Kazalo

1 NAPOVEDNA SPREMENLJIVKA JE OPISNA

V model vključimo napovedno spremenljivko x, ki je opisna/kategorična, recimo, da ima l vrednosti $(a_1, a_2, ...a_l)$. Taka spremenljivka podatke na nek način deli v l skupin. V model jo vključimo tako, da na podlegi njenih vrednosti naredimo $l-1$ regresorjev, ki so umetne spremenljivke z vrednostmi 0 in 1 (*dummy variables*). Označimo jih 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: $\beta_j = \mu_{a_j} - \mu_{a_1}$, $j = 1, ..., l-1$.

Model zapišemo

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

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

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

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:

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

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, $\mu_{a_2} - \mu_{a_1}$. V okviru inference linearnega modela v tem primeru testiramo ničelni domnevi

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

in

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

Primer: vpliv spola na povprečen SKT

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

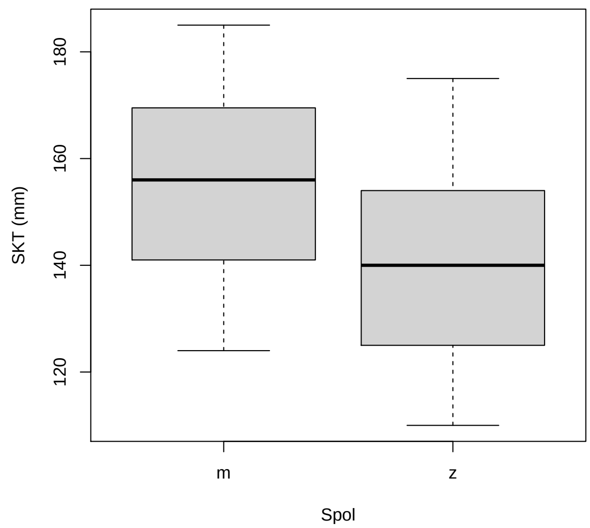
```
> levels(tlak$spol)
```

```
[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

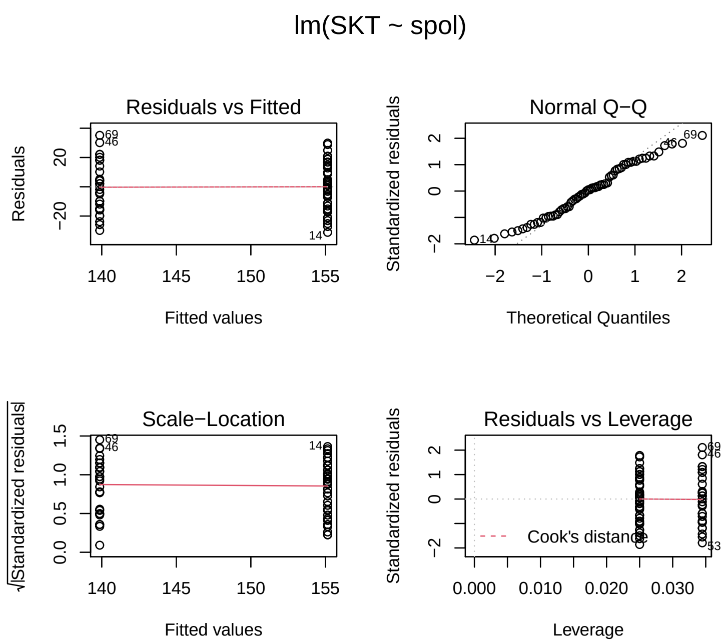
V `lm` model damo kot napovedno spremenljivko opisno spremenljivko `spol`. Funkcija `lm` na podlagi faktorja `spol` naredi regresor `spolz` z vrednostma 0 za moške in 1 za ženske.

```
> 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 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 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)$coeff

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 155.15000   2.682346 57.841156 6.406345e-59
spolz       -15.28793   4.137522 -3.694948 4.445916e-04

> summary(model.spol)$r.squared

[1] 0.1692771
```

Povprečje pri moških (`Intercept`) je 155.2 mm in je statistično značilno različno od 0 ($p < 0.0001$), testiranje te ničelne domneve je vsebinsko nesmiselno. 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:

```
                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

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

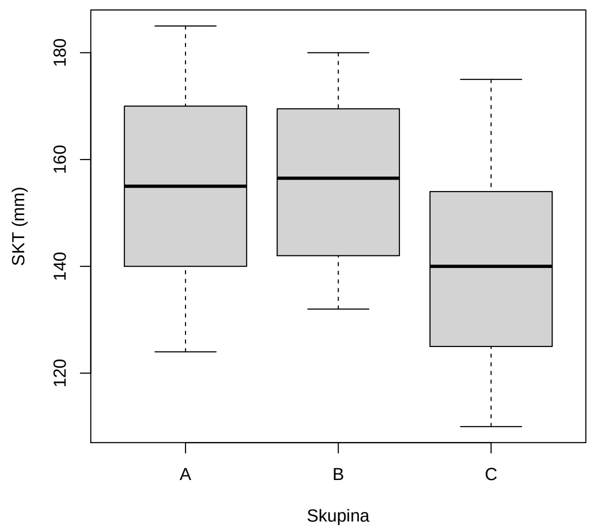
kar pomeni, da je pričakovana vrednost $E(y_i)$ enaka:

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

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
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)$coeff

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 154.85000   3.821683 40.518794 2.363652e-48
skupinaB      0.60000   5.404676  0.111015 9.119414e-01
skupinaC    -14.98793   4.967682 -3.017088 3.622919e-03
```

Ničelne domneve v povzetku `lm` modela so:

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

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.

1.3 Hkratno testiranje 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 $\boldsymbol{\beta}$:

$$H_0 : \mathbf{c_j}^{\mathrm{T}}\boldsymbol{\beta} = \delta, \tag{12}$$

$\mathbf{c_j}$ 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:

$$H_{0j} : \mathbf{c}_j^{\mathrm{T}} \boldsymbol{\beta} = \delta_j, \quad j = 1, ..., m. \tag{13}$$

Izračunamo m testih statistik

$$t_j = \frac{\mathbf{c}_j^{\mathrm{T}} \mathbf{b} - \delta_j}{\hat{\sigma}\sqrt{\mathbf{c}_j^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{c}_j}}. \tag{14}$$

V števcu (14) 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** oziroma asimptotsko, ko so stopinje prostosti ostanka $n - k - 1$ velike, **multivariatna normalna porazdelitev**. Obliko te porazdelitve določajo stopinje prostosti ostanka modela ($df_{residual}$) in korelacijska matrika t_j-statistik $\mathbf{R}$, ki se izračuna takole:

$$\mathbf{R} = \mathbf{D}^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{C}\mathbf{D}. \tag{15}$$

V enačbi (15) je $\mathbf{D}$ diagonalna matrika, $\mathbf{D} = \hat{\sigma} diag(\mathbf{c}_j^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{c}_j)^{-1/2}$, $j = 1, ...m$, 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 oziroma multivariatne normalne porazdelitve, za kateri 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.

1.4 Osnovna matrika primerjav v `lm` modelu

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

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

              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 154.85000   3.821683 40.518794 2.363652e-48
skupinaB      0.60000   5.404676  0.111015 9.119414e-01
skupinaC    -14.98793   4.967682 -3.017088 3.622919e-03
```

```
> 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: $\beta_0 = \mu_A = 0$, H_0: $\beta_1 = \mu_B - \mu_A = 0$ in 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 $\mathbf{R}$ (15) za `model.vec`, za katerega smo ocenili tri parametre. Hkratno želimo testirati tri enostavne ničelne domneve $H_0 : \beta_j = 0$, $j = 0, ..., 2$.

```
> b <- coefficients(model.vec)
> k <- length(b)-1
> 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`.

```
> n <- length(tlak$SKT)
> 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 0.0089
```

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 =0.0089), kar pomeni, daje povprečni SKT v skupini C statistično značilno različen od povprečnega SKT v skupini A.

1.5 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  0.00893 **
---
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 = 2.3509
95% family-wise confidence level


Linear Hypotheses:
                 Estimate lwr      upr
(Intercept) == 0 154.8500 145.8655 163.8345
skupinaB == 0      0.6000 -12.1060  13.3060
skupinaC == 0    -14.9879 -26.6666  -3.3093
```

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

1.6 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: $\beta_1 = \mu_B - \mu_A = 0$ H$_0$: $\beta_2 = \mu_C - \mu_A = 0$ in 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  0.00993 **
mu_C-mu_B == 0  -15.588      4.968  -3.138  0.00714 **
---
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 -12.3521  13.5521
mu_C-mu_A == 0 -14.9879 -26.8927  -3.0831
mu_C-mu_B == 0 -15.5879 -27.4927  -3.6831
```

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.

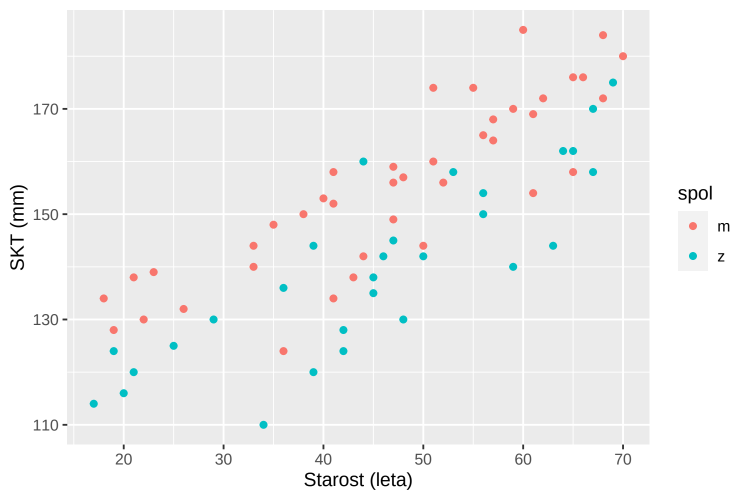
2 OPISNA IN ŠTEVILSKA NAPOVEDNA SPREMENLJIVKA

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

2.1 Dve regresijski premici

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

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



Slika 4: 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:

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 \texttt{starost}_i + \varepsilon_i, \tag{16}$$

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:

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 \texttt{starost}_i + \beta_3 \texttt{starost}_i w_i + \varepsilon_i, \tag{18}$$

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.

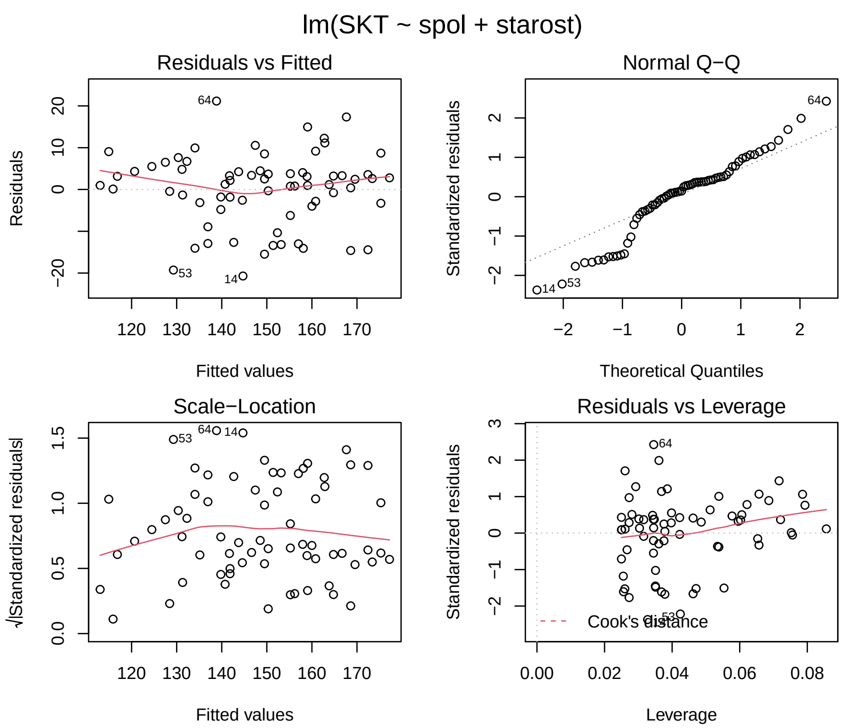
2.1.1 Model brez interakcije

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 5: Grafični prikaz ostankov za `model.vzporedni`

Slika 5 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-07 ***
spolz == 0       -13.51345    2.16932  -6.229   <1e-07 ***
starost == 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)

         Simultaneous Confidence Intervals

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

Quantile = 2.3543
95% family-wise confidence level


Linear Hypotheses:
                 Estimate lwr      upr
(Intercept) == 0 110.2870 101.7214 118.8525
spolz == 0       -13.5135 -18.6207  -8.4062
starost == 0       0.9561   0.7877   1.1245
```

Enačbi vzporednih premic sta:

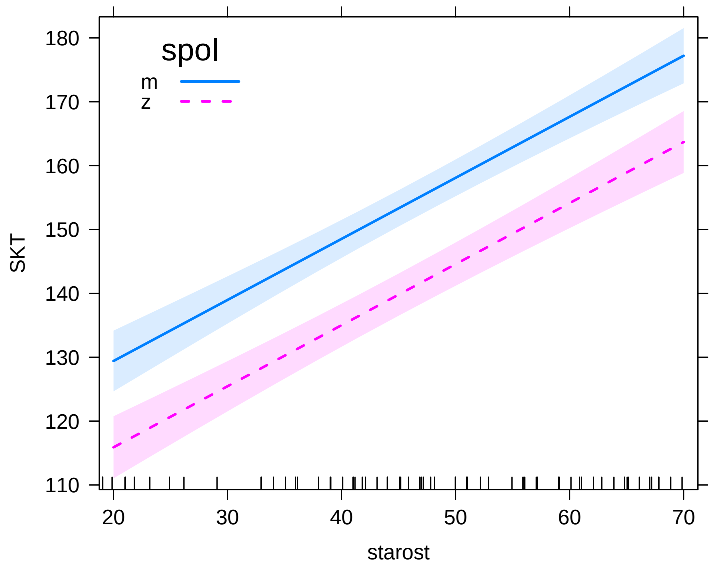
$$\begin{aligned} \text{Moški:} \quad & \widehat{SKT} = 110.29 + 0.96\ \texttt{starost}. \\ \text{Ženske:} \quad & \widehat{SKT} = (110.29 + (-13.51)) + 0.96\ \texttt{starost}. \end{aligned}$$

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))
```



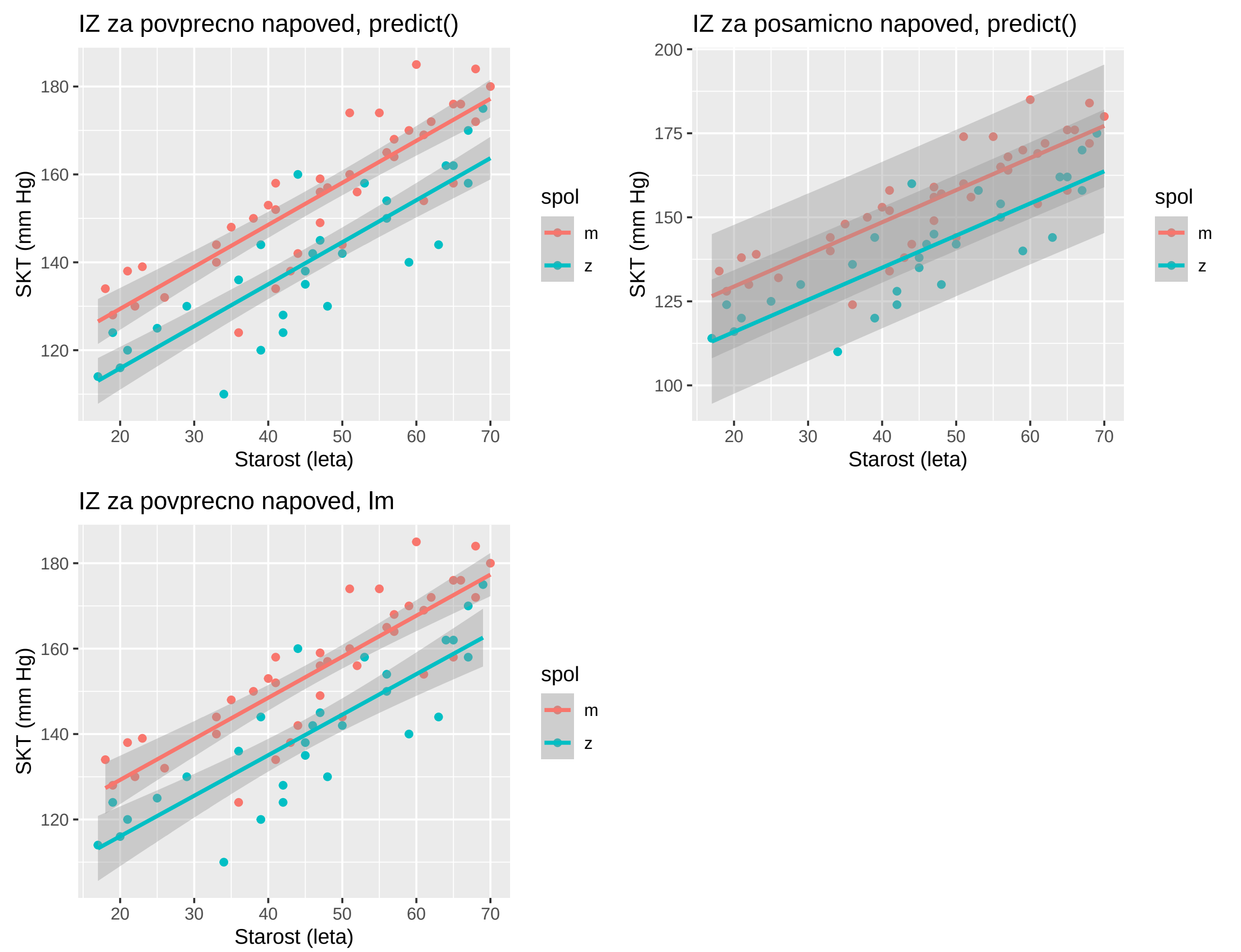
Slika 6: Odvisnost SKT od starosti in spola, premici dobljeni po `model.vzporedni` z 95 % intervali zaupanja za povprečne napovedi

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 7: 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 1.15e-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 161.5275
spolz == 0       -13.5135 -18.7619  -8.2650
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.

2.1.2 Model z interakcijo

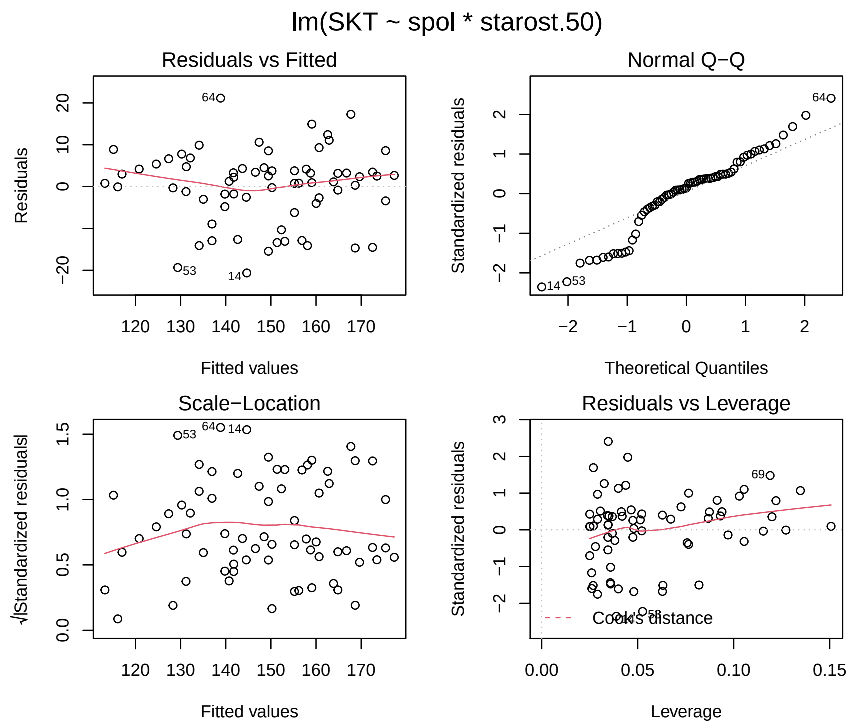
Zanima nas, **kako starost, spol in njuna interakcija vplivajo na SKT.** Za starost bomo upoštevali `starost.50`.

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 8: 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 2.38e-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 = 2.5134
95% family-wise confidence level


Linear Hypotheses:
                      Estimate  lwr       upr
(Intercept) == 0      158.10616 154.47407 161.73825
spolz == 0            -13.56295 -19.25826  -7.86764
starost.50 == 0         0.96135   0.71925   1.20345
spolz:starost.50 == 0  -0.01203  -0.37696   0.35290
```

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

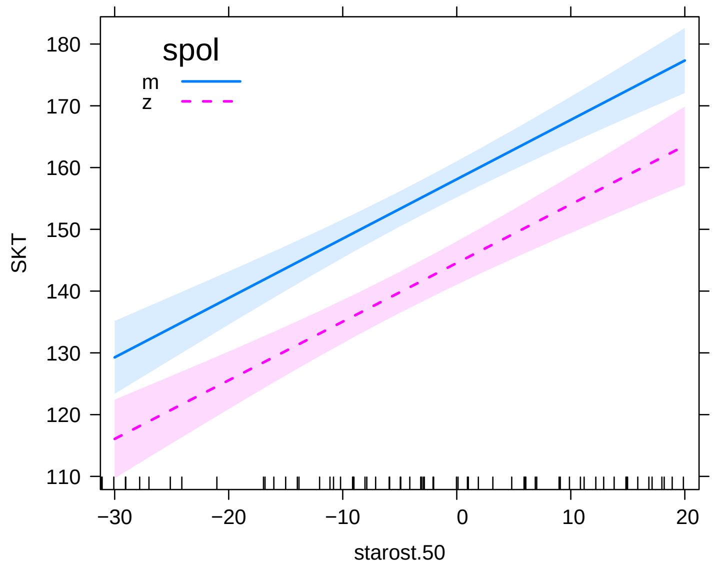
Moški: $\widehat{SKT} = 158.11 + 0.96\ \texttt{starost.50}$,

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

Vsebinsko obrazložite rezultate.

Grafični prikaz napovedi za `model.razlicni` je na Sliki 9:

```
> 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 9: Odvisnost SKT od centrirane starosti, spola in njune interakcije, premici dobljeni po `model.razlicni` z intervali zaupanja za povprečne napovedi

2.2 Več regresijskih premic

2.2.1 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: $\beta_3 = naklon = 0$.

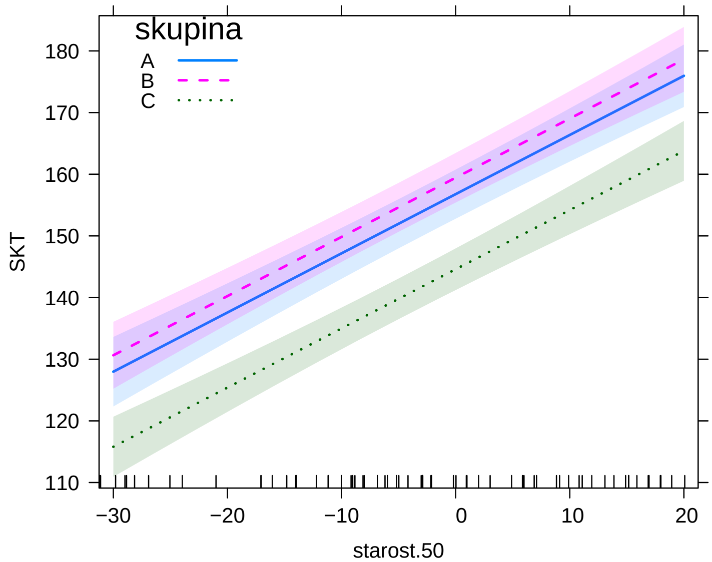
```
> test.vzporedne<-glht(model.vzporedne)
```

```
> 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 = 2.4964
95% family-wise confidence level


Linear Hypotheses:
                 Estimate lwr      upr
(Intercept) == 0 156.7696 151.7970 161.7421
skupinaB == 0      2.6635  -4.3611   9.6881
skupinaC == 0    -12.1748 -18.6430  -5.7066
starost.50 == 0    0.9598   0.7808   1.1388
```

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 10.

```
> 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 10: 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 = 2.5254
95% family-wise confidence level


Linear Hypotheses:
                                     Estimate lwr      upr
povp B|starost - povp A|starost == 0   2.6635  -4.4428   9.7698
povp C|starost - povp A|starost == 0 -12.1748 -18.7182  -5.6313
povp C|starost - povp B|starost == 0 -14.8383 -21.3617  -8.3149
naklon == 0                            0.9598   0.7787   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`.

2.2.2 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    0.922
skupinaC == 0            -12.37731    2.68078  -4.617   <1e-04 ***
starost.50 == 0            1.03526    0.13512   7.662   <1e-04 ***
skupinaB:starost.50 == 0  -0.13881    0.19416  -0.715    0.942
skupinaC:starost.50 == 0  -0.08594    0.17370  -0.495    0.989
---
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 = 2.6321
95% family-wise confidence level
```

```
Linear Hypotheses:
                         Estimate  lwr       upr
(Intercept) == 0         156.92052 151.58328 162.25777
skupinaB == 0              2.24975  -5.41747   9.91696
skupinaC == 0            -12.37731 -19.43350  -5.32112
starost.50 == 0            1.03526   0.67960   1.39092
skupinaB:starost.50 == 0  -0.13881  -0.64988   0.37226
skupinaC:starost.50 == 0  -0.08594  -0.54314   0.37126
```

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-08 ***
naklon_B == 0   0.8965     0.1394   6.429 4.13e-08 ***
naklon_C == 0   0.9493     0.1091   8.697  < 1e-08 ***
---
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.

3 VAJE

3.1 `model.spol`

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

3.2 `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.

3.3 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.

3.4 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.