Exploring Factors Affecting GDP Per Capita

Nathan Dennis , Kevin Wu , Katherine Jackson , Eigard Alstad , Michael Pocress

Abstract

This paper investigates the impact of various health and economic indicators on GDP per capita, focusing on the log-transformed GDP per capita to better model the relationship. GDP per capita is an important indicator of economic development, reflecting the standard of living and economic health of a country. Using a dataset covering the period from 2000 to 2015, we identified significant predictors of log GDP per capita. Through exploratory data analysis (EDA), we explored relationships between health and economic indicators and log GDP per capita. We then addressed multicollinearity and applied model selection techniques (forward, backward, and stepwise) alongside a mixed-effects model. Evaluating based on AIC, the mixed effects model with a random intercept and slope for Country achieved the lowest AIC. Our results show that life expectancy has a significant positive relationship with log GDP per capita, while some other health factors were found to be less impactful. We concluded that a variety of economic and health factors play a crucial role in predicting log GDP per capita.

1 Introduction

Gross Domestic Product (GDP) per capita is a key economic indicator widely used to assess the standard of living, economic performance, and prosperity of nations. It is often associated with many socioeconomic outcomes, including health, education, and quality of life. Understanding the relationship between GDP per capita and factors such as life expectancy is crucial for nations and economists who aim to design strategies to improve economic conditions around the world.

In this paper, we analyze GDP per capita across various countries and regions between the years 2000-2015. Due to the highly skewed nature of GDP per capita, we apply a log transformation to make the relationship between GDP per capita and its predictors more appropriate for modeling. Using a comprehensive dataset with many variables that may influence national health and economic conditions, we aim to identify factors that drive improvements in overall GDP per capita globally. Our investigation will begin with exploratory data analysis of our given dataset, then transition into model building to explore these relationships.

We are interested in assessing the effect of various factors, including economic and health indicators, on GDP per capita. Our primary hypothesis is that a higher life expectancy is positively correlated with a higher GDP per capita, and that this relationship is statistically significant. Previous studies have explored similar questions, such as how increased life expectancy improves overall economic growth [2]. Others have explored the opposite trend, using economic factors, such as GDP per capita, to predict life expectancy [4]. Building on this research, we were also interested in determining whether health factors or economic factors have a stronger influence on GDP per capita. We hypothesize that economic factors such as years of schooling or a country's development status would have a stronger impact on GDP per capita. However, we also explore how health factors such as HIV cases or average BMI affect GDP per capita.

2 Description of Data

We chose to use a life expectancy dataset from Kaggle, where the original data came from the Global Health Observatory data repository under the World Health Organization (WHO). There are 21 columns in this dataset with 2864 observations over the years 2000-2015. There are 179 countries represented in this dataset across these years. This dataset contains information on life expectancy, health metrics such as average BMI and immunization coverage against diseases like Polio and Hepatitis B, economic indicators such as GDP per capita and economic development, and demographic data. For further details on the data, please refer to the appendix section .1.

3 Exploratory Data Analysis

3.1 Missing Data

In this dataset we have no missing data. This is because the dataset was already cleaned and available on Kaggle, which was the source of this data. There were missing values in the original dataset which had 193 countries, while the cleaned data had 179 countries. We decided to investigate which countries these are and why the data was missing.

We discovered that in the original dataset there were many missing values for some countries. Some were more obvious, such as North Korea, which is a country that is already very private, and Vatican City, which is a very small country. Other missing countries such as Andorra and Nauru both had relatively small populations, hence it may be harder to find data for these countries. Most of the missing countries had this commonality of being lesser-known countries with smaller populations. We looked into the data source to determine if there would be any way to fill in the missing data, however, the missing countries simply had too many missing values to include in analysis, leading us to decide against this.

We also noticed that when building the mixed effects model, we had a major outlier when plotting the residuals. After exploring the data, we discovered that there may have been a typo in the data, where Eritrea in 2015 had a sudden spike with a very large GDP per capita of 9011, when previous years were around 700-900. Rather than removing this data point, we filled in the value with 901.

3.2 Data Visualization

In this section, we create multiple visualizations to observe trends and relationships between different variables and GDP per capita. We discover if any transformations could be used based on these visualizations, or if any interaction terms could be interesting to explore in the model. First, we observe the distribution of GDP per capita. If we can see any skewing, we can apply a transformation which would benefit model building and interpretation of relationships.

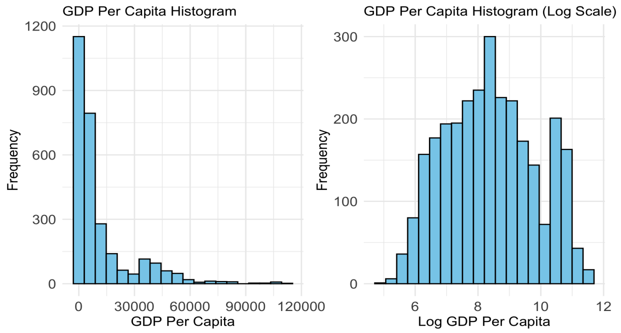

Figure 1: GDP Histogram

We can see the histograms of GDP per capita in Figure 1, where there is an obvious right skew without any transformations. This is something we would like to avoid, as skewed data can violate the assumptions of many models. We have learned that when the response variable has heavy right skew, we can take the log of this variable to fix it. After taking the log transform of GDP per capita we observe a much cleaner normal distribution. This transformation eliminates the skew, allowing for more reliable model building and a clearer interpretation of the relationship between variables with log GDP per capita.

Next, we can observe a correlation matrix with this new log GDP per capita variable. We note the features that are highly correlated with log GDP per capita and check any relationships between variables for multicollinearity. We will further check for multicollinearity when model building through observing variable VIF scores.

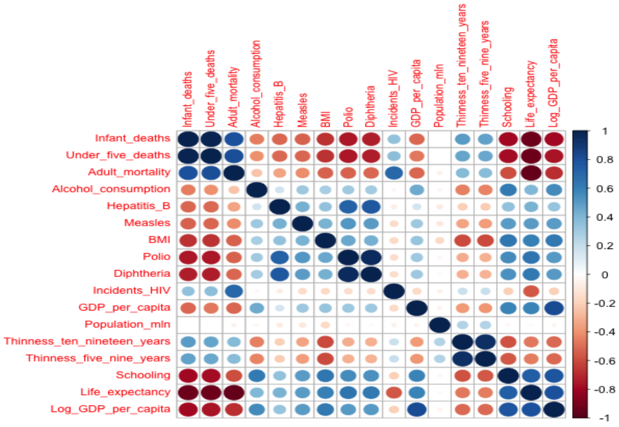

Figure 2: Correlation Matrix

In Figure 2, we see many variables which are highly correlated with log GDP per capita. *infant_deaths* (**ID**), *under_five_deaths* (**UFD**), *Adult_mortaility* (**AM**), and *Life_expectancy* (**LE**) appear to be the most highly correlated with log GDP per capita, and are all influenced by mortality. There are some other variables that have a lower correlation with log GDP per capita, such as *Incidents_HIV* and *Population_mln*.

With regard to multicollinearity, it seems like there is quite a bit in this dataset. Many variables are highly correlated with each other, such as the same 4 variables we noted previously that were highly correlated with log GDP per capita. This makes sense as they are themselves correlated since they all relate to morality. Other factors which are highly correlated include health factors such as Polio and Diptheria, both of which represent immunization coverage (percentage) against these diseases. We observed the correlation coefficients between the 4 variables influenced by mortality (with the abbreviations mentioned in the bold text describing them):

	UFD	**ID**	**AM**	**LE**
UFD	1.0000000	0.9856513	0.8023611	-0.9204191
ID	0.9856513	1.0000000	0.7946609	-0.9200319
AM	0.8023611	0.7946609	1.0000000	-0.9453604
LE	-0.9204191	-0.9200319	-0.9453604	1.0000000

Table 1: Correlation Coefficients of Mortality Variables

We see life expectancy is strongly negatively correlated to the other mortality variables. This makes sense, the other variables are all related to factors that influence death so life expectancy would decrease as these variables increase.

After this analysis, we decided to look into the relationship between log GDP per capita and different variables. First, we visualized a plot between log GDP per capita and life expectancy. We saw in the correlation matrix that there appears to be a strong positive relationship between the variables.

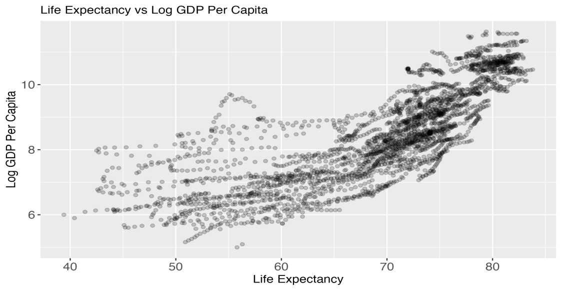

Figure 3: Life Expectancy vs log GDP Per Capita

Looking at Figure 3, we can see there is a strong positive relationship between life expectancy and log GDP per capita. As life expectancy increases, log GDP per capita also tends to increase. This observation can help us in model building, as we already see a clear linear relationship between these variables. We did consider if adding a quadratic term could be necessary since it seems like there may be a quadratic relationship, but decided on using a linear term for this relationship to better answer our research question.

Next, we observed a scatter plot between years of schooling and log GDP per capita. In the correlation matrix, we saw that there may have been a strong positive correlation between the two variables. However, we were interested in seeing if there is any sort of quadratic term necessary to model the relationship between these variables.

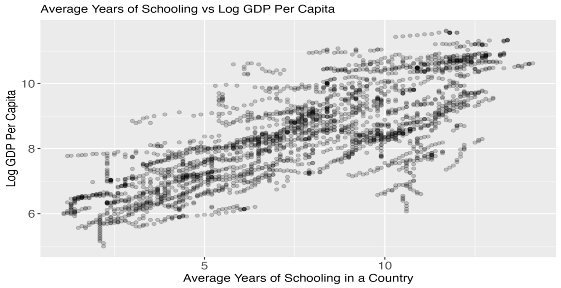

Figure 4: Years of Schooling vs log GDP Per Capita

In Figure 4, we observe the strong positive relationship that was highlighted in Figure 2. As the average number of years people spend in school increases, so does log GDP per capita. This is expected, as higher education levels typically lead to greater productivity and economic growth. A lower average number of years of schooling may indicate a less educated population, which could result in a lower GDP per capita due to lower overall income.

Furthermore, past research has indicated that the relationship between GDP and years of schooling (education) is dependent on the development status of a country [7]. Specifically, the correlation between GDP and education is stronger in developing countries, but tends to be weaker in developed countries. Because of this, we decided to include an interaction term between schooling and the economic development status of a country in model building to explore this relationship.

Next, we observed a scatter plot between log GDP per capita and BMI, a variable we noticed in the correlation matrix that was only moderately correlated with log GDP per capita.

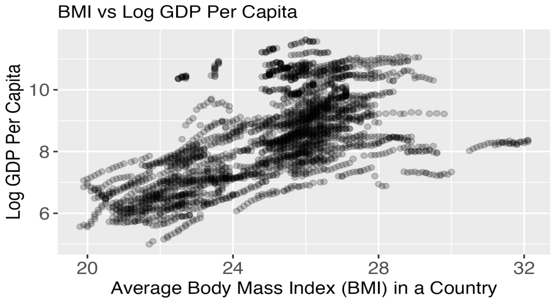

Figure 5: BMI vs log GDP Per Capita

In Figure 5, we can see that there is potentially a moderate positive relationship between average BMI and log GDP per capita. There is a large cluster of points towards the middle of the plot, indicating many BMI observations are within this range. We thought that it may have related to the accessibility of food options in countries, where there are more food choices available in countries with higher GDP and hence a higher BMI. Past studies have shown that GDP is positively correlated with BMI, providing another justification to include this predictor in our final model [1].

After this observation, we were interested in seeing what the relationship would look like if we colored the points based on the development status of a country. We plotted this below:

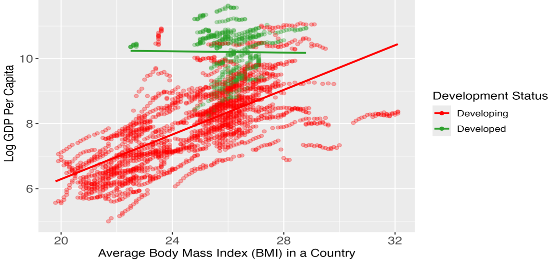

Figure 6: BMI vs log GDP Per Capita, Colored by Development Status

The scatter plot in Figure 6 reveals that the relationship between BMI and log GDP per capita varies by a country's development status. In developing countries, as average BMI increases log GDP per capita tends to increase, suggesting a positive relationship. For developed countries, log GDP per capita remains relatively constant as average BMI increases. Given these differences, we hypothesize that the effect of BMI on log GDP per capita depends on a country's development status, prompting us to add an interaction term between BMI and development status in model building.

We now examine the difference in log GDP per capita across developing and developed nations.

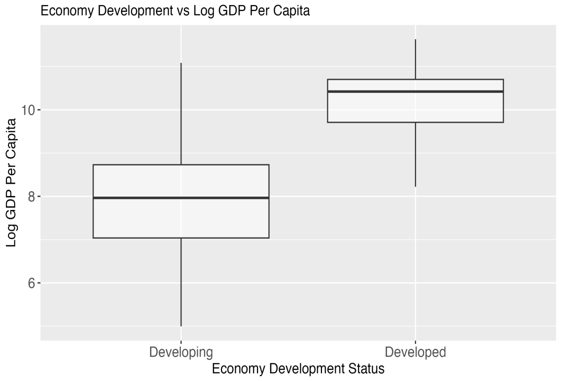

Figure 7: Economy Development vs log GDP Per Capita

From the box plots in Figure 7, we can observe a clear difference in log GDP per capita between the two economy development levels. The GDP per capita for developed countries tends to be much higher than that of developing countries, as can be observed from the significantly higher median log GDP per capita for developed countries.

We also investigated the two variables that we observed in Figure 2 that showed small correlations with log GDP per capita: population of a country in millions and HIV incidents per 1000 between those aged 15-49. We create 2 scatter plots to see why this may have been the case.

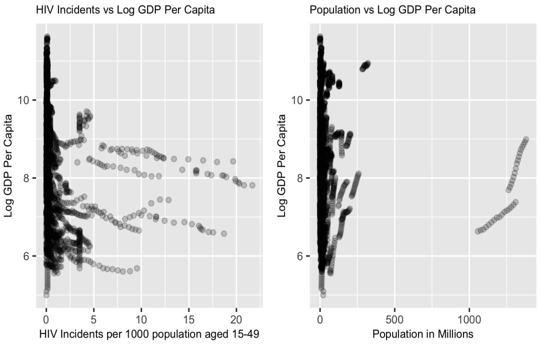

Figure 8: Population and HIV Incidents vs log GDP per capita

As we see in both plots from Figure 8, there is heavy right skewing. In the HIV incidents plot, the majority of the points are clustered on the far left of the graph, indicating a very small number of HIV incidents per 1,000 people. A similar trend is observed in the population graph, where there are only 2 countries which seem to have extremely high populations of over one billion. These are notable outliers in the dataset, and even with transformations the relationship could not be made linear.

We considered against using these variables in the final model selection as they would violate the assumption of linearity. However, research suggests that the prevalence of HIV cases is typically negatively correlated with economic growth, so we decided to include it in the model to observe its effect [6]. However, we decided to exclude population from the model due to the heavy skewing and lack of a linear relationship with log GDP per capita.

Finally, before developing the mixed effects model including a random slope and intercept for each country, we decided that it would be interesting to visualize the trend of log GDP per capita over time across the 16 years in our dataset. We aggregated the mean log GDP per capita across each year and plotted this in Figure 9.

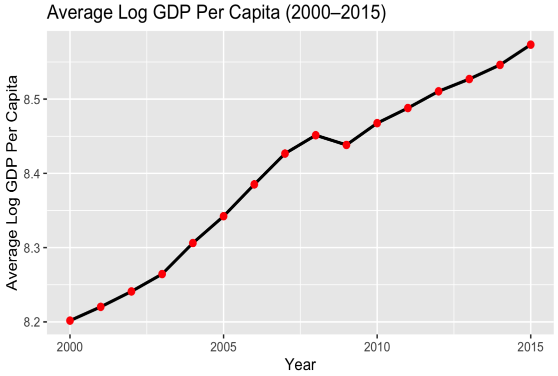

Figure 9: Average Log GDP Per Capita Over Time

As expected, the plot shows that log GDP per capita generally increases over time, with a brief dip in 2009 which could be due to the recession in this time period. Aside from this decrease, average log GDP per capita has risen steadily across the world over the years.

4 Methods

In this section, we provide an overview of the methods we used to address our research question. First, we developed a baseline model based on insights from our EDA and then identified and eliminated multicollinearity in our full baseline model. Next, we used several model building techniques to build models and selected the best model based on AIC.

4.1 Multicollinearity

We now develop a model to predict log GDP per capita. Based on our findings from the previous visualizations, we created a baseline model including all predictors except *Year*, *Economy_status_developing* (the reciprocal of *Economy_status_developed*, and *Region* or *Country* as we are not focused on geographical factors for this first model. We also excluded *population_mln* based on our EDA and *thinness_ten_nineteen_years* since it closely mirrors *thinness_five_nine_years*. We begin with our initial, baseline model:

```
baseline <- lm(Log_GDP_per_capita ~
  Infant_deaths + Under_five_deaths +
  Adult_mortality + Alcohol_consumption +
  Hepatitis_B + Measles +  BMI + Polio +
  Diphtheria + Incidents_HIV + Life_expectancy +
  Thinness_five_nine_years + Schooling +
  as.factor(Economy_status_Developed) +
  BMI * as.factor(Economy_status_Developed) +
  Schooling * as.factor(Economy_status_Developed),
  data=data)
```

Model 1: Full Baseline Model

To determine which features to use in our model, we first aimed to minimize multicollinearity. In the correlation matrix shown in Figure 2, we observed significant multicollinearity among several predictors. To verify these findings, we used the *vif* command in R to calculate the Variance Inflation Factor (VIF) scores. VIF measures the inflation in the variance of a regression coefficient due to collinearity (correlation) with other predictors, and is typically applied to continuous variables. We display the VIF for all of our continuous variables:

Variable Name	VIF
Infant_deaths	45.246351
Under_five_deaths	46.567900
Adult_mortality	24.442049
Alcohol_consumption	2.459010
Hepatitis_B	2.609419
Measles	1.572780
BMI	2.972737
Polio	12.014544
Diphtheria	12.971048
Incidents_HIV	2.809618
Life_expectancy	46.104331
Thinness_five_nine_years	1.973627
Schooling	4.869217

Table 2: Variance Inflation Factors (VIF) for Variables

We can see that there are many variables with high VIF scores. Typically, a VIF score higher than 5 indicates multicollinearity, however, recent studies have suggested that 10 may be a better threshold. [3] We observed that all 4 variables influenced by mortality had VIF scores of greater than 5 and 10: *Infant_deaths*, *Under_five_deaths*, *Adult_mortality*, and *Life_expectancy*. Since all of these variables were related to life expectancy as they dealt with mortality, we decided to remove all variables related to mortality except life expectancy itself to eliminate this multicollinearity.

The last removal choice we made was between *Diphtheria* and *Polio*, where both variables represented percent immunization against these diseases. Since Diphtheria had the highest VIF score, we removed it from the model. After this, we observed the results based on the updated model:

Variable Name	VIF
Alcohol_consumption	2.288046
Hepatitis_B	2.225600
Measles	1.563855
BMI	2.742238
Polio	3.312312
Incidents_HIV	1.920855
Life_expectancy	5.433497
Thinness_five_nine_years	1.917599
Schooling	4.318028

Table 3: VIF Scores for Selected Variables

Based on Table 3, we observed that life expectancy is the only variable with a VIF score slightly above 5. We decided that although it was above the usual threshold, we would keep it in the model. One reason was because we were interested in its effect on log GDP per capita from our research question, so we decided to keep it in the model. Also, 5 is just a typical threshold that some use, but others think this may be too harsh of a penalty and prefer using 10.

4.2 Model Building

We built four different models using the Forward, Backward, and Stepwise Selection procedures, as well as a mixed effects model that included a random intercept and slope for country. For model selection, we used AIC as our selection criteria.

The full model we used consisted of the variables selected in the previous step:

```
full_model <- lm(Log_GDP_per_capita ~
  Alcohol_consumption + Hepatitis_B + Measles +
  BMI + Polio + Incidents_HIV + Life_expectancy +
  Thinness_five_nine_years + Schooling +
  as.factor(Economy_status_Developed) +
  BMI * as.factor(Economy_status_Developed) +
  Schooling * as.factor(Economy_status_Developed),
  data=data)
```

Model 2: Full Model

The model (shown above) includes the categorical economy status and the interaction terms we decided to use based on the EDA. We did not calculate VIF scores for these variables as the development status is categorical, whereas the VIF scores focus on numeric variables.

4.2.1 Forward Selection

The first model building technique we used was Forward Selection which consisted of 5 key steps:

1. Start with the intercept only model, $E[Y|X] = \beta_0$.
2. Consider all models with one more term added.
3. For each model, calculate using the AIC criterion which is the criterion we chose.
4. Include the term that leads to the smallest AIC in the new model.
5. Iterate steps 3 and 4 until no further AIC improvement is possible.

Forward Selection

We began with the intercept only model and iteratively added predictors one by one from of the set of predictors included in Model 2. The process stopped when adding any additional predictors no longer reduced the AIC. The final model selected by forward selection is the model with the lowest AIC among the models created during this process.

4.2.2 Backward Selection

The next model building technique we used was backward selection. It is similar to forward selection with one key difference. It begins with the full model, including all predictors, and iteratively removes predictors. The process continues until no further removal of predictors results in a reduction in AIC.

1. Start with the full model, including all predictors.
2. Consider all models with one term removed.
3. For each model, calculate the AIC value.
4. Remove the term that leads to the smallest AIC in the new model. (Largest AIC reduction)
5. Iterate steps 3 and 4 until no further removal of predictors results in a reduction of AIC.

Backward Selection

In the backwards selection algorithm we began with the full model from Model 2, where every variable we selected in Model 2 was included in the initial model. The algorithm then iteratively removed one predictor in each loop, evaluating the AIC at each step and continued until there was no further reduction in AIC by removing a predictor, choosing the final model with the lowest AIC.

4.2.3 Stepwise Selection

The third model selection approach we employed was stepwise selection. This selection technique is similar to the forward and backwards selection procedures, as it also iteratively adds/removes predictors from the model until some stopping condition is reached. The formula is given in the following figure:

1. Start with the intercept-only model.
2. Do one step of forward selection.
3. Do one step of backward selection.
4. Iterate steps 2 and 3 until the AIC cannot be improved

Stepwise Selection

The stepwise selection method combines both the forward and backward selection techniques. It begins with the intercept model in our formulation and iteratively adds and removes predictors from the 12 possible options based on improvement in AIC (reducing AIC). The process continues until no further improvements in AIC can be made.

4.2.4 Mixed Effects

The final technique we employed was mixed-effects modeling. Mixed-effects models are a special case of multilevel models with 2 levels. In our case, our data structure consists of 16 years worth of data for 179 countries, leading to repeated measures within each country. This was seen in the EDA, where we observe obvious correlations between certain groups of observations in the graphs. Given this structure, we believed it would be interesting to apply a mixed-effects model to our data since it accounts for both fixed effects, the average association between Year and log GDP per capita, and random effects, in our case, Country, with a random intercept and slope for each country. The fixed effects can be interpreted as the average association between Year and log GDP per capita across the data. The random intercept for Country accounts for the variability in baseline log GDP per capita for different countries. The random slope for Country allows the association between Year and log GDP per capita to vary across countries, acknowledging that some countries may experience stronger or weaker trends in GDP growth over time. We can express the model in this standard notation for a random intercept model:

$$Y_{ij} = \alpha_j + \tau_j Year_{i,j} + \sum_{k=1}^{13} \beta_k X_{i,j,k} + \epsilon_{ij}; \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_Y^2)$$
$$i = 1, \ldots, n_j \qquad\qquad j = 1, \ldots, J$$

Where:

$$\begin{aligned} \alpha_j &= \mu_\alpha + \eta_{\alpha,j} \qquad & \eta_{\alpha,j} &\sim N(0, \sigma_\alpha^2) \\ \tau_j &= \mu_\tau + \eta_{\tau,j} \qquad & \eta_{\tau,j} &\sim N(0, \sigma_\tau^2) \end{aligned}$$

The variable Y_{ij} is the log GDP per capita for the i-th observation in the j-th group, the log GDP per capita for the i-th observation in the j-th country with J = 179 countries and $n_j = 16$ years of data. β_k is the fixed-effect coefficient for the k-th predictor in $X_{i,j,k}$, the k-th predictor variable for the i-th observation in the j-th country. Then, α_j is the random intercept for the j-th country, which allows the baseline (intercept) value of the log GDP per capita to vary across different countries. The intercept depends on the fixed average intercept, μ_α, which is the overall average baseline log GDP per capita across all countries, alongside the random effect $\eta_{\alpha,j}$, which represents the deviation of the intercept for country j from the overall mean intercept. τ_j is the random slope for the effect of Year in country j, allowing the relationship between year and log GDP per capita to vary across countries. This depends on the grand mean slope, μ_τ, which represents the overall average effect of year on log GDP per capita across all countries, and the random effect $\eta_{\tau,j}$, which represents the deviation of the slope for country j from the overall mean slope. We build the mixed effects model with the same 12 variables in the original full model, adding an extra year term for the random effect so 13 total fixed effects modeled by $\sum_{k=1}^{13} \beta_k X_{i,j}$.

4.3 Comparison of Models

We compared the AIC values of the four modeling approaches we implemented. We chose to use AIC as the selection criterion as it imposed a larger penalty on models with a larger number of parameters, as indicated by its formula. We wanted to prevent overfitting, so we hoped that by using AIC, it could help us choose the model that would over fit the least.

Furthermore, AIC is widely used in similar health and economic research contexts. For example, in health research the AIC criterion was applied to determine the optimal representation of dietary variables in a longitudinal dental study [5]. This reflects its utility in managing complex datasets, particularly in public health research.

5 Results

5.1 AIC Comparison

Model Name	AIC
Forward	5589.862
Backward	5589.862
Stepwise	5589.862
Mixed-Effects	-5443.377

Table 4: AIC Scores Across Models

From the results displayed in Table 4, we observed that the forward, backward, and stepwise selection all produced identical AIC values of 5589.862. This suggested that all 3 methods produced the same exact model and are equally optimal in terms of balancing model fit and complexity as measured by AIC. We checked and discovered that the models were the exact same, as they used all 12 predictors from Model 2.

We note that the three selection models had the highest AIC values, while the mixed-effects model had a better, lower AIC of -5443.377. This makes sense, as including a random intercept and slope for each country would improve analysis as each country would have a different baseline log GDP per capita, which would grow at different rates.

5.2 Selection Procedure Output

We began by analyzing the first model determined previously, the model developed by the selection procedures which had the highest AIC. First, we observed which predictors were chosen for this model, which revealed that all predictors were used. We then verify that key assumptions, including existence, linearity, independence, homogeneity of variance, and the optional normality assumption were not violated. For simplicity, we refer to *Economy_status_Developed* as ESD, where ESD1 indicates the country has a developed economy.

Variable Name	Estimate	p-value
Intercept	-1.9333174	2e-16
Life_expectancy	0.1040712	2e-16
Schooling	0.0392499	5.92e-07
Incidents_HIV	0.1439772	2e-16
ESD1	3.9888553	4.49e-07
BMI	0.1225751	2e-16
Polio	-0.0150864	2e-16
Measles	0.0052187	9.07e-11
Alcohol_consumption	0.0303222	3.07e-11
Hepatitis_B	0.0028016	0.0121
Thinness_five_nine_years	0.0078622	0.0319
ESD1 · BMI	-0.1713495	2.27e-08
ESD1 · Schooling	0.1060028	1.58e-06

Table 5: Estimates and P-Values of Selection Model

We interpret these coefficients and discuss the results further in our conclusion. Next, we create plots to analyze the assumptions of this model. First, we noted that we assumed the existence of residuals. We also recognize that the independence assumption would be violated because data from the same country across different years would be correlated. This correlation arises since a variable's value in one year is directly influenced by its value in previous years in the same country. Data points within each country are correlated, hence the assumption is not satisfied.

We tested the other assumptions, through analysis of fitted vs residual plots and QQ plot.

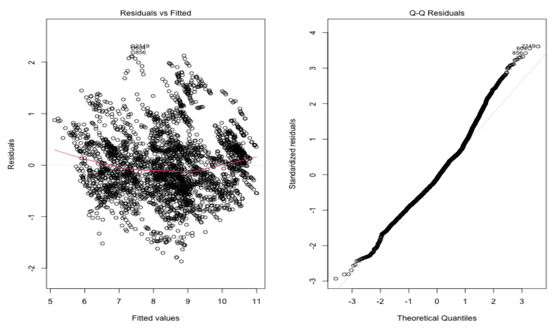

Figure 10: Selection Model Assumptions

First, we checked the residuals vs fitted plot for violation of the linearity assumption. It seems like in the plot the loess curve starts slightly above the 0 residuals line, then dips slightly under towards the middle of the plot and then increases again. However, it does seem that the residuals are evenly distributed around the 0 residual line with mean 0, with little evidence of any patterns of high over and under fitting overall despite the loess curve. This leads us to conclude the linearity assumption is satisfied. We also check the homoskedasticity assumption based on this plot. We see that the residuals generally follow a constant variance/width around the 0 residual line without any noticable pattern, leading us to believe this assumption is satisfied.

We use the QQ plot to check the normality assumption. The residuals seem to generally follow the line representing the theoretical quantiles of the standard normal distribution, except towards the top right of the plot. There is a noticeable deviation from the standard quantiles line in this region, which suggests the normality assumption is not satisfied.

We also check for outliers by observing the studentized residuals across all points in the data:

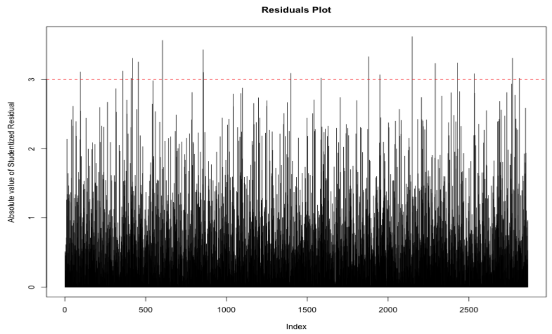

Figure 11: Selection Model Outliers

Based on this plot, we observe that some residuals have studentized residual values greater than 3, suggesting potential outliers. However, these values are not extremely large and do not appear to be major outliers. We found that Equatorial Guinea, United Arab Emirates, and Brunei Darussalam were the most frequent countries associated with these higher residuals, with 18 observations having studentized residuals greater than 3.

5.3 Mixed-Effects Output

We now create and display the results for the mixed effects model, where we have different intercepts and slopes for Year based on the Country in the data. We observe the regression coefficients from the mixed effects model. We used *lmertest* in R to get the p-values of the coefficients. We compare the output of this model to the output from Table 5 in the discussion.

Variable Name	Estimate	p-value
Intercept	2.922	2.59e-08
Life_expectancy	0.01501	9.69e-11
Schooling	0.0457	3.14e-10
Incidents_HIV	0.029	0.000153
ESD1	3.249	0.005919
BMI	0.1348	1.49e-10
Polio	0.00044	0.109589
Measles	0.00076	0.001171
Alcohol_consumption	0.0343	2e-16
Hepatitis_B	0.00091	5.94e-05
Thinness_five_nine_years	0.0013	0.139590
Year	0.0057	0.018319
ESD1 · BMI	-0.067	0.143319
ESD1 · Schooling	-0.0024	0.850281

Table 6: Estimates and P-Values of Mixed Effects

We check the same assumptions for this mixed effects model. We first assumed the existence of residuals assumption. Unlike the previous model, including a random intercept and slope for country helps account for the correlation between observations within the same country across different years. This allows us to conclude that the model properly accounts for this dependency, and the residuals are independent after accounting for these country specific effects. We now check the residuals vs fitted and QQ plot to check the rest of the assumptions.

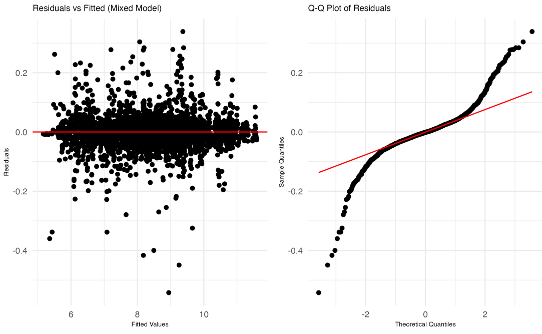

Figure 12: Mixed Effects Model Assumptions

We can see in the residuals vs fitted plot that the points are generally evenly distributed around the red, 0 residuals line. There is no significant patterns of under or over estimation within the residuals across the plot, hence we conclude the linearity assumption is satisfied. Furthermore, the variance of residuals appears relatively constant across the range of fitted values. The potential increase in variance towards the middle of the plot alone is not enough to raise significant concerns, so we conclude the homoscedasticity assumption is satisfied.

As for normality, we look at the qqplot of the residuals. Similar to the previous model, we see that the residuals generally follow the theoretical quantiles line for a normal distribution until the top right and now also bottom left, where there is major deviation away from this line. This causes us to conclude the normality assumption is violated in this model.

Finally, we observe if there are any outliers:

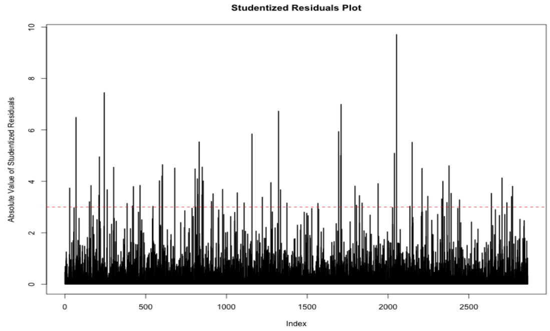

Figure 13: Mixed Effects Model Outliers

Based on the studentized residuals plot, we observe many observations with studentized residuals exceeding a value of 3, with one residual greater than the majority of the others. In total, there were 63 outliers identified. Venezuela, Equatorial Guinea, and Libya were the most frequent outlier countries and the largest residual was Equatorial Guinea in 2000 with a studentized residual value of -9.71. We note that compared to the selection model there are more outliers in this model and one outlier that had a much larger residual than any of the residuals in the previous selection model.

5.4 Model Without Log Output

We also decided to create a model without the log transform of GDP per capita, instead directly predicting GDP per capita using same 12 predictors we used in the selection model previously. We display the coefficients and corresponding p-values, where we will compare this models output to the other models developed. We were interested in comparing its output with the other models and examining how life expectancy specifically is affected to answer our research question further, since throughout this analysis we have used log GDP per capita.

Variable Name	Estimate	p-value
Intercept	-40,880	2e-16
Life_expectancy	909.4	2e-16
Schooling	-18.77	0.888523
Incidents_HIV	1,318	2e-16
ESD1	21,720	0.106856
BMI	-90.6	0.557081
Polio	-120.9	1.00e-06
Measles	-0.5779	0.966348
Alcohol_consumption	-265	0.000654
Hepatitis_B	-5.522	0.772081
Thinness_five_nine_years	-198.5	0.001521
ESD1 · BMI	-2,546	1.13e-06
ESD1 · Schooling	5,744	2e-16

Table 7: Estimates and P-Values of Non-Log Output

6 Conclusion & Discussions

6.1 Selection Model Conclusion

Based on the coefficient values in Table 5, we see which variables were seen as significant in the model to predict log GDP per capita. The only variables which were not significant at a 1% significance level were *Hepatitis_B*, which represents percent immunization coverage of Hepatitis B (HepB3) among 1-year-olds, and *Thinness_five_nine_years*, representing the prevalence of thinness among children aged 5-9 years (BMI < -2 standard deviations below the median). However, at a 5% level these would've been significant, but it is interesting to see they had the largest p-values. This may suggest neither of these health indicators are as useful as other predictors in our model at predicting log GDP per capita.

The rest of the selected variables in the model were found to be significant at a 1% significance level. There were many health factors which were deemed as significant, including key variables we were interested in such as life expectancy and BMI. The positive coefficient for life expectancy indicates that as the life expectancy of a country increases, we expect its log GDP per capita (and hence GDP per capita since the log transform is monotone), to increase as well. In our case, we predict this increase to be about 0.1040712, so we expect on average that a 1 year increase in life expectancy, holding other variables constant, corresponds to a 0.1040712 increase in log GDP per capita.

We also see that the coefficient for the development status of a nation was significant (ESD1) with a p-value of nearly 0, with the baseline level being for developing countries. When the development status of a country is developed rather than developing, we expect on average their log GDP per capita to increase by 3.988, holding other variables constant. Similar to the boxplot in Figure 7, developed countries have higher log GDP per capitas compared to developing countries.

We also note that the interaction terms between the economy status with BMI and Schooling were seen as significant with near 0 p-values. The interaction between economy status and BMI was negative, -0.171, indicating that the effect of average BMI on log GDP per capita is more negative for developed countries than developing countries. If we include the positive main effect of BMI, 0.121323, the overall effect for BMI is negative for developed countries (-0.172 + 0.122 = -0.05). This suggests that for developed countries, an increase in average BMI is associated with a slight decrease in log GDP per capita. For developing countries, the positive main effect of BMI (0.122) suggests that a higher average BMI is associated with an increase in log GDP per capita. The coefficient for the interaction between development status and schooling was positive, 0.106, indicating for developed countries the impact of years of schooling on log GDP per capita is larger than for developing countries.

Other significant variables in our model include immunization against measles and polio, both of which were significant predictors of log GDP per capita with near 0 p-values. Polio had a negative coefficient, indicating that as the immunization coverage of 1-year olds against polio increases, log GDP per capita decreases holding other variables constant. In contrast measles had a positive coefficient, suggesting that as the immunization coverage of 1-year olds against measles increases, log GDP per capita increases, holding other variables constant. While both immunization indicators are significant, their effects on log GDP per capita are opposite in direction.

All of the economic factors were significant in this model, including development status and average years of schooling. We noted previously that 2 health factors, those being *Hepatitis_B* and *Thinness_five_nine_years*, would not be significant at the 1% level. This could imply that certain health factors may not be as impactful in predicting log GDP per capita, but it seems like the economic factors would be important which makes sense since GDP directly relates to the economic status of a country.

6.2 Mixed-Effects Conclusion

One observation we made comparing the fixed effects coefficients in the mixed-effects model (besides year) in Table 6 and selection models coefficients in Table 5 was that many more coefficients were seen as not significant in the mixed effects model. For our primary question regarding life expectancies impact on log GDP per capita, we see that the fixed effects coefficient for life expectancy was positive and the p-value was nearly 0, indicating the effect was significant. In this case since the coefficient was 0.01501, we expect that on average when life expectancy increase by one, log GDP per capita increases by 0.01501 holding all other variables constant. So, we see that life expectancy has a positive effect on log GDP per capita.

At a 5% significance level we noticed health factors such as *Polio*, which represents percentage of 1-year olds immunized against polio, and *Thinness_five_nine_years*, thinness prevalence of those aged between 5 and 9 years old, were not statistically significant in this model. This suggests these variables are not strong predictors of log GDP per capita in this dataset. We also noticed that neither interaction terms are statistically significant in this model with large p values of 0.143 for economy status and BMI, and 0.85 for economy status and Schooling. This could indicate that the relationship between log GDP per capita and predictors BMI and Schooling is consistent across both developed and developing countries, the effects of these variables on log GDP per capita do not significantly differ based on economic status. This contrasts with the previous selection model, which found both interactions statistically significant.

We noticed that the economic factors such as years of schooling, which positively impacts log GDP per capita, and economy status, where developed countries are expected to have a larger log GDP per capita by about 3.249 compared to developing countries, were seen as significant in the model. This is noteworthy, as the main effects that were not significant were all health related factors. Some of these health factors may not be strong predictors of log GDP per capita.

We hypothesize the inclusion of the random effects may have led to a better fit for the data, reducing the significance of terms such as the interaction terms by accounting for differences between countries. We note that since the independence assumption was violated in the selection model the results would be less reliable compared to the mixed-effects model. Failing to account for the dependencies between countries in the selection based model could've lead to more bias estimates overall.

6.3 Non-Log Comparison

When comparing the final model, which used the same 12 predictors to model non-logged GDP per capita, to the previous models, we observe several differences. In this final model there are many variables that were not significant at the 5% significance level, including health indicators such as *BMI* (p-value 0.557), *Measles* (p-value 0.996), and

Hepatitis_B (p-value 0.772), which were all significant in the previous 2 models.

Life expectancy was still significant with a near 0 p-value and positive coefficient, indicating that a higher life expectancy is positively correlated with a larger GDP per capita. Overall, fewer predictors were found to be significant in this model, which could align with our sub-hypothesis regarding whether health or economic factors more strongly contribute to GDP per capita. The lack of significance for many health indicators suggests only specific health factors may be strong associated with GDP per capita. We also note that years of schooling was also deemed insignificant, which could suggest the influence of education on GDP per capita may be weaker than we anticipated.

We also noticed that the interaction terms were both seen as significant in this model, alongside the same sign of the coefficients compared to the selection model. The interaction between development status and BMI was negative, hence the effect of BMI on GDP per capita is more negative for developed countries. The interaction between development status and schooling was again positive, so for developed countries the impact of schooling on GDP per capita is larger than for developing countries. But, we also realize since independence is violated using the same logic as the selection based model, these results may not be accurate.

6.4 Overall Conclusion

From these models we obtained both similar and differing results, which provide valuable insights into the relationship between economic and health factors on log GDP per capita. Every model highlighted the importance of life expectancy as a significant predictor of log GDP per capita, a consistent positive relationship. However, there were differences in the role of interaction terms and significance of certain predictors, particularly health factors, across models. While the mixed-effects model did not find the interaction terms statistically significant, the stepwise model identified both interactions as meaningful. This difference suggests different modeling techniques can yield very different results, emphasizing the need to carefully consider the modeling techniques used when analyzing complex datasets such as this one with the correlated data points based on country. Overall, these results highlight the relationship between economic and health indicators on log GDP per capita and the importance of model selection methods when analyzing data.

6.5 Future Work

In this analysis, we explored how several factors affected GDP per capita of a country, specifically log GDP per capita. While we found many variables to be significant predictors, there are definitely more factors that could further explain a country's economic performance. One key limitation of our analysis was the number of variables used. While this dataset provided valuable insights, it represents only a small subset of the factors that contributes to a country's true GDP per capita. Using a life expectancy dataset may have its limitations, as it primarily focuses on health factors rather than economic ones. While we found many health factors to be significant predictors, economic factors such as infrastructure and trade are likely to have a more direct influence on GDP per capita. We hope that future work can incorporate the factors we found significant in our model, such as life expectancy, towards new models with more variables.

Furthermore, research could be conducted towards the present day to observe if these factors still influence GDP per capita. The data we used ranged from 2000-2015, since these years there have been many changes in the world. The most notable change happened in 2020 during the COVID-19 pandemic. This pandemic heavily affected the overall economy of the world and it would be interesting to see what factors have went in to improve the economy since this pandemic, specifically GDP per capita.

7 Challenges

We faced several challenges during this project, including figuring out what interactions terms we wanted to investigate in our model. We wanted to answer our research question to the best of our ability, which also meant trying to model all the confounders possible in our model and satisfy modeling assumptions. While we decided to include 2 interation terms, there are most likely many more in our dataset that we could've used.

Another challenge we faced was the amount of variables in our dataset. We were interested in investigating the relationship between health factors and GDP per capita, but it may be unconventional using a life expectancy dataset to answer questions regarding GDP per capita. We had to perform a log transformation to get any meaningful results in our EDA to eliminate how skewed GDP per capita was. While we understand the goal of this project was not just to create the best model, we could not create a sufficient model without performing the log transformation, which heavily changed analysis. We wished we could've included more economic indicators in our analysis, as they would be a better fit for this problem. We had some trouble finding data in the first place, but the challenge after finding the data was determining if we had enough relevant predictors to conduct analysis.

References

[1] Garry Egger, Boyd Swinburn, and F.M. Amirul Islam. “Economic growth and obesity: An interesting relationship with world-wide implications”. In: *Economics & Human Biology* 10.2 (2012), pp. 147–153. DOI: 10.1016/j.ehb.2012.01.002. URL: https://doi.org/10.1016/j.ehb.2012.01.002.

[2] Golam Hossain. “Impact of Life Expectancy on Economics Growth and Health Care Expenditures: A Case of Bangladesh”. In: *Universal Journal of Public Health 1(4): 180-186, 2013* 1 (Dec. 2013), pp. 180–186. DOI: 10.13189/ujph.2013.010405.

[3] Katerina M. Marcoulides and Tenko Raykov. “Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods”. In: *Educational and Psychological Measurement* 79.5 (2019), pp. 874–882. DOI: 10.1177/0013164418817803.

[4] Goran Miladinov. “Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries”. In: *Genus* 76.2 (2020), Article 2. DOI: 10.1186/s41118-019-0071-0.

[5] John VanBuren et al. “AIC identifies optimal representation of longitudinal dietary variables”. In: *Journal of Public Health Dentistry* 77.4 (2017), pp. 360–371. DOI: 10.1111/jphd.12220.

[6] Salisu Ibrahim Waziri et al. “Effect of the Prevalence of HIV/AIDS and the Life Expectancy Rate on Economic Growth in SSA Countries: Difference GMM Approach”. In: *Global Journal of Health Science* 8.4 (2015), pp. 212–220. DOI: 10.5539/gjhs.v8n4p212. URL: https://doi.org/10.5539/gjhs.v8n4p212.

[7] Yuyao Wen. “Exploring the GDP-Education Relationship: A Comparative Study of Developing and Developed Nations”. In: *Advances in Economics, Management and Political Sciences* 61 (Dec. 2023), pp. 182–191. DOI: 10.54254/2754-1169/61/20231254.

Appendix

.1 Data Description

- **Country**: Country Observed
- **Region**: The Region in the world the Country is in (Asia, Africa, Oceania, etc.)
- **Year**: Year Observed
- **Infant_deaths**: Represents the infant deaths per 1000 population. (Generally infant is defined as 0-1 years old)
- **Under_five_deaths**: Represents the deaths of children under 5 years old per 1000 population.
- **Adult_mortality**: Represents the deaths of adults per 1000 population.
- **Alcohol_consumption**: Represents the alcohol consumption, recorded in liters of pure alcohol per capita for those 15+ years old.
- **Hepatitis_B**: Represents percent of coverage of Hepatitis B immunization among 1 year olds.
- **Measles**: Represents the percent of first dose of Measles-containing vaccine immunization among 1 year olds.
- **BMI**: BMI (Body Mass Index) represents a measure of nutritional status in adults and is defined as a numerical value calculated based on a person’s weight and height. The exact formula is dividing a person’s weight in kilograms by the square of their height in meters. In this dataset, BMI represents the average BMI of the population in a country.
- **Polio**: Represents the percent coverage of polio immunization among 1 year olds.
- **Diphtheria**: Represents the percent coverage of Diphtheria tetanus toxoid and pertussis immunization among 1 year olds.
- **Incidents_HIV**: Incidents of HIV per 1000 population for those aged 15-49.
- **GDP_per_capita**: The GDP per capita of a country in USD. The GDP per capita in simple terms is calculated by dividing the value of an economy’s GDP by the number of inhabitants.
- **Population_mln**: Total population in millions
- **Thinness_ten_nineteen_years**: Prevalence of thinness among adolescents aged 10-19 years, defined as those with a BMI < -2 standard deviations below the median.

- **Thinness_five_nine_years**: Prevalence of thinness among adolescents aged 5-9 years, defined as those with a BMI < -2 standard deviations below the median.
- **Schooling**: Average years that people aged 25+ spent in formal education.
- **Economy_status_Developed**: Developed Country or not (1 yes, 0 no)
- **Economy_status_Developing**: Developing Country or not (1 yes, 0 no)
- **Life_expectancy**: Average life expectancy in years.