1 KOLINEARNOST

Ukvarjamo se s splošnim linearnim modelom s k napovednimi spremenljivkami na podatkih v n točkah

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

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

V (1) je $\mathbf{y}$ vektor odzivne spremenljivke, $\mathbf{X}$ je modelska matrika reda $(n \times k+1)$, $\boldsymbol{\beta}$ je vektor parametrov modela velikosti $(k+1)$ in $\boldsymbol{\varepsilon}$ je vektor napak velikosti (n), za katerega velja $E(\boldsymbol{\varepsilon}) = 0$ in $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $\mathbf{I}$ je enotska diagonalna matrika reda $n \times n$.
Za (1) smo po metodi najmanjših kvadratov dobili rešitev sistema $k+1$ enačb v obliki

$$\boldsymbol{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{2}$$

Napovedi modala so $\hat{\boldsymbol{y}} = \mathbf{X}\boldsymbol{b}$.

Matrika $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ v (2) mora biti obrnljiva oziroma nesingularna, da rešitev obstaja. Če je modelska matrika polnega ranga, $rang(\mathbf{X}) = k+1$, to pomeni, da nobene napovedne spremenljivke ne moremo napisati kot linearne kombinacije ostalih napovednih spremenljivk. Potem lahko pokažemo, da je tudi $rang(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = k+1$ in matrika $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ je nesingularna. V takem primeru pravimo, da gre za model s polnim rangom in ocene parametrov modela so enolično določene.

Če modelska matrika ni polnega ranga, pravimo, da gre v modelu za **popolno kolinearnost regresorjev** in enolična rešitev sistema enačb za $\boldsymbol{b}$ ne obstaja. V takem primeru lahko dobimo iste napovedane vrednosti $\hat{\boldsymbol{y}}$ z različnimi koeficienti linearne kombinacije regresorjev.

Če so nekateri regresorji tesno korelirani med seboj, pravimo, da gre za **kolinearnost** oziroma **multikolinearnost**. Korelirani regresorji v model prispevajo zelo podobno informacijo. V takem primeru ima matrika $\mathbf{X}$ še vedno polni rang vendar so določeni regresorji skoraj linearna kombinacija ostalih regresorjev. V takem modelu majhne spremembe v podatkih povzročijo velike spremembe v ocenah parametrov, saj so te močno odvisne od drugih regresorjev v modelu.

V primeru kolinearnosti različne linearne kombinacije regresorjev dajo zelo podobne napovedane vrednosti. Drugače povedano, v prostoru parametrov je večje območje vrednosti za parametre, ki dajo zelo podobno modelsko napoved in zato je težko natančno oceniti parametre $\boldsymbol{\beta}$.

Do kolinearnosti v modelu lahko pride tudi v situaciji, ko korelacijski koeficienti med pari regresorjev niso veliki, se pa pokaže, da obstaja tesna povezanost dveh regresorjev ob prisotnosti tretjega regresorja v modelu. v takem primeru govorimo o t. i. multipli povezanosti v kateri je ena spremenljivka korelirana z drugo samo ob prisotnosti tretje. Od tu izhaja tudi izraz multikolinearnost.

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_j služi za ugotavljanje prisotnosti kolinearnosti za posamezno številsko napovedno spremenljivko x_j, $j = 1, ..., k$. Spomnimo se, kako je definirana varianca ocen parametrov:

$$Var(\boldsymbol{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}. \tag{3}$$

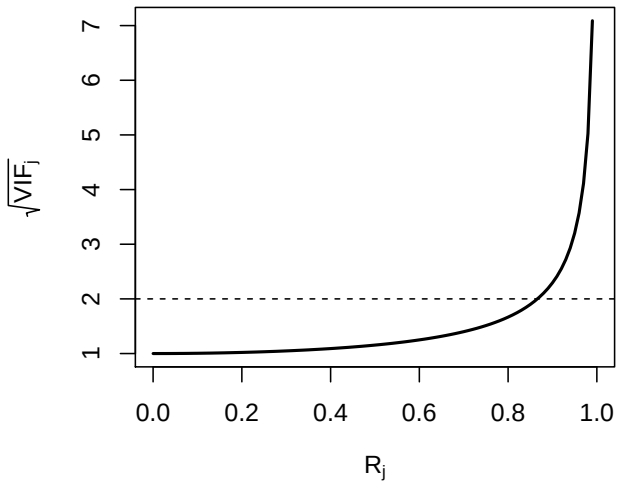
VIF_j temelji na oceni variance za b_j:

$$\widehat{Var}(b_j) = \frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2} \cdot \frac{1}{1 - R_j^2} = \frac{\widehat{\sigma}^2}{SS_{x_j}} \cdot VIF_j. \tag{4}$$

V (4) je $\widehat{\sigma}^2$ z modelom ocenjena varianca napak ($SS_{residuals}/n-k-1$), $SS_{x_j} = \sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2$ 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. Če je koreliranost regresorja x_j z ostalimi regresorji velika, je multipli koeficient korelacije R_j velik in posledično je velika tudi vrednost VIF_j .

V kontekstu standardne napake ocene parametra modela gledamo $\sqrt{VIF_j}$. Ta vrednost pove, 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 $\sqrt{1 - 1/4} = 0.87$. Če je $VIF_j = 9$, kar pomeni trikratno povečanje intervala zaupanja, ima koeficient multiple korelacije vrednost $\sqrt{1 - 1/9} = 0.89$.



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:

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X_1}\boldsymbol{\beta}_1 + \mathbf{X_2}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{5}$$

$\mathbf{y}$ 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 $GVIF_1$:

$$GVIF_1 = \frac{det\mathbf{R}_{11} det\mathbf{R}_{22}}{det\mathbf{R}}, \tag{6}$$

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:

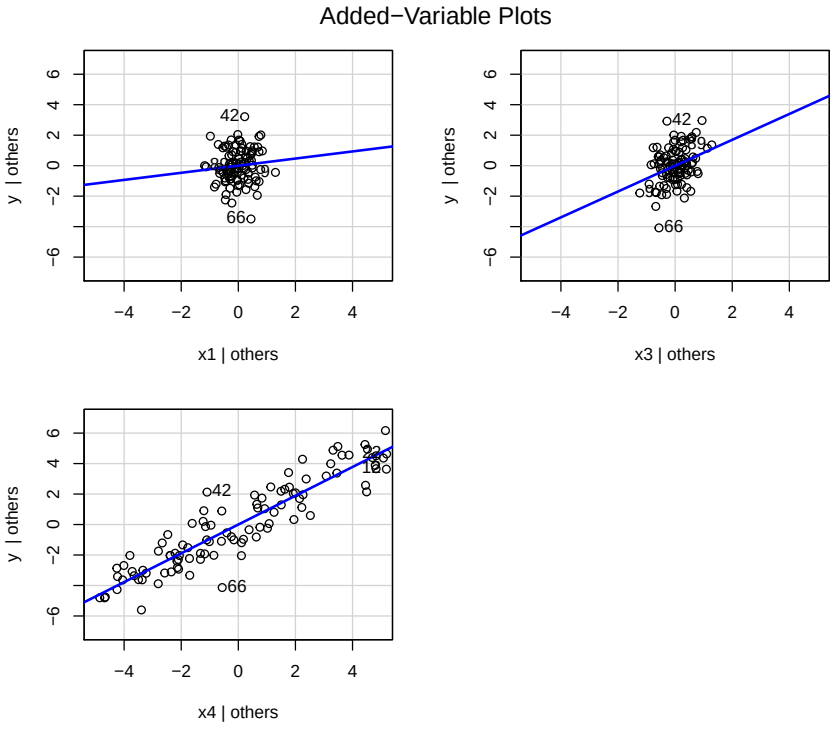
- na podlagi matrike korelacijskih koeficientov in vsebinske presoje izločimo določene napovedne spremenljivke;
- iz več koreliranih napovednih spremenljivk naredimo nove med seboj neodvisne spremenljivke z uporabo metode glavnih komponent (PCA) na napovednih spremenljivkah;
- iz več koreliranih spremenljivk 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 = 0.5)
> 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  0.9880 -0.1101
x2 -1.0000  1.0000 -0.9880  0.1101
x3  0.9880 -0.9880  1.0000 -0.1007
x4 -0.1101  0.1101 -0.1007  1.0000

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

> summary(mod.0)$coef

               Estimate Std. Error     t value     Pr(>|t|)
(Intercept) -0.01598637 0.30486686 -0.05243721 9.582893e-01
x1           0.23349303 0.24857891  0.93931148 3.499282e-01
x3           0.84802112 0.24412830  3.47366993 7.718631e-04
x4           0.94425226 0.04055012 23.28605078 2.844917e-41

> summary(mod.0)$r.squared

[1] 0.9273738

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

[1] 0

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

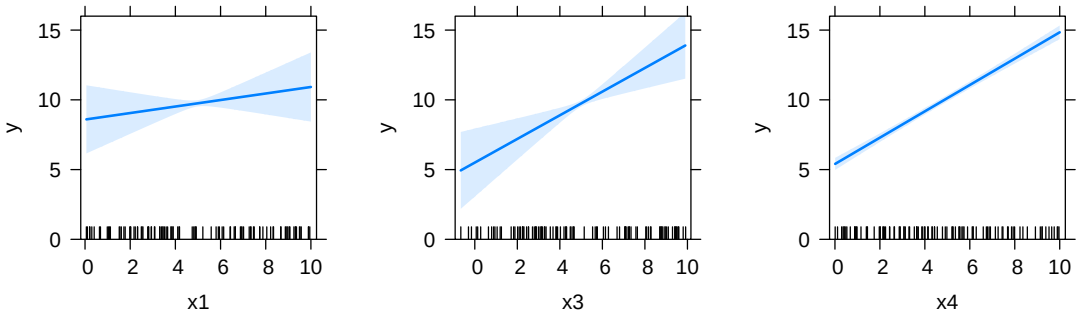
Ilustracija multikolinearnosti:

```
> mod.1<-lm(y~x1+x3+x4)
> vif(mod.1)
       x1        x3        x4
41.989777 41.905678  1.015077
> coef(summary(mod.1))
               Estimate Std. Error     t value     Pr(>|t|)
(Intercept) -0.01598637 0.30486686 -0.05243721 9.582893e-01
x1           0.23349303 0.24857891  0.93931148 3.499282e-01
x3           0.84802112 0.24412830  3.47366993 7.718631e-04
x4           0.94425226 0.04055012 23.28605078 2.844917e-41
> confint(mod.1)
                 2.5 %    97.5 %
(Intercept) -0.6211423 0.5891696
x1          -0.2599322 0.7269183
x3           0.3634303 1.3326120
x4           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))
```



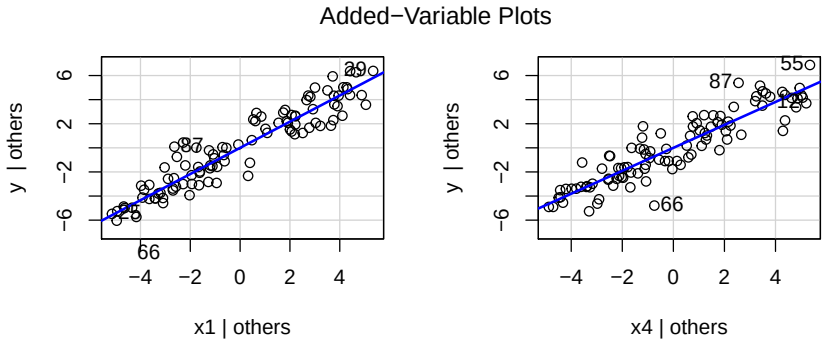
Slika 3: Napovedane vrednosti za y s 95 % intervali zaupanja za povprečno napoved za `mod.1`, pri (`x1` in `x3`) se intervali zaupanja hitro širijo z oddaljenostjo od povprečne vrednosti

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
-4.0818 -0.6179  0.0512  0.6511  2.9623

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.05578    0.32156  -0.173    0.863
x1           1.08650    0.04074  26.670   <2e-16 ***
x4           0.95165    0.04274  22.265   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.202 on 97 degrees of freedom
Multiple R-squared:  0.9182,        Adjusted R-squared:  0.9166
F-statistic: 544.7 on 2 and 97 DF,  p-value: < 2.2e-16
```

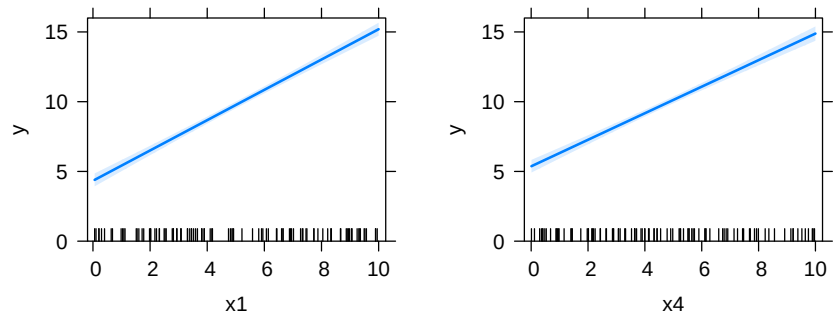
```
> confint(mod.1a)

                 2.5 %    97.5 %
(Intercept) -0.6939884 0.5824292
x1           1.0056484 1.1673563
x4           0.8668220 1.0364828
```

```
> 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`

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.

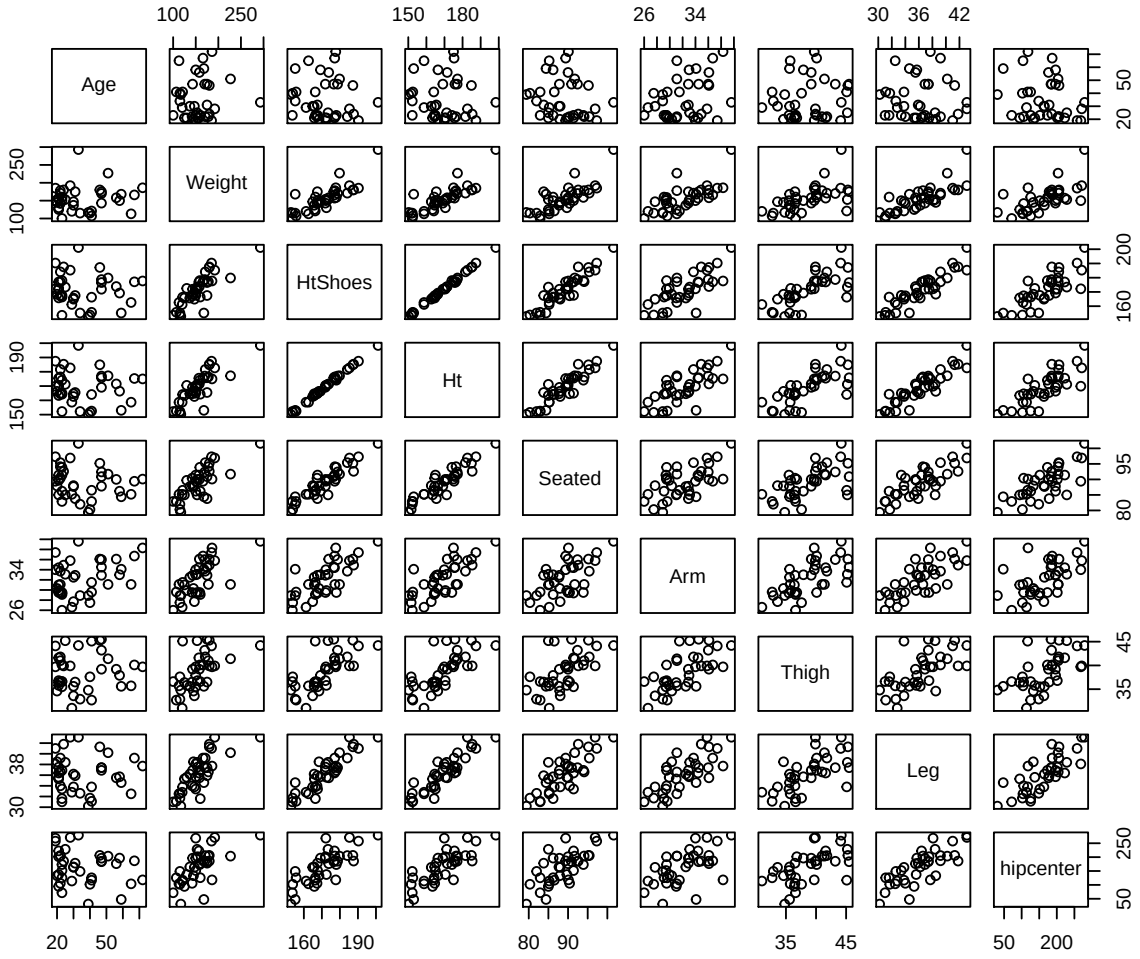
Raziskovalce je zanimala odvisnost `hipcenter` od ostalih spremenljivk. Naredite ustrezni statistični model, izvedite diagnostiko izbranega modela in ga obrazložite.

```
> 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)
> pairs(seatpos)
```



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

```
> round(cor(seatpos, method="spearman"),2)
```

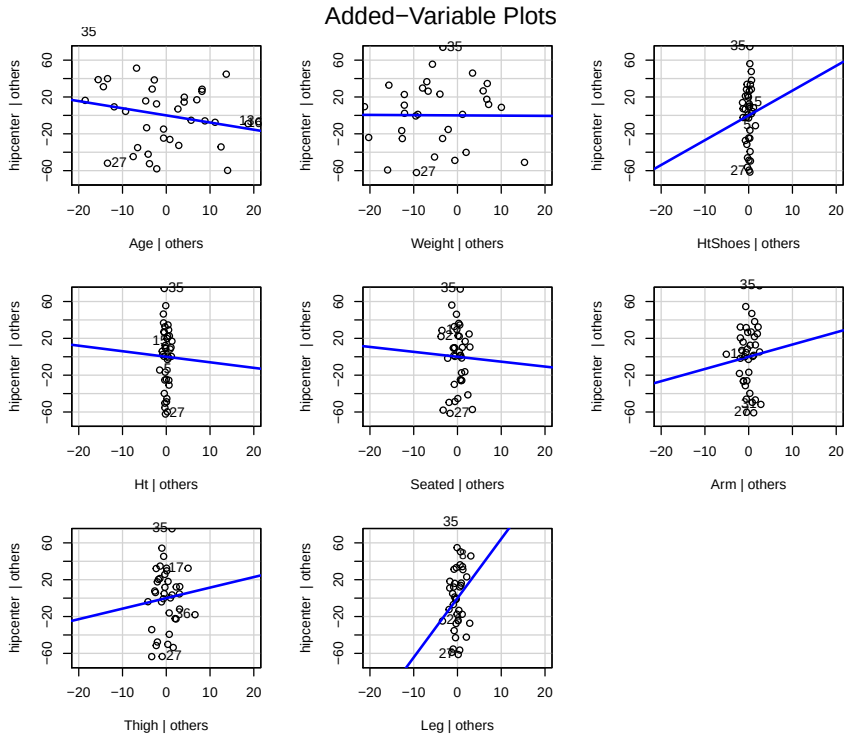
	Age	Weight	HtShoes	Ht	Seated	Arm	Thigh	Leg	hipcenter
Age	1.00	0.07	-0.09	-0.09	-0.21	0.27	0.06	-0.10	-0.19
Weight	0.07	1.00	0.85	0.86	0.76	0.72	0.65	0.79	0.66
HtShoes	-0.09	0.85	1.00	0.99	0.90	0.74	0.77	0.89	0.80
Ht	-0.09	0.86	0.99	1.00	0.90	0.76	0.78	0.90	0.82
Seated	-0.21	0.76	0.90	0.90	1.00	0.56	0.63	0.75	0.68
Arm	0.27	0.72	0.74	0.76	0.56	1.00	0.67	0.74	0.60
Thigh	0.06	0.65	0.77	0.78	0.63	0.67	1.00	0.67	0.66
Leg	-0.10	0.79	0.89	0.90	0.75	0.74	0.67	1.00	0.80
hipcenter	-0.19	0.66	0.80	0.82	0.68	0.60	0.66	0.80	1.00

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	Leg
1.997931	3.647030	307.429378	333.137832	8.951054	4.496368	2.762886	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)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-436.43212823	166.5716187	-2.62008697	0.01384361
Age	-0.77571620	0.5703288	-1.36012113	0.18427175
Weight	-0.02631308	0.3309704	-0.07950283	0.93717877
HtShoes	2.69240774	9.7530351	0.27605845	0.78446097
Ht	-0.60134458	10.1298739	-0.05936348	0.95306980
Seated	-0.53375170	3.7618942	-0.14188376	0.88815293
Arm	1.32806864	3.9001969	0.34051323	0.73592450
Thigh	1.14311888	2.6600237	0.42974011	0.67056106
Leg	6.43904627	4.7138601	1.36598163	0.18244531

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

```
[1] 0.6865535
```

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)$coef

                 Estimate  Std. Error    t value  Pr(>|t|)
(Intercept) -435.80896663 163.9718818 -2.6578274 0.0124859
Age           -0.73677805   0.5440369 -1.3542796 0.1857608
Weight        -0.03278959   0.3250151 -0.1008863 0.9203119
Ht             2.09530039   2.6403640  0.7935650 0.4336809
Seated        -0.40266740   3.6738994 -0.1096022 0.9134548
Arm            1.26841777   3.8337805  0.3308530 0.7430553
Thigh          0.98000131   2.5533222  0.3838142 0.7038230
Leg            6.46851759   4.6395252  1.3942197 0.1734922

> summary(mod.1)$r.squared

[1] 0.6857298
```

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

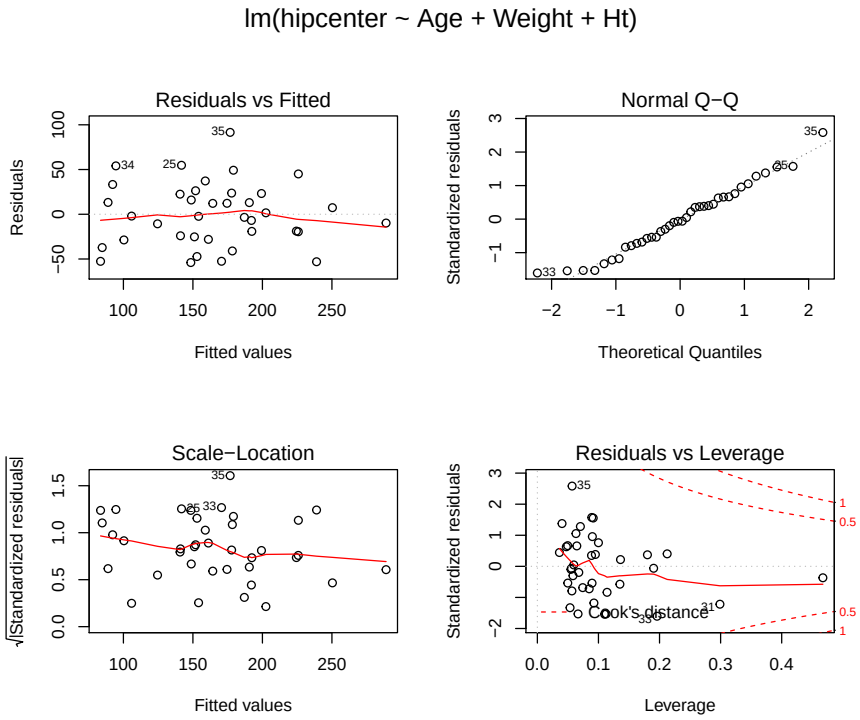
```
> mod.2<-update(mod.1, .~. -Seated -Leg, data=seatpos)
> vif(mod.2)

     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
```

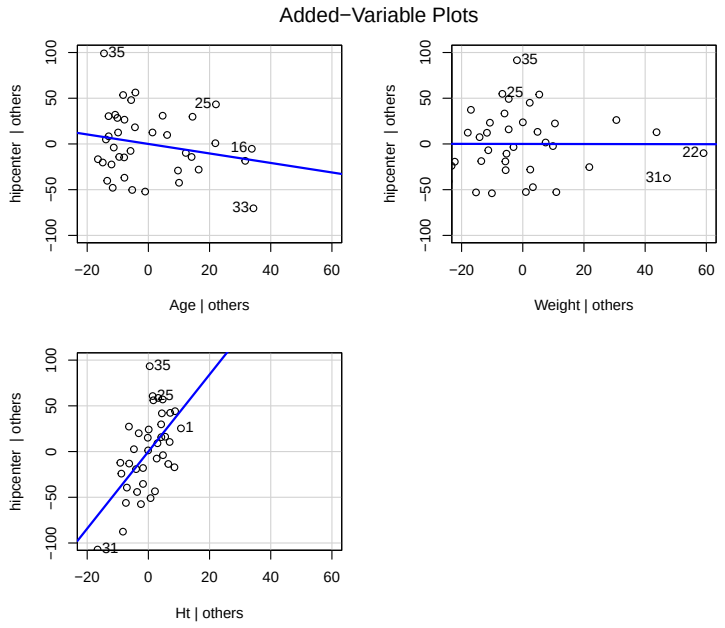
Poglejmo še model, v katerem imamo vključene samo napovedne spremenljivke, katerih vrednosti po navadi poznamo brez dodatnih meritev, to so `Age`, `Weight` in `Ht`.

```
> 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
```

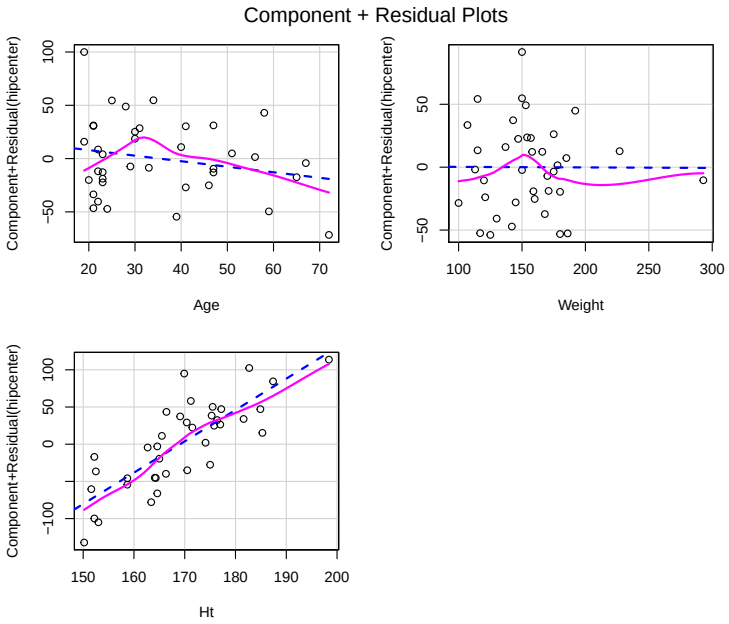
Modela `mod.2` in `mod.3` sta ekvivalentna, zato nadaljujemo z `mod.3`.



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).

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.

Ali lahko iz modela izločimo regresor x_1, če je v modelu prisotna interakcija $x_1 : x_2$?

Če je v modelu interakcija $x_1 : x_2$ značilna, morata ostati v modelu tudi člena x_1 in x_2 ne glede na njuno značilnost; če je v modelu interakcija $x_1 : x_2 : x_3$ značilna, morajo v modelu ostati x_1, x_2 in x_3 in vse njihove dvojne interakcije.

Prav tako velja za člene višjega reda pri polinomski regresiji. Če je značilen kvadratni člen, mora linearni člen ostati v modelu ne glede na njegovo značilnost.