VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

MICAH ASHADEEP EMMANUEL

V01101166

Date of Submission: 23-06-2024

CONTENTS

INTRODUCTION

This project leverages multi-linear regression analysis. We'll construct a model to predict total food consumption based on factors like how often meals are eaten at home, possession of a government ration card, age, and economic indicators potentially reflecting food prices. To ensure the model's credibility, we'll meticulously evaluate its assumptions through diagnostics. This iterative process will culminate in a more robust understanding of the significant differences observed, ultimately unveiling the key socio-economic factors shaping dietary patterns across the population of Odisha.

OBJECTIVES

Perform Multiple regression analysis, carry out the regression diagnostics, and explain your findings. Correct them and revisit your results and explain the significant differences you observe. [data “NSSO68.csv”]Extract data specific to R.

. Finding the missing values and assigning it with mean

. Conducting Multiple Regression

.

RESULTS AND INTERPRETATION

USING R

```
install.packages("dplyr")
library(dplyr)
install.packages("tidyverse")
library(tidyverse)
install.packages("car")
library(car)
install.packages("lmtest")
library(lmtest)
setwd("C:\\Users\\HP\\Documents\\ns")
getwd()
data = read.csv("NSSO68 new.csv")
str(data)
wb_data <- data %>%
  filter(state_1 == "WB")
relevant_columns <- c("foodtotal_q", "Meals_At_Home", "Possess_ration_card",
"Age", "MPCE_URP", "MPCE_MRP")
westbengal _data <- westbengal _data %>%
  select(all_of(relevant_columns)) %>%
  print(westbengal _data)
str(westbengal _data)
sum(is.na(westbengal _data$Meals_At_Home))
sum(is.na(westbengal _data$Possess_ration_card))
sum(is.na(westbengal _data$Age))
sum(is.na(westbengal _data$MPCE_URP))
sum(is.na(westbengal _data$MPCE_MRP))
complete_rows <- complete.cases(westbengal_data$foodtotal_q,
westbengal_data$Possess_ration_card)
```

```
filtered_data <- westbengal_data[complete_rows, ]
model <- lm(foodtotal_q ~ Possess_ration_card, data = filtered_data)


imput_with_mean <- function(data,column) {
  data %>%
    mutate(across(all_of(columns), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))
}
sum(is.na(data$foodtotal_q))
cleaned_data <- na.omit(data)
odisha_data <- cleaned_data %>%
  filter(state_1 == "WB")
nrow(westbengal _data)
model <- lm(foodtotal_q ~ Meals_At_Home + Possess_ration_card + Age +
MPCE_URP + MPCE_MRP, data = data)
summary(model)
```

RESULT

```
Residual standard error: 8.747 on 100424 degrees of freedom
  (1232 observations deleted due to missingness)
Multiple R-squared:  0.1853,                    Adjusted R-squared:  0.1852
F-statistic:  4567 on 5 and 100424 DF,  p-value: < 2.2e-16
```

CODE
USING PYTHON

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
```

```
data = pd.read_csv("/content/NSSO68 new.csv")
```

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8538 entries, 0 to 8537
Columns: 384 entries, slno to fv_tot
dtypes: float64(303), int64(80), object(1)
memory usage: 25.0+ MB
None
```

```
westbengal _data = data[data['state_1'] == "WB"].copy()
```

```
relevant_columns = ["foodtotal_q", "Meals_At_Home",
"Possess_ration_card", "Age", "MPCE_URP", "MPCE_MRP"]
westbengal_data = westbengal_data[relevant_columns].copy()
```

```
print(westbengal_data.head())
```

```
foodtotal_q  Meals_At_Home  Possess_ration_card  Age  MPCE_URP
MPCE_MRP
741     33.110413           60.0                   2.0    31    3455.50
3844.95
742     31.683645           60.0                   1.0    42    2572.67
2377.28
743     25.575244           60.0                   1.0    53    1792.75
2039.86
744     24.920166           60.0                   1.0    60     880.00
970.04
745     24.742780           90.0                   1.0    35     854.50
935.56
```

```
print(westbengal_data.isnull().sum())
```

```
foodtotal_q            0
Meals_At_Home          40
Possess_ration_card    0
Age                    0
MPCE_URP               0
MPCE_MRP               0
dtype: int64
```

```
cleaned_data = westbengal_data.dropna()
```

```
print(cleaned_data.shape[0])
```

```
1013
```

```
X = cleaned_data[["Meals_At_Home", "Possess_ration_card", "Age",
"MPCE_URP", "MPCE_MRP"]]
y = cleaned_data["foodtotal_q"]
```

```
X = sm.add_constant(X)
```

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

```
print(model.summary())
```

```
                     OLS Regression Results
========================================================================
=======
Dep. Variable:            foodtotal_q   R-squared:
0.334
Model:                            OLS   Adj. R-squared:
0.331
Method:                 Least Squares   F-statistic:
101.0
Date:                Sun, 23 Jun 2024   Prob (F-statistic):
2.10e-86
Time:                        10:07:29   Log-Likelihood:
-3404.2
No. Observations:                1013   AIC:
6820.
Df Residuals:                    1007   BIC:
6850.
Df Model:                           5
Covariance Type:            nonrobust
========================================================================
================
                          coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------
----------------
const                   4.6109      1.497      3.081      0.002
1.674       7.548
Meals_At_Home           0.1545      0.012     12.774      0.000
0.131       0.178
Possess_ration_card    -0.0047      0.473     -0.010      0.992      -
0.933       0.924
Age                     0.0752      0.018      4.274      0.000
0.041       0.110
```

```
MPCE_URP                0.0011      0.000      5.160      0.000
0.001       0.001
MPCE_MRP                0.0017      0.000      8.449      0.000
0.001       0.002
========================================================================
=======
Omnibus:                      211.654   Durbin-Watson:
1.512
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
1935.790
Skew:                           0.679   Prob(JB):
0.00
Kurtosis:                       9.635   Cond. No.
2.51e+04
========================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 2.51e+04. This might indicate that
there are

strong multicollinearity or other numerical problems.
```

FINDING

The code analyzes data on food expenditure in Odisha, India. It finds a moderate correlation between food expenditure and factors like frequency of meals at home, age, and monthly expenditure. Interestingly, having a ration card doesn't seem to significantly affect food expenditure in this dataset.