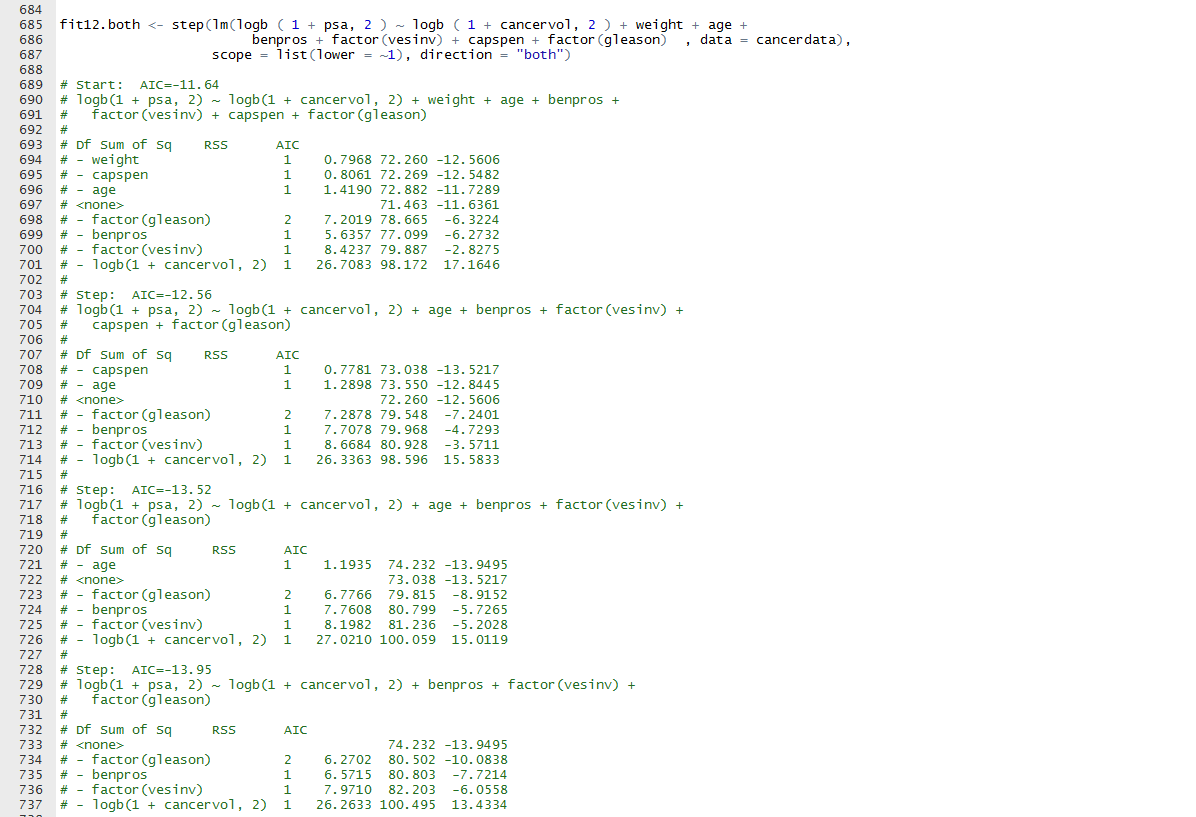
# 5 Automatic plots using AIC:

## 5.1 Backward

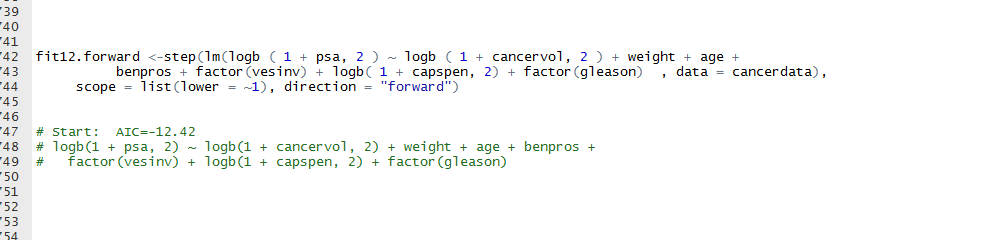




## 5.2 Both



## 5.3 Forward



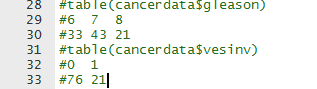
Both, Backward fit gave the same fitness model as we derived. But the forward model of the AIC gave all the factors In the model, which is the overfitted model.

# 

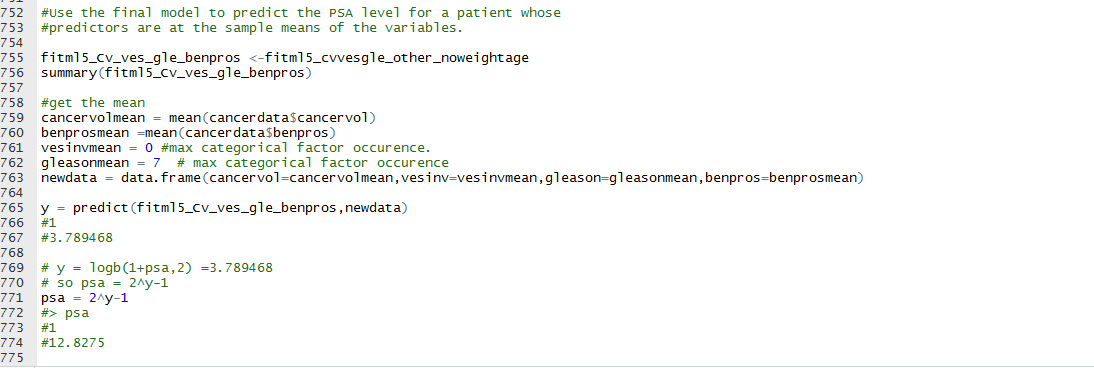
# 6) predict the value from the final model at the mean of the predictors.

Our final fitted Model is: “fitml5\_Cv\_ves\_gle\_benpros”.

The mean for the categorical variable is considered as max occurrence of a variable. So far Vesinv it is 0 and Gleason mean is 7.



The mean of other predictor is calculated using the mean function.

****

For the mean values of the predictors, the Prediction values of the y i.e. logb(1+psa,2) comes out to be 3.789468.

So psa = 2^y -1 = 12.8275.  
psa/log(2) =psa = (2^(y)-1)/log(2)

#18.50617

For the mean values of the predictors, the psa value comes out to be 12.8275.

For a linear regression, the line will pass through y mean for the value of x mean. But for a multi regression with factors considered as maximum frequency, this is not the case.   
By this technique, it avoids the overfitting of the model.