StatSoft

Jak poznat vliv faktorů vizuálně

V tomto článku bychom se rádi věnovali otázce, jak poznat již z grafického náhledu vztahy a závislosti v analýze rozptylu. Pomocí následujících grafických zobrazení byste měli krásně pochopit, jak to vypadá, když jsou významné jednotlivé faktory či interakce a tedy i trochu lépe pochopit samotnou analýzu rozptylu. Není třeba se bát ničeho složitého, cílem je pouze ukázat chování dat za různých situací.

Analýza rozptylu (ANOVA)

Než přistoupíme ke grafickým zobrazením, nejprve shrňme, o co vlastně jde. Analýza rozptylu se zabývá vyšetřováním vztahů mezi spojitou závislou proměnnou a jednou nebo více nezávislými kategorickými proměnnými (také se jim říká faktory). Uveďme příklady úloh analýzy rozptylu: Zkoumání vlivu varianty zkouškového testu (A,B,C) na dosažený výsledek žáků nebo vliv hnojiva a daného pole na množství plodiny při sklizni. Toto je úloha, která nás zajímá.

Typy úloh

ANOVA jednoduchého třídění (anglicky one-way ANOVA)

Takto nazýváme situaci, kdy máme jen jednu nezávislou kategorickou proměnnou (nehledě na počet hladin této proměnné). Pokud bychom chtěli ozdobit text vzorcem, pak rovnice takovéhoto „regresního" modelu bude:

$$Y_{jp} = \mu + \alpha_j + e_{jp}$$

Y značí závislou kvantitativní proměnnou, μ je referenční či souhrnná úroveň (absolutní člen), α_j je parametr vztahující se k j-té hladině nezávislé proměnné, e je náhodná chyba. Je zde i p, což znamená, že pro každou hladinu nezávislé proměnné můžete mít více pozorování.

Příkladem takovéto úlohy může být třeba odhad ceny bytu pouze v závislosti na tom, v kterém je kraji.

ANOVA dvojného třídění pouze s hlavními efekty

Mírně složitější je model dvojného třídění, kdy přidáme další nezávislou proměnnou. Model je tedy takovýto:

$$Y_{jgp} = \mu + \alpha_j + \beta_g + e_{jgp}$$

β_g je parametr pro k g-tou hladinu druhé vysvětlující proměnné.

ANOVA dvojného třídění s interakcemi

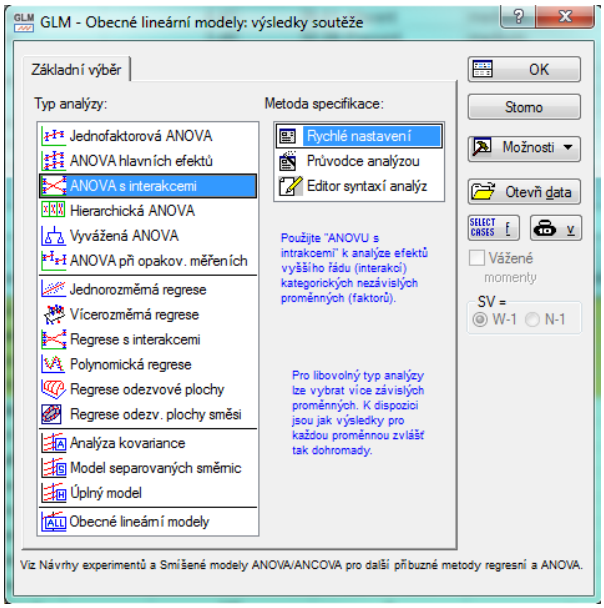
Někdy mohou efekty prvního i druhého faktoru působit složitěji a společně. Pak mluvíme o modelu dvojného třídění s interakcemi:

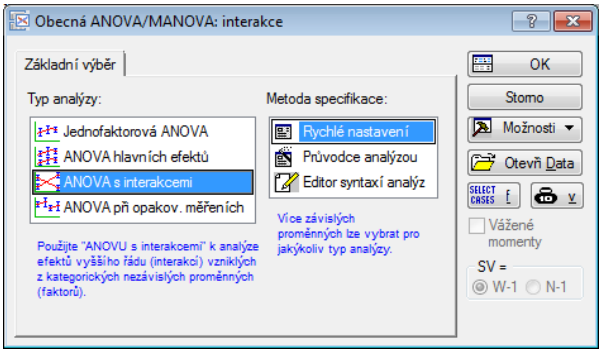
$$Y_{jgp} = \mu + \alpha_j + \beta_g + \lambda_{jg} + e_{jgp}$$

Přibyl člen λ_{jg}, který vnáší společný vliv prvního a druhého faktoru, každá kombinace hladin těchto dvou faktorů tedy může mít jiný unikátní vliv.

Příkladem dvojného třídění může být zkoumání vlivu velikosti bytu (2+1,2+KK,...) a kraje na jeho cenu. To jestli jde o model s interakcemi nebo bez, záleží na tom, jak působí v konkrétním případě faktory na závislou proměnnou. Kompletně vypracovaný příklad analýzy rozptylu dvojného třídění najdete také v **jednom z minulých newsletterů.**

Poznámka: V softwaru *STATISTICA* naleznete jednotlivé modely buď pod tlačítkem *Statistiky-ANOVA* (to





je jednodušší a pro náš článek naprosto dostačující možnost) nebo jako první položky funkcionality GLM – což je zkratka pro Obecné lineární modely, které mají v sobě rozšířenější možnosti lineárních modelů a nalezneme je pod *Statistiky-Pokročilé lineární/nelineárními modely* v menu statistik. Tři úlohy výše zadáte pomocí prvních tří položek v okně *Typ analýz.*

My se nyní budeme snažit tyto situace rozpoznat a popsat. Nyní již bez vzorců a jednoduše.

Grafické výstupy

Připomeneme, že budeme sledovat situaci, kdy vysvětlujeme závislost jednoho kvantitativního znaku (spojitá závislá veličina) na dvou kvalitativních proměnných (faktorech).

Abychom si vše představili, předpokládejme, že máme na pozadí úlohy následující data: závislou proměnnou je výše platu, nezávislou je pohlaví a dosažené vzdělaní. Všechna následující data jsou pouze ilustrativní a nijak neodrážejí reálný stav věcí ohledně platů žen a mužů, jména jsou také smyšlená. Zdrojová data by tedy byla ve tvaru tabulky vpravo:

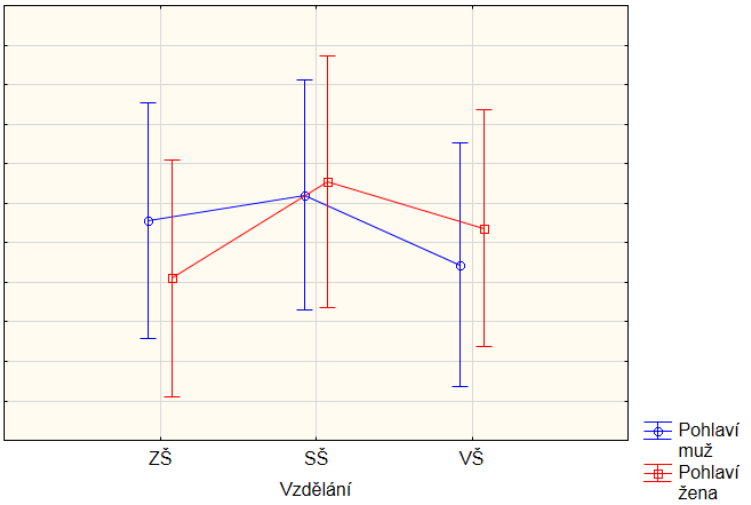
	1 Plat	2 Pohlaví	3 Vzdělání
Eva Vonásková	17169	žena	SŠ
Jarmila Černá	16642	žena	ZŠ
Michal Novák	16157	muž	ZŠ
Petr David	18970	muž	VŠ
Tomáš Procházka	15879	muž	SŠ
Martin Seifert	16067	muž	VŠ
Filip Kukačka	25347	muž	ZŠ

Jeden z výstupů analýzy rozptylu je graf průměrů v jednotlivých skupinách s vykreslením intervalu spolehlivosti pro tento průměr. Z tohoto grafu budeme vycházet v následujícím. Proč právě průměr? Analýza rozptylu má za úkol porovnání středních hodnot v jednotlivých skupinách, klasickým odhadem střední hodnoty je právě průměr a ten využívá ve svých výpočtech i metoda analýza rozptylu.

Začněme nejjednodušším případem:

1. Proměnná Pohlaví, ani Vzdělání nemá vliv na plat

Na následujícím obrázku vidíte, průměry a jejich intervaly spolehlivosti pro všechny kombinace skupin (pohlaví má 2 skupiny, vzdělání 3 skupiny, dohromady šest kombinací a tedy i šest průměrů v grafu.)

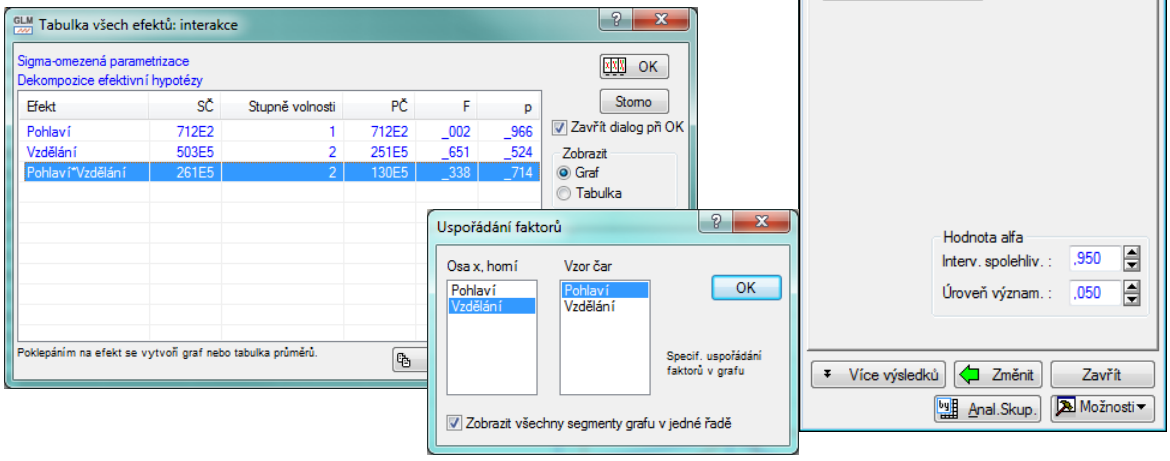


Je vidět, že „anténky", potažmo celé intervaly se příliš neliší – ani modré oproti červeným, ani se nijak nemění spolu se vzděláním. Všech 6 intervalů se hodně překrývá. Je to tedy typický příklad situace, kdy faktory nemají vliv na závislou proměnnou.

Efekt	p
Abs. člen	0,000000
Pohlaví	0,965843
Vzdělání	0,523858
Pohlaví*Vzdělání	0,714315

Pokud bychom pro tato data spočetli analýzu rozptylu a vypočetli významnost koeficientů v modelu dvojného třídění s interakcemi, vyjde podle očekávání, že žádná proměnná ani interakce významné nejsou. Významný je jen absolutní člen, což je vlastně jakási hladina, kde se vyskytují průměrně všechna data a jelikož se jedná o platy, jistě se tato hladina nebude pohybovat okolo 0. Jinak řečeno zamítáme hypotézu, že by byl absolutní člen modelu roven 0.

Poznámka: Pokud byste nevěděli, jak vyprodukovat graf a výsledky výše, postupujte podle následujícího návodu: otevřte dialog *Statistiky-ANOVA-ANOVA s interakcemi-OK*. Proměnné zvolte: *Plat* jako závislou a *Pohlaví* a *Vzdělání* jako kategorické faktory. Klikneme na *OK* a máme výsledky. Pod tlačítkem *Velik. Efektů* vyvoláme testy významnosti. Tlačítkem *Vš. Efekty/grafy* vyvoláme graf. Přesné nastavení pro graf:

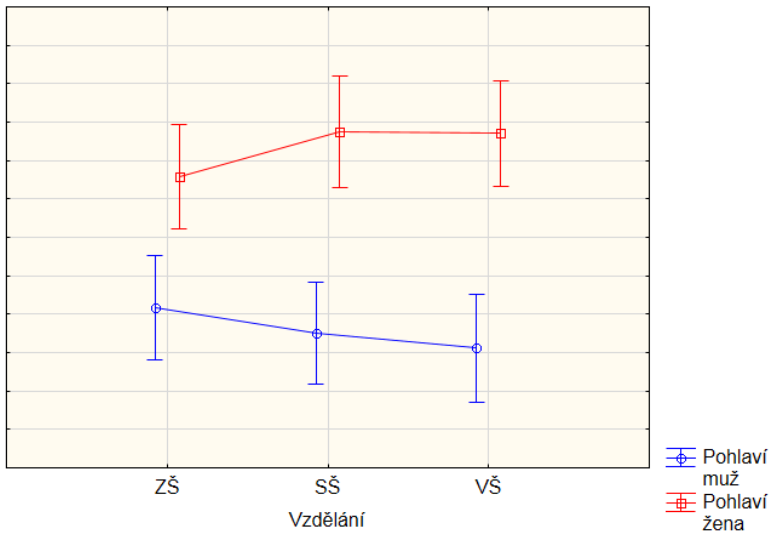


Tento graf lze vyprodukovat i bez analýzy rozptylu přes záložku *Grafy*, jedná se o *Grafy průměrů s odchylkami*. Za grupovací proměnnou je potřeba zvolit *Vzdělání* a v záložce *Kategorizovaný* aktivovat proměnnou pro kategorizaci X a nastavit ji na *Pohlaví,* navíc zvolit rozložení přes sebe.

2. Proměnná Pohlaví má vliv, Vzdělání však vliv nemá

Je vidět, že pro různá pohlaví jsou hladiny odlišné – průběhy intervalů pro muže a pro ženy jsou dokonce úplně odděleny. Pokud vezmeme zvlášť muže, tak se jejich plat pohybuje na stejné hladině (intervaly se hodně překrývají), podobně u žen, vliv vzdělání je tedy zanedbatelný, viz výsledky testů významnosti faktorů:

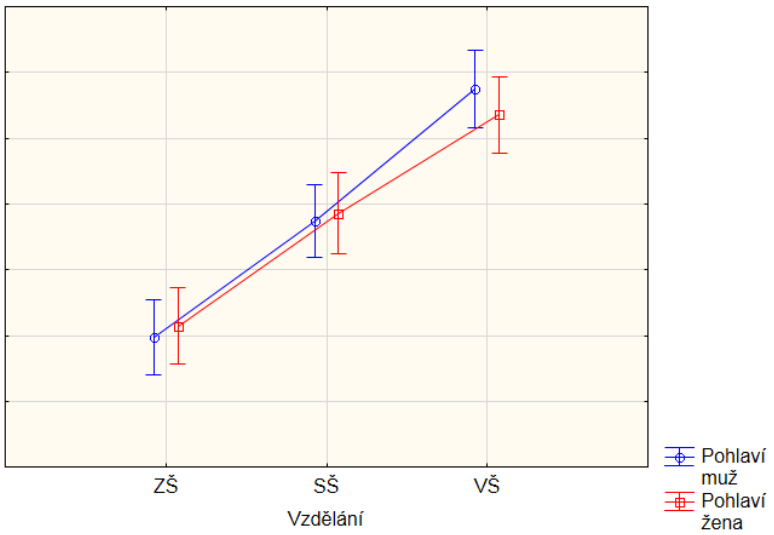
Efekt	p
Abs. člen	0,000000
Pohlaví	0,000000
Vzdělání	0,919679
Pohlaví*Vzdělání	0,239826



3. Proměnná Pohlaví nemá vliv, Vzdělání však vliv má

Tato situace je velmi podobná té předchozí, pouze gtaficky to vypadá jinak, jelikož nyní máme rozdíl u veličiny, která je přímo na ose a ne u veličiny, která je rozlišena barvami. Není asi potřeba moc vysvětlovat, hladiny pro vzdělání se liší (obecně nemusí jen růst, jak je tomu na obrázku, klidně může jít o „zlomené" nebo klesající průběhy). Zatímco hladiny pro pohlaví pro jednotlivé vzdělání jsou takřka stejné.

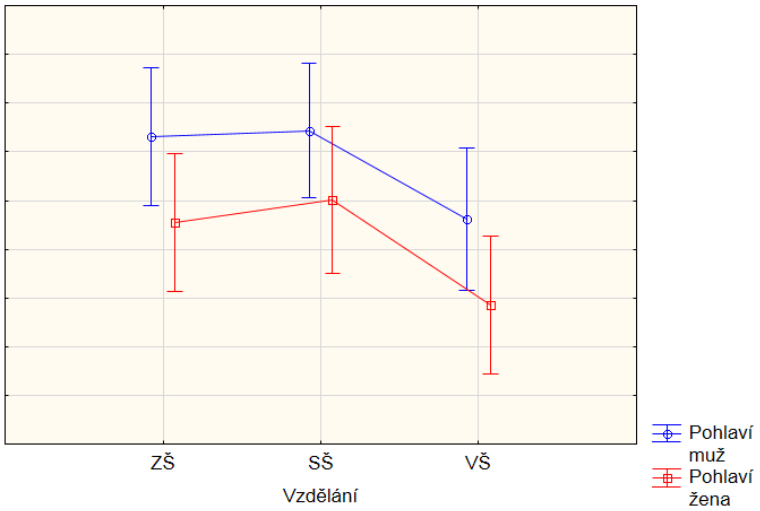
Efekt	p
Abs. člen	0,000000
Pohlaví	0,894989
Vzdělání	0,000000
Pohlaví*Vzdělání	0,553071



4. Má vliv proměnná Pohlaví i proměnná Vzdělání

Pokud nastane situace, kdy průběhy v grafu pro jednotlivá pohlaví (v našem případě modrý a červený graf) mají stejný tvar, ale jsou od sebe posunuty pro jednotlivá pohlaví, pak jde o vliv obou nezávislých veličin zároveň. Čím je interakce nevýznamnější, tím více mají průběhy stejný tvar. V tomto případě tedy o vlivu interakce nemůže být řeč.

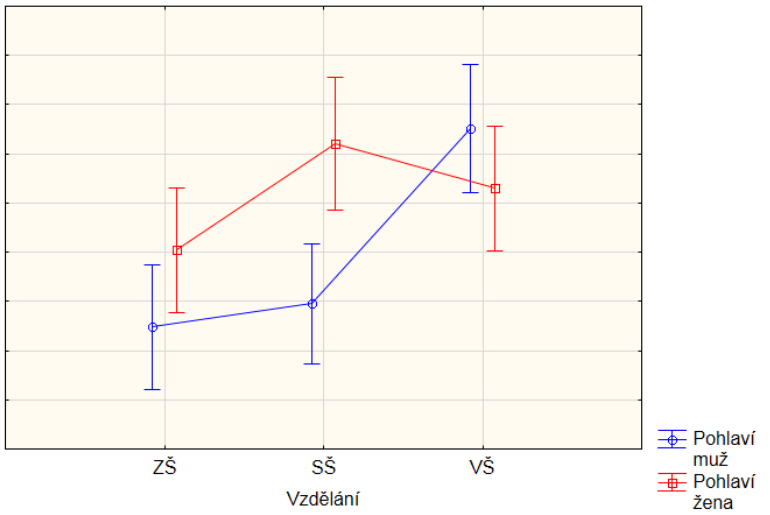
Efekt	p
Abs. člen	0,000000
Pohlaví	0,006064
Vzdělání	0,014644
Pohlaví*Vzdělání	0,965103



5. Významný vliv interakce

Nejsložitější situace nastává, pokud má vliv i interakce, což znamená, že každá kombinace faktorů může mít svou unikátní hladinu. Z obrázku tuto situaci poznáme tak, že průběhy pro jednotlivá pohlaví již nejsou stejná, jinak řečeno, že se křivky lámou pro každé pohlaví jinak.

Efekt	p
Abs. člen	0,000000
Pohlaví	0,024394
Vzdělání	0,000402
Pohlaví*Vzdělání	0,003262

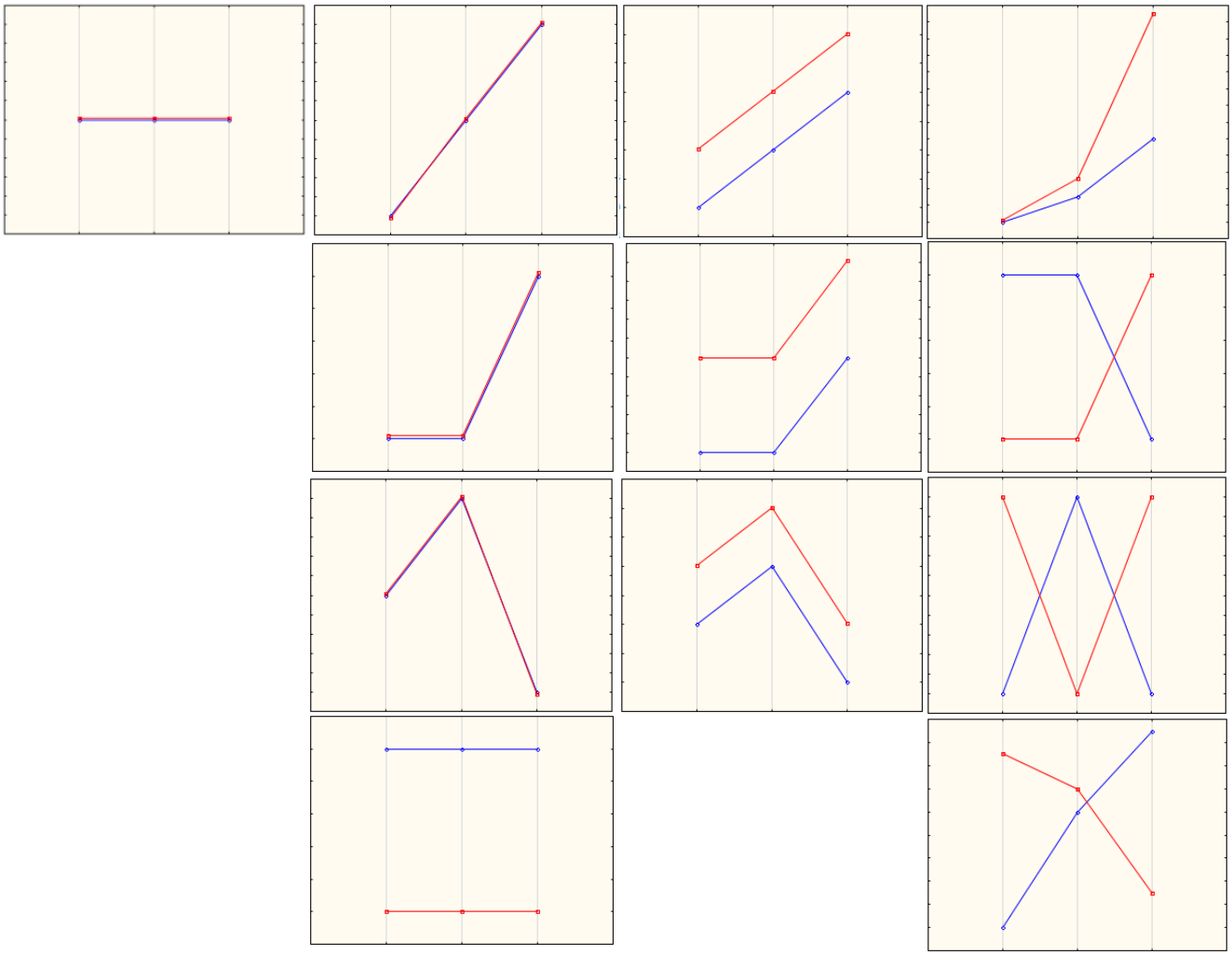


Shrnutí

Naším cílem bylo ukázat situaci a trochu pomoci s pochopením modelu analýzy rozptylu dvojného třídění. Samozřejmě je potřeba upozornit, že není vhodné se rozhodovat pouze podle grafů, nicméně mohou pro Vás být dobrým vodítkem i prezentací toho, co se v datech děje.

Na závěr bychom shrnuli teoretické průběhy pro jednotlivé situace- tedy opravdu jen s vlivy, které zkoumáme. Ostatní považujeme za nulové, což se v praxi nestane, nicméně alespoň je pěkně vidět, co jednotlivé vlivy mohou provést s průměry (opět bereme v úvahu 2 faktory, jeden má 3 hladiny a druhý dvě).

Významný pouze absolutní člen	Vliv má pouze jeden faktor	Vliv mají oba faktory, ale ne interakce	Interakční člen má také vliv



StatSoft STATISTICA

StatSoft CR s. r. o.
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
t +420 233 325 006, **f** +420 233 324 005
e info@statsoft.cz