# The Association between GDP Per Capita and Various Development Indicators

Greg Rafferty

# Introduction to the Research Question

The purpose of this study was to identify the best predictors of a country's GDP per capita from multiple indicators common to development goals, that is, fixed broadband subscription rate, access to improved water sources, internet user rate, mortality rate for under 5-year olds, proportion of seats held by women in national parliaments, rural population rate, urban population rate, and the birth rate.

I spent spent my career living in developing countries while working for an international aid agency. By investigating which development goals most effectively lead to increases in GDP per capita, all aid agencies globally will be better able to focus their activities towards improvement of those indicators, which will indirectly lead to increased incomes for families.

I have witnessed firsthand how a small income boost to impoverished communities can lead dramatic improvement in quality of life in those communities. By focusing development efforts on improving the indicators which are most closely correlated to GDP per capita, aid agencies can have the greatest impact towards their end goals.

# Methods

## Sample

The World Bank data set is a subset of data extracted from the primary World Bank collection of development indicators, compiled from officially-recognized international sources, from the years 2012 and 2013. The sample used in this study contained N = 248 countries with data aggregated from national, regional, and global estimates. Although all variables have valid data observations for a minimum of 190 countries, only data from the year 2012 was used in this study because this was not intended to be a longitudinal study. Furthermore, 2013 data has been reserved to test the regression equation developed by the 2012 data for validity.

## Measures

The GDP per capita, in current US dollars, is the quantitative response variable. Gross Domestic Product is a measure of output of a country, created by taking the monetary value of all finished goods and services produced within the country's borders during a specific time period, and dividing it by the country's population. It is a commonly used indicator of standard of living, with higher GDP per capita equating to a higher standard of living.

Predictors used in the study are all quantitative, and they are:

| Variable | Label |
|---|---|
| x142_2012 | GDP PER CAPITA (CURRENT US$) |
| x126_2012 | FIXED BROADBAND SUBSCRIPTIONS (PER 100 PEOPLE) |
| x156_2012 | IMPROVED WATER SOURCE (% OF POPULATION WITH ACCESS) |
| x167_2012 | INTERNET USERS (PER 100 PEOPLE) |
| x192_2012 | MORTALITY RATE, UNDER-5 (PER 1,000) |
| x243_2012 | PROPORTION OF SEATS HELD BY WOMEN IN NATIONAL PARLIAMENTS (%) |
| x258_2012 | RURAL POPULATION (% OF TOTAL POPULATION) |
| x283_2012 | URBAN POPULATION (% OF TOTAL) |
| x58_2012 | BIRTH RATE, CRUDE (PER 1,000 PEOPLE) |

## Analyses

Because all variables are quantitative, the distributions for the predictors and the GDP per capita response variable were evaluated by calculating the mean, standard deviation, and minimum and maximum.

A scatter plot for each variable was also examined, and a Pearson correlation was used to test bivariate associations between the individual quantitative predictors and the GDP per capita response variable. Many plots appeared to display a logarithmic relationship between the predictor and response, so a Pearson correlation was calculated from the log of the predictor and the response. When the log correlation was more significant than the linear correlation, the log relationship was reported.

Lastly, a lasso regression with the least angle regression selection algorithm was used to identify the subset of variables that best predicted GDP per capita. The lasso regression model was estimated on a training data set consisting of the 2012 data (N = 248) and tested on the 2013 data. 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

## Descriptive Statistics

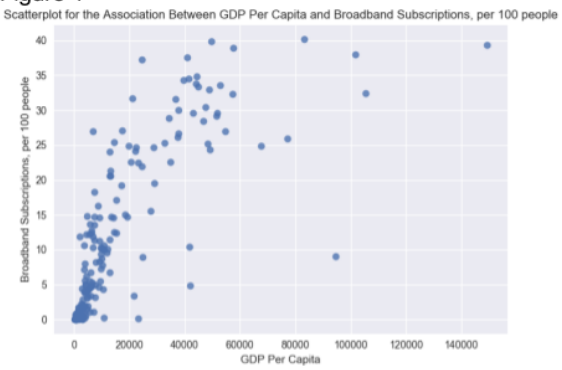
Table 1 shows descriptive statistics for GDP per capita and the quantitative predictors. The average GDP per capita was US$ 14,751 (std dev = 21,612) with a minimum GDP of US$ 244.2 and a maximum of US$ 149,161

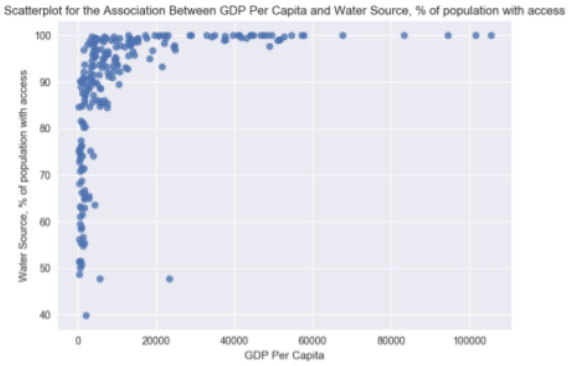
| **Table 1** | | | | | |
|---|---|---|---|---|---|
| Analysis Variables | N | Mean | Std Dev | Min | Max |
| Fixed Broadband Subscriptions, % | 199 | 11.6 | 12.2 | 0.00016 | 43.2 |
| Improved Water Source, % | 227 | 88.3 | 14.5 | 39.9 | 100 |
| Internet Users, % | 234 | 39.9 | 28.0 | 0.00 | 96.2 |
| Mortality Rate, per 1000 | 225 | 36.0 | 35.5 | 2.1 | 172 |
| Women in Parliament, % | 220 | 19.2 | 10.7 | 0.00 | 56.3 |
| Rural Population, % | 245 | 42.2 | 23.5 | 0.00 | 91.2 |
| Urban Population, % | 245 | 57.8 | 23.5 | 8.8 | 100 |
| Birth Rate, per 1000 | 237 | 21.4 | 10.4 | 8.2 | 49.9 |
| GDP Per Capita (US$) | 225 | 14,751 | 21,612 | 244.2 | 149,160 |

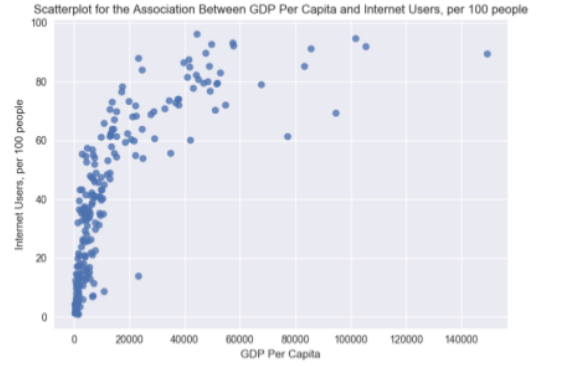
## Bivariate Analyses

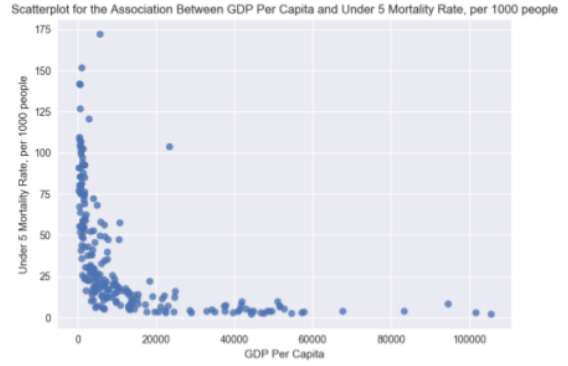
Scatter plots for the association between the GDP per capita response variable and the quantitative predictors (Figure 1) revealed that GDP per capita has the strongest correlation with the rate of fixed broadband subscriptions (Pearson $r = 0.789$, $p < .0001$). All correlations are shown in Table 2.

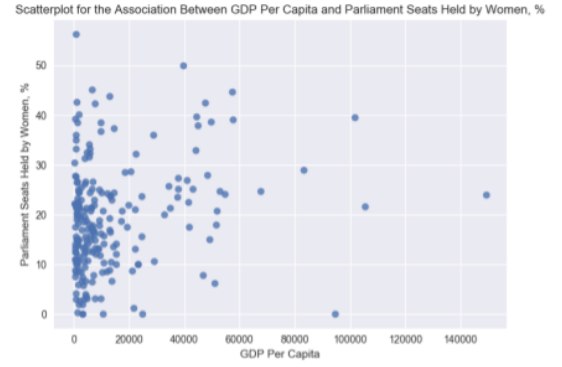
Figure 1


Scatterplot for the Association Between GDP Per Capita and Broadband Subscriptions, per 100 people
Broadband Subscriptions, per 100 people
GDP Per Capita

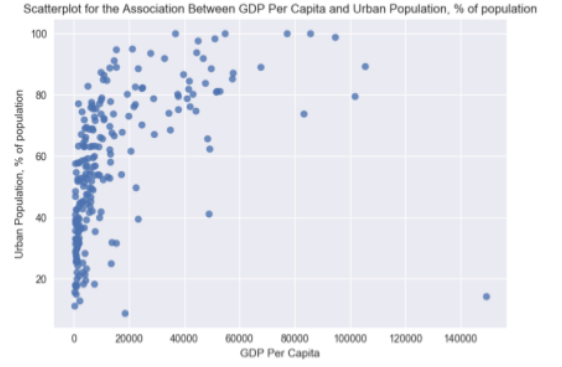

Scatterplot for the Association Between GDP Per Capita and Water Source, % of population with access
Water Source, % of population with access
GDP Per Capita


Scatterplot for the Association Between GDP Per Capita and Internet Users, per 100 people
Internet Users, per 100 people
GDP Per Capita


Scatterplot for the Association Between GDP Per Capita and Under 5 Mortality Rate, per 1000 people
Under 5 Mortality Rate, per 1000 people
GDP Per Capita


Scatterplot for the Association Between GDP Per Capita and Parliament Seats Held by Women, %
Parliament Seats Held by Women, %
GDP Per Capita


Scatterplot for the Association Between GDP Per Capita and Rural Population, % of population
Rural Population, % of population
GDP Per Capita


Scatterplot for the Association Between GDP Per Capita and Urban Population, % of population
Urban Population, % of population
GDP Per Capita


Scatterplot for the Association Between GDP Per Capita and Birth Rate, per 1000 people
Birth Rate, per 1000 people
GDP Per Capita

| Table 2 | | |
|---|---|---|
| Variable | Pearson r | p-value |
| Fixed broadband subscription rate | 0.789 | 1.00e-38 |
| Improved water source rate | 0.458 | 1.76e-16 |
| Internet User rate | 0.776 | 9.78e-37 |
| Log of Under 5 Mortality Rate | -0.710 | 7.68e-28 |
| Proportion of Parliament Seats Held by Women | 0.253 | 0.000692 |
| Rural Population Rate | -0.619 | 5.03e-20 |
| Urban Population Rate | 0.619 | 5.03e-20 |
| Log of Birth Rate | -0.579 | 1.35e-17 |

Note that the *Under 5 Mortality Rate* was transformed to a logarithmic scale in order to linearize it for more accurate regression analysis. Figure 2 shows the scatterplot resulting from this transformation, showing the increased linearity.

Figure 2


## Lasso Regression Analysis

Of the 8 predictor variables analyzed, the lasso regression model dropped 2 of them. The proportion of women in national parliament was not retained (the regression analysis also showed it to have the least significance by a wide margin), and the rural population rate was not retained (the urban population rate captures the same information, in the inverse, and also had a *slightly* higher correlation than the rural population rate; the urban population rate was retained). Table 3 shows the regression coefficients of the lasso analysis.

| Table 3 | |
|---|---|
| Variable | Regression coefficient |
| Fixed broadband subscription rate | 9587 |
| Improved water source rate | -31.82 |
| Internet User rate | 9030 |
| Under 5 Mortality Rate | 397.5 |
| Urban Population Rate | 2329 |
| Birth Rate | 4875 |

GDP per capita increases the most with the fixed broadband subscription rate and the linked variable of internet user rate. The weakest predictor retained in the model was the rate of access to an improved water source. Indeed, nearly all of the countries in the survey have a 100% rate, so not much information was provided by this data point. Together, these 6 variables can explain 68% of the variability of GDP per capita (Training data R-square = 0.6827; Test data R-square = 0.6806). The mean squared error for the training set was 124390237 and for the testing set was 123487565.

## Conclusion and Limitations

This project used lasso regression analysis to identify a subset of development-related variables that best predicted Gross Domestic Product per capita in N = 248 countries and regions in the world during the year 2012. GDPs per capita ranged from $244 to $150,000, indicating that there was considerable variability in the distribution of incomes around the world.

The lasso regression analysis indicated that 6 of the 8 development-related predictor variables were selected in the final model. These 6 predictors accounted for 68% of the observed variability in GDPs per capita. Only the rural population rate and proportion of seats held by women in national parliaments were excluded. THe strongest predictors of GDPs per capita were the fixed broadband subscription rate and the internet user rate. GDPs per capita were greater when both fixed broadband subscription rates and internet user rates increased. Access rates to improved water sources paradoxically emerged as a predictor of *lower* GDPs per capita, but the regression coefficient is small relative to the other predictors, so the effect is negligible.

There was a slight decrease in the mean squared error when the 2012 set lasso regression algorithm was used to predict GDPs per capita in the 2013 set. This suggests that the predictive accuracy of the algorithm may be stable in future samples of GDPs per capita.

The results of this project indicate that efforts to increase internet penetration rates and get more users connected in a country are a priority for achieving consistently higher GDPs per capita. These results should be considered carefully when deciding how to spend development aid money. Although higher GDPs per capita are frequently used as a proxy measure for higher quality-of-life, other indicators such as mortality rates and access to trade markets for goods and services indirectly lead to higher GDPs per capita but directly lead to higher quality-of-life. Therefore, if aid money is to be spent on increasing internet penetration rates, it must be done so minimally so that aid funds remain available to be spent on other indicators.

Although access rates to improved water sources and birth rate were both retained in the final lasso regression model, Table 2 showed that the bivariate analyses indicated these variables to be only weakly correlated with GDP per capita. Table 3 shows that the lasso regression model did not weigh these predictors highly in the final model. This suggests that these two development-related factors may not have a considerable impact on GDPs per capita, and re-analysis of these factors should be conducted on future datasets to determine whether this conclusion is supported.

This project successfully developed a predictive algorithm for GDPs per capita that appears to have little bias and variance in a different sample. In addition, it provides more information on which development-related factors are most likely to have a significant impact on GDPs per capita. However, there are some limitations that should be taken into account when considering changes in the way funds are spent in the international development field based on the results of this project. First, we analyzed only data from a single year, but development of countries is ongoing. So, it is important to test this algorithm on datasets that are surveyed in multiple years to determine whether the algorithm remains relatively unbiased and stable despite these ongoing changes. Second, the analysis was conducted on only a subset of the available development data. Different factors may emerge as the important predictors of GDPs per capita when using the full collection of potential predictors. Therefore, we cannot assume that the predictive algorithm developed in this study will be the best possible when different development indicators available. Finally, there is a large number of development-related factors that could impact GDPs per capita, but the current project examined only a few of these factors. It is possible that the factors identified as important predictors of GDPs per capita 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 GDPs per capita should expand the algorithm by adding more development-related predictors to the statistical model, and evaluating the applicability of the algorithm with this greater dataset.

# Appendix

## Python code

```
'''
Greg Rafferty
The Association between GDP Per Capita and Various Development Indicators
Analysis of World Bank data from 2012, with verification on 2013 data
'''

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


data = pd.read_csv('worldbank.csv', low_memory=False)


# Ignore warnings
import warnings
warnings.filterwarnings("ignore")
# Suppress warnings for chained indexing
pd.options.mode.chained_assignment = None

print ("Number of observations:", len(data)) #number of observations (rows)
print ("Number of variables:", len(data.columns)) # number of variables (columns)
print ('')

# Codebook
'''
Variable      Label
x142_2012     GDP PER CAPITA (CURRENT US$)
x126_2012     FIXED BROADBAND SUBSCRIPTIONS (PER 100 PEOPLE)
x156_2012     IMPROVED WATER SOURCE (% OF POPULATION WITH ACCESS)
x167_2012     INTERNET USERS (PER 100 PEOPLE)
x192_2012     MORTALITY RATE, UNDER-5 (PER 1,000)
x243_2012     PROPORTION OF SEATS HELD BY WOMEN IN NATIONAL PARLIAMENTS (%)
x258_2012     RURAL POPULATION (% OF TOTAL POPULATION)
x283_2012     URBAN POPULATION (% OF TOTAL)
x58_2012      BIRTH RATE, CRUDE (PER 1,000 PEOPLE)
'''

sub1 = data[['x142_2012','x126_2012','x156_2012','x167_2012','x192_2012',
             'x243_2012','x258_2012','x283_2012','x58_2012']]

sub2 = data[['x142_2013','x126_2013','x156_2013','x167_2013','x192_2013',
```

```
            'x243_2013','x258_2013','x283_2013','x58_2013']]

# Map codes to information for titles and captions
info = {'x142_2012':'GDP per capita',
        'x126_2012':'Broadband Subscriptions, per 100 people',
        'x156_2012':'Water Source, % of population with access',
        'x167_2012':'Internet Users, per 100 people',
        'x192_2012':'Under 5 Mortality Rate, per 1000 people',
        'x243_2012':'Parliament Seats Held by Women, %',
        'x258_2012':'Rural Population, % of population',
        'x283_2012':'Urban Population, % of population',
        'x58_2012':'Birth Rate, per 1000 people'}

# Set the variables to numeric
for variable in sub1:
    sub1[variable] = pd.to_numeric(sub1[variable], errors='coerce')


#=============================================================================
# Describe the variables
#=============================================================================
for variable in sub1:
    print (info[variable])
    print (sub1[variable].describe())
    print ('')


#=============================================================================
# Print univariate histograms
#=============================================================================
for variable in sub1:
    plt.figure()
    sns.distplot(sub1[variable].dropna(), kde=False);
    plt.ylabel('Count of countries')
    plt.title('Histogram for ' + info[variable])


#=============================================================================
# Bivariate scatterplots
#=============================================================================
for variable in sub1.columns[1:]:
    plt.figure()
    scat1 = sns.regplot(x="x142_2012", y=variable, fit_reg=False, data=sub1)
    plt.xlabel('GDP Per Capita')
    plt.ylabel(info[variable])
    plt.title('Scatterplot for the Association Between GDP Per Capita and '
              + info[variable])


#=============================================================================
```

```
# Bivariate scatterplots with correlation coefficient for regression analysis
#==============================================================================
# Create dataframe from original data with NaNs removed
data_clean = sub1.dropna()
# Create datafram from logarithms of original data
data_log = np.log(data_clean.astype('float64'))

# Inspect a linear relationship
for variable in sub1.columns[1:]:
    print ('Association between ' + info[variable] + ' and GDP Per Capita')
    print (scipy.stats.pearsonr(data_clean[variable], data_clean['x142_2012']))
    print ('')

    plt.figure()
    scat1 = sns.regplot(x="x142_2012", y=variable, fit_reg=True, data=data_clean)
    plt.xlabel('GDP Per Capita')
    plt.ylabel(info[variable])
    plt.title('Scatterplot for the Association Between GDP Per Capita and '
              + info[variable])


    # Inspect a logarithmic relationship
    print ('Association between log of ' + info[variable] + ' and GDP Per Capita')
    print (scipy.stats.pearsonr(data_log[variable], data_clean['x142_2012']))
    print ('')

    plt.figure()
    scat1 = sns.regplot(x="x142_2012", y=variable, fit_reg=True, data=data_log)
    plt.xlabel('GDP Per Capita')
    plt.ylabel('Log of ' + info[variable])
    plt.title('Scatterplot for the Association Between GDP Per Capita and log of '
              + info[variable])


#==============================================================================
# Run a Lasso Regression Analysis
#==============================================================================
from sklearn import preprocessing
from sklearn.linear_model import LassoLarsCV

# Select predictor variables and target variable as separate data sets
predvar = data_clean[['x126_2012','x156_2012','x167_2012','x192_2012',
                      'x243_2012','x258_2012','x283_2012','x58_2012']]

target = data_clean.x142_2012

# Standardize predictors to have mean=0 and sd=1
predictors = predvar.copy()
for variable in predictors:
    predictors[variable]=preprocessing.scale(predictors[variable]
```

```
    .astype('float64'))


# Create test variables from corresponding 2013 data
data_clean2 = sub2.dropna()
target_test = data_clean2.x142_2013

predvar_test = data_clean2[['x126_2013','x156_2013','x167_2013','x192_2013',
                            'x243_2013','x258_2013','x283_2013','x58_2013']]

predictors_test = predvar_test.copy()
for variable in predictors_test:
    predictors_test[variable]=preprocessing.scale(predictors_test[variable]
    .astype('float64'))

pred_train = predictors
pred_test = predictors_test
tar_train = target
tar_test = target_test


# Specify the lasso regression model
model=LassoLarsCV(cv=10, precompute=False).fit(pred_train,tar_train)

# Print variable names and regression coefficients
print ('Regression coefficients')
for idx in range(0, len(predictors.columns)):
    print (predictors.columns[idx], ':', model.coef_[idx],
           '    (' + info[predictors.columns[idx]] + ')')
#print (dict(zip(predictors.columns, model.coef_)))
print ('')

# Plot coefficient progression
plt.figure()
m_log_alphas = -np.log10(model.alphas_)
ax = plt.gca()
plt.plot(m_log_alphas, model.coef_path_.T)
plt.axvline(-np.log10(model.alpha_), linestyle='--', color='k',
            label='alpha CV')
plt.ylabel('Regression Coefficients')
plt.xlabel('-log(alpha)')
plt.title('Regression Coefficients Progression for Lasso Paths')

# Plot mean square error for each fold
m_log_alphascv = -np.log10(model.cv_alphas_)
plt.figure()
plt.plot(m_log_alphascv, model.cv_mse_path_, ':')
plt.plot(m_log_alphascv, model.cv_mse_path_.mean(axis=-1), 'k',
         label='Average across the folds', linewidth=2)
plt.axvline(-np.log10(model.alpha_), linestyle='--', color='k',
```

```
            label='alpha CV')
plt.legend()
plt.xlabel('-log(alpha)')
plt.ylabel('Mean squared error')
plt.title('Mean squared error on each fold')


# MSE from training and test data
from sklearn.metrics import mean_squared_error
train_error = mean_squared_error(tar_train, model.predict(pred_train))
test_error = mean_squared_error(tar_test, model.predict(pred_test))
print ('training data MSE')
print(train_error)
print ('')
print ('test data MSE')
print(test_error)
print ('')

# R-square from training and test data
rsquared_train=model.score(pred_train,tar_train)
rsquared_test=model.score(pred_test,tar_test)
print ('training data R-square')
print(rsquared_train)
print ('')
print ('test data R-square')
print(rsquared_test)
```