**ECO 321 - Econometrics**

**Project 1 - Linear Regression with Multiple Regressors**

**Due Date: 10/30/2020**

**Group 8**

**Group Member Names:**

Emirhan Akkaya (SBU ID#: 112768575)

Parv Joshi (SBU ID#: 112169570)

Shail Shah (SBU ID#: 112315115)

**Research Question:**

Our research project questions the dependence of factors like a country’s standard of living, military expenditure, government savings, and government debt on a nation’s healthcare expenditure in standard years (non-recession years).

**Our Motivation:**

Our research was motivated by a range of different factors that mainly came about by our group members’ experiences and other current-day factors such as the Covid-19 pandemic. All of these things combined made us think of a topic in health economics. Firstly, some of our group members have been to many countries like India, Turkey, and the United Arab Emirates. The differences in standard of living and the quality of healthcare available is a stark difference from what we see here in the United States or other European countries.

Something else that intrigued us was that there are countries that provide healthcare free of charge, and it got us wondering what factors were in play here that allowed a country to do that. For example, a country like Germany that provides free healthcare has higher tax rates than other countries. Still, on the contrary, a country like Norway that has one of the best healthcare systems in the world has comparatively lower tax rates. To add to that, with the recent coronavirus pandemic, many countries will increase their government debt; thus, we were also looking to see how much of an effect government debt has on healthcare, if at all, and if these effects would be long-term or short-term.

**Review of Key Studies closely related to our Research Question**

While looking at research done by others on similar topics, we came across many sources. The foremost one was a book named *The economics of public health care reform in advanced and emerging economies*, written in 2012 by Clements, B., Coady, D., and Gupta, S. This book talks about how countries in the past have spent about half of their GDP on healthcare and why it will diversify in the future; especially when countries will try to reduce their government debt ratios (Clements et al., 2012, pp. 1–3).

The second study we came across was a research paper in 2005 named *Who’s Going Broke? Comparing Growth in Healthcare Costs in Ten OECD Countries* written by Hagist, C., and Kotlikoff, L. They focused on predicting how demographic factors affect healthcare spending (Hagist & Kotlikoff, 2005).

The last resource we referred to was a 2012 article written by Heijink R, Koolman X, and Westert GP, named *Spending more money, saving more lives? The relationship between avoidable mortality and healthcare spending in 14 countries* also used OLS estimators to determine the relationship of government healthcare spending on four variables. Their independent variables were time trend, age and mortality rates; education and unemployment rates; and alcohol and tobacco consumption. Their regression results on page 534 were similar to what we learned in class and were understandable and intuitive. We used variables of which good and enough data could be found and referred to this for our analysis.

The citations for the three sources were:

Christian Hagist & Laurence J. Kotlikoff, 2009. “Who’s Going Broke? Comparing Growth in Healthcare Costs in Ten OECD Countries,” Hacienda PÃºblica EspaÃ±ola, IEF, vol. 188(1), pages 55-72, March.

Clements BJ, Coady D, Gupta S. *The Economics of Public Health Care Reform in Advanced and Emerging Economies*. International Monetary Fund; 2012.

Heijink R, Koolman X, Westert GP. Spending more money, saving more lives? The relationship between avoidable mortality and healthcare spending in 14 countries. *The European journal of health economics*. 2012;14(3):527-538. doi:10.1007/s10198-012-0398-3

**Brief Description of the data that was used:**

Our group used data that was listed on the World Bank and IMF websites. To organize the data, we selected 2017 as our observational year and considered 127 countries when analyzing the data. The data is collected from credible sources of economic data such as the World Bank and IMF database. We mainly used the USD as our unit and made sure to specify if any other unit was used for the data. As for our independent variables that we had used, we considered GDP per capita (GDPpc) for our data because we wanted to understand better each country’s average person is earning and how that might affect each country’s healthcare spending per resident. We also considered Military Expenditure (ME) to see if there is a correlation where higher military expenditure resulted in lower healthcare spending or vice-versa. Government savings (GS) was used to observe if higher savings resulted in higher healthcare expenditure. Finally, we used the Central Government Debt (CGD) to see if debt negatively affected the amount of money allocated for healthcare. Our dependent variable was Government Healthcare Spending (GHS), where we measured a country’s healthcare expenditure.

**Exploratory Data Analysis**

**Summary of Data:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Government Healthcare Spending (GHS)** | **GDP per Capita (GDPpc)** | **Military Expenditure (ME)** | **Government Savings (GS)** | **Central Government Debt (CGD)** |
| **Minimum** | 19.43 | 773.6 | 0.000e+00 | -1.836e+10 | 2.826 |
| **1st Quartile** | 81.01 | 5021.7 | 1.998e+00 | 6.673e+08 | 32.374 |
| **Median** | 381.11 | 13788.5 | 1.015e+09 | 4.230e+09 | 46.673 |
| **Mean** | 1277.14 | 21852.9 | 1.121e.10 | 3.401e+10 | 54.060 |
| **3rd Quartile** | 1353.30 | 33821.1 | 5.084e+09 | 2.626e+10 | 66.190 |
| **Max** | 10246.14 | 112822.6 | 6.060e+11 | 5.490e+11 | 196.771 |

We notice that all the five variables summarized here are continuous ones and decimal values for the statistics make sense. For the 127 countries, the minimum a country spends on healthcare spending is $19.43 (unit is current USD per capita), and the maximum is $10,246.14, with a mean of 1,277.14. Moreover, the median is only $381.11. This implies that most countries spend less than average for healthcare, making the GHS data right-skewed.

A similar trend goes for GDP per capita, Military Expenditure, and Central Government Debt. They are right-skewed too. However, the data for Military Expenditure has a very high range: Range = max-min = 6.060e+11 - 0.000e+00 = 6.060e+11 = $606000000000 (in current USD). This high range signifies that there is no specific pattern in the military expenditure observations, some have zero expenditure, and some have in billions.

For GDP per capita, we again have the values of the majority of the countries are small, since the mean is $21852.9 (current USD), and the maximum is $112822.6, which is a lot more. The central government debt is measured in a unit of a percentage of GDP. We have countries whose debt per GDP is 2.826% to 196.771%. Most of them have a low debt to GDP ratio since the average is only about 54.060%.

Government Savings is the only variable that is not entirely right-skewed since it has a higher median than the mean. Here, it is interesting to notice is that the minimum of Government Savings is negative. This means that our data consists of countries with a budget deficit (more spending than saving) in 2017.

**Correlation Matrix for Data:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **GHS** | **GDPpc** | **ME** | **GS** | **CGD** |
| **GHS** | 1.00000 | 0.80501 | 0.44644 | 0.40717 | 0.05483 |
| **GDPpc** | 0.80501 | 1.00000 | 0.22619 | 0.28651 | 0.03142 |
| **ME** | 0.44644 | 0.22619 | 1.00000 | 0.66077 | 0.10629 |
| **GS** | 0.40717 | 0.28651 | 0.66077 | 1.00000 | 0.05202 |
| **CGD** | 0.05483 | 0.03142 | 0.10629 | 0.05202 | 1.00000 |

We notice that we have a very significant correlation of 0.80501 between Government Healthcare Spending and GDP per capita. The correlation between Government Healthcare Spending with Military Expenditure and Government Savings 0.44644 and 0.40717, respectively. We can say that these are moderately correlated, and it is questionable if the correlations are significant or not. When it comes to Central Government Debt, there is a very low correlation of 0.05483. Here, we can predict that the Central Government Debt might not affect Government Healthcare Spending; however, it may not be the case.

When we try to check the correlation between our independent variables, we observe that our variables mostly have low correlations (from 0.03142 to 0.28651) except for one case: correlation between Military Expenditure and Government Savings. They have a correlation of 0.66077, which is high and could be significant. This means that there could have been an Omitted Variable Bias if we had not considered both variables. Still, we cannot say for sure since we do not know yet if both variables affect our Government Healthcare Spending or not.

Overall, the correlation matrix gives us an idea that Government Healthcare Spending mostly will depend on GDP per capita, may or may not depend on Military Expenditure and Government Savings, and probably does not depend on Central Government Debt.

**Histograms of the Variables:**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

**Figure 1: Histograms of the Variables**

The histogram for the government healthcare spending per capita is skewed right, and there is an abundance of countries that spend less than 1000 dollars (USD). The two countries that spent the most money on healthcare spending were Switzerland and the United States. Most countries spent less than 1000 dollars, most likely due to smaller gross domestic product and smaller economies. Also, each country’s preferences on what they see as important also contribute to how much each country spends on healthcare.

The histogram for GDP per capita is also roughly skewed right, and for this histogram, most countries’ GDP per capita is less than 2000 dollars. The country with the highest GDP per capita in Luxembourg and the country with the smallest GDP per capita in Burundi.

The bar graph had virtually one bar for military expenditure except for one country between 6e+11 and 7e+11. This country was the United States, and we kept the U.S in our project because it was not an outlier for the other independent variables, and also since it is a significant country to keep in terms of the geopolitical influence it has on the world. The country that had the smallest military expenditure was Costa Rica. The unit for this histogram is also US dollars.

The histogram for government savings is the only histogram that has negative values in it. The negative values here would imply that some countries would be spending more money than saving. The shape of this histogram is not entirely skewed since some of the countries have negative values. If we were not to consider these negative values, then the histogram would most approximately be roughly right-skewed. The unit for this graph is also US Dollars. The country with the lowest government savings, which is negative, is Angola, and the country has the highest government savings in India.

Finally, our last histogram draws out government debt for each respective country. This histogram is also roughly skewed to the right. The unit is a percentage of GDP, so if a country has 100 for government debt, this implies that whatever the government produces in a given year is equivalent to its outstanding debt. Most countries have debt that is less than their total GDP, roughly between 25 and 50% of their GDP. The country with the highest government debt as a percentage of GDP is Japan, and Brunei Darussalam is the country with the lowest government debt as a percentage of GDP. As a percentage of GDP, high government debt may be explained by several factors, including the country’s median age, the country’s geopolitical state, and the banking industry.

**Scatterplots of Government Healthcare Spending with each of GDP per capita, Military Expenditure, Government Savings, and Central Government Debt:**

|  |  |
| --- | --- |
|  |  |
|  |  |

**Figure 2: Scatter Plots for variable GHS (Government Healthcare Spending) against other variables.**

Looking at the graphs, we notice that they look a lot different from each other. The first graph is Government Healthcare Spending versus GDP per capita. When we look at the points, we notice that it is not mostly linear. It looks either like an exponential or an x-squared graph. Since we are focusing on finding a linear model for this project, we can still approximate the line of best fit (the green regression line); however, the nonlinearity might increase our error term.

Our Government Healthcare Spending versus Military Expenditure graph looks mostly like a vertical line, except the one point far away from our other observations. After sorting our data, we realized it is the United States that acts as an outlier in this graph. The reason for that is that the U.S. is the only country with an exceptional military (and its spending) and spends very high on healthcare. However, we decided not to remove the U.S. observation from our data since it is not an outlier for any of our other variables, nor is it an anomaly. There is no definite pattern; no law that prevents countries from having high or low military expenditures. This means it is okay and not incorrect to have the United States in our data despite this. Moreover, the U.S. is an essential country. Here, the regression line changes drastically from the vertical line we would get if we excluded the U.S. This means that here, it acts like an influential variable, which may increase our error.

The Government Healthcare Spending versus Government Savings graph has much concentration of points in the bottom half, indicating that most countries have low savings and healthcare spendings. This is something we already predicted using our histograms and summary statistics. However, our Government Healthcare Spending versus Central Government Debt graph looks much scattered, especially along the horizontal axis. This means that it is difficult to predict the Government Healthcare Spending depending upon the country’s debt. Again, this is something we have already predicted.

We notice that none of the graphs look linear, and hence a non-linear model would be a better option. A linear model may not be perfect and may have a high error. However, since our focus for the project is the variables’ casualty and not finding the best possible model, we will find a linear model.

**Econometric Models we Plan to Use:**

For our project, we used the Ordinary Least Squares (OLS) Econometric model. The OLS estimates the parameters in our regression model and minimizes the sum of the squared residuals. The OLS also has three least squares (LSA) assumptions. These assumptions are E(u|X = x) = 0, which implies that for any given x, the mean of x is equal to zero. The second LSA assumption is that each distribution is independently and identically distributed. This means that each distribution is chosen from the same population or set of data and that the chosen entities are randomly selected from the distribution. Randomness is fundamental in statistical analysis because it allows us to assume no bias in the data we are working with. Finally, the third LSA assumption states that large outliers are very rare. This is important for an OLS fit because an outlier can very dramatically affect the distribution in the original data set or population. In our R code, we used the government healthcare as the dependent variable and all other independent variables as the regressors.

Also, for our project, we assume homoskedasticity but used heteroskedasticity because it was a better fit for our project since we had multiple dependent variables. Heteroskedasticity was a better fit for our data because heteroskedasticity takes into account unequal variances, and we had variances that were not equal to each other in our data.

**Summary of Results Obtained:**

At first, we try to regress our model with only one independent variable. We chose to regress Government Healthcare Spending on GDP per capita only. Here, we hypothesize that Government Healthcare Spending will be dependent on GDP per capita. We believe that the higher the income per person of a country, the better the medical service and equipment they can afford, which may be more expensive. Moreover, this will allow them to help more patients. Let us consider the following null and alternative hypothesis for the initial test:

When we perform simple regression using the linear model (lm) function in R, and assuming homoskedasticity (the default option), we get the following output:

> summary(single\_homo\_model)

Call:

lm(formula = GHS ~ GDPpc, data = data)

Residuals:

Min 1Q Median 3Q Max

-4271.2 -429.3 0.4 283.9 6042.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.010e+02 1.538e+02 -2.608 0.0102 \*

GDPpc 7.679e-02 5.062e-03 15.171 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1204 on 125 degrees of freedom

Multiple R-squared: 0.6481, Adjusted R-squared: 0.6452

F-statistic: 230.2 on 1 and 125 DF, p-value: < 2.2e-16

We see that its p-value is very low, which can be rejected even at a 1% level of significance. This means that Government Healthcare Spending depends on GDP per capita much. Moreover, it had an R-squared value of 0.6481. This means that this model explains only about 64.81% of the variance. Standard Error for GDPpc is 5.062e-03, which is small. The residual standard error for the model is 1204 (with 125 degrees of freedom.), which is high.

Let us correct it by assuming heteroskedasticity instead of homoskedasticity. The output with heteroskedasticity robust standard errors is:

> single\_model

t test of coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -401.032086 141.658804 -2.8310 0.00541 \*\*

GDPpc 0.076794 0.010337 7.4293 1.497e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Here, our p-value is still very small, and our null hypothesis can be rejected at a 1% level of significance. This means a strong causality.

However, let us now look at the predicted values of our Government Healthcare Spending using the mode with one variable and compare it with the summary statistics of the GHS variable. Here is the summary using the predicted model:

**Summary (Data$predicted\_single)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| -341.6 | -15.4 | 657.8 | 1277.1 | 2196.2 | 8263.1 |

Here are the summary statistics of the GHS variable:

**Summary (Data$GHS)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| 19.43 | 81.01 | 381.11 | 1277.14 | 1353.30 | 10246.14 |

If we notice, the mean is almost the same: 1277.1 in both cases. However, the minimums and maximums are very different. For the minimum, the model predicts -341.6, but it actually is 19.43. For the maximum, it predicts 8263.1, but it actually is 10246.14. This shows that we have a substantial error.

To reduce this error, let us consider a regression model with our four chosen independent variables. Our new model becomes:

Here our null hypothesis will be to check if each of the slopes will be zero or not. For that, our hypothesis to be tested is:

Assuming homoskedasticity (default), we get:

> summary(homo\_model)

Call:

lm(formula = GHS ~ GDPpc + ME + GS + CGD, data = data)

Residuals:

Min 1Q Median 3Q Max

-3842.8 -365.0 11.9 281.1 5532.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.961e+02 2.044e+02 -1.938 0.0549 .

GDPpc 7.051e-02 4.760e-03 14.813 < 2e-16 \*\*\*

ME 9.785e-09 2.352e-09 4.160 5.94e-05 \*\*\*

GS 4.517e-10 1.510e-09 0.299 0.7653

CGD 1.344e-01 2.813e+00 0.048 0.9620

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1084 on 122 degrees of freedom

Multiple R-squared: 0.7219, Adjusted R-squared: 0.7128

F-statistic: 79.18 on 4 and 122 DF, p-value: < 2.2e-16

Here we see that only the variables GDPpc and ME are statistically significant (at 1% level of significance) since their p-values are below 0.01. Nevertheless, the others are not even significant at the 10% significance level, which means they are insignificant. Here, the standard error for GDPpc is 4.760e-03, which is smaller than the previous model. This signifies that this model shows a better relationship between Government Healthcare spending and GDP per capita. We here have a multiple R-squared of 0.7219 (72.19% of the variance explained) and an Adjusted R-squared value of 0.7128 (71.28% of the variance explained). This signifies that this is a better model. Moreover, our residual standard error here is 1084 (on 122 degrees of freedom - since we added three variables), which is less than our previous one. Hence, we can conclude that this is a better linear model. Let us now adjust for heteroskedasticity. Using heteroskedasticity robust standard errors, we get the following output:

> model

t test of coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.9607e+02 2.0164e+02 -1.9643 0.05177 .

GDPpc 7.0514e-02 9.5285e-03 7.4003 1.913e-11 \*\*\*

ME 9.7850e-09 1.6708e-09 5.8564 4.097e-08 \*\*\*

GS 4.5171e-10 1.4147e-09 0.3193 0.75004

CGD 1.3435e-01 3.2206e+00 0.0417 0.96679

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Here too, we find that GDPpc and ME are the only significant explanatory variables. We notice that our standard errors have changed, and so has our p-values. Hence this is a better result since it does not assume that the errors have the same variance. From this, we can find our final model:

OR

Here, we can conclude that when GDP per capita or Military Expenditure increases, Government Healthcare Spending increases too since their coefficients are positive. However, Military Expenditure affects Government Healthcare Spending much less than GDP per capita since its coefficient is much smaller. Also, our intercept is -3.9607e+02, which can be written approximately as -396.07. This means that when a country has zero GDP per capita and zero Military Expenditure, its Government Healthcare Spending would be -396.07. This could mean that the country, in that case, borrows about $396.07 from other countries in order to meet its healthcare needs. Also, since Government Spending has an insignificant coefficient, we can say that there would have been no Omitted Variable Bias even when we did not consider both Military Expenditure and Government Savings initially.

Now, let us see how this changes the predicted errors. Here is the summary using the new predicted model with multiple variables:

**Summary (Data$predicted\_multiple)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| -334.06 | -23.68 | 621.77 | 1277.14 | 2094.84 | 10012.85 |

We notice that the summary statistics have changed from before. Let us compare this with our summary statistics of the GHS variable. Here is the table for that:

**Summary (Data$GHS)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| 19.43 | 81.01 | 381.11 | 1277.14 | 1353.30 | 10246.14 |

The mean now is precisely the same. The has been a small decrease in the difference between the predicted and observed minimum from our single-variable model. However, the difference between the predicted and observed maximum has reduced drastically. This means that our new model is indeed a better model for our predictions. However, since there is still a significant difference in predicted and observed values of the minimum, we can say that there must be a better model than this. However, from the data visualization and the interpretations we made of the summary statistics, we can say that such a better model is not likely to be linear. It may have non-linear elements. However, since our focus is mainly on finding causality using a linear model rather than finding the best model, our current model is good enough for this project.

Moreover, we have assumed that Government Healthcare Spending may affect only four variables. Part of the error could be because of Omitted Variable Bias (if any) due to variables we did not consider, such as Effective Average Tax Rates or the country’s political status (whether it is a democracy, republic, monarchy, etcetera). However, we cannot say if an Omitted Variable Bias exists or not.

**Conclusions:**

We conclude that there is causality (with a positive correlation) between our dependent variable and two of our variables, but no causality between the dependent variable and two of our other variables. The two variables with causality are GDP per capita and Military Expenditure, and the two without causality are Government Savings and Central Government Debt. This is in accordance with the predictions we made using data visualization that GDP per capita is likely to affect Government Healthcare Spending, it is unpredictable for Military Expenditure and Government Savings, and Central Government Debt most likely does not affect Government Healthcare Spending.

The overall linear model found using heteroskedasticity robust standard errors, and a level of significance of 1% is:

This model has a multiple R-squared of 0.7219 (it explains 72.19% of the variance) and an Adjusted R-squared value of 0.7128 (it explains 71.28% of the variance). It has a residual standard error of 1084 (on 122 degrees of freedom).

**References**

Christian Hagist & Laurence J. Kotlikoff, 2009. “Who’s Going Broke? Comparing Growth in Healthcare Costs in Ten OECD Countries,” Hacienda PÃºblica EspaÃ±ola, IEF, vol. 188(1), pages 55-72, March.

Clements BJ, Coady D, Gupta S. *The Economics of Public Health Care Reform in Advanced and Emerging Economies* . International Monetary Fund; 2012.

Heijink R, Koolman X, Westert GP. Spending more money, saving more lives? The relationship between avoidable mortality and healthcare spending in 14 countries. *The European journal of health economics*. 2012;14(3):527-538. doi:10.1007/s10198-012-0398-3

International Monetary Fund. (n.d.). *Central Government Debt - Percent GDP*. IMF Data. Retrieved October 12, 2020, from https://www.imf.org/external/datamapper/CG\_DEBT\_GDP@GDD/SWE

World Bank. (n.d.-a). *Adjusted savings: net national savings (current US$) | Data*. World Bank Data. Retrieved October 12, 2020, from https://data.worldbank.org/indicator/NY.ADJ.NNAT.CD

World Bank. (n.d.-b). *Current health expenditure per capita (current US$) - Afghanistan | Data*. World Bank Data. Retrieved October 12, 2020, from https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD?locations=AF

World Bank. (n.d.-c). *GDP per capita, PPP (current international $) - United States | Data*. World Bank Data. Retrieved October 12, 2020, from https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?locations=US

World Bank. (n.d.-d). *Military expenditure (current USD) - Afghanistan | Data*. World Bank Data. Retrieved October 12, 2020, from https://data.worldbank.org/indicator/MS.MIL.XPND.CD?locations=AF