Linearna regresija

Borna Bešić, Tomislav Buhiniček, Nikola Zadravec

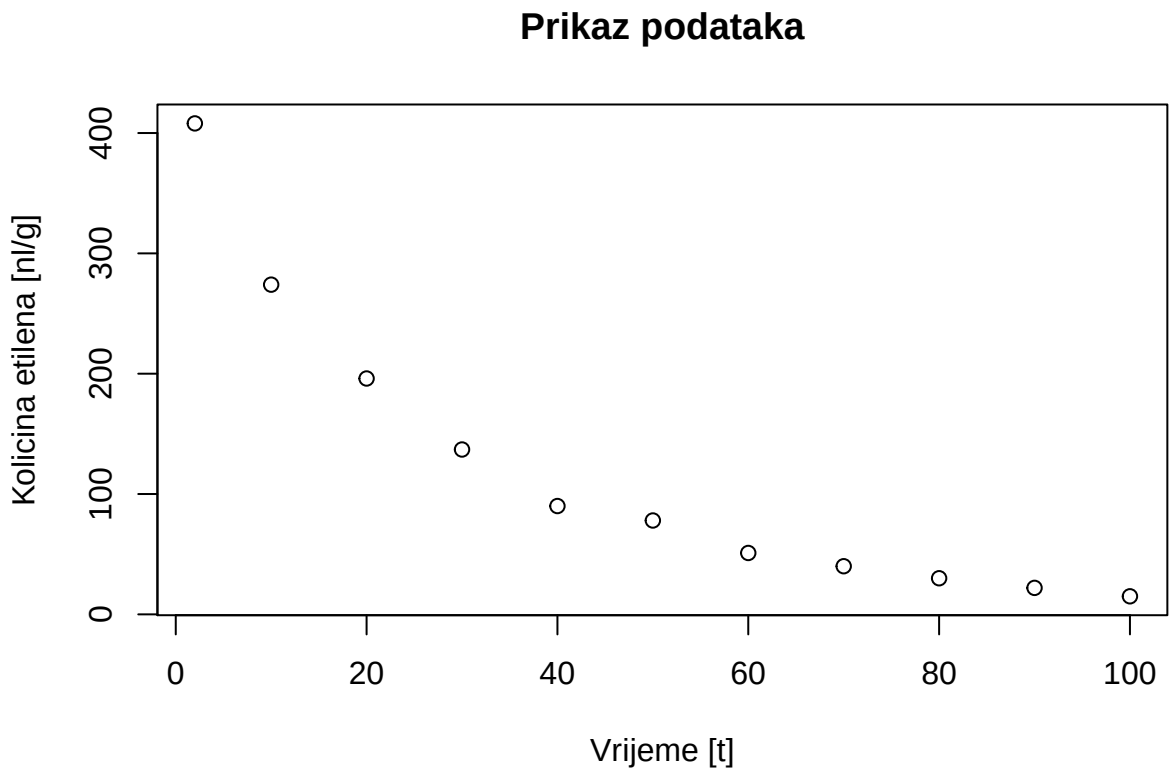
2. lipnja 2017.

Zadatak A

U članku "Ethylene Synthesis in Lettuce Seeds: Its Physiological Significance" (Plant Physiology, 1972., str. 719-722) proučava se količina etilena (y, u nl/g) koju sadrži sjeme salate kao funkcija vremena izlaganja (x, u minutama) tvari koja apsorbira etilen. Podaci se nalaze u datoteci zad51r.dat (Devore, Jay L., Probability and Statistics for Engineering and the Sciences, 1982., Brooks/Cole Publishing Company, Monterey, California, str. 472).

Prikaz podataka u Kartezijevom koordinatnom sustavu

Na slijedećem dijagramu prikazani su parovi podataka (x, y) iz zadanog skupa:



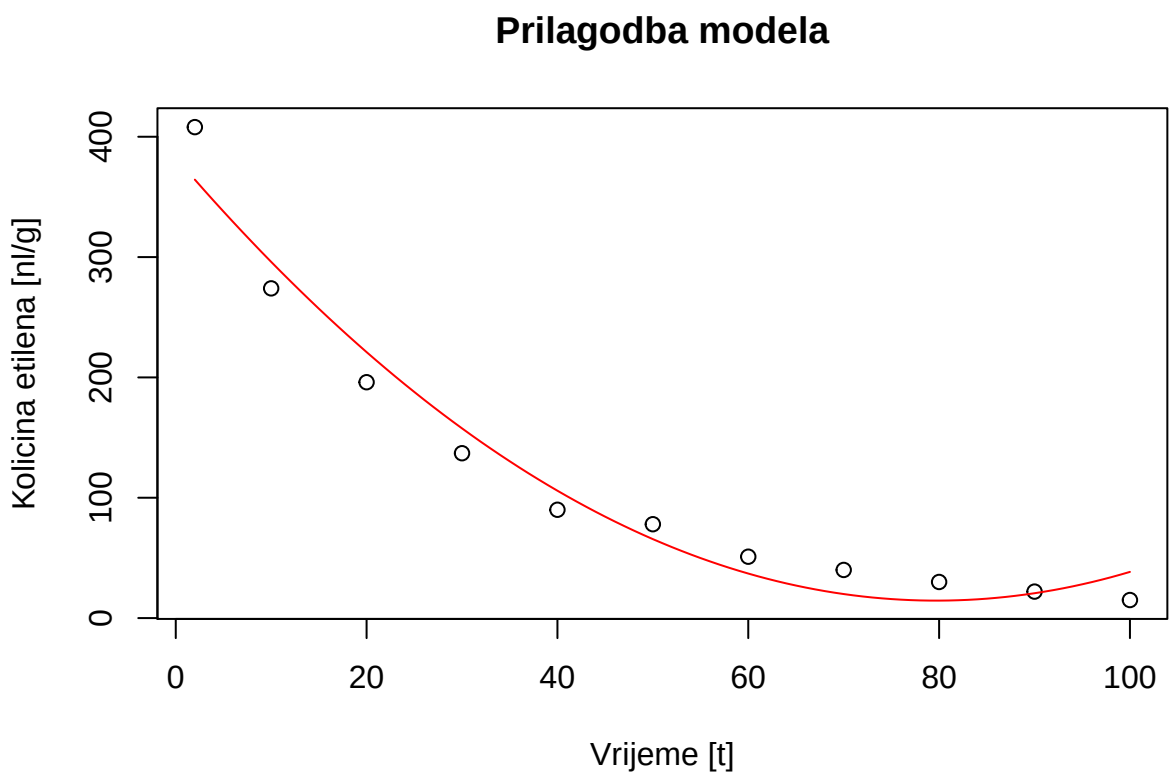
Prilagodba kvadratičnog modela

Prvi model čiju ćemo prilagodbu provesti jest slijedeći kvadratični model: $y = \theta_0 + \theta_1 x + \theta_2 x^2$

```
X <- A.data$x
Y <- A.data$y
X.squared <- X^2

model <- lm(Y ~ X +  X.squared)
```

Slijedeći graf prikazuje parabolu dobivenu prilagodbom navedenog modela zajedno s empirijskim podacima:

```
predict.original <- function(x){
  beta0 <- model$coefficients["(Intercept)"]
  beta1 <- model$coefficients["X"]
  beta2 <- model$coefficients["X.squared"]
  return(beta0 + beta1 * x + beta2 * x^2)
}

plot(X, Y, xlab = "Vrijeme [t]", ylab = "Kolicina etilena [nl/g]",
     main="Prilagodba modela")
x.draw <- min(X):max(X)
y.draw <- predict.original(x.draw)
lines(x.draw, y.draw, col="red")
```




```
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X + X.squared)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.040 -21.335   1.353  14.753  43.637
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 382.605565  20.362603  18.790 6.65e-08 ***
## X            -9.237263   0.946518  -9.759 1.02e-05 ***
## X.squared     0.057950   0.009036   6.413 0.000206 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.58 on 8 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9579
## F-statistic: 114.7 on 2 and 8 DF,  p-value: 1.292e-06
```

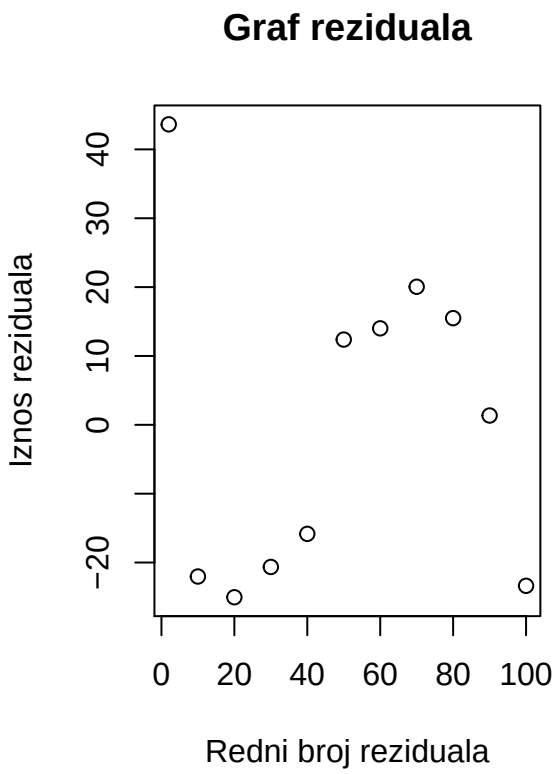
Testiramo slijedeću hipotezu: $\theta_2 = 0$, uz dvostranu alternativu. Kao što je vidljivo, p-vrijednost za parametar θ_2 (koji stoji uz x^2) iznosi 0.000206. Prema tome, uz razinu značajnosti $\alpha = 5\%$, odbacujemo nultu hipotezu u korist alternative.

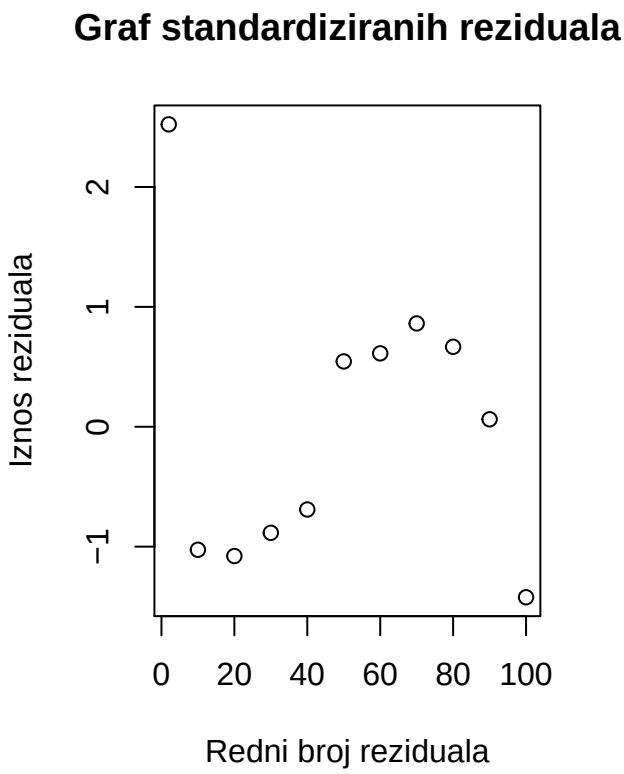
Također, iz priloženog sažetka doznajemo vrijednost statistike R^2 koja iznosi 0.9663. To je prilično zadovoljavajuća vrijednost iako možemo bolje kao što ćemo vidjeti u nastavku.

Provjera normalnosti reziduala

Pretpostavka linearne regresije jest da su reziduali normalno distribuirani. Radi toga radimo provjeru normalnosti na sljedeća dva načina: grafički (QQ plot) te Kolmogorov-Smirnovljevim testom.

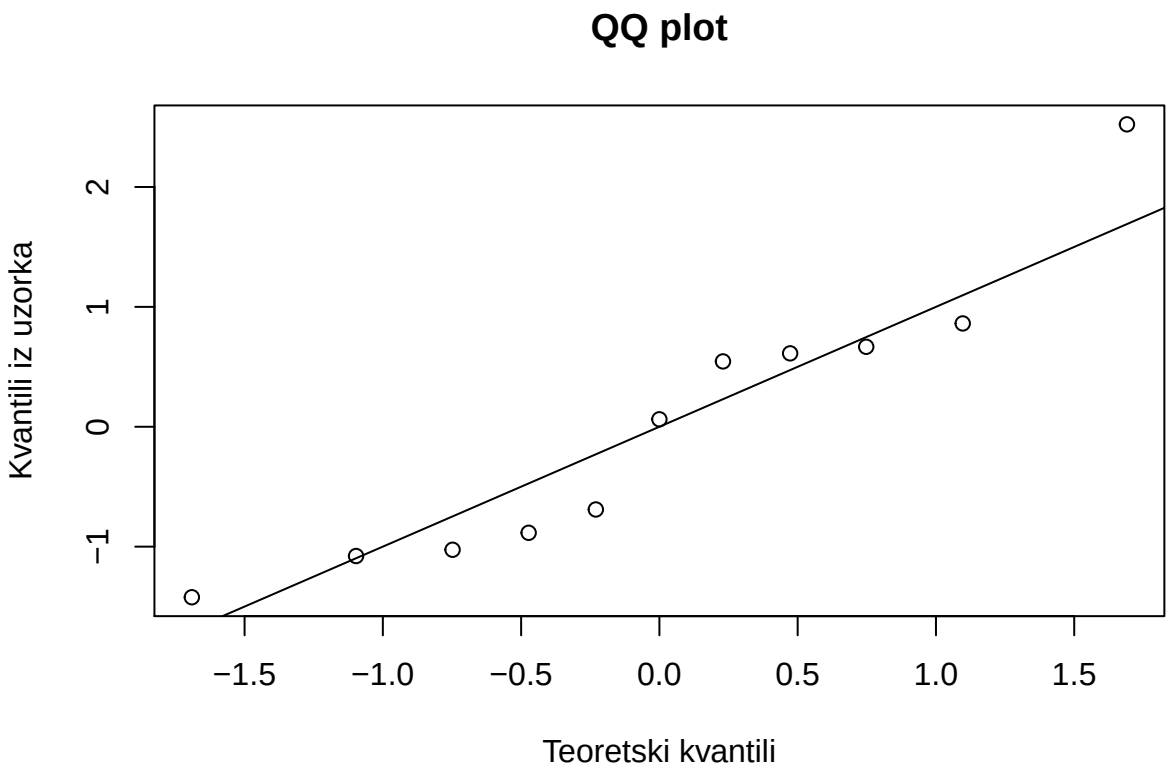
```
par(mfrow=c(1,2))
plot(X, model$residuals, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf reziduala')
residuals.standardized <- rstandard(model)
plot(X, residuals.standardized, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf standardiziranih reziduala')
```





QQ plot

```
qqnorm(residuals.standardized, xlab = 'Teoretski kvantili', ylab = 'Kvantili iz uzorka',
       main = 'QQ plot')
abline(a=0, b=1)
```



Analizom dobivenog QQ plota, iako je uzorak relativno male veličine, može se pretpostaviti da reziduali vrlo vjerojatno ne dolaze iz normalne distribcije.

Kolmogorov-Smirnovljev test

```
ks.test(residuals.standardized, 'pnorm')
```

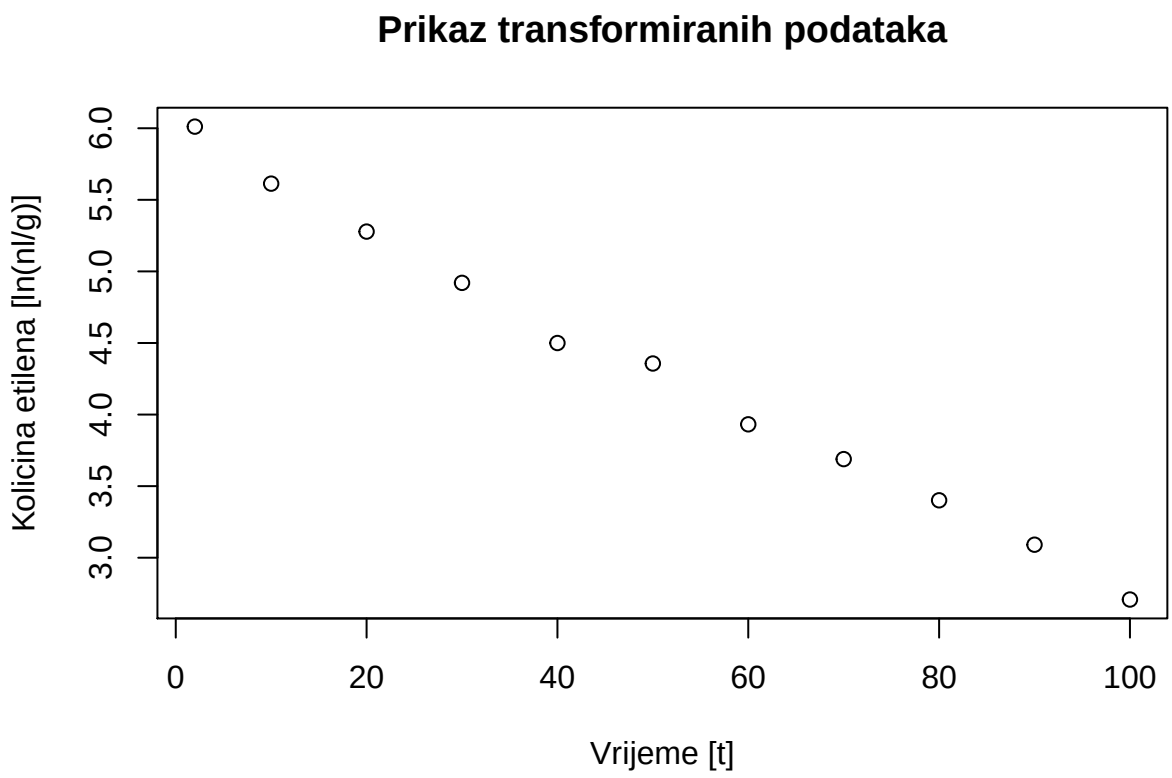
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  residuals.standardized
## D = 0.20947, p-value = 0.6473
## alternative hypothesis: two-sided
```

Provedbom Kolmogorov-Smirnovljevog testa dobivamo p-vrijednost jednaku 0.6473. Sa razinom značajnosti $\alpha = 5\%$ ne možemo odbaciti početnu hipotezu da su reziduali normalno distribuirani u korist dvostrane alternative.

Logaritamska transformacija podataka

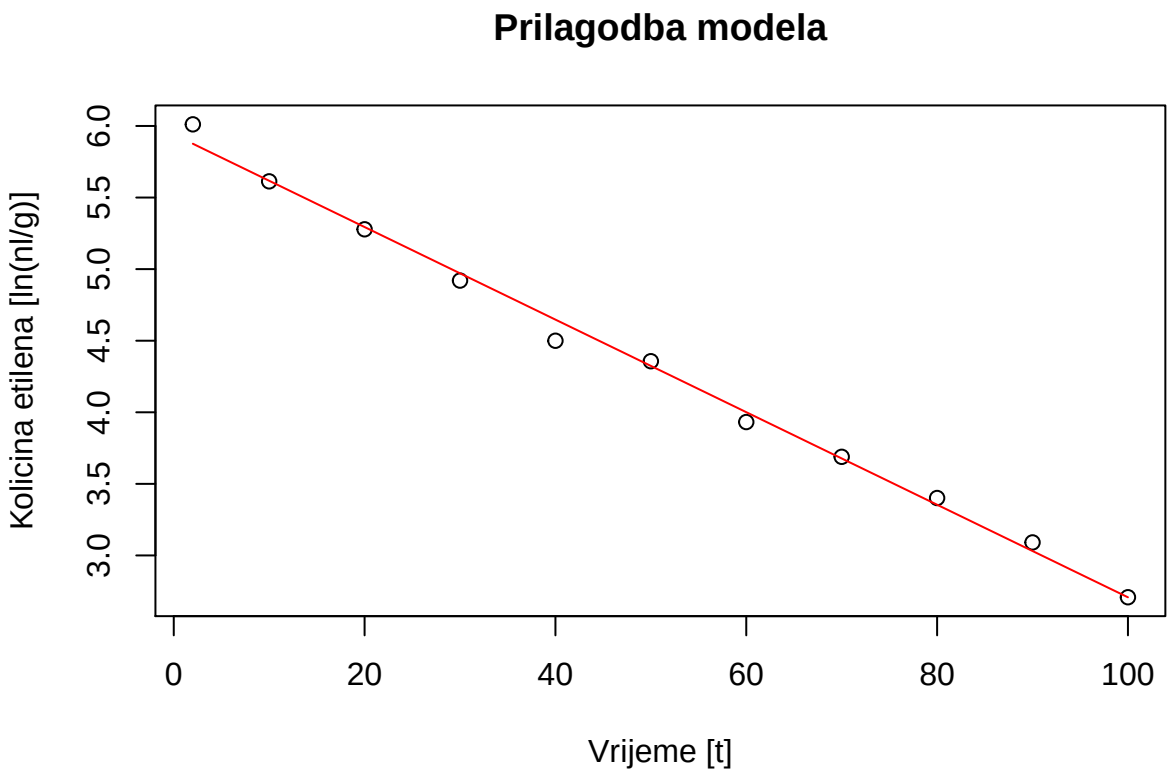
Slijedeće što ćemo napraviti jest transformirati originalni skup podataka kako bi vidjeli ima li transformacija utjecaj na rezultate. Transformacija koju ćemo koristi je slijedeća: $y^0 = ln(y)$.

```r
Y0 <- log(Y)
plot(X, Y0, xlab = "Vrijeme [t]", ylab = "Kolicina etilena [ln(nl/g)]",
     main="Prikaz transformiranih podataka")
```



Kao što je vidljivo na dijagramu raspršenja, nakon transformacije podaci izgledaju puno bolje. Model kojeg ćemo u ovom slučaju iskoristiti glasi: $y^0 = \theta_0 + \theta_1 x$

```r
model.ln <- lm(Y0 ~ X)

predict.ln <- function(x){
  beta0 <- model.ln$coefficients["(Intercept)"]
  beta1 <- model.ln$coefficients["X"]
  return(beta0 + beta1 * x)
}

plot(X, Y0, xlab = "Vrijeme [t]", ylab = "Kolicina etilena [ln(nl/g)]",
     main="Prilagodba modela")
y.ln.draw <- predict.ln(x.draw)
lines(x.draw, y.ln.draw, col="red")
```




```
summary(model.ln)
```

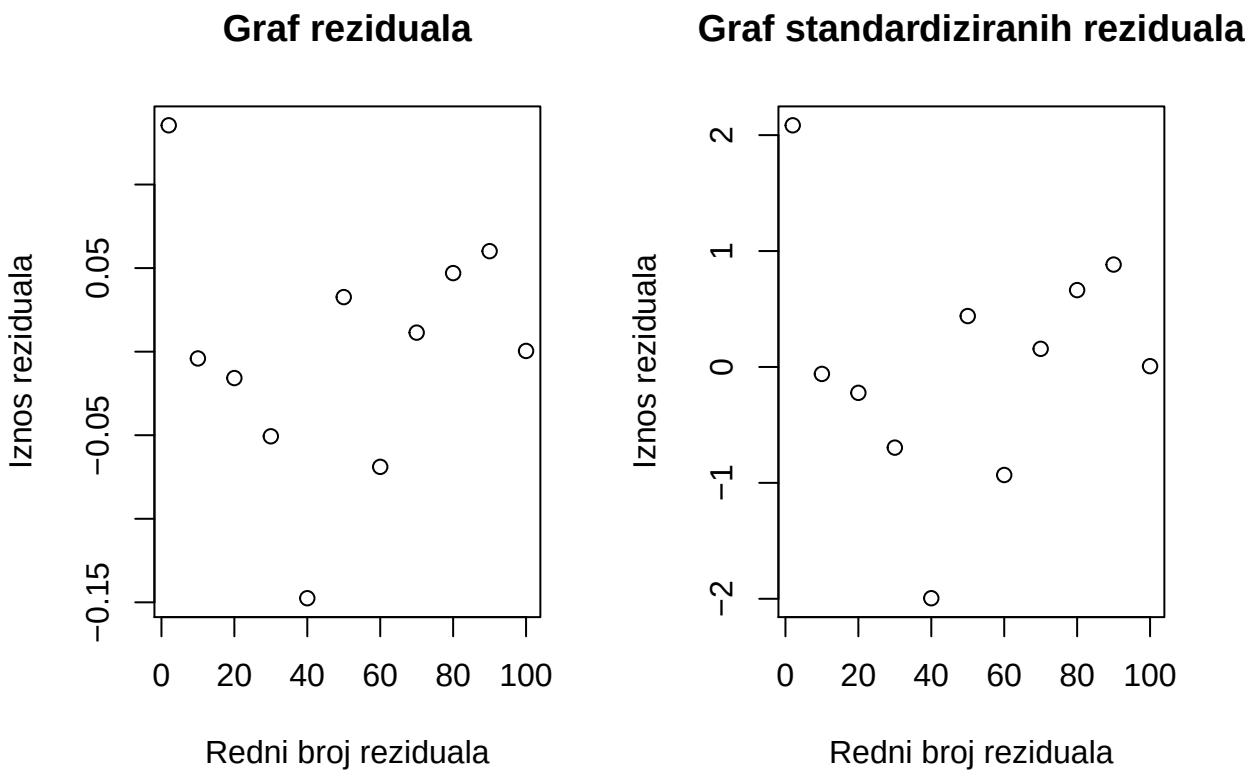
```
##
## Call:
## lm(formula = Y0 ~ X)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.147537 -0.033230  0.000425  0.039823  0.135430
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9404951  0.0443816   133.8 3.68e-16 ***
## X           -0.0323287  0.0007501   -43.1 9.73e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07797 on 9 degrees of freedom
## Multiple R-squared:  0.9952, Adjusted R-squared:  0.9946
## F-statistic:  1857 on 1 and 9 DF,  p-value: 9.734e-12
```

Sada iz sažetka vidimo da je vrijednost R^2 statistike jednaka 0.9952 što je puno bolje od prethodnog slučaja kada smo koristili netrasnformirane podatake. Možemo biti zadovoljni pošto je ova vrijednost vrlo blizu broju 1.

Provjera normalnosti reziduala transformiranih podataka

Kao što smo napravili i za originalni skup podataka, provesti ćemo provjeru normalnosti reziduala, ali sada za transformirane podatke. Koristimo ista dva kriterija: grafički (QQ plot) te Kolmogorov-Smirnovljev test.

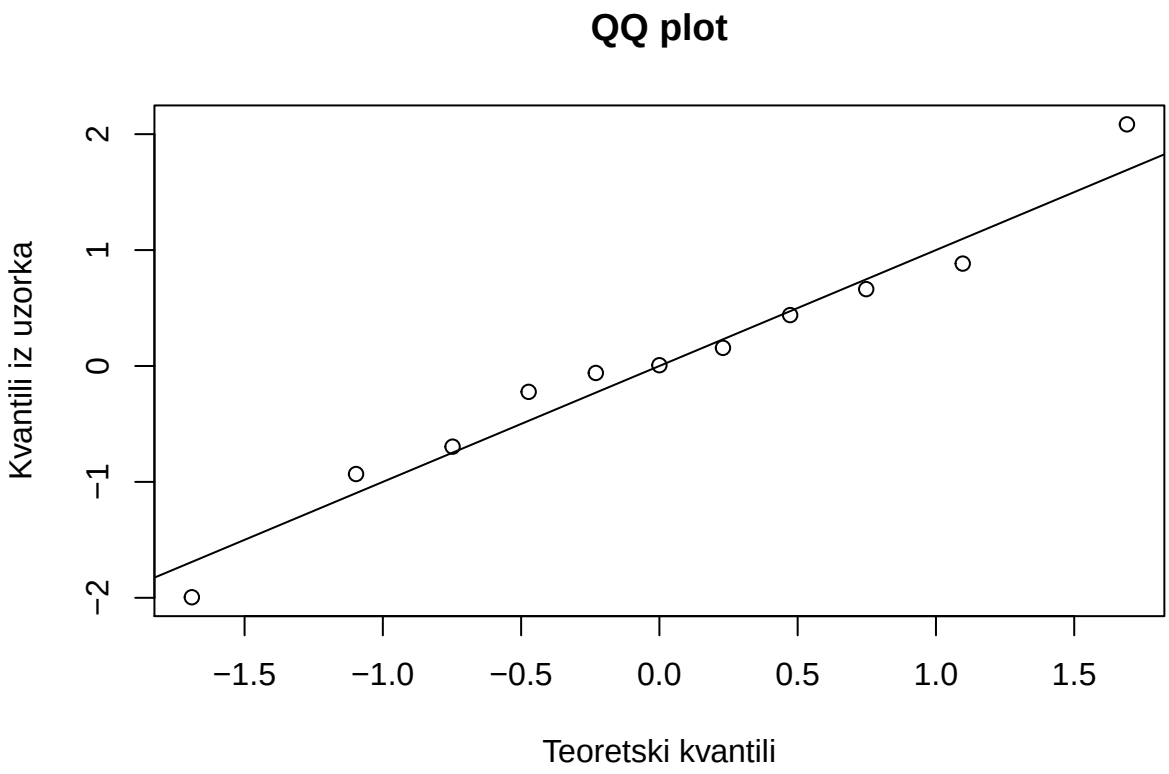
```
par(mfrow=c(1,2))
plot(X, model.ln$residuals, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf reziduala')
residuals.ln.standardized <- rstandard(model.ln)
plot(X, residuals.ln.standardized, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf standardiziranih reziduala')
```



Za razliku od reziduala originalnih podataka, reziduali logaritamski transformiranih podataka pokazuju puno bolju distribuciju kao što je vidljivo iz priloženih grafova.

QQ plot

```
qqnorm(residuals.ln.standardized, xlab = 'Teoretski kvantili', ylab = 'Kvantili iz uzorka',
       main = 'QQ plot')
abline(a=0, b=1)
```



Također, QQ plot reziduala transformiranih podataka pokazuje puno veće podudaranje sa pravcem nego u slučaju netransformiranih podataka.

Kolmogorov-Smirnovljev test

```
ks.test(residuals.ln.standardized, 'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  residuals.ln.standardized
## D = 0.13895, p-value = 0.9644
## alternative hypothesis: two-sided
```

Provedbom Kolmogorov-Smirnovljevog testa nad transformiranim podacima, dobivamo p-vrijednost u iznosu 0.9644. Slijedi da ne možemo na razini značajnosti od $\alpha = 5\%$ odbaciti nultu hipotezu da su reziduali normalno distribuirani u korist dvostrane alternative.

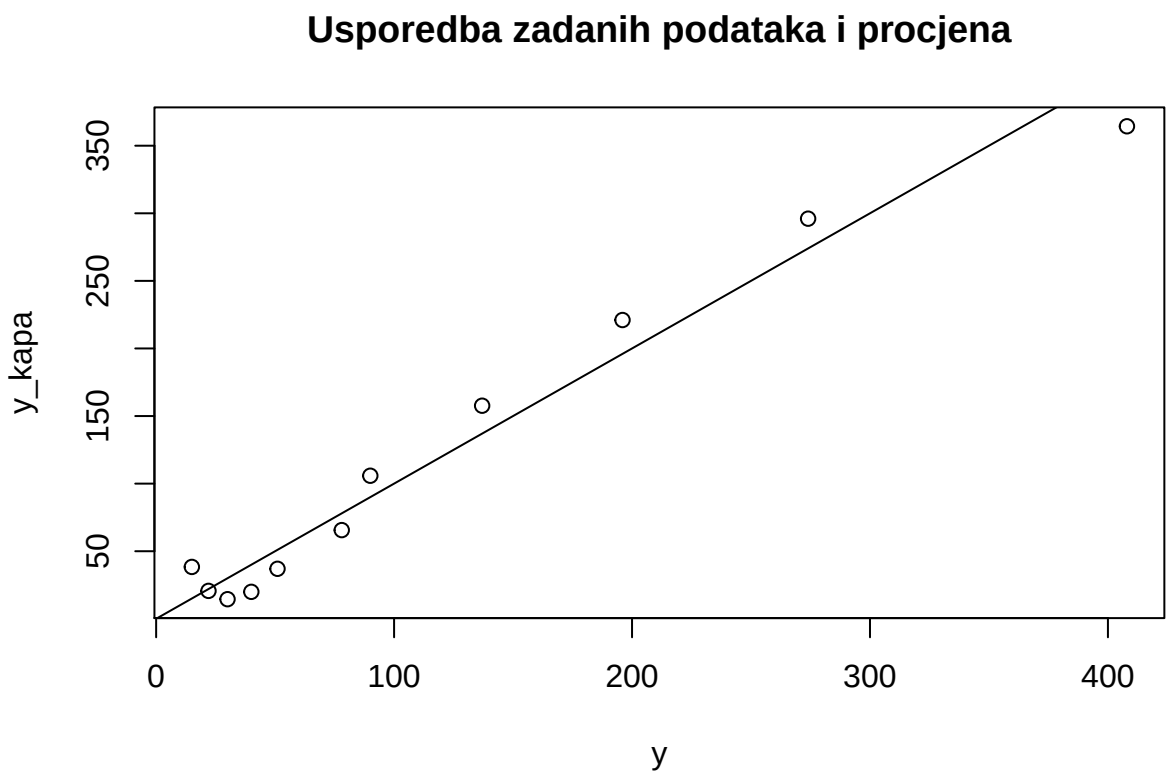
Eksponencijalni model za originalne podatke

Uz pretpostavku da je linearan model dobar za transformirane podatke, njegova formula glasi:

$$\hat{y} = e^{5.9404951 - 0.0323287 \cdot x}$$

Na grafu ispod prikazani su parovi izmjerenih i procijenjenih podataka uz pravac $y = x$.

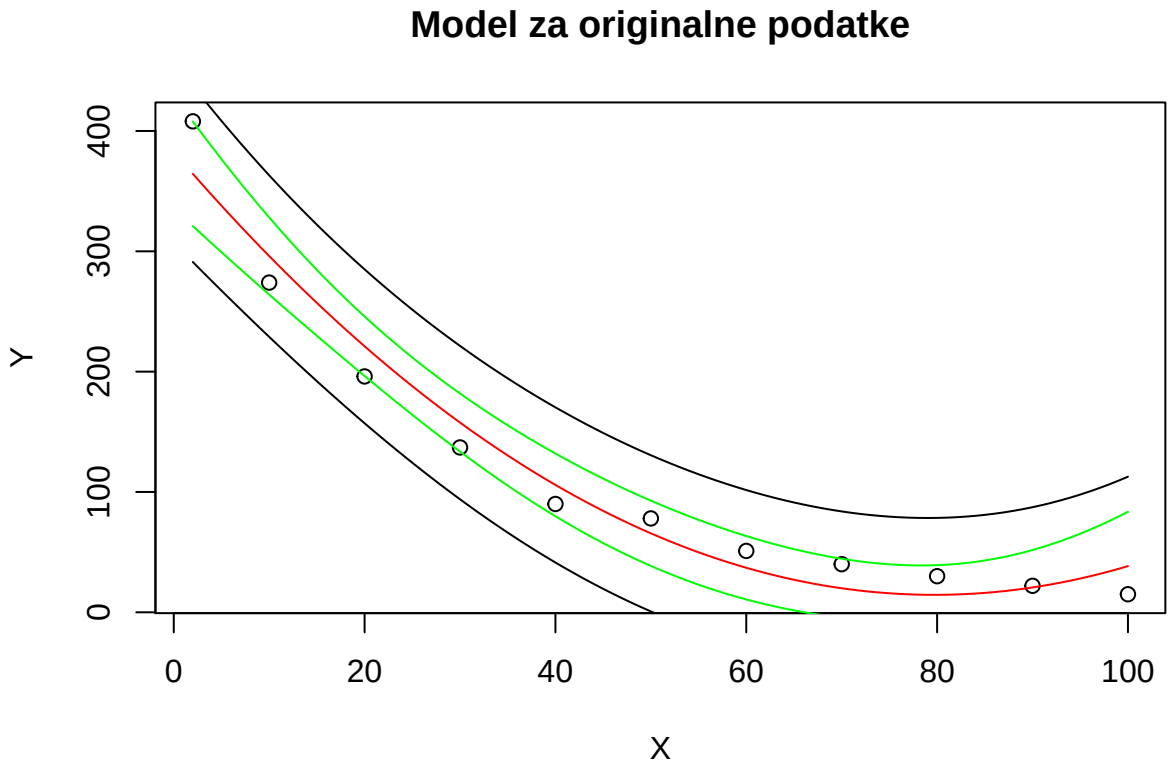
```r
Y.predicted <- predict.original(X)
plot(Y, Y.predicted, xlab="y", ylab="y_kapa",
     main="Usporedba zadanih podataka i procjena")
abline(a=0, b=1)
```



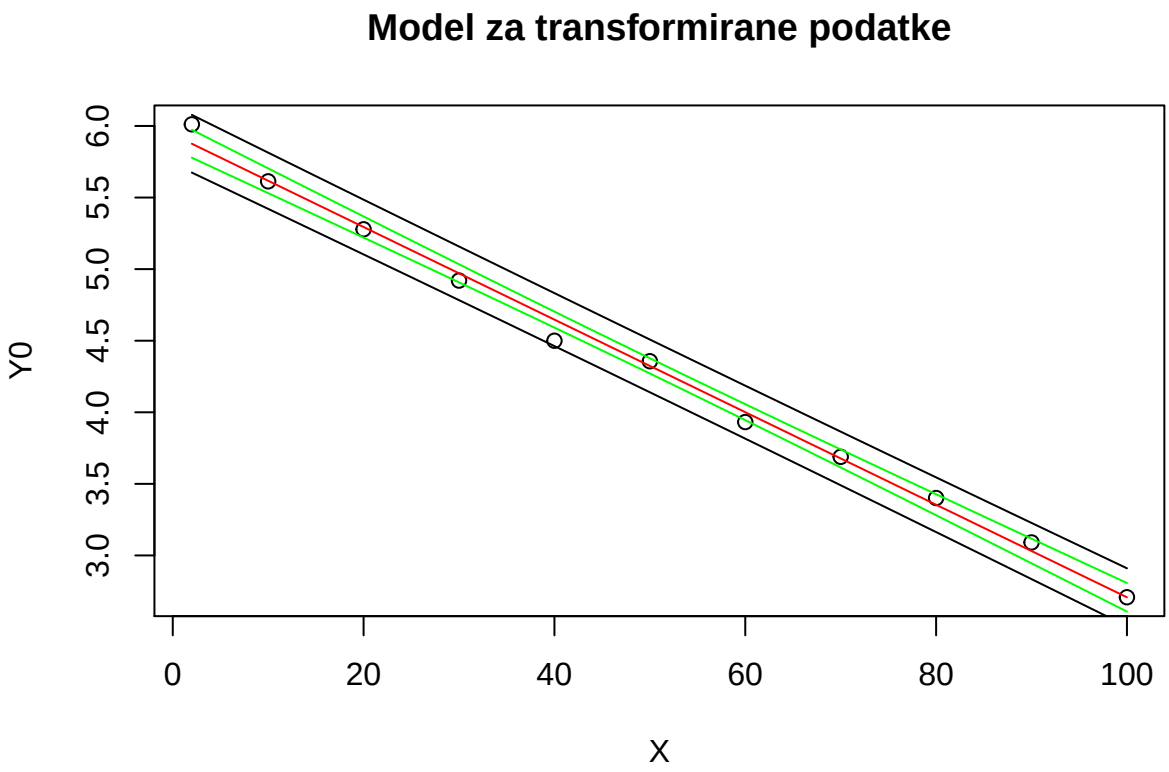
Krivulje 95% pouzdanih intervala

Na slijedećim grafovima prikazani su parovi podataka, i za originalne i za transformirane podatke. Crnom bojom su označene gornja i donja krivulja pouzdanosti za Y dok su zelenom bojom označene gornja i donja krivulja pouzdanosti za $\overline{Y}$.

```r
X.new <- seq(min(X), max(X))
X.squared.new <- X.new^2

prediction <- predict.lm(model, newdata = data.frame(X=X.new, X.squared=X.squared.new),
                         interval = 'prediction', level=0.95)
confidence <- predict.lm(model, newdata = data.frame(X=X.new, X.squared=X.squared.new),
                         interval = 'confidence', level=0.95)
plot(X, Y, main='Model za originalne podatke')
lines(x.draw, y.draw, col="red")
lines(X.new, prediction[,2])
lines(X.new, prediction[,3])
lines(X.new, confidence[,2], col="green")
lines(X.new, confidence[,3], col="green")
```




```
prediction.ln <- predict.lm(model.ln, newdata = data.frame(X=X.new),
                            interval = 'prediction', level=0.95)
confidence.ln <- predict.lm(model.ln, newdata = data.frame(X=X.new),
                            interval = 'confidence', level=0.95)
plot(X, Y0, main='Model za transformirane podatke')
lines(x.draw, y.ln.draw, col="red")
lines(X.new, prediction.ln[,2])
lines(X.new, prediction.ln[,3])
lines(X.new, confidence.ln[,2], col="green")
lines(X.new, confidence.ln[,3], col="green")
```



Linearan model za transformirane podatke odnosno eksponencijalan model za originalne podatke je zasigurno bolji. Osim što je iz grafova vidljiva bolja prilagodba, 95% pouzdani intervali za Y te srednju vrijednost od Y su puno uži. Također, statistika R^2 modela za transformirane podatke iznosi 0.9952 dok za originalne podatke iznosi 0.9663.

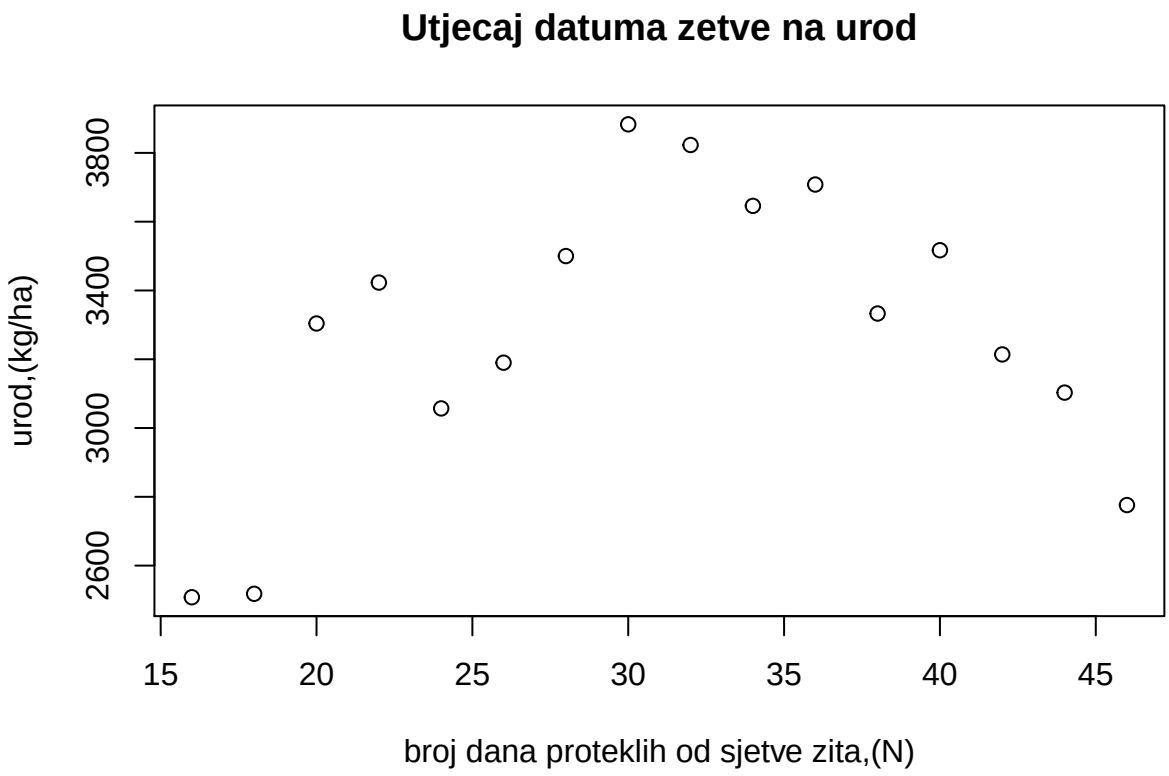
Zadatak B

U članaku "Determination of Biological Maturity and Effect of Harvesting and Drying Conditions on Milling Quality of Paddy" (J. Agricultural Eng. Research, 1975., str. 353-361) obrađuju se podaci o žetvi paddyja, vrste žita u Indiji. Varijabla x predstavlja datum žetve (tj. to je broj dana proteklih od sjetve žita), a y predstavlja urod (u kg/ha). Podaci se nalaze u datoteci zad55r.dat (Devore, Jay L., Probability and Statistics for Engineering and the Sciences, 1982., Brooks/Cole Publishing Company, Monterey, California, str. 478).

Prikaz podataka u Kartezijevom koordinatnom sustavu

Na sljedećem dijagramu prikazani su parovi podataka (x, y) iz zadanog skupa:

```
B.data <- read.table("zad55r.dat", header = TRUE, sep = " ")
B.data <- data.frame(B.data)

plot(B.data, xlab = "broj dana proteklih od sjetve zita,(N)", ylab = "urod,(kg/ha)",
     main="Utjecaj datuma zetve na urod")
```



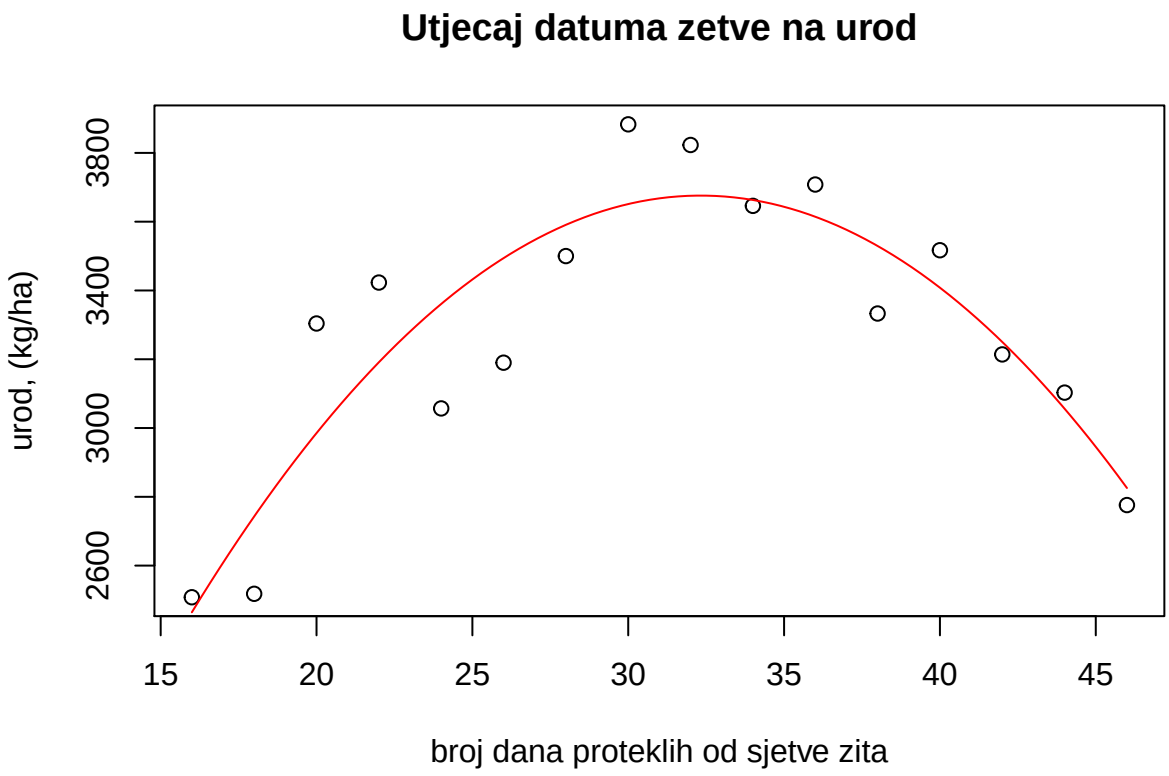
Prilagodba kvadratičnog modela

Prvi model čiju ćemo prilagodbu provesti jest sljedeći kvadratični model: $y = \theta_0 + \theta_1 x + \theta_2 x^2$

```
urod <- B.data$y
N <- B.data$x
N.squared <- I(N^2)

model <- lm(urod ~ N +  N.squared)
```

Sljedeći graf prikazuje parabolu dobivenu prilagodbom navedenog modela s empirijskim podacima:

```
f = function(x, koeficijenti)
  return(koeficijenti[[1]] + koeficijenti[[2]] * x + koeficijenti[[3]] * x^2)

plot(N, urod, xlab = "broj dana proteklih od sjetve zita", ylab = "urod, (kg/ha)",
     main="Utjecaj datuma zetve na urod")
curve(f(x, model$coefficients), add = TRUE, col = "red")
```




```
summary(model)
```

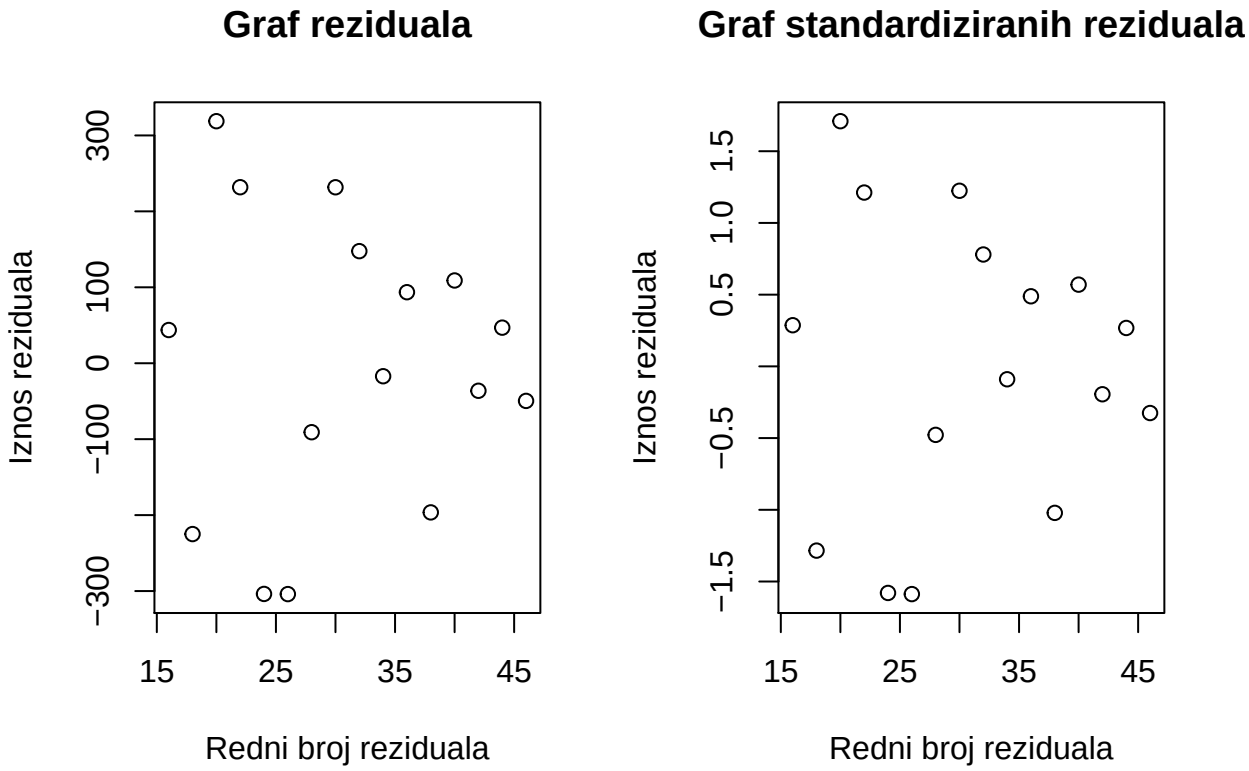
```
##
## Call:
## lm(formula = urod ~ N + N.squared)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304.00 -117.23   13.32  118.63  318.73
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1074.6320   618.0231  -1.739    0.106
## N             293.9240    42.2303   6.960 9.92e-06 ***
## N.squared      -4.5464     0.6753  -6.733 1.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 204.1 on 13 degrees of freedom
## Multiple R-squared:  0.7939, Adjusted R-squared:  0.7622
## F-statistic: 25.03 on 2 and 13 DF,  p-value: 3.483e-05
```

Vrijednost statistike R^2 se može vidjeti pozivom funkcije summary nad modelom i iznosi 0.7939.

Provjera normalnosti reziduala

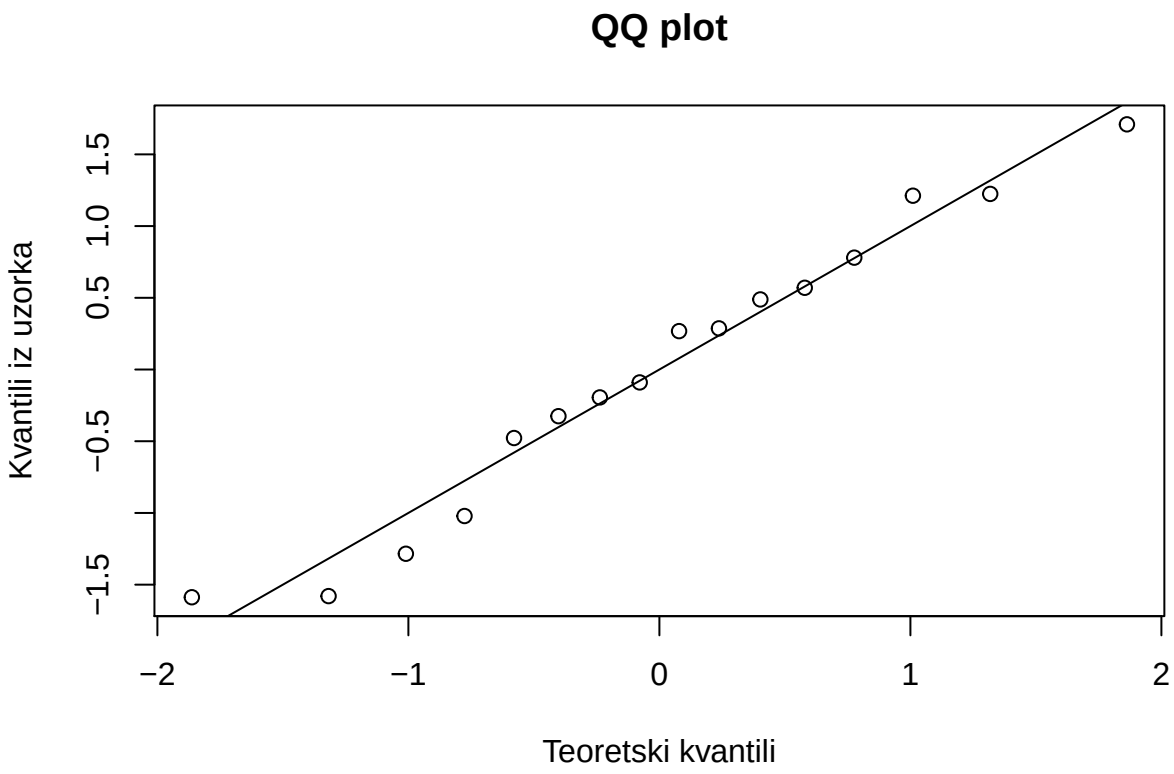
Pretpostavka linearne regresije jest da su reziduali normalno distribuirani. Radi toga radimo provjeru normalnosti na sljedeća dva načina: grafički (QQ plot) te Kolmogorov-Smirnovljevim testom.

```
par(mfrow=c(1,2))

plot(N, model$residuals, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf reziduala')

plot(N, rstandard(model), xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf standardiziranih reziduala')
```



QQ plot

```
qqnorm(rstandard(model), xlab = 'Teoretski kvantili', ylab = 'Kvantili iz uzorka',
       main = 'QQ plot')
abline(a=0, b=1)
```



Analizom dobivenog QQ plota može se pretpostaviti da reziduali dolaze iz normalne distribcije.

Kolmogorov-Smirnovljev test

```
ks.test(rstandard(model), 'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(model)
## D = 0.10553, p-value = 0.9857
## alternative hypothesis: two-sided
```

Provedbom Kolmogorov-Smirnovljevog testa dobivamo p-vrijednost jednaku 0.9857. Test na normalnost za reziduale ne može odbaciti H0 da standardizirani reziduali dolaze iz normalne razdiobe.

Hipoteza $H_0 : \theta_2 = 0$

```
summary(model)
```

```
##
## Call:
## lm(formula = urod ~ N + N.squared)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -304.00 -117.23   13.32  118.63  318.73
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1074.6320   618.0231  -1.739    0.106
## N             293.9240    42.2303   6.960 9.92e-06 ***
## N.squared      -4.5464     0.6753  -6.733 1.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 204.1 on 13 degrees of freedom
## Multiple R-squared:  0.7939, Adjusted R-squared:  0.7622
## F-statistic: 25.03 on 2 and 13 DF,  p-value: 3.483e-05
```

Testiramo sljedeću hipotezu: $H_0 : \theta_2 = 0$, u odnosu na alternativu $H_a : \theta_2 ! = 0$. Kao što je vidljivo iz izlaza funkcije summary, p-vrijednost za parametar θ_2 (koji stoji uz x^2) iznosi $1.40 \cdot 10^{-5}$. Prema tome, uz razinu značajnosti $\alpha = 1\theta_2$ je značajan koeficijent.

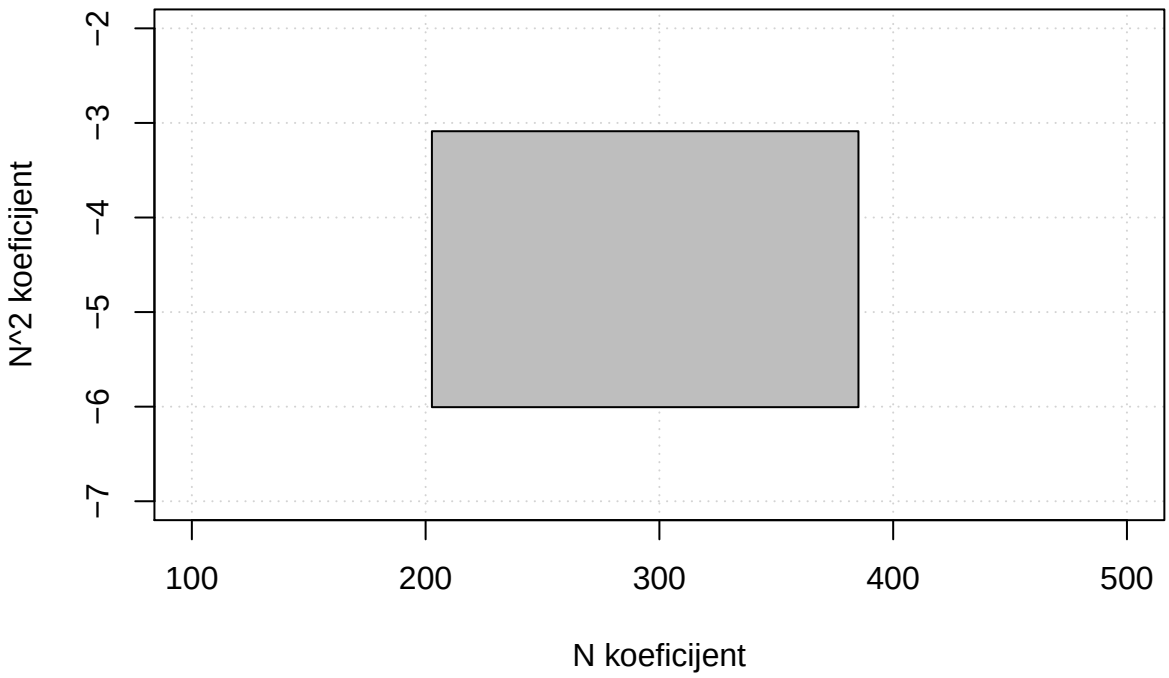
```
N.confidence = confint(model, 'N', level=0.95)
N.confidence
```

```
##     2.5 %   97.5 %
## N 202.691 385.1569
```

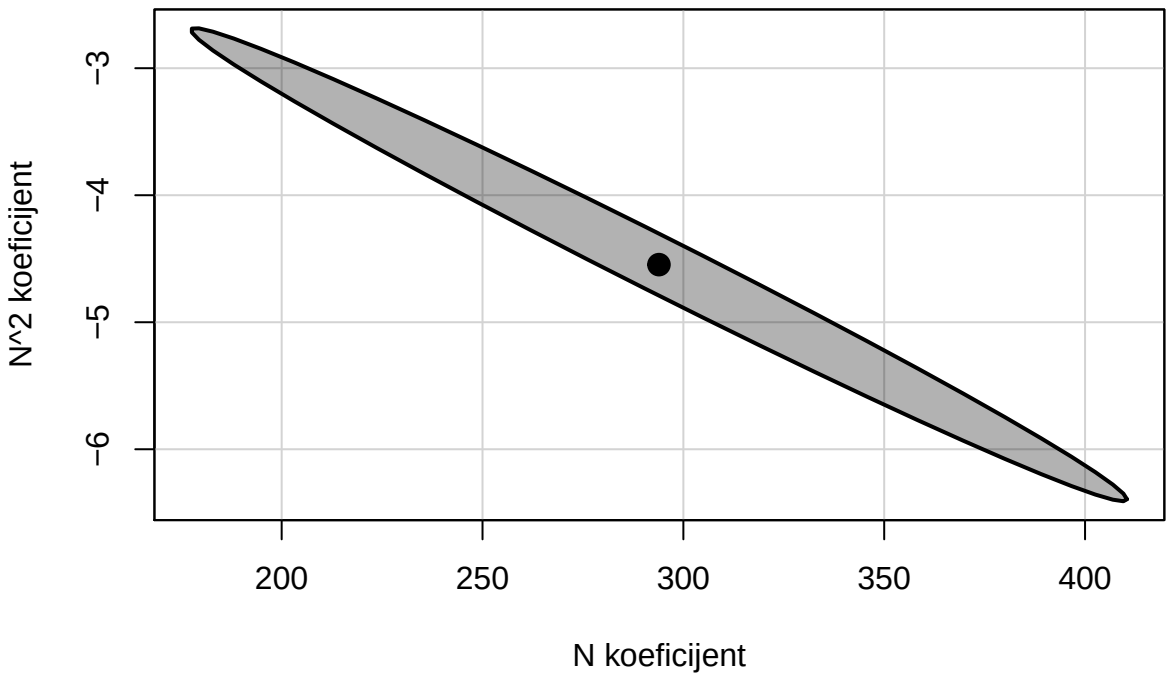
```
N.squared.confidence = confint(model, 'N.squared', level=0.95)
plot(c(100,500), type='n', c(-7,-2), panel.first=grid(),xlab = "N koeficijent", ylab = "N^2 koeficijent
rect(xleft=N.confidence[1,1], ybottom=N.squared.confidence[1,1], xright=N.confidence[1,2], ytop=N.squar

require('car')
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 3.3.3
```



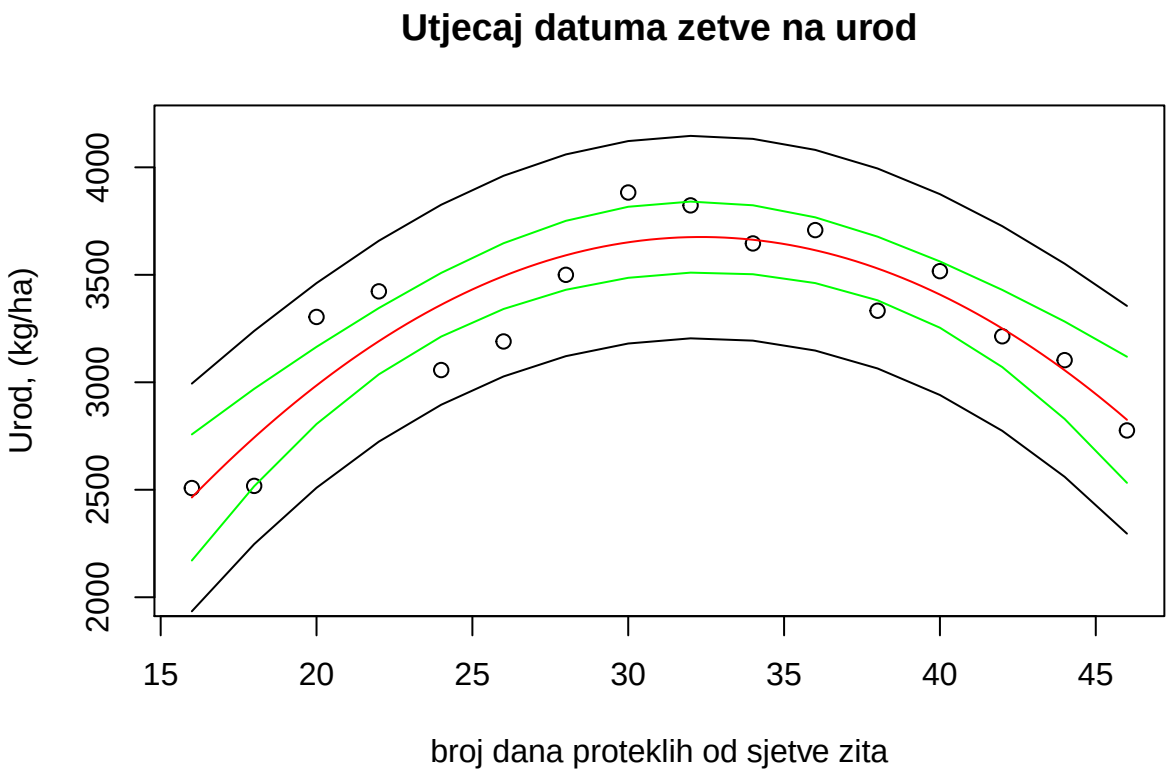

```
confidenceEllipse(model,fill=TRUE,levels=0.95,xlab = "N koeficijent", ylab = "N^2 koeficijent",col='bla
```



Iz prvog grafa u kojem uspoređujemo pouzdane intervale za parametre prihvaćenog modela vidimo kako je interval za koeficijent uz x^2 je za dva reda veličine manji od interval za koeficijent uz x. Usporedbom tog grafa i drugog grafa (prikaz pouzdanog područja za parametre modela) vidimo kako je površina elipse (95% pouzdano područje) manja od umnoška pouzdanih intervala prvog i drugog parametra.

Krivulje 95% pouzdanih intervala

```
plot(N,urod,ylim=c(2000, 4200), xlab="broj dana proteklih od sjetve zita", ylab="Urod, (kg/ha)",main="U
prediction = predict.lm(model,B.data,interval = "prediction")
confidence = predict.lm(model,B.data,interval = "confidence")

curve(f(x, model$coefficients), add = TRUE, col = "red")
lines(N, prediction[,2])
lines(N, prediction[,3])
lines(N, confidence[,2], col="green")
lines(N, confidence[,3], col="green")
```



Zaključak

Iz prethodne analize podataka možemo zaključiti da postoji kvadratna veza između uroda i broj dana proteklih od sjetve žita te time opravdavamo optimalno razdoblje žetve između 28 i 36 nakon cvatnje.

Zadatak C

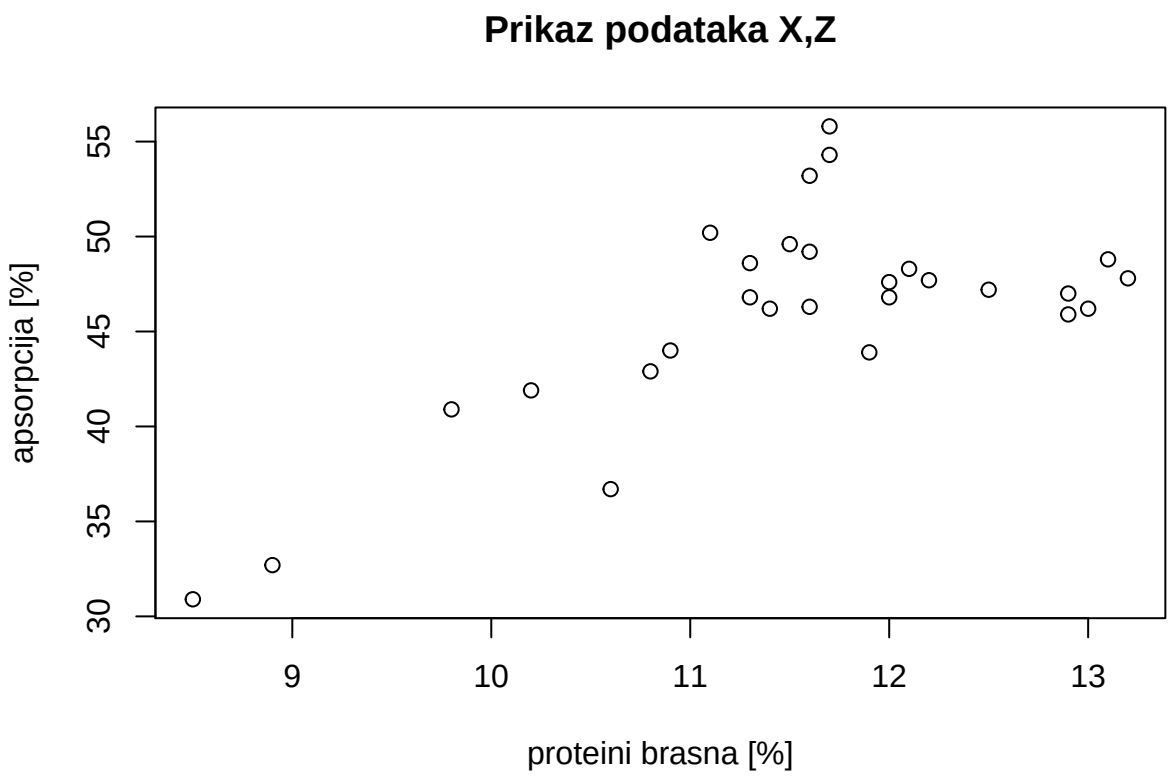
U radu "An Ultracentrifuge Flour Absorption Method" (Cereal Chemistry, 1978., str. 96-101) autori su proučavali odnos između apsorpcije vode pšeničnog brašna i raznih karakteristika tog brašna. Konkretno, promatrali su odnos između apsorpcije z (u %) te proteina brašna x (u %) i gubitka škroba y (u Farrandovim jedinicama). Podaci dobiveni pokusom nalaze se u datoteci zad57r.dat (Devore, Jay L., Probability and Statistics for Engineering and the Sciences, 1982., Brooks/Cole Publishing Company, Monterey, California, str. 490).

Prikaz podataka u Kartezijevom koordinatnom sustavu (2D i 3D)

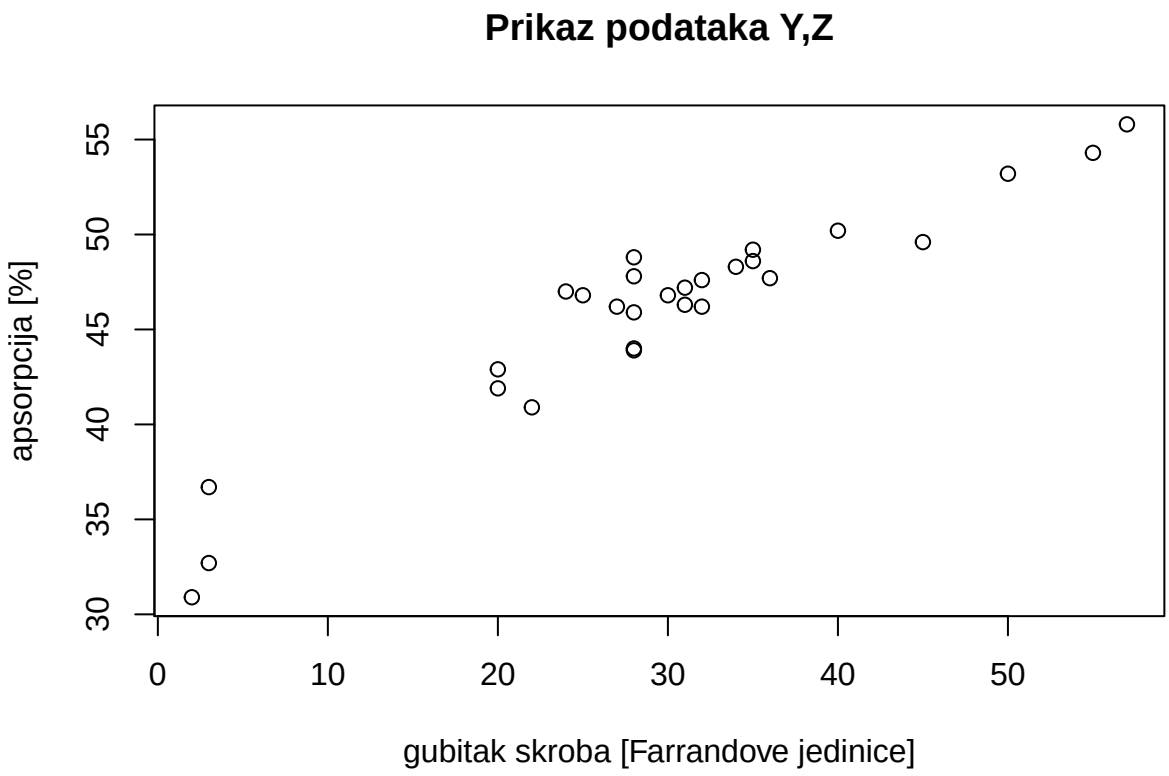
Na početku zadatka ćemo dane podatke prikazati u 2D koordinatnom sustavu (X,Y), (X,Z) i (Y,Z) te zatim sve podatke zajedno u 3D koordinatnom sustavu (X,Y,Z).

```
C.data <- read.table("zad57r.dat", header = TRUE, sep = " ")
C.data <- data.frame(C.data)

X <- C.data$x
```

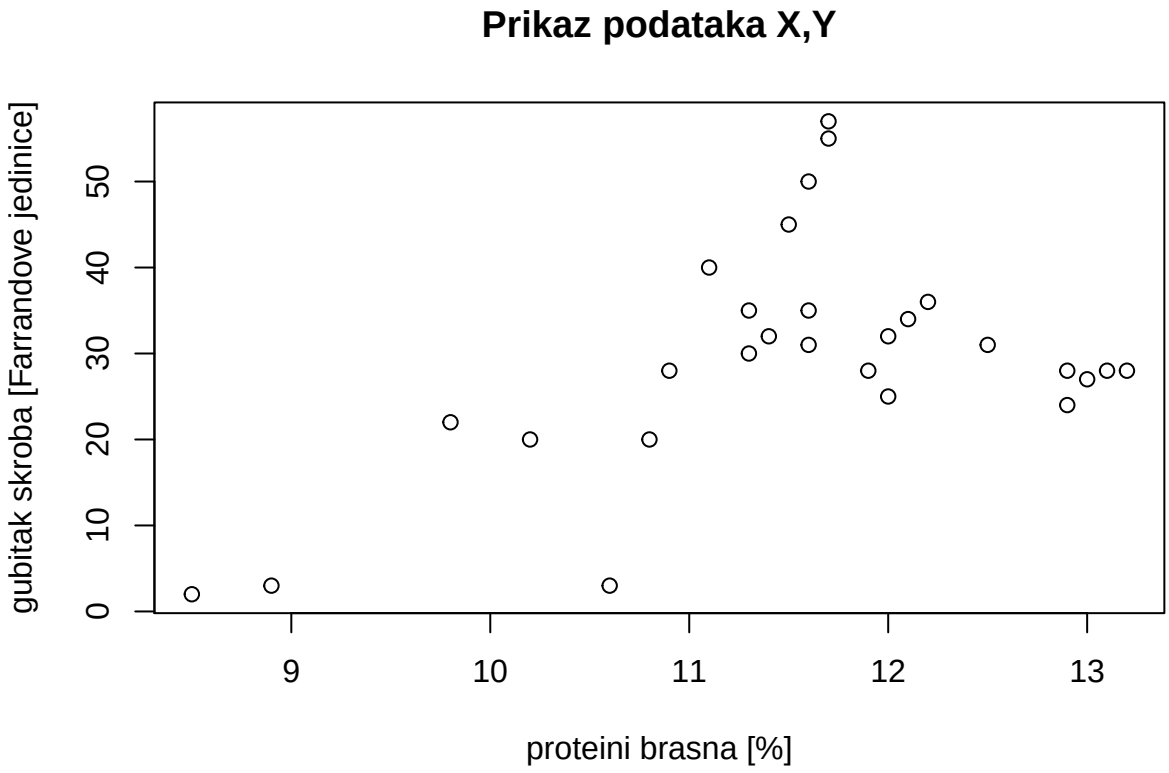
```
Y <- C.data$y
Z <- C.data$z

plot (X,Z, xlab = "proteini brasna [%]",ylab = "apsorpcija [%]",
      main="Prikaz podataka X,Z")
```




```
plot (Y,Z, xlab = "gubitak skroba [Farrandove jedinice]",ylab = "apsorpcija [%]",
      main="Prikaz podataka Y,Z")
```




```
plot (X,Y, xlab = "proteini brasna [%]",ylab = "gubitak skroba [Farrandove jedinice]",
      main="Prikaz podataka X,Y")
```

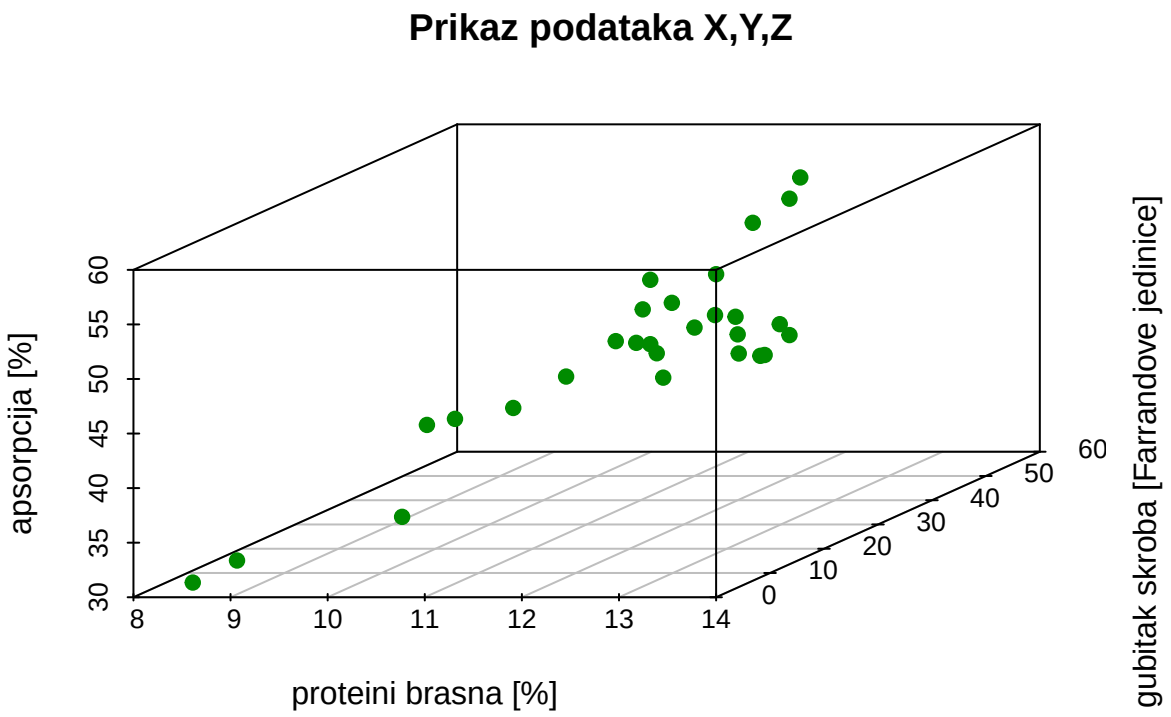



```
require(scatterplot3d)
```

```
## Loading required package: scatterplot3d
```

```
## Warning: package 'scatterplot3d' was built under R version 3.3.3
```

```
scatterplot3d(X,Y,Z, pch = 19, color = "green4",
              xlab="proteini brasna [%]",
              ylab="gubitak skroba [Farrandove jedinice]",
              zlab = "apsorpcija [%]",
              main ="Prikaz podataka X,Y,Z" )
```



Testovi korelacije

Zatim ćemo nad parom (Y,Z) provesti Pearsonov test korelacije, a nad parom (X,Y) Spearmanov test korelacije.

```
cor(Y,Z)
```

```
## [1] 0.946518
```

```
cor.test(Y,Z)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Y and Z
## t = 14.958, df = 26, p-value = 2.751e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8864776 0.9752210
## sample estimates:
##      cor
## 0.946518
```

Nad parom (Y,Z) smo radili Pearsonov test korelacije. Iz dobivenih rezultata vidimo da je $r = 0.946518$ te zaključujemo da su podaci pozitivno korelirani. P-vrijednost iz testa koreliranosti je jako mala ($2.751e-14$) pa možemo odbaciti hipotezu da je koreliranost jednaka 0. Možemo primjetiti povezanost između velike korelacije i male p-vrijednosti.

```
cor(X,Y,method = "spearman")
```

```
## [1] 0.2870111
```

```
cor.test(X,Y,method = "spearman")
```

```
## Warning in cor.test.default(X, Y, method = "spearman"): Cannot compute
## exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  X and Y
## S = 2605.3, p-value = 0.1386
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2870111
```

```
sumY = rnorm(28,0,0.01);
sumX = rnorm(28,0,0.01);

noviX=X+sumX
noviY=Y+sumY

cor(noviX,noviY,method = "spearman")
```

```
## [1] 0.2577997
```

```
cor.test(noviX,noviY,method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  noviX and noviY
## S = 2712, p-value = 0.1847
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2577997
```

Nad parom (X,Y) smo radili Spearmanov test korelacije. Pomoću rezultata naslućujemo da je korelacija između danih podataka jako mala jer je $r = 0.2870111$. P-vrijednost testa je 0.1386 pa ne možemo odbaciti hipotezu da je korelacija jednaka 0. Problem kod Spearmanovog testa je što ne može naći egzaktnu p-vrijednost ako ima jednake vrijendosti. Zbog toga smo napravili šum za obje varijable X i Y koji je iz normalne distribucije $N(0, 0.0001)$. Ponovno smo izračunali stupanj korelacije i napravili Spearmanov test. Novo dobivene vrijednosti se malo razlikuju od prijašnjih pa zbog p-vrijednosti veće od 0.05 ne možemo zaključiti da je korelacija različita od 0.

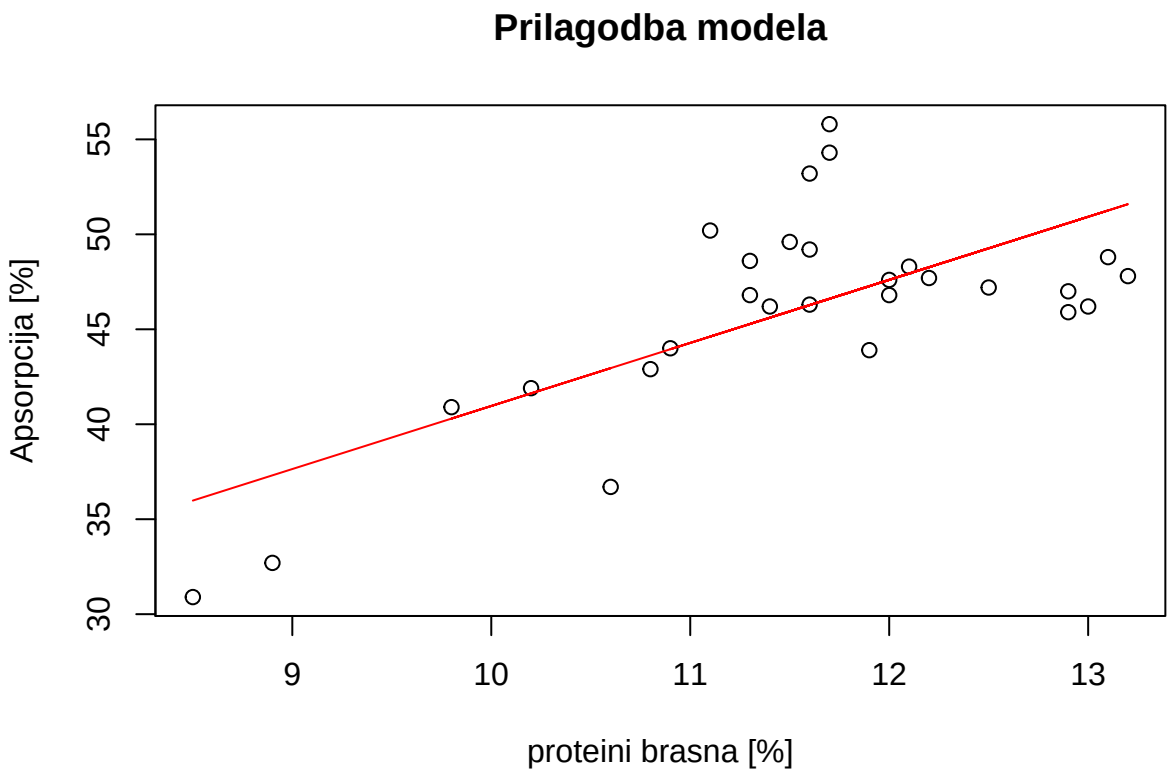
Prilagodba linearnog modela

Sljedeći korak našeg zadatka su prilagodbe linearnih modela. Prvi model čiju ćemo prilagodbu provesti jest slijedeći linearni model: $z = \alpha_0 + \alpha_1 x$

```
X <- C.data$x
Z <- C.data$z

model <- lm (Z ~ X)
```

Sljedeći graf prikazuje pravac dobiven prilagodbom navedenog modela zajedno s empirijskim podacima. Dobili smo pravac $z = 3.321 * x + 7.753$.

```
plot(X, Z, xlab = "proteini brasna [%]", ylab = "Apsorpcija [%]",
     main="Prilagodba modela")
lines(X, model$fitted.values, col="red")
```




```
summary(model)
```

```
##
## Call:
## lm(formula = Z ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2542 -3.4265  0.0108  1.8721  9.1928
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.753      7.877   0.984    0.334
## X              3.321      0.681   4.876 4.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.113 on 26 degrees of freedom
## Multiple R-squared:  0.4777, Adjusted R-squared:  0.4576
## F-statistic: 23.78 on 1 and 26 DF,  p-value: 4.656e-05
```

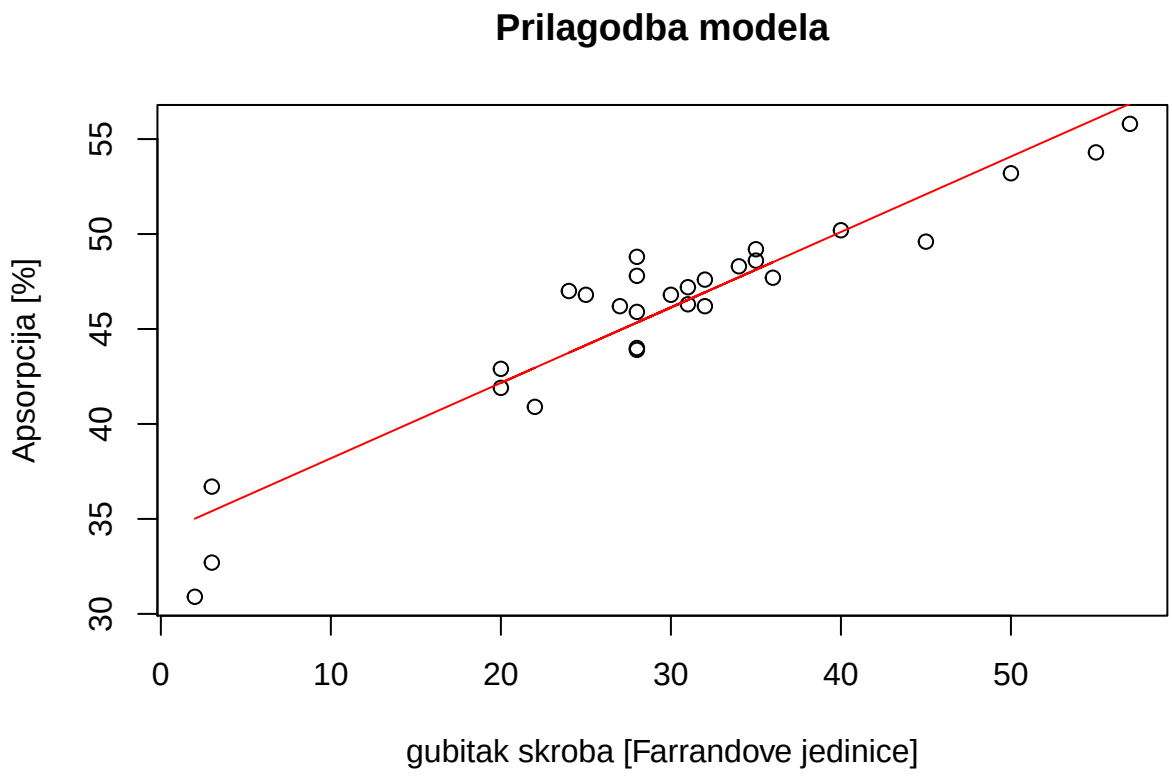
Iz sažetka iznad doznajemo vrijednost statistike R^2 koja iznosi 0.4777. Možemo biti zadovoljni dobivenim rezultatom.

Drugi model čiju ćemo prilagodbu provesti jest slijedeći linearni model: $z = \beta_0 + \beta_1 y$

```
Y <- C.data$y
Z <- C.data$z

model2 <- lm (Z ~ Y)
```

Slijedeći graf prikazuje pravac dobiven prilagodbom navedenog modela zajedno s empirijskim podacima. Dobili smo pravac $z = 0.39722 * y + 34.21809$.

```
plot(Y, Z, xlab = "gubitak skroba [Farrandove jedinice]", ylab = "Apsorpcija [%]",
     main="Prilagodba modela")
lines(Y, model2$fitted.values, col="red")
```




```
summary(model2)
```

```
##
## Call:
## lm(formula = Z ~ Y)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1125 -1.1297  0.2862  0.8230  3.4598
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 34.21809    0.85941   39.82  < 2e-16 ***
## Y            0.39722    0.02655   14.96 2.75e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.836 on 26 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8919
## F-statistic: 223.8 on 1 and 26 DF,  p-value: 2.751e-14
```

Iz sažetka pripradnog modela možemo očitati vrijednost statistike R^2 koja iznosi 0.8959. Primjećujemo da je vrijednost statistike R^2 veća nego za prvi model čime možemo naslutiti da će koeficijent uz varijablu x biti veći nego uz y u zadnjem modelu.

Prilagodba linearnog modela s dvije nezavisne varijable

Nakon što smo radili prilagodbe s jednom nezavisnom varijablom radimo prilagodbu s dvije nezavisne varijable. Sljedeći model čiju ćemo prilagodbu provesti jest sljedeći linearni model: $z = \theta_0 + \theta_1 x + \theta_2 y$

```
X <- C.data$x
Y <- C.data$y
Z <- C.data$z

model3 <- lm (Z ~ Y + X)
```

Značajnost dobivenog modela

```
summary(model3)
```

```
##
## Call:
## lm(formula = Z ~ Y + X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10062 -0.60544 -0.03045  1.00419  1.66205
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.43976    2.18829   8.884 3.30e-09 ***
## Y            0.33563    0.01814  18.507 4.18e-16 ***
## X            1.44228    0.20764   6.946 2.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.094 on 25 degrees of freedom
## Multiple R-squared:  0.9645, Adjusted R-squared:  0.9616
## F-statistic: 339.3 on 2 and 25 DF,  p-value: < 2.2e-16
```

Iz sažetka očitavamo da je vrijednost statistike $R^2 = 0.9645$. Pošto su p-vrijednosti uz sve koeficijente jako male (X,Y) možemo zaključiti da su oba koeficijenta iznimno značajna za naš model.

Uspoređivanje proširenog modela s reduciranima koristeći ANOVU

```
anova(model,model3)
```

```
## Analysis of Variance Table
##
## Model 1: Z ~ X
## Model 2: Z ~ Y + X
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     26 439.93
## 2     25  29.93  1    410.01 342.5 4.181e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model2,model3)
```

```
## Analysis of Variance Table
##
## Model 1: Z ~ Y
## Model 2: Z ~ Y + X
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     26 87.687
## 2     25 29.928  1     57.76 48.25 2.789e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

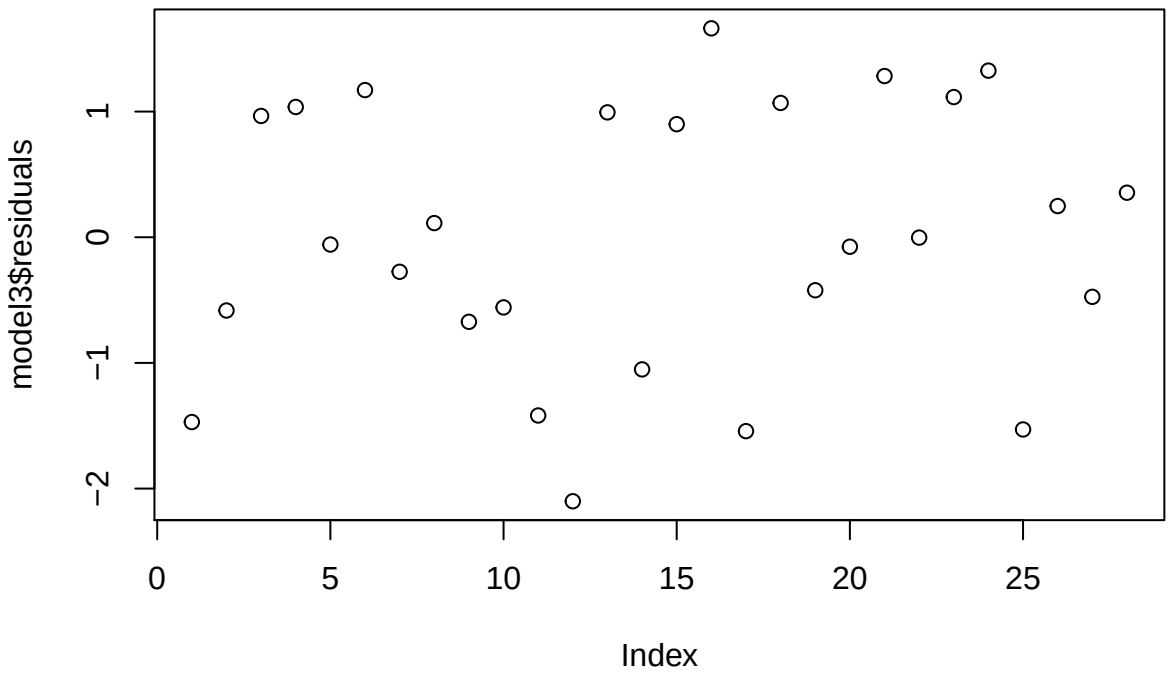
Koristeći ANOVU usporedili smo prvi model $z = \alpha_0 + \alpha_1 x$ i treći model $z = \theta_0 + \theta_1 x + \theta_2 y$. Pošto smo dobili da je p-vrijednost jednaka 4.181e-16 možemo odbaciti hipotezu da je novododani koeficijent jednak 0. Zaključujemo da je prošireni model bolji od reduciranog. Zatim smo uspredili drugi model $z = \beta_0 + \beta_1 y$ i treći model $z = \theta_0 + \theta_1 x + \theta_2 y$. Ponovno smo dobili jaku malu p-vrijednost koja iznosi 2.789e-07 te možemo odbaciti hipotezu da podaci podržavaju reducirani oblik.

Provjera normalnosti reziduala

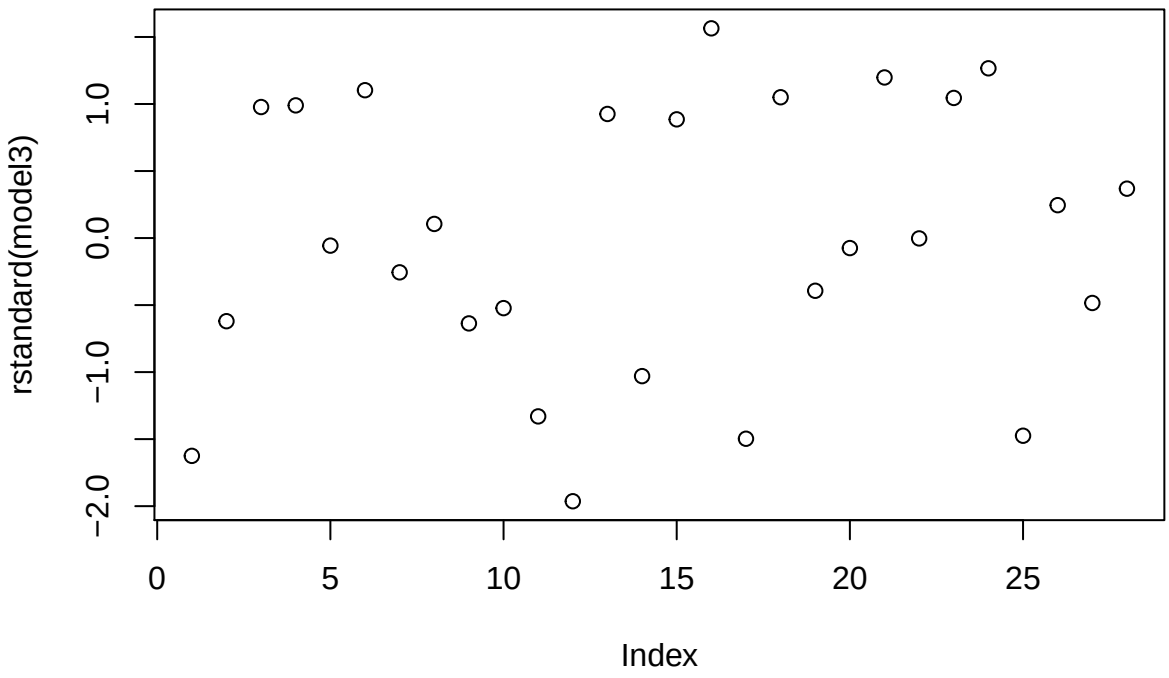
Sljedeće što ćemo napraviti je provjeriti normalnost reziduala. Koristimo ćemo sljedeća dva kriterija: grafički (QQ plot) te Kolmogorov-Smirnovljev test.

Prvo ćemo nacrtati grafove reziduala i standardiziranih reziduala.

```
plot(model3$residuals)
```

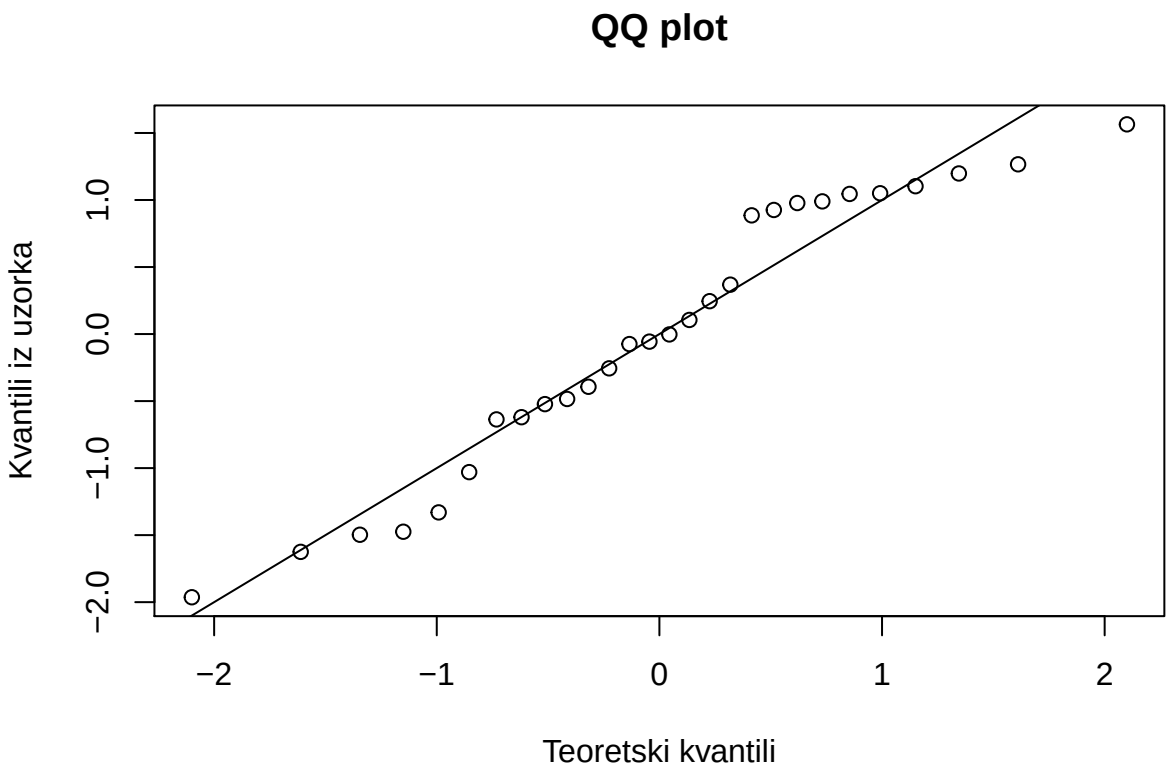



```
plot(rstandard(model3))
```



QQ plot

```
qqnorm(rstandard(model3), xlab = 'Teoretski kvantili', ylab = 'Kvantili iz uzorka',
       main = 'QQ plot')
abline(a=0,b=1)
```



Analizom dobivenog QQ plota možemo naslutiti da reziduale najvjerojatnije dolaze iz normalne razdiobe. Kolmogorov-Smirnovljev test bi to trebao potvrditi jer nam je veličina uzorka jako mala.

Kolmogorov-Smirnovljev test

```
ks.test(rstandard(model3), 'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(model3)
## D = 0.16921, p-value = 0.3584
## alternative hypothesis: two-sided
```

Rezultati Kolmogorov-Smirnov potvrđuju nam da podaci dolaze iz normalne razdiobe pošto p-vrijednost iznosi 0.3584. Interpretacija dane p-vrijednosti je da ne možemo odbaciti hipotezu da podaci dolaze iz normalne razdiobe.

Plohe 95% pouzdanih intervala

Na sljedećem grafu su prikazani originalni podaci zajedno s plohama za intervale predikcije (siva boja) i intervale pouzdanosti (crvena boja).

```
require(scatterplot3d)
require(plot3D)
```

```
## Loading required package: plot3D
```

```
## Warning: package 'plot3D' was built under R version 3.3.3
```

```
#scatterplot3d(X,Y,Z, pch = 19, color = "green4", main="3D Scatterplot")
X.new = seq(min(X), max(X),length.out = 1000)
Y.new = seq(min(Y), max(Y),length.out = 1000)

prediction <- predict.lm(model3, newdata = data.frame(X=X.new, Y=Y.new),
                         interval = 'prediction', level=0.95)
confidence <- predict.lm(model3, newdata = data.frame(X=X.new, Y=Y.new),
                         interval = 'confidence', level=0.95)


scatter3D(X, Y, Z, colvar = NULL, col = "blue",
          pch = 19, cex = 0.5)
scatter3D(X.new,Y.new,prediction[,2], add = TRUE, colkey = FALSE,
          pch = 18, cex = 3, col = "gray")
scatter3D(X.new,Y.new,prediction[,3], add = TRUE, colkey = FALSE,
          pch = 18, cex = 3, col = "gray")
scatter3D(X.new,Y.new,confidence[,2], add = TRUE, colkey = FALSE,
          pch = 18, cex = 3, col = "red")
scatter3D(X.new,Y.new,confidence[,3], add = TRUE, colkey = FALSE,
          pch = 18, cex = 3, col = "red")
```