

Kazalo

1 Praktični postopek linearnega modeliranja	1
1.1 Seznanitev s podatki	2
1.2 Grafično prikazovanje podatkov	3
1.3 Ocenjevanje parametrov linearnega modela	6
1.4 Diagnostika modela	7
1.5 Iskanje ustreznnejšega linearnega modela, če v prejšnjem koraku predpostavke niso izpolnjene	8
1.6 Obrazložitev rezultatov	14
1.7 Diagnostični grafikoni dodane spremenljivke in parcialnih ostankov	21
1.8 Opisna spremenljivka v linearinem modelu	22
1.9 Dve opisni spremenljivki v modelu	27
1.10 Dve opisni spremenljivki in njuna interakcija v modelu	29
1.11 Številska in dve opisni spremenljivki v modelu	31
1.12 Številska, dve opisni spremenljivki ter njihove interakcije v modelu	34

1 Praktični postopek linearnega modeliranja

V uvodnem poglavju na primeru pokažemo osnovne postopke in pravila statističnega modeliranja. Na prvem mestu moramo jasno opredeliti namen statističnega modeliranja (*descriptive, exploratory, prognostic*). Od namena modeliranja je odvisno, kako bomo zbrali podatke in kako bomo interpretirali rezultate modela. Ko so podatki zbrani, je prvi korak modeliranja seznanitev s podatki. Razmisliti moramo, kako jih bomo ustrezeno matematično predstavili. Katere spremenljivke so številske, katere opisne, kako bomo opisne spremenljivke vključili v linearni model. Pomembno vlogo v tej fazi predstavlja ustrezeni grafični prikazi podatkov.

V nadaljevanju določimo začetno obliko dveh osnovnih komponent statističnega modela: sistematična komponenta in slučajna komponenta. V tej fazi se moramo jasno zavedati namena našega modeliranja, ali gre za opis zveze med odzivno spremenljivko in napovednimi, ali gre za iskanje vzročno-posledične zveze, ali pa za napovedovanje odzivne spremenljivke.

Tej fazi sledi ocenjevanje parametrov statističnega modela in preverjanje izpolnjevanja predpostavk. Fazi preverjanja ustreznosti modela pravimo diagnostika modela.

Ko za izbrane podatke izberemo ustrezen model, sledi interpretacija rezultatov, ki pogosto vključuje grafične prikaze napovedanih vrednosti z ocenami njihove natančnosti.

Primer: pljučna kapaciteta

Primer linearnega modeliranja bomo prikazali na podatkovnem okviru `lungcap` iz paketa `GLMsData`. Podatki so bili zbrani za vzorec 654 otrok in mladostnikov v Bostonu sredi sedemdesetih let prejšnjega stoletja (Kahn in Michael, 2005). Kot primer linearnega modeliranja so bili uporabljeni v knjigi *Generalized Linear Models With Examples in R* (Dunn P. K. in Smyth G. K., 2018).

1.1 Seznanitev s podatki

```
library(GLMsData)
data(lungcap)
str(lungcap)

'data.frame': 654 obs. of 5 variables:
 $ Age    : int 3 4 4 4 4 4 4 5 5 5 ...
 $ FEV    : num 1.072 0.839 1.102 1.389 1.577 ...
 $ Ht     : num 46 48 48 48 49 49 50 46.5 49 49 ...
 $ Gender: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Smoke  : int 0 0 0 0 0 0 0 0 0 0 ...

summary(lungcap)
```

Age	FEV	Ht	Gender	Smoke
Min. : 3.000	Min. :0.791	Min. :46.00	F:318	Min. :0.00000
1st Qu.: 8.000	1st Qu.:1.981	1st Qu.:57.00	M:336	1st Qu.:0.00000
Median :10.000	Median :2.547	Median :61.50		Median :0.00000
Mean : 9.931	Mean :2.637	Mean :61.14		Mean :0.09939
3rd Qu.:12.000	3rd Qu.:3.119	3rd Qu.:65.50		3rd Qu.:0.00000
Max. :19.000	Max. :5.793	Max. :74.00		Max. :1.00000

V naboru podatkov `lungcap` je pet spremenljivk: `Age`, starost v dopolnjenih letih, `FEV`, pljučna kapaciteta v litrih (L), `Ht` telesna višina v palcih (1 palec = 2,54 cm, spremenljivko bomo zaradi predstavljevanosti vrednosti preračunali v cm), `Gender`, spol (F: female, M: male)) in `Smoke`, status kajenja (0: ni kadilec/ni kadilka, 1: kadilec/kadilka). `Smoke` je celoštevilska spremenljivka, čeprav označuje dve kategoriji kajenja. Za nadaljnje delo jo spremenimo v spremenljivko tipa `factor` in oznaki sprememimo v `Ne` in `Da`.

```
lungcap$Ht <- lungcap$Ht*2.54
lungcap$Smoke <- factor(lungcap$Smoke, labels=c("Ne", "Da"))
levels(lungcap$Gender)
```

```
[1] "F" "M"

# zamenjamo oznaki za spol za grafične prikaze
levels(lungcap$Gender) <- c("Ženske", "Moški")
summary(lungcap)
```

Age	FEV	Ht	Gender	Smoke
Min. : 3.000	Min. :0.791	Min. :116.8	Ženske:318	Ne:589
1st Qu.: 8.000	1st Qu.:1.981	1st Qu.:144.8	Moški :336	Da: 65
Median :10.000	Median :2.547	Median :156.2		

Mean : 9.931	Mean : 2.637	Mean : 155.3
3rd Qu.: 12.000	3rd Qu.: 3.119	3rd Qu.: 166.4
Max. : 19.000	Max. : 5.793	Max. : 188.0

V podatkovnem okviru imamo podatke za 654 otrok in mladostnikov, 336 jih je moškega in 318 ženskega spola. V vzorcu je 65 kadilcev, veliko več je nekadilcev (589). Najmlajša oseba je stara 3 leta in najstarejša 19 let, polovica je stara 10 let ali manj, ena četrtina pa nad 12 let. Najmanjši otrok je visok 117 cm, polovica jih je manjših ali enakih 156 cm in največji mladostnik je visok 188 cm.

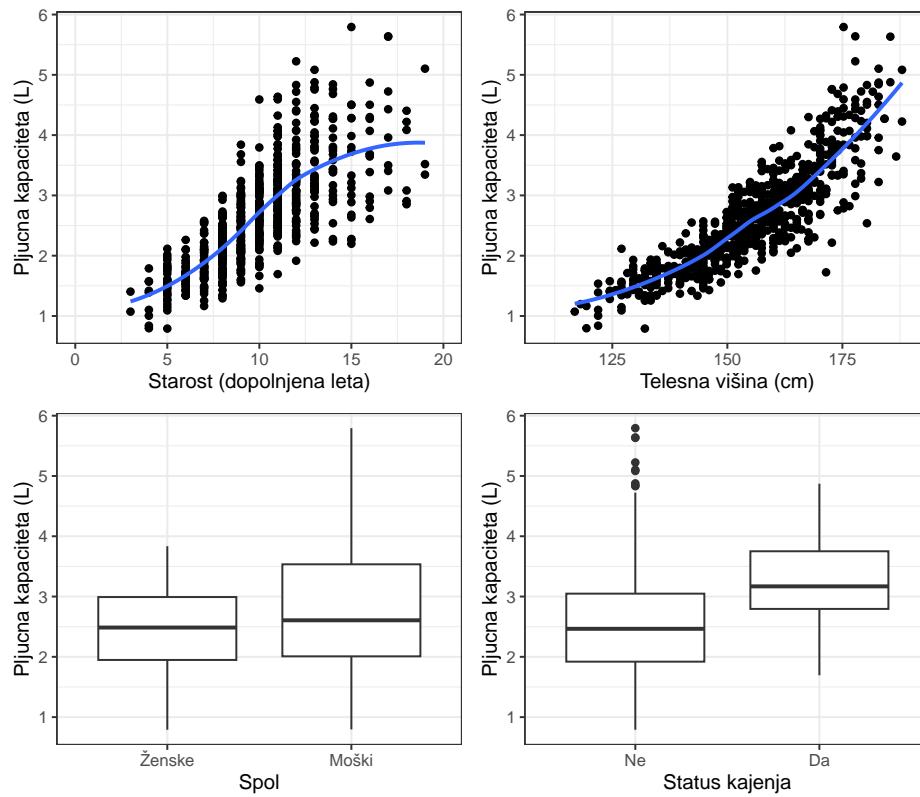
Raziskovalno vprašanje je, kako je pljučna kapaciteta ob upoštevanju spola, starosti in telesne višine, povezana s kajenjem. Raziskava je bila narejena kot opazovalna študija. Ni šlo za načrtovano študijo s kontroliranimi vrednostmi napovednih spremenljivk in temu ustreznim slučajnim izborom otrok in mladostnikov. Vrednosti napovednih spremenljivk niso bile določene vnaprej, ampak so odvisne od enot v vzorcu. Tako pridobljeni podatki omogočajo modeliranje, ki pojasni zvezo med izbranimi napovednimi spremenljivkami in FEV, ne moremo pa oceniti vpliva napovednih spremenljivk na FEV v smislu vzroka in posledice (*cause and effect*). Za modeliranje vzročno-posledičnih zvez pri takih podatkih moramo uporabiti posebne metode, ki jih tekom tega predmeta ne bomo obravnavali.

Glede na raziskovalno vprašanje je **FEV odzivna spremenljivka**, napovedne spremenljivke so štiri: dve **številski**, **Age** in **Ht** ter dve **opisni**, **Gender** in **Smoke**. Obe opisni spremenljivki imata samo dve ravni, kar pomeni, da generirata vsaka po eno umetno/slepo spremenljivko (*dummy variable*).

1.2 Grafično prikazovanje podatkov

Raziščimo odvisnost FEV od napovednih spremenljivk na podlagi grafičnih prikazov (Slika 1).

```
p1 <- ggplot(lungcap, aes(x=Age, y=FEV))+ geom_point() + xlim(c(0,20)) +
  xlab("Starost (dopolnjena leta)") + ylab("Pljučna kapaciteta (L)") + theme_bw() +
  geom_smooth(se=FALSE)
p2 <- ggplot(lungcap, aes(x=Ht, y=FEV))+ geom_point() + xlim(c(110, 190)) +
  xlab("Telesna višina (cm)") + ylab("Pljučna kapaciteta (L)") + theme_bw() +
  geom_smooth(se=FALSE)
p3 <- ggplot(lungcap, aes(x=Gender, y=FEV))+ geom_boxplot() + xlab("Spol") +
  ylab("Pljučna kapaciteta (L)") + theme_bw()
p4 <- ggplot(lungcap, aes(x=Smoke, y=FEV))+ geom_boxplot() + xlab("Status kajenja") +
  ylab("Pljučna kapaciteta (L)") + theme_bw()
ggarrange(p1, p2, p3, p4, ncol=2, nrow=2)
```



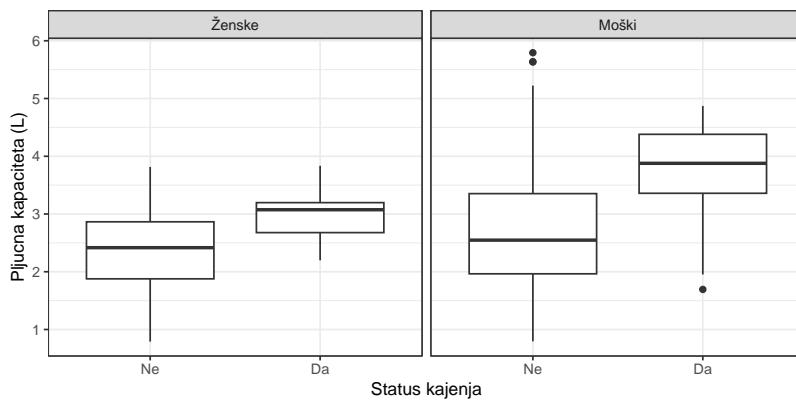
Slika 1: FEV v odvisnosti od Age, Ht (z dodanim gladilnikom), Gender in Smoke

Slika 1 kaže jasno odvisnost FEV od starosti, s starostjo FEV narašča, ne povsem linearno. Tudi telesna višina vpliva na FEV pozitivno, zveza ne izgleda linearna. Tako pri starosti, kot pri telesni višini, se variabilnost FEV z starostjo in s telesno višino povečuje (problem heteroskedastičnosti).

Slika 1 levo spodaj kaže porazdelitev FEV po spolu, vrednosti so nekoliko višje pri moških kot pri ženskah ob tem, da je variabilnost te spremenljivke pri moških precej večja. Mediana FEV kadičev je večja kot pri nekadilcih, kar je malo nenavadno in bi bilo lahko posledica veliko manjšega vzorca za kadičce. Pri interpretaciji teh grafov moramo biti previdni, saj vsak zase prikazuje samo zvezo dveh spremenljivk brez hkratnega upoštevanja vpliva ostalih napovednih spremenljivk. Pri kajenju se pokaže šolski primer *confoundinga*, starost vpliva na pljučno kapaciteto in tudi na status kajenja, zato okvirja z ročaji na Sliki 1 desno spodaj ne odražata prave zveze med FEV in Smoke. V randomizirani študiji bi bila porazdelitev starosti med kadičci in nekadilci enaka, v tem primeru pa ni, zato je bistveno, da starost upoštevamo v modelu. V vzorcih je variabilnost pljučne kapacitete pri nekadilcih precej večja kot pri kadičcih, kar je tudi verjetno povezano s starostjo.

Poglejmo, kakšna je porazdelitev FEV po skupinah določenih s štirimi možnimi kombinacijami vrednosti spremenljivk **Gender** in **Smoke** (Slika 2). Tako pri ženskah kot pri moških je mediana FEV kadičev višja kot pri nekadilcih.

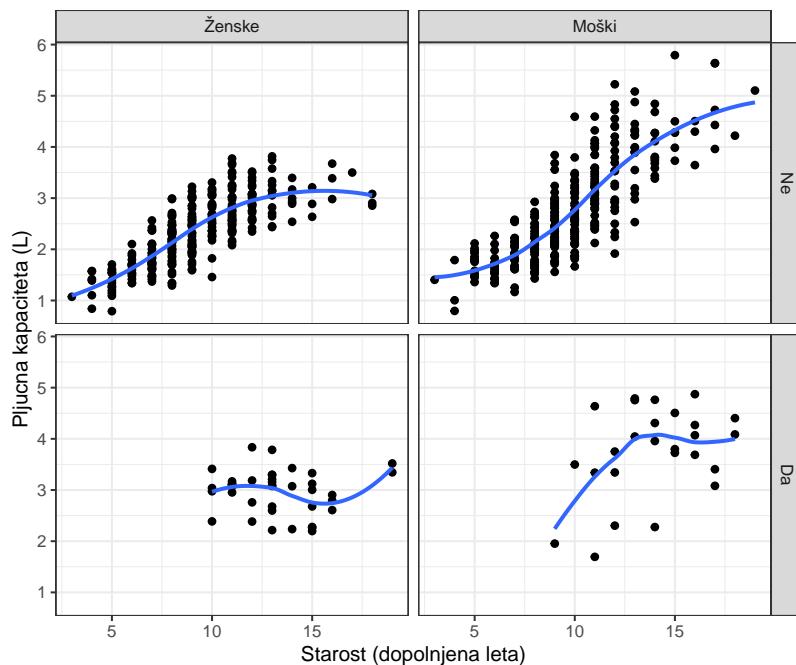
```
ggplot(lungcap, aes(x=Smoke, y=FEV)) + geom_boxplot() + facet_wrap(~ Gender) +
  xlab("Status kajenja") + ylab("Pljučna kapaciteta (L)") + theme_bw()
```



Slika 2: FEV v odvisnosti od Gender in Smoke hkrati

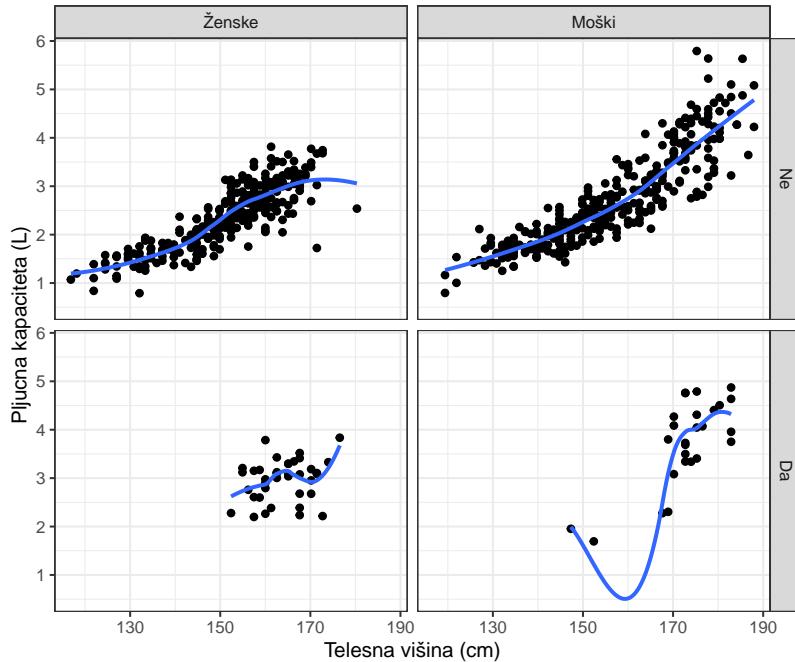
Bolj jasno nelinearno odvisnost FEV od starosti vidimo, če na sliko dodamo gladilnik. Hkratno odvisnost FEV od starosti, spola in statusa kajenja lahko prikažemo tako, da podatke razdelimo v skupine glede na spol in kajenje (Slika 3). Gladilnik pokaže nelinearno odvisnost FEV od starosti tako v skupini nekadilcev kot nekadilk, pri kadilcih in kadilkah pa ni videti neke jasne odvisnosti FEV od starosti. Gladilnika sta v teh dveh skupinah določena z zelo majhnim številom podatkov. Podobno tudi odvisnost FEV od telesne višine ni linearna pri nekadilcih in nekadilkah (Slika 4) in nejasna pri kadilcih in kadilkah.

```
ggplot(lungcap, aes(x=Age, y=FEV)) + geom_point() + geom_smooth(se=FALSE) +
  facet_grid(Smoke ~ Gender) + xlab("Starost (dopolnjena leta)") +
  ylab("Pljučna kapaciteta (L)") + theme_bw()
```



Slika 3: FEV v odvisnosti od Age, Gender in Smoke hkrati

```
ggplot(lungcap, aes(x=Ht, y=FEV)) + geom_point() + geom_smooth(se=FALSE) +
  facet_grid(Smoke ~ Gender) + xlab("Telesna višina (cm)") +
  ylab("Pljučna kapaciteta (L)") + theme_bw()
```



Slika 4: FEV v odvisnosti od Ht, Gender in Smoke hkrati

Na prikazanih slikah, smo videli zveze med FEV in vsako od številskih spremenljivk ob upoštevanju spola in statusa kajenja, še vedno pa ne vemo, kako je FEV povezana s statusom kajenja ob upoštevanju vseh treh ostalih napovednih spremenljivk: starosti, telesne višine in spola. Odgovor na to vprašanje lahko dobimo z analizo linearnega modela za FEV v odvisnosti od vseh štirih napovednih spremenljivk (model multiple regresije).

1.3 Ocenjevanje parametrov linearnega modela

Zapišimo linearni model za i -to osebo:

$$FEV_i = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot Ht_i + \beta_3 \cdot Gender_i + \beta_4 \cdot Smoke_i + \varepsilon_i$$

```
mod1 <- lm(FEV ~ Age + Ht + Gender + Smoke, data=lungcap)
```

Da bomo vedeli, kako sta v model vključeni opisni spremenljivki Gender in Smoke, izpišemo ravnini obih opisnih spremenljivk in nekaj vrstic modelske matrike:

```
levels(lungcap$Gender)
```

```
[1] "Ženske" "Moški"
```

```
levels(lungcap$Smoke)
```

```
[1] "Ne" "Da"
```

Tako imenovana referenčna raven spremenljivke `Gender` je "Ženske" in referenčna skupina spremenljivke `Smoke` je "Ne". V R-ju je v splošnem referenčna vrednost tista, ki je prva po abecedi (enako, kot so po vrsti določene ravni spremenljivke tipa `factor`), ta dobi v slepi spremenljivki vrednost 0. V našem primeru je drugače, ker so bile v osnovi ravni določene glede na angleške izraze vrednosti spremenljivke `Gender` ("F" za female, "M" za male), pri spremenljivki `Smoke`, pa smo vrednosti "0" in "1" prekodirali v "Ne" in "Da".

```
head(model.matrix(mod1)) # prvih šest vrstic modelske matrike X
```

	(Intercept)	Age	Ht	GenderMoški	SmokeDa
1	1	3	116.84	0	0
2	1	4	121.92	0	0
3	1	4	121.92	0	0
4	1	4	121.92	0	0
5	1	4	124.46	0	0
6	1	4	124.46	0	0

```
tail(model.matrix(mod1)) # zadnjih šest vrstic modelske matrike X
```

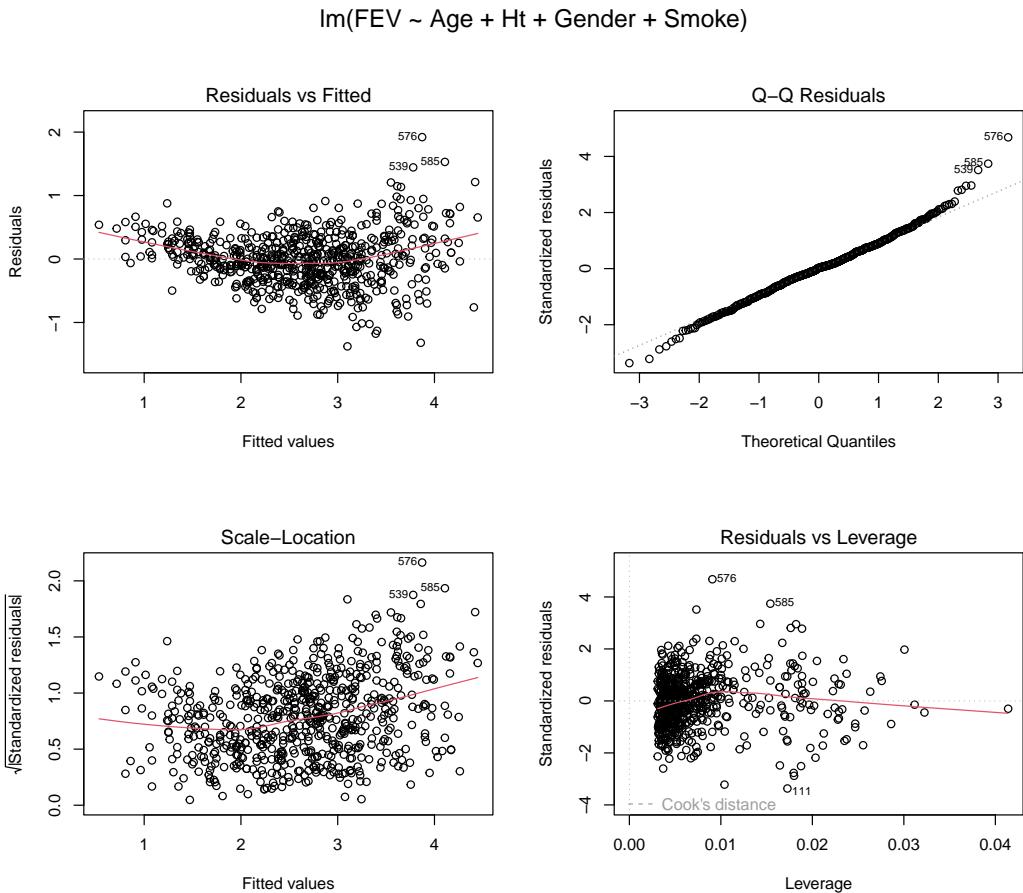
	(Intercept)	Age	Ht	GenderMoški	SmokeDa
649	1	16	176.53	1	1
650	1	16	182.88	1	1
651	1	17	170.18	1	1
652	1	17	175.26	1	1
653	1	18	170.18	1	1
654	1	18	179.07	1	1

Ker sta spremenljivki `Gender` in `Smoke` opisni, vsaka z dvema vrednostma, sta v model vključeni kot slepi spremenljivki `GenderMoški` in `SmokeDa`. Spremenljivka `GenderMoški` ima vrednost 1 za moškega in vrednost 0 za žensko. Spremenljivka `SmokeDa` ima vrednost 1 za kadilca/-ko in vrednost 0 za nekadilca/-ko. Katera vrednost opisne spremenljivke dobi v slepi spremenljivki vrednost 0 ali 1 je odvisno od ravni te vrednosti.

1.4 Diagnostika modela

Preden pogledamo ocene parametrov modela, moramo narediti diagnostiko modela. Diagnostiko za `mod1` naredimo na podlagi grafičnih prikazov ostankov in standardiziranih ostankov.

```
par(mfrow=c(2,2), oma = c(0, 0, 3, 0))
plot(mod1)
```



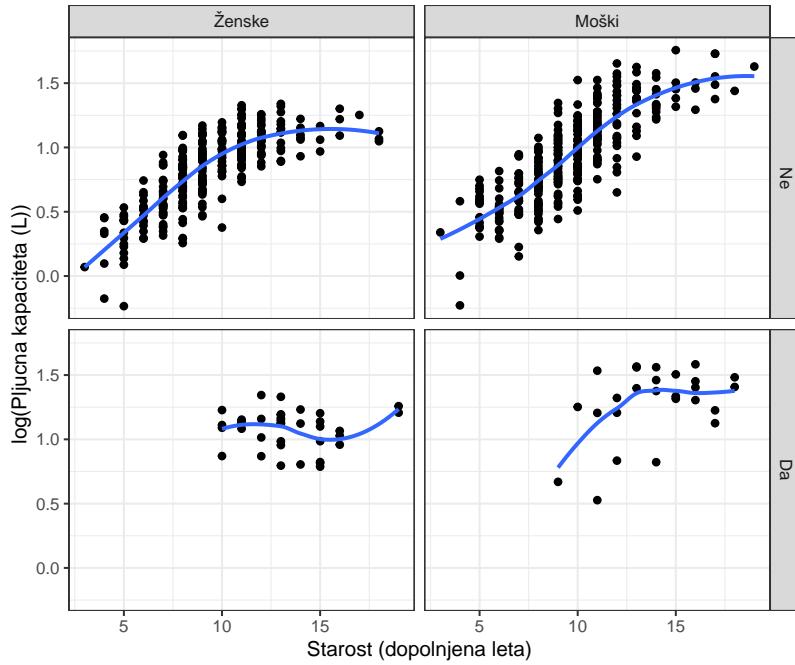
Slika 5: Slike ostankov za mod1

Slika 5 kaže odstopanje od predpostavk linearnega modela: gladilnik na prvi sličici levo zgoraj se ne prilega abscisi, kar odraža nelinearnost. Vidna je tudi nekonstantna varianca, saj je razpršenost točk pri višjih vrednostih z modelom prilagojenih vrednosti \hat{y} (*fitted values*) večja kot pri nižjih vrednostih. Nekonstantna varianca se vidi tudi na spodnji levi sličici, ker gladilnik ni vodoraven. Bistvenega odstopanja porazdelitve standardiziranih ostankov od standardizirane normalne porazdelitve na desni zgornji sličici ni videti. Prav tako ni videti vplivnih točk (Cookova razdalja na desni spodnji sličici ni večja od 1). Model zaradi nelinearnosti in heteroskedastičnosti ni ustrezen.

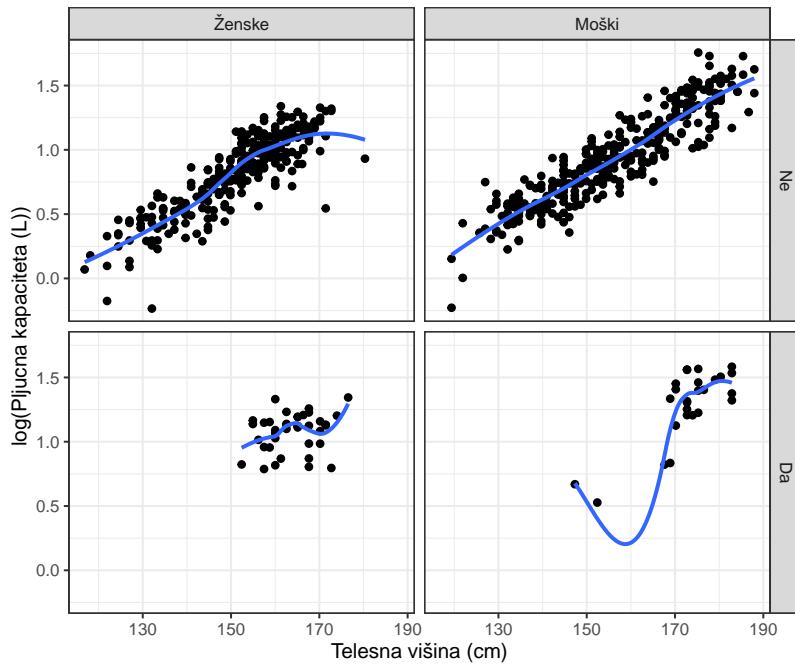
1.5 Iskanje ustrenejšega linearnega modela, če v prejšnjem koraku predpostavke niso izpolnjene

V naslednjem koraku poskusimo modelirati transformirano odzivno spremenljivko: logaritmiramo spremenljivko FEV. Najprej poglejmo grafične prikaze za $\log(\text{FEV})$ (Sliki 6 in 7).

```
ggplot(lungcap, aes(x=Age, y=log(FEV)))+ geom_point() + geom_smooth(se=FALSE) +
  facet_grid(Smoke~ Gender) + xlab("Starost (dopolnjena leta)") +
  ylab("log(Pljučna kapaciteta (L))") + theme_bw()
```

Slika 6: $\log(\text{FEV})$ v odvisnosti od Age, Gender in Smoke hkrati

```
ggplot(lungcap, aes(x=Ht, y=log(FEV))) + geom_point() + geom_smooth(se=FALSE) +
  facet_grid(Smoke ~ Gender) + xlab("Telesna višina (cm)") +
  ylab("log(Pljučna kapaciteta (L))") + theme_bw()
```

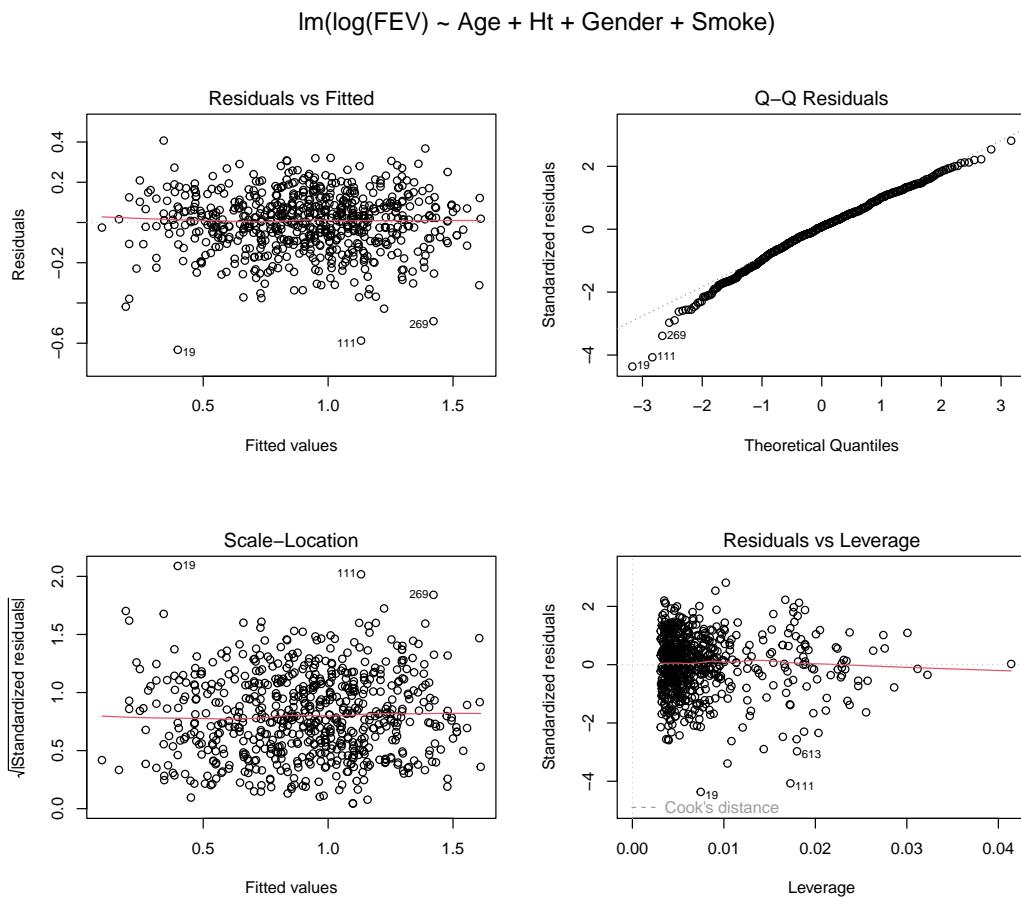
Slika 7: $\log(\text{FEV})$ v odvisnosti od Ht, Gender in Smoke hkrati

Slike 6 in 7 še vedno ne kažeta linearne zveze med $\log(FEV)$ in Age in med $\log(FEV)$ in Ht, težav z nekonstantno varianco pa ni več videti.

Naredimo model za $\log(FEV)$ v odvisnosti od vseh štirih napovednih spremenljivk:

$$\log(FEV_i) = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot Ht_i + \beta_3 \cdot Gender_i + \beta_4 \cdot Smoke_i + \varepsilon_i$$

```
mod2 <- lm(log(FEV) ~ Age + Ht + Gender + Smoke, data=lungcap)
```



Slika 8: Slike ostankov za mod2

Slika 9 ne kaže več kršenja predpostavk linearnega modela.

Peš izračun ocen parametrov modela z matrikami

Poglejmo mod2 s katerim smo modelirali trasformirano odzivno spremenljivko $\log(FEV)$ še v matrični obliki:

```
log(lungcap$FEV) [c(1:3, 654)]
```

```
[1] 0.06952606 -0.17554457 0.09712671 1.48251322
```

```
lungcap$Age[c(1:3, 654)]
```

```
[1] 3 4 4 18
```

```
lungcap$Ht[c(1:3, 654)]
```

```
[1] 116.84 121.92 121.92 179.07
```

```
lungcap$Gender[c(1:3, 654)]
```

```
[1] Ženske Ženske Ženske Moški
```

```
Levels: Ženske Moški
```

```
lungcap$Smoke[c(1:3, 654)]
```

```
[1] Ne Ne Ne Da
```

```
Levels: Ne Da
```

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad (1)$$

$$\mathbf{y} = \begin{pmatrix} 0.0695 \\ -0.1755 \\ \vdots \\ 1.4825 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 3 & 116.8 & 0 & 0 \\ 1 & 4 & 121.9 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 18 & 179.1 & 1 & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

```
Xmat <- model.matrix(~ Age + Ht + Gender + Smoke, data=lungcap)
head(Xmat)
```

	(Intercept)	Age	Ht	GenderMoški	SmokeDa
1	1	3	116.84	0	0
2	1	4	121.92	0	0
3	1	4	121.92	0	0
4	1	4	121.92	0	0
5	1	4	124.46	0	0
6	1	4	124.46	0	0

```
tail(Xmat)
```

	(Intercept)	Age	Ht	GenderMoški	SmokeDa
649	1	16	176.53	1	1
650	1	16	182.88	1	1
651	1	17	170.18	1	1
652	1	17	175.26	1	1
653	1	18	170.18	1	1
654	1	18	179.07	1	1

```
XtX <- t(Xmat) %*% Xmat # t() transponiranje matrike; %*% množenje matrik
y <- log(lungcap$FEV)
inv.XtX <- solve(XtX) # solve() vrne inverzno matriko
XtY <- t(Xmat) %*% y
```

```
beta <- inv.XtX %*% XtY
round(drop(beta), 5)
```

(Intercept)	Age	Ht	GenderMoški	SmokeDa
-1.94400	0.02339	0.01685	0.02932	-0.04607

$$\log(\widehat{FEV}_i) = -1.944 + 0.02339 \cdot Age_i + 0.01685 \cdot Ht_i + 0.02932 \cdot Gender_i - 0.04607 \cdot Smoke_i$$

Učinkovitejša pot računanja ocen parametrov modela je z **direktnim reševanjem sistema linearnih enačb**:

```
beta <- solve(XtX, XtY); round(beta, 5)
```

	[,1]
(Intercept)	-1.94400
Age	0.02339
Ht	0.01685
GenderMoški	0.02932
SmokeDa	-0.04607

Še učinkovitejša pot je uporaba **QR-dekompozicije modelske matrike**:

```
QR <- qr(Xmat)
beta <- qr.coef(QR, y); round(beta, 5)
```

(Intercept)	Age	Ht	GenderMoški	SmokeDa
-1.94400	0.02339	0.01685	0.02932	-0.04607

V vseh treh primerih je rezultat za ocene parametrov enak, funkcija `lm()` uporablja pri izračunu zadnji način izračuna.

Izračunajmo še oceno za varianco napak $\hat{\sigma}^2 = s^2$:

```
y.hat <- Xmat %*% beta
SSost <- sum((y-y.hat)^2); SSost
```

```
[1] 13.73356
s2 <- SSost / (length(lungcap$FEV) - length(beta))
round(c(s=sqrt(s2), s2=s2), 4)
```

s	s2
0.1455	0.0212

Variančno-kovariančna matrika ocen parametrov modela:

```
var.matrix <- s2*inv.XtX; round(var.matrix, 7)
```

	(Intercept)	Age	Ht	GenderMoški	SmokeDa
(Intercept)	0.0061840	0.0001549	-5.0e-05	0.0001390	0.0000422
Age	0.0001549	0.0000112	-1.7e-06	0.0000050	-0.0000208
Ht	-0.0000500	-0.0000017	4.0e-07	-0.0000017	0.0000007

```

GenderMoški  0.0001390  0.0000050 -1.7e-06   0.0001373  0.0000201
SmokeDa      0.0000422 -0.0000208  7.0e-07   0.0000201  0.0004372

var.betaj <- diag(var.matrix)
round(sqrt(var.betaj), 3)

```

(Intercept)	Age	Ht	GenderMoški	SmokeDa
0.079	0.003	0.001	0.012	0.021

Izračunajmo še napoved povprečja $\log(\text{FEV})$ za ženske, ki kadijo, so stare 18 let in visoke 168 cm ter pripadajoči standardni odklon povprečne napovedi:

```

x0.vek <- matrix(c(1, 18, 168, 0, 1), nrow=1) # prva komponenta vektorja je konstanta
y0.x0 <- x0.vek %*% beta
var.y0.x0 <- sqrt(x0.vek %*% solve(t(Xmat) %*% Xmat)) %*% t(x0.vek)*s2
round(c(y0.x0, var.y0.x0, sqrt(var.y0.x0)),3)

```

```
[1] 1.261 0.023 0.153
```

Vse peš izračunane vrednosti dobimo v povzetku linearnega modela, ki ga naredi funkcija `lm()`.

```
names(mod2)
```

```

[1] "coefficients"   "residuals"       "effects"        "rank"
[5] "fitted.values"  "assign"          "qr"             "df.residual"
[9] "contrasts"      "xlevels"         "call"           "terms"
[13] "model"

```

```
names(summary(mod2))
```

```

[1] "call"          "terms"          "residuals"      "coefficients"
[5] "aliased"       "sigma"          "df"             "r.squared"
[9] "adj.r.squared" "fstatistic"    "cov.unscaled"

```

```
round(vcov(mod2), 7)
```

	(Intercept)	Age	Ht	GenderMoški	SmokeDa
(Intercept)	0.0061840	0.0001549	-5.0e-05	0.0001390	0.0000422
Age	0.0001549	0.0000112	-1.7e-06	0.0000050	-0.0000208
Ht	-0.0000500	-0.0000017	4.0e-07	-0.0000017	0.0000007
GenderMoški	0.0001390	0.0000050	-1.7e-06	0.0001373	0.0000201
SmokeDa	0.0000422	-0.0000208	7.0e-07	0.0000201	0.0004372

Standardne napake ocen parametrov modela izračunamo na podlagi diagonalnih elementov variančno-kovariančne matrike ocen parametrov modela:

```
round(sqrt(diag(vcov(mod2))),5)
```

(Intercept)	Age	Ht	GenderMoški	SmokeDa
0.07864	0.00335	0.00066	0.01172	0.02091

Sekvenčni F -testi za model mod2

`anova(mod2)`

```
Analysis of Variance Table
```

```
Response: log(FEV)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	43.210	43.210	2041.9564	< 2.2e-16 ***
Ht	1	15.326	15.326	724.2665	< 2.2e-16 ***
Gender	1	0.153	0.153	7.2451	0.007293 **
Smoke	1	0.103	0.103	4.8537	0.027937 *
Residuals	649	13.734	0.021		

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Katere ničelne domneve se testirajo v vsaki od vrstic zgornjega izpisa, ki ga vrne `anova(mod2)`?

1.6 Obrazložitev rezultatov

Izpišimo povzetek modela in obrazložimo rezultate.

`summary(mod2)`

Call:

```
lm(formula = log(FEV) ~ Age + Ht + Gender + Smoke, data = lungcap)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.63278	-0.08657	0.01146	0.09540	0.40701

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.943998	0.078639	-24.721	< 2e-16 ***
Age	0.023387	0.003348	6.984	7.1e-12 ***
Ht	0.016849	0.000661	25.489	< 2e-16 ***
GenderMoški	0.029319	0.011719	2.502	0.0126 *
SmokeDa	-0.046067	0.020910	-2.203	0.0279 *

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.1455 on 649 degrees of freedom

Multiple R-squared: 0.8106, Adjusted R-squared: 0.8095

F-statistic: 694.6 on 4 and 649 DF, p-value: < 2.2e-16

intervali zaupanja za parametre modela - velikost vpliva posamezne napovedne
spremenljivke ob upoštevanju ostalih napovednih spremenljivk v modelu

`confint(mod2)`

2.5 % 97.5 %

(Intercept)	-2.098414941	-1.789581413
Age	0.016812109	0.029962319
Ht	0.015550757	0.018146715
GenderMoški	0.006308481	0.052330236
SmokeDa	-0.087127344	-0.005007728

Model za povprečno vrednost $\log(FEV)$ zapišemo:

$$\hat{y} = E(\log(FEV)) = -1.944 + 0.023Age + 0.017Ht + 0.029GenderMoski - 0.046SmokeDa. \quad (2)$$

Pomen ocenjenih parametrov modela:

- presečišče $b_0 = -1.944$ predstavlja povprečno vrednost $\log(FEV)$, ko imajo vse napovedne spremenljivke vrednost 0. To je torej presečišče za referenčno skupino ženske nekadilke. Presečišče v `mod2` nima vsebinskega pomena, saj nas ne zanima pljučna kapaciteta novorojenčkov višine 0 cm;
- $b_1 = 0.023$ je ocena parametra, ki pove za koliko se razlikuje povprečna vrednost $\log(FEV)$, pri osebah, ki sta za eno leto narazen ob konstantnih vrednostih ostalih napovednih spremenljivk v modelu. Če želimo obrazložitev podati v osnovnih enotah FEV , torej v litrih, upoštevamo aproksimacijo $E(\log(FEV)) \approx \log(E(FEV))$ in obrazložimo inverzno transformirane parametre: če se Age poveča za 1 leto, se povprečna vrednost FEV poveča za $\exp(b_1) = \exp(0.023) = 1.023$ -krat ob konstantnih vrednostih ostalih napovednih spremenljivk v modelu;
- $b_2 = 0.017$ je ocena parametra, ki pove za koliko se spremeni povprečna vrednost $\log(FEV)$, če se Ht poveča za 1 cm ob konstantnih vrednostih ostalih napovednih spremenljivk v modelu. Ali, če se Ht poveča za 1 cm, se povprečna vrednost FEV poveča za $\exp(b_2) = \exp(0.017) = 1.017$ -krat ob konstantnih vrednostih ostalih napovednih spremenljivk v modelu;
- vpliv Age in Ht na povprečne vrednosti $\log(FEV)$ je v vseh štirih skupinah enak. Modeli za različne skupine se razlikujejo v presečiščih;
- $b_3 = 0.029$ predstavlja razliko med presečiščem v skupini moških in v skupini žensk ne glede na statust kajenja, pri vseh vrednostih Age in Ht . Ker v model ni vključena nobena interakcija med napovednimi spremenljivkami, je to ocena za razliko povprečne vrednosti $\log(FEV)$ med moškimi in ženskami pri katerikoli vrednosti Age in Ht , tako za kadilce kot za nekadilce. Moški imajo v povprečju $\exp(0.029) = 1.029$ -krat večjo povprečno vrednost FEV kot ženske pri katerikoli vrednosti Age in Ht , tako za kadilce kot za nekadilce;
- $b_4 = -0.046$ predstavlja razliko med presečiščem v skupini kadilcev in v skupini nekadilcev ne glede na spol, pri vseh vrednostih Age in Ht . Ker v model ni vključena nobena interakcija med napovednimi spremenljivkami, je to ocena za razliko povprečne vrednosti $\log(FEV)$ med kadilci in nekadilci pri katerikoli vrednosti Age in Ht , tako za moške kot za ženske. Kadilci imajo v povprečju $\exp(-0.046) = 0.955$ -krat manjšo povprečno vrednost FEV kot nekadilci pri katerikoli vrednosti Age in Ht , tako za moške kot za ženske.

Z `mod2` smo modelirali zvezo med $\log(FEV)$ in številskima spremenljivkama Age in Ht za štiri skupine otrok in mladostnikov. Referenčna skupina so **ženske nekadilke**. Za vsako skupino modelske napovedi (2) izračunamo:

- **ženske nekadilke**, $GenderMoki = 0$ in $SmokeDa = 0$:

$$\hat{y} = E(\log(FEV)) = -1.944 + 0.023Age + 0.017Ht.$$

- **ženske kadilke**, $GenderMoki = 0$ in $SmokeDa = 1$:

$$\hat{y} = E(\log(FEV)) = (-1.944 - 0.046) + 0.023Age + 0.017Ht.$$

- **moški nekadilci**, $GenderMoki = 1$ in $SmokeDa = 0$:

$$\hat{y} = E(\log(FEV)) = (-1.944 + 0.029) + 0.023Age + 0.017Ht.$$

- **moški kadilci**, $GenderMoki = 1$ in $SmokeDa = 1$:

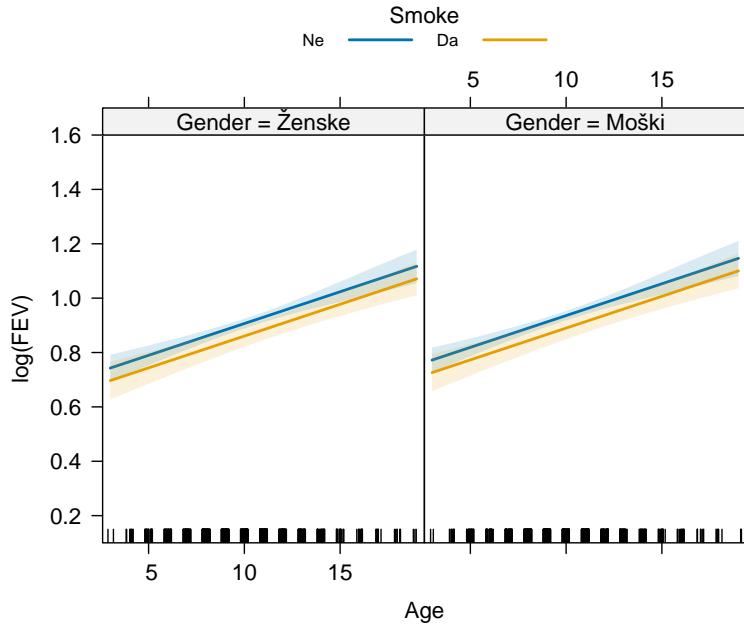
$$\hat{y} = E(\log(FEV)) = (-1.944 + 0.029 - 0.046) + 0.023Age + 0.017Ht.$$

Napovedi za `mod2` so predstavljene na Slikah 9 in 10.

```
mean(lungcap$Ht)
```

```
[1] 155.3047
```

```
library(effects)
plot(Effect(c("Smoke", "Gender", "Age"), mod2), multiline=TRUE,
     ci.style="bands", main="", ylim=c(0.1,1.6))
```

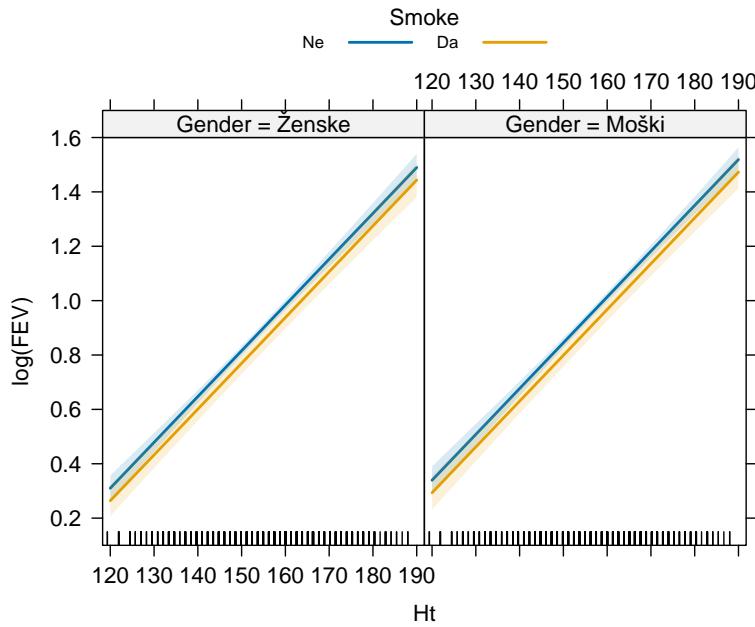


Slika 9: Modelske napovedi za $\log(FEV)$ v odvisnosti od `Age`, `Gender` in `Smoke` hkrati, pri povprečni vrednosti `Ht` za `mod2`

```
mean(lungcap$Age)
```

```
[1] 9.931193
```

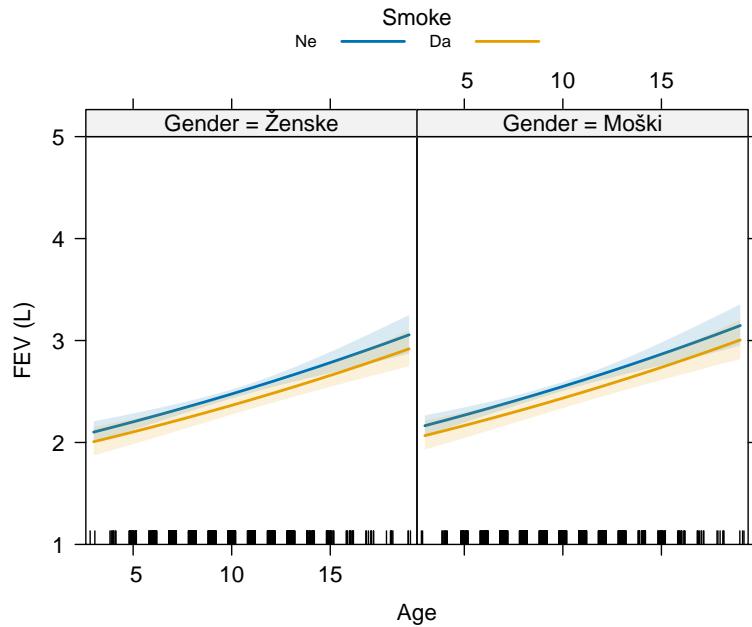
```
plot(Effect(c("Smoke", "Gender", "Ht"), mod2), multiline=TRUE,
    ci.style="bands", main="", ylim=c(0.1,1.6))
```



Slika 10: Modelske napovedi za `log(FEV)` v odvisnosti od `Ht`, `Gender` in `Smoke` hkrati, pri povprečni vrednosti `Age` za `mod2`

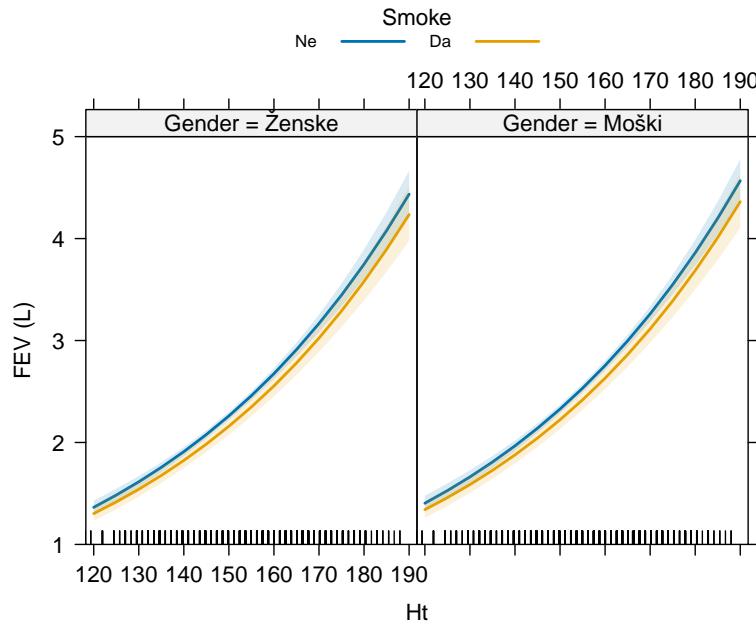
Napovedi za `mod2` originalni skali (FEV (L)) so predstavljene na Slikah 11 in 12.

```
plot(Effect(c("Smoke", "Gender", "Age"), mod2,
            transformation = list(link = log,inverse = exp)),
    axes = list(y = list(lab = "FEV (L)", type = "response")),
    multiline=TRUE, ci.style="bands", main="", ylim=c(1, 5))
```



Slika 11: Modelske napovedi za FEV v odvisnosti od Age, Gender in Smoke hkrati, pri povprečni vrednosti Ht za mod2

```
plot(Effect(c("Smoke", "Gender", "Ht"), mod2,
            transformation = list(link = log,inverse = exp)),
      axes = list(y = list(lab = "FEV (L)", type = "response")), multiline=TRUE,
      ci.style="bands", main="", ylim=c(1, 5))
```



Slika 12: Modelske napovedi za FEV v odvisnosti od Ht, Gender in Smoke hkrati, pri povprečni vrednosti Age za mod2

Izračun povprečne ali posamične napovedi in pripadajoči intervali zaupanja

```
# funkcija predict() ne dela s šummi ki zato ravni Gender spremenimo in ponovimo modeliranje

levels(lungcap$Gender) <- c("Z", "M")
levels(lungcap$Smoke)

[1] "Ne" "Da"

mod2a <- lm(log(FEV) ~ Age + Ht + Gender + Smoke, data=lungcap)

# vrednosti napovednih spremenljivk pri katerih napovedujemo
# povprečno vrednost odzivne spremenljivke
# zapisemo v podatkovni okvir z enakimi imeni spremenljivk

novi.df <- data.frame (Age=c(17, 18, 19), Ht=c(168, 168, 168), Gender=c("Z", "Z", "Z"),
                        Smoke=c("Da", "Da", "Da"))

# povprečne napovedi za log(FEV) s pripadajočimi 95 % IZ

povp.napoved <- predict(mod2a, newdata=novi.df, interval="confidence")
cbind(novi.df, povp.napoved)
```

	Age	Ht	Gender	Smoke	fit	lwr	upr
1	17	168	Z	Da	1.238105	1.195803	1.280406
2	18	168	Z	Da	1.261492	1.215541	1.307442

```

3 19 168      Z   Da 1.284879 1.234681 1.335077
# inverzna transformacija napovedi za FEV in pripadajoči 95 % IZ

cbind(novi.df, round(exp(povp.napoved), 2))

  Age Ht Gender Smoke fit lwr upr
1 17 168      Z   Da 3.45 3.31 3.6
2 18 168      Z   Da 3.53 3.37 3.7
3 19 168      Z   Da 3.61 3.44 3.8

# posamične napovedi za log(FEV) s pripadajočimi 95 % IZ

pos.napoved <- predict(mod2a, newdata=novi.df, interval="prediction")
cbind(novi.df, pos.napoved)

  Age Ht Gender Smoke      fit      lwr      upr
1 17 168      Z   Da 1.238105 0.9493434 1.526866
2 18 168      Z   Da 1.261492 0.9721736 1.550810
3 19 168      Z   Da 1.284879 0.9948558 1.574902

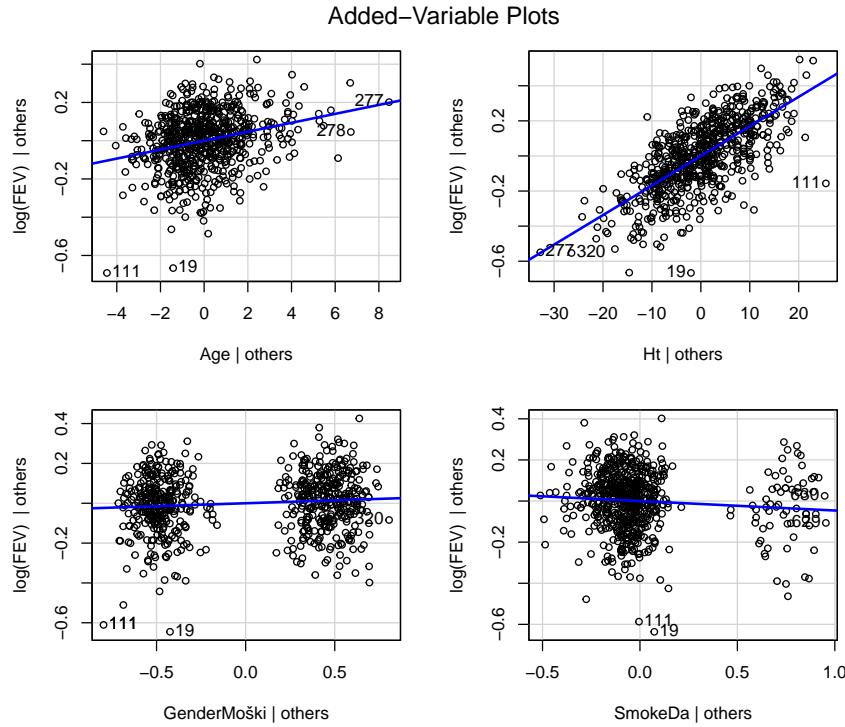
cbind(novi.df, round(exp(pos.napoved), 2))

  Age Ht Gender Smoke fit lwr upr
1 17 168      Z   Da 3.45 2.58 4.60
2 18 168      Z   Da 3.53 2.64 4.72
3 19 168      Z   Da 3.61 2.70 4.83

```

1.7 Diagnostični grafikoni dodane spremenljivke in parcialnih ostankov

```
# library(car)
avPlots(mod2)
```

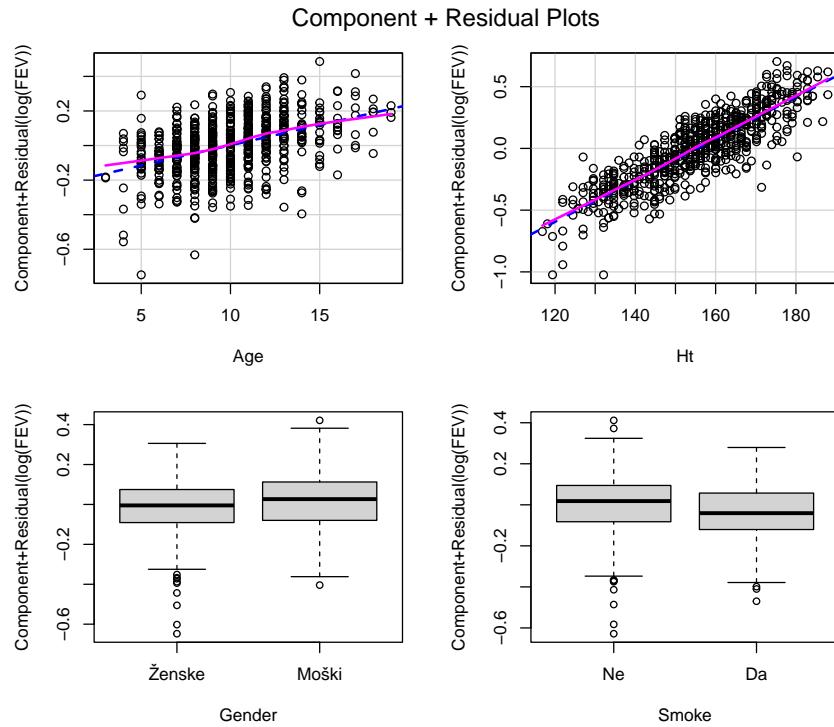


Slika 13: Grafikoni dodane spremenljivke, `avPlots(mod2)`

```
# 3D grafikon za dve številski spremenljivki hkrati
# avPlot3d(mod2, coef1="Age", coef2="Ht")
```

Slika 13 prikazuje grafikone dodane spremenljivke za `mod2`. Leva zgornja sličica prikazuje zvezo med $\log(\text{FEV})$ in `Age` ob upoštevanju ostalih spremenljivk v modelu. Naklon premice je enak oceni parametra za `Age` v `mod2`. Vidimo, da s starostjo $\log(\text{FEV})$ ob upoštevanju ostalih spremenljivk v modelu narašča, označeni sta dve točki z največjo vrednostjo ostanka (19, 111) in dve točki, ki imata največji parcialni vzvod (točki sta najbolj oddaljeni od centra regresorskega prostora za model brez `Age`). Razporeditev točk okoli premice je dokaj enakomerna, ne kaže na prisotnost nekonstantne variance. Podobno lahko komentiramo desno zgornjo sličico za zvezo med $\log(\text{FEV})$ in `Ht` ob upoštevanju ostalih spremenljivk v modelu. Spodnja leva sličica prikazuje zvezo med $\log(\text{FEV})$ in `Gender` ob upoštevanju ostalih spremenljivk v modelu, kaže, da imajo moški v povprečju nekoliko večjo vrednost $\log(\text{FEV})$ kot ženske. Podobno lahko na podlagi desne spodnje sličice rečemo, da imajo kadilci ob upoštevanju vseh ostalih spremenljivk v modelu v povprečju manjšo vrednost $\log(\text{FEV})$ kot nekadilci. Tudi na spodnjih dveh grafikonih je naklon premice enak ocenam parametrov pri `GenderMoški` in pri `SmokeKadilec` za `mod2`.

```
crPlots(mod2)
```



Slika 14: Grafikoni parcialnih ostankov, `crPlots(mod2)`

Slika 14 prikazuje grafikone parcialnih ostankov za posamezen regresor v modelu `mod2`. Za številске regresorje modra črtkana premica prikazuje modelske napovedi $\log(\text{FEV})$ glede na vrednosti posameznega regresorja pri povprečnih vrednostih ostalih regresorjev; točke predstavljajo parcialne ostanke za regresor, ki je na vodoravnji osi. Gladilnik je narisani na podlagi parcialnih ostankov. Gladilnik za parcialne ostanke glede na spremenljivko `Age` se dovolj dobro prilega modelskim napovedim, da lahko privzamemo linearno zvezo med `Age` in $\log(\text{FEV})$ ob upoštevanju `Ht`, `Gender` in `Smoke`. Podobno lahko rečemo za zvezo med $\log(\text{FEV})$ in `Ht`, v tem primeru se gladilnik še bolje prilega napovedanim vrednostim. Ker sta `Gender` in `Smoke` opisni spremenljivki, spodnji dve sličici prikazujeta porazdelitev parcialnih ostankov za vsako od skupin določeno na podlagi opisne spremenljivke. Enako kot na grafikonih dodane spremenljivke (Slika 13), se tudi tu lepo vidi, da imajo moški nekoliko višjo pljučno kapaciteto kot ženske ob upoštevanju starosti, telesne višine ter kajenja. Kadilci pa imajo nekoliko manjšo pljučno kapaciteto kot nekadilci ob upoštevanju ostalih spremenljivk v modelu (primerjajte ta grafikon s prikazom $\log(\text{FEV})$ v odvisnosti od `Smoke`).

1.8 Opisna spremenljivka v linearinem modelu

Ali je povprečna pljučna kapaciteta odvisna od kajenja? Če bi imeli dva slučajna vzorca, enega za kadilce in drugega za nekadilce, bi na to vprašanje odgovorili na podlagi testiranja ničelne domneve o povprečjih:

H_0 : povprečna pljučna kapaciteta kadilcev je enaka povprečni pljučni kapaciteti nekadilcev.

H_1 : povprečna pljučna kapaciteta kadilcev ni enaka povprečni pljučni kapaciteti nekadilcev.

Če za ta primer pozabimo, da so bili podatki lungcap pridobljeni z opazovanjem, ne z načrtovanim izborom kadilcev/kadilk in nekadilcev/nekadilk, lahko zgoraj postavljeno H_0 preverimo z Welchovim t-testom (ne moremo predpostaviti enakih varianc v vzorcih).

```
t.test(FEV~Smoke, data=lungcap, alternative="two.sided", var.equal=FALSE)
```

```
Welch Two Sample t-test

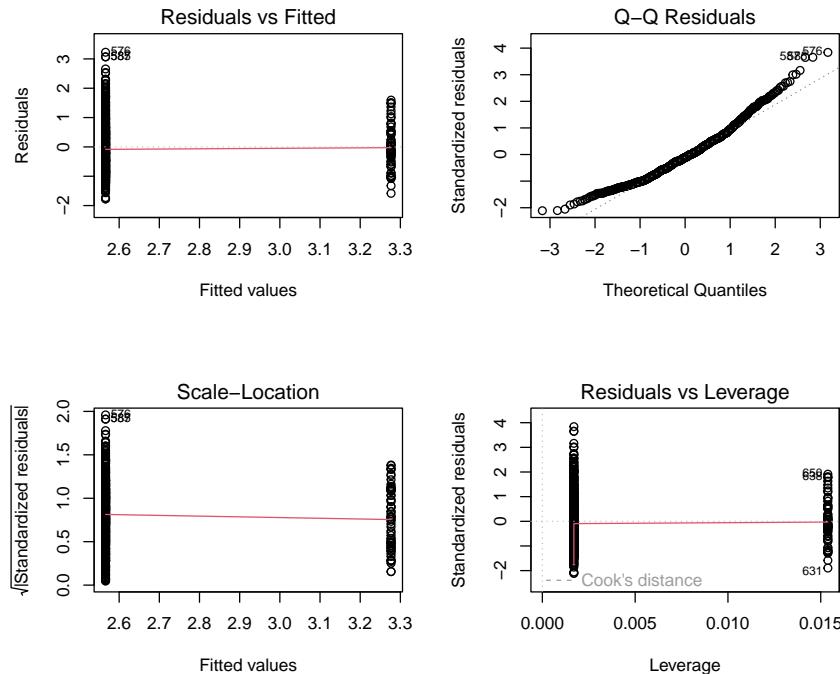
data: FEV by Smoke
t = -7.1496, df = 83.273, p-value = 3.074e-10
alternative hypothesis: true difference in means between group Ne and group Da is not equal to
95 percent confidence interval:
-0.9084253 -0.5130126
sample estimates:
mean in group Ne mean in group Da
2.566143      3.276862
```

Rezultat Welchovega t-testa je statistično značilen ($p < 0,0001$). Pljučna kapaciteta kadilcev je pri 95 % zaupanju od 0,51 L do 0,91 L večja kot pri nekadilcih. Tak rezultat je vsebinsko gledano nepričakovani, iz predhodne analize tega primera pa vemo, da je ta rezultat posledica tega, da v statistični analizi nismo upoštevali drugih dejavnikov, ki tudi vplivajo na FEV.

Isto ničelno domnevo lahko preverimo z linearnim modelom.

```
mod.opisna <- lm(FEV ~ Smoke, data=lungcap)

par(mfrow=c(2,2))
plot(mod.opisna)
```



Slika 15: Diagnostični grafkonci za mod.opisna

Slike ostankov za mod.opisna, v katerega smo vključili eno opisno spremenljivko, kažejo na problem nekonstantne variance. Iz predhodne analize vemo, da je v model potrebno vključiti še druge spremenljivke, odzivno spremenljivko pa je potrebno logaritmirati. Na tem mestu uporabimo mod.opisna za predstavitev pomena parametrov linearnega modela, če je napovedna spremenljivka opisna.

```
mod.opisna$coeff
```

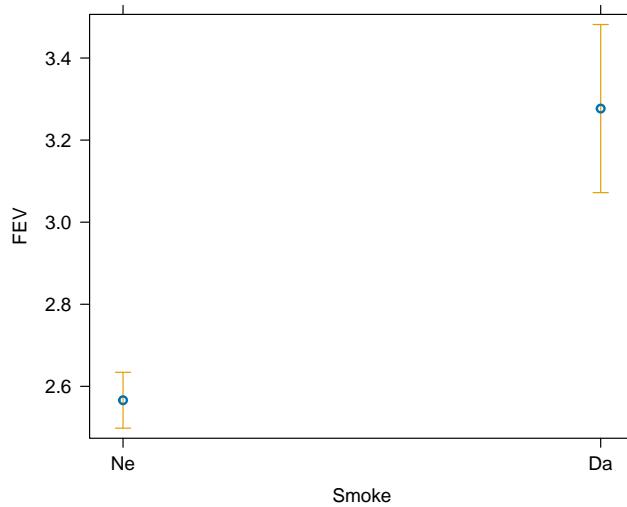
```
(Intercept)      SmokeDa
2.5661426     0.7107189
```

```
summary(mod.opisna)$r.squared
```

```
[1] 0.06023322
```

Povprečna pljučna kapaciteta nekadilcev/nekadilk je 2.57 L. Kadilci/kadilke imajo v povprečju za 0.71 L večjo pljučno kapaciteto kot nekadici/nekadilke (razlika povprečij, ki smo jo dobili pri Welchovem testu: $-2.57 + 3.28 = 0.71$). Intervalov za upanjanja za parametra modela ne izpišemo, ker diagnostika modela tega ne dovoljuje (predpostavke niso izpolnjene). Model pojasnjuje samo 6 % variabilnosti FEV.

```
plot(Effect(c("Smoke")), mod.opisna), multiline=TRUE,
ci.style="bar", main="", lty=0)
```



Slika 16: Modelske napovedi za FEV v odvisnosti od `Smoke` s pripadajočimi 95 % intervali zaupanja za povprečno napoved za `mod.opisna`

Na podlagi spremenljivk `Gender` in `Smoke` naredimo novo spremenljivko `Gender.Smoke` z vrednostmi:

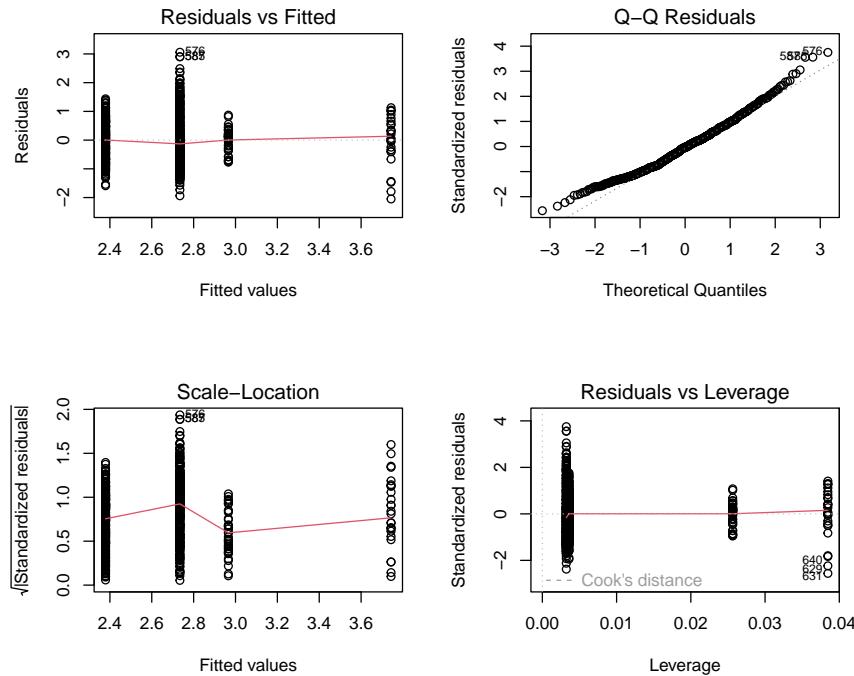
```
lungcap$Gender.Smoke <- lungcap$Gender:lungcap$Smoke
levels(lungcap$Gender.Smoke) # spremenljivka ima 4 vrednosti/kategorije
```

```
[1] "Z:Ne" "Z:Da" "M:Ne" "M:Da"
```

Model za odvisnost FEV od `Gender.Smoke`:

```
mod.opisna.4 <- lm(FEV ~ Gender.Smoke, data=lungcap)
```

```
par(mfrow=c(2,2))
plot(mod.opisna.4)
```



Slika 17: Diagnostični grafkoni za mod.opisna.4

```
mod.opisna.4$coeff
```

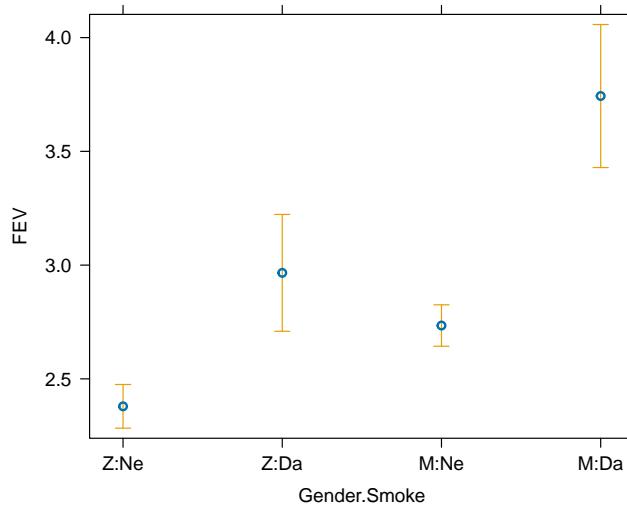
```
(Intercept) Gender.SmokeZ:Da Gender.SmokeM:Ne Gender.SmokeM:Da
2.3792115      0.5867372      0.3551692      1.3640193
```

```
summary(mod.opisna.4)$r.squared
```

```
[1] 0.117164
```

Povprečna pljučna kapaciteta nekadilk je 2.38 L. Kadilke imajo v povprečju za 0.59 L večjo pljučno kapaciteto kot nekadilke. Moški nekadilci imajo v povprečju za 0.36 L večjo pljučno kapaciteto kot nekadilke, moški kadilci imajo v povprečju za 1.36 L večjo pljučno kapaciteto kot nekadilke. Model pojasnjuje 11.7 % variabilnosti FEV.

```
plot(Effect(c("Gender.Smoke")), mod.opisna.4), multiline=TRUE,
ci.style="bar", main="", lty=0)
```



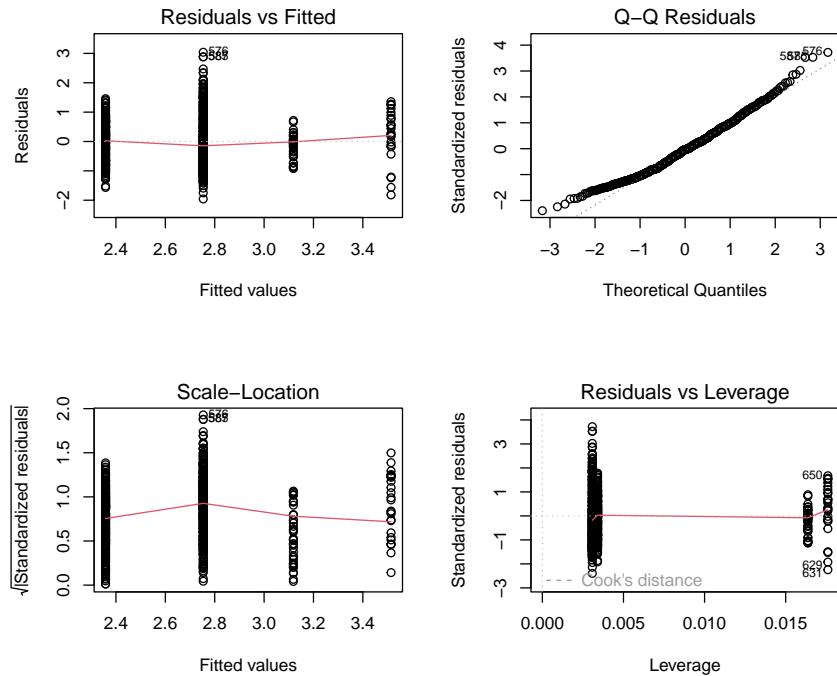
Slika 18: Modelske napovedi za FEV v odvisnosti od `Gender.Smoke` s pripadajočimi 95 % intervali zaupanja za povprečno napoved za `mod.opisna.4`

1.9 Dve opisni spremenljivki v modelu

V model lahko namesto `Gender.Smoke` vključimo dve opisni spremenljivki `Smoke` in `Gender`, najprej predpostavimo, da je zveza med FEV in `Smoke` pri moških in ženskah enaka (ni interakcije med `Smoke` in `Gender`):

```
mod.opisni2 <- lm(FEV ~ Smoke + Gender, data=lungcap)
```

```
par(mfrow=c(2,2))
plot(mod.opisni2)
```

Slika 19: Diagnostični grafikoni za `mod.opisni2`

Diagnostični grafikoni še vedno nakazujejo nekonstantno varianco napak, kljub temu bomo za namen interpretacije parametrov modela izpisali ocene teh parametrov:

```
mod.opisni2$coeff
```

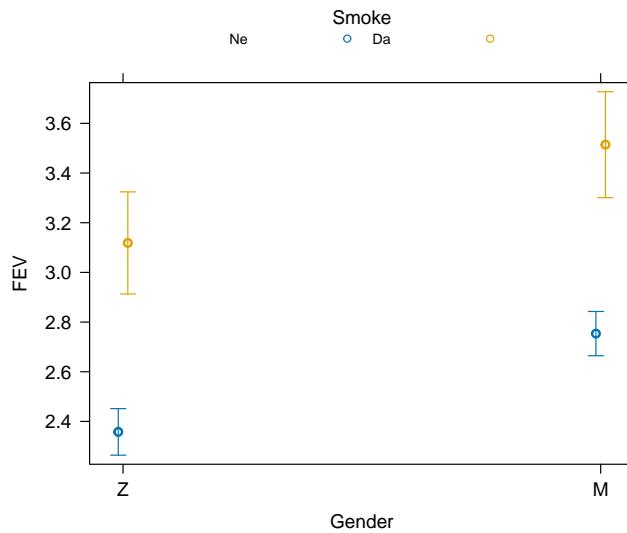
```
(Intercept)      SmokeDa      GenderM
 2.3578761     0.7607029     0.3957065
```

```
summary(mod.opisni2)$r.squared
```

```
[1] 0.1120457
```

Povprečna pljučna kapaciteta nekadilk je 2.36 L. Kadilke/kadilci imajo v povprečju za 0.76 L večjo pljučno kapaciteto kot nekadilke/nekadilci. Moški nekadilci/kadilci imajo v povprečju za 0.40 L večjo pljučno kapaciteto kot nekadilke/kadilke (interakcija med `Gender` in `Smoke` ni predpostavljena). Model pojasnjuje samo 11.2 % variabilnosti `FEV`.

```
plot(Effect(c("Gender", "Smoke"), mod.opisni2), multiline=TRUE,
  ci.style="bar", main="", lty=0)
```



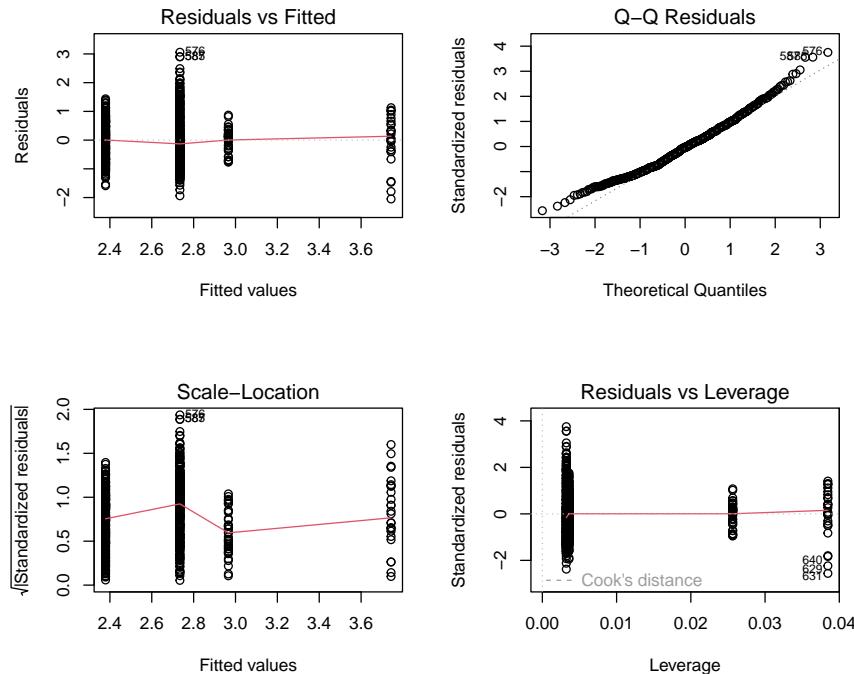
Slika 20: Modelske napovedi za FEV v odvisnosti od `Smoke` in `Gender` s pripadajočimi 95 % intervali zaupanja za povprečno napoved za `mod.opisni.2`

1.10 Dve opisni spremenljivki in njuna interakcija v modelu

Vključimo še interakcijski člen med `Gender` in `Smoke` v model, to pomeni, da predpostavimo, da kajenje drugače vpliva na pljučno kapaciteto pri moških kot pri ženskah:

```
mod.opisni2.int <- lm(FEV~Smoke*Gender, data=lungcap)

par(mfrow=c(2,2))
plot(mod.opisni2.int)
```



Slika 21: Diagnostični grafikoni za mod.opisni2.int

Diagnostični grafikoni še vedno nakazujejo nekonstantno varianco napak, vključitev interakcijskega člena ni bistveno spremenila modela. Kaj parametri modela pomenijo v tem primeru?

```
mod.opisni2.int$coeff
```

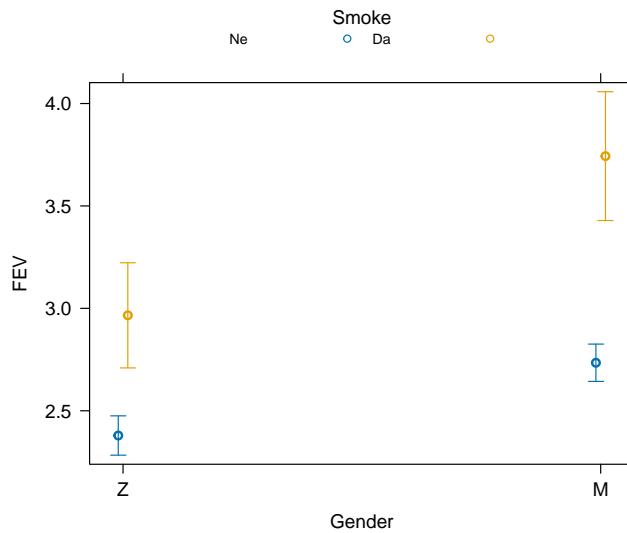
(Intercept)	SmokeDa	GenderM	SmokeDa:GenderM
2.3792115	0.5867372	0.3551692	0.4221129

```
summary(mod.opisni2.int)$r.squared
```

```
[1] 0.117164
```

Povprečna FEV nekadilk je 2.38 L. Kadilke imajo v povprečju za 0.59 L večjo FEV kot nekadilke. Moški nekadilci imajo v povprečju za 0.36 L večjo FEV kot nekadilke, kadilci pa imajo v povprečju za $(0.587+0.355+0.422=1.364)$ L večjo FEV kot nekadilke. Model pojasnjuje 11.7 % variabilnosti FEV, vključitev interakcijskega člena ni vplivala na bistveno povečanje pojasnjene variabilnosti FEV. Ta model je enakovreden modelu mod.opisna.4, razlikuje se le v pomenu zadnjega parametra.

```
plot(Effect(c("Gender", "Smoke"), mod.opisni2.int), multiline=TRUE,
     ci.style="bar", main="", lty=0)
```



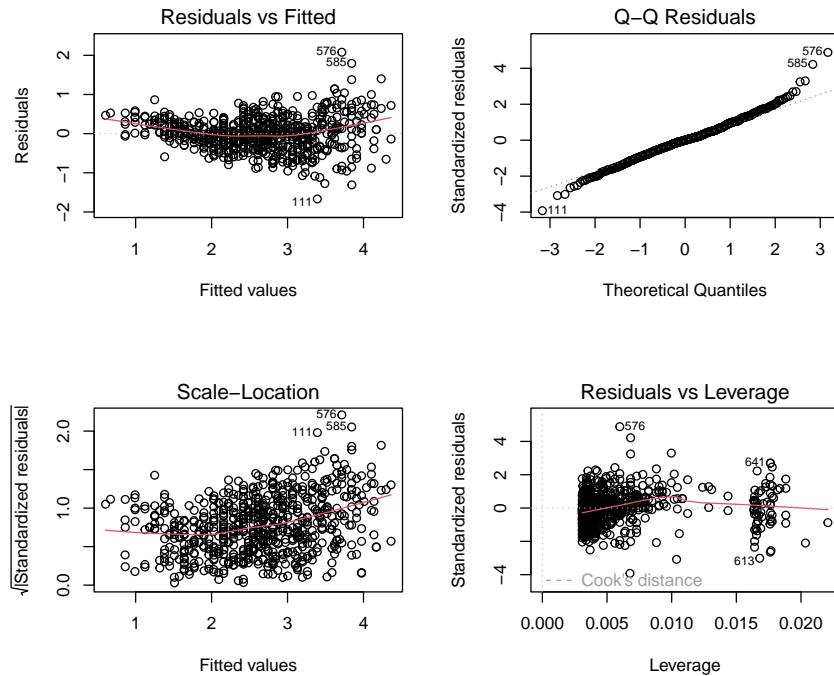
Slika 22: Modelske napovedi za FEV v odvisnosti od `Smoke` in `Gender` ter njune interakcije s pripadajočimi 95 % intervali zaupanja za `mod.opisni.2.int`

1.11 Številska in dve opisni spremenljivki v modelu

V model za FEV poleg `Gender` in `Smoke` vključimo še številsko spremenljivko `Ht`? Vemo že, da je zveza med `FEV` in `Ht` očitna, vendar ne linearна. Kako se to odraža, če sta v modelu tudi spremenljivki `Gender` in `Smoke`?

```
mod3 <- lm(FEV ~ Gender + Smoke + Ht, data=lungcap) # brez interakcij
```

```
par(mfrow=c(2,2))
plot(mod3)
```



Slika 23: Diagnostični grafikoni za mod.opisni2

Z vključitvijo številske spremenljivke Ht v model, se je porazdelitev ostankov in standardiziranih ostankov na diognostičnih grafikonih precej spremenila. Prisotna je nelinearna zveza med ostanki in prilagojenimi vrednostmi ter nekonstantna varianca napak. Vseeno poglejmo pomen parametrov modela.

```
mod3$coeff
```

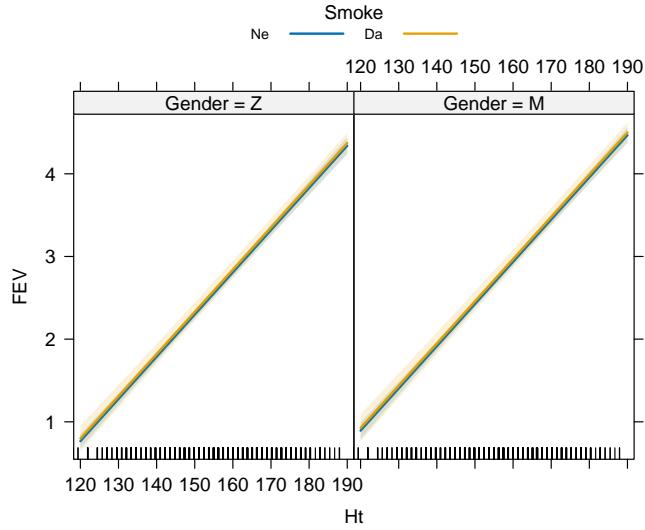
(Intercept)	GenderM	SmokeDa	Ht
-5.36207814	0.12764341	0.03413801	0.05106019

```
summary(mod3)$r.squared
```

```
[1] 0.7588628
```

Ocena presečišča izgubi vsebinski pomen, saj odraža povprečno FEV pri $Ht=0$ za nekadilke. V tem modelu je predpostavljeno, da je zveza med FEV in Ht za vse štiri skupine določene glede na **Gender** in **Smoke** enaka (vzporedne premice). Z upoštevanjem telesne višine v modelu, je postala razlika med napovedanimi vrednostmi za kadihelce in nekadilce minimalna, ampak še vedno pozitivna. Modelske napovedi geometrijsko predstavljajo 4 vzporedne premice.

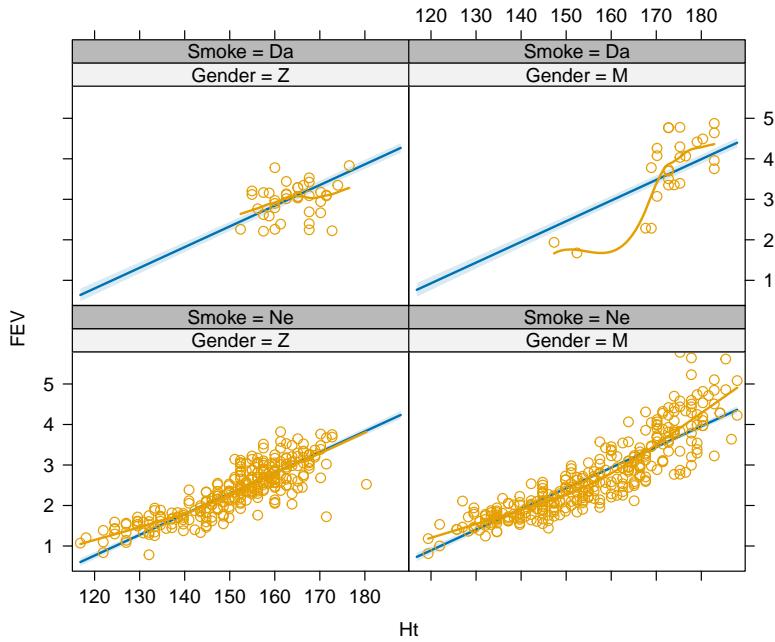
```
plot(Effect(c("Ht", "Smoke", "Gender"), mod3), multiline=TRUE, ci.style="band", main="")
```



Slika 24: Modelske napovedi za FEV v odvisnosti od Ht, Gender in Smoke za mod3

Kot diagnostiko modela poglejmo še grafikon parcialnih ostankov za ta model:

```
plot(Effect(c("Ht", "Gender", "Smoke"), mod3, partial.residuals=TRUE), main="")
```



Slika 25: Modelske napovedi za FEV v odvisnosti od Ht, Gender in Smoke s parcialnimi ostanki

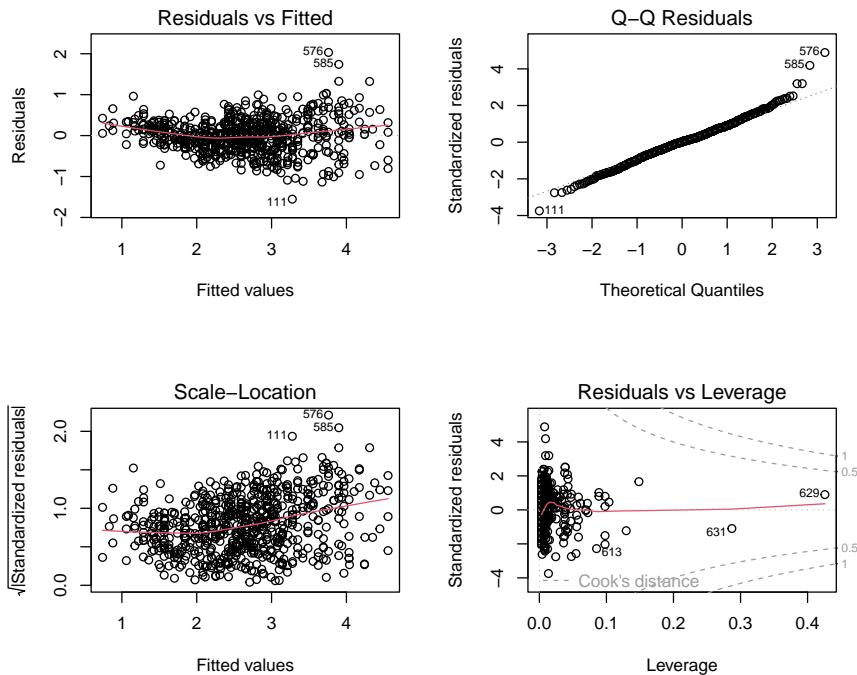
Grafikon kaže prisotnost nelinearnosti, pa tudi drugačno odvisnost FEV od Ht v skupinah Kadi, NeKadi pri moških in pri ženskah. To bi lahko pomenilo, da interakcijski člen med Ht, Smoke in Gender pojasni pomemben del variabilnosti FEV.

1.12 Številska, dve opisni spremenljivki ter njihove interakcije v modelu

V model vključimo interakcijske člene med Ht in $Smoke$ in $Gender$ (tri dvojne in ena trojna interakcija) - predpostavimo, da je zveza med FEV in Ht različna pri kadilcih in nekadilcih, ta razlika je različna pri moških in pri ženskah.

```
mod3.int <- lm(FEV ~ Gender * Smoke * Ht, data=lungcap)
# enak model lahko na dolgo zapišemo:
# mod3.int <- lm(FEV ~ Gender + Smoke + Ht +
#                  Gender : Smoke + Gender : Ht + Smoke : Ht +
#                  Gender : Smoke : Ht, data=lungcap)

par(mfrow=c(2,2))
plot(mod3.int)
```



Slika 26: Diagnostični grafikoni za mod3.int

```
summary(mod3.int)
```

Call:

```
lm(formula = FEV ~ Gender * Smoke * Ht, data = lungcap)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.55289	-0.25070	0.00711	0.24854	2.03200

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.398334	0.315838	-13.926	< 2e-16 ***
GenderM	-1.309992	0.393214	-3.331	0.000913 ***
SmokeDa	4.377015	1.934946	2.262	0.024024 *
Ht	0.044767	0.002080	21.526	< 2e-16 ***
GenderM:SmokeDa	-8.965794	2.625769	-3.415	0.000679 ***
GenderM:Ht	0.009264	0.002559	3.620	0.000318 ***
SmokeDa:Ht	-0.026547	0.011820	-2.246	0.025048 *
GenderM:SmokeDa:Ht	0.053738	0.015663	3.431	0.000640 ***

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 '	1		

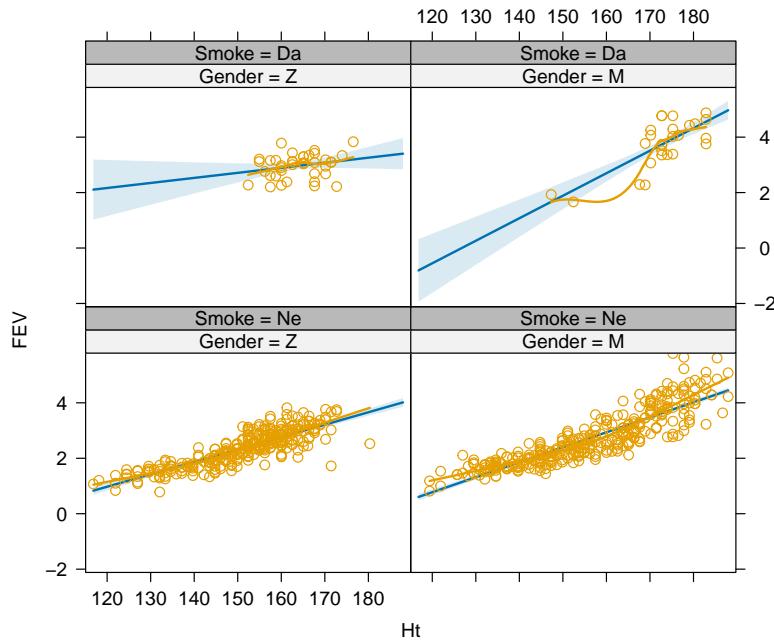
Residual standard error: 0.4173 on 646 degrees of freedom

Multiple R-squared: 0.7708, Adjusted R-squared: 0.7683

F-statistic: 310.4 on 7 and 646 DF, p-value: < 2.2e-16

Z vključitvijo vseh interakcij v model smo pojasnili približno 1 % variabilnosti FEV več. Diagnostika modela na podlagi ostankov pokaže, da so ostanki bližje danim predpostavkam, še vedno imamo dokaj očitno prisotnost nekonstantne variance napak. Grafikoni parcialnih ostankov so tudi ustreznnejši:

```
plot(Effect(c("Ht", "Gender", "Smoke"), mod3.int, partial.residuals=TRUE), main="")
```

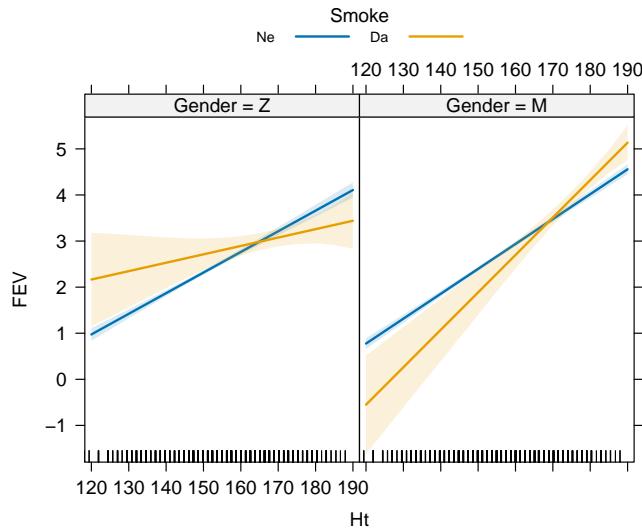


Slika 27: Modelske napovedi za FEV v odvisnosti od Ht, Gender in Smoke s parcialnimi ostanki za mod3.int

Kaj v tem modelu pomenijo ocenjeni parametri? Model mod3.int geometrijsko predstavlja štiri

različne premice (Slika 28).

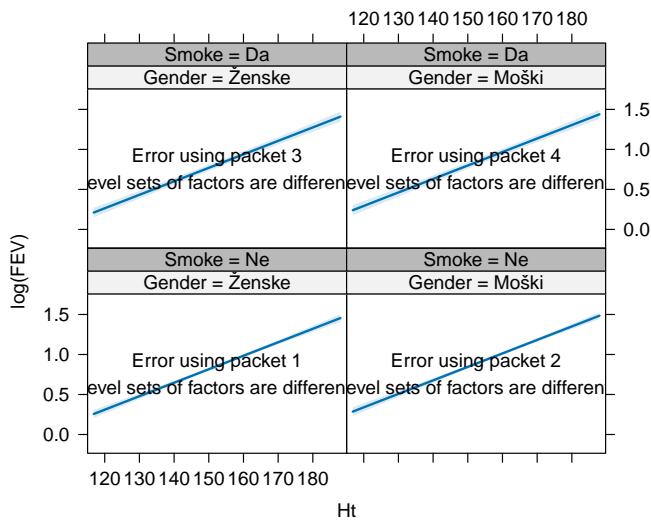
```
plot(Effect(c("Ht", "Smoke", "Gender"), mod3.int), multiline=TRUE,
  ci.style="band", main="")
```



Slika 28: Modelske napovedi za FEV v odvisnosti od Ht, Gender in Smoke za mod3.int

Zdaj pa se vrnimo k modelu za transformirano spremenljivko $\log(\text{FEV})$ (mod2). Ali bi morali tudi v ta model vključiti interakcijske člene? Za vajo uporabite grafikone parcialnih ostankov za mod2 (`plot(Effect(..., mod2, partial.residuals=TRUE))`), da grafično ocenite, ali je vključitev interakcijskih členov potrebna.

```
plot(Effect(c("Ht", "Gender", "Smoke"), mod2, partial.residuals=TRUE), main="")
```

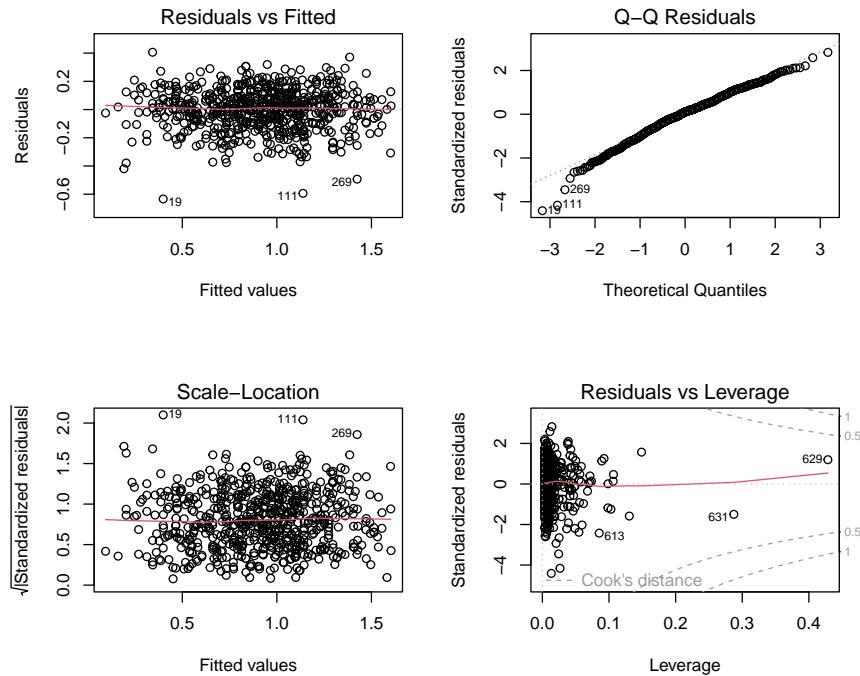


Slika 29: Modelske napovedi za FEV v odvisnosti od Ht, Gender in Smoke za mod2 s parcialnimi ostanki

Naredimo model, ki vključuje Age brez interakcijskih členov z drugimi spremenljivkami, Ht, Gender in Smoke pa z vsemi možnimi interakcijami.

```
mod2.int <- lm(log(FEV) ~ Age+Ht*Gender*Smoke, data=lungcap)

par(mfrow=c(2,2))
plot(mod2.int)
```



Slika 30: Diagnostični grafkoni za mod2.int

Slika 30 kaže, da na podlagi diagnostike ostankov mod2.int ne vidimo kršenja predpostavk linearnega modela.

Zanima nas, ali je model z interakcijskimi členi mod2.int boljši od mod2.

```
anova(mod2,mod2.int)
```

Analysis of Variance Table

```
Model 1: log(FEV) ~ Age + Ht + Gender + Smoke
Model 2: log(FEV) ~ Age + Ht * Gender * Smoke
Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     649 13.734
2     645 13.492  4   0.24109 2.8813 0.02203 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-test za primerjavo dveh gnezdenih modelov (mod2 in mod2.int) pokaže, da interakcijski členi pojasnijo statistično pomemben del variabilnosti log(FEV). Modela mod2.int in mod2 nista ekvivalentna ($p = 0.022$), boljši je kompleksnejši mod2.int.

Poglejmo še rezultate sekvenčnih F-testov:

```
anova(mod2.int)
```

Analysis of Variance Table

Response: log(FEV)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	43.210	43.210	2065.6327	< 2.2e-16 ***
Ht	1	15.326	15.326	732.6643	< 2.2e-16 ***
Gender	1	0.153	0.153	7.3291	0.0069645 **
Smoke	1	0.103	0.103	4.9100	0.0270502 *
Ht:Gender	1	0.006	0.006	0.3029	0.5822792
Ht:Smoke	1	0.001	0.001	0.0490	0.8248592
Gender:Smoke	1	0.001	0.001	0.0269	0.8697814
Ht:Gender:Smoke	1	0.233	0.233	11.1463	0.0008904 ***
Residuals	645	13.492	0.021		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sekvenčni F -testi pokažejo, da je ob upoštevanju vseh spremenljivk in dvojnih interakcij v modelu statistično značilna trojna interakcija Ht:Gender:Smoke. To pomeni, da je zveza med Ht in log(FEV) ob upoštevanju Age različna v štirih skupinah določenih glede na Gender in Smoke (Slika 33). Zveza med log(FEV) in Age, ob upoštevanju Ht pa je v vseh štirih skupinah enaka (Slika 31).

`summary(mod2.int)`

Call:

`lm(formula = log(FEV) ~ Age + Ht * Gender * Smoke, data = lungcap)`

Residuals:

Min	1Q	Median	3Q	Max
-0.63367	-0.08785	0.01486	0.09508	0.40608

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.9667927	0.1236301	-15.909	< 2e-16 ***
Age	0.0224068	0.0033934	6.603	8.42e-11 ***
Ht	0.0170652	0.0009311	18.327	< 2e-16 ***
GenderM	0.0451366	0.1368349	0.330	0.741612
SmokeDa	1.6253091	0.6816185	2.384	0.017391 *
Ht:GenderM	-0.0001156	0.0008915	-0.130	0.896872
Ht:SmokeDa	-0.0102266	0.0041575	-2.460	0.014163 *
GenderM:SmokeDa	-3.0417188	0.9146070	-3.326	0.000932 ***
Ht:GenderM:SmokeDa	0.0182215	0.0054578	3.339	0.000890 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

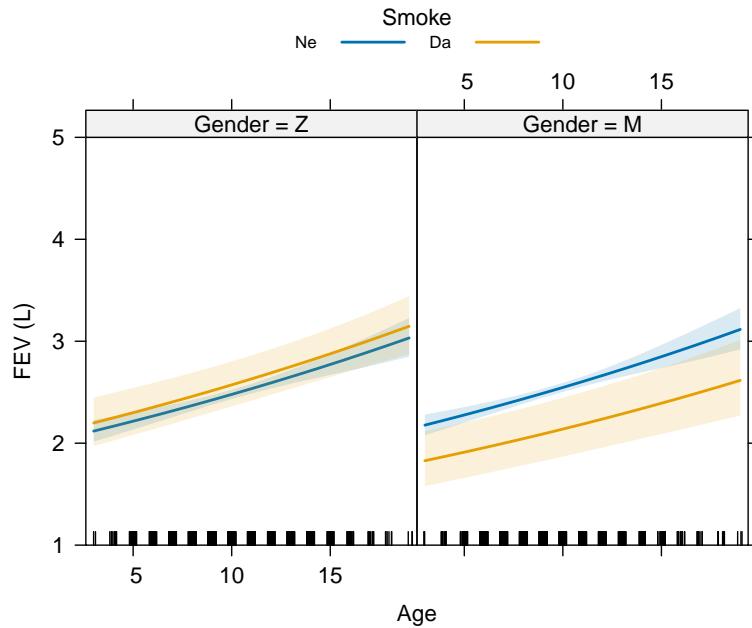
Residual standard error: 0.1446 on 645 degrees of freedom

Multiple R-squared: 0.814, Adjusted R-squared: 0.8117

F-statistic: 352.8 on 8 and 645 DF, p-value: < 2.2e-16

`plot(Effect(c("Smoke", "Gender", "Age")), mod2.int,`
`transformation = list(link = log, inverse = exp)),`

```
axes = list(y = list(lab = "FEV (L)", type = "response")),
multiline=TRUE, ci.style="bands", main="", ylim=c(1, 5))
```

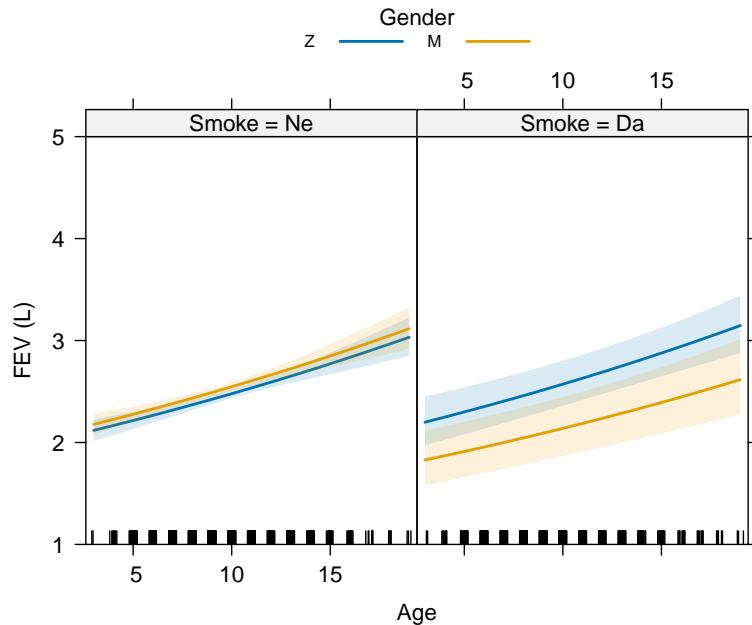


Slika 31: Modelske napovedi za FEV v odvisnosti od Age, Gender in Smoke hkrati, pri povprečni vrednosti Ht za mod2.int

```
mean(lungcap$Ht)
```

```
[1] 155.3047
```

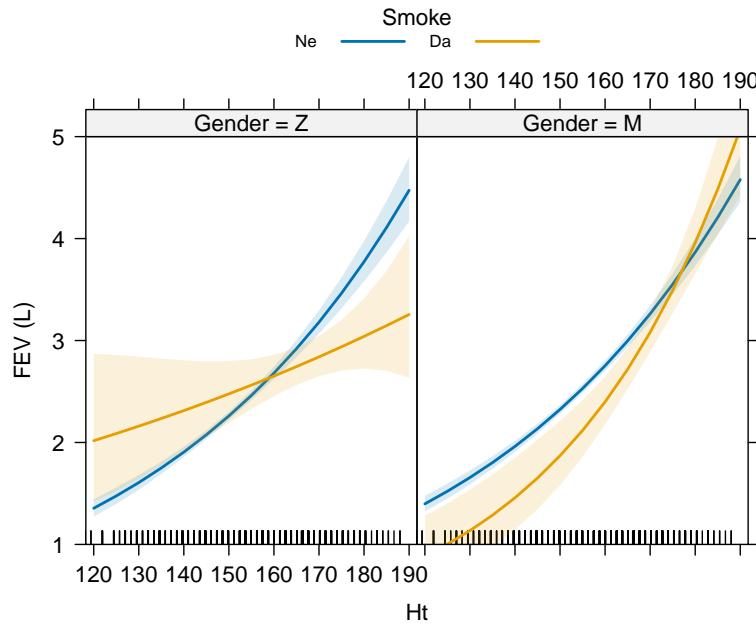
```
plot(Effect(c("Gender", "Smoke", "Age"), mod2.int,
            transformation = list(link = log, inverse = exp)),
      axes = list(y = list(lab = "FEV (L)", type = "response")),
      multiline=TRUE, ci.style="bands", main="", ylim=c(1, 5))
```



Slika 32: Modelske napovedi za FEV v odvisnosti od `Gender`, `Age`, in `Smoke` hkrati, pri povprečni vrednosti `Ht` za `mod2.int`

Slika 31 prikazuje, da je zveza med FEV in `Age` pri povprečni vrednosti `Ht` pozitivna in skoraj linearna, pri ženskah ni pomembne razlike med napovedmi za kadilke in nekadilke, pri moških pa je ta razlika večja, nekadilci imajo večjo napovedano vrednost FEV kot kadilci. Na Sliki 32 se bolj jasno vidi primerjavo napovedi po spolu. Rezultat je čuden - za kadilke model napove večjo FEV kot za kadilce. Kako te nenavadne napovedi pojasnjuje dejstvo, da so izračunane pri povprečni telesni višini 155.3 cm?

```
plot(Effect(c("Smoke", "Gender", "Ht"), mod2.int,
            transformation = list(link = log,inverse = exp)),
      axes = list(y = list(lab = "FEV (L)", type = "response")), multiline=TRUE,
      ci.style="bands", main="", ylim=c(1, 5))
```



Slika 33: Modelske napovedi za FEV v odvisnosti od Ht, Gender in Smoke hkrati, pri povprečni vrednosti Age za `mod2.int`

```
mean(lungcap$Age)
```

```
[1] 9.931193
```

Na Sliki 33 vidimo napovedi FEV glede na Ht, Gender in Smoke pri povprečni starosti 9.9 let. Vidimo, da je zveza med Fev in Ht eksponentno naraščajoča. Širine 95 % intervalov zaupanja za povprečno napoved so odvisne od števila podatkov v posamezni skupini in od oddaljenosti od povprečne telesne višine. Na sliki vidimo prisotnost interakcije med Ht in Gender ter med Ht in Smoke saj krivulje niso vzporedne. Tudi rezultat na tej sliki je videti malo nenavaden, ker prikazuje napovedi pri povprečni starosti 9.9 let. Raziščite, kako bi uporabili funkcijo `Effect()`, da bi se napovedi izračunale pri bolj primerni starosti, glede na to, da proučujemo vpliv kajenja na pljučno kapaciteto ob upoštevanju ostalih spremenljivk v modelu.

Pomen ocenjenih parametrov modela `mod2.int`:

- $b_0 = -1.967$ predstavlja povprečno vrednost $\log(\text{FEV})$, ko imajo vsi regresorji vrednost 0. To je torej presečišče za referenčno skupino ženske nekadilke. Presečišče v `mod2.int` nima vsebinskega pomena, saj nas ne zanima pljučna kapaciteta novorojenčkov višine 0 cm;
- $b_1 = 0.022$ pove za koliko se spremni povprečna vrednost $\log(\text{FEV})$, če se `Age` poveča za 1 leto ob konstantnih vrednostih ostalih regresorjev v modelu, kar pomeni, da je ta zveza enaka v vseh štirih skupinah določenih glede na `Gender` in `Smoke` pri konstantni vrednosti Ht. Obrazložitev v osnovnih enotah `FEV`: če se `Age` poveča za 1 leto, se povprečna vrednost FEV poveča za $\exp(b_1) = \exp(0.0224) = 1.0226$ -krat ob konstantnih vrednostih ostalih napovednih spremenljivk v modelu;
- $b_2 = 0.017$ je ocena parametra, ki pove za koliko se spremni povprečna vrednost $\log(\text{FEV})$, če se Ht poveča za 1 cm ob konstantnih vrednostih ostalih napovednih spremenljivk v modelu,

kar pomeni pri konstantni vrednosti **Age** za ženske nekadilke;

- vpliv **Ht** na povprečne vrednosti $\log(\text{FEV})$ je različen v vsaki od štirih skupin določenih glede na **Gender** in **Smoke**. Modeli za štiri skupine se razlikujejo v presečiščih in v naklonih glede na **Ht** pri konstantni vrednosti **Age**. Geometrijsko model predstavlja štiri ravnine:

- **ženske nekadilke**, $\text{GenderM} = 0$ in $\text{SmokeDa} = 0$:

$$\hat{y} = -1.967 + 0.0224\text{Age} + 0.017\text{Ht}.$$

- **ženske kadilke**, $\text{GenderM} = 0$ in $\text{SmokeDa} = 1$:

$$\hat{y} = (-1.967 + 1.625) + 0.0224\text{Age} + (0.017 - 0.010)\text{Ht}.$$

- **moški nekadilci**, $\text{GenderM} = 1$ in $\text{SmokeDa} = 0$:

$$\hat{y} = (-1.967 + 0.045) + 0.0224\text{Age} + (0.017 - 0.0001)\text{Ht}.$$

- **moški kadilci**, $\text{GenderM} = 1$ in $\text{SmokeDa} = 1$:

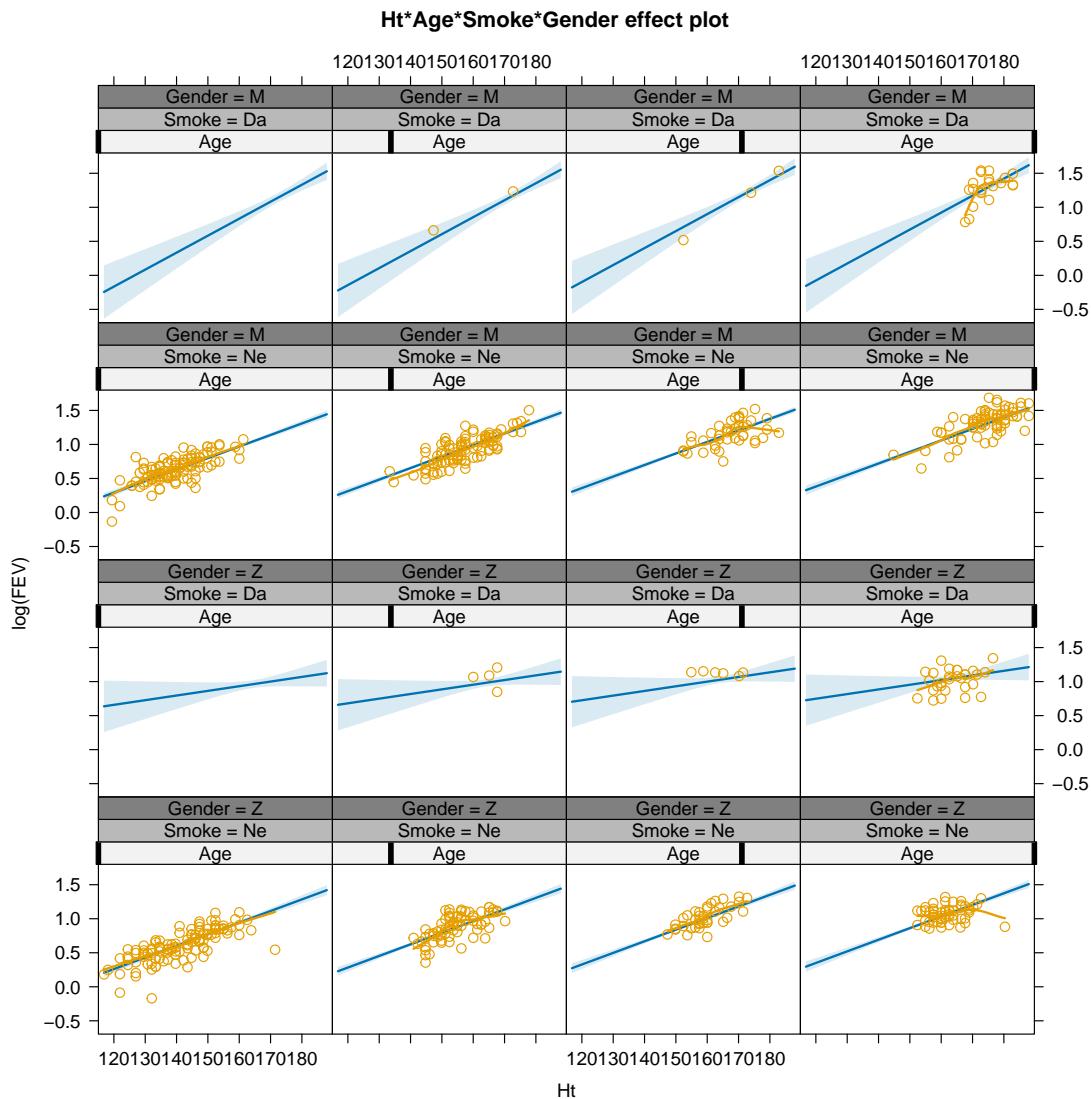
$$\hat{y} = (-1.967 + 0.045 + 1.625 - 3.042) + 0.0224\text{Age} + (0.017 - 0.010 - 0.0001 + 0.018)\text{Ht};$$

- $b_3 = 0.045$ predstavlja razliko med presečiščema za nekadilce in za nekadilke (**Age=0** in **Ht=0**);
- $b_4 = 1.625$ predstavlja razliko med presečiščema za kadilke in za nekadilke (**Age=0** in **Ht=0**);
- $b_5 = -0.0001$ predstavlja razliko v naklonu glede na **Ht** med nekadilci in nekadilkami pri konstantni vrednosti **Age**;
- $b_6 = -0.010$ predstavlja razliko v naklonu glede na **Ht** med kadilkami in nekadilkami pri konstantni vrednosti **Age**;
- vsota $b_3 + b_7 = 1.625 - 3.042$ predstavlja razliko med presečiščema kadilcev in nekadilcev (**Age=0** in **Ht=0**);
- vsota $b_6 + b_8 = -0.010 + 0.018$ predstavlja razliko v naklonu glede na **Ht** kadilcev in nekadilcev pri konstantni vrednosti **Age**;

Izziv

Še za `mod2.int` narišimo sliko parcialnih ostankov na kateri lahko razberemo, ali bi bilo potrebno v model vključiti tudi interakcijske člene z **Age** (Slika 34).

```
plot(Effect(c("Ht", "Age", "Smoke", "Gender"), mod2.int,
            partial.residuals=TRUE))
```



Slika 34: Modelske napovedi za FEV v odvisnosti od Ht, Gender in Smoke hkrati, pri povprečni vrednosti Age

Obrazložite sliko in rezultate, ki sledijo.

```
mod2.int.vse <- lm(log(FEV) ~ Age*Ht*Gender*Smoke, data=lungcap)
anova(mod2.int.vse, mod2.int)
```

Analysis of Variance Table

```
Model 1: log(FEV) ~ Age * Ht * Gender * Smoke
Model 2: log(FEV) ~ Age + Ht * Gender * Smoke
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     638 13.131
2     645 13.492 -7  -0.36158 2.5098  0.015 *
---

```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod2.int.vse)
```

Analysis of Variance Table

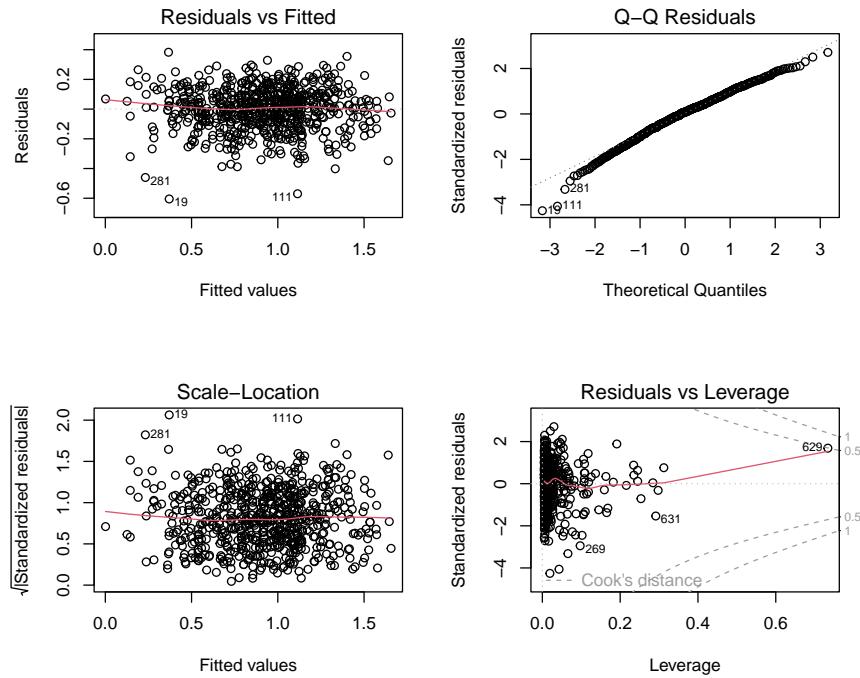
Response: log(FEV)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	43.210	43.210	2099.4787	< 2.2e-16 ***
Ht	1	15.326	15.326	744.6692	< 2.2e-16 ***
Gender	1	0.153	0.153	7.4492	0.0065216 **
Smoke	1	0.103	0.103	4.9904	0.0258335 *
Age:Ht	1	0.001	0.001	0.0707	0.7903929
Age:Gender	1	0.006	0.006	0.2882	0.5915464
Ht:Gender	1	0.004	0.004	0.2071	0.6491763
Age:Smoke	1	0.041	0.041	1.9778	0.1601139
Ht:Smoke	1	0.010	0.010	0.4820	0.4877888
Gender:Smoke	1	0.001	0.001	0.0304	0.8615821
Age:Ht:Gender	1	0.269	0.269	13.0719	0.0003234 ***
Age:Ht:Smoke	1	0.010	0.010	0.4637	0.4961528
Age:Gender:Smoke	1	0.035	0.035	1.7018	0.1925216
Ht:Gender:Smoke	1	0.152	0.152	7.3921	0.0067292 **
Age:Ht:Gender:Smoke	1	0.074	0.074	3.5968	0.0583438 .
Residuals	638	13.131	0.021		

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
```

```
plot(mod2.int.vse)
```



Slika 35: Diagnostični grafikoni za mod2.int.vse

```
summary(mod2.int.vse)
```

Call:

```
lm(formula = log(FEV) ~ Age * Ht * Gender * Smoke, data = lungcap)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.60494	-0.08675	0.01153	0.09379	0.38277

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.546e+00	3.166e-01	-8.043	4.29e-15 ***
Age	1.215e-01	4.111e-02	2.954	0.00325 **
Ht	2.049e-02	2.153e-03	9.515	< 2e-16 ***
GenderM	9.803e-01	4.131e-01	2.373	0.01794 *
SmokeDa	1.205e+01	4.997e+00	2.412	0.01614 *
Age:Ht	-6.003e-04	2.595e-04	-2.313	0.02103 *
Age:GenderM	-1.371e-01	5.095e-02	-2.692	0.00730 **
Ht:GenderM	-5.876e-03	2.793e-03	-2.104	0.03579 *
Age:SmokeDa	-8.366e-01	3.779e-01	-2.214	0.02721 *
Ht:SmokeDa	-7.147e-02	3.035e-02	-2.355	0.01883 *
GenderM:SmokeDa	-1.388e+01	5.811e+00	-2.388	0.01722 *
Age:Ht:GenderM	8.444e-04	3.158e-04	2.673	0.00770 **

```
Age:Ht:SmokeDa      4.931e-03  2.297e-03   2.147  0.03219 *
Age:GenderM:SmokeDa 8.748e-01  4.570e-01   1.914  0.05603 .
Ht:GenderM:SmokeDa 8.236e-02  3.497e-02   2.355  0.01881 *
Age:Ht:GenderM:SmokeDa -5.194e-03 2.739e-03  -1.897  0.05834 .

---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1435 on 638 degrees of freedom
Multiple R-squared:  0.8189,    Adjusted R-squared:  0.8147
F-statistic: 192.4 on 15 and 638 DF,  p-value: < 2.2e-16
```

Kaj bi lahko povedali o modelu `mod2.int.vse`? Ali bi dodali še katarega od grafičnih prikazov, ki bi lahko pokazal interakcijo med `Age` in ostalimi spremenljivkami?

Uvod

Statistični model

$$\text{Odziv} = \text{Signal} + \text{Šum}.$$

Odziv odzivna spremenljivka, odvisna spremenljivka

Signal sistematična komponenta odzivne spremenljivke

Šum slučajna komponenta odzivne spremenljivke

S statističnim modelom želimo čim bolje **oceniti signal in šum**.

Uvod

Linearni model (enostavna linearna regresija):

$$y_i = \mu_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i **odzivna spremenljivka**

ε_i **napaka**, slučajna komponenta odzivne spremenljivke

μ_i **pričakovana vrednost** y_i , sistematična komponenta odzivne spremenljivke

x_i **napovedna/neodvisna spremenljivka**

β_0, β_1 **parametra** linearnega modela

Odzivna spremenljivka y_i je v **linearni zvezi s parametri modela**.

Nelinearni eksponentni model

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

Odzivna spremenljivka y_i

- **ni v linearni zvezi** s parametrom modela β_1
- **je v linearni zvezi** s parametrom β_0 in z napako ε_i .

Uvod

Kaj je dober statistični model?

Statistični model predstavlja redukcijo pogosto obsežnega nabora podatkov na majhno število modelskih parametrov.

- Dober statistični model podatke reducira tako, da lahko na podlagi interpretacije parametrov naredimo smiselne odločitve.
- Model se dobro prilega podatkom, če sistematični del modela dobro opiše variabilnost odzivne spremenljivke, posledično je negotovost majhna.
- Model je dober, če je parsimoničen, kar pomeni, da vsebuje smiselno majhno število parametrov.
- Pri modeliranju je vedno treba nareediti kompromis med kompleksnostjo in interpretabilnostjo modela.

Uvod

Proces statističnega modeliranja

Statistično modeliranje je v grobem zaporedje treh korakov, ki jih ciklično ponavljamo dokler diagnostika modela ne pokaže, da je model sprejemljiv:

- začasna formulacija modela
- ocenjevanje parametrov
- diagnostika modela

Za končni model naredimo **obrazložitev rezultatov modeliranja**.

Uvod

Napovedne ali neodvisne spremenljivke

Za **napovedne spremenljivke** velja:

- so lahko številske ali opisne
- so vnaprej izbrane s strani načrtovalca raziskave
- v načrtovanem poskusu izbira vrednosti napovednih spremenljivk vpliva na statistično sklepanje o vplivih napovednih spremenljivk na odzivno spremenljivko in omogoča vzročno-posledično sklepanje
- če vrednosti napovednih spremenljivk niso izbrane vnaprej, v praksi predpostavimo, da so vsaj točne (brez merskih napak).

Uvod

Namen statističnega modeliranja

Namen statističnega modeliranja:

1. razumevanje izbranega procesa opisanega z odzivno spremenljivko in izbranimi napovednimi spremenljivkami, zanimajo nas povezave med napovednimi spremenljivkami in odzivno spremeljivko (*descriptive model*);
2. proučevanje vzročno-posledične zveze med napovednimi spremenlivkami in odzivno spremeljivko, katere napovedne spremenljivke statistično pomembno vplivajo na proces in kako (*explanatory model*);
3. napovedovanje odzivne spremenljivke na podlagi novih vrednosti napovednih spremenljivk, kjer vrednosti odzivne spremenljivke niso izmerjene/opazovane (*prognostic model*).

<https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf>

Linearni model

Predpostavke linearnega modela

Imamo **odzivno spremenljivko** Y in m **napovednih spremenljivk** $X_j, j = 1, \dots, m$, ki so številske in/ali opisne. m napovednih spremenljivk generira k **regresorjev**, $X_1, \dots, X_k, k \geq m$.

Predpostavke linearnega modela:

1. Y je številska spremenljivka, njene vrednosti so **medsebojno neodvisne**.
2. Pričakovana vrednost Y pogojno na X_1, \dots, X_k je

$$\mathbb{E}(Y|X_1, \dots, X_k) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

$\beta_j, j = 0, \dots, k$ so **parametri modela**. Y je v linearni zvezi s parametri modela.

3. Varianca Y pogojno na X_1, \dots, X_k , je konstantna

$$\text{Var}(Y|X_1, \dots, X_k) = \sigma^2 > 0.$$

Linearni model

Napovedne spremenljivke, regresorji

Iz m napovednih spremenljivk dobimo k regresorjev X_1, \dots, X_k , $k \geq m$, na različne načine:

- **številsko spremenljivko** v model vključimo direktno kot **en regresor**; včasih je ta spremenljivka predhodno **transformirana** (npr. *log*). V določenih primerih je številska spremenljivka vključena v model z **več regresorji** (npr. polinomska regresija, zlepki);
- za **opisno spremenljivko** z d vrednostmi se v model vključi $d - 1$ regresorjev z vrednostmi 0 in 1 (neme spremenljivke, *dummy variables*);
- dodatne regresorje lahko dobimo z vključitvijo **interakcij med napovednimi spremenljivkami** v modelu.

Linearni model

Modeliramo na podatkih iz vzorca, ki ima n enot

Vrednosti odzivne in napovednih spremenljivk so dobljene **na vzorcu, ki ima n enot**. Linearni model za i -to enoto, $i = 1, \dots, n$, zapišemo

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i,$$

ε_i se imenuje **napaka (error)**, njene lastnosti so:

$$\mathbb{E}(\varepsilon_i) = 0,$$

$$Var(\varepsilon_i) = \sigma^2 \quad \text{ali} \quad Var(\varepsilon_i) = \frac{\sigma^2}{w_i}, \quad i = 1, \dots, n,$$

ε_i so medsebojno neodvisni $Cov(\varepsilon_i, \varepsilon_j) = 0$,

$w_i, i = 1, \dots, n$ so **znane pozitivne uteži**.

Linearni model

Normalni linearni model

Normalni linearni model: če je lahko predpostavimo, da je porazdelitev Y pri X_1, \dots, X_k normalna s povprečjem na regresijski hiper-ravnini in varianco σ^2 .

$$Y|X_1, \dots, X_k \sim N(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k, \sigma^2),$$

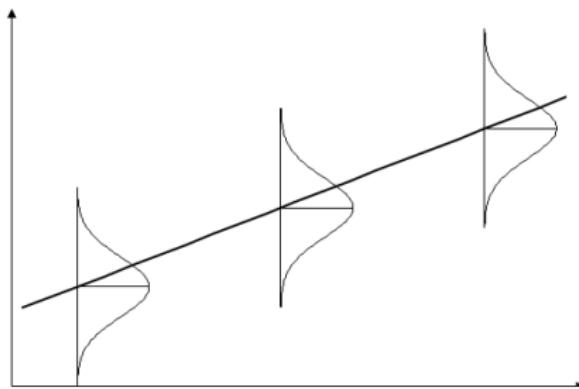
Posledično:

- ε_i so neodvisno enako normalno porazdeljeni $\varepsilon_i \sim iid N(0, \sigma^2)$
- $Var(\varepsilon_i) = \sigma^2$ ali $Var(\varepsilon_i) = \frac{\sigma^2}{w_i}$, varianca σ^2 in uteži w_i so konstante

Linearni model

Predpostavke

Ilustracija predpostavke linearnega modela na primeru enostavne linearne regresije



Linearni model

Ocenjevanje parametrov modela

Parametre linearnega modela ocenjujemo na podlagi podatkov na vzorcu n enot.

Metode:

OLS metoda najmanjših kvadratov (*Ordinary Least Squares*)

WLS tehtana metoda najmanjših kvadratov (*Weighted Least Squares*)

GLS posplošena metoda najmanjših kvadratov
(*Generalised Least Squares*)

ML metoda največjega verjetja (*Maximum likelihood*)

Linearni model

Ocenjevanje parametrov modela, OLS

OLS, minimiramo vsoto kvadratov odklonov y od $\mathbb{E}(y)$:

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}))^2.$$

WLS, minimiramo vsoto tehtanih kvadratov odklonov y od $\mathbb{E}(y)$:

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n w_i (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}))^2.$$

Izraz $S(\beta_0, \beta_1, \dots, \beta_k)$ parcialno odvajamo po parametrih β_j , $j = 0, \dots, k$, in odvode izenačimo z 0. Dobimo **normalni sistem** $k+1$ **linearnih enačb**. Rešitev tega sistema so **cenilke parametrov**, b_j , $j = 0, \dots, k$.

Linearni model

Prilagojene vrednosti, ostanki

Z modelom **prilagojene vrednosti** (*fitted values*) označimo \hat{y}_i :

$$\hat{y}_i = b_0 + b_1 x_{i1} + \dots + b_k x_{ik}, \quad i = 1, \dots, n.$$

Razliko med dejansko vrednostjo y_i in napovedano vrednostjo \hat{y}_i imenujemo **ostanek** (*residual*) , e_i :

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n.$$

Ostanki so e_i nekorelirani z prilagojenimi vrednostmi \hat{y}_i , kar uporabljamo pri analizi modela z grafičnimi prikazi.

Linearni model

Varianca ostankov

Varianca ostanka:

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii}).$$

h_{ii} se imenuje vzvod (*leverage* ali *hat-value*)

h_{ii} je odvisen od $(x_{i1}, x_{i2}, \dots, x_{ik})$, zavzema vrednosti med $1/n$ in 1.

Za izračun $\text{Var}(e_i)$ moramo varianco σ^2 oceniti na podlagi podatkov, cenilko variance označimo s^2 .

$$\widehat{\text{Var}}(e_i) = s^2(1 - h_{ii}).$$

Linearni model

Standardizirani ostanki

Standardizirani ostanek e_{s_i} :

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

Če je $n \gg k$ velja, da je njihova porazdelitev približno $N(0, 1)$.

Ali so predpostavke modela izpolnjene, ugotavljamo z analizo ostankov in standardiziranih ostankov.

Normalni linearni model

O parametrih modela lahko povemo več

Če lahko privzamemo **normalni linearni model**, poznamo verjetnostne porazdelitve parametrov modela in:

- za vsako cenilko parametra lahko izračunamo njen standardno napako;
- izračunamo interval zaupanja za vsak parameter modela;
- testiramo lahko statistične domneve o parametrih modela;
- izračunamo napovedi in intervale zaupanja za povprečno napoved in za posamično napoved.

Lastnosti parametrov modela poglejmo najprej na primeru enostavne linearne regresije.

Enostavna linearna regresija

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

ε_i so neodvisne slučajne spremenljivke s pričakovano vrednostjo $\mathbb{E}(\varepsilon_i) = 0$ in konstantno varianco $Var(\varepsilon_i) = \sigma^2$ za vsak $i = 1, \dots, n$.

β_0, β_1 parametra modela, ki ju bomo ocenili z metodo najmanjših kvadratov OLS.

Enostavna linearna regresija

Ocenjevanje parametrov modela po metodi najmanjših kvadratov, OLS

Izrek 1.1: Po metodi najmanjših kvadratov sta cenilki parametrov β_0 in β_1

$$b_0 = \bar{y} - b_1 \bar{x},$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}.$$

SS_{xx} je vsota kvadratov odklonov za x

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})x_i,$$

SS_{xy} je vsota produktov odklonov x in y

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i(y_i - \bar{y}) = \sum_{i=1}^n y_i(x_i - \bar{x}).$$

Enostavna linearna regresija

Dokaz izreka 1.1

Dokaz:

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

parcialno odvajamo po β_0 in β_1 :

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i),$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i).$$

Ko odvoda izenačimo z 0, dobimo sistem dveh linearnih enačb.

Enostavna linearna regresija

Dokaz izreka 1.1, nadaljevanje

$$\sum_{i=1}^n y_i = nb_0 + b_1 \sum_{i=1}^n x_i,$$

$$\sum_{i=1}^n x_i y_i = b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2.$$

Iz prve enačbe sledi

$$b_0 = \bar{y} - b_1 \bar{x}$$

Ta rezultat uporabimo v drugi enačbi sistema

Enostavna linearna regresija

Dokaz izreka 1.1, nadaljevanje

$$\sum_{i=1}^n (x_i y_i - x_i(\bar{y} - b_1 \bar{x}) - b_1 x_i^2) = \sum_{i=1}^n (x_i y_i - x_i \bar{y} + b_1 x_i \bar{x} - b_1 x_i^2) = 0$$

$$\sum_{i=1}^n (x_i y_i - x_i \bar{y}) = b_1 \sum_{i=1}^n (x_i^2 - x_i \bar{x}).$$

Iz tega sledi:

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n (x_i y_i - x_i \bar{y})}{\sum_{i=1}^n (x_i^2 - x_i \bar{x})} = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sum_{i=1}^n x_i (x_i - \bar{x})} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}} \end{aligned}$$

Enostavna linearna regresija

Enačba regresijske premice

Enačba regresijske premice:

$$\hat{y} = b_0 + b_1 x$$

b_0 presečišče premice z ordinatno osjo

b_1 naklon premice

Model velja na intervalu $[x_{min}, x_{max}]$.

Ob danih predpostavkah sta cenilki b_0 in b_1 funkciji y_i in posledično tudi ε_i .

Enostavna linearna regresija

Cenilka za varianco napak

Cenilka za varianco napak σ^2 je s^2

Napake: $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ ocenimo z **ostanki**:

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i, \quad i = 1, \dots, n.$$

Definiramo **vsoto kvadratov ostankov** ($SS_{residual}$):

$$SS_{residual} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

Nepristranska cenilka za σ^2 :

$$s^2 = \frac{SS_{residual}}{n - 2}.$$

V imenovalcu delimo z $n - 2$ namesto z n .

Enostavna linearna regresija

Povzetek glede y

Glede odzivne spremenljivke smo do sedaj povedali:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\hat{y}_i = b_0 + b_1 x_i$$

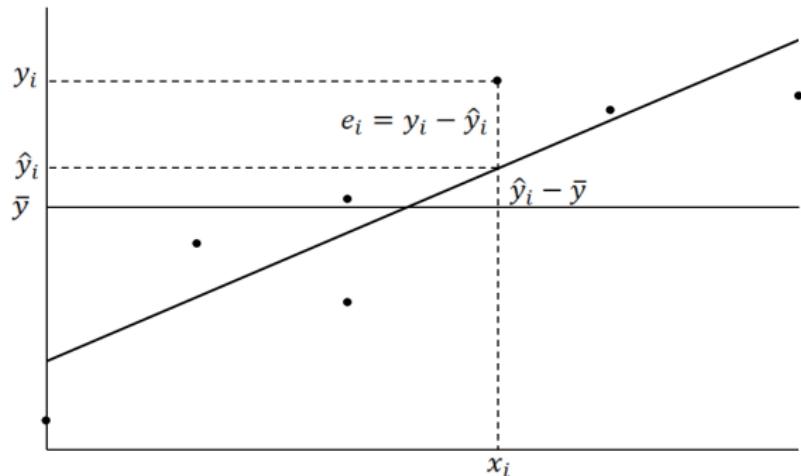
V nadaljevanju bomo videli

$$SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SS_{model} + SS_{residual}$$

Enostavna linearna regresija

$$SS_{yy} = SS_{model} + SS_{residual}$$

Grafični prikaz, ki je osnova za delitev variabilnosti odzivne spremenljivke na dva dela:



Enostavna linearna regresija

$SS_{yy} = SS_{model} + SS_{residual}$, koeficient determinacije

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$
$$SS_{yy} = SS_{model} + SS_{residual}.$$

Koeficient determinacije R^2 je delež variabilnosti za y , ki je pojasnjen z regresijskim modelom:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{SS_{model}}{SS_{yy}}.$$

Koeficient determinacije je enostavna mera za kakovost linearnega regresijskega modela.

Za izračun R^2 ne potrebujemo nobenih predpostavk.

Enostavna linearna regresija

Koeficient determinacije

Lastnosti koeficienteja determinacije:

- je nenegativna vrednost;
- je manjši ali enak 1; ima vrednost 1, če je $SS_{model} = SS_{yy}$, ko so vse točke na premici;
- R^2 je odvisen od zaloge vrednosti napovedne spremenljivke;
- pri uporabi R^2 moramo biti previdni, saj vsak dodani regresor poveča vrednost R^2 , tudi če je vpliv tega regresorja na odzivno spremenljivko statistično nepomemben (multipla regresija).

Enostavna linearna regresija

Diagnostika linearnega modela

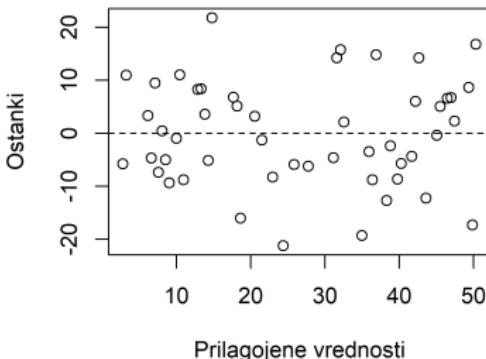
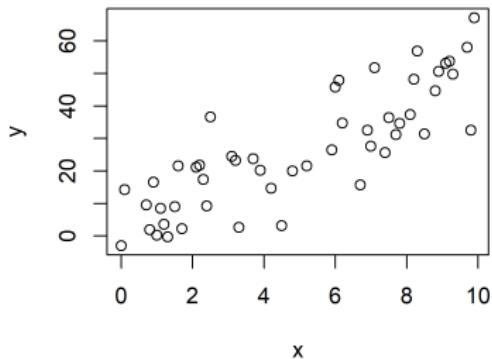
Diagnostika modela je namenjena preverjanju predpostavk linearnega modela. Na podlagi podatkov ocenimo parametre modela in preverimo, ali je bilo tako modeliranje upravičeno. Preverjamo:

- **linearnost** odvisnosti odzivne spremenljivke od napovedne spremenljivke (razsevni grafikon y glede na x , slika ostankov v odvisnosti od prilagojenih vrednosti, odvisnost ne sme biti vidna);
- **varianca napak** oziroma varianca odzivne spremenljivke pogojno na napovedne spremenljivke **je konstantna** (slika ostankov glede na prilagojene vrednosti);
- **pričakovana vrednost napak je 0** (slika ostankov glede na prilagojene vrednosti);
- **napake so medsebojno neodvisne** (težko preveriti, verjamemo, da so bili podatki pridobljeni z ustreznim načinom vzorčenja, analiza avtokorelacije ostankov).

Enostavna linearna regresija

Diagnostika linearnega modela, primer

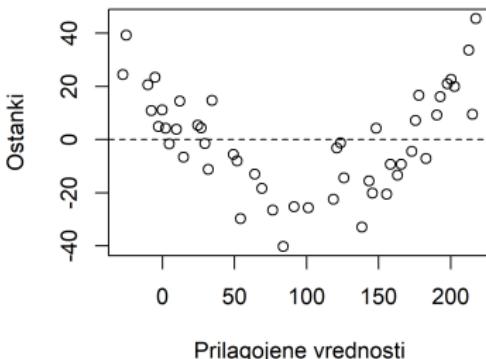
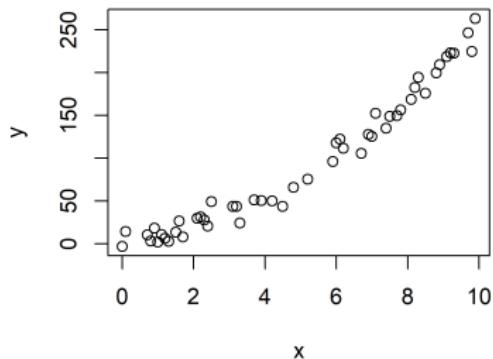
Ni odstopanja od predpostavk linearnega modela



Enostavna linearna regresija

Diagnostika linearnega modela, primer

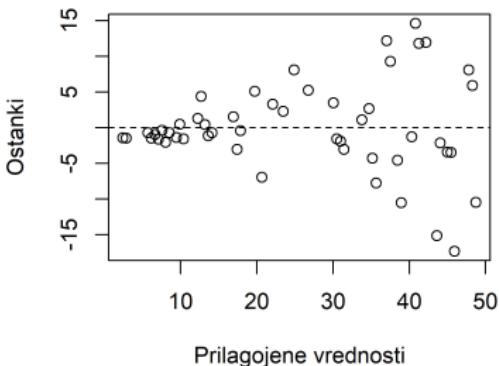
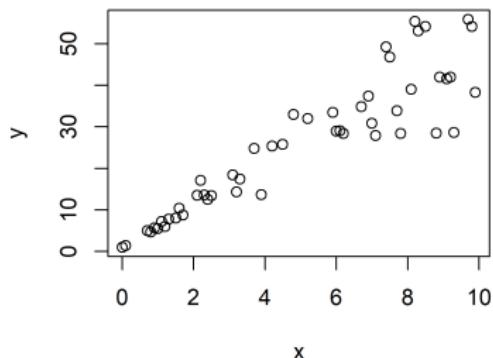
Nelinearnost



Enostavna linearna regresija

Diagnostika linearnega modela, primer

Nekonstantna varianca, heteroskedastičnost



Enostavna linearna regresija

Porazdelitev cenilk b_0 , b_1 in s^2

Ob predpostavki $\varepsilon_i \sim iid N(0, \sigma^2)$ in dejstvu, da sta b_0 in b_1 funkciji normalno porazdeljenih spremenljivk, velja, da je tudi njuna porazdelitev aproksimativno (če je n velik) normalna:

$$b_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)\right),$$
$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{xx}}\right).$$

Cenilki varianc parametrov modela $s_{b_0}^2$ in $s_{b_1}^2$ izračunamo tako, da σ^2 zamenjamo z s^2 .

Za porazdelitev cenilke variance napak s^2 velja

$$\frac{(n - 2)s^2}{\sigma^2} \sim \chi_{n-2}^2$$

Enostavna linearna regresija

Statistično sklepanje o parametrih modela

Izrek 1.4: Če sta $X \sim N(0, 1)$ in $Y \sim \chi_n^2$ neodvisni slučajni spremenljivki, potem velja

$$\frac{X}{\sqrt{Y/n}} \sim t_n$$

Izrek 1.5: Če uporabimo predstavljene lastnosti cenilk in izrek 1.4, lahko pod predpostavko normalne porazdelitve napak ε_i , $i = 1, \dots, n$ izpeljemo

$$\frac{b_0 - \beta_0}{s \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}} \sim t_{n-2} \quad \frac{b_1 - \beta_1}{\frac{s}{\sqrt{SS_{xx}}}} \sim t_{n-2}.$$

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2} \quad \text{in} \quad \frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}.$$

Enostavna linearna regresija

Interval zaupanja za β_0

Interval zaupanja za β_0

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2}$$

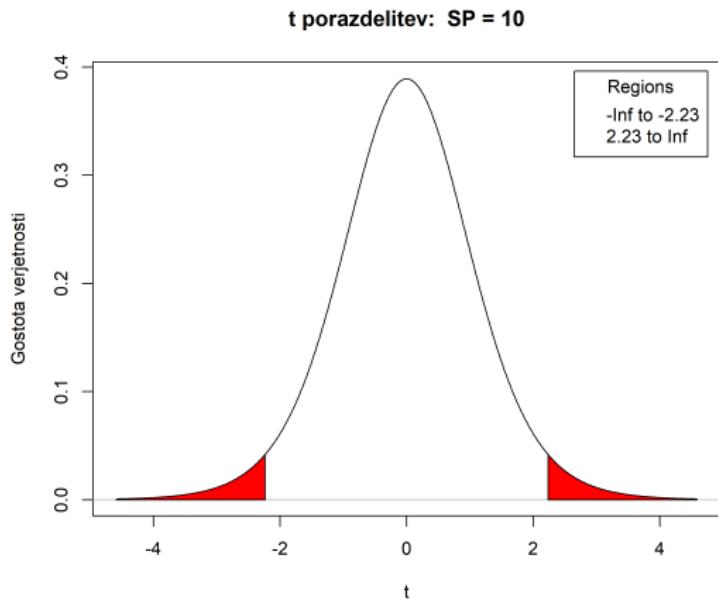
$$P\left(-|t_{\frac{\alpha}{2};n-2}| \leq \frac{b_0 - \beta_0}{s_{b_0}} \leq |t_{\frac{\alpha}{2};n-2}|\right) = 1 - \alpha$$

$$P(b_0 - |t_{\frac{\alpha}{2};n-2}|s_{b_0} \leq \beta_0 \leq b_0 + |t_{\frac{\alpha}{2};n-2}|s_{b_0}) = 1 - \alpha$$

$|t_{\frac{\alpha}{2};n-2}|$ je absolutna vrednost $(\alpha/2)$ -tega kvantila t -porazdelitve s $SP = n - 2$.

Enostavna linearna regresija

Studentova t -porazdelitev



$t_{0,025;10} = -2,23$ je 0,025-ti kvantil t -porazdelitve s $SP = 10$.

Enostavna linearna regresija

Interval zaupanja za β_1

Interval zaupanja za β_1

$$\frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}$$

$$P\left(-|t_{\frac{\alpha}{2};n-2}| \leq \frac{b_1 - \beta_1}{s_{b_1}} \leq |t_{\frac{\alpha}{2};n-2}|\right) = 1 - \alpha$$

$$P(b_1 - |t_{\frac{\alpha}{2};n-2}| s_{b_1} \leq \beta_1 \leq b_1 + |t_{\frac{\alpha}{2};n-2}| s_{b_1}) = 1 - \alpha$$

$|t_{\frac{\alpha}{2};n-2}|$ je absolutna vrednost $(\alpha/2)$ -tega kvantila t -porazdelitve s $SP = n - 2$.

Enostavna linearna regresija

Intervali zaupanja za β_0 in β_1

100(1 - α) % intervala zaupanja za β_0 in β_1 :

$$\left(b_0 - |t_{\frac{\alpha}{2};n-2}|s_{b_0}, \quad b_0 + |t_{\frac{\alpha}{2};n-2}|s_{b_0} \right)$$

$$\left(b_1 - |t_{\frac{\alpha}{2};n-2}|s_{b_1}, \quad b_1 + |t_{\frac{\alpha}{2};n-2}|s_{b_1} \right)$$

Enostavna linearna regresija

Testiranje domnev za β_1

Testiramo ničelno domnevo za β_1

$$H_0 : \beta_1 = \beta \quad H_1 : \beta_1 \neq \beta.$$

Testna statistika je

$$T = \frac{b_1 - \beta}{\frac{s}{\sqrt{SS_{xx}}}} \sim t_{n-2},$$

Ničelno domnevo zavrnemo pri stopnji značilnosti α , če je

$$T < -|t_{\alpha/2;n-2}| \quad \text{ali} \quad T > |t_{\alpha/2;n-2}|$$

Ob tem je verjetnost, da T leži v območju zavrnitve ničelne domneve, ko je ta pravilna, največ α .

Enostavna linearna regresija

Testiranje domneva za β_1 , p -vrednost

Za dani vzorec podatkov izračunamo vrednost testne statistike t , za katero pod ničelno domnevo izračunamo **p -vrednost**

$$p = P(|T| \geq |t| \mid \beta_1 = \beta)$$

Če je $p < \alpha$ ničelno domnevo zavrnemo in če je $p \geq \alpha$ ničelne domenve ne moremo zavrniti.

Enako kot ničelno domnevo z dvostransko alternativno domnevo, lahko testiramo tudi ničelno domnevo z **enostransko alternativno domnevo** $H_1 : \beta_1 < \beta$ ali $H_1 : \beta_1 > \beta$. V tem primeru je p -vrednost polovica p -vrednosti pri testiranju dvostranske alternativne domneve.

Enostavna linearna regresija

Testiranje domnev za β_0

Podobno lahko testiramo ničelno domnevo za β_0

$$H_0 : \beta_0 = \beta \quad H_1 : \beta_0 \neq \beta$$

Testna statistika je

$$T = \frac{b_0 - \beta}{s \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}} \sim t_{n-2}$$

Ničelno domnevo zavrnemo pri stopnji značilnosti α , če je

$$T < -|t_{\alpha/2;n-2}| \quad \text{ali} \quad T > |t_{\alpha/2;n-2}|$$

Ob tem je verjetnost, da T leži v območju zavrnitve ničelne domneve, ko je ta pravilna, največ α .

Enostavna linearna regresija

Testiranje domnev za β_0 in β_1 , povzetek modela v R

Ali je zveza med y in x pomembna?

Kakšna je zveza (naraščajoča/padajoča, tesna/šibka)?

Če y ni odvisen od x , je najboljša napoved za y , $\hat{y} = \bar{y}$ in $\beta_1 = 0$.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

Osnovni povzetek regresijskega modela v R vsebuje rezultat testiranja zgornje domneve in tudi rezultat testiranja ničelne domneve:

$$H_0 : \beta_0 = 0 \quad H_1 : \beta_0 \neq 0$$

Ta ničelna domneva je redko vsebinsko zanimiva.

Enostavna linearna regresija

Analize variance za regresijski model

$$\begin{aligned}SS_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\&= SS_{model} + SS_{residual}\end{aligned}$$

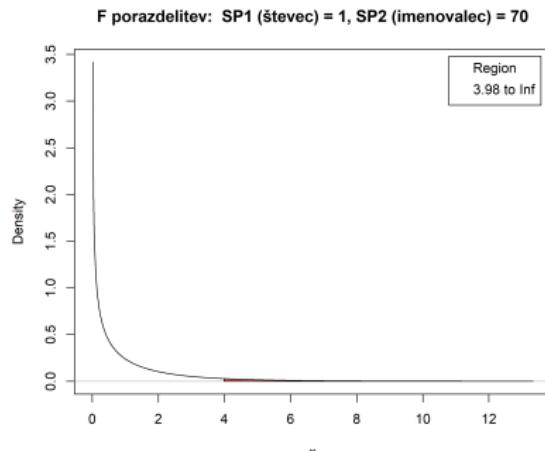
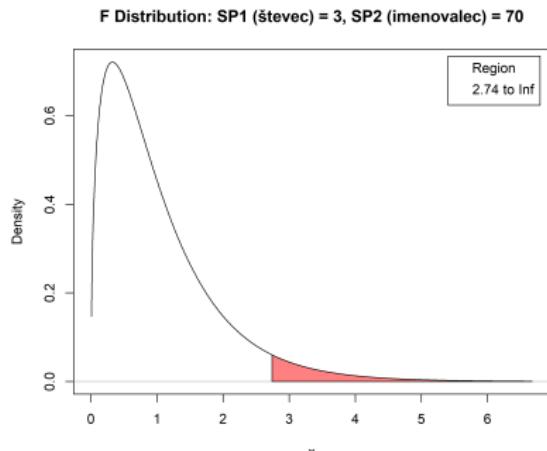
Tabela: Shema tabele ANOVA za enostavni linearni regresijski model

Vir variabilnosti	<i>df</i>	<i>SS</i>	$MS = SS/df$	<i>F</i>
Model	1	SS_{model}	MS_{model}	$MS_{model}/MS_{residual}$
Ostanek (Residual)	$n - 2$	$SS_{residual}$	$MS_{residual}$	
Skupaj	$n - 1$	SS_{yy}		

Enostavna linearna regresija

Analiza variance, F -porazdelitev

Ob predpostavki $\varepsilon_i \sim iid N(0, \sigma^2)$ za F -statistiko velja, da je njena ničelna porazdelitev F -porazdelitev s stopinjami prostosti $SP_{model} = k$ in $SP_{residual} = n - 2$.



Enostavna linearna regresija

Analiza variance

Iz tabele ANOVA dobimo:

- ▶ cenilko za varianco σ^2 , ki jo označimo $s^2 = MS_{residual}$. Količino s imenujemo **standardna napaka regresije** (*Residual standard error*).
- ▶ F -statistika testira domnevo o ničelnem vplivu napovedne spremenljivke:
 $H_0 : \beta_1 = 0,$
 $H_1 : \beta_1 \neq 0.$

Enostavna linearna regresija

Analiza variance

F -test dobi večji pomen v primeru, ko imamo k napovednih spremenljivk v modelu, ker testira ničelno domnevo o hkratni ničnosti vseh parametrov v modelu

$$H_0 : \beta_1 = \dots = \beta_k = 0$$

nasproti alternativni domnevi

$$H_1 : \text{vsaj en } \beta_i \neq 0, \quad i = 1, \dots, k$$

Enostavna linearna regresija

Povezava med T in F statistiko

Izrek 1.6: Če je slučajna spremenljivka X porazdeljena po t -porazdelitvi s stopinjami prostosti v , $X \sim t_v$, potem je slučajna spremenljivka X^2 porazdeljena po F -porazdelitvi s stopinjami prostosti 1 in v , $X^2 \sim F_{1,v}$.

- za testiranje domneve $\beta_1 = 0$ velja

$$T = \frac{b_1}{\frac{s}{\sqrt{SS_{xx}}}} \sim t_{n-2}$$

- če zgornji izraz kvadriramo, dobimo F -statistiko

$$F = \frac{b_1^2 SS_{xx}}{s^2} \sim F_{1,n-2}$$

Enostavna linearna regresija

Povezava med T in F statistiko

- če upoštevamo, $SS_{model} = b_1^2 S_{xx}$ in $s^2 = SS_{residual}/(n - 2)$

$$F = \frac{b_1^2 SS_{xx}}{s^2} = \frac{SS_{model}/1}{SS_{residual}/(n - 2)} \sim F_{1,n-2}.$$

F -statistika je skalirano razmerje vsote kvadratov odklonov modela in ostanka. Če je SS_{model} veliko večja od $SS_{residual}$, bo F -statistika velika in ničelno domnevo, ki pravi, da regresorji niso uporabni pri napovedovanju odzivne spremenljivke, bomo zavnili.

Enostavna linearna regresija

Napovedovanje, povprečna napoved

Na podlagi ocenjenih parametrov modela lahko izračunamo **povprečno napoved** $\hat{y}(x_0)$, x_0 je izbrana vrednost napovedne spremenljivke.

$$\hat{y}(x_0) = b_0 + b_1 x_0,$$

ob tem je prava napoved

$$\mathbb{E}(y(x_0)) = \beta_0 + \beta_1 x_0$$

Enostavna linearna regresija

Napovedovanje, povprečna napoved

Pokažemo lahko, da je porazdelitev povprečne napovedi pri x_0 normalna:

$$\hat{y}(x_0) \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right)\right)$$

in velja, da je statistika

$$\frac{\hat{y}(x_0) - \beta_0 - \beta_1 x_0}{s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}} \sim t_{n-2}$$

100(1 - α)% interval zaupanja za povprečno napoved

$$\left(\hat{y}(x_0) - |t_{\frac{\alpha}{2}; n-2}| s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}, \quad \hat{y}(x_0) + |t_{\frac{\alpha}{2}; n-2}| s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \right)$$

Enostavna linearna regresija

Napovedovanje, posamična napoved

Z y_0 označimo eno izmed možnih vrednosti odzivne spremenljivke pri x_0

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Napoved za y_0 imenujemo **posamična napoved** in je enaka **povprečni napovedi** $\hat{y}(x_0)$:

$$\mathbb{E}(\hat{y}(x_0) - y_0) = \beta_0 + \beta_1 x_0 - \beta_0 - \beta_1 x_0 = 0$$

Varianca posamične napovedi je:

$$\text{Var}(y_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right).$$

Enostavna linearna regresija

Napovedovanje, posamična napoved

Podobno kot pri intervalu zaupanja za povprečno napoved lahko definiramo T -statistiko:

$$\frac{\hat{y}(x_0) - y_0}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}} \sim t_{n-2}$$

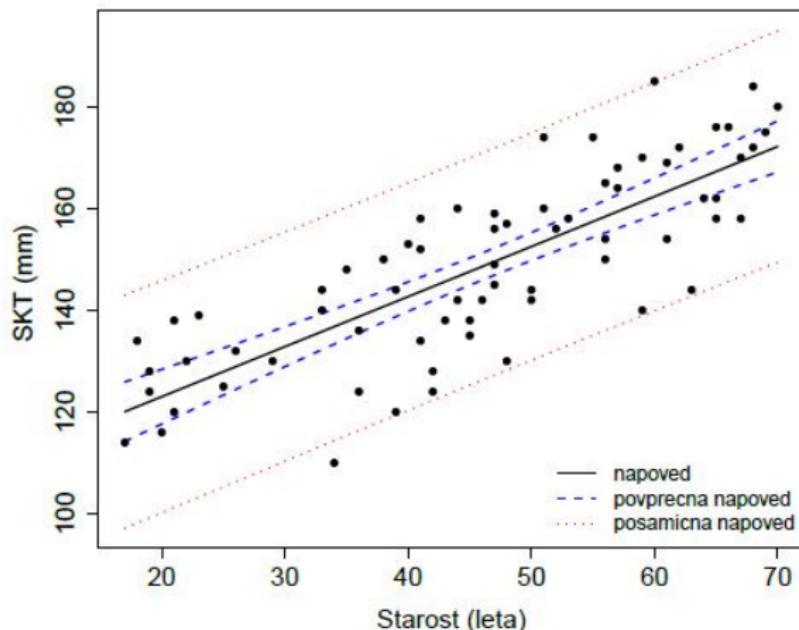
in $100(1 - \alpha)$ % interval zaupanja za posamično napoved je

$$\hat{y}(x_0) \mp |t_{\frac{\alpha}{2}; n-2}| s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

- velikost vzorca n
- varianca napak, ocenjena z s^2
- položaj x_0 , najmanjša je pri povprečju \bar{x} in narašča s kvadratom razdalje od povprečja.

Enostavna linearna regresija

Primer: intervali zaupanja za napovedi



Linearni model v matrični obliki

Odzivno spremenljivko y modeliramo na podlagi k regresorjev (splošni normalni linearni model)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

V matrični obliki:

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

Linearni model v matrični obliki

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

\mathbf{y} vektor odzivne spremenljivke

\mathbf{X} modelska matrika reda $(n \times k + 1)$

$\boldsymbol{\beta}$ vektor parametrov modela velikosti $(k + 1) \times 1$

$\boldsymbol{\varepsilon}$ vektor napak velikosti $(n \times 1)$, $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ in

$Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$, \mathbf{I} je enotska diagonalna matrika reda $n \times n$

Linearni model v matrični obliki

Cenilke parametrov po metodi najmanjših kvadratov

Minimiramo vsoto kvadratov napak:

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}))^2$$

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Parcialno odvajamo po parametrih β_j , $j = 0, \dots, k$, in odvode izenačimo z 0. Dobimo **normalni sistem** $k + 1$ **linearnih enačb**:

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{y}$$

Rešitev obstaja, če je $\mathbf{X}^T \mathbf{X}$ **nesingularna**.

Linearni model v matrični obliki

Cenilke parametrov po metodi najmanjših kvadratov

$\mathbf{X}^T \mathbf{X}$ je nesingularna:

- če je $n \geq k + 1$; to pomeni, da je število enot vsaj tako veliko kot število ocenjevanih parametrov;
- če nobena spremenljivka ni linearnejša kombinacija ostalih spremenljivk, kar pomeni, da ima matrika \mathbf{X} polni rang $k + 1$, gre za **linearni model polnega ranga** (*full rank linear model*).

Linearni model v matrični obliki

Cenilke parametrov po metodi najmanjših kvadratov

Rešitev je vektor cenilk parametrov $\mathbf{b} = (b_0, b_1, \dots, b_k)$:

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

Posamezno cenilko b_j lahko zapišemo

$$b_j = \frac{\sum_{i=1}^n x_{ij}^* y_i}{\sum_{i=1}^n (x_{ij}^*)^2},$$

x_{ij}^* je vrednost spremenljivke x_{ij} po tem, ko je bila prilagojena na vse ostale napovedne spremenljivke x_1, \dots, x_k brez spremenljivke x_j .

Vektor prilagojenih vrednosti: $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$

Vektor ostankov: $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$

Linearni model v matrični obliki

Lastnosti cenilk parametrov, nepristranskost

Izrek 2.1: v linearinem modelu polnega ranga so cenilke parametrov izračunane po metodi najmanših kvadratov
 $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ nepristranske:

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta}$$

variančno-kovariančno matrika vektorja cenilk je

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

Linearni model v matrični obliki

Lastnosti cenilk parametrov, nepristranskost

Dokaz:

$$\mathbb{E}(\mathbf{b}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}(\mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta}$$

Edina predpostavka: $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$. Cenilke parametrov modela so nepristranske tudi, če varianca σ^2 ni konstantna ali če so napake korelirane.

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

Tu smo upoštevali predpostavko konstantne variance $Var(\mathbf{y}) = \sigma^2$.

Linearni model v matrični obliki

Lastnosti cenilk parametrov, Gauss-Markov izrek

Gauss-Markov izrek:

Naj bo \mathbf{b}^* nepristranska cenilka za β in \mathbf{b} cenilka za β po metodi najmanjših kvadratov, potem velja, da je $Var(b_i) \leq Var(b_i^*)$, $i = 1, \dots, k + 1$.

Pravimo, da je **\mathbf{b} najboljša linearna nepristranska cenilka za β** (BLUE, *Best Linear Unbiased Estimator*).

(Brez dokaza.)

Ker so cenilke parametrov linearnega normalnega modela linearne kombinacije odzivne spremenljivke, za katero smo predpostavili normalno porazdelitev, je njihova porazdelitev **večrazsežna normalna porazdelitev**.

Linearni model v matrični obliki

Cenilka za σ^2

Nepristranska cenilka za σ^2

$$s^2 = \hat{\sigma}^2 = \frac{1}{n - k - 1} (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

Linearni model v matrični obliki

Matrika \mathbf{H}

Poglejmo povezavo med $\hat{\mathbf{y}}$ in \mathbf{y} :

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$$

Matrika \mathbf{H} reda $n \times n$ (*hat matrix*) je ključna pri izračunu napovedi $\hat{\mathbf{y}}$:

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

Matrika \mathbf{H} ima lepe lastnosti, pokažemo lahko, da velja:

$$\mathbf{H} = \mathbf{H}^T = \mathbf{H}^2 \text{ (idempotentna matrika).}$$

Linearni model v matrični obliki

Ostanki

Vektor ostankov lahko zdaj zapisemo tudi z matriko \mathbf{H} :

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

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

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

Linearni model v matrični obliki

Statistično sklepanje v linearinem modelu

Glavna predpostavka za statistično sklepanje je:
cenilke parametrov \mathbf{b} so porazdeljene po **večrazsežnostni
normalni porazdelitvi**

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1})$$

Velja tudi, da so cenilke prametrov \mathbf{b} neodvisne od cenilke variance napak $\hat{\sigma}^2$.

Porazdelitev za reskalirano varianco napak $(n - k - 1)\hat{\sigma}^2/\sigma^2$ je **χ^2 -porazdelitev** s stopinjami prostosti $SP = n - k - 1$:

$$\frac{(n - k - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}$$

Linearni model v matrični obliki

Intervalne ocene za parametre modela

Interval zaupanja za posamezen parameter modela β_j , $j = 0, \dots, k$,
ob upoštevanju ostalih regresorjev v modelu imenujemo
parcialni interval zaupanja.

Definiran je na podlagi statistike

$$\frac{b_j - \beta_j}{\hat{\sigma} \sqrt{a_{jj}}}$$

$\sqrt{a_{jj}}$ je diagonalni element matrike $\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1}$.

Kakšna je porazdelitev te statistike?

Linearni model v matrični obliki

Intervalne ocene za parametre modela

Velja, da je statistika:

$$\frac{b_j - \beta_j}{\sigma \sqrt{a_{jj}}} \sim N(0, 1)$$

Zgornji izraz delimo s korenom reskalirane variance napak deljene z $(n - k - 1)$, dobimo statistiko, ki je porazdeljena po t -porazdelitvi z $SP = n - k - 1$ (Izrek 1.4):

$$\frac{b_j - \beta_j}{\sigma \sqrt{a_{jj}}} / \sqrt{\frac{\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2}}{n - k - 1}} \sim t_{n-k-1}$$

Ko zgornji izraz poenostavimo, dobimo

$$\frac{b_j - \beta_j}{\hat{\sigma} \sqrt{a_{jj}}} \sim t_{n-k-1}$$

Linearni model v matrični obliki

Intervalne ocene za parametre modela

Posledično je $100(1 - \alpha) \%$ parcialni interval zaupanja za β_j ob upoštevanju ostalih napovednih spremenljivk v modelu:

$$(b_j - |t_{\frac{\alpha}{2};n-k-1}| \hat{\sigma} \cdot \sqrt{a_{jj}}, \quad b_j + |t_{\frac{\alpha}{2};n-k-1}| \hat{\sigma} \cdot \sqrt{a_{jj}})$$

Funkcija `confint()` vrne parcialne 95 % intervale zaupanja za vse parametre v modelu.

Linearni model v matrični obliki

Testiranje domnev o parametrih modela

Za j -ti parameter lahko zapišemo H_0 in H_1 , $j = 0, \dots, k$:

$H_0 : \beta_j = \gamma_j$ ob upoštevanju vseh ostalih členov v modelu

$H_1 : \beta_j \neq \gamma_j$

Testna statistika je

$$t = \frac{b_j - \gamma_j}{\hat{\sigma} \sqrt{a_{jj}}}$$

ki je pod ničelno domnevo porazdeljena t_{n-k-1} .

Rezultate testiranja posamičnih $k + 1$ ničelnih domnev za parametre modela dobimo v povzetku 1m modela. Ti testi so medsebojno odvisni. Hkratnost testiranja odvisnih ničelnih domnev tu ni upoštevana.

Linearni model v matrični obliki

Tabela analize variance

Spomnimo se

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\begin{aligned} SS_{yy} &= SS_{model} + SS_{residual} \\ &= (\mathbf{b}^T \mathbf{X}^T \mathbf{y} - C) + (\mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{X}^T \mathbf{y}) \end{aligned}$$

kjer je $C = (\sum_{i=1}^n y_i)^2 / n$ je t. i. korekcijski člen.

Tabela: Shema tabele ANOVA za spošni linearni model s k regresorji

Vir variabilnosti	df	SS	$MS = SS/df$	F
Model	k	SS_{model}	MS_{model}	$MS_{model}/MS_{residual}$
Ostanek (Residual)	$n - k - 1$	$SS_{residual}$	$MS_{residual}$	
Skupaj	$n - 1$	SS_{yy}		

Linearni model v matrični obliki

F-test za model

Za linearni model z več napovednimi spremenljivkami na podlagi *F*-statistike testiramo ničelno domnevo, da so parametri $(\beta_1, \beta_2, \dots, \beta_k)$ hkrati enaki nič:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_1 : vsaj en parameter β_j , $j = 1, \dots, k$, je različen od nič

Ničelno domnevo testiramo na podlagi *F*-statistike

$$F = \frac{SS_{model}/k}{SS_{residual}/(n - k - 1)}$$

Ob predpostavki $\varepsilon \sim iid N(0, \sigma^2 \mathbf{I})$ je *F*-statistika porazdeljena $F_{k, n-k-1}$.

Linearni model v matrični obliki

Prilagojen koeficient determinacije

V izpisu povzetka 1m modela najdemo poleg koeficiente determinacije

$$R^2 = SS_{model}/SS_{yy} = 1 - SS_{residual}/SS_{yy}$$

tudi **prilagojeni koeficient determinacije** (*Adjusted R-squared*), ki vsebuje tudi informacijo o stopinjah prostosti:

$$R_a^2 = 1 - \frac{\frac{SS_{residual}}{(n-k-1)}}{\frac{SS_{yy}}{(n-1)}} = 1 - \frac{(n-1)\hat{\sigma}^2}{SS_{yy}}$$

R_a^2 je bolj primeren za primerjavo dveh modelov z različnimi napovednimi spremenljivkami kot R^2 . V primerjavi z ostalimi kriteriji za izbiro ustreznega modela (jih še ne poznamo), je njegova uporaba zastarela.

Linearni model v matrični obliki

Napovedi

Za vsak y_i , $i = 1, \dots, n$, imamo vrednosti k napovednih spremenljivk $(x_{i1}, x_{i2}, \dots, x_{ik})$. Označimo z \mathbf{x}_i vektor $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})^T$ in zapišimo

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

Zapišimo še napovedano vrednost za odzivno spremenljivko y_* pri vrednostih napovednih spremenljivk $\mathbf{x}_* = (1, x_{*1}, x_{*2}, \dots, x_{*k})^T$

$$y_* = \mathbf{x}_*^T \boldsymbol{\beta} + \varepsilon_*$$

napaka ε_* ima pričakovano vrednost 0, varianco σ^2 in je neodvisna od ε_i , $i = 1, 2, \dots, n$. Zanimata nas dva intervala zaupanja, najprej za **povprečno napoved** $\mathbf{x}_*^T \boldsymbol{\beta}$ in nato še za **posamično napoved** y_* .

Linearni model v matrični obliki

Varianca povprečne napovedi in matrika \mathbf{H}

Diagonalnim elementom matrike $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ pravimo **vzvodi**. Označimo jih h_{ii} .

Pokazali smo že, da velja $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, oziroma $\hat{y}_i = h_{ii}y_i$.

Torej vzvod predstavlja neko mero vpliva y_i na \hat{y}_i .

Po drugi strani je vzvod h_{ii} odvisen samo od napovednih spremenljivk:

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$$

$\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})^T$ vsebuje komponente i -te vrstice modelske matrike \mathbf{X} .

Linearni model v matrični obliki

Varianca povprečne napovedi in matrika \mathbf{H}

Za varianco prilagojene vrednosti/povprečne napovedi \hat{y} se pokaže, da je sorazmerna s h_{ii} :

$$\begin{aligned} \text{Var}(\hat{y}_i) &= \text{Var}(\mathbf{x}_i^T \mathbf{b}) \\ &= \mathbf{x}_i^T \text{Var}(\mathbf{b}) \mathbf{x}_i = \\ &= \sigma^2 \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \\ &= \sigma^2 h_{ii} \end{aligned}$$

Linearni model v matrični obliki

Varianca povprečne napovedi in matrika \mathbf{H}

Vzvod ima vrednost med $\frac{1}{n}$ in 1, kar pomeni, da je varianca povprečne napovedi vedno manjša od variance napak σ^2 .

Za enostavno linearno regresijo že vemo, da varianco prilagojene vrednosti pri x_i izrazimo

$$Var(\hat{y}(x_i)) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)$$

kar pomeni, da je vzvod

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

Linearni model v matrični obliki

Interval zaupanja za povprečno napoved

Napoved v točki x_* je $\hat{y}_* = \mathbf{x}_*^T \mathbf{b}$, njena pričakovana vrednost je

$$\mathbb{E}(\mathbf{x}_*^T \mathbf{b}) = \mathbf{x}_*^T \boldsymbol{\beta}$$

in njena varianca

$$Var(\mathbf{x}_*^T \mathbf{b}) = \mathbf{x}_*^T Var(\mathbf{b}) \mathbf{x}_* = \sigma^2 \mathbf{x}_*^T \mathbf{A} \mathbf{x}_*$$

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

Linearni model v matrični obliki

Interval zaupanja za povprečno napoved

Ker je napoved $\mathbf{x}_*^T \mathbf{b}$ linearna kombinacija normalno porazdeljenih spremenljivk, velja, da je porazdeljena normalno

$$\mathbf{x}_*^T \mathbf{b} \sim N(\mathbf{x}_*^T \boldsymbol{\beta}, \sigma^2 \mathbf{x}_*^T \mathbf{A} \mathbf{x}_*)$$

Velja

$$\frac{\mathbf{x}_*^T \mathbf{b} - \mathbf{x}_*^T \boldsymbol{\beta}}{\hat{\sigma} \sqrt{\mathbf{x}_*^T \mathbf{A} \mathbf{x}_*}} \sim t_{n-k-1}$$

in $(1 - \alpha)100\%$ interval zaupanja za povprečno napoved je

$$\left(\mathbf{x}_*^T \mathbf{b} - |t_{\frac{\alpha}{2}; n-k-1}| \hat{\sigma} \cdot \sqrt{\mathbf{x}_*^T \mathbf{A} \mathbf{x}_*}, \mathbf{x}_*^T \mathbf{b} + |t_{\frac{\alpha}{2}; n-k-1}| \hat{\sigma} \cdot \sqrt{\mathbf{x}_*^T \mathbf{A} \mathbf{x}_*} \right)$$

Linearni model v matrični obliki

Interval zaupanja za posamično napoved

Izrazimo razliko med pravo napovedjo in njenou oceno ter varianco te razlike:

$$y_* - \hat{y}_* = \mathbf{x}_*^T \boldsymbol{\beta} + \varepsilon_* - \mathbf{x}_*^T \mathbf{b}$$

Velja $\mathbb{E}(y_* - \hat{y}_*) = 0$ in ε_* in \mathbf{b} sta neodvisna.

$$\begin{aligned} \text{Var}(y_* - \hat{y}_*) &= \text{Var}(\mathbf{x}_*^T \mathbf{b}) + \text{Var}(\varepsilon_*) \\ &= \sigma^2 \mathbf{x}_*^T \mathbf{A} \mathbf{x}_* + \sigma^2 \\ &= \sigma^2 (1 + \mathbf{x}_*^T \mathbf{A} \mathbf{x}_*) \end{aligned}$$

Linearni model v matrični obliki

Interval zaupanja za posamično napoved

Tudi tu lahko pokažemo, da je

$$\frac{y_* - \hat{y}_*}{\hat{\sigma} \sqrt{1 + \mathbf{x}_*^T \mathbf{A} \mathbf{x}_*}} \sim t_{n-k-1}$$

in $(1 - \alpha)100\%$ interval zaupanja za posamično napoved je

$$\left(\hat{y}_* - |t_{\frac{\alpha}{2}; n-k-1}| \hat{\sigma} \sqrt{1 + \mathbf{x}_*^T \mathbf{A} \mathbf{x}_*}, \quad \hat{y}_* + |t_{\frac{\alpha}{2}; n-k-1}| \hat{\sigma} \sqrt{1 + \mathbf{x}_*^T \mathbf{A} \mathbf{x}_*} \right)$$

Linearni model v matrični obliki

Interpretacija ocen parametrov linearnega modela z več regresorji

Interpretacijo ocen parametrov linearnega modela z več regresorji si poglejmo najprej na primeru dveh številskih regresorjev x_1 in x_2 . Zamislimo si pričakovano vrednost tega modela v točki (x_{01}, x_{02}) .

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

in v točki $(x_{01}, x_{02} + 1)$

$$\mathbb{E}(y|x_{01}, x_{02}+1) = \beta_0 + \beta_1 x_{01} + \beta_2(x_{02}+1) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_2,$$

sledi

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

Torej velja, če x_{02} povečamo za eno enoto in ostane izbrana vrednost x_{01} nespremenjena, se pričakovana vrednost y poveča za β_2 .

Linearni model v matrični obliki

Interpretacija ocen parametrov linearnega modela z več regresorji

V linearinem modelu z več regresorji **ima vsak regresor "pogojni vpliv"**: če regresor x_j povečamo za eno enoto, se pogojno na konstantne vrednosti vseh ostalih regresorjev v modelu pričakovana vrednost odzivne spremenljivke poveča za β_j enot.

Pogojni vpliv regresorja x_j v modelu z več regresorji je lahko zelo drugačen, kot je njegov "robni" vpliv na odzivno spremenljivko, ko je x_j edini regresor v modelu. Prisotnost ostalih regresorjev lahko povzroči spremembo velikosti, lahko pa tudi spremembo predznaka parametra β_j .

Linearni model v matrični obliki

Interpretacija ocen parametrov linearnega modela z več regresorji

Geometrijska predstavitev enostavne linearne regresije je **premica** v dvodimenzionalnem prostoru, za model z dvema regresorjema je **ravnina** v tridimenzionalnem prostoru, za model s k regresorji pa je to **hiper ravnina** v $k + 1$ dimenzionalnem prostoru.

V regresijski analizi pogosto modeliramo vpliv izbrane spremenljivke na odzivno spremenljivko ob upoštevanju (*controlling for*) določenih t. i. **motečih spremenljivk** (*confounding variables*) v modelu. Zanima nas vpliv te izbrane napovedne spremenljivke, vendar vemo, da je odzivna spremenljivka odvisna tudi od nekaterih drugih spremenljivk, ki pa niso predmet naše raziskave.

Linearni model v matrični obliki

F-test za primerjavo gnezdenih modelov

Model.1 je **gnezden znotraj** Model.2 če velja:

Model.1

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i,$$

Model.2

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \beta_{k+1} x_{i(k+1)} + \dots + \beta_{k+r} x_{i(k+r)} + \varepsilon_i.$$

Zanima nas, ali sta tako modela ekvivalentna oziroma ali je model z več členi v statističnem smislu boljši.

Linearni model v matrični obliki

F -test za primerjavo gnezdenih modelov

H_0 : Model.1 in Model.2 sta ekvivalentna:

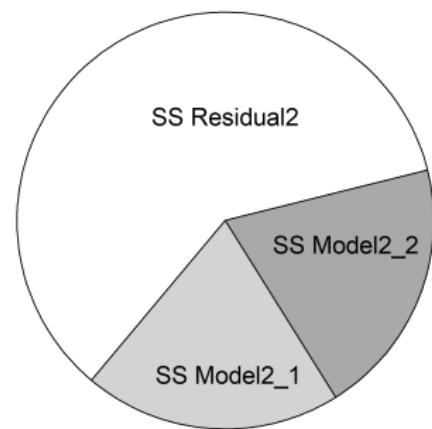
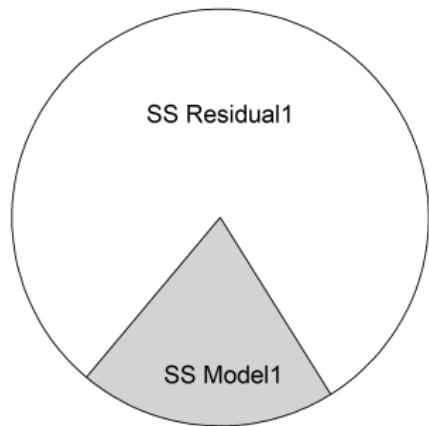
$$H_0 : \beta_{k+1} = \beta_{k+2} = \dots = \beta_{k+r} = 0.$$

H_1 : Model.2 je boljši kot Model.1:

$$H_1 : \text{vsaj en } \beta_{k+j} \neq 0, \quad j = 1, \dots, r.$$

Linearni model v matrični obliki

F-test za primerjavo gnezdenih modelov



Linearni model v matrični obliki

F-test za primerjavo gnezdenih modelov

Statistično sklepanje temelji na F -statistiki:

$$F = \frac{\frac{SS_{\text{residual1}} - SS_{\text{residual2}}}{df_{\text{residual1}} - df_{\text{residual2}}}}{\frac{SS_{\text{residual2}}}{df_{\text{residual2}}}} \sim F_{df_{\text{residual1}} - df_{\text{residual2}}, df_{\text{residual2}}}$$

Če se modela razlikujeta za r parametrov

$$F = \frac{\frac{SS_{\text{residual1}} - SS_{\text{residual2}}}{r}}{\frac{SS_{\text{residual2}}}{n - k - r - 1}}.$$

R: funkcija `anova(model1, model2)`, prvi model je gnezdeni model. Oba modela morata biti narejena na istih podatkih.

Linearni model v matrični obliki

Sekvenčni F -testi funkcije anova

Funkcija anova na 1m modelu z več napovednimi spremenljivkami, vrne **sekvenčne vsote kvadratov ostankov modela** in rezultate **sekvenčnih F -testov**.

Sekvenčni F -test testira vpliv posamezne spremenljivke ob upoštevanju predhodnih spremenljivk v modelu.

Kaj se testira v posamezni vrstici izpisa?

- Prva vrstica: vsota kvadratov ostankov za model
 $y_i = \beta_0 + \beta_1 x_{1i}$, označimo jo $SS_{\beta_1 | \beta_0}$.

Z F -testom testiramo:

$$H_0 : \beta_1 = 0$$

$$H_0 : \text{modela } y_i = \beta_0 \text{ in } y_i = \beta_0 + \beta_1 x_{1i} \text{ sta ekvivalentna.}$$

Linearni model v matrični obliki

Sekvenčni F -testi funkcije anova

- Druga vrstica: razlika vsot kvadratov ostankov za model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ in za model $y_i = \beta_0 + \beta_1 x_{1i}$, označimo jo $SS_{\beta_2|\beta_0,\beta_1}$.

$H_0 : \beta_2 = 0$ ob upoštevanju β_1 v modelu.

H_0 : modela $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ in $y_i = \beta_0 + \beta_1 x_{1i}$ sta ekvivalentna.

Linearni model v matrični obliki

Sekvenčni F -testi funkcije anova

- Če je v modelu k napovednih spremenljivk, se izpiše k vrstic z razlikami vsot kvadratov ostankov:

$$SS_{\beta_1|\beta_0}$$

$$SS_{\beta_2|\beta_0,\beta_1}$$

...

$$SS_{\beta_k|\beta_0,\dots,\beta_{k-1}}$$

- v i -ti vrstici se izvede F -test na podlagi F -statistike,
 $i = 1, \dots, k$:

$$\frac{SS_{\beta_i|\beta_0,\dots,\beta_{i-1}}}{SS_{\beta_{i-1}|\beta_0,\dots,\beta_{i-2}}/(n-i-1)} \sim F_{1,n-i-1}.$$

Testira se ničelna domneva $H_0 : \beta_i = 0$ ob upoštevanju $i - 1$ napovednih spremenljivk v modelu.

Linearni model v matrični obliki

Sekvenčni F -testi funkcije anova

- Če je napovedna spremenljivka opisna z d različnimi vrednostmi, se v modelu ocenjuje $d - 1$ parametrov in z F -testom testiramo ničelno domnevo, da je vseh $d - 1$ parametrov enakih 0:

$$\frac{SS_{\beta_i, \dots, \beta_{i+d-1} | \beta_0, \dots, \beta_{i-1}}}{SS_{\beta_{i-1} | \beta_0, \dots, \beta_{i-2}} / (n - i - d - 1)} \sim F_{d-1, n-i-d-1}.$$

Izpis funkcije `anova()` za linearni model je odvisen od vrstnega reda napovednih spremenljivk v modelu.

Diagnostika linearrega modela

Preverjanje predpostavk linearrega modela:

- **linearost**; enostvana regresija: razsevni grafikon y glede na x ; multipla regresija: "grafikoni parcialnih ostankov";
- **pričakovana vrednost napak je 0**, gladilnik na sliki ostankov glede na prilagojene vrednosti se mora čim bolje prilegati abscisi;
- **varianca napak je konstantna** (slika ostankov ali transformiranih standardiziranih ostankov glede na napovedane vrednosti);
- **porazdelitev napak je normalna** (kvantilni graf za standardizirane ostanke);
- **napake so medsebojno neodvisne** (težko preveriti, ustrezni način pridobivanja podatkov, princip slučajnosti; če so podatki izmerjeni v času, ostanke narišemo glede na čas meritve).

Diagnostika linearrega modela

V postopku diagnostike modela uporabljamo grafične prikaze:

- ostankov e
- standardiziranih ostankov e_s
- studentiziranih ostankov e_t
- posebnih točk (regresiski osamelci, vplivne točke, vzvodne točke)

Diagnostika linearrega modela

Ostanki in njihove lastnosti

Ostanki

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

Pričakujemo, da imajo ostanki podobne lastnosti kot napake $\varepsilon \sim iid N(0, \sigma^2)$: neodvisnost, konstantna varianca, normalna porazdelitev.

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

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

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

Ker velja, da so y_i normalno porazdeljene slučajne spremenljivke, to velja tudi za ostanke.

Diagnostika linearrega modela

Ostanki in njihove lastnosti

Poiščimo zvezo med ostanki in napakami:

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

Ker je $\mathbf{H}\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X}$, sledi zveza med ostanki in napakami:

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

Diagnostika linearrega modela

Ostanki in vzhodi

Posamezen ostanek e_i zapišemo

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

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i.$$

Enostavna linearna regresija:

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

Diagnostika linearrega modela

Ostanki in vzvodi

Za **vzvode** velja:

- h_{ii} , $i = 1, \dots, n$ so diagonalni elementi matrike \mathbf{H}
- vzvod je odvisen od n , od položaja točke v regresorskem prostoru in od SS_{XX}
- h_{ii} zavzemajo vrednosti med $1/n$ in 1
- $\sum_i^n h_{ii} = k + 1$, kjer je $k + 1$ število parametrov v modelu.

Z večanjem n se elementi matrike \mathbf{H} približujejo vrednosti 0 in ostanki postanejo dobra aproksimacija za napake.

Diagnostika linearrega modela

Varianca ostankov

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

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

Varianca ostankov **ni konstantna**, odvisna je od matrike \mathbf{H} .

Diagnostika linearrega modela

Ostanki in njihove lastnosti

Lastnosti variance ostankov:

- varianca ostankov ni konstantna, odvisna je od matrike \mathbf{H} , kar pomeni, da je varianca posameznega ostanka je odvisna od položaja točke v regresorskem prostoru;
- kovarianca ostankov $Cov(e_i, e_j) = -\sigma^2 h_{ij}$, za $i \neq j$ je odvisna od vrednosti h_{ij} , njena vrednost se bliža vrednosti 0, ko velikost vzorca n narašča;
- za ostanke e_i velja, da so v absolutnem smislu manjši kot napake ε_i , saj so vzvodi h_{ii} po definiciji pozitivne vrednosti; tudi varianca ostankov $Var(e_i) = \sigma^2(1 - h_{ii})$ je vedno manjša kot varianca napak $Var(\varepsilon_i)$.

Diagnostika linearrega modela

Ostanki in njihove lastnosti

- točka z velikim vzvodom ima ostanek z majhno varianco in potencialno lahko predstavlja vplivno točko, ki potegne prilegano premico ali ravnino k sebi, da s tem zagotovi manjšo vrednost ostanka;
- zaradi naštetih lastnosti ostanki niso najboljše vrednosti za diagnostiko modela (standardizirani ostanki, studentizirani ostanki);
- ker velja $\text{Cov}(\mathbf{e}, \hat{\mathbf{y}}) = 0$, lahko na podlagi grafa ostankov glede na prilagojene vrednosti preverjamo predpostavko linearnosti zveze med \mathbf{y} in regresorji.

Diagnostika linearnega modela

Ostanki in njihove lastnosti, iz gradiva P2

Izrek 2.2: če je varianca napak $\text{Var}(\varepsilon) = \sigma^2 \mathbf{I}$ so ostanki \mathbf{e} nekorelirani s prilagojenimi vrednostmi odzivne spremenljivke $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ oziroma $\text{Cov}(\mathbf{e}, \hat{\mathbf{y}}) = 0$.

Dokaz:

$$\begin{aligned}\text{Cov}(\mathbf{e}, \hat{\mathbf{y}}) &= \text{Cov}((\mathbf{I} - \mathbf{H})\mathbf{y}, \mathbf{H}\mathbf{y}) \\ &= (\mathbf{I} - \mathbf{H})(\sigma^2 \mathbf{I}) \mathbf{H}^T \\ &= \sigma^2(\mathbf{H}^T - \mathbf{H}\mathbf{H}^T) \\ &= 0\end{aligned}$$

Posledica: razsevni grafikon ostankov glede na prilagojene vrednosti je dobro diagnostično orodje za regresijski model. Če na grafu vidimo odvisnost ostankov od prilagojenih vrednosti, model ne ustreza predpostavkam linearnega modela.

Diagnostika linearrega modela

Standardizirani ostanki

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

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

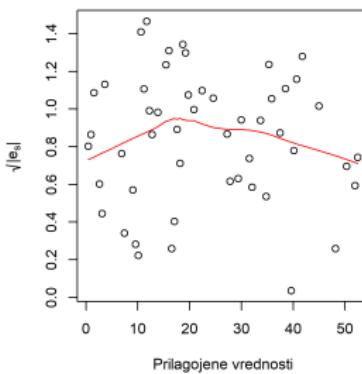
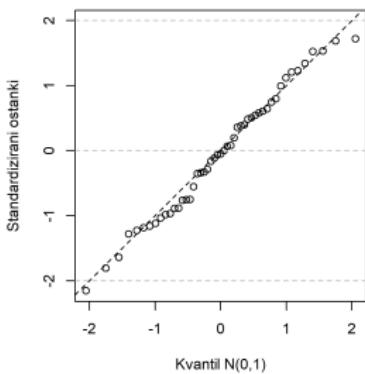
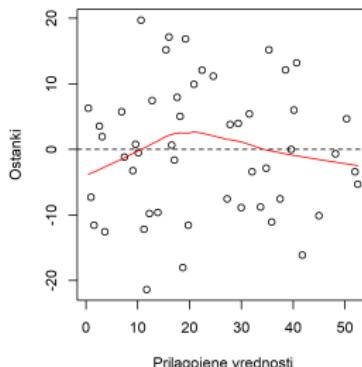
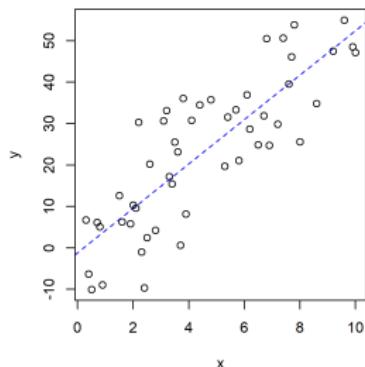
Ker σ v splošnem ne poznamo, jo ocenimo z $\hat{\sigma}$, tako izračunane standardizirane ostanke imenujemo tudi notranje studentizirani ostanki (*internally studentized residuals*).

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

Če so predpostavke modela izponjene, imajo standardizirani ostanki konstantno varianco. Porazdelitev standardiziranih ostankov je približno t_{n-k-1} ; če pa je $n \gg k$, je porazdelitev približno $N(0, 1)$.

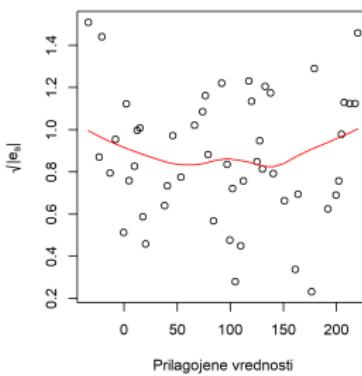
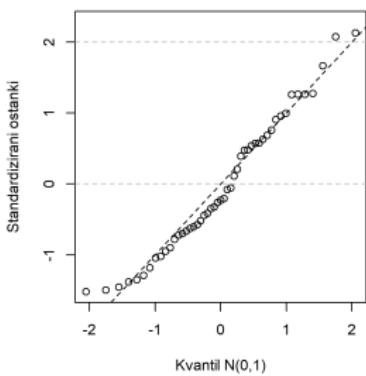
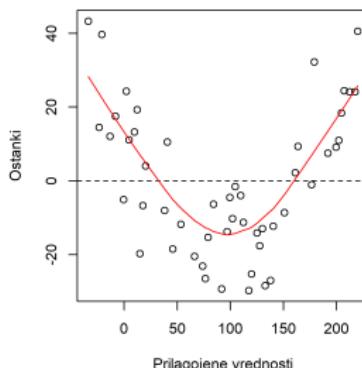
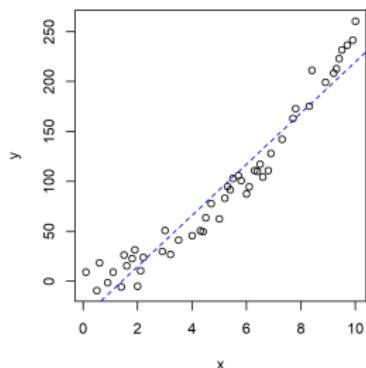
Diagnostika linearega modela

Primer grafičnih prikazov, ko so predpostavke linearega modela izpolnjene



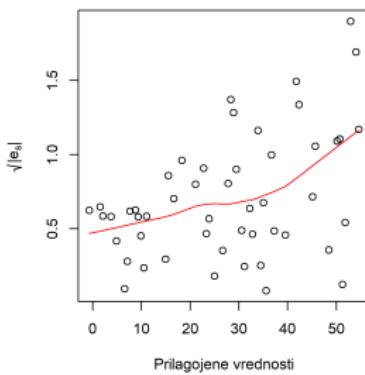
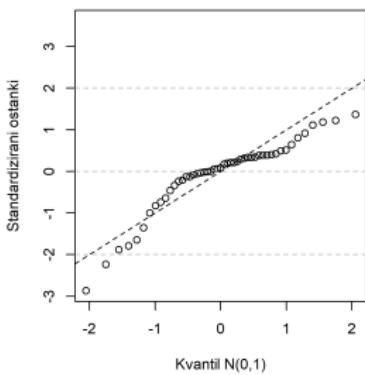
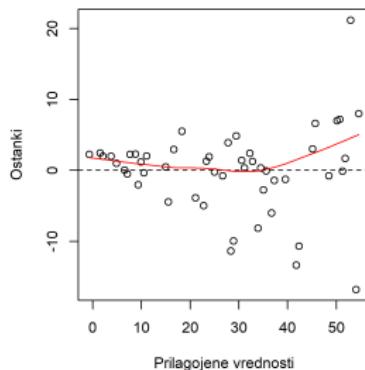
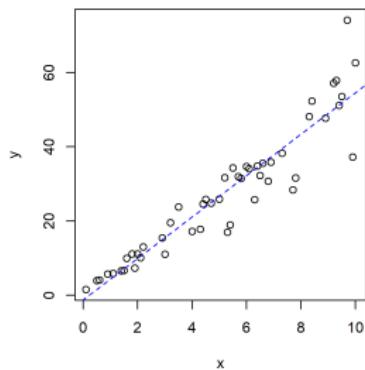
Diagnostika linearrega modela

Predpostavka o linearni zvezi ni izpolnjena



Diagnostika linearega modela

Predpostavka o konstantni varianci ni izpolnjena



Diagnostika linearrega modela

Studentizirani ostanki

Studentizirani ostanki

Povezanosti med števcem in imenovalcem pri e_{s_i} , $i = 1, \dots, n$, se znebimo z izračunom studentiziranih ostankov. Cenilka za σ se izračunana brez upoštevanja i -te točke:

$$e_{t_i} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(-i)} \cdot \sqrt{1 - h_{ii}}} \sim t_{n-k-2}$$

$\hat{\sigma}_{(-i)}$ se izračunana tako, da je v regresijskem modelu i -ta točka izpuščena. Posledično sta števec in imenovalec neodvisna.

Studentizirani ostanki so primernejši za za odkrivanje regresijskih osamelcev kot standardizirani ostanki, saj je $\hat{\sigma}_{(-i)}$ v primeru zelo odstopajoče vrednosti znatno manjša od $\hat{\sigma}$, kar poveča vrednost studentiziranega ostanka.

Diagnostika linearrega modela

Graf dodane spremenljivke (*added variable plot* ali *partial regression plot*)

Graf dodane spremenljivke je prikaz vpliva posameznega regresorja na odzivno spremenljivko ob upoštevanju ostalih regresorjev v modelu.

Za model:

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

graf dodane spremenljivke x_j naredimo na podlagi **ostankov dveh modelov**:

Prvi model za y_i brez regresorja x_j :

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

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

Diagnostika linearrega modela

Graf dodane spremenljivke (*added variable plot* ali *partial regression plot*)

Drugi model za x_j v odvisnosti od ostalih regresorjev:

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

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

Ostanki $e_{i,y}^{(-j)}$ in $e_{i,x_j}^{(-j)}$ predstavljajo vrednosti y in x_j "očiščene" za vpliv ostalih spremenljivk v modelu.

Diagnostika linearrega modela

Graf dodane spremenljivke

Graf dodane spremenljivke narišemo kot razsevni grafikon za odvisnost $e_{i,y}^{(-j)}$ od $e_{i,x_j}^{(-j)}$ (funkcija `avPlot` iz paketa `car`).

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 linearrega modela z več napovednimi spremenljivkami tudi v kontekstu analize nekonstantne variance in vplivnih točk.

Diagnostika linearrega modela

Graf parcialnih ostankov

Graf parcialnih ostankov (*Partial Residual Plots*)

Ta grafikon omogoča preverjanje linearnosti ozziroma prisotnost nelinearnosti v modelu z več napovednimi spremenljivkami (funkcija `crPlots()` iz paketa `car`).

Za model $y_i = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon_i$ izračunamo **parcialne ostanke** za vsako od napovednih spremenljivk x_j , $j = 1, \dots, k$:

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

Diagnostika linearrega modela

Graf parcialnih ostankov

Graf parcialnih ostankov prikazuje parcialne ostanke e_{i,x_j} v odvisnosti od x_{ij} .

Na grafu je tudi gladilnik dobljen z neparametrično regresijo, ki jo izračuna funkcija `lowess()`.

Ta graf pokaže morebitno nelinearnost v zvezi y in x_j , ki je nismo zaobjeli v linearinem 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`.

Primer PACIENTI v gradivu, skriptna datoteka `primerP3.R`.

Diagnostika linearrega modela

Posebne točke

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

Pogledali bomo tri vrste posebnih točk:

- **regresijski osamelci**
- **vzvodne točke**
- **vplivne točke**

Diagnostika linearrega modela

Regresijski osamelec

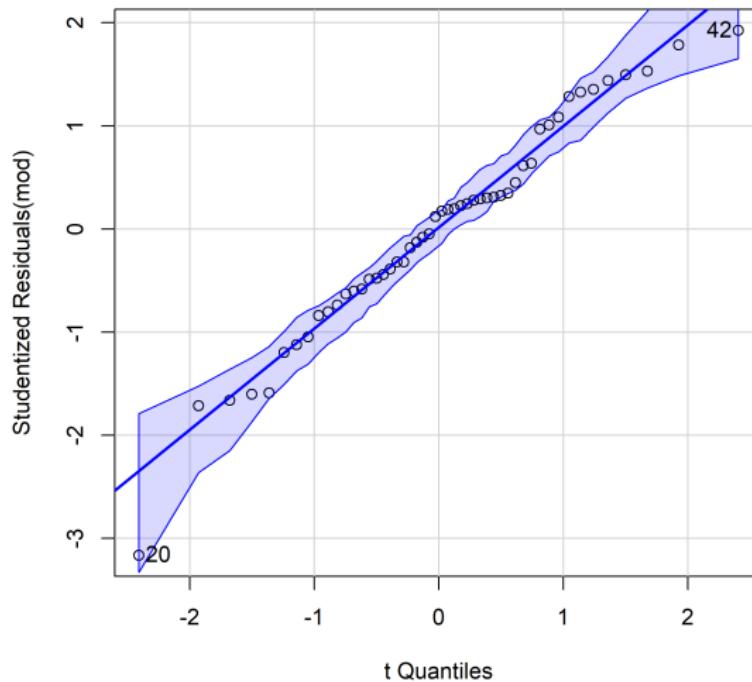
Regresijski osamelec je točka, pri kateri vrednost spremenljivke y_i močno odstopa od pripadajoče napovedane vrednosti \hat{y}_i .

Regresijske osamelce ugotavljamo na osnovi studentiziranih ostankov: grafični način ali z modelom.

Funkcija `qqPlot` iz paketa `car` nariše studentizirane ostanke glede na kvantile t_{n-k-2} in pripadajočo 95 % ovojnicu, ki je izračunana s parametričnim bootstrap pristopom (Aitkinson, 1985). Točke, ki ležijo izrazito izven ovojnici, so regresijski osamelci.

Diagnostika linearrega modela

Regresijski osamelec, qqPlot()



Diagnostika linearrega modela

Regresijski osamelec, določanje z modelom

Model za ugotavljanje regresijskih osamelcev (*Mean-shift outlier model*) za i -to točko:

$$y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \gamma d_i + \varepsilon_i, \quad i = 1, \dots, n,$$

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

Za vsako točko posebej, $i = 1, \dots, n$, preverjamo ničelno domnevo, da :

$$H_{0i} : \gamma = 0 \quad i\text{-ta točka ni regresijski osamelec}$$

$$H_{1i} : \gamma \neq 0 \quad i\text{-ta točka je regresijski osamelec}$$

Če velja $\gamma \neq 0$, se presečišče premakne iz α na $\alpha + \gamma$, ob upoštevanju enake odvisnosti y od (x_1, \dots, x_k) kot velja za ostale točke.

Diagnostika linearrega modela

Regresijski osamelec, določanje z modelom

Teorija pokaže, da je testna statistika pod ničelno domnevo studentizirani ostanek za i -to točko

$$e_{t_i} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(-i)} \cdot \sqrt{1 - h_{ii}}} \sim t_{n-k-2}$$

Naredimo torej n odvisnih testov zato je treba dobljene p -vrednosti prilagoditi. Funkcija `outlierTest` iz paketa `car` vrne po Bonferroniju popravljene p -vrednosti.

Diagnostika linearrega modela

Vzvodne točke

Vzvodne točke so daleč od centra regresorskega prostora, imajo **velik vzvod**, vrednost h_{ii} .

Velja: $\sum_i^n h_{ii} = k + 1$, kjer je $k + 1$ število parametrov v modelu.

Povprečni vzvod je:

$$\bar{h} = \frac{k + 1}{n}.$$

Za i -to točko, ki ima vzvod h_{ii} večji od dvakratnika povprečnega vzvoda, pravimo, da je **vzvodna točka**:

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

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

Diagnostika linearrega modela

Vplivne točke

Točka $(y_i, x_{i1}, \dots, x_{ik})$ je vplivna, če se ocene parametrov modela \mathbf{b} ali pa z modelom prilagojene vrednosti \hat{y}_i , $i = 1, \dots, n$ bistveno spremenijo, če jo izločimo iz modela. Vplivna točka lahko vpliva na statistično sklepanje za parametre modela.

Vplivnost posamezne točke vrednotimo z različnimi merami, ki temeljijo na:

- razlikah $(\mathbf{b}_{(-i)} - \mathbf{b})$, kjer je $\mathbf{b}_{(-i)}$ vektor ocen parametrov v modelu, kjer i – to točko izločimo (**Cookova razdalja, DFBETAS**)
- razlikah napovedi $(\hat{y}_i - \hat{y}_{i(-i)})$, $i = 1, \dots, n$, kjer je $\hat{y}_{i(-i)}$ napoved v i -ti točki za model, ki i – te točke pri oceni parametrov ne upošteva (**DFFITS**).

Diagnostika linearrega modela

Vplivne točke, Cookova razdalja

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

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

$\hat{\sigma}^2$ je ocena za varianco napak.

Ta razdalja je osnovana na podlagi skupnega območja zaupanja za vektor parametrov modela β . Če je Cookova razdalja večjo od 0,5, vektor $\mathbf{b}_{(-i)}$ pade izven 50 % skupnega območja zaupanja za β , za model za vse podatke.

Točka je vplivna, če ima Cookovo razdaljo večjo od , $D_i > 1$.

Diagnostika linearrega modela

Vplivne točke, Cookova razdalja

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

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

Točka z veliko vrednostjo standardiziranega ostaneka in hkrati z velikim vzvodom ima velik vpliv na ocene parametrov in posledično tudi na modelske napovedi.

Diagnostika linearrega modela

Vplivne točke, Cookova razdalja

Cookovo razdaljo lahko zapišemo tudi na osnovi prilagojenih vrednosti:

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

Cookova razdalja je torej skalirana Evklidska razdalja med vektorjem napovedi modela narejenega na vseh podatkih in vektorjem napovedi modela na podatkih, kjer je i -ta točka izločena.

Točke z veliko Cookovo razdaljo identificiramo na četrtem diagnostičnem grafikonu za model. Na razsevnem grafikonu standardiziranih ostankov in vzvodov sta prikazani izoliniji za Cookovo razdaljo z vrednostma 0.5 in 1.

Primer PADAVINE, primerP3.R

Predpostavke niso izpolnjene

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

- transformacije
- modeliranje

Predpostavke niso izpolnjene

Pregled metod

Pregled metod:

- Konstantna varianca napak
 - transformacija odzivne spremenljivke
 - modeliranje: tehtana metoda najmanjših kvadratov, interakcijski členi
- Linearnost
 - transformacija odzivne spremenljivke in/ali napovednih spremenljivk
 - modeliranje: polinomska regresija, interakcijski členi
 - modeliranje: zlepki, aditivni modeli (*splines, additive models*)
- Vplivnost točk (ni predpostavka modela, lahko posledica neizpolnjenih predpostavk)
 - transformacija napovednih spremenljivk
 - modeliranje: tehtana metoda najmanjših kvadratov, interakcijski členi

Predpostavke niso izpolnjene

Pregled metod

- Normalna porazdelitev ostankov
 - transformacija odzivne spremenljivke
 - modeliranje: posplošeni linearni modeli (GLM) (ne bomo obravnavali)
- Neodvisnost napak
 - diferenciranje (časovne vrste)
 - modeliranje: linearni mešani modeli (longitudinalni, hierarhični)
 - modeliranje: GEE modeli *Generalised Estimating Equations*
 - modeliranje: kopule (*copulas*)

Predpostavke niso izpolnjene

Transformacije

Transformacije

Transformiramo lahko odzivno in/ali napovedne spremenljivke.

Za vsako transformacijo je v procesu modeliranja potrebno ugotoviti, ali smo problem res rešili:

- če nas pri modeliranju zanima zveza med odzivno in napovednimi spremenljivkami ali pa gre za statistično sklepanje, potem naredimo model na transformirani spremenljivki in ponovno izvedemo diagnostiko modela;
- če modeliramo z namenom napovedovanja, potem ustreznost izbire transformacije preverimo na podlagi PRESS-ostankov. Primerjamo vsoto kvadratov PRESS ostankov modela na transformiranih in netransformiranih podatkih (o tem bomo več govorili v poglavju o izbiri modela).

Predpostavke niso izpolnjene

Klasične transformacije odzivne spremenljivke

Tabela: 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
$\sigma^2 \propto \frac{\mathbb{E}(y)}{1-\mathbb{E}(y)}$	$\arcsin(\sqrt{y})$, $\text{logit}(y)$	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}	$y \neq 0$

Predpostavke niso izpolnjene

Klasične transformacije odzivne spremenljivke

Transformacija za delež

y je delež, omejena zaloga vrednosti na intervalu $[0,1]$. V ozadju je slučajna spremenljivka z :

$$z \sim b(n, \pi), \quad \mathbb{E}(z) = n\pi, \quad \text{Var}(z) = n\pi(1 - \pi)$$

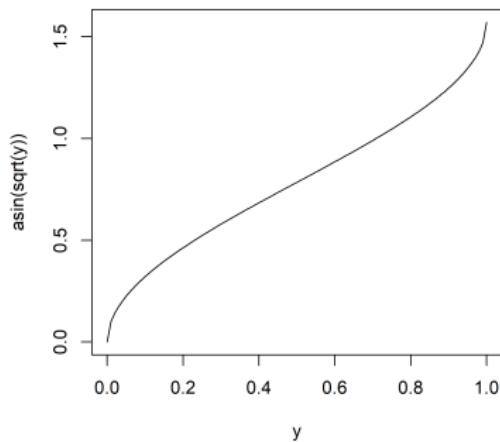
Varianca deležev blizu 0 oz. blizu 1 je manjša od variance deležev blizu $1/2$. Z nekonstantno varianco ni težav, če so vrednosti deležev približno na intervalu $[0.25, 0.75]$.

Transformaciji: $\text{asin}(\sqrt{y})$ in $\text{logit}(y)$.

Predpostavke niso izpolnjene

Klasične transformacije odzivne spremenljivke

$$f(y) = \text{asin}(\sqrt{y}) \quad y \in [0, 1]$$

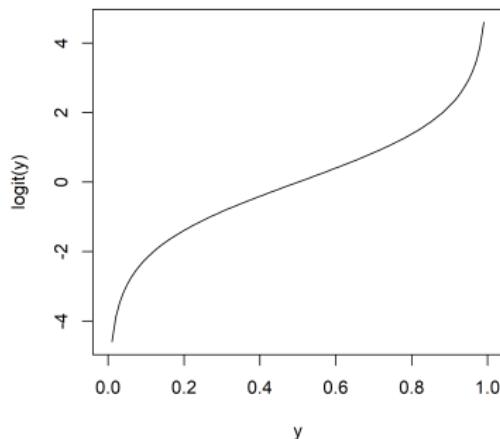


Pokažemo lahko, da je varianca transformiranih vrednosti približno $\frac{1}{4}n$ in je tako neodvisna je od π . To pomeni, da ta transformacija stabilizira varianco.

Predpostavke niso izpolnjene

Klasične transformacije odzivne spremenljivke

$$f(y) = \text{logit}(y) = \ln \frac{y}{1-y} \quad y \in (0, 1)$$



Logit transformacija je osnova logistični regresiji (GLM).

Predpostavke niso izpolnjene

Box-Cox transformacije

Box-Cox transformacije (Box in Cox, 1964)

Družino transformacij za odvisno spremenljivko y , ki je funkcija parametra λ . Za i -to točko $i = 1, \dots, n$ v tem primeru linearni model zapišemo:

$$z_i = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} = (\mathbf{X}\boldsymbol{\beta})_i + \varepsilon_i & , \lambda \neq 0 \\ \ln(y_i) = (\mathbf{X}\boldsymbol{\beta})_i + \varepsilon_i & , \lambda = 0 \end{cases} \quad \varepsilon \sim N(0, \sigma^2 \mathbf{I})$$

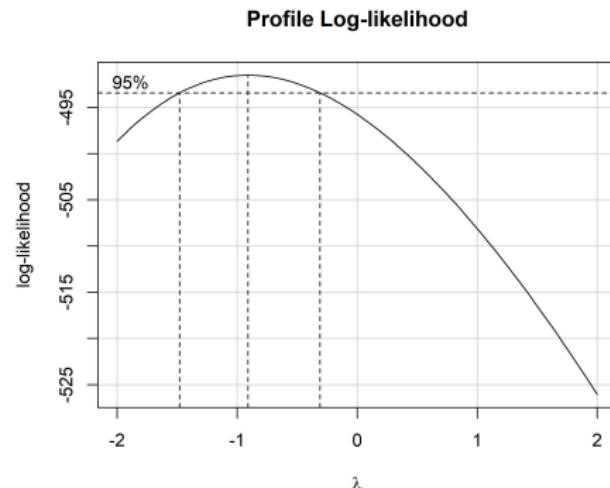
Box-Cox transformacije so definirane za $\mathbf{y} > 0$.

Z metodo največjega verjetja (*maximum likelihood*) hkrati ocenujemo $\boldsymbol{\beta}$ in λ .

Predpostavke niso izpolnjene

Box-Cox transformacije

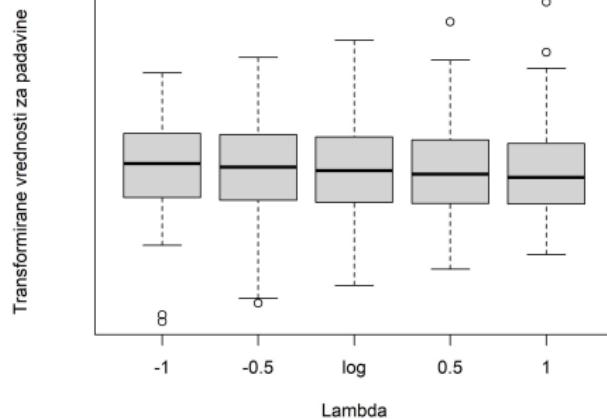
Oceno $\hat{\lambda}$ dobimo z numeričnim postopkom: za različne vrednosti $\hat{\lambda}$ izračunamo verjetje $L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \lambda)$. Izberemo λ , pri kateri ima logaritem verjetja maksimalno vrednost.



Primer: postaje, funkcija `boxCox()`

Predpostavke niso izpolnjene

Box-Cox transformacije



Slika: Okvirji z ročaji za različne transformacije odzivne spremenljivke, približna izbira parametra λ

Primer: postaje, funkcija `symbox()`

Predpostavke niso izpolnjene

Nelinearnost, ki se da linearizirati

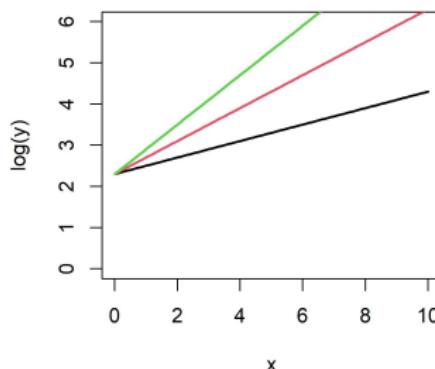
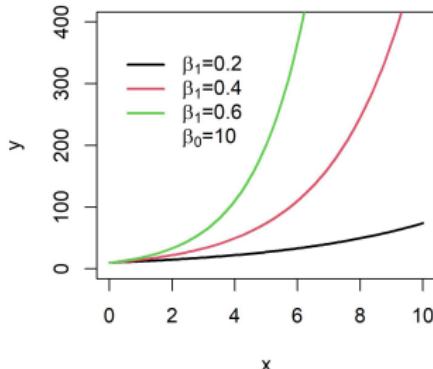
Exponentna zveza

Če zvezo med y in x opišemo z eksponentno funkcijo:

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

z logaritmiranjem izraza dobimo linearno zvezo:

$$\ln(y) = \ln(\beta_0) + \beta_1 x = \beta_0^* + \beta_1 x$$



Predpostavke niso izpolnjene

Nelinearnost, ki se da linearizirati, eksponentna zveza

Pomen parametra β_1 :

$$\beta_1 = \ln(y(x+1)) - \ln(y(x)) = \ln \frac{y(x+1)}{y(x)}$$

$$\frac{y(x+1)}{y(x)} = e^{\beta_1} \quad \frac{y(x+1) - y(x)}{y(x)} = e^{\beta_1} - 1.$$

Če se x poveča za eno enoto, se y spremeni za $100 \cdot (e^{\beta_1} - 1)$ %.

Pri $x = 0$ je povprečna vrednost $\bar{y} = e^{\beta_0^*}$

Predpostavke niso izpolnjene

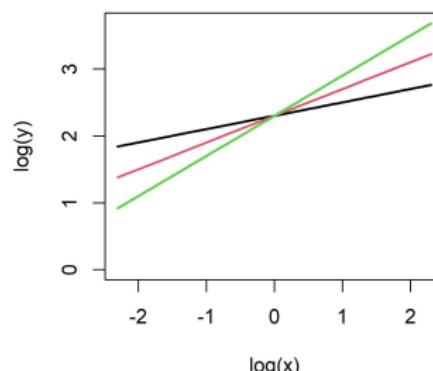
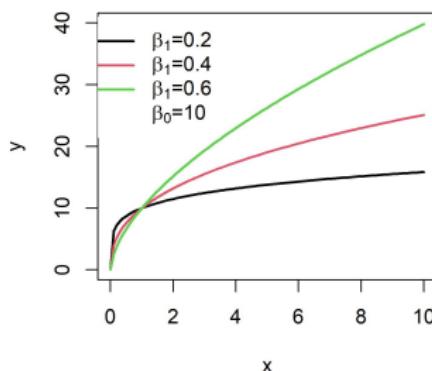
Nelinearnost, ki se da linearizirati, multiplikativna zveza

Multiplikativna zveza

$$y = \beta_0 x^{\beta_1},$$

Lineariziramo z logaritmiranjem:

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



Predpostavke niso izpolnjene

Nelinearnost, ki se da linearizirati, multiplikativna zveza

Pomen parametra β_1 :

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

Če se x poveča za 1 %, se y spremeni za β_1 %.

Predpostavke niso izpolnjene

Nelinearnost, ki se da linearizirati, multiplikativna zveza

Dva regresorja x_1 in x_2 v multiplikativnem odnosu z y :

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

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

Če se x_1 poveča za 1 %, se y spremeni za β_1 %, ko je vrednost x_2 konstantna.

Predpostavke niso izpolnjene

Vloga napake v modelu za transformirane spremenljivke

Eksponentna zveza, kjer je napaka v aditivi zvezi z odzivno spremenljivko:

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

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 .

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

Predpostavke niso izpolnjene

Vloga napake v modelu za transformirane spremenljivke

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

Primer: kovine

Predpostavke niso izpolnjene

Interakcija dveh številskeih napovednih spremenljivk

Linearni model z interakcijo dveh številskeih spremenljivk

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

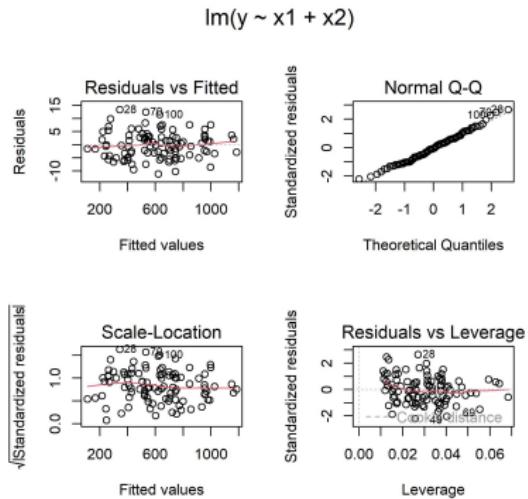
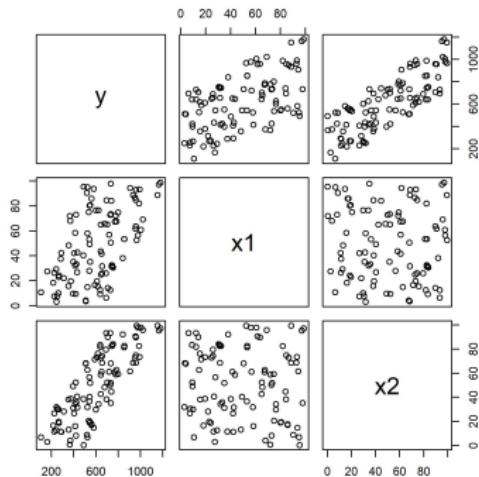
Interakcijske člene v model vključimo iz različnih razlogov:

- poznavanje vpliva izbranih dejavnikov na odzivno spremenljivko
- iskanje utreznih regresorjev v linearinem modelu da izponimo predpostavke.

Predpostavke niso izpolnjene

Primer generiranih podatkov **brez interakcije** dveh napovednih spremenljivk

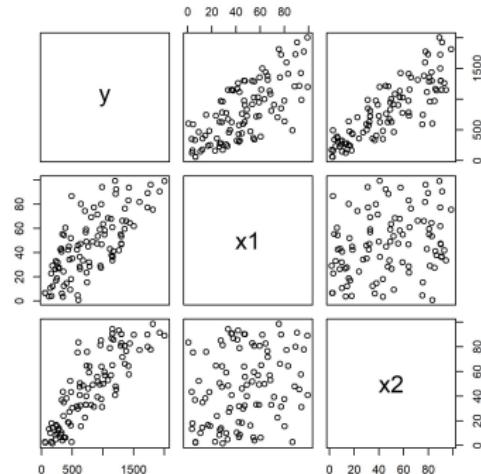
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$



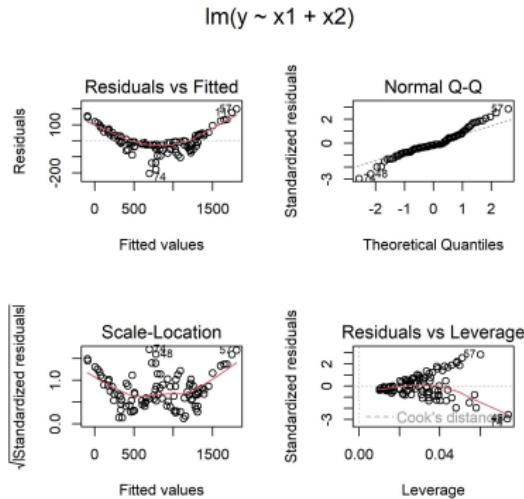
Predpostavke niso izpolnjene

Primer generiranih podatkov z **interakcijo** dveh napovednih spremenljivk

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



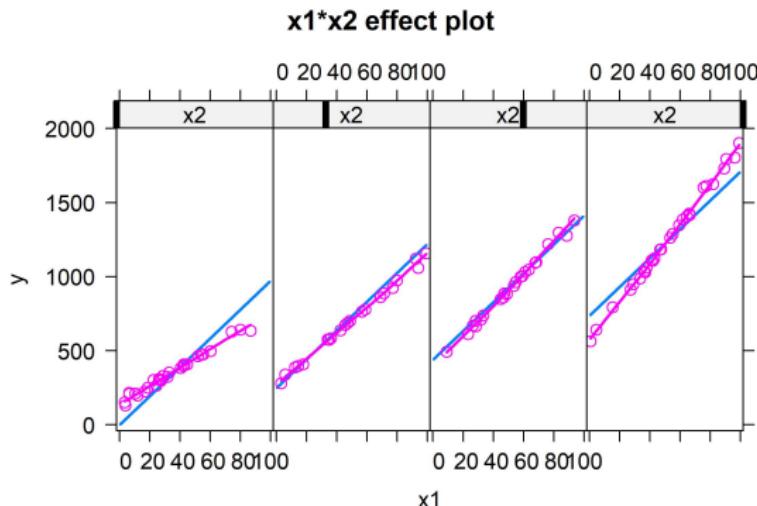
Ostanki modela brez interakcije



Predpostavke niso izpolnjene

Primer generiranih podatkov z **interakcijo** dveh napovednih spremenljivk

Graf parcialnih ostankov za model brez interakcije, y v odvisnosti od x_1 pri izbranih intervalih vrednostih x_2

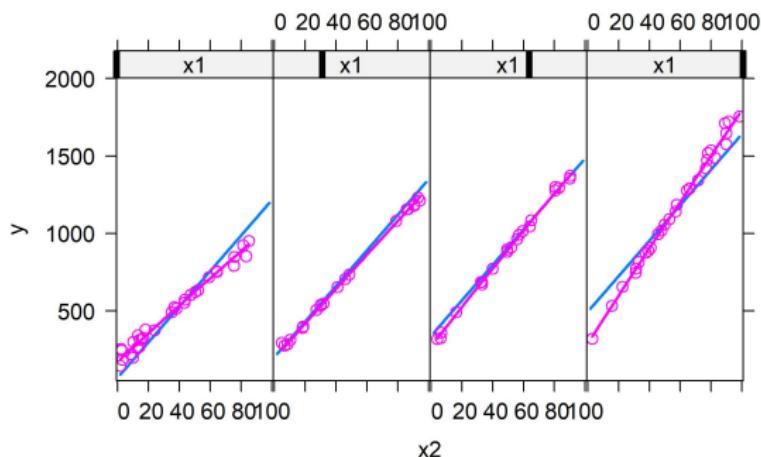


Predpostavke niso izpolnjene

Primer generiranih podatkov z **interakcijo** dveh napovednih spremenljivk, graf parcialnih ostankov za model brez interakcije

Graf parcialnih ostankov za model brez interakcije, y v odvisnosti od x_2 pri izbranih intervalih vrednostih x_1

x^2*x1 effect plot

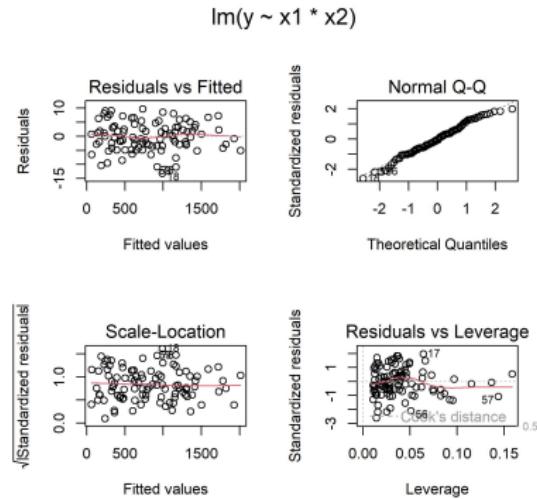
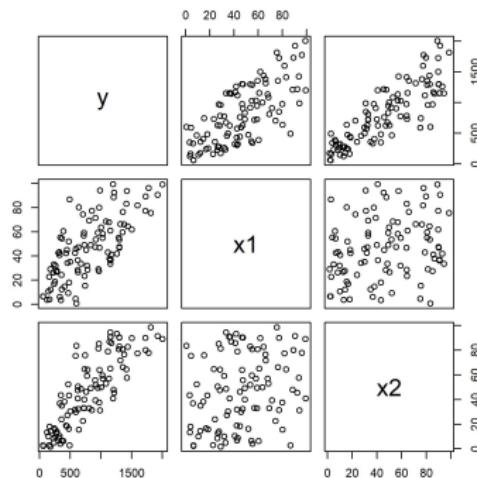


Predpostavke niso izpolnjene

Primer generiranih podatkov z **interakcijo** dveh napovednih spremenljivk, model vključuje interakcijski člen

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

Ostanki modela z interakcijo

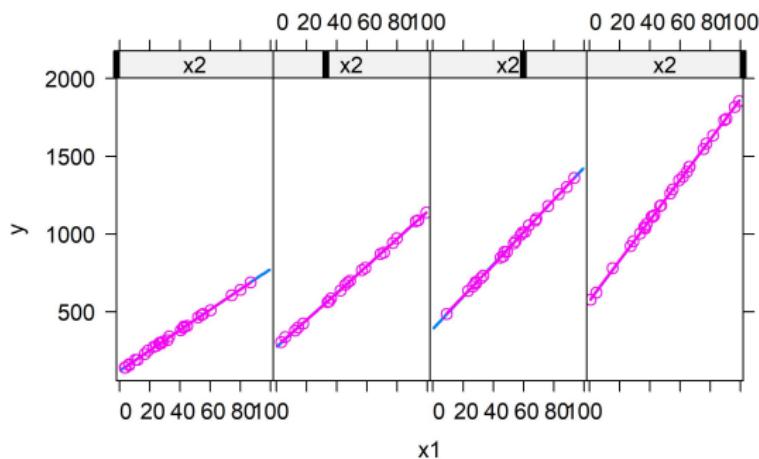


Predpostavke niso izpolnjene

Primer generiranih podatkov z **interakcijo** dveh napovednih spremenljivk, graf parcialnih ostankov za model z interakcijskim členom

Graf parcialnih ostankov za model brez interakcije, y v odvisnosti od x_1 pri izbranih intervalih vrednostih x_2

x1*x2 effect plot

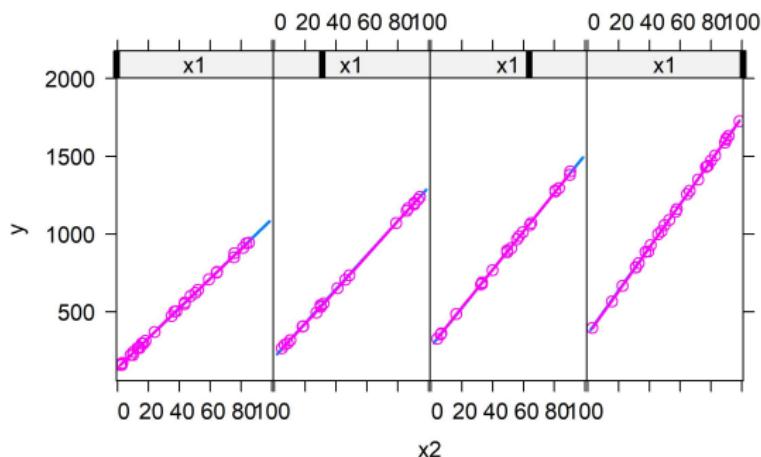


Predpostavke niso izpolnjene

Primer generiranih podatkov z **interakcijo** dveh napovednih spremenljivk, graf parcialnih ostankov za model z interakcijskim členom

Graf parcialnih ostankov za model brez interakcije, y v odvisnosti od x_2 pri izbranih intervalih vrednostih x_1

x^2*x1 effect plot



Predpostavke niso izpolnjene

Interakcija dveh številskih napovednih spremenljivk

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

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

in v točki $(x_{01}, x_{02} + 1)$

$$\begin{aligned} E(y|x_{01}, x_{02} + 1) &= \beta_0 + \beta_1 x_{01} + \beta_2(x_{02} + 1) + \beta_3 x_{01}(x_{02} + 1) \\ &= \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_2 + \beta_3 x_{01} x_{02} + \beta_3 x_{01}. \end{aligned}$$

Razlika:

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

Primer: postaje

Predpostavke niso izpolnjene

Trasformacije, ki zmanjšajo vplivnost točk

Trasformacije, ki zmanjšajo vplivnost točk

Večja vplivnost točk je lahko povzročena zaradi splošnega kršenja predpostavk linearnega modela:

- nekonstantna varianca
- nelinearne zveze med odzivno spremenljivko in regresorji
- nenavadne vrednosti regresorja

S transformacijo odzivne spremenljivke ali regresorjev lahko dosežemo **zmanjšanje** ali pa tudi **povečanje** vplivnosti posameznih točk v modelu.

Primer: mammals

Predpostavke niso izpolnjene

WLS

Metoda tehtanih najmanjših kvadratov (WLS, *Weighted Least Squares*)

Predpostavimo:

- napake ε_i so neodvisne in normalno porazdeljene
- varianca napak ni konstantna:
 - $Var(\varepsilon_i) = \sigma_i^2, i = 1, \dots, n$
 - $Var(\varepsilon_i) = \sigma^2 w_i, i = 1, \dots, n, w_i$ so uteži

Uteži w_i so znane in pozitivne, zapišemo jih v diagonalno matriko \mathbf{W} dimenzije $(n \times n)$.

Predpostavke niso izpolnjene

WLS

Ocene parametrov modela po metodi WLS dobimo z minimiranjem izraza

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

Večja utež W_{ii} pomeni, da ima i -ti podatek večji vpliv na oceno parametrov modela.

Predpostavke niso izpolnjene

WLS kot osnovna oblika GLS

Povezava med utežmi in varianco napak v normalnem linearinem modelu, ko ta ni konstantna:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{V})$$

V je diagonalna **variančna matrika napak** dimenzije $n \times n$.
Vrednosti na diagonali so različne.

Predpostavke niso izpolnjene

WLS kot osnovna oblika GLS

Za oceno parametrov modela in komponent variančne matrike napak uporabimo **metodo največjega verjetja**. Metodo imenujemo tudi **posplošena metoda najmanjših kvadratov** (GLS, *Generalized Least Square*)

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

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

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

Predpostavke niso izpolnjene

WLS kot osnovna oblika GLS

Če primerjamo izraz, ki smo ga dobili po WLS in izraz pri GLS , 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.

Rešitev je

$$\mathbf{b} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}$$

Predpostavke niso izpolnjene

WLS povzetek

- 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.
- Tudi za WLS ocene parametrov modela velja, da so najboljše linearne nepristranske cenilke za β (*BLUE, Best Linear Unbiased Estimator*), njihova varianca je najmanjša (Gauss-Markov izrek).

Primer: andy

Opisne in številske napovedne spremenljivke

Opisna napovedna spremenljivka z dvema vrednostma

Opisna napovedna spremenljivka ima dve vrednosti - naredimo eno umetno spremenljivko

$$w_{1i} = \begin{cases} 0, & x_i = a_1 \\ 1, & x_i = a_2 \end{cases} \quad \text{referenčna vrednost/skupina}$$

Model

$$y_i = \beta_0 + \beta_1 w_{1i} + \varepsilon_i$$

Pričakovana vrednost y_i :

$$\mathbb{E}(y_i) = \begin{cases} \beta_0, & x_i = a_1 \\ \beta_0 + \beta_1, & x_i = a_2 \end{cases}$$

Opisne in številske napovedne spremenljivke

Opisna napovedna spremenljivka z dvema vrednostma

Pomen parametrov:

- β_0 je povprečje y za referenčno vrednost $x = a_1$, μ_{a_1}
- β_1 je razlika $\mu_{a_2} - \mu_{a_1}$

Testiramo ničelni domnevi

$$H_0 : \beta_0 = \mu_{a_1} = \beta$$

$$H_0 : \beta_1 = \mu_{a_2} - \mu_{a_1} = \delta$$

Primer: zveza med FEV in Smoke

Opisne in številske napovedne spremenljivke

t-test za dva neodvisna vzorca, ponovitev

t-test za primerjavo dveh povprečij (*t-test za dva neodvisna vzorca*)

Predpostavke:

- imamo dva neodvisna vzorca
- analiziramo slučajno spremenljivko y , ki je v prvi populaciji porazdeljena $N(\mu_1, \sigma^2)$, v drugi populaciji pa $N(\mu_2, \sigma^2)$
- varianci obeh normalnih porazdelitev sta enaki

Zanima nas, ali sta povprečni vrednosti spremenljivke y v obeh populacijah enaki.

Opisne in številske napovedne spremenljivke

t-test za dva neodvisna vzorca

Testiramo ničelno domnevo:

$$H_0 : \mu_1 = \mu_2 \quad \text{ali} \quad \delta = \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 \neq \mu_2 \quad \text{ali} \quad \delta = \mu_1 - \mu_2 \neq 0$$

Testna statistika

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{s_{sk}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2}$$

Primer: vpliv kajenja Smoke na povprečno FEV

Opisne in številske napovedne spremenljivke

Opisna napovedna spremenljivka z I vrednostmi

Opisna napovedna spremenljivka x ima I vrednosti (a_1, a_2, \dots, a_I) .

Taka spremenljivka podatke deli v I skupin.

V model vključimo $I - 1$ regresorjev z vrednostmi 0 in 1.

$$w_1 = \begin{cases} 0, & x_i = a_1 \\ 1, & x_i = a_2 \\ 0, & x_i = a_3 \\ \dots \\ 0, & x_i = a_I \end{cases} \quad w_{I-1} = \begin{cases} 0, & x_i = a_1 \\ 0, & x_i = a_2 \\ 0, & x_i = a_3 \\ \dots \\ 1, & x_i = a_I \end{cases}$$

Opisne in številske napovedne spremenljivke

Opisna napovedna spremenljivka z I vrednostmi

Model

$$y_i = \beta_0 + \beta_1 w_{1i} + \beta_2 w_{2i} + \dots + \beta_{I-1} w_{Ii} + \varepsilon_i,$$

Pomen parametrov:

- $\beta_0 = \mu_{a_1}$ povprečje odzivne spremenljivke pri referenčni vrednosti opisne napovedne spremenljivke a_1
- razlike med povprečji j -te skupine in referenčne skupine:
$$\beta_j = \mu_{a_j} - \mu_{a_1}, j = 1, \dots, I-1$$

$$\mathbb{E}(y_i) = \begin{cases} \beta_0, & x_i = a_1 \\ \beta_0 + \beta_1, & x_i = a_2 \\ \dots \\ \beta_0 + \beta_{I-1}, & x_i = a_I. \end{cases}$$

Primer: FEV, spremenljivka Gender.Smoke

Opisne in številske napovedne spremenljivke

Dve opisni in ena številska napovedna spremenljivka

Varianta 1: Model brez interakcije opisnih in številske spremenljivke

Kakšna je zveza med FEV, Ht, Gender in Smoke?

Geometrijsko: štiri vzporedne premice (presečišča so različna, nakloni so enaki).

$$y_i = \beta_0 + \beta_1 \text{GenderM}_i + \beta_2 \text{SmokeDa}_i + \beta_3 \text{Ht}_i + \varepsilon_i$$

Opisne in številske napovedne spremenljivke

Dve opisni in ena številska napovedna spremenljivka

$$E(y_i) = \begin{cases} \beta_0 + \beta_3 \text{Ht}_i, & \text{Gender} = Z, \text{ Smoke} = Ne \\ (\beta_0 + \beta_1) + \beta_3 \text{Ht}_i, & \text{Gender} = M, \text{ Smoke} = Ne \\ (\beta_0 + \beta_2) + \beta_3 \text{Ht}_i, & \text{Gender} = Z, \text{ Smoke} = Da \\ (\beta_0 + \beta_1 + \beta_2) + \beta_3 \text{Ht}_i, & \text{Gender} = M, \text{ Smoke} = Da \end{cases}$$

Pomen parametrov:

- β_0 predstavlja povprečni FEV žensk nekadilk pri $\text{Ht}=0$
- β_1 je razlika povprečja FEV za moške in povprečja FEV za ženske nekadilce pri konstantni Ht na $[\text{Ht}_{min}, \text{Ht}_{max}]$
- β_2 je razlika povprečja FEV za kadilke/kadilce in povprečja FEV za nekadilke/nekadilce pri konstantni Ht na $[\text{Ht}_{min}, \text{Ht}_{max}]$
- β_3 je naklon vzporednih premic

Opisne in številske napovedne spremenljivke

Dve opisni in ena številska napovedna spremenljivka

Varianta 2: Model z interakcijo dveh opisnih in številske spremenljivke

V model vključimo Ht, Smoke in Gender ter **interakcijske člene**:

- **tri dvojne interakcije:** Gender : Smoke, Gender : Ht in Smoke : Ht
- **ena trojna interakcija:** Gender : Smoke : Ht

Predpostavimo, da je zveza med FEV in Ht različna pri kadilcih in nekadilcih, ta razlika je različna pri moških in pri ženskah.

Geometrijsko: **štiri različne premice** (presečišča so različna, nakloni so različni).

Vaja 1: Začetna analiza podatkov, enostavni linearni regresijski model, simulacije

1. Primer: Odstotek telesne mašcobe

R paketi, ki jih bomo uporabili na vajah:

```
library(vtable) # summary table (sumtable)
library(reshape2) # reshape data sets for ggplot (melt)
library(ggplot2) # nice plots (ggplot)
library(corrplot) # correlation plots (corrplot)
library(knitr) # for markdown
library(kableExtra) # creates nice latex tables (kable, kable_styling)
library(car) # regression diagnostics
```

Začetna analiza podatkov je pomemben del vsake obdelave podatkov (vir: <https://muse.jhu.edu/pub/56/article/793379/pdf>). Sestavlja jo naslednji koraki:

1. določitev metapodatkov (podatki o podatkih);
2. čiščenje podatkov (odpravljanje napak v podatkih);
3. pregled podatkov (razumevanje lastnosti podatkov);
4. poročanje začetne analize podatkov vsem sodelavcem, vpletenih v analizo;
5. izpopolnjevanje in posodabljanje načrta analize, ki vključuje ugotovitve na podlagi začetne analize podatkov;
6. poročanje začetne analize podatkov (vsebovati mora vse korake, ki vplivajo na interpretacijo rezultatov).

Cilj naše analize bo razviti model za napovedovanje odstotka telesne mašcobe na podlagi spremenljivk, katerih vrednosti lahko dobimo le z uporabo tehtnice in merilnega traku (vir: Roger W. Johnson (1996), "Fitting Percentage of Body Fat to Simple Body Measurements", Journal of Statistics Education, <http://jse.amstat.org/v4n1/datasets.johnson.html>). V podatkih sta dve oceni odstotka telesne mašcobe, dobljeni po Brozokovi in Sirijevi enačbi. V vaji bo slednja spremenljivka naša odzivna spremenljivka.

```
bodyfat <- read.table(url("https://jse.amstat.org/datasets/fat.dat.txt"))
```

```
head(bodyfat)
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
1	1	12.6	12.3	1.0708	23	154.25	67.75	23.7	134.9	36.2	93.1	85.2	94.5	59.0
2	2	6.9	6.1	1.0853	22	173.25	72.25	23.4	161.3	38.5	93.6	83.0	98.7	58.7
3	3	24.6	25.3	1.0414	22	154.00	66.25	24.7	116.0	34.0	95.8	87.9	99.2	59.6
4	4	10.9	10.4	1.0751	26	184.75	72.25	24.9	164.7	37.4	101.8	86.4	101.2	60.1
5	5	27.8	28.7	1.0340	24	184.25	71.25	25.6	133.1	34.4	97.3	100.0	101.9	63.2
6	6	20.6	20.9	1.0502	24	210.25	74.75	26.5	167.0	39.0	104.5	94.4	107.8	66.0
	V15	V16	V17	V18	V19									
1	37.3	21.9	32.0	27.4	17.1									
2	37.3	23.4	30.5	28.9	18.2									
3	38.9	24.0	28.8	25.2	16.6									
4	37.3	22.8	32.4	29.4	18.2									
5	42.2	24.0	32.2	27.7	17.7									
6	42.0	25.6	35.7	30.6	18.8									

Metapodatki

Ta niz podatkov za 252 moških vsebuje informacije o:

Tabela 1: Metapodatki podatkovnega okvira **bodyfat**.

Ime	Pomen	Enote	Merska lestvica	Class	NAs
case	ID			integer	0
brozek	odstotek telesne maščobe (Brozek)	%	številkska	numeric	0
siri	odstotek telesne maščobe (Siri)	%	številkska	numeric	0
density	Gostota, določena s tehtanjem pod vodo	gm/cm^3	številkska	numeric	0
age	Starost	years	številkska	numeric	0
weight	Masa	lbs	številkska	numeric	0
height	Višina	inches	številkska	numeric	0
BMI	Indeks telesne mase	kg/m^2	številkska	numeric	0
fatfreeweight	Masa brez maščobe [weight/(1-brozek/100)]	lbs	številkska	numeric	0
neck	Obseg vratu	cm	številkska	numeric	0
chest	Obseg prsnega koša	cm	številkska	numeric	0
abdomen	Obseg abdomna	cm	številkska	numeric	0
hip	Obseg bokov	cm	številkska	numeric	0
thigh	Obseg stegna	cm	številkska	numeric	0
knee	Obseg kolena	cm	številkska	numeric	0
ankle	Obseg gležnja	cm	številkska	numeric	0
biceps	Obseg bicepsa	cm	številkska	numeric	0
forearm	Obseg podlakti	cm	številkska	numeric	0
wrist	Obseg zapestja	cm	številkska	numeric	0

Tabela 1 prikazuje metapodatke podatkovnega okvira **bodyfat**. Metapodatki so podatki o podatkih oz. informacije o spremenljivkah, na primer: oznaka in pomen posamezne spremenljivke, merska lestvica, enote, intervali, na katerih se vrednosti posameznih spremenljivk lahko nahajajo, informacije o manjkajočih podatkih. Metapodatki pa lahko vključujejo tudi informacije o tem, kako so bili podatki pridobljeni (npr. viri podatkov, metode zbiranja podatkov, kako je definirana ciljna populacija, merila za vključitev in izključitev posameznih enot, metode vzorčenja, čas zbiranja podatkov, ipd.).

Iz metapodatkov lahko vidimo, da sta spremenljivki BMI in **fatfreeweight** izpeljani na podlagi drugih spremenljivk v podatkovnem okviru. Ti dve spremenljivki bomo izločili iz podatkovnega okvira, saj nas bo zanimalo napovedovanje odstotka telesne maščobe na podlagi ‘osnovnih’ spremenljivk.

```
bodyfat <- bodyfat[, -c(8:9)]
```

Spremenljivkam dopišemo imena

```
colnames(bodyfat) <- c("case", "brozek", "siri", "density", "age", "weight",
                      "height", "neck", "chest", "abdomen", "hip", "thigh",
                      "knee", "ankle", "biceps", "forearm", "wrist")
```

in sprememimo enote za spremenljivki masa in višina:

```
## iz lb v kg
bodyfat$weight <- bodyfat$weight * 0.454
## iz in v cm
bodyfat$height <- bodyfat$height * 2.54
```

Čiščenje podatkov

Čiščenje podatkov je sistematičen poskus iskanja napak v podatkih in, če je le mogoče, njihovega odpravljanja. Pogosti primeri so: napačno kodiranje opisnih spremenljivk, nemogoče vrednosti, datumi, izven časovnega okvira študije, manjkajoče vrednosti, osamelci, nemogoče kombinacije vrednosti dveh spremenljivk, podvojene vrstice, ipd.

V naslednjem koraku podatke očistimo vrednosti, ki niso verjetne:

```
summary(bodyfat)
```

case	brozek	siri	density	
Min. : 1.00	Min. : 0.00	Min. : 0.00	Min. : 0.995	
1st Qu.: 63.75	1st Qu.: 12.80	1st Qu.: 12.47	1st Qu.: 1.041	
Median :126.50	Median :19.00	Median :19.20	Median :1.055	
Mean :126.50	Mean :18.94	Mean :19.15	Mean :1.056	
3rd Qu.:189.25	3rd Qu.:24.60	3rd Qu.:25.30	3rd Qu.:1.070	
Max. :252.00	Max. :45.10	Max. :47.50	Max. :1.109	
age	weight	height	neck	
Min. :22.00	Min. : 53.80	Min. : 74.93	Min. :31.10	
1st Qu.:35.75	1st Qu.: 72.19	1st Qu.:173.35	1st Qu.:36.40	
Median :43.00	Median : 80.13	Median :177.80	Median :38.00	
Mean :44.88	Mean : 81.23	Mean :178.18	Mean :37.99	
3rd Qu.:54.00	3rd Qu.: 89.44	3rd Qu.:183.51	3rd Qu.:39.42	
Max. :81.00	Max. :164.87	Max. :197.49	Max. :51.20	
chest	abdomen	hip	thigh	
Min. : 79.30	Min. : 69.40	Min. : 85.0	Min. :47.20	
1st Qu.: 94.35	1st Qu.: 84.58	1st Qu.: 95.5	1st Qu.:56.00	
Median : 99.65	Median : 90.95	Median : 99.3	Median :59.00	
Mean :100.82	Mean : 92.56	Mean : 99.9	Mean :59.41	
3rd Qu.:105.38	3rd Qu.: 99.33	3rd Qu.:103.5	3rd Qu.:62.35	
Max. :136.20	Max. :148.10	Max. :147.7	Max. :87.30	
knee	ankle	biceps	forearm	wrist
Min. :33.00	Min. : 19.1	Min. :24.80	Min. :21.00	Min. :15.80
1st Qu.:36.98	1st Qu.: 22.0	1st Qu.:30.20	1st Qu.:27.30	1st Qu.:17.60
Median :38.50	Median : 22.8	Median :32.05	Median :28.70	Median :18.30
Mean :38.59	Mean : 23.1	Mean :32.27	Mean :28.66	Mean :18.23
3rd Qu.:39.92	3rd Qu.: 24.0	3rd Qu.:34.33	3rd Qu.:30.00	3rd Qu.:18.80
Max. :49.10	Max. : 33.9	Max. :45.00	Max. :34.90	Max. :21.40

#najmanjša oseba je visoka 75 cm

```
which.min(bodyfat$height)
```

```
[1] 42
```

```
bodyfat$weight[which.min(bodyfat$height)] #in tehta 93 kg
```

```
[1] 93.07
```

```
bodyfat$height[bodyfat$case==42] <- 176.53 #glej https://jse.amstat.org/datasets/fat.txt
```

Pregledovanje podatkov

Pregledovanje podatkov nam omogoča razumeti lastnosti podatkov, ki bi lahko vplivale na nadaljnjo analizo in interpretacijo rezultatov. Vključuje korake, ki preverjajo, ali podatki izpolnjujejo določene lastnosti oz. predpostavke, ki so potrebne, da določena analiza da zadovoljive rezultate, vendar izključuje kakršno koli testiranje hipotez. Pregledovanje podatkov vključuje preučevanje univariatnih in multivariatnih porazdelitev spremenljivk, izračun opisnih statistik ter identifikacija manjkajočih vrednosti pri posameznih enotah in spremenljivkah.

V kontekstu linearnih modelov lahko z univariatnimi porazdelitvami spremenljivk odkrijemo prisotnost osamelcev, ki imajo lahko močan vpliv na rezultate analize, ter spremenljivke, ki so asimetrično porazdeljene. Čeprav linerani model ne predpostavlja ničesar o porazdelitvi odzivne spremenljivke (neodvisno od regresorjev) in porazdelitvi regresorjev, je verjetno, da bodo spremenljivke, ki so močno asimetrične, vplivale na porazdelitev

Tabela 2: Opisne statistike za spremenljivke v podatkovnem okviru **bodyfat**.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 50	Pctl. 75	Max
siri	252	19	8	0	12	19	25	48
age	252	45	13	22	36	43	54	81
weight	252	81	13	54	72	80	89	165
height	252	179	7	163	173	178	184	197
neck	252	38	2	31	36	38	39	51
chest	252	101	8	79	94	100	105	136
abdomen	252	93	11	69	85	91	99	148
hip	252	100	7	85	96	99	104	148
thigh	252	59	5	47	56	59	62	87
knee	252	39	2	33	37	38	40	49
ankle	252	23	2	19	22	23	24	34
biceps	252	32	3	25	30	32	34	45
forearm	252	29	2	21	27	29	30	35
wrist	252	18	0.9	16	18	18	19	21

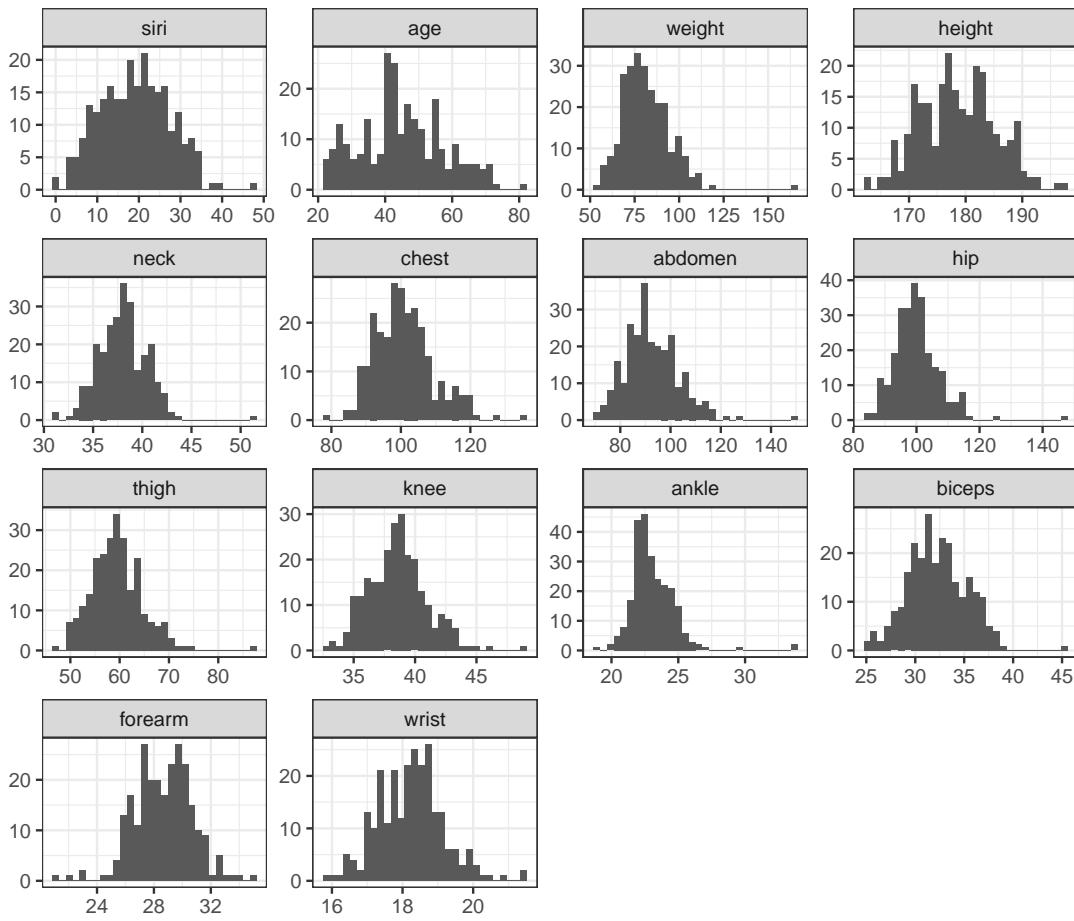
ostankov, za katere linearni model predpostavlja, da so normalno porazdeljeni. V splošnem velja, če je spremenljivka asimetrično porazdeljena, je potrebno analizo temu prilagoditi in raje uporabiti metodo, ki ne temelji na predpostavki o normalni porazdelitvi, lahko pa spremenljivko tudi lineariziramo. Pri asimetrično porazdeljenih spremenljivkah se pogosto zgodi tudi to, da je povezanost z drugimi spremenljivkami lahko le posledica majhnega števila enot, zato je treba to povezanost obravnavati previdno. Če so pri spremenljivki prisotne manjkajoče vrednosti, zaključkov na podlagi analize, omejene na enote, pri katerih manjkajočih vrednosti ni, morda ne bomo mogli posplošiti na celotno populacijo. Preučevanje bivariatnih povezanosti (preko korelacij oz. grafično) nam lahko pomaga pri odkrivanju nelinearnosti, interakcij, osamelcev in pri identificiranju multikolinearnosti.

Poglejmo si univariatne porazdelitve spremenljivk in korelacije med pari napovednih spremenljivk.

```
# napovedne spremenljivke
pred <- c("age", "weight", "height", "neck", "chest", "abdomen", "hip",
"thigh", "knee", "ankle", "biceps", "forearm", "wrist")

sumtable(bodyfat[,c("siri", pred)],
  add.median = T,
  digits = 1,
  title = "Opisne statistike za spremenljivke v podatkovnem okviru
\\texttt{bodyfat}.",
  out = "kable"
)

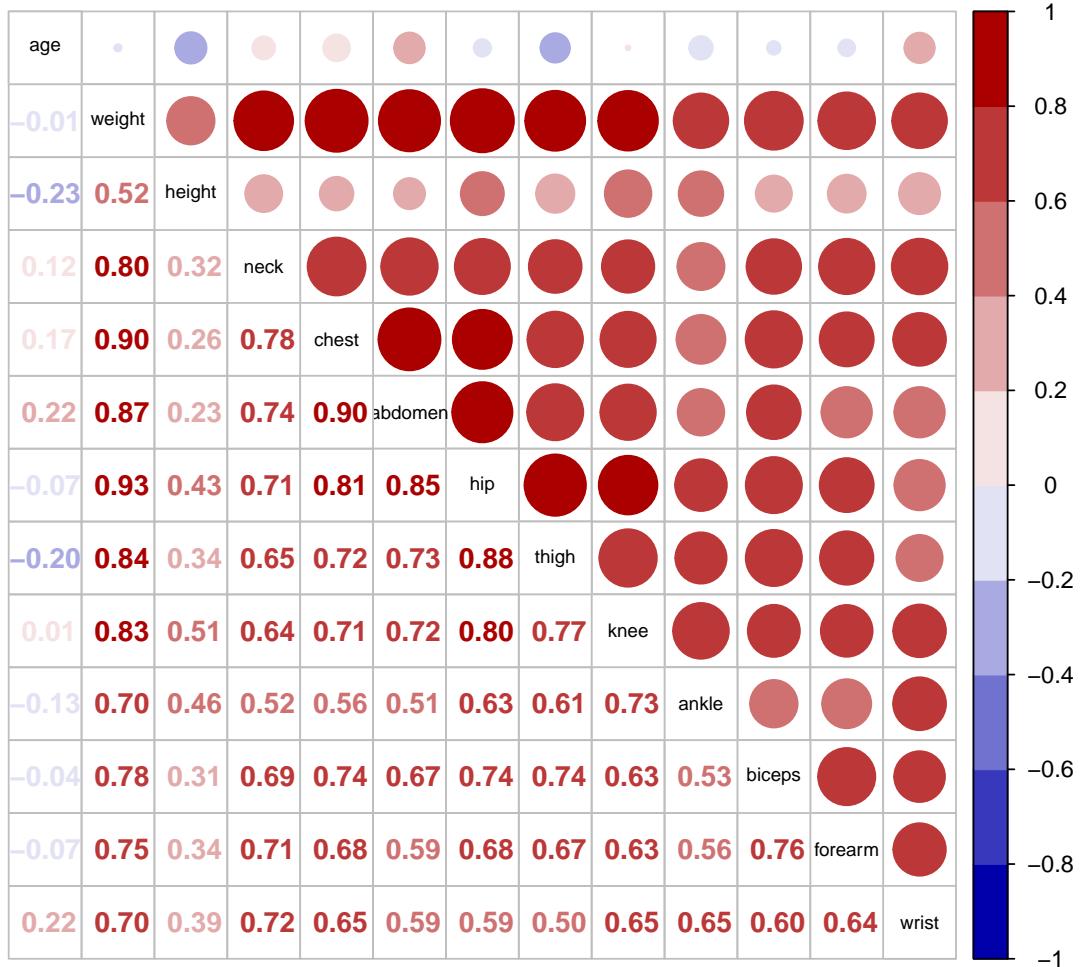
bodyfat_long <- melt(bodyfat[, c("case", "siri", pred)], id.vars = "case")
qplot(value, data = bodyfat_long) +
  facet_wrap(~variable, scales = "free") +
  theme_bw() +
  xlab("")
```



Slika 1: Univariatne porazdelitve spremenljivk v podatkovnem okviru **bodyfat**.

Porazdelitev odzivne spremenljivke **siri** je dokaj blizu normalni porazdelitvi, pri nekaterih napovednih spremenljivkah imamo posamezne osamelce, sicer pa so vrednosti precej normalno porazdeljene po zalogi vrednosti.

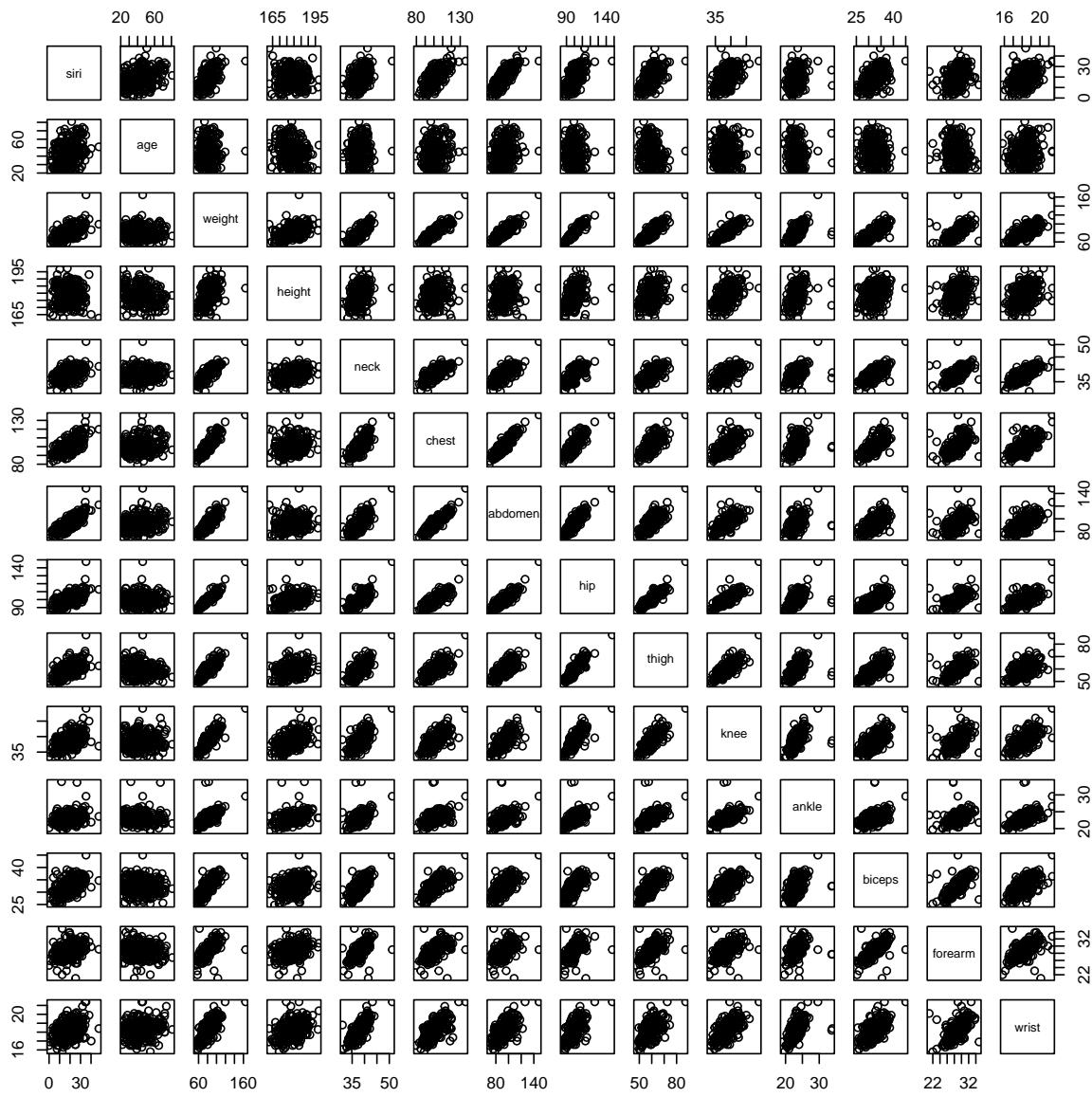
```
colors10 <- colorRampPalette(c("#0000aa", "white", "#aa0000"))(10)
corrplot.mixed(cor(bodyfat[, pred], method="spearman"),
               lower.col = colors10, upper.col = colors10,
               tl.col="black", tl.cex = 0.7)
```



Slika 2: Spearmanovi koeficienti korelacji med napovednimi spremenljivkami v podatkovnem okviru `bodyfat`.

Slika 2 kaže, da so napovedne spremenljivke dokaj tesno pozitivno povezane med seboj, edina napovedna spremenljivka, ki ne kaže povezanosti z ostalimi je `age`. Močne korelacijske kažejo na to, da bi v modelu znali imeti težave s kolinearnostjo.

```
pairs(bodyfat[, c("siri", pred)])
```



Slika 3: Matrika razsevnih grafikonov v podatkovnem okviru **bodyfat**.

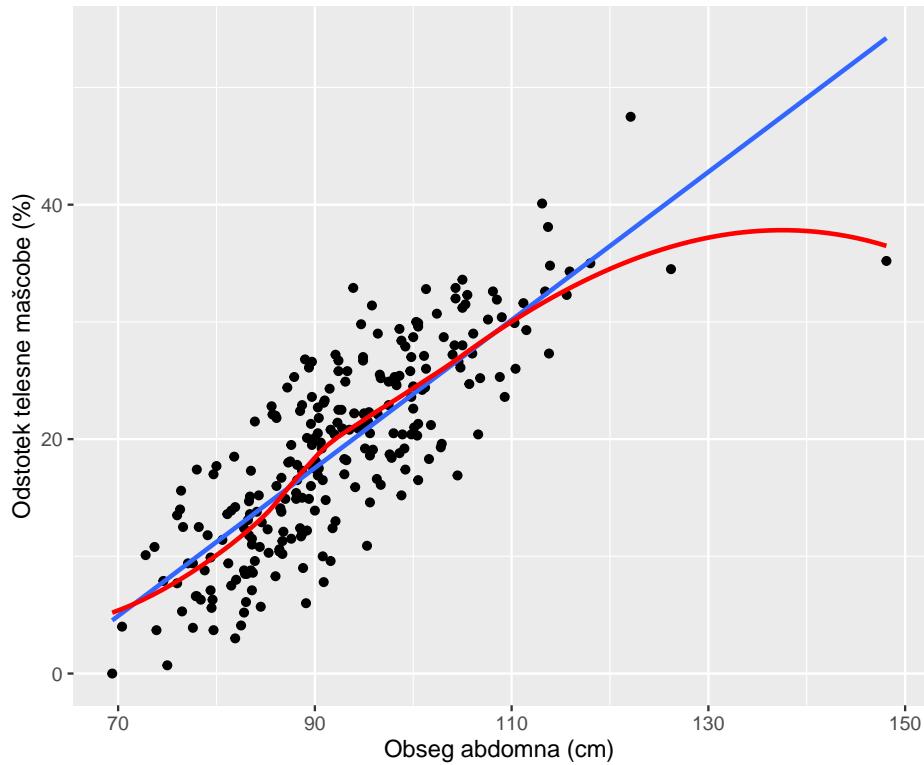
Matrika razsevnih grafikonov na Sliki 3 nam grafično prikaže vse bivariatne zveze med posameznimi pari spremenljivk v podatkovnem okviru **bodyfat**. Vidimo, da so zveze lepo linearne, povezanost med nekaterimi pari spremenljivk pa je zelo močna, kar bi posledično lahko pomenilo težave s kolinearnostjo. Ponekod so opazni osamelci.

Analiza

V današnji vaji se bomo osredotočili na zvezo med spremenljivkama **siri** in **abdomen**. Zanima nas, ali bi **siri** lahko napovedali na podlagi spremenljivke **abdomen**. Če med spremenljivkama obstaja linearna povezanost, potem tako zvezo lahko modeliramo z linearnim modelom:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \quad i = 1, \dots, n.$$

```
ggplot(data=bodyfat, aes(x=abdomen, y=siri)) + geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  geom_smooth(col="red", se=FALSE) +
  xlab("Obseg abdomna (cm)") +
  ylab("Odstotek telesne maščobe (%)")
```



Slika 4: Razsevni grafikon za `siri` in `abdomen` z dodano premico in gladilnikom.

Cilj je najti oceni za parametra β_0 in β_1 , tako da se bo regresijska premica dobro prilegala podatkom:

$$\hat{y}_i = b_0 + b_1 \cdot x_i, \quad i = 1, \dots, n.$$

```
m1 <- lm(siri ~ abdomen, data = bodyfat)
summary(m1)
```

Call:
`lm(formula = siri ~ abdomen, data = bodyfat)`

Residuals:

Min	1Q	Median	3Q	Max
-19.0160	-3.7557	0.0554	3.4215	12.9007

Coefficients:

Estimate	Std. Error	t value	Pr(> t)
----------	------------	---------	----------

```
(Intercept) -39.28018    2.66034   -14.77   <2e-16 ***
abdomen      0.63130    0.02855    22.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 4.877 on 250 degrees of freedom
 Multiple R-squared: 0.6617, Adjusted R-squared: 0.6603
 F-statistic: 488.9 on 1 and 250 DF, p-value: < 2.2e-16

V povzetku modela se testirata dve ničelni hipotezi:

$H_0 : \beta_0 = 0$ proti $H_1 : \beta_0 \neq 0$

in

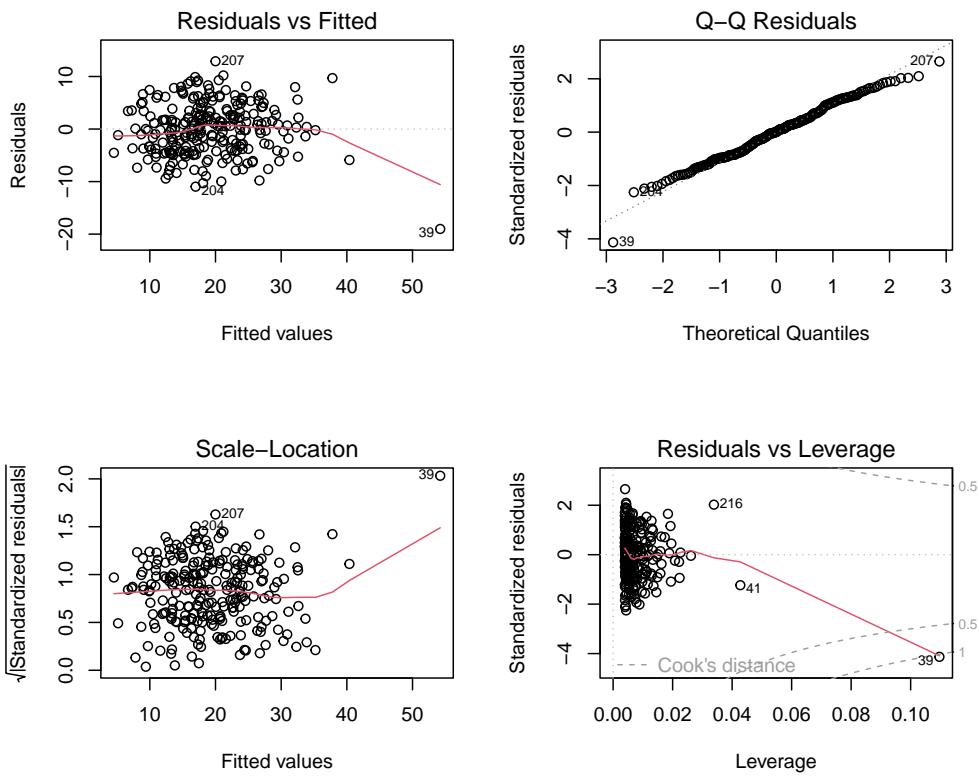
$H_0 : \beta_1 = 0$ proti $H_1 : \beta_1 \neq 0$

Statistično sklepanje in napovedi so veljavne, kadar so predpostavke normalnega linearnega modela izpolnjene, to je kadar:

1. obstaja linearna odvisnost odzivne spremenljivke od napovednih sprmenljivk,
2. imajo napake ε_i za vsako enoto skupno varianco σ^2 (homoskedastičnost),
3. je pričakovana vrednost napak 0,
4. so napake porazdeljene po normalni porazdelitvi,
5. so napake medsebojno neodvisne.

Osnovno diagnostiko modela naredimo na podlagi slik ostankov modela:

```
par(mfrow=c(2,2))
plot(m1)
```



Slika 5: Ostanki za model 1 $\text{siri} \sim \text{abdomen}$.

Na podlagi slik ostankov lahko preverimo predpostavke 1., 2., 3. in 4. Razložite, kako! Kaj pomeni predpostavka 5.?

Zdi se, da se linearni model dobro prilega podatkom, razen za enoto 39.

```
bodyfat[39, ]
```

```
case brozek siri density age weight height neck chest abdomen hip thigh
39 39 33.8 35.2 1.0202 46 164.8701 183.515 51.2 136.2 148.1 147.7 87.3
knee ankle biceps forearm wrist
39 49.1 29.6 45 29 21.4
```

Izkaže se, da večino osamelcev, ki smo jih identificirali na podlagi Slike 1, lahko pripišemo enoti 39.

To enoto bi imelo v nadaljevanju smisel izključiti iz analize in primerjati rezultate obeh modelov. Kasneje bomo v okviru posebnih točk v regresijski analizi videli, da je ta točka t. i. vplivna točka.

```
bodyfat_brez39 <- bodyfat[-which(bodyfat$case==39),]
m2 <- lm(siri ~ abdomen, data = bodyfat_brez39)
summary(m2)
```

```
Call:
lm(formula = siri ~ abdomen, data = bodyfat_brez39)
```

Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

```
-10.9133 -3.6469  0.1914   3.1737  12.7613

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -42.95774    2.71323 -15.83   <2e-16 ***
abdomen      0.67195    0.02921  23.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 4.717 on 249 degrees of freedom
Multiple R-squared: 0.6801, Adjusted R-squared: 0.6788
F-statistic: 529.3 on 1 and 249 DF, p-value: < 2.2e-16

Primerjava ocen obeh modelov:

```
compareCoefs(m1, m2)
```

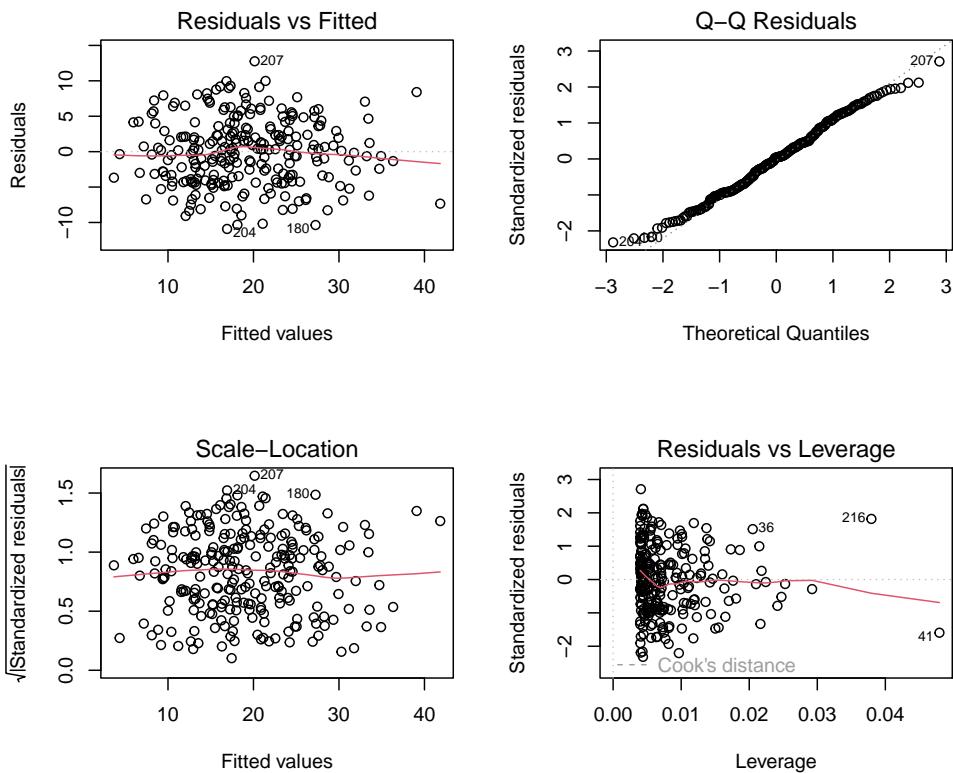
Calls:

```
1: lm(formula = siri ~ abdomen, data = bodyfat)
2: lm(formula = siri ~ abdomen, data = bodyfat_brez39)
```

	Model 1	Model 2
(Intercept)	-39.28	-42.96
SE	2.66	2.71
abdomen	0.6313	0.6720
SE	0.0286	0.0292

Poglejmo ostanke modela m2.

```
par(mfrow=c(2,2))
plot(m2)
```



Slika 6: Ostanki za model 2 $\text{siri} \sim \text{abdomen}$.

Oceni $b_0 = -42.96$ in $b_1 = 0.67$, dobljeni po metodi najmanjših kvadratov, sta nepristranski. Pripadajoči standardni napaki izražata mero natančnosti ocen, na podlagi katerih lahko izračunamo intervale zaupanja.

95% interval zaupanja za obe oceni:

```
confint(m2)
```

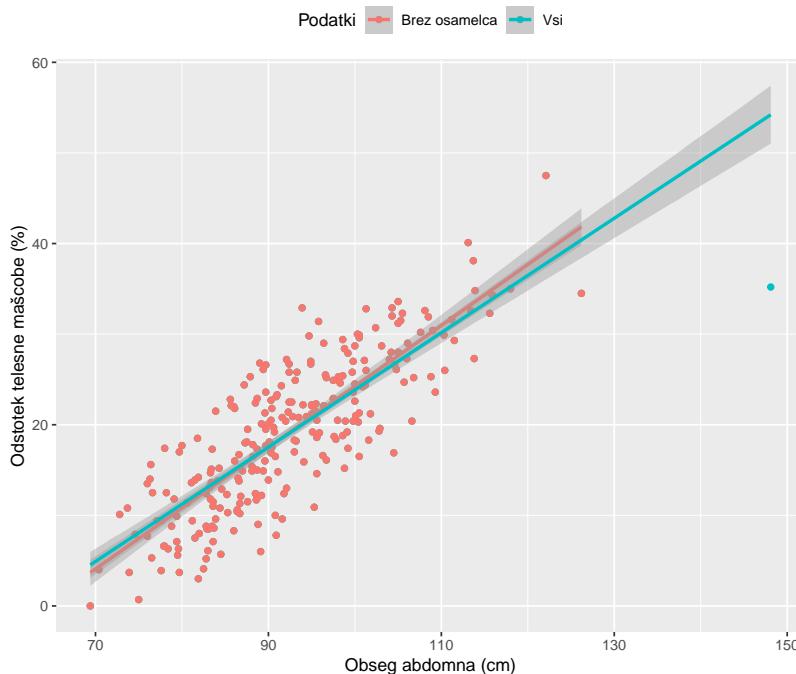
	2.5 %	97.5 %
(Intercept)	-48.3015390	-37.6139418
abdomen	0.6144288	0.7294781

Interpretacija modela: če se obseg abdomna poveča za 1 cm, se telesna maščoba v povprečju poveča za 0.67 %, 95 % IZ: (0.61, 0.73). Z modelom smo pojasnili 68 % variabilnosti odzivne spremenljivke.

```
bodyfat$Podatki <- "Vsi"
bodyfat_brez39$Podatki <- "Brez osamelca"

bodyfat_komb <- rbind(bodyfat, bodyfat_brez39)

ggplot(aes(x=abdomen, y=siri, color=Podatki), data=bodyfat_komb) +
  geom_point() +
  geom_smooth(method = "lm", se=TRUE) +
  xlab("Obseg abdomna (cm)") +
  ylab("Odstotek telesne maščobe (%)") +
  theme(legend.position = "top")
```



Slika 7: Odvisnost **siri** od **abdomen** za dana vzorca 252 oz. 251 moških in regresijski premici na podalgi modelov **m1** in **m2** s 95 % intervali zaupanja za povprečno napoved.

2. Primer: Čas teka Collina Jacksona

V datoteki *COLLIN.txt* so podatki za 21 tekov čez ovire na 110 m tekača Collina Jacksona: hitrost vetra = **windspeed** (m/s) in čas teka= **time** (s) (Vir: Daly et al., str. 525). Podatki so bili dobljeni v poskusu v zaprtem prostoru, hitrost vetra je bila izbrana za vsak tek posebej vnaprej. Negativne vrednosti hitrosti vetra pomenijo, da je veter pihal v prsi tekača. Kako hitrost vetra vpliva na čas teka čez ovire na 110 m?

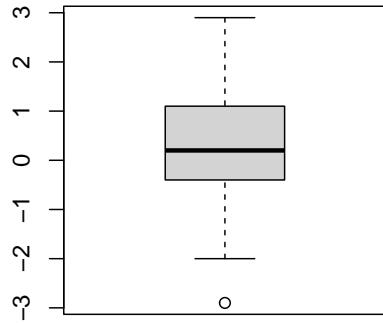
1. Grafično prikažite podatke.
2. Ocenite parametra linearnega regresijskega modela za odvisnost časa teka od hitrosti vetra.
3. Analizirajte ostanke modela na podlagi grafičnih prikazov.
4. Obrazložite cenilki parametrov modela in njuna intervala zaupanja.
5. Obrazložite koeficient determinacije.
6. Izračunajte povprečno in posamično napoved časa teka ter pripadajoče 95 % intervale zaupanja za naslednje hitrosti vetra: -1 m/s, 0 m/s, 1 m/s in 4 m/s. Ali so vse napovedi upravičene? Zakaj?

```
d <- read.table("COLLIN.txt", header = T)
summary(d)
```

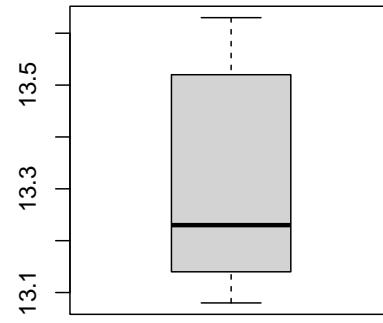
windspeed	time
Min. : -2.9000	Min. : 13.08
1st Qu.: -0.4000	1st Qu.: 13.14
Median : 0.2000	Median : 13.23
Mean : 0.2238	Mean : 13.30
3rd Qu.: 1.1000	3rd Qu.: 13.52
Max. : 2.9000	Max. : 13.63

```
#poglejmo najprej univariatne porazdelitve spremenljivk v podatkovnem okviru
par(mfrow=c(1,2))

boxplot(d$windspeed, main="", xlab="Hitrost vetra [m/s]")
boxplot(d$time, main="", xlab="Čas teka [s]")
```



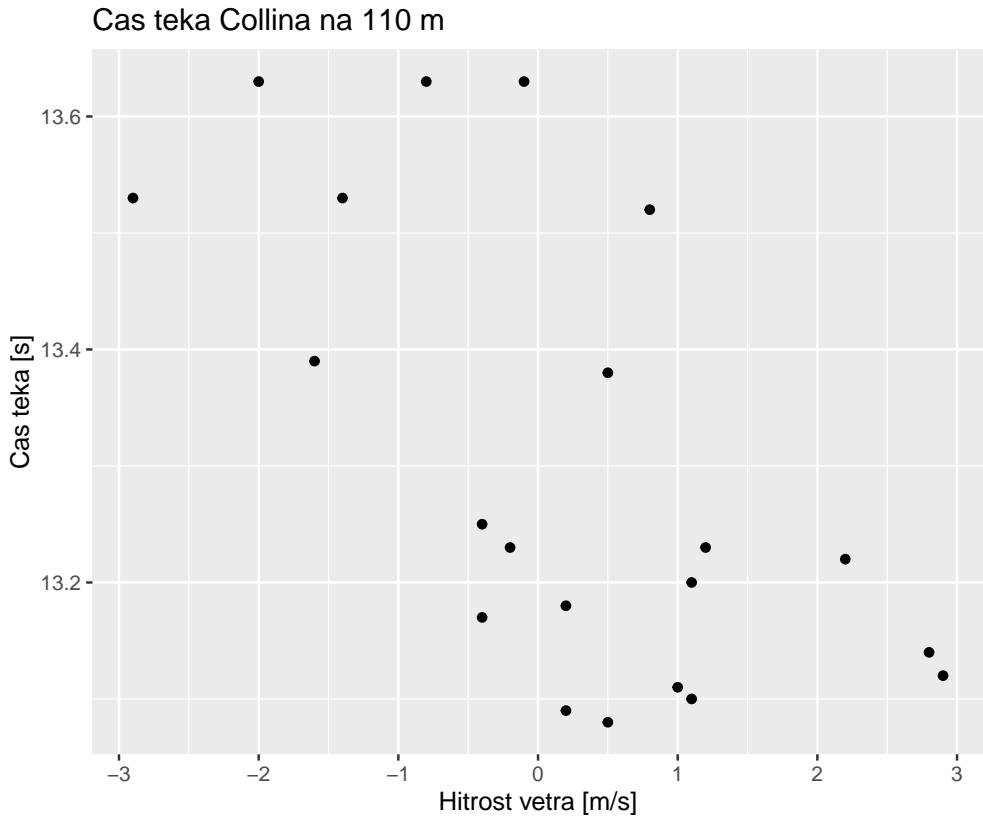
Hitrost vetra [m/s]



Čas teka [s]

Slika 8: Univariatne porazdelitve spremenljivk v podatkovnem okviru COLLIN.

```
#Ali obstaja linearна povezanost med spremenljivkama?
ggplot(data=d, aes(x=windspeed, y=time)) +
  geom_point() +
  xlab("Hitrost vetra [m/s]") +
  ylab("Čas teka [s]") +
  ggtitle("Čas teka Collina na 110 m")
```



Slika 9: Odvisnost časa teka od hitrosti vetra.

Enostavni linearni model za čas teka Collina v odvisnosti od hitrosti vetra:

```
m_collin <- lm(time ~ windspeed, data = d)
summary(m_collin)
```

Call:
`lm(formula = time ~ windspeed, data = d)`

Residuals:

Min	1Q	Median	3Q	Max
-0.21487	-0.12487	-0.02873	0.08976	0.29975

Coefficients:

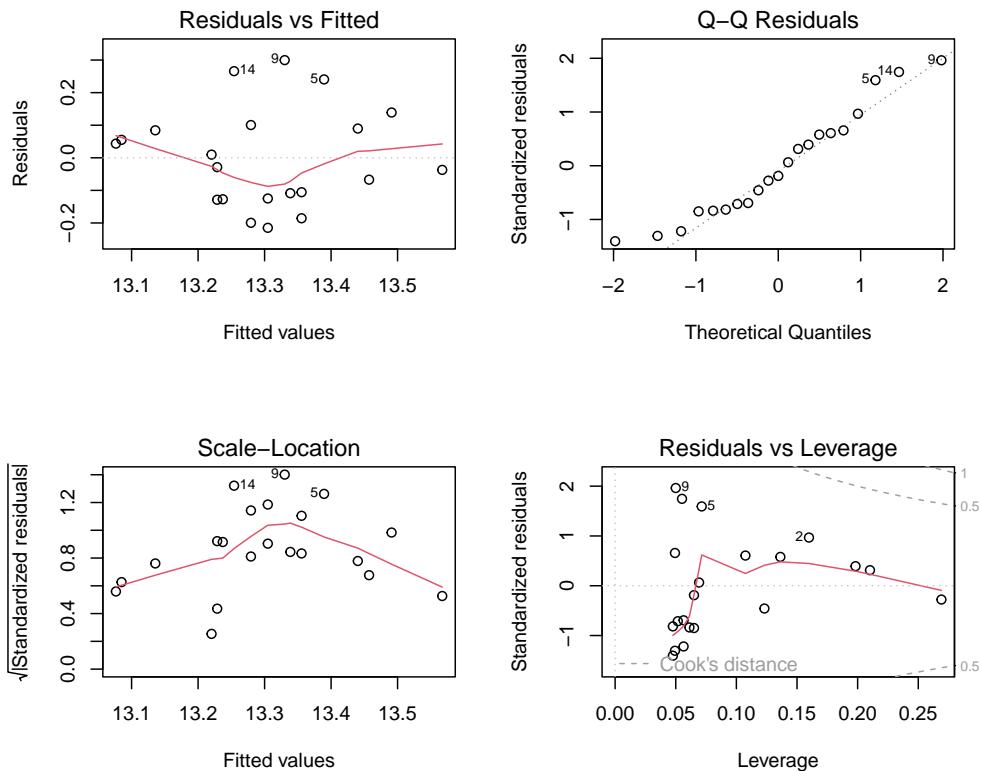
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.32179	0.03460	385.043	< 2e-16 ***
windspeed	-0.08460	0.02361	-3.584	0.00198 **

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	''	'	'	'

Residual standard error: 0.1567 on 19 degrees of freedom
Multiple R-squared: 0.4033, Adjusted R-squared: 0.3719
F-statistic: 12.84 on 1 and 19 DF, p-value: 0.00198

Slike ostankov:

```
par(mfrow=c(2,2))
plot(m_collin)
```



Slika 10: Ostanki za model `time ~ windspeed`.

Na prvi levi sličici Slike 10 so prikazani ostanki (Residuals) v odvisnosti od napovedanih vrednosti (Fitted values). Vidimo, da prihaja do odstopanj gladelnika (rdeča črta, ki jo nariše funkcija loess – local polynomial regression fitting) od vodoravne črte z vrednostjo ostanka 0. Ker je model narejen le na 21 enotah, so odstopanja normalna in še ne pomenijo kršitve predpostavk. Vsaka točka namreč gladelnik potegne v svojo smer, zato smo pri analizah majhnih vzorcev načeloma do odstopanj bolj tolerantni. Zgornja desna slika kaže porazdelitev standardiziranih ostankov v primerjavi z normalno porazdelitvijo (ravna črta v ozadju). Točke se dobro prilegajo ravni črtkani črti, kar pomeni, da lahko privzamemo normalno porazdelitev za standardizirane ostanke. Spodnja leva slika prikazuje koren absolutne vrednosti standardiziranih ostankov v odvisnosti od napovedanih vrednosti. Gladelnik, ki je vodoraven, pomeni, da med ostanki ni prisotne heteroskedastičnosti. Tudi tu lahko odstopanja pripisemo majhnemu vzorcu. Spodnja desna slika identificira vplivne točke, ki jih v tem primeru ni.

```
coef(m_collin)
```

```
(Intercept)  windspeed
13.32179201 -0.08460258
```

```
confint(m_collin)
```

	2.5 %	97.5 %
(Intercept)	13.2493772	13.39420677
windspeed	-0.1340109	-0.03519423

Interpretacija modela: v povzetku modela se testirata dve ničelni hipotezi. Prva testira, ali je čas teka Collina na 110 m v brezveterju enak 0 (ni smiselna). Domnevo zavrnemo: imamo 95 % zaupanje, da je čas teka Collina na 110 m v brezveterju nekje med 13.25 in 13.39 s. Druga ničelna domneva testira, ali je čas teka Collina na 110 m ovisen od hitrosti vetra (smiselna). Tudi to domnevo zavrnemo v prid alternativne: obstaja povezanost med časom teka in hitrotjo vetra. S 95 % zaupanjem lahko trdimo, če se hitrost vetra v hrbot poveča za 1 m/s, se čas teka v povprečju zmanjša med -0.13 in -0.04 s. Z modelom smo pojasnili 40 % variabilnosti odzivne spremenljivke.

```
casi = data.frame(windspeed = c(-1, 0, 1, 4))

povprecne_napovedi = data.frame(predict(m_collin, casi, interval = "confidence"))
povprecne_napovedi = data.frame(cbind(casi,
                                         povprecne_napovedi$fit,
                                         paste0("(", 
                                                 round(povprecne_napovedi$lwr, 2),
                                                 ", ",
                                                 round(povprecne_napovedi$upr, 2), ")"))
                                         ))

colnames(povprecne_napovedi) = c("Hitrost vetra [m/s]", "Povprečna napoved [s]", "95 % IZ")

kable(povprecne_napovedi,
      digits = c(0, 2, 0),
      caption = "Povprečna napoved časa teka.") %>%
  kable_styling("striped", full_width = F)
```

Tabela 3: Povprečna napoved časa teka.

Hitrost vetra [m/s]	Povprečna napoved [s]	95 % IZ
-1	13.41	(13.31, 13.5)
0	13.32	(13.25, 13.39)
1	13.24	(13.16, 13.32)
4	12.98	(12.78, 13.18)

```
posamicne_napovedi = data.frame(predict(m_collin, casi, interval = "prediction"))
posamicne_napovedi = data.frame(cbind(casi,
                                         posamicne_napovedi$fit,
                                         paste0("(", 
                                                 round(posamicne_napovedi$lwr, 2),
                                                 ", ",
                                                 round(posamicne_napovedi$upr, 2), ")"))
                                         ))

colnames(posamicne_napovedi) = c("Hitrost vetra [m/s]", "Posamična napoved [s]", "95% IZ")

kable(posamicne_napovedi,
      digits = c(0, 2, 0),
      caption = "Posamična napoved časa teka.") %>%
  kable_styling("striped", full_width = F)
```

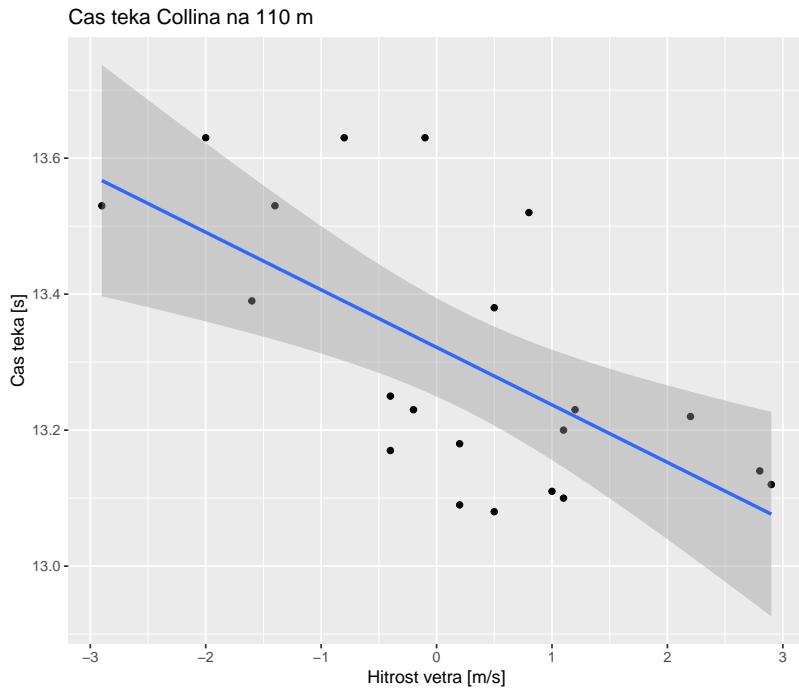
Tabela 4: Posamična napoved časa teka.

Hitrost vetra [m/s]	Posamična napoved [s]	95% IZ
-1	13.41	(13.07, 13.75)
0	13.32	(12.99, 13.66)
1	13.24	(12.9, 13.58)
4	12.98	(12.6, 13.37)

Primer interpretacije za hitrost vetra 1 m/s: napovedana vrednost za čas teka je 13.24 s. Pripadajoči 95 % IZ za povprečni čas teka pri hitrosti vetra 1 m/s je (13.16 s, 13.32 s). Za posamezni tek pri hitrosti vetra 1 m/s je pripadajoči 95 % IZ (12.9 s, 13.58 s).

Napoved za hitrost vetra 4 m/s ni upravičena, saj gre za ekstrapolacijo.

```
ggplot(data=d, aes(x = windspeed, y = time)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "lm", se = TRUE) +
  xlab("Hitrost vetra [m/s]") +
  ylab("Čas teka [s]") +
  ggtitle("Čas teka Collina na 110 m")
```



Slika 11: Odvisnost `time` od `windspeed` za dani vzorec 21 tekov in regresijska premica s 95 % intervali zaupanja za povprečno napoved.

3. Simulacija podatkov iz modela linearne regresije

Zanimajo nas lastnosti cenilk enostavnega linearne regresijskega modela. Osredotočili se bomo na testiranje domneve $H_0 : \beta_1 = 0$. Za izbrani vrednosti parametrov enostavne linearne regresije $\beta_0 = 100$ in $\beta_1 = 1$ bomo izvedli simulacije, ki bodo ilustrirale vpliv velikosti vzorca n in vrednosti variance napak σ^2 na porazdelitev cenilk parametrov in na moč testa pri testiranju domneve $H_0 : \beta_1 = 0$. Za vsako izbrano velikost vzorca n najprej generiramo vrednosti napovedne spremenljivke x na intervalu 15 do 70. Pri tem uporabimo funkcijo `sample` z argumentom `replace=TRUE`: `x<-sample(c(17:70), size=n, replace=TRUE)`. Za tako določene

vrednosti napovedne spremenljivke generiramo vrednosti odzivne spremenljivke, pri čemer upoštevamo da so pogojno na vrednosti napovedne spremenljivke porazdeljene normalno s pričakovano vrednostjo $\beta_0 + \beta_1 x$ in varianco σ^2 : $y_i = 100 + x_i + \varepsilon_i$; napake ε_i , $i = 1, \dots, 50$, generiramo s funkcijo `rnorm()` za porazdelitev $N(0, \sigma^2 = 11^2)$.

Z namenom odgovoriti na naslednja vprašanja:

- Kakšne so porazdelitve ocen parametrov enostavnega linearnega modela?
- Kolikšen delež intervalov zaupanja za β_1 vsebuje pravo vrednost parametra?
- Kolikšna je moč testa pri testiranju ničelne domneve $H_0 : \beta_1 = 0$?

bomo izvedli simulacije, pri čemer bomo podatke generirali 1000-krat in za vsak generirani vzorec izračunali cenilki parametrov enostavnega linearnega modela b_0 in b_1 , 95 % interval zaupanja za β_1 in p -vrednost pri testiranju domneve $H_0 : \beta_1 = 0$.

```
f.reg.sim <- function(x, beta0, beta1, n, sigma, nsim){

  # pripravimo prazne vektorje za rezultate simulacij, cenilki parametrov b0 in b1,
  # p-vrednost za testiranje domneve beta1=0,
  # spodnjo in zgornjo mejo intervala zaupanja za beta1
  b0 <- b1 <- l.b1 <- u.b1 <- p.b1 <- NULL

  for(i in 1:nsim){
    epsilon <- rnorm(n, mean=0, sd=sigma)
    y <- beta0 + beta1*x + epsilon
    m <- lm(y~x)
    b0[i] <- coef(m)[1]
    b1[i] <- coef(m)[2]
    l.b1[i] <- confint(m)[2, 1]
    u.b1[i] <- confint(m)[2, 2]
    p.b1[i] <- summary(m)$coef[2, 4]
  }
  return(data.frame(b0, b1, l.b1, u.b1, p.b1))
}

#parametra modela
beta0 <- 100
beta1 <- 1
#velikost vzorca
n <- 50
#standardno odklon napak
sigma <- 11
#generiramo vrednosti x
x <- sample(c(17:70), size=n, replace=TRUE)

#število simulacij
nsim <- 1000

#nastavimo seme za ponovljivost
set.seed(20)
rez.1000 <- f.reg.sim(x=x, beta0, beta1, n, sigma, nsim)

# 2.5 in 97.5 centil za b1 na podlagi simulacij
(centili <- quantile(rez.1000$b1, probs = c(0.025, 0.975)))
```

2.5% 97.5%

```

0.815409 1.186635
# ocena verjetnosti za napako II. vrste za H0: beta1=0
alfa <- 0.05
sum(rez.1000$p.b1 > alfa)/nsim

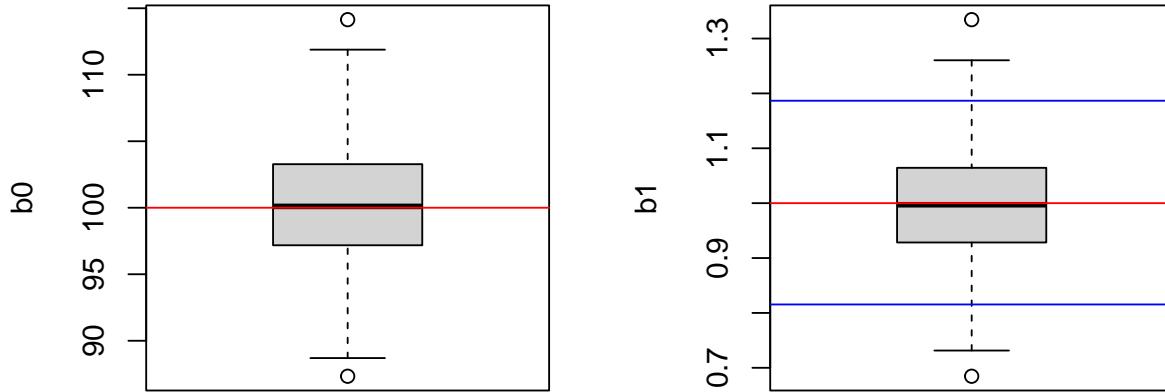
[1] 0

# ocena moči testa na podlagi Nsim simulacij
(moc.testa <- 1 - sum(rez.1000$p.b1 > alfa)/nsim)

[1] 1

par(mfrow=c(1,2))
boxplot(rez.1000$b0, ylab = "b0");
abline(h = beta0, col = "red")
boxplot(rez.1000$b1, ylab = "b1");
abline(h = beta1, col = "red");
abline(h = centili, col = "blue")

```



Slika 12: Porazdelitev cenilk parametrov b_0 (levo) in b_1 (desno) za $\sigma = 11$ in $n = 50$, `set.seed(20)`, rdeča črta kaže pravo vrednost za parameter, modri črti predstavljata 2.5 in 97.5 centil za b_1 .

```

# delež intervalov zaupanja, ki ne vsebujejo prave vrednosti parametra beta1,
# (ocena velikosti testa)
sum(rez.1000$l.b1 > beta1 | rez.1000$u.b1 < beta1)/nsim

[1] 0.052

```

Domača naloga: Simulacije iz modela enostavne linearne regresije

Simulacije ponovite za vse kombinacije:

- različnih velikosti vzorcev n : 10, 15, 50 in 1000 in
- različnih vrednosti σ : 5, 11, 22.

Grafično prikažite:

- odvisnost širine intervala zaupanja za β_1 od n , za vsako vrednost σ ;
- odvisnost širine intervala zaupanja za β_1 od σ , za vsak n ;
- odvisnost moči testa od n , za vsako vrednost σ ;
- odvisnost moči testa od σ , za vsak n

in napišite kratek povzetek vaših ugotovitev.

Vaja 2: Diagnostika linearnega modela

Seznam potrebnih R paketov:

```
library(ggplot2)
library(car)
library(effects)
library(dplyr)
library(knitr)
```

1. Preverjanje predpostavk linearnega regresijskega modela na podlagi ostankov modela

Spodaj je definirana funkcija `f.generiranje.lm.1()` za generiranje n parov podatkov enostavnega linearnega regresijskega modela:

$$y = \beta_0 + \beta_1 \cdot x_1 + \epsilon,$$

kjer:

- $x_{1,i} \sim U(1, 1000)$ (enakomerna diskretna porazdelitev),
- $\epsilon_i \sim N(0, \sigma^2)$.

```
f.generiranje.lm.1 <- function(beta0, beta1, sigma, n) {

  # generiranje vrednosti za x1
  x1 <- sample(1:1000, size = n, replace = TRUE)
  # generiranje napak
  epsilon <- rnorm(n, 0, sigma)
  # izračun napovedne spremenljivke za model (parametri = vhodni argumenti)
  y <- beta0 + beta1 * x1 + epsilon

  # podatke spravimo v podatkovni okvir in vrnemo kot rezultat funkcije
  data.frame(x1 = x1, epsilon = epsilon, y = y)

}
```

a) Pripravite funkcijo `f.generiranje.lm.2()` za generiranje n parov podatkov linearnega regresijskega modela:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \epsilon,$$

kjer:

- $x_{1,i} \sim U(50, 100)$ (enakomerna zvezna porazdelitev),
- $x_{2,i} \sim Poiss(5)$ in
- $\epsilon_i \sim N(0, \sigma^2)$

Funkcija `f.generiranje.lm.2()` naj sprejme naslednje argumente:

- β_0 ('beta0'),
- β_1 ('beta1'),
- β_2 ('beta2'),
- σ ('sigma'),

- velikost vzorca ('n').

```
f.generiranje.lm.2 <- function(beta0, beta1, beta2, sigma, n) {

  # generiranje vrednosti za x1 in x2
  x1 <- runif(n, 50, 100)
  x2 <- rpois(n, 5)

  # generiranje napak
  epsilon <- rnorm(n, 0, sigma)
  # izračun napovedne spremenljivke za model (parametri = vhodni argumenti)
  y <- beta0 + beta1 * x1 + beta2 * x2 + epsilon

  # podatke spravimo v podatkovni okvir in vrnemo kot rezultat funkcije
  data.frame(x1 = x1, x2 = x2, epsilon = epsilon, y = y)

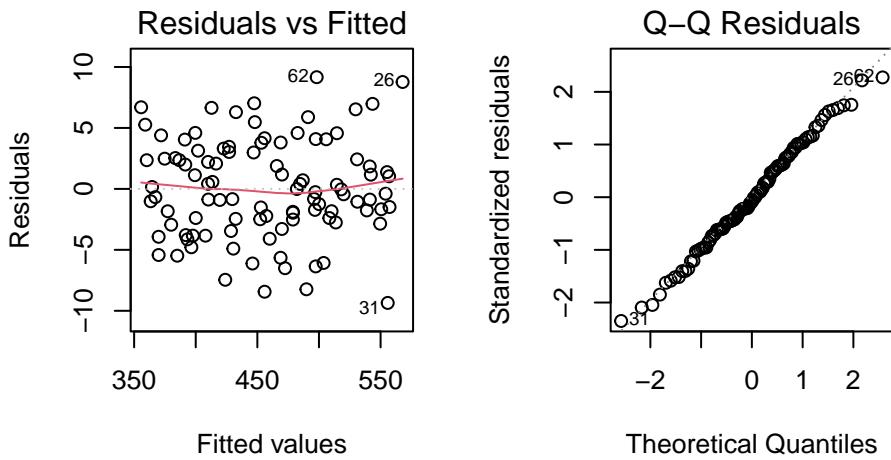
}
```

b) Za parametre:

- $\beta_0 = 150$,
- $\beta_1 = 4$,
- $\beta_2 = 2.5$,
- $\sigma = 4$,
- $n = 100$

desetkrat zaženite funkcijo `f.generiranje.lm.2()`, naredite linearni regresijski model in primerjajte prve tri grafe ostankov. **Opazujte, ali ostanki modela izpolnjujejo predpostavke linearnega regresijskega modela.** Kolikokrat izgleda, kot da ostanki niso v skladu s predpostavkami?

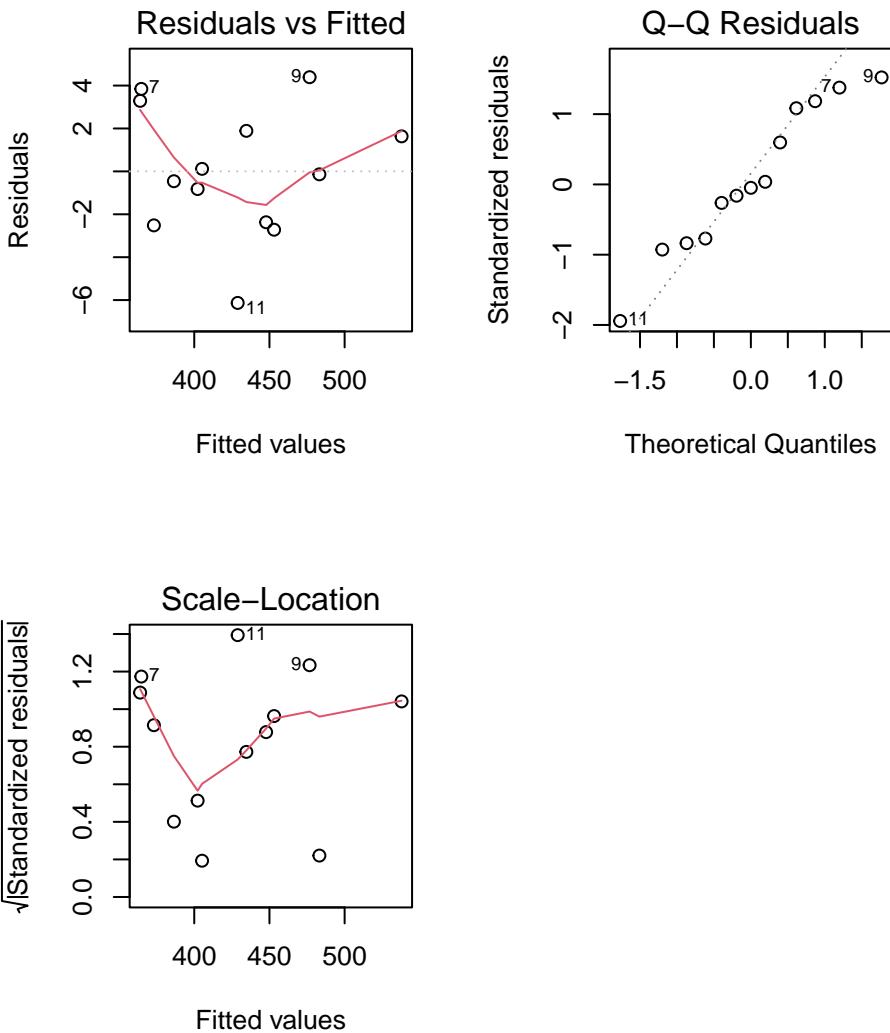
```
# zaženemo funkcijo in rezultate funkcije shranimo v vektor
generirani.podatki <- f.generiranje.lm.2(150, 4, 2.5, 4, 100)
# narišemo ostanke za linearni regresijski model na generiranih podatkih
par(mfrow = c(2,2))
plot(lm(y ~ x1 + x2, data = generirani.podatki), which = c(1:3))
```



Spremenite velikost vzorca na 13 ($n = 13$) in opazujte, kaj se v primeru manjšega vzorca dogaja z ostanki. Za generiranje uporabite vrednosti semena, ki so zapisane v vektorju `semena`.

```
# vektor semen
semena <- c(82, 145, 153, 217, 318, 411, 514, 8106)

set.seed(semena[1])
generirani.podatki <- f.generiranje.lm.2(150, 4, 2.5, 4, 13)
# narišemo ostanke za linearni regresijski model na generiranih podatkih
par(mfrow = c(2,2))
plot(lm(y ~ x1 + x2, data = generirani.podatki), which = c(1:3))
```



Kolikokrat ostanki ne kažejo izpolnjenosti predpostavk v primeru majhnega vzorca? Na kratko napišite povzetek vaših ugotovitev o vplivu velikosti vzorca na grafe ostankov.

c) Definirajte funkcijo `f.generiranje.lm.1.H()`, ki vsebuje elemente funkcije `f.generiranje.lm.1()`, s tem da krši predpostavko o konstantni varianci. Varianca napak naj bo sorazmerna z x_1 .

```
f.generiranje.lm.1.H <- function(beta0, beta1, n) {
  # generiranje vrednosti za x1
  x1 <- sample(1:1000, size = n, replace = TRUE)
  # generiranje napak; napake so sorazmerne z vrednostmi xi
  epsilon <- rnorm(n, 0, x1 * 0.8)
  # izračun napovedne spremenljivke za model (parametri = vhodni argumenti)
  y <- beta0 + beta1 * x1 + epsilon

  # podatke spravimo v podatkovni okvir in vrnemo kot rezultat funkcije
  data.frame(x1 = x1, epsilon = epsilon, y = y)
```

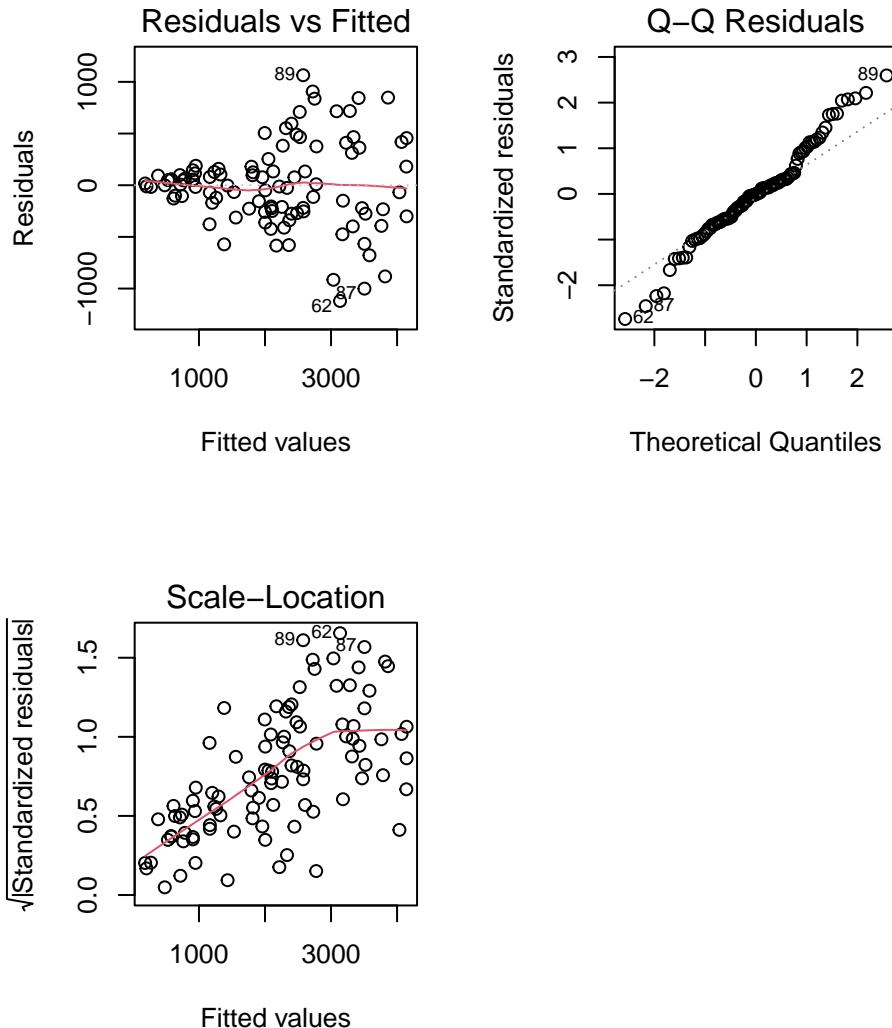
}

Za parametre:

- $\beta_0 = 150$,
- $\beta_1 = 4$,
- $n = 100$

desetkrat zaženite funkcijo `f.generiranje.lm.1.H()`, naredite linearni regresijski model in opazujte prve tri grafe ostankov.

```
# zaženemo funkcijo in rezultate funkcije shranimo v vektor
generirani.podatki.H <- f.generiranje.lm.1.H(150, 4, 100)
# narišemo ostanke za linearni regresijski model na generiranih podatkih
par(mfrow = c(2,2))
plot(lm(y ~ x1, data = generirani.podatki.H), which = c(1:3))
```



Opazujte ostanke prvega in tretjega grafa ob spremjanju odvisnosti variance napak od spremenljivke x_1

(npr. večkratnika x_1). Kakšne so vaše ugotovitve?

d) Definirajte funkcijo `f.generiranje.lm.1.N()` tako, da kršite predpostavko o normalnosti ostankov. Namesto normalne porazdelitve ostankov uporabite eksponentno porazdelitev. (Za vajo lahko poskusite še s katero drugo porazdelitvijo, npr. gama ali beta porazdelitvijo.)

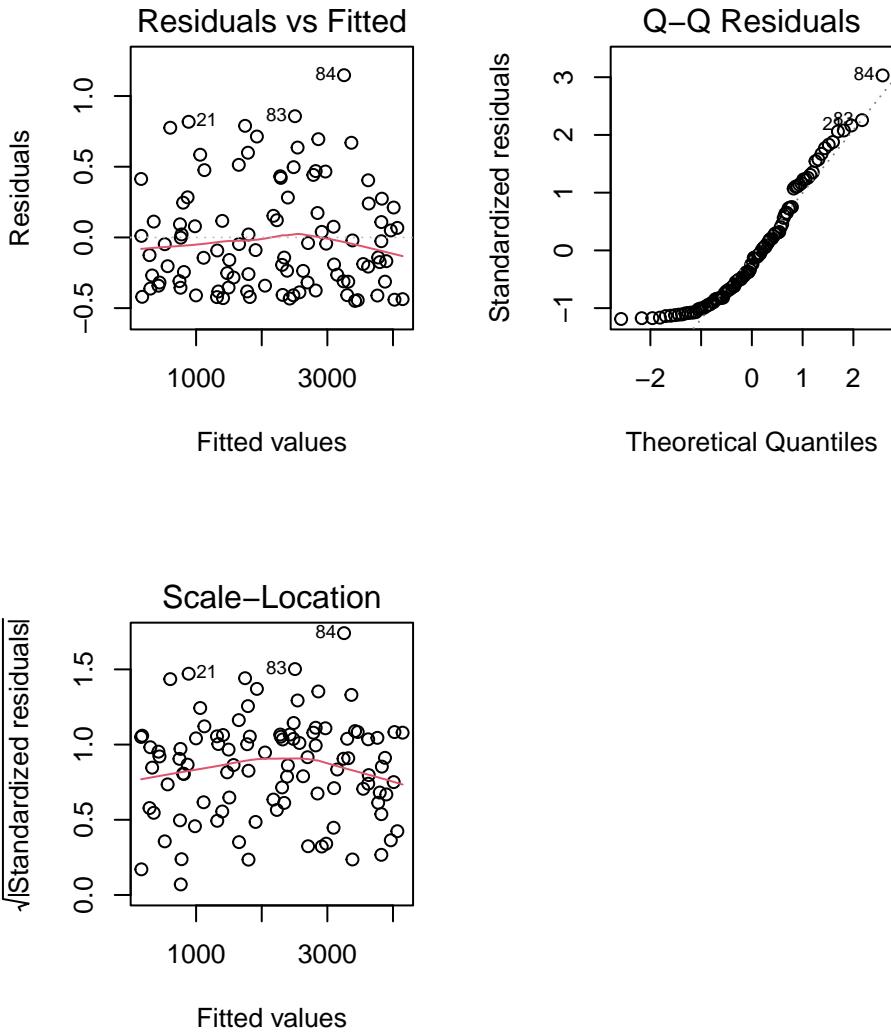
```
f.generiranje.lm.1.N <- function(beta0, beta1, n) {  
  
  # generiranje vrednosti za x1  
  x1 <- sample(1:1000, size = n, replace = TRUE)  
  # generiranje napak  
  epsilon <- rexp(n, rate = 2)  
  #epsilon <- rgamma(n, shape = 2)  
  # izračun napovedne spremenljivke za model (parametri = vhodni argumenti)  
  y <- beta0 + beta1 * x1 + epsilon  
  
  # podatke spravimo v podatkovni okvir in vrnemo kot rezultat funkcije  
  data.frame(x1 = x1, epsilon = epsilon, y = y)  
  
}
```

Za parametre:

- $\beta_0 = 150$,
- $\beta_1 = 4$,
- $n = 100$

desetkrat zaženite funkcijo `f.generiranje.lm.1.N()`, naredite linearni regresijski model in opazujte prve tri grafe ostankov. Kateri graf preverja predpostavko o normalnosti? Kakšna so odstopanja?

```
# zaženemo funkcijo in rezultate funkcije shranimo v vektor  
generirani.podatki.N <- f.generiranje.lm.1.N(150, 4, 100)  
# narišemo ostanke za linearni regresijski model na generiranih podatkih  
par(mfrow = c(2,2))  
plot(lm(y ~ x1, data = generirani.podatki.N), which = c(1:3))
```



Povzemite svoje ugotovitve.

2. Interpretacija modela z večimi napovednimi spremenljivkami

Kadar je v model vključenih več napovednih spremenljivk, postane gradnja modela hitro precej bolj kompleksna. Treba se je odločiti, katere spremenljivke je potrebno vključiti v model, ali so prisotni le glavni vplivi ali tudi interakcije ter kako bomo definirali številske in opisne spremenljivke, da bomo v modelu ustrezno opisali morebitno nelinearnost oz. diskretnost spremenljivk. Tudi interpretacija regresijskih parametrov postane bolj zapletena, saj interpretacija posameznega parametra postane odvisna od drugih spremenljivk v modelu. Načeloma lahko dani parameter interpretiramo kot povprečno oz. pričakovano razliko vrednosti odzivne spremenljivke, če primerjamo dve osebi (enoti), ki se razlikujeta za eno enoto dane napovedne spremenljivke, medtem ko so ostale vrednosti napovednih spremenljivk za obe enoti enake. Torej, posamezen regresijski parameter β_j , $j = 1, \dots, k$ meri pogojni vpliv spremenljivke X_j . To pomeni, da se interpretacija parametra β_j spremeni, če se spremeni nabor napovednih spremenljivk v modelu in obstaja povezanost (korelacija) X_j z drugimi napovednimi spremenljivkami v modelu.

Primer 1: Več koreliranih številskih spremenljivk v modelu

```
## za dani primer bomo izključili tudi osebo 39, za katero smo v prejšnji vaji videli,  
## da ima znaten vpliv na rezultate modela  
bodyfat <- bodyfat[-which(bodyfat$case==39),]
```

Radi bi pojasnili odstotek telesne maščobe s 3 spremenljivkami: telesno težo, višino in obsegom trebuha.

```
bodyfat <- bodyfat %>%  
  select(siri, weight, height, abdomen)  
  
summary(bodyfat)
```

	siri	weight	height	abdomen
Min.	: 0.00	Min. : 53.80	Min. :162.6	Min. : 69.40
1st Qu.	:12.45	1st Qu.: 72.07	1st Qu.:173.4	1st Qu.: 84.55
Median	:19.20	Median : 80.02	Median :177.8	Median : 90.90
Mean	:19.09	Mean : 80.90	Mean :178.6	Mean : 92.33
3rd Qu.	:25.25	3rd Qu.: 89.38	3rd Qu.:183.5	3rd Qu.: 99.20
Max.	:47.50	Max. :119.29	Max. :197.5	Max. :126.20

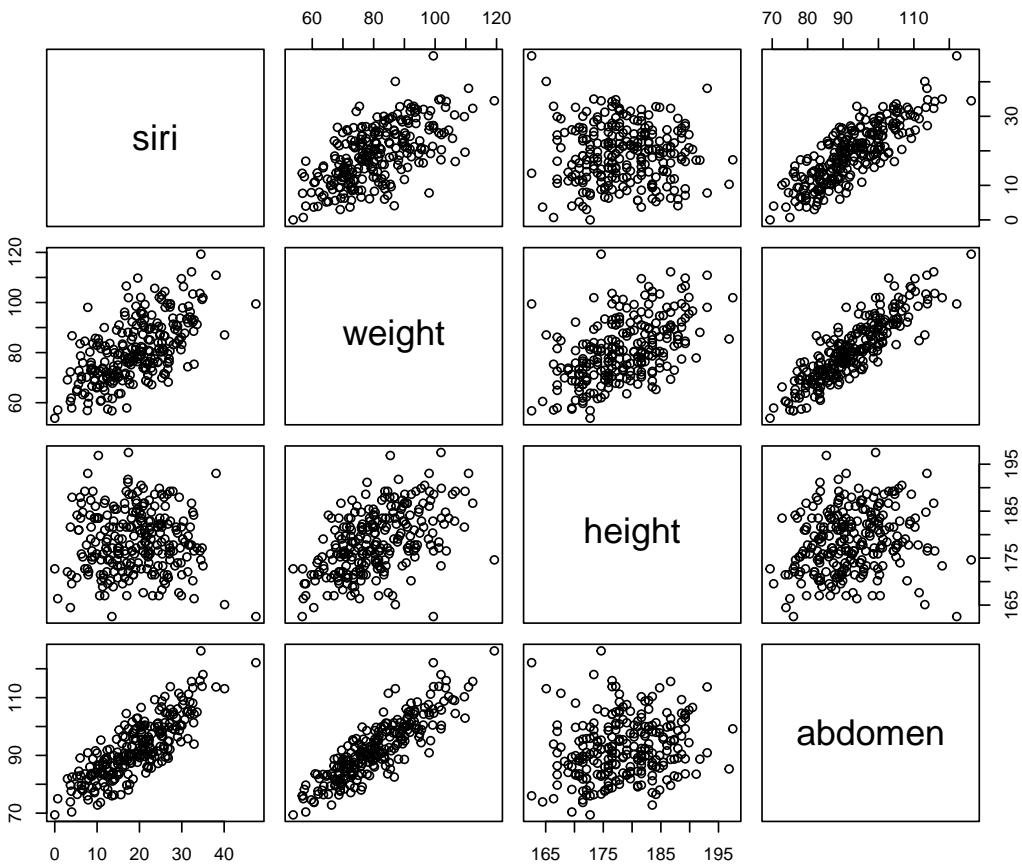
Spomnimo se močnih parnih korelacijs, ki so značilne za dani podatkovni okvir. Poglejmo Pearsonove koeficiente korelacije med posameznimi pari spremenljivk v podatkovnem okviru:

```
kable(cor(bodyfat),  
      digits=2,  
      caption = "Pearsonovi korelacijski koeficienti med pari spremenljivk  
siri, weight, height in abdomen v podatkovnem okviru bodyfat.")
```

Table 1: Pearsonovi korelacijski koeficienti med pari spremenljivk siri, weight, height in abdomen v podatkovnem okviru bodyfat.

	siri	weight	height	abdomen
siri	1.00	0.62	-0.03	0.82
weight	0.62	1.00	0.51	0.87
height	-0.03	0.51	1.00	0.18
abdomen	0.82	0.87	0.18	1.00

```
pairs(bodyfat)
```



Slika 1: Matrika razsevnih grafikonov za izbrane spremenljivke v podatkovnem okviru `bodyfat`.

Na podlagi 3 napovednih spremenljivk naredimo 4 potencialne modele:

```
m1 <- lm(siri~weight, bodyfat)

m2 <- lm(siri~weight + height, bodyfat)

m3 <- lm(siri~weight + abdomen, bodyfat)

m4 <- lm(siri~weight + height + abdomen, bodyfat)

compareCoefs(m1, m2, m3, m4)
```

Calls:

```
1: lm(formula = siri ~ weight, data = bodyfat)
2: lm(formula = siri ~ weight + height, data = bodyfat)
3: lm(formula = siri ~ weight + abdomen, data = bodyfat)
4: lm(formula = siri ~ weight + height + abdomen, data = bodyfat)
```

	Model 1	Model 2	Model 3	Model 4
(Intercept)	-14.89	77.24	-47.67	-31.07
SE	2.76	10.03	2.63	11.55

weight	0.4200	0.5827	-0.2927	-0.2190
SE	0.0337	0.0337	0.0466	0.0682
height		-0.5896		-0.0920
SE		0.0624		0.0623
abdomen		0.9794	0.9130	
SE		0.0560	0.0717	

Vidimo, da je ocena parametra za maso v 4 različnih modelih precej drugačna - ne le da spremeni velikost, temveč celo predznak. To je zato, ker je interpretacija mase v štirih modelih bistveno drugačna. Tudi za višino vidimo, da je bodisi irelevantna bodisi kaže močno povezanost s odstotkom telesne mašcobe, odvisno od tega, ali smo v modelu upoštevali tudi obseg trebuha.

```
round(c(summary(m1)$adj.r.squared,
       summary(m2)$adj.r.squared,
       summary(m3)$adj.r.squared,
       summary(m4)$adj.r.squared), 2)
```

```
[1] 0.38 0.54 0.72 0.72
```

Primerjava vrednosti prilagojenih R^2 4 modelov nakazuje na pomembno vlogo spremenljivke `abdomen` pri pojasnjevanju procenta telesne mašcobe.

Potrebno se je zavedati, da bo vsakršna izbira spremenljivk v (linearni) model, v katerega so vključene skorelirane napovedne spremenljivke, vedno spremenila interpretacijo modela. Tega se moramo zavedati predvsem v situacijah, ko nas zanima interpretacija ocen parametrov modela.

Primer 2: Številska in opisna spremenljivka v modelu

V datoteki `IQ.txt` so podatki o rezultatih IQ testa `kid_score` za 434 otrok in o ocjenjenem IQ-ju njihovih mater `mom_iq`. Za vsako od mater imamo še podatek o tem, ali je končala srednjo šolo ali ne, `mom_hs`. Kadar ocenjujemo multipli regresijski model, nas v praksi pogosto zanima naslednje:

- Ali vsaj ena od napovednih spremenljivk lahko pojasni del variabilnosti otrokovega IQ-ja? $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$
- Če lahko zavrnemo hipotezo iz prve točke: ali vse napovedne spremenljivke pomagajo pojasniti del variabilnosti napovedne spremenljivke ali zadostuje le podmnožica teh spremenljivk (več o tem v poglavju o izbiri modela)?
- Kolikšna je povezanost med napovednimi in odzivno spremenljivko (npr. kolikšno spremembo vrednosti otrokovega rezultata na testu lahko v povprečju pričakujemo, če primerjamo dva otroka mater, ki sta obe končali srednjo šolo, a se njun IQ razlikuje za 1 točko). Kako natančne so naše ocene?
- Kako natančno lahko napovemo rezultat na testu za nove otroke?
- Kako dobro se model prilega podatkom? Je v modelu prisotna nelinearnost? Ali obstaja interakcija med napovednima spremenljivkama?

```
data <- read.table("IQ.txt", header=T)
str(data)
```

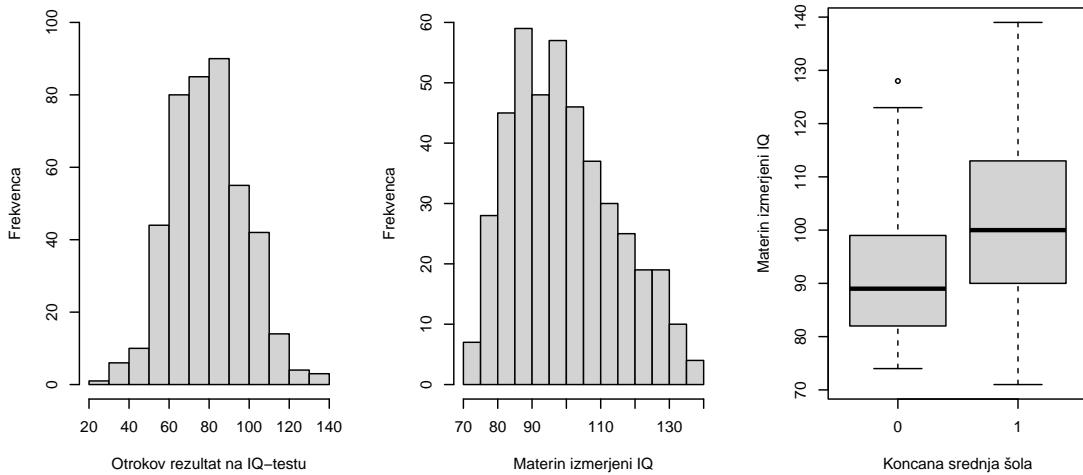
```
'data.frame': 434 obs. of 3 variables:
 $ kid_score: int 113 98 86 97 94 105 102 84 86 74 ...
 $ mom_hs   : int 1 1 1 1 1 0 1 1 1 1 ...
 $ mom_iq   : int 121 89 115 99 93 108 139 125 82 95 ...
```

```

data$mom_hs <- factor(data$mom_hs)

# poglejmo najprej univariatne porazdelitve spremenljivk v podatkovnem okviru
par(mfrow=c(1,3))
hist(data$kid_score, main="", xlab="Otrokov rezultat na IQ-testu", ylab="Frekvenca",
      ylim=c(0,100))
hist(data$mom_iq, main="", xlab="Materin izmerjeni IQ", ylab="Frekvenca", ylim=c(0,60))
boxplot(data$mom_iq ~ data$mom_hs, xlab = "Končana srednja šola",
        ylab = "Materin izmerjeni IQ")

```

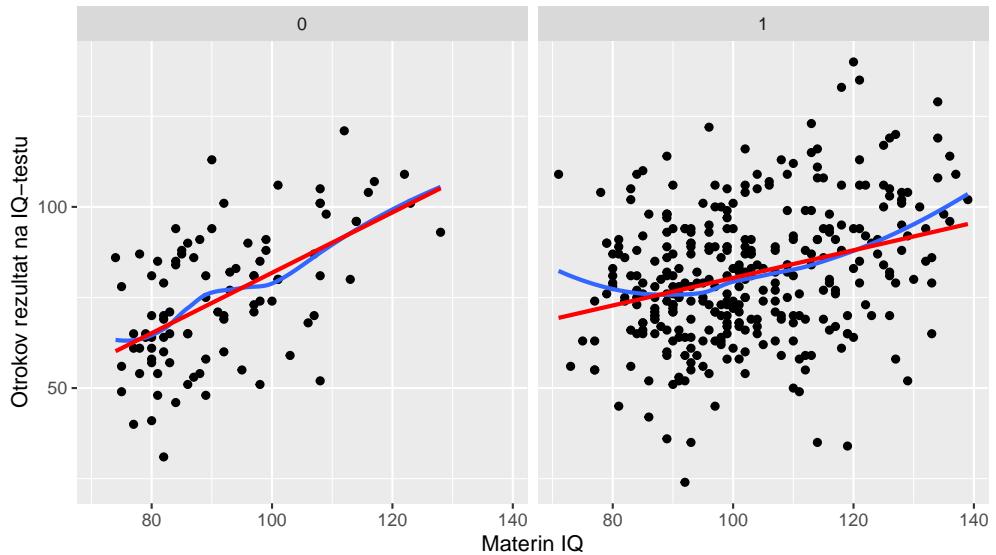


Slika 2: Univariatne porazdelitve spremenljivk v podatkovnem okviru IQ.

```

#Ali obstaja linearна povezanost med spremenljivkama?
ggplot(data=data, aes(x=mom_iq, y=kid_score)) +
  geom_point() +
  geom_smooth(se=FALSE) + geom_smooth(method="lm", se=FALSE, col="red") +
  facet_wrap(~mom_hs) +
  xlab("Materin IQ") +
  ylab("Otrokov rezultat na IQ-testu")

```



Slika 3: Odvisnost otrokovega rezultata na IQ–testu od materinega izmerjenega IQ–ja in materine izobrazbe.

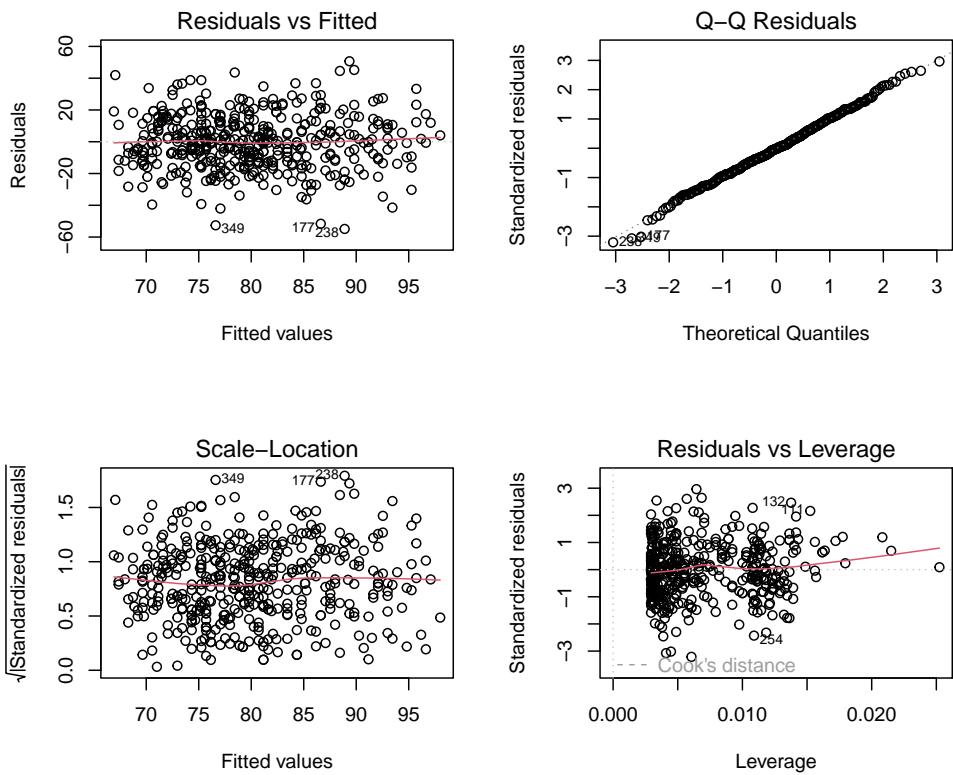
Graf nakazuje, da je zveza med `mom_iq` in `kid_score` drugačna glede na `mom_hs` in za `mom_hs=1` rahlo nelinearna.

Za vajo bomo v prvem modelu predpostavili linearno odvisnost med `kid_score` in `mom_iq`. Vanj bomo vključili le glavne vplive obeh spremenljivk, kar pomeni, da bomo predpostavili, da sta naklona enaka ne glede na `mom_hs`:

```
m1 <- lm(kid_score ~ mom_hs + mom_iq, data=data)
```

Osnovni diagnostični grafi ostankov:

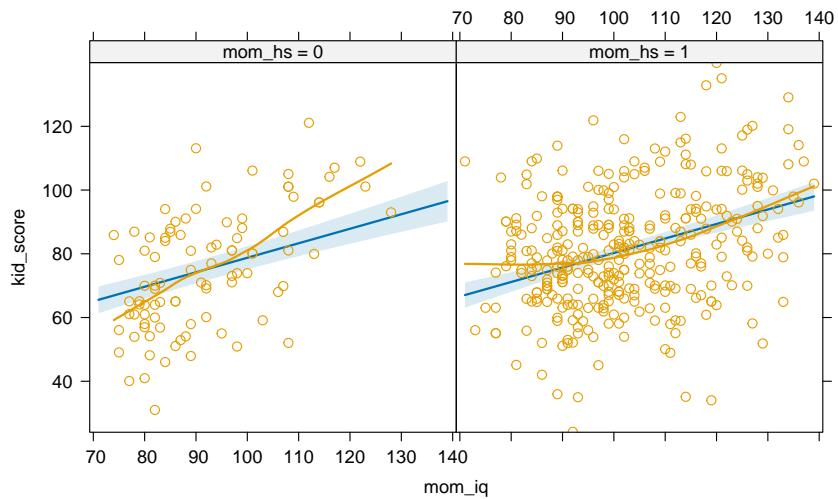
```
par(mfrow=c(2,2))
plot(m1)
```



Slika 4: Ostanki za model m1.

Ostanki za m1 izgledajo sprejemljivi. Poglejmo še grafikon parcialnih ostankov posebej glede na mom_hs, ki nam lahko pomaga pri odkrivanju interakcij ter nelinearnosti zvez.

```
plot(Effect(c("mom_iq", "mom_hs"), m1, partial.residuals=TRUE), main="")
```



Slika 5: Parcialni ostanki za model m1.

Gladilnika kažeta, da se `kid_score` v odvisnosti od `mom_iq` spreminja drugače glede na materino izobrazbo, kar nakazuje prisotnost interakcije med `mom_iq` in `mom_hs`. Poleg tega se gladilnik pri `mom_hs=1` ne prilega dobro premici, kar nakazuje, da bi lahko bila v modelu prisotna nelinearnost v odvisnosti `kid_score` od `mom_iq` in `mom_hs=1`.

Model z eno številsko in eno opisno spremenljivko, ki predpostavlja le glavne (aditivne) vplive na odzivno spremenljivko, bo dal ocene parametrov dveh vzporednih premic.

Čeprav model ni ustrezен, si za vajo oglejmo povzetek modela ter interpretirajmo ocene parametrov:

```
summary(m1)
```

Call:

```
lm(formula = kid_score ~ mom_hs + mom_iq, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-54.90	-11.76	-0.26	11.34	50.65

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	33.24436	5.54363	5.997	4.26e-09 ***							
mom_hs1	1.49673	2.09049	0.716	0.474							
mom_iq	0.45510	0.05714	7.964	1.49e-14 ***							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

Residual standard error: 17.14 on 431 degrees of freedom
Multiple R-squared: 0.1448, Adjusted R-squared: 0.1409
F-statistic: 36.49 on 2 and 431 DF, p-value: 2.282e-15

Zapišimo model `m1: 33.24 + 1.5 * mom_hs + 0.46 * mom_iq`.

Ker ima `mom_hs` dve možni vrednosti, $\text{mom_hs} \in \{0, 1\}$, dobimo oceni za dve regresijski premici z različnima presečiščema in enakima naklonoma:

- `mom_hs=0` (referenčna kategorija): $33.24 + 0.46 * \text{mom_iq}$;
- `mom_hs=1`: $(33.24 + 1.5) + 0.46 * \text{mom_iq} = 34.74 + 0.46 * \text{mom_iq}$.

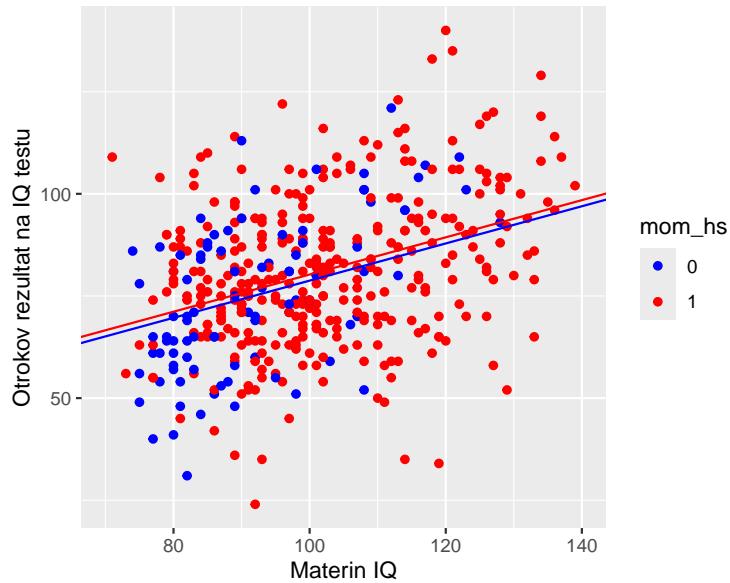
Za vajo interpretirajmo posamezne parametre modela `m1`:

- *Presečišče*: za otroka, katerega mati ima IQ enak 0 in ni končala srednje šole, bi bila povprečna napoved rezultata na testu enaka 33.24. Interpretacija v tem primeru ni smiselna, saj nobena mati nima IQ-ja enakega 0.
- *Koeficinet mom_hs*: če primerjamo otroka, katerih matere imata enak IQ, a je mati prvega končala srednjo šolo, mati drugega pa ne, ima prvi otrok v povprečju za 1.5 točke boljši rezultat na testu.
- *Koeficinet mom_iq*: če primerjamo otroka, katerih matere imata enako vrednost `mom_hs`, a se razlikujeta za 1 točko IQ-ja, ima otrok matere z višjim IQ-jem v povprečju za 0.46 točke boljši rezultat na testu (oz. je povprečna razlika 4.6 točk, če se materi razlikujeta za 10 točk IQ-ja).

in prikažimo povprečne napovedi `kid_score` na podlagi `m1`:

```
ggplot(data, aes(mom_iq, kid_score)) +
  geom_point(aes(color = mom_hs), show.legend = TRUE) +
  geom_abline(intercept = c(coef(m1)[1], coef(m1)[1] + coef(m1)[2]),
              slope = coef(m1)[3],
              color = c("blue", "red")) +
```

```
scale_color_manual(values = c("blue", "red")) + xlim(c(70,140)) +
  labs(x = "Materin IQ", y = "Otrokov rezultat na IQ testu")
```



Slika 6: Odvisnost otrokovega rezultata na IQ testu od materinega izmerjenega IQ-ja in materine izobrazbe. Črti predstavlja povprečno napoved na podlagi modela m1 za otroke, katerih matere so (rdeča) in niso (modra) končale srednjo šolo.

V naslednjem koraku bomo sprostili predpostavko, da sta naklona za mom_iq enaka ne glede na mom_hs:

```
m2 <- lm(kid_score ~ mom_hs * mom_iq, data=data)
```

Ali je interakcija v modelu potrebna ali ne, lahko preverimo z F -testom. Z ukazom `anova(model)` izvedemo sekvenčni F -test, ki testira vpliv posamezne spremenljivke ob upoštevanju predhodnjih spremenljivk v modelu.

```
anova(m2)
```

Analysis of Variance Table

```
Response: kid_score
          Df Sum Sq Mean Sq F value    Pr(>F)
mom_hs       1  2808   2808.3  9.7283  0.001937 **
mom_iq       1 18640  18639.6 64.5690 9.069e-15 ***
mom_hs:mom_iq 1  2521   2520.7  8.7321  0.003298 **
Residuals   430 124131    288.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V prvi vrstici izpisa z F -testom primerjamo model, kjer smo vključili napovedno spremenljivko `mom_hs` z ničelnim modelom, ki vsebuje le presečišče. V drugi vrstici primerjamo model, ki vključuje `mom_hs` in `mom_iq` z modelom, ki vključuje le `mom_hs`. V zadnji vrstici testiramo domnevo, ali je v modelu značilna interakcija med `mom_hs` in `mom_iq`.

Prisotnost interakcije lahko preverimo tudi na podlagi F -testa za primerjavo gnezdenih modelov, ki testira domnevo, da sta modela ekvivalentna. Ničelno domnevo lahko zavrnemo: modela nista ekvivalentna, interakcija je v modelu potrebna. Primerjajte rezultate obeh testov.

```
anova(m1, m2)
```

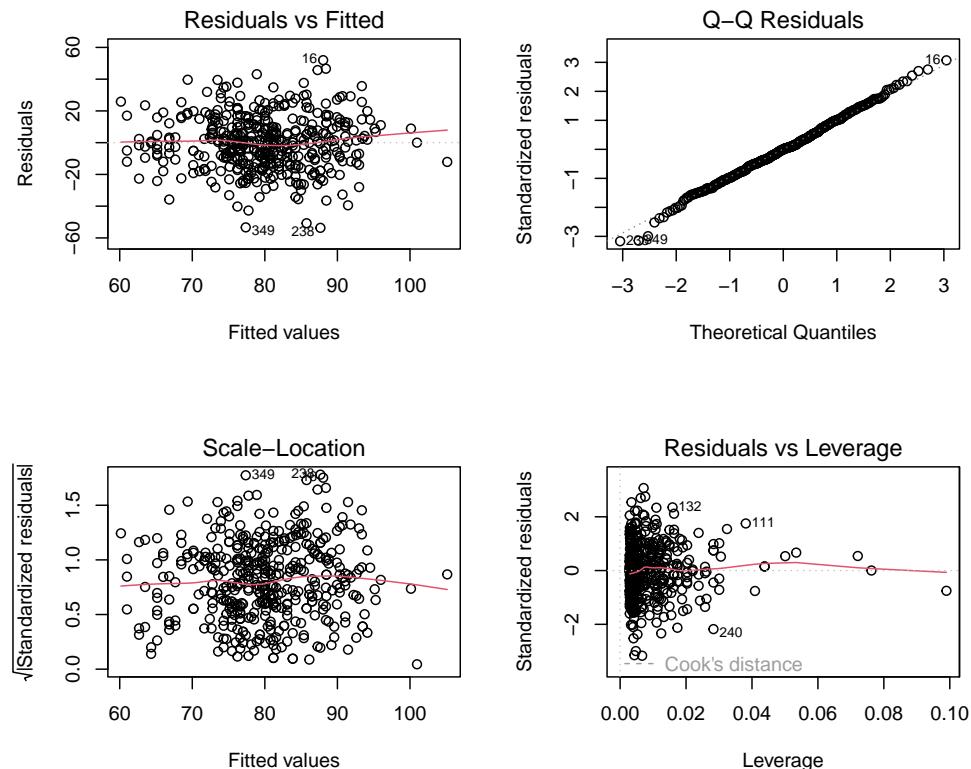
Analysis of Variance Table

```
Model 1: kid_score ~ mom_hs + mom_iq
Model 2: kid_score ~ mom_hs * mom_iq
Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     431 126652
2     430 124131  1      2520.8 8.7321 0.003298 **
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Diagnostika modela:

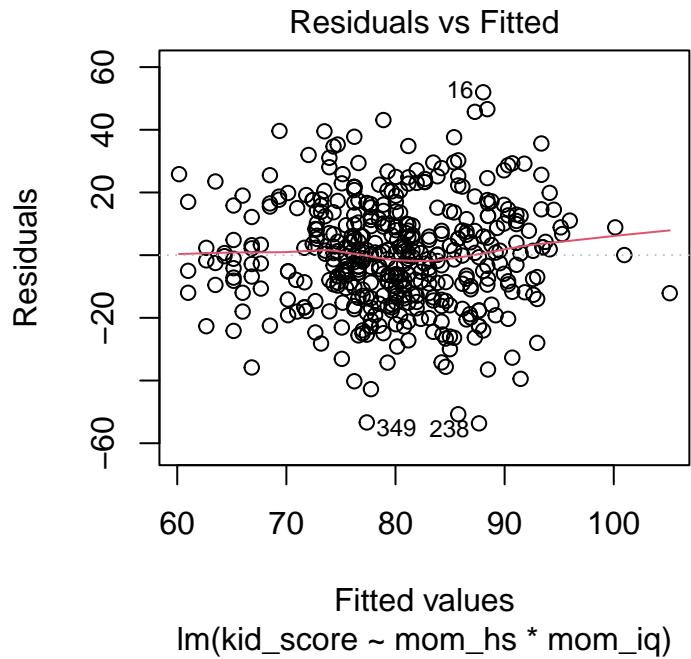
```
par(mfrow=c(2,2))
plot(m2)
```



Slika 7: Ostanki za model m2.

Slike ostankov so sprejemljive, čeprav je na prvi sličici, ki prikazuje ostanke v odvisnosti od napovedanih vrednosti, vidna rahla nelinearnost vpliva napovedne spremenljivke `mom_iq` na `kid_score`. Sliko poglejmo pobliže:

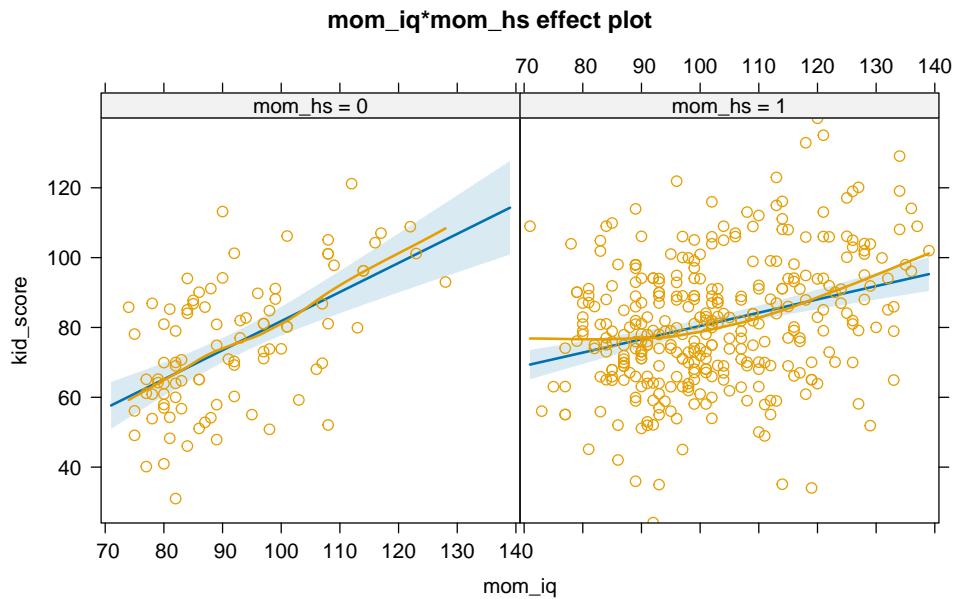
```
plot(m2, which = 1)
```



Slika 8: Ostanki za model m2.

Poglejmo še grafikon parcialnih ostankov:

```
plot(Effect(c("mom_iq", "mom_hs"), m2, partial.residuals=TRUE))
```



Slika 9: Parcialni ostanki za model m2.

Vidimo, da prihaja le do manjših odstopanj gladilnika v repih pri `mom_hs=1`, torej smo z vključeno interakcijo situacijo (vsaj deloma) popravili. V kolikor nas zanima interpretacija ocen parametrov, bi v praksi tak model privzeli kot zadovoljiv; v kolikor bi nas zanimale natančne napovedi, bi model poizkušali izboljšati tako, da bi nelinearnost modelirali s polinomsko regresijo ali zlepki. Na račun večje fleksibilnosti (ter kompleksnosti) modela, s katero bi dobili bolj natančne napovedi, pa bi žrtvovali del njegove interpretabilnosti.

V tej vaji bomo privzeli, da je kljub manjši kršitvi predpostavke o linearnosti, naš model zadovoljiv. Za interpretacijo si poglejmo izpis povzetka modela:

```
summary(m2)
```

Call:
`lm(formula = kid_score ~ mom_hs * mom_iq, data = data)`

Residuals:

Min	1Q	Median	3Q	Max
-53.654	-10.834	-0.049	11.001	51.966

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	-1.4537	12.9641	-0.112	0.91077							
mom_hs1	43.7778	14.4575	3.028	0.00261 **							
mom_iq	0.8327	0.1398	5.958	5.33e-09 ***							
mom_hs1:mom_iq	-0.4518	0.1529	-2.955	0.00330 **							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

Residual standard error: 16.99 on 430 degrees of freedom
Multiple R-squared: 0.1618, Adjusted R-squared: 0.156
F-statistic: 27.68 on 3 and 430 DF, p-value: < 2.2e-16

Ocenjeni model m2 lahko zapišemo: $-1.45 + 43.78 * \text{mom_hs} + 0.83 * \text{mom_iq} - 0.45 * \text{mom_hs} * \text{mom_iq}$.

Najlaže si je rezultate razložiti v smislu dveh premic z različnima presečiščema in naklonoma:

- `mom_hs=0` (referenčna kategorija): $-1.45 + 0.83 * \text{mom_iq}$;
- `mom_hs=1`: $(-1.45 + 43.78) + (0.83 - 0.45) * \text{mom_iq} = 42.32 + 0.38 * \text{mom_iq}$.

Razlaga posameznih parametrov:

- *Presečišče*: predstavlja napovedano vrednost rezultata na IQ-testu za tiste otroke, katerih matere niso končale srednje šole in so imele IQ enak 0 (interpretacija ni smiselna).
- *Koeficient mom_hs*: predstavlja napovedano razliko rezultata na IQ-testu za dva otroka, katerih matere imata IQ enak 0, a se razlikujeta glede na to, ali sta končali srednjo šolo (interpretacija ni smiselna).
- *Koeficient mom_iq*: predstavlja napovedano razliko rezultata na IQ-testu za dva otroka, katerih matere nista končala srednje šole, a se njun IQ razlikuje za 1.
- *Interakcija* predstavlja napovedano razliko naklonov za `mom_iq` za matere, ki so oz. niso končale srednje šole.

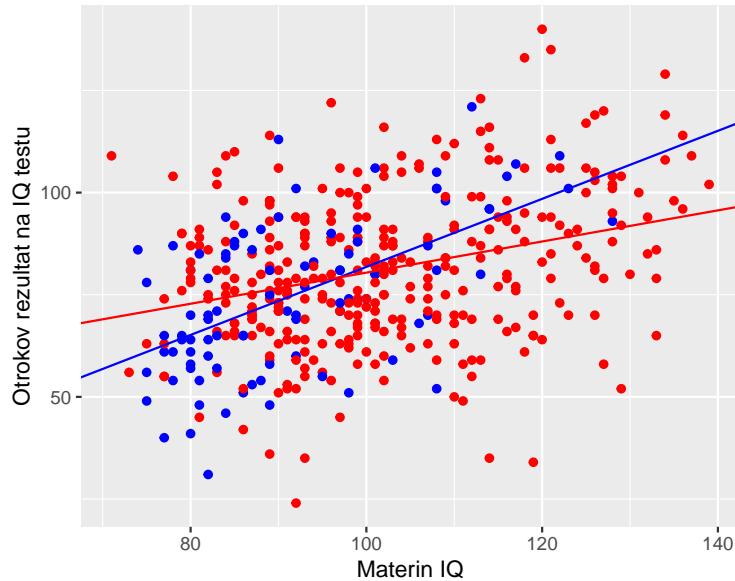
Ocenjene napovedi na podlagi modela m2:

```
ggplot(data, aes(mom_iq, kid_score)) +
  geom_point(aes(color = factor(mom_hs)), show.legend = FALSE) +
  geom_abline(
    intercept = c(coef(m2)[1], sum(coef(m2)[1:2])),
    slope = c(coef(m2)[3], sum(coef(m2)[3:4])),
```

```

color = c("blue", "red")) +
scale_color_manual(values = c("blue", "red")) +
labs(x = "Materin IQ", y = "Otrokov rezultat na IQ testu")

```



Slika 10: Odvisnost otrokovega rezultata na IQ testu od materinega izmerjenega IQ-ja in materine izobrazbe. Črti predstavljata povprečno napoved na podlagi modela m2 za otroke, katerih matere so (rdeča) in niso (modra) končale srednjo šolo.

Videli smo, da je oceno za presečišče težko interpretirati, kadar napovedne spremenljivke ne vključujejo vrednosti nič. Interpretacijo lahko olajšamo tako, da spremenljivke centriramo ali pa uporabimo neko referenčno točko; v našem primeru vemo, da je populacijsko povprečje IQ-ja enako 100, tako da bo 100 naša referenčna točka:

```

#data$mom_iq_centered <- data$mom_iq - mean(data$mom_iq)
#data$mom_hs_centered <- data$mom_hs - mean(data$mom_hs)

data$mom_iq_centered <- data$mom_iq - 100 # odštejemo populacijsko povprečje
#data$mom_hs_centered <- data$mom_hs - 0.5 # odštejemo sredinsko točko

m2.2 <- lm(kid_score ~ mom_hs * mom_iq_centered, data=data)
summary(m2.2)

```

Call:
`lm(formula = kid_score ~ mom_hs * mom_iq_centered, data = data)`

Residuals:

Min	1Q	Median	3Q	Max
-53.654	-10.834	-0.049	11.001	51.966

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	81.8156	2.0948	39.057	< 2e-16 ***
mom_hs1	-1.3995	2.2921	-0.611	0.5418

```

mom_iq_centered      0.8327      0.1398    5.958 5.33e-09 ***
mom_hs1:mom_iq_centered -0.4518     0.1529   -2.955   0.0033 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 16.99 on 430 degrees of freedom
 Multiple R-squared: 0.1618, Adjusted R-squared: 0.156
 F-statistic: 27.68 on 3 and 430 DF, p-value: < 2.2e-16

- Koeficient za mom_hs zdaj predstavlja napovedano razliko rezultata na IQ-testu za dva otroka, katerih matere imata IQ enak 100, a se razlikujeta glede na to, ali sta končali srednjo šolo.
- Koeficient za mom_iq_centered predstavlja napovedano razliko rezultata na IQ-testu za dva otroka, katerih matere nista končala srednje šole, a se njun IQ razlikuje za 1.

Vrednost R^2 za ta model znaša 0.16. Z modeliranjem glavnih vplivov mom_hs in mom_iq ter njune interakcije smo torej uspeli pojasniti 16.18 % variabilnosti odzivne spremenljivke kid_score.

Model na centriranih podatkih brez presečišča:

```
m2.3 <- lm(kid_score ~ -1 + mom_hs * mom_iq_centered , data=data)
summary(m2.3)
```

Call:

```
lm(formula = kid_score ~ -1 + mom_hs * mom_iq_centered, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-53.654	-10.834	-0.049	11.001	51.966

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
mom_hs0	81.8156	2.0948	39.057	< 2e-16 ***
mom_hs1	80.4161	0.9304	86.435	< 2e-16 ***
mom_iq_centered	0.8327	0.1398	5.958	5.33e-09 ***
mom_hs1:mom_iq_centered	-0.4518	0.1529	-2.955	0.0033 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.99 on 430 degrees of freedom
 Multiple R-squared: 0.9575, Adjusted R-squared: 0.9571
 F-statistic: 2422 on 4 and 430 DF, p-value: < 2.2e-16

Namesto napovedane razlike rezultata na IQ-testu za dva otroka, katerih matere imata IQ enak 100, a se razlikujeta glede na to, ali sta končali srednjo šolo, tu dobimo napovedani vrednosti kid_score za otroka, katerega mati ima IQ enak 100 in ni (mom_hs0) oz. je (mom_hs1) končala srednje šolo.

Primerjajmo modela z in brez presečišča:

```
anova(m2.2)
```

Analysis of Variance Table

Response: kid_score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mom_hs	1	2808	2808.3	9.7283	0.001937 **
mom_iq_centered	1	18640	18639.6	64.5690	9.069e-15 ***

```

mom_hs:mom_iq_centered    1    2521   2520.7  8.7321  0.003298 **
Residuals                  430 124131    288.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
anova(m2.3)

```

Analysis of Variance Table

```

Response: kid_score
          Df  Sum Sq Mean Sq  F value    Pr(>F)
mom_hs           2 2775930 1387965 4808.0067 < 2.2e-16 ***
mom_iq_centered 1   18640    18640   64.5690 9.069e-15 ***
mom_hs:mom_iq_centered 1    2521    2521    8.7321  0.003298 **
Residuals        430 124131      289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Vrednost za $SS_{m2.3}$ je presenetljivo velika. Matematično lahko pokažemo, da v modelu brez presečišča izraz SS_y ne razпадa na vsoto $SS_{model} + SS_{residual}$. Tudi povprečje ostankov v takem modelu ni nujno enako 0.

Primerjajmo vrednost $R^2 = SS_{model}/SS_{total} = 1 - SS_{res}/SS_{total}$ v modelu s presečiščem:

```

y_fit_m2.2 <- m2.2$fitted.values
SS.res <- sum((y_fit_m2.2 - data$kid_score)^2)
SS.total <- sum((data$kid_score - mean(data$kid_score))^2)
1-SS.res/SS.total

```

[1] 0.1618412

z vrednost R^2 v modelu brez presečišča:

```

y_fit_m2.3 <- m2.3$fitted.values
SS.res <- sum((y_fit_m2.3 - data$kid_score)^2)
SS.total <- sum((data$kid_score - 0)^2)
# SS.total se računa relativno na vrednost 0!
1-SS.res/SS.total

```

[1] 0.957507

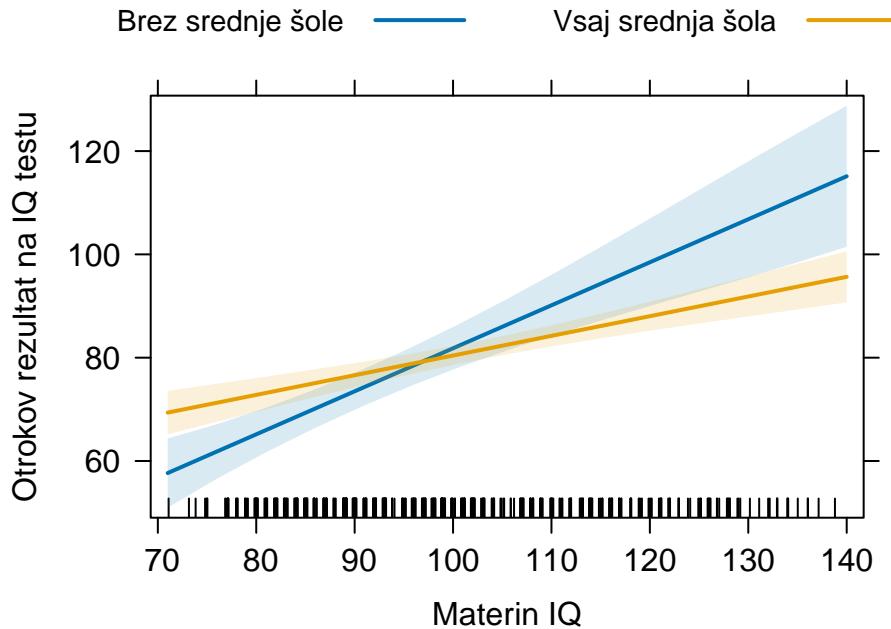
R^2 v modelu brez presečišča tako ne moremo interpretirati kot delež pojasnjene variabilnosti.

Pri interpretaciji modela si lahko pomagamo tudi z grafičnimi prikazi iz paketa **effects**:

```

plot(Effect(c("mom_iq", "mom_hs"), m2),
     multiline = TRUE, ci.style = "bands", main = "",
     xlab = "Materin IQ", ylab="Otrokov rezultat na IQ testu",
     key.args = list(space="top",
                     text = list(c("Brez srednje šole", "Vsaj srednja šola"), cex = .9),
                     title=""))

```



Slika 11: Napovedane vrednosti za `kid_score` v odvisnosti od materinega IQ-ja glede na materino izobrazbo za model `m2`.

Kaj mislite, ali lahko zvezo med `mom_iq` in `kid_score` interpretiramo kot vzročno-posledično?

V običajnem regresijskem kontekstu, kadar je namen modeliranja deskriptiven, se interpretacija nanaša na primerjave med enotami. Pri vzročnem sklepanju pa primerjamo dva potencialna izida (*potential outcomes*) na isti enoti, če bi bila izpostavljena dvem različnim obravnavanjem (vprašanje: *What if?*). Na splošno lahko rezultate regresijske modela interpretiramo v smislu vzorka in poslednice le ob močnih predpostavkah oz. v kontekstu načrtovanih poskusov, saj z načrtovanjem zbiranja podatkov lahko zagotavimo, da je dodelitev obravnavanj posameznim enotam neodvisna od potencialnih izidov (pogojna glede na dejavnike, ki smo jih upoštevali pri načrtovanju poskusa).

V praksi pa načrtovani poskusi niso vedno mogoči zaradi različnih logističnih, etičnih ali finančnih omejitev. Interpretacija vplivov proučevanih dejavnikov v smislu vzroka in posledice je lahko pristranska, če dodelitev obravnavanj posameznim enotam ni slučajen (skupine, ki jih primerjamo, se razlikujejo v mnogih t.i. motečih spremenljivkah, ki tudi vplivajo na izid). Če želimo rezultate kljub temu interpretirati v smislu vzroka in posledice, moramo v regresijskem modelu upoštevati vse moteče dejavnike, ki pojasnjujejo alokacijo enot v posamezna obravnavanja. Glavne težave se pojavi pri vprašanju, katere moteče spremenljivke je potrebno upoštevati v modelu, poleg tega pa se posledično lahko zgodi, da naš končni model vključuje veliko število spremenljivk.

Domača naloga: Povzetek ugotovitev simulacij

Za domačo nalogo zapišite kratek povzetek vaših ugotovitev iz današnjih vaj in ponovite oz. dopолните simulacije. V pomoč so vam lahko naslednja vprašanja:

- Kako velikost vzorca vpliva na diagnostiko grafov ostankov?
- Kako na grafu ostankov zaznamo prisotnost heteroskedastičnosti?

- Kako na grafu opazimo, da ostanki niso porazdeljeni normalno?
- Na kaj vpliva heteroskedastičnost?
- S simulacijami pokažite, kaj kaj se zgodi z velikostjo testa v primeru kršitve predpostavke o konstantni varianci.

Vaja 3: Predpostavke niso izpolnjene

Seznam potrebnih R paketov:

```
library(ggplot2)
library(ggpubr)
library(car)
library(effects)
library(dplyr)
library(ISLR2)
```

1. R^2 v modelu brez presečišča

R^2 v modelu s presečiščem lahko izpeljemo iz izraza, ki razdeli vsoto kvadratov odklonov odzivne spremenljivke SS_{yy} na dva dela: del SS_{model} , ki ga pojasni linearni model, ter del $SS_{residual}$, ki ostane z modelom nepojasnjen:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$
$$SS_{yy} = SS_{model} + SS_{residual},$$

pri čemer smo upoštevali, da je $2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$.

R^2 je delež variabilnosti odzivne spremenljivke, ki je pojasnjen z modelom:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{SS_{model}}{SS_{yy}}.$$

R^2 torej primerja dani model z modelom, ki vsebuje le presečišče.

Kadar model nima presečišča, ga ni smiselno primerjati z modelom, ki vključuje le presečišče. V takem primeru gre regresijska premica namesto skozi \bar{y} skozi izhodišče, in R^2 je enak:

$$R_0^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2}.$$

Vrednosti R^2 dveh modelov - s presečiščem in brez presečišča - ni mogoče neposredno primerjati, saj temeljijo na različnih izračunih. Vrednost R-kvadrata bo na splošno višja v modelu brez presečišča, vendar to nujno ne pomeni, da je ta model boljši. Načeloma model brez presečišča uporabimo le v primerih, ko iz teorije vemo, da je presečišče enako nič.

2. Posebne točke v modelu

Vrnimo se na primer iz prejšnje vaje, kjer smo na podlagi podatkovnega okvira `bodyfat` pojasnjevali odstotek telesne maščobe na podlagi 3 spremenljivk: telesne teže, višine in obsega trebuha. Še enkrat poglejmo, kako izgledajo parne povezanosti z odzivno spremenljivko. Funkcija `scatterplot` v paketu `car` omogoča identifikacijo dveh enot z največjo Mahalanobisovo razdaljo od središča podatkov.

```
scatterplot(siri ~ weight, data=bodyfat,
            smooth=list(smooth=loessLine, border=FALSE, style="none"),
            regLine=TRUE, id=TRUE, boxplots=FALSE,
            cex.axis=1.5, cex.lab=1.5)
```

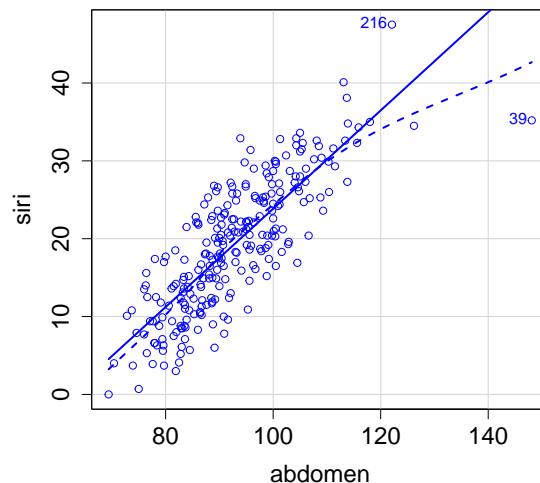
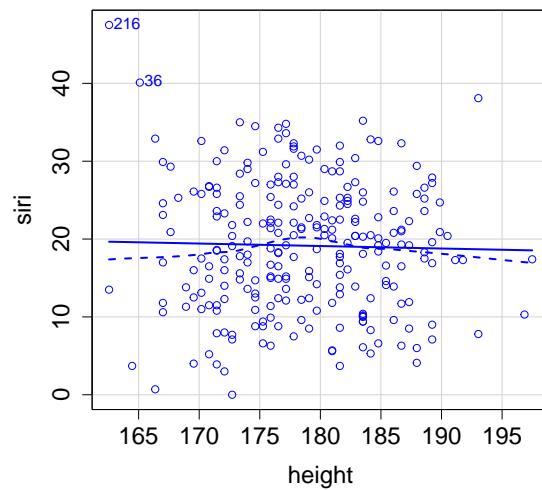
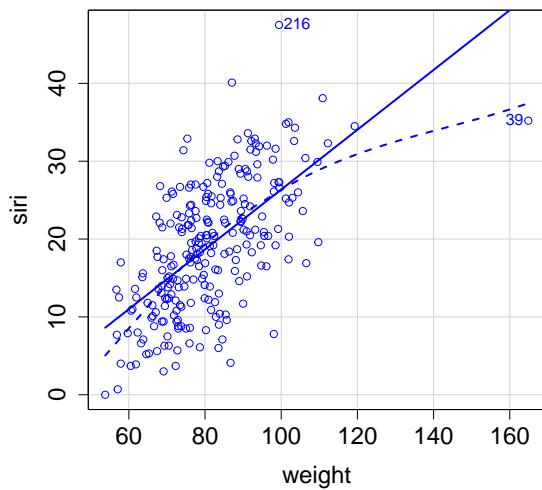
[1] 39 216

```
scatterplot(siri ~ height, data=bodyfat,
smooth=list(smooth=loessLine, border=FALSE, style="none"),
regLine=TRUE, id=TRUE, boxplots=FALSE,
cex.axis=1.5, cex.lab=1.5)
```

[1] 36 216

```
scatterplot(siri ~ abdomen, data=bodyfat,
smooth=list(smooth=loessLine, border=FALSE, style="none"),
regLine=TRUE, id=TRUE, boxplots=FALSE,
cex.axis=1.5, cex.lab=1.5)
```

[1] 39 216

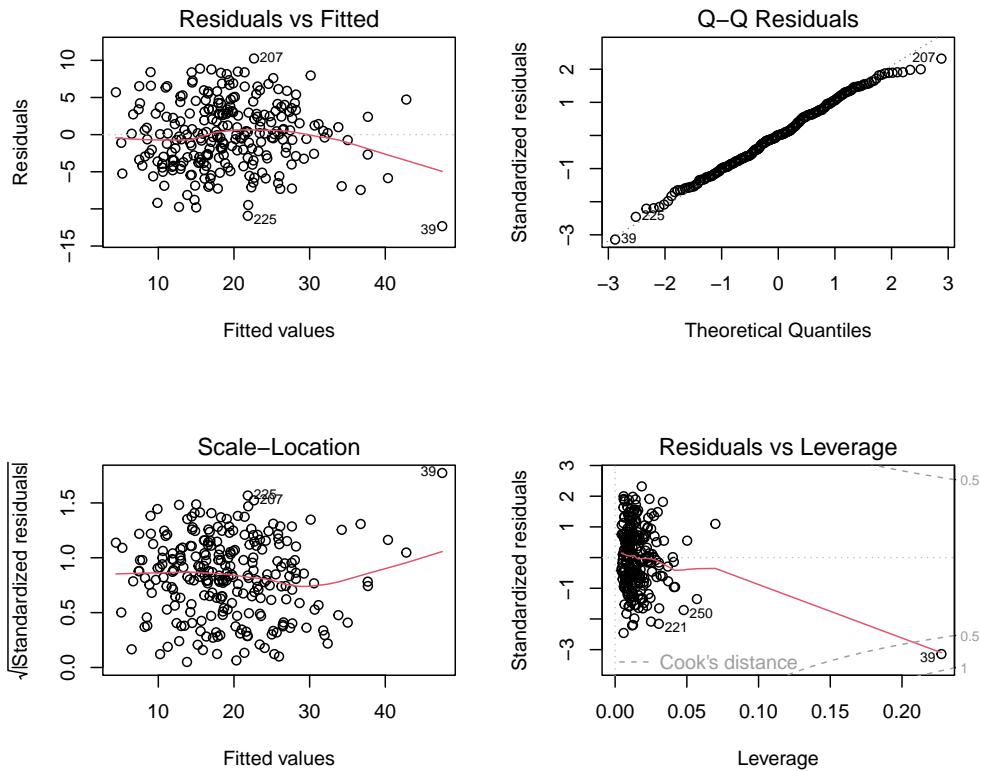


Izstopata predvsem enoti 39 in 216. V prejšnji vaji smo osebo 39 a priori izključili iz modela. Tokrat bomo naredili model na vseh 252 enotah.

```
m.bodyfat <- lm(siri~weight + height + abdomen, bodyfat)
```

Poglejmo, kako izgledajo osnovni diagnostični grafi ostankov za model `m.bodyfat`:

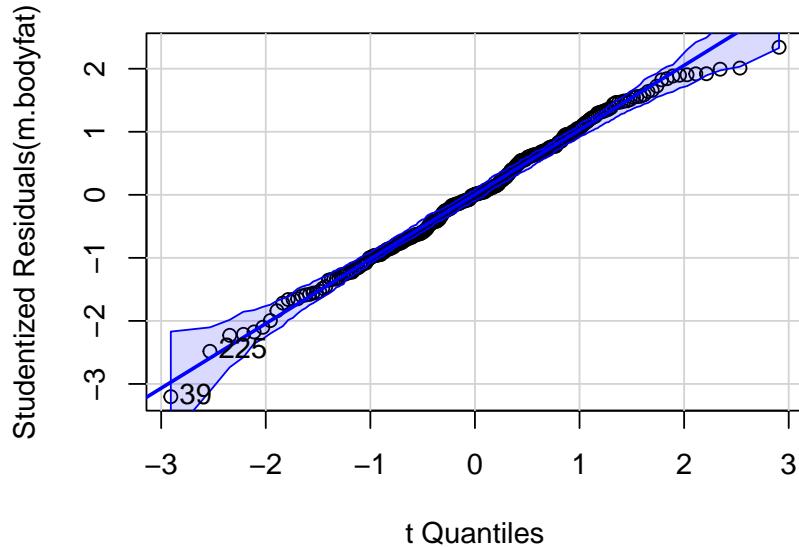
```
par(mfrow=c(2,2))
plot(m.bodyfat)
```



Slika 1: Ostanki za model `m.bodyfat`.

Izstopa 39. enota, ki ima veliko vrednost standardiziranih ostankov ter veliko vrednost Cookove razdalje. Torej bomo nadaljevali z analizo posebnih točk.

```
qqPlot(m.bodyfat)
```



Slika 2: QQ-grafikon za studentizirane ostanke za `m2` s 95 % bootstrap ovojnico.

```
[1] 39 225
```

Slika porazdelitve ostankov kaže, da imamo v modelu nekaj točk, ki imajo studentizirano ostanke po absolutni vrednosti večje od 2. Po privzetih nastavitevah funkcija identificira dve enoti z največjima vrednostima studentiziranih ostankov. Nobena točka ni daleč zunaj 95 % bootstrap ovojnice, kar kaže na to, da v modelu nimamo regresijskih osamelcev.

Naredimo še statistični test, ki temelji na studentiziranih ostankih in testira ničelno domnevo, ki pravi, da i -ta točka, $i = 1, \dots, n$, ni regresijski osamelec:

```
outlierTest(m.bodyfat)
```

```
No Studentized residuals with Bonferroni p < 0.05
```

```
Largest |rstudent|:
```

rstudent	unadjusted	p-value	Bonferroni	p
39	-3.201895		0.0015444	0.3892

Enota 39 je potencialni regresijski osamelec, vendar pa je njena popravljena Bonferroni p -vrednost večja od 0.05. Poglejmo si, kako se dejanska vrednost `siri` pri tej enoti razlikuje od na podlagi modela napovedane vrednosti.

```
bodyfat[39, ]
```

siri	weight	height	abdomen	
39	35.2	164.8701	183.515	148.1

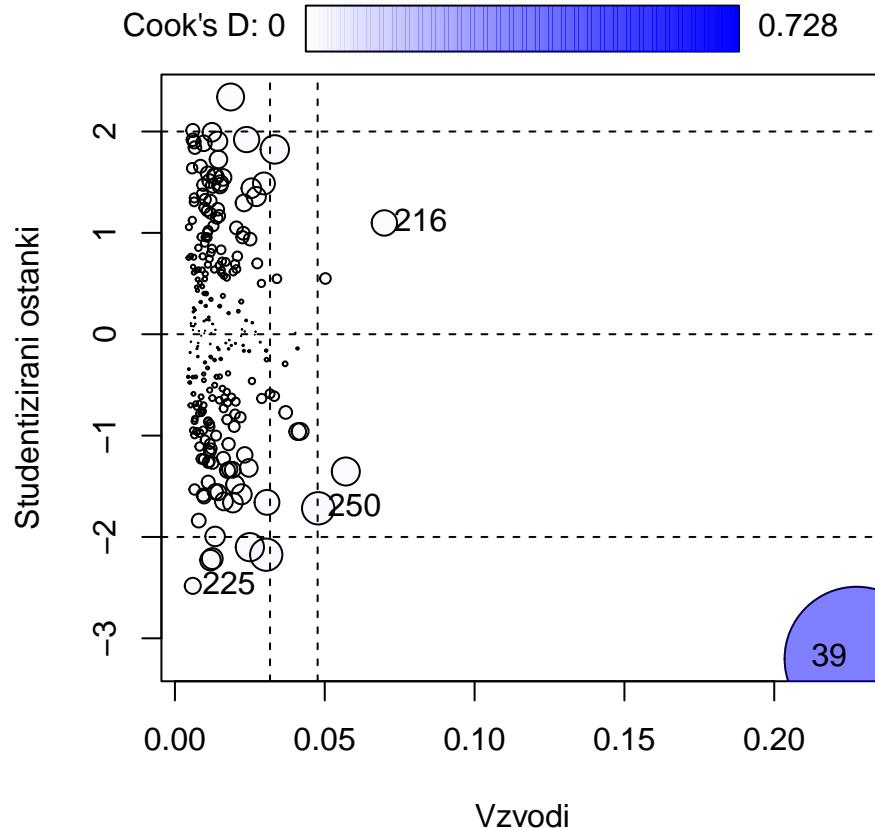
```
fitted(m.bodyfat)[39]
```

```
39  
47.52723
```

Vidimo, da za dano enoto močno precenimo odstotek telesne maščobe. Glede na maso ter obseg abdomna bi na podlagi modela za to enoto pričakovali višji odstotek telesne maščobe.

Poglejmo, ali so v modelu tudi točke, ki imajo velik vzvod:

```
influencePlot(m.bodyfat,
               id = list(method = "noteworthy", n = 2, cex = 1, location = "lr"),
               xlab = "Vzvodi", ylab = "Studentizirani ostanki")
```



Slika 3: Grafični prikaz studentiziranih ostankov, vzvodov in Cookove razdalje (ploščina kroga je sorazmerna Cookovi razdalji) za model `m.bodyfat`.

	StudRes	Hat	CookD
39	-3.201895	0.227490261	0.727621709
216	1.097903	0.069895940	0.022627093
225	-2.482470	0.005955063	0.009041502
250	-1.717273	0.047811915	0.036730979

V modelu je nekaj točk, katerih vzvod presega trikratnik povprečnega vzvoda. Vzvodne točke same po sebi še niso problem, če pa so hkrati tudi regresijski osamelci, so pogosto tudi vplivne točke. Problematična je točka 39, ki je tako vzvodna točka kot tudi točka z veliko vrednostjo studentiziranega ostanka. Čeprav test za regresijske osamelce ni dal značilnega rezultata (kar je lahko tudi posledica konzervativnosti Bonferronijevega popravka), iz izpisa vidimo, da je Cookova razdalja pri tej enoti večja od 0.5. Torej bi točka lahko bila vplivna.

Primerjajmo ocene parametrov modelov z in brez enote 39.

```
m.bodyfat_brez39 <- lm(siri~weight + height + abdomen, bodyfat[-39, ])
summary(m.bodyfat_brez39)
```

Call:
`lm(formula = siri ~ weight + height + abdomen, data = bodyfat[-39,])`

Residuals:

Min	1Q	Median	3Q	Max
-11.101	-3.309	0.023	3.230	10.050

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-31.07071	11.54742	-2.691	0.00762 **
weight	-0.21900	0.06822	-3.210	0.00150 **
height	-0.09202	0.06234	-1.476	0.14120
abdomen	0.91304	0.07171	12.732	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.38 on 247 degrees of freedom
Multiple R-squared: 0.7264, Adjusted R-squared: 0.7231
F-statistic: 218.6 on 3 and 247 DF, p-value: < 2.2e-16

```
compareCoefs(m.bodyfat, m.bodyfat_brez39)
```

Calls:

```
1: lm(formula = siri ~ weight + height + abdomen, data = bodyfat)
2: lm(formula = siri ~ weight + height + abdomen, data = bodyfat[-39, ])
```

	Model 1	Model 2
(Intercept)	-39.2	-31.1
SE	11.5	11.5
weight	-0.2984	-0.2190
SE	0.0647	0.0682
height	-0.0369	-0.0920
SE	0.0610	0.0623
abdomen	0.9635	0.9130
SE	0.0713	0.0717

Primerjajmo grafično pričakovani odstotek telesne maščobe v odvisnosti od `weight` oz. `abdomen` ob upoštevanju ostalih spremenljivk na podlagi obeh modelov:

```
# Napovedi glede na weight
effect_weight_m1 <- as.data.frame(Effect("weight", m.bodyfat))
effect_weight_m2 <- as.data.frame(Effect("weight", m.bodyfat_brez39))

effect_weight_m1$Model <- "m.bodyfat"
effect_weight_m2$Model <- "m.bodyfat_brez39"
```

```

effect_weight <- rbind(effect_weight_m1, effect_weight_m2)

p1 <- ggplot(effect_weight, aes(x = weight, y = fit, color = Model, fill = Model)) +
  geom_line(size = 1.2) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2) +
  labs(
    x = "Masa (kg)",
    y = "Pričakovani odstotek telesne maščobe"
  ) +
  theme_minimal() +
  scale_color_manual(values = c("m.bodyfat" = "red", "m.bodyfat_brez39" = "blue")) +
  scale_fill_manual(values = c("m.bodyfat" = "red", "m.bodyfat_brez39" = "blue"))

# Napovedi glede na abdomen
effect_abdomen_m1 <- as.data.frame(Effect("abdomen", m.bodyfat))
effect_abdomen_m2 <- as.data.frame(Effect("abdomen", m.bodyfat_brez39))

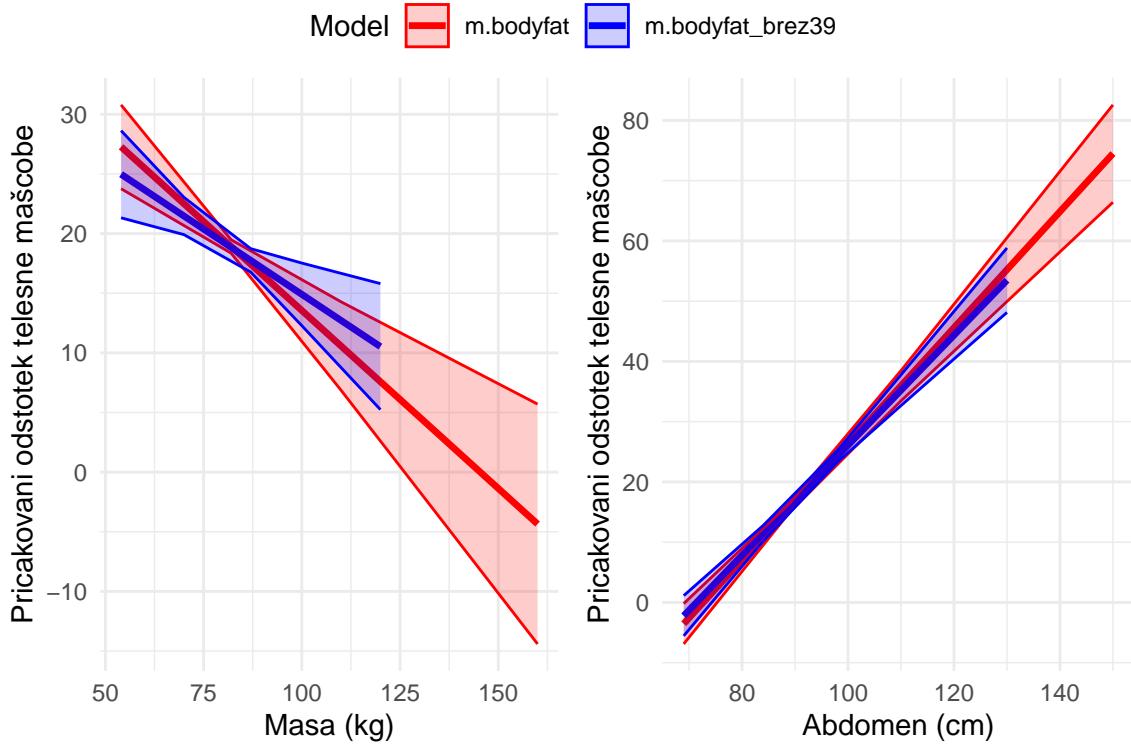
effect_abdomen_m1$Model <- "m.bodyfat"
effect_abdomen_m2$Model <- "m.bodyfat_brez39"

effect_abdomen <- rbind(effect_abdomen_m1, effect_abdomen_m2)

p2 <- ggplot(effect_abdomen, aes(x = abdomen, y = fit, color = Model, fill = Model)) +
  geom_line(size = 1.2) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2) +
  labs(
    x = "Abdomeen (cm)",
    y = "Pričakovani odstotek telesne maščobe"
  ) +
  theme_minimal() +
  scale_color_manual(values = c("m.bodyfat" = "red", "m.bodyfat_brez39" = "blue")) +
  scale_fill_manual(values = c("m.bodyfat" = "red", "m.bodyfat_brez39" = "blue"))

ggarrange(p1, p2, ncol = 2, common.legend = TRUE, legend="top")

```



Slika 4: Napovedi za odstotek telesne maščobe v odvisnosti od `weight` oz. `abdomen` ter ob upoštevanju ostalih spremenljivk na podlagi modelov `m.bodyfat` ter `m.bodyfat_brez39`.

Vidimo, da je v modelu brez enote 39 povezanost med maso in odzivno spremeljivko `siri` ob upoštevanju ostalih spremenljivk šibkejša.

Brez dobrega poznavanja stroke pa ne moremo reči, kateri model je primernejši oz. ustrezejši; zato vplivnih točk ne izločamo kar tako iz modela. Pomembno je, da jih identificiramo!

3. Logaritmska transformacija

Kadar aditivnost in linearost postaneta vprašljivi, včasih situacijo lahko popravimo z nelinearno transformacijo. Če so vrednosti odzivne spremenljivke izključno pozitivne, je vsebinsko smiselno uporabiti logaritemsko transformacijo odzivne spremenljivke. Tak linearni model postane multiplikativnen na originalni skali y_i :

$$\log y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + \epsilon_i.$$

Z inverzno transformacijo dobimo

$$y_i = e^{b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + \epsilon_i} = B_0 B_1^{x_{i1}} B_2^{x_{i2}} \dots E_i,$$

kjer so $B_0 = e^{b_0}$, $B_1 = e^{b_1}$, $B_2 = e^{b_2}, \dots$ eksponencirani regresijski parametri, $E_0 = e^{\epsilon_i}$ pa je eksponencirana napaka. Ker je $\exp(a) > 0$, so napovedi, ki jih da tak model vedno pozitivne.

```
?Wage
head(Wage)
```

Podatkovni okvir `Wage` iz paketa `ISLR2` vsebuje podatke o bruto letnih zaslužkih (1000 \$) za 3000 moških iz srednje-atlantske regije. Osredotočili se bomo na opisovanje povezanosti med plačo in starostjo ter izobrazbo. Tovrstnega modela ne bi mogli uporabiti za razlaganje vzročno-posledičnih zvez med odzivno in napovednima spremenljivkama, saj se npr. mlajši in starejši moški v vzorcu razlikujejo še v marsikateri drugi lastnosti

kot pa le po izobrazbi. Model nam lahko služi za raziskovanje *povezanosti* med zaslužkom ter starostjo in izobrazbo, gre torej za deskriptivni model. Lahko pa bi ga uporabili tudi za napovedovanje bruto letnih zaslužkov za nove enote, čeprav bi za bolj natančne napovedi imelo smisel v modelu modelirati nelinearnost.

Katero metodo ter tudi strategijo modeliranja uporabimo, je odvisno od tega, ali je naš končni cilj napovedovanje, vzročno-posledično sklepanje (inferenca) ali kombinacija obeh. Linearni model omogoča razmeroma preprosto in razumljivo interpretacijo parametrov, vendar pa bo morda dal manj natančne napovedi kot nekateri drugi nelinearni pristopi. Nasprotно pa ti pristopi lahko dajo natančnejše napovedi odzivne spremenljivke, vendar na račun fleksibilnosti dobimo manj razumljiv model, na podlagi katerega je sklepanje lahko zelo zahtevno.

```
Wage <- Wage %>%
  dplyr::select(age, education, wage) #izberemo spremenljivke, ki nas zanimajo
```

```
str(Wage)
```

```
'data.frame': 3000 obs. of 3 variables:
 $ age      : int 18 24 45 43 50 54 44 30 41 52 ...
 $ education: Factor w/ 5 levels "1. < HS Grad",...: 1 4 3 4 2 4 3 3 3 2 ...
 $ wage     : num 75 70.5 131 154.7 75 ...
```

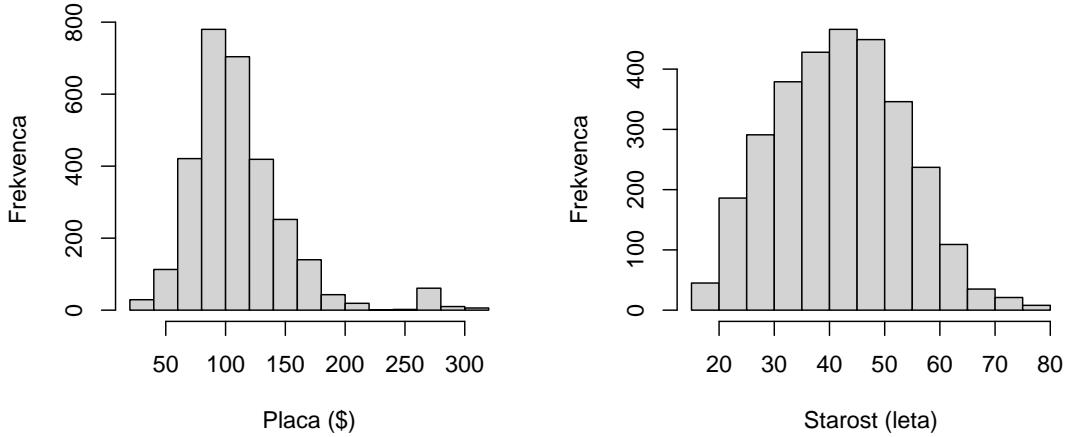
Poglejmo najprej univariatne porazdelitve spremenljivk v podatkovnem okviru, čeprav regresijski model ne predpostavlja ničesar o porazdelitvi napovednih spremenljivk. Te so lahko porazdeljene po porazdelitvah, ki so daleč od normalnosti. Model prav tako ne predpostavlja, da mora biti odzivna spremenljivka normalna, temveč se ta predpostavka nanaša na porazdelitev napak.

```
summary(Wage)
```

age	education	wage
Min. :18.00	1. < HS Grad :268	Min. : 20.09
1st Qu.:33.75	2. HS Grad :971	1st Qu.: 85.38
Median :42.00	3. Some College :650	Median :104.92
Mean :42.41	4. College Grad :685	Mean :111.70
3rd Qu.:51.00	5. Advanced Degree:426	3rd Qu.:128.68
Max. :80.00		Max. :318.34

```
par(mfrow=c(1,2))
hist(Wage$wage, main="",
  xlab="Plača ($)", ylab="Frekvenca",
  breaks=20)
```

```
hist(Wage$age, main="",
  xlab="Starost (leta)", ylab="Frekvenca")
```



Slika 5: Univariatne porazdelitve številskih spremenljivk v podatkovnem okviru `Wage`.

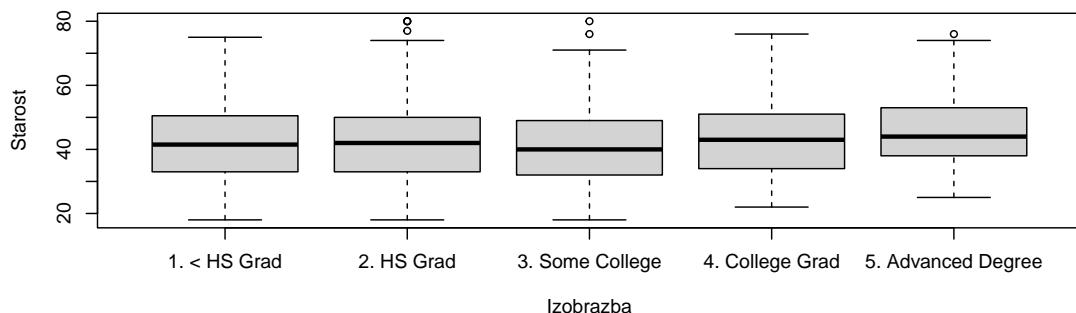
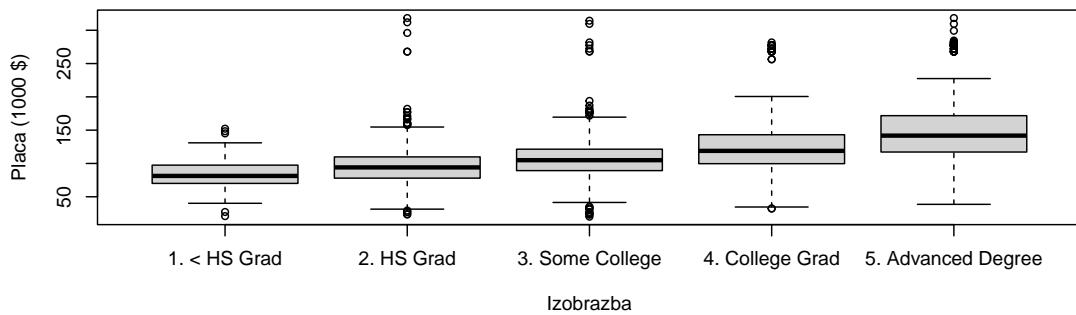
Porazdelitev odzivne spremenljivke `wage` je rahlo asimetrična v desno, vrednosti `wage` pa so izključno pozitivne.

Prikažimo še porazdelitev `wage` in `age` glede na `education`:

```
par(mfrow=c(2, 1))
par(cex=0.8)

boxplot(Wage$wage ~ Wage$education,
        xlab = "Izobrazba", ylab = "Plača (1000 $)")

boxplot(Wage$age ~ Wage$education,
        xlab = "Izobrazba", ylab = "Starost")
```



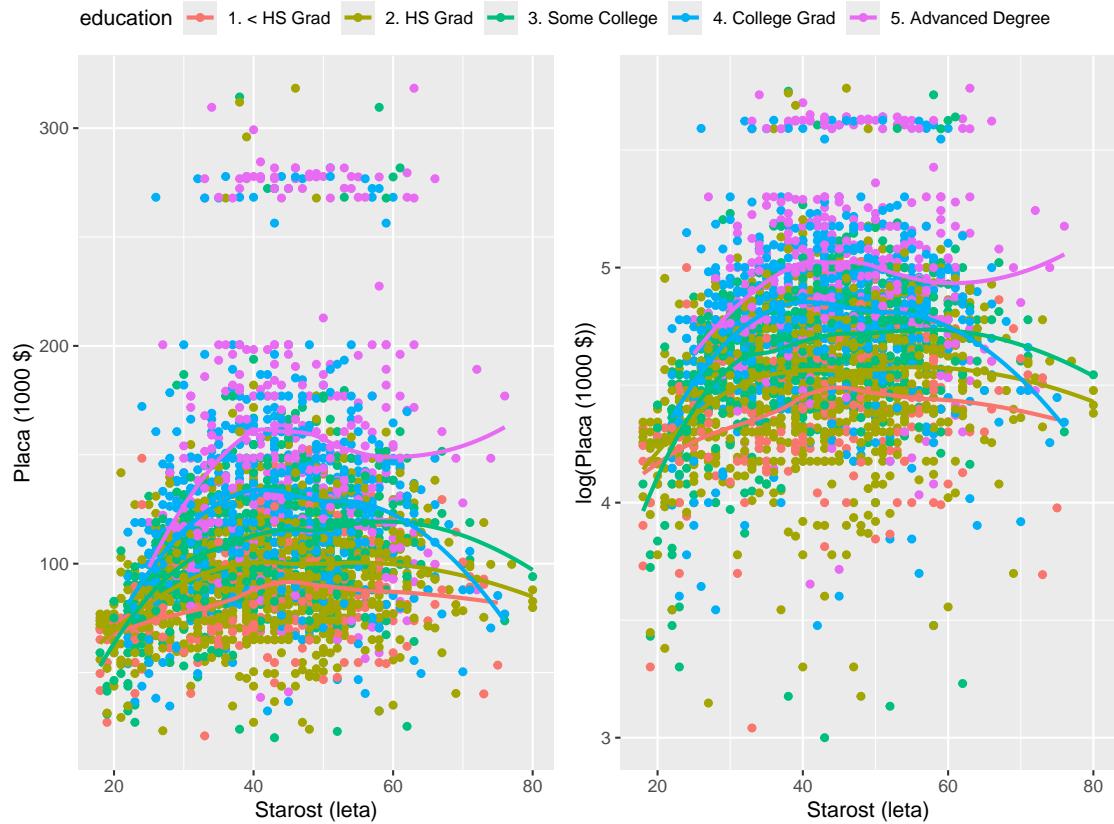
Slika 6: Porazdelitev wage in age glede na education v podatkovnem okviru Wage.

Poglejmo grafično, kako izgleda odvisnost wage od age po education:

```
#Ali obstaja linearna povezanost med spremenljivkama age in wage (glede na izobrazbo)?
p1 <- ggplot(data=Wage, aes(x=age, y=wage, col=education)) +
  geom_point() +
  geom_smooth(se=FALSE) +
  #geom_smooth(method="lm", se=FALSE) +
  xlab("Starost (leta)") +
  ylab("Plača (1000 $)")

p2 <- ggplot(data=Wage, aes(x=age, y=log(wage), col=education)) +
  geom_point() +
  geom_smooth(se=FALSE) +
  #geom_smooth(method="lm", se=FALSE) +
  xlab("Starost (leta)") +
  ylab("log(Plača (1000 $))")

ggarrange(p1, p2, ncol = 2, common.legend = TRUE, legend="top")
```



Slika 7: Odvisnost `wage` oz. `log(wage)` od `age` in `education` v podatkovnem okviru `Wage`.

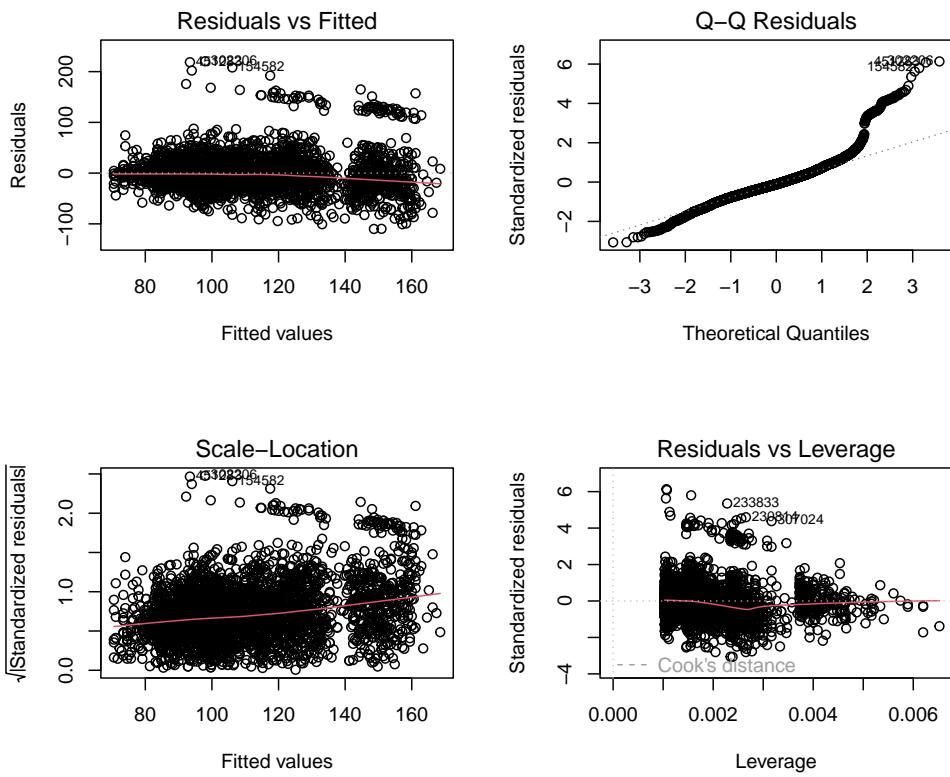
Graf nakazuje, da je vpliv `age` nelinearen in drugačen glede na `education`.

Naredimo prvi model, v katerem predpostavimo linearnost zveze med `wage` in `age` ter aditivnost vplivov `age` in `education`:

```
m0 <- lm(wage ~ age + education, data=Wage)
```

Osnovni diagnostični grafi ostankov za model `m0`:

```
par(mfrow=c(2,2))
plot(m0)
```



Slika 8: Ostanki za model `m0`.

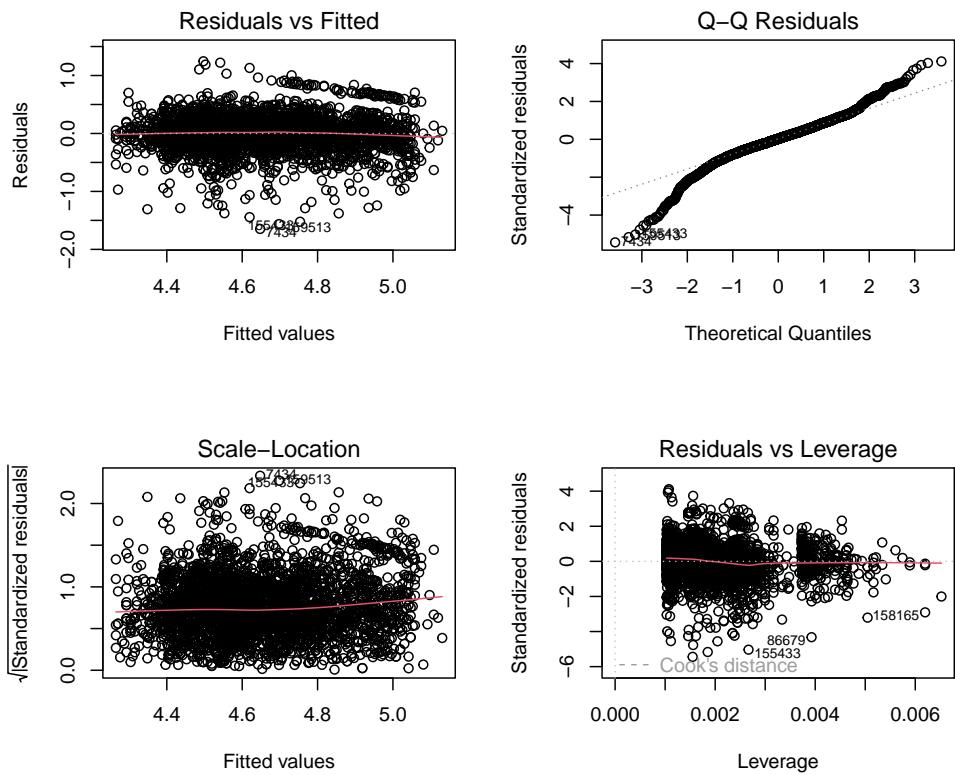
Levi sličici v prvi in drugi vrstici kažeta nekonstantno varianco. Varianca ostankov narašča z napovedanimi vrednostmi (zgornja leva sličica), slika ostankov je podobna klinu: variabilnost ostankov narašča od leve proti desni. Prisotnost nekonstantne variance še bolje pokaže gladilnik na levi spodnji sliki, kjer so na vodoravnih osi napovedane vrednosti, na navpični osi pa korenji absolutnih vrednosti standardiziranih ostankov. Kot smo lahko pričakovali glede na podatke, vidimo tudi, da precej enot izstopa tudi zaradi velikih vrednosti standardiziranih ostankov.

Poglejmo, kako situacija izgleda, če odzivno spremenljivko `wage` logaritmiramo.

```
m1 <- lm(log(wage) ~ age + education, data=Wage)
```

Osnovni diagnostični grafi ostankov za model `m1`:

```
par(mfrow=c(2,2))
plot(m1)
```



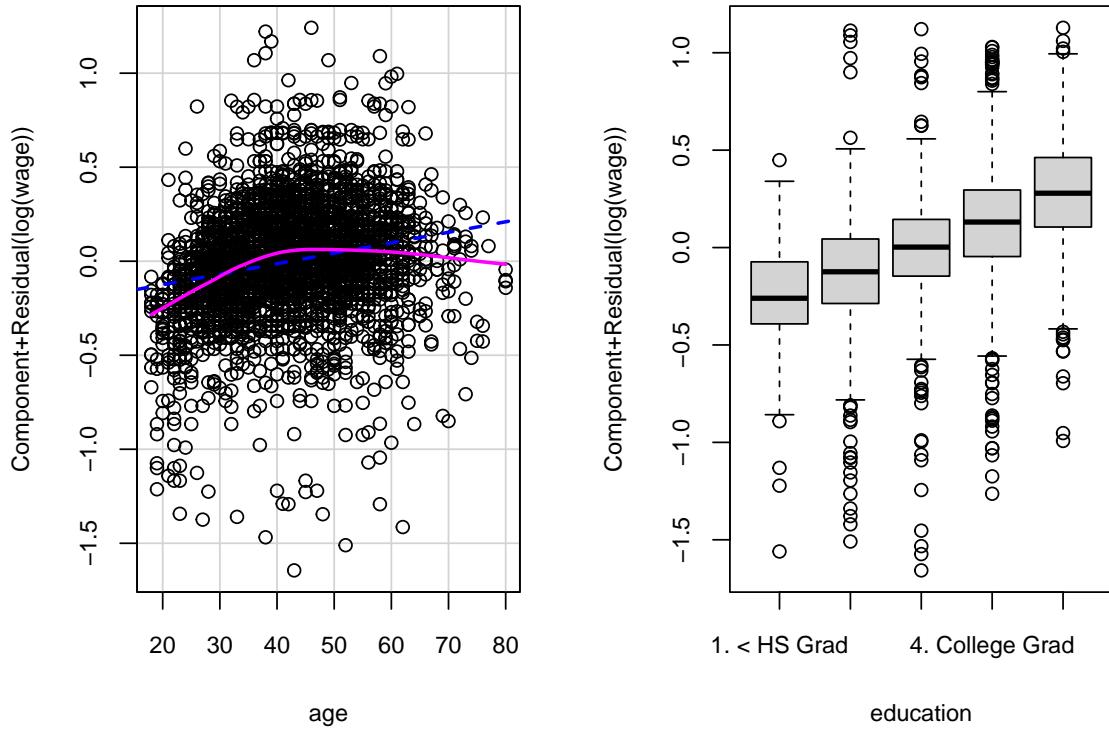
Slika 9: Ostanki za model m1.

Tretji grafikon nakazuje, da smo z z logaritemsko transformacijo odzivne spremenljivke heteroskedastičnost odpravili. Kot je pričakovano, pa je v podatkih še vedno precej enot z veliko vrednostjo standardiziranega ostanka. Porazdelitev le-teh tudi odstopa od standardizirane normalne porazdelitve.

Za boljšo diagnostiko modela si poglejmo grafikon parcialnih ostankov:

```
crPlots(m1, cex.lab=0.8, cex.axis=0.8)
```

Component + Residual Plots



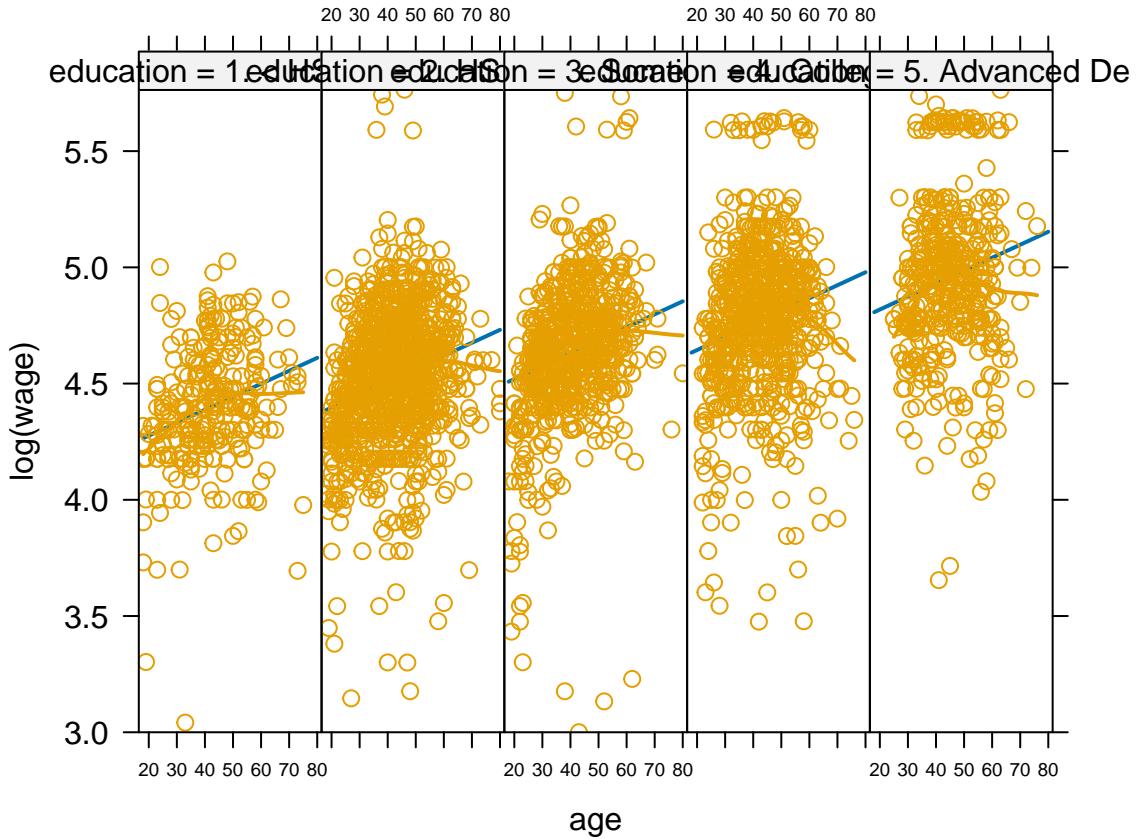
Slika 10: Graf parcialnih ostankov za model `m1`.

Iz prve sličice vidimo, da se gladilnik ne prilega dobro premici: zveza med `log(wage)` in `age` je ob upoštevanju `education` nelinearna.

Poglejmo si še grafikon parcialnih ostankov za model `m1` v odvisnosti od `age` pri različnih vrednostih spremenljivke `education`.

```
plot(Effect(c("age", "education"), m1, partial.residuals = TRUE),
     ci.style = "none",
     lattice = list(layout = c(5, 1)),
     axes = list(x=list(cex=0.6)))
```

age*education effect plot



Slika 11: Graf parcialnih ostankov za model m1.

Predvsem je očitna nelinearnost zveze med `log(wage)` in `age`. Pri modeliranju nelinearnosti bi si lahko pomagali s polinomsko regresijo ali zlepki, vendar več o tem kasneje.

Kljub temu, da se naš model podatkom ne prilega najbolje, bomo za vajo vseeno razmislili o interpretaciji modela, ki ima eno številsko in eno opisno spremenljivko z večimi kategorijami, odzivna spremenljivka pa je logaritmizirana.

```
summary(m1)
```

```
Call:  
lm(formula = log(wage) ~ age + education, data = Wage)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.64758	-0.15373	0.00796	0.17330	1.24577

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.1645351	0.0273609	152.208	< 2e-16 ***
age	0.0055762	0.0004821	11.566	< 2e-16 ***
education2. HS Grad	0.1205914	0.0209082	5.768	8.86e-09 ***

```

education3. Some College    0.2432633  0.0219999  11.057  < 2e-16 ***
education4. College Grad   0.3682739  0.0218360  16.865  < 2e-16 ***
education5. Advanced Degree 0.5424496  0.0236742  22.913  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 0.303 on 2994 degrees of freedom
 Multiple R-squared: 0.2592, Adjusted R-squared: 0.258
 F-statistic: 209.6 on 5 and 2994 DF, p-value: < 2.2e-16

```
confint(m1)
```

	2.5 %	97.5 %
(Intercept)	4.110887058	4.218183125
age	0.004630888	0.006521597
education2. HS Grad	0.079595544	0.161587233
education3. Some College	0.200126939	0.286399695
education4. College Grad	0.325458948	0.411088924
education5. Advanced Degree	0.496030185	0.588868967

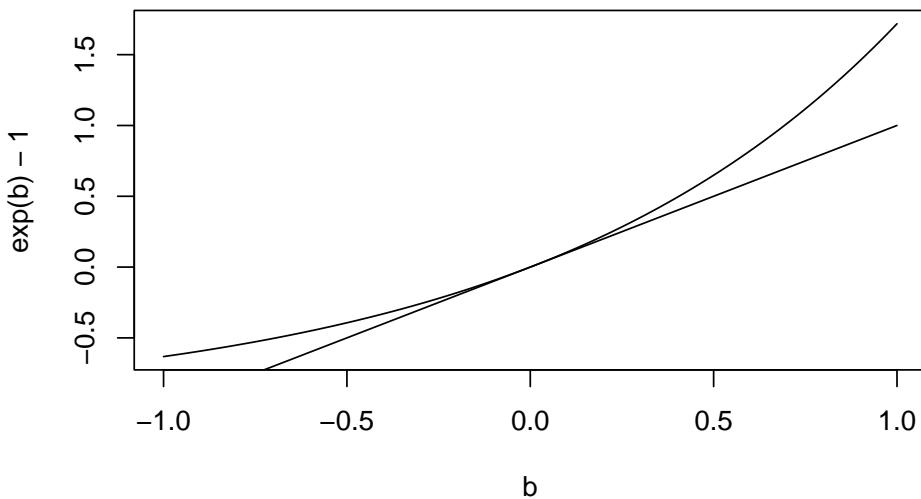
Ob upoštevanju `education` log(`wage`) narašča z `age` ($p < 0.001$), in sicer z vsakim letom starosti se v povprečju log(`wage`) poveča za 0.006 enote oz. se `wage` poveča za 0.6 % ($\exp(\beta_{age}) = 1.0055918$), pripadajoči 95 % IZ je (0.5, 0.7).

V primerjavi z osebo s < HS Grad in enako vrednostjo spremenljivke `age` ima v povprečju oseba s stopnjo izobrazbe:

- 2. HS Grad log(`wage`) višjo za 0.12 enote oz. `wage` višjo za 12.8 % ($\exp(\beta_{HSGrad}) = 1.1281638$), pripadajoči 95 % IZ je (8.3, 17.5);
- 3. Some College log(`wage`) višjo za 0.24 enote oz. `wage` višjo za 27.5 % ($\exp(\beta_{SomeCollege}) = 1.2754044$), pripadajoči 95 % IZ je (22.2, 33.2);
- 4. College Grad log(`wage`) višjo za 0.37 enote oz. `wage` višjo za 44.5 % ($\exp(\beta_{CollegeGrad}) = 1.4452379$), pripadajoči 95 % IZ je (38.5, 50.8),
- 5. Advanced Degree log(`wage`) višjo za 0.54 enote oz. `wage` višjo za 72 % ($\exp(\beta_{AdvancedDegree}) = 1.7202155$), pripadajoči 95 % IZ je (64.2, 80.2).

Z modelom m1 smo uspeli pojasniti 25.92 % variabilnosti log(`wage`).

V modelu, v katerem je odzivna spremenljivka logaritmirana, so ocene regresijskih parametrov običajno majhne. Kot prikazuje spodnjia slika, za majhne vrednosti približek $\exp(x) = 1 + x$ dobro aproksimira relativno razliko.

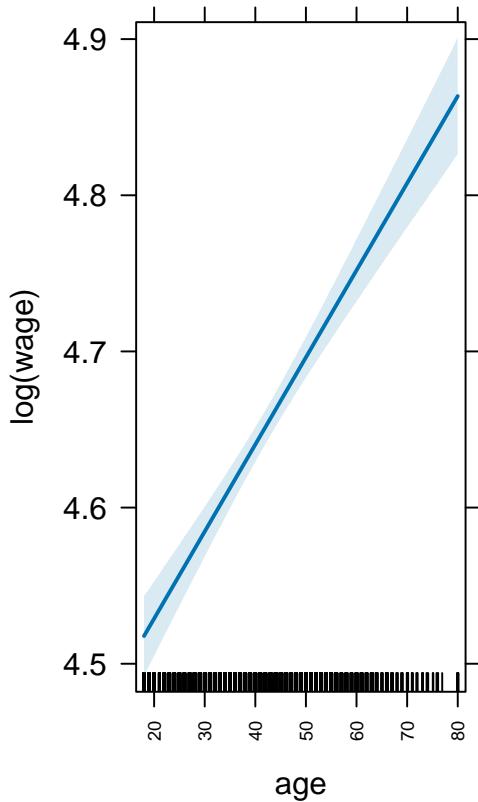


Slika 12: Interpretacija eksponenciranih regresijskih parametrov v regresijskem modelu z logaritmirano odzivno spremenljivko kot relativne razlike (ukriviljena zgornja črta) in približek $\exp(x) = 1 + x$, ki velja za majhne koeficiente x .

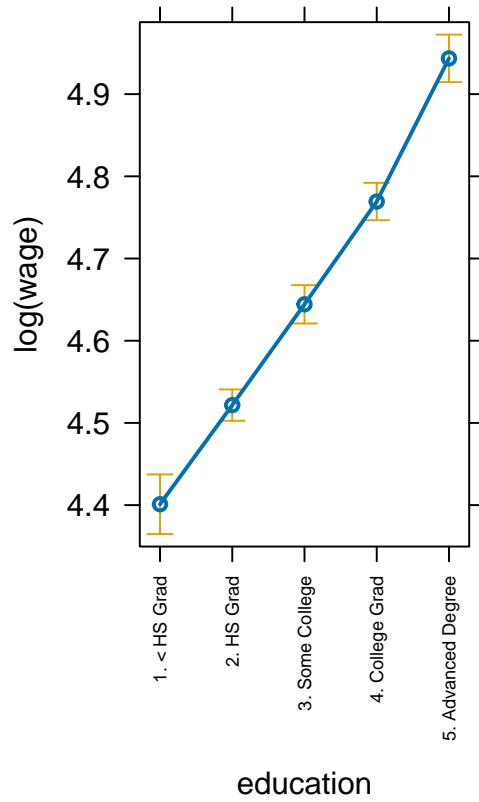
Pri interpretaciji si lahko pomagamo z grafičnimi prikazi iz paketa **effects**:

```
plot(predictorEffects(m1, ~age + education),
     axes = list(x=list(cex=0.6, rotate=90)))
```

age predictor effect plot



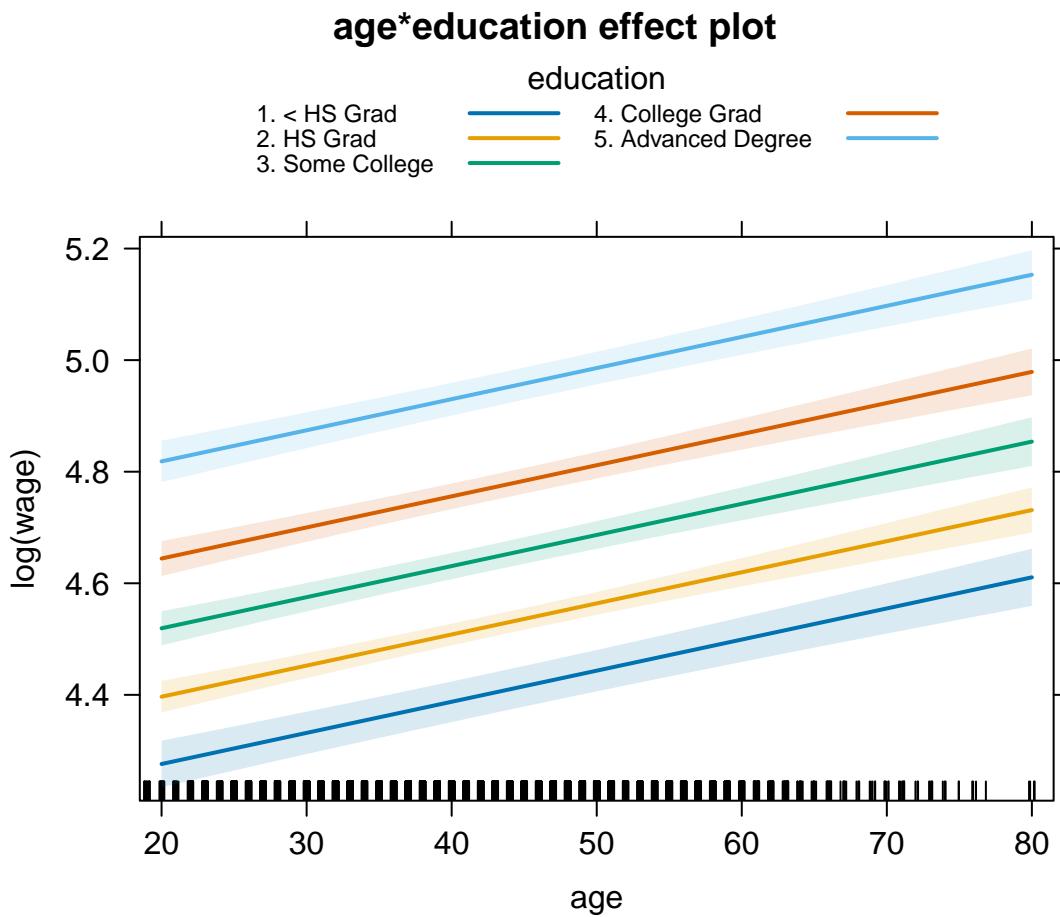
education predictor effect plot



Slika 13: Povprečne napovedi za `log(wage)` na podlagi modela `m1`, ki jih vrne funkcija `predictorEffects` za spremenljivki `age` in `education`.

oziroma:

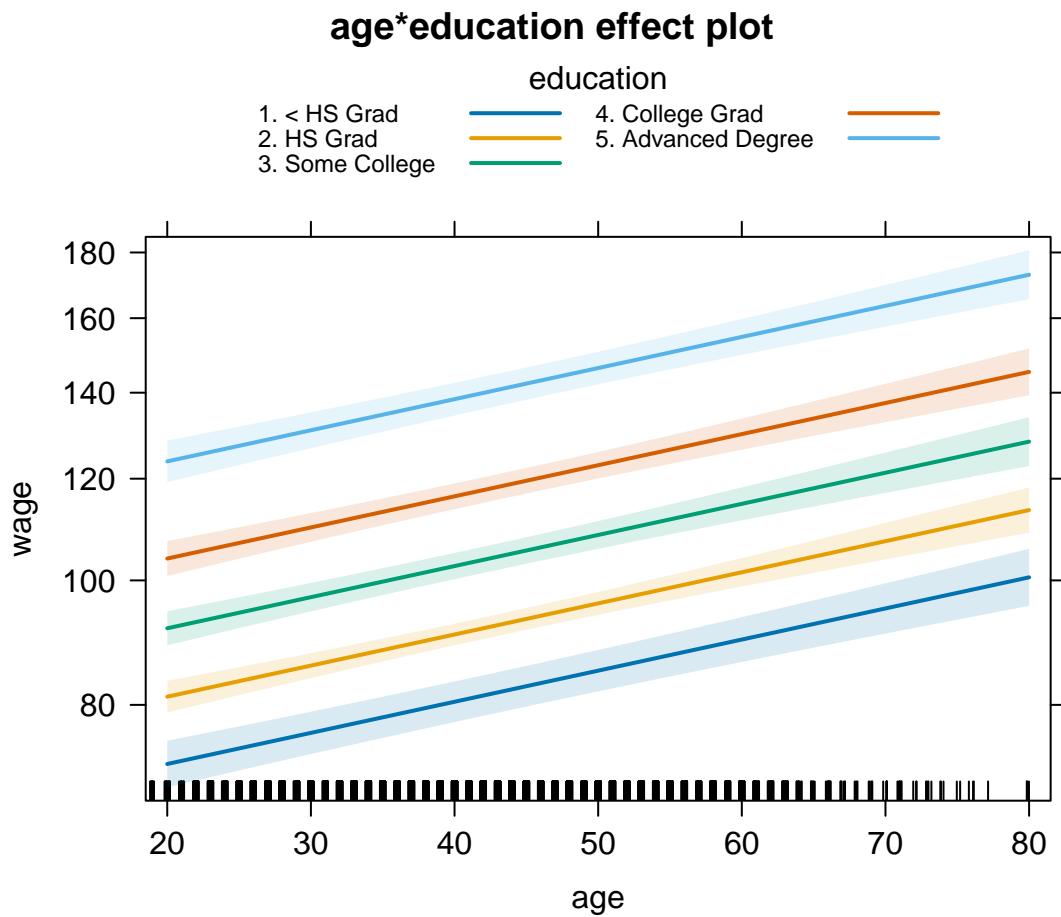
```
plot(Effect(c("age", "education"), m1), multiline=TRUE, ci.style = "bands")
```



Slika 14: Povprečne napovedi za `log(wage)` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

oz. na originalni skali spremenljivke `wage`:

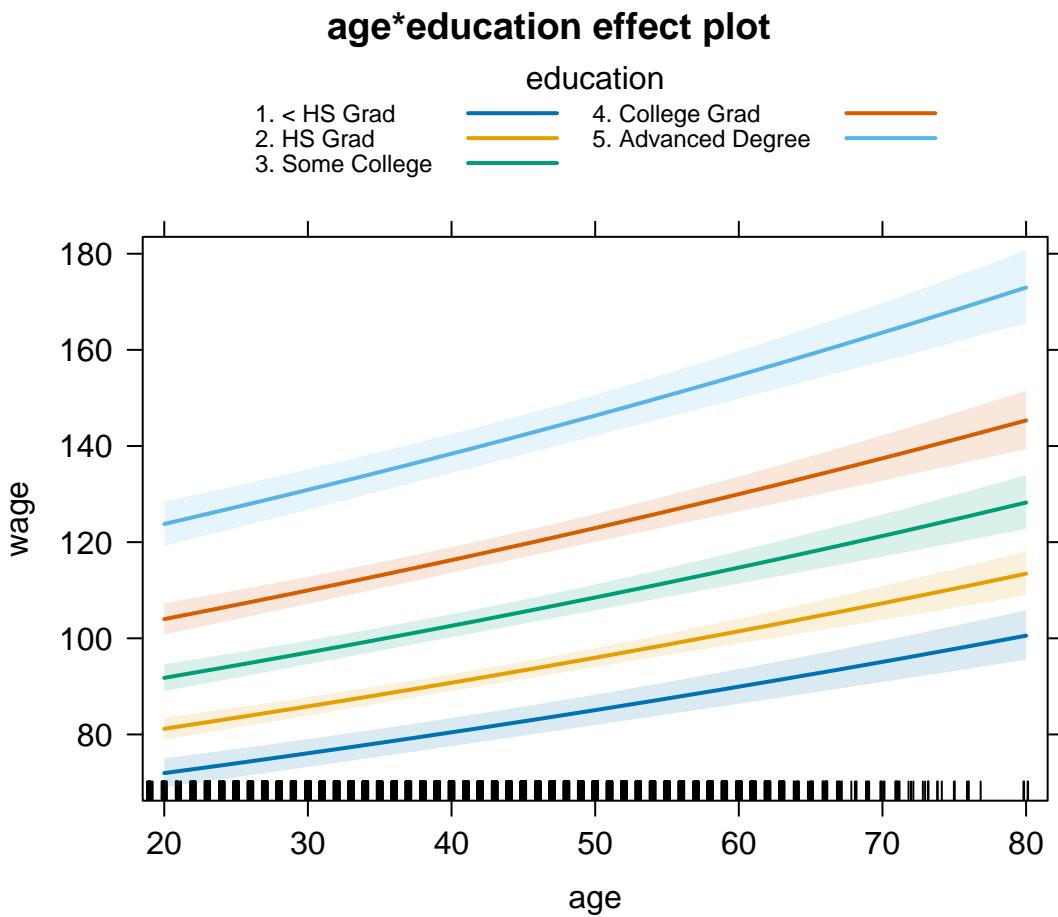
```
plot(Effect(c("age", "education"), m1, transformation = list(link = log, inverse = exp)),
  multiline=TRUE,
  ci.style = "bands",
  axes=list(y=list(lab="wage")))
```



Slika 15: Povprečne napovedi za `wage` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

ali:

```
plot(Effect(c("age", "education"), m1),
     multiline=TRUE,
     ci.style = "bands",
     axes=list(y=list(transform=exp, lab="wage")))
```



Slika 16: Povprečne napovedi za `wage` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

Poglejmo še, kako bi interpretirali model, ki vključuje tudi interakcijo med `age` in `education`.

```
m2 <- lm(log(wage) ~ age*education, data=Wage)
```

```
anova(m1, m2)
```

Analysis of Variance Table

Model 1: `log(wage) ~ age + education`

Model 2: `log(wage) ~ age * education`

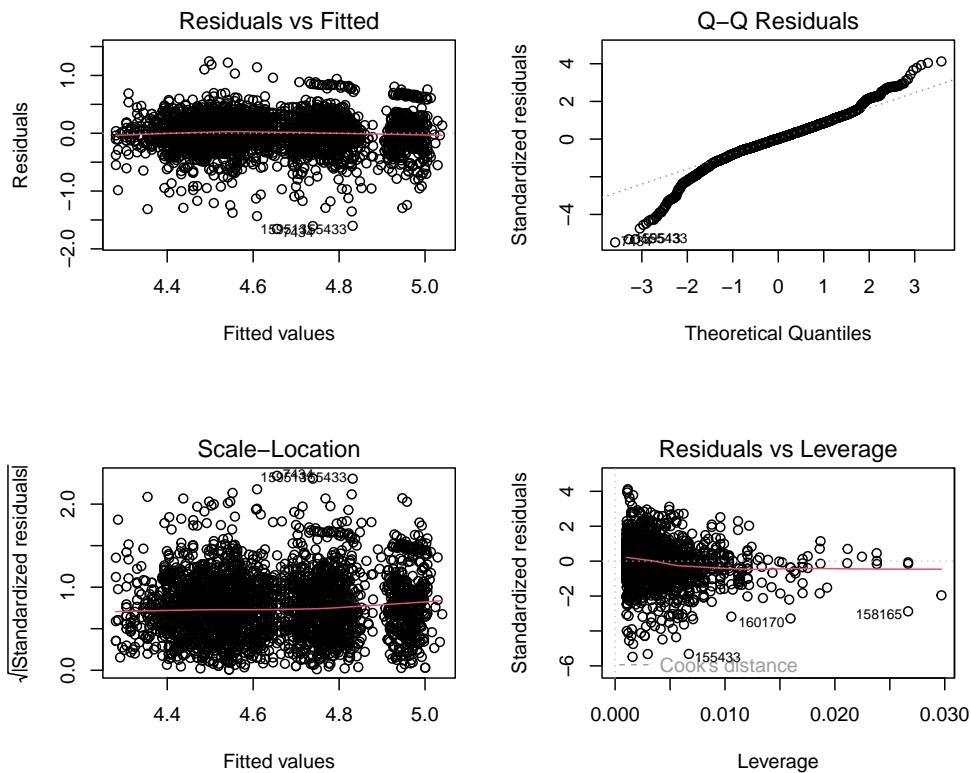
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2994	274.87				
2	2990	273.03	4	1.8363	5.0273	0.0004885 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F-test kaže na to, da interakcija med `age` in `education` izboljša prileganje modela. Poglejmo še osnovne diagnostične grafe ostankov za model `m2`:

```
par(mfrow=c(2,2))
```

```
plot(m2)
```



Slika 17: Ostanki za model m2.

Interpretacija modela m2:

```
summary(m2)
```

Call:

```
lm(formula = log(wage) ~ age * education, data = Wage)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.65539	-0.15483	0.00691	0.17417	1.24528

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.1921197	0.0640086	65.493	< 2e-16 ***
age	0.0049162	0.0014664	3.353	0.000811 ***
education2. HS Grad	0.0979291	0.0731558	1.339	0.180791
education3. Some College	0.0644316	0.0775180	0.831	0.405937
education4. College Grad	0.4160484	0.0792801	5.248	1.65e-07 ***
education5. Advanced Degree	0.6467308	0.0918866	7.038	2.40e-12 ***
age:education2. HS Grad	0.0005434	0.0016738	0.325	0.745466
age:education3. Some College	0.0043591	0.0017917	2.433	0.015033 *
age:education4. College Grad	-0.0011018	0.0018093	-0.609	0.542593
age:education5. Advanced Degree	-0.0022699	0.0020470	-1.109	0.267563

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3022 on 2990 degrees of freedom
Multiple R-squared: 0.2642, Adjusted R-squared: 0.262
F-statistic: 119.3 on 9 and 2990 DF, p-value: < 2.2e-16

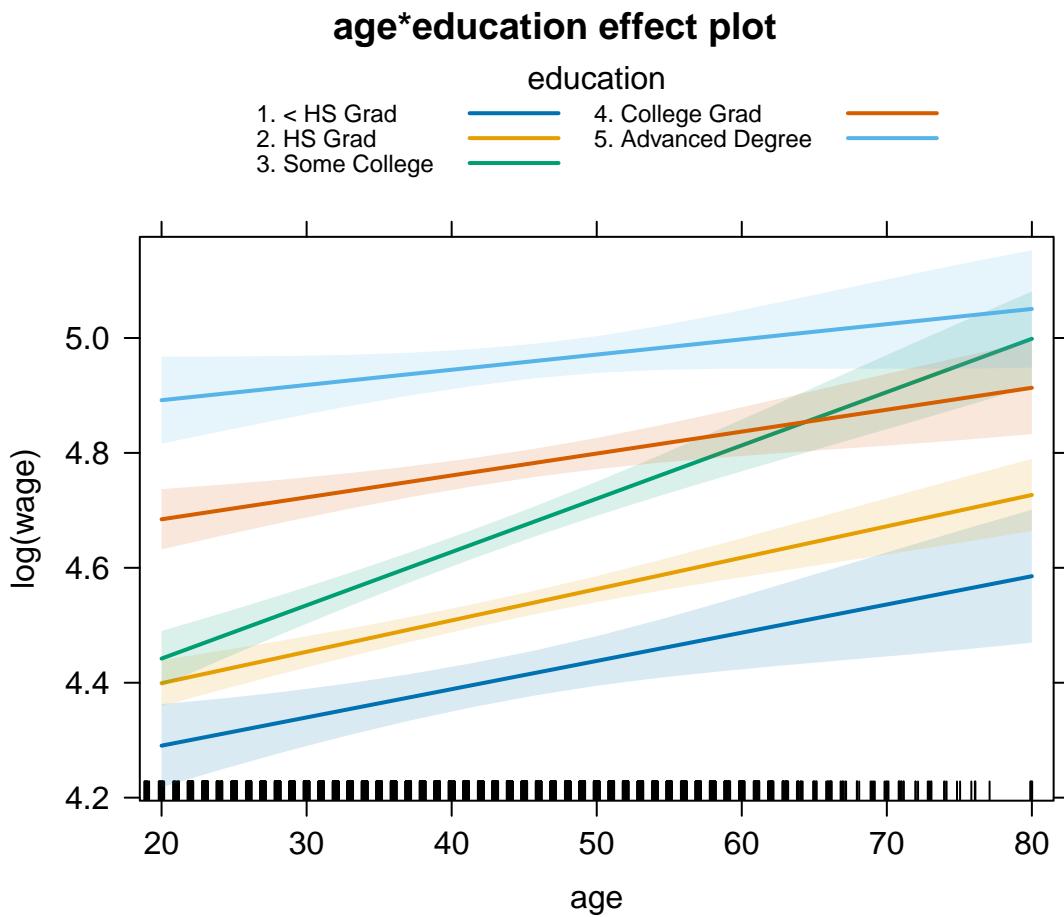
Ob upoštevanju ostalih spremenljivk v modelu je zveza med `log(wage)` in `age` drugačna glede na `education`.

- *Presečišče:* da napoved `log(wage)` za enoto, ki je stara 0 let in ima izobrazbo 1. < HS Grad. Interpretacija v tem primeru ni smiselna, saj nobena enota ni stara 0 let.
- *Koefficient `age`:* nam pove, da `log(wage)` za 1. < HS Grad narašča z `age` ($p = 0.001$), in sicer se z vsakim letom starosti v povprečju `log(wage)` poveča za 0.005 enote oz. se `wage` poveča za 0.5 % ($\exp(\beta_{age}) = 1.0049283$), pripadajoči 95 % IZ je (0.2, 0.8).
- *Koefficienti za `education`:* dajo povprečno razliko `log(wage)` med 1. < HS Grad ter ostalimi stopnjami izobrazbe, če je starost enaka 0. Interpretacija v tem primeru ni smiselna, saj nobena oseba ni stara 0 let.
- *Interakcijski členi:* predstavljajo razlike v naklonih premic, ki napovedujejo `log(wage)` v odvisnosti od `age`, če primerjamo ostale stopnje izobrazbe z 1. < HS Grad. Npr., razlika v starosti enega leta ustreza $e^{\beta_{age:2.HSGrad}} = 0.05$ % večjo razliko v zaslužku pri osebi z izobrazbo 2. HS Grad v primerjavi z osebo 1. < HS Grad, ocenjena napovedana razlika na leto starosti pri osebah z izobrazbo 2. HS Grad pa je $e^{\beta_{age} + \beta_{age:2.HSGrad}} = 0.55$ %.

Model `m2` pojasni 26 % variabilnosti spremenljivke `log(wage)`.

Zveze med `log(wage)` in `age` se v modelu `m2` ne da opisati brez upoštevanja spremenljivke `education` zaradi prisotne interakcije. Če želimo npr. opisati, kako se `log(wage)` spreminja od `age` glede na `education`, si lahko pomagamo z grafičnimi prikazi, ki jih dobimo s funkcijo `predictorEffects` ali `Effect`.

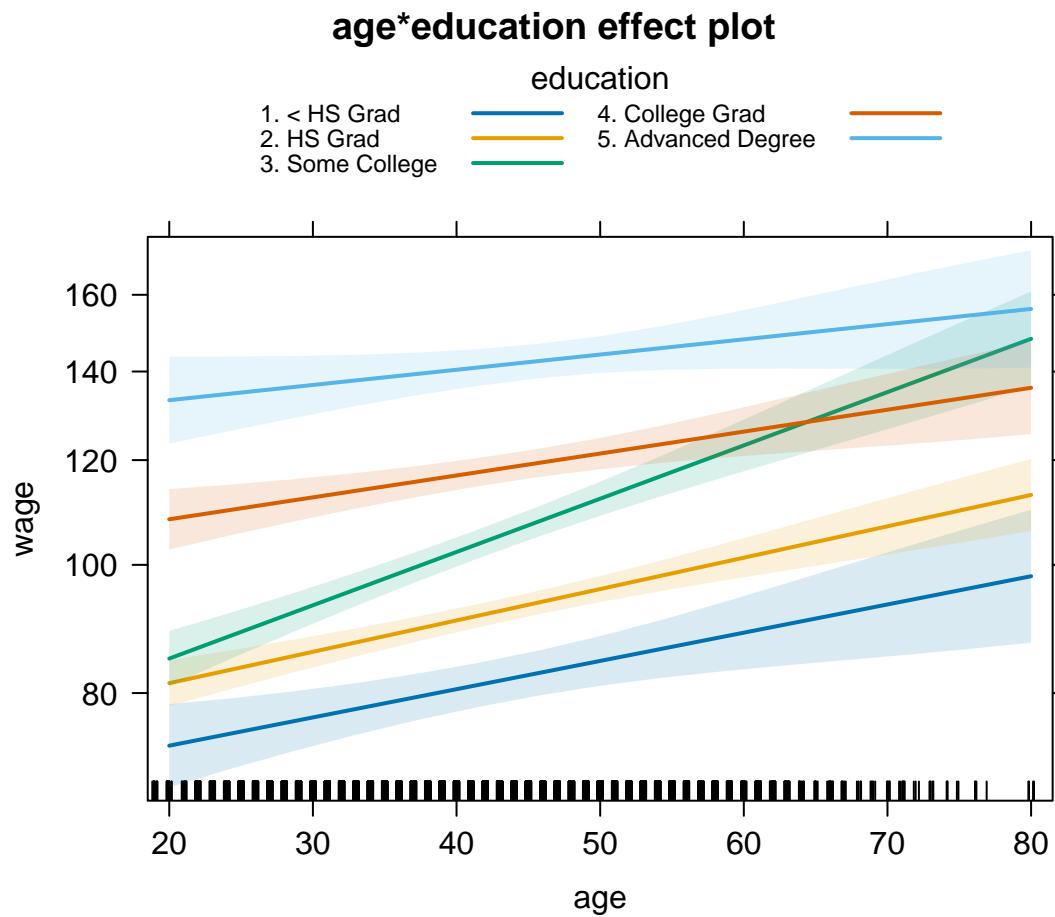
```
plot(Effect(c("age", "education"), m2), multiline=TRUE, ci.style = "bands")
```



Slika 18: Povprečne napovedi za `log(wage)` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

oz. na originalni skali spremenljivke `wage`:

```
plot(Effect(c("age", "education"), m2, transformation = list(link = log, inverse = exp)),
  multiline=TRUE,
  ci.style = "bands",
  axes=list(y=list(lab="wage")))
```



Slika 19: Povprečne napovedi za wage na podlagi modela m1, ki jih vrne funkcija Effect za spremenljivki age in education.

Domača naloga:

1. Interakcija dveh številskih napovednih spremenljivk:

Pripravite funkcijo za generiranje podatkov linearne regresijskega modela:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \epsilon,$$

pri čemer naj bodo vrednosti napovednih spremenljivk porazdeljene:

- $x_{1,i} \sim U(50, 100)$ in
- $x_{2,i} \sim U(50, 100)$.

Generirajte podatke za naslednje parametre:

- $\beta_0 = 10$,
- $\beta_1 = 0.5$,
- $\beta_2 = 0.15$,
- $\beta_3 \in \{0, 0.05, 0.5\}$,
- $\sigma = 10$,
- $n = 100$.

Navodilo:

- Naredite linearni regresijski model, ki ne upošteva interakcije, in preverite, če model izpoljuje predpostavke (pomagajte si z grafi ostankov, grafi dodane spremenljivke in grafi parcialnih ostankov). Zapišite ugotovitve.
- V linearni regresijski model vključite interakcije med napovednima spremenljivkama in preverite, če model izpoljuje predpostavke (pomagajte si z grafi ostankov, grafi dodane spremenljivke in grafi parcialnih ostankov). Zapišite ugotovitve.

2. Log-log model:

V datoteki SNOWGESE.txt so podatki o 45 jatah, za kateri so po dveh različnih metodah (`obs1`, `obs2`) ocenili število gosk v vsaki jati posebej. Za vse jate so zaradi posnete fotografije lahko prešteli dejansko število gosk (`photo`).

Opazujemo dejansko število gosk v jati (`photo`) v odvisnosti od števila gosk v jati, preštetih po drugi metodi (`obs2`).

- Grafično prikažite odvisnost `photo` od `obs2`.
- Ali se vam zdi linearни regresijski model primeren? Naredite linearni regresijski model `model.goske` za prvotne podatke.
- Naredite diagnostiko ostankov modela. Ali so vse predpostavke linearnega modela izpolnjene? Obrazložite svojo trditev.
- Naredite model `model.goske.log`, v katerem logaritmirate odvisno spremenljivko (`photo`) in neodvisno spremenljivko (`obs2`). Naredite diagnostiko ostankov modela. Ali so vse predpostavke linearnega modela izpolnjene? Obrazložite svojo trditev.
- Naredite analizo posebnih točk za `model.goske.log`.
- Interpretirajte ocenjene parametre modela `model.goske.log`, skupaj s pripadajočimi 95% intervali zaupanja in koeficient determinacije.