# M6120 – 8. CVIČENÍ : **M6120cv08** (*Metody regresní diagnostiky*)

## 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.

Pokud se zabýváme detekcí neočekávaných hodnot v datech, v zásadě může jít o

- neočekávané hodnoty vysvětlované proměnné, tzv. **odlehlá pozorování** (*outliers*);
- neočekávané hodnoty vektoru vysvětlujících proměnných, tzv. **vybočující body** (*leverage points*);
- v některých případech může jít dokonce o data, jež lze zařadit do obou skupin.

Jejich výskyt nemusí nutně významně ovlivnit analýzu dat, neboť takováto měření mohou být plně ve shodě s předpokládaným modelem.

Většinou je tomu však naopak a odlehlá pozorování i vybočující body hrají významnou roli pro výsledky regresní analýzy.

Na druhé straně i některé další body mohou mít významný vliv ať již na $\widehat{\boldsymbol{\beta}}$, $\widehat{\beta}_i$, $D\boldsymbol{\beta}$, $\widehat{\mathbf{Y}}$ apod.

Všechny body, jež nějakým způsobem ovlivňují podstatně analýzu dat, tj. některou z charakteristik spojených s odhadem vektoru parametrů v lineárním modelu a testováním hypotéz o nich, nazveme **vlivnými body**.

Základním diagnostickým nástrojem v regresní analýze jsou **rezidua**.

Připomeňme definici a vlastnosti regresního modelu a zaveďme celou škálu diagnostických měr influnce.

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

Protože $Y_i = \mathbf{H}_i'\mathbf{Y} = \sum_{j=1}^{n} h_{ij}Y_j$, 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: $\mathbf{r} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \underbrace{(\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta}}_{=\mathbf{0}} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$
$= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$

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

**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:

- **normovaná rezidua** (*normalized or scaled residuals*) $\boxed{r_{Ni} = \frac{r_i}{s}}$, kde $s^2 = \frac{(\mathbf{Y}-\widehat{\mathbf{Y}})'(\mathbf{Y}-\widehat{\mathbf{Y}})}{n-k} = \frac{SSE}{n-k}$.
- **standardizovaná rezidua** (*standardized or internally studentized residuals*) $\boxed{r_{Si} = \frac{r_i}{s\sqrt{1-h_{ii}}}}$
- **predikovaná rezidua** (*predicted or crossvalidated residuals*) $\boxed{r_{P(i)} = Y_i - \hat{Y}_{(i)}}$, 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 rezidua** (*jackknife or externally studentized residuals*) $\boxed{r_{J(i)} = \frac{r_{P(i)}\sqrt{1-h_{ii}}}{s_{(i)}}}$, 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 $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \Rightarrow \boxed{\mathbf{r} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I}-\mathbf{H}))}$. 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:** $\boxed{r_{P(i)} = \frac{r_i}{1-h_{ii}}}$ ; $\boxed{Dr_{P(i)} = \frac{\sigma^2}{1-h_{ii}}}$

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

**Věta 10:** $\boxed{r_{J(i)} = \frac{r_i}{\sqrt{1-h_{ii}}s_{(i)}}}$ Důkaz: viz. Staudte, Sheather(1990).

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

**Věta 11:** Nechť $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \Rightarrow \boxed{r_{J(i)} \sim t(n-k)}$. 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 $r_{J(i)}$ 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 $|r_{J(i)}| \geq t_{1-\alpha/2}(n-k)$, pak na hladině významnosti $\alpha$ zamítáme hypotézu $H_0$ a tedy $i$-té pozorování na hladině významnosti $\alpha$ je odlehlé.

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

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 $\boxed{|r_{J(i)}| \geq 2}$.

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

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

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

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

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* $\sigma^2$, 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 $\sigma^2$, 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 $\boxed{DFBETAS_{ij} = \frac{\hat{\beta}_i - \hat{\beta}_{j(i)}}{\sqrt{D\hat{\beta}_j}}}$ (doporučená hranice $2\sqrt{k/n}$)

**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

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

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")
   ```

## B. STRATEGIE REGRESNÍ DIAGNOSTIKY

Jednotlivé míry regresní diagnostiky poskytují cenné informace o výskytu vlivných pozorování. V žádném případě však nemá smysl používat je automaticky na všechna data, protože bychom byli brzy zavaleni horou výsledků, z převážné většiny naprosto zbytečných.

Mnohem praktičtější je naopak držet se některé strategie regresní diagnostiky. Ta musí být dostatečně pružná, abychom mohli v každém kroku rozhodnout, zda některá pozorování přidat či vyloučit, aplikovat vhodnou transformaci a vrátit se o několik kroků zpět nebo naopak některý krok přeskočit apod. Je zřejmé, že k tomu musíme mít k dispozici vhodné programové vybavení s interaktivním přístupem.

**GRAFY IDENTIFIKACE VLIVNÝCH BODŮ** - existuje velké množství různých grafů, jako příklad uveďme:

**Graf predikovaných reziduí:**

- osa $x$: predikovaná rezidua $r_{P(i)}$
- osa $y$: rezidua $r_i$

Vybočující body (*leverage points*) jsou snadno identifikovány svou polohou, neboť leží mimo přímku $y = x$.

Odlehlá pozorování (*outliers*) leží sice na přímce $y = x$ nebo v její blízkosti, jsou však dostatečně vzdáleny od ostatních bodů.

**Williamsův graf:**

- osa $x$: vlivy $h_{ii}$
- osa $y$: Jackknife rezidua $r_{J(i)}$

Do grafu lze zakreslit mezní linie pro odlehlá pozorování (*outliers*): $y = t_{1-\alpha}(n-k)$ a jednak mezní linie pro vybočující body (*leverage points*): $x = 2\frac{k}{n}$.

**Pregibonův graf:**

- osa $x$: vlivy $h_{ii}$
- osa $y$: kvadráty modifikovaných normovaných reziduí $r^2_{Mi} = \frac{r_i^2}{SSE} = \frac{r_i^2}{(n-k)s^2}$

Protože platí, že $E(h_{ii} + r^2_{Mi}) = \frac{k}{n}$, lze do grafu zakreslit dvě hraniční přímky: $y = -x + 2\frac{k}{n}$ a $y = -x + 3\frac{k}{n}$.
K rozlišení mezi body platí tato pravidla:

(a) bod je silně vlivný, leží-li nad horní přímkou

(b) bod je pouze vlivný, leží-li mezi oběma přímkami

Může jít jak o vybočující body (*leverage points*), tak i o odlehlá pozorování (*outliers*).

**Indexové grafy:**

- osa $x$: index $i$
- osa $y$: jednotlivé typy reziduí, vlivy $h_{ii}$, $\beta_{(i)}$

**Rankitové Q-Q grafy:**

- osa $x$: kvantily standardizovaného normálního rozložení
- osa $y$: pořádkové statistiky (vzestupně setříděné hodnoty reziduí)

**POSTUP PŘI REGRESNÍ DIAGNOSTICE:** lze jich navrhnout celou řadu

(a) Spočteme základní charakteristiky jednotlivých proměnných $Y, X_1, \ldots, X_k$, tj. průměr, směrodatnou odchylku, šikmost atd. a vykreslíme pro ně histogram, krabicový graf, případně indexový graf apod. Odtud získáme představu o možných odlehlých pozorování v jednotlivých proměnných, tvaru dat apod.

(b) Vykreslíme rozptylové grafy (*scatter plots*) dvojic vektorů $(Y, X_i)$, resp. dvojic vektorů $(X_i, X_j)$.

(c) Zkontrolujeme, zda se v datech nevyskytují vybočující body v prostoru regresorů (*leverage points*), např. pomocí diagonálních prvků projekční mastice $H$. Tyto body v žádném případě automaticky nevylučujeme, neboť nemusí jít nutně o vlivné body. Při další analýze jim však věnujeme zvýšenou pozornost.

(d) Provedeme pečlivou analýzu reziduí. Používáme přitom jak grafické zobrazení reziduí (např. proti indexu), tak u „podezřelých" hodnot užijeme k ověření významnosti vlivu některou z dříve uvedených statistik. Získáme tak představu o odlehlých pozorováních.

(e) Po této všeobecné analýze dále postupujeme dle toho, co nás především zajímá. Je-li to vliv jednotlivých pozorování na hodnotu $\hat{\beta}$, užijeme např. Cookovu vzdálenost. Je-li to vliv jednotlivých pozorování na varianční matici $\hat{\beta}$, užijeme statistiku $COVRATIO_i$ apod. V maximální možné míře přitom též využíváme grafická znázornění, jež nám podstatně urychlují orientaci ve výsledcích.

Výše uvedené kroky představují pouze jednu možnou základní strategii regresní diagnostiky. Při jejím užití vzniká v praxi potřeba dalších kroků a analýz. Patří mezi ně různé transformace, změna modelu, vylučování pozorování apod.

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

PŘÍKLAD 1: Simulovaná data

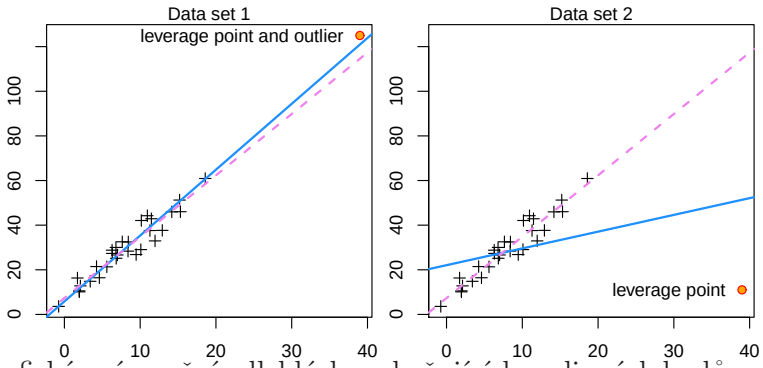
Nejprve na simulovaných datech ukážeme příklad dat, ve kterých se nacházejí pozorování

- **odlehlá** (*outliers*) – $y$-ové hodnoty
- **vybočující** (*leverage points*) – $x$-ové hodnoty
- **vlivná**, jejichž odstraněním se radikálně změní výsledek

```
> k <- 15
> n <- 2 * k - 1
> sigma1 <- 3.5
> xx <- seq(1, k, by = 0.5) + sigma1 * rnorm(n)
> a <- 5
> b <- 3
> sigma2 <- 3.5
> yy <- a + b * xx + sigma2 * rnorm(n)
> x1 <- c(xx, 39)
> y1 <- c(yy, 125)
> x2 <- x1
> y2 <- c(yy, 11)
```

Simulovaná data graficky znázorníme. Doplníme vždy odhadnuté regresní přímky se všemi body (plná čára) a bez posledního bodu (čárkovaně).

```
> xlim <- range(c(x1, x2))
> ylim <- range(c(y1, y2))
> par(mfrow = c(1, 2), mar = c(2, 2, 2, 0.5) + 0.05)
> plot(x1[1:n], y1[1:n], pch = 3, xlim = xlim, ylim = ylim)
> points(x1[n + 1], y1[n + 1], pch = 21, col = "red", bg = "orange")
> abline(lm(y1 ~ x1), col = "dodgerblue", lwd = 2)
> abline(lm(y1[-(n + 1)] ~ x1[-(n + 1)]), col = "violet",
    lwd = 2, lty = 2)
> text(x1[n + 1], y1[n + 1], "leverage point and outlier   ",
    adj = c(1, 0.5))
> mtext("Data set 1")
> plot(x2[1:n], y2[1:n], pch = 3, xlim = xlim, ylim = ylim)
> points(x2[n + 1], y2[n + 1], pch = 21, col = "red", bg = "orange")
> abline(lm(y2 ~ x2), col = "dodgerblue", lwd = 2)
> abline(lm(y2[-(n + 1)] ~ x2[-(n + 1)]), col = "violet",
    lwd = 2, lty = 2)
> text(x2[n + 1], y2[n + 1], "leverage point   ", adj = c(1,
    0.5))
> mtext("Data set 2")
```




Obrázek 1: Grafické znázornění odlehlých, vybočujících a vlivných bodů na simulovaných datech.

Z druhého panelu grafu je patrné, že doplněný bod v druhé sérii simulovaných dat je vlivným bodem.

Grafické zobrazení regresních přímek doplníme také explicitním výpočtem pomocí funkce `lm()` pro oboje simulovaná data. Výsledky zobrazíme pomocí funkce `summary()`.

```
> m1 <- lm(y1 ~ x1)
> summary(m1)


Call:
lm(formula = y1 ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-8.20309 -1.67102  0.07375  3.14365  6.30204

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.8296     1.1903   4.897 3.67e-05 ***
x1            2.9517     0.1026  28.762  < 2e-16 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 4.039 on 28 degrees of freedom
Multiple R-squared: 0.9673, Adjusted R-squared: 0.9661
F-statistic: 827.2 on 1 and 28 DF,  p-value: < 2.2e-16
```

```
> m2 <- lm(y2 ~ x2)
> summary(m2)
```

```
Call:
lm(formula = y2 ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-40.3571  -6.5171   0.2596   6.7593  24.8853

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.0678     3.7758   5.845 2.78e-06 ***
x2            0.7510     0.3255   2.307   0.0287 *
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 12.81 on 28 degrees of freedom
Multiple R-squared: 0.1597, Adjusted R-squared: 0.1297
F-statistic: 5.322 on 1 and 28 DF,  p-value: 0.02867
```

Porovnáme–li oba dva regresní modely, vidíme, že i když se výchozí data liší pouze v jediném bodě, dostali jsme diametrálně odlišné výsledky.

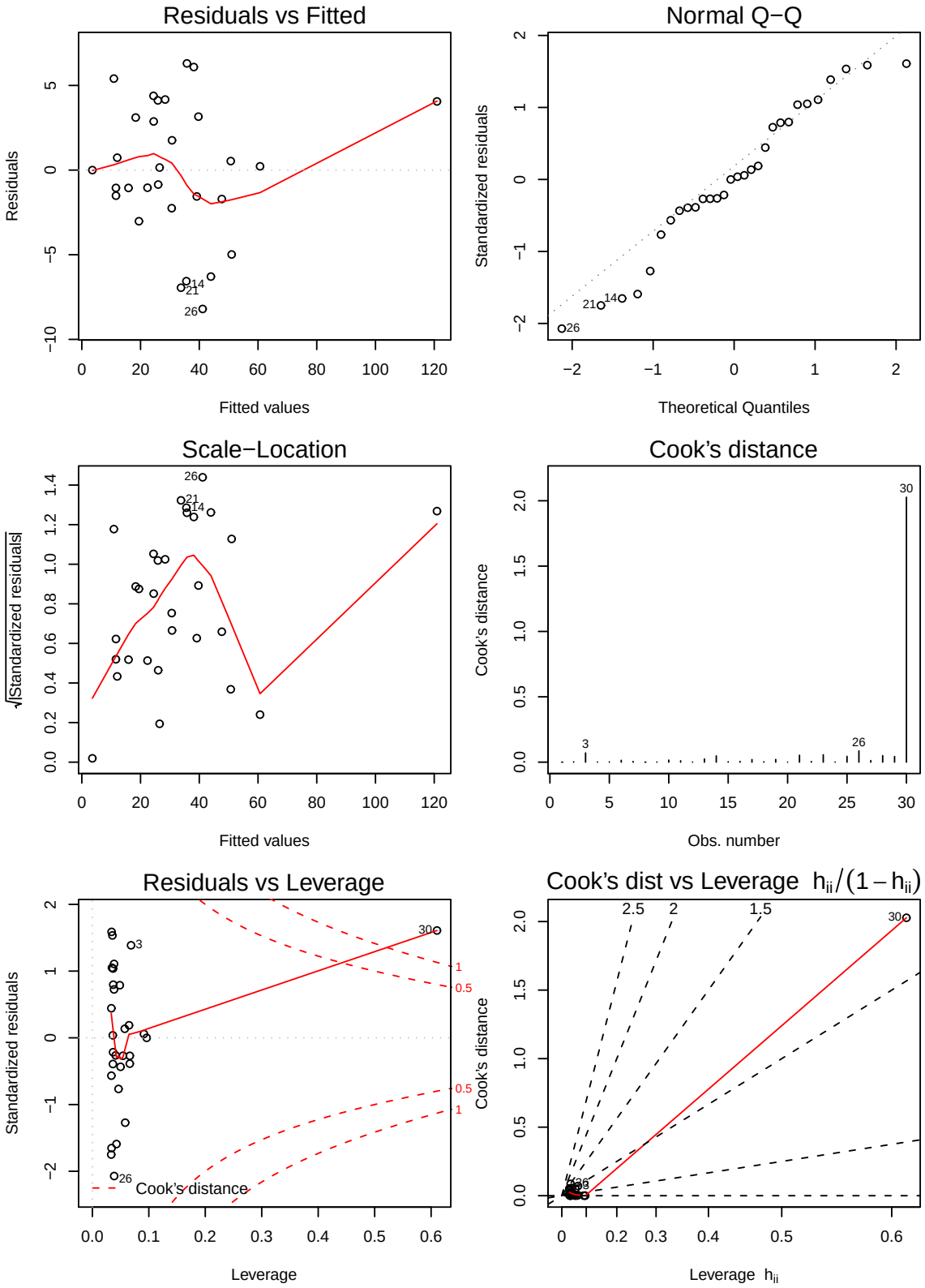
| | | | |
|---|---|---|---|
| 1. model | $y = 5.83 + 2.95\,x$ | upravený koeficient determinace: | 0.97 |
| 2. model | $y = 22.07 + 0.75\,x$ | | 0.16 |

První model vysvětluje téměř 97% variability dat, kdežto druhý pouhých 16%.

Tyto skutečnosti by se měly projevit také v analýze reziduí.

Podíváme se nejprve, jaké typy diagnostických grafů nabízí funkce `plot.lm()` (ale stačí psát `plot()`) pro simulovaná data 1 a 2.

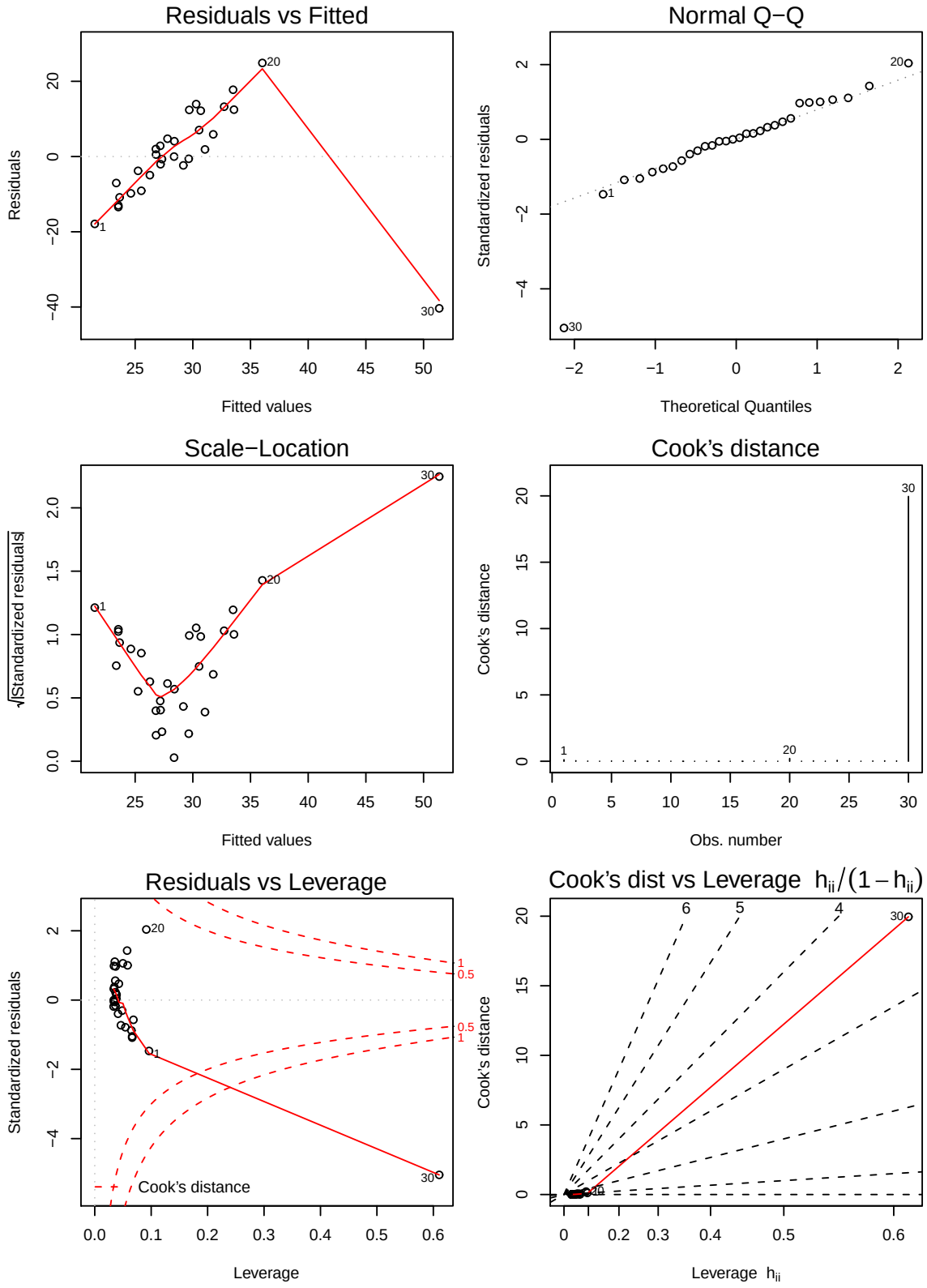
```
> par(mfrow = c(3, 2), mar = c(5, 5, 1.5, 0) + 0.05)
> plot(m1, which = 1:6)
```




Obrázek 2: `plot.lm(.,which=1:6)` pro simulovaná data 1.

```
> par(mfrow = c(3, 2), mar = c(5, 5, 1.5, 0) + 0.05)
```

```
> plot(m2, which = 1:6)
```




Obrázek 3: `plot.lm(.,which=1:6)` pro simulovaná data 2.

Prohlédneme–li si pozorně jednotlivé grafy, vidíme, že v těch grafech, ve kterých figuruje Cookova vzdálenost, se nejlépe odhalí oba dva doplněné koncové body.

Velmi užitečným je také bublinkový graf vytvořený pomocí funkce `influencePlot()`, který nabízí knihovna `car`.

Na ose $x$ jsou vyneseny hodnoty vlivů $h_{ii}$, na ose $y$ hodnoty studentizovaných reziduí $r_{J(i)}$. Každý bod $(h_{ii}, r_{J(i)})$ reprezentuje kruh, jehož obsah je úměrný velikosti Cookovy vzdálenosti $D_i$.

Kromě toho jsou do grafu zaneseny dvě vertikální referenční čáry pro vlivy (konkrétně hodnoty $2\frac{k}{n}$ a $3\frac{k}{n}$) a tři horizontální referenční čáry pro hodnoty $-2, 0, 2$.

Body ležící mezi oběma vertikálními přímkami jsou **vlivné**, ty body, které leží napravo od druhé vertikální přímky jsou pak již **silně vlivné**.

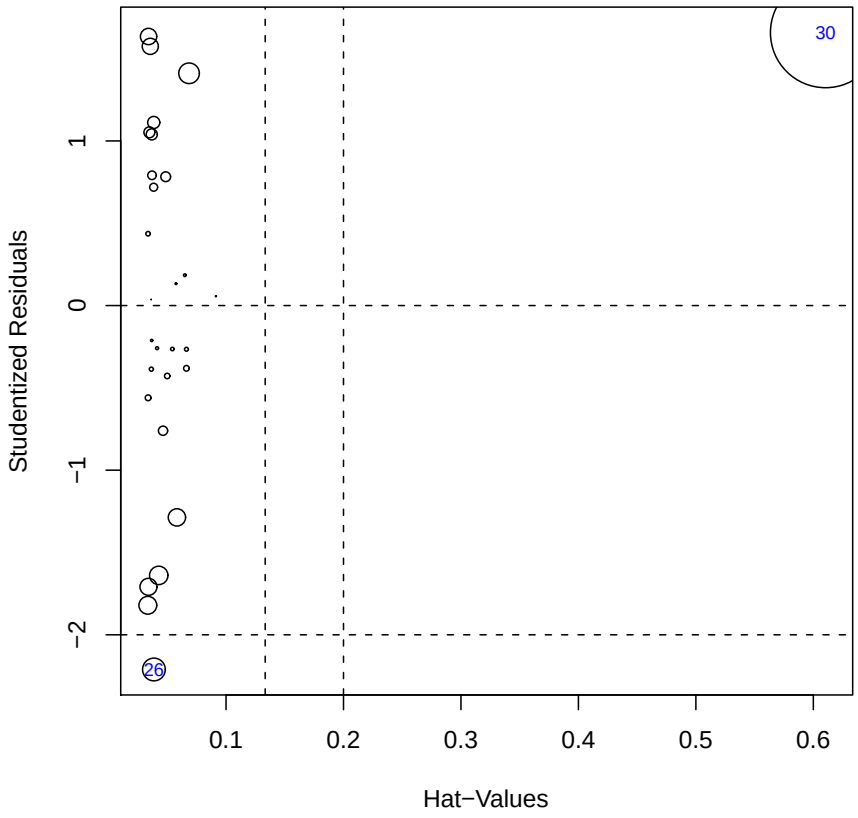
Body, které leží nad horizontální přímkou v bodě 2 nebo pod horizontální přímkou v bodě $-2$ jsou **odlehlé**.

Pokud přidáme do funkce `influencePlot()` argument `id.method="identify"`, můžeme interaktivně provádět identifikaci podezřelých bodů.

Identifikaci podezřelých hodnot neprovedeme interaktivně, ale ještě před vykreslením grafu je sami nalezneme. Po vykreslení grafu pomocí funkce `influencePlot()` pak s využitím příkazu `text()` doplníme číslo indexu, tj. pořadí pozorování.

Vytvořme bublinkové grafy pro oba dva simulované datové soubory, doplňme identifikaci podezřelých bodů a vypišme je.

```
> library(car)
> k <- length(coef(m1))
> n <- length(x1)
> sr <- rstudent(m1)
> Lr2 <- abs(sr) > 2
> hii <- hatvalues(m1)
> Lh2 <- hii > 2 * k/n
> L <- Lr2 | Lh2
> par(mfrow = c(1, 1), mar = c(5, 5, 1.5, 0) + 0.05)
> influencePlot(m1)
> text(hii[L], sr[L], as.character((1:n)[L]), adj = c(0.5,
    0.5), col = "blue", cex = 0.75)
```




Obrázek 4: Graf `influencePlot()` z knihovny `car` pro simulovaná data 1.

Vypíšeme podezřelá pozorování spolu s hodnotami studentizovaných reziduí $r_{J(i)}$, vlivy $h_{ii}$ a Cookovou vzdáleností $D_i$.

```
> Di <- cooks.distance(m1)
> cat(paste("meze pro hii:  2*k/n=", round(2 * k/n, 4),
      "   3*k/n=", round(3 * k/n, 4), "\n", sep = ""))
```

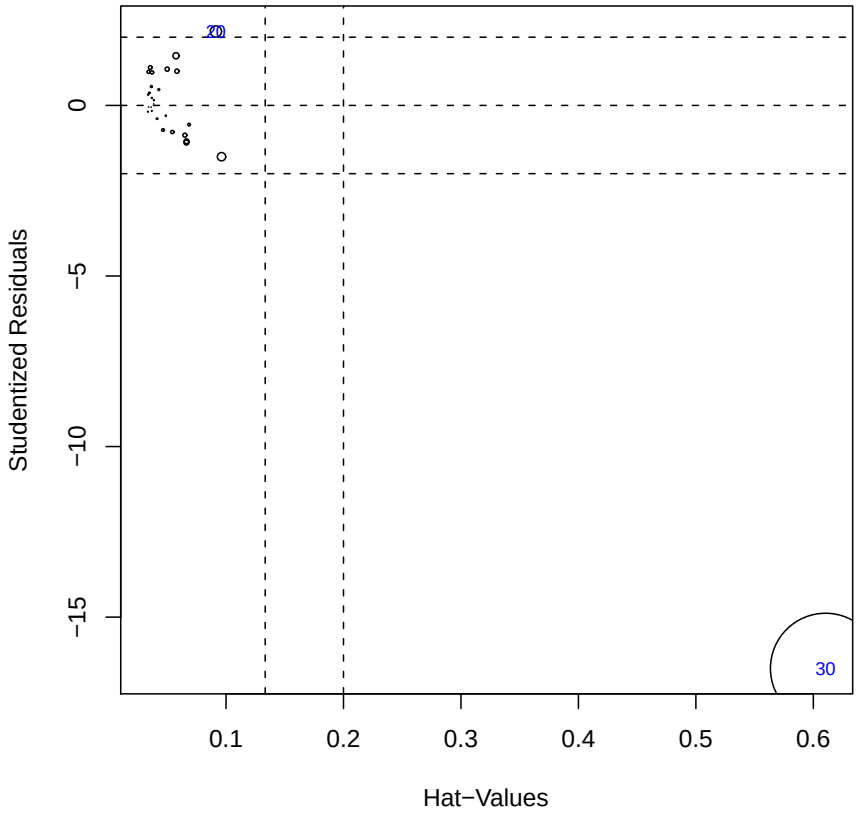
```
meze pro hii:  2*k/n=0.1333   3*k/n=0.2
```

```
> cbind((1:n)[L], x1[L], y1[L], sr[L], hii[L], Di[L])
```

```
   [,1]     [,2]      [,3]      [,4]       [,5]       [,6]
26   26 11.96542  32.94453 -2.210719 0.03861484 0.08618558
30   30 39.00000 125.00000  1.658210 0.61041902 2.02747514
```

```
> library(car)
> k <- length(coef(m2))
> n <- length(x2)
> sr <- rstudent(m2)
> Lr2 <- abs(sr) > 2
> hii <- hatvalues(m2)
> Lh2 <- hii > 2 * k/n
```

```
> L <- Lr2 | Lh2
> par(mfrow = c(1, 1), mar = c(5, 5, 1.5, 0) + 0.05)
> influencePlot(m2)
> text(hii[L], sr[L], as.character((1:n)[L]), adj = c(0.5,
    0.5), col = "blue", cex = 0.75)
```




Obrázek 5: Graf `influencePlot()` z knihovny `car` pro simulovaná data 2.

Vypíšeme podezřelá pozorování spolu s hodnotami studentizovaných reziduí $r_{J(i)}$, vlivy $h_{ii}$ a Cookovou vzdáleností $D_i$.

```
> Di <- cooks.distance(m2)
> cat(paste("meze pro hii:  2*k/n=", round(2 * k/n, 4),
    "   3*k/n=", round(3 * k/n, 4), "\n", sep = ""))
```

```
meze pro hii:  2*k/n=0.1333   3*k/n=0.2
```

```
> cbind((1:n)[L], x2[L], y2[L], sr[L], hii[L], Di[L])
```

```
   [,1]     [,2]     [,3]       [,4]       [,5]       [,6]
20   20 18.58604 60.91134   2.168430 0.09137258  0.2088146
30   30 39.00000 11.00000 -16.502845 0.61041902 19.9574746
```

Oba dva bublinkové grafy dobře odhalily jako podezřelé uměle doplněné body s indexem 30.

Provedeme–li srovnání obou uměle doplněných bodů, dostaneme

| kritérium | bod 30 – 1. sada dat | bod 30 – 2. sada dat |
|---|---|---|
| $r_{J(i)}$ | 1.6582 | $-16.5028 \quad < -2$ |
| $h_{ii}$ | $0.6104 \quad > 3\frac{k}{n}$ | $0.6104 \quad > 3\frac{k}{n}$ |
| $D_i$ | $2.0275 \quad > 1$ | $19.9575 \quad > 1$ |

Z hodnot vyplývá, že oba dva body zřetelně vybočují v $x$-vých hodnotách (identifikuje $h_{ii}$). Bod 30 v první sadě dat je v souladu s předpokládaným regresním modelem (regresní přímkou), proto hodnota studentizovaného rezidua $r_{J(i)}$ nevybočuje. Signifikantní je samozřejmě Cookova vzdálenost $D_i$. Naproti tomu bod 30 ve druhé sadě dat jasně nevyhovuje předpokládanému regresnímu modelu (vysoká hodnota $r_{J(i)}$ a závratná hodnota $D_i$).

V případě bodů 26 v první sadě dat a bodu 20 ve druhé sadě dat máme hodnoty

| kritérium | bod 26 – 1. sada dat | bod 20 – 2. sada dat |
|---|---|---|
| $r_{J(i)}$ | $-2.2107 \quad < -2$ | $2.1684 \quad > 2$ |
| $h_{ii}$ | 0.0386 | 0.0913 |
| $D_i$ | 0.0862 | 0.2088 |

Je vidět, že tyto dva body byly vybrány pouze díky těsně vyšším hodnotám studentizovaných reziduí.

Jako zástupce **indexových grafů** si uvedeme graf `infIndexPlot()`, který je dostupný opět v knihovně `car`.

Tato funkce nabízí pro jednotlivá pozorování ($i = 1, \ldots, n$) samostatně hodnoty

- Cookovy vzdálenosti $D_i$,
- studentizovaná rezidua $r_{J(i)}$
- vlivů $h_{ii}$,
- simultánní hladiny významnosti (na základě Bonferroniho nerovnosti) pro testování hypotézy $H_0: \ E\varepsilon_i = 0$ proti alternativě $H_1: \ E\varepsilon_i \neq 0$.

Poznámka (**Bonferroniho nerovnost**).

Mějme $n$ intervalů spolehlivosti $CI_1, \ldots, CI_n$, pro které platí $P(\gamma(\theta_i) \in CI_i) = 1 - \alpha_i$ pro $\alpha_i \in (0, 1)$. Pak platí

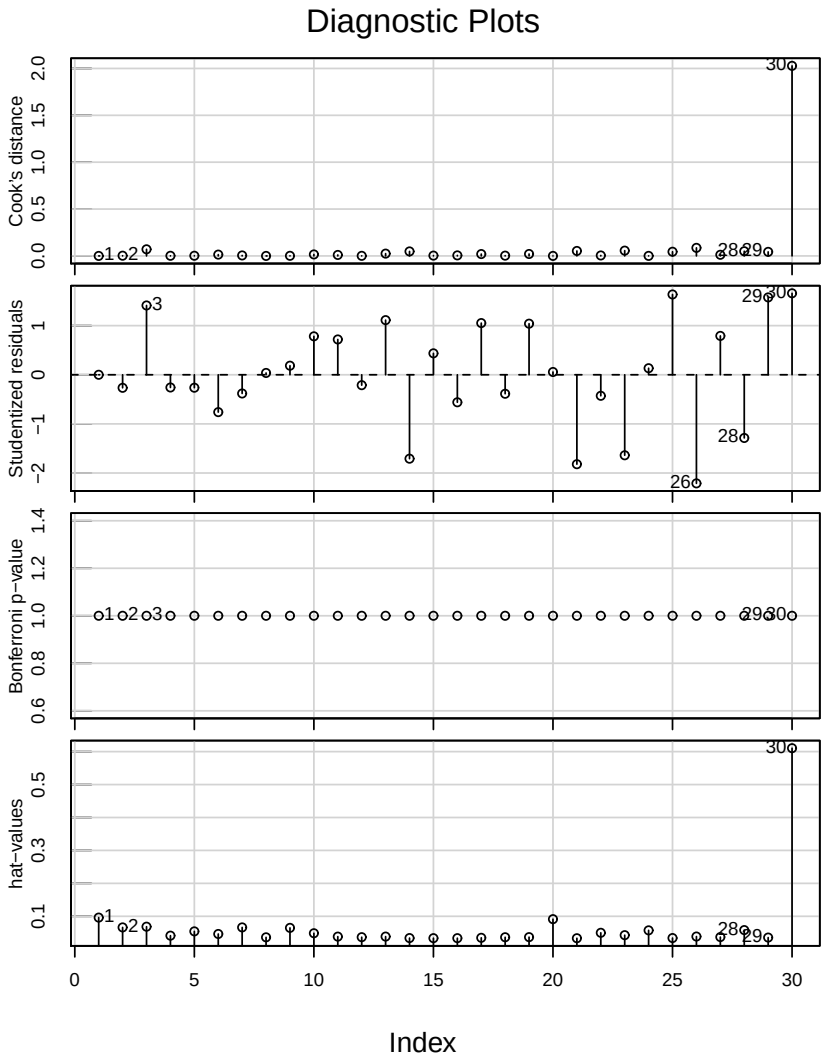
$$P\left(\bigcap_{i=1}^{n} \{\gamma(\theta_i) \in CI_i\}\right) \geq 1 - (\alpha_1 + \cdots + \alpha_n)$$

Pokud zvolíme $\alpha_i = \frac{\alpha}{n}$, pak bude zaručeno, že

$$P\left(\bigcap_{i=1}^{n} \{\gamma(\theta_i) \in CI_i\}\right) \geq 1 - \alpha.$$

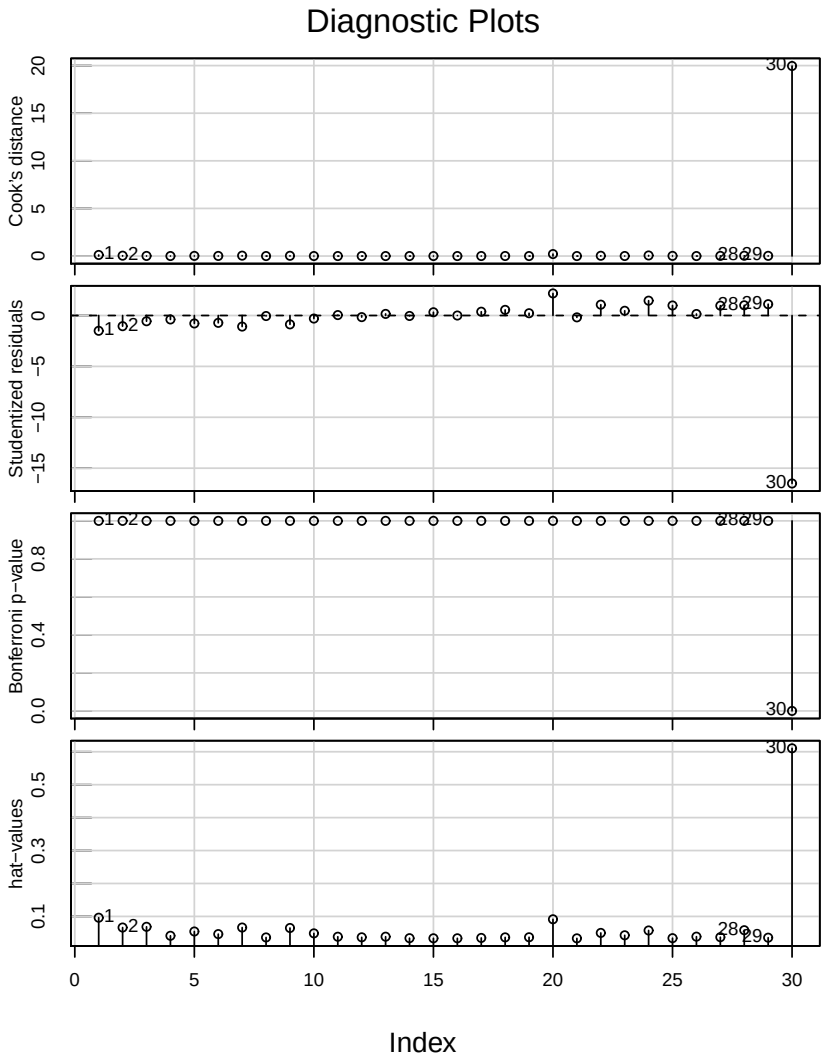
Pro oba dva regresní modely použijeme funkci `infIndexPlot()` z knihovny `car`. Přidáme požadavek na označení 5 bodů $(x_i, y_i)$, které mají největší Mahalanobisovu vzdálenost od výběrových průměrů $(\bar{x}, \bar{y})$.

```
> library(car)
> infIndexPlot(m1, id.method = "mahal", id.n = 5)
```




Obrázek 6: Graf `infIndexPlot()` z knihovny `car` pro simulovaná data 1.

```
> library(car)
> infIndexPlot(m2, id.method = "mahal", id.n = 5)
```




Obrázek 7: Graf `infIndexPlot()` z knihovny `car` pro simulovaná data 2.

Nejdůležitější informací předchozích dvou grafů jsou simultánní hladiny významnosti (na základě Bonferroniho nerovnosti) pro testování hypotézy $H_0 : E\varepsilon_i = 0$ proti alternativě $H_1 : E\varepsilon_i \neq 0$. Z grafu však nezjistíme explicitní p–hodnoty. Tuto informaci získáme díky funcki `outlierTest()` z knihovny `car`.

```
> outlierTest(m1)
```

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
26 -2.210719           0.035713           NA
```

```
> outlierTest(m2)
```

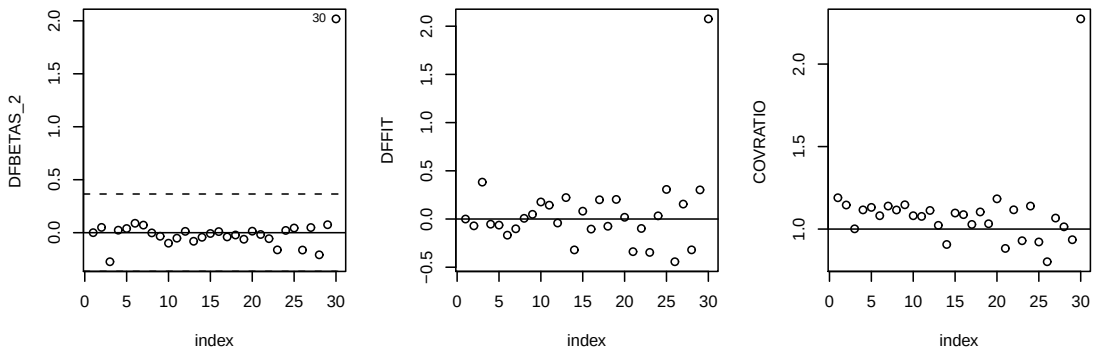
```
    rstudent unadjusted p-value Bonferonni p
30 -16.50284         1.2485e-15   3.7456e-14
```

V prvním modelu byl nalezen jako významný bod 26, ve druhém modelu bod 30.

Dosud jsme nezkoumali $DFFIT$ rezidua, $DFBETAS_{ij}$ a $COVRATIO_i$ ($i = 1, \ldots, n$, $j = 1, \ldots, k$), která odhalují simultánní vliv pozorování na odhady neznámých parametrů $\boldsymbol{\beta}$ a $\sigma^2$. Vytvoříme indexové grafy a provedeme identifikaci podezřelých pozorování.

```
> X <- x1
> M <- m1
> Y <- y1
> IM <- influence.measures(M)
> IDs <- 2:4
> txtY <- c("DFBETAS_2", "DFFIT", "COVRATIO")
> shift <- c(0, 0, 1)
> CutOff <- c(2/sqrt(length(X)), 2 * sqrt(length(coef(M)/length(X))),
    3 * length(coef(M)/length(X)))
> par(mfrow = c(1, 3), mar = c(5, 5, 0.5, 0.5) + 0.05)
> for (id in 1:3) {
    ID <- IDs[id]
    plot(IM$infmat[, ID], type = "p", xlab = "index",
        ylab = txtY[id])
    L <- abs(IM$infmat[, ID] - shift[id]) > CutOff[id]
    abline(h = shift[id])
    abline(h = c(shift[id] - CutOff[id], shift[id] +
        CutOff[id]), lty = 2)
    if (sum(L) > 0) {
        text(as.numeric(rownames(IM$infmat)[L]), IM$infmat[L,
            ID], labels = rownames(IM$infmat)[L], cex = 0.75,
            pos = 2)
        pom <- data.frame(cbind(rownames(IM$infmat)[L],
            X[L], Y[L], IM$infmat[L, ID]))
        names(pom) <- c("index", "x", "y", colnames(IM$infmat)[ID])
        o <- order(-abs(IM$infmat[L, ID]))
        print(pom[o, ], digits = 4)
    }
 }
```

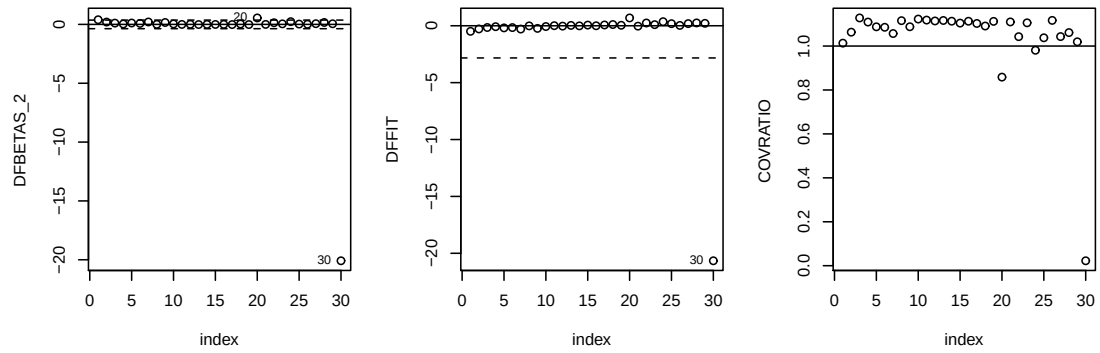
```
  index  x   y           dfb.x1
1    30 39 125 2.01818459863541
```




Obrázek 8: Indexové grafy pro vybrané influenční míry pro simulovaná data 1.

```
> X <- x2
> M <- m2
> Y <- y2
> IM <- influence.measures(M)
> IDs <- 2:4
> txtY <- c("DFBETAS_2", "DFFIT", "COVRATIO")
> shift <- c(0, 0, 1)
> CutOff <- c(2/sqrt(length(X)), 2 * sqrt(length(coef(M)/length(X))),
    3 * length(coef(M)/length(X)))
> par(mfrow = c(1, 3), mar = c(5, 5, 0.5, 0.5) + 0.05)
> for (id in 1:3) {
    ID <- IDs[id]
    plot(IM$infmat[, ID], type = "p", xlab = "index",
        ylab = txtY[id])
    L <- abs(IM$infmat[, ID] - shift[id]) > CutOff[id]
    abline(h = shift[id])
    abline(h = c(shift[id] - CutOff[id], shift[id] +
        CutOff[id]), lty = 2)
    if (sum(L) > 0) {
        text(as.numeric(rownames(IM$infmat)[L]), IM$infmat[L,
            ID], labels = rownames(IM$infmat)[L], cex = 0.75,
            pos = 2)
        pom <- data.frame(cbind(rownames(IM$infmat)[L],
            X[L], Y[L], IM$infmat[L, ID]))
        names(pom) <- c("index", "x", "y", colnames(IM$infmat)[ID])
        o <- order(-abs(IM$infmat[L, ID]))
        print(pom[o, ], digits = 4)
    }
}
```

```
   index                 x                y            dfb.x2
30    30                39               11 -20.0853820508696
20    20   18.5860388797267 60.9113433759351 0.548041322583496
1      1 -0.757673226860099 3.59174629397058 0.396240079235418
  index  x  y             dffit
1    30 39 11 -20.6573198640163
```




Obrázek 9: Indexové grafy pro vybrané influenční míry pro simulovaná data 2.

V prvním modelu má vliv na odhady neznámých parametrů $\beta_1$ a $\sigma^2$ bod 30 a v druhém modelu jsou to body 1, 20 a 30.

PŘÍKLAD 2: Obsah těžkých kovů v řece Moravě v průběhu let 1997 až 2009

Načteme datový soubor `sedimenty.csv` pomocí příkazu `read.csv2()` a vytvoříme proměnnou `data` typu `data.frame`.

```
> fileDat <- paste(data.library, "sedimenty.csv", sep = "")
> data <- read.csv2(fileDat)
> str(data)
```

```
,data.frame,:   144 obs. of  6 variables:
 $ lokalita: int  1 1 1 1 1 1 1 1 1 ...
 $ rok     : int  1997 1997 1998 1998 1999 1999 2000 2000 2001 2001 ...
 $ Pb      : num  24.9 34.1 38 40.3 33.2 33.7 45.9 82.5 68.9 64.1 ...
 $ Cd      : num  0.69 1.22 0.842 0.894 0.548 0.609 1.13 0.78 0.64 0.86 ...
 $ Ni      : num  29.8 36.6 40.3 35.4 21.6 24.1 NA NA 60.2 54.6 ...
 $ Hg      : num  0.056 0.144 0.187 0.086 0.078 0.071 0.22 NA 0.15 0.2 ...
```

Všimněme si, že pro některé těžké kovy jsou v souboru chybějící pozorování. Budeme muset pak být velmi obezřetní při identifikaci podezřelých pozorování podle jejich pořadových čísel.

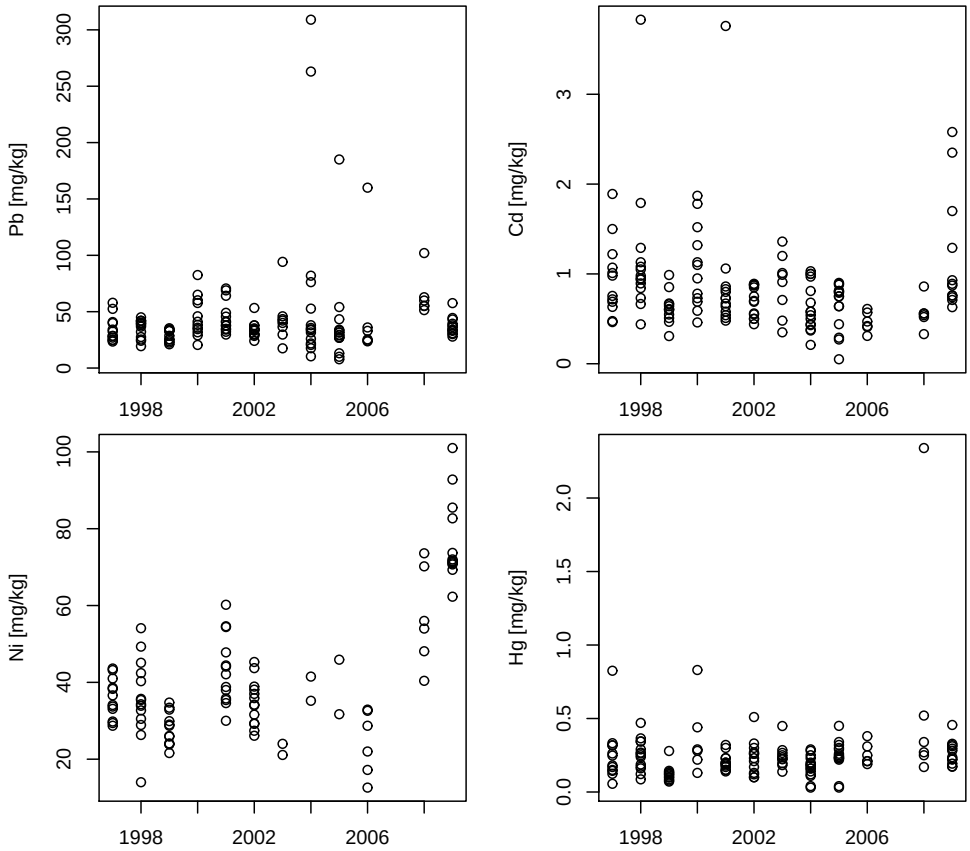
```
> Lna <- apply(is.na(data), 1, any)
> print(data[Lna, ])
```

```
    lokalita  rok    Pb   Cd Ni    Hg
7          1 2000  45.9 1.13 NA 0.220
8          1 2000  82.5 0.78 NA    NA
13         1 2003    NA   NA NA 0.183
14         1 2003  45.9 1.36 NA 0.225
15         1 2004 263.0 0.43 NA 0.040
16         1 2004  10.4 0.21 NA 0.030
17         1 2005  13.1 0.27 NA 0.320
18         1 2005  10.0 0.29 NA 0.030
29         2 2000  60.1 1.52 NA 0.830
30         2 2000  34.5 0.69 NA    NA
35         2 2003    NA   NA NA 0.285
36         2 2004  52.7 0.58 NA 0.160
37         2 2004  21.8 0.37 NA 0.120
38         2 2005  26.9 0.44 NA 0.230
39         2 2005  27.5 0.90 NA 0.220
48         3 2000  41.1 1.10 NA 0.290
49         3 2000  57.7 1.32 NA    NA
54         3 2003  17.5 0.35 NA 0.138
55         3 2003  43.4 0.91 NA 0.449
56         3 2004 309.0 0.81 NA 0.280
57         3 2004  35.6 0.54 NA 0.190
58         3 2005  33.2 0.80 NA 0.340
59         3 2005  29.5 0.81 NA 0.450
70         4 2000  38.3 0.95 NA 0.280
71         4 2000  31.7 0.73 NA    NA
76         4 2003  29.7 0.48 NA 0.196
77         4 2004  76.2 0.50 NA 0.160
78         4 2004  34.3 0.58 NA 0.200
79         4 2004  20.5 0.44 NA 0.110
80         4 2005  43.3 0.88 NA 0.300
```

```
81          4 2005  29.0 0.64 NA 0.290
92          5 2000  65.0 1.78 NA 0.440
93          5 2000  35.8 1.87 NA    NA
100         5 2004  81.9 0.97 NA 0.180
101         5 2004  25.9 1.00 NA 0.230
102         5 2005 185.0 0.81 NA 0.230
103         5 2005  31.5 0.75 NA 0.240
114         6 2000  20.5 0.46 NA 0.130
115         6 2000  29.0 0.59 NA    NA
120         6 2003  36.5 0.99 NA 0.266
121         6 2003  94.2 0.71 NA 0.237
122         6 2004  31.6 0.68 NA 0.180
123         6 2004  17.6 0.38 NA 0.140
124         6 2005  54.1 0.89 NA 0.240
125         6 2005  26.8 0.75 NA 0.230
126         6 2005   7.8 0.05 NA 0.040
132         7 2001    NA   NA NA    NA
135         7 2002    NA   NA NA    NA
140         7 2006    NA   NA NA    NA
141         7 2006    NA   NA NA    NA
```

Data vykreslíme

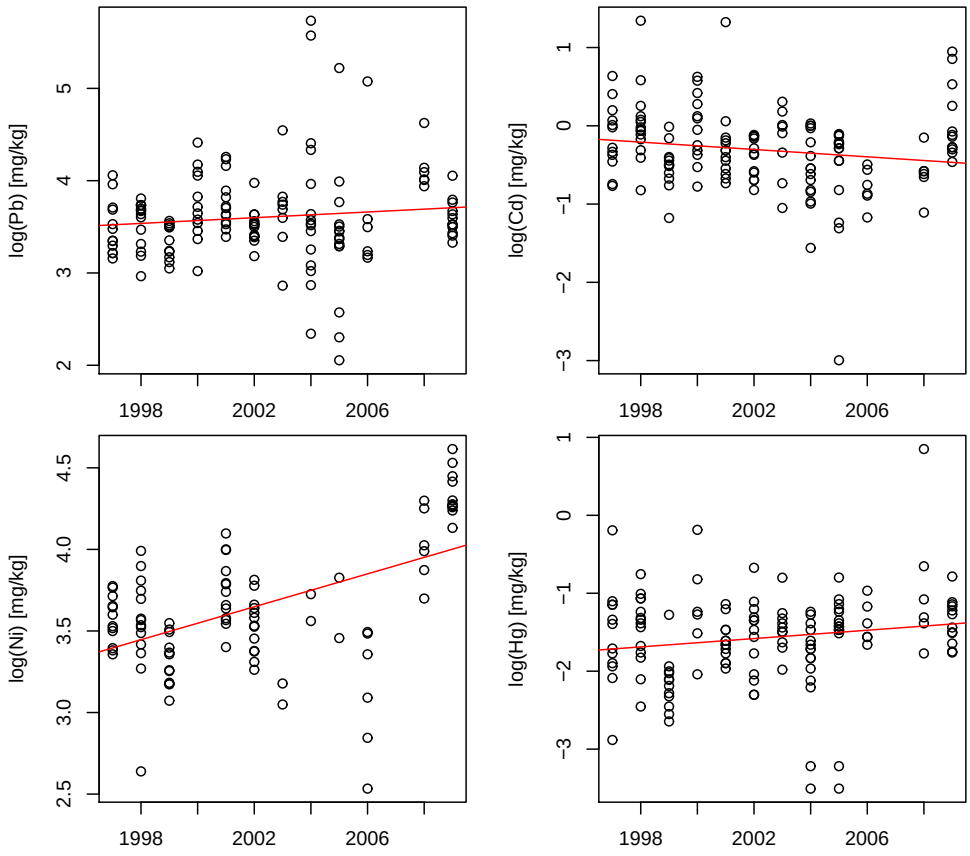
```
> par(mfrow = c(2, 2), mar = c(2, 5, 0.5, 0.5) + 0.05)
> for (id in 3:6) plot(data[, id] ~ data[, 2], ylab = paste(names(data)[id],
    "[mg/kg]"))
```




Obrázek 10: *Obsah těžkých kovů v řece Moravě v průběhu let 1997 až 2009 (v mg/kg).*

Vzhledem k tomu, že jako závisle proměnné budeme brát **logaritmy** obsahů těžkých kovů v mg/kg, vykresleme ještě zlogaritmovaná data. Do grafů pro názornost dokreslíme regresní přímky.

```
> LogNames <- c(names(data)[1:2], paste("log(", names(data)[3:6],
    ") [mg/kg]", sep = ""))
> par(mfrow = c(2, 2), mar = c(2, 5, 0.5, 0.5) + 0.05)
> for (id in 3:6) {
    plot(log(data[, id]) ~ data[, 2], ylab = LogNames[id])
    abline(lm(log(data[, id]) ~ data[, 2]), col = "red")
 }
```




Obrázek 11: *Logaritmy obsahu těžkých kovů v řece Moravě v průběhu let 1997 až 2009 (v mg/kg).*

Pro další analýzu vybereme jeden z těžkých kovů, například kadmium a provedeme regresní analýzu pro logaritmy obsahu těžkých kovů. Protože nezávisle proměnnou jsou roky, které nabývají vysokých hodnot, doporučuje se vždy čas centrovat.

```
> id <- 4
> y <- log(data[, id])
> xshift <- mean(range(data[, 2]))
> x <- data[, 2] - xshift
> ytxt <- LogNames[id]
> Data <- data.frame(x = x, y = y)
> m <- lm(y ~ x)
> summary(m)
```

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

Residuals:
      Min        1Q    Median        3Q       Max
-2.622863 -0.272384  0.005871  0.255382  1.603285

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.32587    0.04521  -7.208 3.55e-11 ***
x           -0.02350    0.01232  -1.907   0.0586 .
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 0.5206 on 136 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared: 0.02606,    Adjusted R-squared: 0.01889
F-statistic: 3.638 on 1 and 136 DF,  p-value: 0.05857
```

Dalším krokem bude grafické zázornění výsledků spolu s intervaly spolehlivosti kolem regresní přímky a také s predikčními intervaly spolehlivosti.

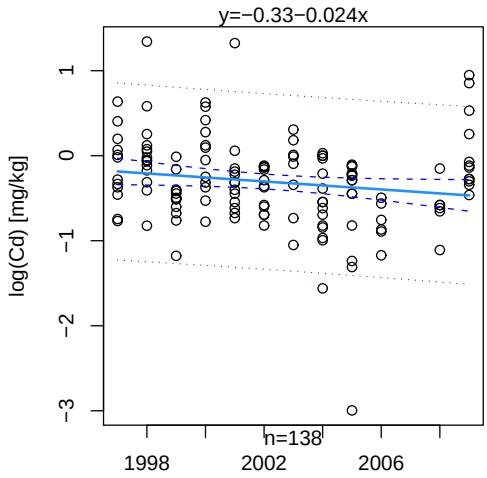
```
> xrange <- range(x)
> xx <- seq(xrange[1], xrange[2], length.out = 100)
> conf.i <- predict(m, newdata = data.frame(x = xx), interval = "confidence")
> str(conf.i)
```

```
 num [1:100, 1:3] -0.185 -0.188 -0.191 -0.193 -0.196 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:100] "1" "2" "3" "4" ...
  ..$ : chr [1:3] "fit" "lwr" "upr"
```

```
> pred.i <- predict(m, newdata = data.frame(x = xx), interval = "prediction")
> str(pred.i)
```

```
 num [1:100, 1:3] -0.185 -0.188 -0.191 -0.193 -0.196 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:100] "1" "2" "3" "4" ...
  ..$ : chr [1:3] "fit" "lwr" "upr"
```

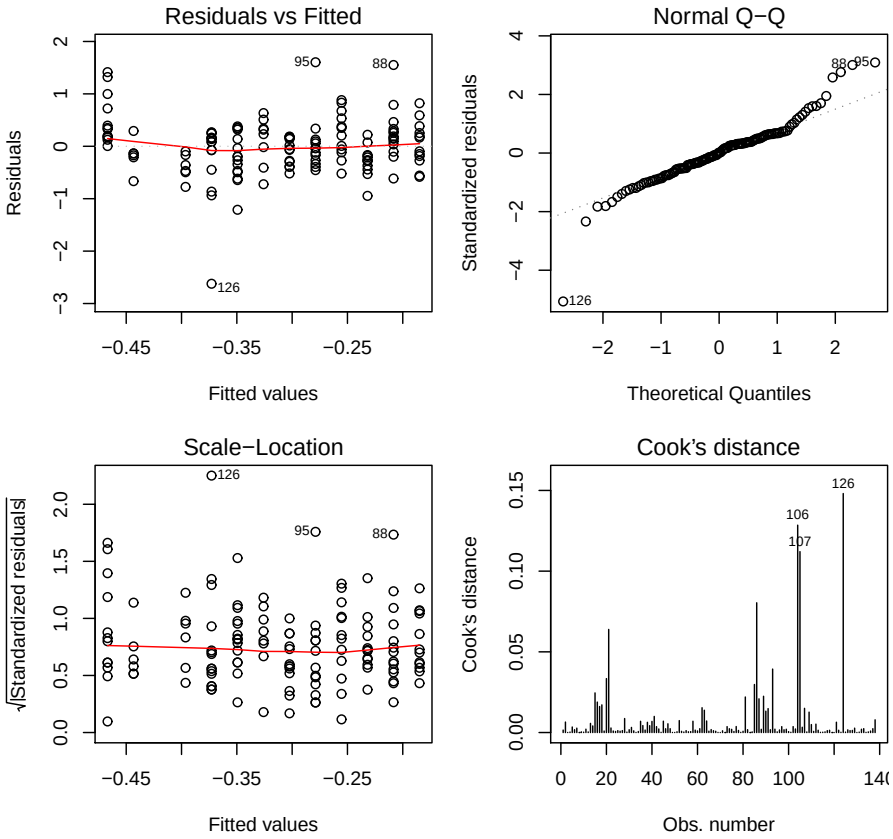
```
> ylim <- range(c(y, pred.i[, 2:3]), na.rm = TRUE)
> par(mfrow = c(1, 1), mar = c(2, 5, 1, 0) + 0.5)
> xt <- x + xshift
> xxt <- xx + xshift
> plot(y ~ xt, ylab = ytxt, ylim = ylim)
> matlines(xxt, cbind(conf.i, pred.i[, -1]), lty = c(1,
    2, 2, 3, 3), lwd = c(2, 1, 1, 1, 1), col = c("dodgerblue",
    "blue", "blue", "darkgreen", "darkgreen"))
> znam <- ifelse(coef(m)[2] < 0, "", "+")
> txtmodel <- paste("y=", round(coef(m)[1], 2), znam, round(coef(m)[2],
    3), "x", sep = "")
> mtext(txtmodel)
> mtext(paste("n=", length(m$residuals), sep = ""), side = 1)
```




Obrázek 12: *Logaritmy obsahu těžkých kovů v řece Moravě v průběhu let 1997 až 2009 (Cd v mg/kg).*

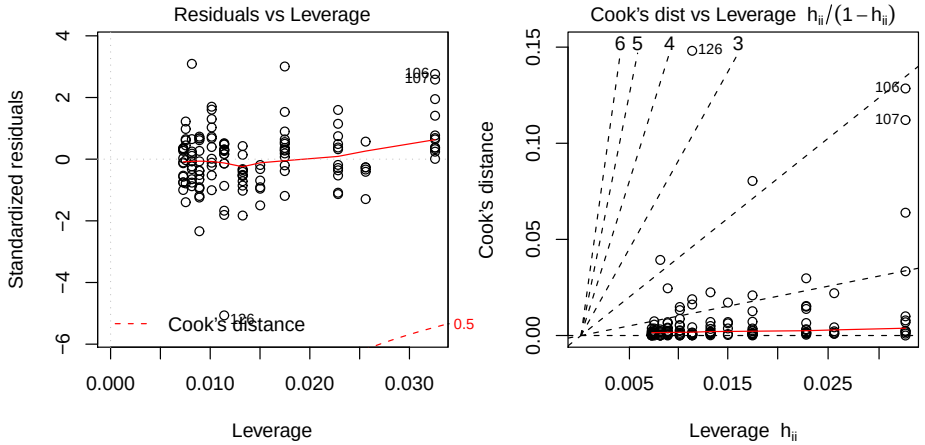
Pro regresní model provedeme analýzu reziduí – využijeme funkci `plot()`.

```
> par(mfrow = c(2, 2), mar = c(5, 5, 1.5, 0) + 0.05)
> plot(m, which = 1:4)
```




Obrázek 13: `plot.lm(.,which=1:4)` pro *Logaritmy obsahu těžkých kovů v řece Moravě v průběhu let 1997 až 2009 (Cd v mg/kg).*

```
> par(mfrow = c(1, 2), mar = c(5, 5, 1.5, 0) + 0.05)
> plot(m, which = 5:6)
```




Obrázek 14: `plot.lm(.,which=5:6)` pro *Logaritmy obsahu těžkých kovů v řece Moravě v průběhu let 1997 až 2009 (Cd v mg/kg).*

V dalším kroku využijeme funkci `influencePlot()` pro vykreslení bublinkového grafu. Na ose

- $x$ jsou vyneseny hodnoty vlivů $h_{ii}$,
- na ose $y$ hodnoty studentizovaných reziduí $r_{J(i)}$.

Každý bod $(h_{ii}, r_{J(i)})$ reprezentuje kruh, jehož obsah je úměrný velikosti Cookovy vzdálenosti $D_i$.

Kromě toho jsou do grafu zaneseny dvě vertikální referenční čáry pro vlivy (konkrétně hodnoty $2\frac{k}{n}$ a $3\frac{k}{n}$) a tři horizontální referenční čáry pro hodnoty $-2, 0, 2$.

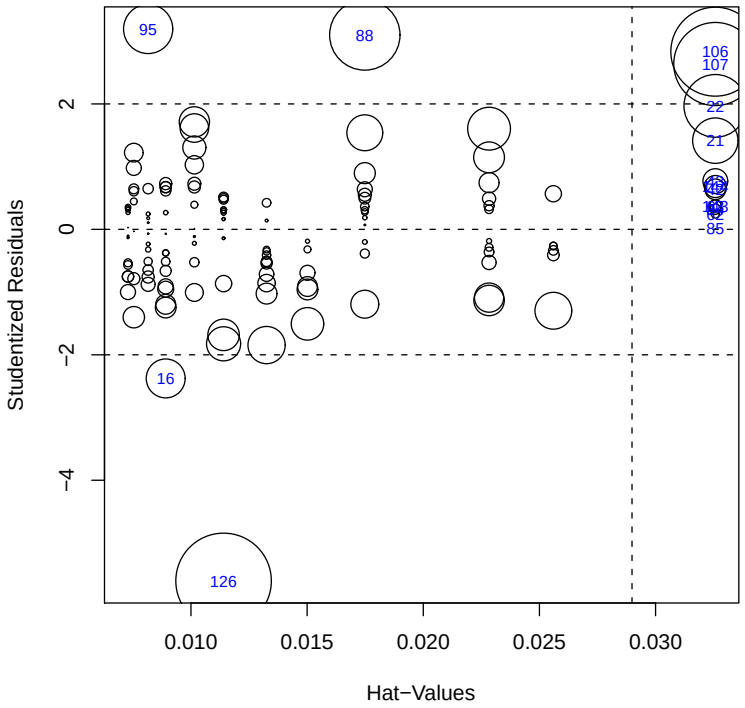
```
> k <- length(coef(m))
> sr <- rstudent(m)
> n <- length(sr)
> str(sr)
```

```
 Named num [1:138] -0.3607 0.7444 0.0703 0.186 -0.7134 ...
 - attr(*, "names")= chr [1:138] "1" "2" "3" "4" ...
```

```
> Lr2 <- abs(sr) > 2
> hii <- hatvalues(m)
> str(hii)
```

```
 Named num [1:138] 0.0228 0.0228 0.0175 0.0175 0.0133 ...
 - attr(*, "names")= chr [1:138] "1" "2" "3" "4" ...
```

```
> Lh2 <- hii > 2 * k/n
> L <- Lr2 | Lh2
> par(mfrow = c(1, 1), mar = c(5, 5, 1.5, 0) + 0.05)
> influencePlot(m)
> text(hii[L], sr[L], names(hii)[L], adj = c(0.5, 0.5),
    col = "blue", cex = 0.75)
```




Obrázek 15: Graf `influencePlot()` z knihovny `car` pro *Logaritmy obsahu těžkých kovů v řece Moravě v průběhu let 1997 až 2009 (Cd v mg/kg).*

Dříve než začneme vypisovat podezřelá pozorování pomocí jejich pořadí, je třeba si všimnout, že pro kadmium místo vstupních 144 pozorování je pouze 138, tedy 6 je chybějících. Proto budeme muset být při výpisu velmi opatrní.

Vypíšeme tedy podezřelá pozorování spolu s hodnotami studentizovaných reziduí $r_{J(i)}$, vlivy $h_{ii}$ a Cookovou vzdáleností $D_i$.

```
> LNA <- with(Data, is.na(y))
> Di <- cooks.distance(m)
> pom <- data.frame(data[!LNA, c(1:2, id)][L, ], y[!LNA][L],
    x[!LNA][L], sr[L], hii[L], Di[L])
> names(pom) <- c(names(data)[c(1:2, id)], "y", "x", "student.res",
    "h_ii", "D_i")
> print(pom, digits = 4)
```

```
    lokalita  rok   Cd        y  x student.res     h_ii       D_i
16         1 2004 0.21 -1.56065  1   -2.376692 0.008912 2.456e-02
21         1 2009 1.29  0.25464  6    1.414210 0.032573 3.342e-02
22         1 2009 1.70  0.53063  6    1.968491 0.032573 6.388e-02
42         2 2009 0.87 -0.13926  6    0.638394 0.032573 6.891e-03
43         2 2009 0.93 -0.07257  6    0.768871 0.032573 9.982e-03
62         3 2009 0.71 -0.34249  6    0.242065 0.032573 9.933e-04
63         3 2009 0.76 -0.27444  6    0.374618 0.032573 2.378e-03
84         4 2009 0.74 -0.30111  6    0.322660 0.032573 1.764e-03
85         4 2009 0.63 -0.46204  6    0.009414 0.032573 1.503e-06
88         5 1998 3.83  1.34286 -5    3.099694 0.017483 8.040e-02
95         5 2001 3.76  1.32442 -2    3.195242 0.008157 3.932e-02
```

```
106        5 2009 2.58  0.94779  6    2.833149 0.032573 1.285e-01
107        5 2009 2.35  0.85442  6    2.636163 0.032573 1.121e-01
126        6 2005 0.05 -2.99573  2   -5.605079 0.011404 1.481e-01
143        7 2009 0.76 -0.27444  6    0.374618 0.032573 2.378e-03
144        7 2009 0.89 -0.11653  6    0.682831 0.032573 7.880e-03
```

```
> cat(paste("meze pro hii:  2*k/n=", round(2 * k/n, 4),
    "   3*k/n=", round(3 * k/n, 4), "\n", sep = ""))
```

```
meze pro hii:  2*k/n=0.029   3*k/n=0.0435
```

Vzhledem k tomu, že nezávisle proměnnou je čas (tj. jednotlivé roky), jejich odlehlost nemá cenu zkoumat.

Body, které je třeba prověřit se tak stávají body, které jsou vybrány na základě hodnot studentizovaných reziduí $r_{J(i)}$. Jde o body, jejich indexy (pořadí) je 126, 106, 107, 88, 95 a 16.

Uvidíme, jak dopadne testování hypotézy $H_0: \ E\varepsilon_i = 0$ proti alternativě $H_1: \ E\varepsilon_i \neq 0$ pomocí simultánních hladin významnosti získaných na základě Bonferroniho nerovnosti.

```
> outlierTest(m)
```

```
     rstudent unadjusted p-value Bonferonni p
126 -5.605079         1.1272e-07   1.5555e-05
```

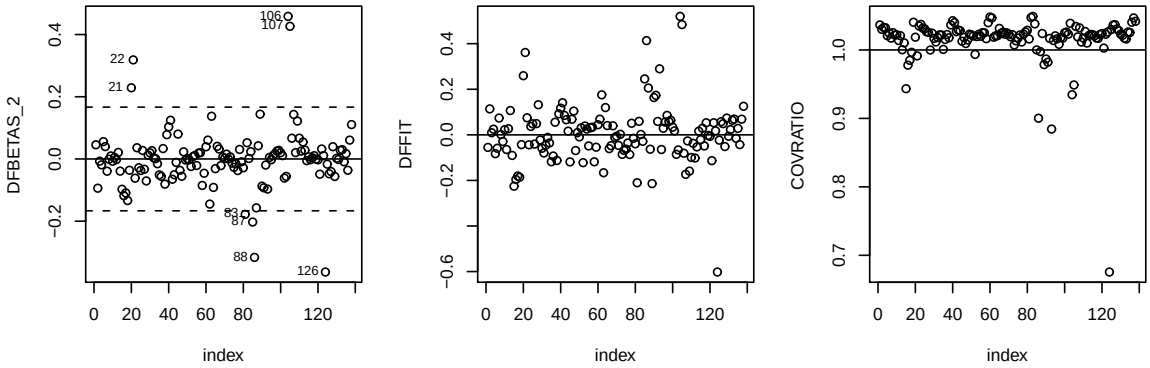
Jako odlehlé pozorování bylo určeno pozorování s indexem 126.

Na závěr ještě prozkoumejme $DFFIT$ rezidua, $DFBETAS_{ij}$ a $COVRATIO_i$ ($i = 1, \ldots, n$, $j = 1, \ldots, k$), která odhalují simultánní vliv pozorování na odhady neznámých parametrů $\boldsymbol{\beta}$ a $\sigma^2$. Vytvoříme proto indexové grafy a provedeme identifikaci podezřelých pozorování. Jejich pořadí setřídíme podle velikosti vlivu.

```
> X <- x
> M <- m
> Y <- y
> IM <- influence.measures(M)
> IDs <- 2:4
> txtY <- c("DFBETAS_2", "DFFIT", "COVRATIO")
> shift <- c(0, 0, 1)
> CutOff <- c(2/sqrt(length(X)), 2 * sqrt(length(coef(M)/length(X))),
    3 * length(coef(M)/length(X)))
> par(mfrow = c(1, 3), mar = c(5, 5, 0.5, 0.5) + 0.05)
> for (id in 1:3) {
    ID <- IDs[id]
    plot(IM$infmat[, ID], type = "p", xlab = "index",
        ylab = txtY[id])
    L <- abs(IM$infmat[, ID] - shift[id]) > CutOff[id]
    abline(h = shift[id])
    abline(h = c(shift[id] - CutOff[id], shift[id] +
        CutOff[id]), lty = 2)
    if (sum(L) > 0) {
```

```
        text(as.numeric(rownames(IM$infmat)[L]), IM$infmat[L,
            ID], labels = rownames(IM$infmat)[L], cex = 0.75,
            pos = 2)
        pom <- data.frame(cbind(rownames(IM$infmat)[L],
            X[L], Y[L], IM$infmat[L, ID]))
        names(pom) <- c("index", "x", "y", colnames(IM$infmat)[ID])
        o <- order(-abs(IM$infmat[L, ID]))
        print(pom[o, ], digits = 4)
    }
}
```

```
    index  x                  y              dfb.x
106   106  3 -0.755022584278033  0.458400139122709
107   107  5 -0.653926467406664  0.426528105228927
126   126  2 -0.116533816255952 -0.363493603762246
22     22  6  0.254642218373581  0.318499569715897
88     88 -6 -0.281037529733112 -0.316400707045973
21     21  5 -0.579818495252942  0.228817552516062
87     87  6 -0.462035459596559  -0.20275156658593
83     83  2 -0.446287102628419 -0.178120421517005
```




Obrázek 16: Indexové grafy pro vybrané influenční míry pro *Logaritmy obsahu těžkých kovů v řece Moravě v průběhu let 1997 až 2009 (Cd v mg/kg).*

Identifikovaná pozorování bude třeba prověřit v tom smyslu, zda nedošlo k nějaké chybě například při přepisu. Rozhodně to neznamená okamžitě je vypustit ze zpracování.

## C. Úkol:

Na základě grafu predikovaných reziduí, Williamsova grafu a Pregibonova grafu proveďte identifikaci odlehlých a vlivných pozorování pro obsah olova v řece Moravě v letech 1997 až 2009.