Kazalo

1 DIAGNOSTIKA LINEARNEGA MODELA

Diagnostika je namenjena preverjanju predpostavk linearnega modela. V praksi na podlagi podatkov ocenimo parametre modela, za tem pa je potrebno preveriti, ali je bilo tako modeliranje upravičeno. Preveriti moramo sledeče:

- linearnost odvisnosti odzivne spremenljivke od napovednih spremenljivk. V primeru enostvane regresije mora razsevni grafikon y glede na x odražati linearno odvisnost, v primeru več napovednih spremenljivk uporabimo "grafikone dodane spremenljivke" in "grafikone parcialnih ostankov";

- varianca napak oziroma varianca odzivne spremenljivke pogojno na napovedne spremenljivke je konstantna (slika ostankov glede na napovedane vrednosti, razporeditev ostankov okoli vrednosti 0 mora biti slučajna, ne sme biti odvisna od napovedanih vrednosti);

- ker je pričakovana vrednost napak 0, se mora gladilnik na sliki ostankov glede na napovedane vrednosti čim bolje prilegati vodoravni osi;

- porazdelitev napak je normalna (QQ graf za standardizirane ostanke);

- napake so medsebojno neodvisne (težko preveriti, verjamemo, da so bili podatki pridobljeni z ustreznim načinom vzorčenja, princip slučajnosti; če so podatki izmerjeni v času, ostanke narišemo glede na čas meritve).

Osnova za diagnostiko modela so ostanki: navadni, standardizirani in studentizirani.

V kontekst diagnostike linearnega modela sodi tudi analiza t. i. posebnih točk. Z analizo posebnih točk ugotavljamo, kako dobro so posamezne vrednosti odzivne spremenljivke y_i opisane z modelom in kako posamezne y_i vplivajo na parametre in napovedi modela. Točke, za katere model ne da dobre napovedi, imenujemo regresijski osamelci, točke, ki znatno vplivajo na vrednosti ocen parametrov in posledično tudi na napovedi modela so t. i. vplivne točke, določamo jih na osnovi različnih mer vplivnosti (Cookova razdalja, DFFITS, DFBETAS,...). Točke, ki imajo velik vzvod, predstavljajo vzvodne točke.

1.1 Ostanki

Vektor ostankov izračunamo kot razliko vektorja odzivne spremenljivke $\mathbf{y}$ in z modelom napovedane vrednosti:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \tag{1}$$

Za ostanke pričakujemo, da bodo imeli podobne lastnosti kot napake: neodvisnost, normalna porazdelitev, konstantna varianca. Izkaže se, da se tem predpostavkam v kontekstu ostankov samo približamo.

V linearnem modelu je cenilka za $\boldsymbol{\beta}$ dobljena po metodi najmanjših kvadratov

$$\boldsymbol{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{2}$$

Videli smo, da z modelom napovedane vrednosti $\hat{\mathbf{y}}$ lahko izrazimo tudi z matriko $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{b} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}, \tag{3}$$

Vektor ostankov posledično lahko izrazimo

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}. \tag{4}$$

Matrika I je identična matrika reda n. Ker velja, da so y_i normalno porazdeljene slučajne spremenljivke, to velja tudi za ostanke. Zapišemo lahko

$$\begin{aligned}\mathbf{e} &= (\mathbf{I} - \mathbf{H})\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \varepsilon) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta} - \mathbf{H}\varepsilon - \boldsymbol{\varepsilon}.\end{aligned}$$

Ker je $\mathbf{HX} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{X}$, sledi

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}.$$

Ostanek ε_i tako zapišemo

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

Z večanjem n se elementi matrike $\mathbf{H}$ približujejo vrenosti 0 in ostanki postanejo dobra aproksimacija za napake. h_{ii}, $i = 1, ...n$ so diagonalni elementi matrike $\mathbf{H}$, tem vrednostim pravimo **vzvodi**. Vzvod je odvisen od velikosti vzorca in od vrednosti regresorjev, torej od položaja točke v regresorskem prostoru. Točke, ki so relativno daleč od centra regresorskega prostora, imajo velik vzvod. Pokažemo lahko, da je vrednost h_{ii} med $1/n$ in 1. Velja, da je $\sum_i^n h_{ii} = k + 1$, kjer je $k + 1$ število parametrov v modelu. Posledično je povprečni vzvod

$$\overline{h} = \frac{k+1}{n}. \tag{5}$$

Varianca ostankov $Var(\mathbf{e})$ je ob predpostavki $Var(\epsilon) = \sigma^2\mathbf{I}$:

$$Var(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\ (\sigma^2\mathbf{I})\ (\mathbf{I} - \mathbf{H})^{\mathrm{T}} = \sigma^2(\mathbf{I} - \mathbf{H})^2, \tag{6}$$

ker je $(\mathbf{I} - \mathbf{H})$ idempotentna matrika velja tudi

$$Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}), \tag{7}$$

To pomeni, da varianca ostankov ni konstantna, temveč je odvisna matrike $\mathbf{H}$, kar pomeni, da je odvisna od položaja točke v regresorskem prostoru.

- v nasprotju z napakami ε_i, $i = 1, ...n$, ostanki e_i niso nujno neodvisni, za $i \neq j$ je $Cov(e_i, e_j) = -\sigma^2 h_{ij}$, vrednosti h_{ij} se bližajo vrednosti 0, ko velikost vzorca n narašča.
- ostanki e_i se nagibajo k temu, da so v absolutnem smislu manjši kot napake ε_i tudi, ko so predpostavke modela izpolnjene, za vzvode h_{ii} velja, da so vedno pozitivne vrednosti, kar pomeni, da je varianca ostankov $Var(e_i) = \sigma^2(1 - h_{ii})$ vedno manjša kot varianca napak $Var(\varepsilon_i)$.
- točka z velikim vzvodom ima ostanek z majhno varianco in potencialno lahko predstavlja vplivno točko, ki potegne prilegano premico ali ravnino k sebi, da s tem zagotovi manjšo vrednost ostanka.
- zaradi naštetih lastnosti ostanki niso najboljše vrednosti za diagnostiko modela.

1.1.1 PRESS ostanki

Boljšo mero za oceno napake napovedi za posamezno točko dobimo s t. i. **PRESS ostanki**:

$$e_{i,-i} = y_i - \hat{y}_{i,-i}. \tag{8}$$

V (8) je $\hat{y}_{i,-i}$ napoved za y_i na podlagi modela, ki je narejen na vseh podatkih brez i-te točke.

Glede na definicijo PRESS ostankov jih izračunamo tako, da prilagodimo za vsako točko en model, torej n modelov. Pokaže se, da to ni potrebno. Izračunamo jih lahko na podlagi vzvodov h_{ii}, $i = 1, ..., n$.

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}}. \tag{9}$$

Ker za ostanke $_i$ velja, da so normalno porazdeljeni in ker so vzvodi odvisni samo od modelske matrike, so tudi PRESS ostanki porazdeljeni normalno. Njihova varianca je

$$Var(e_{i,-i}) = \frac{Var(e_i)}{(1 - h_{ii})^2} = \frac{\sigma^2(1 - h_{ii})}{(1 - h_{ii})^2} = \frac{\sigma^2}{1 - h_{ii}}.$$

To pomeni, da njihova varianca ni konstantna.

PRESS ostanki predstavljajo povečane navadne ostanke modela, to povečanje je odvisno od tega, kako vplivna je posamezna točka v procesu ocenjevanja parametrov modela.

1.1.2 Standardizirani ostanki

Ker varianca ostankov ni konstantna, je smiselno izračunati standardizirane ostanke:

$$\frac{y_i - \hat{y}_i}{\sigma\sqrt{1 - h_{ii}}}, \quad i = 1, ..., n. \tag{10}$$

Podobno izračunamo lahko standardizirane PRESS ostanke, ki so enaki standardiziranim ostankom:

$$\frac{\frac{y_i - \hat{y}_i}{\sigma\sqrt{1-h_{ii}}}}{\frac{\sigma^2}{1-h_{ii}}} = \frac{y_i - \hat{y}_i}{\sigma\sqrt{1 - h_{ii}}}, \quad i = 1, ..., n.$$

Ker σ v splošnem ne poznamo, v praksi σ ocenimo z $\hat{\sigma}$, tako izračunane standardizirane ostanke imenujemo tudi notranje studentizirani ostanki (*internally studentized residuals*).

$$e_{s_i} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \quad i = 1, ..., n. \tag{11}$$

Če je model sprejemljiv, imajo standardizirani ostanki konstantno varianco. Števec in imenovalec pri (11) sta povezana, saj je v oceni $\hat{\sigma}$ upoštevana tudi vrednost števca $y_i - \hat{y}_i$. Zato je

porazdelitev standardiziranih ostankov le približno t_{n-k-1}; če pa je $n >> k$, je porazdelitev približno $N(0,1)$.

Točke, ki imajo po absolutni vrednosti standardiziran ostanek več kot 2, $|e_{s_i}| > 2$, so kandidati za regresijske osamelce. Pri interpretaciji pa moramo biti previdni, saj je vzrok za veliko vrednost e_{s_i} lahko tudi napaka meritve konkretnega podatka, neizpolnjenost predpostavke linearnosti ali konstantne variance.

1.1.3 Studentizirani ostanki

Povezanosti med števcem in imenovalcem v (11), se znebimo z izračunom studentiziranih ostankov. Studentizirani ostanek e_{t_i} je podoben standardiziranemu ostanku e_{s_i}, $i = 1, ..., n$, vendar je ocena za standardno napako regresije izračunana brez upoštevanja i-te točke:

$$e_{t_i} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(-i)} \cdot \sqrt{1 - h_{ii}}}, \tag{12}$$

$\hat{\sigma}_{(-i)}$ je cenilka variance napake, ki je izračunana tako, da je v regresijskem modelu i-ta točka izpuščena. Posledično sta števec in imenovalec neodvisna. Teorija pove, da so studentizirani ostanki porazdeljeni t_{n-k-2}.

Studentizirani ostanki so primerni za odkrivanje regresijskih osamelcev, saj je $\hat{\sigma}_{(-i)}$ v primeru zelo odstopajoče vrednosti znatno manjša od $\hat{\sigma}$.

Za izračun $\hat{\sigma}_{(-i)}$ izrazimo z PRESS ostanki

$$\hat{\sigma}_{(-i)} = \frac{1}{n-k-2} \sum_{j \neq i} e^2_{j,-j}.$$

1.2 Graf dodane spremenljivke

Če imamo v modelu več številskih napovednih spremenljivk, **robni razsevni grafikon** odzivne spremenljivke glede na posamezno napovedno spremenljivko ne prikaže nujno pravega vpliva te spremenljivke na odzivno spremenljivko, saj ne upošteva vpliva ostalih spremenljivk v modelu. Za grafični prikaz vpliva posamezne spremenljivke na odzivno spremenljivko ob upoštevanju ostalih spremenljivk v modelu uporabimo t. i. **graf dodane spremenljivke** (*added variable plots* ali *partial regression plots*), ki ga naredi funkcija `avPlot` iz paketa `car` (Slika 4).

Recimo, da je v linearnem modelu k napovednih spremenljivk x_j

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, ..., n. \tag{13}$$

Graf dodane vrednosti spremenljivke x_j naredimo na podlagi ostankov dveh modelov. S prvim modelom napovemo y v odvisnosti od vseh ostalih napovednih spremenljivk razen x_j:

$$y_i^{(-j)} = \beta_0^{(j)} + \beta_1^{(j)} x_{i1} + \cdots + \beta_{j-1}^{(j)} x_{i,j-1} + \beta_{j+1}^{(j)} x_{i,j+1} + \beta_k^{(j)} x_{ik} + \varepsilon_i, \quad i = 1, ..., n. \tag{14}$$

Za ta model izračunamo ostanke $e_{i,y}^{(-j)}$:

$$e_{i,y}^{(-j)} = y_i - \hat{y}_i^{(-j)}, \quad i = 1, ...n. \tag{15}$$

Z drugim modelom napovemo x_j v odvisnosti od vseh ostalih napovednih spremenljivk:

$$x_{ij}^{(-j)} = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_{j-1} x_{i,j-1} + \gamma_{j+1} x_{i,j+1} + \gamma_k x_{ik} + \varepsilon_i, \quad i = 1, ..., n. \tag{16}$$

Ostanke tega modela označimo $e_{i,x_j}^{(-j)}$:

$$e_{i,x_j}^{(-j)} = x_{ij} - \hat{x}_{ij}^{(-j)}, \quad i = 1, ...n. \tag{17}$$

Ostanki $e_{i,y}^{(-j)}$ in $e_{i,x_j}^{(-j)}$ predstavljajo vrednosti y in x_j "očiščene" za vpliv ostalih spremenljivk v modelu. Graf dodane spremenljivke narišemo kot razsevni grafikon za odvisnost $e_{i,y}^{(-j)}$ od $e_{i,x_j}^{(-j)}$.

Za premico, ki opisuje odvisnost ostankov $e_{i,y}^{(-j)}$ od $e_{i,x_j}^{(-j)}$ velja:

- naklon premice, je enak oceni parametra b_j iz polnega modela;
- ostanki te premice so enaki ostankom polnega modela;
- standardna napaka naklona te premice je skoraj enaka standardni napaki ocene parametra b_j v polnem modelu (razlikuje se zaradi stopinj prostosti ostanka pri izračunu ocene s^2).

Opisane lastnosti grafa dodane spremenljivke omogočajo diagnostiko linearnega modela z več napovednimi spremenljivkami tudi v kontekstu analize nekonstantne variance in vplivnih točk, kar bomo videli na primerih, ki sledijo.

1.3 Graf parcialnih ostankov

Linearnost oziroma prisotnost nelinearnosti v modelu z več napovednimi spremenljivkami analiziramo na podlagi t. i. **grafa parcialnih ostankov** (*Component Plus Residual Plots*), ki jih nariše funkcija `crPlots()` iz paketa `car`.

Za model $y_i = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon_i$ izračunamo t. i. **parcialne ostanke** za vsako od napovednih spremenljivk kot vsoto navadnih ostankov e_i, $i = 1, ...n$, in vrednosti $b_j x_{ij}$, ki izraža z x_j pojasnjen del vrednosti odzivne spremenljivke y_i ob upoštevanju ostalih spremenljivk v modelu:

$$e_{i,x_j} = e_i + b_j x_{ij}. \tag{18}$$

Grafično prikažemo parcialne ostanke e_{i,x_j} v odvisnosti od x_{ij} in na njih prikažemo še gladilnik dobljen z neparametrično regresijo, ki jo izračuna funkcija `lowess()`. Ta graf pokaže morebitno nelinearnost v odnosu y in x_j, ki je nismo zaobjeli v linearnem modelu.

Če je v model vključena interakcija napovednih spremenljivk, funkcija `crPlots()` ni uporabna. Diagnostiko modela naredimo na podlagi grafov parcialnih ostankov s pomočjo funkcije `Effect()` iz paketa `effects`.
(https://socialsciences.mcmaster.ca/jfox/Courses/R/ICPSR/jss2627.pdf).

1.4 Primer: `pacienti`

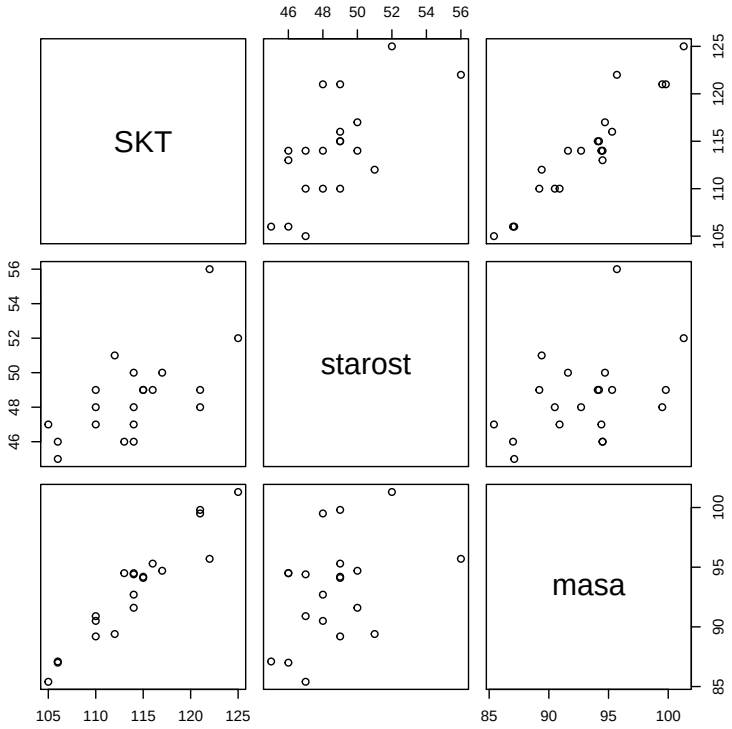
Imamo podatke za 20 moških s povišanim krvnim tlakom: krvni tlak (`SKT`, mm Hg), starost (`starost`, leta), telesna masa (`masa`, kg). Podatki so v datoteki `PACIENTI.txt`.

```
> summary(pacienti)
```

```
      SKT           starost          masa
 Min.   :105.0   Min.   :45.00   Min.   : 85.40
 1st Qu.:110.0   1st Qu.:47.00   1st Qu.: 90.22
 Median :114.0   Median :48.50   Median : 94.15
 Mean   :114.0   Mean   :48.60   Mean   : 93.09
 3rd Qu.:116.2   3rd Qu.:49.25   3rd Qu.: 94.85
 Max.   :125.0   Max.   :56.00   Max.   :101.30
```

Zanima nas, kako je `SKT` odvisen od `starost` in `masa` hkrati.
Matrika razsevnih grafikonov prikazuje, v kakšni zvezi so pari analiziranih spremenljivk. Uporabimo funkcijo `pairs()`. Je `SKT` linearno odvisen od `starost`? Je linearno odvisen od `masa`? Kakšna je povezava med `starost` in `masa`? Odgovori na ta vprašanja govorijo o robni odvisnosti SKT od napovednih spremenljivk, vsaka slika zase ne upošteva prisotnosti drugih napovednih spremenljivk.

```
> pairs(pacienti)
```

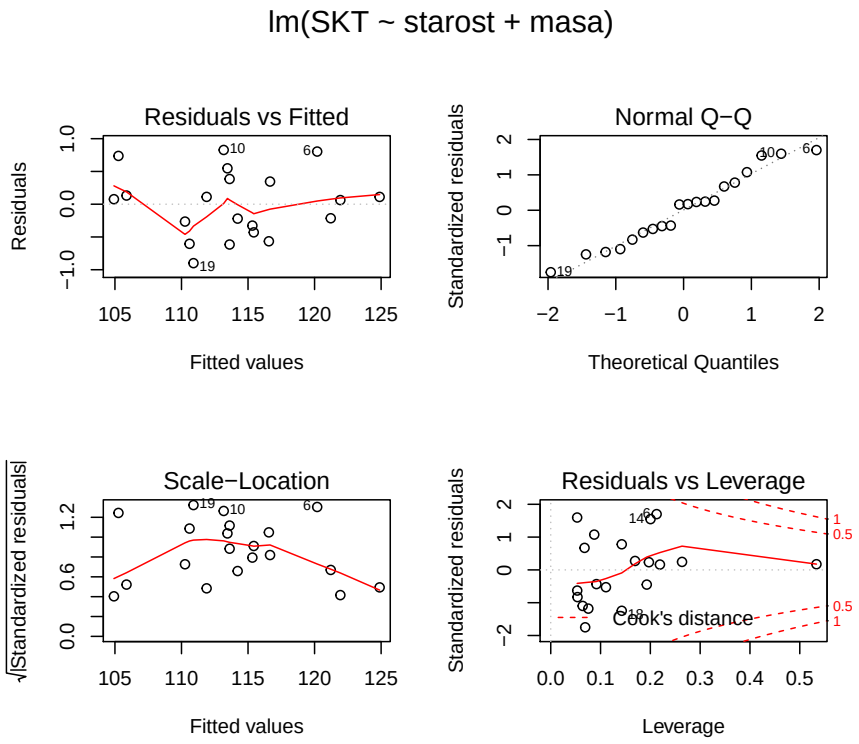


Slika 1: Matrika razsevnih grafikonov za `SKT`, `starost` in `masa` za 20 pacientov

Naredimo linearni regresijski model za napovedovanje `SKT` od `starost` in `masa`. Osnovni diagnostični grafi na Sliki 2 kažejo, da so predpostavke linearnega modela dokaj dobro izpolnjene, ni vplivnih točk niti kandidatov za regresijske osamelce.

```
> model.p<-lm(SKT ~ starost + masa, data=pacienti)
> coef(summary(model.p))
```

```
              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) -16.5793694 3.00745871 -5.51275 3.803805e-05
starost       0.7082515 0.05351399 13.23488 2.217640e-10
masa          1.0329611 0.03115625 33.15422 6.859831e-17
```

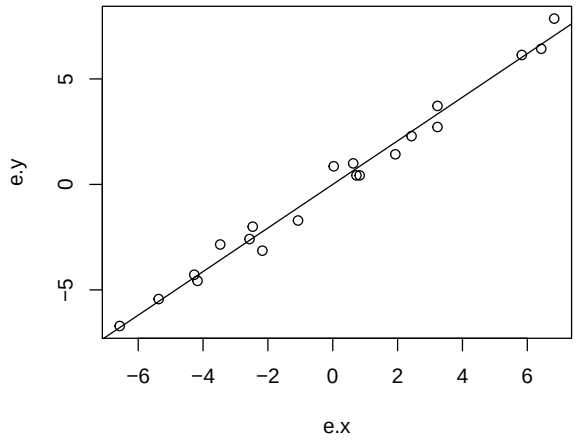


Slika 2: Ostanki za `model.p`

Če imamo v modelu več številskih napovednih spremenljivk, je poleg ostankov na Sliki 2 za prepoznavanje odstopanj od predpostavk linearnega modela informativno prikazati vpliv vsake od napovednih spremenljivk na odzivno spremenljivko ob upoštevanju ostalih spremenljivk v modelu. Za to naredimo grafe dodane spremenljivke. Za ilustracijo naredimo izračune in graf dodane spremenljivke za `masa`, ki prikazuje odvisnost `SKT` od `masa` ob upoštevanju `starost` v modelu:

```
> e.y <- residuals(lm(SKT~starost, data=pacienti))
> e.x <- residuals(lm(masa~starost, data=pacienti))
> mod.e <- lm(e.y~e.x)
> (b.e <- coef(summary(mod.e))[2,1])

[1] 1.032961

> (s.b.e <- coef(summary(mod.e))[2,2])

[1] 0.03027843
```

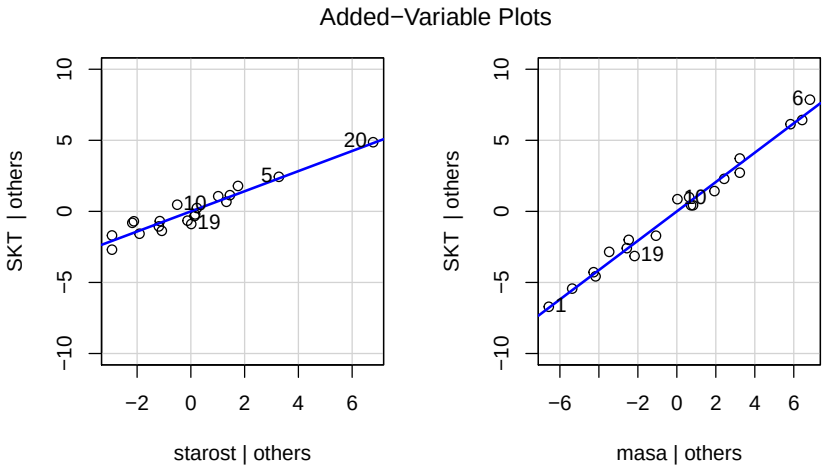
```
> plot(e.x, e.y)
> abline(reg=mod.e)
```



Slika 3: Graf dodane spremenljivke `masa`, odvisnost `SKT` od `masa` ob upoštevanju `starost`

Funkcija `avPlots` iz paketa `car` (Sliki 4) na podlagi `model.p` nariše grafikona dodane spremenljivke za `starost` in za `masa`. Po prednastavitvi velja `id=TRUE`, kar pomeni, da je argument `id` določen takole:
`id=list(method=list(abs(residuals(mod.e, type="pearson")), "x"), n=2)`.
Na grafih sta z vrednostjo `rownames()` označeni dve točki (`n=2`) z največjim standardiziranim ostankom (`abs(residuals(mod.e, type="pearson")`) in dve točki z največjo vrednostjo na x-osi oziroma največjim parcialnim vzvodom (`"x"`). Premici na Sliki 4 se dobro prilegata točkam, kar pomeni, da je odvisnost močna, hkrati ni videti prisotnosti nekonstantne variance, ki bi se odražala v neenakomerni porazdelitvi točk okoli premice.

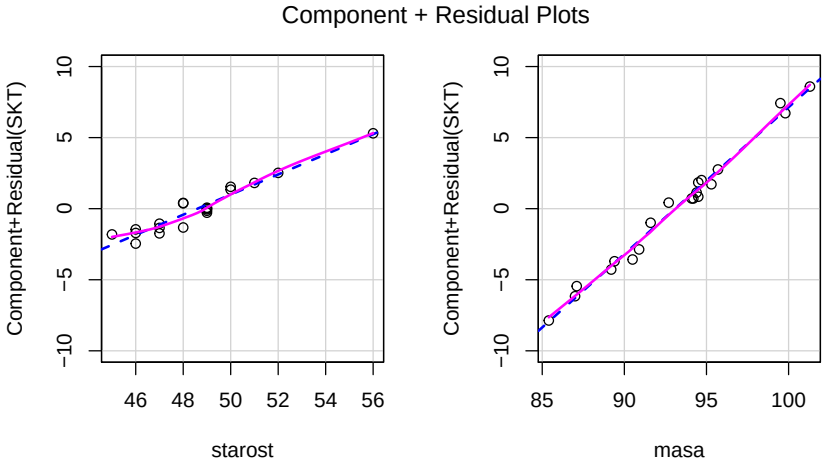
```
> library(car)
> avPlots(model.p, ylim=c(-10, 10))
```



Slika 4: Grafa dodane spremenljivke za `model.p`

Linearnost odvisnosti SKT od posamezne spremenljivke upoštevajoč drugo spremenljivko v modelu preverimo na podlagi grafikonov parcialnih ostankov, ki ga nariše funkcija `crPlot` (Slika 5). Gladilnik se dobro prilega premici, kar pomeni, da ni dodatne nelinearnosti v odvisnosti `SKT` od `starost` in `masa`.

```
> crPlots(model.p, ylim=c(-10, 10))
```



Slika 5: Grafa parcialnih ostankov za `model.p`

Izpišemo R^2 in ocene parametrov s pripadajočimi 95 % parcialnimi intervali zaupanja.

```
> summary(model.p)$r.squared

[1] 0.9913858

> model.p$coeff

(Intercept)     starost        masa
-16.5793694   0.7082515   1.0329611

> confint(model.p)

                 2.5 %      97.5 %
(Intercept) -22.9245526 -10.2341861
starost       0.5953468   0.8211561
masa          0.9672272   1.0986950
```

Sklepi:

- `starost` in `masa` v `model.p` pojasnita 99 % variabilnosti `SKT`;
- obe napovedni spremenljivki `starost` in `masa` sta zelo statistično značilni ($p < 0.0001$);
- izberemo poljubno vrednost za maso na intervalu 85 kg do 102 kg. Pri izbrani vrednosti za maso velja: če se starost poveča za 10 let, se SKT v povprečju poveča za 7.1 mm, pripadajoč 95 % IZ je (6.0 mm, 8.2 mm);
- izberemo poljubno vrednost za starost na intervalu 45 let do 56 let. Pri izbrani vrednosti za starost velja: če se masa poveča za 10 kg, se SKT v povprečju poveča 10.3 mm, pripadajoč 95 % IZ je (9.7 mm, 11.0 mm).

Ker so ocene parametrov linearnega modela porazdeljene po multivariatni normalni porazdelitvi, lahko pokažemo, da $100(1-\alpha)$ % **območje zaupanja za vse parametre modela hkrati** določimo na podlagi F-statistike:

$$F = \frac{(\mathbf{b}-\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{b}-\boldsymbol{\beta})}{(k+1)\hat{\sigma}^2},$$

ki je porazdeljena $F_{k+1,n-k-1}$.

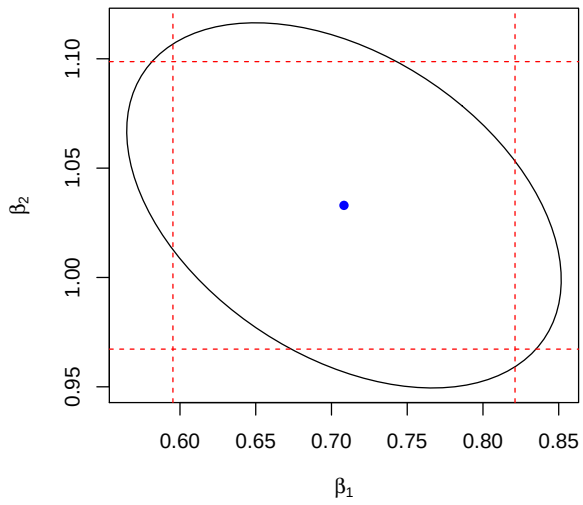
$100(1-\alpha)$ % območje zaupanja za $\boldsymbol{\beta}$ predstavlja vse vrednosti za $\boldsymbol{\beta}$, ki ustrezajo pogoju

$$P\left(\frac{(\mathbf{b}-\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{b}-\boldsymbol{\beta})}{(k+1)\hat{\sigma}^2} \leq F_{\alpha}(k+1, n-k-1)\right) = 1-\alpha. \tag{19}$$

Slika 6 prikazuje območje zaupanja za parametra β_1 in β_2 ob upoštevanju β_0 za `model.p`. Na

obeh slikah sta prikazana tudi parcialna intervala zaupanja. Pri grafičnem prikazu smo uporabili funkcijo `ellipse()` iz istoimenskega paketa, ki za dani model izračuna meje območja glede na 19.

```
> library(ellipse);
> plot(ellipse(model.p, which=c(2,3)), type="l",
+      xlab=expression(beta[1]), ylab=expression(beta[2]))
> abline(v=confint(model.p)[2,], h=confint(model.p)[3,], lty=2, col="red")
> points(model.p$coef[2],model.p$coef[3], pch=16, col="blue")
```



Slika 6: Primer 95 % območja zaupanja za parametra β_1 in β_2 ob upoštevanju β_0 za model `model.p` (elipsa) in meje 95 % parcialnih intervalov zaupanja za β_1 in za β_2 iz istega modela, modra pika označuje cenilki b_1 in b_2

Za ilustracijo izračunajmo variančno kovariančno matriko za `model.p` s funkcijo `vcov` in "peš"

$$Var(\mathbf{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.$$

```
> vcov(model.p)
```

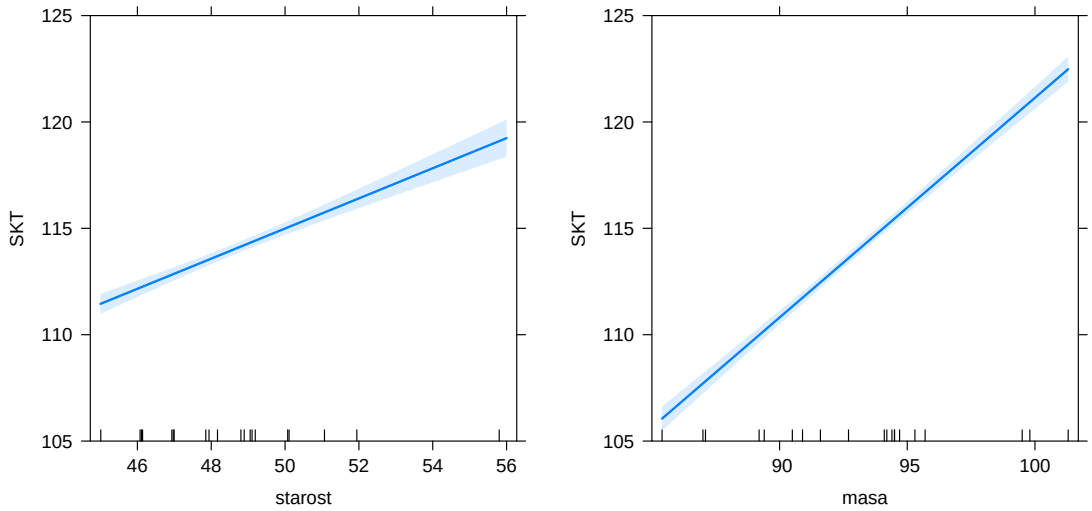
```
             (Intercept)       starost          masa
(Intercept)  9.04480792 -0.0759540289 -0.0573558287
starost     -0.07595403  0.0028637468 -0.0006791714
masa        -0.05735583 -0.0006791714  0.0009707118
```

```
> n <- length(model.p$residuals)
> b <- model.p$coef
```

```
> k <- length(b)-1
> (s2 <- sum(model.p$residuals^2)/(n-k-1))

[1] 0.2837604

> X <- model.matrix(model.p)
> head(X)

  (Intercept) starost masa
1           1      47 85.4
2           1      49 94.2
3           1      49 95.3
4           1      50 94.7
5           1      51 89.4
6           1      48 99.5

> (A <- solve(t(X) %*% X))

            (Intercept)      starost         masa
(Intercept)  31.8748075 -0.267669593 -0.202127676
starost      -0.2676696  0.010092130 -0.002393468
masa         -0.2021277 -0.002393468  0.003420885

> (s2*A)

            (Intercept)       starost          masa
(Intercept)  9.04480792 -0.0759540289 -0.0573558287
starost     -0.07595403  0.0028637468 -0.0006791714
masa        -0.05735583 -0.0006791714  0.0009707118
```

Napovedane vrednosti za `SKT` dobljene z `model.p` ležijo na ravnini. Za grafični prikaz napovedi in pripadajočih intervalov zaupanja za povprečne napovedi uporabimo funkcijo `predictorEffects` iz paketa `effects`. Narisali bomo dve sliki, napovedi za `SKT` glede na `starost` pri povprečni masi 93.1 kg ter napovedi za `SKT` glede na `masa` pri povprečni starosti 48.6 let.

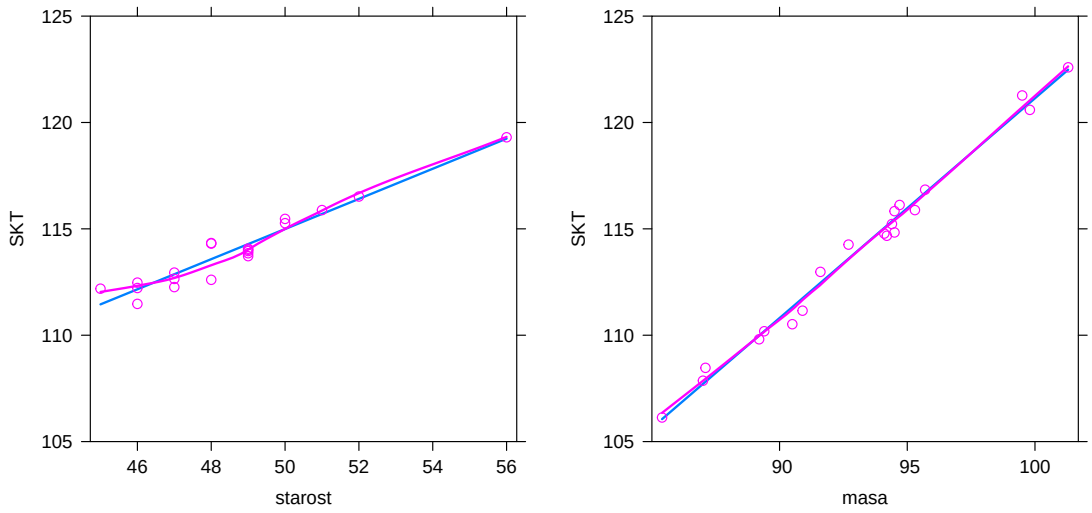
```
> library(effects)
> plot(predictorEffects(model.p, ~.), ylim=c(105,125),main="")
```



Slika 7: Napovedane vrednosti in 95 % intervali zaupanja za povprečen SKT za `model.p`; levo: glede na starost pri povprečni masi 93.1 kg, desno: glede na maso pri povprečni starosti 48.6 let

Funkcijo `predictorEffects()` lahko uporabimo tudi za grafične prikaze parcialnih ostankov. To velja za enostavne in tudi za kompleksnejše linearne modele, v katerih so vključeni tudi interakcijski členi, katerih pomen bomo predstavili v enem od naslednjih poglavij. Kot smo videli, za modele brez interakcijskih členov enakovredne grafične prikaze naredi funkcija `crPlot()`.

```
> library(effects)
> plot(predictorEffects(model.p, ~., partial.residuals=TRUE),
+      ci.style="none", ylim=c(105,125),main="")
```



Slika 8: Napovedane vrednosti za povprečen SKT za `model.p` in parcialni ostanki z gladilnikom; levo: glede na starost pri povprečni masi 93.1 kg, desno: glede na maso pri povprečni starosti 48.6 let

Izračunajmo napoved za `SKT` za paciente stare 50 let z maso 100 kg in 95 % IZ za povprečno napoved in za posamično napoved.

```
> vrednosti<-data.frame(starost=50, masa=100)
> povp.napoved<-predict(model.p, vrednosti, interval="confidence")
> pos.napoved<-predict(model.p, vrednosti, interval="prediction")
> print(data.frame(cbind(vrednosti, povp.napoved, pos.napoved)))

  starost masa      fit      lwr      upr    fit.1    lwr.1    upr.1
1      50  100 122.1293 121.6436 122.6151 122.1293 120.9049 123.3537
```

Za osebe stare 50 let z maso 100 kg je napovedana vrednost za SKT 122.1 mm, 95 % IZ za povprečno napoved je (121.6 mm, 122.6 mm), 95 % IZ za posamično napoved je (120.9 mm, 123.4 mm).

1.5 Posebne točke

Posebne točke v regresijski analizi so enote, ki zelo odstopajo od ostalih glede na določene kriterije. Te točke prispevajo zelo pomembno informacijo o regresijskem modelu, zato je vedno

potrebna njihova analiza. Pogledali bomo tri vrste posebnih točk: **regresijske osamelce**, **vzvodne točke** in **vplivne točke**.

Posebne točke bomo najprej predstavili na primeru.

1.5.1 Primer: POSTAJE, 1. del

Za meteorološke postaje (datoteka POSTAJE.txt) analizirajmo odvisnost letne količine padavin, `padavine`, od nadmorske višine, `z.nv`. Podatki so za leto 1992, `padavine` so izražene v mm, nadmorska višina `z.nv` v metrih. Za večino postaj imamo tudi podatke za geografsko dolžino in širino, `x.gdol` in `y.gsir`; to so Gauss-Krugerjeve koordinate, ki so izražene v metrih.

```
> postaje<-read.table("POSTAJE.txt", header=TRUE, sep="\t")
> head(postaje)
```

```
               Postaja x.gdol y.gsir z.nv padavine
1          Babno polje 464930  56264  756     1643
2            Bizeljsko 554193  97520  170     1048
3 Brezovica pri Topolu 451721 105620  708     1715
4                Brnik 459888 119639  362     1396
5               Bukovo 415040 112316  715     2089
6                Celje 519274 122872  244     1169
```

```
> summary(postaje)
```

```
                 Postaja       x.gdol           y.gsir           z.nv
 Babno polje          : 1   Min.   :387744   Min.   : 36680   Min.   :  92.0
 Bizeljsko            : 1   1st Qu.:416388   1st Qu.: 76441   1st Qu.: 260.0
 Brezovica pri Topolu: 1   Median :442091   Median : 97325   Median : 480.0
 Brnik                : 1   Mean   :457490   Mean   : 97119   Mean   : 520.9
 Bukovo               : 1   3rd Qu.:488515   3rd Qu.:119842   3rd Qu.: 700.0
 Celje                : 1   Max.   :612650   Max.   :165750   Max.   :2514.0
 (Other)              :61   NA's   :2        NA's   :2
    padavine
 Min.   : 807
 1st Qu.:1296
 Median :1541
 Mean   :1633
 3rd Qu.:1891
 Max.   :3207
```

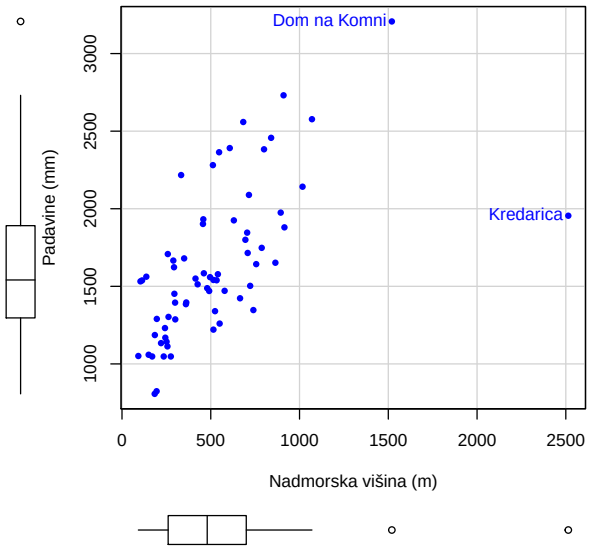
```
> rownames(postaje)<-postaje$Postaja
```

Opomba: z ukazom `rownames` vsaki vrstici damo ime, to ime služi za identifikacijo postaje na slikah in pri določenih izpisih. Dve postaji nimata podatka za `x.gdol` in/ali za `y.gsir`. S funkcijo `is.na` ugotovimo, kateri postaji sta to.

```
> rownames(postaje)[is.na(postaje$x.gdol)]

[1] "Jezersko" "Ozeljan"

> rownames(postaje)[is.na(postaje$y.gsir)]

[1] "Jezersko" "Ozeljan"
```

Slika 9 prikazuje odvisnost padavin od nadmorske višine. V ukazu `scatterplot` iz paketa `car` se z argumentom `id` na sliki izpišeta imeni dveh izstopajočih postaj. Na sliki sta prikazana tudi okvirja z ročaji za `padavine` in za `z.nv`.

```
> library(car)
> scatterplot(padavine~z.nv, regLine=F, smooth=FALSE, boxplots='xy',
+             xlab=c("Nadmorska višina (m)"), ylab=c("Padavine (mm)"),
+             data=postaje, pch=16, id=list(n=2, location="lr")) # id=TRUE

Dom na Komni    Kredarica
          10           23
```

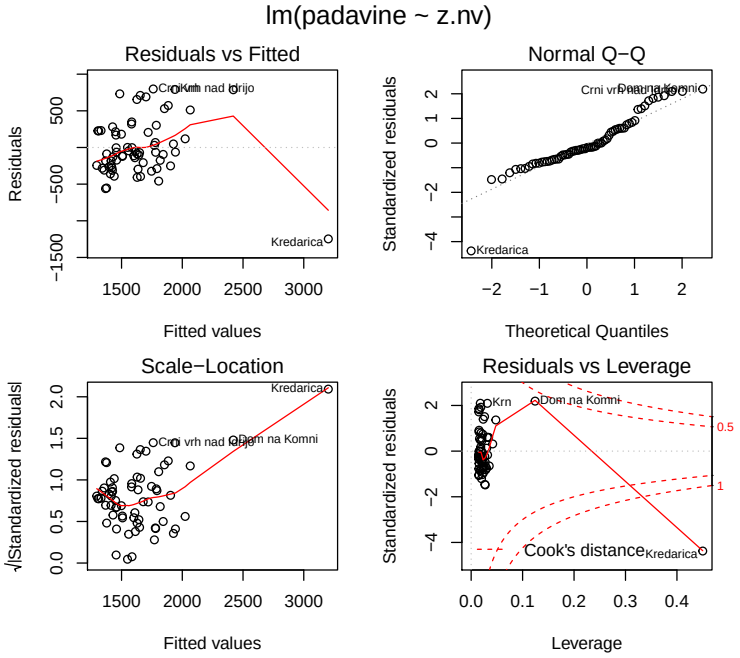


Slika 9: Odvisnost letne količine padavin od nadmorske višine za 67 postaj v Sloveniji, podatki so za leto 1992

Slika 9 kaže, da Kredarica in Dom na Komni najbolj odstopata od ostalih postaj po nadmorski višini, Dom na Komni pa tudi po količini padavin.

Naredimo linearni regresijski model in poglejmo ostanke:

```
> model.0 <- lm(padavine~z.nv, data=postaje)
```



Slika 10: Grafični prikaz ostankov za `model`

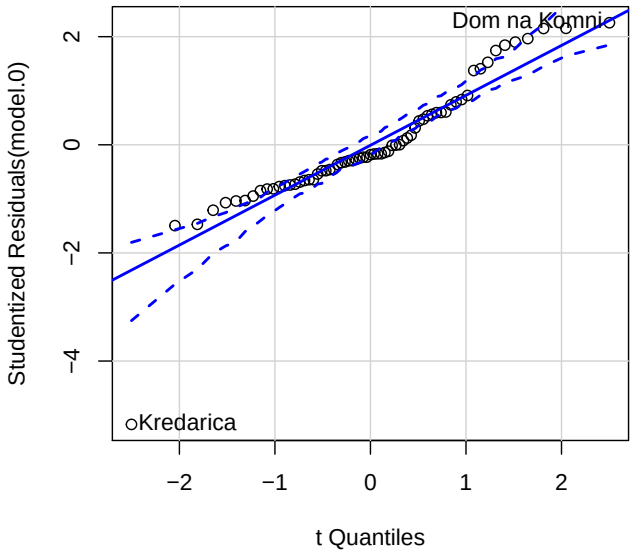
Iz Slike 10 je razvidno, da uporabljeni model ne ustreza podatkom. Poskusimo ugotoviti, kaj povzroča težave. Naredimo analizo posebnih točk.

1.5.2 Regresijski osamelci

Regresijski osamelec je točka, pri kateri vrednost spremenljivke y_i močno odstopa od pripadajoče napovedane vrednosti $\hat{y}_i$. Regresijske osamelce ugotavljamo na osnovi studentiziranih ostankov na dva načina: grafični način in z modelom.

Funkcija `qqPlot` iz paketa `car` nariše studentizirane ostanke glede na kvantile t_{n-k-2} in pripadajočo 95 % točkovno ovojnico. Ovojnica je izračunana s parametričnim bootstrap pristopom (Aitkinson, 1985).

```
> qqPlot(model.0, id=TRUE)

Dom na Komni    Kredarica
          10           23

> # id=list(method="y", n=2, cex=1, col=carPalette()[1], location="lr")
```



Slika 11: QQ grafikon za studentizirane ostanke za `model.0` s 95 % bootstrap ovojnico

Daleč izven ovojnice je Kredarica (Slika 11), kar nakazuje, da je Kredarica regresijski osamelec.

Drugi način za ugotavljanje osamelcev je z modeliranjem. Model za ugotavljanje regresijskih osamelcev (*Mean-shift outlier model*) za i-to točko zapišemo takole:

$$y_i = \alpha + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \gamma d_i + \varepsilon_i, \qquad i = 1, ..., n, \tag{20}$$

kjer je d_i umetna spremenljivka z vrednostjo 1 za i-točko in 0 za ostale točke.

Za vsako točko posebej, $i = 1, ..., n$, preverjamo, ali je regresijski osamelec. Ničelna domneva pravi, da i-ta točka ni regresijski osamelec, $H_{0i} : \gamma = 0$. Alternativna domneva trdi, da i-ta točka je regresijski osamelec, torej $H_{1i} : \gamma \neq 0$, kar pomeni, da se presečišče premakne iz α na $\alpha + \gamma$, ob upoštevanju enake odvisnosti y od $(x_1, ..., x_k)$ kot velja za ostale točke.

Teorija pokaže, da je testna statistika pod ničelno domnevo kar vrednost studentiziranega

ostanka za i-to točko $e_{t_i} = (y_i - \hat{y}_i)/(\hat{\sigma}_{(-i)} \cdot \sqrt{1 - h_{ii}})$, pripadajoča ničelna porazdelitev je Studentova porazdelitev t_{n-k-2}.

Naredimo torej n testov, za vsako točko po enega, za vsakega izračunamo p-vrednost. Ampak ti testi so med seboj odvisni in zato je treba dobljene p-vrednosti prilagoditi. Tu je uporabljen najenostavnejši način prilagoditve p-vrednosti, to je Bonferronijev popravek, ki množi dobljene p-vrednosti s številom testov, torej z n.

Ukaz `outlierTest` iz paketa `car` izpiše vse tiste točke, pri katerih je nepopravljena p-vrednost pod 0.05.

```
> outlierTest(model.0)

           rstudent unadjusted p-value Bonferroni p
Kredarica -5.172079         2.4801e-06   0.00016616
```

Imamo en regresijski osamelec, to je Kredarica, saj je njena popravljena Bonferroni p-vrednost 0.0002. Na Kredarici je vrednost za padavine bistveno nižja, kot bi jo glede na njeno nadmorsko višino pričakovali na osnovi modela.

Ilustracija izračuna Bonferronijevega popravka p-vrednosti za Kredarico:

```
> length(model.0$resid)*outlierTest(model.0)$p  ### to je Bonferronnijev p

   Kredarica
0.0001661642
```

1.5.3 Vzvodne točke

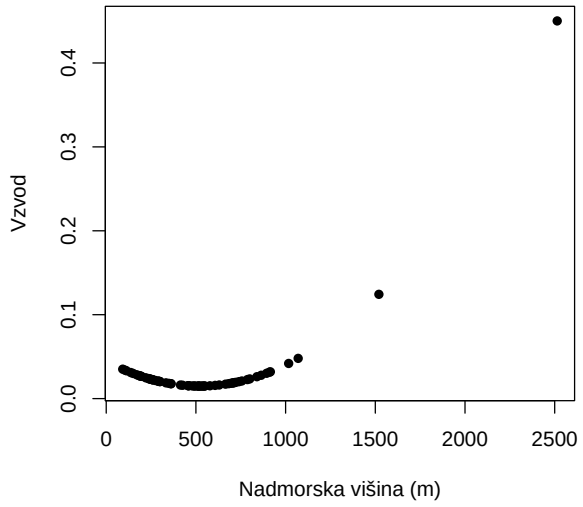
Točke, ki so daleč od centra regresorskega prostora, imajo velik vzvod. Za i-to točko, ki ima vzvod h_{ii} večji od dvakratnika povprečnega vzvoda, pravimo, da je vzvodna točka:

$$h_{ii} > 2\overline{h} = 2 \cdot \frac{k+1}{n}. \tag{21}$$

Glede določitve, kako velik mora biti vzvod, da je točka vzvodna, obstoja tudi bolj ohlapno pravilo: $h_{ii} > 3\overline{h}$.

Vzvode za izbrani model izračunamo z ukazom `hatvalues`. Slika 12 prikazuje kvadratno odvisnost vzvodov (**??**) od nadmorske višine za `model.0`.

```
> plot(postaje$z.nv,hatvalues(model.0), pch=16,
+      xlab=c("Nadmorska višina (m)"), ylab=c("Vzvod"))
```



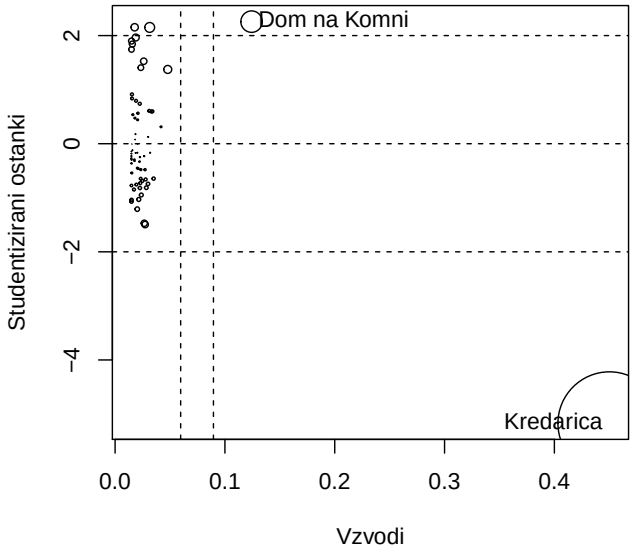
Slika 12: Vzvod v odvisnosti od nadmorske višine za `model.0`

Na Sliki 13 je grafični prikaz studentiziranih ostankov in vzvodov, ki ga dobimo z ukazom `influencePlot` iz paketa `car`. Meji za vzvodne točke sta črtkani navpični črti pri dvakratniku in trikratniku povprečnega vzvoda:

```
> h_povp<-mean(hatvalues(model.0))
> (meja2<-2*h_povp)

[1] 0.05970149

> (meja3<-3*h_povp)

[1] 0.08955224
```

Meji za potencialne regresijske osamelce sta pri vrednostih studentiziranega ostanka -2 in 2, glej vodoravni črti (Slika 13). Za izbrano število identificiranih točk dobimo izpis teh vrednosti.

```
> influencePlot(model.0, id=list(method="noteworthy", n=2, cex=1, location="lr"),
+               xlab="Vzvodi", ylab="Studentizirani ostanki")

               StudRes       Hat     CookD
Dom na Komni  2.257904 0.1242801 0.3403021
Kredarica    -5.172079 0.4501319 7.8423552
```



Slika 13: Grafični prikaz studentiziranih ostankov, vzvodov in Cookove razdalje (ploščina kroga je sorazmerna Cookovi razdalji) za `model.0`

Iz Slike 13 ugotovimo, da sta vzvodni točki Dom na Komni in Kredarica; pri teh dveh postajah nadmorska višina močno odstopa navzgor od povprečne nadmorske višine. Ugotovili pa smo že, da je Kredarica regresijski osamelec.

Vzvodne točke same po sebi niso problem, če pa so hkrati tudi regresijski osamelci, so pogosto tudi vplivne točke, kot bomo videli v nadaljevanju.

1.5.4 Vplivne točke

Izmed posebnih točk so vplivne točke najpomembnejše. Točka $(y_i, x_{i1}, ...x_{ik})$ je vplivna, če velja, da se ocene parametrov modela $\boldsymbol{b}$ ali pa modelske napovedi $\hat{y}_i$, $i = 1, ..., n$ bistveno spremenijo, če jo izločimo iz modela. Vplivna točka lahko vpliva na inferenco modela.

Mer, ki vrednotijo vplivnost posamezne točke, je več. Nekatere izhajajo iz razlike $(\boldsymbol{b}_{(-i)} - \boldsymbol{b})$, kjer je $\boldsymbol{b}_{(-i)}$ cenilka vektorja parametrov v modelu, kjer $i - to$ točko izločimo (Cookova razdalja, DFBETAS). Druge mere vplivnosti i-te točke temeljijo na razlikah napovedi $(\hat{y}_i - \hat{y}_{i(-i)})$, $i = 1, ..., n$, kjer je $\hat{y}_{i(-i)}$ napoved v i-ti točki za model, ki $i - te$ točke pri oceni parametrov ne upošteva (DFFITS).

Cook (1977) je definiral **Cookovo razdaljo** D_i tako, da meri vpliv i-te točke na skupno spremembo ocen parametrov $(\boldsymbol{b}_{(-i)} - \boldsymbol{b})$. Cookova razdalja je definirana

$$D_i = \frac{(\boldsymbol{b}_{(-i)} - \boldsymbol{b})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} (\boldsymbol{b}_{(-i)} - \boldsymbol{b})}{(k+1)\hat{\sigma}^2}. \tag{22}$$

V enačbi (22) je $\hat{\sigma}^2$ cenilka za varianco napak. Ta razdalja je osnovana na podlagi območja zaupanja za vektor parametrov modela $\boldsymbol{\beta}$. Če ima vrednost 0,5, to pomeni, da vektor $\boldsymbol{b}_{(-i)}$ pade izven 50 % območja zaupanja za $\boldsymbol{\beta}$, dobljenega na modelu za vse podatke. Pogosto se kot mejno vrednost za to, da rečemo, da je neka točka vplivna vzame vrednost Cookove razdalje $D_i > 1$.

Pokažemo lahko, da se D_i izrazi s standardiziranim ostankom in vzvodom:

$$D_i = \frac{e_{si}^2}{k+1} \cdot \frac{h_{ii}}{1 - h_{ii}}. \tag{23}$$

Iz zgornjega izraza sledi, da ima točka z veliko vrednostjo standardiziranega ostaneka in velikim vzvodom velik vpliv na modelske napovedi.

Da se pokazati, da Cookovo razdaljo lahko zapišemo še drugače

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(-i)})^2}{(k+1)\hat{\sigma}^2}. \tag{24}$$

Iz zgornjega izraza vidimo, da Cookovo razdaljo lahko izračunamo kod skalirano Evklidsko razdaljo med vektorjem napovedi modela narejenega na vseh podatkih in vektorjem napovedi modela na podatkih, kjer je i-ta točka izločena.
Točke z veliko Cookovo razdaljo najenostavneje identificiramo na četrtem diagnostičnem grafikonu za model (Slika 10 spodaj desno). Na razsevnem grafikonu standardiziraniih ostankov in vzvodov sta prikazani izoliniji za Cookovo razdaljo z vrednostma 0.5 in 1.

Na Sliki 13 je vrednost Cookove razdalje za posamezno točko predstavljena s ploščino kroga. Če uporabimo identifikator točk (`id=list(method="noteworthy", n=a)`), iz pripadajočega

izpisa razberemo a točk z največjo vrednostjo studentiziranega ostanka, a točk z največjim vzvodom in a točk z največjo Cookovo razdaljo (pogosto se točke prekrivajo in jih je v izpisu manj kot $3a$).

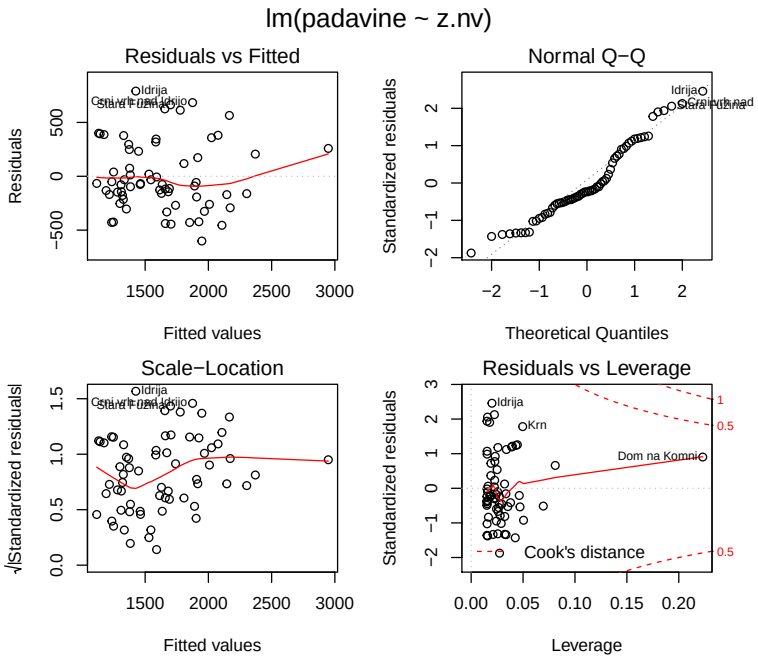
Poglejmo nekaj koristnih izpisov, ki jih dobimo z ukazom `lm.influence`.

```
> vplivne<-lm.influence(model.0)
> names(vplivne)

[1] "hat"          "coefficients" "sigma"        "wt.res"

> # 4 točke z največjimi vzvodi
> sort(vplivne$hat, decreasing=TRUE)[1:4]

   Kredarica Dom na Komni       Vojsko        Mašun
  0.45013195   0.12428014   0.04795426   0.04188555

> # spremembi ocen b_0 in b_1, če Kredarico izločimo
> vplivne$coeff["Kredarica",]

(Intercept)        z.nv
224.1234565  -0.4952036

> # 4 točke z največjo spremembo b_0
> sort(vplivne$coeff[,1],decreasing=TRUE)[1:4]

   Kredarica       Idrija      Podbrdo Stara Fužina
  224.123457    19.125029    10.258344     9.687883

> # 4 točke z največjo spremembo naklona
> sort(vplivne$coeff[,2])[1:4]

  Kredarica      Idrija      Podkum     Ozeljan
-0.49520362 -0.01535168 -0.01123776 -0.01072357

> # izpis standardnih napak regresije brez vključenega podatka
> sort(vplivne$sigma)[1:4] # standardna napaka za model.0 je 384.2 mm

          Kredarica        Dom na Komni Črni vrh nad Idrijo                Krn
           325.1417            372.6180            373.8608            373.8974
```

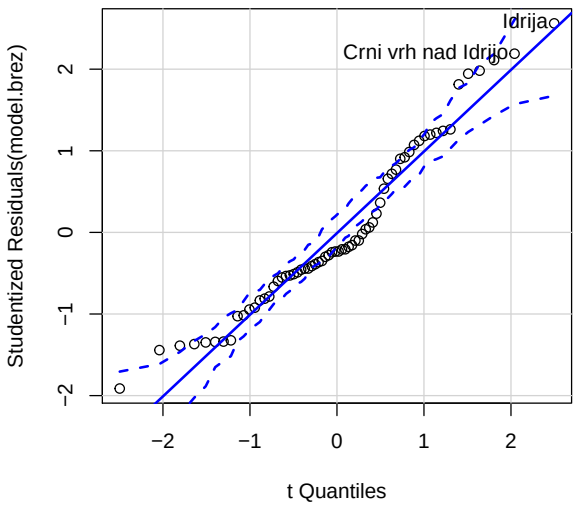
Na osnovi povedanega ugotovimo, da je Kredarica edina vplivna točka; oceni parametrov modela se močno spremenita, če Kredarico izločimo iz podatkov.

Naredimo model znova brez Kredarice in poglejmo ostanke.

```
> postaje.brez<-subset(postaje, subset=postaje$Postaja!="Kredarica")
> model.brez<-lm(padavine~z.nv, data=postaje.brez)
```



Slika 14: Grafični prikaz ostankov za `model.brez`, model brez Kredarice



Slika 15: QQ graf za studentizirane ostanke za `model.brez` s 95 % bootstrap ovojnico

Slike ostankov (Slika 14) kažejo, da `model.brez` nima več vplivnih točk niti regresijskih osamelcev (Slika 15)

```
> summary(model.brez)

Call:
lm(formula = padavine ~ z.nv, data = postaje.brez)

Residuals:
    Min      1Q  Median      3Q     Max
-601.11 -188.58  -76.05  244.03  790.77

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  999.246     81.234  12.301  < 2e-16 ***
z.nv           1.282      0.144   8.902  8.4e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 325.1 on 64 degrees of freedom
Multiple R-squared:  0.5532,        Adjusted R-squared:  0.5462
F-statistic: 79.24 on 1 and 64 DF,  p-value: 8.403e-13

> confint(model.brez)

                  2.5 %      97.5 %
(Intercept) 836.9633919 1161.528407
z.nv          0.9944814    1.570017
```

Z modelom pojasnimo 55.3 % variabilnosti letne količine padavin.

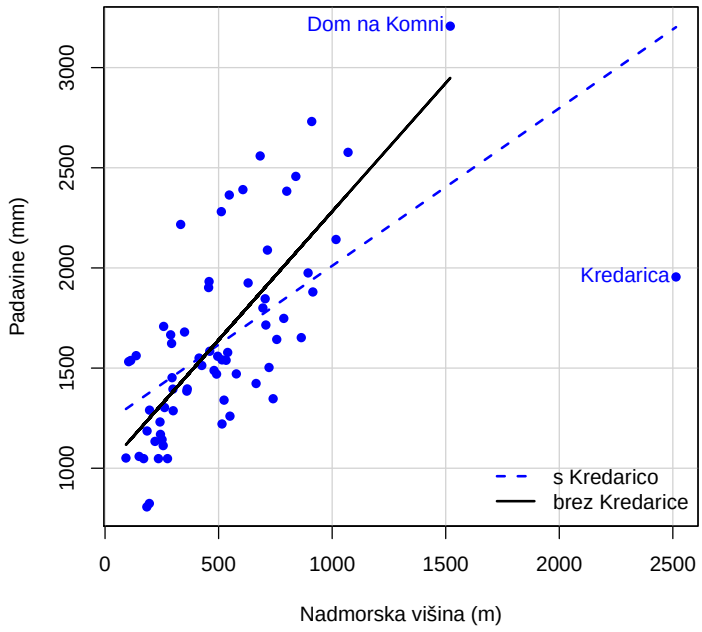
Interpretacija: model ocenjuje, da se letna količina padavin v povprečju poveča za 128.2 mm na vsakih 100 m nadmorske višine, pripadajoči 95 % interval zaupanja je od 99.4 mm do 157.0 mm. Če si dovolimo manjšo ekstrapolacijo, je ocena za letno količino padavin na nadmorski višini 0 m enaka 999 mm (836.9 mm, 1161.5 mm).

Primerjajmo ocene parametrov in njihove standardne napake za `model.0` in `model.brez` z ukazom `compareCoefs` iz paketa `car`:

```
> compareCoefs(model.0, model.brez)

Calls:
1: lm(formula = padavine ~ z.nv, data = postaje)
2: lm(formula = padavine ~ z.nv, data = postaje.brez)

            Model 1 Model 2
(Intercept)  1223.4   999.2
SE             81.2    81.2

z.nv          0.787   1.282
SE            0.127   0.144
```

Vidimo, da sta se obe oceni parametrov bistveno spremenili, pripadajoči standardni napaki pa sta podobni. Obe premici sta predstavljeni na Sliki 16.

```
> scatterplot(padavine~z.nv, regLine=list(lty=2), smooth=FALSE,
+             boxplots=F, xlab=c("Nadmorska višina (m)"), ylab=c("Padavine (mm)"),
+             data=postaje,  pch=16, lwd=2, id=TRUE)
> # dodamo še premico za model.brez
> lines(postaje.brez$z.nv, model.brez$fitted, lwd=2, lty=1)
> legend("bottomright", legend=c("s Kredarico", "brez Kredarice"),
+        bty="n", lty=c(2,1), lwd=2, col=c("blue", "black"))
```



Slika 16: Odvisnost letne količine padavin (mm) od nadmorske višine (m) na podatkih s Kredarico (`model.0`) in na podatkih brez Kredarice (`model.brez`)

Vemo, da se letna količina padavin z nadmorsko višino povečuje, zato imamo tu vse razloge, da dvomimo o pravilnosti podatka za letno količino padavin na Kredarici. Imamo razlago meteorologov, zakaj je bila v letu 1992 tam izmerjena količina padavin prenizka: višje pihajo močnejši vetrovi, merilnik za padavine tisto leto ni bil ustrezno zavarovan pred vetrom, zato je precej padavin odnesel veter.

Za boljše napovedovanje količine padavin manjkajo še druge spremenljivke, npr. geografska dolžina in širina, mikrolokacija, itd.

2 NEKONSTANTNA VARIANCA

Kadar pri linearnem modelu predpostavka o konstantni varianci napak σ^2 ni izpolnjena, govorimo o **nekonstantni varianci** (**heteroskedastičnosti**). Če je varianca σ^2 odvisna od pričakovane vrednosti odzivne spremenljivke $E(y)$, lahko poskusimo z različnimi transformacijami odzivne spremenljivke y. V Tabeli 1 so navedene primerne transformacije pri različnih zvezah med varianco σ^2 in pričakovano vrednostjo $E(y)$. Najbolj uporabni funkciji, ki prideta v poštev pri transformacijah, sta logaritem in kvadratni koren.

Tabela 1: Najpogosteje uporabljene transformacije pri različnih zvezah med varianco σ^2 in pričakovano vrednostjo $E(y)$; znak $\propto$ pomeni sorazmernost

Odnos σ^2 do $E(y)$	Transformacija $T(y)$	Opomba
$\sigma^2 \propto$ *konstanta*	y	ni transformacije
$\sigma^2 \propto$ E(y)	$\sqrt{y}$	y je frekvenca, Poissonova porazdelitev
$\sigma^2 \propto$ E(y)(1-E(y))	arcsin($\sqrt{y}$), *logit*(y)	y je delež, binomska porazdelitev
$\sigma^2 \propto$ E(y)^2	$log(y)$	$y > 0$
$\sigma^2 \propto$ E(y)^4	y^{-1}	$y \neq 0$

Varianca σ^2 je lahko odvisna tudi od ene ali več napovednih spremenljivk, lahko pa od katere druge spremenljivke, ki ni v modelu. V takem primeru lahko pomaga transformacija ustrezne napovedne spremenljivke. Problem nekonstantne variance lahko rešujemo tudi z modeliranjem variance, pri čemer v najpreprostejšem primeru uporabimo tehtano metodo najmanjših kvadratov (WLS, Weighted Least Squares) ali pa kompleksnejšo posplošeno metodo najmanjših kvadratov (GLS, *Generalized Least Squares*). Kadar odzivna spremenljivka ni porazdeljena po normalni porazdelitvi, se težavam z nekonstantno varianco včasih izognemo z uporabo posplošenih linearnih modelov (GLM, *Generalized Linear Model*).

2.1 Box-Cox transformacije

Box in Cox (1964) sta predlagala družino transformacij za odvisno spremenljivko y, ki so v svoji osnovi potenčne transformacije, potenca je označena z λ:

$$T_{BC}(y,\lambda) = y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ ln(y) & , \lambda = 0 \end{cases} . \tag{25}$$

Za $\lambda \neq 0$ gre v bistvu za transfomacije tipa y^λ, saj se od y^λ le odšteje 1 in deli z λ, npr. $\lambda = 0.5$ pomeni korensko transformacijo. Izjema je $\lambda = 0$, ki predstavlja logaritemsko transformacijo.

Box-Cox transformacije so definirane za $y > 0$. Problem nastane pri določenih transformacijah, če so vrednosti za y enake 0 oziroma negativne (npr. `log`, `sqrt`).

Kako ugotoviti, katera vrednost za λ je za podatke ustrezna?

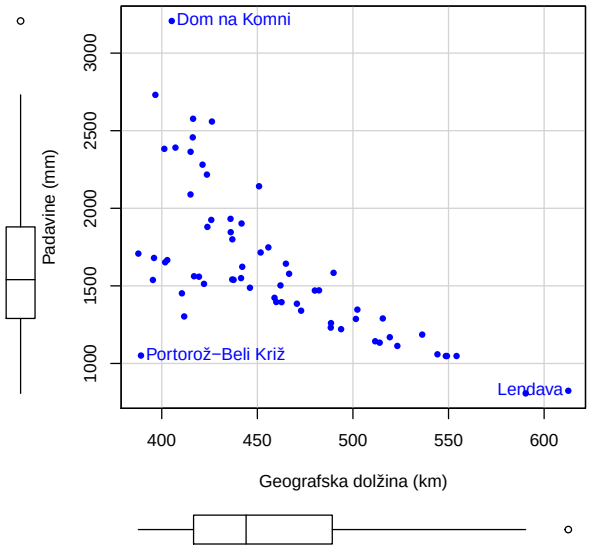
Informativno narišemo porazdelitve transformirane odzivne spremenljivke za smiselno izbrane vrednosti $\lambda = -1, -0.5, 0, 0.5, 1$ (funkcija `symbox` iz paketa `car`). Izberemo λ, pri kateri je porazdelitev najbolj simetrična.

Za analizirani model je najboljša tista transformacija za y, pri kateri je `Residual Sum of Squares`, $SS_{residual}$, najmanjši. V kontekstu posplošenih linearnih modelov (GLM) ima vlogo $SS_{residual}$ funkcija `-log likelihood`. Na podlagi analize odvisnosti logaritma verjetja od vrednosti λ izračunamo optimalno vrednost za λ. Izberemo λ, pri kateri ima logaritem verjetja maksimalno vrednost. Za izračun uporabimo funkcijo `powerTransform` iz paketa `car`, ki vrne optimalni λ in pripadajoči interval zaupanja ter izvede dva informativna testa: $H_0 : \lambda = 0$ (ustrezna je logaritemska transformacija) in $H_0 : \lambda = 1$ (y ni treba transformirati). Grafični prikaz odvisnosti logaritma verjetja od λ dobimo s funkcijo `boxCox` iz paketa `car`.

2.2 Primer: POSTAJE, 2. del

Za meteorološke postaje (datoteka POSTAJE.txt) analizirajmo odvisnost letne količine padavin (`padavine`) od geografske dolžine v Gauss-Krugerjevih koordinatah, ki so izražene v metrih (`x.gdol`).

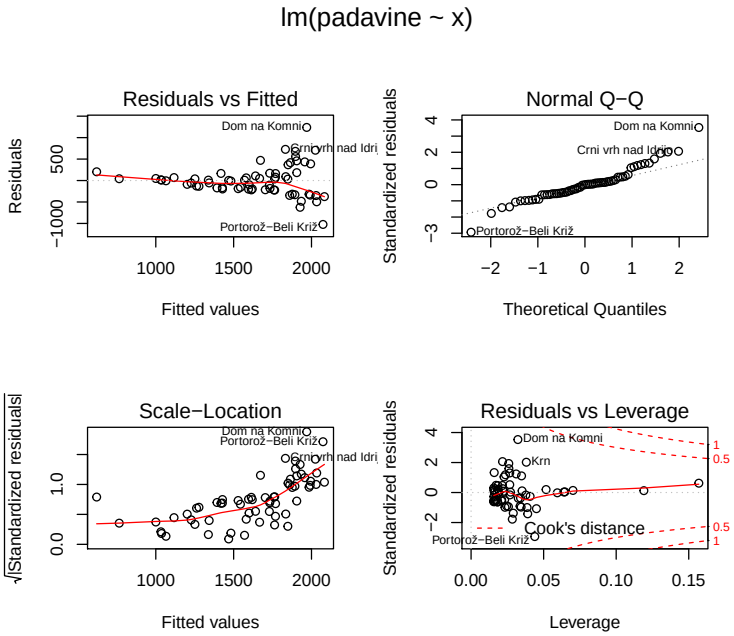
```
> # Kredarico izločimo iz analize (glej primer pri posebnih točkah)
> postaje<-postaje.brez
> # koordinate geografske dolžine izrazimo v km
> postaje$x<-postaje$x.gdol/1000
```



Slika 17: Odvisnost letne količine padavin od geografske dolžine za 64 postaj; podatki so za leto 1992

Naredimo linearni regresijski model in poglejmo ostanke (Slika 18):

```
> model.1 <- lm(padavine~x, data=postaje)
```



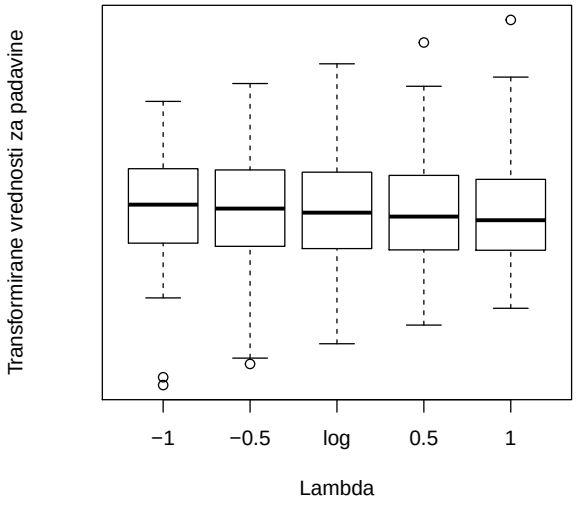
Slika 18: Grafični prikaz ostankov za `model.1`

Levi sličici v prvi in drugi vrstici kažeta nekonstantno varianco. Varianca ostankov narašča z napovedanimi vrednostmi (zgornja leva sličica), slika ostankov je podobna klinu: variabilnost ostankov narašča od leve proti desni. Prisotnost nekonstantne variance še bolje pokaže gladilnik na levi spodnji sliki, kjer so na vodoravni osi napovedane vrednosti, na navpični osi pa koreni absolutnih vrednosti standardiziranih ostankov.

Ocene parametrov so ustrezne, standardne napake pa ne, zato inferenca ni utemeljena.

Slika 19 prikazuje porazdelitve transformiranih vrednosti za `padavine` pri petih različnih vrednostih za λ.

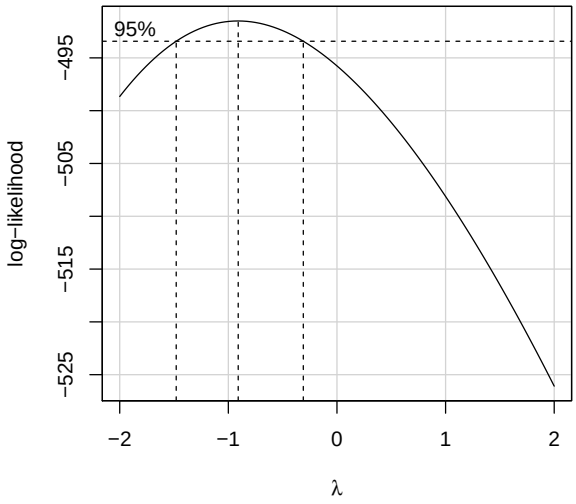
```
> symbox(~padavine, xlab= "Lambda", ylab="Transformirane vrednosti za padavine",
+        data=postaje)
```



Slika 19: Okviri z ročaji za različne transformacije za spremenljivko `padavine`

```
> summary(powerTransform(model.1))

bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.9102          -1      -1.4923      -0.3282

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df      pval
LR test, lambda = (0) 8.509824  1 0.0035323

Likelihood ratio test that no transformation is needed
                           LRT df      pval
LR test, lambda = (1) 33.23255  1 8.177e-09
```

```
> boxCox(model.1)
```



Slika 20: Logaritem verjetja v odvisnosti od λ za `model.1`, optimalna vrednost za λ in njen 95 % interval zaupanja

Rezultati optimizacije funkcije logaritma verjetja kažejo, da za λ izberemo vrednost -1. Dobljena transformacija je neprimerna, saj je spremenljivka 1/`padavine` vsebinsko neobrazložljiva. Za dani primer Box-Cox transformacija ne da ustrezne rešitve. Problem bomo v nadaljevanju rešili z dodatno napovedno spremenljivko `z.nv` in z modeliranjem variance napak.

2.3 Primer: KOVINE

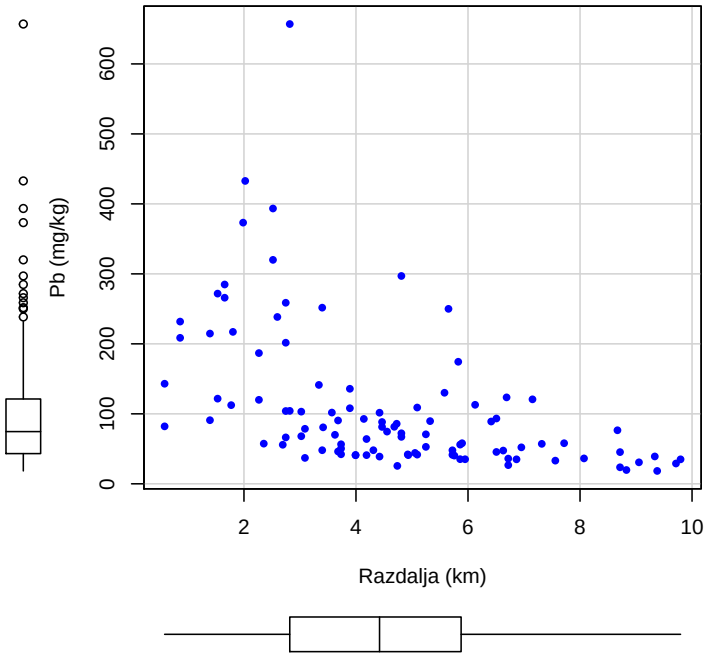
V letu 2000 so raziskovalci ugotavljali vsebnost težkih kovin Cd, Zn, Cu in Pb v tleh na 119 vzorčnih mestih v Celju in okolici; koncentracija je izražena v mg/kg. Za vsako točko je bila ugotovljena tudi razdalja do cinkarne, izražena je v metrih. Ugotoviti želimo, kako se koncentracija Pb spreminja v odvisnosti od razdalje do cinkarne. Razdaljo bomo izrazili v km, upoštevali bomo vzorčne točke z oddaljenostjo do 10 km od cinkarne.

```
> kovine0<-read.table("KOVINE.txt", header=TRUE, sep="\t")
> kovine0$razdalja<-kovine0$razdalja.m/1000
> # izločimo vzorčne točke z oddaljenostjo več kot 10 km
> kovine<-kovine0[kovine0$razdalja<10,]
> dim(kovine)

[1] 103   6
```

```
> summary(kovine[,c("Pb","razdalja")])
```

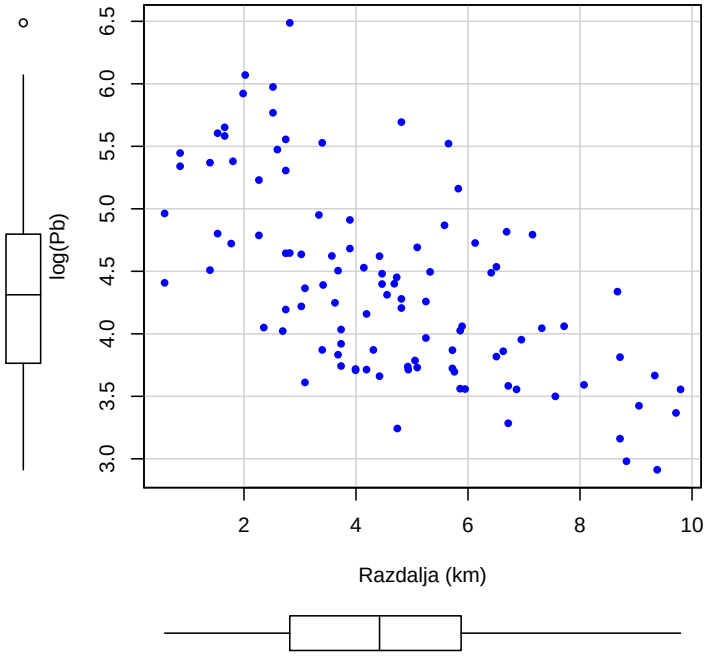
```
       Pb            razdalja
 Min.   : 18.40   Min.   :0.5831
 1st Qu.: 43.15   1st Qu.:2.8178
 Median : 74.60   Median :4.4204
 Mean   :109.80   Mean   :4.5957
 3rd Qu.:121.20   3rd Qu.:5.8770
 Max.   :657.00   Max.   :9.7949
```

Poglejmo najprej grafični prikaz odvisnosti `Pb` od razdalje do cinkarne.



Slika 21: Odvisnost koncentracije Pb od razdalje do cinkarne, podatki Celje 2000

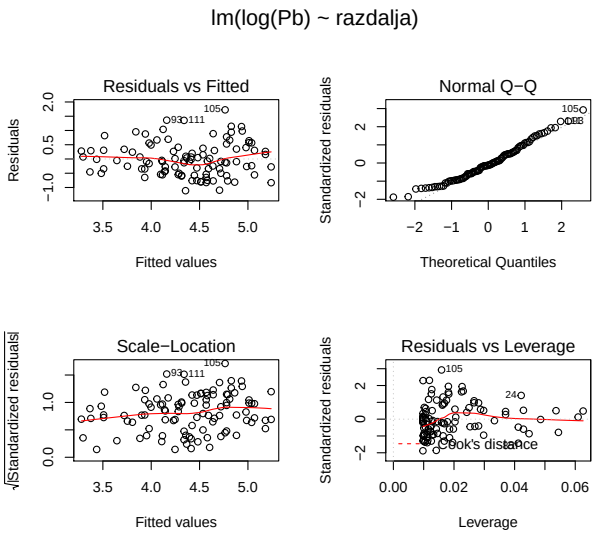
Na Sliki 21 je razvidna velika variabilnost Pb, njegova porazdelitev je asimetrična. Kaže se različna variabilnost za različne oddaljenostih od cinkarne, pri majhnih vrednostih je variabilnost večja kot pri velikih; torej imamo problem nekonstantne variance, tudi predpostavka o linearni odvisnosti je vprašljiva. Poskusimo z logaritemsko transformacijo `Pb`:



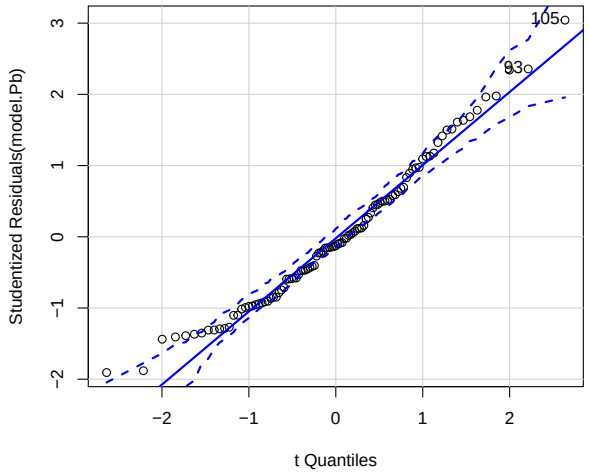
Slika 22: Odvisnost `log(Pb)` od `razdalja` do cinkarne

Slika 22 kaže, da smo z logaritemsko transfomacijo za `Pb` dosegli, da je njegova porazdelitev bistveno bolj simetrična, tudi problem heteroskedastičnosti smo odpravili.

```
> model.Pb <- lm(log(Pb)~razdalja, data=kovine)
```

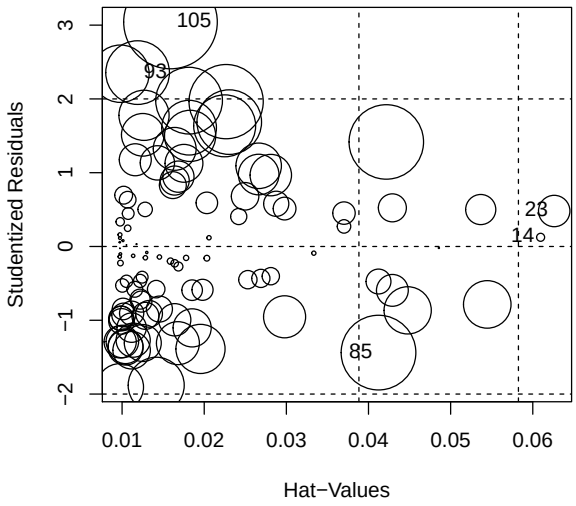


Slika 23: Grafični prikaz ostankov za `model.Pb`



Slika 24: QQ grafikon za studentizirane ostanke za `model.Pb`

```
> influencePlot(model.Pb, id=T)

       StudRes        Hat        CookD
14   0.1245217 0.06095320 0.0005081864
23   0.4807967 0.06260910 0.0077790813
85  -1.4383357 0.04121804 0.0440034051
93   2.3566544 0.01189041 0.0319742915
105  3.0425450 0.01589461 0.0691073177
```



Slika 25: Grafični prikaz studentiziranih ostankov glede na vzvode za `model.Pb`

```
> outlierTest(model.Pb)

No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
105 3.042545          0.0029966      0.30865
```

Ostanki so sprejemljivi. Regresijskih osamelcev in vplivnih točk ni. Če za mejno vrednost vzamemo $3\bar{h}$, imamo dve vzvodni točki, kar pomeni, da imamo dve lokaciji z večjo oddaljenostjo od cinkarne glede na povprečje.

```
> summary(model.Pb)

Call:
lm(formula = log(Pb) ~ razdalja, data = kovine)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11238 -0.47755 -0.07509  0.34428  1.72365

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.36427    0.13419  39.976  < 2e-16 ***
razdalja    -0.21302    0.02628  -8.107 1.26e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.594 on 101 degrees of freedom
Multiple R-squared:  0.3942,        Adjusted R-squared:  0.3882
F-statistic: 65.72 on 1 and 101 DF,  p-value: 1.264e-12

> confint(model.Pb)

                2.5 %    97.5 %
(Intercept)  5.0980795  5.630466
razdalja    -0.2651392 -0.160893
```

S transformacijo odzivne spremenljivke smo naredili t. i. **eksponentni model**, ki je v praksi zelo pogost:

$$y = exp(\beta_0 + \beta_1 x + \varepsilon). \tag{26}$$

Model (26) zlahka lineariziramo: $ln(y) = \beta_0 + \beta_1 x + \varepsilon$. Pomen parametra β_1 ugotovimo z diferenciranjem te enačbe:

$$100\beta_1 = \frac{100\frac{dy}{y}}{dx}. \tag{27}$$

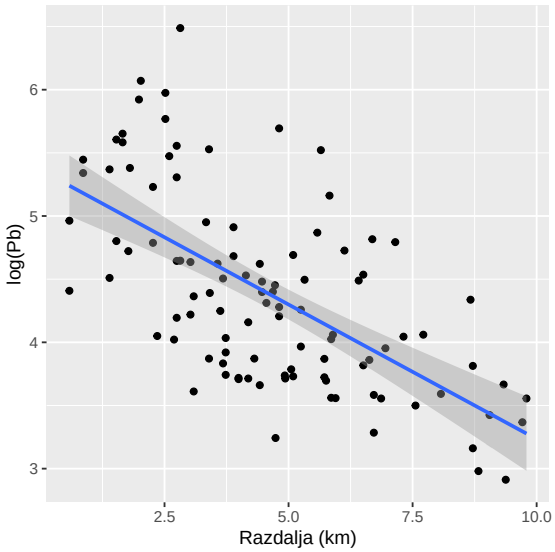
Torej: če se x spremeni za eno enoto, se y spremeni za $100\beta_1$ %.

Interpretacija rezultatov:

- Pri cinkarni (`razdalja` = 0), je `log(Pb)` = 5.364, torej je napovedana vrednost za koncentracijo `Pb` = $exp(5.364)$ =213.636 mg/kg. 95 % IZ za to napoved je (163.7 mg/kg, 278.8 mg/kg).

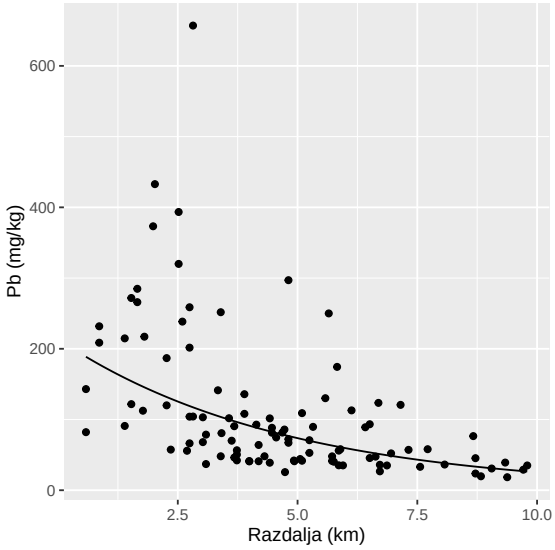
- Če se `razdalja` poveča za 1 km, se koncentracija `Pb` na vsak km v povprečju zmanjša za 21 %. Pripadajoči 95 % interval zaupanja je od 16 % do 27 %.
- Razdalja pojasni cca 39.4 % variabilnosti `log(Pb)`.

```
> library(ggplot2)
> ggplot(data = kovine, aes(x = razdalja, y = log(Pb))) +
+   geom_point() + stat_smooth(method = "lm") +
+   xlab("Razdalja (km)") + ylab("log(Pb)")
```



Slika 26: Odvisnost `log(Pb)` od razdalje do cinkarne in pripadajoča regresijska premica s 95 % intervalom zaupanja za povprečno napoved `log(Pb)`

```
> ggplot(data = kovine, aes(x = razdalja, y = Pb)) +
+   geom_point() + xlab("Razdalja (km)") + ylab("Pb (mg/kg)") +
+   stat_function(fun=function(razdalja) exp(5.36427-0.21302*razdalja) )
```



Slika 27: Odvisnost Pb od razdalje do cinkarne; eksponentni model

3 VAJE

3.1 Koruza

V datoteki KORUZA.txt so rezultati bločnega poskusa s koruzo v letu 1990. Poskus je bil zasnovan v 3 ponovitvah (blokih), v poskusu je bilo 15 različnih gostot setve. Analizirajte, kako gostota setve vpliva na gostoto vznika.

a) Podatke ustrezno grafično prikažite. Sliko na kratko obrazložite.

b) Izberite ustrezen regresijski model za odvisnost gostote setve od gostote vznika.

c) Obrazložite vse korake in končne rezultate modeliranja.

3.2 Sesalci

V datoteki `mammals` v paketu MASS so imena za 62 sesalcev ter podatki o masi telesa in masi možganov zanje. Zanima nas, ali obstaja odvisnost mase možganov `brain` (g) od mase telesa `body` (kg).

```
> library(MASS)
> data(mammals); head(mammals)

                  body brain
Arctic fox       3.385  44.5
Owl monkey       0.480  15.5
Mountain beaver  1.350   8.1
Cow            465.000 423.0
Grey wolf       36.330 119.5
Goat            27.660 115.0
```

a) Grafično prikažite odvisnost `brain` od `body`. Kratko komentirajte sliko.

b) Grafično prikažite porazdelitev spremenljivke `brain`. S katero transformacijo bi dosegli, da bi bila porazdelitev čim bliže normalni porazdelitvi? Zakaj?

c) S katero transformacijo za `body` bi dosegli linearno odvisnost transformirane spremenljivke `brain` od transformirane spremenljivke `body`? Zakaj?

d) Analizirajte ustrezni model in obrazložite rezultate.