Zobecněná analýza rozptylu, více faktorů a proměnných

Menu:	QCExpert	Anova	Více faktorů

Zobecněná analýza rozptylu (ANalysis Of VAriance, ANOVA) umožňuje posoudit do jaké míry ovlivňují kvalitativní proměnné (např. číslo směny, jméno operátora), i kvantitativní měřené proměnné (teplota, tlak, otáčky) zvolenou odezvu (např. přesnost, čistota, prevnost, rozměry, a podobně). Je tedy modul zobecněná Anova důležitým nástrojem pro identifikaci cest vedoucích ke zlepšení jakosti, výtěžku, atd. Zároveň pomůže tento pokročilý modul identifikovat a prokázat příčiny problémů v procesech. Rovněž lze tento postup velmi efektivně využít při hledání a prokazování nejdůležitějších parametrů ovlivňujících předmět zájmu.

Tento modul je zobecněním analýzy rozptylu založené na lineárním regresním modelu s pomocnými (dummy) binárními proměnnými a Moore-Penrose pseudoinverzi. Tato metoda umožňuje kombinaci modelu ANOVA s modelem klasické lineární regrese. V modelu se tedy mohou jako prediktory vyskytovat diskrétní faktory (podobně jako v předchozích dvou modulech), ale současně také diskrétní nebo spojité číselné proměnné. Navíc je možné tohoto modulu využít pro predikci odezvy pro danou kombinaci prediktorů. Analyzuje se vliv faktorů s pevnými efekty a spojitých proměnných na výsledek pozorování. Výsledky pozorování Z_i při n_j různých úrovních faktoru X_j a různých hodnotách proměné Y_k lze popsat regresním modelem s neznámými parametry:

$$Z = \alpha_0 + \sum_j \boldsymbol{\alpha}_j X_j + \sum_k \beta_k Y_k + \varepsilon ,$$

kde α_0 je absolutní člen (celková střední hodnota), $\boldsymbol{\alpha}_j$ je vektor ($n_j \times 1$) latentních parametrů pro j-tý faktor a β_k je regresní koeficient k-té proměnné. Náhodná chyba ε_{ij} má normální rozdělení a definičně nulovou střední hodnotu, $\varepsilon \sim N(0,\sigma^2)$. Latentní parametry slouží pouze k testu významnosti příslušného faktoru. Modul *Anova - více faktorů* vypočítá odhady $\mathbf{a}_j$, b_k, e_i parametrů $\boldsymbol{\alpha}$, β, a chyb ε.

Použití pro MSA: Measurement System Analysis

Analýza systému měření (Measurement system analysis, MSA) je požadován na řadě pracovišť, kde se používají měřidla a čidla ke získávání údajů jak v průběhu technologických procesů, tak i například ve fázi hodnocení parametrů surovin, stavu zařízení, vlastností produktu. Cílem MSA je posoudit kvalitu získávaných naměřených hodnot tím, že se rozdělií a identifikuje jejich variabilita. Podle principu sčítání rozptylů je celkový rozptyl součtem dílčích rozptylů v důsledku náhodných i nenáhodných chyb různého charakteru

$$\sigma^2_{celkový} = \sum_i \sigma_i^2 + \sigma^2_{reziduální} .$$

Příspěvky k celkovému rozptylu σ_i^2 se vysvětlují jako důsledek variability i-té příčiny, například měnící se operátor, směna, podmínky měření, způsob kalibrace, atd. Reziduální rozptyl je pak možné v praxi chápat jako rozptyl způsobený nevysvětlenými nebo neovlivnitelnými náhodnými vlivy. Úkolem MSA je identifikovat a kvantifikovat vlivy na nepřesnost měření na základě existujících dat a případně je formálně rozdělit v R&R studii na chyby reprodukovatelnosti (stejný objekt měření, stejné podmínky) a opakovatelnosti (stejný objekt měření, různé podmínky), případně zahrnout ještě variabilitu procesní (různé objekty měření), což je nevyhnutelné například u destruktivních měření. Zobecněná ANOVA je efektivní nástroj pro takovýto typ úloh. Její hlavní výhoda je v možnosti současného zpracování jak spojitých, tak faktorových prediktorů, jak ilustruje Tabulka 1. V protokolu pak odpovídá *Reziduální rozptyl* rozptylu opakovatelnosti, *Vysvětlený rozptyl* rozptylu reprodukovatelnosti (pokud data zahrnují pouze variabilitu příslušející reprodukovatelnosti) a navíc v odstavci *Anova pro jednotlivé faktory* je podrobný rozbor dílčích příspěvků jednotlivých faktorů k celkové variabilitě v podobě parciálních součtů čtverců příslušejících jednotlivým proměnným. Na faktory s největšími příspěvky k variabilitě je pak třeba soustředit primární úsilí vedoucí ke snížení jejich vlivu.

Data a parametry

Data jsou uspořádána ve sloupcích, na pořadí sloupců nezáleží. Sloupce obsahují hodnoty prediktorů (faktorů a/nebo parametrů), jeden sloupec musí obsahovat hodnoty odezvy. *Faktory* jsou proměnné, které místo čísel mají pouze „úrovně“. Úrovně jsou v tabulce zapsány jako textové hodnoty, například: „malý, střední, velký, Ano, Ne“, apod., mohou být ovšem zapsány i pomocí čísel, jako např. číslo linky, šarže, směny. Proměnné jsou číselné, mohou nabývat reálných hodnot. Sloupec odezvy obsahuje číselné hodnoty pozorované pro odpovídající kombinace úrovní faktorů a proměnných. Faktory mohou mít dvě nebo více úrovní, proměnné mohou nabývat libovolné reálné hodnoty.

Tabulka 1 Příklad uspořádání dat v datové tabulce. Prediktory jsou zastoupeny dvěma faktory: Linka a Operátor a jednou proměnnou: Teplota. Faktor Linka má 3 úrovně, faktor Operátor 2 úrovně. Proměnná Teplota musí obsahovat alespoň dvě odlišné hodnoty. Pro každou kombinaci úrovní a proměnné máme k dispozici 1 pozorování Z nazvané „Výtěžek“.

Prediktory

Faktory *Proměnná* *Odezva*

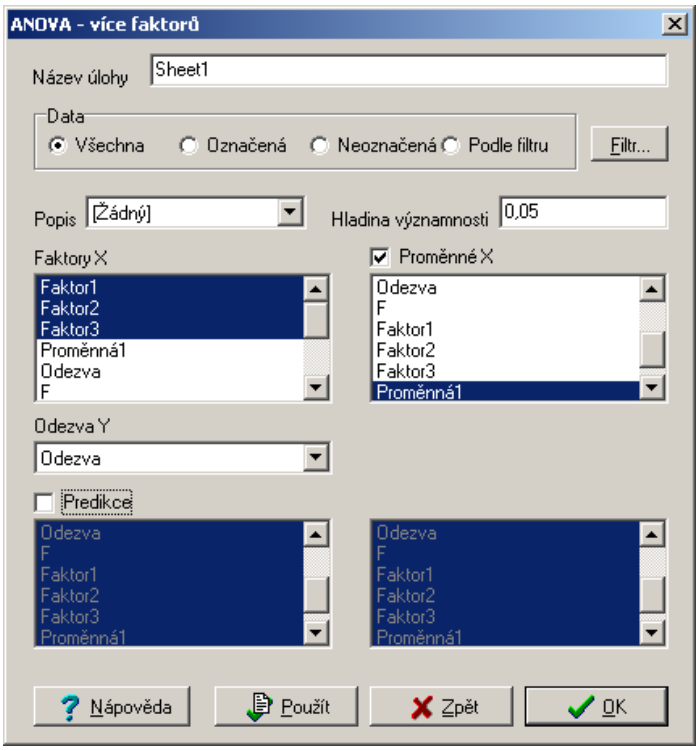
Linka	Operátor	Teplota	Výtěžek
A	Veselý	13.3	14.6
B	Vránová	16.3	17.4
C	Veselý	18.7	13.3
A	Vránová	14.5	12.6
B	Veselý	11.1	17.5
C	Vránová	16.0	14.9
.....	.....	.....	.....

Data musí obsahovat prediktor, avšak tyto prediktory mohou být buď pouze faktory, nebo pouze proměnné, nebo kombinace obou. Pokud ovšem jsou prediktorem pouze numerické proměnné, je samozřejmě vhodnější použít pro analýzu lineární regresi.

Jedna kombinace úrovně faktorů se nazývá cela, nebo buňka. Na rozdíl od předchozích dvou metod Anova pro 1 a 2 faktory není nutné, aby byly změřeny odezvy pro všechny kombinace úrovní faktorů. Naopak, chybějící měření lze s určitou mírou přesnosti dopočítat (predikovat). Obecně lze podle počtu měření při jednotlivých kombinacích, tedy počtu měření v cele rozlišit 3 typy analýzy rozptylu, bez ohledu na počet hodnoty proměnných:

1 pozorování v každé cele – *Vyvážená analýza rozptylu bez opakování*
stejný počet $n_0 > 1$ pozorování v každé cele – *Vyvážená analýza rozptylu s opakováním*
nestejný počet $n_{ij} > 0$ pozorování v každé cele – *Nevyvážená analýza rozptylu*

V dialogovém okně se vyberou sloupce, s faktory a sloupce s proměnnými. Proměnné lze vybrat jen je-li zaškrtnuto políčko *Proměnné*. Nechceme-li vybrat žádný faktor, je možné odznačit faktory myší se stisknutým *Ctrl*. V poli Odezva se vybere jediný sloupec obsahující pozorovanou odezvu, zvolí se hladina významnosti, na níž se budou provádět testy (obvyklá hodnota je 0.05). Po stisku *OK* se provede analýza a výsledky se zapíší do okna *Grafy* a do *Protokolu*.



Obrázek 1 ANOVA více faktorů - Vstupní dialogový panel

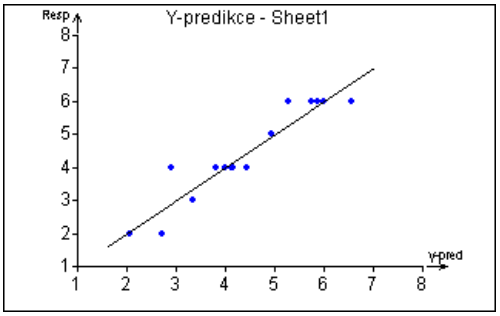
Výstup je popsán v následujících odstavcích.

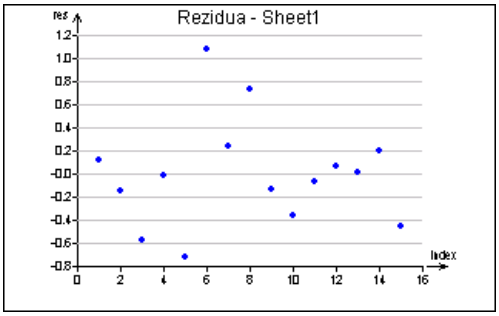
Protokol

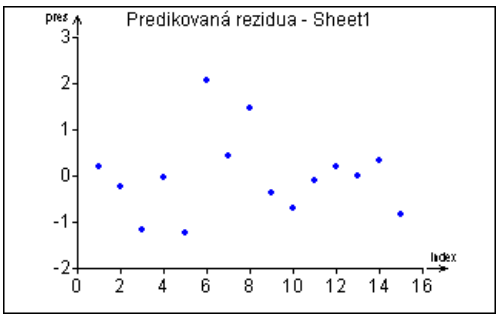
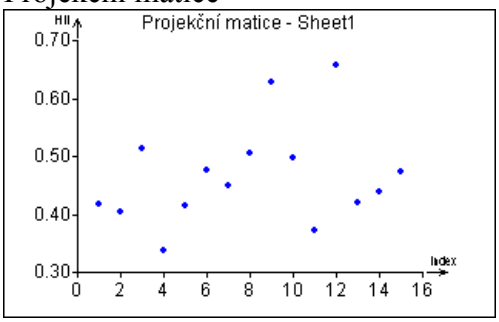
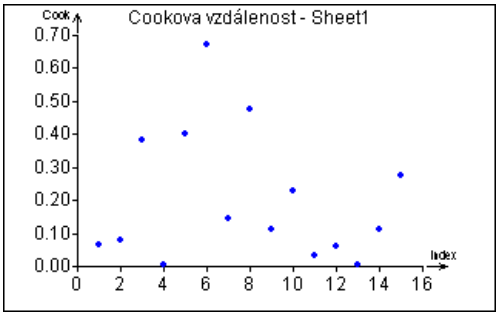
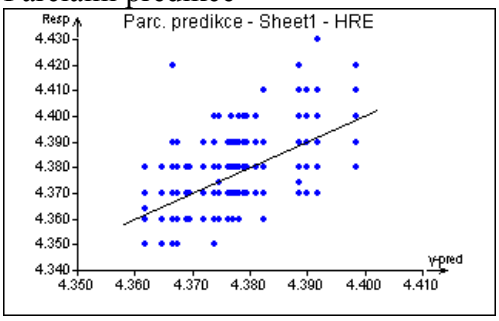
Analýza rozptylu více faktorů	
Počet dat	Celkový počet řádků pro výpočet.
Počet prediktorů celkem	Počet faktorů a proměnných.
Počet faktorů	Z toho počet faktorů.
Počet proměnných	Z toho počet proměnných.
Průměr Y	Aritmetický průměr všech naměřených odezev.
Absolutní člen	Predikce odezvy při nepřítomnosti vlivu faktorů a nulových hodnotách všech proměnných.
Hladina významnosti	Zvolená hladina významnosti pro testy. Doporučená hodnota 0.05.
Počty úrovní	Počty úrovní jednotlivých faktorů.
Celková ANOVA	Celkový test, zda studované prediktory (faktory a proměnné) mají vůbec nějaký vliv na odezvu.
Zdroj	Zdroje variability se zde hodnotí podle toho, jak velká část celkové variability (variability pouze naměřené odezvy bez ohledu na nějaký model) se podařila vysvětlit pomocí celého použitého modelu. Variabilita se měří primárně součtem čtverců.
Stupňů volnosti	Počet stupňů volnosti příslušející jednotlivým zdrojům.
Součet čtverců	Variabilita vyjádřená jako součet čtverců.
Rozptyl	Variabilita vyjádřená jako rozptyl.
F-statistika	Poměr celkového a reziduálního rozptylu.
p-hodnota	Pravděpodobnost odpovídající *F*-statistice. Je-li *p*-hodnota menší, než zadaná hladina významnosti, je model významný.

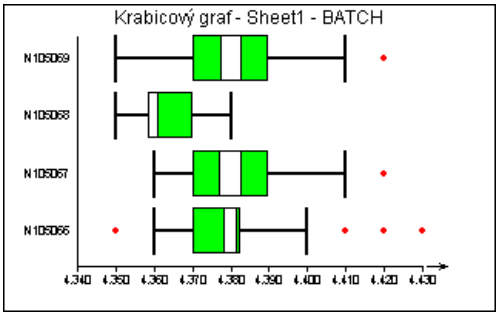
Významnost	Slovní vyjádření výsledku testu – Významný/Nevýznamný.
Zdroj	Jednotlivé zdroje variability
Celková variabilita	Hodnoty odvozené od celkového součtu čtverců $CSC = \sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2$
Vysvětlená variabilita	*CSC – RSC*
Reziduální variabilita	variabilita odvozená od reziduálního součtu čtverců $RSC = \sum_{i=1}^{n}\left[Z_i - \left(\alpha_0 + \sum_j \mathbf{a}_j X_{ij} + \sum_k b_k Y_{ik} + e_i \right) \right]$
ANOVA pro jednotlivé faktory	Tabulka podílu variability vysvětlené jednotlivými členy modelu, tedy jednotlivými faktory a proměnnými.
Prediktor	Název faktoru nebo proměnné.
Parametr	Hodnota odhadu parametru b_k, pokud jde o proměnnou, v případě faktoru je toto políčko prázdné.
Součet čtverců	Parciální součet čtverců, který se podařilo vysvětlit tímto prediktorem.
F-statistika	Odpovídající F-kvantil.
p-hodnota	Odpovídající pravěpodobnost. Je-li *p*-hodnota menší, než zadaná hladina významnosti, je tento prediktor významný.
Významnost	Slovní vyjádření výsledku testu – Významný/Nevýznamný.
Tabulka predikce	Bylo-li v dialogovém okně zaškrtnuto políčko *Predikce*, vypisuje se tabulka predikovaných hodnot odezvy na základě zadaných hodnot prediktorů vybraných v dialogovém okně
Prediktory	Zadané hodnoty faktorů a proměnných.
Predikce	Vypočítané predikované (předpověděné) hodnoty odezvy
Spodní mez	Spodní konfidenční mez predikce
Horní mez	Horní konfidenční mez predikce

Grafy

Y-Predikce 	Základní graf proložení dat. Vynáší se naměřená odezva (osa Y) proti predikované hodnotě (osa X). Čím blíže leží body k přímce y=x, tím úspěšněji popisuje (predikuje) model data. Graf posuzuje model jako celek, rozlišení na jednotlivé faktory a proměnné poskytují parciální grafy predikce.
Rezidua	Graf reziduí představuje svislé vzdálenosti bodů od přímky z předchozího grafu Y-predikce, tedy odchylky odezvy od predikce. Hodnoty výrazně vzdálené od nuly (ve srovnání s ostatními) signalizují pravděpodobné vychýlené odezvy (možné hrubé chyby).

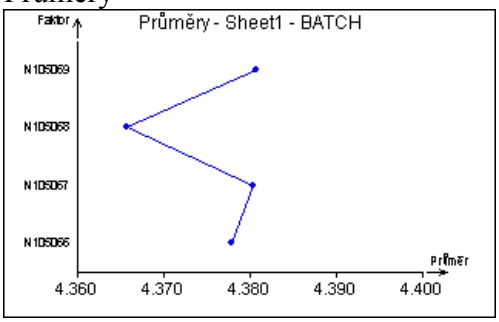


Predikovaná rezidua 	Predikovaná rezidua představují vzdálenost naměřené *i*-té odezvy Z_i od modelu (predikce), který byl však vypočítán z dat mimo *i*-tého bodu (řádku). Jedná se o podobný graf jako předchozí graf reziduí, tento graf je však daleko citlivější na výskyt ojedinělých vybočujících hodnot odezvy (outlierů). Hodnoty výrazně vzdálené od nuly (ve srovnání s ostatními) signalizují pravděpodobné vychýlené odezvy (možné hrubé chyby).
Projekční matice 	Graf prvků projekční matice je analogický s podobným grafem v lineární regresi. Vysoké hodnoty diagnostikují data s vysokým vlivem na výsledky analýzy. U takových bodů musí být věnována zvýšená pozornost správnosti změřené odezvy. Možnou příčinou vysokých hodnot jsou také chyby v hodnotách faktoru nebo proměnné.
Cookova vzdálenost 	Graf pro posouzení vybočujících bodů. Výrazně vysoké hodnoty svědčí o možné hrubé chybě odezvy.
Parciální predikce 	Graf parciální predikce je obdobou parciálního regresního grafu v lineární regresi. Vyjadřuje významnost příspěvku daného faktoru k vysvětlení variability predikce, tedy i statistickou významnost tohoto faktoru v modelu. Čím výrazněji tvoří body lineární závislost, tím je tento faktor významnější. Skutečnou významnost je však třeba ověřit v odstavci *ANOVA pro jednotlivé faktory* v protokolu.
Krabicový graf	Krabicový graf znázorňuje rozdělení naměřených odezev pro jednotlivé úrovně jednoho faktoru při ignorování vlivu



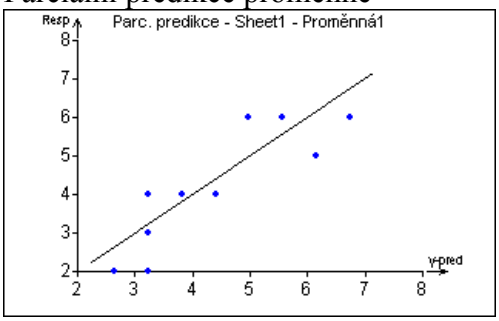
ostatních faktorů a proměnných. Pokud ostatní prediktory ignorovat nelze (jsou-li statisticky významné), je tento graf třeba chápat pouze jako informativní. Význam grafu je shodný s krabicovým grafem pro 1-faktorovou ANOVU.
Větší obdélník ohraničuje vnitřních 50% dat, horní okraj zeleného (vyšrafovaného) obdélníku odpovídá 75% kvantilu, spodní okraj zeleného obdélníku odpovídá 25% kvantilu, střed bílého pruhu v zeleném obdélníku odpovídá mediánu, šířka pruhu odpovídá intervalu spolehlivosti mediánu, dva černé proužky jsou tzv. vnitřní hradby. Data mimo vnitřní hradby jsou označena červeným bodem a lze je považovat za vybočující měření.

Průměry



Graf průměrů je jinou grafickou formou předchozího krabicového grafu. Znázorňuje polohu aritmetického průměru odezvy pro pro jednotlivé úrovně daného faktoru uvedeného v záhlaví grafu.

Parciální predikce proměnné



Obdoba parciálního regresního grafu v lineární regresi. Vyjadřuje významnost příspěvku dané proměnné k vysvětlení variability predikce, tedy i statistickou významnost této proměnné v modelu. Čím výrazněji tvoří body lineární závislost, tím je tato proměnná významnější. Skutečnou významnost je však třeba ověřit v odstavci *ANOVA pro jednotlivé faktory* v protokolu.
