

Prediction of Adjusted Net National Income Per Capita of Countries

Introduction to the Research Question

The purpose of this study was to identify the best predictors for Adjusted Net National Income Per Capita of countries from multiple World Bank development indicators such as Exports of Goods and Services, Food Production Index, Foreign Direct Investment- Net Inflows, Forest Area, Gdp at Market Prices, and Gdp Growth

As someone who has had a keen interest in the development of countries and the work of the World Bank in helping the development of countries, I will like to find out about the best development factors that will lead to an increase or decrease in the Adjusted Net National Income Per Capita of Countries.

Such knowledge will enable countries, governments and stakeholders to make informed decisions and focus on the factors that will help increase their Adjusted Net National Income Per Capita without wasting much time and resources on indicators that do not count much. This, will in effect, help drive forward the development of each country.

Methods

Sample

The sample included N=248 countries and regions of the world for the years 2012 and 2013. The sample is made up of national, regional and global estimates. I am only looking at the 2012 data for this sample data. I will also be looking at all the 86 variables in the World Bank data for 2012 which will help predict the Adjusted Net National Income Per Capita of countries for that year – 2012 and also help make future predictions of subsequent years, such as the year, 2013

Measures

The Adjusted Net National Income Per Capita ,(Current Us\$), quantitative response variable, measures the Gross National Income (GNI) minus consumption of fixed capital and natural resources depletion.

Predictors used in the algorithm are all quantitative and they are:

Predictor	Meaning
x100_2012	Death Rate, Crude (Per 1,000 People)
x121_2012	Exports Of Goods And Services (% Of Gdp)
x125_2012	Fertility Rate, Total (Births Per Woman)
x126_2012	Fixed Broadband Subscriptions (Per 100 People)
x129_2012	Food Production Index (2004-2006 = 100)
x12_2012	Adjusted Savings: Carbon Dioxide Damage (% Of Gni)
x131_2012	Foreign Direct Investment, Net Inflows (% Of Gdp)
x132_2012	Foreign Direct Investment, Net Inflows (Bop, Current Us\$)
x134_2012	Forest Area (% Of Land Area)
x139_2012	Gdp At Market Prices (Current Us\$)
x140_2012	Gdp Growth (Annual %)
x142_2012	Gdp Per Capita (Current Us\$)
x143_2012	Gdp Per Capita Growth (Annual %)
x146_2012	Gross Domestic Savings (% Of Gdp)

x149_2012	Health Expenditure Per Capita (Current Us\$)
x14_2012	Adjusted Savings: Consumption Of Fixed Capital (% Of Gni)
x150_2012	Health Expenditure, Total (% Of Gdp)
x153_2012	Household Final Consumption Expenditure, Etc. (% Of Gdp)
x154_2012	Imports Of Goods And Services (% Of Gdp)
x155_2012	Improved Sanitation Facilities (% Of Population With Access)
x156_2012	Improved Water Source (% Of Population With Access)
x157_2012	Incidence Of Tuberculosis (Per 100,000 People)
x15_2012	Adjusted Savings: Consumption Of Fixed Capital (Current Us\$)
x161_2012	Industry, Value Added (% Of Gdp)
x162_2012	Inflation, Consumer Prices (Annual %)
x163_2012	Intentional Homicides (Per 100,000 People)
x167_2012	Internet Users (Per 100 People)
x169_2012	Labor Force, Female (% Of Total Labor Force)
x16_2012	Adjusted Savings: Education Expenditure (% Of Gni)
x171_2012	Life Expectancy At Birth, Female (Years)
x172_2012	Life Expectancy At Birth, Male (Years)
x173_2012	Life Expectancy At Birth, Total (Years)
x174_2012	Lifetime Risk Of Maternal Death (%)
x179_2012	Manufacturing, Value Added (% Of Gdp)
x187_2012	Mobile Cellular Subscriptions (Per 100 People)
x18_2012	Adjusted Savings: Energy Depletion (% Of Gni)
x190_2012	Mortality Rate, Infant (Per 1,000 Live Births)
x191_2012	Mortality Rate, Neonatal (Per 1,000 Live Births)
x192_2012	Mortality Rate, Under-5 (Per 1,000)

x195_2012	Net Migration	
x19_2012	Adjusted Savings: Energy Depletion (Current Us\$)	
x1_2012	Access To Electricity (% Of Population)	
x204_2012	Out-Of-Pocket Health Expenditure (% Of Total Expenditure On	Health)
x25_2012	Adolescent Fertility Rate (Births Per 1,000 Women Ages 15-19)	
x261_2012	Secure Internet Servers (Per 1 Million People)	
x268_2012	Surface Area (Sq. Km)	
x274_2012	Survival To Age 65, Female (% Of Cohort)	
x275_2012	Survival To Age 65, Male (% Of Cohort)	
x277_2012	Terrestrial And Marine Protected Areas (% Of Total Territorial	Area)
x283_2012	Urban Population (% Of Total)	
x284_2012	Urban Population Growth (Annual %)	
x29_2012	Age Dependency Ratio (% Of Working-Age Population)	
x2_2012	Access To Non-Solid Fuel (% Of Population)	
x31_2012	Agricultural Land (% Of Land Area)	
x35_2012	Agriculture, Value Added (% Of Gdp)	
x36_2012	Agriculture, Value Added (Annual % Growth)	
x37_2012	Air Transport, Passengers Carried	
x38_2012	Air Transport, Registered Carrier Departures Worldwide	
x45_2012	Arable Land (% Of Land Area)	
x47_2012	Armed Forces Personnel (% Of Total Labor Force)	
x205_2012	Percentage Of Students In Primary Education Who Are Female	(%)
x211_2012	Personal Remittances, Paid (Current Us\$)	
x212_2012	Personal Remittances, Received (% Of Gdp)	
x213_2012	Personal Remittances, Received (Current Us\$)	

x218_2012	Population Ages 65 And Above (% Of Total)	
x219_2012	Population Density (People Per Sq. Km Of Land Area)	
x21_2012	Adjusted Savings: Natural Resources Depletion (% Of Gni)	
x220_2012	Population Growth (Annual %)	
x221_2012	Population, Ages 0-14 (% Of Total)	
x222_2012	Population, Ages 15-64 (% Of Total)	
x223_2012	Population, Female (% Of Total)	
x242_2012	Private Credit Bureau Coverage (% Of Adults)	
x243_2012	Proportion Of Seats Held By Women In National Parliaments	(%)
x244_2012	Public Credit Registry Coverage (% Of Adults)	
x253_2012	Renewable Electricity Output (% Of Total Electricity Output)	
x255_2012	Renewable Internal Freshwater Resources Per Capita (Cubic	Meters)
x258_2012	Rural Population (% Of Total Population)	
x48_2012	Armed Forces Personnel, Total	
x49_2012	Automated Teller Machines (Atms) (Per 100,000 Adults)	
x58_2012	Birth Rate, Crude (Per 1,000 People)	
x67_2012	Cause Of Death, By Communicable Diseases And Maternal, Prenatal And Nutrition Conditions (% Of Total)	
x68_2012	Cause Of Death, By Injury (% Of Total)	
x69_2012	Cause Of Death, By Non-Communicable Diseases (% Of Total)	
x86_2012	Commercial Bank Branches (Per 100,000 Adults)	
x9_2012	Adjusted Net National Income (Current Us\$)	

Analyses

The distributions for the predictors and the Adjusted Net National Income Per Capita ,(Current Us\$), response variable were evaluated by calculating the mean, standard deviation and minimum and maximum values for the quantitative predictor variables. This is because all the predictors are quantitative.

Scatter plot was also examined, and Pearson correlation was used to test bivariate associations between individual quantitative predictors and the Adjusted Net National Income Per Capita ,(Current Us\$), response variable.

Lasso regression with the least angle regression selection algorithm was used to identify the subset of variables that best predicted Adjusted Net National Income Per Capita ,(Current Us\$). The lasso regression model was estimated on a training data set consisting of a random sample of 60% of the batches (N=148.8), and a test data set included the other 40% of the batches (N=99.2). All predictor variables were standardized to have a mean=0 and standard deviation=1 prior to conducting the lasso regression analysis. Cross validation was performed using k-fold cross validation specifying 10 folds. The change in the cross validation mean squared error rate at each step was used to identify the best subset of predictor variables. Predictive accuracy was assessed by determining the mean squared error rate of the training data prediction algorithm when applied to observations in the test data set

Results

The data included a lot of predictor variables for the 2012 year; a total of 85 predictor variables.

Therefore, I first of all, I conducted a Lasso regression analyses to only get the predictor variables that are relevant in predicting **The Adjusted Net National Income Per Capita ,(Current Us\$)** response variable.

As a result 74 of the variables were shrunk to zero (0) hence not retained in the model. And 11 of them were retained. The retained 11 are

- I. GDP PER CAPITA (CURRENT US\$),
- II. HEALTH EXPENDITURE PER CAPITA (CURRENT US\$),
- III. SECURE INTERNET SERVERS (PER 1 MILLION PEOPLE),
- IV. FIXED BROADBAND SUBSCRIPTIONS (PER 100 PEOPLE),
- V. ADJUSTED NET NATIONAL INCOME (CURRENT US\$),
- VI. SURVIVAL TO AGE 65, MALE (% OF COHORT),
- VII. POPULATION AGES 65 AND ABOVE (% OF TOTAL),
- VIII. ADJUSTED SAVINGS: NATURAL RESOURCES DEPLETION (% OF GNI),
- IX. RURAL POPULATION (% OF TOTAL POPULATION) ,
- X. ADJUSTED SAVINGS: CARBON DIOXIDE DAMAGE (% OF GNI),
- XI. EXPORTS OF GOODS AND SERVICES (% OF GDP))

Fig1 shows Table1 which gives descriptive Statistics for Data Analytic Variables which were retained in the model.

Fig1:

ANALYSIS VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM	MAXIMUM
x21_2012 - (ADJUSTED SAVINGS: NATURAL RESOURCES DEPLETION (% OF GNI))	55	4.02	5.32	0.00	21.31
x261_2012 - (SECURE INTERNET SERVERS (PER 1 MILLION PEOPLE))	55	414.44	714.58	0.28	3133.61
x142_2012 - (GDP PER CAPITA (CURRENT US\$))	55	17624.47	19493.07	270.09	83208.69
x121_2012 - (EXPORTS OF GOODS AND SERVICES (% OF GDP))	55	44.37	21.40	5.52	107.20
x126_2012 - (FIXED BROADBAND SUBSCRIPTIONS (PER 100 PEOPLE))	55	14.06	12.74	0.01	40.17
x218_2012 - (POPULATION AGES 65 AND ABOVE (% OF TOTAL))	55	10.31	6.28	2.35	21.16
x275_2012 - (SURVIVAL TO AGE 65, MALE (% OF COHORT))	55	73.47	12.02	46.08	90.02
x149_2012 - (HEALTH EXPENDITURE PER CAPITA (CURRENT US\$))	55	1567.58	2032.09	23.96	9070.51
x9_2012 - (ADJUSTED NET NATIONAL INCOME (CURRENT US\$))	55	320914115286.35	626979714689.37	1110573363.00	2990000000000.00
x12_2012 - (ADJUSTED SAVINGS: CARBON DIOXIDE DAMAGE (% OF GNI))	55	0.44	0.33	0.05	1.88
x258_2012 - (RURAL POPULATION (% OF TOTAL POPULATION))	55	38.89	19.86	2.27	84.20
x11_2012 - (ADJUSTED NET NATIONAL INCOME PER CAPITA (CURRENT US\$))	55	14137.21	15674.55	192.66	67688.51

The average Adjusted Net National Income Per Capita ,(Current Us\$) was 14137.21 (sd=15674.55). The predictor variable with the highest average was Adjusted Net National Income (Current Us\$) with a

mean of 320914115286.35 (sd= 626979714689.37) followed by Gdp Per Capita (Current Us\$) with a mean of 17624.47 (sd = 19493.07)

Bivariate Analyses

Scatter plots, Fig2 and Fig3, were generated to visualise the association between Adjusted Net National Income Per Capita ,(Current Us\$) and the quantitative predictor variables and Pearson correlation “r” values were calculated for the associations.

Fig2 Association Between Quantitative Predictors and x11_2012 - Adjusted Net National Income Per Capita ,(Current Us\$)- first 5

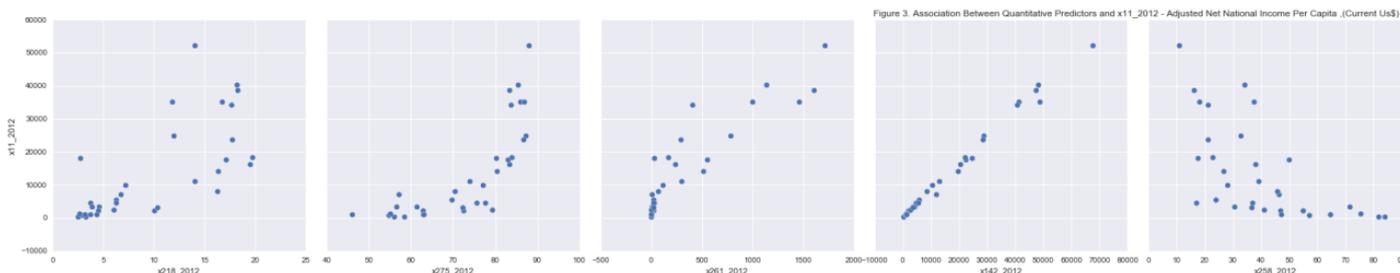
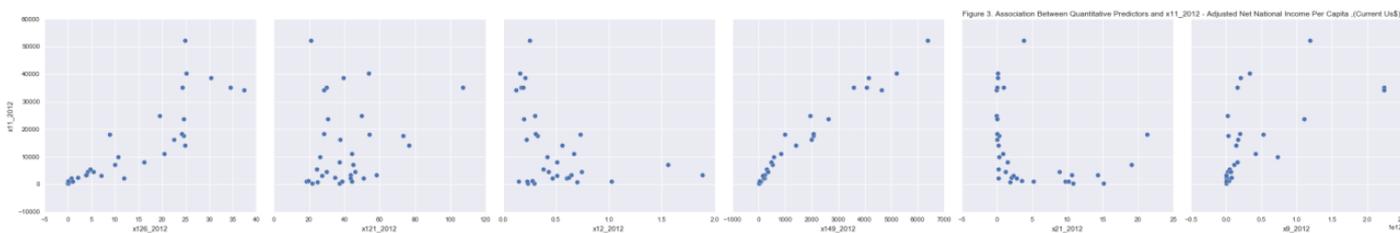


Fig 3 Association Between Quantitative Predictors and x11_2012 - Adjusted Net National Income Per Capita ,(Current Us\$) – last 6



The Pearson correlation “r” values and the associated p-values revealed that **all** the retained variables were significantly associated with The Adjusted Net National Income Per Capita ,(Current Us\$) as can be seen in the Table 2 below.

Table2

Table 2: Pearson Correlation "r" value and relative p-values for the association between quantitative predictor values and The Adjusted Net National Income Per Capita ,(Current Us\$)		
VARIABLE	PEARSON , r	p-value
x261_2012 - (SECURE INTERNET SERVERS (PER 1 MILLION PEOPLE))	0.82	0.0001
x12_2012 - (ADJUSTED SAVINGS: CARBON DIOXIDE DAMAGE (% OF GNI))	-0.42	0.0013
x121_2012 - (EXPORTS OF GOODS AND SERVICES (% OF GDP))	0.33	0.0145
x149_2012 - (HEALTH EXPENDITURE PER CAPITA (CURRENT US\$))	0.99	0.0001
x126_2012 - (FIXED BROADBAND SUBSCRIPTIONS (PER 100 PEOPLE))	0.87	0.0001
x258_2012 - (RURAL POPULATION (% OF TOTAL POPULATION))	-0.63	0.0001
x9_2012 - (ADJUSTED NET NATIONAL INCOME (CURRENT US\$))	0.57	0.0001
x275_2012 - (SURVIVAL TO AGE 65, MALE (% OF COHORT))	0.74	0.0001
x142_2012 - (GDP PER CAPITA (CURRENT US\$))	1.00	0.0001
x218_2012 - (POPULATION AGES 65 AND ABOVE (% OF TOTAL))	0.71	0.0001
x21_2012 - (ADJUSTED SAVINGS: NATURAL RESOURCES DEPLETION (% OF GNI))	-0.40	0.0026

From the Scatter Plots and the Pearson "r" values it can be seen that **GDP PER CAPITA (CURRENT US\$)** ($r = 1.00$, $p = .0001$) is the most strongest relation to the response variable, **Adjusted Net National Income Per Capita ,(Current Us\$)** followed by **HEALTH EXPENDITURE PER CAPITA (CURRENT US\$)** ($r = 0.99$, $p = .0001$).

Multivariable analyses

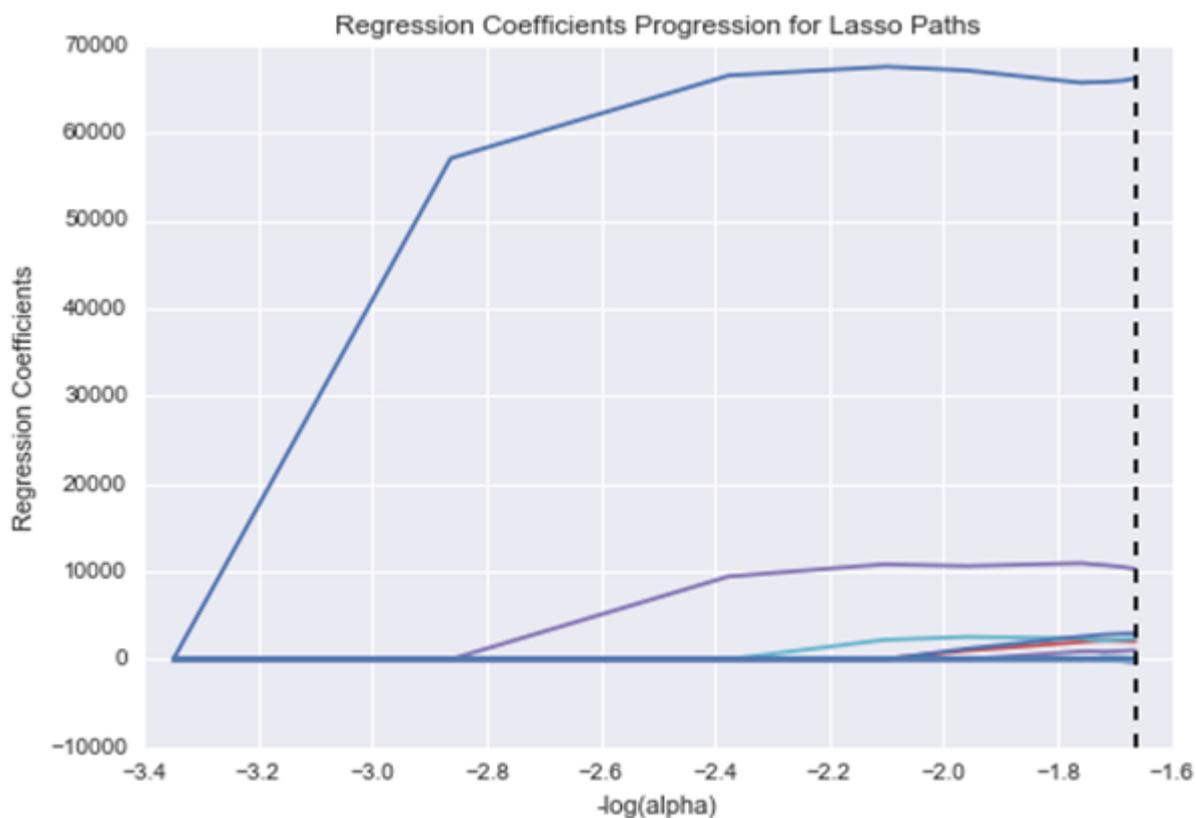
Table 3 shows the retained predictor variables and their respective model co-efficients.

Table 3

Table 3: Model Co-efficients of Retained Predictors	
Predictor	Co-efficient
x142_2012 - (GDP PER CAPITA (CURRENT US\$))	11949.37828
x149_2012 - (HEALTH EXPENDITURE PER CAPITA (CURRENT US\$))	1996.271328
x261_2012 - (SECURE INTERNET SERVERS (PER 1 MILLION PEOPLE))	711.9843858
x126_2012 - (FIXED BROADBAND SUBSCRIPTIONS (PER 100 PEOPLE))	431.2711585
x9_2012 - (ADJUSTED NET NATIONAL INCOME (CURRENT US\$))	353.5999677
x275_2012 - (SURVIVAL TO AGE 65, MALE (% OF COHORT))	164.2523971
x218_2012 - (POPULATION AGES 65 AND ABOVE (% OF TOTAL))	18.70327262
x21_2012 - (ADJUSTED SAVINGS: NATURAL RESOURCES DEPLETION (% OF GNI))	-9.633017391
x258_2012 - (RURAL POPULATION (% OF TOTAL POPULATION))	-17.73558659
x12_2012 - (ADJUSTED SAVINGS: CARBON DIOXIDE DAMAGE (% OF GNI))	-31.87655894
x121_2012 - (EXPORTS OF GOODS AND SERVICES (% OF GDP))	-69.89047621

From the Table 3 and also Fig 4 below:

Fig 4



it can be seen that **GDP PER CAPITA (CURRENT US\$)** is most strongly associated with **Adjusted Net National Income Per Capita ,(Current Us\$)** and this is followed by **HEALTH EXPENDITURE PER CAPITA (CURRENT US\$)** . This corroborates the -Pearson correlation “r” results. The **Adjusted Net National Income Per Capita ,(Current Us\$)** increases when **GDP PER CAPITA (CURRENT US\$)** increases. The response variable also increases when **HEALTH EXPENDITURE PER CAPITA (CURRENT US\$)** increases. The reverse is true. These were the 2 strongest variables that is positively associated with The Adjusted Net National Income Per Capita ,(Current Us\$)- the target variable. And increase in

- SECURE INTERNET SERVERS (PER 1 MILLION PEOPLE),
- FIXED BROADBAND SUBSCRIPTIONS (PER 100 PEOPLE),
- ADJUSTED NET NATIONAL INCOME (CURRENT US\$),
- SURVIVAL TO AGE 65, MALE (% OF COHORT),
- POPULATION AGES 65 AND ABOVE (% OF TOTAL)

all result in an increase in the The Adjusted Net National Income Per Capita ,(Current Us\$) and vice versa. Hence there is a positive relationship between the Adjusted Net National Income Per Capita ,(Current Us\$) and the above stated 7 variables.

However, it can be seen that for

- ADJUSTED SAVINGS: NATURAL RESOURCES DEPLETION (% OF GNI),
- RURAL POPULATION (% OF TOTAL POPULATION),
- ADJUSTED SAVINGS: CARBON DIOXIDE DAMAGE (% OF GNI) and
- EXPORTS OF GOODS AND SERVICES (% OF GDP)

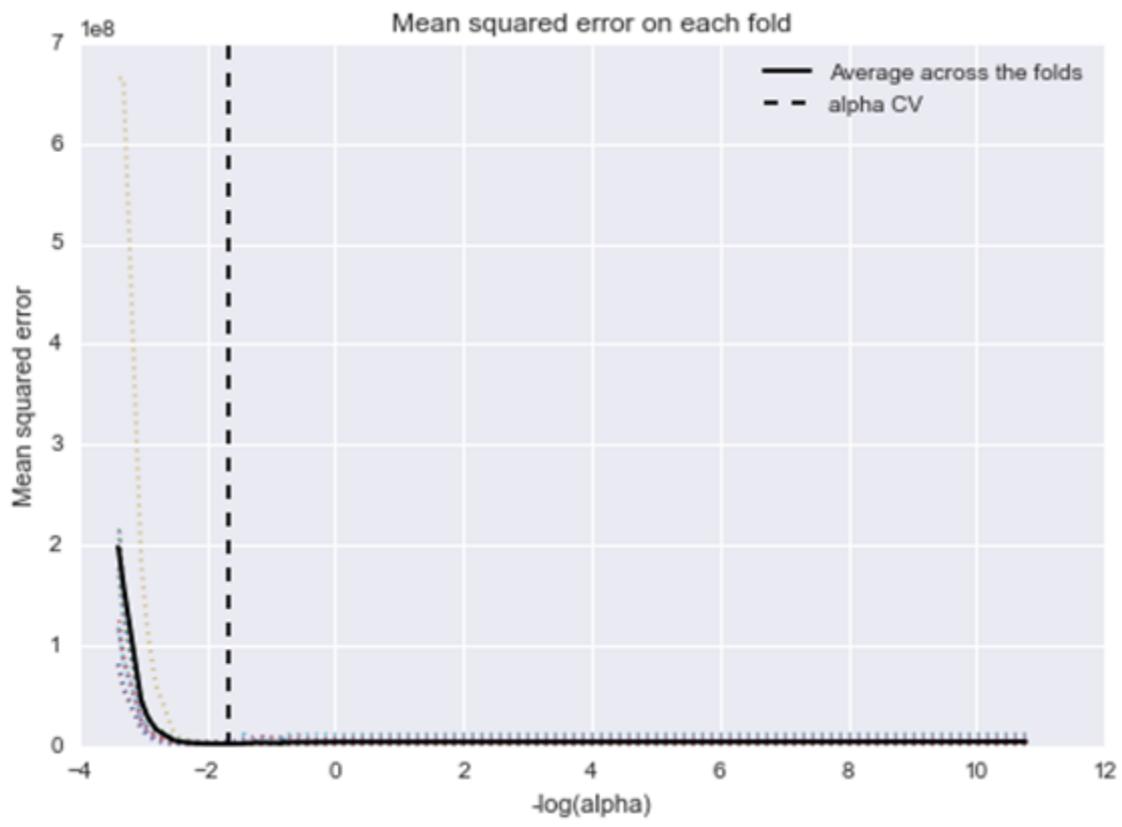
an increase in any of these results in a decrease in the Adjusted Net National Income Per Capita ,(Current Us\$) and vice versa. Hence there is a negative relationship between these 4 predictor variables and **Adjusted Net National Income Per Capita ,(Current Us\$)**

Together, these 11 predictors accounted for 99.5% (almost 100%) of the variance in **The Adjusted Net National Income Per Capita ,(Current Us\$)**. The mean squared error (MSE) for the test data (MSE= 1607604.93409) differed slightly from the MSE for the training data (MSE= 812548.990714).

MSE -TRAINING DATA	812548.990714
MSE -TEST DATA	1607604.93409
R-Square - TRAINING DATA	0.995739422921
R-Square - TEST DATA	0.99525790594

This suggests that predictive accuracy may decline a little bit when the lasso regression algorithm developed on the training data set is applied to predict future **Adjusted Net National Income Per Capita ,(Current Us\$)** as seen in (Figure 5) below.

Fig5



Conclusions/Limitations

Brief Overview Of Key Findings And Implications

This project used lasso regression analysis to identify the best predictors for Adjusted Net National Income Per Capita, (Current Us\$) of countries from multiple World Bank development indicators in N= 248 countries and regions of the world for the year 2012 which is made up of national, regional and global estimates. The **Adjusted Net National Income Per Capita, (Current Us\$)** for this period ranged from **192.66 (Current Us\$)** to **67688.51 (Current Us\$)** indicating that there was considerable variability in the **Adjusted Net National Income Per Capita, (Current Us\$)** of countries and regions of the world for that year.

The prediction accuracy of the model was 0.995739422921 (96%) on the training dataset and 0.99525790594 (99.5%) when ran on the test dataset. Hence, the algorithm helped to identify the best predictors for Adjusted Net National Income Per Capita, (Current Us\$) of countries for the year 2012 given the indicators available in this World Bank dataset.

There was significant increase in the MSE when the training set lasso regression algorithm was used to predict the Adjusted Net National Income Per Capita, (Current Us\$) in the test data set. This suggests that the predictive accuracy of the algorithm may not be very stable for future datasets and hence has to be looked into further by using other analytic methods such as Multiple Regression.

The Pearson correlation “r” values and associated p-values revealed that **all** the retained variables were significantly associated with **Adjusted Net National Income Per Capita ,(Current Us\$)** as can be seen in the Table 2 above. From the Scatter Plots and the Pearson “r” values and also from the multivariate analysis it can be seen that **GDP PER CAPITA (CURRENT US\$)** is the most strongest relation to the **Adjusted Net National Income Per Capita ,(Current Us\$)** followed by **HEALTH EXPENDITURE PER CAPITA (CURRENT US\$)**.

This means that countries and regions should spend more on increasing their GDP PER CAPITA (CURRENT US\$) and also they should make more efforts and allocate more resources to HEALTH EXPENDITURE PER CAPITA (CURRENT US\$) as the more these two variables increase, there is more likelihood of increase in their Adjusted Net National Income Per Capita ,(Current Us\$). Although,

- SECURE INTERNET SERVERS (PER 1 MILLION PEOPLE),
- FIXED BROADBAND SUBSCRIPTIONS (PER 100 PEOPLE),

- ADJUSTED NET NATIONAL INCOME (CURRENT US\$),
- SURVIVAL TO AGE 65, MALE (% OF COHORT),
- POPULATION AGES 65 AND ABOVE (% OF TOTAL)

all indicated a positive co-efficient ,meaning an increase in these variables are likely to lead to increase in the Adjusted Net National Income Per Capita ,(Current Us\$) of countries and regions, due to general scarcity of resources globally, if there is only a few resources at the disposal of countries and regions, it will be advisable to allocate this scarce resources to **GDP PER CAPITA (CURRENT US\$)** and **HEALTH EXPENDITURE PER CAPITA (CURRENT US\$)**. This is because these were the 2 that indicated the most strongest positive association with **Adjusted Net National Income Per Capita ,(Current Us\$)** .

On the other hand, it is advisable for countries and regions to reduce their

- ADJUSTED SAVINGS: NATURAL RESOURCES DEPLETION (% OF GNI),
- RURAL POPULATION (% OF TOTAL POPULATION),
- ADJUSTED SAVINGS: CARBON DIOXIDE DAMAGE (% OF GNI) and
- EXPORTS OF GOODS AND SERVICES (% OF GDP)

as a decrease in these variables have the tendency to increase their Adjusted Net National Income Per Capita ,(Current Us\$) as these factors are negatively associated with the Adjusted Net National Income Per Capita ,(Current Us\$) .

Limitations

Although the p-values associated with the Pearson correlation “r” values indicated that all the retained 11 variables were significantly associated with The Adjusted Net National Income Per Capita ,(Current Us\$) , it can be seen that some of the associations were not linear hence the Pearson correlation could not capture the true relationship between those indicators and The Adjusted Net National Income Per Capita ,(Current Us\$); therefore could not help in interpreting the Adjusted Net National Income Per Capita ,(Current Us\$) variability given those non-linear associations.

In effect, though the model accurately predicts 99.5% (almost 100%) of the variance in The Adjusted Net National Income Per Capita ,(Current Us\$), with all the indicators combined, a further examination of the variables using other analysis like Multiple Regression will further enlighten the true associations between these non-linear curves and The Adjusted Net National Income Per Capita ,(Current Us\$).

What is more, as indicated in the codebook, this World Bank data set is a subset of data extracted from the primary World Bank collection of development indicators, which means there are other development indicators which were not provided and hence not captured in this analysis. It is possible that the indicators identified as important predictors of The Adjusted Net National Income Per Capita ,(Current Us\$), among the set of predictors analyzed in this project are confounded by other factors not considered in this analysis,

As a result, these same factors may not emerge as important factors when other factors are taken into consideration. Therefore, future efforts to develop a solid predictive algorithm for The Adjusted Net National Income Per Capita, (Current Us\$), should expand the algorithm by adding more development indicators to the statistical model, and testing the applicability of the algorithm for more other years.