LSM

Bojana Milošević

11/25/2019

Zaključivanje

Žašto su neke tačke proglašene autlajerom?

- došlo je do greške pri uzimanju podataka, pa nam nije mnogo važno da te tačke uklopimo u model
- pretpostavka o modelu je pogrešna

Šta možemo da uradimo?

- izbacimo autlajere
- primenimo neku robusnu metodu za ocenjivanje parametara

Robusna linearna regresija

Zadržavamo pretpostavku o modelu ali parametre modela ne dobijamo metodom najmanjih kvadrata već minimiziranjem

$$\sum_{i=1}^{n} \rho(Y_i - X\beta)$$

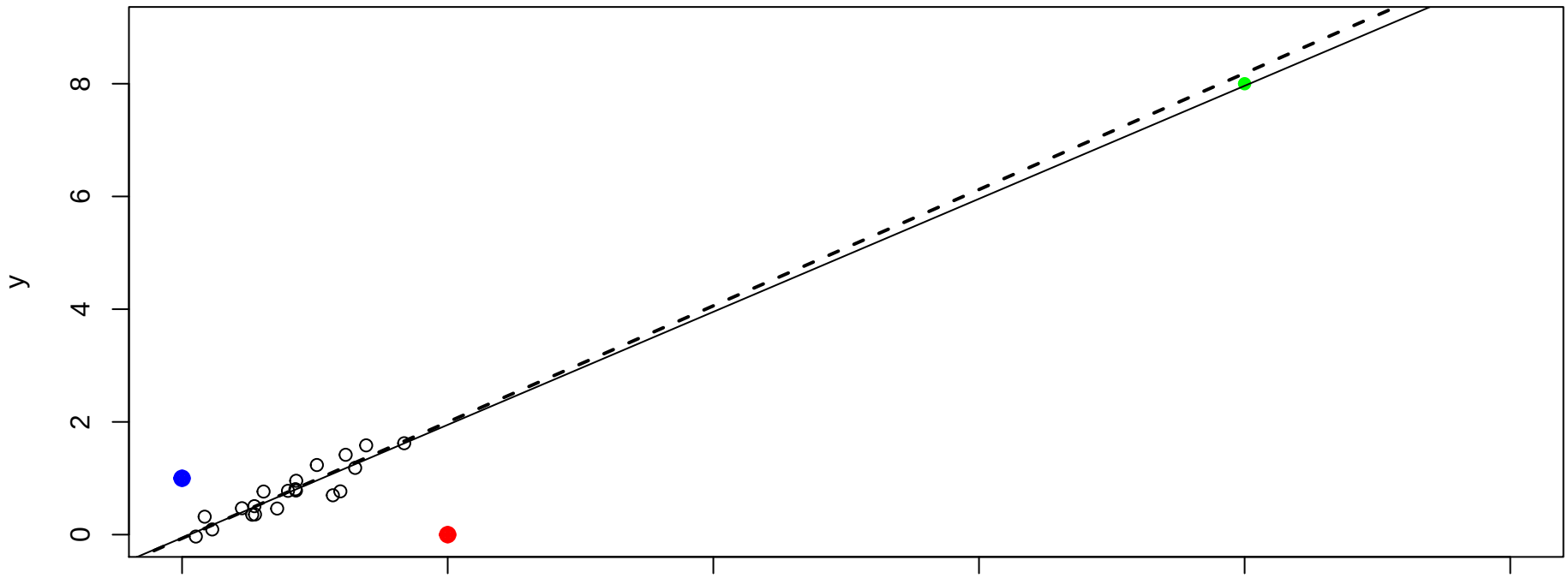
- $\rho(u) = u^2$ LSE (MNK)
- $\rho(u) = |u|$ LAD
- $\rho(u) = \begin{cases} u^2, & |u| < C \\ 2C|u| - C^2, & |u| \geq C \end{cases}$ Hjuberov metod
- $\rho(u) = \begin{cases} u(k - u^2)^2, & |u| < k \\ 0 & |k| \geq C \end{cases}$ Tjukijev metod
- *rlm*() funkcija

Primer

```
set.seed(10)
x <- runif(20)
y <- 2*x + rnorm(n=length(x),mean=0,sd=0.2)
xy <- cbind(x=x,y=y)
outliers <- rbind(c(1,0), c(0,1), c(4,8))
all.points <- data.frame(rbind(outliers,xy))
```

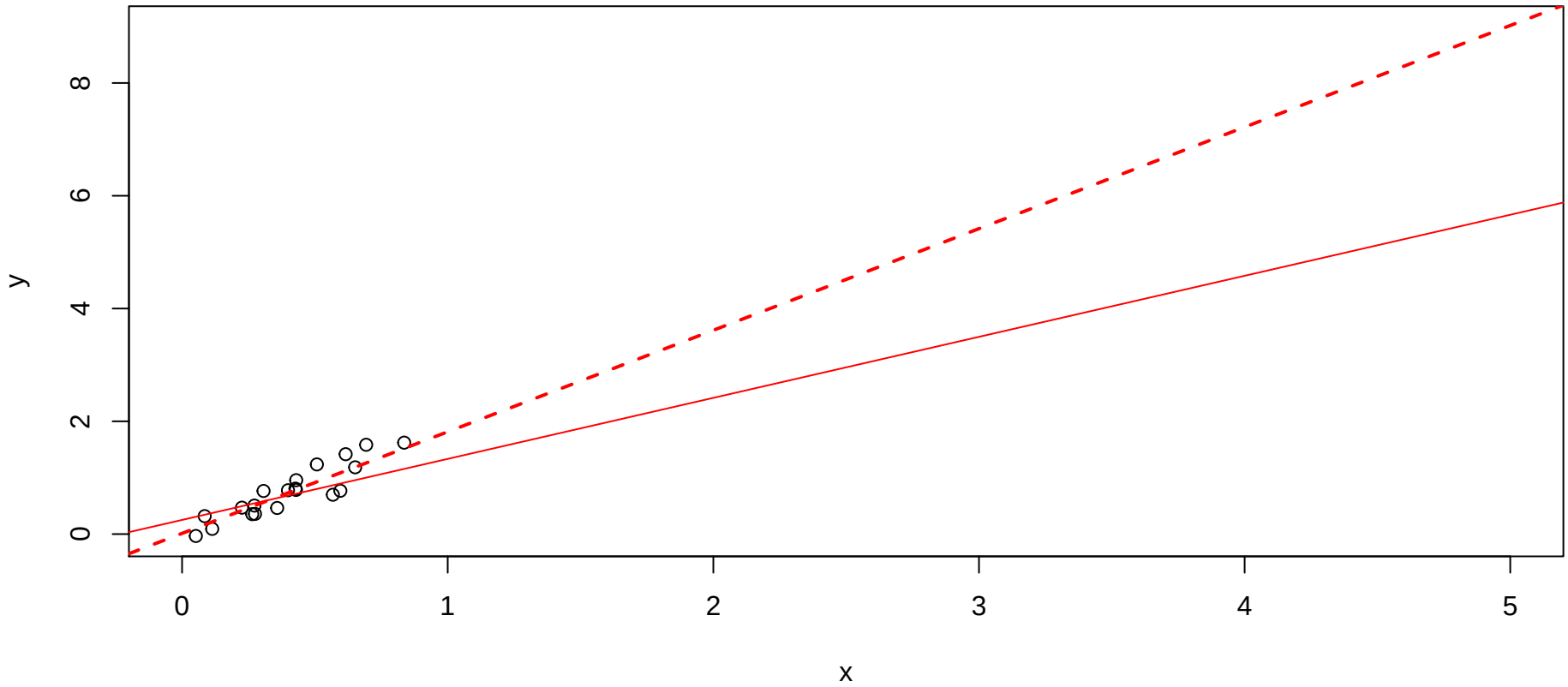
Primer

```
plot(xy, xlim=c(0,5), ylim=c(min(0,min(y)),9))
points(outliers, pch=c(10,10,20), col=c("red","blue","green"),lwd=3)
abline(lm(y~x, data=all.points[-(1:3),]))
library(MASS)
abline(rlm(y~x, data=all.points[-(1:3),]),lty=2,lwd=2)
```



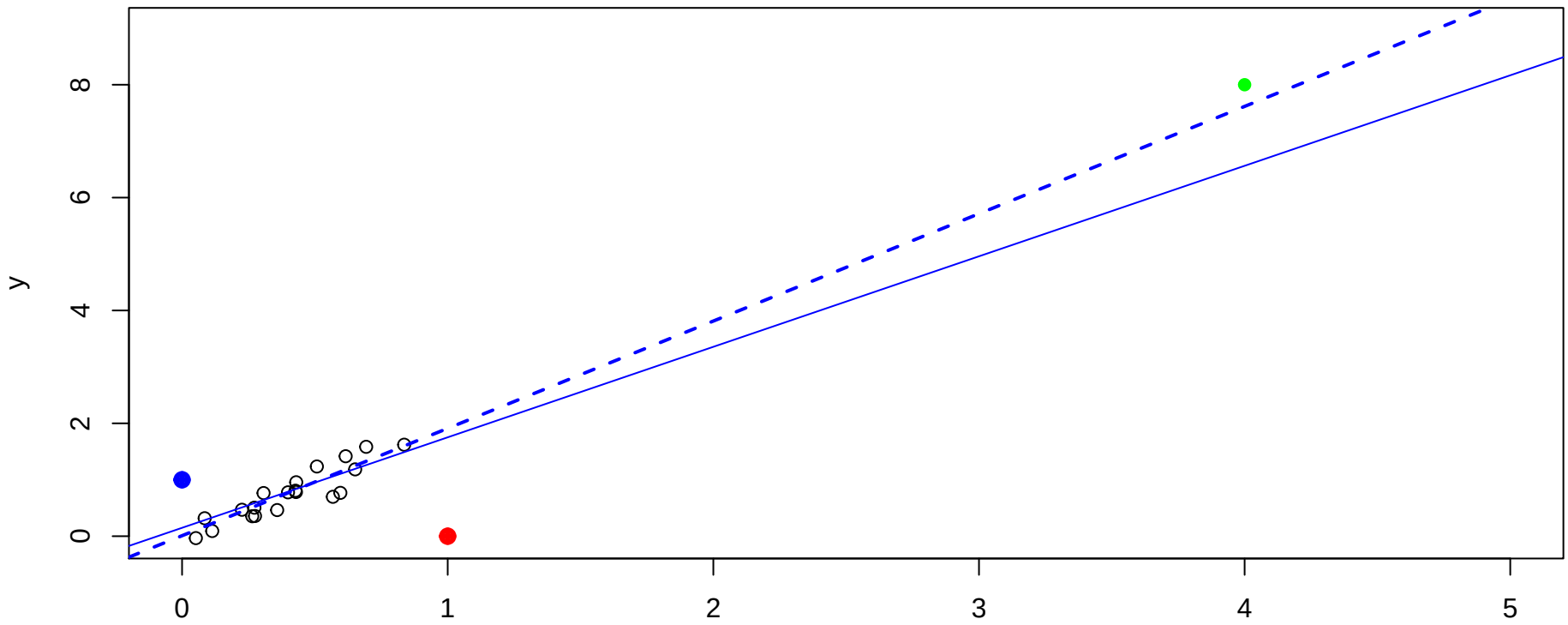
Primer

```
plot(xy, xlim=c(0,5), ylim=c(min(0,min(y)),9))
abline(lm(y~x, data=all.points[-(2:3),]),col="red")
abline(rlm(y~x, data=all.points[-(2:3),]),col="red",lty=2,lwd=2)
```



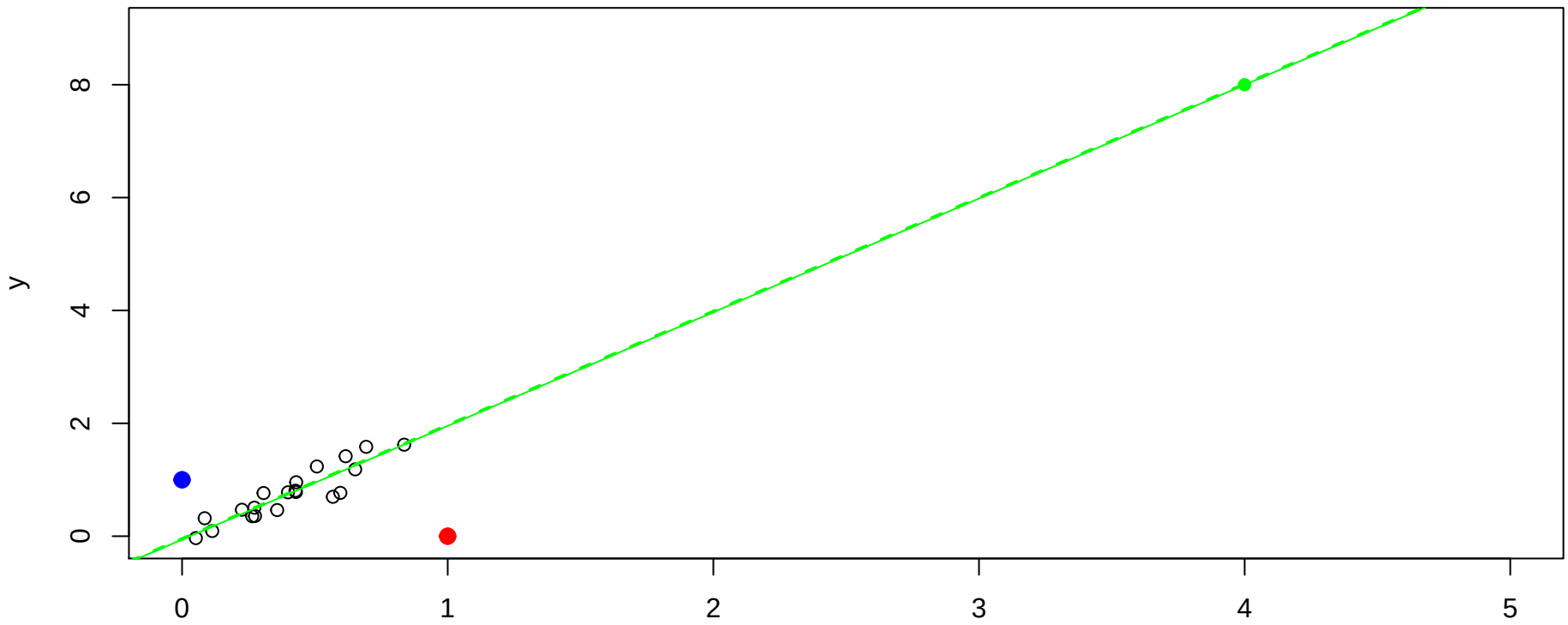
Primer

```
plot(xy, xlim=c(0,5), ylim=c(min(0,min(y)),9))
points(outliers, pch=c(10,10,20), col=c("red","blue","green"),lwd=3)
abline(lm(y~x, data=all.points[-c(1,3),]),col="blue")
abline(rlm(y~x, data=all.points[-c(1,3),]),col="blue",lty=2,lwd=2)
```



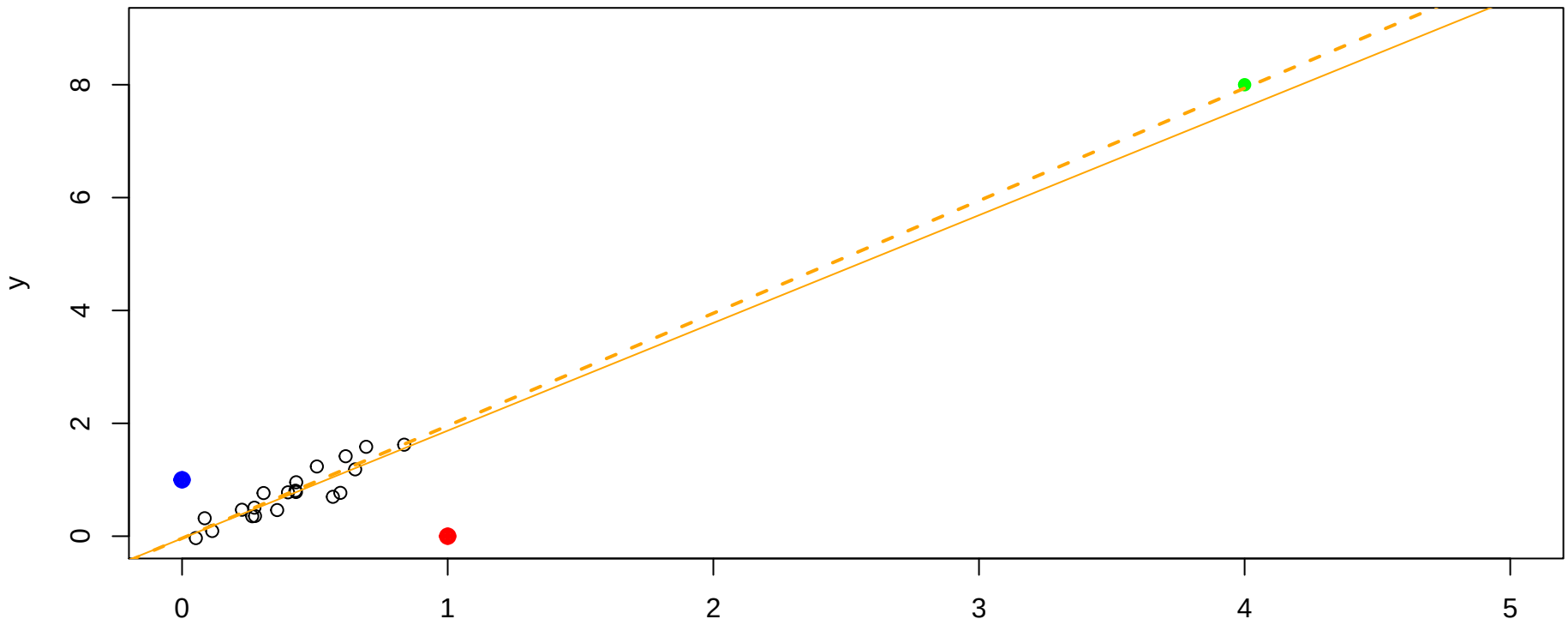
Primer

```
plot(xy, xlim=c(0,5), ylim=c(min(0,min(y)),9))
points(outliers, pch=c(10,10,20), col=c("red","blue","green"),lwd=3)
abline(lm(y~x, data=all.points[-(1:2),]),col="green")
abline(rlm(y~x, data=all.points[-(1:2),]),col="green",lty=2,lwd=2)
```



Primer

```
plot(xy, xlim=c(0,5), ylim=c(min(0,min(y)),9))
points(outliers, pch=c(10,10,20), col=c("red","blue","green"),lwd=3)
abline(lm(y~x, data=all.points),col="orange")
abline(rlm(y~x, data=all.points),col="orange",lty=2,lwd=2)
```



Multikolinearnost

- Ukoliko postoji linearna veza između prediktora matrica $X^T X$ nije ivertibilna
- Ukoliko postoji približno linearna veza između prediktora računanje iverza matrice $X^T X$ nije stabilno

Multikolinearnost

- **Previše prediktora u modelu.** Ovaj problem se javlja često u medicinskim istraživanjima u kojima ima premalo pacijenata u istraživanju.
- **Neprecizna formulacija modela.** Bespotrebno ubacivanje većih stepena prediktora ili sabiraka koji se odnose na njihovu interakciju. Na primer, ukoliko imamo dva prediktora X_1 i X_2 mozhda je X_1X_2 nepotrebno ubaciti u model.
- **Ubacivanje u model prediktora izmedju kojih prirodno postoji linearna veza.** Na primer, ubaciti u model prediktore BRUTO plata, NETO plata i TARA plata.
- **Uzorak na kome se vrše obzervacije je uslovljen nekim ograničenjima u populaciji.** Uzorkovanje vršimo iz "potpopulacije" na kojoj su prediktori veoma korelisani.

Multikolinearnost

Faktor inflacije disperzije $VIF_j = \frac{1}{TOL_j}$, gde je $TOL_j = 1 - R_j^2$ tolerancija, a R_j^2 koeficijent determinacije modela u kome je zavisna promenljiva X_j a nezavisne sve ostale. Jasno je da vrednost VIF_j-a bliska jedinici govori da X_j nije u linearnoj vezi sa ostalim prediktorima. Smatra se da problem multikolinearnosti postoji ukoliko je $VIF_j > 5$.

Može se pokazati da je $D(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2)\sum_{j=1}^{n}(X_j-\bar{X})^2}$,

odnosno dijagonalni elementi matrice $(X^TX)^{-1}$ je $\frac{1}{(1-R_j^2)\sum_{j=1}^{n}(X_j-\bar{X})^2}$

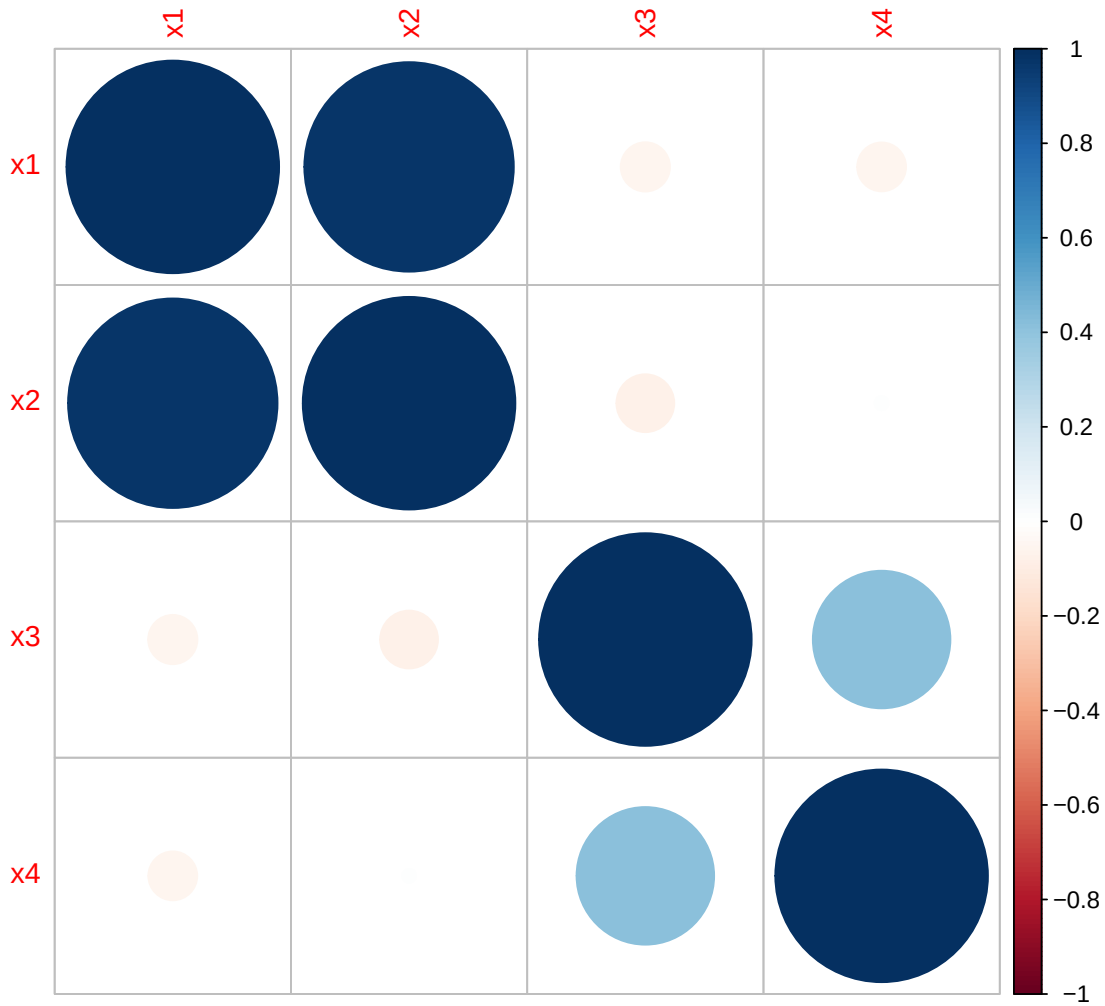
Primer

```
set.seed(10)
x1=runif(50)
x2=x1*0.8+0.2*runif(50)
x3=1+rbeta(50,0.5,0.5)
x4=0.2*x1+x3+rnorm(50,5)
x=cbind(x1,x2,x3,x4)
cor(x)
```

```
##                x1           x2           x3           x4
## x1  1.00000000  0.970587277 -0.05376718 -0.052905524
## x2  0.97058728  1.000000000 -0.07409595  0.004250879
## x3 -0.05376718 -0.074095953  1.00000000  0.418815686
## x4 -0.05290552  0.004250879  0.41881569  1.000000000
```

Primer

```
library("corrplot")
corrplot(cor(x))
```



Primer

VIF

```
library(car)
y=x%*%cbind(c(1,2,3,4))+rnorm(50)
vif(lm(y ~ x1 + x2+x3+x4))
```

```
##        x1        x2        x3        x4
## 19.030242 19.087529  1.271440  1.332006
```

```
vif(lm(y ~ x1 +x3+x4))
```

```
##       x1       x3       x4
## 1.004027 1.214194 1.214082
```

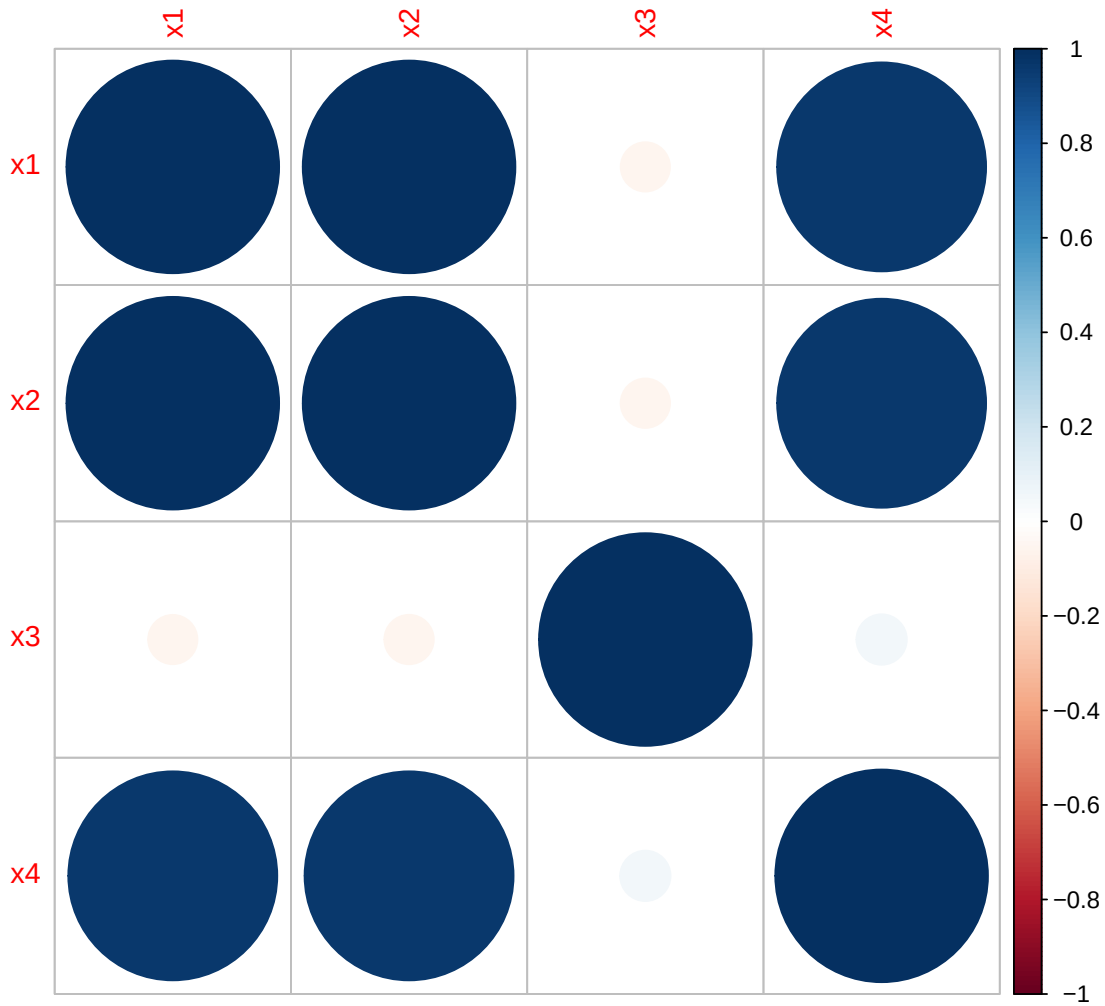
Primer

```
set.seed(10)
x1=runif(50,1,80)
x2=x1*0.8+0.2*runif(50)
x3=1+rbeta(50,0.5,0.5)
x4=0.2*x1+x3+rnorm(50,5)
x=cbind(x1,x2,x3,x4)
cor(x)
```

```
##                x1           x2           x3         x4
## x1  1.00000000  0.99999567 -0.05376718 0.96570882
## x2  0.99999567  1.00000000 -0.05403467 0.96588128
## x3 -0.05376718 -0.05403467  1.00000000 0.05622558
## x4  0.96570882  0.96588128  0.05622558 1.00000000
```

Primer

```
library("corrplot")
corrplot(cor(x))
```



Primer

VIF

```
library(car)
y=x%*%cbind(c(1,2,3,4))+rnorm(50)
vif(lm(y ~ x1 + x2+x3+x4))
```

```
##            x1            x2            x3            x4
## 126942.15662 127844.97767      1.27144     19.70549
```

```
vif(lm(y ~ x3+x4))
```

```
##       x3       x4
## 1.003171 1.003171
```

Otklanjanje problema

- analiza glavnih komponenti
- korišćenje metoda regularizacije
  - nazubljena regresja
  - LASSO regresija

Analiza glavnih komponenti (PCA)

$Z = XA$ linearna transformacija prediktora, odnosno

$$Z_k = a_{1k}X_1 + a_{2k}X_2 + \cdots + a_{pk}X_p$$

$$D(Z_k) = a_k^T \Sigma a_k$$

Bez umanjenja opštosti možemo pretpostaviti da je maksimalna disperzija, uz uslov da je $|a_i| = 1$ baš $D(Z_1)$. Z_1 ćemo zvati prva glavna komponenta. Neka je a_2 vektor za koji je $|a_2| = 1$, Xa_2 je ortogonalno sa Z_1 i $a_2^T \Sigma a_2$ je maksimalno moguće. $Z_2 = Xa_2$ zvaćemo drugom glavnom komponentom. Postupak ponavljamo, pri čemu je svaka od narednih glavnih komponenti ortogonalna na sve prethodne.

Analiza glavnih komponenti

Lema

Neka je Σ kovarijaciona matrica slučajnog vektora X. Neka su $\lambda_1 \geq \lambda_2 \cdot \geq \lambda_p \geq 0$ njene sopstvene vrednosti. Tada je i-ta glavna komponenta data sa $Z_i = v_i^T X$, za $i = 1, 2, ..., p$, gde je v_i i-ti sopstveni vektor.

Primetimo da je tada $D(Z_i) = \lambda_i$, kao i da je za $i \neq j$ Z_i ortogonalno na Z_j, $(Cov(Z) = V^T \Sigma V = Diag(\lambda_1, ..., \lambda_p) = D_\lambda)$.

Analiza glavnih komponenti

Lema

Neka je $Z = XV$. Tada je

$$\sum_{i=1}^{p} D(X_i) = \sum_{i=1}^{p} D(Z_i) = \sum_{i=1}^{p} \lambda_i.$$

Udeo objašnjenog variabiliteta i-tom glavnom komponentom je $\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ i smatra se da treba zadržati bar 80%

Analiza glavnih komponenti

Lema

Koeficijent korelacije između Z_i i X_k je

$$\rho_{Z_i,X_k} = \frac{v_{ki}\sqrt{\lambda_i}}{\sqrt{D(X_k)}}.$$

Dalje, kako je $\Sigma = VDV^T$ zaključujemo i da je $D(X_k) = \sum_{i=1}^{p} \lambda_i v_{ki}^2$

Analiza glavnih komponenti

Standardizovan model $Y = X\delta + \varepsilon$ se može prikazati u obliku $Y = Z\eta + \varepsilon$ gde su Z glavne komponente dobijene od standardizovanih prediktora.

Tada je ocena nepoznatog parametra η dobijena metodom najmanjih kvadrata data sa

$$\hat{\eta} = (Z^T Z)^{-1} Z^T Y = \frac{1}{n-1} D_\lambda^{-1} Z^T Y$$

Iskoristili smo da je uzoračka kovarijaciona matrica $\Sigma = \frac{1}{n-1} X^T X$ i $D_\lambda = \frac{1}{n-1} Z^T Z$.

Važi i

$$D(\eta_j) = \frac{\sigma^2}{(n-1)\lambda_j}$$

Odavde direktno vidimo zašto će prisustvo multikolinearnosti uticati na loše zaključivanje o značajnosti koeficijenata u modelu

Analiza glavnih komponenti

Koje komponente zadržati?

- objašnjeno 80% varijabiliteta
- zadržati samo dve komponente radi grafičkog prikaza
- $\eta_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}$

Analiza glavnih komponenti

Pretpostavimo da smo r glavnih komponenti odlučili da zadržimo, a preostalih $p - r$ da izbacimo.

$$Z = \begin{pmatrix} Z_{(r)} & Z_{(p-r)} \end{pmatrix}$$

$$\eta = (\eta_{(r)}^T \eta_{(n-r)}^T)^T$$

Kako je $Z = XV$ zaključujemo da je $Z_{(r)} = XV_{(r)}$ ($V_{(r)}$ je matrica koja se sastoji od prvih r kolona matrice V, a $V_{(p-r)}$ matrica koja se sastoji od preostalih kolona). Novi model se može prikazati u obliku

$$Y = Z_{(r)}\eta_{(r)} + \tilde{\varepsilon} = Z\eta_r + \tilde{\varepsilon},$$

gde je $\eta_r = \begin{pmatrix} \eta_{(r)} \\ 0 \end{pmatrix}$.

Analiza glavnih komponenti

$$Y = Z_{(r)}\eta_{(r)} + \tilde{\varepsilon} = Z\eta_r + \tilde{\varepsilon},$$

gde je $\eta_r = \begin{pmatrix} \eta_{(r)} \\ 0 \end{pmatrix}$

$$Y = X\delta_r + \tilde{\varepsilon},$$

gde je $\delta_r = V\eta_r$.

Pristrasnost ocene

Nepoznat parametar δ ocenićemo sa $V\hat{\eta}_r = V_{(r)}\hat{\eta}_{(r)}$. Ispitajmo nepristrasnost ocene.

$$E(V_{(r)}\hat{\eta}_{(r)}) = V_{(r)}E(\hat{\eta}_{(r)}) = V_{(r)}\eta_{(r)} = V\eta_r = \begin{pmatrix} V_{(r)} & V_{(p-r)} \end{pmatrix} \begin{pmatrix} \eta_{(r)} \\ \mathbf{0} \end{pmatrix}$$

$$= \begin{pmatrix} V_{(r)} & V_{(p-r)} \end{pmatrix} \begin{pmatrix} V_{(r)} & \mathbf{0} \end{pmatrix}^T \delta = \begin{pmatrix} I_{(r)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \delta$$

$$= \delta - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{(p-r)} \end{pmatrix} \delta \neq \delta.$$

Kovarijaciona matrica ocene

Neka je $\tilde{\delta}$ ocena MNK za δ.

$$Cov(\tilde{\delta}) = \frac{1}{n-1}\sigma^2 V D^{-1} V^T = V \begin{pmatrix} D_r^{-1} & 0 \\ 0 & D_{p-r}^{-1} \end{pmatrix} V^T \frac{\sigma^2}{n-1}$$

$$= \frac{\sigma^2}{n-1}(V_{(r)} D_r^{-1} V_{(r)}^T + V_{(p-r)} D_{p-r}^{-1} V_{(p-r)}^T)$$

Prvi sabirak predstavlja kovarijaciju ocenjenih parametara na osnovu r zadržanih glavnih komponenti a ostatak, deo koji je nestao eliminacijom komponenti koje su bile "višak".

Primer

```
xc=scale(x)
PmodelCentrirano=lm(y~xc)
res.pca <- prcomp(xc, scale = TRUE)
res.pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.71899663 1.00394190 0.19273644 0.00198115
##
## Rotation (n x k) = (4 x 4):
##            PC1         PC2         PC3            PC4
## x1 -0.57971066 -0.02850685 -0.40607668 -0.7058502919
## x2 -0.57974547 -0.02875653 -0.40162134  0.7083562299
## x3  0.01550668  0.99556010 -0.09284073  0.0004686316
## x4 -0.57235503  0.08497352  0.81558729 -0.0025685243
```

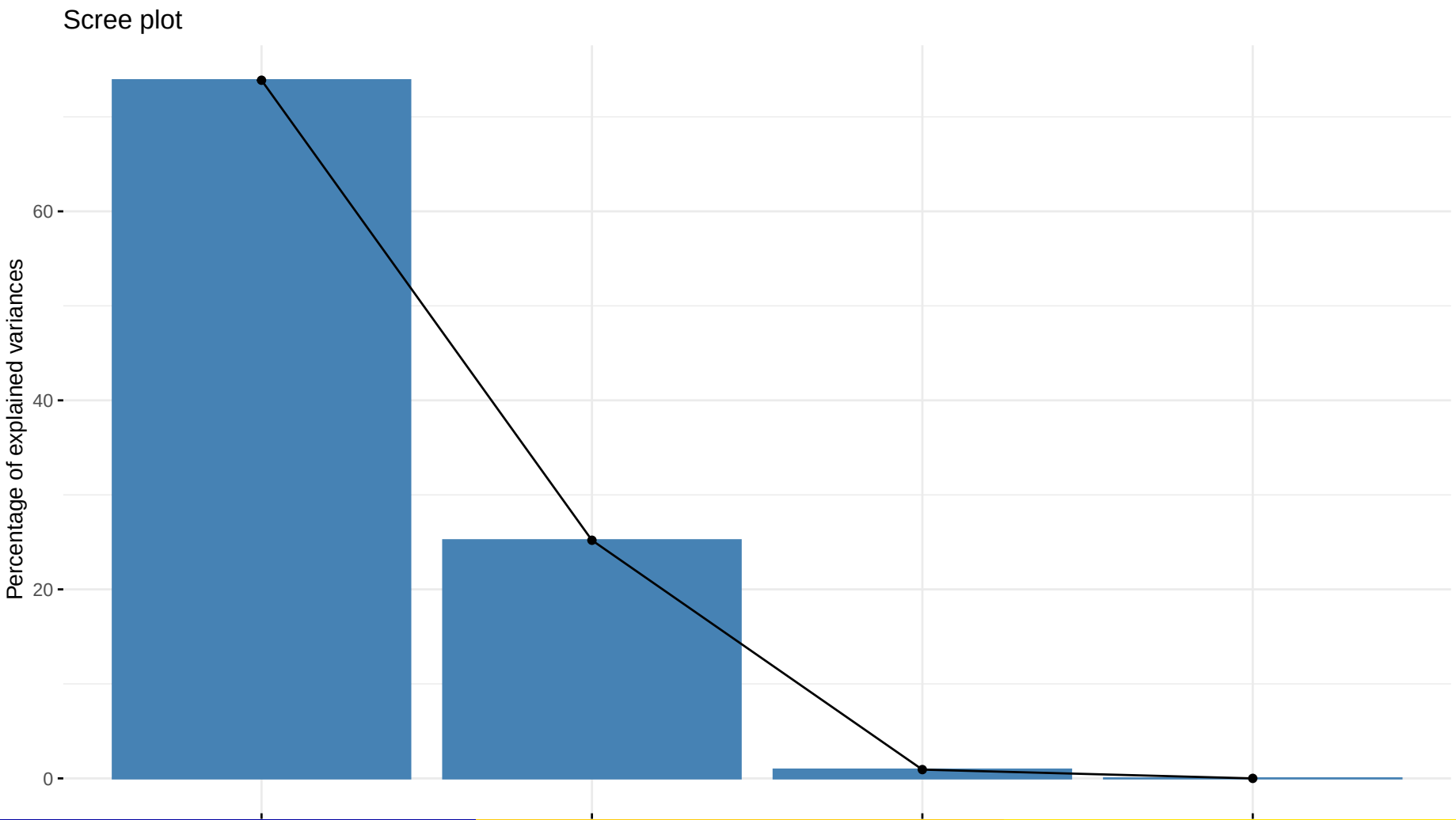
Primer

```
summary(res.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3      PC4
## Standard deviation     1.7190 1.0039 0.19274 0.001981
## Proportion of Variance 0.7387 0.2520 0.00929 0.000000
## Cumulative Proportion  0.7387 0.9907 1.00000 1.000000
```

Primer

```
library("factoextra")
fviz_eig(res.pca)
```



Primer

RMSE:

- bez izbacivanja
- bez poslednje komponente
- bez poslednje dve komponente

Primer

```
Pmodel=lm(y~x)
rmse<-function (x, y) sqrt (mean ( (x-y)^2) )
rmse(y,Pmodel$fitted)
```

```
## [1] 0.9713976
```

```
xnovo=res.pca$x
fitPC1=lm(y~xnovo[,1:3])
rmse(y,fitPC1$fitted)
```

```
## [1] 0.9715441
```

```
fitPC2=lm(y~xnovo[,1:2])
rmse(y,fitPC2$fitted)
```

```
## [1] 1.944205
```

```
Pmodel1=lm(y~x1+x3+x4)
rmse(y,Pmodel1$fitted)
```

```
## [1] 0.9731919
```