Logistička regresija

Višestruka linearna regresija se koristi kada su u pitanju varijable čije su vrednosti kontinualne ili metričke (intervalne vrednosti ili racio). U praksi se često dešava da je zavisna varijabla dihotomna, odnosno da može da uzima samo dve vrednosti, na primer:

- muški ili ženski pol
- kupljena roba A ili roba B
- kupovina je obavljena ili nije obavljena
- proizvod je ispravan ili neispravan

Logistička regresija je pogodna za rešavanje problema kada su u pitanju demografske varijable jer su one uglavnom kategoriče (bračni status, zanimanje, lokacija itd.). Ona je posebno uspešna ako je kategorička varijabla kao zavisna promenljiva sa jakom asimetrijom ili ako ima nelinearnu relaciju sa ostalim varijablama.

Problemi ove vrste se mogu rešiti i preko višestruke linearne regresije tako što bi dve vrednosti varijable obeležili sa dva cela broja, obično sa 0 i 1. Dobili bi smo regresioni model koji bi mogao da predvidi vrednost zavisne varijable, zajedno sa regresionim koeficijentima koji bi pokazivali relativni uticaj svake nezavisne varijable. Ipak, logistička odnosno logit regresija je adekvatnije rešenje.

Sa aspekta predviđanja, želimo da znamo u kojoj od dve moguće grupe spada svaki ispitanik, odnosno jedinica posmatranja (muškarac ili žena, kupljena roba A ili B itd.). Preko višestruke linearne regresije dobićemo rešenje u kojem će zavisna promenljiva imati vrednost negde između 0 i 1. Predviđena vrednost će izgledati kao verovatnoća da će jedinica posmatranja pripasti jednoj ili drugoj grupi. Na primer, ako je sa „0“ obeležen slučaj kupovine robe A a sa „1“ kupovina robe B, a vrednost zavisne promenljive (kupovine) iznosi 0,65 za datog kupca, onda ispada da je veća verovatnoća da će kupac kuputi robu B jer je vrednost bliža jedinici. Pretpostavka je da se kod višestruke linearne regresije dobijena vrednost zavisne promenljive u takvim slučajevima može tretirati kao verovatnoća.

Ipak, problem koji se često javlja jeste da se preko višestruke linearne regresije dobiju vrednosti zavisne promenljive koje su manje od nule a veće od jedinice (na primer, -0,2 ili 1,3). Pošto se ovakve vrednosti ne mogu tumačiti kao verovatnoće, postaje jasno da pomenuti model nije dobro rešenje.

Potrebno je izvršiti određenu vrstu matematičke transformacije zavisne varijable da bi se dobio logistički regresioni model. Transformacija se izvodi tako što se vrednosti zavisne promenljive pomnože sa tzv. prirodnim logaritmom proporcije (natural logarithm of the odds ratio). U pitanju je veoma složena transformacija, ali na sreću za razumevanje logističke regresije i njeno izvođenje uz pomoć statističkog softvera njeno poznavanje nije neophodno.

Osnovna obeležja logističke regresije su sledeća:

- Logistički model treba da se koristi kada je zavisna varijabla dihotomna.
- Zavisna varijabla mora da bude transformisana multiplikacijom sa prirodnim logaritnom proporcije (logit transformacija).
- Nezavisne varijable moraju da budu kontinualne i da su u međusobnoj linearnoj zavisnosti, kao i kod obične regresije. Kategoričke varijable se takođe mogu koristiti (dummy variables).
- Logit i model razvijen metodom najmanjih kvadrata (optimal least squares – OLS) su linearno aditivni modeli.
- Kod predviđanja, logistički model izražava verovatnoću da će data jedinica posmatranja upasti u jednu dihotomnu grupu umesto u drugu.
- Regresioni koeficijenti kod logističke regresije mogu da se interpretiraju kao kod obične regresije – što veći koeficijent to je veći uticaj nezavisne promenljive na zavisnu, pod pretpostavkom male i nikakve kolinearnosti.
- Obična OLS regresija može da se koristi ako je zavisna varijabla dihotomna i ne postoji velika razlika između regresionih koeficijenata OLS i logit regresije. Ipak, veća preciznost se postiže primenom logit regresije.

Model logističke regresije

Matematička transformacija se izvodi tako da se kao rezultat zavisne promenljive uvek dobije vrednost između 0 i 1. Ove vrednosti se tada mogu interpretirati kao verovatnoća da određena jedinica posmatranja pripada jednoj od dve grupe. Transormacija koja se izvodi je sledeća:

gde je:

ln – prirodni logaritam (stepen na koji konstanta *e* (2,72) mora biti podignuta)

p – proporcija uzorka koja pripada jednoj od kategorija zavisne promenljive *Y*

$1-p=q$ – proporcija koja pripada drugoj kategoriji

$X_1 - X_k$ – nezavisne promenljive.

Ova transformisana jednačina mora da se preuredi da bi na levoj strani jednačine bilo iskazano samo *p*, odnosno verovatnoća:

Kao i kod obične regresije, potrebna nam je mera koja će pokazati koliko je tačan logistički model koji je izračunat. Kod obične regresije, ta mera je indeks determinacije R^2, koji pokazuje koji procenat varijacija zavisne varijable je objašnjen nezavisnim varijablama koje se nalaze u modelu. Ne postoji slična mera kod logističkog modela. Najčešći pristup jeste da se izračunava tačnost predviđanja korišćenjem proste proporcije:

gde je:

T – tačnost predviđanja

m – broj tačnih predviđanja

n – ukupan broj jedinica posmatranja

Ako je predviđena verovatnoća veća od 0,5 zaključuje se da jedinica posmatranja pripada grupi koja ima vrednost zavisne promenljive 1, a ako je manja od 0,5 onda pripada grupi koja ima vrednost zavisne promenljive 0. Nakon određivanja modela moguće je izračunati verovatnoću za nove jedinice posmatranja o tome kojoj grupi pripadaju.

Testiranje dobijenog modela može da se izvede tako što se uzorak podeli na dva dela, ako je dovoljno velik, pa se izradi logistički model na jednoj polovini a zatim se primeni na jedinicama iz druge polovine.

Kada su u pitanju ostali aspekti regresione analize, za logističku regresiju oni su identični kao i kod obične regresije:

- ocenjivanje statističke značajnosti regresionih koeficijenata
- interpretacija regresionih koeficijenata
- sirove naspram standardizovane vrednosti regresionih koeficijenata
- kolinearnost
- rešavanje problema vrednosti koje nedostaju, uniformne vrednosti, kategoričke varijable
- stepwise regresija.

Varijable sa više od dve kategorije. Logistički model može da posluži i za varijable koje imaju više od dve kategorije. Na primer:

- kupovina proizvoda A, B, C ili D

- putovanje vozom, automobilom ili autobusom
- odluka investitora da investira u jednoj od pet opština.

Logistički model je u osnovi binomni model i on je pogodan za dve kategorije, ali se može primeniti i u sledećoj situaciji. Na primer, ako imamo četiri proizvoda, izrade se zasebno četiri logistička modela (proizvod A naspram svih ostali, B naspram svih ostalih itd.). Na taj način ćemo dobiti verovatnoću za kupovinu svakog proizvoda pojedinačno. Osim toga, verovatnoća za poslednji proizvod može da se dobije na još jednostavniji način, tako što se od jedinice oduzmu verovatnoće za tri prethodna proizvoda.

Primer: Logistička regresiona analiza

Podaci za navedeni primer se nalaze u Statistica spreadsheet-u pod nazivom „Primeri za knjigu.stw".

Kupac se odlučuje da kupi određeni proizvod na osnovu četiri njegove karakteristike (X_1, X_2, X_3 i X_4). Varijabla „Izbor" nam pokazuje da li se dati kupac odlučio da kupi proizvod (1) ili da ga ne kupi (2). Uzorak broji ukupno 60 kupaca od kojih se 30 odlučilo da kupi proizvod a 30 da ga ne kupi. Dobijen je sledeći regresioni model:

Model: Logistic regression (logit) N of 0's: 30 1's: 30 (Spreadsheet9 in Primeri za knjigu.stw) Dep. var: Izbor Loss: Max likelihood (MS-err. scaled to 1) Final loss: 27,400452727 Chi2(4)=28,377 p=,00001

	Const.B0	X1	X2	X3	X4
Estimate	3,603776	0,1963255	-0,1382724	-0,1623809	0,08056165
Standard Error	13,66688	0,0759929	0,08115838	0,07023905	0,09781007
t(55)	0,2636867	2,583472	-1,703735	-2,311832	0,823654
p-value	0,7930068	0,01246891	0,0940751	0,02455811	0,4136935
-95%CL	-23,78527	0,04403235	-0,3009174	-0,3031431	-0,1154541
+95%CL	30,99282	0,3486187	0,02437263	-0,02161871	0,2765774
Wald's Chi-square	0,06953069	6,674327	2,902714	5,344569	0,6784059
p-value	0,792023	0,009785533	0,08844025	0,02079324	0,4101423
Odds ratio (unit ch)	36,73668	1,216923	0,8708614	0,8501173	1,083896
-95%CL	0,00000000004679375	1,045016	0,7401389	0,7384934	0,8909615
+95%CL		1,417109	1,024672	0,9786133	1,318609
Odds ratio (range)		296,9134	0,04773966	0,004707632	3,089049
-95%CL		3,585644	0,001333187	0,00004523121	0,1986209
+95%CL		24586,25	1,709495	0,4899669	48,04239

U redu „Estimate" navedene su vrednosti regresionih koeficijenata. Crvenom bojom su označene one varijable koje su statistički značajne. p-vrednost (p-value) jasno pokazuje zašto su varijable X_1 i X_2 statistički značajne, jer su manje od $\alpha = 0,05$. Tačnost modela može da se vidi na osnovu sledeće tabele:

Classification of Cases (Spreadsheet9 in Primeri za knjigu.stw) Odds ratio: 16,000 Perc. correct: 80,00%

	Pred. - 1,000000	Pred. - 2,000000	Percent - Correct
1,000000	24	6	80,00000
2,000000	6	24	80,00000

U slučaju kupovine model je tačno prepoznao 80% kupaca, a takođe i u slučaju kada proizvod nije kupljen. Može se reći da model ima dobru moć predviđanja.

Logistička regresija u statističkom paketu Statistica

1. varijanta (jednostavnija procedura)

Pokretanje analize:

Statistics ▸ Advanced Linear/Nonlinear Models ▸ Nonlinear Estimation ▸ Quick Logit regression ▸ OK

Definisanje varijabli:

▸ Variables

Otvara se prozor sa spiskom varijabli od kojih treba odabrati one koje će biti uvrštene u analizu. U polje "Dichotomous dependent variable" se unosti zavisna promenljiva a u polje "Continuous independent variable list" se unose nezavisne varijable. Bitno je da zavisna varijabla bude definisana kao kategorička varijabla prilikom njenog definisanja i unosa podataka.

▸ OK ▸ OK ▸ Advanced

U prozoru "Estimation method" izabrati "Rosenbrock and quasi-Newton". Označiti opciju "Asymptotic standard errors".

▸ OK

Analiza je urađena a model se može pogledati pod sledećom opcijom:

▸ Summary: Parameters & standard errors

Tačnost predviđanja se može pogledati pod sledećom opcijom:

▸ Residuals ▸ Classification of cases & odds ratio.

U modulu postoji čitav niz opcija koje pružaju ostale rezultate analize.

2. varijanta

Ova varijanta omogućava da se uradi i stepwise regresija odnosno da se izvrši selekcija samo onih varijabli koje statistički značajno objašnjavaju zavisnu promenljivu.

Pokretanje analize:

Statistics ▸ Advanced Linear/Nonlinear Models ▸ Generalized Linear/Nonlinear Models ▸ Logit model ▸ OK

Definisanje varijabli:

▸ Variables

Otvara se prozor sa spiskom varijabli od kojih treba odabrati one koje će biti uvrštene u analizu. U polje "Dependent variable" se unosti zavisna promenljiva a u polje "Continuous predictors" se unose nezavisne varijable. Bitno je da zavisna varijabla bude definisana kao kategorička varijabla prilikom njenog definisanja i unosa podataka.

▸ OK ▸ Advanced

Bira se jedna od opcija za stepwise regresiju, najčešće "Forward stepwise".

▸ OK

Analiza je urađena a model se može pogledati pod sledećom opcijom:

▸ Estimates

Tok stepwise regresije i konačni model se mogu pogledati pod sledećom opcijom:

▸ Model building

Tačnost predviđanja se može pogledati pod sledećom opcijom:

▸ Resid. 1 ▸ Class & odds ratio.

U modulu postoji čitav niz opcija koje pružaju ostale rezultate analize.