Multivarijatna obradba podataka

Mladen Petrovečki
Martina Mavrinac

Multivarijatna analiza podataka

- statistička obradba podataka:
  - univarijatna
  - bivarijatna
  - multivarijatna

Multivarijatni modeli

- multivarijatni ili multidimenzionalni modeli, engl. *multivariate*
  - najmanje dvije zavisne varijable (varijate)
  - broj veza među pokazateljima (m):
    - $m = k(k-1)/2$ ⇦ (k = broj pokazatelja)
    - npr. osam pokazatelja (m = 8) ⇨ k = 28
- dvosmjerna analiza varijance
- višestruka regresijska analiza, logistička regresija, Coxov regresijski test
- diskriminacijska analiza
- faktorska analiza
- klasterska analiza
- meta-analiza

Dvosmjerna analiza varijance

- dvosmjerna ANOVA
- npr. analiza vrijednosti glukoze prema dobnim skupinama i spolu ispitanika:
  - razina glukoze (mmol/L)
  - dobna skupina (<20, 20-50, >50 god.)
  - spol (M, Ž)
- uvijek
  - jedan brojčani pokazatelj ⇦ zavisni
  - dva skupna pokazatelja ⇦ nezavisni

Multipla regresijska analiza

- linearna matematička povezanost više pokazatelja
  - x_1-x_n ⇨ nezavisne varijable (prediktori)
  - y ⇨ zavisna brojčana varijabla (kriterij)
- koliko promjena svakog x određuje promjenu y:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

- kolinearnost – pojava visoke korelacije nezavisnih varijabla (loše!)
- R – multipli koeficijent korelacije
- R^2 – multipli koeficijent determinacije

Posebnosti multiple regresije

- postupci odabira pokazatelja u završnu jednadžbu (*model building*):
  - svi (*all, enter*)
  - biranje sprijeda (*forward selection*)
  - unatražno isključivanje (*backward elimination*)
  - postepeno biranje (*stepwise selection*)
  - sve moguće s traženjem najvećeg R^2
- polinomska multipla regresija:

$$y = b_0 + b_1x + b_2x^2 + \ldots$$

Primjer

- povezanost TIBC (*total iron binding capacity*) s UIBC (*unsaturated iron binding capacity*), Fe, feritinom i dobi ispitanika
- statistička obradba podataka:

Parametar		b±SE(b)	t	*P*
• konstanta	b_0	4,5±4,3	0,96	0,341
• UIBC	b_1	0,8±0,1	11,68	<0,001
• Fe	b_2	1,0±0,1	11,01	<0,001
• feritin	b_3	0,01±0,01	0,31	0,760
• dob	b_4	0,01±0,03	-0,97	0,338

- N = 87; *P* = 0,027; R = 0,38; R^2 = 0,144
- TIBC = 0,8×UIBC + 1×Fe (matematički)
- TIBC = 0,8±0,1×UIBC + 1±0,1×Fe (statistički)

Logistička regresija

- povezanost više pokazatelja:
  - x_1-x_n ⇨ nezavisne varijable (prediktori)
  - y ⇨ binarna zavisna brojčana varijabla (binarni kriterij)
- međusobno nezavisna mjerenja
- koliko promjena svakog od x određuje promjenu binarne varijable y:

$$log(p) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

- e^b = omjer izgleda (OR, *odds ratio*)

Primjer

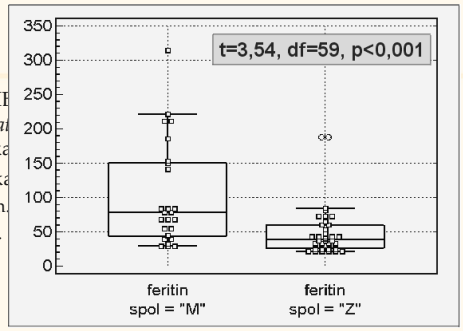
- odnos TIBC (*total iron binding capacity*), UIBC (*unsaturated iron binding capacity*), Fe, feritina i dobi ispitanika sa spolom ispitanika
- statistička obradba podataka:

Param.	b	SE(b)	*P*	OR (95% CI)
• konst.	1,60	0,49	0,003	
• TIBC	0,05	0,09	0,563	
• UIBC	-0,05	0,09	0,615	
• Fe	-0,03	0,11	0,757	
• feritin	-0,02	0,01	0,006	0,98 (0,96 – 0,99)
• dob	-0,02	0,02	0,421	

Primjer

- odnos TIBC (*unsaturated iron binding capacity*), UIBC ispitanika sa spolom ispitanika
- statistička obradba podataka:

Param.	b	SE(b)	*P*	OR (95% CI)
• konst.	1,60	0,49	0,003	
• TIBC	0,05	0,09	0,563	
• UIBC	-0,05	0,09	0,615	
• Fe	-0,03	0,11	0,757	
• feritin	-0,02	0,01	0,006	0,98 (0,96 – 0,99)
• dob	-0,02	0,02	0,421	



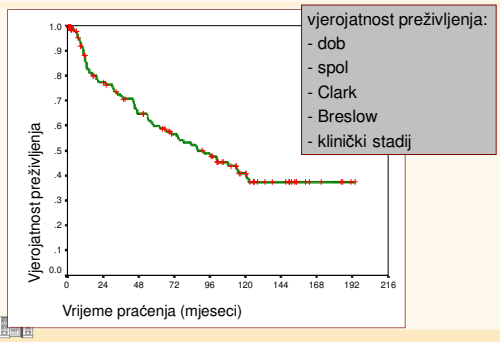
Coxov regresijski test

- povezanost više pokazatelja:
  - x ⇨ nezavisne varijable (prediktori)
  - y ⇨ zavisna varijabla (kriterij) = mjera rizika
- međusobno nezavisna mjerenja
- koliko promjena x određuje y:

$$\lambda_i(t) = \lambda_0(t)\, exp(b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n)$$

- e^b = omjer rizika (RH, *relative hazard*)

Primjer (polipoidni melanom)



Primjer

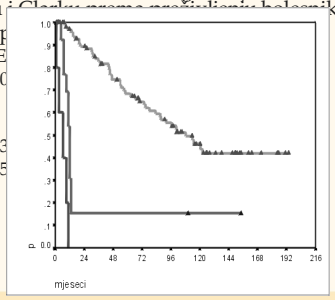
- odnos dobi i spola bolesnika te kliničkog stadija melanoma i stadija po Breslowu i Clarku prema preživljenju bolesnika
- statistička obradba podataka:

Param.	b	SE	*P*	RH (95% CI)
• dob	0,02	0,011	0,019	1,02 (1,001 – 1,04)
• spolm			0,941	
• klin_st			<0,001	
• (1)	1,75	0,35	<0,001	5,81 (2,91 – 11,58)
• (2)	3,70	0,58	<0,001	40,60 (13,02 –126,65)
• clark_st			0,098	
• bres_st			0,433	

Primjer

- odnos dobi i spola bolesnika te kliničkog stadija melanoma i stadija po Breslowu i Clarku prema preživljenju bolesnika
- statistička obradba p

Param.	b	SE
• dob	0,02	0,(
• spolm		
• klin_st		
• (1)	1,75	0,3
• (2)	3,70	0,5
• clark_st		
• bres_st		



Diskriminacijska analiza

- metoda klasifikacije ispitanika ili objekata po skupinama
- analiza varijabli koje razlikuju (diskriminiraju) dvije ili više skupina ispitanika
  - utvrđuje koje se skupine razlikuju u odnosu na aritmetičke sredine određenih varijabla
- uporaba: predviđanje nominalne ili kategoričke varijable
- primjer:
  - prikupljaju se podatci o načinu života i obiteljskoj anamnezi pušača na odvikavanju od pušenja
  - razlog: ustvrditi koje će varijable najbolje predvidjeti:
    - potpuni prestanak pušenja (skupina 1)
    - smanjenje broja popušenih cigareta (skupina 2)
    - nepromijenjenu učestalost pušenja (skupina 3)

Faktorska analiza

- skup statističko-matematičkih postupaka s pomoću kojih se raščlanjuje povezanost između većeg broja varijabli s ciljem da se:
  - smanji broj varijabli
  - odredi faktorska struktura
- sve varijable su nezavisne
- dva modela faktorske analize:
  - eksplanatorni – opisuje međusobnu povezanost varijabli sa faktorom
  - konfirmatorni – potvrđuje ili odbacuje hipoteze ili modele povezanosti

Primjer

1. oblikuje se upitnik o korištenju dopunskog zdravstvenog osiguranja (DZO)
   - sadrži 20 čestica (pitanja, varijabla, engl. *item*)
   - testiranje: ispunjava ga N = 200 ispitanika
2. izračunaju se korelacije među česticama
3. odrede se logička grupiranja
   - npr. grupiranje u dvije skupine (dvo-faktorska struktura)
4. obradba podataka o korelacijama ⇨ cilj: dobivanje linearnih kombinacija, tj. faktora, npr.:
   - faktor 1 = prednosti DZO
   - faktor 2 = nedostaci DZO
5. određuje se broj čestica dovoljan za tumačenje pojedinih faktora
   - npr. 5 čestica za svaki faktor

Primjer, rezultat

- upitnik o prednostima i nedostacima dopunskog zdravstvenog osiguranja
- deset čestica na dva faktora
  - broj čestica određen je uvjetom povezanosti čestica i faktora
  - engl. *factor loading*
  - istovjetan koeficijentu korelacije r
  - npr. r > 0,54
  - rezultat ⇨

Čest-ice	Komponenta 1	2	
v7	,767		
v8	,661		
v3	,585		
v16	,549		
v10	,541		**faktor 1**
v19	,527		
v18	,510		
v15	,502		
v4	,474		
v20	,444		
v5	,358		
v17	,345		
v14	,305		
v9		,719	
v11		,698	
v12		,691	
v13		,666	
v2		,553	**faktor 2**
v6		,518	
v1		-,480	

Klasterska analiza

- postupak klasifikacije pojedinaca ili objekata u skupine
- ne zahtijeva poznavanje skupina pripadnosti niti konačni broj skupina (kao diskriminacijska analiza)
- pronalazi sličnosti kod ispitanika po kojima ih svrstava u skupine, ne tumači zašto
- sve varijable su nezavisne
- grafički prikaz klastera ⇨ dendrogram
- primjer:
  - što reći o dijagnozi neizlječive bolesti: istinu, laž ili prešutjeti odgovor?
  - kome: bolesniku, njegovu prijatelju, kolegi, poslodavcu, članu obitelji, povjereniku zdravstvenog osiguranja, drugom liječniku, studentu medicine?

Primjer, dendrogram

What would you sayabout the diagnosis to:	Subject No. with Cluster	Relative distance between clusters (%) 0 20 40 60 80 100 %
patient	1 - A	
close family member	2 - A	
patient's health insurance agent	6 - B	
another physician	7 - B	
medical student	8 - B	
patient's close friend	3 - C	
patient's employer	4 - C	
patient's colleague	5 - D	

Meta-analiza

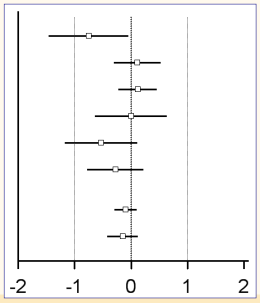
- združivanje srodnih rezultata više nezavisnih istraživanja istog sadržaja
- podrijetlo statističkog postupka ⇨ analiza sustavnih preglednih članaka
- tehnika: Gene V. Glass, 1977.
- temeljne mogućnosti:
  - povećanje statističke snage povećanjem broja ispitanika
  - jasnija procjena utjecaja istraživanog čimbenika
  - razrješenje nesigurnih spoznaja
  - odgovaranje na pitanja koja izvorne studije ne postavljaju
- *Cochrane Database of Systematic Reviews*
  - najpoznatija kolekcija cjelovitih radova koji rabe meta-analizu
  - sadrži više od 4.800 radova

Postupci meta-analize

- analiza proporcija – omjer izgleda
  - hipoteza (usporedba proporcija): omjer izgleda (OR, *odds ratio*) nije značajno različit među studijama
  - test heterogenosti (*heterogeneity test*)
    - $P \geq \alpha$ (npr. $P \geq 0{,}05$) ⇨ studije su homogene ⇨ model utvrđenog učinka (*fixed effects model*)
    - $P < a$ ⇨ studije bi mogle biti heterogene ⇨ model slučajnog učinka (*random effects model*)
  - testiranje H_0: Mantel-Haenszelov test
- analiza brojčanih podataka – standardizirana razlika
  - hipoteza (usporedba prosječnih vrijednosti): standardizirana razlika prosjeka (SMD, *standardized mean difference*) nije značajno različita među studijama
  - test heterogenosti (v. gore)
  - testiranje: Hedgesov g-test

Grafički prikaz

- trakasti grafikon
  - *forest plot*
  - *blobbogram*
  - grafikon granica pouzdanosti



Primjer 1: OR (podatci)

- djelovanje antibiotika A u liječenju upale paranazalnih sinusa specifičnim uzročnikom B u ljudi
- pretraživanje literature (antibiotik A, upala paranazanih sinusa, uzročnik B):
  - 214 istraživanja
  - 24 komparativna istraživanja u ljudi
  - 12 kontroliranih kliničkih pokusa
  - 5 cjelovitih pregleda svih efekata liječenja
- temeljni rezultat istraživanja:

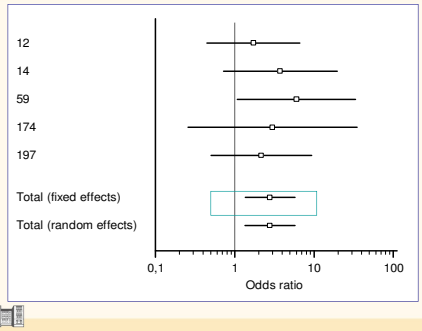
Oznaka	N_{sk}	EF_{sk}	N_{kont}	EF_{kont}
12	73	15	23	3
14	35	7	32	2
59	20	8	20	2
174	12	3	10	1
197	42	6	42	3

Primjer: OR

No.	Skupina	Kontrola	OR (95% CI)
12	15/73	3/23	1,724 (0,452–6,583)
14	7/35	2/32	3,750 (0,717–19,599)
59	8/20	2/20	6,000 (1,082–33,275)
174	3/12	1/10	3,000 (0,260–34,576)
197	6/42	3/42	2,167 (0,504–9,312)
Σ	39/182	11/127	

Utvrđen učinak (*fixed effect*) 2,806 (1,363–5,778)
Slučajni učinak (*random effect*) 2,781 (1,347–5,744)
Test heterogenosti: Q = 1,506; DF = 4; *P* = 0,826

Primjer: grafički prikaz OR



Primjer 2: SMD (podatci)

- djelovanje dvaju hipoglikemika A (noviji) i B (tradicionalni) na snižavanje koncentracije glukoze u krvi
- pretraživanje literature (hipoglikemik A, hipoglikemik B, glukoza u krvi, hiperglikemija, dijabetes):
  - 180 istraživanja
  - 54 komparativna istraživanja u ljudi
  - 9 kontroliranih kliničkih pokusa
  - 6 cjelovitih pregleda svih efekata liječenja
- temeljni rezultat istraživanja (glukoza, mmol/L):

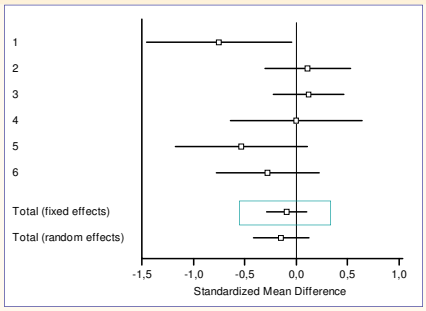
Ozn.	N_{sk}	X ± SD	N_{kont}	X ± SD
1	19	4,7 ± 1,2	17	5,9 ± 1,9
2	45	4,9 ± 1,6	47	4,7 ± 1,9
3	67	5,2 ± 2,1	67	4,9 ± 2,8
4	20	4,4 ± 0,9	20	4,4 ± 1,8
5	24	4,6 ± 1,8	18	5,6 ± 1,9
6	32	4,9 ± 2,1	33	5,5 ± 2,2

Primjer: SMD

No.	N_1	N_2	Σ	SMD (95% CI)
1	19	17	36	-0,748 (-1,452 do -0,0434)
2	45	47	92	0,113 (-0,302 do 0,527)
3	67	67	134	0,121 (-0,222 do 0,463)
4	20	20	40	0,000 (-0,640 do 0,640)
5	24	18	42	-0,532 (-1,174 do 0,110)
6	32	33	65	-0,276 (-0,774 do 0,223)
Σ	207	202	409	

Utvrđeni učinak -0,093 (-0,288 do 0,103)
Slučajni učinak -0,147 (-0,415 do 0,122)
Test heterogenosti: Q = 8,599; DF = 5; *P* = 0,1262

Primjer: grafički prikaz SMD



Podatci u primjerima

- Šimundić AM, ur. Osnove biostatistike u svakodnevnoj praksi – tečaj trajnog usavršavanja medicinskih biokemičara, priručnik. Medicinska naklada, Zagreb, 2008.
- Knežević F, Petrovečki M, Šeparović V. Histological types of polypoid cutaneous melanoma. Croat Med J 1992;33:220-4.
- Statistical methods for multiple variables. U: Dawson-Saunders B, Trapp RG. Basic & clinical biostatistics. Lange Medical Books/McGraw-Hill, New York – Toronto, 2004. Str. 245-279.
- MedCalc for Windows, Statistics for biomedical research – Software manual v 11.2. MedCalc Software, Mariakerke, 2010.
- Patrick Holford and His Own Reality: Part 1, the blobbogram. Dostupno sa: http://holfordwatch.info/2008/04/10/patrick-holford-and-his-own-reality-part-1-the-blobbogram/.
- Pulanić D, Vražić H, Ćuk M, Petrovečki M. Ethics in Medicine: Students' Opinions on Disclosure of True Diagnosis. Croat Med J 2002;43(1):75-79.