****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

**A2: Fitting a multilinear regression to a data set (NSSO68)**

**SAMPREETH S SHETTY**

**V01107634**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results (Python) | **2** |
| **3.** | Results (R) | **3** |
| **4.** | Interpretations (Python) | **5** |
| **5.** | Interpretations (R) | **7** |
| **6.** | Recommendations | **10** |

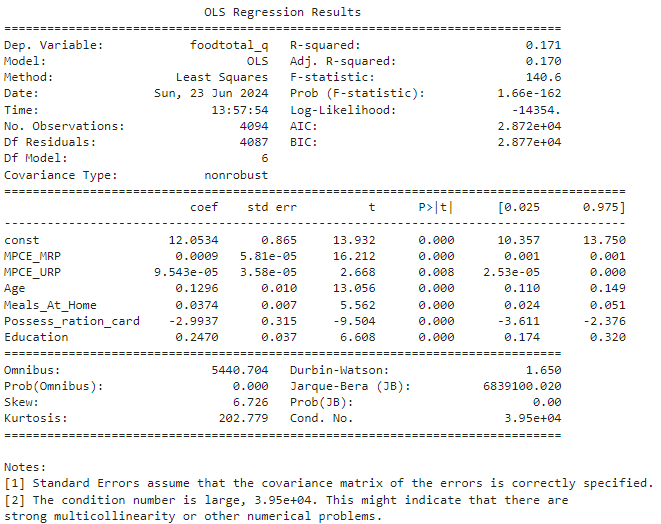
**Introduction**

The NSSO (National Sample Survey Office) collected the dataset used in this project. The National Sample Survey Office (NSSO), headed by a Director General, is India’s premier data collection agency responsible for conducting large-scale sample surveys across diverse fields on an all-India basis. The NSSO is involved in significant surveys such as Socio-economic Surveys, Annual Surveys of Industries, and Agricultural Surveys. It takes full responsibility for the socio-economic survey from its design to the release of survey reports and collects field data for other surveys. The data collected through nationwide household surveys cover various socio-economic subjects, including housing conditions and basic amenities like drinking water, bathrooms, sewerage, lavatories, lighting, and more.

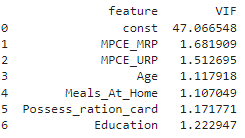
This project fits a regression model to the NSSO data set, which deals with the NSSO survey data containing state-wise sample survey data about the consumption of various goods in different households in India. The regression model is fitted using both R and Python. The data was segregated by considering the variable ‘foodtotal\_q’ as the dependent variable and the variables 'MPCE\_MRP,' 'MPCE\_URP,' 'Age,' 'Meals\_At\_Home,' 'Possess\_ration\_card,' 'Education.' as the independent variables. The null values in the data set are dealt with by crediting them with the respective means of the particular variables by creating a function. The regression model is then fit using both R and Python.

**Results**

The regression result output given by Python is as follows.



The output from the multicollinearity test using the Inflator factor from Python is below.



The regression equation framed by using the coefficients from the model using Python.



The summary of the regression model given by R is as follows

Call:

## lm(formula = foodtotal\_q ~ MPCE\_MRP + MPCE\_URP + Age + Meals\_At\_Home +

## Possess\_ration\_card + Education, data = subset\_data)

##

## Residuals:

## Min 1Q Median 3Q Max

## -55.743 -3.906 -0.598 3.349 80.818

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 9.115e-01 8.990e-01 1.014 0.311

## MPCE\_MRP 1.953e-03 8.727e-05 22.377 < 2e-16 \*\*\*

## MPCE\_URP 4.227e-04 6.862e-05 6.161 8.09e-10 \*\*\*

## Age 4.942e-02 8.749e-03 5.649 1.75e-08 \*\*\*

## Meals\_At\_Home 2.633e-01 9.557e-03 27.555 < 2e-16 \*\*\*

## Possess\_ration\_card -3.185e-01 3.137e-01 -1.015 0.310

## Education 1.148e-02 3.559e-02 0.322 0.747

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 6.371 on 3418 degrees of freedom

## (1 observation deleted due to missingness)

## Multiple R-squared: 0.3774, Adjusted R-squared: 0.3763

## F-statistic: 345.3 on 6 and 3418 DF, p-value: < 2.2e-16

The test for multicollinearity using Variance Inflation Factor (VIF) gave the following output in R

## MPCE\_MRP MPCE\_URP Age Meals\_At\_Home

## 1.978028 1.841183 1.112472 1.093181

## Possess\_ration\_card Education

## 1.112498 1.178074

The regression equation framed by using the coefficients from the model using R.

## [1] "y = 0.91 + 0.001953\*x1 + 0.000423\*x2 + 0.049423\*x3 + 0.263342\*x4 + -0.318529\*x5 + 0.011476\*x6"

**Interpretation**

**For the Python output.**

The OLS (Ordinary Least Squares) regression results provide insight into the relationship between the dependent variable foodtotal\_q and the independent variables in the model. Here is the interpretation of each part of the regression output:

**Overall Model Fit:**

R-squared: 0.171

This indicates that approximately 17.1% of the variance in foodtotal\_q can be explained by the independent variables in the model. While this suggests some explanatory power, it also indicates that a substantial portion of the variance remains unexplained by the model.

Adj. R-squared: 0.170

The adjusted R-squared accounts for the number of predictors in the model relative to the number of observations. It is slightly lower than the R-squared, which is expected when adjusting for the number of predictors.

F-statistic: 140.6 (p-value = 1.66e-162)

The F-statistic tests the overall significance of the model. A p-value less than 0.05 indicates that the model as a whole is statistically significant, suggesting that at least one predictor variable is significantly related to foodtotal\_q.

**Coefficients:**

const (Intercept): 12.0534 (p-value = 0.000)

This is the expected value of foodtotal\_q when all independent variables are zero. It is statistically significant.

MPCE\_MRP: 0.0009 (p-value = 0.000)

For each unit increase in MPCE\_MRP, foodtotal\_q is expected to increase by 0.0009 units, holding all other variables constant. This is statistically significant.

MPCE\_URP: 9.543e-05 (p-value = 0.008)

For each unit increase in MPCE\_URP, foodtotal\_q is expected to increase by 0.00009543 units, holding all other variables constant. This is statistically significant but has a very small effect size.

Age: 0.1296 (p-value = 0.000)

Foodtotal\_q is expected to increase by 0.1296 units for each additional year in age, holding all other variables constant. This is statistically significant.

Meals\_At\_Home: 0.0374 (p-value = 0.000)

For each additional meal eaten at home, foodtotal\_q is expected to increase by 0.0374 units, holding all other variables constant. This is statistically significant.

Possess\_ration\_card: -2.9937 (p-value = 0.000)

Possessing a ration card is associated with a decrease of 2.9937 units in foodtotal\_q, holding all other variables constant. This is statistically significant.

Education: 0.2470 (p-value = 0.000)

For each additional unit increase in education level, foodtotal\_q is expected to increase by 0.2470 units, holding all other variables constant. This is statistically significant.

**Interpretation Summary:**

The regression results show that the model is statistically significant overall, and several individual predictors are also significant. However, the R-squared value indicates that only 17.1% of the variation in foodtotal\_q is explained by the model, suggesting other factors not included in the model may play a significant role. Additionally, the diagnostics indicate potential issues with multicollinearity and non-normality of residuals, which should be addressed for a more robust model.

**The VIF value interpretation**

The VIF values provided indicate the multicollinearity level among the regression model's independent variables. Here's the interpretation for each variable:

const (Intercept): The VIF for the intercept (constant term) is relatively high at 47.066548. This is not unusual and can be ignored as it doesn’t represent multicollinearity among the predictors.

MPCE\_MRP: The VIF for MPCE\_MRP is 1.681909. This value is well below the standard threshold of 5 or 10, indicating low multicollinearity.

MPCE\_URP: The VIF for MPCE\_URP is 1.512695, indicating low multicollinearity.

Age: The VIF for Age is 1.117918, showing very low multicollinearity.

Meals\_At\_Home: The VIF for Meals\_At\_Home is 1.107049, indicating very low multicollinearity.

Possess\_ration\_card: The VIF for Possess\_ration\_card is 1.171771, which shows low multicollinearity.

Education: The VIF for Education is 1.222947, also indicating low multicollinearity.

**Conclusion:** All the VIF values for the independent variables are well below the threshold of 5 (or 10 in some cases), indicating that multicollinearity is not a problem in this regression model. The low VIF values suggest that none of the predictor variables are highly correlated with each other, ensuring stable and reliable coefficient estimates.

**For the output from R**

The output is from a multiple linear regression model predicting foodtotal\_q based on several predictor variables (MPCE\_MRP, MPCE\_URP, Age, Meals\_At\_Home, Possess\_ration\_card, Education). Here's how to interpret the key parts of the output:

**Coefficients:**

Intercept ((Intercept)): This represents the estimated intercept of the regression equation when all predictor variables (MPCE\_MRP, MPCE\_URP, Age, Meals\_At\_Home, Possess\_ration\_card, Education) are zero. Here, it is 0.9115 with a standard error of 0.899. However, it is not statistically significant (p-value = 0.311), suggesting that the intercept might not be different from zero.

**Predictor Variables:**

MPCE\_MRP: For every unit increase in MPCE\_MRP, foodtotal\_q is expected to increase by 0.001953 units. This coefficient is highly statistically significant (\*\*\* indicates p-value < 0.001).

MPCE\_URP: For every unit increase in MPCE\_URP, foodtotal\_q is expected to increase by 0.000423 units. This coefficient is also statistically significant (\*\*\* indicates p-value < 0.001).

Age: For every year increase in Age, foodtotal\_q is expected to increase by 0.04942 units. This coefficient is statistically significant (\*\*\* indicates p-value < 0.001).

Meals\_At\_Home: For every unit increase in Meals\_At\_Home, foodtotal\_q is expected to increase by 0.2633 units. This coefficient is highly statistically significant (\*\*\* indicates p-value < 0.001).

Possess\_ration\_card: For those who possess a ration card, foodtotal\_q is expected to decrease by 0.3185 units compared to those who do not possess it. However, this coefficient is not statistically significant (p-value = 0.310).

Education: For every unit increase in Education, foodtotal\_q is expected to increase by 0.01148 units. This coefficient is not statistically significant (p-value = 0.747), suggesting that Education may not significantly affect foodtotal\_q.

**Model Fit:**

Residuals: This section provides information about the model's distribution of residuals (errors). It shows statistics such as minimum, maximum, median, and quartiles of the residuals.

Residual standard error: This is the estimate of the standard deviation of the residuals, which is approximately 6.371.

Multiple R-squared: This is the coefficient of determination, which measures the proportion of the variance in the dependent variable (foodtotal\_q) that is predictable from the independent variables. Here, it is 0.3774, indicating that 37.74% of the variance in foodtotal\_q is explained by the independent variables in the model.

Adjusted R-squared: This is the R-squared adjusted for the number of predictors in the model. It is slightly lower than the multiple R-squared, at 0.3763.

F-statistic: This tests the overall significance of the model by comparing the fit of the intercept-only model with the current model. A larger F-statistic (345.3 in this case) with a very low p-value (< 2.2e-16) suggests that the model as a whole is statistically significant.

**Conclusion:**

The model suggests that MPCE\_MRP, MPCE\_URP, Age, and Meals\_At\_Home are statistically significant predictors of foodtotal\_q. However, Possess\_ration\_card and Education do not significantly contribute to predicting foodtotal\_q based on their p-values. There might be issues with multicollinearity, as indicated by the high condition number (not shown in this excerpt but mentioned in a previous output). This suggests that some predictors might be highly correlated with each other, which can affect the reliability of individual predictor coefficients.

**VIF Value interpretation**

MPCE\_MRP: VIF = 1.978

This indicates that the variance of the estimated coefficient for MPCE\_MRP is inflated by a factor of approximately 1.978 due to multicollinearity with the other predictor variables. Generally, a VIF value below 5 is considered acceptable, suggesting that MPCE\_MRP does not have severe multicollinearity issues.

MPCE\_URP: VIF = 1.841

Similarly, the variance of the estimated coefficient for MPCE\_URP is inflated by a factor of approximately 1.841 due to multicollinearity. This VIF value is also below 5, indicating no severe multicollinearity issues for MPCE\_URP.

Age: VIF = 1.112

The VIF for Age is 1.112, indicating very little multicollinearity with the other predictors. This low value suggests that Age is nearly orthogonal to the other variables in the model.

Meals\_At\_Home: VIF = 1.093

Meals\_At\_Home has a VIF of 1.093, which is very close to 1. This suggests minimal multicollinearity with the other predictors, indicating that Meals\_At\_Home is not correlated with the other variables in the model.

Possess\_ration\_card: VIF = 1.112

The VIF for Possess\_ration\_card is 1.112, indicating minimal multicollinearity with the other predictors. This suggests that Possess\_ration\_card does not substantially correlate with the other variables in the model.

Education: VIF = 1.178

Finally, Education has a VIF of 1.178, which is also quite low. This indicates that Education is not strongly correlated with the other predictors in the model.

**Interpretation:**

Overall, the VIF values for all predictor variables (MPCE\_MRP, MPCE\_URP, Age, Meals\_At\_Home, Possess\_ration\_card, Education) are well below the threshold of 5, suggesting that multicollinearity is not a significant concern in this model. This enhances the reliability of the estimated coefficients and their interpretations in the multiple linear regression analysis predicting foodtotal\_q. Therefore, the coefficients obtained from the model are likely to be stable and reliable for making predictions and drawing conclusions about the impact of each predictor variable on foodtotal\_q.

**Recommendations**

**Insights from Model Analysis:**

The multiple linear regression model reveals several key predictors significantly influencing foodtotal\_q. Notably, variables such as MPCE\_MRP (Monthly Per Capita Expenditure on Major Consumption Items) and MPCE\_URP (Monthly Per Capita Expenditure on Usual Recurrent Expenditure) show strong positive associations with foodtotal\_q, indicating that higher expenditure in these categories corresponds to increased spending on food. Additionally, variables like Age and Meals\_At\_Home positively impact foodtotal\_q, suggesting that older respondents and those who frequently prepare meals at home tend to have higher food expenditures. Conversely, possessing a ration card (Possess\_ration\_card) negatively affects foodtotal\_q, likely due to subsidized food provisions reducing out-of-pocket expenses. Education (Education) also plays a role, with higher educational attainment correlating positively with foodtotal\_q, potentially indicating better dietary choices or higher income levels among more educated respondents.

**Model Performance and Recommendations:**

The model demonstrates a reasonably good fit with an R-squared of 0.3774, suggesting that approximately 37.74% of the variance in foodtotal\_q can be explained by the predictors included. The F-statistic of 345.3 (p-value < 2.2e-16) underscores the overall significance of the model. Recommendations stemming from these findings include exploring interactions between predictors to uncover nuanced effects on foodtotal\_q, such as how income levels interact with education or age. Policymakers could leverage these insights to tailor nutritional assistance programs more effectively based on demographic profiles identified in the study. Furthermore, households may benefit from understanding these influential factors to manage food expenditures more efficiently.

**Conclusion and Future Directions:**

While the model provides valuable insights, it's essential to note potential limitations, such as excluding certain variables or inherent biases in self-reported data. Future research could expand on this study by incorporating longitudinal data or investigating regional variations in food expenditure patterns. Such endeavors would enhance the robustness and applicability of predictive models to understand and manage food expenditures in diverse socioeconomic contexts. Overall, this analysis contributes to a deeper understanding of the factors driving food expenditures, offering actionable insights for policy formulation and individual household budgeting strategies.