HW3-Matthew Fein-545

Matthew Fein

April 22, 2019

# Global Literacy Rate Analysis

## Executive Summary

The World Bank maintains an extensive database of information from a wide range of categories (i.e economics, education, etc.) collected from individual countries all over the world. Using this database exploratory data analysis (EDA), feature selection and modelling were carried out to analyze what measures may have an impact on literacy rates. The current model failed to yield any meaningful predictors, limitations and proposed improvements are discussed.

require(tidyverse)  
require(gridExtra)  
require(broom)  
require(usdm)

## Introduction

It has been been well documented for over a half century that literacy rates within a country can have profound effects on many things from economic factors to employment markets and far beyond (Blaug, 1966). On an intuitive level this makes sense- depending upon the ability for a nation’s inhabitants to read and write limits the efficiency and productivity of that country. Since literacy rates can have such a strong impact beyond education, having a better sense of what specific indicators are related to literacy rates would be useful. Fortunately, the World Bank maintains an extensive database which may contain this information.The world Bank (<https://data.worldbank.org/indicator/?tab=all>) database has compiled numerous statistics from countries all over the world under various categories such as climate change, economics & growth, education and more. Ten specific independent variables (IVs) have been chosen from the World Bank database with the goal of modeling which, if any, seem to be good predictors of a nation’s literacy rate. This report will go on to describe the data collected, how it was used and what takeaways there are.

## Data

The ten attributes selected for this analysis were: Percent with access to electricity, life expectancy at birth, GDP per capita,Percent of GDP spent on military, workforce size, unemployment rate, percent of population in urban areas, pupil to teacher ratio and percentage of trained teachers. All of these were considered as predictor variables for the dependent measure of literacy rate. In order to predict literacy rate it needed to be determined what countries and years would be used for the analysis. Only individual countries were used rather than regions- this was to avoid any countries potentially being counted more than once. There were only 15 countries who reported literacy rates for any 3-5 year period and 36 for 2 years in a row. In order to maximize datapoints the year 2011 was chosen for all factors due to the fact that this was the year most responses were recorded (96). Of these, the following 56 countries remained:

countries <- read\_csv("countries.csv")

## Parsed with column specification:  
## cols(  
## `Country Name` = col\_character()  
## )

table(countries)

## Methods

All individual predictors were examined for missing values. Those which failed the eyeball test (Jane et al., 2014) for number of missing values were dropped.

data.df <- data.frame(countries)  
literacy <- read\_csv("literacy.csv") #import all data  
assistance <- read\_csv("assrec.csv")  
GDP <- read\_csv("GDPpc.csv")  
workforce <- read\_csv("laborers.csv")  
life <- read\_csv("lifeex.csv")  
MilpcGDP <- read\_csv("military.csv")  
StudRat <- read\_csv("ptratio.csv")  
trained <- read\_csv("trained.csv")  
unemp <- read\_csv("unemployment.csv")  
urbanpct <- read.csv("urban.csv")  
access <- read\_csv("elic.csv")  
# Filter each IV by countries included and count NAs  
lit <- filter(literacy,literacy$`Country Name` %in% data.df$Country.Name)  
lit <- lit[ , "2011"]  
sum(is.na(lit))  
assist <- filter(assistance, assistance$`Country Name` %in% data.df$Country.Name)  
assist <- assist[ ,"2011"]  
sum(is.na(assist))  
GDPpc <- filter(GDP,GDP$`Country Name` %in% data.df$Country.Name)  
GDPpc <-GDPpc[ , "2011"]  
sum(is.na(GDPpc))  
workers <- filter(workforce,workforce$`Country Name` %in% data.df$Country.Name)  
workers <-workers[ , "2011"]  
sum(is.na(workers))  
lifeex <- filter(life,life$`Country Name` %in% data.df$Country.Name)  
lifeex <-lifeex[ , "2011"]  
sum(is.na(lifeex))  
Milpct <- filter(MilpcGDP,MilpcGDP$`Country Name` %in% data.df$Country.Name)  
Milpct <-Milpct[ , "2011"]  
sum(is.na(Milpct))  
ptratio <- filter(StudRat,StudRat$`Country Name` %in% data.df$Country.Name)  
ptratio <-ptratio[ , "2011"]  
sum(is.na(ptratio))  
trnteach <- filter(trained,trained$`Country Name` %in% data.df$Country.Name)  
trnteach <-trnteach[ , "2011"]  
sum(is.na(trnteach))  
unemploy <- filter(unemp,unemp$`Country Name` %in% data.df$Country.Name)  
unemploy <-unemploy[ , "2011"]  
sum(is.na(unemploy))  
pctcity <- filter(urbanpct,urbanpct$Country.Name %in% data.df$Country.Name)  
pctcity <-pctcity[ ,55]  
sum(is.na(pctcity))  
electric <- filter(access,access$`Country Name` %in% data.df$Country.Name)  
electric <-electric[ , "2011"]  
sum(is.na(electric))

This then left the remaining Ivs for inclusion in the model predicting literacy:

variables <- cbind(c(lit, GDPpc, workers, lifeex, unemploy, pctcity, electric ))  
print(variables)

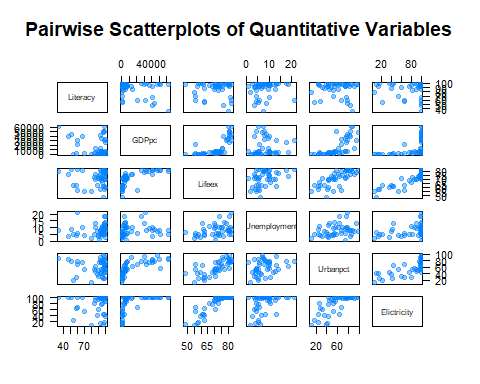
## [,1]   
## 2011 Numeric,56  
## 2011 Numeric,56  
## 2011 Numeric,56  
## 2011 Numeric,56  
## 2011 Numeric,56  
## 23.737   
## 52.163   
## 90.849   
## 63.44   
## 53.406   
## 30.462   
## 72.302   
## 45.558   
## 66.43   
## 84.335   
## 74.961   
## 87.074   
## 63.256   
## 77.964   
## 71.736   
## 67.551   
## 73.753   
## 62.69   
## 78.442   
## 68.094   
## 76.292   
## 51.885   
## 55.155   
## 49.914   
## 30.93   
## 68.327   
## 86.088   
## 30.064   
## 66.757   
## 67.841   
## 100   
## 58.018   
## 77.815   
## 73.568   
## 35.999   
## 94.072   
## 64.14   
## 41.555   
## 41.616   
## 16.768   
## 34.997   
## 60.567   
## 53.829   
## 43.773   
## 100   
## 65.452   
## 54.993   
## 37.533   
## 23.389   
## 66.657   
## 70.825   
## 94.414   
## 88.083   
## 20.078   
## 62.218   
## 33.196   
## 2011 Numeric,56

Following this, further examination was done to remedy any remaining data points. If any country wasmissing more than one value they would be excluded from the analysis. This resulted in the removal of 12 countries and the IV of workforce.

data.good <- data.frame(read\_csv("features.csv"))  
  
df <- filter(data.good, data.good$Country\_Name %in% data.df$Country.Name)  
  
NAs <- is.na.data.frame(df)  
print(NAs)  
  
df <- df[-c(1,4,9,10,21,29,31,46,47,49,52,53), -4]

Following this any issues of collinearity were searched for and remedied.

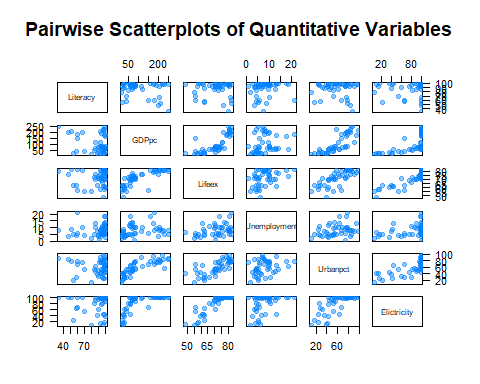
pairs(df[, c("Literacy", "GDPpc", "Lifeex", "Unemployment", "Urbanpct", "Elictricity")],   
 main="Pairwise Scatterplots of Quantitative Variables", las=TRUE, col="#0080ff70", pch=19)



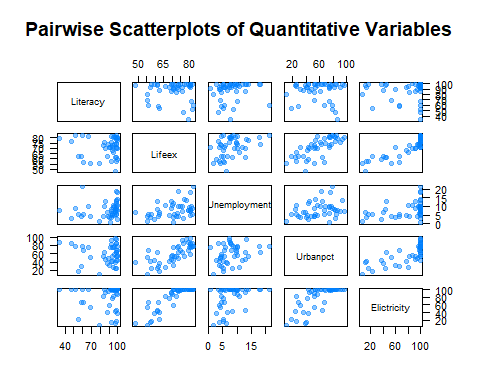
df$GDPpc <- sqrt(df$GDPpc)

GDP per capita showed a parabolic trend with numerous other IVs so a squareroot transformation was conducted. The transformation revealed multicollinearity with urban percentage and access to elictricity and was dropped.

pairs(df[, c("Literacy", "GDPpc", "Lifeex", "Unemployment", "Urbanpct", "Elictricity")],   
 main="Pairwise Scatterplots of Quantitative Variables", las=TRUE, col="#0080ff70", pch=19)

 Following this the data looked as follows:

df <- df[ ,-3]  
  
pairs(df[, c("Literacy","Lifeex", "Unemployment", "Urbanpct", "Elictricity")],   
 main="Pairwise Scatterplots of Quantitative Variables", las=TRUE, col="#0080ff70", pch=19)



round(cor(df[, c("Literacy", "Lifeex", "Unemployment", "Urbanpct")]), digits = 3)

## Literacy Lifeex Unemployment Urbanpct  
## Literacy 1.000 -0.020 0.125 -0.061  
## Lifeex -0.020 1.000 0.392 0.730  
## Unemployment 0.125 0.392 1.000 0.292  
## Urbanpct -0.061 0.730 0.292 1.000

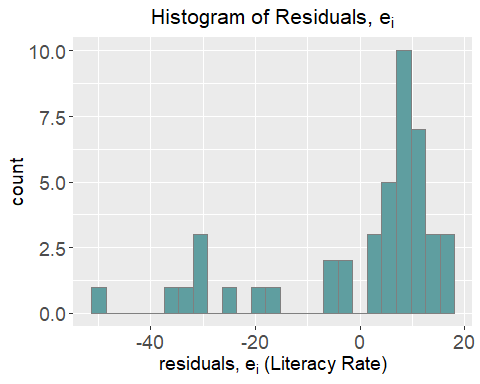
While there was still collinearity between unemployment and urban percentage this was ignored in order to examine the effects of multiple IVs. A multiple regression analysis was concluded which resulted in:

lm.literacy <- lm(Literacy~ .-Country\_Name, data = df)  
summary(lm.literacy)

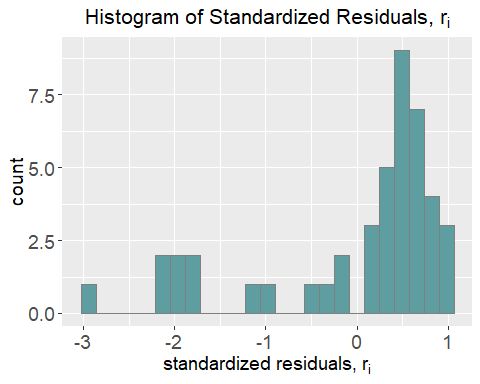
##   
## Call:  
## lm(formula = Literacy ~ . - Country\_Name, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -49.766 -3.582 7.834 10.588 16.781   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 97.90775 32.61000 3.002 0.00466 \*\*  
## Lifeex -0.24591 0.64823 -0.379 0.70648   
## Unemployment 0.56700 0.66179 0.857 0.39681   
## Urbanpct -0.10367 0.18399 -0.563 0.57633   
## Elictricity 0.09766 0.19538 0.500 0.61999   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.68 on 39 degrees of freedom  
## Multiple R-squared: 0.03231, Adjusted R-squared: -0.06694   
## F-statistic: 0.3256 on 4 and 39 DF, p-value: 0.8591

The model achieved poor results with an R-squared = 0.03. Residual plots are:

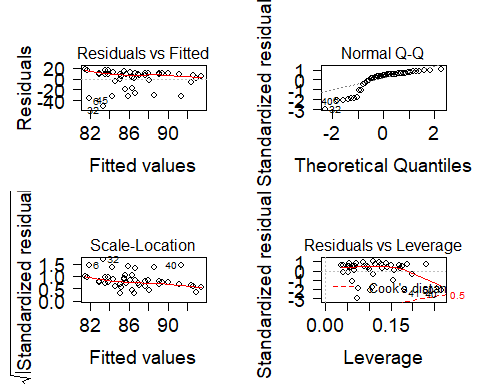
theme.info <- theme(plot.title = element\_text(size=16, hjust=0.5),  
 axis.title = element\_text(size=14),  
 axis.text = element\_text(size=14))  
  
lm.augment <- augment(lm.literacy)  
  
h1 <- lm.augment %>%   
 ggplot(aes(x=.resid)) +  
 geom\_histogram(bins=25, col="gray50", fill="cadetblue") +  
 labs(x=expression(paste("residuals, ", e[i], " (Literacy Rate)", sep=""))) +  
 ggtitle(expression(paste("Histogram of Residuals, ", e[i], sep=""))) +  
 theme.info  
  
h2 <- lm.augment %>%   
 ggplot(aes(x=.std.resid)) +  
 geom\_histogram(bins=25, col="gray50", fill="cadetblue") +  
 labs(x=expression(paste("standardized residuals, ", r[i], sep=""))) +  
 ggtitle(expression(paste("Histogram of Standardized Residuals, ", r[i], sep=""))) +  
 theme.info  
  
h1



h2

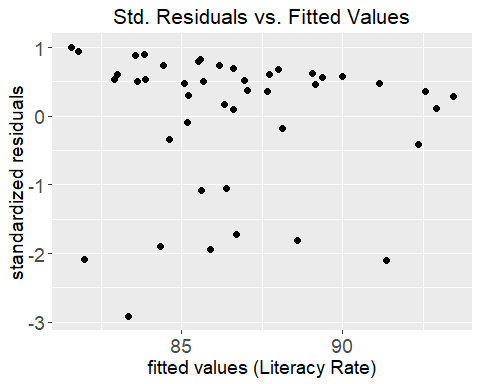
 Even with standarzadization we still have a negatively skewed distribution. Various transformation methods did not resolve this.More model diagnostics follow.

par(mfrow=c(2,2), mar=c(5, 5, 4, 2)+0.1)  
plot(lm.literacy, las=TRUE, cex.main=1.4, cex.axis=1.4, cex.lab=1.4)



full\_join(lm.augment, df) %>%   
 ggplot(aes(x=.fitted, y=.std.resid)) +  
 geom\_point(size=2) +  
 labs(x="fitted values (Literacy Rate)", y="standardized residuals") +  
 ggtitle("Std. Residuals vs. Fitted Values") +  
 theme.info

## Joining, by = c("Literacy", "Country\_Name", "Lifeex", "Unemployment", "Urbanpct", "Elictricity")

 ##DiscussionFurther diagnostics reveal more issues. Constant variance was not upheld in the dataset, and the Q-Q plot reveals a pattern distinctive of a heavy tailed distribution. This makes sense considering the highly skewed residual distribution seen above. Scale-location reveals some possible outliers, however, removal did not yield any significant results in another regression analysis. It seems that a ceiling effect has made this model unsuccessful at capturing anything meaningful. By using such strict inclusion criteria there was very little difference among the literacy rates of the countries included. Furthermore, a sample size of 44 nations is not very well representative of the world. Expanding the sample size may reduce the multicolinearity amongst IVs as well as lead to a better model. Further examination is required.

## References

Blaug, M. (1966). Literacy and economic development. The School Review, 74(4), 393-418.Jain, R., Duval, S., & Adabag, S. (2014). How accurate is the eyeball test? A comparison of physician’s subjective assessment versus statistical methods in estimating mortality risk after cardiac surgery. Circulation: Cardiovascular Quality and Outcomes, 7(1), 151-156.The World Bank (<https://data.worldbank.org/indicator/?tab=all>)