Multiple regression analysis of Total Food Expenditure in Carga Philippines

Nivedita Pandey, MSc(Statistics), Bannasthali Vidyapeeth, Rajasthan

ABSTRACT

This report presents the results of a study examining the relationship between total food expenditure and total household income, total number of family members, total number of employed family members in Carga Philippines. This analysis used the data collected by Family Income and Expenditure Survey (FIES) in the Philippines. The dependent variable was total food expenditure, while the independent variables were the following: total household income, total number of family members, and total number of employed family members. Multicollinearity, heteroskedasticity, autocorrelation is tested for this data.

INTRODUCTION

Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

If we are studying the dependence of a variable on only a single explanatory variable, such a study is known as simple, or two-variable, regression analysis. However, if we are studying the dependence of one variable on more than one explanatory variable, it is known as multiple regression analysis. Multiple regression analysis model is given by:

$$Y_i = \beta_o + \beta_1 X_{1i} \; + \; \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + u_i$$

Where, Y_i is ith observation of dependent variable

$X_{ji}$is the ith observation on jth independent variable

β_j's are regression coefficients

u_i's are disturbance term

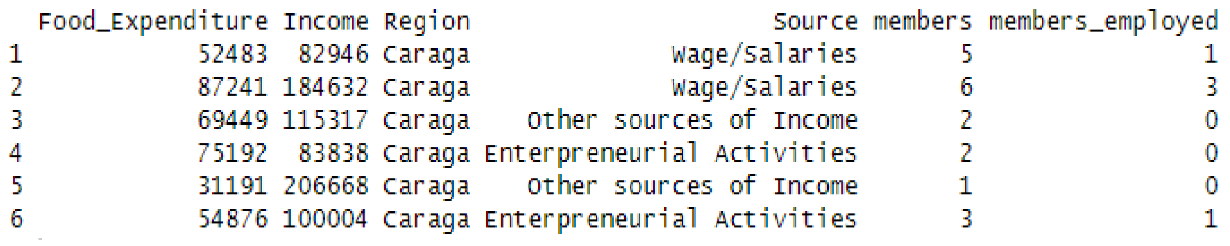
i=1, 2,…, n , j=1,2,…,p

Through multiple regression analysis, our concern was with estimating the average total food expenditure knowing the total household income, total number of family members, and total number of employed family members in Carga Philippines.

METHODOLOGY

This is a descriptive study analyzing whether or not the three independent variables (namely, total household income, total number of family members, and total number of employed family members) are able to estimate the average total food expenditure using multiple regression analysis.

- **DATA-** In this analysis secondary data which is taken from website 'Kaggel'. Inside the data set is some selected variables from the Family Income and Expenditure Survey (FIES) in the Philippines. It contains more than 40k observations and 60 variables which is primarily comprised of the household income and expenditures of that specific household. From 60 variables 6 variables are taken for the analysis of only 'Carga' region of Philippines. The 6 variables are total food expenditure, total household income, region, source of income, total number of family members, and total number of employed family members. Glimpse of data is:



```
  Food_Expenditure Income Region                     Source members members_employed
1            52483  82946 Caraga               Wage/Salaries       5                1
2            87241 184632 Caraga               Wage/Salaries       6                3
3            69449 115317 Caraga    Other sources of Income       2                0
4            75192  83838 Caraga Enterpreneurial Activities       2                0
5            31191 206668 Caraga    Other sources of Income       1                0
6            54876 100004 Caraga Enterpreneurial Activities       3                1
```

- **Bar Graph-**A bar graph shows comparisons among discrete categories .One axis of the chart shows the specific categories being compared, and the axis represents a measure value in this graph the bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

- **Multicollinearity-** The term multicollinearity is due to Ragnar Frisch. Originally it meant the existence of a "perfect," or exact, linear relationship among some or all explanatory variables of a regression model. To detect Multicollinearity some methods are as follows-

  - **Variance-inflating factor (VIF) -** The speed with which variances and covariances increase can be seen with the variance inflation factor. VIF shows how the variance of an estimator is inflated by the presence of multicollinearity. The inverse of the VIF is called tolerance.As a rule of thumb, a VIF exceeding 5 requires further investigation, whereas VIFs above 10 indicate multicollinearity**.** Ideally, the Variance Inflation Factors are below 3.

  - **Correlation matrix-**A correlation matrix (or correlogram) visualizes the correlation between multiple continuous variables. Correlations range always between -1 and +1, where -1 represents perfect negative correlation and +1 perfect positive correlation. Correlations close to-1 or +1 might indicate the existence of multicollinearity. As a rule of thumb, one might suspect multicollinearity when the correlation between two (predictor) variables is below -0.9 or above +0.9.

  - **Eigenvalues and condition index**- We can find the eigen-Values and the condition index, to diagnose multicollinearity. We can derive what is known as the Condition Number k defined as

$$k = \frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}$$

And the condition index (CI) defined as

$$CI = \sqrt{\frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}} = \sqrt{k}$$

Then we have this rule of thumb: If k is between 100 and 1000 there is moderate to strong multicollinearity and if it exceeds 1000 there is severe multicollinearity. Alternatively, if the CI (= √k) is between 10 and 30, there is moderate to strong multicollinearity and if it Exceeds 30 there is severe multicollinearity.

- **Heteroscedasticity-**One of the important assumptions of the classical linear regression Model is that the variance of each disturbance term, conditional on the chosen values of the explanatory variables, is some constant. This is the assumption of homoscedasticity, if the variance is not constant there is Heteroscedasticity. To detect heteroscedasticity there are many tests like-
  - Park Test
  - Glejser Test
  - Spearman's Rank Correlation Test
  - Goldfeld–Quandt Test
  - Breusch–Pagan–Godfrey Test
  - White's General Heteroscedasticity Test

- **Autocorrelation-** The term autocorrelation may be defined as "correlation between members of series of observations ordered in time [as in time series data] or space [as in cross-sectional data].'' in the regression context, the classical linear regression model assumes that such autocorrelation does not exist in the disturbances. To detect autocorrelation there are many test like-
  - The Runs Test
  - Durbin–Watson d Test
  - The Breusch–Godfrey (BG) Test

- **GLS (Generalized least square estimator):** The generalized least squares (GLS) estimator of the coefficients of a linear regression is a generalization of the ordinary least squares (OLS) estimator. It is used to deal with situations in which the OLS estimator is not BLUE (best linear unbiased estimator) because one of the main assumptions of the Gauss-Markov theorem, namely that of homoskedasticity and absence of serial correlation, is violated. In such situations, provided that the other assumptions of the Gauss-Markov theorem are satisfied, the GLS estimator is BLUE.

DATA ANALYSIS & INTERPRETATION

The whole data set has 6 columns namely total food expenditure, total house hold income, region, source of income, total number of family members and total number of employed family members. The column of region indicates that the data is taken only on people living in Carga, Philippines and from the column source of income of is wage /salary, enterpreneurial Activities and other sources.

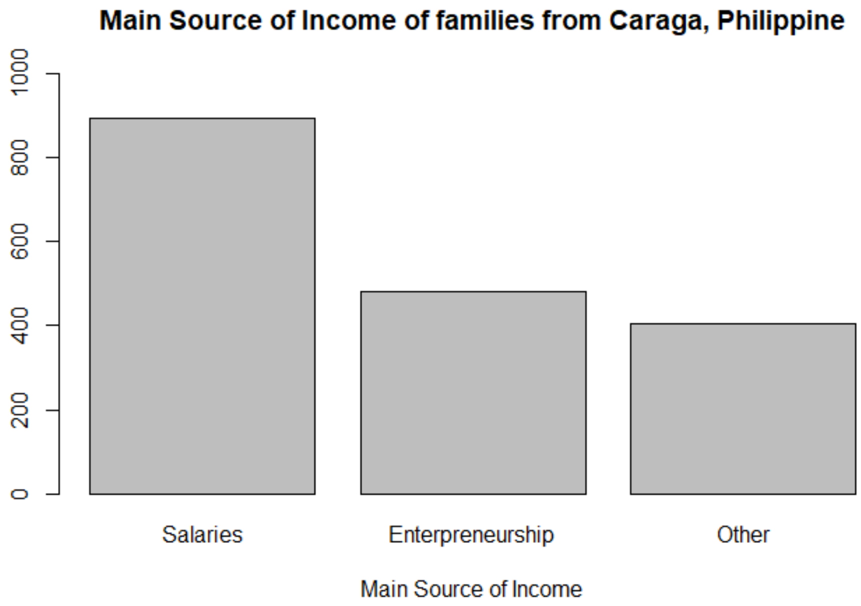
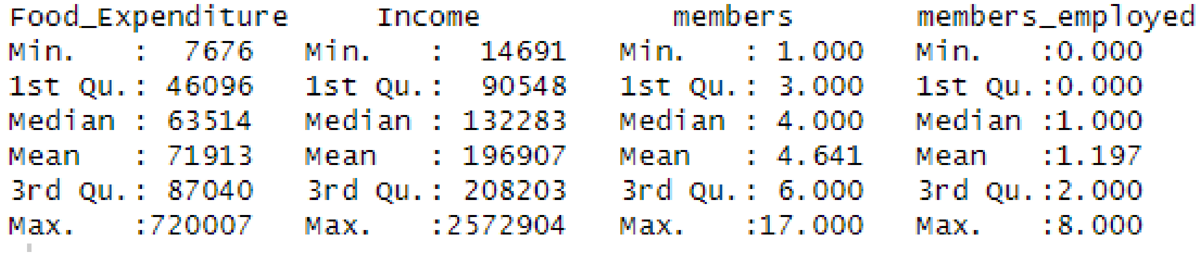


Fig 1: Bar chart of main source of income of families in Caraga, Philippines

From the bar chart in fig 1 we can see that main source of income of maximum families in Caraga, Philippines is wages/salaries and it is approximately double to the enterpreneurial Activities. Summary of remaining columns is:



```
 Food_Expenditure      Income            members      members_employed
 Min.   :  7676   Min.   :  14691   Min.   : 1.000   Min.   :0.000
 1st Qu.: 46096   1st Qu.:  90548   1st Qu.: 3.000   1st Qu.:0.000
 Median : 63514   Median : 132283   Median : 4.000   Median :1.000
 Mean   : 71913   Mean   : 196907   Mean   : 4.641   Mean   :1.197
 3rd Qu.: 87040   3rd Qu.: 208203   3rd Qu.: 6.000   3rd Qu.:2.000
 Max.   :720007   Max.   :2572904   Max.   :17.000   Max.   :8.000
```

Fig 2: Summary of remaining columns of data

From fig 2 we can see that in Caraga, Philippines the average total food expenditure is 71913 PHP (PHP - Philippine Peso, PHP is currency of Philippines), average total household income is 196907 PHP, on an average there are approximately 5 persons in family and on an average approximately one family members employed.

The regression model for the data used for analysis is –

$$Y_i = \beta_o + \beta_1 X_{1i} \ + \ \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Where,

$\mathrm{Y_i}$ is i$^{\text{th}}$ observation of total food expenditure

$\mathrm{X_{1i}}$is the i$^{\text{th}}$ observation on total household income

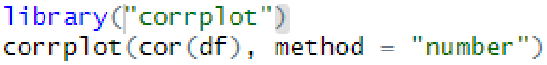
$\mathrm{X_{2i}}$is the i$^{\text{th}}$ observation on total number of family members

$\mathrm{X_{3i}}$is the i$^{\text{th}}$ observation on total number of employed family members

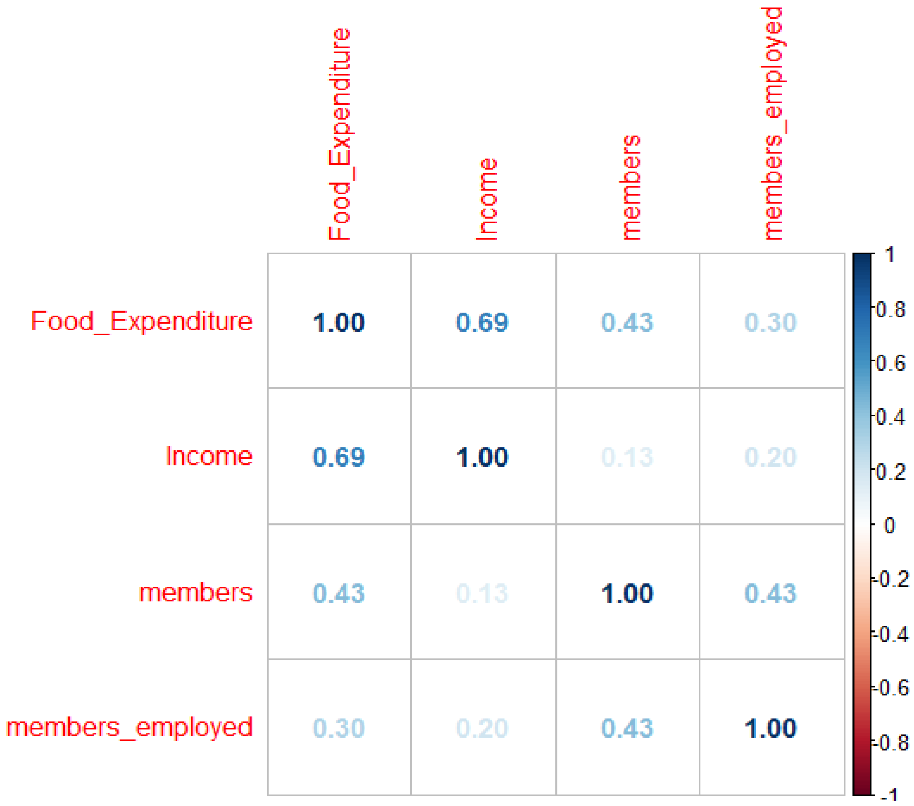
$\beta_0, \beta_1, \beta_2, \beta_3$ are regression coefficients, i=1, 2,…, 1782, $\mathrm{u_i}$'s are disturbance term

- **To check multicollinearity**

Before predicting and estimating the average total food expenditure first, we will check the presence of multicollinearity. To test multicollinearity first we created a correlation matrix.



```
library("corrplot")
corrplot(cor(df), method = "number")
```

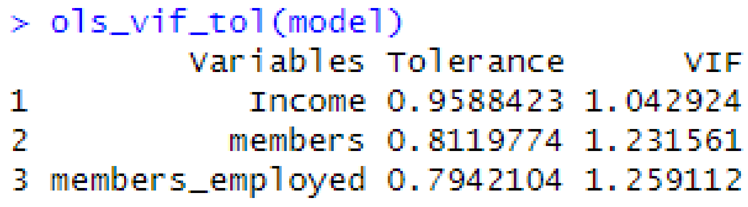


	Food_Expenditure	Income	members	members_employed
Food_Expenditure	1.00	0.69	0.43	0.30
Income	0.69	1.00	0.13	0.20
members	0.43	0.13	1.00	0.43
members_employed	0.30	0.20	0.43	1.00

Fig 3: Correlation matrix of the explanatory variables in the regression model

The image above shows the correlation matrix of the variables that are included in our regression model. There is not so high correlation between any of the explanatory variables that might indicate multicollinearity.

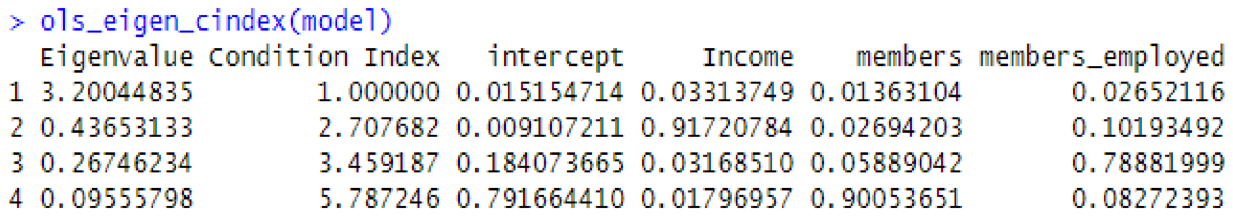
Then we calculated VIF



```
> ols_vif_tol(model)
         Variables Tolerance      VIF
1           Income 0.9588423 1.042924
2          members 0.8119774 1.231561
3 members_employed 0.7942104 1.259112
```

We can see that all the value of VIF is less than 10 and even less than 3 then and value tolerance is far away from 0.1 then it indicates that multicollinearity is not present.

We also checked the multicollinearity using eigenvalues and condition index



```
> ols_eigen_cindex(model)
  Eigenvalue Condition Index   intercept     Income    members members_employed
1 3.20044835        1.000000 0.015154714 0.03313749 0.01363104       0.02652116
2 0.43653133        2.707682 0.009107211 0.91720784 0.02694203       0.10193492
3 0.26746234        3.459187 0.184073665 0.03168510 0.05889042       0.78881999
4 0.09555798        5.787246 0.791664410 0.01796957 0.90053651       0.08272393
```

We can see that the values of condition index number is less than 30 even less than 10 so it is also indicating that there is no multicollinearity.

- **To check heteroscedasticity**

After checking multicollinearity we checked for heteroscedasticity. To check heteroscedasticity first we used the informal method which is graphical method.

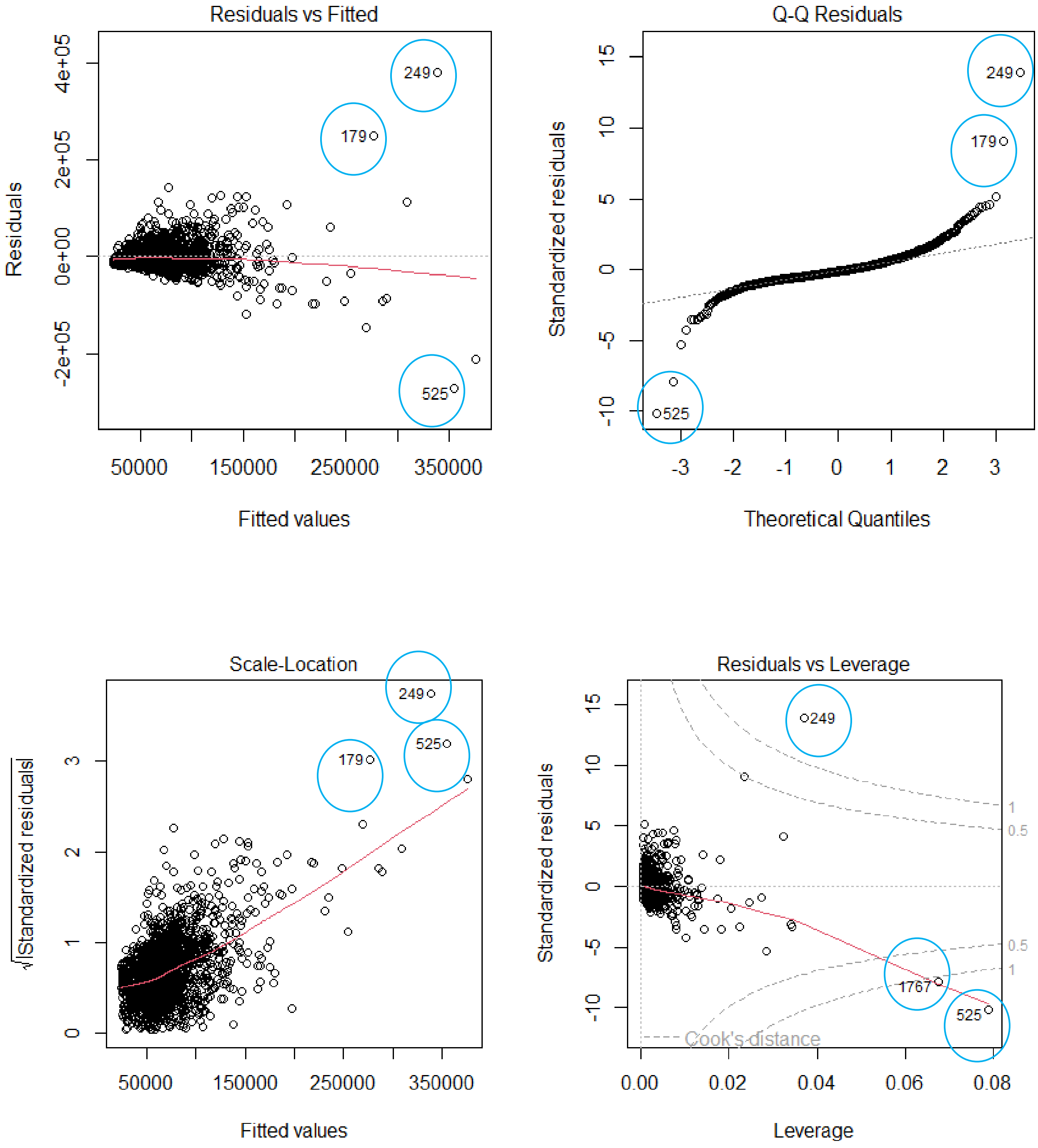


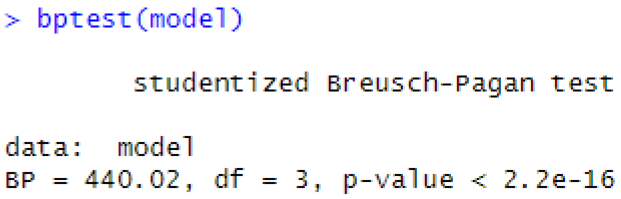
Fig4: Plots to check heteroscedasticity

From the plot in Fig 4 we can see that the outliers are present which are marked with blue circle and we know that presence of outliers is one of the causes of heteroscedasticity. After that we used tests to check heteroscedasticity.

First we used Breusch–Pagan–Godfrey Test. This test checks whether the variance of the residuals depends on the value of the independent variable.Therefore, the hypothesis of the Breusch–Pagan–Godfrey Testis:

H_0 : Residuals are distributed with equal variances (homoscedastic)

H_1 :Residuals are distributed with unequal variances (heteroscedastic)



```
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 440.02, df = 3, p-value < 2.2e-16
```

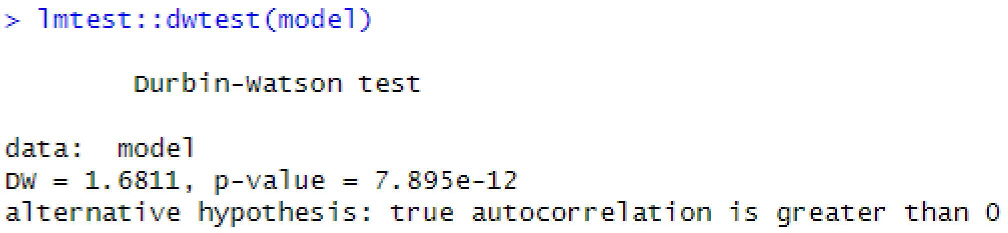
The p-value is less than the level of significance(0.05) so there is enough evidence to reject null hypothesis that means residuals are distributed with unequal variances i.e. heteroscedastic.

- **<u>To check Autocorrelation</u>**

Now we test for autocorrelation for this we used Durbin–Watson d Test**The hypotheses of the Durbin-Watson test are:**

H_0 :First order autocorrelation does not exist

H_1 :First order autocorrelation does exist



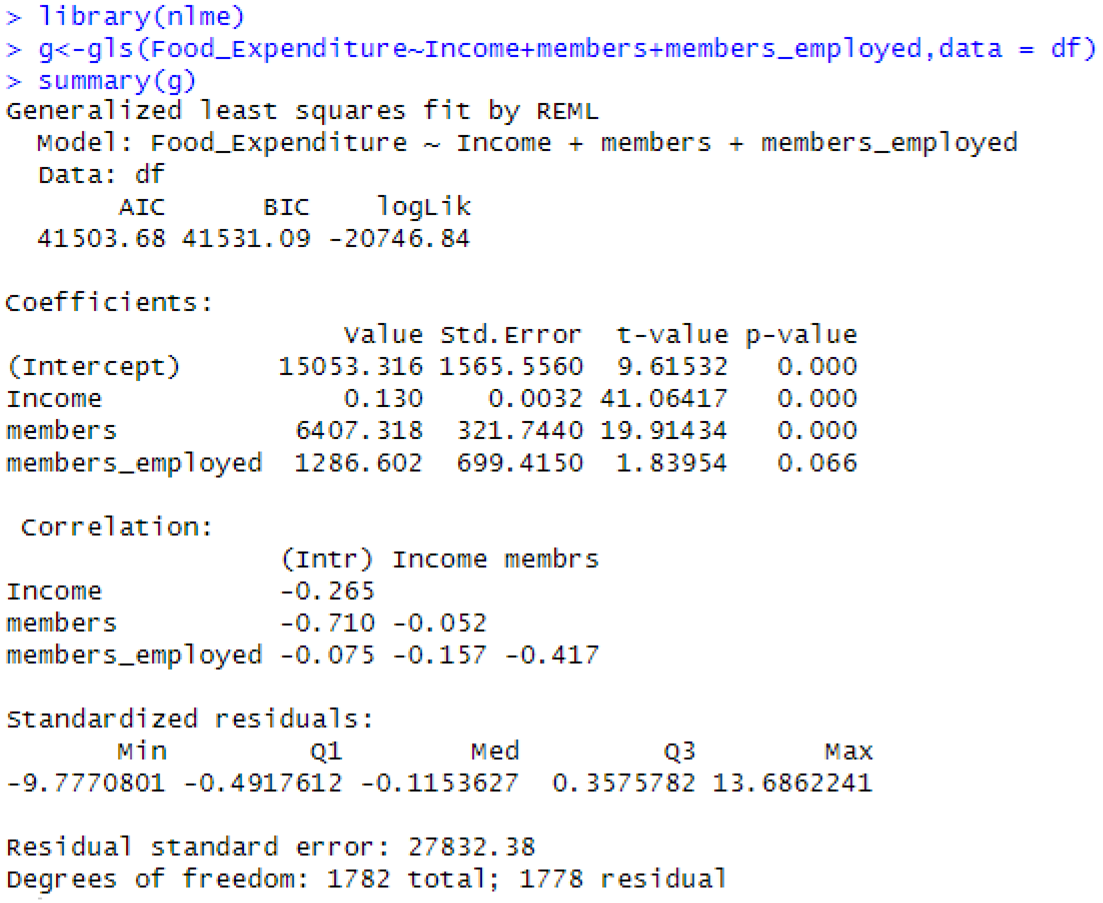
```
> lmtest::dwtest(model)

        Durbin-Watson test

data:  model
DW = 1.6811, p-value = 7.895e-12
alternative hypothesis: true autocorrelation is greater than 0
```

We can see that the DW=1.6811 and from Durbin–Watson table the lower and upper points are1.918, 1.925 respectively. The calculated value 1.6811 lies between o and lower point then there is first order positive autocorrelation. Also the p-value is less than the level of significance

(0.05) so there is enough evidence to reject null hypothesis that means first order autocorrelation exist.

Use of GLS (Generalized least square estimator)

In the data we have seen that there is no multicolliniearity but heterscdasticity and autocorrelation is present. So instead of using OLS (Ordinary Least Square estimator) we will use GLS (Generalized least square estimator)



```
> library(nlme)
> g<-gls(Food_Expenditure~Income+members+members_employed,data = df)
> summary(g)
Generalized least squares fit by REML
  Model: Food_Expenditure ~ Income + members + members_employed
  Data: df
       AIC      BIC    logLik
  41503.68 41531.09 -20746.84

Coefficients:
                     Value Std.Error  t-value p-value
(Intercept)      15053.316 1565.5560  9.61532   0.000
Income               0.130    0.0032 41.06417   0.000
members           6407.318  321.7440 19.91434   0.000
members_employed  1286.602  699.4150  1.83954   0.066

 Correlation: 
                 (Intr) Income membrs
Income           -0.265              
members          -0.710 -0.052       
members_employed -0.075 -0.157 -0.417

Standardized residuals:
       Min         Q1        Med         Q3        Max 
-9.7770801 -0.4917612 -0.1153627  0.3575782 13.6862241 

Residual standard error: 27832.38 
Degrees of freedom: 1782 total; 1778 residual
```

So, if income goes up by 1 PHP then the average food expenditure goes up by 0.13 PHP holding the value of total number of family members and total number of employed family members constant. If one member in the family increases then the average food expenditure goes up by 6407.318 PHP holding the value of total household income and total number of employed family members constant. If one more member in the family gets employed then the average food expenditure goes up by 1286.06 PHP holding the value of total household income and total number of family members constant.

For testing our hypotheses will be-

H_{01}: there is no relationship between total food expenditure and total household income, that is, $\beta_1 = 0$.

H_{02}: there is no relationship between total food expenditure and total number of family members, that is, $\beta_2 = 0$.

H_{03}: there is no relationship between total food expenditure and total number of employed family members, that is, $\beta_3 = 0$.

We can see that p value corresponding to income is less than the level of significance 0.05 ($0.00<0.05$) then we have enough evidence to reject the null hypothesis H_{01} that means total household income have significant effect on total food expenditure at 5% level of significance.

p value corresponding to total number of family members is less than the level of significance 0.05 ($0.00<0.05$) then we have enough evidence to reject the null hypothesis H_{02} that means total number of family members have significant effect on total food expenditure at 5% level of significance.

p value corresponding to total number of employed family members is greater than the level of significance 0.05 ($0.066<0.05$) then we have enough evidence to do not reject the null hypothesis H_{02} that means total number of employed family members do not have significant effect on total food expenditure at 5% level of significance.

CONCLUSION

In this analysis through multiple regression analysis, we estimated the average total food expenditure knowing the total household income, total number of family members, and total number of employed family members in *Carga Philippines* using secondary data which was taken from website '*Kaggel*'. Inside the data set are some selected variables from the Family Income and Expenditure Survey (FIES) in the Philippines. 6 variables were taken for the analysis of only 'Carga' region of Philippines. The 6 variables were total food expenditure, total household income, region, source of income, total number of family members, and total number of employed family members.

We found that main source of income of maximum families in Caraga, Philippines is wages/salaries and it is approximately double to the other source which was enterpreneurial Activities.

In the analysis we found that there was no multicolliniearity but the variance of each disturbance term, conditional on the chosen values of the explanatory variables, was not constant and disturbance term war not unautocorrelated that means there was heteroscedasticity and autocorrelation.

Because of heteroscedasticity and autocorrelation we used GLS (Generalized least square) estimator and found that the total household income and total number of family members have significant effect on total food expenditure at 5% level of significance. But total number of employed family members does not have significant effect on total food expenditure at 5% level of significance.

REFERENCE

- Gujarati, D. N. (2022). Basic econometrics. Prentice Hall.
- https://www.kaggle.com/datasets/grosvenpaul/family-income-and-expenditure/