***REGRESSION***

Estimarea coeficienţilor unui model liniar prin metoda celor mai mici pătrate şi calculul statisticilor necesare testelor statistice asociate sunt efectuate de procedura **Regression**, una dintre cele mai complexe din pachetul de prelucrări statistice dinExcel. Procedura permite şi construirea graficelor necesare pentru aprecierea vizuală a potrivirii modelului liniar. Deşi acestea, din motive evidente, necesită prelucrări suplimentare de scalare înainte de interpretare, existenţa lor este un real ajutor pentru statistician.

***Termeni***

Modelul liniar estimat de procedură

este

Y = α0X0 + α1X1 + α2X2 + … + αp-1Xp-1 + ε,

care exprimă faptul că variabila Y se poate obţine ca o combinaţie liniară a variabilelor X0, X1,…, Xp-1 la care se adaugă o "eroare" ε.

Pentru estimarea parametrilor modelului se consideră disponibile *n* observaţii asupra tuturor variabilelor din model. Valorile sunt structurate ca un tablou dreptunghiular, fiecare variabilă ocupând o coloană (deci o linie este referită drept o observaţie).

Dialogul procedurii **Regression** este prezentat în figura următoare.

***Input***

***Input Y Range***– se precizeazădomeniul (coloana) pe care se aflăvalorile variabilei depen-dente.

***Input X Range***– se precizeazădomeniul pe care se aflăvalorile tuturor variabilelorindependente. Acest domeniu trebuie să fie compact, fiecare variabilă Xi ocupând o coloană.

***Labels***– se marcheazăboxa de control în cazul în care prima linie din tabloul de date este cudenumirile variabilelor (situaţie recomandată).

***Constant Is Zero***– se marcheazăboxa de control dacămodelul care se estimeazăeste fărătermen liber.

***Confidence Level***– se precizează, procentual, siguranţa statisticădorităîn raportareaintervalelor de încredere deci valoarea (1–α)×100, unde α este pragul de semnificaţie. Intervalele obţinute sunt suplimentare, întotdeauna afişându-se cele pentru α = 0,05. Boxa se va marca doar dacă se doreşte şi un alt prag de semnificaţie.

***Output options***

***Output Range, New Worksheet Ply, New Workbook***– Precizeazăzona unde se vor înscrierezultatele. Zona de rezultate este foarte complexă, cuprinde tabele care depind de mărimea modelului, de numărul de observaţii, de numărul graficelor dorite etc. Prin urmare se va prefera o foaie de calcul nouă sau o zonă liberă în dreapta şi în jos.

***Residuals***

***Residuals***– se marcheazăboxa de control în cazul când se doreşte calcularea reziduurilormodelului estimat.

***Residual Plots***– se marcheazăboxa de control în cazul când se doreşte obţinerea diagramelor *reziduuri – variabilă independentă,* adicăvizualizarea punctelor de coordonate(xij, rj), j = 1,…n, având ca abscisă o valoare a variabilei independente Xi, iar ca ordonată reziduul corespunzător.

***Standardized Residuals***– aceastăboxăde control se va marca dacăse doreşte calcululvalorilor standardizate ale reziduurilor. Valorile astfel obţinute provin, teoretic, dintr-o distribuţie normală standard, astfel încât o histogramă a acestor valori trebuie să se apropie de curba normală (clopotul lui Gauss).

 *L****ine Fit Plots*** – se marcheazăaceastăboxăde control dacăse doreşte afişarea diagramelor *Y*

*– variabilă independentă,* prin care se vizualizează, pe un acelaşi grafic, punctele decoordonate (xij, yobs,i), (xij, yest,i), j = 1,…,n, unde abscisele sunt valorile variabilei independente, iar ordonatele sunt valorile observate şi cele estimate ale variabilei dependente. Este desenat câte un grafic pentru fiecare variabilă independentă.

Interpretarea acestor diagrame poate oferi indicaţii asupra adecvanţei modelului, asupra valorilor aberante.

***Normal Probability***

***Normal Probability Plots***– se marcheazădacăse doreşte vizualizarea repartiţiei de sondaj avariabilei Y într-o reţea de probabilitate.

***Exemplu***

**Un set de date cuprinde 25 de observaţii asupra a 4 variabile, notate Y (considerată variabila dependentă) şi X1, X2, X3 (considerate variabile independente). Valorile şi denumirile ocupă în foaia de calcul un domeniu dreptunghiular continuu, B2:E27, valorile Y ocupând prima coloană.**

Pentru a estima modelul liniar

Y = α0 + α1X1 + α2X2 + α3X3 + ε,

cu termen constant, se apelează procedura **Regression**.

1. **Un prim tabel de rezultate, prezentat în figura alăturată, conţine statisticile generale ale ecuaţiei de regresie.**

***Multiple R*** – coeficientul multiplu de corelaţie.

***R Square*** – coeficientul de determinare (este egal cu pătratulcoeficientului de corelaţie multiplă). Poate fi gândit, exprimat procentual, drept proporţia din variaţia variabilei dependente explicată de variaţia variabilelor independente: 60,7% din variaţia lui Y este explicată de variabilele X.

***Adjusted R Square*** – valoarea corectatăa coeficientului dedeterminare. Este introdusă pentru a contracara (parţial)

efectul creşterii mecanice a lui R2 o dată cu numărul variabilelor independente.

***Standard Error*** – eroarea standard a estimaţiei. Se calculeazăca abaterea standard areziduurilor (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.

1. **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: 3 = p – 1, 21 = n – p, 24 = n – 1, unde p = 4 este numărul parametrilor modelului (trei variabile X plus termenul liber) iar n = 25 este numărul de observaţii.

**SS** – sumele de pătrate potrivit descompunerii

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Suma globală | = | Suma de pătrate | + | Suma de pătrate |
| de pătrate | datorată regresiei | 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:α1=α2=α3=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 criticăunilaterală. Dacăvaloarea afişatăeste mai micădecât pragul de semnificaţie fixat, atunci se respinge ipoteza nulă în favoarea ipotezei alternative.

1. **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 referăla variabilele din model, incluzândşi 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 rezultăcămodelul estimat în exemplu este

Y = 11,718 – 1,443\*X1 + 3,135\*X2 – 0,324\*X3.

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 repartiţieicoeficientului).

***t Stat***– statistica t pentru verificarea ipotezei H0:αi= 0 contra ipotezei alternative H1:αi≠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. Adică 2,161 = 11,718/5,421 etc. Utilizarea statisticii este cea uzuală.

***P-value***– probabilitatea criticăbilateralăa testului t cu ipotezele precizate la t Stat. Pentrupragul de semnificaţie α = 0,05 se poate respinge ipoteza de nulitate a termenului liber (0,042 < 0,05) şi a coeficienţilor α1 şi α2 (0,002 şi 0,000 sunt mai mici decât 0,05). Nu se poate respinge ipoteza nulă privind coeficientul α3 (0,069 > 0,05).

***Lower 95%, Upper 95%***– limitele inferioară şi superioarăale intervalului de încredere pentruparametrul respectiv. Limitele la pragul 0,05 sunt calculate automat, indiferent de iniţializarea procedurii Regression.

Se poate deci interpreta că, în populaţie, parametrii modelului liniar sunt cuprinşi în intervalele următoare:

0,444 < α0 < 22,992

−2,271 < α1 < −0,615

...

Se poate observa că ultimul interval cuprinde şi valoarea zero, prin urmare se regăseşte concluzia privind nerespingerea ipotezei nule H0 : α3 = 0.

1. **Studiul reziduurilor se poate face pe baza datelor raportate în tabelul alocat reziduurilor, tabel având structura următoare**:

****

Pentru fiecare observaţie (linie din tabelul de date iniţial) se afişează:

 ***Observation***– numărul de ordine al observaţiei.

***Predicted y***– valoarea y prognozatăpentru observaţia respectivă; se obţine înlocuind valorileX ale observaţiei în modelul estimat.

 ***Residuals***– valoarea erorii de predicţie (diferenţa dintre valoarea observată şi valoareaprognozată).

***Standard Reziduals***– valoarea standardizatăa erorii. Este obţinutăprin împărţireareziduului la abaterea standard a reziduurilor (rezultatul nu este susţinut absolut riguros de teorie).

1. **Analiza calităţii modelului este facilitată şi de graficele construite automat de procedura Regression. Sunt produse două tipuri de diagrame:**

– diagrame reziduuri vs. variabile independente şi

– diagrame variabila dependentă vs. variabile independente.

Graficele necesită, de obicei, prelucrări suplimentare pentru a fi interpretate sau raportate.



|  |
| --- |
| Regiunea reziduurilor |



Diagrama reziduuri – variabilă

În figură se dă un exemplu de diagramă reziduuri – variabilă independentă X. Punctele din figură se pot considera într-o regiune de tip bandă orizontală ceea ce nu contrazice ipotezele de normalitate a erorilor. Forma de bandă uniformă reflectă constanţa dispersiei reziduurilor pentru tot domeniul variabilei independente X1. Alte forme de distribuire a reziduurilor duc la concluzii importante pentru adecvanţa modelului în privinţa variabilei independente implicate:

|  |  |  |  |
| --- | --- | --- | --- |
| Forma regiunii | Interpretare |  |  |
|  |  |  | Situaţia "bună". Nu | se contrazic |
|  |  |  | ipotezele de normalitate făcute asupra erorilor |
|  |  |  |  |  |  |
|  |  |  | Dispersia erorilor nu este constantă (se modifică după valorile X) |
|  |  |  | Se poate ca din model să fie omisă o variabilă de gen "Timp". |
|  |  |  |  |  |
|  |  |  | Modelul liniar nu este adecvat în |
|  |  |  | privinţa | variabilei | independente respective |
|  |  |  | . | Se poate | încerca un  |
|  |  |  | introducerea unui termen pătratic. |
|  |  |  | Situaţia poate să apară în urma unei |
|  |  |  | erori de calcul. Practic ar însemna că |
|  |  |  | nu s-a considerat componenta liniară, |
|  |  |  | adică scopul modelului nu a fost atins. |

