

C U P R I N S

Introducere.....	2
1. Analiza regresională. Generalități	3
2. Metoda celor mai mici pătrate.....	8
3. Metoda celor mai mici pătrate, exemplu realizat	9
4. Evaluarea semnificației ecuației de regresie liniară și a coeficienților ei	11
5. Modelul de regresie liniară multifactorial	17
6. Multicoliniaritatea și atenuarea ei	22
7. Remedierea multicolinearității	24
8. Corelarea în serie (autocorelarea)	28
9. Consecințele corelării în serie	31
10. Eteroschedasticitatea	36
11. Remedierea eteroschedasticității	41
12. Specificația: alegerea variabilelor independente relevante	45
13. Specificația ecuației de regresie: alegerea formei funcționale	52
14. Date economice	59
15. Siseme de ecuații econometrice	63
16. Problema de identificare a modelului sub forma sa redusă.....	68
R E F E R I N Ț E	74
A N E X E	75

Introducere

Materialul prezentat în continuare a servit ca bază a prelegerilor ținute pe parcursul câtorva ani pentru studenții și masteranzii de diferite specialități (matematică și informatică, relații internaționale, management). Obiectivele propuse se referă la expunerea materialului în așa mod ca studenții să obțină însușiri practice la utilizarea reușită a instrumentarului de evaluare cantitativă a proceselor și fenomenelor economice. Accentele au fost puse pe una din abordările econometrice, și anume, pe analiza regresională, ea fiind mai frecvent utilizată în cercetările economice și la modelarea activității economice la diverse niveluri.

O atenție specială se pune pe punctarea etapelor ce precedă realizarea unei regresii, începând de la fundamentarea teoretică a evenimentului cercetat, stabilirea variabilei dependente și a variabilei (variabilelor) relevante independente, alegerea formei funcționale potrivite, și, în sfârșit, colectarea datelor de încredere, ca apoi să fie verificate ipotezele care asigură aplicarea metodei celor mai mici pătrate. În cazul în care una sau mai multe ipoteze sunt violate, să se determine metoda de estimare a coeficienților ecuației de regresie aprobate.

Atunci, când se trag concluzii de rigoare privind metoda de utilizat, se purcede la lansarea ecuației de regresie, folosind Softul specializat, cum ar fi Eviews, sau, în lipsa acestuia, se activează utilitatea Data Anayses din Excell, care oferă posibilitatea lansării modulului Regression. După lansarea ecuației de regresie se efectuează analiza rezultatelor obținute în vederea semnificației atât ecuației în întregime, cât și a fiecărei variabile independente în parte. Se confruntă statisticile Fiser și Durbin-Watson, t-statisticile calculate cu aceleași tabele corespunzătoare gradului de libertate corespunzător și nivelului de semnificație selectat. În cazul în care se confirmă ipotezele respective de luare a deciziilor se trece, în caz de necesitate, la etapa de previziune. Se calculează intervalele de încredere pentru pronosticul punctifer și se stabilește valoarea prognozată pentru variabila dependentă în funcție de valoarea variabilei sau variabilelor independente examinate.

Se propun procedee de eliminare a fenomenelor de multicolaritate, autocorelare în serie și eteroscedasticitate, care pot fi realizate de sinestătător, fără ca să se apeleze la ajutorul Softului specializat.

Și la finele cursului se examinează sisteme de ecuații econometrice care într-un mod mai adecvat descriu procesele și fenomenele economice. Se cercetează problema de identificare a sistemului de ecuații econometrice simultane. O atenție separată se acordă sistemului de ecuații simultane cu identități, care foarte frecvent se întâlnește la modelarea proceselor economice.

Acest curs de lecții este susținut de lucrări practice ce se referă la estimarea ecuațiilor de regresie pentru funcția de producere și pentru cererea la bunuri și servicii de import. La fel sunt prezentate exemple de soluționat pentru lichidarea fenomenelor de multicolaritate, autoregresie, eteroscedasticitate.

Sunt prezentate referințe bibliografice care au constituit baza acestui curs și au servit ca sursă de date, exemple, materiale ilustrative.

1. Analiza regresională. Generalități

1.1. Econometria: definiția și utilizarea.

Econometria poate fi definită ca analiza cantitativă a fenomenelor economice reale. Profesioniștii în domeniu definesc econometria sub forma unui set de tehnici fascinante care permit *măsurarea și analiza fenomenelor economice și previziunea tendințelor economice pe viitor*. Econometria constituie o definiție formală și un conținut vast. Econometria, literal, înseamnă “măsurări economice” și ea se ocupă de măsurarea cantitativă și de analiza economiei reale și a fenomenelor ce țin de business. Ea reprezintă o tentativă de a măsura economia reală și de a construi un pod deasupra prăpastiei ce desparte teoria economică și activitatea de business reală. Econometria ne permite să examinăm datele ce caracterizează firmele din lumea reală și să comăsurăm acțiunile acestor firme cu alți factori, cum ar fi acțiunile consumatorilor și a guvernelor.

Econometria are trei direcții de bază de utilizare:

1. descrierea economiei reale;
2. testarea ipotezelor referitor la teoria economică;
3. pronosticarea activității economice pe viitor.

Cea mai simplă direcție de utilizare a econometriei este descrierea. Econometria ne permite să evaluăm activitatea economică; ea ne permite să introducem numere în ecuații care în prealabil conțineau numai simboluri abstracte. De exemplu, cererea consumatorului pentru un anumit bun poate fi prezentată ca o relație dintre cantitatea cerută (C), prețul bunului (P), prețul bunurilor de substituție (P_s) și venitul disponibil (Y_d). Pentru majoritatea bunurilor relația dintre consum și venitul disponibil se presupune a fi pozitivă, deoarece creșterea venitului disponibil se asociază cu creșterea consumului de bunuri. Econometria ne permite să estimăm această relație în baza consumului, venitului disponibil și prețurilor înregistrate în trecut.

Cu alte cuvinte, o relație funcțională

$$C = f(P, P_s, Y_d) \quad (1.1)$$

Se transformă într-o relație explicativă de felul:

$$C = -60,5 - 0,45 * P + 0,12 * P_s + 12,2 * Y_d \quad (1.2)$$

Această prezentare ne oferă un tablou mult mai specific și descriptiv. Să comparăm ecuațiile (1) și (2). Expresia (1) ne comunică: consumul se așteaptă să crească “odată” cu creșterea venitului disponibil. În timp ce ecuația (2) ne permite să așteptăm o creștere de o cantitate specifică de 12,2 unități la fiecare unitate de creștere a venitului disponibil. Cifra 12,2 se numește coeficientul regresiei estimat. Și abilitatea econometriei de a aprecia acest coeficient este valoarea ei.

Al doilea, și probabil cel mai uzual mod de utilizare a econometriei, este testarea ipotezelor. De exemplu, putem testa, va fi bunul examinat un bun normal (pentru care cererea crește odată cu creșterea venitului disponibil). La prima vedere, se pare că aceasta ipoteză poate fi susținută întrucât semnul coeficientului este pozitiv, însă “semnificația statistică” a acestei estimări urmează a fi investigată înainte de a justifica o atare concluzie. Folosirea econometriei în testarea ipotezelor este, probabil, cea mai importantă funcție.

A treia, și cea mai dificilă modalitate de utilizare a econometriei, este pronosticarea sau previziunea: ce e probabil să se întâmple în trimestrul următor, în anul viitor ori mai departe pe viitor. De exemplu, economiștii folosesc modelele econometrice pentru a face previziuni pentru așa variabile ca: volumul vânzărilor, volumul veniturilor, Produsul Intern Brut, rata inflației etc. Precizia acestor previziuni depinde în ceea mai mare măsură de gradul cu care trecutul dirijează viitorul. De exemplu, vom presupune că președintele companiei, care propune produsul modelat în ecuația (1), dorește să decidă majorarea prețurilor sau păstrarea lor la același nivel. El va pronostica volumul vânzărilor cu și fără creșterea prețurilor ceea ce îl va

ajuta în luarea acestei decizii. În acest mod econometria poate fi utilizată nu numai pentru previziune, dar și pentru analiza politicilor.

1.2. Abordări econometrice de alternativă

Pentru obținerea unui tablou mai reușit al abordării posibile vom puncta etapele necesare de efectuat pentru orice investigație cantitativă:

- a) specificarea modelului sau relației de studiat;
- b) colectarea datelor necesare pentru estimarea modelului;
- c) estimarea modelului cu ajutorul datelor.

Etapele a) și b) sunt similare în investigațiile cantitative iar tehnicile utilizate la etapa c) - estimarea modelului diferă de la o disciplină la altă disciplină. Alegerea tehnicilor pentru evaluarea modelului, în baza unui set de date specifice, de regulă, se referă la “arta” econometrică. Există diferite abordări alternative pentru evaluarea unei și aceeași ecuații, și fiecare abordare poate oferi rezultate ce diferă unul de altul.

În continuare ne vom referi la abordarea ce ține de analiza regresională. Însă important e ca fiecare econometrician să conștientizeze: regresia este numai una din tehnicile folosite în estimarea econometrică.

1.3. Ce este analiza regresională?

Analiza regresională este utilizată pentru efectuarea estimărilor cantitative a relațiilor economice care în prealabil aveau loc doar din punct de vedere pur teoretic. Pentru prezicerea direcției schimbărilor este necesar de cunoscut teoria economică și caracteristicile generale a produsului în examinare (de exemplu, dependența volumului de vânzări a discurilor floppy în funcție de preț). Iar pentru prezicerea schimbărilor în cantitate, sunt necesare un set de date și o metodă de estimare a relației propuse. În econometrie cea mai frecvent utilizată metodă de estimare a acestor relații este analiza regresională.

1.4. Variabile dependente, variabile independente, justificare

Analiza regresională este o tehnică (o metodă) statistică care încearcă să “explice” schimbările unei variabile, variabile dependentă (de explicat) ca funcție de schimbările altei variabile sau set de variabile, așa numite variabile independente (sau explicative), prin evaluarea unei singure ecuații, cum ar fi $C = f(P, P_s, Y_d)$. Aici C este variabilă dependentă (de explicat), iar P, P_s, Y_d – variabile independente (explicative). Analiza regresională este un instrument bine venit pentru economiști deoarece majoritatea afirmațiilor economice pot fi formulate într-o formă funcțională dintr-o singură ecuație.

În economie și business majoritatea afirmațiilor sunt de genul cauză – efect: dacă prețul bunurilor crește cu o unitate, atunci volumul cererii descrește în mediu cu câteva unități în dependență de elasticitatea cererii față de preț. Prin analogie, dacă volumul capitalului utilizat crește cu o unitate, atunci volumul de producție va crește cu câteva unități, în funcție de așa numită productivitate marginală a capitalului. Afirmațiile de acest gen stabilesc relații de tip dacă – atunci, cauzale care postulează logic că schimbările în variabila dependentă sunt cauzate de schimbările într-un număr specificat de variabile independente.

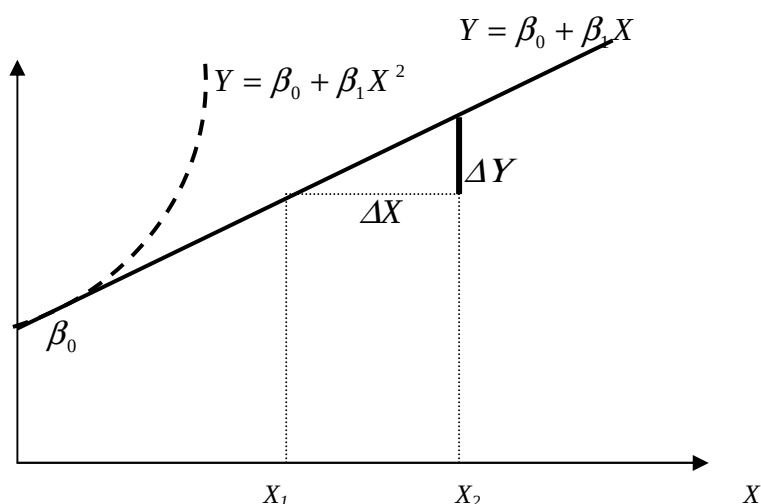
Întrucât multe relații economice sunt după natura sa cauzale, rezultatul regresiei nu poate conta pe semnificația lor, nu poate confirma cauzalitatea. Analiza regresională poate efectua testarea semnificației estimării relațiilor cantitative. Fundamentarea cauzalității relațiilor economice trebuie să includă un suport teoretic solid și un bun simț.

1.5. Modelul liniar de regresie de o singură ecuație

$$Y = \beta_0 + \beta_1 X \quad (1.3)$$

este cel mai simplu model de regresie de o singură variabilă. Prin ecuația (3) se afirmă că variabila dependentă (endogenă) Y este o funcție lineară de o singură variabilă independentă (exogenă) X . Modelul este de o singură ecuație deoarece nu mai sunt alte ecuații pentru Y ca funcție de X (sau de alte variabile). Modelul este linear deoarece sub forma sa grafică reprezintă o linie dreaptă, dar nu o curbă.

β_0, β_1 sunt coeficienții (sau parametrii) care determină coordonatele liniei drepte în orice punct. β_0 este constantă sau termenul de intersecție, el indică valoarea lui Y pentru X egal cu zero. β_1 este coeficientul de înclinație, și el indică valoarea cu care se va schimba Y , când X se schimbă cu o unitate. Coeficientul unghiular β_1 demonstrează reacția (răspunsul) lui Y față de schimbările în X . Pentru a explica și a prezice schimbările în variabila dependentă, ce e obiectivul major în evaluarea relațiilor comportamentale, accentul principal se pune pe coeficientul de înclinație cum ar fi β_1 . Pe desen, de exemplu, dacă X o avut să crească de la X_1 până la X_2 , valoarea lui Y conform ecuației (3) va crește de la Y_1 la Y_2 . În modelele de regresie liniară răspunsul valorilor pronosticate Y la schimbările în X este determinat de o constantă, egală cu coeficientul de înclinație: $\beta_1 = (Y_2 - Y_1) / (X_2 - X_1) = \frac{\Delta Y}{\Delta X}$.



Este necesar să se facă distincție dintre ecuațiile liniare în variabile și ecuațiile liniare în parametri (coeficienți), deoarece regresia liniară trebuie să fie lineară în coeficienți, însă nu neapărat lineară în variabile. Ecuația $Y = \beta_0 + \beta_1 X$ este lineară în variabile, grafic reprezentând o linie dreaptă, în timp ce ecuația $Y = \beta_0 + \beta_1 X^2$ nu este lineară în variabile, deoarece reprezintă grafic o curbă pătratică dar nu o linie dreaptă.

Ecuația este lineară față de coeficienți (parametri) numai în cazul dacă parametrii apar sub forma cea mai simplă – ei sunt ridicați la putere (nu mai mare decât unu), nu se înmulțesc și nu se împart la alți coeficienți și nu fac parte din careva funcții (cum ar fi $\log$ sau $\exp$). Ecuația (3) este lineară în coeficienți, dar $Y = \beta_0 + X^{\beta_1}$ nu este lineară în coeficienți β_0, β_1 , deoarece nu există o transformare a ecuației care s-o aducă la forma lineară. În general, din toate ecuațiile posibile cu o singură variabilă explicativă, numai funcția sub formă generală $f(Y) = \beta_0 + \beta_1 f(X)$ este lineară în coeficienți β_0, β_1 .

Toate cele expuse sunt importante deoarece la aplicarea tehnicii regresiei liniare ecuația necesită să fie lineară în coeficienți. Analiza regresională liniară poate fi aplicată la ecuații care nu sunt liniare în variabile, dar pot fi prezentate în așa mod ca să fie liniare în coeficienți.

1.6. Termenul erorii stocastice. Eroarea de specificație

Schimbările în variabila Y pot fi cauzate nu numai de schimbări în variabila independentă X dar și de schimbări ce parvin din alte surse. Aceste schimbări adiționale apar parțial în urma

omiterii variabilelor explicative ($X_1, X_2, X_3, \dots$), și chiar dacă aceste variabile vor fi introduse în model, Y continue să fie influențat de schimbări care pur și simplu nu pot fi explicate cu ajutorul modelului. Probabil aceste schimbări pot parveni din surse, cum ar fi: 1) influențe omise, 2) erori de măsurare, 3) formă funcțională incorectă sau 4) pur și simplu din cauza unor evenimente aleatoare și complectament imprevizibile.

Econometricienii admit existența unei atare variații esențiale inexplicabile («erori») prin introducerea explicită a unui termen stocastic (aleator) în modelul de regresie. Termenul erorii stocastice este un termen, care se include în ecuația de regresie pentru a reflecta toate schimbările în Y ce nu pot fi explicate prin variabila X . Termenul de eroare, de regulă, este notat prin ε deși și alte simboluri (cum ar fi u sau v) se utilizează frecvent. Includerea termenului de eroare stocastică (sau eroare de specificație) în ecuația (3) rezultă cu ecuația de regresie tipică

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1.4)$$

Ecuația (4) este compusă din două componente: componenta deterministă și componenta stocastică. Expresia $\beta_0 + \beta_1 X$ se numește componentă deterministă, deoarece ea indică valoarea lui Y care este determinată de valorile date a lui X , care se presupune a fi nonstocastice. Această componentă deterministă poate fi prezentată și ca valoarea așteptată a lui Y în conformitate cu X dat, valoarea medie a Y -or asociată cu o valoare distinctă a lui X . Partea deterministă a ecuației poate fi notată prin

$$E(Y / X) = \beta_0 + \beta_1 X \quad (1.5)$$

Spre regret, în realitate valoarea observată a lui Y e puțin probabil să fie egală cu valoarea deterministă așteptată $E(Y / X)$. Ca rezultat elementul stocastic necesită a fi inclus în ecuație

$$Y = E(Y / X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon \quad (1.6)$$

Eroarea de specificație este cauzată cel puțin de patru surse, care produc schimbări în variabila Y diferite de acelea determinate de variabila X .

- Multe influențe minore sunt omise din ecuație (de exemplu, din cauza inutilității datelor).
- Este, de fapt, imposibil de a evita erori de măsurare cel puțin într-o variabilă din ecuație.
- Ecuația teoretică specificată trebuie să aibă altă formă funcțională diferită de aceea aleasă pentru regresie. De exemplu, ecuația specificată trebuie să fie neliniară în variabile pentru regresia liniară (sau vice-versa).
- Toate încercările de a generaliza comportamentul uman trebuie să conțină careva cantități de variații imprevizibile sau pur și simplu aleatoare.

1.7. Extinderea notațiilor

Vom extinde notațiile ca să includem referințe la un număr de observații stabilit și ca să avem posibilitatea de a introduce mai multe variabile independente. Atunci unica ecuație de regresie lineară poate fi scrisă sub formă $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ($i=1,2,\dots,n$), unde

Y_i - observația i a variabilei dependente;

X_i - observația i a variabilei independente;

ε_i - observația i a erorii specificate;

β_0, β_1 - parametrii regresiei;

n - numărul de observații.

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1,$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2,$$

$$Y_3 = \beta_0 + \beta_1 X_3 + \varepsilon_3,$$

$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n.$$

În cazul mai multor variabile independente ecuația de regresie are forma

$$Y_i = \beta_0 + \sum_{m=1}^k \beta_m X_{mi} + \varepsilon_i, \quad (i=1,2,\dots,n),$$

Y_i - observația i a variabilei dependente;

X_{mi} - observația i a variabilei independente m ;

ε_i - observația i a erorii specificate;

β_0, β_m - parametrii regresiei, ($m=1,2,\dots,k$);

n - numărul de observații,

k - numărul de variabile independente.

1.8. Ecuația regresiei evaluată

Odată ce s-a decis specificația ecuației, ea trebuie evaluată, este necesar să se determine parametrii. Aceasta versiune a ecuației de regresie «adevărată» se numește ecuație de regresie estimată și se obține din observațiile Y_s, X_s . În timp ce ecuația adevărată este pur teoretică în natură

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (i=1,2,\dots,n). \quad (1.7)$$

Ecuația de regresie estimată conține numere reale în ea

$$\hat{Y}_i = 103,4 + 6,38X \quad (1.8)$$

Valorile observate a lui X și Y se folosesc la determinarea parametrilor estimați 103,4 și 6,38. Acești parametri s-au folosit la determinarea $\hat{Y}_i$ - valorilor estimate a lui Y_i .

Vom cerceta diferența dintre ecuația de regresie «adevărată» și ecuația de regresie estimată. În primul rând, coeficienții teoretici ai ecuației de regresie β_0, β_1 în ecuația (7) se înlocuiesc cu coeficienții estimați de tipul 103,4 și 6,38 din ecuația (8). Nu e posibil să cunoaștem valorile parametrilor ecuației de regresie «adevărată», de aceea în locul lor se calculează estimările acestor coeficienți folosind datele cunoscute referitor la observațiile variabilelor dependentă și independentă. Parametrii de regresie estimați, notați prin $\hat{\beta}_0, \hat{\beta}_1$, reprezintă o aproximație empirică reușită, obținută din datele observațiilor Y_s, X_s .

În expresia $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$: pentru fiecare set de observații se vor calcula diferite seturi de parametri de regresie estimați. $\hat{Y}_i$ reprezintă valorile estimate a lui Y_i pentru observația i și sunt calculate prin intermediul ecuației de regresie estimată.

Diferența dintre valorile estimate a variabilei dependente ($\hat{Y}_i$) și valoarea reală a variabilei dependente (Y_i) este definită drept reziduală

$$Y_i - \hat{Y}_i = u_i \quad (1.9)$$

Vom nota distincția dintre variabila reziduală u_i și eroarea de specificație $\varepsilon_i = Y_i - E(Y_i / X_i)$.

(1.10)

Variabila reziduală este diferența dintre valorile Y_i observate și cele estimate prin ecuația de regresie $\hat{Y}_i$, în timp ce eroarea de specificație (stocastică) este diferența dintre Y_i observat și ecuația de regresie «adevărată» (valoarea așteptată a variabilei Y). Cu alte cuvinte eroarea stocastică este o valoare teoretică care nici odată nu poate fi observată însă variabila reziduală este o valoare reală care se calculează pentru fiecare observație de fiecare dată când regresia este lansată. Întrădevăr, majoritatea tehnicilor regresionale nu numai evidențiază reziduurile, dar selectează acele valori β_0, β_1 , care le asigură un nivel cât e posibil de mic. Cu cât e mai mică valoarea variabilei reziduale cu atât mai apropiate vor fi valorile estimate $\hat{Y}_s$ de acelea observate

Y_s . O altă cale de a exprima ecuația de regresie estimată constă în combinarea ecuațiilor (2) și (3) și obținerea expresiei $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + u_i$.

2. Metoda celor mai mici pătrate

2.1 Estimarea modelului liniar simplu (de două variabile) folosind metoda celor mai mici pătrate.

Ipoteze

- Modelul este linear în raport cu β_k , ($k=0,1$), ε_i , ($i=1,\dots,n$).
- Valorile X_i sunt considerate fără erori de observație sau de măsurări permanente.
- Termenul erorii stocastice ε_i este normal distribuit de media nulă $E(\varepsilon_i) = 0$, ($i=1,\dots,n$) (în medie modelul este bine specificat).
- Perturbația este omoscedastică: $E(\varepsilon_i^2) = \sigma_i^2 = \sigma = \text{const}$, variația erorilor σ_i^2 este constantă (cu alte cuvinte termenul erorii stocastice este independentă de evoluția variabilei explicative, ceea ce înseamnă ca dispersiile calculate pentru diverse segmente de X_i , nu diferă între ele).
- $E(\varepsilon_i, \varepsilon_j) = 0$, $i, j = 1, 2, \dots, n, i \neq j$. Valorile erorilor stocastice nu sunt autocorelate (sunt independente între ele). Valorile consecutive ale erorilor stocastice nu depind una de alta.
- $E(x_i, \varepsilon_i) = 0$. Valorile erorii stocastice sunt independente de variabila explicativă.

Așadar, în urma observațiilor statistice, avem serii de observații. Problema constă în determinarea parametrilor.

Metoda celor mai mici pătrate (M.C.M.M.P.), în condițiile verificării ipotezelor enunțate, asigură obținerea estimatorilor de maximă verosimilitate, nedeplasați, concordați și eficienți (cu dispersia minimă).

Estimațiile sunt nedeplasate. Aceasata înseamnă că $E(\hat{\beta}_k) = \beta_k$, ($k = 0,1$), prin urmare, estimațiile coeficienților, obținute cu ajutorul metodei celor mai mici pătrate (M.C.M.M.P.), sunt centrate în jurul mulțimii de valori ai coeficienților adevărați.

- Estimațiile sunt efective. Dispersia coeficienților evaluați în jurul valorilor adevărate ale coeficienților este cea mai compactă distribuție, care este posibilă în condițiile dispersiilor nedeplasate. Nici una din metodele liniare existente de estimare a coeficienților nu asigură o dispersie mai mică pentru fiecare din coeficienții estimați decât M.C.M.M.P. Estimațiile sunt de natură BLUE - Best Linear Unbiased Estimators (Teorema Gauss-Marcov).
- Estimațiile sunt concordate. Ceea ce înseamnă, că măbind pînă la infinit numărul de observații, obținem estimări care tind spre valorile adevărate ale coeficienților de regresie. Odată cu creșterea numărului de observații, dispersia devine mai mică, și fiecare estimație tinde spre ceea adevărată.
- Estimațiile sunt de o veridicitate maximă. $\hat{\beta}_s$ este normal distribuit, $N(\hat{\beta}, \text{VAR}[\hat{\beta}])$

Pentru estimarea parametrilor vom aplica metoda celor mai mici pătrate, care se exprimă în modul ce urmează:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n u_i^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Condițiile de ordinul întâi (condiții necesare) se înscriu ca:

$$\partial \sum_{i=1}^n u_i^2 / \partial \beta_0 = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0.$$

$$\partial \sum u_i^2 / \partial \beta_1 = -2 \sum_{i=1}^n (\beta_1 x_i^2 - x_i y_i + \beta_0 x_i) = 0.$$

De unde rezultă sistema de ecuații normale:

$$\sum_{i=1}^n y_i = n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i, \quad (2.1)$$

$$\sum_{i=1}^n x_i y_i = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2. \quad (2.2)$$

$$\text{Împărțind (1) la } n \text{ și rezolvând în raport cu } \hat{\beta}_0, \text{ obținem } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad (2.3)$$

substituind această expresie în (2) îl putem afla pe $\hat{\beta}_1$, $\sum_{i=1}^n x_i y_i = \sum_{i=1}^n x_i (\bar{Y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 \sum_{i=1}^n x_i^2$,

$$\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{Y} = \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{X} \right), \text{ de unde } \hat{\beta}_1 = \frac{\left(\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{Y} \right)}{\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{X} \right)}, \text{ vom împărți numitorul și}$$

numărătorul la n :

$$\hat{\beta}_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} = \frac{Cov(X, Y)}{\sigma_X^2}, \quad \hat{\beta}_0 = \bar{Y} - \bar{X} \frac{Cov(X, Y)}{\sigma_X^2}. \quad (2.4)$$

Așadar, am obținut parametrii $\hat{\beta}_0, \hat{\beta}_1$, care sunt estimatorii pentru β_0, β_1 . Condiția suficientă pentru existența minimului funcționalului este:

$$\frac{\partial^2 \sum_{i=1}^n u_i^2}{\partial \beta_i^2} > 0, i = 0, 1; \quad \left| \begin{array}{cc} \frac{\partial^2 \sum_{i=1}^n u_i^2}{\partial \beta_0^2} & \frac{\partial^2 \sum_{i=1}^n u_i^2}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \sum_{i=1}^n u_i^2}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \sum_{i=1}^n u_i^2}{\partial \beta_1^2} \end{array} \right| > 0 \quad (2.5)$$

$$\frac{\partial^2 \sum_{i=1}^n u_i^2}{\partial \beta_1 \partial \beta_0} = \sum_{i=1}^n x_i \quad \frac{\partial^2 \sum_{i=1}^n u_i^2}{\partial \beta_0^2} = n \quad \frac{\partial^2 \sum_{i=1}^n u_i^2}{\partial \beta_1^2} = \sum_{i=1}^n x_i^2$$

3. Metoda celor mai mici pătrate, exemplu realizat

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}; \quad \hat{\beta}_1 = \frac{Cov(X, Y)}{\sigma_X^2}; \quad Cov(X, Y) = \overline{XY} - \bar{X}\bar{Y}; \quad \sigma_X^2 = \overline{X^2} - \bar{X}^2; \quad \hat{\beta}_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}.$$

Parametrul β_1 se numește coeficient de regresie. Valoarea lui denotă valoarea medie cu care se schimbă variabila de explicat atunci când variabila explicativă se schimbă cu o unitate.

Parametru β_0 denotă valoarea lui Y când $X = 0$. Atunci când valoarea explicativă nu poate primi valoarea 0 explicația precedentă este lipsită de sens. Parametru β_0 poate să nu aibă nici un sens economic. Încercarea de interpretare economică a parametrului β_0 , mai ales atunci când el e mai mic decât 0, $\beta_1 < 0$ poate aduce la situație absurdă.

Este posibil să interpretăm numai semnul pe lângă parametrul β_0 . Dacă $\beta_0 > 0$, atunci schimbarea relativă a variabilei de explicat se petrece cu ritmuri mai reduse decât schimbarea relativă a factorului. Cu alte cuvinte variația rezultatului este mai mică decât variația factorului –

coeficientul de variație după factorul X este mai mare decât coeficientul de variație pentru variabila de explicat Y : $V_X \succ V_Y$. Vom demonstra acest fapt (fenomen) pornind de la comparația schimbărilor relative a variabilelor de explicat Y și explicative - X :

$$\frac{dY}{Y} \prec \frac{dX}{X} \Rightarrow \frac{dY}{dX} \prec \frac{Y}{X}; \quad \frac{\beta_1 dX}{dX} \prec \frac{\beta_0 + \beta_1 X}{X}; \quad \beta_1 X \prec \beta_0 + \beta_1 X \Rightarrow 0 \prec \beta_0.$$

Vom examina un fenomen. Un grup de întreprinderi care produc același produs sunt supuse analizei din punct de vedere a cheltuielilor de producere conform funcției $Y = \beta_0 + \beta_1 X + \varepsilon$. Informația necesară pentru evaluarea parametrilor β_0, β_1 vom prezenta-o sub forma de tabel. Rezultatele obținute în baza efectuării cercetărilor confirmă că numărul observațiilor «n» necesită a fi de 6-7 ori mai mare decât numărul parametrilor pe lângă variabilele independente (explicative) X .

№ întrepr.	Volum de prod. (mii unit.) (X)	Chelt. de prod. (mln. lei) (Y)	$Y \cdot X$	X^2	Y^2	$\hat{Y}_X$
1	1	30	30	1	900	31,6
2	2	70	140	4	4900	67,9
3	4	150	600	16	22500	141,6
4	3	100	300	9	10000	104,7
5	5	170	850	25	28900	178,4
6	3	100	300	9	10000	104,7
7	4	150	600	16	22500	141,6
Îtoto	22,00	770,00	2820,00	80,00	99700,00	770,50

Vom înscrie sistemul de ecuații normale:

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^7 X_i = \sum_{i=1}^7 Y_i \\ \beta_0 \sum_{i=1}^7 X_i + \beta_1 \sum_{i=1}^7 X_i^2 = \sum_{i=1}^7 X_i Y_i \end{cases}, \quad \begin{cases} 7\beta_0 + 22\beta_1 = 770 \\ 22\beta_0 + 80\beta_1 = 2820 \end{cases}$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}; \quad \hat{\beta}_1 = \frac{\overline{YX} - \bar{Y}\bar{X}}{\overline{X^2} - \bar{X}^2}; \quad \beta_0 = -5,79; \beta_1 = 36,84; \hat{Y} = -5,79 + 36,84X.$$

Valorile estimate ale variabilei de explicat sunt prezentate în ultima coloană. Parametrul β_0 este lipsit de careva sens economic.

$$\bar{X} = 3,14; \sigma_X = 1,25; V_X = 39,8\%; \bar{Y} = 110; \sigma_Y = 46,29; V_Y = 42,1\%. V_X = \frac{\sigma_X}{\bar{X}}; V_Y = \frac{\sigma_Y}{\bar{Y}}.$$

Relația $\beta_0 > 0$ corespunde faptului că variabila dependentă se schimbă cu ritmuri mai mari decât variabila independentă $V_Y \succ V_X$. $\beta_0 < 0$ reflectă faptul că variabila independentă se schimbă cu ritmuri mai mari decât schimbarea variabilei dependente $V_X \succ V_Y$.

Dacă vom exprima variabilele X și Y prin devieri de la nivelul mediu, atunci linia regresiei pe grafic va trece prin origine $Y' = Y - \bar{Y}; X' = X - \bar{X}; \hat{Y}' = \beta_1 X'$. Estimarea coeficientului de regresie nu se va schimba..

Estimarea coeficienților ecuației de regresie poate fi obținută într-un mod mai simplu, fara a ne adresa la M.C.M.M.P. Estimația de alternativă a coeficientului β_1 poate fi obținută pornind

de la sensul acestui coeficient: schimbarea variabilei dependente $\Delta Y = Y_n - Y_1$ se confruntă cu schimbarea variabilei independente $\Delta X_n = X_n - X_1$.

În exemplul examinat o atare estimatie de alternativă a parametrului β_1 va constitui $\beta_1' = \frac{170-30}{5-1} = 35$. Această mărime este aproximativă întrucât nu se ține cont de ceea mai mare parte din informația statistică accesibilă. Ea se bazează numai pe valorile variabilelor de mărime maximă sau minimă.

De regulă, ecuația de regresie este urmată de coeficientul de corelație liniară ce caracterizează cât de puternică este dependența liniară între X și Y - r_{XY} . Există mai multe modificări ale formulei pentru coeficientul de corelație liniară. Una dintre ele se prezintă aici.

$r_{XY} = \beta_1 \frac{\sigma_X}{\sigma_Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$, $\beta_1 = \frac{\text{cov}(X,Y)}{\sigma_X^2}$; e cunoscut faptul, că coeficientul liniar de corelație se modifică în diapazonul $-1 < r_{XY} < 1$

Dacă coeficientul $\beta_1 < 0$, $-1 \leq r_{XY} \leq 0$ și invers, dacă $\beta_1 > 0$, $0 \leq r_{XY} \leq 1$

Pentru datele tabelului, valoarea lui este de $r_{XY} = 0,991$, deci e foarte aproape de 1, ceea ce înseamnă că între X și Y exista o legătură liniară puternică, cheltuielile pentru producere sunt puternic dependente de volumul de producție.

Este necesar să se țină cont de faptul că coeficientul liniar de corelație evaluează măsura legăturii puternice între indicii examinați sub forma liniară. Aproximarea valorii coeficientului de regresie către 0 nu înseamnă inexistența legăturii între indici. Dacă modelul va fi specificat în alt mod, legătura dintre indici poate să se adevărească destul de strânsă.

Pentru evaluarea calității funcției liniare alese se calculează pătratul coeficientului linear de corelație R_{YX}^2 care se numește coeficient de determinație - coeficientul de determinație caracterizează cota dispersiei variabilei de explicat Y care este lămurită de regresie în dispersia totală a variabilei de explicat:

$R_{YX}^2 = \frac{\hat{\sigma}_Y^2}{\sigma_Y^2}$. Respectiv, mărimea $1 - R_{YX}^2$ caracterizează cota dispersie Y , explicată de restul

factorilor, care nu se examinează în model.

În exemplul examinat $R_{YX}^2 = 0,982$. Prin urmare ecuația de regresie explică 98,2% din dispersia indicelui rezultat, dar pe seamă altor factori revine numai 1,8% din dispersia totală. Mărimea coeficientului de determinație servește ca unul din criteriile care evaluează calitatea modelului liniar. Cu cât este mai mare cota variației explicată de regresie, cu atât, respectiv, este mai mică influența altor factori și deci modelul liniar bine aproximează datele inițiale și poate fi utilizat pentru pronosticarea valorilor variabilei rezultative (de explicat).

4. Evaluarea semnificației ecuației de regresie liniară și a coeficienților ei

După ce a fost estimată ecuația lineară de regresie se efectuează evaluarea semnificației atât a ecuației în întregime, cât și a fiecărui parametri separat.

Evaluarea semnificației ecuației de regresie în întregime se produce cu ajutorul F-criteriului Fișer, formulând în prealabil ipoteza dependenței direct proporționale $H_0 : \{\beta_0 = 0\}$ și ipoteza independenței variabilelor, $H_0 : \{\beta_1 = 0\}$ contra $H_1 : \left\{ \begin{matrix} \beta_0 \neq 0 \\ \beta_1 \neq 0 \end{matrix} \right\}$ - ipoteza dependenței lineare specificate.

În scopul testării validității modelului se analizează descompunerea sumei pătratelor devierilor a variabilei Y de la medie în două componente: prima, explicată de regresie și a doua neexplicată de regresie:

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

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n ((\hat{Y}_i - \bar{Y}) + u_i)^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n u_i^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})u_i,$$

$$\sum_{i=1}^n u_i = 0 \Rightarrow \sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i \Rightarrow \bar{Y} = \bar{\hat{Y}} = 2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})u_i = 2 \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})u_i = 0.$$

$$\bar{Y} = \bar{\hat{Y}}$$

Suma pătratelor devierilor observațiilor variabilei dependente (de explicat) de la valoarea lor medie $\bar{Y}$ este cauzată de mai multe evenimente: de variabila explicativă și de alți factori. Dacă variabila explicativă nu influențează rezultatul, atunci linia regresiei este paralelă axei OX și $\bar{Y} = \bar{\hat{Y}}$, iar toată dispersia variabilei de explicat este cauzată de alți factori. În cazul când alți factori nu influențează variabila de explicat, atunci Y este legat funcțional de X și suma pătratelor rezidualelor este egală cu 0. Prin urmare, suma pătratelor devierilor explicate de regresie, coincide cu suma totală. Întrucât nu toate punctele câmpului de corelare se află pe linia de regresie, de fiecare dată are loc dispersarea lor cauzată atât de variabila explicativă X (de regresie), cât și de alți factori (neexplicabili de regresie). Linia de regresie este bună pentru previziune, evident atunci, când suma pătratelor devierilor cauzată de regresia va fi cu mult mai mare decât suma pătratelor rezidualelor, atunci ecuația de regresie este semnificativă și variabila explicativă X influențează esențial variabila de explicat Y . În acest caz coeficientul de determinație R_{yx}^2 se va apropia de 1.

Orice sumă a pătratelor devierilor este legată de gradele de libertate a indiciului independent de variație. *Gradele de libertate sunt dependente de numărul de observații «n» și de numărul parametrilor definiți în conformitate cu ele.* Referitor la problema în cauză gradele de libertate trebuie să demonstreze, câte devieri independente din «n» posibile $[(Y_1 - \bar{Y}), (Y_2 - \bar{Y}), \dots, (Y_{n_1} - \bar{Y})]$, sunt necesare pentru formarea sumei pătratelor. Astfel,

pentru suma pătratelor $\sum_{i=1}^n (Y_i - \bar{Y})^2$ sunt necesare $(n-1)$ devieri independente, deoarece din totalitatea de «n» unități după calcularea mediei variază independent numai $(n-1)$ de devieri. De exemplu, avem un șir de valori 1,2,3,4,5. Valoarea medie ale lor este egală cu 3, «n» devieri de la medie sunt: -2; -1; 0; 1; 2. Deoarece $\sum_{i=1}^n (Y_i - \bar{Y}) = 0$, independent variază numai 4 devieri, dar devierea a cincea poate fi determinat, dat fiind cunoscute 4 precedente.

Pentru calcularea sumei pătratelor devierilor explicate de regresie $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ se folosesc valorile variabilei dependente $\hat{Y}_i$ calculate în conformitate cu ecuația de regresie $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

La utilizarea regresiei liniare este adevărată egalitatea $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$ care poate fi confirmată, dacă apelăm la formula coeficientului liniar de corelație $r_{xy} = \beta_1 \frac{\sigma_x}{\sigma_y} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \Rightarrow r_{xy}^2 = \beta_1^2 \frac{\sigma_x^2}{\sigma_y^2}$, de aici rezultă că $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$.

Fiind dat numărul observațiilor pentru X și Y suma pătratelor devierilor variabilei independente de la medie depinde numai de o singura constantă – coeficientul de regresie β_1 și deci suma examinată are numai un grad de libertate. La aceeași concluzie vom ajunge dacă vom examina ecuația $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Parametrul $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, substituind această valoare în ecuația de regresie obținem $\hat{Y}_i = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i = \bar{Y} - \hat{\beta}_1 (X_i - \bar{X})$. Deci, rezultă ca fiind date «n»

observații pentru X și Y , valoarea estimată a lui $\hat{Y}$ în ecuația liniară de regresie este funcție de un singur parametru – coeficientul de regresie. Respectiv și suma pătratelor devierilor variabilei independente (factor) are numai un singur grad de libertate.

Numărul gradelor de libertate a sumei pătratelor devierilor totale este egal cu numărul gradelor de libertate pentru suma pătratelor devierilor explicate de regresie și numărul gradelor de libertate ale sumei pătratelor devierilor reziduale. Numărul gradelor de libertate al sumei pătratelor reziduale în regresia liniară este egală cu $(n-2)$. Numărul gradelor de libertate pentru suma totală a pătratelor devierilor este determinată de numărul observațiilor « n », și, deoarece, se utilizează valoarea medie calculată în baza observațiilor, pierdem un grad de libertate, deci avem $(n-1)$ grade de libertate.

Așadar, avem două egalități:

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

$$(n-1) = 1 + (n-2) .$$

Împărțind fiecare sumă a pătratelor la gradele de libertate ce-i corespund, vom obține pătratul devierilor medii, sau dispersia la un grad de libertate.

$$\hat{\sigma}_{\bar{Y}}^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}) / 1 ; \quad \hat{\sigma}_u^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i) / (n-2); \quad \hat{\sigma}_Y^2 = \sum_{i=1}^n (Y_i - \bar{Y}) / (n-1) .$$

Determinarea dispersiilor racordate la grad de libertate oferă posibilitatea de a efectua comparații între ele. Examinând raportul dintre dispersia explicată de regresie și dispersia reziduală racordate la un grad de libertate, obținem F – criteriul: $F = \frac{\hat{\sigma}_{\bar{Y}}^2}{\hat{\sigma}_u^2}$, F – criteriul de

verificare a ipotezei dependenței liniare a variațiilor. $H_0 : \hat{\sigma}_{\bar{Y}}^2 = \hat{\sigma}_u^2$. Dacă ipoteza H_0 este adevărată, atunci dispersiile nu diferă. Pentru respingerea ipotezei H_0 este necesar ca $\hat{\sigma}_{\bar{Y}}^2 > \hat{\sigma}_u^2$ (de câteva ori). Savantul englez Snedecor a elaborat tabele pentru valorile critice a F – criteriului în raport cu diferite nivele de semnificație a ipotezei H_0 și diferite grade de libertate. Valoarea tabelară a F – criteriului este valoarea maximă a raportului dintre dispersiile $\hat{\sigma}_{\bar{Y}}^2, \hat{\sigma}_u^2$, care poate avea loc la dispersarea aleatoare fiind dată probabilitatea existenței ipotezei H_0 . Valoarea calculată a raportului este veridică, dacă ea este mai mare decât valoarea tabelară. În acest caz ipoteza H_0 se respinge și se trage concluzia că: $F_{\text{efect}} > F_{\text{tab}}$ și se confirmă ipoteza H_1 .

Dacă $F_{\text{efect}} < F_{\text{tab}}$, probabilitatea validării ipotezei H_0 este mai mare decât nivelul indicat (de exemplu, 0,05) și ea nu poate fi respinsă. În acest caz, ecuația de regresie se consideră nesemnificativă, prin urmare, nu se respinge ipoteza H_0 .

În exemplul examinat:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 = 15000; \quad \hat{\sigma}_{\bar{Y}}^2 = \frac{15000}{6};$$

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 = 14735; \quad \hat{\sigma}_u^2 = \frac{265}{5} = 53;$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 265; \quad F = 1435/53 = 278; \quad F_{\alpha=0,005} = 6.61,$$

$$F_{\alpha=0,01} = 16.26; \quad F_{\text{fact}} = 278 > F_{\text{nf}} = 6.61; \quad F_{\text{fact}} = 278 > F_{\text{nf}} = 16.26.$$

Criteriul Fișer este strâns legat cu coeficientul de determinație. $R^2, \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = R^2 \hat{\sigma}_{\bar{Y}}^2 \cdot 1,$

$$\sum_{i=1}^n (\hat{Y}_i - Y_i)^2 = (1 - R^2) \hat{\sigma}_Y^2 / (n-2), \text{ atunci } F = \frac{R^2 \cdot (n-2)}{(1 - R^2)}.$$

Estimarea semnificației ecuației de regresie, de regulă, se prezintă sub forma tabelului analizei disperse.

Surse de variație	Grade de libertate	Suma pătratelor devierilor	Dispersia la un grad de libertate	F - criteriu	
				fact	tabel $\alpha=0.05$
Totală	n-1	15 000	2 500		
Explicată de regresie	1	14 375	14 375	278	6.61
Reziduală	n-2	235	53		

În regresia liniară se analizează nu numai semnificația ecuației în întregime dar și semnificația separată a parametrilor. În acest scop se determină și eroarea standard pentru fiecare parametru: $\hat{\sigma}_{\beta_0}, \hat{\sigma}_{\beta_1}$. Eroarea standard a coeficientului de regresie este definită după formula:

$$\hat{\sigma}_{\beta_1} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / (n-2)}{\sum_{i=1}^n (X_i - \bar{X})^2}} = \sqrt{\frac{S^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Valoarea erorii standard în comun cu t – distribuția Student la (n – 2) grade de libertate se aplică la verificarea semnificației coeficientului de regresie $\hat{\beta}_1$ și pentru calcularea intervalelor de încredere.

Pentru estimarea semnificației coeficientului de regresie, valoarea lui se compară cu eroarea standard, cu alte cuvinte se determină valoarea efectivă a t – criteriul Student: $t_{\beta_1} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}$, care se compară cu valoarea tabelară pentru riscul erorii (α) (nivelul de semnificație) și (n – 2) grade de libertate.

Același rezultat îl obținem când extragem rădăcina pătrată din F – criteriul, și anume, $t_{\beta_1} = \sqrt{F}$. Vom demonstra că $R^2_{\beta_1} = F$.

$$\begin{aligned} (t_{\beta_1})^2 &= \left(\frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}\right)^2 = \frac{\hat{\beta}_1^2}{\sum (Y_i - \hat{Y}_i)^2 / (n-2) / \sum (X_i - \bar{X})^2} = \frac{\hat{\beta}_1^2 \sum (X_i - \bar{X})^2}{\sum (Y_i - \hat{Y}_i)^2 / (n-2)} = \\ &= \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \hat{Y}_i)^2 / (n-2)} = \frac{\hat{\sigma}_Y^2}{\hat{\sigma}_u^2} = F. \end{aligned}$$

Intervalul de încredere pentru coeficientul de regresie se determină ca $\hat{\beta}_1 \pm t_{tab} \cdot \hat{\sigma}_{\beta_1}$, este egal cu valoarea coeficientului estimativă $\pm$ valoarea coeficientului Student tabelă înmulțit cu $\hat{\sigma}_{\beta_1}$.

Întrucât coeficientul de regresie în investigațiile econometrice are o explicație economică clară, intervalele de încredere pentru el nu trebuie să conțină rezultate contradictorii, de exemplu $-10 \leq \hat{\beta}_1 \leq 40$. Ceea ce înseamnă că valoarea adevărată a coeficientului de regresie conține simultan valori pozitive și negative și chiar 0, ce nu poate avea loc.

Eroarea standard a parametrului $\hat{\beta}_0$ se determină prin formula:

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n-2)} \cdot \frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}} = \sqrt{\frac{S^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}}.$$

Evaluarea semnificației se efectuează la fel ca și pentru $\hat{\beta}_1$, $t_{\hat{\beta}_0} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}}$, valoarea t-criteriului calculat se compară cu valoarea tabelară la $(n-2)$ grade de libertate și nivelul de semnificație (α).

Semnificația coeficientului de determinație R se definește în baza valorii erorii coeficientul de determinație $\hat{\sigma}_R = \sqrt{\frac{1-R^2}{n-2}}$.

Valoarea efectivă a t-criteriului student se determină ca: $t_R = \frac{R}{\sqrt{1-R^2}} \sqrt{n-2}$. Aceasta formulă ne mărturisește, că în regresia liniară față de variabile $t_R^2 = F$, deoarece s-a notat ca $F = \frac{R^2}{(1-R^2) \cdot (n-2)}$, plus la aceasta, $t_{\hat{\beta}_1}^2 = F$, a $t_R^2 = t_{\hat{\beta}_1}^2$.

Deci verificarea ipotezei semnificației a coeficienților de regresie și de determinație echivalează cu verificarea ipotezei referitor la validitatea modelului liniar de regresie.

Formula examinată pentru estimarea coeficientului de corelare este recomandată pentru aplicare la un număr mare de observații și dacă r diferă mult de $+1$ sau -1 . În caz contrar distribuția estimațiilor diferă de la aceea normală sau Student, deoarece coeficientul de corelare este limitat de valorile -1 și $+1$. Fisher a introdus o variabilă $z = \frac{1}{2} \ln\left(\frac{1+R}{1-R}\right)$ pentru evalua semnificația R .

4.1. Intervalele de previziune pentru modelul liniar de regresie

În calcule previzionale conform ecuației de regresie se determină valoarea $\hat{Y}_p$ sub formă de previziune punctiferă $\hat{Y}_i$ pentru $X_p = X_k$, substituind în ecuația de regresie $\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_k$ valoarea respectivă a lui X . Dar previziunea punctiferă este evident nereală. De aceea ea este completată cu calculele erorii standard pentru $\hat{Y}_p, \hat{\sigma}_{\hat{Y}}$ și cu estimația intervalului de previziune pentru valoarea Y^* , $\hat{Y}_{x_p} - \hat{\sigma}_{\hat{Y}_x} \leq Y^* \leq \hat{Y}_x + \hat{\sigma}_{\hat{Y}_x}$.

Întru construirea formulei pentru eroarea standard $\hat{\sigma}_{\hat{Y}_x}$ vom apela la ecuația de regresie $\hat{Y}_{x_p} = \hat{\beta}_0 + \hat{\beta}_1 X_p$. Substituind $\hat{\beta}_0$ cu formula pentru calcularea lui $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, vom obține $\hat{Y}_x = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X = \bar{Y} + (X - \bar{X}) \hat{\beta}_1$. De aici rezultă că eroarea standard pentru $\hat{Y} - \hat{\sigma}_{\hat{Y}_x}$ depinde de eroarea $\bar{Y}$ și eroarea coeficientului de regresie $\hat{\beta}_1$, deci $\hat{\sigma}_{\hat{Y}_x}^2 = \hat{\sigma}_{\bar{Y}}^2 + \hat{\sigma}_{\hat{\beta}_1}^2 (X - \bar{X})^2$.

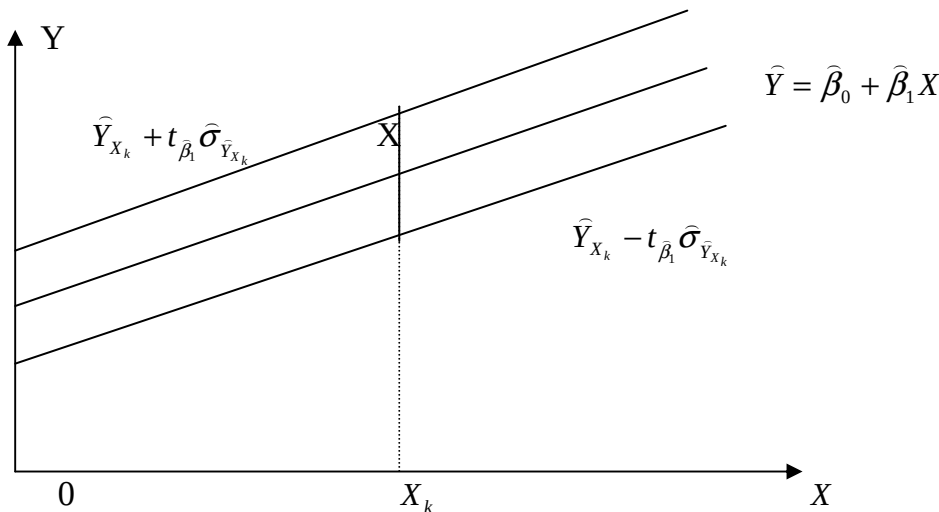
Din teoria selectării este cunoscut faptul că $\hat{\sigma}_{\bar{Y}}^2 = \frac{\sigma^2}{n}$, folosind în calitate de σ^2 dispersia reziduală pentru un grad de libertate S^2 , obținem: $\hat{\sigma}_{\bar{Y}}^2 = \frac{S^2}{n}$. Eroarea standard a coeficientului de regresie este determinată prin formula $\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{S^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$. Considerăm că valoarea

prognozată $X_p = X_k$, atunci în conformitate cu ecuația de regresie obținem următoarea formulă pentru eroarea standard a valorii $\hat{Y}_{X_k}$ prognozată:

$$\hat{\sigma}_{\hat{Y}_{X_k}}^2 = \frac{S^2}{n} + \frac{S^2(X_k - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = S^2 \left(\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right).$$

Respectiv, $\hat{\sigma}_{\hat{Y}_{X_k}} = S \sqrt{\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$. Formula considerată a erorii standard pentru

valoarea medie prognozată a lui $\hat{Y}_X$, fiind dată valoarea X_k , caracterizează eroarea amplasării liniei de regresie. Valoarea erorii standard $\hat{\sigma}_{\hat{Y}_{X_k}}$ atinge minimumul atunci când $X_k = \bar{X}$ și crește cu îndepărtarea punctului X_k de la $\bar{X}$ în orice direcție. Cu alte cuvinte, cu cât este mai mare diferența dintre X_k și $\bar{X}$, cu atât este mai mare eroarea $\hat{\sigma}_{\hat{Y}_{X_k}}$. În baza ei fiind evaluată previziunea valorii medii $\hat{Y}$, pentru valoarea X_k dată. Pot fi așteptate previziuni mai reușite dacă punctul X_k se află în centrul regiunii de observare și nu este cazul să așteptăm rezultate de previziune bune la îndepărtarea a punctului X_k de la punctul $\bar{X}$. În caz că valoarea X_k se află în afară valorilor observate a lui X , folosite la determinarea liniei de regresie, rezultatele previziunii se înrăutățesc pe măsura deplasării X_k de la regiunea valorilor observate pentru variabila explicativă X .



Pentru valoarea prognozată a lui $\hat{Y}_{X_k}$, intervalele de încredere de 95% pentru X_k dat se definesc prin expresia: $\hat{Y}_{X_k} \pm t_{\beta_1} \hat{\sigma}_{\hat{Y}_{X_k}}$. Pe grafic frontierile de încredere pentru $\hat{Y}$ reprezintă două hiperbole situate pe ambele părți de la linia de regresie. Două hiperbole pe ambele părți de la linia de regresie determină intervalele de încredere de 95% pentru valoarea medie a lui $\hat{Y}$ pentru X dat.

Însă valorile observate a lui Y variază în jurul valorii medii a lui $\hat{Y}$. Valorile individuale a lui Y pot fi dispersate de la $\hat{Y}$ în limita valorii erorii aleatoare u , dispersia ei fiind evaluată ca dispersia reziduală pentru un grad de libertate σ_u^2 . Deaceea eroarea valorii individuale

prognozate pentru Y necesită includerea nu numai a erorii standard $\hat{\sigma}_{\hat{Y}_{X_k}}$, dar și a erorii aleatoare σ_u . Eroarea medie $\hat{\sigma}_{\hat{Y}_{X_k}}$ a valorii individuale $\hat{Y}$ prognozată este:

$$\hat{\sigma}_{\hat{Y}_{X_k}} = S \sqrt{1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

Efectuând previziunea în baza ecuației de regresie este necesar să ținem cont de faptul că valoarea prognozată depinde nu numai de eroarea standard a valorii individuale Y , dar și de precizia previziunii valorii variabilei exogene X . Valoarea ei poate fi definită în baza aplicării altor modele, reieșind din situația concretă și analizând dinamica acestui factor. Formula considerată pentru eroarea medie ($\hat{\sigma}_{\hat{Y}_{X_k}}$) a valorii individuale Y poate fi folosită pentru evaluarea semnificației devierii valorii prognozate prin ecuația de regresie și valorii ipotetice înaintate în baza evoluției evenimentelor.

$$t_{\hat{Y}(X_k)} = \frac{(\hat{Y}_{(X_k)} - Y_{hipot})}{\hat{\sigma}_{\hat{Y}_{(X_k)}}}; \quad t_{calc\hat{Y}_{(X_k)}} \succ t_{tab(0,05;n-2)}.$$

5. Modelul de regresie liniară multifactorial

5.1. Crearea și specificarea modelului liniar de regresie genera

Să examinăm cazul când, cu siguranță, mai multe variabile independente pot în plină măsură să explice comportamentul unei variabile dependente. Cazurile când comportamentul variabilei dependente poate fi explicat de o singură variabilă independentă sunt rar întâlnite în realitate. Cererea pentru un careva produs este, cu certitudine, influențată de prețuri, însă această explicație nu este una completă, deoarece reclama, venitul agregat, prețurile produselor de substituție, piețele internaționale, calitatea serviciilor comerciale, diverse capricii a cumpărătorilor, schimbarea preferințelor consumatorilor - toate sunt importante în modelarea reală. Prin urmare, se simte necesitatea vitală de a trece de la modelul regresional de două variabile la modelele de regresie cu mai multe variabile.

Modelul liniar de regresie generalizat cu k variabile independente poate fi prezentat sub forma unei ecuații:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i, \quad (5.1)$$

unde $i = \overline{1, n}$ indică numărul observațiilor, X_{1i} indică observația i a variabilei X_1 , în timp ce X_{2i} indică observația i a variabilei X_2 , k - numărul variabilelor independente, ε_i - termenul erorii stocastice.

Ceea mai mare deosebire dintre modelul de regresie de o singură variabilă și modelul de regresie cu mai multe variabile (multifactorial) constă în interpretarea coeficienților adiționali de înclinație. Acești coeficienți, deseori denumiți coeficienți de regresie parțiali, deoarece coeficienții regresiei multiple corespund derivatelor parțiale după variabilele independente respective. Coeficienții sunt definiți cu scopul de a permite cercetătorului de a distinge impactul unei sau altuia variabile independente. Și anume, coeficientul regresiei multifactoriale indică schimbarea în variabila dependentă examinată în timp ce restul variabilelor independente din ecuație se mențin constante.

Ultima frază subliniată constituie momentul-cheie în înțelegerea regresiei multiple. Coeficientul β_1 măsoară impactul asupra lui Y a creșterii de o unitate în X_1 menținând constante variabilele $X_2, X_3, \dots, X_k$, dar nu este constantă nici una din variabilele relevante omise din ecuație (cum ar fi X_{k+1}). Coeficientul β_0 este valoarea lui Y pentru toți $X_s = 0$ și $\varepsilon_s = 0$,

$s = \overline{1, k}$. Cum a mai fost menționat termenul β_0 se va include în ecuația de regresie, dar în baza lui nu nu pot fi trase concluzii. Fie că:

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_{11} + \hat{\beta}_2 X_{12} + \dots + \hat{\beta}_k X_{1k}$$

$$\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 X_{21} + \hat{\beta}_2 X_{22} + \dots + \hat{\beta}_k X_{2k}$$

.....

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_{k1} + \hat{\beta}_2 X_{k2} + \dots + \hat{\beta}_k X_{kk}$$

.....

$$\hat{Y}_n = \hat{\beta}_0 + \hat{\beta}_1 X_{n1} + \hat{\beta}_2 X_{n2} + \dots + \hat{\beta}_k X_{nk},$$

Atunci modelul liniar de regresie multiplă poate fi înscris sub forma vectorială $Y = X\beta + \varepsilon$, Y - vectorul coloană al variabilei endogene de dimensiunea (n) , X - matricea de dimensiunea $(n \cdot (k+1))$, β - vectorul de dimensiunea $(k+1)$, ε - vectorul de dimensiunea (n) .

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, i = \overline{1, n}; j = \overline{0, k}, \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, u = u \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\beta_1 = \frac{\partial Y}{\partial X_1}; \beta_2 = \frac{\partial Y}{\partial X_2} \dots \beta_k = \frac{\partial Y}{\partial X_k}; \hat{Y} = X\hat{\beta}, u = Y - \hat{Y}.$$

De exemplu, funcția de consum în cele mai multe cazuri este examinată ca un model de forma: $C = f(Y, P, M, Z)$; unde C - consumul; Y - venitul; P - indicele costului de viață; M - banii în numerar; Z - active lichide, $\frac{\partial C}{\partial Y} < 1$. Regresia multiplă este utilizată pe larg la soluționarea problemelor cererii, venitul pe acțiuni, la studierea cheltuielilor de producere, în calcule macroeconomice. În prezent regresia multiplă este una din cele mai răspândite metode în econometrie. **Scopul de bază a regresiei multiple constă în definirea modelului de regresie multiplă, fiind determinată influența fiecărei variabile independente în parte și influența lor în comun asupra variabilei dependente.**

5.2. Ipoteze clasice, M.C.M.M.P.

Ipotezele clasice necesită a fi îndeplinite pentru ca estimatorii obținuți prin metoda celor mai mici pătrate să fie cei mai buni disponibili.

Termenul de eroare care satisface ipotezelor anunțate în continuare se numește termen normal de eroare de tip clasic.

1. Valorile variabilelor $X_{ij}, (i = \overline{1, n}; j = \overline{1, k})$ sunt observate fără erori constante de măsurare.
2. Termenul de eroare ε_i este de media nulă $E(\varepsilon_i) = 0$ (sau variabila ε este de media nulă $(\varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_n) / n = 0$ și normal distribuită).
3. Variația erorilor este constantă pentru fiecare i , $VAR(\varepsilon_i) = \sigma_{\varepsilon_i}^2 = \sigma^2 = const$.
4. Valorile observate ale termenului de eroare nu sunt corelate (are loc independența erorilor $E(\varepsilon_i \varepsilon_j) = 0; i, j = \overline{1, n}; i \neq j$, nu sunt corelații în serie).
5. Erorile sunt independente de variabilele explicative. $COV(X_{ij}, \varepsilon_i) = 0; E(X_{ij} \varepsilon_i) = 0$.
6. Absența coliniarității între variabilele explicative implică o matrice $(X^T X)$ regulată și asigură existența matricei inverse $(X^T X)^{-1}$.
7. Matricea $(X^T X)^{-1} / n$ este o matrice finită ne singulară.

8. Relația $n \succ k$, $n \approx 6(7) \cdot k$ necesită ca numărul de observații să fie superior numărului de variabile explicative.

În cazul respectării ipotezelor enunțate, estimarea parametrilor β_j se face, de regulă, cu ajutorul M.C.M.M.P. care constă în minimizarea sumei pătratelor devierilor valorilor variabilei dependente estimate de la valorile variabilei dependente teoretice.

$$\min_{\beta_i} \sum_{i=1}^n u_i^2 = \min_{\beta_i} (u^T u) = \min_{\beta_i} (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = S(\hat{\beta}),$$

$$S(\hat{\beta}) = (Y^T Y) - 2\hat{\beta}(X^T Y) + \hat{\beta}^T (X^T X) \hat{\beta}.$$

Pentru minimizarea funcționalului $S(\hat{\beta})$, în conformitate cu condițiile necesare de existență a extremumului, efectuăm derivarea lui în raport cu β_s , $s = 0, 1, \dots, k$ și apoi egalăm cu zero expresia obținută:

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = 0; \quad -2(X^T Y) + 2(X^T X) \cdot \hat{\beta} = 0 \Rightarrow (X^T X) \cdot \hat{\beta} = (X^T Y) \Rightarrow \hat{\beta} = (X^T X)^{-1} (X^T Y)$$

5.3. Estimarea parametrilor modelului și proprietățile lor

1. Estimatorul $\hat{\beta}$ este un estimator **BLUE** (Best Linear Unbiased Estimator), adică “cel mai bun estimator liniar nedeplasat” $E(\hat{\beta}) = \beta$.
2. Estimatorul liniar nedeplasat $\hat{\beta}$ are dispersia minimă.
3. Estimatorul $\hat{\beta}$ este consistent: $\hat{\beta} \xrightarrow{n \rightarrow \infty} \beta$.
4. Estimatorul $\hat{\beta}$ este normal distribuit $N(\beta, \text{VAR}(\hat{\beta}))$.

Notății convenționale

Valorile adevărate, dar neobservate		Estimate	
Denumirea	Simbolul	Denumirea	Simbolul
coef.de regresie	β_k	coef.de regresie estimat	$\hat{\beta}_k$
Valoarea așteptată a coef. estimat	$E(\hat{\beta}_k)$		
Variația termenului de eroare	σ^2 sau $\text{VAR}(\varepsilon_i)$	Variația estimată a termenului de eroare	S^2 sau $\hat{\sigma}^2$
Devierea standard a termenului de eroare	σ	Devierea standard a termenului de eroare estimat	S sau $\hat{\sigma}$
Variația coeficienților estimați	$\sigma^2(\hat{\beta}_k)$ sau $\text{VAR}(\hat{\beta}_k)$	Variația estimată a coeficienților estimați	$S^2(\hat{\beta}_k)$ sau $\hat{\sigma}^2(\hat{\beta}_k)$
Devierea standard de la coef.estimați	$\sigma_{\hat{\beta}_k}$ sau $\sigma(\hat{\beta}_k)$	Eroarea standard a coeficienților estimați	$\hat{\sigma}_{\hat{\beta}_k}$ sau $SE(\hat{\beta}_k)$
Termenul erorii stocastice	ε_i	Reziduale(estimarea erorii în sens informal)	u_i

Matricea varianței și covarianței coeficienților de regresie ea forma: $\Omega_{\hat{\beta}} = \sigma_u^2 (X^T X)^{-1}$, iar dispersia reziduală este $\sigma_u^2 = (u^T u)$, dar aceea racordată la un grad de libertate e $\hat{\sigma}_u^2 = \frac{(u^T u)}{n-k-1}$.

Aceste valori depind de unitatea de măsură, de aceea se preferă utilizarea coeficientului de determinație R^2 și a coeficientului de corelație multiplă r .

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

$$r_{YX_1X_2\cdots X_k} = \sqrt{1 - \frac{\sigma_u^2}{\sigma_Y^2}}, \quad \text{unde } \sigma_u^2 = \sum (y_i - \hat{y}_i)^2 \quad \text{este dispersia reziduală} \quad \sigma_Y^2 = \sum (y_i - \bar{Y})^2 \quad \text{este dispersia totală.}$$

5.4. Verificarea modelului cu ajutorul testelor

Sistemul de ecuații normale pentru regresia multiplă este următorul:

[illegible]

$$\Delta = \begin{vmatrix} n & \sum_{i=1}^n X_{1i} & \sum_{i=1}^n X_{2i} & \cdots & \sum_{i=1}^n X_{ki} \\ \sum_{i=1}^n X_{1i} & \sum_{i=1}^n X_{1i} X_{1i} & \sum_{i=1}^n X_{1i} X_{2i} & \cdots & \sum_{i=1}^n X_{1i} X_{ki} \\ \sum_{i=1}^n X_{2i} & \sum_{i=1}^n X_{2i} X_{1i} & \sum_{i=1}^n X_{2i} X_{2i} & \cdots & \sum_{i=1}^n X_{2i} X_{ki} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^n X_{ki} & \sum_{i=1}^n X_{ki} X_{1i} & \sum_{i=1}^n X_{ki} X_{2i} & \cdots & \sum_{i=1}^n X_{ki} X_{ki} \end{vmatrix}$$

$$\Delta \hat{\beta}_0 = \begin{vmatrix} \sum_{i=1}^n Y_i & \sum_{i=1}^n X_{1i} & \sum_{i=1}^n X_{2i} & \cdots & \sum_{i=1}^n X_{ki} \\ \sum_{i=1}^n Y_i X_{1i} & \sum_{i=1}^n X_{1i} X_{1i} & \sum_{i=1}^n X_{1i} X_{2i} & \cdots & \sum_{i=1}^n X_{1i} X_{ki} \\ \sum_{i=1}^n Y_i X_{2i} & \sum_{i=1}^n X_{2i} X_{1i} & \sum_{i=1}^n X_{2i} X_{2i} & \cdots & \sum_{i=1}^n X_{2i} X_{ki} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^n Y_i X_{ki} & \sum_{i=1}^n X_{ki} X_{1i} & \sum_{i=1}^n X_{ki} X_{2i} & \cdots & \sum_{i=1}^n X_{ki} X_{ki} \end{vmatrix}, \quad \hat{\beta}_s = \Delta \hat{\beta}_s / \Delta, s = \overline{0, k}.$$

Judecând în mod analogic, precum și în cazul modelului liniar de regresie simplă, obținem că variația totală este egală cu variația explicată plus variația reziduurilor:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.$$

Aceste valori depind de unitatea de măsură, de aceea se preferă utilizarea coeficientului de determinație R^2 și a coeficientului de corelație multiplă r .

Pentru a aplica testul Fișer vom alcătui următorul tabel al analizei de varianțe.

Sursele de variație	Variația sumei pătratelor abaterilor	Gradul de libertate	Estimatori ai dispersiei în raport cu gradele de libertate
Explicate de regresie	$V_E = \sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 = \sigma_Y^2$	k	$\hat{\sigma}_Y^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2}{k}$
Reziduale	$V_R = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sigma_u^2$	$n - k - 1$	$\hat{\sigma}_Y^2 = \frac{\sum_{i=1}^n (\hat{y}_i - Y_i)^2}{(n - k - 1)}$
Totală	$V_T = \sum_{i=1}^n (y_i - \bar{Y})^2 = \sigma_Y^2$	$n - 1$	$\hat{\sigma}_Y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{(n - 1)}$

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 - \sum_{i=1}^n y_i - \hat{y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sigma_u^2}{\sigma_Y^2};$$

$$1 - R^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \quad \bar{R}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - k - 1)}{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n - 1)} = 1 - \frac{\hat{\sigma}_u^2}{\hat{\sigma}_Y^2}; \quad r = \sqrt{1 - \sigma_u^2 / \sigma_Y^2}.$$

R^2 - măsoară gradul de varianță a lui Y explicat prin regresia Y pe X .

$$F_{calc} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 / (k)}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / (n-k-1)} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \cdot (n-k-1)}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \cdot k}.$$

Regula de decizie pentru un prag de semnificație I_0 pentru α_0 este acceptată, dacă $F_{cfkc} < F_{tab} = F_{1-\alpha; (k-1, n-k)}$ și este acceptată I_1 și respinsă I_0 , dacă $F_{cfkc} > F_{tab} = F_{1-\alpha; (k-1, n-k)}$.

$$F_{cfkc} = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}.$$

6. Multicoliniaritatea și atenuarea ei

Spre deosebire de modelul unifactorial, în cazul modelului multifactorial, ipoteza I_1 presupune independența variabilelor explicative. Ne respectarea ei produce fenomenul de multicoliniaritate, caz în care o variabilă endogenă este explicată de mai multe variabile explicative.

Frecvența relativ ridicată a coliniarității dintre variabilele explicative se datorează gradului sporit de interdependență din economie.

Existența multicoliniarității este semnalată de:

- analogiile în evoluția variabilelor explicative;
- apropierea de zero a determinantului $|X^T X|$;
- mărimea coeficientului de determinație multiplă (R^2), care aproape coincide cu mărimea lui în cazul în care una dintre variabilele cauzale este omisă;
- contrazicere în verificarea testelor și anume: testul F aplicat valorilor teoretice este semnificativ, iar testul t aplicat parametrilor de regresie semnalează nesemnificații în rândul parametrilor.

6.1. Atenuarea multicoliniarității. Procedee de selecție a variabilelor exogene în cazul unui model multifactorial.

- Dacă seriile de date sunt formate dintr-un număr redus de termeni ($n < 10$), atunci se recomandă includerea de termeni suplimentari ca ($n > 15$), astfel încât analogiile întâmplătoare să fie, pe cât posibil, eliminate.
- În cazul corelării intense a 2 variabile exogene, se renunță la una din ele, considerându-se că variabila omisă este exprimată de către cea reținută în model.
- Dacă datele sunt prezentate sub formă de serii cronologice, se poate proceda la calculul diferențelor de ordinul întâi ($\Delta = Y_i - Y_{i-1}$) sau la logaritmare valorilor $Y_i, X_{1i}, X_{2i}, \dots, X_{ni}$ în scopul atenuării colinearității, prilejuite de prezența trendului în date.

Eliminarea fenomenului de colinearitate implică calcularea coeficienților de corelație dintre variabilele exogene r_{X_i / X_j} și r_{Y / X_i} coeficienții de corelație liniară dintre variabila de explicat Y și variabilele sale explicative X_i . Dacă $r_{X_i / X_j} \approx 1$, $i \neq j$, va trebui ca una din cele două variabile să fie eliminată din rândul variabililor exogene.

Criteriul de excludere/includere a 2 variabile exogene care-s corelate liniar. Dacă $r_{Y / X_i} > r_{Y / X_j}$, se exclude X_j și se reține X_i , în caz contrar se exclude X_i și se reține X_j . Astfel, la prima etapă, reținând k variabile exogene liniar independente – fiind posibilă estimarea celor $(k+1)$ parametri se poate trece la etapa în care se continuă operația de selecție a variabililor exogene X_i . În acest sens există mai multe procedee.

6.1.1. Primul procedeu

În model se introduc cele k variabile exogene, ordinea de includere fiind dată de mărimea coeficienților de corelație a variabilei Y în raport cu factorii săi $r_{Y/X_1} \succ r_{Y/X_2} \succ r_{Y/X_3} \succ \dots r_{Y/X_k}$, în așa fel se obțin k modele:

$$M(1): Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \hat{u}_i^1 = \hat{Y}_i^1 + \hat{u}_i^1$$

⋮

$$M(j): Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \dots + \hat{\beta}_j X_{i,j} + \hat{u}_i^j = \hat{Y}_i^j + \hat{u}_i^j$$

⋮

$$M(k): Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \dots + \hat{\beta}_j X_{i,j} + \dots + \hat{\beta}_k X_{i,k} + \hat{u}_i^k = \hat{Y}_i^k + \hat{u}_i^k$$

Este cunoscut că, variația totală a variabilei Y este egală cu suma variației, explicată de regresie, a modelului $M(j)$ și variația reziduală $\sigma_{Y_j}^2 = \sigma_{\hat{Y}_j}^2 + \sigma_{\hat{u}_j}^2$.

Din relația precedentă ușor se obține coeficientul de determinație, care are forma de $R_j^2 = 1 - \frac{\sigma_{\hat{u}_j}^2}{\sigma_Y^2}$

și măsoară ponderea în variația totală a variației variabilei dependente Y_i , explicată de model.

$R_{u_j}^2 = 1 - R_j^2$ este cota parte a variației, neexplicate de regresia $M(j)$, în variația totală. În baza acestor relații se pot formula criterii de selecție a modelului optim. $M(r)$ din grupul de modele

$M(j)$, și anume $\sigma_r^2 = \max_j \sum_{i=1}^n (\hat{Y}_i^j - \bar{Y})^2$ sau $R_r^2 = \max_j R_j^2$, sau $\sigma_r^2 = \min_j \sum_{i=1}^n (Y_i - \hat{Y}_i^j)^2$, gradul

de semnificație a acestor indicatori, fiind verificat anterior prin testul F .

6.1.2. Al doilea procedeu

Al doilea procedeu pornește de la premisa că cei $(k+1)$ factori de influență ai variabilei de explicat Y sunt liniar independenți. În aceste condiții matricea $(X^T X)^{-1}$ există și cu ajutorul ei se estimează parametrii $(\hat{\beta}_j)$ și dispersiile acestora, obținându-se:

$M_k: Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \dots + \hat{\beta}_r X_{i,r} + \dots + \hat{\beta}_k X_{i,k} + \hat{u}_i^k$. Apoi se testează semnificația estimatorilor $(\hat{\beta}_j)$ cu ajutorul testului t cu pragul de semnificație α și $(n - (k+1))$ grade de libertate. Dacă $\frac{|\hat{\beta}_j|}{\sigma_{\hat{\beta}_j}} \geq t_{\alpha, (n - (k+1))}$, atunci $(\hat{\beta}_j)$ este semnificativ diferit de zero, în caz contrar $(\hat{\beta}_j)$ este

nesemnificativ diferit de zero.

Presupunând că $(\hat{\beta}_j)$ diferă de zero pentru $j = 0, 1, \dots, r$ și $(\hat{\beta}_j)$ nu diferă semnificativ de zero pentru $j = r+1, \dots, k$, ceea ce înseamnă că variabilele $(X_j), j \succ r$ nu influențează semnificativ variabila Y și pot fi excluse, astfel modelul va fi construit în baza variabilelor exogene $(X_j), j \leq r$.

6.1.3. Al treilea procedeu: Teste ce determină multicolinearitatea

✓ Testul Klein

Acest test este bazat pe comparația coeficientului de determinație R^2 pentru modelul cu k variabile exogene: $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_k X_k + u$ și coeficienților de corelație simplă r_{X_i/X_j}^2 dintre variabilele explicative pentru $i \neq j$. Dacă $R_Y^2 \prec r_{X_i/X_j}^2$, exista pericolul multicolinearității.

✓ Testul Farrar și Glauber

Etapa 1. Calculăm determinantul matricei coeficienților de corelație

$$D = \begin{vmatrix} 1 & r_{X_1/X_2} & r_{X_1/X_3} & \cdots & r_{X_1/X_k} \\ r_{X_2/X_1} & 1 & r_{X_2/X_3} & \cdots & r_{X_2/X_k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{X_k/X_1} & r_{X_k/X_2} & \cdots & r_{X_k/X_{k-1}} & 1 \end{vmatrix}$$

Dacă valoarea determinantului tinde spre zero, riscul multicolarității e mare. De exemplu, pentru un model de 2 variabile explicative, dacă ambele serii sunt perfect corelate, atunci

determinantul $D = \begin{vmatrix} 1 & r_{X_1/X_2} \\ r_{X_2/X_1} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} = 0$, iar în cazul când seriile sunt ortogonale

determinantul $D = \begin{vmatrix} 1 & r_{X_1/X_2} \\ r_{X_2/X_1} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1$.

Etapă 2. Efectuăm un test χ^2 , verificând următoarele ipotezele:

I_0 : $D = 1$ (seriile sunt ortogonale)

I_1 : $D < 1$ (seriile sunt dependente)

Valoarea empirică χ^2 calculată pentru un eșantion de n observații și K variabile explicative ($K = k + 1$, dacă se include termenul constant) este $\chi^2_{calc} = -[n - 1 - 1/6(2 * k + 5)] * \ln(D)$. Dacă $\chi^2_{calc} \geq \chi^2_{tabel}$ cu $k*(k-1)/2$ grade de libertate și un prag de semnificație α , atunci ipoteza I_0 este respinsă, are loc prezumția multicolarității. Dacă $\chi^2_{calc} < \chi^2_{tabel}$ se acceptă ipoteza de ortogonalitate.

7. Remedierea multicolarității

Ce poate fi întreprins întru reducerea consecințelor care multicolaritatea severă le produce asupra ecuației în examinare? Nu este un răspuns univoc deoarece multicolaritatea este un fenomen care se schimbă de la un set de date la altul, chiar dacă ecuația specificată de regresie este aceeași.

7.1. Nu se fac nimic

La prima etapă când multicolaritatea puternică este depistată este necesar de decis dacă în genere trebuie de făcut ceva. După cum s-a observat, orice remediu contra multicolarității produce șovăeli de anumit gen și, deseori se întâmplă că de a nu face nimic este o acțiune corectă!

Argumente majore în favoarea considerației serioase de a nu face nimic este acelea că multicolaritatea în ecuație nu întotdeauna reduce t-statisticile suficient întrucât ca ele să devină nesemnificative sau să modifice în mare măsură coeficienții $\hat{\beta}_s$ ca ei să difere semnificativ de la acei așteptați. Cu alte cuvinte, simpla existența a multicolarității nu întotdeauna înseamnă ceva. De exemplu, fie că coeficientul simplu de corelație între două variabile explicative este egal cu 0,97, fiecare din ele având t- statistica individuală semnificativă la pragul 95% de încredere. În așa caz nici un remediu nu are rost. În cazul multicolarității severe cel mai simplu remediu constă în eliminarea a unei sau a mai multor variabile din ecuație. Spre nefericire, eliminarea variabilei multicolineare, care conform teoriei aparține ecuației de regresie, este o operație destul de periculoasă, deoarece ecuația modificată va fi supusă deplasărilor de specificație. Dacă se elimină o atare variabilă atunci intenționat se creează deplasări în estimări. Deaceia econometricienii experimentați în cele mai multe cazuri vor păstra variabilele multicolineare în ecuație necâtând la diminuarea potențială a t-statisticilor.

Ultimul argument întru susținerea dezideratului de nu a întreprinde nici o măsură pentru a combate multicolinearitatea este unul teoretic care se aplică la orice ecuație. De fiecare dată când regresia se relansează, noi ne asumăm riscul de a descoperi o specificație care se potrivește, dat fiind întâmplătoare pentru un set de date distinct, dar nu de aceea că ea este adevărată. Mărirea numărului încercărilor mărește șansele de a obține rezultate întâmplătoare. Deci, specificația consecutivă este bine venită atunci când există multicolinearitate severă, deoarece în acest caz estimarea coeficienților este un procedeu sensibil la modificări de specificare. În concluzie, deseori este mai bine să păstrăm ecuația neajustată, înfruntând tot felul de multicolinearitate, dînsă nu multicolinearitate extremă. Oricum acest remediu greu se acceptă de cercetătorii începători atunci când ei se confruntă cu regresia finală, ultima având t-statistici ne semnificative. În comparație cu alternativa posibilă a deplasărilor cauzate de omiterea variabilei importante ori rezultatelor întâmplătoare, t-statisticile joase se par a fi o problemă minoră.

7.2. Eliminarea unei sau mai multor variabile multicolineare

Probabil că cea mai sigură cale a salva ecuația de multicolinearitate semnificativă constă în aruncarea tuturor variabilelor multicolineare. Multicolinearitatea este cauzată de corelarea dintre variabilele explicative; ecuația în lipsa variabilelor multicolineare nu mai este supusă corelării și toate problemele legate de multicolinearitate sunt sistate. Coeficienții variabilelor păstrate măsoară aproape tot impactul comun asupra variabilei dependente a variabilelor explicative multicolineare excluse.

Ca să demonstrăm cum funcționează această metodă vom examina următorul exemplu:

$$\hat{C}_i = -367.83 + 0.5113 * Y_{di} + 0.0427 * LA_i$$

(1.0307) (0.0942)

$$t = \quad \quad 0.496 \quad \quad 0.453 \quad \quad \bar{R}^2 = 0.835$$

C – consumul; Y_d - venitul disponibil; LA - active lichide.

$$\hat{C}_i = -471.43 + 0.9714 * Y_{di}$$

(0.157)

$$t = \quad \quad 6.187 \quad \quad \bar{R}^2 = 0.861$$

$$\hat{C}_i = -199.44 + 0.08876 * LA_i$$

(0.01443)

$$t = \quad \quad 6.153 \quad \quad \bar{R}^2 = 0.860$$

De notat, că eliminarea unei din variabilele multicolineare exclude atât multicolinearitatea între două variabile explicative cât și majorează t-statisticile pe lângă coeficienții variabilelor păstrate. Prin aruncarea Y_d , se majorează t_{LA} de la 0.453 până la 6.153, în același timp aruncarea variabilei schimbă valoarea coeficientului rămas (deoarece variabila aruncată nu se mai menține constantă), așa schimbări dramatice nu sunt de excepție.

Fie că se dorește eliminarea unei variabile, cum se decide care variabilă să se arunce? În cazul multicolinearității severe nu importa care variabilă va fi aruncată. Este lipsit de sens alegerea variabilei pentru aruncare în pofida faptului că ea este cea mai potrivită sau că este de o semnificație sporită (sau are semnul așteptat) în ecuația originală. Din contra, fundamentarea teoretică a modelului va servi drept temei pentru luarea deciziilor în acest sens. În exemplu prezentat există un suport teoretic întru susținerea ipotezei că venitul disponibil determină consumul dar nu activele lichide.

În multe cazuri, simpla soluție – aruncarea unei din variabilele multicolineare este una bună. De exemplu, unii din cercetătorii neexperimentați includ prea multe variabile explicative în ecuațiile de regresie nedorit să se confrunte cu deplasările în variabilele rămase. Prin urmare, ei

au deseori două sau mai multe variabile în ecuație care măsoară în esență aceleași obiecte. În acest caz, variabilele multicolineare nu sunt **irelevante**, deoarece fiecare dintre ele este mult posibil rezonabilă atât teoretic cât și statistic. În schimb, variabilele pot fi numite **inutile**; numai una din ele necesită să reprezinte influența asupra variabilei dependente care o demonstrează fiecare dintre ele. De exemplu, pentru funcția cererii agregate, nu va avea sens să se introducă venitul disponibil și PIB deoarece ambii indicatori măsoară același subiect: venitul. Puțin mai subtilă este concluzia că populația și venitul disponibil nu vor fi ambele incluse în atare funcție agregată a cererii deoarece, din nou, întradevăr măsoară același subiect: volumul pieței agregate. Cu creșterea populației va crește și venitul. Aruncarea variabilelor multicolineare inutile de acest gen nu va face nimic decât se ridice gradul erorii de specificație.

7.3. Transformarea variabilelor multicolineare

Deseori în ecuațiile care se confruntă cu consecințe destul de serioase a multicolarității, ca să autorizăm considerația acțiunilor de remediere, toate variabilele sunt extrem de importante din punct de vedere teoretic. În așa cazuri nici inacțiunea, nici mai ales aruncarea variabilei nu sunt de folos. Oricum, uneori pentru a scăpa cel puțin de unele multicolarități este posibilă transformarea variabilelor din ecuație. Două din cele mai frecvente transformări sunt:

- formarea unei combinații lineare din variabilele multicolineare,
- transformarea ecuației în diferențe finite (sau logaritmi).

Tehnica formării combinației liniare a două sau mai multe variabile multicolineare constă în:

- a) crearea unei variabile noi care este o funcție de variabile multicolineare;
- b) folosirea variabilei obținute pentru a înlocui variabilele vechi în ecuația de regresie.

De exemplu, dacă variabilele X_1 și X_2 sunt puternic multicolineare, o nouă variabilă $X_3 = X_1 + X_2$ (sau în caz general, orice combinație lineară de felul $X_3 = k_1 X_1 + k_2 X_2$ poate fi substituită în modelul reestimat în locul ambelor variabile multicolineare. Această tehnică este utilă atunci când ecuația urmează a fi aplicată pentru date în afara celor observate, deoarece atunci multicolaritatea poate să nu existe sau poate să urmeze același șablon ca și înăuntru eșantionului de date observate. Deseori, în ecuațiile pentru care consecințele multicolarității sunt destul de severe, încât să justifice aplicarea acțiunilor de remediere, toate variabilele sunt extrem de importante din punct de vedere a motivației teoretice. Pentru aceste cazuri nici inacțiunea, nici eliminarea variabilelor nu sunt de folos. Oricum, uneori este posibilă transformarea variabilelor din ecuație într-o reducere a multicolarității. Dezavantajul major al acestei tehnici constă în faptul că ambele variabile au același coeficient în ecuația reestimată. De exemplu, dacă

$$X_{3i} = X_{1i} + X_{2i} \Rightarrow Y_i = \hat{\beta}_0 + \hat{\beta}_3 X_{3i} = \hat{\beta}_0 + \hat{\beta}_3 (X_{1i} + X_{2i}) + u_i.$$

Necesită atenție includerea în combinația lineară a variabilelor care se așteaptă să aibă coeficienți diferiți (cum ar fi diferite semne) sau o diferență extrem de mare a valorilor (cum ar fi mărimi de diferit ordin) fără a ajusta aceste devieri prin folosirea constantelor potrivite (k_3) în ecuația sub formă generală ($X_3 = k_1 X_1 + k_2 X_2$). De exemplu, dacă două variabile multicolineare sunt PIB și rata inflației, atunci o simplă sumă poate inunda complet variația inflației (depinde de unitățile de măsură a variabilelor).

$$X_{3i} = GNP_i + INF_i = 3.250 * GNP_i + 0.08 * INF_i.$$

Să vedem cum se va schimba X_3 atunci când IBP se dublează, tot cu atâta se dublează și X_3 , dar dacă INF se dublează, X_3 aproape deloc nu se schimbă. În majoritatea combinațiilor liniare, necesită a fi întreprins un calcul grijuliu asupra valorilor medii – coeficienților așteptați a variabilelor ce fac parte din combinația liniară. Cu alte cuvinte, variabilele pot să se anuleze una pe alta sau să se inundeze una pe alta după mărime.

Pentru un exemplu de acest gen să formăm o combinație liniară dintre venitul disponibil și activele lichide în funcția de consum și deci să relansăm regresia cu combinația liniară a variabilelor explicative. Pentru a balansa ambele variabile, venitul disponibil poate fi înmulțim cu 10, obținând: $X_{3i} = 10(Yd_i) + LA_i$

Constantele în combinațiile liniare de acest fel se capătă arbitrar, dar ele pot acționa relativ bine. Când X_3 este folosită la înlocuirea ambelor variabile explicative și se estimează ecuația de regresie, obținem:

$$\hat{C}_i = -355.43 + 0.0467 * X_{3i} \\ (0.0073)$$

$$t = 6.362 \quad R_2 = 0.868$$

Comparând rezultatele precedente, observăm că din nou eliminarea multicolinearității semnificativ ridică t-statistica variabilei explicative, în timp ce are un efect mic asupra semnificației ecuației. Este interesant că coeficientul estimat poate fi calculat din estimările precedente a ecuației ca combinație liniară.

Al doilea fel de transformări care poate fi luat în considerație ca un posibil remediu contra multicolinearității severe constă în schimbarea formei funcționale a ecuației. Să vedem cum transformarea ecuației în diferențe finite de ordinul întâi va diminua gradul de multicolinearitate în eșantionul de date (s-ar putea discuta pe marginea transformărilor în log sau alte forme funcționale dar ele sunt efectuate conform aceluiași principiu). Diferențele de ordinul întâi nu sunt alt ceva decât schimbarea în variabila din perioada precedentă și perioada curentă (care se referă ca “delta” sau Δ) $\Delta X_t = X_t - X_{t-1}$.

Dacă ecuația (sau mai multe variabile în ecuație) sunt transformate de la specificarea normală la specificarea în diferențe de ordinul întâi este destul de clar că gradul de multicolinearitate se va reduce esențial pentru două motive. Primul, orice schimbare în definiția variabilelor (cu excepția simplei schimbări liniare) va reduce gradul de multicolinearitate. Al doilea, multicolinearitatea are loc cel mai frecvent (deși bineînțeles nu exclusiv) în seriile temporale de date, pentru care diferențele de ordinul întâi sunt extrem de puțin asemănătoare cu deplasările permanente ascendente pentru agregatele din care ele sunt calculate. De exemplu, PIB crește numai cu 5-6%% anual, iar schimbările în PIB (diferențele de ordinul întâi) pot fluctua auster. Prin urmare, transformarea întregii ecuații sau a unei părți ai ecuației în diferențe prin specificare este în stare să reducă posibilitatea multicolinearității în modelul cu serii temporale.

În timp ce multicolinearitatea severă uneori poate fi diminuată prin trecerea la specificarea ecuației în diferențe de ordinul întâi (sau la alte specificări), schimbarea formei funcționale a ecuației pur și simplu pentru evitarea multicolinearității deseori este în stare să aducă posibile complicații teoretice. De exemplu, modelarea stocurilor nu este de aceeași natură ca și modelarea schimbărilor în stocurile de capital, care reprezintă investițiile, chiar dacă o ecuație derivă din alta. Dacă scopul principal al lansării regresiei constă în modelarea diferențelor de ordinul întâi, atunci modelul poate fi specificat în așa mod. Pe lângă aceasta, la calcularea diferențelor de ordinul întâi, gradul de libertate va fi redus cu o unitate.

7.4. Majorarea numărului de observații.

Altă cale de a trata multicolinearitatea constă în încercarea de a mări volumul setului de observații în așa mod ca să se reducă nivelul multicolinearității.

Ideea ce stă la bază majorării volumului de observații în model constă în faptul că un eșantion de date mai mare (deseori necesitând o nouă colectare de date) va permite estimări mai exacte, în același timp un set mai mare de date, în mod normal, va reduce într-un fel sau altul variațiile coeficienților estimați diminuând impactul multicolinearității chiar dacă gradul de multicolinearitate rămâne același.

Oricum, pentru majoritatea aplicațiilor din economie și business aceasta soluție nu este posibilă. După ce setul de date este completat cu date disponibile care par a fi comparabile, date noi în general sunt greu de găsit sau ele sunt foarte scumpe.

Una din căile de majorare a eșantionului de date constă în comasarea șirurilor temporare și obținerea de date încrucișate. O astfel de combinație a surselor de date, de regulă, constă în completarea cu date încrucișate (de obicei lipsite de multicolinearitate) a seriilor temporare de date multicolineare, astfel deminuând multicolinearitatea în setul comun de date. Problema majoră a acestei reuniuni constă în interpretarea și folosirea estimărilor generale. Până când sunt motive de încredere că modelul teoretic considerat este același în ambele abordări, parametrii estimați obținuți vor fi un fel de funcții asociate ai modelului adevărat în serii temporare și a modelului adevărat încrucișat. În genere, așa combinație a diferitor tipuri de date nu este recomandabilă ca modalitate în reducerea multicolinearității. În majoritatea cazurilor dificultățile unei interpretări necunoscute sunt mai grave decât consecințele cunoscute ale multicolinearității.

7.5. Alegerea unui remediu corect

Nu există o soluție unică în problema cum se va face alegerea remediei contra multicolinearității; o ajustare contra multicolinearității care poate fi de folos pentru o ecuație, este nepotrivită pentru o altă ecuație.

8. Corelarea în serie (autocorelarea)

Corelarea în serie numită tot odată și autocorelare, poate să existe în orice cercetări de investigare în care ordinul observațiilor are careva semnificație. De aceea cel mai frecvent autocorelarea apare în setul de date cu serii temporare. *În esență din corelarea în serie rezultă că termenul erorii stocastice dintr-o perioadă depinde în mod simetric de termenul de eroare stocastice din altă perioadă.* Și deoarece seriile temporare de date se folosesc în multe aplicații econometrice, este important de a înțelege corelarea în serie și consecințele ei pentru estimatorii M.C.M.M.P.

Se va întreprinde o încercare de a răspunde la întrebările:

- Care este esența problemei?
- Care sunt consecințele problemei?
- Cât de periculoasă este problema?
- Ce remedii exista pentru că problema să fie soluționată?

8.1. Corelarea în serie perfectă și imperfectă

Corelarea observațiilor termenului de eroare între ele pe parcursul timpului se numește corelare în serie. În acest compartiment se va discuta descrierea caracterului (naturei) a corelării în serie și deosebirea dintre două forme a fenomenului, corelare în serie “perfectă” și “imperfectă”.

8.1.1. Corelarea în serie perfectă

Corelarea în serie perfectă apare atunci când sunt sfidate ipotezele clasice referitor la faptul că observațiile termenului de eroare nu sunt corelate într-o ecuație specificată corect. Vom aminti că ipoteza clasică afirmă că: $E(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$. Dacă valoarea așteptată a produsului oarecăror două observații arbitrare ai termenului de eroare nu este egală cu “0”, atunci se va spune că termenul de eroare este corelat în serie. Atunci când econometricienii folosesc termenul corelarea în serie fără nici o modificare, ei se referă la corelarea în serie perfectă.

Cel mai frecvent întâlnit tip de corelare în serie se referă la corelarea în serie de ordinul întâi, în care observația curentă a termenului de eroare este o funcție de observația precedentă a termenului de eroare:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad (8.1)$$

unde ε_t este termenul de eroare din ecuația examinată, ρ - este parametru ce descrie relația funcțională între observațiile termenului de eroare; u_t - este termenul de eroare clasic (necorelat în serie). Forma funcțională (8.1.1.1) este așa numita schemă Marcoviană de ordinul întâi și ρ este coeficientul de autocorelare de ordinul întâi. Un astfel mod de corelare în serie se caracterizează prin faptul că una din valorile observațiilor a termenului de eroare afectează direct următoarea valoare observată a termenului de eroare. Mărimea coeficientului ρ indică cât de strânsă este corelarea în serie în ecuație. Dacă ρ este egal cu zero, atunci nu există corelare în serie (deoarece ε_t este egal cu u_t termenul de eroare clasică). Dacă ρ după valoarea absolută nu este mai mare decât unul, valoarea precedentă a termenului de eroare devine mai importantă în determinarea valorii curente a termenului de eroare și un nivel înalt de corelare în serie există. ρ mai mare decât unul după valoarea absolută nu este rezonabil deoarece implică tendința creșterii continue în timp pentru termenul de eroare după valoarea absolută. Prin urmare, se va stabili $-1 < \rho < 1$. Semnul ρ indică caracterul corelării în serie în ecuație. Valoarea ρ pozitivă contribuie la faptul că termenul de eroare va avea tendința pe viitor să-și păstreze semnul pozitiv de la o perioadă la alta. Așa o tendință înseamnă că în caz că ε_t se întâmplă să aibă șansă de a obține valori mari într-o perioadă de timp, următoarele observații vor tinde să rețină o porțiune din valorile originale mari și vor avea același semn ca și observația originală. De exemplu, în modele cu serii temporare un șoc extrem de mare în economie într-o perioadă de timp poate să continue și în câteva perioade ce vor urma. Dacă aceasta se întâmplă, atunci termenul de eroare va tinde să rămână pozitiv pentru un număr de observații, apoi negativ pentru altele câteva, apoi din nou pozitiv. Acest fenomen se numește corelare în serie pozitivă.

Valoarea ρ negativă implică tendința schimbării semnului termenului de eroare în observațiile consecutive de la negativ la pozitiv și invers. Acest fenomen se numește corelare în serie negativă și implică un fel de cicluri (asemănătoare cu mișcarea pendulului) în urma desenării perturbațiilor stocastice. Corelarea în serie negativă de ordinul întâi se caracterizează prin faptul că termenul de eroare tinde să aibă semn opus de la o observație la alta.

De exemplu, corelarea în serie poate să existe în termenul de eroare a ecuației semianuale a cererii pentru careva obiecte sezoniere (cum ar fi luminițe de Crăciun) care nu au variabile "dummy" (fictive) sezoniere. Oricum, în majoritatea aplicațiilor cu serii temporare corelarea în serie negativă se întâlnește mai rar decât corelarea în serie pozitivă.

Corelarea în serie poate lua multe alte forme ce diferă de aceea de ordinul întâi. De exemplu, în modelul trimestrial, termenul de eroare a observației în trimestrul curent poate fi funcțional relatat la observațiile termenului de eroare în același trimestru al anului precedent:

$$\varepsilon_t = \rho \varepsilon_{t-4} + u_t. \quad (8.2)$$

La fel este posibil că termenul de eroare în ecuație să fie funcție de mai multe observații precedente a termenului de eroare: $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + u_t$, așa o formulare este numită corelare în serie de ordinul doi.

8.1.2. Corelarea în serie imperfectă

Prin corelarea în serie *imperfectă* se subînțelege corelarea în serie care *este cauzată de eroarea de specificare cum ar fi variabile omise sau formă funcțională incorectă*. În timp ce corelarea în serie perfectă este cauzată de distribuția fundamentală a termenului de eroare în specificația adevărată a ecuației, care nu poate fi schimbată, corelarea în serie imperfectă este cauzată de eroarea de specificare care deseori poate fi corectată.

Cum se întâmplă că eroarea de specificare cauzează corelare în serie? Vom aminti că termenul de eroare poate fi tratat drept efect de la omiterea variabilelor, nelinearitatea, erori de măsurare și pur și simplu abateri stocastice a variabilei dependente. Aceasta înseamnă că, dacă

noi omitem o variabilă relevantă sau utilizăm o formă funcțională incorectă, atunci o parte din efectul omis care nu poate fi reprezentat de variabilele explicative rămase trebuie să fie absorbit de către termenul de eroare. Termenul de eroare pentru ecuația specificată incorect, prin urmare, include efectul a mai multor variabile omise și/ori o parte din efectul a diferenței dintre forma funcțională proprie și alta aleasă de cercetător. Acest termen de eroare nou poate fi corelat în serie chiar dacă nu este acel adevărat. În acest caz corelarea în serie e cauzată de alegerea specificației de către cercetător și nu de termenul de eroare perfect asociat cu specificarea corectă.

Remediile pentru corelarea în serie depind de tipul de corelare: perfectă sau imperfectă. Nu este surpriză, că cel mai bun remediu pentru corelarea în serie imperfectă, de obicei va fi acela de nu a întreprinde încercări întru omiterea variabilelor din ecuație. Și, prin urmare, majoritatea econometricienilor încearcă să se încredințeze că au obținut specificația cea mai bună posibilă înainte de a petrece o mulțime de timp necăjându-se cu corelarea în serie imperfectă.

Pentru a vedea cum omiterea variabilei poate cauza corelarea în serie a termenului de eroare, să admitem că ecuația adevărată este:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t, \quad (8.3)$$

unde ε_t este termenul de eroare clasică. Dacă X_2 va fi omis accidental din ecuație (sau datele pentru X_2 nu sunt disponibile), atunci $Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t^*$, unde $\varepsilon_t^* = \beta_2 X_{2t} + \varepsilon_t$. Deci termenul de eroare care va fi folosit în cazul omiterii variabilei nu este termenul clasic de eroare ε_t . În schimb el este o funcție de variabila independentă X_2 . Prin urmare, termenul nou de eroare ε_t^* poate fi corelat în serie chiar și atunci când termenul adevărat de eroare ε_t nu este corelat. În special, termenul nou de eroare ε_t^* va tinde spre a fi corelat în serie atunci când:

- variabila X_2 este singură corelată în serie (acest fapt este destul de probabil în seriile temporare);
- mărimea ε_t este mică în comparație cu mărimea $\beta_2 X_{2t}$.

Aceste tendințe se produc, dacă există una sau mai multe variabile omise. Vom menționa primul fapt: eroarea ε_t^* apare cu valoarea diferită de zero deoarece exista corelare în serie imperfectă, estimația M.C.M.M.P. a termenului liber β_0 va fi ajustată la această problemă. Doi: deoarece corelarea în serie imperfectă implică erori de specificare de tipul variabilelor omise, corelarea în serie imperfectă probabil poate fi asociată cu coeficienți estimați deplasați. Atât deplasările cât și corelarea în serie imperfectă vor dispărea odată cu corectarea erorii de specificare.

Vom examina cererea pentru pește pentru a demonstra cum variabilele omise pot cauza corelare în serie în termenul de eroare a ecuației incorect specificate:

$$F_t = \beta_0 + \beta_1 RP_t + \beta_2 \ln Y_{dt} + \beta_3 D_t + \varepsilon_t, \quad (8.4)$$

aici F_t este consumul de pește pe un cap de locuitor într-un an t , RP_t este prețul relativ a peștelui față de carnea de vită în anul t , Y_{dt} este venitul disponibil real pe un cap de locuitor în anul t , D_t este variabila “dummy”, egală cu zero până la decizia Papei și cu o unitate după, și ε_t - termenul clasic de eroare (necorelat în serie). Admitem ca ecuația (8.4) este de o specificație corectă. Ce se va întâmpla cu ecuația în cauză dacă variabila venitul disponibil va fi omisă $F_t = \beta_0 + \beta_1 RP_t + \beta_3 D_t + \varepsilon_t^*$?

Cel mai evident efect va fi acela că coeficienții estimați pe lângă RP_t și D_t vor fi deplasați pe măsura corelării RP_t și D_t cu Y_{dt} . Efectul secundar va fi acela că termenul de eroare acum va include o parte considerabilă din efectul eliminat al veniturii disponibil asupra consumului de pește, faptul că ε_t^* este egal cu $\varepsilon_t + \beta_2 \ln Y_{dt}$. Este rezonabil de așteptat că venitul disponibil (prin urmare și log lui) poate urma un șablon moderat de corelare în serie:

$$\ln Y_{dt} = f(\ln Y_{dt-1}) + u_t \quad (8.5)$$

de ce este probabil? Vom privi graficul Y_{dt} în timp. Observăm că creșterea continuă a venitului disponibil în timp îl determină și pe $\log Y_{dt}$ să acționeze în mod autocorelat sau corelat în serie. Dar, dacă venitul disponibil este corelat în serie (și dacă impactul lui nu este relativ mai mic decât ε_t), atunci ε_t^* este, mult probabil, să fie la fel corelat în serie, ce poate fi exprimat ca: $\varepsilon_t^* = \rho \varepsilon_{t-1}^* + u_t$, unde ρ reprezintă coeficientul de corelare în serie, dar u_t este termenul de eroare clasic. Acest exemplu ne-a demonstrat că într-adevăr este posibil ca variabila eliminată să introducă corelare în serie “imperfectă” în ecuație.

Un alt tip răspândit de corelare în serie este acela, cauzat de forma funcțională incorectă. În această situație alegerea incorectă a formei funcționale poate cauza corelarea în serie a termenului de eroare. Să admitem că ecuația adevărată este prezentată sub forma logaritmică completă:

$$\ln Y_t = \beta_0 + \beta_1 \ln X_{1t} + \varepsilon_t \quad (8.6)$$

dar în locul ei este lansată regresia liniară: $Y_t = \alpha_0 + \alpha_1 X_{1t} + \varepsilon_t^*$. Termenul nou de eroare ε_t^* acum este o funcție de termenul adevărat de eroare ε_t și de la diferența între forma liniară și forma în log completă. Din figura . observăm că aceste diferențe urmează compartimente moderat autoregresive. Diferențele pozitive tind a fi urmate de diferențe pozitive și diferențe negative tind a fi urmate de diferențe negative. Prin urmare, folosirea formei liniare atunci când una neliniară este mai potrivită de obicei rezultă cu corelare în serie pozitivă imperfectă.

9. Consecințele corelării în serie

Consecințele corelării în serie sunt complet diferite după caracter de consecințele ale problemelor discutate anterior. Variabilele omise, variabilele irelevante și multicolinearitatea toate au indicii externi care pot fi recunoscuți complet. Fiecare problemă schimbă coeficienții estimați și erorile standard într-un mod concret, și analizarea acestor schimbări deseori oferă destulă informație ca problema să fie soluționată. Cum vom vedea, corelarea în serie îi mai mult probabil să aibă simptome interne și afectează ecuația estimată pe o cale care nu este ușor de observat prin examinare numai a rezultatelor ca atare.

Există consecințe majore a corelării în serie:

- Corelarea în serie perfectă nu cauzează deplasări în coeficienții estimați;
- Corelarea în serie contribuie la creșterea varianțelor distribuțiilor β .
- Corelarea în serie induce M.C.M.M.P. să subestimeze varianțele (și erorile standarde) a coeficienților.

9.1. Sinteza consecințelor corelării în serie

Existența în ecuație corelării în serie a termenului de eroare violează ipoteza clasică prin urmare estimările ecuației cu M.C.M.M.P. au de suportat cel puțin trei consecințe.

9.1.1. Corelarea în serie perfectă nu cauzează deplasări în coeficienți

Să ne amintim că ceea mai importantă proprietate a tehnicii de estimare prin M.C.M.M.P. constă în faptul că estimatorii liniari nedeplasați au o varianță minimă. Dacă erorile sunt corelate în serie, una din ipotezele teoremei Gauss-Marcov este violată și anume ea cauzează deplasări în coeficienții estimați. Să admitem că este cunoscut faptul că termenul de eroare în ecuația ce urmează

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t \quad (9.1.1)$$

este supus corelării perfecte în serie de ordinul întâi:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad (9.1.2)$$

aici u_t este termenul erorii clasice (necorelat în serie). Dacă ecuația (9.1) este corect specificată și este estimată cu M.C.M.M.P., atunci estimațiile coeficienților obținuți vor fi nedepasate: $E(\hat{\beta}_1) = \beta_1$; $E(\hat{\beta}_2) = \beta_2$. Corelarea perfectă în serie nu introduce deplasări în procedura de estimare. Aceasta concluzie este justă atât pentru corelarea pozitivă în serie cât și corelarea negativă în serie de ordinul unu. Dacă corelarea în serie este imperfectă, oricum deplasările pot fi introduse prin utilizarea specificației incorecte.

Lipsa deplasărilor nu înseamnă cu necesitate că estimațiile M.C.M.M.P. ale coeficienților ecuației corelate în serie vor fi strâns apropiate de valorile adevărate ale coeficienților, deoarece o singură estimație observată în realitate poate parveni dintr-un număr mare al valorilor posibile. Plus la aceasta, eroarea standard a acestor estimații va fi la sigur majorată de corelarea în serie. Această majorare va spori probabilitatea devierii suficiente a valorii $\hat{\beta}$ de la valoarea adevărată β . În acest caz valorile nedepasate cu o distribuție $\hat{\beta}_s$ sunt centrate în jurul valorii adevărate β .

9.1.2. Corelarea în serie mărește varianțele distribuțiilor $\hat{\beta}$

În timp ce violarea ipotezei clasice nu cauzează deplasări, ea poate afecta principala concluzie a teoremei Gauss-Marcov, aceea a varianței minime. În special atunci, când ipoteza clasică este violată, este cu neputință de a dovedi că estimațiile M.C.M.M.P. $\hat{\beta}_s$ au o varianță minimă. Prin urmare, termenul de eroare este corelat în serie atunci, când M.C.M.M.P. nu mai oferă varianță minimă a coeficienților estimați.

Termenul de eroare corelat în serie impune variabila dependentă să fluctueze în același mod în care procedeul de estimare a M.C.M.M.P. o atribuie variabilelor independente. Deci, este mult probabil că M.C.M.M.P. nu oferă estimări adevărate pentru β în fața corelării în serie cauzată de balansare, $\hat{\beta}_s$ rămânând nedepasate deoarece supraestimarea este tot atât de probabilă ca subestimarea; oricum aceste erori majorează varianța distribuției estimațiilor, sporind mărimea cu care orice estimație e probabil să difere de la valoarea adevărată β . Într-adevăr, poate fi demonstrat că în cazul când termenul de eroare este distribuit în felul $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, atunci varianța $\hat{\beta}_s$ este funcție de ρ . Cu cât este mai mare ρ cu atât este mai mare varianța $\hat{\beta}_s$. Efectul corelației în serie asupra distribuției coeficienților în demonstrația sa grafică rezultă cu faptul că distribuția $\hat{\beta}_s$ din ecuația corelată se regăsește în jurul coeficienților β adevărați, dar este mai plată decât distribuția din ecuația fără corelare în serie.

9.1.3. Corelarea în serie cauzează M.C.M.M.P. să subestimeze varianța (și erorile standard) a coeficienților

Dacă corelarea în serie mărește varianțele (la fel și erorile standard) a $\hat{\beta}_s$, atunci putem presupune că $\hat{\sigma}(\hat{\beta}_s)$ obținută prin M.C.M.M.P. tot va crește, însă nu de fiecare dată. În schimb aceste varianțe $\hat{\sigma}(\hat{\beta}_s)$ au tendința de a fi destul de mici. Prin urmare, corelarea în serie majorează devierile standard a coeficienților estimați, dar în așa mod care nu este evidențiat de estimațiile M.C.M.M.P.

M.C.M.M.P. are tendința de a subestima erorile standard a coeficienților ecuației corelate în serie, deoarece corelarea în serie rezultă din compartimentul de observații care permit o aproximare mai bună decât aceea pe care observațiile ecuației necorelate în serie pot s-o justifice. Aproximația mai bună rezultă nu numai din subestimațiile erorilor standard ale $\hat{\beta}_s$, dar și din erorile standard a reziduurilor, încât nici pe t - statistici, nici pe F - statistică nu se poate baza în prezența corelării în serie perfecte.

În special, tendința M.C.M.M.P. de a subestima $\hat{\sigma}(\hat{\beta}_s)$ va contribui la supraestimarea t - statisticilor ale coeficienților estimați, întrucât:

$$t = \frac{(\hat{\beta} - \beta_{H_0})}{\hat{\sigma}(\hat{\beta})}. \quad (9.1.3)$$

Dacă $\hat{\sigma}(\hat{\beta})$ prea mic cauzează t - valoare mare pentru un coeficient distinct, atunci este mult probabil că ipoteza nulă va fi respinsă $H_0: (\beta = 0)$, în timp ce ea este adevărată. În esență, M.C.M.M.P. produce confuzie în vederea semnificației rezultatului concret. Corelarea în serie nu numai majorează devierile standard, dar deseori conduce la concluzii greșite care fac dificilă acapararea acestei creșteri de M.C.M.M.P.

9.2. Testul Durbin-Watson

Cel mai larg utilizat test pentru depistarea corelării în serie este d - testul Durbin-Watson.

9.2.1. Statistica Durbin-Watson

Statistica Durbin-Watson d este calculată prin examinarea reziduurilor ale estimației concrete a ecuației, și se folosește pentru determinarea faptului existenței corelării în serie de ordinul întâi în termenul de eroare inclus în ecuație. Este important ca d - statistica Durbin-Watson să se folosească numai atunci când ipotezele care fundamentează acest fenomen sunt de față:

- Modelul de regresie include termenul de intersecție (termenul liber).
- Corelarea în serie de ordinul întâi există: $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$, unde ρ este coeficientul corelării

în serie și u_t este termenul erorii clasice (necorelat în serie), $\rho = \sum_{t=2}^n u_t u_{t-1} / \sum_{t=2}^n u_{t-1}^2$.

- Modelul de regresie nu trebuie să conțină variabile întârziate dependente în calitate de variabile independente (în acest caz d - statistica este deplasată spre 2, dar poate fi folosit testul n - Durbin - Watson sau altele).

Ecuația pentru d - statistica Durbin-Watson pentru T observații este următoarea:

$$d = \sum_{t=2}^T (u_t - u_{t-1})^2 / \sum_{t=1}^T u_t^2, \quad (9.2.1)$$

aici u_t sunt reziduurile obținute prin M.C.M.M.P. Vom menționa că numărătorul are cu o observație mai puțină decât numitorul, deoarece o observație necesită a fi utilizată pentru calcularea u_{t-1} . d - statistica Durbin-Watson este egală cu zero atunci când există corelație în serie pozitivă extremă, este egală cu doi dacă nu există corelație în serie și este egală cu 4 dacă există corelație în serie negativă extremă. Vom demonstra aceasta, prin a introduce datele reziduurilor respective în ecuația (9.8).

✓ $d = 0$. Corelarea în serie pozitivă extremă. În acest caz, $u_t = u_{t-1}$, deci $(u_t - u_{t-1}) = 0$ și $d = 0$, ($\rho = 1$).

✓ $d \approx 4$. Corelare în serie negativă extremă. În acest caz, $u_t = -u_{t-1}$ și $(u_t - u_{t-1}) = 2u_t$.

Substituind în ecuația (9.8), obținem $d = \sum (2u_t)^2 / \sum u_t^2 \Rightarrow d \approx 4, \rho = -1$.

✓ Nu exista corelare în serie: $d \approx 2, \rho = 0$.

9.2.2. Utilizarea d - testului Durbin-Watson

Testul Durbin-Watson nu este frecvent utilizat din două motive. Primul, econometricienii aproape niciodată nu testează ipoteza unidirecțională zero a existenței corelării în serie negative

în reziduuri deoarece corelarea în serie negativă, după cum a fost menționat mai sus, este foarte greu de explicat teoretic în analiza economică sau de business. Existența ei înseamnă că corelarea în serie imperfectă probabil că este cauzată de eroarea de specificare. Doi, uneori testul Durbin-Watson este neconcludent. În timp ce regula de luarea a deciziei de fiecare dată are numai regiuni de “acceptare” sau de “respingere”, testul Durbin-Watson are a treia posibilitate, numită regiune neconcludentă.

În aceste circumstanțe, utilizarea d - testului Durbin-Watson este aproape similară utilizării t - testului sau F - testului. Pentru a testa corelarea în serie pozitivă sunt necesare următoarele etape:

1. Obținem reziduurile M.C.M.M.P. din ecuația supusă testării și calculăm d - statistica cu ajutorul formulei (9.8).
2. Determinăm volumul eșantionului și numărul variabilelor explicative și apoi consultăm tabelele statisticilor pentru a găsi valorile critice: d_U - maximală și d_L - minimală respectiv.
3. Dat fiind anunțată ipoteza $H_0 : \rho \leq 0$, care infirmă corelarea în serie pozitivă și ipotezele unidirecționale $H_A : \rho > 0$ (există corelare în serie pozitivă).

Ceea mai potrivită regula de luarea a deciziei este:

dacă $d < d_L$ se respinge ipoteza H_0 :

dacă $d > d_U$ nu se respinge ipoteza H_0 :

dacă $d_L \leq d \leq d_U$ ipoteza H_0 : nu e convingătoare.

În unele circumstanțe cel potrivit va fi testul bidirecțional. În acest caz vor fi utilizate numai etapele 1 și 2 dar etapa 3 nu va fi utilizată.

Dat fiind anunțate ipotezele bilaterale de alternativă:

$H_0 : \rho = 0$ (nu-i corelare în serie)

$H_A : \rho \neq 0$ (este corelare în serie),

ceea mai potrivită regula de luare a deciziei va fi:

dacă $d < d_L$ se respinge H_0 :

dacă $d > 4 - d_L$ se respinge H_0 :

dacă $4 - d_U > d > d_U$ nu se respinge H_0 :,

în celelalte cazuri H_0 : nu e concludentă

9.3. Estimarea modelelor cu autocorelarea erorilor

Corelarea în serie de ordinul întâi presupune că valoarea termenului de eroare în momentul t ε_t depinde de valoarea termenului de eroare în momentul $t-1$ ε_{t-1} . Prin urmare există un model de regresie de forma:

$$\varepsilon_t = \beta_0 + \beta_1 \varepsilon_{t-1} + u_t,$$

unde β_0, β_1 sunt parametrii ecuației de regresie. În conformitate cu formulele M.C.M.M.P.

avem: $\beta_0 = \bar{\varepsilon}_t - \beta_1 \bar{\varepsilon}_{t-1}$; $\beta_1 = \frac{\overline{\varepsilon_t \varepsilon_{t-1}} - \bar{\varepsilon}_t \bar{\varepsilon}_{t-1}}{\overline{\varepsilon_{t-1}^2} - \bar{\varepsilon}_{t-1}^2}$, $\bar{\varepsilon}_t = \bar{\varepsilon}_{t-1} = 0$, deoarece ε_t sunt reziduurile obținute din

ecuații cu M.C.M.M.P. și conform ipotezelor întru funcționarea M.C.M.M.P.

$$\sum \varepsilon_t = 0 \Rightarrow \bar{\varepsilon}_t = \bar{\varepsilon}_{t-1} = 0. \text{ Atunci avem } \beta_0 = 0, \text{ iar } \beta_1 = \frac{\overline{\varepsilon_t \varepsilon_{t-1}}}{\overline{\varepsilon_{t-1}^2}} = \frac{\sum_{t=2}^n \varepsilon_t \varepsilon_{t-1}}{\sum_{t=2}^n \varepsilon_{t-1}^2} \approx \rho_1^\varepsilon \text{ care este}$$

coeficientul de autocorelare a reziduurilor de ordinul unu. Deci, avem: $\varepsilon_t = \rho_1^\varepsilon \varepsilon_{t-1} + u_t$, u_t - termenul erorii stocastice clasic. Vom nota că $|\rho_1^\varepsilon| < 1$. Ținând cont de ultima relație, obținem:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \rho_1^\varepsilon \varepsilon_{t-1} + u_t. \quad (9.3.1)$$

Vom considera abordarea de bază $Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t$ la estimarea parametrilor ecuației de regresie în cazul când are loc autocorelarea reziduurilor. Înscriem modelul de regresie enunțat pentru $t = t-1$:

$$Y_{t-1} = \beta_0 + \beta_1 X_{1t-1} + \varepsilon_{t-1}, \quad (9.3.2)$$

vom înmulți ambele părți ale ecuației (9.3.2) la ρ_1^ε ,

$$\rho_1^\varepsilon Y_{t-1} = \rho_1^\varepsilon \beta_0 + \rho_1^\varepsilon \beta_1 X_{1t-1} + \rho_1^\varepsilon \varepsilon_{t-1}, \quad (9.3.3)$$

și vom extrage (9.3.2) din (9.3.1) după ce obținem:

$$Y_t - \rho_1^\varepsilon Y_{t-1} = \beta_0 - \rho_1^\varepsilon \beta_0 + \beta_1 X_{1t} - \rho_1^\varepsilon \beta_1 X_{1t-1} + \varepsilon_t - \rho_1^\varepsilon \varepsilon_{t-1} \quad (9.3.4)$$

$$\text{sau } Y'_t = \beta'_0 + \beta_1 X'_{1t} + u'_t, \quad (9.3.5)$$

$$\text{În (9.3.5)} \quad Y'_t = Y_t - \rho_1^\varepsilon Y_{t-1} \quad (9.3.6)$$

$$X'_t = X_t - \rho_1^\varepsilon X_{t-1} \quad (9.3.7)$$

$$u'_t = \varepsilon_t - \rho_1^\varepsilon \varepsilon_{t-1} \quad (9.3.8)$$

$$\beta'_0 = \beta_0(1 - \rho_1^\varepsilon) \quad (9.3.9)$$

Deoarece u'_t este termen de eroare stocastic necorelat, pentru estimarea parametrilor ecuației (9.3.5) se aplică M.C.M.M.P. simplă.

În concluzie, atunci când erorile ecuației inițiale sunt autocorelate, pentru estimarea parametrilor ecuației de regresie se utilizează M.C.M.M.P.G. și este necesar să se îndeplinească următoarele condiții:

- să se transforme variabilele Y_t și X_t la forma (9.3.6)-(9.3.7),
- să se aplice M.C.M.M.P. la ecuația (9.3.5) entru estimarea parametrilor β'_0, β'_1 ,
- să se calculeze parametrul $\beta'_0 / (1 - \rho_1^\varepsilon) = \beta_0$,
- să se înscrie ecuația inițială.

M.C.M.M.P.G. este o analogie a metodei diferențelor finite. Numai că se scade din Y_t și X_t nu valoarea deplină a Y_{t-1} (sau X_{t-1}), dar numai careva parte din ele - $r_1^\varepsilon Y_{t-1}$ sau $r_1^\varepsilon X_{t-1}$. Atunci când $r_1^\varepsilon = 1$, aceasta metodă este metoda diferențelor finite de ordinul 1, întrucât $Y'_t = Y_t - Y_{t-1}$;

$$X'_t = X_t - X_{t-1}$$

În concluzie, dacă valoarea d - testului Durbin-Watson se apropie de 0, aplicarea metodei diferențelor finite de ordinul întâi este destul de motivată. Dacă $r_1^\varepsilon = 1$, termenul de eroare este corelat negativ în serie, atunci metoda expusă se modifică în felul următor.

$$Y'_t = Y_t - (-1)Y_{t-1} = Y_t + Y_{t-1} \quad (9.3.10)$$

$$X'_t = X_t - (-1)X_{t-1} = X_t + X_{t-1} \quad (9.3.11)$$

$$\text{Deoarece,} \quad \beta'_0 = \beta_0(1 - (-1)) = 2\beta_0 \quad (9.3.12)$$

$$\text{avem} \quad Y_t + Y_{t-1} = 2\beta_0 + \beta(X_t + X_{t-1}) + u_t \quad (9.3.13)$$

$$\text{și} \quad (Y_t + Y_{t-1})/2 = \beta_0 + \beta(X_t + X_{t-1})/2 + u_t/2 \quad (9.3.14)$$

În esență, în modelul (9.3.14) se determină mediile a două perioade pentru fiecare serie și apoi pentru datele medii obținute cu ajutorul M.C.M.M.P.G. se estimează parametrii β_0, β_1 .

Problema principală constă în determinarea estimației r_1^ε , ca să putem aplica aceasta metodă. Există o mulțime de procedee pentru determinarea acestei estimări. Însă abordarea de bază o constituie evaluarea acestui coeficient nemijlocit din estimările obținute pentru ecuația inițială de regresie $r_1^\varepsilon = 1 - d/2$, d este testul Durbin-Watson.

Ipoteze unilaterale

$H_0 : \rho \leq 0$ (lipsește autocorelarea variabilelor reziduale);

$H_A : \rho > 0$ (autocorelarea variabilelor reziduale are loc);

$d < d_L$ (ipoteza H_0 se respinge);

$d_L \leq d \leq d_U$ (domeniul de incertitudine);

$d > d_U$ (ipoteza H_0 se acceptă).

Ipoteze bilaterale de alternativă

$H_0 : \rho = 0$ (lipsește autocorelarea variabilelor reziduale);

$H_A : \rho \neq 0$ (autocorelarea variabilelor reziduale are loc);

$d < d_L$ (ipoteza H_0 se respinge);

$d > 4 - d_L$ (ipoteza H_0 se respinge);

$4 - d_U > d > d_U$ (ipoteza H_0 se acceptă) în restul cazurilor există domeniul de incertitudine.

O transformare grijulie a variabilelor întru corectarea eteroschedasticității, deși nu evita corelarea falsă, datorată valorilor mari, uneori poate fi o abordare reușită în soluționarea acestor probleme. De notat totuși, că nu fiecare variabilă în ecuație este tratată în același mod (spre deosebire de M.C.M.M.P.G. ponderată). Fiecare variabilă în modelul cu date încrucișate poate fi examinată în vederea transformărilor posibile care se vor solda cu interpretări semnificative și complete a ecuației de regresie.

10. Eteroschedasticitatea

Eteroschedasticitatea este rezultatul violării ipotezei clasice referitor la faptul că observațiile termenului de eroare au o varianță constantă (fac parte dintr-o populație cu o varianță constantă). Ipoteza varianței constante pentru diferite observații a termenului de eroare (omoschedasticitatea) nu este de fiecare dată una realistă. De exemplu, în modelul care măsoară înălțimea, să comparăm eroarea cu un inch (2.54 cm) la măsurarea înălțimii unui jucător de basket și eroarea cu un inch la măsurarea înălțimii șoricelului. E mult probabil că termenul de eroare asociat cu înălțimea basketbolistului va parveni din distribuția cu o varianță mai mare decât aceea asociată cu înălțimea șoricelului. Cum se va demonstra, distincția dintre eteroschedasticitate și omoschedasticitate este importantă deoarece M.C.M.M.P. aplicată la modelele eteroschedastice, nu mai este estimator de o varianță minimă (rămânând totuși nedeplasat).

Deseori eteroschedasticitatea apare atunci când datele sunt de așa natură că există o diferență mare dintre valoarea cea mai mare observată și valoarea cea mai mică observată. Devierea mare între mărimile observațiilor în populație contribuie la probabilitatea sporită ca distribuția termenului de eroare să aibă pentru observațiile mai mari o varianță mai mare, în timp ce distribuția termenului de eroare pentru observațiile mici are o varianță mică.

Poate fi ușor obținută o diferență mare între cele mai mici și cele mai mari valori ale variabilelor în mulțimea de date încrucișate. Vom aminti că în modelele cu date încrucișate variabilele sunt observate în același timp, dar pentru diferite obiecte (de exemplu, persoane, state, regiuni etc). Deoarece modelele încrucișate deseori includ observații de diferite mărimi în același exemplu, eteroschedasticitatea este greu de evitat în tematicile economice studiate încrucișat.

Punerea accentului pe modelele încrucișate nu înseamnă că eteroschedasticitatea este imposibilă în modelele cu serii temporare și nici nu se exclude posibilitatea că variabilele omise

pot cauza eteroschedasticitate imperfectă în orice tipuri de date. Oricum la mod general, eteroschedasticitatea cu probabilitate mai mare poate avea loc în modelele cu date încrucișate decât în modelele cu serii temporare.

În acest context, se va încerca să dea răspunsul la câteva întrebări ce țin de eteroschedasticitate, care au fost oglindite pentru multicolinearitate și corelare în serie.

1. Care este esența problemei?
2. Care sunt consecințele problemei?
3. Cum problema se depistează?
4. Ce remedii pentru așa problemă sunt disponibile?

10.1. Eteroschedasticitatea perfectă și imperfectă

Eteroschedasticitatea perfectă este aceea care poate fi cauzată de termenul de eroare a ecuației specificate corect, în timp ce eteroschedasticitatea imperfectă este cauzată de eroarea de specificație cum ar fi variabilele omise.

10.1.1. Eteroschedasticitatea perfectă.

Eteroschedasticitatea perfectă să referă la eteroschedasticitatea care este funcție ce depinde de termenul de eroare a ecuației de regresie *corect specificată*. Utilizarea cuvântului “eteroschedasticitate” fără modificări (cum ar fi perfect sau imperfect) implică eteroschedasticitatea perfectă.

Așa tip de eteroschedasticitate apare atunci când în *ecuația specificată corect* ipoteza clasică care presupune că varianțele termenului de eroare sunt constante, este violată. Vom reaminti, că ipoteza presupune că:

$$\text{VAR}(\varepsilon_i) = \sigma^2, (i = 1, 2, \dots, n) \quad (10.1.1)$$

Dacă aceasta presupunere are loc, toate observațiile ale termenului de eroare pot fi imaginate dat fiind prezentate printr-o distribuție asemănătoare cu valoarea medie zero și varianța σ^2 . Acest σ^2 nu se schimbă de la observație la alta a termenului de eroare; aceasta proprietate se numește homoschedasticitate.

În cazul eteroschedasticității, varianța termenului de eroare nu este constantă, în schimb, varianța distribuției termenului de eroare depinde exact de observația în discuție:

$$\text{VAR}(\varepsilon_i) = \sigma_i^2, (i = 1, 2, \dots, n). \quad (10.1.2)$$

De menționat că diferența dintre (10.1) și (10.2) constă în prezența indicelui "i" pe lângă σ^2 , care denotă faptul că varianța termenului de eroare în condiții de eteroschedasticitate se schimbă în dependență de observație în loc să fie constantă pentru orice observație.

Altă cale de a ilustra eteroschedasticitatea constă în prezentarea grafică a mulțimei în care unele observații a termenului de eroare au distribuții mai plate decât altele. Cea mai simplă situație e aceea pentru care observațiile termenului de eroare pot fi grupate numai în două distribuții diferite, lată și "îngustă". Aceasta versiune simplă a problemei poate fi numită eteroschedasticitate discretă. În acest caz ambele distribuții vor fi centrate în jurul punctului zero, însă una va avea o varianță mai mare decât alta.

Eteroschedasticitatea ea forme mult mai complexe, oricum: numărul diverselor modele cu eteroschedasticitate practic nu este limitat, chiar analiza unui procent mic din aceste alternative este o sarcină grea. În continuare se vor examina principiile generale ale eteroschedasticității, concentrându-se asupra celor mai frecvent specificate modele de eteroschedasticitate perfectă. Ceea ce nu înseamnă că econometricienii se axează numai pe un tip de eteroschedasticitate.

Vom examina un model cu eteroschedasticitate, în care varianța termenului de eroare este relatată la variabila exogenă Z_i . Pentru ecuația de regresie tipică:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad (10.1.3)$$

varianța termenului de eroare clasică ε_i în condițiile propuse va fi egală cu:

$$\text{VAR}(\varepsilon_i) = \sigma^2 Z_i^2, \quad (10.1.4)$$

aici variabila Z_i poate fi egală sau nu poate fi egală cu una din variabilele X_s din ecuație. Variabila Z_i se numește factor de proporționalitate deoarece varianța termenului de eroare se schimbă proporțional cu pătratul de Z_i . Cu cât este mai mare Z_i , cu atât este mai mare varianța distribuției termenului de eroare pentru observația " i ". Pot fi "n" diferite distribuții, câte una pentru fiecare observație. În funcție de numărul valorilor distincte pe care le are variabila Z_i pot fi prezentate observațiile termenului de eroare.

Ce reprezintă în realitate factorul de proporționalitate Z_i ? Cum este posibil ca o variabilă exogenă, cum ar fi Z_i , să schimbe întreaga distribuție a termenului de eroare? Să ne adresăm la funcția care relatează consumul gospodăriilor la venitul lor. Cheltuielile gospodăriilor populației cu venit mic este imposibil să varieze după valoarea absolută ca și cheltuielile gospodăriilor populației cu venit mare, deoarece schimbarea cu 10% în cheltuieli pentru familiile cu venit mare atrage mai mulți bani decât schimbarea cu 10% a acelor cu venit mic. Pe lângă aceasta, cota parte din bugetul familiilor cu venit mic care trebuie să fie cheltuită pentru necesități este mult mai mare decât aceea din bugetul familiilor cu venit mare. În cazul dat Y_i va reprezenta cheltuielile de consum dar factorul de proporționalitate Z_i va oglindi venitul gospodăriilor populației. Dacă venitul gospodăriilor populației crește, atunci și varianța termenului de eroare în ecuație de asemenea va fi de natura să explice cheltuielile respective.

Acest exemplu ne demonstrează faptul că eteroschedasticitatea e mult probabil să apară în modele cu date încrucișate, deoarece există o variație mare în valorile variabilei dependente incluse. Bunăoară, abaterile exogene care pentru familiile cu venit mic se arată a fi semnificative, pentru familiile cu venit mare pot să pară minuscule.

În același timp, eteroschedasticitatea poate să apară în modelele cu serii de date temporare cel puțin în două situații care diferă de acelea din modelele cu date încrucișate cu un număr mare de variații în valorile variabilei dependente:

- Eteroschedasticitatea poate să apară în modelele cu date sub formă de serii temporare cu schimbări mari în variabila dependentă (rata de schimbare a variabilei dependente este mare). Dacă are loc o creștere extrem de mare în industrie, atunci e mult probabil ca varianța termenului de eroare să crească în aceeași măsură. Însă atare fenomen nu are loc în seriile de timp cu o rată joasă de schimbare.
- Eteroschedasticitatea poate să apară în orice model, cu serii de date temporare în care calitatea datelor colectate se schimbă dramatic. Cum numai tehnica de colectare a datelor devine mai bună, varianța termenului de eroare va diminua, deoarece erorile de măsurare iau parte din termenul de eroare. Atunci când erorile de măsurare se micșorează, se micșorează și varianța termenului de eroare.

10.1.2. Eteroschedasticitatea imperfectă

Eteroschedasticitatea cauzată de erori de specificare, cum ar fi variabilele omise, se referă la eteroschedasticitatea imperfectă. În același timp forma funcțională improprie puțin probabil să cauzeze eteroschedasticitatea imperfectă, pe când ea produce corelarea în serie imperfectă, cele două concepte fiind similare sub mai multe aspecte.

Variabila omisă e posibil să conducă la eteroschedasticitatea termenului de eroare deoarece o parte din efectul omis nu este reprezentat nici de o variabilă explicativă prezentă în ecuație, deci este incorporat de termenul de eroare. Dacă acest fenomen are o componentă eteroschedastică, termenul de eroare al ecuației respecificate poate fi eteroschedastic chiar dacă termenul de eroare al ecuației adevărate nu este. Aceasta distincție este importantă deoarece în condițiile eteroschedasticității imperfecte remediul corect constă în încercarea de a găsi variabila eliminată și de a o include în ecuația de regresie. Deci înainte de a purcede la depistarea sau a remedierea eteroschedasticității perfecte este important să existe siguranța specificației corectă.

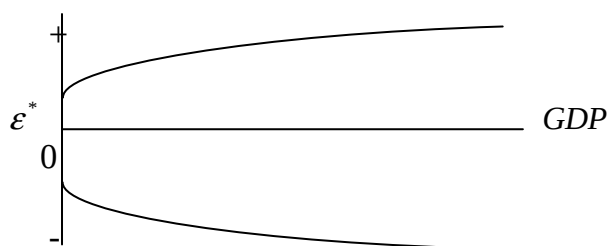
De exemplu, să considerăm un studiu încrucișat al importurilor în anul 1990 al unui număr de națiuni de diferită mărime. Pentru simplitate, se presupune că cel mai bun model a importurilor pentru națiuni în abordarea încrucișată include o relație pozitivă referitor la PIB-ul respectiv și o relație pozitivă privind prețul relativ (care include impactul ratei de schimb) între aceste țări și restul lumii. În aceste circumstanțe modelul “adevărat” va fi următorul:

$$M_i = f(PIB_i, PR_i) = \beta_0 + \beta_1 PIB_i + \beta_2 PR_i + \varepsilon_i, \quad (10.1.5)$$

unde M_i sunt importurile (în \$) a națiunii i_s ; PIB_i este produsul intern brut (în \$) a națiunii i_s ; PR_i este raportul dintre prețul domestic al mărfurilor normal comercializate (convertit în \$ cu ajutorul ratei de schimb) și prețurile mondiale ale acestor bunuri (măsurate în \$) pentru națiunea i_s ; ε_i este termenul erorii clasice.

Acum să admitem că ecuația este lansată fără PIB , atunci ecuația va lua forma: $M_i = \beta_{0i} + \beta_2 PR_i + \varepsilon_i^*$, unde termenul de eroare a ecuației respecificate, ε^* , este funcție de la PIB , variabila aruncată și termenul noneteroschedastic de eroare ε : $\varepsilon_i^* = \beta_0 + \beta_1 PIB_i$.

În măsura în care prețul relativ nu acționează ca o variabilă «proxy» (de înlocuit) față de PIB , termenul de eroare nu încorporează efectul variabilei omise. Dacă acest nou efect are o varianță mai mare pentru valorile PIB mai mari, ce pare a fi probabil, termenul nou de eroare ε^* , este eteroschedastic. Impactul acestui efect în același timp depinde de mărimea $\beta_1 PIB_i$, componenta comparabilă cu valoarea absolută a componentei tipice ε_i . Cu cât este mai mare cota parte a variabilei omise în ε_i^* , cu atât e mai probabilitatea prezenței eteroschedasticității imperfecte. Deci, termenului de eroare ε_i^* prezentat grafic în raport cu PIB , apare în figura de mai jos. Observăm că, valorii mai mari ai PIB -ui îi corespunde varianța mai mare a termenului de eroare.



10.2. Consecințele eteroschedasticității

Fie că s-a stabilit eteroschedasticitatea termenului de eroare din ecuație este, atunci ce impact ar avea acest fenomen asupra estimațiilor coeficienților? Consecințe ale eteroschedasticității în linii mari sunt aproape identice cu acelea ale corelării în serie, deși aceste două probleme sunt complet diferite.

În cazul când termenul de eroare al ecuației este eteroschedastic, există trei consecințe majore:

- Eteroschedasticitatea perfectă nu cauzează deplasări în estimările coeficienților cu metoda celor mai mici pătrate. Deci, putem afirma că o ecuație corect specificată care conține eteroschedasticitate perfectă are următoarele proprietăți: $E(\beta_s) = \beta_s, \forall s$. Ecuația cu eteroschedasticitate imperfectă cauzată de variabilele omise, cu siguranță, va avea o posibilă deplasare de specificare.
- Eteroschedasticitatea mărește varianța distribuțiilor $\hat{\beta}_s$. Dacă termenul de eroare în ecuație este eteroschedastic în funcție de factorul de proporționalitate Z :

$$VAR(\varepsilon_i) = \sigma^2 Z_i^2, \quad (10.2.1)$$

atunci varianța $\hat{\beta}_s$ este funcție de Z :

$$VAR^{**}(\hat{\beta}_s) = f(Z^2) \cdot [VAR(\hat{\beta}_s)], \quad (10.2.2)$$

unde $VAR^{**}(\hat{\beta}_s)$ este varianța cu eteroschedasticitate; $f(Z^2)$ indică o funcție pozitivă de Z , factorul de proporționalitate care “cauzează” eteroschedasticitatea în ecuația (10.7), iar $[VAR(\hat{\beta}_s)]$ este varianța fără eteroschedasticitate. Dacă ipoteza clasică cu privire la eteroschedasticitate este violată, atunci nu poate fi dovedită existența varianței minime din teorema Gauss-Marcov. Eteroschedasticitatea conduce la faptul că M.C.M.M.P. subestimează varianțele (și erorile standard) ale coeficienților.

- Eteroschedasticitatea produce creșterea varianțelor $\hat{\beta}_s$ într-un mod care nu este perceput de estimațiile M.C.M.M.P., deci această metodă aproape de fiecare dată subestimează aceste varianțe. Prin urmare, nici t - statisticile, nici F - statistica nu pot fi de încredere în fața eteroschedasticității incorecte. Rezultă că M.C.M.M.P. se soldează cu t - valori majorate, care pot fi obținute dacă termenul de eroare a fost eteroschedastic, în unele cazuri provocând cercetătorii să respingă ipoteza nulă atunci când ea n-ar trebui să fie respinsă.

Eteroschedasticitatea cauzează un segment specific de consecințe deoarece Z și varianța distribuției termenului de eroare crește, în așa mod că se mărește probabilitatea apariției observațiilor ale termenului de eroare majore (după valoarea absolută). Dacă din întâmplare segmentul acestor observații este pozitiv, dacă una din variabilele independente este suficient mai mare decât media, estimațiile $\hat{\beta}_s$, obținute cu M.C.M.M.P. pentru aceasta variabilă vor tinde spre a fi mai mari în comparație cu valoarea adevărată. Pe de altă parte, dacă segmentul acestor valori mari ale observațiilor termenului de eroare se întâmplă a fi negativ când una din variabilele X_s este suficient mai mică decât media, atunci estimațiile $\hat{\beta}_s$ obținute prin M.C.M.M.P. pentru această variabilă au tendința de a fi mai mici decât acelea adevărate. Deoarece termenul de eroare totuși se presupune a fi independent de toate variabilele explicative, supraestimările sunt tot atât de probabile ca și subestimările iar estimatorul M.C.M.M.P. în prezența eteroschedasticității rămâne nedeplasat. Oricum, eteroschedasticitatea contribuie la faptul că $\hat{\beta}_s$ se îndepărtează de la valorile adevărate, deci varianța distribuției $\hat{\beta}_s$ crește.

10.3. Testarea eteroschedasticității

Nu toți econometricienii utilizează aceleași teste pentru depistarea eteroschedasticității, deoarece eteroschedasticitatea are diferite forme și manifestarea ei exactă în ecuația examinată aproape de fiecare dată este necunoscută. Abordarea problemei cu ajutorul “factorul de proporționalitate Z_i ” este numai una din multe specificații ai formelor de eteroschedasticitate. Prin urmare, nu există o înțelegere universală asupra metodei de testare a eteroschedasticității; manualele de econometrie înscriu mai mult de opt metode diferite pentru o atare testare.

Vom prezenta patru teste diferite pentru depistarea de eteroschedasticitate. Primul va considera testul Park.

10.3.1 Testul Park

Fie $VAR(\varepsilon_i) = \sigma^2 Z_i^2$, unde ε_i este termenul de eroare în ecuația supusă estimării, σ^2 - varianța termenului de eroare homoschedastic și Z_i este factorul de proporționalitate. Testul Park este un procedeu formal care încearcă să testeze reziduurile în vederea eteroschedasticității în același mod în care d - statistica *Durbin-Watson* testează reziduurile în vederea corelării în serie. Testul Park conține trei etape principale. La prima etapă ecuația de regresie este estimată cu ajutorul M.C.M.M.P. și sunt calculate reziduurile. La etapa a doua se efectuează logaritmarea pătratelor rezidualelor, care reprezintă o variabilă dependentă într-o nouă ecuație de regresie cu

unica variabilă explicativă - factorul de proporționalitate Z . În sfârșit, la a treia etapă, rezultatele obținute de la lansarea regresiei adiționale sunt testate în vederea existenței ai eteroschedasticității.

Oricum, nu e necesar să se lanseze testul Park pentru fiecare ecuație estimată. Înainte de a folosi testul Park, este o idee bună de a verifica următoarele probleme.

- Există oare erori de specificare evidente? Dacă ecuația estimată este suspectată în vederea variabilelor omise sau a fost relansată din cauza specificării, testul Park va fi amânat până când specificația este pe cât e posibil bună.
- Este afectat de eteroschedasticitate subiectul cercetărilor deseori? Nu numai studiul modelelor de regresie cu date încrucișate este ceea mai potrivită sursă a eteroschedasticității (de exemplu, este mai mult suspect decât altele modelul cu varianțe majore în valorile variabilei dependente).
- În sfârșit, prezentarea grafică a reziduurilor demonstrează careva dovezi în favoarea eteroschedasticității? Uneori putem economisi mult timp construind graficul reziduurilor în funcție de factorul potențial de proporționalitate Z . Graficul deseori ne demonstrează este probabilă sau nu eteroschedasticitatea fără a fi aplicat testul Park.

Atunci când există unele motive pentru suspectarea eteroschedasticității cel mai potrivit este să se lanseze testul Park. Deoarece testul Park nu se lansează automat de pachetele computerizate de regresie, este necesar a cunoaște cum se lansează testul desinestatător:

Etapă I: La prima etapă estimăm ecuația cu M.C.M.M.P. și apoi obținem rezidualele din estimării: $u_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$.

Etapă II: Folosind reziduurile calculate pentru a alcătui variabila dependentă pentru regresia adițională. Testul Park în special sugerează să lanșați următoarea ecuație de regresie în log compleți: $\ln(u_i^2) = \alpha_0 + \alpha_1 \ln(Z_i) + v_i$, unde u_i sunt reziduurile din prima ecuație; Z_i este cea mai bună selecție pentru posibilul factor de proporționalitate; v_i este termenul de eroare clasic (omoschedastic).

Etapă III: Cu ajutorul t -testului se testează semnificația coeficientului de pe lângă $\ln(Z)$ din ecuația adițională. Ultima etapă constă în testarea semnificației $\ln(Z)$ cu ajutorul t -statisticii întru explicația $\ln(u^2)$ din ecuație. Dacă coeficientul de pe lângă $\ln(Z)$ este semnificativ diferit de zero, se confirmă existența segmentului eteroschedastic în reziduuri în respect cu Z ; în caz contrar, eteroschedasticitatea relatată la acest Z distinct nu este susținută. Oricum, există posibilitatea confirmării că termenul specific de eroare a ecuației este omoschedastic.

Testul Park nu este ușor de utilizat. Problema majoră constă în identificarea factorului de proporționalitate Z . Deși Z adeseori este o variabilă explicativă în ecuația de regresie de origine, acest fapt nu este garantat de fiecare dată. Un Z particular poate fi ales pentru testul Park numai după investigarea tipului de eteroschedasticitate potențială în ecuație. Un Z bun este o variabilă care e probabil să se schimbe odată cu varianța termenului de eroare.

De exemplu, în modelul încrucișat a țărilor sau regiunilor, un Z bun va fi unul care măsoară volumul observațiilor în raport cu variabila dependentă în examinare. Atunci când este greu de specificat cel mai bun Z pentru o ecuație particulară, este deseori bine venit să distingem Z bun de la Z prost.

11. Remedierea eteroschedasticității

Se vor prezint câteva remedii contra eteroschedasticității, totodată se va atrage atenția că există situații în care problema poate fi lăsată neajustată. De arta econometriei ține învățarea de a distinge o situație de alta.

Primul pas în încercarea de a scăpa ecuația de eteroschedasticitate constă în a încerca de a percepe este perfectă sau imperfectă eteroschedasticitatea. Dacă eteroschedasticitatea se

confirmă a fi imperfectă, atunci se determină variabilele omise care cauzează eteroschedasticitatea imperfectă ca apoi ele să fie incluse în ecuație. Dacă eteroschedasticitatea este perfectă, se consideră două remedii generalizate.

1. Utilizarea M.C.M.M.P. ponderate.

Dacă eteroschedasticitatea este perfectă se va considera M.C.M.M.P. ponderată (o formă generalizată a M.C.M.M.P.). Împărțind toți termenii ecuației la factorul de proporționalitate Z (sau la o funcție de Z), care pare a fi relatată la eteroschedasticitate. După împărțire, se reestimează ecuația cu variabilă dependentă ajustată și variabilele independente ajustate.

2. Redefinirea variabilelor.

Efectul eteroschedasticității reziduurilor deseori poate fi eliminat prin redefinirea variabilelor. Aceasta este o abordare directă pentru corectarea eteroschedasticității în timp ce abordarea metodei ponderate este indirectă. Redefinirea variabilelor se va baza pe teoria respectivă și recentrarea ecuației la comportamentul de bază care necesită a fi explicat.

Deci, primul lucru care necesită a fi făcut, dacă testul Park indică posibilitatea apariției eteroschedasticității constă în examinarea minuțioasă a ecuației în vederea erorilor de specificare. Deși nu se va include niciodată variabila explicativă dintr-un simplu motiv că testul Park indică posibilitatea eteroschedasticității, se cuvine o meditare riguroasă prin specificările ecuației. Dacă reeditarea vă permite să descoperiți o variabilă care va trebui să fie în ecuația de regresie de la bun început, atunci aceasta variabilă necesită a fi introdusă în ecuație. Oricum, dacă nu există erori de specificare evidente, atunci eteroschedasticitatea probabil că este firească, și unul din remedii poate fi aplicat.

11.1. *Metoda celor mai mici pătrate ponderată*

Să examinăm o ecuație cu eteroschedasticitate perfectă cauzată de factorul de proporționalitate Z :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad (11.1)$$

unde varianța termenului de eroare în loc să fie constantă este de felul:

$$\text{VAR}(\varepsilon_i) = \sigma_i^2 = \sigma^2 Z_i^2, \quad (11.2)$$

unde Z_i este factorul de proporționalitate, σ^2 este varianța constantă a erorii clasice (omoschedastice) ε_i . Dat fiind că eteroschedasticitatea perfectă există, ecuația (11.1) este următoarea:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + Z_i u_i. \quad (11.3)$$

Termenul de eroare în ecuația (11.3) $Z_i u_i$ este eteroschedastic deoarece $\sigma^2 Z_i^2$ este varianța care nu e constantă. Cum putem ajusta ecuația (11.3) ca ea să devină omoschedastică? Cea mai ușoară metodă constă în împărțirea ecuației în întregime la factorul de proporționalitate Z_i , prin urmare obținând termenul de eroare u_i care are o varianță constantă σ^2 . Ecuația obținută satisface ipotezelor clasice, și lansarea regresiei pentru ecuația nouă mai mult nu va fi suspectată în vederea prezenței termenului de eroare eteroschedastic. Acest remediu generalizat contra eteroschedasticității este numit M.C.M.M.P. ponderată, care de fapt este o versiunea M.C.M.M.P.

M.C.M.M.P. ponderată implică împărțirea ecuației examinate în întregime la oricare variabilă care ar transforma termenul de eroare în unul omoschedastic și apoi relansarea regresiei cu variabilele transformate. Dată fiind forma generală de eteroschedasticitate (11.2), procedeul constă din trei etape:

$$1. \text{ Împărțim ecuația (11.3) la factorul de proporționalitate } Z_i \text{ și obținem:} \\ Y_i / Z_i = \beta_0 / Z_i + \beta_1 X_{1i} / Z_i + \beta_2 X_{2i} / Z_i + u_i \quad (11.4)$$

termenul de eroare u_i în ecuația (11.4) este omoschedastic.

2. Recalculăm datele pentru variabile conform ecuației (11.4).

3. Estimăm ecuația (11.4) cu M.C.M.M.P.

La aplicarea M.C.M.M.P. ponderată se obțin estimările ecuației transformate, care pot fi complet înșelătoare, deoarece detaliile exacte cu privire la complectarea acestei regresii depind de faptul dacă factorul de proporționalitate Z_i este și el variabila explicativă în ecuația (11.1). Dacă Z_i nu este variabilă explicativă în ecuația (11.1), atunci regresia lansată la etapa a 3-a va fi următoarea:

$$Y_i / Z_i = \beta_0 / Z_i + \beta_1 X_{1i} / Z_i + \beta_2 X_i / Z_i + u_i. \quad (11.5)$$

De notat că această ecuație nu are termenul liber. Însă, după cum a fost menționat anterior, omiterea termenului constant scoate în evidență efectul constant al variabilei omise, neliniaritatea și erorile de măsurare asupra altor coeficienți estimați. Pentru a evita situația în care elementul constant forțează schimbările în estimările coeficientului unghiular, o abordare alternativă constă în adăugarea termenului constant în ecuația (11.5) înainte ca ecuația să fie estimată. Deci, când Z_i nu este identic cu una din variabilele X_s în ecuația inițială, atunci este bine venit ca următoarea specificație să fie lansată la etapa a 3 cu M.C.M.M.P. ponderată:

$$Y_i / Z_i = \alpha_0 + \beta_0 / Z_i + \beta_1 X_{1i} / Z_i + \beta_2 X_i / Z_i + u_i. \quad (11.6)$$

Dacă Z este o variabilă explicativă în ecuația (11.6), atunci nu este nevoie ca termenul constant să fie adăugat în ecuație, deoarece unu deja există. Să revenim la ecuația (10.12). Dacă $Z = X_1$ (sau dacă $Z = X_2$), atunci unul din coeficienții unghiulari devine termen constant în ecuația transformată deoarece $X_1 / Z = 1$. $Y_i / Z_i = \beta_0 / Z_i + \beta + \beta_2 X_i / Z_i + u_i$ (11.7)

În cazul în care este folosită aceasta formă a M.C.M.M.P. ponderate, oricum, coeficienții obținuți în urma estimării ecuației (11.6) necesită a fi interpretați foarte grijuliu. Vom nota că $\hat{\beta}_1$ acum este termenul de intersecție a ecuației (11.6) chiar dacă el este coeficient unghiular în ecuația (11.1). Prin urmare, dacă suntem cointeresați în estimarea coeficientului de pe lângă variabila X_1 în ecuația (11.1), va trebui să examinăm termenul de intersecție în ecuația (11.6). Calculatorul va afișa $\hat{\beta}_0$ ca «coeficient unghiular» și $\hat{\beta}_1$ ca “termen constant” atunci când în realitate sunt estimați coeficienții opuși în ecuația (11.1).

Există trei probleme majore în utilizarea M.C.M.M.P. ponderate:

1. Sarcina de identificare a factorului de proporționalitate este, cum a fost accentuat, foarte dificilă.
2. Forma funcțională care relatează factorul Z_i la varianța termenului de eroare a ecuației inițiale în general poate să nu fie funcție pătrată în ecuația (11.2). Atunci când sunt aplicate alte relații funcționale, sunt necesare alte transformări.
3. Uneori M.C.M.M.P. ponderată se aplică la ecuația cu eteroschedasticitate imperfectă. În așa cazuri, poate fi demonstrat că estimările suportă o mică reducere în deplasări atunci când este omisă o variabilă, și estimările sunt inferioare celor care sunt obținute din ecuația corect specificată.

11.1.2. O abordare directă: Redefinirea variabilelor

O altă abordare întru eliminarea eteroschedasticității din ecuație constă în reintoarcerea la teoria de bază corespunzătoare ecuației și redefinirea variabilelor în așa mod ca eteroschedasticitatea să fie evitată. Redefinirea variabilelor deseori este utilă pentru că permite ecuației estimate să se concentreze asupra aspectului comportamental a relației. O atare remediere este un proces dificil și descurajător deoarece apelarea la el ignoră toată munca anterioară. Oricum, odată ce partea teoretică a fost revăzută, abordările de alternativă descoperite posibilele căi pentru evitarea problemelor care la început păreau de nedepășit.

Spre nefericire, este dificil să se specifice procedee pentru o situație mai generală decât “remediarea completă a proiectului de investigare”, se va prezenta un exemplu non numeric privind cele relatate. Abordarea directă de redefinire a variabilelor va fi comparată cu metoda

mult mai formalizată – M.C.M.M.P. ponderată. Să examinăm modelul cu date încrucișate a cheltuielilor totale a guvernelor din diferite orașe. Din punct de vedere logic variabilele pentru examinare în o atare analiză sunt: 1) venitul agregat; 2) populația și 3) salariul mediu în fiecare oraș. Cu cât mai mare va fi venitul rezidenților orașului și a businessului, cu atât mai mari vor fi cheltuielile guvernului orășenesc. În acest caz este foarte clar că orașele mari au venituri mari și respectiv cheltuieli mari (în valoare absolută) decât orașele mici.

Aproximarea acestei funcții cu linia de regresie la fel ne oferă o pondere exagerată pentru orașele mari de oarece în caz contrar ele vor contribui la valori mari a pătratelor reziduurilor. Aceasta este așa deoarece M.C.M.M.P. minimizează suma pătratelor reziduurilor, și deoarece reziduurile pentru orașele mari este posibil să fie mai mari pur și simplu din cauza mărimii orașului, estimațiile regresionale vor fi în special sensibile la reziduurile orașelor mari. Acest fenomen deseori este numit “corelare falsă” datorită mărimii.

În plus, reziduurile pot indica eteroschedasticitatea. Remediu pentru un atare fel de eteroschedasticitate nu constă în utilizarea automată a M.C.M.M.P. ponderate și nici în aruncarea observațiilor referitor la orașele mari. Are sens să se considere reformularea modelului pe o cale care va face reduceri la scara factorului (mărimia orașului) și va accentua comportamentul corespunzător. În acest caz, cheltuielile pe cap de locuitor, va fi o variabilă explicativă logică. O atare transformare este prezentată în figura ce va urma. Forma ecuației transformate plasează New York și Los Angeles pe aceeași scară ca și Pasadena sau New Brunswick și astfel le oferă lor aceeași pondere în estimare. Dacă variabila explicativă nu să va produce a fi funcție de mărimia orașului, în orice caz, ea nu va necesita să fie ajustată la un cap de locuitor. Dacă ecuația include salariul mediu a lucrătorilor din oraș, de exemplu, el nu va fi împărțit la populație în ecuația transformată.

Vom nota, că transformarea în careva sens este similară M.C.M.M.P. ponderate. Diferența constă în aceea că nu există un termen în ecuație reciproc la populație (cum este în M.C.M.M.P. ponderată) și nu toate variabilele explicative se împart la populație. Din ecuația inițială,

$$EXP_i = \beta_0 + \beta_1 POP_i + \beta_2 INC_i + \beta_3 WAGE_i + \varepsilon_i \quad (11.8)$$

versiune celor mai mici pătrate ponderată va fi

$$EXP_i / POP_i = \beta_1 + \beta_0 / POP_i + \beta_2 INC_i / POP_i + \beta_3 WAGE_i / POP_i + u_i \quad (11.9)$$

atunci când ecuația direct transformată va lua forma:

$$EXP_i / POP_i = \alpha_0 + \alpha_1 INC_i / POP_i + \alpha_2 WAGE_i + u_i \quad (11.10)$$

Cum putem observa, M.C.M.M.P. ponderată (11.8) împărțită în întregime la populație, atunci când în aceea, transformată conform teoriei, sunt împărțite la populație numai variabilele de cheltuieli și venit. În timp ce ecuația (11.9) direct transformată întradevăr soluționează eteroschedasticitatea potențială din model, așa o soluție va fi considerată întâmplătoare întru beneficiul remedierii ecuației pe o cale care se concentrează pe faptul examinării comportamentului de bază.

De notat, ca este posibil ca ecuația (11.9) reformulată să posede eteroschedasticitate; varianța termenului de eroare poate fi mai mare pentru observațiile care au valorile mai mari pe cap de locuitor cum ar fi cheltuieli și venituri decât pentru acelea observării care au valori mai mici pe cap de locuitor a cheltuielilor și veniturilor. Deci este legitimată suspectarea și testarea eteroschedasticității chiar și în aceasta ecuație transformată. O atare eteroschedasticitate în ecuația transformată nu este verosimila totuși, deoarece va fi o mică variație în mărimile normal asociate cu eteroschedasticitatea.

O transformare grijulie a variabilelor întru corectarea eteroschedasticității în timp ce totodată nu evita, corelarea falsă, datorată mărimii, poate uneori să fie o abordare reușită în soluționarea acestor probleme. De notat totuși, că nu fiecare variabilă în ecuație este tratată în același mod (spre deosebire de M.C.M.M.P. ponderată). Fiecare variabilă în modelul cu date încrucișate poate fi examinați în vederea transformărilor posibile care se vor solda cu interpretări semnificative și complete a ecuației de regresie.

12. Specificația: alegerea variabilelor independente relevante

12.1. Variabilele omise

Înainte de a estima ecuația de regresie ea necesită a fi specificată complet. Specificarea ecuației econometrice constă din trei etape: 1) alegerea corectă a variabilelor independente, 2) alegerea formei funcționale corecte, 3) alegerea formei corecte a termenului stocastic de eroare. Specificarea erorii rezultă din efectuarea incorectă a unei din etapele menționate.

Ne vom opri la prima etapă – alegerea variabilelor independente. Cercetătorul hotărăște care variabile independente vor fi incluse în ecuație și aceasta reprezintă atât momentul slab cât și momentul forte n economie. Momentul forte îl constituie faptul că ecuația fiind formulată poate fi folosită pentru previziunea obiectivelor necesare individuale, iar momentul slab se referă la posibilitatea estimării a mai multor specificații până când va fi găsită una care susține poziția înaintată în defavoarea altor indicatori care o dezaprobă. Ulterior sarcina principală va ține de demonstrarea modului de selecție al variabilelor pentru ecuația de regresie fără a comite erori ce rezultă din alegerea nereușită.

Primul considerent în deciderea apartenenței variabilei ecuației constă în fundamentarea ei în baza teoriei economice corespunzătoare. Dacă răspunsul este de forma unui “da” ambiguu, atunci variabila va fi introdusă în ecuație definitiv atunci când va avea o semnificație statistică. Ne includerea variabilei relevante în ecuație conduce la majorarea deplasărilor estimațiilor rămase, însă includerea unei variabile ne importante contribuie la majorarea varianței coeficienților estimați.

Fie că la prima specificație a ecuației examinate din diverse motive nu s-a inclus o variabilă independentă importantă. Sau, admitem că nu au fost posibil de găsit datele necesare (sau datele colectate s-au dovedit a fi incomplete) pentru una dintre variabilele exogene relevante. În ambele situații rezultatul este același: variabila defenită ca o variabilă de explicat importantă a fost omisă, a rămas în afara ecuației de regresie. De fiecare dată când există o variabilă omisă, interpretarea și utilizarea ecuației estimate devine suspectă, cum ar fi prețul în ecuația pentru cerereceea ce nu numai a împiedica estimarea coeficientului de regresie pentru preț dar, cu siguranță, a cauza deplasări în estimațiile coeficienților pe lângă variabilele din ecuație. Deplasarea cauzată de eliminarea variabilei relevante din ecuație se numește deplasare de specificație (mai rar, deplasarea variabilei omise). Într-o ecuație cu mai multe variabile, coeficientul $\hat{\beta}_k$ reprezintă schimbarea în variabila dependentă Y , cauzată de schimbarea cu o unitate în variabila independentă X_k , dat fiind restul variabilelor din ecuație neschimbate. Omiterea variabilei cauzează deplasări, ea poate schimba valoarea așteptată a coeficienților estimați de la valoarea adevărată a coeficienților.

12.1.2. Urmările omiterii variabilelor

Fie că modelul de regresie adevărat este următorul:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad (12.1)$$

unde ε_i - este termenul de eroare clasic.

Dacă a fost omisă o variabilă independentă importantă, ecuația devine:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i^*, \quad \varepsilon_i^* = X_{2i} + \varepsilon_i. \quad (12.2)$$

Termenul de eroare nu este independent de variabila de explicat X_{1i} atâta timp, cât variabilele X_{1i} și X_{2i} sunt corelate întrucât odată cu schimbarea variabilei X_{2i} se schimbă și variabila X_{1i} și ε_i^* , (cu alte cuvinte, nu sunt respectate ipotezele clasice referitor la independența variabilelor de explicat față de termenul de eroare), deci până când variabila

omisă nu este corelată cu nici una dintre variabilele incluse în ecuație (ceea ce este aproape nevirosimil).

În general, atunci când unele din ipotezele clasice nu sunt respectate, nu are loc teorema Gauss-Marcov, și estimațiile nu sunt BLUE). Ceea ce înseamnă că estimatorii liniari nu mai sunt nedepasați și de varianță minimă (aceeași varianță pentru toți estimatorii liniari nedepasați) sau nu se îndeplinesc concomitent ambele ipoteze. Estimarea ecuației (12.2) în timp ce atunci este adevărată ecuația (12.1), cauzează deplasări în estimațiile ecuației (12.2). Ceea ce înseamnă că: $E(\hat{\beta}_1) \neq \beta_1$. (12.3)

Deci, lipsa variabilei X_2 din ecuație conduce la deplasarea valorii așteptate a coeficientului $\hat{\beta}_1$ de la valoarea sa adevărată. Dacă variabilele X_1 și X_2 sunt corelate și variabila X_2 este omisă din ecuație, atunci M.C.M.M.P. va atribui variabilei X_1 o varianță eventual cauzată de X_2 , ce se soldează cu estimări deplasate pentru $\hat{\beta}_1$.

Pentru a demonstra faptul că variabila omisă poate cauza estimări deplasate, vom examina funcția de producere care exprimă venitul (Y) ca funcție dependentă de la cantitatea utilizată de muncă (X_1) și de capital (X_2). Fie că din careva considerente dacă lipsesc datele pentru capital și variabila (X_2) va fi omisă din model. Aceasta eliminare, cu certitudine, va deplasa estimatorul coeficientului de pe lângă variabila muncă deoarece este evident că munca și capitalul sunt corelați (creșterea în capital, de regulă, necesită cel puțin câteva brațe de muncă spre utilizare și vice versa). Prin urmare, M.C.M.M.P. va atribui muncii o creștere a volumului de producție de fapt cauzată de capital. Deci deplasarea va fi o funcție de β_2 și coeficientul de corelație dintre capital și muncă

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 f(r_{12}) \quad (12.4)$$

Rezultă că valoarea așteptată a coeficientului de pe lângă variabila inclusă ($\hat{\beta}_1$), atunci când variabila importantă (X_2) este omisă, este egal cu valoarea sa adevărată plus coeficientul adevărat de pe lângă variabila exclusă înmulțit cu o funcție dependentă de coeficientul de corelație simplă dintre variabila inclusă și aceea exclusă:

a) valoarea adevărată β_2 este zero (deci deplasarea nu există, ceea ce înseamnă că variabila X_2 nu este importantă în model); sau b) r_{12} este zero (variabilele X_1 și X_2 sunt perfect incorelate).

Valoarea termenului $\beta_2 f(r_{12})$ determină cantitatea deplasării de specificație introdusă în estimația β_1 prin eliminarea X_2 . Dacă variabila inclusă și aceea exclusă sunt necorelate, nu există deplasări, însă în realitate, aproape de fiecare dată, careva corelație (fie chiar aleatoare) dintre oricare două variabile există și atunci deplasările sunt cauzate frecvent de omiterea variabilei importante.

12.2.1. Un exemplu deplasărilor de specificație.

Fie că $\hat{Y}_t = -0,605 - 0,45PC_t + 0,12PB_t + 12,2 \ln YD_t$

$$\begin{array}{ccc} & (0,07) & (0,05) & (11,2) \\ t = & -6,4 & -2,5 & 10,6 \\ \bar{R}^2 = 0,984; & n = 35 & \text{(anuale 1950-1984)} \end{array}$$

Y_t – consumul de pasare;

PC_t – costul unui kg de carne de pasare;

PB_t – costul unui kg de carne de vita ;

$\ln YD_t$ – venitul disponibil pe cap de locuitor (logaritmnatural).

Dacă estimăm aceasta ecuație fără prețul de substituie, obținem:

$$\begin{array}{ccc} \hat{Y}_t = -80,7 - 0,34PC_t + 15,0 \ln YD_t \\ & (0,06) & (0,42) \end{array}$$

$$t = \quad -5,6 \quad 36,0$$

$$\bar{R}^2 = 0,981; \quad n = 35$$

12.3. Corectarea variabilelor omise

Teoretic soluționarea problemei deplasărilor de specificație se reduce la introducerea variabilei omise în ecuație. Spre regret, aceasta este mult mai ușor de pronunțat decât de executat.

În primul rând, deplasarea variabilelor omise este greu de depistat. Atunci când unii indici ai deplasării de specificație se exprimă explicit (cum ar fi, semnul coeficientului estimat este opus celui care se presupune a fi), alții nu sunt clari. Cel mai reușit indiciu de relevanță a variabilei omise constă în fundamentarea teoretică corectă a modelului. Ce variabile necesită a fi incluse? Ce semne se așteaptă? Se cunosc unele informații referitor intervalele în care se vor afla coeficienții? Posibil că accidental a fost eliminată o variabilă care, cu certitudine, se consideră importantă sub aspect teoretic. Ceea mai bună cale de a evita omiterea variabilelor importante constă examinarea cu precauție a ecuației înainte de a introduce datele în calculator.

A doua sursă de complexitate constă în problema alegerii variabilei cu care va fi suplimentată ecuația odată ce s-au depistat deplasări urmate de variabilele omise. Simultan în ecuație pot fi adăugate toate variabilele importante posibile ce conduce la pierderea preciziei de estimare, sau pot fi testate mai multe variabile și menținută aceea cu statistici mai bune în vederea reducerii deplasărilor. Însă tehnica majorării numărului de variabile cu scopul selectării celor mai reușite rezultate de regresie nu este una reușită deoarece variabila care în cel mai bun mod corectează deplasările de specificație poate contribui în mai mare măsură la schimbare decât la obținerea soluției adevărate a problemei. În așa circumstanțe o ecuație stabilită poate oferi rezultate statistice superbe pentru un set de date în examinare, dar aceste rezultate ele devin teribile când sunt aplicate la alte seturi de date, deoarece ele nu descriu caracteristicile populației adevărate.

Includerea unei variabile adiționale în ecuație nu asigură cu siguranță lichidarea deplasărilor variabilelor omise. Fie că semnul variabilei omise diferă de cel așteptat, atunci el nu poate fi schimbat în direcția dorită prin aruncarea variabilei care are valoarea testului t a coeficientului estimat mai mică (după valoare absolută) decât t - valoarea a coeficientului estimat cu un semn nedorit. Mai mult decât atât, semnul în genere nu poate fi schimbat chiar dacă variabila care va fi eliminată are o valoare de t foarte mare.

Dacă coeficientul estimat semnificativ diferă de la așteptările noastre (atât după semn cât și după amplitudă), atunci, la sigur, în model există unele deplasări de specificație. Deși este adevărat că un set de date prost sau o teorie presupusă slabă pot la fel să ofere semne sau amplitudă semnificative, însă aceste evenimente pot fi uneori eliminate.

O tehnică corectă pentru reducerea numărului variabilelor omise constă în examinarea direcției deplasărilor cauzate de omiterea variabilei din ecuație. Dacă va fi demonstrat că semnul deplasării așteptate este în direcția opusă în raport cu aceea observată, atunci variabila poate fi eliminată din ecuație. Direcția deplasării așteptate poate fi determinată din:

$$E(\hat{\beta}_1) = \beta_2 f(r_{12}), \quad E(\hat{\beta}_{PB}) = \beta_{PB} f(r_{PC,PB}) = (+) \cdot (+) = (+);$$

$$E(\hat{\beta}_{PC}) = \beta_{PD} \cdot f(r_{PC,PD}) = (-)(+) = (-).$$

Tehnic dată va funcționa bine atunci, și numai atunci, când numai o singură variabilă este omisă din ecuație. În cazul omiterii concomitente a mai multor variabile, impactul asupra coeficienților din ecuație este greu de specificat.

12.4. Variabile irelevante (neimportante)

Ce se va întâmpla dacă în ecuație se introduce o variabilă care nu-i aparține? Acest caz, variabilelor neimportante, este unul invers la acel al variabilelor omise, și poate fi analizat

folosind modelul elaborat în paragraful precedent. Acest model cu variabile neimportante conține mai multe variabile în ecuația de estimat decât în ecuația adevărată.

Includerea variabilei în ecuația, căreia ea nu-i aparține, nu cauzează deplasări, dar contribuie la creșterea varianței coeficienților estimați incluși.

12.4.1. Impactul variabilelor neimportante

Regresia adevărată specificată este:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad (12.5),$$

dar cercetătorul din careva motive a inclus o variabilă în plus:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i^{**}, \quad \varepsilon_i^{**} = \varepsilon_i - \beta_2 X_{2i} \quad (12.6).$$

O atare greșală nu va cauza deplasări, dacă coeficientul adevărat al variabilei neimportante este 0. În acest caz, $\varepsilon_i^{**} = \varepsilon_i$ și $\hat{\beta}_i$ este nedeplasat în (11.8), atunci când $\beta_2 = 0$.

Includerea variabilei neimportante va majora varianța coeficienților estimați, prin urmare ea va tinde să diminueze magnitudinea absolută a t -testelor. La fel variabilele neimportante vor contribui la diminuarea $\bar{R}^2$, (dar nu R^2). În modelul cu Y și X_1, X_2 , varianța estimatorului β_1 obținută prin M.C.M.M.P. este:

$$VAR(\hat{\beta}_1) = \frac{\hat{\sigma}_u^2}{(1 - r_{12}^2)} \sum (X_{1i} - \bar{X}_1)^2, \text{ dacă } r_{12} \neq 0$$

iar atunci când $r_{12} = 0$ ea este:

$$VAR(\hat{\beta}_1) = \hat{\sigma}_u^2 \sum (X_{1i} - \bar{X}_1)^2.$$

Deci, chiar dacă variabilele neimportante nu cauzează deplasări, ele cauzează probleme pentru regresie, deoarece reduc precizia regresiei.

Tabel. Sumar al impactului variabilelor omise sau a variabilelor neincluse (neimportante) asupra restului coeficienților

Efectul asupra coef. estimați	Variabila omisă	Variabila introdusă neimportantă
Deplasări	Da *	Nu
Creșterea sau descreșterea varianței	Descrește *	Crește *

12.4.2. Efectuarea alegerii corecte de specificație

Vor fi examinate 4 criterii de validare cu privire la luarea deciziilor în vederea apartenenței variabilei ecuației.

- 1) Teoria: variabila este introdusă în ecuație fără ambiguități și teoretic fundamentat?
- 2) t -testul: coeficientul estimat diferă semnificativ de "0"?
- 3) $\bar{R}^2$: se îmbunătățește aproximarea ecuației (ajustată la gradele de libertate) atunci când o variabilă se adaugă la ecuație?
- 4) Deplasări: se schimbă semnificativ alți coeficienți atunci când o variabilă este inclusă în ecuație?

Dacă toate criteriile enunțate sunt confirmate, variabila aparține ecuației; dacă nici unul din ele nu este just, variabila este neimportantă și poate fi, cu siguranță, exclusă din ecuație. Atunci când o variabilă importantă este inclusă în ecuație, includerea ei e mult probabil să contribuie la creșterea $\bar{R}^2$, la schimbarea altor coeficienți, menținând valoarea t -testului semnificativă. Pe de altă parte, dacă o variabilă neimportantă este introdusă în ecuație, ea va reduce $\bar{R}^2$, având o

valoare a t -testului nesemnificativă, și, prin urmare, un impact mic asupra coeficienților de pe lângă restul variabilelor.

Deseori cele patru criterii nu se acordă. Aceasta se întâmplă atunci când sau variabila are un t -test nesemnificativ, sau variabila, fiind comparativ necorelată cu variabilele prezente în ecuație, are un efect mic asupra coeficienților estimați. Cum de procedat în asemenea circumstanțe?

Singura, și cea mai importantă, justificare în determinarea importanței variabilei este fundamentarea teoretică. Nu cantitatea evidenței statistice servește drept dovadă în favoarea “neimportantei” variabilei, motivația teoretică confirmă această necesitate. Uneori, în lipsa unei alternative mai bune, o variabilă importantă din punctul de vedere teoretic rămâne în afara ecuației, în așa cazuri este limitată utilitatea ecuației.

12.4.3. Căutări de specificație

Una din pozițiile slabe ale econometriei este posibilitatea manipulării exagerate cu seriile de date pentru a obține diferite rezultate, specificând diferite ecuații de regresie, până când nu se obțin estimații cu proprietățile căutate.

Deși problema nu este deloc ușoară, are sens de a face o încercare spre minimizarea numărului de ecuații estimate și a se baza mai mult pe teorie decât pe aproximarea statistică frecventă la alegerea variabilelor. Se va demonstra aceasta prin examinarea a trei, cel mai uzuale, tehnici incorecte pentru specificarea ecuației de regresie.

12.4.4. Explorarea datelor cu scopul maximizării R^2

Cu siguranță, cea mai proastă cale de specificare este încercarea de a formula simultan o serie de regresii și de a alege ecuația care în cel mai reușit mod corespunde rezultatelor pe care se dorește de a obține. În așa situație, se încearcă estimarea practic a tuturor combinațiilor posibile ale variabilelor independente de alternativă și selectarea lor se va efectua în baza rezultatelor obținute.

Practica dată de estimare simultană a unui număr de combinații de variabile independente și selectarea celei mai bune dintre ele nu ține cont de numărul specificațiilor examinate de la prima până la ultima.

Dacă să simplificăm, în cazul când cele 95% de încredere, obținute în urma regresiiilor consecutive, nu sunt ocazionale și au fost lansate mai mult decât 20 de regresii, cât de multă încredere putem avea în rezultatele obținute? Între timp, menținând regresia cu testul t înalt și ignorând aceea cu testul t mic, obținem un test exagerat pentru estimarea semnificației coeficienților.

Mai mult decât atât, așa “explorare de date” și “pescuirea grabei” în obținerea statisticilor necesare pentru ecuația de regresie finală este o metodă în esență lipsită de etica cercetătorilor empirice. Acest procedeu include nu numai combinații de alternativă a variabilelor independente, dar și un număr mare de forme funcționale, structuri de lag, tot ce ne oferă o tehnică “avansată” de estimare, și atunci crește extraordinar șansa de a obține rezultate necesare, însă rezultatul final nu va fi unul de valoare. Cercetătorul nu caută dovezi științifice de a susține ipotezele inițiale; din contra, așteptările anterioare sunt impuse datelor într-un mod care este în esență greșit.

12.4.5. Procedee regresionale iterative

Regresia iterativă implică utilizarea produselor program pentru alegerea variabilei independente folosite la estimarea ecuației specificate. Programul de calculator oferă o lista de variabile independente în baza lor fiind apoi după etape construiește ecuația. În primul rând se alege variabila de explicat, care de una singură explică o mare parte din varianța de la valoarea medie a variabilei dependente. În calitate de a doua variabilă se alege aceia, care cel mai mult

contribuie la majorarea R^2 , ținând cont de faptul că prima variabilă deja este introdusă în ecuație. Procedul iterativ continue până când variabila ce va urma a fi introdusă în ecuație nu izbuteste să atingă careva creștere în R^2 . Contribuția fiecărei variabile independente presupune o creștere în R^2 cauzată de includerea ei în ecuația de regresie.

Spre regret, orice corelare între variabilele independente face această procedură dificilă. În procesul de evaluare a corelației între variabile este greu de separat impactul unei variabile de la impactul altei variabile. Ca rezultat, în prezența multicolinearității este imposibil de determinat contribuția individuală a fiecărei variabile suficientă pentru a afirma că una din ele este mai importantă și deci necesită a fi inclusă în primul rând, și mai rău, nu este o justificare teoretică pentru a alege combinația variabilelor specificate.

Din cauza acestor probleme, deseori se evită procedul iterativ. Primejdia cea mai mare este că coeficienții obținuți pot fi deplasați, valorile “ t ”- calculate nu urmează pe viitor repartiția valorilor “ t ”- tabulare, variabilele importante pot fi excluse din cauza aranjamentului (ordinea în care a avut loc selecția), semnele coeficienților estimați la fazele intermediare sau finale a procedurii pot să difere de la semnele preconizate. Utilizarea procedurii iterativ este o anumită ignoranță față de faptul ce variabile vor fi introduse.

12.4.6. Căutările succesive ale specificației

Din orgoliu majoritatea econometricienilor, evita explorarea datelor și metoda iterativă de specificare a variabilelor. În schimb, se preferă specificarea ecuației prin estimarea ecuației inițiale și apoi eliminarea sau adăugarea consecutivă a variabilelor (sau schimbarea formei funcționale) până când o ecuație verosimilă cu “statistici bune” este obținută. Aflându-se în fața situație când cunoașterea sigură (în baza teoriei) a unor variabile relevante, aparent este o practică general acceptată, când necunoscând sunt sau nu sunt relevante variabilele adăugate, se recurge la verificarea R^2 și a “ t ”- testelor pentru toate variabilele (până la și după selecție). Întrădevăr, ușor se demonstrează, că o atare căutare este cautare a “neadevărului”. După cum se va constata, există o diferență enormă în abordarea căutării de specificație succesivă și abordarea care va fi recomandată. Căutarea succesivă a specificației este o tehnică care permite cercetătorul să estimeze un număr secret de regresii apoi să prezinte alegerea finală (bazată pe o mulțime de așteptări nespecificate în vederea semnelor și semnificației coeficienților) ultima fiind prezentată ca o specificație estimată. Așa o metodă stabilește greșit veridicitatea statistică a rezultatelor regresiei din două motive.

- 1) Semnificația statistică a rezultatelor este supraestimată deoarece estimațiile regresiei precedente sunt ignorate.
- 2) Mulțimea motivațiilor utilizată la alegerea dintre rezultatele diferitor regresii este secretă. Nu e posibil, sub nici o formă, să se cunoască: oferă sau nu rezultate regresiiilor efectuate semne opuse sau coeficienți semnificativi pentru variabilele importante.

Spre regret, nu există o metodă universală acceptată pentru a dirija căutările succesive, în primul rând deaceia că testul potrivit la o etapă a procedurii depinde de testele care au fost îndeplinite anterior și deaceia că testele este foarte greu de inventat. O modalitate de îmbunătățire constă în reducerea gradelor de libertate în ecuația „finală” cu unul pentru fiecare încercare alternativă de specificare. Acest procedeu este departe de a fi exact, dar impune o penalitate explicită pentru căutările de specificare.

În general, se recomandă de a menține numărul regresiiilor estimate cât se poate de mic; de a se concentra la considerente teoretice atunci când se selectează variabilele, formele funcționale și de a dezvălui toate specificațiile investigate. Deci, se recomandă de a combina economia (folosirea teoriei și analizei pentru limitarea numărului de specificații estimate) cu dezvăluirea (informarea referitor la toate ecuațiile estimate).

Și totuși, aceasta este o altă față a istoriei. Unii cercetători simt că modelul menținut, dacă va fi o șansă, va demonstra direct și cele mai bune rezultate statistice (inclusiv semnele și

coeficienții) care cel mai bine se potrivesc cu specificațiile adevărate. Problema acestei psihologii este aceea că elementul de șansă este, în mod normal, foarte puternic pentru orice aplicație. Plus la aceasta persoanele rezonabile adesea „nu sunt de acord cu faptul cum arată modelul adevărat”. Prin urmare, diferiți cercetători vor examina aceleași seturi de date și vor „veni” cu modele „mai bune” extrem de diferite. Deoarece aceasta poate să se întâmple, diferența dintre un econometrician bun și prost nu este atât de clară cum se pare. Atât timp cât se manifestă un respect sănătos față de pericolul căutărilor de specificare, e mult probabil să se procedeze într-un mod rezonabil.

Concluziile obținute sunt absolut clare: cel mai important lucru în specificarea ecuației va fi făcut înainte de orice încercare de estimare a ecuației la calculator. Întrucât o perfecțiune nu este rezonabilă, vor fi perioade când specificații adiționale va fi necesar de estimat. Oricum aceste estimații noi necesită a fi temeinic fundamentale teoretic și explicit luate în vedere atunci când se va testa semnificația și se vor totaliza rezultatele. În așa mod va fi redus pericolul estimărilor statistice incorecte.

12.4.7. Impactul căutărilor succesive de specificare

Să prezentăm un exemplu prin care se va demonstra că eliminarea variabilelor din model în baza „t”-testului introduce deplasări sistematice în ecuația estimată. Fie că modelul ipotetic pentru o variabilă independentă distinctă este:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (12.7)$$

Să admitem pe viitor că, în baza teoriei, există certitudine că variabila independentă X_1 aparține ecuației, iar variabila independentă X_2 nu aparține ecuației. Rămîne să se determine în vederea includerii variabilei X_2 în ecuație; mulți cercetători experimentați utilizează numai testul „t”, care indică că coeficientul β_2 este semnificativ diferit de 0, prin urmare, ei păstrează variabila independentă X_2 în ecuație, obținând forma (11.17) ca model final. În caz ce testul „t” nu indică diferența semnificativă de 0, acești cercetătorii exclud variabila X_2 din ecuație și considera Y ca funcție de o singură variabilă X_1 .

Două tipuri de erori pot fi introduse pornind de la o atare abordare. Prima, variabila X_2 poate fi păstrată în ecuație atunci când ea nu aparține ei, însă greșală de acest gen nu va schimba valoarea așteptată a lui β_1 . A doua, variabila X_2 poate fi aruncată din ecuație deși trebuie să aparțină ei și atunci coeficientul estimat pentru X_1 va fi deplasat de valoarea β_2 în măsura în care variabilele X_1 și X_2 sunt corelate. Cu alte cuvinte, coeficientul β_1 va fi deplasat atâta timp cât variabila X_2 , care trebuie să aparțină ecuației, va fi eliminat din ea, și variabila X_2 va fi eliminată de fiecare dată când coeficientul estimat va fi semnificativ diferit de 0. Atunci, valoarea așteptată a lui $\hat{\beta}_1$ nu va fi egală cu valoarea teoretică β_1 , și va avea deplasări sistematice în ecuația examinată: $E(\hat{\beta}_1) = \beta_1 + \beta_2 \cdot f(r_{X_1/X_2}) \neq \beta_1$. Unde P indică probabilitatea semnificației testului „t”. Aceasta este la fel cazul când „t”- testul calculat pentru $\hat{\beta}_1$ nu mai urmează repartiția „t” tabelar. Cu alte cuvinte, „t”- testul calculat este deplasat prin căutarea succesivă de specificație.

Întrucât, majoritatea cercetătorilor consideră un număr de variabile diferite înaintea stabilirii modelului final, cel care are încredere în „t”- testul calculat se confruntă cu această problemă sistematic. Deci, practica eliminării potențialelor variabile independente pur și simplu din motivul că „t”- testul calculat indică că coeficientul estimat nu este semnificativ diferit de 0 va cauza deplasări sistematice în coeficienții estimați (și „t”- testele lor) pe lângă variabilele rămase.

13. Specificația ecuației de regresie: alegerea formei funcționale

13.1 Forme funcționale alternative

Alegerea formei funcționale pentru ecuația de regresie este o parte vitală în specificarea ecuației de regresie. M.C.M.M.P. la utilizarea sa necesită ca ecuația examinată să fie liniară în conformitate, cu coeficienții, dar există o varietate de forme funcționale care sunt liniare în coeficienți în timp ce nu sunt liniare față de variabile. Se vor prezenta în detalii cele mai frecvent utilizate forme funcționale în scopul de a ajuta utilizatorul în dezvoltarea abilității de a alege corect una din ele la specificarea ecuației.

Alegerea formei funcționale, aproape de fiecare dată, se va baza pe teoria economica sau de business fundamentală și numai rareori pe aceea formă funcțională care furnizează o previziune mai bună. Relația logică dintre variabila dependentă și variabilele independente în examinare se va compara cu proprietățile diferitor forme funcționale, și numai aceea, care în modul cel mai reușit oglindește teoria va fi aleasă. În continuare, cele mai des utilizate forme funcționale vor fi caracterizate în termeni grafici, ecuații și exemple pentru a face comparație între ele.

13.2 Forma liniară

Modelul liniar de regresie, se bazează pe ipoteza că coeficienții unghiulari din relația ce caracterizează variabila dependentă și cele independente sunt constanți și are loc relația $\frac{\partial Y}{\partial X_i} \approx \frac{\Delta Y}{\Delta X_i} = \beta_i$, $i = 1, 2, \dots, k$. Dat fiind constant coeficientul unghiular, elasticitatea variabilei Y în respect cu variabila X (schimbarea în procente în variabila dependentă cauzată de schimbarea cu un procent în variabila independentă, restul variabilelor din ecuație rămânând constante) nu este constantă: $E_{Y, X_i} = \frac{\partial Y}{\partial X_i} \frac{X_i}{Y} = \frac{\partial Y/Y}{\partial X_i/X_i} = \beta_i \frac{X_i}{Y}$. Dacă relația, ce se presupune a fi dintre variabila dependentă Y și variabila independentă, X este de așa natură că coeficientul unghiular (de înclinație) al relației se presupune a fi constant, atunci se va folosi forma funcțională liniară.

Spre regret, teoria în mai multe cazuri indică numai semnul relației, dar nu și forma funcțională. Forma liniară se va fi utilizată de fiecare dată când există un oarecare minim de teorie, care poate fi utilizată la fundamentarea acestei forme, până când nu se vor găsi dovezi stricte că această formă nu este potrivită. Este posibilă utilizarea modelului liniar atât timp cât teoria, bunul simț sau experiența nu justifică folosirea unei alte forme funcționale. Deoarece acest model efectiv se utilizează apriori, la el uneori se referă ca la o formă funcțională “default” implicită.

13.3 Forma exponențială sau forma logaritmică completă

Ceea mai răspândită formă funcțională (neliniară în variabile, dar liniară în coeficienți) este forma logaritmică completă. Forma logaritmică completă este des utilizată la specificarea ecuației de regresie. Spre deosebire de modelul liniar, elasticitățile

$E_{Y, X_i} = \frac{\partial Y}{\partial X_i} \frac{X_i}{Y} = \frac{\partial Y/Y}{\partial X_i/X_i} = \beta_i \frac{X_i}{Y}$ dar nu coeficienții unghiulari sunt constanți în acest model.

Dacă se presupune că elasticitățile sunt constante, atunci rezultă că $E_{Y, X_i} = \beta_i = \text{const}$. Forma funcțională exponențială $Y = e^{\beta_0} X_1^{\beta_1} X_2^{\beta_2} e^\varepsilon$ este aceea care satisface ipoteza conform căreia elasticitățile sunt constante.

Aplicând la ecuația menționată transformarea în logaritmi, prin logaritmarea ambelor părți a ecuației obținem ecuația liniară în logaritmi, care se numește formă funcțională logaritmică completă.

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \varepsilon,$$

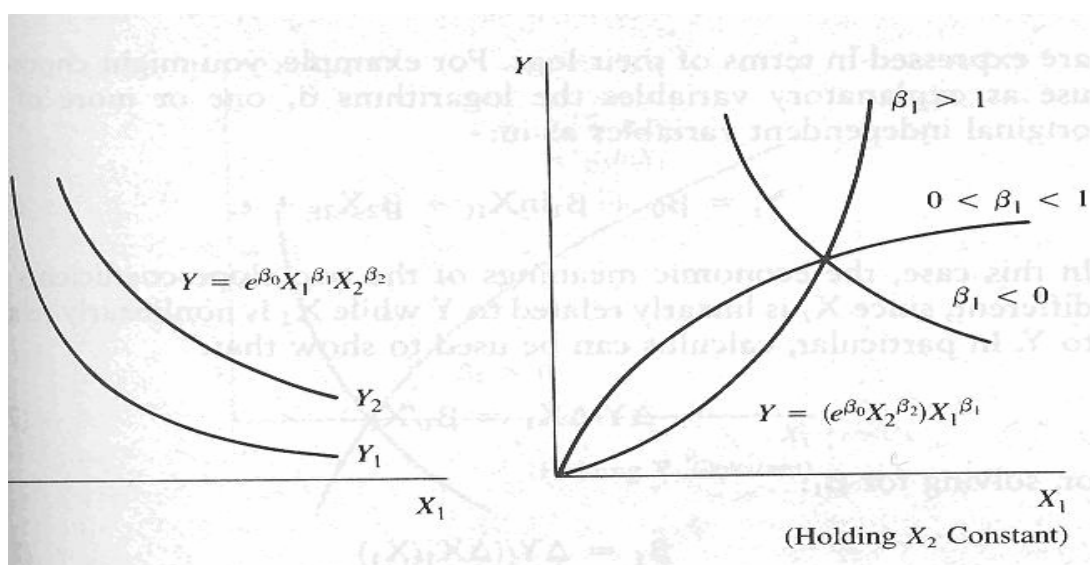
$\ln Y$ - logaritmul natural de la Y . În ecuația logaritmică completă, coeficienții individuali de regresie, de exemplu β_k , pot fi interpretați ca elasticități, deoarece

$$\beta_k = \frac{\partial \ln Y}{\partial \ln X_k} = \frac{\Delta Y / Y}{\Delta X_k / X_k} = E_{Y, X_k}.$$

Dat fiind constanți coeficienții de regresie, ecuația logaritmică completă satisface condiția ca modelul să conțină elasticități constante. Modul de interpretare a parametrilor β_k în ecuația logaritmică completă ține de faptul că la schimbarea variabilei X_k cu un %, restul variabilelor fiind menținute constante, Y se va schimba cu β_k %. În situația când elasticitățile sunt constante, coeficienții de înclinație nu mai sunt constanți.

Desenul din stînga demonstrează conceptul economic al funcției de producere-curbele de indiferență. Izocuantele funcției de producere demonstrează diferite combinații posibile ai factorilor X_1 capitalul și X_2 munca, care pot fi utilizate pentru a fabrica un volum anumit de producție. Atare funcție logaritmică completă de producere se numește funcție de producere de tip Cobb-Douglas. Desenul din dreapta demonstrează relațiile dintre Y și X_1 care există, dacă X_2 se menține constant sau nu a fost inclus în model. De menționat, că în acest caz înclinația curbei depinde de semnul și mărimea coeficientului β_1 .

Înainte de a utiliza modelul logaritmic complet, este necesar să se verifice toate observațiile ca ele să nu conțină valori de 0. Modelul logaritmic complet poate fi utilizat numai în cazul când toate variabilele primesc valori pozitive. Variabilele “dummy”, care pot lua valori 0, nu vor fi percepute chiar dacă vor fi introduse în ecuație.



13.4 Formă semilogaritmică

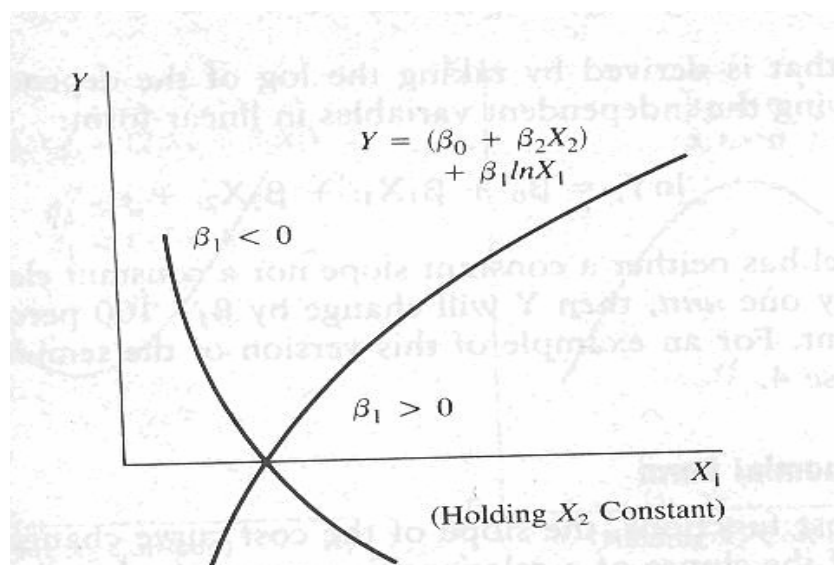
Formă semilogaritmică este o variantă a ecuației logaritmice complete în care unele, dar nu toate variabilele (dependentă și independente), sunt exprimate în termeni de logaritmi. De exemplu, $Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 X_{2i} + \varepsilon_i$. În acest caz, sensul economic al coeficienților unghiulari este diferit, în timp ce variabila X_2 este în dependență liniară față de Y , variabila X_1 este în dependență neliniară față de Y . În special, $\frac{\Delta Y}{\Delta X_1} = \beta_1 / X_1$ sau $\beta_1 = \Delta Y / (\Delta X_1 / X_1)$, fapt care

poate fi demonstrat prin calcule. Cu alte cuvinte, dacă valoarea se va schimba cu 1%, atunci valoarea Y se va schimba cu $\beta_1 / 100$, valoarea X_2 rămânând intactă (valoarea lui X_1 trebuie să fie pozitivă pentru a fi posibilă operația logaritmării). Elasticitatea variabilei Y în respect cu

variabila X_1 ea forma: $E_{Y/X_1} = \frac{\Delta Y}{\Delta X_1} \frac{X_1}{Y} = \frac{\beta_1}{Y}$, și descrește odată cu creșterea lui Y . Pe desenul ce urmează este prezentată relația dintre Y și X_1 , X_2 fiind menținută constantă. De notat că în cazul când $\beta_1 > 0$, impactul schimbărilor în X_1 asupra lui Y se află în declin odată cu creșterea lui X_1 . În concluzie, formă semilogaritmică se va fi folosită atunci, când se presupune o relație descrescătoare dintre Y și X_1 .

În economie și business aplicarea acestei forme semilogaritmice este întâlnită foarte frecvent. De exemplu, majoritatea funcțiilor de consum după un anumit nivel de venit manifestă o creștere cu rată descrescătoare. Aceste, așa numite curbe Engel, tind să devină plate deoarece atunci când venitul crește esențial, un procent mic din venit este îndreptat spre consum și un procent mai mare merge pentru acumulări. Atunci consumul crește cu o rată descrescătoare. Dacă Y este consumul al unui bun, și X_1 este venitul disponibil, (X_2 fiind menținut în locul restului variabilelor independente), atunci utilizarea formei semilogaritmice este justificată de fiecare dată când rata creșterii unui bun se așteaptă a fi în declin atunci când venitul disponibil va crește.

Alt exemplu se referă la varianta funcției semilogaritmice care se obține prin logaritmare variabilei dependente Y , variabilele independente rămânând sub forma lineară: $\ln Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$. În acest model nici coeficientul unghiular nu este constant, nici elasticitate nu sunt constante. Dacă variabila independentă X_1 se va schimba cu o unitate, atunci variabila dependentă Y se va schimba cu $\beta_1 / 100\%$, variabila independentă X_2 menținându-se constantă.



13.5 Forme polinomiale

În majoritatea funcțiilor de cost coeficientul unghiular al curbei de cost se schimbă în același mod cum se schimbă volumul. Dacă se așteaptă că coeficientul unghiular al relației să depindă de nivelul variației însuși, atunci se va considera modelul polinomial. Forma funcțională polinomială exprimă variabila dependentă Y ca funcție de variabilele independente, unele dintre care sunt ridicate la putere mai mare decât unul. De exemplu, în polinomul de gradul 2 (pătratic), măcar una din variabilele independente este ridicată la pătrat:

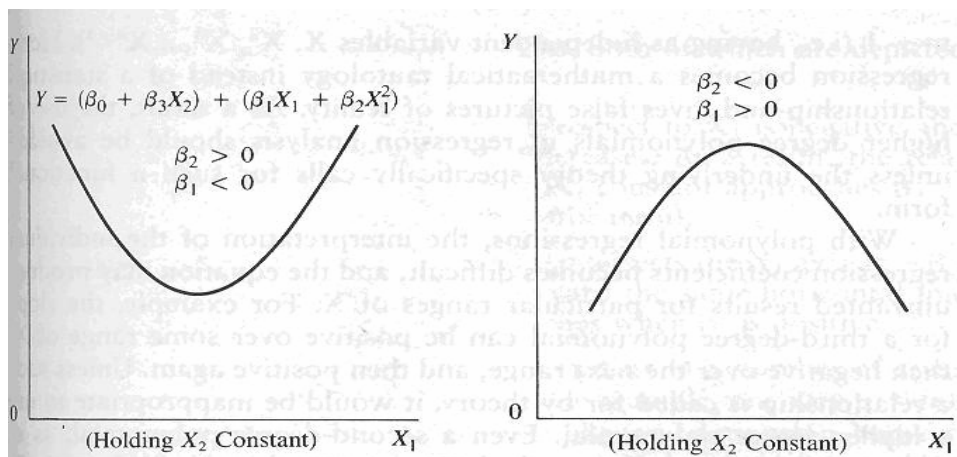
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \beta_3 X_{2i} + \varepsilon_i$. Astfel de model întradevăr poate să producă coeficientul unghiular care se schimbă odată cu schimbarea variabilei independente. În ecuația examinată unghiul de înclinație a variabilei Y în respect cu variabila X_1 este:

$\frac{\Delta Y}{\Delta X_1} = \beta_1 + 2\beta_2 X_1$, iar în respect cu variabila X_2 este $\frac{\Delta Y}{\Delta X_2} = \beta_3$. De menționat că primul coeficient unghiular depinde de valorile variabilei X_1 , iar al doilea coeficient este constant. În cazul unei funcții de cost, Y fiind costul mediu a producției, și X_1 fiind nivelul de producție al firmei, atunci dacă firma are o curbă de cost cu punct de sa (cum ar fi în figura din stângă), e posibil ca β_1 să fie negativ iar β_2 să fie pozitiv, pictată în desenul ce urmează.

Un alt exemplu, se consideră modelul veniturilor anuale ale angajaților în funcție de vârsta fiecărui angajat și de un alt factor ce stimulează productivitatea, cum ar fi educația. Care va fi impactul așteptat al vârstei asupra venitului? De regulă, cu vârsta un tânăr lucrător (el sau ea) câștigă mai mult. Oricum, dincolo de acest punct de vedere, cu vârsta în foarte multe cazuri câștigurile nu cresc deloc, dar cu apropierea vârstei de pensionare se așteaptă că câștigul va începe să descrească. Ca rezultat, relația logică dintre câștiguri și vârstă poate să arate ceva asemănător cu desenul din partea dreaptă din figura de mai jos. Câștigurile vor crește până la un nivel anumit, apoi cu înaintarea în vârstă vor descrește. Așa o relație teoretică poate fi modelată cu ajutorul ecuației pătratice: $Z_i = \beta_0 + \beta_1 V_i + \beta_2 V_i^2 + \dots + \varepsilon_i$. Care se așteaptă a fi semnele pentru β_1, β_2 ? Cu cât este mai în vârstă salariatul cu atât diferența dintre V și V^2 va crește vertiginos, întrucât V^2 va fi foarte mare. Prin urmare, coeficientul pe lângă V va fi mai important pentru o vârstă mai mică decât pentru o vârstă mai majorată. Din contra, coeficientul pe lângă V^2 va fi mai important la o vârstă mai majorată. Deoarece se preconizează că impactul vârstei este în creștere apoi descrește, e de așteptat în acest caz ca β_1 să fie pozitiv, iar β_2 să fie negativ (restul coeficienților având aceleași semne). Anume acest fenomen este exact acela care a fost observat de mulți cercetători în domeniul economiei muncii.

Spre regret, nu toate polinoamele pot fi utilizate ca curbe-mijloace de prognozare. În realitate, orice n observații pot fi approximate exact (toate reziduurile fiind egale cu 0) printr-o curbă de regresie sub formă de polinom de gradul $(n-1)$ (având ca variabilele independente $X, X^2, X^3, \dots, X^{n-1}$). În acest caz regresia se transformă într-o tautologie matematică dar nu relație statistică și ne oferă un tablou fals al realității. În concluzie, folosirea polinoamelor de grad înalt în analiza regresională va fi evitată atâta timp cât teoria corespunzătoare nu e elaborată pentru aceste forme funcționale.

În regresia polinomială interpretarea coeficienților specifici devine dificilă, și ecuația poate produce efecte nedorite pentru domenii speciale ale variabilei X . De exemplu, coeficientul individual pentru polinomul de ordinul 3 poate fi pozitiv pentru un domeniu al variabilei X , apoi negativ pentru alt domeniu ce va urma, și apoi din nou pozitiv. Utilizarea polinoamelor de ordin înalt va fi inoportună atâta timp cât nu există o teorie specială întru susținerea acestui tip de relații. Chiar și polinomul de gradul doi, cum este acela din ecuația examinată, impune un coeficient unghiular simetric ($\cup$ înclinația, sau $\cap$ inversă înclinație) care în unele cazuri poate să nu fie rezonabil. Deci, în orice probă când se folosește ecuația polinomială de regresie este necesară o atitudine de precauție, este necesară certitudinea că forma funcțională atinge acele obiective care sunt susținute din punct de vedere teoretic și nu altele.



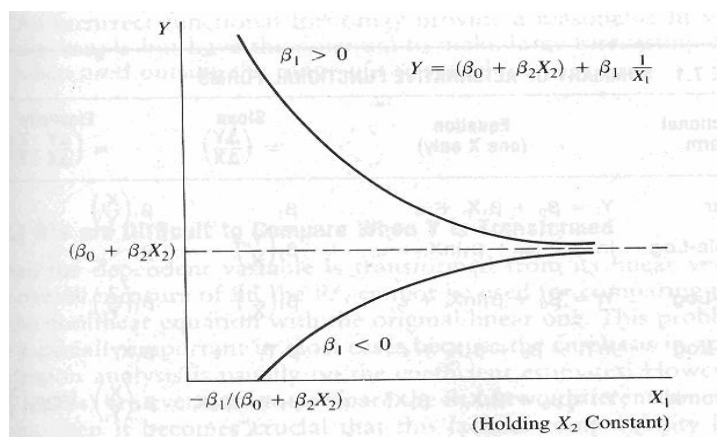
13.6. Forma inversă (hiperbola)

Forma funcțională inversă exprimă variabila dependentă Y ca o funcție inversă de una sau mai multe variabile independente (în cazul examinat numai de o singură variabilă independentă

X_1): $Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_{1i}} \right) + \beta_2 X_{2i} + \varepsilon_i$, forma funcțională inversă se va utiliza atunci când impactul

ei asupra unei variabile dependente se așteaptă a fi aproape de 0, în timp ce variabila independentă crește și eventual se apropie de infinit. Vom nota, că în asemenea circumstanțe, odată cu creșterea variabilei independente X_1 , impactul ei asupra variabilei dependente Y descrește.

În ecuația examinată variabila independentă X_1 nu poate lua valori de 0, deoarece atunci prin împărțire la zero se obține valoarea de infinit, sau o valoare nedeterminată. Coeficienții unghiulare sunt: $\frac{\Delta Y}{\Delta X_1} = -\frac{\beta_1}{X_1^2}$; $\frac{\Delta Y}{\Delta X_2} = \beta_2$; coeficientul unghiular pentru variabila independentă X_1 se încadrează în două categorii, fiecare din ele fiind oglindite pe desen:



1. Atunci când $\beta_1 > 0$, coeficientul unghiular în respect cu variabila X_1 este negativ și descrește după valoarea absolută odată cu creșterea variabilei X_1 . Ca rezultat, relația dintre Y și X_1 (variabila X_2 fiind constant) se apropie de $\beta_0 + \beta_2 X_2$ atunci când X_1 crește.
2. Atunci când $\beta_1 < 0$, curba intersectează axa X_1 în punctul $-\frac{\beta_1}{(\beta_0 + \beta_2 X_2)}$, și coeficientul unghiular este ascendent, apropiindu-se spre o linie orizontală (numită asimptotă), apropiată de valorile coeficientul unghiular și în cazul când $\beta_1 > 0$.

Forma funcțională inversă se aplică în mai multe domenii din teoria economică și din viața reală. De exemplu, să examinăm curba Philips, o relație neliniară dintre rata neangajaților în

câmpul muncii și schimbarea salariului în procente; cu siguranță, schimbările procentuale în salariu (W) se vor reflecta negativ asupra ratei neangajaților în câmpul muncii (U), majorând nivelul de neangajați; creșterea ratei neangajaților va contribui pe viitor la reducerea nivelului creșterii salariului din cauze instituționale și alte cauze. Așa o ipoteză poate fi testată cu ajutorul formei funcționale inverse: $W_t = \beta_0 + \beta_1 \left(\frac{1}{U_t} \right) + \varepsilon_t$.

Estimarea acestei ecuații prin metoda celor mai mici pătrate oferă următoarea ecuație:

$$W_t = 0.00679 + 0.1842(1/U_t); R^2 = 0.397$$

$$(0.0590)$$

$$t = 3.20$$

13.7 Probleme care apar odată cu alegerea formei funcționale incorecte

Ceea mai bună cale de a selecta pentru modelul de regresie o formă funcțională corectă constă în alegerea specificației care în cel mai reușit mod se potrivește teoriei ce stă la fundamentarea ecuației. În majoritatea cazurilor forma liniară va fi o formă adecvată. Iar pentru majoritatea cazurilor rămase, bunul simț a demonstra o alegere concretă dintre formele simple de alternativă prezentate mai jos.

<i>Sumarul formelor funcționale alternative</i>			
Forma funcțională	Ecuția	Coeficientul unghiular $= \left(\frac{\Delta Y}{\Delta X} \right)$	Elasticitatea $= \left(\frac{\Delta Y}{\Delta X} \cdot \frac{X}{Y} \right)$
Liniară	$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$	β_1	$\beta_1 \left(\frac{X_i}{Y_i} \right)$
Log completă	$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$	$\beta_1 \left(\frac{Y_i}{X_i} \right)$	β_1
Semilog ($\ln X$)	$Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$	$\beta_1 \left(\frac{1}{X_i} \right)$	$\beta_1 \left(\frac{1}{Y_i} \right)$
Semilog ($\ln Y$)	$\ln Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$	$\beta_1 Y_i$	$\beta_1 X_i$
Polinomială	$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$	$\beta_1 + \beta_2 X_i$	$\beta_1 \left(\frac{X_i}{Y_i} \right) + 2\beta_2 \left(\frac{X_i^2}{Y_i} \right)$
Inversă	$Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_i} \right) + \varepsilon_i$	$-\beta_1 \left(\frac{1}{X_i^2} \right)$	$-\beta_1 \left(\frac{1}{X_i Y_i} \right)$

Oricum, din când în când, apar circumstanțe în care modelul din considerente logice este neliniar în variabile, însă forma exactă a acestei neliniarități este greu de conceput. În așa caz, forma liniară nu este corectă, și chiar alegerea dintre diverse forme neliniare nu poate fi făcută în baza teoriei economice. Oricum, tocmai în aceste cazuri, rămâne să ne achităm (în termenii înțelegerii relației adevărate) pentru evitarea alegerii formei funcționale numai în baza aproximării reușite. Se pot da două răspunsuri la această problemă:

1. R^2 este greu de comparat atunci când variabilele dependente sunt transformate.
2. Forma funcțională incorectă poate prezenta o aproximare rezonabilă a observațiilor, dar are potențial mare de a produce erori în pronosticare, când se efectuează previziunea în afara regiunii observabile.

13.7.1 R^2 este dificil de comparat când variabila dependentă Y este transformată

Atunci când variabila dependentă este transformată de la versiunea sa liniară, R^2 nu poate fi folosit pentru compararea aproximării ecuației neliniare cu cea liniară de origine. Problema dată, în majoritatea cazurilor, nu este importantă deoarece în analiza regresională aplicată accentul se pune de regulă pe estimarea coeficienților. Oricum, dacă R^2 (sau $\bar{R}^2$) totuși se folosește pentru comparația a două aproximări pentru forme funcționale distincte, atunci aceasta devine esențial deoarece acest gol a comparabilității va fi memorat. Fie ca se face tentativa de a compara ecuația liniară $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ cu versiunea ei în semilogaritmi: $\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$. Vom menționa că unica diferență dintre aceste două ecuații constă în forma prezentării pentru variabila dependentă. Coeficientul de determinație R^2 pentru ecuația respectivă nu poate fi utilizată pentru comparația aproximațiilor a două ecuații din cauza că suma totală a devierilor pătratelor (TSS) a variabilei dependente de la valoarea medie este distinctă pentru aceste două formulări. Deci coeficientul de determinație R^2 nu poate fi comparat întrucât variabilele dependente sunt diferite. Nu există motiv pentru care două variabile dependente distincte să aibă devierea de la valoarea medie identică sau comparabilă. Deoarece (TSS) este diferită, coeficientul de determinație R^2 (sau coeficientul de determinație $\bar{R}^2$ ajustat) nu pot fi comparați.

Calea de a evita această problemă constă în crearea «quazi- R^2 » prin transformarea valorilor pronosticate a variabilei dependente neliniare într-o formă care este în mod direct comparabilă cu variabila de origine dependentă. Aceasta variabilă dependentă transformată este apoi folosită la calcularea «quazi- R^2 ». În esență, «quazi- R^2 » este R^2 care permite comparația aproximărilor ecuației obținute prin forme funcționale distincte, transformând valorile pronosticate ale unei variabile dependente în forma funcțională a altei variabile dependente.

Pentru exemplul expus ar însemna executarea următorilor etape:

- Estimarea ecuației $\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$ și determinarea $\ln \hat{Y}_{pr}$
 - Transformarea $\ln \hat{Y}_{pr}$ prin antilogaritmare *anti* $\ln(\ln \hat{Y}_i) = Y_i$.
 - Calcularea «quazi- R^2 » sau («quazi- $\bar{R}^2$ ») folosind valorile calculate noi ale antilogaritmilor drept $\hat{Y}_i$ pentru a obține reziduurile necesare în ecuația pentru R^2 . «quazi- R^2 »
- $$= 1 - \frac{\sum [Y_i - \text{anti} \ln(\ln \hat{Y}_i)]^2}{\sum [Y_i - \bar{Y}]^2}$$
- Acest «quazi- R^2 » pentru ecuația în logaritmi este în mod direct comparabil cu R^2 tradițional pentru ecuația liniară.

13.7.2 Un exemplu al folosirii incorecte de $\bar{R}^2$ (coeficient de determinație ajustat)

Pentru încorporarea impactului schimbărilor în numărul tuturor variabilelor independente este necesar de utilizat $\bar{R}^2$, care reprezintă coeficientul de determinație ajustat la numărul gradelor de libertate: $\bar{R}^2 = 1 - \frac{\sum u_i^2 / (n - k - 1)}{\sum (Y_i - \bar{Y}) / (n - 1)}$. De menționat, că numai diferența dintre R^2 și

$\bar{R}^2$ ajustat la gradul de libertate este în continuare pierdută la calculele de estimare a coeficientului de înclinație. Este cunoscut faptul că coeficienți de determinație se exprimă unul prin altul $\bar{R}^2 = 1 - (1 - R^2) \frac{(n-1)}{(n-k-1)}$, de unde rezultă că $\bar{R}^2$ va crește, va descrește sau va rămâne intact atunci când se introduce o variabilă adițională în ecuație, depinde de faptul dacă se îmbunătățește aproximația cauzată de includerea variabilei noi în depășirea pierderii gradelor de libertate.

În concluzie, avertizarea e următoarea: de fiecare dată trebuie de ținut cont de faptul că calitatea potrivirii ecuației estimate este numai una din măsurile ce contribuie la calitatea

regresiei în ansamblu. Cum a fost menționat anterior, gradul de estimare a coeficienților în conformitate cu teoria economică și așteptările preconizate cu privire la acești coeficienți sunt la fel de importante ca și potrivirea regresiei. De exemplu, o ecuație estimată cu o potrivire bună în raport cu ecuația de regresie teoretică, dar cu semne confuze pentru coeficienții estimați poate oferi previziuni neverosimile și în acest sens nu va fi o ecuație reușită. Alți factori, cum ar fi relevanța teoretică și utilitatea, la fel joacă un rol important.

Din cele expuse tragem concluzii că o potrivire mai bună pentru ecuația estimată este o soluție din cele mai bune. Însă, mulți cercetători începători presupun că dacă $\bar{R}^2$ (sau R^2, r) sunt buni, atunci ceea mai bună cale de a îmbunătăți calitatea ecuației este atingerea valorii maxime pentru unul din coeficienții nominalizați.

Probabil că examinarea unui exemplu de abuz este cel mai bun exemplu de vizualizare a pericolului încorporat în maximizarea $\bar{R}^2$, fără de a ține cont de sensul economic sau semnificația statistică a ecuației.

14. Date economice

Orice analiză cantitativă necesită a fi precedată de colectarea, organizarea și introducerea datelor în calculator. Uzual această muncă este ingrătă și ocupă mult timp, deoarece este greu de găsit datele, există diferență dintre datele teoretice și cele empirice, este sporită probabilitatea erorilor tipografice și dactilografice. Atunci când sunt cunoscute sursele de informație și datele sunt corect definite, în pofida pierderii de timp la meditarea asupra naturii datelor și colectării lor, e mai mică probabilitatea comiterii greșelilor la utilizarea sau interpretarea rezultatelor regresiei respective,.

14.1 Datele de căutat

Când se alege tema pentru cercetare în primul rând trebuie să existe certitudinea că este posibil de le găsit datele pentru variabila dependentă și toate variabilele independente relevante. În orice caz, verificarea disponibilității înseamnă deciderea asupra specificației variabilelor ce vor fi incluse în studiu. Jumătate din timp care cercetătorii începătorii îl petrec colectând datele este irosit în căutarea variabilelor incorecte din surse greșite. Câteva minute de meditație asupra naturii datelor de căutat vor salva ore de nemulțumire pe viitor.

De exemplu, fie că variabila dependentă este cantitatea televizoarelor solicitate într-un an, atunci și majoritatea variabilelor independente la fel vor fi măsurate anual. Va fi nepotrivit sau pur și simplu greșit să fie definite prețurile TV ca prețuri pentru o lună distinctă. Mai bine înțeles va fi prețul mediu pe parcursul unui an fiind raportat la numărul TV vândute într-o lună. Dacă variabila dependentă include toate televizoarele vândute, indiferent de marcă, atunci prețul cel mai potrivit va fi prețul agregat format în bază prețurilor ale tuturor claselor de TV. Și totuși calcularea unei atare variabile nu este bine venită.

Datele statistice utilizate în analiza regresională reprezintă o verigă de legătură între modelarea economică teoretică și economia reală, care necesită a fi percepută prin intermediul acestei verigi. Majoritatea proceselor și fenomenelor economice poate exact măsurată, prezentată de valori numerice. Datele cantitative pot fi obținute din Anuare Statistice fiind posibilă utilizarea lor directă sau o prelucrare adițională este necară înainte ca ele să fie folosite.

Științele economice și sociale foarte des au de a face cu fenomene cantitative de tip dihotomic. Toate evenimentele care se caracterizează de fenomenul dihotomic pot fi prezentate cu ajutorul variabilelor „dummy”.

Date statistice pot fi de natura ce va urma:

- Date temporale, în care obiectul de cercetare este fixat și este supus examinării în diferite momente de timp.

- Date încrucișate, pentru care pentru în momentul de timp fixat se cercetează diferite obiecte.
- Date de panou, care reprezintă date de chestionare, date experimentale și constituie cele mai veridice date.

14.2 Surse de date economice

- ✓ Anuare statistice ale republicii Moldova.
- ✓ Buletine statistice trimestriale, Chișinău
- ✓ Evoluția social-economică a Republicii Moldova, MER, Chișinău
- ✓ Buletine ale Băncii Naționale din Moldova.
- ✓ Tendințe în economia Moldovei.
- ✓ International Financial Statistics pub
- ✓ International Monetary Fun.
- ✓ Piața financiară, revistă Chișinău
- ✓ World Economic Outlook
- ✓ www.IREX.RU/PUBLICATIONS/POLEMICA/5/ARTY.HTM
- ✓ www.WEFA.com
- ✓ www.CIRS-md.org
- ✓ www.met.dnt.md
- ✓ <http://ecfor.rssi.ru>
- ✓ <http://www.inf.org/external/country/index.htm>

14.3.1 Variabile «proxy» sau variabile delegate

Variabilele «proxy» se utilizează pentru substituirea variabilelor teoretice necesare atunci când datele pentru variabilele respective sunt incomplete sau în genere lipsesc. De exemplu, valoarea investițiilor nete este o variabilă care nu se evaluează într-un număr relativ mare de țări. Prin urmare, cercetătorul poate folosi valorile investițiilor brute ca variabilă «proxy» în condițiile în care se presupune că valoarea investițiilor brute este direct proporțională valorii investițiilor nete. Proporționalitatea este tot ce e necesar, deoarece analiza regresională este în primul rând o relație dintre schimbări în valorile variabilelor decât schimbări în nivelul absolut a variabilelor.

În general variabila «proxy» este o variabilă delegată «bună», atunci când schimbările ei relativ bine reflectă schimbările în variabila teoretică corectă.

Datele lipsesc sau sunt incomplete:

- Date încrucișate: lipsesc careva date din una sau mai multe observații, atunci acestea observații pot fi eliminate.
- Serii temporale: lipsesc date pentru careva perioade de timp, atunci se estimează datele omise prin interpolare, folosind media datelor adiacente.
- Datele sunt trimestriale, dar sunt disponibile numai date anuale, atunci se interpoalează datele trimestriale. În orice caz interpolarea poate fi justificată atunci când schimbarea variabilei este lentă și netedă.

Atunci când datele lipsesc complet problema se agravează. Omiterea variabilelor relevante creează deplasări. De fapt, un proiect regresional reușit poate fi stopat din cauza datelor inadecvate. În multe cazuri chiar și o tehnică regresională simplă nu poate fi aplicată deoarece informația este ironată. Uneori ea este măsurată cu atâtea erori încât utilizatorul numai formează tabele și construiește grafice pentru a trage concluzii. Printre altele, aceste tabele și grafice servesc ca material auxiliar de folos la fundamentarea ecuației de regresie.

14.3.2 Utilizarea întârzierii în timp în Economie și Econometrie

Majoritatea regresiilor studiate sunt instantanee după natura lor. Cu alte cuvinte, ele includ valorile variabilelor dependentă și independente în aceeași perioadă de timp: $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$, indicele t se referă la puncte distincte în timp, și atunci când toate variabilele au același indice, ecuația este instantanee în timp. Dar nu toate situațiile în economie sau business implică relații instantanee în timp între variabila dependentă și variabilele independente. În multe cazuri este necesar de a oferi posibilitatea ca să treacă ceva timp între schimbările în variabila independentă și schimbarea rezultativă în variabila dependentă. Durata acestui timp dintre cauză și efect se numește lag. Multe ecuații econometrice includ una sau mai multe variabile independente cu întârzieri în timp cum ar fi X_{t-1} , unde indicele $t-1$ indică că observația X_{t-1} să referă la perioada de timp ce precedă momentul de timp t , cum ar fi în ecuația ce urmează:

$Y_t = \beta_0 + \beta_1 X_{1t-1} + \beta_2 X_{2t} + \varepsilon_t$. Orice schimbare pe piața agricolă, cum ar fi creșterea în preț care fermierul poate s-o câștige pentru promovarea produsului, are un efect întârziat asupra ofertei produsului, $Cantitatea\ oferită_t = f(\text{Preț}_{t-1}, \text{etc})$. Similar, multe teorii macroeconomice au o structură explicită întârziată încorporată în ele. Durata în timp dintre deciziile luate cu privire la așa politici macroeconomice (ca cheltuieli guvernamentale sau creșterea în oferta de mijloace bănești) și impactul acestora asupra Produsului Intern Brut, angajarea în serviciu sau prețuri, de regulă, se măsoară în ani. Creșterea ofertei de bani stimulează Produsului Intern Brut în partea simulării investițiilor, dar investițiile nu pot crește peste noapte deoarece deciziile necesită a fi luate, planurile necesită a fi formate, lucrători adiționali necesită a fi angajați și așa mai departe. Întrădevăr, cum a notat economistul Milton Friedman, odată estimate schimbările în politica monetară iau de la 6 până la 30 de luni pentru a fi complet plătate în economie.

Dacă este formulată ipoteza pentru un lag simplu, apar dificultăți în folosirea lagului în ecuațiile econometrice. Variabila cu întârziere în timp este introdusă în ecuația econometrică ca și o altă variabilă independentă. De exemplu, ecuația ofertei de bumbac are o formă de:

$$C_t = F(PC_{t-1}, PF_t) = \beta_0 + \beta_1 PC_{t-1} + \beta_2 PF_t + \varepsilon_t, \quad (*)$$

unde

C_t - ofertei de bumbac în anul t ;

PC_{t-1} - prețul bumbacului în anul $t-1$;

PF_t - costul muncii de fermier în anul t .

Semnificația coeficientului de regresie pentru variabilele cu întârziere nu-i aceeași ca pentru variabilele fără întârziere. Coeficientul estimat pe lângă variabilă cu întârziere măsoară schimbările în valoarea Y din anul curent atribuită la schimbarea cu o singură unitate în valoarea variabilei X din anul trecut (fiind păstrate constante valorile pentru altele variabile independente X_s în ecuație). Deci parametrul β_1 măsoară (în ecuația *) numărul unităților de bumbac care vor fi produse în plus în anul curent în urma schimbării cu o unitate a prețurilor la bumbac în anul trecut, prețul muncii fermierului rămânând ne schimbat.

14.3.3 Variabile «dummy»

Unele din variabile (de exemplu - genul) pot fi explicate numai în mod calitativ. Așa variabile în cele mai multe cazuri sunt calificate ca variabile binare sau variabile «dummy». Variabilele «dummy» iau valori de 1 sau 0 în dependență de faptul îndeplinirii sau ne îndeplinirii al anumitei condiții.

Ca să ilustrăm aceasta, vom presupune că Y_i reprezintă salariul învățătorului i școlar și nivelul lui de pregătire depinde în primul rând de gradul obținut și de experiența de învățător. Toți învățătorii au $B.$, $A.$, dar unei din ei au $M.A.$ și atunci, ecuația care reprezintă relația dintre câștigul învățătorului și gradul obținut de el poate avea forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \text{ unde}$$

$$X_{1i} = \begin{cases} 1, & \text{daca invatatorul "i" are } \hat{I} .\hat{A} \\ 0, & \text{incaz contrar} \end{cases}$$

X_{2i} - numărul de ani lucrați ca învățător pentru lectorul i . Variabila X_1 ia numai două valori 0 și 1, și deci X_1 se numește variabilă «dummy» sau pur și simplu «dummy». În cazul dat variabila «dummy» reprezintă condiția existenței gradului $M.A.$.

Coeficientul de regresie ce corespunde variabilei X_1 în această ecuație este interpretat în modul următor.

1) dacă învățătorul are numai gradul de $B., A.$, $X_1 = 0$ și $E\left(\frac{Y_i}{X_i}\right) = \beta_0 + \beta_2 X_{2i}$

2) dacă învățătorul are și gradul de $M.A.$, $X_1 = 1$ și $E\left(\frac{Y_i}{X_i}\right) = \beta_0 + \beta_1 + \beta_2 X_{2i}$

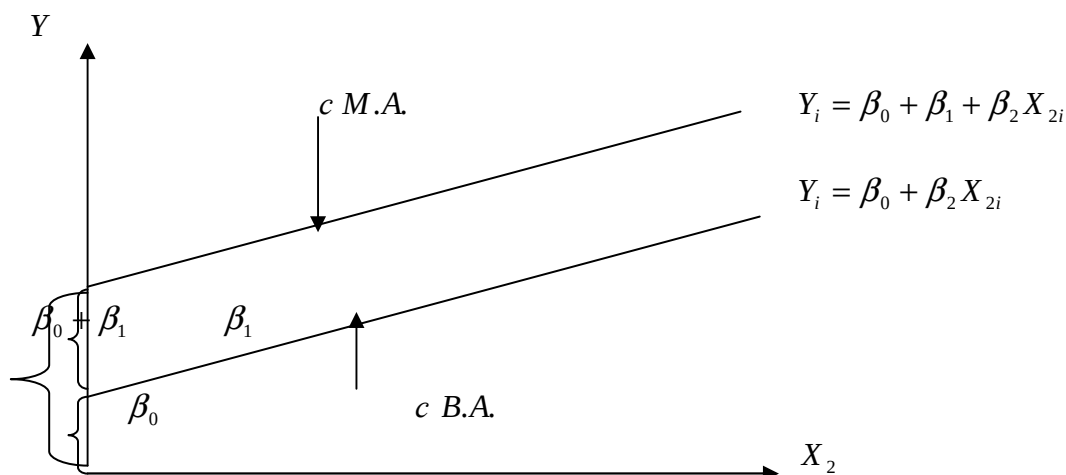
Comparând aceste două situații tragem concluzia că β_1 reprezintă câștigul mediu adițional obținut de învățătorul care are gradul de master în raport cu câștigul învățătorului care are numai de gradul de bacalavr, ambii având aceeași experiență de muncă. Abilitatea de a postula semnul coeficientului de regresie este esențială în analiza regresională. Astfel variabila «dummy» este numită variabilă «dummy» de intersecție deoarece ea în realitate schimbă punctul de intersecție a regresiei în dependență de faptul dacă are învățătorul gradul de master sau nu.

O formulare de alternativă a modelului de regresie va fi dacă vom defini X_{1i} ca:

$$X_{1i} = \begin{cases} 0, & \text{daca invatatorul "i" are } \hat{I} .\hat{A} \\ 1, & \text{incaz contrar} \end{cases}$$

Aceste condiții schimbă politica. În acest caz β_1 va fi interpretat ca diferența dintre câștigul mediu al învățătorului cu gradul de $B., A.$ și acel învățător care are gradul de $M.A.$, și semnul se presupune a fi negativ. Deoarece coeficientul β_1 a variabilei definite vine cu semnul invers, el va avea aceeași amplitudă absolută ca și β_1 pentru variabila originală. Sensul constă în aceea că acești doi coeficienți β măsoară exact același eveniment (dar în direcții opuse). În acest sens definiția variabilelor «dummy» este arbitrară. Însă, odată ce ele au fost definite, o singură interpretare poate fi prezentată.

E de menționat, că în acest exemplu s-a folosit o singură variabilă «dummy» chiar dacă au existat două condiții. Aceasta se întâmplă din cauza că condițiile se construiesc în baza unei singuri variabile. Evenimentul nu este prezentat explicit de variabila «dummy», condiția omisă formează baza cu care condiția inclusă se compară. Astfel, în situațiile duale (de felul $M.A.$ și $B., A.$), numai o singură variabilă «dummy» se include fiind independentă; coeficientul este interpretat ca efectul condiției incluse în raport cu condiția omisă. Dacă a treia condiție (cum ar fi existența $Ph.D$) va fi inclusă, atunci numai două variabile vor fi folosite.



Începătorii deseori greșesc când includ variabile “dummy” în conformitate cu numărul condițiilor, însă așa model este inutil deoarece variabila “dummy” adaugă constante, care sunt perfect multicolineare cu termenul de intersecție care deja este în ecuație. Multicolinearitatea perfectă este definită ca o relație liniară între o parte sau toate variabilele explicative. Aceasta se întâmplă deoarece suma variabilelor “dummy” este egală cu o unitate pentru toate observațiile care au fost executate. Programele de regresie, utilizate la calculator, în acest caz nu vor prezenta nici un rezultat.

Variabila “dummy”, care are o singură observație cu valoarea 1, restul observațiilor având valoare 0 (sau viceversa) va fi evitată. Așa acțiune de o dată “dummy”, pur și simplu elimina această observație din setul de date, evaluarea artificial determinând coeficientul pentru variabila “dummy” egal cu valoarea reziduală pentru aceasta observație. Aceiași estimatie poate fi obținută pentru restul coeficienților, dacă observația va fi exclusă, dar excluderea observației este pur și simplu de fiecare dată mai potrivită.

Uneori variabilele “dummy” se utilizează în calculele variațiilor sezoniere pentru datele în modele cu șiruri temporale. De exemplu, dacă

$$X_{1t} = \begin{cases} 1 & \text{in I trimestru} \\ 0 & \text{in celelalte} \end{cases}$$

$$X_{2t} = \begin{cases} 1 & \text{in trimestru II} \\ 0 & \text{in celelalte} \end{cases},$$

$$X_{3t} = \begin{cases} 1 & \text{in trimestru III} \\ 0 & \text{in celelalte} \end{cases}$$

atunci $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t$, unde X_4 - este variabila independentă diferită de “dummy”, și t este indicele observațiilor trimestriale. Vom nota că numai trei variabile “dummy” sunt necesare pentru prezentarea a patru anotimpuri. În această formulare β_1 indică cu cât valoarea așteptată pentru Y în primul trimestru diferă de valoarea așteptată a lui Y în trimestrul patru, condiția omisă. β_2 și β_3 pot fi interpretate similar.

Procedeeul poate fi aplicat numai în cazul când Y și X_4 nu sunt “ajustate sezonier” înainte de estimare. Includerea variabilelor “dummy” conform anotimpului “desezonarizează” variabila Y , la fel ca și alte variabile independente care nu sunt ajustate după anotimp.

14.4 Etapele în regresia aplicativă

- Trecerea în revistă a literaturii de specialitate
- Specificarea modelului: selectarea variabilei independente și a formei funcționale
- Lansarea ipotezelor referitor la semnul preconizat pentru coeficienți
- Colectarea datelor
- Estimarea și evaluarea ecuației
- Oformarea rezultatelor

15. Siseme de ecuații econometrice

15.1. Noțiuni generale privind sisteme de ecuații folosite în econometrie

În științele sociale obiectul de cercetare statistică este prezentat de sisteme complexe. Evaluarea stricteței legăturilor dintre variabile, construirea ecuațiilor de regresie izolate nu este suficientă pentru descrierea acestor sisteme și pentru explicarea mecanismului de funcționare ale lor. Atunci, când pentru calcule economice se recurge la utilizarea ecuațiilor de regresie separate, în majoritatea cazurilor se presupune că factorii pot fi schimbați independent unul de

Atunci când trecem de la cercetări la nivelul micro la calcule macroeconomice, utilizarea sistemelor de ecuații simultane este o necesitate stringentă. Modelul de funcționare al economiei naționale este un sistem de ecuații ce include funcții de consum; de investiții, de salarizare, identități cu privire la venituri etc. Fiind indicatori agregați, indicatorii macroeconomici, cel mai frecvent sunt interdependenți. Spre exemplu, cheltuielile pentru consumul final în economie depind de Produsul Intern Brut. În timp ce mărimea Produsului Intern Brut se consideră ca funcție de investiții.

[illegible]
$$\begin{aligned} y_1 &= f(x_1, x_2, x_3, x_4, x_5) \\ y_2 &= f(x_1, x_3, x_4, x_5) \\ y_3 &= f(x_2, x_3, x_5) \\ y_4 &= f(x_3, x_4, x_5) \end{aligned}$$

În consecință, sistemul de ecuații independente cu trei variabile dependente și patru variabile independente (patru factori) ia forma:

Însă, dacă variabila dependentă y dintr-o ecuație face parte în calitate de factor x în altă ecuație, atunci recurgem la formarea unui *sistem recursiv de ecuații*:

În acest sistem variabila dependentă y include în fiecare ecuație ulterioară în calitate de factori toate variabilele dependente din ecuațiile anterioare, la fel și mulțimea factorilor x . Ca exemplu unui atare sistem poate servi *modelul de productivitate al muncii și modelul de randament al fondurilor fixe sub forma ce urmează:*

unde y_1 este productivitatea muncii; y_2 este randamentul fondurilor fixe; x_1 este înzestrarea muncii cu fonduri; x_2 este înzestrarea muncii cu energie electrică; x_3 este nivelul de calificare al forței de muncă.

Sistem de ecuații simultane este cel mai frecvent utilizat în cercetările econometrice. În acest sistem unele și aceleași variabile dependente în unele ecuații se regăsesc în partea stângă a ecuației în timp ce în altele se regăsesc în partea dreaptă:

Acest sistem de ecuații interdependente a primit denumirea de sistem de ecuații comune, sau sistem de ecuații simultane. Prin această definiție se scoate în relief, că în sistemul examinat unele și aceleași variabile y simultan se consideră ca dependente în unele ecuații și ca independente în alte ecuații. În econometrie acest sistem de ecuații se mai numește și ca model sub formă structurală. Spre deosebire de sistemele anterioare fiecare ecuație din sistemul de ecuații simultane nu poate fi cercetată de sinistătător, și pentru determinarea parametrilor din model nu poate fi utilizată M.C.M.M.P. tradițională. În acest scop se utilizează metode speciale de evaluare.

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + \varepsilon_2, \end{cases}$$

unde y_1 este ritmul de schimbare al salariului lunar; y_2 este ritmul de modificare al prețurilor; x_1 este procentul neangajaților în câmpul muncii; x_2 este ritmul de schimbare a capitalului fix; x_3 este ritmul de modificare al prețurilor de import la materia primă.

15.2. Formele structurală și redusă ale sistemului de ecuații simultane

Sistemul de ecuații comune, simultane (sau forma structurală a modelului) de regulă este constituită din variabile endogene și exogene.

Variabilele endogene se notează prin y în sistemul de ecuații examinat anterior. Acestea sunt variabile dependente și numărul lor este egal cu numărul ecuațiilor din sistem.

Variabilele exogene se notează, de regulă, prin x . Acestea sunt variabile predeterminate, care influențează variabilele dependente, însă care nu depind de cele din urmă.

Ceea mai simplă formă structurală a modelului se exprimă ca:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + \varepsilon_2, \end{cases}$$

unde y sunt variabile endogene; x sunt variabile exogene.

Clasificarea variabilelor în endogene și exogene depinde de conceptul teoretic adoptat în model. Variabilele economice pot fi interpretate ca variabile endogene într-un model, dar în alte modele aceleași variabile se vor produce ca variabile exogene. În calitate de variabile exogene pot fi considerate valorile variabilelor endogene pentru perioada precedentă de timp (variabile întârziate). Deci, consumul anului curent (y_t) este posibil să depindă nu numai de factori economici, dar și de consumul în anul precedent (y_{t-1}). Variabile extraeconomice (cum ar fi, condițiile climatice) iau parte din sistem în calitate de variabile exogene.

Modelul sub forma sa structurală permite evidențierea impactului oricărei variabile exogene asupra variabilei endogene. Este oportun ca în calitate de variabile exogene să fie selectate variabilele, care pot fi tratate ca instrumente de control. Modificându-le și ghidându-le, avem posibilitatea să obținem anticipat valorile obiectiv ale variabilelor endogene. Modelul sub forma sa structurală încorporează pe lângă variabilele endogene și cele exogene coeficienții a_{is} și b_{ik} , care se numesc coeficienți structurali ai modelului. Toate variabilele modelului sunt exprimate în devieri de la valorile medii, încât prin variabila x se subînțelege $x - \bar{x}$, dar prin variabila y - respectiv $y - \bar{y}$, prin urmare termenul liber lipsește din fiecare ecuație.

Folosirea M.C.M.M.P. pentru estimarea coeficienților structurali, conform teoriei, de regulă, ne oferă, valori deplasate și neconsistente. Deaceia pentru determinarea coeficienților structurali ale modelului, modelul se transformă într-o formă, numită *forma redusă a modelului*. Forma redusă a modelului reprezintă un sistem de ecuații neliniar în raport cu coeficienții de pe lângă variabilele endogene și exogene din modelul sub forma sa structurală:

$$\left\{ \begin{array}{l} \hat{y}_1 = \delta_{11}x_1 + \delta_{12}x_2 + \dots + \delta_{1m}x_m, \\ \hat{y}_2 = \delta_{21}x_1 + \delta_{22}x_2 + \dots + \delta_{2m}x_m, \\ \dots\dots\dots \\ \hat{y}_n = \delta_{n1}x_1 + \delta_{n2}x_2 + \dots + \delta_{nm}x_m, \end{array} \right.$$

unde δ_{ji} sunt coeficienții modelului sub forma sa redusă

După aspectul său modelul sub forma sa redusă nici într-un fel nu se deosebește de la sistemul de ecuații independente, la care deja poate fi aplicată M.C.M.M.P. Prin aplicarea M.C.M.M.P. putem estima coeficienții δ , apoi cu ajutorul lor să evaluăm valorile variabilelor endogene prin valorile variabilelor exogene.

Coeficienții formei reduse a modelului reprezintă funcții neliniare în raport cu coeficienții formei structurale a modelului. Să examinăm acest deziderat în baza unui exemplu simplu al

formei structurale, exprimând coeficienții formei reduse a modelului δ_{ij} prin coeficienții formei structurale a modelului a_{is} și b_{ik} . În scopul simplificării în model nu este inclus termenul erorii stocastice. Pentru modelul structural sub forma:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 \\ y_2 = b_{21}y_1 + a_{22}x_2, \end{cases} \quad \text{forma redusă se prezintă ca:} \quad \begin{cases} \hat{y}_1 = \delta_{11}x_1 + \delta_{12}x_2, \\ \hat{y}_2 = \delta_{21}x_1 + \delta_{22}x_2, \end{cases}$$

Din prima ecuație a modelului structural y_2 poate fi exprimat în felul următor: $y_2 = \frac{y_1 - a_{11}x_1}{b_{12}}$.

Prin urmare sistemul de ecuații simultane va fi prezentat ca: $\begin{cases} y_2 = y_1 - a_{11}x_1/b_{12} \\ y_2 = b_{21}y_1 + a_{22}x_2, \end{cases}$

Din el obținem egalitatea $\frac{y_1 - a_{11}x_1}{b_{12}} = b_{21}y_1 + a_{22}x_2$, și atunci $y_1 - b_{12}b_{21}y_1 = a_{11}x_1 + b_{12}a_{22}x_2$ sau

$$y_1 = a_{11}x_1 / (1 - b_{12}b_{21}) + b_{12}a_{22}x_2 / (1 - b_{12}b_{21}).$$

Deci, prima ecuație din forma structurală este prezentat ca *ecuație a formei de model redusă*:

$y_1 = \delta_{11}x_1 + \delta_{12}x_2$. Din ecuația dată urmează, că coeficienții formei reduse a modelului sunt relații neliniare în raport cu coeficienții formei structurale a modelului, deci $\delta_{11} = a_{11} / (1 - b_{12}b_{21})$, $\delta_{12} = b_{12}a_{22} / (1 - b_{12}b_{21})$.

Prin analogie putem demonstra, că coeficienții din a doua ecuație (δ_{21}, δ_{22}) ai formei reduse a modelului la fel se află într-o relație neliniară față de coeficienții formei structurale a modelului. În acest scop vom exprima variabila y_1 din a doua ecuație structurală ca

$y_1 = \frac{y_2 - a_{22}x_2}{b_{21}}$ sau, dacă înscrinem aceasta expresie în partea stângă a primei ecuații din forma

structurală a modelului, obținem $(y_2 - a_{22}x_2)/b_{21} = b_{12}y_2 + a_{11}x_1$.

De unde avem $y_2 = \frac{a_{11}b_{21}}{1 - b_{21}b_{12}}x_1 + \frac{a_{22}}{1 - b_{21}b_{12}}x_2$, ceea ce corespunde ecuației din forma redusă a

modelului: $y_2 = \delta_{21}x_1 + \delta_{22}x_2$ deci $\delta_{21} = a_{11}b_{21} / (1 - b_{21}b_{12})$ $\delta_{22} = a_{22} / (1 - b_{21}b_{12})$.

Însă în modelele econometrice, de regulă, sunt incluse nu numai ecuații, ce exprimă legăturile între variabile separate și tendințele de evoluție a evenimentului, dar și diverse identități. Așa doar, în anul 1947, cercetând dependența liniară a consumului (c) de la venitul (y), T. Havelmo a propus să fie considerată simultan și identitatea de venit. În acest caz modelul se prezintă ca:

$\begin{cases} c = a + by \\ y = c + x, \end{cases}$ unde x sunt investițiile în capitalul fix și în stocurile de export și import, a, b sunt

parametrii dependenței liniare ai variabilei endogene c de variabila endogenă y . Estimările lor trebuie să țină cont de identitatea de venit în deosebire de estimările parametrilor în regresia liniară obișnuită.

Acest model conține două variabile endogene c, y și o variabilă exogenă x . Sistemul redus

de ecuații va alcătui: $\begin{cases} c = A_0 + A_1x \\ y = B_0 + B_1x \end{cases}$ $x = (y - B_0) / B_1, c = A_0 + A_1(y - B_0) / B_1$

$$c = A_0 - A_1B_0 / B_1 + A_1y / B_1 = a + by, \quad a = A_0 - A_1B_0 / B_1, \quad b = A_1 / B_1.$$

Din el putem obține valorile variabilei endogene c prin valorile variabilei exogene x . Calculând coeficienții modelului (A_0, A_1, B_0, B_1), putem trece la coeficienții modelului sub forma sa structurală a, b , substituind în prima ecuație a modelului sub forma sa redusă expresia pentru x din a doua ecuație a modelului sub forma sa redusă. Forma redusă a modelului, deși permite să obținem valorile variabilei endogene prin valorile variabilei exogene, în sens analitic este

inferioară modelului sub forma sa structurală, deoarece în ea lipsesc legăturile dintre variabilele endogene.

16. Problema de identificare a modelului sub forma sa redusă

Odată cu trecerea de la modelul sub forma sa redusă la modelul sub forma sa structurală, cercetătorul se confruntă cu problema de identificare. Identificarea nu e altceva decât corespunderea univocă dintre forma redusă și forma structurală a modelului.

Vom examina problema de identificare pentru cazul sistemului de ecuații cu două variabile endogene. Fie că modelul sub forma sa structurală este exprimat ca:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + \dots + a_{1m}x_m, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \dots + a_{2m}x_m, \end{cases} \text{ unde } y_1 \text{ și } y_2 \text{ sunt variabile dependente simultane.}$$

Din a doua ecuație putem exprima y_1 prin următoarea formulă: $y_1 = \frac{y_2}{b_{21}} - \frac{a_{21}}{b_{21}}x_1 - \dots - \frac{a_{2m}}{b_{21}}x_m$.

Atunci în sistem avem două ecuații pentru o singură variabilă endogenă y_1 cu același set de variabile exogene dar cu coeficienți diferiți pe lângă ele:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + \dots + a_{1m}x_m, \\ y_1 = y_2 / b_{21} - a_{21}x_1 / b_{21} - a_{22}x_2 / b_{21} - \dots - a_{2m}x_m / b_{21} \end{cases}$$

Existența a două variante de calcul pentru coeficienții structurali ai aceluiași model ține de identificarea incompletă a celei din urmă. Modelul sub forma sa structurală completă conținând în fiecare ecuație din sistem n variabile endogene și m variabile exogene, este constituit din $n(n-1+m)$ parametri. Deci, cu $n=2$ și $m=3$, prezentarea completă a modelului sub forma

sa structurală este: $\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + a_{13}x_3, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3, \end{cases}$. Observăm, că modelul conține opt

coeficienți structurali, ce corespunde expresiei $n(n-1+m)$.

Modelul complet sub forma sa redusă conține nm parametri. Ceea ce pentru ultimul exemplu înseamnă existența a șase coeficienți ai modelului sub forma sa redusă. Acest fapt poate fi confirmat dacă ne adresăm la modelul sub forma sa redusă, care se exprima în felul următor:

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + \delta_{13}x_3, \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3 \end{cases}. \text{ Întradevăr, acest model conține șase coeficienți } \delta_{ij}. \text{ În baza acestor}$$

șase coeficienți ai modelului redus este necesar să determinăm opt coeficienți structurali ai modelului structural, ceea ce, în mod natural, nu poate conduce la o soluție unică. Modelul structural complet conține mai mulți parametri decât modelul redus. Respectiv $n(n-1+m)$ parametri ai modelului structural nu pot fi determinați în mod univoc cu ajutorul nm parametri ai modelului redus.

Pentru a obține soluția unică posibilă pentru modelul structural este necesar să presupunem, că unii dintre coeficienții modelului, demonstrând o relație insuficientă a factorilor cu variabila endogenă din partea stângă a sistemului, sunt egali cu zero. În așa mod se va micșora numărul coeficienților structurali din model. Deci, dacă vom admite, că în modelul examinat $a_{13} = 0, a_{21} = 0$, modelul structural se va prezenta ca:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 \end{cases}. \text{ În acest model numărul coeficienților structurali nu depășește numărul}$$

coeficienților din modelul redus, care este egal cu 6. Micșorarea numărului coeficienților structurali din model este posibilă și în alt mod: de exemplu prin egalarea unor coeficienți între

ei, deci prin admiterea, că impactul lor asupra variabilei endogene este același. Asupra coeficienților structurali pot fi aplicate restricții de felul $b_{ij} + a_{ij} = 0$.

De pe poziția de identificare modelele structurale pot fi împărțite în trei categorii:

- modele ce pot fi identificate;
- modele ce nu pot fi identificate;
- modele ce sunt supraidentificate.

Modelul poate fi identificat, dacă toți coeficienții structurali sunt determinați în mod inivoc, deci numărul de parametri din modelul structural este egal cu numărul de parametri din modelul redus. În acest caz coeficienții structurali din model sunt evaluați prin parametrii modelului redus și este posibil de identificat modelul. Modelul structural cu două variabile endogene și trei variabile exogene, examinat anterior, care conține șase coeficienți structurali reprezintă un model identificat.

Modelul nu poate fi identificat, dacă numărul coeficienților reduși e mai mic decât numărul coeficienților structurali, prin urmare coeficienții structurali nu pot fi estimați cu ajutorul coeficienților modelului sub forma sa redusă. Modelul structural complet, care conține n variabile endogene și m variabile exogene (predeterminate) în fiecare din ecuațiile sistemului, nu poate fi identificat.

Modelul este supraidentificat, dacă numărul coeficienților reduși e mai mare decât numărul coeficienților structurali. În acest caz cu ajutorul coeficienților modelului sub forma sa redusă pot fi obținute două sau mai multe valori pentru un singur coeficient structural. În atare model numărul coeficienților structurali e mai mic decât numărul coeficienților modelului sub forma sa redusă. Deci, dacă în modelul structural complet se admite că unii coeficienți iau valori nule $a_{13} = 0$, $a_{21} = 0$, dar și $a_{22} = 0$, atunci sistemul de ecuații devine supraidentificat:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{23}x_3 \end{cases} . \text{ În acest sistem cinci coeficienți structurali nu pot fi determinați univoc}$$

folosind șase coeficienți din modelul sub forma sa redusă. Modelul supraidentificat, spre deosebire de modelul, care nu poate fi identificat, practic poate fi soluționat, însă necesită procedee speciale pentru calcularea parametrilor.

Modelul structural este un sistem de ecuații simultane, în care orice ecuație necesită a fi verificată privind subiectul de identificare. *Modelul se consideră identificabil, dacă fiecare ecuație din acest sistem poate fi identificată.* În cazul în care cel puțin o ecuație, care face parte din sistem, nu poate fi identificată, și atunci modelul integral se consideră imposibil de identificat. Modelul supraidentificat conține cel puțin o ecuație, care este supraidentificată.

Îndeplinirea condițiilor de identificare ale modelului se verifică pentru fiecare ecuație din sistem. Pentru ca ecuația să fie identificată este necesar *ca numărul variabilelor predeterminate, care nu fac parte din ecuație însă este prezent în sistem, să fie egal cu numărul variabilelor endogene, prezente în ecuația examinată, fără una.*

Dacă să notăm numărul variabilelor endogene în ecuația j prin H , iar numărul variabilelor exogene, care fac parte din sistem, însă nu sunt incluse în ecuația în examinare, prin D atunci condiția de identificare a modelului va lua forma următoarei reguli:

$D + 1 = H$ - ecuația poate fi identificată;

$D + 1 < H$ - ecuația nu poate fi identificată;

$D + 1 > H$ - ecuația este supraidentificată.

Admitem, că se consideră următorul sistem de ecuații simultane:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{33}x_3 + a_{34}x_4. \end{cases}$$

Prima ecuație este exact identificată deoarece în ea sunt prezente trei variabile endogene - y_1, y_2, y_3 , deci $H = 3$, și două variabile exogene - x_1, x_2 , numărul variabilelor exogene absente este egal cu doi - x_3 și x_4 , $D = 2$. Rezultă că se îndeplinește egalitatea: $D + 1 = H$, t.e. $2 + 1 = 3$, ceea ce înseamnă prezența ecuației identificabile.

În a doua ecuație din sistem $H = 2$ (y_1 și y_2) și $D = 1$ (x_4). Are loc egalitatea $D + 1 = H$, $1 + 1 = 2$. Prin urmare, a doua ecuație este identificabilă.

Din a treia ecuație desprindem, că $H = 3$ (y_1, y_2, y_3) iar $D = 2$ (x_1, x_2), deci, în conformitate cu regula de calcul $D + 1 = H$, și aceasta ecuație este identificabilă. În așa mod este identificabil sistemul integral.

Să admitem că în modelul examinat $a_{21} = 0, a_{33} = 0$, atunci sistemul are forma:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{34}x_4. \end{cases}$$

Prima ecuație din acest sistem nu s-a modificat. Sistemul continuie să conțină trei variabile endogene și patru variabile exogene, deaceia pentru această ecuație $D = 2$ și $H = 3$ și ea este identificabilă. A doua ecuație dă dovadă de $H = 2$ și $D = 2$ (x_1, x_4), și regula de calcul ne oferă $2 + 1 > 2$. Prin urmare această ecuație este supraidentificată. La fel se adevărește că este supraidentificată și a treia ecuație, în care $H = 3$ (y_1, y_2, y_3) și $D = 3$ (x_1, x_2, x_3), deci regula de calcul demonstrează inegalitatea: $3 + 1 > 3$ sau $D + 1 > H$. Modelul integral este supraidentificat.

Să presupunem, că ultima ecuație din sistemul cu trei variabile endogene are forma: $y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + a_{34}x_4$, deci spre deosebire de ecuația precedentă în ea au fost incluse încă două variabile exogene, care fac parte din sistem x_1, x_2 . În acest caz ecuația devine neidentificabilă deoarece cu $H = 3$ și $D = 1$, $D + 1 < H$, iar $1 + 1 < 3$. În pofida faptului, că prima ecuație este identificată, a doua ecuație este supraidentificată, de aici rezultă că modelul se consideră neidentificat, deci nu are soluție statistică.

Pentru estimarea coeficienților modelului structural este necesar ca sistemul de ecuații să fie posibil de identificat sau acest sistem de ecuații să fie supraidentificat.

Regula de calcul considerată reprezintă o condiție necesară însă nu și suficientă pentru ca sistemul de ecuații să fie posibil de identificat. O condiție mai perfectă se determină în cazul în care asupra coeficienților matricei formate din parametrii modelului structural se aplică unele condiții. Ecuația poate fi identificată, dacă în baza variabilelor endogene și exogene, care nu fac parte din ecuație, poate fi obținută o matrice din coeficienții ei pe lângă alte ecuații din sistem, determinantul căreia nu este egal cu zero iar rangul matricei nu e mai mic decât numărul variabilelor endogene din sistem fără una.

Oportunitatea verificării condiției de identificare a modelului prin determinantul matricei formate din coeficienți pe lângă variabilele ce lipsesc în ecuația examinată, dar care sunt prezente în alte ecuații ai sistemului, se explică prin faptul că e posibilă situația, pentru care regula de calcul este îndeplinită, însă determinantul matricei pe lângă coeficienții numiți este egal cu zero. În acest caz are loc numai condiția necesară pentru a fi identificată ecuația examinată, în timp ce condiția suficientă este violată.

Să considerăm următorul model structural:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \end{cases}$$

condiției necesare și condiției suficiente pentru a fi identificată ecuația. Pentru prima ecuație are loc: $H = 3$ (y_1, y_2, y_3) și $D = 2$ (x_3, x_4 lipsesc, atunci $D + 1 = H$ și condiția necesară de identificare este satisfăcută, prin urmare, ecuația este exact identificată. Pentru verificarea

condiției suficiente se va completa următorul tabel, format din coeficienții pe lângă variabilele care nu fac parte din prima ecuație. Determinantul acestei matrice este egal cu zero, $\det A = 0$.

Ecuații	Variabile	
	x_3	x_4
2	a_{23}	a_{24}
3	0	0

În a doua ecuație avem: $H = 2$ (y_1, y_2) și $D = 1$ (x_1 lipsește), regula de calcul confirmă faptul, că ecuația este posibil de identificat ($D + 1 = H$). Este îndeplinită și condiția suficientă pentru a fi identificată ecuația în cauză. Coeficienții de pe lângă variabilele, care nu fac parte din a doua ecuație formează următoarea matrice:

Ecuații	Variabile	
	y_3	x_1
2	b_{13}	a_{11}
3	-1	a_{31}

În conformitate cu acest tabel, $\det A \neq 0$, rangul matricei este egal cu 2, ceea ce corespunde următorului criteriu: rangul matricei formate din coeficienți trebuie să fie nu mai mic decât numărul variabilelor endogene fără una, deci a doua ecuație poate fi exact identificată.

A treia ecuație din sistem conține 2 variabile endogene și două variabile exogene, care nu aparțin ecuației și $H = 3$ și $D = 2$, deci în conformitate cu condiția necesară această ecuație este exact identificată ($D + 1 = H$). O concluzie contrară obținem verificând condiția suficientă de identificare. Să alcătuim tabelul coeficienților de pe lângă variabilele, care nu aparțin acestei ecuații, din care conchidem că $\det A = 0$:

Ecuații	Variabile	
	x_3	x_4
2	0	0
3	a_{23}	a_{24}

Din tabel observăm că se violează condiția suficientă de identificare, prin urmare ecuația nu poate fi identificată. Și atunci modelul structural examinat nu poate fi identificat în ansamblu, deoarece, dat fiind satisfăcută condiția necesară, este violată condiția suficientă.

Deseori în modelele econometrice odată cu ecuațiile, parametrii cărora necesită a fi estimați statistic, se folosesc identități de balanță cu participarea variabilelor din model, coeficienții de pe lângă aceste variabile sunt egali cu ± 1 . Pentru acest caz, necătând la faptul că însuși identitățile nu necesită verificare în privința identității, în verificarea ecuațiilor structurale din sistem aceste identități participă.

De exemplu, să parcurgem la examinarea modelului econometric ce descrie economia unei țări:

$$\begin{cases} y_1 = A_{01} + b_{13}y_3 + b_{14}y_4 + \varepsilon_1 \\ y_2 = A_{02} + b_{23}y_3 + a_{21}x_1 + \varepsilon_2, \\ y_3 = A_{03} + b_{34}y_4 + a_{31}x_1 + \varepsilon_3 \\ y_4 = y_1 + y_2 + x_2 \end{cases},$$

unde y_1 sunt cheltuielile de consum final pentru anul curent; y_2 sunt investițiile brute în anul curent; y_3 sunt cheltuielile pentru salarizare în anul curent; y_4 este venitul brut pentru anul curent; x_1 este venitul brut pentru anul precedent; x_2 sunt cheltuielile guvernamentale în anul curent; A_{0i} este termenii liber din ecuația i ; ε_i este eroarea stocastică din ecuația i . Din acest model fac parte patru variabile endogene (y_1, y_2, y_3, y_4), de menționat că una dintre ele, y_4 este definită cu ajutorul unei identități. Și atunci, soluția statistică este necesară numai pentru

primele trei ecuații ai modelului, care este necesar de verificat în vederea identificării. Modelul conține două variabile predeterminate, una dintre care x_2 este exogenă iar alta este întârziată x_1 .

La soluționarea practică a problemei în baza informației statistice pentru un șir de ani sau în baza unei totalități de regiuni pentru un singur an în ecuațiile pentru variabilele endogene y_1 , y_2 , y_3 , de regulă, participă termenul liber A_{0i} , $i=\overline{1,3}$, valoarea căruia cumulează impactul factorilor care nu au fost incluși în model și nu are nici o influență asupra problemei de identificare a modelului.

Deoarece datele reale referitor la variabilele endogene y_1 , y_2 , y_3 pot să difere de la acele teoretice, postulate în model, este oportun ca în model să se introducă componenta aleatoare pentru fiecare ecuație din model, cu excepția identităților. Componenta aleatoare (devierile), notate ca ε_i nu influențează soluționarea problemei de identificare a modelului.

În modelul econometric examinat prima ecuație din sistem poate fi identificată deoarece pentru ea $H=3$, $D=2$ și are loc condiția necesară de identificare ($D+1=H$). Plus la aceasta, este veridică și condiția suficientă de identificare, în așa fel că rangul matricei respective este egal cu 3, iar determinantul ei nu este egal cu zero $\det A \neq 0$.

Ecuatii	y_2	x_1	x_2
2	-1	a_{21}	0
3	0	a_{31}	0
4	1	0	1

La fel și a doua ecuație din sistem este exact identificată deoarece are loc: $H=2$ и $D=1$, deci se realizează regula de calcul ($D+1=H$), în același timp are loc și condiția suficientă de identificare, și anume: rangul matricei examinate este egal cu 3, dar determinantul ei nu este egal cu zero: $\det A = -b_{34}$:

Ecuatii	y_1	y_4	x_2
1	-1	b_{14}	0
3	0	b_{34}	0
4	1	-1	1

În mod analogic se identifică și a treia ecuație din sistem deoarece $H=2$ и $D=1$, deci se îndeplinește regula de calcul ($D+1=H$), în același timp se îndeplinește și condiția suficientă de identificare, care constată că rangul matricei este egal cu 3, iar determinantul ei nu este egal cu zero: $\det A = 1$.

Ecuatii Ecuatii	y_1	y_2	x_2
1	-1	0	0
2	0	-1	0
4	1	1	1

Identificarea ecuațiilor este un procedeu suficient de complicat și nu poate fi limitat la examinarea situațiilor expuse anterior. Coeficienții structurali ai modelului pot fi supuși unor restricții adiționale, de exemplu pentru funcția de producere poate fi lansată ipoteza, că suma elasticităților să fie egală cu zero. Pot fi aplicate restricții asupra dispersiilor și covariațiilor valorilor reziduale.

Coeficienții modelului structural pot fi estimați aplicând diferite metode în dependență de tipul sistemului de ecuații simultane. Cele mai uzuale și mai bine cunoscute din literatura de specialitate sunt următoarele *metode de estimare a coeficienților structurali din model*:

- metoda celor mai mici pătrate indirectă;
- metoda celor mai mici pătrate în două trepte;
- metoda celor mai mici pătrate în trei trepte;

- metoda de maximă veridicitate cu informație completă;
- metoda de maximă veridicitate cu informație incompletă.

Metoda celor mai mici pătrate indirectă M.C.M.M.P.I. se aplică pentru sistemul simultan de ecuații, care poate fi identificat, iar metoda celor mai mici pătrate în două trepte M.C.M.M.P.D.T. se folosește pentru estimarea coeficienților modelului supraidentificat. Metodele rămase de estimare se utilizează și la estimarea sistemelor de ecuații simultane supraidentificate.

Metoda de maximă veridicitate se consideră cea mai generalizată metodă de estimare, rezultatele obținute cu ajutorul ei, dat fiind factorii distribuiți normal, coincid cu acelea obținute prin intermediul M.C.M.M.P. Însă, în cazul când numărul ecuațiilor din sistem este foarte mare, această metoda conduce la procedee de calcul foarte sofisticate. Deaceia în calitate de alternativă se utilizează metoda de maximă veridicitate cu informație incompletă (metoda celui mai mic raport dispersional).

Spre deosebire de metoda de maximă veridicitate cu informație completă în această modificare sunt scoase restricțiile asupra parametrilor, care se referă la funcționarea sistemului în întregime. Ceea ce conduce la o soluție mai simplă, însă volumul de calcul rămâne suficient de înalt. Necâtând la popularitatea sporită a acestei metode în mijlocul anilor 60, ea a fost practic înlocuită cu metoda celor mai mici pătrate în două trepte M.C.M.M.P.D.T. fiind mult mai simplă.

Metoda celor mai mici pătrate în trei trepte M.C.M.M.P.T.T. este o extensiune a M.C.M.M.P.D.T. Această metodă de estimare poate fi aplicată pentru toate tipurile de ecuații ale modelului structural. Însă, în cazul când există restricții asupra parametrilor, mai eficientă se adevărește M.C.M.M.P.D.T.

REFERINȚE

1. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики: Учебник. ЮНИТИ, Москва, 1998.
2. Джонстон Дж. Эконометрические методы. Статистика, Москва, 1980.
3. Эконометрика. Под редакцией члена-корреспондента Российской Академии Наук И. И. Елисеевой. Финансы и статистика, Москва, 2001, 343 с.
4. dr. Nicos Economou. An Estimation of the Potential Output and the Output Gap of the Moldovan Economy. Chisinau, MER, TACIS. Moldovan Economic Trends, 2002, q.4., pp.85-93.
5. Маленко Э. Статистические методы эконометрии. Статистика, Москва, 1976.
6. Моделирование экономических процессов. Под редакцией доктора экономических наук, профессора М. В. Грачевой, кандидата физико-математических наук, доцента Л. Н. Фадеевой, доктора экономических наук, профессора Ю. Н. Черемных. ЮНИТИ, Москва, 2005, 353 с.
7. dr. Apostolos Papaphilippou. An Econometric Estimation of the Import Demand in Moldova. Chisinau, MER, TACIS. Moldovan Economic Trends, 2001, q.3., pp.93-99.
8. Ion Pârțachi, Alexandru Brailă, Natalia Șişcanu. Econometrie Aplicată. A.S.E.M., Chișinău, 1999, 172 p.
9. Pecican E., Econometrie. Editura All, București, 1994.
10. Schatteles T. Metode econometrice moderne, Universitas, Chișinău, 1992.
11. A. H. Studenmund. Using Econometrics (second edition). Washington. HarperCollins Publishers Inc. 1992, 662 p.
12. Титнер Г. Введение в эконометрию. Статистика, Москва, 1978.
13. Четыркин Е. М. Статистические методы прогнозирования. Статистика, Москва, 1977.
14. Zaman C. Econometrie, București, 1998.

LUCRAREA DE LABORATOR Nr. 1 EVALUAREA ECONOMETRICĂ A FUNCȚIEI CERERII PENTRU IMPORT

Etapa I. Regresia statică

Să se estimeze coeficienții de regresie pentru funcția cererii la import, prezentată sub forma logaritmică completă:

$$\ln M_t = \beta_0 + \beta_1 \ln Y_t + \beta_2 \ln P_t, \text{ unde} \quad (1),$$

M_t – importurile reale trimestriale, exprimate în \$ SUA,

Y_t – produsul intern brut trimestrial, exprimat în \$ SUA,

P_t – prețul relativ trimestrial, care este raportul dintre indicele prețului mondial pentru import și indicele prețului de consum $P_t = e_t^* P_t^{*M} / IPC_t$,

P_t^{*M} – indicele prețului mondial pentru import, date trimestriale;

e_t – rata de schimb normată la valoarea primului trimestru 1996;

IPC_t – indicele prețului de consum, date trimestriale.

Să fie:

- a) consultată teoria cu privire la prezentarea analitică a cererii pentru import;
- b) specificat modelul;
- c) examinate semnele fiecărui coeficient a modelului specificat;
- d) selectate datele necesare pentru efectuarea analizei regresionale;
- e) folosită metoda celor mai mici pătrate la estimarea și evaluarea formei funcționale propuse (statisticile Student, coeficienții de determinare, statistica Fisher);
- f) documentate rezultatele.

Să se introducă variabila binară $Dummy_t$ ce va lua în considerație consecințele crizei financiare din Russia, anul 1998, care ia forma:

$$Dummy_t = \begin{cases} 1 & \text{IIItr. 1998, IVtr. 1998, Itr. 1999, IItr. 1999,} \\ 0 & \text{pentru trimestrele rămase.} \end{cases}$$

În continuare să fie examinată în calitate de funcție cererii pentru importa forma semilogaritmică ce urmează:

$$\ln M_t = \beta_0 + \beta_1 \ln Y_t + \beta_2 \ln P_t + \beta_3 Dummy_t \quad (2),$$

Să fie executate punctele a)-f) de mai sus.

Etapa II. Regresia dinamică

Să se estimeze coeficienții de regresie pentru funcția cererii la import, prezentată sub forma semilogaritmică cu întârziere în timp:

$$\ln M_t = \beta_0 + \beta_1 \ln Y_t + \beta_2 \ln P_t + \beta_3 M_{t-1} \quad (3),$$

Să fie executate punctele a)-f) de mai sus.

Și, în final, să se estimeze coeficienții de regresie pentru funcția cererii la import, prezentată sub forma semilogaritmică cu întârziere în timp și cu variabila binară $Dummy_t$:

$$\ln M_t = \beta_0 + \beta_1 \ln Y_t + \beta_2 \ln P_t + \beta_3 M_{t-1} + \beta_4 Dummy_t \quad (4).$$

Trimestrul I al anului 1996 va fi trimestru de bază. În baza ratei de schimb se vor recalcula datele pentru Produsul Intern Brut real. La calcularea Indicelui Prețului de Consum se vor folosi datele pentru inflația trimestrială. Valorile indicatorilor reali se calculează în baza IPC. $M_t \$real = M_t \$nom. / IPC$; $Y_t \$real = Y_t \$nom. / IPC$. $IPC_t = \prod_{k=tr.b}^n IPC_k * (1 + infl_{k+1} / 100)$.

Datele pot fi selectate de pe www.statistica.md.

LUCRAREA DE LABORATOR Nr. 2

ESTIMAREA POTENȚIALULUI ECONOMIC AL MOLDOVEI

Etapa I. Aplicarea filtrului Hodric-Prescott

Să se estimeze trendul care reprezintă PIB real potențial în baza filtrului Hodrick-Prescott (HP). Aceasta se obține prin determinarea trendului ce corespunde PIB real potențial care minimizează simultan media ponderată dintre trendul estimat și valorile PIB real observat în orice moment de timp și rata schimbării dintre trendul estimat în orice moment de timp. Aceasta se obține prin minimizarea funcției obiectiv ce urmează:

$$\sum (\ln Y_t - \ln Y_t^*)^2 + \lambda \sum [(\ln Y_{t+1}^* - \ln Y_t^*) - (\ln Y_t^* - \ln Y_{t-1}^*)]^2 \quad (1),$$

unde $\ln Y_t$ și $\ln Y_t^*$ sunt logaritmele PIB real și trendului estimat respectiv.

$\sum (\ln Y_t - \ln Y_t^*)^2$ este suma pătratelor devierilor dintre actualul PIB - $\ln Y_t$ și trendul corespunzător $\ln Y_t^*$. $\lambda \sum [(\ln Y_{t+1}^* - \ln Y_t^*) - (\ln Y_t^* - \ln Y_{t-1}^*)]^2$ reprezintă funcția de penalitate care penalizează pătratul devieilor de la rata de creștere a componentelor trendului; λ factorul ce prezintă ponderea și care controlează cât de netedă este linia trendului obținut.

Utilizând programul Eviews să se estimeze trendul PIB-ului potențial pentru $\lambda=10;30;100$.

Y_t – produsul intern brut anual, exprimat în MDL pentru anii 1995-2002. Datele le găsiți pe situl : <http://www.statistica.md>.

Rezultatele să fie prezentate atât grafic cât și prin tabele.

Etapa II. Abordarea problemei prin utilizarea funcției de producere

Să se estimeze coeficienții de regresie pentru funcția de producție de tip Cobb-Gouglas cu rentabilitatea la scară constantă și factorii de producere munca N_t și capitalul K_t utilizați:

1. $Y_t = K^{1-\alpha} N^\alpha$ sub forma logaritmică $\ln Y_t = \beta_1 \ln K_t + \beta_2 \ln N_t$
2. Să se estimeze funcția de producere sub forma logaritmică completă pentru valorile PIBreal efectiv și PIBreal potențial.
3. Toți indicatorii ce fac parte din calcule să fie calculați în termeni reali.
4. În baza funcțiilor de producere estimate să se efectueze pronosticul pentru anii 2005-2010, dat fiind rata creșterii capitalului utilizat de 5%,10%,15%,20% și 25% anual respectiv; iar rata creșterii muncii de 10% anual.
5. Rezultatele să fie prezentate atât sub formă de tabele cât și sub formă grafică.

Tab.1 Date istorice cu privire la subiectul studiat

Anii	1995	1996	1997	1998	1999	2000	2001	2002	2003
PIB_tnominalefec	6480	7798	8917	9122	12322	16020	19052	22556	27297
<i>PIB_trealefect</i>									
Infla_t	0,3	0,24	0,12	0,08	0,39	0,31	0,10	0,05	0,012
Defl.PIB₁₉₉₅									
K_tnominal	14450	16138	14743	24702	30926	33598	34325	35827	37782
<i>K_treal</i>									
L_tnominal	2800	3623	4153	4689	5207	7108	9322	12729	18508
<i>L_treal</i>									
PIB_tpotnom	6574	7879	8026	11126	13150	15941	18336	21802	26824
PIB_tpotreal									

LUCRAREA DE LABORATOR Nr. 3-4 EVALUAREA PRODUSULUI INTERN BRUT DIN PARTEA CERERII

Etapa I. Prezentarea funcției de consum sub forma funcției de PIB

Să se estimeze consumul ca:

$$C_t = cY_{t-1} + C_0 \quad (1),$$

unde C_t - consumul în anul t , Y_{t-1} - produsul intern brut în anul $t-1$, C_0 - consumul inițial iar c

– înclinația la limită spre consum, $c = (\sum_{1995}^{2003} C_t / Y_t) / 9$.

Produsul intern brut în anul t se calculează în conformitate cu formula:

$$Y_t = C_t + I_t = cY_{t-1} + C_0 + I_t \quad (2),$$

aici I_t - investițiile brute în anul t , admitem că $I_t = I_0$, $C_0 = \text{const}$.

1. Utilizând datele istorice pentru ultimii ani, estimați coeficientul c - înclinația la limită spre consum.

2. Determinați valoarea I_0 ca valoarea investițiilor nominale în anul de bază.

Datele le puteți găsi pe situl : <http://www.statistica.md>.

Etapa II. Calculați PIB în anul t

6. Estimați produsul intern brut în anul curent t folosind expresia:

$$Y_t = Y_{t_0} c^t + \bar{A} \sum_{l=t_0}^{t-1} c^l$$

aici $\bar{A} = C_0 + I_0$ sunt cheltuielile independente. Consumul inițial se determină ca $C_0 = 0$, investițiile inițiale se presupun să fie egale cu: $I_0 = I_{t_0}$, $t_0 = 2003$ - anul inițial, iar $t_k = 2008$, $t_k = 2012$.

7. Selectând anul 2001 drept an de bază, recalculați Y_t în prețurile anului 2001, presupunând că, începând cu anul 2005, inflația va lua valori din tabelul atașat mai jos, iar în 2004 inflația primește valoare de 7,5%.

Tabelul1. Date istorice

Anii	1995	1996	1997	1998	1999	2000	2001	2002	2003
PIB_{tnomlefec}	6480	7798	8917	9122	12322	16020	19052	22556	27297
Инфляция_t	0,3	0,24	0,12	0,08	0,39	0,31	0,10	0,05	0,12
Дефл. PIB₂₀₀₁									
PIB_{t realefec}									
Anii	2004	2005	2006	2007	2008	2009	2010	2011	2012
Инфляция_t	0,075	0,075	0,075	0,07	0,065	0,06	0,055	0,05	0,05
Дефл. PIB₂₀₀₁									
PIB_{tnominefect}									
PIB_{tpeanprevez}									

Etapa III. Evaluarea creșterii economice în baza modelului Domar și Harrod-Domar

1. Să se calculeze PIB pentru anii 2004-2012 în conformitate cu formula:

$$Y_t = (1 + \sigma \times s)^t Y_{t_0}$$

unde σ sporirea producției pentru o unitate de investiții, iar s este norma acumulărilor, $s = 1 - c$, aici s este înclinația la limită spre consum, Y_{t_0} este produsul intern brut în anul t_0 .

Folosiți valoarea de c , calculată anterior la calcularea s ; calculați $\sigma = \max_t (I_t / Y_t)$ pentru anii 1995-2002.

2. Să se calculeze PIB pentru anii 2004-2012 în conformitate cu formula:

$$Y_t = (1 + \rho_g)^t Y_{t_0}$$

unde ritmul garantat de creștere este $\rho_g = \frac{s}{v - s}$. Aici v este determinat de principiul de

accelerare, adică $I_t = v \times (Y_t - Y_{t-1})$, Y_t este produsul intern brut în anul t , iar Y_{t-1} este produsul intern brut în anul $t - 1$, s este norma acumulărilor, $s = 1 - c$, aici c este înclinația la limită spre consum, Y_{t_0} este produsul intern brut în anul t_0 .

Calculați $v = \max_t I_t / (Y_t - Y_{t-1})$ pentru anii 1995-2002.

Etapa IV. Evaluarea salariului nominal și real în baza funcției de producere Cobb-Douglas.

1. Dat fiind determinați coeficienții funcției de producere $Y_t = K^{1-\alpha} L^\alpha$ în lucrarea precedentă, calculați salariul real ca $w_t = \partial Y_t / \partial L_t = \alpha \times K^{1-\alpha} L^{\alpha-1}$, iar salariul nominal $w_t \text{ nom} = w_t \text{ Defl.PIB}_t$, Defl.PIB_t calculați recent.
2. Păstrând ritmul de creștere al forței de muncă și ritmul de creștere al capitalului, indicat anterior, ritmul de creștere al produsului intern brut prognoza, determinați salariul real și acel nominal prognozate.

Tabelul 2. Date istorice

INDICATORII PRINCIPALI MACROECONOMICI								
	1995	1996	1997	1998	1999	2000	2001	2002
Produsul intern brut (PIB)								
mil.lei	6480	7798	8917	9122	12322	16020	19052	22556
Consumul final:								
mil.lei	5371	7356	8681	9203	11090	16503	19263	23289
Formarea brută de capital:								
mil.lei	1612	1891	2123	2360	2820	3836	4436	4886
Investiții în capital fix:								
mil.lei	844,8	987,4	1202,2	1444,4	1591,8	1759,3	2315,1	2804,2

LUCRAREA DE LABORATOR Nr. 5

LICHIDAREA FENOMENULUI DE ETEROSCEDASTICITATE

Nr.	Prețul Yi (lei)	Volum ul Xi	Yi _{estim.}	u _i	u _i *u _i	ln(u _i *u _i)	ln(Zi)	Yi/Xi
1	4,6	2						
2	4,6	2						
3	4,7	2						
4	5	4						
5	5,2	4						
6	5,4	4						
7	5,6	6						
8	5,8	6						
9	5,9	6						
10	6,3	8						
11	6,5	8						
12	6,8	8						
13	7,4	10						
14	7,6	10						
15	7,1	10						
16	4,8	2						
17	5	2						
18	5,1	2						
19	5,4	4						
20	5,5	4						
21	5,6	4						
22	6,3	6						
23	6,3	6						
24	6,4	6						
25	7,2	8						
26	7,4	8						
27	7,5	8						
28	8,2	10						
29	8,4	10						
30	8,8	10						
Total								
Media								

1. Să se lanseze ecuația de regresie. Să se declare variabila X drept factor de proporționalitate. Să se aplice testul Park pentru a depista fenomenul de eteroscedasticitate.
 2. Să se calculeze $\ln((u_i)^2)$ și $\ln(Z_i)$, d_L și să se lanseze o nouă regresie.
 3. Să se compare t-statistica variabilei Z calculată cu $t_{\text{tabel}}(30-2; 0,05)$.
 - a) dacă $t_{\text{calc}} > t_{\text{tabel}}(30-2; 0,05)$, termenul de eroare este eteroscedastic,
 - b) dacă $t_{\text{calc}} < t_{\text{tabel}}(30-2; 0,05)$, termenul de eroare nu este eteroscedastic.
- Să se efectueze transformările variabilelor $Y^*_i = Y_i/X_i$; $X^*_i = 1/X_i$.
4. Să se estimeze ecuația de regresie transformată. Să se verifice semnificația coeficientului de determinație, coeficientului de corelație, t-statisticile, F-statistica.

Запуск исходного уравнения регрессии

$$Y = b_0 + b_1 \cdot X$$

Запуск преобразованного уравнения регрессии

$$Y^* = Y/X = b_0 \cdot 1/X + b_1$$

$$Y = b_0 + b_1 \cdot X$$

LUCRAREA DE LABORATOR Nr. 6

LICHIDAREA FENOMENULUI DE AUTOCORELARE ÎN SERIE

Anul	Consumul Yt	Venitul Xt	Ytestim	u _t	ut*u _t	ut-ut ₋₁	ut*u _{t-1}	u _{t-1} *u _{t-1}	Y*t	X*t	Y*testim	u* _t	(u _t) ²	u*t-u*t ₋₁
1	84,4	88,0												
2	91,9	94,0												
3	99,2	100,0												
4	104,0	106,0												
5	109,0	110,0												
6	117,8	119,0												
7	122,9	127,0												
8	130,0	135,0												
9	138,7	143,0												
10	149,1	155,0												
11	158,0	167,0												
12	167,5	177,0												
13	177,8	186,0												
14	186,6	197,0												
15	195,7	211,0												
16	208,6	228,0												
17	221,5	239,0												
18	232,1	252,0												
Total														
Media														
DW* ₂	d _U	d _L												
	1,39	1,16												

$$Y_t - r_o * Y_{t-1} = b_0 + b_1 * (X_t - r_o * X_{t-1})$$

$$Y_t = b_0 + b_1 * X_t + r_o * Y_{t-1} - b_1 * r_o * X_{t-1}$$

1. Să se lanseze ecuația de regresie. Să se calculeze statistica $DW = \frac{\sum(u_t - u_{t-1})^2}{\sum(u_t)^2}$. Să se atragă atenția că suma cu termenul întârziat conține cu un termen mai puțin.
2. Să se afle valorile tabelare pentru $d_U = DW(N; k; 0,05)$ și $d_L = DW(N; k; 0,05)$, (N este numărul de observații, k este numărul variabilelor independente) și să se compare cu statistica DW calculată.
 - a) în caz când $DW < d_U$, termenul rezidual este autocorelat,
Să se calculeze $r_o = \frac{\sum(u_t * u_{t-1})}{\sum((u_{t-1}))^2}$,
Să se efectueze transformarea variabilelor conform formulelor $Y^*_1 = (1 - r_o)^{1/2} * Y_1$; $Y^*_t = Y_t - Y_{t-1} * r_o$;
 $X^*_1 = (1 - r_o)^{1/2} * X_1$; $X^*_t = X_t - X_{t-1} * r_o$.
3. Să se estimeze ecuația de regresie transformată.
4. Să se recalculeze statistica $DW = \frac{\sum(u_t - u_{t-1})^2}{\sum(u_t)^2}$. Să se execute etapa 2.
 - b) în caz când $DW < d_U$, termenul rezidual nu este autocorelat.

LUCRAREA DE LABORATOR Nr. 7
LICHIDAREA FENOMENULUI DE MULTICOLINEARITATE

Nr.	Y	X ₁	X ₂	X ₃	X ₄	SumXi
1	74,3	1,0	29,0	15,0	52,0	
2	72,5	1,0	31,0	22,0	44,0	
3	83,8	1,0	40,0	23,0	34,0	
4	93,1	2,0	54,0	18,0	22,0	
5	102,7	3,0	71,0	17,0	6,0	
6	78,5	7,0	26,0	6,0	60,0	
7	95,9	7,0	52,0	6,0	33,0	
8	109,4	10,0	68,0	8,0	12,0	
9	104,3	11,0	56,0	8,0	20,0	
10	87,6	11,0	31,0	8,0	47,0	
11	109,2	11,0	55,0	9,0	22,0	
12	113,3	11,0	66,0	9,0	12,0	
13	115,9	21,0	47,0	4,0	26,0	

1. Să se încredințeze că ecuația de regresie este supusă fenomenului de multicolinearitate. Să se calculeze suma celor patru variabile pentru fiecare observație.
2. Să se lanseze ecuația de regresie. Să se calculeze t- statisticile $b_i = b_i / \sigma(BI)$.
3. Să se determine valoarea minimă F_i , egală cu $F_L = \min(F_i)$, și să se compare cu valoarea tabelară $F(1;n-m-1;\alpha)$ unde: n este numărul observațiilor, m este numărul variabilelor independente α este nivelul de semnificație de 0,05).
4. Să se calculeze valoarea criteriului particular $F_i = (t_{bi})^2$. Sunt posibile variante:
 - a) $F_L < F_0$, variabila independentă se exclude din ecuație. Să se treacă la etapă 5.
 - b) $F_L > F_0$, în acest caz modelul obținut este acel corect.
5. Să se estimeze ecuația de regresie în funcție de variabilele independente păstrate. Să se îndeplinească etapele 1-4.

Y	X ₁	X ₂	X ₄	Y = b₀ + b₁*X₁ + b₂*X₂ + b₃*X₃ - b₄*X₄					
74,3	1,0	29,0	52,0	sigma(b_i)					
72,5	1,0	31,0	44,0	t _{bi}					
83,8	1,0	40,0	34,0	F _i	i =	F _i	<	F _{itabel}	5,32
93,1	2,0	54,0	22,0	FL					
102,7	3,0	71,0	6,0						
78,5	7,0	26,0	60,0						
95,9	7,0	52,0	33,0						
109,4	10,0	68,0	12,0	Y = b₀ + b₁*X₁ + b₂*X₂ - b₃*X₄					
104,3	11,0	56,0	20,0	sigma(b_i[^])					
87,6	11,0	31,0	47,0	t _{bi}					
109,2	11,0	55,0	22,0	F _i	i =	F _i	<	F _{itabel}	5,12
113,3	11,0	66,0	12,0	FL					
115,9	21,0	47,0	26,0						
				Y = b₀ + b₁*X₁ + b₂*X₂					
				sigma(b_i[^])					
				t _{bi}					
				F _i		F _i	<	F _{itabel}	4,96
				FL					