****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

**A2: Regression Analysis – P1**

**Manoranjan Mohankumar**

**V01107254**

**Date of Submission: 25-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2** |
| **3.** | Interpretations | **5** |
| **4.** | Recommendations | **7** |
| **5.** | Codes | **NA** |
| **6.** | References | **7** |

**Introduction**

Background

Multiple regression analysis is a powerful statistical technique that helps us understand the relationship between one dependent and several independent variables. This method is widely used in various fields, such as economics, social sciences, and natural sciences, to make predictions, infer causal relationships, and understand the dynamics within datasets. Researchers and analysts can make informed decisions by analyzing the relationships between variables, identifying trends, and developing strategies based on empirical evidence.

Objective

This report aims to perform a comprehensive multiple regression analysis using two different programming languages, R and Python, on a large dataset from the National Sample Survey Office (NSSO) 68th round. The aim is to:

* Determine the factors influencing food expenditure (foodtotal\_v) among households.
* Compare the results obtained from R and Python to ensure the robustness and reliability of the findings.
* Interpret the results and provide actionable insights based on the analysis.

Dataset Description

The dataset used for this analysis is the NSSO68 dataset, which is a part of the National Sample Survey conducted in India. The dataset includes a wide range of variables capturing various socio-economic aspects of households. For this analysis, the focus is on the following variables:

* foodtotal\_v: Total food expenditure.
* foodtotal\_q: Quantity of food consumed.
* MPCE\_URP: Monthly Per Capita Expenditure under Uniform Reference Period.
* MPCE\_MRP: Monthly Per Capita Expenditure under Mixed Reference Period.

These variables were selected to understand how household expenditure and consumption patterns influence total food expenditure.

Methodology

Both R and Python were employed to perform the regression analysis, leveraging their respective libraries and functions to ensure accurate and comprehensive analysis:

* R:
  + Data loading and preprocessing using readr and dplyr.
  + Regression analysis using the lm function.
  + Diagnostic checks using car package for VIF calculation.
  + Log transformation of the dependent variable to check for improved model fit.
* Python:
  + Data loading and preprocessing using pandas.
  + Regression analysis using statsmodels.
  + Diagnostic checks using variance\_inflation\_factor from statsmodels.
  + Log transformation of the dependent variable to evaluate model performance.

Using both languages allows for cross-validation of results, ensuring that the findings are not an artifact of a specific software environment but are inherent to the data and the relationships it encapsulates.

**Results**

R Code Execution

1. Data Loading and Preparation:
   1. The dataset was loaded using the readr package.
   2. Columns were selected for analysis: state\_1, foodtotal\_v, foodtotal\_q, MPCE\_URP, and MPCE\_MRP.
2. Regression Analysis:
   1. A multiple regression model was fitted using the lm function with foodtotal\_v as the dependent variable.
   2. The model summary indicated that all predictors were significant at the 0.001 level.
   3. Variance Inflation Factor (VIF) was calculated to check for multicollinearity, showing acceptable values (all below 2).
3. Log-Transformation:
   1. The dependent variable foodtotal\_v was log-transformed to handle skewness.
   2. A new regression model was fitted with the log-transformed dependent variable.
   3. The model summary indicated a lower R-squared value compared to the original model, suggesting a less explained variance.
4. Regression Equations:
   1. Original Model: foodtotal\_v = -93.7753 + 24.5783 \* foodtotal\_q - 0.0012 \* MPCE\_URP + 0.0617 \* MPCE\_MRP
   2. Log-Transformed Model: log\_foodtotal\_v = 5.1052 + 0.0463 \* foodtotal\_q - 0.0000 \* MPCE\_URP + 0.0000 \* MPCE\_MRP

Python Code Execution

1. Data Loading and Preparation:
   1. The dataset was loaded using pandas with selected columns: state\_1, foodtotal\_v, foodtotal\_q, MPCE\_URP, and MPCE\_MRP.
2. Regression Analysis:
   1. A multiple regression model was fitted using statsmodels with foodtotal\_v as the dependent variable.
   2. The model summary indicated that all predictors were significant at the 0.001 level.
   3. VIF was calculated using variance\_inflation\_factor, showing acceptable values (all below 2).
3. Log-Transformation:
   1. The dependent variable foodtotal\_v was log-transformed to handle skewness.
   2. A new regression model was fitted with the log-transformed dependent variable.
   3. The model summary indicated a lower R-squared value compared to the original model, suggesting a less explained variance.
4. Regression Equations:
   1. Original Model: foodtotal\_v = -93.7753 + 24.5783 \* foodtotal\_q - 0.0012 \* MPCE\_URP + 0.0617 \* MPCE\_MRP
   2. Log-Transformed Model: log\_foodtotal\_v = 5.1052 + 0.0463 \* foodtotal\_q - 0.0000 \* MPCE\_URP + 0.0000 \* MPCE\_MRP

**Interpretations**

Consistent Findings Across R and Python

Both the R and Python analyses yield consistent results, indicating that the selected independent variables (foodtotal\_q, MPCE\_URP, and MPCE\_MRP) significantly influence the dependent variable (foodtotal\_v). This consistency suggests robustness in the analysis and reliability in the conclusions drawn from the data.

Original Model Interpretation

Coefficients and Their Significance:

* Intercept: The negative intercept (-93.7753) implies that if all predictors (foodtotal\_q, MPCE\_URP, and MPCE\_MRP) were zero, the expected food expenditure would be negative, which is not realistic but indicates a baseline adjustment factor.
* foodtotal\_q (Quantity of Food): The positive coefficient (24.5783) signifies that for each additional unit increase in food quantity, the food expenditure increases by approximately 24.5783 units, holding other factors constant. This strong positive relationship indicates that food quantity is a major driver of food expenditure.
* MPCE\_URP (Monthly Per Capita Expenditure - Uniform Reference Period): The negative coefficient (-0.0012) suggests a slight decrease in food expenditure as MPCE\_URP increases, holding other factors constant. This might indicate that higher overall expenditure could lead to a relative decrease in the proportion spent on food.
* MPCE\_MRP (Monthly Per Capita Expenditure - Mixed Reference Period): The positive coefficient (0.0617) implies that an increase in MPCE\_MRP leads to an increase in food expenditure. This suggests that higher overall income or expenditure levels positively influence food spending.

Model Fit:

* The R-squared value of ~0.726 indicates that approximately 72.6% of the variance in food expenditure can be explained by the independent variables. This high R-squared value demonstrates a good fit of the model to the data.

Log-Transformed Model Interpretation

Coefficients and Their Significance:

* Intercept: The positive intercept (5.1052) in the log-transformed model suggests a baseline level of log-transformed food expenditure.
* foodtotal\_q (Quantity of Food): The coefficient (0.0463) indicates that a 1-unit increase in food quantity leads to approximately 0.0463 increase in the log of food expenditure, holding other factors constant. In practical terms, this suggests a less pronounced but still significant positive relationship.
* MPCE\_URP: The near-zero coefficient (-0.0000) and its non-significance (p-value = 0.069) in the log-transformed model indicate that MPCE\_URP has a negligible impact on the log-transformed food expenditure.
* MPCE\_MRP: The positive coefficient (0.0000) shows a very small but positive impact on the log-transformed food expenditure. This slight effect indicates that as MPCE\_MRP increases, the log of food expenditure also increases marginally.

Model Fit:

* The R-squared value of ~0.434 for the log-transformed model is lower than the original model, indicating that about 43.4% of the variance in the log-transformed food expenditure can be explained by the independent variables. This suggests that the log transformation did not improve the explanatory power of the model significantly and might not be the optimal transformation for this dataset.

Variance Inflation Factor (VIF)

Multicollinearity Check:

* The VIF values for the independent variables were all below 2 in both the R and Python analyses. VIF values below 10 are generally considered acceptable, and values below 2 indicate very low multicollinearity. This suggests that the predictors do not have strong linear relationships with each other and the estimates of the regression coefficients are reliable.

Residuals and Model Diagnostics

Original Model Residuals:

* The residuals (errors) for the original model showed a range indicating that while the model fits well, there are variations that are not captured by the independent variables. The residuals should be further analyzed for patterns that might suggest improvements to the model, such as non-linearity or heteroscedasticity.

Log-Transformed Model Residuals:

* The residuals for the log-transformed model showed a tighter range, which suggests a more normalized distribution of errors. However, the lower R-squared indicates that this transformation may not capture all the nuances in the data.

Summary

* Significance of Food Quantity: Both models consistently show that food quantity (foodtotal\_q) is a significant and strong predictor of food expenditure.
* Income Variables: The impact of the income-related variables (MPCE\_URP and MPCE\_MRP) is less straightforward, with MPCE\_MRP showing a consistent positive effect while MPCE\_URP has a minimal or negative impact.
* Model Fit and Transformations: The original model provides a better fit compared to the log-transformed model, suggesting that the relationship between the predictors and food expenditure is better captured in the original scale.

**Recommendations**

Policy Implications:

* Food expenditure is significantly influenced by the quantity of food (foodtotal\_q), which suggests that policies aimed at increasing food access and availability could effectively increase food expenditure and potentially improve nutrition.
* The negative coefficient for MPCE\_URP in the original model suggests that higher monthly per capita expenditure under Uniform Reference Period (URP) might be associated with lower food expenditure, possibly indicating a shift towards non-food expenditures as income increases.

Practical Applications:

* Businesses and marketers in the food industry can use these insights to target regions and demographics where food quantity significantly drives expenditure.
* Government and non-profits can design interventions to support low-income households in managing their food expenditures more effectively.

**References**

* IPL\_Ball\_by\_Ball dataset
* IPL\_Salaries\_2024 dataset
* ChatGPT