Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 517.98
DOI: https://doi.org/10.24867/26JV02Zarkovic

LINEARNA REGRESIJA

LINEAR REGRESSION

Sandra Žarković, *Fakultet tehničkih nauka, Novi Sad*

Oblast – MATEMATIKA U TEHNICI

Kratak sadržaj – *Upoznajemo se sa pojmom regresione analize i vrstama regresije. Uvode se osnovni pojmovi linearne regresije, vrši se njena podela na prostu i višestruku linearnu regresiju i prikazana je metoda najmanjih kvadrata kojom dobijamo ocenu regresionih parametara. Na kraju su izneseni zaključci.*

Ključne reči:*Linearna regresija*

Abstract – *We familiarize ourselves with the concept of regression analysis and various types of regression. Basic concepts of linear regression are discussed, with a specific focus on its division into simple and multiple linear regression. The least squares method is presented as a means to obtain estimates of regression parameters. In conclusion, findings obtained through the analysis are outlined*

Key words:*Linear regression*

1. UVOD

Regresioni modeli predstavljaju jedan od najznačajnijih oblika modelovanja i istraživanja podataka. Pomoću linearnih regresionih modela se može vršiti predviđanje, mogu se objasniti neki rezultati i izvesti odgovarajući zaključci za posmatrani problem koji se dalje mogu koristiti u ekonomiji, privredi, kao i u mnogim drugim naukama.

Ukoliko tu varijablu nismo u mogućnosti izmeriti, počećemo od informacija o jednoj ili više drugih varijabli koje možemo izmeriti. Varijablu koja je u fokusu našeg interesovanja modelujemo kao kombinaciju drugih varijabli.

Kroz primenu regresije u nauci pokazalo se da je ovaj jednostavan pristup vrlo fleksibilan i vrlo koristan. Osnovne vrste regresije su linearna i nelinearna regresija, a linearna regresija se deli na prostu i višestruku regresiju [3].

Veza između povezanih promenljivih može biti linearna i nelinearna, u zavisnosti od toga da li vezu opisujemo linearnom ili nelinearnom funkcijom. Cilj regresione analize jeste ustanoviti kako se jedna varijabla Y menja u funkcionalnoj zavisnosti od drugih varijabli X_1, X_2,... i šta još pored nezavisnih promenljivih utiče na ishod razmatranja.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio prof. dr Nebojša Ralević.

2. LINEARNA REGRESIJA

Pri prvom susretu sa statistikom, susrećemo se sa slučajnim varijablama, te ih i posmatramo i analiziramo zasebno. Međutim, vrlo brzo se susrećemo s više od jedne slučajne varijable odjednom i zatim počinje razmišljanje o tome kako su one međusobno povezane. Za početak, u najjednostavnijem slučaju, dve varijable mogu biti međusobno nezavisne tj. nepovezane. Tada se ispred nas nalaze dva jednodimenzionalna problema, a ne jedan dvodimenzionalni, te je takve varijable najbolje analizirati zasebno. S druge strane, dve varijable mogu biti povezane, odnosno jedna nam daje informaciju o drugoj. Odnos među takvim varijablama je međusobno zavisan. Često je slučaj prisustvo dvodimenzionalnih podataka (x_1, y_1), … , (x_n, y_n) u kom je svaka od varijabli X i Y delimično zavisna od druge.

Najčešće smo zainteresovani za analizu neke varijable Y, i tada istražujemo na koji način varijabla$X$utiče na varijablu Y. U tom slučaju Y nazivamo varijablom odgovora, a X varijablom objašnjenja ili prediktorskom varijablom (varijablom predviđanja). Treći naziv za X je regresor što objašnjava zašto se cela tema naziva regresija. Dakle, zaključujemo da je regresija statistička metoda koja pronalazi odnos između zavisne varijable i jedne ili više nezavisnih varijabli. Ako zavisna varijabla zavisi samo od jedne nezavisne varijable, onda se radi o prostoj regresiji. S druge strane, ako zavisi od više nezavisnih varijabli, onda da je reč o višestrukoj regresiji. Kada je odnos zavisne i nezavisne varijable linearan, radi se o linearnoj regresiji [1].

Osnovna svrha primene regresione analize je da se na osnovu jedne poznate promenljive može predvideti vrednost druge, nepoznate promenljive i to iz relacije koja pokazuje njihovu zavisnost. Veze između pojava mogu biti funkcionalne (determinističke) i statističke (stohastičke). Glavni zadatak regresione analize jeste otkrivanje zakonitosti i pravilnosti koje vladaju u odnosima među masovnim statističkim pojavama i kreiranje matematičkih modela koji pomoću simbola opisuju ponašanje pojava u stvarnim uslovima funkcionisanja. Kada se u analizi međuzavisnosti definiše koja je promenljiva zavisna a koja nezavisna, onda se koriste metode regresione analize. Zavisnost pojava se utvrđuje prema prethodnim teorijskim i empirijskim saznanjima o prirodi pojava i njihovim odnosima.

U istraživanjima često se interes istraživača usmerava prema problemu povezanosti među promenljivim (obeležjima). Pri tom je od posebnog interesa mogućnost prognoziranja ili predviđanja vrednosti jedne zavisne promenljive na osnovu drugih nezavisnih promenljivih.

Prvi tako formulisani problem potiče od engleskog antropologa Francisa Galtona. On je studirajući zajedno sa Pirsonom nasleđivanje u biologiji, mereći visine očeva i sinova ustanovio neku vrstu paradoksa, odnosno, da visoki očevi imaju visoke sinove ali u proseku ne tako visoke kao što su oni sami, i slično, da niski očevi imaju niske sinove ali opet u proseku ne tako niske kao što su oni. Ovu tendenciju proseka neke karakteristike (u ovom slučaju visine) odabrane grupe da u sledećoj generaciji sinova teži ka proseku populacije a ne proseku njihovih očeva, Galton je nazvao regresijom, tačnije, regresijom prema proseku. Da bi dobio informaciju zavisnosti visine sinova od visine njihovih očeva, Pirson je pretpostavio da se ta zavisnost može izraziti kao funkcija, Y = f(X), pri čemu je Y zavisna promenljiva, odnosno promenljiva koju želimo da objasnimo ili predvidimo (u Galtonovom primeru visina sinova), a X nezavisna promenljiva koju koristimo da objasnimo zavisnu promenljivu (visina očeva) [2].

Opšti oblik modela regresije je:

$$Y = f(X_1, X_2, \ldots\ldots, X_n) + \varepsilon$$

gde je:

- f funkcija zavisnosti a,
- ε stohastički član,slučajna greška tj. rezidual.

Iz navedene relacije vidimo da se model sastoji iz determinističkog dela, koji predstavlja funkciju kojom se izražava zavisnost zavisne promenljive od određenog broja nezavisnih promenljivih, i stohastičkog dela koji predstavlja eventualno odstupanje od te navedene funkcionalne zavisnosti. Modele regresije mozemo podeliti u odnosu na broj nezavisnih promenljivih uključenih u model i u odnosu na oblik funkcije determinističkog dela regresionog modela. Prema obliku funkcije determinističkog dela, modele regresije delimo na linearne i nelinearne regresione modele.

Veza između promenljivih linearnog modela predstavljena je linearnom funkcijom čiji je grafik prava, a veza između promenljivih nelinearnog modela ima oblik neke druge matematičke funkcije čiji je grafik neka kriva linija. Iz tog razloga se nelinearni model naziva još i krivolinijski regresioni model.

Pomoću regresione i korelacione analize određujemo jačinu, smer i oblik veze između posmatranih pojava. Oblik veze, kao što je već navedeno, predstavlja oblik matematičke funkcije iz determinističkog dela modela. Smer veze može biti pozitivan i negativan [3]. Koeficijent korelacije predstavlja meru povezanosti između dve promenljive. Postoje različiti koeficijenti korelacije koji se koriste u različitim slučajevima.

PROSTA LINEARNA REGRESIJA

Najjednostavniji od svih regresionih modela je linearni regresioni model sa dve varijable X i Y. Varijabla X je nezavisna varijabla i u eksperimentu je najčešće pod kontrolom. Vrednosti ove varijable se biraju i za svaku od izabranih vrednosti dobija se jedna ili više vrednosti varijable Y. Varijabla Y je zavisna varijabla pa se govori o regresiji Y nad X.

Pretpostavke za ovaj linearni model su: normalnost, linearnost, nezavisnost i jednakost varijansi.

Normalnost: očekuje se da su vrednosti Y za svaku vrednost nezavisne varijable X normalno raspodeljene da bi procedure zaključivanja (ocenjivanje i testiranje hipoteza) bile validne.

Linearnost: očekuje da su aritmetičke sredine subpopulacija vrednosti Y sve na jednoj pravoj liniji, što se iskazuje jednačinom:

$$\mu_{y/x} = \alpha + \beta x$$

gde je:

- $\mu_{y/x}$ aritmetička sredina subpopulacionih
- Y vrednosti dobijenih za svaku vrednost X,

a α i β su populacioni regresioni koeficijenti.

Geometrijski α i β su odsečak i nagib prave linije na kojoj sve aritmetičke sredine leže.

Nezavisnost: vrednosti varijable Y iz uzorka za jednu vrednost varijable X ni na koji način ne zavise od vrednosti varijable Y dobijene za neku drugu vrednost varijable X.

Jednakost varijansi: očekuje se da su varijanse subpopulacija Y sve među sobom jednake. Sve ove pretpostavke mogu se sumirati u jednu jednačinu poznatu kao regresioni model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

gde:

- β_1 predstavlja nagib prave,
- $\beta_0$predstavlja odsečak na y-osi,
- ε je greška, odnosno, odstupanje y od aritmetičke sredine podskupa na osnovu kojeg se ta varijabla ispituje [4].

2.1. Ocena parametara

Metoda koja se koristi da se izračunaju parametri linearne jednačine iz datih tačaka naziva se metoda najmanjih kvadrata. Ovom metodom se na osnovu uzorka (X_i, Y_i), i=1,…,n određuje zavisnost Y od X, pri pretpostavci da je ona oblika $Y = aX + b + \varepsilon$. Pretpostavimo da svako Y možemo predstaviti kao u modelu (1).

Neka je:

$$F = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X))^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

Traži se rešenje optimizacionog problema (1):

$$\min F = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} \varepsilon_i^2 \qquad (2)$$

Nalaženje minimuma ove funkcije svodi se na rešavanje sistema normalnih jednačina:

$$\frac{\partial F}{\partial \beta_i} = 0,\ i = 0,1$$

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i.$$

Ako označimo:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n,\ \ \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n.$$

$$S_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

$$S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2,$$

rešenje ovog sistema su

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}.$$

Ocenjeni model (fitovani model) je

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1,2,\dots,n. \tag{3}$$

Može se napisati u obliku

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_1 (X_i - \bar{X}) = \hat{\beta}_0 + \hat{\beta}_1 (X_i - \bar{X})$$

gde je $\bar{Y} = \beta_0$.

Ocene greške $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i),\ i = 1,2,\dots n$ date su sa $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

Uvedemo oznake : $S_{YY} = \sum (Y_i - \bar{Y})^2$, $R_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$.

Tada se ocenjeni parametri mogu napisati na sledeći način:

$$\hat{\beta}_1 = R_{XY}\sqrt{\frac{S_{YY}}{S_{XX}}} \quad \text{i} \quad \hat{\beta}_0 = \bar{Y} - \bar{X} R_{XY}\sqrt{\frac{S_{YY}}{S_{XX}}}.$$

a jednačina prave u linearnom modelu (3) može se zapisati u sledeća tri oblika:

$$\hat{Y}_i = \sqrt{\frac{S_{YY}}{S_{XX}}} R_{XY} X_i + \bar{Y} - \bar{X}\sqrt{\frac{S_{YY}}{S_{XX}}} R_{XY},$$

$$\hat{Y}_i - \bar{Y} = \sqrt{\frac{S_{YY}}{S_{XX}}} R_{XY}(X_i - \bar{X}),$$

$$\frac{\hat{Y}_i - \bar{Y}}{\sqrt{S_{YY}}} = R_{XY}\frac{X_i - \bar{X}}{\sqrt{S_{XX}}} \quad i = 1,2,\dots,n.$$

Kvadratni koreni varijansi ocenjenih parametara su njihove standardne greške tj:

$$SE(\widehat{\beta_1}) = s_{\widehat{\beta_1}} = \sqrt{\frac{s^2}{S_{XX}}},$$

$$SE(\widehat{\beta_0}) = s_{\widehat{\beta_0}} = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}$$

Centrirana ocena za varijansu reziduala σ^2 se može dobiti pomoću ocene $\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ za zbir kvadrata reziduala $\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (Y_i - f(X_i))^2$.

$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ se naziva i rezidualna suma kvadrata (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{f}(X_i))^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i))^2$$

$$\widehat{\sigma^2} = s^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \frac{1}{n-2}\sum_{i=1}^{n} (Y_i - \hat{Y}_i))^2 = \frac{1}{n-2} RSS \tag{4}$$

2.2. Višestruka linearna regresija

Model višestruke linearne regresije predstavlja se oblikom:

$$Y = X\beta + \varepsilon \tag{5}$$

gde su:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},\ X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Kod standardnih pretpostavki modela treba dodati da u slučaju višestruke regresije imamo i pretpostavku da su promenljive $X_1, X_2, \dots, X_n$ linearno nezavisne [5].

2.2.1. Metod najmanjih kvadrata

Parametre možemo oceniti različitim metodama, ali najčešće korišćen je metod najmanjih kvadrata, kao i kod proste linearne regresije. Dakle, tražimo takve vrednosti vektora β za koje funkcija:

$$\min_{\beta_k, k=0,..,p} (Y - X\beta)^T (Y - X\beta) \tag{6}$$

ima minimalnu vrednost (2) (Lozanov-Crvenković, 2012b). Rešavanje ovog sistema svodi se na sistem normalnih jednačina i njegovim rešavanjem dobijamo vrednosti ocenjenih parametara koje iznose [5]:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{7}$$

Ove ocene su nepristrasne, jer se može pokazati da je $E(\hat{\beta}) = \beta$. $\hat{\beta}$ ima $\mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$ raspodelu.

Ocena varijanse σ^2reziduala je:

$$\hat{\sigma}^2 = s^2 = \frac{(Y - \hat{Y})^T (Y - \hat{Y})}{n - (p+1)}. \tag{8}$$

2.3. Evaluacija regresionog modela

Kad se izračuna linearna regresiona jednačina mora biti evaluirana da bi se odredilo da li ona adekvatno opisuje odnos između dve varijable i da li se efektivno može koristiti u ocenjivanju i/ili predviđanju. Evaluacija je bazirana na ispitivanju nagiba regresione jednačine i na ispitivanju koeficijenta determinacije. Nagib regresione jednačine trebao bi biti značajno različit od nule, odnosno treba odbaciti nultu hipotezu H_0 da je b=0. U suprotnom, teorijski se pokazuje da su takvi modeli po pravilu nelinearnog tipa ili su linearni, ali bez značaja za predviđanje. Ispitivanje ove hipoteze može se sprovesti primenom analize varijanse i F statistike ili primenom t statistike. Osim nagiba potrebno je proceniti i jačinu regresione jednačine tako da se uporedi rasipanje uzoračkih tačaka oko regresione linije sa njihovim rasipanjem oko $\hat{Y}$, odnosno da se izračuna koeficijent determinacije r^2 [6].

2.3.1. Koeficijent determinacije

Ukupno odstupanje jedne registrovane vrednosti promenljive $Y_i$od srednje vrednosti $\bar{Y}$ se može podeliti na: modelom objašnjeno odstupanje $\hat{Y}_i - \bar{Y}$, i neobjašnjeno odstupanje $Y_i - \hat{Y}_i$ registrovanih vrednosti od vrednosti određenih modelom. To možemo zapisati [5]:

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

Gde je:

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

Ako uvedemo oznake:

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

tada je $S_{YY} = S_R + S_E$. Neka je sa R^2označen količnik odstupanja objašnjenog modelom i ukupnog odstupanja [7]:

$$R^2 = 1 - \frac{S_E}{S_{YY}}.$$

Ovaj količnik predstavlja relativnu meru fitovanja i uvek se nalazi između 0 i 1. Za slučaj kada je R^2 blizu 0, varijabilnost nije u potpunosti objašnjena modelom, već greškom i za ovakav model smatramo da je izuzetno loš. Kada je R^2 blizu 1, varijabilnost u Y je uglavnom objašnjena nezavisnom promenljivom X, što znači da je model dobar.

2.3.2. Koeficijent determinacije i adjungovani koeficijent determinacije

Pored koeficijenta determinacije koji računamo kao i u slučaju proste linearne regresije, u slučaju višestruke regresije imamo i adjungovani (korigovani) koeficijent determinacije koji računamo[5]:

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{n-(p+1)}}{\frac{SST}{n-1}}$$

gde je SST ukupan zbir kvadrata a SSE ukupan zbir kvadrata grešaka.

Uopšteno govoreći, uvodeći dodatne nezavisne promenljive u model, povećava se broj ocenjenih parametara uz nepromenjen obim uzorka, čime se povećava broj stepeni slobode pa se smanjuje pouzdanost ocenjivanja. Stoga je prednost korigovanog koeficijenta determinacije to što on uzima u obzir odnos broja promenljivih i obim uzorka, pa ga je prikladno koristiti za modele koji sadrže različit broj nezavisnih promenljivih [5].

3. ZAKLJUČAK

Linearna regresiona analiza je jedan od najčešće korišćenih modela u analizi relacije kontinuiranih varijabli. U ovom radu obrađene su teorijske osnove generalnog linearnog modela, jednostavne linearne regresione analize i višestruke regresione analize. Predstavljena je metoda najmanjih kvadrata i evaluacija regresionog modela pomoću koeficijenta determinacije. Uočili smo da se pomoću metode najmanjih kvadrata dobijaju ocene nepoznatih parametara u modelu.

4. LITERATURA

[1] Dodge, Y. 2008. The Concise Encyclopaedia of Statistics. Springer.

[2] Montgomery, D. C., and Runger, G. C. (2010). Applied statistics and probability for engineers. John

[3] Spiegel, M. S. (2008). Schaum's outline of theory and problems of Statistics. New York: McGraw-Hill

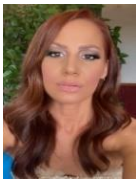
[4] Lozanov-Crvenković, Z. (2012). Statistika. Novi Sad: Prirodno-matematički fakultet u Novom Sadu.

[5] Lozanov-Crvenković, Z. (2012). Višestruka regresija. materijal sa predavanja iz predmeta Statističko modeliranje, Prirodno- matematički fakultet, Univerzitet u Novom Sadu.

[6] Čobanović, K., Nikolić-Dorić, E., Mutavdžić, B. (2000). Teorijski i praktični aspekti modela nelinearne regresije. Acta periodica technologica, 31(1), 351-62.

[7] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). "An Introduction to Statistical Learning with Applications in R", Springer, New York.

Kratka biografija:



Sandra Žarković rođenaje u Novom Sadu 1987. god. Diplomirala na Prirodno-matematičkom fakultetu. Master rad piše na Fakultetu tehničkih nauka iz oblasti Linearne i nelinearne regresije u inženjerstvu.
kontakt: sandrinazarkovic@gmail.com