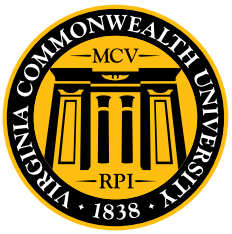


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A2a: Multiple Regression Analysis and Diagnostics on NSSO68 Dataset for the State of Goa

SATHWIK NAG CHANNAGIRI VENKATESH

V01107764

Date of Submission: 23-06-2024

CONTENTS

INTRODUCTION

This analysis aims to process and scrutinize the NSSO68 dataset for Goa, focusing on understanding the determinants of food consumption. Our primary objectives are to conduct a multiple regression analysis to identify the key factors influencing food consumption, diagnose potential issues with the model, and correct any identified problems to ensure robust findings. We start by handling missing values through imputation, ensuring that the dataset is complete and ready for analysis. Then, we fit a multiple regression model to examine the relationship between food consumption and various predictors, including Monthly Per Capita Expenditure (MPCE), age, home meals, ration card possession, and education level.

Objectives:

- Conduct Multiple Regression Analysis
- Handle Missing Data
- Perform Regression Diagnostics
- Explain Findings
- Correct Identified Issues
- Present Findings and Recommendations.

RESULTS & INTERPRETATION

Multiple Regression Analysis Explanation and Findings

Data Preparation

Missing Value Check and Imputation:

- Initially, there were missing values in the Education column

```
> sum(is.na(subset_data$MPCE_MRP))
[1] 0
> sum(is.na(subset_data$MPCE_URP))
[1] 0
> sum(is.na(subset_data$Age))
[1] 0
> sum(is.na(subset_data$Possess_ration_card))
[1] 0
> sum(is.na(data$Education))
[1] 7
```

```
In [9]: # Check for missing values
        print(subset_data['MPCE_MRP'].isna().sum())
        print(subset_data['MPCE_URP'].isna().sum())
        print(subset_data['Age'].isna().sum())
        print(subset_data['Possess_ration_card'].isna().sum())
        print(data['Education'].isna().sum())
```

```
0
0
0
0
7
```

- Imputation was performed using the mean of the Education column. Post-imputation, no missing values remained in the Education column.

```
> # Columns to impute
> columns_to_impute <- c("Education")
> # Impute missing values with mean
> data <- impute_with_mean(data, columns_to_impute)
> sum(is.na(data$Education))
[1] 0
```

```
In [10]: # Impute missing values with mean
         subset_data['Education'].fillna(subset_data['Education'].mean(), inplace=True)
         print(subset_data['Education'].isna().sum())
```

```
0
```

Model Fitting

Regression Model:

- The multiple regression model was fitted using the formula:
- he dependent variable is foodtotal_q, and the independent variables are MPCE_MRP, MPCE_URP, Age, Meals_At_Home, Possess_ration_card, and Education.

```
# Fit the regression model
model <- lm(foodtotal_q~ MPCE_MRP+MPCE_URP+Age+Meals_At_Home+Possess_ration_card+Education, data = subset_data)
```

```
In [19]: # Fit the regression model
         model = ols('foodtotal_q ~ MPCE_MRP + MPCE_URP + Age + Meals_At_Home + Possess_ration_card + Education', data=subset_data).fit()
```

Regression Results

Summary of the Model:

```
> print(summary(model))

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
-24.8627  -3.1793  -0.5373   2.9313  23.3686

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.6802449  3.8911589   2.488  0.01323 *
MPCE_MRP           0.0018788  0.0002439   7.703 8.95e-14 ***
MPCE_URP          -0.0001224  0.0001688  -0.725  0.46860
Age                0.0038182  0.0234588   0.163  0.87078
Meals_At_Home      0.1296228  0.0514332   2.520  0.01208 *
Possess_ration_card -2.2872695  1.3792227  -1.658  0.09796 .
Education          0.2468905  0.0932145   2.649  0.00837 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.618 on 438 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.2865,    Adjusted R-squared:  0.2767
F-statistic: 29.31 on 6 and 438 DF,  p-value: < 2.2e-16
```

```
In [12]: # Print the regression results
         print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            foodtotal_q   R-squared:                       0.286
Model:                            OLS   Adj. R-squared:                  0.277
Method:                 Least Squares   F-statistic:                     29.31
Date:                Sun, 23 Jun 2024   Prob (F-statistic):           1.60e-29
Time:                        20:52:19   Log-Likelihood:                -1396.0
No. Observations:                 445   AIC:                             2806.
Df Residuals:                     438   BIC:                             2835.
Df Model:                           6
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                  9.6802      3.891      2.488      0.013       2.033      17.328
MPCE_MRP                   0.0019      0.000      7.703      0.000       0.001       0.002
MPCE_URP                  -0.0001      0.000     -0.725      0.469      -0.000       0.000
Age                        0.0038      0.023      0.163      0.871      -0.042       0.050
Meals_At_Home              0.1296      0.051      2.520      0.012       0.029       0.231
Possess_ration_card       -2.2873      1.379     -1.658      0.098      -4.998       0.423
Education                  0.2469      0.093      2.649      0.008       0.064       0.430
==============================================================================
Omnibus:                       42.314   Durbin-Watson:                   1.685
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              122.715
Skew:                           0.423   Prob(JB):                     2.25e-27
Kurtosis:                       5.429   Cond. No.                     7.72e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.72e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- **Residuals**:
  - The residuals show a somewhat symmetric distribution with a median close to zero, indicating a reasonably good fit.
- **Coefficients**:
  - MPCE_MRP (Monthly Per Capita Expenditure based on MRP): Positive and highly significant ($p < 0.001$).
  - MPCE_URP (Monthly Per Capita Expenditure based on URP): Negative and not significant ($p = 0.4686$).
  - Age: Positive but not significant ($p = 0.87078$).
  - Meals_At_Home: Positive and significant ($p = 0.01208$).
  - Possess_ration_card: Negative and marginally significant ($p = 0.09796$).
  - Education: Positive and significant ($p = 0.00837$).
- **Model Fit**:
  - Multiple R-squared: 0.2865
  - Adjusted R-squared: 0.2767
  - F-statistic: 29.31 ($p < 2.2e-16$), indicating the model is statistically significant overall.

Diagnostics

Multicollinearity Check:

- Variance Inflation Factor (VIF) values for all variables are below the threshold of 8, indicating no significant multicollinearity issues:

```
> # Check for multicollinearity using Variance Inflation Factor (VIF)
> vif(model) # VIF Value more than 8 its problematic
          MPCE_MRP           MPCE_URP                Age       Meals_At_Home Possess_ration_card
          2.893338           2.675995           1.338726           1.171905           1.096270
         Education
          1.376191
```

```
In [21]: # Check for multicollinearity using Variance Inflation Factor (VIF)
         X = subset_data[['MPCE_MRP', 'MPCE_URP', 'Age', 'Meals_At_Home', 'Possess_ration_card', 'Education']]
         X = sm.add_constant(X)
         vif_data = pd.DataFrame()
         vif_data["feature"] = X.columns
         vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
         print(vif_data)
```

	feature	VIF
0	const	213.483072
1	MPCE_MRP	2.893338
2	MPCE_URP	2.675995
3	Age	1.338726
4	Meals_At_Home	1.171905
5	Possess_ration_card	1.096270
6	Education	1.376191

Regression Equation:

- The constructed regression equation based on the coefficients is:

```
> # Print the equation
> print(equation)
[1] "y = 9.68 + 0.001879*x1 + -0.000122*x2 + 0.003818*x3 + 0.129623*x4 + -2.28727*x5 + 0.246891*x6"
```

```
In [23]: # Construct the equation
equation = f"y = {round(coefficients[0], 2)}"
for i in range(1, len(coefficients)):
    equation += f" + {round(coefficients[i], 6)}*x{i}"
print(equation)
```

```
y = 9.68 + 0.001879*x1 + -0.000122*x2 + 0.003818*x3 + 0.129623*x4 + -2.28727*x5 + 0.246891*x6
```

where:

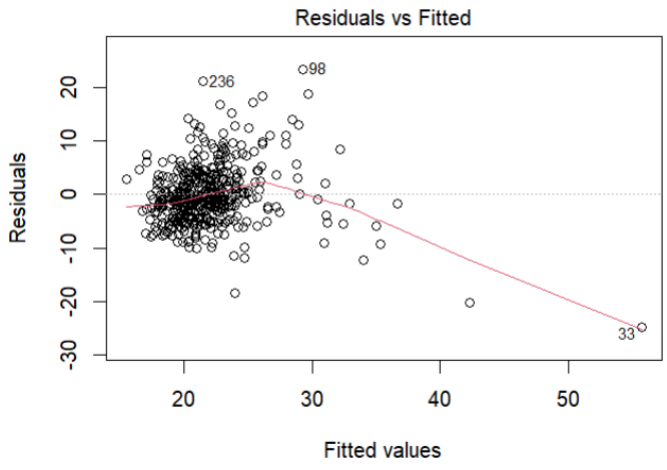
- x1 = MPCE_MRP
- x2 = MPCE_URP
- x3 = Age
- x4 = Meals_At_Home
- x5 = Possess_ration_card
- x6 = Education

- **Significant Predictors**:
  - MPCE_MRP and Meals_At_Home have positive and significant effects on foodtotal_q.
  - Education also has a positive and significant effect, suggesting higher education levels are associated with higher food expenditure.
- **Non-significant Predictors**:
  - MPCE_URP and Age do not significantly influence foodtotal_q in this model.
  - Possess_ration_card shows a marginal negative significance, indicating those with ration cards might have slightly lower food expenditure, though this is not strongly significant.
- **Model Fit**:
  - The R-squared value of 0.2865 indicates that approximately 28.65% of the variability in food expenditure is explained by the model. This is a moderate level of explanatory power.

Interpretation of Residual Plots and Breusch-Pagan Test

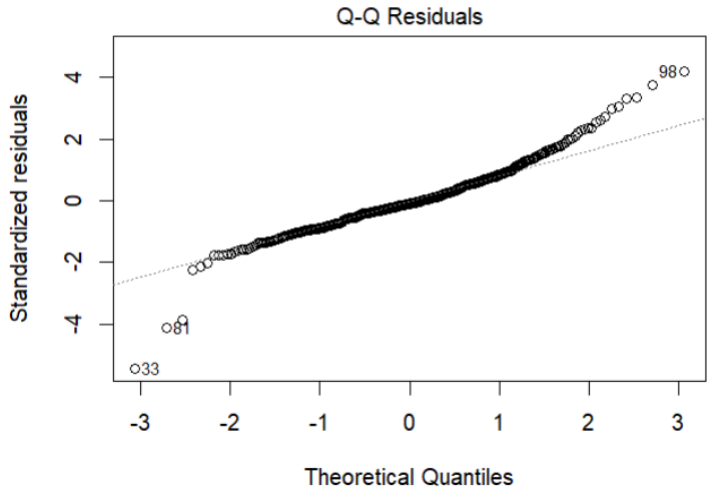
Residual Plots

1. **Residuals vs Fitted Plot**:



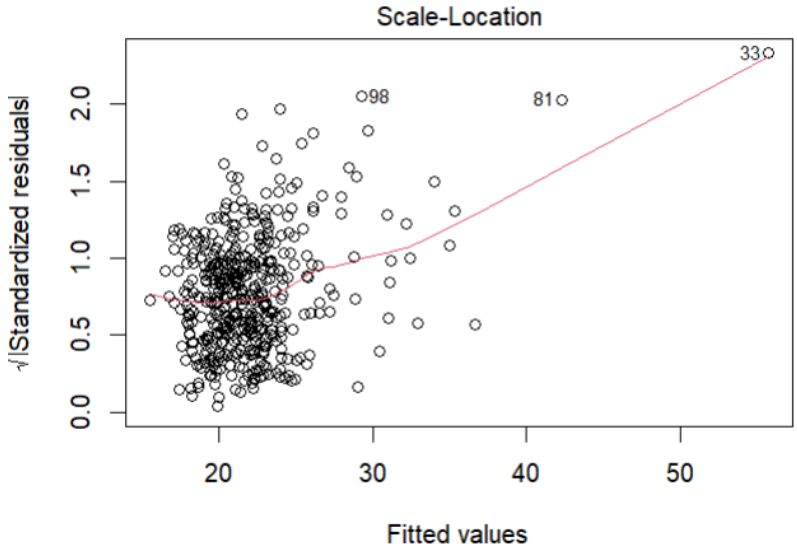
- The residuals appear to fan out as the fitted values increase, which indicates heteroscedasticity (non-constant variance of residuals).

2. **Normal Q-Q Plot**:



- The points mostly follow the diagonal line, suggesting that the residuals are approximately normally distributed. However, there are some deviations at the tails, indicating potential outliers or deviations from normality.

3. **Scale-Location Plot**:



- The red line in the plot is not horizontal and shows an upward trend, further indicating heteroscedasticity.

Breusch-Pagan Test for Testing for Homoscedasticity

The Breusch-Pagan test for heteroscedasticity yielded:

```
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 128.12, df = 6, p-value < 2.2e-16
```

- BP = 128.12
- df = 6
- p-value < 2.2e-16

Since the p-value is significantly less than 0.05, we reject the null hypothesis of constant variance. This confirms the presence of heteroscedasticity.

Transforming the Dependent Variable:

- Log transformation to the dependent variable to stabilize the variance.

```
> summary(log_model)

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.26542 -0.13942  0.00684  0.16331  0.72052

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.374e+00  1.779e-01  13.343  < 2e-16 ***
MPCE_MRP            7.508e-05  1.115e-05   6.732 5.25e-11 ***
MPCE_URP           -1.685e-06  7.716e-06  -0.218  0.82727
Age                 3.814e-04  1.073e-03   0.356  0.72230
Meals_At_Home       7.758e-03  2.352e-03   3.299  0.00105 **
Possess_ration_card -1.089e-01  6.306e-02  -1.727  0.08486 .
Education           1.334e-02  4.262e-03   3.131  0.00186 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2569 on 438 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.2674,    Adjusted R-squared:  0.2574
F-statistic: 26.65 on 6 and 438 DF,  p-value: < 2.2e-16

>
```

- The residuals from the log-transformed model show a more symmetric distribution around zero, indicating a better fit.

Significant Predictors:

- **MPCE_MRP**: Indicates that higher Monthly Per Capita Expenditure based on MRP is associated with higher log-transformed food expenditure.

- **Meals_At_Home**: More meals at home are positively associated with higher log-transformed food expenditure.
- **Education**: Higher education levels are positively associated with higher log-transformed food expenditure.

Non-significant Predictors:

- **MPCE_URP** and **Age**: Do not significantly affect log-transformed food expenditure.
- **Possess_ration_card**: Although not highly significant, it shows a marginally significant negative effect on log-transformed food expenditure.

Model Fit:

- The R-squared value of 0.2674 indicates that approximately 26.74% of the variability in log-transformed food expenditure is explained by the model. This is a moderate level of explanatory power.
- The adjusted R-squared is slightly lower, indicating that the model accounts for the number of predictors.

RECOMMENDATIONS

Based on the findings, the following recommendations are proposed:

Increase Financial Support for Low-Income Families:

- **Rationale**: Higher MPCE is strongly associated with increased food consumption. Financial support, subsidies, or welfare programs can help low-income families afford more and better-quality food, improving their nutritional status.

Enhance Nutritional Education Programs:

- **Rationale**: Education positively influences food consumption. By expanding nutritional education programs, individuals can make more informed food choices, leading to better dietary habits and food security.

Promote Home Gardening and Urban Agriculture:

- **Rationale**: Encouraging home gardening can increase the availability of fresh produce, reduce food costs, and improve food consumption, particularly in urban areas where food access might be limited.

Implement Community Meal Programs:

- **Rationale**: Since meals consumed at home are positively correlated with food consumption, community meal programs can provide nutritious meals to those who might not have adequate food at home, especially targeting vulnerable populations like the elderly and children.

Monitor and Adjust Ration Card Benefits:

- **Rationale**: Reevaluate the benefits provided through ration cards to ensure they meet the nutritional needs of families. Adjusting the types and quantities of food provided can help improve food consumption and nutritional outcomes.