

## Kazalo

<b>1</b>	<b>LINEARNI REGRESIJSKI MODEL</b>	<b>1</b>
<b>2</b>	<b>ENOSTAVNA LINEARNA REGRESIJA</b>	<b>5</b>
2.1	Statistične lastnosti ocen parametrov . . . . .	5
2.2	Koeficient determinacije . . . . .	7
2.3	Analiza variance . . . . .	8
2.4	Napovedovanje . . . . .	8
2.5	Primer: SKT . . . . .	9
2.6	Simulacija po predpostavkah modela enostavne linearne regresije . . . . .	17
<b>3</b>	<b>LINEARNI MODEL V MATRIČNI OBLIKI</b>	<b>23</b>
3.1	Ocene parametrov . . . . .	23
3.2	Pričakovana vrednost ocen parametrov in njihova variančno-kovariančna matrika . .	24
3.3	Matrika $\mathbf{H}$ . . . . .	24
3.4	Ostanki . . . . .	25
<b>4</b>	<b>VAJE</b>	<b>26</b>
4.1	Čas teka Collina Jacksona . . . . .	26
4.2	Simulacije za enostavno linearno regresijo . . . . .	26

## 1 LINEARNI REGRESIJSKI MODEL

*A statistical model is a collection of assumptions that has sufficient structure to allow us to estimate interesting quantities, to use past data to predict future values, and to perform many other tasks. (Fox and Weisberg (2011): An R Companion to Applied Regression, str. 149).*

*... all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind. (Box and Draper (1987), str. 424)*

Predpostavke za linearni regresijski model so:

1. Imamo **odzivno spremenljivko** (*response variable*)  $y$  in  $m$  **napovednih/pojasnjevalnih spremenljivk** (*predictors/explanatory variables*).
2. Odzivna spremenljivka  $y$  je številska, njene vrednosti so medsebojno neodvisne.
3. Napovedne spremenljivke so lahko številske (npr. starost) in/ali opisne (npr. spol, ocena vrednotena na petmestni lestvici). Vse te spremenljivke določajo t. i. **regresorje**; to so številske spremenljivke, ki so vključene v model. Iz napovednih spremenljivk dobimo regresorje na različne načine:
  - številsko spremenljivko v model vključimo direktno kot en regresor; včasih je ta spremenljivka predhodno transformirana (npr. *log*). V določenih primerih je številska spremenljivka vključena v model z več regresorji (npr. polinomska regresija, zleпки);
  - za opisno spremenljivko z  $d$  vrednostmi se v model vključi  $d - 1$  regresorjev z vrednostmi 0 in 1 (neme spremenljivke, *dummy variables*);

- dodatne regresorje lahko dobimo z upoštevanjem interakcij med obstoječimi napovednimi spremenljivkami v modelu.

Torej,  $m$  napovednih spremenljivk generira  $k$  regresorjev, ki jih označimo  $x_1, \dots, x_k$ ,  $k \geq m$ .

4. Pričakovana vrednost odzivne spremenljivke pogojno na regresorje je linearna kombinacija regresorjev  $x_j$ :

$$E(y|x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k. \quad (1)$$

Pri tem so  $\beta_j$ ,  $j = 0, \dots, k$ , parametri modela, ki so v linearnem odnosu z odzivno spremenljivko. Te parametre ocenjujemo v postopku linearnega modeliranja.

5. Varianca odzivne spremenljivke  $y$  pogojno na regresorje  $x_1, \dots, x_k$  je konstantna, ta lastnost se imenuje **homoskedastičnost**:

$$Var(y|x_1, \dots, x_k) = \sigma^2 > 0. \quad (2)$$

Vrednosti odzivne in napovednih spremenljivk so dobljene na vzorcu, ki ima  $n$  enot. Linearni regresijski model za  $i$ -to enoto,  $i = 1, \dots, n$ , zapišemo takole:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad (3)$$

$\varepsilon_i$  se imenuje **napaka** (*error*). Privzete predpostavke se zrcalijo v njenih lastnostih:

$$E(\varepsilon|x_1, \dots, x_k) = 0, \quad (4)$$

$$Var(\varepsilon|x_1, \dots, x_k) = \sigma^2 \quad \text{oziroma} \quad Var(\varepsilon_i|x_1, \dots, x_k) = \frac{\sigma^2}{w_i}, \quad i = 1, \dots, n, \quad (5)$$

$\varepsilon_i$  so medsebojno neodvisni. Varianca napak je v splošnem lahko obratno sorazmerna z znanimi pozitivnimi utežmi  $w_i$ ,  $i = 1, \dots, n$  (apriorne uteži).

Najbolj pogosta dodatna predpostavka je, da je porazdelitev  $y$  pri  $x_1, \dots, x_k$  normalna; ta model imenujemo **normalni linearni model**:

$$y|x_1, \dots, x_k \sim N(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k, \sigma^2), \quad (6)$$

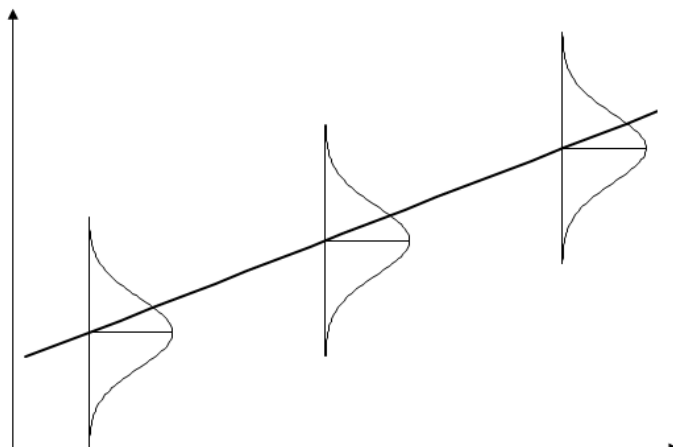
$y$  je slučajna spremenljivka, njena porazdelitev je pri vsakem  $\mathbf{x} = (x_1, \dots, x_k)^T$  normalna s povprečjem na regresijski ravnini in varianco  $\sigma^2$ .

Posledično je porazdelitev napak  $\varepsilon \sim iid N(0, \sigma^2)$ . Oznaka *iid* pomeni neodvisno enako porazdelitev *independent identically distributed*, kar pomeni:

- $\varepsilon$  so porazdeljeni po normalni porazdelitvi;
- homogenost variance:  $Var(\varepsilon_i|x_1, \dots, x_k) = \frac{\sigma^2}{w_i}$ , nepoznani del variance  $\sigma^2$  je konstanta;
- $\varepsilon$  so neodvisni.

Ilustracija predpostavk za normalni regresijski model z enim številskim regresorjem je na Sliki 1.

Normalni linearni model sodi v širšo skupino posplošenih linearnih modelov (*Generalized Linear Models*, GLM), kjer za odzivno spremenljivko privzamemo druge verjetnostne porazdelitve (logistična regresija, Poissonova regresija,...).



Slika 1: Predpostavke za normalni linearni regresijski model z eno napovedno številsko spremenljivko

V praksi ocenjujemo parametre linearnega modela na podlagi vrednosti odzivne spremenljivke in napovednih spremenljivk dobljenih na vzorcu  $n$  enot. Za oceno parametrov uporabimo **metodo najmanjših kvadratov** (*Ordinary Least Squares*, OLS) ali pa **tehtano metodo najmanjših kvadratov** (*Weighted Least Squares*, WLS), s katero minimiramo vsoto kvadriranih odklonov oziroma tehtano vsoto kvadriranih odklonov vrednosti odzivne spremenljivke od njene pričakovane vrednosti. Funkcija  $S(\beta_0, \beta_1, \dots, \beta_k)$  ima  $k + 1$  parametrov:

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}))^2. \quad (7)$$

ali v primeru tehtane vsote kvadratov odklonov

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n w_i (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}))^2. \quad (8)$$

Funkcijo (7) ali (8) parcialno odvajamo po parametrih  $\beta_j$ ,  $j = 0, \dots, k$ , in odvode izenačimo z 0. Dobimo t. i. normalni sistem  $k + 1$  linearnih enačb. Rešitev tega sistema so ocene parametrov, ki jih označimo  $b_j$ ,  $j = 0, \dots, k$ .

Z modelom **napovedane vrednosti** (*fitted values*) označimo  $\hat{y}_i$ :

$$\hat{y}_i = b_0 + b_1 x_{i1} + \dots + b_k x_{ik}, \quad i = 1, \dots, n. \quad (9)$$

Razliko med dejansko vrednostjo  $y_i$  in napovedano vrednostjo  $\hat{y}_i$  imenujemo **ostanek** (*residual*):

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n. \quad (10)$$

Matematična teorija pokaže, da so ostanki  $e_i$  nekorelirani z napovedanimi vrednostmi  $\hat{y}_i$ , kar uporabljamo pri analizi modela z grafičnimi prikazi (dokaz v poglavju 3.4).

Varianca ostanka je (dokaz v poglavju 3.4):

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii}). \quad (11)$$

Pri izračunu moramo varianco  $\sigma^2$  oceniti, oceno variance označimo  $s^2$ . Količina  $h_{ii}$  se imenuje vzvod (*leverage* ali *hat-value*);  $h_{ii}$  je vrednost med  $1/n$  in 1, o tem kasneje.

Če v izrazu (11) uporabimo ocenjeno varianco  $s^2$ , lahko izračunamo **standardizirane ostanke**  $e_{s_i}$ :

$$e_{s_i} = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, \dots, n, \quad (12)$$

za katere pri pogoju, da je  $n \gg k$  velja, da je njihova porazdelitev približno  $N(0, 1)$ .

Ali so predpostavke modela izpolnjene, ugotavljamo z analizo ostankov in standardiziranih ostankov.

Če lahko privzamemo normalni linearni model, matematična statistika pove, da so porazdelitve parametrov normalne z znanimi parametri, kar omogoča dodatne izračune:

- za vsako oceno parametra lahko izračunamo njeno standardno napako;
- izračunamo interval zaupanja za vsak parameter modela;
- testiramo lahko domneve o parametrih modela;
- izračunamo napovedi in intervale zaupanja za povprečno napoved in za posamično napoved.

## 2 ENOSTAVNA LINEARNA REGRESIJA

Če je v modelu (3) samo ena napovedna spremenljivka  $x$ , ki je številska, ta model imenujemo enostavna linearna regresija:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n. \quad (13)$$

Ocenjujemo dva parametra modela  $\beta_0$  in  $\beta_1$ . Vsoto kvadriranih odklonov  $y_i$  od njene pričakovane vrednosti  $(\beta_0 + \beta_1 x_i)$ :

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (14)$$

parcialno odvajamo po  $\beta_0$  in  $\beta_1$ :

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i), \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i). \end{aligned}$$

Ko odvoda izenačimo z 0, dobimo sistem dveh linearnih enačb:

$$\begin{aligned} \sum_{i=1}^n y_i &= n b_0 + b_1 \sum_{i=1}^n x_i, \\ \sum_{i=1}^n x_i y_i &= b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2. \end{aligned}$$

Z rešitvijo sistema enačb dobimo oceni za  $\beta_0$  in  $\beta_1$ , označimo ju  $b_0$  in  $b_1$ :

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x}, \\ b_1 &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2 = SS_{xy} / SS_{xx}, \end{aligned} \quad (15)$$

kjer je  $SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$  in  $SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ . Model enostavne linearne regresije predstavlja premico, ki se po metodi najmanjših kvadratov najbolj prilaga podatkom. Ocena  $b_0$  predstavlja presečišče premice z ordinatno osjo,  $b_1$  pa njen naklon. Enačba regresijske premice je:

$$\hat{y} = b_0 + b_1 x. \quad (16)$$

Model velja na intervalu vrednosti napovedne spremenljivke  $[x_{min}, x_{max}]$ .

### 2.1 Statistične lastnosti ocen parametrov

Ob danih predpostavkah za linearni regresijski model matematična statistika pokaže (Rice, 1995, str. 513-514):

- oceni parametrov  $b_0$  in  $b_1$  sta nepristranski:

$$E(b_0) = \beta_0, \quad E(b_1) = \beta_1. \quad (17)$$

- varianci in kovarianca ocen parametrov so:

$$\begin{aligned} Var(b_0) &= \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right), \\ Var(b_1) &= \frac{\sigma^2}{SS_{xx}}, \\ Cov(b_0, b_1) &= -\frac{\sigma^2 \bar{x}}{SS_{xx}}. \end{aligned} \quad (18)$$

Iz enačb (18) vidimo, da sta varianci in kovarianca ocen parametrov enostavnega linearnega modela odvisni od vrednosti napovedne spremenljivke  $x_i$ ,  $i = 1, \dots, n$ , in od variance napake  $\sigma^2$ . Vrednosti  $x_i$  so znane, varianco napake pa moramo oceniti. Glede na predpostavko linearnega regresijskega modela (5) je  $\sigma^2$  varianca napak  $\varepsilon$  pogojno na regresorje. Napako  $\varepsilon_i$  ocenimo z **ostankom**  $e_i = y_i - \hat{y}_i$  (*residual*) in definiramo **vsoto kvadriranih ostankov** ( $SS_{residual}$ ):

$$SS_{residual} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2. \quad (19)$$

Matematična statistika pokaže (Rice, 1995, str. 540), da je nepristranska ocena za  $\sigma^2$ :

$$s^2 = \frac{SS_{residual}}{n - 2}. \quad (20)$$

V imenovalcu (20) delimo z  $n - 2$  namesto z  $n$ , saj smo dve stopinji prostosti porabili za oceno parametrov modela. Ocenimo varianco parametrov modela  $s_{b_0}^2$  in  $s_{b_1}^2$  izračunamo tako, da v (18)  $\sigma^2$  zamenjamo z  $s^2$ .

Ob predpostavki  $\varepsilon \sim iid N(0, \sigma^2)$  in posledičnem dejstvu, da sta oceni za presečišče in naklon premice linearni kombinaciji normalno porazdeljenih spremenljivk, velja, da je tudi njuna porazdelitev normalna:

$$\begin{aligned} b_0 &\sim N \left( \beta_0, \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right) \right), \\ b_1 &\sim N \left( \beta_1, \frac{\sigma^2}{SS_{xx}} \right). \end{aligned} \quad (21)$$

To omogoča izračun intervalov zaupanja za parametra  $\beta_0$  in  $\beta_1$  ter testiranje domnev o parametrih. Pokažemo namreč lahko, da sta statistiki  $(b_j - \beta_j)/s_{b_j}$ ,  $j = 0, 1$  porazdeljeni po  $t$ -porazdelitvi s  $SP = n - 2$ :

$$\begin{aligned} \frac{b_0 - \beta_0}{s_{b_0}} &\sim t(SP = n - 2), \\ \frac{b_1 - \beta_1}{s_{b_1}} &\sim t(SP = n - 2). \end{aligned}$$

100(1 -  $\alpha$ ) % intervala zaupanja (IZ) za  $\beta_0$  in  $\beta_1$  sta:

$$\begin{aligned} &\left( b_0 - t_{1-\frac{\alpha}{2}}(SP = n - 2)s_{b_0}, \quad b_0 + t_{1-\frac{\alpha}{2}}(SP = n - 2)s_{b_0} \right), \\ &\left( b_1 - t_{1-\frac{\alpha}{2}}(SP = n - 2)s_{b_1}, \quad b_1 + t_{1-\frac{\alpha}{2}}(SP = n - 2)s_{b_1} \right), \end{aligned}$$

$t_{1-\frac{\alpha}{2}}(SP = n - 2)$  je  $(1 - \alpha/2)100$ -ti centil  $t$ -porazdelitve s stopinjami prostosti  $n - 2$ .

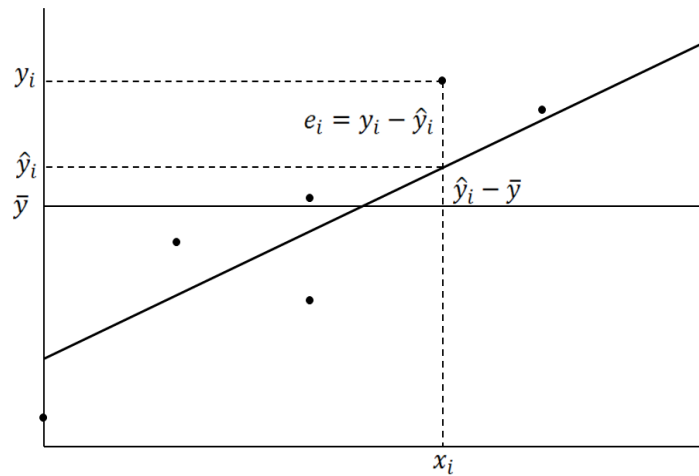
## 2.2 Koeficient determinacije

Koeficient determinacije je enostavna mera, ki opredeljuje kakovost linearnega regresijskega modela; za njen izračun ne potrebujemo nobenih predpostavk. Izhodišče za njegov izračun je razvidno iz Slike 2:  $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$ . To enačbo kvadriramo in seštejemo po vseh enotah,  $i = 1, \dots, n$ . Z upoštevanjem formul za  $b_0$  in za  $b_1$  (15) ugotovimo, da je  $2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$  in ostane:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (22)$$

$$SS_{yy} = SS_{model} + SS_{residual}. \quad (23)$$

Vsota kvadriranih odklonov (*Sum of Squares*) za odzivno spremenljivko  $y$ ,  $SS_{yy}$ , se razdeli na dva dela: na del, ki ga pojasni regresijski model,  $SS_{model}$ , in na del, ki ostane z regresijskim modelom nepojasnen,  $SS_{residual}$ .



Slika 2: Izhodišče za izračun koeficienta determinacije:  $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

Koeficient determinacije  $R^2$  je delež variabilnosti za  $y$ , ki je pojasnjen z regresijskim modelom:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{SS_{model}}{SS_{yy}}. \quad (24)$$

Lastnosti koeficienta determinacije:

- je nenegativna vrednost;
- je manjši ali enak 1; ima vrednost 1, če je  $SS_{model} = SS_{yy}$ , ko so vse točke na premici;
- $R^2$  je odvisen od zaloge vrednosti napovedne spremenljivke;
- previdni moramo biti pri uporabi  $R^2$  v kontekstu multiple regresije, saj vsak dodani regresor poveča vrednost  $R^2$ , tudi če je vpliv tega regresorja na odzivno spremenljivko statistično nepomemben.

## 2.3 Analiza variance

Videli smo, da se vsota kvadriranih odklonov za odzivno spremenljivko  $y$  razdeli na dva dela: na del, ki ga pojasni regresijski model ( $SS_{model}$ ), in na del, ki ostane z regresijskim modelom nepojasnen ( $SS_{residual}$ ). Vsak od teh členov ima pripadajoče stopinje prostosti  $Df$  (*Degrees of freedom*). Stopinje prostosti za skupno variabilnost so  $n - 1$ , kjer je  $n$  število enot, ki so vključene v model. Stopinje prostosti za regresijski model z eno napovedno spremenljivko so  $k = 1$ , kjer je  $k + 1 = 2$  število ocenjenih parametrov v modelu. Stopinje prostosti za ostanek,  $Df_{residual}$  predstavljajo razliko, torej  $n - 2$ .

Opredelimo še srednji kvadrirani odklon (*Mean Square*) za posamezno komponento,  $MS = SS/Df$ . Vse omenjene količine po virih variabilnosti uredimo v t. i. tabelo analize variance (ANOVA).

Tabela 1: Shema tabele ANOVA za enostavni linearni regresijski model

Vir variabilnosti	$Df$	$SS$	$MS = SS/df$	$F$
Model	1	$SS_{model}$	$MS_{model}$	$MS_{model}/MS_{residual}$
Ostanek ( <i>Residual</i> )	$n - 2$	$SS_{residual}$	$MS_{residual}$	
Skupaj	$n - 1$	$SS_{skupaj}$		

Ob predpostavki  $\varepsilon \sim iid N(0, \sigma^2)$  za  $F$ -statistiko velja, da je njena ničelna porazdelitev  $F$ -porazdelitev s stopinjami prostosti  $SP_1 = 1$  in  $SP_2 = n - 2$  (Dokaz: Rice, 1995, str. 448).

Iz tabele ANOVA dobimo:

- oceno za varianco  $\sigma^2$ , ki jo označimo  $s^2$ . Teorija pove, da je po metodi najmanjših kvadratov ta ocena enaka  $MS_{residual}$ . Količino  $s$  imenujemo **standardna napaka regresije** (*Residual standard error*).
- $F$ -statistika testira domnevo o ničelnem vplivu napovedne spremenljivke:  
 $H_0 : \beta_1 = 0$ ,  
 $H_1 : \beta_1 \neq 0$ .

## 2.4 Napovedovanje

Za vsako vrednost  $x_0$  na intervalu  $[x_{min}, x_{max}]$  smemo izračunati napoved  $\hat{y}(x_0) = b_0 + b_1x_0$ ; to je napoved za povprečje  $\bar{y}(x_0)$  in jo imenujemo **povprečna napoved**. Ocena variance povprečne napovedi je:

$$\widehat{Var}(b_0 + b_1x_0) = s^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right). \quad (25)$$

Napoved posamične vrednosti  $y(x_0)$  imenujemo **posamična napoved** in je enaka povprečni napovedi. Ocena variance posamične napovedi pa je za člen  $s^2$  večja od variance povprečne napovedi:

$$\widehat{Var}(b_0 + b_1x_0 + e_0) = s^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right). \quad (26)$$

Varianca za povprečno napoved in za posamično napoved je v obeh primerih odvisna od  $x_0$  in narašča s kvadratom razdalje od povprečja  $\bar{x}$ ; najmanjša je pri povprečju. Meje pripadajočih intervalov zaupanja so na hiperbolah (Slika 6).



## 2.5 Primer: SKT

Analizirajmo odvisnost sistoličnega krvnega tlaka (SKT) od starosti oseb (datoteka SKT.txt).

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

```
      spol SKT starost
1      m 158      41
2      m 185      60
3      m 152      41
4      m 159      47
5      m 176      66
6      m 156      47
```

```
> 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 ...
```

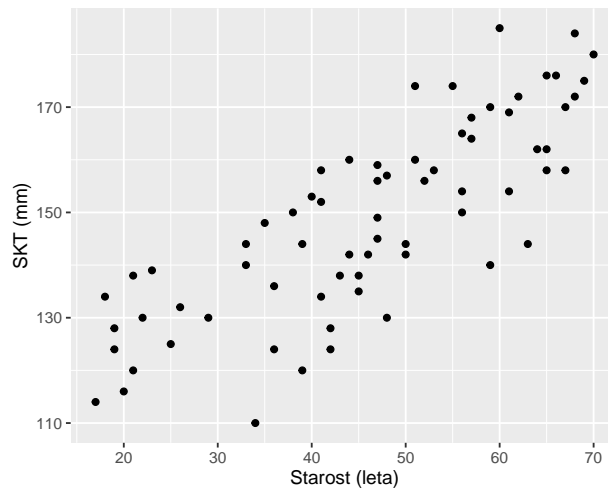
Analiziramo opisne statistike za SKT in starost.

```
> summary(tlak)
```

spol	SKT	starost
m:40	Min. :110.0	Min. :17.00
z:29	1st Qu.:135.0	1st Qu.:36.00
	Median :149.0	Median :47.00
	Mean :148.7	Mean :46.14
	3rd Qu.:162.0	3rd Qu.:59.00
	Max. :185.0	Max. :70.00

Ustrezni grafični prikaz za analizo odvisnosti SKT od starost je razsewni diagram. Slika 3 kaže, da je linearna zveza primerna za te podatke.

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



Slika 3: Odvisnost SKT od starosti za vzorec 69 oseb

Naredimo linearni regresijski model za odvisnost SKT od starosti, poimenovali ga bomo `model.SKT`.

```
> model.SKT <- lm(SKT~starost, data=tlak)
> # Objekt model.SKT ima naslednje komponente
> names(model.SKT)

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

> model.SKT$coef # ni potrebno napisati celega imena "coefficients"

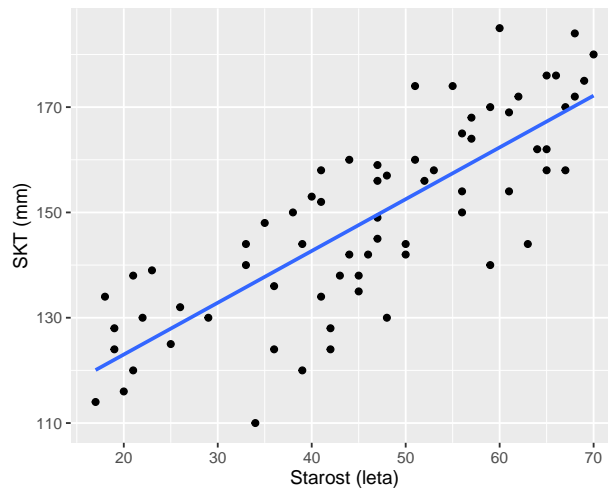
(Intercept)      starost
103.3490547      0.9833276
```

Najbolj osnovni rezultati v objektu `model.SKT` sta ocena za presečišče (`Intercept`) in ocena naklona SKT glede na `starost`. Enačba regresijske premice je (Slika 4):

$$\hat{y} = 103.35 + 0.98 x.$$

- Pričakovan SKT pri novorojenčku je 103.35; ta interpretacija ni vsebinsko smiselna, ker nimamo podatkov za starost oseb mlajših od 17 let.
- Z vsakim letom starosti se SKT v povprečju poveča za 0.98 mm. Bolje povedano: na deset let se SKT v povprečju poveča za 9.8 mm.

```
> ggplot(data=tlak, mapping=aes(x=starost, y=SKT)) +  
+   geom_point() +  
+   geom_smooth(method="lm", se=FALSE) +  
+   xlab("Starost (leta)") +  
+   ylab("SKT (mm)")
```

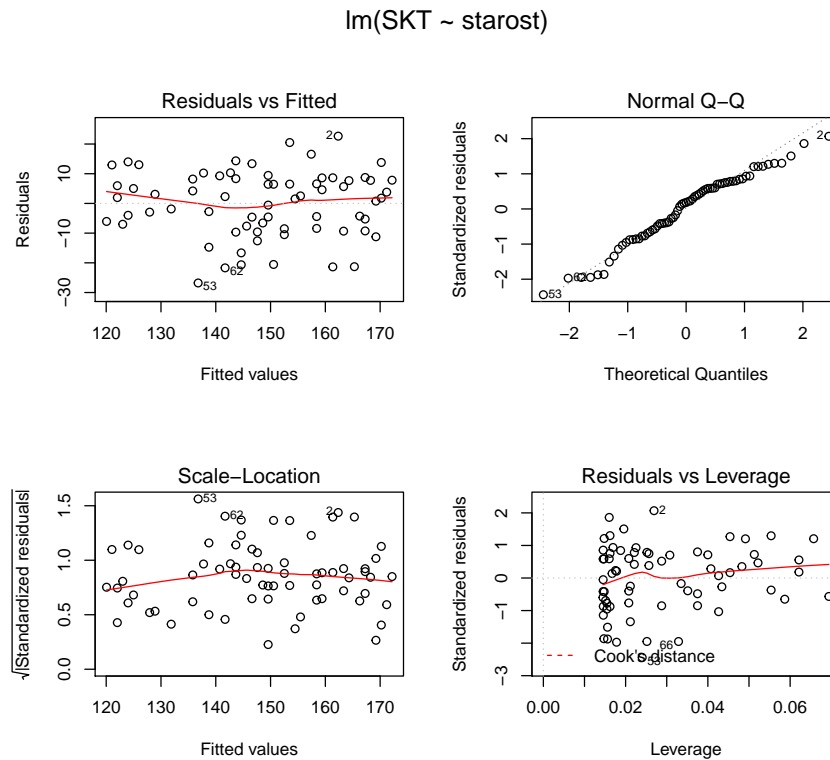


Slika 4: Odvisnost SKT od starosti za vzorec 69 oseb in regresijska premica

Kakovost prilaganja regresijske premice podatkom in izpolnjevanje predpostavk linearnega modela ocenimo na podlagi slik ostankov modela. Z ukazom `plot(model.SKT)` dobimo 4 slike ostankov:

- Graf 1: *Residuals vs Fitted*;
- Graf 2: *Normal Q-Q plot*;
- Graf 3: *Scale-Location plot of  $\sqrt{|\text{residuals}|}$  against fitted values*;
- Graf 4: *Plot of residuals against leverages, and a plot of Cook's distances against leverage/(1-leverage)*.

```
> par(mfrow = c(2, 2), oma=c(0,0,3,0))
> plot(model.SKT)
```



Slika 5: Grafični prikaz ostankov glede na napovedano vrednost (zgoraj levo) ter QQ graf standardiziranih ostankov (zgoraj desno), kvadratni koren absolutne vrednosti standardiziranih ostankov glede na napovedano vrednost (spodaj levo) in standardizirani ostanki glede na vzvod (spodaj desno)

Zaenkrat pogledjmo prva dva grafa. Če model ustreza podatkom, morajo biti točke na Graf 1 razporejene slučajno okoli vrednosti 0, kar pomeni, da je gladilnik približno na osi  $x$ . Če naj bi veljal normalni linearni model, morajo biti standardizirani ostanki porazdeljeni *približno* po  $N(0, 1)$ , to pomeni, da bi morale biti točke na Graf 2 *približno* na črtkani črti. Slika 5 kaže, da model dovolj dobro ustreza podatkom, ni pa idealen.

Povzetek modela dobimo z ukazom `summary(model.SKT)`. Izpis je sestavljen iz treh delov:

- informacija o ostankih;
- informacija o koeficientih regresijskega modela: ocena, standardna napaka,  $t$ -vrednost,  $p$ -vrednost;
- koeficient determinacije in informacija iz tabele ANOVA.

```
> summary(model.SKT)
```

Call:

```
lm(formula = SKT ~ starost, data = tlak)
```

Residuals:

Min	1Q	Median	3Q	Max
-26.782	-7.632	1.968	8.201	22.651

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	103.34905	4.33190	23.86	<2e-16 ***
starost	0.98333	0.08929	11.01	<2e-16 ***

---

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

Residual standard error: 11.1 on 67 degrees of freedom

Multiple R-squared: 0.6441, Adjusted R-squared: 0.6388

F-statistic: 121.3 on 1 and 67 DF, p-value: < 2.2e-16

```
> names(summary(model.SKT)) # za dodatne ali posamezne izpise
```

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

V povzetku modela so standardne napake ocen parametrov (*Std. Error*) izračunane na podlagi variančno-kovariančne matrike ocen parametrov, ki jo vrne funkcija `vcov`. Na njeni diagonali so pripadajoče variance, izven diagonale je kovarianca:

$$\begin{aligned} \text{Var}(b_0) &= s^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right), \\ \text{Var}(b_1) &= \frac{s^2}{SS_{xx}}, \\ \text{Cov}(b_0, b_1) &= -\frac{s^2 \bar{x}}{SS_{xx}}, \end{aligned}$$

$s^2 = SS_{\text{residual}} / (n - 2)$  je ocena za  $\sigma^2$ .

```
> vcov(model.SKT)
```

	(Intercept)	starost
(Intercept)	18.7653188	-0.367938365
starost	-0.3679384	0.007973539

```
> # standardne napake ocen parametrov
```

```
> sqrt(diag(vcov(model.SKT)))
```

(Intercept)	starost
4.33189553	0.08929467

Za posamezen parameter modela testiramo ničelno domnevo, da je njegova vrednost enaka 0. Ilustrirajmo postopek za  $\beta_0$  in  $\beta_1$ .

$H_0 : \beta_0 = 0$ . Regresijska premica gre skozi izhodišče.

$H_1 : \beta_0 \neq 0$ .

```
> t.b0 <- model.SKT$coef[1]/sqrt(vcov(model.SKT)[1,1]); t.b0
```

```
(Intercept)
      23.8577
```

ali pa

```
> (t.b0 <- coef(summary(model.SKT))[1,1]/coef(summary(model.SKT))[1,2])
```

```
[1] 23.8577
```

$$t = \frac{b_0 - 0}{s_{b_0}} = \frac{103.34905}{4.3319} = 23.86.$$

```
> # p-vrednost
```

```
> p.b0 <- 2*pt(abs(t.b0), df=model.SKT$df.residual, lower.tail=FALSE); p.b0
```

```
[1] 1.836799e-34
```

Ničelno domnevo, da gre premica skozi izhodišče, zavrnemo ( $p < 0.0001$ ). Opomba: rezultat ni vsebinsko smiseln. Testiranje te domneve je vedno v izpisu, pogosto pa ta domneva ni raziskovalno zanimiva.

$H_0 : \beta_1 = 0$ . Regresijska premica je vzporedna osi  $x$ ; ni odvisnosti SKT od starosti.

$H_1 : \beta_1 \neq 0$ .

```
> # t.b1<-model.SKT$coef[2]/sqrt(vcov(model.SKT)[2,2]); t.b1
```

```
> (t.b1 <- coef(summary(model.SKT))[2,1]/coef(summary(model.SKT))[2,2])
```

```
[1] 11.01216
```

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{0.98333}{0.08929} = 11.01.$$

```
> p.b1 <- 2*pt(t.b1, df=model.SKT$df.residual, lower.tail=FALSE); p.b1
```

```
[1] 1.116594e-16
```

Ničelno domnevo zavrնemo v korist alternativne domneve; zveza je statistično značilna. SKT je statistično značilno odvisen od starosti ( $p < 0.0001$ ).

V tretjem delu povzetka modela je standardna napaka regresije  $s$ , koeficient determinacije  $R^2$  in povzetek tabele analize variance modela ( $F$ -statistika s pripadajočo  $p$ -vrednostjo).

```
> summary(model.SKT)$r.squared
[1] 0.6441239

> (F <- summary(model.SKT)$fstatistic)
      value      numdf      dendif
121.2678     1.0000     67.0000

> F <- as.numeric(F)
> (p.F <- pf(F[1], df1=F[2], df2=F[3], lower.tail=FALSE))
[1] 1.116594e-16
```

Za vajo pogledjmo izračun koeficienta deteminacije na naših podatkih:

```
> SS_model<-sum((model.SKT$fitted-mean(tlak$SKT))^2);SS_model
[1] 14951.25

> SS_res<-sum(model.SKT$residual^2);SS_res
[1] 8260.514

> R2<-SS_model/(SS_model+SS_res); R2
[1] 0.6441239
```

Približno dve tretjini variabilnosti SKT pojasni starost, ostala variabilnost ostane nepojasnjena. Le ugibamo lahko, kaj poleg starosti še vpliva na variabilnost SKT, morda spol, prehrana, genetika, fizična aktivnost, itd.

Vsote kvadriranih odklonov  $SS_{model}$  in  $SS_{residual}$  dobimo v tabeli analize variance s funkcijo `anova()`:

```
> anova(model.SKT)
```

Analysis of Variance Table

Response: SKT

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
starost	1	14951.3	14951.3	121.27	< 2.2e-16 ***
Residuals	67	8260.5	123.3		

---

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

Ocena za skupno varianco  $\sigma^2$  iz (6) je  $s^2 = 123.29$ , pripadajoča standardna napaka regresije je  $s = 11.1$ . V tem primeru  $F$ -statistika testira domnevo  $H_0 : \beta_1 = 0$ , ki je ekvivalentna domnevi, ki jo testira  $t$ -statistika; velja  $F = t^2 = 121.27$ .

Intervala zaupanja za parametra modela dobimo s funkcijo `confint`:

```
> confint(model.SKT)

                2.5 %      97.5 %
(Intercept) 94.7025550 111.995554
starost      0.8050947  1.161561
```

- Prvi interval zaupanja ni vsebinsko smiseln.
- Če se starost poveča za 10 let, pri 95 % zaupanju pričakujemo, da se bo SKT povečal na intervalu 8.1 mm do 11.6 mm.

Izračunajmo napovedi in njihove 95 % IZ za SKT za osebe stare 30 in 60 let.

```
> starost.napovedi<-data.frame(starost=c(30,60))
> ## povprečne napovedi
> povp.napovedi.SKT<-predict(model.SKT, starost.napovedi, interval="confidence")
> data.frame(cbind(starost.napovedi,povp.napovedi.SKT ))

  starost    fit    lwr    upr
1     30 132.8489 128.9247 136.7731
2     60 162.3487 158.7132 165.9842

> ## posamične napovedi
> pos.napovedi.SKT<-predict(model.SKT, starost.napovedi, interval="prediction")
> data.frame(cbind(starost.napovedi,pos.napovedi.SKT ))

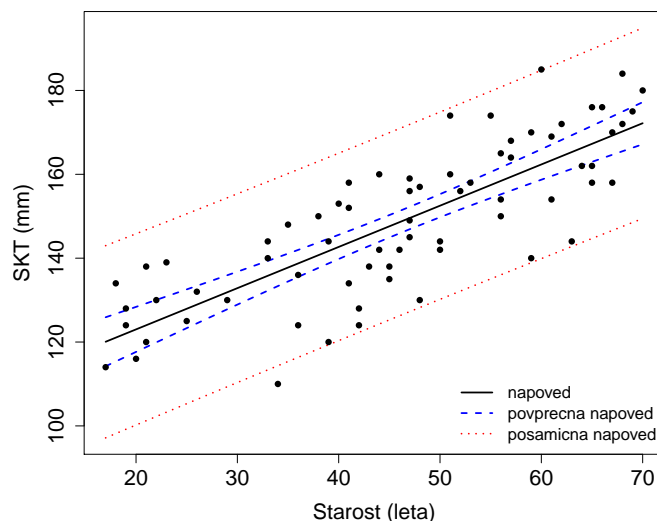
  starost    fit    lwr    upr
1     30 132.8489 110.3412 155.3566
2     60 162.3487 139.8895 184.8079
```

Interpretacija za osebe stare 30 let: napovedana vrednost za tlak 132.8 mm. Pripadajoč 95 % IZ za povprečje oseb s starostjo 30 let je (128.9 mm, 136.8 mm). Za posamezno osebo staro 30 let je pripadajoč 95 % IZ (110.3 mm, 155.4 mm).

Ali bi bila utemeljena napoved za osebo staro 95 let? Ali bi bila utemeljena napoved za osebo staro 10 let? Ne, ker gre za ekstrapolacijo.

Grafični prikaz 95 % intervalov zaupanja za povprečne in za posamične napovedi je na Sliki 6.





Slika 6: Napovedi za SKT, 95 % IZ za povprečne napovedi (notranji hiperboli) in za posamične napovedi (zunanji hiperboli)

## 2.6 Simulacija po predpostavkah modela enostavne linearne regresije

Simulacijo bomo izvedli, da bi ilustrirali statistične lastnosti ocen parametrov linearnega modela ter velikost testa in moč testa pri testiranju domnev o parametrih. Za okvir simulacije vzemimo model odvisnosti SKT od `starost` (`model.SKT`). V tem modelu testiranje domneve  $H_0 : \beta_0 = 0$  ni vsebinsko zanimivo, osredotočili se bomo na testiranje domneve  $H_0 : \beta_1 = 0$ .

V vseh simulacijah obdržimo iste vrednosti za napovedno spremenljivko `starost`. Za vrednosti odzivne spremenljivke SKT upoštevamo, da so pogojno na vrednosti napovedne spremenljivke porazdeljene normalno s pričakovano vrednostjo  $\beta_0 + \beta_1 \text{starost}$  in varianco  $\sigma^2$ , pri čemer za vrednosti  $\beta_0, \beta_1$  in  $\sigma^2$  vzamemo ocene iz `model.SKT`:  $SKT_i = 103 + 0.98 \text{starost}_i + \varepsilon_i$ ; napake  $\varepsilon_i$ ,  $i = 1, \dots, 69$ , generiramo s funkcijo `rnorm()` za porazdelitev  $N(0, \sigma^2 = 11^2)$ .

Za vsak generirani vzorec vrednosti `SKT` izračunamo oceni parametrov enostavnega linearnega modela  $b_0$  in  $b_1$  ter 95 % interval zaupanja za  $\beta_1$ . Izpišemo tudi  $p$ -vrednost pri testiranju domneve  $H_0 : \beta_1 = 0$ .

```
> # vrednosti za starost vzamemo iz podatkovnega okvira tlak
> starost<-tlak$starost
> # velikost vzorca
> n<-length(tlak$starost)
> # standardni odklon napak
> sigma<-11
> # izbrana parametra modela
> beta0<-103
> beta1<-0.98
> # teoretični izračun variance ocene b0
```

```
> var.b0<-sigma^2*(1/n+mean(starost)^2/sum((starost-mean(starost))^2))
> # teoretična standardna napaka za b0
> sqrt(var.b0)
```

```
[1] 4.291455
```

```
> # teoretični izračun variance ocene b1
> var.b1<-sigma^2/sum((starost-mean(starost))^2)
> # teoretična standardna napaka za b1
> sqrt(var.b1)
```

```
[1] 0.08846106
```

```
> set.seed(77) # za ponovljivost rezultatov
> # vrednosti slučajnih napak
> epsilon<-rnorm(n, mean=0, sd=sigma)
> # simulirane vrednosti odzivne spremenljivke
> SKT<-beta0 + beta1*starost + epsilon
```

```
> # model enostavne linearne regresije na simuliranih podatkih
> mod<-lm(SKT~starost)
> (b0<-coef(mod)[1])
```

```
(Intercept)
  105.9049
```

```
> (b1<-coef(mod)[2])
```

```
starost
0.9183129
```

```
> (p.b1<-coefficients(summary(mod))[2,4]) # za H0: beta1=0
```

```
[1] 1.054677e-14
```

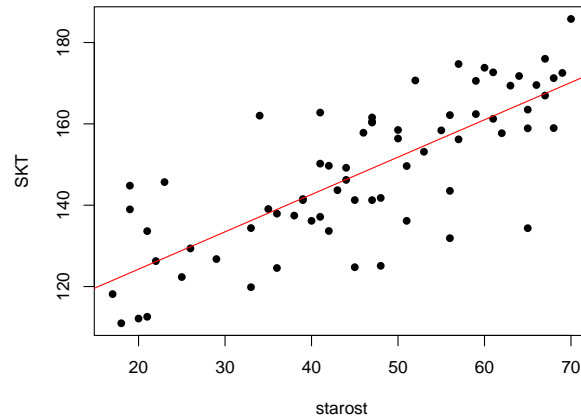
```
> (sp.meja.b1<-confint(mod)[2,1])
```

```
[1] 0.7327401
```

```
> (zg.meja.b1<-confint(mod)[2,2])
```

```
[1] 1.103886
```

```
> plot(starost, SKT, pch=16)
> abline(reg=mod, col="red")
```



Slika 7: Primer simuliranih podatkov in pripadajoča regresijska premica, `set.seed(77)`

Simulacijo bomo velikokrat ponovili. Na podlagi rezultatov simulacij želimo pogledati:

- kakšni sta porazdelitvi ocen parametrov enostavnega linearnega modela;
- kolikšen delež intervalov zaupanja za  $\beta_1$  ne vsebuje prave vrednosti parametra (velikost testa);
- kolikšna je moč testa pri testiranju ničelne domneve  $H_0 : \beta_1 = 0$ .

V ta namen bomo izvedli 1000 simulacij (`Nsim=1000`). Napišimo funkcijo, ki bo izvedla simulacije in za vsako simulacijo shranila v podatkovni okvir oceni parametrov, spodnjo in zgornjo mejo intervala zaupanja za  $\beta_1$  in  $p$ -vrednost pri testiranju ničelne domneve  $H_0 : \beta_1 = 0$ .

```
> Nsim <- 1000
> reg.sim <- function(x, beta0, beta1, sigma, n, Nsim) {
+   # pripravimo prazne vektorje za rezultate simulacij, oceni parametrov b0 in b1,
+   # p-vrednost za testiranje domneve beta1=0,
+   # spodnjo in zgornjo mejo intervala zaupanja za beta1
+   b0 <- numeric(Nsim)
+   b1 <- numeric(Nsim)
+   p.b1 <- numeric(Nsim)
+   sp.meja.b1 <- numeric(Nsim)
+   zg.meja.b1 <- numeric(Nsim)
+
+   for (i in 1:Nsim) {
+     epsilon<-rnorm(n, mean=0, sd=sigma)
+     y<-beta0+beta1*x+epsilon
+     mod<-lm(y~x)
```

```
+   b0[i]<-coef(mod)[1]
+   b1[i]<-coef(mod)[2]
+   p.b1[i]<-coefficients(summary(mod))[2,4]
+   sp.meja.b1[i]<-confint(mod)[2,1]
+   zg.meja.b1[i]<-confint(mod)[2,2]
+ }
+ return(data.frame(b0,b1,p.b1,sp.meja.b1,zg.meja.b1))
+ }
> rez.1000<-reg.sim(x=starost, beta0=103, beta1=0.98, sigma=11, n=30, Nsim=1000)
> head(rez.1000)

           b0           b1           p.b1 sp.meja.b1 zg.meja.b1
1 106.90334 0.9055551 1.168895e-11 0.6844261 1.126684
2 104.78995 0.9275064 7.847449e-17 0.7607452 1.094268
3 99.62207 1.0249480 6.718667e-18 0.8505561 1.199340
4 100.33065 0.9441515 3.556979e-19 0.7933889 1.094914
5 105.65246 0.9421287 1.368567e-22 0.8138692 1.070388
6 93.07662 1.1370228 2.988673e-23 0.9867236 1.287322

> # standardna napaka b0 in b1 na podlagi porazdelitve Nsim ocen parametrov
> (sd(rez.1000$b0))

[1] 4.336861

> (sd(rez.1000$b1))

[1] 0.08443493

> # 2.5 in 97.5 centil za b1 na podlagi simulacij
> (centili<-quantile(rez.1000$b1, probs=c(0.025, 0.975)))

      2.5%      97.5%
0.8207757 1.1485645

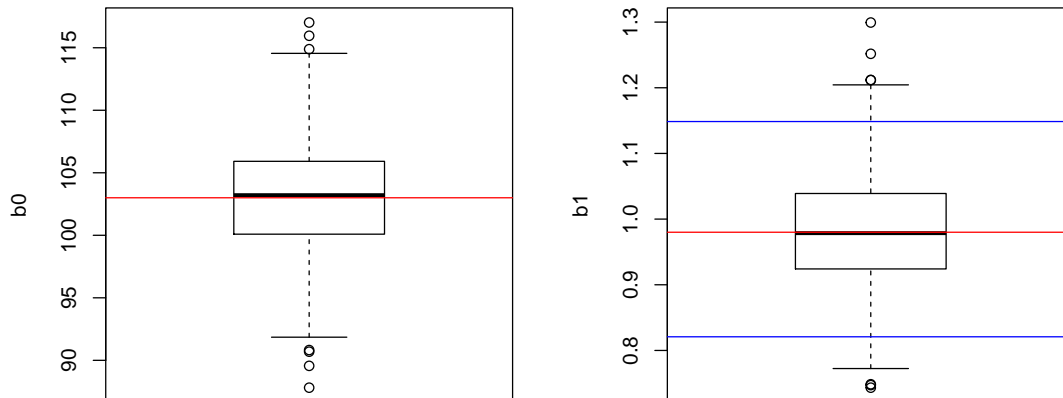
> # ocena verjetnosti za napako II. vrste za H0: beta1=0
> sum(rez.1000$p>0.05)/Nsim

[1] 0

> # ocena moči testa na podlagi Nsim simulacij
> (moc.testa<-1-sum(rez.1000$p>0.05)/Nsim)

[1] 1
```

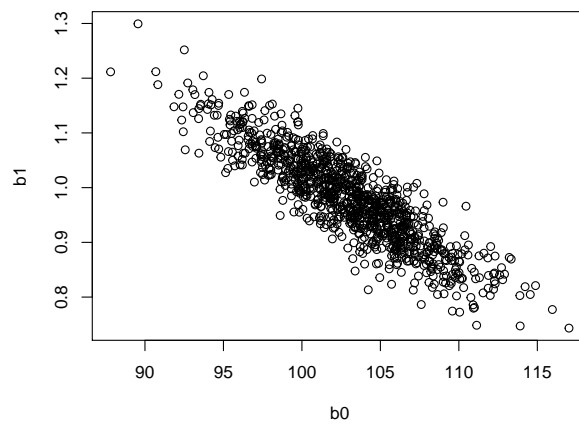
```
> par(mfrow=c(1,2))
> boxplot(rez.1000$b0, ylab="b0");
> abline(h=beta0, col="red")
> boxplot(rez.1000$b1, ylab="b1");
> abline(h=beta1, col="red");
> abline(h=centili, col="blue")
```



Slika 8: Porazdelitev ocen parametrov  $b_0$  (levo) in  $b_1$  (desno) za  $\sigma = 11$  in  $n = 69$ , `set.seed(77)`, rdeča črta kaže pravo vrednost za parameter, modri črti predstavljata 2.5 in 97.5 centil za  $b_1$

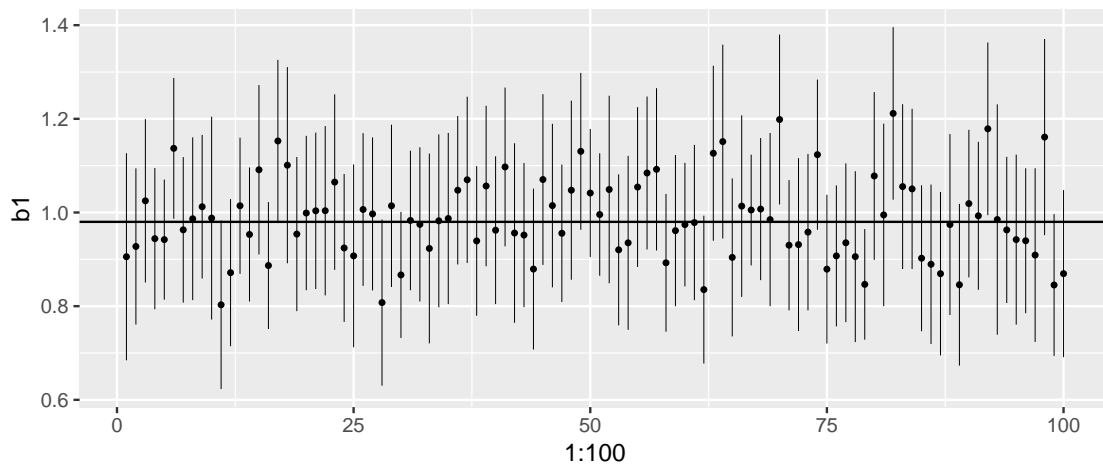
Slika 9 prikazuje negativno povezanost ocen parametrov  $b_0$  in  $b_1$ , ki je teoretično izražena s kovarianco  $Cov(b_0, b_1) = -\sigma^2 \bar{x} / S_{xx}$ .

```
> plot(rez.1000$b0, rez.1000$b1, pch=1, xlab="b0", ylab="b1")
```



Slika 9: Povezanost ocen parametrov  $b_0$  in  $b_1$ , `set.seed(77)`

```
> ggplot(rez.1000[1:100,], aes(x=1:100,y=b1,ymin=sp.meja.b1,ymax=zg.meja.b1)) +  
+ geom_pointrange(size=0.2, shape=16) +  
+ geom_hline(yintercept=beta1)
```



Slika 10: 100 intervalov zaupanja za  $\beta_1$  za  $\sigma = 11$  in  $n = 69$ , `set.seed(77)`

```
> # delež intervalov zaupanja, ki ne vsebujejo prave vrednosti parametra beta1,  
> # to je ocena velikosti testa  
> sum(rez.1000$sp.meja.b1>beta1 | rez.1000$zg.meja.b1<beta1)/Nsim
```

```
[1] 0.033
```

### 3 LINEARNI MODEL V MATRIČNI OBLIKI

Splošni linearni regresijski model lahko zapišemo:  $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ ,  $i = 1, \dots, n$ . Krajši in bolj eleganten je zapis v matrični obliki:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (27)$$

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

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

#### 3.1 Ocene parametrov

Ocene za  $\boldsymbol{\beta}$ , označimo jih  $\mathbf{b}$ , dobimo po metodi najmanjših kvadratov (*Ordinary Least Squares*, OLS). Iščejo minimum izraza:

$$S(\mathbf{b}) = \sum_{i=1}^n e_i^2 = \mathbf{e}^T \mathbf{e}. \quad (28)$$

Z odvajanjem in enačenjem parcialnih odvodov z 0 dobimo sistem normalnih enačb:

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{y}.$$

Rešitev obstaja, če je  $\mathbf{X}^T \mathbf{X}$  nesingularna. To velja,

- če je  $n \geq k + 1$ ; to pomeni, da je število enot vsaj tako veliko kot število ocenjevanih parametrov;
- če nobena spremenljivka ni linearna kombinacija ostalih spremenljivk.

Rešitev je:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (29)$$

Izračun ocen parametrov z matrikami za primer SKT naj služi kot ilustracija in vaja:

```
> X<-model.matrix(model.SKT) # modelska matrika
> X[1:5,] # prvih 5 vrstic

(Intercept) starost
1           1      41
2           1      60
3           1      41
4           1      47
5           1      66

> t(X)%*%X # kaj je v tej matriki: n, vsota x, vsota x*x
```

```

              (Intercept) starost
(Intercept)      69      3184
starost          3184 162388

> t(X)%*%tlak$SKT # kaj je v tej matriki: vsota y, vsota x*y

              [,1]
(Intercept) 10262
starost      488744

> b<-solve(t(X)%*%X) %*% t(X)%*% tlak$SKT; b # oceni parametrov

              [,1]
(Intercept) 103.3490547
starost      0.9833276

```

### 3.2 Pričakovana vrednost ocen parametrov in njihova variančno-kovariančna matrika

Ocene parametrov dobljene po metodi najmanjših kvadratov so nepristranske:

$$E(\mathbf{b}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta}. \quad (30)$$

Edina predpostavka uporabljena pri dokazu je, da je  $E(\boldsymbol{\varepsilon}) = 0$ . Iz tega sledi, da so ocene parametrov modela nepristranske tudi, če varianca  $\sigma^2$  ni konstantna ali če so napake korelirane. Variančno-kovariančna matrika za  $\mathbf{b}$  je vezana na varianco za  $\mathbf{y}$  in na modelsko matriko:

$$Var(\mathbf{b}) = ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) Var(\mathbf{y}) ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}. \quad (31)$$

Porazdelitev parametrov normalnega linearnega modela je večrazsežna normalna porazdelitev.

### 3.3 Matrika $\mathbf{H}$

Poglejmo povezavo med  $\hat{\mathbf{y}}$  in  $\mathbf{y}$ :

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{H}\mathbf{y}. \quad (32)$$

Matrika  $\mathbf{H}$ , t. i. "hat matrix", je ključna pri izračunu napovedi  $\hat{\mathbf{y}}$ :

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T. \quad (33)$$

Pokažemo lahko, da velja:  $\mathbf{H} = \mathbf{H}^T = \mathbf{H}^2$ .

Izračun matrike  $\mathbf{H}$  za primer SKT:

```

> H<-X %*% solve(t(X)%*%X) %*% t(X)
> dim(H)

[1] 69 69

> round(H[1:10,1:10],3) # izpis prvih 10 stolpcev in 10 vrstic matrike H

```



	1	2	3	4	5	6	7	8	9	10
1	0.016	0.010	0.016	0.014	0.008	0.014	0.007	0.016	0.007	0.011
2	0.010	0.027	0.010	0.015	0.032	0.015	0.034	0.012	0.034	0.024
3	0.016	0.010	0.016	0.014	0.008	0.014	0.007	0.016	0.007	0.011
4	0.014	0.015	0.014	0.015	0.016	0.015	0.016	0.014	0.016	0.015
5	0.008	0.032	0.008	0.016	0.040	0.016	0.043	0.010	0.043	0.028
6	0.014	0.015	0.014	0.015	0.016	0.015	0.016	0.014	0.016	0.015
7	0.007	0.034	0.007	0.016	0.043	0.016	0.045	0.010	0.045	0.030
8	0.016	0.012	0.016	0.014	0.010	0.014	0.010	0.015	0.010	0.012
9	0.007	0.034	0.007	0.016	0.043	0.016	0.045	0.010	0.045	0.030
10	0.011	0.024	0.011	0.015	0.028	0.015	0.030	0.012	0.030	0.022

Diagonalne člene matrike  $\mathbf{H}$  imenujemo vzvodi (*leverages* ali *hatvalues*) in jih dobimo z ukazom `hatvalues`:

```
> round(hatvalues(model.SKT),3)
```

	1	2	3	4	5	6	7	8	9	10	11	12	13
0.016	0.027	0.016	0.015	0.040	0.015	0.045	0.015	0.045	0.022	0.037	0.022	0.029	
14	15	16	17	18	19	20	21	22	23	24	25	26	
0.021	0.015	0.015	0.015	0.062	0.052	0.055	0.019	0.017	0.016	0.066	0.016	0.020	
27	28	29	30	31	32	33	34	35	36	37	38	39	
0.037	0.026	0.049	0.051	0.021	0.031	0.016	0.015	0.025	0.017	0.023	0.026	0.041	
40	41	42	43	44	45	46	47	48	49	50	51	52	
0.029	0.018	0.015	0.015	0.037	0.014	0.043	0.016	0.043	0.021	0.035	0.021	0.025	
53	54	55	56	57	58	59	60	61	62	63	64	65	
0.024	0.016	0.015	0.015	0.069	0.059	0.062	0.021	0.015	0.018	0.055	0.015	0.018	
66	67	68	69										
0.033	0.034	0.043	0.048										

### 3.4 Ostanke

V matrični obliki izračunamo ostanke takole:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}. \quad (34)$$

Varianca ostankov:

$$\text{Var}(\mathbf{e}) = (\mathbf{I} - \mathbf{H}) (\sigma^2 \mathbf{I}) (\mathbf{I} - \mathbf{H})^T = \sigma^2 (\mathbf{I} - \mathbf{H}). \quad (35)$$

Kovarianca med  $\mathbf{e}$  in  $\hat{\mathbf{y}}$  je nič:

$$\text{Cov}(\mathbf{e}, \hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H}) (\sigma^2 \mathbf{I}) \mathbf{H}^T = \sigma^2 (\mathbf{H}^T - \mathbf{H}\mathbf{H}^T) = 0. \quad (36)$$

## 4 VAJE

### 4.1 Čas teka Collina Jacksona

V datoteki COLLIN.txt so podatki za 21 tekov na 110 m čez ovire tekača Collina Jacksona: hitrost vetra = `windspeed` (m/s) in čas teka = `time` (s) (Vir: Daly et al., str. 525). Podatki so bili dobljeni v poskusu v zaprtem prostoru, hitrost vetra je bila izbrana za vsak tek posebej vnaprej. Negativne vrednosti hitrosti vetra pomenijo, da je veter pihal v prsi tekača. Kako hitrost vetra vpliva na čas teka na 110 m čez ovire?

- Grafično prikažite podatke.
- Ocenite parametra linearnega regresijskega modela za odvisnost časa teka od hitrosti vetra.
- Analizirajte ostanke modela na podlagi grafičnih prikazov.
- Obrazložite oceni parametra modela in njuna intervala zaupanja.
- Obrazložite koeficient determinacije.
- Izračunajte povprečno in posamično napoved časa teka ter pripadajoče 95 % intervale zaupanja za naslednje hitrosti vetra: -1 m/s, 0 m/s, 1 m/s in 4 m/s. Ali so vse napovedi upravičene? Zakaj?

### 4.2 Simulacije za enostavno linearno regresijo

Za izbrani vrednosti parametrov enostavne linearne regresije  $\beta_0 = 100$  in  $\beta_1 = 1$  izvedite simulacije, ki bodo ilustrirale vpliv velikosti vzorca  $n$  in vrednosti variance napak  $\sigma^2$  na porazdelitev ocen parametrov in na moč testa.

Za vsako izbrano velikost vzorca  $n$  najprej generirajte vrednosti napovedne spremenljivke  $x$  na intervalu 15 do 70. Pri tem uporabite funkcijo `sample` z argumentom `replace=TRUE`: `x<-sample(c(17:70), size=n, replace=TRUE)`. Za tako določene vrednosti napovedne spremenljivke nadaljujte z simulacijami pri izbranih vrednostih  $\sigma^2$ .

Navodilo:

- Izberite naslednje vrednosti za  $n$ : 7, 15, 30, 100 in naslednje vrednosti za  $\sigma$ : 5, 11, 22. Za vsako kombinacijo  $n$  in  $\sigma$  naredite 1000 simulacij.
- Grafično prikažite:
  - odvisnost širine intervala zaupanja za  $\beta_1$  od  $n$ , za vsako vrednost  $\sigma$ ;
  - odvisnost širine intervala zaupanja za  $\beta_1$  od  $\sigma$ , za vsak  $n$ ;
  - odvisnost moči testa od  $n$ , za vsako vrednost  $\sigma$ ;
  - odvisnost moči testa od  $\sigma$ , za vsak  $n$ .
- Napišite kratek povzetek vaših ugotovitev.