VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modeling (SCMA 632)

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

POLAMREDDY MADHUMITHA

V01107517

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**

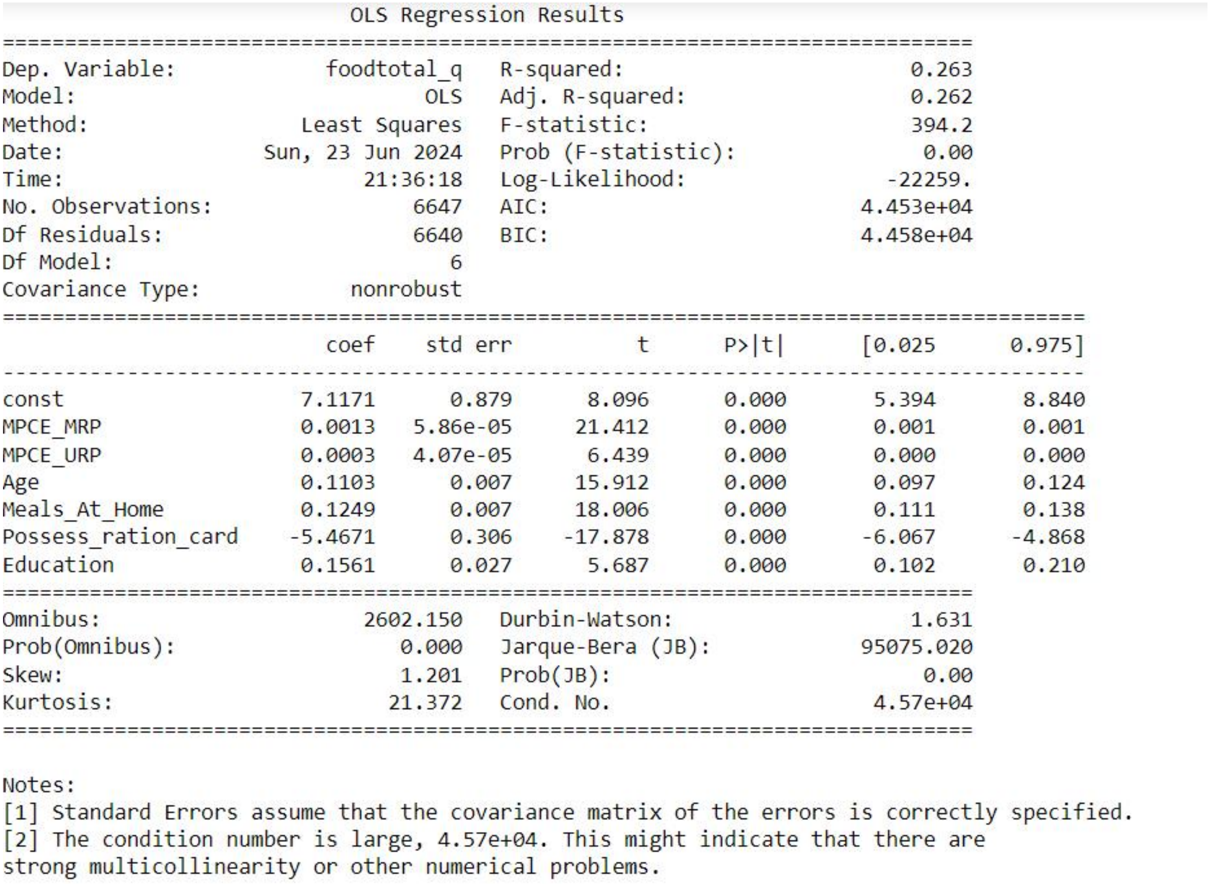
Introduction

The dataset utilized in this project was gathered by the National Sample Survey Office, or NSSO. The principal data collection organization in India, the National Sample Survey Office (NSSO), is led by a Director General and is in charge of carrying out extensive sample surveys throughout the country in a variety of sectors. Important surveys conducted by the NSSO include annual surveys of industries, agricultural surveys, and socioeconomic surveys. It gathers field data for other surveys and assumes complete responsibility for the socio-economic survey, from its conception to the publication of survey findings. Nationwide household surveys are used to gather data on a variety of socioeconomic topics, such as housing conditions and the availability of essential utilities like lights, restrooms, sewage, bathrooms, and drinking water.

In this research, the NSSO survey data—which includes state-by-state sample survey data regarding the consumption of various items in diverse Indian households—is fitted to a regression model. Python and R are both used in the regression model's fitting. The variables "MPCE_MRP," "MPCE_URP," "Age," "Meals_At_Home," "Possess_ration_card," and "Education" were regarded as the independent variables, and the variable "foodtotal_q" as the dependent variable. In order to handle the null values in the data set, a function is created and the corresponding means of the relevant variables are credited to them. Then, R and Python are used to fit the regression model.

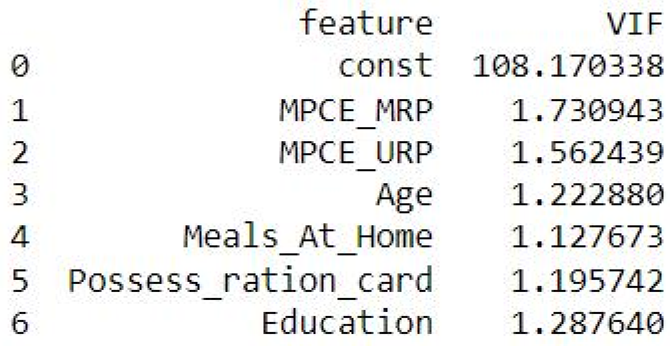
Results

The regression result output given by Python is as follows.



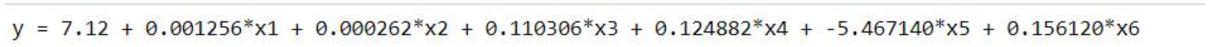
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            foodtotal_q   R-squared:                       0.263
Model:                            OLS   Adj. R-squared:                  0.262
Method:                 Least Squares   F-statistic:                     394.2
Date:                Sun, 23 Jun 2024   Prob (F-statistic):               0.00
Time:                        21:36:18   Log-Likelihood:                -22259.
No. Observations:                6647   AIC:                         4.453e+04
Df Residuals:                    6640   BIC:                         4.458e+04
Df Model:                           6
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   7.1171      0.879      8.096      0.000       5.394       8.840
MPCE_MRP                0.0013   5.86e-05     21.412      0.000       0.001       0.001
MPCE_URP                0.0003   4.07e-05      6.439      0.000       0.000       0.000
Age                     0.1103      0.007     15.912      0.000       0.097       0.124
Meals_At_Home           0.1249      0.007     18.006      0.000       0.111       0.138
Possess_ration_card    -5.4671      0.306    -17.878      0.000      -6.067      -4.868
Education               0.1561      0.027      5.687      0.000       0.102       0.210
==============================================================================
Omnibus:                     2602.150   Durbin-Watson:                   1.631
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            95075.020
Skew:                           1.201   Prob(JB):                         0.00
Kurtosis:                      21.372   Cond. No.                     4.57e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.57e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

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



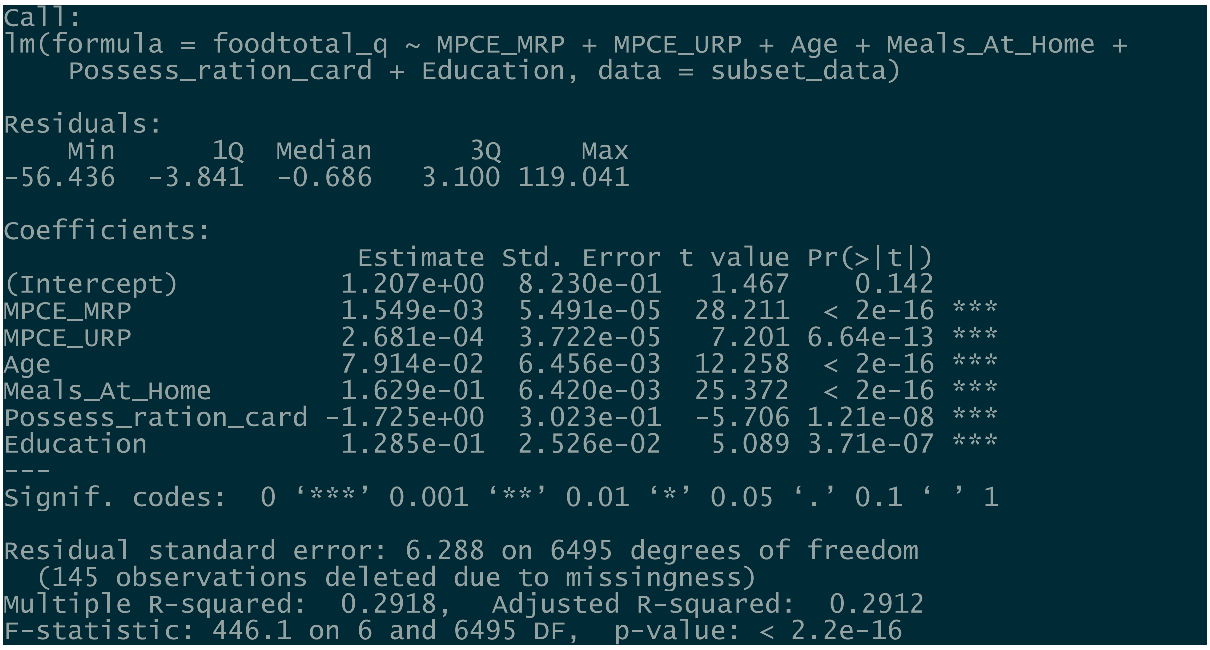
```
               feature         VIF
0                const  108.170338
1             MPCE_MRP    1.730943
2             MPCE_URP    1.562439
3                  Age    1.222880
4        Meals_At_Home    1.127673
5  Possess_ration_card    1.195742
6            Education    1.287640
```

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



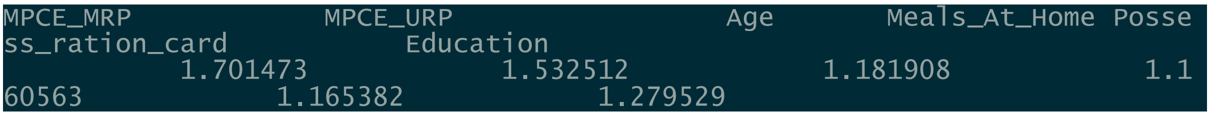
```
y = 7.12 + 0.001256*x1 + 0.000262*x2 + 0.110306*x3 + 0.124882*x4 + -5.467140*x5 + 0.156120*x6
```

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



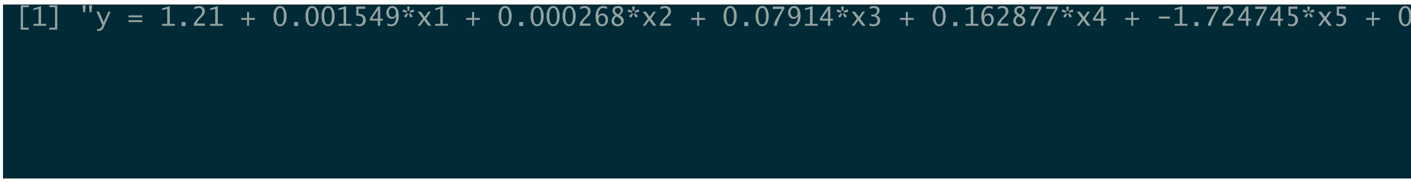
```
Call:
lm(formula = foodtotal_q ~ MPCE_MRP + MPCE_URP + Age + Meals_At_Home +
    Possess_ration_card + Education, data = subset_data)

Residuals:
    Min      1Q  Median      3Q     Max
-56.436  -3.841  -0.686   3.100 119.041

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.207e+00  8.230e-01   1.467    0.142
MPCE_MRP             1.549e-03  5.491e-05  28.211  < 2e-16 ***
MPCE_URP             2.681e-04  3.722e-05   7.201 6.64e-13 ***
Age                  7.914e-02  6.456e-03  12.258  < 2e-16 ***
Meals_At_Home        1.629e-01  6.420e-03  25.372  < 2e-16 ***
Possess_ration_card -1.725e+00  3.023e-01  -5.706 1.21e-08 ***
Education            1.285e-01  2.526e-02   5.089 3.71e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.288 on 6495 degrees of freedom
  (145 observations deleted due to missingness)
Multiple R-squared:  0.2918,   Adjusted R-squared:  0.2912
F-statistic: 446.1 on 6 and 6495 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  Posse
ss_ration_card          Education
           1.701473            1.532512            1.181908            1.1
60563            1.165382            1.279529
```

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



```
[1] "y = 1.21 + 0.001549*x1 + 0.000268*x2 + 0.07914*x3 + 0.162877*x4 + -1.724745*x5 + 0
```

Interpretation

For the Python output.

The findings of the OLS (Ordinary Least Squares) regression shed light on how the independent variables in the model and the dependent variable foodtotal_q relate to one another. This is how each component of the regression output is interpreted.:

Overall Model Fit:

R-squared: 0.2918

This indicates that approximately 29.1% 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: 0.2912

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: 446.1 (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.

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.