DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICAL SCIENCES
UNIVERSITY OF DELHI

PROJECT REPORT

STATISTICAL ANALYSIS OF SOME DATA USING TECHNIQUES FROM LINEAR MODELS, MULTIVARIATE ANALYSIS AND NON-PARAMETRIC INFERENCE

(AS PART OF THE COURSE PAPER 305: PRACTICAL – III PART B: PROBLEM SOLVING USING SPSS – I)

UNDER THE GUIDANCE AND SUPERVISION OF

MR. ABHISHEK K. UMRAWAL AND DR. RANJITA PANDEY
(ASSISTANT PROFESSORS)

SUBMITTED BY:

VISHAL KUMAR
M. SC. STATISTICS, SEMESTER – III

DATE: NOV 16, 2015
PLACE: NEW DELHI

ABSTRACT

To initiate a new way of conducting practicals, case study approach was introduced in Statistics Department of University of Delhi. The main purpose of this approach is to analyse the data using SPSS (Statistical Package for Social Sciences), to draw inferences and make further decisions.

Case study has been divided into 4 parts:

1. **Exploratory Data Analysis:**

This is the initial phase of data analysis, exploratory techniques like histogram, box-plot, stem and leaf, Quantile-Quantile plot and Kolmogorov-Smirnov test were used to explore "iris" dataset. As this was just an exploratory phase no hard inferences were drawn from this part of the case study.

2. **Regression Analysis:**

In this phase model building techniques were used to predict the average miles/gallon of the car based on some available car specifications in "mtcars" dataset. For this a linear model was build, then it was subjected to go through model diagnostics like multicollinearity, autocorrelation and heteroscedasticity, a parsimonious model was selected and finally the most important, normality assumption of errors was tested.

3. **Classification Problem:**

This phase is an important part of machine learning technique used to classify data. Techniques like Naïve-Bayes classifier and logistic regression were used to filter spam emails. Skull types were predicted using logistics regression and discriminant analysis. Sentiment analysis was performed to predict whether a candidate will win the US elections or not? Multiclass classification was used to predict the flower species of "iris" dataset and was compared to the predictions made by logistic regression.

4. **Non-Parametric Inference:**

In this phase no information about the distribution of data was available. Various techniques like Wilcoxon Sign Test, Wolofowitz Run Test, and Mann – Whitney Wilcoxon U – Test, Kolmogorov – Smirnov Test, Chi – Square Tests like Karl Pearson's Goodness of Fit, Mc – Nemar's Test, Cochran – Mantel Heenszel Test were used to analyse data.

ACKNOWLEDGEMENTS

I have given my best efforts to make this a good report. However, it would not have been possible without the kind support, guidance and constant supervision of our professors **MR. ABHISHEK K. UMRAWAL AND DR. RANJITA PANDEY** and I would like to express my gratitude towards them for providing necessary help regarding the concepts and techniques used in this case study.

Lecture notes released by **MR. ABHISHEK K. UMRAWAL,** provided the necessary knowledge to tackle the given problem at hand and to understand the basic "why" concept behind any test and technique used in this case study.

REFERENCES

1. Lecture Notes for Paper 305 (B): IBM SPSS Statistics of M. Sc. Statistics Semester–III, 2015 at Department of Statistics, University of Delhi, New Delhi, INDIA by **MR. ABHISHEK K. UMRAWAL.**
2. R Programming for Data Science by Dr. ROGER D. PENG.
3. Exploratory Data Analysis by R by Dr. ROGER D. PENG.
4. Statistical Inference for Data Science by Dr. Brian Caffo.
5. Regression Models for Data Science in R by Dr. Brian Caffo.