Višestruka regresija i korelacija

Ako se ispituje zavisnost jedne pojave od dve ili više nezavisnih pojava, onda se govori o višestrukoj ili multiploj regresiji. Zadatak regresije je da otkrije što više faktora (nezavisnih promenljivih) koji utiču na zavisnu promenljivu. Polazi se od pretpostavke da što je više nezavisnih varijabli u modelu, sve je manji uticaj latentne promenljive (standardne greške) ε_i, $i = 1,2,...,n$. Veoma je bitno pažljivo birati promenljive koje će biti uključene u model.

Osnovni višestruki regresioni model izgleda na sledeći način:

$$\hat{x}_{i1.23..m} = a_{1.23..m} + b_{12.34..m}\, x_{i2} + b_{13.24..m}\, x_{i3} + ... + b_{1m.23...(m-1)}\, x_{im} + \varepsilon_i\,, \quad i = 1, 2, \ldots n$$

$a_{1.23...m}$ – slobodni član,
$\hat{x}_{i1.23...m}$, $i = 1,2,...,n$ – pojedinačne vrednosti regresije,
$x_{i2},.., x_{ij}$, $i = 1,2,...,n$, $j = 1,2,...,$m – vrednosti nezavisnih promenljivih,
$b_{12.34...m}$, $b_{13.24...m}$, $b_{1m.23...(m-1)}$ – regresioni koeficijenti,
ε_i, $i = 1,2,...,n$ – latentna promenljiva (slučajna greška),
m – broj nezavisnih promenljivih,
n – broj jedinica u uzorku.

Veoma često u literaturi se za označavanje zavisne promenljive koristi simbol „Y“. U ovom tekstu će naizmenično biti upotrebljavana oba simbola.

Ovaj model daje najbolje moguće predviđanje vrednosti zavisne promenljive na osnovu vrednosti nezavisnih promenljivih, ako su sve pretpostavke ispunjene. Na osnovu veličine regresionih koeficijenata možemo zaključiti koliki je relativni uticaj ili važnost svake nezavisne promenljive ako se ti koeficijenti konvertuju u beta koeficijente β. Ovi koeficijenti se dobiju kada se sve vrednosti promenljivih standardizuju.

Jedna od pretpostavki za upotrebu regresione analize jeste postojanje linearne zavisnosti između varijabli. Ona je neophodna jer analiza započinje izračunavanjem koeficijenata proste korelacije (bivarijantnih korelacija) za sve parove varijabli, a sva ova izračunavanja zahtevaju linearan odnos između parova varijabli.

Višestruka regresija će biti prikazana na hipotetičkom primeru sa 7 nezavisnih varijabli. Prilikom korelacione analize, od posebnog interesa je određivanje stepena povezanosti između varijabli. Korelaciona analiza nam pruža sledeće:

- Relativnu važnost svake nezavisne varijable u predviđanju ili uticaju na zavisnu varijablu.
- Stepen do kojeg sve nezavisne varijable kombinovano objašnjavaju varijacije zavisne varijable.

Odgovore na ova pitanja dobijamo preko veličine standardizovanog regresionog koeficijenta β i koeficijenta proste korelacije r. U primeru sa 7 nezavisnih varijabli ovi pokazatelji su izračunati i prikazani u Tabeli 1.

Tabela 1

Varijabla	Koeficijent proste korelacije r	Standardizovani regresioni koeficijent β	Regresioni koeficijent b
1	2	3	4
X_1	0,63	0,55	2,89
X_2	0,52	0,27	10,41
X_3	0,40	0,15	6,62
X_4	0,21	0,17	1,32
X_5	0,11	−0,04	-5,08

X_6	0,06	0,22	3,44
X_7	0,03	0,01	4,45

U drugoj koloni koeficijenti proste korelacije pokazuju jačinu veze između svake nezavisne promenljive posebno sa zavisnom promenljivom Y. Ovaj koeficijent se kreće u intervalu od 0,03 do 0,63. Kada se ove vrednosti podignu na kvadrat dobijaju se koeficijenti determinacije koji objašnjavaju koliko data nezavisna varijabla ima udela u objašnjavanju varijacija nezavisne promenljive. Na primer, ako se prvi regresioni koeficijent podigne na kvadrat, dobija se da nezavisna promenljiva X_1 objašnjava 39,7% varijacija zavisne promenljive Y.

Multikolinearnost. Pošto se koeficijenti korelacije i beta koeficijenti uzimaju kao mere relativne važnosti svake nezavisne varijable, vrednosti u drugoj i trećoj koloni tabele 1 bi trebale da budu proporcionalne ili bar da opadaju isti redom. Međutim, vidi se da to nije slučaj. Razlog leži u multikolinearnosti ili prosto kolinearnosti. Multikolinearnost pokazuje kolika je međuzavisnost između nezavisnih varijabli. Što je veća multikolinearnost, to se više odražava na beta koeficijente i oni sve manje mogu da se upotrebe kao pokazatelji relativnog uticaja svake nezavisne varijable.

Razlog leži u tome što se regresioni koeficijenti, b i β, uvek izračunavaju tako da daju najbolje moguće predviđanje zavisne varijable Y, a ne da pokaže relativnu važnost svake nezavisne promenljive X. Kada je multikolinearnost mala i ne postoji onda su regresioni koeficijenti približno proporcionalni koeficijentima proste korelacije pa i jedni i drugi daju sličnu predstavu o relativnoj važnosti nezavisnih varijabli. Ako postoji značajna multikolinearnost, onda će najznačajnijoj nezavisnoj varijabli biti dodeljena prava vrednost beta koeficijenta, dok će kod ostalih nezavisnih beta vrednost biti mnogo manja da bi se izbegla međuzavisnosti i međusobni uticaj nezavisnih varijabli.

U tabeli 1, pošto se veličine beta koeficijenata nisu proporcionalne sa koeficijentima korelacije, može se zaključiti da postoji značajna multikolinearnost. Na primer, vidimo da najznačajnija nezavisna varijabla X_1 ima visok koeficijent korelacije i beta koeficijent, ali već X_2 ima nešto manji koeficijent korelacije ali duplo manju vrednost beta koeficijenta. Nesrazmera se ponavlja i kod drugih varijabli u modelu. To je zbog toga što se preklapa uticaj nezavisnih varijabli pa su zbog toga beta koeficijenti svih varijabli osim X_1 puno manji. Ovaj problem se može rešiti preko stepwise regresije.

Prihvatljivi nivo multikolinearnosti nije lako odrediti. On zavisi od broja nezavisnih varijabli u modelu, koliko njih je korelisano i u kom obimu. Potrebno je na početku izraditi tablicu prosti koeficijenata korelacije između svih varijabli. Prosti koeficijenti proste korelacije do 0,5 između nekoliko nezavisnih varijabli obično ne bi trebalo da utiču na regresione koeficijente. Ako su pomenuti koeficijenti proste korelacije veći od 0,7 onda je u pitanju ozbiljan problem.

Moguća rešenja su sledeća (Myers & Mullet, 2003, str. 89):

- Izraditi tabelu sa svim varijablama i njihovim koeficijentima proste korelacije. Ako kod nekog para varijabli koeficijent prelazi 0,7, onda se jedna od dve varijable eliminiše, obično ona koja ima manju korelaciju sa zavisnom varijablom Y.
- Ukoliko tri ili više nezavisnih varijabli imaju veliku međusobnu korelaciju, izabere se ona sa najvećom korelacijom sa Y i onda se eliminišu sve ostale ili se izradi nova zajednička varijabla od svih međuzavisnih varijabli (na osnovu vaganih vrednosti ili na osnovu proporcija u korelaciji sa Y).
- Izradi se analiza glavnih komponenti za sve nezavisne varijable. Ova tehnika traži grupu od dve ili više varijabli koje su visoko ili osrednje međusobno korelisane ali su istovremeno nepovezane sa ostalim varijablama. Za svaku od ovih grupa izrađuju se vrednosti koje se zovu faktor skorovi što je vrsta vaganih proseka. Pošto su ovi faktor skorovi nekorelisani i sadrže većinu informacija iz originalnih varijabli, oni mogu da se upotrebe kao novi set nezavisnih varijabli u višestrukom regresionom modelu. Ova opcija je najbolja i preporučuje se posebno ako je u pitanju veliki broj varijabli (preko 50). Ipak, ovim se gubi mogućnost da se posmatra svaka originalna varijabla pojedinačno.

Indeks determinacije. Višestruka regresija takođe pokazuje koliki je jaka međuzavisnost zavisne varijable sa svim nezavisnim varijablama preko indeksa korelacije R. Indeks determinacije R^2 pokazuje koliki je procenat varijabiliteta zavisne promenljive objašnje varijabilitetom nezavisnih promenljivih. U primeru iz tabele 1 indeks determinacije je 48% što je daleko od poželjne veličine od 70% . To znači da neke varijable koje imaju značajnu povezanost sa nezavisnom promenljivom Y nedostaju u modelu, ali nije poznato koje su to varijable. Pošto se indeks korelacije i indeks determinacije računaju na osnovu podataka koji su prikupljeni, dakle post-festum, ne može se ništa učiniti na njegovom poboljšanju. Ipak, u praksi se preporučuje da se prvo uradi pilot istraživanje gde se na manjem uzorku testira što veći broj

varijabli da bi se identifikovale sve one koje imaju najznačajniji uticaj, a zatim se uradi veliko posmatranje na kompletnom uzorku gde se prikupljaju podaci o tim varijablama.

Multikolinearnost se može utvrditi i preko specifičnih pokazatelja kao što je, na primer, nivo tolerancije. Nivo tolerancije je proporcija varijanse varijable koja nije povezana sa ostalim varijablama u regresionom modelu. Visok nivo tolerancije, preko 0,8 znači da je ta varijable relativno nekorelisana sa ostalim varijablama. Nizak nivo tolerancije, do 0,2 ukazuje na veliku multikolinearnost i da ta varijabla malo doprinosti objašnjavanju zavisne varijable u modelu.

Značaj višestruke regresije. Prema tome, na osnovu prethodno rečenog, višestruka regresija se koristi za dobijanje odgovora na sledeća pitanja:

- Koliko dobro sve nezavisne varijable kombinovano objašnjavaju ili im se može pripisati razlog za varijacije zavisne varijable (R^2).
- Kolika je relativna važnost svake nezavisne varijable u objašnjavanju varijacija zavisne varijable (beta koeficijenti), pod uslovom da ne postoji značajna multikolinearnost.
- Koja je najbolja predviđena vrednost zavisne varijablie za bilo koju kombinaciju nezavisnih varijabli.
- Koji se obim promene zavisne varijable može očekivati za svaku jedinicu promene svake nezavisne varijable (koeficijenti proste korelacije).

Pretpostavke na kojima se zasniva model višestruke regresije su slične onima koje važe za prostu regresiju i one glase:

- Oblik zavisnosti između svih varijabli je linearan odnosno prava linija. Ovo je pogotovo važno za odnos nezavisnih varijabli sa zavisnom varijablom.
- Sve varijable su kontinualne.
- Sve varijable imaju interval varijacije, disperziju odnosno varijansu koje imaju smisla, odnosno većina opservacija nije jedna vrednost ili interval.
- U bazi se nalazi barem tri do pet puta više jedinica posmatranja nego što je varijabli jer bi u suprotnom regresioni koeficijenti bili nepouzdani.
- Multikolinearnost između varijabli je mala ili ne postoji.

Testiranje statističke značajnosti. Pre objašnjavanja rezultata potrebno je testirati njihovu statističku značajnost. Ako R, b i β nisu statistički značajne, zaključuje se da nijedna nezavisna varijabla nema stvarnu povezanosti sa zavisnom varijablom. To znači da dobijeni model nema praktičnu vrednost. Većina statističkih softvera ima opciju testiranja.

Ukoliko su svi regresioni koeficijenti b statistički značajni, onda će i indeks korelacije R biti sigurno značajan. U obrnutom slučaju to ne mora da se desi jer je moguće da se zbog velikog broja varijabli dobije statistički značajno R a da b koeficijenti nisu značajni.

Vrednosti koje nedostaju. Često se dešava u praksi da neke vrednosti nedostaju u bazi podataka, odnosno da za neke jedinice posmatranja nije bilo moguće skupiti vrednosti za sve varijable. Na primer, neki ispitanici nisu želeli ili mogli da odgovore na sva pitanja iz upitnika. Ne postoji idealno rešenje, ali postoji nekoliko rešenja koja mogu da umanje ovaj problem:

- Eliminisanje jedinice posmatranje iz baze u celosti. Kada se radi analiza softver automatski izostavlja tu jedinicu. Problem u ovom slučaju jeste da postoji opasnost da se isključi veliki broj jedinica što se odražava na krajnji rezultat.
- Izračunavanje ocenjene vrednosti svake vrednosti koja nedostaje. Postoji više načina da se to uradi:
    - Umetanje na mesto nedostajuće vrednosti srednje vrednosti koja je izračunata za varijablu na osnovu celog uzorka.
    - Umetanje na mesto nedostajuće vrednosti srednje vrednosti koja je izračunata na osnovu svih vrednosti date jedinice posmatranja.
    - Na osnovu izračunate korelacije, umesto nedostajuće vrednosti umeće se vrednost varijable koja je jako korelisana sa varijablom za koju vrednost nedostaje.

Koliko vrednosti sme da nedostaje u celom skupu? Ne postoji tačan odgovor na ovo pitanje, ali se smatra da je prihvatljivo maksimalno do 10%. Neki smatraju da taj procenat može da ide do 15% - 20%, a ako udeo nedostajućih vrednosti prelazi 20% onda u jedinicu treba eliminisati iz analize.

Uniformno ocenjivanje. Još jedan problem koji može da se javi jeste kada za neku jedinicu posmatranja ne postoje varijacije u prikupljenim vrednostima varijabli. Na primer, ispitanik je na sva ili skoro sva pitanja odgovorio istom ocenom (na skali od 1 do 10 on je na sva pitanja zaokružio ocenu 5). Pošto u tom slučaju ne postoje variajcije za datu jedinicu posmatranja, ne dolazi do kovarijacije sa ostalim varijablama i jedinicama posmatranja. Povećava se samo veličina uzorka n ili skupa N ali se ne povećava kovarijansa. Na taj način se veštački snižava korelacija. Ni ovde ne postoji idealno rešenje. Ukoliko su sve vrednosti jednake bolje je takvu jedinicu eliminisati iz analize. Ukoliko je prisutan deo vrednosti koji se ponavlja za datu jedinicu posmatranja može se uraditi sledeće:

- Eliminisati jedinicu posmatranja kod koje ne postoji interval u vrednostima varijabli u dovoljnoj meri. Na primer na mernoj skali sa 10 vrednosti interval za tu jedinicu posmatranja su samo tri susedne vrednosti).
- Eliminisati jedinicu posmatranja kod koje postoji mali broj varijacija u odnosu na najčešću vrednost, na primer do 25% posmatranih varijabli.
- Izračunati standardnu devijaciju svih vrednosti varijabli za svaku jedinicu posmatranja i eliminisati one jedinice posmatranja kod kojih je izračunata vrednosti blizu nule.

Kategoričke vrednosti. U praksi se često dešava da nisu sve varijable izražene na metričkoj skali a da je potrebno izvesti regresionu analizu. Tipičan primer takvih varijabli bračni status, pol, profesija, stručna sprema, mesto stanovanja, država rođenja itd. Jedan način za rad sa takvim varijablama je njihovo prevođenje u kategoričke varijable (dummy variables) na sledeći način:

- Svaka kategorija (modalitet) se posmatra kao posebna nezavisna varijabla.
- Za svaku jedinicu posmatranja se dodeljuje vrednost „1“ ako jedinica poseduje neku karakteristiku a „0“ ako je ne poseduje. Na primer, kod bračnog statusa „1“ za „u braku“ i „0“ za „nije u braku“.
- Nove varijable se ubacuju u regresioni model, ali tako da jedna kategorija iz svake originalne varijable mora biti isključena iz analize.

Razlog za ovo isključivanje je da se izbegne da vrednosti varijable budu međusobna linearna kombinacija. Na primer, ako imamo četiri različita bračna statusa (samac, u braku, razveden(a), udovac-udovica) onda mora jedna kategorija da ima vrednost nula i da bude isključena iz računa. Neki softveri to rade računski ako se varijabla na početku definiše kao kategorička (dummy). Ako želimo da uključimo samce u naš regresioni model, kod bračnog statusa sa četiri modaliteta imali bi četiri varijable obeležene na sledeći način:

Varijable	Unete vrednosti
Samac	1
U braku	0
Razveden(a)	0
Udovac-udovica	0

Stepwise regresija

Višestruka regresija nam daje model u koji su uključene sve varijable sa kojima je analiza i započeta, bez obzira na njihov različiti značaj, a takođe i u slučaju kada je prisutna velika multikolinearnost. Stepwise regresija nam omogućuje da se izborimo sa problemom multikolinearnosti i sa nezavisnim varijablama koje su od malog značaja.

Kada je multikolinearnost velika, onda mnoge varijable imaju slično značenje, pa nije potrebno da sve one budu uključene u model. Stepwise regresija omogućava da se eliminišu varijable koje se preklapaju sa drugima i zbog toga malo ili uopšte ne doprinose tačnosti u predviđanju modela. Kao rezultat ovog pristupa dobija se novi model sa manjim brojem nezavisnih varijabli koji je isto toliko dobar koliko i model u kojem se nalaze sve nezavisne varijable.

Tipični tok stepwise regresije se odvija na sledeći način (Myers & Mullet, 2003, str. 92):

1. Računar izabere jednu nezavisnu varijablu koja ima najveću korelaciju sa zavisnom varijablom.
2. Računar bira između ostalih varijabli onu koja najviše doprinosi tačnosti predviđanja prvoj koja je izabrana. Ovaj korak se izvodi sve dok ne ostane ni jedna varijabla koja bi doprinela još više tačnosti modela.

3. Pri svakom koraku izračunava se test statističke značajnosti za onaj nivo predviđanja koji dodaje nova varijabla. Ako je taj nivo predviđanja ispod značajnosti koju je unapred odredio analitičar, ta varijabla se isključuje iz modela.
4. Računar daje finalni regresioni model sa b koeficijentima. Ako je multikolinearnost bila visoka, model će imati manje varijabli u odnosu na originalni model.