Kazalo

1 NAPOVEDNA SPREMENLJIVKA JE OPISNA

V model vključimo napovedno spremenljivko x, ki je opisna in ima l vrednosti $(a_1, a_2, ...a_l)$. V takem primeru se v `lm` model vključi $l-1$ regresorjev, ki so umetne spremenljivke z vrednostmi 0 in 1 (*dummy variables*). Označili jih bomo z w_j, $j = 1, ..., l-1$.

$$w_1 = \begin{cases} 0, & x_i = a_1 \\ 1, & x_i = a_2 \\ 0, & x_i = a_3 \\ \\ 0, & x_i = a_l \end{cases}$$

in

$$w_{l-1} = \begin{cases} 0, & x_i = a_1 \\ 0, & x_i = a_2 \\ 0, & x_i = a_3 \\ & \\ 1, & x_i = a_l \end{cases}$$

Tak model ima l parametrov $\beta_0, ..., \beta_{l-1}$. Ena od vrednosti opisne spremenljivke x ima vlogo referenčne vrednosti, običajno je to a_1. Z modelom ocenjujemo povprečje odzivne spremenljivke pri referenčni vrednosti opisne napovedne spremenljivke a_1, $\beta_0 = \mu_{a_1}$ ter $l-1$ razlik med povprečji j-te skupine in referenčne skupine: $\beta_{j-1} = \mu_{a_j} - \mu_{a_1}$, $j = 2, ..., l$.

Model zapišemo

$$y_i = \beta_0 + \beta_1 w_{1i} + \beta_2 w_{2i} + ... + \beta_l w_{(l-1)i} + \varepsilon_i, \tag{3}$$

kar pomeni, da je pričakovana vrednost $E(y_i)$ enaka:

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

Parameter modela β_0 je torej povprečje odzivne spremenljivke za referenčno vrednost a_1, μ_{a_1}; β_1 je razlika povprečja odzivne spremenljivke pri vrednosti a_2 in povprečja pri vrednosti a_1, $\mu_{a_2} - \mu_{a_1}$,..., β_{l-1} je razlika $\mu_{a_l} - \mu_{a_1}$.

1.1 Opisna napovedna spremenljivka z dvema vrednostma

Če ima opisna napovedna spremenljivka x dve vrednosti (a_1, a_2), $l = 2$, in je a_1 referenčna vrednost, se v model vključi eno umetno spremenljivko $w_1 = (w_{11}, w_{12}, ..., w_{1n})$, tako da velja:

$$w_{1i} = \begin{cases} 0, & x_i = a_1 \\ 1, & x_i = a_2. \end{cases}$$

Model zapišemo

$$y_i = \beta_0 + \beta_1 w_{1i} + \varepsilon_i, \tag{6}$$

kar pomeni, da je pričakovana vrednost:

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

Parameter modela β_0 je povprečje odzivne spremenljivke za referenčno vrednost a, μ_{a1}; β_1 je razlika povprečja odzivne spremenljivke pri vrednosti a_2 in povprečja odzivne spremenljivke pri vrednosti a_1, $\mu_{a_2} - \mu_{a_1}$.

Primer: vpliv spola na povprečen SKT

```
> tlak<-read.table(file="SKT.txt", header = TRUE)
> str(tlak)

'data.frame':        69 obs. of  3 variables:
 $ spol   : Factor w/ 2 levels "m","z": 1 1 1 1 1 1 1 1 1 ...
 $ SKT    : int  158 185 152 159 176 156 184 138 172 168 ...
 $ starost: int  41 60 41 47 66 47 68 43 68 57 ...
```

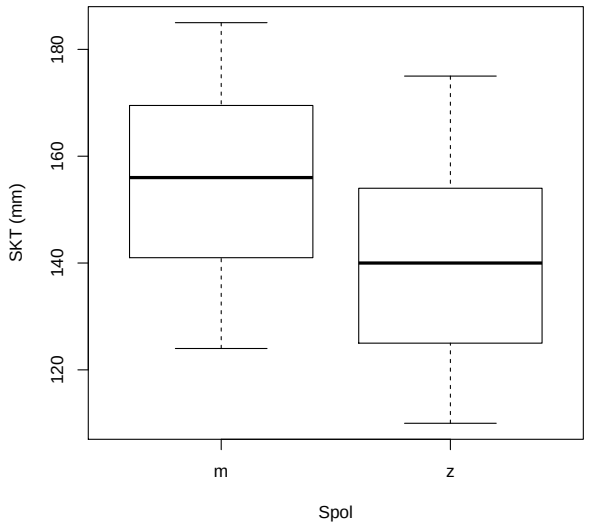
Za razumevanje nadaljnjih izpisov je pomembno, da vemo, v kakšnem vrstnem redu so v analizi urejene vrednosti opisne spremenljivke. Opisna spremenljivka za tako analizo mora biti vrste `factor`, njene vrednosti so urejene po angleški abecedi.

```
> levels(tlak$spol)

[1] "m" "z"
```

Za spremenljivko `spol` je referenčna vrednost `m`.

Zanima nas vpliv spola na SKT. Ali je povprečni SKT po spolu enak? Slika 1 prikazuje porazdelitev SKT po spolu.

```
> boxplot(SKT~spol, data=tlak, ylab=c("SKT (mm)"), xlab="Spol")
```



Slika 1: Odvisnost SKT od spola

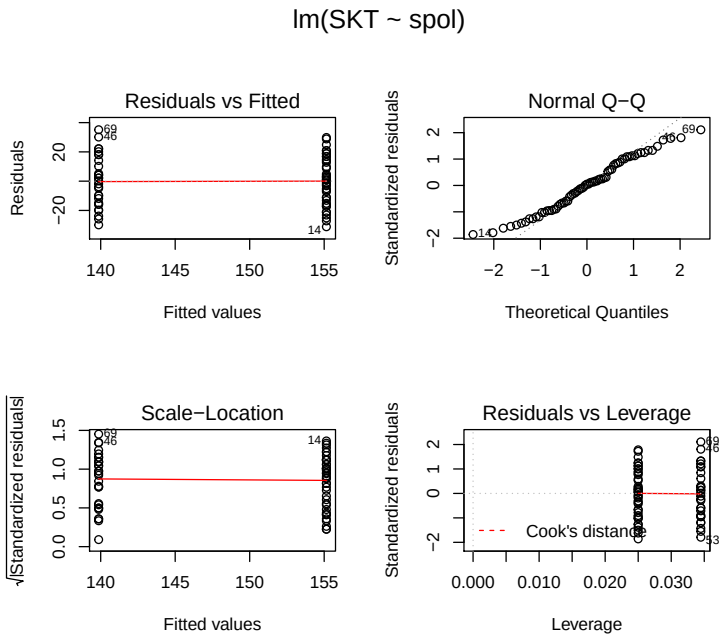
V `lm` model damo kot napovedno spremenljivko opisno spremenljivko `spol`. Funkcija `lm` na podlagi faktorja `spol` naredi regresor `spolz` z vrednostma 0 za moške in 1 za ženske.

```
> model.spol<-lm(SKT ~ spol, data=tlak)
```

Za razumevanje izpisov, ki sledijo, poglejmo vrednosti spremenljivke `spol` in modelsko matriko za `model.spol`:

```
> tlak$spol

 [1] m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m m
[39] m m z z z z z z z z z z z z z z z z z z z z z z z z z z z z z
Levels: m z

> X<-model.matrix(model.spol)
> X[1:5,]

  (Intercept) spolz
1           1     0
2           1     0
3           1     0
4           1     0
5           1     0

> X[38:42,]
```

```
   (Intercept) spolz
38           1     0
39           1     0
40           1     0
41           1     1
42           1     1
```



Slika 2: Grafični prikaz ostankov za `model.spol`

Slika ostankov (Slika 2) kaže, da je predpostavka o konstantni varianci izpolnjena. Tu primerjamo varianci napovedne spremenljivke v dveh skupinah. Da je variabilnost SKT pri moških in pri ženskah približno enaka, prikazuje tudi Slika 1. Porazdelitev ostankov v repih nekoliko odstopa od normalne porazdelitve, vendar ne toliko, da bi morali ukrepati.

```
> summary(model.spol)

Call:
lm(formula = SKT ~ spol, data = tlak)

Residuals:
   Min     1Q Median     3Q    Max
-31.15 -14.86   0.85  14.14  35.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  155.150      2.682  57.841  < 2e-16 ***
```

```
spolz        -15.288      4.138  -3.695 0.000445 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.96 on 67 degrees of freedom
Multiple R-squared:  0.1693,        Adjusted R-squared:  0.1569
F-statistic: 13.65 on 1 and 67 DF,  p-value: 0.0004446
```

Povprečje pri moških (`Intercept`) je 155.2 mm in je statistično značilno različno od 0 ($p < 0.0001$). Ženske imajo za 15.3 mm nižji povprečen SKT, razlika med povprečnima SKT pri moških in pri ženskah je statistično značilna ($p = 0.0004$). `spol` pojasni 17.0 % variabilnosti SKT, standardna napaka regresije je 16.96 mm.

Intervali zaupanja za parametre modela so:

```
> confint(model.spol)
```

```
                2.5 %     97.5 %
(Intercept) 149.79601 160.503985
spolz       -23.54646  -7.029402
```

Pri 95 % zaupanju je povprečni SKT pri moških med 149.8 mm in 160.5 mm, moški imajo od 7.0 mm do 23.5 mm višji povprečni SKT kot ženske.

t-test za primerjavo dveh povprečij

Standardno se tako primerjavo izvede s t-testom. Predpostavke tega testa so: imamo dva neodvisna vzorca, v katerih analiziramo slučajno spremenljivko y, ki je v prvi populaciji porazdeljena $N(\mu_1, \sigma^2)$, v drugi populaciji pa $N(\mu_2, \sigma^2)$; varianci obeh normalnih porazdelitev sta enaki. Zanima nas, ali sta povprečni vrednosti spremenljivke y v obeh populacijah enaki.

Ničelna in alternativna domneva, ki nas zanimata, se izražata z razliko med povprečjema μ_1 in μ_2, to je $\delta = \mu_1 - \mu_2$:

H_0: $\mu_1 = \mu_2$ ali $\delta = \mu_1 - \mu_2 = 0$.
H_1: $\mu_1 \neq \mu_2$ ali $\delta = \mu_1 - \mu_2 \neq 0$.

```
> t.test(SKT~spol, alternative='two.sided', conf.level=.95, var.equal=TRUE,
+   data=tlak)
```

```
        Two Sample t-test

data:  SKT by spol
t = 3.6949, df = 67, p-value = 0.0004446
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.029402 23.546460
sample estimates:
mean in group m mean in group z
       155.1500        139.8621
```

Opomba: z argumentom `var.equal=TRUE` v funkciji `t.test` se izvede standardni t-test, ki predpostavlja enakost varianc po spolu. Welchov test je izpeljanka t-testa, ki ne predpostavlja enakosti varianc.

Primerjejte rezultate t-testa z rezultati `lm` modela. kaj je v `lm` modelu dodano?

1.2 Opisna napovedna spremenljivka z več kot dvema vrednostma

Če ima opisna spremenljivka x tri vrednosti (a_1, a_2, a_3), $l = 3$, z `lm` modelom izračunamo povprečje odzivne spremenljivke za referenčno skupino a_1 in ga primerjamo s povprečjema odzivne spremenljivke za ostali dve skupini. V tem primeru imamo v `lm` modelu dve umetni spremenljivki w_1 in w_2:

$$w_{1i} = \begin{cases} 0, & x_i = a_1 \\ 1, & x_i = a_2 \\ 0, & x_i = a_3 \end{cases}$$

in

$$w_{2i} = \begin{cases} 0, & x_i = a_1 \\ 0, & x_i = a_2 \\ 1, & x_i = a_3. \end{cases}$$

Model zapišemo

$$y_i = \beta_0 + \beta_1 w_{1i} + \beta_2 w_{2i} + \varepsilon_i, \tag{10}$$

kar pomeni, da je pričakovana vrednost $E(y_i)$ enaka:

$$E(y_i) = \begin{cases} \beta_0, & x_i = a_1 \\ \beta_0 + \beta_1, & x_i = a_2 \\ \beta_0 + \beta_2, & x_i = a_3. \end{cases}$$

Parameter modela β_0 je povprečje odzivne spremenljivke za referenčno vrednost a_1, μ_{a_1}; β_1 je razlika povprečja odzivne spremenljivke pri vrednosti a_2 in povprečja pri vrednosti a_1, $\mu_{a_2} - \mu_{a_1}$; β_2 je razlika $\mu_{a_3} - \mu_{a_1}$.

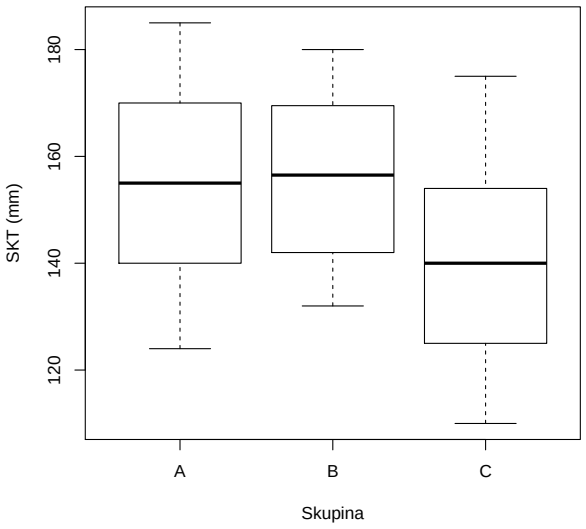
Primer: za podatke SKT dodamo izmišljeno spremenljivko `skupina` s tremi ravnmi A, B in C. V skupini A je prva polovica moških, v skupini B je druga polovica moških, v skupini C pa so ženske.

```
> tlak$skupina<-factor(rep(c("A", "B","C"), times=c(20,20,29)))
> tlak$skupina
```

```
 [1] A A A A A A A A A A A A A A A A A A A A B B B B B B B B B B B B B B B B B B
[39] B B C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C
Levels: A B C
```

```
> tapply(tlak$SKT, tlak$skupina,  mean, na.rm=TRUE)
```

```
       A        B        C
154.8500 155.4500 139.8621
```



Slika 3: `SKT` v odvisnosti od `skupina`

Uporabimo funkcijo `lm` in ocenimo tri parametre modela:

```
> model.vec<-lm(SKT~skupina, data=tlak)
> X<-model.matrix(model.vec)
> X[18:23,]
```

```
   (Intercept) skupinaB skupinaC
18           1        0        0
19           1        0        0
20           1        0        0
21           1        1        0
22           1        1        0
23           1        1        0
```

```
> X[38:43,]
```

```
   (Intercept) skupinaB skupinaC
38           1        1        0
```

```
39          1        1        0
40          1        1        0
41          1        0        1
42          1        0        1
43          1        0        1
```

Ker je v model dejansko vključena ena napovedna spremenljivka s tremi vrednostmi, na podlagi katere naredimo dve umetni spremenljivki, moramo statistično značilnost vpliva te spremenljivke najprej preveriti z F-testom, ki ga najdemo v tretjem delu povzetka modela, celotno tabelo analize variance pa naredimo z ukazom `anova`. Preverjamo ničelno domnevo

$H_0 : \beta_1 = \beta_2 = 0.$

```
> anova(model.vec)

Analysis of Variance Table

Response: SKT
          Df  Sum Sq Mean Sq F value   Pr(>F)
skupina    2  3932.8 1966.41  6.7319 0.002185 **
Residuals 66 19278.9  292.11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To ničelno domnevo zavrnemo, kar pomeni, da ima smisel pogledati rezultate v povzetku modela.

V praksi se lahko zgodi, da ničelno domnevo $H_0 : \beta_1 = \beta_2 = 0$ obdržimo, rezultati v povzetku modela pa dajejo statistično značilne razlike med povprečji. Takih rezultatov ne smemo upoštevati kot statistično značilne.

```
> summary(model.vec)

Call:
lm(formula = SKT ~ skupina, data = tlak)

Residuals:
   Min     1Q Median     3Q    Max
-30.85 -14.86   0.55  14.14  35.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  154.850      3.822  40.519  < 2e-16 ***
skupinaB       0.600      5.405   0.111  0.91194
skupinaC     -14.988      4.968  -3.017  0.00362 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.09 on 66 degrees of freedom
Multiple R-squared:  0.1694,        Adjusted R-squared:  0.1443
F-statistic: 6.732 on 2 and 66 DF,  p-value: 0.002185
```

Ničelne domneve v povzetku `lm` modela so:

H_0: $\beta_0 = \mu_A = 0$, H_0: $\beta_1 = \mu_B - \mu_A = 0$ in H_0: $\beta_2 = \mu_C - \mu_A = 0$.

V povzetku `lm` modela so te domneve testirane z navadnim t-testom, ki ne upošteva hkratnosti primerjav, zato so izračunane p-vrednosti le informativne. V nadaljevanju bomo spoznali, kako v testiranju domnev upoštevamo hkratnost primerjav.

2 INTERVALI ZAUPANJA ZA PARAMETRE MODELA IN TESTIRANJE DOMNEV

V linearnem modelu odzivno spremenljivko y modeliramo na podlagi k regresorjev

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{12}$$

kar zapišemo v matrični obliki takole:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{13}$$

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

V (13) je $\mathbf{y}$ vektor odzivne spremenljivke, $\mathbf{X}$ je modelska matrika reda $(n \times k + 1)$, $\boldsymbol{\beta}$ je vektor parametrov modela reda $(k + 1 \times 1)$ in $\boldsymbol{\varepsilon}$ je vektor napak reda $(n \times 1)$, za katerega velja $E(\boldsymbol{\varepsilon}) = 0$ in $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $\mathbf{I}$ je enotska diagonalna matrika reda $n \times n$.

Ocene parametrov izračunamo po metodi najmanjših kvadratov (OLS) ali po metodi največjega verjetja (ML):

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{14}$$

Lahko pokažemo, da so tako dobljene ocene parametrov nepristranske

$$E(\mathbf{b}) = \boldsymbol{\beta}. \tag{15}$$

Če je matrika $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ nesingularna, kar pomeni, da gre za t. i. model polnega ranga, je varianca ocen parametrov

$$Var(\mathbf{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}. \tag{16}$$

Za ocene parametrov linearnega modela **b**, dobljene po metodi najmanjših kvadratov, lahko pokažemo, da so **najboljše linearne nepristranske cenilke** za $\boldsymbol{\beta}$ (BLUE, *Best Linear Unbiesed Estimator*), kar pomeni, da imajo med linearnimi cenilkami najmanjšo varianco (Gauss-Markov izrek).

Pokažemo lahko, da je nepristranska cenilka za σ^2

$$\hat{\sigma}^2 = \frac{1}{n-k-1}(\mathbf{y}-\mathbf{Xb})^{\mathrm{T}}(\mathbf{y}-\mathbf{Xb}). \tag{17}$$

2.1 Interpretacija ocen parametrov linearnega modela z več regresorji

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

$$E(y|x_{01}, x_{02}) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02}, \tag{18}$$

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

$$E(y|x_{01}, x_{02}+1) = \beta_0 + \beta_1 x_{01} + \beta_2(x_{02}+1) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_2, \tag{19}$$

iz (18) in (19) sledi

$$\beta_2 = E(y|x_{01}, x_{02}+1) - E(y|x_{01}, x_{02}). \tag{20}$$

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

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

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

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

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

2.2 Inferenca v linearnem modelu

Inferenca v linearnem modelu polnega ranga temelji na tem, da so ocene parametrov **b** porazdeljene po multivariatni normalni porazdelitvi

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \quad \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}). \tag{21}$$

Poleg tega velja, da so ocene prametrov **b** neodvisne od ocene variance napak $\hat{\sigma}^2$. Porazdelitev statistike $(n-k-1)\hat{\sigma}^2/\sigma^2$ je χ^2-porazdelitev s stopinjami prostosti $SP = n-k-1$:

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}. \tag{22}$$

2.3 Intervalne ocene za parametre modela

Interval zaupanja za posamezen parameter modela β_j, $j = 0, ..., k$, **ob upoštevanju ostalih regresorjev v modelu** imenujemo **parcialni interval zaupanja**. Definiran je na podlagi statistike $\frac{b_j-\beta_j}{\hat{\sigma}\sqrt{a_{jj}}}$, kjer je $\sqrt{a_{jj}}$ diagonalni element matrike $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, to matriko označimo **A**. Velja, da je statistika:

$$\frac{b_j - \beta_j}{\sigma\sqrt{a_{jj}}} \sim N(0,1). \tag{23}$$

Ko izraz (23) delimo s korenom spremenljivke (22) deljene z n-k-1, kjer je (22) porazdeljena po χ^2_{n-k-1} porazdelitvi, dobimo statistiko, ki je porazdeljena po t-porazdelitvi s $SP = n-k-1$:

$$\frac{b_j - \beta_j}{\sigma\sqrt{a_{jj}}} \Big/ \sqrt{\frac{\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2}}{n-k-1}} \sim t(SP = n-k-1), \tag{24}$$

ko zgornji izraz poenostavimo, dobimo

$$\frac{b_j - \beta_j}{\hat{\sigma}\sqrt{a_{jj}}} \sim t(SP = n-k-1). \tag{25}$$

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

$$(b_j - t_{\alpha/2}(SP = n-k-1)\hat{\sigma}\sqrt{a_{jj}}, \; b_j + t_{\alpha/2}(SP = n-k-1)\hat{\sigma}\sqrt{a_{jj}}). \tag{26}$$

Funkcija `confint()` vrne parcialne 95 % intervale zaupanja za vse parametre v modelu. Pri njihovi interpretaciji se je potrebno zavedati, da so to intervali zaupanja za posamezen parameter ob upoštevanju vseh ostalih členov v modelu.

2.3.1 Testiranje domnev o parametrih modela

Za testiranje ničelne domneve $H_0 : \beta_j = \gamma$ ob alternativni domnevi $H_0 : \beta_j \neq \gamma$ in ob upoštevanju vseh ostalih členov v modelu, uporabimo testno statistiko

$$t = \frac{b_j - \gamma}{\hat{\sigma}\sqrt{a_{jj}}}, \tag{27}$$

ki je pod ničelno domnevo porazdeljena po Studentovi porazdelitvi s $SP = n - k - 1$.

Rezultate testiranja posamičnih $k + 1$ ničelnih domnev za parametre modela dobimo v povzetku `lm` modela.

2.3.2 Območje zaupanja za vse parametre linearnega modela

Ker so ocene parametrov linearnega modela porazdeljene po multivariatni normalni porazdelitvi, lahko pokažemo, da $100(1-\alpha)$ % **območje zaupanja za vse parametre modela hkrati** določimo na podlagi F-statistike:

$$F = \frac{(\mathbf{b} - \boldsymbol{\beta})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{b} - \boldsymbol{\beta})}{(k+1)\hat{\sigma}^2}, \tag{28}$$

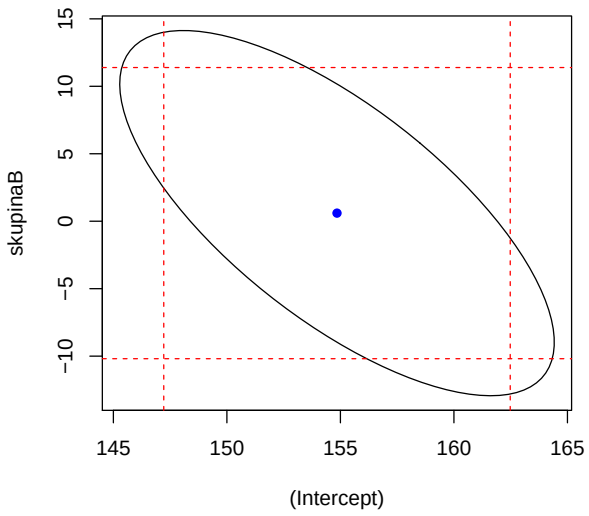
ki je porazdeljena po F-porazdelitvi s stopinjami prostosti $SP_1 = k + 1$ in $SP_2 = n - k - 1$.

$100(1 - \alpha)$ % območje zaupanja za $\boldsymbol{\beta}$ predstavlja vse vrednosti za $\boldsymbol{\beta}$, ki ustrezajo pogoju

$$P\left(\frac{(\mathbf{b} - \boldsymbol{\beta})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{b} - \boldsymbol{\beta})}{(k+1)\hat{\sigma}^2} \leq F_\alpha(k+1, n-k-1)\right) = 1 - \alpha. \tag{29}$$

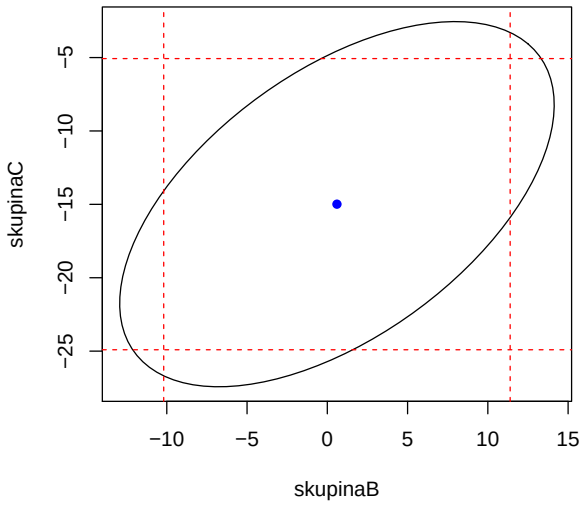
Slika 4 prikazuje območje zaupanja za parametra β_0 in β_1 ob upoštevanju β_2 za model `model.vec`. Slika 5 pa območje zaupanja za parametra β_1 in β_2 ob upoštevanju β_0 za `model.vec`. Na obeh slikah sta prikazana tudi parcialna intervala zaupanja.

```
> library(ellipse); plot(ellipse(model.vec, which=c(1,2)), type="l")
> abline(v=confint(model.vec)[1,], h=confint(model.vec)[2,], lty=2, col="red")
> points(model.vec$coef[1],model.vec$coef[2], pch=16, col="blue")
```



Slika 4: Primer 95 % območja zaupanja za parametra β_0 in β_1 ob upoštevanju β_2 za model `model.vec` (elipsa) in meje 95 % parcialnih intervalov zaupanja za β_0 in za β_1 iz istega modela (črtkane črte)

```
> plot(ellipse(model.vec, which=c(2,3)), type="l")
> abline(v=confint(model.vec)[2,], h=confint(model.vec)[3,], lty=2, col="red")
> points(model.vec$coef[2],model.vec$coef[3], pch=16, col="blue")
```



Slika 5: Primer 95 % območja zaupanja za parametra β_1 in β_2 ob upoštevanju β_0 za model `model.vec` (elipsa) in meje 95 % parcialnih intervalov zaupanja za β_1 in za β_2 iz istega modela

Za ilustracijo izračunajmo variančno kovariančno matriko za `model.vec` s funkcijo `vcov` in po (16):

```
> vcov(model.vec)

            (Intercept)  skupinaB  skupinaC
(Intercept)    14.60526 -14.60526 -14.60526
skupinaB      -14.60526  29.21053  14.60526
skupinaC      -14.60526  14.60526  24.67786

> n <- length(model.vec$residuals)
> b <- model.vec$coef
> k <- length(b)-1
> (s2 <- sum(model.vec$residuals^2)/(n-k-1))

[1] 292.1053

> X <- model.matrix(model.vec)
> (A <- solve(t(X) %*% X))

            (Intercept) skupinaB    skupinaC
(Intercept)        0.05    -0.05 -0.05000000
skupinaB          -0.05     0.10  0.05000000
skupinaC          -0.05     0.05  0.08448276
```

```
> (s2*A)

            (Intercept)  skupinaB  skupinaC
(Intercept)    14.60526 -14.60526 -14.60526
skupinaB      -14.60526  29.21053  14.60526
skupinaC      -14.60526  14.60526  24.67786
```

2.3.3 Napovedi linearnega modela

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

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i. \tag{30}$$

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

$$y_* = \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_*, \tag{31}$$

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

Interval zaupanja za povprečno napoved

Napoved v točki x_* je $x_*^{\mathrm{T}}b$, njena pričakovana vrednost je

$$E(\mathbf{x}_*^{\mathrm{T}}\mathbf{b}) = \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta} \tag{32}$$

in njena varianca

$$Var(\mathbf{x}_*^{\mathrm{T}}\mathbf{b}) = \mathbf{x}_*^{\mathrm{T}}Var(\mathbf{b})\mathbf{x}_* = \sigma^2\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*, \tag{33}$$

$\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, kjer je $\mathbf{X}$ je modelska matrika. Ker je napoved $\mathbf{x}_*^{\mathrm{T}}\mathbf{b}$ linearna kombinacija normalno porazdeljenih spremenljivk, tudi zanjo velja, da je porazdeljena normalno

$$\mathbf{x}_*^{\mathrm{T}}\mathbf{b} \sim N(\mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta}, \quad \sigma^2\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*). \tag{34}$$

Velja

$$\frac{\mathbf{x}_*^{\mathrm{T}}\mathbf{b} - \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}} \sim t(SP = n - k - 1), \tag{35}$$

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

$$\left(\mathbf{x}_*^{\mathrm{T}}\mathbf{b} - t_{1-\frac{\alpha}{2}}(SP = n - k - 1)\hat{\sigma}\sqrt{\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}, \quad \mathbf{x}_*^{\mathrm{T}}\mathbf{b} + t_{1-\frac{\alpha}{2}}(SP = n - k - 1)\hat{\sigma}\sqrt{\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}\right). \tag{36}$$

Interval zaupanja za posamično napoved

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

$$y_* - \hat{y}_* = \mathbf{x}_*^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_* - \mathbf{x}_*^{\mathrm{T}}\mathbf{b}. \tag{37}$$

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

$$\begin{aligned} Var(y_* - \hat{y}_*) &= Var(\mathbf{x}_*^{\mathrm{T}}\mathbf{b}) + Var(\varepsilon_*) \\ &= \sigma^2\mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_* + \sigma^2 \\ &= \sigma^2(1 + \mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*). \end{aligned} \tag{38}$$

Tudi tu lahko pokažemo, da je $y_* - \hat{y}_*$ neodvisen od $(n-k-1)\hat{\sigma}^2/\sigma^2$ in velja

$$\frac{y_* - \hat{y}_*}{\hat{\sigma}\sqrt{1 + \mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*}} \sim t(SP = n-k-1), \tag{39}$$

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

$$\left(y_* - t_{1-\frac{\alpha}{2}}(SP = n-k-1)\hat{\sigma}\sqrt{1 + \mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*},\ \hat{}\ y_* + t_{1-\frac{\alpha}{2}}(SP = n-\hat{k}-1)\hat{\sigma}\sqrt{1 + \mathbf{x}_*^{\mathrm{T}}\mathbf{A}\mathbf{x}_*} \right) \tag{40}$$

2.4 F-test za model

Pri modelu enostavne linearne regresije smo videli, da lahko ničelno domnevo $H_0 : \beta_1 = 0$ preverimo tudi na podlagi F-statistike, ki je pod ničelno domnevo porazdeljena $F(SP_1 = 1, SP_2 = n-k-1)$. Za linearni model z več napovednimi spremenljivkami na podlagi F-statistike testiramo ničelno domnevo, da so parametri $(\beta_1, \beta_2, ..., \beta_k)$ hkrati enaki nič:

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

Alternativna domneva pravi, da je vsaj en parameter β_j, $j = 1, ..., k$, različen od nič.

Vsoto kvadriranih odklonov za odzivno spremenljivko SS_{yy}, enako kot pri enostavni linearni regresiji, razdelimo na dva dela

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SS_{yy} &= SS_{model} + SSresidual \\ &= (\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - C) + (\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}), \end{aligned} \tag{41}$$

kjer je $C = (\sum_{i=1}^{n} y_i)^2/n$ je t. i. korekcijski člen. F-statistika je definirana z razmerjem

$$F = \frac{SS_{model}/k}{SS_{residual}/(n-k-1)}. \tag{42}$$

Ob predpostavki $\varepsilon \sim iid\ N(0, \sigma^2)$ je F-statistika porazdeljena po F-porazdelitvi s $SP_1 = SP_{model} = k$ in $SP_2 = SP_{residal} = n - k - 1$.

Tabela 1: Shema tabele ANOVA za spošni linearni regresijski model s k regresorji

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

Za ilustracijo izračunajmo SS_{yy}, SS_{model} in $SS_{residuals}$ za `model.vec`:

```
> t(b)

     (Intercept) skupinaB  skupinaC
[1,]      154.85      0.6 -14.98793

> (SS_yy<-sum((tlak$SKT-mean(tlak$SKT))^2))

[1] 23211.77

> (SS_model<-t(b) %*% t(X) %*% tlak$SKT-sum(tlak$SKT)^2/n)

        [,1]
[1,] 3932.82

> (SS_residual<-t(tlak$SKT) %*% tlak$SKT-t(b) %*% t(X) %*% tlak$SKT)

         [,1]
[1,] 19278.95

> (F <- (SS_model/k)/(SS_residual/((n-k-1))))

         [,1]
[1,] 6.731853

> anova(model.vec)

Analysis of Variance Table

Response: SKT
          Df  Sum Sq Mean Sq F value   Pr(>F)
skupina    2  3932.8 1966.41  6.7319 0.002185 **
Residuals 66 19278.9  292.11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(model.vec)$fstatistic

    value     numdf     dendf
 6.731853  2.000000 66.000000
```

V izpisu povzetka `lm` modela najdemo poleg koeficienta determinacije $R^2 = SS_{model}/SS_{yy} = 1 - SS_{residual}/SS_{yy}$, tudi prilagojeni koeficient determinacije (*Adjusted R-squared*), ki vsebuje tudi informacijo o stopinjah prostosti:

$$R_a^2 = 1 - \frac{\frac{SS_{residual}}{(n-k-1)}}{\frac{SS_{yy}}{(n-1)}} = 1 - \frac{(n-1)\hat{\sigma}^2}{SS_{yy}}. \tag{43}$$

Za koeficient determinacije velja, da se ob vsaki dodani napovedni spremenljivki v modelu njegova vrednost poveča. Za prilagojeni koeficient determinacije to ne velja, ker namesto $SS_{residual}$ v formuli nastopa ocena $\hat{\sigma}^2$. Zato je bolj primeren za primerjavo dveh modelov z različnimi napovednimi spremenljivkami kot navaden R^2. V primerjavi z ostalimi kriteriji za izbiro ustreznega modela, ki jih bomo spoznali v nadaljevanju, je njegova uporaba zastarela.

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

[1] 0.1694322

> summary(model.vec)$adj.r.squared

[1] 0.1442634
```

2.4.1 F-test za primerjavo gnezdenih modelov

Model 1 vsebuje regresorje $x_1, ..., x_k$, Model 2 pa ima poleg teh regresorjev še dodatne regresorje x_{k+j}, $j = 1, ..., r$. Pravimo, da je Model 1 **gnezden znotraj** Model 2; govorimo o hierarhiji modelov (44) in (45).

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i, \tag{44}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \beta_{k+1} x_{i(k+1)} + ... + \beta_{k+r} x_{i(k+r)} + \varepsilon_i. \tag{45}$$

Statistična inferenca omogoča primerjavo dveh gnezdenih modelov. Zanima nas, ali sta taka modela ekvivalentna oziroma ali je model z več členi v statističnem smislu boljši. Ničelna domeva in alternativna domneva sta:

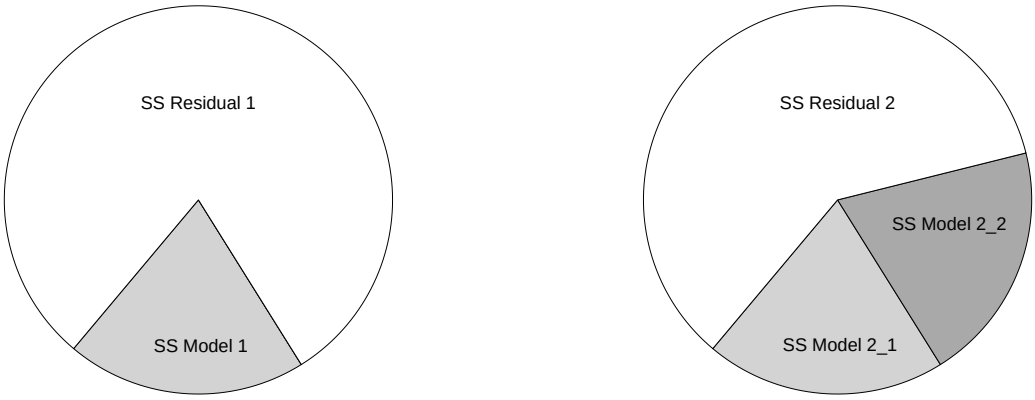
H_0: Model 1 in Model 2 sta ekvivalentna:

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

H_1: Model 2 je boljši kot Model 1:

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

Ilustracija vsote kvadriranih odklonov (SS) za dva gnezdena modela je na Sliki 6. Model 1 je gnezden znotraj Model 2; pojasnjeni del variabilnosti pri Model 2 sestavljata dve rezini, `SS Model 2_1 = SS Model 1` in dodatna rezina `SS Model 2_2`.



(a) Model 1 (b) Model 2

Slika 6: Ilustracija oznak v F-testu

Statistično sklepanje temelji na F-statistiki, ki temelji na informaciji o Model $3_3 inocenivariancezaModel 2$:

$$F = \frac{\frac{SS_{residual1} - SS_{residual2}}{Df_{residual1} - Df_{residual2}}}{\frac{SS_{residual2}}{Df_{residual2}}}. \tag{46}$$

Njena ničelna porazdelitev je $F(Df_{residual1} - Df_{residual2}, Df_{residual2})$. Ugotovimo, da za izračun testne statistike (46) potrebujemo le informacijo o ostankih pri obeh modelih.

F-test za dva gnezdena modela naredi funkcija `anova(model1, model2)`, prvi model je gnezdeni model. To pomeni, da s funkcijo `lm` najprej naredimo oba modela, nato uporabimo funkcijo `anova`. Izvedba tega testa je mogoča samo, če sta oba modela narejena na istih podatkih.

2.4.2 Sekvenčni F-testi funkcije `anova`

Ko uporabimo funkcijo `anova` na `lm` modelu z več napovednimi spremenljivkami, dobimo **sekvenčne vsote kvadriranih ostankov modela** in rezultate **sekvenčnih F-testov**. Sekvenčni F-test testira vpliv posamezne spremenljivke ob upoštevanju predhodnih spremenljivk v modelu.

- V prvi vrstici izpisa je vsota kvadriranih ostankov modela, ki vključuje samo prvo napovedno spremenljivko, označimo jo z $SS_{\beta_1|\beta_0}$. Z F-testom testiramo ničelno domnevo $H_0 : \beta_1 = 0$

oziroma preverjamo domnevo o ekvivalentnosti modela, ki vsebuje samo parameter β_0 in modela $\beta_0 + \beta_1 x_1$.

- V drugi vrstici je zapisana razlika med $SS_{residuals}$ dveh gnezdenih modelov: modela, ki ima dodano drugo napovedno spremenljivko $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, in prvega modela $\beta_0 + \beta_1 x_1$, označimo jo z $SS_{\beta_2|\beta_1,\beta_0}$. Ta vsota predstavlja del variabilnosti odzivne spremenljivke, ki je ob upoštevanju x_1, pojasnjena z x_2. Z F-testom primerjamo ta dva modela, ničelna domneva je $H_0 : \beta_2 = 0$ ob upoštevanju β_1 v modelu.

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

$$SS_{\beta_1|\beta_0}$$
$$SS_{\beta_2|\beta_0,\beta_1}$$
$$SS_{\beta_3|\beta_0,\beta_1,\beta_2}$$

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

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

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

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

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

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

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

2.5 Hkratno testiranja več parcialnih ničelnih domnev

Če je namen linearnega modela testiranje več domnev o parametrih modela hkrati, se soočamo s težavo hkratnega testiranja več domnev na podlagi istih podatkov, kar lahko privede do napačnih zaključkov o statistično značilnem vplivu izbranih spremenljivk (*false positive rate*). V nadaljevanju predstavljamo teorijo, ki omogoča hkratno testiranje več domnev na podlagi multivariatne t-porazdelitve.

Parcialno ničelno domnevo definiramo na podlagi linearne kombinacije $k+1$ parametrov $\boldsymbol{\beta}$:

$$H_0 : \mathbf{c_j}^{\mathrm{T}} \boldsymbol{\beta} = \delta, \tag{47}$$

$\mathbf{c_j}$ so koeficienti linearne kombinacije zapisani v vektor dimenzije $k+1$, δ je vrednost desne strani ničelne domneve, ki je največkrat enaka 0. Pri testiranju parcialne ničelne domneve $H_0 : \beta_0 = 0$ ima vektor $\mathbf{c_0}$ samo eno vrednost različno od 0: $\mathbf{c}_0 = (1, 0, ..., 0)$, pri $H_0 : \beta_1 = 0$ je od nič različna druga komponenta vektorja: $\mathbf{c}_1 = (0, 1, 0, ..., 0)$, ...

Z ničelnimi domnevami določimo vsebinske primerjave med parametri modela. Število vsebinsko zanimivih ničelnih domnev je ponavadi več kot 1. Če hkrati testiramo m ničelnih domnev, jih zapišemo v sistem ničelnih domnev:

$$H_{0j} : \mathbf{c}_j^{\mathrm{T}} \boldsymbol{\beta} = \delta_j, \quad j = 1, ..., m. \tag{48}$$

Izračunamo m testih statistik

$$t_j = \frac{\mathbf{c}_j^{\mathrm{T}} \mathbf{b} - \delta_j}{\hat{\sigma} \sqrt{\mathbf{c}_j^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \mathbf{c}_j}}. \tag{49}$$

V številcu (49) je ocena vrednosti primerjave, ki jo določa ničelna domneva, v imenovalcu pa njena standardna napaka.

V osnovnem povzetku `lm` modela dobimo hkrati preverjenih $k+1$ parcialnih ničelnih domnev na podlagi t-statistik, kjer pri izračunu p-vrednosti hkratnost primerjav ni upoštevana, zato pravimo, da so te p-vrednosti zgolj informativne. V splošnem lahko na podlagi `lm` modela testiramo tudi sestavljene domneve, ki vključujejo več parametrov hkrati.

Pri hkratnem testiranju m ničelnih domnev upoštevamo, da je ničelna porazdelitev testnih statistik $\mathbf{t} = (t_1, ..., t_m)$ **multivariatna t-porazdelitev**. Obliko te porazdelitve določajo stopinje prostosti ostanka modela ($df_{residual}$) in korelacijska matrika t_j-statistik $\mathbf{R}$, ki se izračuna takole:

$$\mathbf{R} = \mathbf{D}^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} Var(\mathbf{b}) \mathbf{C} \mathbf{D}. \tag{50}$$

V enačbi (50) je variančno-kovariančna matrika ocen parametrov $Var(\mathbf{b}) = \hat{\sigma}^2 (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1}$, $\mathbf{D}$ je diagonalna matrika, $\mathbf{D} = diag(\mathbf{c}_j^{\mathrm{T}} Var(\mathbf{b}) \mathbf{c}_j)^{-1/2}$, $j = 1, ...m$, ki ima na diagonali obratne vrednosti standardnih napak primerjav. Matrika primerjav $\mathbf{C}$ je reda $(k+1) \times m$, stolpec te matrike vsebuje koeficiete posamezne primerjave; matrika $\mathbf{X}$ je modelska matrika reda $n \times (k+1)$. Korelacijska matrika $\mathbf{R}$ je reda $m \times m$ in je odvisna od variančno-kovariančne matrike ocen parametrov in od

matrike primerjav $\mathbf{C}$. Pri hkratnem preverjanju več domnev izračunamo p-vrednosti na podlagi multivariatne t-porazdelitve, za katero najprej ocenimo matriko $\mathbf{R}$.

Z `lm` modelom ocenjujemo $k+1$ parametrov in testiramo hkratne domneve, da je vsak posamezen parameter modela enak 0. Matrika primerjav $\mathbf{C}$ reda $(k+1) \times (k+1)$ in je diagonalna matrika z enkami na diagonali. Domneve, ki vsebujejo samo en parameter, imenujemo enostavne domneve.

2.6 Osnovna matrika primerjav v `lm` modelu

Nadaljevanje analize odvisnosti zgornjega krvnega tlaka (`SKT`) od opisne spremenljivke `skupina` (`model.vec`):

```
> summary(model.vec)

Call:
lm(formula = SKT ~ skupina, data = tlak)

Residuals:
   Min     1Q Median     3Q    Max
-30.85 -14.86   0.55  14.14  35.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  154.850      3.822  40.519  < 2e-16 ***
skupinaB       0.600      5.405   0.111  0.91194
skupinaC     -14.988      4.968  -3.017  0.00362 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.09 on 66 degrees of freedom
Multiple R-squared:  0.1694,        Adjusted R-squared:  0.1443
F-statistic: 6.732 on 2 and 66 DF,  p-value: 0.002185

> confint(model.vec)

                2.5 %     97.5 %
(Intercept) 147.21976 162.480237
skupinaB    -10.19078  11.390785
skupinaC    -24.90623  -5.069635
```

Ničelne domneve, ki so v povzetku modela, so enostavne. Testirane so z navadnim t-testom, ki ne upošteva hkratnosti primerjav:

H_0: $\beta_0 = \mu_A = 0$, $\quad H_0$: $\beta_1 = \mu_B - \mu_A = 0 \quad$ in $\quad H_0$: $\beta_2 = \mu_C - \mu_A = 0$.

Izračunane p-vrednosti v povzetku modela so zato le informativne.

Za ilustracijo bomo izračunali korelacijsko matriko t-statistik $\mathbf{R}$ (50) za `model.vec`, za katerega

smo ocenili tri parametre. Hkratno želimo testirati tri enostavne ničelne domneve $H_0 : \beta_j = 0$, $j = 0, ..., 2$.

```
> round(b, 3) # ocene parametrov v lm modelu

(Intercept)    skupinaB    skupinaC
    154.850       0.600     -14.988

> varb<-vcov(model.vec); round(varb, 3) # variančno-kovariančna matrika ocen parametrov

            (Intercept) skupinaB skupinaC
(Intercept)      14.605  -14.605  -14.605
skupinaB        -14.605   29.211   14.605
skupinaC        -14.605   14.605   24.678

> C<-diag(3)  # matrika enostavnih primerjav
> rownames(C)<-c("beta0 = 0","beta1 = 0", "beta2 = 0")
> colnames(C)<-c("c0","c1","c2");C

          c0 c1 c2
beta0 = 0  1  0  0
beta1 = 0  0  1  0
beta2 = 0  0  0  1

> sqrt(diag(C %*% varb %*% t(C))) # vektor ocen standardnih napak ocen parametrov

beta0 = 0 beta1 = 0 beta2 = 0
 3.821683  5.404676  4.967682

> D<-diag(1/sqrt(diag(C %*% varb %*% t(C)))); D

          [,1]     [,2]      [,3]
[1,] 0.2616648 0.000000 0.0000000
[2,] 0.0000000 0.185025 0.0000000
[3,] 0.0000000 0.000000 0.2013011

> t<-D %*% C %*% b; t  # t- statistike za 3 ničelne domneve

          [,1]
[1,] 40.518794
[2,]  0.111015
[3,] -3.017088

> R<-D %*% C %*% varb %*% t(C) %*% t(D); R  #korelacijska matrika t-statistik

           [,1]       [,2]       [,3]
[1,]  1.0000000 -0.7071068 -0.7693093
[2,] -0.7071068  1.0000000  0.5439838
[3,] -0.7693093  0.5439838  1.0000000
```

Korelacije med posameznimi t-statistikami so velike, npr. -0.77, -0.71, kar potrjuje potrebnost izračuna prilagoditve p-vrednosti za testiranje domnev, ki so testirane v povzetku `lm` modela. Prilagojene p-vrednosti izračunamo na podlagi multivariatne t-porazdelitve s funkcijo `pmvt` iz paketa `mvtnorm`.

```
> df<- n - k - 1; df # stopinje prostosti ostanka

[1] 66

> library(mvtnorm)
> # p-vrednosti izračunane po multivariatni t-porazdelitvi
> # numerična integracija (Genz in Bretz, 2009)
> p.mvt1<-sapply(abs(t),
+                function(x) {1 - pmvt(-rep(x, 3), rep(x, 3),
+                                      delta=rep(0, 3), corr = R, df = df)})
> round(p.mvt1,4)

[1] 0.0000 0.9985 0.0091
```

Tako izračunane p-vrednosti so za vse tri ničelne domneve hkrati večje kot v povzetku `lm` modela (`model.vec`). Prvo ničelno domnevo zavrnemo ($p < 0.0001$), v tem modelu nima vsebinskega pomena; drugo ničelno domnevo obdržimo (p =0.9985), ni statistično značilne razlike v povprečnem `SKT` v skupinah A in B; tretjo ničelno domnevo zavrnemo (p =0.0091), kar pomeni, daje povprečni SKT v skupini C statistično značilno različen od povprečnega SKT v skupini A.

2.7 Funkcija `glht`

Funkcija `glht` (*general linear hypotheses testing*) iz paketa `multcomp` (Bretz, Hothorn, Westfall, 2010) izračuna prilagojene p-vrednosti in intervale zaupanja za izbrane primerjave na osnovi multivariatne t-porazdelitve. Funkcija ima dva argumenta, prvi je `ime modela`, ki je lahko rezultat funkcij `lm`, `gls`, `lme`, `glm` in drugi argument je `linfct` (*linear function*), s katerim definiramo hkratne ničelne domneve oziroma primerjave (matrika primerjav `C`). Za izračun verjetnosti multivariatne t-porazdelitve se uporablja Monte Carlo integracija, kar pomeni, da dobimo vsakič, ko uporabimo to funkcijo na istih podatkih, malo drugačne rezultate. Če popravljamo p-vrednosti, ki so v povzetku modela, argumenta `linfct` ni potrebno posebej definirati.

```
> library(multcomp)
> # popravljene p-vrednosti za lm model
> test.0<-glht(model.vec)
> summary(test.0)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ skupina, data = tlak)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0  154.850      3.822  40.519  < 0.001 ***
skupinaB == 0       0.600      5.405   0.111  0.99847
skupinaC == 0     -14.988      4.968  -3.017  0.00936 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test.0)

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ skupina, data = tlak)

Quantile = 2.353
95% family-wise confidence level


Linear Hypotheses:
                  Estimate lwr      upr
(Intercept) == 0 154.8500 145.8577 163.8423
skupinaB == 0      0.6000 -12.1171  13.3171
skupinaC == 0    -14.9879 -26.6768  -3.2991
```

Primerjajte osnovne in popravljene p-vrednosti in intervale zaupanja za parametre modela za `model.vec`.

2.8 Vsebinsko določena matrika primerjav v `lm` modelu

V standardnem povzetku linearnega modela smo preverili prvo ničelno domnevo, da je povprečje v skupini A enako 0. Ta nima vsebinskega pomena, po drugi strani pa ne izvemo, ali obstaja statistično značilna razlika v povprečnem `SKT` med skupinama B in C. Smiselne ničelne domneve za `model.vec` so:

H_0: $\beta_1 = \mu_B - \mu_A = 0$ $\quad$ H_0: $\beta_2 = \mu_C - \mu_A = 0$ $\quad$ in $\quad$ H_0: $\beta_2 - \beta_1 = \mu_C - \mu_B = 0$.

Za argument `linfct` funkcije `glht` določimo matriko primerjav, ki ima posamezno primerjavo zapisano v vrstico. Za hkratno testiranje teh treh domnev je matrika primerjav `C1` taka:

```
> C1<-rbind(c(0, 1, 0), c(0, 0, 1), c(0, -1, 1))
> rownames(C1)<-c("mu_B-mu_A","mu_C-mu_A", "mu_C-mu_B")
> colnames(C1)<-c("beta0","beta1","beta2");C1

          beta0 beta1 beta2
mu_B-mu_A     0     1     0
mu_C-mu_A     0     0     1
mu_C-mu_B     0    -1     1

> test.1<-glht(model.vec,linfct=C1)
> summary(test.1)

         Simultaneous Tests for General Linear Hypotheses
```

```
Fit: lm(formula = SKT ~ skupina, data = tlak)

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
mu_B-mu_A == 0    0.600      5.405   0.111  0.99322
mu_C-mu_A == 0  -14.988      4.968  -3.017  0.00995 **
mu_C-mu_B == 0  -15.588      4.968  -3.138  0.00699 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
> confint(test.1)

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ skupina, data = tlak)

Quantile = 2.3965
95% family-wise confidence level


Linear Hypotheses:
               Estimate lwr      upr
mu_B-mu_A == 0   0.6000 -12.3524  13.5524
mu_C-mu_A == 0 -14.9879 -26.8931  -3.0828
mu_C-mu_B == 0 -15.5879 -27.4931  -3.6828
```

Interpretacija: med skupinama A in B ni statistično značilne razlike med povprečnim `SKT` ($p = 0.9932$). Povprečni `SKT` skupine C je statistično značilno različen od povprečnega `SKT` v skupinah A in B. Pri 95 % zaupanju je povprečni SKT v skupini A od 3.1 mm do 26.9 mm višji kot v skupini C, v skupini B pa od 3.7 mm do 27.5 mm višji kot v skupini C.

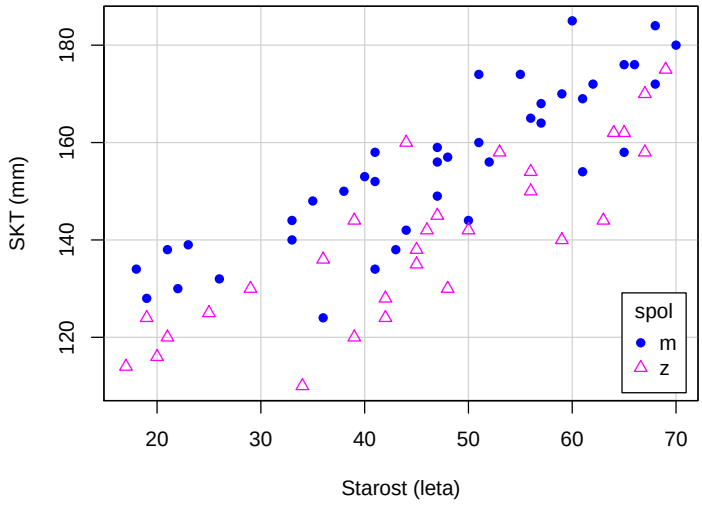
3 OPISNA IN ŠTEVILSKA NAPOVEDNA SPREMENLJIVKA

Na primerih poglejmo `lm` model, ki vključuje eno opisno napovedno spremenljivko in eno številsko napovedno spremenljivko.

3.1 Dve regresijski premici

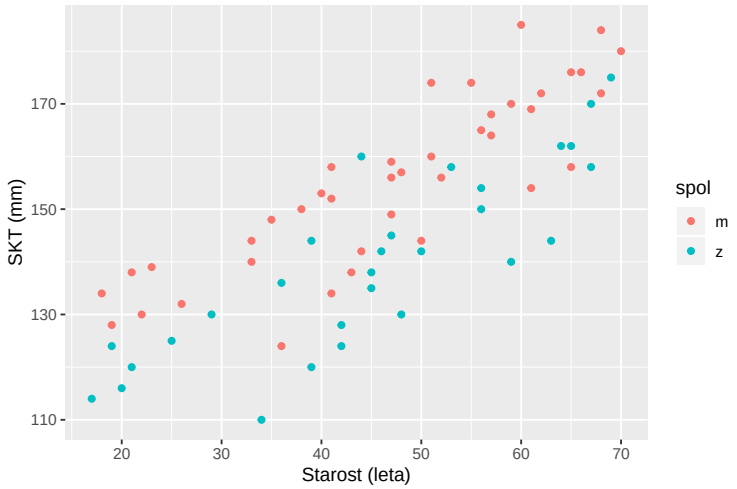
Zanima nas, kako starost in spol hkrati vplivata na SKT (Slika 7).

```
> library(car)
> scatterplot(SKT~starost | spol, regLine=FALSE,
+             smooth=FALSE, boxplots=FALSE, by.groups=TRUE,  pch=c(16,2),
+             xlab="Starost (leta)", ylab="SKT (mm)",
+             legend=list(coords="bottomright"), data=tlak)
```



Slika 7: Odvisnost SKT od starosti po spolu, `scatterplot()`

```
> library(ggplot2)
> ggplot(data=tlak, aes(x=starost, y=SKT, col=spol)) +
+        geom_point() + xlab("Starost (leta)") + ylab("SKT (mm)")
```



Slika 8: Odvisnost SKT od starosti po spolu, `ggplot()`

`SKT` je linearno odvisen od starosti, torej lahko uporabimo linearni regresijski model, ki ga geometrijsko predstavljata dve premici. Poglejmo dve varianti, ki prideta v poštev v tem primeru.

Varianta 1: Model brez interakcije
Zanima nas, kako starost in spol vplivata na SKT. V geometrijskem konteksu gre za dve vzporedni premici (presečišči sta različni, naklona sta enaka).

V tem primeru se v model za spremenljivko `spol` vključi umetno spremenljivko w_i, poleg nje pa še `starost` in ocenjujemo tri parametre modela:

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 \texttt{starost}_i + \varepsilon_i, \tag{51}$$

kar pomeni, da je pričakovana vrednost $E(y_i)$ enaka:

$$E(y_i) = \begin{cases} \beta_0 + \beta_2 \texttt{starost}_i, & \texttt{spol} = m \\ (\beta_0 + \beta_1) + \beta_2 \texttt{starost}_i, & \texttt{spol} = z. \end{cases}$$

Parameter modela β_0 predstavlja povprečni SKT moških pri `starost`=0; β_1 je razlika povprečja SKT za ženske in povprečja SKT za moške pri `starost`=0 in β_2 je naklon vzporednih premic. Glede na to, da sta premici vzporedni, je razlika povprečja SKT za ženske in povprečja SKT za moške za vse vrednosti spremenljivke `starost` enaka β_1.

Varianta 2: Model z interakcijo

Zanima nas, kako starost, spol in njuna interakcija vplivajo na SKT. V geometrijskem konteksu gre za dve različni premici (presečišči sta različni, naklona sta različna).

V tem primeru ocenjujemo štiri parametre modela:

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 \texttt{starost}_i + \beta_3 \texttt{starost}_i w_i + \varepsilon_i, \tag{53}$$

kar pomeni, da je pričakovana vrednost $E(y_i)$ enaka:

$$E(y_i) = \begin{cases} \beta_0 + \beta_2 \texttt{starost}_i, & \texttt{spol} = m \\ (\beta_0 + \beta_1) + (\beta_2 + \beta_3)\texttt{starost}_i, & \texttt{spol} = z. \end{cases}$$

Parameter modela β_0 predstavlja povprečni SKT moških pri `starost`=0; β_1 je razlika povprečja SKT za ženske in povprečja SKT za moške pri `starost`=0; β_2 je naklon premice za moške in β_3 je razlika naklona premice za ženske in naklona premice za moške.

3.1.1 Model brez interakcije

Zanima nas, **kako starost in spol vplivata na SKT.**

```
> model.vzporedni <- lm(SKT ~ spol + starost, data=tlak)
```

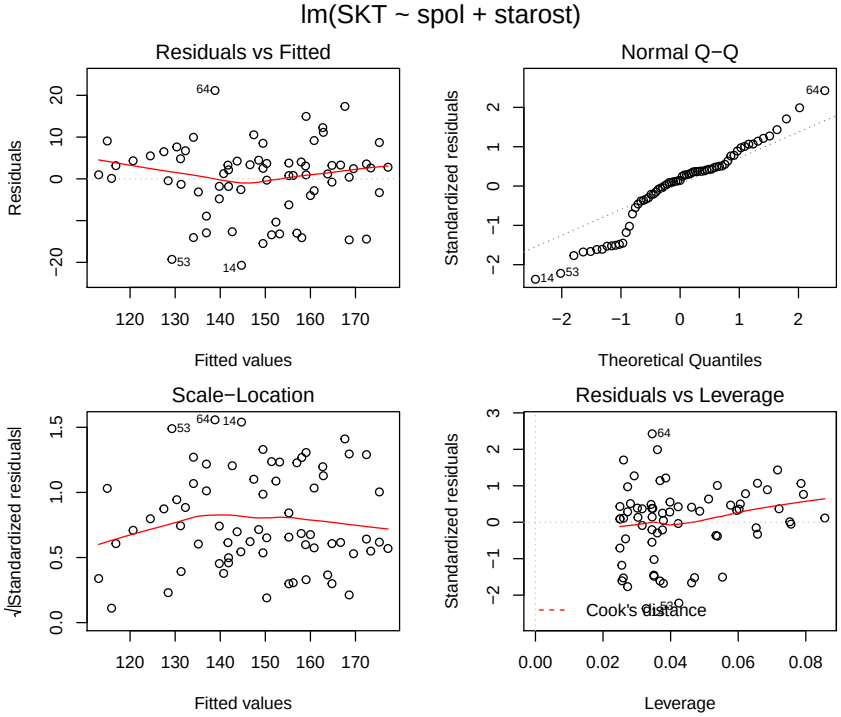
Modelska matrika je v tem primeru reda $n \times 3$, $n = 69$:

```
> X<-model.matrix(model.vzporedni)
> X[18:21,] # za ilustracijo
```

```
   (Intercept) spolz starost
18           1     0      19
19           1     0      22
20           1     0      21
21           1     0      38
```

```
> X[39:42,]
```

```
   (Intercept) spolz starost
39           1     0      26
40           1     0      61
41           1     1      39
42           1     1      45
```



Slika 9: Grafični prikaz ostankov za `model.vzporedni`

Slika 9 levo zgoraj je sprejemljiva, desna zgoraj pa je mejno sprejemljiva. Nadaljujemo z analizo.

V modelu `model.vzporedni` se povprečni SKT pri `spol=z` primerja na referenčno skupino `spol=m` pri `starost=0`, poleg tega zadnji parameter ocenjuje spremembo `SKT` v odvisnosti od `starost`, za katero smo predpostavili, da je enaka pri moških in pri ženskah.

```
> summary(model.vzporedni)

Call:
lm(formula = SKT ~ spol + starost, data = tlak)

Residuals:
    Min      1Q  Median      3Q     Max
-20.705  -3.299   1.248   4.325  21.160

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 110.28698    3.63824  30.313  < 2e-16 ***
spolz       -13.51345    2.16932  -6.229  3.7e-08 ***
starost       0.95606    0.07153  13.366  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.878 on 66 degrees of freedom
Multiple R-squared:  0.7759,        Adjusted R-squared:  0.7691
F-statistic: 114.2 on 2 and 66 DF,  p-value: < 2.2e-16

> confint(model.vzporedni)

                  2.5 %     97.5 %
(Intercept) 103.0230018 117.550959
spolz       -17.8446366  -9.182272
starost       0.8132441   1.098872
```

H_0: $\beta_0 = \mu_{m(starost=0)} = 0$,
H_0: $\beta_1 = \mu_z|starost - \mu_m|starost = 0$,
H_0: $\beta_2 = naklon = 0$.

S funkcijo `glht` popravimo p-vrednosti in intervale zaupanja za parametre modela zaradi hkratnih primerjav:

```
> test.vzporedni<-glht(model.vzporedni)
> summary(test.vzporedni)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ spol + starost, data = tlak)
```

```
Linear Hypotheses:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0 110.28698    3.63824  30.313  < 1e-08 ***
spolz == 0       -13.51345    2.16932  -6.229 5.96e-08 ***
starost == 0       0.95606    0.07153  13.366  < 1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test.vzporedni)

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ spol + starost, data = tlak)

Quantile = 2.352
95% family-wise confidence level


Linear Hypotheses:
                  Estimate lwr      upr
(Intercept) == 0 110.2870 101.7298 118.8441
spolz == 0       -13.5135 -18.6157  -8.4112
starost == 0       0.9561   0.7878   1.1243
```

Enačbi vzporednih premic sta:

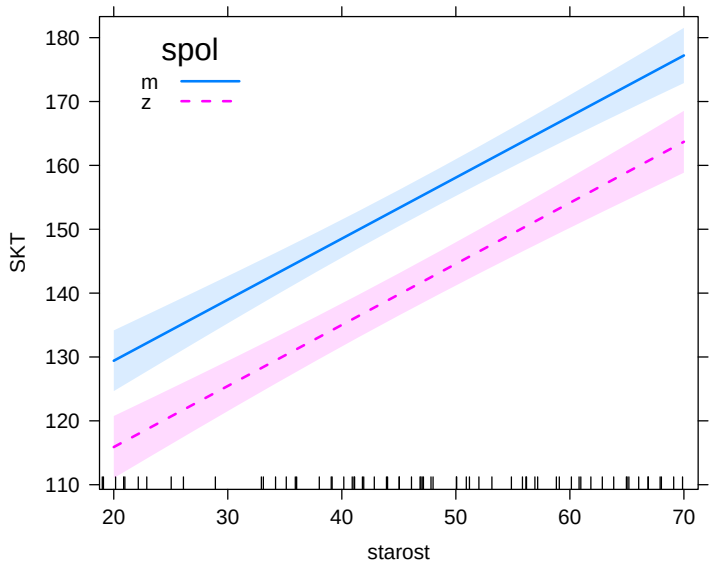
$$\text{Moški:} \quad \widehat{SKT} = 110.29 + 0.96\ \texttt{starost}.$$
$$\text{Ženske:} \quad \widehat{SKT} = (110.29 + (-13.51)) + 0.96\ \texttt{starost}.$$

Ta model pojasni 77.6 % variabilnosti SKT. p - vrednosti v povzetku modela so izračunane tako, da se upošteva testiranje več domnev hkrati, kar pa pri majhnem številu ocenjenih parametrov in tako močni statistični značilnosti rezultatov ne spremeni bistveno. Vsebinska interpretacija:

- Povprečni SKT pri starosti 0 za moške je 110.3 mm in je statistično značilno različen od nič ($p < 0.0001$); ta ocena parametra nima vsebinskega pomena.
- Moški imajo pri vseh analiziranih starostih za 13.5 mm večji SKT kot ženske, ta razlika je statistično značilno različna od 0, pripadajoči 95 % interval zaupanja je (8.4 mm, 18.6 mm).
- Če se starost poveča za 10 let se ob upoštevanju spola SKT v povprečju poveča za 9.6 mm, pripadajoči 95 % interval zaupanja je (7.9 mm, 11.2 mm). Velja za moške in ženske.

Grafični prikaz napovedi s 95 % intervali zaupanja za povprečno napoved lahko naredimo s pomočjo funkcije `Effect` iz paketa `effects`:

```
> library(effects)
> plot(Effect(c("starost", "spol"), model.vzporedni),
+      multiline=T, ci.style="bands",
+      key.args=list(x=0.05, y=0.8, corner=c(0,0)),
+      main="", lty=c(1:2))
```

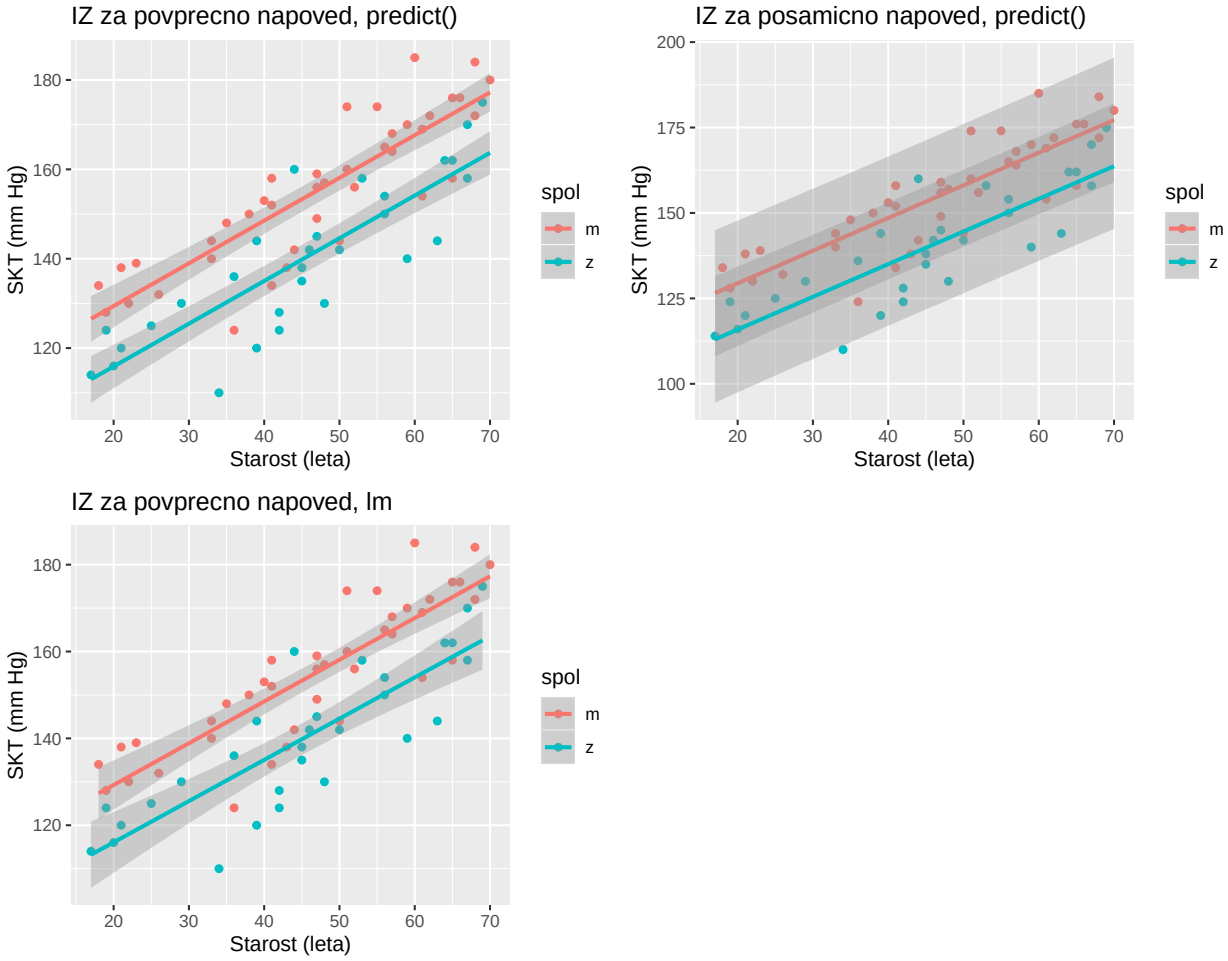


Slika 10: Odvisnost SKT od starosti in spola, premici dobljeni po `model.vzporedni` z 95 % intervali zaupanja za povprečne napovedi

Za grafični prikaz podatkov, napovedi in pripadajočih intervalov zaupanja za povprečne oziroma za posamične napovedi s funkcijo `ggplot()`, moramo za model, ki ima več kot eno napovedno spremenljivko, najprej izračunati napovedi in meje intervalov zaupanja s funkcijo `predict()`.

```
> # napovedi za izbrane vrednosti napovednih spremenljivk starost in spol:
> n1<-max(tlak$starost)- min(tlak$starost)+1
> x <-seq(from=min(tlak$starost), to=max(tlak$starost), by=1)
> nap.x <- data.frame(starost = rep(x, times=2),
+                     spol = rep(c("m","z"), each=n1))
> mod<-model.vzporedni
> # interval zaupanja za povprečno napoved
> conf.int <- cbind(nap.x, predict(mod, nap.x, interval="confidence", level=0.95))
> # interval zaupanja za posamično napoved
> pred.int <- cbind(nap.x, predict(mod, nap.x, interval="prediction", level=0.95))

> library(ggplot2)
> p0 <- ggplot(data=tlak, mapping=aes(x=starost, y=SKT, colour=spol)) +
+       ggtitle("IZ za povprečno napoved, lm") + geom_point() +
+       geom_smooth(method="lm", se=TRUE) +
+       xlab("Starost (leta)") + ylab("SKT (mm Hg)")
```

```
> p1 <- ggplot(conf.int, aes(x = starost, y = fit, col=spol)) +
+       ggtitle("IZ za povprečno napoved, predict()") +
+       geom_point(data = tlak, aes(x = starost, y = SKT, col=spol)) +
+       geom_smooth(data = conf.int, aes(ymin = lwr, ymax = upr), stat = "identity") +
+       xlab("Starost (leta)") + ylab("SKT (mm Hg)")
> p2 <- ggplot(pred.int, aes(x = starost, y = fit, col=spol)) +
+       ggtitle("IZ za posamično napoved, predict()") +
+       geom_point(data = tlak, aes(x = starost, y = SKT, col=spol)) +
+       geom_smooth(data = pred.int, aes(ymin = lwr, ymax = upr), stat = "identity") +
+       xlab("Starost (leta)") + ylab("SKT (mm Hg)")

> library(gridExtra)
> grid.arrange(p1,p2,p0, ncol=2)
```



Slika 11: Odvisnost SKT od starosti in spola, premici dobljeni po `model.vzporedni` z 95 % intervali zaupanja za povprečno napoved (zgoraj levo); z 95 % intervali zaupanja za posamično napoved (zgoraj desno) in 95 % intervali zaupanja za povprečno napoved, če bi vsako od premic modelirali posebej (levo spodaj)

Opozorimo naj, da vrstni red napovednih spremenljivk v formuli modela določa vrstni red ocenjenih parametrov v povzetku modela. Matriko primerjav sestavimo z upoštevanjem tega vrstnega reda. Če v `model.vzporedni` zamenjamo vrstni red napovednih spremenljivk, je vrstni red ocen parametrov modela tak:

```
> model.vzporedni.a<-lm(SKT~starost+spol, data=tlak)
> coefficients(model.vzporedni.a)

(Intercept)     starost       spolz
 110.286980    0.956058  -13.513454
```

Model `model.vzporedni` lahko spremenimo tako, da so vsi parametri vsebinsko obrazložljivi. Če želimo, da presečišče ocenjuje povprečje SKT pri vsebinsko izbrani starosti (npr. 50 let), to dosežemo tako, da od vsake vrednosti za starost odštejemo izbrano vrednost. Novo spremenljivko označimo `starost.50`.

```
> tlak$starost.50<-tlak$starost-50
> model.vzporedni.50 <- lm(SKT ~ spol + starost.50, data=tlak)
```

Ničelne domneve, ki se testirajo v povzetku `model.vzporedni.50`, so:

H_0: $\beta_0 = \mu_{m(starost.50=0)} = 0$,
H_0: $\beta_1 = \mu_z|starost.50 - \mu_m|starost.50 = 0$, to velja za vsako starost,
H_0: $\beta_2 = naklon = 0$,

```
> test.vzporedni.50<-glht(model.vzporedni.50)
> summary(test.vzporedni.50)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ spol + starost.50, data = tlak)

Linear Hypotheses:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0 158.08988    1.42086 111.264  < 1e-07 ***
spolz == 0       -13.51345    2.16932  -6.229 1.29e-07 ***
starost.50 == 0    0.95606    0.07153  13.366  < 1e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test.vzporedni.50)

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ spol + starost.50, data = tlak)
```

```
Quantile = 2.4194
95% family-wise confidence level


Linear Hypotheses:
                  Estimate lwr      upr
(Intercept) == 0 158.0899 154.6522 161.5276
spolz == 0       -13.5135 -18.7620  -8.2649
starost.50 == 0    0.9561   0.7830   1.1291
```

Parameter β_0 ocenjuje povprečni SKT za moške pri starosti 50 let, parameter β_1 ocenjuje razliko med povprečnim SKT žensk in povprečnim SKT moških pri vseh starostih. Ker sta premici vzporedni, je ocena tega parametra enaka kot za `model.vzporedni`. Tudi ocena za naklon ostane ista.

3.1.2 Model z interakcijo

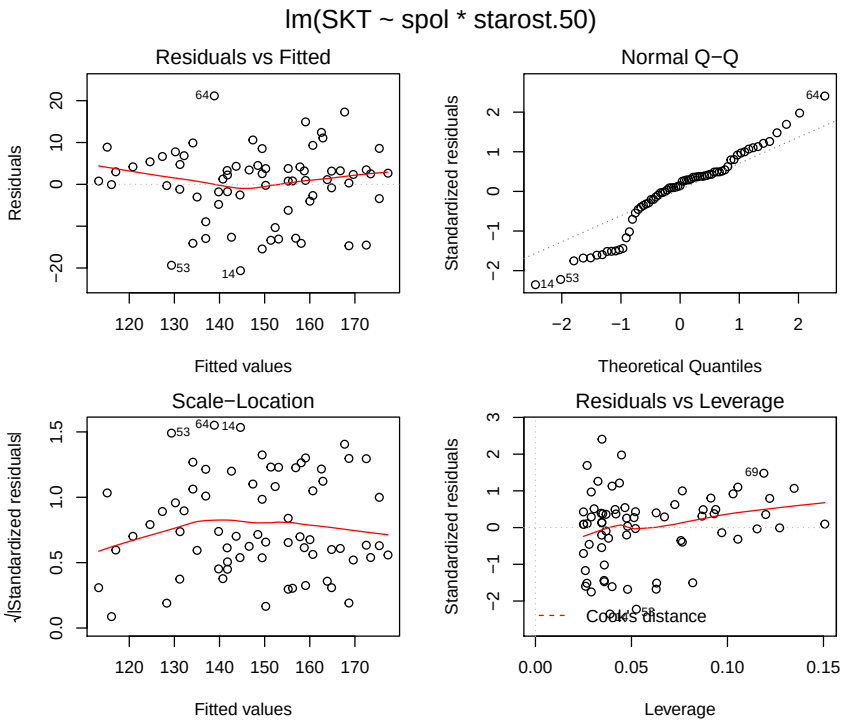
Zanima nas, **kako starost, spol in njuna interakcija vplivajo na SKT.** Za starost bomo upoštevali `starost.50`.

Pri modeliranju vključimo v model `spol`, `starost.50` in njuno interakcijo `spol:starost.50`, kar krajše zapišemo `spol*starost.50`. Zapis `spol*starost.50` je isti kot zapis `spol+starost.50+spol:starost.50`

```
> model.razlicni <- lm(SKT ~ spol*starost.50, data=tlak)
```

Modelska matrika je reda $n \times 4$, $n = 69$:

```
> X<-model.matrix(model.razlicni)
> X[18:21,]

   (Intercept) spolz starost.50 spolz:starost.50
18           1     0        -31                0
19           1     0        -28                0
20           1     0        -29                0
21           1     0        -12                0

> X[39:42,]

   (Intercept) spolz starost.50 spolz:starost.50
39           1     0        -24                0
40           1     0         11                0
41           1     1        -11              -11
42           1     1         -5               -5
```



Slika 12: Grafični prikaz ostankov za `model.razlicni`

```
> summary(model.razlicni)

Call:
lm(formula = SKT ~ spol * starost.50, data = tlak)

Residuals:
    Min      1Q  Median      3Q     Max
-20.647  -3.410   1.254   4.314  21.153

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      158.10616    1.44509 109.409  < 2e-16 ***
spolz            -13.56295    2.26598  -5.985 1.03e-07 ***
starost.50         0.96135    0.09632   9.980 9.63e-15 ***
spolz:starost.50  -0.01203    0.14519  -0.083    0.934
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.946 on 65 degrees of freedom
Multiple R-squared:  0.7759,        Adjusted R-squared:  0.7656
F-statistic: 75.02 on 3 and 65 DF,  p-value: < 2.2e-16
```

Ničelne domneve, ki se testirajo v povzetku `model.razlicni`, so:

H_0: $\beta_0 = \mu_{m(starost.50=0)} = 0$,
H_0: $\beta_1 = \mu_z|starost.50 - \mu_m|starost.50 = 0$,
H_0: $\beta_2 = naklon_m = 0$,
H_0: $\beta_3 = naklon_z - naklon_m = 0$.

```
> test.razlicni<-glht(model.razlicni)
> summary(test.razlicni)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ spol * starost.50, data = tlak)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0      158.10616    1.44509 109.409  < 1e-07 ***
spolz == 0            -13.56295    2.26598  -5.985 2.31e-07 ***
starost.50 == 0         0.96135    0.09632   9.980  < 1e-07 ***
spolz:starost.50 == 0  -0.01203    0.14519  -0.083        1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test.razlicni)

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ spol * starost.50, data = tlak)

Quantile = 2.5131
95% family-wise confidence level


Linear Hypotheses:
                      Estimate  lwr       upr
(Intercept) == 0      158.10616 154.47451 161.73781
spolz == 0            -13.56295 -19.25757  -7.86833
starost.50 == 0         0.96135   0.71928   1.20342
spolz:starost.50 == 0  -0.01203  -0.37691   0.35285
```

To je model dveh premic, ki imata različno izhodišče in različna naklona. Njuni enačbi sta:

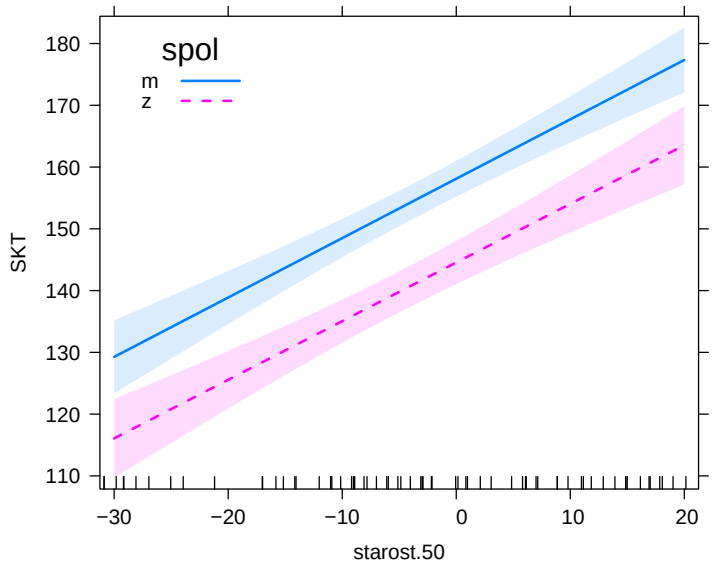
Moški: $\widehat{SKT} = 158.11 + 0.96$ `starost.50`,

Ženske: $\widehat{SKT} = (158.11 + (-13.56)) + (0.96 + (-0.01))$ `starost.50` $= 144.55 + 0.95$ `starost.50`.

Vsebinsko obrazložite rezultate.

Grafični prikaz napovedi za `model.razlicni` je na Sliki 13:

```
> plot(Effect(c("starost.50", "spol"), model.razlicni),
+      multiline=T, ci.style="bands",
+      key.args=list(x=0.05, y=0.8, corner=c(0,0)),
+      main="", lty=c(1:2))
```



Slika 13: Odvisnost SKT od centrirane starosti, spola in njune interakcije, premici dobljeni po `model.razlicni` z intervali zaupanja za povprečne napovedi

3.2 Več regresijskih premic

3.2.1 Model brez interakcije

Analizirajmo model za odvisnost `SKT` od `skupina` in `starost.50`. Uporabimo funkcijo `lm` in ocenimo štiri parametre modela:

```
> model.vzporedne<-lm(SKT~skupina+starost.50, data=tlak)
> X<-model.matrix(model.vzporedne)
> X[18:23,]

   (Intercept) skupinaB skupinaC starost.50
18           1        0        0        -31
19           1        0        0        -28
20           1        0        0        -29
21           1        1        0        -12
22           1        1        0          2
23           1        1        0         -9

> X[38:43,]

   (Intercept) skupinaB skupinaC starost.50
38           1        1        0        -17
39           1        1        0        -24
40           1        1        0         11
41           1        0        1        -11
42           1        0        1         -5
43           1        0        1         -3
```

V modelu `model.vzporedne` se B in C primerjata na referenčno skupino A pri starosti 50 let, poleg tega zadnji parameter ocenjuje spremembo `SKT` v odvisnosti od `starost.50`, za katero smo predpostavili, da je enaka v vseh treh skupinah. Z uporabo `glht` hkratno testiramo eno statistično domnevo več kot pri `model.vzporedni`:

H_0: $\beta_0 = \mu_{A(starost.50=0)} = 0$,
H_0: $\beta_1 = \mu_B|starost.50 - \mu_A|starost.50 = 0$,
H_0: $\beta_2 = \mu_C|starost.50 - \mu_A|starost.50 = 0$,
H_0: $\beta_3 = naklon = 0$.

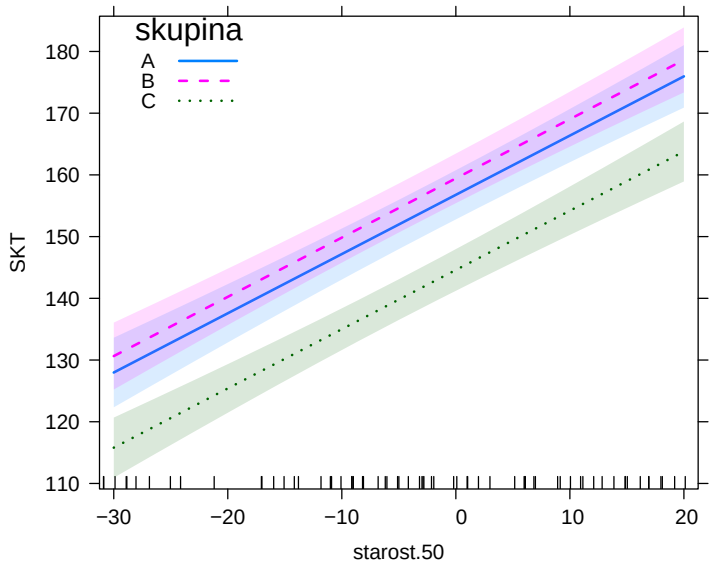
```
> test.vzporedne<-glht(model.vzporedne)
```

```
> summary(test.vzporedne)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ skupina + starost.50, data = tlak)

Linear Hypotheses:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0 156.76955    1.99190  78.703   <1e-04 ***
skupinaB == 0      2.66352    2.81389   0.947    0.739
skupinaC == 0    -12.17480    2.59103  -4.699   <1e-04 ***
starost.50 == 0    0.95978    0.07169  13.387   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test.vzporedne)

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ skupina + starost.50, data = tlak)

Quantile = 2.4989
95% family-wise confidence level


Linear Hypotheses:
                 Estimate lwr      upr
(Intercept) == 0 156.7696 151.7919 161.7472
skupinaB == 0      2.6635  -4.3683   9.6953
skupinaC == 0    -12.1748 -18.6496  -5.7000
starost.50 == 0    0.9598   0.7806   1.1389
```

Ker smo z modelom predpostavili, da je odvisnost `SKT` od `starost.50` za vse tri skupine enaka, lahko interpretacijo ničelnih domnev o presečiščih razširimo na katerokoli analizirano vrednost spremenljivke `starost.50` na analiziranem intervalu. Torej, pri katerikoli vrednosti spremenljivke `starost.50` je povprečen `SKT` v skupini C statistično značilno nižji od povprečnega SKT v skupini A ($p < 0.0001$), med skupinama A in B ni statistično značilnih razlik ($p = 0.739$). Odvisnost `SKT` od starosti je statistično značilna ($p < 0.0001$), v povprečju se `SKT` z vsakim letom starosti poveča za 0.96 mm (0.78 mm, 1.14 mm).

Grafičen prikaz napovedi za `model.vzporedne` je na Sliki 14.

```
> plot(Effect(c("starost.50", "skupina"), model.vzporedne), multiline=T, ci.style="bands",
+      key.args=list(x=0.05, y=0.8, corner=c(0,0)),  main="", lty=c(1:3))
```



Slika 14: Odvisnost `SKT` od `starost.50`, napovedi za tri skupine po `model.vzporedne` z intervali zaupanja za povprečne napovedi

Vaja: testiramo ničelne domneve, ki se nanašajo na parne razlike presečišč in na naklon.

H_0: $\beta_1 = \mu_B|starost.50 - \mu_A|starost.50 = 0$,
H_0: $\beta_2 = \mu_C|starost.50 - \mu_A|starost.50 = 0$,
H_0: $\beta_2 - \beta_1 = \mu_C|starost.50 - \mu_B|starost.50 = 0$,
H_0: $\beta_3 = naklon = 0$.

Ker je to model vzporednih premic, prve tri ničelne domneve primerjajo povprečni SKT med dvema skupinama pri poljubni izbrani vrednosti za `starost.50`.

```
> C2<-rbind(c(0, 1, 0, 0), c(0, 0, 1, 0), c(0, -1, 1, 0), c(0, 0, 0, 1))
> colnames(C2)<-c("beta0","beta1","beta2","beta3")
> rownames(C2)<-c("povp B|starost - povp A|starost",
+                 "povp C|starost - povp A|starost",
+                 "povp C|starost - povp B|starost",
+                 "naklon"); C2
```

```
                               beta0 beta1 beta2 beta3
povp B|starost - povp A|starost     0     1     0     0
povp C|starost - povp A|starost     0     0     1     0
povp C|starost - povp B|starost     0    -1     1     0
naklon                              0     0     0     1
```

```
> test.vzporedne.2<-glht(model.vzporedne, linfct=C2)
> summary(test.vzporedne.2)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ skupina + starost.50, data = tlak)

Linear Hypotheses:
                                     Estimate Std. Error t value Pr(>|t|)
povp B|starost - povp A|starost == 0   2.66352    2.81389   0.947    0.744
povp C|starost - povp A|starost == 0 -12.17480    2.59103  -4.699   <1e-04 ***
povp C|starost - povp B|starost == 0 -14.83831    2.58310  -5.744   <1e-04 ***
naklon == 0                            0.95978    0.07169  13.387   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test.vzporedne.2)

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ skupina + starost.50, data = tlak)

Quantile = 2.525
95% family-wise confidence level


Linear Hypotheses:
                                     Estimate lwr      upr
povp B|starost - povp A|starost == 0   2.6635  -4.4415   9.7685
povp C|starost - povp A|starost == 0 -12.1748 -18.7170  -5.6326
povp C|starost - povp B|starost == 0 -14.8383 -21.3605  -8.3161
naklon == 0                            0.9598   0.7788   1.1408
```

Interpretacija: pri katerikoli vrednosti spremenljivke `starost.50` v opazovanem intervalu je povprečen `SKT` v skupini C statistično značilno nižji od povprečnega SKT v skupini A ($p < 0.0001$) in v skupini B ($p < 0.0001$), med skupinama A in B ni statistično značilne razlike ($p = 0.744$). Odvisnost `SKT` od starosti je statistično značilna ($p < 0.0001$), v povprečju se `SKT` z vsakim letom poveča za 0.96 mm (0.78 mm, 1.14 mm), v vseh treh skupinah enako.

Vaja: oblikujte matriko primerjav za primer hkratnega testiranja istih ničelnih domnev, če je formula modela enaka `SKT` $\sim$ `starost.50+skupina`.

3.2.2 Model z interakcijo

Zanima nas vpliv spremenljivk `skupina` in `starost.50` ter njun medsebojni vpliv na SKT. Z modelom ocenjujemo šest parametrov.

```
> model.razlicne<-lm(SKT~skupina*starost.50, data=tlak)
> # X<-model.matrix(model.razlicne)
> # X[18:21,]
> # X[39:42,]
```

Ničelne domneve, ki se testirajo v povzetku modela, so:

H_0: $\beta_0 = \mu_{A(starost.50=0)} = 0$,
H_0: $\beta_1 = \mu_{B(starost.50=0)} - \mu_{A(starost.50=0)} = 0$,
H_0: $\beta_2 = \mu_{C(starost.50=0)} - \mu_{A(starost.50=0)} = 0$,
H_0: $\beta_3 = naklon_A = 0$,
H_0: $\beta_4 = naklon_B - naklon_A = 0$,
H_0: $\beta_5 = naklon_C - naklon_A = 0$.

Popravek p-vrednosti za hkratno testiranje šestih domnev dobimo takole:

```
> test.razlicne<-glht(model.razlicne)

> summary(test.razlicne)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ skupina * starost.50, data = tlak)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0         156.92052    2.02772  77.388   <1e-04 ***
skupinaB == 0              2.24975    2.91292   0.772   0.9217
skupinaC == 0            -12.37731    2.68078  -4.617   0.0001 ***
starost.50 == 0            1.03526    0.13512   7.662   <1e-04 ***
skupinaB:starost.50 == 0  -0.13881    0.19416  -0.715   0.9422
skupinaC:starost.50 == 0  -0.08594    0.17370  -0.495   0.9888
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test.razlicne)

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ skupina * starost.50, data = tlak)

Quantile = 2.6333
95% family-wise confidence level
```

```
Linear Hypotheses:
                         Estimate  lwr       upr
(Intercept) == 0         156.92052 151.58101 162.26004
skupinaB == 0              2.24975  -5.42074   9.92023
skupinaC == 0            -12.37731 -19.43651  -5.31812
starost.50 == 0            1.03526   0.67945   1.39107
skupinaB:starost.50 == 0  -0.13881  -0.65010   0.37247
skupinaC:starost.50 == 0  -0.08594  -0.54333   0.37146
```

Interpretirajte rezultate.

Primer: želimo preveriti statistično značilnost vseh treh naklonov hkrati.

H_0: $\beta_3 = naklon_A = 0$, H_0: $\beta_4 + \beta_3 = naklon_B = 0$ in H_0: $\beta_5 + \beta_3 = naklon_C = 0$.

```
> nic<-c(0,0,0)
> C3a<-rbind(c(nic, 1, 0, 0), c(nic, 1, 1, 0), c(nic,1, 0, 1))
> rownames(C3a)<-c("naklon_A", "naklon_B", "naklon_C")
> colnames(C3a)<-c("beta0","beta1","beta2","beta3","beta4","beta5"); C3a

         beta0 beta1 beta2 beta3 beta4 beta5
naklon_A     0     0     0     1     0     0
naklon_B     0     0     0     1     1     0
naklon_C     0     0     0     1     0     1

> test.3a<-glht(model.razlicne, linfct=C3a)
> summary(test.3a)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ skupina * starost.50, data = tlak)

Linear Hypotheses:
              Estimate Std. Error t value Pr(>|t|)
naklon_A == 0   1.0353     0.1351   7.662   <1e-07 ***
naklon_B == 0   0.8965     0.1394   6.429   <1e-07 ***
naklon_C == 0   0.9493     0.1091   8.697   <1e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Nakloni v vseh treh skupinah so pozitivni in statistično značilni.

4 VAJE

4.1 model.spol

Na osnovi `model.spol` izračunajte hkratna 95 % intervala zaupanja za povprečni `SKT` za moške in za povprečni `SKT` za ženske.

```
> C<-rbind(c(1,0),c(1,1))
> rownames(C)<-c("moški","ženske")
> confint(glht(model.spol, linfct=C))

         Simultaneous Confidence Intervals

Fit: lm(formula = SKT ~ spol, data = tlak)

Quantile = 2.2862
95% family-wise confidence level


Linear Hypotheses:
            Estimate lwr      upr
moški == 0  155.1500 149.0176 161.2824
ženske == 0 139.8621 132.6600 147.0642
```

S 95 % zaupanjem je povprečni `SKT` za moške na intervalu (149.0, 161.3) in povprečni `SKT` za ženske je na intervalu (132.7, 147.1).

4.2 model.razlicne

Za `model.razlicne` želimo paroma primerjati vsa presečišča in paroma vse naklone ter statistično značilnost naklona referenčne skupine. Testiramo hkrati sedem domnev:

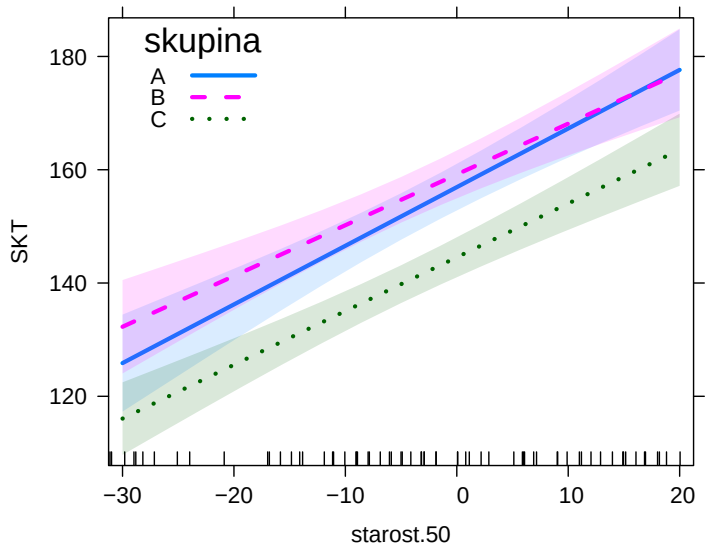
H_0: $\beta_1 = \mu_{B(starost.50=0)} - \mu_{A(starost.50=0)} = 0$,
H_0: $\beta_2 = \mu_{C(starost.50=0)} - \mu_{A(starost.50=0)} = 0$,
H_0: $\beta_2 - \beta_1 = \mu_{C(starost.50=0)} - \mu_{B(starost.50=0)} = 0$,
H_0: $\beta_3 = naklon_A = 0$,
H_0: $\beta_4 = naklon_B - naklon_A = 0$,
H_0: $\beta_5 = naklon_C - naklon_A = 0$,
H_0: $\beta_5 - \beta_4 = naklon_C - naklon_B = 0$.

Napišite ustrezno matriko primerjav in izvedite test ter obrazložite rezultate.

```
> nic<-c(0,0,0)
> C3<-rbind(c(0, 1, 0, nic), c(0, 0, 1, nic), c(0, -1, 1, nic),
+           c(nic, 1, 0, 0), c(nic, 0, 1, 0), c(nic,0, 0, 1), c(nic, 0, -1, 1))
> colnames(C3)<-c("beta0","beta1","beta2","beta3","beta4","beta5")
> rownames(C3)<-c("povp B - povp A","povp C - povp A",
+                 "povp C - povp B", "naklon A",
+                 "naklon B - naklon A", "naklon C - naklon A",
```

```
+                 "naklon C - naklon B")
> C3

                    beta0 beta1 beta2 beta3 beta4 beta5
povp B - povp A         0     1     0     0     0     0
povp C - povp A         0     0     1     0     0     0
povp C - povp B         0    -1     1     0     0     0
naklon A                0     0     0     1     0     0
naklon B - naklon A     0     0     0     0     1     0
naklon C - naklon A     0     0     0     0     0     1
naklon C - naklon B     0     0     0     0    -1     1

> test.razlicne<-glht(model.razlicne, linfct=C3)
> summary(test.razlicne)

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ skupina * starost.50, data = tlak)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
povp B - povp A == 0        2.24975    2.91292   0.772  0.93508
povp C - povp A == 0      -12.37731    2.68078  -4.617  0.00016 ***
povp C - povp B == 0      -14.62706    2.72917  -5.360  < 1e-04 ***
naklon A == 0               1.03526    0.13512   7.662  < 1e-04 ***
naklon B - naklon A == 0   -0.13881    0.19416  -0.715  0.95160
naklon C - naklon A == 0   -0.08594    0.17370  -0.495  0.98940
naklon C - naklon B == 0    0.05287    0.17707   0.299  0.99894
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
> plot(Effect(c("starost.50", "skupina"), model.razlicne), multiline=T, ci.style="bands",
+      key.args=list(x=0.05, y=0.75, corner=c(0,0)),
+      main="", lty=c(1:3), lwd=3)
```



Slika 15: Odvisnost `SKT` od `starost.50` (centrirana spremenljivka `starost`), napovedi za tri premice po `model.razlicne` z intervali zaupanja za povprečne napovedi

Pri starosti 50 let povprečni SKT skupine C statistično značilno odstopa od povprečnega SKT skupin A in B. Ni statistično značilnih razlik v naklonih med skupinami. Vpliv starosti v skupini A je statistično značilen ($p < 0.0001$).

4.3 Pelod

V poskusu so obsevali pelod buč z 8 različnimi odmerki rentgenskega sevanja (100, 200, 300, 350, 400, 500, 600, 700 Gy, *gray* je enota za absorbirano sevanje), v dveh različnih zračnih vlagah (Room humidity, RH, in High Humidity, HH). Za vsako kombinacijo vlage in odmerka sevanja je bilo 9 kapljic, ki so vsebovale pelod buč, torej 9 ponovitev za vsak odmerek sevanja; skupaj je bilo v poskusu 144 kapljic. Izid: kalivost peloda izražena kot delež kalenega peloda v kapljici (to se ugotavlja z mikroskopom). Podatki so v datoteki PELOD.txt in so bili analizirani že v kontekstu uporabe različnih transformacij spremenljivk.

Tokrat nas zanima, kako zračna vlaga (`Vlaga`), odmerek sevanja (`Sevanje`) in njuna interakcija vplivajo na kalivost peloda (`Kalivost`).

a) Grafično prikažite podatke.

b) Uporabite ustrezno transformacijo za `Kalivost` in analizirajte model.

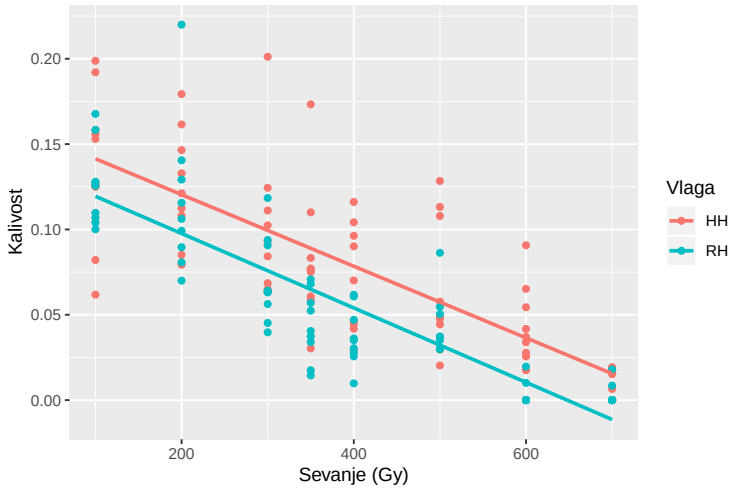
c) Grafično predstavite napovedi modela.

d) Obrazložite rezultate.

```
> pelod<-read.table("PELOD.txt",   header=TRUE)
> str(pelod)

'data.frame':        144 obs. of  3 variables:
 $ Vlaga   : Factor w/ 2 levels "HH","RH": 1 1 1 1 1 1 1 1 1 1 ...
 $ Sevanje : int  100 100 100 100 100 100 100 100 100 200 ...
 $ Kalivost: num  0.192 0.125 0.156 0.153 0.199 ...

> xtabs(~pelod$Vlaga+pelod$Sevanje)

           pelod$Sevanje
pelod$Vlaga 100 200 300 350 400 500 600 700
         HH   9   9   9   9   9   9   9   9
         RH   9   9   9   9   9   9   9   9

> summary(pelod)

 Vlaga      Sevanje         Kalivost
 HH:72   Min.   :100.0   Min.   :0.00000
 RH:72   1st Qu.:275.0   1st Qu.:0.02870
         Median :375.0   Median :0.06075
         Mean   :393.8   Mean   :0.06756
         3rd Qu.:525.0   3rd Qu.:0.10470
         Max.   :700.0   Max.   :0.22000
```

Grafični prikaz podatkov za `Kalivost` v odvisnosti od Sevanje in `Vlaga`:

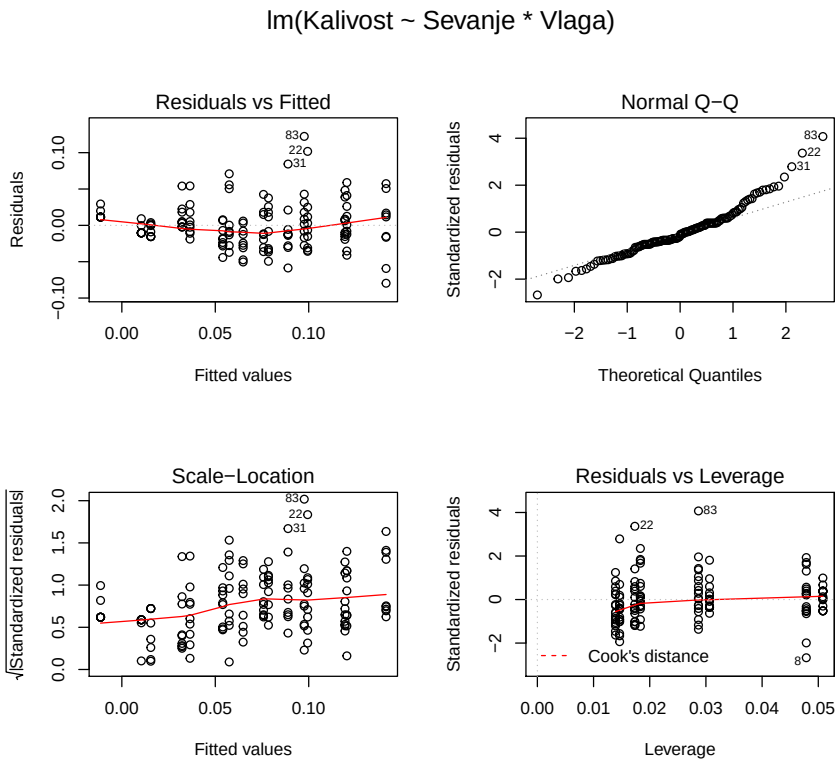
```
> ggplot(data=pelod, aes(x=Sevanje, y=Kalivost, col=Vlaga)) +
+    geom_point() + geom_smooth(method="lm", se=FALSE) +
+    xlab("Sevanje (Gy)") +  ylab("Kalivost")
```



Slika 16: `Kalivost` glede na spremenljivki `Sevanje` in `Vlaga` (HH, RH)

Slika 16 nakazuje, da je variabilnost za kalivost pri različnih odmerkih sevanja različna: pri manjših odmerkih sevanja je kalivost večja in njena variabilnost tudi. Poglejmo, kako to heteroskedastičnost vidimo na ostankih modela.

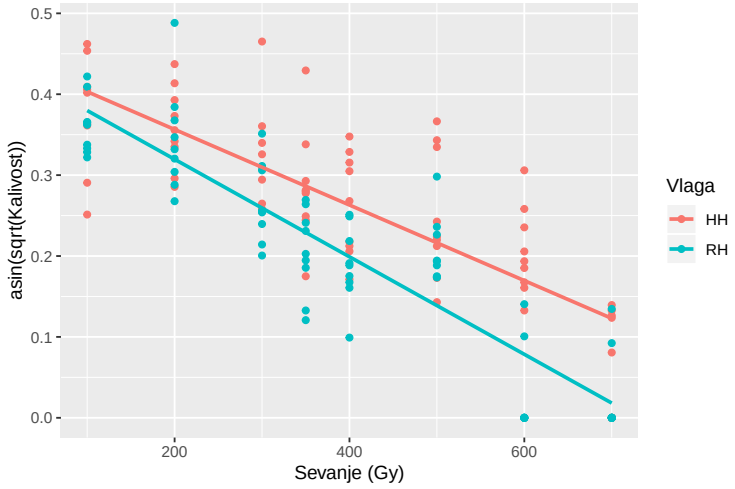
```
> model.p0<-lm(Kalivost~Sevanje*Vlaga, data=pelod)
```



Slika 17: Ostanki za `model.p0`

Za nadaljnjo analizo bomo uporabili `asin(sqrt(p))`, `logit` transformacija ni primerna, ker so med podatki za `Kalivost` ničle. Narišimo izhodiščne podatke z uporabo `asin(sqrt(p))` transfomacije:

```
> ggplot(data=pelod, aes(x=Sevanje, y=asin(sqrt(Kalivost)), col=Vlaga)) +
+    geom_point() +  geom_smooth(method="lm", se=FALSE) +
+    xlab("Sevanje (Gy)") +  ylab("asin(sqrt(Kalivost))")
```

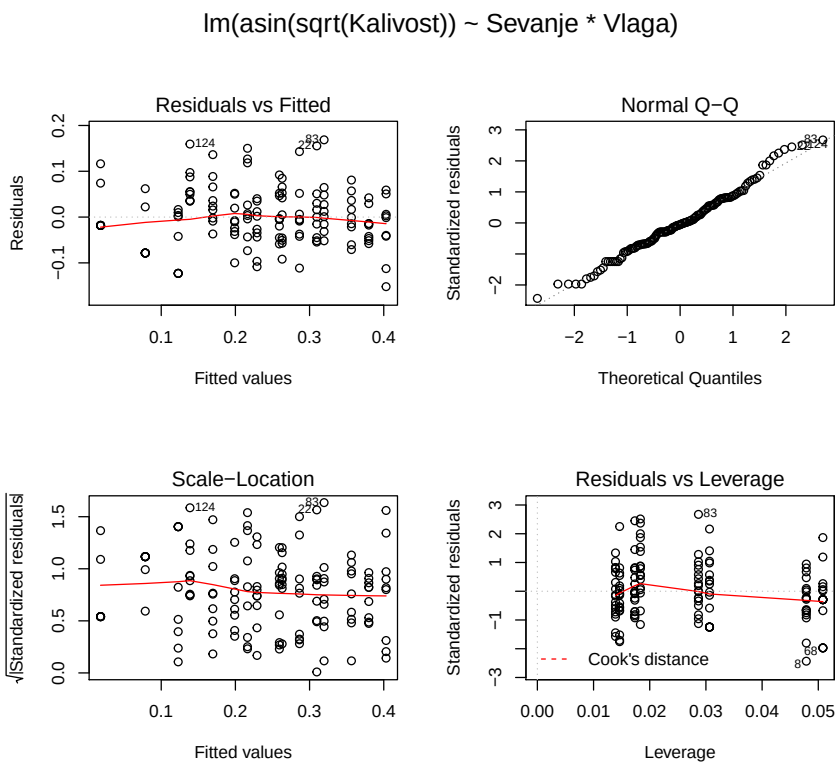


Slika 18: `asin(sqrt(Kalivost))` glede na spremenljivki `Sevanje` in `Vlaga` (HH, RH)

Slika 18 kaže, da je varianca transformirane kalivosti manj problematična. Naredimo model na transformirani spremenljivki:

Ob primerjavi Slik 16 in 18 vidimo, da sta premici na prvi sliki skoraj vzporedni, premici na transformiranih podatkih pa se razlikujeta v naklonih. Naredimo model in ugotovimo, ali je razlika naklonov statistično značilna.

```
> model.p1<-lm(asin(sqrt(Kalivost))~Sevanje*Vlaga, data=pelod)
```



Slika 19: Ostanki za `model.p1`

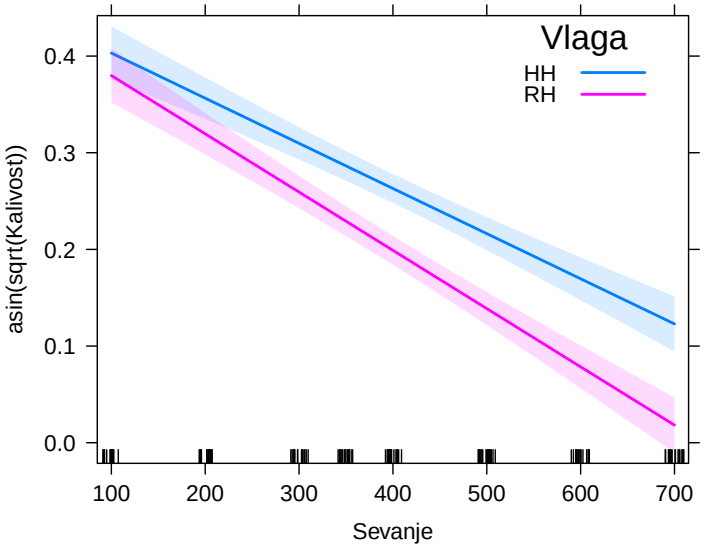
Ostanki na Sliki 19 kažejo, da je s transformacijo heteroskedastičnost odpravljena. Poglejmo povzetek modela:

```
> summary(model.p1)$r.squared

[1] 0.7386538

> summary(glht(model.p1))

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = asin(sqrt(Kalivost)) ~ Sevanje * Vlaga, data = pelod)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0      4.499e-01  1.752e-02  25.682   <0.001 ***
Sevanje == 0         -4.671e-04  4.015e-05 -11.633   <0.001 ***
VlagaRH == 0         -9.743e-03  2.477e-02  -0.393    0.950
Sevanje:VlagaRH == 0 -1.356e-04  5.678e-05  -2.388    0.045 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Z modelom pojasnimo 74 % variabilnosti `asin(sqrt(Kalivost))`.

Linearna odvisnost `asin(sqrt(Kalivost))` od `Sevanje` je statistično značilna, naklona za HH in RH sta statistično različna ($p = 0.0454$), pri RH je naklon večji kot pri HH (Slika 20). Ocena presečišča v modelu vsebinsko ni zanimiva.

```
> plot(Effect(c("Sevanje","Vlaga"), model.p1), multiline=T, ci.style="bands",
+      key.args=list(x=0.7, y=0.8, corner=c(0,0)), main="" )
```



Slika 20: Napovedi za `asin(sqrt(Kalivost)` glede na `Sevanje` in `Vlaga` (RH, HH) s 95 % intervali zaupanja za povprečno napoved za `model.p1`

Zaradi uporabljene nelinearne transformacije vsebinska interpretacija naklona premic ni mogoča. Pri izbranem sevanju in vlagi lahko izračunamo napoved za `asin(sqrt(Kalivost))` in na izračunani vrednosti uporabimo inverzno transformacijo $(sin(\hat{y}))^2$.

4.4 Pljučna kapaciteta

V podatkovnem okviru `lungcap` iz paketa `GLMsData` so podatki o pljučni kapaciteti (litri), starosti (dopolnjena leta), telesni višini (inče), spolu in kajenju za vzorec mladostnikov v Bostonu sredi sedemdesetih let (Kahn in Michael, 2005).

- Naredite statistični povzetek za vse spremenljivke v naboru podatkov, spremenljivke smiselno grafično prikažite in na kratko obrazložite. Podatke za telesno višino preračunajte v cm.
- Analizirajte odvisnost pljučne kapacitete od starosti, spola in kajenja. Za izbrani model naredite diagnostiko in ga obrazložite. Grafično prikažite napovedi modela s 95 % intervali zaupanja za povprečno napoved.