

Regresia statistica - mesaje, simboluri, interpretari

Prof.dr. Stefan PECICAN

Catedra de Statistica si Previziune Economica, A.S.E. Bucuresti

În jurul egalității $y_i = a_0 + a_1 x_{1i} + u_i$, ca și a variantei acesteia $y_i = a_0 + \sum_{j=1}^k a_j X_{ji} + u_i$,

($i=1,2,...,n$) s-a scris o întreaga literatură aparținând domeniului analizei de regresie, cu obiectivul centrat pe estimarea parametrilor și, legat de acesta, pe verificarea soluțiilor din diverse puncte de vedere. O serie de pachete de programe (MICROSOFT, SPSS, SHAZAM, ECONOMETRIC NEWS etc.) au preluat „subiectul”, dat fiind și interesul deosebit al economistilor pentru această analiză, oferind solicitatorului soluții, din ce în ce mai numeroase în ultimul timp, marcate prin simboluri (R_y^2 , s.e., JB, D.W., SSR), termeni (coefficients, heteroscedasticity, intercept) nume proprii (Akaike, Chow, Granger).

Cuvinte cheie: regresie, estimare, parametri, verificare ipoteze, erori.

Pentru început, ne propunem să conturăm, prin termeni specifici, problema regresiei: în condițiile în care dispunem de un esanțion (**sample**) de „n” valori privind **variabila** considerată **dependentă** (efect, regresant) - y și **variabila cauzală** sau, mai frecvent, variabilele cauzale (factori, variabile independente, regresori) - $X_1, X_2, ..., X_k$, se aplică metoda celor mai mici pătrate (**least squares** - **LS**) sau alta metodă de calcul numeric, în vederea estimării parametrilor ecuației liniare

$$y_i = a_0 + a_1 X_{1i} + ... + a_k X_{ki} + u_i.$$

Estimațiile $\hat{a}(j)$ permit, în urma atribuirii de valori factorilor, obținerea de **valori ajustate** ale variabilei efect - y ca și obținerea de **prognoze** - y^* . Abaterile dintre valorile ajustate și cele empirice/adeverate sunt denumite **valori reziduale** sau **erori** dacă se referă la y (asadar, $u=y-y$, respectiv, erori (de estimare) dacă se referă la parametru

($e = \hat{a}_{(j)} - a_{(j)}$). Pentru erorile astfel obținute se măsoară împrăștierea prin dispersie (**variance**) - s^2 sau cu ajutorul abaterii medii pătratice (**standard error**) - s sau s.e.

Estimare și teste

Rezultatele estimării (**parameters estimates, coefficients**) $\hat{a}(j)$, ($j=1,2,...,k$) reprezintă o primă categorie de soluții și ele pot fi considerate „fructul” cel mai pretios al analizei de regresie, data fiind interpretarea

lor economică. Fiecare dintre estimatori poate fi afișat alături de abaterea sa medie pătratică (**s.e.**) - indicator care măsoară, sub formă de medie, împrăștierea parametrilor estimați (dacă am dispune de s esanțioane) în raport cu valoarea adevărată a parametrului - a_j . Prin raportarea estima-

torului $\hat{a}(j)$ la abaterea sa medie pătratică ($\hat{a}(j)/s.e.=t$) se obține nivelul „t” - element-cheie pentru verificarea semnificației estimatției (**t-test**). Dacă „t” astfel obținut este superior valorii corespunzătoare din tabelul repartiției Student, pentru un risc α (poate fi 0,05 sau 0,1 sau 0,01) și gradelor de libertate (**degree of freedom** - **d.f.**) $n-m$ (m =număr de parametri în model), considerăm că estimatorul diferă semnificativ de zero, nefiind deci rodul întâmplării ci al acțiunii unui factor important asupra variabilei-efect.

O verificare globală, privind ansamblul estimațiilor, este semnalată de testul F (**F-statistic, ANOVA**). Ea implică o prealabilă determinare a sumei pătratelor abaterilor datorate regresiei $S(y-\hat{y})^2 = SSR$ sau, uneori, ESS) precum și a sumei pătratelor erorilor ($S(y-\hat{y})^2 = SSE$ sau, în unele lucrări, RSS). Nivelul F rezultă astfel $F = [SSR/(m-1)] : [SSE/(n-m)]$. Comparat cu valoarea corespunzătoare din tabelul repartiției Snedecor, rezultă, pentru $F > F(\alpha, m-1; n-m)$, că, pe ansamblu, estimările sunt semnifi-

cative, modelul este valid întrucât modificările datorate factorilor inclusi în model sunt semnificativ superioare modificărilor variabilei-efect datorate erorii (factorilor accidentali, episodici, minori). Aceasta din urma confirmare este completata si de **coeficientul de determinatie (R^2 , R-square)**. Acest indicator este obtinut ca un raport între SSR si SST ($SST = S(y - \bar{y})^2$) si exprima proportia în care factorii inclusi în model contribuie la modificarea variabilei-efect. Varianta R^2 (**adjusted R^2**) exprima acelasi lucru, dar într-o forma comparabila (indicat pentru variante ale modelului cu numar diferit de factori), motiv pentru care, în relatia sa de calcul, fiecare dintre sume este "amendata" cu numarul de grade de libertate corespunzator.

Estimatiile privind parametrii ca si rezultatele primei „baterii” de teste permit formularea unor elemente de analiza si posibilitatea alegerii între variantele posibile ale modelului. Astfel, daca estimatiile obtinute pentru parametri ($\hat{\alpha}(1)$, $\hat{\alpha}(2)$, ... au rezistat testelor t si F iar R este suficient de apropiat de 1 (sau de 100%) si, atentie, daca si urmatoarele teste, la care urmeaza sa ne referim, confirma rezultatele estimarii, atunci $\hat{\alpha}(j)$, ($j > 0$), exprima, în unitati naturale, cu cât se modifica, în medie, variabila-efect (creste pentru $\hat{\alpha} \geq 0$; scade pentru $\hat{\alpha} < 0$) la o crestere cu o unitate naturala a factorului X_j , în conditiile în care ceilalti factori din model sunt considerati constanti. În ceea ce priveste prognoza (predictia), aceasta rezulta, relativ simplu, înlocuind în ecuatie de regresie variabilele cauzale cu valorile preconizate a fi realizate în viitor (valori anticipate, **expected values**) în conditiile mentinerii parametrilor estimati pe baza valorilor esantionului (date din trecut) $y^* = \hat{\alpha}(0) + \hat{\alpha}(1)X_1^* + \dots + \hat{\alpha}(j)X_j^*$. Mentionam ca pentru estimatiile $\hat{\alpha}(j)$, valorile ajustate y , si valorile prognozate - y^* ni se comunica, eventual, si limitele intervalelor de încredere (**lower confidence interval; upper c.i.**) în care se încadreaza, cu o anumita probabilitate ($P = 1 - \alpha$), a fiind apriori stabilit, nivelul adevarat al parametrului α , respectiv nivelul empiric - y sau cel viitor

y^* . Asadar, prognoza y^* reprezinta un nivel-pivot, în jurul caruia, între anumite limite, si cu o probabilitate suficient de mare ($P = 1 - \alpha$) ne asteptam sa se situeze nivelul viitor al variabilei-efect, presupunând ca valorile preconizate pentru factori se vor realiza si nu se vor ivi situatii noi importante (care sa „impuna” noi factori sau modificari ale parametrilor).

În ceea ce priveste optiunea între eventualele variante ale modelului, avem tot interesul sa alegem acea varianta pentru care a rezultat cea mai mare valoare $\hat{R}^2$ ca si F (mai ales daca urmarim obtinerea de prognoze), respectiv, varianta valida în sensul testului F , cu R^2 maxim, pentru care estimările difera semnificativ de zero în sensul testului t (daca urmarim în primul rând analiza procesului).

Cea de a 2-a baterie de teste se refera la teste ce privesc confirmarea ipotezelor metodei celor mai mici patrute (ipoteze legate îndeosebi de caracteristicile erorii de a fi egal împrastiate în raport cu X , normal distribuita, neautocorelata, dar si de calitatea factorilor din model de a fi independenti între ei, corelati liniar cu variabila -efect, cu valori corect evaluate).

Astfel, **heteroscedasticity-test**, ca de altfel si testele **BG**, **BPG**, **White** se refera la verificarea ipotezei privind egala împrastiere (homoscedasticity) a erorii pe diversele segmente de erori ordonate în raport cu evolutia factorului. Primele 2 dintre testele mentionate se bazeaza pe testul F . Astfel, dupa o prealabila ordonare a valorilor $u(i)$ în raport cu valorile crescatoare ale unuia dintre factori, se procedeaza la segmentarea sirului, rezultând 2 siruri cu numar egal de termeni (testul **BG** presupune excluderea în prealabil a valorilor centrale din sir reducând astfel numarul termenilor la $2n/3$). Se procedeaza, în continuare, la calculul raportului F astfel: $F = (Su^2 \text{ în gr.p}) / (Su^2 \text{ în gr.r})$ (în care suma gr.p > suma gr.r) si se compara rezultatul cu $F\text{-tab}(\alpha, (n/2) - m; (n/2) - m)$. Ipoteza este confirmata daca $F < F\text{-tab}$. Testul **BPG** si **White** se refera la egala împrastiere a erorii în raport cu ansamblul factorilor, motiv pentru care

apeleaza la analiza de regresie a erorii în raport cu factorii. Astfel, **testul BPG** presupune ca pe baza erorii din modelul initial sa se obtina valori $g(i)$ astfel: $g(i) = u(i) / S_{u^2}(i)$ dupa care în urma estimarii ecuatiei $g = b(0) + b(1)X_1 + b(2)X_2$ se obtine $h = S(g - g^2) / (m - 1)$ asimilat variabilei hipotat cu care se compara (hipotat $-tab a$, $m - 1$). Daca h este inferior valorii tabelate, ipoteza egalei împrastieri a erorii este confirmata. Testul **White** implica un procedeu similar, numai ca $u = b(0) + b(1)X_1 + b(2)X_2 + b(3)X_1^2 + b(4)X_2^2 + b(5)X_1X_2 + v$, iar R^2 pentru aceasta din urma regresie este multiplicat cu n si rezultatul comparat cu hipotat. Pentru $(nR) < X^2$ (a numar parametrilor. suplimentari-1) ipoteza este confirmata.

Testele simbolizate **DW** si **BG** se refera la ipoteza metodei celor mai mici patrate privind neautocorelarea erorii. Testul **DW** este conceput pentru cazul liniar, privind posibila corelare între $u(i)$ si $u(i-1)$. El implica determinarea valorii $d = S(u(i) - u(i-1))^2 / S u^2(i)$ care se compara cu valorile tabelate de catre Durbin si Watson (valori $d(1)$ si $d(2)$ de coordonate n si k). Se accepta ca pentru valori d situate în jurul lui 2 ipoteza neautocorelării este confirmata; pentru niveluri d apropiate de zero sau de 4, ipoteza se considera ca fiind infirmata. Testul **Wallis** din unele programe reprezinta o varianta a testului DW indicata în cazul seriilor de date influentate de sezonalitate (diferenta de la numator este $u(i) - u(i-4)$). În ceea ce priveste testul **BG** acesta are un grad mai mare de generalizare referindu-se si la distante mai mari decât unu în ce priveste autocorelatia. Se apeleaza la regresia $u(i) = b(1)u(i-1) + \dots + b(p)u(i-p) + v$ pentru care se obtine R^2 , urmând ca nivelul $q = (n-p) \cdot R^2$ sa se compare cu hipotat $tab(a, p)$. Daca q se situeaza sub valoarea tabelata, ipoteza neautocorelării erorii este confirmata.

Testul **JB** se refera la ipoteza conform careia eroarea $u(i)$ urmeaza legea de repartitie normala. Testul implica determinarea prealabila a coeficientului de asimetrie (s) (**skewness**) si a celui de aplatare (k)

(**kurtosis**) pentru repartitia erorii. Se obtine nivelul $JB = n[(s^2/6) + ((k-3)^2/24)]$ caruia, i se cauta un corespondent (valoarea cea mai apropiata) în tabelul repartitiei hipotat pentru $df=2$. Odata gasita o astfel de valoare, preluam nivelul α din capul respectivei coloane si daca $P=1-\alpha$ este suficient de aproape de 1, consideram ipoteza confirmata. Indicatorul **VIF** se refera la ipoteza privind independenta factorilor (în sensul ca evolutiile lor nu prezinta analogii semnificative). Definit astfel: $VIF = 1/(1 - r_{X_1X_2}^2)$, indicatorul exprima în ce masura dispersia în esantionul de date este accentuata de prezenta multicolinearitatii (termen care se refera la analogii alarmante între evolutiile paralele ale factorilor). Se considera ca pentru $VIF < 6$, ipoteza independentei factorilor este confirmata si multicolinearitatea nu afecteaza estimările.

Elemente de analiza

Rezultatele testelor din aceasta sectiune sunt importante în vederea aprecierii calitatii estimatorilor si implicit utilitatii lor pentru analiza si prognoza. În acest sens, se recomanda ca ipotezele metodei de estimare sa fie confirmate. Neconfirmarea acestora afecteaza îndeosebi precizia (eficienta) estimatiilor $\hat{a}(j)$ ($j=1..k$), acestea încadrându-se în intervale de încredere mult prea largi iar testul „t” nu mai este întotdeauna revelator. Astfel, autocorelarea erorii (semnalata de testele **DW**, **BG**) atrage neîncrederea în analiza de regresie („spurious regression”) chiar daca R -squared si testul „t” prezinta valori satisfactoare. Inegala împrastiere a erorii (heteroscedasticity) afecteaza atât precizia parametrilor estimati cât si precizia prognozei. Multicolinearitatea (semnalata de **VIF**) este deosebit de pagubitoare pentru analiza (nu întotdeauna si pentru prognoza) întrucât maresta imprecizia parametrilor asa încât interpretarea acestora devine superflua.

O consecinta a interesului pentru confirmarea ipotezelor metodei de estimare este reprezentata si de varietatea de metode si variante metodologice de estimare. Astfel, metoda celor mai mici patrate în varianta

obisnuita (**LS** sau **OLS**) este recomandata în situatii în care dependentă la care se referă este liniară, ecuația de regresie este unică, ipotezele metodei de estimare nu ridică probleme (sau nu ne preocupă). În cazurile în care ipoteza egalei împrastieri a erorii (homoscedasticity) este infirmată se recomandă varianta ponderată a metodei (**WLS**) sau metoda variabilelor instrumentale (**IV**) iar unii autori recomandă varianta generalizată a metodei celor mai mici pătrate (**GLS**). Așa numita regresie **RIDGE** este recomandată în vederea atenuării/eliminării multicolarității. În cazul modelelor cu ecuații simultane (modelul poate include zeci sau sute de ecuații de regresie) se recomandă metoda celor mai mici pătrate în 2 stadii (**2-LS**) sau în 3 stadii (**3-LS**) pentru a evita corelarea erorii cu variabilele cauzale. Deseori se fac referiri la alte metode cum ar fi cea a verosimilității maxime (**ML**).

O grupă distinctă de simboluri se referă la unele dintre criteriile de alegere între variantele modelului de regresie. Frecvent invocat în acest sens este **criteriul AIC** pentru care se calculează

$$AIC(g) = \ln(Su^2/n(g)) + (2k(g)/n(g))$$

în cazul în careia dintre cele g „variante pentru a stabili varianta cu AIC minim care va fi cea reținută. Indicatorul AIC este o mărime comparabilă și semnalează varianta cu gradul de precizie mai mare.

Simbolul **RESET** (Regression Specification Error Test) se referă la verificarea corectei elaborări a modelului prin introducerea între factori a valorilor ajustate, obținute în prealabil pentru variabila efect, așa încât $Y = b_0 + b(1)X_1 + b(2)X_2 + b(3)y^2 + b(4)y^3 + v$. După estimare, se determină un nou R^2 pe baza căruia se calculează

$$F = [(R^2_{nou} - R^2) / m^*] / [(1 - R^2_{nou}) / (n - m)],$$

unde m^* = număr de noi parametri. Dacă $F < F_{tabelat}(a, m, n - m)$ modelul este validat.

Granger-test apelează, de asemenea, la testul F și este recomandat în cazurile în care oportunitatea introducerii unui nou factor (W) urmează să fie verificată (mai ales în cazurile în care datele sunt urmărite în timp privind procese inertiabile). Dacă no-

tăm cu $v(i)$ erorile din varianta de model în care apare și variabila W , se ajunge la raportul $F = (Su^2 - Sv^2) / [(Sv^2) / (n - m - 1)]$. Pentru $F < F_{tab}(a, n - (m + 1))$ oportunitatea introducerii variabilei W , în postura de factor, este confirmată.

Testul **Chow** destinat verificării stabilității estimatorilor $\hat{a}(j)$ în cazul apariției de noi valori pentru variabilele modelului, prezintă o deosebită importanță îndeosebi pentru prognoze. Testul F este și de astă dată invocat, într-o manieră asemănătoare precedentei verificări, iar nivelul $F < F_{tab}(a, k, n - 2k)$ confirmă stabilitatea parametrilor.

În ce privește **J-test**, acesta se referă la o modalitate de a alege între 2 variante ale modelului, printr-un „transfer” al variabilei-efect ajustate. Testul poate fi considerat un fel de „arbitru” în cazurile în care precedentele criterii (R^2, F, AIC) oferă informații contradictorii în ce privește cea mai bună alternativă dintre 2 variante ale modelului. Astfel, în cazul variantelor A și B , se procedează la introducerea în varianta A a unui nou factor $y(B)$. Dacă aceasta suplimentare îmbunătățește performanțele variantei A (grad de determinare, semnificație) se consideră că această variantă este „fragilă”, perfectibilă și, ca urmare, este preferată varianta B (decizie „întărită” și de insensibilizarea performanțelor variantei B de către un transfer simetric, a lui $y(A)$ între factorii variantei B).

Opțiunea între variante este recomandată atât în vederea argumentării analizei cât și în vederea pregătirii terenului pentru obținerea de prognoze de calitate.

Bibliografie

- Griffiths, W., Carter-Hill, R., Judge, G. - Learning and Practicing Econometrics, Wiley, 1992
- Halpern, L. - Real exchange rate policy in Hungary, în Economy of Transition, 4-1996
- Pecican, E., St. - Piața valutară, bănci și econometrie, Editura Economica, București, 2000.