M6120 – 10. CVIČENÍ : **M6120cv10** (*Transformace náhodných veličin*)

A. Transformace stabilizující rozptyl

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

$$E_\theta X = \theta.$$

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

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

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

Vzniká otázka, zda lze najít netriviální funkci $\boxed{g}$ tak, aby náhodná veličina $Y = g(X)$ měla **rozptyl nezávisející na** θ. (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** $\boxed{g}$, z Taylorova rozvoje dostaneme aproximaci

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

Potom střední hodnotu lze aproximovat takto

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

a rozptyl

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

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

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

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 $\boxed{g}$ vypočtená podle předchozího vzorce měla výhodný tvar.

Ukázalo se, že takto vypočtená funkce $\boxed{g}$

- nejen **výrazně stabilizuje rozptyl**, takže rozptyl $D_\theta g(X)$ závisí na θ 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.

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

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

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

Chceme najít takovou funkci $\boxed{g}$, aby asymptotické rozdělení náhodné veličiny

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

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**

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

Spočítejme střední hodnotu a rozptyl náhodné veličiny Y:

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

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

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

Binomické rozdělení

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

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

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

Pak pro $n \to \infty$ 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 $\boxed{g}$, aby asymptotické rozdělení náhodné veličiny

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

Pak

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

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

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

Anscombe (1948) opět navrhl stabilnější transformaci:

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

přičemž

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

χ^2 ROZDĚLENÍ

Nechť náhodná veličina X má χ^2 **rozdělení** s parametrem $\nu > 0$, tj. $\boxed{X \sim \chi^2(\nu)}$. Lze spočítat, že $EX = \nu$ a $DX = 2\nu$, tj. $\boxed{\sigma^2(\nu) = 2\nu}$. 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**

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

Spočítejme střední hodnotu a rozptyl náhodné veličiny Y

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

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 θ se střední hodnotou a rozptylem

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

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

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

Položme v dalším

$$\boxed{\lambda = 1 - \vartheta}$$

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
$$c = \sigma \qquad \text{a} \qquad K = \begin{cases} 0 & \lambda = 0 \Rightarrow \vartheta = 1, \\ -\frac{1}{\lambda} = -\frac{1}{1-\theta} & \lambda \neq 0 \Rightarrow \vartheta \neq 1, \end{cases}$$
a odtud
$$\boxed{g(X) = X^{(\lambda)} = \begin{cases} \ln X & \lambda = 0 \ (\vartheta = 1), \\ \frac{X^\lambda - 1}{\lambda} & \lambda \neq 0 \ (\vartheta \neq 1). \end{cases}}.$$
- **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:
$$\boxed{g(X+a) = (X+a)^{(\lambda)} = \begin{cases} \ln(X+a) & \lambda = 0 \ (\vartheta = 1), \\ \frac{(X+a)^\lambda - 1}{\lambda} & \lambda \neq 0 \ (\vartheta \neq 1). \end{cases}}.$$
- **Mocninná transformace se znaménkem** lze opět použít v případě, že náhodné veličiny nejsou kladné:
$$\boxed{g(X) = sign(X)|X|^{(\lambda)} = \begin{cases} sign(X) \ln |X| & \lambda = 0 \ (\vartheta = 1), \\ sign(X) \frac{|X|^\lambda - 1}{\lambda} & \lambda \neq 0 \ (\vartheta \neq 1). \end{cases}}.$$

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
$$\mathbf{Y} = (Y_1 = g(X_1), \ldots, Y_n = g(X_n))'$$
je výběr z normálního rozdělení se střední hodnotou μ a rozptylem σ^2. Označme
$$\mathbf{y} = (y_1, \ldots, y_n)'$$
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 $\frac{\partial l}{\partial \sigma^2} = 0$.

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

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

Upravme nyní logaritmus věrohodnostní funkce takto:

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

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

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

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

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

Nyní toto maximum vyjádřeme 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 vypočtěme 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

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

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

Nyní hledejme maximum funkce $l(\hat{\mu},\hat{\sigma}^2,\lambda) = l(\bar{y},s^2,\lambda)$ pro parametr λ. Protože maximum vzhledem k λ 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 λ, 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 λ:

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

tj. všechna λ splňující nerovnost $l^*(\lambda) \ge D_\alpha$ 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 $\boxed{H_0^1 : \lambda = 1}$. Pokud hypotézu **nezamítneme**, tj. $l^*(1) \ge D_\alpha$, **nemusíme data transformovat**.

(b) Pokud předchozí hypotézu **zamítneme**, můžeme testovat další hypotézu $\boxed{H_0^2 : \lambda = 0}$. Pokud tuto hypotézu **nezamítneme**, tj. $l^*(0) \ge D_\alpha \wedge l^*(1) < D_\alpha$, 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

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

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

Mocninné transformace v prostředí R

V knihovně `car` jsou k dispozici funkce

`bcPower`, `basicPower`	Box-Coxova a Yeo-Johnsonova mocninná transformace
`boxCox`	Box-Coxova transformace pro lineární regresní model
`boxCoxVariable`	výpočet a konstrukce závisle proměnné pro Boxovu-Coxovu transformaci v lineárním regresním modelu
`estimateTransform`, `powerTransform`	nalezení jednorozměrné nebo vícerozměrné mocninné transformace
`plot.powerTransform`	plot pro objekt typu `powerTransform`
`testTransform`	test podílem věrohodností pro jednorozměrnou či vícerozměrnou mocninnou transformaci

Příklad 1

Datový soubor v prvním a druhém řádku obsahuje popis, od čtvrtého řádku jsou samotná data.

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

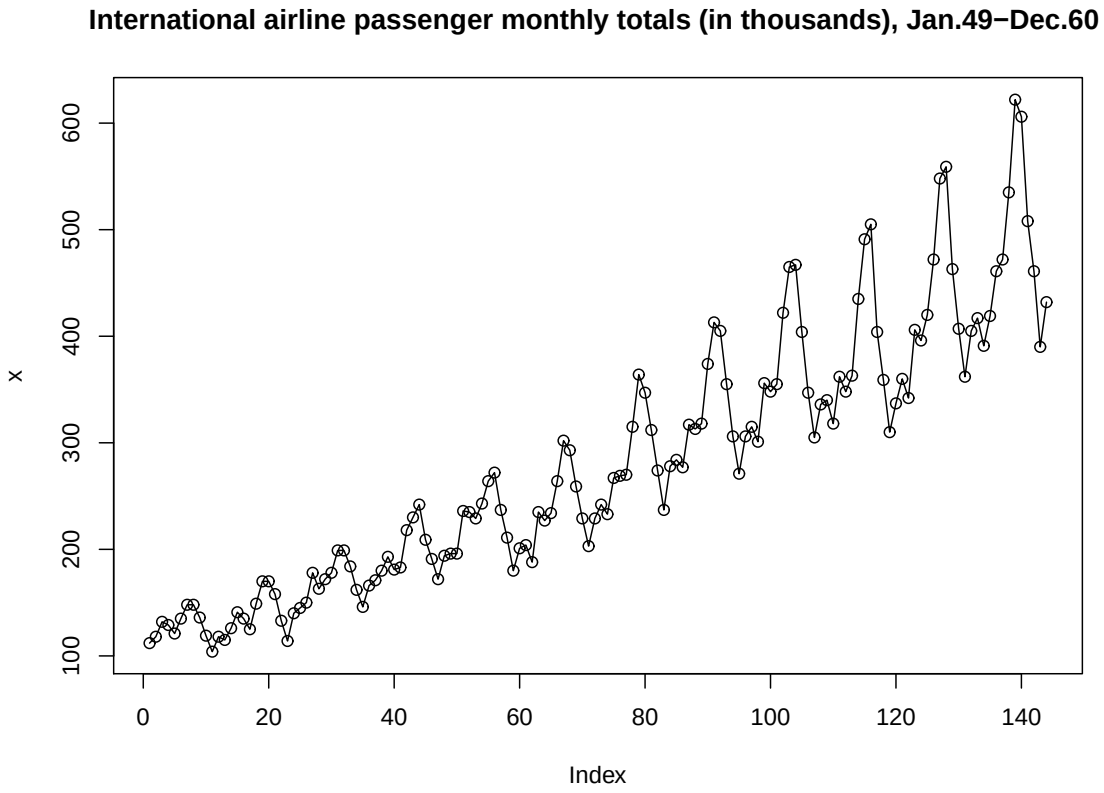
```
[1] "BD Example 1.1.6."
[2] "International airline passenger monthly totals (in thousands), Jan.49-Dec.60 "
```

```
> close(con)
> x <- scan(fileDat, skip = 3)
> str(x)
```

```
 num [1:144] 112 118 132 129 121 135 148 148 136 119 ...
```

Načtená data vykreslíme.

```
> TXT <- POPIS[2]
> plot(x, type = "o", main = TXT)
```



Obrázek 1: Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje

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

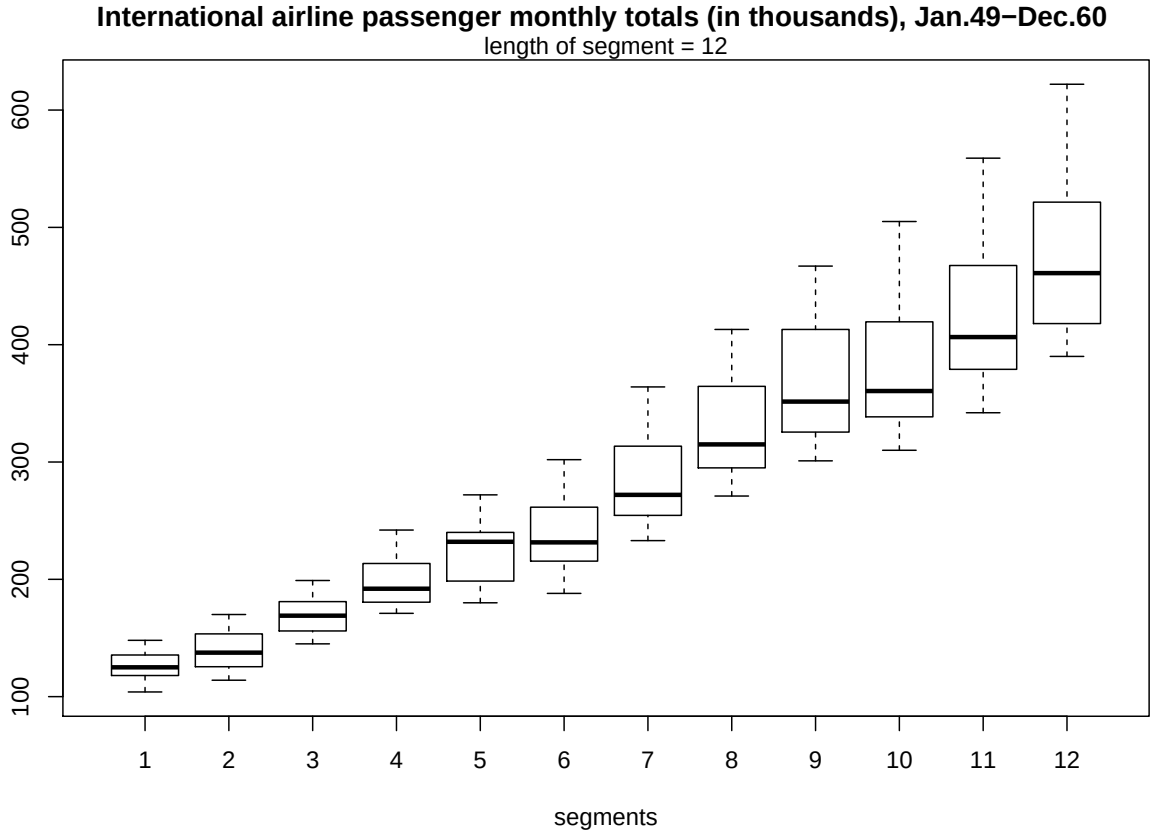
Abychom ještě lépe demonstrovali nestacionaritu v rozptylu vytvoříme nejprve velmi užitečnou funkci `boxplotSegments()`, která pro segmenty dat délky `seglen` vykreslí krabicové grafy.

```
> boxplotSegments <- function(x, seglen = 8, shift = 0, appendL = 0, appendH = 0,
    ...) {
    if (min(dim(as.matrix(x))) > 1)
        stop("x is not a vector")
    x <- as.vector(x)
    n <- length(x)
    if (seglen < 4) {
        warning("segment length too short: 4 will be used")
        seglen <- 4
    }
    nseg <- floor((n - shift)/seglen)
    ns <- nseg * seglen
    n0 <- n - ns - shift
    idStart <- shift + 1
    idEnd <- n - n0
    grL <- NULL
    grH <- NULL
    igr <- 1
    if (shift > 0 & appendL == 2) {
        idStart <- 1
        nseg <- nseg + 1
        igr <- 2
        grL <- rep(1, shift)
    }
    if (shift > 0 & appendL == 1) {
        idStart <- 1
        grL <- rep(1, shift)
    }
    if (n0 > 0 & appendH == 2) {
        idEnd <- n
        grH <- rep(nseg + 1, n0)
    }
    if (n0 > 0 & appendH == 1) {
        idEnd <- n
        grH <- rep(nseg, n0)
    }
    gr <- c(grL, rep(igr:nseg, each = seglen), grH)
    X <- x[idStart:idEnd]
    segments <- factor(gr)
    plot(X ~ segments, ...)
}
```

Ihned nově vytvořenou funkci použijeme. Parametr `seglen` položíme roven 12, protože jde o měsíční data.

Ostatní parametry nemusíme nastavovat, protože naše časová řada začíná v lednu roku 1949 a končí v prosinci 1960, takže první i poslední rok je celý.

```
> seglen = 12
> par(mfrow = c(1, 1), mar = c(4, 2, 3, 0) + 0.05)
> boxplotSegments(x, seglen)
> mtext(paste("length of segment =", seglen))
> title(main = TXT)
```



Obrázek 2: Hetoroskedascita znázorněná pomocí funkce `boxplotSegments()` pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

Z grafu je velmi názorně vidět, jak variabilita se mění se střední hodnotou. Protože o výchozím rozdělení dat nic nevíme, zvolíme mocninnou transformaci. Začneme s jednoduchým přístupem založeným na regresním modelu (funkce `powtr()`), který využívá vztah

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

a neznámé parametry odhaduje **metodou nejmenších čtverců**. Odhadem směrnice regresní přímky získáme parametr ϑ 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`).

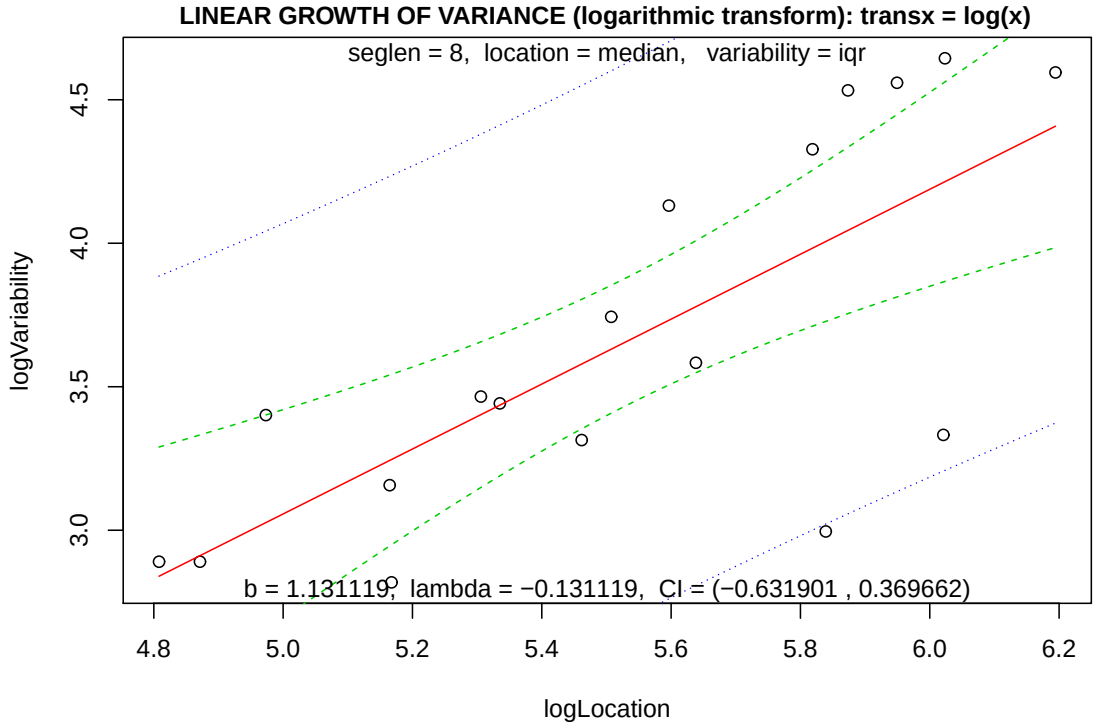
Polohu můžeme odhadnout buď pomocí výběrového průměru (`location="mean"`) nebo výběrového mediánu (`location="median"`).

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.

```
> 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 -0.131
  ..- attr(*, "names")= chr "Estimate"
 $ transfx: num [1:144] 4.72 4.77 4.88 4.86 4.8 ...
 $ txt    : chr "LINEAR GROWTH OF VARIANCE (logarithmic transform): transx = log(x)"
```

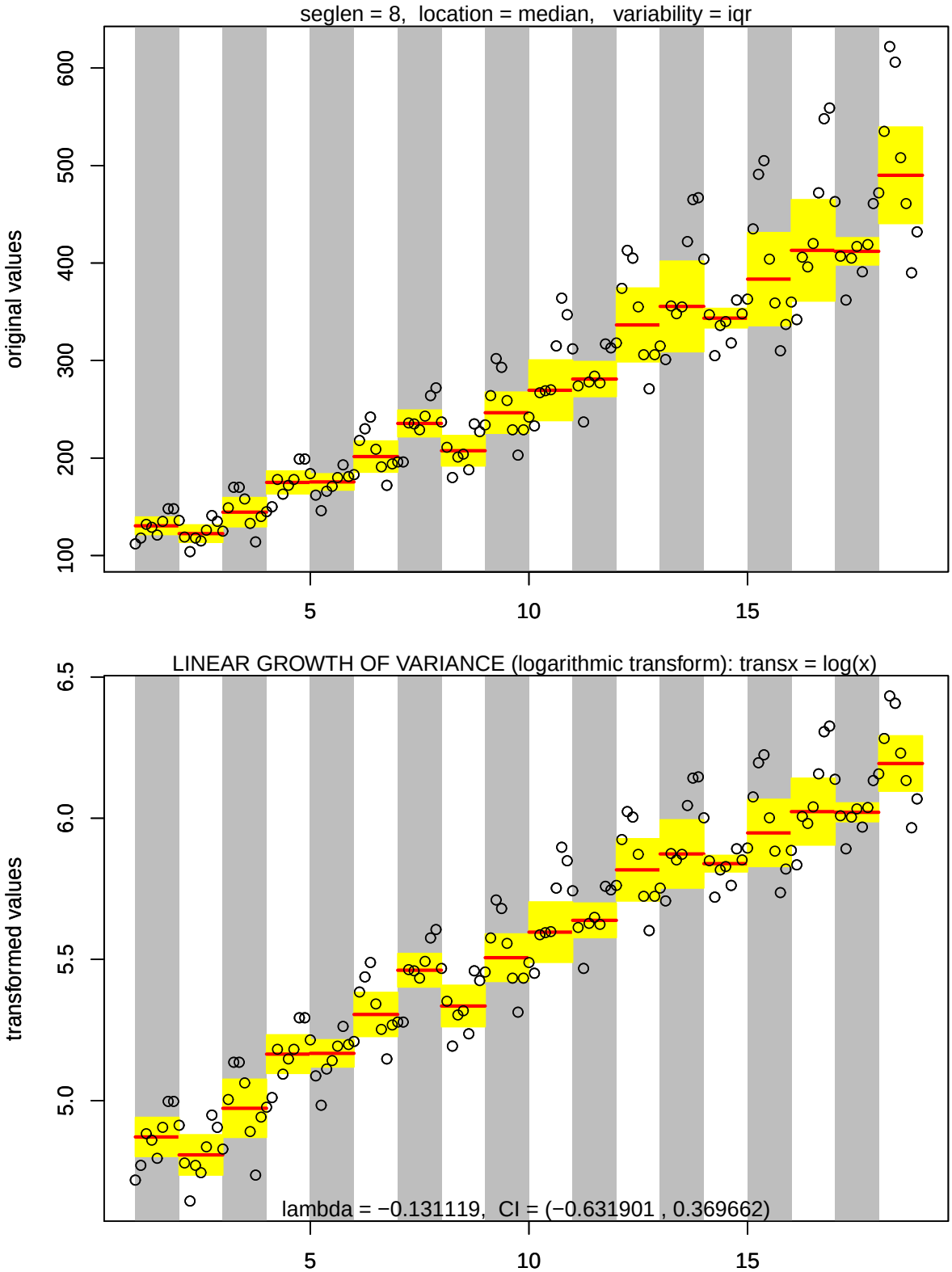


Obrázek 3: Mocninná transformace pomocí funkce `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

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.

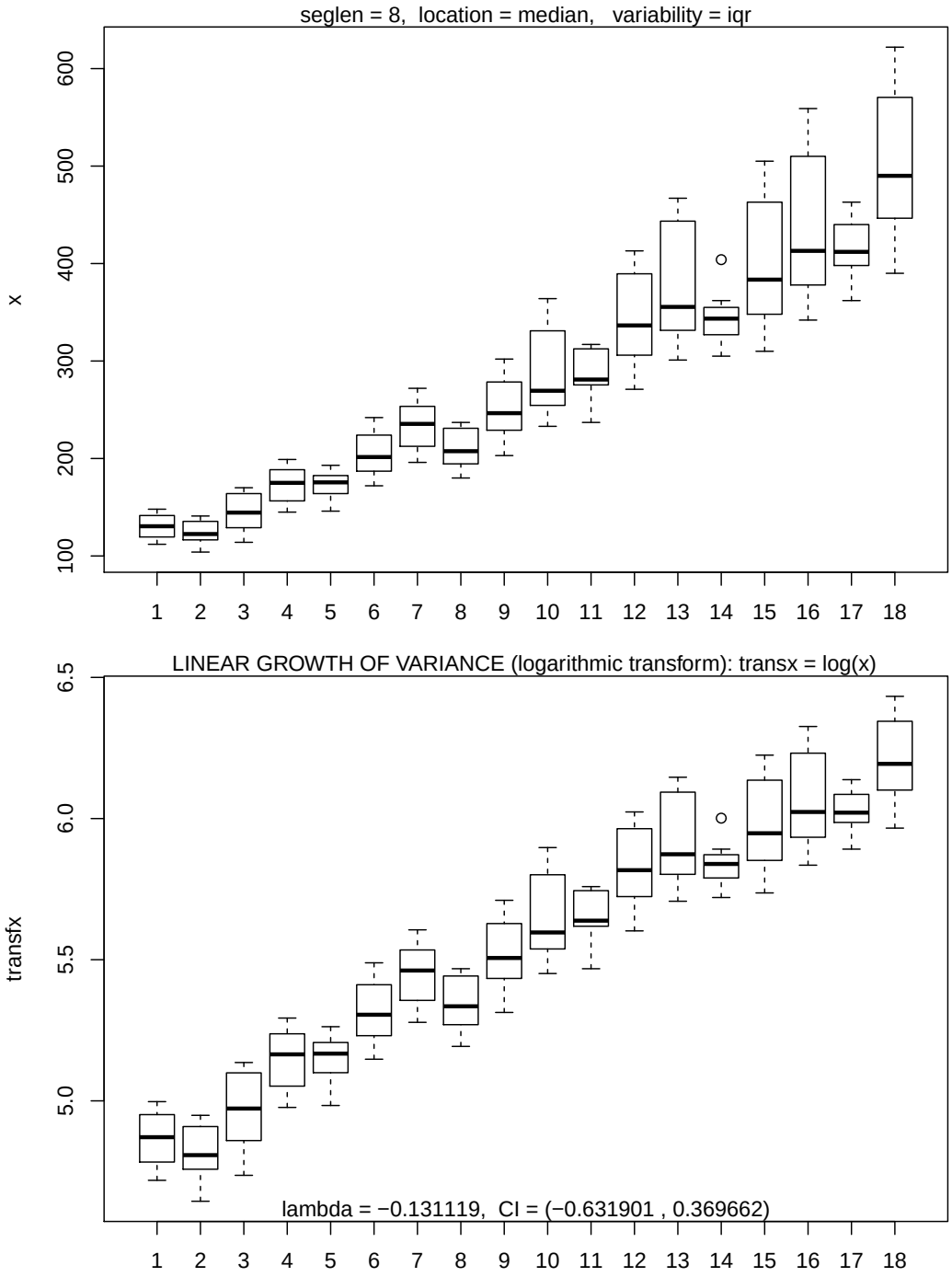
```
> outp <- powtr(x, seglen = seglen, figure2 = TRUE, location = location,
     variability = variability)
```



Obrázek 4: Mocninná transformace pomocí funkce `powtr` – odhady polohy a variability v jednotlivých segmentech (volba `figure2=TRUE`) pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

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

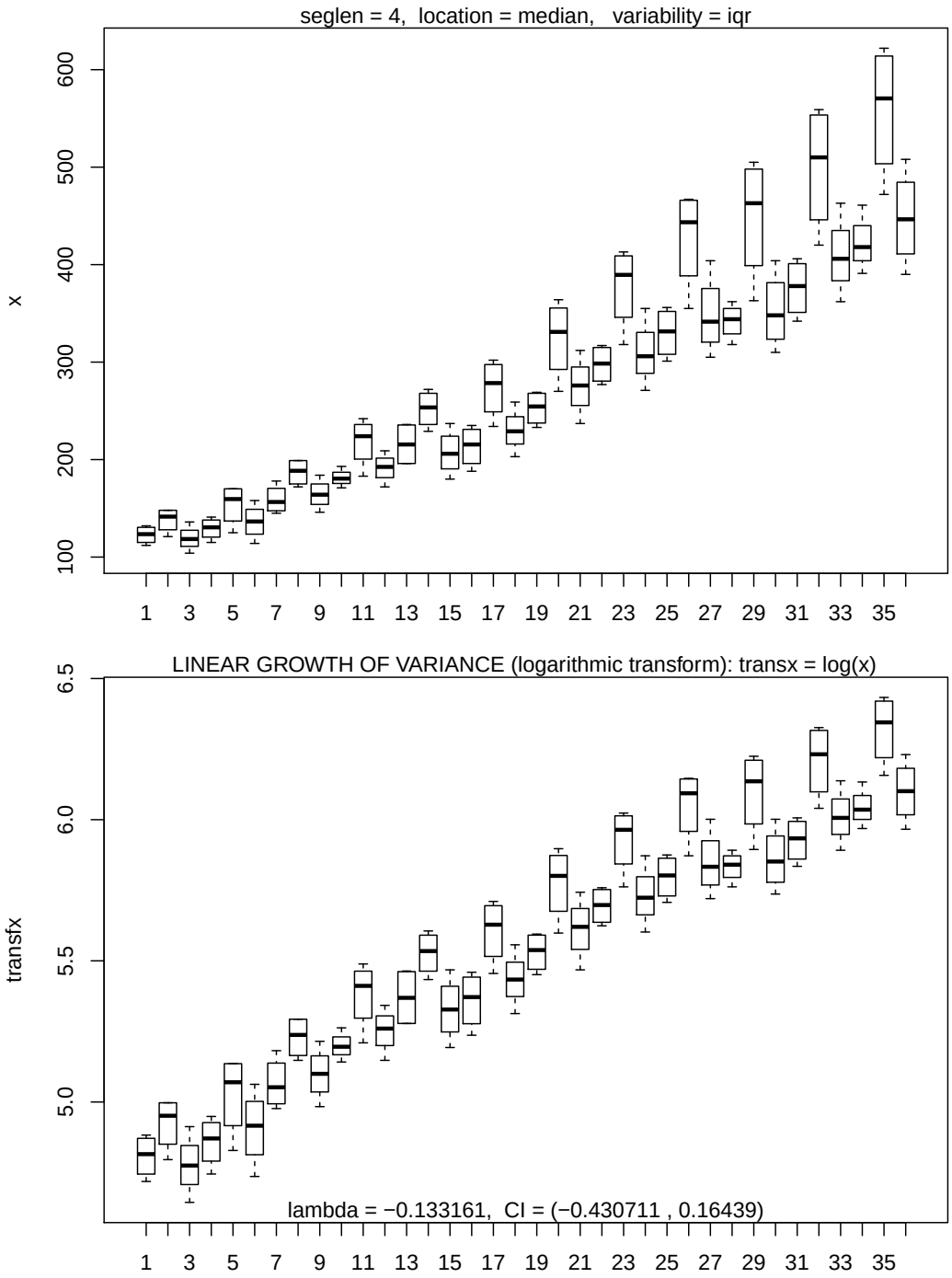
```
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
    variability = variability)
```



Obrázek 5: Mocninná transformace pomocí funkce `powtr` – krabicové grafy v jednotlivých segmentech (volba `figure3=TRUE`) pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

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

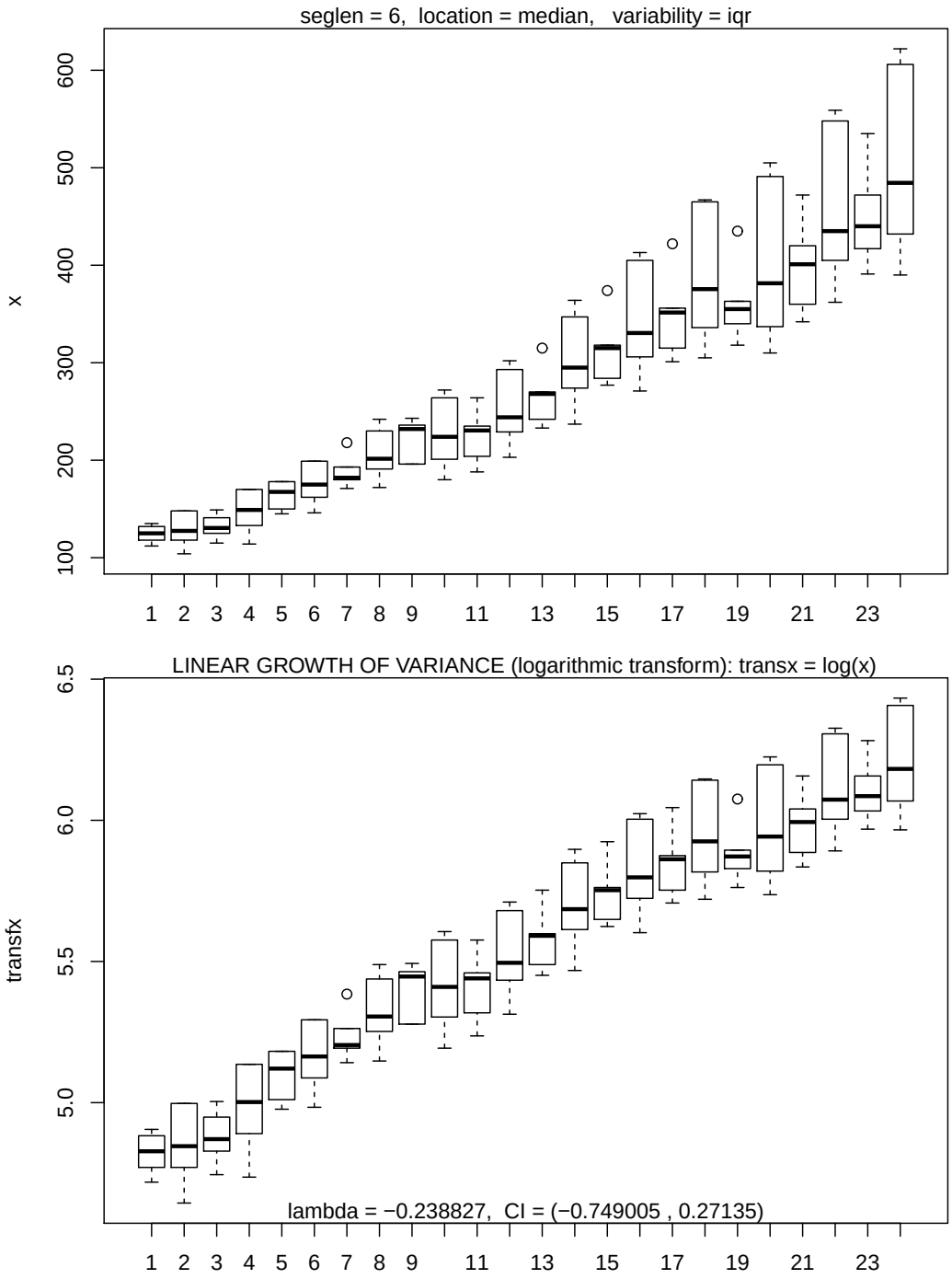
```
> seglen = 4
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
    variability = variability)
```



Obrázek 6: `powtr` – krabicové grafy (`seglen=4` a `figure3=TRUE`) pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

Totéž znovu zopakujeme pro volbu `seglen=6`.

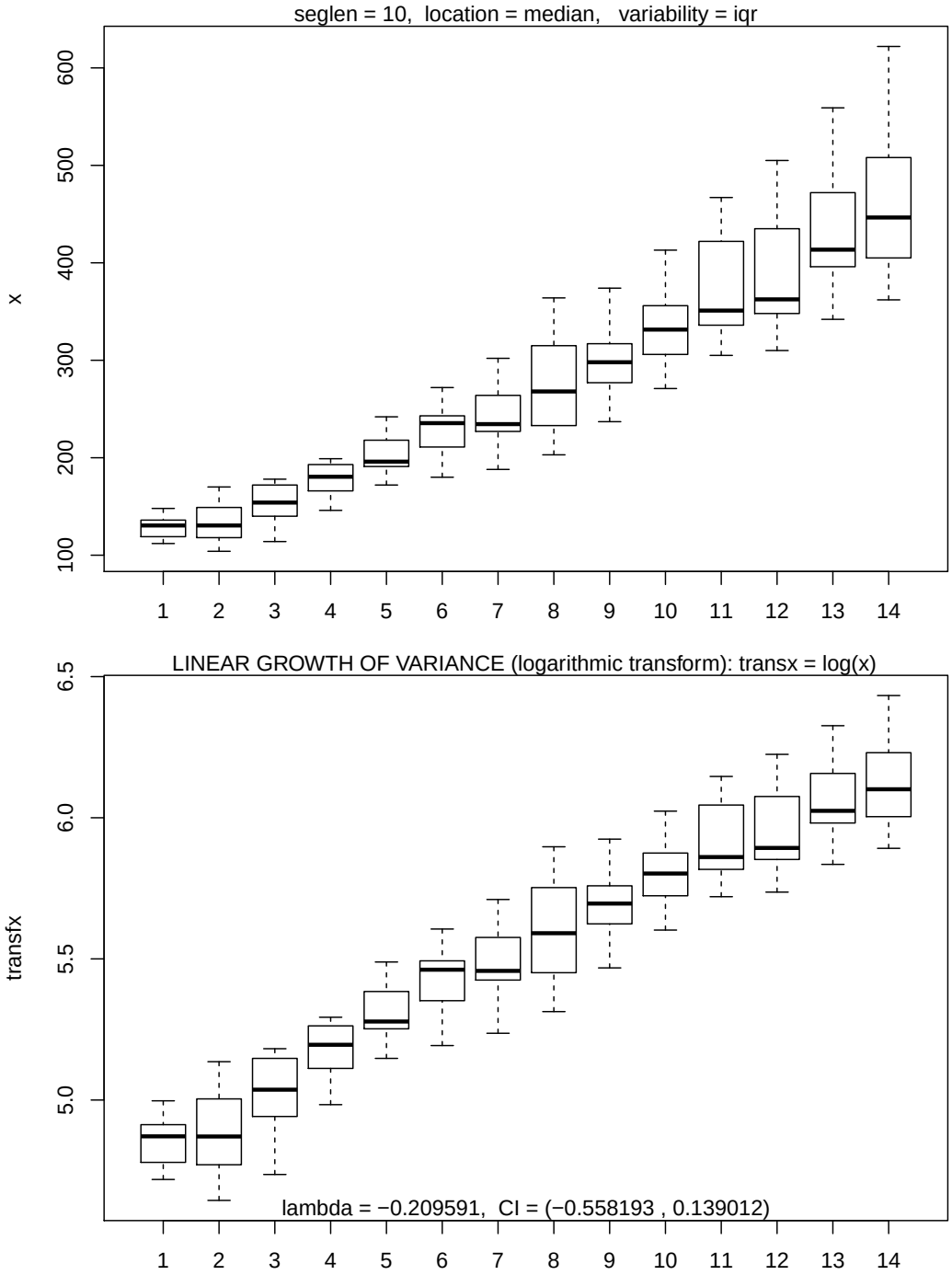
```
> seglen = 6
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
     variability = variability)
```



Obrázek 7: `powtr` – krabicové grafy (`seglen=6` a `figure3=TRUE`) pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

Další volba `seglen=10`.

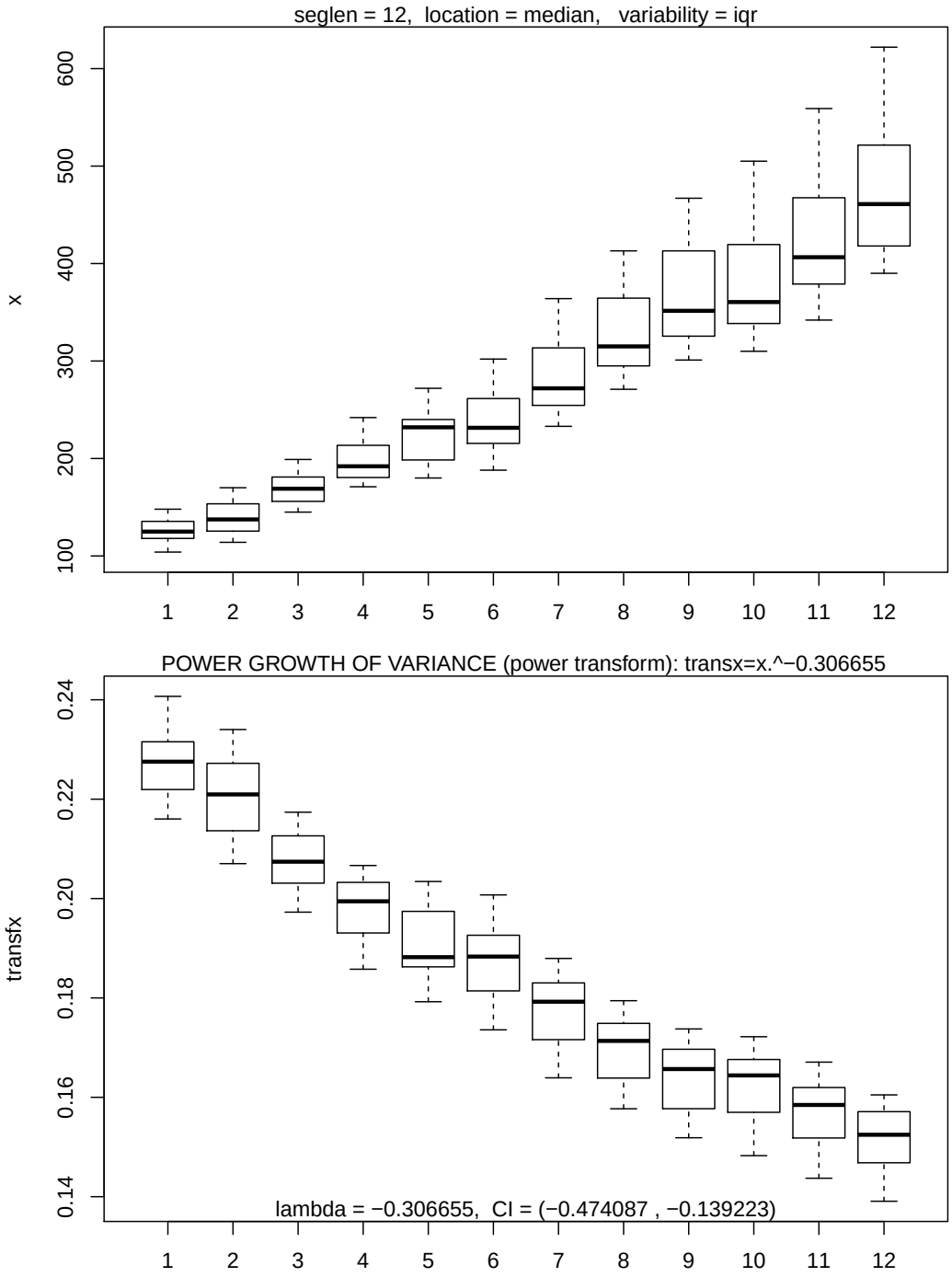
```
> seglen = 10
> outp <- powtr(x, seglen = seglen, figure3 = TRUE, location = location,
      variability = variability)
```



Obrázek 8: `powtr` – krabicové grafy (volba `seglen=10` a `figure3=TRUE`) pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

Poslední volba `seglen=12`.

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



Obrázek 9: `powtr` – krabicové grafy (volba `seglen=12` a `figure3=TRUE`) pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

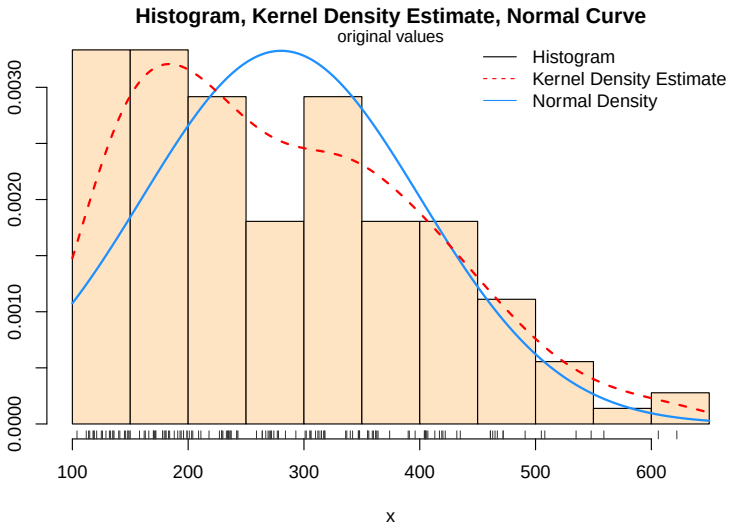
Výsledky přechozích kroků shrňme do tabulky.

`seglen`	dolní mez	odhad $\hat{\lambda}$	horní mez
4	-0.431	-0.133	0.164
6	-0.749	-0.239	0.271
8	-0.632	-0.131	0.370
10	-0.559	-0.210	0.139
12	-0.474	-0.3037	-0.139

Protože kromě jediného případu (pro `seglen=12`) interval spolehlivosti pro λ obsahuje nulu, rozhodneme se pro logaritmickou transformaci dat.

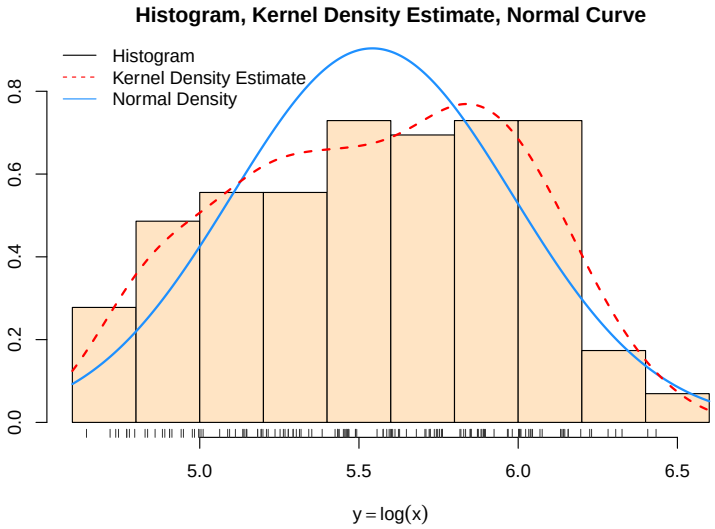
Postupně vykreslíme histogram (spolu s jádrovým odhadem i normální hustotou) pomocí funkce `HistFit()` pro netransformovaná data, následně pro logaritmovaná data.

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



Obrázek 10: Testování normality pro netrasformovaná data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

```
> HistFit(log(x), xlab = expression(y == log(x)))
```



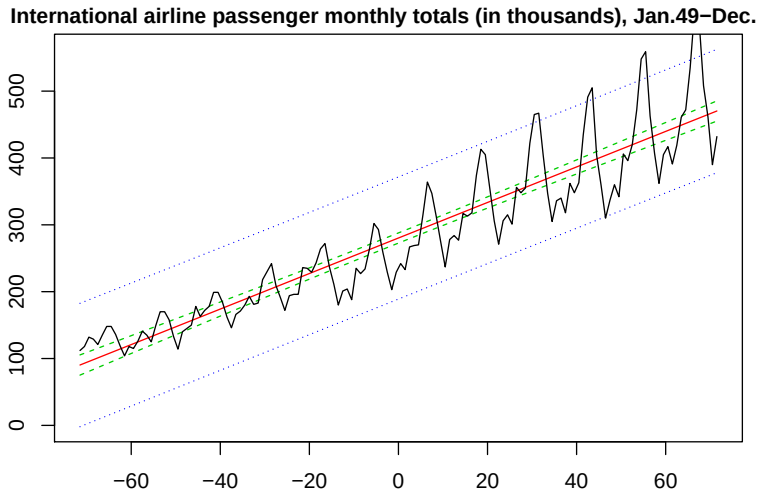
Obrázek 11: Testování normality pro trasformovaná data ($Y = \log(X)$) *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

I když jsme provedli transformaci, přesto nás výsledky grafické kontroly normality neuspokojily.

Z grafů 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 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.

```
> n <- length(x)
> Time <- 1:n
> CenterTime <- Time - mean(Time)
> nn <- 300
> data <- data.frame(CenterTime, x)
> LinTrend <- lm(x ~ CenterTime, data = data)
> print(summary(LinTrend))


Call:
lm(formula = x ~ CenterTime, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-93.858 -30.727  -5.757  24.489 164.999

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 280.29861    3.83810   73.03   <2e-16 ***
CenterTime    2.65718    0.09233   28.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.06 on 142 degrees of freedom
Multiple R-squared: 0.8536,        Adjusted R-squared: 0.8526
F-statistic: 828.2 on 1 and 142 DF,  p-value: < 2.2e-16
```

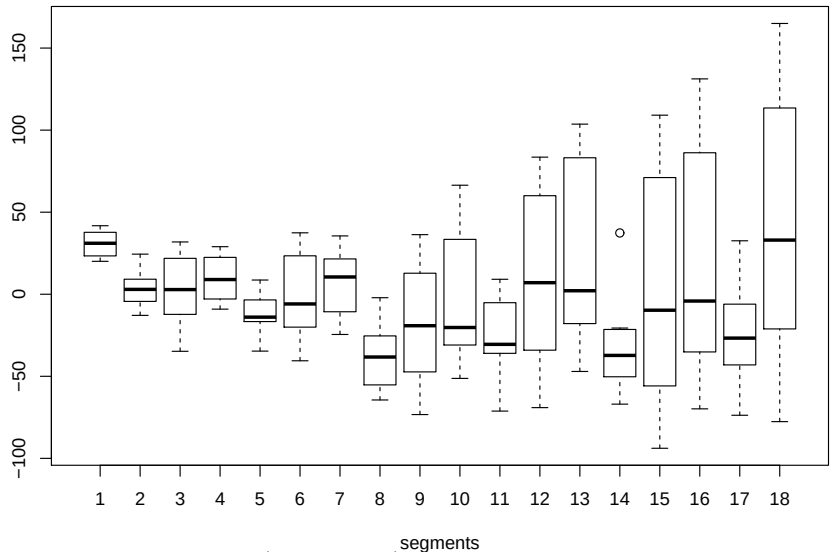
```
> new <- data.frame(CenterTime = seq(CenterTime[1], CenterTime[n], length.out = nn))
> pred.w.plim <- predict(LinTrend, new, interval = "prediction")
> pred.w.clim <- predict(LinTrend, new, interval = "confidence")
> par(mar = c(2, 2, 1, 0) + 0.5)
> matplot(new$CenterTime, cbind(pred.w.clim, pred.w.plim[, -1]), col = c(2,
    3, 3, 4, 4), lty = c(1, 2, 2, 3, 3), type = "l", ylab = "predicted y")
> lines(CenterTime, x)
> title(main = TXT, cex.main = 1)
```



Obrázek 12: Lineární trend pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

Dříve než na rezidua použijeme mocninnou transformaci, podívejme se pomocí funkce `boxplotSegments()`, zda to vůbec nutné.

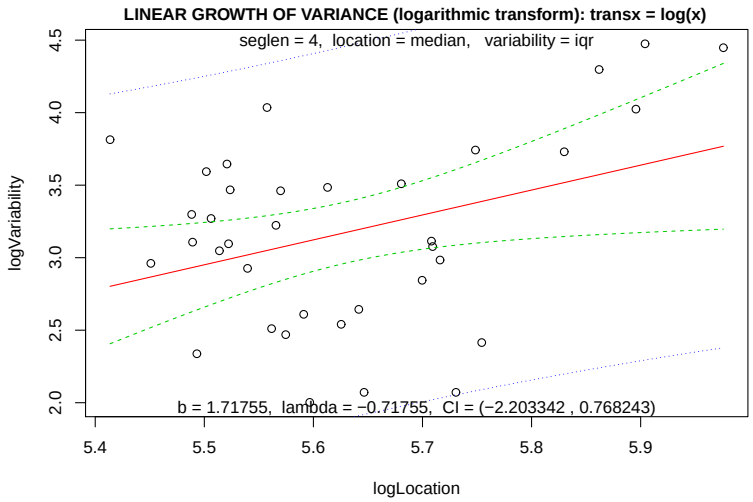
```
> res <- resid(LinTrend)
> par(mfrow = c(1, 1), mar = c(4, 2, 0, 0) + 0.05)
> boxplotSegments(res, seglen = 8)
```



Obrázek 13: `boxplotSegments` (`seglen=8`) pro rezidua po lineárním trendu u dat *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

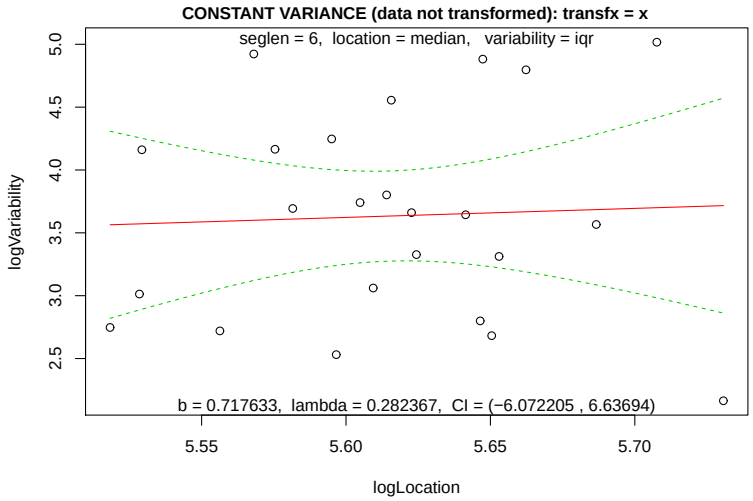
Vidíme, že variabilita velmi kolísá a nejspíše bude problém najít vhodnou transformaci. Přesto to vyzkoušíme pro různé hodnoty parametru `seglen`. Protože pro mocninnou transformaci potřebujeme nezáporné hodnoty, k reziduím přičteme odhadnutý konstantní člen.

```
> res <- resid(LinTrend) + coef(LinTrend)[1]
> seglen = 4
> outp <- powtr(res, seglen = seglen, figure = TRUE, location = "median",
    variability = "iqr")
```



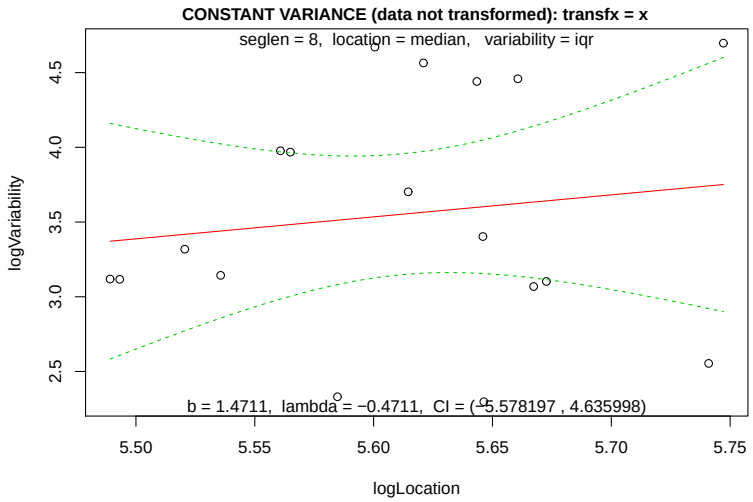
Obrázek 14: `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro rezidua po lineárním trendu u dat *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

```
> seglen = 6
> outp <- powtr(res, seglen = seglen, figure = TRUE, location = "median",
    variability = "iqr")
```



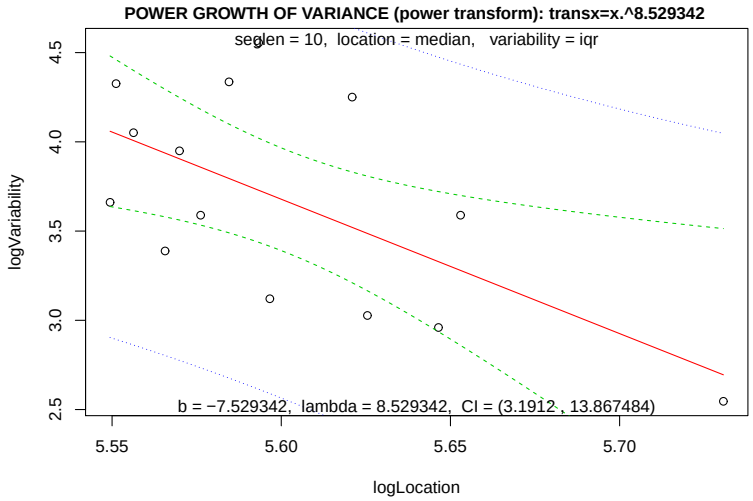
Obrázek 15: `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro rezidua po lineárním trendu u dat *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

```
> seglen = 8
> outp <- powtr(res, seglen = seglen, figure = TRUE, location = "median",
      variability = "iqr")
```



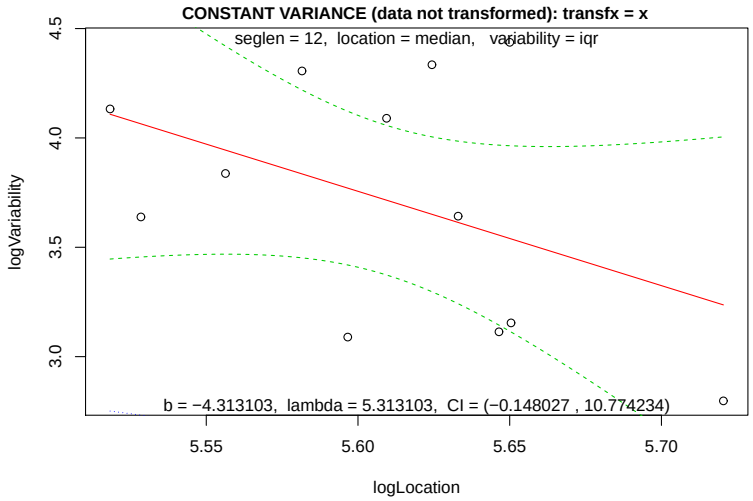
Obrázek 16: `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro rezidua po lineárním trendu u dat *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

```
> seglen = 10
> outp <- powtr(res, seglen = seglen, figure = TRUE, location = "median",
      variability = "iqr")
```



Obrázek 17: `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro rezidua po lineárním trendu u dat *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

```
> seglen = 12
> outp <- powtr(res, seglen = seglen, figure = TRUE, location = "median",
      variability = "iqr")
```



Obrázek 18: `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro rezidua po lineárním trendu u dat *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

Vidíme, jak dostáváme rozporuplné výsledky. Tento postup se rozhodně neosvědčil, protože byl nekorektní. Klasický regresní model předpokládal 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 λ pro Box–Coxovu transformaci.

```
> library(car)
> data <- data.frame(TIME = CenterTime, X = x)
> transf1 <- powerTransform(X ~ TIME, data = data)
> summary(transf1)


bcPower Transformation to Normality

   Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Y1    0.0529   0.0997          -0.1425           0.2482

Likelihood ratio tests about transformation parameters
                            LRT df      pval
LR test, lambda = (0)  0.2812671  1 0.5958719
LR test, lambda = (1) 75.3699483  1 0.0000000


> str(transf1)


List of 13
 $ value      : num 513
 $ 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] 101
```

```
 $ start      : num 0.0529
 $ lambda     : Named num 0.0529
  ..- attr(*, "names")= chr "Y1"
 $ roundlam   : Named num 0
  ..- attr(*, "names")= chr "Y1"
 $ family     : chr "bcPower"
 $ xqr        :List of 4
  ..$ qr   : num [1:144, 1:2] -12 0.0833 0.0833 0.0833 0.0833 ...
  .. ..- attr(*, "assign")= int [1:2] 0 1
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:144] "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 [1:144, 1] 112 118 132 129 121 135 148 148 136 119 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:144] "1" "2" "3" "4" ...
  .. ..$ : NULL
 $ x          : num [1:144, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:144] "1" "2" "3" "4" ...
  .. ..$ : chr [1:2] "(Intercept)" "TIME"
  ..- attr(*, "assign")= int [1:2] 0 1
 $ weights    : num [1:144] 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "class")= chr "powerTransform"
```

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

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

```
Y1
 0
```

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
-93.858 -30.727  -5.757  24.489 164.999

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 280.29861    3.83810   73.03   <2e-16 ***
TIME          2.65718    0.09233   28.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 46.06 on 142 degrees of freedom
Multiple R-squared: 0.8536,        Adjusted R-squared: 0.8526
F-statistic: 828.2 on 1 and 142 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    0.0529   0.0997          -0.1425           0.2482

Likelihood ratio tests about transformation parameters
                            LRT df      pval
LR test, lambda = (0)  0.2812671  1 0.5958719
LR test, lambda = (1) 75.3699483  1 0.0000000
```

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

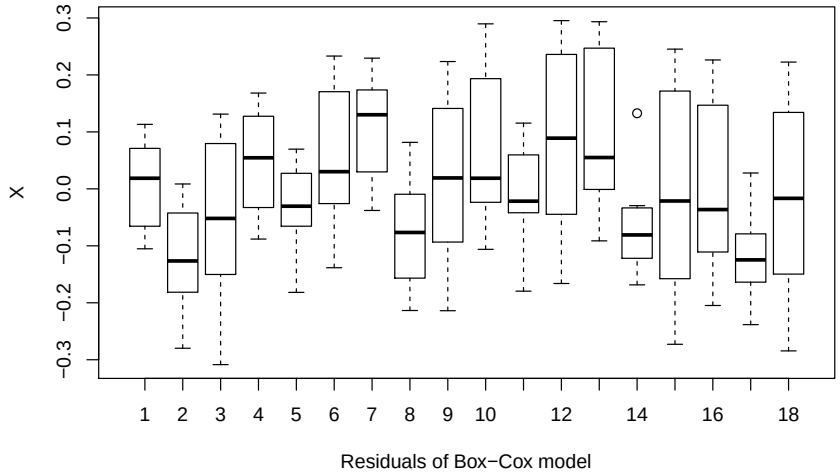
```
Y1
 0
```

```
> 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
-0.30858 -0.10388 -0.01796  0.09738  0.29538

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.5421760  0.0115864  478.33   <2e-16 ***
TIME        0.0100484  0.0002787   36.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.139 on 142 degrees of freedom
Multiple R-squared: 0.9015,        Adjusted R-squared: 0.9008
F-statistic:  1300 on 1 and 142 DF,  p-value: < 2.2e-16
```

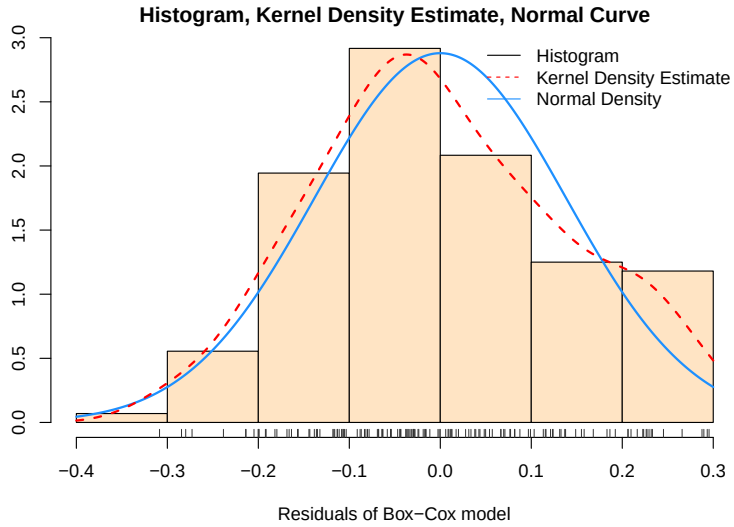
Na závěr se ještě podívejme, jak dopadla rezidua u transformovaného modelu.

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



Obrázek 19: `boxplotSegments` (volba `seglen=8`) rezidua v Box–Coxově modelu pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

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



Obrázek 20: Testování normality reziduí v Box–Coxově modelu pro data *Počty pasažérů (v tisících) na mezinárodní letecké lince (leden 1949 - prosinec 1960) - měsíční údaje*

E. Úkol:

(a) Načtěte soubor informací `srazkyprutok.txt` a dat `srazkyprutok.dat`. Prohledněte si oba soubory.

(b) Proveďte mocninnou transformaci dat.

(c) Ověřte graficky i pomocí testu normalitu transformovaných dat.