Kazalo

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 ustrezno matematično predstavili. Katere spremenljivke so številske, katere opisne, kako bomo opisne spremenljivke vključili v linearni model. Pomembno vlogo v tej fazi predstavljajo ustrezni 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 Generallized 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 predstavljevosti 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 spremenimo 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.

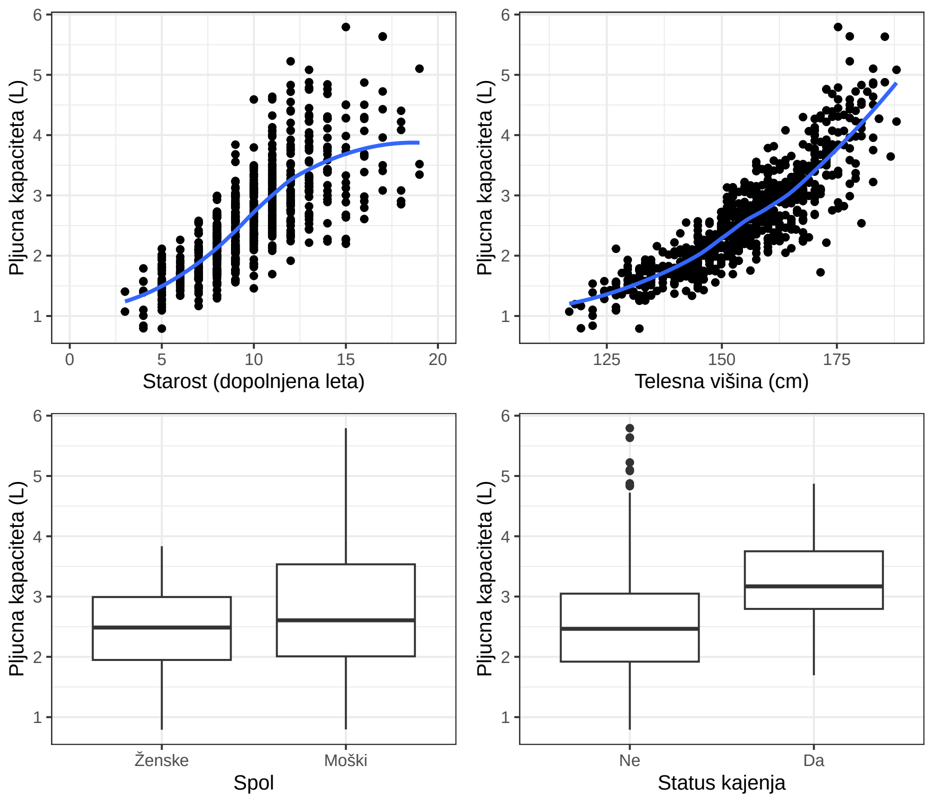
Raziskovalno vprašanje je, kako je pljučna kapaciteta ob upoštevanju spola, starosti in telesne višine, povezana s kajenjem. Raziskava je bila narejena kot opazovalna študija. Ni šlo za načrtovano študijo s kontroliranimi vrednostmi napovednih spremenljivk in temu ustreznim slučajnim izborom otrok in 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).

```r
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)
```



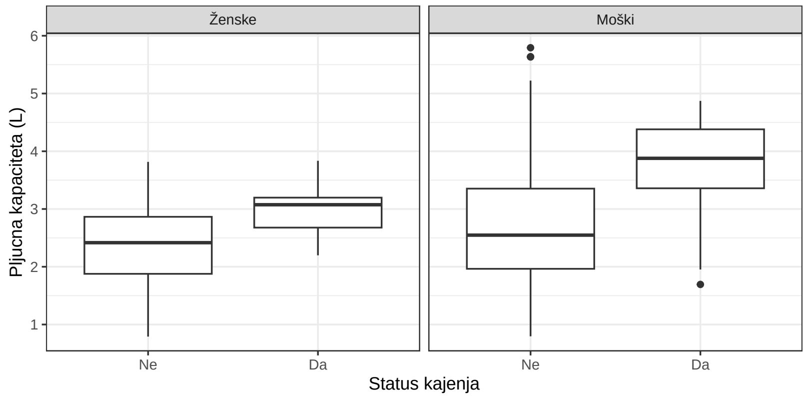
Slika 1: `FEV` v odvisnosti od `Age`, `Ht` (z dodanim gladilnikom), `Gender` in `Smoke`

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` kadilcev je večja kot pri nekadilcih, kar je malo nenavadno in bi bilo lahko posledica veliko manjšega vzorca za kadilce. 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 kadilci 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 kadilcih, 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` kadilcev višja kot pri nekadilcih.

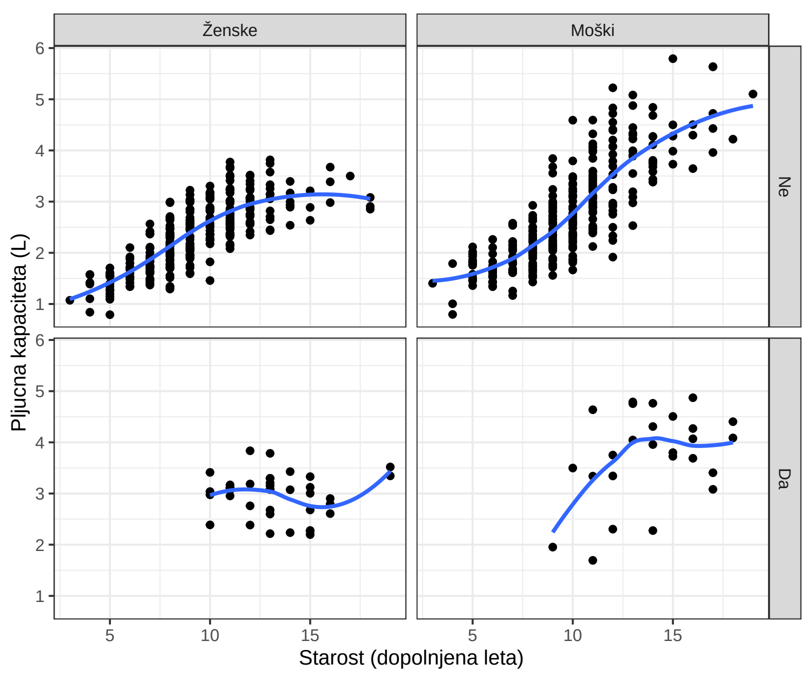
```
ggplot(lungcap, aes(x=Smoke, y=FEV))+ geom_boxplot() +  facet_wrap(.~ Gender) +
  xlab("Status kajenja") + ylab("Pljučna kapaciteta (L)") + theme_bw()
```



Slika 2: `FEV` v odvisnosti od `Gender` in `Smoke` hkrati

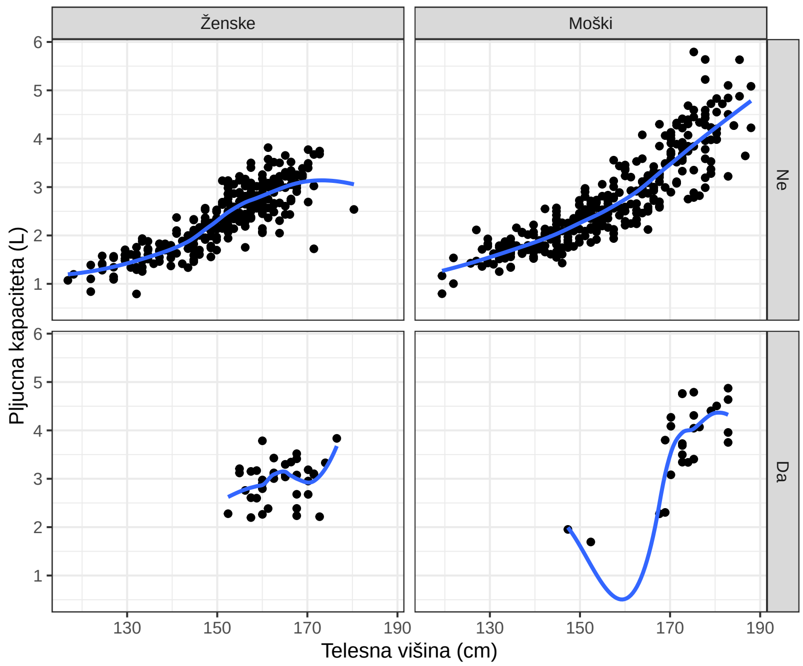
Bolj jasno nelinearno odvisnost `FEV` od starosti vidimo, če na sliko dodamo gladilnik. Hkratno odvisnost `FEV` od starosti, spola in statusa kajenja lahko prikažemo tako, da podatke razdelimo v skupine glede na spol in kajenje (Slika 3). Gladilnik pokaže nelinearno odvisnost `FEV` od starosti tako v 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 ravni obeh 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
```

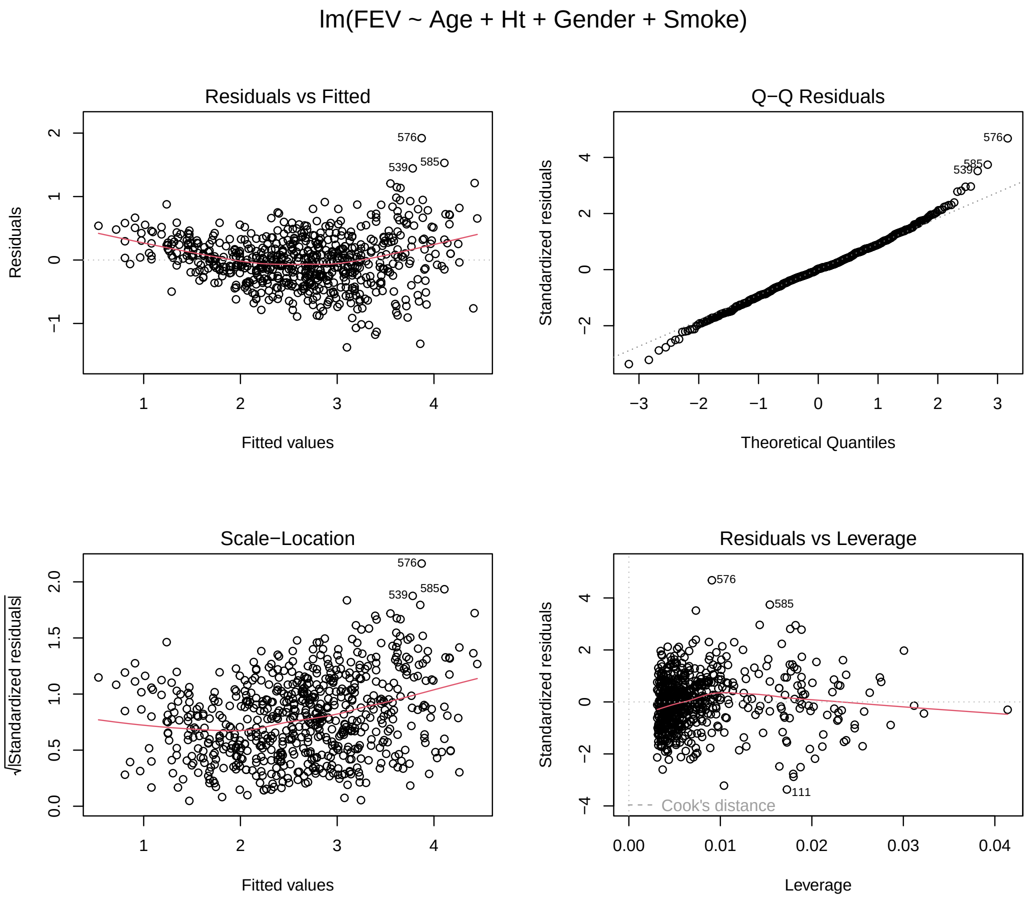
```
    (Intercept) Age     Ht GenderMoški SmokeDa
649           1  16 176.53           1       1
650           1  16 182.88           1       1
651           1  17 170.18           1       1
652           1  17 175.26           1       1
653           1  18 170.18           1       1
654           1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



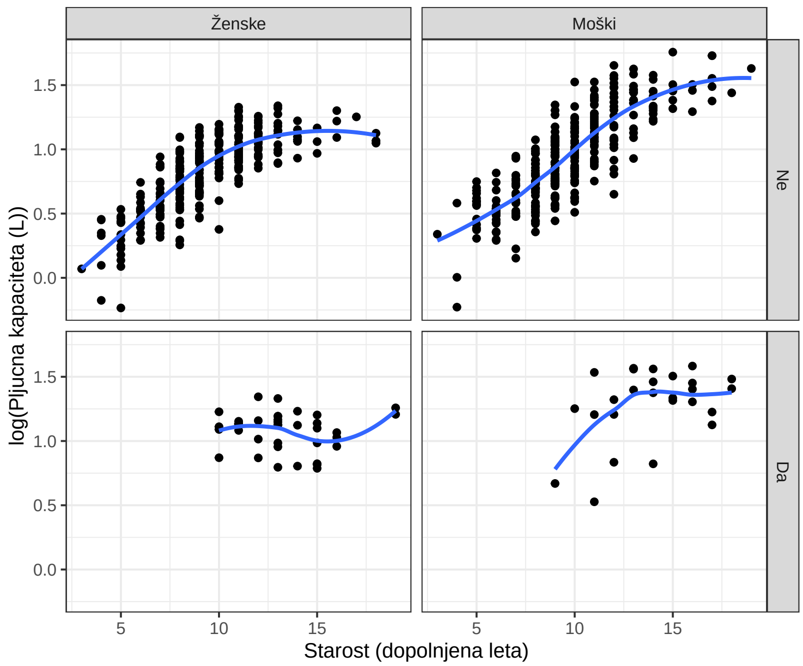
Slika 5: Slike ostankov za `mod1`

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 ustreznejšega linearnega modela, če v prejšnjem koraku predpostavke niso izpolnjene

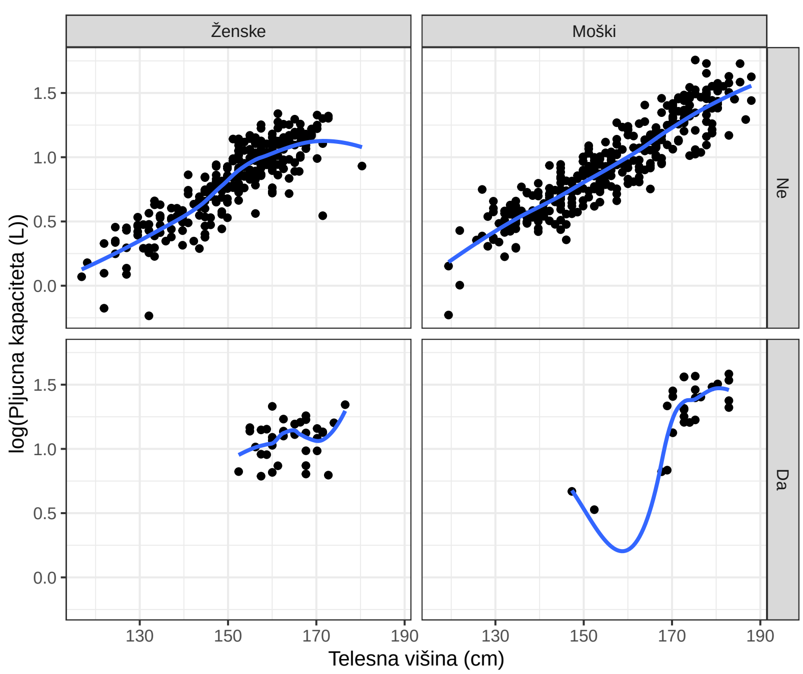
V naslednjem koraku poskusimo modelirati transformirano odzivno spremenljivko: logaritmiramo spremenljivko `FEV`. Najprej poglejmo grafične prikaze za `log(FEV)` (Sliki 6 in 7).

```
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(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()
```



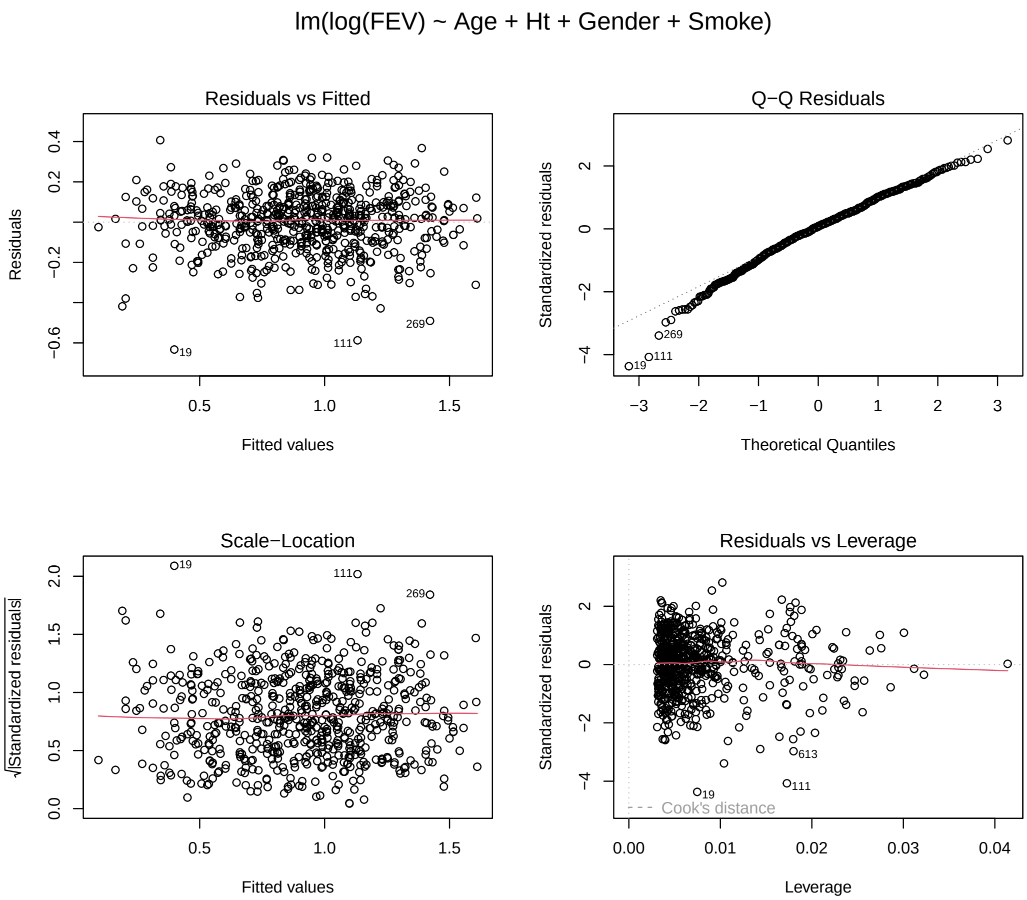
Slika 7: `log(FEV)` v odvisnosti od `Ht`, `Gender` in `Smoke` hkrati

Sliki 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
```

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{1}$$

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

```
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
```

```r
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(FEV)` za ženske, ki kadijo, so stare 18 let in visoke 168 cm ter pripadajoči standardni odklon povprečne napovedi:

```r
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()`.

```r
names(mod2)
```

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

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

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

```r
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:

```r
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`

```r
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.

```r
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
```

```r
# intervali zaupanja za parametre modela - velikost vpliva posamezne napovedne
# spremenljvke 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 spremni 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 presečišiščem 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 presečišiščem 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 $SmokeDa = 1$:

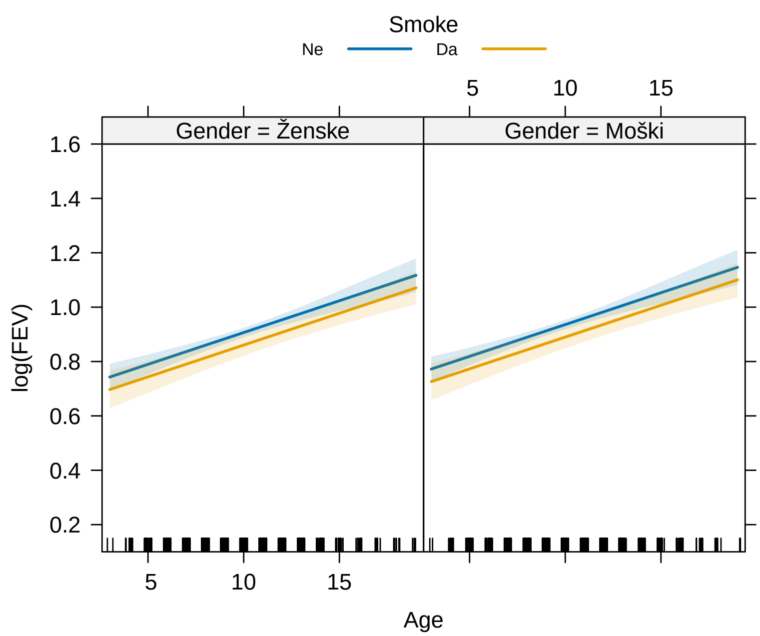
$$\hat{y} = E(log(FEV)) = (-1.944 + 0.029 - 0.046) + 0.023Age + 0.017Ht.$$

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

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

```
[1] 155.3047
```

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

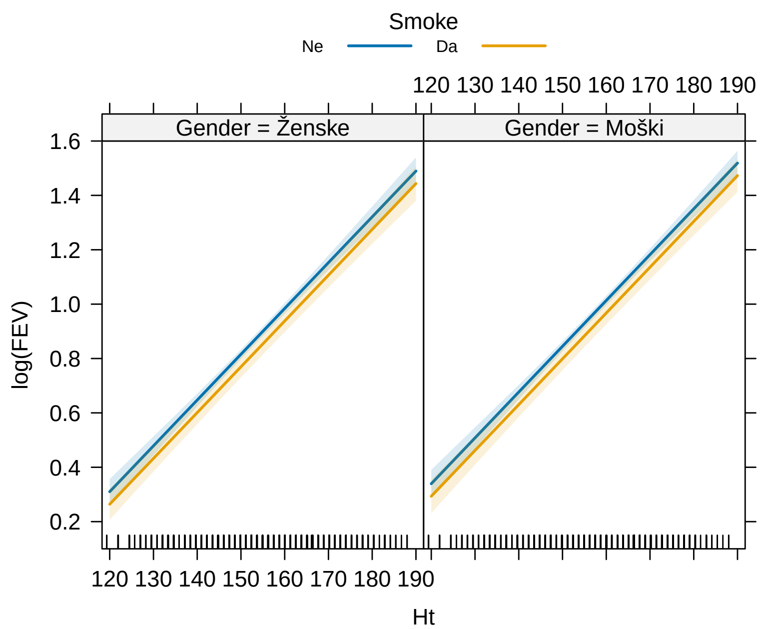


Slika 9: Modelske napovedi za `log(FEV)` v odvisnosti od `Age`, `Gender` in `Smoke` hkrati, pri povprečni vrednosti `Ht` za `mod2`

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

```
[1] 9.931193
```

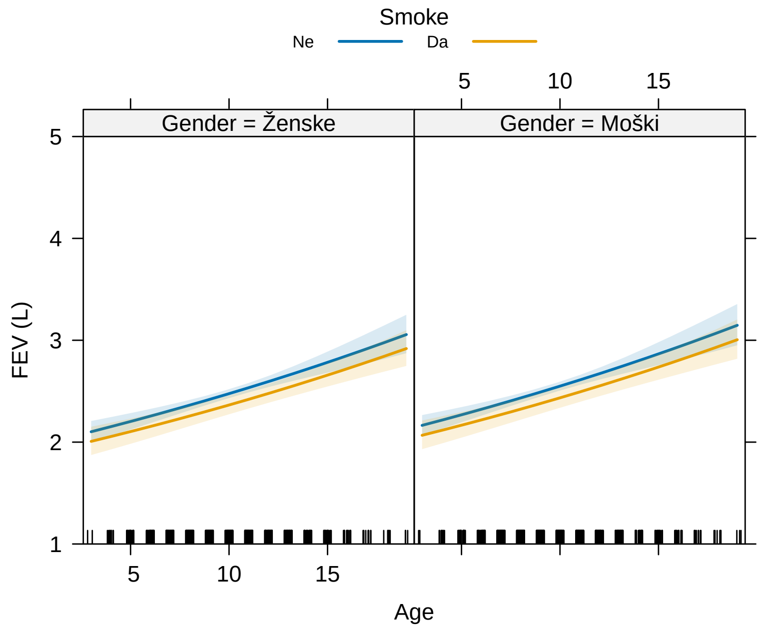
```
plot(Effect(c("Smoke", "Gender", "Ht"), mod2), multiline=TRUE,
     ci.style="bands", main="", ylim=c(0.1,1.6))
```



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

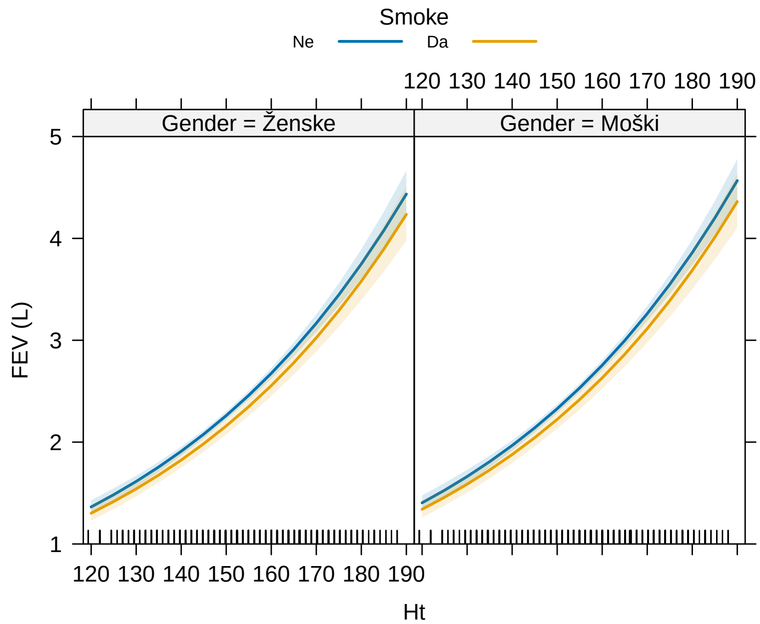
Napovedi za `mod2` originalni skali (FEV (L)) so predstavljene na Slikah 11 in 12.

```
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 šumniki 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
# zapišemo 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
```

```r
#  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
```

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

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

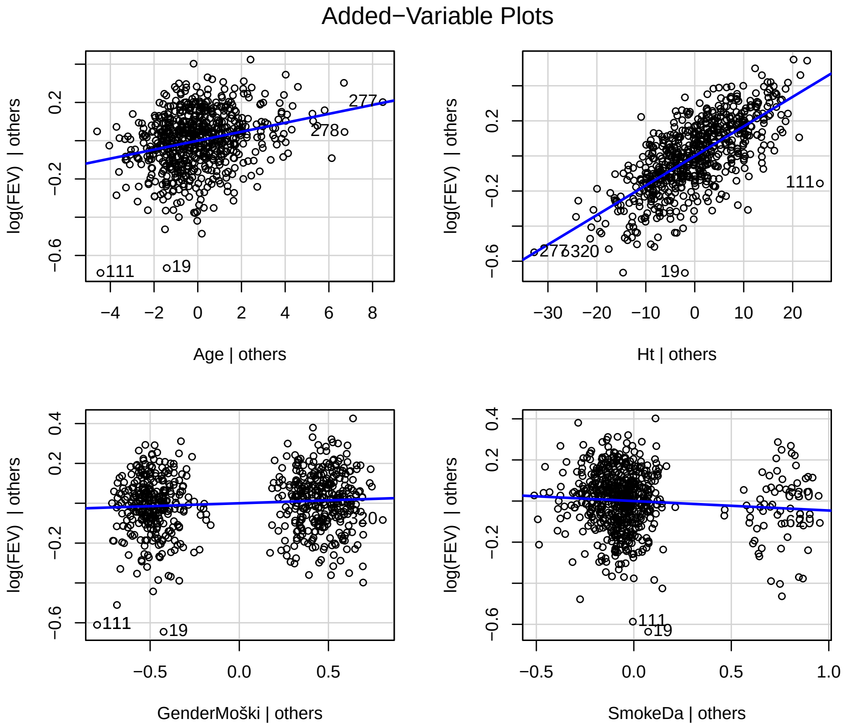
```
  Age  Ht Gender Smoke      fit       lwr      upr
1  17 168      Z    Da 1.238105 0.9493434 1.526866
2  18 168      Z    Da 1.261492 0.9721736 1.550810
3  19 168      Z    Da 1.284879 0.9948558 1.574902
```

```r
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)
```

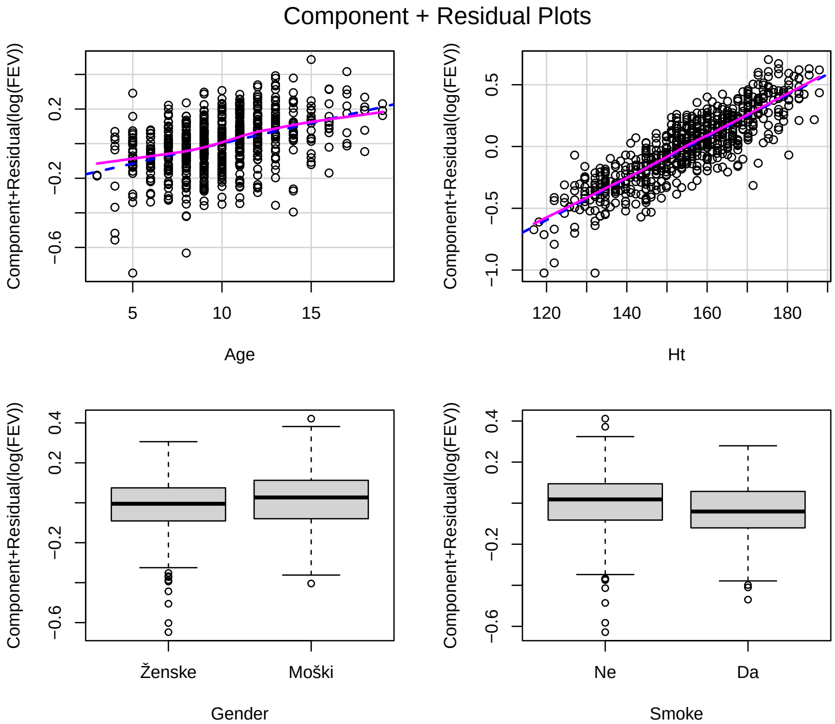


Slika 13: Grafikoni dodane spremenljivke, `avPlots(mod2)`

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

Slika 13 prikazuje grafikone dodane spremenljivke za `mod2`. Leva zgornja sličica prikazuje zvezo med `log(FEV)` in `Age` ob upoštevanju ostalih spremenljivk v modelu. Naklon premice je enak oceni parametra za `Age` v `mod2`. Vidimo, da s starostjo `log(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(FEV)` in `Ht` ob upoštevanju ostalih spremenljivk v modelu. Spodnja leva sličica prikazuje zvezo med `log(FEV)` in `Gender` ob upoštevanju ostalih spremenljivk v modelu, kaže, da imajo moški v povprečju nekoliko večjo vrednost `log(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(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 številske regresorje modra črtkana premica prikazuje modelske napovedi `log(FEV)` glede na vrednosti posameznega regresorja pri povprečnih vrednostih ostalih regresorjev; točke predstavljajo parcialne ostanke za regresor, ki je na vodoravni osi. Gladilnik je narisan 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(FEV)` ob upoštevanju `Ht`, `Gender` in `Smoke`. Podobno lahko rečemo za zvezo med `log(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(FEV)` v odvisnosti od `Smoke`).

1.8 Opisna spremenljivka v linearnem 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:

H0: povprečna pljučna kapaciteta kadilcev je enaka povprečni pljučni kapaciteti nekadilcev.

H1: 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)
```

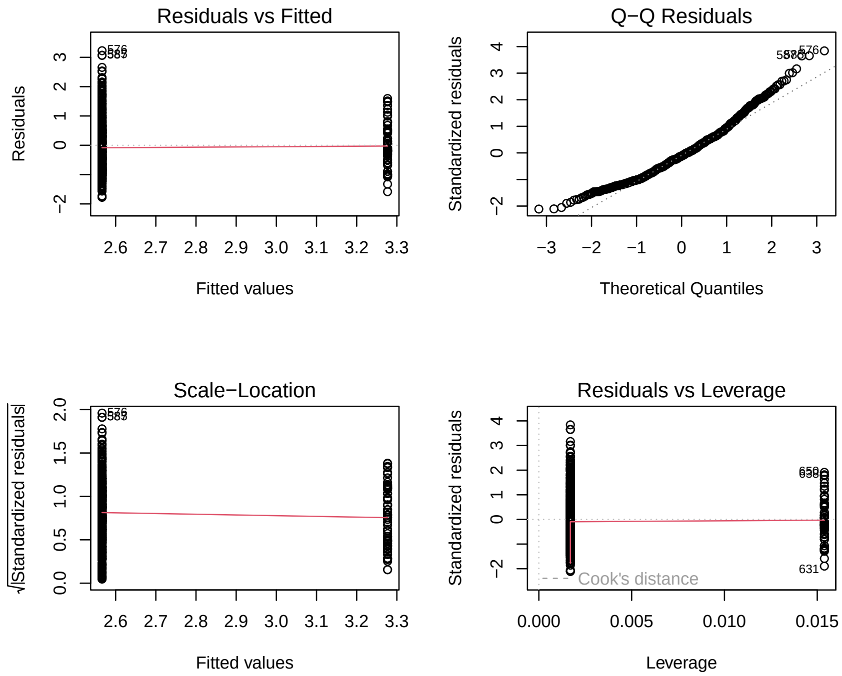
```

    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 nepričakovan, 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 grafikoni 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
```

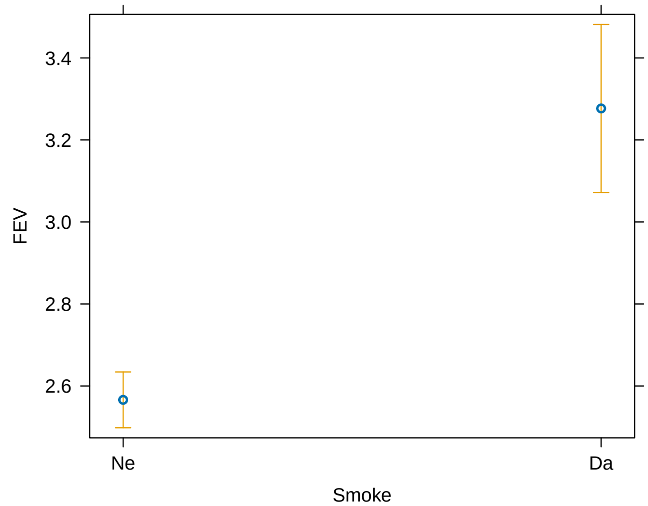
```
(Intercept)     SmokeDa
  2.5661426   0.7107189
```

```
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 Welchove testu:-2.57+3.28=0.71). Intervalov zaupanja 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:

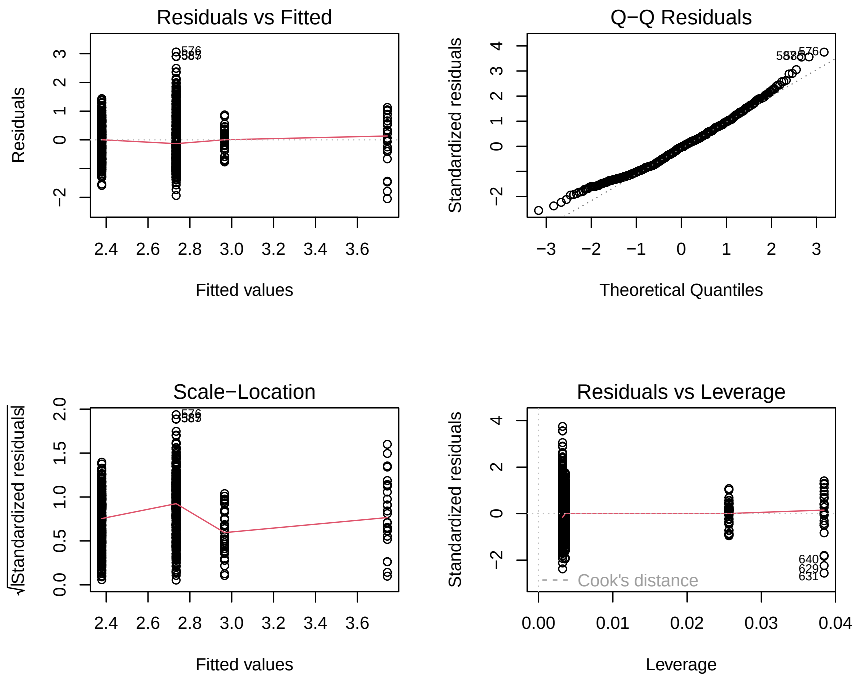
```
lungcap$Gender.Smoke <- lungcap$Gender:lungcap$Smoke
levels(lungcap$Gender.Smoke) # spremenljivka ima 4 vrednosti/kategorije
```

```
[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 grafikoni za `mod.opisna.4`

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

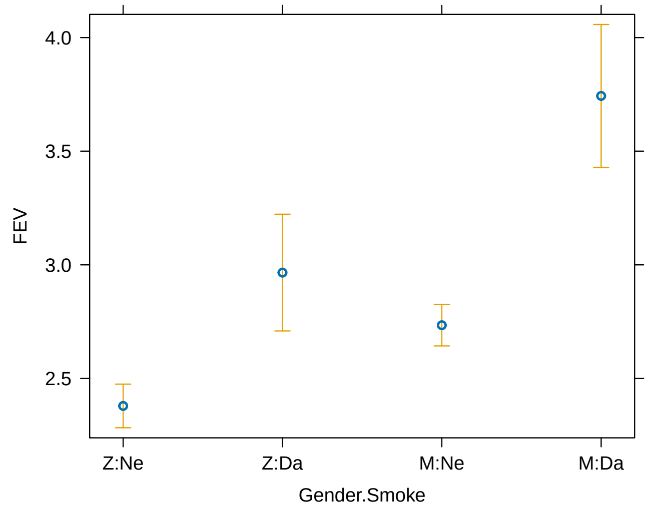
```
(Intercept) Gender.SmokeZ:Da Gender.SmokeM:Ne Gender.SmokeM:Da
  2.3792115        0.5867372        0.3551692        1.3640193
```

```
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)
```



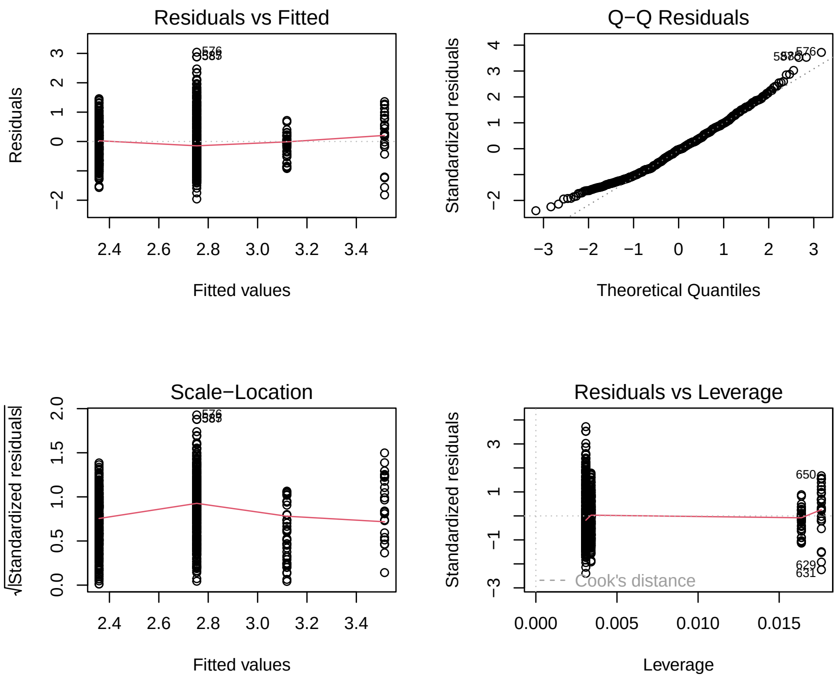
Slika 18: Modelske napovedi za `FEV` v odvisnosti od `Gender.Smoke` s pripadajočimi 95 % intervali zaupanja za povprečno napoved za `mod.opisna.4`

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
```

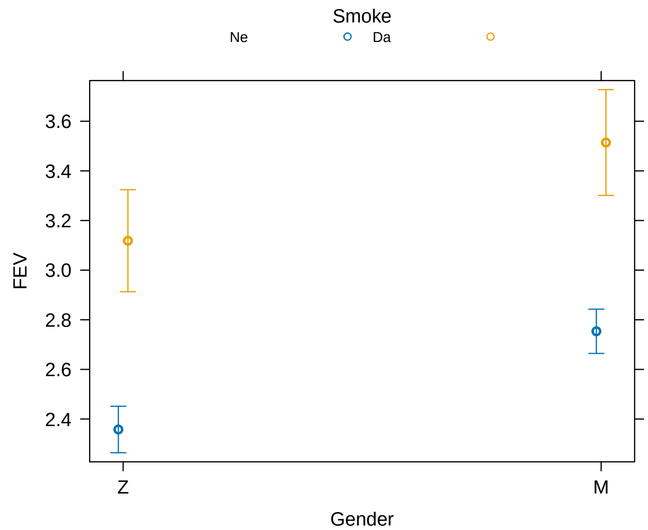
```
(Intercept)     SmokeDa     GenderM
  2.3578761   0.7607029   0.3957065
```

```
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)
```



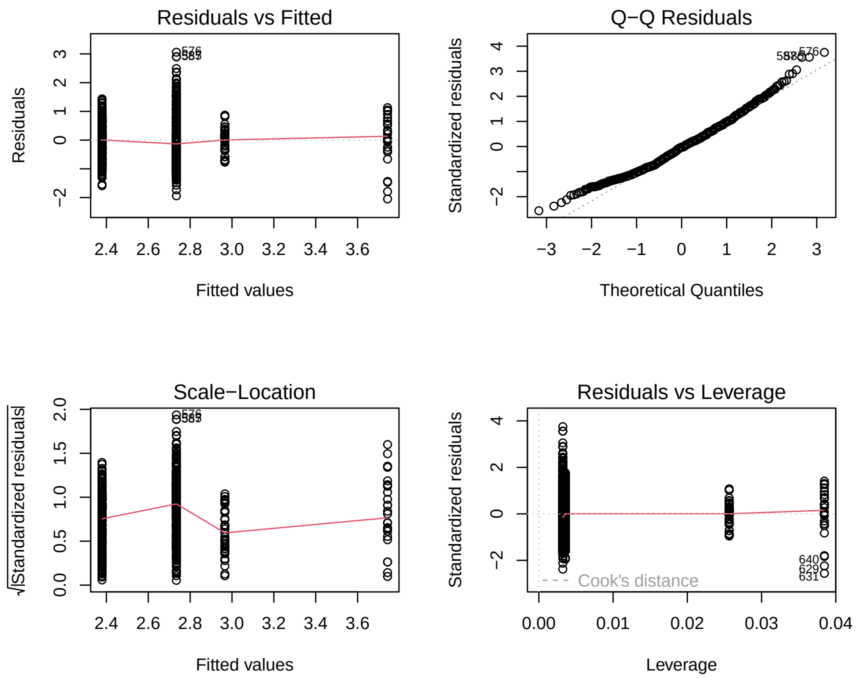
Slika 20: Modelske napovedi za FEV v odvisnosti od Smoke in Gender s pripadajočimi 95 % intervali zaupanja za povprečno napoved za mod.opisni.2

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
```

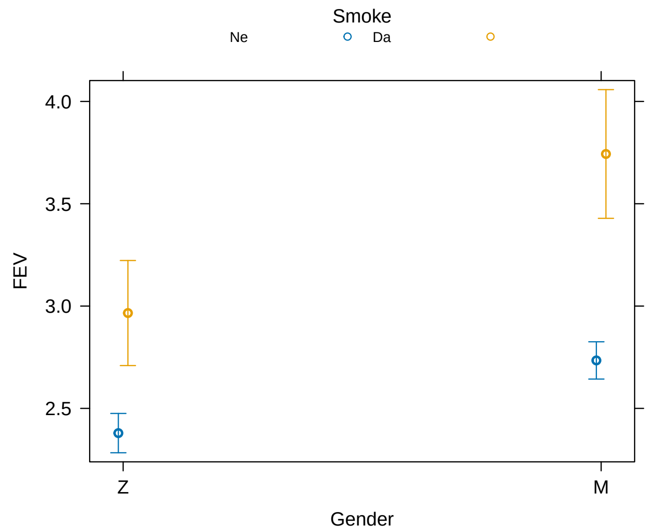
```
    (Intercept)          SmokeDa          GenderM SmokeDa:GenderM
      2.3792115        0.5867372        0.3551692       0.4221129
```

```
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)
```



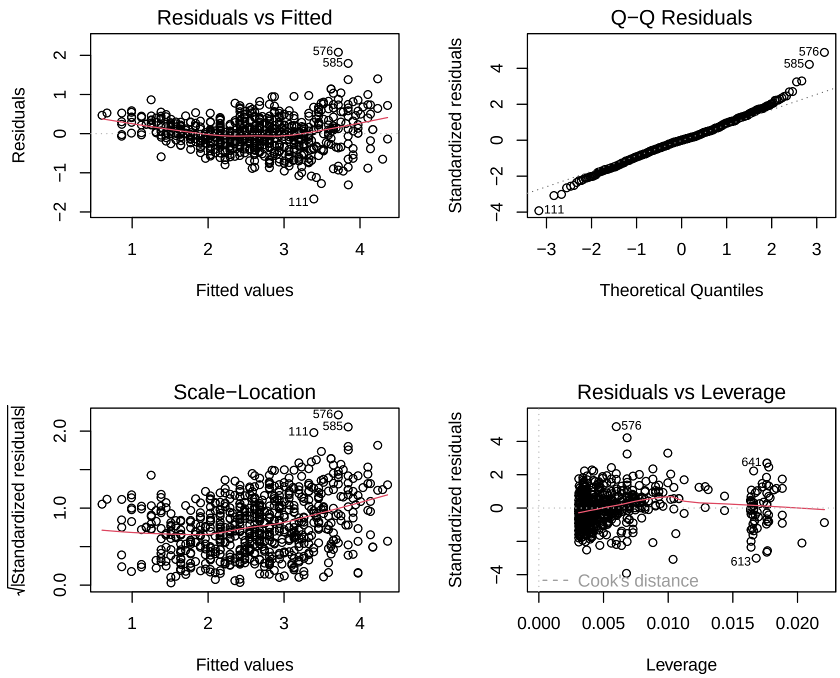
Slika 22: Modelske napovedi za `FEV` v odvisnosti od `Smoke` in `Gender` ter njune inerakcije s pripadajočimi 95 % intervali zaupanja za `mod.opisni.2.int`

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 linearna. 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 diagnostičnih grafikonih precej spremenila. Prisotna je nelinearna zveza med ostanki in prilagojenimi vrednostmi ter nekonstantna varianca napak. Vseeno poglejmo pomen parametrov modela.

```
mod3$coeff
```

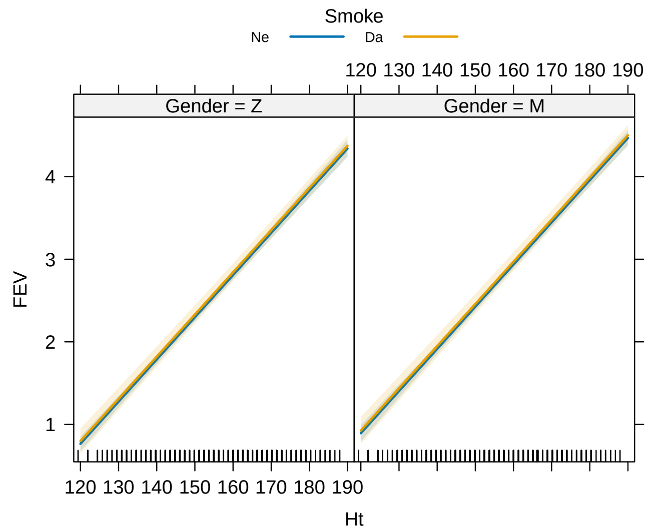
```
(Intercept)     GenderM     SmokeDa          Ht
-5.36207814  0.12764341  0.03413801  0.05106019
```

```
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 kadilce in nekadilce minimala, ampak še vedno pozitivna. Modelske napovedi geometrijsko predstavljajo 4 vzporedne premice.

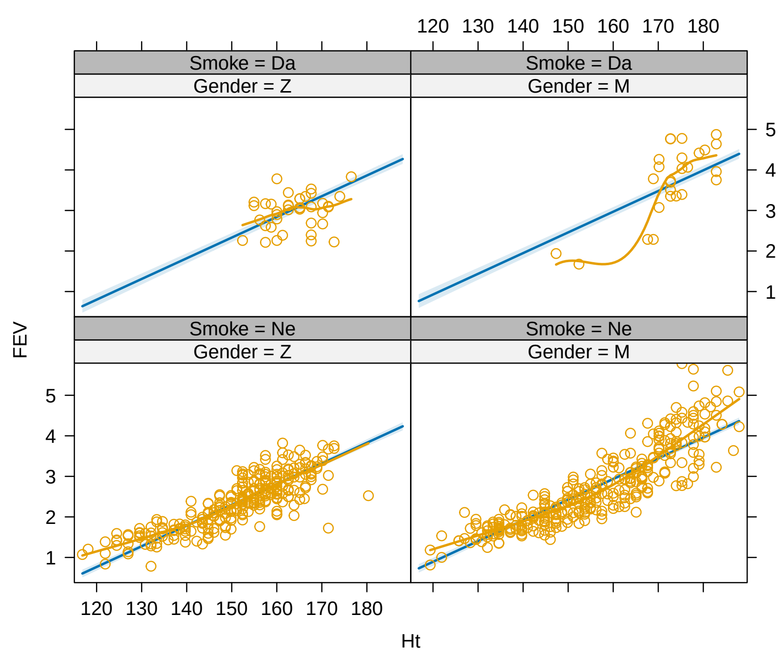
```
plot(Effect(c("Ht", "Smoke","Gender"), mod3), multiline=TRUE, ci.style="band", main="")
```



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

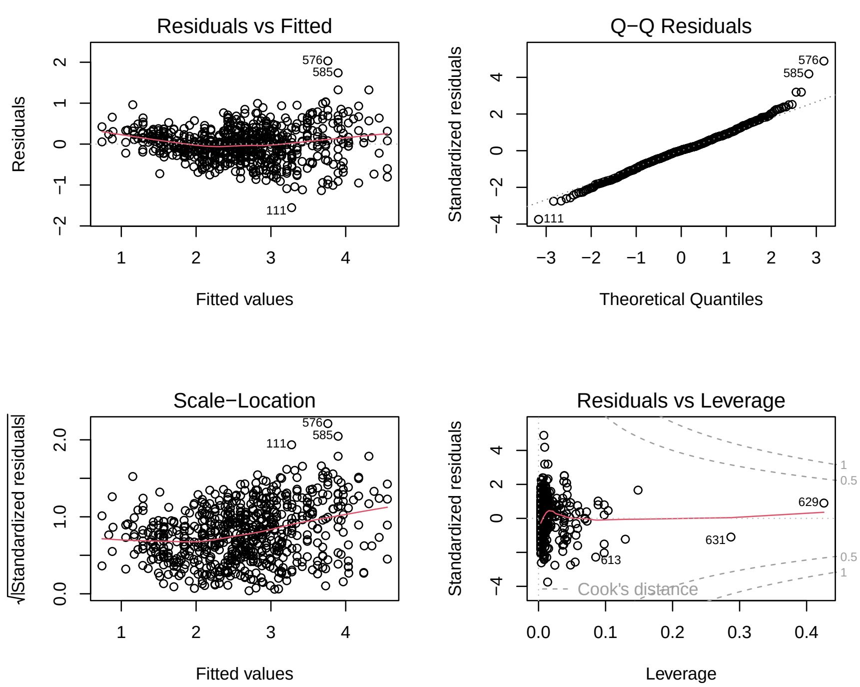
Grafikon kaže prisotnost nelinearnosti, pa tudi drugačno odvisnost `FEV` od `Ht` v skupinah `Kadi`, `NeKadi` pri moških in pri ženskah. To bi lahko pomenilo, da interakcijski člen med `Ht`, `Smoke` in `Gender` pojasni pomemben del variabilnosti `FEV`.

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`

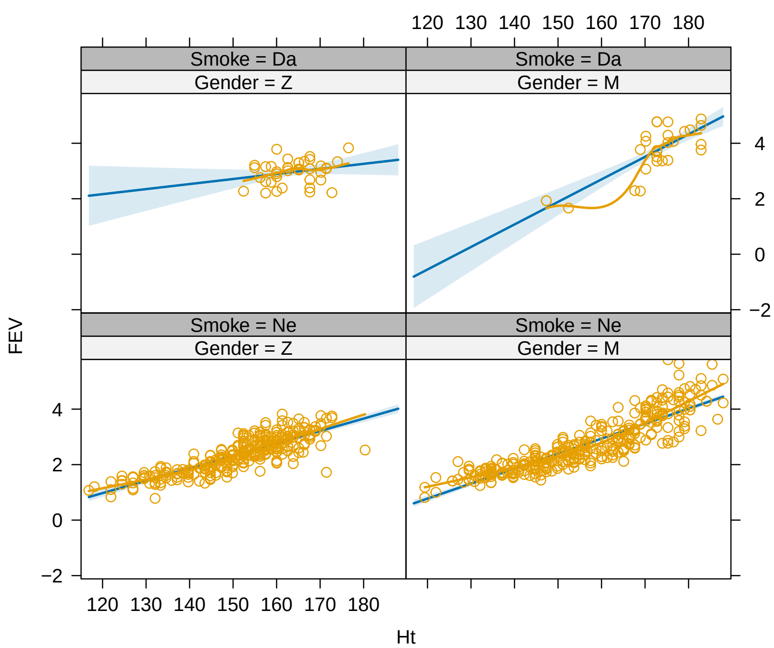
```
summary(mod3.int)
```

```
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 ustreznejši:

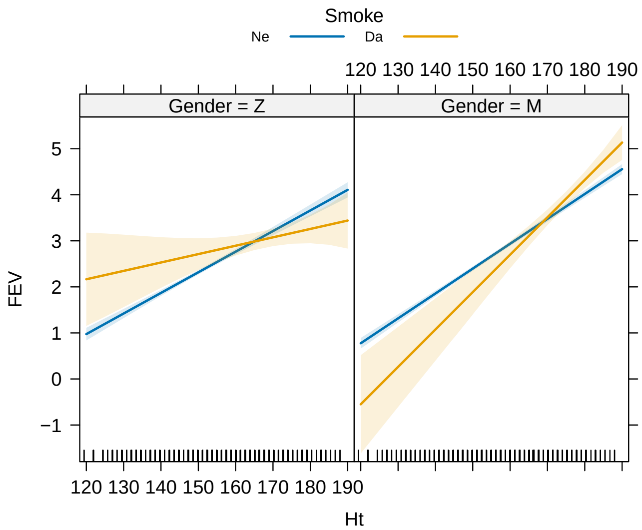
```
plot(Effect(c("Ht","Gender","Smoke"), mod3.int, partial.residuals=TRUE), main="")
```



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).

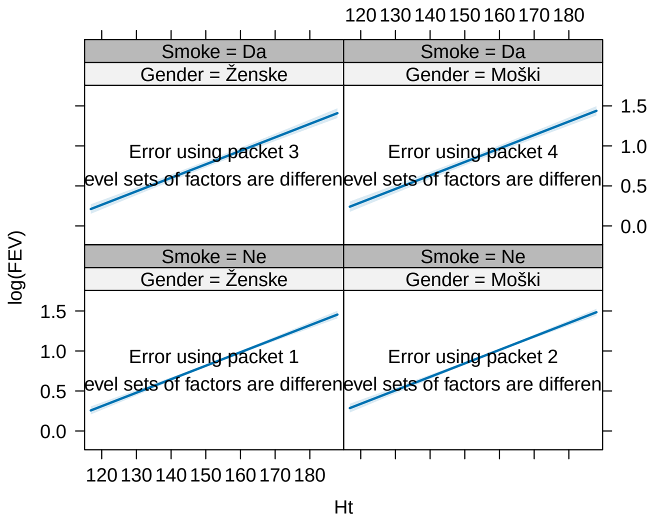
```
plot(Effect(c("Ht", "Smoke","Gender"), mod3.int), multiline=TRUE,
     ci.style="band", main="")
```



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(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="")
```

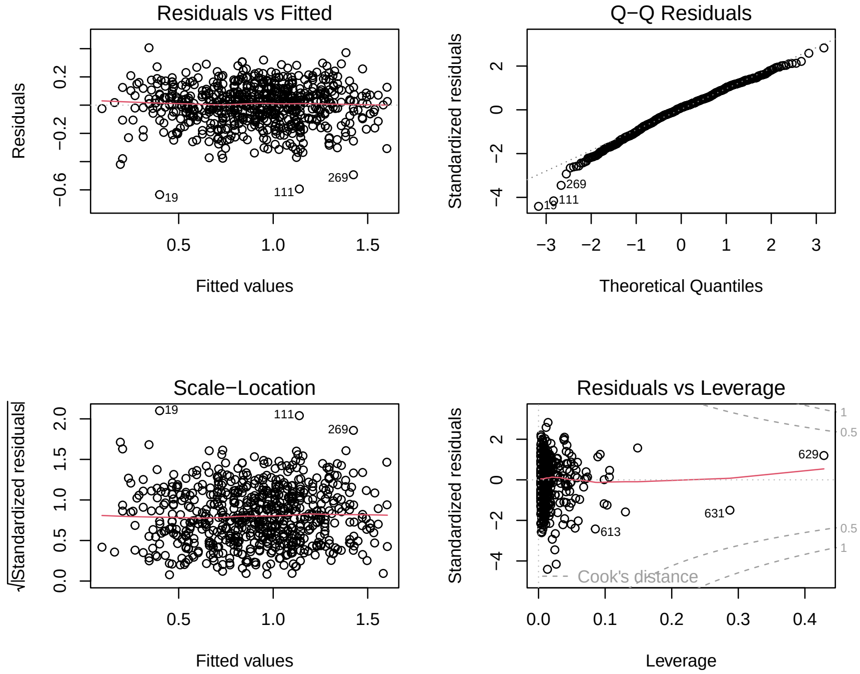


Slika 29: Modelske napovedi za `FEV` v odvisnosti od `Ht`, `Gender` in `Smoke` za `mod2` s parcialnimi ostanki

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 grafikoni 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
```

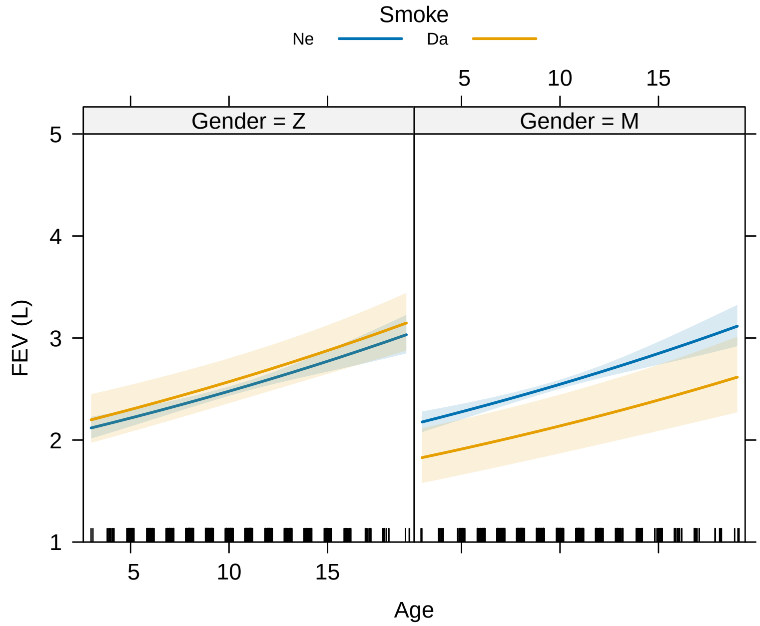
Sekvenčni F-testi pokažejo, da je ob upoštevanju vseh spremenljivk in dvojnih interakcij v modelu statistično značilna trojna interakcija `Ht:Gender:Smoke`. To pomeni, da je zveza med `Ht` in `log(FEV)` ob upoštevanju `Age` različna v štirih skupinah določenih glede na `Gender` in `Smoke` (Slika 33).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

```r
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
```

```r
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))
```

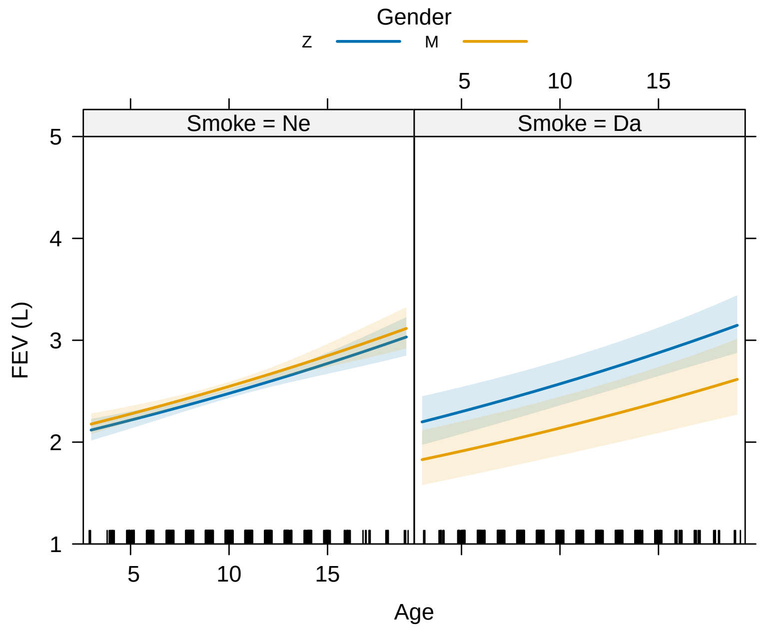


Slika 31: Modelske napovedi za `FEV` v odvisnosti od `Age`, `Gender` in `Smoke` hkrati, pri povprečni vrednosti `Ht` za `mod2.int`

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

```
[1] 155.3047
```

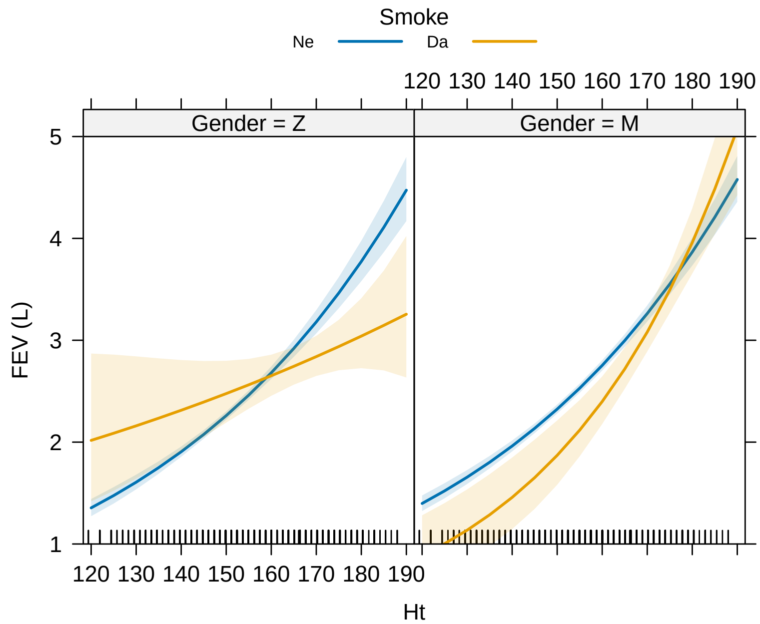
```
plot(Effect(c("Gender","Smoke","Age"), mod2.int,
            transformation = list(link = log,inverse = exp)),
     axes = list(y = list(lab = "FEV (L)", type = "response")),
     multiline=TRUE, ci.style="bands", main="", ylim=c(1, 5))
```



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(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(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(FEV)`, če se `Ht` poveča za 1 cm ob konstantnih vrednostih ostalih napovednih spremenljivk v modelu,

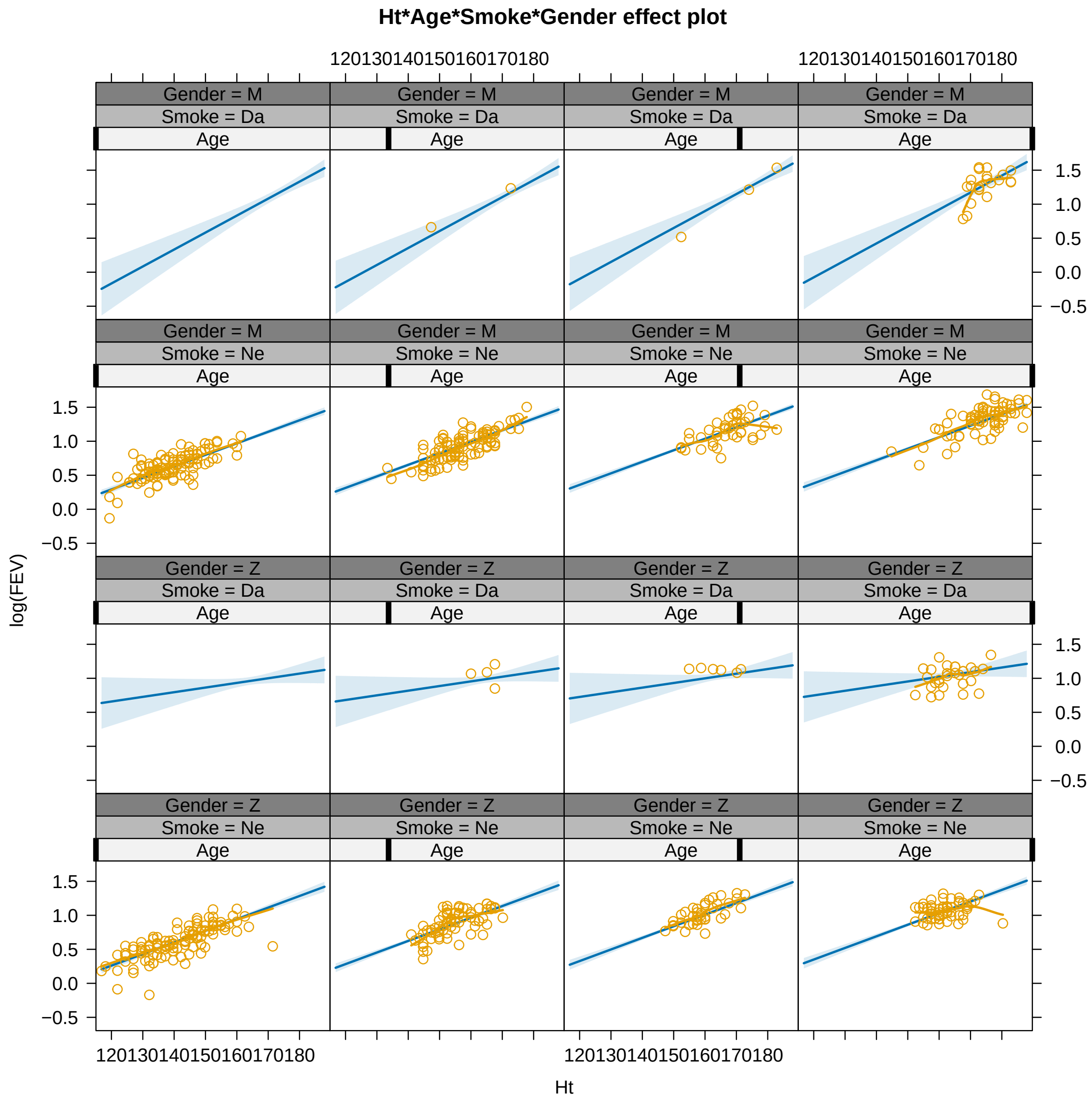
kar pomeni pri konstantni vrednosti `Age` za ženske nekadilke;

- vpliv `Ht` na povprečne vrednosti `log(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**, $GenderM = 0$ in $SmokeDa = 0$:
  $$\hat{y} = -1.967 + 0.0224 Age + 0.017 Ht.$$
  - **ženske kadilke**, $GenderM = 0$ in $SmokeDa = 1$:
  $$\hat{y} = (-1.967 + 1.625) + 0.0224 Age + (0.017 - 0.010) Ht.$$
  - **moški nekadilci**, $GenderM = 1$ in $SmokeDa = 0$:
  $$\hat{y} = (-1.967 + 0.045) + 0.0224 Age + (0.017 - 0.0001) Ht.$$
  - **moški kadilci**, $GenderM = 1$ in $SmokeDa = 1$:
  $$\hat{y} = (-1.967 + 0.045 + 1.625 - 3.042) + 0.0224 Age + (0.017 - 0.010 - 0.0001 + 0.018) Ht;$$
- $b_3 = 0.045$ 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)
```

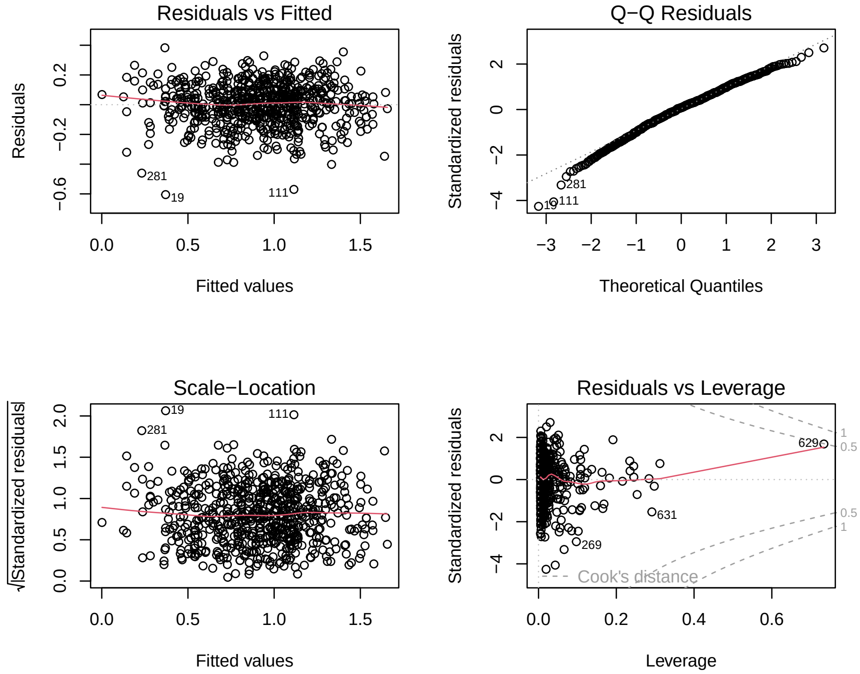
```
Analysis of Variance Table

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
```

```r
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
```

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



Slika 35: Diagnostični grafikoni za `mod2.int.vse`

```r
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?