Ce le face pe oameni fericiți?

Analiza regresiei liniare multiple a nivelului de fericire al persoanelor din 156 țări ale lumii din anul 2019

Student: Bobu Natalia

## Introducere

În analiza regresiei liniare multiple, încercăm să încadrăm un model predictiv la datele noastre și utilizăm acest model pentru a prezice o variabilă rezultat din mai multe variabile predictor independente.

Acest model predictiv utilizează o linie dreaptă pentru a rezuma datele, iar metoda celor mai mici pătrate este utilizată pentru a obține dreapta care oferă descrierea (cea mai potrivită) a datelor.

## Setul de date: Happy 2019

Sursă: <https://www.kaggle.com/unsdsn/world-happiness#2019.csv>

Setul de date pe care l-am ales este Happy 2019, unul dintre seturile de date Kaggle. Raportul despre fericirea mondială este un sondaj reper al stării fericirii globale. Acest set de date oferă scorul de fericire al 156 de țări din întreaga lume, bazate pe șase factori, inclusiv PIB/persoană, suportul social, speranța de viață, libertatea alegerii, generozitatea și corupția. Suma valorilor acestor șase factori, fiecare având tipul de date numeric, ne oferă scorul de fericire și cu cât scorul de fericire este mai mare, cu atât țara este considerată mai fericită.

Scorurile și clasamentele de fericire folosesc date din Sondajul Mondial Gallup . Scorurile se bazează pe răspunsurile la întrebarea principală de evaluare a vieții adresată în sondaj. Această întrebare, cunoscută sub numele de scara Cantril, solicită respondenților să se gândească la o scară cu cea mai bună viață posibilă pentru ei fiind 10 și cea mai rea viață posibilă fiind 0 și să își evalueze propriile vieți curente pe acea scară. Scorurile sunt din eșantioane reprezentative la nivel național și folosesc ponderile Gallup pentru a face reprezentativele estimărilor. Coloanele care urmează scorul de fericire estimează măsura în care fiecare dintre cei șase factori - producția economică, sprijinul social, speranța de viață, libertatea, absența corupției și generozitatea - contribuie la creșterea evaluărilor vieții în fiecare țară.

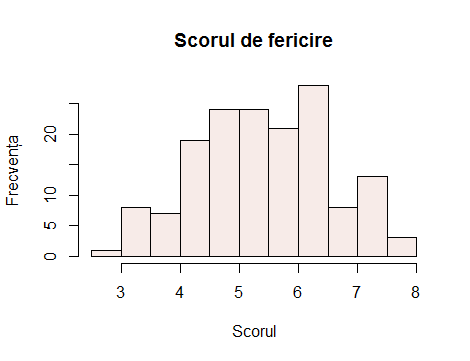
## Scop

Scopul alegerii acestei lucrări este de a afla care factori sunt mai importanți pentru a trăi o viață mai fericită. Drept urmare, oamenii și țările se pot concentra pe factorii mai importanți pentru a atinge un nivel mai mare de fericire.

## Normalitate

summary(Happiness$Score)

hist(Happiness$Score,main="Scorul de fericire",xlab="Scorul",ylab="Frecvența",col=rgb(0.97,0.92,0.91))



kurtosis(Happiness$Score)



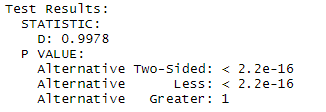
skewness(Happiness$Score)



Valoare Kurtosis este negativă, deci repartizarea este puțin aplatizată în vârf și cu mai puține date în coadă, iar valoare Skewness care este aproape 0, denotă faptul că datele sunt practic repartizate normal.

library(fBasics)

ksnormTest(Happiness$Score,title = "Kolmogorov-Smirnov",description = NULL)



## Sumarul setului de date:

Happiness=Happiness[c(3:9)]

summary(Happiness$GDP.capita)

boxplot(Happiness$GDP.capita,horizontal = TRUE,main="PIB/capita",col=rgb(0.58,0.65,0.81))

A screenshot of a cell phone

Description automatically generated

summary(Happiness$Social.support)

boxplot(Happiness$Social.support,horizontal = TRUE,main="Suportul social",col=rgb(0.79,0.89,0.97))

A picture containing orange, room

Description automatically generated

A screenshot of a cell phone

Description automatically generated

summary(Happiness$Life.expectancy)

boxplot(Happiness$Life.expectancy,horizontal = TRUE,main="Speranța de viață",col=rgb(0.86,0.92,0.71))



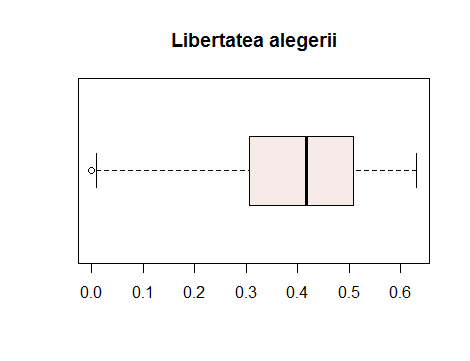
A screenshot of a cell phone

Description automatically generated

summary(Happiness$Freedom.of.choice)

boxplot(Happiness$Freedom.of.choice,horizontal = TRUE,main="Libertatea alegerii",col=rgb(0.97,0.92,0.91))





summary(Happiness$Generosity)

boxplot(Happiness$Generosity,horizontal = TRUE,main="Generozitatea",col=rgb(0.98,0.73,0.76))

A screenshot of a cell phone

Description automatically generated

summary(Happiness$Corruption)

boxplot(Happiness$Corruption,horizontal = TRUE,main="Corupția",col=rgb(0.83,0.80,0.91))

A picture containing orange, red

Description automatically generated

A screenshot of a cell phone

Description automatically generated

## Vizualizare

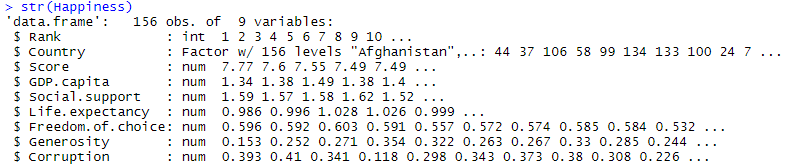
1. Mai întai de toate trebuie să încărcăm și să vizualizăm setul nostru de date pentru a avea o mai bună percepție a acestuia.

setwd("E:/Univ/Info3/Modelare si simulare/ProiectMS/")

Happiness= read.csv("happy2019.csv")

Happiness

str(Happiness)



După cum vedem în imagine, setul nostru conține 156 de înregistrări a 9 variabile, două dintre care sunt descriptive, pentru a detalia resul înregistrărilor și anume țara și poziția ei în rating. Scorul va fi variabila noastră dependentă, iar restul 6 variabile vor fi cele independente.

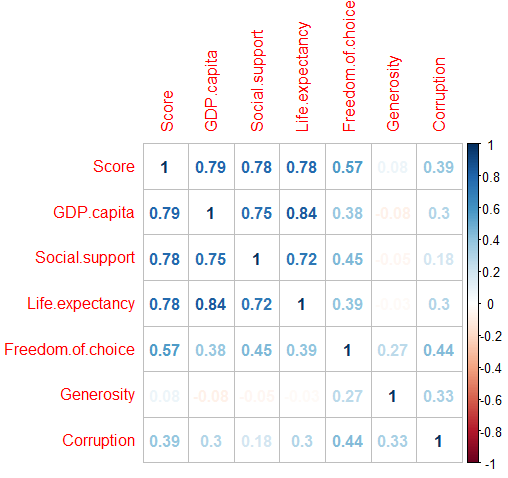
1. Correlation plot ne oferă posibilitatea să vedem corelația dintre variabilele numerice din setul nostru

library(corrplot)

happy = cor(Happiness[c(3:9)])

corrplot(happy, method = 'number')

corrplot(happy, method = 'circle')

 A screenshot of a cell phone

Description automatically generated

După cum se vede, PIB-ul are cel mai mare impact asupra scorului, urmat de suportul social și speranța de viață, care sunt la același nivel. Cea mai mica însemnătate pentru scorul final o are generozitatea.

1. Acum, vom vizualiza corelația dintre scorul de fericire și fiecare dintre cei șase factori separate, prin scatter plot și dreapta de regresie.

happy=Happiness[c(3:9)]

library(ggplot2)

1. Corelația între scorul de fericire și PIB

cor(Happiness$Score,Happiness$GDP.capita)



ggplot(happy, aes(x = GDP.capita, y = Score)) +

geom\_point(size = 3, alpha = 0.8,col=rgb(0.58,0.65,0.81)) +

geom\_smooth(method = "lm", fullrange = TRUE) +

theme\_bw() + labs(title = "Fericire~PIB")A close up of a map

Description automatically generated

1. Corelația între scorul de fericire și Suportul social

cor(Happiness$Score,Happiness$Social.support)



ggplot(happy, aes(x = Social.support, y = Score)) +

geom\_point(size = 3, alpha = 0.8,col=rgb(0.79,0.89,0.97)) +

geom\_smooth(method = "lm", fullrange = TRUE) +

theme\_bw() + labs(title = "Fericire~Suportul social")

A close up of a map

Description automatically generated

1. Corelația între scorul de fericire și Speranța de viață

cor(Happiness$Score,Happiness$Life.expectancy)



ggplot(happy, aes(x = Life.expectancy, y = Score))

+geom\_point(size = 3, alpha = 0.8,col=rgb(0.86,0.92,0.71))

+geom\_smooth(method = "lm", fullrange = TRUE)

+theme\_bw() + labs(title = "Fericire~Speranța de viață")

A close up of a map

Description automatically generated

1. Corelația între scorul de fericire și Libertarea alegerii

cor(Happiness$Score,Happiness$Freedom.of.choice)



ggplot(happy, aes(x = Freedom.of.choice, y = Score)) +

geom\_point(size = 3, alpha = 0.8,col='pink') +

geom\_smooth(method = "lm", fullrange = TRUE) +

theme\_bw() + labs(title = "Fericire~Libertatea alegerii")

A close up of a map

Description automatically generated

1. Corelația între scorul de fericire și Generozitatea

cor(Happiness$Score,Happiness$Generosity)



ggplot(happy, aes(x = Generosity, y = Score)) +

geom\_point(size = 3, alpha = 0.8,col=rgb(1,0.7,0.7)) +

geom\_smooth(method = "lm", fullrange = TRUE) +

theme\_bw() + labs(title = "Fericire~Generozitate")

A screenshot of a cell phone

Description automatically generated

1. Corelația între scorul de fericire și Corupția

cor(Happiness$Score,Happiness$Corruption)



ggplot(happy, aes(x = Corruption, y = Score)) +

geom\_point(size = 3, alpha = 0.8,col=rgb(0.83,0.80,0.91)) +

geom\_smooth(method = "lm", fullrange = TRUE) +

theme\_bw() + labs(title = "Fericire~Coruptie")

A close up of a map

Description automatically generated

# Modelul liniar

În această secțiune o să încercăm să prezicem nivelul de fericire. Urmează să construim modelul nostru pentru a putea ulterior vizualiza datele statistice despre aceasta cu ajutorul funcției summary.

happy\_mlm=lm(Score ~ GDP.capita + Social.support+Life.expectancy

+Freedom.of.choice + Generosity + Corruption, data=happy)

summary(happy\_mlm)

A screenshot of a cell phone

Description automatically generated

După cum vedem în imagine doar 4 dintre cele 6 variabile independente au impact semnififcativ asupra scorului de fericire, aceasta reiese din p-valorile lor, care sunt <0.05 la toate în afară de Generozitate și Corupție. Valoarea Adjusted R-square este 0.7703, care ne arată că modelul nostru va prezice scorul fericirii cu o precizie de aproximativ 70%.

O să încercăm să construim un alt model fără acele două variabile, pentru a compara care dintre ele este mai bun, sau dacă există vreo diferență sau îmbunătățire.

happy\_mlm1=lm(Score ~ GDP.capita + Social.support + Life.expectancy

+ Freedom.of.choice, data=happy)

summary(happy\_mlm1)

A screenshot of a cell phone

Description automatically generated

În acest model am eliminat Generozitatea și Corupția. Putem să observăm că p-valorile tuturor variabilelor independente sunt mai mici decât 0.05, deci au o pondere semnificativă pentru scorul de fericire. În schimb valoarea Adjusted R-squared, s-a modificat doar cu o unitate și este acum 0.7649, deci cele două modele nu diferă mult unul față de altul.

Un al treilea model o să îl contruim eliminând doar Generozitatea.

happy\_mlm2 = lm(Score ~ GDP.capita + Social.support + Life.expectancy

+Freedom.of.choice + Corruption, data=happy)

summary(happy\_mlm2)

A screenshot of a cell phone

Description automatically generated

Acest model are și el p-valorile variabilelor independente mai mici decât 0.05, iar valoarea Adjusted R-squared este și mai bună, 0.7703.

Ca să alegem modelul mai bun, o să aplicăm testul Akaike, care ne va arăta un coeficient – coeficientul Akaike, modelul cel mai bun va fi cel cu coeficientul Akaike cel mai mic.

aic1=AIC(happy\_mlm)

aic2=AIC(happy\_mlm1)

aic3=AIC(happy\_mlm2)

print(aic1)

print(aic2)

print(aic3)

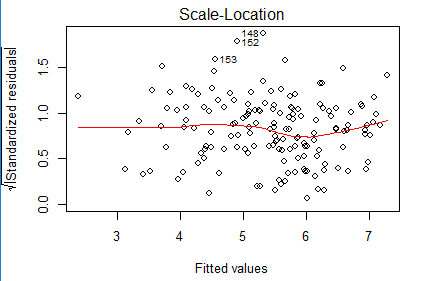
A screenshot of a cell phone

Description automatically generated

Cel mai mic coeficient Akaike îl are ultimul model, deci pe acest îl vom alege ca să prezicem scorul de fercire. Este bine de menționat că în cazul primelor două modele coeficientul Akaike este mai bun la cel cu 6 variabile independente, chiar dacă valoarea Adjusted R-squared este mai bună la cel cu 4.

par(mfrow=c(2,2))

plot(happy\_mlm2)



Linia reprezentată este practic aplatizată, semnificând absența homoscedasticității, deci varianța rezidurilor nu trebuie să crească cu valorile prezise.

car::vif(happy\_mlm2)

A close up of a mans face

Description automatically generated

Observăm că în urma testului de multicolinearitate, nici una dintre variabilele independente nu întrece valoare 5, rezultă că nu există coliniaritate semnificativă în modelul nostru.

# Predicție

Odată ce am ales modelul potrivit, putem să prezicem valoarea scorului de fericire, în dependență de celelalte variabile. În primul rând, o să împărțim dataset-ul nostru în altele două mici, training și test.

Training = head(happy, 124)

test = tail(happy, 32)

happy\_mlm3 = lm(Score ~ GDP.capita + Social.support + Life.expectancy

+Freedom.of.choice, data=training)

prediction=predict(happy\_mlm3, test)

prediction\_vs\_actual = as.data.frame(cbind(Prediction = prediction

,Actual = test$Score))

ggplot(prediction\_vs\_actual, aes(Actual, Prediction,col='pink'))

+geom\_point() + geom\_abline()

+labs(title = "Regresie liniară multiplă", x = "Scorul real"

,y = "Scorul prezis")

A picture containing covered, white, table, large

Description automatically generated A screenshot of a cell phone

Description automatically generated

Acum o să încercăm să prezicem scorul de fericire, având ca variabile independente, valori separate de cele existente în dataset.

newdata=data.frame(GDP.capita=1.35,Social.support=1.52

,Life.expectancy=0.99,Freedom.of.choice=0.55,Corruption=0.11)

predict(happy\_mlm2, newdata)



Valorile pe care le-am introdus pentru variabilele independente au fost selectate aleator de la primele 10 țări din listă. Valoarea scorului de fericire prezisă însă ar corespunde țării de pe poziția 20 din rating.

## Concluzii

Modelul liniar construit prezice datele destul de bine, dar se dorește un model care ar avea o precizie de 100%. Instabilitatea este motivată de colectarea datelor, care a fost în mediul online, posibil cu pe un eșantion mic, dar variabilele independente, unele dintre ele ar fi necer de înlocuit cu altele mult mai semnificative petru scorul final.

# Bibliografie

* *Cursul de Modelare si simulare*, conf. dr. Alina Barbulescu, Univ. Ovidius Constanta, 2020
* *Laboratorul de Modelare si simulare*, lect.dr. Elena Bautu, Univ. Ovidius Constanta, 2020
* <https://www.kaggle.com/unsdsn/world-happiness>
* *R Tutorial with Bayesian Statistics using Stan*, Chi Yau, 2009
* *R Data analysis,* Institute for Digital Research and Education, University of California, Los Angeles, 2020
* *Regression Models for Categorical and Limited Dependent Variables.* Long, J. Scott, 1997.
* *A Handbook of Statistical Analyses Using R (1st Edition),* Brian S. Everitt and Torsten Hothorn, 2017