MODELLING OF WORLD DEVELOPMENT INDICATORS INFLUENCING THE GLOBAL ECONOMIC MOVEMENTS USING STATISTICAL TECHNIQUES.

[1]Apoorva R. Pali

[1]Assistant Professor, Department of Statistics, MES Abasaheb Garware College, Pune, India.

Abstract: World Development Indicators is a collection of internationally comparable statistics from officially-recognized international sources presenting the most accurate global development data. In this paper, the current and historical data containing nineteen prime development indicators for 96 countries are studied. Multiple linear regression analysis is used to identify the significant development indicators affecting a country's GNI per capita and to predict its future value based on the significant indictors. The model is improved by using Stepwise variable selection method, Box-Cox transformation and removing the influential observations. Further as a remedy to multicollinearity, ridge regression is used and an optimal model satisfying all the assumptions of multiple regression with lesser residual error is obtained. All these results will help countries in policy making and achieving developmental goals.

Keywords: World Development Indicators, Gross National Income, Multiple Linear Regression Analysis, Stepwise variable selection method, Box-Cox transformation, Influential Observations, Multicollinearity, Ridge Regression.

I. INTRODUCTION

All countries in the world are striving to improve their economies sustainably in order to alleviate poverty and advance towards a mound of targets. The economy of any country depends on various vital parameters which directly or implicitly influence it. The basic objective of this paper is to study the crucial factors responsible for the economic development of any country.

World Development Indicators (WDI) is the chief compilation of development indicators of World Bank. WDI represents the most recent global data with highest accuracy. It comprises of internationally comparable statistics of global development and the fight against poverty like health indicators, demography statistics, economic factors. The database includes national, regional, and global estimates and is compiled from officially-recognized sources.

1.1 Overview of the variables under study

The various indicators considered in our study are as follows:

1) Gross National Income per capita

Gross National Income (GNI) is used as an indicator of a nation's wealth. GNI is a nation's total money earned by its people and businesses. It also includes money earned from oversees in terms of foreign investment and economic development aid. Thus, GNI is the gross domestic product, plus incomes earned by foreign residents, minus income earned in the domestic economy by non-residents. GNI of a country divided by its mid-year population is the country's GNI per capita.

2) Age dependency ratio (% of working-age population)

A country's dependents-age population is the count of people younger than 15 or older than 64 years of age. And working-age population is the total number of people between 15 to 64 years of age. Age dependency ratio of a country is the ratio of its dependents to its working-age population.

3) Agricultural land (sq. km)

Land under permanent crops is land used for cultivation of such crops that occupy the land for long periods and which are not necessarily replanted after each harvest. However, permanent pasture is land used for five or more years for forage. Agricultural land is defined as the land area that is arable, under permanent crops, and under permanent pastures.

4) Agricultural machinery (tractors per 100 sq. km of arable land)

Agricultural machinery is the number of tractors in use per 100 sq. km in agriculture in a country at the end of a calendar year.

5) Crude birth rate (per 1,000 people)

Crude birth rate is the number of live births per 1,000 population occurring during the year.

6) Cereal production (metric tons)

Cereal production is defined as the production of crops harvested for dry grain during the year. The cereal crops harvested for hay or harvested green for food, feed, or silage and those used for grazing are not included while calculating the cereal production of any country.

7) CO_2 emissions (metric tons per capita)
Carbon dioxide (CO_2) emissions are the emissions occurring due to the burning of fossil fuels and the manufacturing of cement. Carbon dioxide emitted during the consumption of solid, liquid, and gas fuels is also included in it.

8) Crop production index (2004-2006 = 100)
Crop production index is the index for agricultural production of each year taking 2004-2006 as the base period.

9) Death rate, crude (per 1,000 people)
Crude death rate of a country is defined as the number of deaths per 1,000 population occurring in that country during the year.

10) Domestic credit provided by financial sector (% of GDP)
Domestic credit provided by the financial sector includes all credit to various sectors.

11) Total fertility rate (births per woman)
Total fertility rate is defined as the number of children that would be born to a woman if she were to live to the end of her childbearing years.

12) Food production index (2004-2006 = 100)
Any country's food production index is the production index of edible and nutritive food crops taking 2004-2006 as the base period. As coffee and tea have no nutritive value, although edible, they are excluded from this index.

13) Land under cereal production (hectares)
The land under cereal production of a country is the total harvested area in that country. Cereals like rice, wheat, barley, maize, buckwheat, oats etc. are considered for this indicator.

14) Life expectancy at birth, total (years)
Life expectancy at birth is one of the most important indicators of a country. It is the number of years a newborn infant is expected to live if the prevailing conditions of mortality at the time of its birth were to remain the same throughout its life.

15) Livestock production index (2004-2006 = 100)
The livestock production index is the production index of meat and milk, honey, wool, raw silk and dairy products such as cheese etc. by taking 2004-2006 as the base period.

16) Merchandise exports (current US$)
Merchandise exports of a country is the value of goods exported to other countries valued in current U.S. dollars.

17) Mortality rate, adult, male (per 1,000 male adults)
Adult male mortality rate is defined as the probability of a 15-year-old male dying before reaching age 60, subject to the prevalent age-specific mortality rates of that year.

18) Mortality rate, adult, female (per 1,000 male adults)
Adult female mortality rate is defined as the probability of a 15-year-old female dying before reaching age 60, subject to the prevalent age-specific mortality rates of that year.

1.2 Objectives of the study
The prime objective of our study is to identify the variables which influence GNI per capita of any country significantly. After identifying the significant indicators, we tried to establish a model for GNI per captia based on the significant indicators influencing its variations. The aim is to remove all the model deficiencies and obtain an optimized model for future prediction of GNI per capita for any country.

II. LITERATURE REVIEW
In past few decades plenty of work has been done by many researches on economic parameters of countries. A brief review of literature is given below to throw light on the work done by researchers on study of the global economic factors.

K. S. Bhanu and A. R. Pali, 2019 [1] in their study found the significant determinants affecting BSE Sensex fluctuations in India. They used the method of ridge regression to optimize the model. de Vlaming, R., & Groenen, P. J.,2015 [2] gauged the current and future potential of ridge regression and predicted human traits using genome-wide SNP data. They concluded that under favorable conditions the predictive accuracy of ridge regression is slightly higher than the approach based on repeated simple regression for their data. Pallavi Kudal, 2010 [3] studied the impact of various economic factors on the fluctuations of stock market in India using multiple linear regression. Robert H. Rasche and Harold T. Shapiro, 1968 [5] tried to estimate the relationship in variations of an economic factor using macroeconomic approach. These studies inspired us to carry out further research in this area.

III.STATISTICAL METHODOLOGIES & DATA COLLECTION
The different methods and data used in this paper are described in this section.

2.1 Statistical methods used in the study
Various statistical techniques used are described in brief below

1) Karl Pearson Correlation Coefficient

The Karl Pearson's correlation coefficient measures the extent of a linear relationship between two variables and is denoted by r or r_{xy} (x and y being the two variables involved).

2) Multiple Linear Regression

Multiple linear regression is used to model the relationship between two or more independent variables and a dependent variable by fitting a linear equation to the given data.

3) Multicollinearity

When one or more regressor variables in a multiple regression model are highly correlated then this is known as multicollineairty. When the assumption of linear independency of regressor variables gets violated then the problem of multicollinearity arises.

4) Variable Selection

From the set of regressor variables on which data is collected, some can be dropped before a final working model is built. Variable selection method is used to find an optimal set of regressor variables. We have used Stepwise selection method in the study.

5) Ridge Regression

It is one of the procedures to deal with the problem of multicollinearity. The ridge estimator is found by solving a slightly modified version of the normal equations. The modified normal equation in calculating Ridge estimator is given by:

$$(X'X + kI)\ \hat{\beta}_R = X'Y \qquad (1)$$

On solving Equation (1) we get the Ridge estimator $\hat{\beta}_R$ as follows:

$$\hat{\beta}_R = (X'X + kI)^{-1} X'Y \qquad (2)$$

where $k \geq 0$ is a constant selected by the analyst.

As the ridge estimator has less variance than the ordinary least square estimator, $\hat{\beta}_R^*$ is a more stable estimator of β than is the unbiased estimator estimator $\hat{\beta}$ (OLS estimator).

2.2 Data and data collection

The data considered for the study is a secondary quantitative data collected from authentic World Bank site. The dataset comprises of 19 variables mentioned above viz. GNI per capita, Age dependency ratio (% of working-age population), Agricultural land (sq. km), etc. The data is an annual data for 96 countries collected over a period of 56 years i.e. from 1961 to 2016.

IV. RESULTS AND DISCUSSION

Table 1: Notations used for variables considered in our study

Variables	Factor
Y	GNI per capita
X_1	Age dependency ratio (% of working-age population)
X_2	Agricultural land (sq. km)
X_3	Agricultural machinery (tractors per 100 sq. km of arable land)
X_4	Crude birth rate (per 1000 people)
X_5	Cereal production (metric tons)
X_6	CO2 emissions (metric tons per capita)
X_7	Crop production index (2004-2006 = 100)
X_8	Crude death rate (per 1000 people)
X_9	Domestic credit provided by financial sector (% of GDP)
X_{10}	Fertility rate, total (births per woman)
X_{11}	Food production index (2004-2006 = 100)
X_{12}	Land under cereal production (hectares)
X_{13}	Total life expectancy at birth (years)
X_{14}	Livestock production index (2004-2006 = 100)
X_{15}	Merchandise exports (current US$)
X_{16}	Mortality rate (per 1000 male adults)
X_{17}	Mortality rate (per 1000 female adults)
X_{18}	Population ages 15-64 (% of total population)

The notations used for statistical analysis are given in Table 1.

The statistical analysis has been carried out in following four steps:

STEP I: Multicollinearity Diagnostics

Table 2: Eigen values and condition indices of X'X matrix

Regressor	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Eigen value	6.47	5.72	2.81	0.89	0.64	0.52	0.33	0.24	0.19
Condition index	1.0	1.13	2.30	7.26	10.12	12.48	19.62	26.52	34.32
Regressor	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
Eigen value	0.14	0.04	0.02	0.002	0.001	0.0001	0.0001	0.0001	0.0001
Condition index	45.86	181.03	416.87	416.87	**329290**	**962430**	**2.03×10^7**	**2.17×10^{15}**	**5.68×10^{15}**

Eigen values and condition indices of X'X matrix are calculated and shown in Table 2, where X is the design matrix of regressor variables.

DISCUSSION

It is observed from Table 2 that some of the regressor variables have very large values of condition indices (shown in bold) suggesting severe problem of multicollinearity.

STEP II: Variable Selection Method (Stepwise Regression)

Table 3: Stepwise variable selection method

	Estimate	Standard error	t value	Pr(>|t|)
(Intercept)	1768.004	4481.913	0.394	0.694185
X_1	132.898	61.786	2.151	0.034221 *
X_6	868.415	106.075	8.187	1.92e-12 ***
X_8	745.657	188.493	3.956	0.000154 ***
X_{17}	-25.106	10.867	-2.310	0.023209 *
X_7	-55.405	32.274	-1.717	0.089549 .
X_3	9.762	1.252	7.798	1.20e-11 ***
X_4	-312.597	40.008	-7.813	1.11e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model Summary I

Residual standard error: 3830 on 88 degrees of freedom
Multiple R-squared: 0.8389, Adjusted R-squared: 0.8261
F-statistic: 65.49 on 7 and 88 DF, p-value: < 2.2e-16

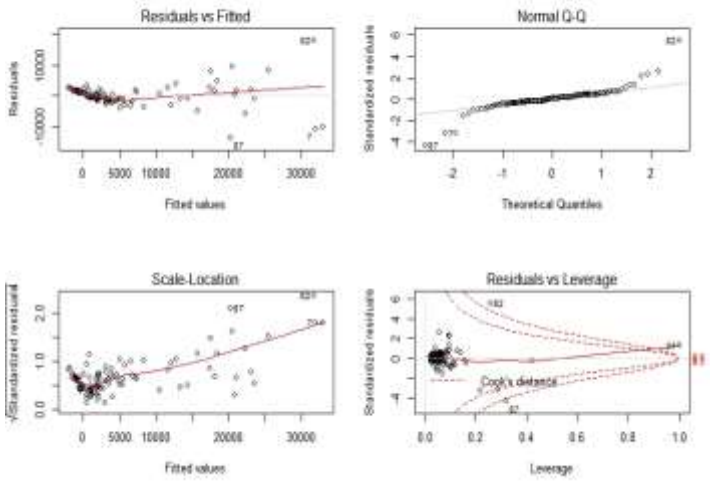

Figure 1: Residual plot for model I

After detecting the presence of multicollineariy, an optimal subset of regressors are obtained using Stepwise variable selection method. The result is given in Table 3. Residuals plots of the stepwise regression model is shown in Figure 1.

DISCUSSION

From Table 3, it is observed that Multiple R-squared is 0.8389 i.e. 83% variation in the response variable Y is explained by the model. It is also observed that repressors X_1, X_6, X_8, X_{17}, X_7, X_3 and X_4 comprise the optimal set of regressors for obtaining a better fit. From Table 3 it is seen that the intercept of stepwise regression model is not significant. Also, residual vs fitted plot shown in Figure 1 shows heteroscedasticity. The plot is not random within a band but is funnel shaped suggesting non constant variance. This is violation of the assumptions of multiple linear regression. Also, the residual plots show few outliers. The Normal Q-Q plot in Figure 1 shows heavy tails on both sides indicating non normality of the residuals. In the residuals vs leverage plot in Figure 1, observation 24 is going beyond the Cook's distance. This indicates that 24 is an influential observation and should be removed to improve the predictability of the model. Thus, there is a need to overcome the above drawbacks for further model optimization.

STEP III: Box-Cox Transformation and omission of influential observations

As from Figure 1 it is observed that the assumptions of normality and homoscedasticity are getting violated, we used Box-Cox transformation for model improvement.

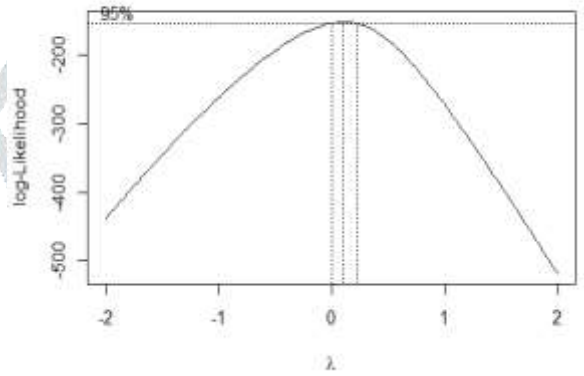

Figure 2: Box-Cox Transformation

It is observed from Figure 2 that the likelihood is getting maximized at $\lambda = 0$ (approx..). Thus, we transform the old variable Y to a new variable $Z = \log_e Y$. On applying the natural log transformation, stepwise regression and removing the influential observations we fit the model again. The results of the improved model are given in Table 4.

Table 4: Improved model

	Estimate	Standard error	t value	Pr(>|t|)
(Intercept)	-4.7105235	1.6956091	-2.778	0.006669 **
X_{13}	0.1591795	0.0198445	8.021	3.93e-12 ***
X_6	0.1012122	0.0120676	8.387	6.95e-13 ***
X_8	0.0911775	0.0254639	3.581	0.000558 ***
X_{16}	0.0035215	0.0013760	2.559	0.012180 *
X_3	0.0003414	0.0001471	2.321	0.022598 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model Summary II

Residual standard error: 0.4628 on 89 degrees of freedom
Multiple R-squared: 0.9086, Adjusted R-squared: 0.9035
F-statistic: 177 on 5 and 89 DF, p-value: < 2.2e-16

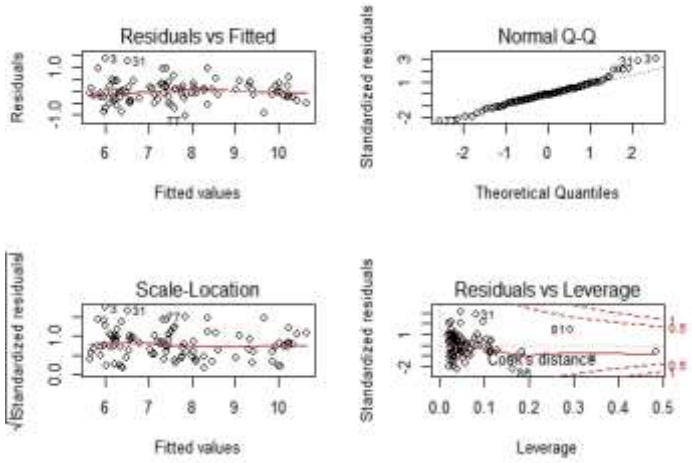

Figure 3: Residual plots of improved model

DISCUSSION

From Table 4 it is observed that Multiple R-squared of the model has increased to 0.9086 i.e. 90.86 % of variation in Z is explained by the model. All regressors and the intercept are significant now. The residuals vs fitted plot in Figure 3 shows random points within a band with no pattern. This indicates that the assumption of homoscedasticity is now valid for the model. Normal Q-Q plot also shows the validation of the assumption of normality. There are no influential points observed from the residuals vs leverage plot. Thus, the model has shown a drastic improvement.

STEP IV: Ridge regression as a remedy to multicollinearity

We note that even in the final model multicollinearity is present as few of the regressors are still linearity related. To deal with this problem we use ridge regression method as a remedy to the problem of multicollinearity.

The ridge estimator is obtained on solving a slightly modified version of the normal equations. The ridge estimator $\hat{\beta}_R$ is defined as a solution to the following equation:

$$(X'X + kI)\ \hat{\beta}_R = X'Y \qquad \text{From (1)}$$

The ridge estimator is as follows,

$$\hat{\beta}_R = (X'X + kI)^{-1} X'Y \qquad \text{From (2)}$$

where X is the design matrix of regressors, Y is the response vector and $k \geq 0$ is a non-negative constant selected by the analyst.

Starting with k = 0.001 to k = 1, we found the regression coefficient estimates from equation (2) for different values of k. Then we have plotted these values of ridge estimates against k to obtain the Ridge trace. As k increases, $\hat{\beta}_R$ (ridge regression coefficient) becomes more stable.

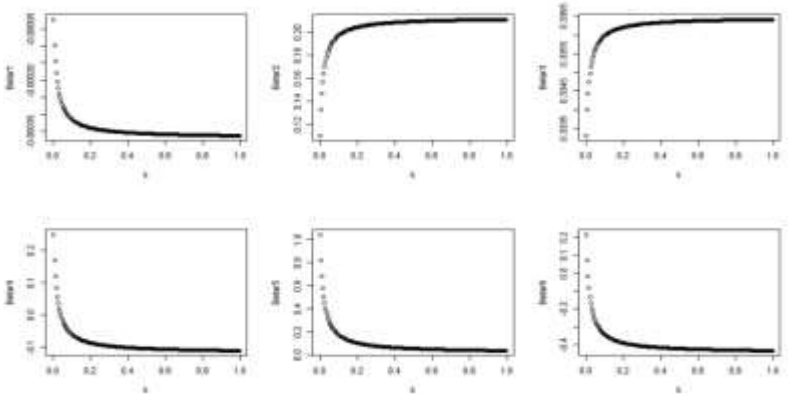

Figure 4: Ridge Trace

It is observed from Figure 4 that between k = 0.4 and k = 0.6, $\hat{\beta}_R$ has been found to be reasonably stable and thus we have taken k=0.5.

$\hat{\beta}_R = (\hat{\beta}_0, \hat{\beta}_3, \hat{\beta}_6, \hat{\beta}_8, \hat{\beta}_{13}, \hat{\beta}_{16})$, the vector of regression coefficients from equation (2) for k=0.3 is given in Table 5.

Table 5: Ridge estimates of regression coefficients

	(Intercept)	X_3	X_6	X_8	X_{13}	X_{16}
Ridge estimate	-0.0004	0.2094	0.3364	-0.1059	0.0498	-0.4212

Thus, our final model using the Ridge regression estimators obtained in Table 8 is given by:

$$Z = (-0.0004) + (0.2094 \times X_3) + (0.3364 \times X_6) + (-0.1059 \times X_8) + (0.0498 \times X_{13}) + (-0.4212 \times X_{16}) \quad (3)$$

DISCUSSION

On fitting the model given in equation (3) using the ridge estimates given in Table 5, the residual standard error has reduced from 0.4628 to 0.0412. Thus, the model has been improved by removing all the violations present and a highly optimized model has been obtained.

V. CONCLUSION

It is concluded that the factors viz. Agricultural machinery (tractors per 100 sq. km of arable land), CO2 emissions (metric tons per capita), Crude death rate (per 1000 people), Total life expectancy at birth (years) and Mortality rate (per 1000 male adults) significantly influence the variations in GNI per capita of a country. The optimized final model is given by $Z = (-0.0004) + (0.2094 \times X_3) + (0.3364 \times X_6) + (-0.1059 \times X_8) + (0.0498 \times X_{13}) + (-0.4212 \times X_{16})$ and can be used for future prediction of GNI per capita from the values of the explanatory variables for any country.

REFERENCES

[1] Bhanu, K. and Pali, A. 2019. "A Study on Determinants Affecting BSE Sensex in India: A Macroeconomic Approach", International Journal of Scientific Research in Mathematical and Statistical Sciences, Vol.6, Issue.2, pp.121-133.

[2] de Vlaming, R. and Groenen, P. J. 2015. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics, *BioMed Research International*, *2015*, 143712. doi:10.1155/2015/143712.

[3] Kudal, P. 2010. Impact of Macroeconomic Variables on Indian Stock Market and Strategies for Investors- Post Global Financial Crisis, *Apotheosis: Tirpude's National Journal Of Business Research (TNBJR)*,Vol 4, Issue 1, Pg 39-55.

[4] Harville D. A. 1983. Discussion on a section on interpolation and estimation, *Statistics: An Appraisal*, Pg 281–286.

[5] Rasche, R. and Shapiro, H. 1968. The F.R.B.-M.I.T. Economic model: its special features, *American Economic Review*, LVIII, 2, Pg 136-137.

[6] Malkeil, B. 1963. Equity yields, growth, and the structure of share prices, *American Economic Review*, LIII, 5, 1004-1031, Pg 13-25.

[7] Aronszajn, N. 1950. Theory of reproducing kernels, *Transactions of the American Mathematical Society*, 68, Pg 337–404, doi: 10.1090/s0002-9947-1950-0051437-7.