MSDS 6372 Project 1

Kaggle Dataset: “Life Expectancy (WHO): Statistical Analysis on factors influencing life expectancy”.

Kevin Boyd & Shikha Pandey

Introduction:

The World Health Organization (WHO) has kept track of health status as well as many other related factors for every country for many years. This data is stored in a repository called The Global Health Observatory and is made publicly available for the purpose of health data analysis. Our goal is to explore this data set and try to build and interoperate a model that helps to explain the key relationships we observe. Our main interest is to perform statistical analysis on the factors influencing life expectancy. This study is specifically helpful because it incorporates a period from 2000 to 2015 for all countries and includes not only demographic variables and income composition but also focuses on immunization factors, mortality rates, and other health related factors.

Data Description:

Information in this dataset related to life expectancy, health factors for 193 countries has been collected from the WHO data repository website and its corresponding economic data was collected from United Nation website for the period 2000-2015.

This data set lists data for countries from 2000-2015 for various immunizations - Hepatitis B, Polio and Diphtheria, diseases - Measles, HIV/AIDS, health factors - Alcohol, BMI, mortality factors – adult mortality, infant deaths, under-five deaths, economic factors – GDP, percentage expenditure, Total expenditure, Income composition of resources and social factors – population, schooling, status, and other factors – thinness 1-19 years, thinness 5-9 years.

Initial look at the data shows most missing values are for population, Hepatitis B, GDP. We have decided to move forward with keeping all the data to do our Exploratory Data Analysis.

Exploratory Data Analysis:

Chart, box and whisker chart

Description automatically generated

Life Expectancy Distribution for Developed and Developing Countries:

Chart, histogram

Description automatically generated

GDP:

Chart, box and whisker chart

Description automatically generated

Plot showing multiple variables we considered:

Calendar

Description automatically generated with medium confidence

Objective 1:

Our first objective is to build a regression model with the purpose of identifying key relationships and interpreting those relationships. After our EDA we have identified some variables that appear to be related to life expectancy, but we need to be careful that they are not highly correlated with each other because they could be trying to describe the same thing. One approach we took was to run models with all the explanatory variables and use the VIF to remove one variable at a time and re-run the model. Once we stopped seeing high correlation between explanatory variables, we ran a stepwise selection to narrow down our predictors even more. To select our model, we looked over the variables we chose in our stepwise selection to make sure they all made sense and were statistically significant. We looked at residual plots, Q-Q plots, and checked for influential points that could be outliers in this analysis. Parameter interpretation.............

Plot for the variables in the model we chose:

Plot showing histogram of residuals:

Plots showing linear regression model output (residual plots, QQ plot, Cooks D):

Plots showing AIC, BIC, ASE, R2, Adjusted R2:

Objective 2:

Our second objective is to build a model with additional complexity so that it can be used to better predict. This may come at the expense of the interpretation, but the main purpose of this model is to make the most accurate prediction of future data. Using the train and test sets we established for our first model, we made another linear regression model that included an interaction......... or something like that............

We also wanted to build another regression model, but this time we decided to use a non-parametric technique to build it. We ended up selecting a k-nearest neighbors’ regression model because...... or a regression trees model because..........

**Appendix**

# load and look at the data  
lifeExpectancy <- read.csv("Life Expectancy Data.csv", header = T)  
head(lifeExpectancy, n=5)

## Country Year Status Life.expectancy Adult.Mortality infant.deaths  
## 1 Afghanistan 2015 Developing 65.0 263 62  
## 2 Afghanistan 2014 Developing 59.9 271 64  
## 3 Afghanistan 2013 Developing 59.9 268 66  
## 4 Afghanistan 2012 Developing 59.5 272 69  
## 5 Afghanistan 2011 Developing 59.2 275 71

## Alcohol percentage.expenditure Hepatitis.B Measles BMI under.five.deaths  
## 1 0.01 71.279624 65 1154 19.1 83  
## 2 0.01 73.523582 62 492 18.6 86  
## 3 0.01 73.219243 64 430 18.1 89  
## 4 0.01 78.184215 67 2787 17.6 93  
## 5 0.01 7.097109 68 3013 17.2 97

## Polio Total.expenditure Diphtheria HIV.AIDS GDP Population  
## 1 6 8.16 65 0.1 584.25921 33736494  
## 2 58 8.18 62 0.1 612.69651 327582  
## 3 62 8.13 64 0.1 631.74498 31731688  
## 4 67 8.52 67 0.1 669.95900 3696958  
## 5 68 7.87 68 0.1 63.53723 2978599  
## thinness..1.19.years thinness.5.9.years Income.composition.of.resources  
## 1 17.2 17.3 0.479  
## 2 17.5 17.5 0.476  
## 3 17.7 17.7 0.470  
## 4 17.9 18.0 0.463  
## 5 18.2 18.2 0.454  
## Schooling  
## 1 10.1  
## 2 10.0  
## 3 9.9  
## 4 9.8  
## 5 9.5

summary(lifeExpectancy)

## Country Year Status Life.expectancy  
## Afghanistan : 16 Min. :2000 Developed : 512 Min. :36.30   
## Albania : 16 1st Qu.:2004 Developing:2426 1st Qu.:63.10   
## Algeria : 16 Median :2008 Median :72.10   
## Angola : 16 Mean :2008 Mean :69.22   
## Antigua and Barbuda: 16 3rd Qu.:2012 3rd Qu.:75.70   
## Argentina : 16 Max. :2015 Max. :89.00   
## (Other) :2842 NA's :10   
## Adult.Mortality infant.deaths Alcohol percentage.expenditure  
## Min. : 1.0 Min. : 0.0 Min. : 0.0100 Min. : 0.000   
## 1st Qu.: 74.0 1st Qu.: 0.0 1st Qu.: 0.8775 1st Qu.: 4.685   
## Median :144.0 Median : 3.0 Median : 3.7550 Median : 64.913   
## Mean :164.8 Mean : 30.3 Mean : 4.6029 Mean : 738.251   
## 3rd Qu.:228.0 3rd Qu.: 22.0 3rd Qu.: 7.7025 3rd Qu.: 441.534   
## Max. :723.0 Max. :1800.0 Max. :17.8700 Max. :19479.912   
## NA's :10 NA's :194   
## Hepatitis.B Measles BMI under.five.deaths  
## Min. : 1.00 Min. : 0.0 Min. : 1.00 Min. : 0.00   
## 1st Qu.:77.00 1st Qu.: 0.0 1st Qu.:19.30 1st Qu.: 0.00   
## Median :92.00 Median : 17.0 Median :43.50 Median : 4.00   
## Mean :80.94 Mean : 2419.6 Mean :38.32 Mean : 42.04   
## 3rd Qu.:97.00 3rd Qu.: 360.2 3rd Qu.:56.20 3rd Qu.: 28.00   
## Max. :99.00 Max. :212183.0 Max. :87.30 Max. :2500.00   
## NA's :553 NA's :34   
## Polio Total.expenditure Diphtheria HIV.AIDS   
## Min. : 3.00 Min. : 0.370 Min. : 2.00 Min. : 0.100   
## 1st Qu.:78.00 1st Qu.: 4.260 1st Qu.:78.00 1st Qu.: 0.100   
## Median :93.00 Median : 5.755 Median :93.00 Median : 0.100   
## Mean :82.55 Mean : 5.938 Mean :82.32 Mean : 1.742   
## 3rd Qu.:97.00 3rd Qu.: 7.492 3rd Qu.:97.00 3rd Qu.: 0.800   
## Max. :99.00 Max. :17.600 Max. :99.00 Max. :50.600   
## NA's :19 NA's :226 NA's :19   
## GDP Population thinness..1.19.years  
## Min. : 1.68 Min. :3.400e+01 Min. : 0.10   
## 1st Qu.: 463.94 1st Qu.:1.958e+05 1st Qu.: 1.60   
## Median : 1766.95 Median :1.387e+06 Median : 3.30   
## Mean : 7483.16 Mean :1.275e+07 Mean : 4.84   
## 3rd Qu.: 5910.81 3rd Qu.:7.420e+06 3rd Qu.: 7.20   
## Max. :119172.74 Max. :1.294e+09 Max. :27.70   
## NA's :448 NA's :652 NA's :34   
## thinness.5.9.years Income.composition.of.resources Schooling   
## Min. : 0.10 Min. :0.0000 Min. : 0.00   
## 1st Qu.: 1.50 1st Qu.:0.4930 1st Qu.:10.10   
## Median : 3.30 Median :0.6770 Median :12.30   
## Mean : 4.87 Mean :0.6276 Mean :11.99   
## 3rd Qu.: 7.20 3rd Qu.:0.7790 3rd Qu.:14.30   
## Max. :28.60 Max. :0.9480 Max. :20.70   
## NA's :34 NA's :167 NA's :163

#Looking for missing values in dataset  
gg\_miss\_var(lifeExpectancy) + labs(title = 'Missing Values', x = 'Dataset Columns')

Table

Description automatically generated

sapply(lifeExpectancy, function(x) sum(is.na(x)))

## Country Year   
## 0 0   
## Status Life.expectancy   
## 0 10   
## Adult.Mortality infant.deaths   
## 10 0   
## Alcohol percentage.expenditure   
## 194 0   
## Hepatitis.B Measles   
## 553 0   
## BMI under.five.deaths   
## 34 0   
## Polio Total.expenditure   
## 19 226   
## Diphtheria HIV.AIDS   
## 19 0   
## GDP Population   
## 448 652   
## thinness..1.19.years thinness.5.9.years   
## 34 34   
## Income.composition.of.resources Schooling   
## 167 163

**EDA**

Plot showing multiple variables we considered.

lifeExpectancy %>%   
 select(Life.expectancy, Adult.Mortality, BMI, Schooling, Status, GDP, Alcohol, thinness..1.19.years, Income.composition.of.resources) %>%  
 ggpairs(aes(color = Status, alpha = 0.5))

Calendar

Description automatically generated with medium confidence

Plot for the variables in the model we chose.

lifeExpectancy %>%   
 select(Life.expectancy,Adult.Mortality,Income.composition.of.resources,Schooling,HIV.AIDS,Diphtheria,percentage.expenditure,BMI,Polio,Year,Measles,Status) %>%  
 ggpairs(aes(color = Status, alpha = 0.5))

Chart

Description automatically generated

# look at specific variables   
ggplot(lifeExpectancy, aes(x=Life.expectancy, fill=Status)) +  
 geom\_density(alpha = 0.5) +  
 ggtitle("Distribution of Life Expectancy for Developed and Developing Countries") + theme\_classic()

Chart, histogram

Description automatically generated

ggplot(lifeExpectancy, aes(x=Status, y=GDP, fill = Status)) +  
 geom\_boxplot() +  
 ggtitle("Distribution of GDP for Developed and Developing Countries") +  
 theme\_classic()

Chart, box and whisker chart

Description automatically generated

ggplot(lifeExpectancy, aes(x=Status, y=Life.expectancy, fill = Status)) +  
 geom\_boxplot() +  
 ggtitle("Distribution of Life Expectancy for Developed and Developing Countries") + theme\_classic()

Chart, box and whisker chart

Description automatically generated

ggplot() +  
 geom\_density(data = lifeExpectancy, aes(x=HIV.AIDS), fill = "darkorange", color = "darkorange", alpha = 0.5, size = 1) +  
 geom\_density(data = lifeExpectancy, aes(x=Measles), fill = "darkblue", color = "darkblue", alpha = 0.5, size = 1) +

geom\_density(data = lifeExpectancy, aes(x=Polio), fill = "green", color = " green", alpha = 0.5, size = 1) +  
 geom\_density(data = lifeExpectancy, aes(x=Diphtheria), fill = "steelblue", color = "steelblue", alpha = 0.5, size = 1) +

xlim(0,10) +  
 labs(title = "HIV/AIDS, Measles, Diphtheria, and Polio Distributions",

caption = " HIV/AIDS = Orange, Measles = Dark Blue, Diphtheria = Light Blue, Polio = Green") +

xlab("Diseases") +

theme\_classic()

Chart

Description automatically generated

**Modeling**

# Check VIF for all variables excluding Country   
lifeExpectancy <- lifeExpectancy[,-1] # Removing Country  
  
# Make linear regression model to check for VIFs  
VIFmodel <- lm(Life.expectancy~.,data=lifeExpectancy)  
vif(VIFmodel) # Really high VIF for infant.deaths (213.611125) and under.five.deaths (203.591539)

## Year Status   
## 1.160199 1.835451   
## Adult.Mortality infant.deaths   
## 1.824077 213.611125   
## Alcohol percentage.expenditure   
## 2.396838 12.908188   
## Hepatitis.B Measles   
## 1.690982 1.517667   
## BMI under.five.deaths   
## 1.803542 203.591539   
## Polio Total.expenditure   
## 1.722587 1.125873   
## Diphtheria HIV.AIDS   
## 2.096348 1.502306   
## GDP Population   
## 13.710846 1.943602   
## thinness..1.19.years thinness.5.9.years   
## 7.611033 7.588487   
## Income.composition.of.resources Schooling   
## 3.029205 3.557069

# linear regression model after removing infant.deaths to check for VIFs  
lifeExpectancy <- lifeExpectancy[,-5] # Removing infant.deaths  
VIFmodel <- lm(Life.expectancy~.,data=lifeExpectancy)  
vif(VIFmodel) # GDP is still at 13.709051 > 10

## Year Status   
## 1.152460 1.835438   
## Adult.Mortality Alcohol   
## 1.800843 2.321353   
## percentage.expenditure Hepatitis.B   
## 12.905739 1.690848   
## Measles BMI   
## 1.407729 1.801148   
## under.five.deaths Polio   
## 2.722234 1.719379   
## Total.expenditure Diphtheria   
## 1.125718 2.076238   
## HIV.AIDS GDP   
## 1.501606 13.709051   
## Population thinness..1.19.years   
## 1.829132 7.607654   
## thinness.5.9.years Income.composition.of.resources   
## 7.543426 3.007084   
## Schooling   
## 3.548483

lifeExpectancy <- lifeExpectancy[,-15] # Removing GDP  
VIFmodel <- lm(Life.expectancy~.,data=lifeExpectancy)  
vif(VIFmodel) # All VIFs under 10 now but lets take out thinness.1.19 (7.607806) as well

## Year Status   
## 1.141683 1.828083   
## Adult.Mortality Alcohol   
## 1.799106 2.325654   
## percentage.expenditure Hepatitis.B   
## 1.415571 1.682075   
## Measles BMI   
## 1.407688 1.793609   
## under.five.deaths Polio   
## 2.719712 1.710104   
## Total.expenditure Diphtheria   
## 1.125079 2.053225   
## HIV.AIDS Population   
## 1.500842 1.828964   
## thinness..1.19.years thinness.5.9.years   
## 7.607806 7.542788   
## Income.composition.of.resources Schooling   
## 2.989325 3.522145

lifeExpectancy <- lifeExpectancy[,-16] # Removing thinness.1.19  
VIFmodel <- lm(Life.expectancy~.,data=lifeExpectancy)  
vif(VIFmodel) # All VIFs way under 10 now

## Year Status   
## 1.140928 1.826850   
## Adult.Mortality Alcohol   
## 1.797348 2.316201   
## percentage.expenditure Hepatitis.B   
## 1.415368 1.681162   
## Measles BMI   
## 1.407412 1.788927   
## under.five.deaths Polio   
## 2.702100 1.703712   
## Total.expenditure Diphtheria   
## 1.125019 2.050767   
## HIV.AIDS Population   
## 1.500452 1.827843   
## thinness.5.9.years Income.composition.of.resources   
## 1.950988 2.984263   
## Schooling   
## 3.509259

# Check summary statistics  
summary(VIFmodel)

##   
## Call:  
## lm(formula = Life.expectancy ~ ., data = lifeExpectancy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.8332 -2.2011 0.0452 2.3382 12.1761   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.393e+02 4.680e+01 7.250 6.41e-13 \*\*\*  
## Year -1.428e-01 2.337e-02 -6.111 1.24e-09 \*\*\*  
## StatusDeveloping -9.086e-01 3.421e-01 -2.656 0.00799 \*\*   
## Adult.Mortality -1.698e-02 9.579e-04 -17.725 < 2e-16 \*\*\*  
## Alcohol -1.790e-01 3.377e-02 -5.300 1.31e-07 \*\*\*  
## percentage.expenditure 4.463e-04 6.059e-05 7.365 2.80e-13 \*\*\*  
## Hepatitis.B -3.070e-03 4.533e-03 -0.677 0.49832   
## Measles 1.373e-05 1.054e-05 1.302 0.19305   
## BMI 3.296e-02 6.056e-03 5.443 6.04e-08 \*\*\*  
## under.five.deaths -2.770e-03 9.044e-04 -3.062 0.00223 \*\*   
## Polio 9.228e-03 5.166e-03 1.786 0.07424 .   
## Total.expenditure 8.774e-02 4.123e-02 2.128 0.03347 \*   
## Diphtheria 1.767e-02 5.884e-03 3.003 0.00271 \*\*   
## HIV.AIDS -4.525e-01 1.820e-02 -24.865 < 2e-16 \*\*\*  
## Population 2.900e-09 1.720e-09 1.687 0.09189 .   
## thinness.5.9.years -3.152e-02 2.690e-02 -1.172 0.24144   
## Income.composition.of.resources 1.117e+01 8.457e-01 13.204 < 2e-16 \*\*\*  
## Schooling 9.203e-01 6.003e-02 15.332 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.639 on 1639 degrees of freedom  
## (1281 observations deleted due to missingness)  
## Multiple R-squared: 0.8301, Adjusted R-squared: 0.8283   
## F-statistic: 470.9 on 17 and 1639 DF, p-value: < 2.2e-16

# Objective 1: Build regression models and identify key relationships and observe those relationships.  
  
splitPerc = .85  
trainIndices = sample(1:dim(lifeExpectancy)[1],round(splitPerc \* dim(lifeExpectancy)[1]))  
train = lifeExpectancy[trainIndices,]  
test = lifeExpectancy[-trainIndices,]  
  
# make linear regression model  
model1 <- lm(Life.expectancy~.,data=train)  
summary(model1)

##   
## Call:  
## lm(formula = Life.expectancy ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.4358 -2.1473 0.0769 2.3129 11.4281   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.628e+02 5.030e+01 7.214 8.93e-13 \*\*\*  
## Year -1.548e-01 2.512e-02 -6.165 9.26e-10 \*\*\*  
## StatusDeveloping -7.539e-01 3.717e-01 -2.028 0.04273 \*   
## Adult.Mortality -1.580e-02 1.034e-03 -15.273 < 2e-16 \*\*\*  
## Alcohol -1.781e-01 3.697e-02 -4.817 1.62e-06 \*\*\*  
## percentage.expenditure 4.226e-04 6.602e-05 6.401 2.11e-10 \*\*\*  
## Hepatitis.B -4.081e-03 4.865e-03 -0.839 0.40166   
## Measles 1.658e-05 1.086e-05 1.527 0.12699   
## BMI 3.485e-02 6.601e-03 5.279 1.51e-07 \*\*\*  
## under.five.deaths -3.250e-03 1.038e-03 -3.132 0.00177 \*\*   
## Polio 8.686e-03 5.454e-03 1.593 0.11148   
## Total.expenditure 1.111e-01 4.458e-02 2.491 0.01284 \*   
## Diphtheria 1.854e-02 6.308e-03 2.939 0.00335 \*\*   
## HIV.AIDS -4.776e-01 2.032e-02 -23.501 < 2e-16 \*\*\*  
## Population 3.631e-09 1.855e-09 1.957 0.05058 .   
## thinness.5.9.years -2.899e-02 2.945e-02 -0.984 0.32519   
## Income.composition.of.resources 1.177e+01 9.175e-01 12.825 < 2e-16 \*\*\*  
## Schooling 8.986e-01 6.493e-02 13.840 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.597 on 1377 degrees of freedom  
## (1102 observations deleted due to missingness)  
## Multiple R-squared: 0.832, Adjusted R-squared: 0.83   
## F-statistic: 401.3 on 17 and 1377 DF, p-value: < 2.2e-16

plot(model1)

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

Description automatically generatedA picture containing scatter chart

Description automatically generated

# Stepwise feature selection  
m3 <- ols\_step\_both\_p(model1)  
m3

Stepwise Selection Summary

----------------------------------------------------------------------------------------------------------------

Added/ Adj.

Step Variable Removed R-Square R-Square C(p) AIC RMSE

----------------------------------------------------------------------------------------------------------------

1 Adult.Mortality addition 0.737 0.736 1697.9110 14085.1728 4.8202

2 Income.composition.of.resources addition 0.797 0.797 768.9010 13471.3188 4.2298

3 Schooling addition 0.819 0.819 416.0310 13095.7264 3.9844

4 HIV.AIDS addition 0.823 0.823 355.0160 13043.3758 3.9391

5 Diphtheria addition 0.826 0.825 312.0030 12891.3671 3.9087

6 percentage.expenditure addition 0.828 0.827 283.9450 12866.5618 3.8870

7 BMI addition 0.829 0.828 271.2380 12855.3324 3.8768

8 Polio addition 0.830 0.829 260.0670 12845.4210 3.8677

9 Status addition 0.830 0.830 251.9820 12838.2581 3.8608

10 Measles addition 0.830 0.829 279.1360 12100.5992 3.8937

11 Year addition 0.831 0.830 275.6040 12097.6383 3.8902

----------------------------------------------------------------------------------------------------------------

plot(m3)

Graphical user interface, application

Description automatically generated

# Model from Stepwise feature selection  
model2 <- lm(Life.expectancy~Adult.Mortality+Income.composition.of.resources+Schooling+HIV.AIDS+Diphtheria+percentage.expenditure+BMI+Polio+Status+Year+Measles,data=train)  
summary(model2)

##   
## Call:  
## lm(formula = Life.expectancy ~ Adult.Mortality + Income.composition.of.resources +   
## Schooling + HIV.AIDS + Diphtheria + percentage.expenditure +   
## BMI + Polio + Status + Year + Measles, data = train)  
##   
Residuals:

Min 1Q Median 3Q Max

-21.3798 -2.2000 -0.0681 2.2405 19.9345

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.632e+02 3.684e+01 4.430 9.84e-06 \*\*\*

Adult.Mortality -1.706e-02 8.587e-04 -19.865 < 2e-16 \*\*\*

Income.composition.of.resources 8.199e+00 6.768e-01 12.114 < 2e-16 \*\*\*

Schooling 9.107e-01 4.648e-02 19.594 < 2e-16 \*\*\*

HIV.AIDS -5.010e-01 1.799e-02 -27.851 < 2e-16 \*\*\*

Diphtheria 3.313e-02 4.833e-03 6.854 9.17e-12 \*\*\*

percentage.expenditure 2.700e-04 4.612e-05 5.855 5.44e-09 \*\*\*

BMI 3.334e-02 5.104e-03 6.533 7.90e-11 \*\*\*

Polio 2.312e-02 4.792e-03 4.825 1.49e-06 \*\*\*

StatusDeveloping -8.813e-01 2.723e-01 -3.237 0.001225 \*\*

Year -5.565e-02 1.841e-02 -3.022 0.002538 \*\*

Measles -2.835e-05 7.722e-06 -3.672 0.000246 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.861 on 2303 degrees of freedom

(182 observations deleted due to missingness)

Multiple R-squared: 0.8304, Adjusted R-squared: 0.8296

F-statistic: 1025 on 11 and 2303 DF, p-value: < 2.2e-16

plot(model2)

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

vif(model2)

Adult.Mortality Income.composition.of.resources

1.760692 2.992762

Schooling HIV.AIDS

3.418962 1.430644

Diphtheria percentage.expenditure

1.922423 1.375546

BMI Polio

1.592288 1.893781

Status Year

1.625848 1.112729

Measles

1.061709

# model metrics  
hist(model2$residuals, col = "darkslateblue", main = "Histogram of Residuals")

Chart, histogram

Description automatically generated

sqrt(sum((model2$residuals)^2)) # Just the comment added by Shikha - value for this is 185.2804

## [1] 185.2804

# predictions  
preds <- predict(model2, newdata = test, interval = "prediction")  
preds <- as.data.frame(preds)  
test$Life.expectancy2 <- preds[,1]  
preds2 <- test[,c(3,19)]  
head(preds2)

Life.expectancy Life.expectancy2

14 56.2 61.66236

18 77.5 76.78103

27 73.5 73.06785

34 75.4 76.51815

42 73.4 72.28075

43 72.9 71.41124

ggplot() +  
 geom\_density(data = preds2, aes(x=Life.expectancy)) +  
 geom\_density(data = preds2, aes(x=Life.expectancy2), color = "Red") +  
 ggtitle("Predictions vs Actual Values") +  
 xlab("Life Expectancy") +  
 theme\_classic()

Chart, line chart, histogram

Description automatically generated