Linearna višestruka regresija i korelacija

U praksi je često ponašanje neke pojave pod uticajem ne samo jedne pojave već niza drugih pojava. Jedna pojava se posmatra kao zavisna veličina (Y), a ostale koje imaju uticaj na nju su nezavisne promenljive veličine (X_1, X_2, …, X_n). Višestruka regresija je statistička procedura koja mnogo bolje opisuje međuzavisnost pojava u realnosti.

Izračunavanje parametara višestruke regresije, ne predstavlja problem obzirom da su nam na raspolaganju razni softveri i preporuka je da se primeri iz ove oblasti odrađuju korišćenjem nekih od softvera koji to omogućavaju.

Linearna višestruka regresija i korelacija je samo proširenje proste linearne regresije i korelacije.

Primer, deterministički deo trodimenzionalnog modela višestruke regresije i korelacije jednak je:

$$Y_1 = b_0 + b_1 X_1 + b_2 X_2$$

Nepoznate parameter navedenog modela višestruke regresije možemo odrediti kristeći metod najmanjih kvadrata gde se dobija sistem jednačina:

$$b_0\, n + b_1 \Sigma X_1 + b_2 \Sigma X_2 = \Sigma Y$$

$$b_0 \Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2 = \Sigma X_1 Y$$

$$b_0 \Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2 = \Sigma X_2 Y$$

Rešavanjem ovog sistema jednačina dobijaju se ocenjeni parametri $b_0,\ b_1,\ b_2$.

Standardna greška višestruke regresije S_E pokazuje koliko, u proseku, odstupaju empirijske vrednosti obeležja Y u odnosu na regresionu ravan Y_i

To je *apsolutna* mera neobjašnjenog varijabiliteta.

Standardna greška jednaka je nuli ako su sve tačke na ravni Y_i. To je idealan slučaj višestruke linearne regresije i veoma je redak. Standardna greška se računa se po:

$$S_E = \sqrt{\frac{\Sigma (Y - Y_i)^2}{n-3}}$$

R^2 je koeficijent **višestruke determinacije** koji pokazuje procenat varijacija zavisne promenljive Y koji je objašnjen zajedničkim uticajem nezavisnih promenljivih uključenih u model. R^2 se u opštem slučaju dobija kao:

$$R^2 = \frac{objašnjeni\ varijabilitet}{ukupni\ varijabilitet} = \frac{\sum(Y_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}.$$

$0 \leq R^2 \leq 1$, što je bliži jedinici veće je učešće objašnjenog varijabiliteta u ukupnom, odnosno ocenjena regresiona j-na bolje reprezentuje empirijske podatke

Nedostatak koeficijenta višestruke determinacije je u tome što zavisi od broja promenljivih i veličine uzorka. Zbog toga je potrebno da ga korigujemo.

Korigovani koeficijent višestruke determinacije je:

$$\bar{R}^2 = 1 - \frac{n-1}{n-(k+1)}(1-R^2);\ \bar{R}^2 \leq R^2$$

gde je k - broj promenljivih u uzorku.

Ova razlika se smanjuje sa povećanjem uzorka, uz nepromenjen broj nezavisnih promenljivih u posmatranom modelu.

Koeficijent višestruke linearne korelacije (označavamo ga sa R) je *relativna mera* kao i koeficijent proste linearne korelacije i pokazuje stepen linearnog slaganja varijacija u uzorku između posmatrane zavisne promenljive Y i grupe nezavisnih promenljivih $X_1, X_2, \ldots, X_n$.

Za razliku od koeficijenta proste linearne korelacije r, on nikada ne može biti negativan, tj. $0 \leq R \leq +1$. ($R = +\sqrt{R^2}$).

Kako bi izračunavanje višestruke korelacije bilo što preciznije, potrebno je koristiti što veći uzorak sa više vrednosti promenljivih.

Koeficijent višestruke linearne korelacije ne može dati informaciju o *relativnom pojedinačnom značaju nezavisnih promenljivih*. To se postiže pomoću ***koeficijenta delimične ili parcijalne korelacije.***

Koeficijent delimične korelacije pokazuje stepen linearnog slaganja varijacija zavisne promenljive i jedne nezavisne promenljive pri čemu je uticaj druge nezavisne promenljive isključen.

Slično, koeficijentu proste linearne korelacije, koeficijent delimične korelacije može uzeti vrednost od -1 do +1, i predstavlja ocenu koeficijenta delimične korelacije u osnovnom skupu.

Koeficijent delimične korelacije se dobija pomoću koeficijenta proste korelacije između promenljivih navedenih u indeksu $r_{yx_i}, r_{yx_j}, r_{x_i \cdot x_j}$:

$$r_{yx_i \cdot x_j} = \frac{r_{yx_i} - r_{yx_j} \cdot r_{x_i \cdot x_j}}{\sqrt{\left(1 - r_{yx_j}{}^2\right)\left(1 - r_{x_i \cdot x_j}{}^2\right)}},$$

Ako je $r_{yx_i \cdot x_j} < r_{yx_j \cdot x_i}$ možemo zaključiti da vrednost slučajne promenljive X_1 ima veći relativni značaj od vrednosti promenljive X_2 na zavisnu promenljivu Y.

Primer je dat u materijalu za vežbe.

INDEKSNI BROJEVI

Društvene pojave, (posebno) ekonomske, u toku vremena se manje ili više menjaju. Njihova dinamika se prati pomoću vremenskih serija koje predstavljaju niz podataka za nivo (veličinu) posmatranih pojava u sukcesiji vremenskih intervala.

Vremenske serije se posmatraju kao empirijske funkcije koje izražavaju zavisnost pojava od vremena, tj. vreme se uzima kao nezavisna promenljiva, a veličina pojave čije kretanje se prati kao zavisna promenljiva ili funkcija. Vreme koje označavamo sa (x) nije samo okvir posmatranja neke pojave (y) nego i faktor koji uzrokuje promene u njihovim odnosima.

Da bi se došlo do ispravnih zaključaka o dinamici posmatranih pojava i faktora koji ih opredeljuju, vremenske serije moraju da budu homogene, tj. sastavljene od uporedivih podataka. To znači, da ista pojava mora da bude definisana i merena na isti način za sve vreme njenog posmatranja. Mogu se upoređivati samo podaci koji se odnose na iste vremenske jedinice.

Uređivanjem statističkih podataka koji se odnose na dva perioda ili više njih nastaje vremenski statistički niz.

Vremenski niz čine hronološki uređene pojave $y_1, y_2, ..., y_t, \dots, y_n$ određene u vremenskim trenucima ili intervalima vremena.

t (vreme) – obeležje, y_t frekvencija

Razlikujemo :

- intervalne vremenske nizove,

- trenutne vremenske nizove.

Intervali posmatranja kod intervalnih vremenskih nizova su na pr. dani, nedelje, meseci, godine. Vrednosti članova ovih nizova nastaju sabiranjem vrednosti pojava po tim intervalima i imaju svojstvo kumulativnosti.

Grafički se intervalni nizovi prikazuju površinskim ili linijskim grafikonima.

Primer intervalnog vremenskog niza za proizvodnju povrća u tonama u jednoj opštini.

Proizvodnja povrća u t u jednoj opštini		
Godina	Proizvodnja	Kumulativ
2004	13678	13678
2005	18620	32298
2006	16678	48976
2007	9395	58371
2008	8524	66895
2009	11322	78217
2010	10071	88288

Trenutni vremenski nizovi iskazuju brojčano stanje pojave u odabranim, većinom jednako udaljenim trenutcima i nemaju svojstvo kumulativnosti.

Primer trenutnog vremenskog niza

Broj nezaposlenih žena u jednoj opštini	
Godina	Broj žena
2005	2832
2006	2462
2007	2018
2008	2050
2009	2024
2010	2033

Statistička analiza vremenskih nizova sastoji se od njihovog grafičkog prikazivanja i primene različitih brojčanih postupaka radi uočavanja obeležja razvoja pojave u vremenu.

Osnovni brojčani pokazatelji relativnih promena

U praksi se često pojavljuje problem analiziranja i upoređivanja dinamike kretanja najmanje dve različite pojave. Najčešće se računaju vrednosti niza uzastopnih apsolutnih promena (oznaka: Δy_t) neke pojave. Te se vrednosti dobiju tako što se od vrednosti pojave u tekućem periodu oduzme vrednost pojave u prethodnom periodu, tj.

$\Delta y_i = y_i - y_{i-1}$ za svaki $i = 1,2,...,n$.

Posmatrani vremenski period može biti najčešće godina, kvartal itd.

Istaknimo da se apsolutne promene mogu računati u odnosu na isti (bazni) vremenski period. Odgovarajući pokazatelji se dobiju tako što se od frekvencije pojave u tekućem periodu oduzme frekvencija pojave u baznom periodu:

$\Delta y_i^* = y_i - y_b$ za svaki $i = 1,2,...,n$.

Vrednosti tih pojava mogu se značajno razlikovati po svojoj veličini. Zbog toga se umesto apsolutnih promena izračunavaju i interpretiraju *relativni pokazatelji* promena tih pojava.

Jedna od najčešćih mera relativnih promena je koeficijent dinamike (oznaka: K_i).

Koeficijent dinamike je jednak odnosu apsolutne promene frekvencija dva uzastopna perioda i frekvencije pojave u prvom od ta dva perioda:

$K_i = \frac{\Delta y_i}{y_{i-1}} - \frac{y_i}{y_{i-1}} - 1$, za svako $i = 1, 2, ..., n$.

Usko vezana uz koeficijent dinamike je pojedinačna stopa promene (oznaka: S_i) neke pojave.

Pojedinačna stopa promene iskazuje relativnu promenu vrednosti pojave u dva uzastopna vremenska perioda, a dobija se množenjem koeficijenta dinamike sa 100:

$S_i = \left(\frac{y_i}{y_{i-1}} - 1\right) \cdot 100 = K_i \cdot 100$, za svako $i = 1, 2, ..., n$.

Pojedinačna stopa promene predstavlja procenat promene frekvencije neke pojave u *tekućem* vremenskom periodu u odnosu na prethodni vremenski period. Iz tog razloga se uz numeričku vrednost pojedinačne stope promene uvek piše znak %.

Osim na ovaj način pojedinačna stopa promene se može računati i usporeďivanjem sa vrednostima u fiksiranom vremenskom periodu. U takvim slučajevima pojedinačnu stopu promene vrednosti pojave u tekućem periodu u odnosu na vrednost te pojave u proizvoljnom, ali fiksiranom- baznom periodu y_b dobijamo iz izraza:

$S_i^* = \left(\frac{y_i}{y_b} - 1\right) \cdot 100.$

Primer 1

U tablici su prikazani podaci o kretanju broja prodatih automobila u četiri godine.

period	broj prodatih automobila
2008	3 724
2009	3 567
2010	3 240
2011	3 500

Potrebno je izračunati koeficijente dinamike promene broja prodatih automobila u uzastopnim vremenskim periodima, kao i odgovarajuće pojedinačne stope promene. Objasniti značenje dobijenih rezultata za 2010. i 2011. godinu.

Potom uporediti brojeve prodatih automobila u svakoj godini s brojem prodatih automobila u 2011. godini. Objasniti značenje pokazatelja koji odgovara 2008 . godini.

Rešenje:

Najpre ćemo izračunati koeficijente dinamike i njima odgovarajuće stope promene:

Godina	i	broj prodatih automobila	K_i	S_i
2008	0	3 724	-	-
2009	1	3 567	-0,04216	-4,2159 %
2010	2	3 240	-0,09167	-9,16737 %
2011	3	3 500	0,080247	8,024691 %

Iz dobijene tabele uočavamo da je broj prodatih automobila u 2010. godini bio približno 9,17% manji nego 2009. godine.

Sada izračunavamo pojedinačne stope promene uzimajući 2011. godinu kao *bazni period*:

Godina	i	broj prodatih automobila	$Si*$
2008	0	3 724	6.4 %

2009	1	3 567	1.91429 %
2010	2	3 240	-7.42857 %
2011	3	3 500	0 %

Pokazatelj koji odgovara 2008. godini jednak je 6,4%, što znači da je broj prodatih automobila 2008. godine bio za 6,4% veći nego 2011. godine.

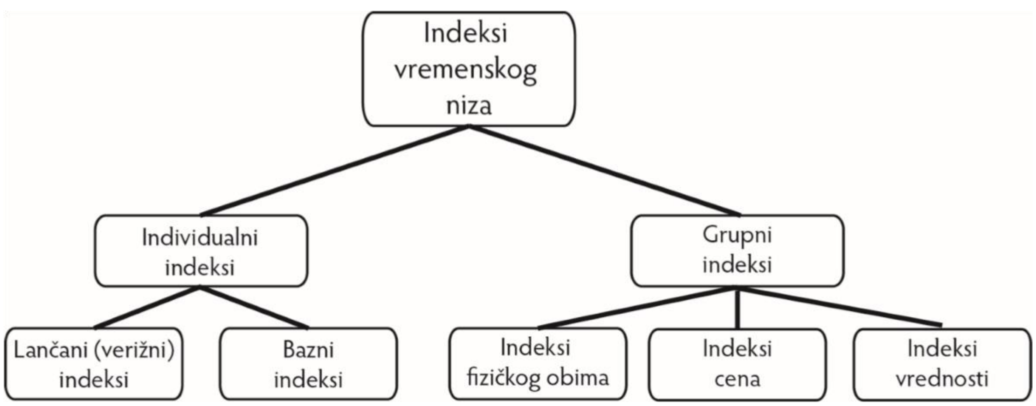
Individualni indeksi

Relativne varijacije posmatrane pojave u različitim vremenskim intervalima ili trenucima iskazuju se relativnim brojevima – *indeksima.* Indeksi predstavljaju odnos nivoa pojave u posmatranom periodu i baznom periodu.

Ako se pomoću indeksa prati razvoj jedne pojave u vremenu tada je reč o *individualnim indeksima*, dok *grupni indeksi* prate razvoj skupa (grupe) pojava.

U zavisnosti od toga da li je baza stalna ili promenljiva individualni indeksi se dele na:

- bazne indekse,
- lančane (verižne) indekse.



Lančani indeksi

Lančani indeksi su relativni brojevi (u %) koji pokazuju promene stanja pojave u uzastopnim periodima, odnosno pokazuju za koliko se procenata vrednost pojave u jednom periodu promenila u odnosu na prethodni vremenski period.

Ove indekse računamo tako da se vrednost *i*-tog perioda podeli s vrednošću prethodnog *i-1* perioda, a zatim taj odnos pomnoži sa sto.

$$L_i = \frac{y_i}{y_{i-1}} \cdot 100\ \%.$$

Lančani indeks pokazuje koliko jedinica pojave u *i*- tom vremenskom periodu dolazi na svakih sto jedinica pojave *i-1* perioda. Lančani indeks za prvi period se ne može izračunati pa se za taj period stavlja – (ne 0, ili 100). Vrednost indeksa je broj koji može biti 100, manji ili veći od 100.

Koeficijent dinamike i pojedinačnu stopu promene u uzastopnim vremenskim periodima možemo izračunati pomoću odgovarajućeg lančanog indeksa:

$K_i = \frac{L_i}{100} - 1$, $S_i = L_i - 100 = K_i \cdot 100$, za svako $i = 1,2,...,n$.

Pomoću vrednosti pojave u jednom periodu i svih lančanih indeksa možemo izračunati vrednosti pojave u svakom od preostalih n–1 perioda na temelju sledećih relacija:

Za godine koje prethode izabranoj baznoj $y_{i-1} = \frac{y_i}{L_i} \cdot 100$

Za godine koje slede odabranoj bazi $y_i = \frac{L_i \cdot y_{i-1}}{100}$

Primer 1

U sledećoj tabeli navedeni su lančani indeksi promene prosečne isplaćene mesečne plate u periodu od 2006. do 2010. godine u jednom preduzeću:

Godina	i	L_i
2006	0	-
2007	1	105,9
2008	2	105,9
2009	3	105.2
2010	4	105,2

Ako je prosečna mesečna isplaćena neto-plata u 2009. godini iznosila 46030,00 din, odrediti prosečne mesečne isplaćene neto-plate u preostalim godinama.

Rešenje:

Vremenski indeks vrednosti $i = 3$ odgovara 2009. godini, pa je

$y_3 = 46030,00$.

Za $i < 3$, odnosno $i = 0, 1, 2$ (to su godine koje prethode 2009-toj) dobijamo redom:

$$y_{2008} = y_2 = \frac{y_3}{L_3} \cdot 100 = \frac{46030,00}{105,2} \cdot 100 \approx 43754,75;$$

$y_{2007} = y_1 = \frac{y_2}{L_2} \cdot 100 = \frac{43754,75}{105,9} \cdot 100 \approx 41317,04;$

$y_{2006} = y_0 = \frac{y_1}{L_1} \cdot 100 = \frac{41317,04}{105,9} \cdot 100 \approx 39015,15.$

Za i = 4 (godina koja sledi posle 2009)

$y_{2010} = y_4 = \frac{L_4 y_3}{100} = \frac{105,2 \cdot 46030,00}{100} \approx 48423,56.$ ■

Prosečna stopa promene je konstanta kojom se zamenjuje niz pojedinačnih stopa. Ona je prosečna i relativna promena vrednosti neke pojave u ukupno posmatranom vremenskom periodu (iskazuje se u %).

Prosečna stopa promene računa se pomoću *geometrijske sredine lančanih indeksa* i može se iskoristiti za predviđanje razvoja pojave za periode koji slede nakon poslednjeg u posmatranom vremenskom nizu.

$G = \sqrt[n-1]{L_1 L_2 \cdot \ldots \cdot L_n}$

Iz definicije lančanih indeksa sledi da njihovu geometrijsku sredinu možemo izračunati i prema formuli:

$G = \sqrt[n-1]{\frac{y_n}{y_0}} \cdot 100.$

Možemo zaključiti da za alternativno računanje vrednosti geometrijske sredine lančanih indeksa moramo znati samo prvi (y_0) i poslednji (y_n) član vremenskog niza. To je ujedno i najveći nedostatak geometrijske sredine kao mere prosečnog tempa promene, jer ne uzima u obzir sve frekvencije vremenskog niza, nego isključivo prvu i poslednju frekvenciju u vremenskom nizu, tako da se u praksi navedene vrednosti prognoziraju preciznije na temelju nekoga od modela trenda.

Pomoću spomenute geometrijske sredine lančanih indeksa računa se i ***prosečna stopa promene vrednosti*** posmatrane pojave (oznaka: *S*):

$S = (G - 100)$ %.

Na temelju nje se okvirno mogu prognozirati frekvencije posmatrane pojave i u budućim vremenskim periodima uz nužnu pretpostavku da se dinamika pojave ne menja u odnosu na dinamiku pojave u posmatranom periodu.

Primer 2.

Na osnovu podataka iz Primera 1. Izračunati prosečnu godišnju stopu promene prosečne mesečne isplaćene neto zarade u periodu od 2006. do 2010. godine, i na osnovu toga proceniti prosečnu mesečnu isplaćenu neto zaradu za 2011, 2012 . i 2013. godinu.

Rešenje:

Geometrijska sredina lančanih indeksa jednaka je:

$$G = \sqrt[4]{\frac{48423{,}56}{39015{,}15}} \cdot 100 = 105{,}55$$

Sledi prosečna godišnja stopa promene je:

$S = G - 100 = 105{,}55 - 100 = 5{,}55\%$,

Dobijeni rezultat nam ukazuje da su u posmatranom periodu prosečne mesečne zarade rasle za prosečno 5,55% godišnje.

Budući da 2011. godini odgovara vrednost vremenske promenljive $t = 5$, 2012. godini $t = 6$, a 2013. godini $t = 7$, prognozu za 2011., 2012. i 2013. godinu dobijamo na sledeći način :

$$\hat{y}_5 = y_0 \cdot \left(\frac{G}{100}\right)^5 = 39015{,}15 \cdot \left(\frac{105{,}55}{100}\right)^5 = 51112{,}19;$$

$$\hat{y}_6 = y_0 \cdot \left(\frac{G}{100}\right)^6 = 39015{,}15 \cdot \left(\frac{105{,}55}{100}\right)^6 = 53948{,}91;$$

$$\hat{y}_7 = y_0 \cdot \left(\frac{G}{100}\right)^7 = 39015{,}15 \cdot \left(\frac{105{,}55}{100}\right)^7 = 56943{,}08.$$

Dakle, prognoza za prosečnu mesečnu neto zaradu za 2011. je 51112,19; za 2012. je 53948,91, a za 2013. godinu je 56943,08 dinara.

■

6.2.2. Bazni indeksi

Bazni indeksi izražavaju procentualnu promenu nivoa pojave, u određenom vremenskom periodu u odnosu na njen nivo u jednom fiksiranom - baznom periodu y_0.

Indeks sa stalnom bazom, zapravo predstavlja procenat (udeo) frekvencije posmatrane pojave u tekućem periodu u odnosu na frekvenciju posmatrane pojave u baznom periodu.

Bazni indeksi tumače se u procentima tako da se od njih oduzima 100. Oni pokazuju koliko jedinica pojave u i-tom periodu dolazi na svakih 100 jedinica pojave u baznom periodu. Mogu biti veći od 100, jednaki 100 ili manji od 100.

Opšti obrazac je:

$$B_i = \frac{y_i}{y_0} \cdot 100\%.$$

Najčešće bazne indekse posmatramo kao:

- Indekse fizičkog obima, $B_{q_i} = \frac{q_i}{q_0} \cdot 100\%$;
- Indekse cena, $B_{p_i} = \frac{p_i}{p_0} \cdot 100\%$;
- Indekse vrednosti $B_{pq_i} = \frac{p_i q_i}{p_0 q_0} \cdot 100\%$.

Sve frekvencije neke pojave možemo odrediti pomoću svih unapred zadatih baznih indeksa i jedne unapred zadate frekvencije te pojave. Razlikujemo sledeća dva slučaja:

1. Ako je zadata frekvencija pojave u baznom periodu *b* (tj. vrednost y_b), tada vrednosti pojave u svim ostalim periodima dobijamo iz jednakosti:

$$y_i = \frac{B_i y_b}{100}, \ i = 0,1,2,...,n-1.$$

2. Ako je zadata frekvencija pojave y_t, u nekom drugom periodu koji nije bazni, tada se prvo računa frekvencija pojave u baznom period y_b, pa se sve ostalo svodi na predhodni slučaj.

$$y_b = \frac{y_i}{B_t} \cdot 100.$$

Kao i za lančane indekse, i za bazne indekse se može računati odgovarajuća pojedinačna stopa promene

$$S_i^* = B_i - 100 \text{ za svako } i = 0, 1, 2, \ldots, n-1.$$

Napomenimo da i iz baznih indeksa nezavisno od izbora baze *b* možemo izračunati prosečnu stopu promene frekvencija neke pojave u posmatranom periodu, prema formuli:

$$\bar{S} = \left(\sqrt[n-1]{\frac{B_{n-1}}{B_0}} - 1\right) \cdot 100.$$

Jedna od značajnih prednosti individualnih indeksa jeste mogućnost preračunavanja svake pojedine vrste indeksa u preostalu vrstu indeksa bez poznavanja originalnih emipirijskih frekvencija vremenskog niza.

Pretvaranje individualnih indeksa možemo podeliti u tri grupe:

1. Pretvaranje baznih indeksa sa bazom b_0 u nove bazne indekse sa nekom drugom bazom $b_1 \neq b_0$;

$B_{b_{0i}} = \frac{B_{0i}}{B_{b_0 b_1}} \cdot 100\%$, za svako $i = 1, 2, \ldots, n$.

2. Pretvaranje baznih indeksa sa bazom b_0 u lančane indekse;

$L_i = \frac{B_{0i}}{B_{b_{01}}} \cdot 100\%$, za svako $i = 1, 2, \ldots, n$.

3. Pretvaranje lančanih indeksa u bazne indekse sa bazom b_0.

- za $i = b_0$, $B_i = 100$;
- za $i < b_0$, $B_{i-1} = \frac{B_i}{L_i} \cdot 100$;
- za $i > b_0$, $B_i = \frac{L_i \cdot B_{i-1}}{100}$.

dr Slavica Dabetić