Dvojfaktorová analýza rozptylu s využitím programu SAS

Martina Lakatová
Ekonomická Univerzita
Fakulta hospodárskej informatiky
Inžiniersky stupeň štúdia, 2. ročník

Analýza rozptylu

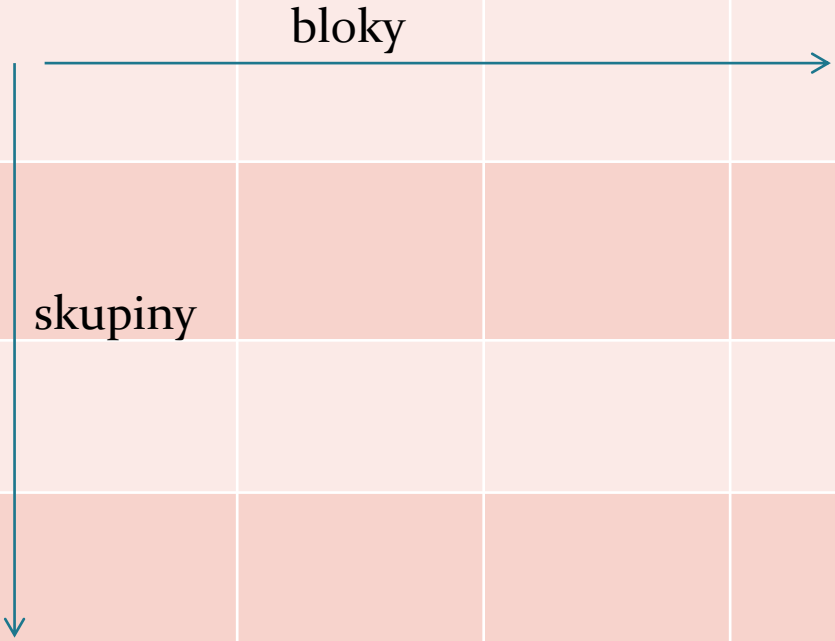
- posúdenie vplyvu faktora (faktorov) na závislú premennú,
- zistenie existencie štatisticky významných rozdielov medzi úrovňami faktorov

Použitie pri:

- analýze vzťahov medzi vysvetľujúcou a vysvetľovanými premennými, alebo pri
- analýze zdrojov variability v štatistických modeloch.

- 1 faktor – jednofaktorová ANOVA
- viac faktorov – viacfaktorová ANOVA
  - bez opakovania / s opakovaním
  - s interakciou / bez interakcie
  - nevyvážený/vyvážený model
- MANOVA – súčasné pôsobenie 1 alebo viacerých faktorov na 2 alebo viacero nezávislých premenných

Faktor B A	1	2	…	G	Priemer
1					$y_{1\bullet}$
2					$y_{2\bullet}$
…					…
H					$y_{H\bullet}$
Priemer	$y_{\bullet 1}$	$y_{\bullet 2}$	…	$y_{\bullet G}$	$y_{\bullet\bullet}$



Dvojfaktorová analýza rozptylu (bez interakcie)

Vychádzame celkového súčtu štvorcov SS_T

$$SS_T = SS_A + SS_B + SS_E$$

kde

$$SS_T = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{r}(y_{ghi} - \bar{y}_{\bullet\bullet})^2$$

vyjadruje celkovú variabilitu; súčty štvorcov

$$SS_A = G^* r \sum_{h=1}^{H}(\bar{y}_{h\bullet} - \bar{y}_{\bullet\bullet})^2 \qquad SS_B = H^* r \sum_{g=1}^{G}(\bar{y}_{\bullet g} - \bar{y}_{\bullet\bullet})^2$$

popisujú rozdiely v úrovni skupín vzniknutých triedením podľa faktora A (faktora B) a vzťah

$$SS_E = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{r}(y_{hgi} - \bar{y}_{hg})^2$$

predstavuje reziduálny súčet štvorcov.

Dvojfaktorová analýza rozptylu (s interakciou)

Vychádzame celkového súčtu štvorcov SS_T

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

a uvažujeme aj s efektom interakcie

$$SS_{AB} = r \sum_{h=1}^{H} \sum_{g=1}^{G} (\bar{y}_{hg} - \bar{y}_{h\bullet} - \bar{y}_{\bullet g} + \bar{y}_{\bullet\bullet})^2$$

Dvojfaktorová analýza rozptylu (bez interakcie)

Hypotéza o účinku hlavných efektov (bez interakcie)

- nulová hypotéza a alternatívna pre efekt faktora A

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_H = 0$ pre všetky $h = 1, 2, \ldots, H$

H_1: nie všetky α_h sa rovnajú 0

- nulová a alternatívna hypotéza pre efekt faktora B

$H_0: \beta_1 = \beta_2 = \ldots = \beta_G = 0$ pre všetky $g = 1, 2, \ldots, b)$

H_1: nie všetky β_g sa rovnajú 0

Dvojfaktorová analýza rozptylu (s interakciou)

Hypotéza o účinku hlavných efektov (bez interakcie)

- nulová hypotéza a alternatívna pre efekt faktora A

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_H = 0 \quad \text{pre všetky} \quad h = 1, 2, \ldots, H$$

$$H_1: \text{nie všetky } \alpha_h \text{ sa rovnajú 0}$$

- nulová a alternatívna hypotéza pre efekt faktora B

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_G = 0 \quad \text{pre všetky} \quad g = 1, 2, \ldots, b)$$

$$H_1: \text{nie všetky } \beta_g \text{ sa rovnajú 0}$$

Hypotéza o efekte interakcie

- nulová hypotéza

$$H_0: (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{HG} = 0$$

- alternatívna hypotéza

$$H_1: \text{nie všetky } \alpha\beta_{hg} \text{ sa rovnajú 0}$$

Dvojfaktorová analýza rozptylu (bez interakcie)

- Testovacia charakteristika pre hlavné efekty

$$F_A = \frac{H*G(r-1)}{H-1} * \frac{G*r\sum_{h=1}^{H}(\bar{y}_{h\bullet} - \bar{y}_{\bullet\bullet})^2}{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{r}(y_{hgi} - \bar{y}_{hg})^2}$$

$$F_B = \frac{H*G(r-1)}{G-1} * \frac{H*r\sum_{g=1}^{G}(\bar{y}_{\bullet g} - \bar{y}_{\bullet\bullet})^2}{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{r}(y_{hgi} - \bar{y}_{hg})^2}$$

- Hlavný efekt faktora A (faktora B) považujeme za preukázaný, ak hodnota F prekročí hodnotu kritickej oblasti $F_{1-\alpha(H-1,H*G(r-1))}$, respektíve $F_{1-\alpha(G-1,H*G(r-1))}$,

Dvojfaktorová analýza rozptylu (s interakciou)

- Testovacia charakteristika pre testovanie interakcie

$$F = \frac{H * G(r-1)}{(H-1)(G-1)} * \frac{r\sum_{h=1}^{H}\sum_{g=1}^{G}(\bar{Y}_{hg} - \bar{Y}_{h\bullet} - \bar{Y}_{\bullet g} + \bar{Y}_{\bullet\bullet})^2}{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{r}(Y_{hgi} - \bar{Y}_{hg})^2}$$

- Kritická oblasť $F_{1-\alpha((H-1)*(G-1),\ H*G(r-1))}$.

Výsledky práce

Použitý program na analýzu:

- pomocou rôznych programov – Excel, STATGRAPHICS Plus, produkty SAS

- SAS Enterprise Guide – časť ponuky *Task*-> *ANOVA*
  - *One-Way ANOVA* (PROC ANOVA)
  - *Linear Models* (PROC GLM)

Databáza

	Drug	Disease	BPChange
1	1	A	42
2	1	A	44
3	1	A	36
4	1	A	13
5	1	A	19
6	1	A	22
7	1	B	33
8	1	B	40
9	1	B	
10	1	B	
11	1	B	
12	1	B	
60	4	A	15
61	4	B	27
62	4	B	12
63	4	B	12
64	4	B	-5
65	4	B	16
66	4	B	15
67	4	C	22
68	4	C	7
69	4	C	25
70	4	C	5
71	4	C	12
72	4	C	7

vysvetľovaná premenná

- zmena systolického krvného tlaku

vysvetľujúce premenné

- typ choroby – Choroba X, Choroba Y, Choroba Z,
- druh podávaného lieku – Liek 1, Liek 2, Liek 3, Liek 4

Tab. 1 Tabuľka viacfaktorovej analýzy rozptylu *s interakciou (Zdroj: SAS, EG, 5.1)*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	6 353,00	577,5455	6,03	0,0001
Error	60	5 749,00	95,8167		
Corrected Total	71	12 102,00			

R-Square	Coeff Var	Root MSE	Zmena_KT Mean
0,5250	51,0710	9,7886	19,1667

Tab. 2 Overenie štatistickej významnosti faktorov a interakcie *(Zdroj: SAS, EG, 5.1)*

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Liek	3	4 543,22	1 514,41	15,81	0,0001
Choroba	2	809,08	404,54	4,22	0,0193
Liek*Choroba	6	1 000,69	166,78	1,74	0,1271

Tab.3 Tabuľka viacfaktorovej analýzy rozptylu *bez interakcie (Zdroj: SAS, EG, 5.1)*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5 352.31	1070.46	10.47	0.0001
Error	66	6 749.69	102.27		
Corrected Total	71	12 102.00			

R-Square	Coeff Var	Root MSE	BPChange Mean
0.442266	52.76227	10.11277	19.16667

Tab. 6 Overenie štatistickej významnosti faktorov *(Zdroj: SAS, EG, 5.1)*

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Disease	2	809.08	404.54	3.96	0.0238
Drug	3	4543.22	1514.41	14.81	0.0001

- Data
- Model
- Model Options
- Advanced Options
- Post Hoc Tests
  - Least Squares
  - Arithmetic
- Plots
- Predictions
- Titles
- Properties

Class effects to use	
Drug	True
Disease	True
Drug*Disease	False
Comparisons	
Show p-values for differences	None
Adjustment method for comparison	Scheffe
Confidence limits	
Show confidence limits	
Covariates	
Covariate to use in computation	

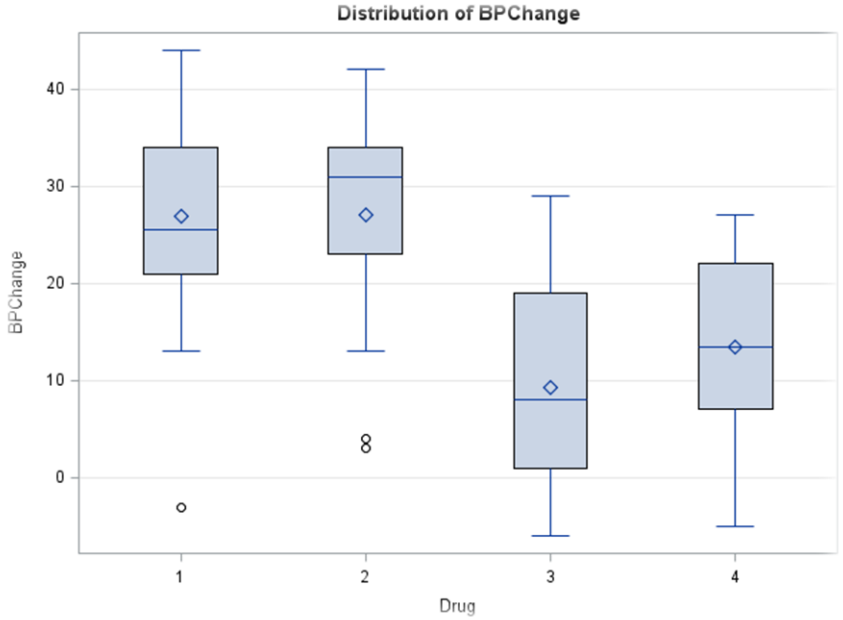
Default
Bonferroni
Dunnett
Scheffe
Sidak
Studentized maximum modulus
Tukey
No adjustment

Class effects to use	
Drug	True
Disease	True
Drug*Disease	False
Comparisons	
Comparison method	Default
Error mean square	
Error effect to use	
Type of mean square	
Means options	
Show means for	
Join nonsignificant subsets	
Sort means in descending order	
Confidence intervals	
Show for the means	No
Show for all pairwise differences	No
Homogeneity of variance	
Homogeneity test	None

Default
Bonferroni t-tests
Duncan multiple range tests
Gabriel multiple comparison tests
Pairwise t-tests
Ryan-Einot-Gabriel-Welsh multiple ran
Scheffe's multiple comparison procedur
Sidak's adjusted pairwise t-test
Student-Newman-Keuls multiple range
Tukey's studentized range test
Waller-Duncan K-ratio t-test

Tab. 7 Schefféova metóda pre faktor Druh lieku *(Zdroj: SAS, EG, 5.1)*

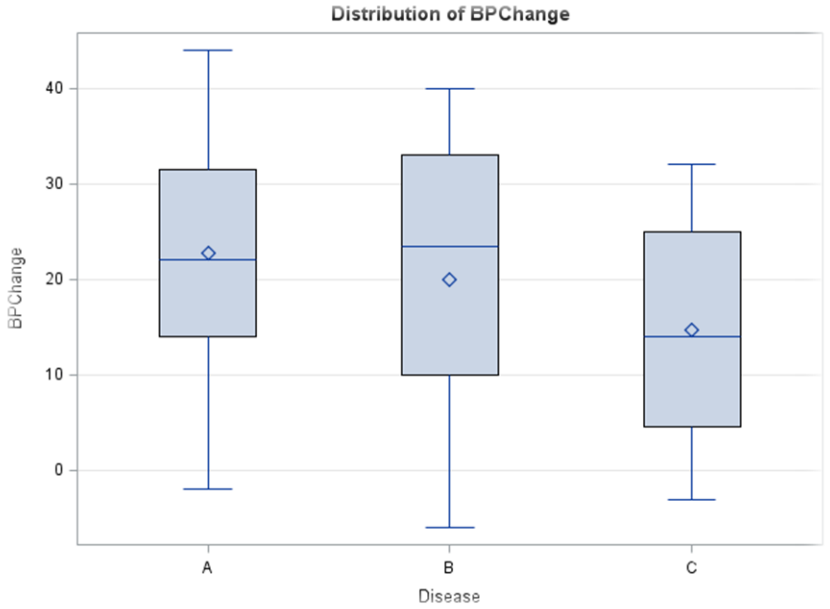
Drug	BPChange LSMEAN	Least Squares Means for effect Drug Pr > |t| for H0: LSMean(i)=LSMean(j)				
		i/j	1	2	3	4
1	26.89	**1**		1.0000	0.0001	0.0025
2	27.06	**2**	**1.0000**		0.0001	0.0021
3	9.28	**3**	**0.0001**	**0.0001**		0.6773
4	13.44	**4**	**0.0025**	**0.0021**	**0.6773**	



Obr. 1 Rozdiely pre faktor Druh lieku *(Zdroj: SAS, EG, 5.1)*

Tab. 7 Schefféova metóda pre faktor Druh choroby*(Zdroj: SAS, EG, 5.1)*

Disease	BPChange LSMEAN	Least Squares Means for effect Disease Pr > |t| for H0: LSMean(i)=LSMean(j)			
		i/j	1	2	3
X	22.79	1		0.6350	0.0266
Y	20.00	2	0.6350		0.2012
Z	14.71	3	0.0266	0.2012	



Obr. 2 Rozdiely pre faktor Typ choroby *(Zdroj: SAS, EG, 5.1)*

Jednofaktorová analýza rozptylu

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{\bar{y}})^2=\sum_{i=1}^{k}n_i(\bar{y}_i-\bar{\bar{y}})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

$$SST = SSM + SSE$$

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}$$

Dvojfaktorová analýza rozptylu (bez interakcie)

- Testovacia charakteristika pre hlavné efekty

$$F_A = \frac{H*G(r-1)}{H-1} * \frac{G*r\sum_{h=1}^{H}(\bar{y}_{h\bullet}-\bar{y}_{\bullet\bullet})^2}{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{r}(y_{hgi}-\bar{y}_{hg})^2}$$

$$F_B = \frac{H*G(r-1)}{G-1} * \frac{H*r\sum_{g=1}^{G}(\bar{y}_{\bullet g}-\bar{y}_{\bullet\bullet})^2}{\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{r}(y_{hgi}-\bar{y}_{hg})^2}$$

Ďakujem za pozornosť