ECON F342: Applied Econometrics

Cross-Sectional Data Analysis

on

"Determinants of Health Expenditure in Countries"

Anirudh Kathera	2020B3A71927H
Arjun Khandelwal	2020B3A70848H
Madhumita Phadke	2020B3AA0441H
Pralhad Riswadkar	2020B3AA1274H
Ritul Mohan Sharma	2020B3AA0832H
M Varun Srinivas	2020B3AA0497H
Tushar Komali Reddy	2020B3PS1381H
Shreyash Mahapatra	2020B3A71167H

Submitted to Dr. Rishi Kumar

Introduction

Economic growth possesses multiple determinants, including natural resources, technology, formation of capital, however one determinant - human capital, exerts a direct influence on all other factors. The productivity of human capital is a driver of economic growth, and thus the paper discusses one important determinant of human capital productivity: a country's expenditure on healthcare. A healthier workforce has been empirically proven to be a more efficient workforce with enhanced worker productivity.

Thus the aim of this paper is to find determinants of healthcare expenditure from a multitude of factors such as initial levels of health among individuals, the income of individuals in a state and their preferences for urban living, foreign intervention in country's economies, the composition of the labor force itself and an indicator of state priorities through defense expenditure. This paper using data from the world bank indicators from the year 2018, runs an OLS model trying to estimate the determinants of healthcare expenditure in countries.

Literature Review

Xu, K., Saksena, P., & Holly, A. (2011). The determinants of health expenditure: A country-level panel data analysis. Social Science & Medicine

The study examines the determinants of health expenditure at the country level using panel data analysis. The authors find that income, population age structure, and government expenditure on health are positively associated with health expenditure, while inflation and debt have a negative effect. The study also suggests that health expenditure is affected by factors such as economic development, demographic changes, and government policies. The authors conclude that policies aimed at improving health outcomes should consider these determinants of health expenditure.

Cantarero, D., & Lago-Peñas, S. (2013). Decomposing the determinants of health care expenditure: The case of Spain. The European Journal of Health Economics

The study analyses the determinants of healthcare expenditure in Spain using a decomposition approach. The authors find that the main drivers of healthcare expenditure in Spain are population ageing, income growth, and technological change. They also find that the demand for health care services has a significant impact on health care expenditure, while the effect of supply-side factors is relatively small. The study suggests that policies to control healthcare expenditure in Spain should focus on improving efficiency and reducing the demand for healthcare services. The authors also propose several policy recommendations, including the promotion of healthy lifestyles and the implementation of measures to reduce the use of unnecessary medical services.

Okunda, A., (2005). Analysis and Implications of the Determinants of Healthcare Expenditure
in African Countries.
The study examined the determinants of health expenditure in African countries using 1995 data and found that GDP and ODA had the most significant impact on health expenditure. Other factors, such as income inequality, also had an impact. The study suggested that future research should consider incorporating informal healthcare spending into their analysis. They also highlighted the decline in healthcare spending in Africa during the SAP era, which may have contributed to the deterioration of basic needs such as housing, education, and food. The study suggested that policies should focus on increasing stable finance for population health and productivity improvements in Africa, private sector cooperation with government and NGO providers of health services, and reducing income inequalities.

Piatti-Fünfkirchen, M., Lindelow, M., & Yoo, K. (2018). What are governments spending on health in East and Southern Africa? Health systems & reform
The article examines government spending on health in East and Southern Africa. The authors find that while overall government spending on health has increased in the region, it remains low compared to other regions of the world. Additionally, there is significant variation in spending across countries in the region, with some countries spending much more on health than others. The authors argue that increasing government spending on health is necessary to improve health outcomes in the region, but that this will require significant policy changes and increased political will.

Afonso, A., Hankins, M., & Soares, J. (2019). Relationship between female mortality and health care expenditure: Sustainable assessment of health care system.
The authors explain that while increased healthcare expenditure can lead to improved health outcomes and reduced mortality rates, it is not always sustainable in the long term.
They discuss the concept of "health care cost-effectiveness" and the importance of prioritising interventions that provide the most benefit for the lowest cost. They also use the variable Female Mortality to assess the sustainability of the healthcare system.
The conclusion is that a sustainable healthcare system must balance the need for quality care with cost-effectiveness and must prioritise policy interventions that provide health benefits for the population as a whole.

All these papers have used some standard variables like GDP per capita, % GDP health care, Infant Mortality, population per doctor, Life expectancy etc. We take insights from these papers and include new variables like refugee population and Foreign Direct Investment. We also try to find a correlation between defense expenditure of countries and their Health care expenditure.

Data

In this report, we have used a database of cross-sectional data to evaluate the relationship of several varied factors with health expenditure. For this analysis, 2018 has been taken as the year for all factors from each country, as the largest database was available for the same.

In the regression, the dependent variable is health expenditure with the proxy chosen being Current health expenditure per capita, PPP (current international $). The model includes 8 independent variables, which are listed below.

- Foreign Direct Investment as a proxy for economic growth and trade openness (specifically Foreign direct investment, net inflows as a percentage of GDP).

- Female Labour Force Participation as a proxy for gender and social development (specifically Labor force, female as a percentage of total labor force).
- GDP per capita as a proxy for the economic conditions of the country (specifically GDP per capita, PPP in current international $).
- Refugee Population as a proxy for government stability and social condition (specifically Refugee population by country or territory of origin).
- Life expectancy as a proxy for general health conditions of the population (specifically Life expectancy at birth, female in years).
- Defense Expenditure as a proxy for public sector investment and expenditure (specifically Military expenditure as a percentage of GDP).
- Urban Population as a proxy for urbanization (specifically Urban population as a percentage of total population).
- Maternal Mortality Rate as a proxy for health infrastructure availability and health awareness in society (specifically Maternal mortality ratio as modeled estimate, per 100,000 live births).

After clearing the database of countries with data unavailable for any of the variables for the year 2018, we obtained a database with 166 countries and groups of countries, and 8 independent variables along with health expenditure. We used this cross-sectional data for the regression analysis.

Methodology

We run two linear models to gauge the determinants of health expenditure. The first model is described as below:

$$Log(Health\ Expenditure) = \beta^{\circ} + \beta^{1} Log(\text{GDP per capita}) + \beta^{2}\ Refugee\ Population + \beta^{3} Urban\ Population + \beta^{4} Female\ Labour\ Force\ Participation\ + \beta^{5} Defense\ Expenditure + \beta^{6}\ Maternal\ Mortality\ Rate + \beta^{7}\ \text{Life expectancy} + \text{u}$$

Model 2 includes a categorical variable on the income groups of countries as defined by the World Bank based on the GNI:

1. Low-income: \$1,085 or less
2. Lower middle-income: \$1,086 and \$4,255;
3. Upper middle-income: \$4,256 and \$13,205;
4. High-income economies: \$13,205 or more.

$$Log(Health\ Expenditure) = \beta^{\circ} + \beta^{1} Refugee\ Population + \beta^{2} Urban\ Population\ + \beta^{3} Female\ Labour\ Force\ Participation\ + \beta^{4}\text{Life expectancy} + \beta^{5} Defense\ Expenditure + \beta^{6}\ Maternal\ Mortality\ Rate + \beta^{7}\text{Income group} + \text{u}$$

This report present the results of the cross-sectional regression analysis, including estimates of the coefficients, standard errors, t-values, and p-values. This is followed by the interpret the coefficients and explain how they relate to the research question. The overall fit of the model and statistical tests used to evaluate the goodness of fit are also highlighted.

Results

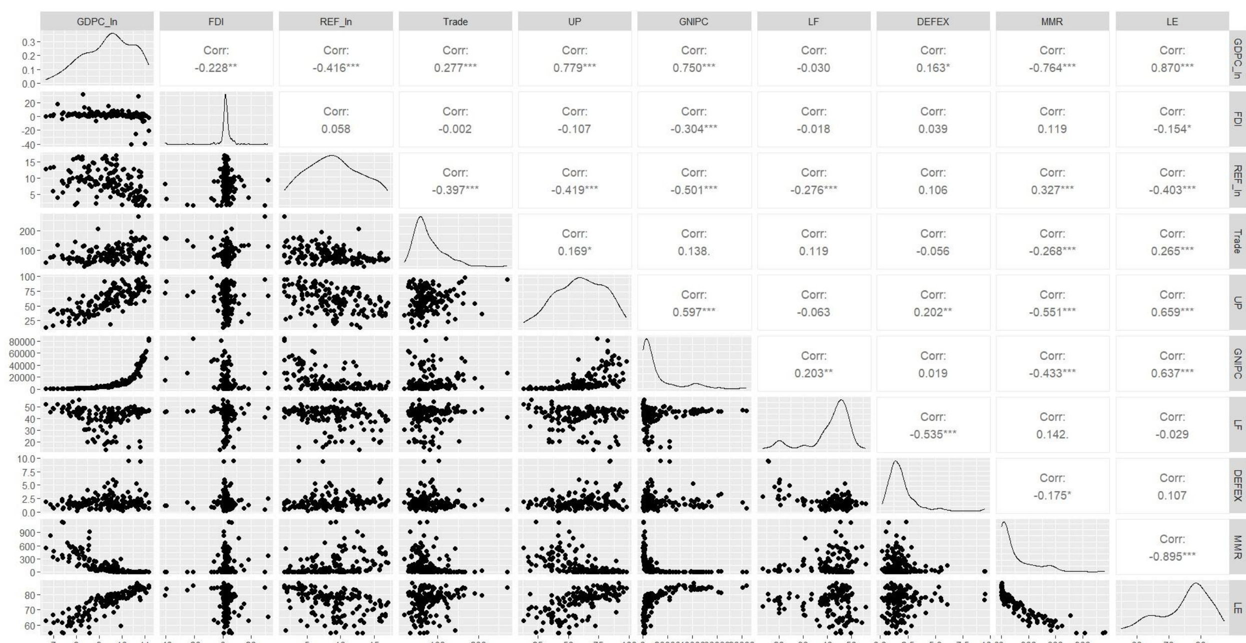
Variable Name	Min.	1st Quartile	Median	Mean	3rd Quartile	Max
HEX_ln	3.425	5.569	6.672	6.643	7.571	9.271
GDPC_ln	6.659	8.569	9.518	9.397	10.289	11.177
FDI_ln	-40.414	1.241	1.966	2.012	3.232	31.921
REF_ln	1.609	6.059	9.004	9.044	12.401	17.07
Trade	17.93	53.28	66.5	77.79	93.33	269.95
UP	13.03	43.63	59.6	59.76	75.78	98
GNIPC	280	1902	5390	12431	14180	8370
LF	13.27	39.21	44.22	41.36	46.54	55.96
DEFEX	0.1501	1.1696	1.6237	1.892	2.206	9.543
MMR	2	13	60	162.2	238.5	1140
LE	54.99	69.95	77.64	74.89	80.16	87.32
RefP	5	428	8136	1335016	243272	25905096

Summary Statistics of Variables:

The above table possesses the summary statistics of all variables in the dataset.

Variable Graph Matrix

Below is a graph matrix for the correlation between all independent variables for our model.



1. **Country:**
   This variable is an indicator of the number of observations in the dataset. There are 166 observations in our dataset. Its a character type variable, and can be interpreted as our index variable.

2. **factor(IG):**
We add a categorical variable on the basis of GNI-Per Capita according to the following conditions. The variable takes the values ranging from 0 to 4 according to the value of GNI-Per Capita. This variable is added in order to split the data set into multiple categories and obtain potential information with respect to control and treatment effects. This splitting gives us an idea of potential health expenditure based on how economically well off a country is in terms of its GNI-Per capita.

```
IG = 0
IG[GNIPC <= 1085] = 1
IG[GNIPC >= 1086 & GNIPC <= 4255 ] = 2
IG[GNIPC >= 4256 & GNIPC <= 13205] = 3
IG[GNIPC > 13205] = 4
```

Regression Results:

Model 1:

```
Call:
lm(formula = HEX_ln ~ GDPC_ln + RefP + UP + LF + DEFEX + MMR +
    LE, data = ModelData)
```

We use the above Linear Regression model, where our dependent variable is HEX_ln. Our independent variables are as listed below:

- GDPC_ln
- RefP
- UP
- LF
- DEFEX
- MMR
- LE

Upon fitting the above model, we obtain the following results:

Coefficients

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.576e+00  6.206e-01 -10.596  < 2e-16 ***
GDPC_ln      9.310e-01  5.747e-02  16.201  < 2e-16 ***
RefP         1.023e-08  6.709e-09   1.525 0.129320
UP           2.535e-03  2.112e-03   1.200 0.231746
LF           6.775e-03  3.485e-03   1.944 0.053643 .
DEFEX        4.287e-02  2.255e-02   1.901 0.059174 .
MMR          1.018e-03  2.661e-04   3.824 0.000188 ***
LE           5.045e-02  9.590e-03   5.261 4.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Intercept: An intercept value of -6.576 indicates that when all other variables are zero, there is a net negative value of ln(Health Expenditure) by the government.
GDPC_ln: A coefficient of 9.310e-03 indicates that for every one percent increase in GDP per capita, PPP (in current US dollars), there is a 0.009310 percent increase in health expenditure.
RefP: For every one unit rise in refugee population, there is a 1.023x10^-8 percent increase in (health expenditure keeping all other variables constant. The very small value of this coefficient can be attributed to the fact that the units of Refugee population are linear and in absolute numbers.
UP: For every one unit rise in urban population, there is a 0.00254 percent increase in health expenditure.
LF: For every one percent increase in total female labor force participation rate, there is a 0.0068 percent increase in ln(health expenditure) keeping all other variables constant.
DEFEX: For a one unit increase in defence expenditure, there is a 0.0438% increase in health expenditure, this can be attributed to the fact that governments spend similar amount in the respective sectors when there is a general increase in income or ability tro spend.
MMR: For every one unit increase in maternal mortality rate, there is a 0.00102 percent increase in health expenditure, this can be attributed to the fact that governments wish to provide better healthcare facilities in all cases and wish to prevent adverse cases of mortality from happening.
LE: For every one unit increase in life expectancy, there is a 0.0517 percent increase in health expenditure. This can be interpreted from the fact that for every individual in the country, if there is a one unit increase in life expectancy, then, there are overall better healthcare and sanitation facilities in the country and this would imply an increase in health expenditure spending.

Test of Significance: On observing the P-values for the different independent variables in our model 1, we see that Intercept, GDPC_ln, MMR, LE are significant at 100% level i.e we reject the null hypothesis that these variables are individually insignificant.
LF and DEFEX are significant at the 90% level and we can reject the null hypothesis that this variable is individually insignificant at this level.

Model 2:

```
Call:
lm(formula = HEX_ln ~ RefP + UP + LF + DEFEX + MMR + LE + factor(IG),
    data = ModelData)
```

We use the above Linear Regression model, where our dependent variable is HEX_ln. Our independent variables are as listed below:

- RefP
- UP
- LF
- DEFEX
- MMR
- LE
- factor(IG)2
- factor(IG)3
- factor(IG)4

Upon fitting the above model, we obtain the following results:

Coefficient:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.288e-01  9.849e-01   0.334 0.738939    
RefP        1.465e-08  8.688e-09   1.687 0.093663 .  
UP          9.162e-03  2.602e-03   3.521 0.000564 ***
LF          1.356e-03  4.696e-03   0.289 0.773124    
DEFEX       2.439e-02  2.950e-02   0.827 0.409691    
MMR         1.893e-04  3.686e-04   0.514 0.608302    
LE          6.208e-02  1.250e-02   4.967 1.77e-06 ***
factor(IG)2 4.346e-01  1.303e-01   3.335 0.001065 ** 
factor(IG)3 1.054e+00  1.667e-01   6.323 2.58e-09 ***
factor(IG)4 1.915e+00  2.042e-01   9.377  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Intercept: While keeping all other variables zero, Health Expenditure is expected to be: $e^{0.3288}$ i.e 1.38$ per capita.
RefP: A coefficient of 1.465e-08 indicates that for every one percent increase in RefP there is a 1.47*10^-8 percent increase in health expenditure. The very small value of this coefficient can be attributed to the fact that the units of refugee population are linear and in absolute be attributed to the fact that the units of refugee population are linear and in absolute numbers.
UP: For every one unit rise in Urban Populaiton, there is a 0.0092 percent increase in health expenditure keeping all other variables constant.
UP: For every one unit rise in urban population, there is a 0.00136 percent increase in health expenditure.
DEFEX: For a one unit increase in defence expenditure, there is a 0.0247% increase in health expenditure, this can be attributed to the fact that governments spend similar amount in the respective sectors when there is a general increase in income or ability to spend.

MMR: For every one unit increase in maternal mortality rate, there is a 0.000189 percent increase in health expenditure, this can be attributed to the fact that governments wish to provide better healthcare facilities in all cases and wish to prevent adverse cases of mortality from happening.
LE: For every one unit increase in life expectancy, there is a 0.064 percent increase in health expenditure. This can be interpreted from the fact that for every individual in the country, if there is a one unit increase in life expectancy, then, there are overall better healthcare and sanitation facilities in the country and this would imply an increase in health expenditure spending.
Factor(IG)2: Compared to the countries in the category IG1 i.e GNIPC < 1085, there is a 0.544% increase in the health expenditure per capita per year.
Factor(IG)3: Compared to the countries in the category IG1 I.e GNIPC < 1085, there is a 1.87% increase in the health expenditure per capita per year.
Factor(IG)4: Compared to the countries in the category IG1 I.e GNIPC < 1085, there is a 5.79% increase in the health expenditure per capita per year.

Test of Significance: On observing the P-values for the different independent variables in our model 2, we see that UP, LE, factor(IG)3, factor(IG)4 are significant at 100% level i.e we reject the null hypothesis that these variables are individually insignificant.

factor(IG)2 is significant at the 99% level and we can reject the null hypothesis at this level.
RefP is significant at the 90% level and we can reject the null hypothesis that this variable is individually insignificant at this level.

Regression Diagnostics

Model 1:

```
Model1 <- lm(HEX_ln ~ GDPC_ln + RefP + UP + LF + DEFEX + MMR + LE, data = ModelData)
summary(Model1)
```

<u>Breusch-Pagan Test for heteroskedasticity:</u>

The Breusch-Pagan determines whether or not the model's error variance is constant. If the estimated test statistic is higher than the chi-square distribution's critical value, there may be strong evidence of heteroscedasticity.

Null hypothesis:
Ho : Homoskedasticity in data

H1: Heteroskedasticity in data
Here, we can see that the P value is 0.001308 which is < 0.05, so we reject the null hypothesis, i.e our data suffers from heteroskedasticity ; $E(u/x) \neq 0$.
Due to heteroskedasticity the estimator no longer remains BLUE, as variance increases the predictions become inefficient.
Also, we can not claim a causation between health expenditure and the independent variables.

<u>Testing for Joint significance - F Test or ‘anova’ test:</u>

```
anova(Model1, Model1r)
```

F test is a test for multiple restrictions simultaneously, and hence is called Joint Hypothesis

```
> anova(Model1, Model1r)
Analysis of Variance Table

Model 1: HEX_ln ~ GDPC_ln + RefP + UP + LF + DEFEX + MMR + LE
Model 2: HEX_ln ~ 1
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    158  16.434
2    165 263.625 -7   -247.19 339.51 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Null Hypothesis H0 : $\beta1, \beta2, \beta3 \ldots \beta7 = 0$
Alternate Hypothesis: : $\beta1, \beta2, \beta3 \ldots \beta7 \neq 0$
Here we get the P value to be 2.2 x e-16 which is clearly less than 0.05, as well as the F statistic turns out to be 339.51 which is pretty large and so we can conclude that the null hypothesis is rejected and the independent variables are jointly significant.

```
> bptest(Model1)

        studentized Breusch-Pagan test

data:  Model1
BP = 23.658, df = 7, p-value = 0.001308
```

<u>Testing for Multicollinearity – Variance Inflation Factor:</u>

```
vif(Model1)
```

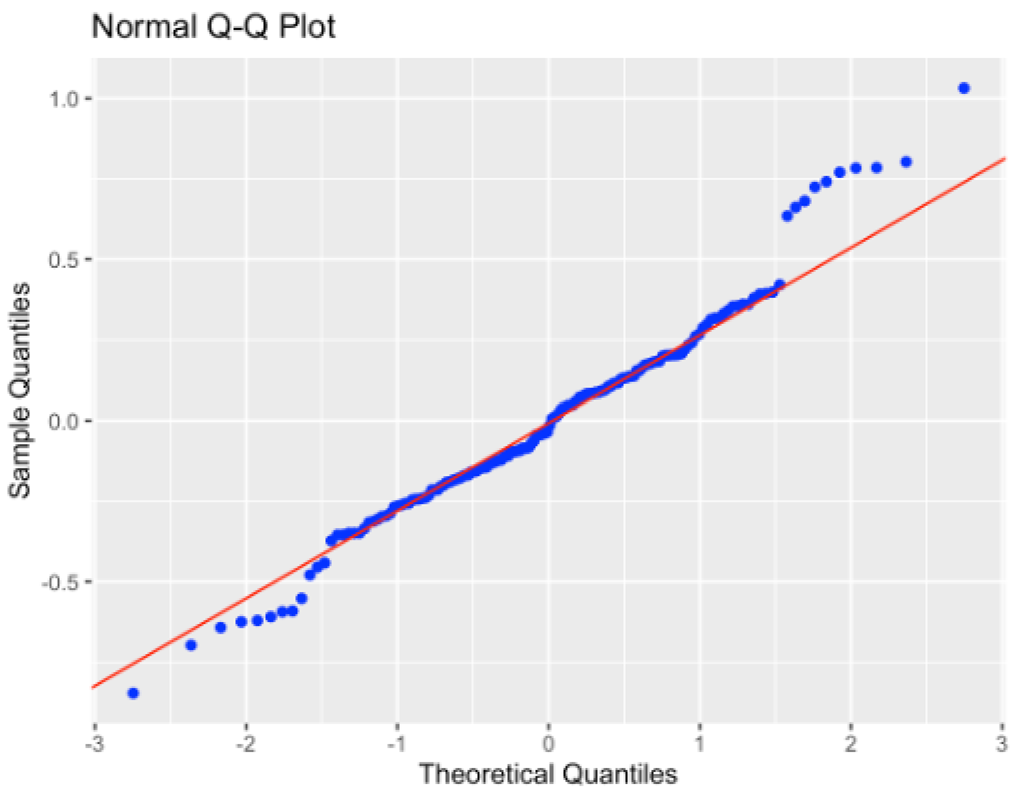
```
> vif(Model1)
 GDPC_ln     RefP       UP       LF    DEFEX      MMR       LE 
6.066290 1.124387 2.731210 1.520741 1.476889 5.598490 9.057465 
```

Here all the variables have a VIF <10, implying there is no multicollinearity.

<u>Testing for Normality : Data and Residuals</u>

1. *Q-Q plot – Normality of residuals*

   The univariate normality of the dataset is indicated by points on the Normal QQ plot. If the distribution of the data is correct, the points will be on the 45-degree reference line. If the data is not distributed properly, the points will deviate from the reference line.



   As most of the observations lie on the 45 ° reference line mentioned above, we can say that the residuals are normally distributed.

2. *Shapiro Wilk test – Normality of sample data*

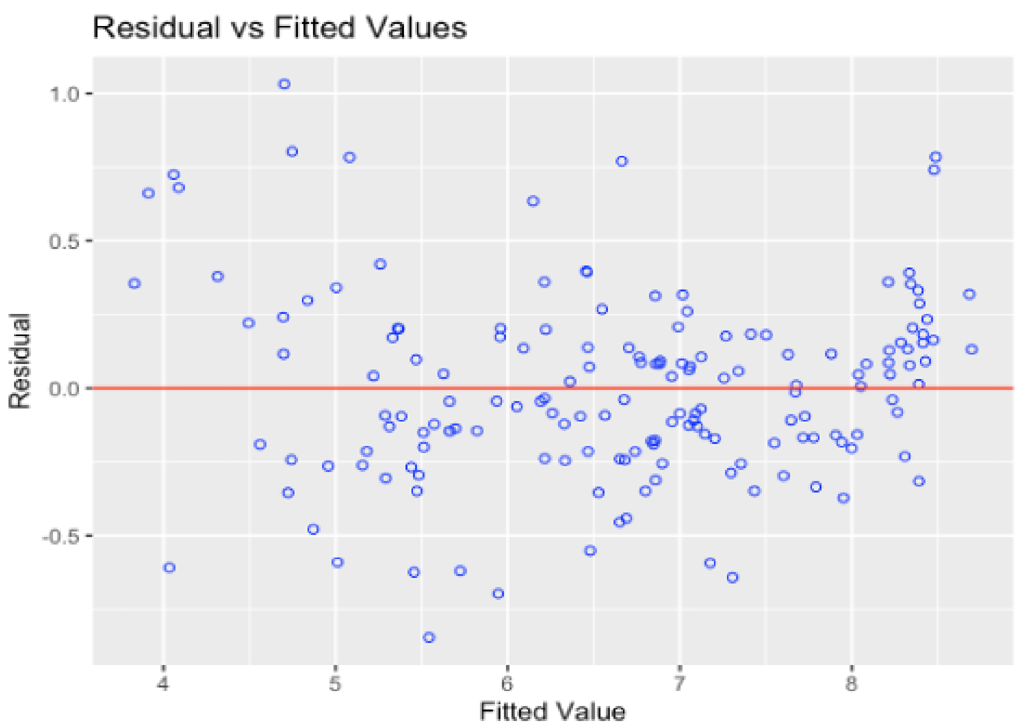
This test checks how close the sample data is to a normal distribution,
Null Hypothesis-
H0 : Data is Normally distributed
H1 : Data is not Normally distributed

Test	Statistic	pvalue
Shapiro-Wilk	0.9766	0.0066
Kolmogorov-Smirnov	0.069	0.4088
Cramer-von Mises	29.3603	0.0000
Anderson-Darling	1.1011	0.0067

We can see that the p value for the SWILK test turns out to be 0.066 which is < 0.05, and hence we reject the null hypothesis.

3. *Residual vs Fitted – Normality of residuals:*

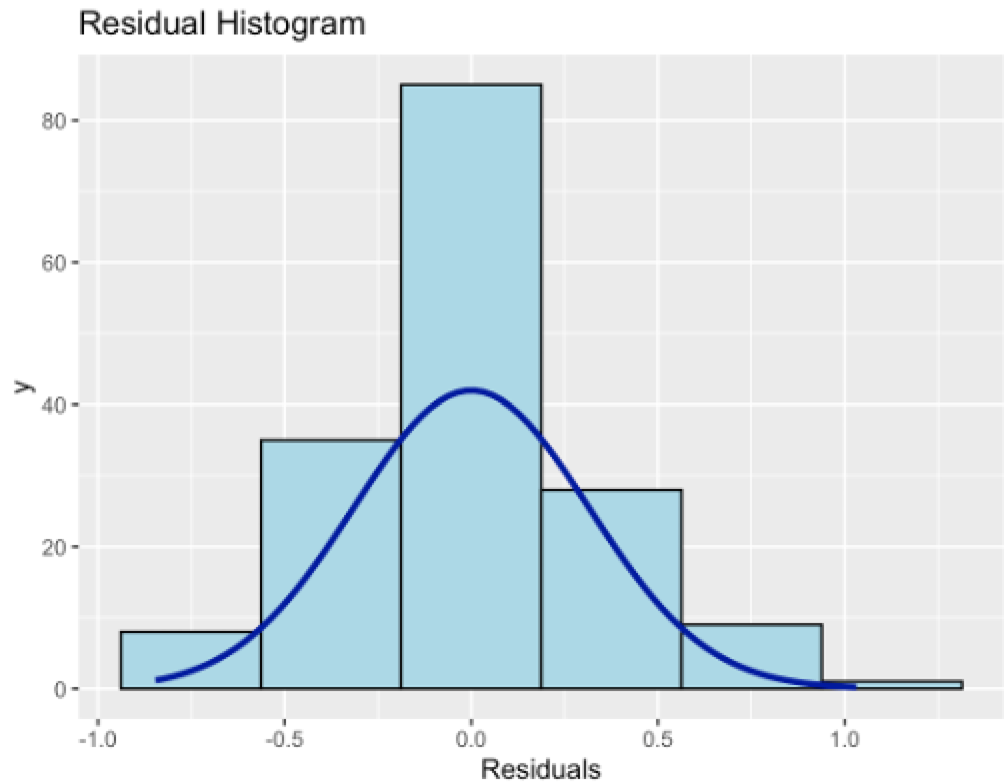
```
ols_plot_resid_fit(Model1)
```



We observe a very random distribution of residuals along the fitted values, thus implying that the value of the residuals are independent of the fitted value.

$E(\hat{\beta}) = \beta$

4. *Residual Histogram – Normality of residuals:*



Clearly, the distribution of residuals is normal.

<u>Testing for alternative functional forms: Resettest (Ramsey test)</u>

```
resettest(Model1, power = 2:3, type = "regressor")
```

It determines if the fitted values can be used in non-linear ways to further explain the dependent variable.

```
	RESET test

data:  Model1
RESET = 5.4557, df1 = 14, df2 = 144, p-value = 2.588e-08
```

Null Hypothesis :
Ho : Model is linear in mentioned variables
H1 : Model is non-linear in mentioned variables
We get a very small p- value of 2.58 x e-8, and hence we reject the null hypothesis, implying that a model with square or cube terms will end up in a better fit.

MODEL 2 : with income group dummy and dropping GDPC

```
Model2 <- lm(HEX_ln ~ RefP + UP + LF + DEFEX + MMR + LE + factor(IG), data = ModelData)
summary(Model2)
```

Breusch Pagan Test for heteroskedasticity:

```
> bptest(Model2)

	studentized Breusch-Pagan test

data:  Model2
BP = 14.15, df = 9, p-value = 0.1171
```

We get a P value of 0.1171, and hence fail to reject the null hypothesis i.e. there exists homoskedasticity in data.

Testing for Multicollinearity – Variance Inflation Factor

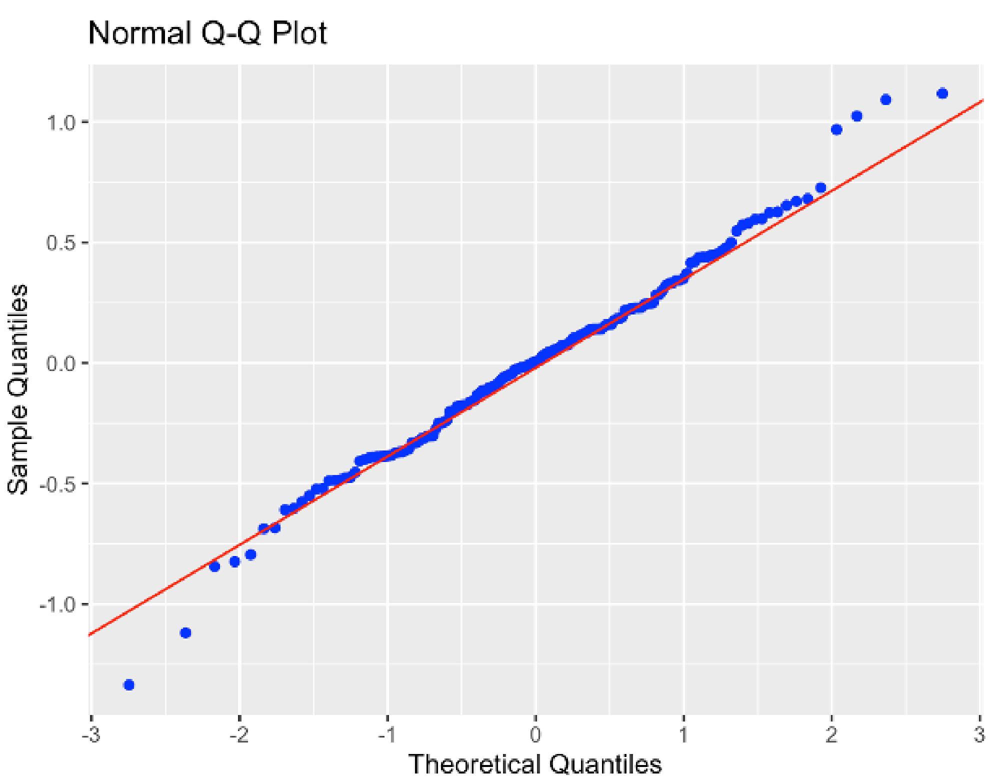
```
> vif(Model2)
                  GVIF Df GVIF^(1/(2*Df))
RefP          1.183620  1        1.087943
UP            2.603502  1        1.613537
LF            1.733381  1        1.316579
DEFEX         1.585952  1        1.259346
MMR           6.741468  1        2.596434
LE            9.658618  1        3.107832
factor(IG)    6.241313  3        1.356894
```

As the VIF in <10 for each case, we can say that there is no multicollinearity between the independent variables.

Testing for Normality : Data and Residuals

1. *Q-Q plot – Normality of residuals:*

   Since most of the observations lie on the 45° reference line, the distribution of residuals is normal.

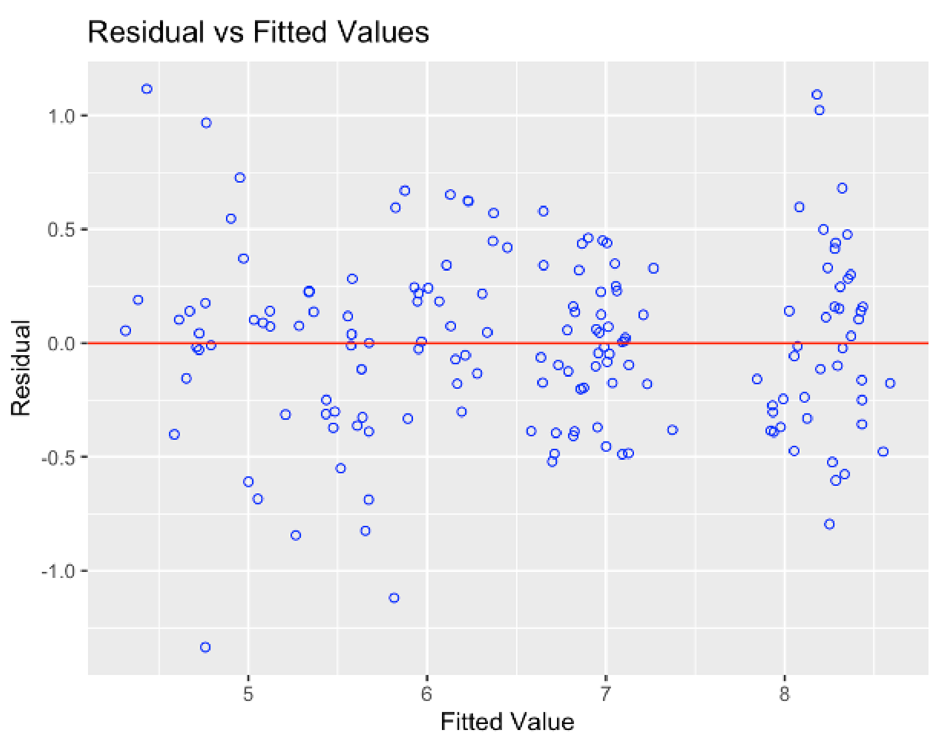


2. Shapiro-Wilk Test – Normality of sample data

Test	Statistic	pvalue
Shapiro-Wilk	0.9896	0.2615
Kolmogorov-Smirnov	0.053	0.7400
Cramer-von Mises	24.9324	0.0000
Anderson-Darling	0.4103	0.3394

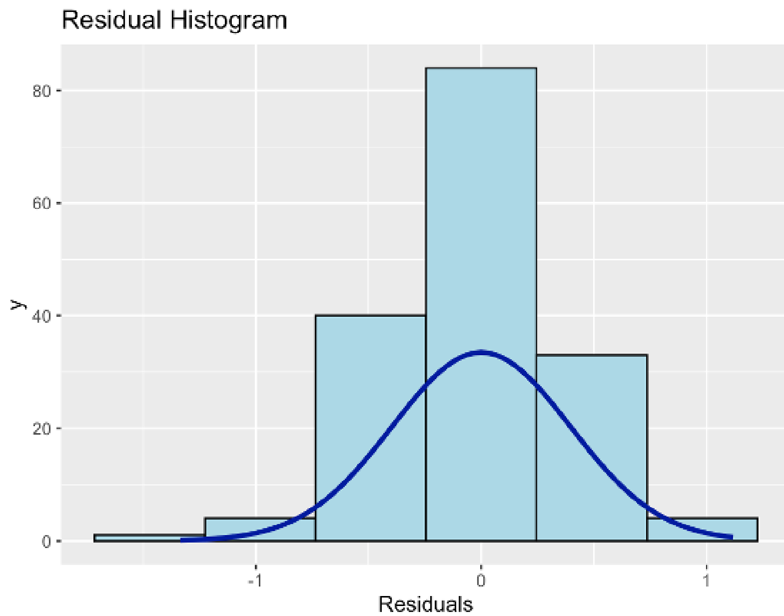
Here the p value for the SWILK test comes out to be 0.2615 (>0.05), so we accept the null hypothesis i.e. the sample data is now normally distributed.

3. Residual vs Fitted – Normality of residuals :



We observe a very random distribution of residuals along the fitted values, thus implying that the value of the residuals are independent of the fitted value

4. Residual Histogram– Normality of residuals:



Clearly, the distribution of residuals is normal.

Conclusion

In model 1: Intercept, GDPC_ln, MMR, LE are significant at 100% level and only these variables majorly impact the health expenditure of a country.
The other variables,LF and DEFEX are significant at the 90% level and we fail to prove their insignificance 10% of the times.
In model 2: UP, LE, factor(IG)3, factor(IG)4 are significant at 100% level. UP and LE are significant at 100%level and impact the health expenditure of a country.
Compared to the reference income group level i.e 1, the values of health expenditure of income group 3 and 4 are significantly different from the reference level 1 at 100% level.
factor(IG)2 is significant at the 99% level and it is gnificantly different from the reference level 1 at 99% level.
RefP is significant at the 90% level and we fail to prove their insignificance 10% of the times.

There have been several improvements in moving from Model 1 to Model 2, namely Homoskedasticity, Normality of sample data, Multicollinearity, normality of residuals.
The distribution of the residuals in Model 2 has lesser outliers and this distribution follows a normal distribution more closely than model 1,
Moving on to the reduction in multicollinearity**,** We see the value of variance inflation factor has decreased significantly for almost all variables.
For Homoscedasticity, Clearly there has been an increase in the p-value proving model 2 to be homoskedastic, which allows us to establish a causal relationship between the dependent and independent variables.
Moving on to Normality of sample data; Normality of sample data is an important assumption for performing t-test and ANOVA, as absence of normality may lead these tests to produce inaccurate results.

Observing the increase in the p-value for the SWILK test we fail to reject the null hypothesis, implying that the sample data is normally distributed after dropping GDPC and including income group dummies in the model.

References and Citations:

- https://rpubs.com/iabrady/residual-analysis
- https://stats.stackexchange.com/
- Zasuwa, G., Marek, A., Wesolowski, G., & Niewiadoma, J. (2020, November 1). Consumer Activism in Times of Economic Crisis and Recovery: A Cross-Country aAnalysis of the Role of Social Capital in Boycotting Products. *EUROPEAN RESEARCH STUDIES JOURNAL*, *XXIII*(Special Issue 2), 928–938. https://doi.org/10.35808/ersj/1907
- Hansen, P. & King, A., 1996. The determinants of health care expenditure: a cointegration approach. *Journal of Health Economics*, 15(1), pp.127-137.
- https://www.kaggle.com/datasets/ploverbrown/world-bank-indicators-collec