M0130 – 9. PRAKTIKUM : **M0130pr09** (*Analýza reziduí*)

A. Míry influence v regresní diagnostice

V praxi se často můžeme setkat s jevem, že v souboru dat se vyskytují některé hodnoty výrazně se lišící od hodnot ostatních. V literatuře se v průběhu minulých let rozvinuly dva směry, které se svým způsobem snaží s jejich existencí vyrovnat.

- metody robustní statistiky
- metody regresní diagnostiky

Regresní diagnostika jde cestou detekovat více či méně ojedinělá data a dát původci dat možnost rozhodnout se, jak s nimi v případě výskytu dále naložit, tj. zda je v souboru ponechat či vyloučit, věnovat jim menší váhu při zpracování, popřípadě je vhodně transformovat apod. V rámci regresní diagnostiky se budeme zabývat dvěma základními úlohami

- jak detekovat mezi daty neočekávané hodnoty
- jak rozhodnout, zda mohou významně ovlivnit statistickou analýzu, případně jakým způsobem.

Regresní model $\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ \wedge\ E\boldsymbol{\varepsilon} = 0 \ \wedge\ var\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}_n \ \wedge\ h(\mathbf{X}) = k = p+1}$.

Odhady metodou nejmenších čtverců:

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{Y}.$$

Projekční matice $\boxed{\mathbf{H}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ je idempotentní $\mathbf{H}^2 = \mathbf{H}$, symetrická $\mathbf{H}' = \mathbf{H}$

Označme matici
$$\mathbf{H} = \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ \underbrace{h_{n1}}_{\mathbf{H}_1} & \cdots & \underbrace{h_{nn}}_{\mathbf{H}_n} \end{pmatrix} = (\mathbf{H}_1, \cdots, \mathbf{H}_n) \quad . \text{ Pak } \widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}.$$

Protože $\quad Y_i = \mathbf{H}_i'\mathbf{Y} = \sum_{j=1}^{n} h_{ij}Y_j,\quad$ vidíme, jak pozorování Y_j ovlivňuje i-tou odhadovanou hodnotu $\hat{Y}_i$.

Definice 1: **Sloupce** matice $\mathbf{H}$ se nazývají **vlivové vektory** a $\|\mathbf{H}_j\|^2$ se nazývá **vliv j-tého pozorování** na odhad $\widehat{\mathbf{Y}}$, stručně **j-tý vliv**.

Věta 1: $\boxed{\|\mathbf{H}_j\|^2 = h_{jj}}$.

Důkaz: $\|\mathbf{H}_j\|^2 = \mathbf{H}_j'\mathbf{H}_j = \sum_{i=1}^{n} h_{ij}h_{ij} \wedge \mathbf{H}^2 = \mathbf{H} \Rightarrow h_{jj} = \mathbf{H}_j'\mathbf{H}_j$

Věta 2: $\boxed{\bigvee_{i=1}^{n} 0 \le h_{ii} \le 1}$.

Důkaz: $\mathbf{H}^2 = \mathbf{H} \Rightarrow h_{ii} > 0;\ \ h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ij} = h_{ii}^2 + \sum_{i\neq j} h_{ij}^2 \Rightarrow h_{ii} > h_{ii}^2 \Rightarrow h_{ii} < 1$

Věta 3: $\boxed{tr\mathbf{H} = k}$

Důkaz: $\mathbf{H}^2 = \mathbf{H}$ je idempotentní $\Rightarrow tr\mathbf{H} = h(\mathbf{H})$. Užitím věty o součinu matic lze odvodit, že $h(\mathbf{H}) = h(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = k$.

Věta 4: Průměrný vliv pozorování $Y_1, \cdots, Y_n$ je roven $\boxed{\frac{k}{n}}$.

Důkaz: $\sum_{i=1}^{n} h_{ii} = k \Rightarrow \bar{h} = \frac{1}{n}(||\mathbf{H}_1||^2 + \cdots + ||\mathbf{H}_n||^2) = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{k}{n}$.

Definice 2: Definujme **vektor reziduí**: $\boxed{\mathbf{r} = \mathbf{Y} - \widehat{\mathbf{Y}}}$.

Věta 5: $\boxed{\mathbf{r} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}}$

Důkaz: $\mathbf{r} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \underbrace{(\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta}}_{=\mathbf{0}} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$
$= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$

Pak maticově
$$\begin{pmatrix} r_1 \\ \vdots \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} 1-h_{11} & -h_{12} & \dots & h_{1n} \\ -h_{21} & 1-h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{n1} & \dots & \cdots & 1-h_{nn} \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j = \varepsilon_i(1 - h_{ii}) - \sum_{j \neq i} h_{ij}\varepsilon_j$.

Poznámka 1: Pokud jsou prvky $h_{ii} \approx 1$ (blízké k 1) $\Rightarrow 1 - h_{ii} \approx 0$. Pak neočekávaně velká chyba pozorování Y_i (velká chyba ε_i) se nemusí odrážet v r_i. Na ostatní rezidua však vliv mít může.

Věta 6: $\boxed{D\widehat{\mathbf{Y}} = \sigma^2\mathbf{H}}$; $\boxed{D\mathbf{r} = \sigma^2(\mathbf{I} - \mathbf{H})}$.

Důkaz: $D\widehat{\mathbf{Y}} = D(\mathbf{HY}) = \mathbf{H}D\mathbf{Y}\mathbf{H}' = \sigma^2\mathbf{HH}' = \sigma^2\mathbf{H}^2 = \sigma^2\mathbf{H}$;
$D\mathbf{r} = D(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon} = (\mathbf{I}-\mathbf{H})D\boldsymbol{\varepsilon}(\mathbf{I}-\mathbf{H})' = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})' = \sigma^2(\mathbf{I}-\mathbf{H})^2 = \sigma^2(\mathbf{I}-\mathbf{H})$

Poznámka 2: Pokud matice $\mathbf{H}$ není diagonální maticí, pak rezidua r_i jsou korelovaná. Z předchozího je vidět, že rezidua r_i v některých situacích nemusí dobře identifikovat odlehlá pozorování. Proto se v literatuře zavádějí a používají další typy reziduí.

Značení: Symbol (i) bude znamenat vynechání i-tého řádku, symbol $[j]$ vynechání j-tého sloupce.

Definice 3: Definujme:

- **standardizovaná rezidua** (také **normovaná rezidua**)
  $\boxed{r_i^* = \frac{r_i}{s\sqrt{1-h_{ii}}}}$, kde $s^2 = \frac{(\mathbf{Y}-\widehat{\mathbf{Y}})'(\mathbf{Y}-\widehat{\mathbf{Y}})}{n-k} = \frac{SSE}{n-k}$.
- i**-té predikované reziduum** $\boxed{r_{(i)} = Y_i - \hat{Y}_{(i)}}$ kde v lineárním modelu vynecháme i-té pozorování a značíme matici plánu $\mathbf{X}_{(i)}$, vektor $\mathbf{Y}_{(i)}$, odhad metodou nejmenších čtverců $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}\mathbf{Y}_{(i)}$ a i-té pozorování odhadneme pomocí $\hat{\boldsymbol{\beta}}_{(i)}$ takto $\hat{Y}_{(i)} = \mathbf{x}_i\hat{\boldsymbol{\beta}}_{(i)}$, kde $\mathbf{x}_i$ je i-tý řádek původní matice plánu.
- **studentizovaná (jackknife) rezidua** $\boxed{r_{(i)}^* = \frac{r_{(i)}\sqrt{1-h_{ii}}}{s_{(i)}}}$, kde
  $s_{(i)}^2 = \frac{(\mathbf{Y}_{(i)}-\widehat{\mathbf{Y}}_{(i)})'(\mathbf{Y}_{(i)}-\widehat{\mathbf{Y}}_{(i)})}{n-k-1}$.
- i**-té DFFIT reziduum** $\boxed{d_i = \frac{\hat{Y}_i - \widehat{Y}_{(i)}}{\sqrt{h_{ii}}s_{(i)}}}$.
- i**-té parciální reziduum** $\boxed{r_{i[j]} = Y_i - \hat{Y}_{i[j]}}$ kde v lineárním modelu vynecháme j-tý regresor, odpovídající matici plánu označíme $\mathbf{X}_{[j]}$, odhad metodou nejmen-

ších čtverců $\hat{\boldsymbol{\beta}}_{[j]} = (\mathbf{X}'_{[j]}\mathbf{X}_{[j]})^{-1}\mathbf{X}_{[j]}\mathbf{Y}$ a i-té pozorování odhadneme pomocí $\hat{\boldsymbol{\beta}}_{[j]}$ takto $\hat{Y}_{i[j]} = \mathbf{x}_{i[j]}\hat{\boldsymbol{\beta}}_{[j]}$, kde $\mathbf{x}_{i[j]}$ je i-tý řádek matice plánu $\mathbf{X}_{[j]}$.

Věta 7: Nechť $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \Rightarrow \boxed{\mathbf{r} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I}-\mathbf{H}))}$.

Důkaz: viz. Anděl(1978).

Poznámka 3: Protože $s^2(1-h_{ii})$ je odhad rozptylu $Dr_i = \sigma^2(1-h_{ii})$, pak standardizovaná rezidua mají rozptyl přibližně roven 1. Pokud nastane situace, že chyba je příliš velká oproti modelu, pak pomocí příslušného standardizovaného rezidua je lze snadněji identifikovat.

Věta 8: $\boxed{r_{(i)} = \frac{r_i}{1-h_{ii}}}$; $\boxed{Dr_{(i)} = \frac{\sigma^2}{1-h_{ii}}}$

Věta 9: $\boxed{(n-k)s^2_{(i)} = (n-k)s^2 - \frac{r_i^2}{1-h_{ii}}}$;

Věta 10: $\boxed{r^*_{(i)} = \frac{r_i}{\sqrt{1-h_{ii}}s_{(i)}}}$

Důkazy: viz. Staudte, Sheather(1990).

Poznámka 4: Předchozí vzorce umožňují vypočítat statistiky $r_{(i)}$, $s^2_{(i)}$ a $r^*_{(i)}$ pouze z hodnot známých z celého regresního modelu.

Věta 11: Nechť $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \Rightarrow \boxed{r^*_{(i)} \sim t(n-k)}$.

Důkaz: viz. Staudte, Sheather(1990).

Definice 4: Řekneme, že i-té pozorování Y_i **je odlehlé**, jestliže $\boxed{E\varepsilon_i \neq 0}$.

Poznámka 5: Z věty 11 plyne, že pomocí i-tého studentizovaného rezidua $r^*_{(i)}$ lze testovat nulovou hypotézu $H_0 : E\varepsilon_i = 0$, že i-té pozorování **není odlehlé** proti alternativě $H_1 : E\varepsilon_i \neq 0$, tj. že je odlehlé. Pokud $|r^*_{(i)}| \geq t_{1-\alpha/2}(n-k)$, pak na hladině významnosti α zamítáme hypotézu H_0 a tedy i-té pozorování na hladině významnosti α je odlehlé.

Věta 12: DFFITS rezidua d_i pro $i = 1, \cdots, n$ lze vyjádřit vztahem $\boxed{d_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} r^*_{(i)}}$.

Důkaz: viz. Staudte, Sheather(1990).

Poznámka 6: Je-li $n-k > 30$, je na hladině významnosti $\alpha = 0.05$ kvantil Studentova t rozdělení přibližně roven 2 a lze tedy v praktických situacích na základě věty 11 považovat i-té pozování za odlehlé na hladině významnosti $\alpha = 0.05$, když $\boxed{|r^*_{(i)}| \geq 2}$.

Toto odpovídá také empirických zkušenostem (viz. Staudte, Sheather(1990)).

Posuzujeme-li odlehlé pozorování pomocí i-tého DFFITS rezidiua, můžeme využít vztah z věty 12 a navíc uplatnit vliv i-tého pozorování h_{ii} a jeho průměrnou hodnotu. Pak platí

$$|d_i| = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} |r^*_{(i)}| > 2\left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}}.$$

Uvážíme-li, že průměrný vliv je $\frac{k}{n}$ a dosadíme-li jej za h_{ii}, dostaneme

$$|d_i| = 2\left(\frac{\frac{k}{n}}{1-\frac{k}{n}}\right)^{\frac{1}{2}} = 2\left(\frac{k}{n-k}\right)^{\frac{1}{2}} > 2\left(\frac{k}{n}\right)^{\frac{1}{2}}.$$

Posledně uvedená nerovnost se v praxi užívá pro posouzení, zda i-té pozorování je odlehlé na základě DFFIT reziduí.

Cookova vzdálenost. Pro měření **vlivu i-tého pozorování na hodnotu odhadu vektoru $\boldsymbol{\beta}$** navrhl Cook použít statistiku

$$\boxed{D_i = \frac{(\widehat{\mathbf{Y}}-\widehat{\mathbf{Y}}_{(i)})'(\widehat{\mathbf{Y}}-\widehat{\mathbf{Y}}_{(i)})}{ks^2} = \frac{r_{(i)}^{*2}}{k}\frac{h_{ii}}{1-h_{ii}}}.$$

Cookova vzdálenost souvisí s konfidenčním elipsoidem odhadů, což umožňuje její porovnání s kvantily F-rozdělení s k a $n-k$ stupni volnosti. Jde zde však o posun odhadů, který vznikl vynecháním i-tého bodu. Orientačně platí, že pro $\boxed{D_i > 1}$ posun přesahuje 50%ní konfidenční oblast a daný bod je proto **vlivný**.

Další možné vysvětlení Cookovy vzdálenosti vychází z toho, že jde o eukleidovskou vzdálenost mezi vektorem predikce $\widehat{\mathbf{Y}}$ z metody nejmenších čtverců a vektorem predikce $\widehat{\mathbf{Y}}_{(i)}$, který odpovídá odhadům stanoveným metodou nejmenších čtverců při vynechání i-tého bodu.

Cookova vzdálenost vyjadřuje vliv i-tého bodu pouze na odhady parametrů $\boldsymbol{\beta}$. Pokud proto i-tý bod neovlivní odhady regresních parametrů $\boldsymbol{\beta}$ výrazně, bude hodnota Cookovy vzdálenosti malá.

Takový bod však může silně ovlivnit *odhad reziduálního rozptylu* σ^2, kde $D\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}_n$.

Welschova-Kuhova vzdálenost. Pro měření **vlivu i-tého pozorování simultánně jak na odhad $\boldsymbol{\beta}$**, tak na odhad σ^2, zvolili Welsch a Kuh statistiku

$$\boxed{DFFITS_i = d_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{h_{ii}}s_{(i)}}}$$

(viz. def. 3, věta 12 a pozn. 6).

Parciální vliv. Pro měření **vlivu i-tého pozorování na j-tou složku vektoru $\hat{\boldsymbol{\beta}}$** navrhli Belsley, Kuh a Welsch statistiku $\boxed{DFBETAS_{ij} = \frac{\hat{\beta}_i - \hat{\beta}_{j(i)}}{\sqrt{D\hat{\beta}_j}}}$

Variační poměr. Pro stanovení **míry vlivu i-tého pozorování na matici $D\hat{\boldsymbol{\beta}}$** je navržena statistika

$$\boxed{COVRATIO_i = \frac{(s_{(i)}^2/s^2)^k}{1-h_{ii}}}.$$

Velké hodnoty statistiky signalizují vlivné body (vzhledem k $D\hat{\boldsymbol{\beta}}$).

V literatuře se za vlivná pozorování doporučuje považovat ta, pro něž

$$\boxed{|COVRATIO_i - 1| > 3\frac{k}{n}}.$$

LITERATURA:

Anděl, J.(1978): *Matematická statistika*, Praha, SNTL.

Antoch, J., Vorlíčková, D. (1992): *Vybrané metody statistické analýzy dat*, Academia Praha

Staudte, R.G.,Sheather, S.J.(1990): *Robust Estimation and Testing*, New York, Wiley

Míry influence v prostředí R:

K dispozici je především funkce `influence.measures()`, kterou lze použít pouze na objekt třídy `lm`. Výsledkem je objekt třídy `infl`, který je tvořen ze tří prvků: `infmat`, `is.inf` a `call`.

Matice `infmat` :

- každý řadek odpovídá jednomu pozorování $(Y_i, \mathbf{x}_i)$
- prvních k sloupců tvoří matici statistik DFBETAS, sloupce jsou označeny `dbf.` a za tečkou následuje většinou přiměřeně zkrácený název příslušného regresoru
- následuje sloupec statistik DFFITS označený `dffit`.
- další sloupce nazvané `cov.r`, `cook.d` `hat` obsahují statistiky COVRATIO, D_i a diagonální prvky matice $\mathbf{H}$.

Samostatně lze jednotlivá rezidua a další statistiky získat z objektu typu `lm` pomocí funkcí

```
rstandard()  dffits()          dfbetas()    covratio()
rstudent()   cooks.distance()  hatvalues()
```

Parciální rezidua v prostředí R získáme například tímto postupem

1. Nejprve vytvoříme pomocnou matici, která je typu $n \times (k-1)$ (pokud model obsahuje konstatní člen vždy značený jako `(Intercept)`). Matice parciálních reziduí obsahuje tolik sloupců, kolik je regresorů, které lze vynechat.
   `pom.parc.rez <- residuals(model.lm,type = "partial")`
2. Protože tato rezidua jsou modifikována tak, aby každý sloupec měl nulový průměr, je třeba přičíst jistou konstantu, kterou prostředí R spolu s modifikovanými parciálními rezidui nabízí.
   `parc.rez <- pom.parc.rez + attr(pom.parc.rez,"constant")`

Příklad 1: **Návštěvnost v hromadných ubytovacích zařízeních v ČR v letech 2000–2010**

V datovém souboru `NavstevnostHromadUbytZar2000_2010Q.dat` jsou obsažena čtvrtletní data, která se týkají návštěvnosti v hromadných ubytovacích zařízeních v ČR v letech 2000–2010. Hromadným ubytovacím zařízením se rozumí zařízení s minimálně pěti pokoji nebo deseti lůžky.

Čtvrtletní data v jednotlivých sloupcích obsahují

1. sloupec – počet hostů
2. sloupec – počet přenocování

Datový soubor načteme pomocí příkazu `read.table()`. Vzhledem k tomu, že obsahuje v prvním řádku názvy proměnných, v příkazu `read.table` nesmíme zapomenout nastavit `header=TRUE`. Příkazem `str()` vypíšeme strukturu datového rámce.

```
> fileDat <- paste(data.library, "NavstevnostHromadUbytZar2000_2010Q.dat",
     sep = "")
> navstevnost <- read.table(fileDat, header = TRUE)
```

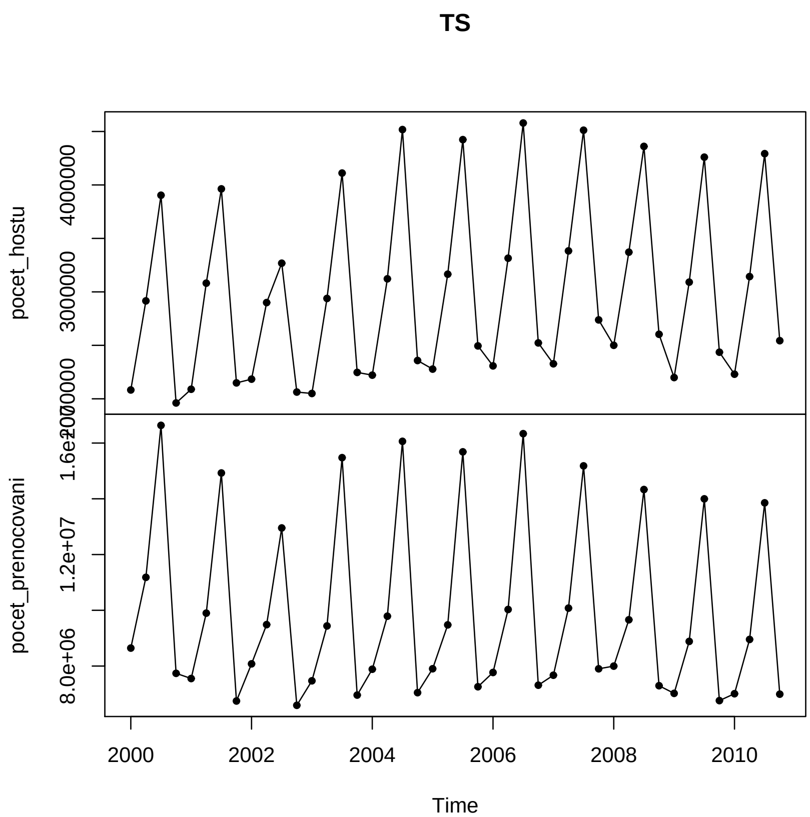
Vedle načteného datového rámce vytvoříme i vícerozměrnou časovou řadu, pomocí příkazu `str()` se podíváme na její strukturu.

```
> TS <- ts(navstevnost, start = 2000, frequency = 4)
> str(TS)
```

```
 int [1:44, 1:2] 2082827 2915857 3903838 1961250 2089561 3081402 3963317 2148905 2183879 2899998 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr [1:2] "pocet_hostu" "pocet_prenocovani"
 - attr(*, "tsp")= num [1:3] 2000 2011 4
 - attr(*, "class")= chr [1:2] "mts" "ts"
```

Časové řady vykreslíme pomocí příkazu `plot()`.

```
> plot(TS, type = "o", pch = 20, cex = 3)
```



Obrázek 1: *Návštěvnost v hromadných ubytovacích zařízeních v ČR v letech 2000–2010*

Pro další zpracování vybereme druhou časovou řadu, která se týká počtu přenocování v hromadných ubytovacích zařízeních. Protože data nabývají hodnot v řádu miliónů, budeme je dále uvažovat v tisících.

```
> xts <- TS[, 2]/1000
```

Protože jde o sezónní model, pro dekompozici časové řady si můžeme vybrat například (modifikovanou) metodu malého trendu

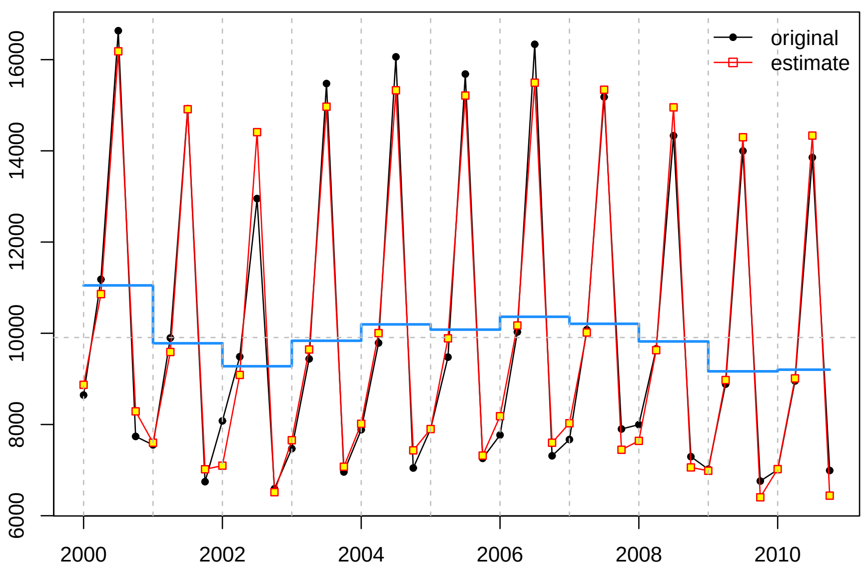
$$\boxed{M_I^{modif}}: \; Y_{jk} = \mu + m_j + s_k + \varepsilon_{jk} \qquad \varepsilon_{jk} \sim WN(0, \sigma^2), \; j = 1, \ldots, r, \; k = 1, \ldots, d$$

s dodatečnými podmínkami

$$s_1 + \cdots + s_d = 0 \qquad \text{a} \qquad m_1 + \cdots + m_r = 0.$$

Dekompozici provedeme pomocí funkce `SzSmallTrendModif()`.

```
> vysl <- SzSmallTrendModif(xts)
```



Obrázek 2: Metoda malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování).*

Pomocí funkcí `summary()` a `anova()` vypíšeme výsledný regresní dekompoziční model s příslušnými statistikami.

```
> summary(vysl$model)


Call:
lm(formula = x ~ grY + grS, data = data, contrasts = list(grY = contr.sum,
    grS = contr.sum))

Residuals:
     Min       1Q   Median       3Q      Max
-1456.56  -237.44   -52.61   331.61   981.06

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9907.01      79.08 125.271  < 2e-16 ***
grY1         1142.89     250.09   4.570 7.82e-05 ***
grY2         -126.47     250.09  -0.506  0.61677
grY3         -629.55     250.09  -2.517  0.01740 *
```

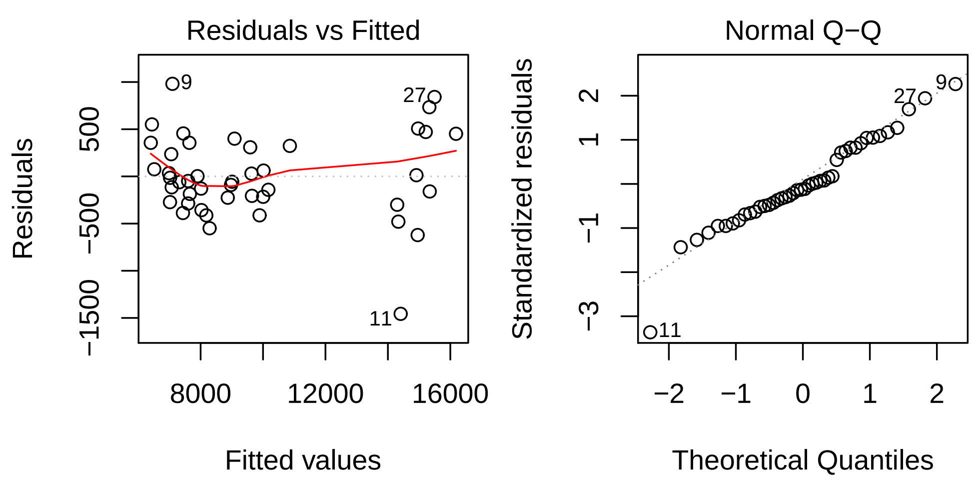
```
grY4          -71.20     250.09  -0.285  0.77783
grY5          288.16     250.09   1.152  0.25832
grY6          173.11     250.09   0.692  0.49414
grY7          454.94     250.09   1.819  0.07889 .
grY8          300.76     250.09   1.203  0.23854
grY9          -86.14     250.09  -0.344  0.73291
grY10        -741.46     250.09  -2.965  0.00589 **
grS1        -2179.80     136.98 -15.913 3.61e-16 ***
grS2         -191.05     136.98  -1.395  0.17334
grS3         5132.67     136.98  37.471  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 524.6 on 30 degrees of freedom
Multiple R-squared: 0.9815,       Adjusted R-squared: 0.9735
F-statistic: 122.7 on 13 and 30 DF,  p-value: < 2.2e-16
```

```
> anova(vysl$model)
```

```
Analysis of Variance Table

Response: x
          Df    Sum Sq   Mean Sq  F value    Pr(>F)
grY       10  12753103   1275310   4.6343 0.0005118 ***
grS        3 426359506 142119835 516.4395 < 2.2e-16 ***
Residuals 30   8255749    275192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

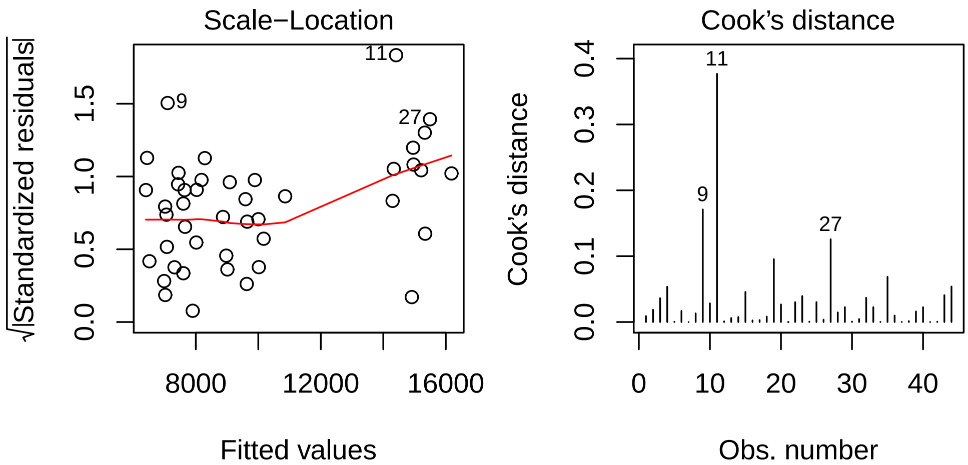
Podíváme se, jakou grafickou regresní diagnostiku nabízí funkce `plot.lm()` (stačí psát pouze `plot()`).

```
> par(mfrow = c(1, 2), mar = c(5, 5, 2, 0) + 0.05)
> plot(vysl$model, which = 1:2)
```



Obrázek 3: Analýza reziduí pomocí funkce `plot()` – grafy 1 a 2 u metody malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování).*

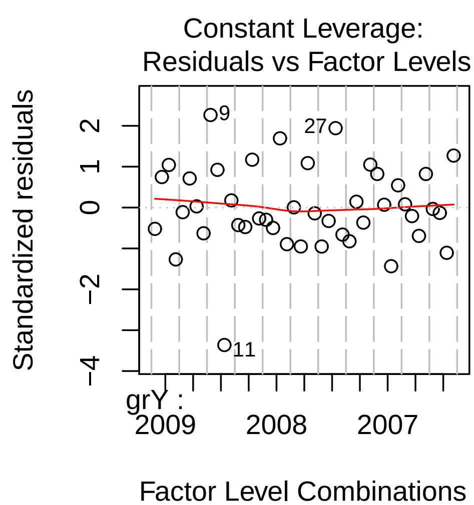
```
> par(mfrow = c(1, 2), mar = c(5, 5, 2, 0) + 0.05)
> plot(vysl$model, which = 3:4)
```



Obrázek 4: Analýza reziduí pomocí funkce `plot()` – grafy 3 a 4 u metody malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování)*

Graf Cookových vzdáleností $\widehat{\mathbf{Y}}$ od $\widehat{\mathbf{Y}}_{(i)}$ odhaluje 3 odlehlá pozorování.

```
> par(mfrow = c(1, 1), mar = c(5, 5, 2, 0) + 0.05)
> plot(vysl$model, which = 5)
```

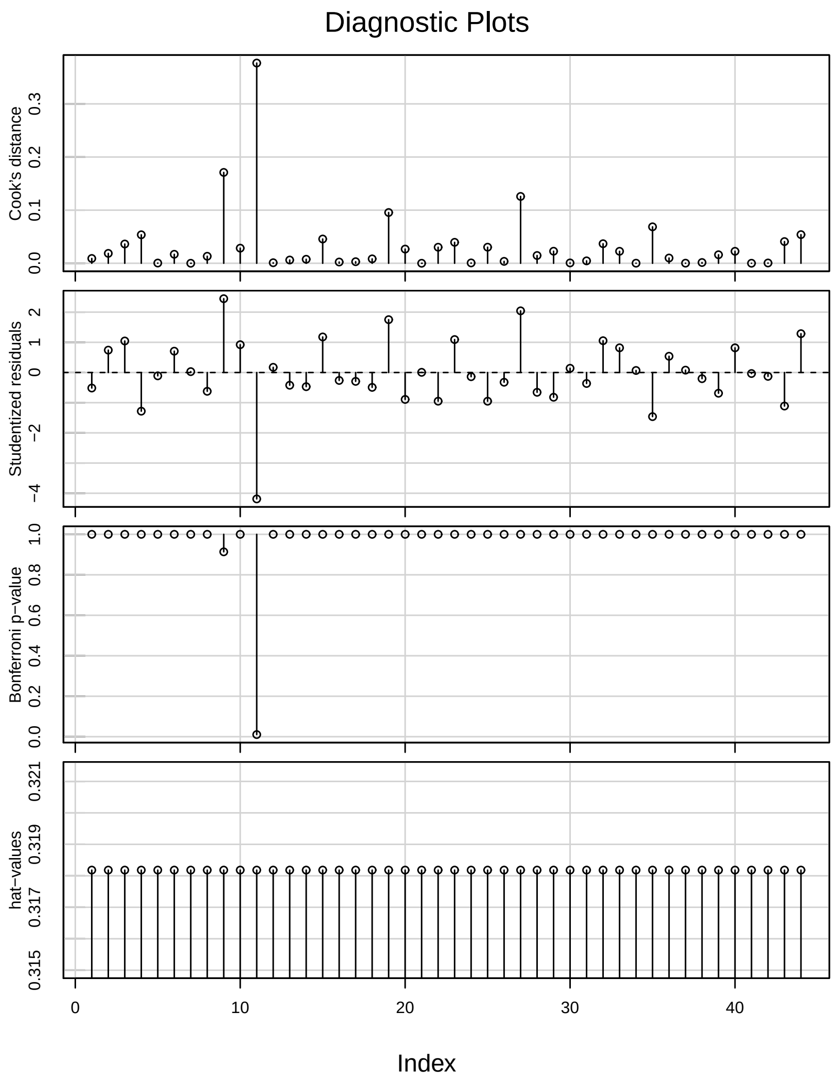


Obrázek 5: Analýza reziduí pomocí funkce `plot()` – graf 5 u metody malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování)*

Zajímavé grafy lze najít v knihovně `car` (`influencePlot()`, `infIndexPlot`). První je určen spíše pro klasickou regresi, druhý je již vhodnější pro naše účely.

```
> library(car)
> par(mfrow = c(1, 1), mar = c(5, 5, 2, 0) + 0.05)
```

```
> infIndexPlot(vysl$model, vars = c("Cook", "Studentized",
    "Bonf", "hat"))
```



Obrázek 6: Analýza reziduí pomocí funkce `infIndexPlot()` – metoda malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování).*

B. Testování normality

Existuje celá řada testů zabývající se normalitou.

Jedna skupina testů je založena na empirických distribučních funkcích, jako zástupce můžeme uvést Kolmogorův–Smirnovův test, popř. Shapiro–Wilksův test.

Další testy jsou založeny na momentových charakteristikách, především na šikmosti či špičatosti. Příkladem může být d'Agostinův test, popř. Jarque–Bera test.

V základním balíku R-base najdeme dva známé testy normality: Shapiro-Wilkův test a Kolmogorův-Smirnovův test.

Shapiro–Wilkův test pro testování normality

Shapiro–Wilkův test je založen na statistice

$$W = \frac{\left(\sum\limits_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2},$$

kde $X_{(i)}$ jsou pořádkové statistiky a a_i jsou váhy, které jsou odvozeny ze středních hodnot a varianční matice pořádkových statistik prostého náhodného výběru z $N(0,1)$ rozsahu n. Tyto hodnoty bývají tabelovány.

Na testovou statistiku W lze pohlížet jako na korelaci mezi pozorovanými hodnotami a jejich normálními skóry.

Testová statistika dosahuje hodnoty 1 v případě, že data vykazují perfektní shodu s normálních rozdělením. Je–li W statisticky významně nižší než 1, zamítáme nulovou hypotézu o shodě s normálním rozdělením.

V knihovnách `tseries`, popř. `FitAR` lze najít tzv. Jarque–Bera test normality, který je založen na výběrové šikmosti a špičatosti.

Jarque–Bera test pro testování normality

Označme	výběrový průměr	$\bar{X} = M_1 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$
	výběrový k-tý centrální moment	$M_k = \frac{1}{n}\sum\limits_{i=1}^{n} (X_i - \bar{X})^k$
	výběrová šikmost	$B_1 = \frac{M_3}{M_2^{3/2}}$
	výběrová špičatost	$B_2 = \frac{M_4}{M_2^2}$

Pak pro Jarque–Bera statistiku platí

$$JB = \frac{n}{6}\left\{B_1^2 + \frac{B_2 - 3}{4}\right\} \overset{A}{\sim} \chi^2(2)$$

Příklad 1 (pokračování): Návštěvnost v hromadných ubytovacích zařízeních v ČR v letech 2000–2010

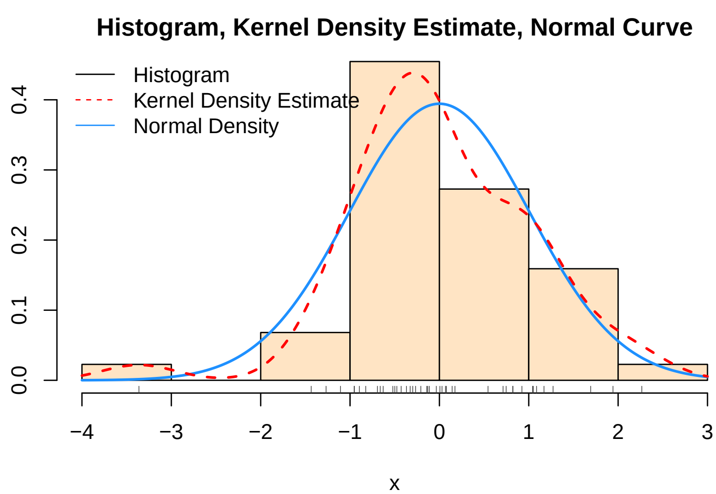
Testy normality budeme provádět na standardizovaných reziduích modifikovaného modelu malého trendu.

```
> library(tseries)
> res.standard <- rstandard(vysl$model)
> shapiro.test(res.standard)
```

```
        Shapiro-Wilk normality test

data:  res.standard
W = 0.957, p-value = 0.1007


> jarque.bera.test(res.standard)


        Jarque Bera Test

data:  res.standard
X-squared = 5.1166, df = 2, p-value = 0.07743
```

Normalitu sice nezamítáme, ale přesto výsledky testu nejsou příliš přesvědčivé. Podíváme se ještě graficky, co způsobila 3 odlehlá pozorování.

```
> par(mfrow = c(1, 1), mar = c(2, 2, 2, 0) + 0.05)
> HistFit(res.standard)
```



Obrázek 7: Grafické testování normality u standardizovaných reziduí metody malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování)*

C. Testování homoskedascity rozptylu

K testování homoskedascity rozptylu se často používá Breusch–Paganův test. Použijeme variantu vhodnou pro časové řady.

Test je založen na myšlence uvažovat variabilitu reziduí jako regresní model ve tvaru

$$\log \sigma_t^2 = a + bt$$

a následně testovat

$$\boxed{H_0} : b = 0 \quad vs \quad \boxed{H_1} : b \neq 0$$

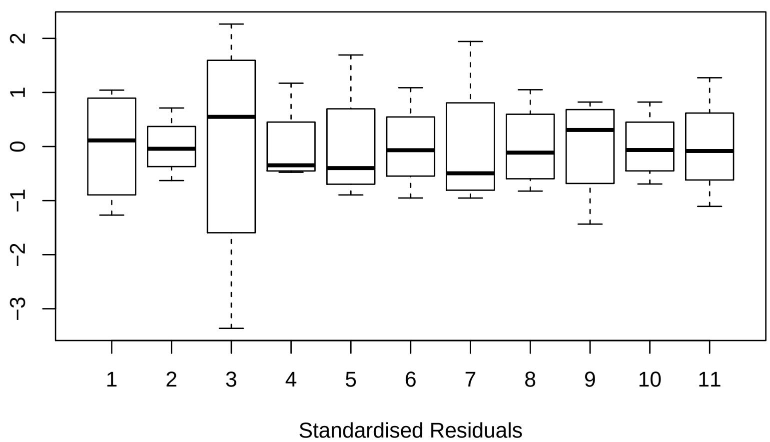
Příklad 1 (pokračování): **Návštěvnost v hromadných ubytovacích zařízeních v ČR v letech 2000–2010**

Testování homoskedascity budeme provádět na standardizovaných reziduích modifikovaného modelu malého trendu.

```
> library(lmtest)
> Time <- 1:length(res.standard)
> bptest(res.standard ~ Time)

        studentized Breusch-Pagan test

data:  res.standard ~ Time
BP = 1.086, df = 1, p-value = 0.2974
```

Homoskedascita rozptylu nebyla zamítnuta, podívejme se ještě graficky na tento problém.

```
> par(mfrow = c(1, 1), mar = c(5, 2, 2, 0) + 0.05)
> boxplotSegments(res.standard, seglen = 4, xlab = "Standardised Residuals")
```



Obrázek 8: Grafické testování homoskedascity u standardizovaných reziduí metody malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování)*

D. Testování nezávislosti reziduí

Pokud je regresní model vyhovující, rezidua by měla být přibližně normálním bílým šumem.

K testování bílého šumu se používají tzv. **Portmanteaovy statistiky**, nejčastěji Ljung–Boxův nebo Box–Piercův statistiky přibližně s χ^2 rozdělením založené na vztazích

$$
\begin{aligned}
BP &= n \sum_{i=1}^{h} r_i^2 \\
LB &= n(n+2) \sum_{i=1}^{h} \frac{r_i^2}{n-i}
\end{aligned}
$$

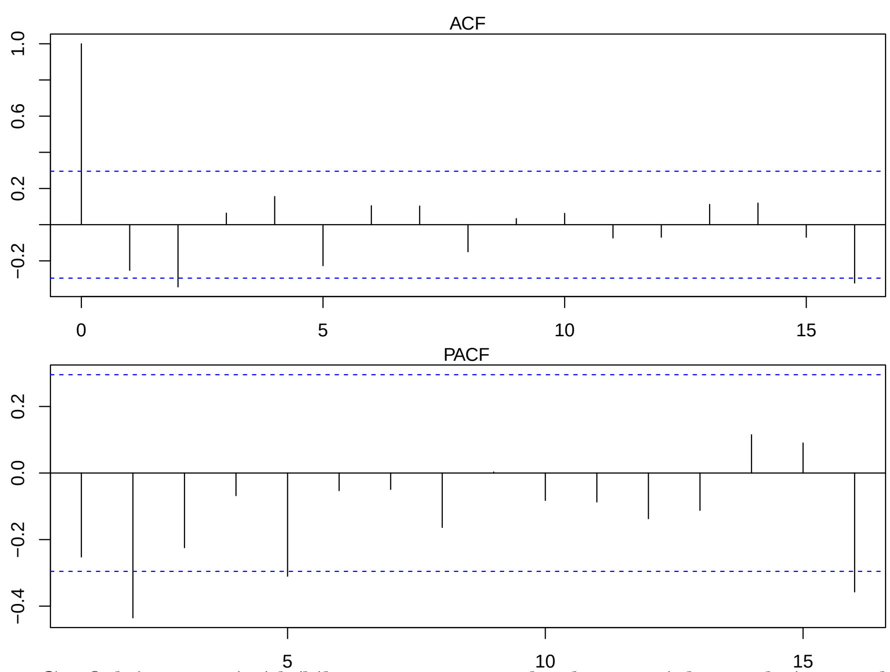
PŘÍKLAD 1 (POKRAČOVÁNÍ): **Návštěvnost v hromadných ubytovacích zařízeních v ČR v letech 2000–2010**

Testování reziduí jako bílého šumu budeme provádět na standardizovaných reziduích modifikovaného modelu malého trendu.

```
> Box.test(res.standard)

        Box-Pierce test

data:  res.standard
X-squared = 2.7996, df = 1, p-value = 0.09429

> Box.test(res.standard, type = "Ljung-Box")

        Box-Ljung test

data:  res.standard
X-squared = 2.995, df = 1, p-value = 0.08352
```

Ani jeden test nezamítl hypotézu, že jde o bílý šum. Tento výsledek by měl potvrdit graf ACF a PACF.

```
> par(mfrow = c(2, 1), mar = c(2, 2, 1, 0) + 0.05)
> acf(res.standard)
> mtext("ACF")
> acf(res.standard, type = "partial")
> mtext("PACF")
```



Obrázek 9: Grafické testování bílého šumu u standardizovaných reziduí metody malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování)*

E. Autokorelace reziduí

V regresních modelech pro časové řady je třeba věnovat velkou pozornost problematice autokorelovaných reziduí. Ve většině případů se u časových řad s autokorelací reziduí setkáme, neboť hodnota pozorování v časovém okažiku t velmi pravděpodobně ovlivní následující hodnoty.

Pro testování autokorelace reziduí prvního řadu je používán Durbin–Watsonův test

Durbin–Watsonův test autokorelace reziduí 1. řádu

Durbin-Watsonova statistika je definována vztahem

$$D = \frac{\sum_{i=2}^{n} (r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2}.$$

Protože platí $(a-b)^2 \leq 2a^2 + 2b^2$, dostáváme

$$D \leq \frac{2\sum_{i=2}^{n} r_i^2 + 2\sum_{i=2}^{n} r_{i-1}^2}{\sum_{i=1}^{n} r_i^2} \leq 4 \qquad \Rightarrow \qquad \boxed{0 \leq D \leq 4}.$$

Vzhledem k tomu, že $Er = 0$, bude pro větší hodnoty n platit

$$\sum_{i=2}^{n} r_i^2 \doteq \sum_{i=1}^{n} r_i^2 \doteq \sum_{i=1}^{n-1} r_{i+1}^2.$$

Označme <u>výběrový autokorelační koeficient</u>:

$$\hat{\rho}(1) = \frac{\widehat{E}(r_i r_{i+1})}{\sqrt{\widehat{D}r_i \widehat{D}r_{i+1}}} = \frac{\sum_{i=1}^{n-1} r_{i+1} r_i}{\sqrt{\sum_{i=1}^{n-1} r_i^2 \sum_{i=1}^{n-1} r_{i+1}^2}} \Rightarrow D \approx 2(1 - \hat{\rho}_1) \quad \text{nebo} \quad \hat{\rho}(1) \approx 1 - \tfrac{D}{2}.$$

Pokud budou **rezidua málo korelovaná**, hodnota D se bude pohybovat **kolem 2**.

Kladná korelace způsobí, že $D \in (0,2)$ a **záporná korelace** způsobí, že $D \in (2,4)$.

Přesné rozdělení statistiky D závisí na tvaru matice plánu $\mathbf{X}$, proto jsou tabelovány intervaly d_L a d_U, ve kterých se nachází kritické hodnoty (pro různá n, k a α).

Dolní a horní hranice Durbin-Watsonova testu na 5% hladině významnosti										
	k=1		k=2		k=3		k=4		k=5+	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
100+	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

kde k je počet nezávisle proměnných v regresní rovnici.

Pro **rychlé posouzení autokorelace prvního řadu** vystačíme s následující tabulkou:

Pokud hodnota Durbin-Watsonovy statistiky D bude v mezích				
0 až d_L	d_L až d_U	d_U až $(4-d_U)$	$(4-d_U)$ až $(4-d_L)$	$(4-d_L)$ až 4
Zamítáme H_0 KLADNÁ autokorelace	Ani nezamítáme ani nepřijímáme H_0	Nezamítáme nulovou hypotézu H_0	Ani nezamítáme ani nepřijímáme H_0	Zamítáme H_0 NEGATIVNÍ autokorelace

Příklad 1 (pokračování): **Návštěvnost v hromadných ubytovacích zařízeních v ČR v letech 2000–2010**

Testování autokorelace reziduí budeme provádět na standardizovaných reziduích modifikovaného modelu malého trendu.

```
> library(lmtest)
> Time <- 1:length(res.standard)
> (DWtest <- dwtest(res.standard ~ Time, alternative = "two.sided"))


        Durbin-Watson test

data:  res.standard ~ Time
DW = 2.4615, p-value = 0.1608
alternative hypothesis: true autocorelation is not 0
```

Vidíme, že v tomto případě se autokorelace reziduí prvního řádu neprokázala.

Abychom to lépe pochopili, je třeba si uvědomit, že pokud pro rezidua (standardizovaná s nulovou střední hodnotou) platí autokorelace prvního řádu, pak můžeme psát

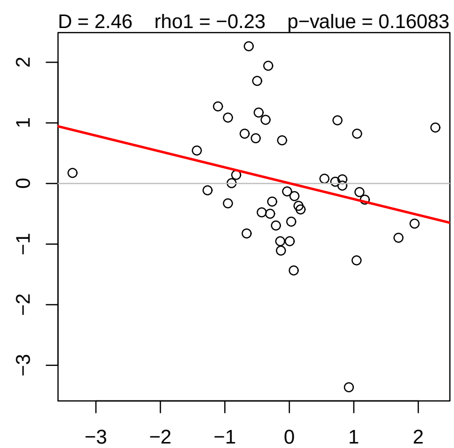
$$r_i = \varphi r_{i-1} + \eta_i.$$

Proto použijeme k prozkoumání tohoto vztahu jednoduchý regresní model. I když předchozí regresní vztah neobsahuje konstantní člen, je lepší ho nevynechávat kvůli interpretaci koeficientu determinace. Absolutní člen (označovaný `(Intercept)`) by se neměl výzamně lišit od nuly.

```
> n <- length(res.standard)
> x <- res.standard[1:(n - 1)]
> y <- res.standard[2:n]
> m.res <- lm(y ~ x)
> summary(m.res)


Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1253 -0.5870 -0.1440  0.6574  2.0957

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.004374   0.152233   0.029   0.9772
x           -0.261747   0.153340  -1.707   0.0954 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9978 on 41 degrees of freedom
Multiple R-squared: 0.06635,        Adjusted R-squared: 0.04358
F-statistic: 2.914 on 1 and 41 DF,  p-value: 0.0954
```

Z výsledků je patrné, že podle očekávání absolutní člen se významně neliší od nuly. Ani směrnice není významná (na 5% hladině). Také p–hodnota F testu je velmi malá, koeficient derminace velmi nízký, proto Durbin–Watsonův test nezamítl hypotézu, že rezidua jsou nekorelovaná. Výsledky znázorníme graficky.

```
> par(mfrow = c(1, 1), mar = c(2, 2, 1, 0) + 0.05)
> txt <- paste("D =", round(DWtest$statistic, 2), "   rho1 =",
     round(1 - 0.5 * DWtest$statistic, 2), "   p-value =",
     round(DWtest$p.value, 5))
> plot(x, y)
> abline(h = 0, col = "gray")
> abline(m.res, col = "red", lwd = 2)
> mtext(txt)
```



Obrázek 10: Grafické testování autokorelace u standardizovaných reziduí metody malého trendu pro data *Návštěvnost (v tisících) v hromadných ubytovacích zařízeních v ČR v letech 2000–2010 (počet přenocování)*

F. Úkol:

Pro všechny regresní modely, které jste dříve vytvořili, proveďte analýzu reziduí z hlediska

- vybočujících pozorování
- normality
- nezávislosti
- heteroskedascity
- autokorelace reziduí