Sushobhan Parajuli

MTH 4130 : Mathematics of Data Analysis

Professor Giulio Trigila

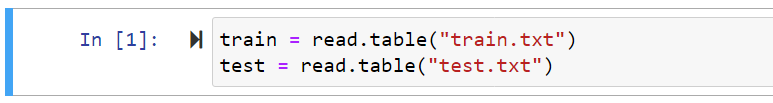
Final Project

12/18/2021

**Linear Regression Model with R**

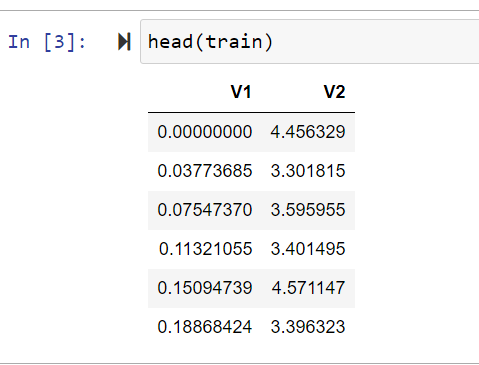
For this project, a file train.txt was given and it contained two columns of data. The first column represented the independent variable that we interpreted as time and the second column was relative to the price of some good, say ice cream. I used train.txt to train the linear regression model. A file test.txt was also given which I used, in the end, to test the linear regression model on. I coded with R in the Jupyter Notebook.

First, training and testing datasets were imported using R code :



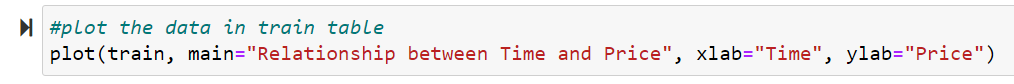
I worked with the training dataset here after till model building.

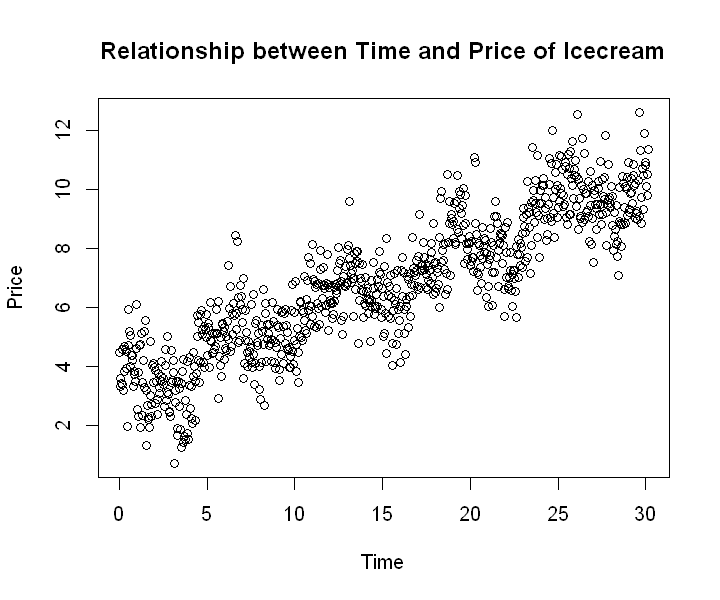
Then, first few entries in the training dataset were inspected:



The first column V1 contains Time, the independent variable and the second column V2 contains Price, the response variable.

A scatter plot showing the relationship between time and price was plotted:

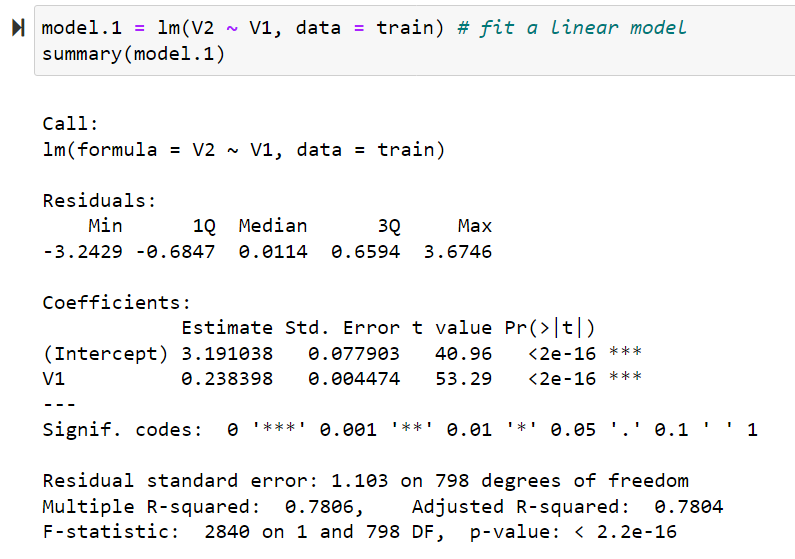




There seems to be a linear relationship between time and price.

**Simple Linear Regression Model:**

A simple linear regression is a way to model the linear relationship between two quantitative variables, predictor variable (independent variable) and response variable. Here, I used R to fit a linear model y = β0 + β1x to the data contained in train.txt and printed a summary table.



According to the above summary table, the estimates βˆ0 and βˆ1 were 3.19 and 0.239 respectively. Estimates βˆ0 and βˆ1 are the sample parameters and are not sufficient to conclude that the entire population follows a linear trend. So, I carried out hypothesis testing to determine the linear trend in the population.

For population slope parameter,

Null hypothesis: β1 = 0 (i.e. there is no relationship)

Alternative hypothesis: β1 =/= 0

I took the industry standard 0.05 significance level.

From the above summary table, the p-value (Pr(>|t|)) of V1 was observed to be 2e-16 which is less than 0.05. Thus the null hypothesis was rejected in favor of the alternative hypothesis. In other words, β1 is not equal to zero and therefore there is a linear relationship between time and price.

Similarly, for population y-intercept parameter,

Null hypothesis: β0 = 0

Alternative hypothesis: β0 =/= 0

I took the industry standard 0.05 significance level.

From the above summary table, the p-value (Pr(>|t|)) of Intercept was observed to be 2e-16 which is less than 0.05. Thus the null hypothesis was rejected in favor of the alternative hypothesis and it concluded that β0 is not equal to zero.

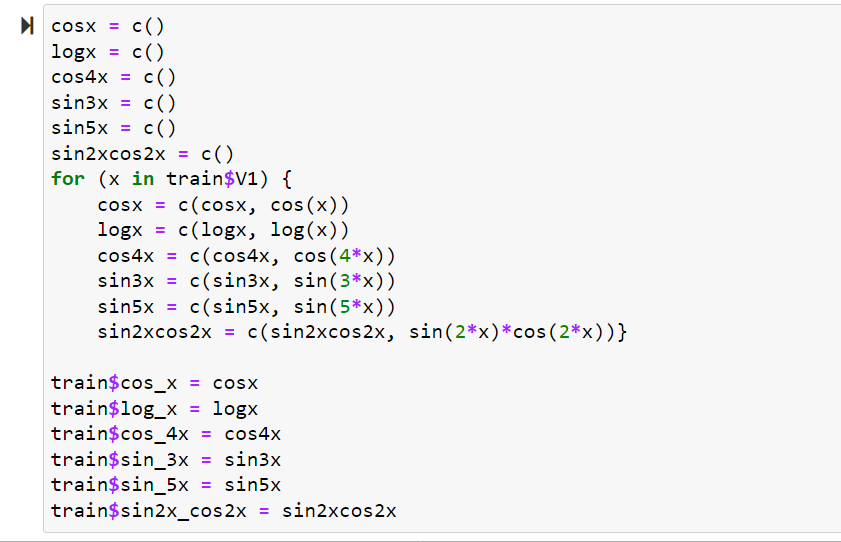
Above simple linear regression model had an adjusted R-squared value of 0.7804. Which can be interpreted as that time explains 78% of variation in price. The R-squared, also known as the coefficient of determination, explains the degree to which independent variable explains the variation in response variable. In multiple regression, adjusted R-squared value is preferred as it includes whether the additional variables are contributing to the model. The adjusted R-squared value ranges from 0 to 1. A higher adjusted R-squared value infers that the model is better.

**Multiple Linear Regression Model:**

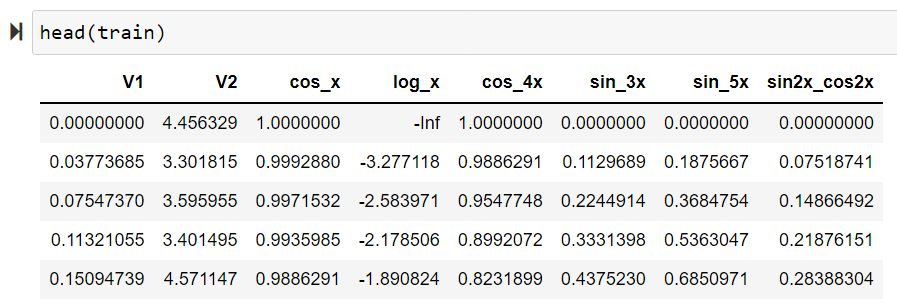
A multiple linear regression is a way to model the linear relationship between one quantitative response variable and more than one predictor variable or feature.

Non-linear features were added in the above simple linear regression model to improve it, as a fellow economist suggested. The features provided were cos(x), log(x), cos(4x), sin(3x), sin(5x) and sin(2x) × cos(2x).

These features were added in the training dataset using following code:

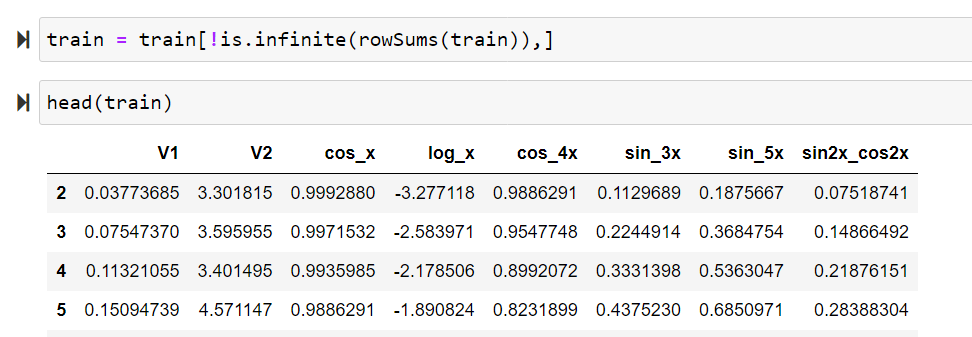


Then first few entries in the training dataset were inspected:



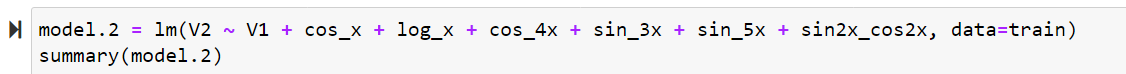
*Note: values for trigonometry functions are in radian.*

The first entry of feature log(x) contained -inf, so the first row was removed with the following code:



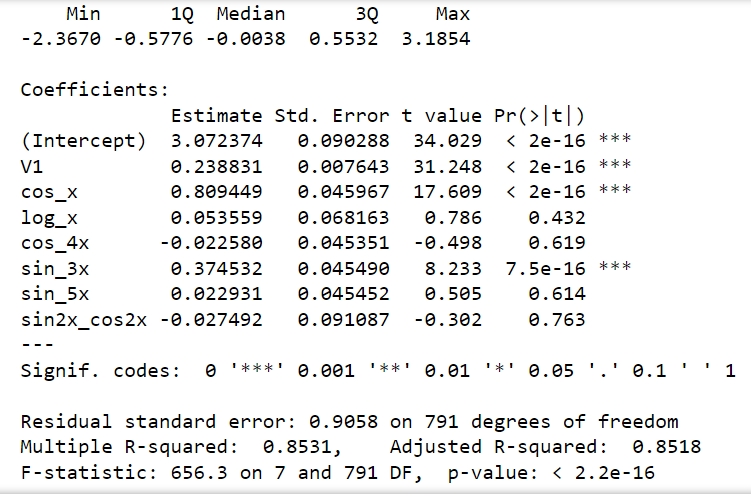
A multiple linear regression model was built using all the features against price and the summary table was calculated to inspect statistically significant features.

The following code was used to build the model:



Executing summary(model.2) printed the following summary table for model.2.

The p-values Pr(>|t|) were inspected and analyzed to select the right features.

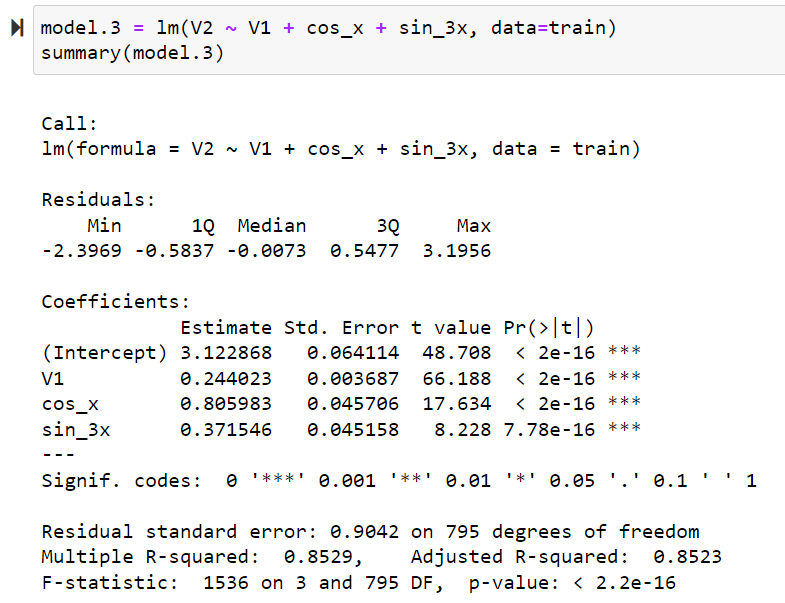


The p-values for features log(x), cos(4x), sin(5x), and sin(2x)cos(2x) were greater than 0.05. It was now understood that the regression parameters were equal to zero and thus the null hypothesis for each feature failed to be rejected. Thus no relationship between these 4 features and price was concluded. Such features were dropped in order to avoid redundancy in the model.

On the other hand, the p-values for features cos(x) and sin(3x) were less than 0.05. Thus these features were concluded to be statistically significant and their null hypothesis of regression parameters being zero was rejected in the favor of alternative hypotheses. These features had a relationship with the price and were chosen in the model to improve the model.

Finally, the multiple linear regression model was built with features V1, cos(x), and sin(3x).

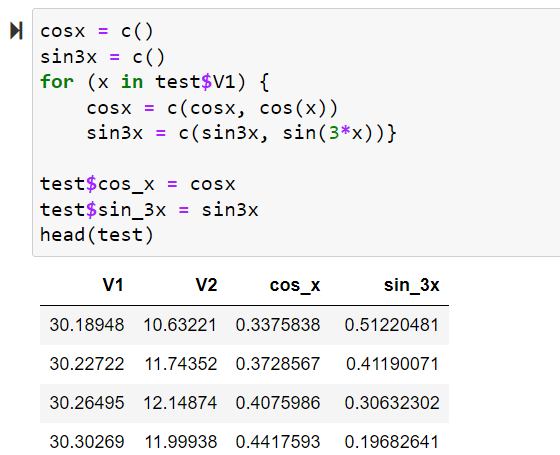
The R code along with the summary table is as follows:



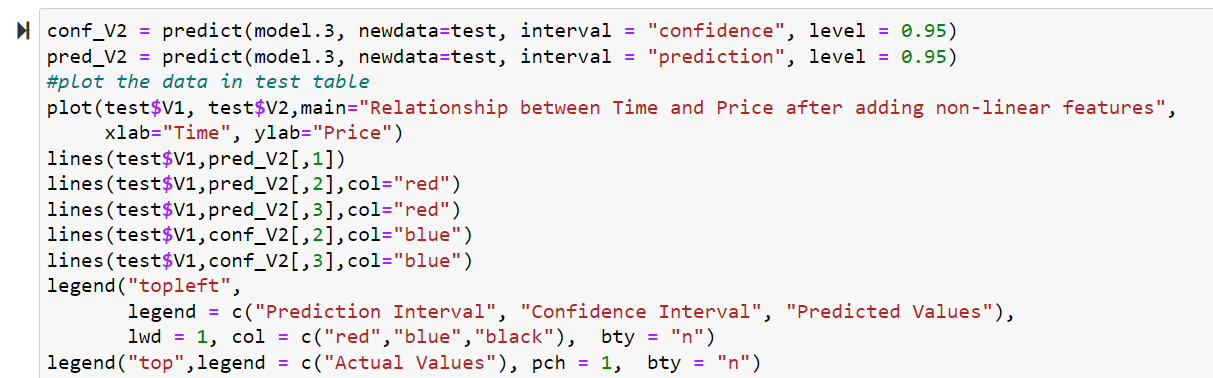
From the above summary table, the adjusted R-squared value was observed to be 0.8523 which is greater than the simple linear regression model. Therefore, this multiple linear regression model was concluded to be better than the previous simple linear regression model.

**Testing the Model:**

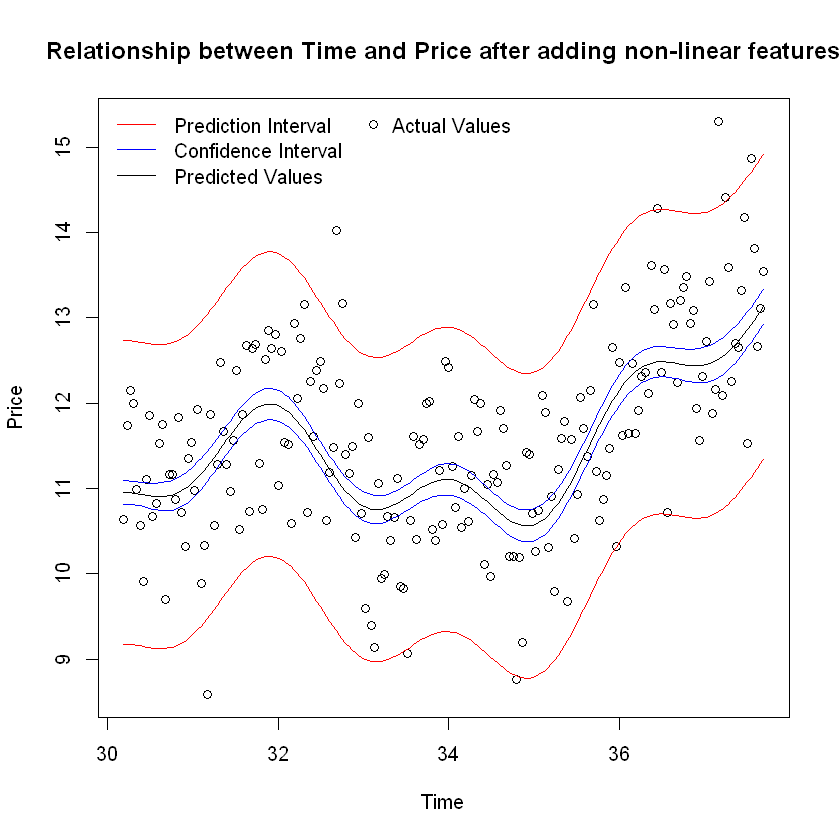
First, the features cos(x) and sin(3x) were added in the test dataset.



Then the predicted values, prediction interval and confidence interval were calculated with the following code:



Following is the plot for actual values, predicted values line, predicted and confidence intervals:



From the above plot, we can be 95% confident that the mean price of all ice creams at a given time falls within blue lines. And we can be 95% confident that a single observation of price at a given time falls within red lines. The line for predicted values is within the confidence interval (blue lines) and most of the actual values are within the prediction interval (red lines). Thus, there is a good agreement between the model (model.3) and the data in the test dataset.

The regression model, with the integration of non-linear features cos(x) and sin(3x), is better than the simple linear regression model and can be used to predict the price of Icecream for a given time.