Vaja 3: Predpostavke niso izpolnjene

Seznam potrebnih `R` paketov:

```
library(ggplot2)
library(ggpubr)
library(car)
library(effects)
library(dplyr)
library(ISLR2)
```

1. R^2 v modelu brez presečišča

R^2 v modelu s presečiščem lahko izpeljemo iz izraza, ki razdeli vsoto kvadratov odklonov odzivne spremenljivke SS_{yy} na dva dela: del SS_{model}, ki ga pojasni linearni model, ter del $SS_{residual}$, ki ostane z modelom nepojasnjen:

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

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

pri čemer smo upoštevali, da je $2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0$.

R^2 je delež variabilnosti odzivne spremenljivke, ki je pojasnjen z modelom:

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

R^2 torej primerja dani model z modelom, ki vsebuje le presečišče.

Kadar model nima presečišča, ga ni smiselno primerjati z modelom, ki vključuje le presečišče. V takem primeru gre regresijska premica namesto skozi $\overline{y}$ skozi izhodišče, in R^2 je enak:

$$R_0^2 = \frac{\sum_{i=1}^{n}\hat{y}_i^{\,2}}{\sum_{i=1}^{n} y_i^2}.$$

Vrednosti R^2 dveh modelov - s presečiščem in brez presečišča - ni mogoče neposredno primerjati, saj temeljijo na različnih izračunih. Vrednost R-kvadrata bo na splošno višja v modelu brez presečišča, vendar to nujno ne pomeni, da je ta model boljši. Načeloma model brez presečišča uporabimo le v primerih, ko iz teorije vemo, da je presečišče enako nič.

2. Posebne točke v modelu

Vrnimo se na primer iz prejšnje vaje, kjer smo na podlagi podatkovnega okvira `bodyfat` pojasnjevali odstotek telesne maščobe na podlagi 3 spremenljivk: telesne teže, višine in obsega trebuha. Še enkrat poglejmo, kako izgledajo parne povezanosti z odzivno spremenljivko. Funkcija `scatterplot` v paketu `car` omogoča identifikacijo dveh enot z največjo Mahalanobisovo razdaljo od središča podatkov.

```
scatterplot(siri ~ weight, data=bodyfat,
  smooth=list(smoother=loessLine, border=FALSE, style="none"),
  regLine=TRUE, id=TRUE, boxplots=FALSE,
  cex.axis=1.5, cex.lab=1.5)
```

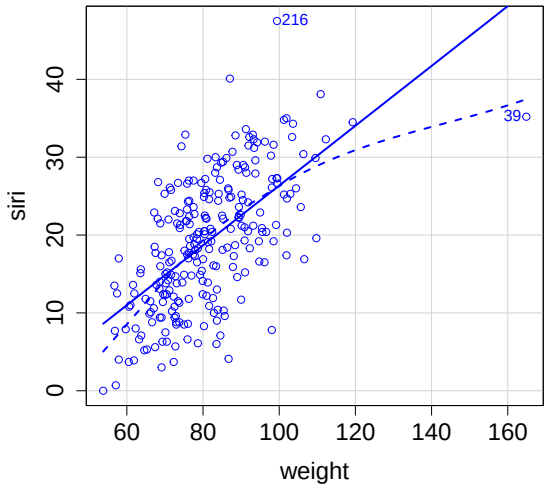
```
[1]  39 216
```

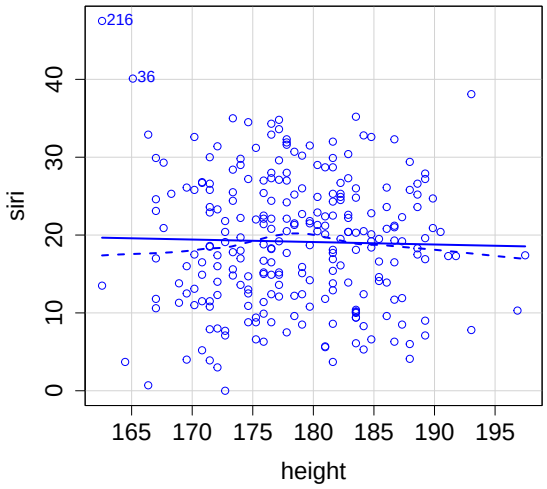
```
scatterplot(siri ~ height, data=bodyfat,
  smooth=list(smoother=loessLine, border=FALSE, style="none"),
  regLine=TRUE, id=TRUE, boxplots=FALSE,
  cex.axis=1.5, cex.lab=1.5)
```

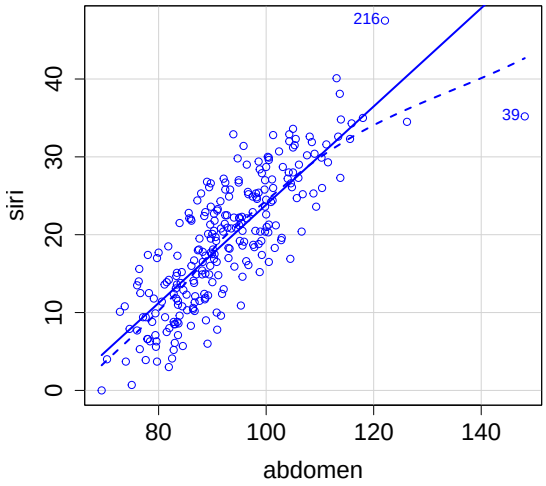
```
[1]  36 216
```

```
scatterplot(siri ~ abdomen, data=bodyfat,
  smooth=list(smoother=loessLine, border=FALSE, style="none"),
  regLine=TRUE, id=TRUE, boxplots=FALSE,
  cex.axis=1.5, cex.lab=1.5)
```

```
[1]  39 216
```





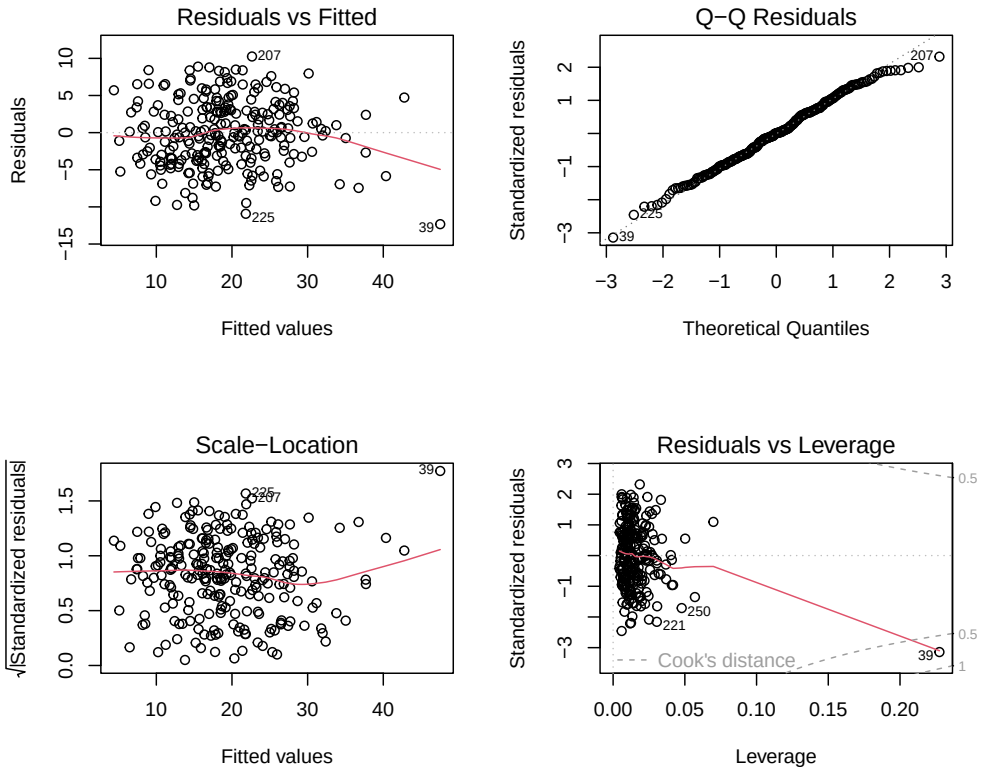


Izstopata predvsem enoti 39 in 216. V prejšnji vaji smo osebo 39 a priori izključili iz modela. Tokrat bomo naredili model na vseh 252 enotah.

```
m.bodyfat <- lm(siri~weight + height + abdomen, bodyfat)
```

Poglejmo, kako izgledajo osnovni diagnostični grafi ostankov za model `m.bodyfat`:

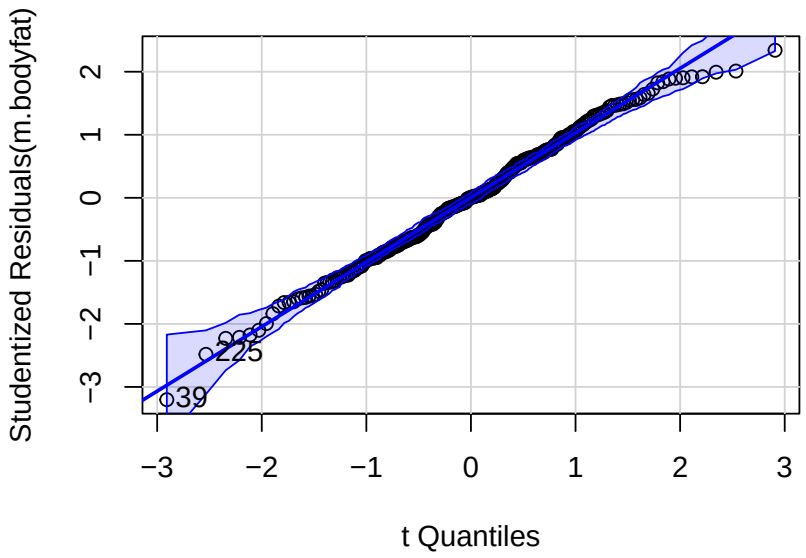
```
par(mfrow=c(2,2))
plot(m.bodyfat)
```



Slika 1: Ostanki za model `m.bodyfat`.

Izstopa 39. enota, ki ima veliko vrednost standardiziranih ostankov ter veliko vrednost Cookove razdalje. Torej bomo nadaljevali z analizo posebnih točk.

```
qqPlot(m.bodyfat)
```



Slika 2: QQ-grafikon za studentizirane ostanke za `m2` s 95 % bootstrap ovojnico.

```
[1]  39 225
```

Slika porazdelitve ostankov kaže, da imamo v modelu nekaj točk, ki imajo studentizirano ostanke po absolutni vrednosti večje od 2. Po privzetih nastavitvah funkcija identificira dve enoti z največjima vrednostima studentiziranih ostankov. Nobena točka ni daleč zunaj 95 % bootstrap ovojnice, kar kaže na to, da v modelu nimamo regresijskih osamelcev.

Naredimo še statistični test, ki temelji na studentiziranih ostankih in testira ničelno domnevo, ki pravi, da i-ta točka, $i = 1, \ldots, n$, ni regresijski osamelec:

```
outlierTest(m.bodyfat)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
39 -3.201895          0.0015444       0.3892
```

Enota 39 je potencialni regresijski osamelec, vendar pa je njena popravljena Bonferroni p-vrednost večja od 0.05. Poglejmo si, kako se dejanska vrednost `siri` pri tej enoti razlikuje od na podlagi modela napovedane vrednosti.

```
bodyfat[39, ]
```

```
   siri   weight  height abdomen
39 35.2 164.8701 183.515   148.1
```

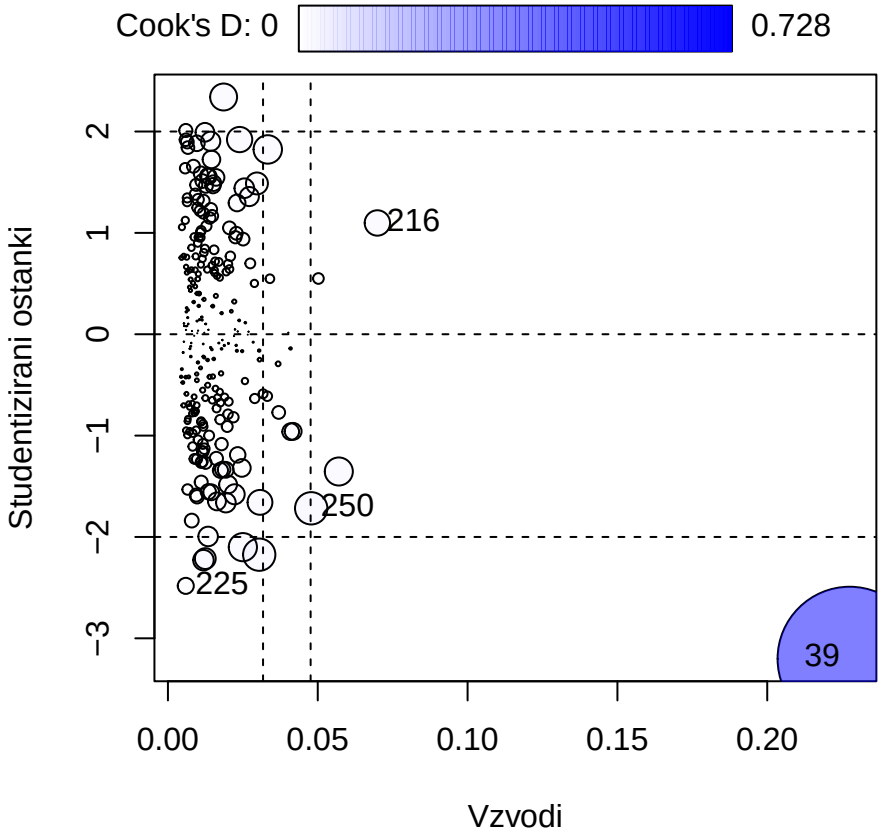
```
fitted(m.bodyfat)[39]
```

```
      39
47.52723
```

Vidimo, da za dano enoto močno precenimo odstotek telesne maščobe. Glede na maso ter obseg abdomna bi na podlagi modela za to enoto pričakovali višji odstotek telesne maščobe.

Poglejmo, ali so v modelu tudi točke, ki imajo velik vzvod:

```
influencePlot(m.bodyfat,
              id = list(method = "noteworthy", n = 2, cex = 1, location = "lr"),
              xlab = "Vzvodi", ylab = "Studentizirani ostanki")
```



Slika 3: Grafični prikaz studentiziranih ostankov, vzvodov in Cookove razdalje (ploščina kroga je sorazmerna Cookovi razdalji) za model `m.bodyfat`.

```
      StudRes         Hat       CookD
39  -3.201895 0.227490261 0.727621709
216  1.097903 0.069895940 0.022627093
225 -2.482470 0.005955063 0.009041502
250 -1.717273 0.047811915 0.036730979
```

V modelu je nekaj točk, katerih vzvod presega trikratnik povprečnega vzvoda. Vzvodne točke same po sebi še niso problem, če pa so hkrati tudi regresijski osamelci, so pogosto tudi vplivne točke. Problematična je točka 39, ki je tako vzodna točka kot tudi točka z veliko vrednostjo studentiziranega ostanka. Čeprav test za regresijske osamelce ni dal značilnega rezultata (kar je lahko tudi posledica konzervativnosti Bonferronijevega popravka), iz izpisa vidimo, da je Cookova razdalja pri tej enoti večja od 0.5. Torej bi točka lahko bila vplivna.

Primerjajmo ocene parametrov modelov z in brez enote 39.

```r
m.bodyfat_brez39 <- lm(siri~weight + height + abdomen, bodyfat[-39, ])

summary(m.bodyfat_brez39)
```

```
Call:
lm(formula = siri ~ weight + height + abdomen, data = bodyfat[-39,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-11.101  -3.309   0.023   3.230  10.050

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -31.07071   11.54742  -2.691  0.00762 **
weight       -0.21900    0.06822  -3.210  0.00150 **
height       -0.09202    0.06234  -1.476  0.14120
abdomen       0.91304    0.07171  12.732  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.38 on 247 degrees of freedom
Multiple R-squared:  0.7264,    Adjusted R-squared:  0.7231
F-statistic: 218.6 on 3 and 247 DF,  p-value: < 2.2e-16
```

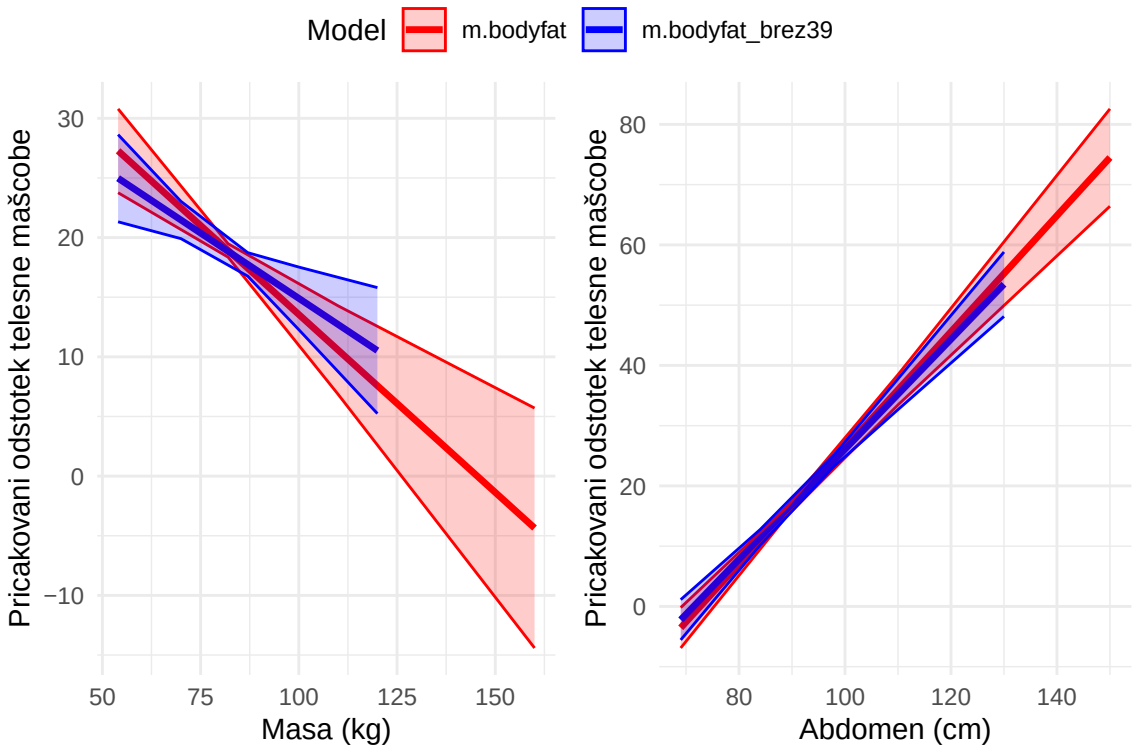
```r
compareCoefs(m.bodyfat, m.bodyfat_brez39)
```

```
Calls:
1: lm(formula = siri ~ weight + height + abdomen, data = bodyfat)
2: lm(formula = siri ~ weight + height + abdomen, data = bodyfat[-39, ])

            Model 1 Model 2
(Intercept)   -39.2   -31.1
SE             11.5    11.5

weight      -0.2984 -0.2190
SE           0.0647  0.0682

height      -0.0369 -0.0920
SE           0.0610  0.0623

abdomen      0.9635  0.9130
SE           0.0713  0.0717
```

Primerjajmo grafično pričakovani odstotek telesne maščobe v odvisnosti od `weight` oz. `abdomen` ob upoštevanju ostalih spremenljivk na podlagi obeh modelov:

```r
# Napovedi glede na weight
effect_weight_m1 <- as.data.frame(Effect("weight", m.bodyfat))
effect_weight_m2 <- as.data.frame(Effect("weight", m.bodyfat_brez39))

effect_weight_m1$Model <- "m.bodyfat"
effect_weight_m2$Model <- "m.bodyfat_brez39"
```

```r
effect_weight <- rbind(effect_weight_m1, effect_weight_m2)

p1 <- ggplot(effect_weight, aes(x = weight, y = fit, color = Model, fill = Model)) +
  geom_line(size = 1.2) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2) +
  labs(
    x = "Masa (kg)",
    y = "Pričakovani odstotek telesne maščobe"
  ) +
  theme_minimal() +
  scale_color_manual(values = c("m.bodyfat" = "red", "m.bodyfat_brez39" = "blue")) +
  scale_fill_manual(values = c("m.bodyfat" = "red", "m.bodyfat_brez39" = "blue"))

# Napovedi glede na abdomen
effect_abdomen_m1 <- as.data.frame(Effect("abdomen", m.bodyfat))
effect_abdomen_m2 <- as.data.frame(Effect("abdomen", m.bodyfat_brez39))

effect_abdomen_m1$Model <- "m.bodyfat"
effect_abdomen_m2$Model <- "m.bodyfat_brez39"

effect_abdomen <- rbind(effect_abdomen_m1, effect_abdomen_m2)

p2 <- ggplot(effect_abdomen, aes(x = abdomen, y = fit, color = Model, fill = Model)) +
  geom_line(size = 1.2) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2) +
  labs(
    x = "Abdomen (cm)",
    y = "Pričakovani odstotek telesne maščobe"
  ) +
  theme_minimal() +
  scale_color_manual(values = c("m.bodyfat" = "red", "m.bodyfat_brez39" = "blue")) +
  scale_fill_manual(values = c("m.bodyfat" = "red", "m.bodyfat_brez39" = "blue"))

ggarrange(p1, p2, ncol = 2, common.legend = TRUE, legend="top")
```



Slika 4: Napovedi za odstotek telesne maščobe v odvisnosti od `weight` oz. `abdomen` ter ob upoštevanju ostalih spremenljivk na podlagi modelov `m.bodyfat` ter `m.bodyfat_brez39`.

Vidimo, da je v modelu brez enote 39 povezanost med maso in odzivno spremeljivko `siri` ob upoštevanju ostalih spremenljivk šibkejša.

Brez dobrega poznavanja stroke pa ne moremo reči, kateri model je primerjnejši oz. ustreznejši; zato vplivnih točk ne izločamo kar tako iz modela. Pomembno je, da jih identificiramo!

3. Logaritmska transformacija

Kadar aditivnost in linearnost postaneta vprašljivi, včasih situacijo lahko popravimo z nelinearno transformacijo. Če so vrednosti odzivne spremenljivke izključno pozitivne, je vsebinsko smiselno uporabiti logaritemsko transformacijo odzivne spremenljivke. Tak linearni model postane multiplikativen na originalni skali y_i:

$$\log y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + \epsilon_i.$$

Z inverzno transformacijo dobimo

$$y_i = e^{b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + \epsilon_i} = B_0 B_1^{x_{i1}} B_2^{x_{i2}} \ldots E_i,$$

kjer so $B_0 = e^{b_0}$, $B_1 = e^{b_1}$, $B_2 = e^{b_2}, \ldots$ eksponencirani regresijski parametri, $E_0 = e^{\epsilon_i}$ pa je eksponencirana napaka. Ker je $\exp(a) > 0$, so napovedi, ki jih da tak model vedno pozitivne.

```
?Wage
head(Wage)
```

Podatkovni okvir `Wage` iz paketa `ISLR2` vsebuje podatke o bruto letnih zaslužkih (1000 $) za 3000 moških iz srednje-atlantske regije. Osredotočili se bomo na opisovanje povezanosti med plačo in starostjo ter izobrazbo. Tovrstnega modela ne bi mogli uporabiti za razlaganje vzročno-posledičnih zvez med odzivno in napovednima spremenljivkama, saj se npr. mlajši in starejši moški v vzorcu razlikujejo še v marsikateri drugi lastnosti

kot pa le po izobrazbi. Model nam lahko služi za raziskovanje *povezanosti* med zaslužkom ter starostjo in izobrazbo, gre torej za deskriptivni model. Lahko pa bi ga uporabili tudi za napovedovanje bruto letnih zaslužkov za nove enote, čeprav bi za bolj natančne napovedi imelo smisel v modelu modelirati nelinearnost.

Katero metodo ter tudi strategijo modeliranja uporabimo, je odvisno od tega, ali je naš končni cilj napovedovanje, vzročno-posledično sklepanje (inferenca) ali kombinacija obeh. Linearni model omogoča razmeroma preprosto in razumljivo interpretacijo parametrov, vendar pa bo morda dal manj natančne napovedi kot nekateri drugi nelinearni pristopi. Nasprotno pa ti pristopi lahko dajo natančnejše napovedi odzivne spremenljivke, vendar na račun fleksibilnosti dobimo manj razumljiv model, na podlagi katerega je sklepanje lahko zelo zahtevno.

```
Wage <- Wage %>%
  dplyr::select(age, education, wage) #izberemo spremenljivke, ki nas zanimajo

str(Wage)
```

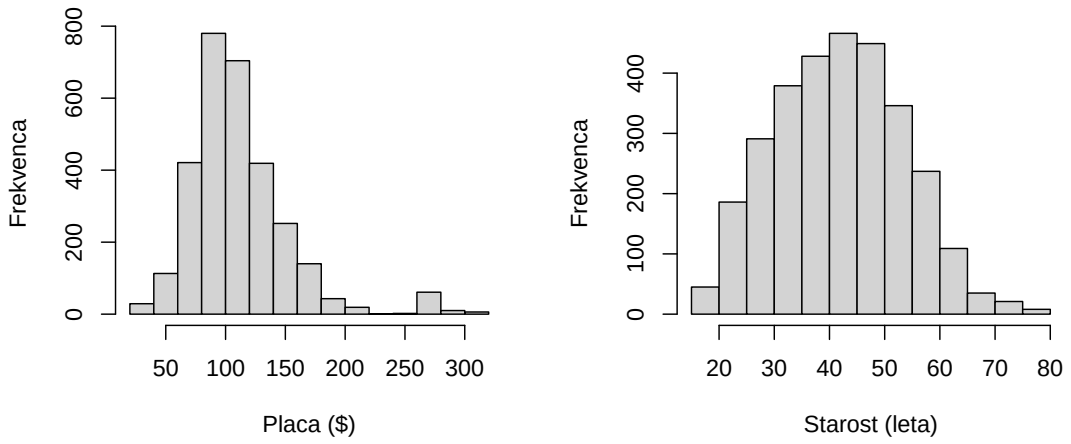
```
'data.frame':   3000 obs. of  3 variables:
 $ age      : int  18 24 45 43 50 54 44 30 41 52 ...
 $ education: Factor w/ 5 levels "1. < HS Grad",..: 1 4 3 4 2 4 3 3 3 2 ...
 $ wage     : num  75 70.5 131 154.7 75 ...
```

Poglejmo najprej univariatne porazdelitve spremenljivk v podatkovnem okviru, čeprav regresijski model ne predpostavlja ničesar o porazdelitvi napovednih spremenljivk. Te so lahko porazdeljene po porazdelitvah, ki so daleč od normalnosti. Model prav tako ne predpostavlja, da mora biti odzivna spremenljivka normalna, temveč se ta predpostavka nanaša na porazdelitev napak.

```
summary(Wage)
```

```
      age                       education        wage
 Min.   :18.00   1. < HS Grad      :268   Min.   : 20.09
 1st Qu.:33.75   2. HS Grad        :971   1st Qu.: 85.38
 Median :42.00   3. Some College   :650   Median :104.92
 Mean   :42.41   4. College Grad   :685   Mean   :111.70
 3rd Qu.:51.00   5. Advanced Degree:426   3rd Qu.:128.68
 Max.   :80.00                            Max.   :318.34
```

```
par(mfrow=c(1,2))
hist(Wage$wage, main="",
     xlab="Plača ($)", ylab="Frekvenca",
     breaks=20)

hist(Wage$age, main="",
     xlab="Starost (leta)", ylab="Frekvenca")
```

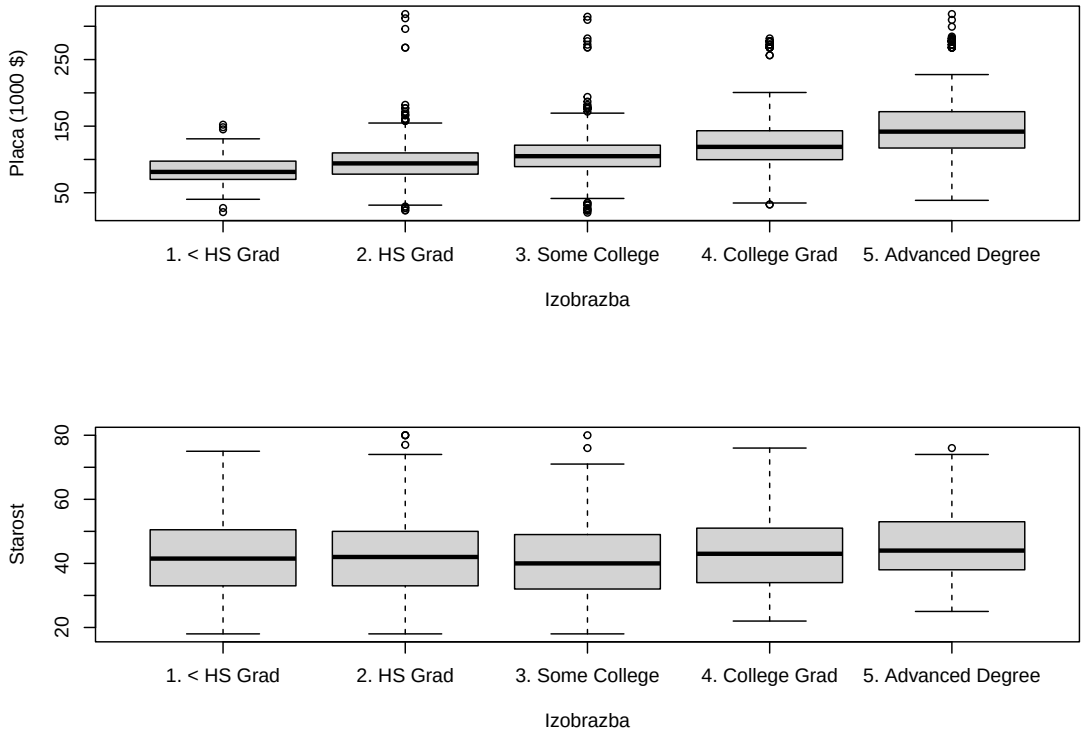


Slika 5: Univariatne porazdelitve številskih spremenljivk v podatkovnem okviru `Wage`.

Porazdelitev odzivne spremenjivke `wage` je rahlo asimetrična v desno, vrednosti `wage` pa so izključno pozitivne.

Prikažimo še porazdelitev `wage` in `age` glede na `education`:

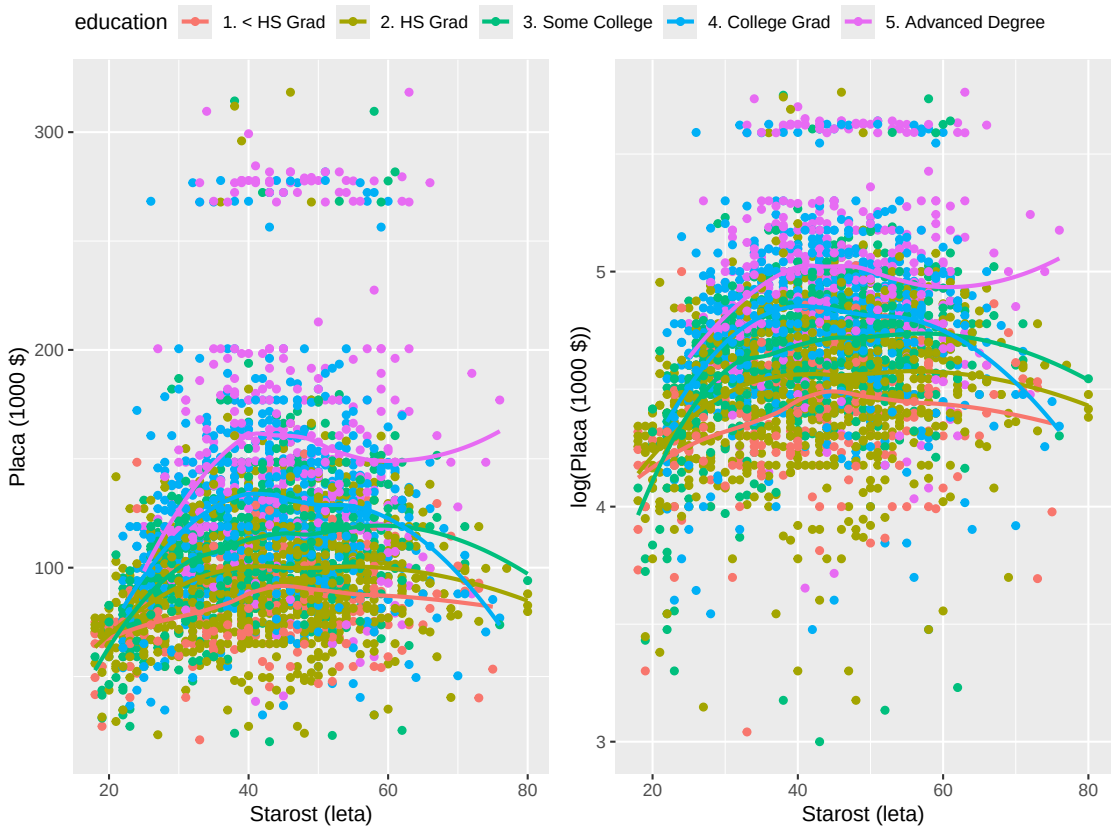
```
par(mfrow=c(2, 1))
par(cex=0.8)

boxplot(Wage$wage ~ Wage$education,
        xlab = "Izobrazba", ylab = "Plača (1000 $)")

boxplot(Wage$age ~ Wage$education,
        xlab = "Izobrazba", ylab = "Starost")
```



Slika 6: Porazdelitev `wage` in `age` glede na `education` v podatkovnem okviru `Wage`.

Poglejmo grafično, kako izgleda odvisnost `wage` od `age` po `education`:

```
#Ali obstaja linearna povezanost med spremenljivkama age in wage (glede na izobrazbo)?
p1 <- ggplot(data=Wage, aes(x=age, y=wage, col=education)) +
  geom_point() +
  geom_smooth(se=FALSE) +
  #geom_smooth(method="lm", se=FALSE) +
  xlab("Starost (leta)") +
  ylab("Plača (1000 $)")

p2 <- ggplot(data=Wage, aes(x=age, y=log(wage), col=education)) +
  geom_point() +
  geom_smooth(se=FALSE) +
  #geom_smooth(method="lm", se=FALSE) +
  xlab("Starost (leta)") +
  ylab("log(Plača (1000 $))")

ggarrange(p1, p2, ncol = 2, common.legend = TRUE, legend="top")
```



Slika 7: Odvisnost `wage` oz. `log(wage)` od `age` in `education` v podatkovnem okviru *Wage*.

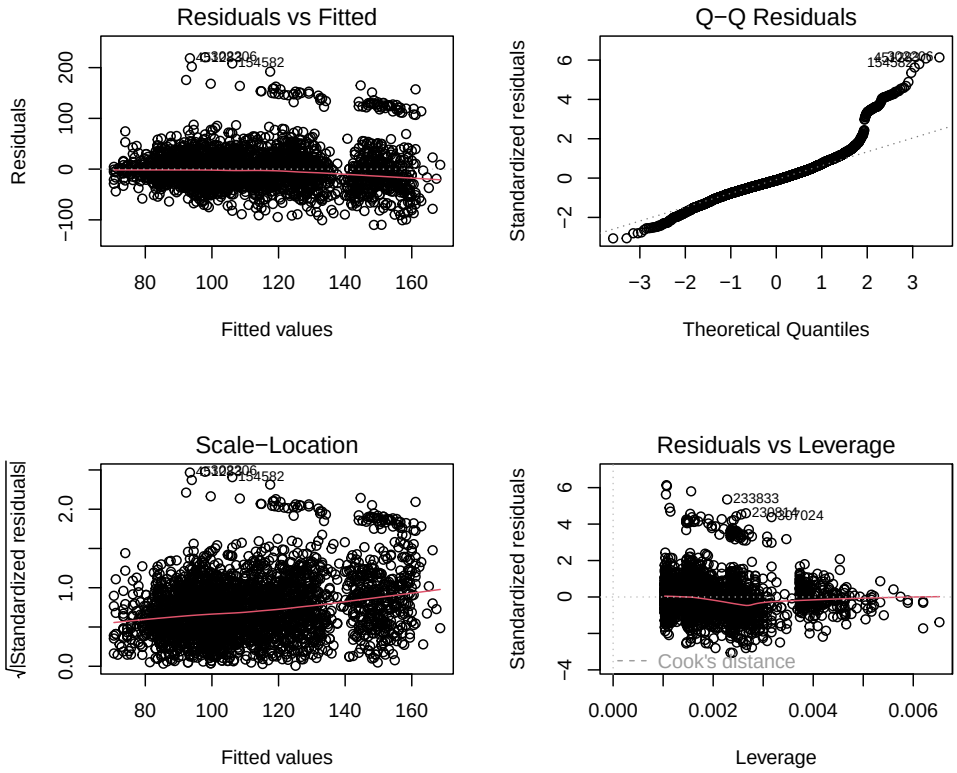
Graf nakazuje, da je vpliv `age` nelinearen in drugačen glede na `education`.

Naredimo prvi model, v katerem predpostavimo linearnost zveze med `wage` in `age` ter aditivnost vplivov `age` in `education`:

```
m0 <- lm(wage ~ age + education, data=Wage)
```

Osnovni diagnostični grafi ostankov za model `m0`:

```
par(mfrow=c(2,2))
plot(m0)
```



Slika 8: Ostanki za model `m0`.

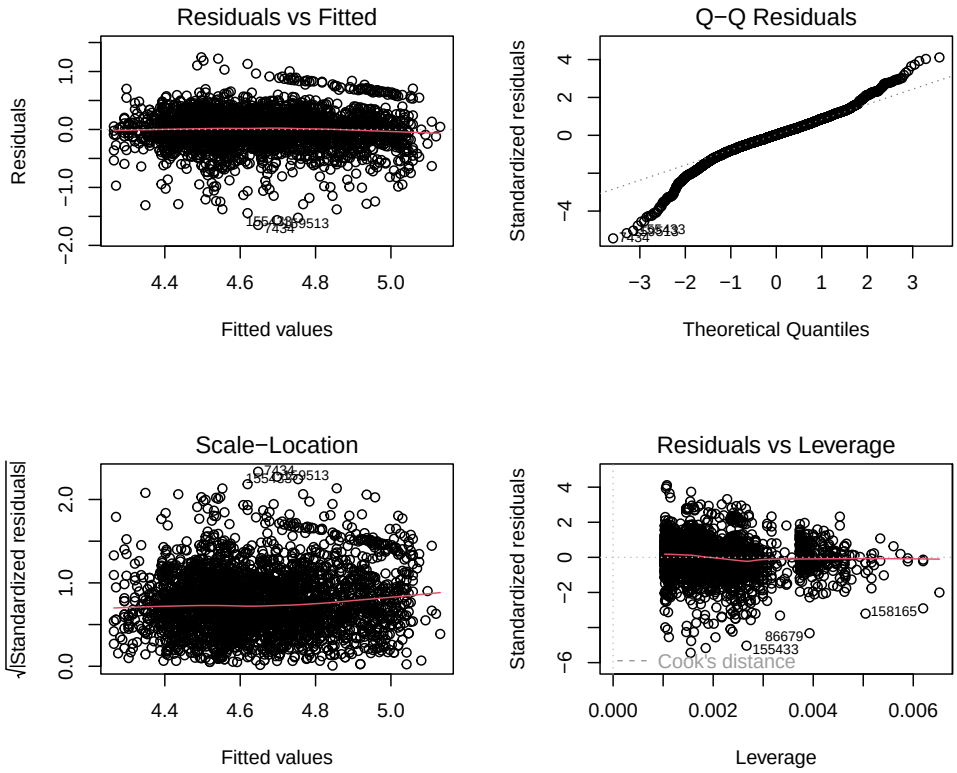
Levi sličici v prvi in drugi vrstici kažeta nekonstantno varianco. Varianca ostankov narašča z napovedanimi vrednostmi (zgornja leva sličica), slika ostankov je podobna klinu: variabilnost ostankov narašča od leve proti desni. Prisotnost nekonstantne variance še bolje pokaže gladilnik na levi spodnji sliki, kjer so na vodoravni osi napovedane vrednosti, na navpični osi pa koreni absolutnih vrednosti standardiziranih ostankov. Kot smo lahko pričakovali glede na podatke, vidimo tudi, da precej enot izstopa tudi zaradi velikih vrednosti standardiziranih ostankov.

Poglejmo, kako situacija izgleda, če odzivno spremenljivko `wage` logaritmiramo.

```
m1 <- lm(log(wage) ~ age + education, data=Wage)
```

Osnovni diagnostični grafi ostankov za model `m1`:

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

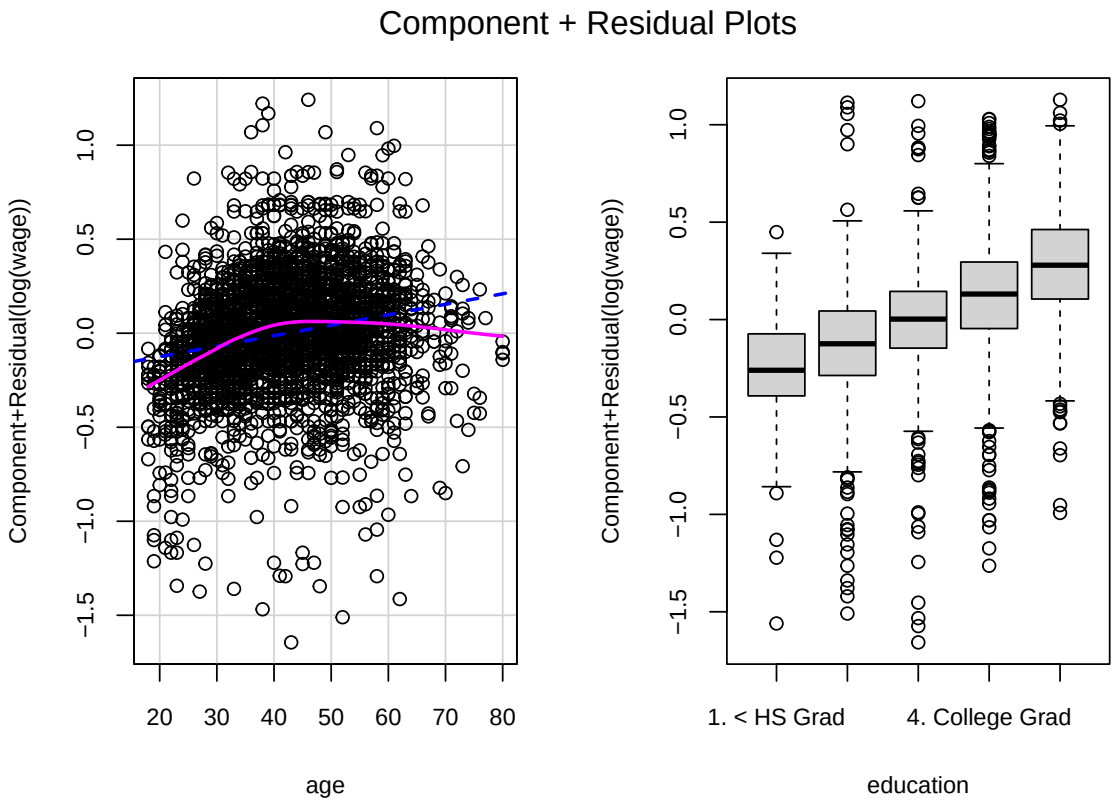


Slika 9: Ostanki za model `m1`.

Tretji grafikon nakazuje, da smo z z logaritemsko transformacijo odzivne spremenljivke heteroskedastičnost odpravili. Kot je pričakovano, pa je v podatkih še vedno precej enot z veliko vrednostjo standardiziranega ostanka. Porazdelitev le-teh tudi odstopa od standardizirane normalne porazdelitve.

Za boljšo diagnostiko modela si poglejmo grafikon parcialnih ostankov:

```
crPlots(m1, cex.lab=0.8, cex.axis=0.8)
```

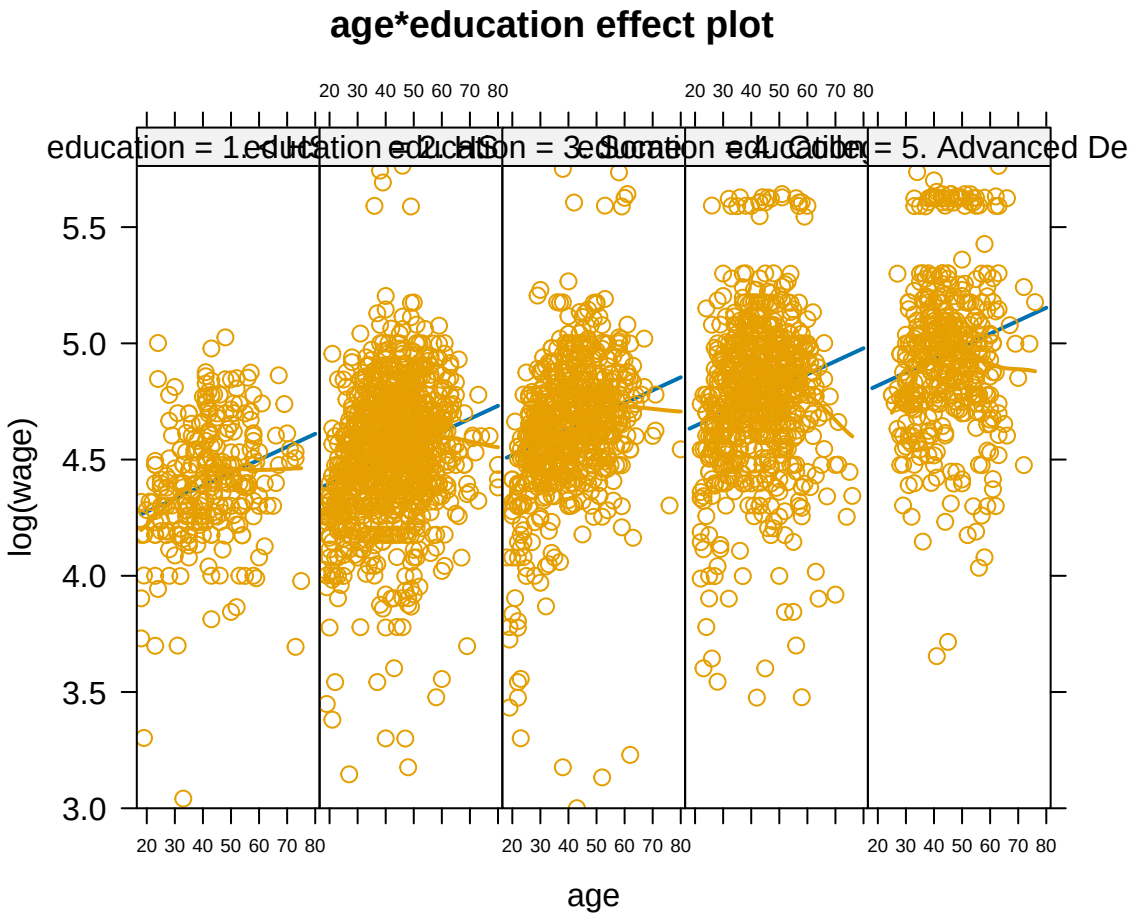


Slika 10: Graf parcialnih ostankov za model `m1`.

Iz prve sličice vidimo, da se gladilnik ne prilega dobro premici: zveza med `log(wage)` in `age` je ob upoštevanju `education` nelinearna.

Poglejmo si še grafikon parcialnih ostankov za model `m1` v odvisnosti od `age` pri različnih vrednostih spremenljivke `education`.

```
plot(Effect(c("age", "education"), m1, partial.residuals = TRUE),
     ci.style = "none",
     lattice = list(layout = c(5, 1)),
     axes = list(x=list(cex=0.6)))
```



Slika 11: Graf parcialnih ostankov za model `m1`.

Predvsem je očitna nelinearnost zveze med `log(wage)` in `age`. Pri modeliranju nelinearnosti bi si lahko pomagali s polinomsko regresijo ali zlepki, vendar več o tem kasneje.

Kljub temu, da se naš model podatkom ne prilega najbolje, bomo za vajo vseeno razmislili o interpretaciji modela, ki ima eno številsko in eno opisno spremenljivko z večimi kategorijami, odzivna spremenljivka pa je logaritmirana.

```
summary(m1)
```

```
Call:
lm(formula = log(wage) ~ age + education, data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64758 -0.15373  0.00796  0.17330  1.24577

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.1645351  0.0273609 152.208  < 2e-16 ***
age                      0.0055762  0.0004821  11.566  < 2e-16 ***
education2. HS Grad      0.1205914  0.0209082   5.768 8.86e-09 ***
```

```
education3. Some College    0.2432633  0.0219999  11.057  < 2e-16 ***
education4. College Grad    0.3682739  0.0218360  16.865  < 2e-16 ***
education5. Advanced Degree 0.5424496  0.0236742  22.913  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.303 on 2994 degrees of freedom
Multiple R-squared:  0.2592,    Adjusted R-squared:  0.258
F-statistic: 209.6 on 5 and 2994 DF,  p-value: < 2.2e-16
```

```
confint(m1)
```

```
                                2.5 %      97.5 %
(Intercept)                 4.110887058 4.218183125
age                         0.004630888 0.006521597
education2. HS Grad         0.079595544 0.161587233
education3. Some College    0.200126939 0.286399695
education4. College Grad    0.325458948 0.411088924
education5. Advanced Degree 0.496030185 0.588868967
```

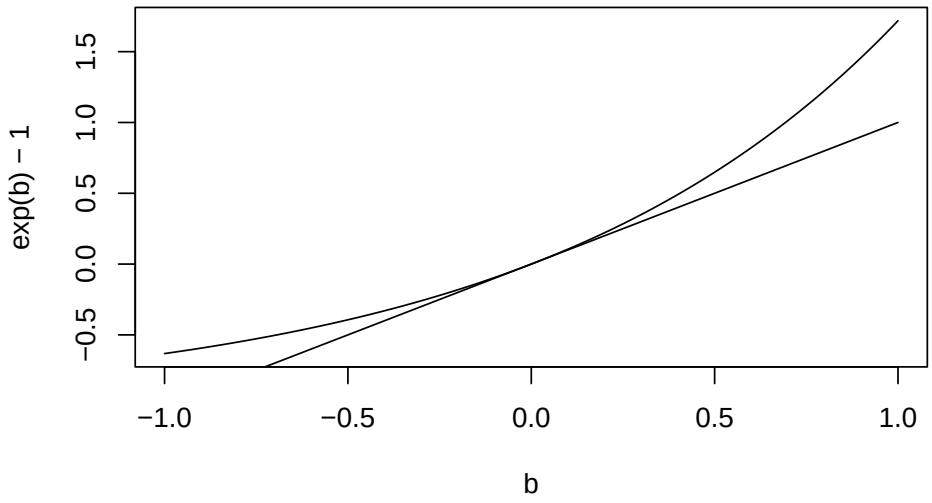
Ob upoštevanju `education` `log(wage)` narašča z `age` ($p < 0,001$), in sicer z vsakim letom starosti se v povprečju `log(wage)` poveča za 0.006 enote oz. se `wage` poveča za 0.6 % ($\exp(\beta_{age}) = 1.0055918$), pripadajoči 95 % IZ je (0.5, 0.7).

V primerjavi z osebo s `< HS Grad` in enako vrednostjo spremenljivke `age` ima v povprečju oseba s stopnjo izobrazbe:

- `2. HS Grad` `log(wage)` višjo za 0.12 enote oz. `wage` višjo za 12.8 % ($\exp(\beta_{HSGrad}) = 1.1281638$), pripadajoči 95 % IZ je (8.3, 17.5);
- `3. Some College` `log(wage)` višjo za 0.24 enote oz. `wage` višjo za 27.5 % ($\exp(\beta_{SomeCollege}) = 1.2754044$), pripadajoči 95 % IZ je (22.2, 33.2);
- `4. College Grad` `log(wage)` višjo za 0.37 enote oz. `wage` višjo za 44.5 % ($\exp(\beta_{CollegeGrad}) = 1.4452379$), pripadajoči 95 % IZ je (38.5, 50.8),
- `5. Advanced Degree` `log(wage)` višjo za 0.54 enote oz. `wage` višjo za 72 % ($\exp(\beta_{AdvancedDegree}) = 1.7202155$), pripadajoči 95 % IZ je (64.2, 80.2).

Z modelom `m1` smo uspeli pojasniti 25.92 % variabilnosti `log(wage)`.

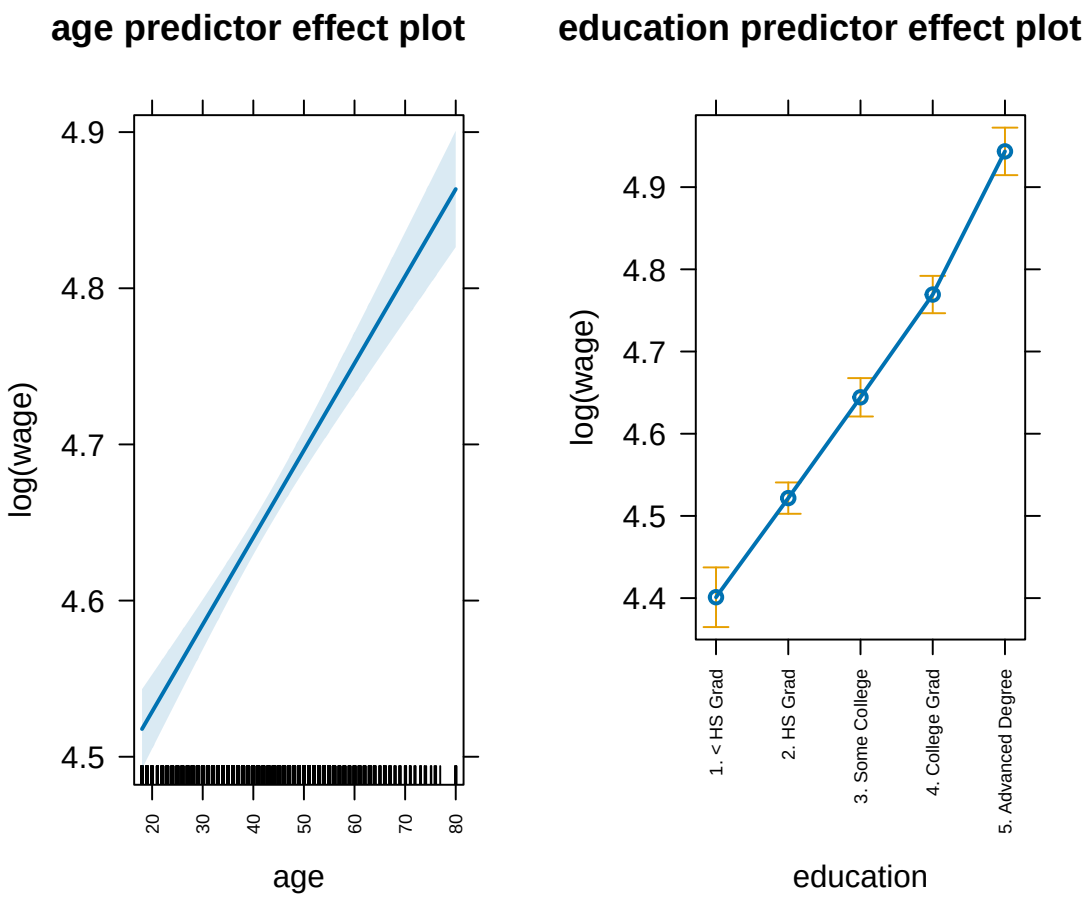
V modelu, v katerem je odzivna spremenljivka logaritmirana, so ocene regresijskih parametrov običajno majhne. Kot prikazuje spodnja slika, za majhne vrednosti približek $\exp(x) = 1 + x$ dobro aproksimira relativno razliko.



Slika 12: Interpretacija eksponenciranih regresijskih parametrov v regresijskem modelu z logaritmirano odzivno spremenljivko kot relativne razlike (ukrivljena zgornja črta) in približek $exp(x) = 1 + x$, ki velja za majhne koeficiente x.

Pri interpretaciji si lahko pomagamo z grafičnimi prikazi iz paketa `effects`:

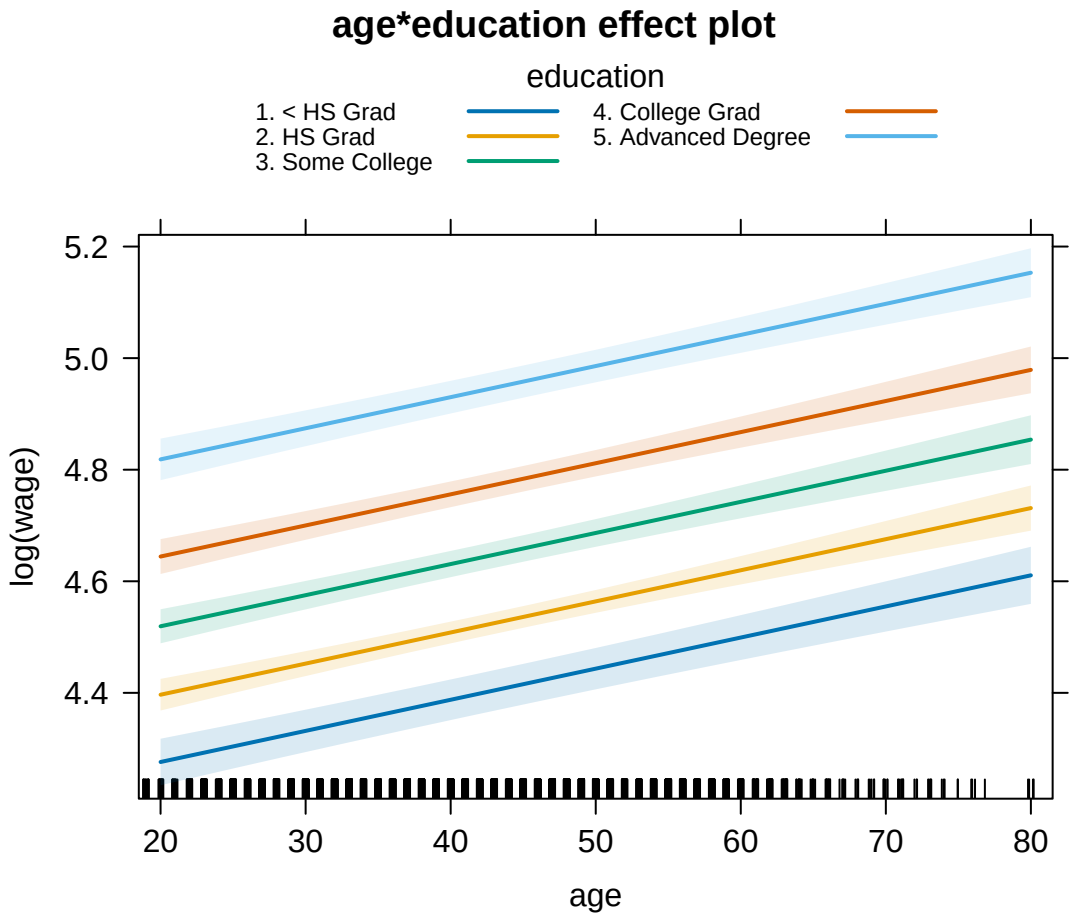
```
plot(predictorEffects(m1, ~age + education),
     axes = list(x=list(cex=0.6, rotate=90)))
```



Slika 13: Povprečne napovedi za `log(wage)` na podlagi modela `m1`, ki jih vrne funkcija `predictorEffects` za spremenljivki `age` in `education`.

oziroma:

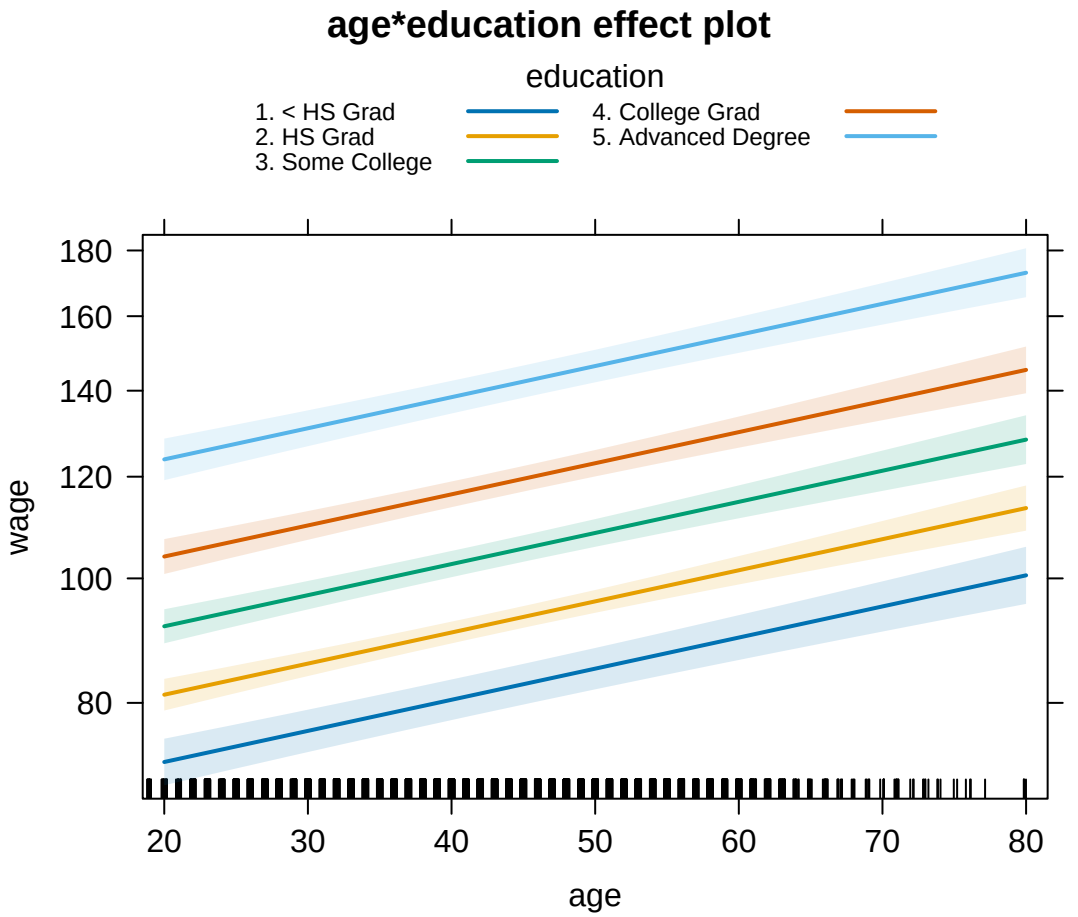
```
plot(Effect(c("age", "education"), m1), multiline=TRUE, ci.style = "bands")
```



Slika 14: Povprečne napovedi za `log(wage)` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

oz. na originalni skali spremenljivke `wage`:

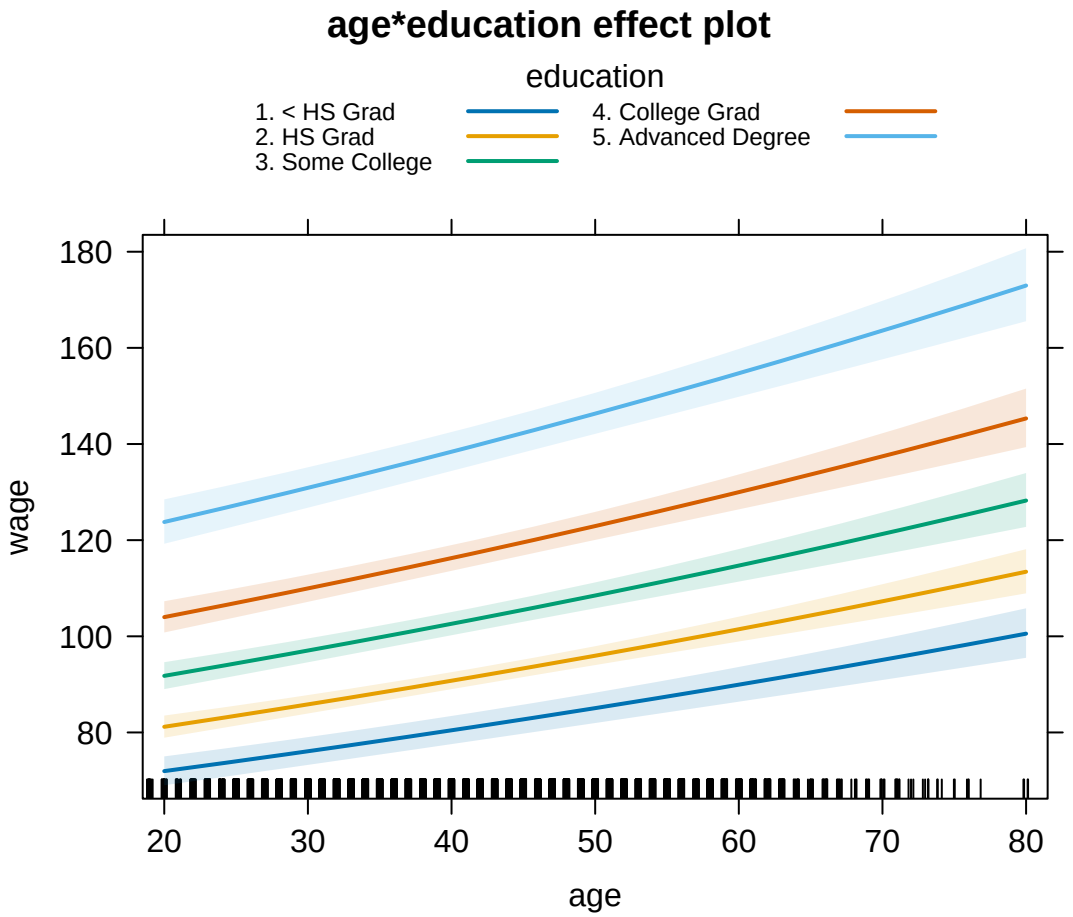
```
plot(Effect(c("age", "education"), m1, transformation = list(link = log, inverse = exp)),
     multiline=TRUE,
     ci.style = "bands",
     axes=list(y=list(lab="wage")))
```



Slika 15: Povprečne napovedi za `wage` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

ali:

```
plot(Effect(c("age", "education"), m1),
     multiline=TRUE,
     ci.style = "bands",
     axes=list(y=list(transform=exp, lab="wage")))
```



Slika 16: Povprečne napovedi za `wage` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

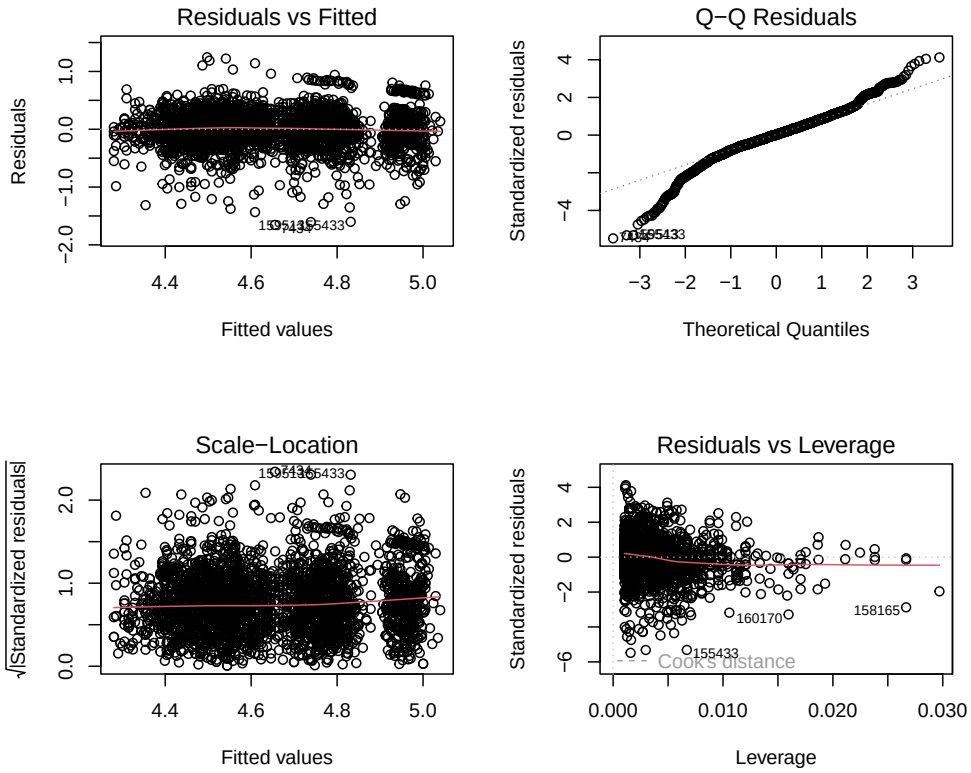
Poglejmo še, kako bi interpretirali model, ki vključuje tudi interakcijo med `age` in `education`.

```
m2 <- lm(log(wage) ~ age*education, data=Wage)

anova(m1, m2)
```

```
Analysis of Variance Table

Model 1: log(wage) ~ age + education
Model 2: log(wage) ~ age * education
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2994 274.87
2   2990 273.03  4    1.8363 5.0273 0.0004885 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-test kaže na to, da interakcija med `age` in `education` izboljša prileganje modela. Poglejmo še osnovne diagnostične grafe ostankov za model `m2`:

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



Slika 17: Ostanki za model `m2`.

Interpretacija modela `m2`:

```
summary(m2)
```

```
Call:
lm(formula = log(wage) ~ age * education, data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65539 -0.15483  0.00691  0.17417  1.24528

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.1921197  0.0640086  65.493  < 2e-16 ***
age                             0.0049162  0.0014664   3.353 0.000811 ***
education2. HS Grad             0.0979291  0.0731558   1.339 0.180791
education3. Some College        0.0644316  0.0775180   0.831 0.405937
education4. College Grad        0.4160484  0.0792801   5.248 1.65e-07 ***
education5. Advanced Degree     0.6467308  0.0918866   7.038 2.40e-12 ***
age:education2. HS Grad         0.0005434  0.0016738   0.325 0.745466
age:education3. Some College    0.0043591  0.0017917   2.433 0.015033 *
age:education4. College Grad   -0.0011018  0.0018093  -0.609 0.542593
age:education5. Advanced Degree -0.0022699  0.0020470  -1.109 0.267563
```

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

Residual standard error: 0.3022 on 2990 degrees of freedom
Multiple R-squared:  0.2642,    Adjusted R-squared:  0.262
F-statistic: 119.3 on 9 and 2990 DF,  p-value: < 2.2e-16
```

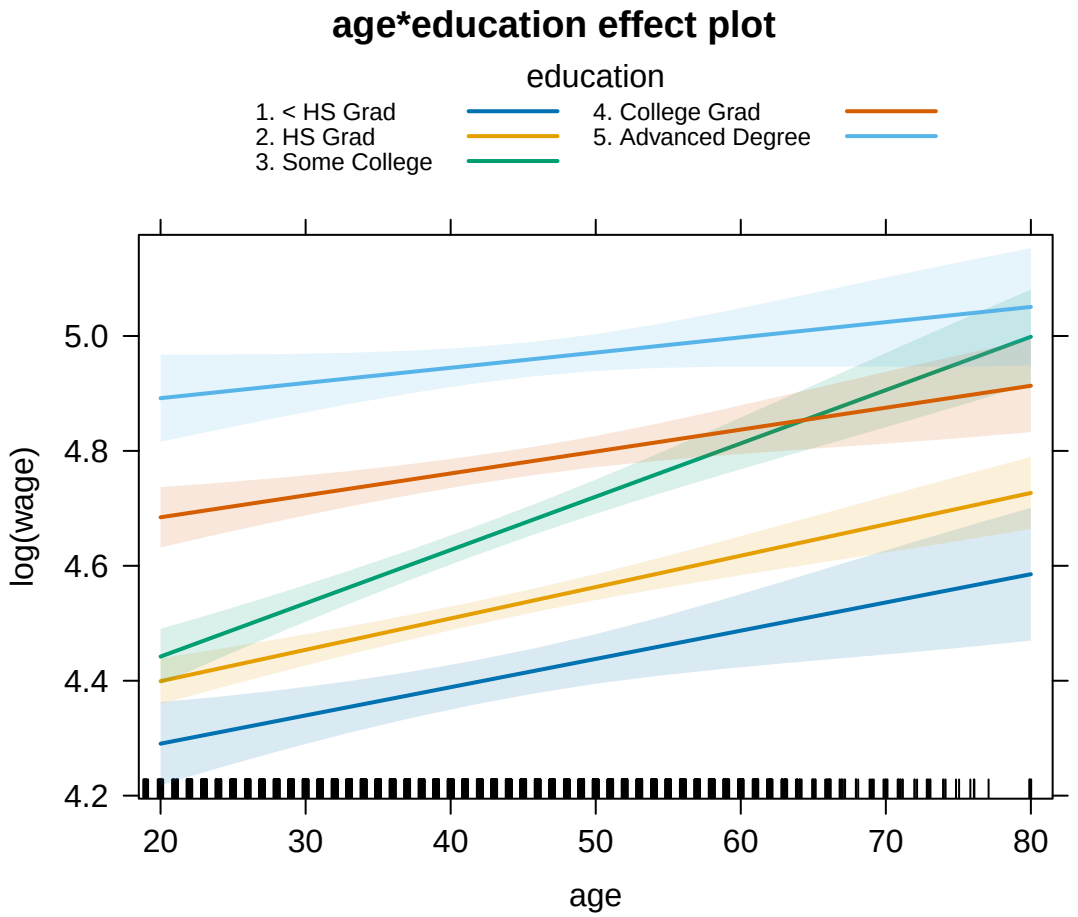
Ob upoštevanju ostalih spremenljivk v modelu je zveza med `log(wage)` in `age` drugačna glede na `education`.

- *Presečišče*: da napoved `log(wage)` za enoto, ki je stara 0 let in ima izobrazbo `1. < HS Grad`. Interpretacija v tem primeru ni smiselna, saj nobena enota ni stara 0 let.
- *Koeficient* `age`: nam pove, da `log(wage)` za `1. < HS Grad` narašča z `age` ($p = 0.001$), in sicer se z vsakim letom starosti v povprečju `log(wage)` poveča za 0.005 enote oz. se `wage` poveča za 0.5 % ($\exp(\beta_{age}) = 1.0049283$), pripadajoči 95 % IZ je (0.2, 0.8).
- *Koeficienti* za `education`: dajo povprečno razliko `log(wage)` med `1. < HS Grad` ter ostalimi stopnjami izobrazbe, če je starost enaka 0. Interpretacija v tem primeru ni smiselna, saj nobena oseba ni stara 0 let.
- *Interakcijski členi*: predstavljajo razlike v naklonih premic, ki napovedujejo `log(wage)` v odvisnosti od `age`, če primerjamo ostale stopnje izobrazbe z `1. < HS Grad`. Npr., razlika v starosti enega leta ustreza $e^{\beta_{age:2.HSGrad}} = 0.05$ % večjo razliko v zaslužku pri osebi z izobrazbo `2. HS Grad` v primerjavi z osebo `1. < HS Grad`, ocenjena napovedana razlika na leto starosti pri osebah z izobrazbo `2. HS Grad` pa je $e^{\beta_{age}+\beta_{age:2.HSGrad}} = 0.55$ %.

Model `m2` pojasni 26 % variabilnosti spremenljivke `log(wage)`.

Zveze med `log(wage)` in `age` se v modelu `m2` ne da opisati brez upoštevanja spemenljivke `education` zaradi prisotne interakcije. Če želimo npr. opisati, kako se `log(wage)` spreminja od `age` glede na `education`, si lahko pomagamo z grafičnimi prikazi, ki jih dobimo s funkcijo `predictorEffects` ali `Effect`.

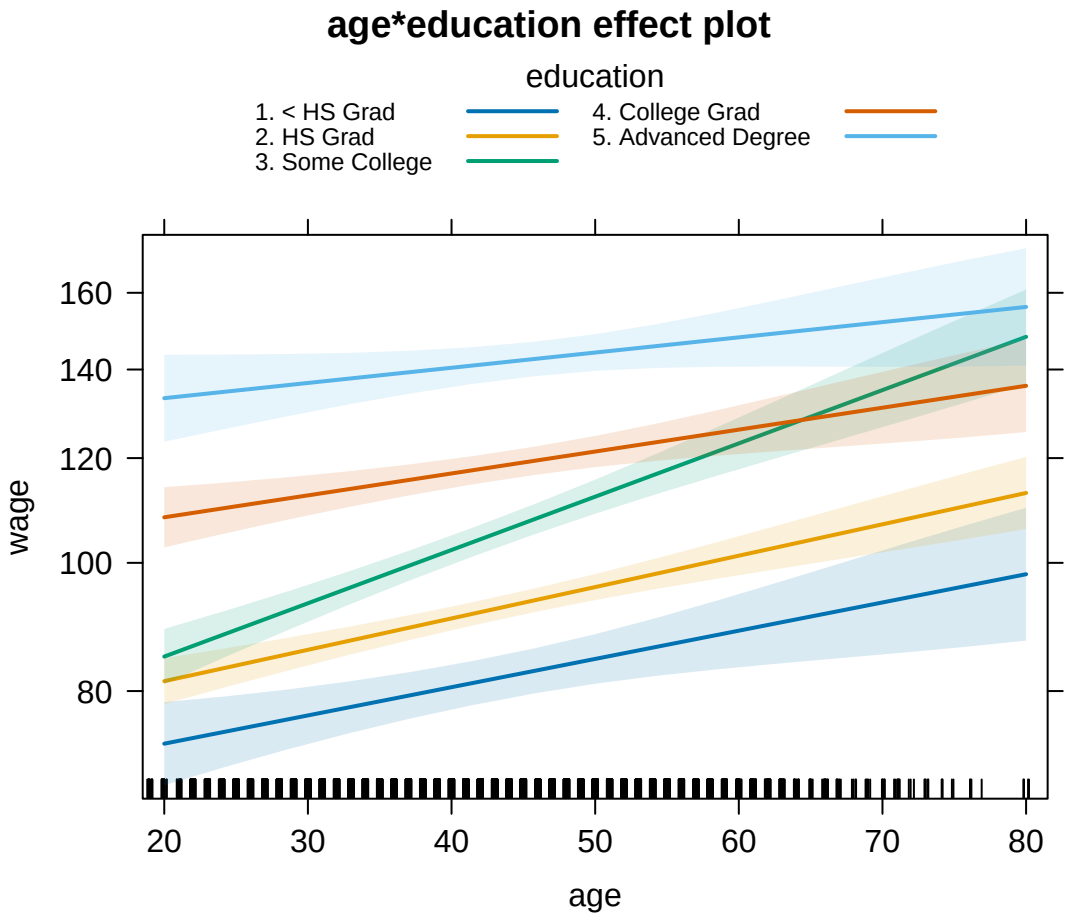
```
plot(Effect(c("age", "education"), m2), multiline=TRUE, ci.style = "bands")
```



Slika 18: Povprečne napovedi za `log(wage)` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

oz. na originalni skali spremenljivke `wage`:

```
plot(Effect(c("age", "education"), m2, transformation = list(link = log, inverse = exp)),
     multiline=TRUE,
     ci.style = "bands",
     axes=list(y=list(lab="wage")))
```



Slika 19: Povprečne napovedi za `wage` na podlagi modela `m1`, ki jih vrne funkcija `Effect` za spremenljivki `age` in `education`.

Domača naloga:

1. Interakcija dveh številskih napovednih spremenljivk:

Pripravite funkcijo za generiranje podatkov linearnega regresijskega modela:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \epsilon,$$

pri čemer naj bodo vrednosti napovednih spremenljivk porazdeljene:

- $x_{1,i} \sim U(50, 100)$ in
- $x_{2,i} \sim U(50, 100)$.

Generirajte podatke za naslednje parametre:

- $\beta_0 = 10$,
- $\beta_1 = 0.5$,
- $\beta_2 = 0.15$,
- $\beta_3 \in \{0, 0.05, 0.5\}$,
- $\sigma = 10$,
- $n = 100$.

Navodilo:

- Naredite linearni regresijski model, ki ne upošteva interakcije, in preverite, če model izpolnjuje predpostavke (pomagajte si z grafi ostankov, grafi dodane spremenljivke in grafi parcialnih ostankov). Zapišite ugotovitve.
- V linearni regresijski model vključite interakcije med napovednima spremenljivkama in preverite, če model izpolnjuje predpostavke (pomagajte si z grafi ostankov, grafi dodane spremenljivke in grafi parcialnih ostankov). Zapišite ugotovitve.

2. Log-log model:

V datoteki SNOWGEESE.txt so podatki o 45 jatah, za kateri so po dveh različnih metodah (`obs1`, `obs2`) ocenili število gosk v vsaki jati posebej. Za vse jate so zaradi posnete fotografije lahko prešteli dejansko število gosk (`photo`).

Opazujemo dejansko število gosk v jati (`photo`) v odvisnosti od števila gosk v jati, preštetih po drugi metodi (`obs2`).

- Grafično prikažite odvisnost `photo` od `obs2`.
- Ali se vam zdi linearni regresijski model primeren? Naredite linearni regresijski model `model.goske` za prvotne podatke.
- Naredite diagnostiko ostankov modela. Ali so vse predpostavke linearnega modela izpolnjene? Obrazložite svojo trditev.
- Naredite model `model.goske.log`, v katerem logaritmirate odvisno spremenljivko (`photo`) in neodvisno spremenljivko (`obs2`). Naredite diagnostiko ostankov modela. Ali so vse predpostavke linearnega modela izpolnjene? Obrazložite svojo trditev.
- Naredite analizo posebnih točk za `model.goske.log`.
- Interpretirajte ocenjene parametre modela `model.goske.log`, skupaj s pripadajočimi 95% intervali zaupanja in koeficient determinacije.