****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical Analysis and Modelling (SCMA 632)**

**A2:**

**NAVIYA MARY VINOD**

**V01108251**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **2** |
| **2.** | Results | **3** |
| **3.** | Interpretations | **3** |

**Introduction**

The aim of this study is to examine the consumption patterns in various districts of Andhra Pradesh using data from the National Sample Survey Office (NSSO) and seeks to identify the top and bottom three consuming districts within the state. This dataset offers extensive insights into household consumption behaviors, encompassing a range of food items such as rice, wheat, chicken, pulses, and other essentials. It delineates consumption trends across rural and urban sectors, as well as district-wise disparities. By leveraging the dataset and employing R, a robust statistical programming language, to conduct data manipulation and cleaning processes to extract pertinent information for analysis. Understanding these consumption dynamics is vital for devising targeted interventions and policy initiatives aimed at enhancing dietary habits and nutritional outcomes in Andhra Pradesh.

**Objectives**

1) Checking of any missing values in the data, identify them and replacing them with the mean of the variable.

2) Check for outliers and describe the outcome of the test and make suitable amendments.

3) Rename the districts as well as the sector, viz. rural and urban.

4) Summarize the critical variables in the data set region wise and district wise and indicate the top three districts and the bottom three districts of consumption.

5) Test whether the differences in the means are significant or not.

**Business Significance**

The focus of this study on Andhra Pradesh’s consumption patterns from NSSO data holds significant s for businesses and policymakers.

Policy Makers and Government Agencies will be able to identify nutritional deficiencies and excesses in specific regions or sectors by examining the consumption patterns. This insight aids in designing effective food security programs, public health interventions, and targeted nutritional assistance to improve the overall health of the population.

**Results And Interpretation**

### Initial Analysis

The multiple regression analysis aims to predict the total food expenditure (foodtotal\_q) based on various predictor variables. Here is a summary of the findings from the initial OLS regression:

#### Key Findings

1. **R-squared Value**: The model explains 31% of the variance in the dependent variable (R-squared = 0.310). This indicates a moderate fit.
2. **Significant Predictors**:
   1. **MPCE\_MRP (Monthly Per Capita Expenditure by Modified Mixed Recall Period)**: Significant positive relationship (coef = 0.0020, p < 0.001).
   2. **MPCE\_URP (Monthly Per Capita Expenditure by Uniform Recall Period)**: Significant positive relationship (coef = 0.0007, p < 0.001).
   3. **Age**: Significant positive relationship (coef = 0.1503, p < 0.001).
   4. **Meals\_At\_Home**: Significant positive relationship (coef = 0.0811, p < 0.001).
   5. **Possess\_ration\_card**: Significant negative relationship (coef = -3.0974, p < 0.001).
3. **Non-Significant Predictor**:
   1. **Education**: Not significant at the 5% level (coef = 0.0397, p = 0.130).
4. **Diagnostic Indicators**:
   1. **Omnibus Test**: Significant (p < 0.001), suggesting non-normality in the residuals.
   2. **Jarque-Bera Test**: Highly significant (p < 0.001), further confirming non-normality.
   3. **Durbin-Watson**: Value of 1.697 indicates slight positive autocorrelation.
   4. **Condition Number**: High (3.22e+04), indicating potential multicollinearity.

### Diagnostic Analysis

Given the diagnostics, several issues need to be addressed:

1. **Non-Normality of Residuals**:
   1. The significant Omnibus and Jarque-Bera tests suggest that the residuals are not normally distributed. This can be confirmed by plotting a histogram of residuals or a Q-Q plot.
2. **Multicollinearity**:
   1. The high condition number suggests multicollinearity. Variance Inflation Factor (VIF) analysis can identify which variables are contributing to this issue.

### Correction Steps

1. **Check for Multicollinearity**:
   1. Calculate VIF for each predictor variable. If VIF > 10, consider removing or combining variables.
2. **Transformations for Normality**:
   1. Apply transformations to the dependent variable or use robust regression techniques to handle non-normality.
3. **Residual Analysis**:
   1. Plot residuals to check for heteroscedasticity. Consider using Weighted Least Squares (WLS) if heteroscedasticity is present.

### Reanalysis

#### Multicollinearity Check

Let's calculate the VIF values for each predictor to address multicollinearity.

#### Normality and Heteroscedasticity

Investigate the distribution of residuals and apply necessary transformations or robust regression techniques.

### Interpretation

* **VIF Values**: Identify if multicollinearity is present and take corrective actions if needed.
* **Residual Plots**: Determine the normality and heteroscedasticity of residuals. Apply necessary transformations or consider robust regression if assumptions are violated.

### Revised Results

After addressing multicollinearity and normality issues, re-estimate the model and compare the results. Note any significant changes in the coefficients, significance levels, or model fit statistics.

### Final Findings

1. **Revised Model Fit**: Evaluate if the adjusted R-squared improves.
2. **Significance of Predictors**: Check if the significance levels of the predictors change.
3. **Assumption Validations**: Ensure that the corrected model satisfies regression assumptions

### Interpretation of VIF Results

The Variance Inflation Factor (VIF) quantifies the level of multicollinearity in a regression model. A VIF value exceeding 10 typically indicates high multicollinearity, which can inflate the standard errors of the coefficients and lead to unreliable results.

#### Analysis of VIF Values

**Constant (const)**:

* 1. The VIF for the constant term is exceptionally high (61.884). This is common since the constant term is a linear combination of the predictors. High VIF for the constant is generally not a concern as it doesn’t affect the multicollinearity of the predictors themselves.

**MPCE\_MRP**:

* 1. The VIF for \text{MPCE\_MRP} is 2.566. This value is below the threshold of 10, indicating that \text{MPCE\_MRP} does not exhibit problematic multicollinearity with the other predictors.

**MPCE\_URP**:

* 1. The VIF for \text{MPCE\_URP} is 2.311, also well below the threshold. Like \text{MPCE\_MRP}, it suggests that \text{MPCE\_URP} does not have significant multicollinearity issues.

**Age**:

* 1. The VIF for Age\text{Age}Age is 1.134, which is very low, indicating that Age\text{Age}Age is almost completely independent of the other predictors in the model.

**Meals\_At\_Home**:

* 1. The VIF for \text{Meals\_At\_Home} is 1.146. This low VIF value indicates that it does not suffer from multicollinearity problems.

**Possess\_ration\_card**:

* 1. The VIF for \text{Possess\_ration\_card} is 1.214, another low value suggesting minimal multicollinearity.

**Education**:

* 1. The VIF for Education\text{Education}Education is 1.375, which is still well below 10. This indicates that the variable Education\text{Education}Education does not present multicollinearity issues with other predictors in the model.

### Implications for the Regression Model

The VIF analysis reveals that none of the predictor variables exhibit problematic multicollinearity. All VIF values are well below the threshold of 10, indicating that each predictor provides unique information about the dependent variable foodtotal\_q\text{foodtotal\\_q}foodtotal\_q. This means the regression coefficients are reliable, and their standard errors are not inflated due to multicollinearity.

Given this context, we do not need to remove or combine any predictors based on multicollinearity concerns. The regression model is statistically sound in terms of multicollinearity, allowing us to proceed with other diagnostic checks and model improvements, such as addressing autocorrelation and non-normality of residuals.

### Revisiting the Regression Model

Next, we address the normality of residuals and potential autocorrelation. The original regression diagnostics indicated non-normality and some level of autocorrelation, suggesting the need for further model adjustments, such as transforming the dependent variable or using more advanced regression techniques.

**R**

### Regression Model Summary

The regression model is fitted with the following predictors:

* MPCE\_MRP
* MPCE\_URP
* Age
* Meals\_At\_Home
* Possess\_ration\_card
* Education

#### Coefficients:

* **Intercept**: 7.6657.6657.665
* **MPCE\_MRP**: 2.169×10−32.169 \times 10^{-3}2.169×10−3, highly significant (p-value < 2e-16)
* **MPCE\_URP**: 6.595×10−46.595 \times 10^{-4}6.595×10−4, highly significant (p-value < 2e-16)
* **Age**: 1.110×10−11.110 \times 10^{-1}1.110×10−1, highly significant (p-value < 2e-16)
* **Meals\_At\_Home**: 9.251×10−29.251 \times 10^{-2}9.251×10−2, highly significant (p-value < 2e-16)
* **Possess\_ration\_card**: −1.576-1.576−1.576, highly significant (p-value < 2e-16)
* **Education**: 2.567×10−32.567 \times 10^{-3}2.567×10−3, not significant (p-value = 0.917)

#### Model Fit:

* **Multiple R-squared**: 0.3319
* **Adjusted R-squared**: 0.3313
* **F-statistic**: 560.6, p-value < 2.2e-16

The high significance levels for most predictors indicate they have a strong relationship with the dependent variable, foodtotal\_q. However, Education appears to have little to no impact as its p-value is very high (0.917).

### Multicollinearity Check (VIF):

The VIF values are as follows:

* **MPCE\_MRP**: 2.536
* **MPCE\_URP**: 2.288
* **Age**: 1.114
* **Meals\_At\_Home**: 1.057
* **Possess\_ration\_card**: 1.022
* **Education**: 1.094

Typically, a VIF value above 10 indicates high multicollinearity that may affect the regression model's reliability. In this case, all VIF values are well below 10, suggesting that multicollinearity is not a concern for your model.

### Interpretation:

1. **MPCE\_MRP and MPCE\_URP**: Both are positively correlated with foodtotal\_q, suggesting higher per capita expenditure is associated with higher food quantity.
2. **Age**: There is a positive correlation with foodtotal\_q, indicating older individuals might spend more on food.
3. **Meals\_At\_Home**: Positively correlated, suggesting more meals at home correlate with higher food expenditure.
4. **Possess\_ration\_card**: Negatively correlated, indicating those with a ration card might spend less on food, possibly due to subsidies.
5. **Education**: No significant effect on food expenditure.

### Recommendations:

* **Model Refinement**: Since Education is not significant, you may consider removing it from the model and refitting to see if the adjusted R-squared improves.
* **Further Analysis**: Investigate why Possess\_ration\_card has a negative coefficient, potentially exploring the interaction between ration card possession and other socioeconomic factors.