Noţiuni de biostatistică. Corelația și regresia

Corelația: Întrebări referitoare la cercetare

Exemplu

- Există o asociere între vârstă şi tensiunea arterială?

Semnificație

- Evaluează dacă două variabile sunt asociate, adică, dacă valorile unei variabile tind să fie mai mari (sau mai mici) decât valorile celeilalte variabile.

- Asocierile între două variabile continue.

Corelația

- Măsoară puterea asocierii liniare între două variabile continue,
- Poate fi pozitivă sau negativă,
- Poate varia între -1 şi +1,
- Nu implică cauzalitate (poate exista alt factor care să explice asocierea).

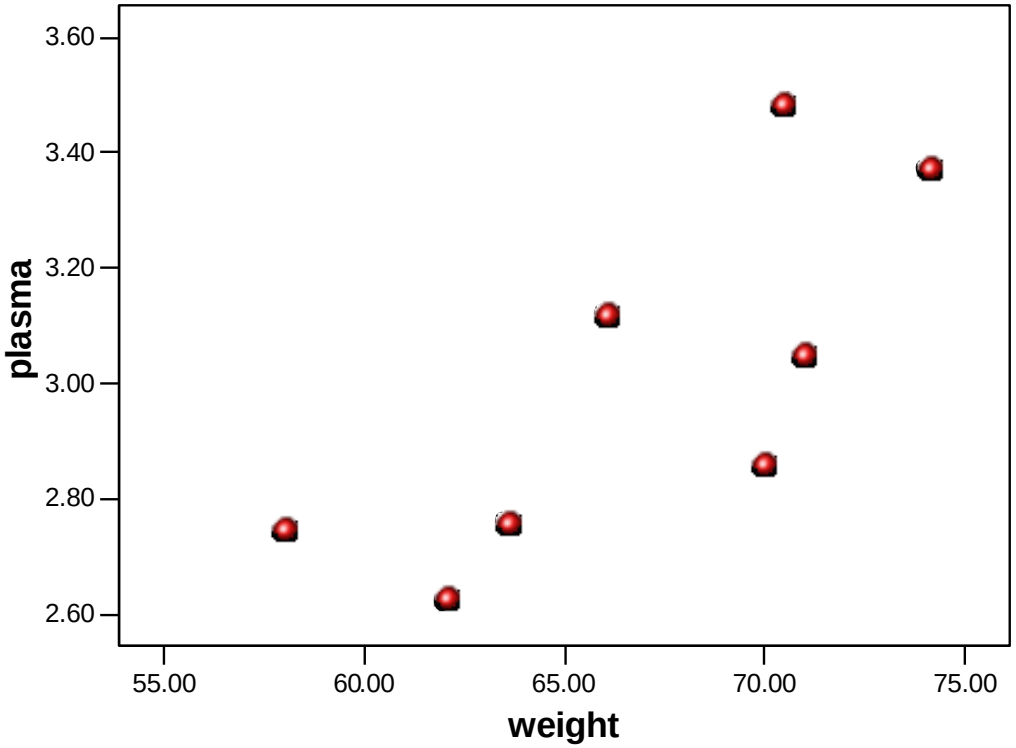
Observații privind corelația

- Nu înseamnă că o variabilă o determină pe cealaltă.
- Consumul de cafea şi accidentele rutiere sunt într-o legătură puternică (r = 0.61), dar acest lucru nu indică faptul că ingerarea cafelei produce accidente de circulație.

Coeficient de corelație Pearson

Există o asociere între greutatea corporală şi volumul plasmei?

	(X) greu. corp.	(Y) volum plasma
1	58.0	2.75
2	70.0	2.86
3	74.0	3.37
4	63.5	2.76
5	62.0	2.62
6	70.5	3.49
7	71.0	3.05
8	66.0	3.12



Coeficientul de corelație

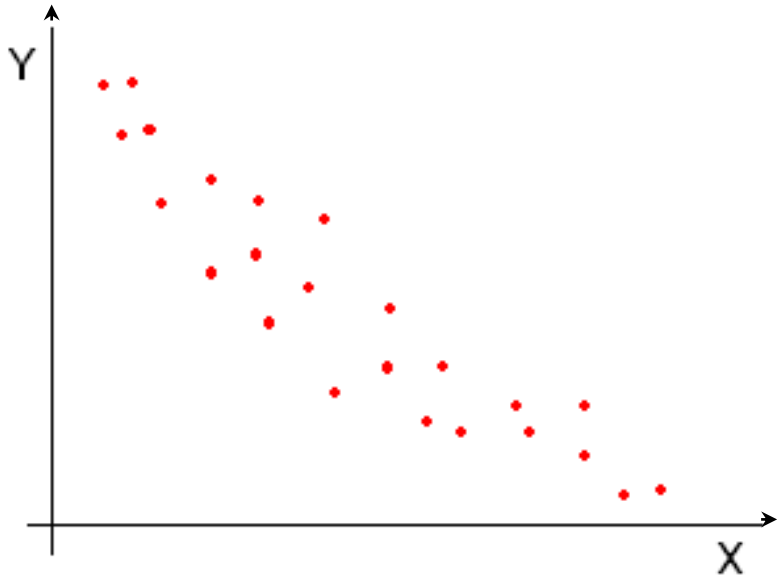
- **Coeficientul de corelație (r) se calculează cu formula:**

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

Coeficientul de corelație Pearson

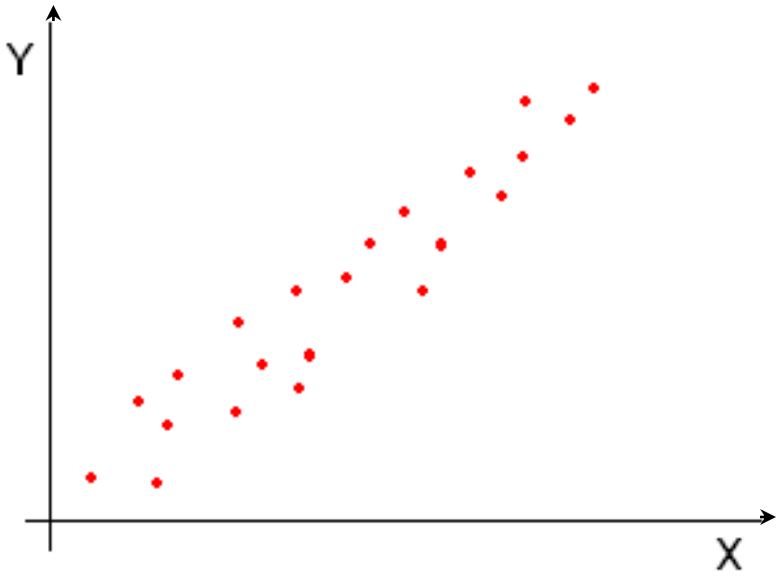
r = -1	Legătură liniară negativă puternică Pe măsură ce valoarea lui X creşte valoarea lui Y scade.
r = 0	Fără legătură liniară între X şi Y
r = +1	Legătură liniară pozitivă puternică Pe măsură ce valoarea lui X creşte, valoarea lui Y creşte (fie scad amândouă).

Coeficientul de corelație Pearson



r = -1

r = +1



Testarea ipotezei pentru coeficientul de corelație

- Semnificația coeficientului de corelație depinde de mărimea coeficientului de corelație şi de numărul de observații din eşantion,

- Validitatea acestui test necesită ca variabilele care sunt observate să fie pe un eşantion randomizat de indivizi şi cel puțin una din variabile să aibă o distribuție normală.

Corelații ne-parametrice

- Când datele sunt categorice (nu sunt continue) sau datele nu au o distribuție normală, poate fi aplicată o metodă de corelație a rangurilor (*Corelația rangurilor a lui Spearman*).

PRESUPUNERILE PRIVIND DATELE NU SUNT VALABILE
TESTELE PARAMETRICE NU POT FI FOLOSITE

TESTE NE-PARAMETRICE

Exemplu: Corelația rangurilor Spearman

- A fost realizat un studiu pentru a investiga relația între scorurile de anxietate ale copiilor şi evaluarea mamelor referitoare la anxietatea copiilor lor.

- Scorurile de anxietate ale copiilor au fost măsurate pe o scală de intervale, scorurile mamelor au fost măsurate pe o scală între 1-7.

H_0 : nu există nici o legătură între evaluările copiilor şi mamelor în ceea ce priveşte anxietatea

$H_0 : \rho = 0$

Exemplu:
Corelația rangurilor Spearman

- Coeficientul de corelație a rangurilor se calculează la fel ca şi coeficientul Pearson, cu excepția faptului că se lucrează pe ranguri şi nu pe valori reale.

- Are valori între -1 to +1 şi are aceeaşi interpretare,

- Nu este necesar ca datele să aibă o distribuție normală (ne-parametric).

Corelația rangurilor Spearman

Correlations

			anxiety score(mother)	anxiety score (child)
Spearman's rho	anxiety score (mother)	Correlation Coefficien	1.000	.638*
		Sig. (2-tailed)	.	.035
		N	11	11
	anxiety score (child)	Correlation Coefficien	.638*	1.000
		Sig. (2-tailed)	.035	.
		N	11	11

*. Correlation is significant at the .05 level (2-tailed).

Corelația este semnificativă la un nivel de 5% (P < 0.05), astfel că ipoteza nulă este respinsă, ceea ce înseamnă că există o legătură între evaluările copiilor şi mamelor referitoare la anxietate.

Regresia liniară simplă

Întrebări privind cercetarea

- Cum se modifică tensiunea arterială sistolică pe măsură ce vârsta creşte?
- Poate fi tensiunea arterială sistolică prezisă plecând de la vârsta subiectului?
- Poate fi prezisă cantitatea de grăsime a organismului prin măsurarea circumferinței abdominale?

Regresia Liniară Simplă

- Regresia liniară simplă descrie legătura între două variabile continue,
- Regresia liniară simplă dă ecuația liniei drepte care descrie cel mai bine asociația dintre două variabile continue,
- Permite predicția unei variabile folosind informații de la cealaltă variabilă.

Tipuri de variabile în regresia liniară

- Variabila dependentă este variabila care va fi prezisă (adică rezultatul care ne interesează),

- Variabila independentă sau variabila explicativă este variabila folosită pentru predicția unui anumit rezultat.

Ecuația unei linii drepte

- Ecuația unei linii drepte este:

$$y' = a + bx$$

y' este valoarea prezisă (variabila dependentă),

a este intersecția cu ordonata,

b este panta (sau gradientul) liniei,

x este varabila independentă (explicativă).

Least Squares (Pătratele minime)

- Valorile lui a şi b sunt calculate pentru a minimiza suma pătratelor distanțelor verticale între linia de regresie şi variabila dependentă. Aceasta se numeşte "least squares fit" (potrivirea pătratelor minime).

- Aceasta reprezintă diferența între valoarea reală a variabilei dependente şi valoarea prezisă față de linia de regresie pentru fiecare valoarea a variabilei independente.

Coeficientul de regresie

- Panta, b, este adesea numită coeficientul de regresie,
- Are acelaşi semn ca şi coeficientul de corelație,
- Când nu există nici o corelație între x şi y, atunci coeficientul de regresie, b, va fi egal cu 0.
- Intersecția cu ordonata, a, este valoarea prezisă a lui y când x este egal cu zero.

Valorile reziduale

$$y = a + bx + \varepsilon$$

ε – se numeşte valoarea reziduală

- Valoarea reziduală este diferența între valoarea prezisă (calculată din ecuația de regresie) şi valoarea observată (y' – y),
- O valoare reziduală este calculată pentru fiecare observație,
- Metoda pătratelor minime încearcă să minimizeze suma pătratelor valorilor reziduale,
- Tehnici matematice sunt folosite pentru a găsi valorile lui a şi b care satisfac potrivirea pătratelor minime.

Valoarea prezisă (y')

- Valoarea prezisă, y', depinde de variația de eşantionare,
- Precizia sa poate fi estimată (eroarea predicției) cu ajutorul erorii standard,
- Cu cât eroarea standard este mai mare, cu atât este mai mare dispersia valorilor prezise ale lui y în jurul liniei de regresie şi ca o consecință eroarea predicției este mai mare.

Exemplu

- O firmă de gimnastică de întreținere doreşte să evalueze grăsimea corporală a clienților. Ei ar dori să fie capabili să prezică cantitatea de grăsime corporală pe baza unor măsurători care să poată fi obținute uşor.

- La 252 de bărbați li s-a măsurat circumferința abdominală.

Testarea ipotezelor

H_0: Nu există nici o relație liniară între grăsimea corporală şi circumferința abdominală în populație,

H_1: Există o relație liniară între grăsimea corporală şi circumferința abdominală în populație.

<u>Reformulare</u>:

H_0: Circumferința abdominală nu contează, indiferent de variabilitatea în grăsimea corporală la nivelul populației,

H_1: Circumferința abdominală contează pentru o anumită variabilitate în grăsimea corporală la nivelul populației.

Regresia liniară simplă

- Variabila dependentă
  - > grăsimea corporală
- Variabila independentă
  - > circumferința abdominală

Coeficienții de regresie calculați pe baza unui eşantion de observații (a şi b) sunt estimări ale coeficienților de regresie de la nivel populațional (α şi β),

Testarea ipotezei şi intervalele de încredere pot fi construite folosind estimările eşantionului pentru a face inferențe referitoare la coeficienții de regresie în populație,

Pentru ca aceste inferențe să fie valide, este necesar să verificăm distribuția datelor (liniaritate, normalitate, constanța varianței).

Exemplu

Grăsimea corporală (%) & **Circumferința abdominală**

(dependentă) **(independentă)**

H_0: Nu există nici o legătură liniară între grăsimea corporală şi circumferința abdominală la nivelul populației

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.814[a]	.662	.661	4.5144

a. Predictors: (Constant), ABDOMEN

coeficientul de corelație

proporția variației în % de grăsime corporală explicată de model

Regresia liniară

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9984.086	1	9984.086	489.903	.000[a]
	Residual	5094.931	250	20.380		
	Total	15079.017	251			

a. Predictors: (Constant), ABDOMEN

b. Dependent Variable: BODYFATB

Interpretarea tabelului ANOVA

- **O proporție semnificativă statistic a variabilității în măsurarea grăsimii corporale poate fi atribuită modelului de regresie ($p<0.001$).**

H_0: variabila independentă (circumferința abd) nu contează pentru variabilitatea grăsimii corporale în populație,

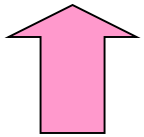
Ceea ce este echivalent cu H_0: $\beta = 0$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9984.086	1	9984.086	489.903	.000[a]
	Residual	5094.931	250	20.380		
	Total	15079.017	251			

a. Predictors: (Constant), ABDOMEN

b. Dependent Variable: BODYFATB



$$\text{F ratio} = \frac{\text{MSReg}}{\text{MSRes}} = 489.9$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.814[a]	.662	.661	4.5144

a. Predictors: (Constant), ABDOMEN

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9984.086	1	9984.086	489.903	.000[a]
	Residual	5094.931	250	20.380		
	Total	15079.017	251			

a. Predictors: (Constant), ABDOMEN

b. Dependent Variable: BODYFATB

$R^2 = 9984.09/15079.02 = 0.66$

Modelul de regresie liniară simplă: $Y = \alpha + \beta X$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-35.197	2.462		-14.294	.000
	ABDOMEN	.585	.026	.814	22.134	.000

a. Dependent Variable: BODYFATB

Estimări ale parametrilor de regresie

Rezultatele T-test pentru ipotezele
H_0: $\alpha = 0$ şi H_0: $\beta = 0$

$$y' = a + bx = -35.197 + 0.585\,x$$

unde,
y este % grăsime corporală, iar x este circ. abd.

Predicția

- Cum pot folosi regresia liniară pentru predicție?
- Ecuația de regresie îmi permite să prezic valoarea variabilei dependente (Y) pentru o anumită valoare a varibilei independente (X),
- Predicția grăsimii corp = -35.197 + (0.585 x circ abd)
- Care este valoarea predicției grăsimii corporale pentru un bărbat cu circum. abd. de 100cm?
- Predicția grăsimii corp. = -35.197 + (0.585 x 100)
  = -35.197 + 58.5
  = 23.3%

Presupuneri ale regresiei liniare

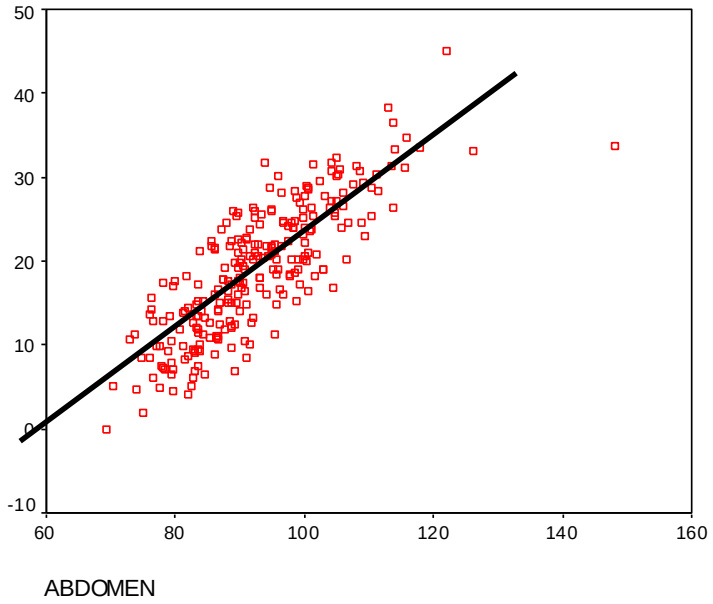
- Ar trebui să existe o relație liniară între variabila dependentă şi variabila independentă,
- Pentru orice valoare a variabilei independente valorile variabilei dependente ar trebui să aibă o distribuție Normală (adică, valori reziduale cu distribuție normală),
- Varianța valorilor variabilei dependente ar trebui să fie aceeaşi pentru toate valorile variabilei independente.

Verificarea presupunerilor:

- **După ce modelul de regresie a fost realizat este esențial să verificăm că presupunerile de regresie liniară nu au fost încălcate.**

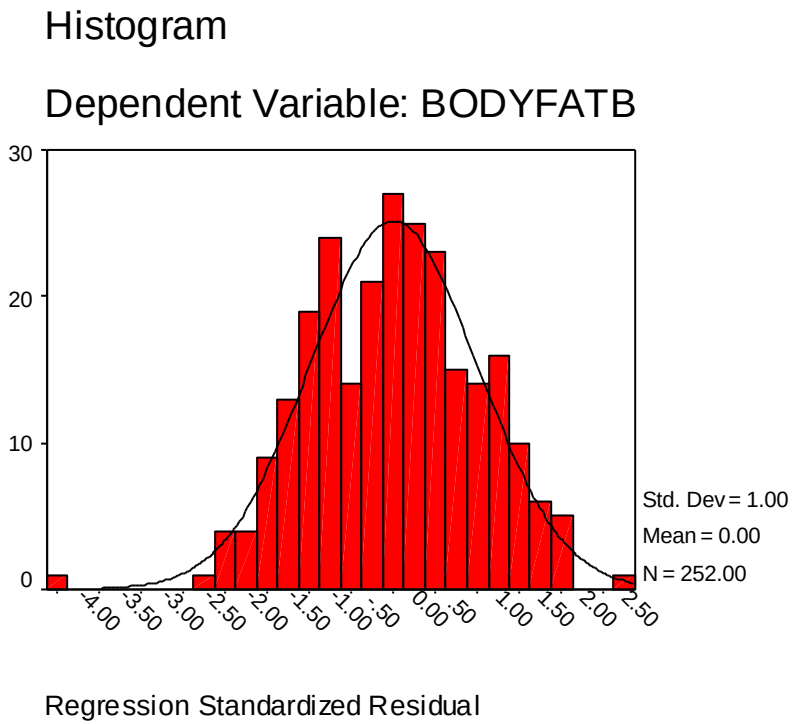
- **Dacă oricare din aceste presupuneri a fost încălcată atunci modelul de regresie foarte probabil nu este valabil.**

Presupuneri: Liniaritatea (1)



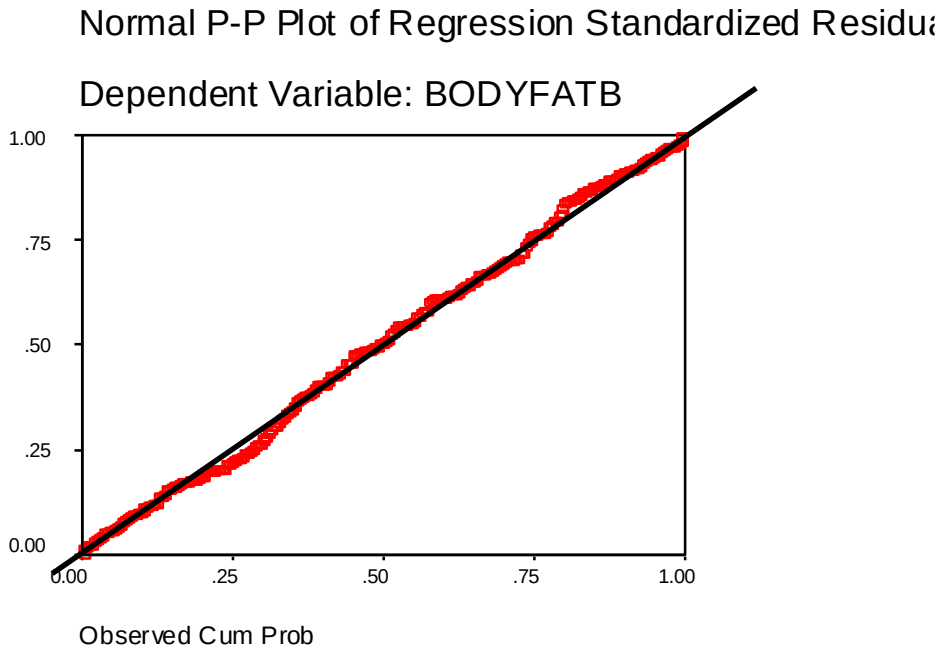
- Se reprezintă grafic variabila dependentă în funcție de variabila independentă pentru evaluarea relației liniare.
- Presupunere satisfăcută

Presupuneri: Valori reziduale normale



- **Valorile reziduale cu distribuție normală pot fi testate analizând histograma lor.**

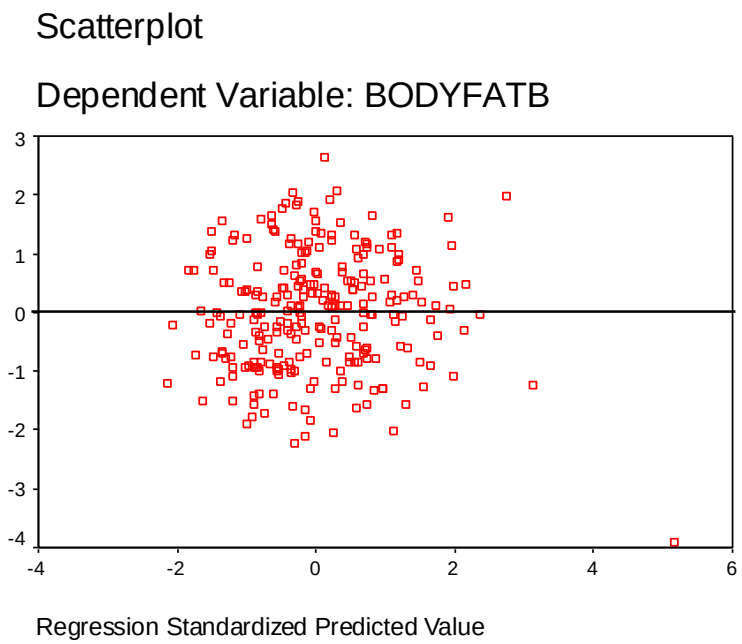
Presupuneri: Valori reziduale normale (2)



- **Valorile reziduale cu distribuție normală pot fi testate analizând graficul din imagine.**

- **Presupunere satisfacută** 

Presupunere: varianță constantă



- **Varianța constantă a valorilor reziduale poate fi apreciată reprezentând grafic valorile reziduale vs. valorile prezise,**

- **Ar trebui să existe o distribuție uniformă în jurul valorii zero.**

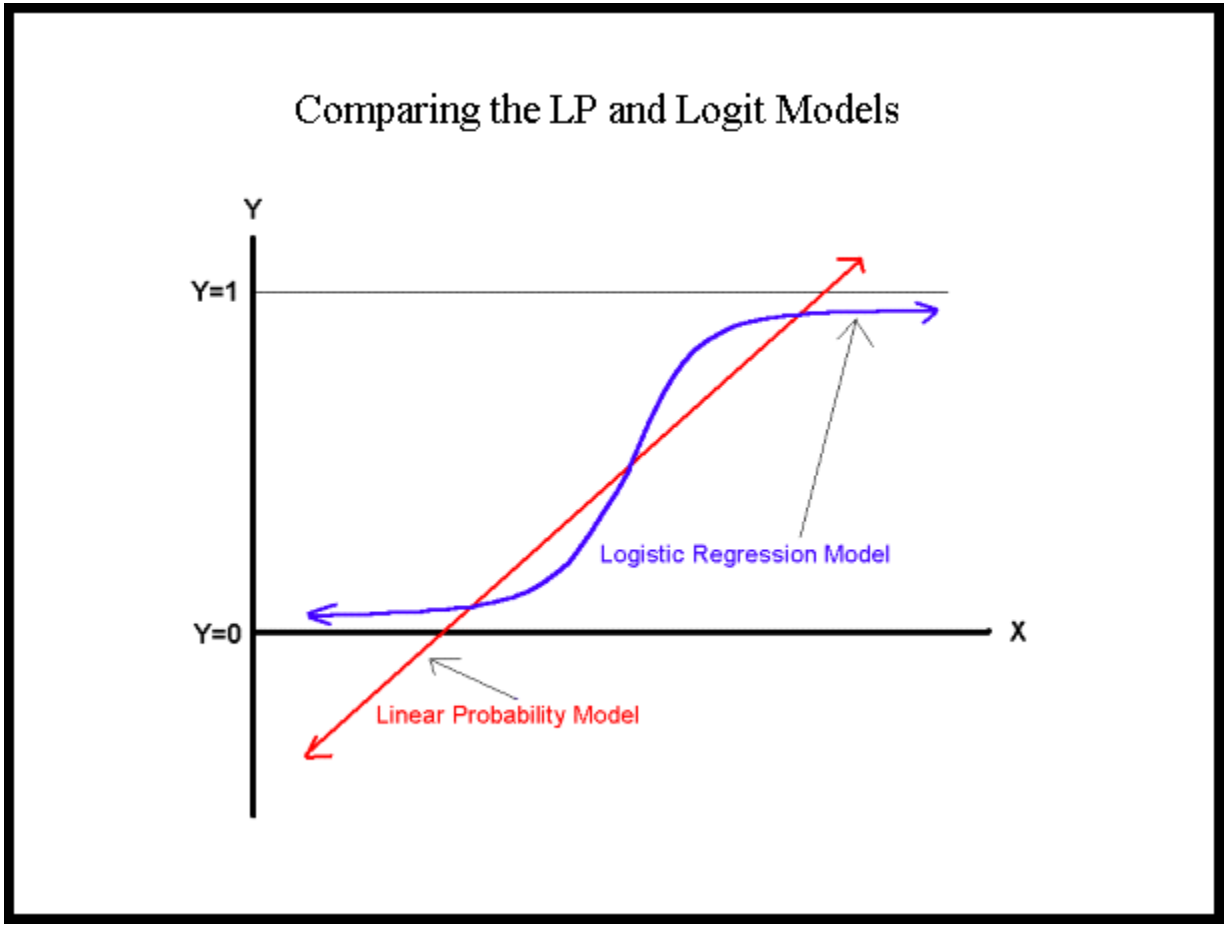
Presupunere: varianță constantă

- Această presupunere nu ar fi satisfăcută dacă răspândirea valorilor reziduale ar creşte sau ar descreşte pe măsură ce valorile prezise ar creşte ca mărime
- Graficul ar trebui să ilustreze o relație randomizată
- Presupunere satisfăcută 

Regresia logistică

- Acest tip de regresie este folosit atunci când variabila dependentă este binară (dichotomică) de ex. prezența sau absența bolii,
- În această situație "clasica" regresie liniară multiplă nu este potrivită,
- Această metodă poate fi folosită pentru a compara caracteristicile subiecților cu sau fără o anumită boală,
- Foarte des folosită în studiile epidemiologice.

Comparaţie regresia liniară-regresia logistică



Exemplu de Regresie Logistică

- Există disponibile informații referitoare la 111 pacienți consecutivi admişi la UPU. Cercetătorul doreşte să investigheze relația dintre statusul vital al pacienților (viu/decedat) şi caracteristicile pacienților la internare.

Strategia pentru analiză:

- Tabele 2x2,
- Calcularea şansei (odds) a unui eveniment de a se produce,
- Regresia Logistică.

Statistică Descriptivă

	Viu	Decedat
	N=71 (64%)	N=40 (36%)
Sex:		
masculin	43 (61%)	24 (60%)
(feminin)	28 (39%)	16 (40%)
Tipul de internare:		
obişnuită	15 (21%)	2 (5%)
(urgență)	56 (79%)	38 (95%)
Vârstă		
<50	26 (37%)	4 (10%)
51-69	21 (30%)	18 (45%)
70+	24 (34%)	18 (45%)

Tabel 2x2

Testarea asocierii între tipul de internare şi statusul vital (tabel 2x2).

Type of admission * vital status Crosstabulation

			vital status		Total
			lived	died	
Type of admission	elective	Count	15	2	17
		% within Type of admission	88.2%	11.8%	100.0%
	Emergency	Count	56	38	94
		% within Type of admission	59.6%	40.4%	100.0%
Total		Count	71	40	111
		% within Type of admission	64.0%	36.0%	100.0%

Calculul şanselor de deces (odds)

Type of admission * vital status Crosstabulation

Count

		vital status		
		lived	died	Total
Type of admission	elective	15	2	17
	Emergency	56	38	94
Total		71	40	111

- **Şansele unui pacient internat în urgenţă de a deceda la UPU = (38/94)/(1-38/94) = 38/56 = 0.679**
- **Şansele unui pacient internare obişnuită de a deceda la UPU = (2/17)/(1-2/17) = 2/15 = 0.133**
- **Astfel, OR pentru tipul de internare = 0.679/0.133 = 5.11**
- **Interpretare: şansele (odds) ale unui pacient de a deceda la UPU este de 5 ori mai mare dacă pacientul a fost internat ca urgenţă comparativ cu situaţia când nu a fost o urgenţă.**

Regresia Logistică

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	TYPE	1.625	.781	4.330	1	.037	5.079
	Constant	-3.638	1.519	5.736	1	.017	.026

a. Variable(s) entered on step 1: TYPE.

Type:
0=obişnuită
1=urgenţă

- **Coeficientul pentru TIPUL de internare este 1.625. Aceasta indică diferenţa în "log odds" între cei internaţi în urgenţă şi nu ca fiind 1.625,**
- **Creşterea în şansele (ODDS) pentru o internare de urgenţă este cu un factor de exp{1.625},**
- **OR (urgenţă/normal) este exp{1.625} = 5.079 (prezentat în ultima coloană din tabelul de mai sus),**
- **Rezultatul este aproape identic cu cel obţinut în urma tabelului 2x2.**

Cât de bun este modelul pentru predicția efectului?

Classification Table[a]

Observed			Predicted		
			vital status		Percentage Correct
			lived	died	
Step 1	vital status	lived	38	33	53.5
		died	6	34	85.0
	Overall Percentage				64.9

a. The cut value is .500

- **Tabelul pentru clasificare ne arată cât de bun este modelul de regresie logistică în predicția rezultatului,**
- **Modelul este mai bun pentru predicția celor ca au decedat şi pe ansamblu, 65% din cazuri au avut o predicție corectă numai pe baza tipului de internare.**