STA6166 Final Project

Final Project Report for STA6166
Spring 2017

Biswas, Sayak

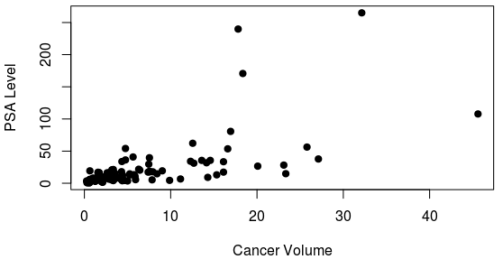
April 23, 2017

1 PSA Level Modeling

1.1 Introduction

The problem is to model the prostate-specific antigen(PSA) with respect to some clinical measurements. The provided number of data points in the sample is 97. I calculated some descriptive statistics on the sample provided. The mean PSA level of the sample is 23.73 mg/ml with a standard deviation of 40.78 mg/ml and a median value of 13.33 mg/ml.

Since we are to model PSA Level, we take it as the response variable and the rest of the variables as the predictors. To get an initial idea about the kind of relationship PSA Level has with the other variables we plot scatter graphs. Unfortunately, with this data set te scatterplots don't quite paint a clear picture of the relationship between the variables except *PSA Level vs. Cancer Volume* which seems to present a slight positive relationship. So, next we create a correlation matrix of the various variables. We can see that there is a moderately positive correlation between *Cancer Volume* and *Capsular Penetration*.



	ID	PSA_lvl	Cancer_vol	Weight	Age	BPH	SVI	CP	Gleason_score
ID	1.000	0.603	0.621	0.114	0.197	0.165	0.567	0.477	0.538
PSA_lvl	0.603	1.000	0.624	0.026	0.017	-0.016	0.529	0.551	0.430
Cancer_vol	0.621	0.624	1.000	0.005	0.039	-0.133	0.582	0.693	0.481
Weight	0.114	0.026	0.005	1.000	0.164	0.322	-0.002	0.002	-0.024
Age	0.197	0.017	0.039	0.164	1.000	0.366	0.118	0.100	0.226
BPH	0.165	-0.016	-0.133	0.322	0.366	1.000	-0.120	-0.083	0.027
SVI	0.567	0.529	0.582	-0.002	0.118	-0.120	1.000	0.680	0.429
CP	0.477	0.551	0.693	0.002	0.100	-0.083	0.680	1.000	0.462
Gleason_score	0.538	0.430	0.481	-0.024	0.226	0.027	0.429	0.462	1.000

1.2 Initial Model

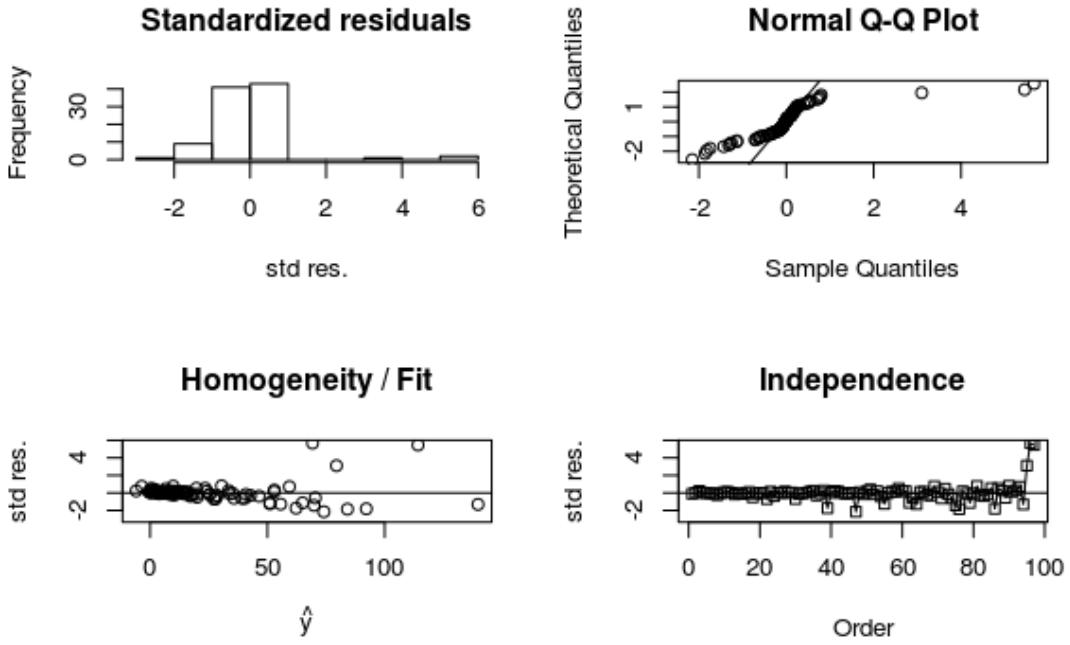
Keeping the previous observations in mind, we now develop a full linear model using all the predictors and without any interaction terms. Assuming all the model assumptions are met(we will verify this later), we test the overall fit of the model with the below hypotheses:

$$H_0 : \beta_1 = ... = \beta_7 = 0 \text{ vs } H_a : \text{at least one of them} \neq 0$$

This initial model has $R^2 = 45.85\%$ and $R^2_{adj} = 41.59\%$. This means that this model is able to explain a large portion of the variability. Also the test statistic value is $T.S. = 10.77$ with an associated p-value of almost 0 found using an $F_{7,39}$ distribution. So, at least one predictor is significant. We will use this information to check if the model can be reduced in any way as the AIC value is around 952.17.

1.3 Model Assumptions

For now let's validate if the model assumptions are met. We make use of the automated *check* function provided by the professor. The graphs generated are shown below:

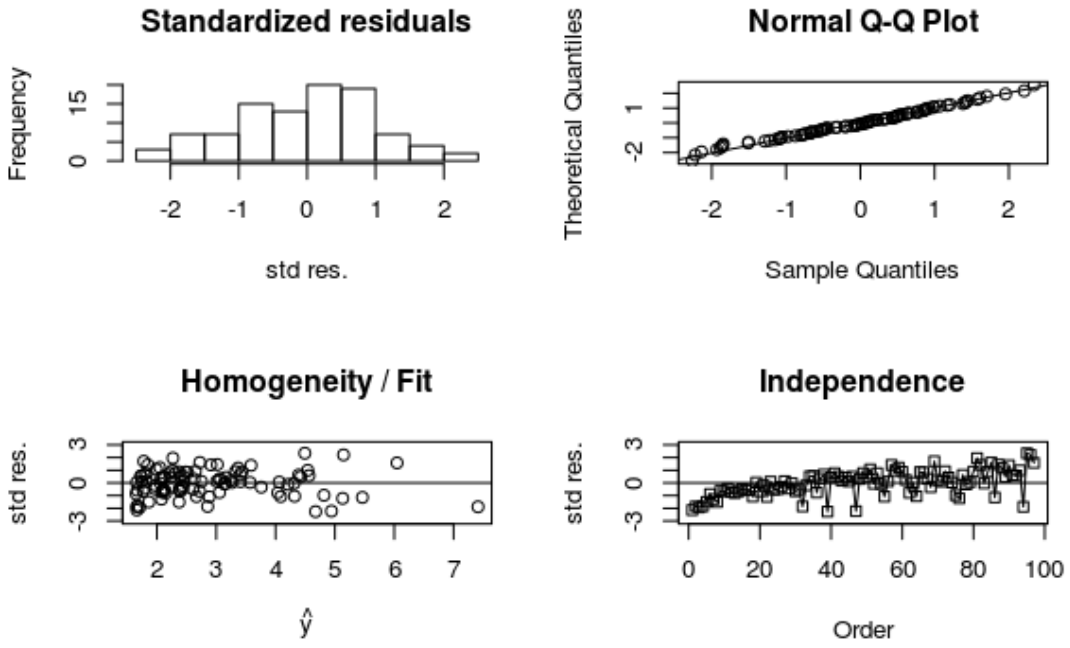


- We can see that the QQ plot and the histogram show heavier left and right tails compared to the normal. Also, the Shapiro-Wilk test which gives a p-value of almost 0 and hence confirms violation of normality.

- As there is no discernible pattern in the time-series plot of the data, we can safely conclude the data to be independent.
- From the graph, variance doesn't seem to constant at all places but the fit of the model seems to be almost correct.

1.4 Transformation

In light of the above, we perform Box-Cox transformation with an estimated power of 0.1004 which fixes the violations as can be seen below:

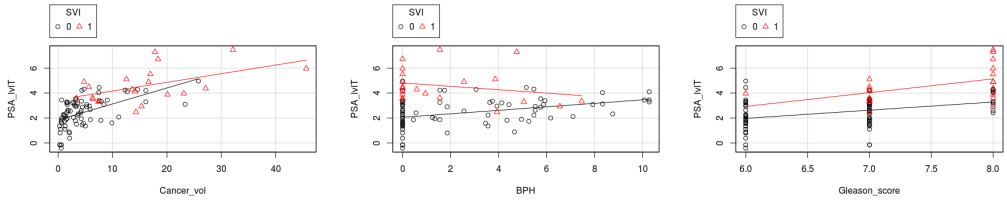


1.5 Model Reduction

Now, we perform automatic reduction of model using the stepT function provided to us. This removes the predictors with high p-values and the ones which are highly correlated. Thus *Age*, *Weight* and *Capsular Penetration* are removed. We now see that the R^2 has increased to 60.71% and ${R^2}_{adj} = 59\%$. Also, the AIC value has reduced to 273.99.

1.6 Interaction Model

We still haven't addressed the fact that we have a qualitative predictor i.e. *Seminal vesicle invasion*. So, we use scatterplots with regression line enabled to check for interaction between this and the other variables.



From the plots, there seems to be an interaction between *Seminal vesicle invasion* and the other predictors. So, we add the corresponding interaction terms to the model. This improves the model a bit and we get $R^2 = 63.33\%$ and $R^2{}_{adj} = 60.89\%$. More importantly, the AIC reduces to 271.28. Next, we perform one more round of automatic model selection, which removes the *Seminal vesicle invasion* predictor while leaving the interaction terms as per the higher p-value. This has marginal increase in the R-squared and the R-square adjusted values and a decrease in the AIC value.

1.7 Final Model

We pick this as our final model as it has the lowest AIC and highest R-squared adjusted.
$\ln(PSA\_lvl) = -0.67 + 0.11 * cancervol + 0.14 * bph + 0.33 * gscore - 0.06 * cancervol * svi - 0.21 * bph * svi + 0.28 * gscore * svi$
This model has $R^2 = 0.6384, R^2{}_{adj} = 0.6143$ and $AIC = 269.94$.

1.8 Prediction Interval

We create a 90% prediction interval for the new data provided using R:

$$ln(PSA\_lvl(mg/ml)) = (0.4114, 3.5446)$$

The point estimate for the given data comes out to be in logarithmic terms:

$$ln(PSA\_lvl(mg/ml)) = 1.978$$

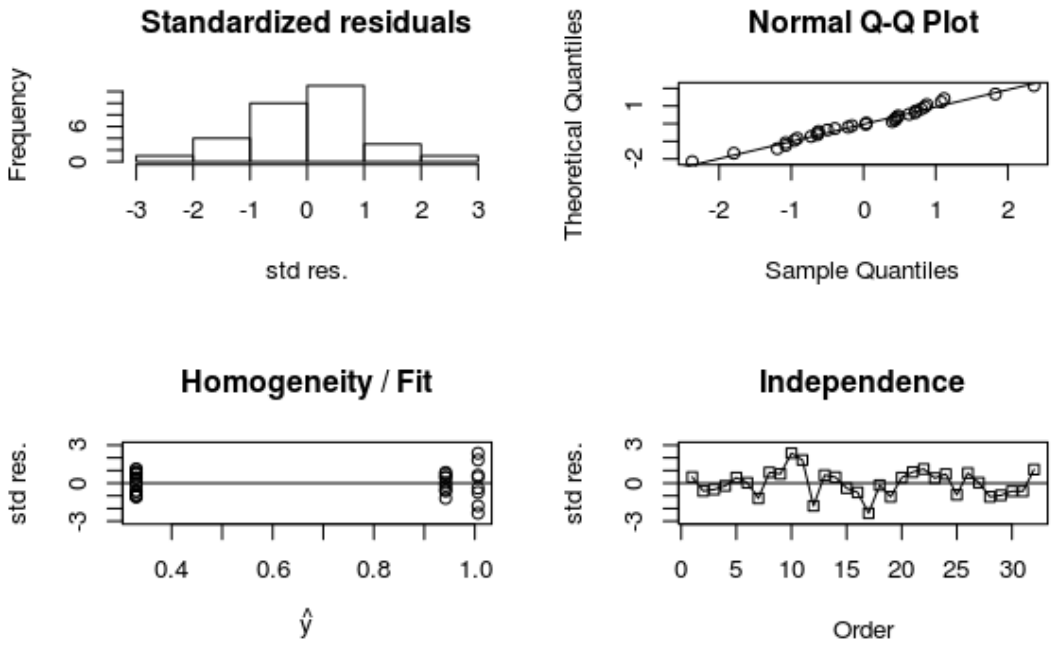
2 Completely Randomized Design

We are required to find the differences if any in the mean sorption rate of three different chemical cleaning solvents. We read the data set in R. As the lengths of the data for each solvent is different, we use factor to load the dataframe.

We find the mean sorption rate of each solvent:

- Aromatics = 0.9422
- Chloroalkanes = 1.0063
- Esters = 0.33

We create the analysis of variance model and check for model assumptions.



- Shapiro-Wilk test gives a p-value 0.88 suggesting data might be normal. This is confirmed by the QQ plot and the histogram.
- We conclude independence as there is no discernible pattern in the time-series plot.
- Data points are evenly distributed about the 0-line, so model fit can be concluded.

The 1-way ANOVA gives a p-value of approximately 0, hence we can conclude that not all solvents have the same mean. Next, we perform the Bonferroni test which shows that only means of Aromatics and Chloroalkanes don't differ. Similarly, Tukey's HSD test has a 0 in the interval for the pair. The pairwise differences are as follows:

- Aromatics - Esters: 0.612
- Chloroalkanes - Esters: 0.676