Kazalo

1 DIAGNOSTIKA LINEARNEGA MODELA Z VEČ REGRESORJI

1.1 Graf dodane spremenljivke

Če imamo v modelu več številskih napovednih spremenljivk, **robni razsevni grafikon** odzivne spremenljivke glede na posamezno napovedno spremenljivko ne prikaže nujno pravega vpliva te spremenljivke na odzivno spremenljivko, saj ne upošteva vpliva ostalih spremenljivk v modelu. Za grafični prikaz vpliva posamezne spremenljivke na odzivno spremenljivko ob upoštevanju ostalih spremenljivk v modelu uporabimo t. i. **graf dodane spremenljivke** (*added variable plots* ali *partial regression plots*), ki ga naredi funkcija `avPlot` iz paketa `car` (Slika 5).

Recimo, da je v linearnem modelu k napovednih spremenljivk x_j

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

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

$$y_i^{(-j)} = \beta_0^{(j)} + \beta_1^{(j)} x_{i1} + \cdots + \beta_{j-1}^{(j)} x_{i,j-1} + \beta_{j+1}^{(j)} x_{i,j+1} + \beta_k^{(j)} x_{ik} + \varepsilon_i, \quad i = 1, ..., n. \tag{2}$$

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

$$e_{i,y}^{(-j)} = y_i - \hat{y}_i^{(-j)}, \quad i = 1, ...n. \tag{3}$$

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

$$x_i^{(-j)} = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_{j-1} x_{i,j-1} + \gamma_{j+1} x_{i,j+1} + \gamma_k x_{ik} + \varepsilon_i, \quad i = 1, ..., n. \tag{4}$$

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

$$e_{i,x_j}^{(-j)} = x_{ij} - \hat{x}_i^{(-j)}, \quad i = 1, ...n. \tag{5}$$

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

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

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

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

1.2 Graf parcialnih ostankov

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

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

$$e_{i,x_j} = e_i + b_j x_{ij}. \tag{6}$$

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

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

1.3 Primer: `pacienti`

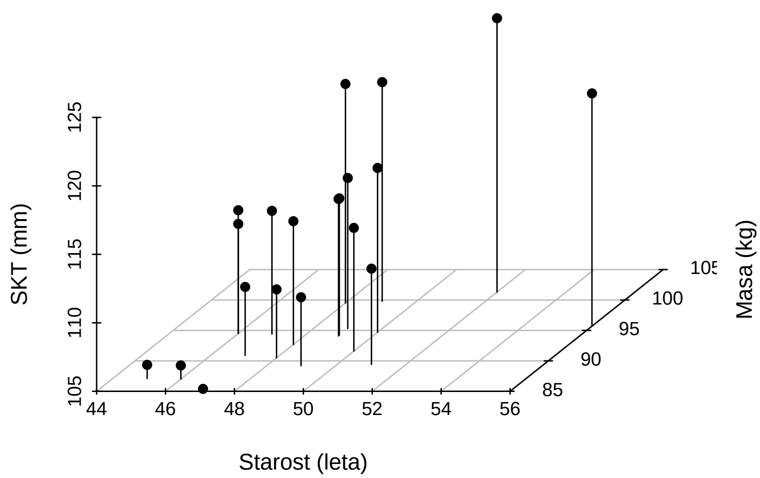
Imamo podatke za 20 moških s povišanim krvnim tlakom: krvni tlak (`SKT`, mm Hg), starost (`starost`, leta), telesna masa (`masa`, kg). Podatki so v datoteki `PACIENTI.txt`.

```
> summary(pacienti)
```

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

Zanima nas, kako je `SKT` odvisen od `starost` in `masa` hkrati. Ker imamo dve številski napovedni spremenljivki, je smiselno podatke prikazati s tridimenzionalnim grafikonom. Za grafični prikaz podatkov v 3D uporabimo funkcijo `scatterplot3d` iz paketa `scatterplot3d`. Na navpični osi je odzivna spremenljivka `SKT`, na vodoravnih oseh pa napovedni spremenljivki `starost` in `masa`. Vsak pacient predstavlja točko v 3D prostoru, na sliki je zaradi boljše vidljivosti vsaki točki dodana navpičnica (Slika 1).

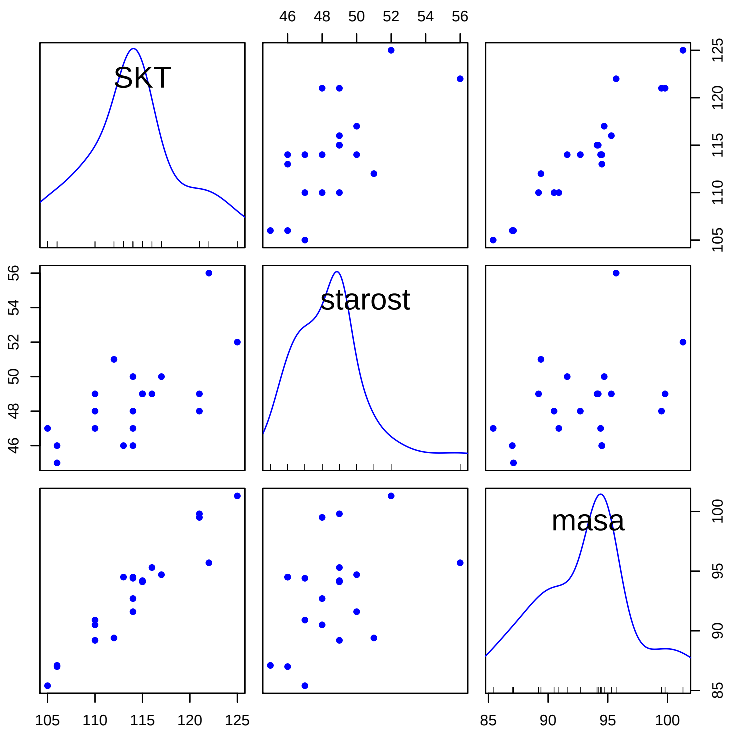
```
> library(scatterplot3d)
> scatterplot3d(pacienti$starost, pacienti$masa, pacienti$SKT, pch=16,
+               type="h", xlab="Starost (leta)", ylab="Masa (kg)",
+               zlab="SKT (mm)",  box=F)
```



Slika 1: Odvisnost `SKT` od `starost` in `masa`

Matrika razsevnih grafikonov prikazuje, v kakšni zvezi so analizirane spremenljivke. Uporabimo funkcijo `scatterplotMatrix()` iz paketa car. Je `SKT` linearno odvisen od `starost`? Je linearno odvisen od `masa`? Kakšna je povezava med `starost` in `masa`? Odgovori na ta vprašanja govorijo o robni odvisnosti SKT od napovednih spremenljivk, vsaka slika zase ne upošteva prisotnosti drugih napovednih spremenljivk.

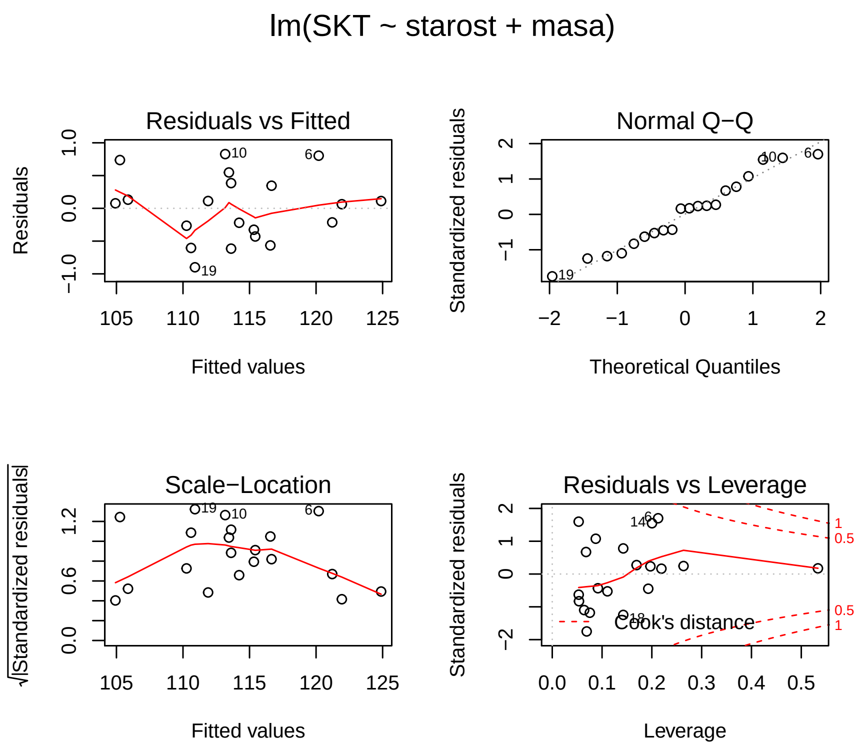
```
> library(car)
> scatterplotMatrix(~SKT + starost + masa, regLine=FALSE, smooth=FALSE,
+                    id=list(n=0), pch=16, data=pacienti)
```



Slika 2: Matrika razsevnih grafikonov za `SKT`, `starost` in `masa` za 20 pacientov, na diagonali so empirične gostote verjetnosti za analizirane spremenljivke

Naredimo linearni regresijski model za napovedovanje `SKT` od `starost` in `masa`. Osnovni diagnostični grafi na Sliki 3 kažejo, da so predpostavke linearnega modela dokaj dobro izpolnjene, ni vplivnih točk niti kandidatov za regresijske osamelce.

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

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

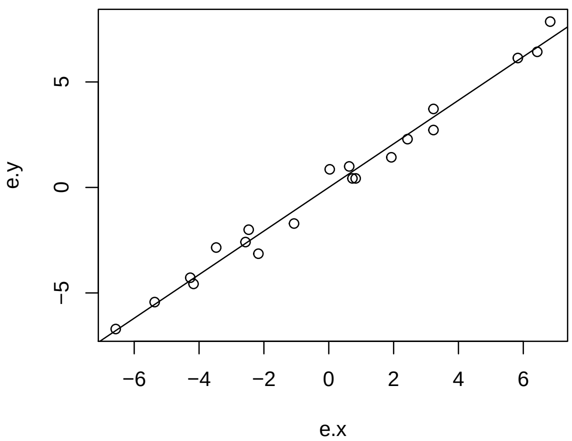


Slika 3: Ostanki za `model.p`

Če imamo v modelu več številskih napovednih spremenljivk, je poleg ostankov na Sliki 3 za prepoznavanje odstopanj od predpostavk linearnega modela informativno prikazati vpliv vsake od napovednih spremenljivk na odzivno spremenljivko ob upoštevanju ostalih spremenljivk v modelu. Za to naredimo grafe dodane spremenljivke. Za ilustracijo naredimo izračune in graf dodane spremenljivke za `masa`, ki prikazuje odvisnost `SKT` od `masa` ob upoštevanju `starost` v modelu:

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

[1] 1.032961

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

[1] 0.03027843
```

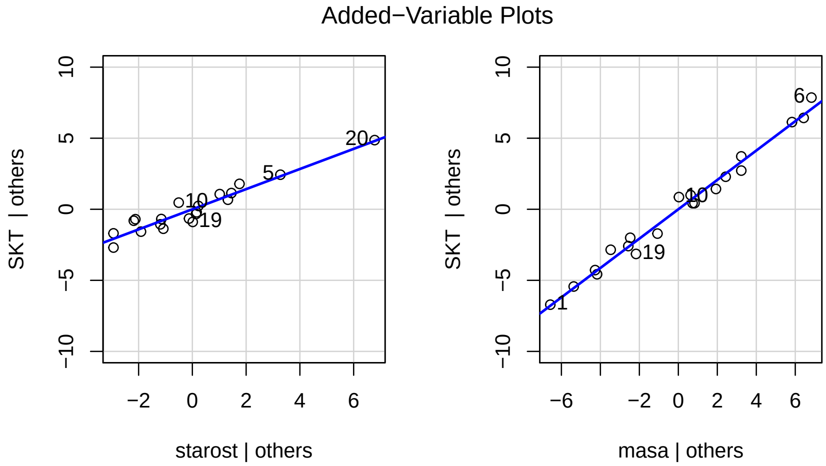
```
> plot(e.x, e.y)
> abline(reg=mod.e)
```



Slika 4: Graf dodane spremenljivke `masa`, odvisnost `SKT` od `masa` ob upoštevanju `starost`

Funkcija `avPlots` iz paketa `car` (Sliki 5) na podlagi `model.p` nariše grafikona dodane spremenljivke za `starost` in za `masa`. Po prednastavitvi velja `id=TRUE`, kar pomeni
`id=list(method=list(abs(residuals(mod.e, type="pearson")), "x"), n=2, cex=1, col=carPalette()[1` `location="lr")`.
Na grafih sta z vrednostjo `rownames()` označeni dve točki z največjim ostankom in dve točki z največjo vrednostjo na x-osi oziroma največjim parcialnim vzvodom. Premici na Sliki 5 se dobro prilegata točkam in ni videti prisotnosti nekonstantne variance, prav tako ni vplivnih točk.

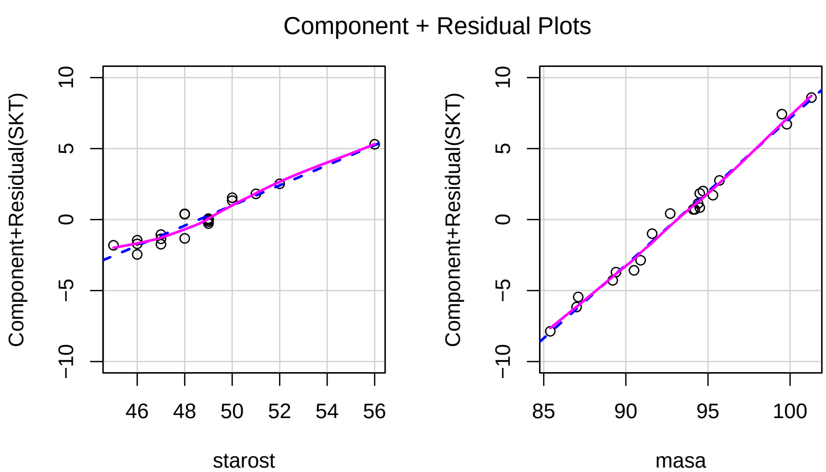
```
> avPlots(model.p, ylim=c(-10, 10))
```



Slika 5: Grafa dodane spremenljivke za `model.p`

Linearnost odvisnosti SKT od posamezne spremenljivke upoštevajoč drugo spremenljivko v modelu preverimo na podlagi grafikonov parcialnih ostankov, ki ga nariše funkcija `crPlot` (Slika 6). Gladilnik se dobro prilega premici, kar pomeni, da ni dodatne nelinearnosti v odvisnosti `SKT` od `starost` in `masa`.

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



Slika 6: Grafa parcialnih ostankov za `model.p`

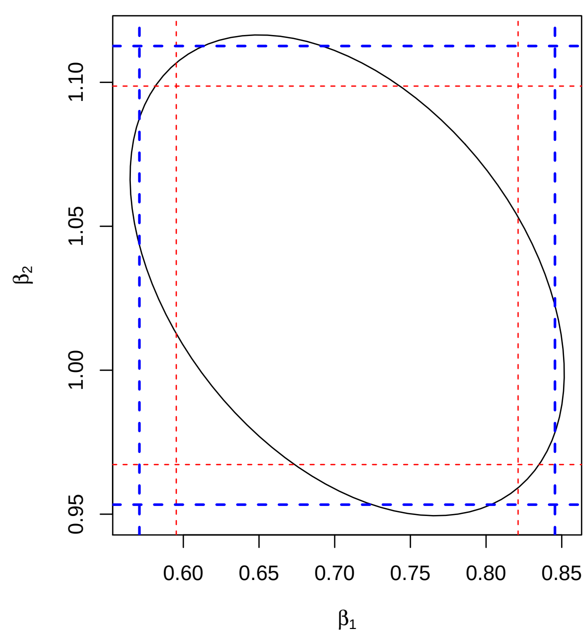
Izpišemo R^2 in ocene parametrov s pripadajočimi 95 % parcialnimi intervali zaupanja, nato pa še s hkratnimi intervali zaupanja za parametre modela.

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

[1] 0.9913858

> Confint(model.p)

               Estimate       2.5 %      97.5 %
(Intercept) -16.5793694 -22.9245526 -10.2341861
starost       0.7082515   0.5953468   0.8211561
masa          1.0329611   0.9672272   1.0986950

> # če uporabimo funkcijo Confint() iz paketa car namesto confint(),
> # se poleg IZ izpišejo tudi ocene parametrov
> library(multcomp)
> confint(glht(model.p))$confint

               Estimate         lwr        upr
(Intercept) -16.5793694 -24.2913344 -8.8674044
starost       0.7082515   0.5710266  0.8454763
masa          1.0329611   0.9530678  1.1128544
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 2.56428
```

Sklepi:

- `starost` in `masa` v `model.p` pojasnita 99 % variabilnosti `SKT`;
- obe napovedni spremenljivki `starost` in `masa` sta zelo statistično značilni ($p < 0.0001$);
- izberemo poljubno vrednost za maso na intervalu 85 kg do 102 kg. Pri izbrani vrednosti za maso velja: če se starost poveča za 10 let, se SKT v povprečju poveča za 7.1 mm, pripadajoč 95 % IZ je (5.7 mm, 8.5 mm);
- izberemo poljubno vrednost za starost na intervalu 45 let do 56 let. Pri izbrani vrednosti za starost velja: če se masa poveča za 10 kg, se SKT v povprečju poveča 10.3 mm, pripadajoč 95 % IZ je (9.5 mm, 11.1 mm).

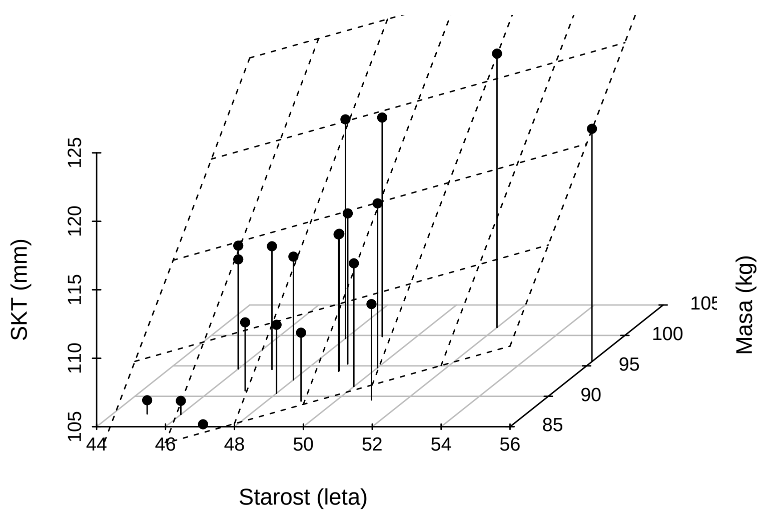
```
> library(ellipse)
> plot(ellipse(model.p, which=c(2,3)), type="l",
+      xlab=expression(beta[1]), ylab=expression(beta[2]))
> abline(v=confint(model.p)[2,], lty=2, col="red")
> abline(h=confint(model.p)[3,], lty=2, col="red")
> abline(v=confint(glht(model.p))$confint[2,2:3], lty=2, lwd=2, col="blue")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



Slika 7: 95 % območje zaupanje za parametra β_1 in β_2 modela `model.p` (elipsa) in meje 95 % parcialnih intervalov zaupanja za vsak parameter modela posebej (rdeče črtkane črte) in hkratnih intervalov zaupanja (modre črtkane črte)

Napovedane vrednosti za `SKT` dobljene z `model.p` ležijo na ravnini (Slika 8).

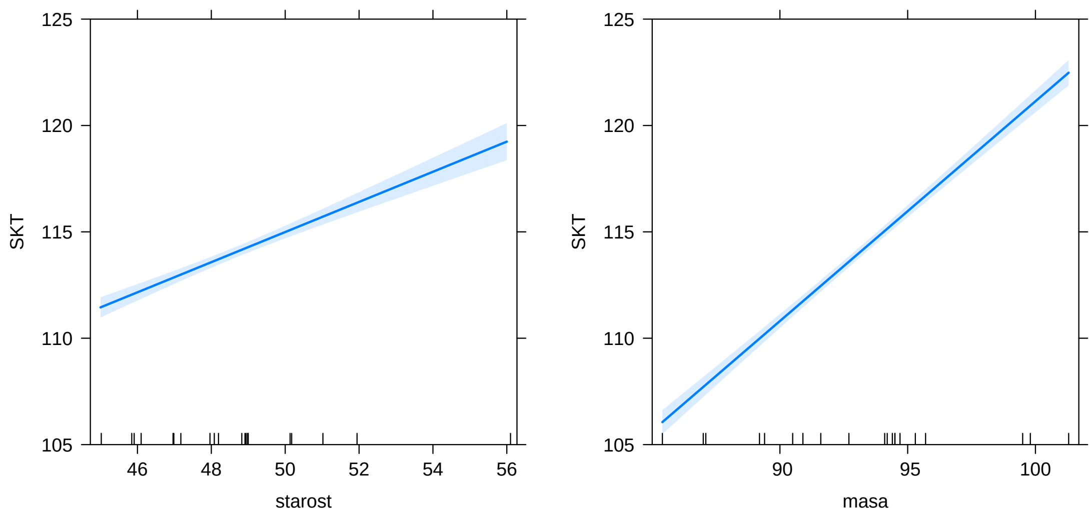
```
> s3d<-scatterplot3d(pacienti$starost, pacienti$masa, pacienti$SKT, pch=16,
+                type="h", xlab="Starost (leta)", ylab="Masa (kg)",
+                zlab="SKT (mm)",  box=F)
> s3d$plane(model.p) # slika ravnine
```



Slika 8: Odvisnost `SKT` od `starost` in `masa`, napovedi izračunane na osnovi `model.p` ležijo na ravnini

Narišimo napovedi in intervale zaupanja za povprečne napovedi s funkcijo `predictorEffects` iz paketa `effects`. Narisali bomo dve sliki, napovedi za `SKT` glede na `starost` pri povprečni masi 93.1 kg ter napovedi za `SKT` glede na `masa` pri povprečni starosti 48.6 let.

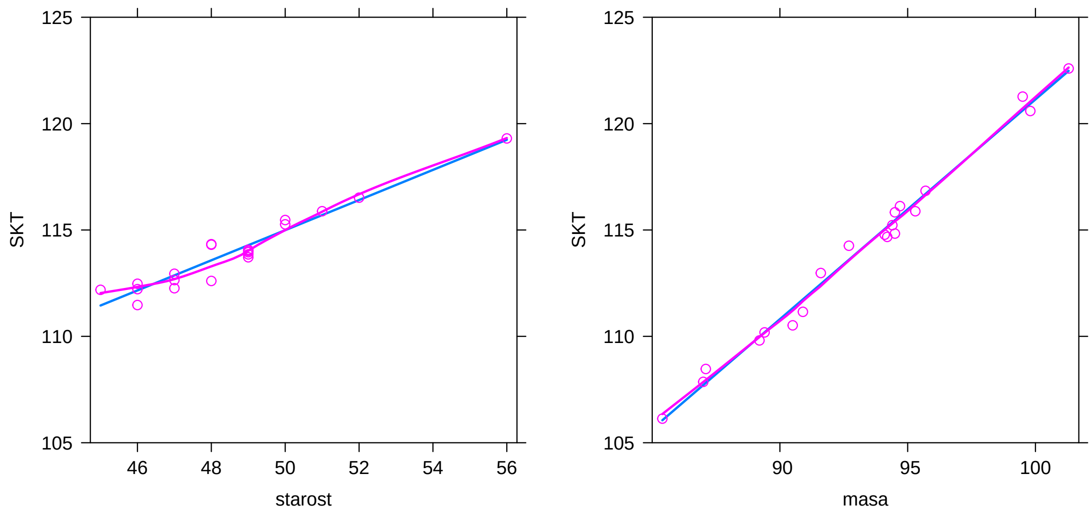
```
> library(effects)
> plot(predictorEffects(model.p, ~.), ylim=c(105,125),main="")
```



Slika 9: Napovedane vrednosti in 95 % intervali zaupanja za povprečen SKT za `model.p`; levo: glede na starost pri povprečni masi 93.1 kg, desno: glede na maso pri povprečni starosti 48.6 let

Funkcijo `predictorEffects()` lahko uporabimo tudi za grafične prikaze parcialnih ostankov tako kot v enostavnih kot tudi v kompleksnejšil linearnih modelih, v katerih so vključeni tudi interakcijski členi. Kot smo videli, za modele brez interakcijskih členov to naredi tudi funkcija `crPlot()`.

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



Slika 10: Napovedane vrednosti za povprečen SKT za `model.p` in parcialni ostanki z gladilnikom; levo: glede na starost pri povprečni masi 93.1 kg, desno: glede na maso pri povprečni starosti 48.6 let

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

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

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

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

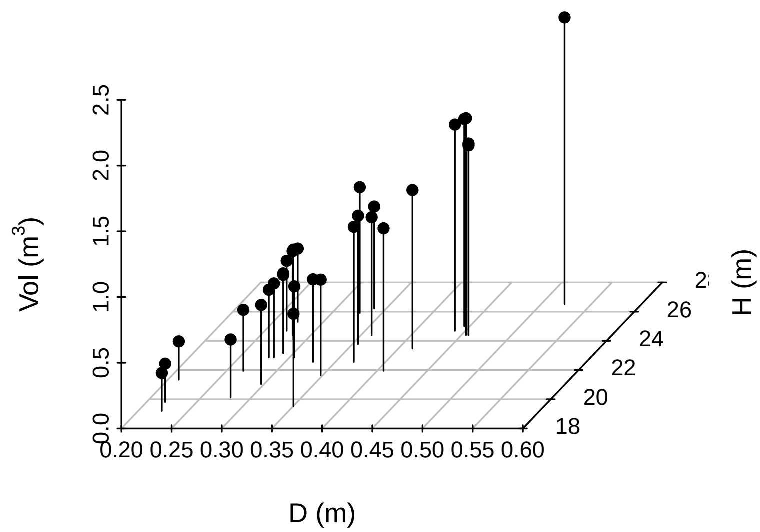
1.4 Primer: `trees`

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 odvisnost `Vol` od `D` in `H` (Slika 11).

```
> scatterplot3d(drevesa$D,drevesa$H, drevesa$Vol,pch=16, type="h", xlab="D (m)",
+               ylab="H (m)", zlab=expression(paste("Vol (",m^3,")")), box=F)
```



Slika 11: 3D prikaz podatkov za `drevesa`

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

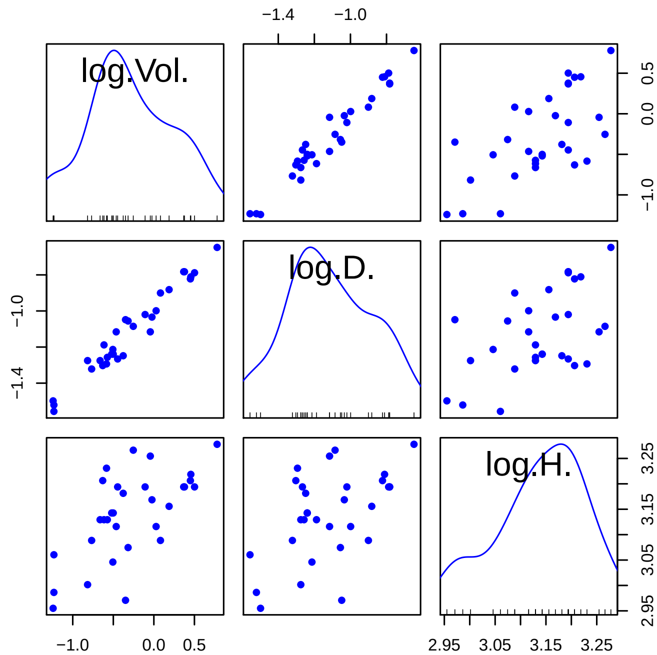
$$Vol = konst \cdot D^2 \cdot H. \tag{7}$$

To je multiplikativni model, saj se D^2 in H množita. Z logaritmiranjem (7) dobimo aditivni izraz, ki je primeren za analizo z linearnim modelom:

$$log(Vol) = log(konst) + 2 \cdot log(D) + 1 \cdot log(H). \tag{8}$$

Narišimo najprej matriko razsevnih grafikonov na logaritmiranih spremenljivkah.

```
> scatterplotMatrix(~log(Vol)+log(D)+log(H), regLine=F, smooth=FALSE,
+                      diagonal=list(method="density"), pch=16, data=drevesa)
```

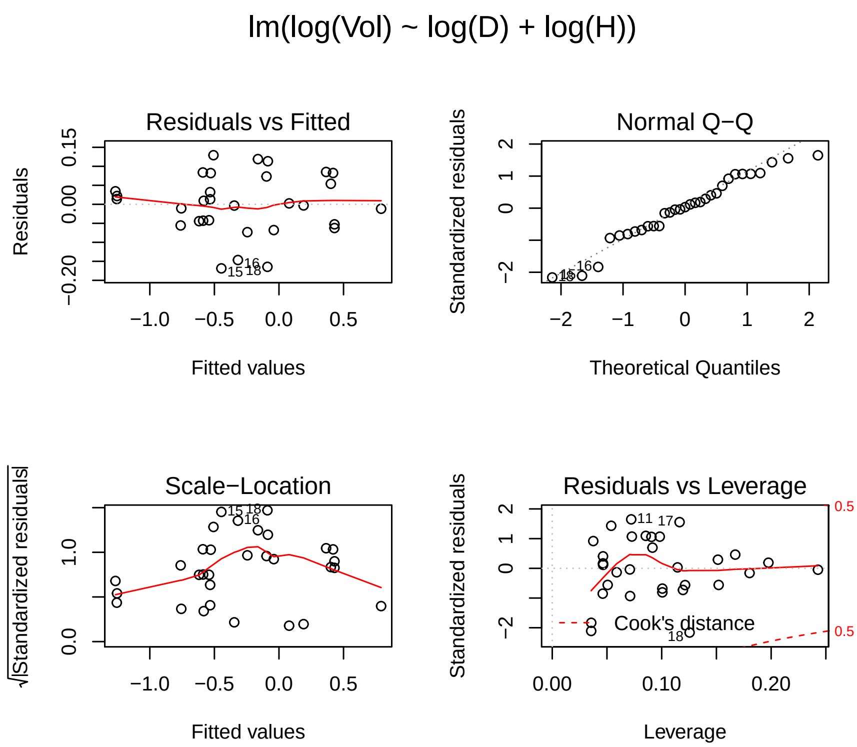


Slika 12: Matrika razsevnih grafikonov za logaritmirane spremenljivke iz podatkovnega okvira `drevesa`

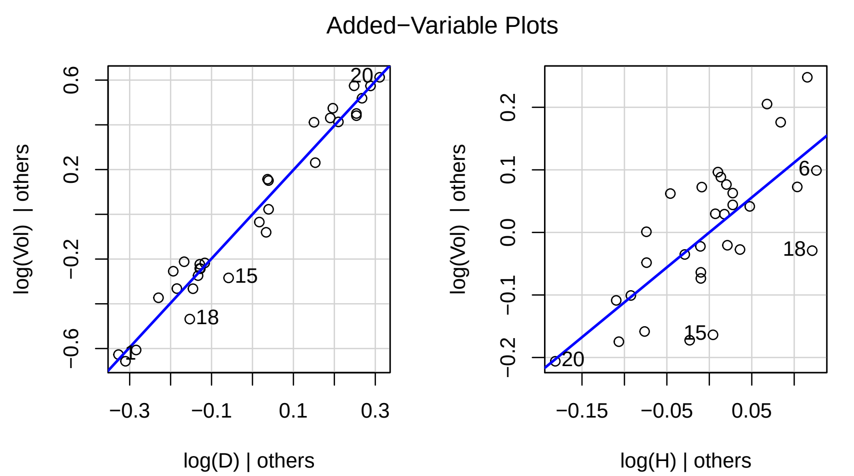
Slika 12 kaže, da je robna odvisnost `log(Vol)` od `log(D)` in od `log(H)` linearna. Naredimo model na logaritmiranih spremenljivkah in preverimo, ali so podatki v skladu z geometrijskim modelom. Če je tako, je parameter modela β_D pri `log(D)` enak 2, parameter β_H pri `log(H)` pa 1. V tem primeru gre za hkratno testiranje dveh domnev:

$$H_0: \quad \beta_D = 2, \qquad H_0: \quad \beta_H = 1. \tag{9}$$

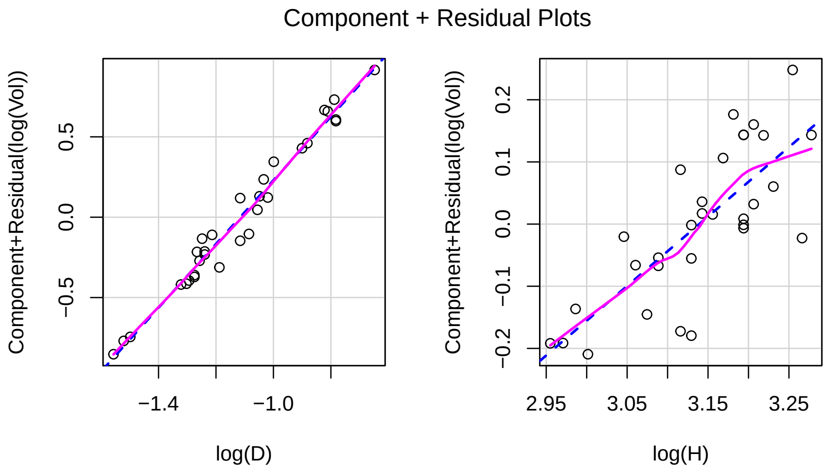
```
> model.d<- lm(log(Vol)~log(D)+log(H), data=drevesa)
```



Slika 13: Ostanki za `model.d`



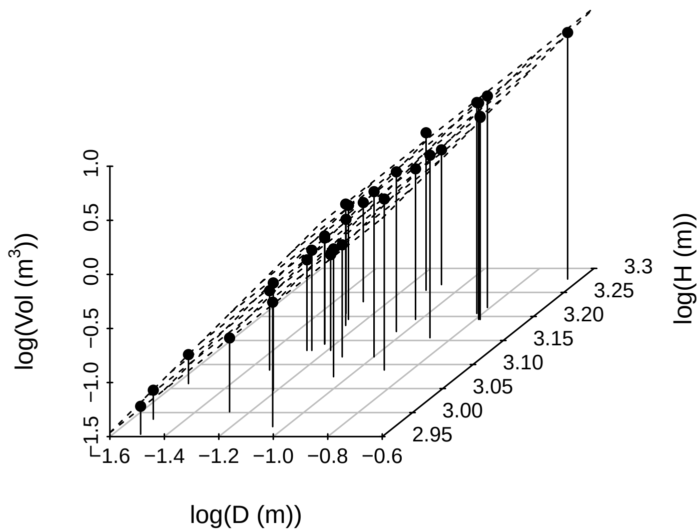
Slika 14: Grafi dodane spremenljivke za `model.d`



Slika 15: Grafi parcialnih ostankov za `model.d`

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

[1] 0.9776784

> drevesa.izpis<-glht(model.d)
> confint(drevesa.izpis)

         Simultaneous Confidence Intervals

Fit: lm(formula = log(Vol) ~ log(D) + log(H), data = drevesa)

Quantile = 2.3381
95% family-wise confidence level


Linear Hypotheses:
                 Estimate lwr     upr
(Intercept) == 0 -1.5864  -3.1996  0.0268
log(D) == 0       1.9826   1.8073  2.1580
log(H) == 0       1.1171   0.6391  1.5951
```

Napovedane vrednosti za `log(Vol)` ležijo na ravnini, ki jo določata `log(H)` in `log(D)` (Slika 16).



Slika 16: 3D prikaz podatkov za `drevesa` in napovedi za `model.d`

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.

Za hkratno testiranje domnev ničelni domnevi (9) zapišemo v obliki linearne kombinacije parametrov modela, desna stran obeh enačb tokrat ni enaka 0:

$$
\begin{aligned}
H_0: &\quad 0 \cdot \beta_0 + 1 \cdot \beta_D + 0 \cdot \beta_H = 2, \\
H_0: &\quad 0 \cdot \beta_0 + 0 \cdot \beta_D + 1 \cdot \beta_H = 1.
\end{aligned} \tag{10}
$$

Koeficiente linearne kombinacije zapišemo v matriko `K`, desno stran ničelnih domnev pa v vektor z argumentom `rhs` (*right hand side*). Za izračunavanje p-vrednosti se uporabi multivariatna t-porazdelitev s stopinjami prostosti ostanka modela.

```
> K<-rbind(c(0,1,0),c(0,0,1))
> rownames(K)<-c("beta_D", "beta_H")
> colnames(K)<-c("beta0", "beta_D", "beta_H")
> K

       beta0 beta_D beta_H
beta_D     0      1      0
beta_H     0      0      1
```

```
> trees.ht<-glht(model.d, linfct=K, rhs=c(2,1))
> summary(trees.ht)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = log(Vol) ~ log(D) + log(H), data = drevesa)

Linear Hypotheses:
            Estimate Std. Error t value Pr(>|t|)
beta_D == 2  1.98265    0.07501  -0.231    0.961
beta_H == 1  1.11712    0.20444   0.573    0.789
(Adjusted p values reported -- single-step method)
```

S korektnim testiranjem obeh ničelnih domnev hkrati ugotovimo, da se obe ničelni domnevi obdrži, kar pomeni, da nimamo razloga, da bi sumili v ustreznost geometrijskega modela (7).

1.5 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 $x_1 x_2$. Zamislimo si pričakovano vrednost tega modela v točki (x_{01}, x_{02}).

$$E(y|x_{01}, x_{02}) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_3 x_{01} x_{02}, \tag{11}$$

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

$$E(y|x_{01}, x_{02}+1) = \beta_0+\beta_1 x_{01}+\beta_2(x_{02}+1)+\beta_3 x_{01}(x_{02}+1) = \beta_0+\beta_1 x_{01}+\beta_2 x_{02}+\beta_3 x_{01} x_{02}+\beta_2+\beta_3 x_{01}. \tag{12}$$

Iz (11) in (12) sledi

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

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.

1.6 Primer: `postaje`

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`.

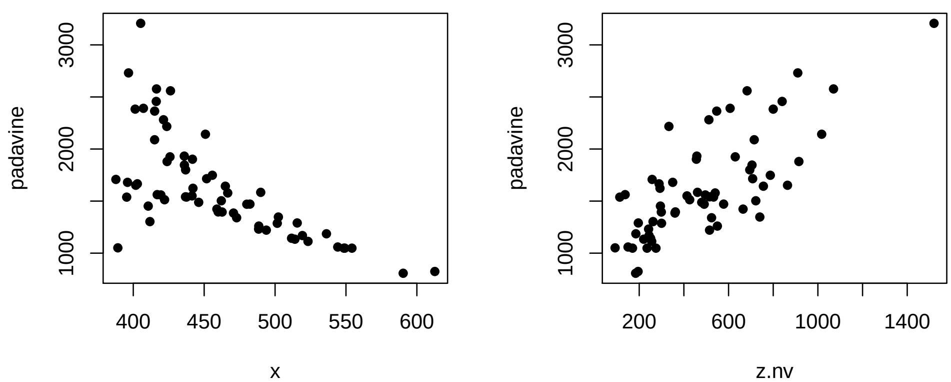
Geografska dolžina in geografska širina sta izraženi v Gauss-Kruggerjevem koordinatnem sistemu v metrih, zaradi lažje interpretacije ju bomo izrazili v km. Ker je podatek za Kredarico napačen, bomo Kredarico izločili, za dve postaji nimamo podatka za `x.gdol`. Analiza bo narejena na tistih postajah, kjer imamo vse podatke. Teh postaj je 64.

```
> postaje<-read.table("POSTAJE.txt", header=TRUE, sep="\t")
> rownames(postaje)<-postaje$Postaja
> postaje.brez<-subset(postaje, subset=postaje$Postaja!="Kredarica")
> postaje64<-na.omit(postaje.brez)
> postaje64$x<-postaje64$x.gdol/1000
> postaje64$y<-postaje64$y.gsir/1000
> summary(postaje64[, c("padavine", "x","z.nv")])
```

```
    padavine          x              z.nv
 Min.   : 807   Min.   :387.7   Min.   :  92.0
 1st Qu.:1289   1st Qu.:416.8   1st Qu.: 261.0
 Median :1540   Median :444.1   Median : 470.5
 Mean   :1625   Mean   :458.2   Mean   : 490.5
 3rd Qu.:1854   3rd Qu.:488.9   3rd Qu.: 686.0
 Max.   :3207   Max.   :612.6   Max.   :1520.0
```

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

```
> par(mfrow=c(1,2))
> with(postaje64, plot(x, padavine,  pch=16))
> with(postaje64, plot(z.nv, padavine,  pch=16))
```

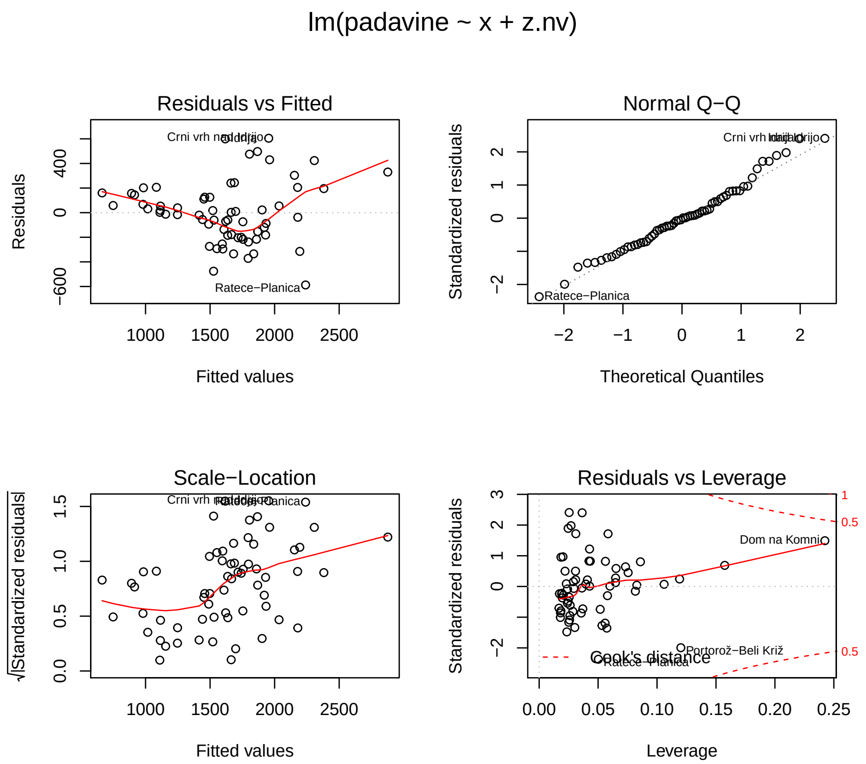


Slika 17: `padavine` v odvisnosti od `z.nv` in od `x`

Kakšna je odvisnost spremenljivke `padavine` od `x`? Kakšna je odvisnost od `z.nv`? Kaj sličice povedo o variabilnosti? Kje se nakazuje nekonstantna variabilnost?

Pri napovedovanju količine padavin je interakcija med nadmorsko višino in geografsko dolžino v Sloveniji pričakovana. Iz pedagoških razlogov bomo najprej naredili model `model.m1` brez interakcijskega člena in analizirali ostanke.

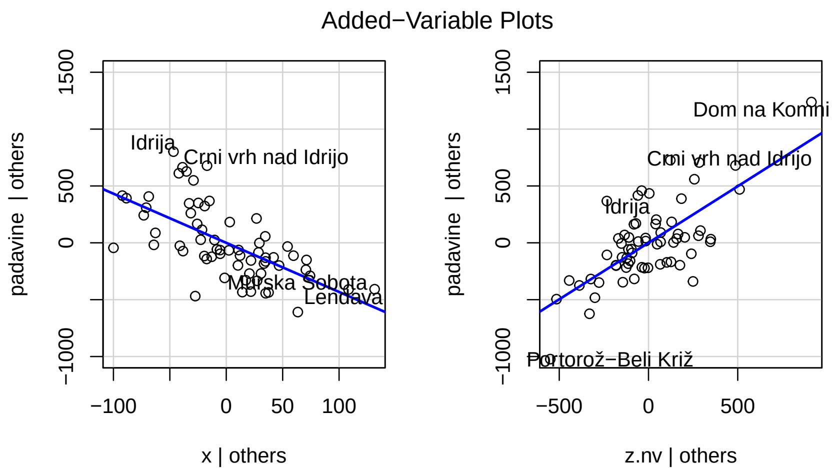
```
> model.m1<-lm(padavine~ x + z.nv, data=postaje64)
```



Slika 18: Ostanki za `model.m1`

V modelu je prisotna heteroskedastičnost, ki je razvidna iz grafa levo spodaj.

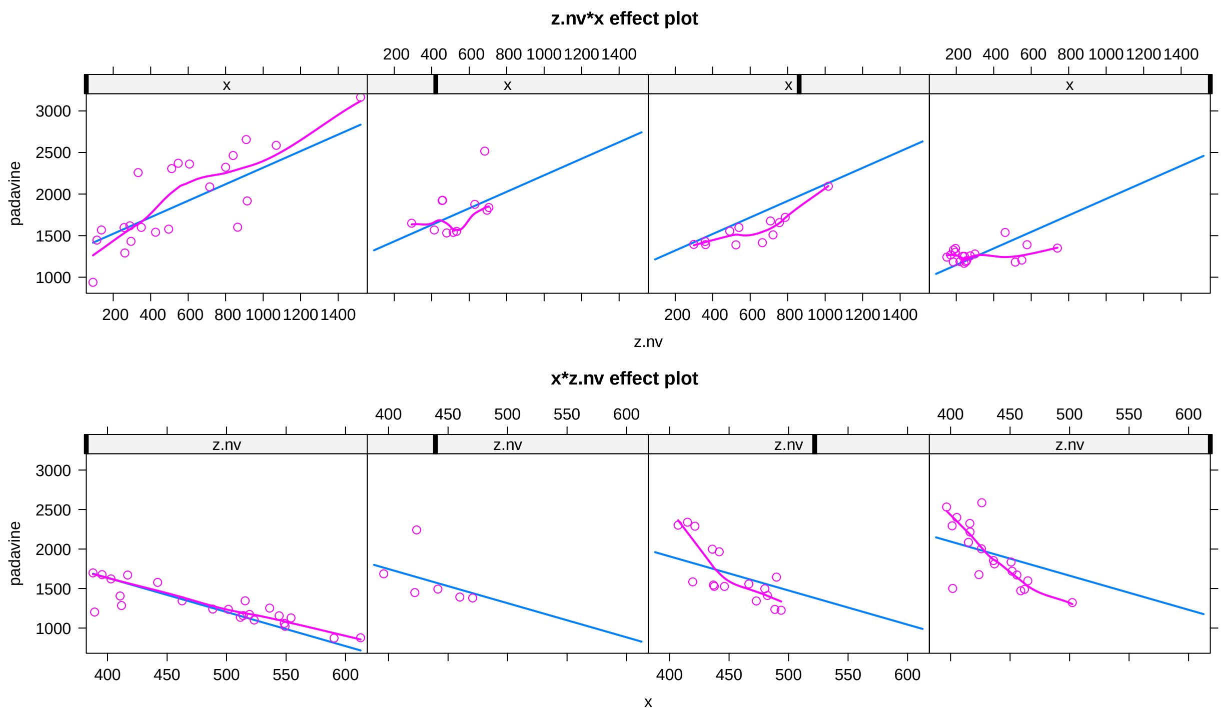
```
> avPlots(model.m1, ylim=c(-1000,1500), id=list(location="avoid"))
```



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. Gladilniki nakazujejo, da je med `x` in `z.nv` prisotna interakcija.

```
> 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: Napovedane vrednosti za `padavine` za `model.m1`; v odvisnosti od nadmorske višine pri samodejno izbranih vrednostih geografske dolžine (zgoraj) in v odvisnosti od geografske dolžine pri samodejno izbranih vrednostih nadmorske višine (spodaj), prikazani so tudi parcialni ostanki z gladilnikom

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

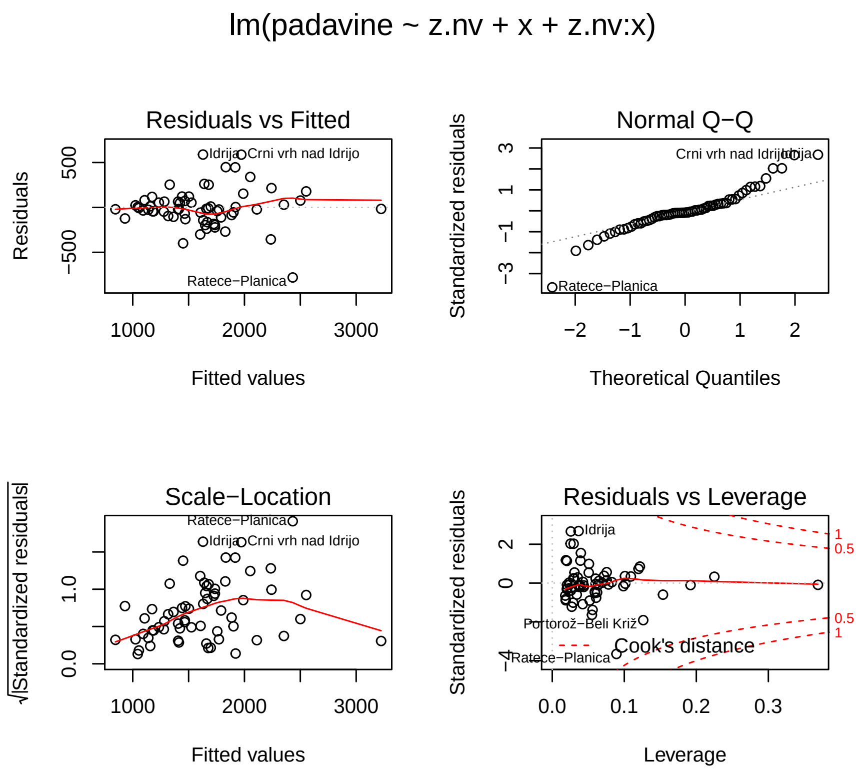
```
> model.m2<-lm(padavine~ z.nv + x + z.nv:x , data=postaje64)
> # model.m2<-lm(padavine~ z.nv * x  , data=postaje64) # krajši zapis
> model.m2$coeff

 (Intercept)           z.nv              x         z.nv:x
```

```
1736.34007572    6.05218444   -1.07842493   -0.01181133
```

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

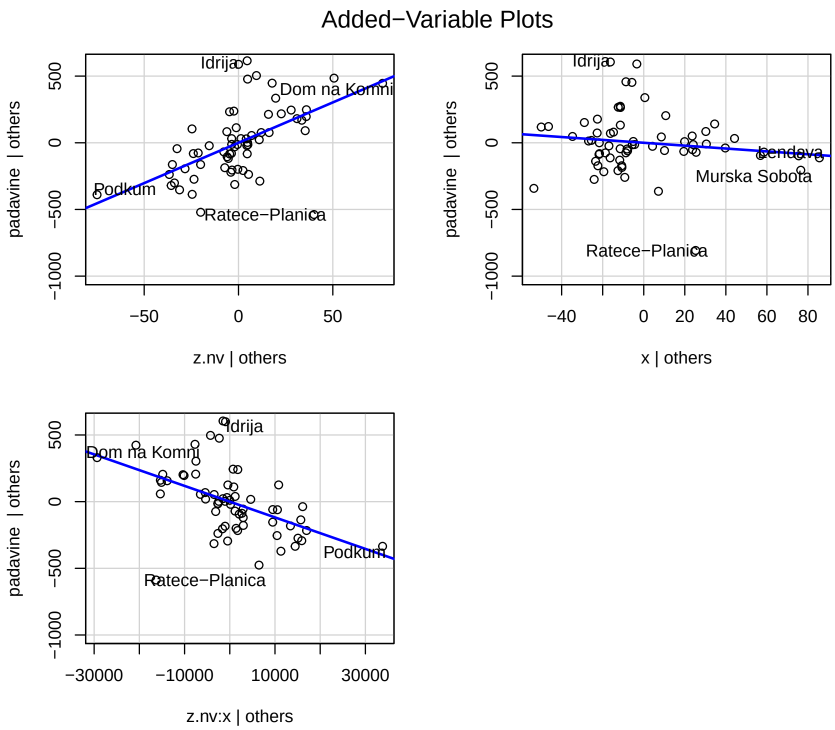
```
> X.m2<-model.matrix(model.m2)
> head(X.m2, n=3)
```

```
                     (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
```



Slika 21: Ostanki za `model.m2`

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



Slika 22: Grafi dodane spremenljivke za `model.m2`

```
> outlierTest(model.m2)
```

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

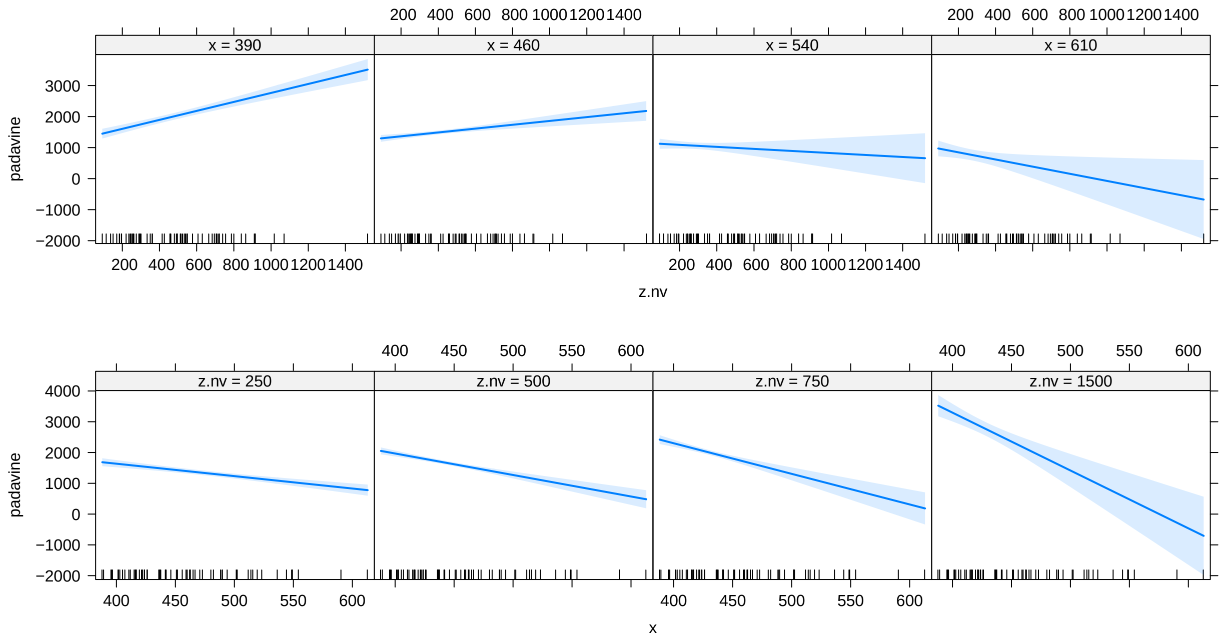
Slika 21 in 22 še vedno kažeta heteroskedastičnost. Kljub očitni heteroskedastičnosti izpišimo povzetek modela z namenom, da se naučimo interpretirati ocene parametrov modela z dvema številskima napovednima spremenljivkama in njuno interakcijo. S funkcijo `glht` dobimo ustrezne izpise za povzetek modela in ustrezne intervale zaupanja.

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

[1] 0.8001414

> postaje.izpis<-glht(model.m2)
> confint(postaje.izpis)

         Simultaneous Confidence Intervals

Fit: lm(formula = padavine ~ z.nv + x + z.nv:x, data = postaje64)

Quantile = 2.2576
95% family-wise confidence level


Linear Hypotheses:
                 Estimate   lwr        upr
(Intercept) == 0  1.736e+03  7.484e+02  2.724e+03
z.nv == 0         6.052e+00  3.421e+00  8.684e+00
x == 0           -1.078e+00 -3.232e+00  1.075e+00
z.nv:x == 0      -1.181e-02 -1.793e-02 -5.695e-03
```

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 23 zgoraj); vpliv geografske dolžine na količino padavin se zmanjšuje z večjo nadmorsko višino (Slika 23 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 23: 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)

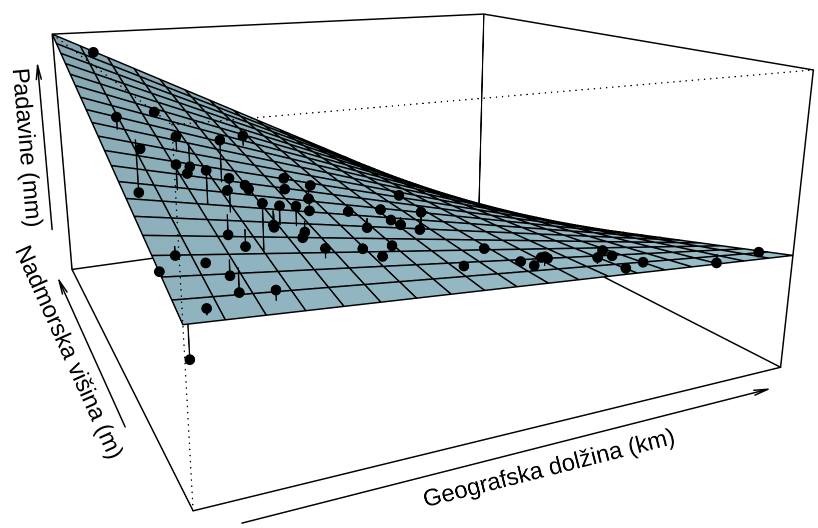
- če se premaknemo za 50 km proti vzhodu in ostanemo na isti nadmorski višini, se količina padavin v povprečju spremeni za:

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

  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
```

Napovedi dobljene z modelom `model.m2` so grafično predstavljene v 3D na Sliki 24. Napovedi so izračunane v mreži točk določeni z razponom vrednosti geografske dolžine in geografske širine. Ker območje Slovenije ni pravokotne oblike in ker meteorološke postaje niso enakomerno porazdeljene po nadmorski višini, veliko napovedi predstavlja ekstrapolacijo.



Slika 24: Napovedane padavine za `model.m2` narisane v 3D

Poglejmo si 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 povedanose 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 ($p < 0.0001$).

1.7 Primer: `Carseats`

V podatkovnem okviru `Carseats` v paketu `ISLR` so podatki za 400 trgovskih centrov v ZDA, ki prodajajo tudi avtomobilske otroške stolčke. Upoštevali bomo naslednje spremenljivke, podatki so za določeno koledarsko leto:

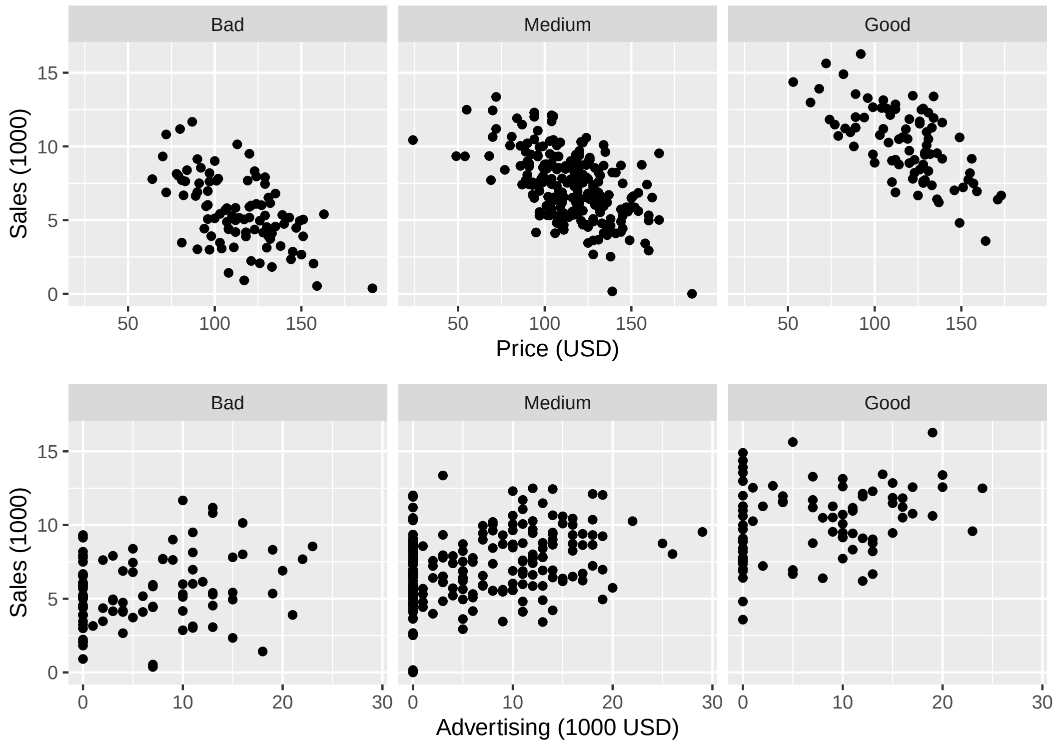
- `Sales` predstavlja letno število prodanih stolčkov (v 1000) za posamezni trgovski center;
- `Price` (USD) je cena posameznega stolčka v trgovskem centru;
- `Advertising` predstavlja letne stroške oglaševanja (1000 USD) za posamezni trgovski center;
- `ShelveLoc`, ki je kakovost police, na kateri prodajajo otroške stolčke, njene vrednosti so: Bad, Medium, Good.

Zanima nas odvisnost `Sales` od `Price`, `Advertising` in `ShelveLoc`.

```
> library(ISLR)
> data(Carseats)
> #zamenjamo vrstni red ravni faktorja ShelveLoc
> Carseats$ShelveLoc<-factor(Carseats$ShelveLoc, levels=c("Bad","Medium","Good"))
```

V tem primeru bomo v model vključili dve številski in eno opisno napovedno spremenljivko. Za preliminarni grafični prikaz v tem primeru uporabimo razsevni grafikon za odvisnost `Sales` od `Price` za vsako vrednost `ShelveLoc` posebej in enako za odvisnost `Sales` od `Advertising` (Slika 25).

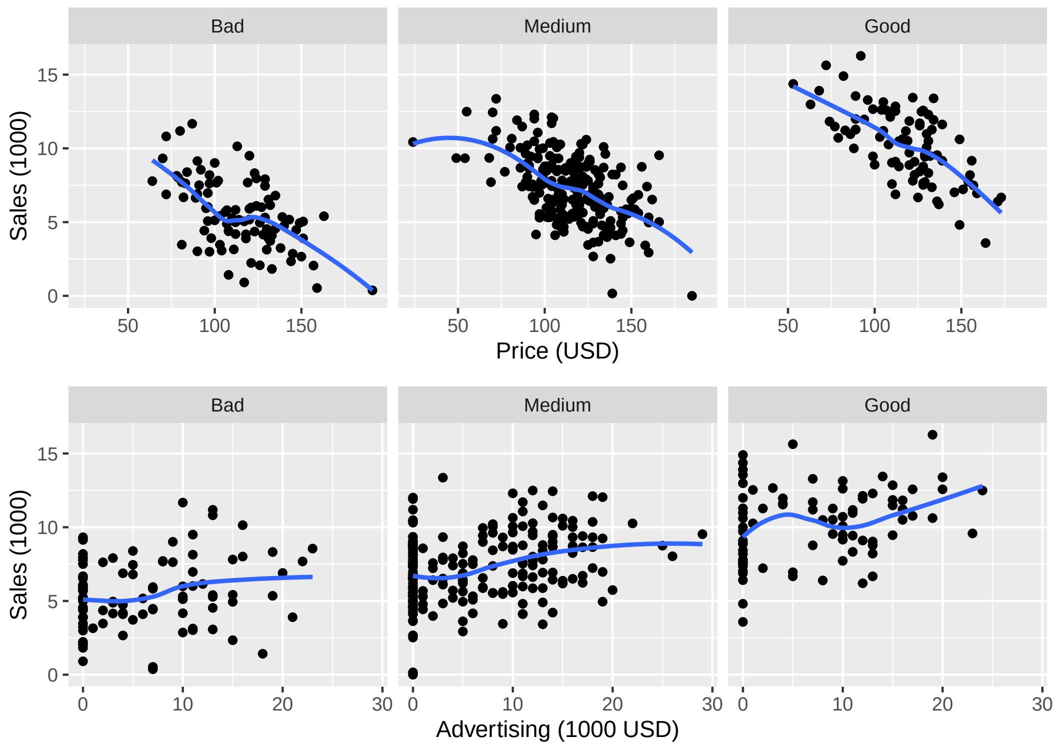
```
> library(ggplot2)
> p1<-ggplot(data=Carseats, aes(x=Price, y=Sales)) +
+   facet_grid(.~ShelveLoc) + geom_point() +
+   xlab("Price (USD)")  +   ylab("Sales (1000)")
> p2<-ggplot(data=Carseats, aes(x=Advertising, y=Sales)) +
+   facet_grid(.~ShelveLoc) +  geom_point() +
+   xlab("Advertising (1000 USD)")  +   ylab("Sales (1000)")
> grid.arrange(p1,p2, nrow=2, ncol=1)
```



Slika 25: `Sales (1000)` v odvisnosti od `Price (USD)` in `ShelveLoc` (zgoraj), `Sales (1000)` v odvisnosti od `Advertising (1000 USD)` in `ShelveLoc` (spodaj)

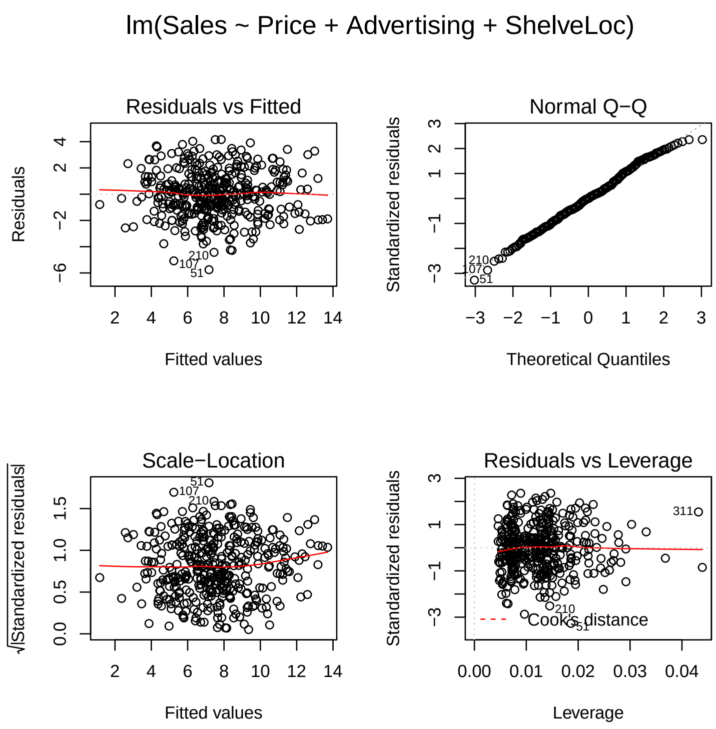
Pogosto na začetne grafikone dodamo gladilnik, ki ga naredi neparametrična regresija in nam kaže naravo odvisnosti (linearnost ali nelinearnost):

```
> p1<-p1+geom_smooth(se=FALSE)
> p2<-p2+geom_smooth(se=FALSE)
> grid.arrange(p1,p2, nrow=2, ncol=1)
```



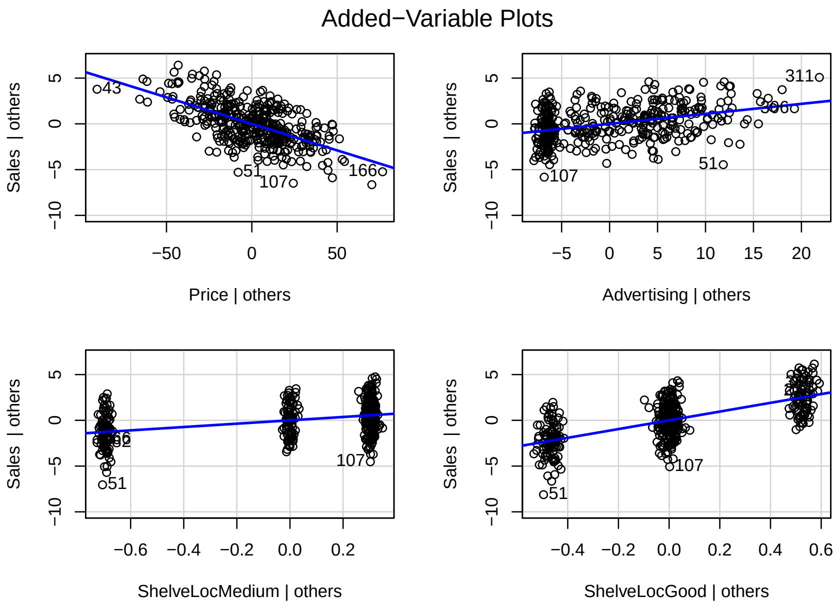
Slika 26: `Sales (1000)` v odvisnosti od `Price (USD)` glede na `ShelveLoc` (zgoraj), `Sales (1000)` v odvisnosti od `Advertising (1000 USD)` glede na `ShelveLoc` (spodaj) z dodanimi gladilniki neparametrične regresije

```
> model.stol<-lm(Sales~ Price + Advertising + ShelveLoc, data=Carseats)
```



Slika 27: Ostanki za `model.stol`

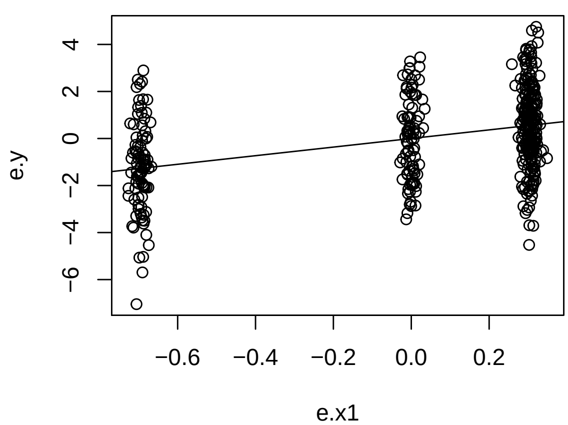
```
> avPlots(model.stol, ylim=c(-10,7))
```



Slika 28: Grafi dodane spremenljivke za `model.stol`

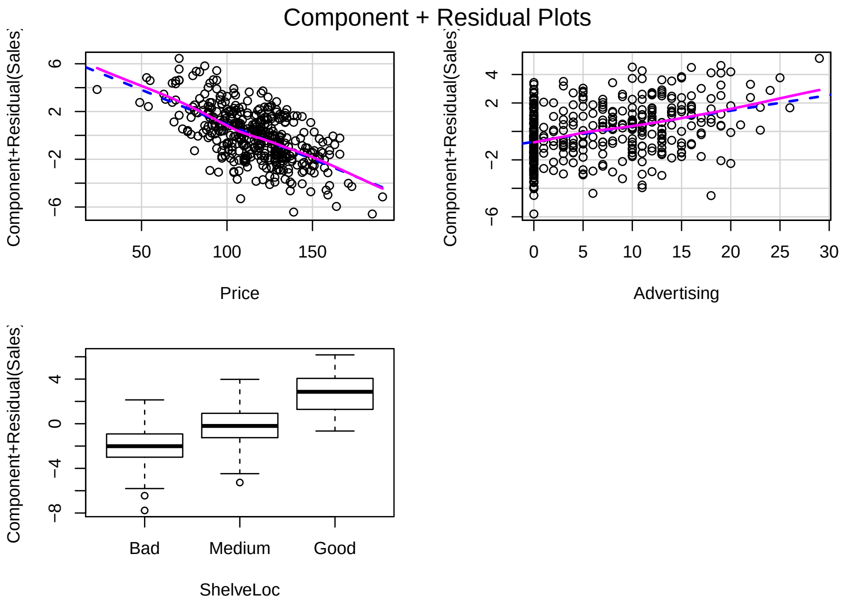
Pokažimo, kaj predstavlja graf dodane spremenljivke, če je ta opisna. Spremenljivka `ShelveLoc` ima tri vrednosti, kar pomeni, da sta v modelu dve umetni spremenljivki `ShelveLocMedium` in `ShelveLocGood`. Izračunanjmo ustrezne ostanke za `ShelveLocMedium`:

```
> Carseats$x.1 <-model.matrix(model.stol)[,"ShelveLocMedium"]
> Carseats$x.2 <-model.matrix(model.stol)[,"ShelveLocGood"]
> e.y <- residuals(lm(Sales~ Price + Advertising + x.2, data=Carseats))
> e.x1 <- residuals(lm(x.1~Price + Advertising + x.2, data=Carseats))
> mod.e1 <- lm(e.y~e.x1)
> (b.e1 <- coef(summary(mod.e1))[2,1])

[1] 1.828803

> (s.b.e1 <- coef(summary(mod.e1))[2,2])

[1] 0.2166708

> # b in standardna napaka na podlagi polnega modela
> (b <- coef(summary(model.stol))[4,1]);(s.b <- coef(summary(model.stol))[4,2])

[1] 1.828803

[1] 0.2174921
```



Slika 29: Graf dodane spremenljivke `ShelveLocMedium`, odvisnost `Sales` od `ShelveLocMedium` ob upoštevanju `Advertising`, `Price` and `ShelveLocGood`

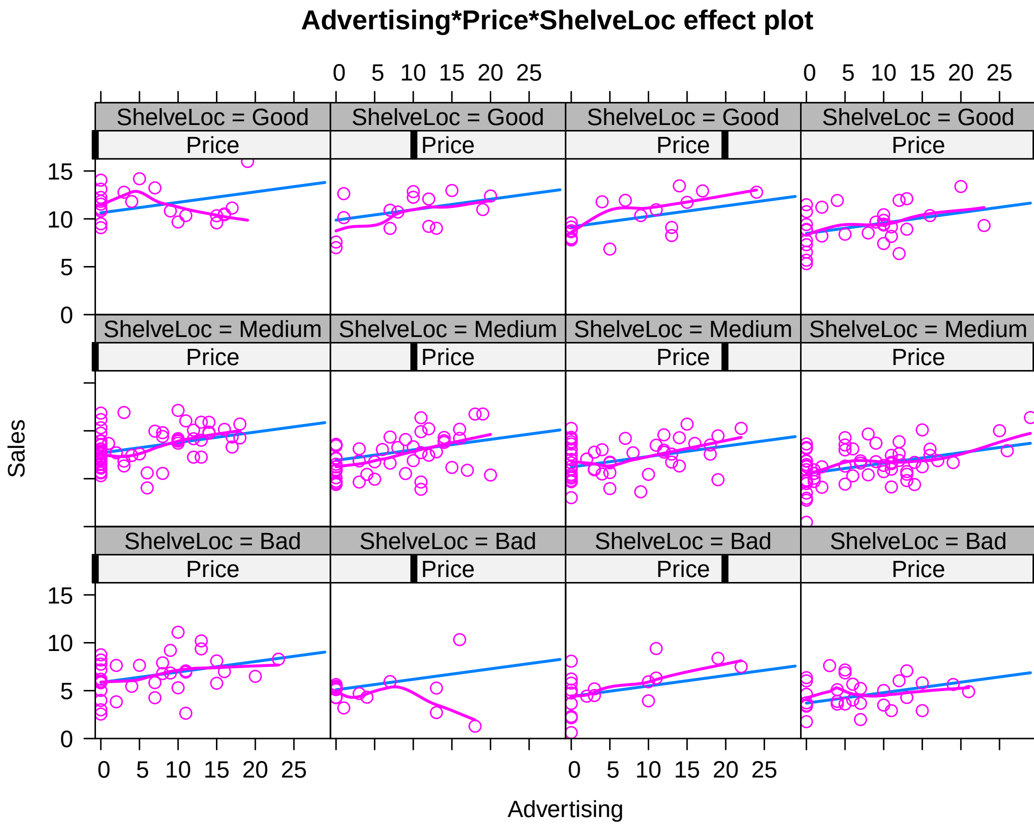
```
> crPlots(model.stol)
```



Slika 30: Grafi parcialnih ostankov za `model.stol`

Ali bi bilo potrebno v model vključiti tudi interakcijski člen `Price:Advertising`? Narišimo grafe parcialnih ostankov s funkcijo `Effect()` (Slika 31).

```
> plot(Effect(c("Advertising","Price", "ShelveLoc"), model.stol, partial.residuals=TRUE),
+       ci.style="none")
```



Slika 31: Grafi parcialnih ostankov za `model.stol` s funkcijo `Effect()`

Gladilniki parcialnih ostankov za `Advertising` na Sliki 31 pri različnih vrednostih `Price` in `ShelveLoc` so dokaj vzporedni in se v večini primerov prilegajo primici, kar pomeni, da interakcijskega člena ni potrebno vključiti v model. Po vseh kriterijih diagnostike linearnega modela je `model.stol` sprejemljiv. Nadaljujemo z analizo povzetka modela.

Za preverjenje statistične značilnosti opisne spremenljivke `ShelveLoc` ob upoštevanju ostalih spremenljivk v modelu lahko uporabimo sekvenčni F-test, ker je ta spremenljivka v `model.stol` na zadnjem mestu.

```
> anova(model.stol)

Analysis of Variance Table

Response: Sales
             Df  Sum Sq Mean Sq F value    Pr(>F)
Price         1  630.03  630.03 199.743 < 2.2e-16 ***
Advertising   1  266.91  266.91  84.621 < 2.2e-16 ***
ShelveLoc     2 1039.42  519.71 164.767 < 2.2e-16 ***
```

```
Residuals   395 1245.91    3.15
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vpliv `ShelveLoc` se pokaže kot močno statistično značilen ($p < 0.0001$). Če je vrstni red napovednih spremenljivk v modelu tak, da izbrana opisna spremenljivka ni zadnja v formuli modela (npr. `model.stol.a`), moramo narediti še model brez izbrane opisne spremenljivke `model.stol.0` in ga primerjati z `model.stol.a`. Uporabimo F-test za ekvivalentnost dveh modelov, ki ga naredi funkcija `anova(model.stol.0, model.stol.a)`.

```
> model.stol.a<-lm(Sales~ ShelveLoc + Price + Advertising, data=Carseats)
> model.stol.0<-lm(Sales~ Price + Advertising, data=Carseats)
> anova(model.stol.0, model.stol.a)
```

```
Analysis of Variance Table

Model 1: Sales ~ Price + Advertising
Model 2: Sales ~ ShelveLoc + Price + Advertising
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    397 2285.3
2    395 1245.9  2    1039.4 164.77 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V nadaljevanju v `model.stol` popravimo p-vrednosti in intervale zaupanja za parametre modela zaradi hkratnega testiranja domnev.

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

```
[1] 0.6084832
```

```
> stol.izpis<-glht(model.stol)
> summary(stol.izpis)
```

```
         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = Sales ~ Price + Advertising + ShelveLoc, data = Carseats)

Linear Hypotheses:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0     11.468018   0.470930  24.352   <1e-10 ***
Price == 0           -0.057975   0.003764 -15.404   <1e-10 ***
Advertising == 0      0.109305   0.013405   8.154   <1e-10 ***
ShelveLocMedium == 0  1.828803   0.217492   8.409   <1e-10 ***
ShelveLocGood == 0    4.776488   0.265261  18.007   <1e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

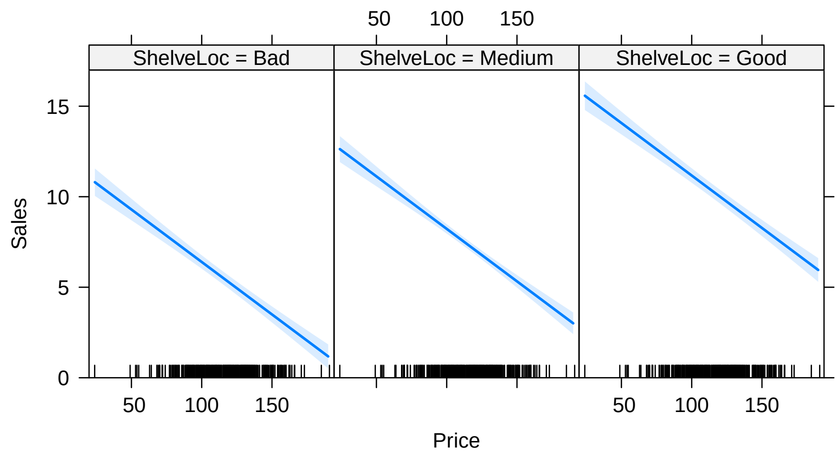
```
> confint(stol.izpis)
```

```
         Simultaneous Confidence Intervals

Fit: lm(formula = Sales ~ Price + Advertising + ShelveLoc, data = Carseats)

Quantile = 2.5269
95% family-wise confidence level


Linear Hypotheses:
                     Estimate lwr      upr
(Intercept) == 0     11.46802 10.27801 12.65803
Price == 0           -0.05797 -0.06749 -0.04846
Advertising == 0      0.10931  0.07543  0.14318
ShelveLocMedium == 0  1.82880  1.27921  2.37839
ShelveLocGood == 0    4.77649  4.10619  5.44679
```

Sklepi:

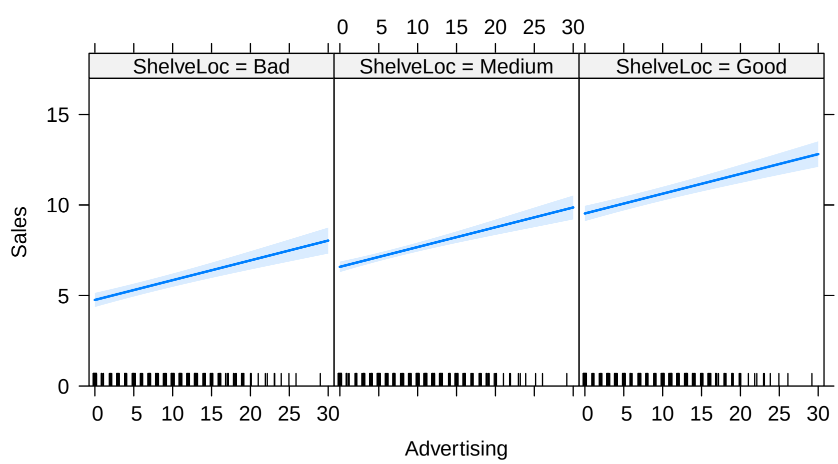
- z modelom je pojasnjene 61 % variabilnosti `Sales`;
- `Price`, `Advertising` in `ShelveLoc` so močno statististično značilni ($p < 0.0001$), njihov vpliv je aditiven;
- ob upoštevanju `Price in Advertising` v modelu je `Sales` na `ShelveLoc=Medium` za 1.829 enot večji kot na `ShelveLoc=Bad`, pripadajoč 95% IZ je (1.28 enot, 2.378 enot); na `ShelveLoc=Good` pa za 4.776 enot večji kot na `ShelveLoc=Bad`, pripadajoč 95 % IZ je (4.106, enot, 5.447 enot);
- ob upoštevanju `ShelveLoc` in `Advertising` velja: če se `Price` poveča za 10 USD, se `Sales` zmanjša za za 0.58 enot, pripadajoč 95 % IZ je (4.8 enot, 6.7 enot);
- ob upoštevanju `ShelveLoc` in `Price` velja: če se `Advertising` poveča za 1 enoto, se `Sales` poveča za 0.109 enot, pripadajoč 95 % IZ je (0.075 enot, 0.143 enot).

```
> plot(Effect(c("Price", "ShelveLoc"), model.stol), main="",
+       layout=c(3,1), ylim=c(0, 17))
```



Slika 32: Napovedane povprečne vrednosti za `Sales` v odvisnosti od `Price` and `ShelveLoc` pri povprečni vrednosti za `Advertising` z ovojnico 95 % intervalov zauapanja za povprečno napoved za `model.stol`

```
> plot(Effect(c("Advertising", "ShelveLoc"), model.stol), main="",
+       layout=c(3,1), ylim=c(0, 17))
```



Slika 33: Napovedane vrednosti za `Sales` v odvisnosti od `Advertising` and `ShelveLoc` pri povprečni vrednosti za `Price` z ovojnico 95 % intervalov zauapanja za povprečno napoved za `model.stol`

Vaja: napišite matriko primerjav, s katero hkrati testirate vse pare razlik položajev polic (`ShelveLoc`) in statistično značilnost `Price` in `Advertising`.

2 VAJE

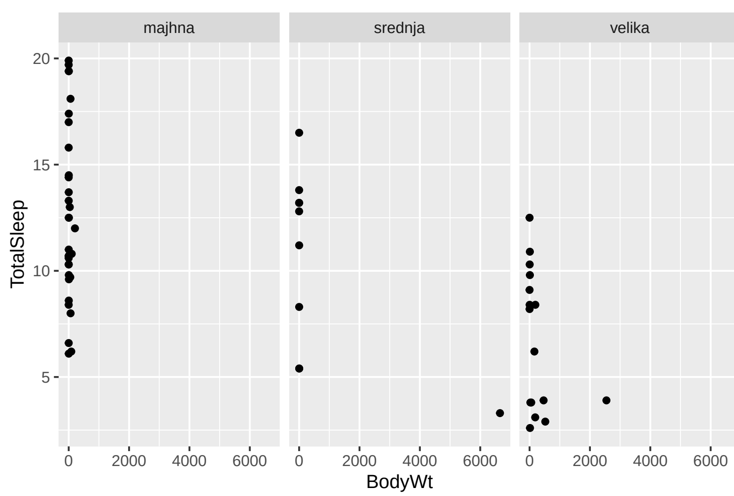
2.1 Spanje

Na datoteki SLEEP.txt (manjkajoči podatki označeni z NA) so podatki za 62 sesalcev. Glej `http://www.statsci.org/data/general/sleep.html`. Delno to informacijo poznamo iz podatkovnega okvira `mammals`. Analizirajte, kako je `TotalSleep` (h/day) odvisen od `BodyWt` (kg) in `Danger3`.

```
> spanje<-read.table("SLEEP.txt", header=T, sep="\t", na.string="NA", dec=".")
> head(spanje)
```

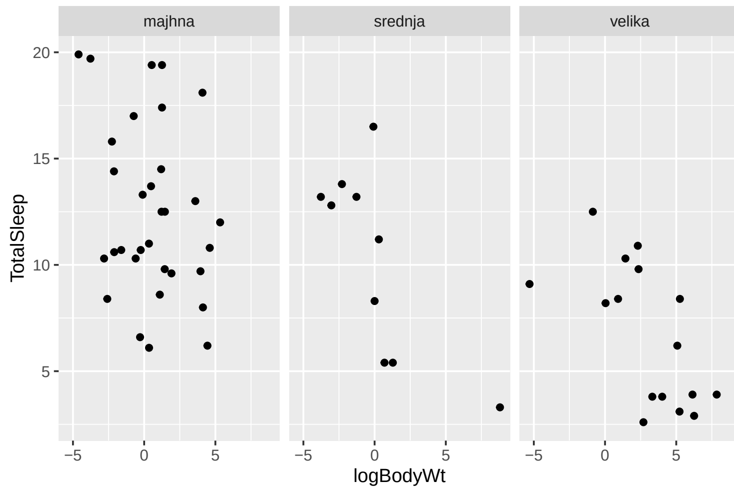
```
                  Species   BodyWt BrainWt TotalSleep LifeSpan Gestation Danger3
1        Africanelephant 6654.000  5712.0        3.3     38.6       645 srednja
2 Africangiantpouchedrat    1.000     6.6        8.3      4.5        42 srednja
3              ArcticFox    3.385    44.5       12.5     14.0        60  majhna
4   Arcticgroundsquirrel    0.920     5.7       16.5       NA        25 srednja
5          Asianelephant 2547.000  4603.0        3.9     69.0       624  velika
6                 Baboon   10.550   179.5        9.8     27.0       180  velika
```

```
> rownames(spanje)<-spanje$Species
```

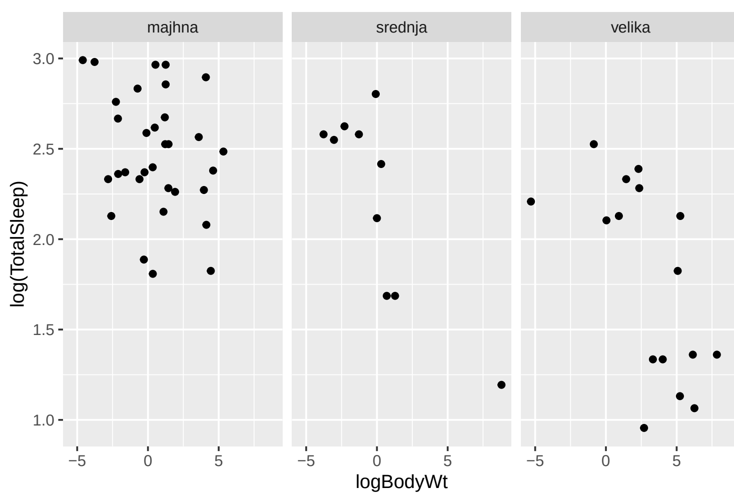


Slika 34: `TotalSleep` v odvisnosti od `BodyWt` za različne vrednosti spremenljivke `Danger3`

Vrednosti napovedne spremenljivke `BodyWt` so med minimalno in maksimalno vrenostjo zelo neenakomerno razporejene, velikih vrednosti je malo, na majhnem razponu majhnih vrednosti pa so zelo različne vrednosti `TotalSleep`, zato poskusimo `BodyWt` logaritmirati. Slika 34 kaže odvisnost `TotalSleep` od `logBodyWt`, ki je precej bolj primerna za modeliranje z normalnim linearnim modelom.



Slika 35: `TotalSleep` v odvisnosti od `logBodyWt` za različne vrednosti spremenljivke `Danger3`



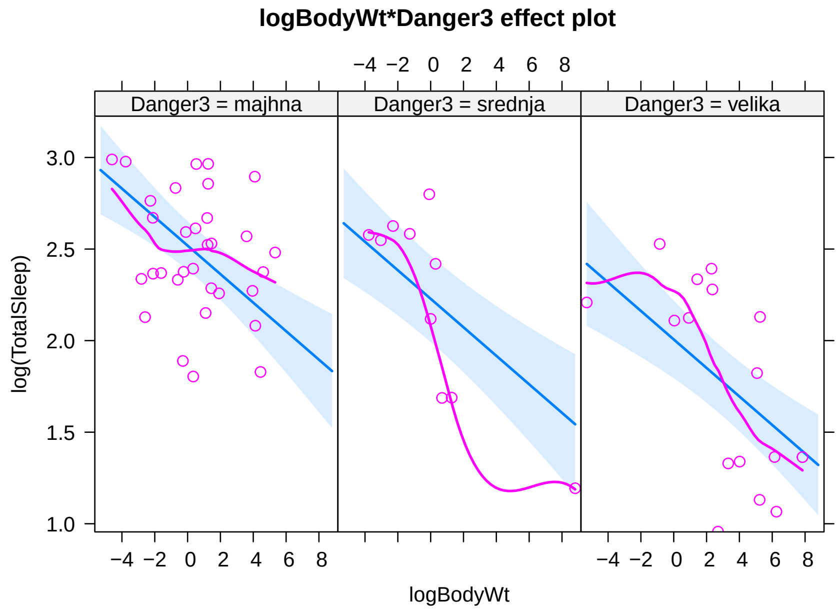
Slika 36: `log(TotalSleep)` v odvisnosti od `logBodyWt` za različne vrednosti spremenljivke `Danger3`

Slika 35 kaže, na težave z nekonstantno varianco, zato transformiramo tudi odzivno spremenljivko.

```
> mod.1 <- lm(log(TotalSleep) ~ Danger3 + logBodyWt, data=spanje)
```

Ali je v model smiselno vključiti interakcijo med `Danger3` in `logBodyWT`? Najprej poglejmo sliko parcialnih ostankov za `mod.1`.

```
> plot(Effect(c("logBodyWt", "Danger3"), mod.1, partial.residuals=TRUE),
+       span=0.8, layout=c(3,1))
```

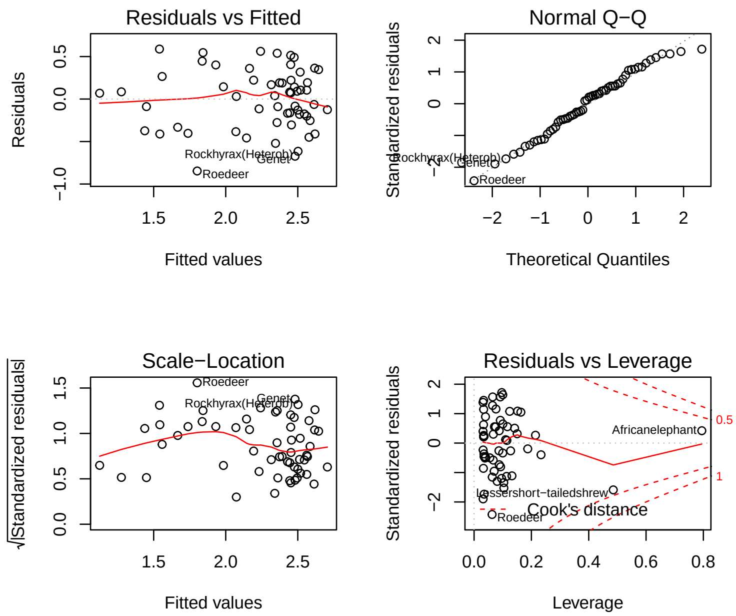


Slika 37: Parcialni ostanki za `mod.1`

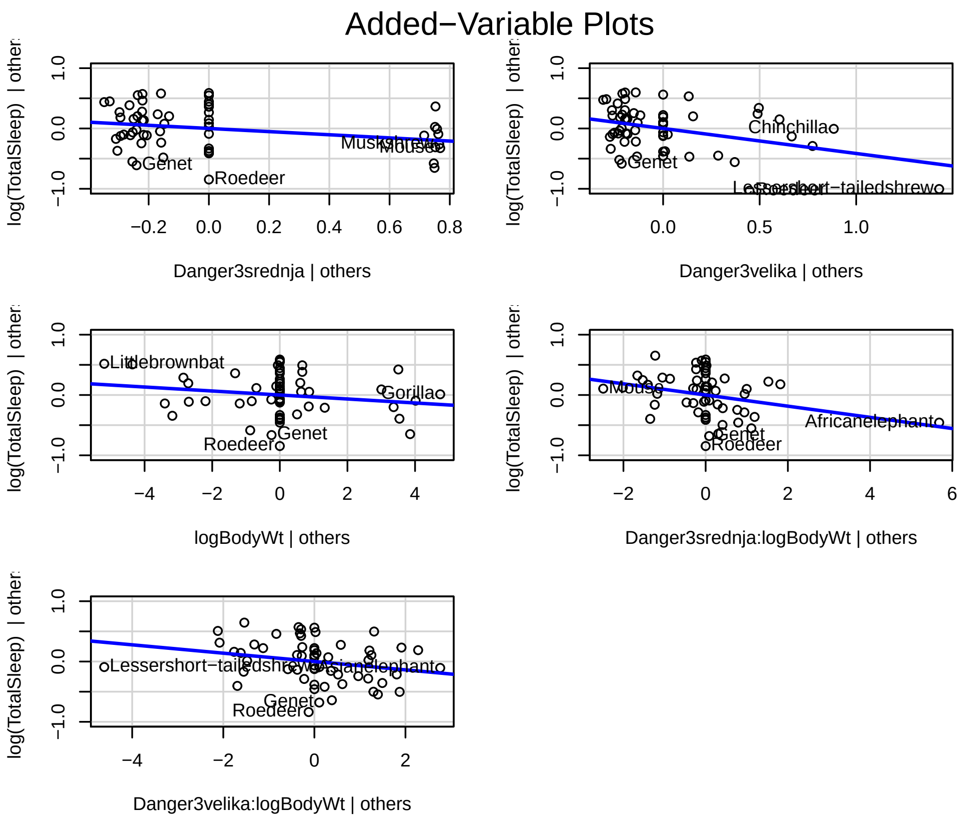
Gladilniki za parcialne ostanke na Sliki 37 so zaradi majhnega števila podatkov precej valoviti vseeno pa kažejo, da odvisnost od `logBodyWt` po kategorijah `Danger3` ni enaka, kar kaže na smiselnost vključitve interakcijskega člena.

```
> mod.1.int <- lm(log(TotalSleep) ~ Danger3 * logBodyWt, data=spanje)
> anova(mod.1, mod.1.int)

Analysis of Variance Table

Model 1: log(TotalSleep) ~ Danger3 + logBodyWt
Model 2: log(TotalSleep) ~ Danger3 * logBodyWt
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     54 7.5144
2     52 6.7579  2   0.75646 2.9103 0.06338 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interakcija je mejno statistično značilna ($p = 0.063$). Nadaljujemo z analizo modela v katerega vključimo interakcijo.



Slika 38: Ostanki za `mod.1.int`



Slika 39: Grafi dodane spremenljivke za `mod.1.int`

```
> outlierTest(mod.1.int)

No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
         rstudent unadjusted p-value Bonferroni p
Roedeer -2.550775           0.013793      0.79997
```

Sliki 38 in 37 ne kažeta večjega odstopanja od predpostavk linearnega modela, prav tako ni vplivnih točk in regresijskih osamelcev.

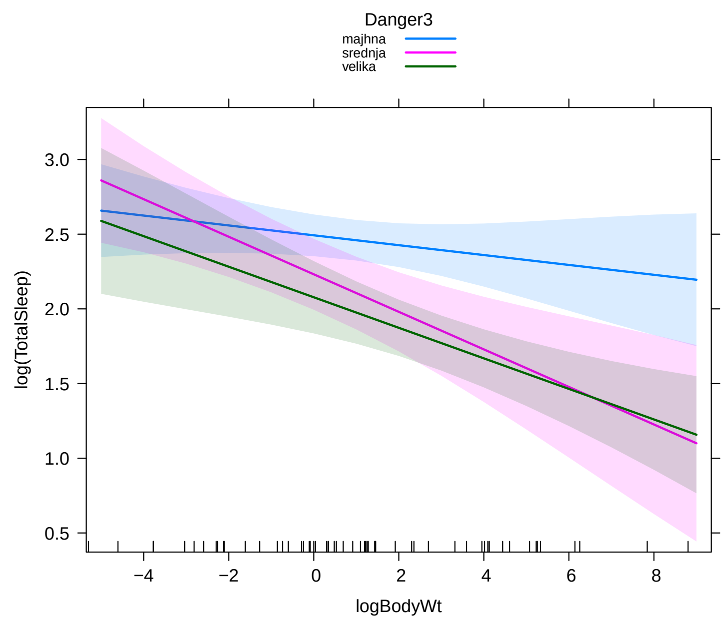
```
> summary(mod.1.int)$r.squared

[1] 0.5646614

> coef(summary(mod.1.int))

                            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)               2.49238213 0.06543382 38.090123 1.162609e-39
Danger3srednja           -0.26091891 0.13146167 -1.984753 5.245946e-02
```

```
Danger3velika           -0.41470506 0.13910099 -2.981324 4.358745e-03
logBodyWt               -0.03304175 0.02522350 -1.309959 1.959673e-01
Danger3srednja:logBodyWt -0.09262741 0.04263438 -2.172599 3.439062e-02
Danger3velika:logBodyWt  -0.06917822 0.03808380 -1.816474 7.506374e-02
```

S hkratnim testiranjem preverimo ničelne domneve o naklonu vsake od treh premic in o razlikah med nakloni:

```
> C<-rbind(c(0,0,0,1,0,0), c(0,0,0,1,1,0), c(0,0,0,1,0,1),
+          c(0,0,0,0,1,0), c(0,0,0,0,0,1), c(0,0,0,0,-1,1))
> rownames(C)<-c("naklon majhna", "naklon srednja", "naklon velika",
+                "naklon srednja-majhna", "naklon velika-majhna","naklon velika-srednja")
> test<- glht(mod.1.int, linfct=C)
> summary(test)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = log(TotalSleep) ~ Danger3 * logBodyWt, data = spanje)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
naklon majhna == 0         -0.03304    0.02522  -1.310  0.53598
naklon srednja == 0        -0.12567    0.03437  -3.656  0.00289 **
naklon velika == 0         -0.10222    0.02853  -3.582  0.00381 **
naklon srednja-majhna == 0 -0.09263    0.04263  -2.173  0.13234
naklon velika-majhna == 0  -0.06918    0.03808  -1.816  0.25761
naklon velika-srednja == 0  0.02345    0.04467   0.525  0.94746
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test)

         Simultaneous Confidence Intervals

Fit: lm(formula = log(TotalSleep) ~ Danger3 * logBodyWt, data = spanje)

Quantile = 2.6141
95% family-wise confidence level


Linear Hypotheses:
                           Estimate lwr      upr
naklon majhna == 0         -0.03304 -0.09898  0.03290
naklon srednja == 0        -0.12567 -0.21552 -0.03581
naklon velika == 0         -0.10222 -0.17681 -0.02763
naklon srednja-majhna == 0 -0.09263 -0.20408  0.01883
naklon velika-majhna == 0  -0.06918 -0.16873  0.03038
naklon velika-srednja == 0  0.02345 -0.09333  0.14023
```

Izbrani model `mod.1.int` predstavlja tri različne premice (Slika 40), z večanjem `logBodyWt` se `log(TotalSleep)` linearno zmanjšuje pri živalih v srednji in veliki nevarnosti. Razlike med nakloni niso statistično značilno različne. Z modelom je pojasnjene 56.5 % variabilnosti `log(TotalSleep)`.



Slika 40: Napovedi za `mod.1.int` s pripadajočimi 95 % intervali zaupanja