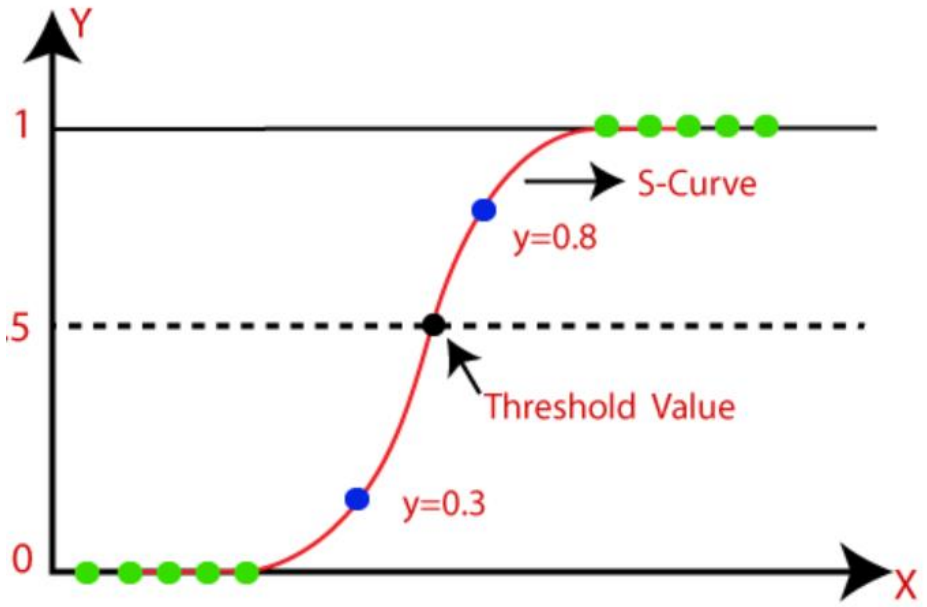
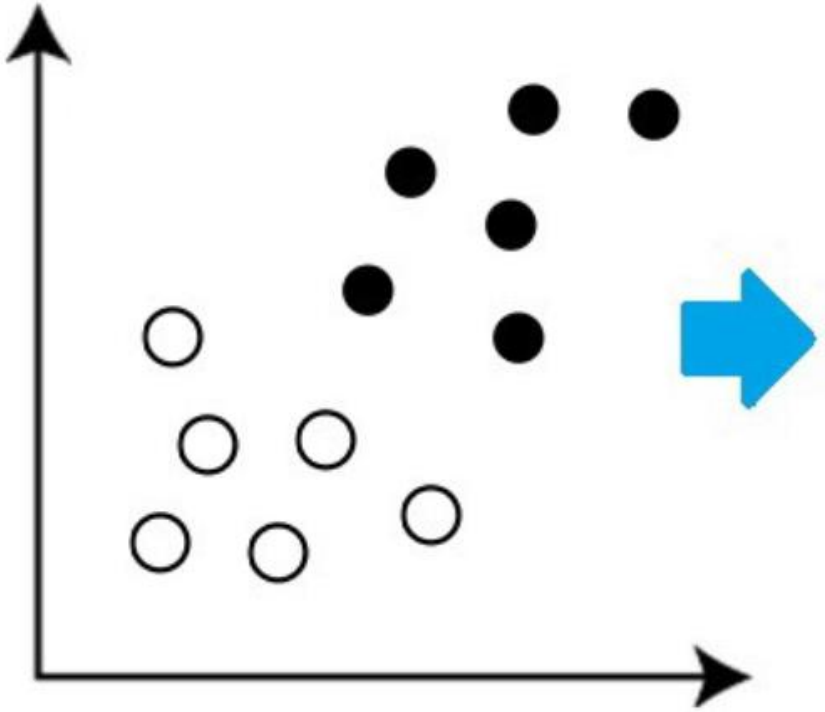




Logistična regresija

Doc. dr. Petra Povalej Bržan

Namen logistične regresije

- Napovedovanje kategorične (nominalne) spremenljivke na osnovi enega ali več numeričnih ali nominalnih napovednikov(prediktorjev).
- Primeri:
  - Napovedovanje bolezni (diagnoze) – pozitivna ali negativna (sladkorna bolezen)
  - Napovedovanje uspešnosti terapije – uspešna ali neuspešna
  - Napovedovanje odločitve za smer študija – medicina / tehnične vede / družboslovne vede

- BINARNA (dve kategoriji v odvisni spremenljivki)
- MULTINOMIALNA (več kategorij v odvisni spremenljivki)

Kaj je razmerje obetov (OR)?

Contingency Tables

	SKUPINA		
MCV_disk	alkoholik	nealkoholik	Total
patološki	357	75	432
normalen	62	191	253
Total	419	266	685

$$odds_{alkoholik} = \frac{število\ alkoholikov}{število\ nealkoholikov} = \frac{419}{266} = 1{,}58$$

$$odds_{alkoholik\ ob\ patološkem\ MCV} = \frac{število\ alkoholikov\ s\ patološkim\ izvidom}{število\ nealkoholikov\ s\ patološkim\ izvidom} = \frac{357}{75} = 4{,}76$$

$$odds_{alkoholik\ ob\ normalnem\ MCV} = \frac{število\ alkoholikov\ z\ normalnim\ izvidom}{število\ nealkoholikov\ z\ normalnim\ izvidom} = \frac{62}{191} = 0{,}32$$

$$OR = \frac{odds_{dogodka\ ob\ prisotnosti\ dejavnika\ tveganja}}{odds_{dogodka\ ob\ odsotnosti\ dejavnika\ tveganja}}$$

$$OR = \frac{odds_{alkoholik\ ob\ patološkem\ MCV}}{odds_{alkoholik\ ob\ normalnemMCV}} = \frac{4{,}76}{0{,}32} = 14{,}66$$

OR>1 => če se prediktor poveča, se verjetnost, da se bo dogodek zgodil poveča

OR<1 => če se prediktor poveča, se verjetnost, da se bo dogodek zgodil zmanjša

OR=1 => če se prediktor poveča, to ne spremeni verjetnosti, da se bo dogodek zgodil

Enačba modela

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right)= b_0 + b_1x_1 \qquad \ln\left(\frac{P(Y)}{1-P(Y)}\right)=b_0 + b_1x_1+ \ldots +b_nx_n$$

b_0 – log odds če so vsi prediktorji enaki 0 (npr. log odds če je izvid=0 – normalen)
Z exp(b0) se znebimo log odds
Exp(b) – razmerje obetov(odds) da je oseba alkoholik, če

$$b_0 = odds_{alkoholik\ ob\ normalnem\ MCV} = \frac{število\ alkoholikov\ z\ normalnim\ izvidom}{število\ nealkoholikov\ z\ normalnim\ izvidom} = \frac{62}{191}=0{,}32$$

$$Exp(b) = \frac{odds\ po\ spremembi\ prediktorja\ za\ eno\ enoto}{originalni\ odds}$$

Kdaj uporabimo logistično regresijo?

- Če je **Y zvezna** spremenljivka (npr. teža otroka ob porodu) => **Linearna regresija** (predpostavka: obstaja linearna povezava med neodvisnimi in odvisno spremenljivko)

- Če je **Y binarna spremenljivka** (npr. mrtev/živ, da/ne, bolan/zdrav…) => **Binarna logistična regresija**

Primer 1

- Baza: alkoholiki.csv
- Namen: Želimo izdelati model, ki bo kar se da natančno napovedoval ali je oseba alkoholik ali ne na osnovi krvnega markerja MCV.
- Odvisna spremenljivka: Skupina (1-alkoholik (preiskovana), 0-nealkoholik (kontrolna)) - dihotomna
- Neodvisna spremenljivka: MCV_disk - dihotomna

$$Skupina = b_0 + b_1 * MCV\_disk$$

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = b_0 + b_1 * MCV\_disk$$

Kodiranje nominalnih spremenljivk

Odvisna spremenljivka:

1- dogodek (alkoholik)

0- ni dogodka (nealkoholik)

SKUPINA

Filter	Value	Label
✓	0	nealkoholik
✓	1	alkoholik

Neodvisna spremenljivka MCV_disk

0 – kontrolna skupina (MCV nepatološki)

1 – testna skupina (mcv patološki)

MCV_disk

Filter	Value	Label
✓	0	normalen
✓	1	patološki

Koeficienti modela

Coefficients

				Wald Test		
	Estimate	Standard Error	z	Wald Statistic	df	p
(Intercept)	−1.125	0.146	−7.698	59.254	1	< .001
MCV_disk	2.685	0.194	13.867	192.306	1	< .001

Note. SKUPINA level 'alkoholik' coded as class 1.

- Estimate – koeficienti modela (b_0 in b_1)
- Potrebujemo Exp(b) za lažjo interpretacijo modela. (Statistics ->Regression Coefficients ->Odds ratios)

Exp(b) – odds ratio

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−1.125	0.146	0.325	−7.698	59.254	1	< .001
MCV_disk	2.685	0.194	14.664	13.867	192.306	1	< .001

Note. SKUPINA level 'alkoholik' coded as class 1.

- Če se spremeni MCV_disk iz 0 (nepatološko) v 1 (patološko), se verjetnost za dogodek (oseba je alkoholik) poveča OR=14,664
- Exp(b0) – pomeni, da je verjetnost za dogodek v primeru, da je prediktor=0 (nepatološki MCV izvid) OR=0,325

Zapis rezultatov

Confusion matrix

	Predicted	
Observed	nealkoholik	alkoholik
nealkoholik	191	75
alkoholik	62	357

Performance metrics ▼

	Value
AUC	0.785
Sensitivity	0.852
Specificity	0.718

S pomočjo logistične regresije smo preverili ali obstaja povezava med patološko vrednostjo MCV in alkoholizmom. Neodvisna spremenljivka MCV izvid je bila testirana z namenom, da se preveri predpostavka linearnosti logit modela. Ugotovili smo, da spremenljivka MCV izvid statistično značnilno prispeva k modelu (b=2,685; β=14,664; p<0,001). Senzitivnost modela znaša 85,3%, specifičnost pa 71,8%. Področje pod ROC krivuljo (AUC) znaša 78,5%.

Poskusimo izboljšati model – dodamo spremenljivko GLDH_disk in interakcijo

$$Skupina = b_0 + b_1 * MCV\_disk + b_2 * GLDH\_disk + b_3 * interakcija$$

GLDH_disk

Filter	Value	Label
✓	0	normalen
✓	1	patološki

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = b_0 + b_1 * MCV\_disk + b_2 * GLDH\_disk + b_3 * interakcija$$

Ocena novega modela

- Preverimo še klasifikacijsko tabelo (Confusion matrix / crosstabs)

Performance Diagnostics

Confusion matrix

Observed	Predicted	
	nealkoholik	alkoholik
nealkoholik	153	91
alkoholik	28	265

Confusion matrix ▼

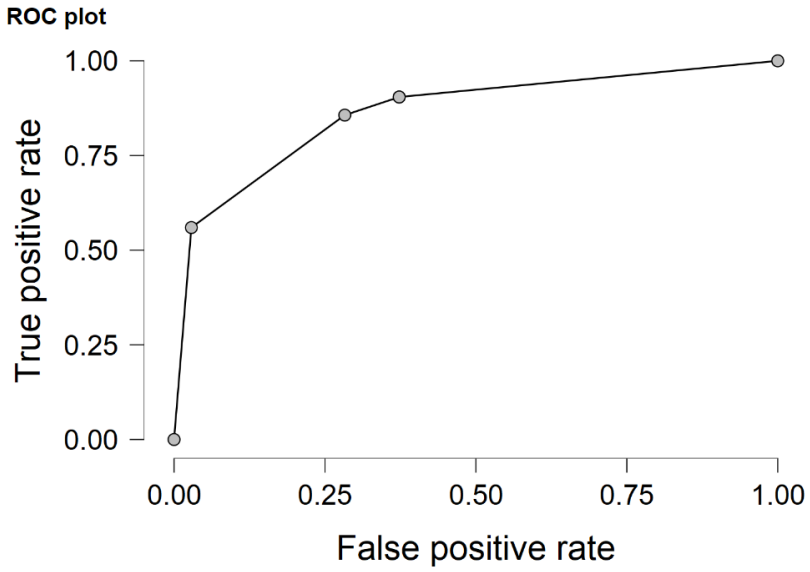
Observed	Predicted	
	nealkoholik	alkoholik
nealkoholik	0.285	0.169
alkoholik	0.052	0.493

Proportions

Ocena modela s pomočjo Senzitivnosti, specifičnosti in AUC

Performance metrics

	Value
AUC	0.864
Sensitivity	0.904
Specificity	0.627



Koeficienti modela

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−1.939	0.203	0.144	−9.555	91.304	1	< .001
MCV_disk	2.425	0.238	11.299	10.205	104.136	1	< .001
GLDH_disk	2.077	0.260	7.981	7.981	63.696	1	< .001

Note. SKUPINA level 'alkoholik' coded as class 1.

- Ugotavljamo, da se tako ob patološkem rezultatu izvida MCV, kot tudi ob patološkem izvidu GLDH, statistično značilno poveča verjetnost za to, da je oseba alkoholik. Interpretiramo Odds Ratio – Exp(b).

Lahko dodajmo v model še interakcijo med neodvisnima spremenljivkama

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−1.698	0.206	0.183	−8.262	68.260	1	< .001
MCV_disk	2.037	0.264	7.668	7.706	59.384	1	< .001
GLDH_disk	1.246	0.399	3.477	3.124	9.760	1	0.002
MCV_disk * GLDH_disk	1.569	0.579	4.802	2.708	7.334	1	0.007

- Interakcija ja prav tako statistično značilna!

Zapis rezultatov

Coefficients

	Estimate	Robust Standard Error	Odds Ratio	z	Wald Test Wald Statistic	df	p
(Intercept)	−1.698	0.206	0.183	−8.262	68.260	1	< .001
MCV_disk	2.037	0.264	7.668	7.706	59.384	1	< .001
GLDH_disk	1.246	0.399	3.477	3.124	9.760	1	0.002
MCV_disk * GLDH_disk	1.569	0.579	4.802	2.708	7.334	1	0.007

Note. SKUPINA level 'alkoholik' coded as class 1.

Confusion matrix

Observed	Predicted: nealkoholik	Predicted: alkoholik
nealkoholik	0.326	0.128
alkoholik	0.078	0.467

Performance metrics ▼

	Value
AUC	0.864
Sensitivity	0.857
Specificity	0.717

S pomočjo logistične regresije smo preverili ali obstaja povezava med patološko vrednostjo MCV ter GLDH in alkoholizmom. Preverili smo predpostavke o linearnosti in neodvisnosti. Ugotovili smo, da spremenljivka MCV izvid statistično značnilno prispeva k modelu (b=2,037; β=7,668; p<0,001). Prav tako statistično značilno vpliva na model izid izvida GLDH (b=1,246; β=3,477; p=0,002). Interakcija obeh spremenljivk prav tako statisticno značilno vpliva na model (b=1,569; β=4,802; p=0,007). Senzitivnost modela znaša 85,7%, specifičnost pa 71,7%. Področje pod ROC krivuljo (AUC) znaša 86,4%.