Prediction of Population Growth Rate using Multiple Linear Regression

Heena Chopra
MSc in Data Analytics
Student ID: *x19205309*

***Abstract*—This project describes a multiple linear regression-based model for predicting the population growth for a set of countries listed in the United Nations (UN) Population growth Dataset. The model analysis is done using UN datasets for various factors such as Mortality Rate, Fertility Rate, Human Development Index, Life Expectancy, Net Migration Rate etc. We perform a number of checks on the assumptions that need to be accomplished before applying Regression analysis to the model. The model is built considering data for 147 countries in the year 2012. Data cleaning for the model was done using python while the Development was done using IBM SPSS (Statistical Package for Social Sciences).**

I. Introduction

Population Growth refers to increase in number of people in a population. According to Wikipedia, the global population has grown from 1 Billion in 1800 to 7.8 billion in 2020 [1]. Population Growth is basically affected by the number of births and the number of deaths. Better health is increasing the population growth as it decreases the mortality rate. Also, the trend in couples these days for having fewer number of children is decreasing the population growth. These factors transition for population growth shows that these population explosions are temporary. It is dependent on a number of factors that affect the population growth directly or indirectly [2].

It is generally said that countries with higher population growth have a lower standard of living whereas the countries with a lower population growth have a higher standard of living as lesser population leads to better employment opportunities, accelerating urbanization and so on. Considering all of these factors that affect the population growth, a model is made on multiple-linear regression that analyses whether these factors affect the population growth directly or indirectly.

II. Model Theory

Multiple Linear Regression Multiple Linear Regression allows us to predict a dependent variable(also known as the response variable) based on it's relationship with other independent variables(also known as the predictor variable). It is an extension of the linear regression model to include multiple independent variables into consideration while estimating the response variable. The equation for the multiple regression model can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon \qquad (1)$$

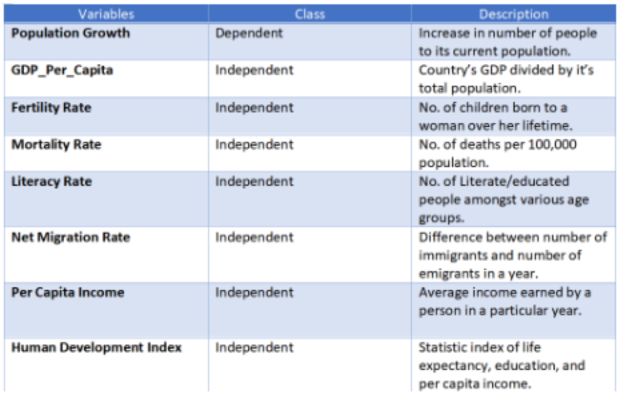
Where, for $i = n$ observations:
y_i = Response Variable
x_i = Predictor Variables
β_0 = Y- Intercept
β_p = Slope of pth predictor variable
ϵ = Error Term

We have certain assumptions that need to be considered for implementing a successful multiple linear regression model. The assumptions are as mentioned below:

1) The relationship between the Independent variables(IVs) and the dependent variables(DV) should be linear.
2) There shouldn't be any multicollinearity in our data.
3) The values of the residuals should be independent.
4) The variance of the residuals should be constant.
5) The values of the residuals should be normally distributed.
6) There shouldn't be any influential cases biasing our model.

III. Data Sources, Cleaning and Data Preparation

A. Variables Used



Variables	Class	Description
Population Growth	Dependent	Increase in number of people to its current population.
GDP_Per_Capita	Independent	Country's GDP divided by it's total population.
Fertility Rate	Independent	No. of children born to a woman over her lifetime.
Mortality Rate	Independent	No. of deaths per 100,000 population.
Literacy Rate	Independent	No. of Literate/educated people amongst various age groups.
Net Migration Rate	Independent	Difference between number of immigrants and number of emigrants in a year.
Per Capita Income	Independent	Average income earned by a person in a particular year.
Human Development Index	Independent	Statistic index of life expectancy, education, and per capita income.

Fig. 1. Description of independent and dependent variables used

B. Sources of Data

Datasets for all of these variables (Dependent and Independent) were downloaded from UN Dataset Repository in csv format.

C. Data Cleaning and Preparation

The dataset for dependent and the independent variables were downloaded from UN Data website in the csv format.

These datasets were further imported in Jupyter notebook using Python libraries such as Numpy and Pandas.

Each of the dataset that was downloaded was first cleaned and transformed to make it appropriate as per the model. For E.g.: Population Growth dataset that was imported earlier had few of the records and attributes that were unnecessary for the model such as Indicator, Gender, Value Footnotes which were removed for further evaluation. Many records had NULL values for fewer columns which were filled accordingly. Many attributes were renamed according to the requirements, for e.g.: Location was renamed to Country, Period was renamed to Year. This procedure was followed for all other datasets also.

Next step after cleaning all the datasets was to transform and merge them into a single dataset. Since, these datasets had data for quite a number of years and different countries and merging data for so many years was time consuming. The year 2012 was chosen for the analysis since it contained minimum no. of empty records as compared to other years. All the datasets were concatenated together by the common column 'Country'.The countries for which the data for all the predictor variables was available were kept and all other rows were removed. The resultant dataset formed had 121 rows and 9 Columns and was later used for analysis in SPSS. The cleaning and transformation of the datasets successfully removed the outliers also, which was clearly seen using Scatter Plots.

IV. Model Outline and Analysis

Many regression models were generated and the assumptions(mentioned in section II) for each of the models were checked.

A. Model-1

1) *Model Equation:*
   Population Growth = -2.786 - 0.448 (*Fertility Rate*) - 0.047(*Net Migration Rate*) - 0.019 (*Literacy Rate*) + 0.003 (*Mortality Rate*) - 4.442×10^{-5} (*GDP Per Capita*) + 5.663 (*HDI*) + 2.799 $\times 10^{-5}$ (*Per Capita Income*)
2) Adjusted R^2 = 0.805
3) Pearson Correlation Values:
   - Per Capita Income: GDP Per Capita = 0.959
   - HDI: Fertility Rate = -0.857
4) p-value (ANOVA Table) $<$ 0.005
5) VIF Values and p-values of the coefficients:
   - VIF-Value (GDP Per Capita) = 14.476
   - VIF-Value (HDI) = 10.196
   - VIF-Value (Per Capita Income) = 19.575
   - p-value (Per Capita Income) = 0.204
6) Coefficient Correlation Values
   - GDP per Capita : Per Capita Income = -0.916
7) Normal P-P Plot: Refer to Fig.2
8) Residual Scatter Plot: Refer to Fig.3
9) Analysis: The value of R^2 is 0.805 which is optimal but the Pearson Correlation Values indicate a strong correlation between Per Capita Income and GDP Per Capita, also with HDI and Fertility Rate. The p-value for (ANOVA Table) is less than 0.005 which is optimum but the coefficient p- value for Per Capita Income is more than 0.2. Also, *VIF* Values for fewer variables ¿10 which contradicts the assumptions for the model. In the Final model, Per Capita Income variable is removed making it a model with minimum correlations and maximum normality of residuals and *homoscedasticity*.

B. Model-2

1) *Model Equation:*
   Population Growth = -3.018 - 0.457 (*Fertility Rate*) - 0.047(*Net Migration Rate*) - 0.020 (*Literacy Rate*) + 0.003 (*Mortality Rate*) - 1.960×10^{-5} (*GDP Per Capita*) + 6.250 (*HDI*)
2) Adjusted R^2 = 0.804
3) Pearson Correlation Values:
   - HDI: Fertility Rate = -0.857
4) p-value (ANOVA Table) $<$ 0.005
5) VIF Values and p-values of the coefficients:
   - All VIF Values $<$ 10
   - All p-values $<$ 0.05
6) Coefficient Correlation Values
   - All Correlation values $<$ 0.8
7) Normal P-P Plot: Refer to Fig.4
8) Residual Scatter Plot: Refer to Fig.5
9) Analysis: This model is the final multiple regression model for this analysis. In this model, all the values and plots are up to the mark and as per the assumptions.

V. Figures and Plots

The following plots were obtained in the regression analysis done in SPSS.

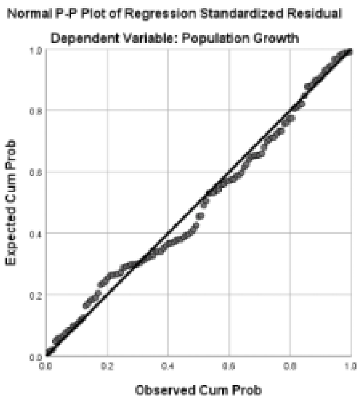


Fig. 2. Model-1: Normal P-P plot

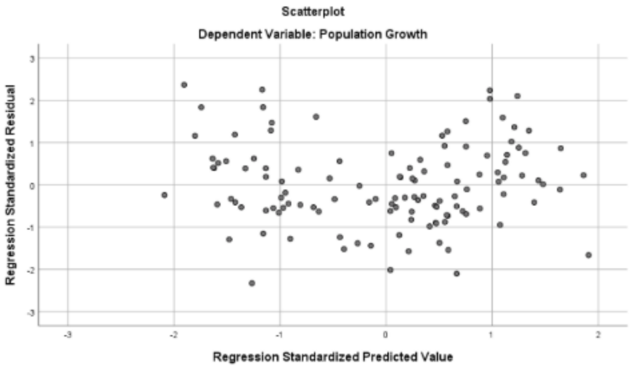


Fig. 3. Model-1: Scatter plot

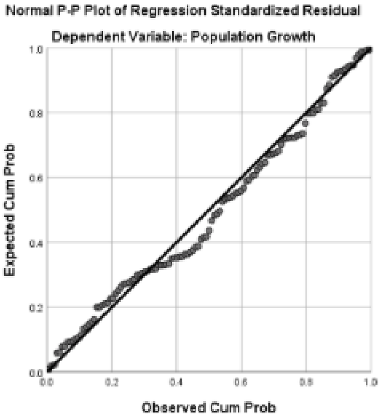


Fig. 4. Model-2: Normal P-P plot

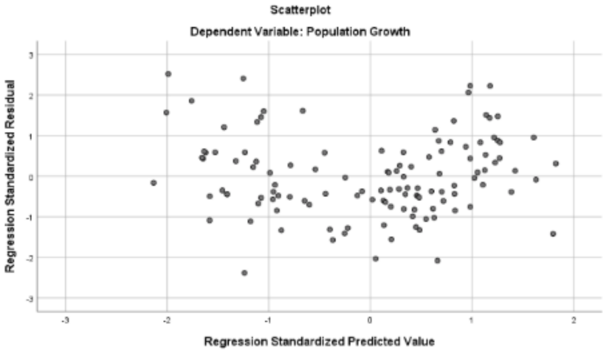


Fig. 5. Model-2: Scatter plot

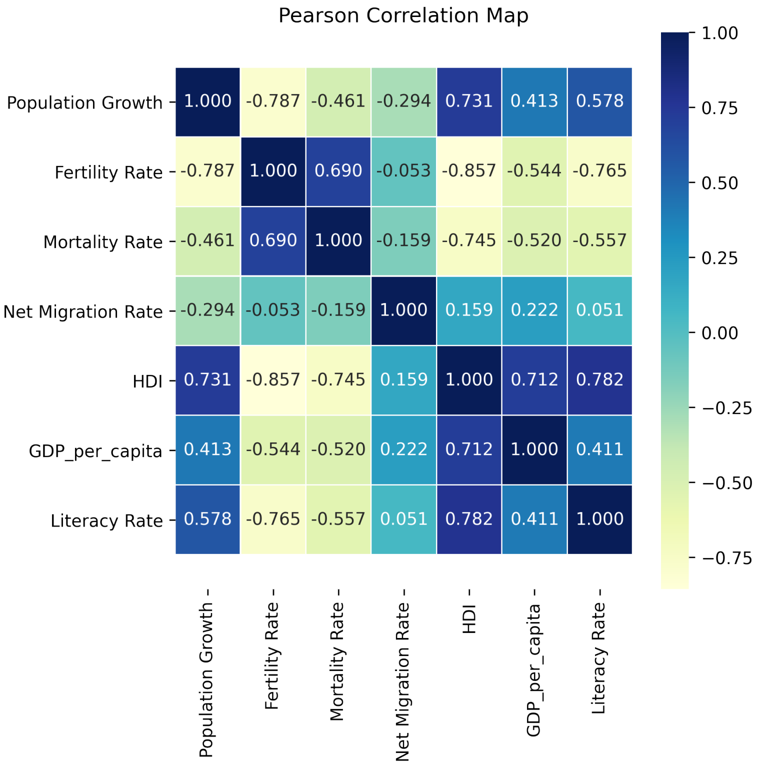


Fig. 6. Pearson Correlation Values Colormap

VI. Conclusion

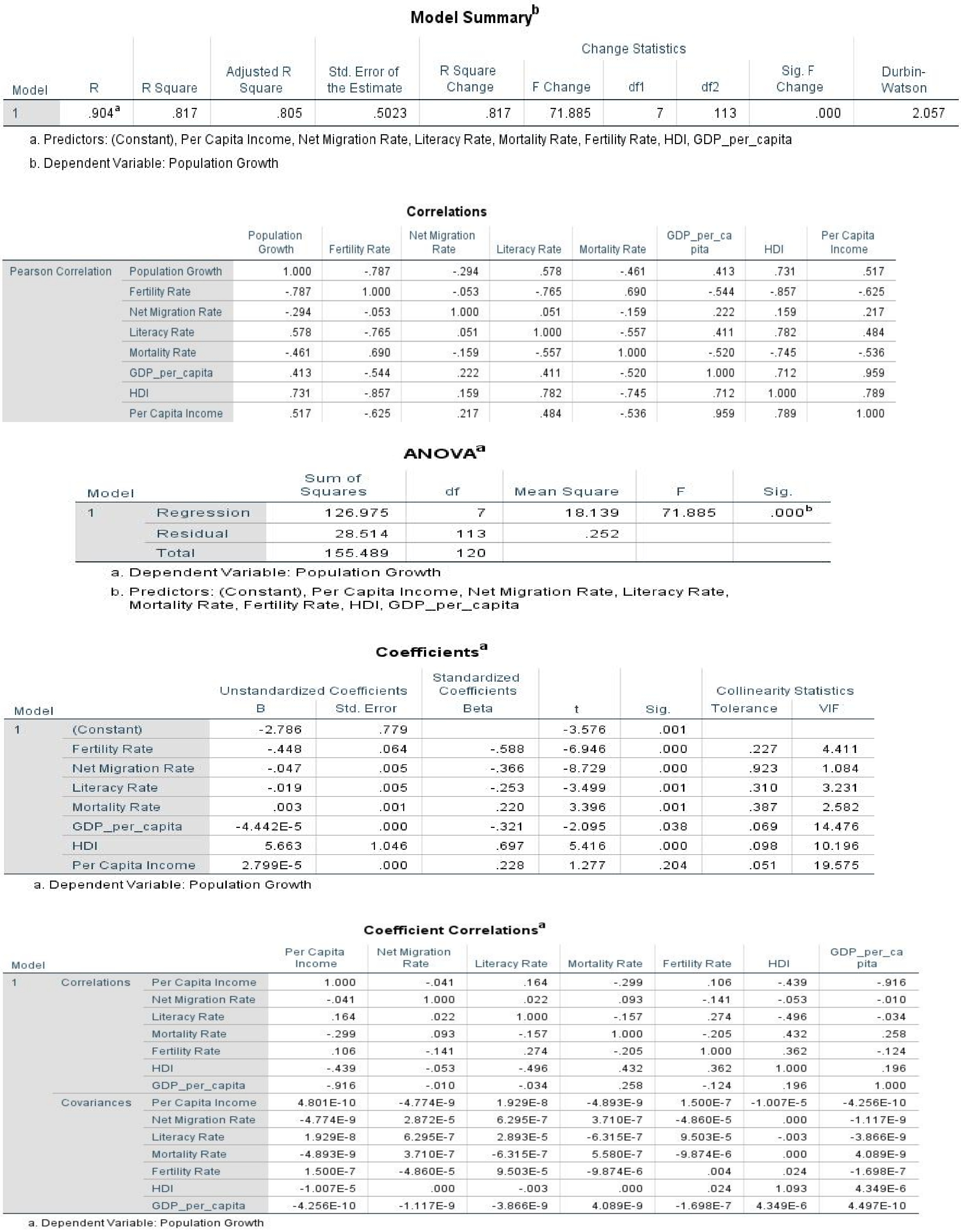
Multiple models were taken into consideration for regression analysis in SPSS. The final model chosen met all the assumptions except for the low correlation between independent variables which was voilated by the variable HDI. Although the HDI shows a correlation with Fertility Rate, it was included into the final model because the VIF and tolerance values were optimum. HDI could not be excluded out of the final model because of it's major contribution in the estimation of the dependent variable. The population growth is a complex phenomenon and it can't be completely explained by the limited no. of variables. The final model equation obtained is:

$$\textit{Population Growth} = \text{-3.018} - 0.457\ (\textit{Fertility Rate}) - 0.047(\textit{Net Migration Rate}) - 0.020\ (\textit{Literacy Rate}) + 0.003\ (\textit{Mortality Rate}) - 1.960 \times 10^{-5}\ (\textit{GDP Per Capita}) + 6.250\ (\textit{HDI})$$

References

[1] (2020) Population Growth Rate. https://en.wikipedia.org/wiki/Population_growth.

[2] (2020) Future Population Growth. https://ourworldindata.org/future-population-growth.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.904[a]	.817	.805	.5023	.817	71.885	7	113	.000	2.057

a. Predictors: (Constant), Per Capita Income, Net Migration Rate, Literacy Rate, Mortality Rate, Fertility Rate, HDI, GDP_per_capita

b. Dependent Variable: Population Growth

Correlations

		Population Growth	Fertility Rate	Net Migration Rate	Literacy Rate	Mortality Rate	GDP_per_capita	HDI	Per Capita Income
Pearson Correlation	Population Growth	1.000	-.787	-.294	.578	-.461	.413	.731	.517
	Fertility Rate	-.787	1.000	-.053	-.765	.690	-.544	-.857	-.625
	Net Migration Rate	-.294	-.053	1.000	.051	-.159	.222	.159	.217
	Literacy Rate	.578	-.765	.051	1.000	-.557	.411	.782	.484
	Mortality Rate	-.461	.690	-.159	-.557	1.000	-.520	-.745	-.536
	GDP_per_capita	.413	-.544	.222	.411	-.520	1.000	.712	.959
	HDI	.731	-.857	.159	.782	-.745	.712	1.000	.789
	Per Capita Income	.517	-.625	.217	.484	-.536	.959	.789	1.000

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	126.975	7	18.139	71.885	.000[b]
	Residual	28.514	113	.252		
	Total	155.489	120			

a. Dependent Variable: Population Growth

b. Predictors: (Constant), Per Capita Income, Net Migration Rate, Literacy Rate, Mortality Rate, Fertility Rate, HDI, GDP_per_capita

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-2.786	.779		-3.576	.001		
	Fertility Rate	-.448	.064	-.588	-6.946	.000	.227	4.411
	Net Migration Rate	-.047	.005	-.366	-8.729	.000	.923	1.084
	Literacy Rate	-.019	.005	-.253	-3.499	.001	.310	3.231
	Mortality Rate	.003	.001	.220	3.396	.001	.387	2.582
	GDP_per_capita	-4.442E-5	.000	-.321	-2.095	.038	.069	14.476
	HDI	5.663	1.046	.697	5.416	.000	.098	10.196
	Per Capita Income	2.799E-5	.000	.228	1.277	.204	.051	19.575

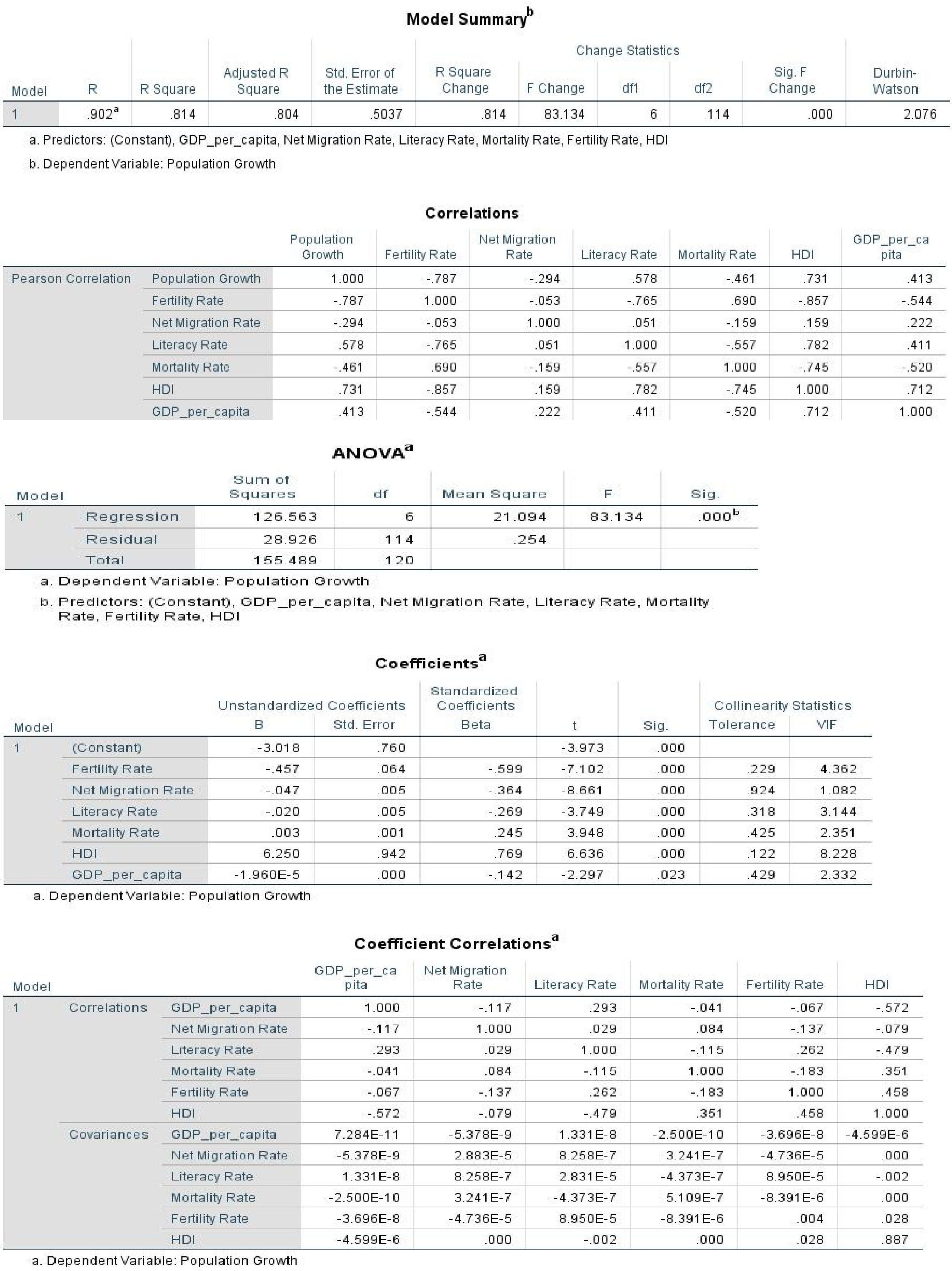
a. Dependent Variable: Population Growth

Coefficient Correlations[a]

Model			Per Capita Income	Net Migration Rate	Literacy Rate	Mortality Rate	Fertility Rate	HDI	GDP_per_capita
1	Correlations	Per Capita Income	1.000	-.041	.164	-.299	.106	-.439	-.916
		Net Migration Rate	-.041	1.000	.022	.093	-.141	-.053	-.010
		Literacy Rate	.164	.022	1.000	-.157	.274	-.496	-.034
		Mortality Rate	-.299	.093	-.157	1.000	-.205	.432	.258
		Fertility Rate	.106	-.141	.274	-.205	1.000	.362	-.124
		HDI	-.439	-.053	-.496	.432	.362	1.000	.196
		GDP_per_capita	-.916	-.010	-.034	.258	-.124	.196	1.000
	Covariances	Per Capita Income	4.801E-10	-4.774E-9	1.929E-8	-4.893E-9	1.500E-7	-1.007E-5	-4.256E-10
		Net Migration Rate	-4.774E-9	2.872E-5	6.295E-7	3.710E-7	-4.860E-5	.000	-1.117E-9
		Literacy Rate	1.929E-8	6.295E-7	2.893E-5	-6.315E-7	9.503E-5	-.003	-3.866E-9
		Mortality Rate	-4.893E-9	3.710E-7	-6.315E-7	5.580E-7	-9.874E-6	.000	4.089E-9
		Fertility Rate	1.500E-7	-4.860E-5	9.503E-5	-9.874E-6	.004	.024	-1.698E-7
		HDI	-1.007E-5	.000	-.003	.000	.024	1.093	4.349E-6
		GDP_per_capita	-4.256E-10	-1.117E-9	-3.866E-9	4.089E-9	-1.698E-7	4.349E-6	4.497E-10

a. Dependent Variable: Population Growth

Fig.7. Model-1 : Full Summary



Model Summary[b]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.902[a]	.814	.804	.5037	.814	83.134	6	114	.000	2.076

a. Predictors: (Constant), GDP_per_capita, Net Migration Rate, Literacy Rate, Mortality Rate, Fertility Rate, HDI

b. Dependent Variable: Population Growth

Correlations

		Population Growth	Fertility Rate	Net Migration Rate	Literacy Rate	Mortality Rate	HDI	GDP_per_capita
Pearson Correlation	Population Growth	1.000	-.787	-.294	.578	-.461	.731	.413
	Fertility Rate	-.787	1.000	-.053	-.765	.690	-.857	-.544
	Net Migration Rate	-.294	-.053	1.000	.051	-.159	.159	.222
	Literacy Rate	.578	-.765	.051	1.000	-.557	.782	.411
	Mortality Rate	-.461	.690	-.159	-.557	1.000	-.745	-.520
	HDI	.731	-.857	.159	.782	-.745	1.000	.712
	GDP_per_capita	.413	-.544	.222	.411	-.520	.712	1.000

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	126.563	6	21.094	83.134	.000[b]
	Residual	28.926	114	.254		
	Total	155.489	120			

a. Dependent Variable: Population Growth

b. Predictors: (Constant), GDP_per_capita, Net Migration Rate, Literacy Rate, Mortality Rate, Fertility Rate, HDI

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-3.018	.760		-3.973	.000		
	Fertility Rate	-.457	.064	-.599	-7.102	.000	.229	4.362
	Net Migration Rate	-.047	.005	-.364	-8.661	.000	.924	1.082
	Literacy Rate	-.020	.005	-.269	-3.749	.000	.318	3.144
	Mortality Rate	.003	.001	.245	3.948	.000	.425	2.351
	HDI	6.250	.942	.769	6.636	.000	.122	8.228
	GDP_per_capita	-1.960E-5	.000	-.142	-2.297	.023	.429	2.332

a. Dependent Variable: Population Growth

Coefficient Correlations[a]

Model			GDP_per_capita	Net Migration Rate	Literacy Rate	Mortality Rate	Fertility Rate	HDI
1	Correlations	GDP_per_capita	1.000	-.117	.293	-.041	-.067	-.572
		Net Migration Rate	-.117	1.000	.029	.084	-.137	-.079
		Literacy Rate	.293	.029	1.000	-.115	.262	-.479
		Mortality Rate	-.041	.084	-.115	1.000	-.183	.351
		Fertility Rate	-.067	-.137	.262	-.183	1.000	.458
		HDI	-.572	-.079	-.479	.351	.458	1.000
	Covariances	GDP_per_capita	7.284E-11	-5.378E-9	1.331E-8	-2.500E-10	-3.696E-8	-4.599E-6
		Net Migration Rate	-5.378E-9	2.883E-5	8.258E-7	3.241E-7	-4.736E-5	.000
		Literacy Rate	1.331E-8	8.258E-7	2.831E-5	-4.373E-7	8.950E-5	-.002
		Mortality Rate	-2.500E-10	3.241E-7	-4.373E-7	5.109E-7	-8.391E-6	.000
		Fertility Rate	-3.696E-8	-4.736E-5	8.950E-5	-8.391E-6	.004	.028
		HDI	-4.599E-6	.000	-.002	.000	.028	.887

a. Dependent Variable: Population Growth

Fig.8. Model 2 : Full Model Summary