Applied Regression Analysis (Final Report)

Supti Biswas

The package 'mclust' in R contains a dataset named 'thyroid' which has 215 observations and 6 variables.

- **Variables** – TSH, T3, T4, DTSH, RT3U and Diagnosis
- **TSH** – Secretion of Thyroid Stimulating Hormone in a person (quantitative)
- **T3** – Amount of 'Triiodothyronine' hormone in a person (quantitative)
- **T4** – Amount of 'Thyroxine' hormone in a person (quantitative)
- **DTSH** - Maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value (quantitative)
- **RT3U** – A blood test performed as a part of evaluation of thyroid function of a person (quantitative)
- **Diagnosis** – A person's secretion level of 'thyroid' hormone which are 'Hypothyroidism', 'Normal' and 'Hyperthyroidism' (categorical/qualitative)

- Now the most important variables are **T3** and **T4** because these two hormones stimulate in a human body to secrete **TSH.** We want to know that how T3 and T4 are associated with TSH. Also, how the other independent variables are associated with TSH.

Regression Model 1 - After taking all the variables and fitting a multiple linear regression model, only the 'Diagnosis' variable is significant (after F-test) at 5% level of significance. The R^2 value is 45.85% and AIC value is 1272.075.

Regression Model 2 – After removing the categorical variable and taking all the quantitative variables, the model is still no better than the previous one but only T4 is significant at 5% level of significance. The R^2 value is 31.31% and AIC value is 1319.236.

Regression Model 3 – After taking the quadratic effect of T3, it is significant at 5% level of significance. This model seems the best among these 3 models as the R^2 value is 47.48% and AIC value 1267.519.

So, the final regression model is –

$$TSH = 5.42447 - 2.09796T3 + 0.21578T3^2 - 0.21659T4 + 0.09820DTSH + 0.03787RT3U -$$

$$2.91569Diagnosis\ (Hypo) + 4.42344Diagnosis\ (Normal)$$

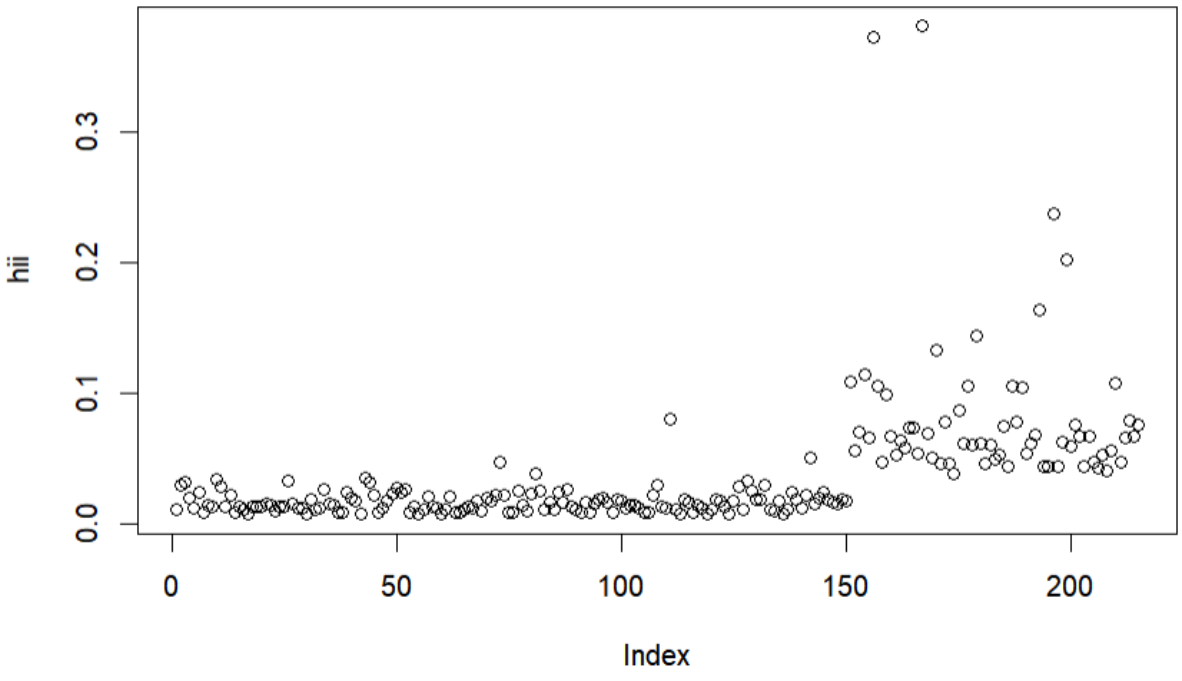
From this model, we see that when T3 is entering in the model with a quadratic term, it has significant effect on the response variable. And the R-squared is also higher than the previous models which is 47.48%. So, we can conclude that the variable T3 has a quadratic effect on the response.

- Though we've fitted a linear regression model, but if we take 'log' of the response variable, R square is higher than all the previous model, 63.57%. But our interest T3 and T4 are not significant.

Checking Outliers - After t-test, we found 3 outliers in the data –

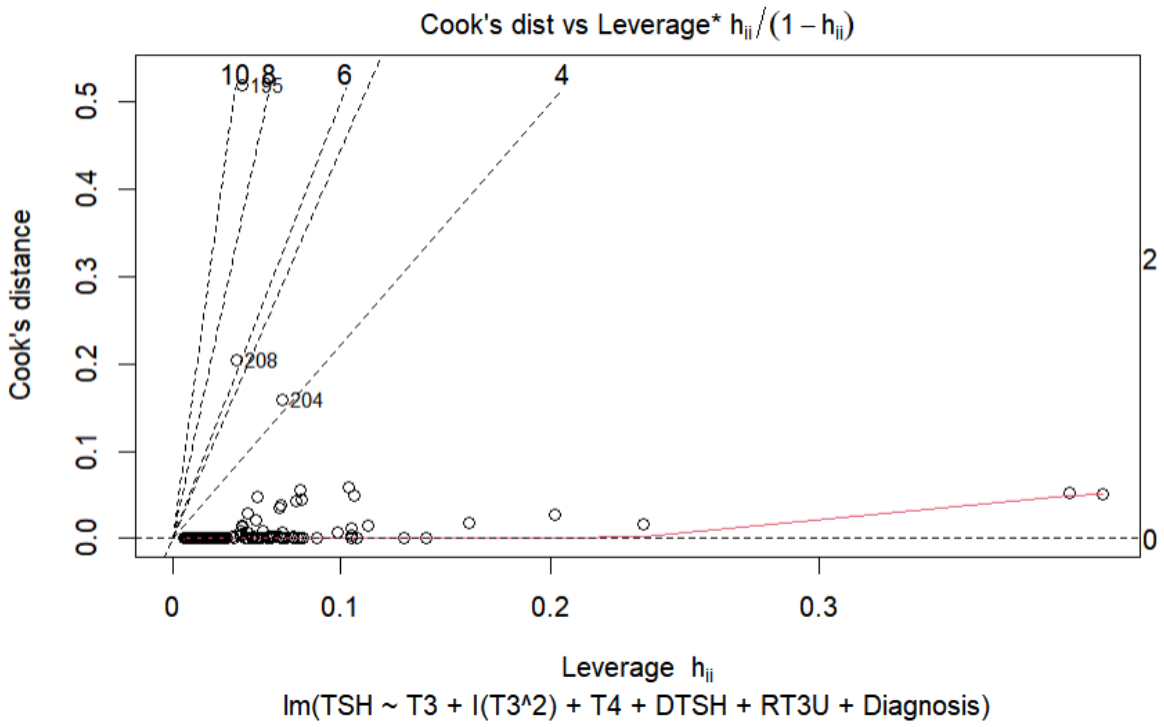
Data Point	TSH Level
195	56.4
204	32.6
208	41.0

Checking High Leverage Points – There are some high leverage points in the data which are 111 151 154 156 157 159 167 170 172 175 177 179 187 188 189 193 196 199 201 210 213 215. The graph is below -



The points lying on the upper side than the other points are the high leverage points.

Checking Influential Points – After calculating Cook's distance, there is no influential points in the dataset. The graph is below –



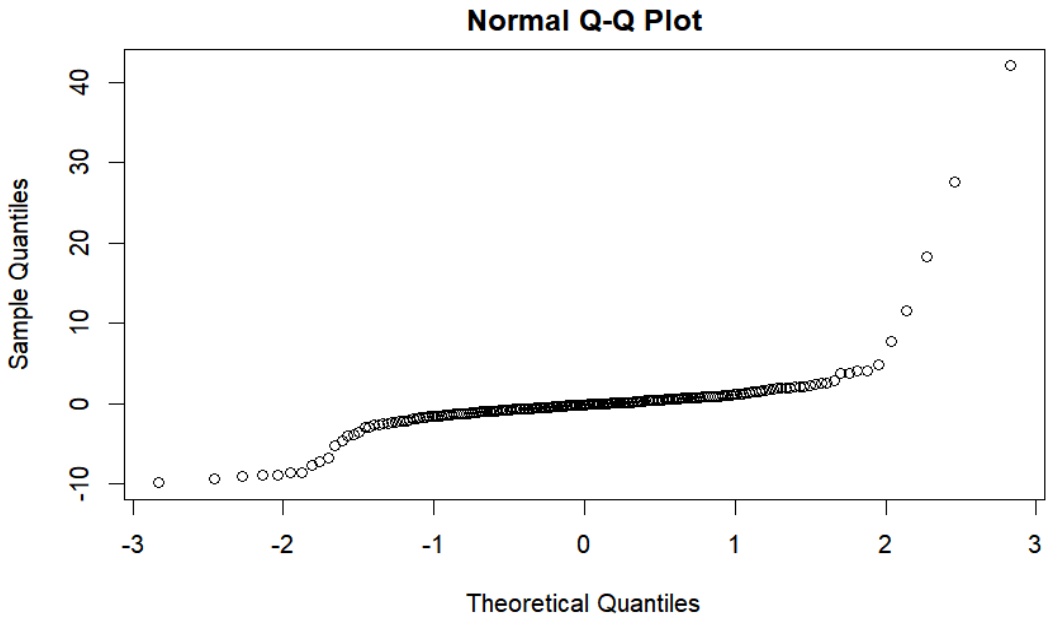
- As there is high leverage points and outlier, a robust regression model could give us a better fit for this dataset –
- **Least Absolute Deviation regression output** –

	coefficients	lower bd	upper bd
(Intercept)	8.28645	6.95726	10.80928
T3	-0.27032	- 0.44077	-0.00495
I(T3^2)	0.02531	0.00454	0.04587
T4	-0.00716	-0.03581	0.02989
DTSH	0.05551	0.01380	0.07858
RT3U	0.00396	-0.00608	0.00796
Diag(Normal)	-7.19148	-10.97760	-3.99431

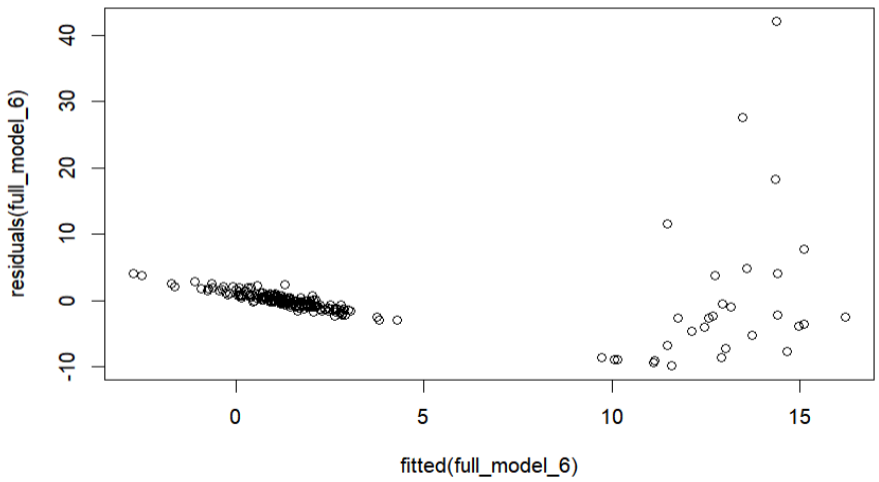
Diag(Hyper) -6.94718 -8.95336 -6.18947

- If we run a robust regression model, all the variables are significant as the coefficients are lying between the lower and upper bound of the interval. So the reason of not finding T3 and T4 significant in the previous models could be the outliers and high leverage points.

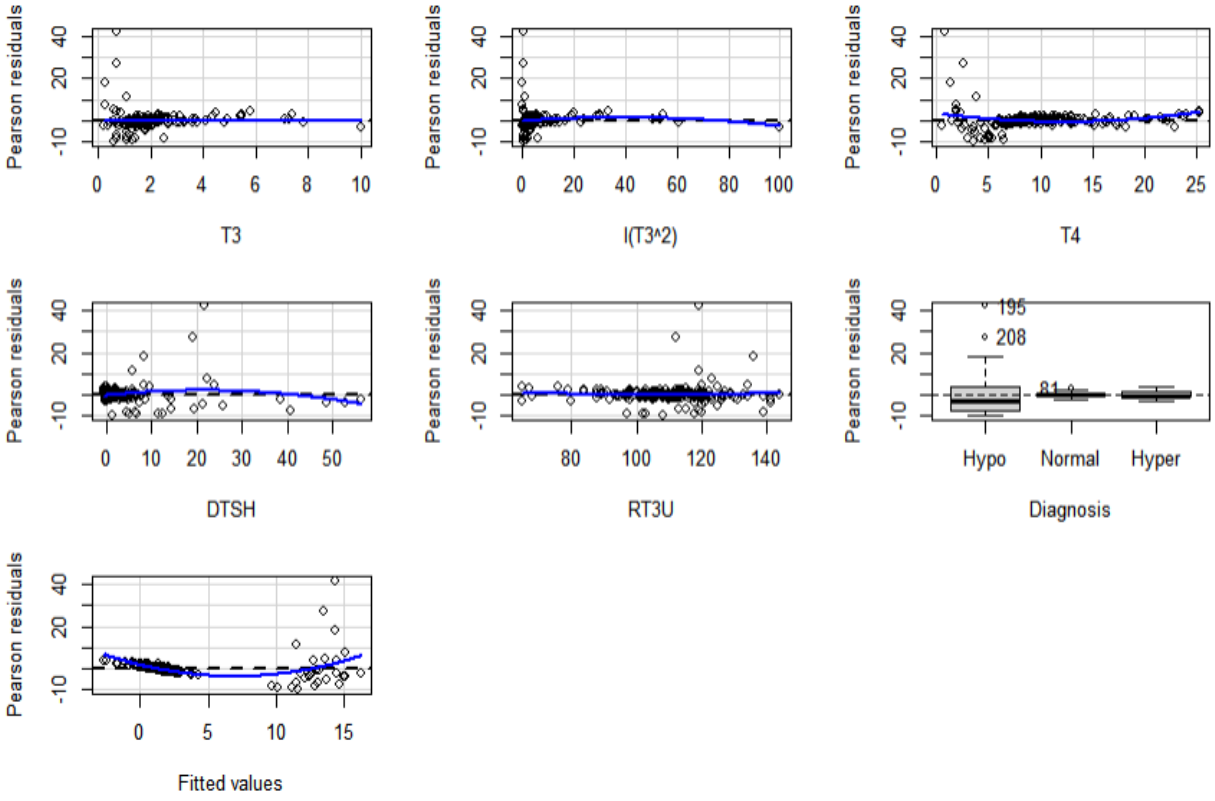
Checking Normality Assumptions – In the linear model, the residuals are not following a normal distribution. The graph is below -



Checking Homoscedasticity - The error variances are not equal from the graph



Non linear effect – From the residual plots, we see some pattern and there might be some non linear effect as well in the data. The graph is below –



- There could be some further analysis of the data using logistic regression where we can take the ‘Diagnostic’ variable as the response which is a categorical variable and take other variables as independent variables. Then we can see if a person has hypothyroidism/hyperthyroidism/normal level of Thyroid Stimulating Hormone in their body.