M5201 – 11. CVIČENÍ:
Ověřování vhodnosti modelů pomocí analýzy reziduí a vhodných testů

1 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.

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

Vyrovnané hodnoty na základě odhadů metodou nejmenších čtverců dostaneme ze vztahu:

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

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

Díky symetrii platí:

$$Y_i = \mathbf{H}_i'\mathbf{Y} = \sum_{j=1}^{n} h_{ij}Y_j$$

a odtud vidíme, jak pozorování Y_j ovlivňuje skrze $h_{ij}i$-tou vyrovnanou 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: $\boxed{\|\mathbf{H}_j\|^2 = h_{jj}}$.

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

Věta 3: $\boxed{tr\mathbf{H} = k}$

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

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

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

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

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 $s^2_{(i)} = \frac{(\mathbf{Y}_{(i)}-\widehat{\mathbf{Y}}_{(i)})'(\mathbf{Y}_{(i)}-\widehat{\mathbf{Y}}_{(i)})}{n-k-1}$.
- **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 nejmenších 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 $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \Rightarrow \boxed{\mathbf{r} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I}-\mathbf{H}))}$.

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

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

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

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

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

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

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

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

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

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

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

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()`	

PŘÍKLAD 1: **Koncentrace oxidu uhličitého v okolí sopky Mauna Loa**

V datovém souboru `CO2.dat` jsou obsažena měsíční data, která udávají koncentraci oxidu uhličitého v okolí havajské sopky Mauna Loa od ledna 1965 do prosince 1980. Na prvním řádku souboru je uveden popis dat a data jsou pod ním ve dvanácti sloupcích.

Nejjednodušší způsob, jak datový soubor načíst, je použít příkaz `scan()`. Při načítání bude nutné přeskočit první řádek, tudíž přidáme argument `skip=1`.

```
> fileDat <- paste(data.library, "CO2.dat", sep = "")
> co2 <- scan(fileDat, skip = 1)
```

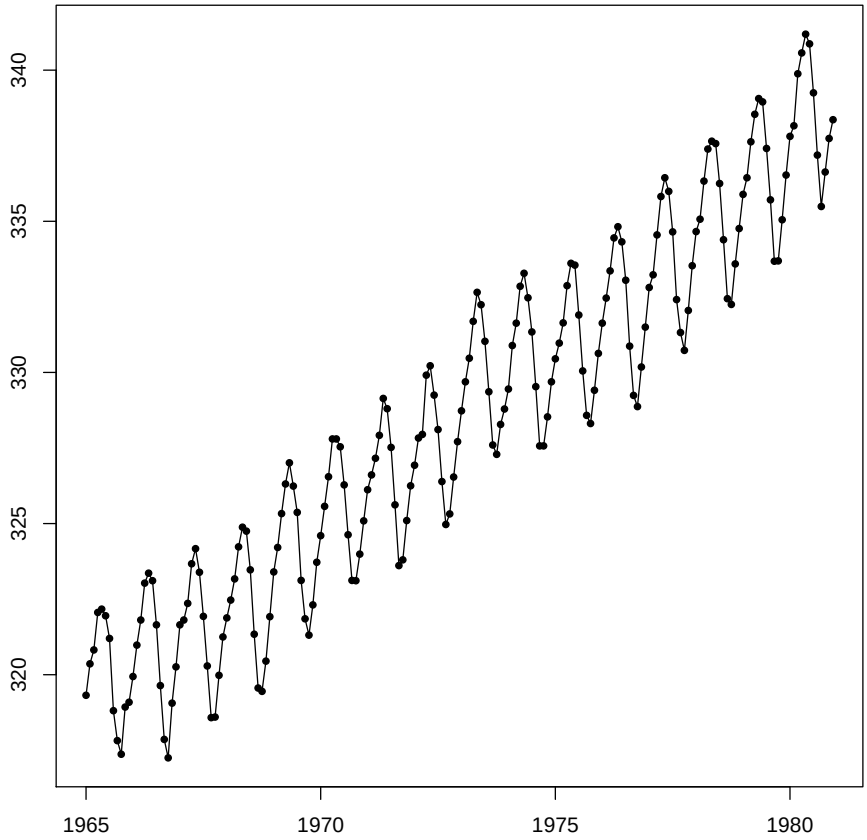
Příkazem `scan` jsme data načetli do vektoru. Pro další práci bude výhodné data převést na časovou řadu (pomocí `ts()`). U nově vytvořené časové řady si prohlédneme její strukturu.

```
> TS <- ts(co2, start = 1965, frequency = 12)
> str(TS)
```

```
 Time-Series [1:192] from 1965 to 1981: 319 320 321 322 322 ...
```

Časovou řadu vykreslíme pomocí příkazu `plot()`.

```
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(TS, type = "o", pch = 20, cex = 3)
```



Obrázek 1: *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980*

Protože data hned na první pohled jasně vykazují sezónní chování, bude třeba k jejich dekompozici použít model zahrnující sezónnost. Pro ten účel zvolíme například (modifikovanou) metodu malého trendu

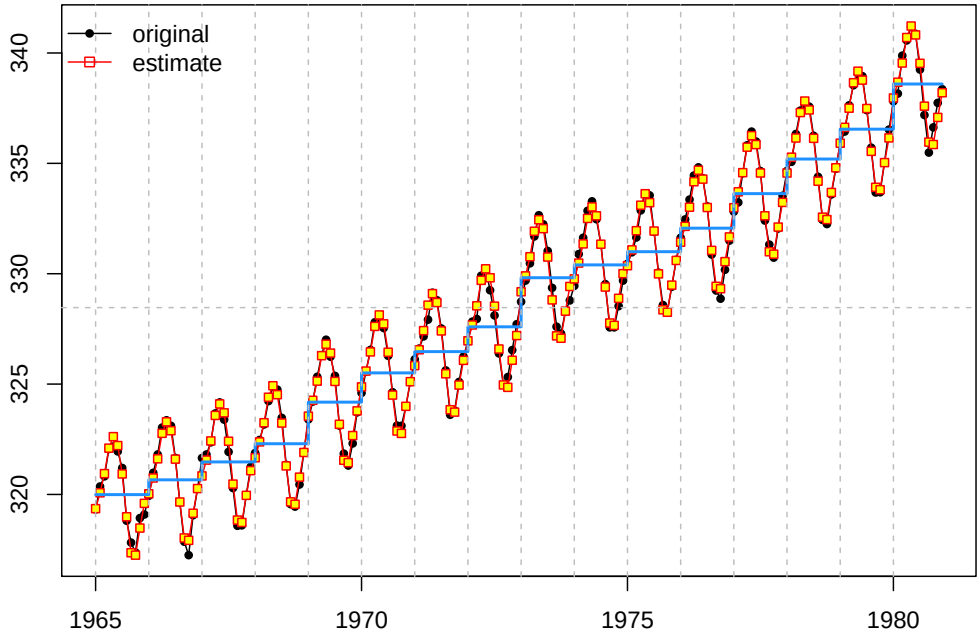
$$\boxed{M_I^{modif}}: \; Y_{jk} = \mu + m_j + s_k + \varepsilon_{jk} \qquad \varepsilon_{jk} \sim WN(0, \sigma^2), \;\; j = 1, \ldots, r, \;\; k = 1, \ldots, d$$

s dodatečnými podmínkami

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

Dekompozici provedeme pomocí funkce `SzSmallTrendModif()`.

```
> fileFun <- paste(data.library, "FunkceM5201.R", sep = "")
> source(fileFun)
> vysl <- SzSmallTrendModif(TS)
```



Obrázek 2: Metoda malého trendu pro data *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980.*

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
-0.67042 -0.17260 -0.00146  0.18635  0.81125

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 328.46396    0.02054 15989.952   <2e-16 ***
grY1         -8.47229    0.07956  -106.491   <2e-16 ***
grY2         -7.80146    0.07956   -98.059   <2e-16 ***
grY3         -6.99063    0.07956   -87.868   <2e-16 ***
grY4         -6.16646    0.07956   -77.509   <2e-16 ***
grY5         -4.28229    0.07956   -53.826   <2e-16 ***
grY6         -2.95729    0.07956   -37.171   <2e-16 ***
grY7         -1.99312    0.07956   -25.052   <2e-16 ***
grY8         -0.86979    0.07956   -10.933   <2e-16 ***
grY9          1.35437    0.07956    17.024   <2e-16 ***
grY10         1.93604    0.07956    24.335   <2e-16 ***
grY11         2.53354    0.07956    31.845   <2e-16 ***
grY12         3.59854    0.07956    45.231   <2e-16 ***
grY13         5.16354    0.07956    64.903   <2e-16 ***
grY14         6.73187    0.07956    84.615   <2e-16 ***
```

```
grY15         8.08437    0.07956   101.616   <2e-16 ***
grS1         -0.63458    0.06813    -9.314   <2e-16 ***
grS2          0.08292    0.06813     1.217    0.225
grS3          0.95104    0.06813    13.959   <2e-16 ***
grS4          2.10542    0.06813    30.903   <2e-16 ***
grS5          2.62667    0.06813    38.554   <2e-16 ***
grS6          2.22292    0.06813    32.628   <2e-16 ***
grS7          0.93667    0.06813    13.748   <2e-16 ***
grS8         -1.00458    0.06813   -14.745   <2e-16 ***
grS9         -2.63333    0.06813   -38.652   <2e-16 ***
grS10        -2.74208    0.06813   -40.248   <2e-16 ***
grS11        -1.51458    0.06813   -22.231   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2846 on 165 degrees of freedom
Multiple R-squared: 0.998,        Adjusted R-squared: 0.9977
F-statistic:  3217 on 26 and 165 DF,  p-value: < 2.2e-16
```

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

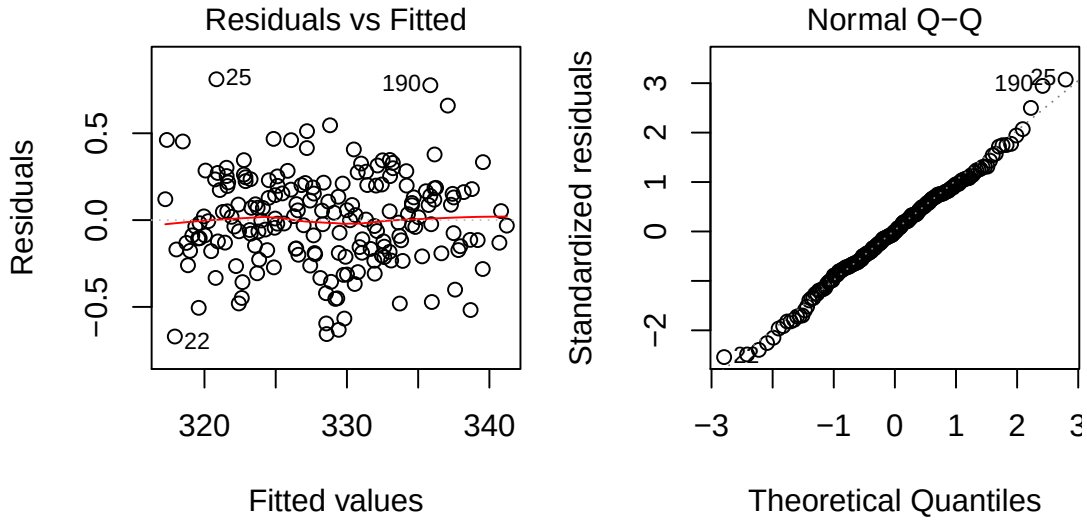
```
Analysis of Variance Table

Response: x
           Df Sum Sq Mean Sq F value    Pr(>F)
grY        15 6195.3  413.02 5097.88 < 2.2e-16 ***
grS        11  582.1   52.91  653.12 < 2.2e-16 ***
Residuals 165   13.4    0.08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nyní, když máme pro data odhadnutý model, bychom chtěli ověřit, zda se pro naše data skutečně hodí, a také to, zda některá data nemají neobvykle velký vliv na odhady parametrů modelu. Proto postupně provedeme analýzu reziduí. Nejprve začneme tím, že se pomocí regresní diagnostiky pokusíme najít případná odlehlá a vybočující pozorování.

Základní diagnostické grafy dostaneme v R tak, že na odhadnutý model zavoláme funkci `plot.lm()`. Stačí psát zkrácenou verzi `plot()`. Příkaz vykresluje šest různých diagnostických grafů, zajímá-li nás pouze některý z nich, použijeme argument `which="`*číslo grafu*`"`.

Prohlédněme si nejprve první dva grafy.

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



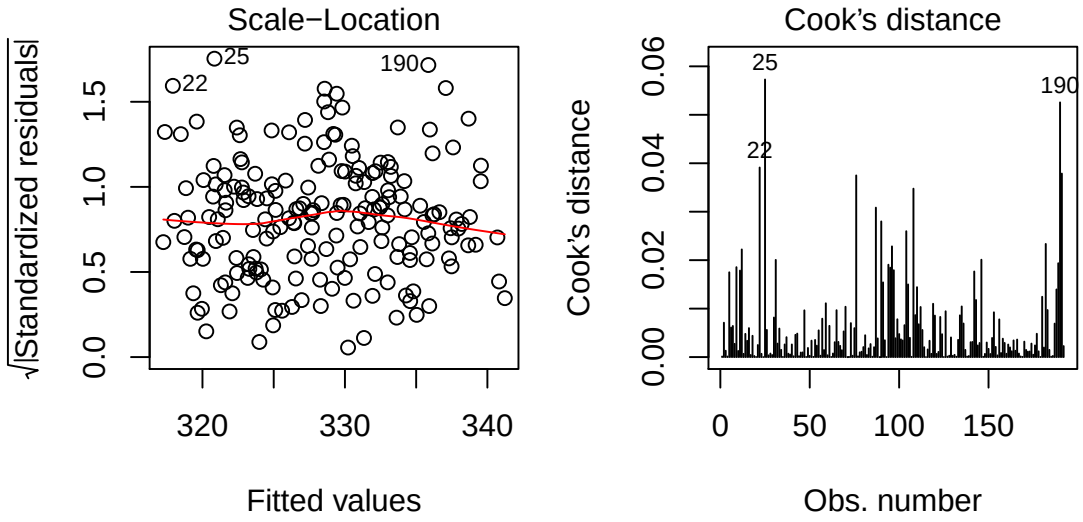
Obrázek 3: Analýza reziduí pomocí funkce `plot()` – grafy 1 a 2 u metody malého trendu pro data *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980.*

Na prvním grafu jsou znázorněny na x-ové ose vyrovnané hodnoty a na ose y proti nim příslušná rezidua. Rezidua by neměla být nijak závislá na velikosti odpovídajících vyrovnaných hodnot, tudíž by neměla vykazovat žádný trend a měla by kolísat kolem nuly. V grafu je zobrazena pomocná červená linie, která představuje neparametrický odhad střední hodnoty reziduí. První diagnostický graf pro naše rezidua vypadá přibližně tak, jak bychom očekávali.

Druhým diagnostickým grafem je Q-Q plot, který slouží k vizuálnímu posouzení pravděpodobnostního rozložení zkoumaných dat. U reziduí nás obvykle zajímá, zda mají normální rozložení. V tom případě by body Q-Q plotu (konstruovaného pro normální rozdělení) měly ležet přibližně na přímce. Podíváme-li se na Q-Q plot pro naše rezidua, vidíme, že body na přímce zhruba leží, takže nic nenasvědčuje tomu, že by rezidua nemohla mít normální rozdělení.

Prohlédněme si další dva diagnostické grafy.

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



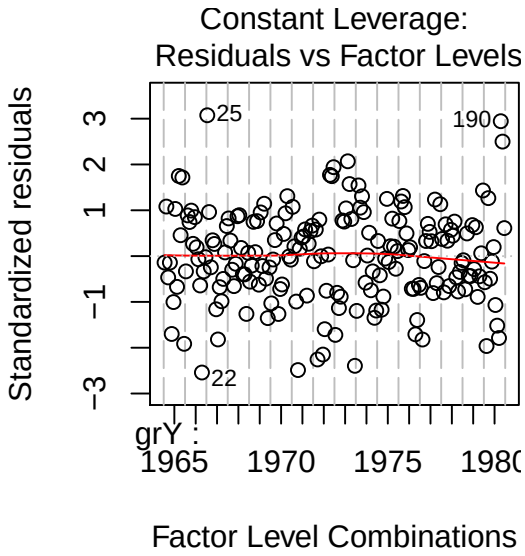
Obrázek 4: Analýza reziduí pomocí funkce `plot()` – grafy 3 a 4 u metody malého trendu pro data *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980*

Na třetím grafu vidíme odmocniny ze standardizovaných reziduích v závislosti na vyrovnaných hodnotách. Opět by mělo platit (podobně jako u prvního grafu), že by odmocniny ze standardizovaných reziduí neměly vykazovat závislost na vyrovnaných hodnotách.

Čtvrtým grafem je graf Cookových vzdáleností $\widehat{\mathbf{Y}}$ od $\widehat{\mathbf{Y}}_{(i)}$. Čím větší má některé pozorování Cookovu vzdálenost, tím větší je jeho vliv na odhady regresních koeficientů. V grafu jsou čísly označena pozorování 22, 25 a 190 s největšími Cookovými vzdálenostmi, která se nabízí jako vlivná pozorování. Z grafu však vidíme, že největší hodnota Cookovy vzdálenosti je $0,06$, což je velmi malá hodnota v porovnání s hodnotou 1, která se orientačně používá jako dolní mez pro vlivná pozorování.

Ještě se podívejme na pátý diagnostický graf, který znázorňuje závislost reziduí na faktorech vysvětlující proměnné.

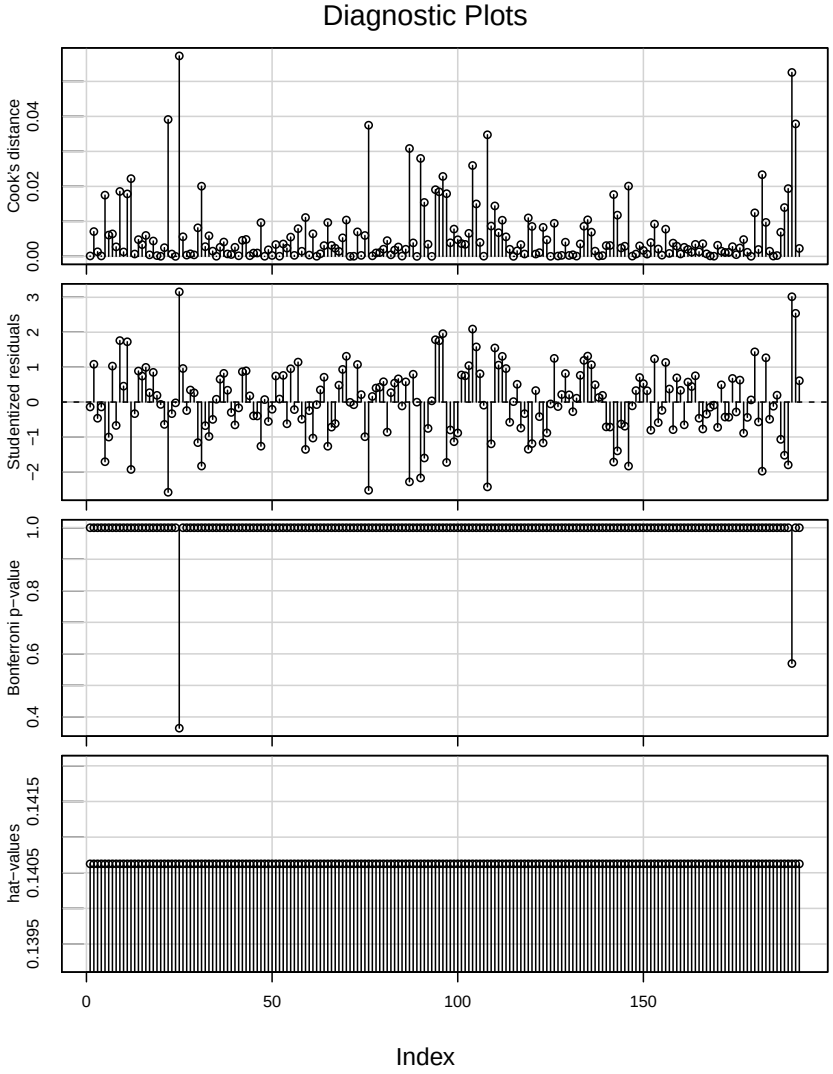
```
> par(mfrow = c(1, 1), mar = c(5, 5, 2, 0) + 0.05)
> plot(vysl$model, which = 5)
```



Obrázek 5: Analýza reziduí pomocí funkce `plot()` – graf 5 u metody malého trendu pro data *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980*

Další zajímavé grafy lze najít v knihovně `car` (příkazy `influencePlot()`, `infIndexPlot()`). První z nich je určen spíše pro klasickou regresi, druhý je již vhodnější pro naše účely, proto ho použijeme na náš model. Argumentem `vars` můžeme určovat, které grafy chceme vykreslit. Zvolíme graf Cookových vzdáleností, studentizovaných reziduí, Bonferroniho p-hodnot a vlivů h_{ii}.

```
> 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 *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980.*

První graf s Cookovými vzdálenostmi je prakticky totožný se čtvrtým diagnostickým grafem funkce `plot.lm()`.

Na druhém grafu vidíme studentizovaná rezidua. Pokud je studentizované reziduum v absolutní hodnotě větší než 2, jemu odpovídající pozorování se považuje za odlehlé. V grafu najdeme takových reziduí hned několik (přibližně 10).

Mezní hodnota 2 pro posouzení odlehlosti pozorování na základě studentizovaných reziduí je odvozena na základě testu s 5% hladinou významnosti, z jehož povahy vyplývá, že vždy 5% největších reziduí přesáhne tuto hranici. Bude se sice jednat o 5% „nejextrémnějších“ pozorování, ale neznamená to, že všechna jsou skutečně výjimečná. Abychom zjistili, která z nich jsou skutečně odlehlá, použijeme Bonferroniho test. Jeho p-hodnoty jsou ve třetím grafu. Pokud je p-hodnota menší než zvolená hladina významnosti (obvykle $0,05$), považuje se odpovídající pozorování za odlehlé. Z grafu vidíme,

že p-hodnota není u žádného pozorování menší než $0,05$, tudíž žádné pozorování není skutečně odlehlé.

V posledním grafu jsou vykresleny vlivy h_{ii} (diagonální prvky projekční matice $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$). Ideálně by měly být vlivy všech pozorování přibližně stejné a pohybovat se kolem k/n. V grafu hledáme hlavně pozorování s neobvykle velkým vlivem. Pro náš model bychom takové pozorování v grafu našli těžko.

2 Testování normality

V klasickém regresním modelu se předpoládá, že náhodné odchylky ε_i mají normální rozdělení. 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–Wilkův 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.

2.1 Shapiro–Wilkův test pro testování normality

Shapiro–Wilkův test je založen na statistice

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

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.

Hlavní myšlenka stojící za testem spočívá v tom, že zjišťujeme, zda body v Q-Q plotu pro testovaná data jsou významně odlišné od regresní přímky proložené těmito body.

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.

2.2 Jarque–Bera test pro testování normality

Jedná se o test založený na porovnání výběrové šikmosti a špičatosti s teoretickou šikmostí a špičatostí pro normální rozdělení.

Označme	výběrový průměr	$\bar{X} = M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$
	výběrový k-tý centrální moment	$M_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$
	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): **Koncentrace oxidu uhličitého v okolí sopky Mauna Loa**

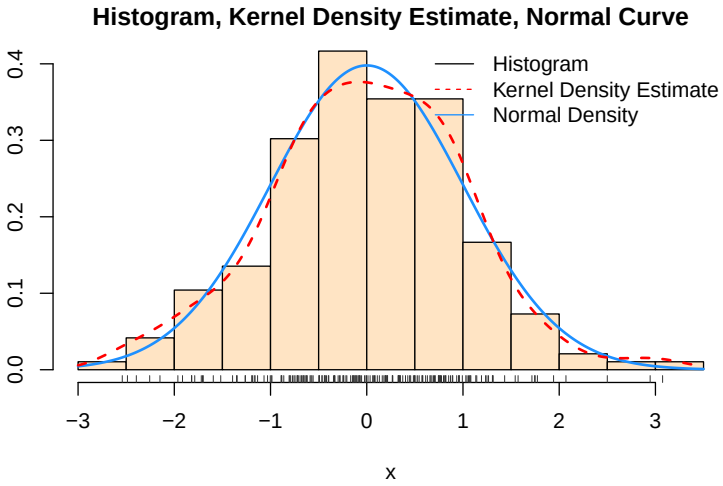
Testy normality budeme provádět na standardizovaných reziduích modifikovaného modelu malého trendu. Ta získáme příkazem `rstandard()`. Pro Shapiro-Wilkův test normality standardizovaných reziduí použijeme funkci `shapiro.test()`, Jarque-Berův test provedeme příkazem `jarque.bera.test()`.

```
> library(tseries)


    'tseries' version: 0.10-25

    'tseries' is a package for time series analysis and
    computational finance.

    See 'library(help="tseries")' for details.


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


        Shapiro-Wilk normality test

data:  res.standard
W = 0.9934, p-value = 0.5531


> jarque.bera.test(res.standard)


        Jarque Bera Test

data:  res.standard
X-squared = 0.5919, df = 2, p-value = 0.7438
```

U prvního testu jsme dostali p-hodnotu $0,55$, u druhého dokonce $0,74$. Obě čísla jsou výrazně větší než hladina významnosti testu $0,05$, tudíž nezamítáme hypotézu o tom, že standardizovaná rezidua mají normální rozdělení.

Rozložení standardizovaných reziduí si můžeme prohlédnout, když si vykreslíme jejich histogram. Když k vykreslení použijeme funkci `HistFit()`, vykreslí se nám do grafu kromě samotného histogramu také jádrový odhad hustoty reziduí a hustota normálního rozdělení se střední hodnotou a rozptylem, které dostaneme jako odhad střední hodnoty a rozptylu reziduí.

```
> 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 *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980*

3 Testování homoskedasticity rozptylu

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

Jestliže jsou rezidua homoskedastická, jejich rozptyl je v čase konstantní. Homoskedasticita je porušena například v případě, kdy rozptyl reziduí σ_t^2 s časem lineárně roste nebo klesá. Breusch-Paganův test je založen na myšlence popsat variabilitu reziduí pomocí regresního modelu ve tvaru

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

a následně testovat hypotézu

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

Pokud zamítneme nulovou hypotézu, prokázali jsme, že rozptyl reziduí není konstantní a tudíž jsou rezidua heteroskedastická.

PŘÍKLAD 1 (POKRAČOVÁNÍ): **Koncentrace oxidu uhličitého v okolí sopky Mauna Loa**

Testování homoskedascity budeme opět provádět na standardizovaných reziduích modifikovaného modelu malého trendu. Breush-Paganův test homoskedaticity lze v R provést příkazem `bptest()`, který najdeme v balíčku `lmtest`.

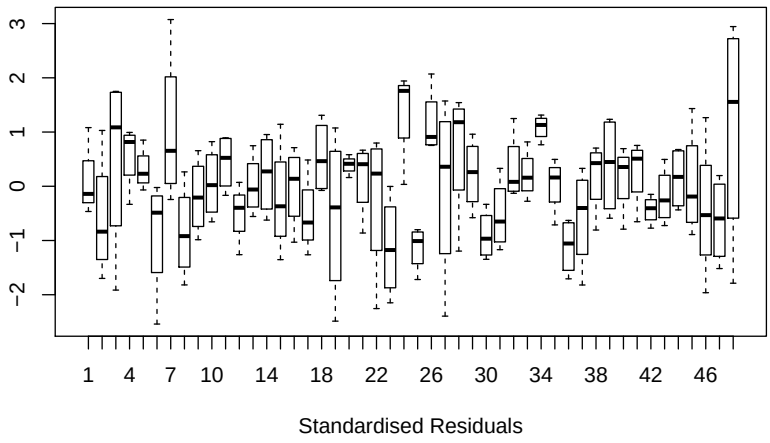
```
> library(lmtest)
> Time <- 1:length(res.standard)
> bptest(res.standard ~ Time)
```

```
        studentized Breusch-Pagan test

data:  res.standard ~ Time
BP = 0.0191, df = 1, p-value = 0.8902
```

Homoskedasticita rozptylu nebyla zamítnuta, protože p-hodnota testu $0,89$ je větší než hladina významnosti $0,05$.

Podívejme se na problém homoskedasticity ještě s využitím grafu. Použitím příkazu `boxplotSegments()` dostaneme krabicové grafy pro krátké úseky dat o čtyřech pozorováních. Z jejich rozměrů lze vidět, jak se variabilita reziduí mění v čase.

```
> 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 homoskedasticity u standardizovaných reziduí metody malého trendu pro data *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980*

4 Testování nezávislosti reziduí

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

K testování bílého šumu se používají tzv. **Portmanteauovy statistiky**, nejčastěji Ljung–Boxova nebo Box–Piercova statistika. Mají přibližně χ^2 rozdělení a jsou 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): **Koncentrace oxidu uhličitého v okolí sopky Mauna Loa**

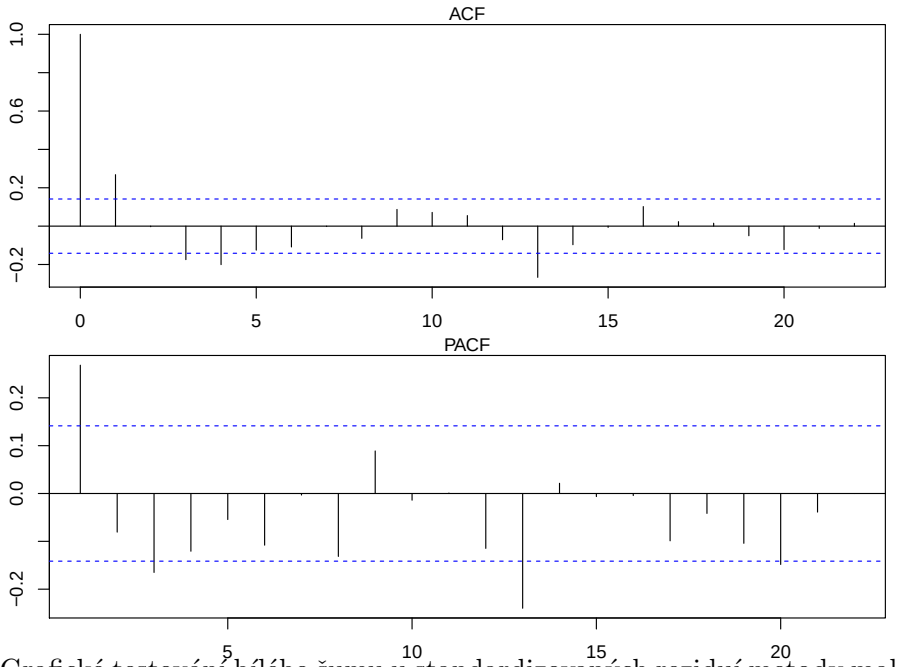
Testování reziduí jako bílého šumu budeme provádět na standardizovaných reziduích modifikovaného modelu malého trendu. K provedení Box-Piercova testu je v R funkce `Box.test()`. Chceme-li provést Ljung-Boxův test, použijeme tutéž funkci a jako její argument uvedeme `type="Ljung-Box"`.

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


        Box-Pierce test

data:  res.standard
X-squared = 13.8135, df = 1, p-value = 0.0002019
```

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


        Box-Ljung test

data:  res.standard
X-squared = 14.0305, df = 1, p-value = 0.0001799
```

Oba testy dávají jako výsledek p-hodnoty výrazně menší než hladina významnosti $0,05$, a proto v obou případech zamítáme nulovou hypotézu, že rezidua jsou bílým šumem. Tento výsledek by se měl rovněž projevit v grafech autokorelační a parciální autokorelační funkce.

```
> 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 *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980*

Pokud by byla rezidua bílým šumem, měla by ACF významně nenulovou hodnotu pouze pro $k = 0$ a PACF by měla všechny hodnoty přibližně nulové.

5 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

5.1 Durbin–Watsonův test autokorelace reziduí 1. řádu

Durbin-Watsonova statistika je definována vztahem

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

Lze odvodit, že statistika D může nabývat pouze hodnot z intervalu $(0, 4)$. 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): **Koncentrace oxidu uhličitého v okolí sopky Mauna Loa**

Testování autokorelace reziduí budeme provádět na standardizovaných reziduích modifikovaného modelu malého trendu. Durbin-Watsonův test provedeme příkazem `dwtest()` z balíčku `lmtest`. Při použití funkce musíme zadat, že chceme testovat autokorelaci reziduí v závislosti na čase, oboustrannou alternativní hypotézu zadáme argumentem `alternative="two.sided"`.

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

        Durbin-Watson test

data:  res.standard ~ Time
DW = 1.4615, p-value = 0.0001313
alternative hypothesis: true autocorelation is not 0
```

Vidíme, že p-hodnota je výrazně menší než hladina významnosti $0,05$, takže se prokázala autokorelace reziduí prvního řádu.

Abychom lépe pochopili, co představuje autokorelace 1. řádu, je třeba si uvědomit, že pokud pro rezidua (standardizovaná s nulovou střední hodnotou) platí autokorelace prvního řádu, pak můžeme psát

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

kde η_i jsou nekorelované náhodné veličiny s nulovou střední hodnotou.

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. Odhadnutý absolutní člen (označovaný `(Intercept)`) by se ale v tomto případě neměl významně lišit od nuly.

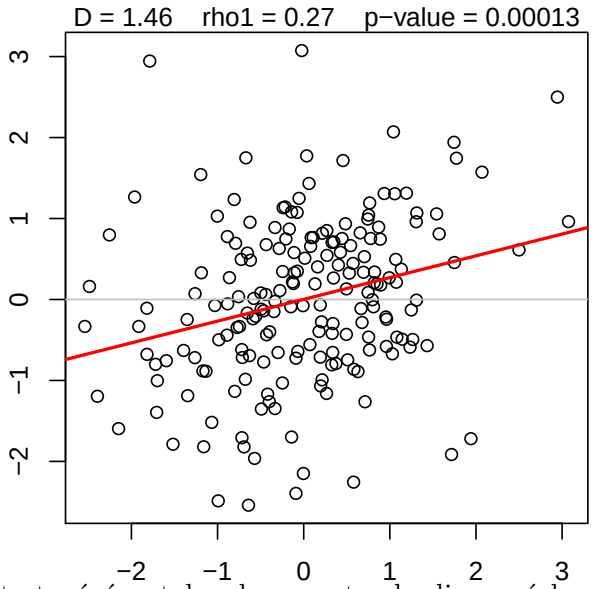
```
> n <- length(res.standard)
> x <- res.standard[1:(n - 1)]
> y <- res.standard[2:n]
> m.res <- lm(y ~ x)
> summary(m.res)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4131 -0.5804 -0.0001  0.6429  3.4238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.001597   0.070248   0.023 0.981889
x           0.268755   0.070133   3.832 0.000173 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9708 on 189 degrees of freedom
Multiple R-squared: 0.0721,        Adjusted R-squared: 0.06719
F-statistic: 14.68 on 1 and 189 DF,  p-value: 0.0001729
```

Z výsledků je patrné, že podle očekávání se absolutní člen významně neliší od nuly. Ale směrnice již významná je (na 5% hladině), protože má p-hodnotu $0,000173$, která je menší než $0,05$. To je také důvod, proč Durbin–Watsonův test zamítl 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 *Koncentrace CO_2 u sopky Mauna Loa v letech 1965–1980*

6 Úkol:

Načtěte si datový soubor `zoo_jihlava.txt` (měsíční údaje o počtu návštěvníků v ZOO Jihlava v letech 2006-2010), na načtená data aplikujte modifikovanou metodu malého trendu a získaná a rezidua analyzujte z hlediska

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