PREDICTING LIFE EXPECTANCY USING REGRESSION

Regression is a technique which is used to predict a target variable using a single or a set of independent variables. Thus, the relation between target and predictor variables can be analyzed using regression. Simple linear regression is used when there is a linear relation between the dependent variable and a single independent variable. In the given data set, there are more than one independent variables. So, Multiple linear regression strategies, also called as multivariate regression models can be used to predict the life expectancy of the people in various countries depending on the indicator variables of world bank.

The linear regression line for n explanatory variables x1, x2, ..., xn is defined to be :

$Y= \beta_0+ \beta_1X_1+ \beta_2X_2+\ldots. + \beta_nX_n.$

This line describes how the target Y changes with the explanatory variables. There are some assumptions of a multiple linear regression like:

- The line of best fit passing through the data points is a straight line.
- The data follows a normal distribution.

The given dataset requires to analyze and predict the life expectancy of the individuals in different countries based on several factors like total population, mortality rate, literacy rate, employment, health expenditure etc. Therefore, a multiple linear regressor could be built by training the given data and predictions could be made on the life expectancy of the people in the countries which are not listed in the given dataset. Based on this analysis, it would be more convenient for a country to identify the factors which are resulting in a lower life expectancy and effectively improve the circumstances.

The model to assess the life expectancy is built followed the below procedure.

- The dataset is loaded into R using read.csv () function. The libraries like dplyr, mice which are useful in building a regression model are loaded into R.
- The data set is checked for missing values. Missing values cause certain problems on the decisions taken on analyzing the data. However, excluding missing data from the evaluation causes loss of data and results in inaccurate

conclusions. The ideal way to deal with missing values is to impute them with certain values which complete the dataset without disturbing the original structure.

- In this analysis, MICE (Multivariate Imputation by Chained Equations) imputation technique is used to replace the missing data. This feature automatically detects the columns with missing values and imputes them accordingly using methods such as predictive mean matching, logistic regression, depending on type of the variable.
- When the data is free of all the missing values, correlation among the variables is calculated using corrplot (). Correlation coefficient which ranges between 1 and -1 gives the relation between two variables. If the value is close to +1 or -1, it means that two variables are highly correlated and thus one variable can be removed from the analysis.
- In the given dataset, the variables whose correlation coefficient is +/-0.8 are considered as highly correlated.
- The correlation coefficient between male and female mortality rates is 0.94. Hence, male mortality rate is not considered in building a model.
- Like wise infant mortality rate, GNP per capita, birth rate crude are highly correlated with female mortality rate , GDP per capita and access to electricity respectively.
- Therefore, the former variables are ignored from the analysis.
- Multiple linear regressor model is built on the remaining variables after eliminating the correlated variables using lm () function.
- Summary of the model provides information about parameters like F-statistic, p-value and adjusted r squared which help in analyzing the model. The adjusted R squared value is used to compare the goodness of the fit of model. For the model built in our analysis, the adjusted R squared value is 0.92.
- We also get the significant variables in the summary of the model. The model can be further simplified by removing the least contributing variables one at a time and checking the adjusted r squared each time.

- In the analysis, adjusted net national income, population growth, unemployment, GDP growth are identified as least significant variables and can be eliminated from the model.
- The adjusted R squared after removing the insignificant variables is 0.92. It can be observed that there is no change after removing the less contributing variables and the model is further simplified.
- We can now apply this model to analyze and predict the life expectancies.

We can evaluate the model built on the given data set as follows:

- The given dataset is divided into training and test sets using split() function of the catools package in R
- The usual split ratio is 80% of the data will be training set and the remaining 20% will be the test set.
- The model is trained using the regression model on training data and in the summary of the model, the adjusted r squared value is 0.90.
- The predictions are made using the regressor test set and the predictions are evaluated using RMSE(Root Mean Squared error) from Metrics package in R.
- The RMSE is considered as good metric for numerical predictions. The value obtained in the analysis is 1.42. Small values of RMSE denote good fit.

The same model can be used to predict the life expectancy of the citizens of any country. In the life expectancy data 2 file , there are 11 countries for which the life expectancy is to be predicted. This dataset also contains missing values which are imputed using mice package and by replacing certain missing observations with mean. The dataset is further simplified by eliminating the least significant variables.