M5VM05 Statistické modelování

6. Ověřování předpokladů v klasickém modelu lineární regrese – I

Jan Koláček (kolacek@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

Motivace

Možnost použití statistických testů je podmíněna nějakými předpoklady o datech. Velmi často je to předpoklad o typu rozložení, z něhož získaná data pocházejí. Mnoho testů je založeno na předpokladu normality. Opomíjení předpokladů o typu rozložení může v praxi vést i ke zcela zavádějícím výsledkům, proto je nutné věnovat tomuto problému patřičnou pozornost.

Testování normality

Graficky

1. Histogram
   - třídicí intervaly $(u_1, u_2\rangle, \ldots, (u_r, u_{r+1}\rangle$
   - doporučuje se volit r blízké $\sqrt{n}$.

   Četnostní hustota j-tého třídicího intervalu je definována vztahem

$$f_j = \frac{p_j}{d_j}$$

kde $d_j = u_{j+1} - u_j$. Soustava obdélníků sestrojených nad třídicími intervaly, jejichž plochy jsou rovny relativním četnostem, se nazývá histogram.

Testování normality

2. Quantile - quantile plot (Q-Q plot)

   Q-Q plot konstruujeme tak, že na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ a na vodorovnou osu kvantily $K_{\alpha_j}(X)$ vybraného rozložení, kde

   $$\alpha_j = \frac{j - r_{adj}}{n + n_{adj}},$$

   přičemž r_{adj} a n_{adj} jsou korigující faktory $\leq 0,5$. Implicitně se klade $r_{adj} = 0,375$ a $n_{adj} = 0,25$. Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadují z dat, nebo se volí na základě teoretického modelu. Body $(K_{\alpha_j}(X), x_{(j)})$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím lepší je soulad mezi empirickým a teoretickým rozložením.
   Jsou-li některé hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.

Testování normality

3. Graf výběrové distribuční funkce
   Položme
   $$z_{(i)} = \frac{x_{(i)} - \bar{x}}{s}, \quad i = 1, \ldots, n, \; s = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$
   x-ová osa: hodnoty $z_{(i)}$
   y-ová osa: hodnoty distribuční funkce $N(0,1)$ $\phi(z_{(i)})$ porovnat s hodnotami výběrové distribuční funkce $F_n(z_{(i)}) = \frac{i}{n}$, $i = 1, \ldots n$.

Testování normality

Výpočtem

1. Kolmogorovův – Smirnovův test
   Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - \Phi(x)|.$$

   Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota. Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem

$$\sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}.$$

Testování normality

2. Shapirův – Wilkův test normality
   Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení $N(\mu, \sigma^2)$. Test je založen na zjištění, zda body v Q-Q plotu jsou významně odlišné od regresní přímky proložené těmito body. Shapirův – Wilkův test se používá především pro výběry menších rozsahů, $n < 50$.

Testování normality

3. Testy dobré shody
   H_0 : „náhodný výběr $X_1, \ldots, X_n$ pochází z rozdělení s distr. funkcí $\Phi(x)$"
   - Je-li distribuční funkce spojitá, pak data rozdělíme do r třídicích intervalů $(u_j, u_{j+1}\rangle, j = 1, \ldots, r$. Zjistíme absolutní četnost n_j j-tého třídicího intervalu a vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.
   - Má-li distribuční funkce nejvýše spočetně mnoho bodů nespojitosti, pak místo třídicích intervalů použijeme varianty $x_{[j]}, j = 1, \ldots, r$. Pro variantu $x_{[j]}$ zjistíme absolutní četnost n_j a vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat variantou $x_{[j]}$. Platí-li nulová hypotéza, pak

$$p_j = \Phi\left(x_{[j]}\right) - \lim_{x \to x_{[j]}^-} \Phi(x) = P\left(X = x_{[j]}\right). \tag{1}$$

Testová statistika:

$$K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j} \approx \chi^2(r - 1 - p). \tag{2}$$

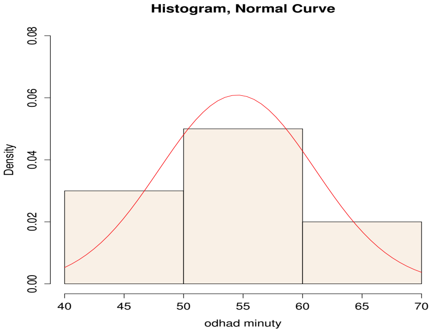
Aproximace se považuje za vyhovující, když $np_j \geq 5, j = 1, \ldots, r$.

Příklad

Příklad 1

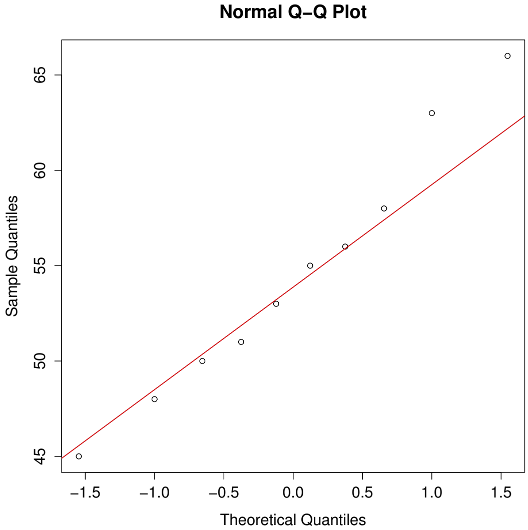
Deset pokusných osob mělo nezávisle na sobě bez předchozího nácviku odhadnout, kdy od daného signálu uplyne jedna minuta. Výsledky pokusu jsou uloženy v souboru „`minuta.RData`". Testujte graficky i výpočtem, zda se jedná o výběr z normálního rozdělení.

Řešení Histogram a teoretická hustota



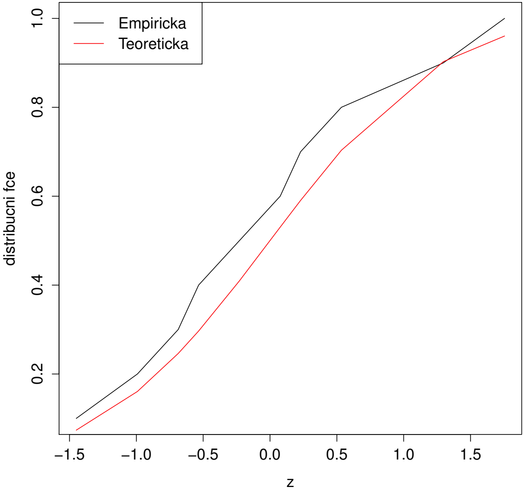
Řešení

Q–Q plot



Řešení

Výběrová distribuční funkce



Řešení

Výpočtem

- Kolmogorovův – Smirnovův test

$$p - value = 0,9985$$

- Shapirův–Wilkův test

$$p - value = 0,9164$$

- Test dobré shody

$$p - value = 0,9189$$

Autokorelace

V některých případech (často v časových řadách) hodnoty náhodné chyby ε_i závisí na předchozích hodnotách ε_{i-k}, $k = 1, 2, \ldots$, což má za následek, že efekt náhodných chyb není okamžitý, ale je pociťován i v budoucnosti. Tento případ se nazývá **autokorelace**.

Nejjednodušší typ: **autoregrese 1. řádu** – ozn. $AR(1)$

$$\varepsilon_i = \theta \varepsilon_{i-1} + u_i,$$

kde θ je neznámý parametr, $|\theta| < 1$, $Eu_i = 0$, $cov(u_i, u_j) = \begin{cases} \sigma_u^2 & i = j, \\ 0 & \text{jinak.} \end{cases}$

$AR(1)$

$\varepsilon_i = \theta\varepsilon_{i-1} + u_i = \theta(\theta\varepsilon_{i-2} + u_{i-1}) + u_i = \theta^2\varepsilon_{i-2} + \theta u_{i-1} + u_i = \cdots = \sum\limits_{j=0}^{\infty} \theta^j u_{i-j}$

$E\varepsilon_i = E\sum\limits_{j=0}^{\infty} \theta^j u_{i-j} = \sum\limits_{j=0}^{\infty} \theta^j E u_{i-j} = 0$

$D\varepsilon_i = D\sum\limits_{j=0}^{\infty} \theta^j u_{i-j} = \sum\limits_{j=0}^{\infty} \theta^{2j} D u_{i-j} = \sigma_u^2 \sum\limits_{j=0}^{\infty} \theta^{2j} = \frac{\sigma_u^2}{1-\theta^2}$

$cov(\varepsilon_i, \varepsilon_{i-j}) = \sum\limits_{r=0}^{\infty}\sum\limits_{s=0}^{\infty} \theta^r\theta^s cov(u_{i-r}, u_{i-j-s}) = \theta^j\sigma_u^2 \sum\limits_{r=0}^{\infty} \theta^{2r} = \frac{\theta^j\sigma_u^2}{1-\theta^2}$ pro $j > 0$

Tedy

$$D\boldsymbol{\varepsilon} = \frac{\sigma_u^2}{1-\theta^2}\begin{pmatrix} 1 & \theta & \theta^2 & \dots & \theta^{n-1} \\ \theta & 1 & \theta & \dots & \theta^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \theta \\ \theta^{n-1} & \dots & \theta^2 & \theta & 1 \end{pmatrix} = \underbrace{\frac{\sigma_u^2}{1-\theta^2}}_{\sigma_\epsilon^2}\mathbf{W}.$$

Máme tedy model tvaru:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ E\boldsymbol{\varepsilon} = \mathbf{0},\ D\boldsymbol{\varepsilon} = \sigma_\epsilon^2\mathbf{W},\quad \text{píšeme}\quad \mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma_\varepsilon^2\mathbf{W})$$

Rozšířený lineární model

Věta 1 (Aitkenův odhad)

Mějme regresní model $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W})$ *plné hodnosti, kde* $\mathbf{W} > 0$. *Pak odhad pomocí metody nejmenších čtverců je roven*

$$\widehat{\boldsymbol{\beta}}_W = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y}.$$

Z věty tedy plyne, že pokud známe parametr θ, můžeme v uvedeném modelu najít odhady $\widehat{\boldsymbol{\beta}}$.

Detekce autokorelace

1. Graficky
   Označme $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}$. Do grafu postupně vykreslíme hodnoty $\hat{\varepsilon}_i$ v závislosti na $\hat{\varepsilon}_{i-1}$, $i = 2, \ldots, n$. Bude-li z grafu zřejmá přibližná lineární závislost, svědčí to o autokorelaci 1. řádu nebo o špatné volbě modelu.
2. Test hypotézy $H_0 : \theta = 0$ proti $H_1 : \theta \neq 0$
   (a) Asymptotický test: Pro dostatečně velká n ($n \geq 30$) platí

$$U_{\hat{\theta}} = \frac{\hat{\theta} - \theta}{\sqrt{\frac{1-\theta^2}{n}}} \stackrel{A}{\sim} N(0,1).$$

Za platnosti hypotézy má tedy statistika

$$\sqrt{n}\hat{\theta} \stackrel{A}{\sim} N(0,1).$$

Pak nulovou hypotézu zamítáme, pokud $|\sqrt{n}\hat{\theta}| > u_{1-\frac{\alpha}{2}}$.

Detekce autokorelace

(b) Durbin – Watsonův test: je založen na statistice

$$D = \frac{\sum_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

Pokud budou residua málo korelovaná, hodnota D se bude pohybovat kolem 2. Kladná hodnota způsobí, že $D \in (0, 2)$ a záporná korelace způsobí, že $D \in (2, 4)$. Přesné hodnoty kritických oborů pro test nalezneme v tabulkách.

Odhad parametru θ

Odhad parametru θ

1. Odhadujeme jako regresní koeficient v modelu

$$\hat{\varepsilon}_i = \theta \hat{\varepsilon}_{i-1} + u_i, \; i = 2, \dots, n$$

metodou nejmenších čtverců. Odtud pak

$$\hat{\theta} = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum_{i=2}^{n} \hat{\varepsilon}_{i-1}^2}.$$

2. Pomocí Durbin – Watsonovy statistiky:

$$\hat{\theta} = 1 - \frac{D}{2}.$$

Odstranění autokorelace 1. řádu

Postup:

1. Nalezneme odhad $\hat{\theta}$
2. Vytvoříme nový model

$$Y_i^* = Y_{i+1} - \hat{\theta} Y_i; \; X_{ij}^* = X_{i+1,j} - \hat{\theta} X_{ij}, \; i = 1, \dots, n-1, \; j = 1, \dots, k,$$

tj. vznikne model

$$\mathbf{Y}^* = \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*, \; E\boldsymbol{\varepsilon}^* = \mathbf{0}, \; D\boldsymbol{\varepsilon}^* = \sigma_{\varepsilon^*}^2 \mathbf{I}_n$$

a hledáme odhady $\hat{\boldsymbol{\beta}}^*$ standardním způsobem.

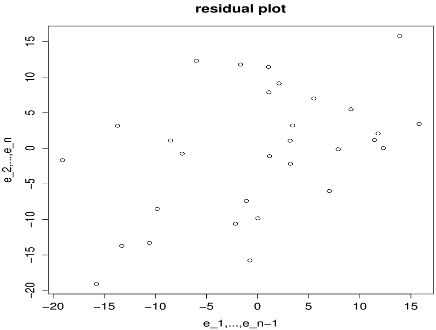
Příklad

Příklad 2

V letech 1953 – 1983 byly měřeny ztráty vody při distribuci do domácností. Výsledky měření jsou uloženy v souboru „`voda.RData`". Proměnná x označuje množství vyrobené vody, proměnná Y ztrátu. Ověřte, zda se v datech vyskytuje autokorelace 1. řádu a případně ji odstraňte.

Řešení

Řešení Graficky



Z grafu je patrná lineární závislost.

Řešení

(a) Asymptotický test:

$$U_{\hat{\theta}} = |\sqrt{n}\hat{\theta}| = 2,339.$$

Nulovou hypotézu tedy **zamítáme**, neboť $|\sqrt{n}\hat{\theta}| > u_{1-\frac{\alpha}{2}} = 1,96.$

(b) Durbin – Watsonův test:

$$D = \frac{\sum\limits_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum\limits_{i=1}^{n} \hat{\varepsilon}_i^2} = 1,082$$

a p-hodnota testu je $0,0016$, takže také **zamítáme** nulovou hypotézu.

Řešení

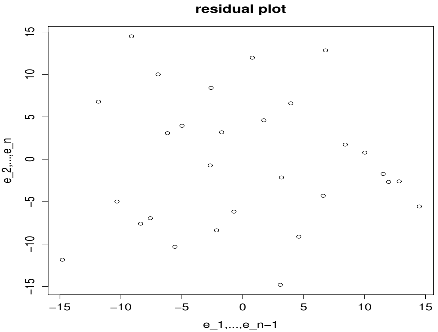
Odstranění autokorelace:

Odhady $\hat{\theta}$ jsou velmi podobné.

Metodou nejmenších čtverců: $\hat{\theta} = 0,42$

Z D–W statistiky: $\hat{\theta} = 0,459$

V nově vzniklém modelu vykreslíme residua:



Také D–W test již **nezamítá** nulovou hypotézu (p-hodnota $= 0,4$).

Úlohy k procvičení

Příklad 1

V souboru „`studenti.RData`“ jsou uloženy údaje o 96 studentech VŠE v Praze. Hodnoty v prvním sloupci značí hmotnost studentů v kg (proměnná Y), ve druhém sloupci je výška studentů v cm (proměnná X_1) a ve třetím sloupci je indikátor pohlaví studenta (proměnná X_2, 0 – žena, 1 – muž). Předpokládejte regresní model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Odhadněte parametry modelu a ověřte normalitu residuí. Dále pak testujte přítomnost autokorelace 1. řádu, případně ji odstraňte.

[Odhady parametrů: $\hat{\beta}_0 = -53,67$, $\hat{\beta}_1 = 0,6648$, $\hat{\beta}_2 = 6,3323$, normalita se nezamítá, autokorelace 1. řádu se zamítá.]

Úlohy k procvičení

Příklad 2

V proměnné „`LakeHuron`“[a] jsou uloženy roční údaje o hloubce jezera Huron (ve stopách) v letech 1875 – 1972. Nalezněte vhodný regresní model a ověřte, zda se v datech vyskytuje autokorelace 1. řádu. Případně se ji pokuste odstranit. Zkoumejte také normalitu residuí.

[a] datový soubor implementovaný v jazyce R

[Vhodný model: polynom 7. stupně, autokorelace 1. řádu se nezamítá, normalita residuí u nového modelu se nezamítá.]