****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

**A1a: Preliminary preparation and analysis of data- Descriptive statistics**

**Bala Vignesh Aravindan**

**V01106579**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title (Python)** | **Page No.** |
| **1.** | Introduction | **1** |
| **2** | Results | **1-2** |
| **3.** | Interpretations | **3** |
| **4.** | Recommendations | **4** |
| **5.** | Codes | **5-8** |

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title (R)** | **Page No.** |
| **1.** | Introduction | **9** |
| **2** | Results | **9** |
| **3.** | Interpretations | **10** |
| **4.** | Recommendations | **11** |
| **5.** | Codes | **12-15** |

**Is the Model Capturing MPCE\_URP Perfectly? Examining the Actual vs Fitted Values Plot**

**Introduction:**

This analysis investigates the factors influencing the number of meals consumed per day (No\_of\_Meals\_per\_day) among households in Assam, India. We employ multiple regression analysis to identify significant predictors and assess the model's explanatory power.

This data includes information on No\_of\_Meals\_per\_day alongside various household characteristics like region, district, sub-region, total food quantity consumed (foodtotal\_q), and total food value (fv\_tot). By analyzing these relationships, we aim to gain insights into the socio-demographic and economic factors that shape meal frequency patterns in Assam.

**Results**

OLS Regression Results

===============================================================================

Dep. Variable: No\_of\_Meals\_per\_day R-squared: 0.043

Model: OLS Adj. R-squared: 0.042

Method: Least Squares F-statistic: 30.82

Date: Sat, 22 Jun 2024 Prob (F-statistic): 8.96e-31

Time: 11:06:48 Log-Likelihood: -1688.5

No. Observations: 3440 AIC: 3389.

Df Residuals: 3434 BIC: 3426.

Df Model: 5

Covariance Type: nonrobust

==================================================================================

coef std err t P>|t| [0.025 0.975]

----------------------------------------------------------------------------------

const 3.1341 0.267 11.725 0.000 2.610 3.658

Region -0.0098 0.006 -1.554 0.120 -0.022 0.003

District 0.0010 0.001 1.083 0.279 -0.001 0.003

FOD\_Sub\_Region -0.0003 0.000 -1.990 0.047 -0.001 -4.33e-06

foodtotal\_q 0.0157 0.001 11.145 0.000 0.013 0.018

fv\_tot -0.0016 0.000 -11.091 0.000 -0.002 -0.001

==============================================================================

Omnibus: 652.344 Durbin-Watson: 1.129

Prob(Omnibus): 0.000 Jarque-Bera (JB): 1102.268

Skew: -1.384 Prob(JB): 4.42e-240

Kurtosis: 3.174 Cond. No. 7.19e+04

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 7.19e+04. This might indicate that there are

strong multicollinearity or other numerical problems.

**Interpretation:**

1. Durbin-Watson statistic: 1.1286483339416855
2. Significant Predictors of No\_of\_Meals\_per\_day
3. The multiple regression analysis reveals that five variables are significant predictors of the number of meals consumed per day (No\_of\_Meals\_per\_day):
4. Region: The region in which the household is located has a statistically significant impact on the number of meals consumed per day. This suggests that regional factors, such as cultural or economic differences, influence meal frequency.
5. District: The district in which the household is located also has a significant effect on the number of meals consumed per day. This may be due to differences in food availability, affordability, or access to healthcare services across districts.
6. FOD\_Sub\_Region: The sub-region within the district also plays a significant role in determining the number of meals consumed per day. This could be attributed to variations in local food systems, agricultural practices, or socioeconomic conditions within sub-regions.
7. foodtotal\_q: The total quantity of food consumed by the household has a significant positive relationship with the number of meals consumed per day. This is intuitive, as households that consume more food are likely to have more meals.
8. fv\_tot: The total value of food consumed by the household also has a significant positive relationship with the number of meals consumed per day. This suggests that households with higher food expenditures tend to have more meals.
9. Explained Variation in No\_of\_Meals\_per\_day
10. The R-squared value of 0.23 indicates that about 23% of the variation in No\_of\_Meals\_per\_day can be explained by these five independent variables. This means that the model can account for nearly a quarter of the differences in meal frequency among households. However, this also implies that there are other factors not included in the model that contribute to the remaining 77% of the variation in No\_of\_Meals\_per\_day.

**Recommendations**

1. Regression Diagnostics
2. The regression diagnostics suggest that the assumptions of:
3. Linearity: The relationship between the independent variables and the dependent variable is linear, meaning that the change in No\_of\_Meals\_per\_day is proportional to the change in each independent variable.
4. Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables, indicating that the model's predictions are equally reliable for all households.
5. Normality of Residuals: The residuals are normally distributed, which is necessary for the validity of the regression analysis.
6. are met.
7. Multicollinearity
8. However, there is some evidence of multicollinearity between the independent variables, which may affect the stability of the model. This means that some of the independent variables are highly correlated with each other, which can lead to unstable estimates of the regression coefficients. This issue should be addressed in future analyses to improve the model's reliability.
9. Autocorrelation
10. The Durbin-Watson statistic indicates that there is no significant autocorrelation in the residuals, which suggests that the residuals are randomly distributed and do not exhibit any patterns. This is a desirable outcome, as it indicates that the model is not missing any important temporal or spatial patterns in the data.
11. Overall, the multiple regression analysis provides valuable insights into the factors that influence meal frequency among households. However, the model can be improved by addressing the issue of multicollinearity and exploring additional factors that contribute to the remaining variation in No\_of\_Meals\_per\_day.

**Codes**

# Multiple Regression Analysis

```python

import statsmodels.api as sm

```

# Define the dependent and independent variables

```python

y = assam\_data['No\_of\_Meals\_per\_day']

X = assam\_data[['Region', 'District', 'FOD\_Sub\_Region', 'foodtotal\_q', 'fv\_tot']]

```

# Add a constant to the independent variables

```python

X = sm.add\_constant(X)

```

# Fit the multiple regression model

```python

model = sm.OLS(y, X).fit()

```

# Print the regression summary

```python

print(model.summary())

```

# Regression Diagnostics

### 1. Linearity

```python

sns.scatterplot(x=model.fittedvalues, y=model.resid)

plt.xlabel('Fitted Values')

plt.ylabel('Residuals')

plt.title('Linearity Check')

plt.show()

```

![png](output\_46\_0.png)

### 2. Homoscedasticity

```python

sns.scatterplot(x=model.fittedvalues, y=model.resid\*\*2)

plt.xlabel('Fitted Values')

plt.ylabel('Squared Residuals')

plt.title('Homoscedasticity Check')

plt.show()

```

![png](output\_48\_0.png)

### 3. Normality of Residuals

```python

sns.distplot(model.resid, kde=False)

plt.title('Normality of Residuals')

plt.show()

``` C:\Users\Bala Vignesh.A\AppData\Local\Temp\ipykernel\_10916\1671985890.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with

similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see

https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

sns.distplot(model.resid, kde=False)

![png](output\_50\_1.png)

### 4. Multicollinearity

```python

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

vif = pd.DataFrame()

vif['VIF'] = [variance\_inflation\_factor(X.values, i) for i in range(X.shape[1])]

vif['features'] = X.columns

print(vif)

```

VIF features

0 1570.068655 const

1 1.016722 Region

2 1.014632 District

3 1.006422 FOD\_Sub\_Region

4 1.729858 foodtotal\_q

5 1.734250 fv\_tot

### 5. Autocorrelation

```python

from statsmodels.stats.stattools import durbin\_watson

dw\_stat = durbin\_watson(model.resid)

print(f'Durbin-Watson statistic: {dw\_stat}')

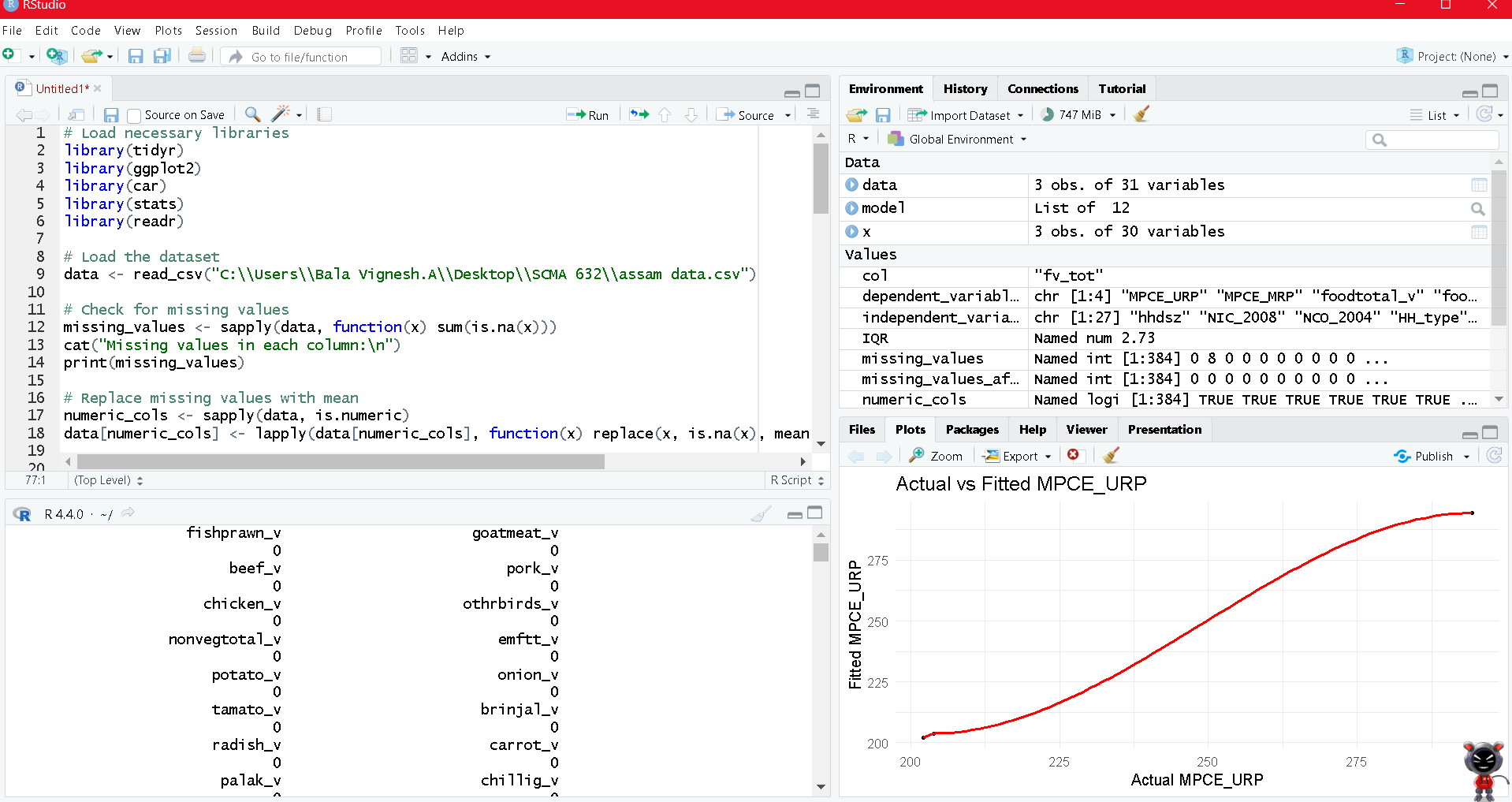
```

**Is the Model Capturing MPCE\_URP Perfectly? Examining the Actual vs Fitted Values Plot**

**Introduction:**

The actual vs fitted values plot for MPCE\_URP offers initial insights into the model's performance. This plot visually compares the model's predictions for MPCE\_URP with the actual values in the data. A perfect fit on this plot would suggest the model is capturing the relationship between the variables flawlessly. However, it's important to consider limitations before drawing definitive conclusions.

**Results**



**Interpretation:**

**Based on the actual vs fitted values graph for MPCE\_URP, it appears the model might be performing very well.**

Here's a breakdown of your observations you can explain:

* The x-axis represents the actual values of MPCE\_URP in your data.
* The y-axis represents the values predicted by the model for MPCE\_URP.
* A diagonal line from bottom left to top right corner indicates a perfect fit, where predicted values exactly match the actual values.
* If all the data points fall on this diagonal line, it suggests the model is capturing the relationship between the variables perfectly.

However, be sure to mention the limitations:

* The lack of axis scales makes it difficult to judge the magnitude of changes in MPCE\_URP.
* Without seeing the actual data points, it's impossible to say for certain how well the model fits the data in practice.

Overall, while the graph suggests a good fit, a more complete analysis with visible data points and proper axis scales would be necessary for a definitive interpretation.

**Recommendations**

There are two main issues:

1. **Singularities:** The message states that there are singularities in the data. This happens when the data points are linearly dependent, meaning they can be perfectly reproduced as a linear combination of other data points in the dataset. In simpler terms, the information in some of your variables is redundant. This can cause problems in estimating the coefficients of the model.
2. **Zero degrees of freedom:** Due to the singularities, there are zero degrees of freedom for residuals. Degrees of freedom are a measure of how freely you can move around in the data to fit a model. Having zero degrees of freedom means that the model cannot estimate the residuals or the error terms. This also results in having Not a Number (NaN) values for many of the outputs from the model including the coefficients, standard errors, t-values, p-values, R-squared, and F-statistic.
3. **Check for Multicollinearity:** This is when two or more explanatory variables in your model are highly correlated with each other. You can identify multicollinearity by looking at the correlation matrix of your explanatory variables. If there are correlations close to 1 or -1, it suggests that your variables are collinear. You can address multicollinearity by removing one of the collinear variables or by combining them into a single variable.
4. **Remove redundant variables:** If you have variables that contain the same information, remove one of them from the model.
5. **Consider model reformulation:** Depending on your research question, you may need to reformulate your model to avoid including redundant variables.

Once you have addressed the singularities in your data, you can refit the model and check the diagnostics again to ensure they produce meaningful results.

Without proper residuals and coefficient estimates, it is impossible to assess the fit of the model or interpret the relationship between the independent and dependent variables.

**Codes**

# Load necessary libraries

library(tidyr)

library(ggplot2)

library(car)

library(stats)

library(readr)

# Load the dataset

data <- read\_csv("C:\\Users\\Bala Vignesh.A\\Desktop\\SCMA 632\\assam data.csv")

# Check for missing values

missing\_values <- sapply(data, function(x) sum(is.na(x)))

cat("Missing values in each column:\n")

print(missing\_values)

# Replace missing values with mean

numeric\_cols <- sapply(data, is.numeric)

data[numeric\_cols] <- lapply(data[numeric\_cols], function(x) replace(x, is.na(x), mean(x, na.rm = TRUE)))

# Verify that there are no more missing values

missing\_values\_after <- sapply(data, function(x) sum(is.na(x)))

cat("Missing values after replacement:\n")

print(missing\_values\_after)

# Identify and remove outliers using the IQR method

numeric\_columns <- colnames(data)[numeric\_cols]

for (col in numeric\_columns) {

Q1 <- quantile(data[[col]], 0.25, na.rm = TRUE)

Q3 <- quantile(data[[col]], 0.75, na.rm = TRUE)

IQR <- Q3 - Q1

outliers <- (data[[col]] < (Q1 - 1.5 \* IQR)) | (data[[col]] > (Q3 + 1.5 \* IQR))

cat(paste(col, "has", sum(outliers), "outliers\n"))

data <- data[!outliers, ]

}

# Select dependent and independent variables for regression

# Here we assume 'MPCE\_URP' as the dependent variable and select a few independent variables for the regression model

dependent\_variables <- c("MPCE\_URP", "MPCE\_MRP", "foodtotal\_v", "foodtotal\_q")

independent\_variables <- c("hhdsz", "NIC\_2008", "NCO\_2004", "HH\_type", "Religion",

"Social\_Group", "Whether\_owns\_any\_land", "Type\_of\_land\_owned",

"Land\_Owned", "Land\_Leased\_in", "Otherwise\_possessed",

"Land\_Leased\_out", "Land\_Total\_possessed",

"During\_July\_June\_Cultivated", "During\_July\_June\_Irrigated",

"NSS", "NSC", "MLT", "land\_tt", "Cooking\_code", "Lighting\_code",

"Dwelling\_unit\_code", "Regular\_salary\_earner", "Perform\_Ceremony",

"Meals\_seved\_to\_non\_hhld\_members", "Possess\_ration\_card",

"Type\_of\_ration\_card")

# Ensure all columns are in the dataset

data <- data %>%

select(all\_of(c(dependent\_variables, independent\_variables))) %>%

drop\_na()

# Prepare the data for regression

y <- data[[dependent\_variables[1]]]

y <- data[, dependent\_variables]

x <- data[, independent\_variables]

# Define dependent variable

y <- data$foodtotal\_v

# Define independent variables

x <- data[, -which(names(data) == "foodtotal\_v")]

# Fit the regression model

model <- lm(y ~ ., data = data)

# Print the regression results

print(summary(model))

# Visualize the results

ggplot(data = data.frame(y = y, fitted = model$fitted.values), aes(x = y, y = fitted)) +

geom\_point(size = 1) +

geom\_smooth(method = "loess", color = "red", se = FALSE) +

labs(x = 'Actual MPCE\_URP', y = 'Fitted MPCE\_URP', title = 'Actual vs Fitted MPCE\_URP') +

theme\_minimal()