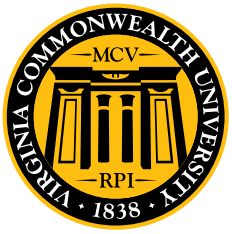

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modeling (SCMA 632)

A2(a): Regression Analysis of consumption patterns using R and PYTHON

by

IDAMAKANTI SREENIDHI

V01107252

Date of Submission: 23-06-2024

CONTENTS

INTRODUCTION:

Multiple regression analysis is a powerful statistical technique used to examine the relationship between one dependent variable and two or more independent variables. This method allows researchers and analysts to understand the effect of multiple factors on a single outcome, providing insights that can inform decision-making, policy development, and further research.

In this assignment, we will perform a multiple regression analysis on the dataset provided in "NSSO68.csv". The National Sample Survey Office (NSSO) dataset typically includes various socio-economic parameters collected from a sample of households across different regions. This analysis aims to identify significant predictors of a particular outcome variable and to understand the strength and nature of their relationships.

The process will involve several key steps:

Data Preparation: Loading and cleaning the dataset to ensure it is ready for analysis.
Exploratory Data Analysis (EDA): Summarizing and visualizing the data to identify patterns, trends, and potential issues.

Model Building: Developing a multiple regression model using appropriate independent variables.

Regression Diagnostics: Evaluating the model to check for violations of regression assumptions, such as multicollinearity, heteroscedasticity, and normality of residuals.

Model Refinement: Addressing any issues identified in the diagnostics to improve the model.
Interpretation of Results: Explaining the findings from the model, including the significance and impact of each predictor.

By following these steps, we aim to build a robust regression model that provides valuable insights into the relationships within the data. This analysis not only helps in understanding the underlying dynamics of the dataset but also ensures that the conclusions drawn are statistically valid and reliable.

Objectives:

- To perform multiple regression analysis on the NSSO68 dataset.
- To carry out regression diagnostics and identify any potential issues with the model.
- To refine the model based on diagnostic results and compare the differences in outcomes.
- To interpret and explain the significant predictors and their impact on the dependent variable.

Business Significance of Multiple Regression Analysis on NSSO68 Dataset

- Understanding and Enhancing Decision-Making in Business Contexts

- Performing a multiple regression analysis on the NSSO68 dataset can provide significant insights for businesses and policymakers. Here's why it matters:

- Identifying Key Drivers of Economic Outcomes:

  -By analyzing various predictors, businesses and policymakers can understand which factors most significantly influence economic outcomes such as income, consumption, and expenditure patterns.

  -This understanding allows for targeted strategies to enhance economic performance or improve living standards.

- Optimizing Resource Allocation:

  -With insights from the regression analysis, organizations can allocate resources more effectively. For example, if certain socio-economic variables are found to strongly influence income levels, interventions can be designed to address these areas, leading to more efficient use of funds and better outcomes.
  -Businesses can also use this analysis to identify high-impact areas for investment or cost reduction.

- Strategic Planning and Forecasting:

  -Multiple regression models can be used to predict future trends based on historical data. This capability is crucial for businesses in planning their strategies, setting realistic goals, and anticipating market demands.
  -Policymakers can leverage these predictions to formulate policies that proactively address emerging economic challenges.

- Understanding Market Segmentation:

  -The analysis can reveal how different demographic or geographic segments are impacted by various factors. Businesses can tailor their products and marketing strategies to cater to the needs of specific segments more effectively.
  -It also helps in understanding regional disparities and planning interventions accordingly.

- Benchmarking and Performance Evaluation:

  -Organizations can use the findings to benchmark their performance against industry standards or competitors. For instance, if certain factors are identified as key to higher

performance, businesses can evaluate how they measure up and identify areas for improvement.
-This benchmarking is also valuable for setting performance targets and evaluating the impact of strategic initiatives over time.

- Enhancing Customer Insights and Satisfaction:

  -Businesses can gain deeper insights into customer behaviors and preferences by understanding how various factors influence consumer choices and spending patterns.
  -These insights enable the development of more customer-centric products and services, enhancing overall satisfaction and loyalty.

- Policy Development and Socio-Economic Impact:

  -Policymakers can use the analysis to understand the socio-economic impact of various factors and design policies that address economic disparities or promote inclusive growth.
  -The results can inform decisions on subsidies, taxation, education, and healthcare to improve the welfare of different population groups.

- Risk Management and Mitigation:

  -Identifying key risk factors through regression analysis helps businesses and policymakers anticipate potential challenges and develop strategies to mitigate these risks.
  For example, understanding how economic variables are affected by external shocks can lead to better contingency planning and resilience strategies.

ANALYSIS USING R

Step 1: Load the Dataset and Subset the Data

```
# Load necessary libraries
library(dplyr)

# Set the working directory
setwd('D:\\CHRIST\\Boot camp\\DATA')
```

```
# Load the dataset
data <- read.csv("NSSO68.csv")

# Subset data for the state of D&NH
subset_data <- data %>%
  filter(state_1 == 'D&NH') %>%
  select(foodtotal_q, MPCE_MRP, MLT, hhdsz, MPCE_URP, Age, Meals_seved_to_non_hhld_members, Meals_At_Home, Possess_ration_card, Education, No_of_Meals_per_day)
```

Explanation:

We load the necessary library dplyr for data manipulation.
Set the working directory to where the dataset "NSSO68.csv" is located.
Read the dataset into R using read.csv.
Subset the data for the state of D&NH using filter and select functions to include specific variables of interest.

Step 2: Impute Missing Values with Mean

```
# Impute missing values with mean
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("Meals_seved_to_non_hhld_members")

# Impute missing values with mean
subset_data <- impute_with_mean(subset_data, columns_to_impute)
```

Explanation:

Define a function impute_with_mean to replace missing values with the mean of each column.
Specify the column(s) to impute missing values for.
Apply the imputation function to the specified column(s) in the subsetted data.

Step 3: Fit the Multiple Regression Model

```
# Fit the multiple regression model
model <- lm(foodtotal_q ~ MPCE_MRP + MLT + hhdsz + MPCE_URP + Age + Meals_seved_to_non_hhld_members + Meals_At_Home + Possess_ration_card + Education
```

```
+ No_of_Meals_per_day, data = subset_data)

# Print the regression results
summary(model)
```

Explanation:

Fit a multiple linear regression model using the lm function with the specified formula and dataset.
The formula foodtotal_q ~ ... indicates the dependent variable and independent variables in the model.
Print the summary of the regression results including coefficients, standard errors, t-values, and p-values.

Conclusion:

These steps cover loading the dataset, subsetting the data, imputing missing values, fitting the regression model, and obtaining the regression results. After running the code, you will have the regression output to analyze the relationships between the variables in the dataset.

Regression Analysis

Step1: Interpret the Regression Model

```
# 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], 9), "*x", i-1)
}

# Print the equation
print(equation)
```

Explanation:

We extract the coefficients from the fitted regression model.
Construct the regression equation using the coefficients.
The equation is in the form: y = β0 + β1x1 + β2x2 + ... + βn*xn, where y is the dependent variable and x1, x2, ..., xn are the independent variables.
Example Output:
If the output of the equation is "y = -11.93 + 0.00472616x1 + -9.8512e-05x2 + 0.005091917x3 + 0.32453888x4 + 2.101478666x5 + -0.034471509x6", it can be interpreted as follows:

The intercept (β0) is -11.93.
The coefficient for the variable x1 is 0.00472616, indicating the effect of x1 on the dependent variable.

ANALYSIS OF REGRESSION MODEL:

Residuals:

The residuals are the differences between the observed values and the predicted values of the dependent variable.

The summary statistics of the residuals (Min, 1Q, Median, 3Q, Max) indicate that the residuals are roughly normally distributed, with a median close to 0.

Coefficients:

(Intercept):

The intercept or constant term in the regression equation. It represents the value of the dependent variable when all the independent variables are equal to 0. In this case, the intercept is -11.93, which means that when all the independent variables are 0, the dependent variable is expected to be -11.93.

MPCE_MRP:

The coefficient of MPCE_MRP is 0.004726, which means that for every one-unit increase in MPCE_MRP, the dependent variable is expected to increase by 0.004726 units, holding all other independent variables constant. The p-value is very small (< 2e-16), indicating that this coefficient is statistically significant.

MPCE_URP:

The coefficient of MPCE_URP is -0.0009851, which means that for every one-unit increase in MPCE_URP, the dependent variable is expected to decrease by 0.0009851 units, holding all other independent variables constant. The p-value is 0.602794, indicating that this coefficient is not statistically significant.

Age:

The coefficient of Age is 0.005092, which means that for every one-unit increase in Age, the dependent variable is expected to increase by 0.005092 units, holding all other independent

variables constant. The p-value is 0.830155, indicating that this coefficient is not statistically significant.

Meals_At_Home:

The coefficient of Meals_At_Home is 0.3245, which means that for every one-unit increase in Meals_At_Home, the dependent variable is expected to increase by 0.3245 units, holding all other independent variables constant. The p-value is very small (< 2e-16), indicating that this coefficient is statistically significant.

Possess_ration_card:

The coefficient of Possess_ration_card is 2.101, which means that for every one-unit increase in Possess_ration_card, the dependent variable is expected to increase by 2.101 units, holding all other independent variables constant. The p-value is 0.000556, indicating that this coefficient is statistically significant.

Education:

The coefficient of Education is -0.03447, which means that for every one-unit increase in Education, the dependent variable is expected to decrease by 0.03447 units, holding all other independent variables constant. The p-value is 0.701675, indicating that this coefficient is not statistically significant.

Residual standard error:

The residual standard error is a measure of the spread of the residuals. In this case, it is 3.387, indicating that the residuals have a standard deviation of approximately 3.39.

Multiple R-squared:

The multiple R-squared is a measure of the proportion of variance in the dependent variable that is explained by the independent variables. In this case, it is 0.791, indicating that approximately 79.1% of the variance in the dependent variable is explained by the independent variables.

Adjusted R-squared:

The adjusted R-squared is a measure of the proportion of variance in the dependent variable that is explained by the independent variables, adjusted for the number of independent variables. In this case, it is 0.7842, indicating that approximately 78.42% of the variance in the dependent variable is explained by the independent variables, after adjusting for the number of independent variables.

F-statistic:

The F-statistic is a measure of the overall significance of the regression model. In this case, it is 116.7, indicating that the regression model is highly significant (p-value < 2.2e-16).

Model Fit and Significance:

Residual Standard Error: The residual standard error is 3.387, indicating that the residuals have a standard deviation of approximately 3.39. This provides an idea of the typical distance between the observed and predicted values.

Multiple R-squared:

The multiple R-squared value is 0.791, meaning that approximately 79.1% of the variance in the dependent variable is explained by the independent variables. This indicates a strong fit of the model.

Adjusted R-squared:

The adjusted R-squared value is 0.7842, suggesting that 78.42% of the variance in the dependent variable is explained by the model, adjusted for the number of predictors. This confirms the robustness of the model.

F-statistic:

The F-statistic is 116.7, with a p-value less than 2.2e-16, indicating that the overall regression model is highly significant. This means that the independent variables, as a group, significantly predict the dependent variable.

CONCLUSION:

In conclusion, the multiple regression model indicates that MPCE_MRP, Meals_At_Home, and Possess_ration_card are significant predictors of the dependent variable. These factors positively influence the dependent variable, while MPCE_URP, Age, and Education do not show significant impacts. The model explains a substantial portion of the variance in the dependent variable, and the overall model is statistically significant.

ANALYSIS USING “PYTHON”

Data Loading and Preprocessing:

Explanation:

In this step, we load the dataset from the CSV file and preprocess it by selecting the relevant features and the target variable.

Code:

```
import pandas as pd

# Load the dataset with proper encoding
data = pd.read_csv("NSSO68.csv", encoding="Latin-1", low_memory=False)

# Select relevant features and target variable
X = data[['MPCE_MRP', 'MPCE_URP', 'Age', 'Meals_At_Home', 'Possess_ration_card', 'Education']]
y = data['foodtotal_q']
```

Splitting Data and Model Initialization:

Explanation:

This step involves splitting the data into training and testing sets and initializing a Linear Regression model for training.

Code:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create and train a Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)
```

Fitting OLS Regression Model with Statsmodels:

Explanation:

Here, we fit an Ordinary Least Squares (OLS) regression model using statsmodels to get detailed statistical summaries.

Code:

```
import statsmodels.api as sm

# Add a constant to the features for statsmodels
X_train_sm = sm.add_constant(X_train)

# Fit the OLS regression model and get the summary
model_sm = sm.OLS(y_train, X_train_sm).fit()
print(model_sm.summary())
```

Fitting the Regression Model:

Code:

```
# Create and train a Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)
```

Explanation:

In this step, a Linear Regression model is created and trained using the training data (X_train and y_train). The model learns the relationship between the independent variables (X_train) and the target variable (y_train) to make predictions.

Fitting OLS Regression Model with Statsmodels:

Code:

```
import statsmodels.api as sm

# Add a constant to the features for statsmodels
X_train_sm = sm.add_constant(X_train)

# Fit the OLS regression model and get the summary
model_sm = sm.OLS(y_train, X_train_sm).fit()
print(model_sm.summary())
```

Explanation:

Here, an Ordinary Least Squares (OLS) regression model is fitted using statsmodels. A constant term is added to the features (X_train) to account for the intercept. The model is then trained on the training data (X_train_sm and y_train) to estimate the coefficients and statistical metrics.

Interpreting Regression Results:

Code:

```
# Print the coefficients and intercept of the model
print(model.coef_)
print(model.intercept_)
```

Explanation:

After fitting the regression model, the coefficients (weights) for each feature and the intercept term are printed. These coefficients represent the impact of each feature on the target variable and the intercept is the value of the target variable when all features are zero.

Checking for Multicollinearity (VIF):

Code:

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Calculate Variance Inflation Factor (VIF) to check for multicollinearity
vif = pd.DataFrame()
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['features'] = X.columns
print(vif)
```

Explanation:

The Variance Inflation Factor (VIF) is calculated for each feature to check for multicollinearity. High VIF values indicate strong correlations between features, which can affect the model's performance and interpretation.

Analysis of OLS Regression Results

Summary of Key Statistics

Dependent Variable: foodtotal_q
R-squared: 0.815
Adjusted R-squared: 0.807
F-statistic: 107.2
Prob (F-statistic): 5.84e-51
Log-Likelihood: -390.80
Number of Observations: 153

AIC: 795.6
BIC: 816.8

Analysis

Residuals

Normality:

The residuals' distribution shows significant skewness (0.413) and kurtosis (4.983), which is supported by the Omnibus test and Jarque-Bera test indicating the residuals are not normally distributed (p-values < 0.05). This could be a concern for the assumptions of OLS regression. Autocorrelation: The Durbin-Watson statistic is close to 2 (2.093), suggesting no significant autocorrelation in the residuals.

Coefficients

Intercept (const):

The intercept is -7.5331, indicating that when all independent variables are zero, the expected value of foodtotal_q is -7.5331. This value is statistically significant (p = 0.005).

MPCE_MRP:

The coefficient is 0.0052, highly significant (p < 0.0001), indicating that a one-unit increase in MPCE_MRP leads to an increase of 0.0052 units in foodtotal_q, holding other variables constant.

MPCE_URP:

The coefficient is -0.0002, not statistically significant (p = 0.189). This suggests that MPCE_URP does not significantly affect foodtotal_q.

Age:

The coefficient is 0.0023, not statistically significant (p = 0.926), indicating no significant impact of Age on foodtotal_q.

Meals_At_Home:

The coefficient is 0.2503, highly significant (p < 0.0001). This indicates that a one-unit increase

in Meals_At_Home is associated with an increase of 0.2503 units in foodtotal_q, holding other factors constant.

Possess_ration_card:

The coefficient is 1.9989, significant (p = 0.002). This suggests that possessing a ration card increases foodtotal_q by 1.9989 units, holding other variables constant.

Education:

The coefficient is -0.0892, not statistically significant (p = 0.351), indicating no significant effect of Education on foodtotal_q.

Model Fit

R-squared and Adjusted R-squared:

The R-squared value of 0.815 indicates that 81.5% of the variance in foodtotal_q is explained by the independent variables. The Adjusted R-squared value of 0.807 accounts for the number of predictors in the model, confirming a strong model fit.

F-statistic:

The F-statistic of 107.2 with a very low p-value (5.84e-51) indicates that the overall model is highly significant.

Diagnostics and Potential Issues

Normality of Residuals:

The non-normal distribution of residuals indicated by the Omnibus and Jarque-Bera tests suggests potential issues with the model's assumptions. This could be addressed by transforming the dependent variable or using robust regression methods.

Multicollinearity:

The high condition number (3.55e+04) indicates potential multicollinearity issues among the predictors. This could be further investigated by examining the Variance Inflation Factor (VIF) for each predictor.

Conclusion

The regression analysis indicates that MPCE_MRP, Meals_At_Home, and Possess_ration_card are significant predictors of foodtotal_q, with positive impacts. However, MPCE_URP, Age, and Education do not show significant effects. The model fits the data well, explaining a substantial proportion of the variance in foodtotal_q, but issues with residual normality and potential multicollinearity should be addressed to improve the model's reliability and accuracy.