Introduction

Monthly expenditure of families living in a neighborhood was collected and stored in a dataset for experimental purposes. The data is stored in the dataset 'Household data' in an .xlxs format. Along with the variable expenditure, the dataset also contains other four features also, features that may affect the expenditure of the family. The dataset is composed of 50 observations of 50 families with 5 different features or columns. So we need to set up a perfect regression model to explain the monthly expenditure using other variables.

Project Objective.

The dataset has information on monthly expenditure, annual income of the household, monthly income, household size (number of members in the household), and monthly EMI. So we need to set up a perfect regression model to explain the monthly expenditure using other variables or features in the dataset. This regression model can further use for most efficient prediction of monthly expenditure based on the other variables selected in the model. Here we need to set up a best-fit regression model based on a regression equation and need to calculate the fittest and balanced coefficient values corresponding to the regressor variables.

Assumptions

We assume that the data provided the dataset Household Data is non-biased, error-free, contain no missing values and no placeholders. The data is related to the families living in a locality or a neighborhood. We also assume that the dataset will be reflective of the reality and the features in the dataset are linked to the financial characteristics of the families. The 50 rows of the dataset correspond to 50 unique families and it is also important to assume that the features in the dataset are chosen significantly and contain relevant data.

The following data dictionary is considered for 5 features in the dataset.

|  |  |  |
| --- | --- | --- |
| Sr. no | Column Name | Description |
| 1 | Annual Income  ($1000s) | Annual income of the household or the family in 1000 dollar |
| 2 | Monthly Income ($1000s) | Monthly income of the household or family in 1000 dollar |
| 3 | Household  Size | Number of members in the household |
| 4 | Amount  Charged ($) | Amount charged is the monthly expenditure of the household in full figure. |
| 5 | monthly EMI ($1000s) | and monthly EMI in 1000 dollar |

Assumptions on Regression Model

Regression is a parametric technique, so it makes assumptions. The assumptions we make are as follows.

1. There exists a linear and additive relationship between dependent (DV) and independent variables (IV). By linear, it means that the change in DV by 1 unit change in IV is constant. By additive, it refers to the effect of X on Y is independent of other variables.

2. Residuals or errors are normally distributed with N(0, 𝜎2). Homoscedasticity

3. Errors and explanatory variables are not correlated, i.e., Cor(𝑋,𝜖) = 0

4. The error terms must be uncorrelated. Presence of correlation in error terms is known as Autocorrelation. It drastically affects the regression coefficients and standard error values since they are based on the assumption of uncorrelated error terms.

5. There is no perfect linear relationship among the explanatory variables (not a problem in bivariate regression) --- also known as problem of multicollinearity

Methodology------------------------------------------

The process of data analysis can start with exploratory data analysis, which will give us the outline of the data set. It includes setting up the working directory, fetching information on number of rows and column, listing features in the data set and its corresponding datatypes. It also includes checking for the data set and fetching data summary. Then we will perform the descriptive analysis. Descriptive analysis will help us to understand, what these sample data say. Using descriptive analysis we can get overview or summary statistic of the data, which includes- measurement of center tendency, averages, mean, standard deviation, histogram, boxplots etc.

In the next step we will perform the data transformations or variable transformation on the dataset. Variable Transformation and Feature Creation are part of Feature Engineering, and can be performed once we are done with the earlier steps of Variable Identification, descriptive statistics, Missing Value imputation and Outlier detection. These two techniques are vital in data exploration and have a remarkable impact on the power of prediction. In data modelling, transformation refers to the replacement of a variable by a function. In other words, transformation is a process that changes the distribution or relationship of a variable with others.

Regression is a parametric technique used to predict continuous (dependent) variable given a set of independent variables. it makes certain assumptions ,If the data set follows those assumptions, regression gives incredible results. Otherwise, it struggles to provide convincing accuracy.

Mathematically, regression uses a linear function to approximate (predict) the dependent variable given as: Y = βo + β1X + ∈

Where, Y - Dependent variable, this is the variable we predict

X - Independent variable, this is the variable we use to make a prediction

βo – Intercept, this is the intercept term. It is the prediction value when X = 0

β1 – Slope, It explains the change in Y when X changes by 1 unit.

∈ - Error, this represents the residual value, i.e. the difference between actual and predicted values.

βo and β1 are known as coefficients. This is the equation of simple linear regression. It's called 'linear' because there is just one independent variable (X) involved. Error is an inevitable part of the prediction-making process. No matter how powerful the algorithm we choose, there will always remain an (∈) irreducible error which reminds us that the "future is uncertain".

A regression with two or more explanatory variables is called a multiple regression. Rather than modeling the mean response as a straight line, as in simple regression, it is now modeled as a function of several explanatory variables. The function “lm” can be used to perform multiple linear regression in R and much of the syntax is the same as that used for fitting simple linear regression models.

Steps to Conduct Regression Analysis

1. Regression equation formulation
2. Running regression analysis & interpretation of result
3. Testing assumptions and result explanation
4. Data transformation and sensitivity analysis (if needed)
5. Robust regression analysis
6. Parsimony (selecting best fit model)
7. Best regression model and result interpretation
8. Prediction

Data Analysis and Reporting---------------------------------

The entire process of data analysis can be divided into following steps. We can follow step by step approach to arrive at the conclusion.

1) Exploratory Data Analysis

2) Descriptive Statistics

3) Data Transformations

4) Regression Analysis

5) Conclusion & Summary

Exploratory Data Analysis

Dataset is a data frame with 50 observations of 5 variables, in other words - the dataset Household Data consist of 50 (fifty) rows and 5 (5) columns. The dataset contains 5 features as follows in the exact order.

|  |  |  |
| --- | --- | --- |
| Feature Code | Type | Continuous/ Discrete |
| Annual Income ($1000s) | numeric | Continuous |
| Monthly Income ($1000s) | numeric | Continuous |
| Household Size | numeric | Discrete |
| Amount Charged ($) | numeric | Continuous |
| Monthly EMI ($1000s) | numeric | Continuous |

Using R code we can check for missing values in the data set - output indicating that there is no missing values or place holders in the data. Here all features are numeric data types. We can observe that all features except “Household Size” are continues in nature but “Household Size” is a discrete variable. Also we can notice that every column name is not in an efficient and user-friendly, to be used in further statistical analysis process. So we need to change the column name in to more easy & useable format. The variable Amount charged is in 1000 multiple but other variables are in fraction of 1000, so it is better to go for data transformation.

|  |  |
| --- | --- |
| Old variable name | New variable name |
| Annual Income ($1000s) | annual\_income |
| Monthly Income ($1000s) | monthly\_income |
| Household Size | member\_no |
| Amount Charged ($) | expenditure |
| Monthly EMI ($1000s) | EMI |

Descriptive Statistics---------------------------

Descriptive Statistics provides simple summaries about the sample and the measures. Together with simple graphics analysis. Using descriptive analysis we can analyze the measures of Central Tendency and measure of dispersion of continues variables.

Measure of Central Tendency

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | annual\_income | monthly\_income | member\_no | expenditure | EMI |
| Mean | 43.48 | 3.623 | 3.42 | 3964 | 0.862 |
| Median | 42.00 | 3.500 | 3.00 | 4090 | 0.800 |
| Minimum | 21.00 | 1.750 | 1.00 | 1864 | 0.200 |
| Maximum | 67.00 | 5.583 | 7.00 | 5678 | 2.100 |

Measure of Dispersion

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | annual\_income | monthly\_income | member\_no | expenditure | EMI |
| Range | 46.00 | 3.833 | 6.00 | 3814 | 1.900 |
| 1st Quartile | 30.25 | 2.521 | 2.00 | 3130 | 0.600 |
| 3rd Quartile | 54.75 | 4.562 | 4.75 | 4733 | 1.200 |
| IQR | 24.50 | 2.041 | 2.75 | 1603 | 0.6 |
| Variance | 211.7241 | 1.470306 | 3.024082 | 871411.2 | 0.2309755 |
| Standard Deviation | 14.55074 | 1.212562 | 1.738989 | 933.4941 | 0.4805991 |

Data Visualization

The dataset is composed of 50 observations of 5 features. The minimum value of variable annual\_income is 21 and maximum value is 67.00. Mean of the variable is 43.48 and the median is observed as 42. The variable monthly\_income having a variance of 1.47 and standard deviation 1.212 with mean and median as 3.62 and 3.5 respectively. The minimum value of monthly\_income 1.75 and maximum value is 5.583

The maximum members in a household is 7 and minimum members is 1. Except from other variables “expenditure is given as 1000 multiple while others as fraction of thousand dollars. Its minimum value is 1864 and the maximum expenditure in the data set is 5678 dollars, mean of the expenditure is 3969 and the median is 4090. Range of the annual\_income is 46 while it is 3.833 for monthly\_income. Standard deviation of the variable member\_no is 1.739 and corresponding variance is 3.024. Inter quartile range (IQR) monthly\_income is 2.041, whereas expenditure having an IQR 1603.

Most households contain member numbers in the slab value of 1 to 2, which is far greater than the frequencies of other member\_no values. From the analysis of boxplots of each individual variables we can observe that all variables except “EMI” is free of outliers. From histogram and density plot of the variable expenditure we can conclude that variable is following a normal distribution. In scatter plot of variables, it is clear that the variable annual income and monthly income are strongly associated with each other, the relationship between the variables is linear in nature and possessing a strong correlation. It can be an indication of multicollinearity.

Analysis on the scatter plot reveals that the distribution between expenditure and monthly\_income is very similar to the distribution between expenditure and annual\_income. So we can make an assumption, it is a case of multicollinearity between the variables annual\_income and monthly\_income. The plot against member\_no against variable EMI is also showing a very strong positive correlation. Plot of each variable against the dependent variable expenditure is almost linear in nature and having a positive correlation. Exploring the other plots between independent variables reveals that there is no particular correlation between member\_no & monthly\_income, member\_no & annual\_income, EMI & annual\_income and EMI & monthly\_income.

Data Transformations--------------------------

In data modelling, transformation refers to the replacement of a variable by a function. For instance, replacing a variable x by the square / cube root or logarithm x is a transformation. In other words, transformation is a process that changes the distribution or relationship of a variable with others.

In the dataset, variables annual\_income, monthly\_income and EMI are stored in fractions of 1000 dollars scale. But feature expenditure is the full figure format. So to adjust the difference the scale of the variable it is better to perform a variable transformation. Here we want to change the scale of a variable or standardize the values of a variable for better understanding. While this transformation is a must if you have data in different scales, this transformation does not change the shape of the variable distribution.



To adjust the scale of the variable expenditure we divided each observations in the variable using 1000 and stored it into new variable exp1000. Also we remove the previous expenditure variable from the data frame. So in context we created new feature along with the transformation and remove the initial variable used for the transformation.

We are also performing logarithmic transformation on each variable and storing it in a newly created variables in the same data frame mydata. We can use this variables in future if we needed. Log-log method is using to transform complex non-linear relationships into linear relationships.



Regression Analysis-----------------------------------------