In the pursuit of happiness

Contents

Introduction

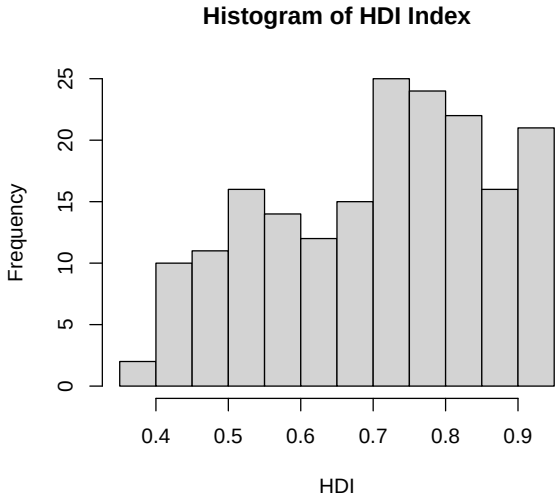
Our project today is to determine what socio-economic variables have statistically significant effects on a country's level of development as measured by the HDI. We intend to do this through different types of regression models, applied to a dataset of 241 countries in 2014.

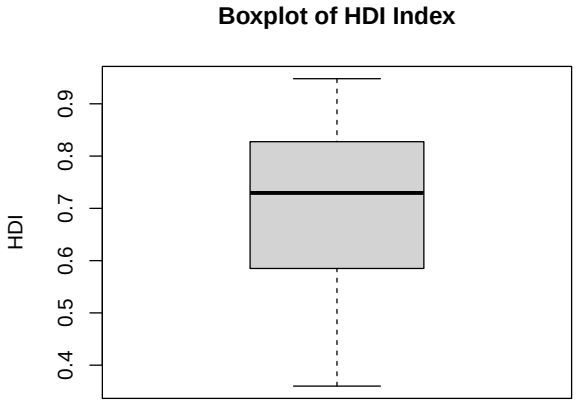
```
library(tidyverse)
library(GGally)
library(ggplot2)
library(caret)
library(MASS)
library(glmnet)
library(factoextra)
```

```
df=readRDS("./merged_gapminder_happiness.rds")%>%filter(!is.na(hdiindex))
df_variables=df%>%dplyr::select(-country,-year)
```

Exploratory Data Analysis

Target variable

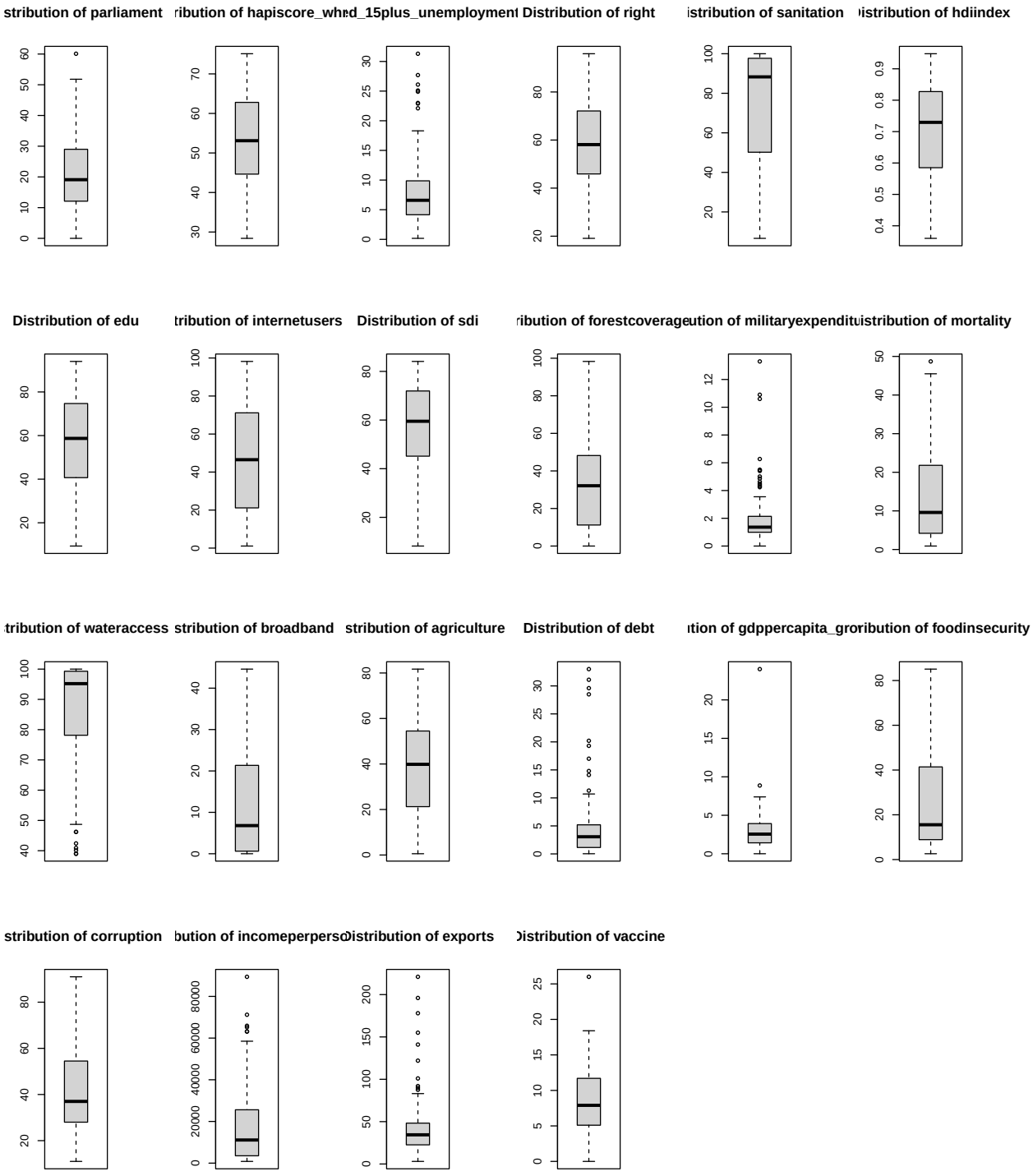





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3600  0.5850  0.7295  0.7054  0.8273  0.9480
```

hdiindex: Human Development Index (HDI)
distribution - concave down increasing, wide spread.
IQR: 0.02125, variance: 9.375379e-04, standard deviation: 3.061924e-02

Independent variables



hapiscore_whr: This is the national average response to the question of life evaluations asking the following "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" This measure is

also referred to as Cantril life ladder. Gapminder has converted this indicator's scale from 0 to 100.
Distribution - mild hump/stretched-out normal curve with wide spread
IQR: 9.52500, variance: 6.312879e+01, standard deviation: 7.945363e+00

parliament: Percentage of national parliamentary seats held by women. Lower and upper houses combined.
Distribution - skewed right, normal.

aged_15plus_unemployment_rate_percent: Percentage of total population, age group above 15, that has been registered as unemployed during the given year.
Distribution - normal
IQR: 11.77500, variance: 8.424851e+01, standard deviation: 9.178698e+00

right: Fundamental Rights in the form of liberal and social rights support both fair representation and the vertical mechanism of accountability that the first attribute seeks to achieve. This attribute is composed of three subattributes: access to justice, civil liberties, and social rights and equality. The three subattributes were aggregated into the Fundamental Rights index using BFA
Distribution: normal/hump, wide spread
IQR: 6.95000, variance: 7.135659e+01, standard deviation: 8.447283e+00

sanitation: percentage of people using improved sanitation facilities that are not shared with other households.
distribution - uniform/linear then with sharp jump/exponential increase at 100%.
IQR: 22.82500, variance: 1.784627e+02, standard deviation: 1.335899e+01

edu: Education index calculated based on Avg years of schooling, taking values 0 as minimum and 15 as maximum
Distribution - wide spread.
IQR: 12.12500, variance: 8.859174e+01, standard deviation: 9.412319e+00

internetusers: Percentage of Population with stable Internet access
Distribution - multimodal, wide spread.
IQR: 19.22500, variance: 2.372645e+02, standard deviation: 1.540339e+01

sdi: The Sustainable Development Index is an efficiency metric, designed to assess the ecological efficiency of nations in delivering human development. It is calculated as the quotient of two figures: (1) a "development index" based on the Human Development Index, calculated as the geometric mean of the life expectancy index, the education index, and a modified income index; and (2) an "ecological impact index" calculated as the extent to which consumption-based CO2 emissions and material footprint exceed per-capita shares of planetary boundaries.
Distribution - seemingly linearly increasing, or skewed left.
IQR: 5.17500, variance: 1.093839e+02, standard deviation: 1.045867e+01

forestcoverage: Percentage of total land area that has been covered with forest
Distribution - Seemingly decreasing.
IQR: 18.10500, variance: 3.058635e+02, standard deviation: 1.748895e+01

militaryexpenditure: Military expenditure as a percentage of GDP.
Distribution - Right skewed normal with outliers at larger values.
IQR: 0.78975, variance: 2.334034e+00, standard deviation: 1.527755e+00

mortality: Newborn mortality rate per 1000
Distribution - decreasing
IQR: 6.22500, variance: 1.591720e+01, standard deviation: 3.989636e+00

wateraccess: Percentage of people with at least basic water services
Distribution - exponentially increasing.
IQR: 7.25000, variance: 2.993909e+01, standard deviation: 5.471663e+00

broadband: Broadband subscribers per 100 people
Distribution - steep drop/potentially exponentially decreasing.
IQR: 9.98500, variance: 4.516839e+01, standard deviation: 6.720743e+00

agriculture: Percentage of land area that is arable
Distribution - wide-ranging
IQR: 13.02500, variance: 2.622336e+02, standard deviation: 1.619363e+01

debt: Total debt service (% of GNI)
Distribution - Sharp decrease then flat
IQR: 6.64000, variance: 3.385360e+01, standard deviation: 5.818385e+00

gdppercapita_growth: GDP per capita
Distribution - decreasing with outliers
IQR: 2.13000, variance: 2.543614e+00, standard deviation: 1.594871e+00

foodinsecurity: Prevalence of moderate or severe food insecurity in the population
Distribution - normal skewed right
IQR: 9.55000, variance: 1.128336e+02, standard deviation: 1.062232e+01

Exports:% of GNP: exports of goods & services, include the value of merchandise, freight, insurance, transport, travel
Data source: https://data.worldbank.org/indicator/NE.EXP.GNFS.ZS/ IQR: 20.17500, variance: 2.426063e+02, standard deviation: 1.557582e+01

Corruption Perception Index (CPI): Transparency International's score of perceptions of corruption. Higher value indicates less corruption.
Data source: http://www.transparency.org/research/cpi/ IQR: 3.25000, variance: 8.818182e+00, standard deviation: 2.969542e+00

Income: Income per person adjusted for purchasing power
Data Source: World Bank, http:llgapm.io/dgdppc/ IQR: 5350.00000, variance: 2.689152e+07, standard deviation: 5.185703e+03

Vaccine: Proportion of people who disagree that vaccines are effective for children to have
Data Source: http://gapm.io/dvaccine_confidence/ IQR: 5.47500, variance: 3.927970e+01, standard deviation: 6.267352e+00

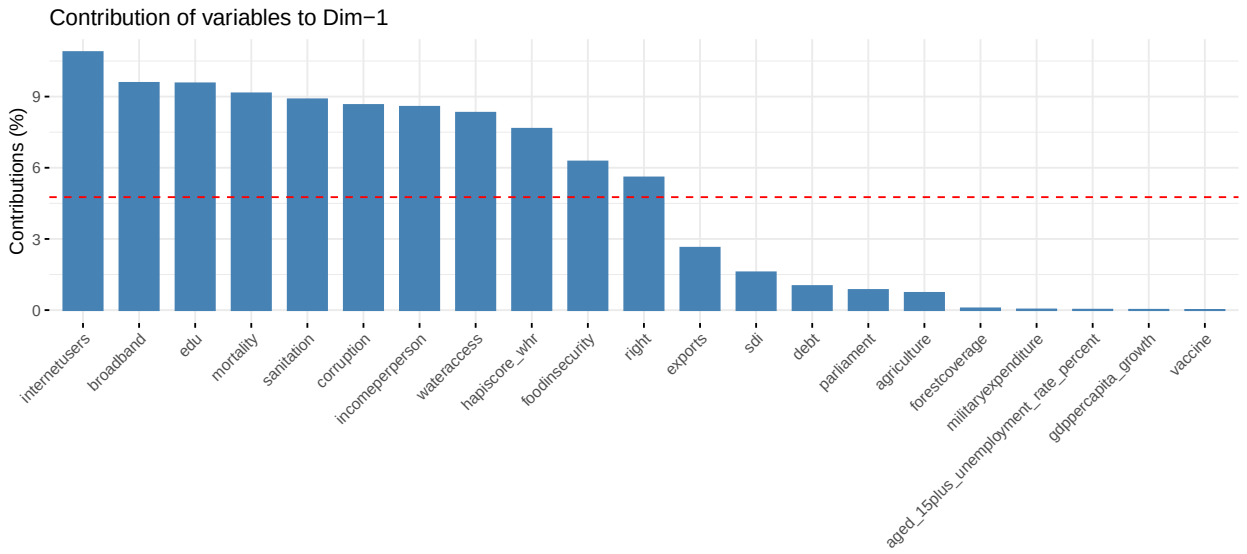
General observations

There are some variables with significant left skew i.e beyond a certain threshold, the variable will most likely not affect the target variable any more (e.g water, sanitation). There are some variables with significant right skew i.e below a certain threshold, the variable will most likely not affect the target variable any more (e.g debt). There are a lot of variables with substantial numbers of missing values,which we fill with mean imputation.

Regression

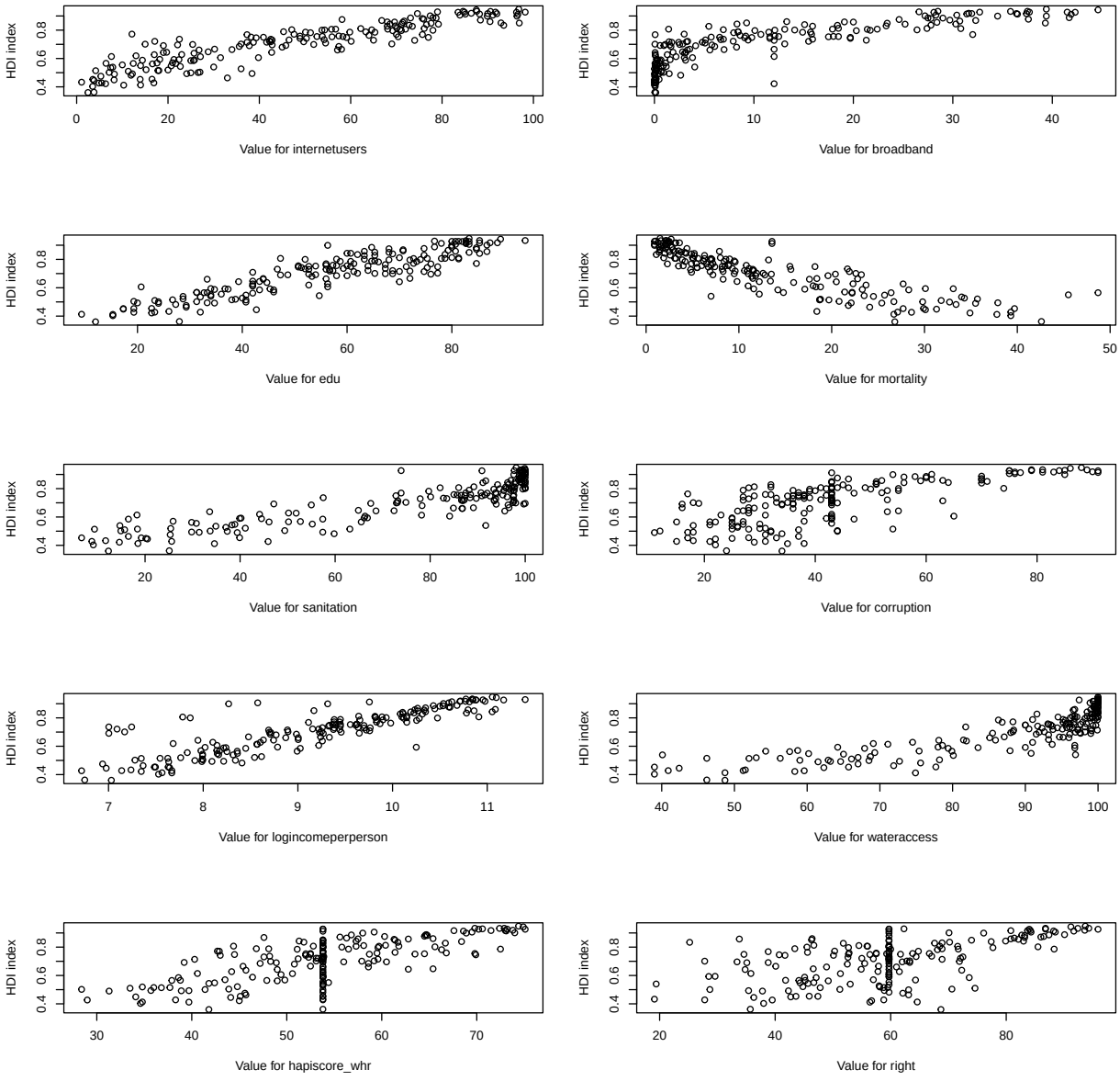
Dimension reduction via PCA To reduce the number of variables to apply to our regression model, we propose applying PCA to all feature variables, picking the first PC which accounts for the largest variation in the data, and choosing the variables with above average contribution.

```
pr_hdi=prcomp(df_variables%>%dplyr::select(-hdiindex),scale=T,center=T)
```

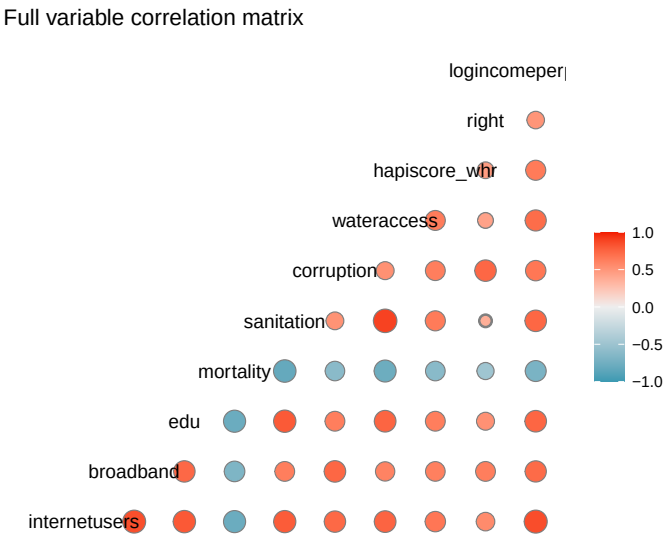


Based on the proposed plan, we choose to select the following variable: internetusers, broadband, edu, mortality, sanitation, corruption, incomeperperson,wateraccess, hapiscore_whr and right.

Linearity assumptions Before we apply the regression model, we check the linearity assumption for these variables, and see that most variables have a linear relationship with HDI, except for incomeperperson, which has a linear relationship once a log transformation has been applied to it.



Correlation matrix



Based on our correlation plot, most of our variables are strongly correlated with each other. As such, if we are to introduce them as they are right now, the model would suffer from multicollinearity. Consequently, we decide to conduct an backwards regression model in order to remove some of the variables based on the Aikaike Information Criterion.

Step backwards regression

To reduce the number of regression variables further to the most important factors affecting a country's development level, we use step backwards regression with an aim of getting the most 3-5 important predictors. We combine this with 10-fold cross-validation to get the most robust iteration.

```
train.control <- trainControl(method = "cv", number = 10)
res.lm <- lm(hdiindex ~., data = df_variables_2)
# Train the model
set.seed(123)
step.model <- train(hdiindex ~., data = df_variables_2,
                    method = "leapBackward",
                    trControl = train.control,
                    trace = FALSE,
                    tuneGrid = data.frame(nvmax = 3:5)
                    )
```

Our step backwards model indicates that a 5 regressor model has the lowest AIC through the corss-validation process.

```
step.model$bestTune
```

```
##   nvmax
## 3     5
```

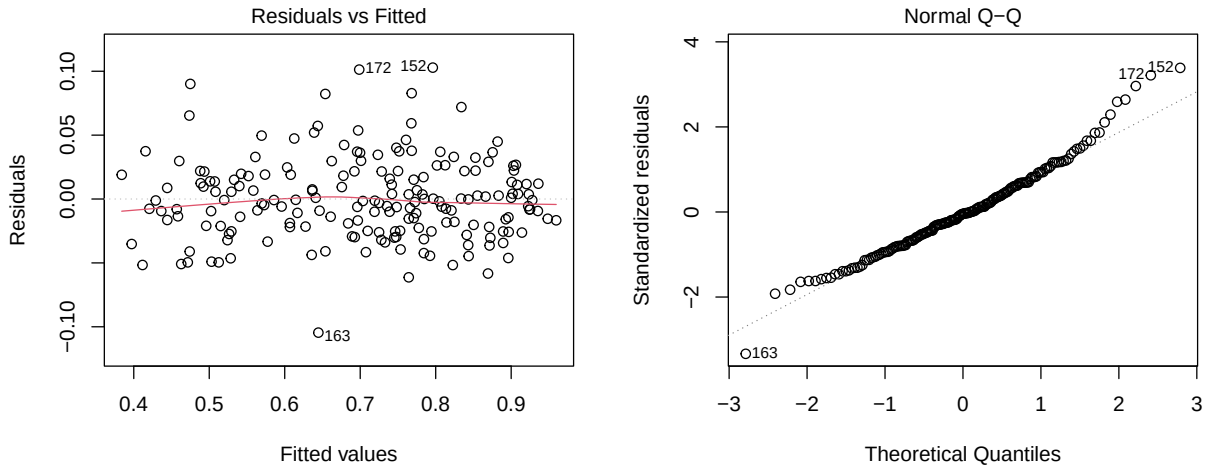
```
summary(step.model$finalModel)
```

```
## Subset selection object
## 10 Variables  (and intercept)
##                     Forced in Forced out
## internetusers           FALSE      FALSE
## broadband               FALSE      FALSE
## edu                     FALSE      FALSE
## mortality               FALSE      FALSE
## sanitation              FALSE      FALSE
## corruption              FALSE      FALSE
## wateraccess             FALSE      FALSE
## hapiscore_whr           FALSE      FALSE
## right                   FALSE      FALSE
## logincomeperperson      FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: backward
##          internetusers broadband edu mortality sanitation corruption
## 1  ( 1 ) " "           " "       "*" " "       " "        " "
## 2  ( 1 ) " "           " "       "*" " "       " "        " "
## 3  ( 1 ) " "           "*"       "*" " "       " "        " "
## 4  ( 1 ) " "           "*"       "*" " "       " "        " "
## 5  ( 1 ) " "           "*"       "*" "*"       " "        " "
##          wateraccess hapiscore_whr right logincomeperperson
## 1  ( 1 ) " "         " "           " "   " "
## 2  ( 1 ) "*"         " "           " "   " "
## 3  ( 1 ) "*"         " "           " "   " "
## 4  ( 1 ) "*"         " "           " "   "*"
## 5  ( 1 ) "*"         " "           " "   "*"
```

Our five variable models include broadband, edu, mortality, logincomeperperson, and wateraccess.

```
final_model_aic=lm(hdiindex~broadband+edu+mortality+logincomeperperson+wateraccess,data=df_variables_2)
```

Model diagnostics
The residual plots seem to exhibit homoskedasticity, and most of the residuals follow a normal distribution. This satisfies the linearity assumption.



We check the VIF of the model to see if the strong correlations between the chosen variables have introduced multicollinearity into the model.

```
##          broadband                edu          mortality logincomeperperson
##           2.636064           3.805438            3.467074           2.996607
##        wateraccess
##           3.125356
```

Based on the VIF values, the model does not have severe multicollinearity issues. Regardless, it may be better to try tree-based models that does not mind does not mind the multicollinear nature of our variables. To that end, we start with a random forest model.

Random Forest

We utilize random forest as a form of variable reduction, given that it allows for variables to be ranked in terms of their importance in dissecting the data.

```
set.seed(123)
df.RF.cv <- train(hdiindex~hapiscore_whr+parliament+aged_15plus_unemployment_rate_percent+right+sanitat
                  method="rf", importance=TRUE,
                  trControl =trainControl(method = "cv", number=10))
df.RF.cv$finalModel
```

```
##
## Call:
##  randomForest(x = x, y = y, mtry = min(param$mtry, ncol(x)), importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
```

```
## No. of variables tried at each split: 11
##
##           Mean of squared residuals: 0.0008516981
##                     % Var explained: 96.31
```

Important variables
Our random forest model indicates that sdi, edu, incomeperperson, corruption and broadband are the top 5 variables that affect HDI the most:

```
##                   Overall
## sdi             100.00000
## edu              84.55666
## incomeperperson  74.83959
## corruption       68.76989
## broadband        62.61783
```

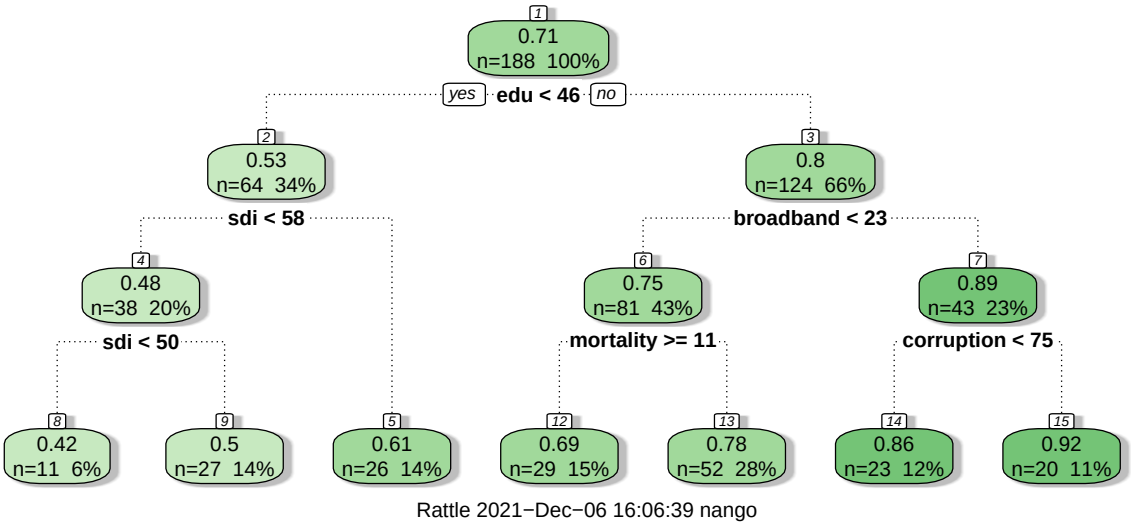
We also construct a single regression tree:

Regression Tree

```
set.seed(123)
df.cart <- train(hdiindex ~hapiscore_whr+parliament+aged_15plus_unemployment_rate_percent+right+sanitat
                 data=df_variables,
                 method="rpart1SE",
                 trControl =trainControl(method = "cv", number=10) )
df.cart
```

```
## CART
##
## 188 samples
##  21 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 169, 168, 170, 171, 169, 169, ...
## Resampling results:
##
##   RMSE        Rsquared   MAE
##   0.05015177  0.8928641  0.03960654
```

We draw the regression tree to get an indication of the top 5 variables by variable importance:



Important variables
Our tree model indicates that the top 5 variables that affect HDI are incomeperperson, internetusers, wateraccess, broadband and mortality.

```
##                   Overall
## incomeperperson 100.00000
## internetusers    97.95956
## wateraccess      72.26290
## broadband        69.60759
## mortality        64.20442
```

Our tree model indicates that the top 5 variables that affect HDI are incomeperperson, internetusers, wateraccess, broadband and mortality.

Model comparisons:

```
final_model_comps
```

```
##                   Model       RMSE R_squared
## 1 5 variable regression 0.03173857 0.9563353
## 2         Random Forest 0.01179182 0.9655732
## 3                  CART 0.03893235 0.8928641
```

Summary

In summary, our three models propose different variables that have significant effects on a country's level of development: 1. 5-variable regression: broadband, edu, mortality,incomeperperson, and wateraccess. 2. Random forest: sdi, edu, incomeperperson, corruption and broadband. 3. CART: incomeperperson, internetusers, wateraccess, broadband and mortality. In all 3 models, broadband and education are present. This makes sense: HDI is a proxy for standards of living, which tends to improve if an individual has more income to spend to improve said standards of living. Additionally, in the 21st century with the pace of

technological innovation, standards of living are increasingly tied to digital participation, making broadband important.