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 Vzvodi

Videli smo že, da je vzvod i-te točke, h_{ii}, diagonalni element matrike $\mathbf{H}$, ki povezuje $\mathbf{y}$ in $\hat{\mathbf{y}}$, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}.$$

Vzvod je odvisen le 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.

V enostavni linearni regresiji se vzvod izračuna takole:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}. \tag{1}$$

Za splošni regresijski model velja, da je $\sum_i^n h_{ii} = k + 1$, kjer je $k + 1$ število parametrov v modelu. Posledično je povprečni vzvod

$$\overline{h} = \frac{k+1}{n}. \tag{2}$$

1.2 Ostanki

Predpostavka normalnega linearnega modela je, da so napake porazdeljene normalno: $\varepsilon \sim iid\ N(0, \sigma^2)$. Za ostanke $e_i = y_i - \hat{y}_i$, $i = 1, ..., n$, velja, da so porazdeljeni $N(0, \sigma^2(1 - h_{ii}))$. To pomeni, da njihova varianca ni konstantna, temveč je odvisna od vzvoda, ta pa je, kot smo videli, odvisen od položaja točke v regresorskem prostoru. Ostanki e_i so nepovezani z z modelskimi napovedmi $\hat{y}_i$, $Cov(e_i, \hat{y}_i) = 0$, zato je razsevni grafikon ostankov glede na napovedane vrednosti dobro diagnostično orodje za regresijski model.

1.2.1 Standardizirani ostanki

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

$$e_{s_i} = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, ..., n. \tag{3}$$

Če je model sprejemljiv, imajo standardizirani ostanki konstantno varianco. Števec in imenovalec pri (3) sta povezana, saj je v oceni s upoštevana tudi vrednost števca $y_i - \hat{y}_i$. Zato je porazdelitev standardiziranih ostankov le približno Studentova s $SP = n-k-1$; če pa je $n >> k$, je porazdelitev približno $N(0, 1)$.

Standardizirani ostanki so povezani z modelskimi napovedmi $\hat{y}_i$, $Cov(e_{s_i}, \hat{y}_i) \neq 0$. Točke, ki imajo po absolutni vrednosti standardizirani ostanek več kot 2, $|e_{s_i}| > 2$, so kandidati za regresijske osamelce.

1.2.2 Studentizirani ostanki

Da se znebimo povezanosti med števcem in imenovalcem v (3), izračunamo studentizirane ostanke. 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:

$$e_{t_i} = \frac{y_i - \hat{y}_i}{s_{(-i)} \cdot \sqrt{1 - h_{ii}}}, \tag{4}$$

$s_{(-i)}$ je standardna napaka regresije, 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 po t-porazdelitvi s $SP = n - k - 2$.

1.3 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 predstavili na primeru.

1.3.1 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")
> str(postaje)

'data.frame':        67 obs. of  5 variables:
 $ Postaja : Factor w/ 67 levels "Babno polje",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ x.gdol  : int  464930 554193 451721 459888 415040 519274 407149 426241 515610 405223 ...
 $ y.gsir  : int  56264 97520 105620 119639 112316 122872 101315 87226 46928 126911 ...
 $ z.nv    : int  756 170 708 362 715 244 607 683 196 1520 ...
 $ padavine: int  1643 1048 1715 1396 2089 1169 2391 2559 1290 3207 ...

> 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)]
```

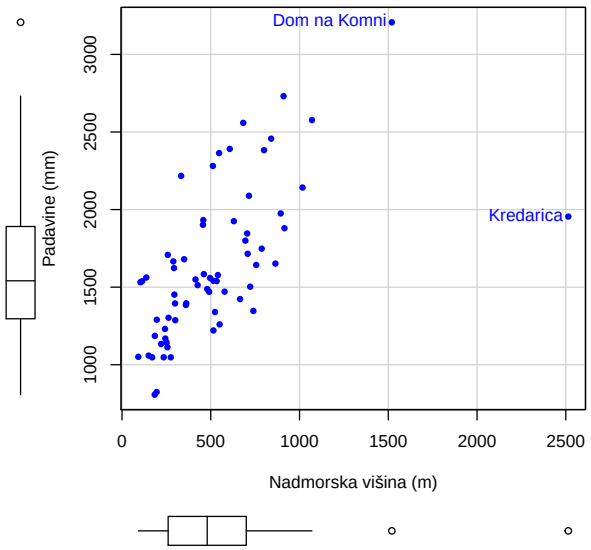
```
[1] "Jezersko" "Ozeljan"
```

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

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

Slika 1 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
```

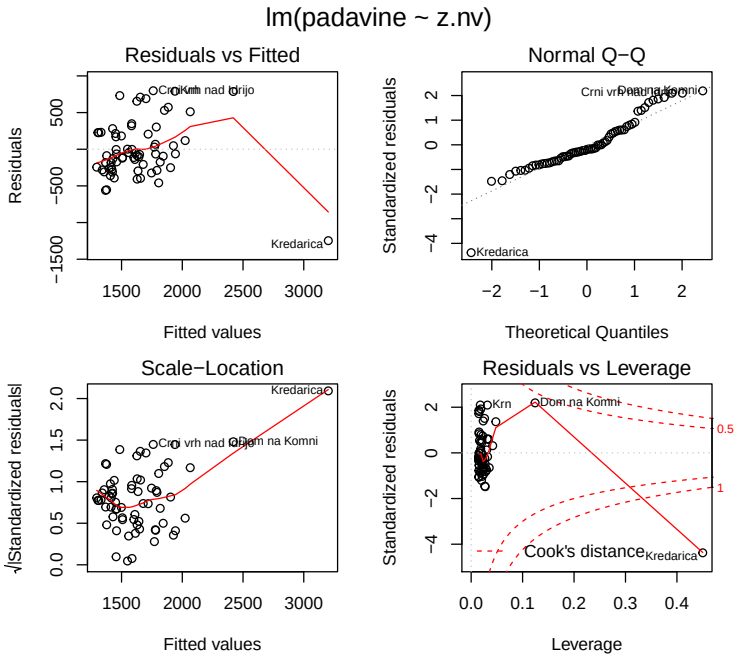


Slika 1: Odvisnost letne količine padavin od nadmorske višine za 67 postaj; podatki so za leto 1992

Slika 1 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 2: Grafični prikaz ostankov za `model`

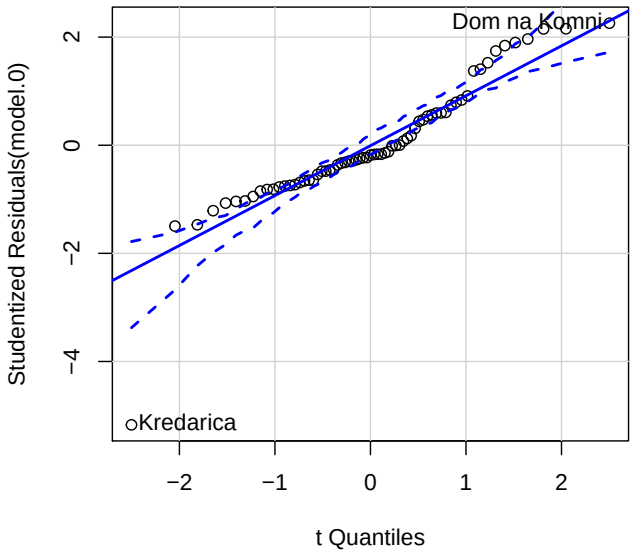
Iz Slike 2 je razvidno, da uporabljeni model ne ustreza podatkom. Poskusimo ugotoviti, kaj povzroča težave. Naredimo analizo posebnih točk.

1.3.2 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-porazdelitve s $SP = n - k - 2$ 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 3: QQ grafikon za studentizirane ostanke za `model.0` s 95 % bootstrap ovojnico

Daleč izven ovojnice je Kredarica (Slika 3), 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:

$$y_i = \alpha + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \gamma d_i + \varepsilon_i, \qquad i = 1, ..., n, \tag{5}$$

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 $e_{t_i} = (y_i - \hat{y}_i)/(s_{(-i)} \cdot \sqrt{1 - h_{ii}})$, 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
```

1.3.3 Vzvodne točke

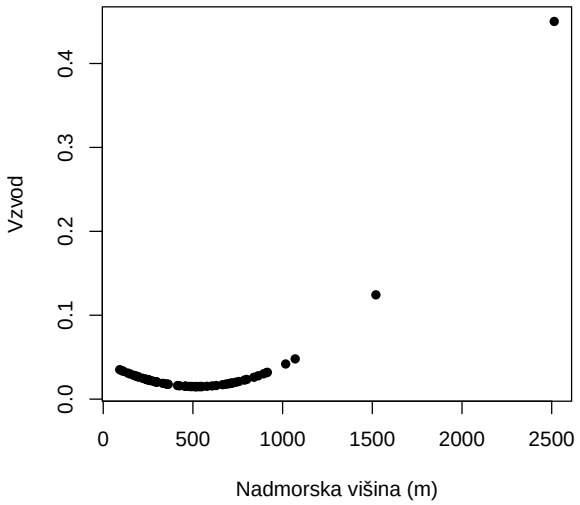
Točke, ki so daleč od centra regresorskega prostora, imajo velik vzvod. Za i-to točko, ki ima vzvod h_{ii} večji od dvakratnika povprečnega vzvoda, pravimo, da je vzvodna točka:

$$h_{ii} > 2\overline{h} = 2 \cdot \frac{k+1}{n}. \tag{6}$$

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 4 prikazuje kvadratno odvisnost vzvodov (1) od nadmorske višine za `model.0`.

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



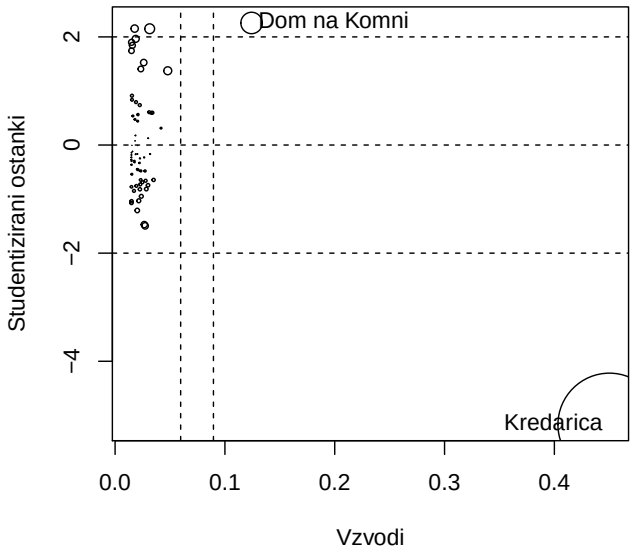
Slika 4: Vzvod v odvisnosti od nadmorske višine za `model.0`

Na Sliki 5 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 5). 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 5: Grafični prikaz studentiziranih ostankov, vzvodov in Cookove razdalje (ploščina kroga je sorazmerna Cookovi razdalji) za `model.0`

Iz Slike 5 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.

1.3.4 Vplivne točke

Izmed posebnih točk so najpomembnejše vplivne točke. Točka $(y_i, x_{i1}, ...x_{ik})$ je vplivna, če velja, da se ocene parametrov modela $\boldsymbol{b}$ bistveno spremenijo, če jo izločimo iz modela; v tem primeru označimo ocene $\boldsymbol{b}_{(-i)}$. Mera vplivnosti i-te točke je sorazmerna izrazu $(\boldsymbol{b}_{(-i)} - \boldsymbol{b})$; večja razlika pomeni, da je točka bolj vplivna.

Mer, ki vrednotijo vplivnost posamezne točke, je več. Nekatere izhajajo iz razlike $(\boldsymbol{b}_{(-i)} - \boldsymbol{b})$, taka je DFBETAS. Druge mere vplivnosti i-te točke temeljijo na razlikah napovedi $(\hat{y}_j - \hat{y}_{j(-i)})$, $j = 1, ..., n$, $\hat{y}_j$ je napoved osnovnega modela v j-ti točki in $\hat{y}_{j(-i)}$ je napoved v j-točki za model, ki $i - te$ točke pri oceni parametrov ne upošteva. Taki meri sta DFFITS in Cookova razdalja.

Cook (1977) je definiral Cookovo razdaljo D_i tako, da meri vpliv i-te točke na vse napovedane vrednosti. Razlika med napovedano vrednostjo $\hat{y}_j$ in pripadajočo napovedano vrednostjo $\hat{y}_{j(-i)}$, ki jo dobimo, če se i-ta točka izloči, se kvadrira in sešteje po vseh točkah, $j = 1, ..., n$. Cookova razdalja se izračuna takole:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(-i)})^2}{k+1} \cdot \frac{1}{s^2}. \tag{7}$$

V enačbi (7) je s^2 ocena za varianco napak.

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

$$D_i = \frac{e_{si}^2}{k+1} \cdot \frac{h_{ii}}{1 - h_{ii}}. \tag{8}$$

Torej: točka, ki ima hkrati velik standardizirani ostanek in velik vzvod, izraža potencialno velik vpliv na modelske napovedi.

Cookova ohlapna definicija: točka je vplivna točka, če je $D_i > 1$.

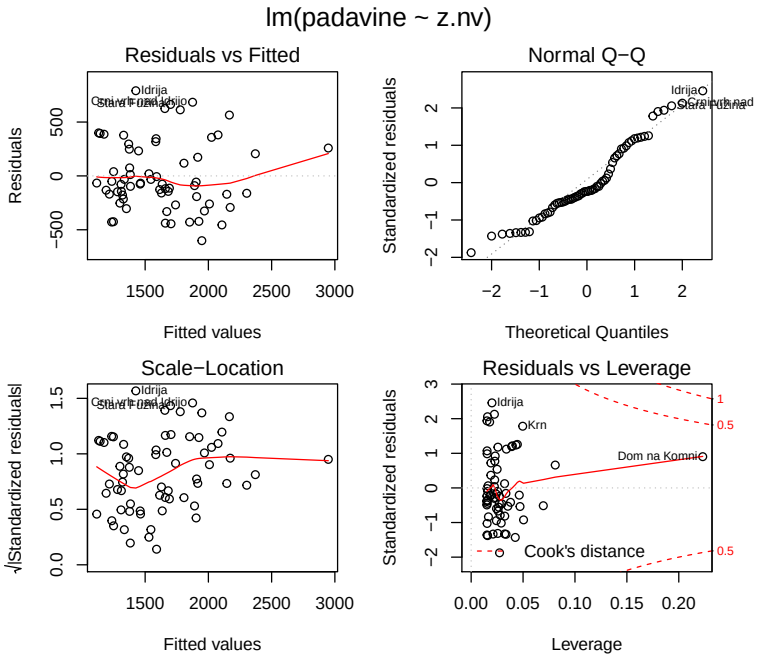
Kako identificiramo vplivne točke na sliki? Na Sliki 2 spodaj desno vidimo izoliniji za Cookovo razdaljo z vrednostma 0.5 in 1.

Na Sliki 5 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 $3a$).

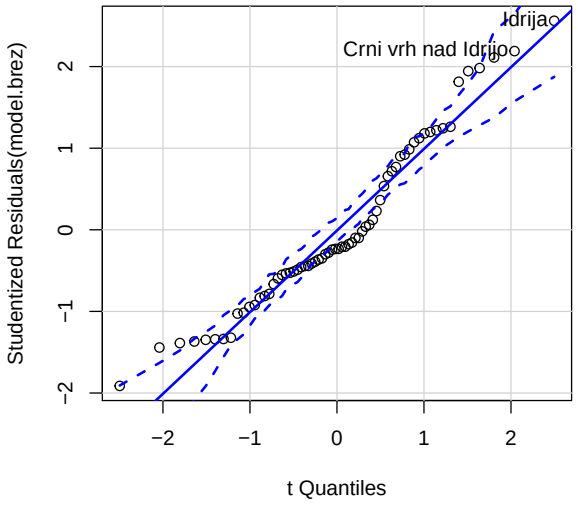
Na osnovi povedanega ugotovimo, da je Kredarica edina vplivna točka; oceni parametrov modela se močno spremenita, če Kredarico izločimo iz podatkov.

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 6: Grafični prikaz ostankov za `model.brez`, model brez Kredarice



Slika 7: QQ graf za studentizirane ostanke za `model.brez` s 95 % bootstrap ovojnico

Slike ostankov kažejo, da `model.brez` nima več vplivnih točk.

```
> 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.

Če v modelu hkratno preverjamo dve ali več ničelnih domnev, se moramo zavedati, da so rezultati testiranja medsebojno odvisni, zato so p-vrednosti v povzetku modela praviloma premajhne, intervali zaupanja za parametre modela pa posledično preozki. Zato moramo upoštevati, da so pri hkratnem testiranju več domnev p-vrednosti v povzetku modela le informativne. Pri ocenjevanju samo dveh parametrov ta problem v splošnem ni prisoten.

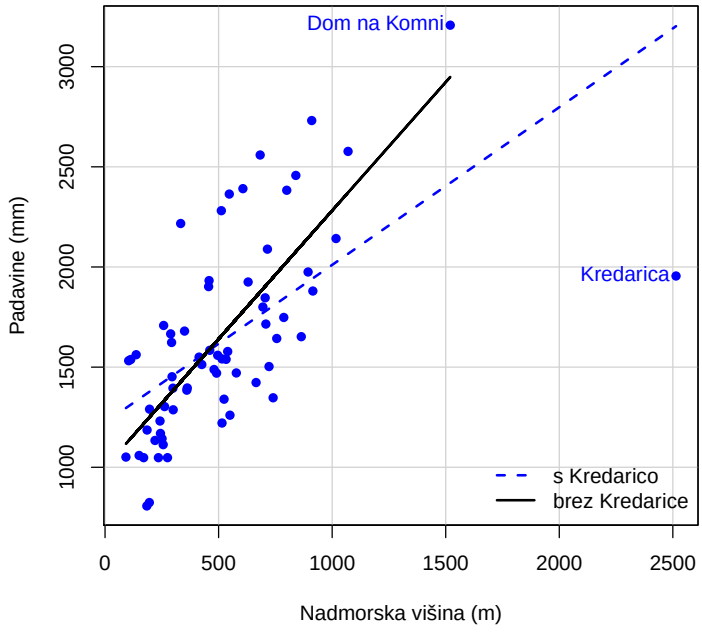
Interpretacija: model ocenjuje, da se letna količina padavin v povprečju poveča za 128.2 mm na vsakih 100 m nadmorske višine, pripadajoči 95 % interval zaupanja je od 99.4 mm do 157.0 mm. Če si dovolimo manjšo ekstrapolacijo, je ocena za letno količino padavin na nadmorski višini 0 m enaka 999 mm (836.9 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 8.

```
> 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 8: 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:

$$T_{BC}(y,\lambda) = y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ ln(y) & , \lambda = 0 \end{cases} . \tag{9}$$

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`).

Če so vrednosti za y tudi negativne, se uporabi družina Yeo-Johnson transfomacij (2000) $T_{YJ}(y,\lambda)$, ki rešujejo ta problem preko $T_{BC}(y,\lambda)$ takole:

$$T_{YJ}(y,\lambda) = \begin{cases} T_{BC}(y+1,\lambda) & , y \geq 0 \\ T_{BC}(-y+1, 2-\lambda) & , y < 0 \end{cases} . \tag{10}$$

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

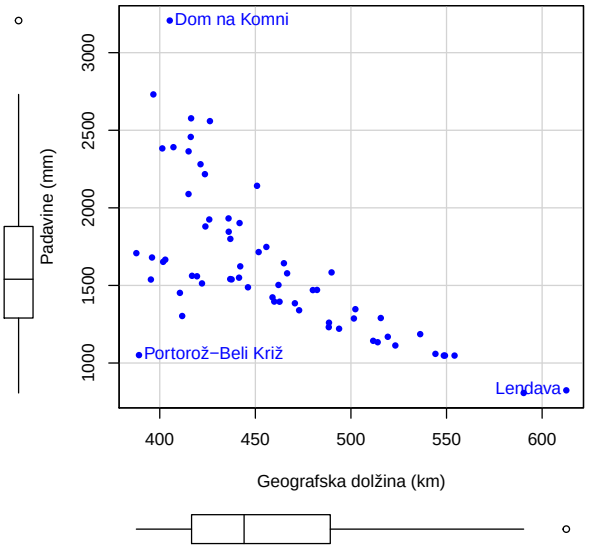
Informativno narišemo porazdelitve transformirane odzivne spremenljivke za smiselno izbrane vrednosti $\lambda = -1, -0.5, 0, 0.5, 1$ (funkcija `symbox` iz paketa `car`). Izberemo λ, pri kateri je porazdelitev najbolj simetrična.

Za analizirani model je najboljša je 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`).

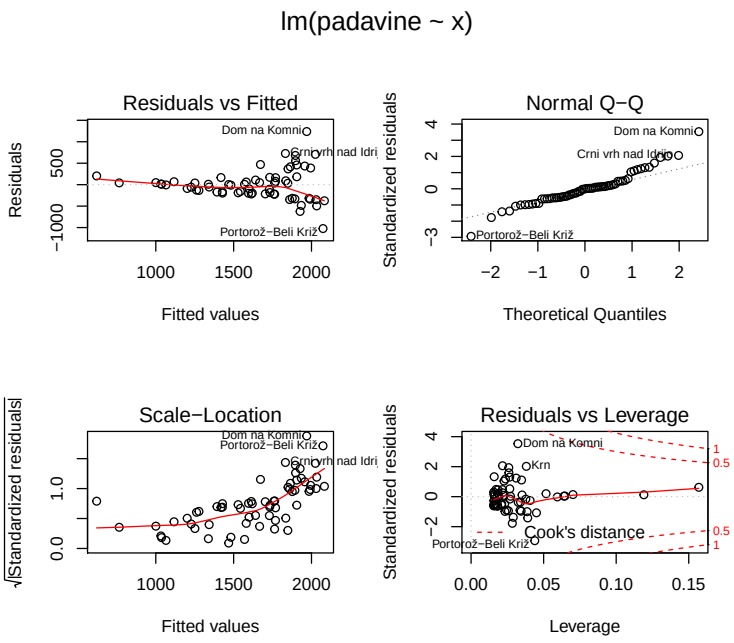
```
> # Kredarico izločimo iz analize (glej primer pri posebnih točkah)
> postaje<-postaje.brez
> # koordinate geografske dolžine izrazimo v km
> postaje$x<-postaje$x.gdol/1000
```



Slika 9: Odvisnost letne količine padavin od geografske dolžine za 64 postaj; podatki so za leto 1992

Naredimo linearni regresijski model in poglejmo ostanke (Slika 10):

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



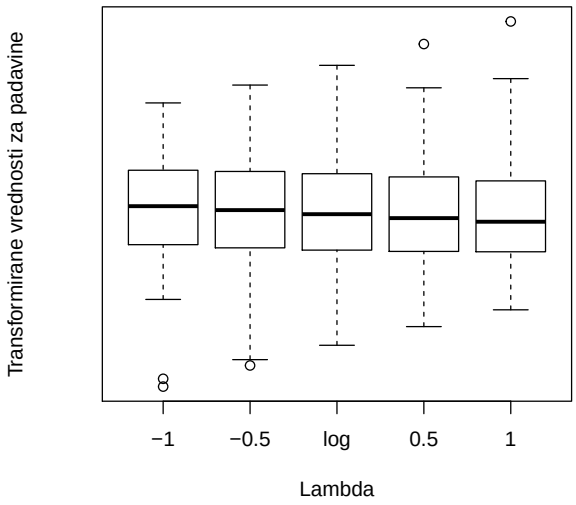
Slika 10: Grafični prikaz ostankov za `model.1`

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

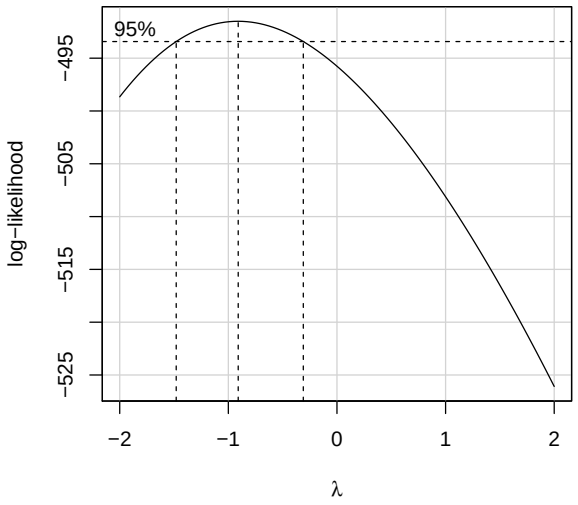
```
> symbox(~padavine, xlab= "Lambda", ylab="Transformirane vrednosti za padavine",
+        data=postaje)
```



Slika 11: 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 12: 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

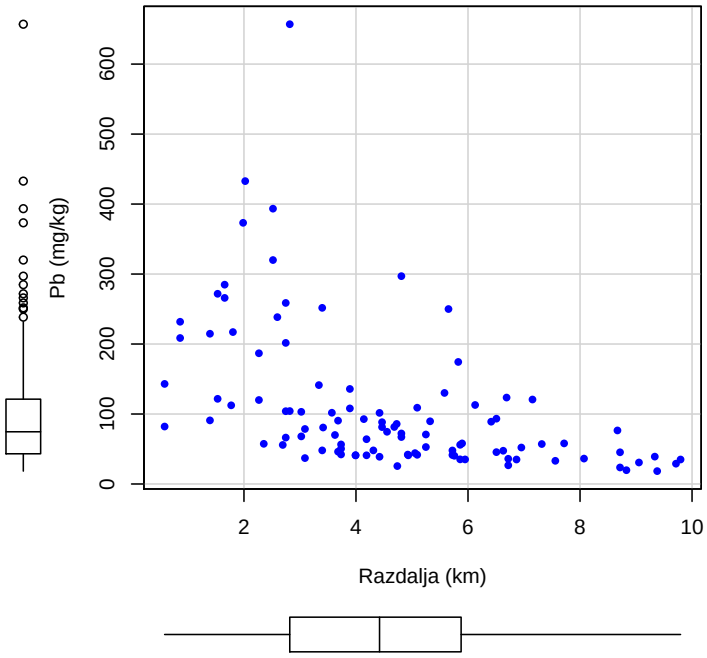
V letu 2000 so raziskovalci ugotavljali vsebnost težkih kovin Cd, Zn, Cu in Pb v tleh na 119 vzorčnih mestih v Celju in okolici; koncentracija je izražena v mg/kg. Za vsako točko je bila ugotovljena tudi razdalja do cinkarne, izražena je v metrih. Ugotoviti želimo, kako se koncentracija Pb spreminja 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")])
```

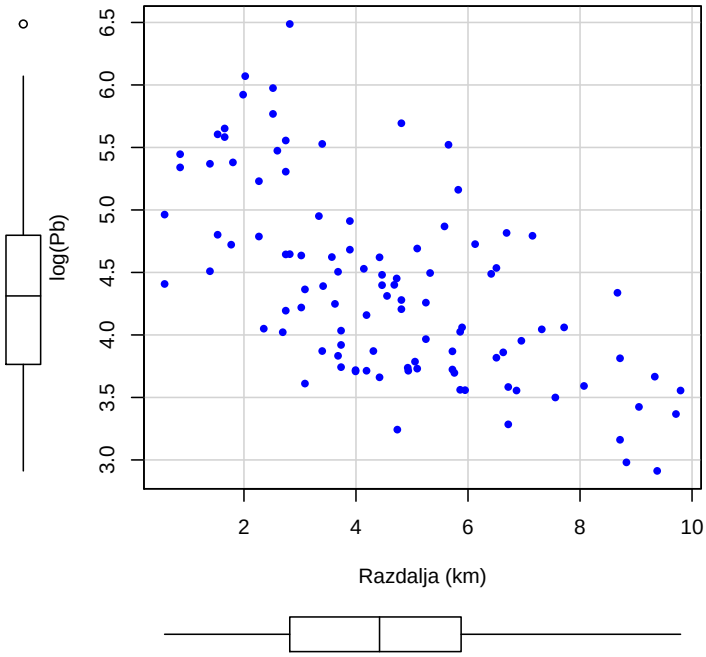
```
       Pb            razdalja
 Min.   : 18.40   Min.   :0.5831
 1st Qu.: 43.15   1st Qu.:2.8178
 Median : 74.60   Median :4.4204
 Mean   :109.80   Mean   :4.5957
 3rd Qu.:121.20   3rd Qu.:5.8770
 Max.   :657.00   Max.   :9.7949
```

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



Slika 13: Odvisnost koncentracije Pb od razdalje do cinkarne, podatki Celje 2000

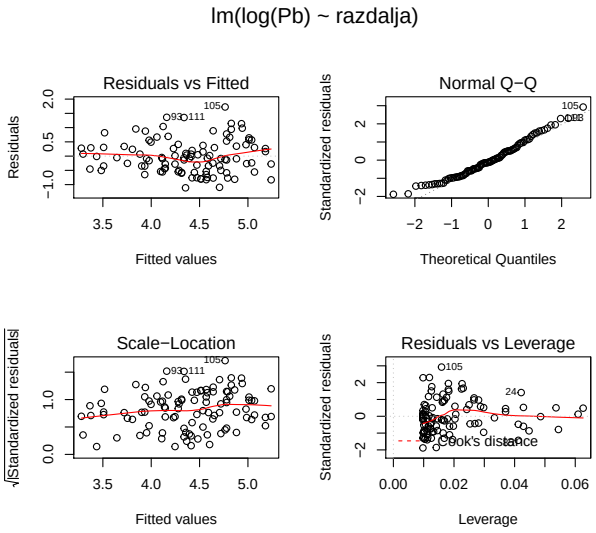
Na Sliki 13 je razvidna velika variabilnost Pb, njegova porazdelitev je asimetrična. Kaže se različna variabilnost za različne oddaljenostih od cinkarne, pri majhnih vrednostih je variabilnost večja kot pri velikih; torej imamo problem nekonstantne variance, tudi predpostavka o linearni odvisnosti je vprašljiva. Poskusimo z logaritemsko transformacijo `Pb`:



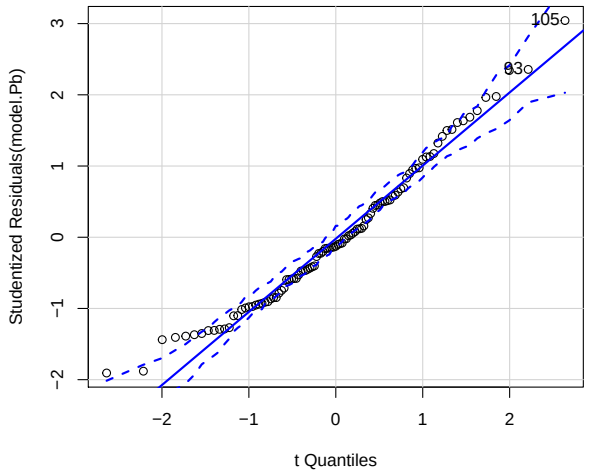
Slika 14: Odvisnost `log(Pb)` od `razdalja` do cinkarne

Slika 14 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)
```

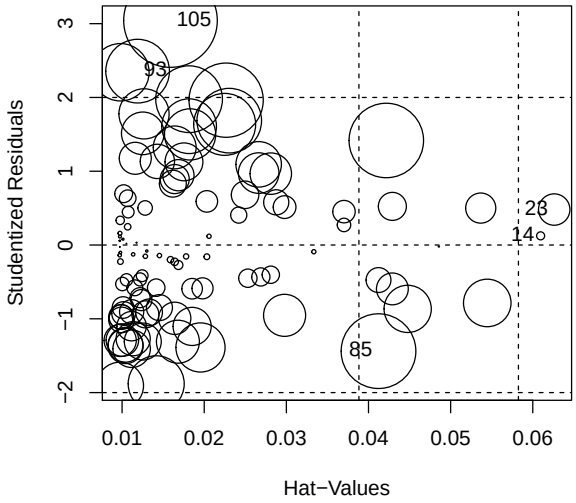


Slika 15: Grafični prikaz ostankov za `model.Pb`



Slika 16: 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 17: 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:

$$y = exp(\beta_0 + \beta_1 x + \varepsilon). \tag{11}$$

Model (11) zlahka lineariziramo: $ln(y) = \beta_0 + \beta_1 x + \varepsilon$. Pomen parametra β_1 ugotovimo z diferenciranjem te enačbe:

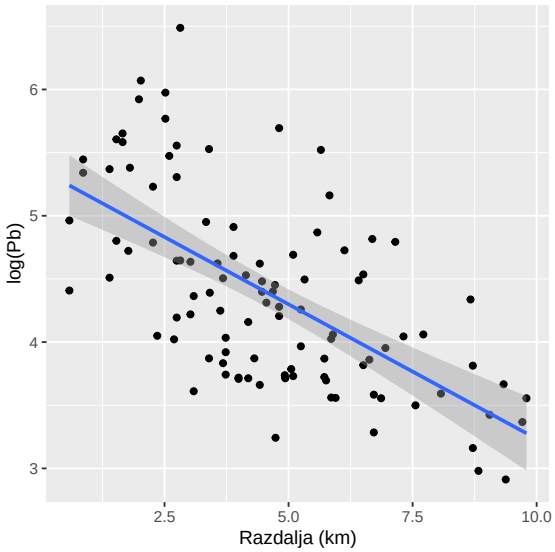
$$100\beta_1 = \frac{100\frac{dy}{y}}{dx}. \tag{12}$$

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

Interpretacija rezultatov:

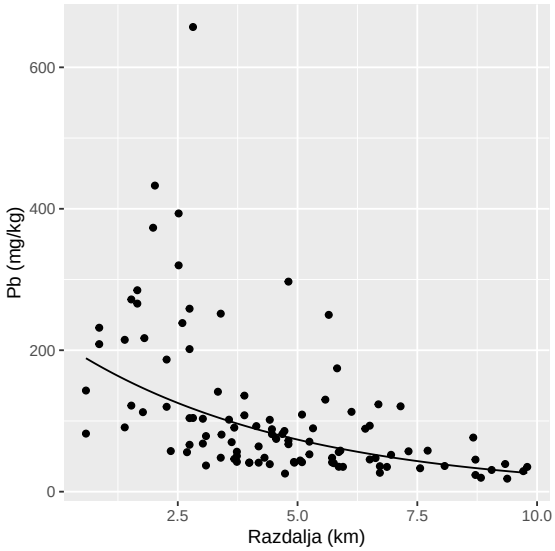
- Pri cinkarni (`razdalja` = 0), je `log(Pb)` = 5.364, torej je napovedana vrednost za koncentracijo `Pb` = $exp(5.364)$ =213.636 mg/kg. 95 % IZ za to napoved je (163.7 mg/kg, 278.8 mg/kg).
- Če se `razdalja` poveča za 1 km, se koncentracija `Pb` na vsak km v povprečju zmanjša za 21 %. Pripadajoči 95 % 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 18: 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 19: Odvisnost Pb od razdalje do cinkarne; eksponentni model

2.4 Transformacije za delež

V praksi je pogosto y spremenljivka z omejeno zalogo vrednosti, najbolj pogosto je to inteval [0,1], pri čemer y predstavlja delež. V ozadju je slučajna spremenljivka, ki je porazdeljena po binomski porazdelitvi $b(n, \pi)$. Varianca deležev blizu 0 oz. blizu 1 je manjša od variance deležev blizu 1/2. Obstojata dve transformaciji, s katerima se lahko odpravi ta nadloga, to sta: $asin(\sqrt{y})$ in $logit(y)$.

Pri analizi deležev problemi z nekonstantno varianco ne nastopijo, če so vrednosti deležev približno na intervalu [0.25, 0.75], tedaj transformacija ni potrebna.

Za transformacijo $asin(\sqrt{y})$ velja, da je standardni odklon transformiranih vrednosti približno enak $0.5/\sqrt{n}$ in je torej neodvisen od parametra binomske porazdelitve π. Ta transformacija stabilizira varianco in odpravi heteroskedastičnost.

Poglejmo vpliv tranformacije `asin`, pri tej tranfomaciji vrednost 0 ne povzroča nobenih težav:

```
> y<-seq(from=0, to=1, by=0.1)
> asin<-asin(sqrt(y))
> print(data.frame(y,asin))

     y      asin
1  0.0 0.0000000
2  0.1 0.3217506
3  0.2 0.4636476
4  0.3 0.5796397
5  0.4 0.6847192
6  0.5 0.7853982
7  0.6 0.8860771
8  0.7 0.9911566
9  0.8 1.1071487
10 0.9 1.2490458
11 1.0 1.5707963
```

Alternativna transformacija je `logit` (`log-odds`):

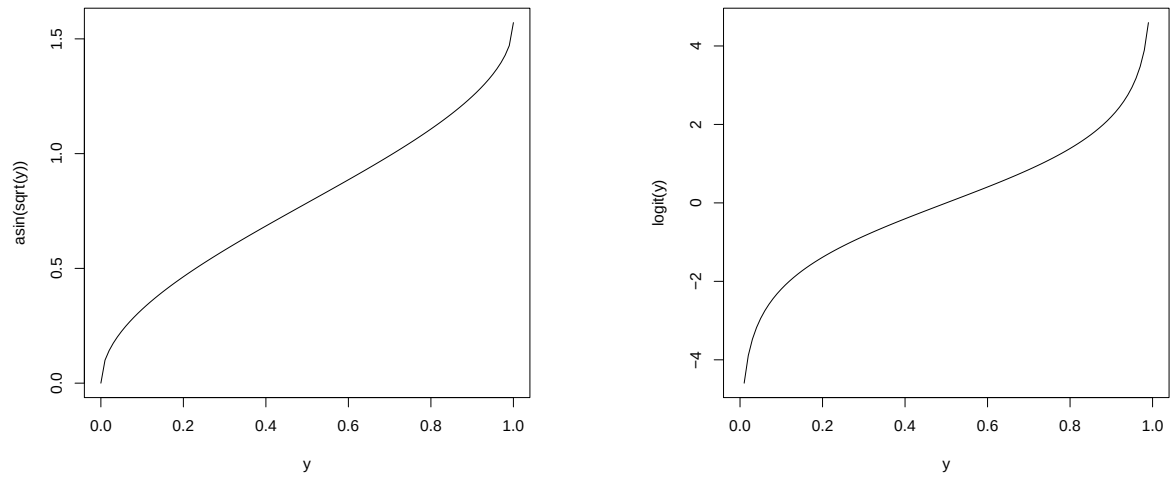
$$logit(y) = ln\frac{y}{1-y}. \tag{13}$$

Ta transformacija je osnova logistični regresiji, ki sodi v okvir posplošenih linearnih modelov. Poudariti velja, da ni definirana za $y = 0$ in za $y = 1$.

```
> y<-seq(from=0, to=1, by=0.1)
> logit<-log(y/(1-y))
> print(data.frame(y,logit))
```

```
     y      logit
1  0.0       -Inf
2  0.1 -2.1972246
3  0.2 -1.3862944
4  0.3 -0.8472979
5  0.4 -0.4054651
6  0.5  0.0000000
7  0.6  0.4054651
8  0.7  0.8472979
9  0.8  1.3862944
10 0.9  2.1972246
11 1.0        Inf
```

Grafični prikaz obeh transformacij je na Sliki 20. Oba grafa sta sigmoidne oblike (S-oblike), zalogi vrednosti transformirane spremenljivke pa sta različni.



(a) `asin` transformacija

(b) `logit` transfomacija

Slika 20: Transformaciji, ki ju uporabljamo, če je odvisna spremenljivka delež; skali na ordinatah sta različni

2.5 Primer: PELOD

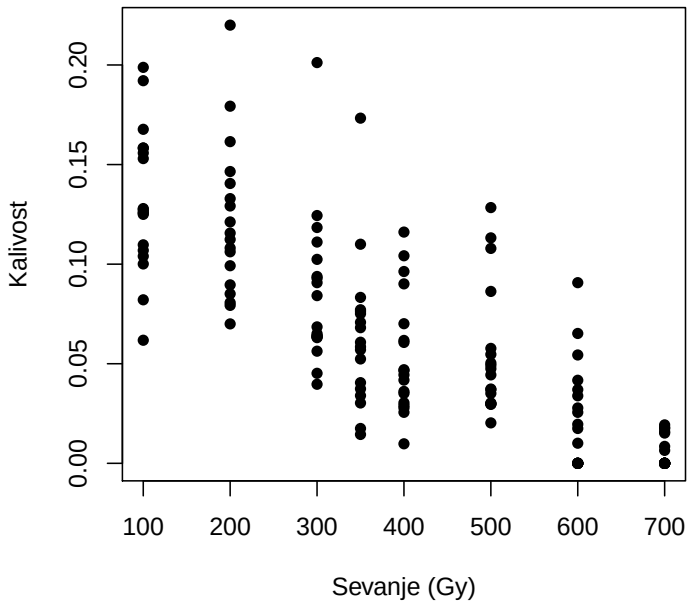
V poskusu so obsevali pelod buč z osmimi različnimi odmerki rentgenskega sevanja (`Sevanje`: 100, 200, 300, 350, 400, 500, 600, 700 Gy, *gray* je enota za absorbirano sevanje). Obsevanje je bilo izvedeno pri dveh različnih zračnih vlagah (`Vlaga`: Room humidity, RH, in High Humidity, HH). Za vsako kombinacijo vlage in odmerka sevanja je bilo 9 kapljic, ki so vsebovale pelod buč; skupaj je bilo v poskusu 144 kapljic. Ugotavljali so kalivost peloda (`Kalivost`), ki je izražena kot delež kalenega peloda v kapljici (to se ugotavlja z mikroskopom). Podatki so v datoteki PELOD.txt.

Kako odmerek sevanja (`Sevanje`) vpliva na kalivost peloda (`Kalivost`)? Zaenkrat napovedne spremenljivke `Vlaga` ne vključimo v model.

```
> pelod<-read.table("PELOD.txt",   header=TRUE)
> str(pelod)

'data.frame':        144 obs. of  3 variables:
 $ Vlaga   : Factor w/ 2 levels "HH","RH": 1 1 1 1 1 1 1 1 1 1 ...
 $ Sevanje : int  100 100 100 100 100 100 100 100 100 200 ...
 $ Kalivost: num  0.192 0.125 0.156 0.153 0.199 ...

> summary(pelod)

 Vlaga      Sevanje         Kalivost
 HH:72   Min.   :100.0   Min.   :0.00000
 RH:72   1st Qu.:275.0   1st Qu.:0.02870
         Median :375.0   Median :0.06075
         Mean   :393.8   Mean   :0.06756
         3rd Qu.:525.0   3rd Qu.:0.10470
         Max.   :700.0   Max.   :0.22000
```

Grafični prikaz podatkov:

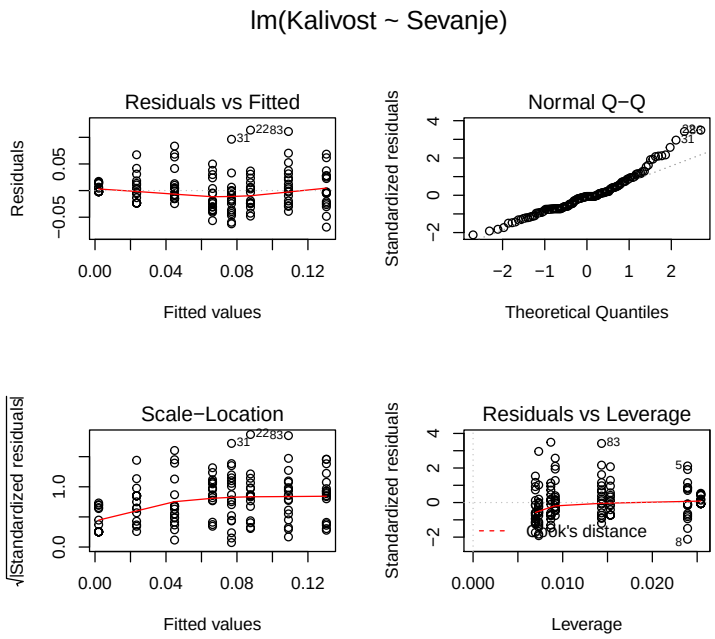
```
> plot(pelod$Sevanje,pelod$Kalivost, pch=16, xlab="Sevanje (Gy)", ylab="Kalivost")
```



Slika 21: Kalivost semen glede na `Sevanje`

Slika 21 nakazuje, da je variabilnost za kalivost pri različnih odmerkih sevanja različna: pri manjših odmerkih sevanja je kalivost večja in njena variabilnost tudi. Poglejmo, kako nekonstantno varianco vidimo na ostankih modela.

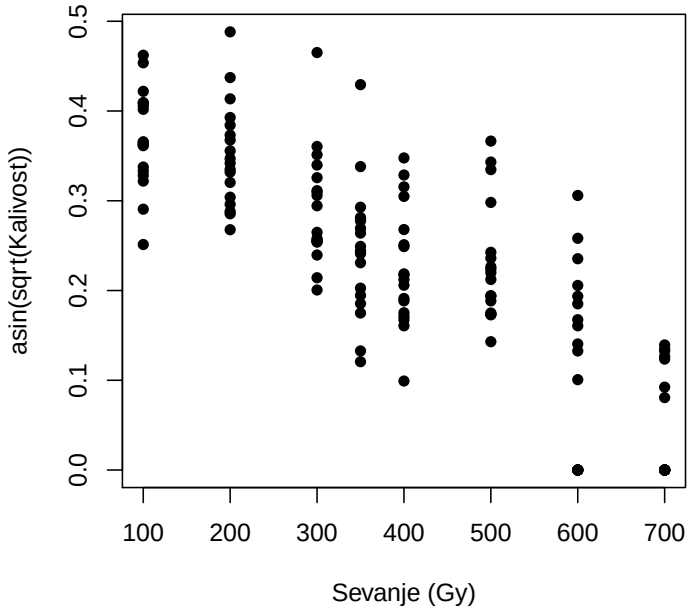
```
> model.p0<-lm(Kalivost~Sevanje, data=pelod)
```



Slika 22: Ostanki za `model.p0`

Za nadaljno analizo bomo uporabili `asin(sqrt(p))`, `logit` transformacija ni primerna, ker so med podatki za `Kalivost` ničle. Narišimo izhodiščne podatke z uporabo `asin(sqrt(p))` transformacije:

```
> plot(pelod$Sevanje,asin(sqrt(pelod$Kalivost)), pch=16,
+      xlab="Sevanje (Gy)", ylab="asin(sqrt(Kalivost))")
```

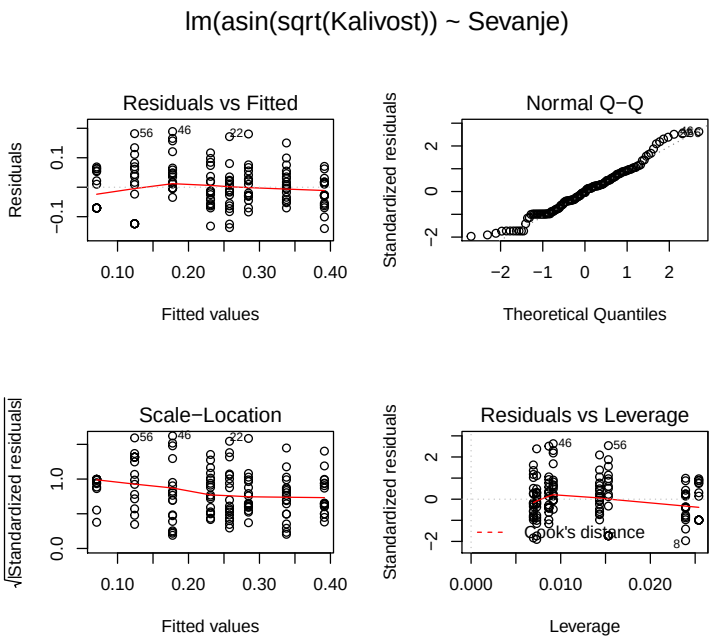


Slika 23: `asin(sqrt(Kalivost))` glede na `Sevanje`

Slika 23 kaže, da je varianca transformirane kalivosti manj problematična. Naredimo model na transformirani spremenljivki:

```
> model.p1<-lm(asin(sqrt(Kalivost))~Sevanje, data=pelod)
```

Ostanki na Sliki 24 kažejo manjšo heteroskedastičnost.



Slika 24: Ostanki za `model.p1`

Analizo tega primera bomo nadaljevali v naslednjih poglavjih, ko bomo v model vključili tudi opisno spremenljivko `Vlaga`.

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.

```
> koruza<-read.table(file="KORUZA.txt", header = TRUE)
> head(koruza)

  blok gostsetve gostvznika prid.ha prid.rast
1    1    65.230      51.85    5880     0.090
```

```
2    1    23.610      23.51    2936     0.124
3    1   129.900     123.80    6962     0.054
4    1    50.480      49.52    5152     0.102
5    1    47.620      41.27    4129     0.087
6    1     4.967       4.04     720     0.145
```

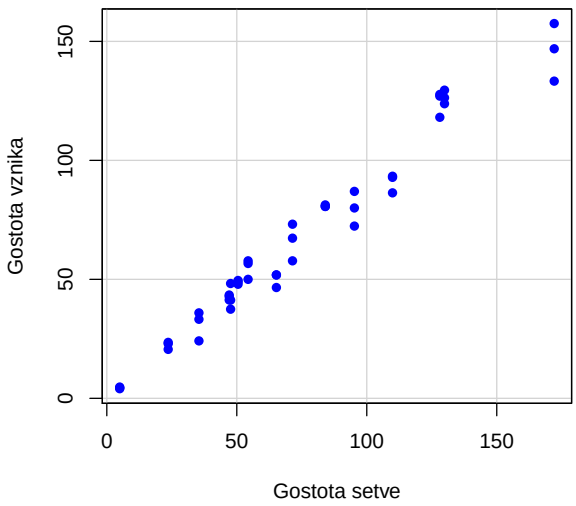
```
> str(koruza)
```

```
'data.frame':        45 obs. of  5 variables:
 $ blok      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ gostsetve : num  65.2 23.6 129.9 50.5 47.6 ...
 $ gostvznika: num  51.9 23.5 123.8 49.5 41.3 ...
 $ prid.ha   : int  5880 2936 6962 5152 4129 720 5219 6481 3508 6719 ...
 $ prid.rast : num  0.09 0.124 0.054 0.102 0.087 0.145 0.073 0.038 0.101 0.071 ...
```

```
> summary(koruza)
```

```
      blok     gostsetve         gostvznika        prid.ha       prid.rast
 Min.   :1   Min.   :  4.967   Min.   :  4.04   Min.   : 717   Min.   :0.0380
 1st Qu.:1   1st Qu.: 47.080   1st Qu.: 41.34   1st Qu.:4129   1st Qu.:0.0600
 Median :2   Median : 65.230   Median : 56.67   Median :5176   Median :0.0840
 Mean   :2   Mean   : 74.632   Mean   : 67.37   Mean   :5095   Mean   :0.0816
 3rd Qu.:3   3rd Qu.:109.900   3rd Qu.: 86.96   3rd Qu.:6595   3rd Qu.:0.0940
 Max.   :3   Max.   :172.100   Max.   :157.50   Max.   :8433   Max.   :0.1560
```

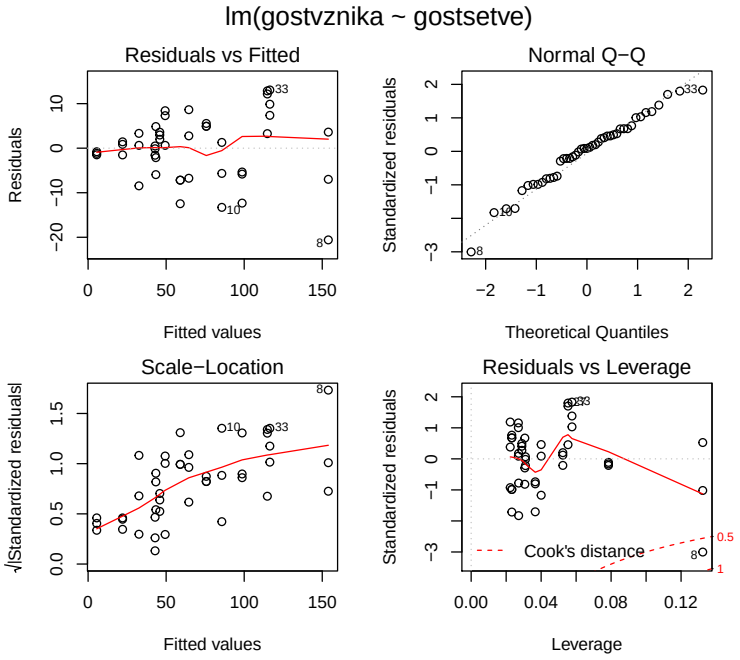
```
> scatterplot(gostvznika~gostsetve, regLine=FALSE, smooth=FALSE,
+             boxplots=FALSE, data=koruza, pch=16,
+             xlab="Gostota setve", ylab="Gostota vznika")
```



Slika 25: Odvisnost gostote vznika od gostote setve

Slika 25 kaže, da lahko privzamemo model za enostavno linearno regresijo. Analizirajmo rezultate:

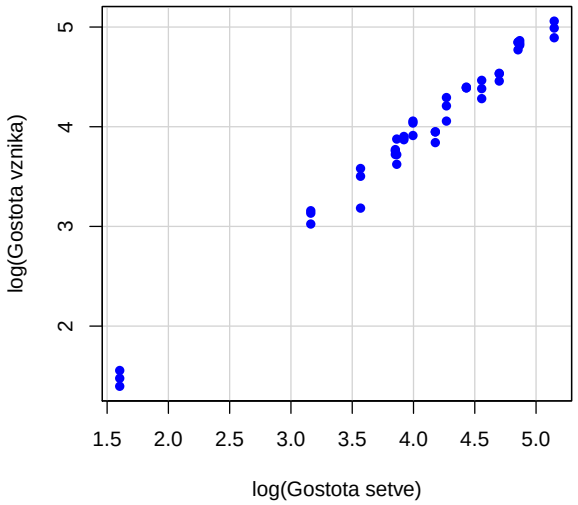
```
> mod <- lm(gostvznika~gostsetve, data=koruza)
```



Slika 26: Ostanki za `mod`

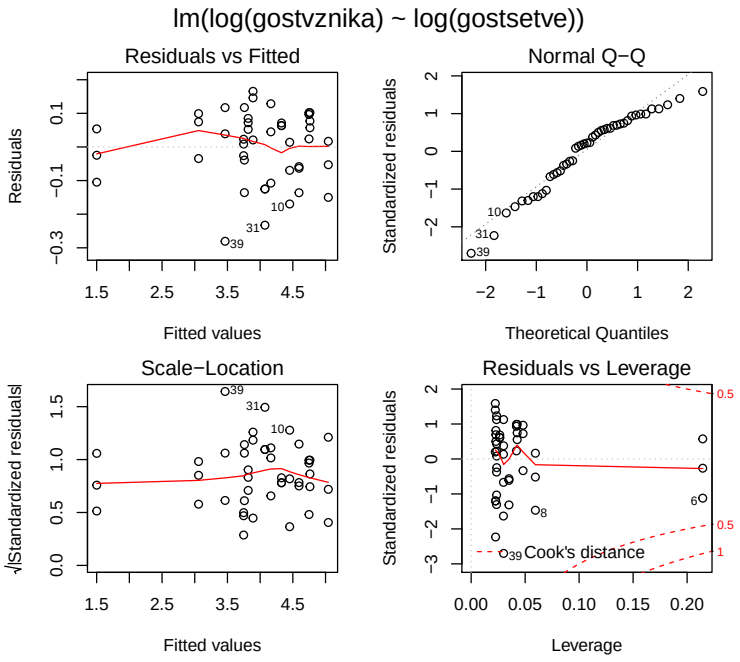
Graf 1 na Sliki 26 kaže, da so točke razporejene v obliki klina. To nakazuje na problem nekonstantne variance. Graf 3 to potrjuje, gladilnik je naraščajoč. Ker sta spremenljivki, ki ju obravnavamo v modelu, kvocienta (gostoti), skušamo problem nekonstantne variance rešiti s tem, da ju logaritmiramo in naredimo t. i. `log-log` model.

```
> scatterplot(log(gostvznika)~log(gostsetve), regLine=F,
+             smooth=FALSE, boxplots=FALSE,  data=koruza, pch=16,
+             xlab="log(Gostota setve)",
+             ylab="log(Gostota vznika)")
```



Slika 27: Odvisnost logaritmirane gostote vznika od logaritmirane gostote setve

```
> mod.log <- lm(log(gostvznika)~log(gostsetve), data=koruza)
```



Slika 28: Ostanki za `mod.log`

Slika 28 kaže, da je nekonstantna varianca odpravljena. Poglejmo povzetek modela.

```
> summary(mod.log)

Call:
lm(formula = log(gostvznika) ~ log(gostsetve), data = koruza)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28056 -0.06348  0.02296  0.07461  0.16547

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.10011    0.07796  -1.284    0.206
log(gostsetve)  0.99896    0.01879  53.173   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1055 on 43 degrees of freedom
Multiple R-squared:  0.985,        Adjusted R-squared:  0.9847
F-statistic:  2827 on 1 and 43 DF,  p-value: < 2.2e-16

> confint(mod.log)

                    2.5 %     97.5 %
(Intercept)    -0.2573256 0.05711203
log(gostsetve)  0.9610710 1.03684529
```

V splošnem log-log model zapišemo takole:

$$y = \beta_0 \cdot x^{\beta_1} \cdot \varepsilon. \tag{14}$$

Ta model se imenuje tudi multiplikativni model, saj se v kontekstu več napovednih spremenljivk zapiše kot produkt, tudi ε je del produkta. Npr. za dve številski napovedni spremenljivki x_1 in x_2 ima multiplikativni model naslednjo obliko:

$$y = \beta_0 \cdot x_1^{\beta_1} \cdot x_2^{\beta_2} \cdot \varepsilon. \tag{15}$$

Model (14) zlahka lineariziramo:

$$log(y) = log(\beta_0) + \beta_1 log(x) + log(\varepsilon).$$

Pomen parametra β_1 dobimo z diferenciranjem in je naslednji:

$$\beta_1 = \frac{dy/y}{dx/x}. \tag{16}$$

Torej: če se x spremeni za 1 %, se y spremeni za β_1 %.

Vsebinska interpretacija naklona za `mod.log`: če se gostota setve poveča za 1 %, se gostota vznika poveča za 0.999 %, pripadajoč 95 % interval zaupanja je [0.961 %, 1.037 %].

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)
```

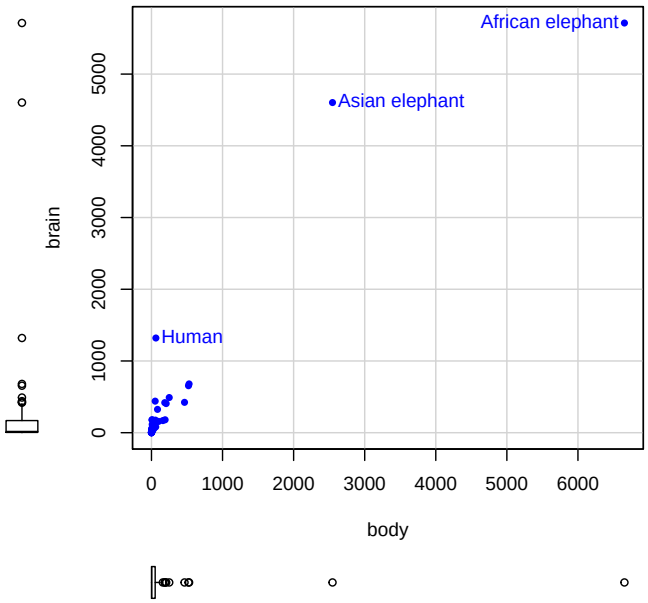
```
                  body brain
Arctic fox       3.385  44.5
Owl monkey       0.480  15.5
Mountain beaver  1.350   8.1
Cow            465.000 423.0
Grey wolf       36.330 119.5
Goat            27.660 115.0
```

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?

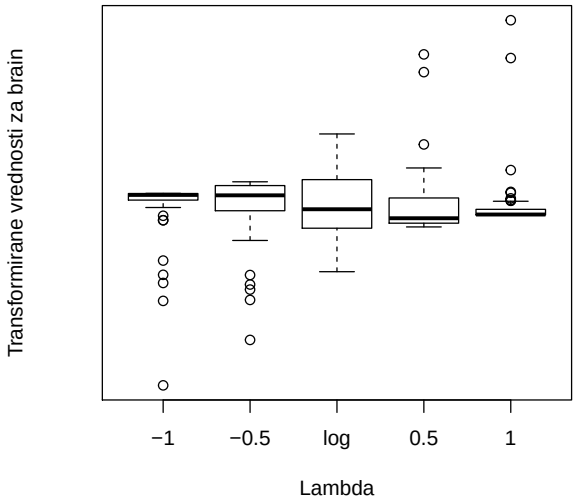
d) Analizirajte ustrezni model in obrazložite rezultate.



Slika 29: Odvisnost `brain` od `body` za 62 sesalcev

Iz slike vidimo, da sta porazdelitvi za `body` in za `brain` zelo asimetrični. Večina točk je v levem spodnjem kotu slike. Poskusimo dobiti ustrezno transfomacijo za `brain`, da bi bila porazdelitev bolj simetrična.

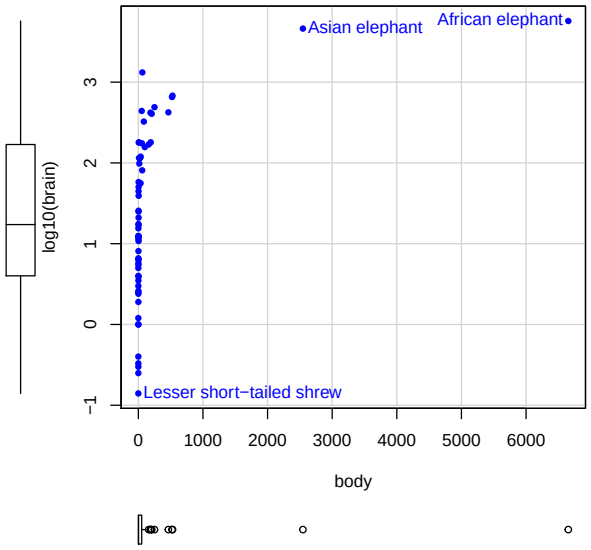
```
> symbox(~brain, xlab= "Lambda", ylab="Transformirane vrednosti za brain",
+        data=mammals)
```



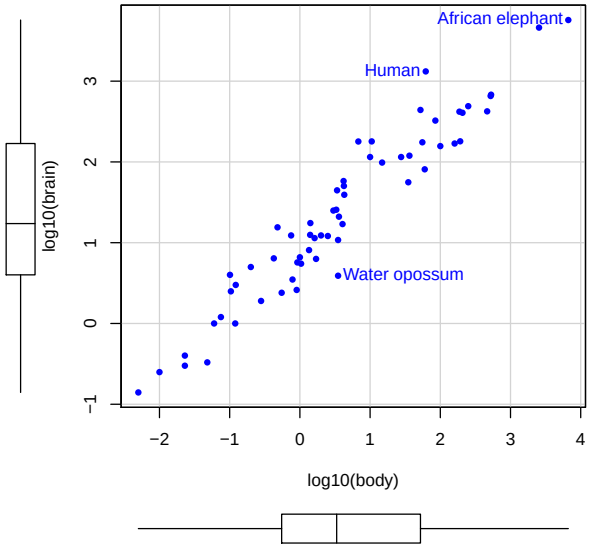
Slika 30: Okviri z ročaji za različne transformacije za spremenljivko `brain`

Na Sliki 30 so prikazani okviri z ročaji za pet izbranih transformacij za spremenljivko `brain`. Slika kaže, da pride v poštev transfomacija $\lambda = 0$, to je logaritemska transformacija. Za lažjo vsbinsko interpretacijo logaritmiranih vrednosti na sliki je smiselno uporabiti desetiški logaritem.

Slika 31 kaže odvisnost `log10(brain)` od `body`. Poskusimo doseči linearnost z uporabo logaritemske transfomacije tudi za `body` (Slika 32).



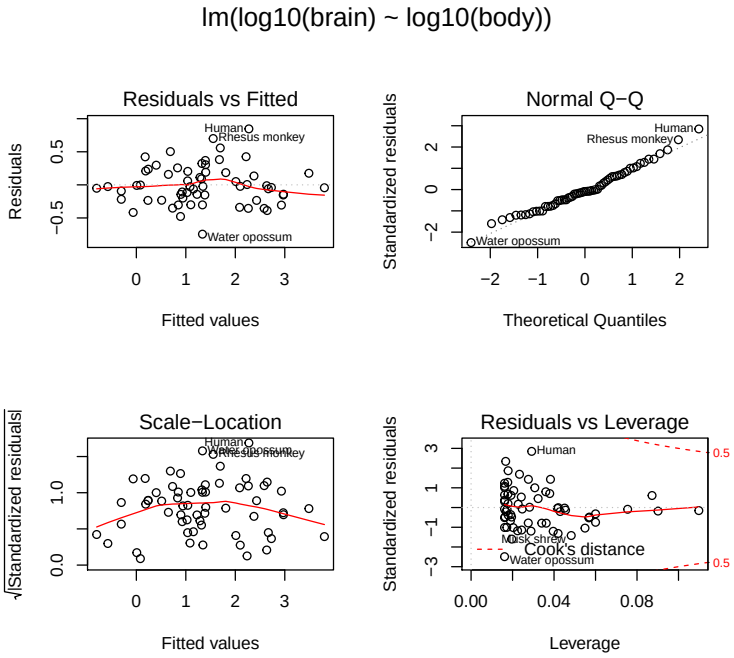
Slika 31: Odvisnost `log10(brain)` od `body` za 62 sesalcev



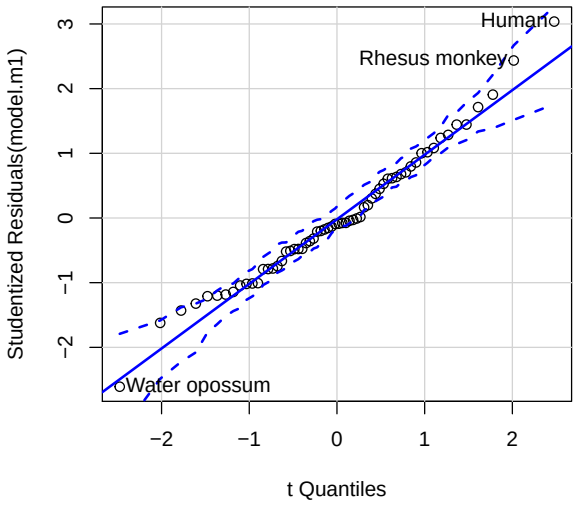
Slika 32: Odvisnost `log10(brain)` od `log10(body)` za 62 sesalcev

Slika 32 kaže, da je odvisnost med logaritmiranima spremenljivkama linearna. Model bomo analizirali v sklopu log-log modelov.

```
> model.m1<- lm(log10(brain)~log10(body), data=mammals)
```



Slika 33: Grafični prikaz ostankov za `model.m1`



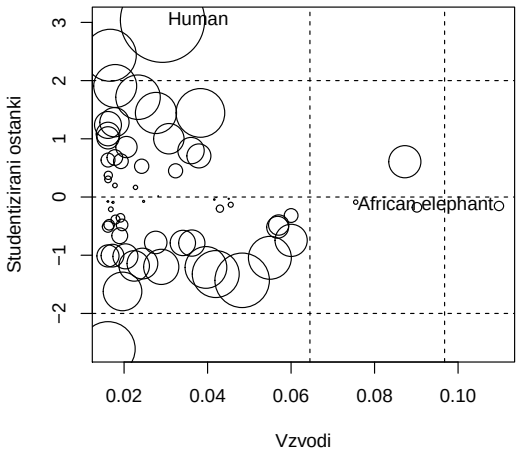
Slika 34: QQ grafikon za studentizirane ostanke za `model.m1` s 95 % bootstrap ovojnico

Sliki 33 kaže, da je porazdelitev ostankov sprejemljiva, vplivnih točk ni, tri točke imajo standardizirane ostanke po absolutni vrednosti večje od 2: `Human`, `Rhesus monkey` in `Water oppusum`, vendar

so te točke na Sliki 34 znotraj 95 % bootstrap ovojnice, kar pomeni, da ne predstavljajo regresijskih osamelcev. Tudi statistični test, ki temelji na sudentiziranih ostankih, ne pokaže statistične značilnosti.

```
> outlierTest(model.m1)

No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferroni p
Human 3.036941          0.0035554      0.22043
```

```
> influencePlot(model.m1,id=list(n=1), xlab="Vzvodi",ylab="Studentizirani ostanki")

                    StudRes        Hat       CookD
Human             3.0369408 0.02920799 0.122022105
African elephant -0.1537331 0.10979911 0.001481634
```

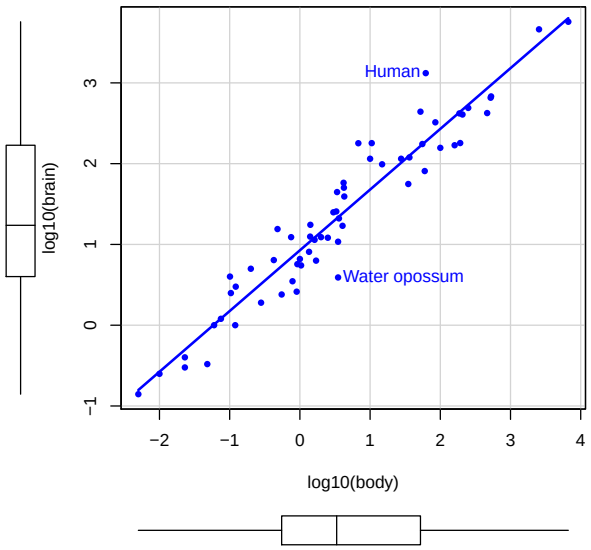


Slika 35: Grafični prikaz studentiziranih ostankov, vzvodov in Cookove razdalje za `model.m1`

```
> summary(model.m1)

Call:
lm(formula = log10(brain) ~ log10(body), data = mammals)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74503 -0.21380 -0.02676  0.18934  0.84613

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.92713    0.04171   22.23   <2e-16 ***
```

```
log10(body)  0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3015 on 60 degrees of freedom
Multiple R-squared:  0.9208,        Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

```
> confint(model.m1)

                2.5 %    97.5 %
(Intercept) 0.8436923 1.0105616
log10(body) 0.6947503 0.8086215
```



Slika 36: Odvisnost `log10(brain)` od `log10(body)` za 62 sesalcev in napovedi za `model.m1`

Interpretacija rezultatov:

- Če se masa telesa poveča za 1 %, se masa možganov poveča 0.75 %; pripadajoč IZ je od 0.7 % do 0.8 %.
- `log(body)` pojasni cca 92 % variabilnosti `log(brain)`.
- Kandidat za regresijskega osamelca je `Human`, vendar test pokaže, da ni osamelec. Vrednost za `log(brain)` pri `Human` je bistveno večja, kot jo napove model.
- Ob upoštevanju trikratnega povprečnega vzvoda za mejo za vzvodne točke je vzvodna točka le `African elephant`.