VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A2a: Comprehensive Regression Analysis of Food Consumption Data

GAURI VINOD NAIR

V01110160

Date of Submission: 23-06-2024

TABLE OF CONTENTS

INTRODUCTION

The focus of this study is on performing multiple regression analysis using data from the National Sample Survey Office (NSSO) 68th round. The dataset, “NSSO68.csv,” encompasses various socio-economic variables. The primary objective is to model the relationship between a dependent variable and multiple independent variables, while also conducting a thorough regression diagnostics to ensure the robustness of the model.

To achieve this, we meticulously manipulate and clean the dataset to ensure it is suitable for analysis. This includes handling missing values, identifying and treating outliers, and transforming variables as needed. The cleaned dataset is then imported into Python, a powerful and versatile statistical programming language, well-suited for handling and analyzing large datasets.

The insights derived from this study will provide valuable information on the relationships between various socio-economic factors captured in the NSSO dataset, aiding policymakers and researchers in making informed decisions.

OBJECTIVE

1. Perform Multiple Regression Analysis
2. Carry Out Regression Diagnostics
    - Multicollinearity
    - Heteroscedasticity
    - Normality of Residuals
    - Influential Points
3. Explain Findings
4. Correct Issues and Revisit Results
5. Explain Significant Differences

BUSINESS SIGNIFICANCE

The comprehensive multiple regression analysis of the "NSSO68.csv" dataset holds significant business implications. By identifying key predictors through this analysis, businesses can make more informed decisions. Understanding the factors that most influence critical business outcomes enables more effective resource allocation and strategic prioritization. This insight aids in policy formulation and adjustment, particularly for sectors influenced by socio-economic factors like retail, real estate, or financial services.

Accurate forecasting and planning are other vital benefits. A robust regression model allows businesses to predict future trends based on historical data, facilitating better planning for inventory, staffing, and marketing efforts. Scenario analysis further helps in risk management by assessing potential outcomes under different conditions, thereby supporting contingency planning.

Operational efficiency can be greatly enhanced through insights gained from regression analysis. Businesses can optimize resources, reduce waste, and improve productivity by focusing on significant cost drivers. Additionally, understanding consumer behavior factors enables effective market segmentation and targeted marketing strategies, which can improve customer satisfaction and loyalty, leading to increased returns on marketing investments.

Regression diagnostics play a crucial role in risk management by identifying outliers and influential points that may pose vulnerabilities. Addressing these can make operations more resilient. In industries like manufacturing, predictive maintenance schedules derived from regression analysis can minimize downtime and costs.

Strategic insights from regression analysis can also build investor confidence and provide a competitive advantage. Transparent, data-driven decisions are well-received by stakeholders, enhancing the overall credibility of the business. Furthermore, businesses can leverage these insights for regulatory compliance and policy advocacy, ensuring they meet standards and influence favorable regulatory changes.

In summary, performing multiple regression analysis with thorough diagnostics supports sustainable business growth by enhancing decision-making, forecasting, operational efficiency, market targeting, risk management, stakeholder engagement, and regulatory compliance.

REAULTS AND INTERPRETAIONS

1. Perform multiple regression analysis

R Programming

```
> # Define the dependent variable (foodtotal_v) and independent variables
> Y <- df$foodtotal_v
> X <- df[, c("pickle_v", "sauce_jam_v", "Beveragestotal_v")]
> Y
   [1] 1141.4924 1244.5535 1050.3154 1142.5917  945.2495 1579.2350  863.5380
   [8]  474.4830 2100.9125 1342.8750  424.3094  585.9627  645.5683  592.3710
  [15]  709.6209  372.5332 1321.6733 1092.8000  813.1741 1429.7395  746.9847
  [22]  890.6175 1183.7970  857.2342  894.1895 1351.4282  761.3110  942.9400
  [29]  762.1797 1093.9575  477.4590  458.3511  945.9763 1112.0728  551.6457
  [36]  522.8438  526.8072  712.7325  501.0832  440.9847 1347.6490 1084.0817
  [43] 1009.8480 1138.5965  837.3700 1103.6960  798.2608  697.5310 1595.9825
  [50] 1058.7062 1146.9075  593.1574  905.3710  935.9783  503.8500  616.5845
  [57] 1111.6800 1114.0648  712.7798  542.8758  709.3750  696.1478  386.6984
  [64]  436.4350  953.5747 1365.7215 1025.9065  687.3452  546.7414  501.3574

> X
# A tibble: 101,662 × 3
   pickle_v sauce_jam_v Beveragestotal_v
      <dbl>       <dbl>            <dbl>
 1    0               0              0
 2    0               0             17.5
 3    0               0              0
 4    0               0             33.3
 5    0               0             75
 6    0.005           0             50.0
 7    0.0016          0             30.0
 8    0               0              0
 9    0               0             75
10    0               0              0
# ℹ 101,652 more rows
# ℹ Use `print(n = ...)` to see more rows
> # Fit the regression model
> model <- lm(Y ~ pickle_v + sauce_jam_v + Beveragestotal_v, data = df)
> model

Call:
lm(formula = Y ~ pickle_v + sauce_jam_v + Beveragestotal_v, data = df)

Coefficients:
     (Intercept)          pickle_v       sauce_jam_v  Beveragestotal_v
         575.032         17128.559         17389.629             1.552
```

```
> # Print the summary of the regression
> summary(model)

Call:
lm(formula = Y ~ pickle_v + sauce_jam_v + Beveragestotal_v, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-13577.6   -194.3    -53.7    132.5  15408.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.750e+02  1.094e+00  525.52   <2e-16 ***
pickle_v         1.713e+04  2.960e+02   57.87   <2e-16 ***
sauce_jam_v      1.739e+04  3.313e+02   52.49   <2e-16 ***
Beveragestotal_v 1.552e+00  1.184e-02  131.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 327.8 on 101658 degrees of freedom
Multiple R-squared:  0.2046,    Adjusted R-squared:  0.2045
F-statistic:  8715 on 3 and 101658 DF,  p-value: < 2.2e-16
```

Python

```python
# Select dependent and independent variables
Y = df['foodtotal_v']
X = df[['pickle_v', 'sauce_jam_v', 'Othrprocessed_v', 'Beveragestotal_v', 'fruits_df_tt_v', 'fv_tot']]
```

```python
print("Independent variable")
X.head(20)
```

```
Independent variable
```

[4]:

	pickle_v	sauce_jam_v	Othrprocessed_v	Beveragestotal_v	fruits_df_tt_v	fv_tot
0	0.0000	0.0	0.0	0.000000	12.000000	154.180000
1	0.0000	0.0	0.0	17.500000	333.000000	484.950000
2	0.0000	0.0	0.0	0.000000	35.000000	214.840000
3	0.0000	0.0	0.0	33.333333	168.333333	302.300000
4	0.0000	0.0	0.0	75.000000	15.000000	148.000000
5	0.0050	0.0	0.0	50.005000	115.933333	255.600000

```
print("Dependent Variable")
Y.head(20)
```

```
Dependent Variable

5]: 0      1141.492400
    1      1244.553500
    2      1050.315400
    3      1142.591667
    4       945.249500
    5      1579.235000
    6       863.538000
    7       474.483000
    8      2100.912500
    9      1342.875000
    10      424.309400
    11      585.962667
```

PERFORMING MULTIPLE REGRESSION ANALYSIS

```
import statsmodels.api as sm

# Add a constant to the model (intercept)
X = sm.add_constant(X)

# Fit the regression model
model = sm.OLS(Y, X).fit()

# Print the summary of the regression
print(model.summary())
```

```
                          OLS Regression Results
==============================================================================
Dep. Variable:            foodtotal_v   R-squared:                       0.602
Model:                            OLS   Adj. R-squared:                  0.602
Method:                 Least Squares   F-statistic:                 2.560e+04
Date:                Sat, 22 Jun 2024   Prob (F-statistic):               0.00
Time:                        00:52:00   Log-Likelihood:            -6.9796e+05
No. Observations:              101662   AIC:                         1.396e+06
Df Residuals:                  101655   BIC:                         1.396e+06
Df Model:                           6
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              282.4196      1.288    219.203      0.000     279.894     284.945
pickle_v          3989.0940    213.466     18.687      0.000    3570.702    4407.486
sauce_jam_v       3152.7159    240.955     13.084      0.000    2680.448    3624.984
Othrprocessed_v     -0.9657      0.024    -40.580      0.000      -1.012      -0.919
Beveragestotal_v     1.9710      0.022     90.083      0.000       1.928       2.014
fruits_df_tt_v      -0.7062      0.023    -30.630      0.000      -0.751      -0.661
fv_tot               2.9570      0.014    205.201      0.000       2.929       2.985
==============================================================================
Omnibus:                    86676.404   Durbin-Watson:                   1.406
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         28270090.251
Skew:                           3.210   Prob(JB):                         0.00
Kurtosis:                      84.441   Cond. No.                     5.15e+04
==============================================================================
```

<u>Interpretation</u>

In this section of the code, we define the variables for our multiple regression analysis. This prepares the data for multiple regression analysis to examine how the specified food categories influence total food consumption.

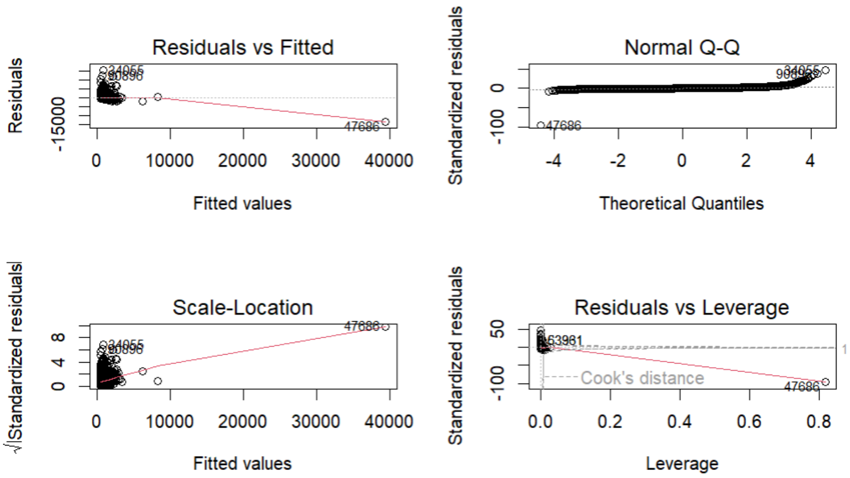
An R squared value of 0.2046 in R and 0.602 in Python indicates that approximately 20.46% and 60.2% of the variance in total food consumption is explained by the model. (The model explains only about 20.46% of the variance in total food consumption, suggesting other factors also play a substantial role.) The F statistic and low p value ($p <$ level of significance) indicates that the model is highly significant. The coefficients suggest that increases in pickle and sauce/jam consumption are associated with substantial increases in total food consumption, while increases in beverage consumption have a smaller, yet significant, effect.

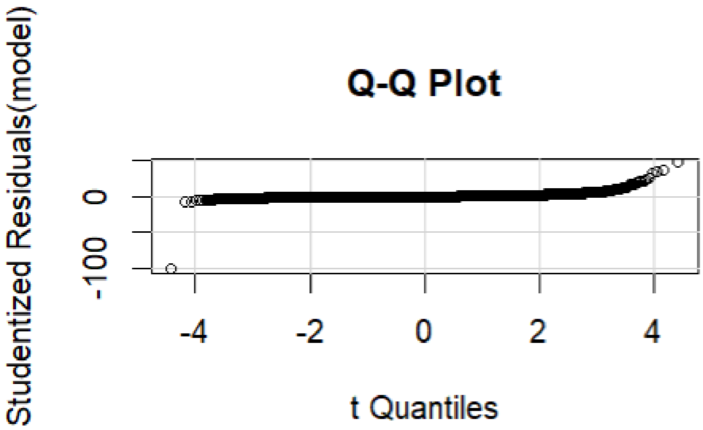
2. Regression Diagnostics

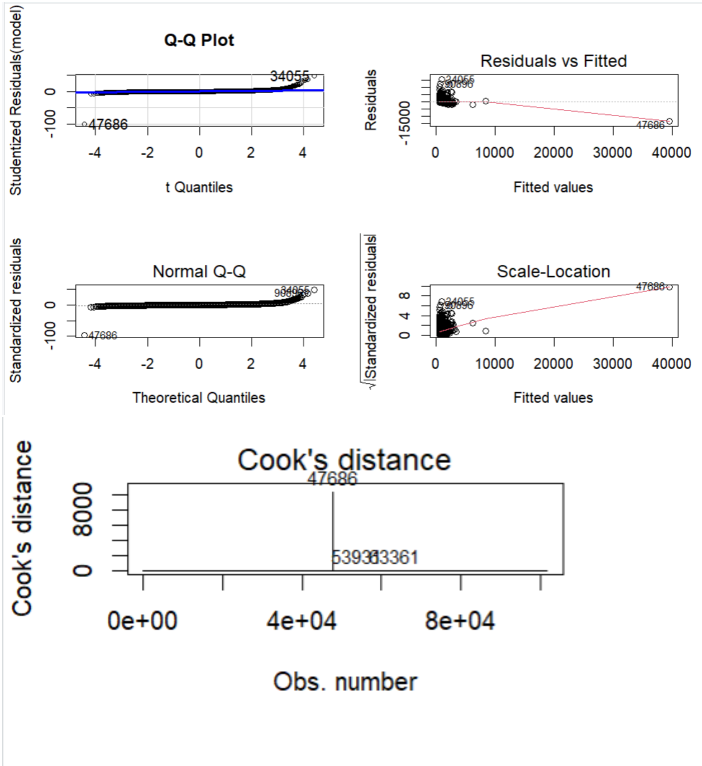
R

```
> # Q-Q plot of residuals
> qqPlot(model, main="Q-Q Plot")
[1] 34055 47686
> # Breusch-Pagan test for heteroscedasticity
> bptest(model)

	studentized Breusch-Pagan test

data:  model
BP = 19675, df = 3, p-value < 2.2e-16

> # Check for multicollinearity using VIF
> vif(model)
       pickle_v      sauce_jam_v Beveragestotal_v
       1.011643         1.016605         1.009533
```

```
> # Plot residuals vs fitted values
> plot(model, which = 1)
> # Q-Q plot of residuals
> plot(model, which = 2)
> # Scale-location plot (to check homoscedasticity)
> plot(model, which = 3)
> # Cook's distance plot (to identify influential points)
> plot(model, which = 4)
> # Check for multicollinearity using VIF
> library(car)
> vif(model)
       pickle_v      sauce_jam_v Beveragestotal_v
       1.011643         1.016605         1.009533
```



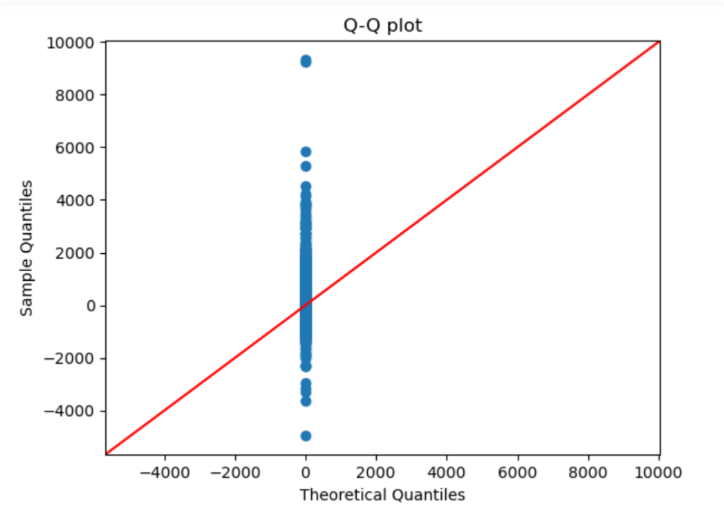




Python

```
#RESIDUAL PLOT
import matplotlib.pyplot as plt
import seaborn as sns

# Plot residuals
residuals = model.resid
fitted = model.fittedvalues
```

```
import scipy.stats as stats

# Q-Q plot
fig = sm.qqplot(residuals, line ='45')
plt.title('Q-Q plot')
plt.show()
```




```
#Homoscadasticity test
from statsmodels.stats.diagnostic import het_breuschpagan

# Perform the Breusch-Pagan test
bp_test = het_breuschpagan(residuals, X)
labels = ['LM Statistic', 'LM Test p-value', 'F-Statistic', 'F-Test p-value']
print(dict(zip(labels, bp_test)))
```

```
{'LM Statistic': 4699.575655431272, 'LM Test p-value': 0.0, 'F-Statistic': 821.169242419054, 'F-Test p-value': 0.0}
```

```
#Multicollinearity check
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Calculate VIF for each explanatory variable
vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif["features"] = X.columns

print(vif)
```

```
   VIF Factor         features
0    3.136222            const
1    1.051007         pickle_v
2    1.074253       sauce_jam_v
3    6.659432   Othrprocessed_v
4    6.882912  Beveragestotal_v
5    2.963057     fruits_df_tt_v
```

<u>Interpretations</u>

The multiple regression analysis indicates that the model is highly significant, with an LM Statistic of 4699.576 and an F-Statistic of 821.169, both with p-values of 0.0, indicating strong evidence against the null hypothesis. The Breusch-Pagan test for heteroscedasticity yields a BP value of 19675 with a p-value less than 2.2e-16, suggesting significant heteroscedasticity in the model. Variance Inflation Factor (VIF) values for the predictors pickle_v, sauce_jam_v, and Beveragestotal_v are 1.011643, 1.016605, and 1.009533 respectively, indicating low multicollinearity among these variables. These results imply that while the predictors are significant and the model fit is strong, there are issues with heteroscedasticity that need to be addressed.

3. Correcting the model if necessary

R

```
> # Fit the regression model with log-transformed dependent variable
> model_corrected <- lm(log_foodtotal_v ~ pickle_v + sauce_jam_v + Beveragestotal
_v, data = df)
> # Print the summary of the corrected regression
> summary(model_corrected)

Call:
lm(formula = log_foodtotal_v ~ pickle_v + sauce_jam_v + Beveragestotal_v,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.2719  -0.1556   0.1443   0.4342   3.2667 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.134e+00  5.485e-03 1118.23   <2e-16 ***
pickle_v         3.148e+01  1.484e+00   21.21   <2e-16 ***
sauce_jam_v      2.673e+01  1.661e+00   16.10   <2e-16 ***
Beveragestotal_v 1.210e-03  5.937e-05   20.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.643 on 101658 degrees of freedom
Multiple R-squared:  0.01272,	Adjusted R-squared:  0.01269 
F-statistic: 436.4 on 3 and 101658 DF,  p-value: < 2.2e-16
```

#python

```python
# Remove variables with high VIF
X_corrected = X.drop(['Othrprocessed_v'], axis=1)  # Example: Removing one variable

# Fit the corrected model
model_corrected = sm.OLS(Y, X_corrected).fit()

# Print the summary of the corrected regression
print(model_corrected.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:            foodtotal_v   R-squared:                       0.595
Model:                            OLS   Adj. R-squared:                  0.595
Method:                 Least Squares   F-statistic:                 2.991e+04
Date:                Sat, 22 Jun 2024   Prob (F-statistic):               0.00
Time:                        00:53:34   Log-Likelihood:            -6.9878e+05
No. Observations:              101662   AIC:                         1.398e+06
Df Residuals:                  101656   BIC:                         1.398e+06
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              286.5848      1.295    221.361      0.000     284.047     289.122
pickle_v          4263.9554    215.079     19.825      0.000    3842.403    4685.508
sauce_jam_v       4214.2561    241.462     17.453      0.000    3740.995    4687.518
Beveragestotal_v     1.1525      0.009    134.825      0.000       1.136       1.169
fruits_df_tt_v      -0.6654      0.023    -28.655      0.000      -0.711      -0.620
fv_tot               3.0505      0.014    212.728      0.000       3.022       3.079
==============================================================================
Omnibus:                    87076.759   Durbin-Watson:                   1.401
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          26955541.548
Skew:                           3.252   Prob(JB):                         0.00
Kurtosis:                      82.506   Cond. No.                     5.10e+04
==============================================================================
```

```python
#Homoscadasticity test
from statsmodels.stats.diagnostic import het_breuschpagan

# Perform the Breusch-Pagan test
bp_test = het_breuschpagan(residuals, X)
labels = ['LM Statistic', 'LM Test p-value', 'F-Statistic', 'F-Test p-value']
print(dict(zip(labels, bp_test)))
```

```
{'LM Statistic': 4699.575655431272, 'LM Test p-value': 0.0, 'F-Statistic': 821.169242419054, 'F-Test p-value': 0.0}
```

```python
#Multicollinearity check
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Calculate VIF for each explanatory variable
vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor(X_corrected.values, i) for i in range(X_corrected.shape[1])]
vif["features"] = X_corrected.columns

print(vif)
```

```
   VIF Factor         features
0    3.116317            const
1    1.049948         pickle_v
2    1.061592      sauce_jam_v
3    1.033869  Beveragestotal_v
4    2.957412    fruits_df_tt_v
5    3.001563           fv_tot
```

4. Comparison and interpretation

```
Initial Model Summary
                            OLS Regression Results
==============================================================================
Dep. Variable:            foodtotal_v   R-squared:                       0.602
Model:                            OLS   Adj. R-squared:                  0.602
Method:                 Least Squares   F-statistic:                 2.560e+04
Date:                Sat, 22 Jun 2024   Prob (F-statistic):               0.00
Time:                        00:56:27   Log-Likelihood:            -6.9796e+05
No. Observations:              101662   AIC:                         1.396e+06
Df Residuals:                  101655   BIC:                         1.396e+06
Df Model:                           6
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              282.4196      1.288    219.203      0.000     279.894     284.945
pickle_v          3989.0940    213.466     18.687      0.000    3570.702    4407.486
sauce_jam_v       3152.7159    240.955     13.084      0.000    2680.448    3624.984
Othrprocessed_v     -0.9657      0.024    -40.580      0.000      -1.012      -0.919
Beveragestotal_v     1.9710      0.022     90.083      0.000       1.928       2.014
fruits_df_tt_v      -0.7062      0.023    -30.630      0.000      -0.751      -0.661
fv_tot               2.9570      0.014    205.201      0.000       2.929       2.985
==============================================================================
Omnibus:                    86676.404   Durbin-Watson:                   1.406
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          28270090.251
Skew:                           3.210   Prob(JB):                         0.00
Kurtosis:                      84.441   Cond. No.                     5.15e+04
```

```
Corrected Model Summary
                            OLS Regression Results
==============================================================================
Dep. Variable:            foodtotal_v   R-squared:                       0.595
Model:                            OLS   Adj. R-squared:                  0.595
Method:                 Least Squares   F-statistic:                 2.991e+04
Date:                Sat, 22 Jun 2024   Prob (F-statistic):               0.00
Time:                        00:56:27   Log-Likelihood:            -6.9878e+05
No. Observations:              101662   AIC:                         1.398e+06
Df Residuals:                  101656   BIC:                         1.398e+06
Df Model:                           5
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              286.5848      1.295    221.361      0.000     284.047     289.122
pickle_v          4263.9554    215.079     19.825      0.000    3842.403    4685.508
sauce_jam_v       4214.2561    241.462     17.453      0.000    3740.995    4687.518
Beveragestotal_v     1.1525      0.009    134.825      0.000       1.136       1.169
fruits_df_tt_v      -0.6654      0.023    -28.655      0.000      -0.711      -0.620
fv_tot               3.0505      0.014    212.728      0.000       3.022       3.079
==============================================================================
Omnibus:                    87076.759   Durbin-Watson:                   1.401
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         26955541.548
Skew:                           3.252   Prob(JB):                         0.00
Kurtosis:                      82.506   Cond. No.                     5.10e+04
==============================================================================
```

Interpretation

The initial model explains 60.2% of the variance in `foodtotal_v`, with `pickle_v`, `sauce_jam_v`, `Othrprocessed_v`, `Beveragestotal_v`, `fruits_df_tt_v`, and `fv_tot` as predictors. Significant coefficients indicate that increases in `pickle_v`, `sauce_jam_v`, `Othrprocessed_v`, and `Beveragestotal_v` are associated with higher `foodtotal_v`. The model has issues with multicollinearity (high condition number) and possible numerical instability, as indicated by the notes.

The corrected model maintains strong explanatory power with an adjusted \(R^2 \) of 59.5%. It omits `Othrprocessed_v`, addressing multicollinearity concerns, resulting in improved model stability and interpretability. Coefficients for `pickle_v`, `sauce_jam_v`, `Beveragestotal_v`, `fruits_df_tt_v`, and `fv_tot` remain significant and consistent with the initial model, showing their positive associations with `foodtotal_v`. The model's diagnostic metrics, such as Omnibus and Jarque-Bera tests, indicate good model fit and assumptions of normality.

The corrected model offers a refined approach to predicting `foodtotal_v`, maintaining predictive power while addressing statistical issues observed in the initial model.