SEMINARSKI RAD

ZADATAK 53

15.12.2018.

STATISTIČKI PRAKTIKUM 1

MARIJA BABIĆ

Članak "Changes in Growth Hormone Status Related to Body Weight of Growing Cattle" (*Growth*, 1977., str. 241-247) proučava vezu između težine tjela x i određenog svojstva metabolizma y(=metabolic clearance rate/body weight). Podaci o mjerenjima se nalaze u datoteci zad53r.dat (DEVORE, JAY L., *Probability and Statistics for Engineering and the Science*, 1982., Brooks/Cole Publishing Company, Monterey, California, str. 468).

a) Prikažite podatke (x,y) u Kartezijevom koordinatnom sustavu. Zatim isto napravite za slijedeće transformacije originalnih podataka:

(1) $(x',y')=(x,ln(y))$
(2) $(x',y')=(ln(x),ln(y))$
(3) $(x',y')=(ln(x),y)$
(4) $(x',y')=(1/x,ln(y))$
(5) $(x',y')=(1/x,y)$

Da li je moguće pretpostaviti linearnu zavisnost između podataka u nekom od ovih (šest) slučajeva?

Dobila sam 28 podataka, od kojih 14 predstavlja težinu tijela i spremljeni su u varijablu x, a 14 ih predstavlja određeno svojstvo metabolizma i spremljeni su u varijablu y:

x: 110 110 110 230 230 230 360 360 360 360 505 505 505 505

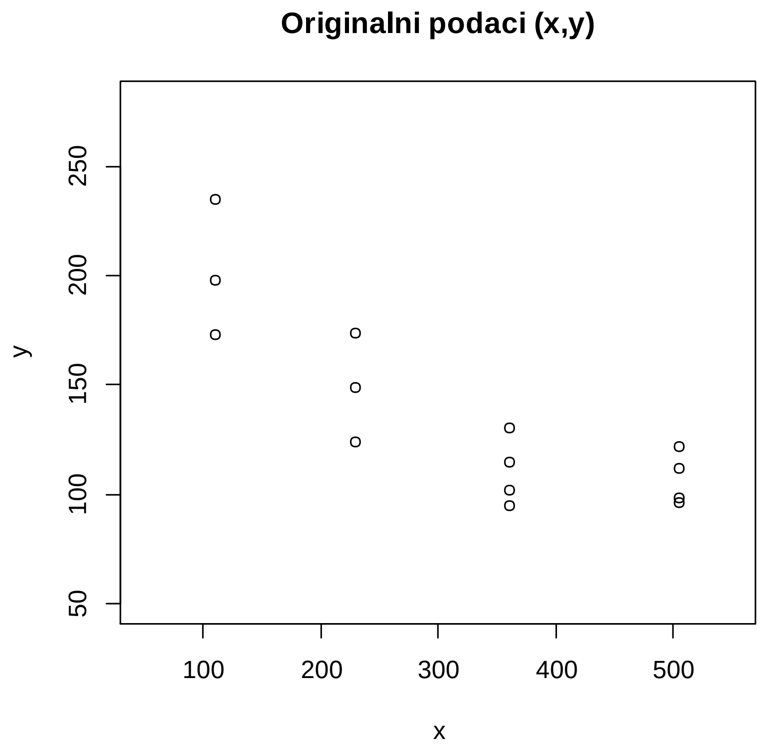
y: 235 198 173 174 149 124 115 130 102 95 122 112 98 96

```
> podaci<-read.table("zad53r.dat",skip=2)
> x<-podaci[,1]
> y<-podaci[,2]
> n<-length(y)
```

Nakon što sam učitala podatke i spremila ih u varijable x i y, prikazala sam svih 6 zadanih modela(originalni model i 5 transformiranih) u Kartezijevom koordinatnom sustavu:

(0) (x,y) - originalni podaci

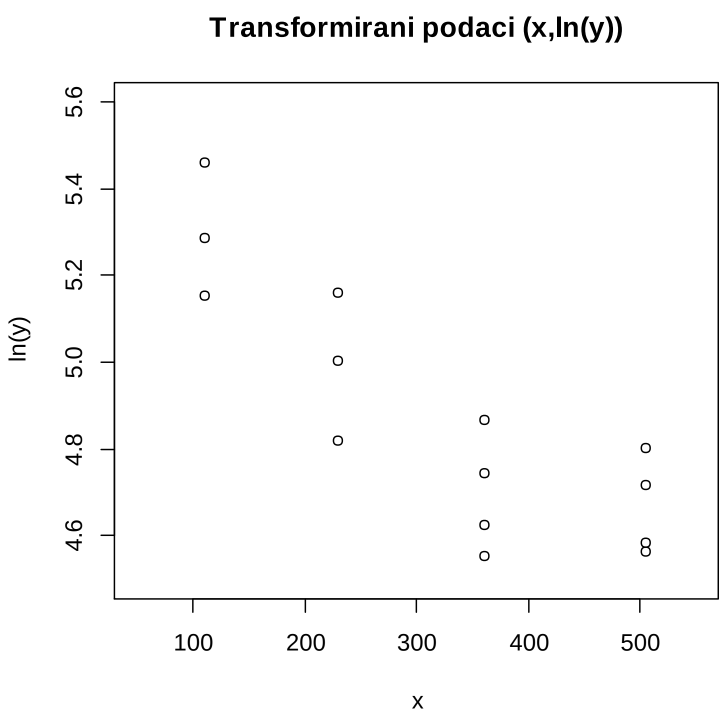
```
>plot(x,y,xlab="x",ylab="y",main="Originalni podaci (x,y)",xlim=c(50,550),ylim=c(50,280))
```



(1) $(x',y')=(x,ln(y))$

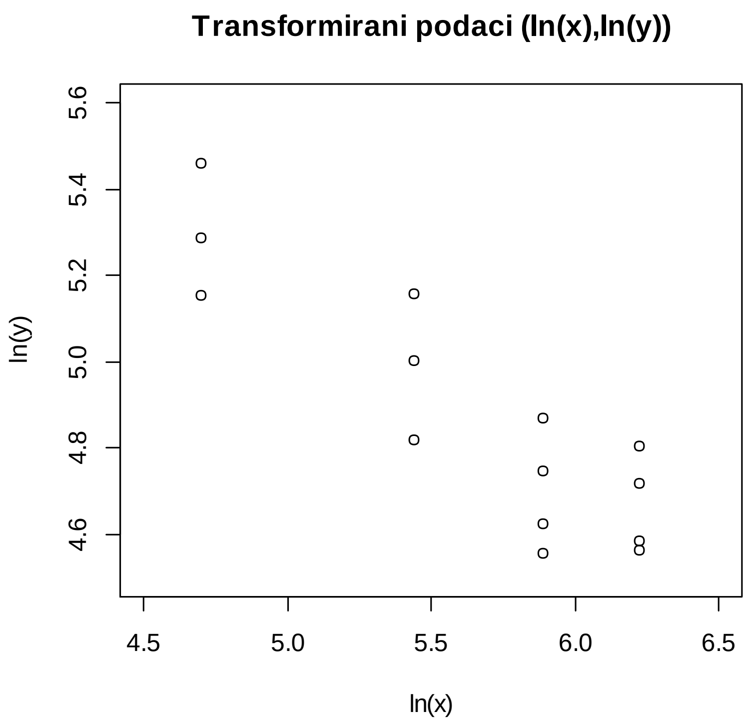
> yn<-log(y)

> plot(x,yn,xlab="x",ylab="ln(y)",main="Transformirani podaci (x,ln(y))",xlim=c(50,550),ylim=c(4.5,5.6))

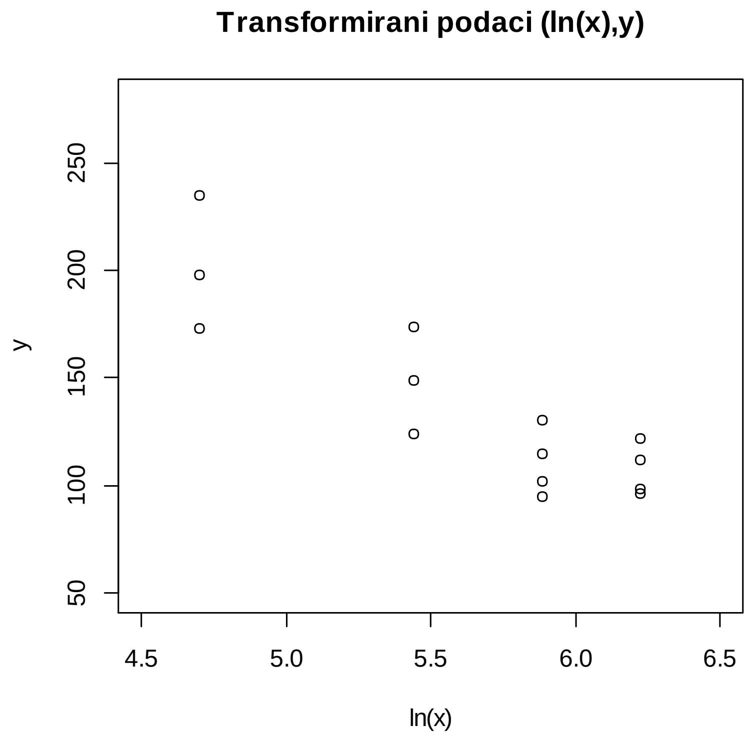


(2) $(x',y')=(ln(x),ln(y))$

> xn<-log(x)
> yn<-log(y)
> plot(xn,yn,xlab="ln(x)",ylab="ln(y)",main="Transformirani podaci (ln(x),ln(y))",xlim=c(4.5,6.5),ylim=c(4.5,5.6))



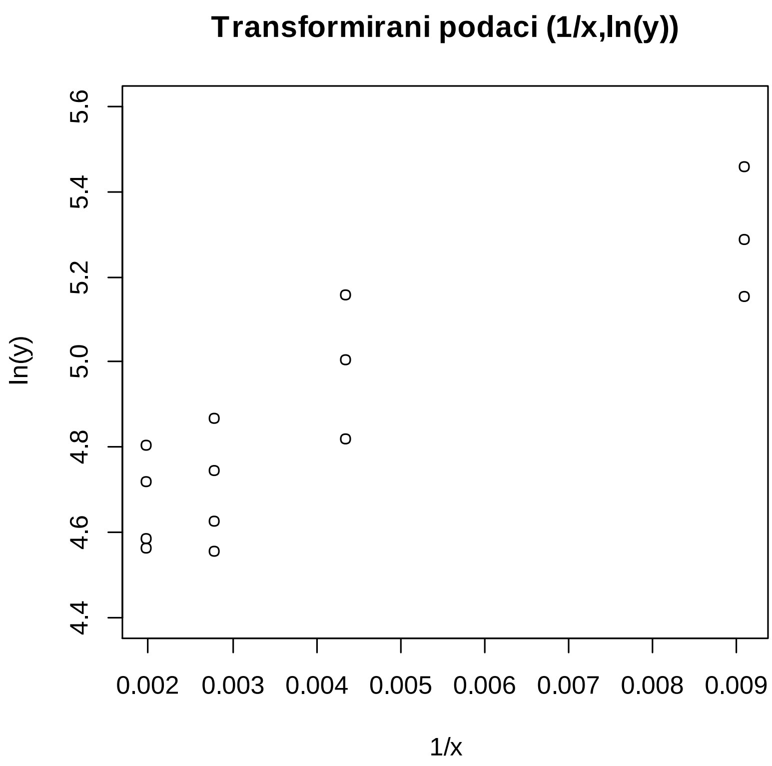
(3) $(x',y')=(ln(x),y)$
> plot(xn,y,xlab="ln(x)",ylab="y",main="Transformirani podaci (ln(x),y)",xlim=c(4.5,6.5),ylim=c(50,280))



(4) $(x', y') = (1/x\ ,\ ln(y))$

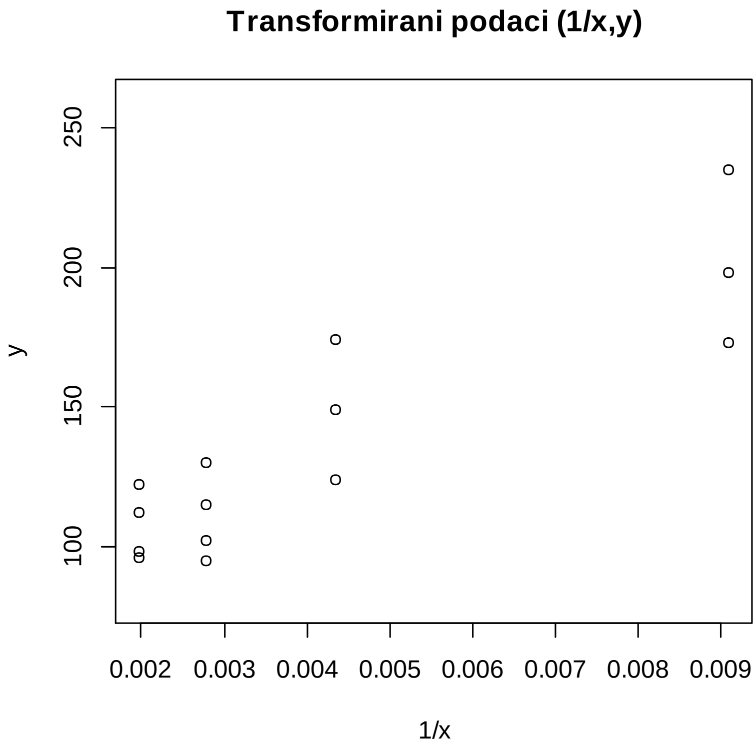
> xn<-x^(-1)

> plot(xn,yn,xlab="1/x",ylab="ln(y)",main="Transformirani podaci (1/x,ln(y))",ylim=c(4.4,5.6))



(5) $(x',y')=(1/x\,,y)$

```
> xn<-x^(-1)
> plot(xn,y,xlab="1/x",ylab="y",main="Transformirani podaci (1/x,y)",ylim=c(80,260))
```



Iz ovih grafova je teško odrediti postoji li linearna zavisnost među podacima, zato što je uzorak takav da sadrži mali broj različitih težina x te "velika" odstupanja pri istoj težini x. Kod grafa originalnih podataka(model (0)) i modela (1), ne izgleda kao da bi podatke dobro mogli aproksimirati pravcem, već kao da bi bolje odgovarala parabola, funkcija e^{-x} ili slična krivulja. Modeli (2), (3), (4) i (5) se čine otprilike jednako dobri, ali se unaprijed iz njih ne može pretpostavili linearna zavisnost podataka već treba provesti testove i provjeriti.

b)Uzmite dva modela iz a) za koja najviše sumnjate u linearnu zavisnost podataka te za oba modela napravite prilagodbu linearnog modela $y' = \theta_0 + \theta_1 * x'$ transformiranim podacima i dobiveni pravac prikažite na istom grafu zajedno s transformiranim podacima. Izračunajte statistiku R^2, te provedite test adekvatnosti modela u oba slučaja. Odaberite bolji model.

Odabrala sam modele (3) i (5). Pravac prilagođavam metodom najmanjih kvadrata, odnosno minizacijom funkcije :

$$L(\theta_0,\theta_1)=\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 * x_i)$$

Koeficijente θ_0 *i* θ_1 za koje $L(\theta_0,\theta_1)$ poprima minimum izračunat ću rješavanjem sustava:

$$\frac{\partial L}{\partial \theta_0}(\theta_0,\theta_1) = 0$$

$$\frac{\partial L}{\partial \theta_1}(\theta_0,\theta_1) = 0$$

Slijedi da su formule za θ_0 *i* θ_1 dane izrazima:

$$\theta_0 = \frac{S_{xy}}{S_{xx}} \qquad \theta_1 = \bar{y}\text{-}\theta_1 * \bar{x}$$

Koeficijentom determinacije R^2 $\left(= \frac{SSR}{S_{yy}} \ \epsilon[0,1] \right)$ izražavamo jačinu linearne povezanosti. On utvrđuje koliko je promjene zavisne varijable objašnjeno promjenom nezavisne varijable. Model je reprezentativniji što je koeficijent determinacije bliže jedinici.

Prvo računam koeficijente θ_0 *i* θ_1 te koeficijent determinacije za model (3):

```
> xn<-log(x)
> yn<-y
> sxx<-sum(xn^2)-n*mean(xn)^2
> syy<-sum(yn^2)-n*mean(yn)^2
> sxy<-crossprod(xn,yn)-n*mean(xn)*mean(yn)
> theta1<-sxy/sxx
> theta1<-c(theta1)
> theta0<-mean(yn)-theta1*mean(xn)
> ssr<-(theta1^2)*sxx
> R2<-ssr/syy
```

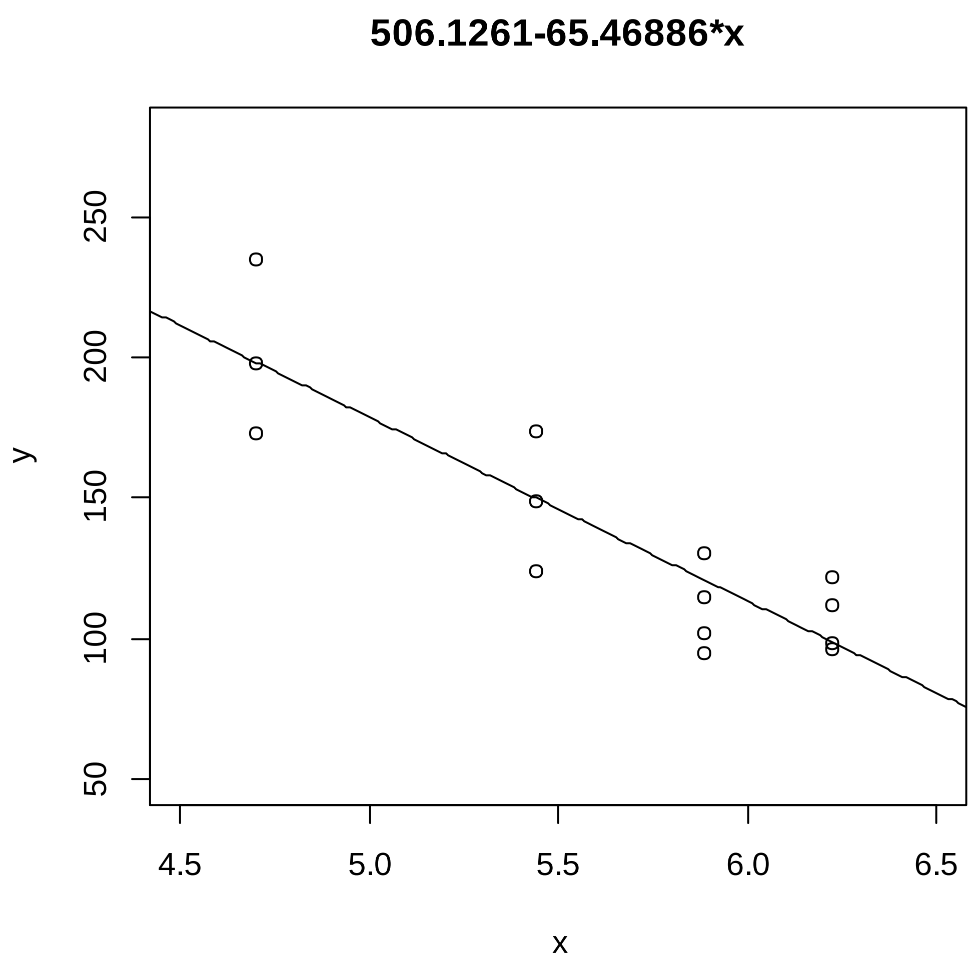
Dobila sam da S_{xx} iznosi 4.378772, S_{yy} iznosi 23875.21, te da S_{xy} iznosi -286.6732. Koeficijenti θ_0 i θ_1 iznose 506.1261, odnosno -65.46886, pa regresijska funkcija glasi:

$$\hat{y} = 506.1261 - 65.46886 * x$$

Koeficijent determinacije za model (3) iznosi 0.7860942 što znači da u tom modelu 21.39058% ukupnog rasipanja potječe od slučajnih pogrešaka, a 78.60942% od danog modela. Za svaku varijablu x imamo nekoliko vrijednosti y, pa je zato postotak koji potječe od slučajnih grešaka toliki.

Prikaz podataka zajedno s pravcem y = 506.1261 - 65.46886 * x :

```
> min(xn)
[1] 4.70048
> max(xn)
[1] 6.224558
> xkoord<-seq(4,7,0.01)
> ykoord<-theta0+theta1*xkoord
> plot(xkoord,ykoord,xlim=c(4.5,6.5),ylim=c(50,280),xlab="x",ylab="y",main="506.1261-65.46886*x",type="l")
> points(xn,yn)
```



Nakon toga računam koeficijente θ_0 *i* θ_1 te koeficijent determinacije za model (5):

```
> xn<-1/x
> yn<-y
> sxx<-sum(xn^2)-n*mean(xn)^2
> syy<-sum(yn^2)-n*mean(yn)^2
> sxy<-crossprod(xn,yn)-n*mean(xn)*mean(yn)
> theta1<-sxy/sxx
> theta1<-c(theta1)
> theta0<-mean(yn)-theta1*mean(xn)
> ssr<-(theta1^2)*sxx
> R2<-ssr/syy
```

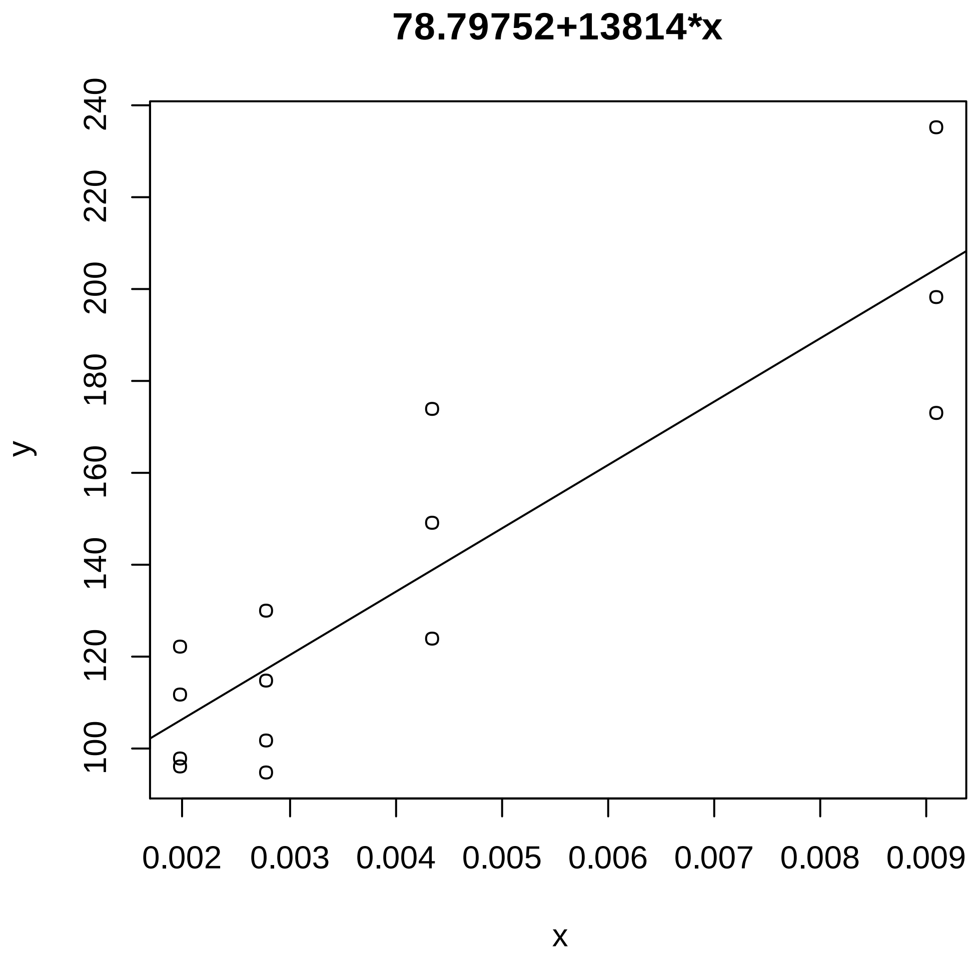
Dobila sam da S_{xx} iznosi 9.960802e-05, S_{yy} iznosi 23875.21, te da S_{xy} iznosi 1.375985. Koeficijenti θ_0 i θ_1 iznose 78.79752, odnosno 13814, pa regresijska funkcija glasi:

$$\hat{y} = 78.79752+13814 * x$$

Koeficijent determinacije za model (5) iznosi 0.7961334, što znači da u tom modelu 20.38666% ukupnog rasipanja potječe od slučajnih pogrešaka, a 79.61334% od danog modela.

Prikaz podataka zajedno s pravcem $y = 78.79752+13814 * x$:

```
> max(xn)
[1] 0.009090909
> min(xn)
[1] 0.001980198
> plot(xn,yn,main="78.79752+13814*x",xlab="x",ylab="y")
> abline(theta0,theta1)
```



Test adekvatnosti linearnog modela :

Među n danih vrijednosti x, točno l ih je različito.. Označit ću ih sa $z_{i,}$, gdje je i=1...l. Za svaki z_i se y mjeri točno n_i puta. Neka su $y_{i,j}$ dobivene realizacije, a $Y_{i,j}$ slučajne varijable za j=1...n_i, i=1...l.

Testiram:

H_0 : linearni model je adekvatan
H_1 : ne H_0

Vrijedi da je :

$$SSE_P=\sum_{i=1}^{l}\sum_{j=1}^{ni}(y_{i,j}-\bar{y}_{ni})$$

Statistika F glasi:

$$F=\frac{\frac{SSE-SSE_p}{l-k-1}}{\frac{SSE_p}{n-l}} \sim F(l\text{-}k\text{-}1,n\text{-}l)$$

Prvo računam realizaciju F statistike za određeni model . Kako ta statistika ima F-distribuciju, jednostavno se dobije p-vrijednost, te pomoću toga mogu zaključiti za koje razine značajnosti mogu odbaciti hipotezu H_0, a za koje razine značajnosti to ne mogu.

Prvo provodim test adekvatnosti za model (3):

```
> n<-14
> ni<-c(3,3,4,4)
> l=4
> k=1
> sse<-syy-ssr
> j<-1
> px<-numeric(l)
> for(i in 1:l){
+ px[i]<-xn[j]
+ j=j+ni[i]}
> j<-1
> py<-matrix(0,l,max(ni))
> for(i in 1:l){
+ for( k in 1:ni[i]){
+ py[i,k]<-yn[j]
+ j<-j+1}}
> sred<-c(0,0,0,0)
> for(i in 1:l){
+ for(j in 1:ni[i]){
+ sred[i]<-sred[i]+py[i,j]}
+ sred[i]<-sred[i]/ni[i]}
> ssep<-0
> for(i in 1:l){
+ for(j in 1:ni[i]){
+ ssep<-ssep+((py[i,j]-sred[i])^2)}}
> f<-(sse-ssep)*10/(2*ssep)
> pv<-1-pf(f,2,10)
> pv
[1] 0.4540264
```

Kako je p vrijednost jednaka 0.4540264, ne mogu odbaciti hipotezu H_0 na razini značajnosti α za $\alpha<0.4540264$. No, kako je svaka standardna razina značajnosti sigurno manja od 45% (na vježbama najčešće koristimo razine značajnosti do 10%), mogu zaključiti da ne mogu odbaciti hipotezu H_0 ni za jednu standardnu razinu značajnosti α.

Zatim provodim isti test za model (5):

```
> n<-14
> ni<-c(3,3,4,4)
> l=4
> k=1
> sse<-syy-ssr
> j<-1
> px<-numeric(l)
> for(i in 1:l){
+ px[i]<-xn[j]
+ j=j+ni[i]}
> j<-1
> py<-matrix(0,l,max(ni))
> for(i in 1:l){
+ for( k in 1:ni[i]){
+ py[i,k]<-yn[j]
+ j<-j+1}}
> sred<-c(0,0,0,0)
> for(i in 1:l){
+ for(j in 1:ni[i]){
+ sred[i]<-sred[i]+py[i,j]}
+ sred[i]<-sred[i]/ni[i]}
> ssep<-0
> for(i in 1:l){
+ for(j in 1:ni[i]){
+ ssep<-ssep+((py[i,j]-sred[i])^2)}}
> f<-(sse-ssep)*10/(2*ssep)
> pv<-1-pf(f,2,10)
> pv
[1] 0.5773824
```

Dobila sam da p-vrijednost iznosi 0.5773824, pa kao u modelu (3) zaključujem da ne mogu odbaciti hipotezu H_0 ni za jednu standardnu razinu značajnosti α.
Kako je u modelu (5) p -vrijednost u test adekvatnosti veća te kako je vrijednost statistike R^2 veća , slijedi da je model (5) bolji od modela (3).

c)Za model odabran u b) nacrtajte graf reziduala, graf standardiziranih reziduala te provjerite da li (standardizirani) reziduali dolaze iz jedinične normalne distribucije.

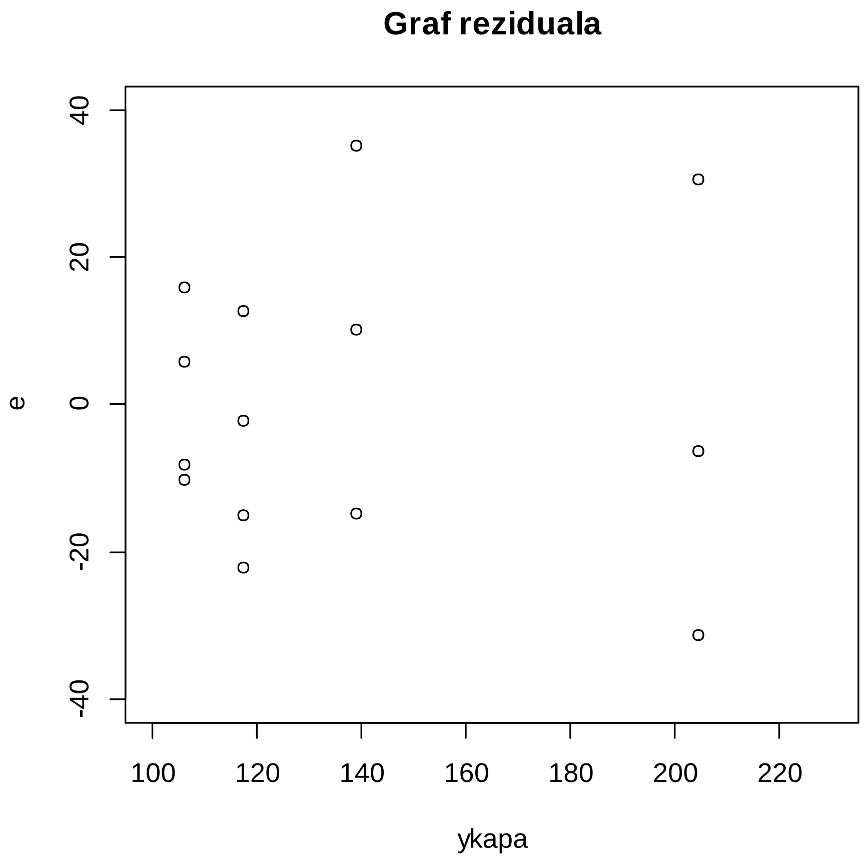
Reziduali su slučajne varijable $E_i=Y_i-\widehat{Y_i}$, odnosno njihove realizacije e_i, za i=1...n. Standardizirani reziduali su slučajne varijable

$$E_i^S = \frac{E_i}{\hat{\sigma} * \sqrt{(1-h_{ii})}}$$

odnosno njihove realizacije e^s_i za i=1...n, gdje je $H=X(X^T X)^{-1}X^T=[h_{ij}]$, te procjena $\hat{\sigma} = \frac{SSE}{n-k-1}$.

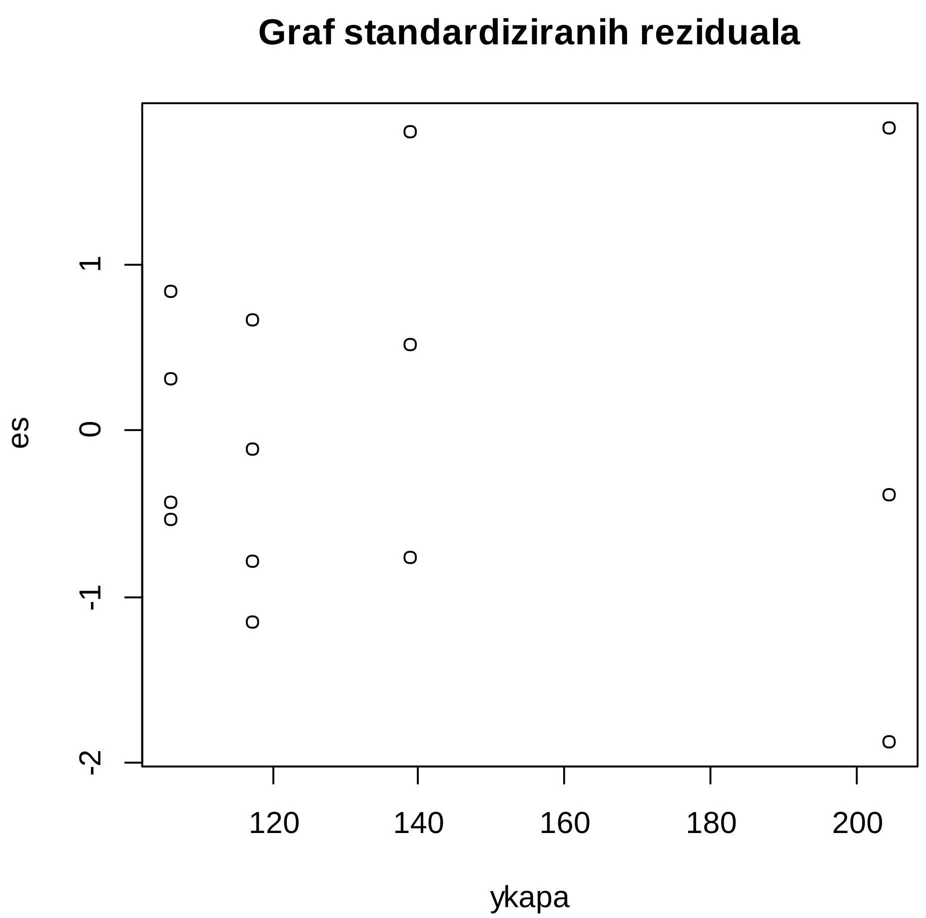
Prvo računam reziduale za model (5) i crtam graf reziduala:

```
> ykapa<-theta0+theta1*xn
> e<-yn-ykapa
> e
 [1]  30.620673  -6.379327 -31.379327  35.141614  10.141614 -14.858386
 [7]  -2.169741  12.830259 -15.169741 -22.169741  15.848025   5.848025
[13]  -8.151975 -10.151975
> plot(ykapa,e,xlim=c(100,230),ylim=c(-40,40),main="Graf 
reziduala",xlab="ykapa",ylab="e")
```



Zatim računam standardizirane reziduale i crtam graf standardiziranih reziduala:

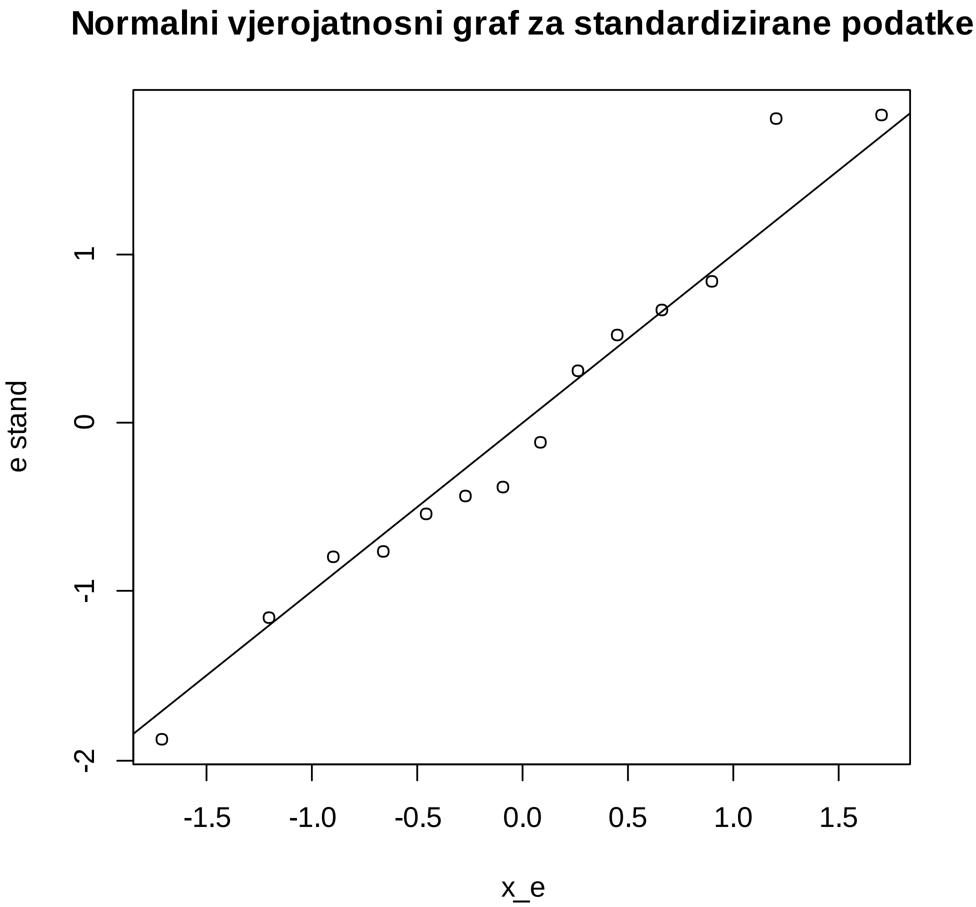
```
> X<-matrix(0,14,2)
> X[,1]<-1
> X[,2]<-xn
> H<-X%*%solve(t(X)%*%X)%*%t(X)
> hii<-diag(H)
> sigmakapa<-sqrt(sse/12)
> es<-e/(sigmakapa*sqrt(1-hii))
> plot(ykapa,es,main="Graf standardiziranih reziduala")
```



S obzirom na grafove, mogu zaključiti da je prilagodba linearnom modelu zadovoljavajuća, jer su točke slučajno grupirane oko 0 po x-osi.

Testiram dolaze li standardizirani reziduali iz jedinične normalne distribucije pomoću normalnog vjerojatnosnog grafa :

```
> ye<-sort(es)
> i<-1:14
> xe<-qnorm((i-3/8)/(14+1/4))
> xe
 [1] -1.70755309 -1.20534492 -0.89943491 -0.66075113 -0.45498114 -0.26699413 -0.08806557  0.08806557
 [9]  0.26699413  0.45498114  0.66075113  0.89943491  1.20534492  1.70755309
> plot(xe,ye,xlab="x_e",ylab="e stand",main="Normalni vjerojatnosni graf")
> abline(0,1)
```



U idealnom slučaju, uzorak bi trebao ležati što bliže pravcu y=x, bez prevelikog odudaranja u rubovima. Na grafu se vidi da se podaci dobro podudaraju s danim pravcem, te da nema odudaranja u rubovima, osim jednog outliera u gornjem desnom rubu (predzadnji element), pa zaključujem da reziduali dolaze iz jedinične normalne distribucije.

d)Nađite 95% pouzdane intervale za parametre θ_0 i θ_1 prilagođenog lineranog modela (odabranog u b)), te 95% pouzdano područje za (θ_0 , θ_1) i prikažite grafički dobiveno pouzadno područje.

Vrijedi:

$$\frac{\widehat{\theta_0}-\theta_0}{\hat{\sigma}*\sqrt{\left(\frac{1}{n}+\frac{\bar{x}*\bar{x}}{S_{xx}}\right)}} \sim t(n\text{-}2) \qquad \frac{\widehat{\theta_1}-\theta_1}{\hat{\sigma}*\sqrt{\left(\frac{1}{S_{xx}}\right)}} \sim t(n\text{-}2)$$

Iz toga slijedi da su formule za konstrukciju (1-α)100% pouzdanih intervala za θ_0 i θ_1 :

$$[\ \widehat{\theta_0} - t_{\frac{\alpha}{2}}(n-2) * \hat{\sigma} * \sqrt{\left(\frac{1}{n} + \frac{\bar{x}*\bar{x}}{S_{xx}}\right)}\quad ,\ \widehat{\theta_0} + t_{\frac{\alpha}{2}}(n-2) * \hat{\sigma} * \sqrt{\left(\frac{1}{n} + \frac{\bar{x}*\bar{x}}{S_{xx}}\right)}\ ]$$

$$[\ \widehat{\theta_1} - t_{\frac{\alpha}{2}}(n-2) * \hat{\sigma} * \sqrt{\left(\frac{1}{S_{xx}}\right)}\ ,\ \widehat{\theta_1} + t_{\frac{\alpha}{2}}(n-2) * \hat{\sigma} * \sqrt{\left(\frac{1}{S_{xx}}\right)}\ ]$$

```
> sigmakapa<-sqrt(sse/12)
> t<-qt(1-0.05/2,12)
> theta0+t*sigmakapa*sqrt(1/14+mean(xn)^2/sxx)
[1] 100.8186
> theta0-t*sigmakapa*sqrt(1/14+mean(xn)^2/sxx)
[1] 56.77647
```

95% pouzdani interval za parametar θ_0 je [56.77647 , 100.8186]

```
> theta1+t*sigmakapa*sqrt(1/sxx)
[1] 18210.72
> theta1-t*sigmakapa*sqrt(1/sxx)
[1] 9417.278
```

95% pouzdani interval za parametar θ_1 je [9417.278, 18210.72]

Pouzdano područje za (θ_0 , θ_1) konstruiram koristeći:

$$F= \frac{1}{(k+1)*\hat{\sigma}*\hat{\sigma}}((X^T X)(\theta-\hat{\theta}), \theta-\hat{\theta}) \sim F(k+1, n-k-1)$$

Zbog $P(F<=f_\alpha)=1-\alpha$ se iz $F- f_\alpha<=0$ dobije $(1-\alpha)100\%$ pouzdano područje za $\theta= (\theta_0 , \theta_1)$. Kako je k=1, kao rješenje se dobije unurašnjost elipse.

Računam koeficijente elipse:

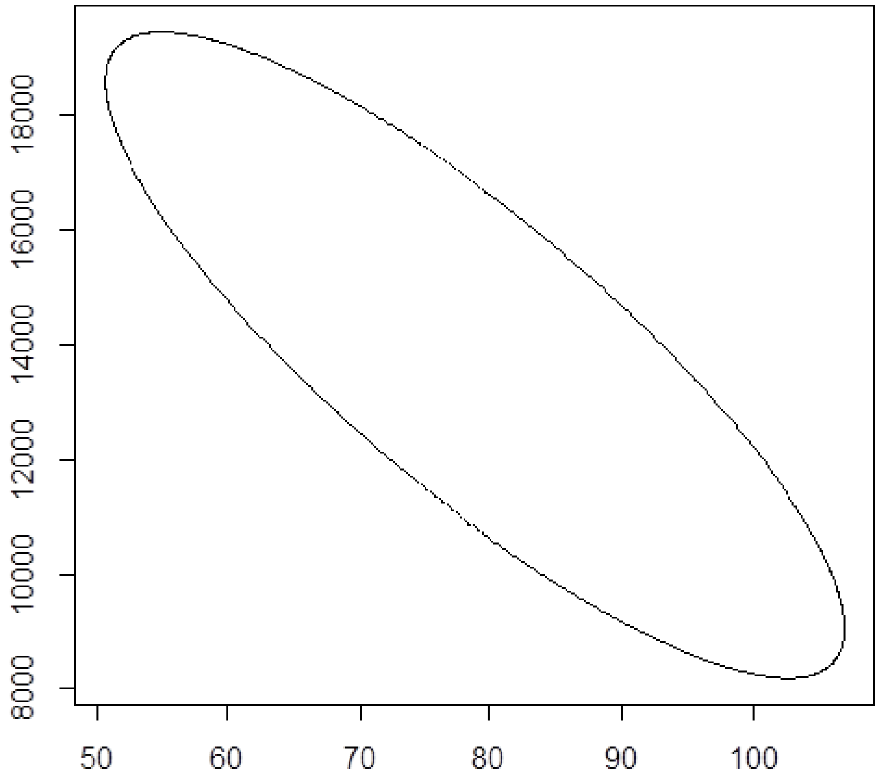
```
> k<-1
> n<-14
> f<-qf(0.95,2,12)
> f
```

```
[1] 3.885294
> koef<-numeric(6)
> m<-t(X)%*%X
> koef[1]<-m[1,1]/(2*sigmakapa^2)
> koef[2]<-m[2,2]/(2*sigmakapa^2)
> koef[3]<-(m[1,2]+m[2,1])/(2*sigmakapa^2)
> koef[4]<-(-2*m[1,1]*theta0-theta1*(m[1,2]+m[2,1]))/(2*sigmakapa^2)
> koef[5]<-(-theta0*(m[1,2]+m[2,1])-2*theta1*m[2,2])/(2*sigmakapa^2)
> koef[6]<-
(m[1,1]*theta0^2+(m[1,2]+m[2,1])*theta0*theta1+m[2,2]*theta1^2)/(2*sigmakapa^2)-f
```

Rješenje je elipsa zadana jednadžbom:

0.01725782*x^2+4.329168e-07*y^2+0.000146317*x*y-4.740969*x-0.02349004*y+345.1487=0

```
> plot.ellipse <- function (a, b, c, d, e, f, n.points = 1000) {
+  A <- matrix(c(a, c / 2, c / 2, b), 2L)
+  B <- c(-d / 2, -e / 2)
+  mu <- solve(A, B)
+  r <- sqrt(a * mu[1] ^ 2 + b * mu[2] ^ 2 + c * mu[1] * mu[2] - f)
+  theta <- seq(0, 2 * pi, length = n.points)
+  v <- rbind(r * cos(theta), r * sin(theta))
+  z <- backsolve(chol(A), v) + mu
+  plot(t(z), type = "l")
+  }
>  plot.ellipse(koef[1],koef[2],koef[3],koef[4],koef[5],koef[6])
```



e) Kako glasi model (regresijska funkcija) za originalne podatke (x,y) (iz a))?

Uz objašnjenje kao u b) dijelu zadatka, računam koeficijente θ_0 i θ_1 za model (0):

```
> sxx<-sum(x^2)-n*mean(x)^2
> syy<-sum(y^2)-n*mean(y)^2
> sxy<-crossprod(x,y)-n*mean(x)*mean(y)
> theta1<-sxy/sxx
> theta1<-c(theta1)
> theta0<-mean(y)-theta1*mean(x)
> ssr<-(theta1^2)*sxx
```

Dobila sam da S_{xx} iznosi 299900, S_{yy} iznosi 23875.21, te da S_{xy} iznosi -70630.
Koeficijenti θ_0 i θ_1 iznose 212.7209, odnosno -0.2355118, pa regresijska funkcija glasi:

$$\hat{y} = 212.7209 - 0.2355118 * x$$

Prikažite te podatke zajedno sa regresijskom funkcijom i krivuljama koje definiraju gornje i donje 95% pouzdane intervale, prvo za srednju vrijednost Y (uz dano x), te zatim i za Y(uz dano x).

Formule za konstrukciju pouzadnih intervala:

(1-α)100% pouzdani interval za E[Y|x=x_0]:

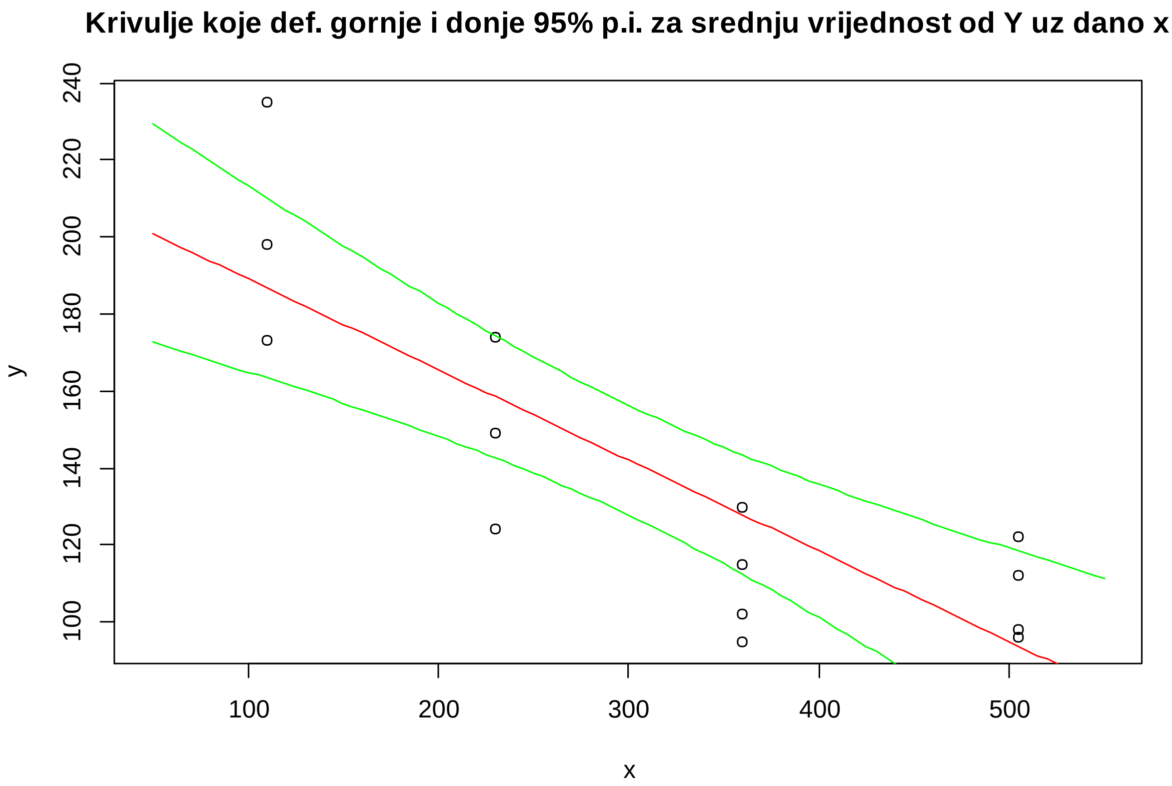
$$\hat{E}[Y|x = x_0] \pm t_{\frac{\alpha}{2}}(n-1) * \hat{\sigma} * \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

(1-α)100% pouzdani za Y u x=x_0:

$$\hat{Y} \pm t_{\frac{\alpha}{2}}(n-1) * \hat{\sigma} * \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

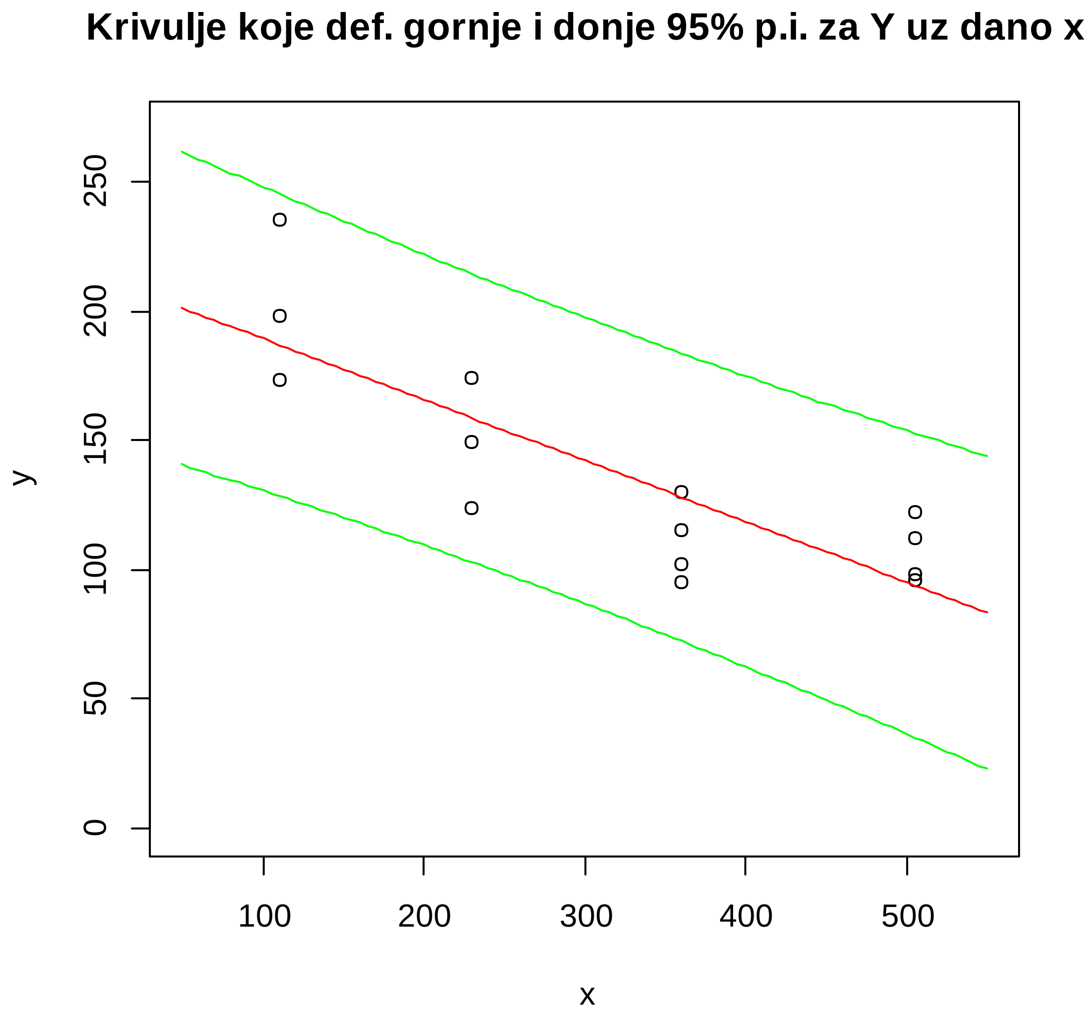
Prikaz podataka s regresijskom funkcijom i krivuljama koje definiraju gornja i donje 95% pouzdane intervale za srednju vrijednost Y uz dano x:

```
> sse<-syy-ssr
> sigmakapa<-sqrt(sse/12)
> t<-qt(1-0.05/2,12)
> donja<-function(x){
+ theta0+theta1*x-t*sigmakapa*sqrt(1/n+(x-mean(x))^2/sxx)}
> gornja<-function(x){
+ theta0+theta1*x+t*sigmakapa*sqrt(1/n+(x-mean(x))^2/sxx)}
> pravac<-function(x){
+ theta0+theta1*x}
> plot(x,y,main="Krivulje koje def. gornje i donje 95% p.i. za srednju vrijednost od Y uz dano x",xlim=c(50,550))
> curve(pravac,col="red",add=T)
> curve(donja,col="green",add=T)
> curve(gornja,col="green",add=T)
```



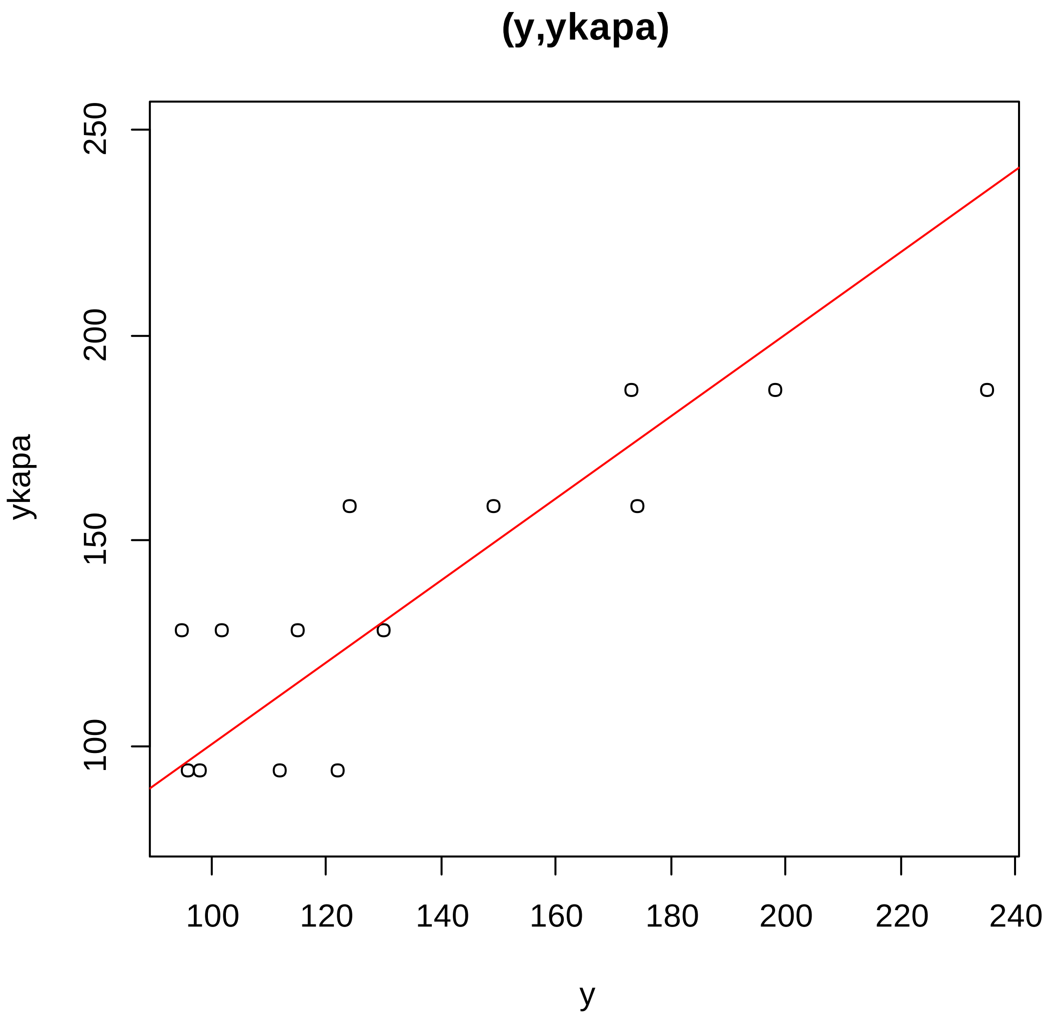
Prikaz podataka s regresijskom funkcijom i krivuljama koje definiraju gornja i donje 95% pouzdane intervale za Y uz dano x:

```
donja<-function(x){
theta0+theta1*x-t*sigmakapa*sqrt(1+1/n+(x-mean(x))^2/sxx)}
gornja<-function(x){
theta0+theta1*x+t*sigmakapa*sqrt(1+1/n+(x-mean(x))^2/sxx)}
pravac<-function(x){
theta0+theta1*x}
plot(x,y,main="Krivulje koje def. gornje i donje 95% p.i. za Y uz dano
x",xlim=c(50,550),ylim=c(0,270))
curve(pravac,col="red",add=T)
curve(donja,col="green",add=T)
curve(gornja,col="green",add=T)
```



Također, prikažite točke (y, $\hat{y}$) zajedno s pravcem y=x u Kartezijevom koordinatnom sustavu ($\hat{y}$ je procjena od y na osnovu modela za originalne podatke) te odgovorite na pitanje je li taj model dobar za originalne podatke.

```
> ykapa<-theta0+theta1*x
> ykapa
 [1] 186.81463 186.81463 186.81463 158.55321 158.55321 158.55321 127.93667
127.93667 127.93667
[10] 127.93667  93.78745  93.78745  93.78745  93.78745
> plot(y,ykapa,main="(y,ykapa)",ylim=c(80,250))
> abline(0,1,col="red")
```



Točke (y, $\hat{y}$) bi trebale ležati što bliže pravcu y=x jer je u tom slučaju procjena bolja, tj $\theta_0+\theta_1*x$ bolje aproksimira y . Na grafu se vidi da te točke nisu grupirane oko pravca, već su dosta udaljene od njega, pa se ne može tvrditi da je model dobar.