2. LINEARNI REGRESIONI MODELI

2.1. Uvod

Osnovni problem u kvantitativnom opisivanju ekonomskih pojava je, kao što je istaknuto u delu 1., izbor promenljivih koje su bitne za željeni opis i njihovo povezivanje u obliku matematičke relacije. U opštem slučaju, relacije koje povezuju jednu promenljivu Y koja se naziva zavisna (ili merena, posmatrana, ona koja se objašnjava) promenljiva, sa u principu većim brojem nezavisnih (ili kontrolisanih, objašnjavajućih) promenljivih X_i , i = 1,2,...,k, putem relacije:

$$Y = f(X_i, \beta_i) \qquad (2.1)$$

gde su β_i parametri modela, se nazivaju regresioni modeli.

Najjednostavniji slučaj, koji će se prvo razmatrati, je relacija koja povezuje jednu zavisnu i jednu nezavisnu promenljivu. Broj parametara u relacijama ovakvog tipa je obično dva i ovde će se označavati sa α i β. Da bi se mogla definisati ovakva relacija potreban je:

a) skup $\{X_j, Y_j\}$, j = 1,2,...,N parova koji čine observacije vrednosti promenljivih, i to N parova ukupno;

b) matematički oblik relacije koja vezuje zavisnu i nezavisnu promenljivu i parametre α i β:

$$Y = f(X, \alpha, \beta) \qquad (2.2)$$

gde veza može biti linearna ili nelinearna bilo po promenljivama bilo po parametrima; i

c) statistička ocena parametara α i β koji se pojavljuju u relaciji (2.2).

Zadatak ekonometrije je da statističkim metodama odredi ocenu parametara α i β u relaciji (2.2), odnosno u opštem slučaju parametara β_i iz relacije (2.1). Isto tako, zadatak ekonometrije je da izvrši testiranje relacije (2.1) ili (2.2) sa ocenjenim vrednostima parametara u odnosu na realne podatke i na taj način doprinese bližem razumevanju posmatrane ekonomske pojave.

Kad god se želi oceniti jedna promenljiva Y u funkciji druge promenljive X, faktički se određuje statistička veličina E(Y|X), koja označava uslovno matematičko očekivanje za promenljivu Y, a za datu vrednost promenljive X tj.:

$$E(Y|X) = \int_{-\infty}^{\infty} y \cdot f(y|x)dy \qquad (2.3)$$

gde $f(y|x)$ označava raspodelu uslovnih verovatnoća za y pri datoj vrednosti x, za slučaj kontinualne raspodele, i:

$$E(Y|X) = \sum_{j=1}^{N} y_j f(y_j|x_i) \qquad (2.4)$$

za slučaj diskretne raspodele.

Uslovno matematičko očekivanje (2.3) ili (2.4) je funkcija slučajne promenljive X i ta funkcija se naziva regresija (ili regresiona kriva).

Na taj način, zavisna promenljiva Y se može izraziti kao:

$$Y = E(Y|X) + \varepsilon \qquad (2.5)$$

gde ε označava slučajnu promenljivu koja obuhvata odstupanja koja nastaju zbog:

a) netačnosti u specifikaciji izraza za $E(Y|X)$;

b) grešaka u određivanju observacija promenljivih Y i X; i

c) slučajnih elemenata svojstvenih svim pojavama sa prisutnim subjektivnim faktorom.

Da bi se regresija (2.5) mogla praktično koristiti za određivanje ocene promenljive Y, a za datu vrednost promenljive X, potrebno je uvesti neke pretpostavke o prirodi slučajne promenljive ε. Ove pretpostavke su ključne u određivanju ocena parametara regresije, kako sa stanovišta metode koja se koristi za ocenjivanje parametara, tako i sa stanovišta njihove tačnosti. Pošto se slučajna promenljiva ε ne može direktno meriti pretpostavljaju se ili oblik raspodele po kojoj se ponaša ε ili samo neki od karakterističnih parametara populacije, kao što su srednja vrednost, varijansa, kovarijansa i slično. Tačnost učinjenih pretpostavki o karakteru slučajne promenljive ε se proverava na osnovu slaganja vrednosti Y dobijenih regresijom, sa vrednostima observacija za promenljivu Y.

Regresioni modeli kod kojih se uslovno matematičko očekivanje (2.3 i 2.4) može izraziti kao linearna funkcija promenljivih X_i i parametara tj.:

$$E(Y|X) = \alpha + \beta X \qquad (2.6)$$

za slučaj jedne nezavisne promenljive, odnosno:

$$E(Y|X_1, X_2, \ldots X_k) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad (2.7)$$

za slučaj k promenljivih, ili:

$$Y = \alpha + \beta X + \varepsilon \qquad (2.8)$$

odnosno:

$$Y_j = b_1 X_{1j} + b_2 X_{2j} + \ldots + b_{kj} X_{kj} + E_j \qquad (2.9)$$

se nazivaju linearni regresioni modeli.

Pored svoje analitičke jednostavnosti, linearni regresioni modeli su pogodni za opisivanje ekonomskih pojava i iz sledećih razloga: (TINBERGEN, 1940)

1) Dobro je poznata matematčka istina da se skoro svaka funkcija može aproksimirati linearnom u dovoljno malom intervalu. Ova propozicija ne važi jedino za izuzetne funkcije koje obično i nisu od praktičnog interesa.

2) Nije redak slučaj da linearna zavisnost i stvarno postoji u ponašanju nekih pojava.

3) Takođe, sasvim je prirodno početi studiju neke pojave čineći najjednostavniju pretpostavku koja je saglasna sa opštom teorijom.

4) Osim navedenog u prilog opravdanosti linearnih modela govori i činjenica da je zajednička reakcija velikog broja pojedinaca linearnija od reakcija jednog pojedinca.

Isto tako, kao što će se pokazati kasnije, izvesna klasa nelinearnih modela se može transformisati u linearni regresioni model.

Od svih metoda koje mogu koristiti za ocenu parametara linearnih regresionih modela tipa (2.8) ili (2.9), metoda najmanjih kvadrata se najčešće koristi i ovde će se i najviše razmatrati.

Pretpostavke o karakteru ε

Metod najmanjih kvadrata se bazira na sledećem skupu pretpostavki o karakteru slučajnih odstupanja ε:

1) Matematičko očekivanje ili srednja vrednost odstupanja E_i je jednaka nuli, tj.:

$$E(\varepsilon_i) = 0, \quad i = 1,2,\ldots N$$

2) Varijansa odstupanja ε_i je konstantna, tj homoskedastična:

$$V(\varepsilon_i) = S^2, \qquad i = 1,2,\ldots N$$

Ako V(ε) nije konstantna, tad imamo heteroskedastičnost.

3) Kovarijansa odstupanja ε_i i ε_j je jednaka nuli, tj. greške nisu korelisane (ili autokorelisane):

$$E(\varepsilon_i \varepsilon_j) = 0, \quad i=1,...,N;\ j=1,\ldots,N;\ i \neq j$$

4) Odstupanje ε_i nije korelisano sa promenljivom X_j tj.:

$$E(\varepsilon_i X_j) = 0, \quad i = 1,2,...,N; j = 1,2,...,k$$

Ova pretpostavka tvrdi da promenljiva X nije slučajna promenljiva. Metod najamanjih kvadrata se sastoji u minimizaciji sume kvadrata:

$$Q = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \qquad (2.10)$$

gde $\hat{Y}_i$ označava ocenjenu vrednost promenljive Y_i, koja se dobija na osnovu regresionog modela:

$$\hat{Y} = a + bX \qquad (2.11)$$

za slučaj jedne nezavisne promenljive, ili:

$$\hat{Y}_i = b_1 X_1 + b_2 X_2 + ... + b_k X_k \qquad i = 1,2,...,N \qquad (2.12)$$

za slučaj više nezavisnih promenljivih.

Veličine a,b odnosno $b_1,b_2,...b_k$ označavaju ocene parametara α, β odnosno β_1, β_2 , ... , β_k , respektivno dobijene metodom najmanjih kvadrata, tj. minimizacijom sume kvadrata (2.10).

Veličina:

$$e_i = Y_i - \hat{Y}_i \qquad (2.13)$$

označava ocenu slučajnih odstupanja ε_i, ili drugim rečima razliku između stvarne vrednosti promenljive Y_i i vrednosti ocenjene regresijom $\hat{Y}_i$, i ponekad se naziva rezidualom.

Vrednost metode najmanjih kvadrata u ocenjivanju parametara linearnih regresionih modela leži u činjenici da dobijene ocene imaju osobine nepristrasnosti, minimalne varijanse i konzistentnosti. Takođe, uz dodatnu pretpostavku da su slučajna odstupanja ε data normalnom raspodelom, ocene dobijene metodom najmanjih kvadrata se poklapaju sa ocenama dobijenim metodom maksimalne verodostojnosti, odnosno poseduju i ostale "lepe" osobine ovih ocena.

U daljem tekstu će se razmotriti detaljnije, korišćenje metode najmanjih kvadrata prvo za slučaj linearnog regresionog modela sa jednom nezavisnom promenljivom, a zatim i za slučaj sa više nezavisnih promenljivih.

2.2. Linearni regresioni model sa dve promenljive

Najjednostavniji slučaj linearnih regresionih modela je model sa dve promenljive, tj. jednom zavisnom i jednom nezavisnom promenljivom. U tom slučaju uslovno očekivanje $E(Y|X)$ ima oblik:

$$E(Y|X) = \alpha + \beta X \qquad (2.14)$$

odnosno zavisna promenljiva se izražava relacijom:

$$Y = \alpha + \beta X + \varepsilon \qquad (2.15)$$

Određivanje regresije Y na X se svodi na nalaženje ocena a i b, parametara α i β, kao i reziduala e_j koji predstavljaju ocenu odgovarajućih vrednosti slučajnih odstupanja ε_j u datom uzorku koji čine parovi observacija $\{X_j, Y_j\}$, j = 1,2,...,N gde sa N kao i dosad označavamo ukupni broj observacija uzorka na osnovu kojeg ocenjujemo regresioni model.

Pre nego što se pređe na primenu linearnog regresionog modela sa dve promenljive preporučljivo je konstruisati dijagram zavisnosti (raspršenosti). Dijagram zavisnosti se konstruiše u pravouglom koordinatnom sistemu, ucrtavanjem svih parova podataka (X_i, Y_i), i = 1,2,...,N, pri čemu se na apcisu nanose vrednosti za nezavisnu promenljivu X, a na ordinatu jedinice zavisne promenljive Y.

Iz dijagrama zavisnosti se može se sagledati:

- Da li između zavisne i nezavisne promenljive postoji veza;
- Ako veza postoji, da li je pravolinijska ili krivolinijska;
- Ako veza postoji i pravolinijska je, da li je direktna ili inverzna.

Primer: Dati su podaci o poslovanju jednog preduzeća koji se odnose na ostvareni profit i izdatke za reklamu u prethodnih 10 godina.

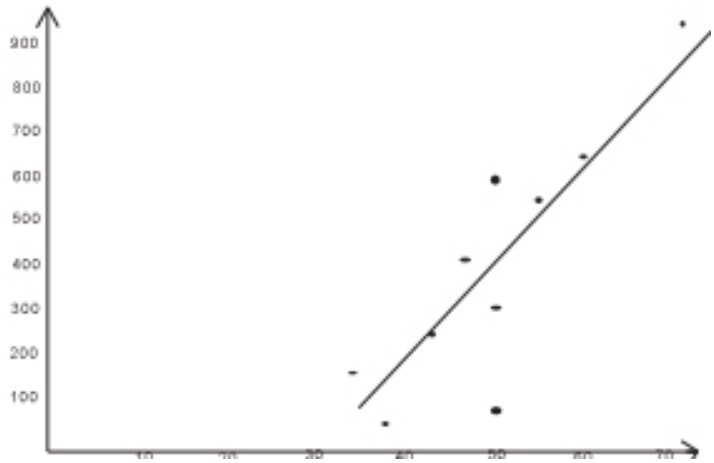
Godina	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
Profit	325	444	268	605	569	190	946	75	100	661
Izdaci za reklamu	51	47	44	50	56	45	71	38	52	61

Nacrtati dijagram zavisnosti profita u odnosu na izdatke za reklamu i na osnovu njega utvrditi eventualno postojanje, oblik i jačinu veze između promenljivih.

Rešenje:

X - Izdaci za reklamu (nezavisna promenljiva);

Y - Profit (zavisna promenljiva);



2.2.1. Ocenjivanje parametara metodom najmanjih kvadrata

U ovom slučaju, suma kvadrata (2.10) ima oblik:

$$Q = \sum_{i=1}^{N} (Y_i - a - bX_i)^2 \qquad (2.16)$$

Minimizacija sume kvadrata (2.16) se vrši u pogledu na parametre a i b i to izjednačavanjem odgovarajućih parcijalnih izvoda sa nulom, tj.:

$$\frac{\partial Q}{\partial a} = \sum_{i=1}^{N} 2(Y_i - a - bX_i)(-1) = 2(Na + b\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} Y_i) = 0$$

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{N} 2(Y_i - a - bX_i)(-X_i) = 2(a\sum_{i=1}^{N} X_i + b\sum_{i=1}^{N} X_i^2 - \sum_{i=1}^{N} X_i Y_i) = 0 \qquad (2.17)$$

odnosno:

$$\sum_{i=1}^{N} Y_i = Na + b\sum_{i=1}^{N} X_i$$
$$\sum_{i=1}^{N} X_i Y_i = a\sum_{i=1}^{N} X_i + b\sum_{i=1}^{N} X_i^2 \qquad (2.18)$$

Sistem jednačina (2.18) se naziva normalne jednačine i njegovim rešavanjem se dolazi do ocena a i b. Tako se za b dobija:

$$b = \frac{N\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2} \; ili$$
$$b = \frac{\frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \overline{XY}}{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \overline{X}^2} \qquad (2.19)$$

Ako definišemo:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}, \quad \overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} \qquad (2.20)$$

tad na osnovu prve od normalnih jednačina (2.18) sledi:

$$\overline{Y} = a + b \cdot \overline{X} \qquad (2.21)$$

odnosno, regresiona prava prolazi kroz tačku $(\overline{X}, \overline{Y})$ određenu srednjim vrednostima observacija X i Y respektivno.

Jednačina (2.21) takođe služi za određivanje ocene a tj.:

$$a = \overline{Y} - b\overline{X} \qquad (2.22)$$

Ocenjena regresiona prava je:

$$\hat{Y} = a + bX \qquad (2.23)$$

Uvodeći smenu promenljivih:

$$x = X - \overline{X}; \qquad y = Y - \overline{Y}; \qquad \hat{y} = \hat{Y} - \overline{Y} \qquad (2.24)$$

i oduzimanjem (2.22) od (2.23) dobijamo:

$$\hat{y} = bx \tag{2.25}$$

pa se do ocene b može doći minimizacijom sume kvadrata:

$$\begin{aligned} Q &= \sum_{i=1}^{N} (y_i - bx_i)^2 \\ \frac{\partial Q}{\partial b} &= 2\sum_{i=1}^{N} (y_i - bx_i)(-x_i) = 2\sum_{i=1}^{N} (x_i y_i - bx_i^2) = 0 \end{aligned} \tag{2.26}$$

tj.

$$b = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \tag{2.27}$$

U sledećem će se pokazati da su izvedene ocene za a i b metodom najmanjih kvadrata najbolje nepristrasne ocene odgovarajućih parametara α i β linearnog regresionog modela (2.14).

Dokaz da je b nepristrasna ocena :

Prvo pokažimo da je b nepristrasna linearna ocena parametra β. Naime, iz relacije (2.27) sledi:

$$b = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} = \frac{\sum_{i=1}^{N} x_i Y_i}{\sum_{i=1}^{N} x_i^2} - \frac{\overline{Y}\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2} = \frac{\sum_{i=1}^{N} x_i Y_i}{\sum_{i=1}^{N} x_i^2}$$

odnosno:

$$b = \sum_{i=1}^{N} w_i Y_i \tag{2.28}$$

jer je

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} (X_i - \overline{X}) = \sum_{i=1}^{N} X_i - N\frac{\sum_{i=1}^{N} X_i}{N} = 0 \tag{2.29}$$

gde w_i označava

$$w_i = \frac{x_i}{\sum_{i=1}^{N} x_i^2} \tag{2.30}$$

Relacija (2.28) pokazuje da je b linearna kombinacija observacija Y_i . Na osnovu izloženog sledi:

$$b = \sum_{i=1}^{N} w_i Y_i = \sum_{i=1}^{N} w_i (\alpha + \beta X_i + \varepsilon_i) = \alpha \sum_{i=1}^{N} w_i + \beta \sum_{i=1}^{N} w_i X_i + \sum_{i=1}^{N} w_i \varepsilon_i$$

ili:

$$b = \beta + \sum_{i=1}^{N} w_i \varepsilon_i \tag{2.31}$$

jer je, na osnovu (2.29), (2.30) i (2.24):

$$\sum_{i=1}^{N} w_i = 0; \qquad \sum_{i=1}^{N} w_i X_i = \sum_{i=1}^{N} w_i X_i - \overline{X} \sum_{i=1}^{N} w_i = \sum_{i=1}^{N} w_i (X_i - \overline{X}) = \sum_{i=1}^{N} w_i x_i = \frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} x_i^2} = 1$$

prema tome, matematičko očekivanje dobijene ocene b je:

$$E(b) = E(\beta + \sum_{i=1}^{N} w_i \varepsilon_i) = E(\beta) + E(\sum_{i=1}^{N} w_i \varepsilon_i) = E(\beta) + \sum_{i=1}^{N} w_i * E(\varepsilon_i)$$

odnosno, kako je po pretpostavci o karakteru slučajnih odstupanja $E(\varepsilon_i) = 0$ sledi:

$$E(b) = \beta \tag{2.32}$$

Na osnovu relacije (2.32) se zaključuje da je b nepristrasna ocena parametra β, ili drugim rečima da je raspodela za b centrirana na vrednosti β, te se zbog toga b zove i centrirana ocena.

Na sličan način se dokazuje da je i:

$$E(a) = \alpha \tag{2.33}$$

tj. da je a nepristrasna linearna ocena za α.

Dokaz da je b najbolja ocena

Da su ocene a i b parametara α i β respektivno, dobijene metodom najmanjih kvadrata i najbolje nepristrasne ocene, u smislu da od svih mogućih nepristrasnih ocena imaju minimalnu varijansu, pokazaće se na primeru parametra b. Naime, varijansa b je:

$$V(b) = E(b-\beta)^2$$

Na osnovu (2.31) sledi:

$$V(b) = E((\sum_{i=1}^{N} w_i \varepsilon_i)^2)$$

odnosno, na osnovu pretpostavki:

$$V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_\varepsilon^2$$

i,

$$E(\varepsilon_i \varepsilon_j) = 0, \qquad i \neq j$$

sledi:

$$V(b) = E(\sum_{i=1}^{N} w_i^2 \varepsilon_i^2 + 2\sum_{i=1}^{N}\sum_{i\neq j}(w_i\varepsilon_i)(w_j\varepsilon_j)) = \sum_{i=1}^{N} E(w_i^2\varepsilon_i^2) + 2\sum_{i=1}^{N}\sum_{i\neq j}(w_i w_i)(\varepsilon_j\varepsilon_j) =$$

$$= \sum_{i=1}^{N} E(w_i^2)E(\varepsilon_i^2) + 2\sum_{i=1}^{2} E(w_i w_j)E(\varepsilon_i\varepsilon_j) = \sum_{i=1}^{N} w_i^2 E(\varepsilon_i^2) = \sigma_\varepsilon^2 \sum_{i=1}^{N} w_i^2$$

Kako je,

$$\sum_{i=1}^{N} w_i^2 = \frac{1}{\sum_{i=1}^{N} x_i^2}$$

to se za varijansu ocene b dobija:

$$V(b) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{N} x_i^2} \qquad (2.34)$$

Da bi pokazali da je varijansa V(b) i minimalna varijansa, posmatraće se proizvoljna ocena b' koja je isto nepristrasna odnosno linearna ocena u smislu da se može izraziti kao:

$$b' = \sum_{i=1}^{N} c_i Y_i$$

gde su c_i konstante koje se mogu izraziti kao:

$$c_i = w_i + d_i \qquad (2.35)$$

s tim što je w_i konstanta koja se sračunava na osnovu (2.30) a d_i je proizvoljna konstanta.

$$b' = \sum_{i=1}^{N} c_i(\alpha + \beta X_i + \varepsilon_i) = \alpha\sum_{i=1}^{N} c_i + \beta\sum_{i=1}^{N} c_i X_i + \sum_{i=1}^{N} c_i\varepsilon_i$$

Slično dosadašnjem izvođenju se pokazuje da je:

$$E(b©) = \alpha\sum_{i=1}^{N} c_i + \beta\sum_{i=1}^{N} c_i X_i$$

Prema tome, da bi b' bilo nepristrasna ocena parametra b, tj. da je:

$$E(b') = \beta$$

mora da je:

$$\sum_{i=1}^{N} c_i = 0 \qquad \text{i} \qquad \sum_{i=1}^{N} c_i X_i = 1$$

što je jedino moguće, na osnovu (2.35) i definiciji za w_i (2.30), ukoliko važi:

$$\sum_{i=1}^{N} d_i = 0 \qquad \text{i} \qquad \sum_{i=1}^{N} d_i X_i = 0$$

jer je:

$$\sum_{i=1}^{N} c_i = \sum_{i=1}^{N}(w_i + d_i) = \sum_{i=1}^{N} w_i + \sum_{i=1}^{N} d_i = 0$$

$$\sum_{i=1}^{N} c_i x_i = \sum_{i=1}^{N} w_i x_i + \sum_{i=1}^{N} d_i x_i = 1$$

Varijansa ove proizvoljne ocene b' je:

$$V(b') = E((\sum_{i=1}^{N} c_i\varepsilon_i)^2) = \sigma_\varepsilon^2\sum_{i=1}^{N} c_i^2$$

$$= \sigma_\varepsilon^2\sum_{i=1}^{N}(w_i + d_i)_i^2 = \sigma_\varepsilon^2\sum_{i=1}^{N} w_i^2 + \sigma_\varepsilon^2\sum_{i=1}^{N} d_i^2 + 2\sigma_\varepsilon^2\sum_{i=1}^{N} w_i d_i$$

Na osnovu (2.35) i činjenice da je:

$$\sum_{i=1}^{N} w_i d_i = 0$$

sledi:

$$V(b') = V(b) + \sigma_\varepsilon^2 \sum_{i=1}^{N} d_i^2$$

Kako je suma kvadrata

$$\sum_{i=1}^{N} d_i^2$$

sigurno nenegativna, sledi da je uvek:

$$V(b') \geq V(b)$$

odnosno da je ocena b sa najmanjom varijansom od svih mogućih linearnih nepristrasnih ocena.

Slično, kao i u slučaju varijanse ocene b dobija se i varijansa ocene a:

$$V(a) = \frac{\sum_{i=1}^{N} X_i^2}{N \sum_{i=1}^{N} x_i^2} \sigma_\varepsilon^2 \qquad (2.36)$$

Takođe, jednostavnim algebarskim transformacijama za kovarijansu ocena a i b dobija se:

$$E((a-\alpha)(b-\beta)) = \frac{-\overline{X}}{\sum_{i=1}^{N} x_i^2} \sigma_\varepsilon^2 \qquad (2.37)$$

I za varijansu ocene a (2.36) se pokazuje da je minimalna u odnosu na sve ostale moguće lnearne nepristrasne ocene za parametar a.

Kako u izrazima za varijanse ocena a i b, kao i odgovarajućem izrazu za kovarijansu, figuriše izraz za varijansu člana koji opisuje slučajna odstupanja ε tj.:

$$V(\varepsilon_i) = \sigma_\varepsilon^2$$

potrebno je izvršiti njegovu ocenu. S obzirom da slučajna odstupanja ne podležu direktnom merenju odnosno nije moguće konstruisati skup observacija $\{\varepsilon_i\}$, ocenjivanje $V(\varepsilon_i)$ odnosno σ_ε^2 se može izvršiti kao:

$$\sigma_e^2 = \frac{\sum_{i=1}^{N} e_i^2}{N-2} \qquad (2.38)$$

gde je:

$$\hat{Y}_0 = X_0 B$$

a σ_e^2 označava ocenu varijanse σ_ε^2 .

$$\sigma_e^2 = \frac{\sum_{i=1}^{N} Y_i^2 - a\sum_{i=1}^{N} Y_i - b\sum_{i=1}^{N} X_i Y_i}{N-2}$$

Može se dobiti:

$$\sigma_e = \sqrt{\sigma_e^2}$$

Ocena (2.38) za σ_ε^2 se dobija na osnovu činjenice da veličine e_i nisu linearno nezavisne kao što su to veličine ε_i . Naime, veličine e_i su povezane sa 2 normalne jednačine tako da imaju samo (N-2) stepena slobode. Drugim rečima, uz pomoć (N-2) vrednosti e_i i 2 normalne jednačine moguće je sračunati preostale dve vrednosti za e_i , od ukupno N.

Primer 2.2.1. Za ilustraciju primene linearne regresije sa jednom zavisnom i jednom nezavisnom promenljivom razmotrimo slučaj nekog proizvoda Y u zavisnosti od dohotka X sa parovima observacija Y_i , X_i datim u sledećoj tabeli sa N = 12:

Y	236	254	267	281	290	311	325	335	355	375	401	431
X	257	275	293	309	319	337	350	364	385	405	437	469

Na osnovu dobijenih podataka sledi:

$$\sum_{i=1}^{N} Y_i = 3861; \sum_{i=1}^{N} X_i = 4200; \sum_{i=1}^{N} X_i^2 = 1516510; \sum_{i=1}^{N} X_i Y_i = 1394495$$

odnosno:

$$\overline{X} = 350; \overline{Y} = 321,75$$

Pa se za ocene parametara dobija:

b = 0,9277 iz (2.19) $\qquad a = \overline{Y} - b\overline{X} = -3,0$

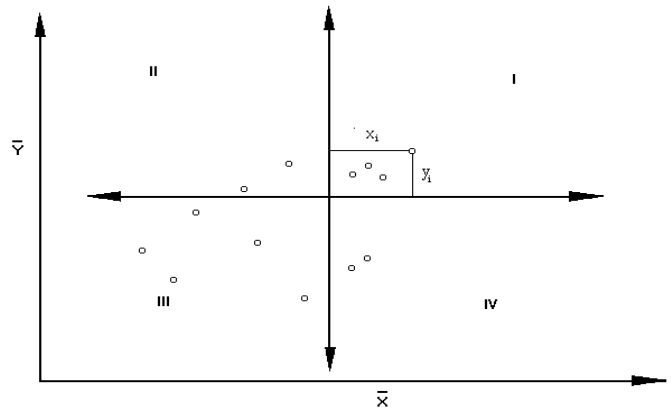
pa je ocenjena regresiona prava:

$$Y = -3,0 + 0,9277X$$

2.2.2. Koeficijenat korelacije i determinacije

Iz dosadašnje diskusije se moglo zaključiti da je regresiona analiza moćno sredstvo za studiju zavisnosti jedne promenljive od druge (ili više drugih). Međutim, u slučajevima kad se ne može utvrditi striktna zavisnost jedne promenljive od druge a ipak postoji nekakva veza između njih, u smislu da im se vrednosti promenljivih povezuju na neki način, odnosno kako se to kaže da su vrednosti korelisane, kao stepen korelacije odnosno povezanosti promenljivih koristi se takozvani *koeficijenat korelacije*. Prema tome, regresiona analiza daje matematičku funkciju koja opisuje zavisnost dveju promenljivih a korelaciona analiza daje jedan broj, koeficijenat korelacije, koji svojom veličinom određuju meru te zavisnosti. Očigledno je dakle, da regresiona analiza pruža više informacija o ponašanju promenljivih i da se na osnovu njenih rezultata može zaključivati i o koeficijentu korelacije dok obrnuto ne važi. U sledećem će se dati veza između koeficijenta korelacije i parametra regresije.

Pojam korelacije dveju promenljivih ilustrujmo na sledećem grafiku.



Sl. 2.3.

Kao mera stepena korelacije promenljivih Y i X može se uzeti suma:

$$\sum_{i=1}^{N} x_i y_i$$

jer ukoliko je ona pozitivna većina tačaka se nalazi u I i III kvadrantu sa Sl. 2.3., ukoliko je negativna tad je većina tačaka u II i IV kvadrantu i ukoliko je bliska nuli tad su tačke ravnomerno raspoređene po svim kvadrantima. Međutim, numerička vrednost gornje sume zavisi od broja N tačaka kao i jedinica u kojima se mere vrednosti promenljivih, te

je u izvesnom smislu proizvoljna i direktno nepogodna za meru korelacije. Zbog toga se za meru korelacije uvodi takozvani Pearson-ov koeficijenat korelacije:

$$r = \frac{\sum_{i=1}^{N} x_i y_i}{N s_x s_y} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{N}}{\sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}}\sqrt{\frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N}}} \qquad (2.47)$$

gde su:

$$s_x = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{N} y_i^2}{N}}$$

ili alternativno:

$$r = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2}} \qquad (2.48)$$

Na osnovu izraza (2.27) za parametar b linearna regresija i izraza (2.48) odnosno (2.47), dobija se:

$$b = r\frac{s_y}{s_x} \qquad (2.49)$$

jer je:

$$b = \frac{s_{XY}}{s_X^2} = \frac{s_{XY}}{s_X s_Y}\frac{s_Y}{s_X} = r\frac{s_y}{s_x}$$

Iz definicije regresione prave sledi da je:

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

dakle

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

ili ako kvadriramo i sumiramo:

$$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} \hat{y}_i^2 + \sum_{i=1}^{N} e_i^2 + 2\sum_{i=1}^{N} \hat{y}_i e_i$$

i kako je, na osnovu (2.25):

$$\sum_{i=1}^{N} \hat{y}_i e_i = 0$$

dobijamo:

$$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} \hat{y}_i^2 + \sum_{i=1}^{N} e_i^2 \qquad (2.50)$$

Iz relacije (2.50) se može zaključiti da se ukupna varijacija vrednosti promenljive Y oko srednje vrednosti $\overline{Y}$ može podeliti u dve komponente. Prva komponenta opisuje varijacije ocenjenih vrednosti $\hat{Y}$ oko njihove srednje vrednosti $\overline{\hat{Y}} = \overline{Y}$. Ova komponenta se označava kao "objašnjena" linearnim uticajem promenljive X. Druga komponenta je takozvana rezidualna ili "neobjašnjena" varijacija Y koja se pripisuje slučajnim odstupanjima.

Odnos "objašnjenog" dela i ukupne varijacije Y se određuje kao:

$$\frac{\sum_{i=1}^{N} \hat{y}_i^2}{\sum_{i=1}^{N} y_i^2} = b^2 \frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} y_i^2} = r^2 \qquad (2.51)$$

i naziva se koeficijenat determinacije i ustvari predstavlja kvadrat koeficijenta korelacije.

$$r^2 = b^2 \frac{\sum_{i=1}^{N} X_i^2 - N\overline{X}^2}{\sum_{i=1}^{N} Y_i^2 - N\overline{Y}^2}$$

$$r = \sqrt{r^2}$$

Iz relacije (2.50) i (2.51) sledi:

$$r^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} y_i^2}$$

dakle,

$$r^2 = \frac{OV}{UV} = 1 - \frac{NV}{UV}$$

gde je OV-objašnjeni varijalibilitet, NV-neobjašnjeni varijabilitet, UV-ukupan varijabilitet;

odnosno zaključuje se da je maksimalna vrednost koeficijenta determinacije jednaka 1 odnosno da se koeficijenat korelacije kreće u granicama:

$$-1 \le r \le +1$$

Na osnovu definicije koeficijenta determinacije vidi se da on predstavlja proporciju varijacije promenljive Y objašnjenih regresionom pravom. Tako na primer, na osnovu podataka iz primera 2.2.1. sračunavamo:

$$r = \frac{3595,42}{\sqrt{3875,83 \cdot 3339,02}} = 0,9994$$

odnosno koeficijenat determinacije ima vrednost:

r = 0,9988

ili drugim rečima, regresiona prava određena u primeru 2.2.1. opisuje (objašnjava) 99,88% varijacija datih podataka Y .

2.2.3. Statistički testovi

Po dobijanju statističkih ocena parametara regresije potrebno je utvrditi u kojoj meri ocenjena regresiona prava odgovara stvarnim podacima. Testiranje pouzdanosti regresije se obavlja na osnovu tri tipa informacija koje određuju efikasnost regresionog modela.

1) A priori informacije čine teorijska ili, iskustvena znanja koja se poseduju o datoj pojavi koja se objašnjava regresionim modelom. Ekonomska teorija je glavni izvor ovih znanja. Ove informacije se odnose pre svega na znak i red veličine parametara. Na primer, ukoliko je poznato da pozitivne promene nezavisne promenljive uvek dovode do pozitivnih promena zavisne promenljive, a regresionom analizom se dobija negativna vrednost parametra b, očigledno je da dobijena regresiona prava loše objašnjava zavisnost promenljivih.

2) Direktnim poređenjem stvarnih vrednosti promenljive Y sa ocenjenim vrednostima $\hat{Y}_0 = a + b\hat{X}_0$ može se direktno zaključivati o kvalitetu regresione prave.

3) Takođe je moguće koristiti razne statističke testove koji određuju intervale poverenja parametara kao i definišu značajnost određenih hipoteza. U ovom delu ćemo se pozabaviti statističkim testovima kvaliteta regresije.

Da bi se mogli sprovesti statistički testovi potrebno je pretpostaviti normalnost raspodele slučajanih odstupanja ε čime se imlicira i normalnost raspodele regresionih parametara.

Prema tome, u slučaju parametra a imamo da je on raspodeljen kao:

$$a: \quad N(\alpha, V(a)) \tag{2.52}$$

gde je varijansa V(a) data izrazom (2.36).

Onda je:

$$z = \frac{(a-\alpha)}{\sqrt{V(a)}} = \frac{a-\alpha}{\sqrt{\sigma_\varepsilon^2 \dfrac{\sum_{i=1}^{N} X_i^2}{N\sum_{i=1}^{N} x_i^2}}} \tag{2.53}$$

raspodeljeno kao:

z : N(0,1)

Međutim, kako u izrazu za V(a) figuriše nepoznata varijansa σ_ε^2 uvodi se promenljiva:

$$v^2 = \frac{(N-2)\sigma_e^2}{\sigma_\varepsilon^2} \tag{2.54}$$

koja ima χ^2 raspodelu sa (N-2) stepena slobode tj.:

$$v^2 : \chi^2(N-2)$$

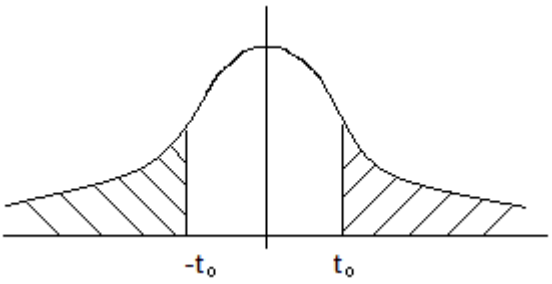
pa promenljiva:

$$t = \frac{(a-\alpha)}{\sigma_e^2}\sqrt{\frac{N\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} X_i^2}} \qquad (2.55)$$

ima Student - ovu t raspodelu sa (N-2) stepena slobode, tj.:

t : t(N-2)

$$t = \frac{b}{\sqrt{V(b)}} : t_{N-2} \qquad V(b) = \frac{\sigma_e^2}{\sum_{i=1}^{N} X_i^2 - N\overline{X}} \qquad P\{|t| > t_0\} = \alpha$$



Napominje se da se ovim transformacijama eliminisala nepoznata varijansa σ_ε^2 i dobila test funkcija koja zavisi jedino od observacija X i Y i hipotetičke vrednosti parametra α.

Hipoteza da je vrednost parametra α=0 se proverava sračunavanjem vrednosti t_0 iz izraza (2.55) sa α=0. Pretpostavljajući nivo značajnosti η i nalaženjem odgovarajuće vrednosti $t_{\eta-2}$ iz tabele t raspodele sa (N-2) stepena slobode, dvostrani test hipoteze se formuliše kao:

a) ako je: $|t| \geq t_0$, hipoteza se odbacuje

b) ako je: $|t| < t_0$, hipoteza se usvaja.

Česta je praksa da se kao značajni parametri, tj. oni koji su dovoljno različiti od nule, uzimaju oni čije je $|t| \geq 2{,}0$.

Interval poverenja sa granicama poverenja od 100(1-η/2) procenata za parametar α je:

$$a \pm t_0\sqrt{V(a)} \qquad (2.56)$$

Analogno se izvodi i statistički test parametra b i za odgovarajući interval se dobija:

$$b \pm t_0 \sqrt{V(b)} \qquad (2.57)$$

U izrazima (2.56) i (2.57) odgovarajuće varijanse se sračunavaju sa ocenjenom vrednosti za varijansu slučajnih odstupanja, tj. σ_e^2.

Tako, u primeru 2.2.1. imamo:

$$\sigma_e^2 = \frac{44,76}{10} = 4,476$$

$$\sqrt{V(a)} = 3,49$$
$$\sqrt{V(b)} = 0,0098$$

Sa nivoom poverenja od 95% (η = 0,05) zaključuju se da se prava vrednost parametra nalazi u intervalu:

$$\alpha = -3,0 \pm 2,228 \cdot 0,010$$

Kako ovaj interval uključuje i vrednost 0 sa poverenjem od 95% se može zaključiti da se prava vrednost parametra a ne razlikuje značajno od nule.

Isto tako 95% interval poverenja za parametar b je:

$$\beta = 0,928 \pm 2,228 \cdot 0,010$$

Navedeni testovi su se odnosili na nezavisno testiranje parametara. Međutim, moguće je izvršiti zajednički test oba parametra uvođenjem kvadratne forme:

$$Q = \frac{1}{2}\left(N(a-\alpha)^2 + 2N\overline{X}(a-\alpha)(b-\beta) + \sum_{i=1}^{N} X_i^2 (b-\beta)^2 \right)$$

koja ima χ^2 raspodelu sa 2 stepena slobode. Na osnovu činjenice da (2.54) ima χ^2 raspodelu su (N-2) stepena slobode sledi:

$$F = \frac{\frac{Q}{2}}{\frac{\sigma_e^2}{\sigma_\varepsilon^2}}$$

ima Fisher-ovu F raspodelu sa 2 i (N-2) stepena slobode. U izgrazu (2.58) potire se nepoznata varijansa σ_ε^2 te ostaju nepoznati samo parametri α i β. Na sličan način kao i u dosadašnjim testovima, koristeći tabelu F-raspodele za dati nivo značajnosti η se može proveriti važnost hipoteze $\alpha=\alpha_0$ ι $\beta=\beta_0$ zamenom ovih vrednosti u (2.58) i ukoliko dobijena vrednost F je veća od tablične vrednosti F_η hipoteza se odbacuje. S obzirom da se zajednički testiraju dva parametra za interval poverenja se ustvari dobija elipsa.

Test ocenjene vrednosti σ_ε^2 se izvodi koristeći χ^2 raspodelu i za granice poverenja za varijansu slučajnih odstupanja σ_ε^2 dobijamo:

$$\frac{(N-2)\sigma_e^2}{\chi^2_{\frac{\eta}{2}}} \le \sigma_\varepsilon^2 \le \frac{(N-1)\sigma_e^2}{\chi^2_{1-\frac{\eta}{2}}} \tag{2.59}$$

2.2.4. Analiza varijacija

Dalja provera pouzdanosti regresionog modela se može izvršiti analizom varijacija a na osnovu rezultata izraženog jednačinom (2.50). Naime, pokazuje se da veličina:

$$F = \frac{(b-\beta)^2 \sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} \frac{e_i^2}{(N-2)}} = \frac{(b-\beta)^2 \sum_{i=1}^{N} x_i^2}{\sigma_e^2} \tag{2.60}$$

ima Fisher - ovu raspodelu sa (1, N-2) stepeni slobode. Uz pomoć izraza (2.60) se može testirati važnost hipoteze da ne postoji linearna veza između promenljivih Y i X tj. da je β=0. Tačnije, sračunavajući izraz (2.60) stavljajući β=0 dobija se vrednost F koju za dati nivo značajnosti η poredimo sa tabličnom vrednosti F_η i hipotezu odbacujemo u slučaju $F > F_\eta$.

Za β=0 izraz (2.60) ima vrednost:

$$F = \frac{Q_1}{\frac{Q_2}{(N-2)}} : F_{1,N-2} = \frac{b^2 \sum_{i=1}^{N} x_i^2}{\sigma_e^2} \tag{2.61}$$

gde je:

$Q_1 = b^2 \sum_{i=1}^{N} x_i^2$ "objašnjena" suma kvadrata

$Q_2 = \sum_{i=1}^{N} e_i^2$ "neobjašnjena" suma kvadrata.

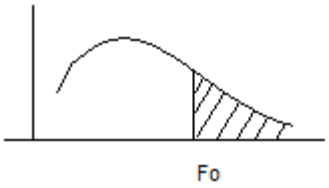
tj.

$$Q_1 = b^2(\sum_{i=1}^{N} X_i^2 - N\overline{X}^2)$$

$$Q_2 = \sum_{i=1}^{N} Y_i^2 - a\sum_{i=1}^{N} Y_i - b\sum_{i=1}^{N} X_i Y_i$$

Značajnost regresije – F-test:

$H_o : \beta = 0$
$H_1 : \beta \neq 0$ $\quad P\{F > F_0\} = \alpha$



kritična oblast $C_0 = (F_0, +\infty)$

Uobičajena praksa je da se rezultati ovih sračunavanja daju u obliku tabele analize varijacija koja ima oblik:

Izvor varijacija	Suma kvadrata	Broj st. slobode	Srednja vrednost
X	Q_1	1	Q_1
ε	Q_2	N-2	$Q_2/(N-2)$
Total	Q_1+Q_2	N-1	

Na osnovu podataka iz primera 2.2.1. imamo:

Izvor varijacija	Suma kvadrata	Broj st. slobode	Srednja vrednost
X	40021,86	1	40021,86
ε	44,8	10	4,48
Total	40066,66	11	

F(1,10) = 8933,5

2.2.5. Svođenje nekih nelinearnih na linearni model

U celokupnom dosadašnjem izlaganju razmatrali su se samo linearni modeli. Sasvim je izvesno da postoje i druge veze po kojima promenljiva Y zavisi od X. Neke od njih pe u ovom poglavlju biti diskutovane, kao što će se i pokazati da standardna regresiona procedura vrlo često se može iskoristiti i u slučaju nelinearnosti. Naravno u mnogim slučajevima to nije moguće. Da bi se to sve i pokazalo, poći će se od dva primera i uporediti ih sa osnovnim linearnim modelom:

$$Y = \alpha + \beta X + \varepsilon$$

Prvi primer: $Y = \alpha + \beta X^2 + \varepsilon$

Drugi primer: $Y = X^{\beta}\varepsilon$

Jasno je, da je u oba primera veza promenljivih Y i X nelinearna, ali postoje znatne razlike u tim nelinearnostima. Naime, u prvom slučaju nelinearnost je u promenljivoj X, dok su parametri α i β prikazani na linearan način, tako da se tehnike objašnjenje ranije mogu i ovde primeniti. Za drugi primer nelinearnost je u parametru β, odnosno u parametru koji treba da se ocenjuje, pa problem postaje znatno složeniji. Tu se pokušava iznaći takva transformacija, koja će problem prevesti u linearni domen, kada se mogu primeniti već razvijene metode.

A) Slučaj nelinearnosti u promenljivoj a ne i u parametrima

Kao primer ovakvog slučaja, predpostavimo model oblika:

$$Y = \alpha + \beta Z^2 + \varepsilon \qquad (2.62)$$

Uvođenjem nove promenljive:

$X = Z^2$

jednačina (2.62) svodi se na dobro poznati oblik linearnog modela

$$Y = \alpha + \beta X + \varepsilon$$

za koji se vrlo lako ocenjuju parametri $\hat{\beta}$ i $\hat{\alpha}$.

Za svaki konkretan primer procedura pri ocenjivanju se svodi na tri koraka:

a) Transformisanje promenljivih da bi se problem sveo na linearni.

b) Regresira se promenljiva Y u odnosu na novu promenljivu X, tj. ocenjuju se parametri $\hat{\beta}$ i $\hat{\alpha}$.

c) Vrši se re-transformacija na staru promenljivu, kada se dobija i stvarni izgled krive.

B) Slučaj nelinearnosti u parametrima

Pretpostavimo funkciju nekog procesa u obliku:

$$P = LM^{\beta}N^{1-\beta}V \qquad (2.63)$$

gde promenljive P,M i N predstavljaju promenljive kojima je opisan proces, dok V predstavlja član slučajne greške, a α i β su parametri koje treba oceniti. Za ovaj slučaj vrlo je pogodno primeniti logaritamsku transformaciju, tj. treba izvršiti logaritmovanje obe strane jednačine (2.63), kada se dobija:

$$\ln P = \ln L + \beta \ln M + (1-\beta)\ln N + \ln V$$

odnosno:

$$\ln P - \ln N = \ln L + \beta(\ln M - \ln N) + \ln V$$

Uvodeći smene:

$$Y = \ln P - \ln N$$
$$\alpha = \ln L$$
$$X = \ln M - \ln N$$
$$\varepsilon = \ln V$$

dobijamo, uz pretpostavku da V ima svoju raspodelu, klasični model oblika:

$$Y = \alpha + \beta X + \varepsilon$$

Prednost logaritamske transformacije je što tačno predefiniše parametre u parametre, odnosno promenljive u promenljive i što član greške daje u vidu zbira a ne proizvoda.

2.2.6. Predviđanja

Jedna od osnovnih namena regresionih modela je predviđanje tj. određivanje vrednosti zavisne promenljive Y_0 na osnovu date vrednosti promenljive X_0. U slučaju da se data vrednost nezavisne promenljive X_0 nalazi u intervalu između najmanje Xmin i najveće Xmax observacije tad imamo slučaj interpolacije. U slučaju da je data vrednost X_0 van navedenog intervala imamo slučaj ekstrapolacije. Može se pokazati da se kao najbolja nepristrasna linearna ocena za Y_0 dobija

$$\hat{Y}_0 = a + b\hat{X}_0 \qquad (2.64)$$

gde su a i b ocene regresionih parametara dobijene metodom najmanjih kvadrata.

Za varijansu $V(Y_0)$ se dobija:

$$V(\hat{Y}_0) = E((\hat{Y}_0 - Y_0)^2) = E((a - \alpha + bX_0 - \beta X_0)^2)$$

odnosno, zamenom odgovarajućih izraza:

$$V(\hat{Y}_0) = \sigma_\varepsilon^2 \left[\frac{1}{N} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{N} x_i^2} \right] \qquad (2.65)$$

Takođe se može pokazati da se nivo značajnosti η dobija interval poverenja za ocenjenu vrednost $\hat{Y}_0$ tj.:

$$\hat{Y}_0 \pm t_{\frac{\eta}{2}} \sqrt{\frac{1}{N} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{N} x_i^2}} \qquad (2.66)$$

Razmatrajući izraz (2.65) za varijansu ocenjene vrednosti Y_0 se može zaključiti da ona potiče od doprinosa varijanse V(a) i varijanse V(b). Isto tako se može zaključiti da što je tačka X_0 , u kojoj se vrši predviđanje, dalja od srednje vrednosti $\overline{X}$ to će i varijansa biti veća. Prema tome, može se u principu zaključiti, da se interpolacija može tačnije izvršiti od ekstrapolacije. Osim ovog, da kažemo matematičkog ograničenja na tačnost ekstrapolacije, treba uvek imati na umu da se regresija ocenjuje na osnovu datog skupa observacija Y_i i X_i i da za vrednosti X_0 jako udaljene od pomenutog skupa observacija, ocenjena regresija ne mora uopšte da važi.

Nešto drugačiji problem predviđanja je slučaj kad se želi oceniti da li par (Y_0 ,X_0) pripada linearnom regresionom modelu ocenjenom na osnovu parova observacija {Y_i ,X_i }. U ovom slučaju se za varijansu dobija:

$$E((Y_0 - \hat{Y}_0)^2) = \sigma_\varepsilon^2 W \qquad (2.67)$$

gde W označava:

$$W = 1 + \frac{1}{N} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{N} x_i^2}$$

Za interval poverenja sa nivoom značajnosti η imamo:

$$\hat{Y}_0 \pm t_{\frac{\eta}{2}} \sigma_e^2 \sqrt{W} \qquad (2.68)$$

Interval poverenja za ocenu prosečne vrednosti zavisne promenljive:

$$\hat{Y}_p - t_0\sigma_{\hat{y}p} \le E(Y_p) \le \hat{Y}_p + t_0\sigma_{\hat{y}p}$$

$$\sigma_{\hat{y}p} = \sigma_e \sqrt{\frac{1}{N} + \frac{(X_p - \overline{X})^2}{\sum_{i=1}^{N} X_i^2 - N\overline{X}^2}}$$

Interval poverenja za predviđenu vrednost zavisne promenljive:

$$\hat{Y}_p - t_0\sigma_{Yp} \le Y_p \le \hat{Y}_p + t_0\sigma_{Yp}$$

$$\sigma_{Yp} = \sigma_e \sqrt{1 + \frac{1}{N} + \frac{(X_p - \overline{X})^2}{\sum_{i=1}^{N} X_i^2 - N\overline{X}^2}}$$

2.3. Linearni regresioni model sa više promenljivih

U ovom delu će se dati generalizacija rezultata dobijenih za slučaj linearne regresije sa dve promenljive u delu 2.2., a za slučaj više nezavisnih promenljivih.

Regresioni model koji izražava vezu između zavisne promenljive Y i k nezavisnih promenljivih X_1 ,X_2,...,X_k ima opšti oblik:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \qquad (2.69)$$

gde je: i = 1,2,...,N a N označava ukupni broj observacija.

Da bi se zadržalo prisustvo konstantnog člana standardna pretpostavka u regresionim modelima oblika (2.69) je da je:

$X_{1i} = 1$; i = 1,2,...,N

Linearni regresioni modeli za više promenljivih se javljaju uvek kad veći broj promenljivih utiče na ponašanje neke pojave koja se opisuje kao zavisna promenljiva. Za ilustraciju, neka Y predstavlja veličinu prodaje nekog proizvoda a promenljive X_2 i X_3 označavaju troškove ekonomske propagande i cenu proizvoda, respektivno.

Sistem jednačina (2.69) se može pogodno prikazati u matričnom obliku kao:

$$Y = X\beta + \varepsilon \qquad (2.70)$$

gde su u skladu sa usvojenom notacijom imamo:

$Y = \{Y_1, Y_2, ...Y_N\}^T$ vektor kolona

$$X = \begin{bmatrix} 1 & X_{21} & X_{31} & ... & X_{k1} \\ 1 & X_{22} & X_{32} & ... & X_{k2} \\ ... & ... & ... & ... & ... \\ 1 & X_{2N} & X_{3N} & ... & X_{kN} \end{bmatrix}$$ matrica reda (n × k)

$\beta = \{\beta_1, \beta_2, ...\beta_k\}^T$ vektor kolona

$\varepsilon = \{\varepsilon_1, \varepsilon_2, ...\varepsilon_N\}^T$ vektor kolona

2.3.1. Ocenjivanje parametara metodom najmanjih kvadrata

Pretpostavke koje omogućavaju ocenu nepoznatih parametara β metodom najmanjih kvadrata su identične već navedenim pretpostavkama u slučaju regresije dve promenljive. Ovde će se ponovo formulisati koristeći matričnu notaciju:

1) $E(\varepsilon) = 0$

2) $E(\varepsilon\varepsilon^T) = \sigma_\varepsilon^2 I_N$

gde ε^T označava transponovani vektor vrstu vektora kolone ε, a I_N označava kvadratnu jediničnu matricu reda N.

$$E\left(\begin{bmatrix} \varepsilon_1 \\ ... \\ \varepsilon_N \end{bmatrix}_{Nx1} \begin{bmatrix} \varepsilon_1 ... \varepsilon_N \end{bmatrix}_{1xN}\right) = E\left(\begin{bmatrix} \varepsilon_1\varepsilon_1 ... \ \varepsilon_1\varepsilon_N \\ ... \\ \varepsilon_N\varepsilon_1 ... \ \varepsilon_N\varepsilon_N \end{bmatrix}_{NxN}\right) = \begin{bmatrix} \sigma_\varepsilon^2 ... \ 0 \\ ... \\ 0 ... \ \sigma_\varepsilon^2 \end{bmatrix}_{NxN}$$

Primećuje se da pretpostavka 2) obuhvata i konstantnost varijanse, s obzirom da je varijansa data članovima na glavnoj dijagonali matrice $\sigma_\varepsilon^2 I_N$ a koji su svi jednaki σ_ε^2 , i činjenicu da je kovarijansa slučajnih odstupanja identički jednaka nuli. Naime, svi elementi gornje matrice koji se nalaze van glavne dijagonale su identički jednaki nuli.

3) $E(x_i \varepsilon^T) = 0$, gde je: $x_i = \{X_{i1} X_{i2} ... X_{iN}\}$, i=1,2,...,k, čime se izražava činjenica da je svaka od nezavisnih promenljivih X_{ij} , i = 1,2,...,k; j = 1,2,...,N nezavisna od slučajnih odstupanja ε.

4) Matrica X ima rang k < N, što znači da ne postoji linearna veza između bilo koje od nezavisnih promenljivih i da je broj observacija N veći od broja k parametara koji se ocenjuju.

Ako sledeći dosadašnju praksu sa $\hat{Y}$ označimo ocenjene vrednosti vektora Y, sa b ocenjenu vrednost vektora β i sa e ocenjenu vrednost vektora ε, tada važi:

$$Y = Xb + e \qquad (2.71)$$

$$\hat{Y} = Xb$$

Ocenjeni vektor b se dobija metodom najmanjih kvadrata minimizacijom kvadratne forme:

$$ee^T = (Y - Xb)^T(Y - Xb) \qquad (2.72)$$

Diferenciranjem (2.72) po vektoru parametara b, vodeći računa o matričnoj i vektorskoj prirodi svih veličina, dobija se:

$$\frac{\partial(ee^T)}{\partial b} = -2X^TY + 2X^TXb = 0 \qquad (2.73)$$

odnosno za ocenu b dobijamo:

$$b = (X^TX)^{-1}X^TY \qquad (2.74)$$

jer, na osnovu pretpostavke 4) sledi da je matrica X^TX nesingularna te prema tome postoji njena inverzna matrica $(X^TX)^{-1}$.

Sračunavanje matrice $(X^TX)^{-1}$ u slučaju visoko korelisanih nezavisnih promenljivih postaje veoma teško, odnosno sračunate vrednosti za b su nepouzdane. Ova pojave se naziva multikolinearnost i biće kasnije detaljnije razmatrana.

Na osnovu (2.74) sledi da je:

$$\begin{aligned} b &= (X^TX)^{-1}X^T(X\beta + \varepsilon) \\ &= \beta + (X^TX)^{-1}X^T\varepsilon \end{aligned} \qquad (2.75)$$

čime se izražava činjenica da se ocena b može predstaviti kao linearna kombinacija slučajnih odstupanja ε. Drugim rečima ako se uvede dodatna pretpostavka da su slučajna odstupanja raspodeljena po normalnoj raspodeli sledi da i ocena b ima normalnu raspodelu. Napominje se da su ovo raspodele za više nezavisnih slučajnih promenljivih.

Za $\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i}$ biće:

$$b_0 = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$$b_1 = \frac{\sum_{i=1}^{N} x_{2i}^2 \sum_{i=1}^{N} x_{1i} y_i - \sum_{i=1}^{N} x_{1i} x_{2i} \sum_{i=1}^{N} x_{2i} y_i}{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i}^2 - (\sum_{i=1}^{N} x_{1i} x_{2i})^2}$$

$$b_2 = \frac{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i} y_i - \sum_{i=1}^{N} x_{1i} x_{2i} \sum_{i=1}^{N} x_{1i} y_i}{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i}^2 - (\sum_{i=1}^{N} x_{1i} x_{2i})^2}$$

Da je ocena b data izrazom (2.74) i nepristrasna ocena vidimo iz sledećeg:

$$\begin{aligned} E(b) &= E(\beta + (X^T X)^{-1} X^T \varepsilon) \\ &= E(\beta) + (X^T X)^{-1} X^T E(\varepsilon) \\ &= \beta \end{aligned}$$

koristeći pretpostavku 1).

Matrica varijansi i kovarijansi za vektor ocena b je:

$$\begin{aligned} E((b-\beta)(b-\beta)^T) &= E\left\{\left[\left(X^T X\right)^{-1} X^T \varepsilon\right]\left[\left(X^T X\right)^{-1} X^T \varepsilon\right]^T\right\} \\ &= E\left(\left(X^T X\right)^{-1} X^T \varepsilon\varepsilon^T X \left(X^T X\right)^{-1}\right) \\ &= \left(X^T X\right)^{-1} X^T E(\varepsilon\varepsilon^T) X \left(X^T X\right)^{-1} \\ &= \sigma_\varepsilon^2 \left(X^T X\right)^{-1} \end{aligned} \qquad (2.76)$$

Prema tome, varijansa parametra b_i je i-ti elemenat a_{ii} na glavnoj dijagonali matrice $\left(X^T X\right)^{-1}$ pomnožen sa varijansom slučajnih odstupanja σ_ε^2.

Za k=2:

$$v(b_0) = \sigma_e^2 (\frac{1}{N} + \frac{\overline{X}_1^2 \sum_{i=1}^{N} x_{2i}^2 + \overline{X}_2 \sum_{i=1}^{N} x_{1i}^2 - 2\overline{X}_1 \overline{X}_2 \sum_{i=1}^{N} x_{1i} x_{2i}}{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i}^2 - (\sum_{i=1}^{N} x_{1i} x_{2i})^2})$$

$$v(b_1) = \sigma_e^2 \frac{\sum_{i=1}^{N} x_{2i}^2}{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i}^2 - (\sum_{i=1}^{N} x_{1i} x_{2i})^2}$$

$$v(b_2) = \sigma_e^2 \frac{\sum_{i=1}^{N} x_{1i}^2}{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i}^2 - (\sum_{i=1}^{N} x_{1i} x_{2i})^2}$$

Varijansa slučajnih odstupanja ε se ocenjuje kao:

$$\sigma_e^2 = \frac{e^T e}{N-k}$$

s obzirom da k normalnih jednačina smanjuje broj stepena slobode na (N-k).

Za k=2, bilo bi:

$$\sigma_e^2 = \frac{\sum_{i+1}^{N} y_i^2 - b_1 \sum_{i=1}^{N} x_{1i} y_i - b_2 \sum_{i=1}^{N} x_{2i} y_i}{N-(k+1)}$$

$$\sigma_e = \sqrt{\sigma_e^2}$$

gde je σ_e standardna greška regresije.

2.3.2. Korelaciona matrica

Generalizacijom izraza dobijenog za koeficijent determinacije u slučaju dve promenljive, dobija se:

$$r^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} y_i^2} \qquad (2.77)$$

Broj stepeni slobode sume $\sum_{i=1}^{N} e_i^2$ iznosi: (N-k) dok je broj stepeni slobode sume $\sum_{i=1}^{N} y_i^2$ jednak (N-1).

Za k=2:

$$r^2 = \frac{b_1 \sum_{i=1}^{N} x_{1i} y_i + b_2 \sum_{i=1}^{N} x_{2i} y_i}{\sum_{i=1}^{N} y_i^2}$$

U opštem slučaju ovo može da unese proizvoljnost u poređenju nekoliko regresija sa različitim stepenima slobode. Zbog toga se koristi takozvani podešeni koeficijent determinacije:

$$\bar{r}^2 = 1 - (1-r)^2 \cdot \frac{N-1}{N-k} \qquad (2.78)$$

a za k=2:

$$\bar{r}^2 = 1 - (1-r)^2 \cdot \frac{N-1}{N-(k+1)}$$

Kvadratni koren iz koeficijenta determinacije se naziva koeficijent višestruke korelacije.

Ako se sračunaju sve proste korelacije između promenljivih

$Y, X_1, X_2, ..., X_k$ a na osnovu izraza (2.48) i uredimo ih u matričnu formu dobijamo takozvanu korelacionu matricu:

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{bmatrix}$$

gde r_{ij} , j = 2,3,...,k označava korelaciju Y i X_j , r_{ii} = 1 za i = 1,2,...,k i r_{ij}, i = 2,...,k; j = 2,...,k je koeficijenat korelacije između nezavisnih promenljivih X_i i X_j .

Parcijalni koeficijenat korelacije Y sa promenljivom X se definiše kao:

$$r_{1i,2,3,\dots,1-i,i+1,\dots,k} = -\frac{R_{1i}}{\sqrt{(R_{11} R_{ii})}} \qquad (2.80)$$

gde su R_{ij} odgovarajući kofaktori matrice (2.79). Parcijalni koeficijenat korelacije je mera linearne veze Y i i-te nezavisne promenljive X_i , pretpostavljajući da se ostale nezavisne promenljive održavaju fiksnim.

2.3.3. Statistički testovi

Statistički testovi kod linearne regresije sa više promenljivih se izvode na identičan način kao i u slučaju linearne regresije sa dve promenljive.

Test značajnosti ocenjenih vrednosti parametara b_i se izvodi na osnovu pretpostavke da greška e ima normalnu raspodelu pored već uvedenih pretpostavki. Kompaktni zapis ovih pretpostavki je:

$$\varepsilon: N(0, \sigma_\varepsilon^2 I_N)$$

Dakle, b_i ima raspodelu:

$$b_i: N(\beta_i, \sigma_\varepsilon^2 a_{ii})$$

Veličina:

$$t = \frac{b_i - \beta_i}{\sqrt{\frac{a_{ii}\sum_{i=1}^{N} e_i^2}{N-k}}} \qquad (2.81)$$

gde a_{ii} kao i dosad označava i-ti elemenat na glavnoj dijagonali matrice $(X^T X)^{-1}$, sledi t - raspodelu.

$$H_0 : \beta_1 = 0 \qquad H_0 : \beta_2 = 0$$
$$H_1 : \beta_1 \neq 0 \qquad H_1 : \beta_2 \neq 0$$
$$t = \frac{b_1}{\sqrt{v(b_1)}} : t_{N-3} \qquad t = \frac{b_2}{\sqrt{v(b_2)}} : t_{N-3}$$

Hipoteza da je neki regresioni parametar β_i jednak nuli tj. da nezavisna promenljiva X_i ne utiče na zavisnu promenljivu Y, se može testirati sračunavanjem izraza (2.81) za $\beta_i = 0$ sa daljim zaključivanjem identičnim onom iznetom u delu 2.2.4.

I ostali statistički testovi diskutovani u delu 2.2.4. se generišu na analogan način i primenjuju u slučaju linearne regresije sa više promenljivih.

S obzirom na definiciju koeficijenta determinacije, "objašnjena" suma kvadrata se može izraziti kao:

$$bX^T y = y^T y r^2$$

a "neobjašnjena" suma kvadrata kao:

$$e^T e = y^T y r^2$$

Tada veličina

$$F = \frac{\frac{r^2}{k-1}}{\frac{1-r^2}{N-k}}$$

sledi F - raspodelu sa (k-1,N-k) stepeni slobode i koristi se analogno u delu 2.2.5.

Izvor varijacija	Suma kvadrata	Broj step. slobode	Srednja vrednost
$X_2,X_3,...,X_k$	bX^Ty	k-1	bX^Ty / (k-1)
ε	e^Te	N-k	e^Te / (N-k)
Total	y^Ty	N-1	

2.3.4. Predviđanje

Pretpostavimo da želimo da odredimo očekivanu vrednost za promenljivu Y asociranu sa skupom vrednosti X_o koji nije u skupu observacija koji definiše regresioni model tj.:

$$X_o = \{1 \quad X_{20} \quad X_{30} \ldots \ X_{k0}\}$$

Najbolji nepristrasni prediktor za odgovarajuću vrednost Y je:

$$\hat{Y}_0 = X_0 B$$

Varijansa ove predikcije je:

$$V(\hat{Y}_0) = X_0 (X^T X)^{-1} X_0$$

Slično kao u delu 2.2.7. interval poverenja za dati nivo značajnosti η za predikciju Y je:

$$Y_0 \pm t_{\frac{\eta}{2}} \sigma_e \sqrt{X_0 (X^T X) X_0}$$

Analogno delu 2.2.7. ukoliko se želi proveriti da li par vrednosti (Y_o ,X_o) pripada datom regresionom modelu, sračunava se ocena Y i sračunava veličina:

$$t = \frac{\hat{Y}_0 - Y_0}{\sigma_e (1 + X_0^T (X^T X)^{-1} X_0)^{\frac{1}{2}}}$$

i ako sračunata vrednost za t prevazilazi neku unapred određenu vrednost za dati nivo značajnosti, tad se zaključuje da par vrednosti (Y_o ,X_o) pripada nekoj drugoj strukturi.

2.4. Vežbe

1) Pokazati da u regresiji dve promenljive, regresija Y na X je različita od regresije X na Y, u opštem slučaju. Objasniti zbog čega nastaje razlika i putem koje regresije je ocenjivanje bolje.

2) U sledećoj tabeli je data zavisnost tražnje nekog proizvoda od cene:

Q	12	10	13	11,5	12	13	12	12	13	13,5	14	13,5	14,5
P	0,54	0,51	0,49	0,49	0,48	0,48	0,48	0,47	0,44	0,43	0,42	0,41	0,40

Oceniti parametre regresije:

$$Q = \alpha P^{\beta} \varepsilon$$

diskutovati značenje koeficijenata α i β; testirati i objasniti ekonomsko značenje hipoteze $\beta=0$.

3) Pokazati da ako se za korelacioni koeficijenat N parova (X_i,Y_i) dobije vrednost r tada se ova ista vrednost dobija i za korelacioni koeficijenat N parova ($aX_i + b$, $cY_i + d$) , gde su a, b, c i d konstante.

4) Dat je uzorak od 20 observacija koji daje sledeće vrednosti:

$$\sum_{i=1}^{20} Y_i = 21,9; \sum_{i=1}^{20} X_i = 186,2; \sum_{i=1}^{20} (X - \overline{X})(Y - \overline{Y}) = 106,4;$$
$$\sum_{i=1}^{20} (X - \overline{X})^2 = 86,9; \sum_{i=1}^{20} (Y - \overline{Y})^2 = 215,4$$

Oceniti parametre α i β u linearnoj regresiji:

$$Y = \alpha + \beta X + \varepsilon$$

kao i odgovarajuće varijanse i intervale poverenja 95%. Takođe oceniti predikciju srednje vrednosti za Y_0 kad je $X_o = 10$ i naći njen 95% interval poverenja.

5) U ekonometrijskim studijama često je moguće koristiti apriorna znanja o vrednostima nekih parametara regresije. Posmatrajmo linearni model:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Oceniti parametre b gornjeg modela uz pretpostavku da važe sledeći uslovi:

a) $\beta_1 = \alpha_0$

b) $\beta_2 = \alpha_1$

c) $\beta_1 = \beta_2$

6) Dat je uzorak od 89 observacija koji daje sledeće vrednosti:

$$\overline{Y} = 5,8; \overline{X}_2 = 2,9; \overline{X}_3 = 3,9$$
$$\sum_{i=1}^{89} (Y - \overline{Y})^2 = 113,6; \sum_{i=1}^{89} (X_2 - \overline{X}_2)^2 = 50,5; \sum_{i=1}^{89} (X_3 - \overline{X}_3)^2 = 967,1$$
$$\sum_{i=1}^{89} (Y - \overline{Y})(X_2 - \overline{X}_2) = 36,8; \sum_{i=1}^{89} (Y - \overline{Y})(X_3 - \overline{X}_3) = 39,1$$
$$\sum_{i=1}^{89} (X_2 - \overline{X}_2)(X_3 - \overline{X}_3) = -66,2$$

Oceniti parametre linearnog regresionog modela koji povezuje gornje promenljive. Formirati tabelu analize varijacija i razmotriti smanjenje ukupne sume kvadrata regresijom prvo samo na X_2 a potom i zajedno na X_2 i X_3 .