Vaja 1: Začetna analiza podatkov, enostavni linearni regresijski model, simulacije

1. Primer: Odstotek telesne maščobe

`R` paketi, ki jih bomo uporabili na vajah:

```
library(vtable) # summary table (sumtable)
library(reshape2) # reshape data sets for ggplot (melt)
library(ggplot2) # nice plots (ggplot)
library(corrplot) # correlation plots (corrplot)
library(knitr)  # for markdown
library(kableExtra) # creates nice latex tables (kable, kable_styling)
library(car) # regression diagnostics
```

Začetna analiza podatkov je pomemben del vsake obdelave podatkov (vir: https://muse.jhu.edu/pub/56/article/793379/pdf). Sestavljajo jo naslednji koraki:

1. določitev metapodatkov (podatki o podatkih);
2. čiščenje podatkov (odpravljanje napak v podatkih);
3. pregled podatkov (razumevanje lastnosti podatkov);
4. poročanje začetne analize podatkov vsem sodelavcem, vpletenih v analizo;
5. izpopolnjevanje in posodabljanje načrta analize, ki vključuje ugotovitve na podlagi začetne analize podatkov;
6. poročanje začetne analize podatkov (vsebovati mora vse korake, ki vplivajo na interpretacijo rezultatov).

Cilj naše analize bo razviti model za napovedovanje odstotka telesne maščobe na podlagi spremenljivk, katerih vrednosti lahko dobimo le z uporabo tehtnice in merilnega traku (vir: Roger W. Johnson (1996), "Fitting Percentage of Body Fat to Simple Body Measurements", Journal of Statistics Education, http://jse.amstat.org/v4n1/datasets.johnson.html). V podatkih sta dve oceni odstotka telesne maščobe, dobljeni po Brozokovi in Sirijevi enačbi. V vaji bo slednja spremenljivka naša odzivna spremenljivka.

```
bodyfat <- read.table(url("https://jse.amstat.org/datasets/fat.dat.txt"))
head(bodyfat)
```

```
  V1   V2   V3     V4 V5     V6    V7   V8    V9  V10   V11   V12   V13  V14
1  1 12.6 12.3 1.0708 23 154.25 67.75 23.7 134.9 36.2  93.1  85.2  94.5 59.0
2  2  6.9  6.1 1.0853 22 173.25 72.25 23.4 161.3 38.5  93.6  83.0  98.7 58.7
3  3 24.6 25.3 1.0414 22 154.00 66.25 24.7 116.0 34.0  95.8  87.9  99.2 59.6
4  4 10.9 10.4 1.0751 26 184.75 72.25 24.9 164.7 37.4 101.8  86.4 101.2 60.1
5  5 27.8 28.7 1.0340 24 184.25 71.25 25.6 133.1 34.4  97.3 100.0 101.9 63.2
6  6 20.6 20.9 1.0502 24 210.25 74.75 26.5 167.0 39.0 104.5  94.4 107.8 66.0
   V15  V16  V17  V18  V19
1 37.3 21.9 32.0 27.4 17.1
2 37.3 23.4 30.5 28.9 18.2
3 38.9 24.0 28.8 25.2 16.6
4 37.3 22.8 32.4 29.4 18.2
5 42.2 24.0 32.2 27.7 17.7
6 42.0 25.6 35.7 30.6 18.8
```

Metapodatki

Ta niz podatkov za 252 moških vsebuje informacije o:

Tabela 1: Metapodatki podatkovnega okvira `bodyfat`.

Ime	Pomen	Enote	Merska lestvica	Class	NAs
case	ID			integer	0
brozek	odstotek telesne maščobe (Brozek)	%	številska	numeric	0
siri	odstotek telesne maščobe (Siri)	%	številska	numeric	0
density	Gostota, določena s tehtanjem pod vodo	gm/cm^3	številska	numeric	0
age	Starost	years	številska	numeric	0
weight	Masa	lbs	številska	numeric	0
height	Višina	inches	številska	numeric	0
BMI	Indeks telesne mase	kg/m^2	številska	numeric	0
fatfreeweight	Masa brez maščobe [weight/(1-brozek/100)]	lbs	številska	numeric	0
neck	Obseg vratu	cm	številska	numeric	0
chest	Obseg prsnega koša	cm	številska	numeric	0
abdomen	Obseg abdomna	cm	številska	numeric	0
hip	Obseg bokov	cm	številska	numeric	0
thigh	Obseg stegna	cm	številska	numeric	0
knee	Obseg kolena	cm	številska	numeric	0
ankle	Obseg gležnja	cm	številska	numeric	0
biceps	Obseg bicepsa	cm	številska	numeric	0
forearm	Obseg podlakti	cm	številska	numeric	0
wrist	Obseg zapestja	cm	številska	numeric	0

Tabela 1 prikazuje metapodatke podatkovnega okvira `bodyfat`. Metapodatki so podatki o podatkih oz. informacije o spremenljivkah, na primer: oznaka in pomen posamezne spremenljivke, merska lestvica, enote, intervali, na katerih se vrednosti posameznih spremenljivk lahko nahajajo, informacije o manjkajočih podatkih. Metapodatki pa lahko vključujejo tudi informacije o tem, kako so bili podatki pridobljeni (npr. viri podatkov, metode zbiranja podatkov, kako je definirana ciljna populacija, merila za vključitev in izključitev posameznih enot, metode vzorčenja, čas zbiranja podatkov, ipd.).

Iz metopodatkov lahko vidimo, da sta spremenljivki `BMI` in `fatfreeweight` izpeljani na podlagi drugih spremenljivk v podatkovnem okviru. Ti dve spremenljivki bomo izločili iz podatkovnega okvira, saj nas bo zanimalo napovedovanje odstotka telesne maščobe na podlagi 'osnovnih' spremenljivk.

```
bodyfat <- bodyfat[, -c(8:9)]
```

Spremenljivkam dopišemo imena

```
colnames(bodyfat) <- c("case", "brozek", "siri", "density", "age", "weight",
                       "height", "neck","chest", "abdomen", "hip", "thigh",
                       "knee", "ankle",  "biceps", "forearm", "wrist")
```

in spremenimo enote za spremenljivki masa in višina:

```
## iz lb v kg
bodyfat$weight <- bodyfat$weight * 0.454
## iz in v cm
bodyfat$height <- bodyfat$height * 2.54
```

Čiščenje podatkov

Čiščenje podatkov je sistematičen poskus iskanja napak v podatkih in, če je le mogoče, njihovega odpravljanja. Pogosti primeri so: napačno kodiranje opisnih spremenljivk, nemogoče vrednosti, datumi, izven časovnega okvira študije, manjkajoče vrednosti, osamelci, nemogoče kombinacije vrednosti dveh spremenljivk, podvojene vrstice, ipd.

V naslednjem koraku podatke očistimo vrednosti, ki niso verjetne:

```
summary(bodyfat)
```

```
      case            brozek           siri          density      
 Min.   :  1.00   Min.   : 0.00   Min.   : 0.00   Min.   :0.995  
 1st Qu.: 63.75   1st Qu.:12.80   1st Qu.:12.47   1st Qu.:1.041  
 Median :126.50   Median :19.00   Median :19.20   Median :1.055  
 Mean   :126.50   Mean   :18.94   Mean   :19.15   Mean   :1.056  
 3rd Qu.:189.25   3rd Qu.:24.60   3rd Qu.:25.30   3rd Qu.:1.070  
 Max.   :252.00   Max.   :45.10   Max.   :47.50   Max.   :1.109  
      age            weight          height           neck      
 Min.   :22.00   Min.   : 53.80   Min.   : 74.93   Min.   :31.10  
 1st Qu.:35.75   1st Qu.: 72.19   1st Qu.:173.35   1st Qu.:36.40  
 Median :43.00   Median : 80.13   Median :177.80   Median :38.00  
 Mean   :44.88   Mean   : 81.23   Mean   :178.18   Mean   :37.99  
 3rd Qu.:54.00   3rd Qu.: 89.44   3rd Qu.:183.51   3rd Qu.:39.42  
 Max.   :81.00   Max.   :164.87   Max.   :197.49   Max.   :51.20  
     chest           abdomen            hip            thigh      
 Min.   : 79.30   Min.   : 69.40   Min.   : 85.0   Min.   :47.20  
 1st Qu.: 94.35   1st Qu.: 84.58   1st Qu.: 95.5   1st Qu.:56.00  
 Median : 99.65   Median : 90.95   Median : 99.3   Median :59.00  
 Mean   :100.82   Mean   : 92.56   Mean   : 99.9   Mean   :59.41  
 3rd Qu.:105.38   3rd Qu.: 99.33   3rd Qu.:103.5   3rd Qu.:62.35  
 Max.   :136.20   Max.   :148.10   Max.   :147.7   Max.   :87.30  
      knee            ankle           biceps         forearm          wrist      
 Min.   :33.00   Min.   :19.1   Min.   :24.80   Min.   :21.00   Min.   :15.80  
 1st Qu.:36.98   1st Qu.:22.0   1st Qu.:30.20   1st Qu.:27.30   1st Qu.:17.60  
 Median :38.50   Median :22.8   Median :32.05   Median :28.70   Median :18.30  
 Mean   :38.59   Mean   :23.1   Mean   :32.27   Mean   :28.66   Mean   :18.23  
 3rd Qu.:39.92   3rd Qu.:24.0   3rd Qu.:34.33   3rd Qu.:30.00   3rd Qu.:18.80  
 Max.   :49.10   Max.   :33.9   Max.   :45.00   Max.   :34.90   Max.   :21.40  
```

```
#najmanjša oseba je visoka 75 cm
which.min(bodyfat$height)
```

```
[1] 42
```

```
bodyfat$weight[which.min(bodyfat$height)] #in tehta 93 kg
```

```
[1] 93.07
```

```
bodyfat$height[bodyfat$case==42] <- 176.53  #glej https://jse.amstat.org/datasets/fat.txt
```

Pregledovanje podatkov

Pregledovanje podatkov nam omogoča razumeti lastnosti podatkov, ki bi lahko vplivale na nadaljno analizo in interpretacijo rezultatov. Vključuje korake, ki preverjajo, ali podatki izpolnjujejo določene lastnosti oz. predpostavke, ki so potrebne, da določena analiza da zadovoljive rezultate, vendar izključujoč kakršno koli testiranje hipotez. Pregledovanje podatkov vključuje preučevanje univaritnih in multivaritnih porazdelitev spremenljivk, izračun opisnih statistik ter identifikacija manjkajočih vrednosti pri posameznih enotah in spremenljivkah.

V kontekstu linearnih modelov lahko z univariatnimi porazdelitvami spremenljivk odkrijemo prisotnost osamelcev, ki imajo lahko močan vpliv na rezultate analize, ter spremenljivke, ki so asimetrično porazdeljene. Čeprav linerani model ne predpostavlja ničesar o porazdelitvi odzivne spremenljivke (neodvisno od regresorjev) in porazdelitvi regresorjev, je verjetno, da bodo spremenljivke, ki so močno asimetrične, vplivale na porazdelitev

Tabela 2: Opisne statistike za spremenljivke v podatkovnem okviru `bodyfat`.

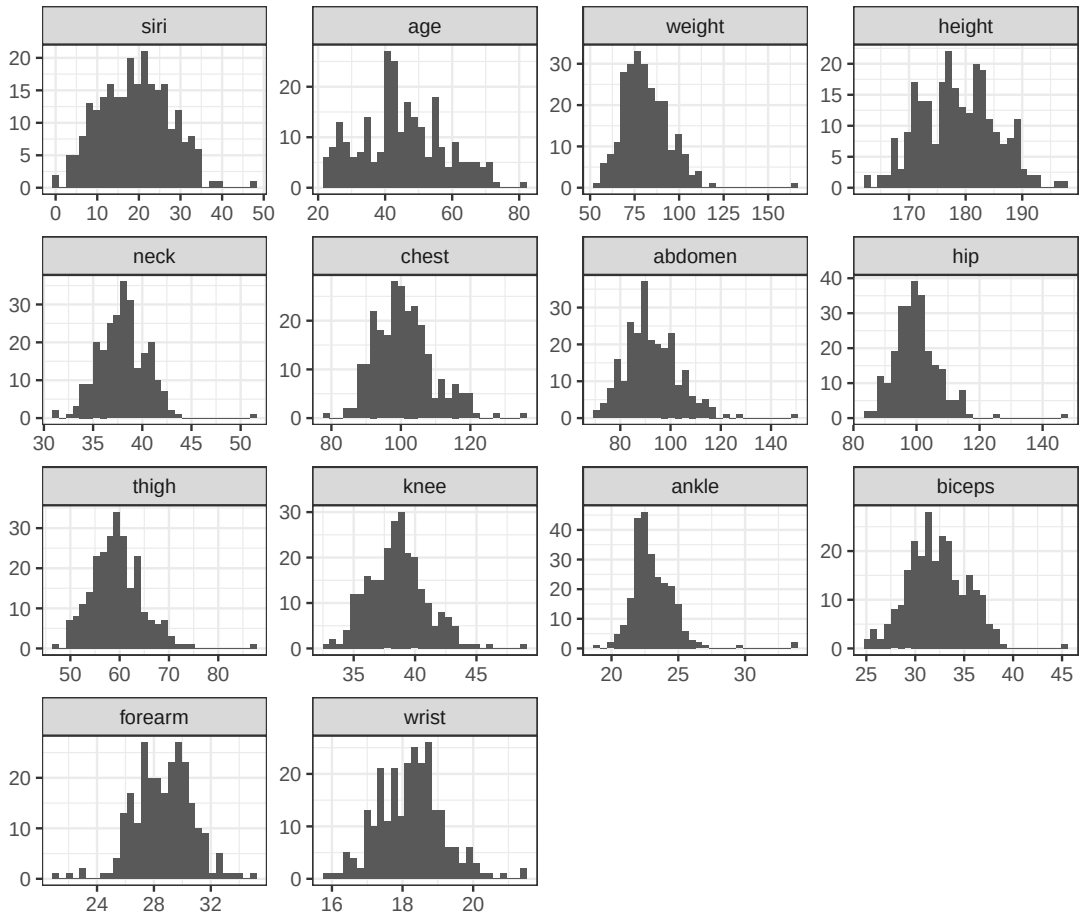
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 50	Pctl. 75	Max
siri	252	19	8	0	12	19	25	48
age	252	45	13	22	36	43	54	81
weight	252	81	13	54	72	80	89	165
height	252	179	7	163	173	178	184	197
neck	252	38	2	31	36	38	39	51
chest	252	101	8	79	94	100	105	136
abdomen	252	93	11	69	85	91	99	148
hip	252	100	7	85	96	99	104	148
thigh	252	59	5	47	56	59	62	87
knee	252	39	2	33	37	38	40	49
ankle	252	23	2	19	22	23	24	34
biceps	252	32	3	25	30	32	34	45
forearm	252	29	2	21	27	29	30	35
wrist	252	18	0.9	16	18	18	19	21

ostankov, za katere linearni model predpostavlja, da so normalno porazdeljeni. V splošnem velja, če je spremenljivka asimetrično porazdeljena, je potrebno analizo temu prilagoditi in raje uporabiti metodo, ki ne temelji na predpostavki o normalni porazdelitvi, lahko pa spremenljivko tudi lineariziramo. Pri asimetrično porazdeljenih spremenljivkah se pogosto zgodi tudi to, da je povezanost z drugimi spremenljivkami lahko le posledica majhnega števila enot, zato je treba to povezanost obravnavati previdno. Če so pri spremenljivki prisotne manjkajoče vrednosti, zaključkov na podlagi analize, omejene na enote, pri katerih manjkajočih vrednosti ni, morda ne bomo mogli posplošiti na celotno populacijo. Preučevanje bivariatnih povezanosti (preko korelacij oz. grafično) nam lahko pomaga pri odkrivanju nelinearnosti, interakcij, osamelcev in pri identificiranju multikolinearnosti.

Poglejmo si univariatne porazdelitve spremenljivk in korelacije med pari napovednih spremenljivk.

```
# napovedne spremenljivke
pred <- c("age", "weight", "height", "neck", "chest", "abdomen", "hip",
"thigh", "knee", "ankle", "biceps", "forearm", "wrist")

sumtable(bodyfat[,c("siri", pred)],
         add.median = T,
         digits = 1,
         title = "Opisne statistike za spremenljivke v podatkovnem okviru
         \\texttt{bodyfat}.",
         out = "kable"
         )
```

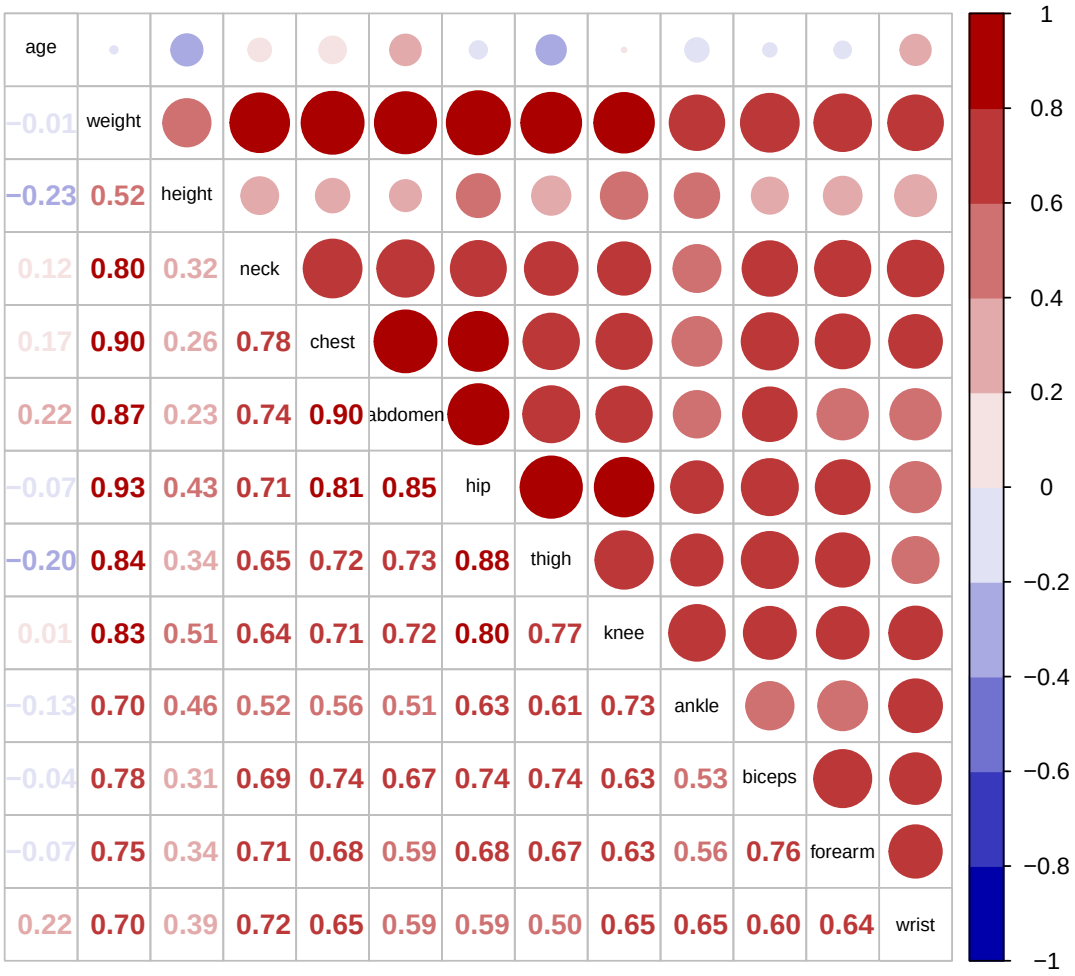
```
bodyfat_long <- melt(bodyfat[, c("case", "siri", pred)], id.vars = "case")
qplot(value, data = bodyfat_long) +
  facet_wrap(~variable, scales = "free") +
  theme_bw() +
  xlab("")
```



Slika 1: Univariatne porazdelitve spremenljivk v podatkovnem okviru `bodyfat`.

Porazdelitev odzivne spremenljivke `siri` je dokaj blizu normalni porazdelitvi, pri nekaterih napovednih spremenljivkah imamo posamezne osamelce, sicer pa so vrednosti precej normalno porazdeljene po zalogi vrednosti.

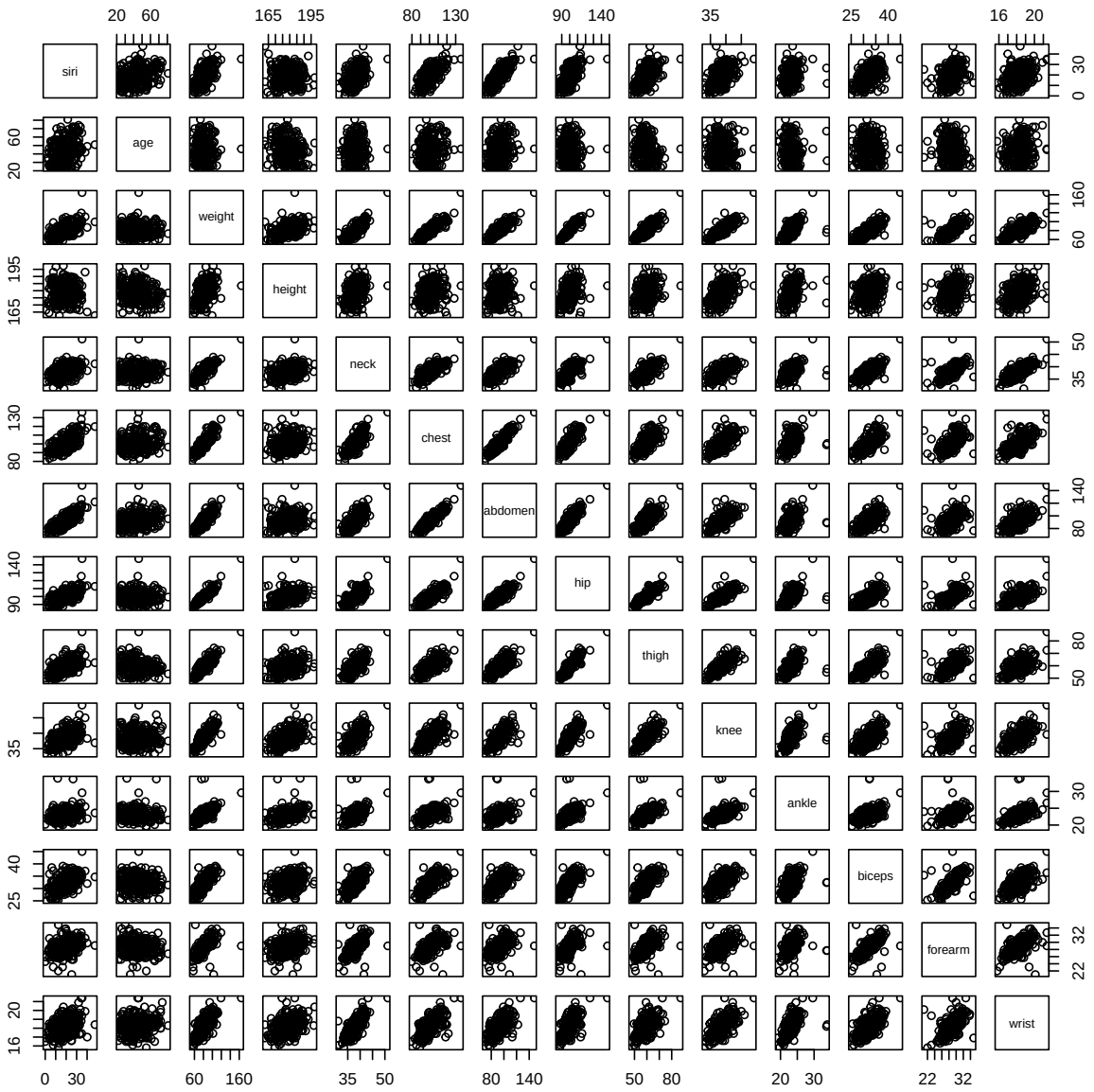
```
colors10 <- colorRampPalette(c("#0000aa","white","#aa0000"))(10)
corrplot.mixed(cor(bodyfat[, pred], method="spearman"),
               lower.col = colors10, upper.col = colors10,
               tl.col="black",   tl.cex = 0.7)
```



Slika 2: Spearmanovi koeficienti korelacije med napovednimi spremenljivkami v podatkovnem okviru `bodyfat`.

Slika 2 kaže, da so napovedne spremenljivke dokaj tesno pozitivno povezane med seboj, edina napovedna spremenljivka, ki ne kaže povezanosti z ostalimi je `age`. Močne korelacije kažejo na to, da bi v modelu znali imeti težave s kolinearnostjo.

```
pairs(bodyfat[, c("siri", pred)])
```



Slika 3: Matrika razsevnih grafikonov v podatkovnem okviru `bodyfat`.

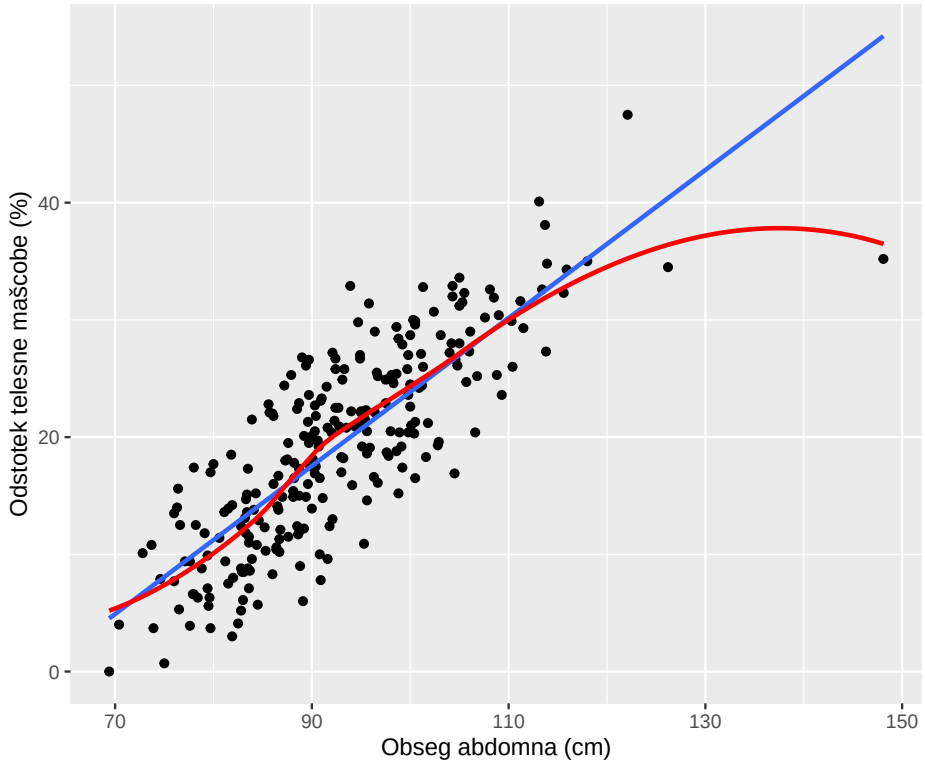
Matrika razsevnih grafikonov na Sliki 3 nam grafično prikaže vse bivariatne zveze med posameznimi pari spremenljivk v podatkovnem okviru `bodyfat`. Vidimo, da so zveze lepo linearne, povezanost med nekaterimi pari spremenljivk pa je zelo močna, kar bi posledično lahko pomenilo težave s kolinearnostjo. Ponekod so opazni osamelci.

Analiza

V današnji vaji se bomo osredotočili na zvezo med spremenljivkama `siri` in `abdomen`. Zanima nas, ali bi `siri` lahko napovedali na podlagi spremenljivke `abdomen`. Če med spremenljivkama obstaja linearna povezanost, potem tako zvezo lahko modeliramo z linearnim modelom:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \quad i = 1, ..., n.$$

```
ggplot(data=bodyfat, aes(x=abdomen, y=siri))+geom_point()+
  geom_smooth(method="lm", se=FALSE) +
  geom_smooth(col="red", se=FALSE)+
  xlab("Obseg abdomna (cm)") +
  ylab("Odstotek telesne maščobe (%)")
```



Slika 4: Razsevni grafikon za `siri` in `abdomen` z dodano premico in gladilnikom.

Cilj je najti oceni za parametra β_0 in β_1, tako da se bo regresijska premica dobro prilegala podatkom:

$$\hat{y}_i = b_0 + b_1 \cdot x_i, \quad i = 1, ..., n.$$

```
m1 <- lm(siri ~ abdomen, data = bodyfat)
summary(m1)
```

```
Call:
lm(formula = siri ~ abdomen, data = bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-19.0160  -3.7557   0.0554   3.4215  12.9007

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -39.28018    2.66034  -14.77   <2e-16 ***
abdomen       0.63130    0.02855   22.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.877 on 250 degrees of freedom
Multiple R-squared:  0.6617,    Adjusted R-squared:  0.6603
F-statistic: 488.9 on 1 and 250 DF,  p-value: < 2.2e-16
```

V povzetku modela se testirata dve ničelni hipotezi:

$H_0 : \beta_0 = 0$ proti $H_1 : \beta_0 \neq 0$

in

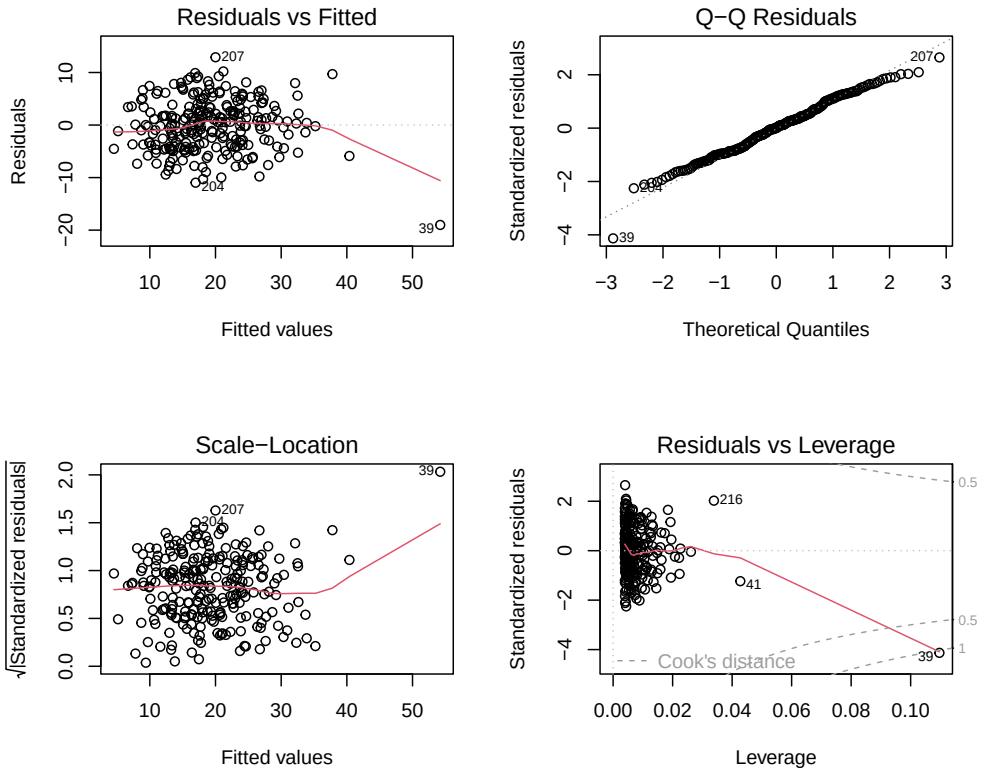
$H_0 : \beta_1 = 0$ proti $H_1 : \beta_1 \neq 0$

Statistično sklepanje in napovedi so veljavne, kadar so predpostavke normalnega linearnega modela izpolnjene, to je kadar:

1. obstaja linearna odvisnost odzivne spremenljivke od napovednih sprmenljivk,
2. imajo napake ε_i za vsako enoto skupno varianco σ^2 (homoskedastičnost),
3. je pričakovana vrednost napak 0,
4. so napake porazdeljene po normalni porazdelitvi,
5. so napake medsebojno neodvisne.

Osnovno diagnostiko modela naredimo na podlagi slik ostankov modela:

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



Slika 5: Ostanki za model 1 `siri` ∼ `abdomen`.

Na podlagi slik ostankov lahko preverimo predpostavke 1., 2., 3. in 4. Razložite, kako! Kaj pomeni predpostavka 5.?

Zdi se, da se linearni model dobro prilega podatkom, razen za enoto 39.

```
bodyfat[39, ]
```

```
   case brozek siri density age   weight  height neck chest abdomen   hip thigh
39   39   33.8 35.2  1.0202  46 164.8701 183.515 51.2 136.2   148.1 147.7  87.3
   knee ankle biceps forearm wrist
39 49.1  29.6     45      29  21.4
```

Izkaže se, da večino osamelcev, ki smo jih identificirali na podlagi Slike 1, lahko pripišemo enoti 39.

To enoto bi imelo v nadaljevanju smisel izključiti iz analize in primerjati rezultate obeh modelov. Kasneje bomo v okviru posebnih točk v regresijski analizi videli, da je ta točka t. i. vplivna točka.

```
bodyfat_brez39 <- bodyfat[-which(bodyfat$case==39),]
m2 <- lm(siri ~ abdomen, data = bodyfat_brez39)
summary(m2)
```

```

Call:
lm(formula = siri ~ abdomen, data = bodyfat_brez39)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-10.9133  -3.6469   0.1914   3.1737  12.7613

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.95774    2.71323  -15.83   <2e-16 ***
abdomen       0.67195    0.02921   23.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.717 on 249 degrees of freedom
Multiple R-squared:  0.6801,    Adjusted R-squared:  0.6788
F-statistic: 529.3 on 1 and 249 DF,  p-value: < 2.2e-16
```

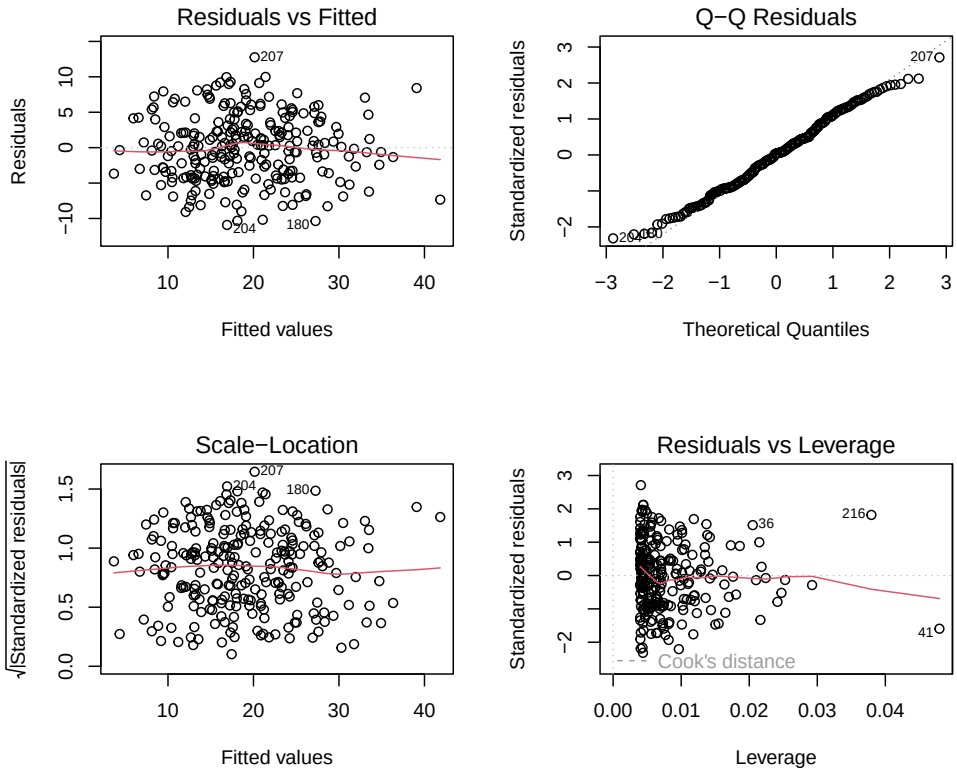
Primerjava ocen obeh modelov:

```
compareCoefs(m1, m2)
```

```
Calls:
1: lm(formula = siri ~ abdomen, data = bodyfat)
2: lm(formula = siri ~ abdomen, data = bodyfat_brez39)

            Model 1 Model 2
(Intercept)  -39.28  -42.96
SE             2.66    2.71

abdomen      0.6313  0.6720
SE           0.0286  0.0292
```

Poglejmo ostanke modela `m2`.

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



Slika 6: Ostanki za model 2 `siri` ∼ `abdomen`.

Oceni b_0 = -42.96 in b_1 = 0.67, dobljeni po metodi najmanjših kvadratov, sta nepristranski. Pripadajoči standardni napaki izražata mero natančnosti ocen, na podlagi katerih lahko izračunamo intervale zaupanja.

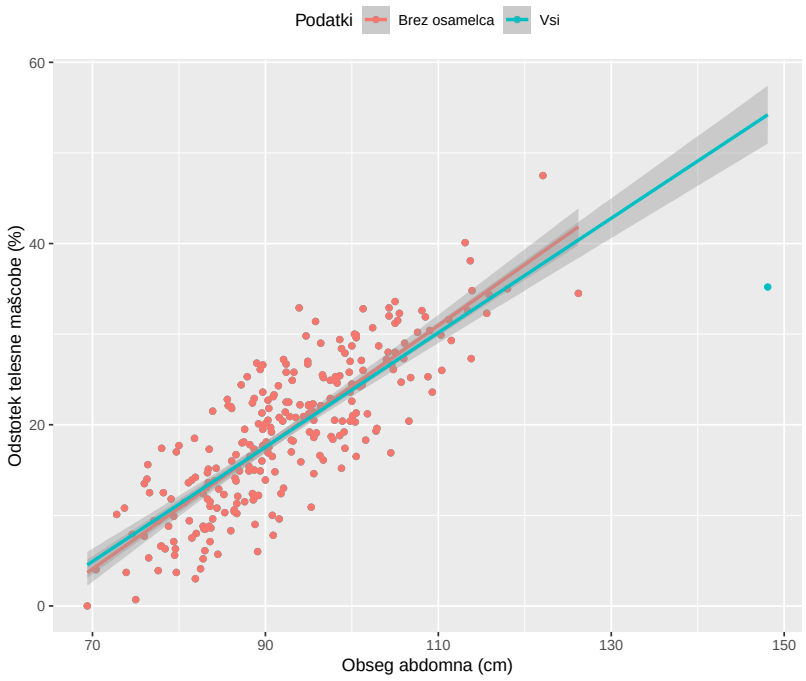
95% interval zaupanja za obe oceni:

```
confint(m2)
```

```
                2.5 %      97.5 %
(Intercept) -48.3015390 -37.6139418
abdomen       0.6144288   0.7294781
```

Interpretacija modela: če se obseg abdomna poveča za 1 cm, se telesna maščoba v povprečju poveča za 0.67 %, 95 % IZ: (0.61, 0.73). Z modelom smo pojasnili 68 % variabilnosti odzivne spremenljivke.

```
bodyfat$Podatki <- "Vsi"
bodyfat_brez39$Podatki <- "Brez osamelca"

bodyfat_komb <- rbind(bodyfat, bodyfat_brez39)

ggplot(aes(x=abdomen, y=siri, color=Podatki), data=bodyfat_komb) +
  geom_point() +
  geom_smooth(method = "lm", se=TRUE) +
  xlab("Obseg abdomna (cm)") +
  ylab("Odstotek telesne maščobe (%)") +
  theme(legend.position = "top")
```



Slika 7: Odvisnost `siri` od `abdomen` za dana vzorca 252 oz. 251 moških in regresijski premici na podalgi modelov `m1` in `m2` s 95 % intervali zaupanja za povprečno napoved.

2. Primer: Čas teka Collina Jacksona

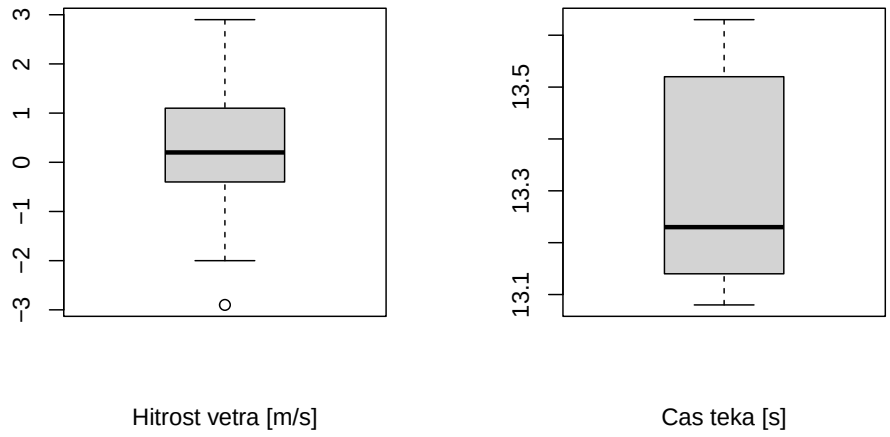
V datoteki *COLLIN.txt* so podatki za 21 tekov čez ovire na 110 m tekača Collina Jacksona: hitrost vetra = `windspeed` (m/s) in čas teka= `time` (s) (Vir: Daly et al., str. 525). Podatki so bili dobljeni v poskusu v zaprtem prostoru, hitrost vetra je bila izbrana za vsak tek posebej vnaprej. Negativne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

1. Grafično prikažite podatke.
2. Ocenite parametra linearnega regresijskega modela za odvisnost časa teka od hitrosti vetra.
3. Analizirajte ostanke modela na podlagi grafičnih prikazov.
4. Obrazložite cenilki parametrov modela in njuna intervala zaupanja.
5. Obrazložite koeficient determinacije.
6. 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?

```
d <- read.table("COLLIN.txt", header = T)
summary(d)
```

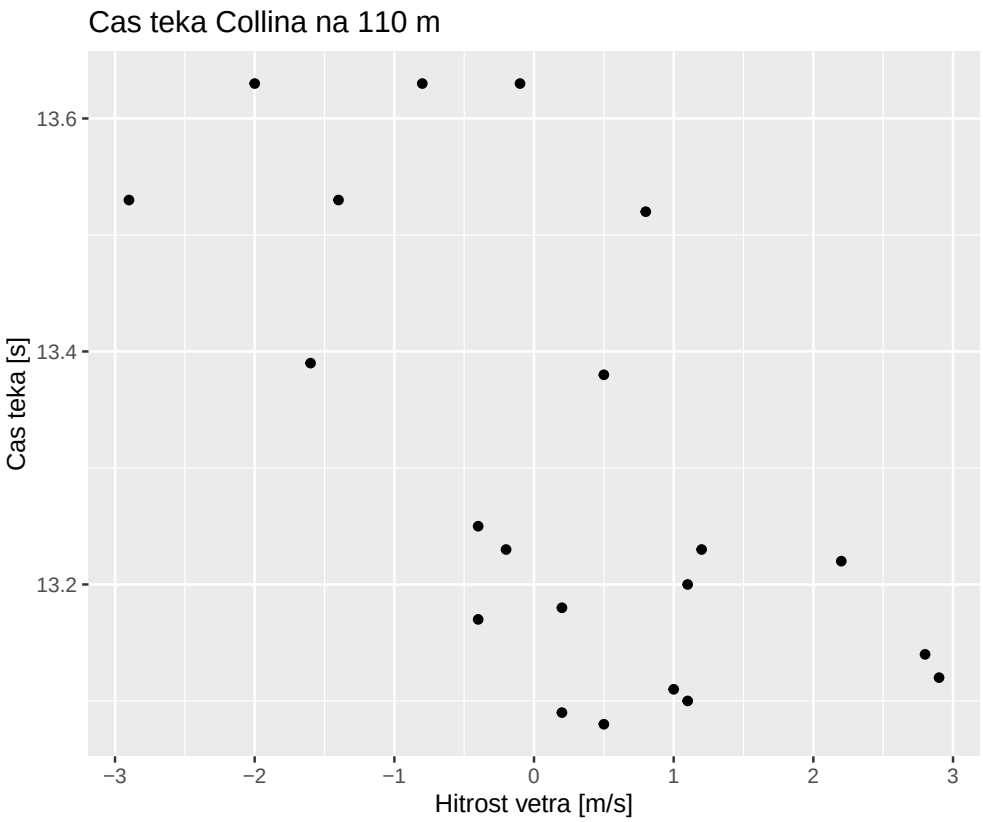
```
   windspeed            time
 Min.   :-2.9000   Min.   :13.08
 1st Qu.:-0.4000   1st Qu.:13.14
 Median : 0.2000   Median :13.23
 Mean   : 0.2238   Mean   :13.30
 3rd Qu.: 1.1000   3rd Qu.:13.52
 Max.   : 2.9000   Max.   :13.63
```

```r
#poglejmo najprej univariatne porazdelitve spremenljivk v podatkovnem okviru
par(mfrow=c(1,2))

boxplot(d$windspeed, main="", xlab="Hitrost vetra [m/s]")
boxplot(d$time, main="",  xlab="Čas teka [s]")
```



Slika 8: Univariatne porazdelitve spremenljivk v podatkovnem okviru `COLLIN`.

```r
#Ali obstaja linearna povezanost med spremenljivkama?
ggplot(data=d, aes(x=windspeed, y=time)) +
  geom_point() +
  xlab("Hitrost vetra [m/s]") +
  ylab("Čas teka [s]") +
  ggtitle("Čas teka Collina na 110 m")
```



Slika 9: Odvisnost časa teka od hitrosti vetra.

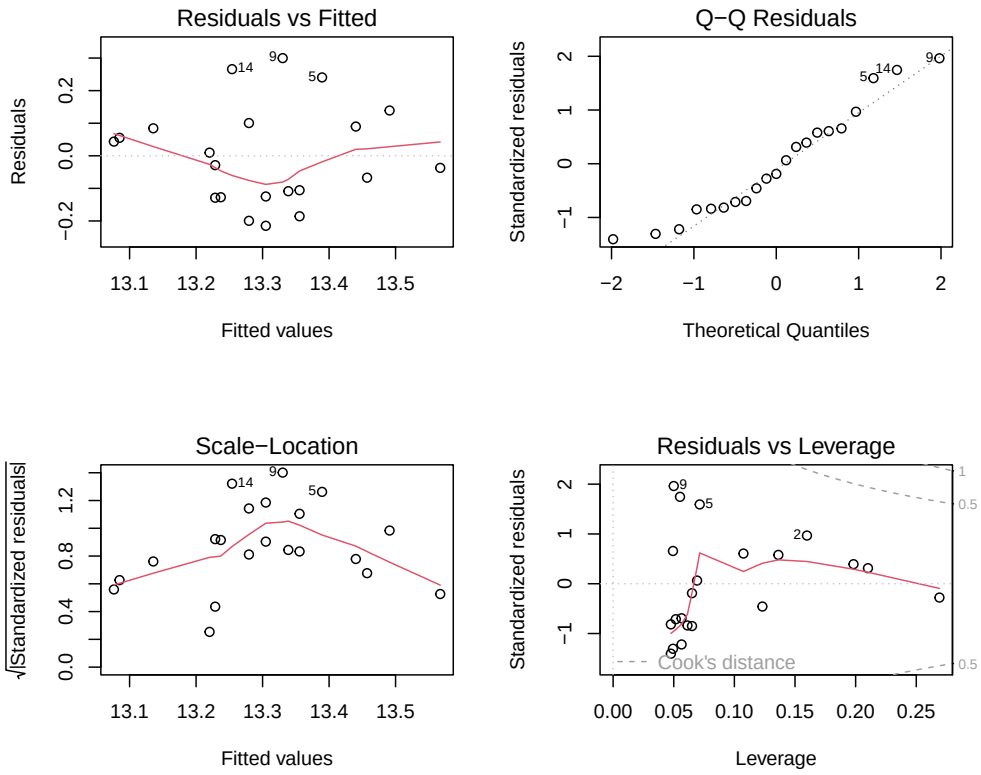
Enostavni linearni model za čas teka Collina v odvisnosti od hitrosti vetra:

```
m_collin <- lm(time~windspeed, data = d)
summary(m_collin)
```

```
Call:
lm(formula = time ~ windspeed, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21487 -0.12487 -0.02873  0.08976  0.29975

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.32179    0.03460 385.043  < 2e-16 ***
windspeed   -0.08460    0.02361  -3.584  0.00198 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1567 on 19 degrees of freedom
Multiple R-squared:  0.4033,    Adjusted R-squared:  0.3719
F-statistic: 12.84 on 1 and 19 DF,  p-value: 0.00198
```

Slike ostankov:

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



Slika 10: Ostanki za model `time` ∼ `windspeed`.

Na prvi levi sličici Slike 10 so prikazani ostanki (Residuals) v odvisnosti od napovedanih vrednosti (Fitted values). Vidimo, da prihaja do odstopanj gladilnika (rdeča črta, ki jo nariše funkcija loess – local polynomial regression fitting) od vodoravne črte z vrednostjo ostanka 0. Ker je model narejen le na 21 enotah, so odstopanja normalna in še ne pomenijo kršitve predpostavk. Vsaka točka namreč gladilnik potegne v svojo smer, zato smo pri analizah majhnih vzorcev načeloma do odstopanj bolj tolerantni. Zgornja desna slika kaže porazdelitev standardiziranih ostankov v primerjavi z normalno porazdelitvijo (ravna črta v ozadju). Točke se dobro prilegajo ravni črtkani črti, kar pomeni, da lahko privzamemo normalno porazdelitev za standardizirane ostanke. Spodnja leva slika prikazuje koren absolutne vrednosti standardiziranih ostankov v odvisnosti od napovedanih vrednosti. Gladilnik, ki je vodoraven, pomeni, da med ostanki ni prisotne heteroskedastičnosti. Tudi tu lahko odstopanja pripišemo majhnemu vzorcu. Spodnja desna slika identificira vplivne točke, ki jih v tem primeru ni.

```
coef(m_collin)
```

```
(Intercept)   windspeed
13.32179201 -0.08460258
```

```
confint(m_collin)
```

```
                2.5 %      97.5 %
(Intercept) 13.2493772 13.39420677
windspeed   -0.1340109 -0.03519423
```

Interpretacija modela: v povzetku modela se testirata dve ničelni hipotezi. Prva testira, ali je čas teka Collina na 110 m v brezveterju enak 0 (ni smiselna). Domnevo zavrnemo: imamo 95 % zaupanje, da je čas teka Collina na 110 m v brezveterju nekje med 13.25 in 13.39 s. Druga ničelna domneva testira, ali je čas teka Collina na 110 m odvisen od hitrosti vetra (smiselna). Tudi to domnevo zavrnemo v prid alternativne: obstaja povezanost med časom teka in hitrotjo vetra. S 95 % zaupanjem lahko trdimo, če se hitrost vetra v hrbet poveča za 1 m/s, se čas teka v povprečju zmanjša med -0.13 in -0.04 s. Z modelom smo pojasnili 40 % variabilnosti odzivne spremenljivke.

```
casi = data.frame(windspeed = c(-1, 0, 1, 4))

povprecne_napovedi = data.frame(predict(m_collin, casi, interval = "confidence"))
povprecne_napovedi = data.frame(cbind(casi,
                                      povprecne_napovedi$fit,
                                      paste0("(",
                                             round(povprecne_napovedi$lwr, 2),
                                             ", ",
                                             round(povprecne_napovedi$upr, 2), ")")
                                      ))

colnames(povprecne_napovedi) = c("Hitrost vetra [m/s]", "Povprečna napoved [s]", "95 % IZ")

kable(povprecne_napovedi,
      digits = c(0, 2, 0),
      caption = "Povprečna napoved časa teka.") %>%
  kable_styling("striped", full_width = F)
```

Tabela 3: Povprečna napoved časa teka.

Hitrost vetra [m/s]	Povprečna napoved [s]	95 % IZ
-1	13.41	(13.31, 13.5)
0	13.32	(13.25, 13.39)
1	13.24	(13.16, 13.32)
4	12.98	(12.78, 13.18)

```
posamicne_napovedi = data.frame(predict(m_collin, casi, interval = "prediction"))
posamicne_napovedi = data.frame(cbind(casi,
                                      posamicne_napovedi$fit,
                                      paste0("(",
                                             round(posamicne_napovedi$lwr, 2),
                                             ", ",
                                             round(posamicne_napovedi$upr, 2), ")")
                                      ))

colnames(posamicne_napovedi) = c("Hitrost vetra [m/s]", "Posamična napoved [s]", "95% IZ")

kable(posamicne_napovedi,
      digits = c(0, 2, 0),
      caption = "Posamična napoved časa teka.") %>%
  kable_styling("striped", full_width = F)
```

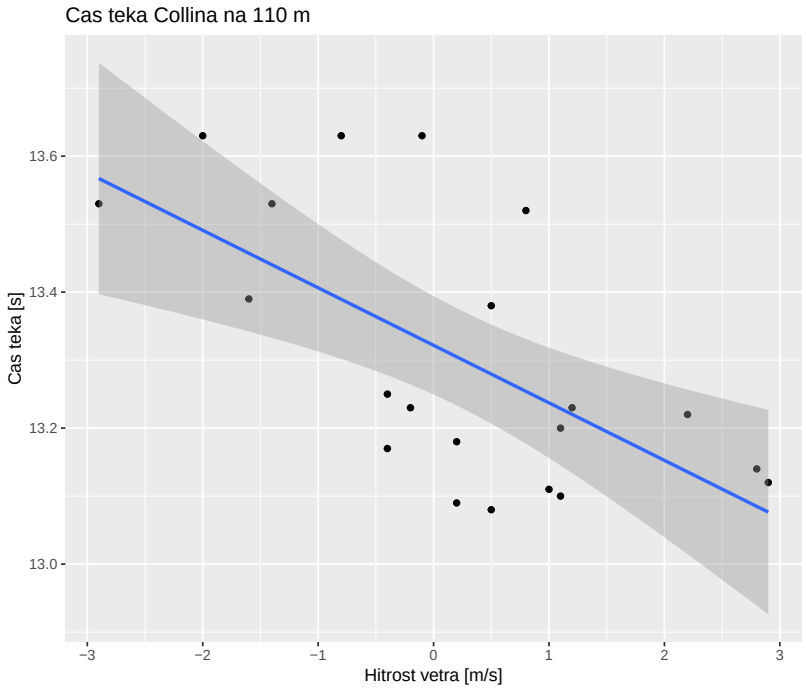
Tabela 4: Posamična napoved časa teka.

Hitrost vetra [m/s]	Posamična napoved [s]	95% IZ
-1	13.41	(13.07, 13.75)
0	13.32	(12.99, 13.66)
1	13.24	(12.9, 13.58)
4	12.98	(12.6, 13.37)

Primer interpretacije za hitrost vetra 1 m/s: napovedana vrednost za čas teka je 13.24 s. Pripadajoči 95 % IZ za povprečni čas teka pri hitrosti vetra 1 m/s je (13.16 s, 13.32 s). Za posamezni tek pri hitrosti vetra 1 m/s je pripadajoči 95 % IZ (12.9 s, 13.58 s).

Napoved za hitrost vetra 4 m/s ni upravičena, saj gre za ekstrapolacijo.

```
ggplot(data=d, aes(x = windspeed, y = time)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = "lm", se = TRUE) +
  xlab("Hitrost vetra [m/s]") +
  ylab("Čas teka [s]") +
  ggtitle("Čas teka Collina na 110 m")
```



Slika 11: Odvisnost `time` od `windspeed` za dani vzorec 21 tekov in regresijska premica s 95 % intervali zaupanja za povprečno napoved.

3. Simulacija podatkov iz modela linearne regresije

Zanimajo nas lastnosti cenilk enostavnega linearnega regresijskega modela. Osredotočili se bomo na testiranje domneve $H_0 : \beta_1 = 0$. Za izbrani vrednosti parametrov enostavne linearne regresije $\beta_0 = 100$ in $\beta_1 = 1$ bomo izvedli simulacije, ki bodo ilustrirale vpliv velikosti vzorca n in vrednosti variance napak σ^2 na porazdelitev cenilk parametrov in na moč testa pri testiranju domneve $H_0 : \beta_1 = 0$. Za vsako izbrano velikost vzorca n najprej generiramo vrednosti napovedne spremenljivke `x` na intervalu 15 do 70. Pri tem uporabimo funkcijo `sample` z argumentom `replace=TRUE`: `x<-sample(c(17:70), size=n, replace=TRUE)`. Za tako določene

vrednosti napovedne spremenljivke generiramo vrednosti odzivne spremenljivke, pri čemer upoštevamo da so pogojno na vrednosti napovedne spremenljivke porazdeljene normalno s pričakovano vrednostjo $\beta_0 + \beta_1$`X` in varianco σ^2: $y_i = 100 + x_i + \varepsilon_i$; napake ε_i, $i = 1, ..., 50$, generiramo s funkcijo `rnorm()` za porazdelitev $N(0, \sigma^2 = 11^2)$.

Z namenom odgovoriti na naslednja vprašanja:

- Kakšne so porazdelitve ocen parametrov enostavnega linearnega modela?
- Kolikšen delež intervalov zaupanja za β_1 vsebuje pravo vrednost parametra?
- Kolikšna je moč testa pri testiranju ničelne domneve $H_0 : \beta_1 = 0$?

bomo izvedli simulacije, pri čemer bomo podatke generirali 1000-krat in za vsak generirani vzorec izračunali cenilki parametrov enostavnega linearnega modela b_0 in b_1, 95 % interval zaupanja za β_1 in p-vrednost pri testiranju domneve $H_0 : \beta_1 = 0$.

```
f.reg.sim <- function(x, beta0, beta1, n, sigma, nsim){

  # pripravimo prazne vektorje za rezultate simulacij, cenilki parametrov b0 in b1,
  # p-vrednost za testiranje domneve beta1=0,
  # spodnjo in zgornjo mejo intervala zaupanja za beta1
  b0 <- b1 <- l.b1 <- u.b1 <- p.b1 <- NULL

  for(i in 1:nsim){
    epsilon <- rnorm(n, mean=0, sd=sigma)
    y <- beta0 + beta1*x + epsilon
    m <- lm(y~x)
    b0[i] <- coef(m)[1]
    b1[i] <- coef(m)[2]
    l.b1[i] <- confint(m)[2, 1]
    u.b1[i] <- confint(m)[2, 2]
    p.b1[i] <- summary(m)$coef[2, 4]
  }
  return(data.frame(b0, b1, l.b1, u.b1, p.b1))
}
```

```
#parametra modela
beta0 <- 100
beta1 <- 1
#velikost vzorca
n <- 50
#standardno odklon napak
sigma <- 11
#generiramo vrednosti x
x <- sample(c(17:70), size=n, replace=TRUE)

#število simulacij
nsim <- 1000

#nastavimo seme za ponovljivost
set.seed(20)
rez.1000 <- f.reg.sim(x=x, beta0, beta1, n, sigma, nsim)
```

```
# 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%
```

```
0.815409 1.186635
```

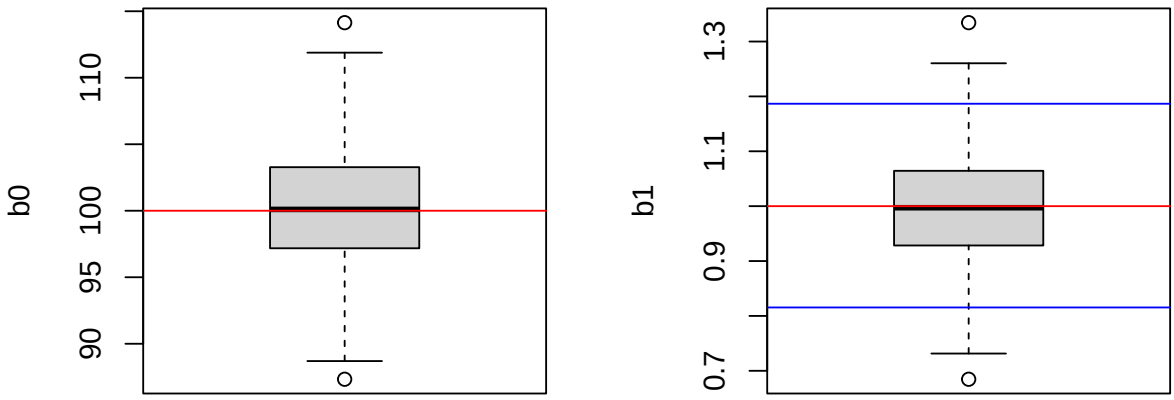
```
# ocena verjetnosti za napako II. vrste za H0: beta1=0
alfa  <- 0.05
sum(rez.1000$p.b1 > alfa)/nsim
```

```
[1] 0
```

```
# ocena moči testa na podlagi Nsim simulacij
(moc.testa <- 1 - sum(rez.1000$p.b1 > alfa)/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 12: Porazdelitev cenilk parametrov b_0 (levo) in b_1 (desno) za $\sigma = 11$ in $n = 50$, `set.seed(20)`, rdeča črta kaže pravo vrednost za parameter, modri črti predstavljata 2.5 in 97.5 centil za b_1.

```
# delež intervalov zaupanja, ki ne vsebujejo prave vrednosti parametra beta1,
# (ocena velikosti testa)
sum(rez.1000$l.b1 > beta1 | rez.1000$u.b1 < beta1)/nsim
```

```
[1] 0.052
```

Domača naloga: Simulacije iz modela enostavne linearne regresije

Simulacije ponovite za vse kombinacije:

- različnih velikosti vzorcev n: 10, 15, 50 in 1000 in
- različnih vrednosti σ: 5, 11, 22.

Grafično prikažite:

- odvisnost širine intervala zaupanja za β_1 od n, za vsako vrednost σ;
- odvisnost širine intervala zaupanja za β_1 od σ, za vsak n;
- odvisnost moči testa od n, za vsako vrednost σ;
- odvisnost moči testa od σ, za vsak n

in napišite kratek povzetek vaših ugotovitev.