Ruoheng Du

Regression and Multivariate Data Analysis

Prof. Jeff Simonoff

Feb 28, 2023

Relationship Between Health Expenditure and GDP Per Capita

Health expenditure is often considered a critical index of a country's overall healthcare system. In general, it includes spending on healthcare infrastructure, medical personnel, medical equipment, medications, and other health-related services. Given that it provides resources to ensure access to necessary medical care for the country's population, the proportion of health expenditure in a country's gross domestic product (GDP) is an important economic indicator of a country's commitment to its healthcare system and the well-being of its citizens (Xu et al. 204). Besides, as a result of the COVID-19 pandemic exposing weaknesses in healthcare systems, there is a heightened awareness of the need to prioritize health spending and examine the appropriate level of health expenditure to protect public health. Therefore, this report aims to investigate the factors that influence the amount of money spent by the government on health care and it will look specifically at the relationship between health expenditure and GDP per capita.

Broadly speaking, academics have explored that there could be a positive association between health expenditure and the economic indicator of per capita GDP (Raghupathi and Raghupathi 13). This relationship is described as complex and iterative. On the one hand, higher GDP per capita tends to be associated with higher health expenditure, as wealthier countries have more resources to devote to health advancement. On the other hand, a healthy population contributes to increased labor productivity and economic prosperity (Zon and Muysken 25). In

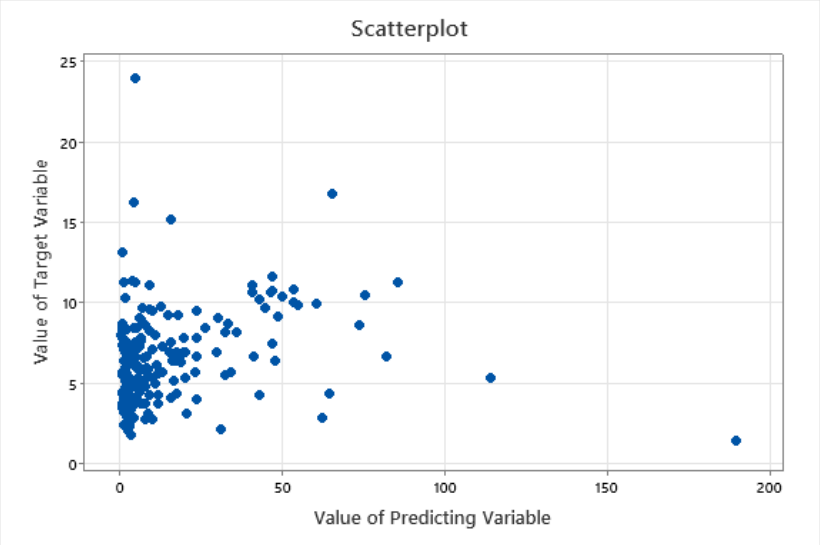
other words, it is reasonable to believe that there is a positive relationship between health expenditure and GDP per capita and it can be expressed by a simple linear regression model:

$$Health\ expenditure = \beta_o + \beta_1 * GDP\ per\ capita + random\ error$$

The following analysis is based on data from a sample of 187 countries in the health expenditure statistical table from UNdata (http://data.un.org/) collected by United Nations Statistics Division, a database listed on the GlobalEdge Statistics Data Sources (https://globaledge.msu.edu/global-resources/statistical-data-sources). There are three noteworthy aspects of the data in this report. Firstly, this report will use GDP per capita, or the total economic output divided by its population, as a representative of the country's economy. This is because that there are countries with different population sizes, which is an important factor in impacting the level of resources and services required to maintain a functional healthcare system, it would be biased to use GDP as the predicting variable. Furthermore, due to the relatively large absolute values of the GDP per capita data, all of the GDP per capita data in this report has been manually updated and will be expressed in thousands in order to study the above relationship. Besides, the GDP per capita discussed in this report is all expressed in US dollars. Secondly, health expenditure as a percentage of GDP is a better measure than the absolute amount of health expenditure. Using the latter can be deceptive because countries with greater GDPs and larger economies are more likely to have higher levels of health expenditure, which does not necessarily imply that they are investing a larger proportion of their resources in healthcare. Therefore, this report will use health expenditure (% of GDP) as the target variable and GDP per capita (in thousands) as the predicting variable in the following analysis. Thirdly, this report will use 2019 data on health expenditure and GDP to lessen the impact of COVID on each country's health expenditure and GDP per capita.

Here is the descriptive statistics and the scatterplot of the two variables.

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Health expenditure (% of GDP)	187	6.583	0.222	3.029	1.500	4.400	6.200	8.300	24.000
GDP per capita (in thousands)	187	15.43	1.71	23.35	0.29	2.12	6.18	17.99	189.51



The target variable, health expenditure (% of GDP), has a mean of 6.583 and a standard deviation of 3.029, which means that on average, countries spend 6.583% of GDP on health. The predicting variable, GDP per capita (in thousands), has a mean of 15.43 and a standard deviation of 23.35, which means that the average GDP per capita of all the listed countries is 1543 US dollars. In addition, as can be seen from the scatterplot, there does appear to be a positive linear relationship between health expenditure and GDP per capita, although there are a number of caveats. Therefore, I expect the model to find evidence to reject the null hypothesis that $\beta_1 = 0$ and accept the alternative hypothesis that $\beta_1 > 0$. And the below is the least squares regression with the two variables.

Regression Equation

Health expenditure (% of GDP) = 6.208 + 0.02429 GDP per capita (in thousands)

Coefficients

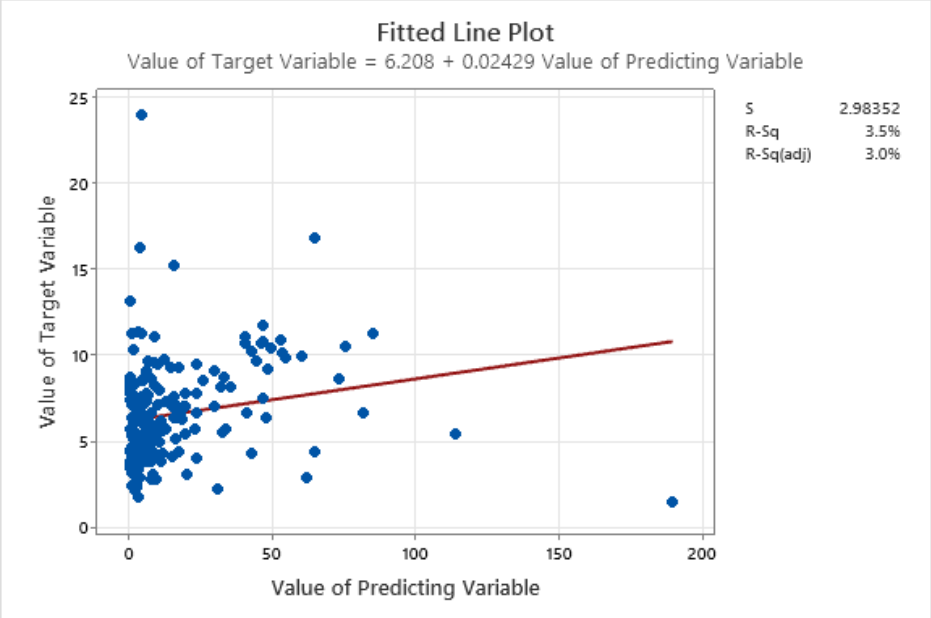
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.208	0.262	23.72	0.000	
Value of Predicting Variable	0.02429	0.00937	2.59	0.010	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.98352	3.50%	2.98%	0.00%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	59.81	59.810	6.72	0.010
Value of Predicting Variable	1	59.81	59.810	6.72	0.010
Error	185	1646.75	8.901		
Lack-of-Fit	184	1635.23	8.887	0.77	0.744
Pure Error	1	11.52	11.520		
Total	186	1706.57			



The regression is quite weak as the R-squared value is only 3.50%. In other words, GDP per capita (in thousands) fails to account for the observed variability in health expenditure (% of GDP) well. The intercept indicates that when the GDP per capita (in thousands) equals zero (i.e., the country has zero GDP), the estimated expected health expenditure (% of GDP) is 6.208 (that

is, the country is expected to spend 6.208% of its GDP on health-related services). But this point does not have any practical interpretation, since it is meaningless to discuss a country with zero GDP. The slope coefficient indicates that a one-thousand US dollars change in GDP per capita is associated with an estimated expected 0.02429 percentage point change in health expenditure (% of GDP). The standard error of the estimate of 2.98 says that this model could be used to predict health expenditure (% of GDP) within ±5.96 percentage points, roughly 95% of the time.

Both coefficients are statistically significant. As for the slope, with the null hypothesis of $\beta_1 = 0$ and the alternative hypothesis of $\beta_1 > 0$, the t–statistic for the slope is:

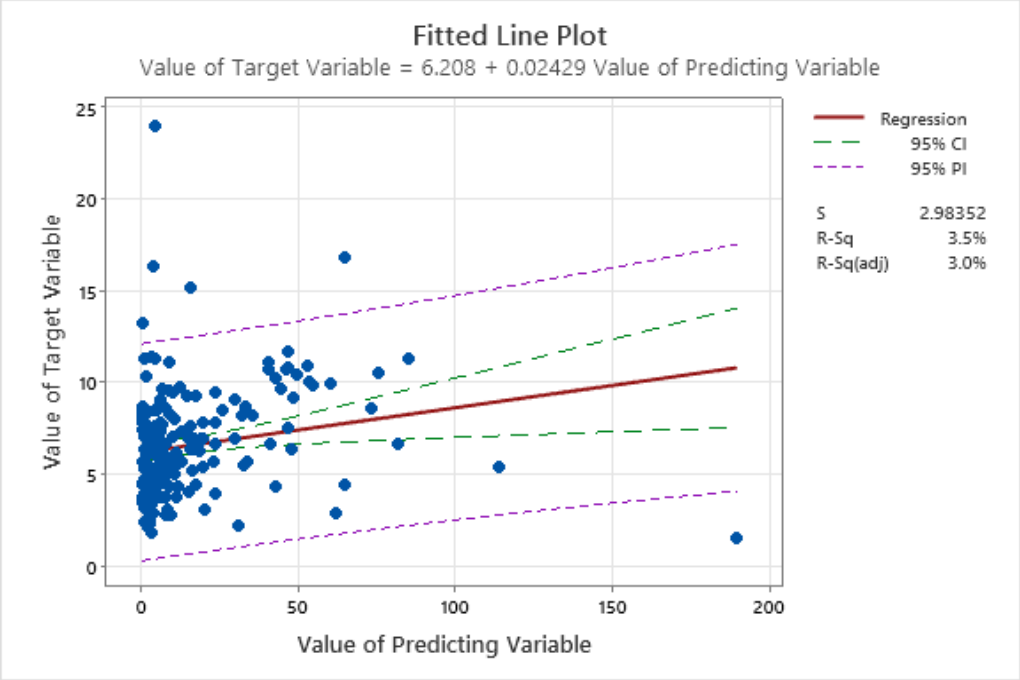
$$t = \frac{\widehat{\beta_1} - Hypothesized\ Value}{\delta_{\widehat{\beta_1}}} = \frac{0.02429 - 0}{0.00937} = 2.5923$$

This is statistically significant compared with the critical value of 1.96, so, we are 95% confident that the null hypothesis could be rejected and the alternative hypothesis of $\beta_1 > 0$ could be accepted. The other variable, the intercept, is also statistically significant, as its t-statistics with the null hypothesis of $\beta_0 = 0$ can be calculated as below:

$$t = \frac{\widehat{\beta_0} - Hypothesized\ Value}{\delta_{\widehat{\beta_0}}} = \frac{6.208 - 0}{0.262} = 23.72$$

So, it's obvious that this value is far greater than the critical value of 1.96, and we are 95% confident that the alternative hypothesis of $\beta_0 > 0$ could be accepted.

The following plot illustrates the confidence interval and the prediction interval of the regression model. Here, the confidence interval can be used to provide an estimate of the average health expenditures for all countries for a given GDP per capita with a certain level of variability, whereas the prediction interval can provide a prediction of the health expenditure for a particular country for a given GDP per capita (Simonoff 4).



Three aspects should be highlighted from the plot. The first thing worth mentioning is that both intervals, particularly the confidence interval, are substantially widen as they approach to the right, indicating that the accuracy of the predictions are lower when the data is further away from the majority of the points. Further, the prediction interval is wider than the confidence interval because prediction for an individual country have greater uncertainty compared to the prediction for the average. Besides, as prediction interval could indicate where the data is likely to lie, it is reasonable to expect 95% of all data points to lie within the prediction interval. Interestingly, in this plot, a few countries do have data points distributed outside the prediction interval, most notably Tuvalu in the top left and Monaco in the bottom right.

Let's use China's GDP per capita data to further illustrate how the confidence interval and the prediction interval can be applied to a particular value.

Regression Equation

Health expenditure (% of GDP) = 6.208 + 0.02429 GDP per capita (in thousands)

Settings

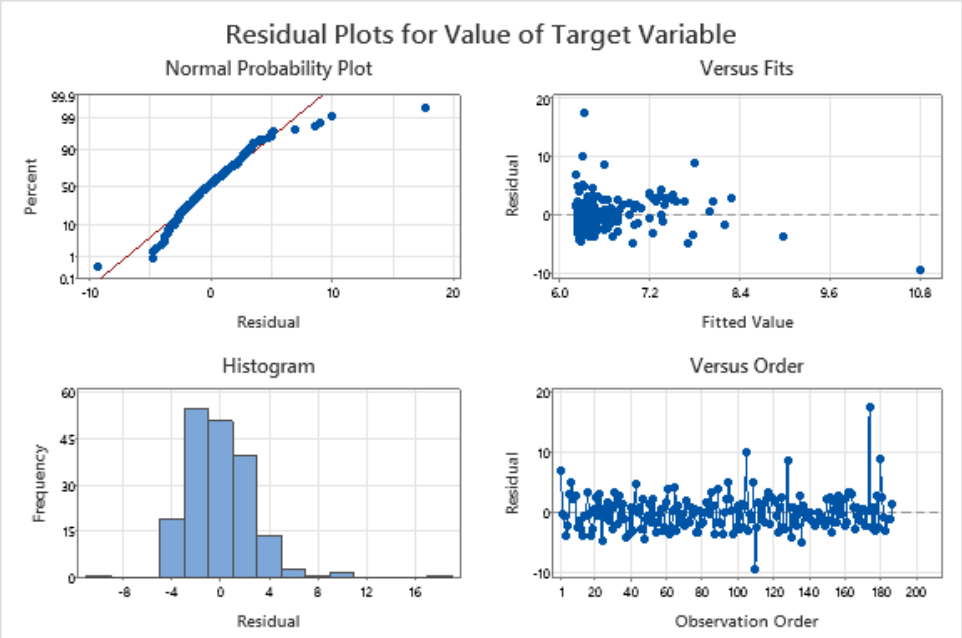
Variable	Setting
GDP per capita (in thousands)	9.96

Prediction

Fit	SE Fit	95% CI	95% PI
6.45001	0.224118	(6.00786, 6.89217)	(0.547335, 12.3527)

As the above table shows, when the GDP per capita is set with 9960 US dollars, we are 95% confidence that the true average health expenditure for all the countries is between 6.00786% of GDP and 6.89217% of GDP, and the true health expenditure for a particular country is between 0.547335% of GDP and 12.3527% of GDP. However, although China's actual health expenditure as percentage of GDP, 5.4, falls within the above 95% prediction interval, the range of this interval is too wide to be statistically significant in practice.

Now, let's take a look at the "four in one plot" to see if the prior interesting points are truly unusual, as well as if our assumptions of least squares regression hold true.



There's insufficient evidence to say that the first assumption is violated and the error has a non-zero expected value because there does not exist specific subgroups that systematically appear to be below or above the regression line. Countries from different regions are evenly distributed above and below the regression line, and there is no particular region, such as Asia, where countries have significantly higher health expenditures (% of GDP) than countries in another particular region. As for the second assumption of homoscedasticity, the plot of residuals versus fitted values does not exhibit a lack of pattern, therefore indicating nonconstant variance and violation of this assumption. The third assumption suggests that the errors should be uncorrelated with each other, and this is not violated because health expenditure (% of GDP) is a relatively individualized matter that is very much influenced by each country's own realities, so, it is impossible to know the error of one specific country given the error of another country. However, the fourth assumption, that errors are normally distributed, is partially violated. As we can see from the normal probability plot, some certain points deviate significantly from the line. Similarly, as shown in the histogram, while there is a concentration on zero, the distribution is skewed to the right, and both sides have fat tails, indicating non-normality.

As aforementioned, one of the obviously unusual points is Monaco, the bottom-right one in the plot of residuals versus fitted values. It is a leverage point that has an unusual predicting variable value as it's GDP per capita is 189507 US dollars, which is about 80.5 standard deviations higher than the average GDP per capita of all the 187 countries. This is also an outlier point because Monaco's target value, health expenditure as percentage of GDP, is obviously lower than that of other countries. This is problematic because it may have a considerable effect on the fitted regression, including measures like R-squared, t-statistics, and the F-statistic (Simonoff 5). As we all know, GDP per capita is influenced by both population size and the total

GDP. In this case, on the one hand, the lower the denominator value, the higher the GDP per capita. Monaco has a very small population of 0.037 million in 2019, according to the World Bank, which contributes to its high GDP per capita. On the other hand, Monaco's tertiary industry is well developed. Following the establishment of the Monaco Monte Carlo Casino in 1865, Monaco's gambling industry grew rapidly, and this industry has played an important role in fostering local economic development. Furthermore, Monaco is also famous for its finance industry. Many well-known banks and corporations, such as Barclays, have established branches here, bringing a substantial amount of liquidity to this country. As a result, Monaco's high GDP per capita is a special case among all the countries, so we need to consider in particular the impact of this leverage point on the regression model. So, I try to remove Monaco from the data set and form a new regression model without any application to it. The following are the statistics and some plots.

Regression Equation

Health expenditure (% of GDP) = 5.821 + 0.0570 GDP per capita (in thousands)

Coefficients

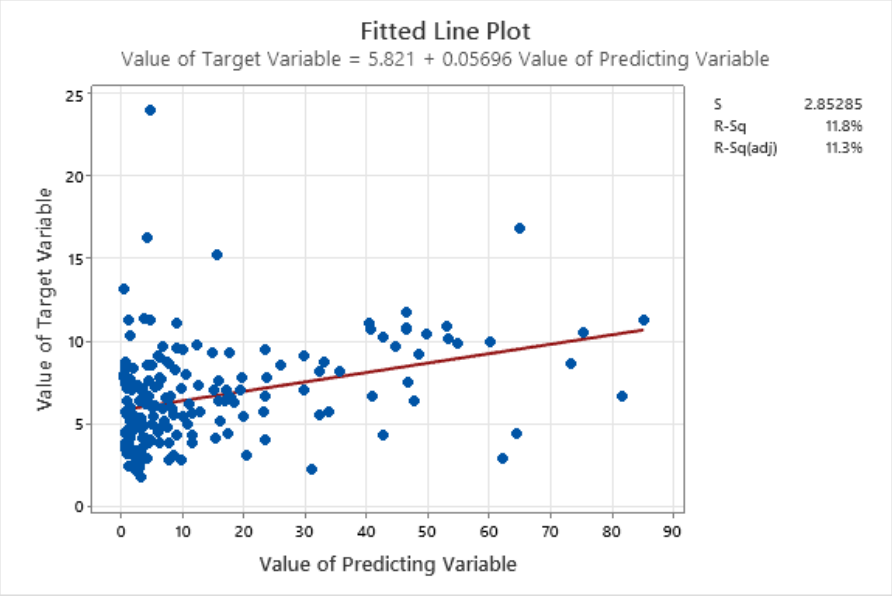
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.821	0.265	21.98	0.000	
Value of Predicting Variable	0.0570	0.0116	4.93	0.000	1.00

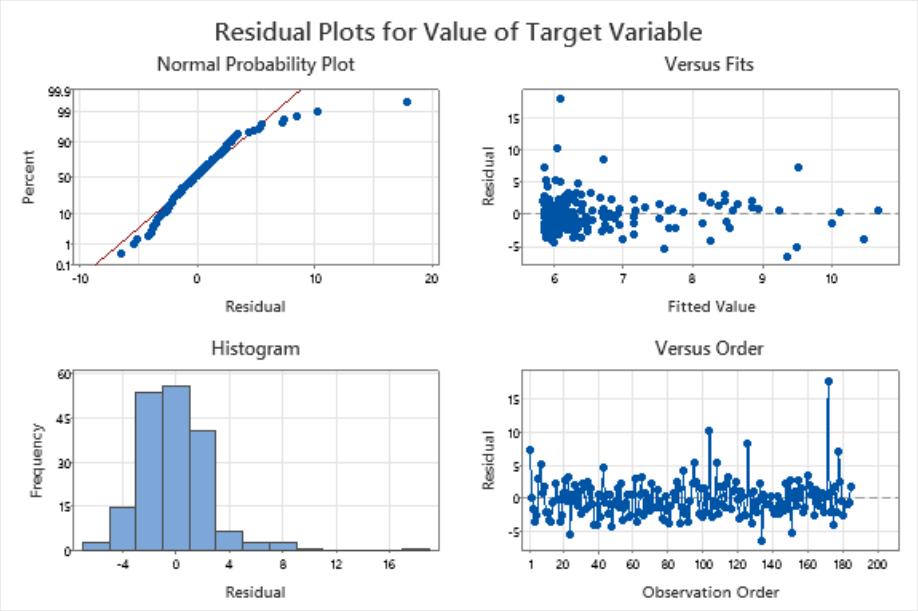
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.85285	11.78%	11.29%	9.60%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	197.72	197.718	24.29	0.000
Value of Predicting Variable	1	197.72	197.718	24.29	0.000
Error	182	1481.25	8.139		
Lack-of-Fit	181	1469.73	8.120	0.70	0.765
Pure Error	1	11.52	11.520		
Total	183	1678.97			





As we can see from the regression equation, omitting Monaco will result in a new R-squared value of 11.78%, which has improved a lot compared with the previous R-squared value of 3.50% and this indicates that this new regression without Monaco is much strong than before. The intercept is smaller than before, and this means that in the current regression model, the health expenditure of a country with zero GDP per capita will account for 5.821% of its GDP. But this is not economically practical since it's meaningless to discuss a country's health

expenditure without the GDP statistic. The slope coefficient of 0.0570 has increased, and this means that a one-thousand US dollars change in GDP per capita is now associated with an estimated expected 0.0570 percentage point change in health expenditure (% of GDP). Both coefficients are statistically significant. The t-statistic of 4.93 for the GDP per capita rejects the null hypothesis of $\beta_1 = 0$ and the t-statistic of 21.98 for the constant also significantly rejects the null hypothesis of $\beta_0 = 0$.

However, this data set without Monaco still display non-normality. Tuvalu, the top-left point in the fitted line plot of the previous data set, is now the right-most one that deviate dramatically from the line in normal probability plot and the top-left one in the plot of residuals versus fitted values, indicating that this is an outlier point worth discussing. This is a country with GDP per capita of 4652 US dollars and health expenditure of 24% of GDP, which is about 5.75 standard deviations higher than the average health expenditure (% of GDP) of all the countries. This is a special case because this country's huge proportion of health expenditure in GDP is a result of its unique environmental and social circumstances. Tuvalu is a small island country in the Pacific with limited resources. The cost of healthcare services in Tuvalu may be relatively high due to factors such as its remote location and high transportation costs. What's more, Tuvalu has a high burden of non-communicable diseases such as diabetes, hypertension, and cardiovascular disease (Global Nutrition Report). This may be due to the relatively unhealthy diet of Tuvaluans, which is mostly highly processed and imported foods. Also, the problem of salt water intrusion due to sea level rise can affect the quality of fresh water supply, which may increase the risk of waterborne diseases. These diseases are often chronic and require ongoing medical care and preventive measures, which can be expensive. Therefore, I try to exclude the influence of Tuvalu in the regression model and here's the new statistics.

Regression Equation

Health expenditure (% of GDP) = 5.685 + 0.0597 GDP per capita (in thousands)

Coefficients

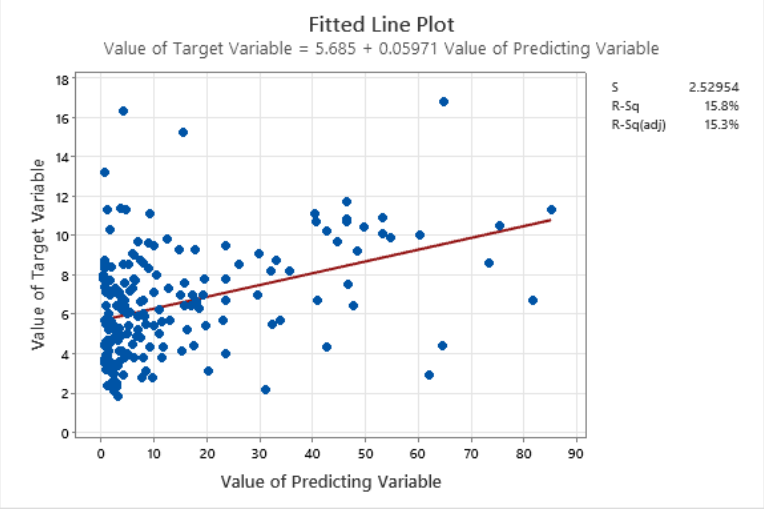
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.685	0.236	24.13	0.000	
Value of Predicting Variable	0.0597	0.0103	5.82	0.000	1.00

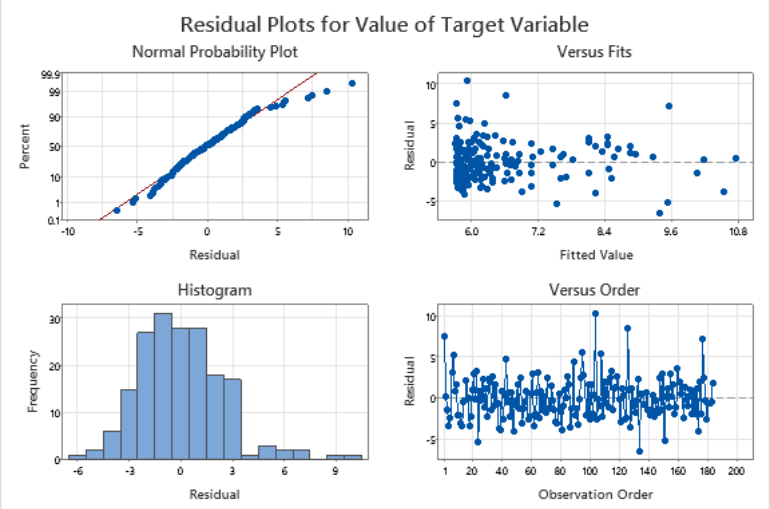
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.52954	15.78%	15.31%	13.43%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	216.92	216.924	33.90	0.000
Value of Predicting Variable	1	216.92	216.924	33.90	0.000
Error	181	1158.14	6.399		
Lack-of-Fit	180	1146.62	6.370	0.55	0.820
Pure Error	1	11.52	11.520		
Total	182	1375.07			





The above plots are based on the data set without Monaco and Tuvalu. We can see that the new R-squared value is 15.78%, indicating that Monaco and Tuvalu both falsely make the model look weak. The new intercept is 5.685 with a statistically significant t-statistic of 24.13 and the new slope coefficient is 0.0597 with t-statistic of 5.82. The normal probability plot still has a few points that depart from the straight line, but these data points are reasonable representations of specific countries in the real world. For instance, the cluster of data points in the upper right corner includes a number of island countries in Oceania. While their relatively high health expenditures are driven by their particular climate, they do not experience the same extreme examples of migration as Tuvalu. It would be useless to keep eliminating these data; instead, think about including new variables may improve the regression model. Therefore, taking GDP per capita (in thousands) as the predicting variable and health expenditure (% of GDP) as the target variable would result in the below regression equation:

$$\text{Health expenditure (\% of GDP)} = 5.685 + 0.0597\ \text{GDP per capita (in thousands)}$$

We can learn from the model that the fact that health expenditure (% of GDP) always has an intercept larger than zero proves that it is a rigid expenditure for countries. It is the fundamental expense that must be covered to keep a working healthcare system in place and guarantee that the general public has the right to access essential medical care. If these demands are not met, it may have a detrimental impact on individual's health and slow down the economic growth. Hence, government budgets frequently assign health expenditure a high priority, independent of the level of economic growth of any country. Also, as countries progress to produce more wealth, given the benefits and significance of health expenditure, they may possibly consider raising their percentage of healthcare expenditure in GDP to better protect the welfare of their residents and the country's future development.

Works Cited

Muysken, Joan. “Health as a Principal Determinant of Economic Growth.” No. 024. Maastricht University, Maastricht Economic Research Institute on Innovation and Technology (MERIT), 2003.

Penghui, Xu, et al. “Direct and Indirect Effects of Health Expenditure on Economic Growth in China.” *Eastern Mediterranean Health Journal*, vol. 28, no. 3, Mar. 2022, pp. 204–12. *EBSCOhost*, https://doi.org/10.26719/emhj.22.007.

Raghupathi, Viju, and Wullianallur Raghupathi. “Healthcare Expenditure and Economic Performance: Insights From the United States Data.” *Frontiers in Public Health*, vol. 8 156, May 2020. EBSCOhost, https://doi.org/10.3389/fpubh.2020.00156.

Simonoff, Jeff. Purchasing Power Parity: Is It True?

Zon, Adriaan van and Joan Muysken. “Health as a principal determinant of economic growth.” *Research Memorandum*, vol. 21, 2003. http://collections.unu.edu/view/UNU:1144#viewMetadata

“Population, Total - Monaco.” *The World Bank*, https://data.worldbank.org/indicator/SP.POP.TOTL?locations=MC. Accessed 26 Feb. 2023.

“Tuvalu Nutrition Profiles.” *Global Nutrition Report*, https://globalnutritionreport.org/resources/nutrition-profiles/oceania/polynesia/tuvalu/. Accessed 26 Feb. 2023.