****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

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

**DEEPTHI ANNA ALEX**

**V01101949**

**Date of Submission: 24-06-2024**

**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** |

**I**

**Introduction:**

The present study aims to examine the variables that impact the quantity of meals eaten daily (No\_of\_Meals\_per\_day) in households located in Arunachal Pradesh, India. We evaluate the explanatory power of the model and find important factors using multiple regression analysis.   
Together with other household characteristics including region, district, sub-region, total food quantity consumed (foodtotal\_q), and total food value (food\_total), this data also includes information on the number of meals eaten each day. We hope to learn more about the sociodemographic and economic factors influencing meal frequency trends in Arunachal Pradesh by examining these linkages.

**Results:**

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

Residual standard error: 6.866 on 1669 degrees of freedom

(4 observations deleted due to missingness)

Multiple R-squared: 0.4478, Adjusted R-squared: 0.4458

F-statistic: 225.6 on 6 and 1669 DF, p-value: < 2.2e-16

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

Coefficients:

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

(Intercept) 19.4440516 1.3614264 14.282 < 2e-16 \*\*\*

MPCE\_MRP 0.0033176 0.0002401 13.816 < 2e-16 \*\*\*

MPCE\_URP 0.0006212 0.0002215 2.805 0.00509 \*\*

Age -0.0264196 0.0174447 -1.514 0.13009

Meals\_At\_Home -0.0231747 0.0106126 -2.184 0.02912 \*

Possess\_ration\_card -0.1149288 0.4336014 -0.265 0.79100

Education -0.2214534 0.0505248 -4.383 1.24e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

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. Significant Predictors of No\_of\_Meals\_per\_day
2. The multiple regression analysis reveals that five variables are significant predictors of the number of meals consumed per day (No\_of\_Meals\_per\_day):
3. 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.
4. 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.
5. 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.
6. The R-squared value of 0.4478 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 56% of the 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}')

```

**Introduction:**

The MPCE\_URP plot of real versus fitted values provides some preliminary information about the model's performance. The model's predictions for MPCE\_URP are shown in this plot together with the actual values found in the data. A perfect fit on this plot would indicate that the model is doing an excellent job of capturing the relationship between the variables. It's crucial to take constraints into account before making firm judgments, though.

**Results**

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

-32.800 -4.516 -0.821 3.526 32.204

Coefficients:

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

(Intercept) 19.4440516 1.3614264 14.282 < 2e-16 \*\*\*

MPCE\_MRP 0.0033176 0.0002401 13.816 < 2e-16 \*\*\*

MPCE\_URP 0.0006212 0.0002215 2.805 0.00509 \*\*

Age -0.0264196 0.0174447 -1.514 0.13009

Meals\_At\_Home -0.0231747 0.0106126 -2.184 0.02912 \*

Possess\_ration\_card -0.1149288 0.4336014 -0.265 0.79100

Education -0.2214534 0.0505248 -4.383 1.24e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.866 on 1669 degrees of freedom

(4 observations deleted due to missingness)

Multiple R-squared: 0.4478, Adjusted R-squared: 0.4458

F-statistic: 225.6 on 6 and 1669 DF, p-value: < 2.2e-16

Y ~ a1 + b1\*x1+ b2\*x2 + b3\*x3 + b4\*x4 + b5\*x5 + b6\*x6 + ui (error)

foodtotal\_q ~ 19.4440516+ 0.0033176\* MPCE\_MRP+ 0.0006212\*MPCE\_URP -0.0264196\* Age -0.0231747\* Meals\_At\_Home -0.1149288\*Possess\_ration\_card -0.2214534\* Education

VIF more than 8 its problematic

there is no multicollinerity as all the independent variables values are less than 8

regression equation "y = 19.44 + 0.003318\*x1 + 0.000621\*x2 + -0.02642\*x3 + -0.023175\*x4 + -0.114929\*x5 + -0.221453\*x6"

9.117317

MPCE\_MRP \*\*\* significant by 1 percent level

**Interpretation:**

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

LINEAR MODEL(LM)

Food total is a dependent variable

Other variables MPCE\_URP,MPCE\_MRP,Age,Meals\_At\_Home , Possess\_ration\_card,Education these are independent variables.

**R SQUARED 0.4478 44 PERCENT RELATION INDEPENT VARIABLE ARE ONLY TO EXPLAIN 44 PERCENT RELATION IN THE DEPENDENT VARIABLE**

**Star means its shows the significance of independent variables**

significant variable

if probability of t is less than 0.05 its significant at 5 percent level

less than 0.01 its 1 percent level

less than 0.1 its 10 percent level

MPCE\_MRP \*\*\* significant by 1 percent level

Education \*\*\* significant by 1 percent level

VIF more than 8 its problematic

there is no multicollinerity as all the independent variables values are less than 8

**Codes**

# Load necessary libraries

library(dplyr)

# Load the dataset

data <- read\_csv("C:\\Users\\HP\\OneDrive\\Desktop\\deepthi asignments scma\\NSSO68 (2).csv")

#NSSO

library(dplyr)

setwd('E:\\Assignments\_SCMA632\\Data')

getwd()

# Load the dataset

data <- read.csv("NSSO68.csv")

unique(data$state\_1)

# Subset data to state assigned

subset\_data <- data %>%

filter(state\_1 == 'KA') %>%

select(foodtotal\_q, MPCE\_MRP, MPCE\_URP,Age,Meals\_At\_Home,Possess\_ration\_card,Education, No\_of\_Meals\_per\_day)

print(subset\_data)

sum(is.na(subset\_data$MPCE\_MRP))

sum(is.na(subset\_data$MPCE\_URP))

sum(is.na(subset\_data$Age))

sum(is.na(subset\_data$Possess\_ration\_card))

sum(is.na(data$Education))

impute\_with\_mean <- function(data, columns) {

data %>%

mutate(across(all\_of(columns), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

}

# 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))

# Fit the regression model

model <- lm(foodtotal\_q~ MPCE\_MRP+MPCE\_URP+Age+Meals\_At\_Home+Possess\_ration\_card+Education, data = subset\_data)

# Print the regression results

print(summary(model))

library(car)

# Check for multicollinearity using Variance Inflation Factor (VIF)

vif(model) # VIF Value more than 8 its problematic

# Extract the coefficients from the model

coefficients <- coef(model)

# Construct the equation

equation <- paste0("y = ", round(coefficients[1], 2))

for (i in 2:length(coefficients)) {

equation <- paste0(equation, " + ", round(coefficients[i], 6), "\*x", i-1)

}

# Print the equation

print(equation)

head(subset\_data$MPCE\_MRP,1)

head(subset\_data$MPCE\_URP,1)

head(subset\_data$Age,1)

head(subset\_data$Meals\_At\_Home,1)

head(subset\_data$Possess\_ration\_card,1)

head(subset\_data$Education,1)

head(subset\_data$foodtotal\_q,1)