UNIVERSITATEA POLITEHNICA BUCUREȘTI
FACULTATEA DE ȘTIINȚE APLICATE
AN III, GRUPA 1331

SESIUNEA DE COMUNICĂRI ȘTIINȚIFICE

UTILIZĂRI ALE REGRESIEI ÎN PROCESELE EDUCAȚIONALE

Coordonatori,
Prof. univ. dr. **TÂRCOLEA CONSTANTIN**
Prof. univ. dr. **PARIS ADRIAN STERE**

Studenți,
CRISTESCU VLAD
SÂIA THEODORA

București, 2015

CUPRINS

I. INTRODUCERE

Regresia multiplă este o metodă de predicție a valorilor unei variabile dependente pornind de la valorile mai multor variabile independente. În psihologie situația cea mai întâlnită este aceea a examenelor de selecție. În acest caz avem un set de variabile independente (numite și "predictori"), care sunt scoruri la diferite teste utilizate și o variabilă dependentă (numită și "criteriu"), ale cărei valori vrem să le estimăm pornind de la relațiile acesteia cu toate variabilele independente. În esență, regresia multiplă este o procedură similară regresiei simple. Așa cum regresia simplă se bazează pe corelația dintre două variabile, regresia multiplă se bazează pe corelația multiplă dintre variabilele implicate. Dacă în cazul regresiei simple căutăm o linie care să aproximeze cel mai bine distribuția punctelor de intersecție pentru două variabile, în regresia multiplă căutăm o linie care să aproximeze cel mai bine tendința norului de puncte al unei distribuții cu mai multe variabile simultan.

Formula de mai jos exprimă ecuația dreptei de regresie simplă:

$$Y' = a_{xy} + b_{xy} * X$$

Unde

Y' este valoarea estimată

a_{xy} este punctul de origine al liniei (valoarea lui Y' pentru X=0) și este o expresie a erorii de estimare (valorile reziduale)

b_{xz} este coeficientul care dă unghiul de înclinare a liniei (panta)

X este valoarea variabilei predictor

Ecuația de regresie multiplă va fi una similară celei de mai sus, cu singura deosebire că vom avea mai mulți coeficienți b, sau, în terminologia consacrată pentru regresia multiplă, beta (β). În plus, aceștia vor fi calculati pe baza coeficientului de corelație partială, după ce a fost eliminată influența pe care o exercită variabilele introduse anterior în ecuație.

$$Y' = a_i + b_1 * X_1 + b_2 * X_2 + b_3 * X_3 + ... + b_k * X_k$$

Unde

Y' este valoarea estimată pentru variabila criteriu (dependentă)

a_i este punctul de origine al liniei

b_1, b_2, b_3... b_k sunt coeficienții beta pentru cele k variabile predictor

X_1, X_2, X_3.... X_k sunt valorile celor k variabile predictor

1. **FORMULE UTILIZATE**

a) **Multiple R** – coeficientul multiplu de corelaţie, se determină ca fiind radical din R^2

b) **R Square** – coeficientul de determinare (este egal cu pătratul coeficientului de corelaţie multiplă). Poate fi gândit, exprimat procentual, drept proporţia din variaţia variabilei dependente explicată de variaţia variabilelor independente.

$$R^2 = \mathbf{c}^\top R_{xx}^{-1}\mathbf{c}, \quad \mathbf{c} = (r_{x_1y}, r_{x_2y}, \ldots, r_{x_Ny})^\top,$$

$$R_{xx} = \begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} & \cdots & r_{x_1x_N} \\ r_{x_2x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ r_{x_Nx_1} & \cdots & & r_{x_Nx_N} \end{pmatrix}.$$

c) **Adjusted R Square** – valoarea corectată a coeficientului de determinare. Este introdusă pentru a contracara (parţial) efectul creşterii mecanice a lui R^2 o dată cu numărul variabilelor independente.

$$\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-p-1} = R^2-(1-R^2)\frac{p}{n-p-1}$$

$$\bar{R}^2 = 1-\frac{SS_{\text{res}}/df_e}{SS_{\text{tot}}/df_t}$$

$$R^2 = 1-\frac{VAR_{\text{res}}}{VAR_{\text{tot}}}, \quad VAR_{\text{res}} = SS_{\text{res}}/n, \; VAR_{\text{tot}} = SS_{\text{tot}}/n$$

d) **Standard Error** – eroarea standard a estimaţiei. Se calculează ca abaterea standard a reziduurilor (pentru numărul gradelor de libertate utilizat se va vedea tabloul ANOVA, în continuare) şi este estimaţia abaterii standard a erorilor ε (în ipoteza normalităţii acestora).

$$\text{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

2.ASPECTE CHEIE ÎN FUNDAMENTAREA REGRESIEI MULTIPLE

(1) În cazul regresiei simple, linia de regresie "caută" cea mai bună traiectorie pentru a minimiza eroarea de estimare. Aceasta este definită printr-o metodă care asigură cea mai mică sumă a pătratelor distanțelor dintre variabila "predictor" și variabila "criteriu". În mod natural, acest deziderat este asigurat de mărimea coeficientului de corelație Pearson dintre cele două variabile. Cu cât corelația este mai mare, cu atât norul de puncte se apropie mai mult de linia de regresie, la limită, pentru o corelație de 1, punctele respective se plasează chiar pe dreapta de regresie. Situația se prezintă în mod similar și in cazul regresiei multiple. Doar că de data aceasta nu ne bazăm pe corelația simplă, dintre două variabile ci pe corelația

multiplă, dintre mai multe variabile, simbolizată prin litera ***R***. Corelația multiplă este similară corelației Pearson și ne spune câtă informație cu privire la o variabilă este conținută în combinația simultană a mai multor variabile cu care se află în asociere. Mai mult, la fel ca și în cazul corelației simple, avem și pentru corelația multiplă un coeficient de determinare (R^2) care are o interpretare similară: procentul de variație din variabila dependentă determinat de variația simultană a variabilelor independente. Semnificația lui *R* este calculată cu ajutorul unui test de varianță (F)

(2) Un alt aspect important contextul regresiei multiple este **multicoliniaritatea**. Acesta este un concept opus ortogonalității și exprimă nivelul corelației dintre variabilele independente. Informația împartașită în comun de variabilele independente reduce contribuția lor la explicarea variației variabilei dependente. Cu alte cuvinte, cu cât acestea corelează mai intens între ele cu atât corelația multiplă cu variabila dependentă (criteriu) este mai mică. În plus, multicoliniaritatea amplifică variabilitatea coeficienților de regresie, fapt care are ca efect o imprecizie mai mare a predicției. Din acest motiv, analiza de regresie trebuie precedată de evaluarea multicoliniarității. Una dintre metode este aceea de a analiza matricea de intercorelații dintre variabilele independente. Corelațiile mari sunt un indicator al liniarității. În principiu, variabilele independente a căror corelație este mai mare de 0.1 ridică problema multicoliniarității. O altă metodă este analiza "toleranței", o opțiune oferită de programele de prelucrări statistice. "Toleranța" este o măsură specifică pentru coliniaritate care ia valori între 0 și 1. Valorile apropiate de 0 sunt un semn al coliniarității. Variabilele pentru care "toleranța" este mai mică de 0.1 ridică o problemă de coliniaritate care ar trebui rezolvată. Principalele soluții posibile în legatură cu variabilele cu probleme de coliniaritate sunt două: eliminarea lor sau, combinarea lor, din moment ce aduc același tip de informație (aceasta în cazul în care corelația lor este de 0.80 sau mai mare).

(3) Ecuația de regresie multiplă are drept finalitate predicția variabilei criteriu. Verificarea potențialului real de predicție este ceea ce se numește validarea ecuației de regresie. Este evident că **modelul de validare** prezintă o importanță aparte. Coeficientul de corelație multiplă (*R*) are o valoare maximă pe eșantionul pe care a fost calculată ecuația de regresie. Dacă nivelul corelației scade dramatic pe alt eșantion, atunci ecuația de regresie nu prezintă utilitatea care a fost estimată. Obținerea unei ecuații sigure ține în mod cert de raportul (15/1) între volumul eșantionului (N) și numărul variabilelor predictor (k). O altă recomandare sugerează utilizarea unui eșantion $N \geq 50+8k$ pentru testarea corelației multiple și $N \geq 104+k$, pentru testarea predictorilor individuali. Evaluarea validității se poate face fie într-o procedură decalată în timp, pe un alt eșantion extras din aceeași populatie, fie prin utilizarea simultană a două eșantioane, unul pentru calcularea ecuației de regresie, altul pentru validarea acesteia. În ambele cazuri se va urmări respectarea criteriilor de constituire a eșantionului enunțate mai sus.

(4) Ultimul aspect care trebuie luat în considerare este efectul valorilor extreme asupra ecuației de regresie, care poate fi considerabil. Uneori chiar și una sau două valori excesive pot influența analiza de regresie. De aceea aceste valori vor fi identificate și tratate corespunzător înaintea calculării ecuației de regresie multiplă.

II. NOȚIUNI TEORETICE

1. CLASIFICARE

O importantă deosebire o reprezintă **alegerea modelului de analiză** care să permită selectarea unui set de predictori având maximum de putere de predicție asupra variabilei criteriu. Scopul nu este acela de a aduna informație de la toate variabilele disponibile ci doar de la acelea care aduc contribuția cea mai consistentă. O primă recomandare, cu caracter preliminar, este aceea de a avea în vedere un anumit raport între numărul de subiecți și numărul variabilelor independente. Acest raport este cifrat la valoarea 15/1, adică pentru un eșantion de 150 de subiecți se poate miza pe cel mult 10 variabile independente. După ce setul de variabile predictor a fost fixat, se va trece la adoptarea uneia dintre metodele de introducere a acestora în ecuația de regresie:

- *Regresia multiplă standard*. Toate variabilele predictor sunt incluse în ecuație, efectul fiecareia fiind evaluat după și independent de efectul tuturor celorlalte variabile introduse anterior. Fiecare variabilă independentă este evaluată numai prin prisma contribuției proprii la explicarea variabilei dependente.

- *Regresia multiplă secvențială* (numită și *regresie ierarhică*). Variabilele independente sunt introduse în ecuație într-o anumită ordine, în functie de opțiunile analistului. Atunci când acesta are motive să creadă că o anumită variabilă are o influență mai mare, o poate introduce în ecuație înaintea altora.
- *Regresia multiplă pas cu pas*. Este utilizată adesea în studii exploratorii, atunci când există un număr mare de predictori despre care nu se știe exact care este contribuția fiecareia la corelatția de ansamblu cu variabila dependentă. Exista trei variante ale acestui tip de analiză:
  - Selecția anterogradă. Toate variabilele independente sunt corelate cu variabila dependentă după care variabila care are corelația cea mai mare este introdusă prima în ecuație. Următoarea variabilă introdusă în ecuație este cea care are corelația cea mai mare, dupa ce a fost eliminat efectul variabilei anterioare. Procesul continuă până ce nivelul contribuției variabilelor independente este prea mic pentru a mai fi luat în considerare.
  - Selecția pas cu pas. Este o variantă a metodei anterioare. Diferența constă în faptul că dacă o variabilă nou introdusă are o contribuție mai consistentă asupra variabilei dependente va determina eliminarea unei variabile anterioare dar care se dovedește mai puțin predictivă.
  - Selecția retrogradă. Pasul inițial al acestei metode este acela de calculare a unei ecuații de regresie în care toate variabilele predictor sunt incluse. Ulterior, pentru fiecare variabilă predictor este efectuat un test de semnificație "F", pentru a se evalua contribuția fiecărui predictor la corelația de ansamblu. Valorile testului F sunt comparate cu o valoare limită prestabilită, variabilele care nu trec acest prag fiind eliminate din ecuație. Pe măsură ce o variabilă este eliminată, o nouă ecuație este calculată și un nou test F este efectuat pentru variabilele rămase, urmat de eventuala eliminare a unei alte variabile. Procesul continuă până când doar variabilele semnificative ramân în ecuație.

Este evident că metoda "secvențială" și cea "pas cu pas" sunt superioare metodei "standard". Între primele două diferența constă în faptul că, în cazul metodei secvențiale, decizia de selecționare a variabilelor introduse în ecuație aparține cercetătorului în timp ce în cazul metodei pas cu pas, programul este cel care face în mod automat selecția, în funcție de parametri fixați de analist.

2. OBIECTIVE DE CERCETARE SPECIFICE ANALIZEI DE REGRESIE MULTIPLĂ

Analiza de regresie multiplă este utilizabilă în situații de predicție. Un caz tipic este acela în care dorim să selectăm candidați pentru o anumită profesie pe baza performanțelor la un set de teste psihologice. Performanța profesională, măsurată prin una din metodele posibile (aprecierea pe baza de experți, apreciere interpersonală, productivitate, etc.) este variabilă printr-un criteriu (dependență). Indicatorii de performanță la teste reprezintă variabilele predictor (independente). Scopul esențial este că, o dată stabilită ecuația de regresie pentru eșantionul studiat, să putem utiliza bateria de teste pentru a face predicții de adaptare profesională în cazul altor subiecți. Este evident că o astfel de procedură este una de durată și urmăreste ceea ce se numește "validarea testelor de selecție". Într-un astfel de caz, subiectii eșantionului ar fi supuși testării psihologice înaintea angajării după care, la un interval adecvat de timp, ar urma să fie evaluați sub aspectul performanței profesionale. Ulterior, dacă rezultatele analizei de regresie justifică aceasta, rezultatele la teste vor putea fi utilizate pentru selecție.

Într-o situație de cercetare ca cea descrisă, întrebarile pe care și le pune cercetătorul, atunci când alege să introducă în ecuația de regresie toți indicatorii testelor, sunt, în mod explicit, urmatoarele:

- *Care dintre indicatorii testelor utilizate are capacitatea de predicție cea mai ridicată? Există indicatori care nu au relevanță pentru predicția performanței profesionale? Are ecuația de regresie astfel obținută o capacitate sigură de predicție?*

Dacă modelul de analiză este unul secvențial sau pas cu pas, atunci întrebările obiectivele implicite vor fi:

- *Care dintre indicatorii testelor utilizate pot fi incluse în ecuația de predicție a performanței profesionale? Are ecuația de regresie, astfel obținută, o capacitate sigură de predicție?*

3. CONDIȚII ȘI LIMITĂRI

Efectuarea analizei de regresie multiplă presupune o serie de condiții prealabile. Acestea se referă la variabile și la distribuția valorilor reziduale.

Variabilele analizate:

- trebuie să fie măsurate pe scala de interval raport, cu respectarea condițiilor de aplicare a testului de corelație (normalitatea distribuției, în special);
- sunt fixe, ele urmează a fi păstrate în orice studiu de replicare;
- vor fi măsurate fără erori, iar cazurile extreme vor fi analizate și tratate corespunzător;
- se supun unui model de corelație liniară;

Valorile reziduale (erorile de predicție):

- media valorilor reziduale în studii de replicare să fie zero;
- erorile din cazul unei variabile independente nu au nici o legatură cu erorile altei sau altor variabile independente;
- erorile nu corelează cu variabilele independente;
- varianta valorilor reziduale pe toată distribuția variabilelor independente este omogenă (homoscedasticitate);
- erorile au o distribuție normală;

Verificarea acestor condiții presupune îndeplinirea tuturor procedurilor de analiză preliminară a datelor, așa cum au fost deja prezentate anterior.

4. RAPORTAREA REZULTATELOR

În lucrare vor fi incluse cele mai importante dintre caracteristicile datelor preliminare precum și datele obtinute prin prelucrare:

- datele inițiale şi eventualele eliminări sau transformări efectuate
- indicatorii statistici descriptivi (medii, abateri standard), matricile de corelație, graficele ilustrative pentru diferitele distribuții
- coeficienții de regresie și semnificațiile lor (R^2, R^2_{adj} și gradele de libertate)
- se vor trage concluzii de ansamblu

Rezultatele studiului demonstrativ de mai sus pot fi sintetizate în felul următor (facem precizarea că datele prezentate nu au nicio legatură cu vreun studiu real pe aceasta tema, având doar o semnificație didactică):

III. IPOTEZE

1. Ipoteze statistice clasice asupra modelului de regresie simplă

Ipotezele statistice clasice asupra modelului de regresie sunt:

- **Liniaritatea modelului**. Relaţia între Y şi X este liniară.
Această ipoteză este necesară pentru estimarea parametrilor modelului;

- **Normalitatea erorilor**. Variabila ε este distribuită normal: $\varepsilon \equiv N(0, \sigma_\varepsilon^2)$;

- **Homoscedasticitatea**. Varianţele $v(\varepsilon)$ sunt constante, oricare ar fi valorile variabilei X, adică, $V(\varepsilon) = \sigma^2$;

- **Necorelarea erorilor**. Erorile sunt necorelate între ele: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$;

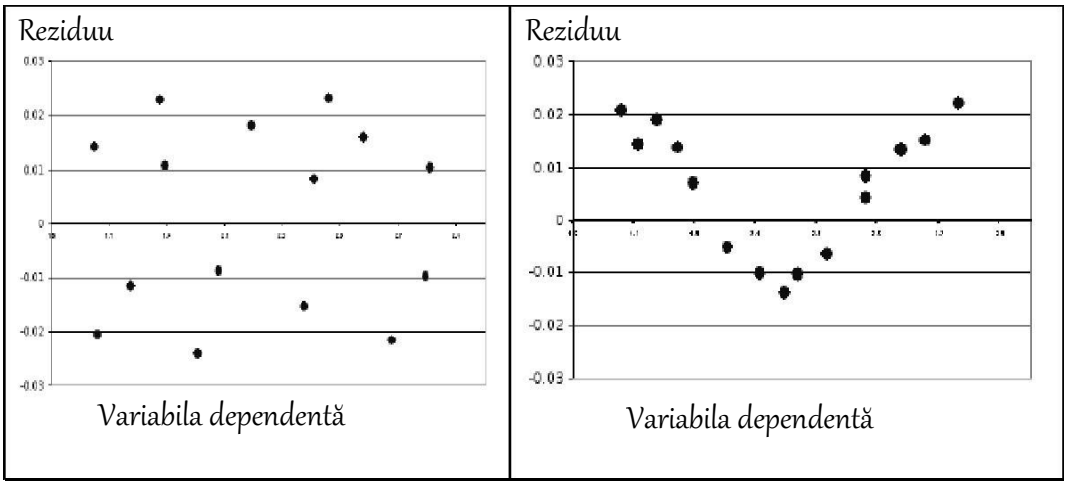
- Independenţa erorilor de valorile variabilei X. Valorile variabilei ε sunt independente de valorile variabilei explicative X, adică $\text{cov}(\varepsilon, x) = 0$.

Încălcarea ipotezelor poate afecta calitatea estimatorilor.

1.1 **Testarea liniarităţii modelului propus**

Liniaritatea relaţiei dintre variabila dependentă şi variabila independentă este importantă atât pentru acurateţea predictivă a modelului cât şi pentru validitatea coeficienţilor estimaţi. Verificarea liniarităţii se poate efectua grafic, folosind: scatterplots; diagrama reziduurilor din regresie.

Diagrama reziduurilor (reziduu = diferenţa dintre un punct observat şi punctul prezis pe dreaptă) din regresie se construieşte luând pe ordonată variabila reziduu şi pe abscisă variabila dependentă (Figura 1.1.1). Dacă reziduurile apar dispersate aleator, de o parte şi de alta a valorii zero (Figura 1.1.1.a), atunci relaţia poate fi modelată cu ajutorul regresiei liniare. Dacă reziduurile apar dispersate în blocuri deasupra sau sub valoarea zero (Figura 1.1.1.b), atunci relaţia dintre variabilele considerate nu poate fi modelată cu ajutorul regresiei liniare.



(a) (b)

Figura 1.1.4:Distribuţia reziduurilor în cazul relaţiei de tip liniar (a) şi a relaţiei de tip neliniar (b)

În cazul unor relaţii neliniare, se poate gândi la o adecvare la un model liniar, utilizând o transformare logaritmică etc., sau pot fi tratate ca atare. În exemplul considerat, distribuţia reziduurilor de regresie validează ipoteza modelului de regresie liniar, reziduurile plasându-se aleator de o parte şi de alta a valorii zero .

1.2. **Testarea ipotezei de normalitate a erorilor**

Pentru variabila aleatoare reziduu, ε, dintr-un model de regresie simplă liniară verificăm ipotezele de: normalitate, homoscedasticitate, necorelare şi independenţă a erorilor.

Ipoteza de normalitate a erorilor presupune că variabila ε urmează o lege normală de medie 0 şi varianţă σ^2: $\varepsilon_i \sim N(0, \sigma^2)$.

Efectele încălcării acestei ipoteze :

Ipoteza de normalitate a erorilor este importantă pentru stabilirea proprietăţilor estimatorilor parametrilor modelului de regresie. Dacă $\varepsilon_i \sim N(0, \sigma^2)$,atunci estimatorii parametrilor modelului de regresie urmează, de asemenea, o lege normală.

Dacă ipoteza de normalitate este încălcată, proprietăţile estimatorilor

construiţi pe baza metodei celor mai mici pătrate au doar proprietăţi asimptotice, adică necesită eşantioane sau seturi mari de date.

Verificarea acestei ipoteze implică şi testarea ipotezei că, în medie, modelul este bine specificat.

Testarea ipotezei de normalitate a erorilor se poate realiza cu ajutorul procedeelor grafice (histograma, box-plot, P-P-plot, diagrama reziduurilor) sau a procedeelor numerice (testul Kolmogorov-Smirnov, testul Jarque- Bera).

1.3. **Homoschedastic vs Heteroschedastic**

1.3.1 Definiţii

A. Homoschedasticitate

La date bivariate, variabila y prezintă homoscedasticitate dacă împrăştierea valorilor y nu depinde de x. Grafic, secţiunile verticale în diagrama de împrăştiere prezintă distribuţii similare ale norilor de puncte.

B. Heteroschedasticitate

La date bivariate, variabila y prezintă heteroscedasticitate dacă împrăştierea valorilor y depinde de x. Grafic, secţiunile verticale în diagrama de împrăştiere prezintă distribuţii diferite ale norilor de puncte.

Consecinţe ale ignorării fenomenului de heteroscedasticitate a erorilor

a) Estimatorii parametrilor din model sunt nedeplasaţi şi consistenţi
b) Estimatorii parametrilor din model nu sunt eficienţi (există estimatori care au o dispersie mai mică).
c) Estimatorii calculaţi pentru dispersia şi covarianţa parametrilor sunt deplasaţi, nu sunt consistenţi şi nu sunt eficienţi.
d) Testul t Student aplicat pentru analiza semnificaţiei estimatorilor nu este valid. Dacă dispersia erorilor şi variaţia factorului explicativ sunt pozitiv corelate, atunci dispersia corectă a parametrului a1 este subestimată, astfel încât calculele sugerează o precizie a estimării mai bună decât este în realitate.
e) Estimatorii parametrilor nu au proprietatea de maximă verosimilitate.

1.3.2. Testarea heteroscedasticităţii : Testul Goldfeld–Quandt

Se identifică, o variabilă notată Z, de care este (potenţial) legată dispersia erorilor. Se aranjează toate observaţiile din eşantionul reţinut pentru analiză în ordinea crescătoare a valorilor Zt . Eşantionul de dimensiune n se divide în două părţi de dimensiuni n1 şi n2 , după eliminarea a m observaţii (între 1/6 şi 1/5 dintre observaţii) situate la mijlocul eşantionului. Se calculează dispersia reziduurilor din modelul estimat pentru primele n1 observaţii şi dispersia reziduurilor din modelul estimat pentru ultimele n2 observaţii. Dacă raportul supraunitar al acestor dispersii este mai mic decât valoarea critică din tabelul distribuţiei teoretice F cu n1 -(k+1) şi

n2 -(k+1) grade de libertate, atunci ipoteza nulă, a lipsei heteroscedasticităţii erorilor nu este respinsă.

1.4. **Testarea ipotezei de autocorelare a erorilor**

Ipoteza necorelării erorilor $\mathrm{cov}(\varepsilon i, \varepsilon j) = 0$ presupune lipsa unei corelaţii între termenii variabilei eroare din modelul de regresie, adică eroarea asociată unei valori a variabilei dependente nu este influenţată de eroarea asociată altei valori a variabilei dependente.

Pentru testarea acestei ipoteze se pot utiliza: *testul Durbin Watson şi Runs test.*

IV. INTERPRETAREA DATELOR

1. **PROGNOZA MEDIEI PENTRU STUDENȚII DIN ANUL 4**

Pornind de la un set de date care conține notele studenților din primii trei ani, ne propunem să facem o prognoză în ceea ce privește media acestora pentru anul IV. Vom alege ca variabilă independentă , media notelor din primii trei ani iar ca variabile independente absența la cursuri și statutul pe piața muncii a studenților (notăm 1 pentru angajat și 0 pentru șomer).

Tabelul **Summary Output** prezintă valoarea raportului de corelaţie (R), valoarea raportului de determinaţie (R^2), valoarea ajustată a lui R şi eroarea standard a estimaţiei. Pentru exemplul considerat, **Summary Output** este prezentat în Tabelul 1.1.10.

Prognoza media in anul 4 pentru studentii

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,97609665
R Square	0,952764671
Adjusted R Square	0,948041138
Standard Error	0,363104878
Observations	23

98/100 din cazuri se bazeaza pe aceasta predictie

Absenta la cursuri si faptul ca un student lucreaza afecteaza nota intr-un procent de peste 90 %

Tabelul 1.1.1

Valoarea R arată dacă există sau nu o corelaţie între variabila dependentă (rezultativa y) şi variabila independentă (factoriala x). Acest indicator au valori între 0 şi 1.

Interpretarea modelului. În interpretarea modelului se foloseşte coeficientul de determinaţie, $\mathbf{R^2}$.

Raportul de determinaţie, R^2, arată proporţia variaţiei variabilei dependente explicate prin modelul de regresie şi este folosit pentru a evalua calitatea ajustării (alegerea modelului).

R^2 ia valori între 0 şi 1. Dacă R^2 este egal cu 0 sau are o valoare foarte mică, atunci modelul de regresie ales nu explică legătura dintre variabile, relaţia dintre variabila dependentă şi variabila independentă nu coincide cu modelul ales, de exemplu, liniar. Dacă R^2 este egal cu 1, atunci toate observaţiile cad pe linia de

regresie, deci, modelul de regresie explică perfect legătura dintre variabile. Ca urmare, R^2 este folosit pentru a stabili care model de regresie este cel mai bun. Această metodă de alegere a modelului de regresie potrivit este recomandată pentru modelele care nu conţin un număr mare de variabile.

Pentru exemplul considerat a rezultat o valoare R^2=0.952, ceea ce înseamnă că între media celor 3 ani(y), absenţa la cursuri(x_1) şi faptul că un student lucrează în timpul facultăţii(x_2), există o legătură liniară, directă, foarte strânsă.

Tabelul Regression ANOVA prezintă rezultatele analizei varianţei variabilei dependente sub influenţa factorului de regresie şi a factorului reziduu. Adică, prezintă informaţii asupra sumei

pătratelor abaterilor variabilei dependente, datorate modelului de regresie şi factorului reziduu, gradele de libertate, estimaţiile varianţelor datorate celor două surse de variaţie (regresie şi reziduu), raportul F şi Sig. (vezi Tabelul 1.1.2).

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	53,18790226	26,59395	201,7059	5,52929E-14
Residual	20	2,63690305	0,131845		
Total	22	55,82480531			

modelul ales este semnificat deoarece eroarea este mai mica de 5% (am stabilit nivelul de incredere 95 %)

Tabelul 1.1.2

Statistica test F se obţine ca raport între media pătratelor abaterilor datorate regresiei şi media pătratelor abaterilor datorate reziduului, calculate cu gradele de libertate corespunzătoare. Această statistică test este folosită pentru testarea modelului de regresie.

Dacă testul F ia o valoare mare, iar valoarea Sig. corespunzătoare statisticii F este mică (mai mică decât 0,05), atunci variabila independentă explică variaţia variabilei dependente şi invers.

În exemplul considerat, valoarea Sig. pentru F este mai mică decât 0,05, deci relaţia liniară dintre cele două variabile considerate este semnificativă (vezi Tabelul 1.1.2).

Pentru exemplul dat, valoarea este mai mică decât 0.05, arătând că panta dreptei de regresie este semnificativ diferit de zero şi corespunde unei legături semnificative între cele două variabile.

2. FACTORI CARE INFLUENŢEAZĂ BURSA DE MERIT

Analizăm un set de date care conţine notele studenţilor din primii trei ani şi ne propunem să verificăm factorii care influenţează nota acestora. Vom alege ca variabilă independentă , media notelor din primii trei ani iar ca variabile independente timp liber şi timp de odihnă. Tabelul **Summary Output** prezintă valoarea raportului de corelaţie (R), valoarea raportului de determinaţie (R^2), valoarea ajustată a lui R şi eroarea standard a estimaţiei. Pentru exemplul considerat, **Summary Output** este prezentat în Tabelul 1.2.1, de mai jos.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,9995
R Square	0,9989
Adjusted R Square	0,9979
Standard Error	0,0079
Observations	5

Tabelul 1.2.1

Valoarea R arată dacă există sau nu o corelaţie între variabila dependentă (rezultativa y) şi variabila independentă (factoriala x). Acest indicator au valori între 0 şi 1.

Interpretarea modelului. În interpretarea modelului se foloseşte coeficientul de determinaţie, $\mathbf{R^2}$.

Raportul de determinaţie, R^2, arată proporţia variaţiei variabilei dependente explicate prin modelul de regresie şi este folosit pentru a evalua calitatea ajustării (alegerea modelului).

R^2 ia valori între 0 şi 1. Dacă R^2 este egal cu 0 sau are o valoare foarte mică, atunci modelul de regresie ales nu explică legătura dintre variabile, relaţia dintre variabila dependentă şi variabila independentă nu coincide cu modelul ales, de exemplu, liniar. Dacă R^2 este egal cu 1, atunci toate observaţiile cad pe linia de regresie, deci, modelul de regresie explică perfect legătura dintre variabile. Ca urmare, R^2 este folosit pentru a stabili care model de regresie este cel mai bun. Această metodă de alegere a modelului de regresie potrivit este recomandată pentru modelele care nu conţin un număr mare de variabile.

Pentru exemplul considerat a rezultat o valoare R^2=0.999 , ceea ce înseamnă că între media celor 3 ani(y), timpul liber (x_1) şi timpul de odihnă (x_2), există o legătură liniară, directă, foarte strânsă.

Tabelul Regression ANOVA prezintă rezultatele analizei varianţei variabilei dependente sub influenţa factorului de regresie şi a factorului reziduu. Adică, prezintă informaţii asupra sumei

pătratelor abaterilor variabilei dependente, datorate modelului de regresie şi factorului reziduu, gradele de libertate, estimaţiile varianţelor datorate celor două surse de variaţie (regresie şi reziduu), raportul F şi Sig. (vezi Tabelul 1.2.2).

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0,1162	0,058122093	932,173	0,00107
Residual	2	0,0001	6,23512E-05		
Total	4	0,1164			

Tabelul 1.2.2

Statistica test F se obţine ca raport între media pătratelor abaterilor datorate regresiei şi media pătratelor abaterilor datorate reziduului, calculate cu gradele de libertate corespunzătoare. Această statistică test este folosită pentru testarea modelului de regresie.

Dacă testul F ia o valoare mare, iar valoarea Sig. corespunzătoare statisticii F este mică (mai mică decât 0,05), atunci variabila independentă explică variaţia variabilei dependente şi invers.

În exemplul considerat, valoarea Sig. pentru F este mai mică decât 0,05, deci relaţia liniară dintre cele două variabile considerate este semnificativă (vezi Tabelul 1.2.2).

Pentru exemplul dat, valoarea este mai mică decât 0.05, arătând că panta dreptei de regresie este semnificativ diferit de zero şi corespunde unei legături semnificative între cele două variabile.

3. FACTORI CARE INFLUENȚEAZĂ BURSA SOCIALĂ

Tabelul **Summary Output** prezintă valoarea raportului de corelaţie (R), valoarea raportului de determinaţie (R^2), valoarea ajustată a lui R şi eroarea standard a estimaţiei. Pentru exemplul considerat, **Summary Output** este prezentat în Tabelul 1.1.10.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0,99952792
R Square	0,999056063
Adjusted R Sq	0,997168189
Standard Error	0,009493557
Observations	4

Tabelul 1.3.1

Valoarea R arată dacă există sau nu o corelaţie între variabila dependentă (rezultativa y) şi variabila independentă (factoriala x). Acest indicator au valori între 0 şi 1.

Interpretarea modelului. În interpretarea modelului se foloseşte coeficientul de determinaţie, $\mathbf{R^2}$.

Raportul de determinaţie, R^2, arată proporţia variaţiei variabilei dependente explicate prin modelul de regresie şi este folosit pentru a evalua calitatea ajustării (alegerea modelului).

R^2 ia valori între 0 şi 1. Dacă R^2 este egal cu 0 sau are o valoare foarte mică, atunci modelul de regresie ales nu explică legătura dintre variabile, relaţia dintre variabila dependentă şi variabila independentă nu coincide cu modelul ales, de exemplu, liniar. Dacă R^2 este egal cu 1, atunci toate observaţiile cad pe linia de regresie, deci, modelul de regresie explică perfect legătura dintre variabile. Ca urmare, R^2 este folosit pentru a stabili care model de regresie este cel mai bun. Această metodă de alegere a modelului de regresie potrivit este recomandată pentru modelele care nu conţin un număr mare de variabile.

Pentru exemplul considerat a rezultat o valoare R^2=0.952, ceea ce înseamnă că între media celor 3 ani(y), absența la cursuri(x_1) şi faptul că un student lucrează în timpul facultății(x_2), există o legătură liniară, directă, foarte strânsă.

Tabelul Regression ANOVA prezintă rezultatele analizei varianţei variabilei dependente sub influenţa factorului de regresie şi a factorului reziduu. Adică, prezintă informaţii asupra sumei

pătratelor abaterilor variabilei dependente, datorate modelului de regresie şi factorului reziduu, gradele de libertate, estimaţiile varianţelor datorate celor două surse de variaţie (regresie şi reziduu), raportul F şi Sig. (vezi Tabelul 1.3.2).

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0,095390428	0,047695	529,1964	0,030724
Residual	1	9,01276E-05	9,01E-05		
Total	3	0,095480556			

Tabelul 1.3.2

Statistica test F se obţine ca raport între media pătratelor abaterilor datorate regresiei şi media pătratelor abaterilor datorate reziduului, calculate cu gradele de libertate corespunzătoare. Această statistică test este folosită pentru testarea modelului de regresie.

Dacă testul F ia o valoare mare, iar valoarea Sig. corespunzătoare statisticii F este mică (mai mică decât 0,05), atunci variabila independentă explică variaţia variabilei dependente şi invers.

În exemplul considerat, valoarea Sig. pentru F este mai mică decât 0,05, deci relaţia liniară dintre cele două variabile considerate este semnificativă (vezi Tabelul 1.3.2).

Pentru exemplul dat, valoarea este mai mică decât 0.05, arătând că panta dreptei de regresie este semnificativ diferit de zero şi corespunde unei legături semnificative între cele două variabile.

CONCLUZII

În concluzie, setul de date analizate s-a dovedit a fi unul corect atât din punct de vedere al verificării corelației coeficienților cât și al respectării pragului de încredere stabilit.

Pentru prognoza notei pentru studenții din anul 4, s-a dovedit că studenții care au avut rezultate bune până în prezent vor continua să le mențină, urmând un trend care urmează ecuația dreptei liniare, dovedind astfel încă o dată faptul că am analizat o regresie multiplă liniară.

În cazul factorilor care influențează bursa de merit, s-a validat ipoteza în care atât timpul liber cât și cel de odihnă sunt factori care influențează obținerea notelor.

Despre factori care influențează bursa socială, putem spune că un timp mai îndelungat folosit pentru a sta la calculator și un număr mare de restanțe vor determina o menținerea unei prestații școlară scăzută, venită și pe fondul unor lipsuri materiale prezente în familia celor în cauză.

În această lucrare, ne-am propus să prezentăm atât factorii care influențează o notă mai bună cât și factorii care vor contribui la obținerea unei note mai slabe. Putem spune, în urma celor prezentate că o prezență ridicată la cursuri, suficient timp pentru odihnă și timp liber utilizat pentru relaxare, vor contribui la creșterea sau menținerea unor note mai bune, în timp ce un număr mai mare de ore petrecute la calculator și un număr mare de restanțe vor determina scăderea notei.

VI. ANEXĂ
NOTAȚII EXCEL

a) Primul tabel
Multiple R – coeficientul multiplu de corelaţie.
R Square – coeficientul de determinare (este egal cu pătratul coeficientului de corelaţie multiplă). Poate fi gândit, exprimat procentual, drept proporţia din variaţia variabilei dependente explicată de variaţia variabilelor independente.
Adjusted R Square – valoarea corectată a coeficientului de determinare. Este introdusă pentru a contracara (parţial) efectul creşterii mecanice a lui R^2 o dată cu numărul variabilelor independente.
Standard Error – eroarea standard a estimaţiei. Se calculează ca abaterea standard a reziduurilor (pentru numărul gradelor de libertate utilizat se va vedea tabloul ANOVA, în continuare) şi este estimaţia abaterii standard a erorilor ε (în ipoteza normalităţii acestora).
Observations – numărul de observaţii din eşantion.

b) Al doilea tabel de rezultate cuprinde tabloul de analiză a varianţei asociat regresiei estimate.
Coloanele acestui tablou au semnificaţiile uzuale într-un tablou ANOVA:
Sursa de variaţie – arată descompunerea variaţiei totale în variaţia explicată de regresie şi cea reziduală (neexplicată).
df – numărul gradelor de libertate: 2 = p – 1, 20 = n – p, 22 = n – 1, unde p = 3 este numărul parametrilor modelului (două variabile X plus termenul liber) iar n = 23 este numărul de observaţii.
SS – sumele de pătrate potrivit descompunerii

Suma globală de pătrate	=	Suma de pătrate datorată regresiei	+	Suma de pătrate reziduală

MS – media sumelor de pătrate: SS împărţită la numărul respectiv de grade de libertate. Valoarea de pe linia a doua (Residual) este estimaţia dispersiei pentru repartiţia erorilor şi este pătratul erorii standard a estimaţiei.
F – valoarea statisticii F pentru testul caracterizat de
H0 : $\alpha 1 = \alpha 2 = \alpha 3 = 0$
H1 : există cel puţin un coeficient αi diferit de zero.

Acest test se referă la ansamblul variabilelor independente (este de remarcat că H0 nu se extinde şi asupra termenului liber). Datorită înţelesului ipotezei nule, se consideră că prin acest test se verifică semnificaţia întregii regresii.
Significance F – este probabilitatea critică unilaterală. Dacă valoarea afişată este mai mică decât pragul de semnificaţie fixat, atunci se respinge ipoteza nulă în favoarea ipotezei alternative.

c) Al treilea tablou de rezultate conţine valorile estimate pentru coeficienţii modelului, precum şi statisticile necesare verificării ipotezelor uzuale asupra coeficienţilor. De remarcat că, spre deosebire de testul F, testele asupra coeficienţilor sunt individuale.
Liniile tabelului se referă la variabilele din model, incluzând şi termenul liber. Coloanele tabelului sunt următoarele:

(prima coloană) – sunt afişate denumirile existente în tabloul de date sau create automat pentru variabilele independente implicate. Intercept este denumirea pentru termenul liber (constant) al modelului.

Coefficients – conţine valorile estimate ale coeficienţilor. Din valorile afişate rezultă că modelul estimat în exemplu este

$$Y = a + b1*x1 + b2*x2$$

În ipotezele distribuţionale ale modelului liniar, valorile calculate ale coeficienţilor provin din repartiţii normale, fiind astfel posibile verificări statistice ale coeficienţilor.

Standard Error – eroarea standard a coeficientului (abaterea standard a repartiţiei coeficientului).

t Stat – statistica t pentru verificarea ipotezei $H0 : \alpha i = 0$ contra ipotezei alternative $H1 : \alpha i \neq 0$. În condiţiile ipotezei nule se demonstrează că raportul dintre coeficient şi eroarea standard a coeficientului urmează o repartiţie Student cu (n – p) grade de libertate. Acest raport este tocmai valoarea raportată drept t Stat.

P-value – probabilitatea critică bilaterală a testului t cu ipotezele precizate la t Stat.

Lower 95%, Upper 95% – limitele inferioară şi superioară ale intervalului de încredere pentru parametrul respectiv.

Se poate observa că ultimul interval cuprinde şi valoarea zero, prin urmare se regăseşte concluzia privind nerespingerea ipotezei nule $H0 : \alpha 3 = 0$.

BIBLIOGRAFIE

- Design and Analysis of Experiments – Douglas Montgomery, 2013, SAS Institute Inc., Cary, North Carolina, USA
- Applied Multiple Regression Correlation Analysis for the Behavioral Sciences - Jacob Cohen, 3rd Edition, Lawrence Erlbaum Associates, Publishers, 2002
- http://www.scritub.com/stiinta/fizica/Regresia-multipla64684.php
- https://www.kendallhunt.com/uploadedFiles/Kendall_Hunt/Content/Higher_Education/Uploads/Gibson-Dillard_Section%209.5.pdf
- http://www.gbv.de/dms/ilmenau/toc/348809573.PDF
- https://www.scribd.com/doc/209600011/Aplicatie-Regresie-multipla
- http://thor.info.uaic.ro/~val/statistica/StatGloss.htm