MOCNINNÁ A BOX-COXOVA TRANSFORMACE V PRŮZKUMOVÉ EXPLORATORNÍ ANALÝZE DAT

POWER AND BOX-COX TRANSFORMATION IN THE EXPLORATORY DATA ANALYSIS

Milan Meloun a Jiří Militký
Univerzita Pardubice, 532 10 Pardubice a
Technická univerzita Liberec, 461 17 Liberec

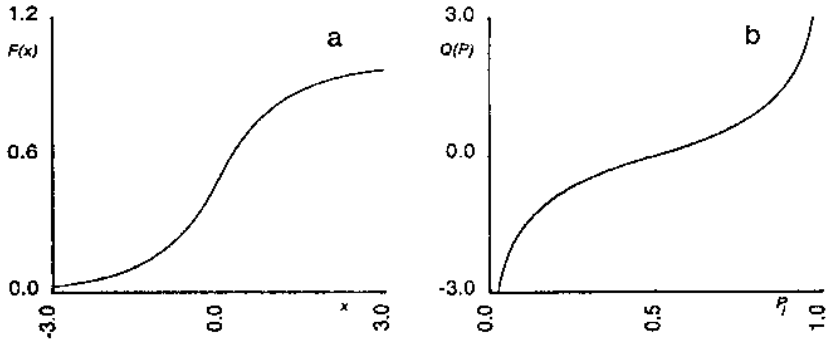
KEY WORDS: *exploratory data analysis, power transformation, Box-Cox transformation, Hines-Hines selection graph, plot of logarithm of likelihood function, re-transformed mean*

SUMMARY:
Exploratory Data Analysis *provides the first contact with the data and serves to uncover unexpected departures from familiar (Gaussian) models. When the data does not fulfil all assumption about the sample i. e. the sample distribution differs from the Gaussian, normal one the user is faced with the problem how to analyze the data.* ***Power*** *or* ***Box-Cox transformation*** *involves finding a scale that can clarify the analysis of the data or simplify the behavior of the data. It help to promote symmetry, constancy of variability, linearity, or additivity of effect, depending on the structure of the data. The proper transformation leads to symmetric distribution of data, stabilizes the variance, or makes the distribution closer to normal.*

ÚVOD

Účelem průzkumové (exploratorní) analýzy dat EDA výsledků každé analytické metody je třeba odhalit zvláštnosti a ověřit předpoklady o výběru pro následné statistické zpracování. Tak lze zabránit provádění numerických výpočtů bez hlubších statistických souvislostí.

Před vlastní analýzou je proto nezbytné vyšetřit platnost základních předpokladů, tj. nezávislost, homogenitu a normalitu výběru. Reprezentativní náhodný výběr je popsán následujícími vlastnostmi: prvky výběru x_i jsou vzájemně nezávislé, výběr je homogenní, jde o normální rozdělení pravděpodobnosti, prvky souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru. Vychází se z ***pořádkových statistik***, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, a ***pořadové pravděpodobnosti*** $P_i = \dfrac{i}{n+1}$, pro kterou platí, že $100P_i$ procentní výběrový kvantil je hodnota, pod kterou leží $100P_i$ procent prvků výběru. 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.



Obr. 1 (a) Distribuční funkce $F(x_i)$ a (b) kvantilová funkce $Q(P_i)$ Laplaceova rozdělení s nulovou střední hodnotou a rozptylem rovným 2

METODICKÁ ČÁST

1. Postup analýzy dat

Data ve spektrofotometrii se často vyznačují nekonstantním rozptylem, malou četností, asymetrickým rozdělením a obecně porušením základních předpokladů, kladených na výběr. Uveďme proto nejprve obecnou osnovu analýzy výběru dat.

1. Průzkumová analýza dat:

Diagnostické grafy: *stupeň symetrie rozdělení*
lokální koncentrace dat
vybočující data

2. Ověření předpokladů výběru dat:

Diagnosticky, testy: *ověření normality*
ověření nezávislosti
ověření homogenity
určení minimální četnosti

3. Transformace dat:

Analýza dat: *originální data*
data po mocninné transformaci
data po Box-Coxově transformaci

4. Parametry polohy, rozptýlení a tvaru:

Analýza 1 výběru:
klasické odhady - *průměr*
- *rozptyl*

robustní odhady - *medián*
- *uřezané průměry*
- *winsorizovaný rozptyl*
- *interkvantilové rozpětí*

adaptivní odhady

5. Testování dvou výběrů:

a) Testy polohy
b) Testy rozptýlení

2. Transformace dat

Pokud se na základě analýzy dat zjistí, že rozdělení výběru dat se příliš odlišuje od rozdělení normálního, vzniká problém, jak data vůbec vyhodnotit. V řadě případů lze nalézt ***vhodnou transformaci***, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě. Vychází se z představy, že zpracovávaná data jsou nelineární transformací normálně rozdělené náhodné veličiny x. Hledá se k nim pak inverzní transformace g(x).

1. **Stabilizace rozptylu** vyžaduje nalezení transformace y = g(x), ve které je již rozptyl $\sigma^2(y)$ konstantní. Pokud je rozptyl původní proměnné x funkcí typu $\sigma^2(x) = f_1(x)$, lze rozptyl $\sigma^2(y)$ určit

$$\sigma^2(y) \approx \left[\frac{dg(x)}{dx}\right]^2 f_1(x) = C$$

kde C je konstanta. Hledaná transformace g(x) je pak řešením diferenciální rovnice

$$g(x) \approx C \int \frac{dx}{\sqrt{f_1(x)}}$$

U řady instrumentálních metod a přístrojů je zajištěna konstantnost relativní chyby $\delta(x)$. To znamená, že rozptyl $\sigma^2(x)$ je dán funkcí $f_1(x) = \delta^2(x)\, x^2 = \text{konst}\, x^2$. Po dosazení vyjde $g(x) = \ln x$. Optimální je pro tento případ logaritmická transformace původních dat. Z toho vyplývá také vhodnost použití geometrického průměru. Pokud je závislost $\sigma^2(x) = f_1(x)$ mocninná, bude optimální transformace g(x) také mocninná. Jelikož pro normální rozdělení je střední hodnota na rozptylu nezávislá, bude transformace stabilizující rozptyl také zajišťovat přiblížení k normalitě.

2. **Zesymetričtění rozdělení** výběru je možné provést jednoduchou mocninnou transformací

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

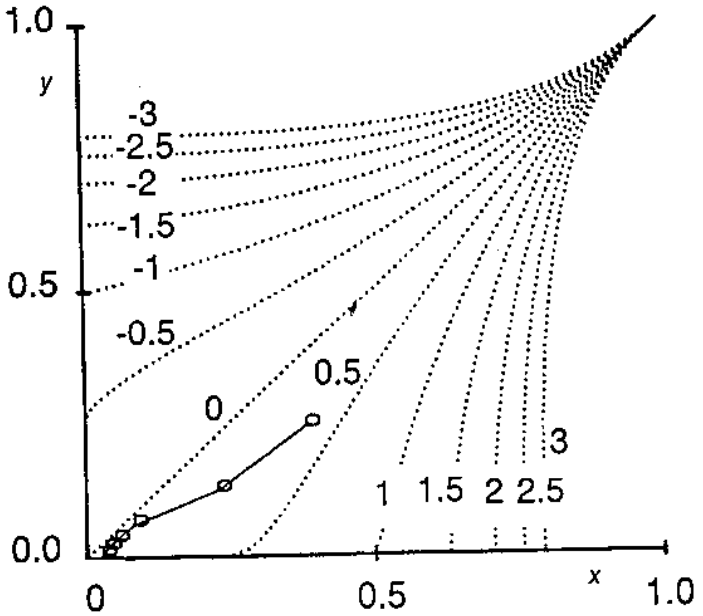
Tato transformace však nezachovává měřítko, není vzhledem k hodnotě λ všude spojitá a hodí se pouze pro kladná data. Optimální odhad $\hat{\lambda}$ se hledá s ohledem na minimalizaci vhodných charakteristik asymetrie. Kromě šikmosti $\hat{g}_1(y)$ je možné užít i robustní verzi šikmosti definovanou výrazem

$$\hat{g}_{1R}(y) = \frac{(\tilde{y}_{0.75} - \tilde{y}_{0.50}) - (\tilde{y}_{0.50} - \tilde{y}_{0.25})}{\tilde{y}_{0.75} - \tilde{y}_{0.25}}$$

Pro symetrická rozdělení je $\hat{g}_p(y) = 0$, stejně jako $\hat{g}_1(y)$ a $\hat{g}_{1R}(y)$. Hodnotu $\hat{\lambda}$ lze hledat pomocí rankitového grafu. Pro optimální $\hat{\lambda}$ budou ležet kvantily $y_{(i)}$ přibližně na přímce.

3. Hines - Hinesův selekční graf (osa x: $\tilde{x}_{0.5}/x_{1-P_i}$, osa y: $\tilde{x}_{P_i}/\tilde{x}_{0.5}$)

Diagnostickou pomůckou pro odhad optimálního parametru λ je selekční graf.



Obr. 2 Selekční graf pro výběr z lognormálního rozdělení

Vychází z požadavků symetrie jednotlivých kvantilů kolem mediánu

$$\left(\frac{\tilde{x}_{P_i}}{\tilde{x}_{0.5}}\right)^{\lambda} + \left(\frac{\tilde{x}_{0.5}}{\tilde{x}_{1-P_i}}\right)^{-\lambda} = 2$$

kde pro pořadové pravděpodobnosti jsou obvykle voleny písmenové hodnoty, $P_i = 2^{-i}$, $i = 2, 3$. K porovnání průběhu experimentálního bodů s ideálním (teoretickým) pro zvolené λ se do grafu zakreslují i řešení rovnice $y^{\lambda} + x^{-\lambda} = 2$ pro $0 \le x \le 1$ a $0 \le y \le 1$:

a) pro $\lambda = 0$ je řešením přímka $y = x$,

b) pro $\lambda < 0$ je řešením vztah $y = (2 - x^{-\lambda})^{1/\lambda}$,

c) pro $\lambda > 0$ je řešením vztah $x = (2 - y^{\lambda})^{-1/\lambda}$.

Podle umístění experimentálních bodů na teoretických křivkách selekčního grafu lze odhadovat velikost λ a posuzovat kvalitu transformace v různých vzdálenostech od mediánu.

Pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti se užívá ***Boxovy-Coxovy transformace***

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

Boxova-Coxova transformace má tyto vlastnosti:

1. Transformace g(x) jsou vzhledem k veličině λ spojité, protože platí

$$\lim_{\lambda \to 0} \frac{x^{\lambda} - 1}{\lambda} = \ln x$$

2. Všechny transformace procházejí bodem [y = 0; x = 1] a mají v tomto bodě společnou směrnici.

3. Mocninné transformace s exponenty -2; -3/2; -1; -1/2; 0; 1/2; 1; 3/2; 2 jsou co do křivosti rovnoměrně rozmístěné.

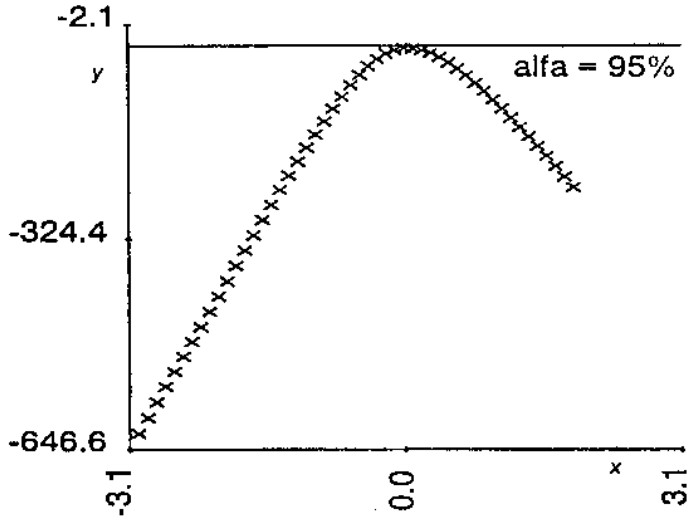
Boxova-Coxova transformace je použitelná pouze pro kladná data. Rozšíření této transformace na oblast, kdy rozdělení dat začíná od prahové hodnoty x_0, spočívá v náhradě x rozdílem $(x - x_0)$, který je vždy kladný.

4. 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 $\ln L = f(\lambda)$ lze znázornit ve zvoleném intervalu např. $-3 \leq \lambda \leq 3$ a identifikovat i maximum $\hat{\lambda}$.



Obr. 3 Graf logaritmu věrohodnostní funkce pro výběr z lognormálního rozdělení

Pro asymptotický $100(1 - \alpha)\%$ní interval spolehlivosti parametru λ platí

$$2 \left[\ln L(\hat{\lambda}) - \ln L(\lambda)\right] \leq \chi^2_{1-\alpha}(1)$$

kde $\chi^2_{1-\alpha}$ (1) je kvantil χ^2-rozdělení s jedním stupněm volnosti. V tomto intervalu spolehlivosti leží všechna λ, pro která platí nerovnost

$$\ln L(\lambda) \geq \ln L(\hat{\lambda}) - 0.5\chi^2_{1-\alpha}(1)$$

Čím bude konfidenční interval širší, tím je mocninná Boxova-Coxova transformace méně výhodná. Pokud obsahuje tento interval i hodnotu $\lambda = 1$, není transformace ze statistického hlediska přínosem.

5. Zpětná transformace

Pokud se podaří nalézt vhodnou transformaci, která vede k přibližné normalitě, lze určit $\bar{y}$, $s^2(y)$, interval spolehlivosti $\bar{y} \pm t_{1-\alpha/2}(n-1)\, s(y)/\sqrt{n}$ a provádět i statistické testování. Problém však spočívá v tom, že všechny statistické charakteristiky je třeba určit pro původní proměnné.

1. ***Nekorektní přístup*** spočívá v prosté zpětné transformaci $\bar{x}_R = g^{-1}(\bar{y})$. Pro jednoduchou mocninnou transformaci vede zpětná transformace na obecný průměr definovaný vztahem

$$\bar{x}_R = \bar{x}_\lambda = \left[\frac{\sum_{i=1}^{n} x_i^\lambda}{n}\right]^{1/\lambda}$$

Pro $\lambda = 0$ se místo x^λ používá $\ln x$ a místo $x^{1/\lambda}$ pak e^x. Hodnota $\bar{x}_R = \bar{x}_{-1}$ představuje ***harmonický průměr***, $\bar{x}_R = \bar{x}_0$ ***geometrický průměr***, $\bar{x}_R = \bar{x}_1$ ***aritmetický průměr*** a $\bar{x}_R = \bar{x}_2$ ***kvadratický průměr***. Tento způsob zpětné transformace vede často ke zkreslujícím výsledkům.

2. ***Korektní (přesnější) přístup*** zpětné transformace vychází z Taylorova rozvoje funkce $y = g(x)$ v okolí $\bar{y}$. Pro retransformovaný průměr $\bar{x}_R$ lze pak odvodit přibližný vztah

$$\bar{x}_R \approx g^{-1}\left[\bar{y} - \frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)\right]$$

Pro rozptyl vyjde

$$s^2(x_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$$

Zde jednotlivé derivace jsou vyčísleny v bodě $x = \bar{x}_R$. Pro $100(1 - \alpha)\%$ní interval spolehlivosti střední hodnoty původního souboru dat x platí

$$\bar{x}_R - I_D \leq \mu \leq \bar{x}_R + I_H$$

kde

$$I_D = g^{-1}\left(\bar{y} + G - t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right)$$

$$I_H = g^{-1}\left(\bar{y} + G + t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right)$$

$$G = -0.5 \frac{d^2 g(x)}{dx^2} \left(\frac{dg(x)}{dx} \right)^{-2} s^2(y)$$

Symbolem $t_{1-\alpha/2}(n - 1)$ je označen $100(1-\alpha/2)\%$ní kvantil Studentova rozdělení s (n-1) stupni volnosti. Při znalosti hodnot konkrétní transformace $y = g(x)$ a odhadů $\bar{y}$, $s^2(y)$ je snadné vyčíslit hodnoty $\bar{x}_R$ a $s^2(x_R)$:

a) Pro speciální případ $\lambda = 0$, tzn. logaritmickou transformaci typu $g(x) = \ln x$, bude

$$\bar{x}_R \approx \exp[\bar{y} + 0.5\, s^2(y)]$$

Rozptyl se určí $s^2(x_R) \approx \bar{x}_R^2\, s^2(y)$

b) Pro případ $\lambda \neq 0$ a Boxovy-Coxovy transformace bude $\bar{x}_R$ jedním z kořenů kvadratické rovnice, pro které platí

$$\bar{x}_{R,1,2} = [0.5\,(1 + \lambda\,\bar{y}) \pm$$

$$\pm\, 0.5 \sqrt{1 + 2\,\lambda\,(\bar{y} + s^2(y)) + \lambda^2\,(\bar{y}^2 - 2\,s^2(y))}\,]^{1/\lambda}$$

Jako odhad $\bar{x}_R$ se pak bere kořen $\bar{x}_{R,i}$, který je nejbližší mediánu $\tilde{x}_{0.5} = g^{-1}(\tilde{y}_{0.5})$. Při znalosti retransformovaného průměru lze z vyčíslit i odpovídající rozptyl

$$s^2(x) = \bar{x}_R^{-2\,\lambda+2}\, s^2(y)$$

ZÁVĚR

Symetrizující mocninné transformace a normalizující Boxovy-Coxovy transformace dat slouží k určení objektivních parametrů polohy. Vlastní výpočet má následující postup:

1. Pro mocninnou transformaci se počítají různé míry symetrie a výběrová špičatost v rozmezí $-3 \leq \lambda \leq 3$ s krokem 0.1. Jsou tištěny optimální hodnoty těchto měr. Je možno kreslit Hinesů-Hinesův selekční graf k určení optimální hodnoty λ. Na základě těchto informací se zadává zvolená hodnota λ. V této transformaci se pak vyčíslí $\bar{y}$, $s^2(y)$, šikmost a špičatost.
2. Pro transformaci se počítá $\ln L(\lambda)$, různé míry symetrie a výběrová špičatost v rozmezí $-3 \leq \lambda \leq 3$ s krokem 0.1. Jsou tištěny optimální hodnoty těchto měr. Je kreslen graf závislosti $\ln L$ na λ spolu s 95%ním konfidenčním intervalem. Na základě těchto informací se zadává zvolená hodnota λ. V této transformaci se počítá $\bar{y}$, $s^2(y)$, šikmost a špičatost. Jsou určeny i retransformované hodnoty $\bar{x}_R$ a 95%ní konfidenční interval pro retransformovanou střední hodnotu μ.

Literatura:

1. M. Meloun, J. Militký: *Statistické zpracování experimentálních dat,* Plus Praha 1994.
2. ADSTAT, TriloByte Statistical Software s. r. o., Pardubice 1990.