Introduction

Indian Premier League (IPL) is a professional 20-20 cricket tournament in India under BCCI. IPL is consist of almost all popular players all around the world. Best player become a part of a team based on the official auction. Team franchises can acquire the players during the auction by paying the auction money. The base price of the players are fixing based on their ranking and other ratings. Obviously, franchises pay higher salaries for quality players based on their past performance, current match form and match fitness. So analyzing the players based on past performances and ranking them will help the franchisees to choose the best players at best price.

Here the performance measures and other sports parameters of the players during the session 2012 is given in the datasets. Record of these parameter are available in two different data set. Past record of batsmen are stored and available in the dataset “batting\_bowling\_ipl\_bat.csv” and the record of bowing related parameters of bowlers are given in the dataset “batting\_bowling\_ipl\_bowl.csv”. Here we need to conduct Principal component analysis on these dataset separately and we need to rank the plays based on the principal component analysis scores.

Project Objective.

The objective is to use principal component analysis (PCA) to rank the batmen and bowlers based on the IPL-2012 session’s performance data. The recorded values of performance parameters of batsmen and bowlers are available in the datasets “batting\_bowling\_ipl\_bat.csv” and “batting\_bowling\_ipl\_bowl.csv” respectively. Here we need to perform the principal component analysis on these two data set independently and rank those players based on the PCA score. Here as a part of principal component analysis (PCA) we need to perform the following steps:

* Interpret loadings
* Interpret Communality
* Number of components to be retained
* Total variance extracted
* Check whether rotation is necessary
* Label the components
* Use the PC Scores to rank the players

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 performance of IPL players in the season 2012. We also assume that the dataset will be reflective of the reality and the features in the dataset are perfect reflection of 2012 season performances of players. Performance records of batmen is stored in the dataset “batting\_bowling\_ipl\_bat.csv” and the bowler data is recorded in the dataset “batting\_bowling\_ipl\_bowl.csv”. Here we assuming that dataset are independent of each other and contain most precise values for each players.

The following is considered for 7 features in the dataset: batting\_bowling\_ipl\_bowl.csv

|  |  |  |
| --- | --- | --- |
| Sr. no | Column Name | Description |
| 1 | Name | Name of batsman |
| 2 | Runs | Number of Runs scored in the tournament |
| 3 | Ave | Tournament batting average |
| 4 | SR | Tournament strike rate |
| 5 | Fours | Number of “fours” earned during the tournament |
| 6 | Sixes | Number of “sixes” earned during the tournament |
| 7 | HF | Number of “half-centuries earned in the tournament |

The following is considered for 5 features in the dataset: batting\_bowling\_ipl\_bat.csv

|  |  |  |
| --- | --- | --- |
| Sr. no | Column Name | Description |
| 1 | Name | Name of batsman |
| 2 | Wkts | Number of wickets taken during the tournament |
| 3 | Ave | The average number of runs conceded per wicket. (Ave = Runs/W) |
| 4 | Econ | The average number of runs conceded per over. (Econ = Runs/Overs bowled) |
| 5 | SR | The average number of balls bowled per wicket taken. (SR = Balls/W) |

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.

Principal component analysis (PCA) allows us to summarize and to visualize the information in a data set containing individuals/observations described by multiple inter-correlated quantitative variables. Principal component analysis is used to extract the important information from a multivariate data table and to express this information as a set of few new variables called principal components. These new variables correspond to a linear combination of the originals. The number of principal components is less than or equal to the number of original variables. In other words, PCA reduces the dimensionality of a multivariate data to two or three principal components that can be visualized graphically, with minimal loss of information.

The Bartlett Sphericity Test will be conducted to check whether the data dimensionality reduction is possible or not in the dataset. If P-value of the test is < 0.05 we reject the null hypothesis and we can perform the data dimensionality reduction using PCA. The reduced data (principal components) can be used in regression or data mining techniques.

A principal component can be mathematically written as:

𝑃𝐶1 = 𝛼1𝑉1 + 𝛼2𝑉2 + 𝛼3𝑉3 + 𝛼4𝑉4 + 𝛼5𝑉5 + 𝛼6𝑉6

Here variables, {V1, V2, V3, V4, V5, V6} are linearly combined using optimal weights in each of the principal components. The weights {𝛼, 𝛾, 𝛿, 𝜃, 𝜋, 𝜗} with subscripts on each variable in each equation are called loadings. If there is N variables in the data then there will be N principal components. The degree of variable grouping (multicollinearity) will guide us in determining how many of the N principal components to be retained. Higher the values of the weights (or loadings) higher is the importance of that variable in that component. The loadings are correlations between the principal components and the individual variables.

Coefficient of determination is square of correlation, it is the measure of common variance. So the total variance extracted by PC1: 𝛼12 + 𝛼22 + 𝛼32+ 𝛼42 + 𝛼52 + 𝛼62. If the variance extracted by each of the components is greater than 1. You may decide to retain all those components which extract at least 75% of the total variance in the model. We perform the Principal Component Analysis in R-studio using the R-programming interface.

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

4) Principal Component Analysis

5) Ranking & Summary

1. Exploratory Data Analysis: Batting Data

Preliminary step of exploratory analysis indicating a large amount of missing values in the dataset.



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | Name | Runs | Ave | SR | Fours | Sixes | HF |
| Missing value count | 0 | 90 | 90 | 90 | 90 | 90 | 90 |

We can remove the missing values or each ninety empty rows using the code:



The transformed data frame “bat” is consist of 90 observations of 7 variables related to the batting performance of each players. The dataset contains 7 features as follows in the exact order.

|  |  |  |
| --- | --- | --- |
| Feature Code | Type | Continuous/ Discrete |
| Name | Factor | Categorical |
| Runs | integer | Continuous |
| Ave | numeric | Continuous |
| SR | numeric | Continuous |
| Fours | integer | Continuous |
| Sixes | integer | Continuous |
| HF | integer | Continuous |

All variable except “Name” is continues variables. The variable “Name” is not required in the principal component analysis, because the variable is insignificant. So we will remove the feature from the data set using the subletting method.

2. Exploratory Data Analysis: Bowling Data

Features related to the performances of each bowlers is stored in a dataset “batting\_bowling\_ipl\_bat” in CSV format. As part of analysis data is imported in to the data frame bowls. Analysis indicating that, large numbers of missing values is present in the data set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables | Name | Wkts | Ave | Econ | SR |
| Missing value count | 0 | 90 | 90 | 90 | 90 |

After removing all missing values from the data: we can observe that data is solid and very apt for the PCA. The modified data frame “bowl” is consist of 83 observations of 5 variables related to the bowling performance of each players. The dataset contains 5 features as follows in the exact order.

|  |  |  |
| --- | --- | --- |
| Feature Code | Type | Continuous/ Discrete |
| Name | Factor | Categorical |
| Wkts | integer | Continuous |
| Ave | numeric | Continuous |
| Econ | numeric | Continuous |
| SR | numeric | Continuous |

Same as above, all variable except “Name” is continues variables. The variable “Name” is not required in the principal component analysis, because the variable is insignificant. So we will remove the feature from the data set using the subletting method.

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. In this case we can perform the descriptive statistics on each dataset separately.

1. Descriptive analysis: Batting Data

The batting dataset is consist of the six “batting variables”: Runs, Ave, SR, Fours, Sixes and HF. In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical matrices and visualization methods as shown below:

Measure of Central Tendency

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Runs | Ave | SR | Fours | Sixes | HF |
| Mean | 219.9 | 24.73 | 119.16 | 19.79 | 7.578 | 1.189 |
| Median | 196.5 | 24.44 | 120.14 | 16.00 | 6.000 | 0.500 |
| Minimum | 2.0 | 0.50 | 18.18 | 0.00 | 0.000 | 0.000 |
| Maximum | 733 | 81.33 | 164.10 | 73.00 | 59.000 | 9.000 |

Measure of Dispersion

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Runs | Ave | SR | Fours | Sixes | HF |
| Range | 731 | 80.83 | 145.92 | 73.00 | 59.000 | 9.000 |
| 1st Quartile | 98.0 | 14.66 | 108.75 | 6.25 | 3.000 | 0.000 |
| 3rd Quartile | 330.8 | 32.20 | 132.00 | 28.00 | 10.000 | 2.000 |
| IQR | 232.8 | 17.54 | 23.25 | 21.75 | 7.000 | 2.000 |
| Variance | 24415.21 | 185.483 | 559.6322 | 268.9549 | 64.02197 | 2.851561 |
| SD | 156.2537 | 13.6192 | 23.6565 | 16.3998 | 8.0014 | 1.6886 |

1. Descriptive analysis: Bowling Data

The bowling dataset is consist of the six “bowling variables”: Name, Wkts, Ave, Econ and SR. In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical matrices and visualization methods as shown below:

Measure of Central Tendency

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Wkts | Ave | Econ | SR |
| Mean | 8.88 | 34.51 | 7.66 | 26.33 |
| Median | 8.00 | 29.00 | 7.53 | 21.60 |
| Minimum | 1.00 | 12.20 | 5.40 | 12.00 |
| Maximum | 25.00 | 161.00 | 11.65 | 96.00 |

Measure of Dispersion

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Runs | Ave | SR | Fours | Sixes | HF |
| Range | 731 | 80.83 | 145.92 | 73.00 | 59.000 | 9.000 |
| 1st Quartile | 98.0 | 14.66 | 108.75 | 6.25 | 3.000 | 0.000 |
| 3rd Quartile | 330.8 | 32.20 | 132.00 | 28.00 | 10.000 | 2.000 |
| IQR | 232.8 | 17.54 | 23.25 | 21.75 | 7.000 | 2.000 |
| Variance | 24415.21 | 185.483 | 559.6322 | 268.9549 | 64.02197 | 2.851561 |
| SD | 156.2537 | 13.6192 | 23.6565 | 16.3998 | 8.0014 | 1.6886 |

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-----------------------------------------

In regression first we create linear composite that will express the relationship between set of predictors (independent variable, explanatory variable, or regressor) and a criterion variable (dependent variable, explained variable or regressand). A regression equation is linear in parameters.

Our data frame is consist of 5 variables- annual\_income, monthly\_income, member\_no, expenditure and EMI. Based on the project objective we need to create a linear multivariate regression model for predicting the monthly “expenditure” of the house hold based of effect of other variables.

So our basic linear regression equation is: Y = β0 + β1 X1 + β2 X2 + β3 X3 + β4 X4 + Є

Based on the above formula we can formulate our first regression equation as: expenditure = β0 + β1 \* annual\_income+ β2 \* monthly\_income + β3 \* member\_no+ β4 \* EMI+ Є

Linear composite is also called explained part and Є is the residual error (unexplained part). An error term is a variable in regression equation which is created when the model doesn’t fully represent the actual relationship between the independent variable and the dependent variable.

The values of β is not available directly, so we need to find and choose the best possible value for them.

β > 0: then +ve association, independent variable positively influence the dependent variable.

* + β < 0: then -ve association, independent variable negatively influence the dependent variable.
  + β = 0: then no association, independent variable have no influence over dependent variable.

We can setup, execute and interpret a regression model very easily using R programming language. We will perform regression analysis using our first basic but full variable regression equation formulated above. At last we will choose a best regression model from all the formulated equations with maximum efficiency and effectiveness.

The command to run a regression model in R is as follows:



The output consist of 3 different section

1. The result variable will give us the coefficient values of each independent variables including the intercept.
2. ANOVA table tell us how the variation or variance is distributed between explained part and unexplained part. In the result Sum of Squares is the explained variance in the dependent variable corresponding to the independent variable, it is called as explained sum of squares. Sum of Squares of Residuals gives the amount of variance in dependent variable based on residuals, it is called as residual sum of squares - (TSS= ESS + RSS).
3. Summary of regression result gives us the F-statistic, P-value, residual, standard errors, coefficient estimates, degree of freedom, significance codes and R2 value.

R2 is the measure of how much of variance in the dependent variable has been explained by the independent variable (R2 =ESS/TSS). Higher R2 value indicating that relationship is solid and regression model is a good one. If P-value <0.05, then the regression equation is significant to explain the dependent variable. If P-value of the independent variable is greater than 0.05 then we will exclude the variable from the equation (Ho: independent variable in the model doesn’t explain the variance in the dependent variable). The value of R2 increases with number of explanatory variable, then we will consider the adjusted R2 value which is penalizes for adding variables that don’t really explain the dependent variable. Standard error of the model is the measure of how wrong the regression model on average using the unit of the response variable. Smaller the value of standard error better the regression model because it indicating that the observations are closer to the fitted line.

Model-1: expenditure = β0 + β1 \* annual\_income+ β2 \* monthly\_income + β3 \* member\_no+ β4 \* EMI+ Є



Output Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| variable name | Coefficient value | P-value | Standard-error | Sum of Squares |
| annual\_income | 0.032208 | 1.13e-09 | 0.004234 | 16.999745 |
| monthly\_income | NA | NA | NA | NA |
| member\_no | 0.409730 | 2.82e-05 | 0.088114 | 18.251011 |
| EMI | -0.205876 | 0.516 | 0.314153 | 0.068897 |
| Intercept | 1.339843 | 5.00e-08 | 0.205888 | ------------------- |



Here P-value of the model is < 0.05 so we reject the null hypothesis, which indicating that current model is significant. From the summary table indicating that the two variables are replication of one another and here exists a strong multicollinearity or singularity between the variable so we can reject one of the variable from the model (there is no difference between the variables annual\_income and monthly\_income, so we can exclude anyone of this variable from the model- it won’t matter which one we excluded - [result1a<-lm(exp1000~monthly\_income+annual\_income+member\_no+EMI, data=mydata) ]). P-value of the variables monthly\_income or annual\_income and member\_no are less than 0.05 so these are significant in the model. The coefficient value of EMI is -0.205876 and P-value is >0.05 so we accept the null hypothesis which means that variable EMI is nonsignificant in the model and can be excluded after test of assumptions. The multiple R-squared value is 0.8272 and adjusted R-squared value is 0.8159 – which indicating that model is a good one. Based on the output we can conclude that the variables monthly\_income (or annual\_income), member\_no and intercept are significant in the model.

Test of Assumptions

Presence of the assumptions make regression quite restrictive. The performance of a regression model is conditioned on fulfillment of these assumptions. Once these assumptions get violated, regression makes biased, erratic predictions.

Test 1: Mean of the residuals is zero

In a solid regression model sum of residual terms is always zero. If we take the absolute values of the distances from each and every data points to the fitted line and added those errors, the sum total of the errors is zero.

> mean(result1$residuals)

# -4.929867e-18

The mean value is a very small number almost equal to zero, so the first assumption is holding.

Test 2: Homoscedasticity of residuals

Plots will help to check the assumptions of normality and homoscedasticity.

> par(mfrow=c(2,2)) # set 2 rows and 2 column

> plot layout plot(result1)

> dev.off()

Normal Q-Q plot (quantile-quantile plot): As the name suggests, this plot is used to determine the normal distribution of errors. It uses standardized values of residuals. Ideally, this plot should show a straight line. If you find a curved, distorted line, then your residuals have a non-normal distribution (problematic situation). Here plot is a straight line which means whole thing is normally distributed. There are some outliers at the start and end points of the plot, but it is not a major violation. So residuals are normally distributed and assumption is holding.

Residual vs. Fitted Values Plot: Ideally, this plot shouldn't show any pattern. But if you see any shape (curve, U shape), it suggests non-linearity in the data set. In addition, if you see a funnel shape pattern, it suggests your data is suffering from heteroscedasticity, i.e. the error terms have non-constant variance. If the graph gives a straight line which means residuals follows homoscedastic distribution. The current graph contain a straight line which means residuals follows a linear distribution and homoscedastic in nature.

Scale Location Plot: This plot is also useful to determine heteroscedasticity. Ideally, this plot shouldn't show any pattern. Presence of a pattern determine heteroscedasticity. Here scale-location plot contain a curved line which and there is no particular pattern present in the distribution, so there is no heteroscedasticity among the residuals.

Above two test of assumption indicating that residuals having a mean value of zero and follows homoscedasticity.

Test 3: Errors and explanatory Variables are uncorrelated

We can test this assumption using the R function: cor.test (IV, residuals). Here we have four independent variable so we need to run the test against each explanatory variables separately.

Ho: Errors and Explanatory variables are uncorrelated

Ha: Errors and Explanatory variables are correlated

cor.test(annual\_income,result1$residuals) #### correlation value: -4.048694e-17

cor.test(monthly\_income,result1$residuals) #### correlation value: -6.178369e-17

cor.test(member\_no,result1$residuals) #### correlation value: -2.298867e-16

cor.test(EMI,result1$residuals) #### correlation value: 5.496769e-17

Here P-value of each test is =1 which is >0.05 so we reject the alternative hypothesis and accept the null hypothesis. Also the correlation values of each test is very small and nearly equal to zero, so the assumption holds, errors and explanatory variables are uncorrelated.

Test 4: There is no perfect linear relationship among the explanatory variables

Presence of correlation in independent variables lead to Multicollinearity. If variables are correlated, it becomes extremely difficult for the model to determine the true effect of IVs on DV.

Correlation table: To investigate possible multicollinearity, first look at the correlation coefficients for each pair of continuous (scale) variables. Correlations of 0.8 or above suggest a strong relationship and only one of the two variables is needed in the regression analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | annual\_income | monthly\_income | member\_no | EMI | exp1000 |
| annual\_income | 1.00 | 1.00 | 0.17 | 0.03 | 0.63 |
| monthly\_income | 1.00 | 1.00 | 0.17 | 0.03 | 0.63 |
| member\_no | 0.17 | 0.17 | 1.00 | 0.92 | 0.75 |
| EMI | 0.03 | 0.03 | 0.92 | 1.00 | 0.61 |
| exp1000 | 0.63 | 0.63 | 0.75 | 0.61 | 1.00 |

The correlation values of the variable annual\_income & monthly\_income is = 1, which means there exists a perfect singularity between the variable or there is high degree of correlation or multicollinearity among the variable. Same way the correlation between the variables EMI and member\_no is =0.92, which implies that there is a strong correlation between the variables and the multicollinearity exists.

In the case of first variable we can reject any variable which we prefer, because both have same correlation between the dependent variable. But in the case of member\_no and EMI correlation between the member\_no and exp1000 is = 0.75 and correlation between is EMI and exp1000 is =0.61, which is less than the member\_no, so preferably it is more apt to exclude the EMI from the model.

Note: we can also check the correlation using the function:

vif(lm(exp1000~monthly\_income+member\_no+EMI, data=mydata))

vif(lm(exp1000~annual\_income+member\_no, data=mydata))

Here variation inflation parameters of the variables - member\_no and EMI is larger in first model. But when we remove the EMI variable and run the second variable the VIF of the parameters is low and normal. If vif measure is high between 2 variables while comparing to other variables, it indicating the case of existence of multicollinearity.

======================================

Model-2: expenditure = β0 + β1 \* monthly\_income + β2 \* member\_no + β3 \* EMI+ Є



Result of the above regression model is as follows:

P-value: < 2.2e-16 so the model is significant, but P-value of the variable EMI is 0.516 and estimate is –ve value, same as in previous model it strongly suggesting the removal of the variable from the model.

Residual standard error: 0.4005 on 46 degrees of freedom.

Multiple R-squared: 0.8272.

Adjusted R-squared: 0.8159

Model-3: expenditure = β0 + β1 \* monthly\_income + β2 \* member\_no + Є



Based on the suggestions from previous models and conclusions from the test of assumptions, we are running a most suitable model after excluding the variables “annual\_income” and “EMI”

Output Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| variable name | Coefficient value | P-value | Standard-error | Sum of Squares |
| monthly\_income | 0.3976 | 7.68e-11 | 0.04761 | 16.999745 |
| member\_no | 0.3563 | 3.12e-14 | 0.03320 | 18.251011 |
| Intercept | 1.3049 | 3.29e-08 | 0.19765 | ------------------- |



Analysis on the output indicating that it is the most efficient, effective and best fit model that can be created. In this model we removed all the unnecessary variables and created the best regression equation and corresponding model for the most relevant prediction. In this model the coefficient values of each model is increased and the standard error of individual variable and the residual standard error of the model itself decreased. The multiple R-squared value (0.8256) and adjusted R-squared value (0.8181) are very high and almost equal to the Model-1 values. The current model achieve such a good R-squared value even after decrease in the number of independent variable.

*Note: Refer appendix section for other model created and tested.*

Robust Linear Regression

It is using to address the issues like outlier, leverage variable etc. The variable EMI in the model-2 contains very few outliers and leverages, but the variables in the model-3 is free of leverages and outliers. Even though we can running the robust linear regression on model-2 for analysis and better understandability (we can’t run the robust regression on the model-1 because of the perfect singularity between the two variable)



The difference in the current model from the previous model-2 is mainly in the residual standard error, here the residual standard error is decreased to 0.3026 on 46 degree of freedom. Here we can see changes in the coefficient values also. The sign of the variable EMI is still –ve and the P-value is >0.05 and have no significance in the model.

There is no need of robust regression analysis on the model-3, because the variables in the model is free of leverages and outliers. The model-3 has no need of any data transformation and we don’t have to construct a log –log model for model-3, because the model-3 and its variables completely holding every assumptions of a perfect linear regression model. So far it is the best and most efficient regression model available.

Parsimony

A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible. If we have large set of predictors (or regressors) then the number of regressors to be included in the model need to be chosen. Parsimonious means the simplest model/theory with the least assumptions and variables but with greatest explanatory power

We can use different types frame work for the purpose of choosing the least predictor variables with best correlation.

1. Forward selection
2. Backward selection
3. Stepwise regresson

Stepwise Regression



Output

* Output of the parsimony model is also suggesting model-3 as the best fit model. Which is: result3<-lm(exp1000~monthly\_income+member\_no, data=mydata) and the corresponding linear regression equation is : expenditure = β0 + β1 \* monthly\_income + β2 \* member\_no + Є, where β0 = 1.3049, β1 = 0.3976 and β2 = 0.3563.

Prediction

Main focus of building a regression model and regression analysis is always creating a best model capable of delivering most accurate prediction according to most relevant independent variable data. Here from the process of regression analysis we can selects the model-2: expenditure = β0 + β1 \* monthly\_income + β2 \* member\_no + Є as the best and true regression model and it can be used in the prediction.

We often use our regression models to estimate the mean response or predict future values of the response variable for certain values of the response variables. The function predict() can be used to make both confidence intervals for the mean response and prediction intervals. To make confidence intervals for the mean response use the option interval=”confidence”. To make a prediction interval use the option interval=”prediction”. By default this makes 95% confidence and prediction intervals. If you instead want to make a 99% confidence or prediction interval use the option level=0.99.



Conclusion

Multilinear regression was carried out to investigate the relationship between the dependent variable “expenditure” and independent variables “annual\_income”, “monthly\_income”, “household members” and “EMI”. The regression analysis on complete set of variables revels that the variables annual\_income (or monthly\_income) and EMI are not significant. So these variables are not required in an improved model. Remaining variable “monthly\_income” and “member\_no” are in strong correlation with dependent variable “expenditure” and these variables together forming the most effective and best fit regression model which is named as model-3. The regression model-3 s not violating any of the regression assumptions and the model is solid and sound in terms of effectiveness and predictive accuracy.

The variables “annual\_income and “monthly\_income are perfectly singular to each other and created the multicollinearity. The variables “member\_no” and “EMI” also violating the assumption of multicollinearity because of the correlational value = 0.92.

The model-3: expenditure = β0 + β1 \* monthly\_income + β2 \* member\_no + Є, have a P-value < 2.2e-16. The R-squared value of the model is =0.856 and adjusted R-squared value is = 0.8181, so these figures strongly supporting the fact that the model-3 is the best fit regression model that can deliver maximum correlation between dependent and independent variable with minimum standard error. The incept (β0) value of model-3 is 1.3049 and the coefficient values of variables monthly\_income (β1) and member\_no (β2) are0.3976 and 0.3563 respectively. The positive coefficient values showing that – the independent variables are positively influencing the dependent variable “expenditure”. The parsimony model output also supporting the model-3 as a best model. The adjusted R-squared value of the model-3 is 0.8181, which means: 81% of the variation in the dependent variable “expenditure” can be explained using the variables monthly\_income and member\_no.

APPENDIX

setwd("C:/Users/SHYAM KRISHNAN K/Desktop/PJ-3/regression")

getwd()

library("readxl")

library(psych)

library(car)

mydata <- read\_excel("Household Data.xlsx")

################################ Exploratory data analysis

dim(mydata)

names(mydata)

str(mydata)

head(mydata)

tail(mydata)

colSums(is.na(mydata))

colnames(mydata)<-c("annual\_income","monthly\_income","member\_no","expenditure","EMI")

############################## Descriptive data analysis.

attach(mydata)

str(mydata)

summary(mydata)

range(annual\_income)

range(monthly\_income)

range(member\_no)

range(EMI)

range(expenditure)

sd(annual\_income)

sd(monthly\_income)

sd(member\_no)

sd(EMI)

sd(expenditure)

var(annual\_income)

var(monthly\_income)

var(member\_no)

var(EMI)

var(expenditure)

########## data transformation

mydata<- transform(mydata, exp1000 = expenditure / 1000)

mydata$expenditure<- NULL

attach(mydata)

################### Data visualization

plot(mydata, pch=1, col="brown", main="Scatterplot of complete variables")

library(lattice)

plot(density(expenditure),col='green', main="density plot - expenditure")

hist(expenditure,main = 'Monthly Expenditure',xlab = 'expenditure',ylab = 'Frequency',col = 'orange')

par(mfrow=c(2,2))

hist(annual\_income,main = 'Annual Income',xlab = 'annual\_income',ylab = 'Frequency',col = 'green')

hist(member\_no,main = 'Member Number',xlab = 'member\_no',ylab = 'Frequency',col = 'pink')

hist(monthly\_income,main = 'Monthly Income',xlab = 'monthly\_income',ylab = 'Frequency',col = 'brown')

hist(EMI,main = 'EMI',xlab = 'EMI',ylab = 'Frequency',col = 'blue')

dev.off()

boxplot(EMI, main="EMI", col = 'green',xlab = 'emi')

############# Multivariate regression ###########

result1<-lm(exp1000~annual\_income+monthly\_income+member\_no+EMI, data=mydata)

result1

aov(result1)

summary(result1)

result2<-lm(exp1000~monthly\_income+member\_no+EMI, data=mydata)

result2

aov(result2)

summary(result2)

result5<-lm(exp1000~monthly\_income+EMI, data=mydata)

result5

aov(result5)

summary(result5)

result3<-lm(exp1000~monthly\_income+member\_no, data=mydata)

result3

aov(result3)

summary(result3)

#################### Test of assumptions #################

####Mean of the residuals is zero

mean(result1$residuals)

#######Homoscedasticity of residuals

par(mfrow=c(2,2))

plot(result1,col="blue")

dev.off()

############ Errors and Explanatory Variables are uncorrelated

cor.test(annual\_income,result1$residuals)

cor.test(monthly\_income,result1$residuals)

cor.test(member\_no,result1$residuals)

cor.test(EMI,result1$residuals)

############multi-collinearity

library(car)

vif(result1)

alias( lm(exp1000~annual\_income+monthly\_income+member\_no+EMI, data=mydata))

vif(lm(exp1000~annual\_income+monthly\_income+member\_no+EMI, data=mydata))

vif(lm(exp1000~monthly\_income+member\_no+EMI, data=mydata))

vif(lm(exp1000~annual\_income+member\_no, data=mydata))

#################### testing variables correlation ########

round(cor(cbind(annual\_income,monthly\_income,member\_no,EMI,exp1000)),2)

############ Robust Regression #############

library(foreign)

library(MASS)

r\_result2<-rlm(exp1000~monthly\_income+member\_no+EMI, data=mydata)

r\_result2

aov(r\_result2)

summary(r\_result2)

########log-log model###########

mydata$lna\_income<-log(mydata$annual\_income)

mydata$lnm\_income<-log(mydata$monthly\_income)

mydata$lnmembe<-log(mydata$member\_no)

mydata$lnEMI<-log(mydata$EMI)

mydata$lnexp1000<-log(mydata$exp1000)

newlogdf <- c("lna\_income", "lnm\_income", "lnmembe","lnEMI","lnexp1000")

newdata <- mydata[newlogdf]

summary(newdata)

plot(newdata)

res\_log<-lm(mydata$lnexp1000~mydata$lna\_income+mydata$lnm\_income+mydata$lnmembe+mydata$lnEMI)

res\_log

aov(res\_log)

summary(res\_log)

res\_log1<-lm(mydata$lnexp1000~mydata$lna\_income+mydata$lnmembe)

res\_log1

aov(res\_log1)

summary(res\_log1)

res\_log2<-lm(mydata$lnexp1000~mydata$lna\_income+mydata$lnmembe+mydata$lnEMI)

res\_log2

aov(res\_log2)

summary(res\_log2)

############## parsimony

######### stepwise regression

library(leaps)

Null<-lm(exp1000~1, data=mydata)

Full<-lm(exp1000~annual\_income+monthly\_income+member\_no+EMI, data=mydata)

step(Null, scope = list(upper=Full), data=mydata, direction="both")

#########prediction

finalmodel<-lm(formula = exp1000 ~ member\_no + monthly\_income, data = mydata)

finalmodel

anova(finalmodel)

summary(finalmodel)

predict(finalmodel,data.frame(member\_no=3, monthly\_income=4.5),interval="confidence")

predict(finalmodel,data.frame(member\_no=3, monthly\_income=4.5),interval="prediction")