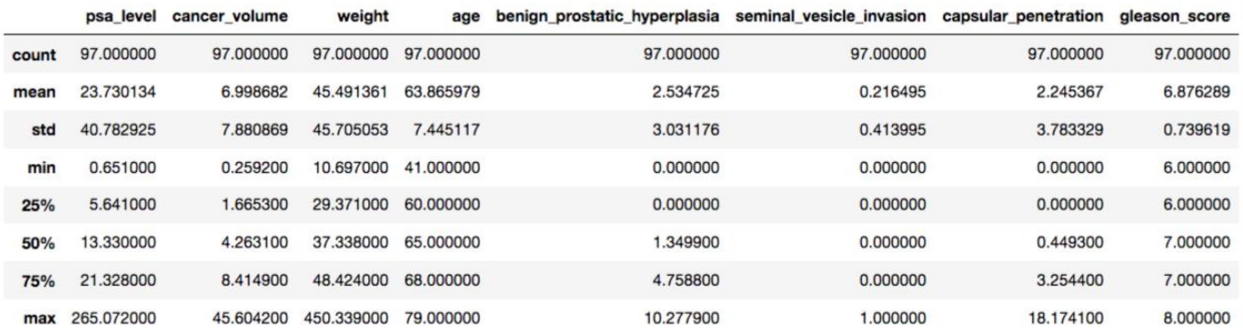
Question 1:

Descriptive statistics for the original dataset:



	psa_level	cancer_volume	weight	age	benign_prostatic_hyperplasia	seminal_vesicle_invasion	capsular_penetration	gleason_score
count	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000
mean	23.730134	6.998682	45.491361	63.865979	2.534725	0.216495	2.245367	6.876289
std	40.782925	7.880869	45.705053	7.445117	3.031176	0.413995	3.783329	0.739619
min	0.651000	0.259200	10.697000	41.000000	0.000000	0.000000	0.000000	6.000000
25%	5.641000	1.665300	29.371000	60.000000	0.000000	0.000000	0.000000	6.000000
50%	13.330000	4.263100	37.338000	65.000000	1.349900	0.000000	0.449300	7.000000
75%	21.328000	8.414900	48.424000	68.000000	4.758800	0.000000	3.254400	7.000000
max	265.072000	45.604200	450.339000	79.000000	10.277900	1.000000	18.174100	8.000000

Multiple Regression:

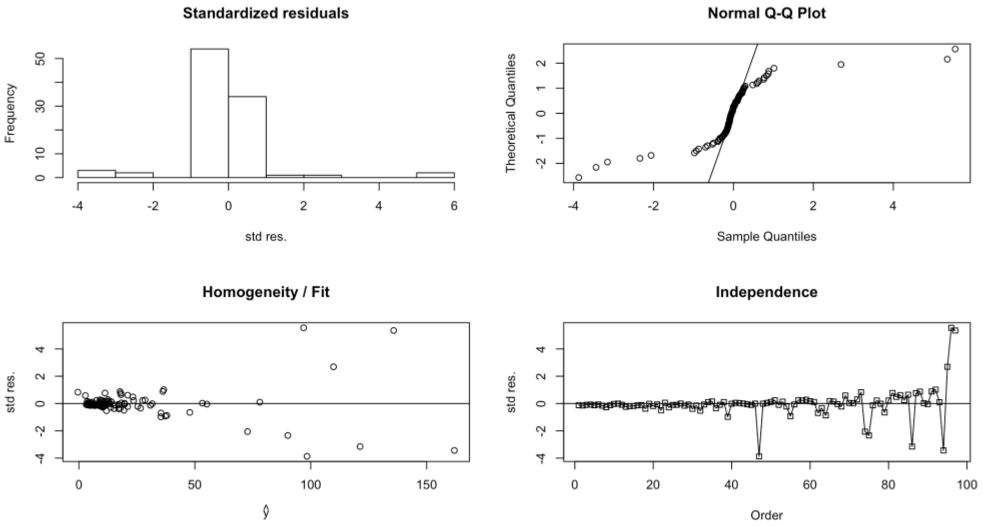
Based on the description of the data, we can see that "seminal_vesicle_invasion" is a categorical variable, and others can be considered numerical variables.

So, for linear regression model, assume the full model is:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * x_4 + b_5 * x_5 + b_6 * x_6$$
$$+ b_7 * x_7 + b_8 * x_5 * (x_1 + x_2 + x_3 + x_4 + x_6 + x_7)$$

(y is psa_level; x_1 is cancer_volumn; x_2 is weight; x_3 is age; x_4 is benign_prostatic_hyperplasia; x_5 is seminal_vasicle_invasion; x_6 is capsular_penetration; x_7 is gleason_score)

Fit the model and check the assumption:



This is obviously not eligible for any further analysis (histogram and Q-Q plot show that it did not meet the normality requirement; and the spread of residuals were not constant; further, the pattern of independence check was almost discernable).

Although converting the response to logarithmic value may lead it to linear, we wouldn't be able to get the predict interval of original response after such a transformation.

So, I plan to apply another model: Random Forest.

Random Forest:

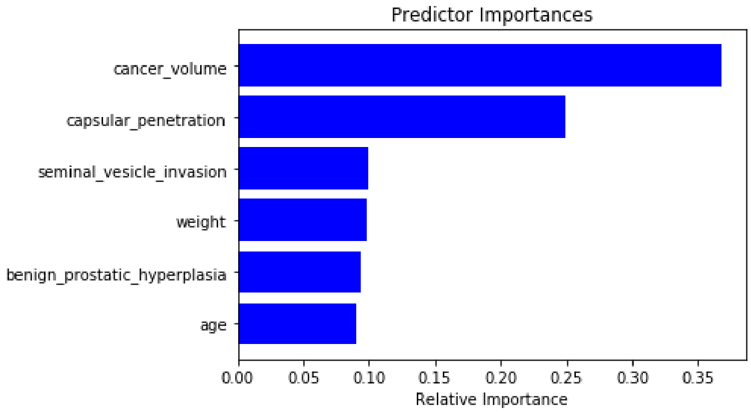
Firstly, I did recursive feature elimination with k-fold (k=3) cross validation (RFECV) and default hyper-parameters for Random Forest Regression model. The reason of using 3-fold is that a small k value would lead to a small variance for cross validation result, which is good for model comparison and selection.

The RFECV shows the best predictor set would be: "age", "weight", "cancer_volumn", "benign_prostatic_hyperplasia", "seminal_vesicle_invasion" and "capsular_penetration", which means "gleason_score" was not important and was abandoned.

Then, I used the selected predictors to do grid-search with 3-fold cross validation for hyper-parameter tuning. Since there are only 6 selected predictors, I didn't tune the maximum of random selected features, and only tuned for the number of trees. The parameter grid was {number of trees: 25, 50, 100, 200, 400, 800, 1600}. The result showed that 400 was the best choice. And the out of bag R-square was 0.2882.

I then trained Random Forest with tuned hyper parameter and selected predictors, and using the full data set (97 samples), got an adjusted R-square of 0.8984 on the full data set.

Following figure is the predictor importance output by the trained model:



Interpretation: if "cancer_volumn" was removed, the R-square would decrease more than 0.35; if "capsular_penetration" was removed, the R-square would decrease about 0.25; If “seminal_vesicle_invasion” or “weight” or “benign_prostatic_hyperplasia” or “age” was removed, the R-square would decrease about 0.09.

To get the predict interval of input {cancer_volumn = 4.2633; weight = 22.783; age = 68; benign_prostatic_hyperplasia = 1.35; seminal_vesicle_invasion = 0; capsular_penetration = 0; gleason_score = 6}, I applied each of the decision trees (400 trees in total) in the random forest model to get 400 predict values, then generated 90% predict interval for it.

The result shows the 90% predict interval of “psa_level” is from 3.857 to 14.296.

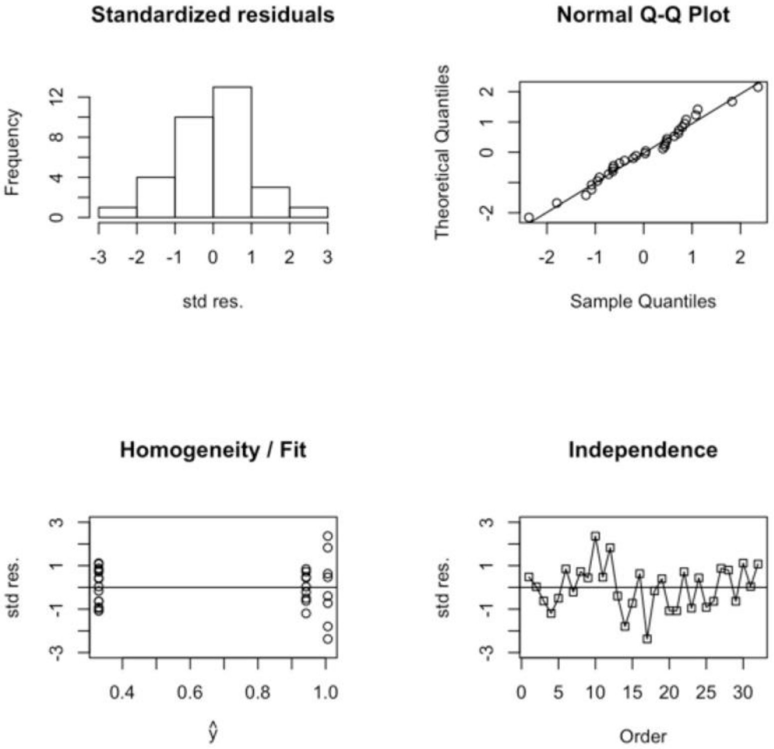
Question 2:

Descriptive Statistics for the original data:

Group	Data															Mean	Std.
Aromatics	1.06	0.95	0.79	0.65	0.82	1.15	0.89	1.12	1.05							0.94	0.17
Chloroalkanes	1.58	1.12	1.45	0.91	0.57	0.83	1.16	0.43								1.01	0.40
Esters	0.29	0.43	0.06	0.06	0.51	0.09	0.44	0.10	0.17	0.55	0.53	0.17	0.61	0.34	0.60	0.33	0.21

Analysis of Variance:

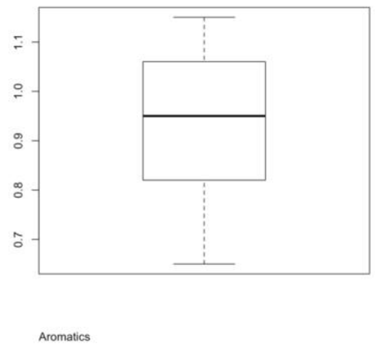
Fit the data to regression model and check the assumption:

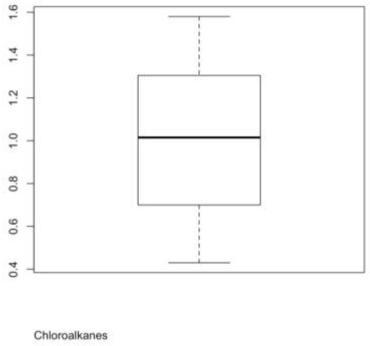


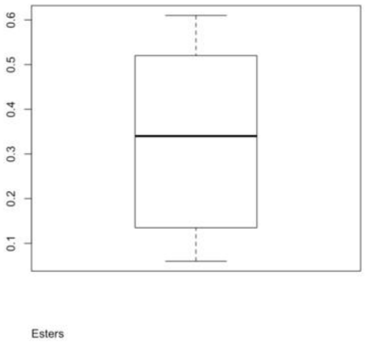
The assumption of homogeneity is obviously violated. The variance of Chloroalkanes is much larger than that of Aromatics and Esters. So, change the comparison strategy to t-test.

T-test:

Check the normality of three groups of data:







The boxplots are symmetric with whiskers of approximately the same length. There are no obvious violations of the assumptions.

T-test result:

Comparison Groups	P-Value	95% Confidence Interval
Aromatics ~ Chloroalkanes	0.6842	[-0.41, 0.28]
Aromatics ~ Esters	1.53*10^{-7}	[0.45, 0.77]
Chloroalkanes ~ Esters	0.0016	[0.33, 1.02]

Therefore, based on the t-test, the mean values are the same between Aromatics and Chloroalkanes. But the mean values of Aromatics and Esters are different, that of Chloroalkanes and Esters are also different.