Fitting a Regression Model on ‘thyroid’ data using R

Data description

- 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)

Question from the dataset

- Secretion of Thyroid Stimulating Hormone in a person is dependent on another two important thyroid hormone (Triiodothyronine and Thyroxine) in human body.

- How is the secretion of Thyroid Stimulating Hormone associated with the other independent variables?

- Also, how does the two most important independent variables T3 and T4 are associated with TSH?

Findings from Running a Linear Regression Model

(Intercept)	T3	T4	DTSH	RT3U
10.16710334	-0.10634921	-0.25665564	0.08185691	0.01934854

Diagnosis(Normal)	**Diagnosis(Hyper)**
-8.65123900	-6.02693064

- Only the ‘Normal’ and ‘Hyper’ category from variable ‘Diagnosis’ are significant (after F-test, p-value is 1.787e-11)
- The main factors (T3, T4) causing secretion of TSH are not significant
- 45.85% of total variation of ‘TSH’ are explained by the explanatory variables
- AIC value is 1272.075

What happens if we remove the effect of 'Diagnosis' from the data?

(Intercept)	**T4**	T3	**DTSH**	RT3U
0.13557211	-0.39922535	0.47572447	0.28110852	0.04108708
- One of the most important factors 'T4' is significant and the 'DTSH' variable has significant effect on secreting TSH in human body (after F-test, p-value is 2.5e-11)
- 31.31% of total variation of 'TSH' are explained by the explanatory variables
- AIC is 1319.236
- **The model is no better compared to the previous one**
- There could be a quadratic effect of a variable on response

Fitting a Linear Regression Model with a Quadratic Effect of an Explanatory Variable we are interested

- Taking the square of variable ‘T3’ and adding the variable in the first model
- **(Intercept)** **T3** **I(T3^2)** T4 DTSH

  9.29203649 -2.09795684 0.21577626 -0.21658509 0.09819652

  RT3U **Diagnosis(Normal)** Diagnosis(Hyper)

  0.03786722 -7.47929005 -4.12340248
- The variables we are interested in ‘T3’ is significant and has a quadratic effect in the model (after F-test, p-value is 0.0121)
- Surprisingly, 47.48% of total variation is explained by this model
- **AIC value is 1267.519 (lowest among the 3 models)**

What happens if we transform the response variable?

- Let’s make the response variable as ‘log(TSH)’
-

(Intercept)	T3	T4	**DTSH**	RT3U
1.265034568	0.008258517	-0.012768191	0.018345368	0.004640523
Diagnosis(Normal)	**Diagnosis(Hyper)**			
-1.517019063	-1.655956542			

- We cannot conclude that ‘T3’ and ‘T4’ are associated with ‘TSH’
- But **63.57%** of total variation is explained by the explanatory variables in this model

Model Selection

- As we are interested in the most important factors ('T3' and 'T4') for the secretion of 'TSH' in human body, the final model based on 'AIC' criteria is –

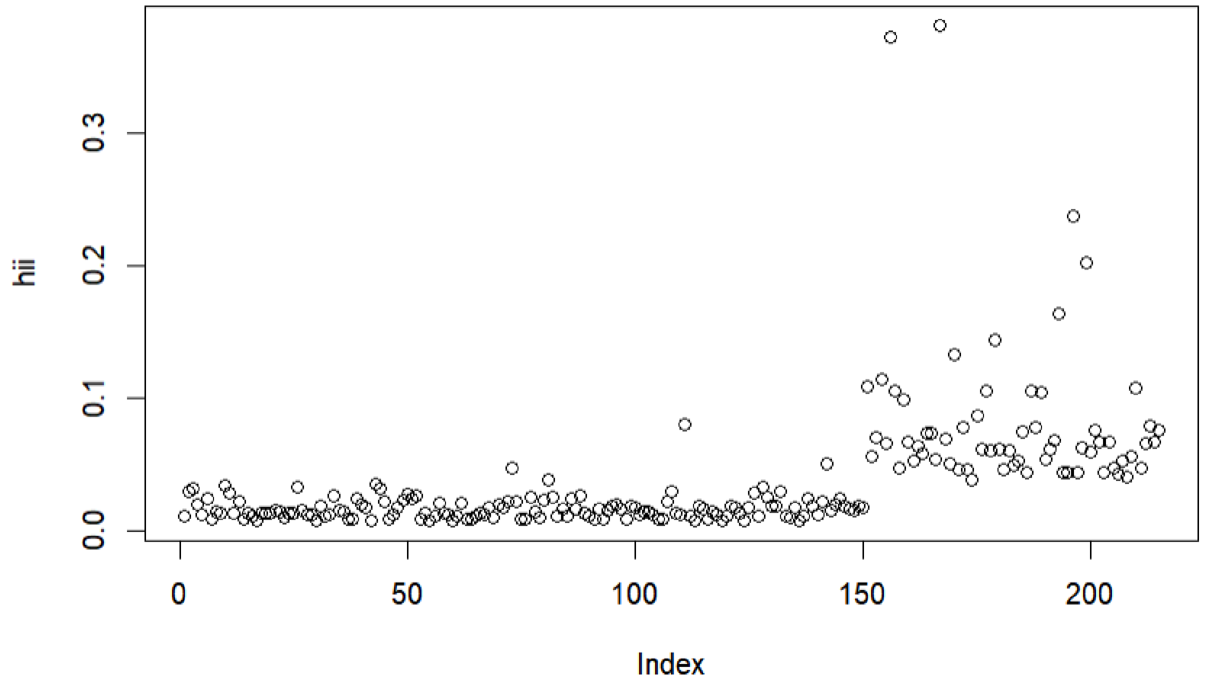
TSH = 9.290 – **2.098T3** + **0.216T3**$\boldsymbol{T3^2}$ - 0.216T4 + 0.098DTSH + 0.038RT3U – **7.489Diag(Normal)** - 4.123Diag(Hyper)

Checking Outliers and High Leverage Points

- After t-test, we found 3 outliers in the data –
- Data Point TSH Level

Data Point	TSH Level
195	56.4
204	32.6
208	41.0

- High Leverage Points (111 151 154 156 157 159 167 170 172 175 177 179 187 188 189 193 196 199 201 210 213 215)



Checking Influential Points

- After calculating Cook's distance, we found no influential points in the dataset
- Output from R - **"named integer(0)"**

Fitting a Robust Regression Model

- As there exists outliers and high leverage points, a robust regression model could give us another set of robust estimates along with confidence intervals –
- **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

Results from Least Absolute Deviation Regression

- All the explanatory variables' coefficients are lying within the lower bound and upper bound range.
- The variables we're interested in 'T3' and 'T4' are significant and are associated with 'TSH'
- Secretion of 'Triiodothyronine' and 'Thyroxine' are associated with secretion of 'Thyroid Stimulating Hormone' in human body

Fitting Ridge or Lasso Regression?

- Sample size is not greater than the number of explanatory variables, it’s not wise to fit a Lasso or Ridge regression

Exhaustive Search

- After exhaustive searching to find which variables are best to run the regression model and taking the BIC criteria, we found 2 variables and the coefficients are –

(Intercept)	DiagnosisNormal	DiagnosisHyper
12.92000	-11.60333	-11.94571

- Both are categories of the variable ‘Diagnosis’

Thank You