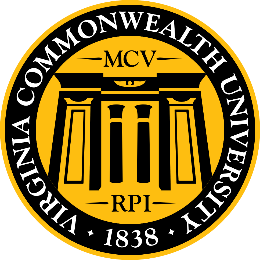
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A2: REGRESSION ANALYSIS**

**KATHRYN SHAJU**

**V01111327**

**Date of Submission: 23-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Business Significance | **1** |
| **3.** | Objectives | **2** |
| **3.** | Results & Interpretations | **2-4** |
| **4.** | Recommendations | **5** |

**INTRODUCTION**

This study is a two part study that focuses on using Regression analysis model to compute results for two datasets, they are, NSSO68 dataset and the IPL datasets that comprise of IPL ball by ball update2024, and IPL salaries 2024. The regression analysis is conducted using both software’s R and Python.

Using the NSSO68 dataset we have specifically looked at the state of Punjab wherein we have taken our dependent variable as foodtotal\_q which comprises of the total food quantity of the state, and we have selected 8 independent variable to see the cause and effect relation between the dependent and independent variables. Over we have used the method of multiple regression model due to the number of the independent variables.

Similarly, using the IPL datasets we run the OLS regression model to detect the relation between the player's performance and payment he receives. Over here we explore the relation between only one dependent and independent variable.

Regression analysis running both R and Python allow us to get better and accurate results, while simultaneously being able to compare the subtle differences occurring in the same. The major advantage with this is the forecasting ability that help in making plans for not only the business counterparts but also equally important in the decision making process for policy makers and economists.

**BUSINESS SIGNIFICANCE**

A strong statistical method for analyzing the relationship between one dependent variable and one or more independent variables is regression analysis. It has many uses in business, offering insightful information that can influence strategy and decision-making.

Regression analysis in particular can be the most helpful in forecasting for any future events that can be based on the relationship tested.

Forecasting helps not only business owners to gauge the market be it in terms of IPL owners or in terms of policymakers who can easily forecast any issues regarding food production in the state of Punjab when it can predict the variables that can have an effect on it and take necessary measures to prevent a catastrophe.

By comprehending the ways in which various elements like, consumer inclinations, and market circumstances impact sales, companies can enhance their pricing tactics to optimize earnings.

**OBJECTIVES**

a) To perform multiple regression analysis, to carry out the regression diagnostics, and explain them. To explain the significant differences found. [Data “NSSO68.csv”]

b) To use IPL data, to establish the relationship between the player's performance and payment he receives. And to Analyze the Relationship between Salary and Performance over the Last Three Years (Regression Analysis)

**RESULTS AND INTERPRETATIONS.**

1. To perform multiple regression analysis, to carry out the regression diagnostics, and explain them. To explain the significant differences found.

Using the NSSO68 dataset, we firstly set the directory and load the data set, following which we firstly filter the dataset according to the state, which is Punjab.

We then create the subset of the data where we include our dependent and independent variables, which are as follows ; foodtotal\_q, MPCE\_MRP, MPCE\_URP,Age,Meals\_At\_Home,Possess\_ration\_card,Education, No\_of\_Meals\_per\_day. Here foodtotal\_q is our dependent variable and the rest are chosen based on relatively theoretical relation.

**Regression Diagnostics:**

One of the assumptions of a regression analysis is that the data chosen should be homoscedastic in nature, which mean that there should not be any missing values nor should there be outliers, this could lead to a skewed distribution making the data inaccurate. Hence we check for any missing values in our data set and found 3 variables which have values missing, they are Education, No\_of\_Meals\_per\_day, Meals\_At\_Home.

Once we identify these variable we have imputed them with mean using function to impute with mean. This ensure that now the data is homoscedastic in nature.

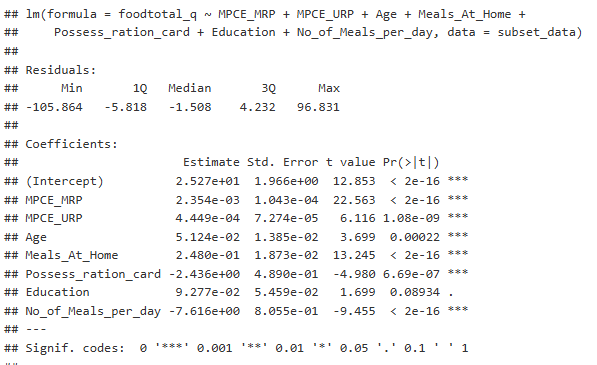
We also check for multicollinearity problem using the Variance Inflation Factor. If the VIF value is <8 we conclude that there is no problem of multicollinearity.

In our analysis we get the VIF value for all the variable less than 8, hence we can conclude that there is no problem of multicollinearity.

Now fitting the regression model we use

*model <- lm(foodtotal\_q~MPCE\_MRP+MPCE\_URP+Age+Meals\_At\_Home+Possess\_ration\_card+Education+No\_of\_Meals\_per\_day, data = subset\_data)*

This allows us to get the regression model. Which looks something like this:



Over here we observe that the all the variables mostly have less than 1% significance and thus we can conclude that the relation between the variables chosen may not be significant.

Multiple R-squared: 0.3082, Adjusted R-squared: 0.3066

We also note that the Multiple R squared is only 0.3, which means that the model is not a good fit to measure.

Our results for the data in the Python Program also results in the same interpretations.

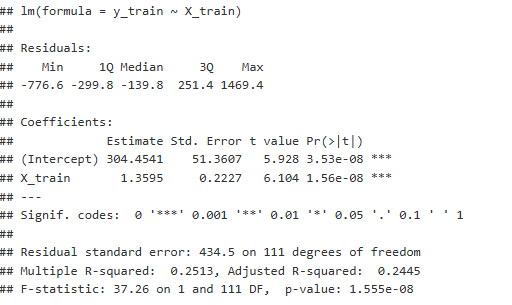
1. To use IPL data, to establish the relationship between the player's performance and payment he receives. And to Analyze the Relationship between Salary and Performance over the Last Three Years (Regression Analysis)

Here again after setting the directory, we load the 2 datasets, both IPL ball by ball update2024, and IPL salaries 2024 in the software’s. Once we load we group the data by Season, Striker and Bowler.

Over here when considering the performance of the players, based on the dataset we have two indicators for this, one is runs\_scored and wicket\_confirmation, hence we calculate total runs scored for each payer in each season likewise for bowler we calculate each wickets taken in each season.

Now to estimate the salary part of our problem we make separate data of salary for striker and now merge the data with the data of runs scored.

Now we can run regression;



Here our R squared value is 0.2, which mean that there is only a 20% significance of the player’s performance on his salary, which mean that the model is not a good fit of measure.

The reasons for the low R squared value could be due to the less amount of variable,

Which could indicate that the apart from the performance of the players which comprises of only runs scored and wickets taken, include more parameters that need to be tested. Taking only for the last three years could also mean that there is a need to include more number of seasons as well. This means that the data is insufficient for measuring a relationship between the performance and salary of the players.

We conclude the same results and interpretations from Python as well.

**RECOMMENDATIONS**

For our first objective, we might need to add more variables, and look also into whether there are any outliers or if there is a case of autocorrelation, which could lead to R square value being low.

Similarly from the results of the second objective, we should definitely include more variables as the data seems to be insufficient, due to there being less parameters that define the performance of the player. Also Salary can be a repetitive variable which could lead to the problem of not only multicollineraity but also autocorrelation which could lead to the skewed distribution of data causing R square to be low.