Kazalo

1 KOLINEARNOST

Kolinearnost v linearnem modelu pomeni, da so nekateri regresorji tesno korelirani med seboj in v model dodajo zelo podobno informacijo. V primeru kolinearnosti različne kombinacije regresorjev dajo zelo podobne napovedane vrednosti. Kolinearnost zato predstavlja večji problem za interpretacijo modela, kot pa za napovedovanje. V literaturi se za **kolinearnost** pogosto uporablja izraz **multikolinearnost**. Ta izraz povdarja, da ni nujno, da gre za povezanost napovednih spremenljivk po parih temveč za t. i. multiplo povezanost v kateri je ena spremenljivka korelirana z drugo samo ob prisotnosti tretje.

Kadar je en regresor linearno popolnoma odvisen od ostalih regresorjev, pravimo, da gre za **popolno kolinearnost**. V takem primeru matrika $\mathbf{X}$ nima polnega ranga in sistem enačb, ki ga rešujemo po metodi najmanjših kvadratov nima enolične rešitve. Če odvisnost enega regresorja od ostalih ni popolna, kar pomeni, da je linearna kombinacija regresorjev blizu 0, govorimo o kolinearnosti, in sistem normalnih enačb, ki ga rešujemo ob ocenjevanju parametrov modela, ima enolično rešitev. V takem modelu majhne spremembe v podatkih povzročijo velike spremembe v ocenah parametrov, saj so te močno odvisne od drugih regresorjev v modelu.

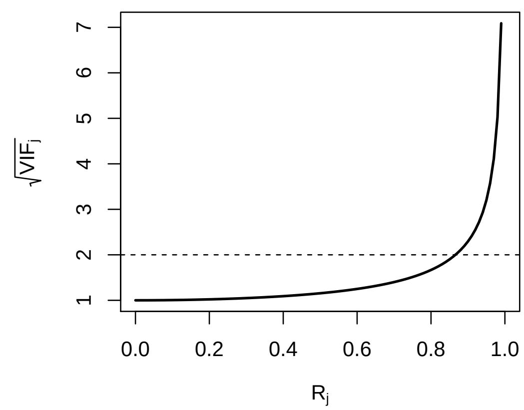
Prisotnost kolinearnosti v modelu se lahko kaže na različne načine:

- v matriki korelacijskih koeficientov številskih napovednih spremenljivk so nekatere vrednosti blizu 1 ali -1;
- vse napovedne spremenljivke so neznačilne, hkrati je vrednost koeficienta determinacije velika;
- na diagonali matrike $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ so velike vrednosti, kar se lahko odraža v velikih standardnih napakah in širokih intervalih zaupanja za nekatere parametre;
- zaloga vrednosti ostankov na vodoravnih oseh grafov dodane spremenljivke (`avPlots`) je manjša pri napovednih spremenljivkah, ki so korelirane z drugimi napovednimi spremenljivkami;
- velike vrednosti statistike VIF (*variance inflation factor*) oziroma $GVIF$ (*generalized variance inflation factor*).

Statistika VIF služi za ugotavljanje prisotnosti kolinearnosti za posamezno številsko napovedno spremenljivko. VIF_j temelji na zapisu ocene variance za oceno b_j:

$$\widehat{Var}(b_j) = \frac{\widehat{\sigma}^2}{SS_{xx_j}} \cdot \frac{1}{1-R_j^2} = \frac{\widehat{\sigma}^2}{SS_{xx_j}} \cdot VIF_j. \tag{1}$$

V (1) je $\widehat{\sigma}^2$ z modelom ocenjena varianca napak, SS_{xx_j} je vsota kvadratov odklonov od povprečja za x_j, R_j je koeficient multiple korelacije, ki ga dobimo z regresijo x_j na vse ostale x_i, $i \neq j$. Člen $1/(1-R_j^2)$ se imenuje VIF_j (*variance inflation factor*) in je mera nadlog, ki jih povzroči kolinearnost pri spremenljivki x_j. $\sqrt{VIF_j}$ meri, kolikokrat je interval zaupanja za β_j povečan relativno na situacijo, kjer kolinearnosti med x_j in ostalimi regresorji v modelu ne bi bilo. Na Sliki 1 je prikazana odvisnost $\sqrt{VIF_j}$ od koeficienta multiple korelacije R_j. Za dvakrat povečan interval zaupanja za β_j mora imeti VIF vrednost $VIF_j = 4$, ta vrednost ustreza vrednosti koeficienta multiple korelacije $\sqrt{3/4} = 0.87$. Če je $VIF_j = 9$, kar pomeni trikratno povečanje intervala zaupanja, ima koeficient multiple korelacije vrednost $\sqrt{8/9} = 0.89$.



Slika 1: $\sqrt{VIF_j}$ v odvisnosti od koeficienta multiple korelacije R_j

V literaturi obstoja več kriterijev za vrednost VIF, pri kateri se lahko pojavijo problemi zaradi kolinearnosti. Največkrat je kot opozorilna vrednost za VIF omenjena vrednost 4 ali 5, kolinearnost pa lahko zahteva poseg v model pri vrednostih nad 10.

Če imamo v model, kjer ocenjujemo $k+1$ parametrov, vključeno opisno napovedno spremenljivko z l vrednostmi, analiza kolinearnosti temelji na povezanosti pripadajočih $l-1$ regresorjev s skupino preostalih regresorjev. V takem primeru linearni model v matrični obliki zapišemo v treh delih:

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X_1}\boldsymbol{\beta}_1 + \mathbf{X_2}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{2}$$

$\mathbf{y}$ je vektor odzivne spremenljivke, $\mathbf{1}$ je enotski vektor reda $n \times 1$, $\mathbf{X_1}$ je modelska matrika opisne napovedne spremenljivke reda $n \times (l-1)$, $\boldsymbol{\beta}_1$ je vektor parametrov vezanih na opisno napovedno spremenljivko reda $(l-1) \times 1$; $\mathbf{X_2}$ je modelska matrika ostalih regresorjev reda $n \times (k-l+1)$, $\boldsymbol{\beta}_2$ je vektor parametrov vezanih na ostale regresorje reda $(k-l+1) \times 1$ in $\boldsymbol{\varepsilon}$ je vektor napak, za katerega velja $E(\boldsymbol{\varepsilon}) = 0$ in $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $\mathbf{I}$ je enotska diagonalna matrika reda $n \times n$. Fox in Monette (1992) sta pokazala, da se VIF skupine regresorjev v $\mathbf{X_1}$ v takem primeru izrazi kot $GVIF_1$:

$$GVIF_1 = \frac{det\mathbf{R}_{11}det\mathbf{R}_{22}}{det\mathbf{R}}, \tag{3}$$

kjer je $\mathbf{R}_{11}$ korelacijska matrika za $\mathbf{X_1}$, $\mathbf{R}_{22}$ korelacijska matrika za $\mathbf{X_2}$ in $\mathbf{R}$ korelacijska matrika za vse regresorje hkrati. Fox in Monette sta pokazala, da je vrednost $GVIF^{1/(2SP)}$ analogna vrednosti $\sqrt{VIF}$, pri čemer je $SP = l - 1$ število stopinj prostosti opisne napovedne spremenljivke. V primeru, da ima napovedna spremenljivka samo eno stopinjo prostosti, je $VIF = GVIF$. Opozorilne vrednosti za prisotnost kolinearnosti so za **kvadrirano vrednost** $GVIF^{1/(2SP)}$ enake kot pri VIF.

Način izračunavanja smo pokazali za primer opisne napovedne spremenljivke z l vrednostmi, postopek je enak v primeru polinomske regresije reda l ali v primeru uporabe regresijskih zlepkov z $l + 1$ vozlišči.

Kako odpravimo kolinearnost:

- na podlagi matrike korelacijskih koeficientov in vsebinske presoje izločimo določene napovedne spremenljivke;
- iz več koreliranih napovednih spremenljivk naredimo nove med seboj neodvisne spremenljivke z uporabo metode glavnih komponent (PCA) na napovednih spremenljivkah;
- iz več koreliranih spremenljivk naredimo eno novo spremenljivko (npr. telesna masa in telesna višina sta ponavadi korelirani, izračunamo indeks telesne mase $ITM = masa/visina^2$, masa v kg in višina v m)
- uporaba Ridge regresije.

Za ilustracijo poglejmo primer popolne kolinearnosti.

```
> set.seed(777)
> x1<- runif(100, min = 0, max = 10)
> x2<-(-x1)
> x3<- x1 + rnorm(100, mean = 0, sd = 1)
> x4<-runif(100, min = 0, max = 10)
> y<-x1 + x2 + x3 + x4 + rnorm(100, mean = 0, sd = 1)
> # korelacijska matrika napovednih spremenljivk
> round(cor(cbind(x1, x2, x3, x4)), 4)
```

```
        x1      x2      x3      x4
x1  1.0000 -1.0000  0.9547 -0.1101
x2 -1.0000  1.0000 -0.9547  0.1101
x3  0.9547 -0.9547  1.0000 -0.0896
x4 -0.1101  0.1101 -0.0896  1.0000
```

```
> mod.0<-lm(y~x1+x2+x3+x4)
```

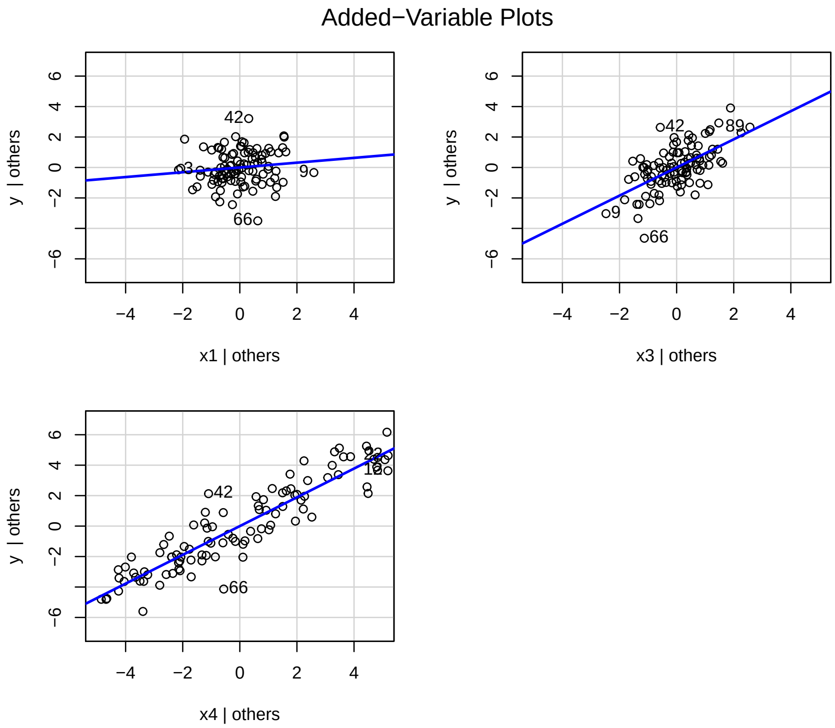
```
> summary(mod.0)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6019 -0.7605 -0.0638  0.7980  3.1619

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01599    0.30487  -0.052    0.958
x1           0.15750    0.12939   1.217    0.226
x2                NA         NA      NA       NA
x3           0.92401    0.12206   7.570 2.28e-11 ***
x4           0.94425    0.04055  23.286  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.139 on 96 degrees of freedom
Multiple R-squared:  0.9306,	Adjusted R-squared:  0.9285
F-statistic: 429.3 on 3 and 96 DF,  p-value: < 2.2e-16

> X.0<-model.matrix(mod.0)
> det(t(X.0)%*%X.0)

[1] 0

> library(car)
> # vif(mod.0) se ne izračuna
```

Ilustracija multikolinearnosti:

```
> mod.1<-lm(y~x1+x3+x4)
> vif(mod.1)

       x1        x3        x4
11.376815 11.329811  1.015077

> coef(summary(mod.1))

               Estimate Std. Error     t value     Pr(>|t|)
(Intercept) -0.01598637 0.30486686 -0.05243721 9.582893e-01
x1           0.15750359 0.12939064  1.21727190 2.264846e-01
x3           0.92401056 0.12206415  7.56987679 2.284579e-11
x4           0.94425226 0.04055012 23.28605078 2.844917e-41

> Confint(mod.1)
```

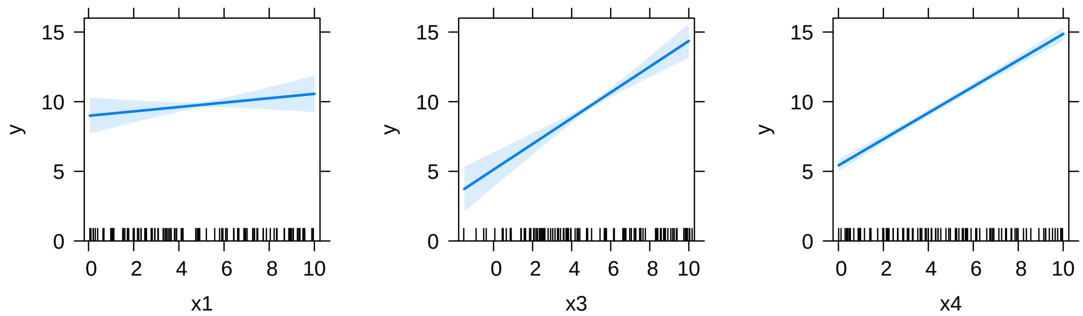
```
                Estimate      2.5 %    97.5 %
(Intercept) -0.01598637 -0.6211423 0.5891696
x1           0.15750359 -0.0993348 0.4143420
x3           0.92401056  0.6817151 1.1663060
x4           0.94425226  0.8637609 1.0247436
```

```
> avPlots(mod.1, ylim=c(-7,7), xlim=c(-5, 5))
```



Slika 2: Grafi dodane spremenljivke za `mod.1`, interval vrednosti ostankov na osi x je pri spremenljivkah z visoko vrednostjo VIF (`x1` in `x3`) veliko ožji kot pri `x4`

```
> library(effects)
> plot(predictorEffects(mod.1, ~.), rows=1, cols=3, main="", ylim=c(0,16))
```



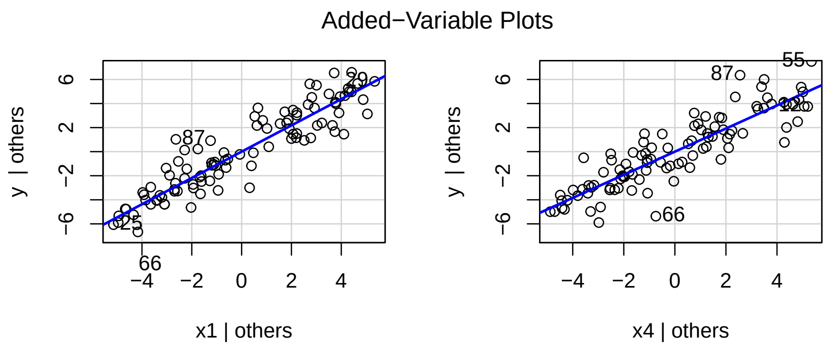
Slika 3: Napovedane vrednosti za `y` s 95 % intervali zaupanja za povprečno napoved za `mod.1`, pri (`x1` in `x3`) se intervali zaupanja hitro širijo z oddaljenostjo od povprečne vrednosti

Zaradi kolinearnosti izločimo spremenljivko `x3` iz modela (isto bi lahko naredili z `x1`):

```
> mod.1a<-lm(y~x1+x4)
> vif(mod.1a)

      x1       x4
1.012276 1.012276

> summary(mod.1a)

Call:
lm(formula = y ~ x1 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6477 -0.8227  0.0049  0.7323  3.9056

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.10270    0.38299  -0.268    0.789
x1           1.09238    0.04852  22.514   <2e-16 ***
x4           0.96038    0.05091  18.865   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.432 on 97 degrees of freedom
Multiple R-squared:  0.8892,	Adjusted R-squared:  0.8869
F-statistic: 389.3 on 2 and 97 DF,  p-value: < 2.2e-16

> confint(mod.1a)
```

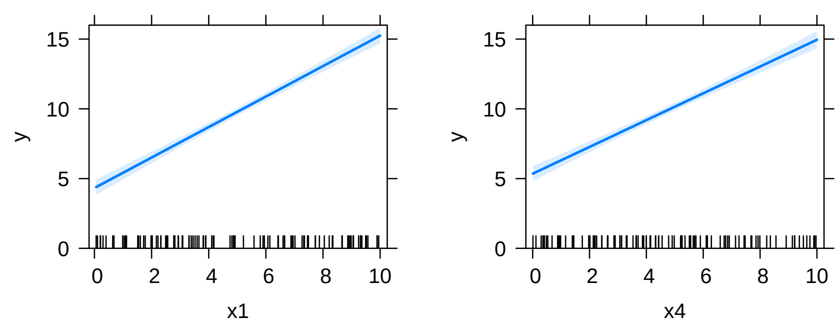
```
                 2.5 %    97.5 %
(Intercept) -0.8628432 0.6574342
x1           0.9960834 1.1886857
x4           0.8593416 1.0614161
```

```
> avPlots(mod.1a, ylim=c(-7,7))
```



Slika 4: Grafi dodane spremenljivke za `mod.1a`

```
> plot(predictorEffects(mod.1a, ~.), rows=1, cols=2, main="", ylim=c(0,16))
```



Slika 5: Napovedane vrednosti za `y` za `mod.1a`

1.1 Primer: `seatpos`

V paketu `faraway` so v podatkovnem okviru `seatpos` naslednji podatki: oddaljenost sredine med kolkoma voznika od fiksne točke v avtu (`hipcenter` v mm), starost voznika (`Age` v letih), telesna masa (`Weight` v funtih), telesna višina voznika z obutimi čevlji (`HtShoes` v cm), telesna višina z bosimi nogami (`Ht` v cm), razdalja od stola do vrha glave šoferja (`Seated` v cm), dolžina roke od komolca navzdol (`Arm` v cm), dolžina stegna (`Thigh`, v cm), dolžina noge od kolena navzdol (`Leg` v cm). Podatke za 38 voznikov so zbrali v HuMoSim laboratoriju na University of Michigan.

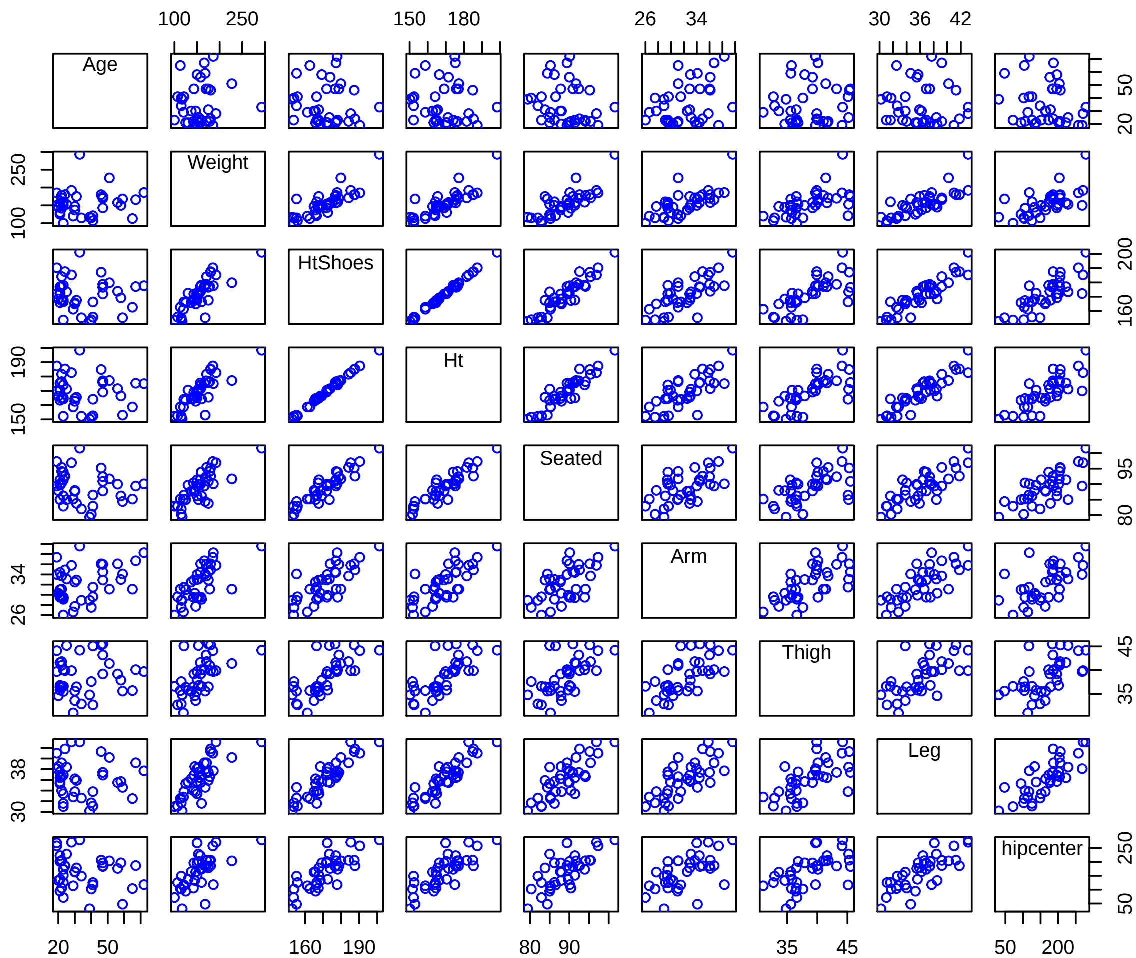
Raziskovalce je zanimala odvisnost `hipcenter` od ostalih spremenljivk. Naredite ustrezni statistični model, izvedite diagnostiko izbranega modela in ga obrazložite.

```
> library(faraway)
> data(seatpos)
> summary(seatpos)
```

```
      Age            Weight         HtShoes           Ht
 Min.   :19.00   Min.   :100.0   Min.   :152.8   Min.   :150.2
 1st Qu.:22.25   1st Qu.:131.8   1st Qu.:165.7   1st Qu.:163.6
 Median :30.00   Median :153.5   Median :171.9   Median :169.5
 Mean   :35.26   Mean   :155.6   Mean   :171.4   Mean   :169.1
 3rd Qu.:46.75   3rd Qu.:174.0   3rd Qu.:177.6   3rd Qu.:175.7
 Max.   :72.00   Max.   :293.0   Max.   :201.2   Max.   :198.4
     Seated            Arm             Thigh            Leg
 Min.   : 79.40   Min.   :26.00   Min.   :31.00   Min.   :30.20
 1st Qu.: 85.20   1st Qu.:29.50   1st Qu.:35.73   1st Qu.:33.80
 Median : 89.40   Median :32.00   Median :38.55   Median :36.30
 Mean   : 88.95   Mean   :32.22   Mean   :38.66   Mean   :36.26
 3rd Qu.: 91.62   3rd Qu.:34.48   3rd Qu.:41.30   3rd Qu.:38.33
 Max.   :101.60   Max.   :39.60   Max.   :45.50   Max.   :43.10
   hipcenter
 Min.   :-279.15
 1st Qu.:-203.09
 Median :-174.84
 Mean   :-164.88
 3rd Qu.:-119.92
 Max.   : -30.95
```

```
> # vrednosti za hipcenter v podatkovnem okviru setpos so negativne
> # interpretacija je lažja, če so pozitivne
> seatpos$hipcenter<-(-1)*seatpos$hipcenter
```

```
> scatterplotMatrix(seatpos, regLine=FALSE,
+                   diagonal=FALSE, smooth=FALSE, data=seatpos)
```



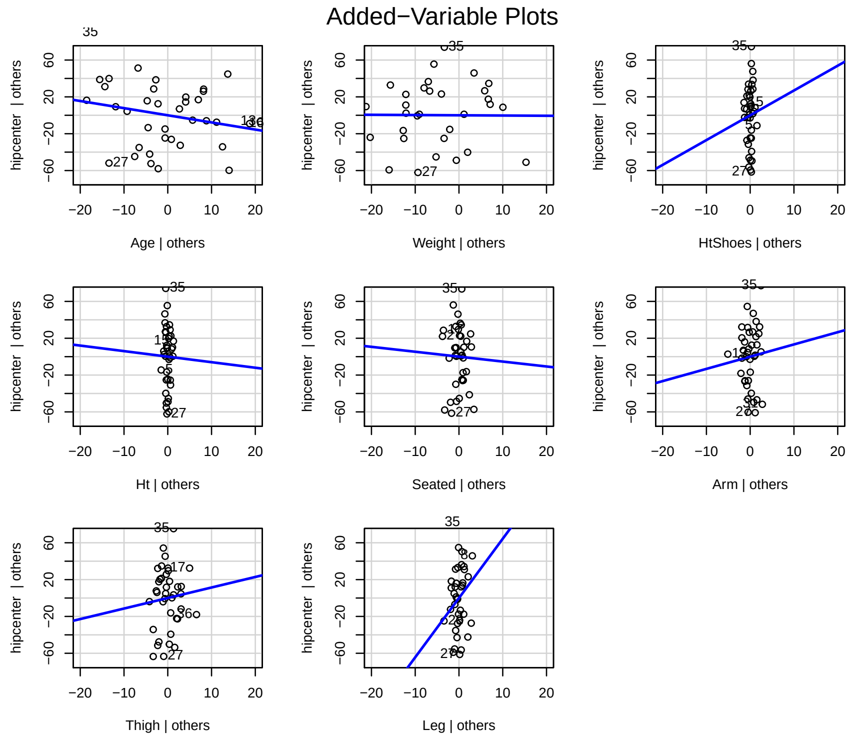
Slika 6: Matrika razsevnih grafikonov za vse številske spremenljivke podatkovnega okvira `seatpos`

```
> round(cor(seatpos, method="spearman"),3)

             Age Weight HtShoes     Ht Seated   Arm Thigh    Leg hipcenter
Age        1.000  0.071  -0.093 -0.094 -0.215 0.275 0.062 -0.099    -0.186
Weight     0.071  1.000   0.848  0.857  0.757 0.719 0.648  0.792     0.664
HtShoes   -0.093  0.848   1.000  0.991  0.901 0.741 0.767  0.892     0.798
Ht        -0.094  0.857   0.991  1.000  0.898 0.759 0.776  0.900     0.819
Seated    -0.215  0.757   0.901  0.898  1.000 0.564 0.627  0.748     0.683
Arm        0.275  0.719   0.741  0.759  0.564 1.000 0.671  0.744     0.603
Thigh      0.062  0.648   0.767  0.776  0.627 0.671 1.000  0.669     0.659
Leg       -0.099  0.792   0.892  0.900  0.748 0.744 0.669  1.000     0.799
hipcenter -0.186  0.664   0.798  0.819  0.683 0.603 0.659  0.799     1.000
```

Napovedne spremenljivke z izjemo `Age` so medsebojno močno korelirane. Pričakujemo težave zaradi kolinearnosti.

```
> mod.0<-lm(hipcenter~., data=seatpos)
> vif(mod.0)

       Age     Weight    HtShoes         Ht     Seated        Arm      Thigh
  1.997931   3.647030 307.429378 333.137832   8.951054   4.496368   2.762886
       Leg
  6.694291
```

```
> avPlots(mod.0, ylim=c(-70,70), xlim=c(-20, 20))
```



Slika 7: Grafi dodane spremenljivke za `mod.0`, interval vrednosti ostankov na osi x je pri spremenljivkah z visoko vrednostjo VIF (`x1` in `x3`) veliko ožji kot pri `x4`

```
> summary(mod.0)

Call:
lm(formula = hipcenter ~ ., data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max 
-62.337 -25.017   3.678  22.833  73.827 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -436.43213  166.57162  -2.620   0.0138 *
Age           -0.77572    0.57033  -1.360   0.1843  
Weight        -0.02631    0.33097  -0.080   0.9372  
HtShoes        2.69241    9.75304   0.276   0.7845  
Ht            -0.60134   10.12987  -0.059   0.9531  
Seated        -0.53375    3.76189  -0.142   0.8882  
Arm            1.32807    3.90020   0.341   0.7359  
Thigh          1.14312    2.66002   0.430   0.6706  
Leg            6.43905    4.71386   1.366   0.1824  
---
```

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

Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared:  0.6866,	Adjusted R-squared:  0.6001
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Z `mod.0` je pojasnjene 68.66 % variabilnosti odzivne spremenljivke, vendar ni statistično značilna nobena napovedna spremenljivka. Standardni napaki pri `HtShoes` in `Ht` sta zelo veliki. VIF spremenljivk `HtShoes` in `Ht` je ogromen. Tudi njun Spearmanov koeficient korelacije je zelo velik (0.991). Poglejmo, kako se spremenijo VIF vrednosti, če iz modela izločimo `HtShoes`:

```
> mod.1<-update(mod.0, .~. -HtShoes, data=seatpos)
> vif(mod.1)

      Age    Weight        Ht    Seated       Arm     Thigh       Leg
 1.875729  3.628705 23.352154  8.808440  4.482567  2.626556  6.690858

> summary(mod.1)

Call:
lm(formula = hipcenter ~ Age + Weight + Ht + Seated + Arm + Thigh +
    Leg, data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max
-61.595 -24.739   5.471  21.565  74.570

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -435.80897  163.97188  -2.658   0.0125 *
Age           -0.73678    0.54404  -1.354   0.1858
Weight        -0.03279    0.32502  -0.101   0.9203
Ht             2.09530    2.64036   0.794   0.4337
Seated        -0.40267    3.67390  -0.110   0.9135
Arm            1.26842    3.83378   0.331   0.7431
Thigh          0.98000    2.55332   0.384   0.7038
Leg            6.46852    4.63953   1.394   0.1735
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.13 on 30 degrees of freedom
Multiple R-squared:  0.6857,	Adjusted R-squared:  0.6124
F-statistic: 9.351 on 7 and 30 DF,  p-value: 4.157e-06
```

Še vedno so prisotne težave s kolinearnostjo. Ker je `Ht` lažje dostopna spremenljivka, v naslednjem koraku izločimo `Seated` in `Leg`.

```
> mod.2<-update(mod.1, .~. -Seated -Leg, data=seatpos)
> vif(mod.2)
```

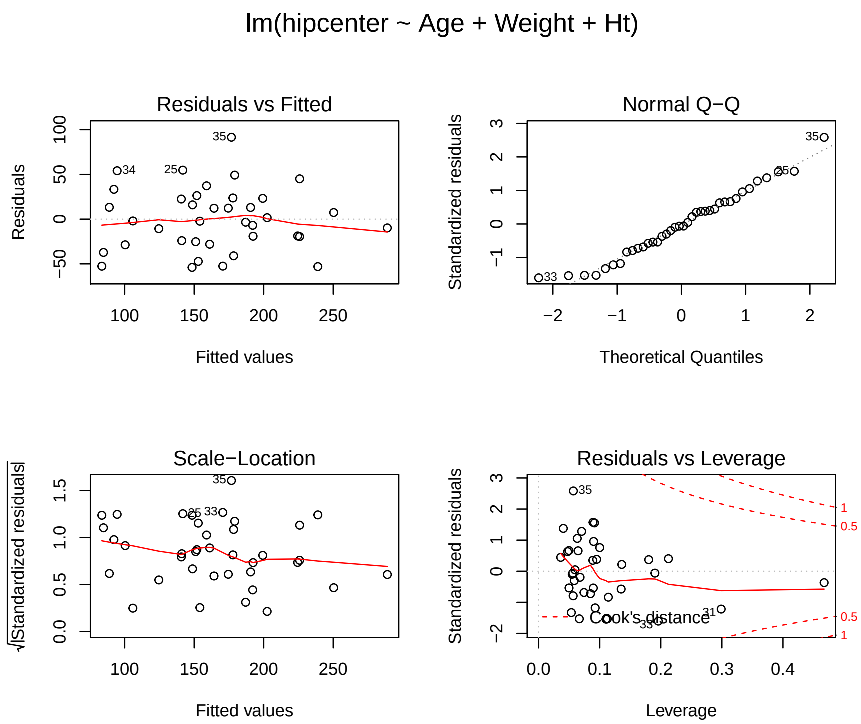
```
     Age   Weight       Ht      Arm    Thigh
1.847327 3.574090 7.260856 4.105119 2.432315
```

```
> summary(mod.2)

Call:
lm(formula = hipcenter ~ Age + Weight + Ht + Arm + Thigh, data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max
-57.945 -25.935   0.301  24.368  81.891

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.899e+02  1.460e+02  -3.356  0.00205 **
Age         -8.109e-01  5.407e-01  -1.500  0.14354
Weight       2.932e-03  3.231e-01   0.009  0.99281
Ht           3.366e+00  1.475e+00   2.283  0.02924 *
Arm          2.796e+00  3.675e+00   0.761  0.45235
Thigh        6.127e-01  2.461e+00   0.249  0.80498
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.19 on 32 degrees of freedom
Multiple R-squared:  0.6637,	Adjusted R-squared:  0.6112
F-statistic: 12.63 on 5 and 32 DF,  p-value: 8.141e-07
```

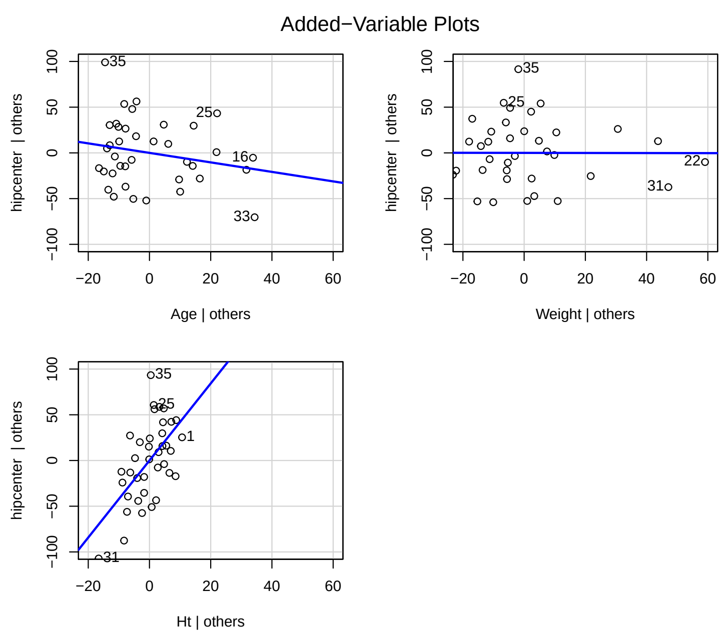
Poglejmo še model, v katerem imamo vključene samo napovedne spremenljivke, katerih vrednosti po navadi poznamo brez dodatnih meritev, to so `Age`, `Weight` in `Ht`.

```
> mod.3<-update(mod.2, .~. -Arm - Thigh, data=seatpos)
> vif(mod.3)

     Age   Weight       Ht
1.093018 3.457681 3.463303

> anova(mod.3, mod.2)

Analysis of Variance Table

Model 1: hipcenter ~ Age + Weight + Ht
Model 2: hipcenter ~ Age + Weight + Ht + Arm + Thigh
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     34 45262
2     32 44266  2    995.88 0.36 0.7005
```

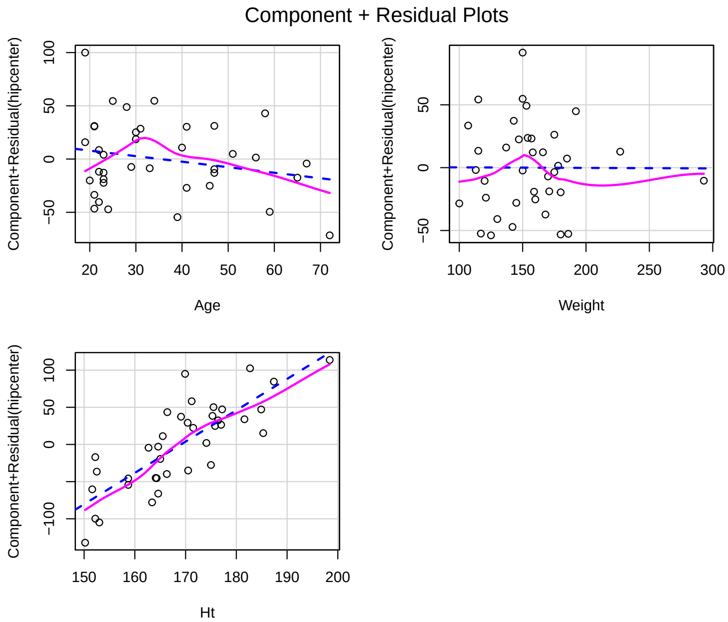
Modela `mod.2` in `mod.3` sta ekvivalentna, zato nadaljujemo z `mod.3`.



Slika 8: Ostanki za `mod.3`



Slika 9: Grafi dodane spremenljivke za `mod.3`



Slika 10: Grafi parcialnih ostankov za `mod.3`

```
> library(multcomp)
> izpis<-glht(mod.3)
> confint(izpis)$confint
                 Estimate          lwr          upr
(Intercept) -5.282977e+02 -861.2375559 -195.3579020
Age         -5.195041e-01   -1.5234899    0.4844817
Weight      -4.270689e-03   -0.7712627    0.7627214
Ht           4.211905e+00    1.7537100    6.6700997
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 2.460517
```

V `mod.3` je samo `Ht` močno statistično značilna napovedna spremenljivka.

Ob upoštevanju starosti in mase voznika je položaj voznikovega sedeža v avtu (`hipcenter`) statistično značilno odvisen samo od telesne višine voznika. Če se `Ht` poveča za 1 cm, se povprečna razdalja med kolki in fiksno točko v avtu (`hipcenter`) poveča za 4.2 mm, 95 % IZ je (1.8 mm, 6.7 mm).

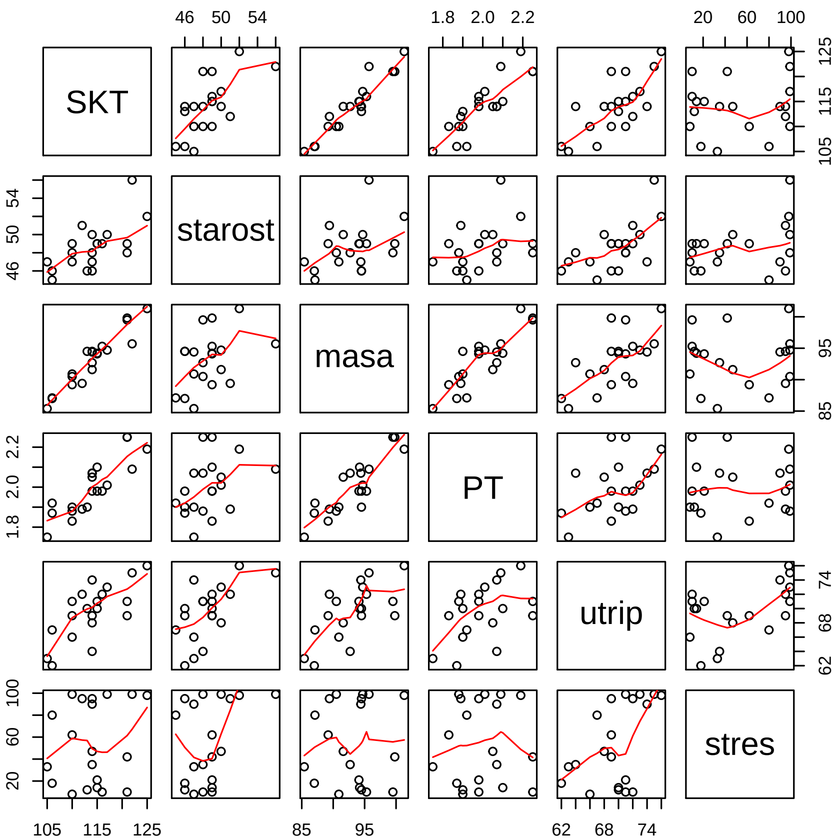
1.2 Primer: `pacienti`

v datoteki `PACIENTI1.txt` so podatki za 20 pacientov s povišanim krvnim tlakom. Za vsakega pacienta so navedene vrednosti naslednjih spremenljivk: zgornji krvni tlak (`SKT`, mm Hg), starost (`starost`, leta), telesna masa (`masa`, kg), površina telesa (`PT`, m^2), bazalni srčni utrip (`utrip`, število utripov na minuto) in stresni indeks (`stres`). Zanima nas odvisnost `SKT` od vseh ostalih spremenljivk.

```
> pacienti<-read.table("PACIENTI1.txt", header=T, sep="\t")
> str(pacienti)

'data.frame':        20 obs. of  7 variables:
 $ zap    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ SKT    : int  105 115 116 117 112 121 121 110 110 114 ...
 $ starost: int  47 49 49 50 51 48 49 47 49 48 ...
 $ masa   : num  85.4 94.2 95.3 94.7 89.4 99.5 99.8 90.9 89.2 92.7 ...
 $ PT     : num  1.75 2.1 1.98 2.01 1.89 2.25 2.25 1.9 1.83 2.07 ...
 $ utrip  : int  63 70 72 73 72 71 69 66 69 64 ...
 $ stres  : int  33 14 10 99 95 10 42 8 62 35 ...

> pacienti$zap<-NULL  # izločimo zaporedno številko pacienta
> summary(pacienti)

      SKT           starost           masa               PT
 Min.   :105.0   Min.   :45.00   Min.   : 85.40   Min.   :1.750
 1st Qu.:110.0   1st Qu.:47.00   1st Qu.: 90.22   1st Qu.:1.897
 Median :114.0   Median :48.50   Median : 94.15   Median :1.980
 Mean   :114.0   Mean   :48.60   Mean   : 93.09   Mean   :1.998
 3rd Qu.:116.2   3rd Qu.:49.25   3rd Qu.: 94.85   3rd Qu.:2.075
 Max.   :125.0   Max.   :56.00   Max.   :101.30   Max.   :2.250
     utrip           stres
 Min.   :62.00   Min.   : 8.00
 1st Qu.:67.75   1st Qu.:17.00
 Median :70.00   Median :44.50
 Mean   :69.60   Mean   :53.35
 3rd Qu.:72.00   3rd Qu.:95.00
 Max.   :76.00   Max.   :99.00
```

Analizirajmo povezanost vseh spremenljivk v podatkovnem okviru `pacienti`. Narišimo najprej matriko razsevnih grafikonov z gladilniki za vse spremenljivke.

```
> pairs(pacienti,panel=panel.smooth)
```



Slika 11: Matrika razsevnih grafikonov za spremenljivke v podatkovnem okviru `pacienti`

```
> round(cor(pacienti, method ="spearman", use="complete"),2)
```

	SKT	starost	masa	PT	utrip	stres
SKT	1.00	0.64	0.93	0.87	0.68	0.14
starost	0.64	1.00	0.41	0.40	0.61	0.38
masa	0.93	0.41	1.00	0.81	0.65	0.07
PT	0.87	0.40	0.81	1.00	0.43	0.04
utrip	0.68	0.61	0.65	0.43	1.00	0.45
stres	0.14	0.38	0.07	0.04	0.45	1.00

Slika 11 in Spearmanovi koeficienti korelacije kažejo, da obstaja močna korelacija med nekaterimi napovednimi spremenljivkami, največja je med `masa` in `PT`. Poglejmo, kako se ta povezanost odraža na VIF vrednostih za posamezne spremenljivke v linearnem modelu, ki opisuje odvisnost `SKT` od navedenih spremenljivk.

Ukaz `vif` iz paketa `car` izračuna VIF ali $GVIF$ za vsako napovedno spremenljivko v modelu.

```
> model.SKT.0<-lm(SKT ~ starost + masa + PT + utrip + stres, data=pacienti)
> library(car)
> vif(model.SKT.0)

 starost     masa       PT    utrip    stres
1.733157 8.415955 5.321477 4.330443 1.815882
```

Najvišjo vrednost VIF ima spremenljivka `masa`, za katero smo že videli, da je tesno korelirana s `PT`. Vemo, da je `PT` izračunana na podlagi `masa` in telesne višine, ki je sicer med podatki ni, po formuli:

$$PT = 0.007184 \cdot visina^{0.725} \cdot masa^{0.425}.$$

Zato se odločimo, da bomo v naslednjem koraku izločili `masa`. Nov model naredimo z ukazom `update`.

```
> model.SKT.1<-update(model.SKT.0,.~.-masa)
> vif(model.SKT.1)

 starost       PT    utrip    stres
1.674407 1.424390 2.268171 1.485726
```

S tem popravkom modela smo se znebili kolinearnosti. Poglejmo, kako so se spremenile ocene parametrov pri napovednih spremenljivkah, ki so ostale v modelu:

```
> compareCoefs(model.SKT.0, model.SKT.1)

Calls:
1: lm(formula = SKT ~ starost + masa + PT + utrip + stres, data = pacienti)
2: lm(formula = SKT ~ starost + PT + utrip + stres, data = pacienti)

             Model 1  Model 2
(Intercept)   -13.52     5.50
SE              2.60     8.97

starost       0.7123   0.5730
SE            0.0509   0.1981

masa          0.9709
SE            0.0653

PT              3.69    24.49
SE              1.63     3.35

utrip        -0.0745   0.4683
SE            0.0529   0.1516

stres        0.00606 -0.01621
SE           0.00351  0.01258
```

Pri modelu `model.SKT.1` so ocene parametrov pri `PT`, `utrip` in `stres` precej drugačne kot v `model.SKT.0`, standardna napaka ocene parametra za `PT` je relativno manjša ($1.63/3.69 = 0.44$, $3.35/24.49 = 0.14$) v primerjavi s tisto v `model.SKT.0`.

```
> library(multcomp)
> summary(glht(model.SKT.1))

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = SKT ~ starost + PT + utrip + stres, data = pacienti)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept) == 0  5.50082    8.96912   0.613   0.9513
starost == 0      0.57301    0.19805   2.893   0.0470 *
PT == 0          24.48547    3.34670   7.316   <0.001 ***
utrip == 0        0.46830    0.15156   3.090   0.0318 *
stres == 0       -0.01621    0.01258  -1.289   0.6068
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

V modelu, ki vsebuje več napovednih spremenljivk, vplive posamezne napovedne spremenljivke obrazložimo ob upoštevanju ostalih spremenljivk v modelu oziroma ob konstantni vrednosti ostalih spremenljivk v modelu. V takem kontekstu so vplivi `starost`, `PT` in `utrip` na `SKT` pozitivni in statistično značilni, vpliv `stres` pa je negativen in statistično neznačilen.

Za oceno pomembnosti posameznih vplivov (velikosti ocen parametrov in pripadajočih intervalov zaupanja) bi potrebovali strokovnjaka s področja medicine. Zavedati se moramo, da je bilo v vzorcu le 20 pacientov in relativno veliko napovednih spremenljivk (4 spremenljivke). Za analizo odvisnosti `SKT` od vseh danih spremenljivk, na podlagi katere bi lahko korektno sklepali na populacijo, bi potrebovali večji vzorec pacientov.

Koliko parametrov je lahko največ v modelu?

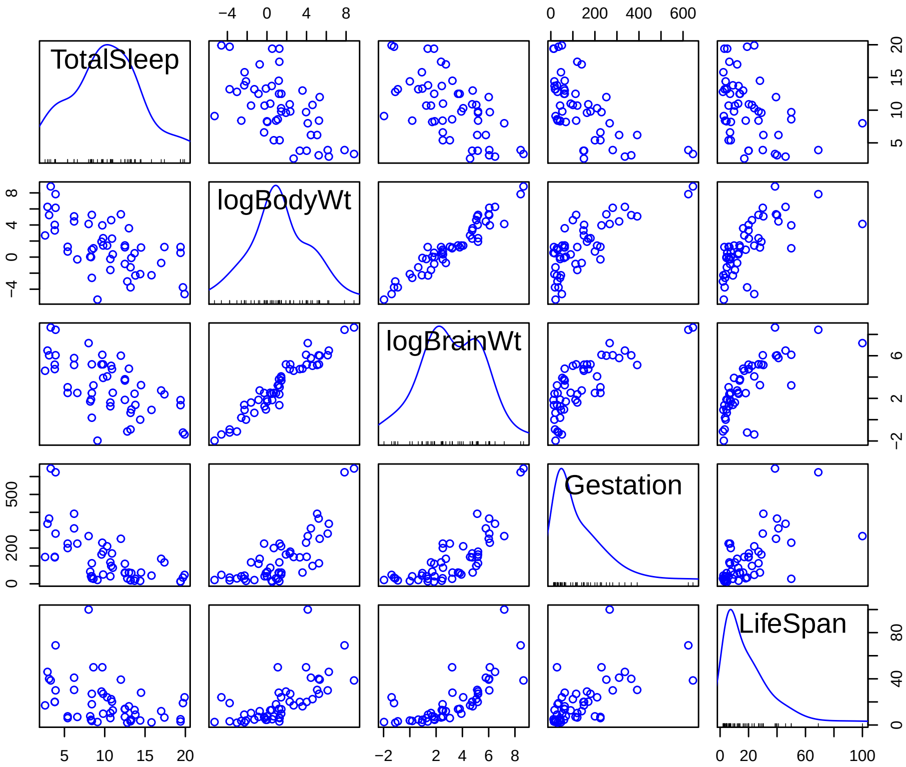
Če je v modelu preveč parametrov, pride do t. i. preprileganja (*overfitting*). To pomeni, da napovedne spremeljivke pojasnijo tudi t. i. slučajno napako, ne samo odvisnost y od napovednih spremenljivk. Pri takem modelu se del slučajne variabilnosti odzivne spremenljivke pripiše napovednim spremenljivkam, posledično je napovedna moč modela slaba. Največje dopustno število parametrov v modelu je vezano na število enot v podatkih.

1.3 Primer: spanje

V datoteki SLEEP.txt (manjkajoči podatki označeni z NA) so podatki za 62 sesalcev. Glej `http://www.statsci.org/data/general/sleep.html`. Delno to informacijo poznamo iz podatkovnega okvira `mammals`. Analizirajmo, kako je `TotalSleep` (h/dan) odvisen od `logBodyWt` (kg), `log(BrainWt)` (g), `Gestation` (dnevi), `LifeSpan` (leta) in `Danger3`. Pričakujemo, da med napovednimi spremenljivkami obstaja povezanost. Najprej izračunamo matriko Spearmanovih korelacijskih koeficientov.

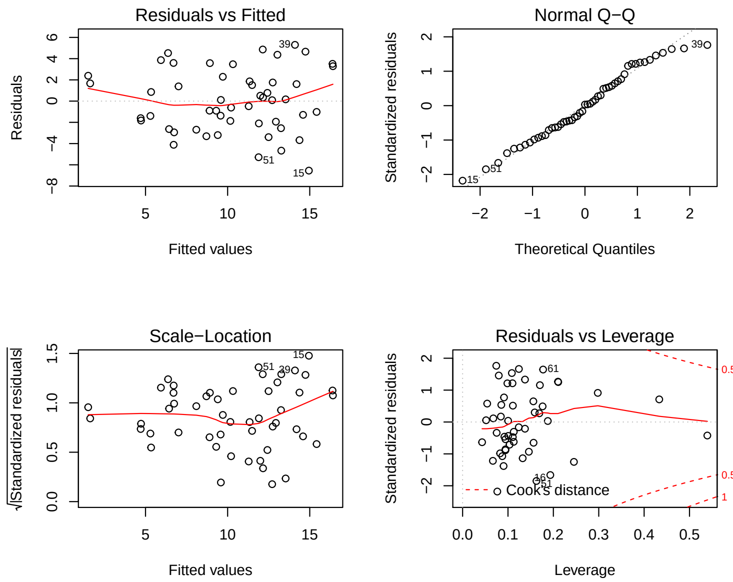
```
> spanje <- read.table("SLEEP.txt",header=TRUE, sep="\t", na.strings="NA")
> str(spanje)

'data.frame':        62 obs. of  7 variables:
 $ Species   : Factor w/ 62 levels "Africanelephant",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ BodyWt    : num  6654 1 3.38 0.92 2547 ...
 $ BrainWt   : num  5712 6.6 44.5 5.7 4603 ...
 $ TotalSleep: num  3.3 8.3 12.5 16.5 3.9 9.8 19.7 6.2 14.5 9.7 ...
 $ LifeSpan  : num  38.6 4.5 14 NA 69 27 19 30.4 28 50 ...
 $ Gestation : num  645 42 60 25 624 180 35 392 63 230 ...
 $ Danger3   : Factor w/ 3 levels "majhna","srednja",..: 2 2 1 2 3 3 1 3 1 1 ...

> spanje$logBodyWt<-log(spanje$BodyWt)
> spanje$logBrainWt<-log(spanje$BrainWt)

> round(cor(spanje[,c("TotalSleep", "logBodyWt","logBrainWt","Gestation","LifeSpan")],
+           use="complete", method = "spearman"), 2)

           TotalSleep logBodyWt logBrainWt Gestation LifeSpan
TotalSleep       1.00     -0.59      -0.62     -0.66    -0.44
logBodyWt       -0.59      1.00       0.95      0.74     0.76
logBrainWt      -0.62      0.95       1.00      0.81     0.83
Gestation       -0.66      0.74       0.81      1.00     0.68
LifeSpan        -0.44      0.76       0.83      0.68     1.00
```

Izrazito močna povezanost obstaja med `logBodyWt` in `logBrainWt` ($r = 0.95$). Tudi ostali korelacijski koeficienti so veliki (Slika 12).

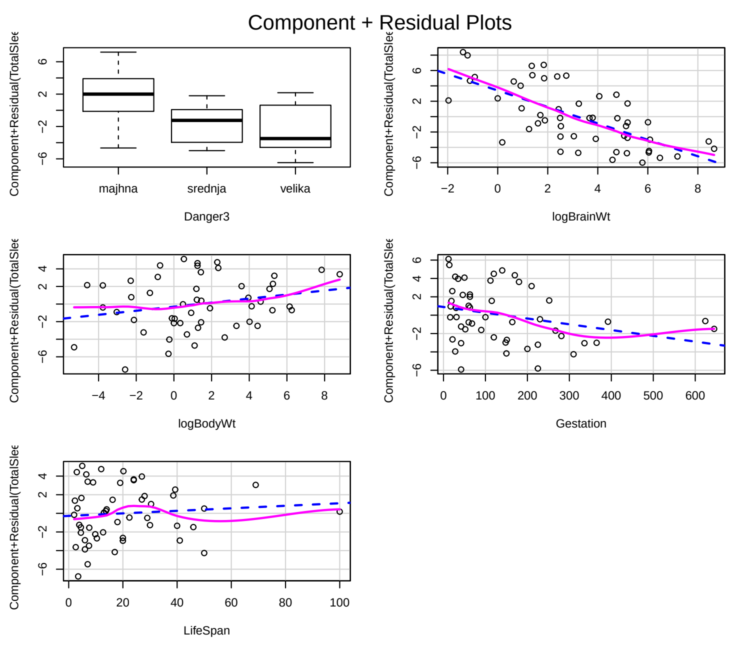


Slika 12: Matrika razsevnih grafikonov za vse številske spremenljivke podatkovnega okvira `spanje`

```
> mod.1 <- lm(TotalSleep ~ Danger3 + logBrainWt + logBodyWt + Gestation + LifeSpan ,
+             data=spanje)
```



Slika 13: Ostanki za `mod.1`



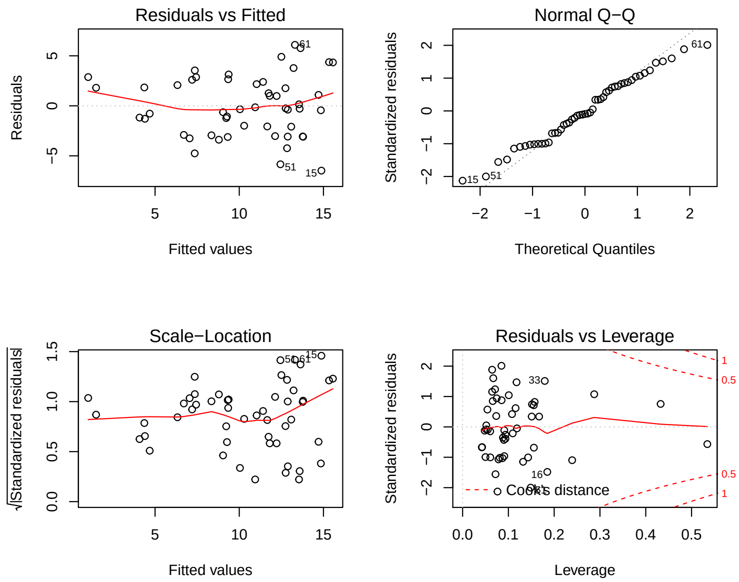
Slika 14: Grafi parcialnih ostankov za `mod.1`

```
> vif(mod.1) # OPOZORILO: funkcija v rnw datoteki ne vrne GVIF!???

Danger3srednja  Danger3velika     logBrainWt      logBodyWt       Gestation
      1.337007       1.292060      16.451325      13.765041        3.138170
      LifeSpan
      2.522080
```

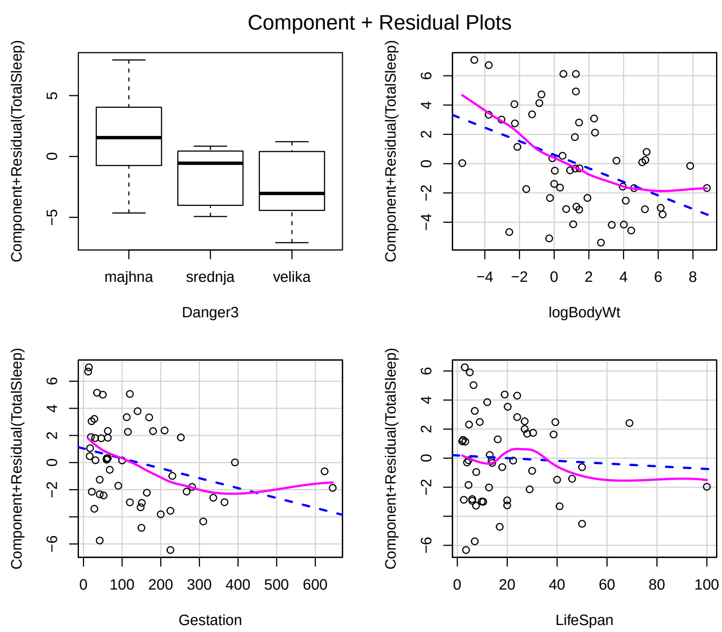
V `mod.1` je prisotna kolinearnost, `logBrainWt` in `logBodyWt` imata zelo visoki vrednosti VIF. Slika 12 kaže njuno tesno povezanost. Ker velja, da so v splošnem laže dostopni podatki za `logBodyWt`, poskusimo iz modela izločiti `logBrainWt`.

```
> mod.2 <- lm(TotalSleep ~ Danger3 +logBodyWt + Gestation + LifeSpan, data=spanje)
> vif(mod.2)  # OPOZORILO: funkcija v rnw datoteki ne vrne GVIF!???

Danger3srednja  Danger3velika      logBodyWt      Gestation        LifeSpan
      1.326074       1.288715       2.864541       3.091764        2.094627

> compareCoefs(mod.1, mod.2)

Calls:
1: lm(formula = TotalSleep ~ Danger3 + logBrainWt + logBodyWt + Gestation +
  LifeSpan, data = spanje)
2: lm(formula = TotalSleep ~ Danger3 + logBodyWt + Gestation + LifeSpan,
  data = spanje)

                Model 1  Model 2
(Intercept)      15.954   14.017
SE                1.533    0.848

Danger3srednja    -3.57    -3.38
SE                 1.39     1.40

Danger3velika     -4.24    -4.16
SE                 1.07     1.08

logBrainWt       -1.069
SE                0.709

logBodyWt         0.230   -0.463
SE                0.516    0.239

Gestation      -0.00629 -0.00729
SE              0.00549  0.00553

LifeSpan        0.01355 -0.00933
SE              0.03685  0.03405
```

Z izločitvijo `logBrainWt` iz modela se predznak ocene parametra spremenljivke `logBodyWt` spremeni, standardna napaka te ocene je manjša kot pri `mod.1`.



Slika 15: Ostanki za `mod.2`



Slika 16: Graf parcialnih ostankov za `mod.2`

Ostanki za `mod.2` (Slika 15) ne kažejo na očitna odstopanja od predpostavk linearnega modela.

Preden se lotimo obrazložitve `mod.2`, ugotovimo, ali ima opisna napovedna spremenljivka `Danger3` statistično značilen vpliv na `TotalSleep`. V ta namen naredimo `mod.2a` brez nje in preverimo ničelno domnevo $H_0 : \beta_1 = \beta_2 = 0$, parametra sta iz modela `mod.2`:

```
> mod.2a <- lm(TotalSleep ~ logBodyWt + Gestation + LifeSpan, data=spanje)
> anova(mod.2a, mod.2)

Analysis of Variance Table

Model 1: TotalSleep ~ logBodyWt + Gestation + LifeSpan
Model 2: TotalSleep ~ Danger3 + logBodyWt + Gestation + LifeSpan
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     47 614.23
2     45 451.37  2    162.86 8.1185 0.0009761 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ničelno domnevo zavrnemo, kar pomeni, da je `Danger3` v modelu potreben. Poglejmo še, ali smo z vključitvijo spremenljivk `LifeSpan` in `Gestation` pojasnili statistično značilno večji del variabilnosti odzivne spremenljivke.

```
> mod.2b <- lm(TotalSleep ~  Danger3 + logBodyWt, data=na.omit(spanje))
> anova(mod.2b, mod.2)

Analysis of Variance Table

Model 1: TotalSleep ~ Danger3 + logBodyWt
Model 2: TotalSleep ~ Danger3 + logBodyWt + Gestation + LifeSpan
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 475.71
2     45 451.37  2    24.344 1.2135 0.3067
```

Modela `mod.2b` in `mod.2` sta ekvivalentna, kar pomeni, da sprejmemo enostavnejšega `mod.2b`, ki mu glede na vajo v prejšnjem poglavju manjka interakcijski člen.

2 VAJE

2.1 Poraba goriva na avtocestah

Raziskovalno vprašanje: kako je poraba goriva na avtocestah odvisna od lastnosti avtomobila?

Raziskovalne domneve: poraba goriva na avtocestah je odvisna od

- tehničnih karakteristik avta (masa, prostornina, moč): večji avti imajo večjo porabo;
- od tipa avta: večji avti imajo večji upor in s tem večjo porabo;
- od porekla avta: avti iz ZDA imajo večjo porabo kor ne-ZDA avti.

Podatki: v paketu MASS je datoteka `Cars93` s karakteristikami avtomobilov, glej `help(Cars93)`.

```
> library(MASS)
> # help(Cars93) ## Data from 93 Cars on Sale in the USA in 1993
> # names(Cars93)
```

Izbrane spremenljivke `MPG.highway, Weight, EngineSize, Horsepower, Type` in `Origin` spremenimo v nam razumljive merske enote in uporabimo slovenska imena spremenljivk.

```
> Cars93$Poraba<-235.21/Cars93$MPG.highway  # v l/100 km
> Cars93$Masa<-Cars93$Weight*0.45359/100    # v 100 kg
> Cars93$Prostornina<-Cars93$EngineSize     # v litih
> Cars93$Moc<-Cars93$Horsepower             # v KM
> Cars93$Poreklo<-Cars93$Origin
> Cars93$Tip<-Cars93$Type
```

Naredimo nov podatkovni okvir `avti` z izbranimi spremenljivkami.

```
> avti <- subset(Cars93, select=c(Poraba, Masa, Prostornina, Moc, Poreklo, Tip))
> rownames(avti)<-Cars93$Make   ### identifikator vozila na slikah
```

Tip `Van` je v več pogledih drugačen od ostalih tipov avtov (večja površina in drugačne lastnosti motorja), vse ostale tipe avtov bi radi primerjali s tipom `Van`, zato ga vzamemo za referenčno skupino.

```
> avti$Tip<-relevel(avti$Tip, ref="Van")
```

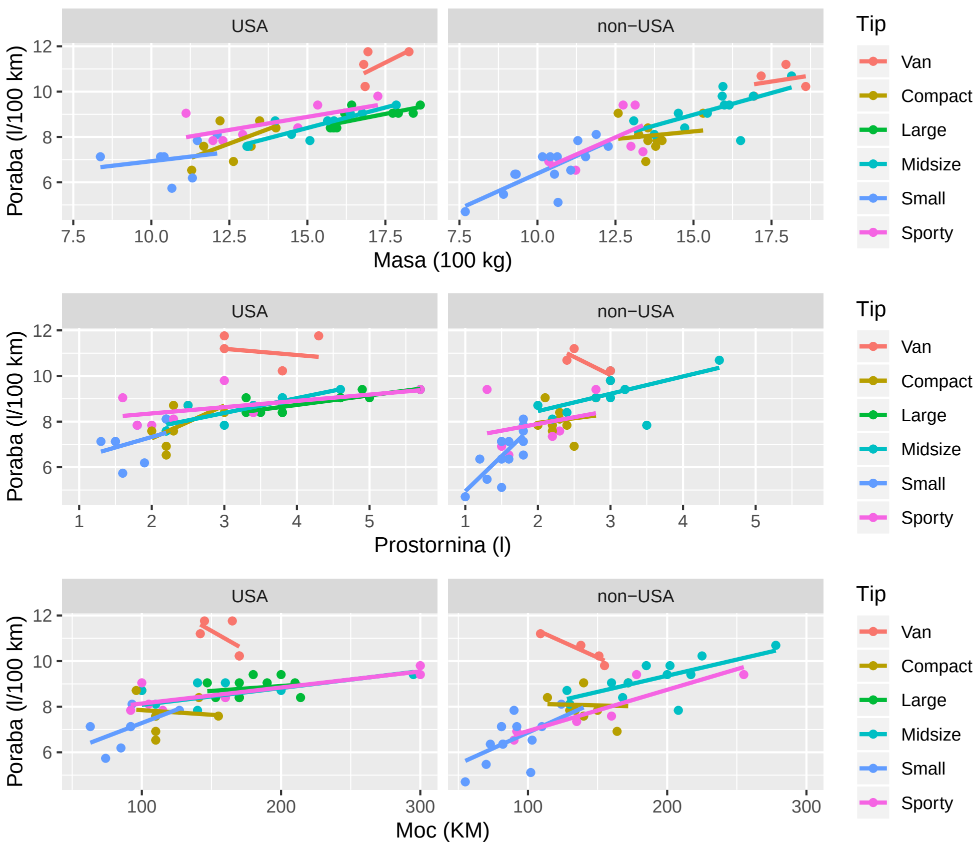
Naredite ustrezni model in ga obrazložite.

Za začetek poglejmo osnovne opisne statistike za analizirane spremenljivke. Za napovedovanje `Poraba` bomo uporabili tri številske spremenljivke (`Masa, Prostornina, Moc`) in dve opisni spremenljivke (`Poreklo`, `Tip`).

```
> summary(avti)

     Poraba            Masa          Prostornina          Moc             Poreklo
 Min.   : 4.704   Min.   : 7.688   Min.   :1.000   Min.   : 55.0   USA    :48
 1st Qu.: 7.587   1st Qu.:11.884   1st Qu.:1.800   1st Qu.:103.0   non-USA:45
 Median : 8.400   Median :13.789   Median :2.400   Median :140.0
 Mean   : 8.330   Mean   :13.938   Mean   :2.668   Mean   :143.8
 3rd Qu.: 9.047   3rd Qu.:15.989   3rd Qu.:3.300   3rd Qu.:170.0
 Max.   :11.761   Max.   :18.620   Max.   :5.700   Max.   :300.0
      Tip
 Van    : 9
 Compact:16
 Large  :11
 Midsize:22
 Small  :21
 Sporty :14
```

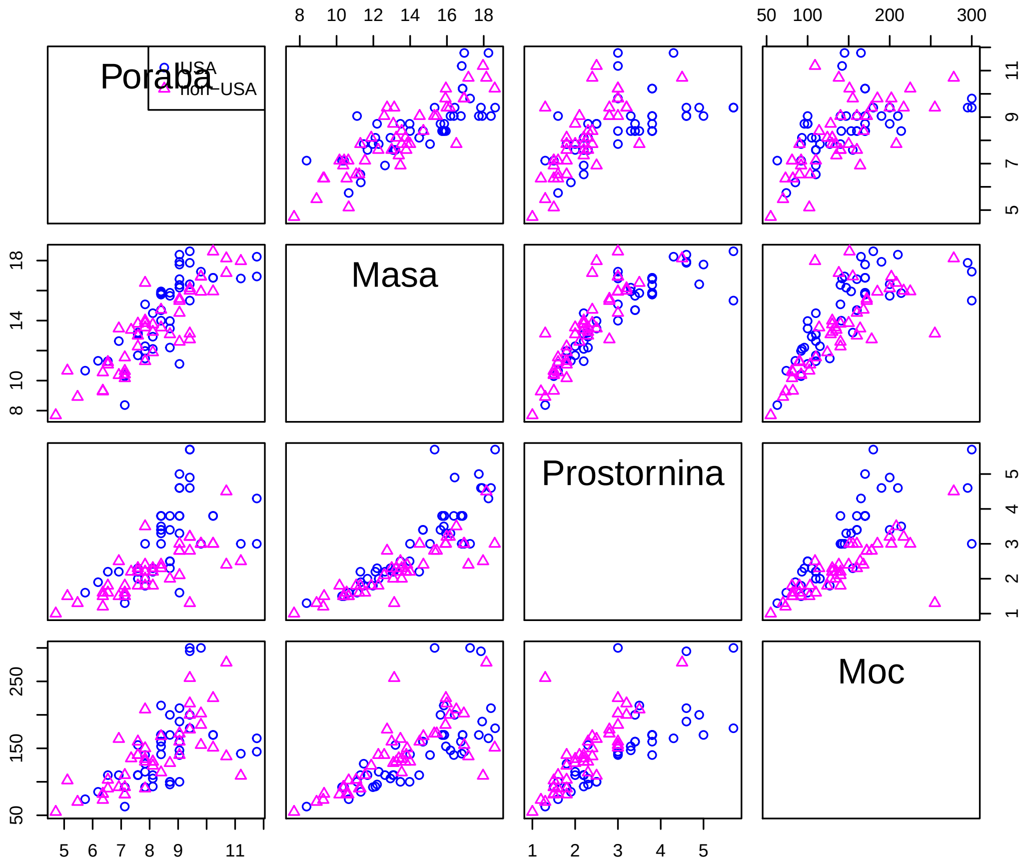
Najprej narišemo nekaj grafičnih prikazov za bivariatno analizo. Narišimo slike, ki kažejo odvisnost `Poraba` od `Masa`, `Poraba` od `Prostornina`, `Poraba` od `Moc` in vsebujejo tudi informacijo o poreklu avtomobila `Poreklo` oziroma o tipu avtomobila `Tip` (Slika 17) in poglejmo povezanost napovednih spremenljivk na podlagi matrike razsevnih grafikonov (Sliki 18 in 19).

```
> library(ggplot2)
> p1<-ggplot(data=avti, aes(x=Masa, y=Poraba, col=Tip)) +
+           facet_grid(.~Poreklo) + geom_point() + geom_smooth(method="lm", se=FALSE) +
+           xlab("Masa (100 kg)") + ylab("Poraba (l/100 km)")
> p2<-ggplot(data=avti, aes(x=Prostornina, y=Poraba, col=Tip)) +
+           facet_grid(.~Poreklo) + geom_point() + geom_smooth(method="lm", se=FALSE) +
+           xlab("Prostornina (l)") + ylab("Poraba (l/100 km)")
> p3<-ggplot(data=avti, aes(x=Moc, y=Poraba, col=Tip)) +
+           facet_grid(.~Poreklo) + geom_point() + geom_smooth(method="lm", se=FALSE) +
+           xlab("Moč (KM)") + ylab("Poraba (l/100 km)")
> library(gridExtra)
> grid.arrange(p1, p2, p3)
```

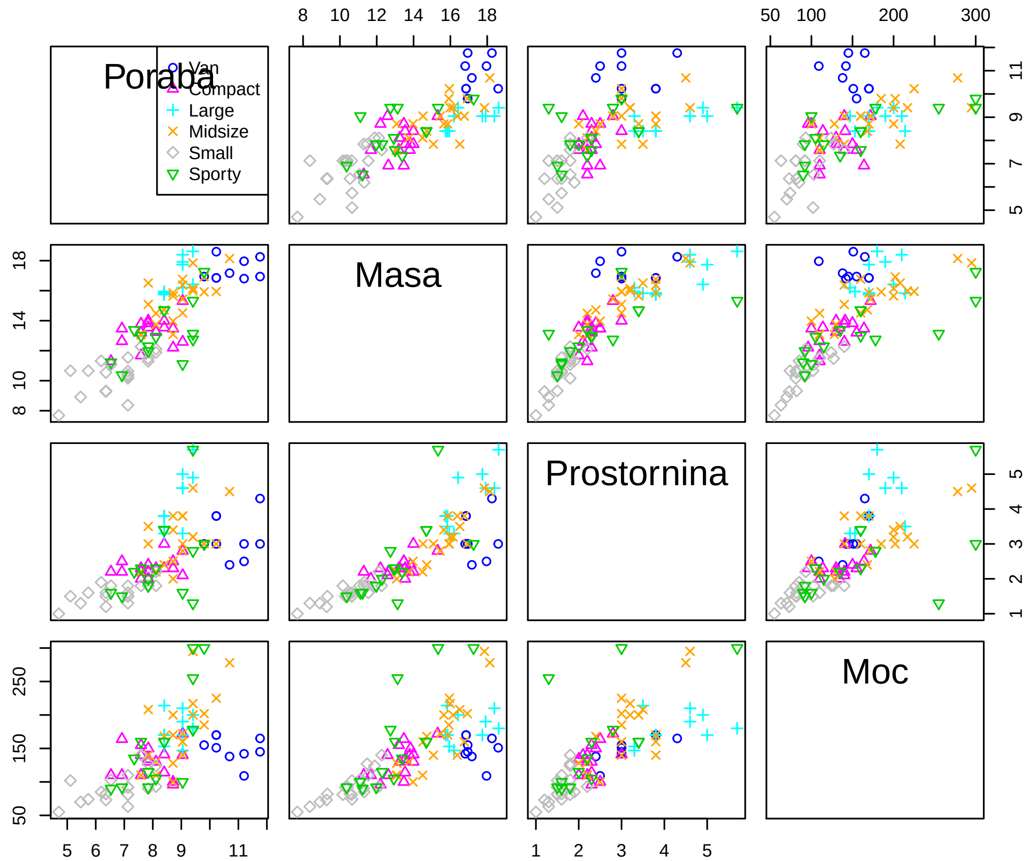


Slika 17: `Poraba (l/100 km)` v odvisnosti od `Masa (100 kg)`, `Prostrornina (l)` in `Moc(KM)`, glede na `Poreklo`

```
> scatterplotMatrix(~Poraba+Masa+Prostornina+Moc|Poreklo, regLine=FALSE,
+                   legend=TRUE, diagonal=FALSE, smooth=FALSE,
+                   data=avti)
```



Slika 18: Matrika razsevnih grafikonov za vse številske spremenljivke podatkovnega okvira `avti` z upoštevanje opisne spremenljivke `Poreklo`



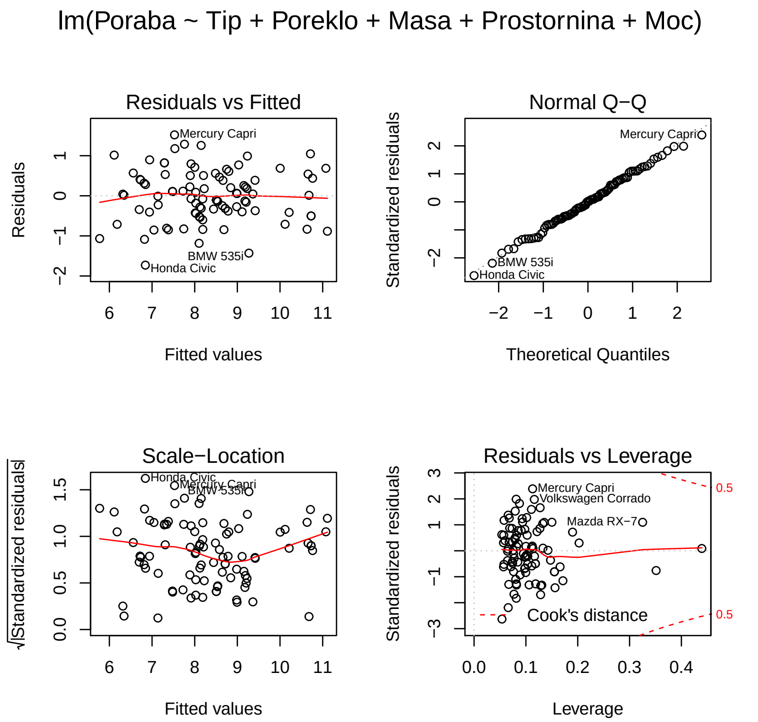
Slika 19: Matrika razsevnih grafikonov za vse številske spremenljivke podatkovnega okvira `avti` z upoštevanjem opisne spremenljivke `Tip`

```
> # matrika korelacijskih koeficientov
> cor(avti[,c("Masa" ,"Prostornina", "Moc")], method="spearman")

                 Masa Prostornina       Moc
Masa        1.0000000   0.8976191 0.8042527
Prostornina 0.8976191   1.0000000 0.8141756
Moc         0.8042527   0.8141756 1.0000000
```

Iz sličic ugotovimo, da je odvisnost `Poraba` od ostalih številskih napovednih spremenljivk po skupinah, ki jih določata opisni spremenljivki, dovolj blizu linearnosti. Številske napovedne spremenljivke so dokaj tesno povezane med seboj, Spearmanov koeficient korelacije je največji med `Masa` in `Prostornina` (0.89). Naredimo model, ki ga določa postavljeno vprašanje in poglejmo vrednosti VIF.

```
> model.0<-lm(Poraba~ Tip + Poreklo + Masa + Prostornina + Moc, data=avti)
> vif(model.0) # ne izpiše se GVIF zaradi težav prevajanja v rnw datoteki

  TipCompact     TipLarge   TipMidsize     TipSmall   TipSporty
    5.026433     2.833733     4.261249    10.372084    5.661495
```

```
Poreklonon-USA          Masa    Prostornina           Moc
      1.455586     12.248412       6.230114      3.904329
```



Slika 20: Ostanki za `model.0`

```
> coef(summary(model.0))

                   Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)     5.225415159 1.225580930  4.2636231 5.291556e-05
TipCompact     -1.587371354 0.416235333 -3.8136391 2.623254e-04
TipLarge       -1.889227420 0.365250285 -5.1724187 1.574995e-06
TipMidsize     -1.446058053 0.340362934 -4.2485768 5.591643e-05
TipSmall       -1.799373842 0.539723668 -3.3338798 1.281509e-03
TipSporty      -1.360076287 0.466232271 -2.9171646 4.542736e-03
Poreklonon-USA -0.128418003 0.169163385 -0.7591359 4.499218e-01
Masa            0.299574441 0.092146649  3.2510617 1.662279e-03
Prostornina    -0.101155561 0.169510338 -0.5967516 5.522975e-01
Moc             0.004940903 0.002657866  1.8589740 6.657484e-02
```

Spremenljivka `Masa` ima vrednost za VIF nad 10. Koeficient korelacije med `Masa` in `Prostornina` je visok. Ker ima v `model.0` spremenljivka `Prostornina` neznačilen vpliv, jo poskusimo prvo izločiti iz modela.

```
> model.1<-update(model.0, .~. - Prostornina)
> vif(model.1)  # ne izpiše se GVIF zaradi težav prevajanja v rnw datoteki
```

```
   TipCompact       TipLarge     TipMidsize       TipSmall      TipSporty
     4.923723       2.397966       4.165348      10.045030       5.573853
Poreklonon-USA          Masa            Moc
     1.206393       9.403582       3.553033
```

```
> summary(model.0)$r.squared
```

```
[1] 0.7868276
```

```
> summary(model.1)$r.squared
```

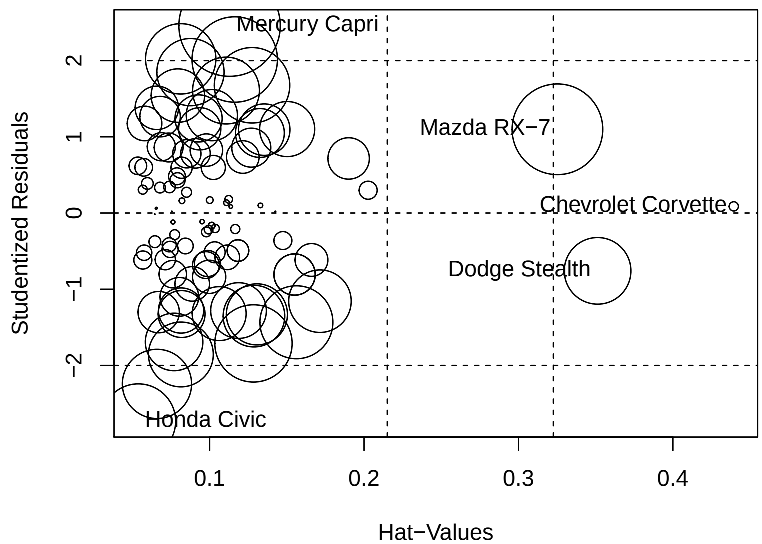
```
[1] 0.7859129
```

```
> compareCoefs(model.0, model.1)
```

```
Calls:
1: lm(formula = Poraba ~ Tip + Poreklo + Masa + Prostornina + Moc, data =
  avti)
2: lm(formula = Poraba ~ Tip + Poreklo + Masa + Moc, data = avti)

               Model 1 Model 2
(Intercept)       5.23    5.41
SE                1.23    1.18

TipCompact      -1.587  -1.623
SE               0.416   0.410

TipLarge        -1.889  -1.975
SE               0.365   0.335

TipMidsize      -1.446  -1.477
SE               0.340   0.335

TipSmall        -1.799  -1.857
SE               0.540   0.529

TipSporty       -1.360  -1.395
SE               0.466   0.461

Poreklonon-USA -0.1284 -0.0866
SE              0.1692  0.1534

Masa            0.2996  0.2731
SE              0.0921  0.0804

Prostornina     -0.101
SE               0.170

Moc            0.00494 0.00447
SE             0.00266 0.00253
```

Modela `model.0` in `model.1` sta glede ocen parametrov in njihovih standardnih napak skoraj enakovredna, kar pomeni, da vpliva kolinearnosti nismo zaznali, hkrati sta ekvivalentna glede pojasnjene variabilnosti, zato obdržimo model `model.0`.

```
> influencePlot(model.0, id=list(n=2))

                        StudRes       Hat       CookD
Chevrolet Corvette  0.08742688 0.43942386 0.000606405
Dodge Stealth      -0.75891280 0.35124505 0.031342800
Honda Civic        -2.73239502 0.05381436 0.039393894
Mazda RX-7          1.09889943 0.32522751 0.058057857
Mercury Capri       2.45956918 0.11292921 0.072596783
```



Slika 21: Grafični prikaz posebnih točk `model.0`

```
> outlierTest(model.0)

No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
             rstudent unadjusted p-value Bonferroni p
Honda Civic -2.732395          0.0076985      0.71596
```

Regresijskih osamelcev ni, vplivnih točk ni.

```
> confint(glht(model.0))

         Simultaneous Confidence Intervals

Fit: lm(formula = Poraba ~ Tip + Poreklo + Masa + Prostornina + Moc,
```

```
    data = avti)

Quantile = 2.7255
95% family-wise confidence level


Linear Hypotheses:
                    Estimate  lwr       upr
(Intercept) == 0     5.225415  1.885114  8.565717
TipCompact == 0     -1.587371 -2.721814 -0.452929
TipLarge == 0       -1.889227 -2.884711 -0.893744
TipMidsize == 0     -1.446058 -2.373712 -0.518404
TipSmall == 0       -1.799374 -3.270382 -0.328365
TipSporty == 0      -1.360076 -2.630785 -0.089368
Poreklonon-USA == 0 -0.128418 -0.589470  0.332634
Masa == 0            0.299574  0.048430  0.550719
Prostornina == 0    -0.101156 -0.563153  0.360842
Moc == 0             0.004941 -0.002303  0.012185
```

Sklepi:

- z modelom je pojasnjene 79 % variabilnosti porabe;
- statistično značilni napovedni spremenljivki v modelu sta `Tip` in `Masa`;
- ob upoštevanju spremenljivk `Poreklo, Tip, Prostornina` in `Moc` v modelu se pri avtu, ki ima 100 kg več, poraba goriva poveča v povprečju za 0.30 l/100 km, pripadajoč 95 % interval zaupanja je (0.05 l/100 km, 0.55 l/100 km);
- ob upoštevanju spremenljivk `Poreklo, Masa, Prostornina` in `Moc` v modelu je poraba goriva pri vseh tipih statistično značilno nižja od porabe v referenčni skupini `Van, USA`. Npr. poraba v skupini `Sporty` je v povprečju za 1.36 l/100 km nižja od porabe v referenčni skupini, pripadajoč 95 % interval zaupanja je 0.09 l /100 km do 2.63 l/100 km; poraba v skupini `Large` je v povprečju za 1.90 l/100 km nižja od porabe v referenčni skupini, pripadajoč 95 % interval zaupanja je 0.90 l /100 km do 2.89 l/100 km. Rezultati pri ostalih tipih so podobni.