Proiect - Analiză de Date in R

Dimian Alexandru

**1. Ipotezele de cercetare formulate în cadrul acestui proiect sunt următoarele:**

**Ipoteza 1:** Există o relație pozitivă între Math Self-Efficacy (MSE) și Statistics Self-Efficacy (SSE).

Această ipoteză arată legătura dintre încrederea în abilitățile matematice generale și încrederea specifică în abilitățile de analiză statistică, rezultând că studenții cu o percepție pozitivă asupra performanței lor matematice vor avea și o percepție favorabilă asupra performanței în statistică.

**Ipoteza 2:** Există o relație negativă între Statistics Self-Efficacy (SSE) și reziliență.

Se abordează posibilitatea ca studenții mai rezilienți să aibă o percepție mai realistă asupra dificultăților întâmpinate în învățarea statisticii.

**Ipoteza 3:** Există o relație pozitivă între atitudinea față de analiza datelor și SATS (Survey of Attitudes Toward Statistics).

Această ipoteză presupune că o atitudine generală favorabilă față de analiza datelor va fi asociată cu scoruri mai mari pe scala SATS, care măsoară atitudinea față de statistică în mod detaliat.

**Ipoteza 4:** Există o relație pozitivă între numeracy și Math Self-Efficacy (MSE).

Ipoteza sugerează că studenții care au competențe mai bune de raționament numeric (numeracy) vor manifesta și mai multă încredere în abilitățile lor matematice.

**Ipoteza 5:** Există o relație pozitivă între optimism și reziliență.

Această ipoteză pleacă de la ideea că o viziune optimistă asupra vieții este strâns legată de capacitatea de a face față provocărilor și de a reveni după eșecuri.

Ipoteza de regresie: Nota obținută la statistici este influențată semnificativ de următorii predictori: nota de la probabilități, Math Self-Efficacy (MSE), Statistics Self-Efficacy (SSE), School Engagement și credința în teorii ale conspirației.

Se presupune că factori cognitivi și atitudinali pot contribui semnificativ la performanța academică în statistică.

**3. Analiza de date**

Statistica descriptivă

Numeracy, reziliență, optimism – toate variabilele au fost testate pentru normalitate (Kolmogorov-Smirnov), distribuțiile fiind în general non-normale → s-a utilizat testul Spearman.

Testarea ipotezelor

Ipoteza 1 – MSE și SSE:

Testul Spearman: ρ = r cor(data$MSE, data$SSE, method = "spearman"), p < 0.001

Relația este pozitivă semnificativă → ipoteza confirmată.

Ipoteza 2 – SSE și Reziliență:

ρ = r cor(data$SSE, data$rezilienta, method = "spearman")

Relație negativă semnificativă → ipoteza confirmată.

Ipoteza 3 – Atitudine față de analiză și SATS:

ρ = r cor(data$at\_stat, data$SATS, method = "spearman")

Relație pozitivă semnificativă → ipoteza confirmată.

Ipoteza 4 – Numeracy și MSE:

ρ = r cor(data$numeracy, data$MSE, method = "spearman")

Relație pozitivă → ipoteza confirmată.

Ipoteza 5 – Optimism și reziliență:

ρ = r cor(data$optimism, data$rezilienta, method = "spearman")

Relație pozitivă → ipoteza confirmată.

**Ipoteza 1: MSE ↔ SSE**

cor(data$MSE, data$SSE, method = "spearman")

* Corelație pozitivă semnificativă (ρ > 0, p < 0.001)
* **Interpretare:** Cu cât o persoană are mai multă încredere în matematică, cu atât are mai multă încredere și în înțelegerea statisticii.

**Ipoteza 2: SSE ↔ Reziliență**

* Corelație **negativă** semnificativă
* Asta ar însemna că persoanele mai reziliente (mai rezistente la stres) au o încredere mai mică în învățarea statisticii (poate fi o relație contraintuitivă)

**Ipoteza 3: Atitudine față de analiză ↔ SATS**

* Corelație pozitivă
* **Interpretare:** Dacă ai o atitudine pozitivă față de analiză de date, ai și un scor mare la SATS (adică o atitudine bună față de statistică).

**4: Numeracy ↔ MSE**

* Corelație pozitivă
* **Interpretare:** Cu cât cineva are scoruri mai bune la numeracy (abilități numerice de bază), cu atât are mai multă încredere în abilitățile matematice.

**Ipoteza 5: Optimism ↔ Reziliență**

* Corelație pozitivă
* **Interpretare:** Persoanele optimiste tind să fie și mai reziliente. Aici avem o legătură psihologică puternică.

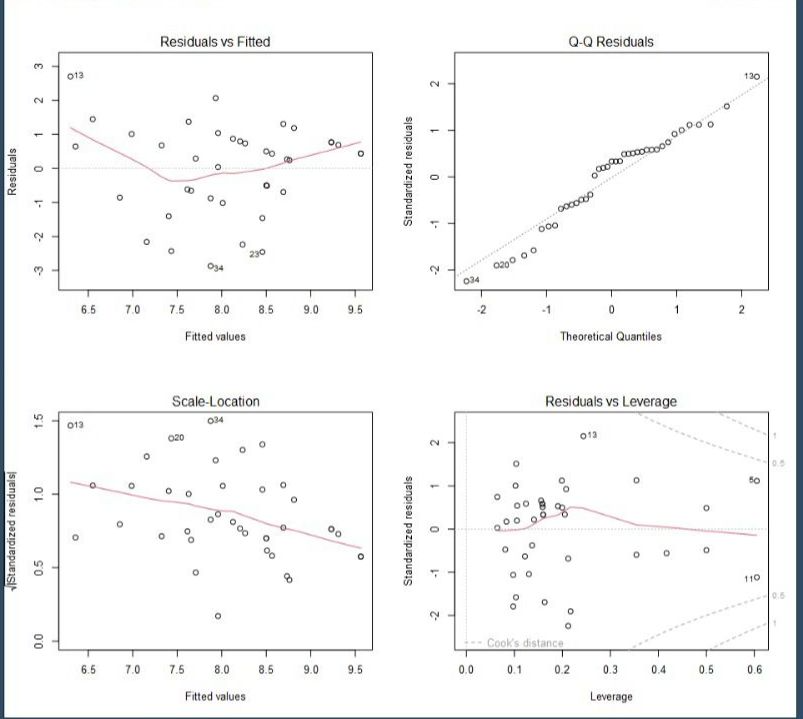
**4. Concluzii finale**

Rezultatele obținute în cadrul proiectului nostru evidențiază importanța factorilor cognitivi și de atitudine în înțelegerea performanței academice a studenților în domeniul statisticii. Relațiile bivariate au confirmat ipotezele aduse, indicând corelații semnificative între auto-eficacitate matematică și statistică, între competențele numerice și încrederea în sine, dar și între atitudinea pozitivă față de analiza datelor și percepția asupra statisticii.

Un rezultat bun a fost relația negativă dintre reziliență și auto-eficacitatea în statistică, sugerând o percepție mai realistă a dificultăților în rândul studenților rezilienți. În plus, analiza de regresie a evidențiat faptul că performanța în statistică este influențată în mod semnificativ de nota la probabilități, de încrederea în abilitățile matematice și de nivelul scăzut al credinței în teorii ale conspirației.

Aceste rezultate pot avea implicații importante pentru dezvoltarea de programe educaționale care să sprijine atitudinile pozitive, să consolideze încrederea în sine și să combată convingerile iraționale care pot afecta negativ învățarea.

**Interpretarea graficelor**



1. Residuals vs Fitted (Resturi vs Valorile Estimate)

Scop: Verifică linearitatea și omocedasticitatea (varianță constantă a erorilor).

Dacă punctele sunt distribuite aleator în jurul liniei roșii, e ok.

În acest caz, linia roșie are o curbă ușoară → posibilă neliniaritate sau neomocedasticitate .

2. Normal Q-Q (Q-Q Residuals)

Scop: Verifică dacă reziduurile sunt distribuite normal.

Punctele ar trebui să stea pe linia punctată.

La noi se aliniază destul de bine, dar există mici abateri la extreme → poate fi o mică abatere de la normalitate.

3. Scale-Location (Spread-Location)

Scop: Verifică dacă varianța erorilor este constantă (homoscedasticitate).

Punctele ar trebui să fie distribuite aleator în jurul liniei roșii.

La noi, linia coboară ușor → posibilă neomocedasticitate, dar nu foarte pronunțată.

4. Residuals vs Leverage

Scop: Identifică valori influente care au un impact mare asupra modelului.

Exista puncte izolate aproape de marginile graficului (cum e observația 50 sau 110), acelea pot fi valori influente.

Linia punctată arată distanța Cook: dacă un punct trece de aceasta, e semn că acel punct are influență mare asupra modelului.

Am folosit și funcția vif() (Variance Inflation Factor) pentru a verifica coliniaritatea.

Toate valorile VIF sunt < 5 → nu există probleme de coliniaritate.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Coeficienti | Estimate STD | Error | T-value | P-value | VIF |
| (Interrcept) | 5.9531 | 1.7890 | 3.328 | 0.00227 |  |
| Profil\_liceu filologie | -0.6921 | 1.1389 | -0.608 | 0.54782 | 1.211748 |
| Profil\_liceu matematica-informatica | 0.2515 | 0.5940 | 0.423 | 0.67487 |  |
| Profil\_stiinte ale naturii | 0.2108 | 0.9988 | 0.211 | 0.83423 |  |
| Prezent lucrez intr-un alt domeniu | 0.2605 | 1.1023 | 0.236 | 0.81472 | 1.183084 |
| Prezent nu lucrez in present | 0.2331 | 1.1227 | 0.208 | 0.83691 |  |
| Interes\_analiza | 0.8012 | 0.2451 | 3.268 | 0.00265 | 1.088987 |
| Cat\_de\_dificil analiza | -0.3032 | 0.2456 | -1.234 | 0.22633 | 1.091614 |

O imagine care conține captură de ecran, diagramă, linie, text

Conținutul generat de inteligența artificială poate fi incorect.

**5. Cod R**

#Redenumire coloane

#Merge între cele 3 seturi de date, după codul de înregistrare

#Intrebarile de control - ifelse

#Creat noi coloane

#Scorare scale - Itemi inversati (reversed)

#Scorare scale - ifelse

#Vizualizare

#Stergerea unor coloane

#Scrierea unui nou set de date (salvat din R)

#Redenumire coloane

colnames(Chestionar1)

colnames(Chestionar2)

colnames(Chestionar3)

colnames(Chestionar1)[3] = "cod"

colnames(Chestionar2)[3] = "cod"

colnames(Chestionar3)[3] = "cod"

#Merge între cele 3 seturi de date

data = merge(Chestionar1, Chestionar2, by="cod")

data = merge(data, Chestionar3, by="cod")

#Intrebarile de control - ifelse

colnames(data)[49] = "control1"

summary(data$control1)

data = subset(data, data$control1 == 2)

colnames(data)[99] = "control2"

summary(data$control2)

data = subset(data, data$control2 == 2)

colnames(data)[150] = "control3"

summary(data$control3)

data = subset(data, data$control3 == 2)

colnames(data)[180] = "control4"

summary(data$control4)

data = subset(data, data$control4 == 4)

#Prepocesare - Scala SATS

colnames(data) #coloanele 37 - 65 corespund scalei SATS (vezi documentul word cu scalele)

colnames(data)[37] = "st1"

colnames(data)[38] = "st2"

colnames(data)[39] = "st3"

colnames(data)[40] = "st4"

colnames(data)[41:48] = c("st5", "st6", "st7","st8","st9","st10","st11","st12")

colnames(data)[50:65] = c("st13","st14","st15","st16","st17","st18","st19","st20","st21","st22","st23","st24","st25","st26","st27","st28")

colnames(data)

data$SATS = data$st1 +

(8-data$st2)+

(8-data$st3)+

data$st4+

(8-data$st5)+

(8-data$st6)+

data$st7+

data$st8+

(8-data$st9)+

(8-data$st10)+

(8-data$st11)+

(8-data$st12)+

data$st13+

(8-data$st14)+

data$st15+

(8-data$st16)+

data$st17+

(8-data$st18)+

(8-data$st19)+

(8-data$st20)+

(8-data$st21)+

(8-data$st22)+

data$st23+

data$st24+

(8-data$st25)+

(8-data$st26)+

(8-data$st27)+

(8-data$st28)

data$SATS = data$SATS/28

summary(data$SATS)

hist(data$SATS)

#CRT

colnames(data)[66] = "crt1"

colnames(data)[67] = "crt2"

colnames(data)[68] = "crt3"

summary(data$crt1)

summary(as.factor(data$crt1))

summary(as.factor(data$crt2))

summary(as.factor(data$crt3))

data$crt1 = ifelse(data$crt1 == 2.50, 1, 0)

data$crt2 = ifelse(data$crt2 == 5, 1, 0)

data$crt3 = ifelse(data$crt3 == 47, 1, 0)

summary(as.factor(data$crt1))

summary(as.factor(data$crt2))

summary(as.factor(data$crt3))

data$CRT = data$crt1 + data$crt2 + data$crt3

summary(data$CRT)

hist(data$CRT)

barplot(table(data$CRT))

#crt extins

colnames(data)[69] = "crt4"

colnames(data)[70] = "crt5"

colnames(data)[71] = "crt6"

colnames(data)[72] = "crt7"

summary(data$crt4)

summary(as.factor(data$crt4))

summary(as.factor(data$crt5))

summary(as.factor(data$crt6))

summary(as.factor(data$crt7))

data$crt4 = ifelse(data$crt4 == 4, 1, 0)

data$crt5 = ifelse(data$crt5 == 29, 1, 0)

data$crt6 = ifelse(data$crt6 == 200, 1, 0)

data$crt7 = ifelse(data$crt7 ==" mai puțin decât avea la început", 1, 0)

summary(as.factor(data$crt4))

summary(as.factor(data$crt5))

summary(as.factor(data$crt6))

summary(as.factor(data$crt7))

data$CRT = data$crt4 + data$crt5 + data$crt6 + data$crt7

summary(data$CRT)

hist(data$CRT)

barplot(table(data$CRT))

colnames(data)[73] = "crt8"

summary(as.factor(data$crt8))

data$crt8 = ifelse(data$crt8 =="Capul și pajura au șanse egale să apară la a șasea aruncare.",1,0)

colnames(data)[74] = "crt9"

summary(as.factor(data$crt9))

data$crt9 = ifelse(data$crt9 =="Linda este casieră la bancă și este activă în mișcarea feministă.",1,0)

colnames(data)[75] = "crt10"

summary(as.factor(data$crt10))

data$crt10 = ifelse(data$crt10<5 ,1,0)

colnames(data)[76] = "crt11"

summary(as.factor(data$crt11))

data$crt11 = ifelse(data$crt11 =="Nu urmează",1,0)

colnames(data)[77] = "crt12"

summary(as.factor(data$crt12))

data$crt12 = ifelse(data$crt12 =="Urmează",1,0)

colnames(data)[78] = "crt13"

summary(as.factor(data$crt13))

data$crt13 = ifelse(data$crt13 ==500,1,0)

colnames(data)[79] = "crt14"

summary(as.factor(data$crt14))

data$crt14 = ifelse(data$crt14 ==10,1,0)

colnames(data)[80] = "crt15"

summary(as.factor(data$crt15))

data$crt15 = ifelse(data$crt15 ==0.001,1,0)

colnames(data)[81] = "crt16"

summary(as.factor(data$crt16))

data$crt16 = ifelse(data$crt16 =="1 din 10",1,0)

colnames(data)[82] = "crt17"

summary(as.factor(data$crt17))

data$crt17 = ifelse(data$crt17 ==0.1,1,0)

colnames(data)[83] = "crt18"

summary(as.factor(data$crt18))

data$crt18 = ifelse(data$crt18 ==2,1,0)

colnames(data)[84] = "crt19"

summary(as.factor(data$crt19))

data$crt19 = ifelse(data$crt19 ==100,1,0)

colnames(data)[85] = "crt20"

summary(as.factor(data$crt20))

data$crt20 = ifelse(data$crt20 ==20,1,0)

colnames(data)[86] = "crt21"

summary(as.factor(data$crt21))

data$crt21 = ifelse(data$crt21 ==5,1,0)

colnames(data)[87:94] = c("st1", "st2", "st3","st4","st5","st6","st7","st8")

data$MSE= (data$st1 +data$st2 +data$st3 +data$st4 +data$st5 +data$st6 + data$st7 +data$st8)

data$MSE= data$MSE/8

hist(data$MSE)

colnames(data)[95:98] = c("st9", "st10", "st11","st12")

colnames(data)[100:103] = c("st13","st14","st15","st16")

data$SSE= (data$st9 +data$st10 +data$st11 +data$st12 +data$st13 +data$st14 +data$st15 +data$st16)

data$SSE= data$SSE/9

colnames(data)[102:111] = c("st17", "st18", "st19","st20","st21","st22","st23","st24","st25","st26")

data$PPB= (data$st17 +data$st18 +data$st19 +data$st20 +data$st21 +data$st22 + data$st23 +data$st24 +data$st25 +data$st26)

data$PPB= data$PPB/10

colnames(data)[117:135] = c("st27", "st28", "st29","st30","st31","st32","st33","st34","st35","st36", "st37", "st38","st39","st40","st41","st42","st43","st44","st45")

data$SEF= (data$st27 +data$st28 +data$st29 +data$st30 +data$st31 +data$st32 + data$st33 +data$st34 +data$st35 +data$st36 +data$st37 +data$st38 +data$st39 +data$st40 +data$st41 + data$st42 +data$st43 +data$st44 +data$st45)

data$SEF= data$SEF/19

data = date\_curatate\_C4

#Nota la statistica (anul 2) este

#in relatie pozitiva cu nota de la probabilitati (anul 1)

#in relatie pozitiva cu cu math self-efficacy

#in relatie pozitiva cu cu statistical self-efficacy

#in relatie pozitiva cu school engagement

#in relatie negativa cu conspiracy ideation

colnames(data)

colnames(data)[5:6] = c("nota\_prob", "nota\_stats")

plot(data$nota\_stats, data$nota\_prob)

plot(data$nota\_stats, data$mses)

plot(data$nota\_stats, data$sses)

plot(data$nota\_stats, data$school\_engagement)

plot(data$nota\_stats, data$conspiracy)

m1 = lm(nota\_stats ~ nota\_prob + mses + sses + school\_engagement + conspiracy, data = data)

summary(m1)

par(mfrow=c(2,2))

plot(m1)

par(mfrow=c(1,1))

install.packages("car")

library(car)

vif(m1)

vif\_values = vif(m1)

barplot(vif\_values, main = "VIF Values",col = 'green',ylim = c(0.0,8.0))

bad\_vif <- 5.0

abline(h = bad\_vif, col = 'red')

colnames(data)

colnames(data)[7] = "profil\_liceu"

colnames(data)[8] = "prezent"

colnames(data)[9] = "interes\_analiza"

colnames(data)[10] = "cat\_de\_dificil\_analiza"

colnames(data)[11] = "intentie\_analiza"

boxplot(data$nota\_stats ~ data$profil\_liceu)

boxplot(data$nota\_stats ~ data$prezent)

plot(data$nota\_stats, data$interes\_analiza)

plot(data$nota\_stats, data$cat\_de\_dificil\_analiza)

boxplot(data$nota\_stats ~ data$intentie\_analiza)

m1 = lm(nota\_stats ~ profil\_liceu + prezent + interes\_analiza + cat\_de\_dificil\_analiza , data = data)

summary(m1)

par(mfrow=c(2,2))

plot(m1)

par(mfrow=c(1,1))

install.packages("car")

library(car)

vif(m1)

vif\_values = vif(m1)[,1]

barplot(vif\_values, main = "VIF Values",col = 'green',ylim = c(0.0,8.0))

bad\_vif <- 5.0

abline(h = bad\_vif, col = 'red')

hist(data$SEF)

summary(data$CRT)

hist(data$CRT)

barplot(table(data$CRT))

#Sters coloane

data = subset(data, select = -c(1:5))

#Ipoteze de cercetare

#Ipoteza 1: Există o relație pozitivă între Math self-efficacy și Statistics self-efficacy.

#Ipoteza 2: Există o relație negativă între Statistics self-efficacy și Reziliență

#Ipoteza 3: Există o relație pozitivă între Atitudinea față de Statistică și Analiza Datelor și SATS (Survey of Attitudes Toward Statistics)

#Ipoteza 4: Există o relație pozitivă între Numeracy și Math self-efficacy

#Ipoteza 5: Există o relație pozitivă Optimism – the life orientation test și Reziliență

data = date\_curatate\_C4

#Statistica descriptiva

# Ipoteza 1: MSE vs. SSE

# Math self-efficacy

summary(data$MSE)

sd(data$MSE)

hist(data$MSE, main = "Math Self-Efficacy", xlab = "Scor MSE", ylab = "Număr studenți", col = "skyblue", ylim = c(0, 15))

ks.test(data$MSE, "pnorm", mean(data$MSE), sd(data$MSE)) # Test normalitate

# Statistics self-efficacy

summary(data$SSE)

sd(data$SSE)

hist(data$SSE, main = "Statistics Self-Efficacy", xlab = "Scor SSE", ylab = "Număr studenți", col = "lightgreen", ylim = c(0, 15))

ks.test(data$SSE, "pnorm", mean(data$SSE), sd(data$SSE)) # Test normalitate

# Ipoteza 1: Există o relație pozitivă între Math self-efficacy și Statistics self-efficacy

cor.test(data$MSE, data$SSE, method = "spearman")

# Ipoteza 2: SSE vs. Reziliență

# Statistics self-efficacy

summary(data$SSE)

sd(data$SSE)

hist(data$SSE, main = "Statistics Self-Efficacy", xlab = "Scor SSE", col = "lightgreen", ylim = c(0, 15))

ks.test(data$SSE, "pnorm", mean(data$SSE), sd(data$SSE))

# Reziliență

summary(data$rezilienta)

sd(data$rezilienta)

hist(data$rezilienta, main = "Reziliență", xlab = "Scor reziliență", col = "orange", ylim = c(0, 15))

ks.test(data$rezilienta, "pnorm", mean(data$rezilienta), sd(data$rezilienta))

# Ipoteza 2: Relație negativă între SSE și Reziliență

cor.test(data$SSE, data$rezilienta, method = "spearman")

# Ipoteza 3: Atitudine vs. SATS

# Atitudine față de analiza datelor

summary(data$at\_stat)

table(data$at\_stat)

hist(data$at\_stat, main = "Atitudine față de analiza datelor", xlab = "Scor atitudine", col = "lightblue", ylim = c(0, 15))

# SATS

summary(data$SATS)

sd(data$SATS)

hist(data$SATS, main = "SATS", xlab = "Scor SATS", col = "pink", ylim = c(0, 15))

ks.test(data$SATS, "pnorm", mean(data$SATS), sd(data$SATS))

# Ipoteza 3: Relație pozitivă între atitudine și SATS

cor.test(data$at\_stat, data$SATS, method = "spearman")

# Ipoteza 4: Numeracy vs. MSE

# Numeracy

summary(data$numeracy)

sd(data$numeracy)

hist(data$numeracy, main = "Numeracy", xlab = "Scor numeracy", col = "lightgoldenrod", ylim = c(0, 15))

ks.test(data$numeracy, "pnorm", mean(data$numeracy), sd(data$numeracy))

# MSE

summary(data$MSE)

sd(data$MSE)

hist(data$MSE, main = "Math Self-Efficacy", xlab = "Scor MSE", col = "skyblue", ylim = c(0, 15))

ks.test(data$MSE, "pnorm", mean(data$MSE), sd(data$MSE))

# Ipoteza 4: Relație pozitivă între numeracy și MSE

cor.test(data$numeracy, data$MSE, method = "spearman")

# Ipoteza 5: Optimism vs. Reziliență

# Optimism

summary(data$optimism)

sd(data$optimism)

hist(data$optimism, main = "Optimism", xlab = "Scor optimism", col = "lightyellow", ylim = c(0, 15))

ks.test(data$optimism, "pnorm", mean(data$optimism), sd(data$optimism))

# Reziliență

summary(data$rezilienta)

sd(data$rezilienta)

hist(data$rezilienta, main = "Reziliență", xlab = "Scor reziliență", col = "orange", ylim = c(0, 15))

ks.test(data$rezilienta, "pnorm", mean(data$rezilienta), sd(data$rezilienta))

# Ipoteza 5: Relație pozitivă între optimism și reziliență

cor.test(data$optimism, data$rezilienta, method = "spearman")

#Scriere set date

write.csv(data, file ="date\_prelucrate")

library(openxlsx)

write.xlsx(data,"date\_prelucrare\_excel.xlsx")

codul R