Kazalo

1 PREDPOSTAVKE NISO IZPOLNJENE, KAKO NAPREJ?

V prejšnjem poglavju smo predstavili, kako za izbrani linearni model preverimo, ali so predpostavke izpolnjene. V tem poglavju bomo predstavili, kako postopamo, ko so predpostavke kršene.

Metode odpravljanja kršitev predpostavk linearnega modela v splošnem delimo v dve skupini:

- transformacije,
- modeliranje.

V preteklosti so bile trasformacije glavni način reševanja tovrstnih problemov in uporabni statistiki jih pri statističnih analizah še vedno pogosto uporabljajo. V zadnjih dveh destletjih pa vse bolj stopajo v ospredje metode, ki uporabljajo osnovne, netransformirane spremenljivke in problem obidejo z ustreznim, drugačnim načinom modeliranja.

Najprej naredimo pregled metod, ki se uporabljajo ob kršitvi posamezne predpostavke linearnega modela:

1. Konstantna varianca napak
   - transformacija odzivne spremenljivke
   - modeliranja: tehtana metoda najmanjših kvadratov, interakcijski členi
2. Linearnost
   - transformacija odzivne spremenljivke in/ali napovednih spremenljivk
   - modeliranje:polinomska regresija, interakcijski členi
   - modeliranje: zlepki, aditivni modeli (*splines*, *additive models*)

3. Vplivnost (to ni prepostavka modela, je pa lahko posledica neizpolnjenih predpostavk)
   - transformacija napovednih spremenljivk
   - modeliranje: tehtana metoda najmanjših kvadratov, interakcijski členi
4. Normalna porazdelitev ostankov
   - transformacija odzivne spremenljivke
   - modeliranje: posplošeni linearni modeli (GLM) (ne bomo obravnavali)
5. Neodvisnost napak
   - diferenciranje (časovne vrste)
   - modeliranje: linearni mešani modeli (longitudinalni, hierarhični)
   - modeliranje: *generalised estimating equations*
   - modeliranje: kopule (*copulas*)

1.1 Transformacije

V naslednjih poglavjih bomo spoznali različne trasformacije odzivne in/ali napovedne spremenljivke ter situacije v katerih jih uporabimo. Pri uporabi transformacij moramo biti previdni, ker to lahko povzroči kršenje, katere od drugih predpostavk linearnega modela. Za vsako transformacijo je v procesu modeliranja potrebno ugotoviti, ali smo problem res rešili.

- Če nas pri modeliranju zanima inferenca (testiranje domnev, intervali zaupanja), potem naredimo model na transformirani spremenljivki in ponovno izvedemo diagnostiko modela;
- Če modeliramo z namenom napovedovanja, potem je eden od načinov, da primerjamo PRESS ostanke modela na transformiranih in netransformiranih podatkih. Različne skale podatkov se rešimo z inverzno transformacijo PRESS ostankov:
  - uporabimo transformacijo $z = f(y)$ (f je obrnljiva funkcija) in modeliramo spremenljivko z;
  - izračunamo PRESS ostanke za model $z_i - \hat{z}_{i,-i}$, kjer je $\hat{z}_{i,-i}$ napoved modela narejenega na podatkih brez i-te točke;
  - ker nas zanima napoved v originalni skali, na $\hat{z}_{i,-i}$ uporabimo inverzno funkcijo in izračunamo $y_i - f^{-1}(\hat{z}_{i,-i})$;
  - Izračunamo t. i. $PRESS$-statistiko za model na originalnih podatkih

$$\sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2$$

in $PRESS$-statistiko za model na transformiranih podatkih

$$\sum_{i=1}^{n}(y_i - f^{-1}(\hat{z}_{i,-i}))^2,$$

če smo s trasformacijo izboljšali napovedi, bo vrednost tako izračunane $PRESS$-statistike manjša.

– Povdariti velja, da primerjave modela na originalnih in modela na transformiranih podatkih ne smemo narediti na osnovi koeficienta determinacije R^2 ali standardnega odklona napak regresije. Poskrbeti moramo, da primerjavo delamo na podlagi vrednosti v originalni skali.

V splošnem je zaželeno, da so vrednosti odzivne spremenljivke definirane na celi realni osi, ker to po definiciji linearnega modela velja za njihove napovedi $\hat{y}_i = \mathbf{X}\boldsymbol{\beta}$. Nekatere transformacije odzivne spremenljivke spremenijo razpon vrednosti v tej smeri:

- če $y \in (0, \infty)$, velja $log(y) \in (-\infty, \infty)$
- če $y \in (0, 1)$, velja $log(\frac{y}{1-y}) \in (-\infty, \infty)$

1.1.1 Varianca je sorazmerna $f(\mathbb{E}(y))$

Če je varianca σ^2 odvisna od pričakovane vrednosti odzivne spremenljivke $\mathbb{E}(y)$, lahko poskusimo z različnimi transformacijami odzivne spremenljivke y. Modeliranje in inferenco v tem primeru izvedemo na transformiranih podatkih.

V Tabeli 1 so navedene primerne transformacije pri različnih zvezah med varianco σ^2 in pričakovano vrednostjo $\mathbb{E}(y)$. V praksi najpogosteje uporabljeni funkciji sta logaritem in kvadratni koren.

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

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

Če y predstavlja delež, to pomeni, da ima omejeno zalogo vrednosti na intevalu [0,1]. V ozadju je slučajna spremenljivka, ki je porazdeljena po binomski porazdelitvi $b(n, \pi)$ za katero velja, da je $\mathbb{E}(y) = n\pi$ in $Var(y) = n\pi(1-\pi)$. Varianca deležev blizu 0 oz. blizu 1 je manjša od variance deležev blizu 1/2. Obstojata dve primerni transformaciji: $asin(\sqrt{y})$ in $logit(y)$.

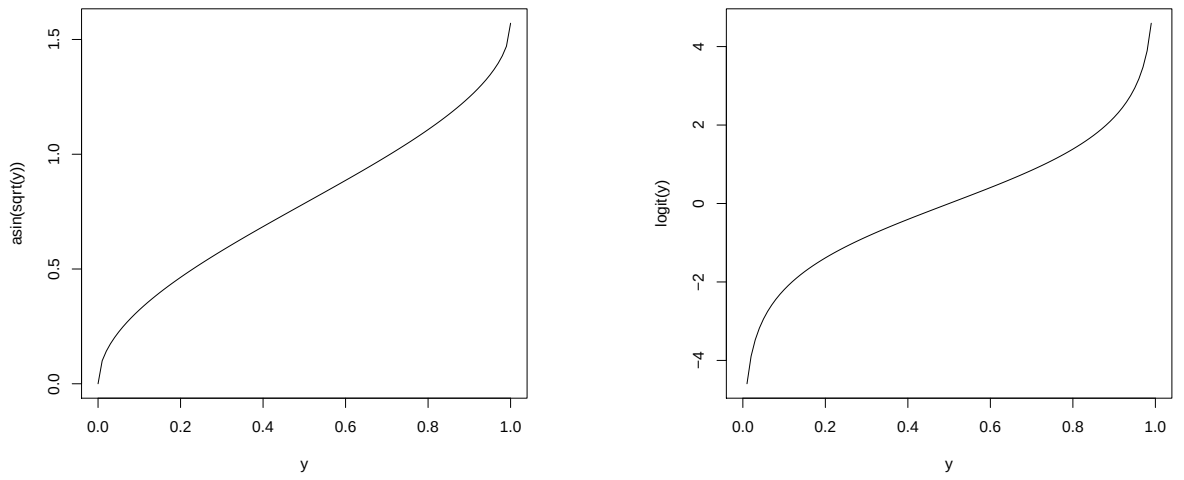
Za transformacijo $asin(\sqrt{y})$ velja, da je standardni odklon transformiranih vrednosti približno enak $1/2\sqrt{n}$ in je torej neodvisen od parametra binomske porazdelitve π. Ta transformacija stabilizira varianco in odpravi heteroskedastičnost.

Alternativna transformacija je `logit`:

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

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

Grafični prikaz obeh transformacij je na Sliki 10. Oba grafa sta sigmoidne oblike (S-oblike), zalogi vrednosti transformirane spremenljivke pa sta različni.



(a) `asin` transformacija

(b) `logit` transfomacija

Slika 1: Transformaciji, ki ju uporabljamo, če je odvisna spremenljivka delež; skali na ordinatah sta različni

Pri analizi deležev se pokaže, da z nekonstantno varianco ni težav, če so vrednosti deležev približno na intervalu [0.25, 0.75], tedaj transformacija ni potrebna.

1.1.2 Box-Cox transformacije

Box in Cox (1964) sta predlagala družino transformacij za odvisno spremenljivko y, ki je funkcija parametra λ. Za vsako točko $i = 1, ..., n$ v tem primeru linearni model zapišemo:

$$z_i = \begin{cases} \frac{y^\lambda - 1}{\lambda} = (\mathbf{X}\boldsymbol{\beta})_i + \varepsilon_i & , \lambda \neq 0 \\ ln(y) = (\mathbf{X}\boldsymbol{\beta})_i + \varepsilon_i & , \lambda = 0 \end{cases} . \tag{2}$$

kjer velja $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$. Box-Cox transformacije so definirane za $\mathbf{y} > 0$. Poleg ocen vektorja parametrov $\boldsymbol{\beta}$ hkrati ocenjujemo tudi parameter λ. Za ocenjevanje parametrov uporabimo metodo največjega verjetja (ML, *maximum likelihood*), poiščemo parametre pri katerih je verjetje za $\mathbf{y}$ največje. Oceno $\hat{\lambda}$ dobimo z numeričnim postopkom kjer za različne vrednosti $\hat{\lambda}$ izračunamo verjetje za $\mathbf{y}$. Izberemo λ, pri kateri ima logaritem verjetja maksimalno vrednost.

Za izračun lahko uporabimo funkcijo `powerTransform` iz paketa `car`, ki vrne optimalni λ in pripadajoči interval zaupanja ter izvede dva informativna testa:

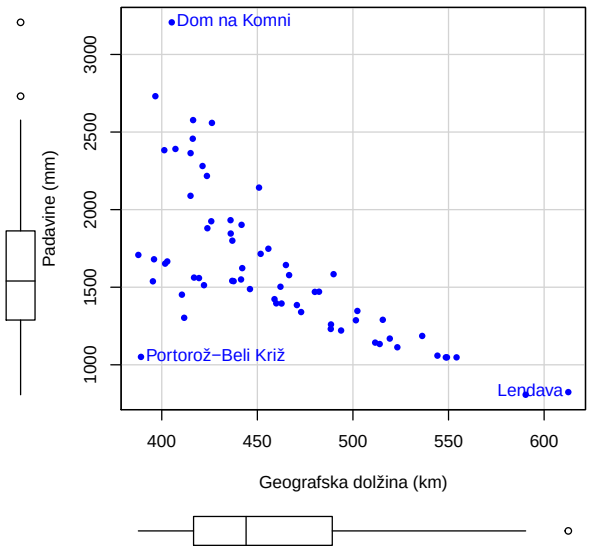
- $H_0 : \lambda = 0$ (ustrezna je logaritemska transformacija),
- $H_0 : \lambda = 1$ (y ni treba transformirati).

Grafični prikaz odvisnosti logaritma verjetja od λ dobimo s funkcijo `boxCox` iz paketa `car`.

Informativno λ ocenimo na podlagi grafičnega prikaza 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.

Primer: `postaje`, 2. del

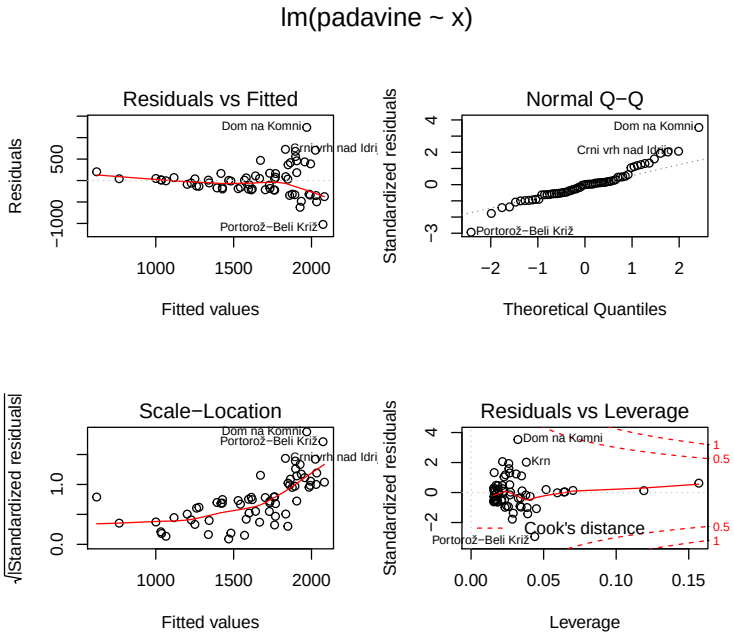
Za meteorološke postaje (datoteka POSTAJE.txt) analizirajmo odvisnost letne količine padavin (`padavine`) od geografske dolžine v Gauss-Krugerjevih koordinatah, ki so izražene v metrih (`x.gdol`).



Slika 2: Odvisnost letne količine padavin od geografske dolžine za 64 postaj; podatki so za leto 1992

Naredimo linearni regresijski model in poglejmo ostanke (Slika 3):

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



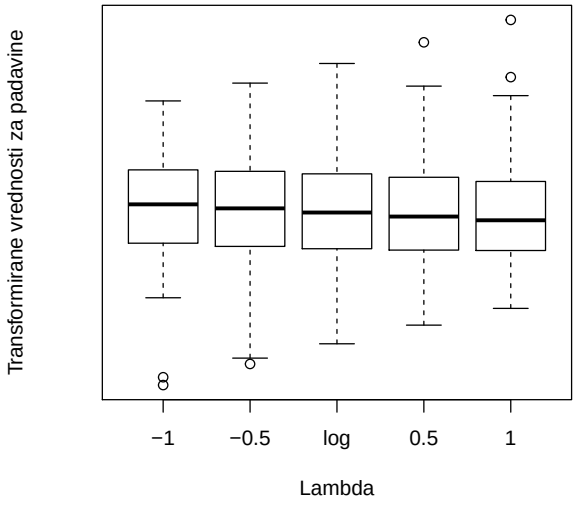
Slika 3: Grafični prikaz ostankov za `model.1`

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.

V tem primeru so ocene parametrov sicer ustrezne (pri dokazu nepristranskosti cenilk smo pokazali, da ne potrebujemo predpostavke o konstantni varianci), njihove standardne napake pa ne, zato kakršnakoli inferenca za ta model ni utemeljena.

Slika 4 prikazuje porazdelitve transformiranih vrednosti za `padavine` pri petih različnih vrednostih za λ.

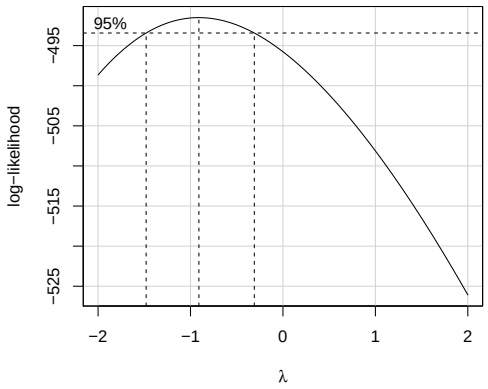
```
> symbox(~padavine, xlab= "Lambda", ylab="Transformirane vrednosti za padavine",
+        data=postaje)
```



Slika 4: Okviri z ročaji za različne transformacije za spremenljivko `padavine`

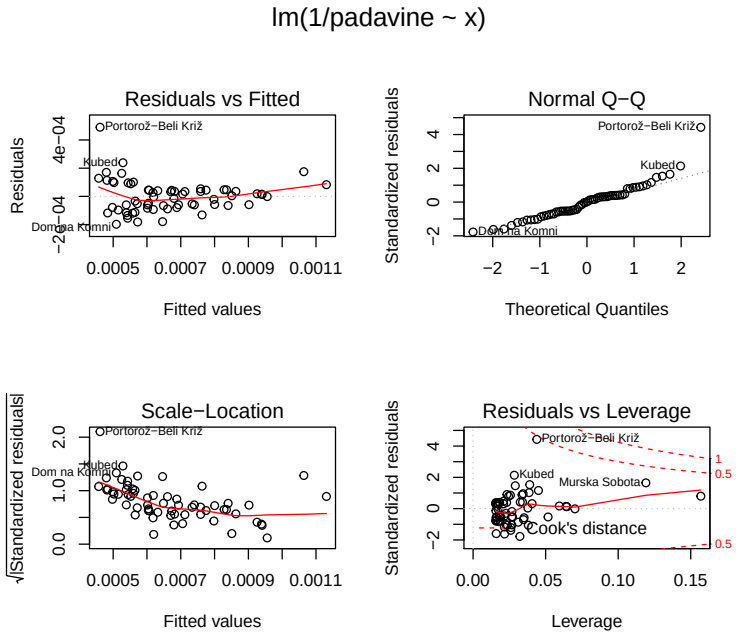
```
> summary(powerTransform(model.1))

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
```

```
> par(mar=c(4,4,1,1))
> boxCox(model.1)
```



Slika 5: 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. Naredimo model za tako transformirano količino padavin in analiza ostankov tega modela (Slika 6) pokaže, da se problema nekonstantne variance nismo rešili. 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.



Slika 6: Grafični prikaz ostankov za `model.2`

1.1.3 Nelinearnost, ki se da linearizirati

V nekaterih situacijah lahko z ustrezno transformacijo odzivne spremenljivke, ali napovedne spremenljivke, ali tudi obeh, nelinearno zvezo spremenimo v linearno.

Exponentna zveza

Če zvezo med y in x opišemo z eksponentno funkcijo:

$$y = \beta_0 e^{\beta_1 x}, \tag{3}$$

z logaritmiranjem izraza dobimo linearno zvezo:

$$ln(y) = ln(\beta_0) + \beta_1 x = \beta_0^* + \beta_1 x. \tag{4}$$

Pomen parametra β_1 ugotovimo z diferenciranjem enačbe (4):

$$\beta_1 dx = \frac{dy}{y}.$$

Torej: $\beta_1 = \frac{dy}{y}/dx$ izraža relativno spremembo odzivne spremenljivke ob povečanju napovedne spremenljivke za eno enoto. Drugače to povemo, če se x poveča za eno enoto, se y spremeni za $100\beta_1$ %.

Multiplikativna zveza

V določenih primerih je odzivna spremenljivka v multiplikativni zvezi z eno ali več regresorji:

$$y = \beta_0 x_1^{\beta_1}, \tag{5}$$

za dva regresorja x_1 in x_2

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2}. \tag{6}$$

Modela (5) in (6) zlahka lineariziramo:

$$log(y) = log(\beta_0) + \beta_1 log(x),$$

oziroma

$$log(y) = log(\beta_0) + \beta_1 log(x_1) + \beta_2 log(x_2).$$

Pomen parametra β_1 v (5) je:

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

Za enak povem gre v (6), ko je vrednost x_2 konstantna. Torej gre za relativne spremembe tako regresorja kot napovedne spremenljivke. Če se x poveča za 1 %, se y spremeni za β_1 %.

Pri uporabi tovrstnih trasformacij je potrebno razmisliti, kakšno vlogo v modelu imajo napake. Poglejmo enostaven primer eksponentne zveze, kjer je napaka lahko v aditivi zvezi z odzivno spremenljivko:

$$y_i = \beta_0 e^{\beta_1 x_i} + \varepsilon_i. \tag{7}$$

Logaritmiranje v tem primeru ne privede do linearnega modela, ker za $log(\varepsilon_i)$ ne moremo predpostaviti normalne porazdelitve, če normalna porazdelitev velja za ε_i.

$$log(y_i) \neq log(\beta_0) + \beta_1 x_i + log(\varepsilon_i),$$

Drugače je, če napaka v izrazu nastopa multiplikativno:

$$y_i = \beta_0 e^{\beta_1 x_i + \varepsilon_i},$$

$$log(y) = log(\beta_0) + \beta_1 x_i + \varepsilon_i.$$

Ker v praksi ponavadi ne vemo, kateri model napake je pravi, je v splošnem nelinearne zveze bolje modelirati z nelinearnimi modeli.

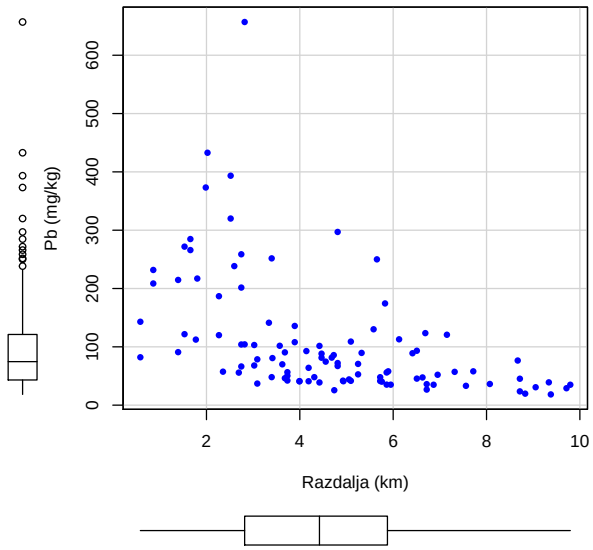
Primer: `kovine`

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 z razdaljo 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")])
```

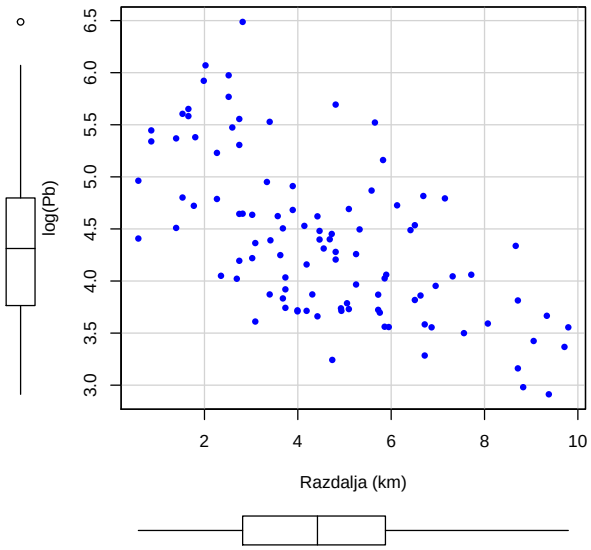
```
       Pb             razdalja
 Min.   : 18.40   Min.   :0.5831
 1st Qu.: 43.15   1st Qu.:2.8178
 Median : 74.60   Median :4.4204
 Mean   :109.80   Mean   :4.5957
 3rd Qu.:121.20   3rd Qu.:5.8770
 Max.   :657.00   Max.   :9.7949
```

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



Slika 7: Odvisnost koncentracije Pb od razdalje do cinkarne, podatki Celje 2000

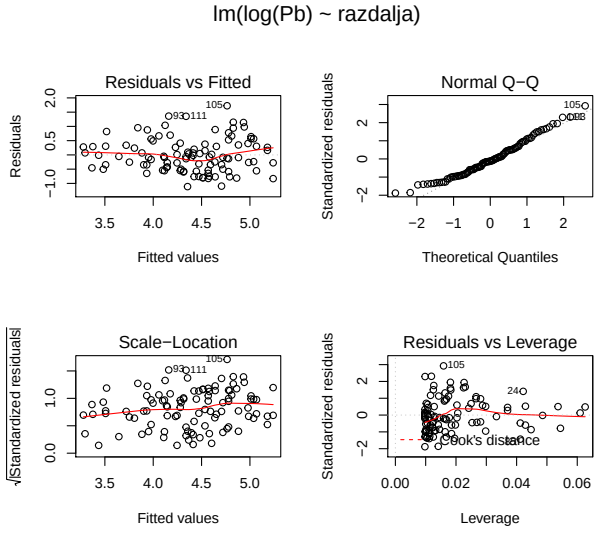
Na Sliki 7 je razvidna velika variabilnost Pb, njegova porazdelitev je asimetrična. Kaže se različna variabilnost za različne oddaljenostih od cinkarne, pri majhnih vrednostih je variabilnost večja kot pri velikih; torej imamo problem nekonstantne variance, tudi predpostavka o linearni odvisnosti je vprašljiva. Poskusimo z logaritemsko transformacijo `Pb`:



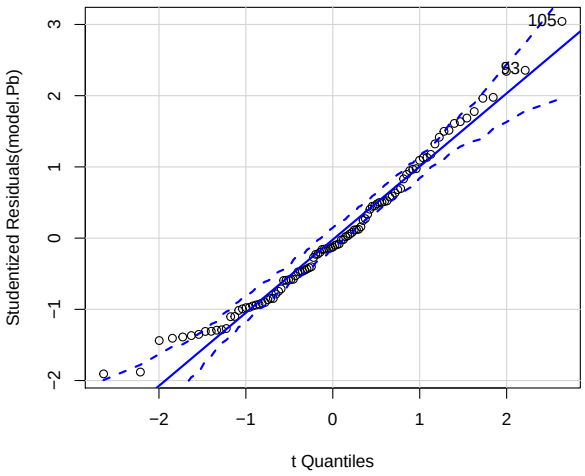
Slika 8: Odvisnost `log(Pb)` od `razdalja` do cinkarne

Slika 8 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)
```

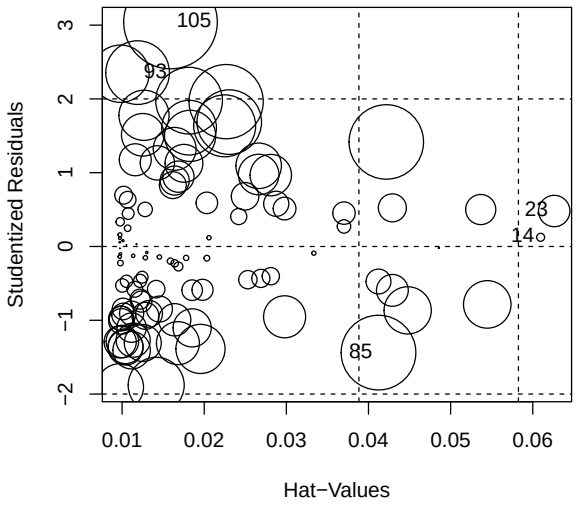


Slika 9: Grafični prikaz ostankov za `model.Pb`



Slika 10: 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 11: 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.

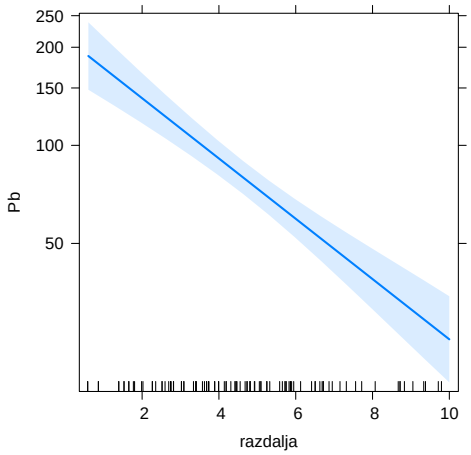
```
> (b <- coefficients(model.Pb))

(Intercept)    razdalja
  5.3642729  -0.2130161

> summary(model.Pb)$r.squared

[1] 0.3942114

> confint(model.Pb)

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

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)`.

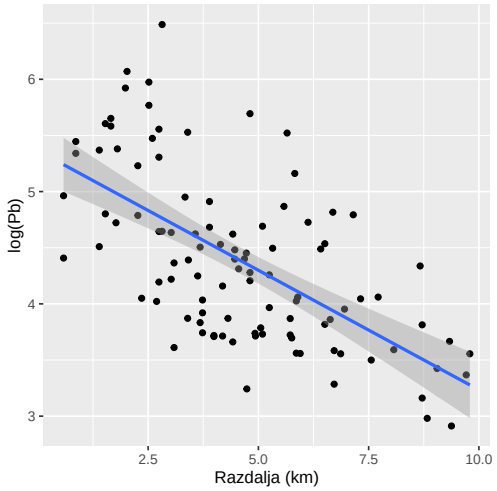
Slike 12, 13 in 14 na različne načine prikazujejo napovedi za `model.Pb`.

```
> library(effects)
> plot(effect(c("razdalja"), model.Pb, transformation=list(link=log, inverse=exp)),
+      axes=list(y=list(lab="Pb")), main="")
```



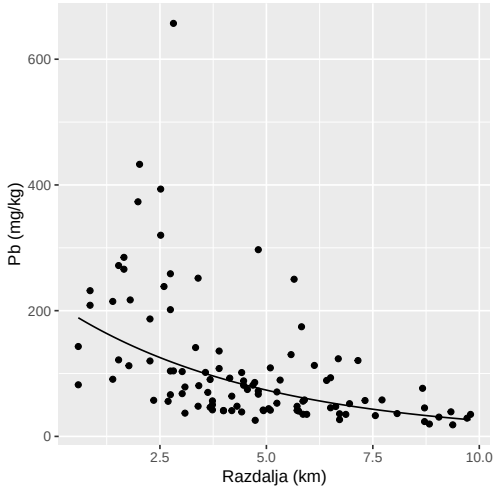
Slika 12: Odvisnost `Pb` od razdalje do cinkarne in pripadajoča regresijska premica s 95 % intervalom zaupanja za povprečno napoved `Pb`, skala za Pb je logaritemska

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



Slika 13: 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(b[1]+b[2]*razdalja) )
```



Slika 14: Odvisnost Pb od razdalje do cinkarne; eksponentni model

Primer: `trees`

V paketu `car` je podatkovni okvir `trees` s podatki za višino debla (`Height`), premer debla (`Girth`) in volumen debla (`Volume`) za 31 dreves (glej `help(trees)`). Najprej naredimo nov podatkovni okvir `drevesa` s podatki za premer debla `D` v metrih, višina drevesa `H` v metrih in volumen drevesa `Vol` v m^3.

```
> names(trees)

[1] "Girth"  "Height" "Volume"

> k1<-0.30480   ## feet -> m
> k2<-0.0254    ## inches -> m
> H<-trees$Height*k1
> D <-trees$Girth*k2
> Vol <-trees$Volume*(k1^3)
> drevesa<-data.frame(cbind(Vol, H, D))
> summary(drevesa)

      Vol                H                D
 Min.   :0.2888   Min.   :19.20   Min.   :0.2108
 1st Qu.:0.5493   1st Qu.:21.95   1st Qu.:0.2807
 Median :0.6853   Median :23.16   Median :0.3277
 Mean   :0.8543   Mean   :23.16   Mean   :0.3365
 3rd Qu.:1.0562   3rd Qu.:24.38   3rd Qu.:0.3874
 Max.   :2.1804   Max.   :26.52   Max.   :0.5232
```

Zanima nas, ali podatki podpirajo geometrijski model izračunavanja volumna telesa:

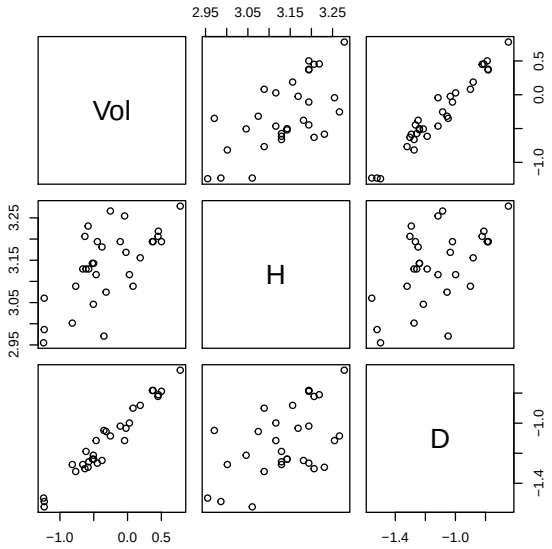
$$Vol = konst \cdot D^2 \cdot H. \tag{8}$$

To je multiplikativni model, saj se D^2 in H množita. Z logaritmiranjem (8) dobimo aditivni izraz, ki je primeren za analizo z linearnim modelom:

$$log(Vol) = log(konst) + 2 \cdot log(D) + 1 \cdot log(H). \tag{9}$$

Narišimo najprej matriko razsevnih grafikonov na logaritmiranih spremenljivkah.

```
> pairs(log(drevesa))
```

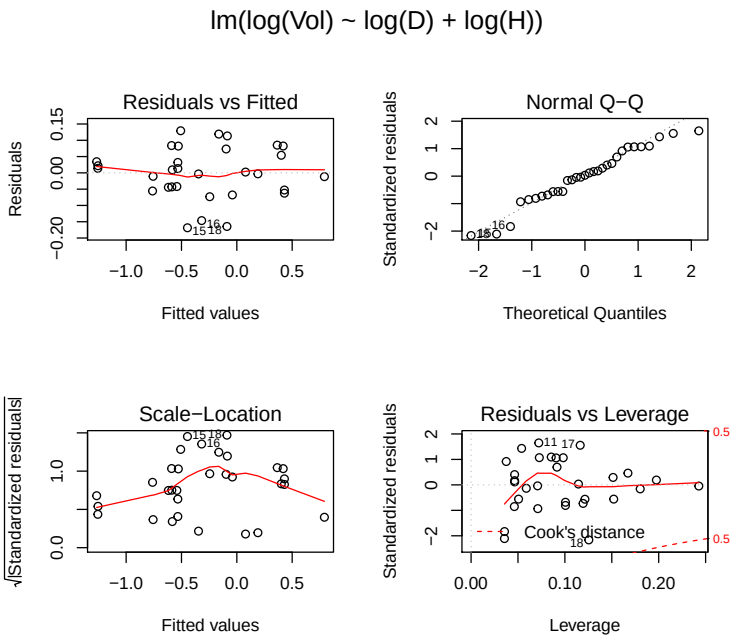


Slika 15: Matrika razsevnih grafikonov za logaritmirane spremenljivke iz podatkovnega okvira `drevesa`

Slika 15 kaže, da je robna odvisnost `log(Vol)` od `log(D)` in od `log(H)` linearna. Naredimo model na logaritmiranih spremenljivkah in preverimo, ali so podatki v skladu z geometrijskim modelom. Če je tako, je parameter modela β_D pri `log(D)` enak 2, parameter β_H pri `log(H)` pa 1. V tem primeru gre za hkratno testiranje dveh domnev:

$$H_0: \quad \beta_D = 2, \qquad H_0: \quad \beta_H = 1. \tag{10}$$

```
> model.d<- lm(log(Vol)~log(D)+log(H), data=drevesa)
```



Slika 16: Ostanki za `model.d`

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

[1] 0.9776784

> confint(model.d)

                2.5 %     97.5 %
(Intercept) -2.999654 -0.1730991
log(D)       1.828998  2.1363022
log(H)       0.698353  1.5358937
```

Po vseh kriterijih je model ustrezen, koeficent determinacije je izjemno visok. V intervalu zaupanja za β_D je vrednost 2, v intervalu zaupanja za β_H je vrednost 1. Na osnovi intervalov zaupanja za parametra lahko sklepamo, da obe ničelni domnevi obdržimo. To pomeni, da ni razlogov, da bi dvomili v ustreznost multiplikativnega modela.

1.1.4 Interakcija dveh številskih napovednih spremenljivk

Interakcijo dveh številskih spremenljivk v linearnem modelu najlažje razložimo na podlagi interpretacije ocen parametrov modela z dvema številskima napovednima spremenljivkama x_1 in x_2 ter njuno interakcijo, ki jo v model vključimo kot njun produkt x_1x_2.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i \tag{11}$$

Zamislimo si pričakovano vrednost tega modela v točki (x_{01}, x_{02}).

$$E(y|x_{01}, x_{02}) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_3 x_{01} x_{02}, \tag{12}$$

in v točki $(x_{01}, x_{02} + 1)$, kar pomeni, da se pri spremenljivki x_2 premaknemo za eno enoto naprej

$$E(y|x_{01}, x_{02}+1) = \beta_0+\beta_1 x_{01}+\beta_2(x_{02}+1)+\beta_3 x_{01}(x_{02}+1) = \beta_0+\beta_1 x_{01}+\beta_2 x_{02}+\beta_3 x_{01} x_{02}+\beta_2+\beta_3 x_{01}. \tag{13}$$

Iz (12) in (13) sledi, da je razlika pričakovanih vrednosti odvisna od vrednosti spremenljivke x_{01}:

$$E(y|x_{01}, x_{02} + 1) - E(y|x_{01}, x_{02}) = \beta_2 + \beta_3 x_{01}. \tag{14}$$

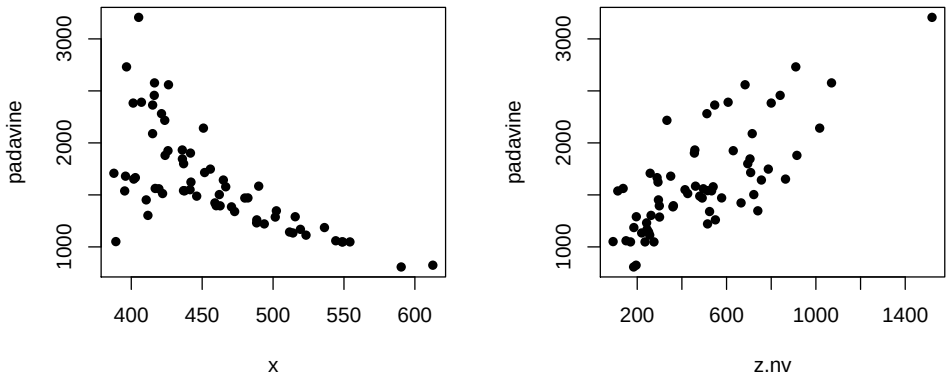
Torej velja, če x_{02} povečamo za eno enoto in ostane izbrana vrednost x_{01} nespremenjena, se pričakovana vrednost y poveča za $\beta_2 + \beta_3 x_{01}$. To pomeni, da je ta sprememba pri različnih vrednostih napovedne spremenljivke x_1 različna. Enako velja za spremembo y, če za eno enoto povečamo vrednost spremenljivke x_1, odvisna je od vrednosti x_2.

Interakcijske člene v model vključimo iz različnih razlogov. Prvi razlog je vsekakor poznavanje vpliva izbranih dejavnikov na odzivno spremenljivko. Drugi razlog je iskanje utreznih regresorjev v linearnem modelu pod pogojem, da bodo predpostavke izpolnjene.

Primer: `postaje`**, 3. del**

Videli smo že, kako je letna količina padavin v Sloveniji odvisna od nadmorske višine ter kako je odvisna od geografske dolžine. Zdaj želimo napovedati količino padavin s hkratnim upoštevanjem nadmorske višine `z.nv` in geografske dolžine `x.gdol`.

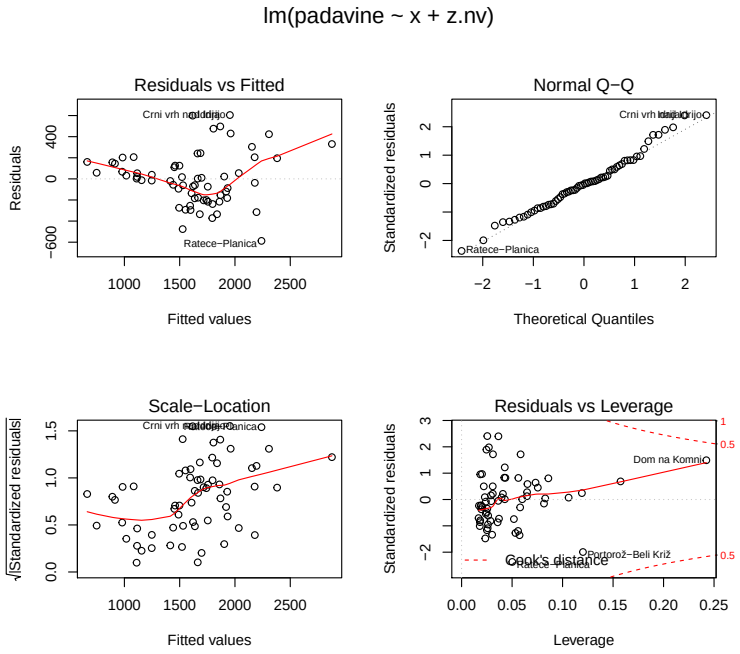
Slika 17 kaže odvisnost padavin od geografske dolžine in od nadmorske višine.



Slika 17: `padavine` v odvisnosti od `z.nv` in od `x`

Pri napovedovanju količine padavin je interakcija med nadmorsko višino in geografsko dolžino v Sloveniji pričakovana. Vsebinsko interakcija v tem primeru pomeni, da se količina padavin z nadmorsko višino spreminja drugače na zahodu kot na vzhodu Slovenije. Modeliramo postopno, najprej naredimo model brez interakcijskega člena in analiziramo ostanke (`model.m1`).

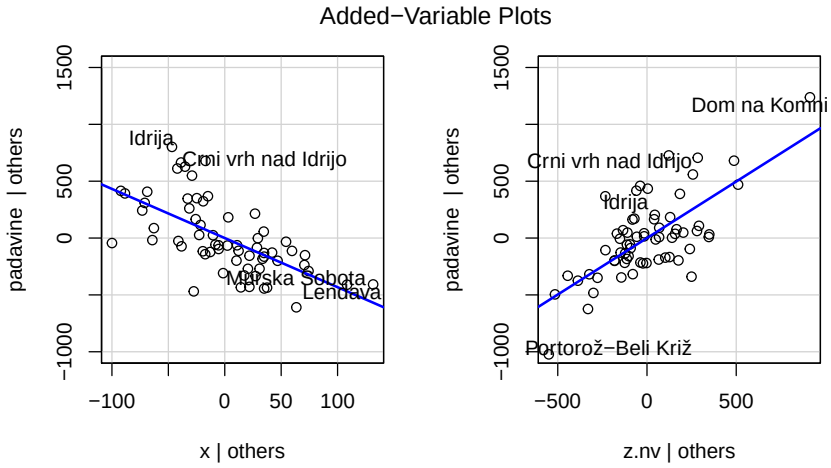
```
> model.m1<-lm(padavine~ x + z.nv, data=postaje)
```



Slika 18: Ostanki za `model.m1`

V modelu je prisotna nekonstantna varianca, ki je razvidna iz obej levih grafikonov.

```
> avPlots(model.m1, ylim=c(-1000,1500), id=list(location="avoid"))
```

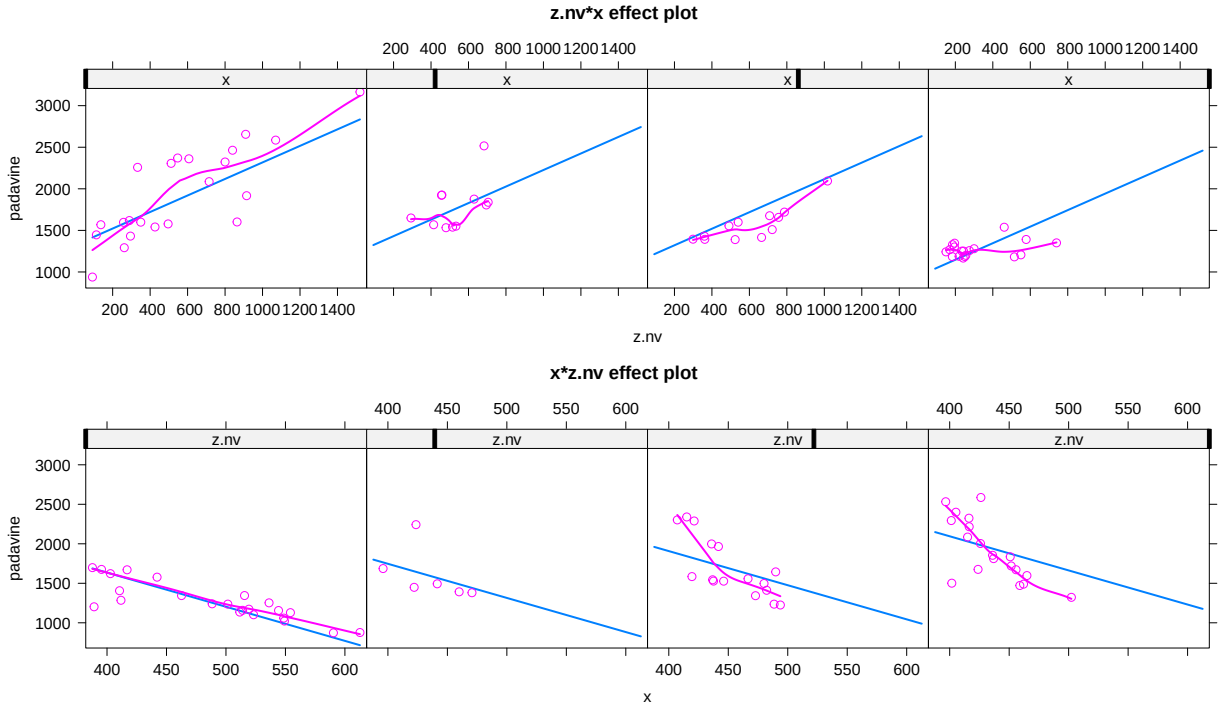


Slika 19: Grafa dodane spremenljivke za `model.m1`

Iz Slike 19, levo, je razvidna nekonstantna varianca glede na spremenljivko `x`, desno pa glede

na spremnljivko `z.nv`. Ko gremo od zahoda proti vzhodu, se variabilnost manjša in ko se povečuje nadmorska višina, se variabilnost veča. Vplivnih točk ni. Slika 20 prikazuje napovedi za `model.m1` in parcialne ostanke z gladilnikom pri različnih vrednostih druge spremenljivke v modelu. Tak grafični prikaz omogoča identifikacijo interakcije med dvema številskim spremenljivkama. Premice, ki predstavljajo napovedi modela, so na vseh razdelkih grafikona vzporedne. Če se gladilniki parcialnih ostankov prilegajo tem premicam, to pomeni, da ni interakcije med spremenljivkama, če pa je naklon gladilnikov v razdelkih različen, to nakazuje interakcijo. Na Sliki 20 gladilniki nakazujejo, da se količina padavin z nadmorsko višino spreminja hitreje na zahodu kot na vzhodu, kar pomeni, da je med `x` in `z.nv` prisotna interakcija.

```
> library(effects)
> graf1 <- plot(Effect(c("z.nv", "x"), model.m1, partial.residuals=TRUE),
+               ci.style="none", lattice=list(layout=c(4, 1)))
> graf2 <- plot(Effect(c("x","z.nv"), model.m1, partial.residuals=TRUE),
+               ci.style="none",lattice=list(layout=c(4, 1)))
> library(gridExtra)
> grid.arrange(graf1, graf2)
```



Slika 20: Parcialni ostanki z gladilnikom (roza črte) za `model.m1` v odvisnosti od nadmorske višine pri izbranih vrednostih geografske dolžine (zgoraj) in v odvisnosti od geografske dolžine pri izbranih vrednostih nadmorske višine (spodaj), prikazane so tudi napovedane vrednosti za `padavine` (modre črte)

V model dodamo še interakcijski člen med nadmorsko višino in geografsko dolžino in izvedimo diagnostiko modela.

```
> model.m2<-lm(padavine~ z.nv + x + z.nv:x , data=postaje)
> # model.m2<-lm(padavine~ z.nv * x  , data=postaje) # krajši zapis
```

Modelska matrika za `model.m2` ima v zadnjem stolpcu produkte vrednosti `z.nv` in `x`.

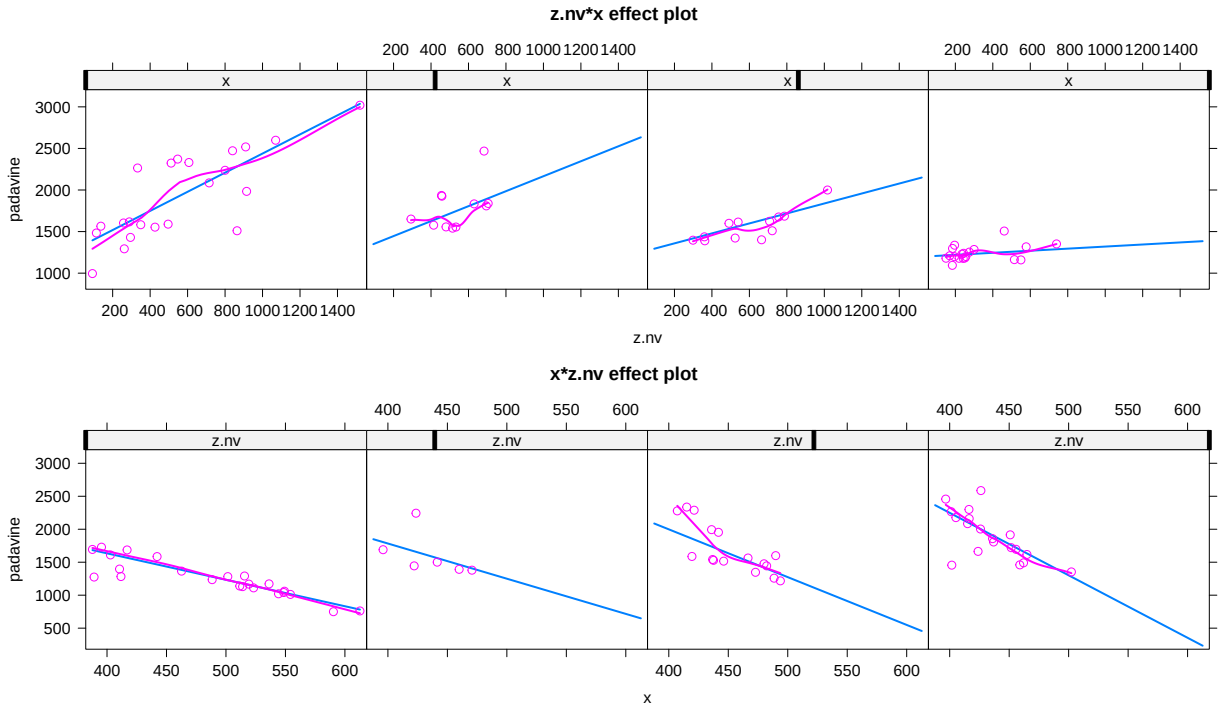
```
> X.m2<-model.matrix(model.m2)
> head(X.m2, n=3)
```

```
                    (Intercept) z.nv       x    z.nv:x
Babno polje                   1  756 464.930 351487.08
Bizeljsko                     1  170 554.193  94212.81
Brezovica pri Topolu          1  708 451.721 319818.47
```

V modelu so ocenjeni trije parametri:

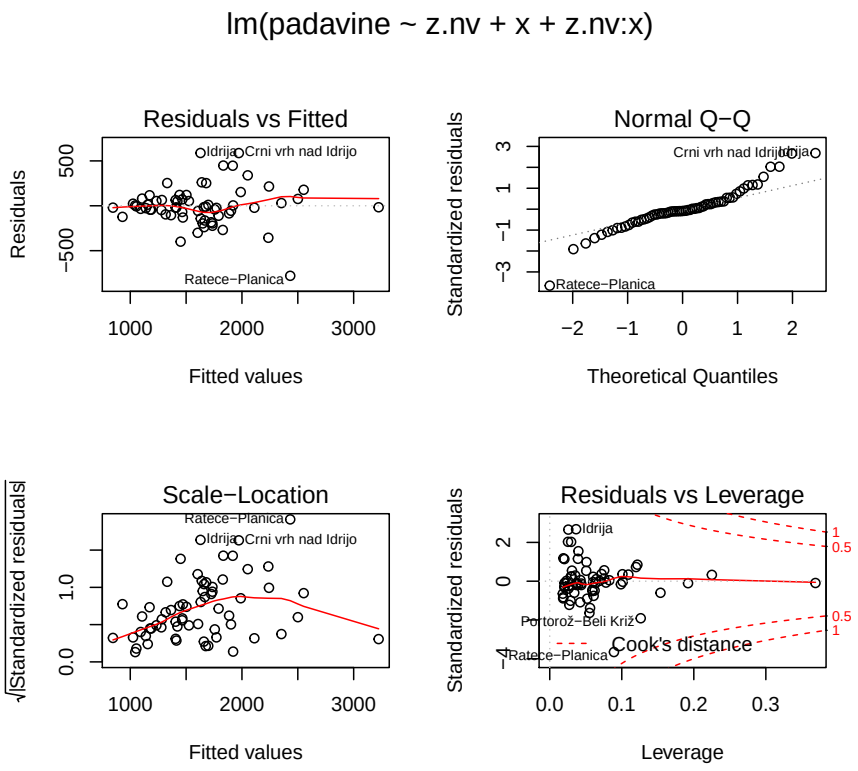
```
> model.m2$coeff
```

```
  (Intercept)          z.nv             x          z.nv:x
1736.34007572    6.05218444   -1.07842493   -0.01181133
```



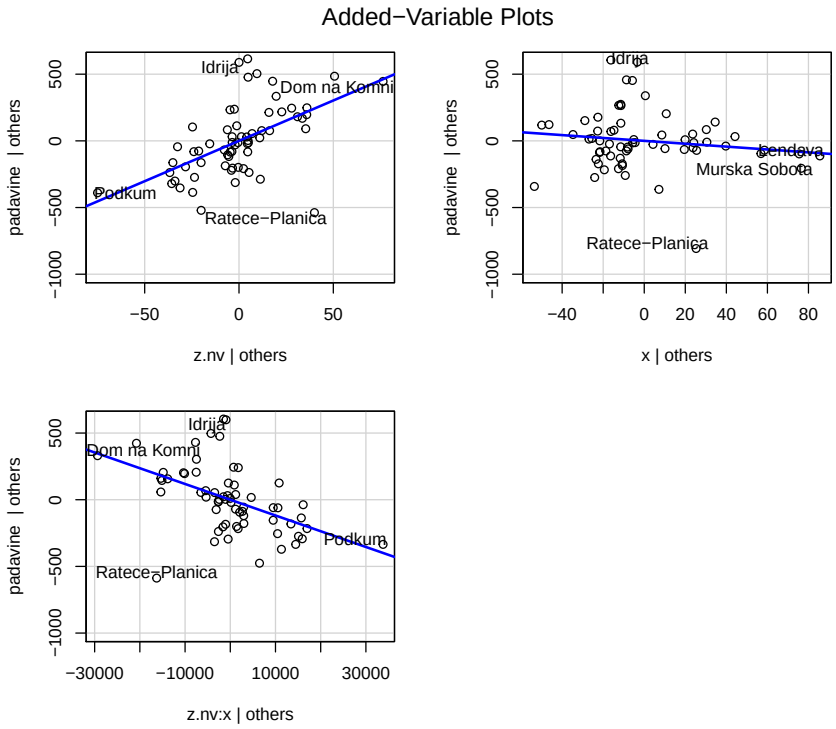
Slika 21: Parcialni ostanki z gladilnikom (roza črte) za `model.m2` v odvisnosti od nadmorske višine pri izbranih vrednostih geografske dolžine (zgoraj) in v odvisnosti od geografske dolžine pri izbranih vrednostih nadmorske višine (spodaj), prikazane so tudi napovedane vrednosti za `padavine` (modre črte)

Grafični prikaz parcialnih ostankov na Sliki 21 kaže, da je interakcija v `model.m2` ustrezno opisana, saj se gladilniki parcialnih ostankov dobro prilegajo modelskim napovedim.



Slika 22: Ostanki za `model.m2`

```
> avPlots (model.m2, ylim=c(-1000,600), id=list(location="avoid"))
```



Slika 23: Grafi dodane spremenljivke za `model.m2`

```
> outlierTest(model.m2)

               rstudent unadjusted p-value Bonferroni p
Rateče-Planica -4.111354         0.00012336    0.0078951
```

Sliki 22 in 23 še vedno kažeta nekonstantno varianco, ki jo bomo v končni fazi modelirali z modelom, kjer parametre ocenjujemo po posplošeni metodi najmanjših kvadratov (GLS). Kljub neizpolnjevanju ene od predpostavk linearnega modela, izpišimo povzetek modela z namenom, da se naučimo interpretirati ocene parametrov modela z dvema številskima napovednima spremenljivkama in njuno interakcijo. Izognili se bomo interpretaciji intervalov zaupanja, ker so ti zaradi nekonstantne variance v modelu pristranski.

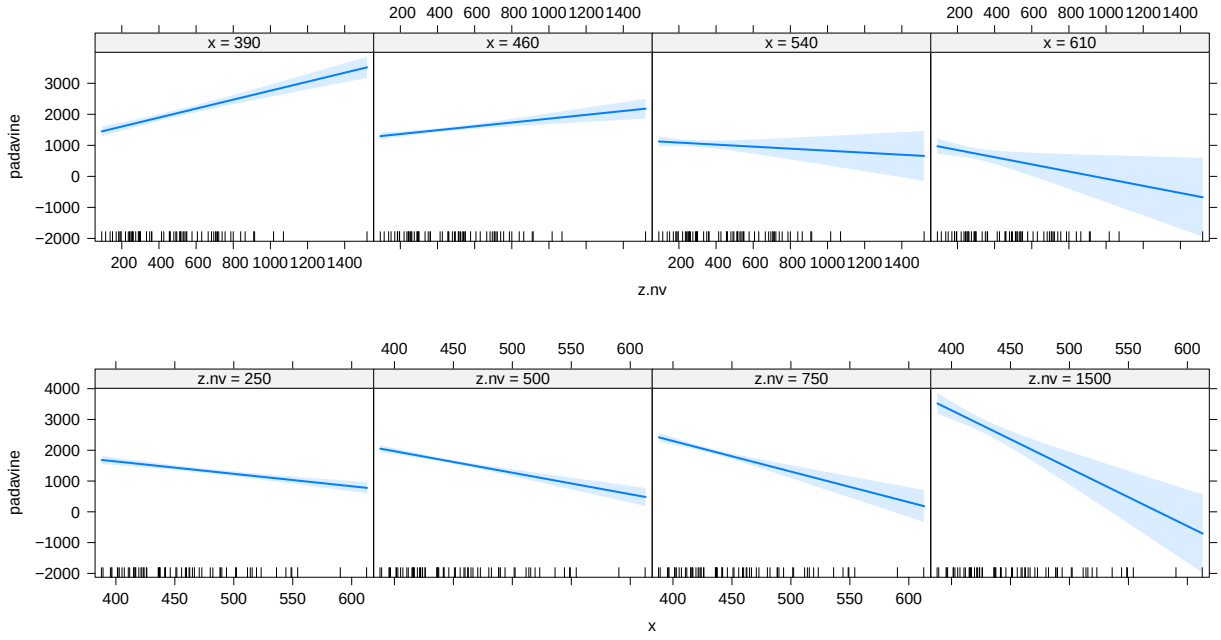
```
> summary(model.m2)$coeff

                 Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) 1736.34007572 4.376140e+02  3.967744 1.962836e-04
z.nv           6.05218444 1.165574e+00  5.192448 2.604035e-06
x             -1.07842493 9.540633e-01 -1.130349 2.628274e-01
z.nv:x        -0.01181133 2.709062e-03 -4.359934 5.185183e-05
```

```
> summary(model.m2)$r.squared

[1] 0.8001414
```

Če bi v modelu `model.m2` ne bila prisotna heteroskedastičnost, bi bili sklepi naslednji:

- `model.m2` pojasni 80 % variabilnosti za količino padavin;
- postaja Rateče-Planica je regresijski osamelec, modelska napoved močno preceni izmerjeno količino padavin;
- pri različnih geografskih dolžinah je vpliv nadmorske višine na količino padavin različen, interakcija `x:z.nv` je negativna in statistično značilna ($p < 0.0001$). To pomeni, da se vpliv nadmorske višine na padavine zmanjšuje od zahoda proti vzhodu (Slika 21 zgoraj); vpliv geografske dolžine na količino padavin se zmanjšuje z večjo nadmorsko višino (Slika 24 spodaj).

```
> plot(predictorEffects(model.m2, ~.,
+                       xlevels=list(x=4, z.nv=c(250, 500, 750, 1500))),
+      rows=2, cols=1, main="", layout=c(4,1))
```



Slika 24: Napovedane vrednosti za `padavine` za `model.m2`; v odvisnosti od nadmorske višine pri samodejno izbranih vrednostih geografske dolžine (zgoraj) in v odvisnosti od geografske dolžine pri izbranih vrednostih nadmorske višine (spodaj)

- če se premaknemo za 50 km proti vzhodu in ostanemo na isti nadmorski višini, se

količina padavin v povprečju spremeni za:

$$\hat{y}(x_0+50|z.nv) - \hat{y}(x_0|z.nv) =$$
$$= (b_0 + b_1 \cdot z.nv + b_2 \cdot (x_0+50) + b_3 \cdot (x_0+50) \cdot z.nv) - (b_0 + b_1 \cdot z.nv + b_2 \cdot x_0 + b_3 \cdot x_0 \cdot z.nv) =$$
$$= b_2 \cdot 50 + b_3 \cdot 50 \cdot z.nv$$

Izračunajmo to spremembo količine padavin pri nadmorskih višinah 100 m, 200 m, ..., 500 m.

```
    z razlika
1 100  -113.0
2 200  -172.0
3 300  -231.1
4 400  -290.1
5 500  -349.2
```

Poglejmo si še rezultat sekvenčnih F-testov, ki jih dobimo z ukazom `anova()`.

```
> anova(model.m2)

Analysis of Variance Table

Response: padavine
          Df  Sum Sq Mean Sq F value    Pr(>F)
z.nv       1 8446324 8446324 168.907 < 2.2e-16 ***
x          1 2615133 2615133  52.297 1.007e-09 ***
z.nv:x     1  950563  950563  19.009 5.185e-05 ***
Residuals 60 3000352   50006
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V prvi vrsti zgornjega izpisa se z F-testom testira domneva $H_0 : \beta_1 = 0$. Z drugimi besedami povedano se ničelna domneva glasi: modela $y_i = \beta_0$ in $y_i = \beta_0 + \beta_1 \dot{z}.nv_i + \varepsilon_i$ sta enakovredna. Ničelno domnevo zavrnemo ($p < 0.0001$).

V drugi vrsti z F-testom primerjamo modela $y_i = \beta_0 + \beta_1 \dot{z}.nv_i$ in $y_i = \beta_0 + \beta_1 \dot{z}.nv_i + \beta_2 \dot{x}_i$, testiramo ničelno domnevo $H_0 : \beta_2 = 0$, tudi to ničelno domnevo zavrnemo ($p < 0.0001$). Ob upoštevanju nadmorske višine ima geografska dolžina statistično značilen vpliv na količino padavin.

V zadnji vrsti izpisa testiramo ničelno domnevo H_0: ni interakcije med x in $z.nv$. Tudi to ničelno domnevo zavrnemo ($p < 0.0001$).

1.1.5 Trasformacije, ki zmanjšajo vplivnost točk

Vemo, da je večja vplivnost točk lahko odvisna od kršenja predpostavk linearnega modela. Vplivne točke se lahko pojavijo zaradi nekonstantne variance, nelinearne zveze med odzivno spremenljivko in regresorji ali pa zaradi nenavadne vrednosti regresorja.

Če so vrednosti regresorja zelo asimetrično porazdeljene oziroma, če neenakomerno pokrivajo regresorski prostor, to povzroči prisotnost vzvodnih točk v modelu. V takem primeru lahko ustrezna transformacija regresorja zmanjša velike vzvode točk in s tem tudi njihovo vplivnost.

V splošnem lahko s transformacijo odzivne spremenljivke ali regresorjev dosežemo zmanjšanje ali pa tudi povečanje vplivnosti posameznih točk v modelu.

Primer: `mammals`

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)
> 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
```

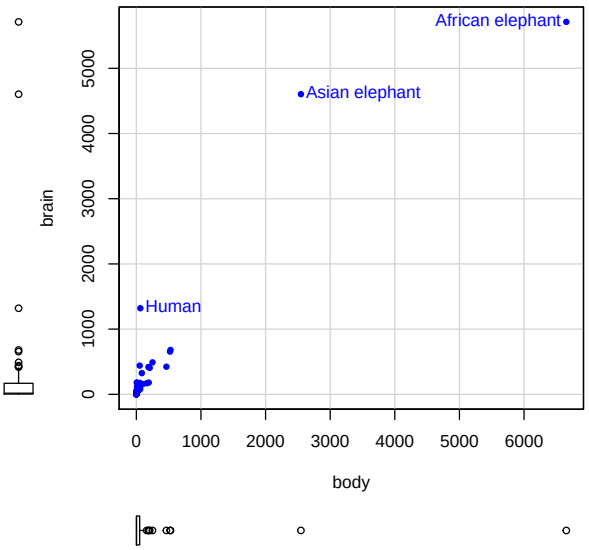
Tega primera se bomo lotili po naslednjih korakih:

a) Grafično prikažite odvisnosti `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?

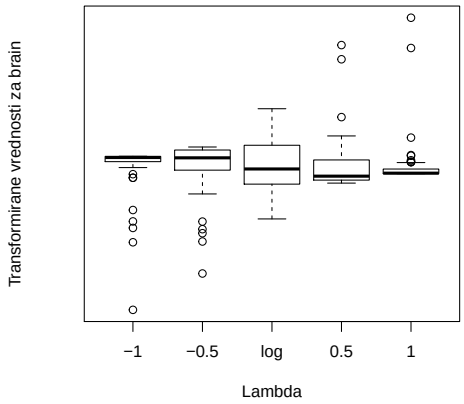
d) Analizirajte ustrezni model in obrazložite rezultate.



Slika 25: Odvisnost `brain` od `body` za 62 sesalcev

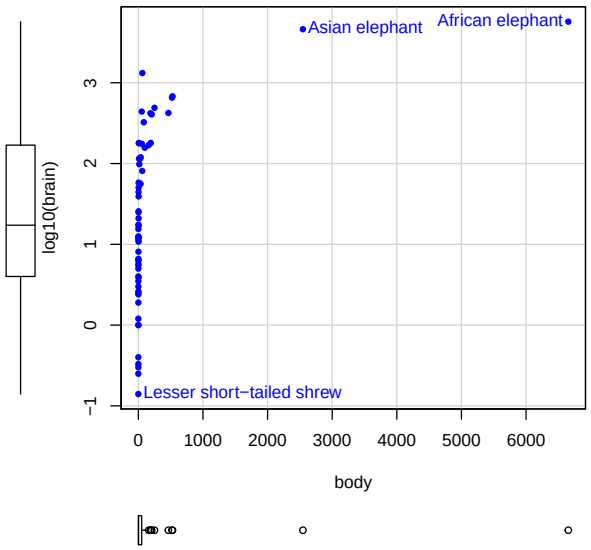
Iz slike vidimo, da sta porazdelitvi za `body` in za `brain` zelo asimetrični. Večina točk je v levem spodnjem kotu slike. Poskusimo dobiti ustrezno transfomacijo za `brain`, da bi bila porazdelitev bolj simetrična.

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

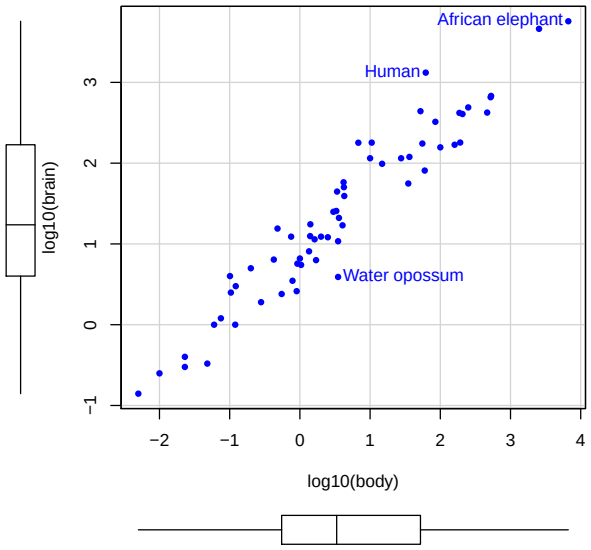


Slika 26: Okviri z ročaji za različne transformacije za spremenljivko `brain`

Na Sliki 26 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 vsebinsko interpretacijo logaritmiranih vrednosti na sliki je tokrat smiselno uporabiti desetiški logaritem. Slika 27 kaže odvisnost `log10(brain)` od `body`. Poskusimo doseči linearnost z uporabo logaritemske transfomacije tudi za `body` (Slika 28).



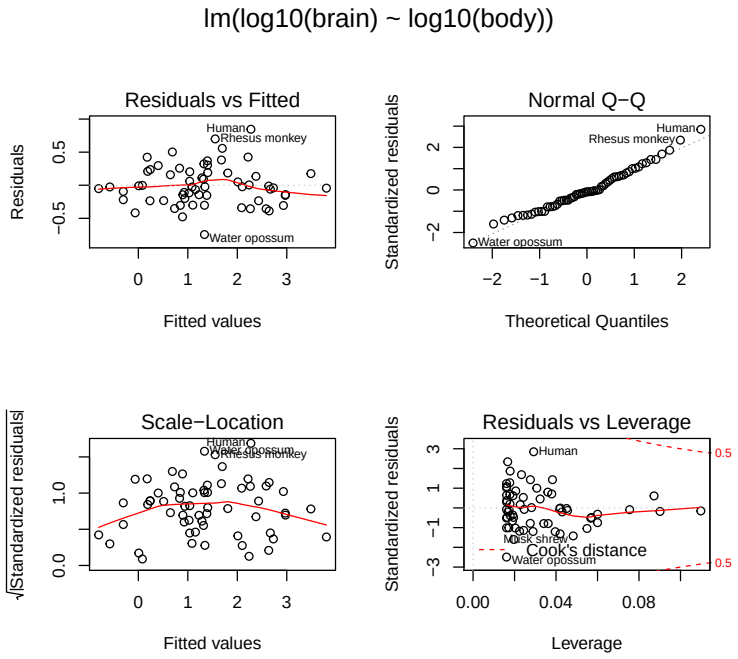
Slika 27: Odvisnost `log10(brain)` od `body` za 62 sesalcev



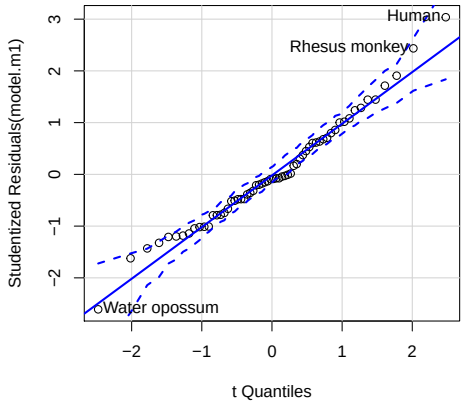
Slika 28: Odvisnost `log10(brain)` od `log10(body)` za 62 sesalcev

Slika 28 kaže, da je odvisnost med logaritmiranima spremenljivkama linearna, kar pomeni, da gre za multiplikativni model.

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



Slika 29: Grafični prikaz ostankov za `model.m1`

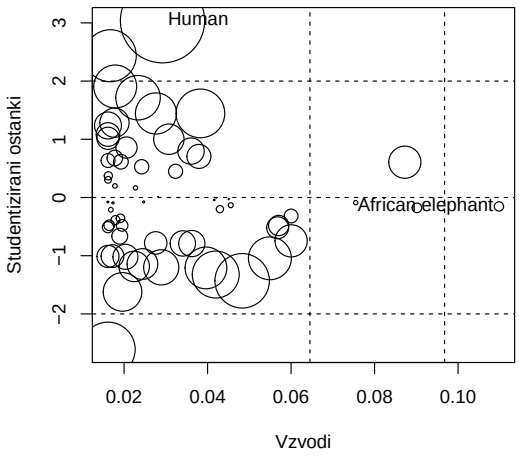


Slika 30: QQ grafikon za studentizirane ostanke za `model.m1` s 95 % bootstrap ovojnico

Sliki 29 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 30 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
```

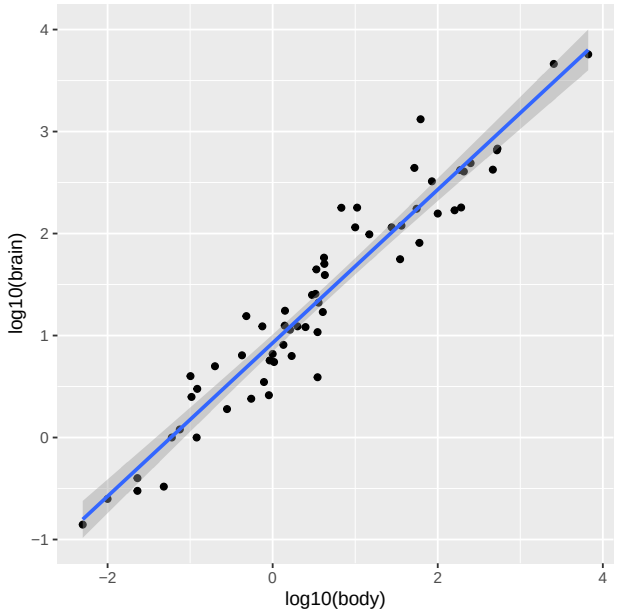


Slika 31: Grafični prikaz studentiziranih ostankov, vzvodov in Cookove razdalje za `model.m1`

```
> summary(model.m1)$coeff

             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.9271269 0.04171111 22.22734 1.183207e-30
log10(body) 0.7516859 0.02846356 26.40871 9.835792e-35
```

```
> summary(model.m1)$r.squared
```

```
[1] 0.9207837

> confint(model.m1)

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



Slika 32: 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`.

1.2 Metoda tehtanih najmanjših kvadratov

Predpostavimo, da poznamo variance napak ε_i $Var(\varepsilon_i) = \sigma_i^2$, ali, da poznamo variance napak izražene z znanimi utežmi w_i, $Var(\varepsilon_i) = \sigma^2 w_i$, $i = 1, ..., n$, kjer je σ^2 neznana. V tem primeru parametere linearnega modela ocenimo po metodi **tehtanih najmanjših kvadratov** (WLS, *Weighted Least Squares*). Uteži za variance napak morajo biti pozitivne, zapišemo jih v diagonalno matriko $\mathbf{W}$. Ocene parametrov modela po metodi tehtanih najmanjših kvadratov dobimo z minimiranjem izraza

$$S(\boldsymbol{\beta}, \mathbf{W}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ii}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2. \tag{15}$$

Večja utež W_{ii} pomeni, da ima i-ti podatek večji vpliv na oceno parametrov modela. Poglejmo, kakšna je povezava med utežmi in varianco napak v linearnem modelu

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kjer za napake predpostavimo, da so porazdeljene normalno, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{V})$. $\mathbf{V}$ je diagonalna **variančna matrika napak** dimenzije $n \times n$. Če so vrednosti na diagonali različne, imamo opravka z nekonstantno varianco napak.

Za oceno parametrov v tem primeru uporabimo metoda največjega verjetja. Funkcijo verjetja v tem primeru zapišemo:

$$L(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi det\mathbf{V}^{\frac{n}{2}}} exp\left(\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right), \tag{16}$$

Zaenkrat predpostavimo, da je $\mathbf{V}$ znana. Maksimiranje zgornjega izraza je enakovredno minimiranju **posplošene vsote kvadratov napak**:

$$\sum_{i=1}^{n} \mathbf{V}_{ii}^{-1}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{17}$$

Če primerjamo izraz, ki smo ga dobili po metodi tehtanih najmanjših kvadratov (15) in izraz pri posplošeni metodi najmanjših kvadratov (17), vidimo, da je $\mathbf{W} = \mathbf{V}^{-1}$. Utež za posamezen podatek je torej obratno sorazmerna z njegovo varianco, kar pomeni, da damo podatku z večjo varianco manj pomembnosti pri ocenjevanju parametrov modela.

Če poznamo variance σ_i^2 ali uteži w_i ali če podatki omogočajo, da variance oziroma uteži ocenimo, je ocenjevanje parametrov z WLS primernejše kot transformacija podatkov. Podatki ostanejo v osnovnih enotah, kar omogoča lažjo interpretacijo dobljenega modela.

V posameznih primerih so uteži lahko določene tudi na podlagi vrednosti izbranih napovednih

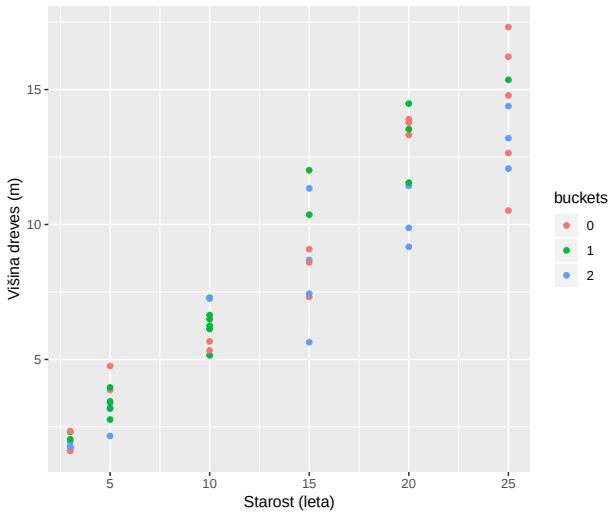
spremenljivk (ene ali več). V primerjavi z OLS ocenami parametrov imajo WLS ocene parametrov v splošnem manjšo varianco.

Primer: `ANDY`

Drevesničar Andy je želel ugotoviti, kako namakanje vpliva na višino dreves ob upoštevanju njihove starosti. Naredil je večletni poskus, v katerem je drevesa v poletnem času dnevno namakal s tremi različnimi količinami vode (0, 1 in 2 vedra vode). Ob koncu poskusa je za vsako drevo zabeležil starost, višino in količino namakanja. Za vsako starost je imel v poskus vključenih več dreves. Podatki so v podatkovnem okviru ANDY.txt, višina dreves (`height`) je izražena v čevljih, starost (`age`) je v letih in namakanje (`buckets`) v številu veder vode.

```
> andy<-read.table("ANDY.txt", header=T)
> str(andy)

'data.frame':        54 obs. of  3 variables:
 $ height : num  5.6 12.7 23.9 28.5 45.6 34.5 5.3 11.1 16.9 18.5 ...
 $ age    : int  3 5 10 15 20 25 3 5 10 15 ...
 $ buckets: int  2 0 1 2 0 0 0 2 1 2 ...

> andy$height<-andy$height/3.2808 # višino dreves izrazimo v metrih
> andy$buckets<-factor(andy$buckets) # buckets naj bo opisna spremenljivka, ker je za

> summary(andy)

     height            age       buckets
 Min.   : 1.615   Min.   : 3.0   0:18
 1st Qu.: 3.399   1st Qu.: 5.0   1:18
 Median : 7.270   Median :12.5   2:18
 Mean   : 7.816   Mean   :13.0
 3rd Qu.:11.895   3rd Qu.:20.0
 Max.   :17.313   Max.   :25.0
```

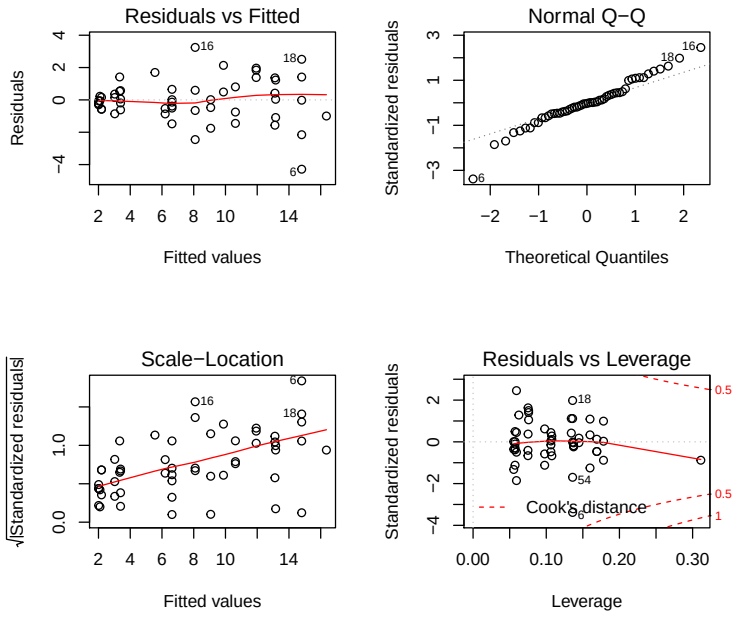
Tokrat bomo dali v model eno opisno in eno številsko spremenljivko ter njuno interakcijo. Zanima nas, ali količina namakanja vpliva na rast dreves. Primerjali bomo odvisnost višine dreves od starosti pri treh količinah namakanja. Več o opisni spremenljivki v linearnem modelu bomo povedali drugič.



Slika 33: `height` v odvisnosti od `age` in `buckets`

Slika 33 kaže odvisnost `height` od `age` za tri različne količine namakanja (0, 1, 2 vedra vode dnevno). Na sliki se vidi, da se variabilnost podatkov s starostjo dreves povečuje, kar pomeni, da predpostavka o konstantni varianci ni izpolnjena.

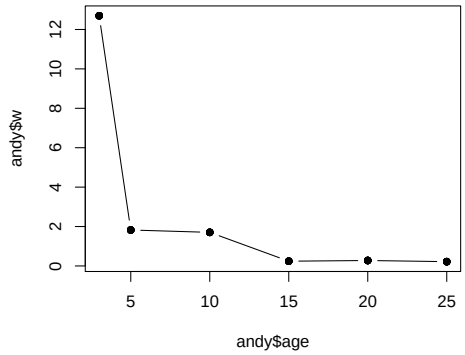
```
> mod.OLS<-lm(height ~ age * buckets, data=andy)
> anova(mod.OLS)

Analysis of Variance Table

Response: height
            Df  Sum Sq Mean Sq  F value  Pr(>F)
age          1 1050.60 1050.60 565.1371 < 2e-16 ***
buckets      2   16.51    8.25   4.4395 0.01702 *
age:buckets  2    9.34    4.67   2.5131 0.09163 .
Residuals   48   89.23    1.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Slika 34: Ostanki za `mod.OLS`

Uporabili bomo metodo tehtanih najmanjših kvadratov. Ker imamo več podatkov pri isti starosti, lahko določimo uteži na podlagi varianc za `height` pri posameznih starostih:

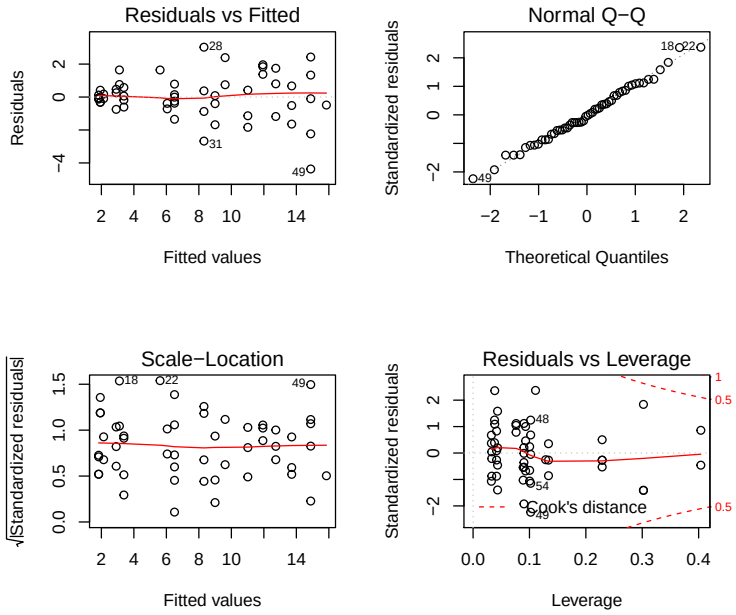
```
> library(dplyr)
> andy <- as_tibble(andy)
> utez <- andy %>%
+   group_by(age) %>%
+   summarise(w=1/var(height))
> andy <- merge(andy, utez, by="age")
> head(andy)

  age   height buckets        w
1   3 1.706901       2 12.69631
2   3 2.316508       1 12.69631
3   3 1.615460       0 12.69631
4   3 2.346989       0 12.69631
5   3 1.767861       2 12.69631
6   3 2.042185       1 12.69631
```



Slika 35: Uteži, kot obratne vrednosti variance `height` enako starih dreves v odvisnosti od `age`

Metodo tehtanih najmanjših kvadratov izvedemo s funkcijo `lm()` in argumentom `weights`:

```
> mod.WLS<-lm(height ~ age * buckets, weights = w, data=andy)
```

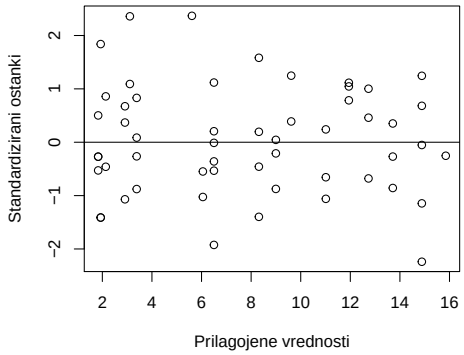


Slika 36: Ostanki za `mod.WLS`

Pri modeliranju variance smo malo spremenili ostanke, veliko bolj pa standardizirane ostanke,

kar se kaže na Sliki 36 na grafu desno zgoraj in levo spodaj. Na grafu levo zgoraj je še vedno videti nekonstantno varianco v ostankih. Če hočemo grafično prikazati standardizirane ostanke v odvisnosti od prilagojenih vrednosti, moramo to narediti peš (Slika 37).

```
> plot(fitted(mod.WLS), rstandard(mod.WLS),
+      xlab="Prilagojene vrednosti",ylab="Standardizirani ostanki")
> abline(h=0)
```



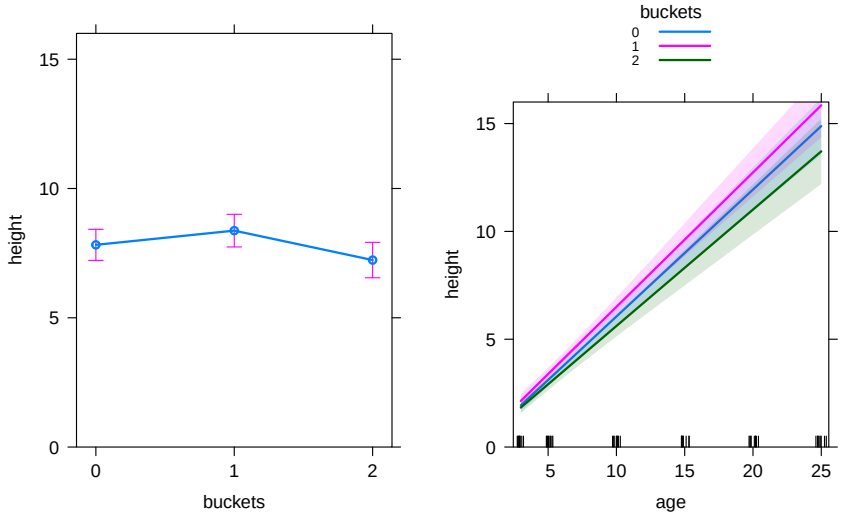
Slika 37: Standardizirani ostanki za `mod.WLS`

```
> anova(mod.WLS)

Analysis of Variance Table

Response: height
            Df Sum Sq Mean Sq  F value Pr(>F)
age          1 829.34  829.34 899.9925 <2e-16 ***
buckets      2   5.12    2.56   2.7792 0.0721 .
age:buckets  2   2.46    1.23   1.3367 0.2723
Residuals   48  44.23    0.92
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V primerjavi z `mod.OLS` vidimo, da se rezultati sekvenčnega testiranja domnev s funkcijo `anova()` spremenijo. Ko upoštevamo uteži, ki so obratno sorazmerne z varianco `height` pri posamezni vrednosti `age`, količina namakanja `buckets` ob upoštevanju `age` nima več statistično značilnega vpliva na `height`.

```
> summary(mod.WLS)

Call:
lm(formula = height ~ age * buckets, data = andy, weights = w)
```

```
Weighted Residuals:
     Min       1Q   Median       3Q      Max
-2.03668 -0.57355 -0.02891  0.63673  2.22073

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.16697    0.20161   0.828    0.412
age           0.58869    0.03158  18.644   <2e-16 ***
buckets1      0.10112    0.32097   0.315    0.754
buckets2      0.04469    0.27642   0.162    0.872
age:buckets1  0.03454    0.04875   0.709    0.482
age:buckets2 -0.04863    0.04747  -1.025    0.311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9599 on 48 degrees of freedom
Multiple R-squared:  0.9498,        Adjusted R-squared:  0.9446
F-statistic: 181.6 on 5 and 48 DF,  p-value: < 2.2e-16
```



Slika 38: Povprečne napovedi `height` glede na `buckets` pri povprečni vrednosti `age` s pripadajočimi 95% intervali zaupanja (levo) in povprečne napovedi `height` glede na `age` pri posamezni vrednosti `buckets` za `mod.WLS`

Analiza pokaže, da je vpliv namakanja na višino dreves ob upoštevanju starosti zanemarljiv. Nakloni premic na Sliki 38 niso statistično značilno različni. Z modelom je pojasnjene 95 % variabilnosti višine dreves.

Vaja: analizo vpliva namakanja na višino dreves ob upoštevanju starosti ponovite z uporabo ustrezne transformacije podatkov.

2 VAJE

2.1 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.