Introduction to Statistical Methods for Data Science

Modelling and Analysis of Gene Expression Data

Lecturer: Dr Fei He

Submitted By: Mariam Khalid

Coursework Type: Individual Assignment

Module Code: 7089CEM

Faculty of Engineering, Environment and Computing

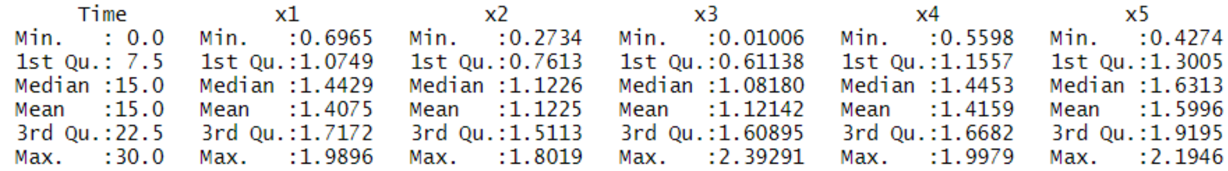
Introduction:

Gene is a unit of heredity which synthesizes gene products which includes proteins and RNAs. The synthesis process evolving to such products is called gene expression. In the subject work, I investigate and explore gene data seeking understanding of the biological process and discover the gene regulation, correlations and prediction of their respective growth over time. For achieving the goal of exploring the data and implementing models I am using **R** language. The report is organized in three different sections as per assignment brief:

- Task1: Preliminary Data Analysis
- Task2: Dimensionality Reduction
- Modelling Gene Regulation (Non-linear Regression)

Task1: Preliminary Data Analysis

The first step I am taking for exploring the given data is to extract general statistics of whole data and individual features. Below image is referring the console output with statistics of data.

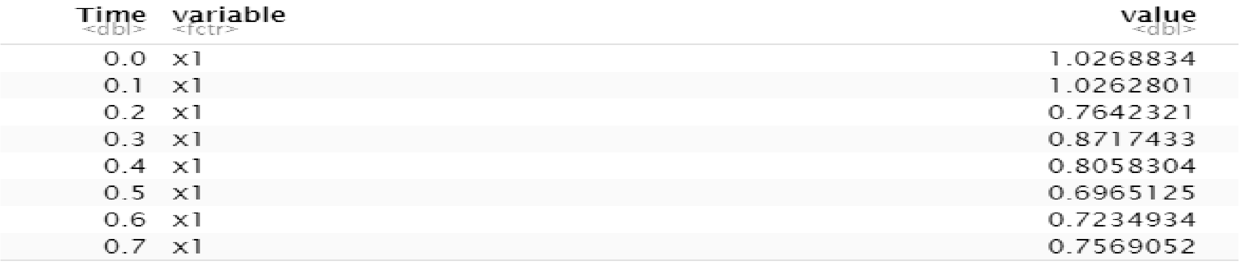


```
      Time            x1              x2              x3               x4              x5
 Min.   : 0.0   Min.   :0.6965   Min.   :0.2734   Min.   :0.01006   Min.   :0.5598   Min.   :0.4274
 1st Qu.: 7.5   1st Qu.:1.0749   1st Qu.:0.7613   1st Qu.:0.61138   1st Qu.:1.1557   1st Qu.:1.3005
 Median :15.0   Median :1.4429   Median :1.1226   Median :1.08180   Median :1.4453   Median :1.6313
 Mean   :15.0   Mean   :1.4075   Mean   :1.1225   Mean   :1.12142   Mean   :1.4159   Mean   :1.5996
 3rd Qu.:22.5   3rd Qu.:1.7172   3rd Qu.:1.5113   3rd Qu.:1.60895   3rd Qu.:1.6682   3rd Qu.:1.9195
 Max.   :30.0   Max.   :1.9896   Max.   :1.8019   Max.   :2.39291   Max.   :1.9979   Max.   :2.1946
```

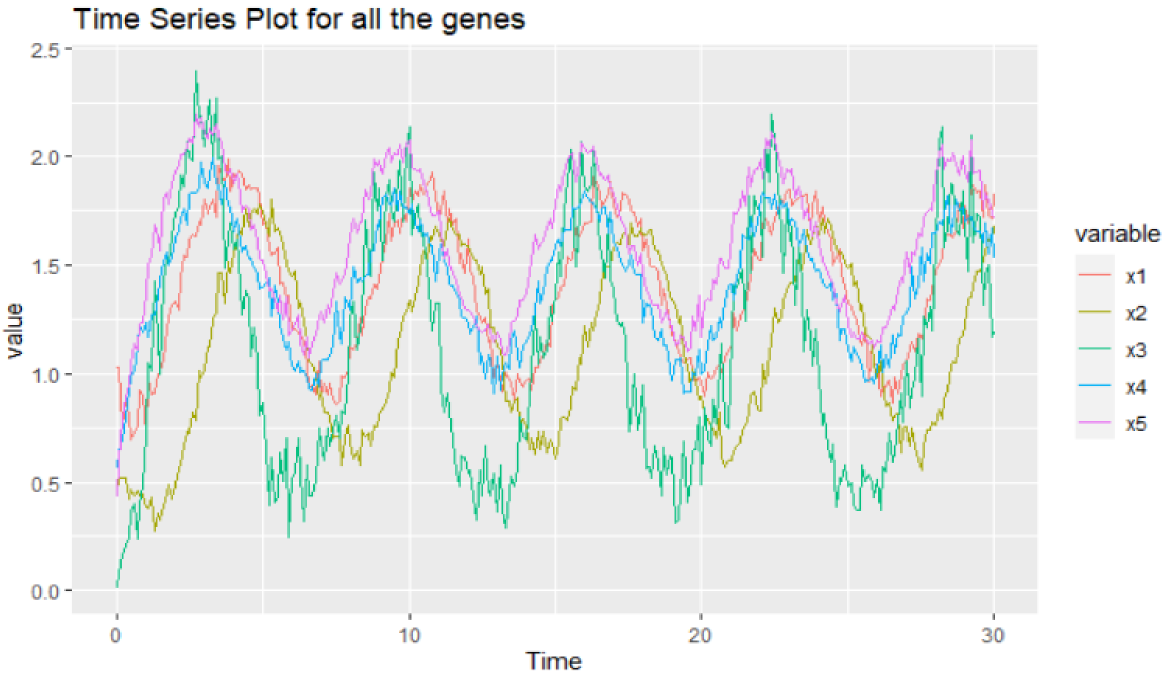
Time Series Plots

To explore the features progression over time and plotting each feature against time feature. I am using **ggplot2 R** graph gallery. Instead of creating plot for each feature, I am using **reshape** R package and melting the data with respect to time as id using the **melt** function for getting single plot with time series progression of each feature.

Following image shows the melted data head and time series progression plots:

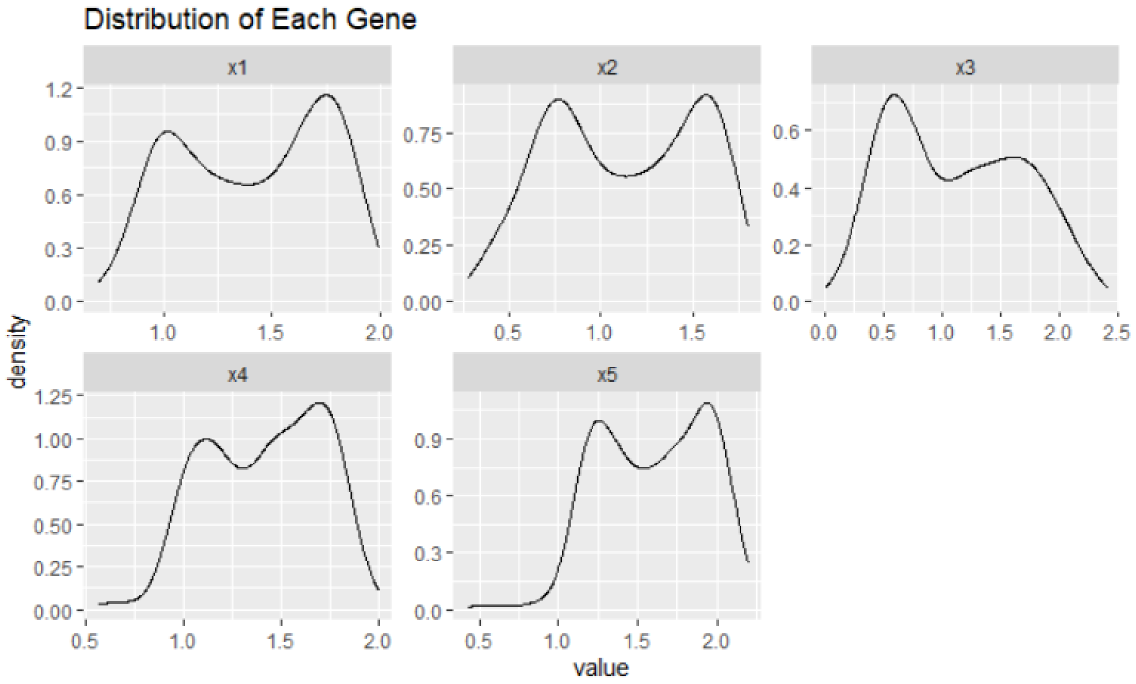


Time <dbl>	variable <fctr>	value <dbl>
0.0	x1	1.0268834
0.1	x1	1.0262801
0.2	x1	0.7642321
0.3	x1	0.8717433
0.4	x1	0.8058304
0.5	x1	0.6965125
0.6	x1	0.7234934
0.7	x1	0.7569052



Distribution of Each Gene

For plotting the distribution of each gene, I first extracted all gene excluding time feature from data frame. Kept only numeric columns and used **gather ()** to convert to key-value pairs and used **facet wrap** to show distribution in separate panels as shown in below image.

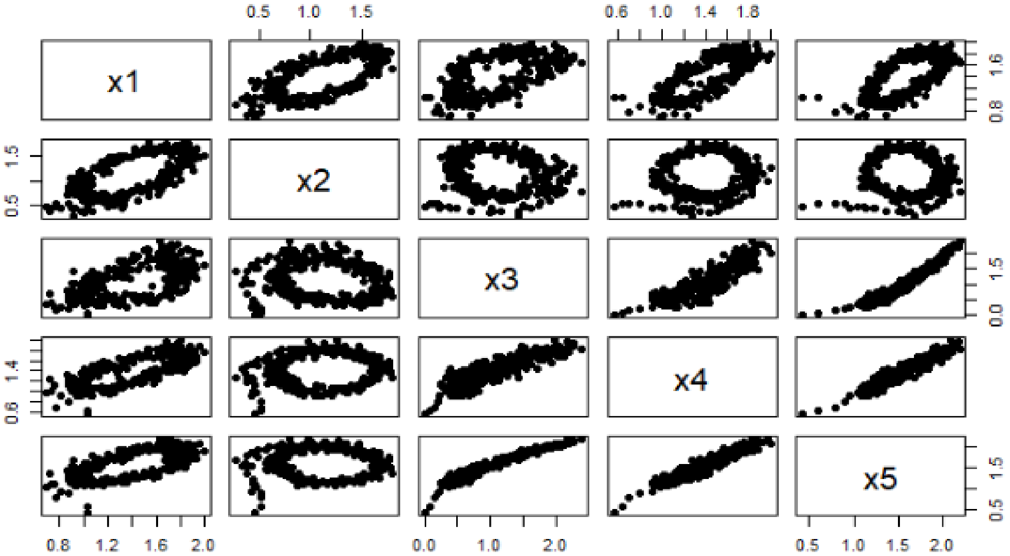


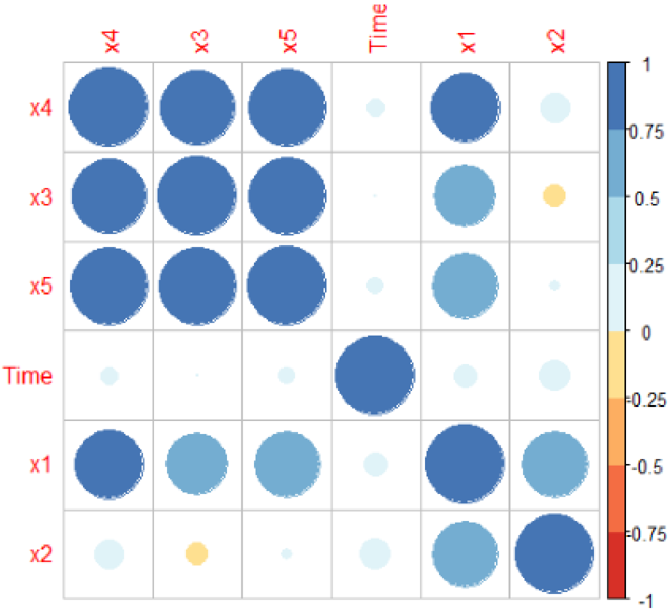
We can clearly view the regulating factor of one gene to another by their respective distribution.

Correlation and Scatter Plot

To review the correlation between genes and regulating factor. We find out the significant correlation between features. We can confirm by below scatter plot and correlation heat map of features that $\boldsymbol{x3, x4\ and\ x5}$ regulates each other significantly.

Scatter and Heat map:

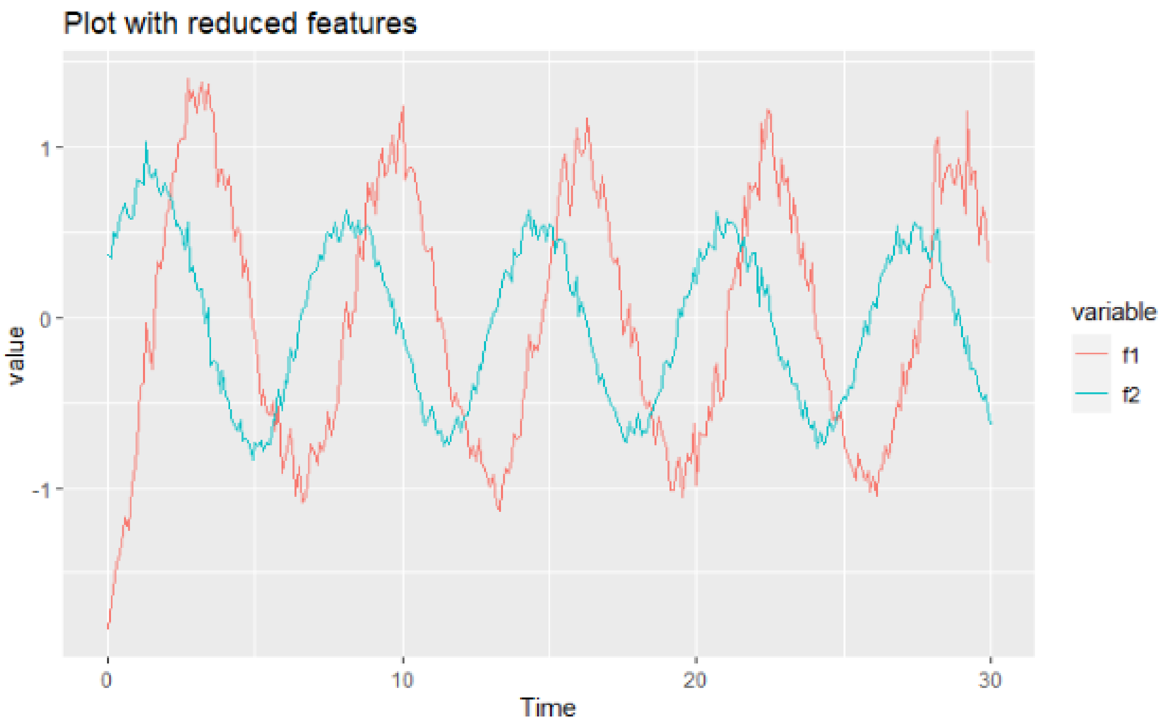




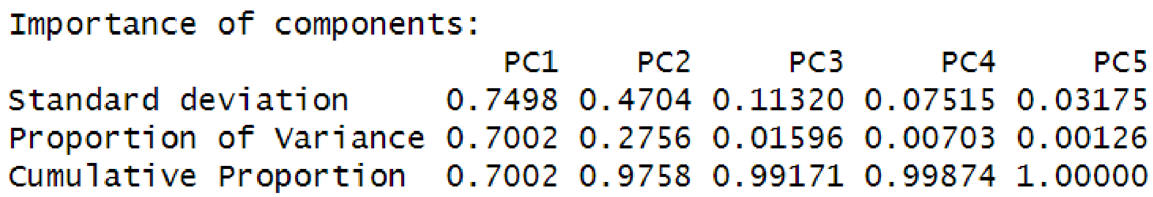
Task2: Dimensionality Reduction

In this section, we will reduce the dimensions of the data i.e. from **5 (gene)** dimensions to **2** dimensions. Here we have used single value decomposition as a method of choice. In R, **prcomp** uses single value decomposition for dimensionality reduction **[1]**.

Firstly we extract all the genes and apply **PCA** using **prcomp** and re-added the time to plot the reduced features. Followed the melt process here as well. Plot with reduced features is as follow:



Importance of principle components is as follow:



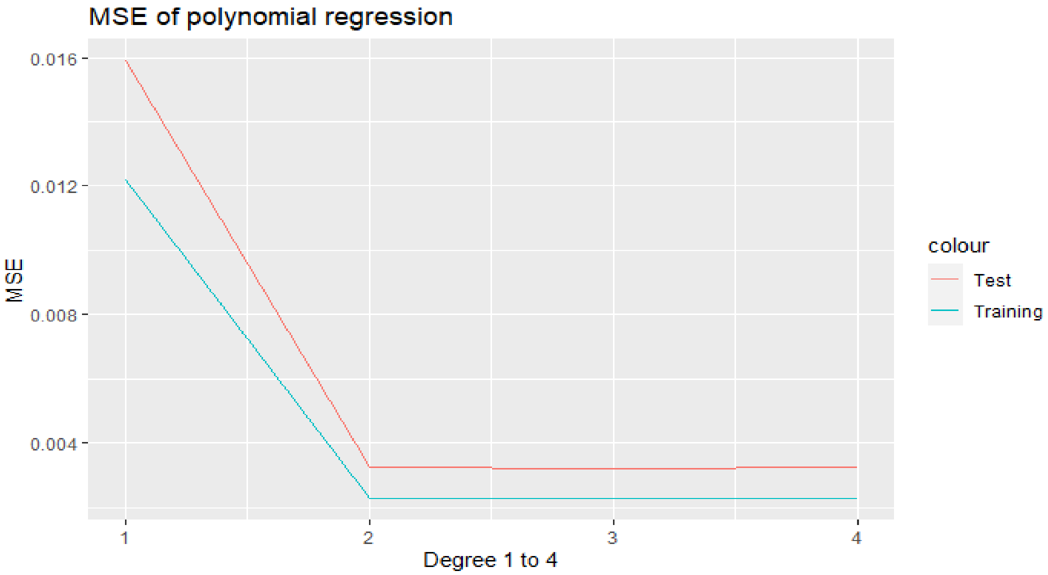
```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     0.7498 0.4704 0.11320 0.07515 0.03175
Proportion of Variance 0.7002 0.2756 0.01596 0.00703 0.00126
Cumulative Proportion  0.7002 0.9758 0.99171 0.99874 1.00000
```

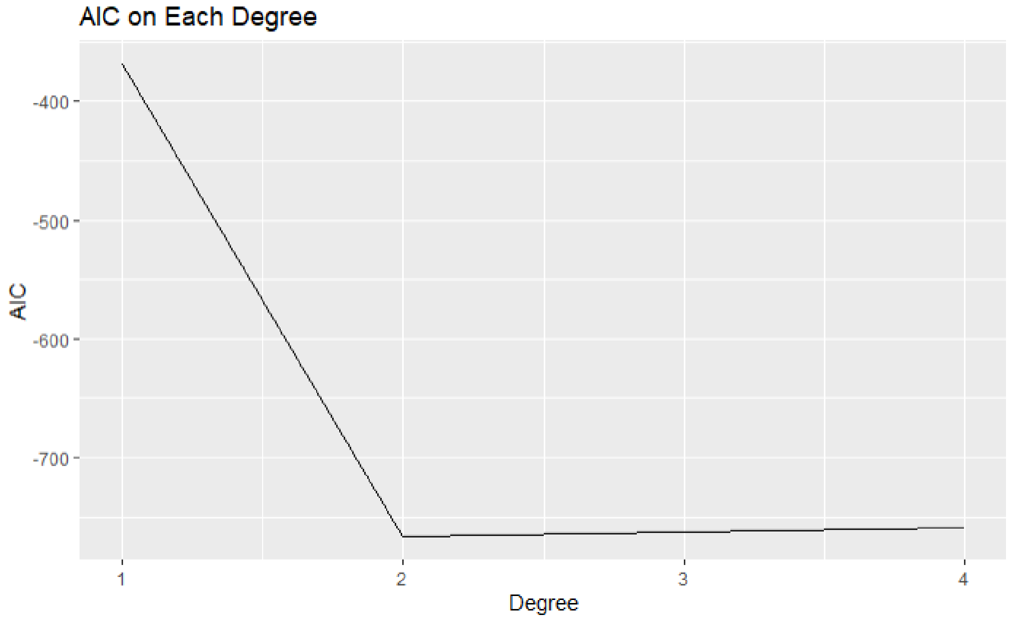
Task 3: Nonlinear Regression – Modelling Gene Regulation

In this section, we will do modeling using nonlinear regression. The first step that we take is to apply test train split on given data. The ratio used is **80/20** for training and testing of model.

To identify and explore different model structures and model with good **MSE**. I have used the iterative approach with finding the respective **AIC** score on different possible combination of polynomial degree.

After the iterations on polynomial degree, I plotted the **MSE** and **AIC** to find the best model equation. Plots are as follow: (code in attached Rmd file)



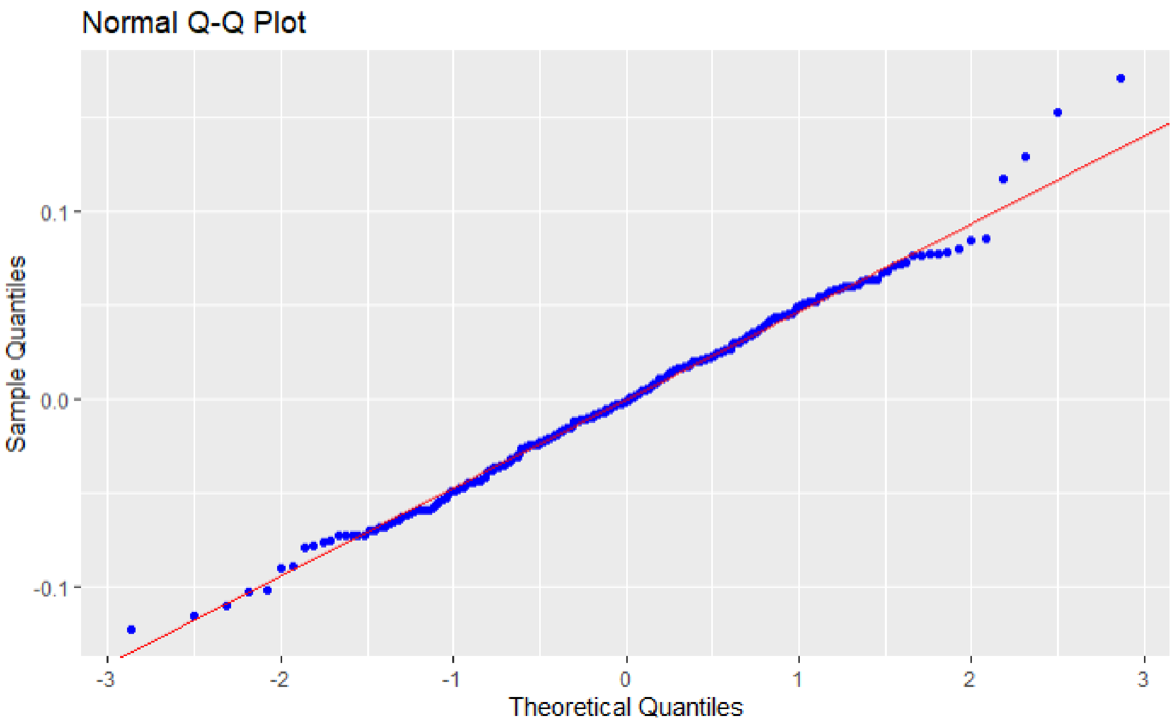


Here we can see that the polynomial degree of two gives the lowest MSE on test and trained data. Therefore we will select the degree of two and the model structure would be:

$$x_3 = w_0 + a_1 x_4 + a_2 {x_4}^2 + b_1 x_5 + b_2 {x_5}^2 + e, \text{ where } e \text{ is Gaussian noise.}$$

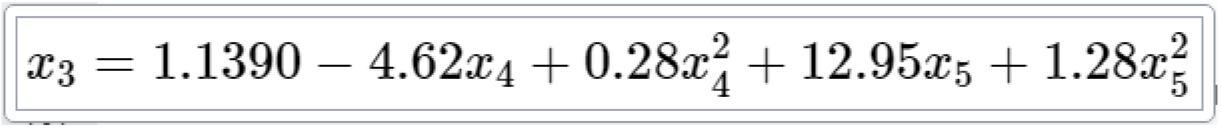
The AIC score is also lowest for model structure of degree **2** and therefore is the best fit.

For validation, first I used **Residuala normality test** to see if our dependent variable have correct functional form. To see of the residuals are Gaussian and for that we do Q-Q plot and we can see from plot below the residuals are near Gaussian. Plot is as follow:



Parameter Estimation

For parameter estimation in R, I used **lm ()** method which uses ordinary least square for parameter estimation and hence the parameters estimated are:

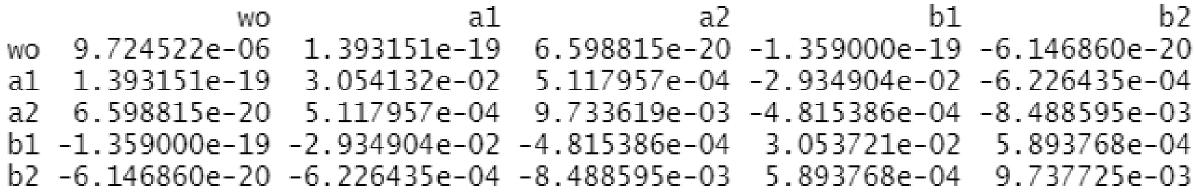


$$x_3 = 1.1390 - 4.62x_4 + 0.28x_4^2 + 12.95x_5 + 1.28x_5^2$$

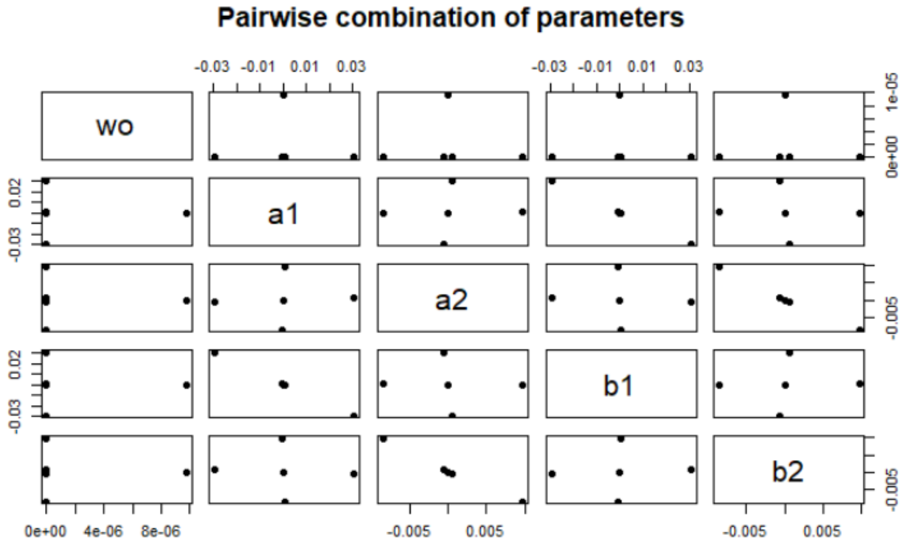
Covariance Matrix and Pair-Wise Combination of Parameters

Using the **vcov ()** we get the covariance matrix of parameters (w_0, a_1, a_2, b_1, b_2). Further we plot the pair-wise combination of parameters.

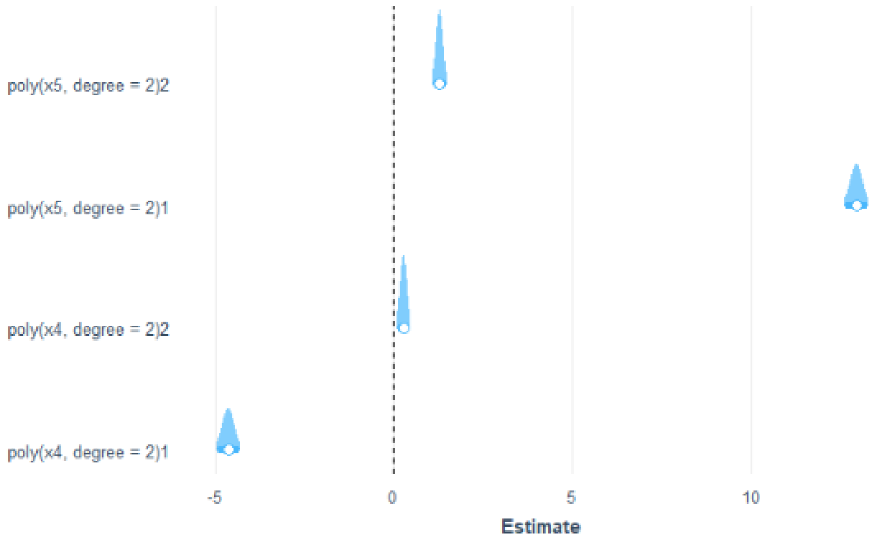
The matrix and plot is as follow:



```
              wo            a1            a2            b1            b2
wo  9.724522e-06  1.393151e-19  6.598815e-20 -1.359000e-19 -6.146860e-20
a1  1.393151e-19  3.054132e-02  5.117957e-04 -2.934904e-02 -6.226435e-04
a2  6.598815e-20  5.117957e-04  9.733619e-03 -4.815386e-04 -8.488595e-03
b1 -1.359000e-19 -2.934904e-02 -4.815386e-04  3.053721e-02  5.893768e-04
b2 -6.146860e-20 -6.226435e-04 -8.488595e-03  5.893768e-04  9.737725e-03
```

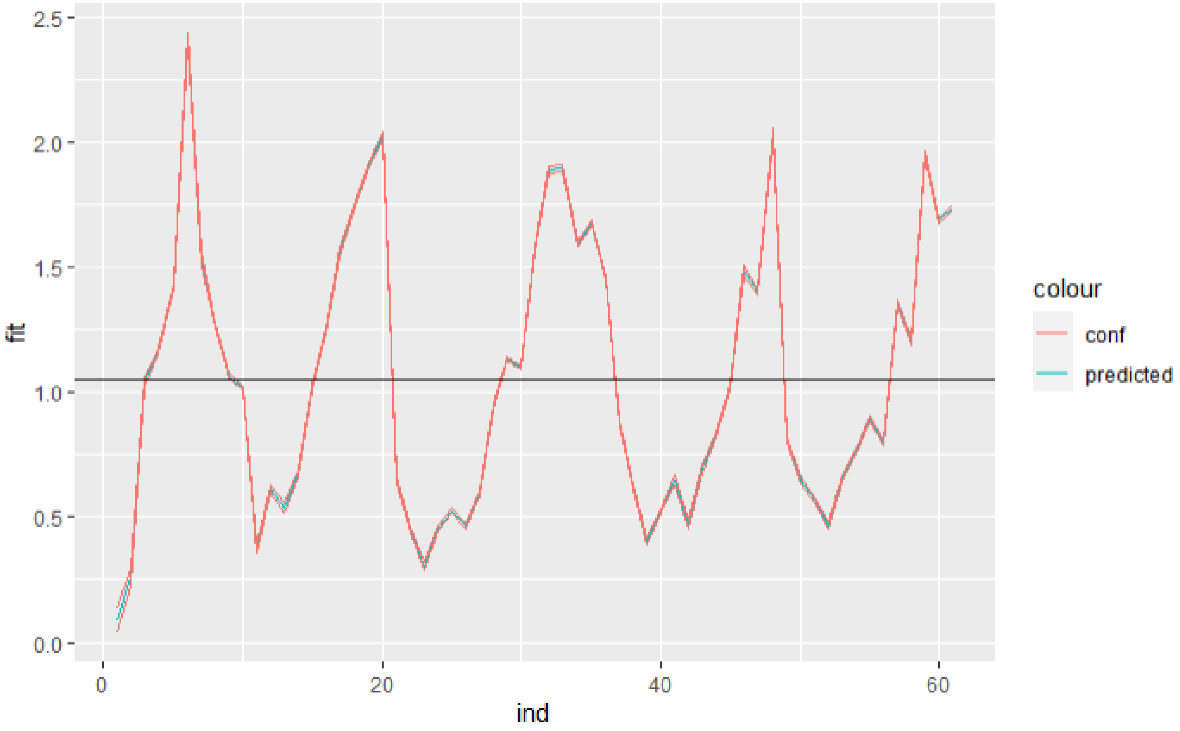


We also plot the degree wise estimate of each dependent variable.



Prediction

We now predict using our model on test data with **95 %** confidence interval. We had testing **61** samples. First we predict using our model and extract the indicator for plotting purpose. Prediction plot is as follow:



The red line shows the confidence within which the predictions from model should lies. The blue line shows the predicted information. The plot clearly shows our prediction coinciding with the confidence mapping. We can say with confidence that our model has predicted good result.

Model Validation

For validation of our model, we will use the k-fold cross validation technique with **10** iterations.

Following is the result of validation.

Linear Regression

301 samples
  2 predictor

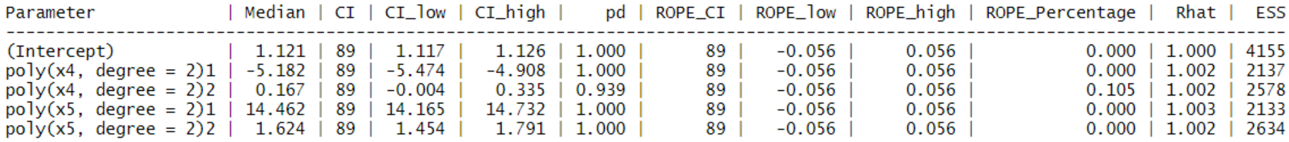
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 272, 269, 270, 272, 269, 271…
Resampling results:

```
  RMSE        Rsquared   MAE
  0.04997768  0.9926197  0.03966265
```

The small value of RMSE shows that our model prediction was right.

Approximate Bayesian Computation (ABC)

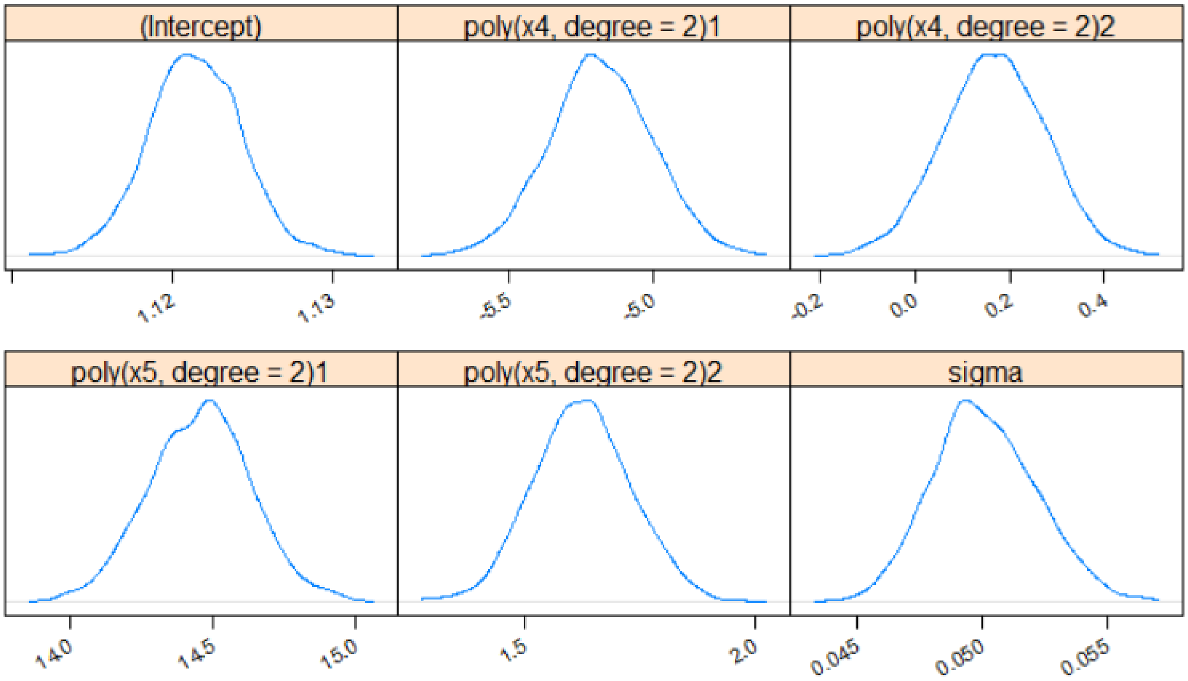
We use the **rstanarm** and **bayestestR** for **ABC** implementation to get posterior probabilities given our model.



```
Parameter             | Median | CI | CI_low | CI_high |    pd | ROPE_CI | ROPE_low | ROPE_high | ROPE_Percentage |  Rhat |  ESS
--------------------------------------------------------------------------------------------------------------------------------
(Intercept)           |  1.121 | 89 |  1.117 |   1.126 | 1.000 |      89 |   -0.056 |     0.056 |           0.000 | 1.000 | 4155
poly(x4, degree = 2)1 | -5.182 | 89 | -5.474 |  -4.908 | 1.000 |      89 |   -0.056 |     0.056 |           0.000 | 1.002 | 2137
poly(x4, degree = 2)2 |  0.167 | 89 | -0.004 |   0.335 | 0.939 |      89 |   -0.056 |     0.056 |           0.105 | 1.002 | 2578
poly(x5, degree = 2)1 | 14.462 | 89 | 14.165 |  14.732 | 1.000 |      89 |   -0.056 |     0.056 |           0.000 | 1.003 | 2133
poly(x5, degree = 2)2 |  1.624 | 89 |  1.454 |   1.791 | 1.000 |      89 |   -0.056 |     0.056 |           0.000 | 1.002 | 2634
```

Posterior description of model using ABC is shown in above image. (Code attached in Rmd file)

Using the **latticeExtra** from R the marginal distribution of 2nd degree model is as follow:



References

[1] Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), pp.37-52.