

M0130 – 6. PRAKTIKUM : M0130pr06 (*Transformace časových řad*)**A. Transformace stabilizující rozptyl**

Nechť náhodná veličina  $X$  má rozdělení, které závisí na nějakém parametru  $\theta$ . Předpokládejme, že tento parametr je zvolen tak, aby platilo

$$E_{\theta}X = \theta.$$

Ve většině případů (ne však u normálního rozdělení) na  $\theta$  závisí i rozptyl veličiny  $X$ , takže můžeme psát

$$D_{\theta}X = \sigma^2(\theta).$$

Přitom  $\sigma(\theta)$  bývá obvykle hladká funkce proměnné  $\theta$ .

Vzniká otázka, zda lze najít netriviální funkci  $g$  tak, aby náhodná veličina  $Y = g(X)$  měla **rozptyl nezávislejší na  $\theta$** . (Požadavkem netriviality se vylučují konstantní funkce  $g$ , které by vedly k veličinám s nulovým rozptylem).

Uvedená úloha v obecném případě nemá řešení. Používá se však určitých aproximací, které se ukázaly velmi užitečné.

Pokud se zabýváme jen dostatečně **hladkými funkcemi**  $g$ , z Taylorova rozvoje dostaneme aproximaci

$$g(X) \approx g(\theta) + g'(\theta)(X - \theta).$$

Potom střední hodnotu lze aproximovat takto

$$E_{\theta}g(X) \approx E[g(\theta) + g'(\theta)(X - \theta)] = g(\theta)$$

a rozptyl

$$D_{\theta}[g(X)] \approx [g'(X)]^2 D_{\theta}X = [g'(\theta)]^2 \sigma^2(\theta).$$

Chceme, aby po transformaci byl **rozptyl konstantní** a nezávisel na střední hodnotě, tj.

$$c^2 = D_{\theta}[g(Y_t)] = [g'(\theta)]^2 \sigma^2(\theta) \quad \Rightarrow \quad g'(\theta) = \frac{c}{\sigma(\theta)},$$

kde  $c$  je nějaká konstanta. Odtud snadno dostaneme tvar transformace stabilizující rozptyl

$$g(\theta) = c \int \frac{1}{\sigma(\theta)} d\theta + K.$$

Konstanty  $c$  a  $K$  se volí tak, aby funkce  $g$  vypočtená podle předchozího vzorce měla výhodný tvar.

Ukázalo se, že takto vypočtená funkce  $g$

- nejen **výrazně stabilizuje rozptyl**, takže rozptyl  $D_{\theta}g(X)$  závisí na  $\theta$  jen velmi málo,
- ale zároveň také **rozdělení náhodné veličiny  $Y = g(X)$  bývá již velmi blízké normálnímu**, i když třeba samotné rozdělení veličiny  $X$  je výrazně nenormální.

## B. Příklady transformací stabilizujících rozptyl

### POISSONOVO ROZDĚLENÍ

Nechť náhodná veličina má **Poissonovo** rozdělení s parametrem  $\lambda > 0$ , tj.

$X \sim Po(\lambda)$  s pravděpodobnostní funkcí  $p_X(x) = P(X = x) = \frac{\lambda^x}{x!}$  pro  $x = 0, 1, 2, \dots$

Lze spočítat, že  $EX = DX = \lambda$ , tj.  $\sigma^2(\lambda) = \lambda$ . Pak pro  $\lambda \rightarrow \infty$  platí

$$\frac{X - \lambda}{\sqrt{\lambda}} \xrightarrow{L} U \sim N(0, 1).$$

Chceme najít takovou funkci  $g$ , aby asymptotické rozdělení náhodné veličiny

$$g(X) - g(\lambda) \xrightarrow{L} Y \sim N(0, c) \quad c > 0.$$

Pak

$$g(\lambda) = c \int \frac{1}{\sigma(\lambda)} d\lambda + K = c \int \frac{1}{\sqrt{\lambda}} d\lambda + K = 2c\sqrt{\lambda} + K.$$

Obvykle se volí  $c = \frac{1}{2}$ ,  $K = 0$  a pracuje se s velmi známou **odmocninovou transformací**

$$Y = g(X) = \sqrt{X}.$$

Spočítejme střední hodnotu a rozptyl náhodné veličiny  $Y$ :

$$\begin{aligned} EY &= Eg(X) \approx g(\lambda) = \sqrt{\lambda} \\ DY &= Dg(X) \approx [g'(\lambda)]^2 \sigma^2(\lambda) = \left[ \frac{1}{2\sqrt{\lambda}} \right]^2 \lambda = \frac{1}{4}. \end{aligned}$$

Poznamenejme, že Anscombe (1948) navrhl **stabilnější transformaci** (ve smyslu, že rozptyl transformované náhodné veličiny je méně závislý na střední hodnotě) ve tvaru

$$Y = \sqrt{X + \frac{3}{8}},$$

přičemž

$$\begin{aligned} EY &= E\sqrt{X + \frac{3}{8}} = \sqrt{\lambda + \frac{3}{8}} - \frac{1}{8\sqrt{\lambda}} + \frac{1}{64\lambda^{\frac{3}{2}}} - \dots \\ DY &= D\sqrt{X + \frac{3}{8}} = \frac{1}{4} \left( 1 + \frac{1}{16\lambda^2} \right). \end{aligned}$$

## BINOMICKÉ ROZDĚLENÍ

Nechť náhodná veličina  $Z$  má **binomické** rozdělení s parametry  $n \in \mathbb{N}$ ,  $\theta \in (0, 1)$  tj.  $Z \sim Bi(n, \theta)$  s pravděpodobnostní funkcí  $p_Z(z) = P(Z = z) = \binom{n}{z} \theta^z (1 - \theta)^{n-z}$  pro  $z = 0, 1, 2, \dots, n$ . Lze spočítat, že  $EZ = n\theta$  a  $DZ = n\theta(1 - \theta)$ .

**Relativní četnost úspěchů v  $n$  nezávislých pokusech**  $X = \frac{Z}{n}$  má střední hodnotu a rozptyl

$$\begin{aligned} EX &= \theta \\ DX &= D\frac{Z}{n} = \frac{1}{n^2} DZ = \frac{\theta(1-\theta)}{n}, \quad \text{tj.} \quad \sigma^2(\theta) = \frac{\theta(1-\theta)}{n}. \end{aligned}$$

Pak pro  $n \rightarrow \infty$  platí

$$\frac{X - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} = \sqrt{n} \frac{X - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{L} U \sim N(0, 1).$$

Chceme najít takovou funkci  $g$ , aby asymptotické rozdělení náhodné veličiny

$$g(X) - g(\theta) \xrightarrow{L} Y \sim N(0, c) \quad c > 0.$$

Pak

$$\begin{aligned} g(\theta) &= \int \frac{cd\theta}{\sigma(\theta)} + K = \int \frac{cd\theta}{\sqrt{\theta(1-\theta)}} + K = \left| \frac{1}{2} \frac{1}{\sqrt{\theta}} d\theta = du \right| \\ &= 2c\sqrt{n} \int \frac{du}{\sqrt{1-u^2}} + K = 2c\sqrt{n} \arcsin \sqrt{\theta} + K. \end{aligned}$$

Zvolíme-li

$$c = \frac{1}{2\sqrt{n}}, K = 0,$$

dostaneme známou **arcussinovou transformaci**

$$Y = g(X) = \arcsin \sqrt{X} = \arcsin \sqrt{\frac{Z}{n}}$$

se střední hodnotou a rozptylem

$$\begin{aligned} EY &= Eg(X) \approx g(\theta) = \arcsin \sqrt{\theta} \\ DY &= Dg(X) \approx [g'(\theta)]^2 \sigma^2(\theta) = \left[ \frac{1}{\sqrt{1-\theta}} \frac{1}{2} \frac{1}{\sqrt{\theta}} \right]^2 \frac{\theta(1-\theta)}{n} = \frac{1}{4n}. \end{aligned}$$

Anscombe (1948) opět navrhl stabilnější transformaci:

$$Y = g(X) = \arcsin \sqrt{\frac{X + \frac{3}{8n}}{1 + \frac{3}{4n}}} = \arcsin \sqrt{\frac{Z + \frac{3}{8n}}{n + \frac{3}{4}}},$$

přičemž

$$EY \approx \arcsin \sqrt{\frac{\theta + \frac{3}{8n}}{1 + \frac{3}{4n}}} \quad \text{a} \quad DY = \frac{1}{4n + 2}.$$

$\chi^2$  ROZDĚLENÍ

Necht' náhodná veličina  $X$  má  $\chi^2$  **rozdělení** s parametrem  $\nu > 0$ , tj.  $X \sim \chi^2(\nu)$ .  
Lze spočítat, že  $EX = \nu$  a  $DX = 2\nu$ , tj.  $\sigma^2(\nu) = 2\nu$ . Pak

$$g(\nu) = c \int \frac{1}{\sigma(\nu)} d\nu + K = c \int \frac{1}{\sqrt{2\nu}} d\nu + K = c\sqrt{2\nu} + K.$$

Obvykle se volí  $c = 1$ ,  $K = 0$  a pracuje se s velmi známou **odmocninovou transformací**

$$Y = g(X) = \sqrt{2X}.$$

Spočítejme střední hodnotu a rozptyl náhodné veličiny  $Y$

$$EY = Eg(X) \approx g(\nu) = \sqrt{2\nu}$$

$$DY = Dg(X) \approx [g'(\nu)]^2 \sigma^2(\nu) = \left[ \frac{1}{2} \frac{1}{\sqrt{\nu}} \sqrt{2} \right]^2 2\nu = 1.$$

R. A. Fisher doporučil raději užívat transformaci

$$Y = g(X) = \sqrt{2X} - \sqrt{2\nu - 1},$$

jejíž rozdělení se blíží normálnímu rozdělení  $N(0, 1)$ .

Poznamenejme, že dnes se často užívá transformace

$$Y = g(X) = 3\sqrt{\frac{\nu}{2}} \left( \sqrt[3]{\frac{X}{\nu}} + \frac{2}{9\nu} - 1 \right),$$

jejíž rozdělení se blíží standardizovanému normálnímu rozdělení ještě rychleji (viz Rao 1978).

**C. Mocninné transformace**

Mějme kladnou náhodnou veličinu  $X$  z rozdělení, které závisí na parametru  $\theta$  se střední hodnotou a rozptylem

$$E_{\mu}X = \mu$$

$$D_{\mu}X = \sigma^2(\mu) = (\sigma\mu^{\vartheta})^2 \quad \sigma \in \mathbb{R}, \quad \text{tj.} \quad X \sim \mathcal{L}(\mu, \sigma^2\mu^{2\vartheta}).$$

Podle **obecného vzorce** se transformace stabilizující rozptyl vypočítá takto:

$$g(\mu) = \int \frac{cd\mu}{\sigma(\mu)} + K = \frac{c}{\sigma} \int \frac{d\mu}{\mu^{\vartheta}} + K = \begin{cases} \frac{c}{\sigma} \ln |\mu| + K & \vartheta = 1, \\ \frac{c}{1-\vartheta} \mu^{1-\vartheta} + K & \vartheta \neq 1. \end{cases}$$

Položme v dalším

$$\lambda = 1 - \vartheta$$

a tento parametr nazvěme **transformačním parametrem** pro mocninnou transformaci.

Různou volbou  $c$  a  $K$  dostaneme následující často užívané transformace

- **Box-Coxova mocninná transformace** pro kladné náhodné veličiny při volbě

$$c = \sigma \quad \text{a} \quad K = \begin{cases} 0 & \lambda = 0 \Rightarrow \vartheta = 1, \\ -\frac{1}{\lambda} = -\frac{1}{1-\vartheta} & \lambda \neq 0 \Rightarrow \vartheta \neq 1, \end{cases}$$

a odtud

$$g(X) = X^{(\lambda)} = \begin{cases} \ln X & \lambda = 0 (\vartheta = 1), \\ \frac{X^\lambda - 1}{\lambda} & \lambda \neq 0 (\vartheta \neq 1). \end{cases}$$

- **Box-Coxova mocninná transformace s posunutím** se použije v případě, že hodnoty náhodné veličiny nejsou kladné. Nalezneme proto takové reálné číslo  $a$  tak, aby pro všechny realizace platilo  $x + a > 0$  a transformace bude mít tvar:

$$g(X + a) = (X + a)^{(\lambda)} = \begin{cases} \ln(X + a) & \lambda = 0 (\vartheta = 1), \\ \frac{(X+a)^\lambda - 1}{\lambda} & \lambda \neq 0 (\vartheta \neq 1). \end{cases}$$

- **Mocninná transformace se znaménkem** lze opět použít v případě, že náhodné veličiny nejsou kladné:

$$g(X) = \text{sign}(X)|X|^{(\lambda)} = \begin{cases} \text{sign}(X) \ln |X| & \lambda = 0 (\vartheta = 1), \\ \text{sign}(X) \frac{|X|^\lambda - 1}{\lambda} & \lambda \neq 0 (\vartheta \neq 1). \end{cases}$$

## D. Odhad transformačního parametru mocninné transformace

- **Parametrický přístup pomocí metody maximální věrohodnosti.** Mějme nezávislé realizace náhodné veličiny

$$X \sim \mathcal{L}(\mu, \sigma^2 \mu^{2\vartheta}).$$

Předpokládejme, že existuje takové

$$\lambda = 1 - \vartheta,$$

že transformovaný náhodný vektor

$$\mathbf{Y} = (Y_1 = g(X_1), \dots, Y_n = g(X_n))'$$

je výběr z normálního rozdělení se střední hodnotou  $\mu$  a rozptylem  $\sigma^2$ . Označme

$$\mathbf{y} = (y_1, \dots, y_n)'$$

realizaci náhodného výběru.

Hledejme maximum **věrohodnostní funkce** pro  $\boldsymbol{\theta} = (\mu, \sigma^2)'$ , tj. pro funkci

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^n \left[ -\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right\} \right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \left( \frac{y_i - \mu}{\sigma} \right)^2 \right\}, \end{aligned}$$

což je stejná úloha jako hledat maximum **logaritmu věrohodnostní funkce**

$$l(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \left( \frac{y_i - \mu}{\sigma} \right)^2.$$

Maxima nalezneme, položíme-li  $\frac{\partial l}{\partial \mu} = 0$  a  $\frac{\partial l}{\partial \sigma^2} = 0$ .

$$0 = \frac{\partial l}{\partial \mu} = \frac{2}{2\sigma^2} \sum_{i=1}^n (y_i - \mu) \quad \Rightarrow \quad \boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}}$$

$$0 = \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2 \quad \Rightarrow \quad \boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

Upravme nyní logaritmus věrohodnostní funkce takto:

$$\begin{aligned} l(\mu, \sigma^2) &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [(y_i - \bar{y}) + (-\bar{y} - \mu)]^2 \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \left\{ \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \right\} \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} [ns^2 + n(\bar{y} - \mu)^2] \end{aligned}$$

Nyní dokažme, že funkce  $l(\mu, \sigma^2)$  nabývá v bodě  $(\hat{\mu}, \hat{\sigma}^2) = (\bar{y}, s^2)$  svého maxima. Platí

$$l(\bar{y}, s^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(s^2) - \frac{n}{2},$$

Ověřme, zda platí nerovnost

$$\begin{aligned} l(\mu, \sigma^2) &\stackrel{?}{\leq} l(\bar{y}, s^2) \\ -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{ns^2 + n(\bar{y} - \mu)^2}{2\sigma^2} &\stackrel{?}{\leq} -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(s^2) - \frac{n}{2} \\ -\frac{1}{2} \ln(\sigma^2) - \frac{s^2}{2\sigma^2} - \frac{(\bar{y} - \mu)^2}{2\sigma^2} &\stackrel{?}{\leq} -\frac{n}{2} \ln(s^2) - \frac{1}{2} \\ 0 &\stackrel{?}{\leq} \underbrace{\left[ \left( \frac{s^2}{2\sigma^2} - \frac{1}{2} \right) - \ln \frac{s}{\sigma} \right]}_{\text{1. člen}} + \underbrace{\frac{(\bar{y} - \mu)^2}{2\sigma^2}}_{\geq 0} \end{aligned}$$

Protože pro všechna kladná  $x = \frac{s}{\sigma} > 0$  platí  $\ln x < \frac{x^2 - 1}{2}$ , je první i druhý člen nezáporný a nerovnost platí.

Celkově jsme tedy dostali, že

$$\max_{\mu, \sigma^2} l(\mu, \sigma^2) = l(\bar{y}, s^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(s^2) - \frac{n}{2}$$

a

$$\max_{\mu, \sigma^2} L(\mu, \sigma^2) = L(\bar{y}, s^2) = (2\pi s^2)^{-\frac{n}{2}} e^{-\frac{n}{2}}.$$

Nyní toto maximum vyjádříme v původních proměnných  $x_i$ , kdy

$$y_i = g(x_i) = \begin{cases} \ln x_i & \lambda = 0, \\ \frac{x_i^\lambda - 1}{\lambda} & \lambda \neq 0. \end{cases}$$

Nejprve vypočítáme jakobián této transformace:

$$|J| = \prod_{i=1}^n \left| \frac{dy_i}{dx_i} \right| = \prod_{i=1}^n \frac{\lambda x_i^{\lambda-1}}{\lambda} = \prod_{i=1}^n x_i^{\lambda-1}.$$

Pak

$$\begin{aligned} \max_{\mu, \sigma^2, \lambda} L(\mu, \sigma^2) &= (2\pi s^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}|J|} \\ &= (2\pi s^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^n x_i^{\lambda-1} \\ &= (2\pi s^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^n e^{(\lambda-1)\ln x_i} \\ &= (2\pi s^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2} + (\lambda-1) \sum_{i=1}^n \ln x_i} \end{aligned}$$

$$\max_{\mu, \sigma^2} l(\mu, \sigma^2, \lambda) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(s^2(\lambda)) - \frac{n}{2} + (\lambda - 1) \sum_{i=1}^n \ln x_i.$$

Nyní hledíme maximum funkce  $l(\hat{\mu}, \hat{\sigma}^2, \lambda) = l(\bar{y}, s^2, \lambda)$  pro parametr  $\lambda$ . Protože maximum vzhledem k  $\lambda$  nezávisí na konstantách, budeme maximalizovat funkci

$$l^*(\lambda) = -\frac{n}{2} \ln(s^2(\lambda)) + (\lambda - 1) \sum_{i=1}^n \ln x_i.$$

Teoretickým odvozením maximálně věrohodného odhadu parametru  $\lambda$ , se zde nebudeme zabývat, ale ukážeme si **jednodušší přístup**: pro různé hodnoty  $\lambda \in (\lambda_1, \lambda_2)$  ( $\lambda_1, \lambda_2 \in \mathbb{R}$ ,  $\lambda_1 < \lambda_2$ ) se vykreslí do grafu hodnoty  $l^*(\lambda)$  a hledá se maximum  $\hat{\lambda}$  v daném intervalu.

V tomto případě Box-Cox (1964) odvodili asymptotické rozdělení statistiky

$$K = -2 \left[ l^*(\lambda) - l^*(\hat{\lambda}) \right] \xrightarrow{L} \chi^2(1)$$

**Interval spolehlivosti pro parametr  $\lambda$ :**

$$1-\alpha = P(K < \chi_{1-\alpha}^2(1)) = P\left(-2 \left[ l^*(\lambda) - l^*(\hat{\lambda}) \right] < \chi_{1-\alpha}^2(1)\right) = P\left(\underbrace{l^*(\hat{\lambda}) - \frac{1}{2} \chi_{1-\alpha}^2(1)}_{=D_\alpha} \leq l^*(\lambda)\right)$$

tj. všechna  $\lambda$  splňující nerovnost  $l^*(\lambda) \geq D_\alpha$  leží v intervalu spolehlivosti a jsou tedy přijatelná.

**Testování hypotéz typu  $H_0 : \lambda = \lambda_0$  proti alternativě  $H_1 : \lambda > \lambda_0$ :**

- (a) Budeme testovat hypotézu  $H_0^1 : \lambda = 1$ . Pokud hypotézu **nezamítneme**, tj.  $l^*(1) \geq D_\alpha$ , **nemusíme data transformovat**.
- (b) Pokud předchozí hypotézu **zamítneme**, můžeme testovat další hypotézu  $H_0^2 : \lambda = 0$ . Pokud tuto hypotézu **nezamítneme**, tj.  $l^*(0) \geq D_\alpha \wedge l^*(1) < D_\alpha$ , transformace bude tvaru

$$y_i = \ln x_i.$$

Pokud však se  $l^*(0) < D_\alpha \wedge l^*(1) < D_\alpha$ , provedeme transformaci

$$y_i = \frac{x_i^{\hat{\lambda}} - 1}{\hat{\lambda}}.$$

• **Jednoduchý algoritmus v praktických úlohách** – funkce `powtr()`

- (a) Algoritmus nejprve zkontroluje vstupní data tak, aby byla **nezáporná**, tj. případně přičte kladnou konstantu. Upravený vektor dat rozdělí (podle nějakého dalšího kritéria, pokud nejsou opakovaná pozorování; např. u časových řad jsou data uspořádána podle časového kritéria) na krátké úseky o délce 4 až 12 údajů. V každém úseku dat se provede pokud možno robustní odhad polohy  $\hat{\mu}$  (průměr, medián) a robustní odhad variability  $\hat{\sigma}^2$  (např. max-min, interkvartilové rozpětí *IQR*). Protože předpokládáme, že

$$\sigma(\mu) = \sigma\mu^{\vartheta} \quad \Rightarrow \quad \ln(\sigma(\mu)) = \ln\sigma + \vartheta \ln(\mu),$$

neznámé  $\vartheta$  odhadneme **metodou nejmenších čtverců**.

- (b) Pro odhad  $\hat{\vartheta} = 1 - \hat{\lambda}$  pomocí t-statistiky zkonstruujeme interval spolehlivosti  $I(\hat{\vartheta})$ .
- Pokud tento interval bude obsahovat **nulu**, tj.  $0 \in I(\hat{\vartheta})$  data se **nebudou transformovat**

$$y_i = x_i.$$

- Pokud  $0 \notin I(\hat{\vartheta}) \wedge 1 \in I(\hat{\vartheta})$ , volí se logaritmická transformace

$$y_i = \ln x_i.$$

- Jinak se volí mocninná transformace

$$y_i = x_i^{\hat{\lambda}}.$$

### PŘÍKLAD 1

Datový soubor v prvním a druhém řádku obsahuje popis časové řady. Od třetího řádku jsou samotná data.

```
> fileN <- "AustralianElectricity.txt"
> fileDat <- paste(data.library, fileN, sep = "")
> con <- file(fileDat)
> (POPIS <- readLines(con, n = 2))
```

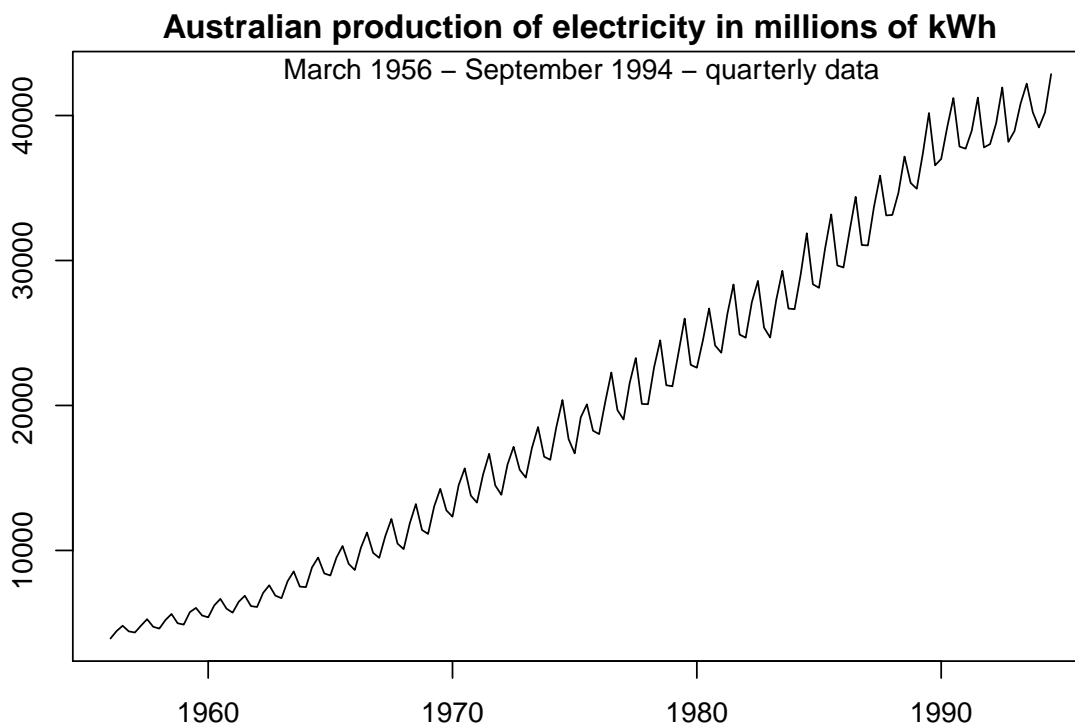
```
[1] "Australian production of electricity in millions of kWh"
[2] "March 1956 - September 1994 - quarterly data"
```

```
> close(con)
> Aelec <- scan(fileDat, skip = 2)
```

Z načtených dat vytvoříme časovou řadu a vykreslíme ji.

```
> AelecTS <- ts(Aelec, start = 1956, frequency = 4)
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(AelecTS, main = POPIS[1], cex.main = 1.15)
> mtext(POPIS[2], side = 3, line = -1)
```





Obrázek 1: Australská produkce elektřiny v milionech kWh (březen 1956 - září 1994) - čtvrtletní údaje

Z grafu je patrné, že data bude třeba transformovat, neboť s rostoucí střední hodnotou roste také variabilita.

Zvolíme mocninnou transformaci. Začneme s jednoduchým přístupem založeným na regresním modelu (funkce `powtr()`), který využívá vztah

$$\sigma(\mu) = \sigma\mu^{\vartheta} \quad \Rightarrow \quad \ln(\sigma(\mu)) = \ln\sigma + \vartheta\ln(\mu),$$

a neznámé parametry odhaduje **metodou nejmenších čtverců**. Odhadem směrnice regresní přímky získáme parametr  $\vartheta$  a tím také parametr  $\lambda = 1 - \vartheta$ .

Funkce `powtr()` rozdělí nezávisle proměnnou, tj. čas, na subintervaly o velikosti, který určuje parametr `seglen` (doporučuje se volit číslo mezi 4 až 12).

Pro každý subinterval se provede odhad polohy a variability (parametry `location` a `variability`).

Polohu můžeme odhadnout buď pomocí výběrového průměru (`location="mean"`) nebo výběrového mediánu (`location="median"`).

Odhad variability lze provést pomocí výběrové směrodatné odchylky (`variability="sd"`), výběrového interkvartilového rozpětí (`variability="iqr"`) nebo pomocí rozpětí, tj. rozdílu mezi maximem a minimem (`variability="range"`).

Nejprve položíme parametr `seglen=8` a pro odhad polohy a variability zvolíme medián a interkvartilové rozpětí. Volbou `figure=TRUE` získáme výsledek v grafické podobě.

```

> x <- Aelec
> seglen <- 8
> location <- "median"
> variability <- "iqr"
> outp <- powtr(x, seglen = seglen, figure = TRUE, location = location,
  variability = variability)
> str(outp)

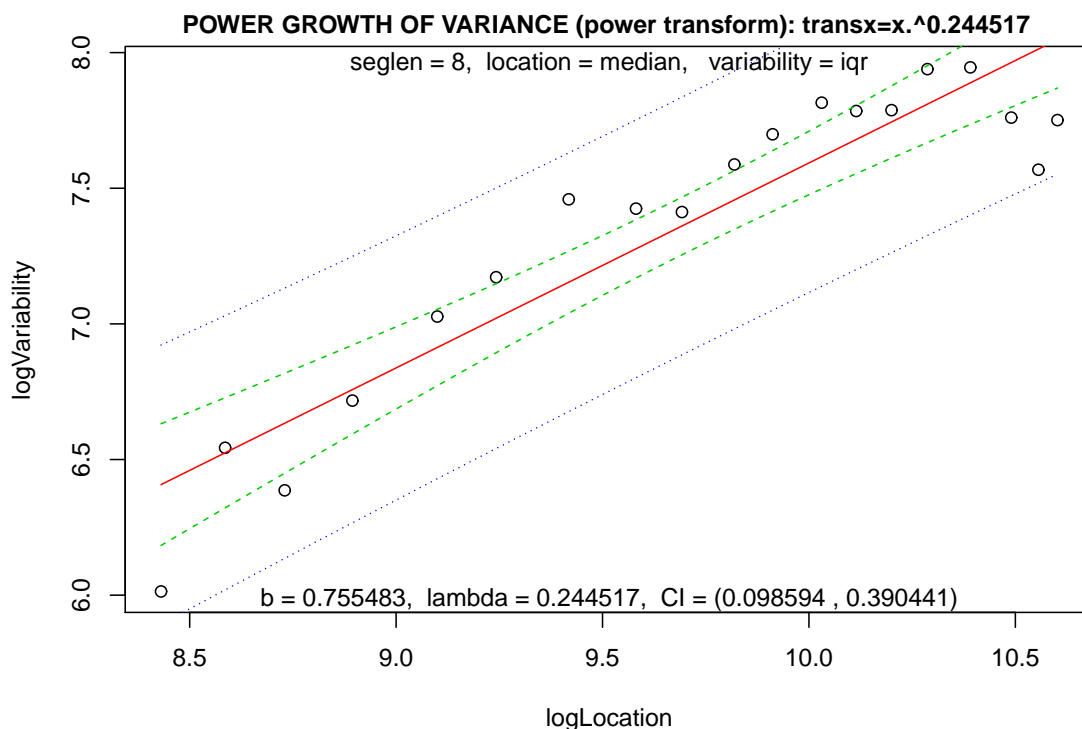
```

List of 3

```

$ lambda : Named num 0.245
  .. attr(*, "names")= chr "Estimate"
$ transfx: num [1:155] 7.56 7.79 7.95 7.79 7.75 ...
$ txt    : chr "POWER GROWTH OF VARIANCE (power transform): transx=x.^0.244517"

```



Obrázek 2: Mocninná transformace pomocí funkce `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

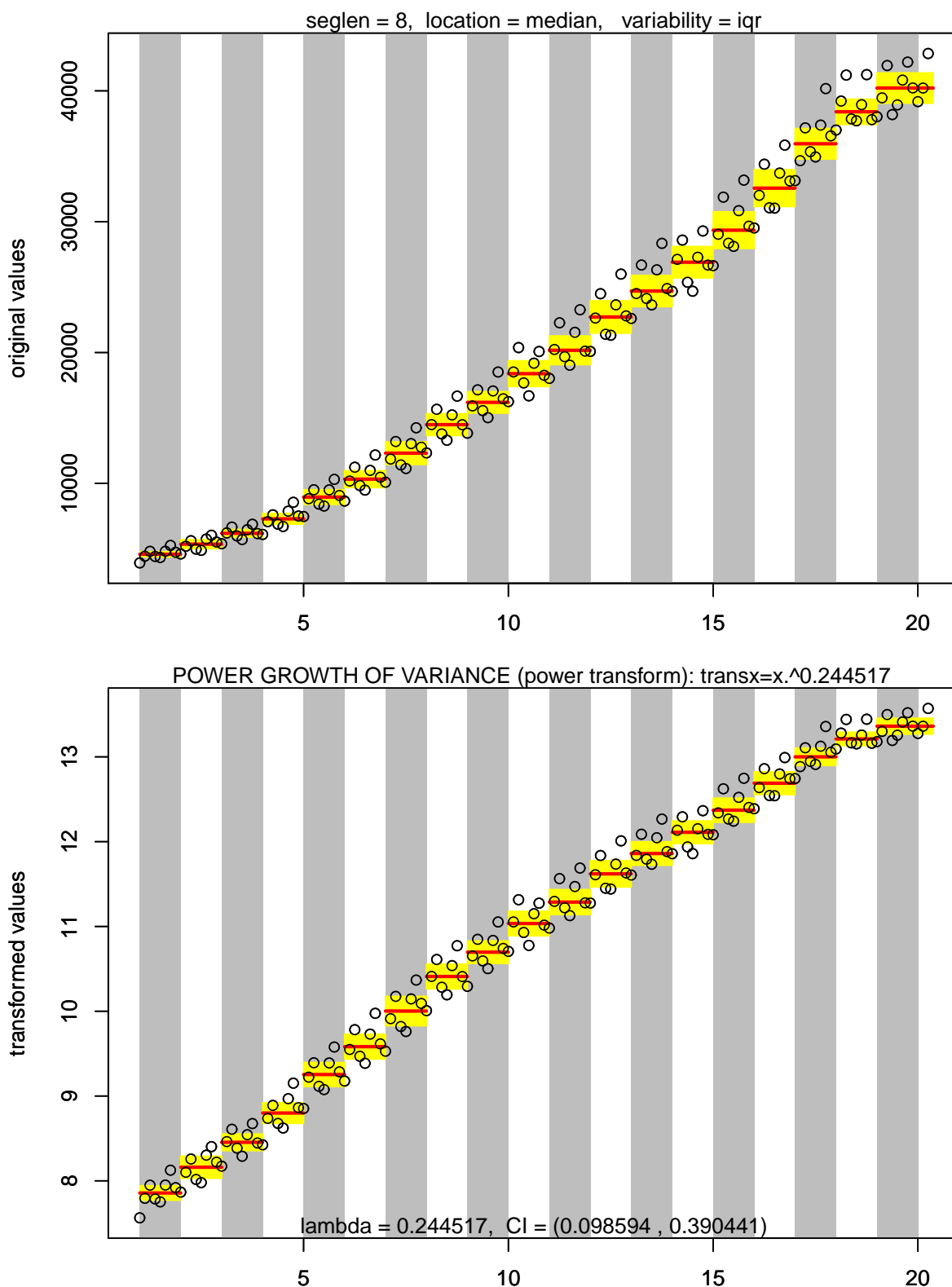
Funkce `powtr` nabízí další zajímavý graf, který názorně ukazuje, jak vypadá variabilita dat v jednotlivých segmentech před a po transformaci (`figure2=TRUE`). Tento graf by měl také ukázat, zda je vůbec mocninná transformace vhodná.

Pokud nedojde ke stabilizaci rozptylu ani po transformaci, pak bude třeba hledat jiný typ transformace než je mocninná.

```

> outp <- powtr(x, seglen = seglen, figure2 = TRUE, location = location,
  variability = variability)

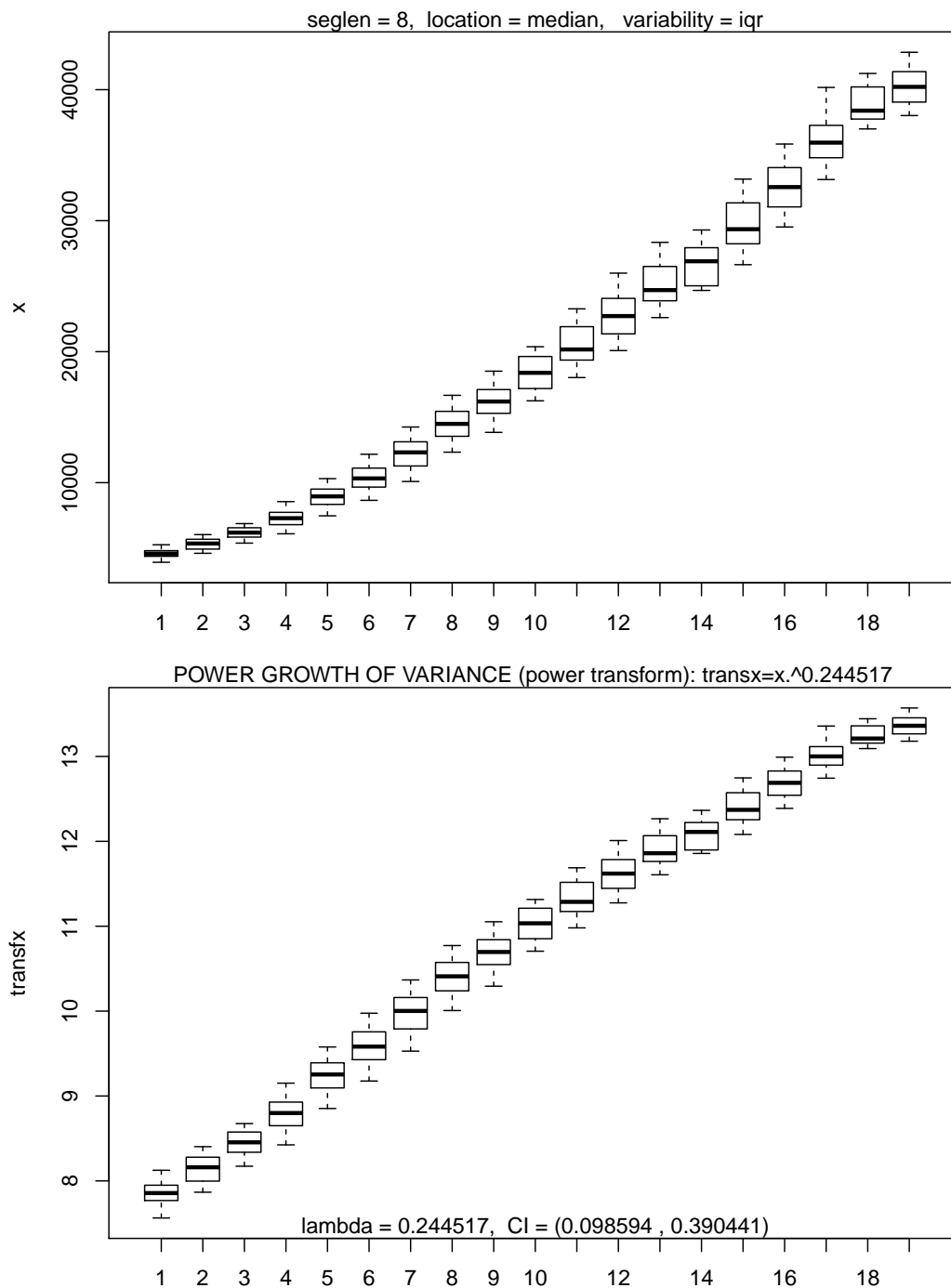
```



Obrázek 3: Mocninná transformace pomocí funkce `powtr` – odhady polohy a variability v jednotlivých segmentech (volba `figure2=TRUE`) pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

Obdobný výstup, ale vyjádřený pomocí krabicových grafů (boxplotů) za jednotlivé segmenty vstupních dat, získáme volbou `figure3=TRUE`.

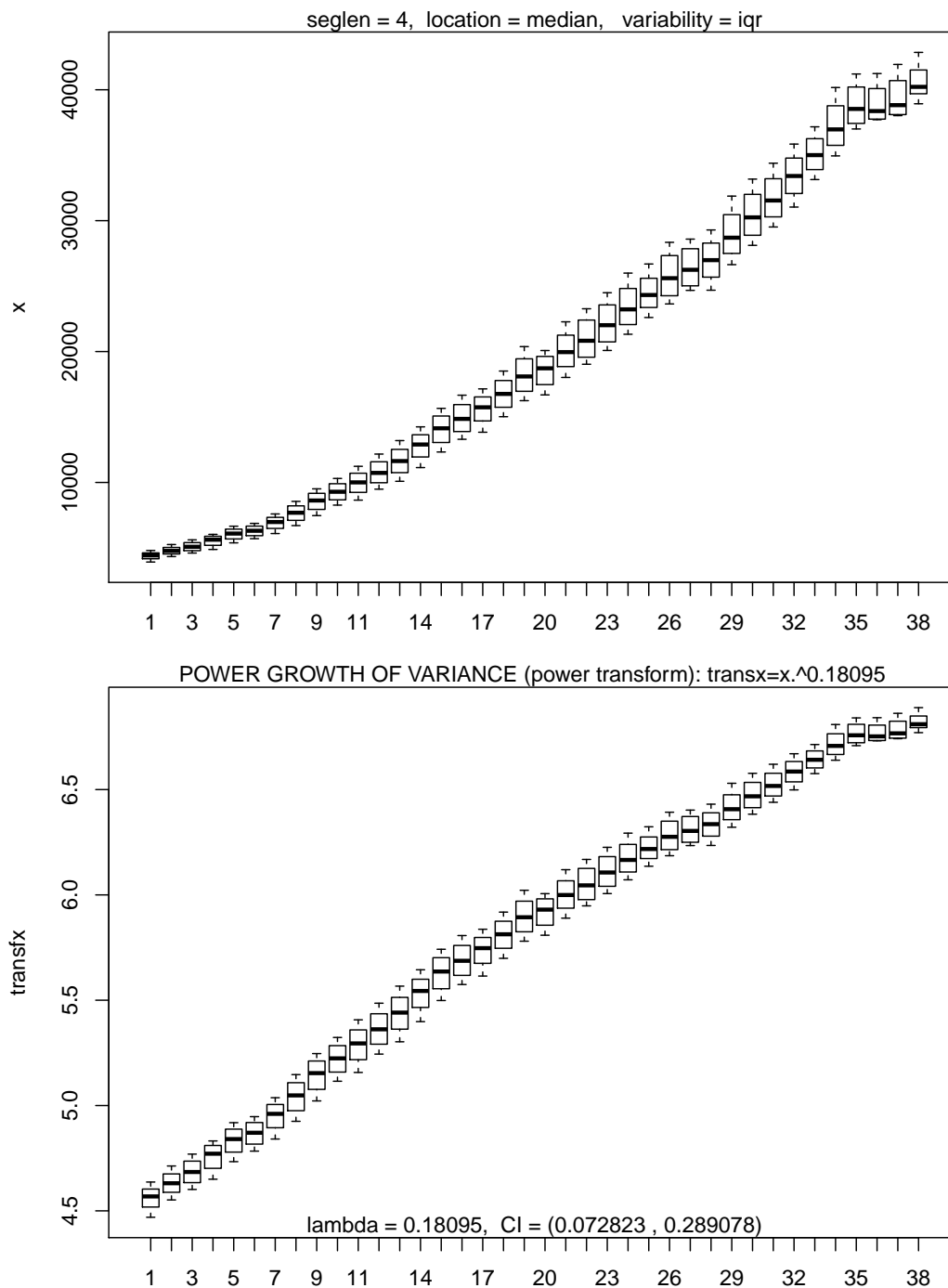
```
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
  variability = variability)
```



Obrázek 4: Mocniná transformace pomocí funkce `powtr` – krabicové grafy v jednotlivých segmentech (volba `figure3=TRUE`) pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

Abychom vyzkoušeli robustnost funkce `powtr()` na našich datech, provedeme odhad mocninné transformace postupně pro `seglen=4,6,10,12`, ostatní parametry necháme nezměněny a vykreslíme třetí graf.

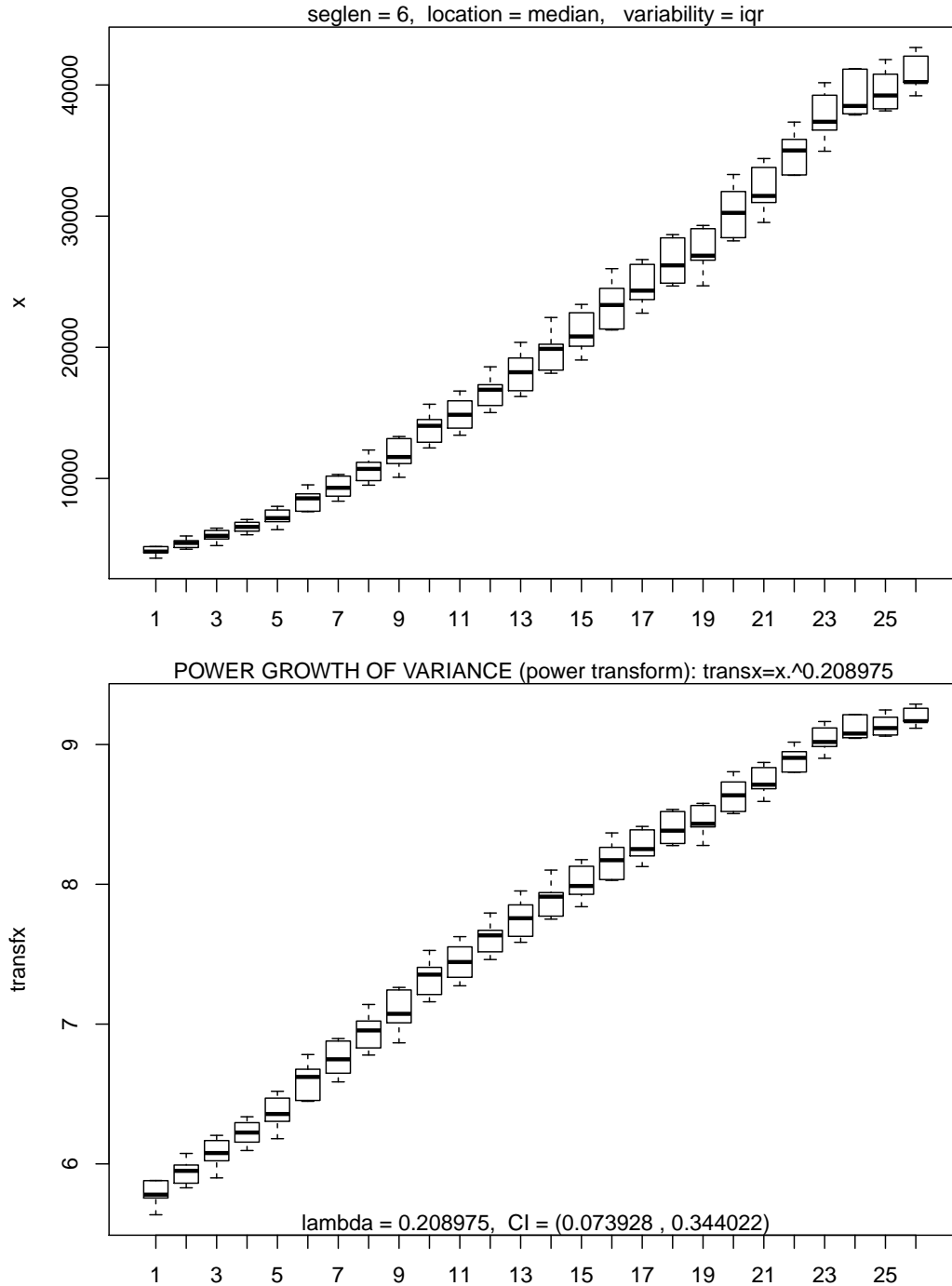
```
> seglen = 4
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
  variability = variability)
```



Obrázek 5: `powtr` – krabicové grafy (volba `seglen=4` a `figure3=TRUE`) pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

Totéž znovu zopakujeme pro volbu `seglen=6`.

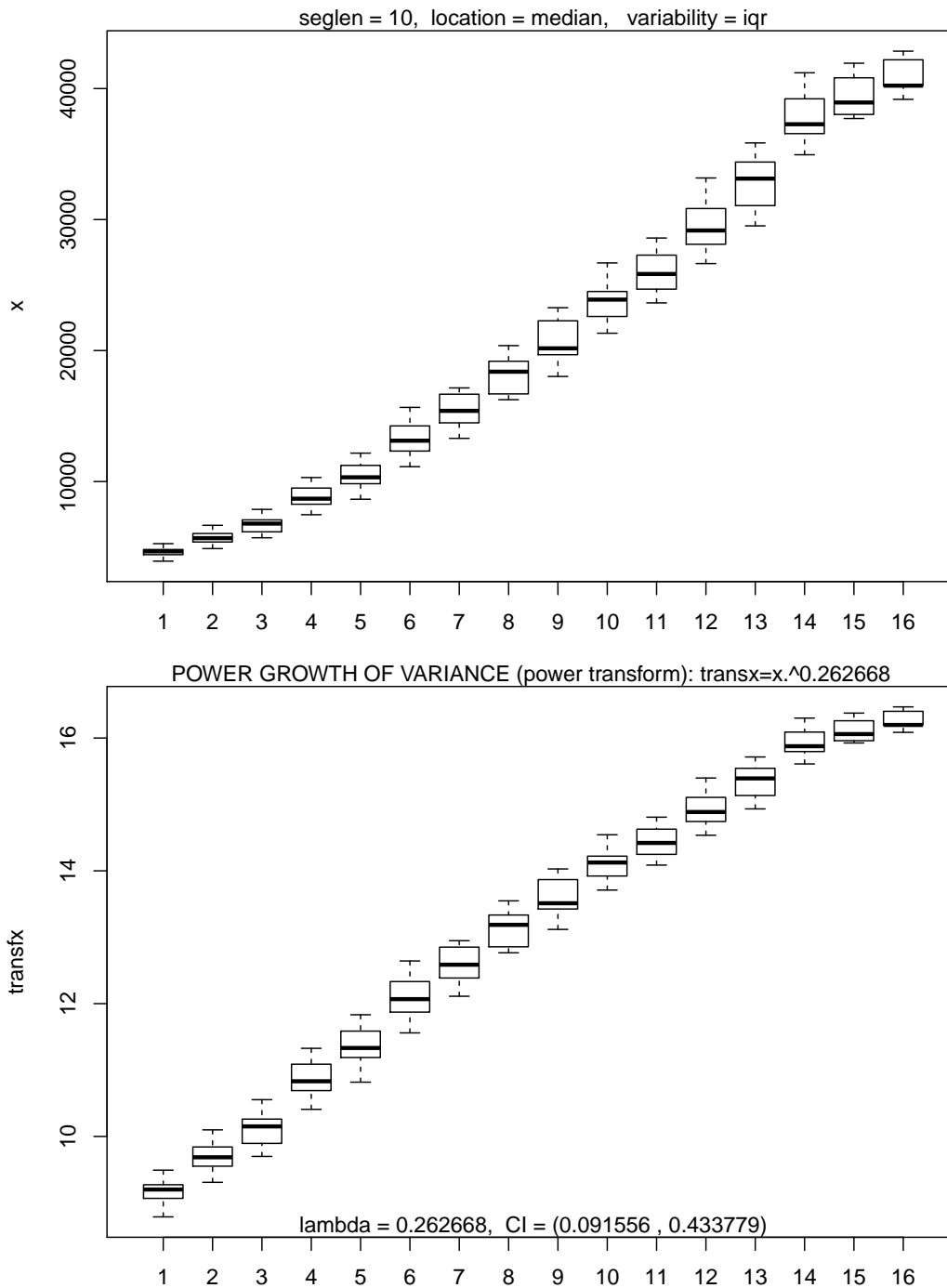
```
> seglen = 6
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
  variability = variability)
```



Obrázek 6: `powtr` – krabicové grafy (volba `seglen=6` a `figure3=TRUE`) pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

Další volba `seglen=10`.

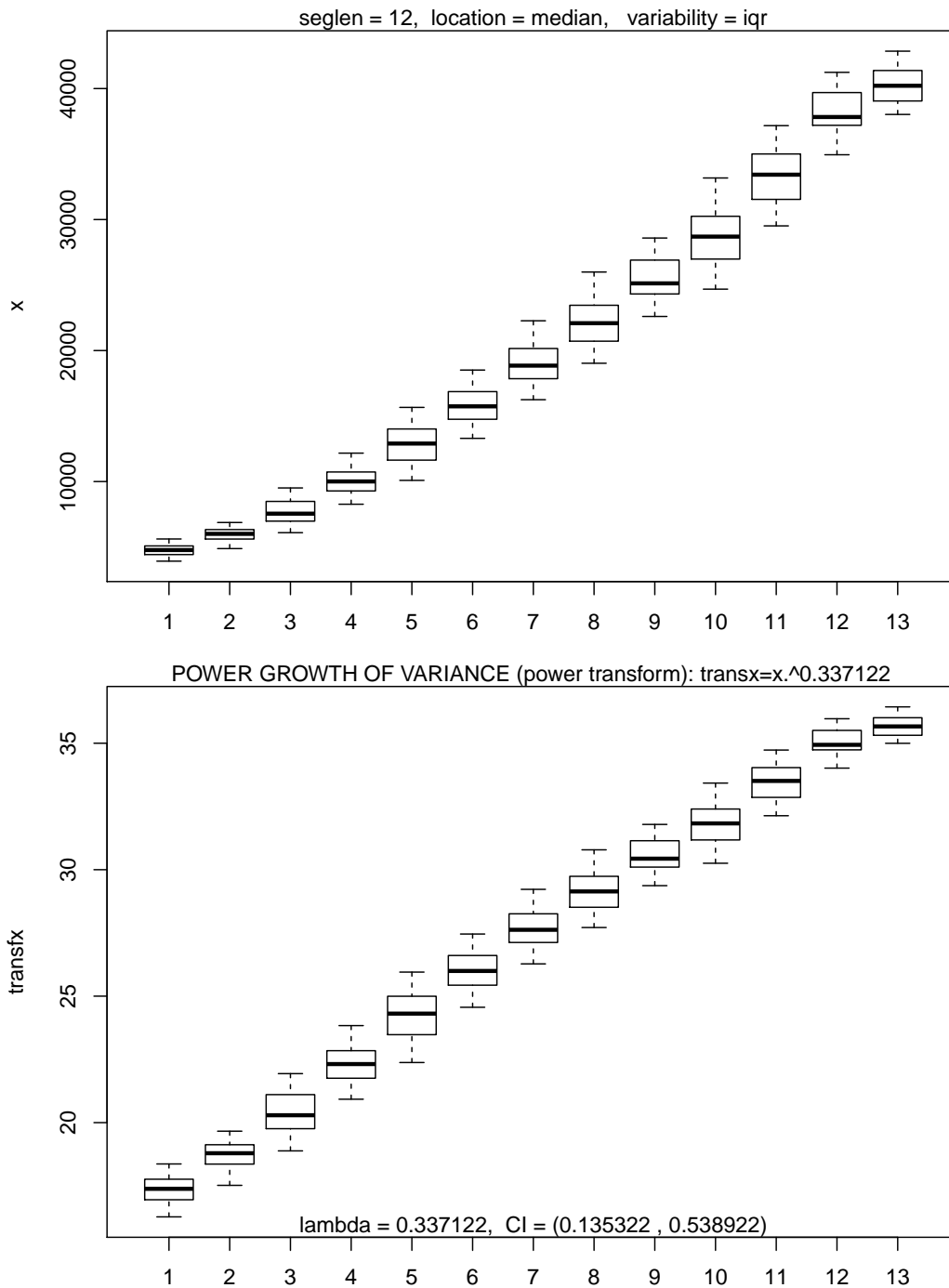
```
> seglen = 10
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
  variability = variability)
```



Obrázek 7: `powtr` – krabicové grafy (volba `seglen=10` a `figure3=TRUE`) pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

Poslední volba `seglen=12`.

```
> seglen = 12
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
  variability = variability)
```



Obrázek 8: `powtr` – krabicové grafy (volba `seglen=12` a `figure3=TRUE`) pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“



Výsledky přechozích kroků shrňme do tabulky.

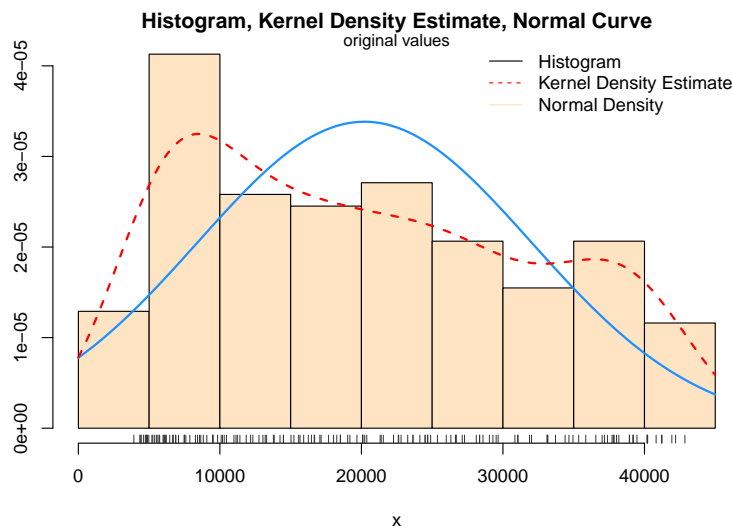
| seglen | dolní mez | odhad $\hat{\lambda}$ | horní mez |
|--------|-----------|-----------------------|-----------|
| 4      | 0.073     | 0.181                 | 0.289     |
| 6      | 0.074     | 0.209                 | 0.344     |
| 8      | 0.099     | 0.245                 | 0.390     |
| 10     | 0.092     | 0.263                 | 0.434     |
| 12     | 0.135     | 0.337                 | 0.539     |

Protože ani jednou interval spolehlivosti neobsahuje nulu, rozhodně nebudeme logaritmovat.

Bývá zvykem vybírat transformace z hodnot, které mají rozumnou interpretaci. V našem případě připadají v úvahu například hodnoty  $\frac{1}{5}$ ,  $\frac{1}{4}$  nebo  $\frac{1}{3}$ .

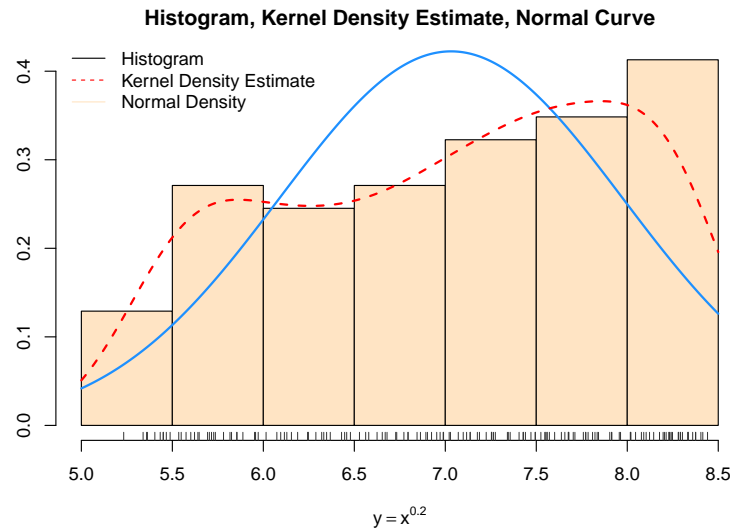
Postupně vykreslíme histogram (spolu s jádrovým odhadem i normální hustotou) pomocí funkce `HistFit()` pro netransformovaná data, následně pro jednotlivé transformace.

```
> x <- Aelec
> HistFit(x)
> mtext("original values", side = 3, line = -0.5, cex = 0.95)
```



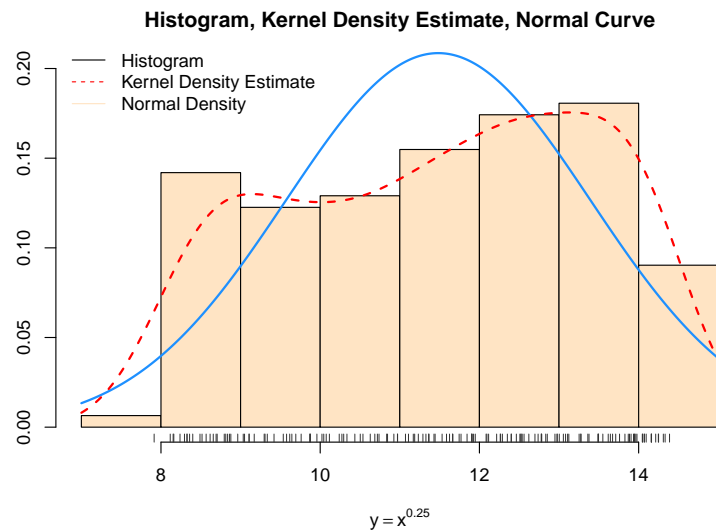
Obrázek 9: Testování normality pro netrasformovaná data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

```
> HistFit(x^0.2, xlab = expression(y == x^0.2))
```



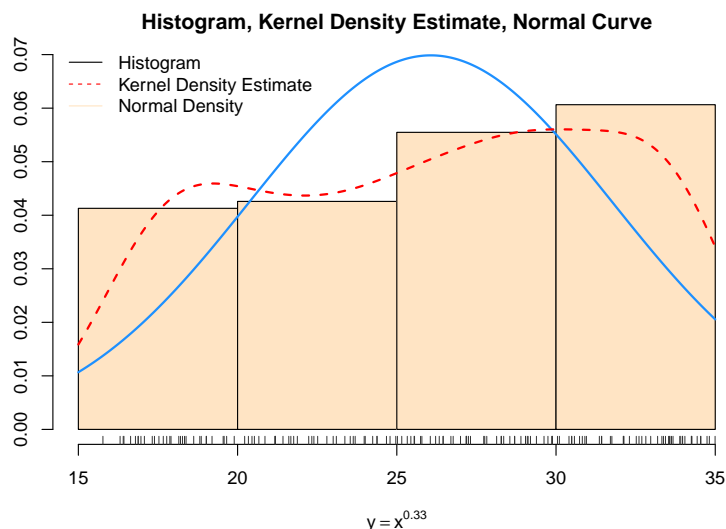
Obrázek 10: Testování normality pro transformovaná data ( $Y = \sqrt[5]{X}$ ) „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

```
> HistFit(x^0.25, xlab = expression(y == x^0.25))
```



Obrázek 11: Testování normality pro transformovaná data ( $Y = \sqrt[4]{X}$ ) „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

```
> HistFit(x^(1/3), xlab = expression(y == x^0.33))
```



Obrázek 12: Testování normality pro transformovaná data ( $Y = \sqrt[3]{X}$ ) „Australská produkce elektriny (březen 1956 - září 1994) - čtvrtletní údaje“

Z grafu je ihned vidět, že se odhadnuté hustoty nepřibližují k normalitě. Je třeba si však uvědomit, že to není ani tak volbou transformace, jako spíše faktem, že časová řada má výrazný deterministický lineární trend, který pak přehluší stochastické vlastnosti kolísání kolem trendu. Ihned nás napadne myšlenka nejprve odstranit lineární trend a teprve pro rezidua hledat vhodnou mocninnou transformaci.

```
> y <- as.vector(AelecTS)
> x <- as.vector(time(AelecTS))
> n <- length(y)
> nn <- 300
> data <- data.frame(x, y)
> LinTrend <- lm(y ~ x, data = data)
> print(summary(LinTrend))
```

Call:

```
lm(formula = y ~ x, data = data)
```

Residuals:

| Min     | 1Q      | Median | 3Q     | Max    |
|---------|---------|--------|--------|--------|
| -3561.3 | -1640.1 | -58.8  | 1247.4 | 5188.0 |

Coefficients:

|             | Estimate   | Std. Error | t value | Pr(> t )   |
|-------------|------------|------------|---------|------------|
| (Intercept) | -2.026e+06 | 2.795e+04  | -72.50  | <2e-16 *** |
| x           | 1.036e+03  | 1.415e+01  | 73.23   | <2e-16 *** |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1970 on 153 degrees of freedom

Multiple R-squared: 0.9723, Adjusted R-squared: 0.9721

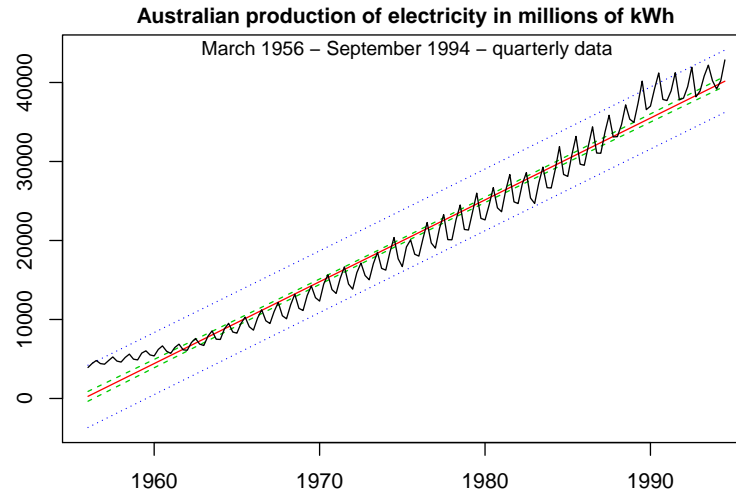
F-statistic: 5362 on 1 and 153 DF, p-value: < 2.2e-16

```
> new <- data.frame(x = seq(x[1], x[n], length.out = nn))
> pred.w.plim <- predict(LinTrend, new, interval = "prediction")
```

```

> pred.w.clim <- predict(LinTrend, new, interval = "confidence")
> par(mar = c(2, 2, 1, 0) + 0.5)
> matplot(new$x, cbind(pred.w.clim, pred.w.plim[, -1]), col = c(2,
  3, 3, 4, 4), lty = c(1, 2, 2, 3, 3), type = "l", ylab = "predicted y")
> lines(x, y)
> title(main = POPIS[1], cex.main = 1)
> mtext(POPIS[2], side = 3, line = -1, cex = 0.95)

```



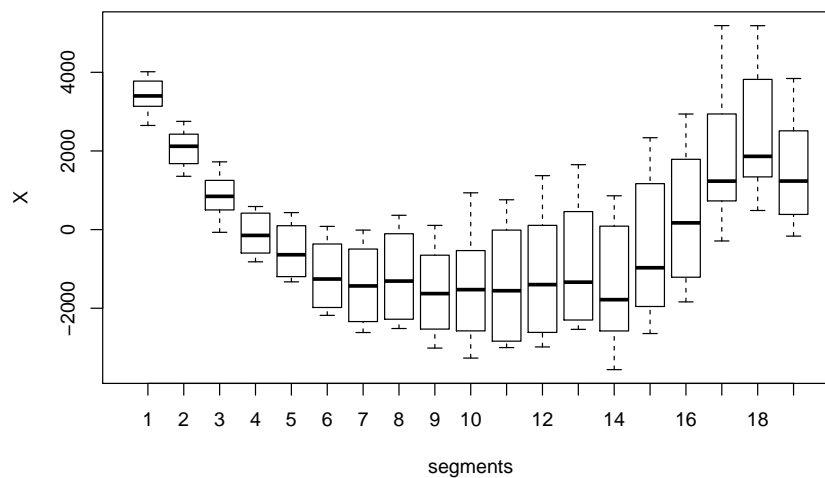
Obrázek 13: Lineární trend pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

Dříve než použijeme mocninovou transformaci, podívejme se pomocí funkce `boxplotSegments()`, zda to pro rezidua má vůbec smysl.

```

> x <- resid(LinTrend)
> boxplotSegments(x, seglen = 8)

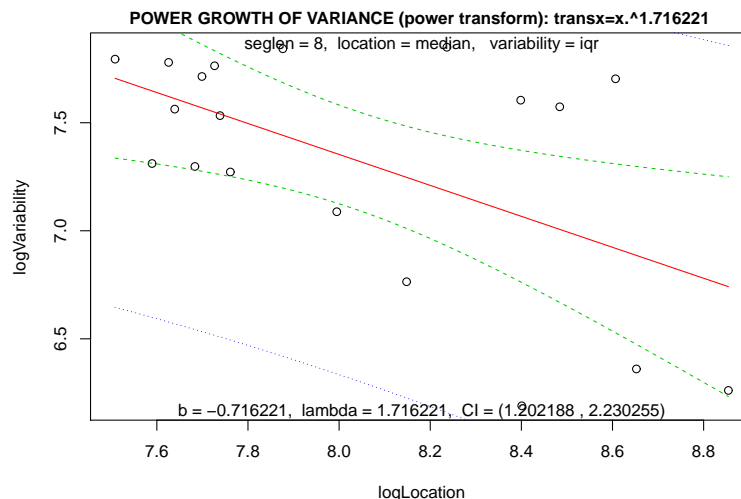
```



Obrázek 14: `boxplotSegments(seglen=8)` pro rezidua po lineárním trendu u dat „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

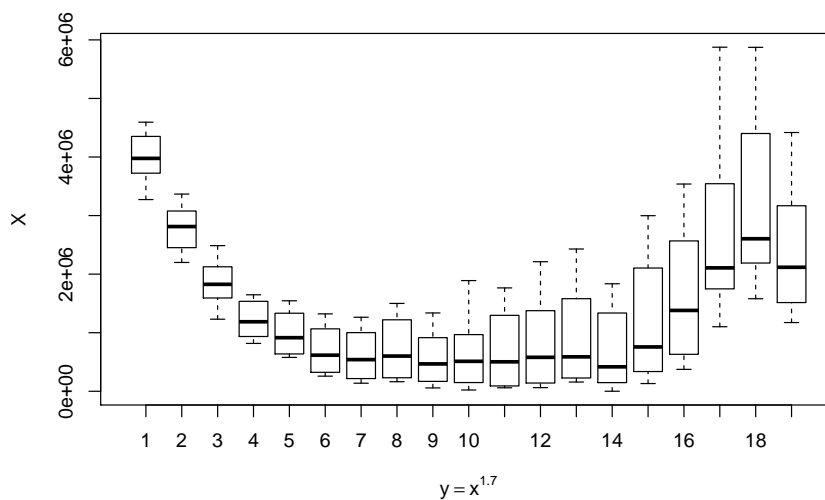
Vyzkoušíme mocninnou transformaci (i když to nejspíše nebude mít smysl) a pomocí funkce `boxplotSegments()` si prohlédneme variabilitu pro transformovaná data.

```
> x <- resid(LinTrend) + coef(LinTrend)[1]
> outp <- powtr(x, seglen = 8, figure = TRUE, location = "median",
  variability = "iqr")
```



Obrázek 15: `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro rezidua po lineárním trendu u dat „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

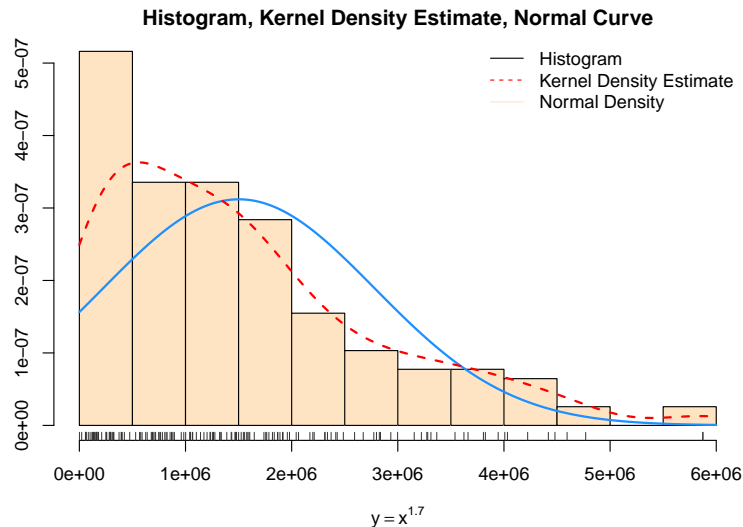
```
> x <- outp$transfx
> boxplotSegments(x, seglen = 8, xlab = expression(y == x^1.7))
```



Obrázek 16: `boxplotSegments` (volba `seglen=8`) pro transformovaná data ( $Y = X^{1.7}$ ) „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

Variabilita se rozhodně nezlepšila. Podívejme se na normalitu pomocí příkazu `HistFit()`.

```
> HistFit(outp$transfx, xlab = expression(y == x^1.7))
```



Obrázek 17: Testování normality pro transformovaná data ( $Y = X^{1.7}$ ) „Australská produkce elektriny (březen 1956 - září 1994) - čtvrtletní údaje“

Tento postup se rozhodně neosvědčil, protože byl nekorektní. Klasický regresní model předpokládá homoskedastická rezidua, což evidentně nebylo splněno.

Naštěstí existují postupy, které v jednom kroku hledají v regresním modelu všechny neznámé parametry. V prostředí R balíček `car` nabízí funkci `powerTransform()`, která hledá parametr  $\lambda$  pro Box–Coxovu transformaci.

```
> library(car)
> TT <- time(AelecTS)
> data <- data.frame(TIME = TT - mean(TT), X = Aelec)
> transf1 <- powerTransform(X ~ TIME, data = data)
> summary(transf1)
```

bcPower Transformation to Normality

|    | Est.Power | Std.Err. | Wald Lower Bound | Wald Upper Bound |
|----|-----------|----------|------------------|------------------|
| Y1 | 0.5079    | 0.0294   | 0.4504           | 0.5654           |

Likelihood ratio tests about transformation parameters

|                       | LRT      | df | pval |
|-----------------------|----------|----|------|
| LR test, lambda = (0) | 176.5429 | 1  | 0    |
| LR test, lambda = (1) | 162.1765 | 1  | 0    |

```
> str(transf1)
```

List of 13

```
$ value      : num 1094
$ counts     : Named int [1:2] 3 3
..- attr(*, "names")= chr [1:2] "function" "gradient"
$ convergence: int 0
$ message    : chr "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
$ hessian    : num [1, 1] 1161
$ start      : num 0.508
```

```

$ lambda      : Named num 0.508
..- attr(*, "names")= chr "Y1"
$ roundlam    : Named num 0.5
..- attr(*, "names")= chr "Y1"
$ family      : chr "bcPower"
$ xqr         :List of 4
..$ qr        : num [1:155, 1:2] -12.4499 0.0803 0.0803 0.0803 0.0803 ...
.. ..- attr(*, "assign")= int [1:2] 0 1
.. ..- attr(*, "dimnames")=List of 2
.. .. ..$ : chr [1:155] "1" "2" "3" "4" ...
.. .. ..$ : chr [1:2] "(Intercept)" "TIME"
..$ rank      : int 2
..$ qraux     : num [1:2] 1.08 1.13
..$ pivot     : int [1:2] 1 2
..- attr(*, "class")= chr "qr"
$ y           : num [1:155, 1] 3923 4436 4806 4418 4339 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:155] "1" "2" "3" "4" ...
.. ..$ : NULL
$ x           : num [1:155, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:155] "1" "2" "3" "4" ...
.. ..$ : chr [1:2] "(Intercept)" "TIME"
..- attr(*, "assign")= int [1:2] 0 1
$ weights     : num [1:155] 1 1 1 1 1 1 1 1 1 1 ...
- attr(*, "class")= chr "powerTransform"

```

Všimněme si, že `transf$roundlam` nabízí vhodnou volbu parametru  $\lambda$ .

```
> print(transf1$roundlam)
```

```
Y1
0.5
```

Nyní ukážeme trochu jiný, ale ekvivalentní postup.

```
> m1 <- lm(X ~ TIME, data = data)
> summary(m1)
```

Call:

```
lm(formula = X ~ TIME, data = data)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-3561.3 -1640.1  -58.8  1247.4  5188.0
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  20218.51     158.27  127.75 <2e-16 ***
TIME         1036.11      14.15   73.23 <2e-16 ***
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1970 on 153 degrees of freedom

Multiple R-squared: 0.9723, Adjusted R-squared: 0.9721

F-statistic: 5362 on 1 and 153 DF, p-value: < 2.2e-16

```
> outT <- powerTransform(m1)
> summary(outT)
```

```
bcPower Transformation to Normality
```

|    | Est.Power | Std.Err. | Wald Lower Bound | Wald Upper Bound |
|----|-----------|----------|------------------|------------------|
| Y1 | 0.5079    | 0.0294   | 0.4504           | 0.5654           |

```
Likelihood ratio tests about transformation parameters
```

|                       | LRT      | df | pval |
|-----------------------|----------|----|------|
| LR test, lambda = (0) | 176.5429 | 1  | 0    |
| LR test, lambda = (1) | 162.1765 | 1  | 0    |

```
> print(outT$roundlam)
```

```
Y1
0.5
```

```
> m2 <- update(m1, basicPower(outT$y, outT$roundlam) ~ .)
> summary(m2)
```

```
Call:
```

```
lm(formula = basicPower(outT$y, outT$roundlam) ~ TIME, data = data)
```

```
Residuals:
```

| Min     | 1Q      | Median  | 3Q     | Max     |
|---------|---------|---------|--------|---------|
| -9.4596 | -3.7626 | -0.3627 | 3.2606 | 10.2915 |

```
Coefficients:
```

|             | Estimate  | Std. Error | t value | Pr(> t )   |
|-------------|-----------|------------|---------|------------|
| (Intercept) | 135.49208 | 0.36621    | 370.0   | <2e-16 *** |
| TIME        | 3.83463   | 0.03274    | 117.1   | <2e-16 *** |

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.559 on 153 degrees of freedom
```

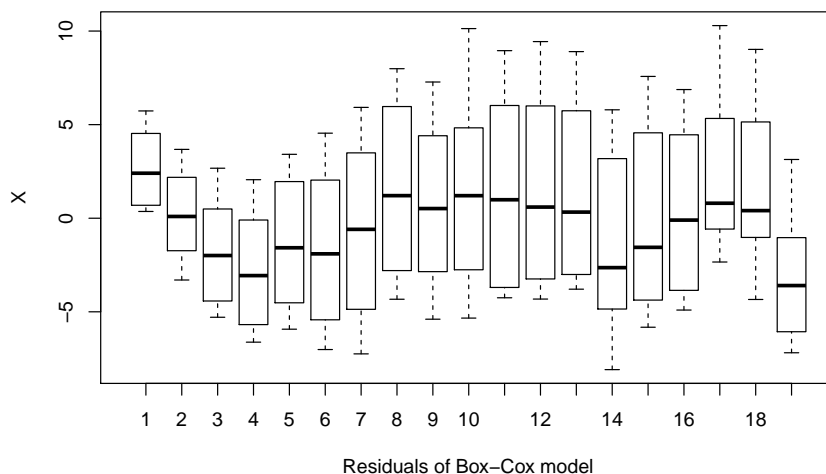
```
Multiple R-squared: 0.989, Adjusted R-squared: 0.9889
```

```
F-statistic: 1.372e+04 on 1 and 153 DF, p-value: < 2.2e-16
```

Na závěr se ještě podíváme, jak dopadla rezidua u transformovaného modelu.

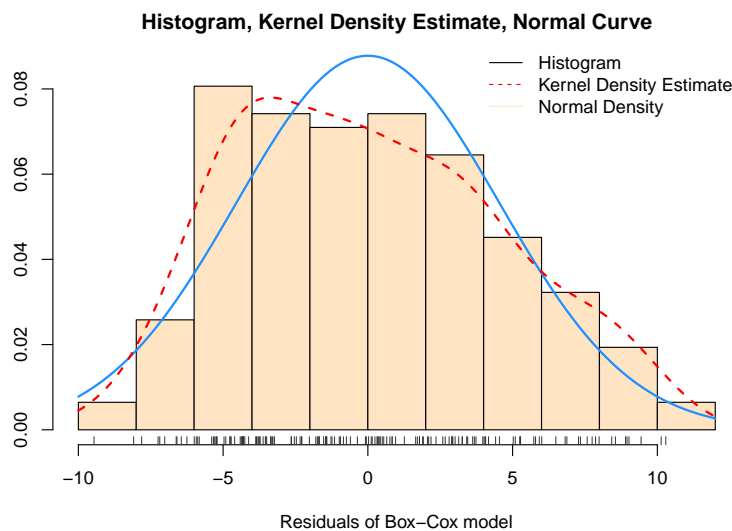
```
> res <- resid(m2)
> boxplotSegments(res, seglen = 8, xlab = "Residuals of Box-Cox model")
```





Obrázek 18: `boxplotSegments` (volba `seglen=8`) rezidua v Box–Coxově modelu pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

```
> HistFit(res, xlab = "Residuals of Box-Cox model")
```



Obrázek 19: Testování normality reziduí v Box–Coxově modelu pro data „Australská produkce elektřiny (březen 1956 - září 1994) - čtvrtletní údaje“

Na závěr můžeme říci, že jsme obdrželi nejlepší výsledek.

## E. Úkol:

Pro časové řady, které jste si našli v úvodním praktiku, zjistěte, zda je třeba data transformovat. V kladném případě zvolte vhodnou transformaci.