II.B

ANALIZA CORELAȚIEI ȘI REGRESIEI LINIARE. APLICAȚII CU SPECIFIC DE MEDICINĂ DENTARĂ ÎN EXCEL

În această lucrare ne propunem să analizăm asocierile liniare dintre datele numerice ale 30 de pacienţi aflaţi în atenţia centrului de diabet şi boli de nutriție şi să prelucrăm datele pacienților dintr-un studiu stomatologic, care au avut restaurări protetice pe diferiți dinți.

1. Introducere

Coeficientul de corelaţie sau coeficientul Pearson este un indice numeric ce dă o măsură a relaţiei dintre două variabile cantitative continue sau discrete, este un indicator independent de unităţile de măsură ale celor două variabile numerice studiate.

Dintre proprietăţile coeficientului de corelaţie menţionăm:

- Coeficientul de corelaţie este un număr cuprins între -1 şi 1.
- Cu cât coeficientul de corelaţie se apropie de 1 în valoare absolută cu atât mai mult "intensitatea" relaţiei liniare între cele două variabile va fi mai mare. Când r este pozitiv relaţia între variabilele X şi Y este "pozitivă", corelația se zice directă, adică o creştere a lui X determină în general o creştere a lui Y.

Când r < 0 relaţia între cele două variabile este "negativă", corelația se zice inversă, adică o creştere a lui X are în general ca şi consecinţă o diminuare a lui Y.

Colton (1974) sugerează următoarele reguli empirice privind interpretarea coeficientului de corelaţie:

- un coeficient de corelaţie ce are valori de la -0,25 la 0,25 sugerează o corelaţie slabă sau nulă,
- un coeficient de corelaţie de la 0,25 la 0,50 (sau de la -0,25 la -0,50) înseamnă un grad de asociere acceptabil, medie

- un coeficient de corelaţie de la 0,5 la 0,75 (sau de la -0,5 la -0,75) înseamnă o corelaţie moderată spre bună
- un coeficient de corelaţie mai mare decât 0,75 (sau mai mic decât -0,75) înseamnă o foarte bună asociere sau corelaţie puternică.

2. Aplicații. Analiza regresiei și corelației

Introduceţi în Excel următorul tabel de date:

VARSTA	GREUTATE	INALTIME	TAS	TAD	GLICEMIE
59	95	1,70	140	100	100
68	85	1,56	150	100	103
70	54	1,57	160	80	99
29	74	1,69	110	60	84
29	61	1,59	120	70	82
52	82	1,89	120	80	72
43	67	1,64	130	80	89
47	86	1,72	140	100	80
30	69	1,57	110	50	76
47	107	1,80	130	90	108
41	84	1,83	110	80	85
41	104	1,75	110	70	122
60	60	1,58	120	70	80
67	74	1,63	160	90	93
73	61	1,53	160	80	95
68	77	1,72	140	80	104
49	109	1,69	160	100	89
50	88	1,66	130	90	123
40	64	1,79	120	80	68
48	78	1,73	140	80	93
38	60	1,65	90	40	73
44	108	1,71	140	100	89
26	75	1,73	110	60	89
47	87	1,80	120	80	87
26	96	1,76	130	90	81
29	83	1,75	120	70	85
33	83	1,75	100	70	71
41	81	1,67	120	90	90
52	73	1,68	140	100	88
43	90	1,63	100	70	82

Exerciţiul 1

Calculaţi coeficientul de corelaţie Pearson dintre *Varsta* şi *Greutate* cu ajutorul funcţiei CORREL.

Indicaţii

1. Copiaţi *Vârsta* şi *Greutatea* într-o altă foaie de lucru (Sheet 2, de exemplu).

2. Introduceţi în Sheet 2 următorul tabel:

	Coeficientul de corelatie Pearson
Varsta si Greutate	

Figura 1. Tabelul creat

şi selectaţi celula unde vom calcula coeficientul de corelaţie.

3. Din meniul **Insert** alegeţi opţiunea **Function** şi selectaţi funcţia **Correl**.

4. În rubrica **Array1** introduceţi referinţele domeniului unde se găseşte variabila Varsta: **A2:A31**. In rubrica **Array2** introduceţi referinţele domeniului unde se găseşte variabila Greutate: **B2:B31**. Click pe **OK** (figura 2).

5. Tabelul de la pct. 2 se va completa cu valoarea coeficientului, -0,14 (dacă alegem o reprezentare cu 2 zecimale). Acest coeficient poate fi interpretat astfel: între *Varstă* şi *Greutate* corelaţia este inversă şi extrem de slabă. Pentru a determina semnificaţia acestei asocieri, e necesar să calculăm valoarea probabilităţii (p) ca ipoteza de nul H_0: ”nu există corelație liniară între varstă și greutate” să se accepte.

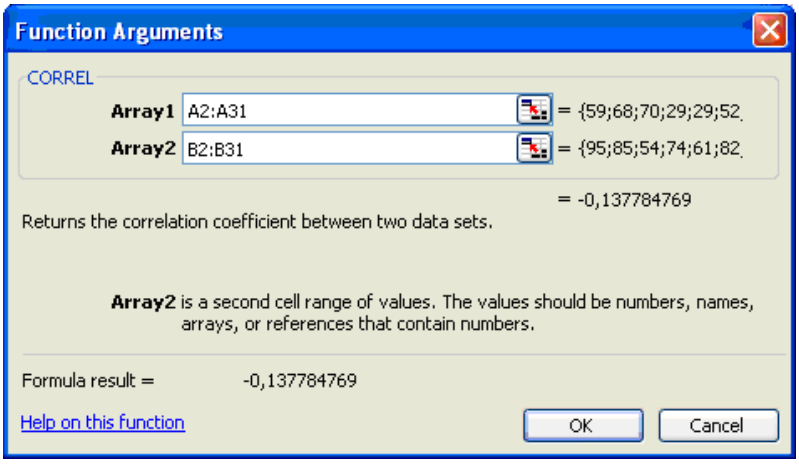

Figura 2. Funcția CORREL

Exerciţiul 2

Calculaţi **indicele de masă corporală** IMC după formula $IMC = \frac{Greutate}{Inaltime^2}$

Indicaţii

Introduceţi formula adaptată la Excel **IMC=Greutate/Inaltime^2** în coloana D. Calculaţi pentru primul pacient şi apoi umpleţi coloana folosind Fill-Down.

Exerciţiul 3

Calculaţi matricea de corelaţii a variabilelor: *Varsta, Greutate, IMC, TAS, TAD şi Glicemie.*

Indicaţii

1. Copiaţi variabilele din listă într-o foaie nouă. Atenţie: IMC se copiază cu **Paste Special – Values**

2. Alegeţi opţiunea **Data Analysis** din meniul **Tools - d**in fereastra care apare alegeți **Correlation**, apoi **Ok**.

3. La **Input Range** selectaţi domeniul unde se găsesc valorile variabilelor Varsta, Greutate, IMC, TAS, TAD şi Glicemie **A1:F31**.

4. **Grouped by**: se va selecta **Columns**.

5. Selectăm **Labels in first row**.

6. **Opţiunile Output** se referă la locul amplasării coeficientului de corelaţie. Selectaţi opţiunea **Output Range**, iar în rubrica de lângă introduceţi J2. Matricea de corelaţii va fi afişată începând cu celula J2.

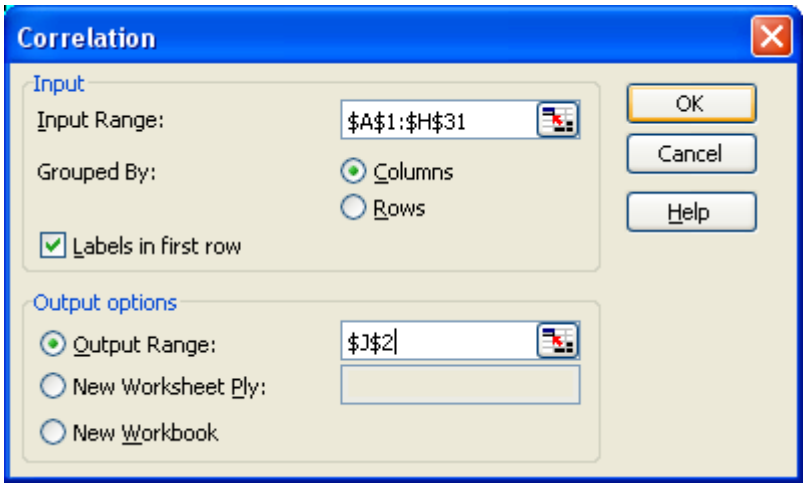

Figura 3. Fereastra Correlation cu setările descrise mai sus

7. Rezultatul va fi o matrice de coeficienți de corelaţie (figura 4).

8. Reprezentați coeficienții cu 2 zecimale şi interpretați rezultatele.

	I	J	K	L	M	N	O	P
1								
2			*VARSTA*	*GREUTATE*	*IMC*	*TAS*	*TAD*	*GLICEMIE*
3		VARSTA	1					
4		GREUTATE	-0,13778	1				
5		IMC	0,051764	0,853890594	1			
6		TAS	0,715771	0,066455169	0,249824	1		
7		TAD	0,462438	0,494847105	0,507708	0,720649	1	
8		GLICEMIE	0,409769	0,362771316	0,474754	0,387625	0,345454	1

Figura 4. Matricea cu coeficienții de corelație calculați

Exerciţiul 4

Reprezentaţi grafic dependenţa (corelaţia) dintre *Vârstă* şi *TAS*, adăugaţi pe grafic dreapta de regresie asociată, calculaţi coeficientul de determinare d şi determinaţi ecuaţia dreptei de regresie.

Indicaţii

1. Selectați valorile celor două coloane (cu tasta CTRL apăsată) și țineți cont că prima variabilă va fi reprezentată pe axa Ox.

2. Din meniu selectați **Insert – Scatter**

3. La opțiunea **Design** alegeți formatul de reprezentare a graficului cu dreapta de regresie și ecuația ei, f_x

4. un exemplu de reprezentare este dat in figura 5

5. Observați că diagrama de dispersie are o tendinţă crescătoare, deci dependenţa dintre *TAS* şi *Vârstă* este pozitivă: o creştere a Vârstei implică o creştere a TAS. 51% din variaţia TAS se datorează relaţiei liniare cu vârsta (51% reprezintă coeficientul de determinare, R^2). Ecuația dreptei de regresie este TAS=0,9935xVarsta+81,633.

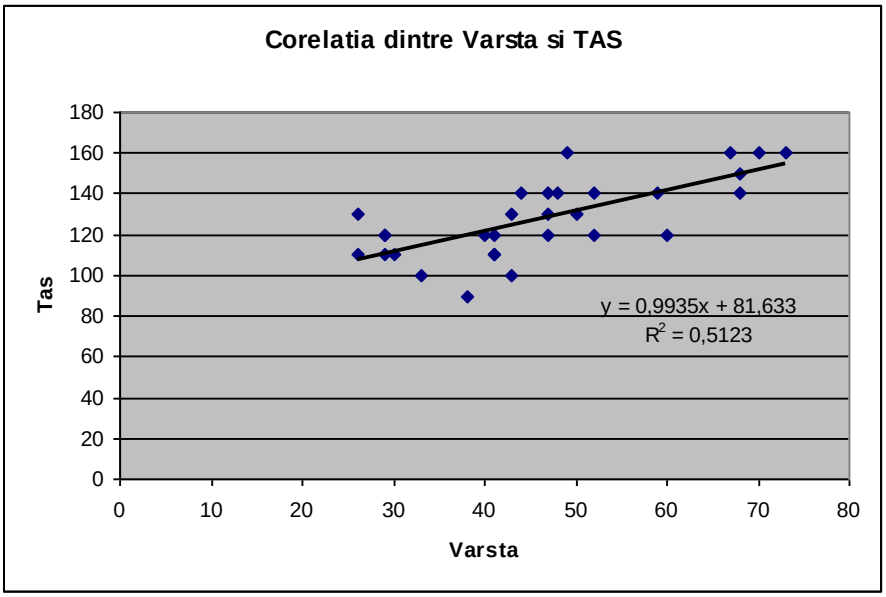

Figura 5. Graficul dependenței variabilei TAS de variabila Varsta

Exercițiul 5

Realizați analiza regresiei dintre cele două variabile de la exercițiul 4.

Indicaţii

1. Alegeţi **Regression** din opţiunile din fereastra **Data Analysis**

2. Selectaţi domeniul valorilor variabilei TAS, de exemplu B1:B20 ca variabilă dependentă (**Input Y Range**), selectaţi variabila Vârstă, de exemplu A1:A20 ca variabilă independentă (**Input X Range**), bifaţi **Labels**, pentru obţinerea intervalelor de încredere bifaţi opţiunea **Confidence Level** (cu nivelul de semnificaţie de **95%**) ca în figura 6

3. Rezultatele sunt reprezentate în figura 7.

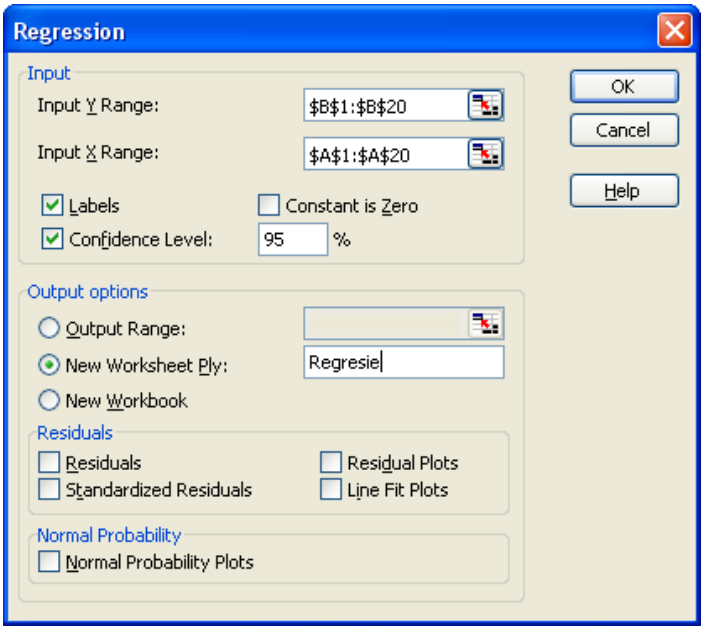

Figura 6. Analiza regresiei liniare

1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0,715771								
5	R Square	0,512328								
6	Adjusted R Square	0,494911								
7	Standard Error	13,5468								
8	Observations	30								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	1	5398,228	5398,228	29,41562807	8,72199E-06				
13	Residual	28	5138,438	183,5157						
14	Total	29	10536,67							
15										
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*	
17	Intercept	81,6327	8,840703	9,233734	5,4269E-10	63,52331613	99,7420746	63,5233161	99,74207458	
18	VARSTA	0,993539	0,183187	5,423618	8,72199E-06	0,618295946	1,36878195	0,61829595	1,368781953	
19										

Figura 7. Rezultatele analizei regresiei liniare

Multiple R = 0,71 este coeficientul de corelaţie multiplu, dar în cazul nostru este coeficientul de corelaţie Pearson.

R Square = 0,51 este coeficientul de determinare multiplu R^2 reprezintă proporţia variaţiei lui Y explicată de relaţia liniară cu X.

p-value = 8,721E-06 = 0,000008721 în acest caz se respinge ipoteza nulă (p-value<0,001), adică corelaţia dintre cele două variabile este extrem de semnificativă.

Coefficients – pentru Intercept (constanta) valoarea este 81,6327, iar pentru coeficientul ”a” valoarea este 0,993539. Deci dreapta de regresie Y=aX+b în cazul nostru este Y=0,993539X+81,6327.

3. Colectarea și prelucrarea datelor cu specific de medicină dentară

S-a efectuat un studiu pe un eșantion de pacienți care au avut restaurări protetice pe diferiți dinți și au fost evaluați după o anumită perioadă de timp (măsurată în luni de zile).

Datele de interes care au fost colectate sunt prezentate în tabelul următor:

Vârsta	Sex	Perioadă de timp	Status la evaluare
36	F	39	SUCCES
61	F	56	SUCCES
25	F	41	SUCCES
29	M	21	SUCCES
37	M	56	ESEC
32	F	33	ESEC
33	M	33	SUCCES
33	M	46	SUCCES
32	M	33	SUCCES
34	F	58	SUCCES
30	M	59	SUCCES
33	F	62	SUCCES
59	F	28	ESEC
26	F	32	SUCCES
33	M	32	SUCCES
45	F	54	SUCCES
34	M	25	SUCCES
45	F	63	SUCCES
17	F	25	SUCCES
56	F	22	SUCCES
45	F	61	SUCCES
48	F	38	SUCCES
41	F	29	SUCCES
33	M	25	SUCCES
45	F	56	SUCCES
31	M	39	SUCCES
27	M	49	ESEC
23	F	30	SUCCES
45	F	30	SUCCES
41	F	24	SUCCES
29	M	32	SUCCES

Exerciţiul 1

Precizați care este numărul pacienților pentru care s-a efectuat studiul.

indicație:transferați tabelul cu date în Excel

Exerciţiul 2

Calculați media de vârstă și media perioadei de evaluare (cu o zecimală).

indicație: funcția AVERAGE; format cells – number, 1 decimal

Exerciţiul 3

Precizați câți pacienți de sex masculin/ feminin avem în studiu.

indicație: funcția COUNTIF…

Exerciţiul 4

Trasați un grafic al vârstelor, comparativ, după sexe.

indicație:reprezentați pe același grafic mediile vârstelor, pentru ambele sexe; sortați mai întâi datele după sex, pentru a calcula mai ușor mediile (sau aplicați un filtru pentru coloana sex).

Exerciţiul 5

Care este perioada minimă, maximă și totală de evaluare?

indicație:funcțiile MIN(), MAX() și SUM()…

Exerciţiul 6

Precizați rata de succes/ eșec total.

indicație:calculați procentul ce-l reprezintă numărul de succese/eșecuri din total.

Exerciţiul 7

Trasați graficul perioadei de evaluare după status.

indicație:reprezentați mediile perioadei pentru success și eșec, pe același grafic.

Exercițiul 8

Reprezentați grafic rata de succes/eșec pentru fiecare sex în parte.

indicație:realizați în prealabil un tabel de distribuție al pacienților după sexe și status, utilizând filter pentru cele 2 coloane de interes.

Exercițiul 9

În tabelul urmator, cazurile au fost defalcate după cele 4 tipuri de restaurări protetice în funcție de tipul de material: RPIC (restaurare protetică integral ceramică), RPZr (restaurare protetica ceramică pe suport zirconia), RPMC (restaurare protetică metalo-ceramică), RPT (restaurare protetică partial mobilizabilă telescopată):

Status evaluare	Tip de restaurare protetică coronară			
	RPIC	RPZr	RPMC	RPT
ESEC	2	1	1	1
SUCCES	9	12	18	13

1. Să se reprezinte grafic aceste cazuri, comparativ, dupa status şi tipul de restaurare protetică.
2. Să se calculeze (printr-o reprezentare grafică) rata de eșec/succes pe fiecare tip de restaurare protetică.

4. Concluzii

În aceast material am reuşit să:

- calculăm coeficientul de corelație Pearson;
- realizăm graphic asocierile dintre două variabile numerice;
- facem și să interpretăm rezultatele analizei regresiei liniare;
- aplicăm funcții statistice;
- realizăm grafice suggestive pentru prelucrările aplicate.

Referințe

Vernic CV, Timar B, Mada L, Apostol SA. Informatică medicală şi metode de biostatistică aplicate în nursing. Editura Eurostampa, 2014: 1-388.

Mihalas GI, Lungeanu D. Introducere in informatica medicala si biostatistica, Editura „Victor Babes", 2009: 1-225.