Study on Literacy Rates
SDGB 7840

Sumi Choudhury

04/06/17

i. **Executive Summary**:

In this research, we will model the literacy rate across countries in order to better understand which socioeconomic factors may be related to the literacy rate. Our data comes from the World Bank website which provides a repository of indicators grouped into categories such as agriculture, education, climate change, health, infrastructure, and economy. Based on some further reading and understanding of the literacy rate, we will select the variables which we will include in our model. We will use a multiple linear regression technique to model the literacy rate (using the statistical software R), provide the accompanying data and statistics used in our model, and finally draw our conclusions.

ii. **Introduction**:

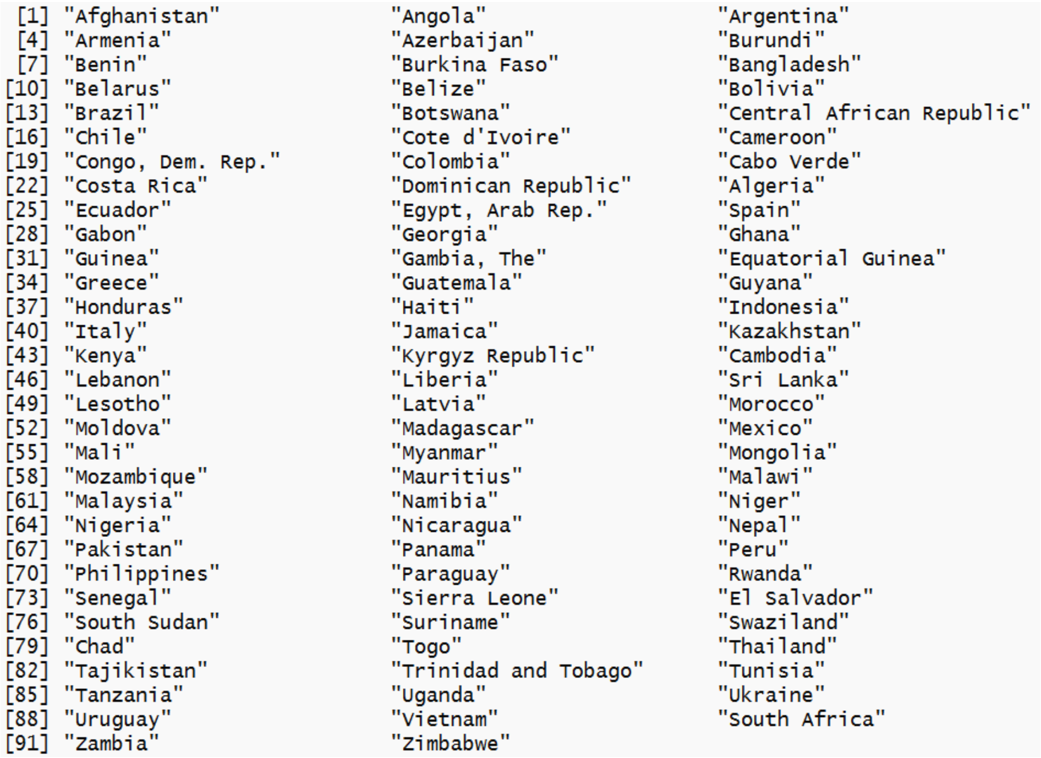
This study will focus on the literacy rate of adults aged 15 years and older who can read and write. The global literacy rate for all people aged 15 and above is 86.3%. The global literacy rate for all males is 90.0% and the rate for all females is 82.7%. As it is expected, the literacy rate for developed countries exceeds the literacy rate for poorer and underdeveloped countries. Ninety-five percent of the world's illiterate people live in developing countries. Of all illiterate adults, women represent about two-thirds of the population. As of 2015, over 75% of the world's 781 million illiterate adults lived in South Asia, West Asia and sub-Saharan Africa. The disparity in the literacy rate across countries is a reflection of the varying social, political, and economic factors in each country.

Why should we study the literacy rate? The United Nations Educational, Scientific and Cultural Organization (UNESCO) defines literacy as a human right and also as an instrument for achieving other rights. It is a direct benefit realized from the right to an education. There are also immense benefits derived from literacy as it is fundamental to informed decision-making, personal empowerment, and active participation in the local and global community.

A country that can effictively increase its literacy rate can further empower itself and foster the well-being of its society as a whole. It also increases the opportunities for global citizens to effectively interact with other nations and thus has the potential to increase understanding and awareness of other cultures and societies. Therefore, it is worthwhile to identify the factors that impact the literacy rate.

iii. Data:

After performing data cleaning and including only the records that have complete cases, we are left with the following ninety-two countries which are included in the final data set. All the records are based on year 2015 data:



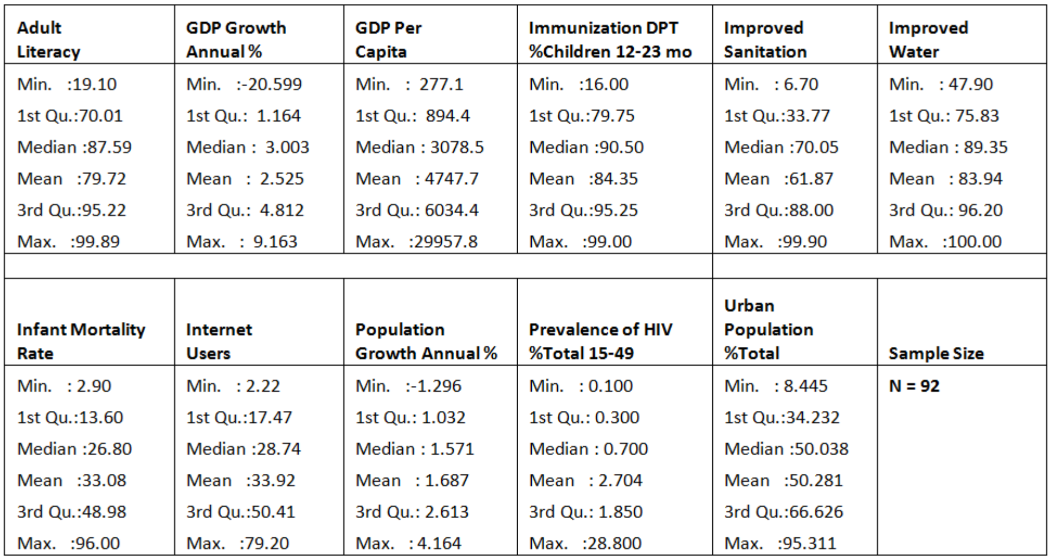
```
 [1] "Afghanistan"              "Angola"                   "Argentina"
 [4] "Armenia"                  "Azerbaijan"               "Burundi"
 [7] "Benin"                    "Burkina Faso"             "Bangladesh"
[10] "Belarus"                  "Belize"                   "Bolivia"
[13] "Brazil"                   "Botswana"                 "Central African Republic"
[16] "Chile"                    "Cote d'Ivoire"            "Cameroon"
[19] "Congo, Dem. Rep."         "Colombia"                 "Cabo Verde"
[22] "Costa Rica"               "Dominican Republic"       "Algeria"
[25] "Ecuador"                  "Egypt, Arab Rep."         "Spain"
[28] "Gabon"                    "Georgia"                  "Ghana"
[31] "Guinea"                   "Gambia, The"              "Equatorial Guinea"
[34] "Greece"                   "Guatemala"                "Guyana"
[37] "Honduras"                 "Haiti"                    "Indonesia"
[40] "Italy"                    "Jamaica"                  "Kazakhstan"
[43] "Kenya"                    "Kyrgyz Republic"          "Cambodia"
[46] "Lebanon"                  "Liberia"                  "Sri Lanka"
[49] "Lesotho"                  "Latvia"                   "Morocco"
[52] "Moldova"                  "Madagascar"               "Mexico"
[55] "Mali"                     "Myanmar"                  "Mongolia"
[58] "Mozambique"               "Mauritius"                "Malawi"
[61] "Malaysia"                 "Namibia"                  "Niger"
[64] "Nigeria"                  "Nicaragua"                "Nepal"
[67] "Pakistan"                 "Panama"                   "Peru"
[70] "Philippines"              "Paraguay"                 "Rwanda"
[73] "Senegal"                  "Sierra Leone"             "El Salvador"
[76] "South Sudan"              "Suriname"                 "Swaziland"
[79] "Chad"                     "Togo"                     "Thailand"
[82] "Tajikistan"               "Trinidad and Tobago"      "Tunisia"
[85] "Tanzania"                 "Uganda"                   "Ukraine"
[88] "Uruguay"                  "Vietnam"                  "South Africa"
[91] "Zambia"                   "Zimbabwe"
```

We consider ten primary variables to include in this research. All of the data is obtained from the World Bank and is easily accessible on http://data.worldbank.org/indicator. The following is an explanation of the selected variables:

- Adult literacy rate, population 15+ years, both sexes (%) - Percentage of the population ages 15 and older who can, with understanding, both read and write a short simple statement on their everyday life. This is our response variable - the variable which we are trying to predict.
- GDP growth (annual %) - Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2000 U.S. dollars. GDP growth can indicate that a country is productive and experiencing financial growth.

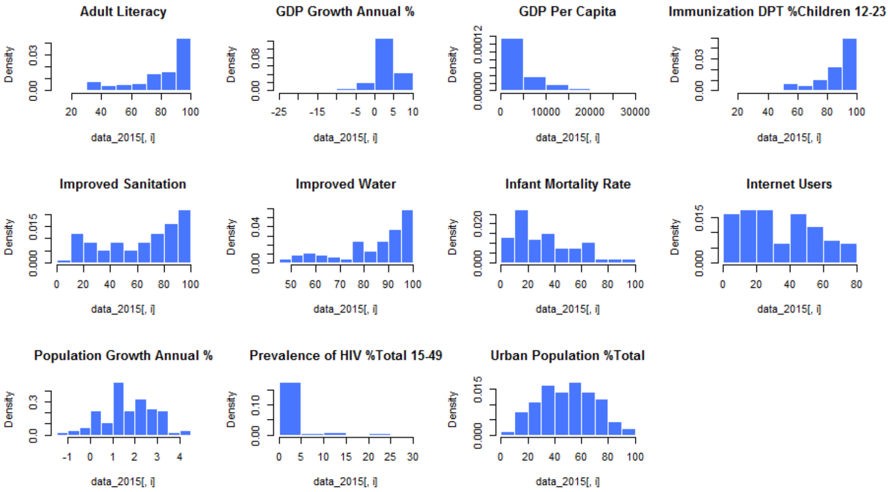
- GDP Per Capita (current US$) - GDP is in current U.S. dollars per person. Data are derived by converting GDP in national currency to U.S. dollars and then dividing it by total population. This tells us how much of the GDP can be rationed per individual.
- Immunization, DPT (% of children ages 12-23 months) - Child immunization measures the percentage of children ages 12-23 months who received vaccinations before 12 months or at any time before the survey. Immunization regulations can help citizens stay healthy and attend school.
- Improved sanitation facilities (% of population with access) - Access to improved sanitation facilities refers to the percentage of the population using improved sanitation facilities. Improved sanitation facilities are important for the health of the individual and society.
- Improved water source (% of population with access) - Access to an improved water source refers to the percentage of the population using an improved drinking water source. Improved water source is important for the health and well-being of individuals.
- Mortality rate, infant (per 1,000 live births) - The infant mortality rate is defined as the number of deaths of children under one year of age, expressed per 1000 live births. A high infant mortality rate indicates poor health and living standards.
- Internet users (per 100 people) - Internet users are individuals who have used the Internet (from any location) in the last 12 months. Access to the internet can help individuals in learning.
- Population growth (annual %) - Annual population growth rate for year t is the exponential rate of growth of midyear population from year t-1 to t, expressed as a percentage. A high population growth rate can result in limited resources for underdeveloped nations.
- Prevalence of HIV, total (% of population ages 15-49) - Prevalence of HIV refers to the percentage of people ages 15-49 who are infected with HIV. Prevalence of HIV is an important indicator of the health and living standards of a nation.
- Urban population (% of total) - Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects. Urban population as a percent of the total can indicate a greater proportion of the population in developed regions, thus can impact the literacy rate.

Summary Statistics:



Adult Literacy	GDP Growth Annual %	GDP Per Capita	Immunization DPT %Children 12-23 mo	Improved Sanitation	Improved Water
Min. :19.10	Min. :-20.599	Min. : 277.1	Min. :16.00	Min. : 6.70	Min. : 47.90
1st Qu.:70.01	1st Qu.: 1.164	1st Qu.: 894.4	1st Qu.:79.75	1st Qu.:33.77	1st Qu.: 75.83
Median :87.59	Median : 3.003	Median : 3078.5	Median :90.50	Median :70.05	Median : 89.35
Mean :79.72	Mean : 2.525	Mean : 4747.7	Mean :84.35	Mean :61.87	Mean : 83.94
3rd Qu.:95.22	3rd Qu.: 4.812	3rd Qu.: 6034.4	3rd Qu.:95.25	3rd Qu.:88.00	3rd Qu.: 96.20
Max. :99.89	Max. : 9.163	Max. :29957.8	Max. :99.00	Max. :99.90	Max. :100.00
Infant Mortality Rate	**Internet Users**	**Population Growth Annual %**	**Prevalence of HIV %Total 15-49**	**Urban Population %Total**	**Sample Size**
Min. : 2.90	Min. : 2.22	Min. :-1.296	Min. : 0.100	Min. : 8.445	**N = 92**
1st Qu.:13.60	1st Qu.:17.47	1st Qu.: 1.032	1st Qu.: 0.300	1st Qu.:34.232	
Median :26.80	Median :28.74	Median : 1.571	Median : 0.700	Median :50.038	
Mean :33.08	Mean :33.92	Mean : 1.687	Mean : 2.704	Mean :50.281	
3rd Qu.:48.98	3rd Qu.:50.41	3rd Qu.: 2.613	3rd Qu.: 1.850	3rd Qu.:66.626	
Max. :96.00	Max. :79.20	Max. : 4.164	Max. :28.800	Max. :95.311	

Distribution Graphs (Histograms):

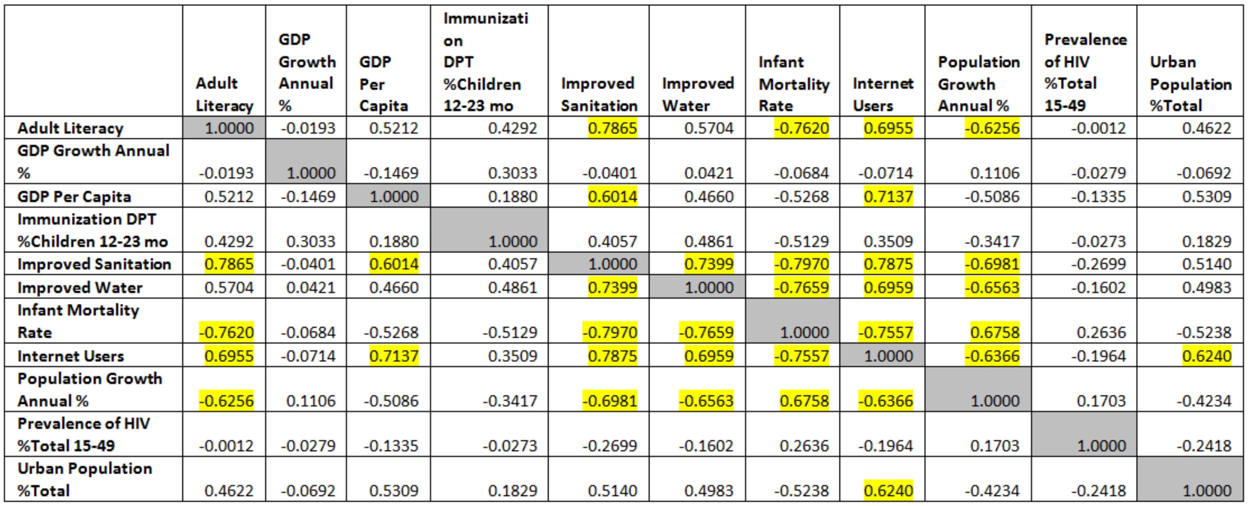


We can see that for the most part, the response and predictor variables are not normally distributed. To obtain a better distribution, we will perform natural log transformations to each of the variables used in our modeling.

iv. <u>Methods</u>:

The first step that we should take in building our model is perform variable selection. In order to do this, we need to examine the correlation matrices which tell us the strength of the linear relationship between the variables. We can also examine scatterplots to help determine the direction and strength of the correlations.

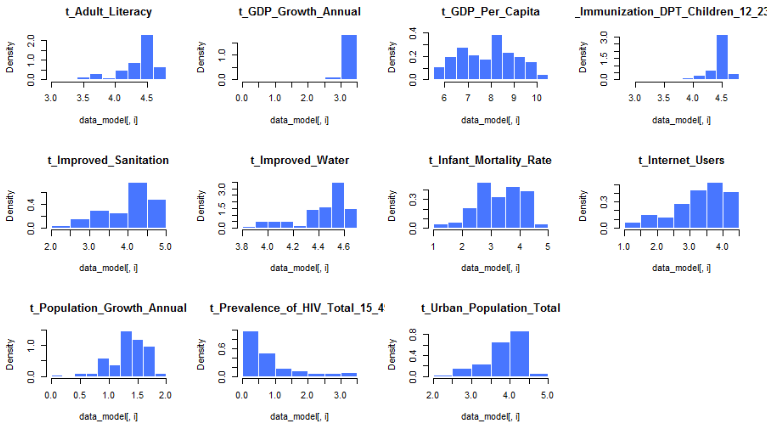
Below is a chart outlining the correlation coefficients between variables. Moderate to strong correlations are highlighted in the chart:



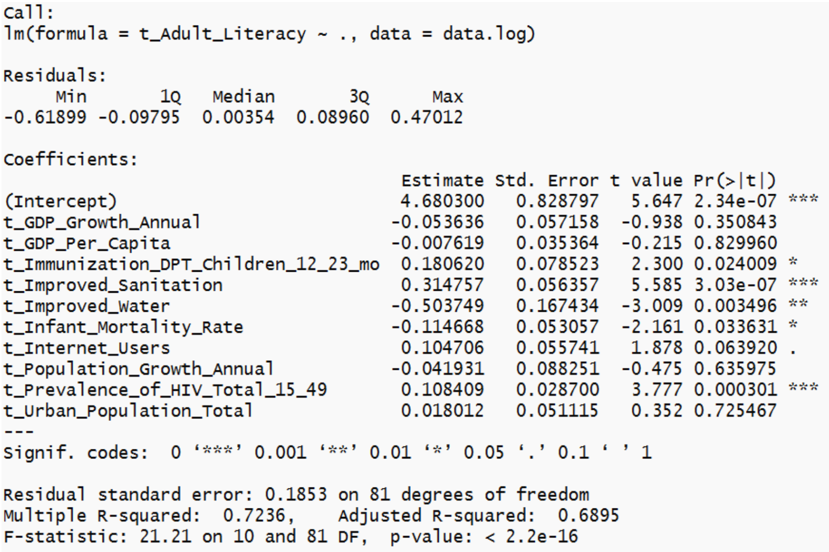
	Adult Literacy	GDP Growth Annual %	GDP Per Capita	Immunization DPT %Children 12-23 mo	Improved Sanitation	Improved Water	Infant Mortality Rate	Internet Users	Population Growth Annual %	Prevalence of HIV %Total 15-49	Urban Population %Total
Adult Literacy	1.0000	-0.0193	0.5212	0.4292	0.7865	0.5704	-0.7620	0.6955	-0.6256	-0.0012	0.4622
GDP Growth Annual %	-0.0193	1.0000	-0.1469	0.3033	-0.0401	0.0421	-0.0684	-0.0714	0.1106	-0.0279	-0.0692
GDP Per Capita	0.5212	-0.1469	1.0000	0.1880	0.6014	0.4660	-0.5268	0.7137	-0.5086	-0.1335	0.5309
Immunization DPT %Children 12-23 mo	0.4292	0.3033	0.1880	1.0000	0.4057	0.4861	-0.5129	0.3509	-0.3417	-0.0273	0.1829
Improved Sanitation	0.7865	-0.0401	0.6014	0.4057	1.0000	0.7399	-0.7970	0.7875	-0.6981	-0.2699	0.5140
Improved Water	0.5704	0.0421	0.4660	0.4861	0.7399	1.0000	-0.7659	0.6959	-0.6563	-0.1602	0.4983
Infant Mortality Rate	-0.7620	-0.0684	-0.5268	-0.5129	-0.7970	-0.7659	1.0000	-0.7557	0.6758	0.2636	-0.5238
Internet Users	0.6955	-0.0714	0.7137	0.3509	0.7875	0.6959	-0.7557	1.0000	-0.6366	-0.1964	0.6240
Population Growth Annual %	-0.6256	0.1106	-0.5086	-0.3417	-0.6981	-0.6563	0.6758	-0.6366	1.0000	0.1703	-0.4234
Prevalence of HIV %Total 15-49	-0.0012	-0.0279	-0.1335	-0.0273	-0.2699	-0.1602	0.2636	-0.1964	0.1703	1.0000	-0.2418
Urban Population %Total	0.4622	-0.0692	0.5309	0.1829	0.5140	0.4983	-0.5238	0.6240	-0.4234	-0.2418	1.0000

We can see that there are moderate to strong postive correlations between the response variable (Adult Literacy) and each of the explanatory variables - Improved Sanitation and Internet Users. However, Internet Users and Improved Sanitation are also highly correlated, so we need to be careful about multicollinearity. Population Growth Annual % and Infant Mortality Rate each have a moderate to strong negative linear relationship with Adult Literacy. In addition, we need to note that the two explanatory variables - Infant Mortality Rate and Improved Water have a strong negative correlation with each other.

Prior to building our model, we will transform the variables using natural log. Below are the histogram distributions of the variables after transformation:



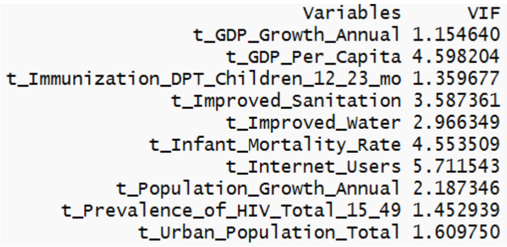
The distributions here are slightly better than in the previous histogram chart. We will continue building our model and check all of our assumptions for the multiple regression linear model. Our population model has the form $y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdot \cdot \cdot + \beta k\ xk + \varepsilon$, and our sample (estimated) model takes the form $\hat{y} = \hat{\beta}0 + \hat{\beta}1x + \hat{\beta}2x2 + \cdot \cdot \cdot + \hat{\beta}\ k\ xk$. We will first build a model which includes all of our predictor variables. Here are the results:



```
Call:
lm(formula = t_Adult_Literacy ~ ., data = data.log)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.61899 -0.09795  0.00354  0.08960  0.47012 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        4.680300   0.828797   5.647 2.34e-07 ***
t_GDP_Growth_Annual               -0.053636   0.057158  -0.938 0.350843    
t_GDP_Per_Capita                  -0.007619   0.035364  -0.215 0.829960    
t_Immunization_DPT_Children_12_23_mo  0.180620   0.078523   2.300 0.024009 *  
t_Improved_Sanitation              0.314757   0.056357   5.585 3.03e-07 ***
t_Improved_Water                  -0.503749   0.167434  -3.009 0.003496 ** 
t_Infant_Mortality_Rate           -0.114668   0.053057  -2.161 0.033631 *  
t_Internet_Users                   0.104706   0.055741   1.878 0.063920 .  
t_Population_Growth_Annual        -0.041931   0.088251  -0.475 0.635975    
t_Prevalence_of_HIV_Total_15_49    0.108409   0.028700   3.777 0.000301 ***
t_Urban_Population_Total           0.018012   0.051115   0.352 0.725467    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1853 on 81 degrees of freedom
Multiple R-squared:  0.7236,	Adjusted R-squared:  0.6895 
F-statistic: 21.21 on 10 and 81 DF,  p-value: < 2.2e-16
```

We should note that not all of the variables included in this model are significant. More importantly, we can see that Improved Water now has a negative coefficient, whereas before we saw that it has a positive relationship with Adult Literacy. Prevalence of HIV has also changed its sign from negative to positive. This is an indication that multicollinearity is present in our data. The next step is to use the vif() function in R to check the variance inflation factor, as this will indicate the level of inflation that we will see in our estimated coefficients for the predictor variables. This is directly related to the high correlations between variables in the correlaton matrix shown earlier.

Upon calculating the VIFs, we obtain the following results:



```
                           Variables      VIF
                 t_GDP_Growth_Annual 1.154640
                    t_GDP_Per_Capita 4.598204
t_Immunization_DPT_Children_12_23_mo 1.359677
               t_Improved_Sanitation 3.587361
                    t_Improved_Water 2.966349
             t_Infant_Mortality_Rate 4.553509
                    t_Internet_Users 5.711543
          t_Population_Growth_Annual 2.187346
     t_Prevalence_of_HIV_Total_15_49 1.452939
            t_Urban_Population_Total 1.609750
```

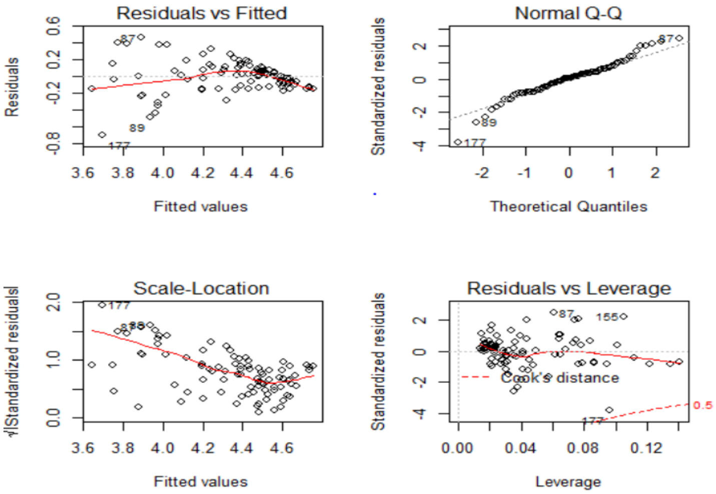
All VIFs are under 10 but a few are closer to 5, which means that multicollinearity is moderate. As we have seen that Improved Water has changed its sign, we should actually remove Improved Water from our model. We should also remove insignificant variables and Prevalence of HIV. Doing so yields the following results:

```
Call:
lm(formula = t_Adult_Literacy ~ t_Immunization_DPT_Children_12_23_mo + 
    t_Improved_Sanitation + t_Infant_Mortality_Rate, data = data.log)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.79386 -0.09150  0.01600  0.09805  0.42714 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           2.95555    0.47418   6.233 1.54e-08 ***
t_Immunization_DPT_Children_12_23_mo  0.08996    0.08141   1.105   0.2722    
t_Improved_Sanitation                 0.31287    0.04944   6.329 1.01e-08 ***
t_Infant_Mortality_Rate              -0.07695    0.04166  -1.847   0.0681 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2086 on 88 degrees of freedom
Multiple R-squared:  0.6198,	Adjusted R-squared:  0.6068 
F-statistic: 47.82 on 3 and 88 DF,  p-value: < 2.2e-16
```

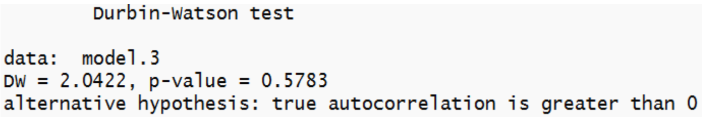
When we remove those variables, we see that Immunization DPT (% children ages 12-23 months) is no longer significant. So we can remove this variable as well. Doing so yields the following results:

```
Call:
lm(formula = t_Adult_Literacy ~ t_Improved_Sanitation + t_Infant_Mortality_Rate, 
    data = data.log)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.79868 -0.09073  0.01294  0.10486  0.45303 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              3.35402    0.30832  10.878  < 2e-16 ***
t_Improved_Sanitation    0.31902    0.04919   6.486 4.81e-09 ***
t_Infant_Mortality_Rate -0.08482    0.04110  -2.064    0.042 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2088 on 89 degrees of freedom
Multiple R-squared:  0.6145,	Adjusted R-squared:  0.6059 
F-statistic: 70.94 on 2 and 89 DF,  p-value: < 2.2e-16
```

Now that we have selected our variables, we need to check our assumptions:



The plot reveals that there is no discernable trend in the residuals (exhibits linearity). The residuals do not quite exhibit constant variance (lacks homoskedasticity). The normal Q-Q Plot shows heavy tails (residuals are not quite normal). The Cook's Distance Plot shows us that all cases are well inside of the Cook's distance lines, meaning that there are no influential cases. In order to check if the errors are independent of one another, we can conduct a Durbin-Watson test.



```
        Durbin-Watson test

data:  model.3
DW = 2.0422, p-value = 0.5783
alternative hypothesis: true autocorrelation is greater than 0
```

The result of the Durbin-Watson test indicates that since p-value of the d-statistic is insignificant, we will fail to reject the null hypothesis that rho = 0; there is zero autocrorrelation of the residuals. Linear regression is robust to violations of the constant variance and normality assumptions, so we can accept this model. The p-value for the overall F-statistic in the final model is also less than $\alpha = 0.05$, so we can say that our overall model is significant.

v. <u>Results</u>:

Our final regression model is $\hat{y}_{\text{Adult Literacy Rate}}$ = 3.35402 % + 0.31902 % / % $x_{\text{improved sanitation}}$ – 0.08482 % / % $x_{\text{infant mortality rate}}$.

This can be interpreted as: holding all other variables constant, a 1% increase in Improved Sanitation % of Population with Access is associated with a 0.32% increase in the adult literacy rate and a 1% increase in the Infant Mortality Rate (per 1,000 live births) is associated with a 0.085% decrease in the adult literacy rate.

We also split the data into train and test data sets at 50/50 split. Upon evaluating the results from training and testing, a random sampling of the training data used in the model shows an R-squared of 0.6626 which means that 66.26% of the variability in the response variable can be explained by variability in the predictor variables. The predictions made on a random sampling of the testing data set shows an R-squared of 0.5279 or 52.79%.

Improvements to this model could have been made had there been more pertinent data available which could influence the adult literacy rate, such as 2015 data on enrollment and government spending on education.

vi. **References**:

1) The World Bank:
http://data.worldbank.org/indicator

2) Wikipedia:
https://en.wikipedia.org/wiki/List_of_countries_by_literacy_rate

3) UNESCO:
http://www.unesco.org/education/GMR2006/full/chapt5_eng.pdf

4) Index Mundi:
https://www.indexmundi.com/facts/indicators