The effect of wealth on Incidence of Malaria Across Developing Countries

---

EC 204 Empirical Economics II - Spring 2020

Profesor Ekaterina Gnedenko

Group 2

Shakthi Iyer, Chris-Emio Raymond & Xenia Vrettakou

Abstract

This paper will discuss the effect of GDP per capita on the incidences of malaria per 1000 people at risk. The main focus of our paper is developing countries in the year 2016, and we collected our data from the World Bank Indicators. Since this is an empirical paper, we will state our null and alternative hypothesis, and will answer the economic problem through statistical measures, using modeled regressions. In addition to the key regressor, GDP per capita, this paper will also explain how other control factors, such as death cause and population density, affect the dependent variable, the incidence of malaria. As the paper progresses, the regression models discussed will be transformed and modified to improve its goodness of fit so it can be the best representation of the data collected.

Outline

1. **Abstract**
2. **Introduction**
3. **Literature Review**
4. **Data Description**
5. **Econometric Model**
    i. *Empirical Analysis*
6. **Summary and Potential Extensions**
7. **References**
8. **Appendices**

2. Introduction

In a world that uncertainty is at its peak due to the recent pandemic adhering with the name COVID-19, the leaders of the developing world are trying to see the magnitude of the implications that another pandemic will potentially bring to their country's economy. Developing countries are still up to today facing the leftovers of older communicable diseases such as malaria, diarrhoeal disease, HIV/AIDS, and tuberculosis, which until today are listed by WHO as the top 10 causes of death in low-income countries as of 2018. Thus, we decided as the new economists of the future to see the wealth effect on the incidents of Malaria per 1000 people at risk across developing countries. Malaria is categorized under the type of communicable disease, as stated by the World Health Organization (WHO)[1]. According to the Center for Disease Control and Prevention (CDC), Malaria is a mosquito-borne disease that transmits a parasite to those infected, allowing the parasite to feed on humans[2]. Such disease has specific symptoms that can cause severe illness, and if not treated on time can lead to death. GDP per Capita is known as the measure of a country's economic output divided by its total population[3]. Looking at the worldwide effects of Malaria and its high infection rate among developing countries, we decided to look at the slowdown period of the world's incidences of Malaria per 1000 people at risk by focusing our data collection on the 2016 fiscal year. This was primarily because that year had enough available data, which is essential when conducting our regression.

[1] "Communicable Diseases (CDS)." *World Health Organization*, World Health Organization, 5 Mar. 2018, www.who.int/about/structure/organigram/htm/en/.
[2] "CDC - Parasites - Malaria." *Centers for Disease Control and Prevention*, CDC, 1 Apr. 2020, www.cdc.gov/parasites/malaria/index.html.
[3] Chappelow, Jim. "Per Capita GDP Definition." *Investopedia*, Investopedia, 29 Jan. 2020, www.investopedia.com/terms/p/per-capita-gdp.asp.

The current COVID-19 global pandemic fostered our topic of interest, and we wanted to see how a measure of economic growth can impact it. We will be focusing on a pervasive and deadly disease that developing countries have struggled with in the past and try to see its economic impacts. By the end of this paper, we aim to relate our results with the potential consequences that COVID-19 will have in the developing world. However, it is essential to mention that although both diseases fall under the communicable diseases umbrella, there is a crucial difference in the way that each of them spreads. We will be taking that into account when making our impact comparisons.

We have decided to use GDP per Capita as our key regressor because we want to see the economic effect that it has on the incidences of Malaria, as it has a vast pool of variables that can be explained by its effect when placed along with them. We expect there to be a negative relationship between GDP per Capita and the incidences of Malaria. Thus, when GDP per Capita increases, there is an inverse effect on the number of malaria incidences per 1000 people. We have concluded in our economic hypothesis for our regression model based on the key regressors' true population coefficient that the null hypothesis is greater than or equal to zero $(H_0 : a_1 \geq 0),$ and we want to prove that it is the opposite by stating our alternative hypothesis of the population parameter coefficient to be less than zero $(H_1 : a_1 < 0)$. We believe that a reduction in the malaria incidences will show better results on the Gross Domestic Product on average for the developing countries that we have accounted for in our project.

In regards to the hypothesis in this empirical paper, we use the data in GDP per capita across 77 developing countries, as a proxy for wealth. We test the statistical hypothesis around the statistical significance level of the coefficient of GDP per capita. We will then be able to

explain the connection between the question and hypothesis. The coefficient must be statistically significant between GDP per capita terms and the incidence of malaria. Moreover, this paper will explore the direction of the effect, whether its negative or positive, by running a one-tail hypothesis test of the statistical significance of the coefficient.

3. Literature Review

Health is one of the most critical factors in which governments are trying to improve to reduce death rates by communicable diseases. In terms of malaria and its effect on different aspects of developing nations' economies, we investigated two separate pieces of research conducted in the past by experts in the field of development economics. Their research falls under the same classification as ours, but they have used different regressors and methods of measurement for the variable of malaria cases.

Starting with the first research by “Saurabh C. Datta and Jeffrey J. Reimer” (2013), they had two different equations, and they ran six different regressions for each econometric equation. The first regression had GDP per capita at purchasing power parity (PPP) as the dependent variables and Malaria cases per million population as the key regressor, and for the second, they swapped the dependent and independent variables. We will focus on the second one as in our research, we also use Malaria incidences but per 1000 people as our dependent variable, and we later discover in our research that we also experience a reverse causality issue between our dependent variable and our key regressor. They discovered that their result in terms of elasticity is a one percent increase in income per capita (holding other things constant) leads to a decrease of 0.897% in Malaria cases per million. Meaning, they also found a negative relationship between GDP per capita and Malaria cases as we do in our attempt to find the relationship

between the two variables. Their variables were all statistically significant at the five percent and the one percent significance levels, and they used an OLS three-stage least squares (3SLS) method because it was consistent and efficient compared to the (2SLT); which was not efficient and did not meet their criteria. They concluded that after accounting for simultaneity and incidental associations, they found that a one percent increase in income per capita leads to a 1.1% decrease in Malaria cases, which is pretty close to their original estimation before accounting for all the other problems.

Moreover, the secondary research was a little bit different but very interesting as it shows a different perspective of how malaria can affect income per capita. This research conducted by Hoyt Bleakley (2010) looked at how malaria eradicated regions of the Americas, but the difference she focuses on specific ages for the victims of malaria. As her key regressor, she used exposure from childhood on malaria cases and not the incidences of malaria on the entire population. Her results were pretty significant as she found that the persistent childhood malaria infection leads to a 50% reduction in adult income. The perspective used by Bleakley to see how the historical effects of malaria exposure during childhood affect the future career of individuals is reflected in their GDP per capita estimates. However, it is important to note for both pieces of research collected data for their malaria cases from the 1990s and 1980s. This encouraged us to use more recent and more accurate data than the ones they used. Lastly, it is as equally essential to mention that both pieces of research raise the issue of reverse causality as they have placed their dependent variables and key regressors in the reverse order compared to ours, a problem that we also discover later in our research and we try to account for.

4. Data Description

For this economic research, our primary focus is on developing countries. We gathered our data from the World Bank Database. To decide what classifies as a developing country, we looked at income levels. The World Bank's threshold for developing countries is a GDP per capita equal to or lower than $12,000.

We will be looking at cross-sectional data focusing on the year of 2016 across developing countries. We assembled 134 observations of developing countries(Figure 1), but after we ran the summarize if e(sample) command, we ended up with 77 observations (Table 1). However, having 77 observations should not be a problem because, according to Peter Kennedy, to satisfy the CLT and be sufficient to use as an approximation for the 2SLS regression, there needs to be a minimum of 30 observations. Our dependent variable is the incidence of Malaria (per 1000 people at risk), and the key regressor of our independent variables is GDP per capita (in current US dollar). In addition to the key regressor, we also included a few control variables such as; cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (percentage of total death), population density (people per sq. km of land area). Included here is a brief mention of our instrumental variables, but it will be discussed in more depth later in this paper. Our instrumental variable is access to electricity (percentage of population) and the density population, as mentioned earlier.

Table 1 : Summary Statistics

```
. sum Malaria GDP_pc Death_Cause Access_Electr Density_pop lnGDP_pc lnMalaria if e(sample)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Malaria	77	111.0374	142.5816	.01	554.51
GDP_pc	77	2627.194	2438.139	282.1931	9817.741
Death_Cause	77	39.45584	18.36681	7.9	65.3
Access_Ele~r	77	63.3417	28.81237	9.298458	100
Density_pop	77	117.7852	171.6794	2.864168	1213.573
lnGDP_pc	77	7.445077	.9612117	5.642591	9.191946
lnMalaria	77	2.630587	2.974597	-4.60517	6.318085

As depicted in Table 1, the highest number of observed cases of Malaria (labeled as Malaria in the table) incidences is in Rwanda with 554.51 cases and looking at the average number of cases being around 111. The lowest number with malaria incidences is 0 as of 2016 from the following countries: Sri Lanka, Azerbaijan, Paraguay, Algeria, Georgia, Turkey, Tajikistan, Iraq, Uzbekistan. El Salvador is the closest non-zero value at 0.01. Remember those numbers are per 1000; therefore, 0.01 is 10 cases in actuality. With that perspective in mind, the high number of cases in Rwanda is very alarming. For our key regressor (GDP_pc), the highest value is $11,666.46 in Costa Rica and the lowest $282.1931 in Burundi.

Taking a closer look into our data set and the scatter plot between the incidence of malaria and GDP per capita (Figure 2) we observed that our values skewed to the right which means that the lowest income countries among the developing countries that we have accounted for in our data set, seem to have the highest incidents of malaria per 1000 people. We also constructed a graph matrix (Figure 3) that depicts scatter plots among all of our variables to detect multicollinearity problems and other linear relationships. We further used the correlation command output (Table 8) as an additional test for multicollinearity.

5. Econometrics model

$Y = a_0 + a_1 \bullet X_1 + a_2 \bullet X_2 + a_3 \bullet X_3 + \varepsilon$

Y - Incidence of malaria per 1000 people at risk

X_1 - Gross Domestic product per capita (current US $)[4]

X_2 - Cause of death by communicable disease

X_3 - Population density (people per sq. km of land area)

X_4 - Access to electricity

ε - Error term

All the variables are continuous. Our first model was a simple regression model with no transformations of our data. However, the R^2 did not explain a big part of malaria, and the SER value was relatively large. Additionally, after looking at the scatterplot of our key regressor against the dependent variable, we noticed that the relationship seemed to be hyperbolic, a form of a log-log model. In models 2-5, we used the log transformation of our key regressor and our other variables. We transformed the variables goodness of fit of our first model was not satisfactory enough for our research. In our final model, we decided to use the log transformation of the dependent variable, incidence of malaria, and the key regressor, GDP per capita. We expect the coefficient of the key regressor to be negative, as it tends to be a negative relationship between GDP per capita and incidence of malaria. We are also under the knowledge that a more developed country with a higher GDP with capita would have fewer cases of malaria compared

[4]It is an endogenous variable and we later instrument it using Access to Electricity and Density Population

to countries with a lower GDP per capita. This can be due to healthcare expenditure and access to appropriate and affordable healthcare.

i) Empirical Analysis

Table 2: Models of Malaria Incidents

Models of Malaria Incidents

Variable	model1	model2	model3	model4	model5
GDP_pc	-0.020				
	0.005				
	0.000				
lnGDP_pc		-1.827	-0.932	2.869	-1.129
		0.286	0.327	7.076	0.495
		0.000	0.006	0.686	0.025
lnDeath_Ca~e			2.601	8.745	
			0.524	5.564	
			0.000	0.120	
lnDensity_~p			-0.352		
			0.190		
			0.067		
lnGDP_pc2				-0.057	
				0.341	
				0.867	
lnGDP_pcln~e				-0.726	
				0.688	
				0.295	
Death_Cause					0.081
					0.021
					0.000
_cons	157.126	16.253	1.812	-27.542	7.831
	20.242	2.151	4.260	36.332	4.370
	0.000	0.000	0.672	0.451	0.077
N	88	78	77	78	77
rmse	127.625	2.403	1.990	2.019	2.019
r2	0.155	0.349	0.570	0.558	0.551
r2_a	0.145	0.340	0.552	0.534	0.539
F	15.781	40.671	32.246	23.051	47.038

legend: b/se/p

Malaria = 157.12 - 0.0204 • GDP_pc (Model 1)

In the first model (Table 3), we chose GDP per capita to be our key regressor X_1 and incidents of Malaria, our dependent variable Y. For our first model, we constructed a simple

linear regression aiming to see the effect of GDP per capita on the Incidents of Malaria. The estimated coefficient of GDP per capita is -0.020, which means for an additional dollar on GDP per capita, there is a decrease in the incidence of malaria by 0.020 per 1000 people. Since the slope coefficient is negative, it represents a negative relationship between the incidence of malaria and GDP per capita, a relationship that we have predicted in our original hypothesis. Moreover, the p-value of our key regressor is 0.000, making it statistically significant at both 5% and 1% significance levels. Thus, we can reject the null hypothesis that the population parameter of a_1 is greater than or equal to zero ($H_0 : a_1 \geq 0$). Similarly, we get a t-statistic of -3.97, which proves to be statistically significant as it is higher than the critical value (-1.64) of the left-sided test and, as expected, leads us to the same conclusion in rejecting the null.

Looking at the goodness of fit measures, the squared residuals of the regression (SER), in model 1, is equal to 127.6, a number that is considered very high as this model suffers from omitted variable bias. Such bias takes place when variables relevant to the regression are excluded from the regression; hence it is violating the first OLS critical assumption. This led to the estimated slope coefficients to be biased and insufficient. This high value alarms us about the goodness of fit of the model, as there is a high discrepancy between the values and the estimated regression model. The R^2 is equal to 0.1550, meaning 15.5% of the variation in the incidence of malaria is explained by GDP Per Capita in the regression. The SER is high because we were under the assumption that there is a linear relationship between the incidence of malaria and GDP Per Capita. However, after looking at the scatter plot for GDP Per Capita against the incidence of malaria (Figure 2), we came to notice that the relationship between the two variables is not linear and is in the form of a log-log relationship.

lnMalaria = 16.25341 - 1.827003 • lnGDP_pc (Model 2)

Hence, we conducted the simple nonlinear regression model in a log-log form as our second model (Table 4), after transforming the dependent variable, Incidents of Malaria, and independent variable, GDP Per Capita, into log formats. The estimated coefficient of GDP per capita is -1.83, which means that for a 1% increase in GDP per capita, there is a decrease in the incidence of malaria by 1.83% per 1000 people. Moreover, the p-value of our key regressor is 0.000, making it statistically significant at both 5% and 1% significance levels. Thus, we can reject the null hypothesis that the population parameter of a_1 is greater than or equal to zero $(H_0 : a_1 \geq 0)$. The rejection of the null hypothesis can be proven by the t-statistic, which is -3.97, as it is statistically significant as it is higher than the critical value (-1.64) of the left-sided one-tailed test.

Looking at the model, we can determine its goodness of fit by the SER and R^2 values. For the second model (Table 4), the SER is 2.40, which shows that the squared error of residuals is small and that the variation of the difference between the true value and the estimated regression line is small. This proves that the log-log model is a better goodness of fit for our data. The R^2 of model 2 is 0.349, which means that 34.9% of the incidence of malaria cases can be explained by GDP Per Capita. In a log-log model the coefficient is the elasticity, which is why economists interpret it as the percent change rather than unit change. Using the log-log model, we can observe a small SER, but we still thought it was necessary to add control variables in order to account for the omitted variable bias and improve the goodness of fit, as will be observed in the next model.

lnMalaria = 1.811 - 0.932 • lnGDP_pc + 2.60 • lnDeath_cause - 0.352 • lnDensity_pop

(Model 3)

To continue improving the goodness of fit of the models, we decided to add the control variables into the third model (Table 5), to reduce omitted variable bias by constructing a multiple nonlinear regression. The two control variables we decided to add on are death cause and density population. This model shows us the effect of GDP per capita on the incidence of Malaria while holding the other variables constant. After doing the two previous regressions, we can conclude that a log-log regression was better for the data, and hence for model 3, we kept the log transformations, and we transformed the control variables into log formats. The estimated coefficient of GDP per capita is -0.93, which means that a 1% increase in GDP per capita would lead to a decrease in incidences of malaria by 0.932% per 1000 people, while holding everything else constant.

Moreover, the p-value of our key regressor is 0.006, making it statistically significant at both 5% and 1% significance levels. Thus, we can reject the null hypothesis that the population parameter of a_1 is greater than or equal to zero $(H_0 : a_1 \geq 0)$. Rejecting the null hypothesis can also be proven by the t-statistic equal to -2.86, which can be said to be statistically significant as it is higher than the critical value (-1.64) of the left tailed test. The same applies to one of the control variables we included lnDeath Cause with p-value 0.000. However, what drove our attention the most is that the variable lnDensity population was insignificant.

While looking at the model's goodness of fit, the R^2 is 0.57, meaning that GDP per capita explains 57% of the malaria incidents, compared to the previous model, the R^2 value has increased, implying that this could be a better model for the data because omitted variable bias

declined, after we added the control variables in the regression. In the third regression model, the square residuals value (SER) is relatively small, with a value of 1.990. In comparison to the previous model, we observe a decrease, showing a better fit, as there are lesser discrepancies between the true population parameter values and the estimated regression line. Although by including the control variables, such as death cause and population density, it accounted for part of the omitted variable bias being present in the previous model and made the regression a better fit for our data, we still believe that the model can be further improved. As the control variable, Density_pop turned out to be insignificant, as mentioned above. Thus, we decided to create a fourth model and try to improve it by dropping the log of population density variable and generating other variables.

$$lnMalaria = -27.54 + 2.86 \cdot lnGDP\_pc - 0.05 \cdot lnGDP\_pc^{2} + 8.74 \cdot lnDeath\_Cause - 0.72 \cdot lnGDP\_pc \cdot lnDeatch\_Cause \qquad \textit{(Model 4)}$$

In the fourth model (Table 6), we constructed a multiple nonlinear regression with all the variables in a log format. We included the interaction term between lnGDP per capita and lnCause of Death. We also raised our key regressor to the second power. This model shows us the effect of GDP per capita on the incidence of Malaria while holding the other variables constant. The estimated coefficient of GDP per capita is 2.86, so for a 1% increase in GDP per capita would lead to an increase in incidences of malaria by 2.86% per 1000 people, while holding everything else constant. This model just by looking at the GDP per capita slope coefficient is not good as we see an increase in malaria incidents when there is a 1% increase in GDP per capita an adverse effect from all the previous models.

Moreover, the p-value of our key regressor is 0.68, which makes it statistically insignificant at the 5% significance level. The same applies to all the other estimated coefficients of the regressors, as they are all statistically insignificant. Thus, we fail to reject the null hypothesis that the population parameter of a_1 is greater than or equal to zero $(H_0 : a_1 \geq 0)$ for every estimated coefficient accounted in model 3.

The R^2 is 0.5581, which means that 55.81% of cases of the incidence of malaria can be explained by GDP Per Capita. The SER is 2.019, which means that this model has a low discrepancy between the estimated regression line and its true data values. After including the interaction term and raising the log transformation of GDP per capita to the second power, we observed that although the measure of best fit SER and R^2 were more or less as good as model 3, that this model is not considered good one due to the statistical insignificance of the slope coefficients.

After running multiple models of regressions, aiming to find the most efficient representation of regressors, we discovered that the regression in Table 5 seems to be our best bet. However, one of our control variables, specifically the estimated coefficient of the lnDensity_pop variable turned out to be insignificant, as it failed to reject the null hypothesis. This led us to think of potential endogeneity problems and a causal relationship between GDP Per Capita and the incidence of Malaria, as their roles could potentially be switched, in other words, the problem of reverse causality. Thus, we decided to run an instrumental variable regression aiming to face the above problems, which is our fifth and last model (Table 7).

In order to find the best instruments that will not directly affect incidences of malaria, but affect it through the key regressor GDP Per Capita, we looked at the graph matrix and

concluded that the Access to Electricity as a percent of the population would be our first instrumental variable. Then due to the failure to reject the null hypothesis at the 5% level of significance in the previous model of the lnDensity population variable, we decided to use it as our second instrument. It does not affect the independent variable directly, and it makes the instrumental variable exogenous. The variable density population does not meet the rigorous definition of the instrumental variable because it is directly affecting the dependent variable, incidences of malaria. As we, unfortunately, observe with the spread of the coronavirus, we can say that the spread of the virus is much more significant in densely populated cities than rural areas, and this situation applies to malaria as well. The two-stage least squares (2SLS) method led us to better results than the regression model 3.

In the first stage, we can see that the instruments Access to electricity and Density population p-values are 0.000 and 0.008, respectively, leading us to the conclusion that is statistically significant with the endogenous variable lnGDP per capita. Also, the overall F-statistic is equal to 43.42, which tests the joint statistical significance of the estimated slope coefficients, and at the 5% significance level, we thus reject the null hypothesis. However, in order for the two instrumental variables to be valid, they should be correlated with the independent variable GDP per capita and should be uncorrelated with the error term. Also, we see that the R^2 is 0.64, indicating this model is a good fit as a lot of the variation is explained by the variables chosen.

In the second stage of Model 5 (Table 7), the estimated slope coefficient of lnGDP per capita is -1.13, which means that for a 1% increase in GDP per capita there will be a 1.13% decrease in the incidents of malaria, while holding everything else constant. Moreover, the

p-value of GDP per capita, is 0.025, which makes it statistically significant at the 5% significance level. Thus, we can reject the null hypothesis that the population parameter of a_1 is greater than or equal to zero $(H_0 : a_1 \geq 0)$. Similarly, the rejection of the null hypothesis can be proven by the t-statistic, -2.28, which can be said to be statistically significant as it is higher than the critical value (-1.64) of the left-sided one-tailed test.

Concerning the model's goodness of fit, the R^2 is 0.5515, which means that lnGDP per capita explains 55.2% incidence of malaria cases. Compared to model 2, we observe that the R^2 of model 5 reduces the bias by 20% just by looking at the R^2 difference of the two models. This implies that the key regressor explains a higher percentage of the independent variable. The SER value is 2.0188, which implies that the squared errors of residuals are quite low and that there are low discrepancies between the true data points and the estimated regression line.

To see the validity of the instruments that we used in our regression, we decided to run some tests. The first one was the instrument exogeneity test (Table 9) that aimed at discovering whether or not the error term is correlated with our two instruments, access to electricity, and population density. According to the results, both of the instruments' p-values are statistically insignificant, which leads us to the failure to reject the null hypothesis. This is a positive effect as it is proof that the instruments are not correlated with the residual. Thus, our assumption about the validity of the instruments is not violated by the exogeneity test because they are proven to be exogenous variables.

In addition to that test, we also run the F-test in order to check if our instruments are weak after running the first stage of instrumental variable regression. Thus, our results showed us an F-statistic is equal to 100.21 (Table 10) which exceeds ten and implies that our instruments

are not at all weak.

We also tried to conduct the Hausman test (Table 11) to see the endogeneity of the regressor lnGDP Per Capita, which will help us see if the OLS regression method is consistent or not with the IV regression method. However, due to our relatively small number of observations as we are focusing on developing countries and a lot of them do not have the financial ability to collect such data, the Hausman test was unsuccessful in our case. Therefore, we have concluded that even though the Hausman test did not work in our case, but the exogeneity test did and provided us with positive results and the F tests too we believe that model 5 is our most robust regression model in testing our hypothesis of the effect of wealth on the incidents of malaria across developing countries. Lastly, keeping in mind all the above, our model is still not perfect, and it has much room for improvement, which the time constraint did not allow us to improve our regression model further.

6. Summary and Potensial Extensions

To summarize our findings, with our 77 cross-sectional observations from 2016, we ran multiple regression models to get the best possible result. We conclude that multiple nonlinear regression shows the best relationship between GDP Per Capita and the incidence of malaria per 1000 people at risk and the rest of the control variables. As we expected, there is overall a negative relationship between our incidents of malaria and GDP per capita. From our instrumental variable (2SLS) regression, we see that the estimated coefficient for log GDP Per Capita is -1.13. It shows that when GDP Per Capita increases by one percent, the incidence of malaria per 1000 of people at risk decreases by 1.13 percent, which is a small percentage. The p-value of the coefficient for lnGDP Per Capita is 0.025, which means there is statistical

significance at the 5% dignificance level between the two variables. As the other pieces of research showed that if we invert the variables, there will still be a negative relationship. Malaria, a communicable disease, has the potential to harm the GDP of developing countries. Using our result, we can infer that COVID-19 may or may not have a similar impact on the GDP of developing countries. Our speculation of the impact of COVID-19 is a bit uncertain because COVID-19, while it is a communicable disease, does not spread the same way as malaria. Malaria's infection rate is greatly affected by the geographical location and spreads through mosquito bites. While COVID-19 can spread globally with ease and is transmitted through close human contact (within six feet). Interacting with contaminated surfaces is likely to transmit the disease to a new person. Also, simply being near an infected person increases the risk of catching it because when they cough, sneeze or talk, they are exposing others to their infected respiratory droplets. If COVID-19 were to have an impact on GDP, it would undoubtedly be of a higher magnitude. Possible extension to our research could be using a communicable disease that spreads similarly to COVID-19 so the conclusion could be more reliable. We could also further our research by using panel data because the relatively small sample size limits us and by using panel data we can decrease the probability of experiencing omitted variable bias.

7. References

1. Bleakley Hoyt., (2010). “Malaria Eradication in the Americas: A Retrospective Analysis of Childhood Exposure” *American Economic Journal*, 2(2), 1-45.
2. “CDC - Parasites - Malaria.” *Centers for Disease Control and Prevention*, CDC, 1 Apr. 2020, www.cdc.gov/parasites/malaria/index.html.
3. Chappelow, Jim. “Per Capita GDP Definition.” *Investopedia*, Investopedia, 29 Jan. 2020, www.investopedia.com/terms/p/per-capita-gdp.asp.
4. “Communicable Diseases (CDS).” *World Health Organization*, World Health Organization, 5 Mar. 2018, www.who.int/about/structure/organigram/htm/en/.
5. Datta, Saurabh C., and Jeffrey J. Reimer., (2013). "Malaria and economic development." *Review of Development Economics,* 17(1), 1-15.
6. Kennedy, P. (2003). *A guide to econometrics*. MIT press.

8. Appendices

Appendix A → Do-file

import excel "/Users/xeniavrettakou/Desktop/Chris_Emio_newdata.xlsx", sheet("Data") firstrow

(5 vars, 134 obs)

// Our Do-File with all of our commands and some explanations to go along with it.

DATA AND TABLES

// We first pulled the data from the World Bank database and turned them into an excel-file

// All of the destring commands above turn the excel file strings to integers so we could manipulate them with mathematical commands.

destring Incidenceofmalariaper1000, generate (Malaria)

destring GDPpercapitacurrentUS, generate (GDP_pc)

destring Populationdensitypeoplepers, generate (Density_pop)

destring Causeofdeathbycommunicable, generate (Death_Cause)

// Using the integers from the excel that we just converted. We created variables for them with the generate command.

generate lnGDP_pc = log(GDP_pc)

generate lnMalaria = log(Malaria)

generate lnGDP_pc2 = lnGDP_pc^2

generate lnDeath_Cause = log(Death_Cause)

generate lnDensity_Pop = log(Density_pop)

*generate lnGDP_pclnDeath_Cause = lnGDP_pc*lnDeath_Cause*

// We generate a scatter plot between our independent and dependent variable before messing with anything else.

scatter Malaria GDP_pc

twoway (fpfit Malaria GDP_pc) (scatter Malaria GDP_pc)

// We generated a summary statistic table between the common variable to describe the data.

sum Malaria GDP_pc Death_Cause Access_Electr Density_pop lnGDP_pc lnMalaria if e(sample)

// We generated a graph matrix to look for linear correlation between our regressors.

graph matrix Malaria GDP_pc Death_Cause Density_pop Access_Electr

MODELS

// The reg command generated a regression according to the variable we imputed

// A simple linear regression

reg Malaria GDP_pc

// A simple nonlinear regression

reg lnMalaria lnGDP_pc

// A Multiple nonlinear regression

reg lnMalaria lnGDP_pc lnDeath_Cause lnDensity_Pop

// Multiples Nonlinear Regression (log-log, Interaction Term and Quadratic) Model

reg lnMalaria lnGDP_pc lnGDP_pc2 lnDeath_Cause lnGDP_pclnDeath_Cause

// The ivreg command generated a first-stage (2SLS) regression for our suspected endogenous variable

ivreg lnMalaria (lnGDP_pc=Access_Electr Density_pop), first

// corr command showed a numerical representation of the correlation between our variables similar to the graph matrix command

corr Malaria GDP_pc Death_Cause Density_pop Access_Electr

// The following commands generate a model for our regressions to we can compile them into one table at the end

quietly reg Malaria GDP_pc

estimates store model1

quietly reg lnMalaria lnGDP_pc

estimates store model2

quietly reg lnMalaria lnGDP_pc lnDeath_Cause lnDensity_Pop

estimates store model3

quietly reg lnMalaria lnGDP_pc lnGDP_pc2 lnDeath_Cause lnGDP_pclnDeath_Cause

estimates store model4

quietly reg lnMalaria Death Cause (lnGDP_pc=Access_Electr Density_pop), first

estimates store model5

// Here is the compiled table of models, a panel data, with their p-value showing their level of significance

estimates table model1 model2 model3 model4 model5, b(%9.3f) star se(%6.3f) stats(N rmse r2 r2_a F) title ("Models of Malaria Incidents")

EXOGENEITY TEST FOR IV

//We tested our instrumental variables for exogeneity. We then conducted an F-test to see the joint statistical significance of our instruments.

predict uhat, residuals

reg uhat Access_Electr Density_pop Death_Cause

test Access_Electr Density_pop

// We then used the Haufsman Test to look for endogeneity of the regressors and examined the difference between the OLS and IV regression estimates of the representative coefficients (unsuccessful).

reg lnMalaria lnGDP_pc lnDensity_pop lnDeath_Cause

estimates store ols

ivreg Malaria Death_Cause (GDP_pc= Access_Electr Density_pop), first

estimates store ivreg

hausman ivreg ols, constant sigmamore

Appendix B

Figure 1: Developing Countries our Number of Observations

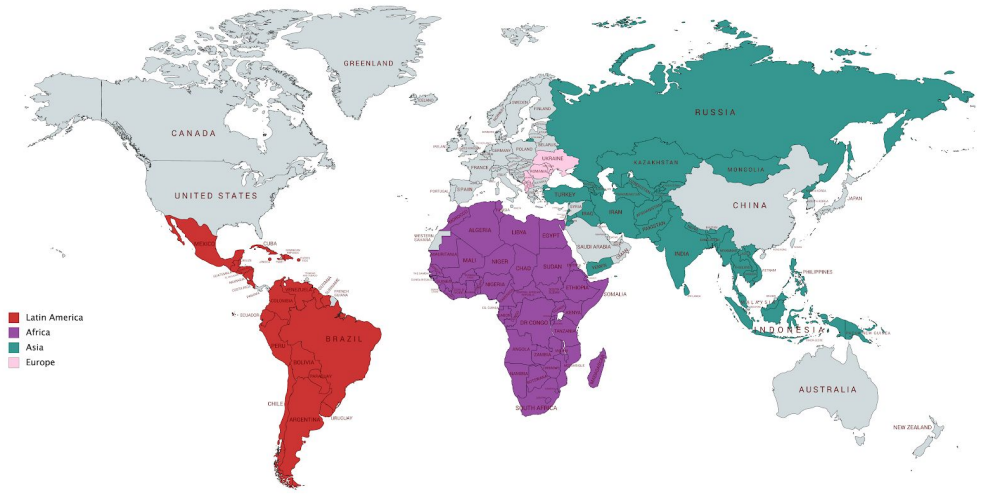


Table 3: Simple Linear Regression Model

```
. reg Malaria GDP_pc

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(1, 86)        =     15.78
       Model |  257041.699         1  257041.699   Prob > F        =    0.0001
    Residual |  1400773.17        86  16288.0601   R-squared       =    0.1550
-------------+----------------------------------   Adj R-squared   =    0.1452
       Total |  1657814.87        87  19055.3433   Root MSE        =    127.62

------------------------------------------------------------------------------
     Malaria |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      GDP_pc |  -.0204339   .0051438    -3.97   0.000    -.0306595   -.0102084
       _cons |   157.1259   20.24214     7.76   0.000     116.8859    197.366
------------------------------------------------------------------------------
```

We constructed a simple regression model based on our dependent and independent variables. Incidences of malaria per 1000 cases versus GDP Per Capita, there are no linear transformations included in this model. Since the slope coefficient is negative, it represents a negative relationship between the incidence of malaria and GDP per capita.

Table 4: Simple Nonlinear Regression (Log-Log) Model

```
. reg lnMalaria lnGDP_pc

      Source |       SS           df       MS      Number of obs   =        78
-------------+----------------------------------   F(1, 76)        =     40.67
       Model |  234.760276         1  234.760276   Prob > F        =    0.0000
    Residual |  438.680388        76  5.77211037   R-squared       =    0.3486
-------------+----------------------------------   Adj R-squared   =    0.3400
       Total |  673.440664        77  8.74598265   Root MSE        =    2.4025

------------------------------------------------------------------------------
   lnMalaria |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnGDP_pc |  -1.827003    .28648    -6.38   0.000    -2.397577   -1.256429
       _cons |   16.25341   2.151373     7.55   0.000     11.96858    20.53824
------------------------------------------------------------------------------
```

After looking at the scatter plot for GDP per capita against the incidence of malaria, we came to notice that the relationship between the two variables resembles a log-log relationship. Hence, we conducted the same simple regression model but as a log-log model for this table. It is a simple nonlinear regression still only using our dependent and independent variable. This did prove that the log-log model has better goodness of fit for our data, and it still shows a negative relationship between GDP Per Capita and malaria cases.

Table 5: Multiple Nonlinear Regression (log-log) Model

```
. reg lnMalaria lnGDP_pc lnDeath_Cause ln_Density_Pop

      Source |       SS           df       MS      Number of obs   =        77
-------------+----------------------------------   F(3, 73)        =     32.25
       Model |  383.254615         3  127.751538   Prob > F        =    0.0000
    Residual |  289.210639        73  3.96178957   R-squared       =    0.5699
-------------+----------------------------------   Adj R-squared   =    0.5523
       Total |  672.465254        76  8.84822702   Root MSE        =    1.9904

------------------------------------------------------------------------------
   lnMalaria |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnGDP_pc |  -.9324825   .3265403    -2.86   0.006    -1.583276   -.2816885
lnDeath_Cause |    2.60099    .523964     4.96   0.000     1.556731    3.645248
ln_Density_Pop |  -.3523109   .1897396    -1.86   0.067    -.7304613    .0258396
       _cons |   1.811682   4.259845     0.43   0.672    -6.678177    10.30154
------------------------------------------------------------------------------
```

After accounting for the nonlinear shape of our data and implementing the log-log

transformation on the dependent variable and the key regressor, we realized that still, our simple nonlinear model was not perfect. We realized that there is a high possibility that we are facing an omitted variable bias problem. Thus, we decided to add two more independent variables as our control variables after transforming them to log variables, and these are the cause of death and density population variables. We then composed a multiple nonlinear regression, which we see that there was an improvement on the goodness of fit according to the R^2 estimation, something that shows an improvement on the omitted variable bias identified in the previous model. Also, we can see that the p-values of the estimated coefficients are statistically significant except the ln cause of death variable.

Table 6: Multiples Nonlinear Regression (log-log, Interaction Term and Quadratic) Model

```
. reg lnMalaria lnGDP_pc lnGDP_pc2 lnDeath_Cause lnGDP_pclnDeath_Cause
```

Source	SS	df	MS	Number of obs	=	78
				F(4, 73)	=	23.05
Model	375.865423	4	93.9663557	Prob > F	=	0.0000
Residual	297.575241	73	4.07637316	R-squared	=	0.5581
				Adj R-squared	=	0.5339
Total	673.440664	77	8.74598265	Root MSE	=	2.019

lnMalaria	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnGDP_pc	2.868589	7.076272	0.41	0.686	-11.2344	16.97158
lnGDP_pc2	-.0572429	.3406779	-0.17	0.867	-.7362129	.6217272
lnDeath_Cause	8.744861	5.563608	1.57	0.120	-2.343394	19.83311
lnGDP_pclnDeath_Cause	-.725545	.6884281	-1.05	0.295	-2.09758	.6464902
_cons	-27.54227	36.33236	-0.76	0.451	-99.95255	44.86801

After we decided to keep all of our variables in their ln form, we decided to see if we could further improve the model because the ln cause of death coefficient was insignificant. We believe that we might be facing some multicollinearity problems, and that is the explanation

behind the lnDeath_Cause variable. Thus we decided to raise the key regressor to the second power and add an interaction term between the transformations of GDP per capita and Death Cause. This model did not turn out good at all as it made all of the estimated coefficients p-values insignificant.

Table 7: Instrumental Variable Regression (log-log) 2SLS Model

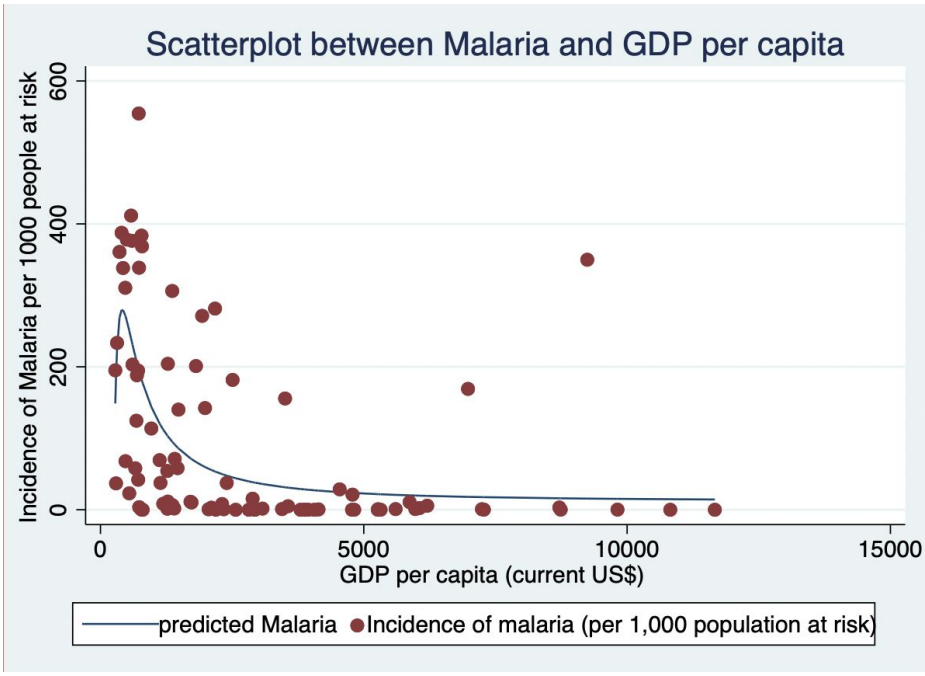
```
. ivreg lnMalaria Death_Cause (lnGDP_pc=Access_Electr Density_pop), first

First-stage regressions
-----------------------

      Source |       SS           df       MS      Number of obs   =        77
-------------+----------------------------------   F(3, 73)        =     43.41
       Model |  44.9957032         3  14.9985677   Prob > F        =    0.0000
    Residual |  25.2228231        73  .345518124   R-squared       =    0.6408
-------------+----------------------------------   Adj R-squared   =    0.6260
       Total |  70.2185263        76  .923927978   Root MSE        =    .58781

------------------------------------------------------------------------------
      lnGDP_pc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
  Death_Cause |  -.0029052   .0064497    -0.45   0.654    -.0157594    .0099489
Access_Electr |   .0248395   .0040759     6.09   0.000     .0167162    .0329628
  Density_pop |  -.0010807   .0003988    -2.71   0.008    -.0018756   -.0002858
        _cons |   6.113625   .5018077    12.18   0.000     5.113524    7.113727
------------------------------------------------------------------------------
```

```
Instrumental variables (2SLS) regression

      Source |       SS           df       MS      Number of obs   =        77
-------------+----------------------------------   F(2, 74)        =     47.04
       Model |  370.863747         2  185.431874   Prob > F        =    0.0000
    Residual |  301.601506        74  4.07569603   R-squared       =    0.5515
-------------+----------------------------------   Adj R-squared   =    0.5394
       Total |  672.465254        76  8.84822702   Root MSE        =    2.0188

------------------------------------------------------------------------------
   lnMalaria |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnGDP_pc |  -1.129317   .4950869    -2.28   0.025      -2.1158   -.1428351
 Death_Cause |   .0812974   .0207408     3.92   0.000     .0399703    .1226244
       _cons |   7.830785   4.370244     1.79   0.077    -.8771166    16.53869
------------------------------------------------------------------------------
Instrumented:  lnGDP_pc
Instruments:   Death_Cause Access_Electr Density_pop
------------------------------------------------------------------------------
```

In the table above, we conducted a 2SLS instrumental variable regression, that assesses the exogeneity of the instrumental variables on the regression. By doing this, we wanted to improve the model further by including more variables in the regression.

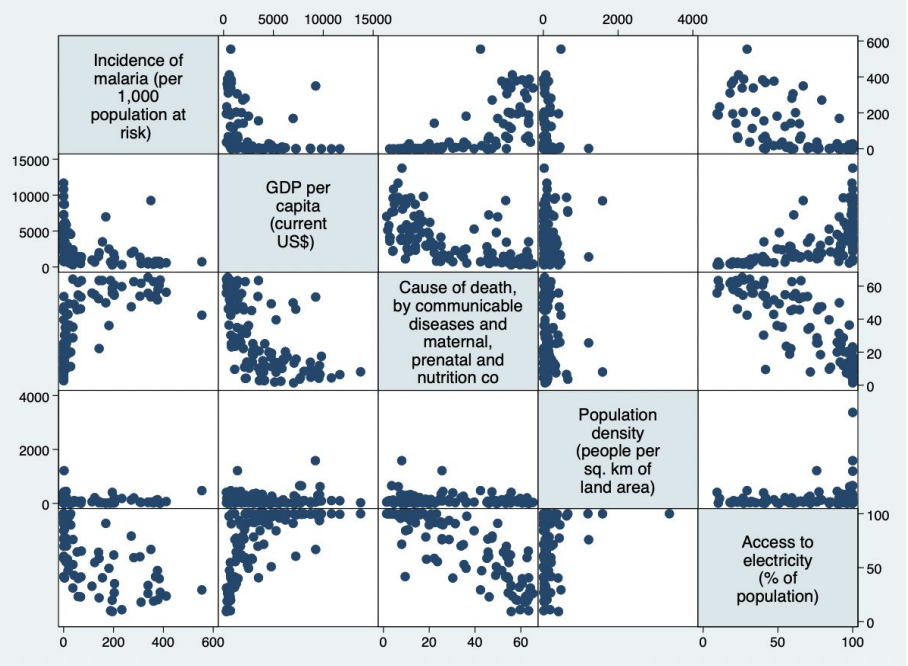
Appendix C

Figure 2: Scatter Plot



We generated a scatterplot between our dependent variable the incidence of Malaria and the key regressor GDP Per Capita to see what is the relationship between them. We discovered that the line of best fit looks like a hyperbola, so we proceeded by using the log-log transformation model. There are two extreme values that we decided to keep in our regression model because we still believe they are relevant to our case and hypothesis.

Figure 3: Graph Matrix



We constructed the graph matrix aiming to help us see the different relationships between our variables and detect any linear relationships or correlations between our regressor. We used it throughout the process of our research and especially while deciding which instruments to use in our instrumental variable regression.

Table 8: Correlation among Variables

```
. corr Malaria GDP_pc Death_Cause Density_pop Access_Electr
(obs=87)

             |  Malaria   GDP_pc Death_~e Densit~p Access~r
-------------+---------------------------------------------
     Malaria |   1.0000
      GDP_pc |  -0.3953   1.0000
 Death_Cause |   0.6731  -0.5721   1.0000
 Density_pop |  -0.0581  -0.1596  -0.1134   1.0000
Access_Ele~r |  -0.6856   0.6273  -0.8428   0.0401   1.0000
```

This table helps provide additional context to the graph matrix. It shows the correlation between every pair of independent variables. There is no coefficient high enough, which means higher than 0.8, to suggest that there is significantly influential multicollinearity in our model.

Appendix D

Table 9: Exogeneity Test for Instrumental Variables

```
. predict uhat, residuals
(56 missing values generated)

. reg uhat Access_Electr Density_pop Death_Cause
```

Source	SS	df	MS	Number of obs	=	77
				F(3, 73)	=	1.30
Model	15.2600698	3	5.08668994	Prob > F	=	0.2820
Residual	286.341434	73	3.9224854	R-squared	=	0.0506
				Adj R-squared	=	0.0116
Total	301.601504	76	3.96844085	Root MSE	=	1.9805

uhat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Access_Electr	-.0129415	.0137331	-0.94	0.349	-.0403115	.0144286
Density_pop	-.00244	.0013438	-1.82	0.074	-.0051183	.0002382
Death_Cause	-.0199042	.0217312	-0.92	0.363	-.0632143	.0234059
_cons	1.892472	1.690762	1.12	0.267	-1.477213	5.262157

The table above represents a test conducted to assess the exogeneity of the instrumental variables. Exogeneity of the instrument variables means that the variables are uncorrelated with the error term, and as a result is also uncorrelated with residuals. We noticed that all the variables are statistically insignificant.

Table 10: Instrumental Weakness Test via F-Test

```
. test Access_Electr Density_pop

 ( 1)  Access_Electr = 0
 ( 2)  Density_pop = 0

       F(  2,   122) =  100.21
            Prob > F =    0.0000
```

This table represents the test conducted to help us understand the statistical significance of the estimated coefficients of the instrumental variables used in this project.

Table 11:Hausman Test

```
. hausman ivreg ols, constant sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     ivreg         ols         Difference          S.E.
-------------+----------------------------------------------------------------
       _cons |   81.04049     1.811682        79.22881               .
------------------------------------------------------------------------------
                        b = consistent under Ho and Ha; obtained from ivreg
         B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                        =  -390.86    chi2<0 ==> model fitted on these
                                      data fails to meet the asymptotic
                                      assumptions of the Hausman test;
                                      see suest for a generalized test
```

The table above represents the Hausman test that was conducted to examine the difference between the LS and IV coefficient estimates, to determine the exogeneity of the instrumental variables.