METODE CORELATIONALE
Regresia si Corelatia

Stud. Master - AMP

ISAIC- MANIU ALEXANDRU
web www.amaniu.ase.ro
e-mail AL.ISAIC-MANIU@CSIE.ASE.RO

Context

- Intre metodele cantitative de cercatare utile sunt si cele de studiere a **dependenţelor** dintre două sau mai multe variabile,incluse în **"analiza de regresie şi corelaţie".**
- În cadrul acesteia se studiază dependenţa dintre o variabilă (caracteristică) rezultativă (**y**) şi una sau mai multe variabile (caracteristici) independente (**x**).
- Caracteristica rezultativă se mai numeşte caracteristica *dependentă, endogenă sau efect*, iar caracteristica independentă se mai numeşte caracteristica *factorială, exogenă sau cauză*.
- **Regresia** ne arată **cum** (ca formă analitică) o variabilă este dependentă de altă variabilă (sau de alte variabile), iar **corelaţia** ne arată **gradul** (cat) în care o variabilă este dependentă de o altă variabilă (sau alte variabile).

I. REGRESIA
Clasificări

a) după numărul caracteristicilor independente:
- legături simple
- legături multiple

b) după direcţia legăturilor, acestea pot fi:
- legături directe
- legături inverse

c) după expresia analitică a legăturilor:
- legături liniare
- legături neliniare

d) după tipul legăturii
- legături funcţionale
- legături stohastice

e) după modul de manifestare în timp:
- legături concomitente
- legături cu decalaj

METODE ELEMENTARE DE CARACTERIZARE A LEGĂTURILOR DINTRE VARIABILE

- 1. metoda seriilor paralele independente
- 2. metoda grupărilor
- 3. metoda tabelului de corelaţie
- 4. metoda grafică

1. Metoda seriilor paralele interdependente

Se ordonează observaţiile în funcţie de caracteristica independentă x (crescător sau descrescător) şi se urmează modul în care se aranjează valorile lui y.

Concluzii:

- caracteristica y se ordonează aproximativ crescător - rezultă că putem aprecia că între cele două variabile există o legătură directă;
- caracteristica y se ordonează aproximativ descrescător rezultă că putem aprecia că între cele două variabile există o legătură inversă;
- caracteristica y nu înregistrează o tendinţă de ordonare (crescător sau descrescător) - rezultă că putem aprecia că între cele două variabile nu există legătură.

2. Metoda grupărilor

- Se repartizează unităţile în grupe omogene în funcţie de o caracteristică independentă.
- Pentru fiecare grupă astfel constituită se centralizează datele numerice referitoare la caracteristica rezultativă şi se calculează medii pe fiecare grupă şi mărimi relative.
- Prin comparaţia variaţiei caracteristicii independente cu indicatorii calculaţi pentru caracteristica rezultată se poate aprecia existenţa şi forma legăturilor dintre cele două variabile.

3. Metoda tabelului de contingenţă

- Tabelul de contingenţă este un tabel cu dublă intrare şi prezintă o grupare a unităţilor unei colectivităţii în funcţie de două caracteristici: una dependentă şi alta independentă.
- Se foloseşte în special în cadrul unui număr mare de observaţii.
- Dacă considerăm două variabile, de exemplu "naţionalitate" şi "religie", atunci tabelul poate fi de forma:

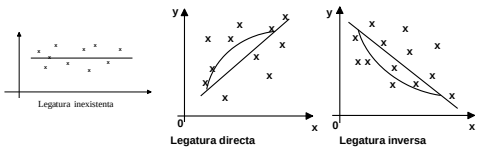
Exemplu - Tabel de contingenţă

Religia/ Naţiona-litatea	Română	Maghiară	.	.	General	.	.	Slovaci	Altele	Total
Ortodoxă	n_{11}	n_{12}	.	.	n_{1j}	.	.	n_{1p-1}	n_{1p}	$n_{1.}$
Romano-catolică	n_{21}	n_{22}	.	.	n_{2j}	.	.	n_{2p-1}	n_{2p}	$n_{2.}$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
General	n_{i1}	n_{i2}	.	.	n_{ij}	.	.	n_{ip-1}	n_{ip}	$n_{i.}$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
Musulma-nă	n_{r-11}	n_{r-12}	.	.	n_{r-1j}	.	.	n_{r-1p-1}	n_{r-1p}	$n_{r-1.}$
Altele	n_{r1}	n_{r2}	.	.	n_{rj}	.	.	n_{rp-1}	n_{rp}	$n_{r.}$

4. Metoda grafică

Graficul se construieşte pornind de le perechile de valori observate (**x, y**) care se reprezintă în sistemul de axe rectangulare – ***corelogramă*** *(nor de puncte)*.

Pe axa OX se reprezintă variabila independentă x, iar pe axa OY variabila dependentă y.



5. MODELUL UNIFACTORIAL LINIAR

Modelul probabilist la nivelul colectivităţii generale:

unde $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$

- (x_i, y_i) reprezintă valorile numerice ale variabilelor cauză şi efect înregistrate la nivelul unităţii statistice „*i*";
- α, β = parametri constanţi
- α = punctul de intersecţie al dreptei de regresie cu axa Oy;
- β = panta dreptei, se mai numeşte şi „*coeficient de regresie*" şi arată cu câte unităţi de măsură se modifică **Y** dacă **X** se modifică cu o unitate
- ε_i = componenta reziduală (eroare aleatoare) pentru unitatea statistică „*i*".

5.1.MODELUL UNIFACTORIAL LINIAR

Valoarea reală y_i a caracteristicii **Y** din modelul probabilistic cuprinde:

- *componenta teoretică, deterministă* ( ), adică $\hat{y}_i$ partea din valoarea reală y_i care se poate determina pe baza modelului pentru o anumită valoare x_i:

$$\hat{y}_i = \alpha + \beta \cdot x_i$$

- *componenta aleatoare (reziduală)*, numită şi *eroarea aleatoare* (ε_i) reprezentând acea parte din valoarea reală a lui Y care nu se poate cuantifica:

$$y_i = \hat{y}_i + \varepsilon_i$$

5.2. MODELUL UNIFACTORIAL LINIAR

- Dacă datele disponibile provin dintr-un eşantion, avem *n* perechi de observaţii reale:

(x_1,y_1), (x_2, y_2), ... , (x_n, y_n), pe baza cărora se vor estima parametrii ecuaţiei, α şi β.

- Modelul de regresie în eşantion va fi $y_i = a + b \cdot x_i + e_i$

cu $\hat{y}_i = a + b \cdot x_i$

unde:

- a = estimatorul parametrului α din colectiv. generală;
- b = estimatorul parametrului β din colectiv. generală;
- e_i = valoarea reziduală pt. unitatea "*i*" în eşantion.

$$e_i = y_i - \hat{y}_i$$

5.3.Estimarea parametrilor modelului unifactorial liniar

Metoda celor mai mici pătrate presupune maximizarea similitudinii, a gradului de asemănare a valorilor teoretice cu valorile reale, deci minimizarea erorilor.

Cum erorile se pot produce intr-un sens sau în altul faţă de valorile reale, ea presupune ***minimizarea sumei pătratelor reziduurilor***:

5.3.-b Estimarea parametrilor modelului unifactorial liniar

$$S = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - b \cdot x_i)^2 \to \text{minim}$$

Condiţiile de ordin 1 de minimizare a funcţiei sunt:

$$\begin{cases} \frac{dS}{da} = 0 \\ \frac{dS}{db} = 0 \end{cases} \Rightarrow \begin{cases} \sum 2(y_i - a - b \cdot x_i)(-1) = 0 \\ \sum 2(y_i - a - b \cdot x_i)(-x_i) = 0 \end{cases} \Rightarrow \begin{cases} \sum y_i - na - b\sum x_i = 0 \\ \sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0 \end{cases}$$

$$\begin{cases} na + b\sum x_i = \sum y_i \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i \end{cases}$$

5.3.-c. Estimarea parametrilor modelului unifactorial liniar

Aplicând metoda determinanţilor, se obţine:

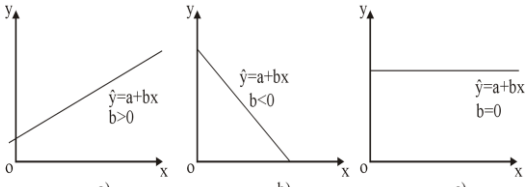
$$\Delta a = \begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{vmatrix} \qquad \Delta b = \begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix}$$

$$\Delta = \begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}$$

$$a = \frac{\Delta a}{\Delta} = \frac{\sum y_i \cdot \sum x_i^2 - \sum x_i \cdot \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{\Delta b}{\Delta} = \frac{n \cdot \sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

5.3.-f. Estimarea parametrilor modelului unifactorial liniar



Linii de regresie cu a) pantă pozitivă b) pantă negativă c) pantă egală cu zero

6. Modele neliniare de regresie

Modelul exponenţial transformat al ecuaţiei exponenţiale are la bază ecuaţia:

$y = a\ b^x$ care se estimează folosind modelul:

$Y = a\ b^x + \varepsilon$

Prin logaritmare, modelul se poate transforma într-un model liniar de forma:

$\lg Y = \lg a + x \lg b$

Făcând următoarele înlocuiri:

$Y' = \lg Y$; $a' = \lg a$; $b' = \lg b$, rezultă ecuaţia unei drepte, respectiv:

$y' = a' + b' x$

Interpretări

- ***Coeficientul „a"*****,** care poate lua atât valori pozitive cât şi negative, reprezintă ordonata la origine, respectiv este valoarea lui „**y**" când **"x**" este egal cu zero.

- ***Coeficientul „b"*** - denumit coeficient de regresie - arată măsura în care variază caracteristica dependentă în cazul în care caracteristica independentă se modifică cu o unitate.
- În funcţie de ***semnul coeficientului de regresie***, putem aprecia tipul de legătură:
  - în cazul corelaţiei directe, coeficientul are o valoare pozitivă;
  - în cazul corelaţiei inverse, valoarea lui este negativă;
  - în cazul în care **b** = 0, se apreciază că variabilele (y şi x) sunt independente.
- În graficul de corelaţie coeficientul „**b**" indică panta liniei drepte.

II . Metoda corelaţiei

- A. Corelaţia parametrică (variabile măsurate pe scala de raport)
- B. Corelaţia neparametrică (variabile măsurate pe scala nominală, ordinală sau de interval)

A. Corelaţia parametrică

Metoda corelaţiei prezintă avantajul că oferă o măsură sintetică a legăturilor dintre variabilele statistice.

Indicatorii care măsoară intensitatea legăturii sunt: covarianţa, coeficientul de corelaţie şi raportul de corelaţie.

1. COVARIANŢA

Covarianţa se calculează sub forma mediei aritmetice simple a produselor abaterilor celor două variabile corelate, ***x*** şi ***y***, de la mediile lor aritmetice $\bar{x}$ şi $\bar{y}$, conform relaţiei:

$$\operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)$$

1. Covarianţa

Covariaţia ***nulă*** - variabilele sunt independente (lipsa legăturii de corelaţie)

- Valoarea sa absolută cov (x,y) ***nu are limită superioară***. Pe măsură ce intensitatea corelaţiei creşte şi covariaţia creşte.
- Indicatorul reprezintă avantajul că se calculează destul de uşor. În acelaşi timp, prezintă şi dezavantajul că depinde de unităţile în care se măsoară variabilele aleatoare. Deci nu este comparabil de la o variabilă la alta.
- Indicatorul ia valori pozitive dacă legătura dintre variabile este directă şi valori negative în coz contrar.
- Valori apropiate de zero semnifică lipsa oricărei legături între x şi y; valori ridicate ale indicatorului arată o legătură puternică.

2. COEFICIENTUL DE CORELAŢIE LINIARĂ SIMPLĂ

Măsoară numai intensitatea legăturii de ***tip liniar*** dintre două variabile *x* şi *y*. Se calculează ca o medie aritmetică a produsului abaterilor normale normate ale celor două variabile.

Notând *abaterile normale normate* ale variabilelor *x* şi *y*:

$$z_x = \frac{x_i - \bar{x}}{s_x} \quad ; \quad z_y = \frac{y_i - \bar{y}}{s_y}$$

rezultă următoarea relaţie de calcul (în care „*n* " este numărul observaţiilor-perechi)

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \cdot s_x \cdot s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Faţă de covarianţă rezultă că relaţia:

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \cdot s_x \cdot s_y}$$

se transformă în coeficientul de corelaţie liniară simplă.

2. COEFICIENTUL DE CORELAŢIE LINIARĂ SIMPLĂ

În practică se utilizează relaţia:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}$$

Coeficientul de corelaţie simplă se mai poate calcula şi cu relaţia:

$$r_{xy} = b\frac{s_x}{s_y},$$

în care:

b - este coeficientul de regresie simplă;

s_x *- abaterea medie pătratică a caracteristicii factoriale;*

s_y *- abaterea medie pătratică a caracteristicii rezultative.*

2.a. COEFICIENTUL DE CORELAŢIE LINIARĂ SIMPLĂ

- Coeficientul de corelaţie poate lua ***valori*** cuprinse între -1 şi +1, adică satisface inegalităţile: $-1 \le r_{yx} \le 1$, iar semnul său, ca şi cel al coeficientului de regresie, semnifică tipul de legătură: semnul minus indică legătura inversă, semnul plus indică legătura directă.
- Cu cât coeficientul de corelaţie are valori mai apropiate de 1 sau –1, cu atât corelaţia liniară dintre variabilele *x* şi *y* este mai puternică.
- Pe măsură ce coeficientul de corelaţie se apropie de 0, scade şi *intensitatea* legăturii dintre cele două variabile.
- În cazul în care r_{yx}=0, variabilele sunt independente ori necorelate liniar, iar pentru r_{yx}=1 rezultă dependenţa funcţională între cele două variabile.

3. - Raportul de corelaţie

Denumit şi **coeficientul de corelaţie** Pearson, acest indicator măsoară atât intensitatea legăturilor liniare, cât şi neliniare. Se defineşte cu relaţia:

$$R_{y/x} = \sqrt{1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}}$$

Coeficientul de determinatie

R^2

3.1. - Raportul de corelaţie

Dispersiile au următoarele semnificaţii:

- măsoară acţiunea tuturor factorilor care au influenţat asupra variabilei rezultative:
- măsoară variaţia valorilor *y* sub influenţa tuturor celorlalţi factori necuprinşi în model, a căror acţiune e considerată constantă; este denumită şi dispersia reziduală;
- măsoară numai influenţa variabilei independente sau factoriale *x* asupra variabilei *y*. Cu cât ponderea acestei dispersii în cadrul dispersiei generale va fi mai mare, cu atât legătura dintre cele două variabile va fi mai puternică.

3.2. - Raportul de corelaţie

Interpretare

Raportul de corelaţie poate lua valori între 0 şi 1. Cu cât valoarea raportului este mai apropiată de 1 cu atât legătura de corelaţie este mai puternică şi invers.

- Daca R→1 legatura dintre X si Y este puternica.
- Daca R →0 legatura dintre X si Y este slaba.

În cazul corelaţiei liniare, raportul de corelaţie este egal cu coeficientul de corelaţie luat în valoare absolută şi această relaţie poate fi considerată ca un test de verificare a liniarităţii legăturii.

- In cazul legaturilor liniare:

$$R = |r_{xy}|$$

B. Corelaţia neparametrică

Dacă metodele de analiză a corelaţiei nu mai folosesc parametrii distribuţiilor, se poate vorbi despre o corelaţie neparametrică (sau liberă de distribuţie), iar acest lucru se întâmplă în următoarele ***situaţii***:

- Variabilele sunt de natură **calitativă** (măsurate pe scală nominală sau ordinală);
- Datele **nu** provin dintr-o populaţie cu *distribuţie normală* de probabilitate sau aproximativ normală;
- Sunt *informaţii insuficiente* pentru a putea presupune normalitatea distribuţiei (de exp, provin din eşantioane de volum redus)

1. Coeficientul de asociere Q

Distribuţia persoanelor în funcţie de naţionalitate şi religie

Religia/ Naţionalitatea	Român	Non-român	**Total**
Ortodox	n_{11}	n_{12}	$\mathbf{n_{1.}}$
Non-ortodox	n_{21}	n_{22}	$\mathbf{n_{2.}}$
Total	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	**n**

Coeficientul Q

A fost introdus de statisticianul englez **Yule**. Formula sa de calcul este:

$$Q = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{n_{11} \cdot n_{22} + n_{12} \cdot n_{21}}$$

Acesta ia valori tot între –1 şi 1.

În cazul unor variabile independente valoarea coeficientului Q este nulă.

2. Coeficienţii de corelaţie ai rangurilor

- a. Coeficientul de corelaţie a rangurilor introdus de **Spearman** se bazează pe analiza concordanţei rangurilor acordate pentru fiecare din cele n unităţi statistice, după variabila *X* şi după variabila *Y*.

$$C_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

- b. Coeficientul rangurilor **Kendall** necesită ordonarea unităţilor după rangurile acordate variabilei *X* şi înscrierea în paralel a rangurilor acordate variabilei *Y*.

$$C_K = \frac{2\sum S_i}{n(n-1)}$$

Exemplu

Pe baza datelor din anuarul statistic, s-au înregistrat datele următoare pentru 10 judete.

Nr.	Jud.	Supraf. (km2)	Nr. Comunelor
1	AB	6242	66
2	AG	6826	93
3	AR	7754	67
4	BC	6621	79
5	BH	7544	86
6	BN	5355	53
7	BR	4766	39
8	BT	4986	68
9	BV	5363	43
10	BZ	6103	81

Să se stabilească dacă există o legătură între suprafaţa totală şi numărul comunelor, utilizând coeficienţii de corelaţie a rangurilor a lui Spearman şi Kendall.