EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Statistická analýza dat

Učební texty k semináři

Autor:

Prof. RNDr. Milan Meloun, DrSc.

Datum:

25. 2. 2011

Centrum pro rozvoj výzkumu pokročilých řídicích a senzorických technologií CZ.1.07/2.3.00/09.0031

TENTO STUDIJNÍ MATERIÁL JE SPOLUFINANCOVÁN EVROPSKÝM SOCIÁLNÍM FONDEM A STÁTNÍM ROZPOČTEM ČESKÉ REPUBLIKY

OBSAH

1. INTERAKTIVNÍ ANALÝZA JEDNOROZMĚRNÝCH DAT

1.1. Úvod

Otázka spolehlivosti a správného vyhodnocení experimentálních dat se v době osobních počítačů ocitá u každého měření dat na prvním místě. V kontrolní laboratoři, ať už vodohospodářské, chemické, biologické, fyzikální či jakékoliv jiné, tvoří základ experimentální práce měření na přístroji. V laboratořích dnes představují instrumentální metody spojovací článek mezi přírodovědnými a technickými obory, protože moderní počítačem řízené přístroje používá každá laboratoř. Navíc na každém psacím stole laboratoře nacházíme počítač, většinou nejvyšší kvality, kapacity a rychlosti, vybavený moderním software. Je proto neomluvitelné vyhodnocovat naměřená data zjednodušenými, aproximativními postupy pozůstalými z dob kalkulaček. Kontrolní orgány, komisaři akreditačních komisí ale především konkurenční pracoviště v zahraničí se předhánějí při vyhodnocování dat v užívání špičkového software s rigorózními matematickými postupy, ve kterých není žádného zjednodušení či zanedbání nějakých statistických předpokladů. Výsledky dosažené těmito náročnějšími postupy se pak berou za validní a jedině správné a přijatelné třeba v okružním testu. Ukažme si zde proto jeden z novějších postupů *interaktivní statistické analýzy dat,* který je založen na diagnostikování v dialogu s osobním počítačem čili na interaktivní analýze a který nabízí uživateli hlubší pohled do všech tajemství, ukrytých v experimentálních datech. S tímto problémem souvisí obvykle i vhodný software, který zajistí bezproblémové a přátelské prostředí a "nechá data promluvit". Nezapomeňme přitom na důležité pravidlo, že *"úroveň užívaného software dnes prozrazuje úroveň celého pracoviště"*.

1.2. Postup interaktivní analýzy dat

Obecný postup náročnější statistické analýzy jednorozměrných dat lze vyjádřit následujícím schématem. Interaktivní přístup uvedený postup ulehčuje, protože většina statistického software obsahuje uvedené statistické diagnostiky a testy.

1. **Průzkumová (exploratorní) analýza dat (EDA)** vyšetřuje především stupeň symetrie a špičatosti rozdělení, lokální koncentraci dat a odhaluje také vybočující a podezřelá data.

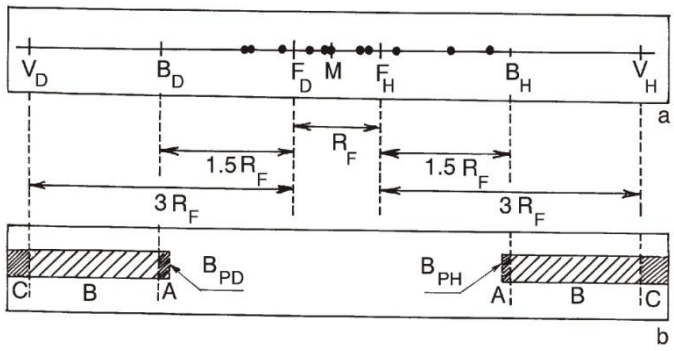
2. **Ověření základních předpokladů o výběru dat** se týká ověření normality, ověření nezávislosti, ověření homogenity a konečně i určení minimální četnosti analyzovaných dat.

3. **Transformace dat** následuje v případě porušení některého z předpokladů o výběru. Patří sem mocninná, exponenciální transformace a Boxova-Coxova transformace.

4. **Vyčíslení nejlepších odhadů parametrů polohy, rozptýlení a tvaru** se týká vyčíslení jednak klasických odhadů (aritmetický průměr a rozptyl), jednak robustních odhadů (medián, uřezané průměry, winsorizovaný rozptyl) a konečně i adaptivních *M*-odhadů. Retransformovaný průměr po transformaci dat se přesto obvykle jeví jako nejlepší odhad střední hodnoty.

1.3. Exploratorní diagnostiky v analýze jednorozměrných dat

Prvním krokem v analýze jednorozměrných dat je průzkumová, exploratorní analýza. Jejím cílem je odhalit statistické zvláštnosti v datech a ověřit předpoklady o výběru pro následné rigorózní statistické zpracování. Jedině tak lze zabránit provádění numerických výpočtů bez hlubších statistických souvislostí.



Obr. 1-1 Konstrukce barierově-číslicového schématu indikujícího vybočující hodnoty: a) diagram rozptýlení s mediánem *M*, kvartily F_D (dolní) a F_H (horní), vnitřní hradby B_D (dolní) a B_H (horní), vnější hradby V_D (dolní) a V_H (horní); b) oblast vybočujících hodnot: A přilehlé (B_{PD} je blízké B_D a B_{PH} je blízké B_H), B značí oblast vnějších a C vzdálených bodů.

Z různých typů výběru se v laboratoři nejvíce uplatňuje **reprezentativní náhodný výběr**, $\{x_i\}$, $i = 1, ..., n$, který má čtyři základní vlastnosti:

(1) Jednotlivé prvky výběru x_i jsou vzájemně nezávislé.

(2) Výběr je homogenní, tj. všechna x_i pocházejí ze stejného rozdělení pravděpodobnosti s konstantním rozptylem.

(3) Předpokládá se také, že jde o normální rozdělení pravděpodobnosti.

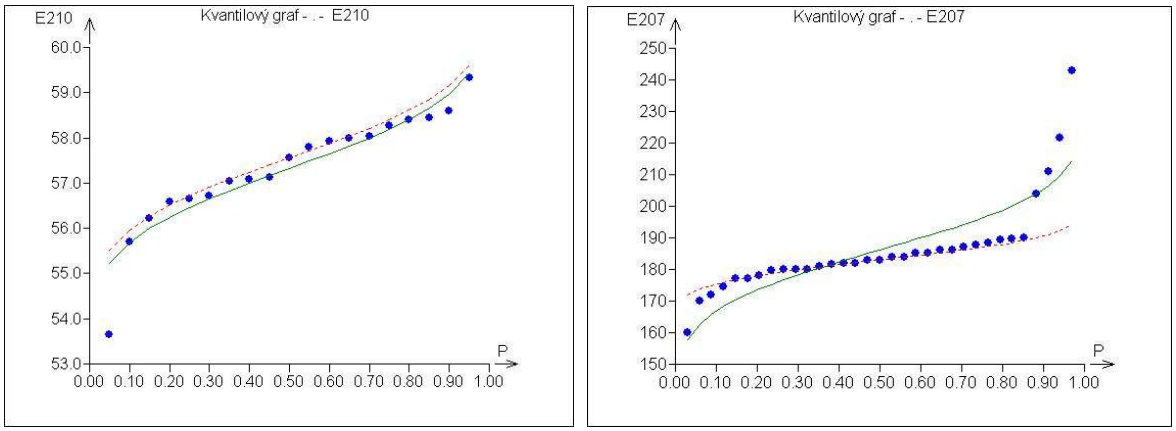
(4) Všechny prvky souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru.

Před vlastní analýzou je vždy nezbytné ověřit platnost základních předpokladů, tj. nezávislost, homogenitu a normalitu výběru. Využívá se k tomu robustních kvantilových charakteristik, které umožňují sledování lokálního chování dat a které jsou vhodné pro malé nebo středně velké výběry. Vychází se z **pořádkových statistik** výběru $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. Platí, že střední hodnota i-té pořádkové statistiky je rovna $100P_i$ procentnímu kvantilu výběrového rozdělení $F^{-1}(P_i) = Q(P_i)$, kde $F(x)$ označuje distribuční funkci a $Q(P_i)$ kvantilovou funkci výběru. Symbol $P_i = i/(n + 1)$ označuje **pořadovou pravděpodobnost**. Připomeňme, že $100P_i$ procentní výběrový kvantil je hodnota, pod kterou leží $100P_i$ procent prvků výběru. Optimální hodnoty P_i závisí na předpokládaném rozdělení výběru. Pro normální rozdělení se doporučuje volba $P_i = (i - 3/8)/(n + 1/4)$. Vynesením hodnot $x_{(i)}$ proti P_i, $i = 1, ..., n$, se získá hrubý odhad **kvantilové funkce** $Q(P)$. Ta je inverzní k funkci distribuční a jednoznačně charakterizuje rozdělení výběru. V průzkumové analýze se často používá speciálních **kvantilů** L pro pořadové pravděpodobnosti $P_i = 2^{-i}$, $i = 1, 2, ...$, které se také nazývají **písmenové hodnoty**.

Tabulka 1-1. Označení písmenových hodnot

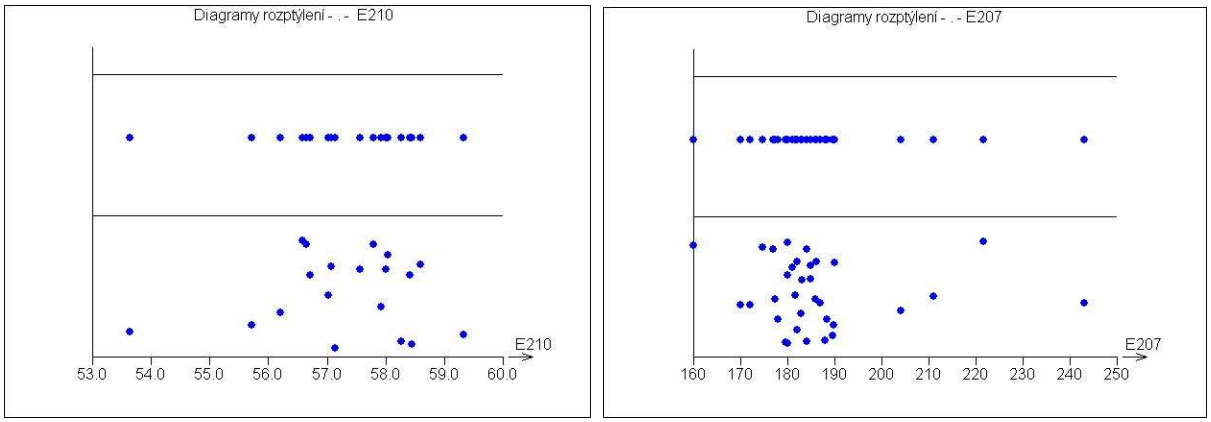
i	i-tý kvantil	Pořadová pravděpodobnost P_i	Symbol písmenové hodnoty L	Hodnota kvantilu u_{Pj}
1	Medián	$2^{-1} = 1 / 2$	M	0
2	Kvantity	$2^{-2} = 1 / 4$	F	-0.674
3	Oktily	$2^{-3} = 1 / 8$	E	-1.15
4	Sedecily	$2^{-4} = 1 / 16$	D	-1.53

Symbol u_{P_i} označuje kvantil normovaného normálního rozdělení $N(0, 1)$. Kromě **mediánu** ($i = 1$) existují pro každé $i > 1$ dvojice kvantilů, a to dolní a horní písmenová hodnota L_D a L_H. Dolní písmenová hodnota je pro pořadovou pravděpodobnost $P_i = 2^{-i}$, zatímco horní je pro $P_i = 1 - 2^{-i}$. Počet písmenových hodnot závisí na rozsahu výběru. Pro velikost výběru n lze určit n_L písmenových hodnot včetně mediánu dle vztahu $n_L = 1.44 \ln (n + 1)$.



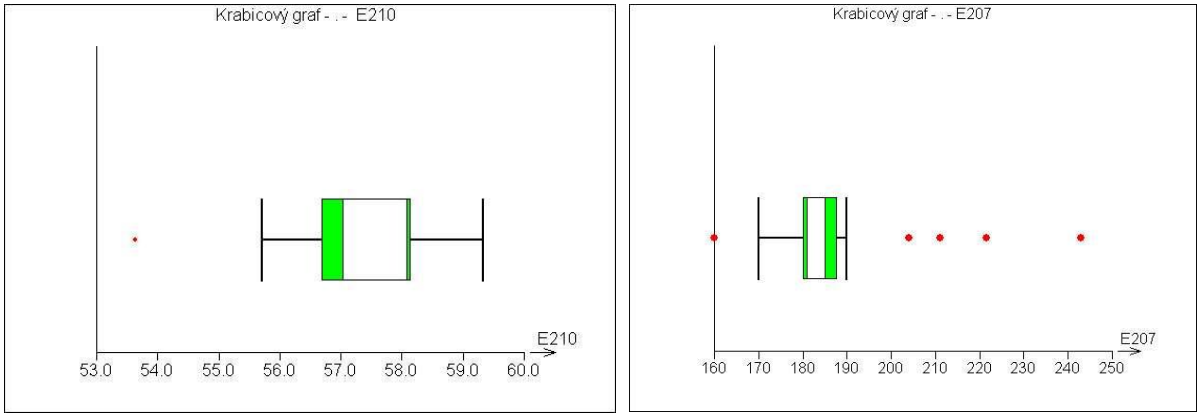
Obr. 1-2 Kvantilové grafy (robustní --- a klasické) pro výběry z rozdělení (a) normálního, symetrického rozdělení Úlohy E2.10, (b) asymetrického rozdělení Úlohy E2.07.

Kvantilový graf (osa x: pořadová pravděpodobnost P_i , osa y: pořádková statistika $x_{(i)}$) umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, který může být symetrický, sešikmený k vyšším nebo nižším hodnotám. Ke snadnějšímu porovnání s normálním rozdělením se do tohoto grafu zakreslují i kvantilové funkce normálního rozdělení $N_{P_i} = \hat{\mu} + \hat{\sigma} u_{P_i}$, pro $0 \le P_i \le 1$, a to: (1) klasických odhadů parametrů polohy a rozptýlení $\hat{\mu} = \bar{x}$ a $\hat{\sigma}$ = s, a (2) robustních odhadů $\hat{\mu} = \tilde{x}_{0.5}$ a $\hat{\sigma} = R_F / 1.349$.



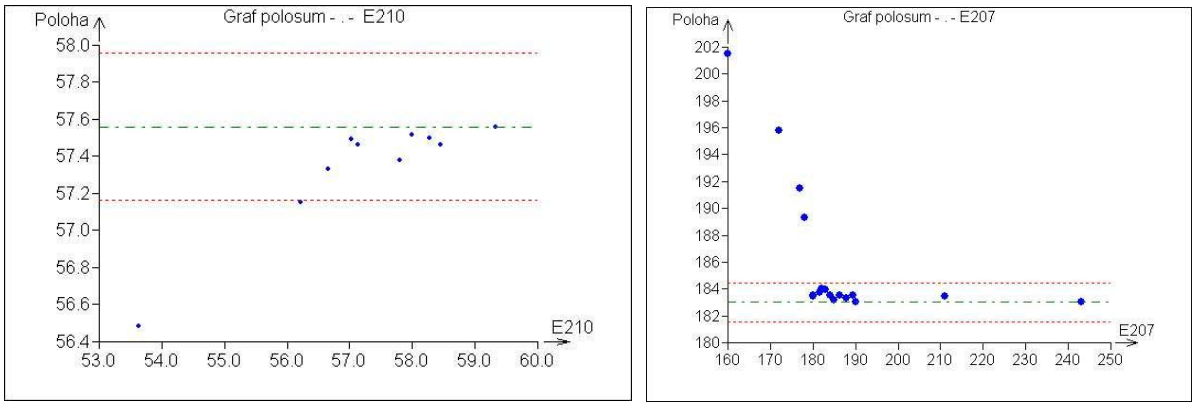
Obr. 1-3 Konstrukce (a) diagramu rozptýlení a (b) rozmítnutého diagramu rozptýlení pro výběry z rozdělení (a) normálního, symetrického rozdělení, (b) asymetrického rozdělení.

Diagram rozptýlení (osa *x*: hodnoty *x*, osa *y*: libovolná úroveň, obyčejně *y* = 0) představuje jednorozměrnou projekci kvantilového grafu do osy *x*, zatímco **rozmítnutý diagram rozptýlení** představuje týž graf, ale body jsou vhodně rozmítnuté ve směru *y*-nové osy. I při své jednoduchosti tento diagram názorně ukazuje na lokální koncentraci dat a indikuje i podezřelá a vybočující měření.



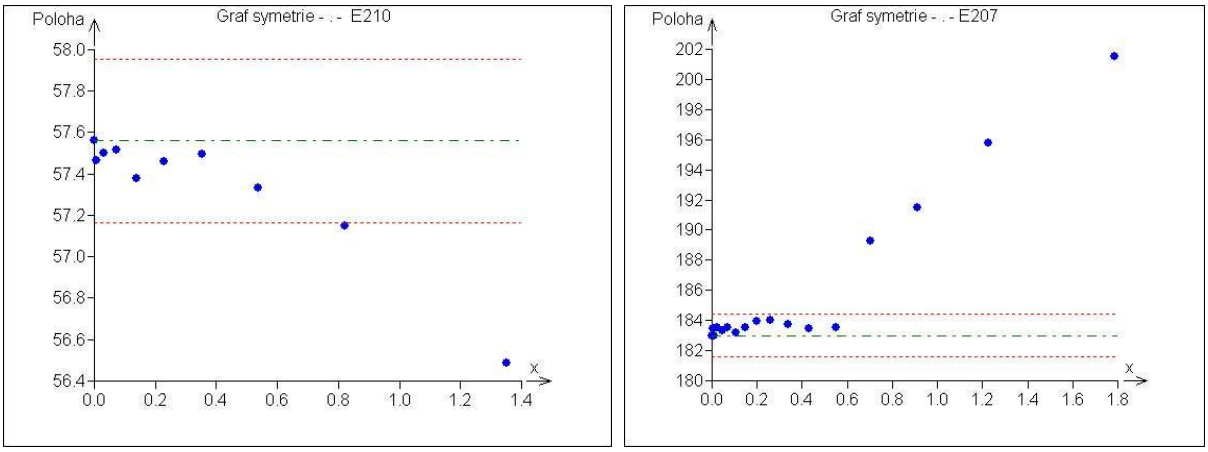
Obr. 1-4 Konstrukce (a) krabicového grafu, a (b) vrubového krabicového grafu z dat diagramu rozptýlení pro výběry z rozdělení (a) normálního, symetrického rozdělení, (b) asymetrického rozdělení. Prázdná kolečka indikují vybočující hodnoty.

Krabicový graf (osa *x*: úměrná hodnotám *x*, osa *y*: libovolný interval) umožňuje vedle znázornění robustního odhadu polohy, mediánu *M* také posouzení symetrie v okolí kvantilů a posouzení symetrie u konců rozdělení a často i identifikaci odlehlých dat. Jde o obdélník délky $R_F = F_H - F_D = \tilde{x}_{0.75} - \tilde{x}_{0.25}$ s vhodně zvolenou šířkou, která je úměrná hodnotě $\sqrt{n}$. V místě mediánu je vertikální čára. Od obou protilehlých stran tohoto obdélníku pokračují úsečky. Ty jsou ukončeny *přilehlými hodnotami* B_{PH} a B_{PD}, ležícími uvnitř *vnitřních hradeb* nejblíže k jejich hranicím B_H, B_D, tj. $B_H = F_H + 1.5R_F$ a $B_D = F_D - 1.5R_F$. Pro data pocházející z normálního rozdělení platí $B_H - B_D = 4.2$. Prvky výběru mimo vnitřní hradby jsou považovány za podezřelá měření (kroužky). Obdobou je **vrubový krabicový graf**, který umožňuje i posouzení variability mediánu, vyjádřenou robustním intervalem spolehlivosti $I_D \le M \le I_H$.



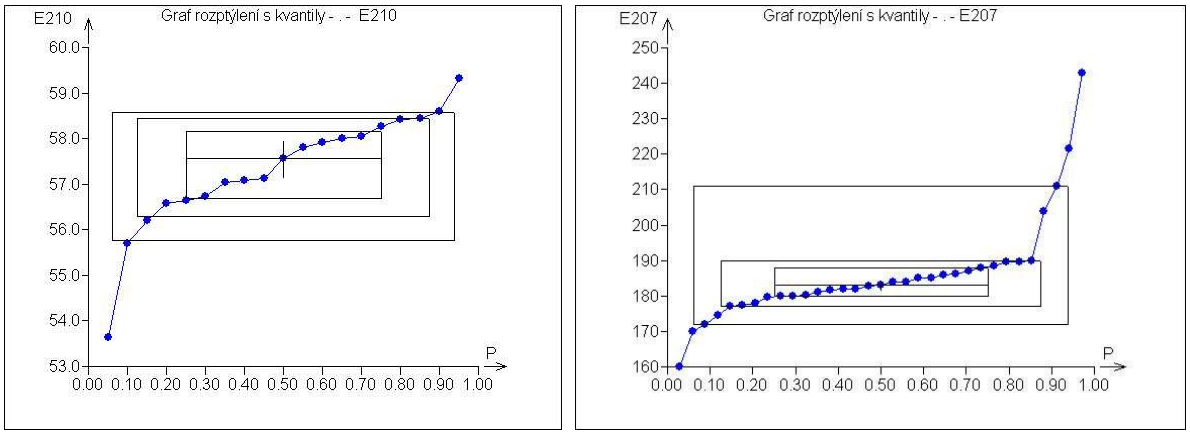
Obr. 1-5 Grafy polosum pro výběry z rozdělení (a) normálního, symetrického rozdělení, (b) asymetrického rozdělení.

Graf polosum (osa *x*: pořádkové statistiky $x_{(i)}$, osa *y:* $Z_i = 0.5(x_{(n+1-i)} + x_{(1)})$ diagnostikuje tak, že pro symetrické rozdělení je grafem horizontální přímka, určená rovnicí $\tilde{x}_{0.5} = M$.



Obr. 1-6 Grafy symetrie pro výběry z rozdělení (a) normálního, symetrického rozdělení, (b) asymetrického rozdělení.

Graf symetrie *(osa x:* $u_{P_i}^2 / 2$ pro $P_i = i/(n+1)$, *osa y:* $Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$ je obdobou předešlého grafu, u kterého symetrická rozdělení vykazují horizontální přímku $y = \tilde{x}_{0.5} = M$. Pokud tato přímka nemá nulovou směrnici, je směrnice odhadem parametru šikmosti, asymetrie.



Obr. 1-7 Konstrukce grafu rozptýlení s kvantily pro výběry z rozdělení (a) normálního, symetrického rozdělení, (b) asymetrického rozdělení.

Graf rozptýlení s kvantily (osa x: P_i, osa y: $x_{(i)}$) představuje vlastně kvantilový graf, který se získá spojením bodů $\{x_{(i)}, P_i\}$ lineárními úseky a pro symetrická rozdělení nabývá tato kvantilová funkce sigmoidálního tvaru. Pro rozdělení sešikmená k vyšším hodnotám je konvexně rostoucí a pro rozdělení sešikmená k nižším hodnotám konkávně rostoucí. Do kvantilového grafu se zakreslují tři obdélníky *F, E* a *D*:

(1) *Kvartilový obdélník F:* na ose *x* pravděpodobnosti $P_2 = 2^{-2} = 0.25$ a $1 - 2^{-2} = 0.75$.

(2) *Oktilový obdélník E*: na *y* oktily E_D a E_H a na ose *x* $P_3 = 2^{-3} = 0.125$ a $1 - 2^{-3} = 0.875$.

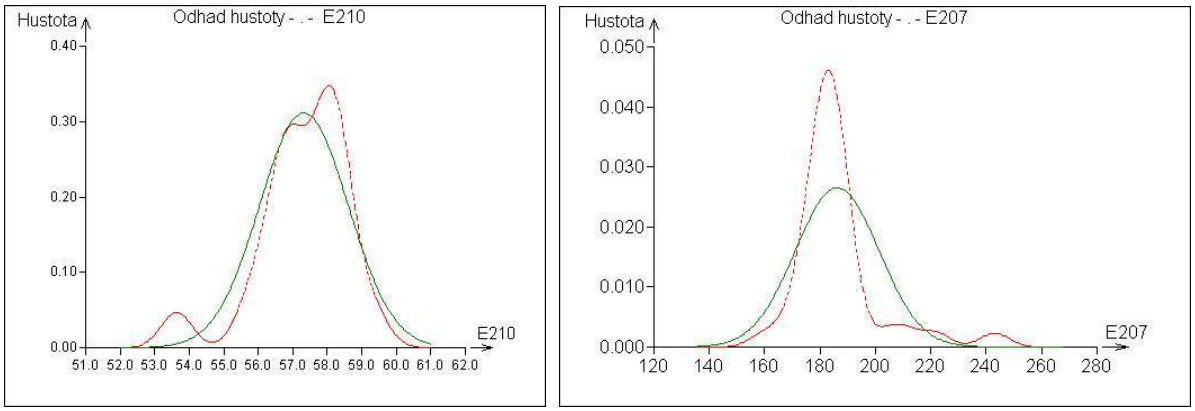
(3) *Sedecilový obdélník D*: na *y* sedecily D_D, D_H a na *x* $P_4 = 2^{-4} = 0.0625$ a $1 - 2^{-4} = 0.9375$.

Tato pomůcka může diagnostikovat i určité anomálie:

(a) Symetrické unimodální rozdělení výběru obsahuje obdélníky symetricky uvnitř sebe.

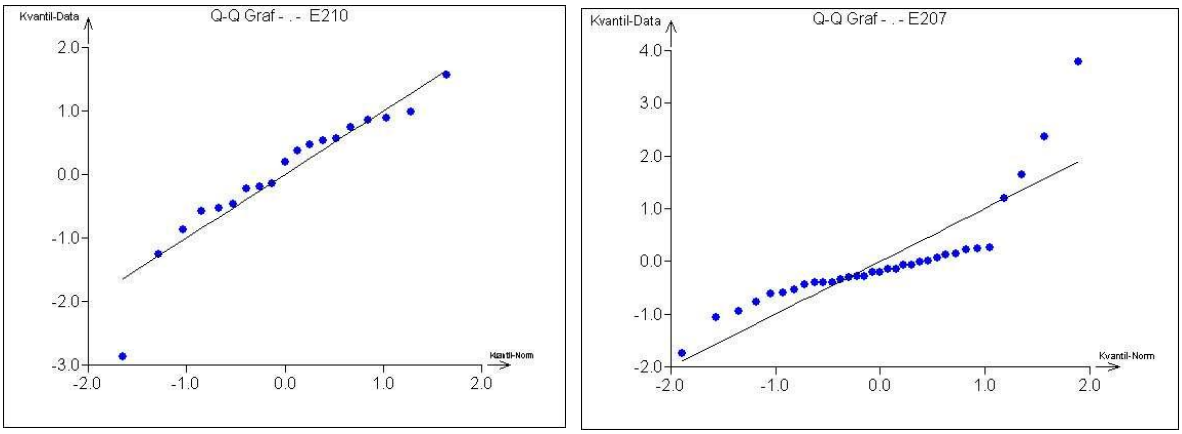
(b) Nesymetrická rozdělení mají pro rozdělení sešikmené k vyšším hodnotám vzdálenosti mezi dolními hranami obdélníků *F, E* a *D* výrazně kratší než mezi jejich horními hranami.

(c) Odlehlá pozorování jsou indikována tím, že na kvantilové funkci mimo obdélník *F* se objeví náhlý vzrůst.



Obr. 1-8 Jádrové odhady hustoty pravděpodobnosti pro výběry z rozdělení (a) normálního, symetrického rozdělení, (b) asymetrického rozdělení . Čárkovaně je znázorněna hustota Gaussova rozdělení s parametry $\bar{x}$ a s^2 a plnou čarou jádrový odhad hustoty pravděpodobnosti empirického rozdělení výběru.

Jádrový odhad hustoty pravděpodobnosti (osa *x*: *x*, osa *y*: hustota pravděpodobnosti) a **histogram** patří k nejužívanějším pomůckám a histogram pak k nejstarším diagramům hustoty pravděpodobnosti. U histogramu jde o obrys sloupcového grafu, kde jsou na ose *x* jednotlivé třídy, definující šířky sloupců, a výšky sloupců odpovídají empirickým hustotám pravděpodobnosti. Kvalitu histogramu ovlivňuje ve značné míře volba počtu tříd *L* a všech délek intervalů Δx_j. Pro přibližně symetrická rozdělení výběru lze vyčíslit *L* podle vztahu $L = \mathrm{int}(2\sqrt{n})$, kde funkce int(*x*) označuje celočíselnou část čísla *x,* nebo je možné užít výraz $L = \mathrm{int}(2.46(n-1)^{0.4})$.



Obr. 1-9 Grafy *Q-Q* pro porovnání rozdělení výběru normálního rozdělení s teoretickým rozdělením.

Kvantil-kvantilový graf (graf Q-Q) (osa *x*: $Q_T(P_i)$, osa *y*: $x_{(i)}$) umožňuje posoudit shodu výběrového rozdělení, charakterizovaného kvantilovou funkcí $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$. Za odhad

kvantilové funkce výběru se užívají pořádkové statistiky $x_{(i)}$. Při shodě výběrového rozdělení se zvoleným teoretickým rozdělením musí platit přibližná rovnost kvantilů $x_{(i)} = Q_T(P_i)$, kde P_i je pořadová pravděpodobnost. Pokud je rozdělení výběru shodné se zvoleným teoretickým rozdělením, je závislost $x_{(i)}$ na $Q_T(P_i)$ lineární a výsledná závislost se nazývá graf *Q-Q*. Těsnost lineární závislosti experimentálními body lze posoudit korelačním koeficientem a využít ho jako rozhodčí kritérium při hledání typu rozdělení.

1.4. Mocninná a Boxova-Coxova transformace dat

Pokud se na základě analýzy dat zjistí, že rozdělení výběru dat se systematicky odlišuje od rozdělení normálního, vzniká problém, jak data vůbec vyhodnotit. Často je pak nejlepším řešením vhodná transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě rozdělení. Zesymetričtění rozdělení výběru je možné provést užitím prosté **mocninné transformace**

$$y = g(x) = \begin{cases} x^{\lambda} & (\lambda > 0) \\ \ln x \quad pro & (\lambda = 0) \\ -x^{-\lambda} & (\lambda < 0) \end{cases},$$

která však nezachovává měřítko a není vzhledem k exponentu λ všude spojitá a proto se hodí pouze pro kladná data. Optimální odhad exponentu λ se hledá s ohledem na optimalizaci charakteristik asymetrie (šikmosti) a špičatosti. Pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti je vhodná **Boxova-Coxova transformace**

$$y = g(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ pro & \\ \ln x & (\lambda = 0) \end{cases},$$

která je použitelná rovněž pouze pro kladná data. Rozšíření této transformace na oblast, kdy rozdělení dat začíná od prahové hodnoty x_0, spočívá v náhradě x rozdílem $(x - x_0)$, který je vždy kladný.

Graf logaritmu věrohodnostní funkce (osa *x*: λ, osa *y*: ln *L*). Pro odhad parametru λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti s tím, že pro $\lambda = \hat{\lambda}$ je rozdělení transformované veličiny *y* normální, $N(\mu_y, \sigma^2(y))$. Po úpravách bude logaritmus věrohodnostní funkce ve tvaru

$$\ln L(\lambda) = -\frac{n}{2}\ln s^2(y) + (\lambda - 1)\sum_{i=1}^{n} \ln x_i,$$

kde $s^2(y)$ je výběrový rozptyl transformovaných dat *y*. Průběh věrohodnostní funkce ln *L*(λ) lze znázornit ve zvoleném intervalu, např. $-3 \le \lambda \le 3$, a identifikovat maximum křivky, jejíž souřadnice *x* indikuje odhad $\hat{\lambda}$. Dva průsečíky křivky ln *L*(λ) s rovnoběžkou s osou *x* indikují 100(1-α)% interval spolehlivosti parametru λ. Čím bude interval spolehlivosti $+\lambda_D$, λ_H, širší, tím je mocninná nebo Boxova-Coxova transformace méně výhodná. Pokud obsahuje interval $+\lambda_D$, λ_H, i hodnotu λ = 1, není transformace ze statistického hlediska přínosem.

Zpětná transformace: Po vhodné transformaci se vyčíslí $\bar{y}$, $s^2(y)$ a potom pomocí zpětné transformace využitím Taylorova rozvoje v okolí $\bar{y}$ se odhadnou retransformované parametry polohy a rozptýlení $\bar{x}_R$ a $s^2(\bar{x}_R)$ původních dat. Uvedený postup vede vesměs k nejlepším odhadům polohy $\bar{x}_R$ a rozptýlení $s^2(\bar{x}_R)$ a je zvláště vhodný v případech asymetrického rozdělení výběru.

1.5. Intervalový odhad parametrů

Představuje interval, ve kterém se bude se zadanou pravděpodobností či statistickou jistotou (1 - α) nacházet skutečná hodnota čili "pravda" daného parametru μ. Neznámý parametr μ odhadujeme dvěma číselnými hodnotami L_D a L_H, které tvoří **meze** tzv. *intervalu spolehlivosti* čili konfidenčního intervalu. Interval spolehlivosti pokryje parametr μ s předem zvolenou, statistickou jistotou čili dostatečně velikou pravděpodobností *P* = (1 - α), což lze vyjádřit vztahem $P(L_D < \mu < L_H) = 1 - \alpha$, nazvanou *koeficient spolehlivosti* (čili konfidenční koeficient, statistická jistota). Je obyčejně roven 0.95 nebo 0.99. Parametr α se nazývá *hladina významnosti*. Interval spolehlivosti vyjadřuje tvrzení: *"Statistická*

jistota, s jakou bude "pravda" μ *ležet v náhodných mezích* L_D, L_H *je rovna právě 1 - α".* Vlastnosti intervalu spolehlivosti:

(1) Čím je rozsah výběru *n* větší, tím je interval spolehlivosti užší.

(2) Čím je odhad přesnější a má menší rozptyl, tím je interval spolehlivosti užší.

(3) Čím je vyšší statistická jistota (1 - α), tím je interval spolehlivosti širší.

Konstrukce intervalových odhadů: Postup konstrukce intervalu spolehlivosti střední hodnoty μ normálního rozdělení $N(\mu, \sigma^2)$:

1. **Velký výběr $n \geq 30$**: Když nejlepším bodovým odhadem střední hodnoty μ je výběrový průměr $\bar{x}$ s rozdělením $N(\mu, \sigma^2/n)$, pak v intervalu $\bar{x} \pm 1.96\sigma/\sqrt{n}$ leží přibližně 95% hodnot náhodných veličin výběru o rozsahu *n* a *100(1-α)%ní interval spolehlivosti střední hodnoty* μ bude vyčíslen vztahem

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}},$$

kde hodnota 1.96 je 100(1 - 0.05/2) = 97.5%ní kvantil normovaného normálního rozdělení $u_{0.975}$.

2. **Malý výběr $n \leq 30$:** v praxi obvykle neznáme směrodatnou odchylku σ ale pouze její odhad *s* a je-li $t_{1-\alpha/2}(n-1)$ je 100(1 - α/2)%ní kvantil Studentova rozdělení bude *100(1 - α)%ní interval spolehlivosti střední hodnoty* μ roven

$$\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

Meze intervalu spolehlivosti závisí vedle chyby *s* i na rozsahu výběru *n*. Pro větší rozsahy výběru ($n > 30$) lze použít místo kvantilu $t_{1-\alpha/2}$ kvantilu normovaného normálního rozdělení $u_{1-\alpha/2}$ a *100(1 - α)%ní oboustranný interval spolehlivosti rozptylu* σ^2 se vypočte dle

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)},$$

kde $\chi^2_{1-\alpha/2}(n-1)$ je horní a $\chi^2_{\alpha/2}(n-1)$ dolní kvantil rozdělení χ^2. Robustní *interval spolehlivosti mediánu* se přibližně vyčíslí

$$\tilde{x}_{0.5} - u_{1-\alpha/2}\frac{0.707s}{\sqrt{n}} \leq med \leq \tilde{x}_{0.5} + u_{1-\alpha/2}\frac{0.707s}{\sqrt{n}}$$

1.6. Analýza malých výběrů

Předem je třeba si uvědomit, že závěry z malých výběrů jsou vždy zatíženy značnou mírou nejistoty. Malých rozsahů proto užijeme jen tam, kde skutečně není možné zvýšit počet měření.

Hornův postup pro malé výběry, $4 \leq n \leq 20$ je založený na pořádkových statistikách. Nejprve se určí hloubka pivotu je $H = (\text{int}((n + 1)/2))/2$ nebo $H = (\text{int}((n + 1)/2 + 1)/2$, pak dolní pivot jako $x_D = x_{(H)}$ a horní pivot dle $x_H = x_{(n+1-H)}$. Odhadem parametru polohy je potom *pivotová polosuma* $P_L = (x_D + x_H)/2$ a a odhadem parametru rozptýlení je *pivotové rozpětí* $R_L = x_H - x_D$. Lze definovat i náhodnou veličinu k testování $T_L = P_L/R_L$, která má přibližně symetrické rozdělení, jehož vybrané kvantily jsou dostupné v tabulce 1-1. Potom se *95%ní interval spolehlivosti střední hodnoty* vypočte vztahem

$$P_L - R_L t_{L,0.975}(n) \leq \mu \leq P_L + R_L t_{L,0.975}(n).$$

1.7. Test správnosti výsledku

Testy hypotéz o parametrech μ a σ^2 normálního rozdělení: soubor s $N(\mu, \sigma^2)$, výběr rozsahu n a vypočteme průměr $\bar{x}$ a směrodatnou odchylku s. Testy správnosti výsledku měření lze provést pomocí intervalu spolehlivosti dle pravidla: pokud $100(1 - \alpha)$ %ní interval spolehlivosti parametru μ obsahuje zadanou hodnotu μ_0, nelze na hladině významnosti α zamítnout hypotézu $H_0 : \mu = \mu_0$.

1.8. Závěr analýzy jednorozměrných dat

V postupu statistického vyhodnocení výsledků měření slouží průzkumová analýza dat EDA jako výhodná pomůcka k vyšetření zvláštností statistického chování dat. Z nejdůležitějších pomůcek jsou to vedle kvantilového grafu a grafu rozptýlení s kvantily i diagram rozptýlení a rozmítnutý diagram rozptýlení, krabicový graf, vrubový krabicový graf, graf polosum a symetrie, kvantil-

kvantilový graf, jádrový odhad hustoty pravděpodobnosti a histogram k určení tvaru rozdělení. U malých výběrů $4 \leq n \leq 20$ poskytuje správné odhady střední hodnoty Hornův postup pivotů. Pivotová polosuma a pivotové rozpětí umožňují vyčíslit i intervalový odhad střední hodnoty a navíc jsou oba odhady dostatečně robustní vůči asymetrii rozdělení malého výběru a i vůči odlehlým hodnotám. Studentův *t*-test správnosti analytického výsledku je ekvivalentní vůči intervalu spolehlivosti. Nachází-li se totiž hodnota μ_0 (tj. "pravda", správná hodnota, norma, standard) v intervalu spolehlivosti $[L_D; L_H]$, je stanovení správné. Exploratorní analýza předurčí volbu, zda k testu správnosti využijeme intervalový odhad aritmetického průměru v případě symetrického rozdělení nebo retransformovaného průměru v případě asymetrického rozdělení. Interaktivní statistická analýza při užití vhodného software umožňuje jednoznačně vyšetřit správnost analytického výsledku.

1.9. Literatura

[1] M. Meloun, J. Militký: *Statistické zpracování experimentálních dat,* Plus Praha 1994 (1. vydání), East Publishing 1996 (2. vydání), Academia Praha 2004 (3. vydání).

[2] M. Meloun, J. Militký: *Kompendium statistického zpracování dat,* Academia Praha 2002.

[3] ADSTAT, TriloByte Statistical Software s. r. o., Pardubice 1990.

2. METODOLOGIE POČÍTAČOVÉ ANALÝZY ROZPTYLU, ANOVA

2.1. Úvod

Analýza rozptylu, označovaná ANOVA (z anglického **An**alysis **o**f **Va**riance), se v technické praxi používá buď jako samostatná technika nebo jako postup umožňující analýzu zdrojů variability u statistických modelů. ANOVA jako samostatná technika umožňuje posouzení významnosti zdrojů variability v datech, vlivu přípravy vzorků na výsledek analýzy, vlivu typu přístroje, lidského faktoru a obsluhy na výsledek měření. Podstatou analýzy rozptylu je rozklad celkového rozptylu dat na *složky objasněné*, jež představují známé zdroje variability a *složku neobjasněnou*, náhodnou čili šum. Následně se testují hypotézy o významnosti jednotlivých zdrojů variability. Podle konkrétního uspořádání experimentu existuje řada variant analýzy rozptylu. Přehled základních technik lze nalézt v řadě článků[1,2] a monografií[3-6]. Často se ANOVA vyskytuje v technické praxi v souvislosti s technikami plánovaných experimentů. Omezíme se zde na jednodušší techniky, vhodné k řešení běžných vodohospodářských úloh.

2.2. Základní pojmy

Historicky se analýza rozptylu začala rozvíjet zejména při vyhodnocování dat v zemědělství. Její terminologie je proto poněkud speciální. Vedle kvalitativních faktorů se vyskytují také *faktory kvantitativní*, jako jsou fyzikální a chemické veličiny. Jednotlivé faktory se vyskytují na jistých *úrovních* Z_1, Z_2, Z_3, jež se označují jako *zpracování*. Tyto úrovně mohou být opět kvalitativní nebo kvantitativní. Zdrojem variability výsledků měření y_{ij} jsou jednotlivé úrovně faktoru. Tomu odpovídá jednoduchý model $y_{ij} = \mu_i + \varepsilon_{ij}$, kde μ_i je skutečná hodnota výsledků analýz a ε_{ij} pak označuje náhodnou chybu. Veličina μ_i se skládá ze složky odpovídající celkovému průměru μ ze všech úrovní faktoru a *efektu* i-té úrovně daného faktoru α_i, tj. $\mu_i = \mu + \alpha_i$, kde μ_i je střední hodnota pro i-tou úroveň. Účelem analýzy rozptylu je testování shody jednotlivých úrovní, čili nulové hypotézy H_0: $\mu_1 = \mu_2 = \mu_3$, nebo jinak vyjádřeno významnosti efektů α_i čili nulové hypotézy H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Pokud jsou předmětem zájmu

pouze rozdíly mezi danými úrovněmi, jde o modely s *pevnými efekty*. Pokud jsou jednotlivé úrovně pouze *výběrem* z konečného či nekonečného souboru, jde o modely s *náhodnými efekty*. Výběr mezi pevnými a náhodnými efekty závisí na vlastním záměru analýzy rozptylu a může se podle něho měnit. Je-li sledován pouze jeden faktor, jde o *jednofaktorovou analýzu rozptylu*, čili třídění dle jednoho faktoru. Často se však sleduje i vliv několika faktorů, kdy jde o *vícefaktorovou analýzu rozptylu*. Jako u jednofaktorové analýzy rozptylu, můžeme provést rozklad μ_{ij} na celkovou střední hodnotu, složky α_i odpovídající efektům faktoru *Z*, složky β_j odpovídající efektům faktoru *L* a interakce τ_{ij}, $\mu_{ij} = \mu + \alpha_i + \beta_j + \tau_{ij}$. Člen τ_{ij} označuje efekt *interakce* úrovní Z_i a L_j. Používá se v případech, kdy nelze objasnit variabilitu y_{ijk} pouze aditivním působením jednotlivých faktorů. Pro vlastní zpracování modelů analýzy rozptylu je důležité, zda je při všech kombinacích faktorů proveden stejný počet měření čili opakování. Kombinace úrovní jednotlivých faktorů, např. Z_i L_j se pak označuje jako *cela*. Pro stejný počet opakování ve všech celách se *experimenty* označují jako *vyvážené*, zatímco pro nestejný počet opakování jako *nevyvážené*. Postupy analýzy nevyvážených experimentů jsou komplikovanější a navíc může při extrémních rozdílech mezi počty opakování dojít při malých odchylkách od základních předpokladů, např. normality, ke značnému zkreslení výsledků testů[5].

2.3. Jednofaktorová analýza rozptylu

Při třídění podle jednoho faktoru se zkoumá jeho vliv na výsledek experimentu. Pro případ dvou úrovní jde o porovnání dvou výběrů. Zajímavý bude obecnější případ, kdy daný faktor *A* má celkem *K* různých úrovní A_1, ..., A_K. Na každé úrovni A_i je provedeno n_i měření $\{y_{ij}\}$, $j = 1, \ldots, n_i$. Celkový počet měření je $N = \sum_{i=1}^{K} n_i$ Přehlednější je uspořádání dat v Tabulce 2-1.

Tabulka 2-1. Uspořádání dat pro jednofaktorovou analýzu rozptylu

	Úroveň faktoru						
	A_1	A_2	...	A_i	...	A_K	Celek
	y_{11}	y_{21}	...	y_{i1}	...	y_{K1}	
	y_{12}	y_{22}	...	y_{i2}	...	y_{K2}	
	...	...	...	...	...	...	
	...	...	...	...	...	...	
Opakování Měření	y_{1n_1}	y_{2n_2}	....	y_{in_i}	...	y_{Kn_K}	
Průměry	$\hat{\mu}_1$	$\hat{\mu}_2$	...	$\hat{\mu}_i$	...	$\hat{\mu}_K$	$\hat{\mu}$
Počet	n_1	n_2	...	n_i	...	n_K	N

Sloupcový průměr $\hat{\mu}_i$ představuje součet prvků sloupce pro A_i dělený počtem opakování n_i, $\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$. *Celkový průměr* $\hat{\mu}$ je součet všech hodnot dělený celkovým počtem dat $\hat{\mu} = \frac{1}{K}\sum_{i=1}^{K} \hat{\mu}_i$. Pro výpočet odhadů efektů α_i lze pak použít vztah $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}$. Při zavedení μ_i vznikne přeurčený model, obsahující o jeden parametr více. Proto se při odhadu efektů α_i používá ještě jedna omezující

podmínka $\sum_{i=1}^{K} n_i \alpha_i = 0$. Pro případ vyvážených experimentů lze použít zjednodušenou podmínku $\sum_{i=1}^{K} \alpha_i = 0$. Vlastní analýza rozptylu, tj. rozklad celkového rozptylu, závisí také na tom, zda jde o modely s pevnými nebo náhodnými efekty.

Základním předpokladem statistické analýzy je fakt, že náhodné chyby ε_{ij} jsou nezávislé a náhodné veličiny s normálním rozdělením $N(0, \sigma^2)$. Střední hodnota chyb je rovna nule a rozptyl σ^2 je konstantní. Součet čtverců odchylek od celkového průměru $\hat{\mu}$, definovaný vztahem $S_c = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu})^2$ se rozloží s využitím μ_i na dvě složky $S_c = \sum_{i=1}^{K} \sum_{j=1}^{n_i} [(y_{ij} - \hat{\mu}_i) + (\hat{\mu}_i - \hat{\mu})]^2 = S_A + S_R$, kde S_A představuje součet čtverců odchylek mezi jednotlivými úrovněmi daného faktoru $S_A = \sum_{i=1}^{K} n_i (\hat{\mu}_i - \hat{\mu})^2$ a S_R je reziduální součet čtverců odchylek uvnitř jednotlivých úrovní, $S_R = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2$. Jednotlivé součty čtverců resp. složky rozptylu se zapisují do tabulky, která má pro jednofaktorovou analýzu rozptylu s pevnými efekty tvar Tabulky 2-2.

Tabulka 2-2. Tabulka analýzy rozptylu pro jednoduché třídění u modelu s pevnými efekty

Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	Očekávaná hodnota
Mezi úrovněmi SA	K - 1	$\frac{S_A}{K-1}$	$\sigma_e^2 + \frac{\sum_{i=1}^{K} n_i \alpha_i^2}{K-1}$
Reziduální SR	N - K	$\frac{S_R}{N-K}$	σ_e^2
Celkový Sc	N - 1	-	-

Poslední sloupec tabulky obsahuje očekávanou hodnotu průměrného čtverce. Nevychýleným odhadem rozptylu chyb σ_e^2 je průměrný reziduální čtverec $\sigma_e^2 = \frac{S_R}{N - K}$. Cílem je především testování, zda jsou efekty α_i nulové, tedy zda jednotlivé úrovně daného faktoru vedou ke statisticky nevýznamným rozdílům ve výsledcích. Nulová hypotéza H_0: $\alpha_i = 0$, $i = 1, ..., K$, se ověřuje proti alternativní hypotéze H_A: $\alpha_i \neq 0$, $i = 1, ..., K$. Při testování se využívá faktu, že veličina S_A / σ_e^2 má χ^2-rozdělení s $(K - 1)$ stupni volnosti a veličina S_R / σ_e^2 má nezávislé χ^2-rozdělení s $(N - K)$ stupni volnosti. Jejich podíl má pak F-rozdělení s $(K - 1)$ a $(N - K)$ stupni volnosti. Testovací Fisherova statistika F_e má tvar $F_e = \frac{S_A\,(N - K)}{S_R\,(K - 1)}$. Při platnosti nulové hypotézy H_0 má F_e statistika Fisherovo F-rozdělení s $(K - 1)$ a $(N - K)$ stupni volnosti. Vyjde-li F_e větší než kvantil Fisherova rozdělení $F_{1-\alpha}(K - 1, N - K)$, je nutné nulovou hypotézu H_0 na hladině významnosti α zamítnout a efekty považovat za nenulové a statisticky významné.

2.3.1. *Technika vícenásobného porovnání*

Pokud vyjde vliv jednotlivých efektů jako statisticky významný, jsou rozdíly mezi průměry μ_i, μ_j, $i \neq j$ rovněž významné. Pro hlubší analýzu se používá řady metod, například Scheffého metoda vícenásobného porovnání[5], pro kterou se zamítá hypotéza H_0: $\mu_i = \mu_j$ pro všechny dvojice (i, j), pro které platí

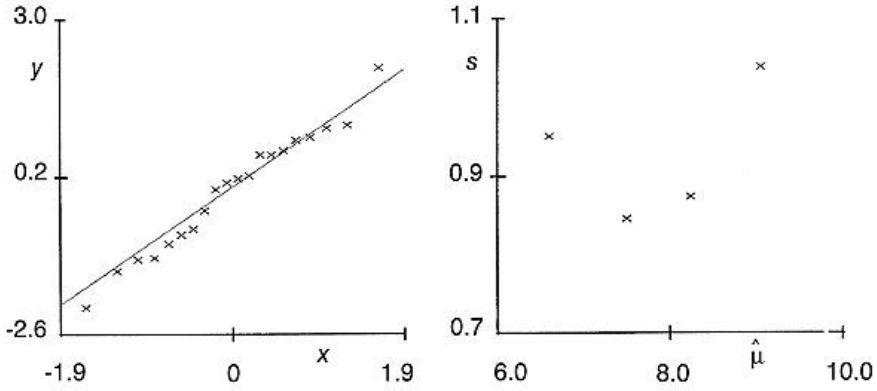
$$\left|\hat{\mu}_i - \hat{\mu}_j\right| \geq \sqrt{(K - 1)\;\;\hat{\sigma}^2\;\;F_{1-\alpha}(K - 1,\, N - K)\left[\frac{1}{n_i} + \frac{1}{n_j}\right]},$$

kde $\hat{\sigma}^2$ je reziduální rozptyl $\hat{\sigma}_e^{\,2}$. Tento vztah se používá pro všechny možné dvojice indexů (i, j). V některých případech je třeba testovat pouze zvolený *lineární kontrast* q definovaný vztahem $q = \sum_{i=1}^{K} C_i\,\mu_i$ se známými konstantami C_i, pro které platí $\sum_{i=1}^{K} C_i = 0$, $\sum_{i=1}^{K} C_i^2 > 0$. Odhadem lineárního kontrastu q je veličina $\hat{q} = \sum_{i=1}^{K} C_i\,\hat{\mu}_i$. Mají-li výsledky měření y_{ij} normální rozdělení $N(\mu_i, \sigma^2)$, lze

testovat nulovovu hypotézu H_0: $q = 0$ pomocí statistiky $F_q = \frac{\hat{q}^2}{\hat{\sigma}^2 \sum_{i=1}^{K} \frac{C_i^2}{n_i}}$. Při platnosti nulové hypotézy H_0 má tato testovací statistika *F*-rozdělení s 1 a (*N* - *K*) stupni volnosti. Hypotéza H_0 se zamítá, pokud F_q je větší než kvantil $F_{1-\alpha}(1, N - K)$. Dosavadní postupy analýzy rozptylu jsou správné jen za předpokladu, když jednotlivé hodnoty y_{ij} jsou vzájemně nezávislé, a když chyby ε_{ij} mají normální rozdělení s konstantním rozptylem. V praxi však bývá důležité tyto předpoklady rovněž ověřit.

2.3.2. *Ověření normality chyb*

Pro posouzení normality chyb lze použít především rankitové grafy. Výhodné je v těchto grafech užití standardizovaných reziduí $\hat{e}_{Si} = \frac{\hat{e}_{ij}}{\hat{\sigma}\sqrt{1 - \frac{1}{n_i}}}$. V případě platnosti předpokladů klasické analýzy rozptylu mají standardizovaná rezidua přibližně normální rozdělení $N(0, 1)$. Pokud platí podmínka, že $\varepsilon_{ij} \,.\, N(0, \sigma^2)$, vznikne v rankitovém grafu lineární závislost s nulovým úsekem a jednotkovou směrnicí.



Obr. 2-1. Rankitový graf pro Jacknife rezidua

V řadě případů je možné zlepšit rozdělení dat ve smyslu přiblížení k normalitě s využitím vhodné transformace. Častým případem je, že data jsou zešikmená směrem k vyšším hodnotám. Pak je vhodné použít např. *posunutou logaritmickou transformaci* $y^* = \ln(y + C)$. Optimální hodnota *C* se volí tak, aby rezidua byla přibližně symetrická se špičatostí blízkou hodnotě Gaussova

rozdělení tj. třem. Pro účely identifikace vybočujících hodnot je však výhodné použít Jackknife reziduí e_{Jij}, která jsou definována vztahem $\hat{e}_{Jij} = \hat{e}_{Sij}\sqrt{\frac{N-K-1}{N-K-\hat{e}^2_{Sij}}}$.

Za předpokladu normality vykazují tato rezidua Studentovo rozdělení s $(N-K-1)$ stupni volnosti. Orientačně platí, že pokud $\hat{e}^2_{Jij} > 10$, lze danou hodnotu y_{ij} považovat za velmi silně vybočující.

2.3.3. *Ověření konstantnosti rozptylu (homoskedasticity)*

Předpoklad konstantnosti rozptylu (homoskedasticity) lze ověřit stejnými metodami jako u lineárních regresních modelů. U nevyvážených plánů je třeba uvažovat s nekonstantností rozptylu klasických reziduí způsobem nestejného počtu měření na jednotlivých úrovních. Pokud je k dispozici dostatečný počet opakování při jednotlivých úrovních daného faktoru, lze kromě průměru $\hat{\mu}_i$ počítat také výběrové rozptyly s_i^2. Předpoklad konstantnosti rozptylu lze pak ověřit na základě grafu s_i vs. $\hat{\mu}_i$. Pokud vznikne náhodný shluk bodů, lze považovat předpoklad shody rozptylů u všech úrovní za přijatelný. Jinak je možné použít vhodnou transformaci stabilizující rozptyl.

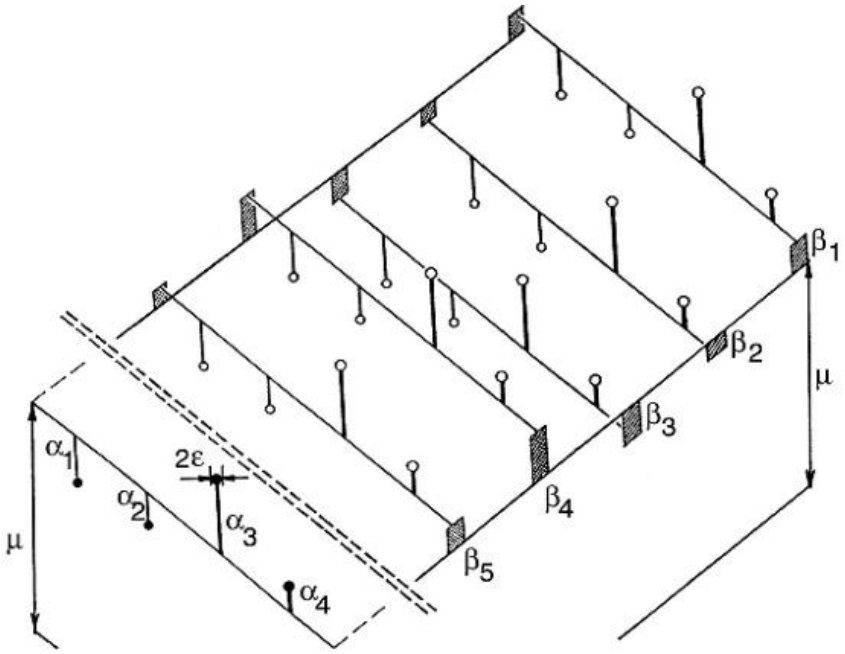
2.4. Dvoufaktorová analýza rozptylu

Při dvoufaktorové analýze rozptylu se provádí experimenty na různých úrovních dvou faktorů A a B. Kombinace úrovní faktorů tvoří typickou mřížkovou strukturu, jejímž elementem je tzv. cela. Platí, že (i, j)-tá cela odpovídá kombinaci úrovně A_i faktoru A a B_j faktoru B. Schematicky je mřížková struktura znázorněna v Tabulce 2-3: V každé cele je obecně n_{ij} pozorování. Často se však setkáváme s případem bez opakování, kdy v každé cele je pouze jediné pozorování, $n_{ij} = 1$. Pro případ analýzy rozptylu bez opakování dojde ke zjednodušení zápisu $y_{ij} = \mu_{ij} + \varepsilon_{ij}$, kde μ_{ij} lze rozložit tak, že kromě řádkových αi a sloupcových $ß_j$ efektů se zde vyskytuje také interakční člen τ_{ij}. Tento člen je pak důsledkem různých kombinací sloupcových a řádkových efektů.

Tabulka 2-3. Uspořádání dat pro dvoufaktorovou analýzu rozptylu

	B_1	B_2	...	B_M	
A_1	.	.	...	.	
A_2	.		...	.	
.	.	.	...	.	cela A_2B_2
.	.	.	...	.	
.	.	.	...	.	
A_N	.	.	...	.	

Nejjednodušším je *Tukeyův model interakce*, vyjádřený tvarem $\tau_{ij} = C\,\alpha_i\,\beta_j$, kde *C* je konstanta. Složitější jsou *řádkově lineární* modely interakcí, vyjádřené tvarem $\tau_{ij} = \gamma_i\,\beta_j\,C_R$ nebo *sloupcově lineární* modely interakcí ve tvaru $\tau_{ij} = \alpha_i\,C_K\,\delta_j$. Kompletnější je *aditivně-multiplikativní* model interakcí $\tau_{ij} = \gamma_i\,\delta_j\,C_W$. Uvedené vztahy obsahují kromě sloupcových a řádkových konstant δ_j a γ_i i obecné konstanty C_R, C_K, C_W. Omezme se zde pouze na nejjednodušší Tukeyův model interakce. Vzhledem ke své speciální definici obsahuje tento model pouze jeden parametr C, a proto se označuje jako model *neaditivity s jedním stupněm volnosti.* Použití Tukeyova modelu interakce je výhodné zejména v případech, kdy je v každé cele pouze jedno pozorování, obr. 2-2.



Obr. 2-2. Model dvojného třídění $\mu_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, kde α_i je řádkový efekt a β_j sloupcový efekt a náhodná chyba ε_{ij} odpovídá poloměru kolečka.

U modelů bez opakování měření obsahuje každá cela právě jednu hodnotu y_{ij}. O chybách ε_{ij} se předpokládá, že jsou to nezávislé a shodně rozdělené náhodné veličiny s nulovou střední hodnotou a konstantním rozptylem. Při testování se navíc předpokládá, že rozdělení chyb je normální. Pokud v ANOVA modelu provedeme rozklad, je model analýzy rozptylu přeurčený. Proto se definují omezující podmínky $\sum_{i=1}^{N}\alpha_i = 0;\ \sum_{j=1}^{M}\beta_j = 0;\ \sum_{i=1}^{N}\tau_{ij} = 0;\ \sum_{j=1}^{M}\tau_{ij} = 0$. V případě pouze aditivního působení jednotlivých faktorů je $\tau_{ij} = 0$ pro všechna $i = 1, ..., N$ a $j = 1, ..., M$. Odhady parametrů μ, α_i, β_j lze pak určit ze vztahů $\hat{\mu} = \frac{1}{N\,M}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$, $\alpha_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij} - \hat{\mu}$, $\hat{\beta}_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij} - \hat{\mu}$. Pro rezidua $\hat{e}_{ij}$ platí $\hat{e}_{ij} = y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. K určení interakcí můžeme využít skutečnosti, že $\tau_{ij} = E(y_{ij}) - \mu - \alpha_i - \beta_j$ a pro odhad interakcí platí přibližně $\hat{\tau}_{ij} \approx \hat{e}_{ij}$. Lze snadno identifikovat *Tukeyův model* interakce. Platí-li tento model, vyjde na grafu $\hat{e}_{ij}$ vs. $\hat{\alpha}_i\ \hat{\beta}_j$ lineární trend. Ze

směrnice odpovídající regresní přímky se odhadne parametr *C*. Platí pro něj výraz $\hat{C} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{e}_{ij}\ \hat{\alpha}_i\ \hat{\beta}_j}{\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\alpha}_i^2\ \ \hat{\beta}_j^2}$. Graf $\hat{e}_{ij}$ vs. $\hat{\alpha}_i\ \hat{\beta}_j\ /\ \hat{\mu}$ se označuje jako *graf neaditivity*.

Pokud vyjde v tomto grafu nenáhodný trend, znamená to, že je třeba uvažovat interakce.

Tabulka 2-4. Tabulka analýzy rozptylu pro dvojné třídění s interakcí Tukeyova typu

Součet čtverců pro	Stupně volnosti	Průměrný čtverec	Kritérium *F*
Faktor *A*, $S_A = M\sum_{i=1}^{N}\alpha_i^2$	*N* - 1	$M_A=S_A/(N-1)$	$F_A = M_A/M_{AB}$
Faktor *B*, $S_B = N\sum_{j=1}^{M}ß_j^2$	*M* - 1	$M_B=S_B/(M-1)$	$F_B = M_B/M_{AB}$
Interakce (Tukey) S_T	1	$M_T = S_T$	$F_T = M_T/M_E$
Reziduální $S_R = S_{AB}-S_T$	*N M - N - M*	$M_E = S_R/(NM-N-M)$	-
Celkový $S_C = \sum_{(i)}\sum_{(j)}(\hat{\mu} - y_{ij})^2$	*N M* - 1	-	-

V tabulce 2-4 představuje S_T součet čtverců odchylek odpovídající Tukeyově interakci[3]

$$S_T = \frac{\left(\sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \ \hat{\alpha}_i \ \hat{\beta}_j \right)^2}{\sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\alpha}_i^2 \ \beta_j^2}.$$

Symbol S_{AB} označuje reziduální součet čtverců pro případ bez interakcí $S_{AB} = \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2$ a průměrný čtverec je $M_{AB} = \frac{S_{AB}}{(N-1)\,(M-1)}$. Hodnota M_{AB} je nevychýleným odhadem rozptylu σ^2. Pomocí *F*-kritéria lze opět provádět statistické testy. Začíná se testováním nulové hypotézy H_0: "Tukeyova interakce je nevýznamná", pro kterou lze použít testovací statistiku F_T z tabulky 2-4. Za předpokladu platnosti nulové hypotézy H_0 má tato testovací statistika *F*-rozdělení s 1 a (*N M* - *N* - *M*) stupni volnosti. Pokud nelze tuto hypotézu zamítnout, testuje se nulová hypotéza H_0: $\alpha_i = 0$, $i = 1, ..., N$ (efekty řádků, čili faktoru *A*, jsou nevýznamné) pomocí statistiky F_A nebo nulová hypotéza H_0: $\beta_j = 0$, $j = 1, ..., M$ (efekty sloupců, čili faktoru *B*, jsou nevýznamné) pomocí statistiky F_B. Obě tyto testovací statistiky jsou uvedeny v Tabulce 2-4. Za předpokladu platnosti hypotézy H_0 má statistika F_A *F*-rozdělení s $(N - 1)$ a $(N - 1)(M - 1)$ stupni volnosti a statistika F_B také *F*-rozdělení s $(M - 1)$ a $(N - 1)(M - 1)$ stupni volnosti. Pokud však vyjde F_T vyšší než odpovídající kvantil *F*-rozdělení, je efekt Tukeyovy interakce významný. V některých případech je výhodné provést *eliminaci neaditivity* s využitím mocninné transformace

$$y_{ij}^{*} = \begin{cases} \dfrac{(y_{ij} + K)^{\lambda} - 1}{\lambda} & \text{pro } \lambda \neq 0 \\ \ln\,(y_{ij} + K) & \text{pro } \lambda = 0 \end{cases},$$

kde *K* je vhodně volená konstanta zajišťující, aby platilo $(y_{ij} + K) > 0$, a kde λ je parametr transformace.

2.4.1. *Vyvážené modely*

Pro vyvážené modely platí, že v každé cele je $n_{ij} = n$ pozorování. Odhadem μ_{ij} jsou aritmetické průměry $\hat{\mu}_{ij} = \frac{1}{n} \sum_{k=1}^{n} y_{ijk}$. Pro odhady ostatních parametrů se použijí vztahy

$$\hat{\mu} = \frac{1}{N\,M}\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\mu}_{ij}\,,\ \hat{\alpha}_i = \frac{1}{M}\sum_{j=1}^{M}\hat{\mu}_{ij} - \hat{\mu}\,,\ \hat{\beta}_j = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_{ij} - \hat{\mu}\,.$$

Rezidua vyjádříme vztahem $\hat{e}_{ijk} = y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. Podobně lze i v tomto případě definovat odhad interakcí $\hat{\tau}_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\alpha}_i - \beta_j$. Povšimněme si, že tento vztah se liší jen tím, že se místo veličin y_{ij} používá průměrů $\hat{\mu}_{ij}$. Pro ověření Tukeyova modelu interakce neaditivity lze vynášet graf $\hat{\tau}_{ij}$ vs. $\hat{\alpha}_i\ \hat{\beta}_j$. Náhodný obrazec bodů zde dokazuje aditivní působení obou faktorů. Součty čtverců modelu analýzy rozptylu pro obecný případ interakcí jsou uvedeny v Tabulce 2-5.

Odpovídající průměrné hodnoty (očekávané hodnoty) průměrných čtverců jsou

$$E(M_A) = \sigma^2 + \frac{n\,M\sum_{i=1}^{N}\alpha_i^{\,2}}{(N-1)\,\sigma^2} = \sigma^2 + n\,M\,\sigma_A^2\,,$$

$$E(M_B) = \sigma^2 + \frac{n\,N\sum_{j=1}^{M}\beta_j^{\,2}}{(M-1)\,\sigma^2} = \sigma^2 + n\,N\,\sigma_B^2$$

a $$E(M_{AB}) = \sigma^2 + \frac{n\sum_{i=1}^{N}\sum_{j=1}^{M}\tau_{ij}^2}{(N-1)\,(M-1)\,\sigma^2} = \sigma^2 + n\,\sigma_{AB}^2\,.$$

Očekávaná hodnota $E(M_R) = \sigma^2$ ukazuje, že rozptyl M_R je nevychýleným odhadem σ^2 rozptylu chyb. Rozptyly σ_A^2, σ_B^2 a σ_{AB}^2 odpovídají efektům řádků, sloupců a interakcí.

Tabulka 2-5. Tabulka analýzy rozptylu pro dvojné třídění a vyvážený experiment

Součet čtverců pro	Stupně volnosti	Průměrný čtverec	Kritérium *F*
Faktor *A*,	*N* - 1	$M_A = \frac{S_A}{N-1}$	

$S_A = n M \sum_{i=1}^{N} \hat{\alpha}_i^2$			$F_A = \frac{M_A}{M_R}$

Faktor *B,*

$S_B = n N \sum_{j=1}^{M} \hat{\beta}_j^2$	M - 1	$M_B = \frac{S_B}{M - 1}$	$F_B = \frac{M_B}{M_R}$

Interakce *AB,*

$S_{AB} = n \sum_{i=1}^{N} \sum_{j=1} \hat{\tau}_{ij}^2$	(N - 1)(M - 1)	$M_{AB} = \frac{S_{AB}}{(N - 1)\ (M - 1)}$	$F_{AB} = \frac{M_{AB}}{M_R}$

Reziduální

$S_R = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{n} (y_{ijk} - \hat{\mu}_{ij})^2$	$M N (n$ - 1)	$M_R = \frac{S_R}{M\ \ N\ (n - 1)}$	-

Celkový

$S_C = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{n} (y_{ijk} - \hat{\mu})^2$	$M N n$ - 1	-	-

Těchto vztahů lze využít i v případech, kdy se hledají odhady rozptylů příslušející faktorům a interakcím. Pak se místo středních hodnot *E*(.) dosazují přímo průměrné čtverce a místo rozptylu σ^2 reziduální rozptyl $\hat{\sigma}^2$. Důležité je, že průměrné čtverce nejsou přímo odhady odpovídajících rozptylů. Také v případě analýzy rozptylu definované Tabulkou 2-5 se využitím statistik F_{AB}, F_B a F_A testuje, zda je možné považovat sloupcové a řádkové efekty, resp. interakce, za nevýznamné.

Pro test nulové hypotézy H_0: $\tau_{ij} = 0$, $i = 1, ..., N$ a $j = 1, ..., M$, lze použít testační statistiku F_{AB}, která má za předpokladu platnosti hypotézy H_0 F-rozdělení s $\{(N - 1)(M - 1)\}$ a $\{M\ N\ (n - 1)\}$ stupni volnosti. Při testování významnosti řádkových efektů faktoru A je H_0: $\alpha_i = 0$, $i = 1, ..., N$. Pokud nulová hypotéza platí, má testovací F_A statistika F-rozdělení s $(N - 1)$ a $\{M\ N\ (n - 1)\}$ stupni volnosti. Analogicky při testování významnosti sloupcových efektů faktoru B je H_0: $\beta_j = 0$, $j = 1, ..., M$. Pokud nulová hypotéza platí, má testovací F_B statistika F-rozdělení s $(M - 1)$ a $\{M\ N\ (n - 1)\}$ stupni volnosti. Nevychýleným odhadem rozptylu je M_R. Výhodou vyvážených experimentů je fakt, že jednotlivé složky modelů analýzy rozptylu jsou vzájemně nezávislé. Proto je možno sčítat jednotlivé dílčí součty čtverců v tabulce 2-5 a 2-4, což umožňuje současné testování několika hypotéz. V podstatě lze tímto jednoduchým postupem ověřovat platnost různých submodelů analýzy rozptylu od nejjednoduššího $y_{ijk} = \mu + \varepsilon_{ijk}$ přes všechny dílčí (obsahující jen některé z parametrů α, β a τ). Sčítání součtů čtverců se doporučuje i v případech, kdy se vliv některých faktorů či interakcí prokáže jako nevýznamný. Pak se příslušný součet čtverců přičte k reziduálnímu a v modelu se odpovídající členy vypustí.

2.5. Souhrn: Postup při analýze rozptylu

Obecně lze postup analýzy rozptylu rozdělit do těchto kroků:

1. Provede se odhad parametrů základního modelu ANOVA.

2. Provede se testování jeho významnosti a konstrukce různých submodelů u modelů s pevnými efekty.

3. Provede se ověření předpokladů normality, homogenity rozptylů a přítomnosti silně vybočujících pozorování. Ne vždy se pro tyto účely hodí klasicky definovaná rezidua a využívají se proto i jiné typy reziduí.

4. Provede se interpretace výsledků s ohledem na zadání dat a jejich případné úpravy.

2.6. Doporučená literatura:

[1] Searle S. R.: *Biometrics* **27**, 1 (1971).

[2] Bartlett M. S., Kendall D. G.: *J. Roy. Stat. Soc.* **B8**, 128 (1946).

[3] Scheffé H.: *The Analysis of Variance.* J. Wiley, New York 1959.

[4] Searle S. R.: *Linear Models.* J. Wiley, New York 1971.

[5] Miller P. G.: *Beyond ANOVA, Basics of Applied Statistics.* J. Wiley, New York 1986.

[6] Speed T. P.: *Annals of Statist.* **15**, 885 (1987).

[7] Emerson J. D., Hoaglin D. C., Kempthorne P. I.: *J. Amer. Statist. Assoc.* **79**, 329 (1984).

[8] Bradu D., Hawkins D. M.: *Technometrics* **24**, 103 (1982).

[9] Bloomfield P., Steiger W.: *Least Absolute Deviations: Theory, Applications and Algorithms.* Birkhäuser, Boston, 1983.

[10] Gabriel K. R.: *J. R. Stat. Soc.* **B40**, 186 (1978).

[11] Cressie N. A. C.: *Biometrics* **34**, 505 (1978).

[12] Potocký R a kol.: *Zbierka úloh z pravdepodobnosti a matematickej štatistiky,* ALFA-SNTL Bratislava 1986.

[13] Anderson R. L.: *Practical Statistics for Analytical Chemists,* van Nostrand Reinhold Comp., New York 1987.

[14] Miller J. C., Miller J. N.: *Statistics for Analytical Chemistry*, Ellis Horwood Chichester 1984.

[15] Liteanu C., Rica I.: *Statistical Theory and Methodology of Trace Analysis*, Ellis Horwood Chichester 1980.

[16] Rice J. A.: *Mathematical Statistics and Data Analysis*, Wadsworth & Brooks, California 1988, str. 397.

[17] Meloun M., Militký J.: *Statistické zpracování experimentálních dat,* Plus Praha 1994, resp. East Publishing Praha 1998, Academia Praha 2004.

[18] Meloun M., Militký J.: *Kompendium statistického zpracování dat,* Academia Praha 2002, Academia Praha 2006.

[19] ADSTAT 1.25, 2.0 a verze 3.0, TriloByte Statistical Software Pardubice, 1992, 1993, 1999.

3. INTERVALOVÉ ODHADY A MÍRY PŘESNOSTI V KALIBRACI

3.1. Úvod

Kalibrace patří k základním úlohám, které se řeší s využitím regresních metod. Používá se při konstrukci snímačů fyzikálních veličin, adjustaci přístrojů a při vývoji nových instrumentálních metod. Skládá se ze dvou fází: (a) sestavení kalibračního modelu, a (b) použití kalibračního modelu. Sestavení kalibračního modelu je totožné s úlohou hledání regresního modelu, tj. pro závislost signálu *y* na cílové veličině *x*.Ve druhé fázi se řeší úloha inverzní, tj. pro naměřenou odezvu *y** se hledá odpovídající hodnota *x** a její statistické charakteristiky. Z numerického hlediska jde o úlohu hledání kořene nelineární funkce. Ze statistického hlediska jde o problém nelineárního vychýleného odhadu.

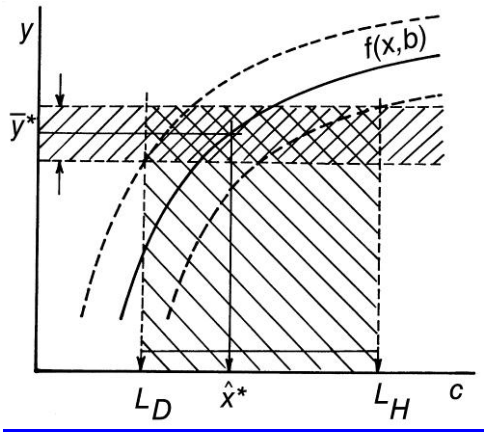
3.2. Druhy kalibrace

Rozdělení kalibračních úloh podle různých hledisek, určujících způsob statistického zpracování, je uvedeno v práci Rosenblatta a Spiegelmana [1]:

(1) *Absolutní kalibrace* je nejběžnější v technické praxi. Při sestavení kalibračního modelu se hledá vztah mezi měřitelnou veličinou η, nazývanou signál (potenciál, napětí článku, proud, elektrický odpor, pH, absorbance, atd.), a cílovou veličinou ξ, která určuje stav nebo vlastnosti systému (obsah, koncentrace, objem, teplota, čas, atd.). Ta může být obtížněji měřitelná a bývá proto určována jako výsledek druhé fáze kalibrace. Příkladem může být kalibrace absorbance roztoku (η) na jeho koncentraci (ξ). Při *kalibračním experimentu* se u *n* vzorků se známými nebo přesně měřitelnými hodnotami proměnné ξ změří odpovídající veličiny η. Vzhledem k tomu, že jsou ve většině případů obě veličiny naměřené, bude pro *n* bodů $\{x_i, y_i\}$, $i = 1, ..., n$, platit $y_i = \eta_i + \varepsilon_i$, $x_i = \xi_i + \delta_i$, kde ε_i a δ_i jsou experimentální chyby. Pokud je proměnná ξ_i měřena přesně, nebo je užito definovaných standardů, je $\delta_i = 0$, $i = 1, ..., n$. Veličina η_i je aproximována kalibračním modelem $f(x, \boldsymbol{\beta})$ a úloha zpracování dat vede na odhad parametrů $\boldsymbol{\beta}$. Ve fázi *použití kalibračního modelu* se pro naměřené hodnoty při *M* opakování měření signálu $\{y_j^*\}$, $j = 1, ..., M$ určuje průměrná hodnota vlastnosti $\hat{x}^*$ a její interval spolehlivosti. Případ

kalibrace absorbance na koncentraci je znázorněn na obr. 3-1 u kalibrační křivky a obr. 3-2 u kalibrační přímky, kde symboly L_D a L_H značí dolní a horní mez asymetrického nebo symetrického konfidenčního intervalu koncentrace. V tomto článku o kalibraci budeme věnovat pozornost především absolutní kalibraci.

(2) *Komparativní kalibrace* je postup, při kterém se jeden přístroj kalibruje vůči druhému a je libovolné, který se užije jako standard. Příkladem může být měření koncentrace z *absorbance* dle Beerova zákona (1. metoda) a *titračně* (2. metoda). Porovnávají se hodnoty absorbance vůči objemu titračního činidla. Chyby δ_i nejsou zanedbatelné a je třeba při sestavení kalibračního regresního modelu užít ortogonální regrese, tj. regrese pro případ obou proměnných zatížených náhodnými chybami, šumem.



Obr. 3-1 Zadání absolutní kalibrace a postup při určení koncentrace pro průměrnou hodnotu signálu z kalibrační křivky. L_D, L_H jsou dolní a horní mez konfidenčního intervalu. Šrafovaně je vyznačen konfidenční pás signálu a konfidenční oblast koncentrace.

S ohledem na vlastní práci s kalibračním modelem je možné uvažovat: (a) *jedno použití*, kdy se z kalibračního modelu, vzniklého z n bodů $\{x_i, y_i\}$, $i = 1, ..., n$, počítá odhad cílové veličiny $\hat{x}^*$ se svým intervalem spolehlivosti pro jednu hodnotu y^*; (b) *vícenásobné použití*, kdy se pro různé hodnoty signálů určují odhady $\hat{x}^*$ na základě jednoho kalibračního modelu; (c) jedno nebo více použití *v kombinaci* s dalšími měřeními, kdy se výsledek druhé fáze kalibrace užívá spolu s dalšími údaji k určení veličiny, která je funkcí více proměnných. Zde se velikost vychýlení odhadů $\hat{x}^*$ projeví v systematické chybě výsledku.

3.3. Rozptyl predikce cílové veličiny *x**

Složitost řešení úlohy kalibrace souvisí zejména s užitým kalibračním modelem. Pro lineární regresní modely lze vyjádřit konfidenční pásy kolem modelu pomocí známých kvadratických rovnic. Složky vektoru ***x*** jsou funkcemi cílové veličiny, obvykle koncentrace. Obyčejně se uvažují polynomické modely, u kterých jednotlivé složky odpovídají mocninám měřené vlastnosti. Při hledání hodnoty cílové veličiny $\hat{x}^*$ se řeší úloha hledání kořene polynomu. Pro nelineární regresní modely se hledá řešení ve tvaru $\hat{x}^* = f^{-1}(y^*)$. Na základě Taylorova rozvoje této funkce lze nalézt přibližnou formuli pro rozptyl $D(\hat{x}^*)$ ve tvaru uvedeném v citaci [2], tj. $D(\hat{x}^*) \approx \left[\frac{\delta f(x,b)}{\delta x}\right]^{-2} \left[\frac{D(y^*)}{M} + D(f(x,b)\right]$, kde *D(y*)* je rozptyl *y**-ových hodnot, který je obyčejně roven σ^2, a $D(f(x, \boldsymbol{b})) = D(\hat{y})$ je rozptyl predikce, který se určuje také z Taylorova rozvoje funkce *f(x, **b**)*. Pro lineární regresní modely je rozptyl predikce roven

$$D(\hat{y}) = \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{(y^*-\bar{y})^2}{b_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

kde b_1 představuje odhad směrnice regresní přímky. Po dosazení dostaneme pro rozptyl odhadované koncentrace $D(\hat{x}^*) \approx \frac{\sigma^2}{b_1^2}\left[\frac{1}{M} + \frac{1}{n} + \frac{(y^*-\bar{y})^2}{b_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$.

Problémem je, že rozdělení veličiny $\hat{x}^*$ je obecně nesymetrické. Jedině pro případ kalibrační přímky a malý reziduální rozptyl lze rozdělení veličiny $\hat{x}^*$ považovat za přibližně symetrické a normální [3]. Za předpokladu, že jak *y*-nové, tak i *y**-ové hodnoty jsou náhodné proměnné s normálním rozdělením platí, že rozdíl $\Delta = \bar{y}^* - f(x^*, \boldsymbol{b})$ bude mít také normální rozdělení. Standardizovaná náhodná veličina $\Delta/\sqrt{D(\Delta)}$ má pak Studentovo rozdělení se stupni volnosti, které byly užity při určení $D(\Delta)$. Pro 100(1 - α)%ní konfidenční interval hodnoty odhadované koncentrace $\hat{x}^*$ platí, že

$$D(\Delta) = D(\bar{y}^*) + \sum_{j=1}^{m}\left[\frac{\delta f(\hat{x}^*,b)}{\delta b_j}\right]^2 D(b_j) + 2\sum_{i=1}^{m-1}\sum_{j>i}^{m}\frac{\delta f(\hat{x}^*,b)}{\delta b_i}\,\frac{\delta f(\hat{x}^*,b)}{\delta b_j}\,\mathrm{cov}(b_i,b_j),$$

kde *m* je počet regresních parametrů. Speciálně pro model kalibrační přímky $y = b_1(x - \bar{x}) + \bar{y}$ vyjde dosazením $D(\Delta) = \hat{\sigma}^2 \left[\frac{1}{M} + \frac{1}{n} + \frac{(\hat{x}^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$. Při znalosti $D(\Delta)$ lze již nalézt krajní meze konfidenčního intervalu pro $\hat{x}^*$. Jde o úlohu hledání kořenů kvadratické rovnice

$[\bar{y}^* - \bar{y} - b_1(\hat{x}^* - \bar{x})]^2 = F_{1-\alpha}(1, n-2)\ D(\Delta)$ vzhledem k proměnné $\hat{x}^*$.

3.4. Kalibrační přímka

Model kalibrační přímky patří v laboratořích k nejpoužívanějším. Předpokládá se obvykle, že tento model vyhovuje v celém sledovaném rozsahu proměnných *x* a *y*. Příkladem může být Lambertův-Beerův zákon, $A = \varepsilon\ d\ c$, vyjadřující lineární vztah mezi absorbancí *A* a koncentrací *c*, kde molární absorpční koeficient ε a délka kyvety *d* jsou konstanty. V některých případech však model kalibrační přímky platí pouze v omezeném intervalu a nad hraničním bodem $\{x_A, y_A\}$ dochází k výrazným odchylkám od linearity. Příkladem může být Kubelkův-Munkův vztah mezi funkcí remise $(1 - R^2)/2R$ a koncentrací *c*, který platí jen pro nízké koncentrace. Konečně i výše citovaný Lambertův-Beerův zákon platí pro některé roztoky jen do určité prahové koncentrace, od které pak dochází k výrazným nelineárním odchylkám. Z hlediska statistického zpracování lze užít kalibračního modelu $y_i = \beta_0 + \beta_1 x + \varepsilon_i,\ i = 1\ldots, n$, pro určení velikosti signálu neznámé koncentrace $y_j^* = b_0 + b_1\kappa,\ j = 1, \ldots, M$. Úlohou kalibrace je potom nalezení odhadu $\hat{x}^*$ parametru κ jako primárního a odhadů parametrů β_1, β_2 jako doplňkových. Vychází se opět z představy normality chyb ε_i a ε_j^*. Odhad $\hat{x}^*$ a jeho odpovídající konfidenční interval je možné určit několika způsoby:

(1) *Přímý odhad* parametru κ ve tvaru $\hat{x}^* = \bar{x} + y^* - \bar{y}\ / b_1$, kde *y** je měřená hodnota signálu (průměr $\bar{y}^*$ pro $M > 1$ opakovaných měření) a b_1 je odhad směrnice kalibrační přímky. Tento odhad je obecně vychýlený.

(2) Korekci na vychýlení lze provést pomocí *Naszodiho modifikovaného odhadu*

$$\hat{x}_B^* = \bar{x} + \frac{(y^* - \bar{y})\, b_1}{b_1^2 + \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

(3) Kručkov [6] navrhl *inverzní odhad* $\hat{x}_I^* = \bar{x} + (y^* - \bar{y}) \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, který vychází z inverzního regresního modelu $E(x/y) = \alpha_1 (y - \bar{y}) + \alpha_0$. Na základě rozboru odhadu $\hat{x}_I^*$ bylo zjištěno, že jde také o vychýlený odhad, který není lepší než přímý odhad $\hat{x}^*$. Navíc se při odhadování parametrů α_0 a α_1 chybně předpokládá, že *y*-ové hodnoty jsou měřeny se zanedbatelnými chybami vůči *x*-ovým hodnotám.

(4) V práci Schwartze [4] je navržen nelineární odhad

$$\hat{x}_N^* = \left(\sum_{i=1}^{n} x_i \exp\left[\frac{-\left(y^* - b_0 - b_1 x_i\right)^2}{2\,\hat{\sigma}^2} \right] \right) \Big/ \left(\sum_{i=1}^{n} \exp\left[\frac{-\left(y^* - b_0 - b_1 x_1\right)^2}{2\hat{\sigma}^2} \right] \right),$$

který je však založen na předpokladu normality reziduí.

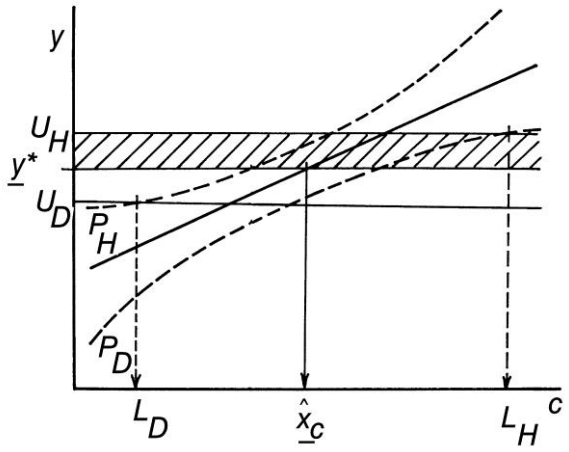
3.5. Nelineární model kalibrační křivky

Kromě nelineárních modelů, vycházejících z fyzikálně chemických vztahů mezi signálem a odezvou se v praxi používají také nelineární empirické modely. Značného zjednodušení lze dosáhnout využitím speciálních modelů, které vedou k odhadům, získaným lineární metodu nejmenších čtverců. Mezi takové dostatečně flexibilní modely patří například, polynomické spline funkce: u *modelu kalibrační křivky* se k aproximaci často volí lokálně definované funkce, které budou v místech vzájemného styku, tj. v uzlech, spojité ve funkčních hodnotách a hodnotách zadaných derivací, obr. 3-1. Vhodné interpolační funkce tohoto typu jsou složeny z polynomických úseků a platí pro ně, že jsou ze třídy $C^m[a, b]$. Obecně jsou funkce třídy $C^m[a, b]$ na intervalu $[a, b]$ spojité v prvních *m* derivacích a funkčních hodnotách. Hladké jsou všechny funkce od třídy C^1. Pro funkce třídy C^m platí, že *m*-tá derivace je lineární lomená závislost,

(*m*+1) derivace je po částech konstantní a (*m* + 2) derivace je po částech nulová, tj. není definovaná v uzlových bodech ξ_i.

3.6. Intervalové odhady cílové veličiny *x*

Při konstrukci intervalů spolehlivosti cílové veličiny *x* pro odhady $\hat{x}^*$ nebo $\hat{x}^*_B$ u silněji rozptýlených dat je nejjednodušší užít k určení $D(\hat{x}^*)$ předpokladu asymptotické normality. Meze 95%ního intervalu spolehlivosti se pak vypočtou dle *aproximativního vztahu,* a to pro dolní mez $L_D = \hat{x}^* - 1.96\sqrt{D(\hat{x}^*)}$ a pro horní mez $L_H = \hat{x}^* + 1.96\sqrt{D(\hat{x}^*)}$.



Obr. 3-2 Určení intervalu spolehlivosti [L_D, L_H] koncentrace *c* z kalibrační přímky. Šrafování značí poloviční šíři intervalu spolehlivosti signálu

Grafický způsob určení intervalu spolehlivosti pro cílovou veličinu *x* je na obr. 3-2. Z obrázku je patrné, že v případě opakování měření signálu *y* a určení střední hodnoty $\bar{y}^*$ je třeba stanovit konfidenční přímky U_D a U_H, a řešit úlohu hledání průsečíku U_H s dolní konfidenční parabolou P_D kalibrační přímky, kdy výsledkem je bod L_H, respektive průsečíku přímky U_D s horní konfidenční parabolou P_H kalibrační přímky, kdy výsledkem je bod L_D. Při znalosti rozptylu měření σ^2 lze snadno definovat 100(1 - α)%ní interval spolehlivosti pro signál *y** ve tvaru $U_{D,H} = \bar{y}^* \pm u_{1-\alpha/2}\sigma$, kde $u_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení. Pokud σ^2 není známo, lze využít nerovnosti $\sigma^2 < [(n-2)\hat{\sigma}^2]/[\chi^2_{\alpha/2}(n-2)M]$, kde $\chi^2_{\alpha/2}$ je dolní kvantil χ^2-rozdělení. Interval spolehlivosti signálu $U_{D,H}$ se potom vypočte ze vztahu $U_{D,H} = \bar{y}^* \pm u_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{M}}\sqrt{\frac{n-2}{\chi^2_{\alpha/2}(n-2)}}$. Místo kvantilu $u_{1-\alpha/2}$ se v této

rovnici pro M = 1 užívá kvantil Studentova rozdělení $t_{1-\alpha/2}(n-2)$ a rozptyl σ^2 je nahrazen jeho odhadem $\hat{\sigma}^2$. Pro celou regresní přímku budou hraniční 100(1 - α)%ní paraboloidy

$$P_{D,H} = b_1 x + b_0 \pm \hat{\sigma}\left\{2F_{1-\alpha}(2, n-2)\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\right\}^{1/2}$$

Hraniční hodnota L_H je potom řešením rovnice $U_H = P_D$, vzhledem k proměnné x. Hraniční hodnota L_D je řešením rovnice opět vzhledem k proměnné x, tj. $U_D = P_H$. Obě rovnice jsou vzhledem k proměnné x kvadratické. Z obr. 3-2 je také patrné, že v některých případech nemusí průsečík přímky s parabolou existovat nebo v jiném případě protne konfidenční přímka signálu parabolu kalibrační přímky ve dvou bodech. To ukazuje na příliš velký rozptyl v datech, kdy směrnice kalibrační přímky je málo významná. Taková kalibrační přímka pak není vhodná.

Kvalita intervalu spolehlivosti cílové veličiny je příznivě ovlivněna:
(1) Opakováním měření signálu y^*, čili růstem M.

(2) Zúžení konfidenčních parabol lze dosáhnout eliminací vlivných bodů, což bude ukázáno v příštím sdělení této publikační řady.

(3) Zmenšením reziduálního rozptylu $\hat{\sigma}^2$, a tedy buď zpřesněním měření, nebo užitím správného kalibračního modelu.

3.7. Přesnost kalibrace

K vyjádření přesnosti kalibrace se definují limitní hodnoty, které souvisejí s takovou úrovní koncentrace, pro kterou je signál ještě statisticky významně odlišný od šumu. V souvislosti s vyjádřením přesnosti a citlivosti kalibračních metod se definují tři specifické úrovně signálu:

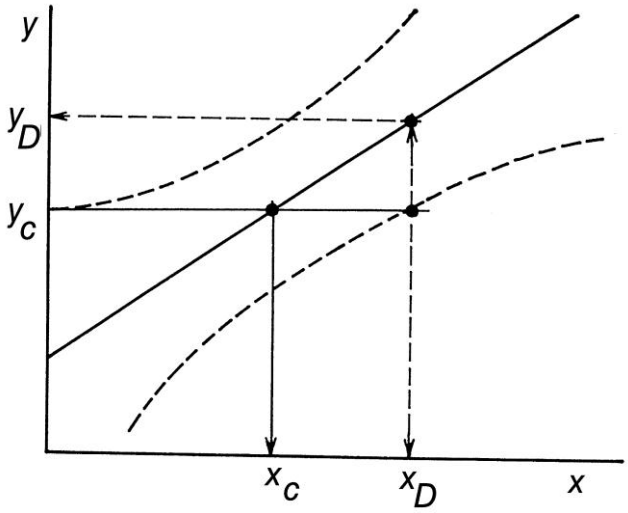
1. **Kritická úroveň** y_c představuje horní mez 100(1 - α)%ního intervalu spolehlivosti predikce signálu z kalibračního modelu pro koncentraci rovnou

nule, tzv. *slepý pokus*. Pro kritickou úroveň y_c platí vztah $y_c = \bar{y} - b_1\bar{x} + \hat{\sigma}\, t_{1-\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$.

Potom platí, že nad hodnotou y_c lze signál odlišit od šumu. Koncentrace x_c, odpovídající hodnotě kritické úrovně, se určí z kalibračního modelu pomocí vztahu $x_c = \frac{y_c - \bar{y}}{b_1} + \bar{x}$.

2. **Limita detekce** y_D odpovídá hodnotě koncentrace, pro kterou je dolní mez 100(1 - α)%ního intervalu spolehlivosti predikce signálu z kalibračního modelu rovna y_c, obr. 3-3. Pro lineární kalibrační model lze psát

$$y_D = y_c + \hat{\sigma}\, t_{1-\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_D-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$



Obr. 3-3 Definice kritické úrovně y_c, limity detekce y_D a jim odpovídající koncentrace x_c a x_D

Odpovídající koncentrace x_D se vypočte podle vztahu $x_D = \frac{y_D - \bar{y}}{b_1} + \bar{x}$. Limita detekce udává skutečnou úroveň signálu, která umožňuje ještě *detekci koncentrace*. Číselná velikost x_D udává minimální koncentraci, kterou lze ještě s pravděpodobností (1 - α) odlišit od nulové hodnoty.

3. **Limita stanovení** y_s je nejmenší hodnota signálu, pro kterou je relativní směrodatná odchylka predikce z kalibračního modelu dostatečně malá a rovna

číslu *C*. Pro číslo *C* se volí obyčejně hodnota *C* = 0.1. Označme predikci v místě x_s výrazem $y(x_s) = \bar{y} + b_1(x_s - \bar{x})$. Podmínka pro určení y_s je pak rovna $\sqrt{D(y(x_s))} / \hat{y}(x_s) = C$. Dosazením bude $y_s = \frac{\hat{\sigma}}{C}\sqrt{1+\frac{1}{n}+\frac{(x_s-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$. K praktickým výpočtům v laboratoři se však často užívá aproximace $y_s \approx \frac{\hat{\sigma}}{C}\sqrt{1+\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$. Odpovídající koncentrace x_s je pak $x_s = \frac{y_s - \bar{y}}{b_1} + \bar{x}$. Z uvedených charakteristik lze snadno konstruovat limitu detekce y_D a limitu stanovení y_s i pro nelineární kalibrační modely a pro případy dat, kdy rozptyly měření nejsou konstantní. Obecně platí porovnání limit, že $y_c \le y_D \le y_s$.

3.8. Závěry kalibrace

Bylo ukázáno, že kalibrační úloha se rozpadá do několika částí:

(1) U kalibračních úloh se pro kalibrační data nalezne kalibrační model, lineární či nelineární. Za kritérium vhodnosti navrženého modelu se bere těsnost proložení experimentálních bodů vypočtenou kalibrační křivkou.

(2) Pro kalibrační model se pro daný signál *y** neznámého vzorku vypočte bodový a intervalový odhad cílové hodnoty *x**.

(3) Před vlastním užitím kalibračního modelu (lineárního i nelineárního) je vhodné vyčíslit limitu detekce a limitu stanovení eventuelně i hodnotu slepého pokusu čili kritickou úroveň, které určují dolní hranici ještě použitelného kalibračního modelu.

(4) Často je vhodné aplikovat v rámci vyhodnocení kalibrace i regresní diagnostiku regresního tripletu (data, model, metoda), jež bude předmětem příštího sdělení a především eliminovat vlivné body, tj. odlehlé hodnoty v kalibrační závislosti a současně ověřit splnění všech předpokladů nejmenších čtverců.

3.9. Doporučená literatura

[1] Rosenblatt J.R. a Spiegelman C. H.: *Technometrics* **23**, 329 (1981).

[2] Ebel S. a Becht U.: *Fresenius Z. Anal. Chem.*, **158** (1987).

[3] Schwartz L. M.: *Anal. Chem.* **48**, 2287 (1976).

[4] Schwartz L. M.: *Anal. Chem.* **49**, 2062 (1977).

[5] Naszodi L. J.: *Technometrics* **20**, 201 (1978).

[6] Krutchkoff R. G.: *Technometrics* **9**, 425 (1967).

[7] Meloun M., Militký J.: *Statistické zpracování experimentálních dat,* Academia Praha 2004 (4. vydání).

[8] Meloun M., Militký J.: *Kompendium statistického zpracování experimentálních dat*, Academia Praha 2002 (1. vydání), 2006 (2. rozšířené vydání).

[9] ADSTAT, TriloByte Statistical Software s. r. o., Pardubice 1990.

4. VÝSTAVBA REGRESNÍHO MODELU REGRESNÍM TRIPLETEM

4.1. Úvod

Při výstavbě regresních modelů se užívá metody nejmenších čtverců MNČ. Tato metoda poskytuje postačující odhady parametrů jenom při současném splnění všech sedmi předpokladů o datech a regresním modelu. Pokud předpoklady nejsou splněny, ztrácí metoda nejmenších čtverců své vlastnosti.

4.1.1. *Základní předpoklady metody nejmenších čtverců (MNČ):*

Statistické vlastnosti odhadů $\hat{y}_P, \hat{e}$, ***b*** závisí na splnění jistých sedmi předpokladů. Pokud platí předpoklady I až IV, představují odhady ***b*** parametrů **β** **n**ejlepší, **n**estranné a **l**ineární **o**dhady (značeno metoda NNLO). Navíc mají asymptoticky normální rozdělení. Pokud platí ještě předpoklad VII, mají odhady ***b*** normální rozdělení i pro konečné výběry.

I. ***Regresní parametry*** β ***mohou nabývat libovolných hodnot.*** V praxi však často existují omezení parametrů, která vycházejí z jejich fyzikálního smyslu.

II. ***Regresní model je lineární v parametrech*** a platí aditivní model měření.

III. ***Matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných X má hodnost rovnou právě m***. To znamená, že žádné její dva sloupce $\boldsymbol{x_j}$, $\boldsymbol{x_k}$ nejsou *kolineární* čili rovnoběžné vektory. Tomu odpovídá i formulace, že matice $\boldsymbol{X}^T\boldsymbol{X}$ je symetrická regulární matice, ke které existuje inverzní matice a jejíž determinant je větší než nula.

IV. ***Náhodné chyby*** ε_i' ***mají nulovou střední hodnotu*** $E(\varepsilon_i') = 0$. To musí u korelačních modelů platit vždy. U regresních modelů se může stát, že $E(\varepsilon_i') = K$, $i = 1, ..., n$, což znamená, že model neobsahuje absolutní člen. Po jeho zavedení bude $E(\varepsilon_i') = 0$, kde $\varepsilon_i' = y_i - \hat{y}_{P,i} - K$.

V. ***Náhodné chyby*** $\boldsymbol{\varepsilon_i}$ ***mají konstantní a konečný rozptyl*** $E(\varepsilon_i^2) = \sigma^2$. Také podmíněný rozptyl $D(y/\boldsymbol{x}) = \sigma^2$ je konstantní a jde o *homoskedastický* případ.

VI. ***Náhodné chyby ε_i jsou vzájemně nekorelované a platí*** $cov(\varepsilon_i\, \varepsilon_j) = E(\varepsilon_i\, \varepsilon_j) = 0$. Pokud mají chyby normální rozdělení, jsou nezávislé. Tento požadavek odpovídá požadavku nezávislosti měřených veličin *y*.

VII. ***Chyby ε_i mají normální rozdělení*** $N(0, \sigma^2)$. Vektor ***y*** má pak vícerozměrné normální rozdělení se střední hodnotou ***Xβ*** a kovariančí maticí $\sigma^2\boldsymbol{E}$, kde ***E*** je jednotková matice.

4.1.2. *Regresní diagnostika*

Metoda nejmenších čtverců MNČ nezajišťuje ještě nalezení přijatelného modelu, a to jak ze statistického, tak i z fyzikálního hlediska. Musí být totiž splněny podmínky, odpovídající složkám tzv. *regresního tripletu* [data, model, metoda odhadu]. Regresní diagnostika obsahuje pomůcky a postupy k identifikaci

a) vhodnosti dat pro navržený regresní model (*kritika dat*),

b) vhodnosti modelu pro daná data (*kritika modelu*),

c) splnění základních předpokladů MNČ (*kritika metody*).

Základní rozdíl mezi regresní diagnostikou a klasickými testy spočívá v tom, že u regresní diagnostiky není třeba přesně formulovat alternativní hypotézu. Tímto pojetím se regresní diagnostika blíží spíše k *exploratorní regresní analýze*. Počítač slouží jako nástroj analýzy dat, modelu a metody odhadu. Model je navrhován v interakci uživatele s programem. Tím by měl být omezen vznik formálních regresních modelů, které nemají fyzikální smysl a jsou v technické praxi obyčejně jen omezeně použitelné.

4.2. Kritika dat

Mezi základní techniky diagnostiky patří stanovení rozmezí dat, jejich variability a přítomnosti vybočujících pozorování. K tomu lze využít grafů rozptýlení s kvantily a řady postupů průzkumové analýzy jednorozměrných dat. Přes svoji jednoduchost umožňuje diagnostika identifikovat ještě před vlastní regresní analýzou:

a) *nevhodnost dat* čili malé rozmezí nebo přítomnost vybočujících bodů,

b) *nesprávnost navrženého modelu* čili skryté proměnné,

c) *multikolinearitu*,

d) *nenormalitu* v případě, kdy jsou vysvětlující proměnné náhodné veličiny.

Kvalita dat úzce souvisí s užitým regresním modelem. Při posuzování se sleduje především výskyt vlivných bodů (VB), které mohou být hlavním zdrojem řady problémů, jako je zkreslení odhadů a růst rozptylů až k naprosté nepoužitelnosti regresních modelů. Podle toho, kde se vlivné body vyskytují, lze provést jejich dělení na

a) vybočující pozorování (outliers), které se liší v hodnotách vysvětlované (závisle) proměnné y od ostatních, a

b) extrémy (high leverage points),

které se liší v hodnotách vysvětlujících (nezávisle) proměnných x nebo v jejich kombinaci (v případě multikolinearity) od ostatních bodů. Vyskytují se však i body, které jsou jak vybočující, tak i extrémní. K identifikaci vlivných bodů typu vybočujícího pozorování se využívá zejména různých typů reziduí a k identifikaci extrémů pak diagonálních prvků H_{ii} projekční matice H, bližší lze nalézt v citaci [1]. Výskyt vlivných bodů (VB) je zdrojem řady problémů a způsobuje zkreslení odhadů a růst rozptylů odhadů parametrů. Vlivné body se dělí dle charakteru:

a) hrubé chyby, což jsou vybočující pozorování,

b) body s vysokým vlivem (tzv. golden points), které rozšiřují predikční schopnosti modelu.

c) zdánlivě vlivné body, jež jsou důsledkem nesprávného regresního modelu.

4.2.1. *Statistická analýza reziduí*

dělí rezidua na následujících několik druhů:

1. Klasická rezidua $\hat{e}_i = y_i - \boldsymbol{x}_i \boldsymbol{b}$. Existují nesprávné představy o reziduích, že

a) rozdělení reziduí je stejné jako rozdělení chyb,

b) vlastnosti reziduí jsou shodné s vlastnostmi chyb,

c) čím je reziduum $\hat{e}_i$ větší, tím je bod vlivnější, a tím spíše by se měl z dat vyloučit. Rozepsáním definičního vztahu

$$\hat{e}_i = (1 - H_{ii}) y_i - \sum_{j \neq i}^{n} H_{ij}\, y_i = (1 - H_{ii})\, \varepsilon_i - \sum_{j=i}^{n} H_{ij} \varepsilon_j$$

je zřejmé, že každé reziduum $\hat{e}_i$ je vlastně lineární kombinací všech chyb ε_i. Rozdělení reziduí je proto závislé na rozdělení chyb, na prvcích projekční matice ***H***, na velikosti výběru *n*. Klasická rezidua mají vlastnosti:

a) rozptyl reziduí je *nekonstantní*, i když rozptyl chyb $\hat{\sigma}^2$ je konstantní.

b) rezidua jsou korelovaná: existuje *párový korelační koeficient* r_{ij} mezi dvěma rezidui e_i a e_j formulovaný jako $r_{ij} = \frac{-H_{ij}}{\sqrt{(1 - H_{ii})\ (1 - H_{jj})}}$, i když chyby ε_i a ε_j jsou nezávislé.

c) rezidua *neindikují* extrémní hodnoty,

d) rezidua jsou normálnější než chyby (effekt supernormality),

e) u malých výběrů nemusí správně indikovat model.

2. Normovaná rezidua $\hat{e}_N$ definovaná vztahem $\hat{e}_{Ni} = \hat{e}_i / \hat{\sigma}$ jsou normálně rozdělené veličiny $\hat{e}_{Ni} - N(0, 1)$. Platí u nich diagnostické pravidlo 3σ, které říká, že rezidua větší než $\forall 3\hat{\sigma}$ indikují vybočující body.

3. Standardizovaná rezidua $\hat{e}_S$ definovaná vztahem $\hat{e}_{Si} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - H_{ii}}}$ mají konstantní rozptyl a maximální hodnota $\hat{e}_{Si}$ je ohraničena velikostí $\sqrt{n - m}$. Platí u nich diagnostické pravidlo, že jsou vhodná především k indikaci heteroskedasticity.

4. Jackknife rezidua $\hat{e}_J$ čili **"plně studentizovaná",** jsou definovaná vztahem, ve kterém se užije místo $\hat{\sigma}$ odhadu směrodatné odchylky $\hat{\sigma}_{(-i)}$, a dostane se $\hat{e}_{Ji} = \hat{e}_{Si}\sqrt{\frac{n - m - 1}{n - m - \hat{e}_{Si}^2}} = \sqrt{n - m}\ \mathrm{cotg}\ \Theta_i$. Rezidua mají Studentovo rozdělení s (*n* - *m* -

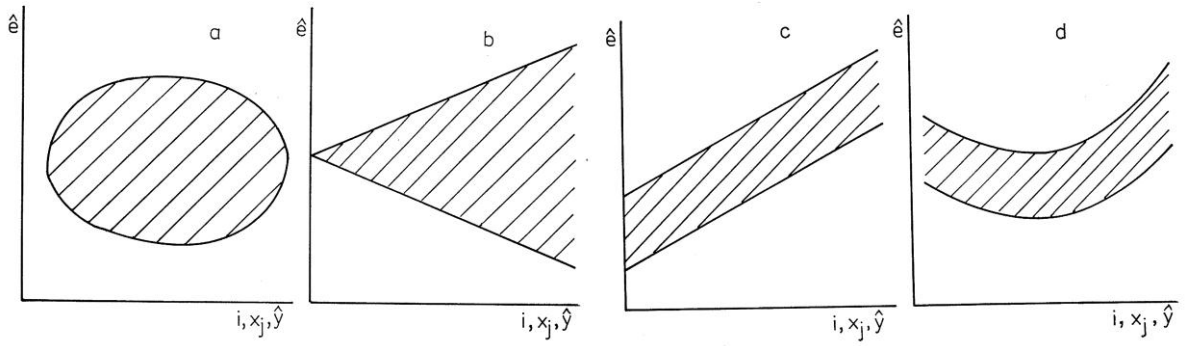
1) stupni volnosti. Platí u nich diagnostické pravidlo, že jsou vhodná především k identifikaci vybočujících bodů (outliers).

5. Predikovaná rezidua $\hat{e}_P$ jsou definovaná vztahem $\hat{e}_{Pi} = y_i - x_i \boldsymbol{b}_{(i)} = \frac{\hat{e}_i}{1 - H_{ii}}$, kde $\boldsymbol{b}_{(i)}$ jsou MNČ odhady ze všech bodů kromě *i*-tého. Platí u nich diagnostické pravidlo, že indikují vybočující hodnoty (outliers).

6. Rekurzívní rezidua $\hat{e}_R$ jsou definována vztahy $\hat{e}_{Ri} = 0, i = 1, \ldots, m,$ $\hat{e}_{Ri} = \frac{y_i - \boldsymbol{x}_i \boldsymbol{b}_{i-1}}{\sqrt{1 + \boldsymbol{x}_i (\boldsymbol{X}_{i-1}^T \boldsymbol{X}_{i-1})^{-1} x_i^T}}$, $i = m + 1, \ldots, n,$ kde $\boldsymbol{b}_{i-1}$ jsou odhady získané z prvních (i - 1) bodů. Platí pro ně diagnostické pravidlo, že indikují autokorelaci a nestabilitu regresního modelu v čase.

4.2.2. *Obrazce v diagnostických grafech:*

Pokud se v diagnostických grafech reziduí objeví tvar `mrakuA bodů (obr. 4-1), je detekována správnost metody nejmenších čtverců. Jiné obrazce bodů v grafu reziduí indikují vesměs nesprávnost v datech či nesprávnost modelu: tvar výseče odhaluje nekonstantnost rozptylu čili heteroskedasticitu v datech, tvar pásu indikuje chybu ve výpočtu nebo nepřítomnost x_j v modelu, tvar pásu může být i důsledkem vybočujících hodnot nebo indikuje chybný výpočet, kdy v regresním modelu chybí absolutní člen. Nelineární tvar ukazuje na nesprávně navržený model.

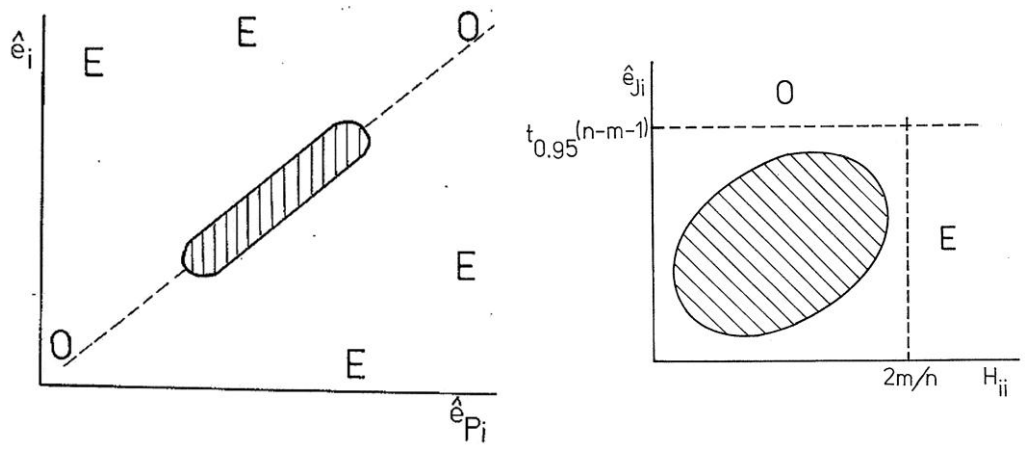


Obr. 4-1 Nejčastější tvary obrazce bodů v grafické analýze reziduí: a) tvar mraku, b) tvar výseče, c) tvar pásu a d) nelineární tvar.

4.2.3. *Grafy identifikace vlivných bodů:*

Existuje pět grafických diagnostik vlivných bodů, které současně testují charakter vlivných bodů, zatímco ostatní diagnostické grafy spíše odhalují podezřelé body.

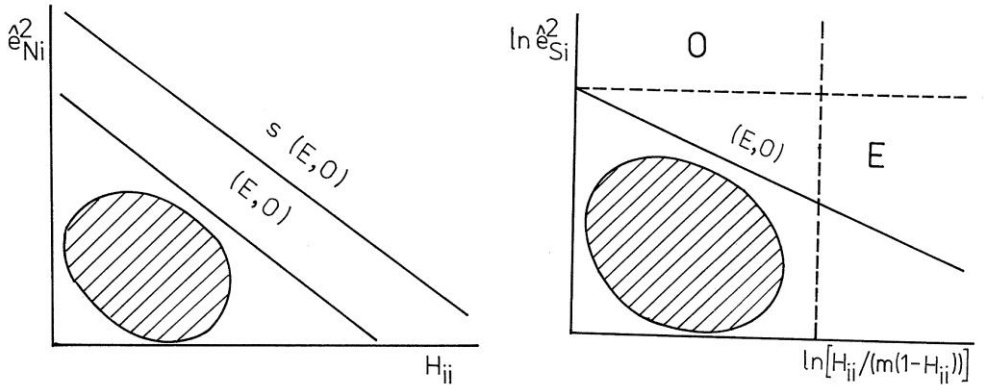
1. Graf predikovaných reziduí GPR, (osa *x*: $\hat{e}_{Pi}$, osa *y:* $\hat{e}_i$). V grafu jsou extrémy snadno identifikovány svou polohou, neboť leží mimo přímku $y = x$. Vybočující body leží sice na přímce $y = x$ nebo v její blízkosti, jsou však dostatečně vzdáleny od mraku, shluku ostatních bodů (obr. 4-2a).



Obr. 4-2 Grafy vlivných bodů: a) Graf predikovaných reziduí (GPR), b) Williamsův graf: E značí extrém a O značí vybočující bod.

2. Williamsův graf WG, (osa *x*: prvky H_{ii}, osa *y*: $\hat{e}_{Ji}$). Do tohoto grafu se zakreslují jednak mezní linie pro vybočující body, $y = t_{0.95}(n - m - 1)$, a jednak mezní linie pro extrémy, $x = 2\,m / n$. Zde $t_{0.95}(n - m - 1)$ je 95% kvantil Studentova rozdělení s $n - m - 1$ stupni volnosti (obr. 4-2b).

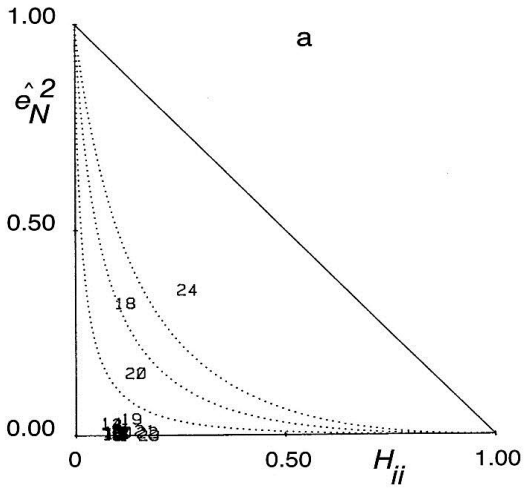
3. Pregibonův graf PG, (osa *x*: prvky H_{ii}, osa *y*: $\hat{e}^2_{Ni}$). Protože platí, že $E(H_{ii} + \hat{e}^2_{Ni}) = (m + 1)/n$, lze do tohoto grafu zakreslit dvě hraniční přímky $y = -x + 2\,(m + 1)/n$ a $y = -x + 3\,(m + 1)/n$. K rozlišení mezi body platí pravidlo: bod je silně vlivný, leží-li nad horní přímkou. Bod je pouze vlivný, leží-li mezi oběma přímkami. Může jít ale jak o extrém, tak o vybočující bod (obr. 4-3a).



Obr. 4-3 Grafy vlivných bodů: a) Pregibonův graf (PG), kde (E, O) značí vlivné body, s(E,O) značí silně vlivné body, b) McCullohův-Meeterův graf (MMG), kde E značí extrém a O vybočující bod, (E, O) obojí vlivné body.

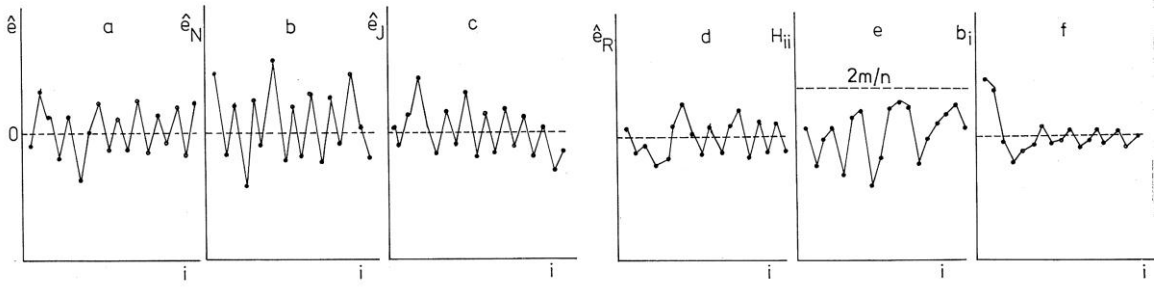
4. McCullohův-Meeterův graf MMG, (osa *x*: ln [H_{ii} /(*m* (1 - H_{ii}))], osa *y*: ln $\hat{e}^2_{Si}$). Linie v tomto grafu představují přímky stejného vlivu přímky se směrnicí -1. Například pro 90% konfidenční čáru je *y* = - *x* - ln [$F_{0.9}$(*n* - *m*, *m*)]. Omezující čára pro extrémy je $x = \ln \frac{2}{n-2m}$ a omezující čára pro vybočující body je *y* = ln [(*n* - *m*) $t^2_{0.95}$ ($t^2_{0.95}$ + *n* - *m*)], kde $t_{0.95}$ je 95% kvantil Studentova rozdělení s *n* - *m* - 1 stupni volnosti a $F_{0.9}$ (*n* - *m* - 1) je 90% kvantil *F*-rozdělení s *n* - *m* a *m* stupni volnosti (obr. 4-3b).

5. L-R grafy, (osu *y*: $\hat{e}^2_{Ni}$ = $\hat{e}^2_i$/*RSC*, na osu *x*: prvky H_{ii}), obr. 4-4. Body nad jednotlivými lemniskátami jsou vlivné, a to ve směru horního rohu trojúhelníka jde o vybočující body a ve směru pravého rohu trojúhelníka jde o extrémy.



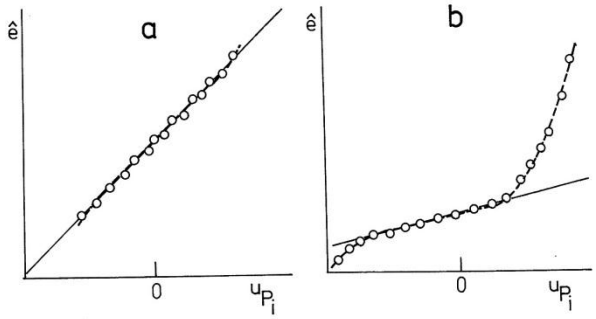
Obr. 4-4 Grafy vlivných bodů: L-R graf.

6. Indexové grafy IG, (osa *x*: index *i*, osa *y*: rezidua $\hat{e}_i$, $\hat{e}_{Si}$, $\hat{e}_{Ni}$, $\hat{e}_{Pi}$, $\hat{e}_{Ji}$, $\hat{e}_{Ri}$, nebo prvky H_{ii} či H^*_{ii} a nebo odhady $\boldsymbol{b}_i$). Mohou zde být i rekurzivní odhady regresních parametrů $\boldsymbol{b}_i$ (obr. 4-5). Tyto grafy indikují pouze podezřelé body, nejsou schopny nikterak testovat vybočující body jako je tomu v pěti předešlých grafech.



Obr. 4-5 Rozličné typy indexových grafů znázorňujících indexovou závislost: a) $\hat{e}_i$, b) $\hat{e}_{Ni}$, c) $\hat{e}_{Ji}$, d) $\hat{e}_{Ri}$, e) H_{ii}, f) b_i na *i*.

7. Rankitové grafy Q-Q, (osa *x*: u_{P_i} pro $P_i = i / (n + 1)$, osa *y*: $\hat{e}_{(i)}$, $\hat{e}_{S(i)}$, $\hat{e}_{N(i)}$, $\hat{e}_{P(i)}$, $\hat{e}_{J(i)}$, $\hat{e}_{R(i)}$). Na osu *x* se vynášejí kvantily normovaného normálního rozdělení u_P pro $P_i = i / (n + 1)$ a na osu *y* pořádkové statistiky čili vzestupně setříděné hodnoty reziduí, (obr. 4-6). Cílem je ověřit normalitu rozdělení reziduí.



Obr. 4-6 Rankitový graf *Q-Q* reziduí indikuje: a) přibližně normální rozdělení, b) rozdělení s dlouhými konci.

4.3. Kritika modelu

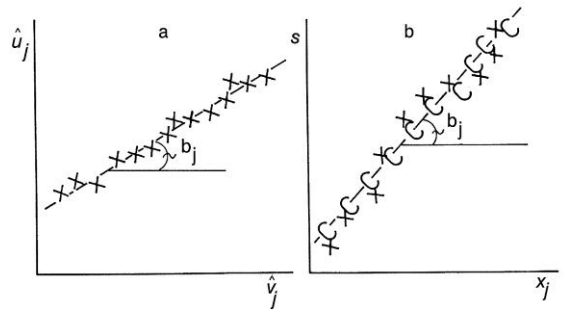
Kvalitu regresního modelu lze posoudit v případě jedné vysvětlující proměnné *x* přímo z rozptylového grafu závislosti *y* na *x*. V případě více vysvětlujících proměnných a multikolinearity mohou rozptylové grafy *mylně indikovat* nelineární trend i u lineárního modelu. Z řady různých grafů k posouzení vztahu

y a x_j se zde omezíme na a) parciální regresní grafy, a b) parciální reziduální grafy.

Belseyho parciální regresní grafy umožňují posouzení kvality navrženého regresního modelu, indikují přítomnost vlivných bodů, nesplnění předpokladů klasické MNČ a konečně vyjadřují závislost mezi ***y*** a zvolenou proměnnou $\boldsymbol{x}_j$ při statisticky neměnném vlivu ostatních $\boldsymbol{X}_{(j)}$. Přitom matice $\boldsymbol{X}_{(j)}$ vznikne z matice ***X*** vynecháním j-tého sloupce $\boldsymbol{x}_j$, odpovídajícího j-té vysvětlující proměnné. Grafy mají vlastnosti:

a) Směrnice přímky v parciálním regresním grafu je stejná jako odhad b_j v neděleném modelu a úsek je roven nule. Lineární závislost platí pouze v případě, že navržený model je správný.

b) Korelační koeficient mezi oběma proměnnými parciálního regresního grafu odpovídá parciálnímu korelačnímu koeficientu $\hat{R}_{yx_j(x)}$.Vychází se z regresního modelu $y = \boldsymbol{X}_{(j)}\beta^* + \boldsymbol{x}_j c + \boldsymbol{\varepsilon}$, kde $\boldsymbol{\beta}^*$ má rozměr (m - 1) H 1, c je regresní parametr příslušející j-té proměnné.



Obr. 4-7 Kritika modelu pomocí a) parciálního regresního grafu, b) parciálního reziduálního grafu.

Parciální reziduální grafy jsou zvané také grafy "komponenta + reziduum". Rovnici $\hat{\boldsymbol{e}} = \hat{\boldsymbol{u}}_j - \hat{\boldsymbol{v}}_j.b_j$ lze přepsat do tvaru $\hat{\boldsymbol{u}}_j = \hat{\boldsymbol{e}} + b_j(E - H_{(j)})x_j$ jako závislost $\hat{\boldsymbol{e}} + b_j(E - H_{(j)})x_j$ na $(E - H_{(j)})x_j$. Jedná se o závislost parciálních reziduí ***s*** na proměnné $\boldsymbol{x}_j$, kde $s = \hat{\boldsymbol{e}} + bx_j = y - \sum_{k \neq j}^{m} x_k b_k$. V grafu se znázorňuje deterministická komponenta $c_{ij} = (x_{ij} - \bar{x}_j)b_j$,i = 1,…, m , která se v grafu značí písmenem "C" a parciální reziduum $s_i = c_{ij} + \hat{e}_i$, i = 1, ..., n, které se zde označuje křížkem "+". Uvedená diagnostika má vlastnosti:

a) Směrnice závislosti s na x_j je rovna b_j a úsek je nulový. Lineární závislost ukazuje na vhodnost navržené x_j v modelu.

b) Rezidua přímky jsou přímo rezidua $\hat{e}_i$ pro nedělený model.

c) Pokud je úhel mezi x_j a některými sloupci matice $X_{(j)}$ malý (multikolinearita), ukazuje parciální reziduální graf nesprávně malý rozptyl kolem regresní přímky $b_j x_j$ a dochází i k potlačení efektu vlivných bodů.

Znaménkový test vhodnosti modelu. Nenáhodnost reziduí čili trend v reziduích lze testovat znaménkovým testem postupem: 1. Určuje se počet sekvencí n_U, kde mají rezidua stejná znaménka, (např. pro rezidua -1, -1, 1, -1, 1, 2, 1 je počet sekvencí roven $\hat{n}_U$ = 4). 2. Stanoví se počet reziduí kladných (n_+) a záporných (n_-). 3. Počet sekvencí n_t a jeho rozptyl D_t se vyčíslí dle $n_t = 1 + \frac{2n_+n_-}{n_+ + n_-} \approx 1 + \frac{n}{2}$, $D_t = \frac{2n_+n_-(2n_+n_- - n_+ - n_-)}{(n_+ + n_-)^2(n_+ + n_- - 1)} \approx \frac{n}{4}$.

Vlastní testování spočívá ve vyšetření podmínky: je-li $n_U < n_t - C1.96\sqrt{D_t}$, existuje v reziduích trend a model je nesprávně navržen. Konstanta C = 0.5 je korekcí na spojitost.

4.4. Kritika metody

V praxi bývají některé předpoklady MNČ porušeny, což vede k použití modifikací metody MNČ. K porušení předpokladů dochází v základních případech:

a) Na parametry jsou kladena omezení, což vede na užití metody podmínkových nejmenších čtverců (MPNČ).

b) Kovarianční matice chyb není diagonální (autokorelace), případně data mají nestejný rozptyl (heteroskedasticita), což vede na užití *metody zobecněných nejmenších čtverců* (MZNČ), respektive *metody vážených nejmenších čtverců* (MVNČ).

c) Rozdělení dat nelze považovat za normální nebo se v datech vyskytují vlivné body. V takovém případě se místo kritéria metody nejmenších čtverců užije *robustního* kritéria, které je na porušení předpokladu o rozdělení chyb a na

vlivné body málo citlivé. Pro odhad parametrů *b* se užívá *iterační metody vážených nejmenších čtverců* (IVNČ).

d) Také proměnné *x* mohou být zatížené náhodnými chybami, což vede na užití *metody rozšířených nejmenších čtverců* (MRNČ). Pro případ stejných rozptylů vede metoda k odhadům minimalizujícím kolmé vzdálenosti (*orthogonální regrese*).

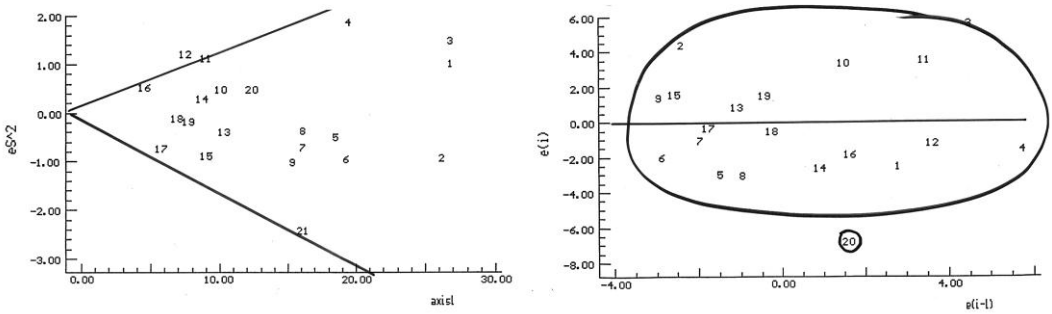
e) Pro špatně podmíněné matice X^TX se používá *metoda racionálních hodností* (RH), vedoucí k systému vychýlených odhadů, kde vychýlení je řízeno jedním parametrem.

1. Hetero/homoskedasticita čili nekonstantnost rozptylu lze popsat následovně: rozptyl veličiny y_i v *i*-tém bodě je popsán vztahem $\sigma_i^2 = \sigma^2 \exp(\lambda \boldsymbol{x}_i \boldsymbol{\beta})$, kde $\boldsymbol{x}_i$ je *i*-tý řádek matice $\boldsymbol{X}$. Předmětem tohoto testu homoskedasticity je ověření nulové hypotézy H_0: $\lambda = 0$ za použití Cookova-Weisbergova testačního kritéria $S_f = \dfrac{\left[\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_p)\,\hat{e}_i^2\right]^2}{2\,\hat{\sigma}^4\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_P)^2}$, kde $\hat{y}_P = \dfrac{1}{n}\sum_{i=1}^{n}\hat{y}_i$. Výsledkem testování je závěr: a) Je-li $S_f < \chi^2(1)$, H_0 (homoskedasticita) je přijata. b) Pro homoskedasticitu tvoří diagnostický graf $\hat{e}_{Si}^2$ v závislosti na $(1 - H_{ii})\,\hat{y}_i$ náhodný elipsovitý mrak bodů. Pro heteroskedasticitu vznikne v tomto grafu typický klínový obrazec, (obr. 4-8a).

2. Autokorelace: Data časových řad mají náhodné chyby ε_i vzájemně korelované. Nejčastější je případ autokorelace prvního řádu $\varepsilon_i = \rho_1\varepsilon_{i-1} + u_i$, kde $u_i - N(0, \sigma^2)$.

a) Pro $\rho_1 = 1$ případ kumulativních chyb, který se v laboratorních datech vyskytuje často.

b) Pro $\rho_1 \leq 1$ jde o autokorelační koeficient 1. řádu. Waldův test pak testuje nulovou hypotézu H_0: $\rho_1 = 0$ *vs.* H_A: $\rho_1 \neq 0$. Je-li Waldovo testační kritérium $W_a = \dfrac{n\,\hat{\rho}_1^2}{1 - \hat{\rho}_1^2} < \chi^2(1)$, je H_0 o homoskedasticitě přijata a v datech není prokázána autokorelace, (obr. 4-8b).



Obr. 4-8 a) Klínový obrazec indikuje heteroskedasticitu v datech, b) elipsa indikuje data bez autokorelace.

3. Normalita chyb: normalitu náhodných chyb lze testovat dvojím způsobem, a to rankitovým grafem a numericky:

a) *Rankitový (Q-Q) graf,* ($\hat{e}_{(i)}$ nebo $\hat{e}_{Ri}$ na u_{Pi} pro $P_i = i / (n + 1)$),

b) *Test normality* testuje nulovou hypotézu H_0: normalita *vs.* H_A: nenormalita. Slouží k tomu Jarque-Berrova testační statistika, která je formulována vztahem

$$L(\hat{e}) = n\left[\frac{\hat{u}_3^2}{6\ \hat{u}_2^3} + \frac{\frac{\hat{u}_4}{\hat{u}_2} - 3}{24}\right] + n\left[\frac{3\hat{u}_1^2}{2\hat{u}_2} - \frac{\hat{u}_3\hat{u}_1}{\hat{u}_2}\right]$$

, kde $\hat{u}_j$ je *j*-tý moment reziduí $\hat{u}_j = \frac{\sum_{i=1}^{n} \hat{e}_i^j}{n}$.

Test normality spočívá ve vyšetření nerovnosti: je-li $L(\hat{e}) > \chi^2_{1-\alpha}(2) = 5.99$, je H_0 (normalita) zamítnuta. Je třeba podotknout, že test není vhodný pro malé výběry.

4.5. Postup výstavby lineárního regresního modelu [1, 4]

1. Návrh modelu: Začíná se vždy od nejjednoduššího modelu, u kterého vystupují jednotlivé vysvětlující proměnné v prvních mocninách a nevyskytují se žádné interakční členy typu $x_j x_k$.

2. Předběžná analýza dat: Sleduje se proměnlivost jednotlivých proměnných a možné párové vztahy. Užívá se proto rozptylových diagramů závislosti x_j na x_k nebo indexových grafů závislosti x_j na *j*. Posuzuje se významnost proměnných s ohledem na jejich proměnlivost a přítomnost multikolinearity. Přibližně lineární vztah mezi proměnnými v rozptylových grafech závislosti x_j na x_k indikuje multikolinearitu. Uživatel může volit polynomickou transformaci zadáním

stupně polynomu. Provádí se sestavení korelační matice ***R*** a její rozklad na vlastní čísla a vlastní vektory.

3. Odhadování parametrů: Odhadování parametrů modelu se provádí metodou nejmenších čtverců MNČ nebo metodou racionálních hodností RH. Ze zobecněné inverzní matice $\boldsymbol{R}^{-1}$ jsou určovány odhady parametrů ***b***, jejich směrodatné odchylky $\sqrt{D(b_j)}$ a velikosti testačních statistik Studentova *t*-testu významnosti pro β_j = 0. Jsou provedeny testy významnosti odhadů b_j, vícenásobného korelačního koeficientu *R* a koeficientu determinace *D*. Je vhodné sledovat rozhodčí kritéria hypotézy regresního modelu jako je střední kvadratická chyba predikce *MEP* a Akaikovo informační kritérium *AIC*.

4. Regresní diagnostika: S využitím pěti rozličných grafů je prováděna identifikace vlivných bodů, a to *grafem predikovaných reziduí, grafy Wiliamsovým, Pregibonovým, McCulloh-Meeterovým,* a *L-R grafem*. Pak následuje ověření předpokladů metody nejmenších čtverců jako jsou homoskedasticita, nepřítomnost autokorelace a normalita rozdělení chyb. V případě více vysvětlujících proměnných se posoudí vhodnost jednotlivých proměnných a jejich funkcí s využitím parciálních regresních grafů nebo grafů "komponenta + reziduum". Je uveden odhad autokorelačního koeficientu reziduí prvního řádu $\hat{\rho}_1$. *Tabulka vlivných bodů* obsahuje rozličná rezidua, u kterých jsou hvězdičkou označeny hodnoty silně vlivných bodů, které by měly být z dat odstraněny.

5. **Konstrukce zpřesněného modelu**: S využitím

a) *metody vážených nejmenších čtverců* (MVNČ) při nekonstantnosti rozptylů,

b) *metody zobecněných nejmenších čtverců* (MZNČ) při autokorelaci,

c) *metody podmínkových nejmenších čtverců* (MPNČ) při omezeních na parametry,

d) *metody racionálních hodností* (RH) u multikolinearity,

e) *metody rozšířených nejmenších čtverců* (MRNČ) pro případ, že všechny proměnné jsou zatížené náhodnými chybami,

f) *robustní metody* pro jiná rozdělení dat než normální a data s vybočujícími hodnotami a extrémy jsou odhadovány parametry zpřesněného modelu.

4.6. Doporučená literatura:

[1] M. Meloun, J. Militký: *Statistické zpracování experimentálních dat,* Plus Praha 1994, (1. vydání), East Publishing Praha 1998 (2. vydání), Academia Praha 2004 (3. vydání).

[2] M. Meloun, J. Militký: *Statistické zpracování experimentálních dat - Sbírka úloh*, Univerzita Pardubice 1996.

[3] ADSTAT 1.25, 2.0 a verze 3.0, TriloByte Statistical Software Pardubice, 1992, 1993, 1999.

[4] M. Meloun, J. Militký: *Kompendium statistického zpracování experimentálních dat,* Academia Praha 2002 (1. vydání), Academia Praha 2006 (2. rozšířené vydání).

PŘÍLOHY

Centrum pro rozvoj výzkumu pokročilých řídicích a senzorických technologií
CZ.1.07/2.3.00/09.0031

Ústav automatizace a měřicí techniky
VUT v Brně
Kolejní 2906/4
612 00 Brno
Česká Republika

http://www.crr.vutbr.cz

info@crr.vutbtr.cz