Linearna regresija

Borna Bešić, Tomislav Buhiniček, Nikola Zadravec

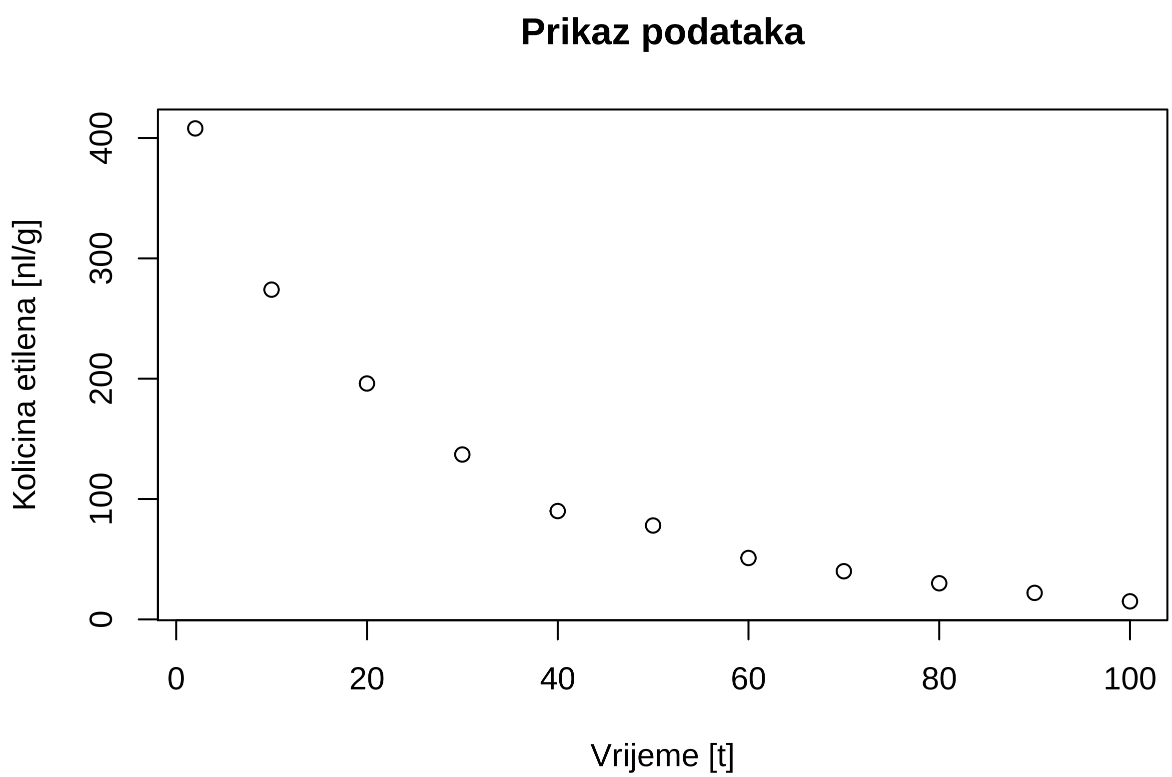
26. svibnja 2017.

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:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$

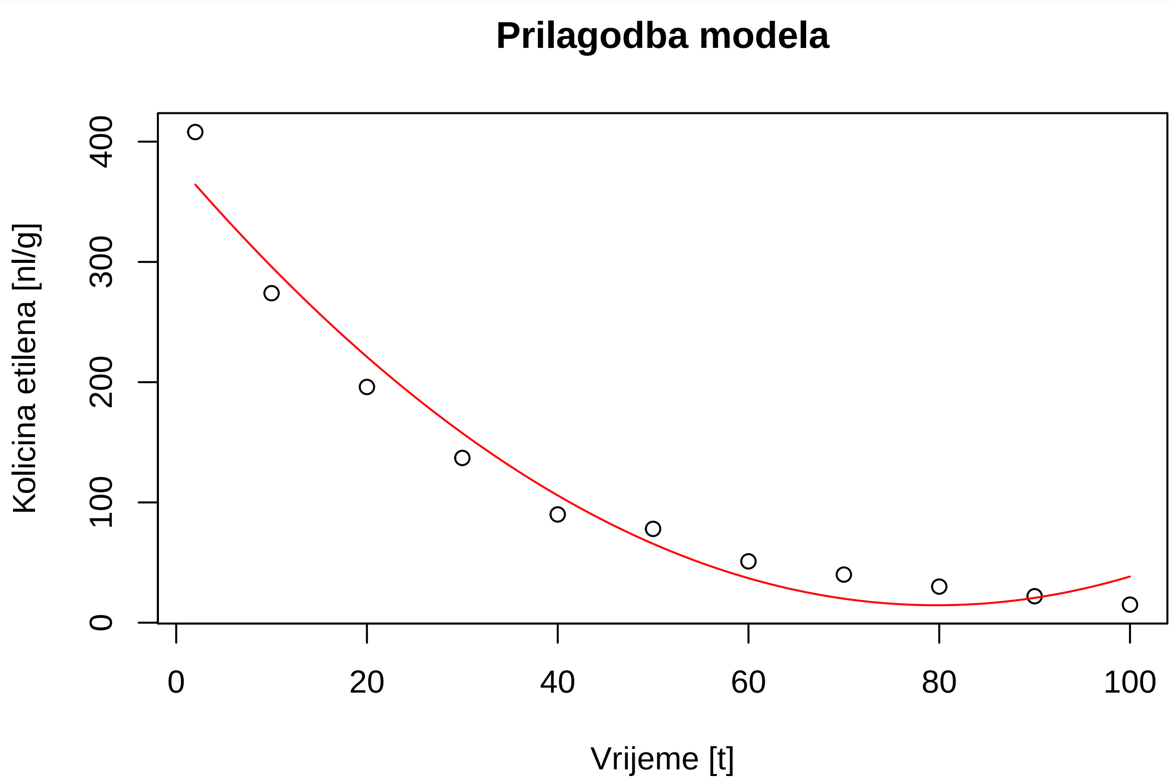
```
X <- A.data$x
Y <- A.data$y
X.squared <- X^2

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
```

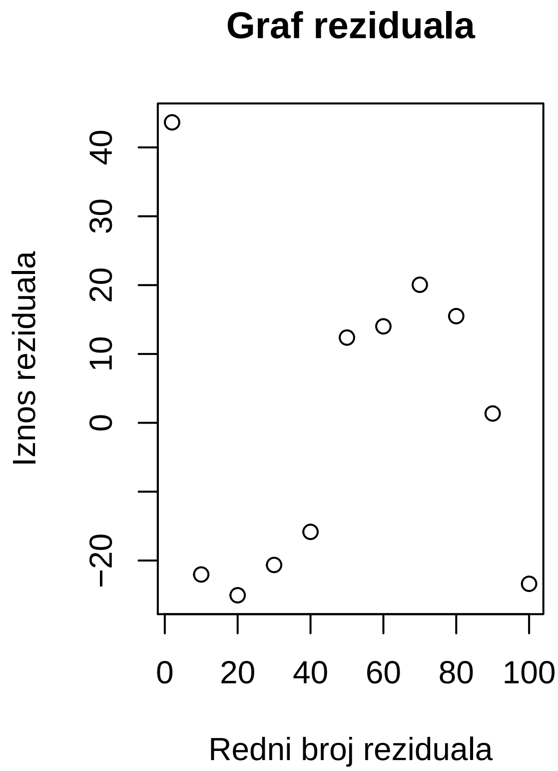
Testiramo slijedeću hipotezu: $\theta_2 = 0$, uz dvostranu alternativu. Kao što je vidljivo, p-vrijednost za parametar θ_2 (koji stoji uz x^2) iznosi $6.65 \cdot 10^{-8}$. Prema tome, uz razinu značajnosti $\alpha = 5\%$, odbacujemo nultu hipotezu u korist alternative.

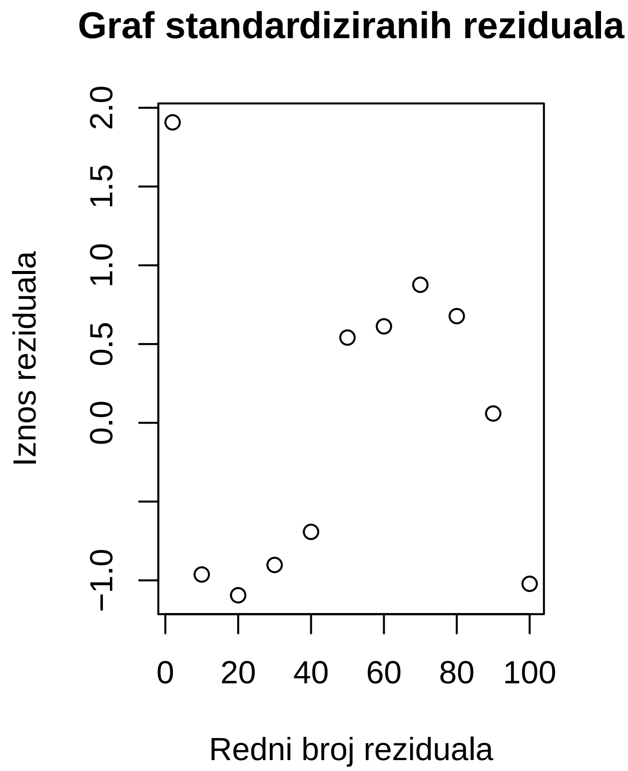
Također, iz priloženog sažetka doznajemo vrijednost statistike R^2 koja iznosi 0.9663. To je prilično zadovoljavajuća vrijednost iako možemo bolje kao što ćemo vidjeti u nastavku.

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.

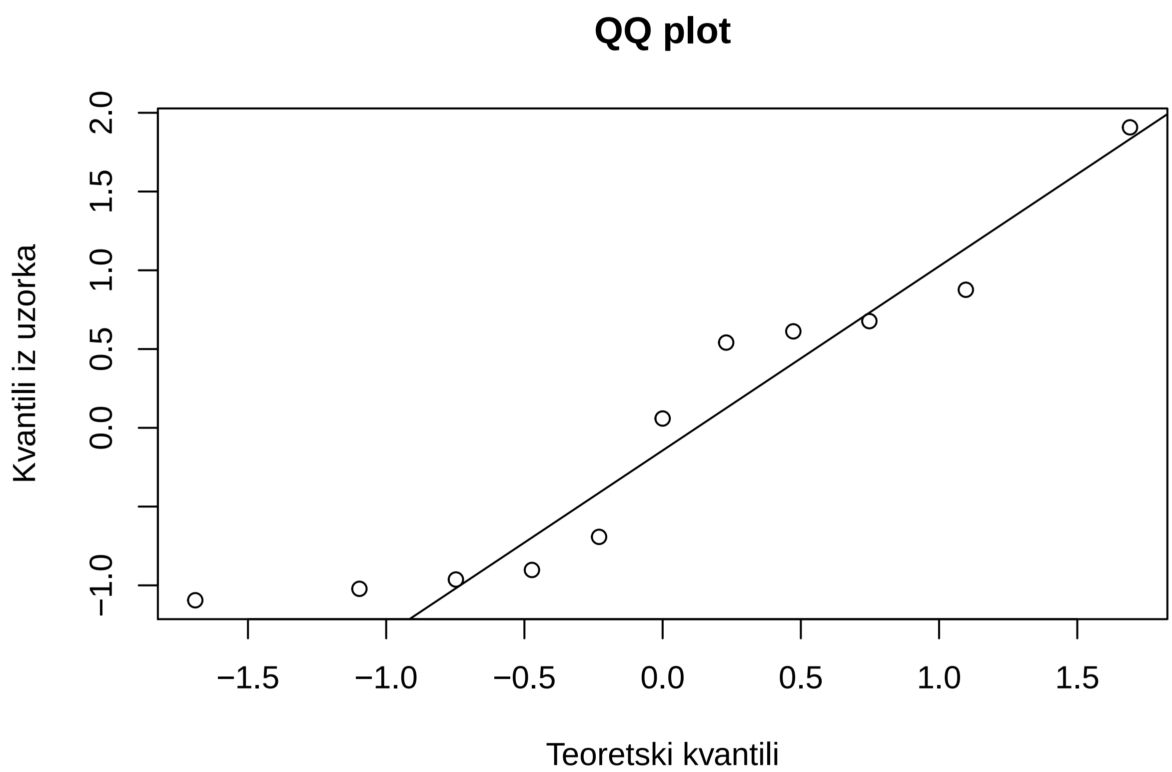
```r
par(mfrow=c(1,2))
plot(X, model$residuals, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf reziduala')
residuals.standardized <- scale(model$residuals)
plot(X, residuals.standardized, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf standardiziranih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





QQ plot

```
qqnorm(residuals.standardized, xlab = 'Teoretski kvantili', ylab = 'Kvantili iz uzorka',
       main = 'QQ plot')
qqline(residuals.standardized)
```



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(model$residuals, 'pnorm')
```

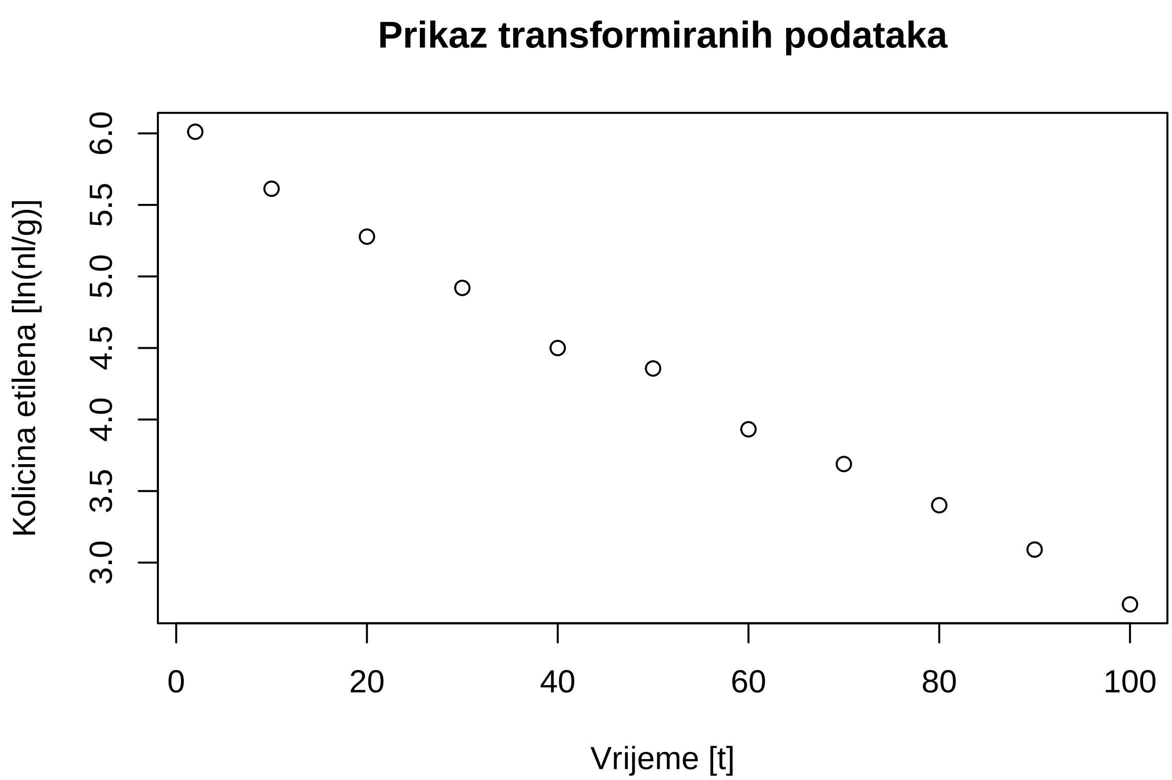
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  model$residuals
## D = 0.45741, p-value = 0.01267
## alternative hypothesis: two-sided
```

Provedbom Kolmogorov-Smirnovljevog testa dobivamo p-vrijednost jednaku 0.01267. Sa razinom značajnosti $\alpha = 5\%$ možem 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: $y^0 = ln(y)$.

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



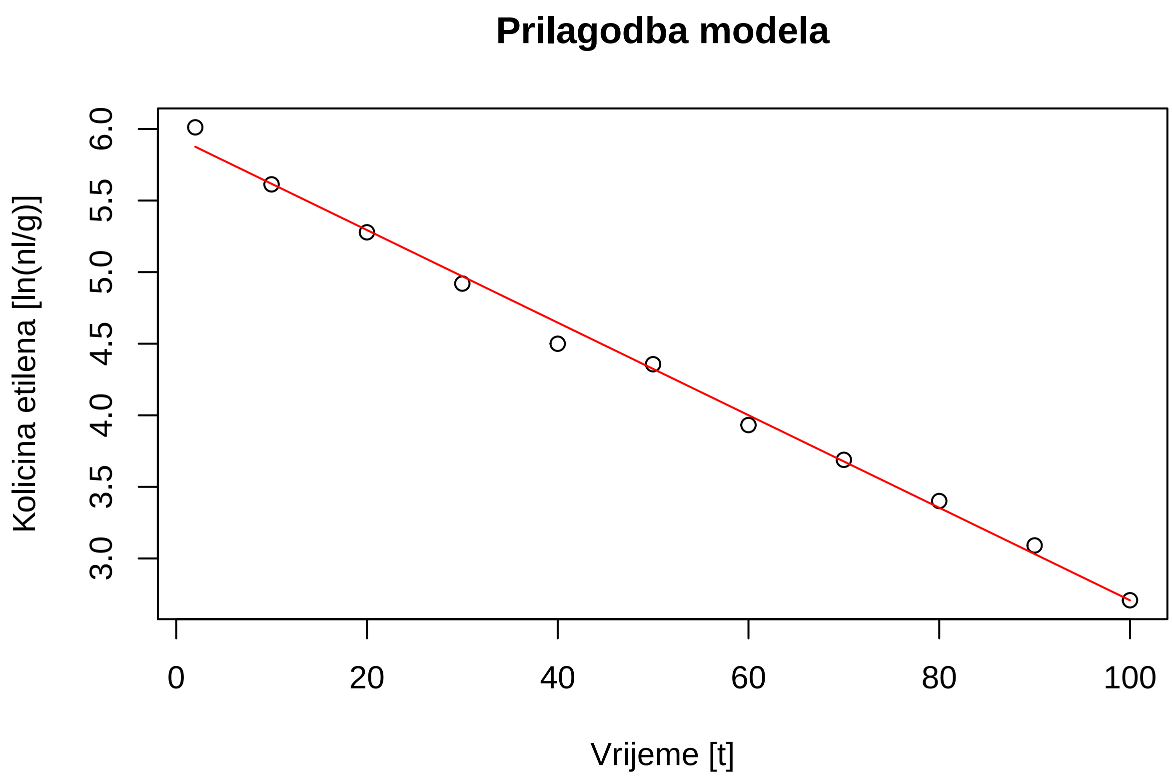
Kao što je vidljivo na dijagramu raspršenja, nakon transformacije podaci izgledaju puno bolje. Model kojeg ćemo u ovom slučaju iskoristiti glasi:

$$y^0 = \theta_0 + \theta_1 x$$

```
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")
```




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

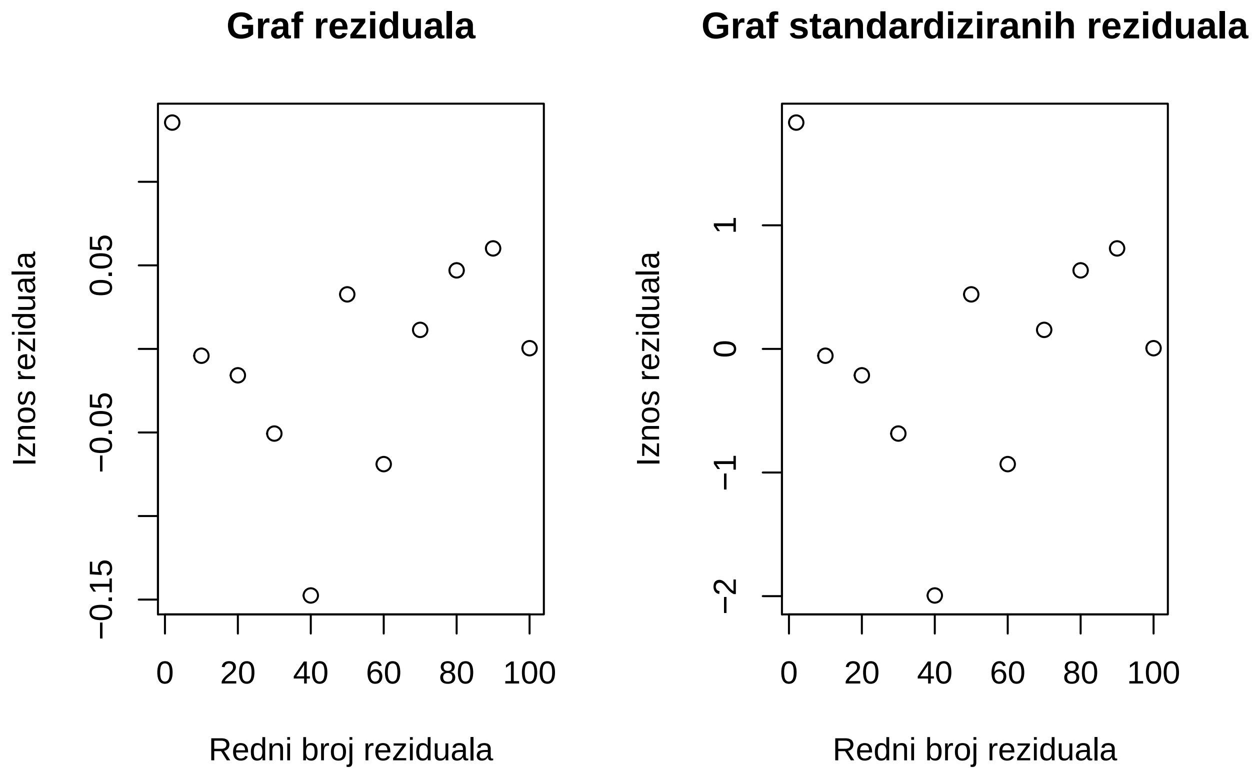
```
##
## Call:
## lm(formula = Y0 ~ X)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.147537 -0.033230  0.000425  0.039823  0.135430
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9404951  0.0443816   133.8 3.68e-16 ***
## X           -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.

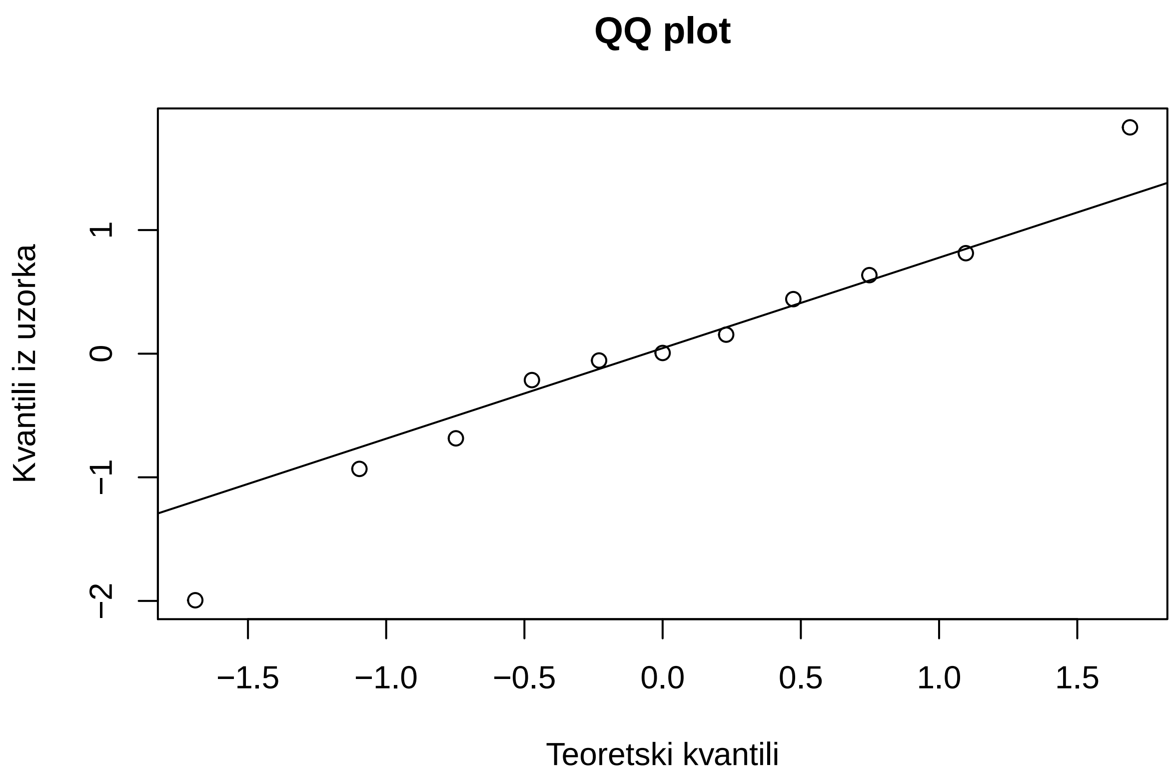
```
par(mfrow=c(1,2))
plot(X, model.ln$residuals, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf reziduala')
residuals.ln.standardized <- scale(model.ln$residuals)
plot(X, residuals.ln.standardized, xlab='Redni broj reziduala', ylab = 'Iznos reziduala',
     main = 'Graf standardiziranih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



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

Kolmogorov-Smirnovljev test

```
ks.test(model.ln$residuals, 'pnorm')
```

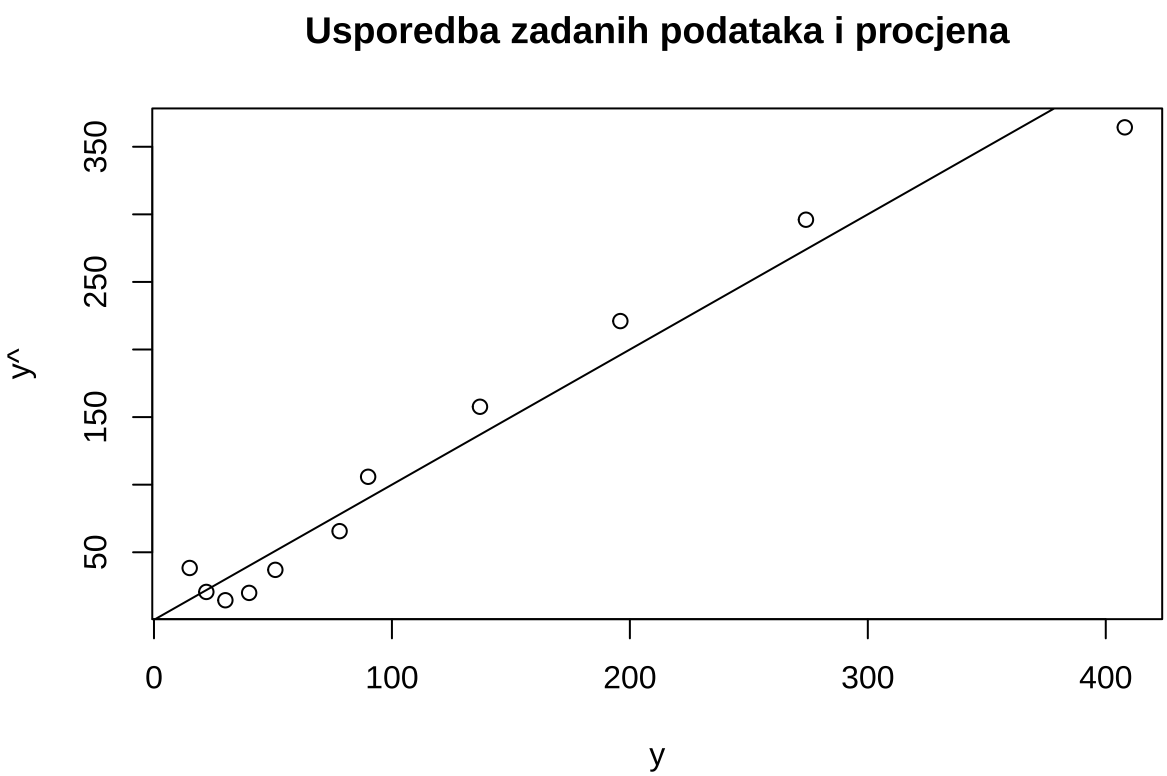
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  model.ln$residuals
## D = 0.44614, p-value = 0.0163
## alternative hypothesis: two-sided
```

Ipak, provedbom Kolmogorov-Smirnovljevog testa nad transformiranim podacima, dobivamo p-vrijednost u iznosu 0.0163. Slijedi da ćemo unatoč svemu na razini značajnosti od $\alpha = 5\%$ odbaciti nultu hipotezu da su reziduali normalno distribuirani u korist dvostrane alternative.

Model za originalne podatke

$$\hat{y} = 382.60556492 - 9.23726255 \cdot x + 0.05795002 \cdot x^2$$

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

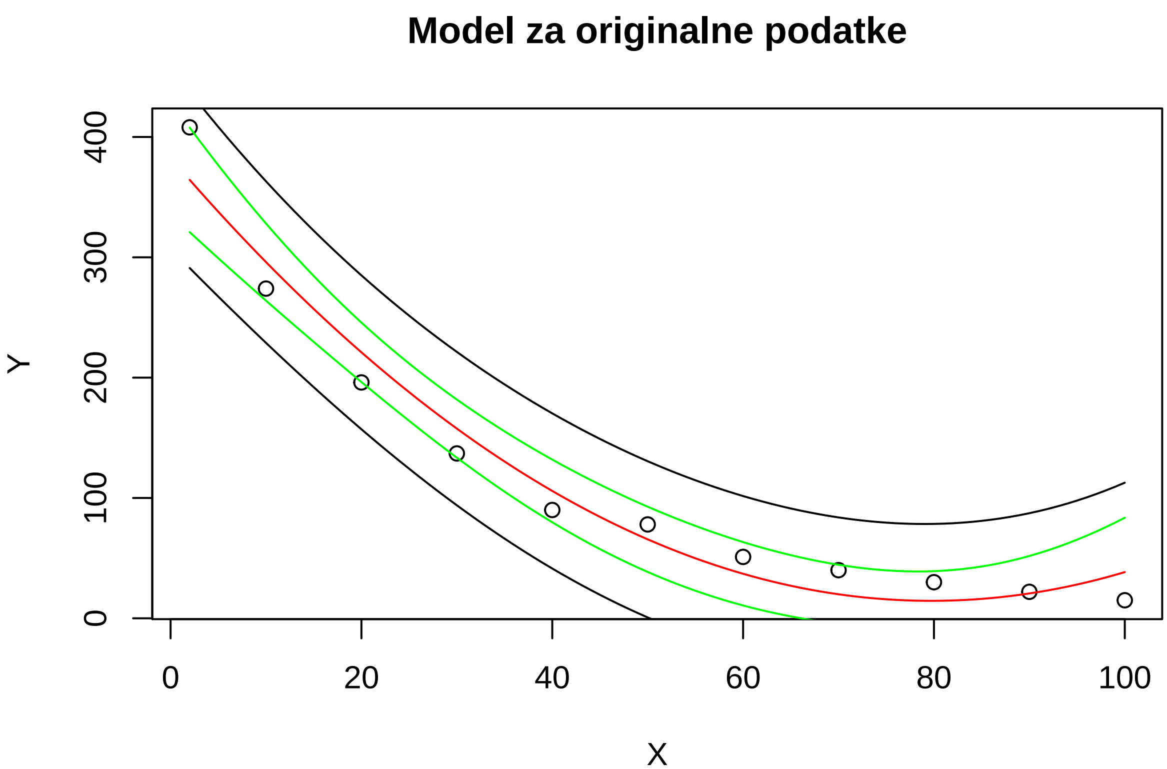


Krivulje 95% pouzdanih intervala

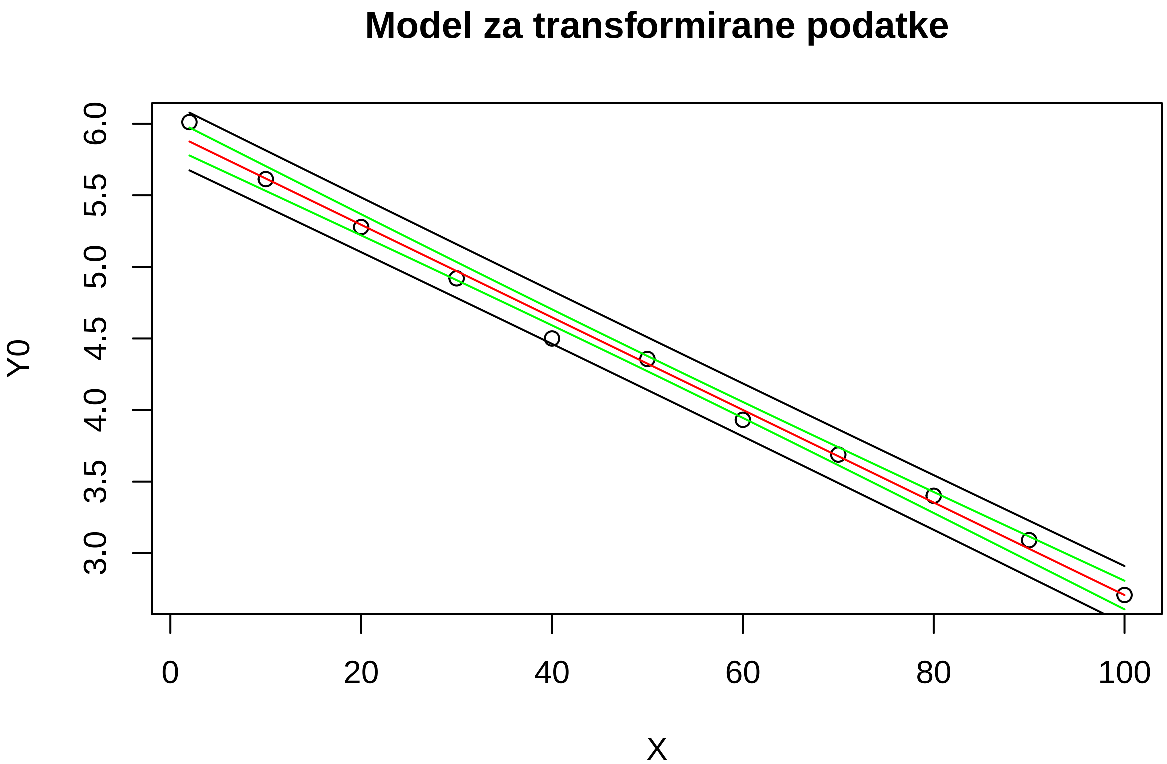
Na slijedećim grafovima prikazani su parovi podataka, i za originalne i za transformirane podatke. Crnom bojom su označene gornja i donja krivulja pouzdanosti za Y dok su zelenom bojom označene gornja i donja krivulja pouzdanosti za $\overline{Y}$.

```
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")
```




```r
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")
```



Zadatak B

Zadatak C