# Applied Statistical Modeling & Inference: Homework #4

Due on April 10, 2015 at 5:00pm

*Professor Ying Lu & Professor Daphna Harel*

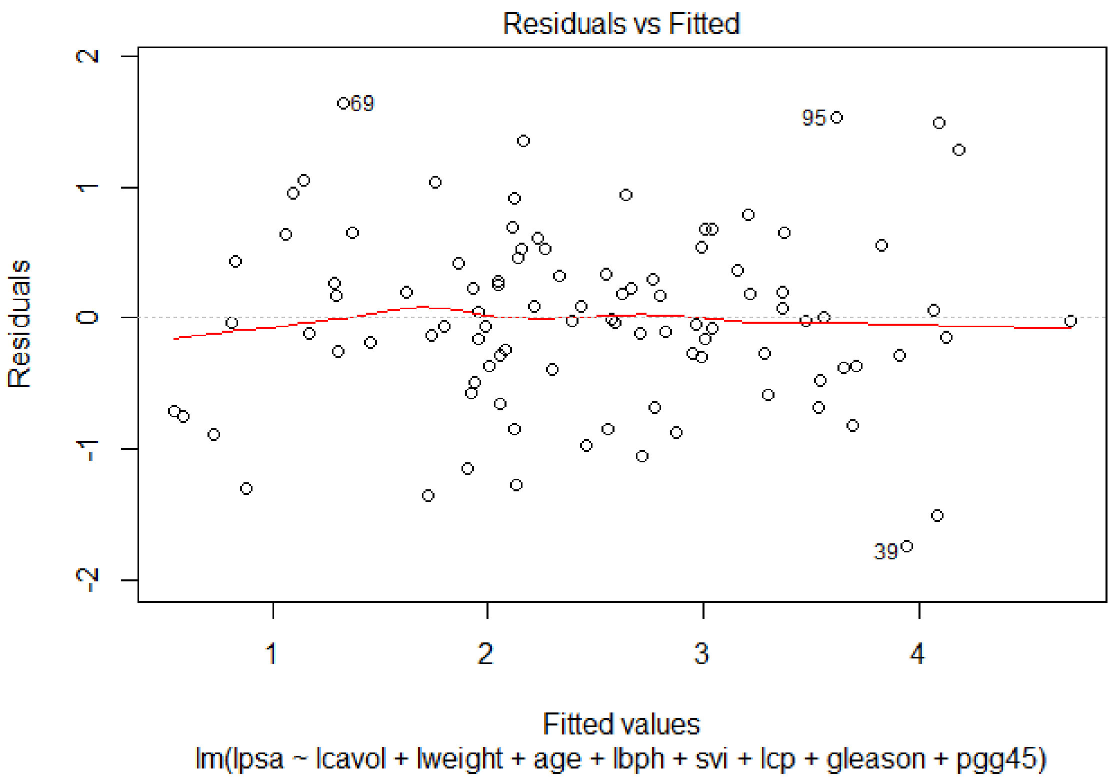
**Sriniketh Vijayaraghavan**

## Problem 1

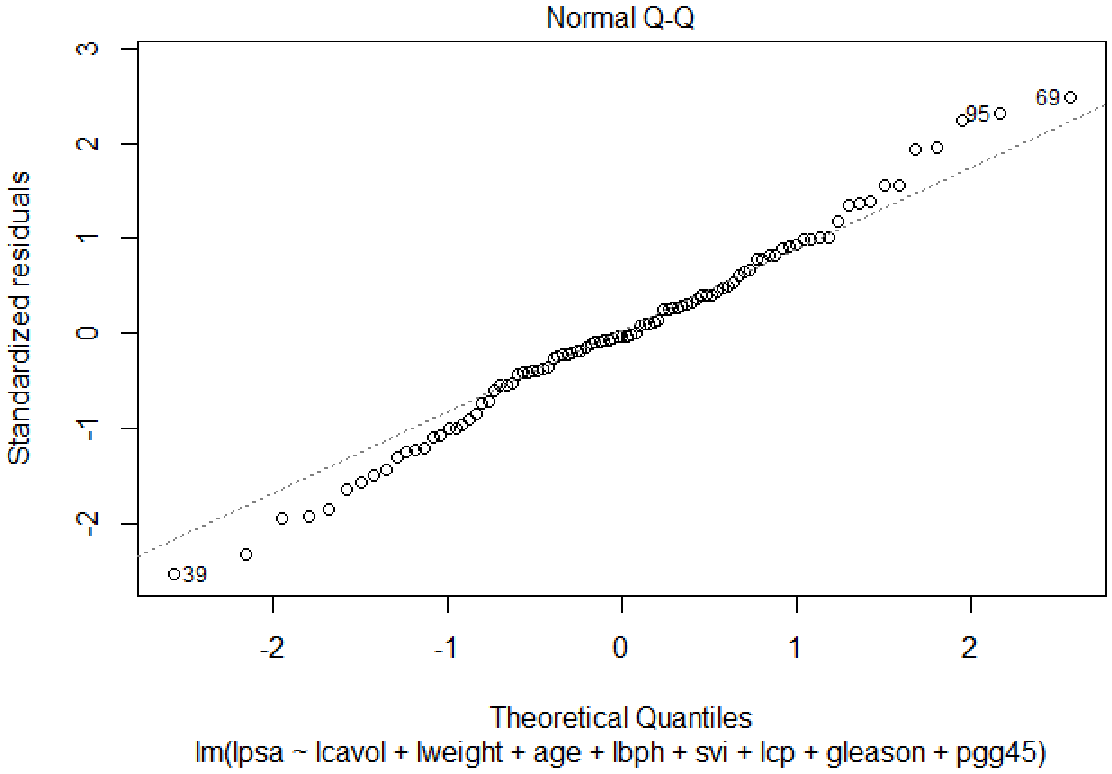
a. In the equation, $y_i = \beta_0 + \sum_{j=1}^{8} \beta_j x_{ij} + \epsilon_i, i = 1, ..., n$, the interpretations of each of the parameters $\beta_j$, for $j = 0, 2, 3, 8$ are as follows:
$\beta_0$ represents the intercept on the Y-axis having the value 0.669. This basically means that the function passes through the Y-axis at $X = 0$, i.e. the independent variable has a value of 0.669 when dependent variable has a value 0.
$\beta_2$ is the slope between the log of the prostate weight and the log of the prostate specific antigen.
$\beta_3$ is the slope between the age of the patient with the log of the prostate specific antigen.
$\beta_8$ is the slope between the percentage of Gleason scores of 4 or 5 with the log of the prostate specific antigen.

b. I have completed the summary statistic and have included the code in my R file. It includes the values of the estimates of the coefficients and intercepts, their standard error, the value of the residuals and the $R^2$ value.

c. The 2 plots can be seen below as follows(code is included in the separate R file)



Problem 1 continued on next page. . .



d. We use a stepwise AIC procedure which prunes through our model which is bidirectional in either adding or subtracting a term from the model in a stepwise fashion. At the end we are left with just a few features that describe the model. The final equation which we get that describes the model is as follows:

lm(formula = lpsa lcavol + lweight + age + lbph + svi, data = Prostate)

e. In this part we try to take the two-way interactions into account and we generate the new model along with the stepwise AIC procedure. After taking these into account, we get the final formula as follows:

lm(formula = lpsa lcavol + lweight + age + lbph + svi + lcp + pgg45 + lcp:pgg45 + lweight:lbph + lcavol:pgg45 + age:lcp + lweight:svi, data = Prostate)

f. We notice that the difference in (d) and (e) is the level of complexity in the model. It can be seen that the second model has a higher $R^2$ value and is also more explanatory. This would be useful for computers to get more accurate results. However, the first model would be much more useful to explain to a human. This is because that model in (d) is very simplified and has removed all the unnecessary features. So, both of these models have their own unique uses.

g. The AIC takes in a base model at the start and tries to minimize information loss in a step by step manor while trying to retain the least number of features. The AIC can work either forwards or backwards, but is by default backwards. In this method, we see that each time there is a score that is being calculated by the AIC, and this score is minimized at each step. So, this score is effectively compared with the other scores that are obtained in the stepwise process and the model decides on the best one with the least number of features.

   The AIC rewards goodness of fit, which is assessed by the likelihood function, but adds an extra condition which penalizes increase in the number of estimated parameters. This is an increasing function, which also is one of the reasons why AIC is preferred over Likelihood.

   In the above parts of the question, we ran AIC on our dataset with and without the interaction terms. We obtain the number from the AIC and also note down the summary function. The number generated by the AIC is used to compare it with other models. The lower the number, the better the model.

# Problem 2

a. The code to implement this has been included in the R code. Here, I will talk about the results that are seen in the summary function of the R code. We see that the equation is of the form:

$$Time = 144.36 + 5.46T1 + 2.03T2$$

This is the fit which we get from the estimates of the coefficients in the summary. We also notice that the standard error for T1 and T2 is very minimal, while the standard error for the Intercept is very large is bigger than the actual value of the intercept. We also notice that the distribution has its median very close to 0, but has a high variance which can be seen from the fact that the quantiles and the Min-Max are so far spread out from each other.
We can see that the adjusted $R^2$ value is 0.90. This seems to be a high value and could be seen as a measure of model adequecy. We also notice that the time taken to complete the work, i.e. labor is almost linearly increasing with the number of transactions of each type.

b. We see that A and D are getting the values NA and are missing. This is because these variables are either partially or completely linearly dependent on the other variables. So, R's lm() function automatically excludes such variables from the result.

c. In the 4 models, we notice that the $\beta_0$ values, the $R^2$, F-statistic, T-statistic and the residuals are the same. The only things that are different are the values of the coefficients of T1,T2,A and D.

d. It seems like the value of T2 is different between M1 and M3, but it is in fact the same. This is because part of the information which was contained in T2 is now comprised as a coefficient to D. If we bring down D to its basic form, we notice that we get back the same estimates for T1 and T2.

# Problem 3

a. We can see that the data can be simulated by using x1 and x2 are provided in the R code. I have calculated the confidence intervals along with a t distributed error with 2 degrees of freedom in the code. We see how the confidence coverage works out while using lm and tlm. We use confint to determine the confidence interval while using lm and effect to determine the confidence interval on tlm.

b. The same exercise is performed on both the fits and si given in the code.