M5201 – 5. CVIČENÍ:
Transformace časových řad stabilizující rozptyl

1 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 funkce g v bodě θ dostaneme aproximaci

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

Na základě rozvoje lze střední hodnotu aproximovat takto

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

a rozptyl

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

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.

2 Příklad transformace stabilizující rozptyl – Poissonovo rozdělení

Nechť náhodná veličina X má **Poissonovo** rozdělení s parametrem $\lambda > 0$, tj. $X \sim Po(\lambda)$ s pravděpodobnostní funkcí

$$p_X(x) = P(X = x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} \text{ pro } x = 0, 1, 2, \ldots.$$

Lze spočítat, že $EX = DX = \lambda$, tj. $\boxed{\sigma^2(\lambda) = \lambda}$. 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**

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

Ověřme střední hodnotu a rozptyl náhodné veličiny Y:

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

3 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:

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

Dále budeme pracovat s parametrem

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

a budeme ho nazývat **transformační parametr** 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:

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

3.1 Odhad transformačního parametru mocninné transformace

3.1.1 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 $\frac{\partial l}{\partial \sigma^2} = 0$.

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

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

$$\max_{\mu,\sigma^2,\lambda} L(\mu,\sigma^2) \qquad = \cdots = \left(2\pi s^2(\lambda)\right)^{-\frac{n}{2}} e^{-\frac{n}{2}+(\lambda-1)\sum_{i=1}^{n}\ln x_i}$$

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

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 dále nebudeme zabývat, ale vysvětlíme 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.

Pro tento případ odvodili Box a Cox (1964) asymptotické rozdělení statistiky

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

Na jejím základě lze odvodit **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** $H_1 : \lambda > \lambda_0$:

1. Nejprve budeme testovat hypotézu $\boxed{H_0^1 : \lambda = 1}$. Pokud hypotézu **nezamítneme**, tj. $l^*(1) \ge D_\alpha$, **nemusíme data transformovat**.
2. 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

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

3.1.2 Jednoduchý algoritmus v praktických úlohách – funkce `powtr()`

Další možností, jak odhadnout transformační parametr λ je následující algoritmus:

1. 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**.

2. 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

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

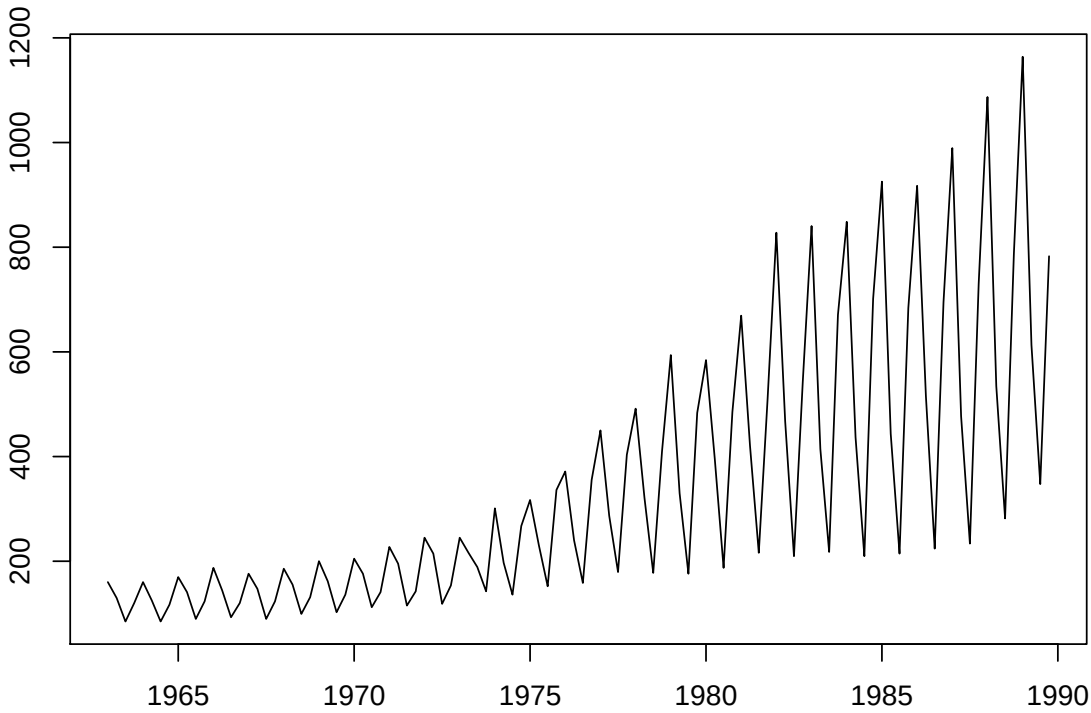
Příklad 1

Datový soubor `UKgas.txt`, jímž se budeme v tomto příkladu zabývat, obsahuje čtvrtletní údaje o množství spotřebovaného zemního plynu ve Velké Británii. Datový soubor obsahuje na prvním řádku popis dat a od druhého řádku následuje jediný sloupec s daty. Nejprve načteme popis a poté do vektoru `UKgas` pomocí funkce `scan` načteme samotná data.

```
> fileDat<-paste(data.library,"UKgas.txt",sep="")
> con<-file(fileDat)
> (popis<-readLines(con,n=1))

[1] "UK Quarterly Gas Consumption"

> close(con)
> UKgas <- scan(fileDat,skip=1)
```

Z vektoru `UKgas` vytvoříme objekt typu časové řady a data vykreslíme.

```
> UKgasTS<-ts(UKgas,start=1963,frequency=4)
> par(mar=c(2,2,1,0)+0.5)
> plot(UKgasTS)
```

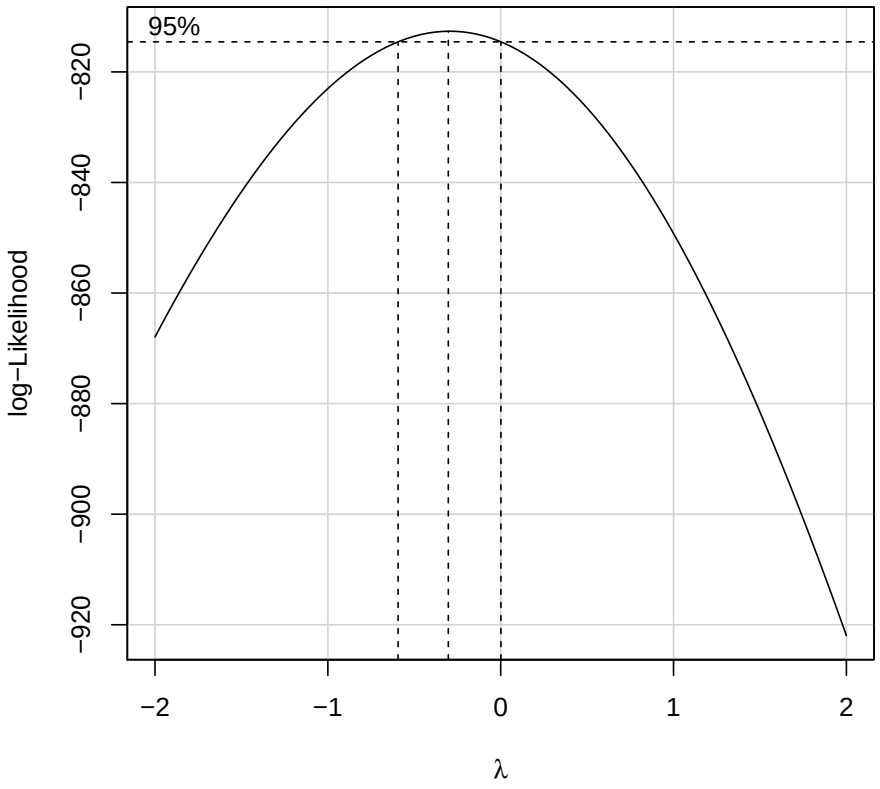


Obrázek 1: Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje

Z grafu je na první pohled patrné, že s rostoucí střední hodnotou dat roste také jejich variabilita, a tudíž bude nutné data transformovat.

K tomu použijeme mocninnou transformaci. Nejprve si můžeme udělat představu o transformačním parametru λ, který bychom získali při odhadu metodou maximální věrohodnosti. Obrázek s věrohodnostní funkcí, vyznačeným odhadem λ metodou maximální věrohodnosti a 95% intervalem spolehlivosti získáme pomocí funkce `boxCox()` z balíčku `car`.

```
> library(car)
> boxCox(UKgasTS ~1)
```



Prostřední svislá přerušovaná čára v obrázku vyznačuje maximálně věrohodný odhad λ, který odpovídá maximu věrohodnostní funkce. Dvě krajní přerušované čáry vymezují 95% interval spolehlivosti pro odhad $\hat{\lambda}$, který má přibližně meze -0.5 a 0. Vidíme, že interval ještě těsně obsahuje nulu, tudíž není vyloučena možnost použití logaritmické transformace dat.

Dále při odhadu parametru $\lambda = 1 - \vartheta$ vyzkoušíme jednoduchý přístup založený na regresním modelu (v R je implementován ve funkci `powtr()`).

Funkce `powtr()` rozdělí nezávisle proměnnou, tj. čas, na subintervaly obsahující takový počet pozorování, jaký určíme parametrem `seglen` (doporučuje se volit číslo mezi 4 až 12).

V každém subintervalu (tj. na základě 4 až 12 pozorování v závislosti na hodnotě parametru `seglen`) se provede odhad polohy a variability (tj. odhad μ a σ). Pomocí parametrů `location` a `variability` specifikujeme, jaký typ statistiky se při tom má použít. Možné volby jsou:

`location="mean"`	odhad polohy pomocí průměru
`location="median"`	odhad polohy pomocí mediánu
`variability="sd"`	odhad variability pomocí směrodatné odchylky
`variability="iqr"`	odhad variability pomocí interkvartilového rozpětí
`variability="range"`	odhad variability pomocí rozdílu mezi maximem a minimem

Funkce `powtr()` 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 odhad parametru ϑ a tím také parametr $\lambda = 1 - \vartheta$.

Výstupem funkce `powtr()` je seznam se třemi položkami, pro nás jsou zajímavé položky `lambda` a `transfx`. Jak už název napovídá, `lambda` je odhad transformačního parametru λ. Druhá položka `transfx` obsahuje vektor s transformovanými daty.

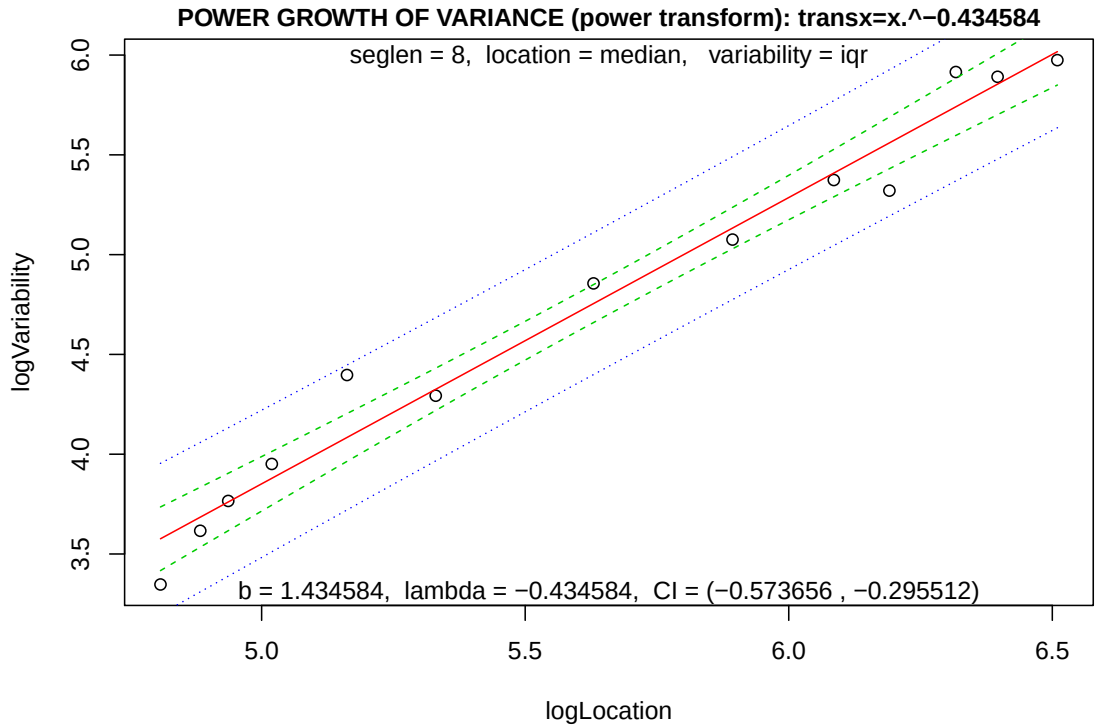
Transformace dat probíhá takto. Při běhu funkce `powtr()` se kromě odhadu λ vypočítají také meze 95%ního intervalu spolehlivosti pro $\hat{\lambda}$. Jestliže interval obsahuje jedničku, data není třeba transformovat a do `transfx` se uloží původní data bez jakékoli změny. Pokud interval obsahuje nulu, přichází v úvahu logaritmická transformace, a proto se do `transfx` uloží zlogaritmovaná data. Ve všech ostatních případech se do `transfx` uloží data umocněná na `lambda`.

V našem příkladu nejprve položíme parametr `seglen=8` a pro odhad polohy a variability zvolíme medián a interkvartilové rozpětí. Pokud navíc ve funkci `powtr()` zadáme volbu `figure=TRUE`, získáme výsledek v grafické podobě.

Protože funkce `powtr()` není součástí R, ale je uložena v souboru `FunkceM5201.R`, musíme nejprve zmíněný soubor načíst.

```
> fileSkript<-paste(data.library,"FunkceM5201.R",sep="")
> source(fileSkript)
> outp<-powtr(UKgas,seglen=8,figure=TRUE,location="median",variability="iqr")
> str(outp)
```

```
List of 3
 $ lambda : Named num -0.435
  ..- attr(*, "names")= chr "Estimate"
 $ transfx: num [1:108] 0.11 0.121 0.145 0.125 0.11 ...
 $ txt    : chr "POWER GROWTH OF VARIANCE (power transform): transx=x.^-0.434584"
```



Obrázek 2: Mocninná transformace dat pomocí funkce `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

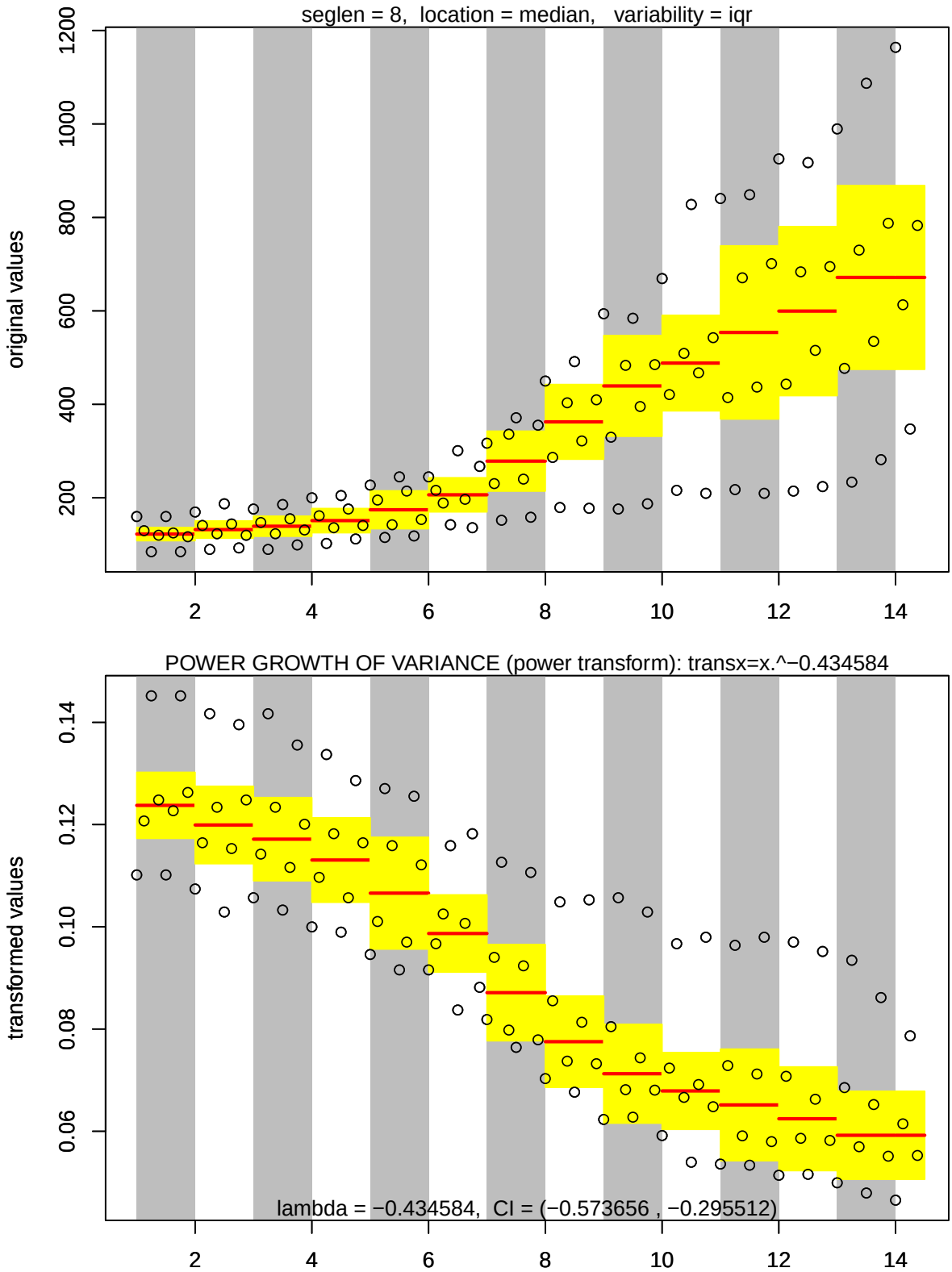
Na obrázku vidíme body, jejichž x-ové souřadnice jsou dány jako logaritmy odhadů polohy a y-ové souřadnice jako logaritmy odhadů variability pro úseky dat o délce 8 pozorování. Body je proložena regresní přímka. Výsledky regrese jsou uvedeny v obrázku dole – odhad směrnice regresní přímky b, odhad transformačního parametru $\lambda = 1 - b$ a 95%ní interval spolehlivosti pro λ.

Protože interval spolehlivosti pro $\hat{\lambda}$ neobsahuje ani jedničku, ani nulu, byla provedena mocninná transformace s odhadnutým parametrem $\hat{\lambda} = -0.434584$.

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

Pokud by nedošlo ke stabilizaci rozptylu ani po transformaci, pak by bylo třeba hledat jiný typ transformace než je mocninná.

```
> outp<-powtr(UKgas,seglen=8,figure2=TRUE,location="median",variability="iqr")
```

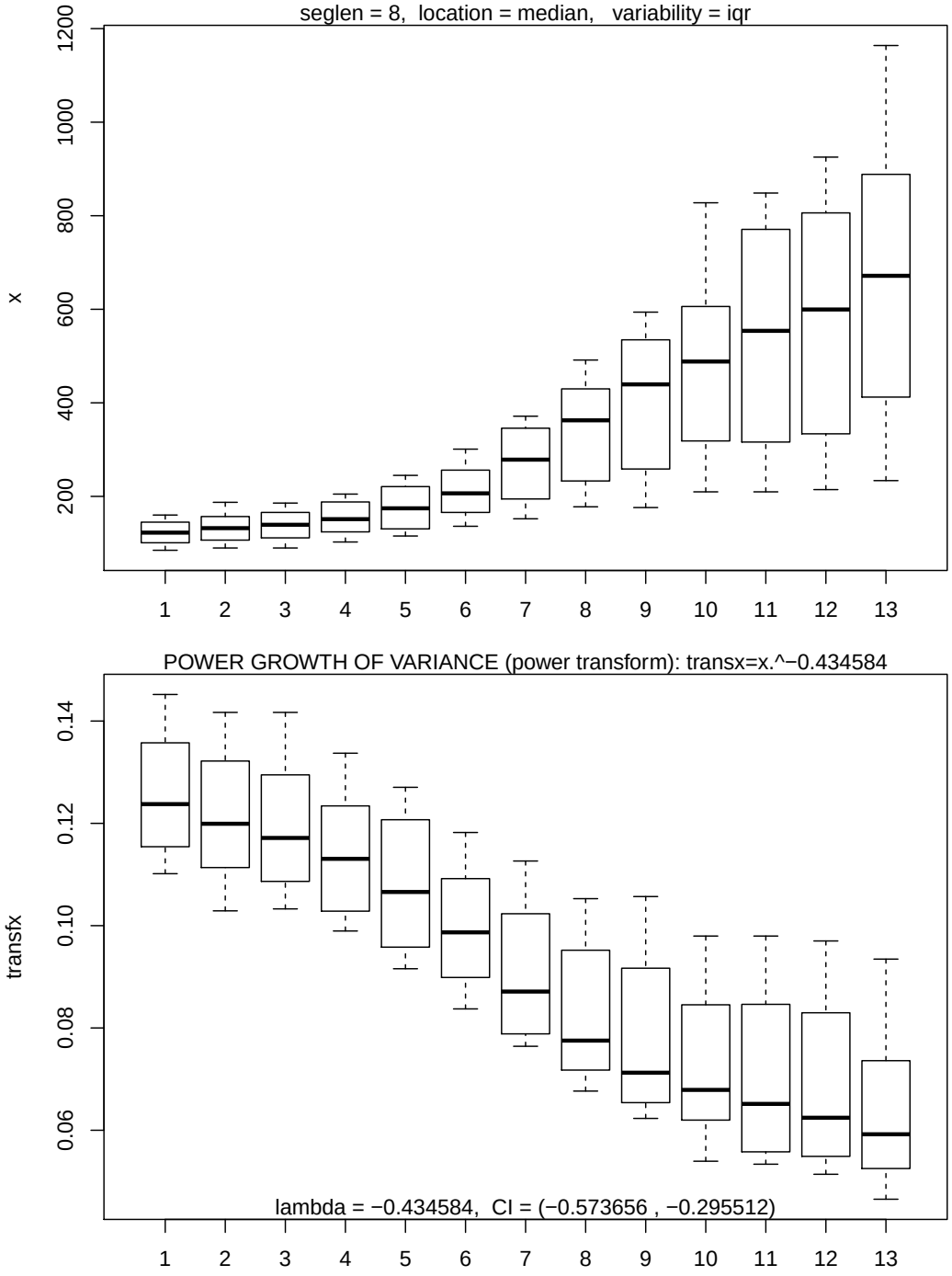


Obrázek 3: Mocninná transformace pomocí funkce `powtr` – odhady polohy a variability v jednotlivých segmentech (volba `figure2=TRUE`) pro data *„Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje“*

Obrázek naznačuje, že variabilita transformovaných dat je nezávislá na střední hodnotě v jednotlivých segmentech a je víceméně konstantní.

Obdobný výstup, pouze vyjádřený pomocí krabicových grafů za jednotlivé segmenty vstupních dat, získáme volbou `figure3=TRUE`.

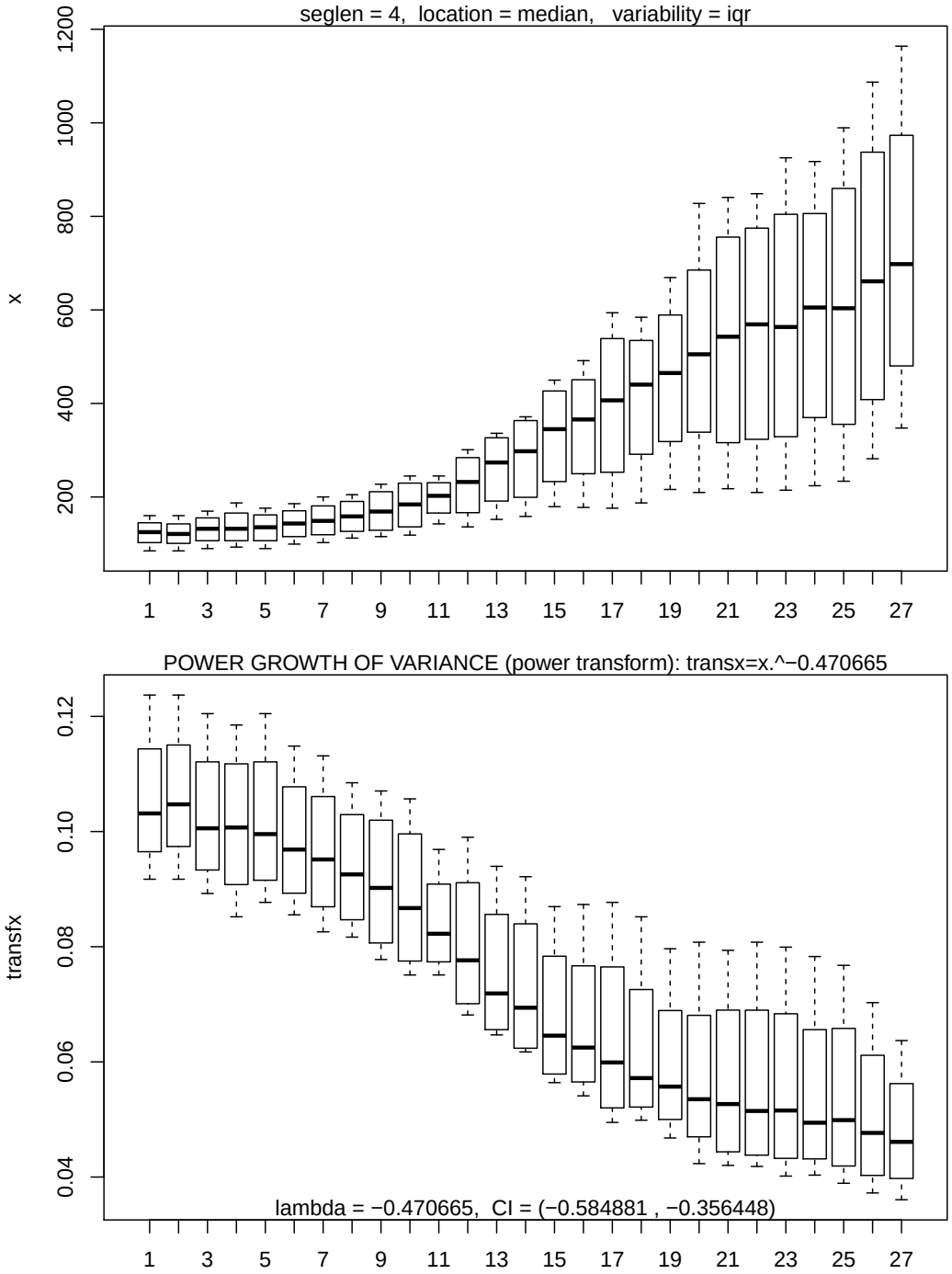
```
> outp<-powtr(UKgas,seglen=8,figure3=TRUE,location="median",variability="iqr")
```



Obrázek 4: Mocninná transformace pomocí funkce `powtr` – krabicové grafy v jednotlivých segmentech (volba `figure3=TRUE`) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Abychom vyzkoušeli robustnost funkce `powtr()` na našich datech (neboli zda je odhad λ citlivý na to, jakou hodnotu `seglen` zvolíme), provedeme odhad mocninné transformace postupně pro `seglen=4,6,10,12`. Ostatní parametry při tom necháme beze změny. K znázornění výsledků vybereme třetí typ grafu.

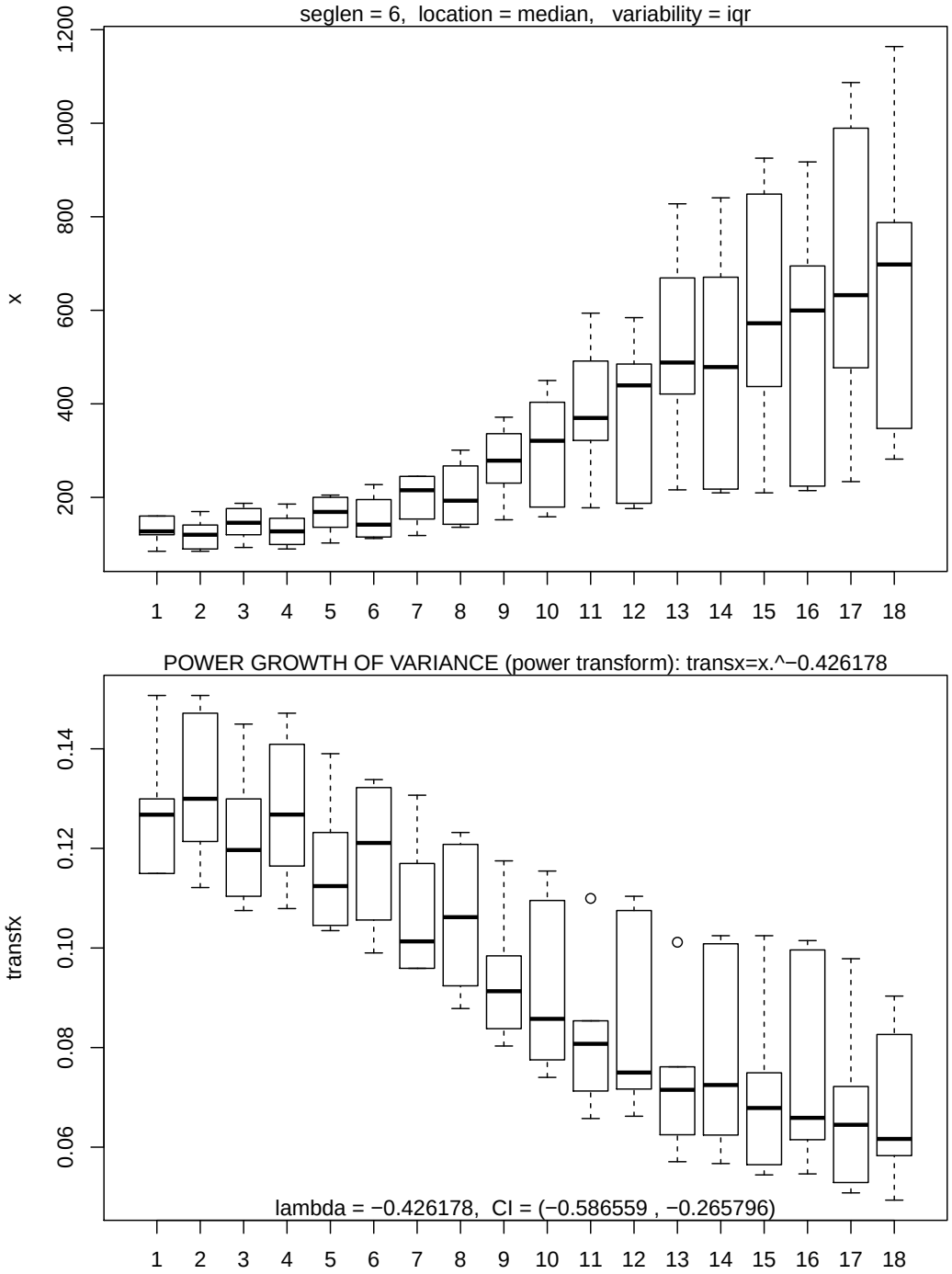
```
> outp<-powtr(UKgas,seglen=4,figure3=TRUE,location="median",variability="iqr")
```



Obrázek 5: `powtr` – krabicové grafy (volba `seglen=4` a `figure3=TRUE`) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Totéž znovu zopakujeme pro volbu `seglen=6`.

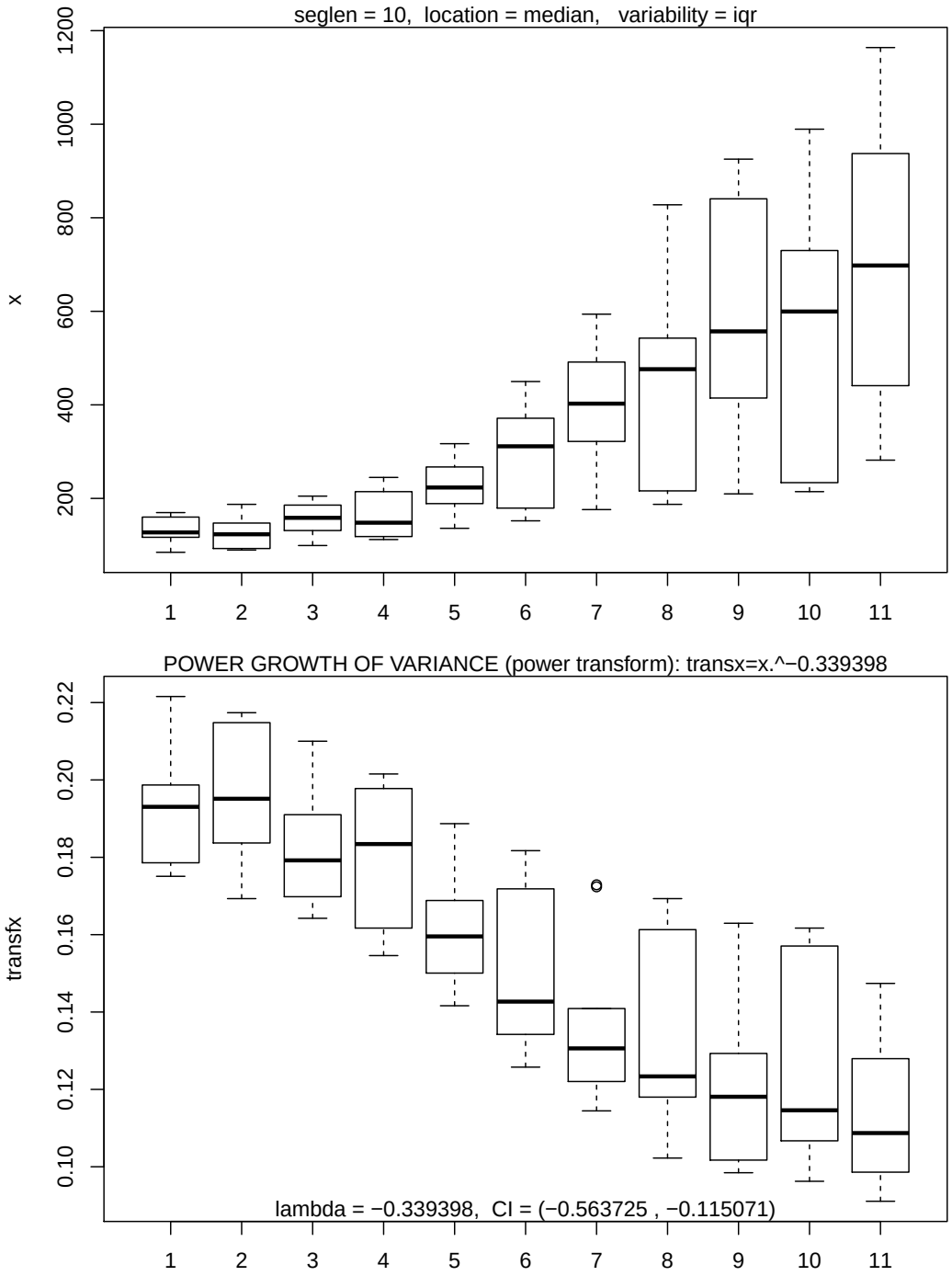
```
> outp<-powtr(UKgas,seglen=6,figure3=TRUE,location="median",variability="iqr")
```



Obrázek 6: `powtr` – krabicové grafy (volba `seglen=6` a `figure3=TRUE`) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Další volba je `seglen=10`.

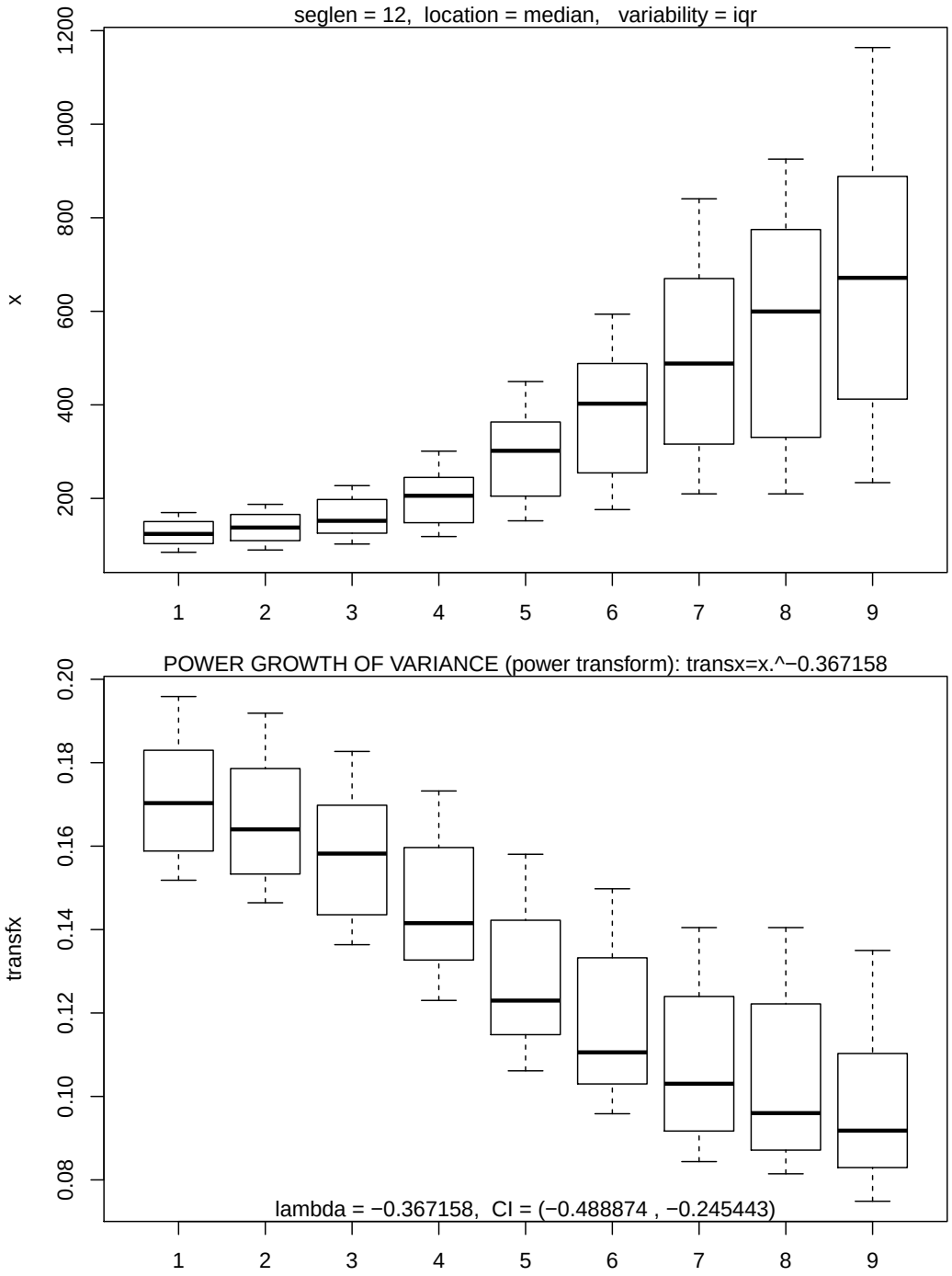
```
> outp<-powtr(UKgas,seglen=10,figure3=TRUE,location="median",variability="iqr")
```



Obrázek 7: `powtr` – krabicové grafy (volba `seglen=10` a `figure3=TRUE`) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

A konečně `seglen=12`.

```
> outp<-powtr(UKgas,seglen=12,figure3=TRUE,location="median",variability="iqr")
```



Obrázek 8: `powtr` – krabicové grafy (volba `seglen=12` a `figure3=TRUE`) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Výsledky předchozích kroků shrňme do tabulky.

`seglen`	dolní mez	odhad $\hat{\lambda}$	horní mez	doporučená transformace
4	-0.585	-0.471	-0.356	mocninná
6	-0.587	-0.426	-0.266	mocninná
8	-0.574	-0.435	-0.296	mocninná
10	-0.564	-0.339	-0.115	mocninná
12	-0.489	-0.367	-0.245	mocninná

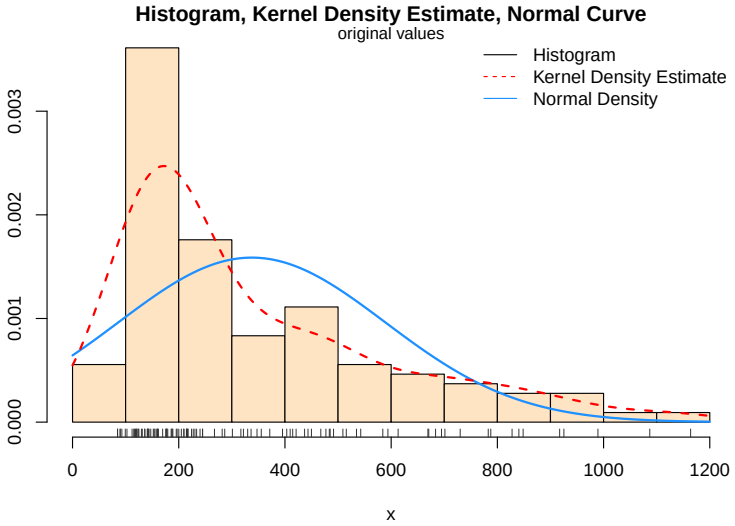
Protože žádný interval neobsahuje jedničku, budeme muset provést transformaci. Žádný interval také neobsahuje nulu, takže můžeme vyloučit logaritmickou transformaci a v úvahu tak přichází mocninná transformace.

Protože bývá zvykem vybírat transformace z hodnot, které mají rozumnou interpretaci, připadají v úvahu například hodnoty $-\frac{1}{2}$, $-\frac{1}{3}$ nebo $-\frac{1}{4}$, které leží uvnitř intervalů spolehlivosti.

Jednotlivé tranformace postupně porovnáme. Pro každou vykreslíme pomocí funkce `boxplotSegments` graf s boxploty pro posouzení variability transformovaných dat a poté vykreslíme histogram (s jádrovým odhadem hustoty a s odhadem normální hustoty), abychom mohli posoudit, zda se rozdělení transformovaných dat přiblížilo normálnímu rozdělení.

Pro porovnání nejprve vykresleme histogram původních, netransformovaných dat.

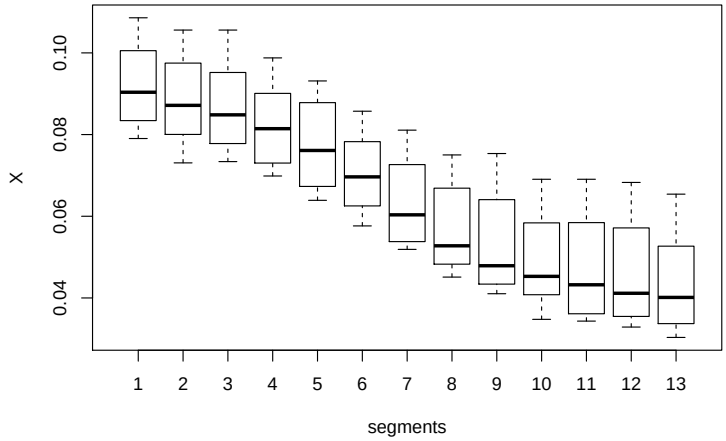
```
> HistFit(UKgas)
> mtext("original values",side=3,line=-0.5,cex=0.95)
```



Obrázek 9: Testování normality pro netrasformovaná data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

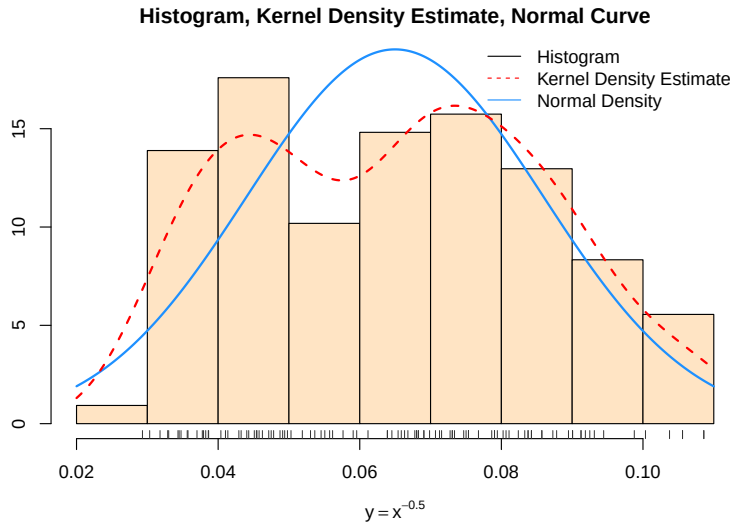
Nejprve výsledky pro mocninnou transformaci s $\lambda = -0.5$.

```
> boxplotSegments(UKgas^(-0.5),seglen=8)
```



Obrázek 10: Porovnání variability pro trasformovaná data ($Y = \frac{1}{\sqrt{X}}$) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

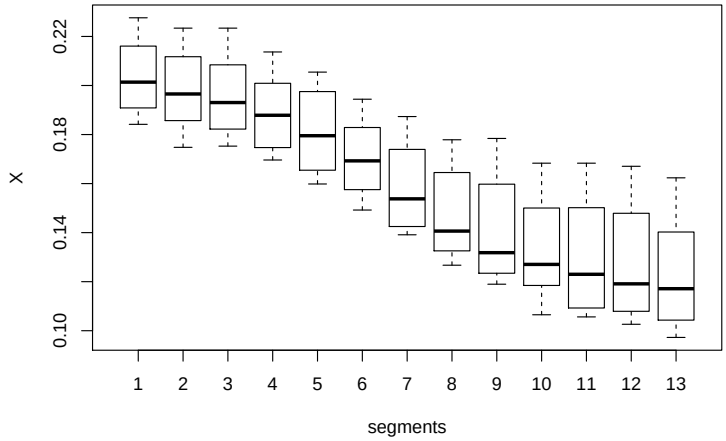
```
> HistFit(UKgas^(-0.5),xlab=expression(y == x^-0.5))
```



Obrázek 11: Testování normality pro trasformovaná data ($Y = \frac{1}{\sqrt{X}}$) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

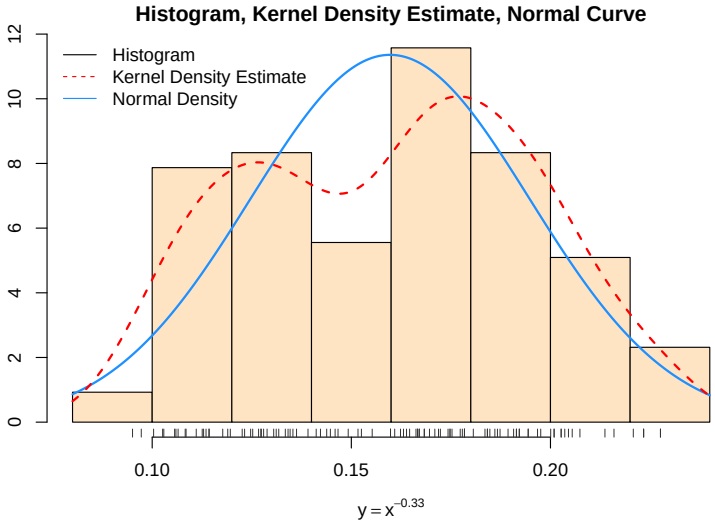
Obdobně dostaneme výsledky pro transformaci $Y = X^{-1/3}$.

```
> boxplotSegments(UKgas^(-1/3),seglen=8)
```



Obrázek 12: Porovnání variability pro trasformovaná data ($Y = \frac{1}{\sqrt[3]{X}}$) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

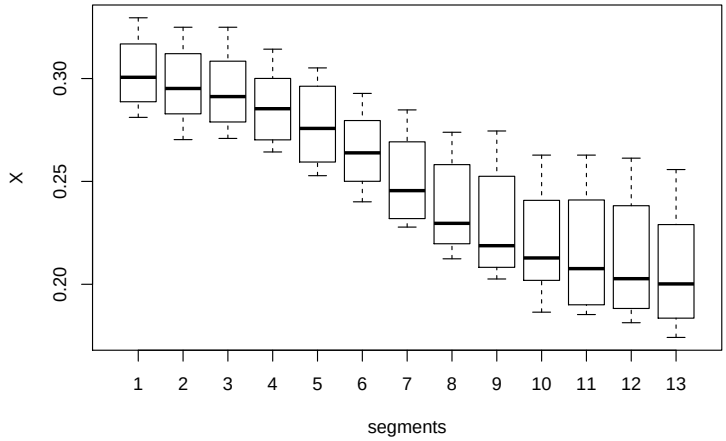
```
> HistFit(UKgas^(-1/3),xlab=expression(y == x^-0.33))
```



Obrázek 13: Testování normality pro trasformovaná data ($Y = \frac{1}{\sqrt[3]{X}}$) „*Spotřeba plynu ve Velké Británii – čtvrtletní údaje*“

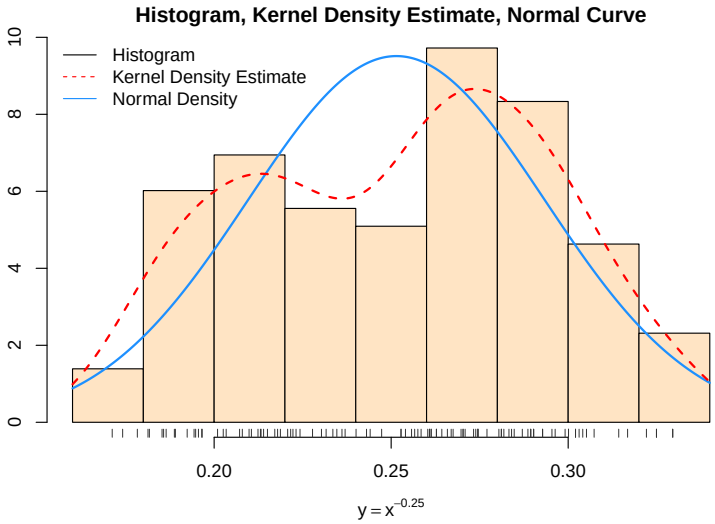
A konečně výsledky pro transformaci $Y = X^{-1/4}$.

```
> boxplotSegments(UKgas^(-0.25),seglen=8)
```



Obrázek 14: Porovnání variability pro trasformovaná data ($Y = \frac{1}{\sqrt[4]{X}}$) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

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

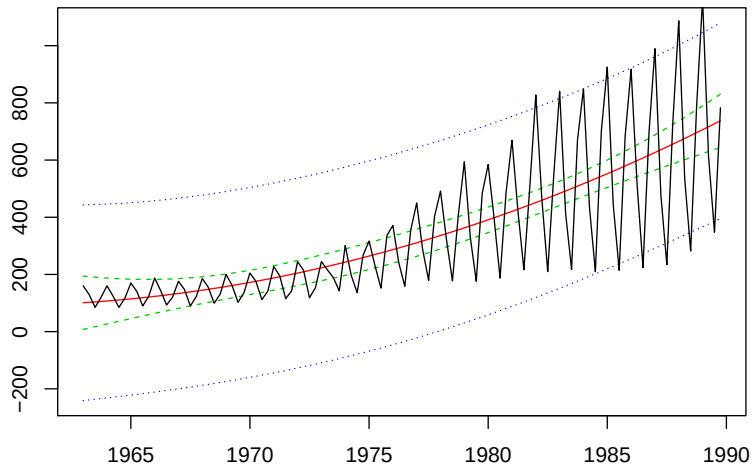


Obrázek 15: Testování normality pro trasformovaná data ($Y = \frac{1}{\sqrt[4]{X}}$) pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Z grafů je ihned vidět, že ačkoli je rozdělení transformovaných dat méně sešikmené, stále má k normalitě daleko. Je třeba si však uvědomit, že to nemusí být 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 trend a teprve pro rezidua hledat vhodnou mocninnou transformaci.

Tento postup nyní vyzkoušíme. Pokud si pozorně prohlédneme graf dat, napadne nás, že trend by mohl být kvadratický. Zcela korektní postup by vyžadoval, abychom stupeň prokládaného polynomu určili exaktnějším způsobem než je pohled „od oka“ (viz např. 3. cvičení), ale tím se zde nyní nebudeme zabývat a rovnou odhadneme model s kvadratickým trendem.

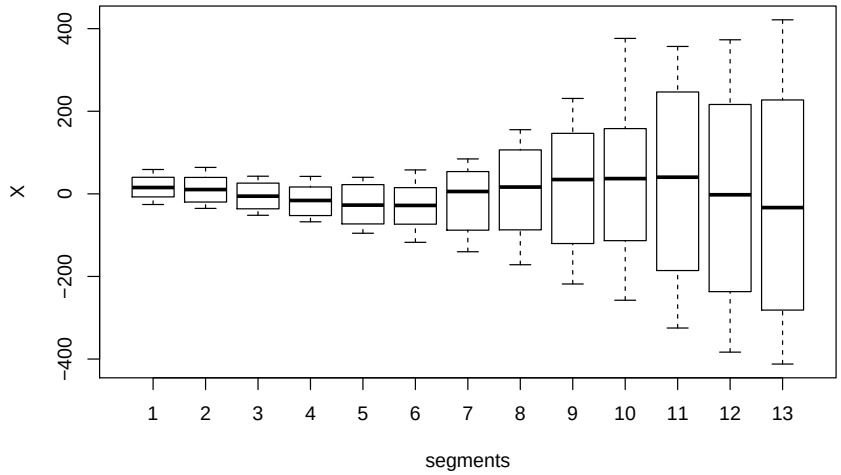
```
> y <- as.vector(UKgasTS)
> x <- as.vector(time(UKgasTS))
> n<-length(y)
> nn<-300
> data<-data.frame(x,y)
> LinTrend <- lm(y ~ x+I(x^2),data=data)
> print(summary(LinTrend))


Call:
lm(formula = y ~ x + I(x^2), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-412.12  -65.52    2.89   58.20  457.79

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.652e+06  1.149e+06   2.309   0.0229 *
x           -2.707e+03  1.162e+03  -2.329   0.0218 *
I(x^2)       6.910e-01  2.941e-01   2.349   0.0207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 166 on 105 degrees of freedom
Multiple R-squared: 0.5717,        Adjusted R-squared: 0.5636
F-statistic: 70.09 on 2 and 105 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)
```



Obrázek 16: Lineární trend pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

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

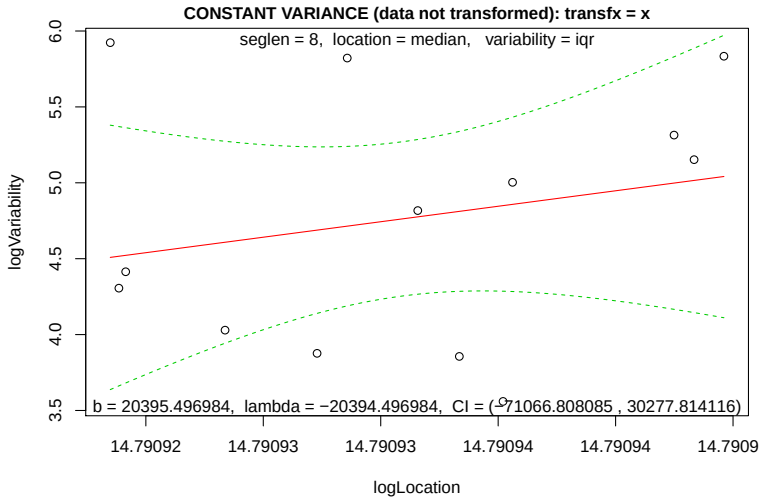
```
> x<-resid(LinTrend)
> boxplotSegments(x,seglen=8)
```



Obrázek 17: `boxplotSegments` (`seglen=8`) pro rezidua po lineárním trendu u dat „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Vyzkoušíme mocninnou transformaci (i když to nejspíše nebude mít smysl) a pomocí funkce `boxplotSegments()` si prohlédneme variabilitu pro transformovaná data. Ještě předtím, než provedeme transformaci, přičteme k reziduím absolutní člen z odhadnuté regresní přímky, abychom prováděli transformaci pro nezáporná data.

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



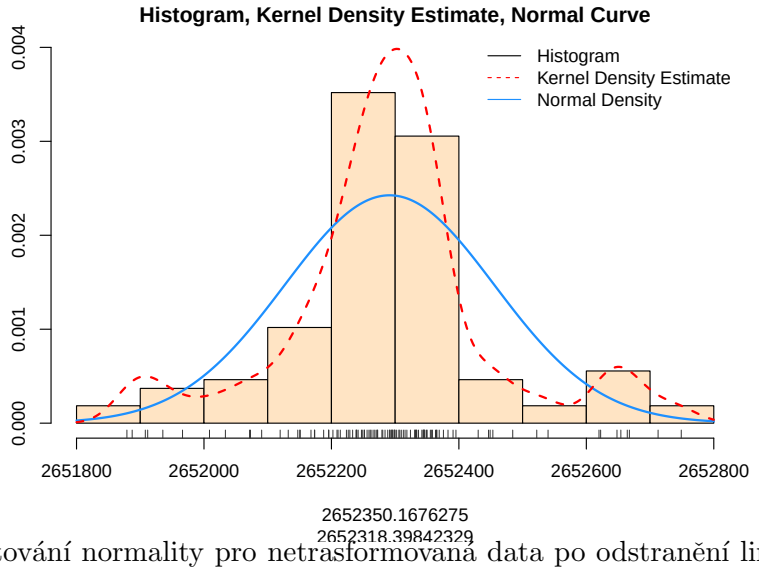
Obrázek 18: `powtr` – regresní přímka pro logaritmy polohy a variability (volba `figure=TRUE`) pro rezidua po lineárním trendu u dat „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Protože interval spolehlivosti obsahuje jedničku, nemělo by být třeba provádět transformaci. Všimněme si ale dole v obrázku, jak velkých hodnot nabývá odhad λ a meze intervalu

spolehlivosti.

Nyní se podívejme na normalitu pomocí příkazu `HistFit()`.

```
> HistFit(outp$transfx,xlab=x)
```



Obrázek 19: Testování normality pro netrasformovaná data po odstranění lineárního trendu $(Y = X)$ „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Ačkoli by se třeba pouhým pohledem mohlo zdát, že rezidua mají k normálnímu rozdělení trochu blíže než data po předchozích transformacích, je třeba si uvědomit, že hlavním cílem bylo stabilizovat rozptyl, což se nám, jak je patrné z obrázku s krabicovými grafy pro rezidua, nepovedlo. Tento postup se tedy rozhodně neosvědčil. Byl totiž od samotného počátku 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 λ pro Box-Coxovu transformaci.

```
> library(car)
> TT<-time(UKgasTS)
> data<-data.frame(TIME=TT-mean(TT),X=UKgas)
> transf1<-powerTransform(X ~ TIME+I(TIME^2), data=data)
> summary(transf1)


bcPower Transformation to Normality

   Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Y1    -0.362   0.1278          -0.6126          -0.1115

Likelihood ratio tests about transformation parameters
                            LRT df        pval
LR test, lambda = (0)   7.947835  1 0.004814494
LR test, lambda = (1) 102.969916  1 0.000000000


> str(transf1)
```

```
List of 13
 $ value      : num 499
 $ 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] 61.2
 $ start      : num -0.362
 $ lambda     : Named num -0.362
  ..- attr(*, "names")= chr "Y1"
 $ roundlam   : Named num -0.5
  ..- attr(*, "names")= chr "Y1"
 $ family     : chr "bcPower"
 $ xqr        :List of 4
  ..$ qr   : num [1:108, 1:3] -10.3923 0.0962 0.0962 0.0962 0.0962 ...
  .. ..- attr(*, "assign")= int [1:3] 0 1 2
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:108] "1" "2" "3" "4" ...
  .. .. ..$ : chr [1:3] "(Intercept)" "TIME" "I(TIME^2)"
  ..$ rank : int 3
  ..$ qraux: num [1:3] 1.1 1.15 1.15
  ..$ pivot: int [1:3] 1 2 3
  ..- attr(*, "class")= chr "qr"
 $ y          : num [1:108, 1] 160.1 129.7 84.8 120.1 160.1 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:108] "1" "2" "3" "4" ...
  .. ..$ : NULL
 $ x          : num [1:108, 1:3] 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:108] "1" "2" "3" "4" ...
  .. ..$ : chr [1:3] "(Intercept)" "TIME" "I(TIME^2)"
  ..- attr(*, "assign")= int [1:3] 0 1 2
 $ weights    : num [1:108] 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "class")= chr "powerTransform"
```

Všimněme si, že `transf$roundlam` nabízí vhodnou volbu parametru λ. Nejedná se přímo o odhad $\hat{\lambda}$, ale o takovou hodnotu, která leží v intervalu spolehlivosti a má rozumnou interpretaci.

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

```
  Y1
-0.5
```

Nyní ukážeme ještě trochu jiný, ale zcela ekvivalentní postup.

```
> m1 <- lm(X ~ TIME+I(TIME^2), data=data)
> summary(m1)
```

```
Call:
lm(formula = X ~ TIME + I(TIME^2), data = data)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-412.12  -65.52    2.89   58.20  457.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 295.6571    23.9670  12.336   <2e-16 ***
TIME         23.7878     2.0499  11.604   <2e-16 ***
I(TIME^2)     0.6910     0.2941   2.349   0.0207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 166 on 105 degrees of freedom
Multiple R-squared: 0.5717,        Adjusted R-squared: 0.5636
F-statistic: 70.09 on 2 and 105 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    -0.362   0.1278          -0.6126          -0.1115

Likelihood ratio tests about transformation parameters
                             LRT df        pval
LR test, lambda = (0)   7.947835  1 0.004814494
LR test, lambda = (1) 102.969916  1 0.000000000
```

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

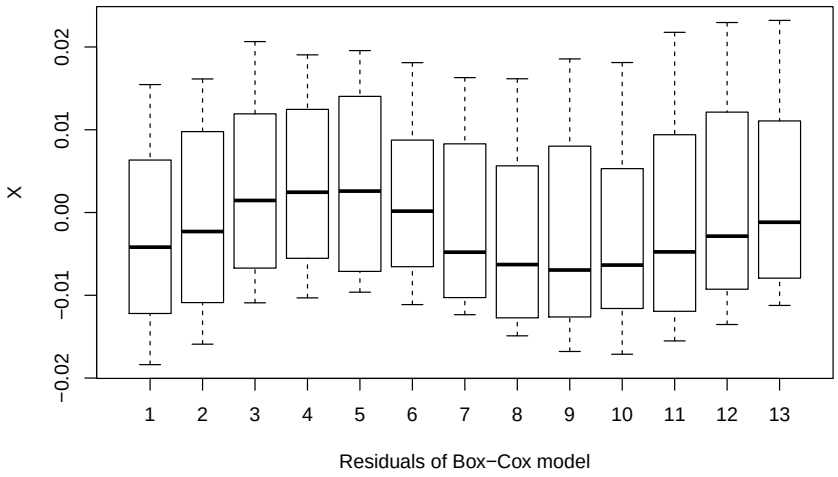
```
  Y1
-0.5
```

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

```
Call:
lm(formula = basicPower(outT$y, outT$roundlam) ~ TIME + I(TIME^2),
    data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.018379 -0.009724 -0.003079  0.010398  0.023211

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.341e-02  1.736e-03   36.53   <2e-16 ***
TIME        -2.195e-03  1.485e-04  -14.78   <2e-16 ***
I(TIME^2)    2.599e-05  2.130e-05    1.22    0.225
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01203 on 105 degrees of freedom
Multiple R-squared: 0.677,        Adjusted R-squared: 0.6708
F-statistic:   110 on 2 and 105 DF,  p-value: < 2.2e-16
```

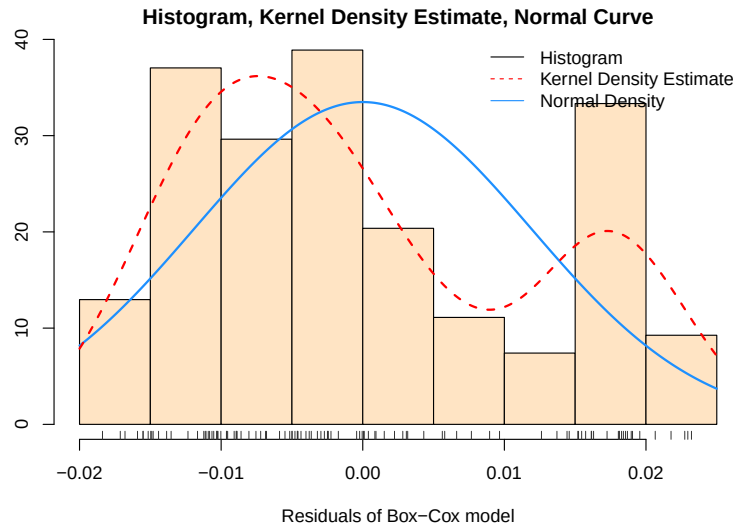
Na závěr se ještě podívejme, jak dopadla rezidua u transformovaného modelu.

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



Obrázek 20: `boxplotSegments` (volba `seglen=8`) rezidua v Box–Coxově modelu pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

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



Obrázek 21: Testování normality reziduí v Box–Coxově modelu pro data „*Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje*“

Na závěr můžeme říci, že se nám podařilo najít transformaci, po níž mají data konstantní rozptyl. Normalita se ale ani v tomto případě příliš nezlepšila.

Protože z výpisu informací o modelu `m2` je vidět, že koeficient u druhé mocniny času se neliší od nuly na hladině významnosti 5 %, vyzkoušíme ještě jako poslední možnost odstranit z dat pouze lineární trend a zároveň najít vhodný transformační parametr.

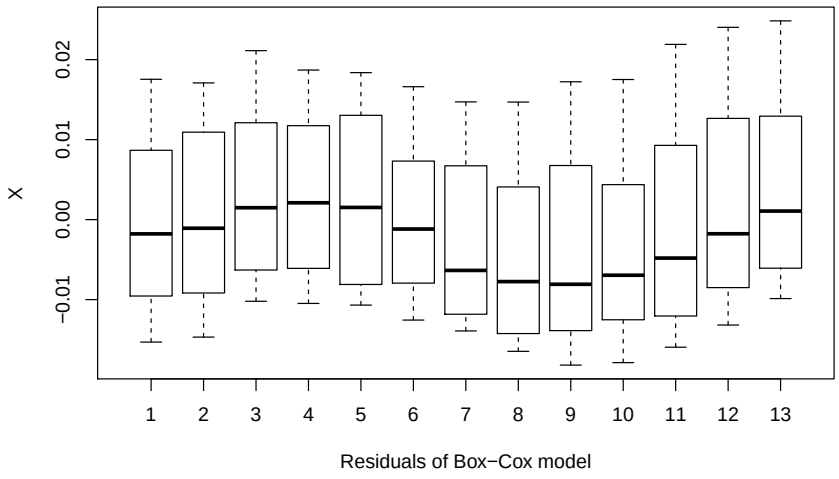
```
> m3 <- lm(X ~ TIME, data=data)
> summary(m3)


Call:
lm(formula = X ~ TIME, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-368.57  -85.50    1.20   69.86  525.95

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  337.631     16.314   20.70   <2e-16 ***
TIME          23.788      2.093   11.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 169.5 on 106 degrees of freedom
Multiple R-squared: 0.5492,        Adjusted R-squared: 0.545
F-statistic: 129.2 on 1 and 106 DF,  p-value: < 2.2e-16


> outT <- powerTransform(m3)
> summary(outT)


bcPower Transformation to Normality

   Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Y1   -0.3275    0.121          -0.5647          -0.0903

Likelihood ratio tests about transformation parameters
                             LRT df        pval
LR test, lambda = (0)   7.299312  1 0.006898103
LR test, lambda = (1) 107.836554  1 0.000000000


> print(outT$roundlam)


  Y1
-0.5


> m4<-update(m3,basicPower(outT$y,outT$roundlam)~.)
> summary(m4)


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

Residuals:
      Min        1Q    Median        3Q       Max
-0.018188 -0.010034 -0.002811  0.009518  0.024849

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0649838  0.0011598   56.03   <2e-16 ***
TIME        -0.0021950  0.0001488  -14.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01205 on 106 degrees of freedom
Multiple R-squared: 0.6724,        Adjusted R-squared: 0.6693
F-statistic: 217.6 on 1 and 106 DF,  p-value: < 2.2e-16
```

I v případě jednoduššího lineárního trendu je navržena stejná transformace jako v případě kvadratického trendu.

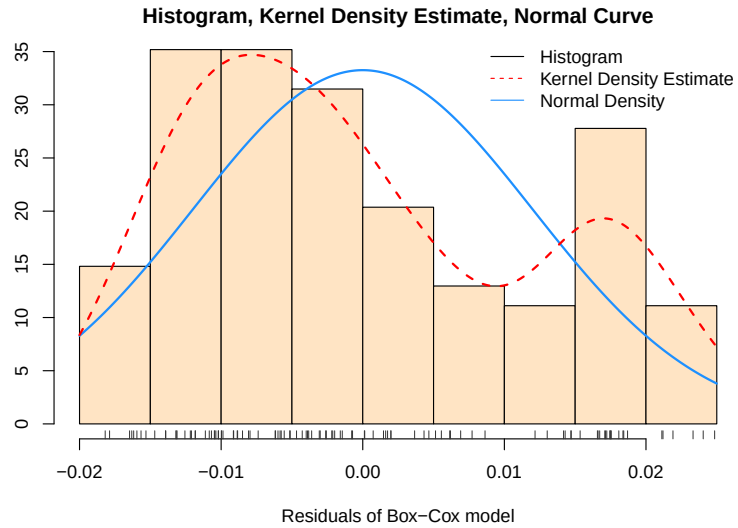
Na závěr si prohlédněme transformovaná rezidua pro lineární trend.

```
> res<-resid(m4)
> boxplotSegments(res,seglen=8,xlab="Residuals of Box-Cox model")
```



Obrázek 22: `boxplotSegments` (volba `seglen=8`) rezidua v Box–Coxově modelu pro data *„Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje“*

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



Obrázek 23: Testování normality reziduí v Box-Coxově modelu pro data *„Spotřeba zemního plynu ve Velké Británii – čtvrtletní údaje“*

Oproti modelu s kvadratickým trendem není na první pohled patrný žádný výrazný rozdíl.

4 Úkol:

Načtěte data uložená v souboru `jj.dat`.

- Data vykreslete do grafu a posuďte, zda bude třeba je transformovat.
- Použijte funkci `boxCox()` ke zjištění odhadu parametru mocninné transformace λ metodou maximální věrohodnosti. Obsahuje interval spolehlivosti některou z hodnot 1 nebo 0, což by znamenalo, že data není třeba transformovat, resp. lze použít logaritmickou transformaci?
- Dále použijte funkci `powtr` s různými volbami parametru `seglen` a na základě výsledků zvolte vhodný transformační parametr λ.
- Na závěr pomocí funkce `powerTransform` zároveň odhadněte lineární trend i transformační parametr λ.
- U všech možných transformací vykreslete histogram a porovnávejte, zda se rozdělení dat blíží normálnímu.