MA012 Statistika II

7. Regresní diagnostika

Ondřej Pokora (pokora@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

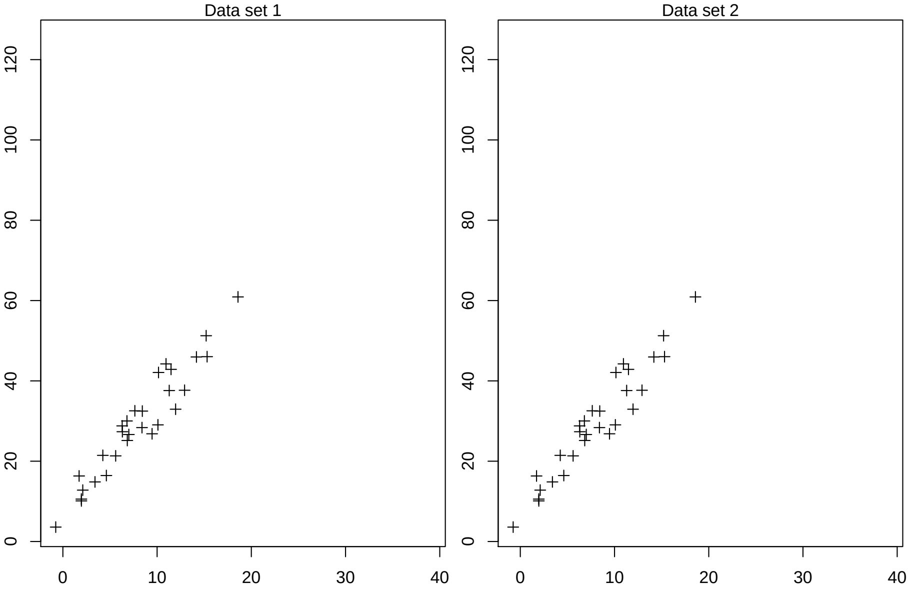
(podzim 2015)

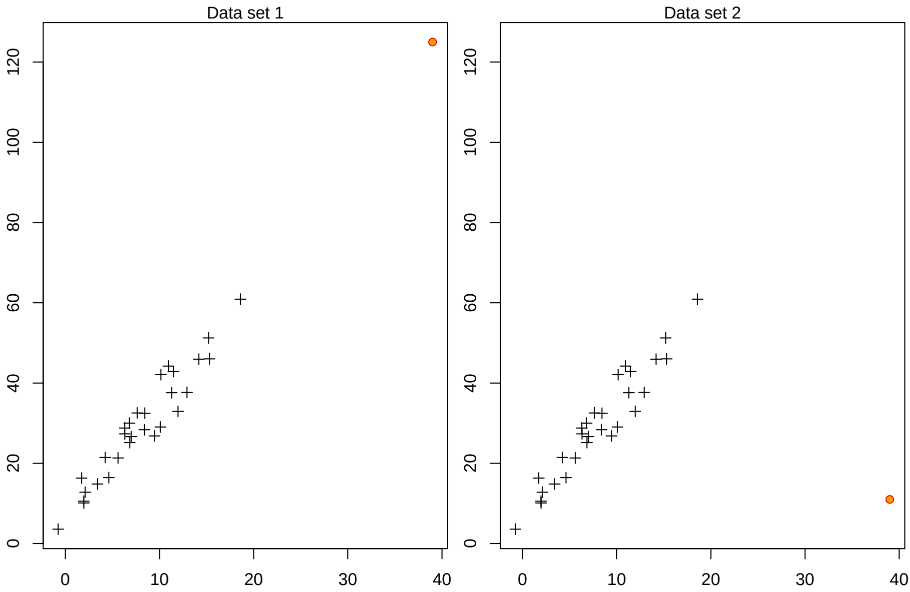
Motivace

V praxi se sestkáváme s jevem, že v souboru dat se vyskytují některé hodnoty výrazně se lišící od hodnot ostatních. V literatuře se rozvinuly dva směry, které se snaží s existencí takových hodnot vyrovnat:

- metody robustní statistiky,
- metody regresní diagnostiky.

Robustní diagnostika používá odlišné modely, testy, apod., které jsou navrženy tak, aby se v nich pokud možno eliminoval vliv výrazně odlišných hodnot v datovém souboru.
Regresní diagnostika, kterou budeme využívat, se snaží detekovat podezřelá data a pomocí vhodných statistických indikátorů dává statistikovi možnost rozhodnout se, jak s takovými hodnotami dále naloží.


Data set 1
Data set 2
0
20
40
60
80
100
120
0
10
20
30
40


Data set 1
Data set 2
0
20
40
60
80
100
120
0
10
20
30
40

Regresní diagnostika

V rámci regresní diagnostiky se zabýváme dvěma základními úlohami:

- detekcí neočekávaných hodnot v datovém souboru,
- rozhodnutím, zda takové hodnoty mohou významně ovlivnit statistickou analýzu datového souboru.

Definice 1 (Neočekávané hodnoty)

- **odlehlá pozorování (outliers)** – neočekávané hodnoty vysvštlované proměnné,
- **vybočující body (leverage points)** – neočekávané hodnoty vysvětlujících proměnných,
- data, která lze zařadit do obou výšeuvedených skupin.

Neočekávané hodnoty v datovém souboru

Výskyt odlehlých pozorování či vybočujících bodů nemusí nutně výzznamně ovlivnit analýzu datového souboru, neboť i taková měření mohou být v souladu s předpokládaným matematicko-statistickým modelem. Většinou však odlehlá či vybočujcící pozorování významně ovlivňují výsledky analýzy a proto je vhodné se v regresní diagnostice zabývat.

Významný vliv na výsledky však mohou mít i jiné body, než jen odlehlá či vybočující pozorování.

Definice 2 (Vlivný bod)

Jako **vlivné body** se označují všechny hodnoty datového souboru, které nějakým způsobem podstatně ovlivňují analýzu datového souboru, tj. některou z charakteristik spojených s odhadem vektoru parametrů v lineárním regresním modelu či s testováním hypotéz o parametrech.

Základním nástrojem regresní diagnostiky jsou rezidua.

Lineární regresní model

Pozorujeme dvojice $(x_i, Y_i),\ i = 1, \ldots, n$.
Předpokládáme

$$Y_i = m(x_i) + \varepsilon_i$$

x_i ... body pevného plánu,
Y_i ... naměřené hodnoty
ε_i ... chyby měření, $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$, $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$.

Hledáme **odhad** regresní funkce m.

Lineární regresní model

Celkově jsme dostali model

$$\begin{aligned} Y_1 &= \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ &\vdots \\ Y_n &= \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{aligned} \quad \underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}}_{\boldsymbol{X}\text{(matice plánu)}} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

O náhodných chybách $\varepsilon_1, \ldots, \varepsilon_n$ budeme předpokládat, že jsou

- **nesystematické**, což lze matematicky vyjádřit požadavkem, že $E\varepsilon_i = 0$, $i = 1, \ldots, n$, tj. $E\boldsymbol{\varepsilon} = \mathbf{0}$ a tedy $E\boldsymbol{Y} = \boldsymbol{X\beta}$
- **homogenní v rozptylu**, tj. že $D\varepsilon_i = \sigma^2 > 0$ pro $i = 1, \ldots, n$;
- jednotlivé náhodné chyby jsou **nekorelované**, tj. že $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j, i, j = 1, \ldots, n$, tj. $D\boldsymbol{Y} = D\boldsymbol{\varepsilon} = \sigma^2 \boldsymbol{I}_n$, takže i měření jsou nekorelovaná.

Terminologie

Používá se následující **terminologie** a značení

- parametry $\beta_1, \ldots, \beta_k$ se nazývají **regresní koeficienty** (**regression coefficients**);
- matice $\boldsymbol{X}$ obsahuje nenáhodné prvky x_{ij} a nazývá se **regresní maticí** nebo **maticí plánu** (**design matrix**);
- popsaný model souhrnně zapíšeme jako $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.

Takto zavedený model budeme nazývat **lineární regresní model**. Dále budeme předpokládat, že $n > k$ a o hodnosti matice $\boldsymbol{X}$ budeme předpokládat, že je rovna k, tj. $h(\boldsymbol{X}) = k$. Bude-li tento přepoklad splněn, budeme říkat, že jde **lineární regresní model plné hodnosti**. V tom případě jsou sloupce matice $\boldsymbol{X}$ nezávislé. V opačném případě, by bylo možné daný sloupec matice $\boldsymbol{X}$ napsat jako lineární kombinaci ostatních sloupců, což je možné interpretovat tak, že proměnná odpovídající danému sloupci je nadbytečná, protože ji lze vyjádřit jako lineární funkci ostatních proměnných.

Metoda nejmenších čtverců

Definice 3

Řekneme, že odhad $\widehat{\boldsymbol{\beta}}$ je odhadem parametru $\boldsymbol{\beta}$ **metodou nejmenších čtverců** (**Least Squares method**), jestliže

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2$$

Vlastnosti

Věta 4

V lineárním regresním modelu $\mathbf{Y} \sim \mathcal{L}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ *platí:*

$$\mathrm{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}, \qquad \mathrm{D}\boldsymbol{Y} = \sigma^2\, \boldsymbol{I}_n,$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

$$\mathrm{E}\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}, \qquad \mathrm{D}\widehat{\boldsymbol{\beta}} = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1},$$

$$S_e = S(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y},$$

$$s^2 = \frac{S_e}{n-k} \text{ je nestranným odhadem rozptylu } \sigma^2,$$

přičemž

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'.$$

Projekční matice

Definice 5 (Projekční matice)

Matice $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ se nazývá **projekční matice (hat matrix)**.

Věta 6

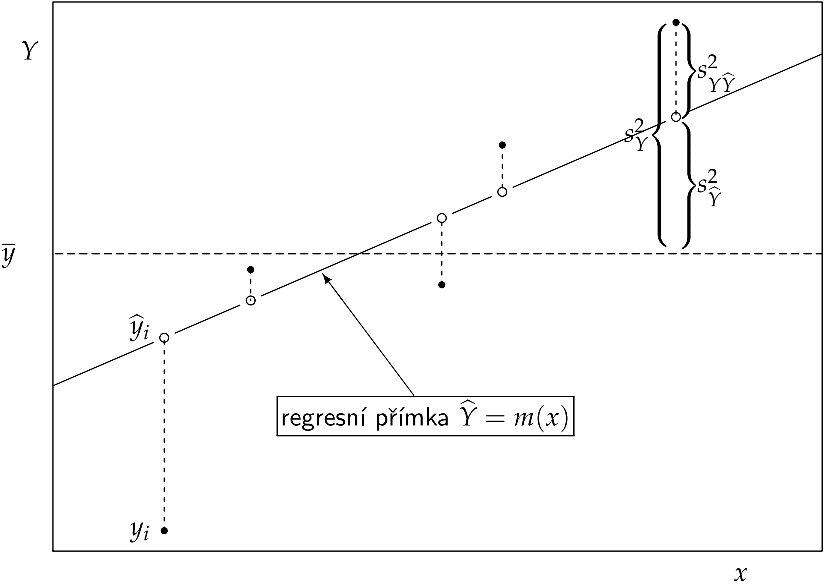
Pro odhady $\widehat{\boldsymbol{Y}}$ platí

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

Důkaz:

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

Graficky


Y
$\overline{y}$
$s^2_{Y\widehat{Y}}$
s^2_Y
$s^2_{\widehat{Y}}$
$\widehat{y}_i$
y_i
regresní přímka $\widehat{Y} = m(x)$
x

Vlastnosti projekční matice

Věta 7

Projekční matice $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$*:*

- *je symetrická řádu* n*, tj.* $\boldsymbol{H}' = \boldsymbol{H}$*,*
- *je* **idempotentní***, tj.* $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$*,*
- *má na hlavní diagonále hodnoty* $h_{ii} \in [0;1]$*,* $(i = 1, \ldots, n)$*,*
- *má* **stopu (Trace)** *rovnou počtu parametrů,* $Tr\,\boldsymbol{H} = \sum_{i=1}^{n} h_{ii} = k$*.*

Definice 8

Jednotlivé sloupce projekční matice $\boldsymbol{H}$ se nazývají **vlivové vektory**.
Číslo h_{ii} se nazývá **vliv** pozorování Y_i, $i = 1, \ldots, n$.

Věta 9

Průměrný vliv pozorování $Y_1, \ldots, Y_n$ *je rovný* $\overline{h} = \dfrac{k}{n}$*.*

Rezidua a jejich vlastnosti

Definice 10 (Rezidua)

Vektor **reziduí (residuals)** je vektor rozdílů skutečných hodnot a odhadů,

$$\boldsymbol{r} = (r_1, \ldots, r_n)' = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (Y_1 - \widehat{Y}_1, \ldots, Y_n - \widehat{Y}_n)'.$$

Věta 11

Pro vektor reziduí $\boldsymbol{r}$ a pro odhady $\widehat{\boldsymbol{Y}}$ platí:

$$\boldsymbol{r} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\varepsilon},$$

$$\mathrm{D}\boldsymbol{r} = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}),$$

$$\mathrm{D}\widehat{\boldsymbol{Y}} = \sigma^2\boldsymbol{H}.$$

Důkaz:

$$\boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} = \underbrace{(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta}}_{\boldsymbol{0}} + (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\varepsilon} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\varepsilon}.$$

Využití reziduí v regresní diagnostice

První tvrzení předchozí věty tedy říká, že rezidua $\boldsymbol{r}$ souvisí s náhodnými odchylkami $\boldsymbol{\varepsilon}$ od regresního modelu. Chybu ε_i měření Y_i bychom tedy mohli detekovat pomocí rezidua r_i.

Avšak v případě, že $h_{ii} \approx 1$, je odpovídající hodnota na hlavní diagonále matice $\boldsymbol{I}_n - \boldsymbol{H}$ rovna $1 - h_{ii} \approx 0$. To znamená, že v případě velkého vlivu pozorování Y_i se chyba tohoto pozorování nemusí projevit v reziduu r_i. Rezidua ostatních měření však ovlivnit může, pokud projekční matice $\boldsymbol{H}$ není diagonální.

V regresní diagnostice se proto zavádí a používají další typy reziduí.

Další typy reziduí

Definice 12

Normované rezidum (normalized/scaled) je

$$r_{\mathrm{N}\,i} = \frac{r_i}{s}.$$

Modifikované normované reziduum je

$$r_{\mathrm{M}\,i} = \frac{r_{\mathrm{N}\,i}}{\sqrt{n-k}} = \frac{r_i}{s\sqrt{n-k}}.$$

Standardizované reziduum je

$$r_{\mathrm{S}\,i} = \frac{r_{\mathrm{N}\,i}}{\sqrt{1-h_{ii}}} = \frac{r_i}{s\,\sqrt{1-h_{ii}}}.$$

Další typy reziduí

Definice 13

Predikované reziduum (predicted/crossvalidated) je reziduum v modelu bez i-tého pozorování, tzn.

$$r_{\mathrm{P}\,(i)} = Y_i - \widehat{Y}_{(i)},$$

kde $\widehat{Y}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)}$, a vektor parametrů $\widehat{\boldsymbol{\beta}}_{(i)}$ odhadneme v LRM bez i-tého pozorování.

Studentizované rezidum (jackknife residual) je

$$r_{\mathrm{J}\,(i)} = r_{\mathrm{P}\,(i)} \frac{\sqrt{1 - h_{ii}}}{s_{(i)}},$$

kde $s^2_{(i)}$ je odhad rozptylu náhodných chyb v modelu bez i-tého pozorování (tzn. s $n - k - 1$ stupni volnosti).

DFFIT reziduum je

$$d_i = \frac{\widehat{Y}_i - \widehat{Y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}}.$$

Další typy reziduí

Definice 14

Vektor **Parciálních reziduí** $\mathbf{r}_{[j]} = (r_{1[j]}, \ldots, r_{n[j]})'$ je tvořen rezidi v modelu bez j-tého parametru β_j, tzn.

$$r_{i[j]} = Y_i - \widehat{Y}_{i[j]},$$

kde $\widehat{Y}_{i[j]} = \boldsymbol{X}_{[j]} \widehat{\boldsymbol{\beta}}_{[j]}$ je odhad v modelu, v němž ve vektoru parametrů $\widehat{\boldsymbol{\beta}}_{[j]}$ chybí parametr β_j a v matici plánu $\boldsymbol{X}_{[j]}$ chybí odpovídající sloupec.

Vlastnosti reziduí

Následující tvrzení ukazuje, že jednotlivá rezidua je možno počítat

Věta 15

$$r_{\mathrm{P}\,i} = \frac{r_i}{1 - h_{ii}},$$

$$r_{\mathrm{J}\,(i)} = \frac{r_i}{s_{(i)}\sqrt{1 - h_{ii}}}, \qquad s_{(i)}^2 = s^2 - \frac{r_i^2}{(n-k)(1-h_{ii})},$$

$$d_i = r_{\mathrm{J}\,(i)}\sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

Pokud mají veličiny $Y_1, \ldots, Y_n$ normální rozdělení pravděpodobnosti, potom

$$\boldsymbol{r} \sim \mathrm{N}_n\left(\boldsymbol{0},\ \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})\right),$$

$$r_{\mathrm{J}\,(i)} \sim t(n-k).$$

Odlehlá pozorování

Definice 16

Pozorování Y_i je **odlehlé**, jestliže $\mathrm{E}\varepsilon_i \neq 0$.

Věta 17

Hypotézu $H_0 : \mathrm{E}\varepsilon_i = 0$ proti $H_1 : \mathrm{E}\varepsilon_i \neq 0$ zamítneme na hladině významnosti α, tzn. pozorování Y_i je odlehlé, pokud

$$\left| r_{\mathrm{J}\,(i)} \right| \geq t_{1-\alpha/2}(n-k).$$

Pro $n-k > 30$ lze použít aproximaci podmínky ve tvaru $\left| r_{\mathrm{J}\,(i)} \right| \geq 2$.

Lze využít také DFFIT reziduí.
Pro $n-k > 30$ detekujeme jako odlehlá ta pozorování, pro něž platí

$$|d_i| > 2\sqrt{\frac{k}{n}}.$$

Vliv pozorování – Cookova vzdálenost

Definice 18 (Cookova vzdálenost)

Pro měření vlivu i-tého pozorování na hodnotu odhadu $\widehat{\boldsymbol{\beta}}$ se používá tzv. **Cookova vzdálenost**

$$D_i = \frac{(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(i)})'(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(i)})}{k\,s^2} = \frac{r^2_{\mathrm{J}\,(i)}}{k}\,\frac{h_{ii}}{1-h_{ii}}.$$

Cookova vzdálenost je euklidovská vzdálenost mezi vektory predikce $\widehat{\boldsymbol{Y}}$ ze všech porozorvání a predikce $\widehat{\boldsymbol{Y}}_{(i)}$ při vynechání i=tého pozorování, a souvisí s konfidenčním elipsoidem odhadů. Vyjadřuje vliv i-tého bodu pouze na odhady $\widehat{\boldsymbol{\beta}}$, nikoliv na odhad rozptylu σ^2 náhodných chyb.

Věta 19

V praxi se obvykle za vlivný bod označuje takový, pro nějž je $D_i > 1$.

Vliv pozorování – další možnosti

Definice 20 (Welschova-Kuhova vzdálenost)

Pro měření vlivu i-tého pozorování simultánně na hodnotu odhadu $\widehat{\boldsymbol{\beta}}$ i na odhad rozptylu σ^2 náhodných chyb se používá statistika

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

Definice 21 (Parciální vliv)

Pro měření vlivu i-tého pozorování na hodnotu j-té složky $\widehat{\beta}_j$ odhadu $\widehat{\boldsymbol{\beta}}$ je navržena statistika

$$DFBETAS_{ij} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j\,(i)}}{\sqrt{\mathrm{D}\widehat{\beta}_j}}.$$

Věta 22

Obvykle se parciální vliv považuje za prokázaný, pokud $DFBETAS_{ij} > 2\sqrt{\frac{k}{n}}$.

Vliv pozorování – variační poměr

Definice 23 (Variační poměr)

Pro měření vlivu i-tého pozorování na kovariační matici $\mathrm{D}\widehat{\boldsymbol{\beta}}$ (spec. na rozptyly) vektoru odhadů $\widehat{\boldsymbol{\beta}}$ je navržena statistika

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

Věta 24

Jako pozorování mající vliv na kovariační matici $\mathrm{D}\widehat{\boldsymbol{\beta}}$ odhadů $\widehat{\boldsymbol{\beta}}$ se doporučuje považovat ta, pro něž je $|COVRATIO_i - 1| > 3\frac{k}{n}$.

Grafické nástroje

Graf predikovaných reziduí:

- osa x: predikovaná rezidua $r_{\mathrm{P}\,(i)}$
- osa y: rezidua r_i
- vybočující body jsou identifikovány polohou výrazně mimo přímku $y = x$
- odlehlá pozorování sice leží na přímce $y = x$ či v její blízkosti, ale jsou výrazně vzdálená od ostatních pozorování

Williamsův graf:

- osa x: vlivy h_{ii}
- osa y: jackknife rezidua $r_{\mathrm{J}\,(i)}$
- mezní linie pro odlehlá pozorování: $y = t_{1-\alpha/2}(n-k)$
- mezní linie pro vybočující body: $x = 2\frac{k}{n}$
- bublinkový graf: obsah bublinek reprezentujících jentolivá data je úměrný Cookově vzdálenosti D_i

Grafické nástroje

Pregibonův graf:

- osa x: vlivy h_{ii}
- osa y: kvadráty $r^2_{\mathrm{M}\,(i)}$ modifikovaných normovaných reziduí
- hraniční linie: $y = -x + 2\frac{k}{n}$ a $y = -x + 3\frac{k}{n}$
- pozorování zobrazená mezi oběma přímkami jsou vlivná, pozorování zobrazená nad horní přímkou jsou silně vlivná

Graf:

- osa x: vlivy h_{ii}
- osa y: Cookova vzdálenost D_i

Q-Q plot:

- osa x: teoretické kvantily standardizovaného normálního rozdělení $N(0;1)$
- osa y: standardizovaná rezidua $r_{\mathrm{S}\,i}$

Grafické nástroje

Indexové grafy:

- osa x: index i pozorování
- osa y: jednotlivé typy reziduí nebo vlivy h_{ii} nebo odhady $\beta_{(i)}$ nebo vzdálenosti (např. Cookova D_i)

Graf:

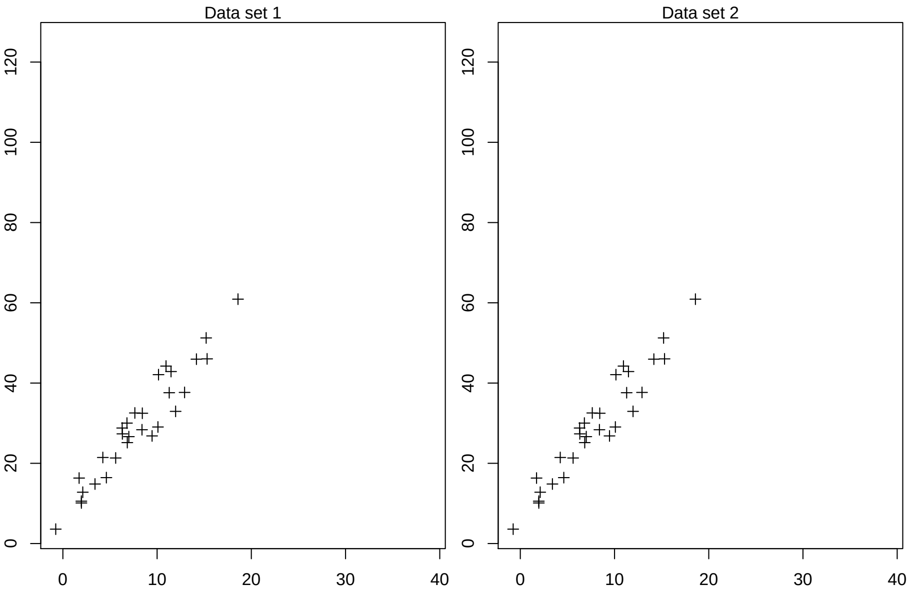
- osa x: odhady $\widehat{Y}_i$
- osa y: rezidua r_i

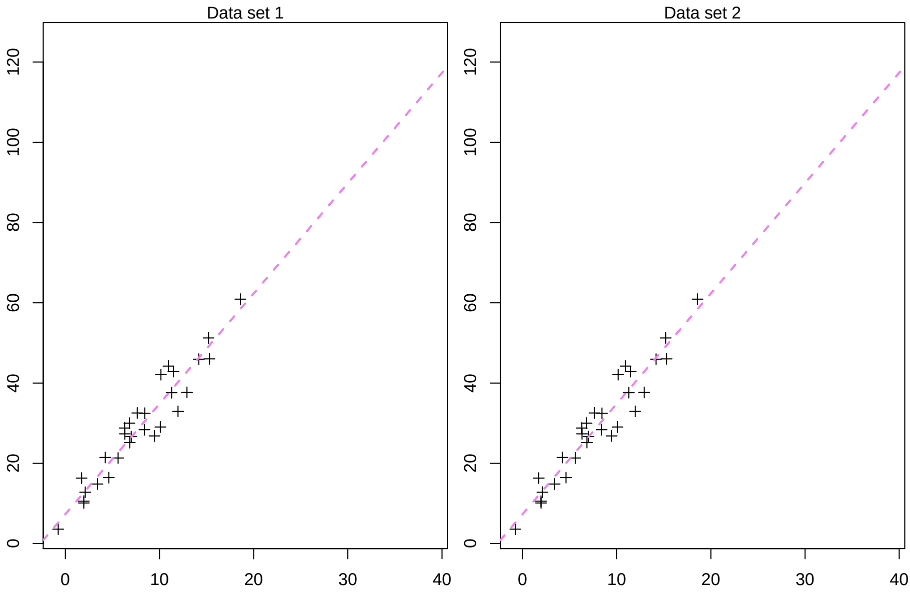
Graf:

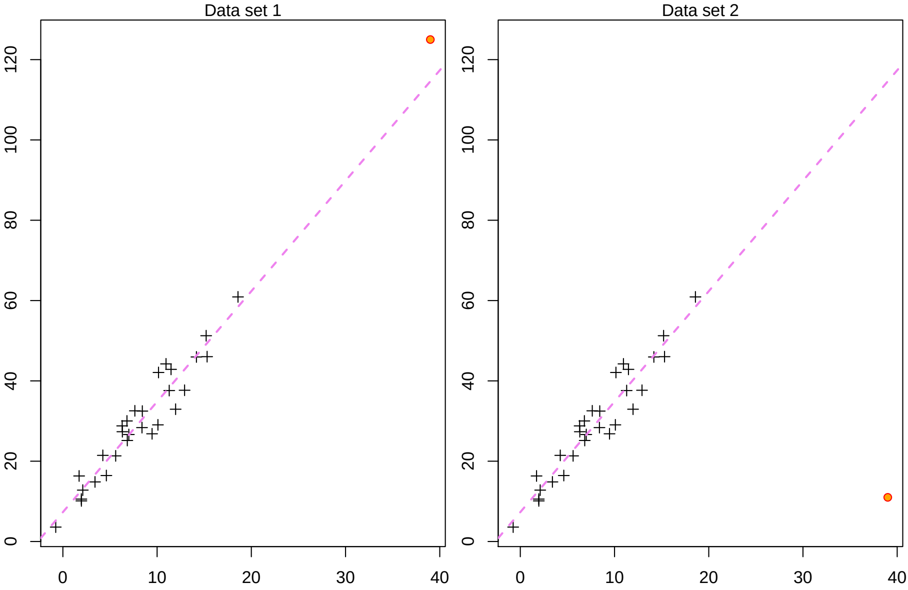
- osa x: odhady $\widehat{Y}_i$
- osa y: odmocniny absolutních hodnot standardizovaných reziduí, $\sqrt{|r_{\mathrm{S}i}|}$

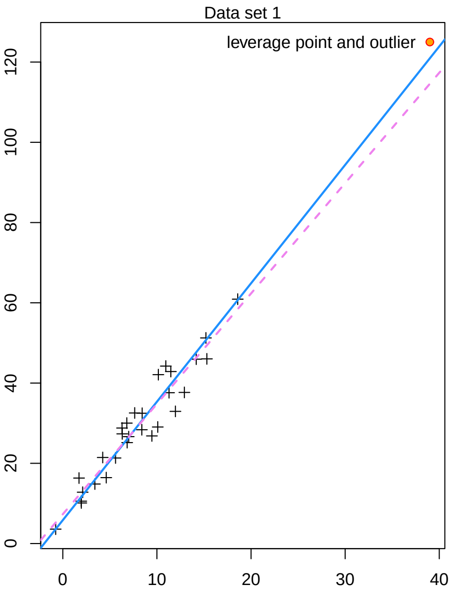
Scatter-plot:

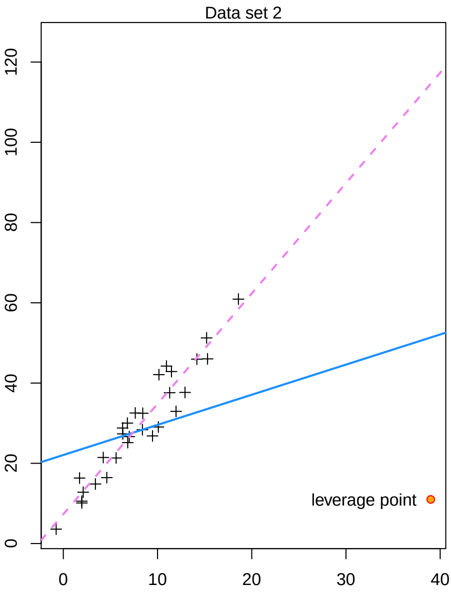
- osa x: nezávisle proměnná
- osa y: závisle proměnná


Data set 1
Data set 2
0
20
40
60
80
100
120
0
10
20
30
40


Data set 1
Data set 2
0
20
40
60
80
100
120
0
10
20
30
40


Data set 1
Data set 2
0
20
40
60
80
100
120
0
10
20
30
40


Data set 1
leverage point and outlier
0
20
40
60
80
100
120
0
10
20
30
40


Data set 2
leverage point
0
20
40
60
80
100
120
0
10
20
30
40

LRM pro Dataset 1

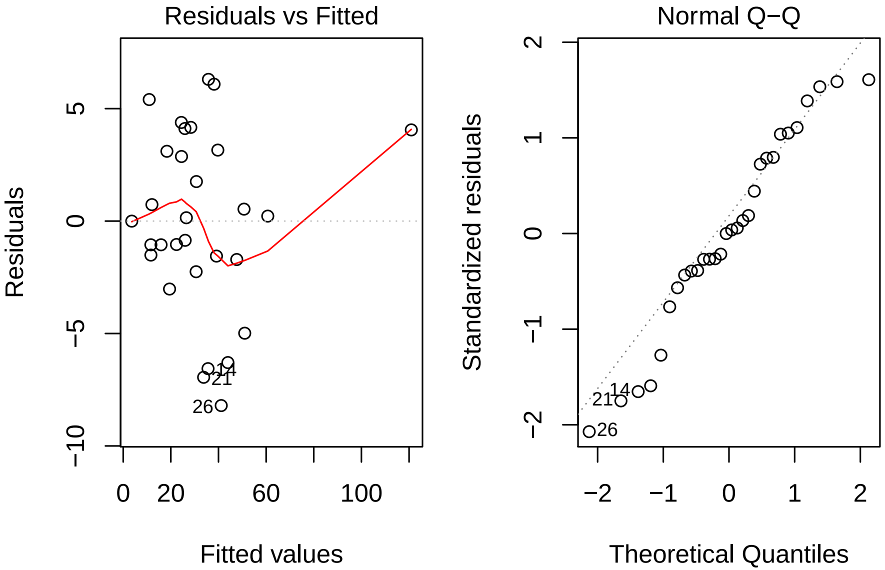
```
m1<-lm(y1~x1)
summary(m1)
```

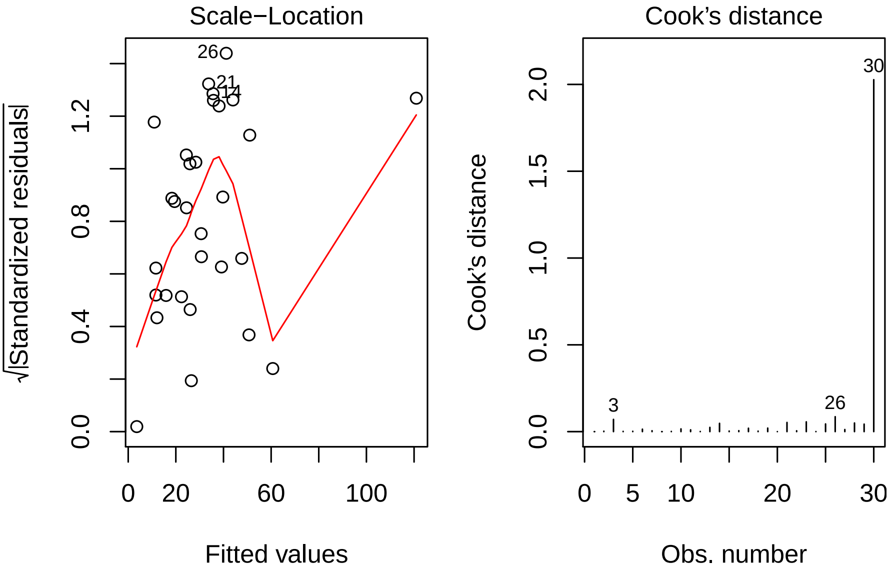
```
Residuals:
    Min      1Q  Median      3Q     Max 
-8.2031 -1.6710  0.0737  3.1437  6.3020 

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 ***

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
```

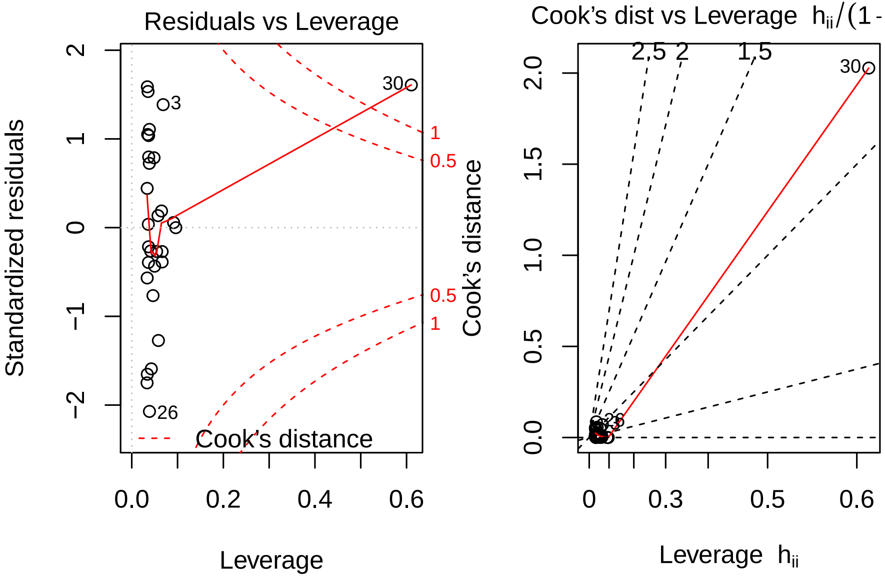
LRM pro Dataset 1

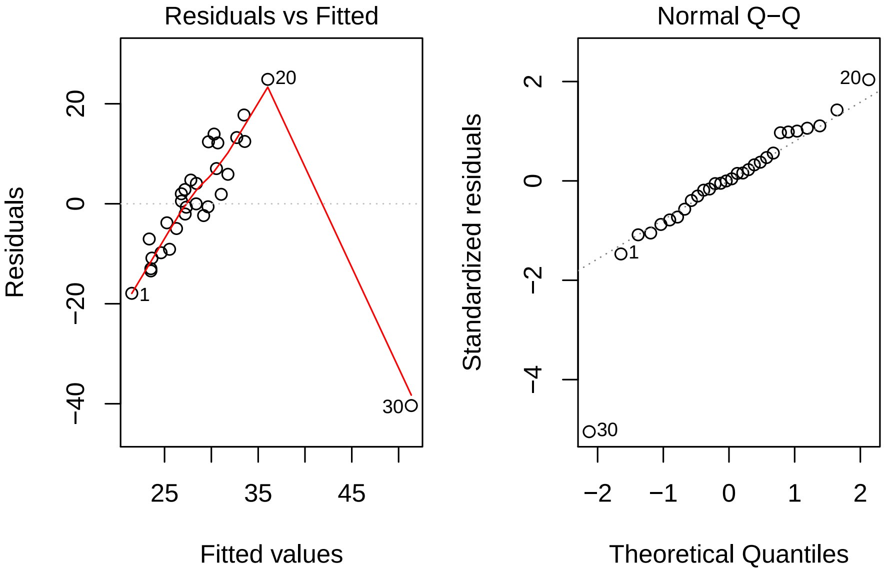
```
m2<-lm(y2~x2)
summary(m2)
```

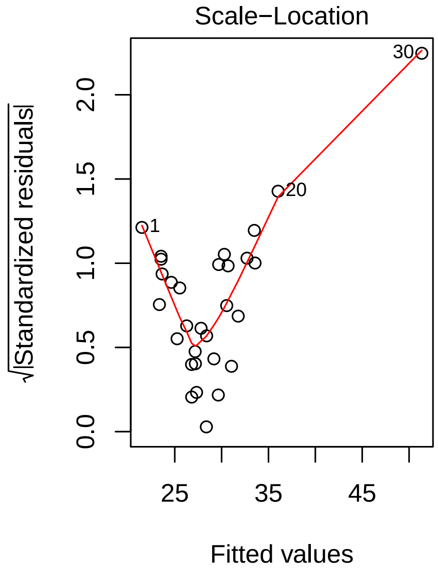
```
Residuals:
    Min      1Q  Median      3Q     Max
-40.357  -6.517   0.260   6.759  24.885

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 *

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
```

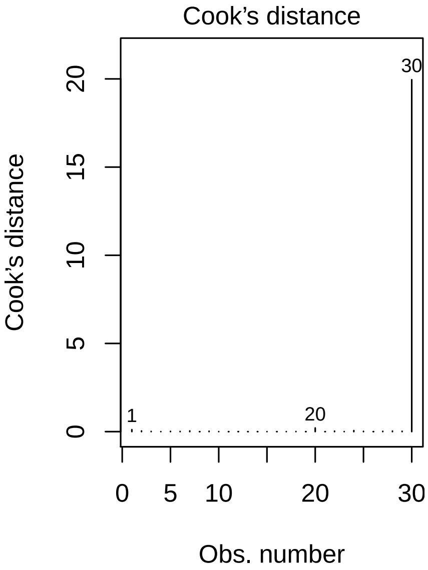

Residuals vs Fitted
Residuals
Fitted values
5
0
−5
−10
0
20
60
100
14
21
26
Normal Q−Q
Standardized residuals
Theoretical Quantiles
2
1
0
−1
−2
−2
−1
0
1
2
21
14
26

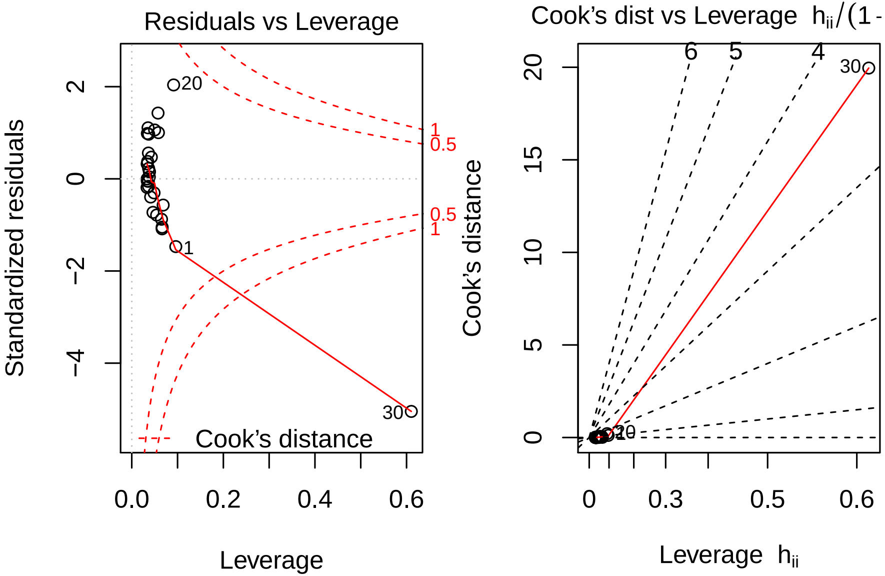

Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
0
20
60
100
Fitted values
26
21
14
Cook's distance
Cook's distance
0.0
0.5
1.0
1.5
2.0
0
5
10
20
30
Obs. number
3
26
30

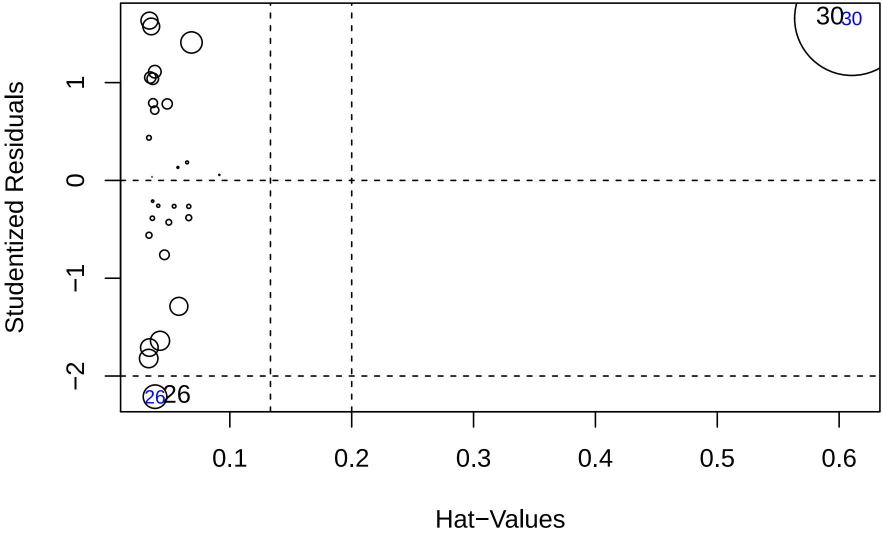

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
30
3
26
1
0.5
0.5
1
Cook's dist vs Leverage $h_{ii}/(1-$
Cook's distance
Leverage h_{ii}
2.5
2
1.5
30
0.0
0.5
1.0
1.5
2.0
0
0.3
0.5
0.6
−2
−1
0
1
2
0.0
0.2
0.4
0.6

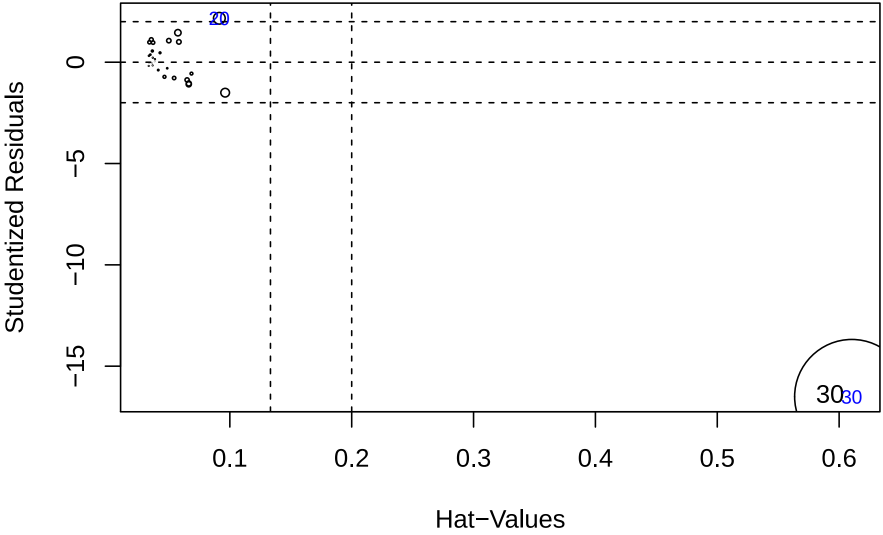

Residuals vs Fitted
Residuals
Fitted values
20
0
−20
−40
25
35
45
20
1
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
2
0
−2
−4
−2
−1
0
1
2
20
1
30


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
25
35
45
Fitted values
1
20
30


Cook's distance
Cook's distance
0
5
10
15
20
0
5
10
20
30
Obs. number
1
20
30


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
20
1
30
0.5
1
Cook's dist vs Leverage $h_{ii}/(1-$
Leverage h_{ii}
6
5
4
30


Studentized Residuals
Hat−Values
30
30
26
26
1
0
−1
−2
0.1
0.2
0.3
0.4
0.5
0.6


Studentized Residuals
0
−5
−10
−15
20
30
30
0.1
0.2
0.3
0.4
0.5
0.6
Hat−Values

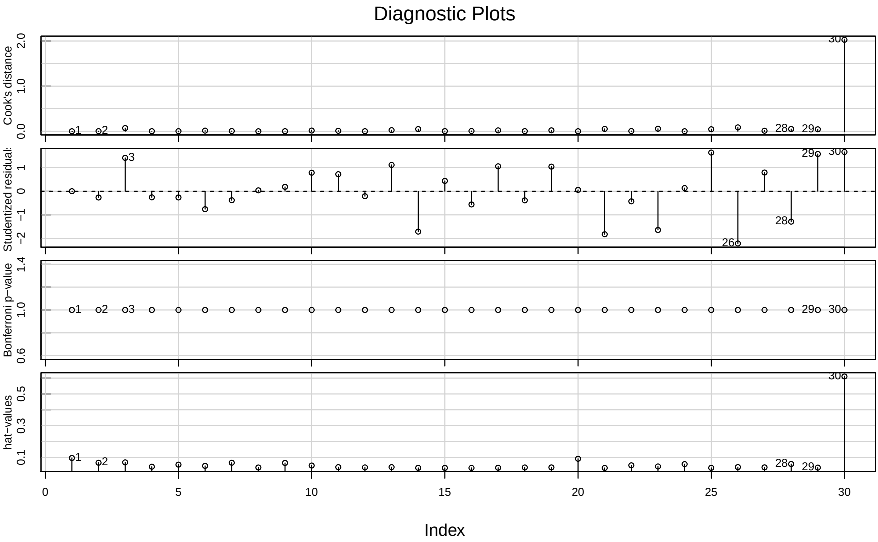
Podezřelá pozorování pro Dataset 1

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

   [,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
```

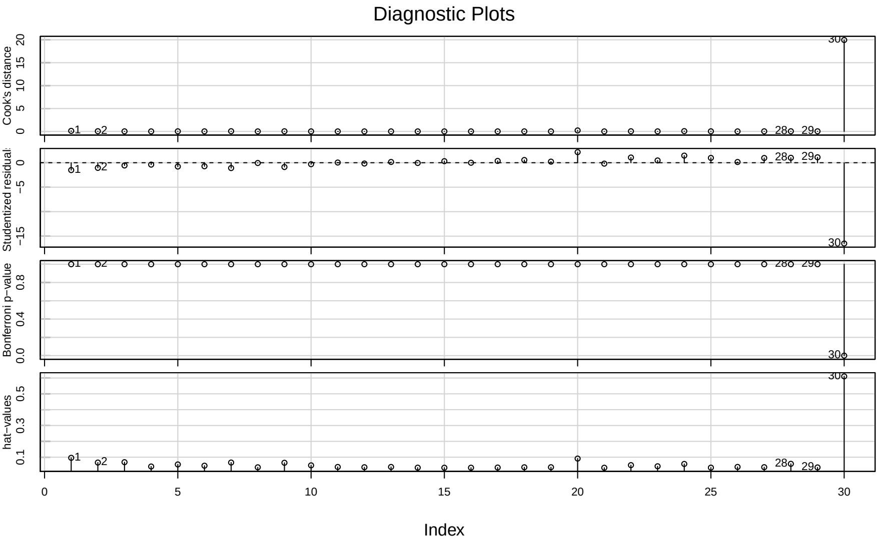
Podezřelá pozorování pro Dataset 2

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

   [,1]      [,2]     [,3]      [,4]       [,5]       [,6]
20   20 18.58604 60.91134   2.16843 0.09137258  0.2088146
30   30 39.00000 11.00000 -16.50284 0.61041902 19.9574746
```

Struktura tabuky: $i \mid x_i \mid Y_i \mid r_{\mathrm{J}\,(i)} \mid h_{ii} \mid D_i$



Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p−value
hat−values
Index
0
5
10
15
20
25
30
0.0
1.0
2.0
−2
−1
0
1
0.6
1.0
1.4
0.1
0.3
0.5
1
2
3
26
28
29
30

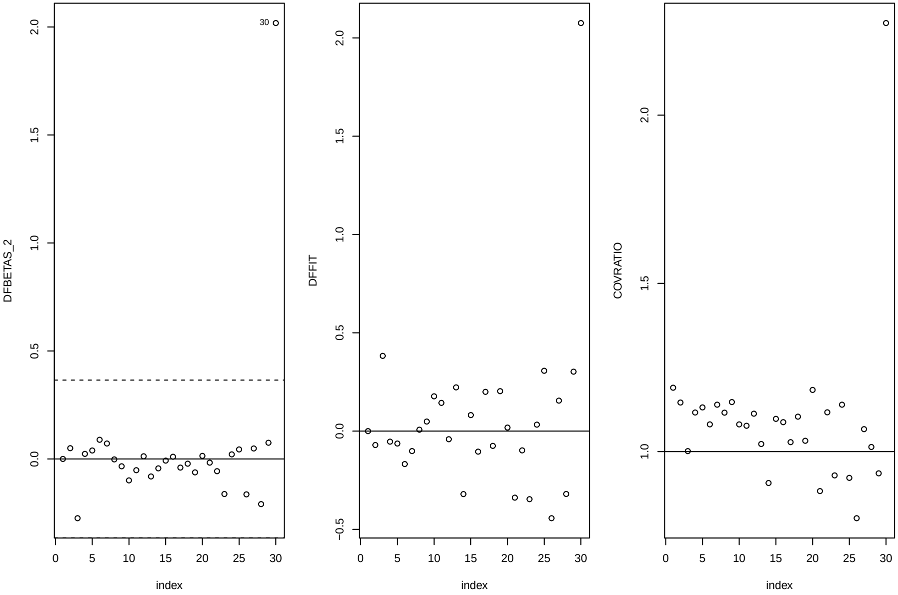

Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p−value
hat−values
Index

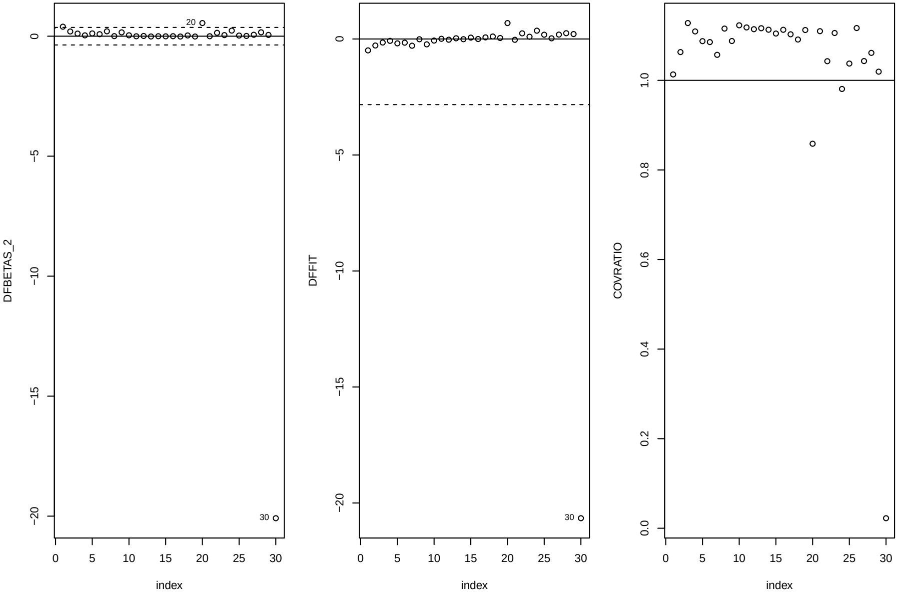
Test nulovosti středních hodnot reziduí Dataset 1

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

Test nulovosti středních hodnot reziduí Dataset 1

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


DFBETAS_2
DFFIT
COVRATIO
index
index
index
30



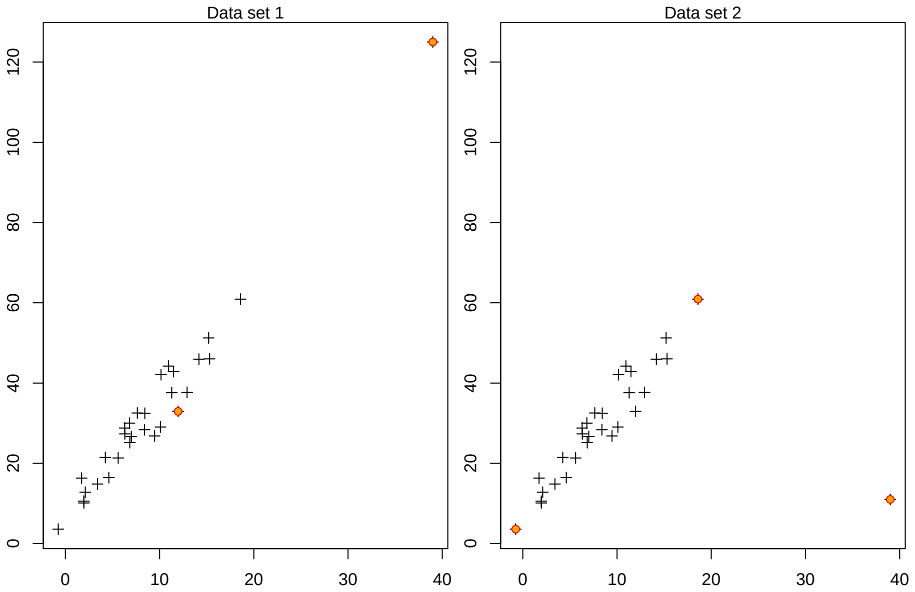
DFBETAS_2
index
20
30
DFFIT
index
30
COVRATIO
index

Podezřelá pozorování pro Dataset 1

```
  index  x   y  dfb.x1
1    30 39 125 2.01818
```

Podezřelá pozorování pro Dataset 1

```
   index          x        y     dfb.x2
30    30         39       11 -20.08538
20    20   18.58604 60.91134 0.5480413
1      1 -0.7576732 3.591746 0.3962400

  index  x  y     dffit
1    30 39 11 -20.65732
```


Data set 1
Data set 2
0
20
40
60
80
100
120
0
10
20
30
40

Regresní diagnostika v R

`library (car)`	
`model <-lm (... ~..., ...)`	
`plot (model, which=...)`	diagnostické grafy
`influence.measures (model)`	vlivy a příbuzné statistiky
`rstandard (model)`	standardizovaná rezidua
`rstudent (model)`	jackknife rezidua
`dffits (model)`	statistika DFFITS
`dfbetas (model)`	statistika DFBETAS
`covratio (model)`	variační poměr COVRATIO
`hatvalues (model)`	vlivy
`cooks.distance (model)`	Cookova vzdálenost
`influencePlot (model)`	diagnostické grafy pro vlivy
`infIndexPlot (model, ...)`	indexové grafy
`outlierTest (model)`	simultánní test odlehlých pozorování