CURSUL VI

ANALIZA LEGĂTURILOR DINTRE PROCESELE ŞI FENOMENELE ECONOMICE

6.1 Tipologia legaturilor statistice

Este cunoscut faptul că variabilitatea fenomenelor social-economice este, în majoritatea cazurilor, determinată de acţiunea simultană a mai multor factori; o parte dintre aceşti factori favorizează evoluţia unui fenomen, alţii o frânează sau acţionează chiar în sens invers.

Sensurile şi intensităţile influenţelor diferiţilor factori se schimbă în condiţii timp şi spaţiu, astfel că evoluţia fenomenelor dependente înregistrează şi ea tendinţe diferite faţă de cele anterioare.

Raporturile de cauzalitate dintre fenomenele social-economice pot fi cuantificate şi analizate cu ajutorul corelaţiei. Informaţiile obţinute sunt deosebit de utile, mai ales pentru faptul că, metodele specifice pe care statistica le pune la dispoziţia cercetătorului, oferă posibilitatea cunoaşterii, în principal, a următoarelor aspecte:

- existenţa raporturilor de cauzalitate dintre fenomene;
- contribuţia fiecărui factor la variabilitatea globală a fenomenelor efect;
- intensitatea legăturilor cauzale dintre fenomenele şi procesele social-economice;
- tendinţele evolutive ale corelaţiei dintre fenomene.

Se remarcă faptul că analiza corelaţiei oferă o paletă mult mai largă de informaţii, fiind preferată deşi determinarea indicatorilor specifici corelaţiei este mult mai dificilă.

Corelaţia *este expresia sintetică a intensităţii legăturilor cauzale dintre fenomene.*

Cuplul corelativ poate cuprinde două sau mai multe variabile, din care una este variabila efect, cunoscută sub numele de variabilă **rezultativă**, simbolizată în general cu Y, iar celelalte sunt variabile cauză**,** cunoscute sub numele de variabile **factoriale** - fiind simbolizate cu $X_1,X_2,... ,X_n$.

Se deosebesc mai multe tipuri de corelaţie.

a) După numărul de variabile din cuplul corelativ, se disting:

- **corelaţia simplă**, în care cuplul corelativ cuprinde două variabile, din care una este variabila rezultativă Y şi cealaltă - variabila factorială X.

- **corelaţie multiplă**, în care cuplul corelativ cuprinde trei sau mai multe variabile, din care una este variabila rezultativă Y şi celelalte sunt variabilele factoriale X_1 , X_2 , etc.

b) După sensul legăturilor factoriale, se deosebesc:

- **corelaţia directă**, în care variabilitatea rezultativei Y se produce în acelaşi sens cu variabilitatea factorialei X (sau factorialelor).

- corelaţia inversă, în care variabilitatea rezultativei Y se produce în sens invers, comparativ cu variabilitatea factorilor determinanţi.

c) După forma legăturilor cauzale, se disting :

- Corelaţie liniară. Constă în aceea că variabila rezultativă Y înregistrează o tendinţă lineară, ca urmare a influenţei factorilor determinanţi.

- Corelaţie neliniară. În acest caz, variabilitatea rezultativei Y se integrează într-o tendinţă de tip nelinear (parabolic, exponenţial, etc).

Cunoaşterea formei corelaţiilor prezintă un interes deosebit în estimarea tendinţelor evolutive ale fenomenelor efect în strânsă legătură cu variabilitatea factorilor determinanţi. Evident, o tendinţă lineară a rezultativei Y, sugerează faptul că nivelurile acestei variabile cresc sau descresc aproximativ în progresie aritmetică; tendinţa neliniară, evidenţiază modificarea acestor niveluri în progresie geometrică, exponenţială, etc. Aceste informaţii sunt deosebit de utile în calculele de previziune precum şi în modelarea tendinţelor evolutive.

Dintre **metodele elementare** de analiză a corelaţiei, prezentate în literatura de specialitate, se folosesc mai frecvent: tabelul de corelaţie şi metoda grafică.

- **Tabelul de corelaţie.** Este de fapt tabelul cu dublă intrare folosit în prezentarea seriilor de distribuţie bidimensionale; rolul variabilei factoriale este preluat de caracteristica principală X şi cel al variabilei rezultative este preluat de caracteristica secundară Y.

Cu ajutorul tabelului de corelaţie pot fi evidenţiate următoarele aspecte: existenţa corelaţiei, sensul corelaţiei, forma corelaţiei şi intensitatea corelaţiei.

Tabelul de corelaţie serveşte, în majoritatea cazurilor în analizele profesionale, ca bază preliminară a folosirii metodelor analitice în studiul corelaţiei.

- **Metoda grafică**. Presupune construirea şi folosirea **corelogramei** în analiza corelaţiei. Pentru a construi corelograma, pe abscisă se trec valorile scării de reprezentare a variabilei factoriale X iar pe ordonată valorile variabilei rezultative Y. Prin unirea punctelor corespunzătoare coordonatelor XY se obţine corelograma.

Cu ajutorul acestei metode se pot evidenţia următoarele aspecte: existenţa corelaţiei, sensul corelaţiei, forma corelaţiei şi intensitatea corelaţiei.

Şi metoda grafică este frecvent folosită în analiza corelaţiei, fiind utilizată în foarte multe cazuri ca element esenţial în alegerea funcţiilor statistico-matematice pentru analiza regresiei şi intensităţii corelaţiei.

În continuare va fi abordată metodologia aplicării funcţiilor statistico-matematice în studiul regresiei simple, precum şi metodologia cuantificării intensităţii corelaţiei simple.

6.2 Metoda regresiei

În cadrul analizei corelaţiei se au în vedere, în principal, două aspecte esenţiale :

- regresia - cu ajutorul căreia, prin folosirea şi interpretarea coeficienţilor de regresie ai diferitelor funcţii statistico-matematice, se determină contribuţiile factorilor determinanţi la variabilitatea fenomenelor efect;

- intensitatea corelaţiei - sintetizată cu ajutorul coeficienţilor de corelaţie.

Pentru corelaţia simplă, primul aspect poate fi evidenţiat cu ajutorul funcţiilor : lineară - pentru legăturile cauzale de tip linear; parabolică de ordin superior, hiperbolică, exponenţială, logaritmică, semilogaritmică, logistică, etc - pentru legături cauzale de tip nelinear.

Corelaţia simplă, după forma pe care o au tendinţele legăturilor cauzale, poate fi **lineară** şi **neliniară.**

Modelul **linear** de ordinul 1 al regresiei simple este dat de expresia:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

în care

y = variabila dependentă (reyultativă);

x = variabila independentă(factorială);

β_0 = valoarea iniţială a lui y;

β_1 = modificarea lui y ca urmare a modificării lui x egală cu unitatea;

ε = eroarea variabilei.

Pentru determinarea parametrilor β_0 şi β_1 se folosesc ca estimatori parametrii funcţiei lineare:

$$\hat{y} = b_0 + b_1 x$$

in care

b_0-valorile variabilei rezultative Y determinate în afara influenţei variabilei factoriale X;

b_1- coeficientul de regresie; el sintetizează creşterea (când are semnul +) sau descreşterea (când are semnul -) înregistrată de variabila rezultativă Y corespunzătoare unei creşteri sau descreşteri a variabilei X egală cu unitatea;

x- valorile $x_1, x_2, x_3, \ldots, x_n$ ale variabilei factorialei X.

Parametrul b_1 are un rol esenţial în cadrul analizei regresiei, deoarece cu ajutorul său se determină contribuţia factorialei X la variabilitatea rezultativei Y.

Aplicând metoda celor mai mici pătrate, $\sum(y - \hat{y})^2 \Rightarrow \min$, cei doi parametri b_0 şi b_1 se pot determina astfel:

$$b_1 = \frac{\operatorname{cov}(X,Y)}{s_x^2} \quad ; \quad b_0 = \bar{y} - b_1 \bar{x}$$

Elementele folosite ca bază de calcul pentru cei doi parametri se determină astfel:

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} yi}{n}$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

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

În analiza fenomenelor social-economice sunt foarte frecvente cazurile în care valorile empirice y_i şi valorile teoretice $\hat{y}_i$ există diferenţe; acestea sunt determinate de acţiunea factorilor întâmplători şi constituie valoarea reziduală a estimării, cunoscută şi sub numele de eroare $(y_i - \hat{y}_i)$.

Eroarea medie (sau eroarea standard) a estimaţiei, se calculează cu formula:

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

unde *SEE* reprezintă suma pătratelor erorilor de estimaţie, şi are următoarea relaţie de calcul:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Pentru a stabili că b_1 estimează corect pe β_1, trebuie verificată ipoteza nulă: H_0:β_1= 0. Alternativa ei este: H_0 : $\beta_1 \neq 0$. În acest scop se poate folosi testul t, determinat cu ajutorul relaţiei:

$$t = \frac{b_1 - \beta_1}{S_{b1}}$$

unde

S_{b1} = eroarea standard a lui b_1.

$$Sb_1 = \frac{S_e}{\sqrt{(n-1)s_x^2}}$$

Dacă erorile variabilei sunt normal distribuite, testul statistic t (Student), considerat ca nivel referenţial, este cel căruia îi corespund n-2 grade de libertate.

Regiunea de respingere a ipotezei nule este:

$$t > t_{\alpha/2,n-2} \text{ sau } t < -t_{\alpha/2,n-2}$$

Spre exemplu, dacă n =100 şi α = 0,05 atunci zona de respingere este:

$$t > t_{0.025,98} \cong 1.984 \quad sau \quad t < -t_{0.025,98} \cong -1.984$$

6.3 Coeficientul de determinaţie

În practică, pentru determinarea contribuţiei unuia dintre factorii de influenţă la variabilitatea fenomenului dependent, se foloseşte coeficientul de determinaţie, a cărui relaţie de calcul este:

$$R^2 = \frac{[\text{cov}(X,Y)]^2}{s_x^2 s_y^2}$$

sau

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SSE}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Având în vedere faptul că abaterile $(y_i - \bar{y})$ evidenţiază variaţia totală a variabilei Y, determinată de toţi factorii care au influenţat-o şi ştiind că o parte din această variaţie este preluată de funcţia de regresie şi evidenţiată prin abaterile $(\hat{y}_i - \bar{y})$, iar cealaltă parte - care constituie variaţia reziduală sau eroarea estimaţiei - este evidenţiată prin abaterile $(y_i - \hat{y}_i)$, înseamnă că:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Dar şi suma pătratelor acestor abateri verifică această relaţie. Deci:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

Considerând:

$$SS_y=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

rezultă că pentru calculul coeficientului de determinaţie se pot folosi mai multe variante de calcul:

$$R^2=1-\frac{SSE}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$R^2=\frac{SS_y-SSE}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$R^2=\frac{SSR}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

În practică R^2 se exprimă şi în procente.

6.4 Coeficientul de corelaţie al lui Pearson.

Coeficientul de determinaţie R^2 evidenţiază proporţia variaţiei variabilei dependente *Y* determinată de influenţa variaţiei variabilei idependente *X*.

Coeficientul de corelaţie *r* evidenţiază intensitatea legăturii cauzale dintre cele două variabile. Ambii coeficienţi au la bază relaţia lui Pearson; coeficientul de determinaţie este pătratul coeficientului de corelaţie. Deci:

$$r=\sqrt{R^2}=\frac{\operatorname{cov}(X,Y)}{s_x s_y}$$

Coeficientul de corelaţie *r* ia valori cuprinse între -1 şi 1 cu următoarea semnificaţie:

- valorile cuprinse între 0 şi 1 evidenţiază o corelaţie directă, din ce în ce mai intensă pe măsură ce acestea se apropie de 1;
- valorile cuprinse între 0 şi -1 evidenţiază o corelaţie inversă, din ce în ce mai intensă pe măsură ce se apropie de -1;
- valoarea zero evidenţiază faptul că între cele două variabile nu există nici o legătură.

De regulă, în practică se rafinează intervalul dintre -1 şi 1 astfel:

- dacă $0 \leq r < 0{,}2$ nu avem o legătură semnificativă între variabile;
- dacă $0{,}20 \leq r < 0{,}50$ legătura dintre variabile este slabă;
- dacă $0{,}50 \leq r < 0{,}75$ legătura dintre variabile este medie;
- dacă $0{,}75 \leq r < 0{,}95$ legătura dintre variabile este puternică;
- dacă $0{,}95 \leq r \leq 1$ între cele două variabile există o legătură de tip funcţional.

Pentru testarea semnificaţiei coeficientului de corelaţie se are în vedere verificarea ipotezei nule (H_0) - ce se face cu ajutorul *testului t*- potrivit căreia între cele două variabile nu este nici o legătură lineară.

Testul t (Student) folosit în acest caz are următoarea relaţie de calcul:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

unde

r - coeficientul de corelaţie;

n - numărul perechilor de valori x şi y.

Valoarea calculată a lui t se compară cu valoarea teoretică obţinută din tabelul *t (Student)*, pentru n-2 grade de libertate şi nivelul de semnificaţie stabilit.

Regiunea de respingere a ipotezei H_0, (pentru n-2 grade de libertate) se stabileşte în acelaşi mod ca cel de la parametrii funcţiei de regresie, adică:

$$t > t_{\alpha/2,n-2} \text{ sau } t < -t_{\alpha/2,n-2}$$

Dacă H_0 se respinge, se trage concluzia, cu un risc asumat (de regulă 5%), că valoarea coeficientului de corelaţie nu este egală cu 0, adică între cele două variabile există o legătură semnificativă sau altfel spus, coeficientul de corelaţie este semnificativ statistic.

În cazul că H_0 se acceptă, coeficientul de corelaţie nu este semnificativ, şi deci între variabile există o legătură întâmplătoare.

În practică este foarte des întâlnit şi coeficientul de corelaţie a rangurilor al lui Spearman, care are la bază relaţia coeficientului lui Pearson. În locul valorilor x,y Spearman a folosit rangurile corespunzătoare acestor valori, ajungând la următoarea relaţie:

$$r_s = \frac{\operatorname{cov}(a,b)}{s_a s_b}$$

unde

a = rangurile valorilor x

b = rangurile valorilor y

Coeficientul lui Spearman ia valori cuprinse între -1 şi 1, având aceeaşi semnificaţie ca cel a lui Pearson.

Pentru verificarea ipotezei nule $H_0 : \rho = 0$ se foloseşte testul statistic z, care are următoarea relaţie:

$$z = \frac{r_s - 0}{1/\sqrt{n-1}} = r_s\sqrt{n-1}$$

Se compară z calculat cu valoarea tabelară corespunzătoare unui anumit nivel de semnificaţie şi dacă valoarea calculată este mai mare decât z_{tab} sau mai mică decât $-z_{tab}$, atunci ipoteza H_0 se respinge.