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‘).

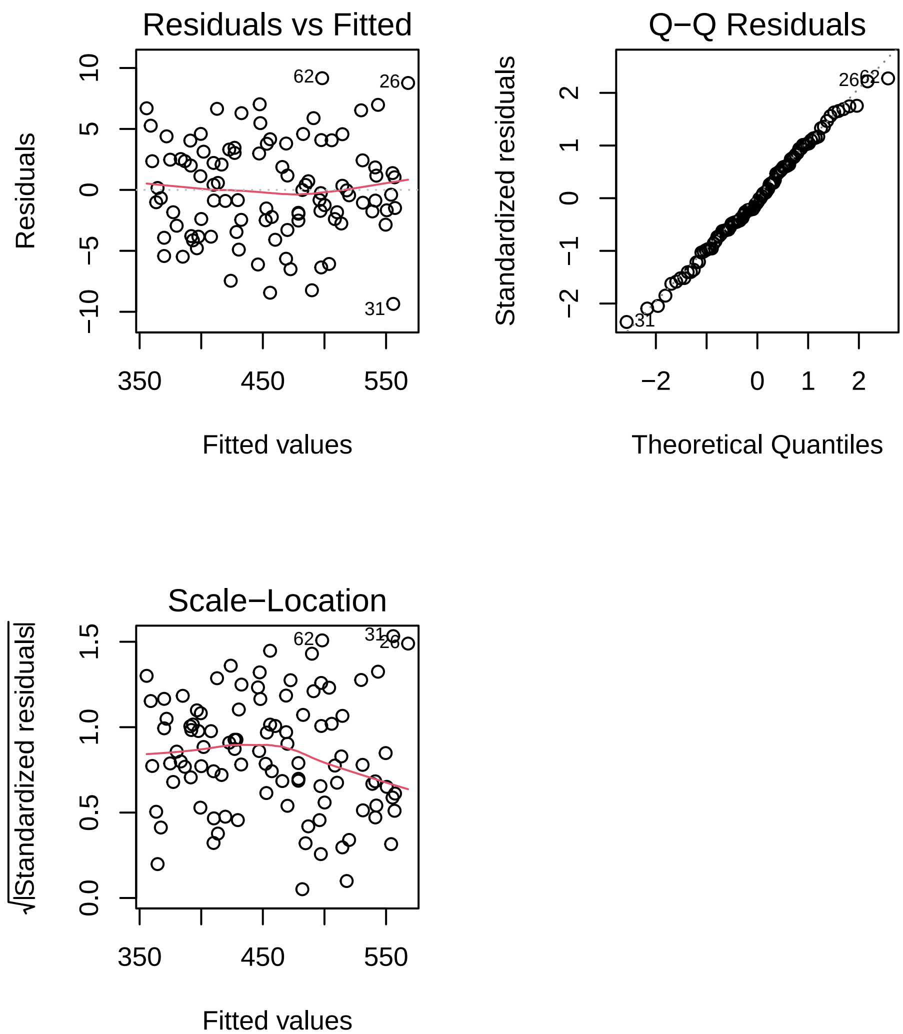
```
f.generiranje.lm.2 <- function(beta0, beta1, beta2, sigma, n) {

  # 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?

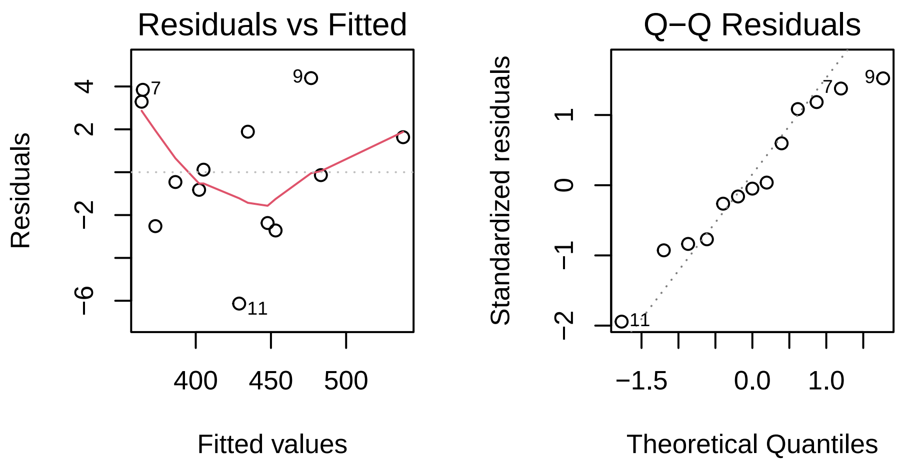
```
# zaženemo funkcijo in rezultate funkcije shranimo v vektor
generirani.podatki <- f.generiranje.lm.2(150, 4, 2.5, 4, 100)
# narišemo ostanke za linearni regresijski model na generiranih podatkih
par(mfrow = c(2,2))
plot(lm(y ~ x1 + x2, data = generirani.podatki), which = c(1:3))
```

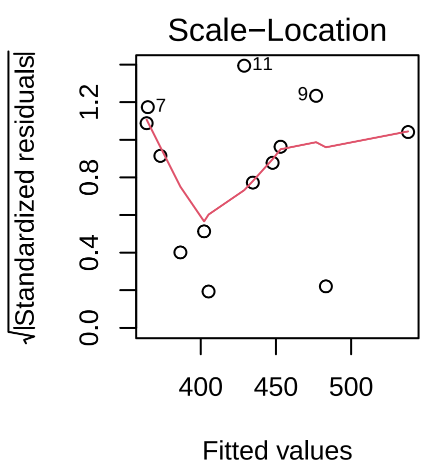


Spremenite velikost vzorca na 13 ($n = 13$) in opazujte, kaj se v primeru manjšega vzorca dogaja z ostanki. Za generiranje uporabite vrednosti semena, ki so zapisane v vektorju `semena`.

```
# 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 regresijki 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)
```

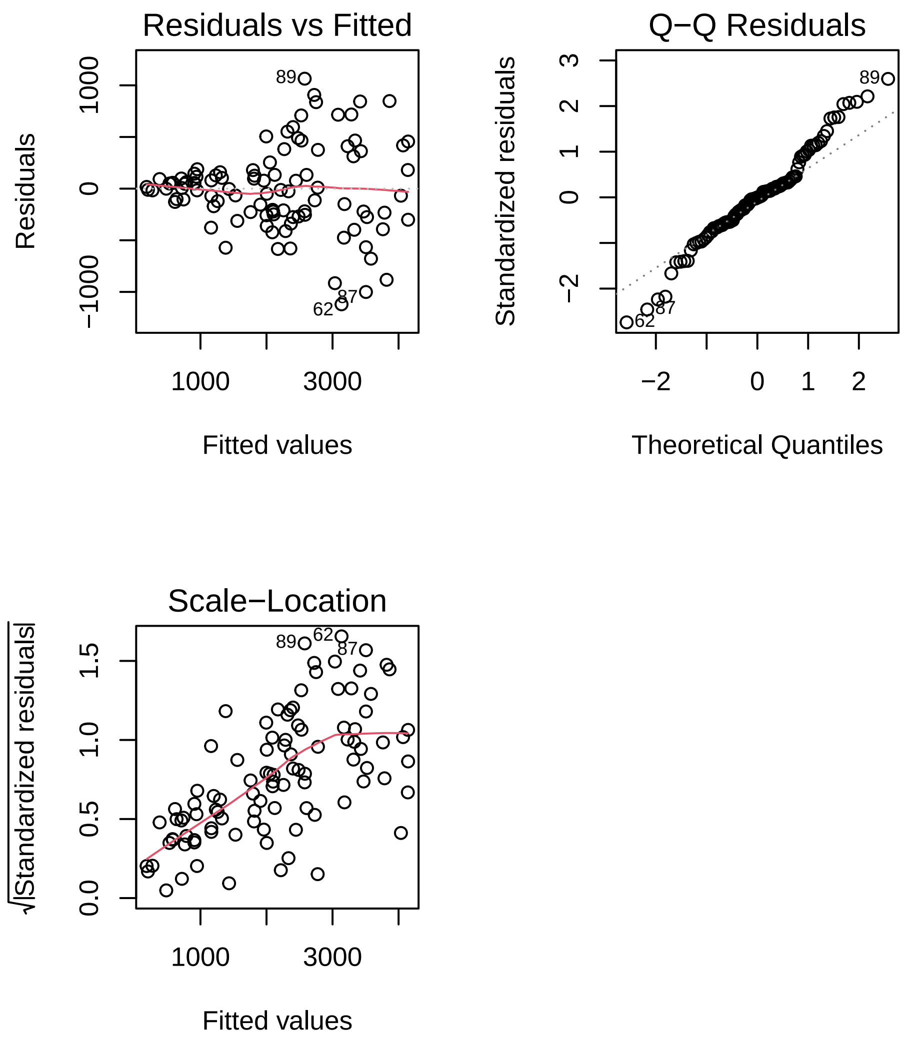
```
}
```

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 spreminjanju 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.)

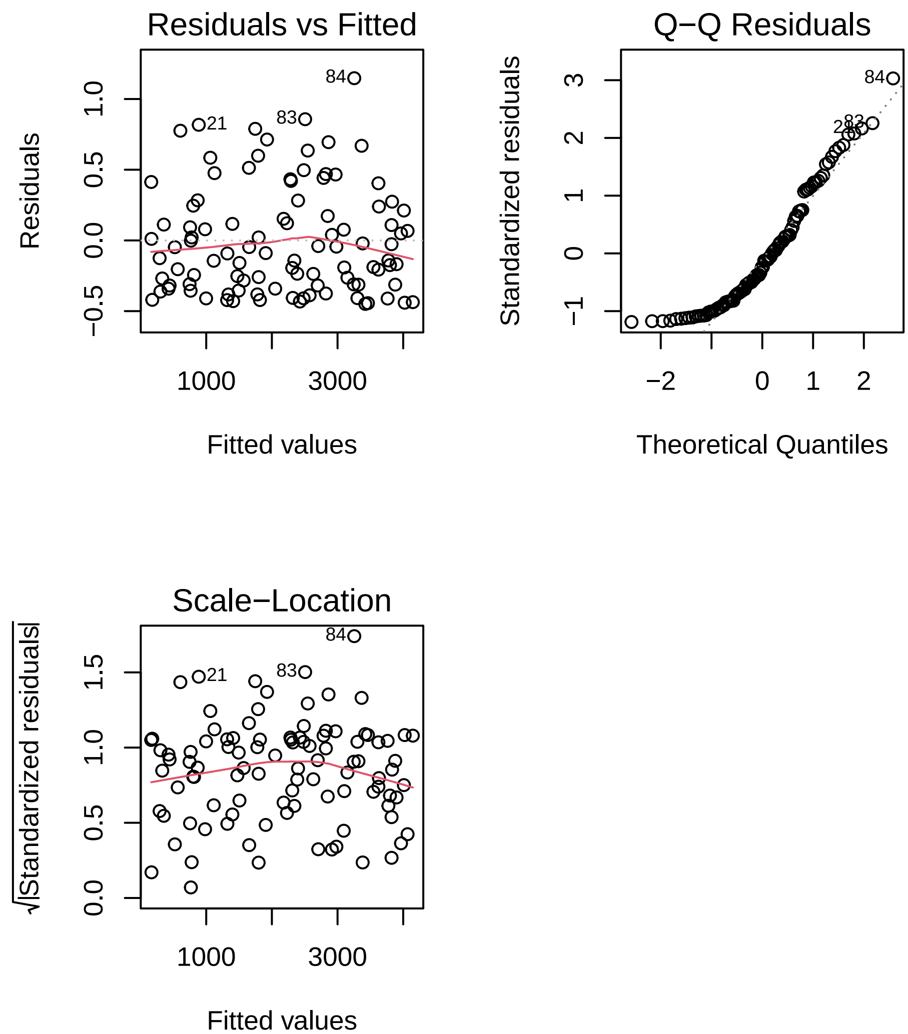
```
f.generiranje.lm.1.N <- function(beta0, beta1, n) {

  # 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 regresijki 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, $j = 1, \ldots, k$ 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  
```

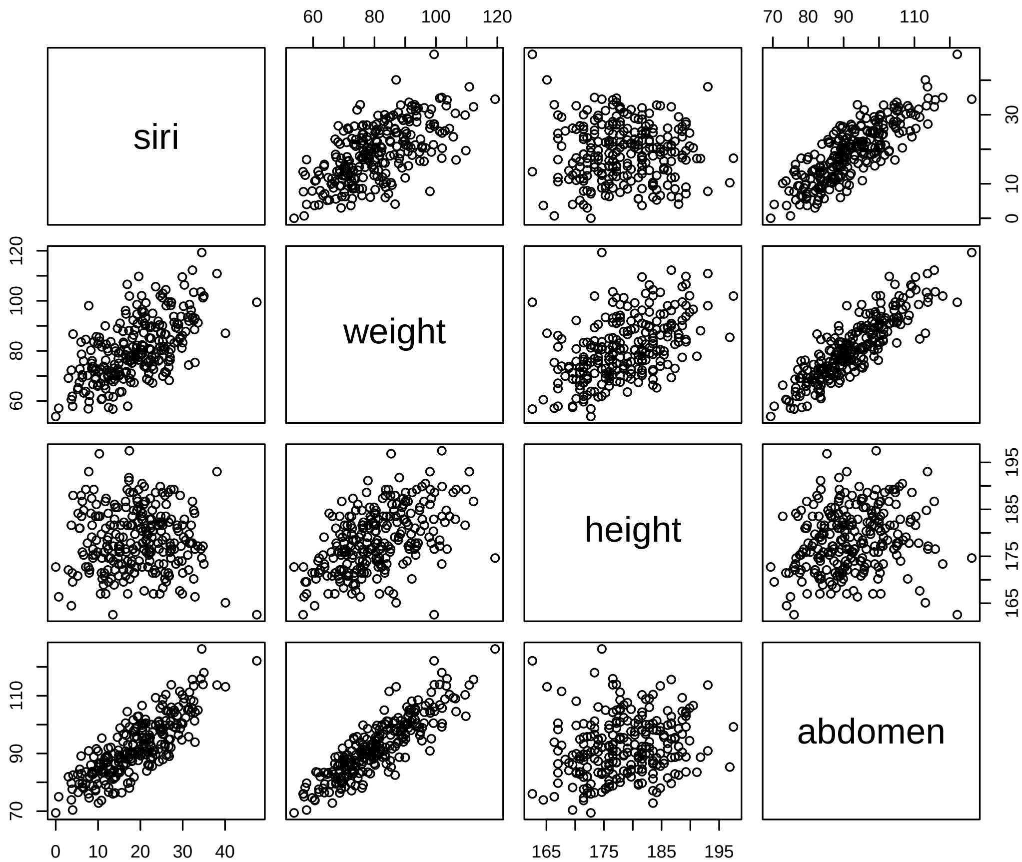
Spomnimo se močnih parnih korelacij, ki so značilne za dani podatkovni okvir. Poglejmo Pearsonove koeficiente korelacije med posameznimi pari spremenljivk v podatkovnem okviru:

```
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 maščobe, odvisno od tega, ali smo v modelu upoštevali tudi obseg trebuha.

```r
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 maščobe.

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 ocenjenem 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:

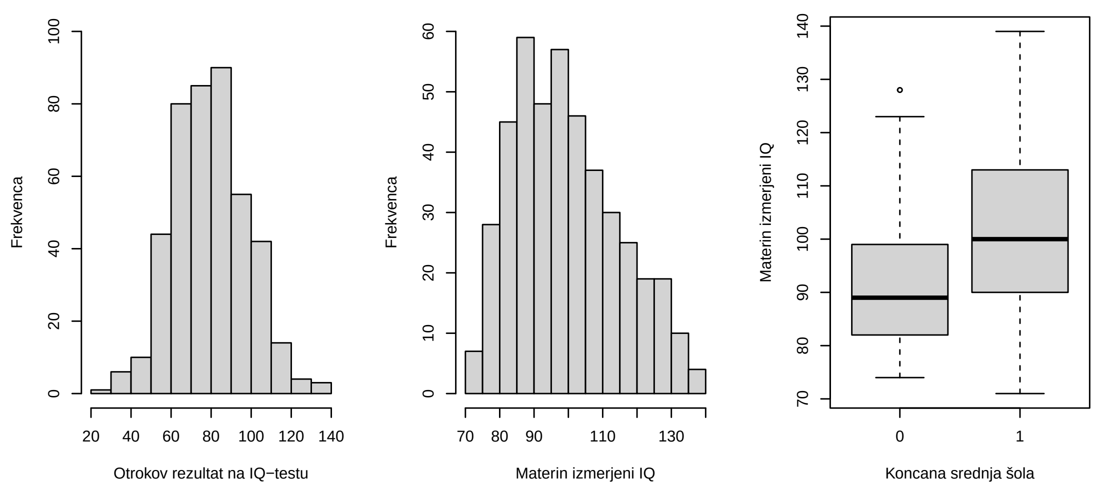
1. Ali vsaj ena od napovednih spremenljivk lahko pojasni del variabilnosti otrokovega IQ-ja? $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
2. Če lahko zavrnemo hipotezo iz prve točke: ali vse napovedne spremenljivke pomagajo pojasniti del variabilnosti napovedne spremenljivke ali zadostuje le podmnožica teh spremenljivk (več o tem v poglavju o izbiri modela)?
3. 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?
4. Kako natančno lahko napovemo rezultat na testu za nove otroke?
5. Kako dobro se model prilega podatkom? Je v modelu prisotna nelinearnost? Ali obstaja interakcija med napovednima spremenljivkama?

```r
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 ...
```

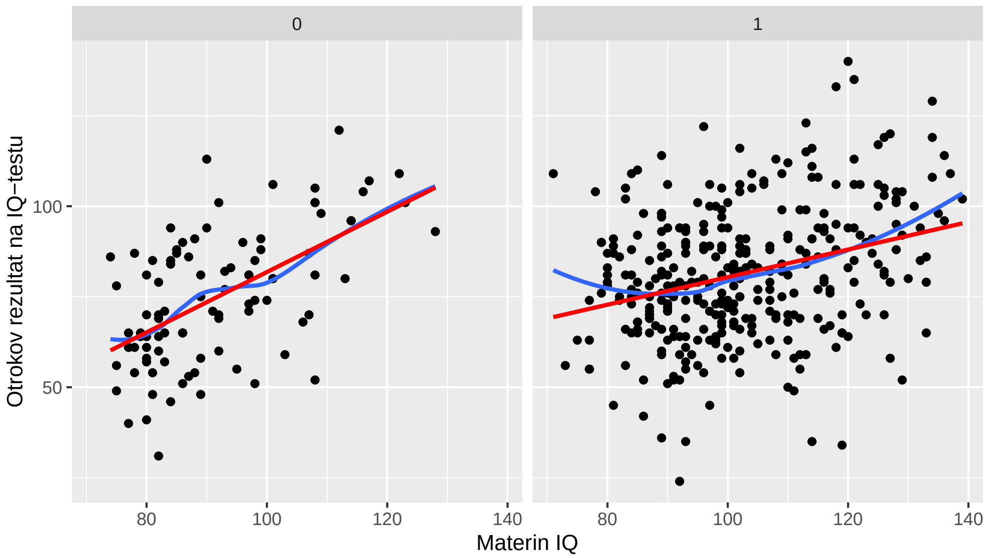
```
data$mom_hs <- factor(data$mom_hs)
```

```
# 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 linearna 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.

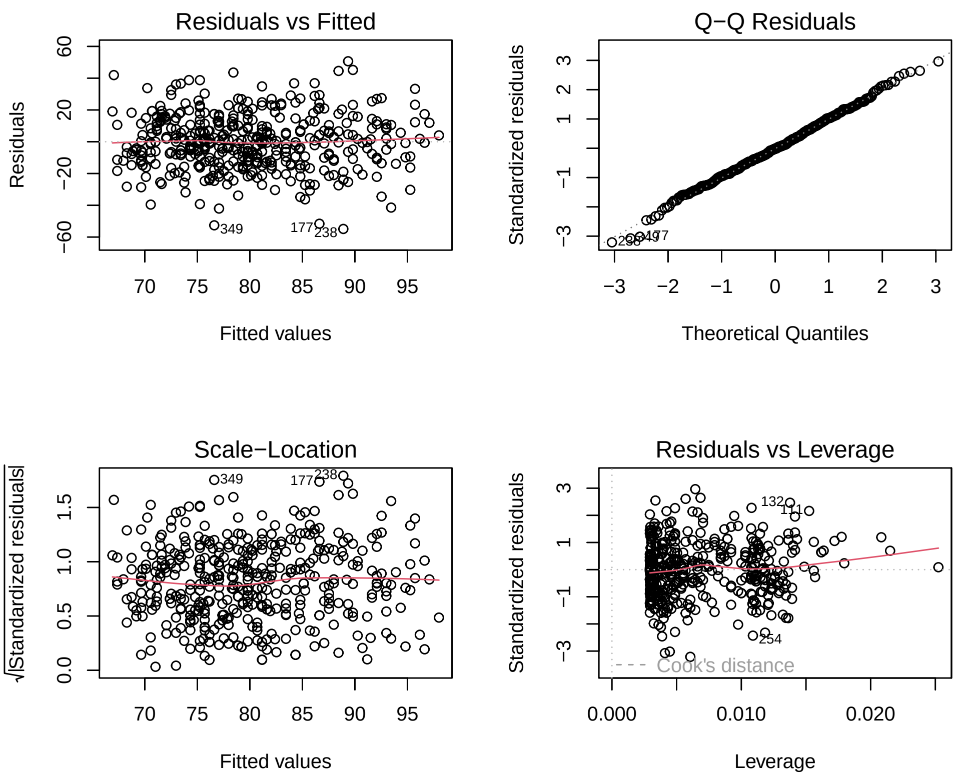
Graf nakazuje, da je zveza med `mom_iq` in `kid_score` drugačna glede na `mom_hs` in za `mom_hs=1` rahlo nelinearna.

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:

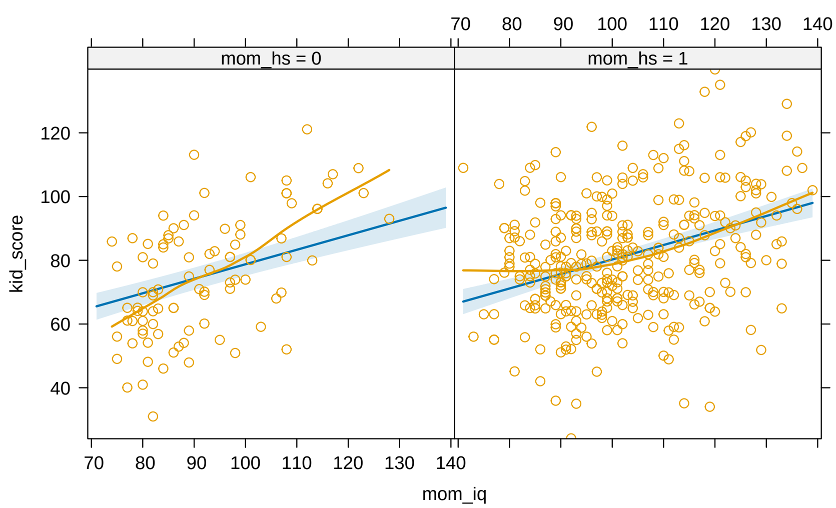
```
par(mfrow=c(2,2))
plot(m1)
```



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 ustrezen, si za vajo oglejmo povzetek modela ter interpretirajmo ocene parametrov:

```r
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`: $33.24 + 1.5 * \texttt{mom\_hs} + 0.46 * \texttt{mom\_iq}$.

Ker ima `mom_hs` dve možni vrednosti, $\texttt{mom\_hs} \in \{0, 1\}$, dobimo oceni za dve regresijski premici z različnima presečiščema in enakima naklonoma:

- `mom_hs=0` (referenčna kategorija): $33.24 + 0.46 * \texttt{mom\_iq}$;
- `mom_hs=1`: $(33.24 + 1.5) + 0.46 * \texttt{mom\_iq} = 34.74 + 0.46 * \texttt{mom\_iq}$.

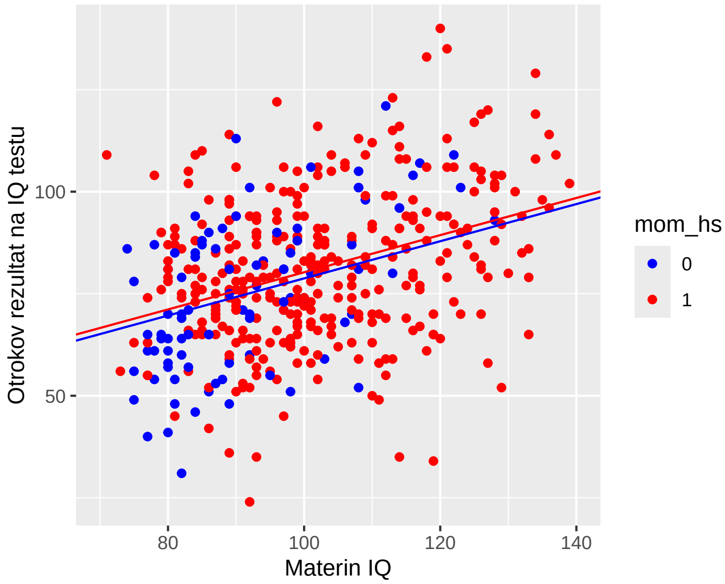
Za vajo interpretirajmo posamezne parametre modela `m1`:

- *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`:

```r
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 predstavljata 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`.

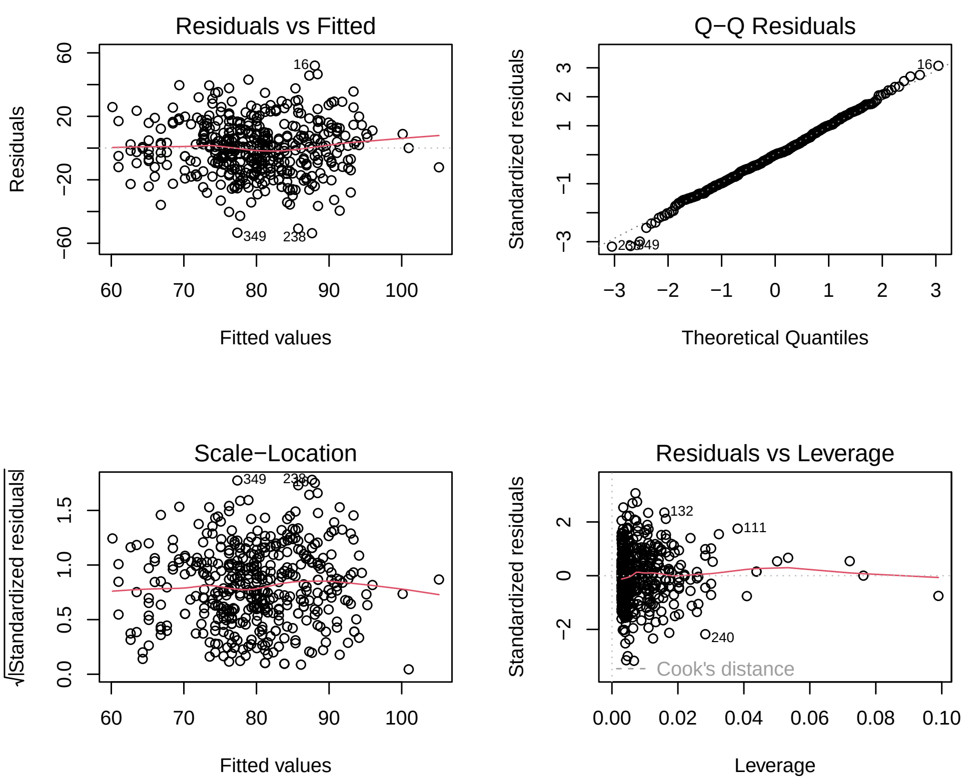
Prisotnost interakcije lahko preverimo tudi na podlagi F-testa za primerjavo gnezdenih modelov, ki testira domnevo, da sta modela ekvivalentna. Ničelno domnevo lahko zavrnemo: modela nista ekvivalentna, interakcija je v modelu potrebna. Primerjajte rezultate obeh testov.

```r
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:

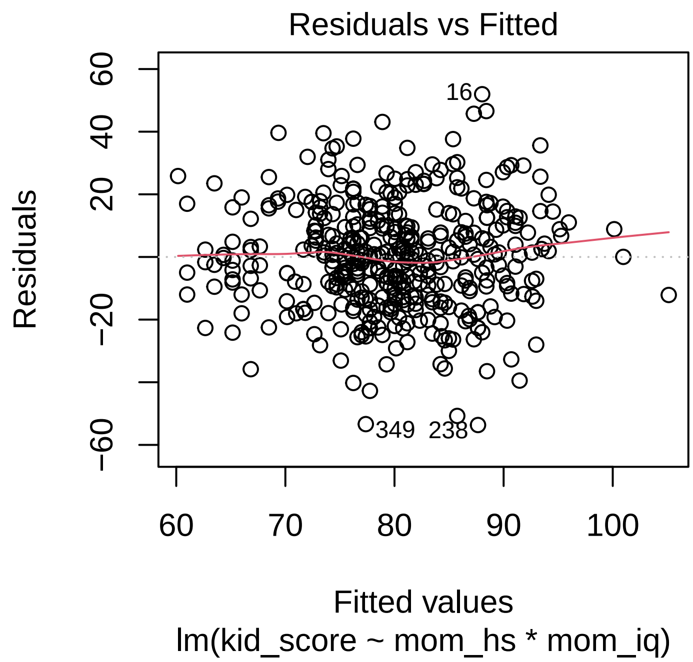
```r
par(mfrow=c(2,2))
plot(m2)
```



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:

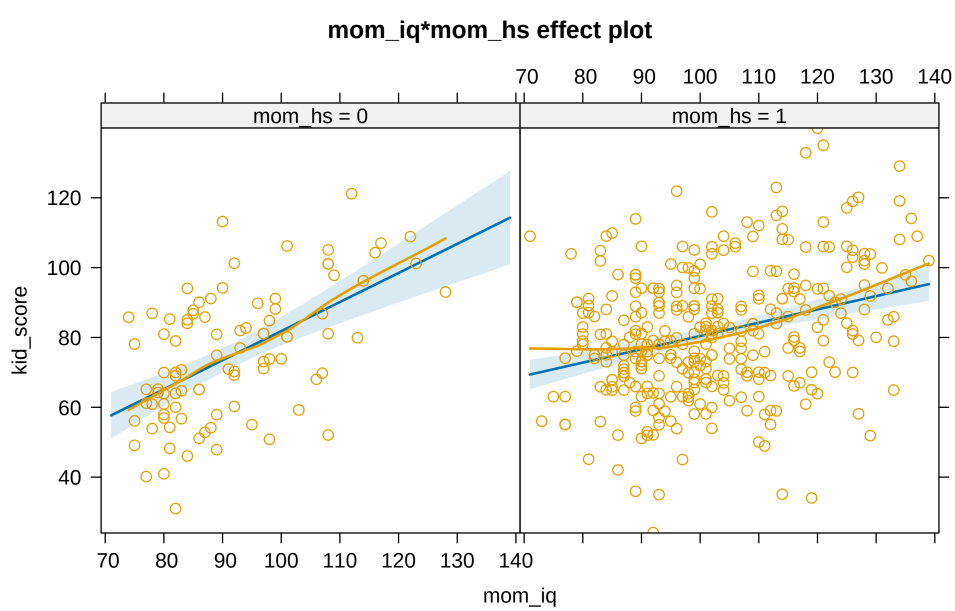
```r
plot(m2, which = 1)
```



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:

```r
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 $*$ `mom_hs` + 0.83 $*$ `mom_iq` - 0.45 $*$ `mom_hs` $*$ `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 $*$ `mom_iq`;
- `mom_hs=1`: (-1.45 + 43.78) + (0.83 - 0.45) $*$ `mom_iq` = 42.32 + 0.38 $*$ `mom_iq`.

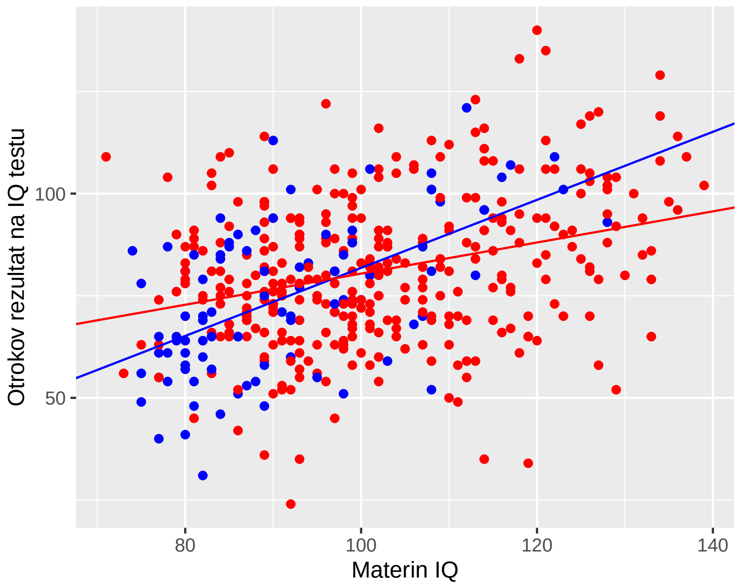
Razlaga posameznih parametrov:

- *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`:

```r
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
```

```r
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 razpade 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:

```r
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:

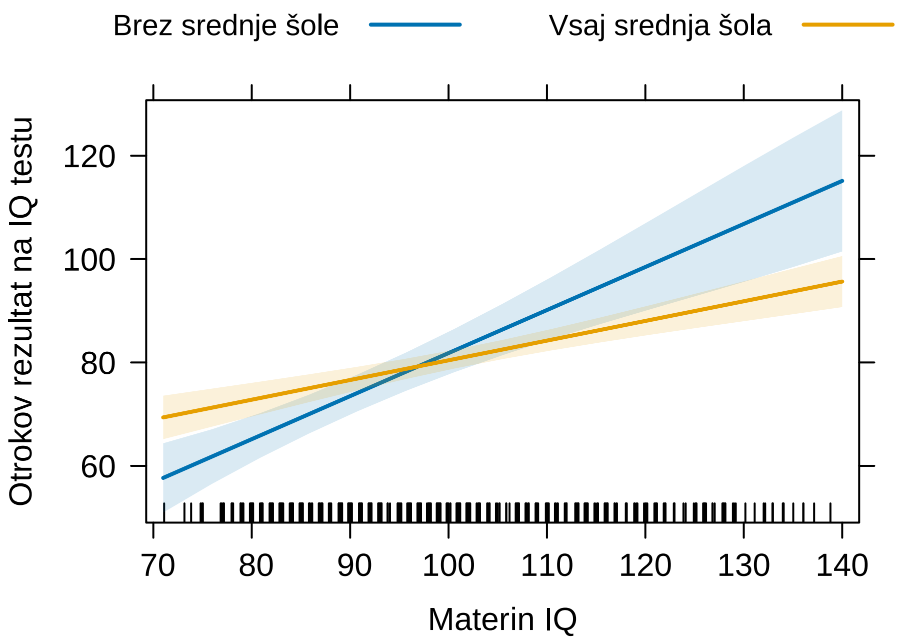
```r
y_fit_m2.3 <- m2.3$fitted.values
SS.res <- sum((y_fit_m2.3 - data$kid_score)^2)
SS.total <- sum((data$kid_score - 0)^2)
# SS.total se računa relativno na vrednost 0!
1-SS.res/SS.total
```

```
[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`:

```r
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 posledice 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 pojavijo 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. dopolnite 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.