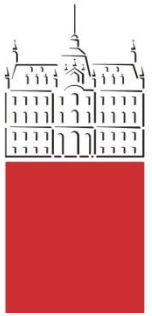
Univerza v Ljubljani
Fakulteta za farmacijo



Logistična regresija

Izbrana poglavja iz biomedicinske informatike 2011/2012, 2.letnik LBM2

Asist. dr. Igor Locatelli, *mag. farm.*

Ljubljana, 9. 12. 2011

Logistična regresija

- Kaj je že (linearna) regresija?
- Kaj pa multipla linearna regresija?
- Kaj pa posplošeni linearni modeli (ang. GLM)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- Pri binarni logistični regresiji je odvisna spremenljivka dihotomna (vrednosti 0 ali 1).
  - Napovedujemo verjetnost za izid 1.
- Multinominalna logistična regresija.
- Verjetnosti imajo zalogo vrednosti med 0 in 1. Kako uporabiti regresijski model?

Obeti in logit

- o = p/(1-p)

Primer: če verjetnost za nek dogodek enaka 75%, potem je obet za ta dogodek enak 0,75/0,25 = 3

Prednost: zaloga vrednosti obeta je med 0 in ∞

- Logit(p) = log (p/(1-p))

Če obete logaritmiramo, dobimo vrednosti med $-\infty$ in ∞ (Zakaj?)

V logistični regresijskem modelu uporabimo logit(p) namesto p, ta trasformacija nam omogoča poenostavitev v GLM.

Nekaj enačb

$$\ln \frac{P(Y=1 \mid x)}{1-P(Y=1 \mid x)} = \text{logit}\, P(Y=1) = \beta_0 + \beta_1 x$$

$$P(Y=1 \mid x) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)}$$

Multipla logistična regresija:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i})}}.$$

Kaj pomenita koeficienti a in b oz. β_0 in β_1?

Lin. regresija: $b = y_{(x+1)} - y_{(x)}$

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = a + bx = \log it$$

Log. regresija: $b = \log it_{(x+1)} - \log it_{(x)}$

Binarni X (0/1) => x=0, x+1=1.

Obet da bo y=1:

$$O_{x=1} = \frac{p(1)}{1-p(1)} \qquad O_{x=0} = \frac{p(0)}{1-p(0)}$$

Logaritmiranje:

$$\log it_{x=1} = \ln\left[\frac{p(1)}{1-p(1)}\right] \qquad \log it_{x=0} = \ln\left[\frac{p(0)}{1-p(0)}\right]$$

Razmerje obetov

Log. razmerja obetov:

$$RO = \frac{O_{x=1}}{O_{x=0}} = \frac{\left[\frac{p(1)}{1-p(1)}\right]}{\left[\frac{p(0)}{1-p(0)}\right]}$$

$$O_{x=1} = \frac{p(1)}{1-p(1)}$$

$$O_{x=0} = \frac{p(0)}{1-p(0)}$$

$$\ln RO = \ln \frac{\left[\frac{p(1)}{1-p(1)}\right]}{\left[\frac{p(0)}{1-p(0)}\right]} = \log it(1) - \log it(0)$$

$$\ln RO = b = \log it_{(1)} - \log it_{(0)}$$

$$RO = e^{b}$$

Ocenjevanje parametrov a in b

- p(x)? =>b in a ?

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = a + bx = \log it$$

- Lin. regresija: metoda najmanjših kvadratov ostankov.
- Log. regresija: metoda največjega verjetja (maximum likelihood method).
  - Funkcija največjega verjetja
  - Oz. logaritem funkcije verjetja ("log likelihood"):
    - Nelinearna funkcija parametrov modela a in b.

Interval zaupanja za b

Interval zaupanja za RO iz b (spodnjo/zgornjo mejo):

$$e^{b \pm z^* SE_b}$$

Multipla logistična regresija

- Matematični model:

$$p(x) = \frac{e^{a+b_1x_1+b_2x_2+\ldots+b_kx_k}}{1+e^{a+b_1x_1+b_2x_2+\ldots+b_kx_k}}$$

- b(k): sprememba v logitu, ki spremlja spremembo X(k) za 1 enoto, medtem ko se ostale sprem. ne spreminjajo.
- Regresijski koef. b = logaritem razmerja obetov
- Razmerje obetov= antilogaritem b: e^b
- Metoda največjega verjetja

Goodness of fit

- Models are compared by taking 2 times the difference between the models log-likelihoods.

$$\chi^2 = 2[(\text{log-likelihood for bigger model}) - (\text{log-likelihood for smaller model})]$$

Note: models must be nested in order to be compared. Nested means that all components of the smaller model must be in the larger model.

Verjetje (*l*) in logaritem verjetja (L)

$$l = \prod_{i=1}^{n} \pi_i^{w_i y_i} \left(1 - \pi_i\right)^{w_i (1 - y_i)}$$

$$L = \ln(l) = \sum_{i=1}^{n} \left(w_i y_i \ln(\pi_i) + w_i (1 - y_i) \ln(1 - \pi_i) \right)$$

y opazovane vrednosti za dogodek
π verjetnost za dogodek

Testiranje pomembnosti modela kot celote

- Ali se log verjetja modela s spremenljivkami X statistično značilno poveča v primerjavi z log verjetja modela brez njih.
- G- statistika- test razmerja dveh verjetij:

$$G = 2\ln\left[\frac{verjetje_{SPREM.X}}{verjetje_0}\right]$$

- Porazdeljuje po hi- kvadrat, df: k-1
- k: št. vseh spremenljivk v modelu
- V SPSS-u: "Omnibus test"

Vrednotenje prileganja modela

- Nagelkerkejev R^2

- Test hi- kvadrat
- Hosmer- Lemeshov test:
  - Oba primerjata opazovano število enot, pri katerem se je opazovani dogodek zgodil, s pričakovanim številom, ki temelji na enačbi logistične regresije.

Testiranje pomembnosti b

- Ho: b=0, Ha: b≠ 0
- Testi:
  - z: $$z = \frac{b - 0}{SE_b}$$
  - Waldova statistika: $$\chi^2 = \left(\frac{b}{SE_b} \right)^2$$
  - G- statistika: $$G = 2 \ln \left[\frac{verjetje_x}{verjetje_0} \right]$$

G- statistika: Model, ki vključuje X, pove več o Y.

Opisne sprem.v regresiji

- Opisne sprem. dve vrednosti => kodiranje 0/1
- Linearna regresija:
  - Enačbi za pričakovani povprečni vrednosti Y:

$$\overline{y}_{x=0} = a + b_1 * 0 = a$$

$$\overline{y}_{x=1} = a + b_1 * 1 = a + b_1$$

- Podobno analogija tudi za log. regresijo

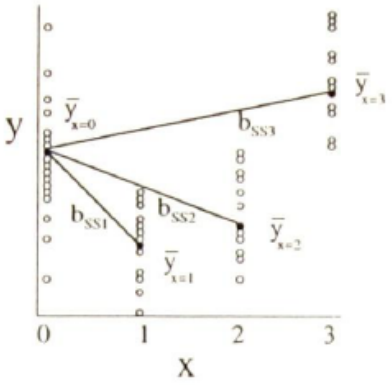
Slepe spremenljivke ("dummy var.")

- Opisne sprem. >2 vrednosti
- Podobno kot pri binarni sprem.?!
- Nove sprem.: vrednosti 0/1

Kodiranje slepih spremenljivk

- Primerjalna kategorija (nima svoje SS)
- b(ss): razlika med aritm. sredino Y pri opazovani kategoriji in aritm. sredino primerjalna kategorije oz. konstanto a
- Za nominalne opisne sprem. z več kategorijami

	slepe spremenljivke			enačba
kategorija osn. sprem.	SS1	SS2	SS3	
0	0	0	0	$\bar{y}_0 = a$
1	1	0	0	$\bar{y}_1 = a + b_{SS1}$
2	0	1	0	$\bar{y}_2 = a + b_{SS2}$
3	0	0	1	$\bar{y}_3 = a + b_{SS3}$
reg. koef.	b_{SS1}	b_{SS2}	b_{SS3}	

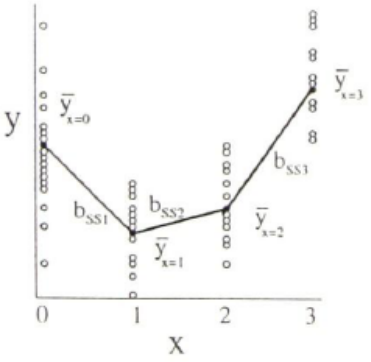


Slika 12

Sekvenčno kodiranje slepih spremenljivk

- Osnovna kategorija (nima svoje SS)
  - Ni več vedno primerjalna
- b(ss): razlika med aritm. sredino Y pri opazovani kategoriji in aritm. sredino prejšnje kategorije
- Za opisne sprem. ordinalnega tipa

	slepe spremenljivke			enačba
kategorija osn. sprem.	SS1	SS2	SS3	
0	0	0	0	$\bar{y}_0 = a$
1	1	0	0	$\bar{y}_1 = a + b_{SS1}$
2	1	1	0	$\bar{y}_2 = a + b_{SS1} + b_{SS2}$
3	1	1	1	$\bar{y}_3 = a + b_{SS1} + b_{SS2} + b_{SS3}$
reg. koef.	b_{SS1}	b_{SS2}	b_{SS3}	



Slika 13

Volitve za predsednika US 1992

Logistic Regression

Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	1847	100,0
	Missing Cases	0	,0
	Total	1847	100,0
Unselected Cases		0	,0
Total		1847	100,0

a If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
no	0
yes	1

Kodiranje kategoričnih spremenljivk v modelu

Categorical Variables Codings

		Frequency	Parameter coding			
			(1)	(2)	(3)	(4)
RS HIGHEST DEGREE	lt high school	189	,000	,000	,000	,000
	high school	947	1,000	,000	,000	,000
	junior college	132	,000	1,000	,000	,000
	bachelor	386	,000	,000	1,000	,000
	graduate degree	193	,000	,000	,000	1,000
age categories	lt 35	438	,000	,000	,000	
	35 - 44	444	1,000	,000	,000	
	45 - 64	617	,000	1,000	,000	
	65 +	348	,000	,000	1,000	
RESPONDENTS SEX	male	804	,000			
	female	1043	1,000			

Block 1

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	71,408	8	,000
	Block	71,408	8	,000
	Model	71,408	8	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2488,558	,038	,051

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3,446	7	,841

H0: Odvisna spremenljivka je generirana na osnovi predpostavljenega modela.

H1: Odvisna spremenljivka ni generirana na osnovi predpostavljenega modela

Rezultat: Z modelom lahko pojasnimo statistično značilen del variance odvisne spremenljivke.

Rezultati

Classification Table(a)

			Predicted		
			Vote for Clinton in 92		Percentage Correct
	Observed		no	yes	
Step 1	Vote for Clinton in 92	no	549	390	58,5
		yes	381	527	58,0
	Overall Percentage				58,3

a The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	AGECAT			9,225	3	,026	
	AGECAT(1)	,162	,138	1,380	1	,240	1,176
	AGECAT(2)	,312	,129	5,876	1	,015	1,367
	AGECAT(3)	,420	,155	7,336	1	,007	1,523
	DEGREE			21,780	4	,000	
	DEGREE(1)	-,651	,174	13,938	1	,000	,522
	DEGREE(2)	-,705	,245	8,263	1	,004	,494
	DEGREE(3)	-,626	,196	10,169	1	,001	,535
	DEGREE(4)	-,163	,221	,543	1	,461	,850
	SEX(1)	,548	,096	32,359	1	,000	1,730
	Constant	-,033	,199	,028	1	,868	,967

a Variable(s) entered on step 1: AGECAT, DEGREE, SEX.