Logistic Regression

Aissata Bah   Brice Laurent   Linh Vu

2023-12-10

From the UN website, a country is classified as low human development when HDI is less than 0.550. So, I will create a logisitic regression for this using our choosen variables over the countries in 2021.

```
# Load Libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(knitr)

data = read.csv("./data/data_clean.csv")
```

```
selected_columns = data[, c("country","year", "region", "hdi", "x1.6","x3.2", "x5.1", "x6.4", "x7.3")]
selected_columns_2021 = selected_columns[selected_columns$year == "2021",]
selected_columns_2021$low_HD = as.numeric(selected_columns_2021$hdi <= 0.550)

# Fit logisitic regression
logreg = glm(low_HD~x1.6+x3.2+x5.1+x6.4+x7.3+region,data=selected_columns_2021,family="binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(logreg)
```

```
##
## Call:
## glm(formula = low_HD ~ x1.6 + x3.2 + x5.1 + x6.4 + x7.3 + region,
##     family = "binomial", data = selected_columns_2021)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.69800  -0.06671  -0.00005  -0.00001   2.22328
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -16.4894  3438.9872  -0.005   0.9962
## x1.6                                  8.8816     4.5466   1.953   0.0508 .
## x3.2                                 -3.8248     6.4076  -0.597   0.5506
## x5.1                                 -1.0829     4.3682  -0.248   0.8042
## x6.4                                 -7.2821     5.2367  -1.391   0.1644
## x7.3                                 -9.2245     4.4534  -2.071   0.0383 *
## regionEastern Europe & Central Asia    0.6260  5778.0978   0.000   0.9999
## regionEU + EFTA + North America        2.9538  4440.9708   0.001   0.9995
## regionLatin America & Caribbean       16.5993  3438.9861   0.005   0.9961
## regionMiddle East & North Africa       0.7979  6464.6016   0.000   0.9999
## regionSouth Asia                      19.8033  3438.9859   0.006   0.9954
## regionSub-Saharan Africa              21.7462  3438.9859   0.006   0.9950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.522  on 137  degrees of freedom
## Residual deviance:  43.444  on 126  degrees of freedom
## AIC: 67.444
##
## Number of Fisher Scoring iterations: 19
```

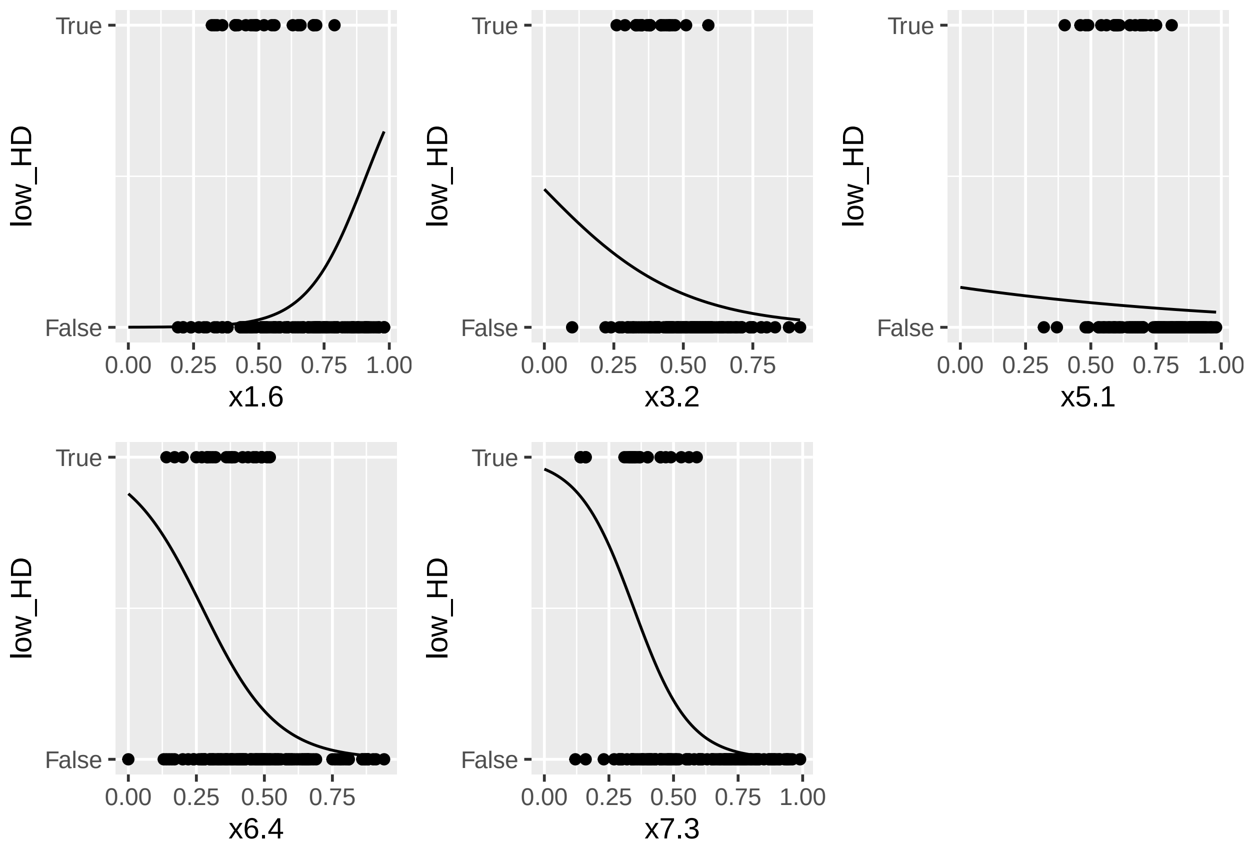
```r
kable(summary(logreg)$coefficients)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.4894055	3438.987249	-0.0047948	0.9961743
x1.6	8.8815832	4.546560	1.9534733	0.0507635
x3.2	-3.8247623	6.407582	-0.5969119	0.5505662
x5.1	-1.0829325	4.368209	-0.2479123	0.8042023
x6.4	-7.2820756	5.236693	-1.3905867	0.1643508
x7.3	-9.2244655	4.453396	-2.0713330	0.0383277
regionEastern Europe & Central Asia	0.6259511	5778.097818	0.0001083	0.9999136
regionEU + EFTA + North America	2.9537822	4440.970799	0.0006651	0.9994693
regionLatin America & Caribbean	16.5993435	3438.986100	0.0048268	0.9961488
regionMiddle East & North Africa	0.7978667	6464.601585	0.0001234	0.9999015
regionSouth Asia	19.8032694	3438.985864	0.0057585	0.9954054
regionSub-Saharan Africa	21.7461763	3438.985874	0.0063234	0.9949547

```r
logreg_noRegion = glm(low_HD~x1.6+x3.2+x5.1+x6.4+x7.3,data=selected_columns_2021,family="binomial")

anova(logreg_noRegion, logreg, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: low_HD ~ x1.6 + x3.2 + x5.1 + x6.4 + x7.3
## Model 2: low_HD ~ x1.6 + x3.2 + x5.1 + x6.4 + x7.3 + region
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       132     79.457
## 2       126     43.444  6   36.012 2.742e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Plot regression
dummy_x1.6 = seq(0,max(selected_columns_2021$x1.6,na.rm=T),0.01)
dummy_x3.2 = seq(0,max(selected_columns_2021$x3.2,na.rm=T),0.01)
dummy_x5.1 = seq(0,max(selected_columns_2021$x5.1,na.rm=T),0.01)
dummy_x6.4 = seq(0,max(selected_columns_2021$x6.4,na.rm=T),0.01)
dummy_x7.3 = seq(0,max(selected_columns_2021$x7.3,na.rm=T),0.01)

average_x1.6 = mean(selected_columns_2021$x1.6,na.rm=T)
average_x3.2 = mean(selected_columns_2021$x1.6,na.rm=T)
average_x5.1 = mean(selected_columns_2021$x1.6,na.rm=T)
average_x6.4 = mean(selected_columns_2021$x1.6,na.rm=T)
average_x7.3 = mean(selected_columns_2021$x1.6,na.rm=T)

yhat_x1.6 = predict(logreg,new=data.frame(x1.6=dummy_x1.6, x3.2 = rep(average_x3.2, length(dummy_x1.6))
yhat_x3.2 = predict(logreg,new=data.frame(x1.6=rep(average_x1.6, length(dummy_x3.2)), x3.2 = dummy_x3.2
yhat_x5.1 = predict(logreg,new=data.frame(x1.6=rep(average_x1.6, length(dummy_x5.1)), x3.2 = rep(average
yhat_x6.4 = predict(logreg,new=data.frame(x1.6=rep(average_x1.6, length(dummy_x6.4)), x3.2 = rep(average
yhat_x7.3 = predict(logreg,new=data.frame(x1.6=rep(average_x1.6, length(dummy_x7.3)), x3.2 = rep(average

phat_x1.6 = exp(yhat_x1.6)/(1+exp(yhat_x1.6))
phat_x3.2 = exp(yhat_x3.2)/(1+exp(yhat_x3.2))
phat_x5.1 = exp(yhat_x5.1)/(1+exp(yhat_x5.1))
phat_x6.4 = exp(yhat_x6.4)/(1+exp(yhat_x6.4))
phat_x7.3 = exp(yhat_x7.3)/(1+exp(yhat_x7.3))

p1 = ggplot() + geom_point(data = selected_columns_2021, aes(x=x1.6, y=low_HD)) + geom_line(aes(x=dummy_
p2 = ggplot() + geom_point(data = selected_columns_2021, aes(x=x3.2, y=low_HD)) + geom_line(aes(x=dummy_
p3 = ggplot() + geom_point(data = selected_columns_2021, aes(x=x5.1, y=low_HD)) + geom_line(aes(x=dummy_
p4 = ggplot() + geom_point(data = selected_columns_2021, aes(x=x6.4, y=low_HD)) + geom_line(aes(x=dummy_
p5 = ggplot() + geom_point(data = selected_columns_2021, aes(x=x7.3, y=low_HD)) + geom_line(aes(x=dummy_

grid.arrange(p1,p2,p3,p4,p5, nrow = 2)
```



Steep slope, more significant

Thereis equation for shift and slope, that can be meaning

Try random forest, explain which will trust more.