Proceedings of CONF-MPCS 2024 Workshop: Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations

DOI: 10.54254/2753-8818/51/2024CH0195

Analysis of the potential correlation between infant mortality rate and GDP per capita among 217 countries

Zeyu Li

College of Art and Science, Boston University, Massachusetts, 02215, United States

lizeyu@bu.edu

Abstract. This research explores the connection between GDP per capita and infant mortality rates (IMR) by examining data collected from 217 countries between 2000 and 2022. Regression analysis is utilized to examine the associations among different factors, encompassing healthcare, education, and unemployment, previously identified as influencing the infant mortality rate (IMR). The analysis reveals an inverse relationship between GDP per capita and IMR. Specifically, one dollar growth in GDP per capita corresponds to a 0.107 reduction in infant death, with all other variables held constant. The study also examined interaction terms, revealing a positive correlation between certain variable combinations and infant mortality. These findings indicate that policymakers should prioritize improving reproductive education and allocating more funding to healthcare in order to further decrease infant death. Nevertheless, the study has constraints, such as the absence of certain observations and the necessity for supplementary social and economic factors to acquire a more comprehensive understanding of the components impacting infant death.

Keywords: Infant mortality, GDP per capita, multiple linear regression, interaction effects.

1. Introduction

Although humanity has made progress in medical care, concerns related to public health continue to exist in the 21st century. Infant death is one of the most significant worldwide concerns. Recently, numerous researchers have been investigating the potential factors that influence infant mortality, particularly in low-income developing nations. Khan et al. utilized a frailty model on panel data from Bangladesh and found that infant sex, maternal age at birth, and parental education are significant drivers of this risk [1]. Baraki and Akalu found that newborn sex, multiparity, and residency strongly influenced infant mortality [2]. Using parametric survival models on a group of over 200,000 kids in seven nations in sub-Saharan Africa, Simmons et al. determined that the delivery fees and the concentration of private institutions affect infant mortality [3].

Even in developed nations, infant mortality cannot be neglected, because it not only contributes to distinct causes compared to developing countries but is also an indicator of underlying societal structural problems. For instance, diarrhoea causes a distinct seasonal pattern in infant mortality in the U.S. [4]. Premature birth rates also significantly influence infant mortality. Despite this, Bairoliya et al. used multivariate logistic and random effects models to discover that full-term infants in the United States still show a higher risk of death due to Sudden Infant Death Syndrome and asphyxiation relative to European countries [5]. Furthermore, infant mortality rate (IMR) might serve as indicators of societal

DOI: 10.54254/2753-8818/51/2024CH0195

concerns that extend beyond public health. Marked variations exist in infant death rate across racial and ethnic groups in the United States [6]. The non-Hispanic Black community experiences an infant mortality rate of 10.8, whereas the rate is 9.4 for the Native Hawaiian and Other Pacific Islander population [6]. The Indian population has an IMR of 8.2 [6]. Over more than a century, these inequalities have consistently grown [7]. This implies that medical and social advancements have not equally benefited individuals of all races and ethnicities [7].

Infant mortality rates are impacted by various factors on a global level, such as healthcare, education, employment, social structure, and other variables. Researchers have found a positive association between adolescent motherhood and maternal mortality with baby mortality [8]. Owusu et al. employed panel quantile regression with bootstrapping to analyse data from 177 countries spanning the years 2000 to 2015. Their investigation revealed a correlation between increased healthcare expenditure and a reduction in newborn mortality rates, which varied from 0.19% to 1.45% [9]. Furthermore, according to Zakir and Wunnava, fertility, female labour force involvement, and gross national product (GNP) per capita have a notable influence on infant mortality [10]. The allocation of government funds towards healthcare did not exert a significant influence on infant mortality rates [10].

This study initially examines the potential relationships between the infant mortality rate and various factors, including GDP per capita, education, healthcare, resource accessibility, and unemployment. The World Bank data is used to create empirical regression models. Multiple linear regression and interaction terms are used.

2. Methodology

2.1. Data source

This study uses panel data at the national level spanning from 2000 to 2022. Data has been gathered for each variable from 217 nations, resulting in a total of 4995 observations. All data are sourced from the World Bank Open Data.

2.2. Variable selection

The dataset contains seven independent variables and one dependent variable (infant mortality rate). The rate of infant mortality is measured by counting deaths among infants for every 1,000 live births. The GDP per capita is calculated using a fixed value of 2015 US dollars to avoid the bias caused by inflation. The variable logograms and meanings are shown in Table 1.

Table 1. List of Variables

Variable	Logogram	Meaning
Mortalityrate	Y	The rate of infant deaths (per 1,000 live births)
GDPpercapita	X1	Gross domestic product per person (adjusted to 2015 US$)
Eduspending	X2	Total public education expenditure (% of GDP)
Healspending	X3	Current health expenditure per capita (current US$)
Unempfemale	X4	Unemployment, female (% of female labour force) (modelled ILO estimate)
Accesscooking	X5	People using clean energy sources and technologies for cooking (% of population)
Acceselectri	X6	Availability of electrical power (% of population)

The summary statistics are shown below in table 2.

DOI: 10.54254/2753-8818/51/2024CH0195

Table 2. Summary Statistics

VARIABLES	mean	sd	min	max
Y	27.33	25.42	1	138.6
X1	15,283	22,478	255.1	228,668
X2	4.447	1.879	0.127	15.59
X3	916.6	1,600	4.448	11,702
X4	9.381	7.362	0.172	42.55
X5	63.11	39.26	0	100
X6	80.94	29.23	0.796	100

2.3. Method introduction

The paper employs a multiple linear regression model to contrast the scenario when interaction terms are taken into account versus when they are not. The primary objective of this section is to assess the relative importance of the two models and evaluate the precision of the outcomes. Eventually, it will enable the streamlined processing of models. The multiple linear regression model is a statistical model that includes multiple independent variables to predict the dependent variable. The general form of multiple linear regression is expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + e \quad (1)$$

The multiple linear regression is employed to elucidate the potential linear correlation between the dependent variable and independent variables. Furthermore, the fundamental concept of this method is to calculate a group of parameters using Ordinary Least Squares (OLS) in order to minimize the sum of squared differences between the dependent and independent variables.

3. Results and discussion

3.1. Multiple linear regression

The analysis reveals that numerous factors exert influence on IMR. As the table 3 shows:

Table 3. Relevance Analysis Between Dependent and Independent Variables

	Y	X1	X2	X3	X4	X5	X6
Y	1						
X1	-0.477**	1					
X2	-0.231**	0.061**	1				
X3	-0.462**	0.811**	0.224**	1			
X4	-0.067**	-0.166**	0.116**	-0.132**	1		
X5	-0.800**	0.504**	0.214**	0.469**	0.128**	1	
X6	-0.836**	0.401**	0.173**	0.368**	0.130**	0.856**	1

Table 3 displays the Pearson correlation coefficient between these independent variables and the infant mortality rate. The research data indicates that access to electricity has the strongest positive correlation with IMR. Access to cooking inversely impacts the IMR. Even though spending on education and healthcare is inversely correlated with the result, these factors are not as impactful as those discussed earlier. Contrary to expectations, there is an inverse relationship between female unemployment and infant death. After analysing the Pearson correlation, a multiple regression was conducted (table 4).

DOI: 10.54254/2753-8818/51/2024CH0195

Table 4. Regression result

	Unstandardized Coefficients		Standardized Coefficients	T	P	Covariance Diagnosis	
	B	Std. Error	Beta			VIF	Tolerance
Constant	81.391	0.901	-	90.292	0.000**	-	-
X1	-0.000	0.000	-0.107	-4.944	0.000**	6.089	0.164
X2	-1.191	0.137	-0.083	-8.694	0.000**	1.188	0.842
X3	-0.000	0.000	-0.021	-1.005	0.315	5.529	0.181
X4	0.070	0.037	0.018	1.911	0.056	1.125	0.889
X5	-0.152	0.013	-0.235	-11.429	0.000**	5.500	0.182
X6	-0.485	0.015	-0.584	-31.687	0.000**	4.409	0.227
R^2	0.793						
Adj R^2	0.793						
F	F (6,2689)=1718.286,p=0.000						
D-W	0.208						

The p-values for four factors-gross domestic product per person, education spending, access to cooking, and access to electricity-were below 0.01. Thus, these four factors exert a substantial influence on the dependent variable Y. Health spending and female unemployment have p-values of 0.315 and 0.056 respectively, suggesting that the effect of these two factors on infant mortality is not substantial. Using the data provided, the appropriate multiple linear regression equation can be derived as:

$$y = 81.391 - 0.107x_1 - 0.083x_2 - 0.021x_3 + 0.018x_4 - 0.235x_5 - 0.584x_6 \quad (2)$$

The model yields a R-squared of 0.793, and the adjusted R-squared is 0.793, meaning that all factors mentioned above can explain 79.3% of the variation in the IMF. Additionally, the validity is confirmed, given that the model passes the F-test F (6,2689) =1718.286, p=0.000.

3.2. Interaction term model

Interactions among certain independent variables can influence IMR, and their interactions are referred to as interaction terms. The GDP per capita is likely related to female unemployment, education spending, and health spending. In other words, the effect of GDP per capita on IMR depends on the values taken by unemployment, education spending, and health spending. Similarly, education spending is probably linked to female unemployment. To solve the problem, the interaction terms are multiplied and then added to the equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_1 x_4 + \beta_{10} x_2 x_4 + e$.

This equation represents a linear regression model with multiple variables. The regression coefficients are denoted as β_i, where i ranges from 1 to 10. The terms $x_1 x_2$, $x_1 x_3$, and $x_2 x_4$ represent the interaction between variables.

If the regression coefficients for the interaction term are significant and negative, it suggests that a larger GDP per capita is likely to lead to lower infant deaths when there is an increase in access to electricity and access to clean cooking resources.

The regression coefficient for x_1 is $\beta_1 + \beta_7 X_2 + \beta_8 X_3$. Therefore, the significance of β_1 alone does not indicate whether the overall effect of x_1 on y is significant. The result is shown in Table 5.

DOI: 10.54254/2753-8818/51/2024CH0195

Table 5. Multiple Linear Regression Model analysis results with interaction terms

	Unstandardized Coefficients		Standardized Coefficients	T	p	Covariance Diagnosis	
	B	Std. Error	Beta			VIF	Tolerance
Constant	85.973	1.114	-	77.183	0.000**	-	-
X1	-0.000	0.000	-0.319	-8.045	0.000**	21.445	0.047
X2	-2.488	0.232	-0.174	-10.703	0.000**	3.589	0.279
X3	-0.003	0.000	-0.229	-7.507	0.000**	12.726	0.079
X4	-0.381	0.096	-0.097	-3.951	0.000**	8.225	0.122
X5	-0.102	0.014	-0.158	-7.358	0.000**	6.260	0.160
X6	-0.489	0.015	-0.588	-32.208	0.000**	4.541	0.220
X1X2	0.000	0.000	0.185	5.022	0.000**	18.455	0.054
X1X3	0.000	0.000	0.254	8.908	0.000**	11.065	0.090
X1X4	-0.000	0.000	-0.033	-1.640	0.101	5.414	0.185
X2X4	0.103	0.019	0.148	5.558	0.000**	9.725	0.103
R^2	0.803						
Adj R^2	0.803						
F	F (10,2685)=1096.283,p=0.000						
D-W value	0.227						

All variables show a high level of significance except the interaction term of GDP per capita and female unemployment. The single variables exhibit a substantial negative impact on y, while the three interaction terms demonstrate a notable positive impact on y. Hence, a modified model can be obtained as follows: $y = 85.973 - 0.319x_1 - 0.174x_2 - 0.229x_3 - 0.097x_4 - 0.158x_5 - 0.588x_6 + 0.185x_1x_2 + 0.254x_1x_3 - 0.033x_1x_4 + 0.148x_2x_4$.

After incorporating the interaction term, gross domestic product per person has a significant coefficient of -0.319, which is tripled compared with the coefficient in the previous model. Among the four interaction terms, only the interaction term of GDP per capita and female unemployment does not have significant effect. Other three interaction terms exhibited significance at the 1% level. This demonstrates that the impact of GDP per capita on the newborn death is contingent upon education spending and health spending.

In addition, the model demonstrates an adjusted R^2 value of 0.803, meaning that it describes 80.3% of the variation in the infant mortality rate by the GDP per capita and the interaction term, which is slightly higher than it is in the multiple linear regression without interaction terms. The model successfully passes the F-test F(10,2685)=1096.283,p=0.000, which confirms its validity.

3.3. Log-log model

The log-log model, also referred to as the constant elasticity model, involves taking the logarithm of both the dependent variable and the main independent variable. It aims to examine the potential correlation between GDP per capita and infant mortality, but not in conventional units. Instead, it analyses the relationship in terms of percentage change. The mathematical model is as follows:

$$lny = \beta_0 + \beta_1 lnx_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + e \quad (3)$$

Table 6 below presents the outcomes.

DOI: 10.54254/2753-8818/51/2024CH0195

Table 6. Log-log Regression analysis results

	POOL Model	FE Model	RE Model	Fixed Effect
Constant	6.289** (17.719)	5.761** (13.213)	5.621** (18.247)	3.713** (8.195)
lnX1	-0.294** (-5.561)	-0.161** (-2.998)	-0.167** (-3.759)	-0.073 (-1.481)
X2	-0.019 (-1.036)	-0.023 (-1.563)	-0.022 (-1.554)	-0.012 (-0.804)
X3	-0.000** (-2.724)	-0.000** (-4.459)	-0.000** (-4.803)	-0.000 (-0.004)
X4	0.017** (2.686)	0.010 (1.500)	0.012* (2.268)	0.005 (0.896)
X5	-0.007** (-3.067)	-0.014** (-5.024)	-0.011** (-5.204)	-0.007** (-2.802)
X6	-0.007** (-3.043)	-0.007** (-3.133)	-0.007** (-3.444)	0.001 (0.538)
R^2	0.780	0.756	0.771	0.423
R^2(within)	0.240	0.264	0.261	0.109
F	F (6,2689)=183.947, p=0.000	F (6,2524)=25.980, p=0.000	χ2(6)=498.658, p=0.000	F (6,2503)=2.719, p=0.012

The models were estimated using the method of robust standard errors. Table 6 presents the outcomes of the mixed model, fixed effect model, and random effect model. Multiple tests are performed to ascertain the most favorable model. The F test, performed with a significance level of 5%, produced a result of F (165,2524) = 39.196, with a p-value of 0.000, which is smaller than 0.05. This suggests that FE model is more advantageous compared to POOL model. The blood pressure (BP) test demonstrated statistical significance at a 5% significance level, with a chi-square value of 12345.522 and a p-value of 0.000, which is lower than the threshold of 0.05. This indicates that RE model is more advantageous compared to POOL model. The Hausman test yielded a non-significant result, indicating that both the chi-square value of 2.885 and the p-value of 0.718 are higher than the significance level of 0.05. This implies that RE model is more advantageous compared to FE model. Thus, RE model is chosen as the ultimate outcome.

According to RE model, all variables demonstrate a significant level of importance. Variables $x_2, x_3, x_5, and\ x_6$ demonstrate a negative correlation with y, whereas x_4 displays a prominent positive impact on y. Thus, a revised model can be obtained as follows: $lny = 5.621 - 0.167lnx_1 - 0.022x_2 + 0.012x_4 - 0.011x_5 - 0.007x_6$

The coefficient for the logged GDP per capita is -0.167, and it is significant. This finding indicates that a 1% rise in GDP per capita is associated with a 16.7% newborn-death reduce. The logged term model accounts for 77.1% of the variance in infant mortality rate, taking into consideration GDP per capita and other controlled variables.

3.4. Quadratic model

The quadratic model is employed to investigate the presence of a diminishing reverse correlation between infant death and GDP per capita while accounting for country-level and time-fixed effects. The mathematical model is as follows:

$$y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + e \quad (4)$$

The term x_1^2 denotes the GDP per capita raised to the exponent of two. If the regression coefficients for the quadratic term are both statistically significant and negative, it indicates that as GDP per capita increases, the infant mortality rate decreases. Furthermore, the strength of this relationship weakens as GDP per capita increases.

DOI: 10.54254/2753-8818/51/2024CH0195

Table 7. Quadratic Term Regression Model analysis results

	POOL Model	FE Model	RE Model	Fixed Effect
Constant	80.862** (18.176)	94.670** (18.728)	88.639** (22.970)	58.301** (13.308)
$X1^2$	0.000 (0.390)	-0.000 (-1.453)	-0.000 (-0.696)	0.000 (0.199)
X2	-1.112** (-2.621)	-1.272** (-4.547)	-1.231** (-4.502)	-0.903** (-3.634)
X3	-0.002** (-3.758)	-0.000** (-3.623)	-0.001** (-4.054)	0.003** (5.413)
X4	0.097 (0.743)	0.082 (0.851)	0.117 (1.322)	-0.093 (-0.866)
X5	-0.167** (-3.201)	-0.238** (-3.699)	-0.184** (-3.691)	-0.034 (-0.587)
X6	-0.481** (-6.787)	-0.594** (-10.256)	-0.580** (-11.107)	-0.366** (-8.494)
R^2	0.791	0.741	0.773	0.446
R^2(within)	0.497	0.525	0.522	0.295
F	F (6,2689)=97.427, p=0.000	F (6,2524)=45.783, p=0.000	$\chi 2(6)$=598.688, p=0.000	F (6,2503)=39.910, p=0.000

Table 7 presents the coefficients of the quadratic model. The method of robust standard error is once again utilized. As with the previous steps, a model test is performed to determine the best model. The initial F test shows a significance level of 5% with F (165,2524) =66.315, p=0.000<0.05, suggesting that FE model is better than POOL model. Furthermore, BP test demonstrates a significance level of 0.05 with F (165,2524) =66.315, p=0.000<0.05, providing additional evidence that FE model is superior to POOL model. BP test yielded a statistically significant result at a significance level of 5%. This implies that RE model offers more benefits compared to POOL model. Hausman test yielded a non-significant result, with a chi-square statistic of 3.970 and a p-value of 0.554, which is higher than the significance level of 0.05. This suggests that RE model is more beneficial compared to FE model. Thus, RE model is the final result. Using the given data, this paper can derive the quadratic regression equation as follows: $y = 88.639 - 1.231x_2 - 0.001x_3 - 0.184x_5 - 0.58x_6$.

The quadratic term has a p-value of 0.487, indicating that it does not have a statistically significant effect on infant death. The insignificance of x_4 has also been demonstrated. The independent variables $x_2, x_3, x_5, and\ x_6$ all exhibit p-values below 0.01 in the T-test, as well as a negative regression coefficient. Thus, it can be deduced that these four factors - education spending, health spending, access to cooking resources, and access to electricity - have a substantial adverse effect on infant mortality.

The model demonstrates an adjusted R^2 value of 0.773, signifying that it explains 77.3% of the variation in the infant mortality rate by considering the GDP per capita and the interaction term. This value is marginally higher than the one observed in the log-log regression model.

4. Conclusion

The study extracted a total of 4995 samples from the data set spanning from 2000 to 2022. The data set comprises 7 variables. The method employed, multiple linear regression analysis, is highly precise, efficient, and thorough. A thorough analysis is performed by taking various factors into account and subsequently determining the Pearson correlation for each variable.

At the beginning of the analysis phase, the multiple linear regression model is utilized in order to arrive at a conclusion regarding the possible correlation. The result shows that, for every 1 dollar rise in GDP per capita, infant mortality decreases by 0.107, assuming all other factors remain unchanged.

When attempting to acquire a more profound comprehension, it is imperative to consider the effect of interaction effects and to incorporate interaction terms that contain coefficients into the equation. An

DOI: 10.54254/2753-8818/51/2024CH0195

inverse relationship was identified, and the GDP per capita coefficient increased threefold compared to the earlier model. While the six individual variables still have a significant negative effect on IMR as demonstrated in the previous model, the three interaction terms have a positive relationship with infant death.

According to the log-log model, a one percent increase in GDP per capita is strongly correlated with a sixteen-point seven percent decrease in infant mortality rate. All the other independent variables exhibit a significant negative correlation with y, except for x4, which displays a significant positive coefficient.

Regarding the quadratic model, the primary independent variable, gross domestic product per person, is found to be statistically insignificant. Conversely, newborn death is adversely impacted by education spending, health spending, access to cooking resources, and access to electricity.

Policymakers should consider enhancing foundational education on reproductive understanding and allocating more resources toward healthcare expenditures. Consequently, there are improved healthcare settings for pregnant women and babies, which possess the capacity to decrease infant death rates.

Regarding the models tested above, there are still some shortcomings to consider. It is necessary to locate a panel data set with a reduced number of missing values for future research. In my present dataset, the missing percentage for both infant mortality and GDP per capita is less than 15 percent. However, health spending and education spending have more than 20 percent missing values. Furthermore, it is necessary to include more social and economic variables to further explore the interaction effect between various variables.

References

[1] As-Khan J R and Awan N 2017 A comprehensive analysis on child mortality and its determinants in Bangladesh using frailty models. Archives of Public Health, 75(1).

[2] Baraki A G, Akalu T Y, Wolde H F, Lakew A M and Gonete K A 2020 Factors affecting infant mortality in the general population: evidence from the 2016 Ethiopian demographic and health survey (EDHS); a multilevel analysis. BMC Pregnancy and Childbirth, 20(1).

[3] Simmons R A, Anthopolos R and O'Meara W P 2021 Effect of health systems context on infant and child mortality in sub-Saharan Africa from 1995 to 2015, a longitudinal cohort analysis. Scientific Reports, 11(1).

[4] Lentzner H R 2024 Seasonal Patterns Of Infant And Child Mortality In New York, Chicago, And New Orleans: 1870-1919 (Illinois, Louisiana). Proquest.

[5] Bairoliya N and Fink G 2018 Causes of death and infant mortality rates among full-term births in the United States between 2010 and 2012: An observational study. PLoS Medicine, 15(3), 1002531.

[6] Jang C and Lee H 2022 A Review of Racial Disparities in Infant Mortality in the US. Children, 9(2), 257.

[7] MacDorman M F and Mathews T J 2011 Understanding racial and ethnic disparities in U.S. infant mortality rates. PubMed, 74, 1-8.

[8] Okobi O E, et al. 2023 Trends and Factors Associated With Mortality Rates of Leading Causes of Infant Death: A CDC Wide-Ranging Online Data for Epidemiologic Research (CDC WONDER) Database Analysis. Cureus.

[9] Owusu P A, Sarkodie S A and Pedersen P A 2021 Relationship between mortality and health care expenditure: Sustainable assessment of health care system. PloS One, 16(2), 247413.

[10] Zakir M and Wunnava P V 1999 Factors affecting infant mortality rates: evidence from cross–sectional data. Applied Economics Letters, 6(5), 271-273.