ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA

DEPARTMENT OF STATISTICS

SYNOPSIS

REGRESSION ANALYSIS OF CAR PRICE DATA

Name- Arnab Roy

Roll No.-411

Supervisor - Dr. Ayan Chandra

Regression Analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables. Here, a secondary data on price of the car (in dollars) in America and data on their features are collected. Price of the cars is taken as the response variable for the regression analysis of the data.

Given this data in hand, the primary objective is to find –

(i) Which variables play significant role in predicting the price of a car
(ii) How well those variables describe the price of a car

In other words, here our main objective is to extract maximum knowledge from this data and to fit a regression model to predict the price of the cars.

To fulfil these purposes, at first, categorical variables are analyzed with the help of suitable diagrams like boxplot or barplot. To analyze the numerical variables or predictors, scatterplots of each predictor against the response variable (i.e. price) are drawn and from the plots, the predictors which do not seem to have significant negative or positive correlation are considered as dummy variables and discarded. Then, among the remaining predictors, multicolinearity is checked by calculating partial correlation coefficient between two predictors ignoring the effect of the other predictors. If the presence of multicolinearity is detected between two predictors, then we remove one of the predictors. Now, if necessary we reduce the skewness of the remaining predictors and response variable by taking suitable transformations and fit a suitable regression model of price on the predictors. Then, by checking the p values, we take the predictors which have a significant effect on the response at α level of significance and rebuild the regression model. After rebuilding the model, we plot the residuals against the fitted values of the response variable and check whether the plot is random or not to justify the fitted regression model.