**IDS 575 - Project Final Report**

**Quick Summary:**

The goal of this project was to understand how data analytics is performed in real world processes by using various statistical models. Personally, this project taught me a lot from the first of the steps, i.e.; data selection, data processing to the part of inferring what a particular model can or cannot do for a certain kind of dataset. I have worked on both of my project proposals, and learnt the parts of data processing from the first one and the parts of statistical modelling from the second one.

**Introduction:**

I have chosen two ideas for my project, one with a big enough data set, for which I have decided to build my own dataset, and the other one with a simple data set for which I would build a predictive model.

* *Proposal 1*: Drug Prediction for a patient using multivariate regression based on vital signs and clinical notes details from the EMR data.
* *Proposal 2*: ​ Prediction of time taken for filling a vending machine based on product already stocked​ ​ and​ ​ the​ ​ distance​ ​ to​ ​ be​ ​ travelled.

**Project Learnings:**

**Data Processing for Idea 1:**

The project had 13 variables and 2 target variables and 500 observations were built manually for the data set. All the factors needed for this dataset were collected, like for example, it was managed to figure out the kind of clinical notes, observations, drugs, etc needed for the dermatology practice that was chosen for the prescription prediction. Data was cooked up for the variables like Blood Pressure, age, pulse, respiration rate, etc. Thus the raw data set was prepared with all the variables necessary.

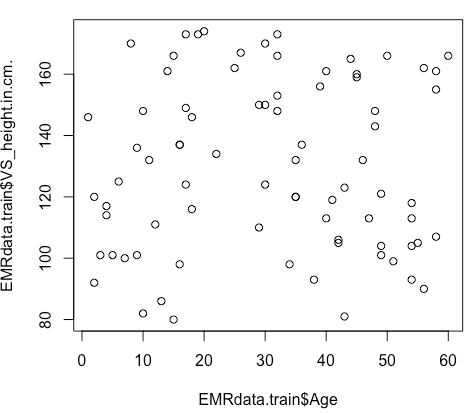
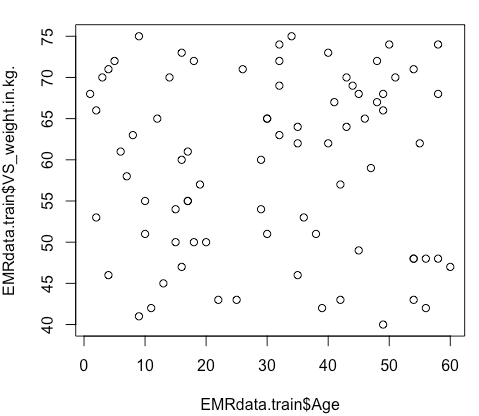
Once this was done, the missing values and outliers needed were also included for each of the variables. A complete preprocessing was done over this dataset. Data was cleaned by means of imputing missing values, removing outliers, and eliminating unnecessary variables. A data extraction step was applied by taking only the subset of variables that were needed for the prediction of target.

Also, categorization of the various variables into various bins was done for doing a categorical analysis. In detail, the values were converted to the categorical values like low BP, normal BP and high BP, similarly for others. This was done for a meaningful prediction of what drug has to be used for these levels of vital sign observations made.

The plan was to make use of multivariate linear regression model since this particular data set involves more than one predictor variables. The best prediction model was supposed to be determined by using the least squares method and should have been validated by using the test data.

**Results and Learnings:**

The entire data building and data preprocessing itself took a lot of time because it involved a lot of efforts to collect the necessary data first. The data set belongs to the healthcare sector and internet could provide information only until a certain extent. For example, it was easy enough to find the diseases, symptoms and the corresponding medicines advised for them, but it was difficult to find the dosages that they were supposed to be taken in. For a data set to give a good predictive model, it is necessary that it has to be meaningful. The data set I built was a bit random, because I used random functions in excel sheet to generate the random Blood Pressure, age, pulse, height, weight, etc values. These resulted in a not so good basic plot of the data.



The above two plots show the age versus weight and age versus height data sets. These are highly skewed data plots with no meaningful interpretation. In reality with age, height and weight increase and have to have a linear relationship. But in my dataset they are completely random and have no correlation whatsoever. Thus one lesson to take is that every and any data set cannot be cooked randomly especially in the condition that the variables are dependent on each other.

Hence, it was difficult at every step for most of the variables as they are cooked and they had to be modified at every such stage. With this, it was not possible to build the model for my idea 1.

**Statistical modelling for Idea 2:**

The second idea was simple, the goal in this one was to use a simple dataset but to work and learn about statistical modelling. This is the reason why such simple project was chosen. The data set consisted of two variables and one target variable, namely Products and Distance were the two variables via which we would be predicting the time taken to fill the vending machine, which is the target variable for our project.

As there are only a few observations, it just needs eye balling to understand that it does not contain any missing values and thus requires no procedure to work on them. The next step was to see if there were any outliers in the given data set. Once the data was plotted it was seen that it was not that skewed either, so it was understood that there were no outliers from the data. Also the data provided for the variables like Products, Distance and Time was pretty straightforward and is appropriate for building a linear model as planned. Hence there is no necessity for any kind of data modifications in this data set. Thus no data cleansing, no data editing, no data reduction was applied over this data set.

**Various regression models used and their results:**

Multiple models were applied for the chosen data set to understand how the data set was and what patterns it followed. The data set was divided into training and test data sets in the ratio 80:20 percentages. The training set was used to build the model and the test data set was used to validate the prediction of the model that was built. That was how the RMSE or the root mean squared error value was computed.

***Baseline Model:***

This model is being used as a metric to compare how my actual intended models would work and predict for the data set. This gave an **RMSE** value of **7.134603**. This would act as the reference with respect to the Root mean squared error value for the rest of the prediction models used further. It is expected that the models that will be used next will have to perform better than this baseline model.

***Linear Regression Model:***

For the initial interpretation, a linear model was built between the target variable and each of the independent variables. One plot was between number of products to be filled and the time taken to fill them. It is seen that the data is a bit skewed but it still had the scope of a strong linear relationship that can be obtained between both the variables. Upon plotting a linear model and drawing the summary of this model the model information was obtained. It was deduced that the relation between both the variables would be **Time = 3.0354 + 2.1790\*Products**. The plotted model helped in inferring a few details like the QQ plot showed that the samples seem to be closely distributed along the ideal normal distribution with mean and standard deviation of the sample.

Similarly a linear model has been plotted for the target variable Time and the independent variable Distance and the relationship was deduced as **Time = 4.961841 + 0.043885\*Distance**. The interpretation from the QQ plot was pretty much the same as above, that the samples seem to be lying along the ideal normal distribution with mean and standard deviation of the sample.

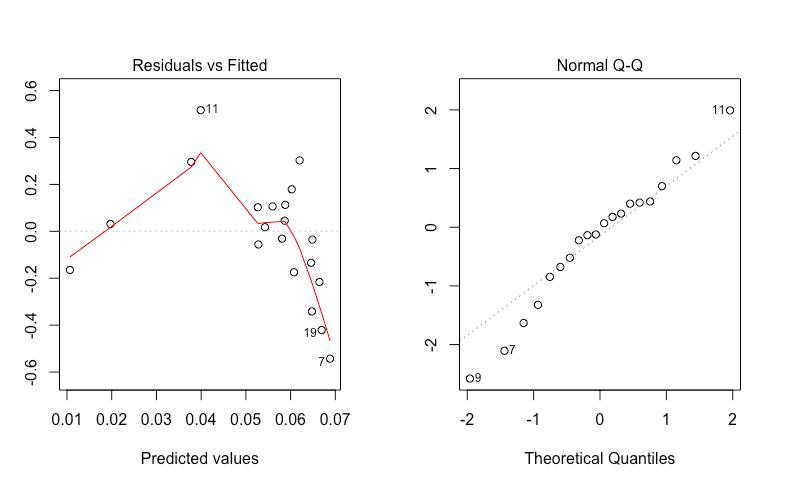
***Multiple Linear Regression Model:***

As a next thought, it was assumed that the target variable Time would be dependent on both the variables Products and Distance which is why we would be needing a multiple linear regression method as our model for this dataset. Upon plotting a model across both the variables, the relationship between the target variable and both the independent variables was seen to be **Time = 2.189193 + 1.613891\*Products + 0.014640\*Distance**. The **RMSE** value was found for this model over the test data and is valued to be **0.7624004**. The QQ plot for this model depicts the fact that the linear model is a decent one, because most points seem to drift away from the ideal normal distribution line shown.

By using the vif function in R, the multi collinearity of the multiple regression model is found out and the values are seen to be same for both the variables Products and Distance and is 3.32684. Hence we need to take the interaction factor as well. Now a multiple linear regression model is plotted for the combination of the independent variables and the interaction factor which is seen as **Time = 7.4309462 + 0.9954094\*Products + 0.0042695\*Distance + 0.0008022 \*(Products\*Distance)**. The **RMSE** value was found for this model over the test data and is valued to be **1.466159**. The QQ plot for this linear model also shows that the linear model is good, because again, only few of the points are seeming to lie far away from the ideal normal distribution line shown in the plot. The model does not seem to be that a decent prediction model also because of its little high RMSE value.

***Gamma Regression Model:***

As part of the mid project future plan, the gamma regression model was also implemented to determine better predictions for the dataset. A plot was built for this gamma regression model across both the variables, the relationship between the target variable and both the independent variables was deduced to be **Time = 7.300e-02 - 1.998e-03\*Products -1.652e-06\*Distance**. The **RMSE** value was found for this model over the test data and is valued to be **2.05024**. The plots can be seen as below. The QQ plot for this model depicts the fact that the linear model is a decent one, because even more points seem to drift away from the ideal normal distribution line shown.



**Learnings:**

Upon trying multiple varieties of linear models for the chosen dataset, it was seen that all the models were performing in their own manner giving predictions with different accuracies. The plots of different models told different stories and gave different inferences, some as expected, some unexpected. The RMSE values for all the models that were used are tabulated below:

|  |  |
| --- | --- |
| **Model** | **RMSE values** |
| Baseline | **7.134603** |
| Multiple Linear Regression | **0.7624004** |
| Multiple Linear Regression with interaction factor | **1.466159** |
| Gamma Linear Regression | **2.05024** |

**Conclusions:**

From above values, it can be seen that the best RMSE values obtained are in the order of Multiple Linear Regression first, followed by Multiple Linear Regression with interaction factor and Gamma Linear regression. It can clearly be seen that all these models have performed better than the baseline model which was our initial reference, as expected. It is also seen that the best model according to our analysis so far for this data set, is the multiple linear regression without interaction factor. This can be understood from the RMSE values as well as the plots. In conclusion, it can be said that in this particular analysis, multiple linear regression has performed the best but more analysis with many more models can lead to different results.

**References:**

[**http://www.statsci.org/data/general/softdrin.html**](http://www.statsci.org/data/general/softdrin.html)

**Montgomery, D. C., and Peck, E. A. (1992). *Introduction to Regression Analysis*. Wiley, New York.**