Multiple Linear Regression Analysis of Human Development Index (HDI) using R

Rohan Narayan Koli
MSc. in Data Analytics 2020-21
Student Id : x19224842
December 02, 2020

***Abstract*—The Human development index (HDI) is used to measure an individuals standard of living and ultimately the socio-economic achievement of a country. This paper reviews the use of alternate predictors for HDI calculation such as: per capita GDP, expenditure on health (in % of GDP), homicide rates, proportion of safely managed drinking water sources, percentage of individuals using the internet, Consumer price index (CPI) and carbon-dioxide emissions per capita metric tons using multiple linear regression. Various regression models were evaluated and analyzed to achieve a trimmed-down desirable model with only three factors with most accuracy.**

***Index Terms*—Multiple Linear Regression, HDI, Gauss Markov assumptions**

I. Introduction

Human Development Index (HDI) is a composite index used to capture three main aspects of human development namely a better standard of living, a health life coupled with longevity and access to knowledge [1]. The health dimension is measured by life expectancy at birth, the education dimension is measured by mean schooling years for adults older than 25 years coupled with expected schooling years for children of school going age. lastly, standard of living dimension is measured by gross national income per capita. Consequently, the HDI is calculated by taking the normalized geometric mean of these three dimensions.

$$HDI = \sqrt[3]{LEI \times EI * II} \tag{1}$$

where, LEI is the Life Expectancy Index, EI is the Education Index and II is the Income Index [2]. For this study, we shortlisted proxy components to analyze the correlations with other socio-economic factors which can also be used to depict the original model proposed by the economist, Mahbub ul Haq. The factors implemented in this model are as follows:

- Per capita GDP
- Expenditure on health as a percentage of GDP
- Homicide rates per 1,00,000
- Proportion of safely managed drinking water sources as a proportion of population
- Percentage of individuals using the internet
- Consumer price index (CPI)
- Carbon-dioxide emissions per capita metric tons

II. DATA

A. Data Description

For the purpose of our study, the data from the year 2017 was collated from the website https://data.un.org. The data was merged using Microsoft Excel and missing values were removed.

1) Dependent Variable: Human development index (HDI) is selected as our dependent variable. Its value ranges from 0 to 1 and the countries are categorized in four categories viz. from 0.8 and above as very high, between 0.7 to 0.8 as high,between 0.555 to 0.7 as medium and below 0.55 as low. The HDI depicts the standard of living of an individual in a country, the higher the value states better is the standard of living.

2) Independent variables: The first independent variable is *GDP per capita* is the monetary worth of a country divided among the total population. It is expressed in US dollar terms. The GDP per capita closely represents the average holding of an individual in an economy [3]. The second independent variable considered is the *governments expenditure on health facilities* expressed as a percentage of the total GDP. This factor defines how important the health of an individual is for a country. The third factor is the *homicide rates per 1,00,000 individual* which depicts the safety and security of an individual living in a country. The fourth factor is the *proportion of safely managed drinking water sources* expressed as a proportion of total population. More availability of drinking water corresponds to better living conditions in a country [4]. The fifth factor is the *percentage of individuals accessing the internet*, which signifies the access to technology and technological advances for an individual which in current times is necessary for obtaining knowledge. The sixth factor is the *consumer price index (CPI)*, which reflects the basic cost of living and the periods of inflation or deflation in an economy. Lastly, the seventh factor is the *carbon dioxide emissions* expressed as per capita metric tons. This factor reflects the extent of air and water pollution.

All the dependent and independent variables are continuous. According to the thumb rule, we should have *50+8m* records for testing multiple correlations and *104+m* records for testing individual predictors, where m is the number of independent variables [5]. We have a record of total 160 observations

which suffice the conditions. Fig. 1 summarizes the variables collected for 160 countries.

```
> summary(data)
      HDI            Emission            GDP              Health      
 Min.   :0.3700   Min.   : 0.0500   Min.   :  0.300   Min.   : 1.180  
 1st Qu.:0.6350   1st Qu.: 0.9975   1st Qu.:  2.165   1st Qu.: 4.758  
 Median :0.7600   Median : 2.5700   Median :  6.165   Median : 6.420  
 Mean   :0.7302   Mean   : 4.2459   Mean   : 14.980   Mean   : 6.616  
 3rd Qu.:0.8425   3rd Qu.: 5.9825   3rd Qu.: 17.938   3rd Qu.: 8.193  
 Max.   :0.9500   Max.   :30.3600   Max.   :108.430   Max.   :17.060  
    Homicide          Water           Internet           CPI        
 Min.   : 0.000   Min.   :  7.07   Min.   :  2.66   Min.   : 14.75  
 1st Qu.: 1.250   1st Qu.: 70.21   1st Qu.: 33.77   1st Qu.:110.57  
 Median : 2.520   Median : 91.90   Median : 61.42   Median :119.16  
 Mean   : 7.390   Mean   : 80.99   Mean   : 56.80   Mean   :134.60  
 3rd Qu.: 7.595   3rd Qu.: 98.00   3rd Qu.: 80.12   3rd Qu.:145.63  
 Max.   :61.710   Max.   :100.00   Max.   :100.00   Max.   :459.03  
```

Fig. 1. Summary Statistics

B. Data Pre-processing

The data accumulated in our study is in a non-standardized form, meaning the range and mean for the dependent and independent variables are different for different variables. Also, the units of measurement are different in each cases. Hence, to standardize the data, Z-score standardization is used using the *scale()* function in R.

$$Z = \frac{x - \mu}{\sigma}$$

where, μ is the mean and σ is the standard deviation. Fig. 2 represents the summary after standardizing the variables.

```
> data[1:8] <- scale(data[1:8])
> summary(data)
      HDI              Emission            GDP               Health       
 Min.   :-2.4692   Min.   :-0.8808   Min.   :-0.7482   Min.   :-2.12571  
 1st Qu.:-0.6525   1st Qu.:-0.6819   1st Qu.:-0.6532   1st Qu.:-0.72664  
 Median : 0.2044   Median :-0.3518   Median :-0.4493   Median :-0.07648  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000  
 3rd Qu.: 0.7699   3rd Qu.: 0.3646   3rd Qu.: 0.1507   3rd Qu.: 0.61670  
 Max.   : 1.5069   Max.   : 5.4822   Max.   : 4.7631   Max.   : 4.08456  
    Homicide            Water             Internet            CPI         
 Min.   :-0.65351   Min.   :-3.3170   Min.   :-2.0141   Min.   :-2.6157  
 1st Qu.:-0.54298   1st Qu.:-0.4839   1st Qu.:-0.8570   1st Qu.:-0.5244  
 Median :-0.43067   Median : 0.4895   Median : 0.1717   Median :-0.3370  
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.01809   3rd Qu.: 0.7632   3rd Qu.: 0.8672   3rd Qu.: 0.2407  
 Max.   : 4.80328   Max.   : 0.8529   Max.   : 1.6068   Max.   : 7.0805  
```

Fig. 2. Summary of Standardized data

III. Model Building and Diagnostics

Before proceeding with the regression analysis, we shall check the correlations of the dependent variable (HDI) with our selected independent variables. We use the cor() function in R to get the Pearson's correlation for our data. On investigating the matrices in Fig. 3 and 4, we find that there is a strong positive correlation between HDI and CO_2 emissions, per capita GDP, proportion of safely managed drinking water resources and percentage of individuals using the internet. A weak positive correlation is detected between HDI and government health spending on health and a weak negative correlation is detected between HDI and homicide rates and CPI for the country. Also, we find a positive correlation between the independent variables CO_2 and per capita GDP and Internet usage, between per capita GDP and Internet usage and between internet usage and proportion of safely managed drinking water.

```
> round(cor(data),digits = 2)  # rounded to 2 decimals
           HDI Emission   GDP Health Homicide Water Internet   CPI
HDI       1.00     0.63  0.72   0.34    -0.22  0.68     0.91 -0.33
Emission  0.63     1.00  0.62   0.08    -0.23  0.46     0.68 -0.21
GDP       0.72     0.62  1.00   0.39    -0.24  0.48     0.71 -0.30
Health    0.34     0.08  0.39   1.00    -0.14  0.27     0.30 -0.13
Homicide -0.22    -0.23 -0.24  -0.14     1.00 -0.07    -0.18 -0.02
Water     0.68     0.46  0.48   0.27    -0.07  1.00     0.62 -0.27
Internet  0.91     0.68  0.71   0.30    -0.18  0.62     1.00 -0.32
CPI      -0.33    -0.21 -0.30  -0.13    -0.02 -0.27    -0.32  1.00
```

Fig. 3. Pearson correlation matrix

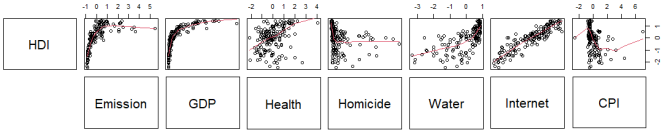

Fig. 4. Scatter plot

A. Multiple Linear regression

The main objective of simple regression is to position a line which represents the relationship between two variables using the least square principle i.e to estimate the value of dependent variable (y), given the value of independent variable(x). On the other hand, *Multiple linear regression* incorporates more than one independent variables which can be expressed as-

$$\hat{y} = a + b_1x_1 + b_2x_2 + +b_3x_3 + ... + b_kx_k + \epsilon$$

where, a is the y-intercept, b_j are the co-efficient of independent variables and ϵ are the error/residuals [6].

On running the multiple linear regression to generalize the population, it is essential that we follow assumptions laid by the Gauss Markov theorem which states that the ordinary least square estimator (OLS) is the best linear unbiased (BLU) estimator. Regressing all the independent variables against dependent variable (HDI), we get the results and summary as shown in fig.5.

At first, on examining the F-statistic, we get a significant p-value of $2.2x10^{-16}$, rejecting the null hypothesis (which states that all regression co-efficient are equal to 0). Hence, the data provides sufficient evidence that the regression model fits data better rather than with no predictor variables.

Secondly, on interpreting the t-statistic for individual predictors at 95% confidence interval, we notice that parameters like CO_2 emissions, Homicide rates, Govt. Health Expenditure and CPI fail to reject the null hypothesis (b_k=0). Hence, the inclusion of these independent variables do not significantly contribute to our model.

Lastly, we can clearly notice that the errors are not randomly distributed in the Residuals vs Fitted plot. A *funnel-in* pattern is visible which states that there are non-linear associations in the data. The same is evident on running Breusch-Pagan test (acquired by running the *bptest()* function in R, present under the *lmtest* library), we get a p-value of 0.0002564, which essentially states that heteroscedasticity is present (we reject the null hypothesis i.e presence of homoscedasticity).

```
> m0 <- lm(HDI~., data=data)
> summary(m0)

Call:
lm(formula = HDI ~ ., data = data)

Residuals:
      Min       1Q   Median       3Q      Max
-0.95991 -0.25846  0.06017  0.20988  1.45378

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.993e-16  2.951e-02   0.000   1.0000
Emission    -4.217e-02  4.424e-02  -0.953   0.3421
GDP          1.203e-01  4.637e-02   2.594   0.0104 *
Health       1.296e-02  3.381e-02   0.383   0.7020
Homicide    -5.952e-02  3.100e-02  -1.920   0.0567 .
Water        1.750e-01  3.837e-02   4.562 1.04e-05 ***
Internet     7.278e-01  5.151e-02  14.130  < 2e-16 ***
CPI         -1.886e-02  3.171e-02  -0.595   0.5528
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3733 on 152 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8606
F-statistic: 141.3 on 7 and 152 DF,  p-value: < 2.2e-16
```

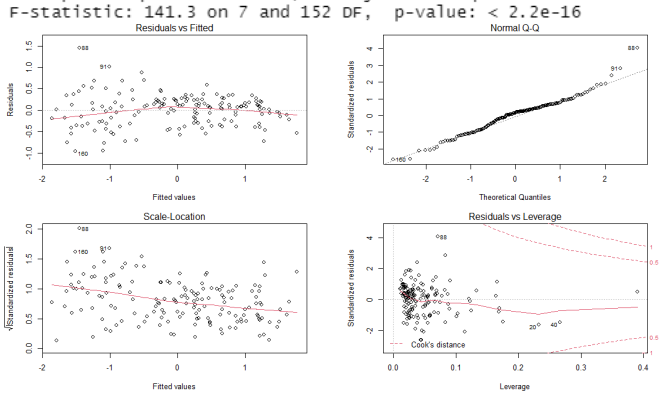

Fig. 5. Model-1

To address these issues, we run a modified multiple linear regression model, by removing the independent variables CO_2 Emissions, homicide rates, govt. health expenditure and CPI. We introduce interacting terms along with their square terms and run a step function in R which tries to lower the value of AIC by removing the insignificant terms. AIC refers to the Akaike information criterion which is a log-likelihood value expressed as-

$$AIC = -2 * logL + k * npar$$

where, *npar* is the number of parameters and *K* is log of number of observations [7].

In the second model Fig.6, on observing the Residuals vs Fitted plot, there is no evidence of any pattern. For the Breusch-Pagan test, we get a p-value of 0.0555 which is better than Model 1, concluding that we addressed the problem of heteroscedasticity. On checking the t-statistic, we notice that the term (GDP^2) has a very low significance at 5% confidence interval. Also, we have acquired a better adjusted R-squared value (co-efficient of multiple determination) of 0.9158 which states that 91.58% variation in HDI (dependent variable) is explained by the predictors (independent variables) in the current model.

In the final model (Fig.7), we removed the GDP^2 term. The term *Internet* cannot be eliminated due to the hierarchical

```
> model <- lm(HDI~GDP*Water*Internet+I(GDP^2)+I(Water^2)+I(Internet^2), data=data)
> modelstep <- step(model)
Start:  AIC=-384.63
HDI ~ GDP * Water * Internet + I(GDP^2) + I(Water^2) + I(Internet^2)

                     Df Sum of Sq    RSS     AIC
- I(water^2)          1   0.02269 12.623 -386.34
<none>                             12.601 -384.63
- I(GDP^2)            1   0.19084 12.791 -384.22
- I(Internet^2)       1   0.49960 13.100 -380.41
- GDP:Water:Internet  1   1.67753 14.278 -366.63

Step:  AIC=-386.34
HDI ~ GDP + Water + Internet + I(GDP^2) + I(Internet^2) + GDP:Water +
    GDP:Internet + Water:Internet + GDP:Water:Internet

                     Df Sum of Sq    RSS     AIC
<none>                             12.623 -386.34
- I(GDP^2)            1   0.20816 12.832 -385.72
- I(Internet^2)       1   0.66817 13.292 -380.09
- GDP:Water:Internet  1   1.67301 14.296 -368.43
> summary(modelstep)

Call:
lm(formula = HDI ~ GDP + Water + Internet + I(GDP^2) + I(Internet^2) +
    GDP:Water + GDP:Internet + Water:Internet + GDP:Water:Internet,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83460 -0.14429  0.02722  0.16131  0.88602

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.74591    0.09008   8.281 6.26e-14 ***
GDP                 1.42975    0.16535   8.647 7.48e-15 ***
Water              -0.36151    0.11805  -3.062  0.00260 **
Internet           -0.10506    0.10509  -1.000  0.31908
I(GDP^2)           -0.04630    0.02944  -1.573  0.11789
I(Internet^2)      -0.15551    0.05519  -2.818  0.00549 **
GDP:Water          -0.90317    0.21013  -4.298 3.09e-05 ***
GDP:Internet       -0.85636    0.17814  -4.807 3.68e-06 ***
Water:Internet      0.66012    0.12904   5.116 9.42e-07 ***
GDP:Water:Internet  0.74570    0.16725   4.459 1.61e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2901 on 150 degrees of freedom
Multiple R-squared:  0.9206,    Adjusted R-squared:  0.9158
F-statistic: 193.3 on 9 and 150 DF,  p-value: < 2.2e-16
```

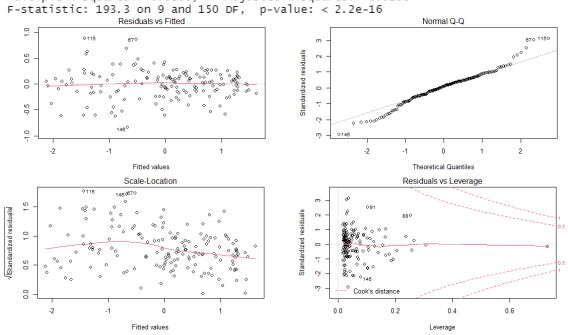

Fig. 6. Model-2

principle which states that, if an interaction term is included in a model the root term should also be included even if the t-test statistic implies that the term is insignificant.

B. Diagnostics

In order to validate the selected model or to validate our ordinary least square estimator is BLUE (best linear unbiased estimator), we check if they meet the Gauss-Markov assumptions.

1) Linearity (Correct functional form): Linearity states that the parameters have a linear relationship to the dependent variable. To validate this claim, we look at the Residuals vs Fitted plot in the Fig.7 and notice that the distribution is white noise or a random scatter plot. Therefore, the multiple linear regression according to the model can be written as below:

$$HDI = \beta_0 + \beta_1.GDP + \beta_2.Water + \beta_3.Internet + \beta_4.Internet^2 +$$

$$\beta_5.(GDP \times Water) + \beta_6.(GDP \times Internet) + \beta_7.(Water \times Internet) +$$

$$+ \beta_8.(GDP \times Water \times Internet)$$

where, β_0 is 0.73309, β_1 is 1.43192, β_2 is -0.34386, β_3 is -0.11830, β_4 is -0.10908, β_5 is -0.88912, β_6 is -0.98998, β_7 is 0.67203 and β_8 is 0.74830.

```
> modelstep <- lm(HDI ~ GDP + Water + Internet + I(Internet^2) +
+                   GDP:Water + GDP:Internet + Water:Internet +
+                   GDP:Water:Internet, data = data)
> summary(modelstep)

Call:
lm(formula = HDI ~ GDP + Water + Internet + I(Internet^2) + GDP:Water +
    GDP:Internet + Water:Internet + GDP:Water:Internet, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81617 -0.14343  0.01401  0.17526  0.89797

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.73309    0.09015   8.132 1.42e-13 ***
GDP                 1.43192    0.16615   8.618 8.51e-15 ***
Water              -0.34386    0.11809  -2.912  0.00414 **
Internet           -0.11830    0.10527  -1.124  0.26288
I(Internet^2)      -0.10908    0.04686  -2.328  0.02124 *
GDP:Water          -0.88912    0.21097  -4.215 4.29e-05 ***
GDP:Internet       -0.98998    0.15734  -6.292 3.22e-09 ***
Water:Internet      0.67203    0.12944   5.192 6.63e-07 ***
GDP:Water:Internet  0.74830    0.16805   4.453 1.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2915 on 151 degrees of freedom
Multiple R-squared:  0.9193,    Adjusted R-squared:  0.915
F-statistic:   215 on 8 and 151 DF,  p-value: < 2.2e-16
```

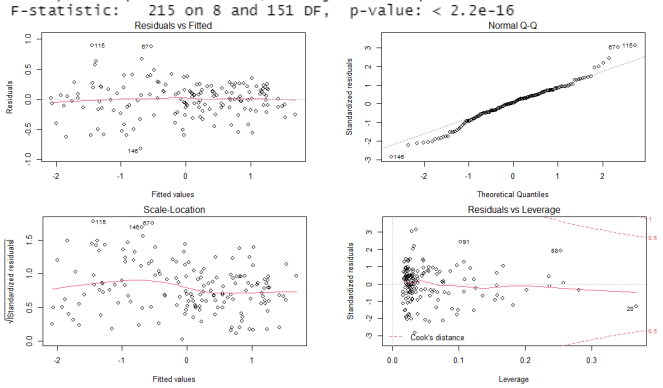

Fig. 7. Model-3

2) Homoscedasticity: To check if errors have a constant variance, we perform Breusch-Pagan test and get a p-value of 0.057. Therefore, we accept the null hypothesis that the error terms have a constant variance.

3) Autocorrelation: To check for auto-correlation meaning the error terms are not correlated, we use the Durbin-Watson test. We get a statistic value of 1.63 which is under an acceptable range and suggests minor positive auto-correlation.

4) Normality of Errors: In Fig.7, the Normal Q-Q plot suggests that the probability plot of standardized residuals fall on the 45-degree line which suffice the normality assumption. Also, checking histogram coupled with the density probability plot of the residuals suggests a normal distribution.

5) Multi-collinearity: VIF is calculated by regressing a predictor against every other predictors in the model. It is given by -

$$VIF = \frac{1}{1 - R_i^2}$$

where, R^2 is the co-efficient of determination. On running the variance inflation factor(VIF) in R, we get high values (greater than 10). Since we have included the square terms and interactions between predictors, the values as observed in fig.9, can be safely ignored and assumed that there are no multi-collinearities between predictors [8].

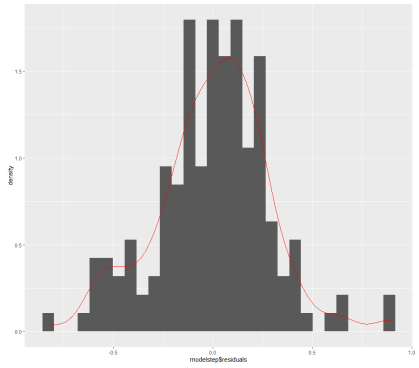

Fig. 8. Model-3

```
> vif(modelstep)
              GDP             Water          Internet      I(Internet^2)
        51.653085         26.093400         20.733665          3.830571
       GDP:Water      GDP:Internet   Water:Internet GDP:Water:Internet
        49.553845         52.631971         28.360656         71.407443
```

Fig. 9. VIF values of Predictors

6) Influential data points: To test if any outliers in a given set of predictors has a high influence or validate and outlier, we use the Cook's distance. On using *max(cooks.distance(modelstep))* in R, we get a maximum value of 0.139 which is well under the threshold value of 1. The same is evident in fig.7 by investigating the Residuals vs Leverage scatter plot.

IV. Conclusion

The goal of this study was to use multiple linear regression to create a model capable of estimating a country's Human Development Index (HDI) given certain factors such as per capita GDP, expenditure on health as a percentage of GDP, Homicide rates per 1,00,000, proportion of safely managed drinking water sources as a proportion of population, percentage of individuals using internet, Consumer price index (CPI) and Carbon-dioxide emissions per capita metric tons were known. Notably, in the process of data building we had to remove many observations due to unavailability of data and restricting our research to the year 2017. Using the regression modelling iterating the process, we used the step() function to eliminate factors which were not statistically significant in estimating the HDI of a country. We were able to trim down the traditional definition of HDI with the help of only three factors namely, per capita GDP of a country, proportion of safely managed drinking water and percentage of individuals using the internet using the regression model as evident in model 3 fig.7. We achieved an adjusted R^2 of 0.915 which states 91.5 percent of the variation in HDI is accounted by the three factors. Lastly, we were able to meet all the assumptions suggested by Gauss Markov theorem.

Thus, based on the regression analysis, in order to improve their HDI and subsequently their world HDI ranking, countries can alternatively focus on the improving above three aforementioned factors. The suggested model can therefore be used in estimating the HDI of a country.

REFERENCES

[1] "Human development index (hdi) — human development reports," 2020. [Online]. Available: http://hdr.undp.org/en/content/human-development-index-hdi

[2] W. Contributors, "Human development index," Wikipedia, 11 2020. [Online]. Available: https://en.wikipedia.org/wiki/Human_Development_Index

[3] "Per capita gdp definition," Investopedia, 2020. [Online]. Available: https://www.investopedia.com/terms/p/per-capita-gdp.asp

[4] W. H. O. WHO, "2.1 billion people lack safe drinking water at home, more than twice as many lack safe sanitation," Who.int, 07 2017. [Online]. Available: https://www.who.int/water_sanitation_health/publications/jmp-2017/en/

[5] S. B. Green, "How many subjects does it take to do a regression analysis," *Multivariate Behavioral Research*, vol. 26, no. 3, pp. 499–510, 1991, pMID: 26776715. [Online]. Available: https://doi.org/10.1207/s15327906mbr2603_7

[6] D. Lind, W. Marchal, and S. Wathen, *Basic Statistics for Business & Economics*, ser. Irwin-McGraw-Hill Series: Operations and Decision Sciences. McGraw-Hill Higher Education, 2008. [Online]. Available: https://books.google.ie/books?id=kYbeGgAACAAJ

[7] "R: Akaike's an information criterion," Ethz.ch, 2020. [Online]. Available: https://stat.ethz.ch/R-manual/R-patched/library/stats/html/AIC.html

[8] "When can you safely ignore multicollinearity? — statistical horizons," Statisticalhorizons.com, 2012. [Online]. Available: https://statisticalhorizons.com/multicollinearity