Linearna regresija - Zadatak B

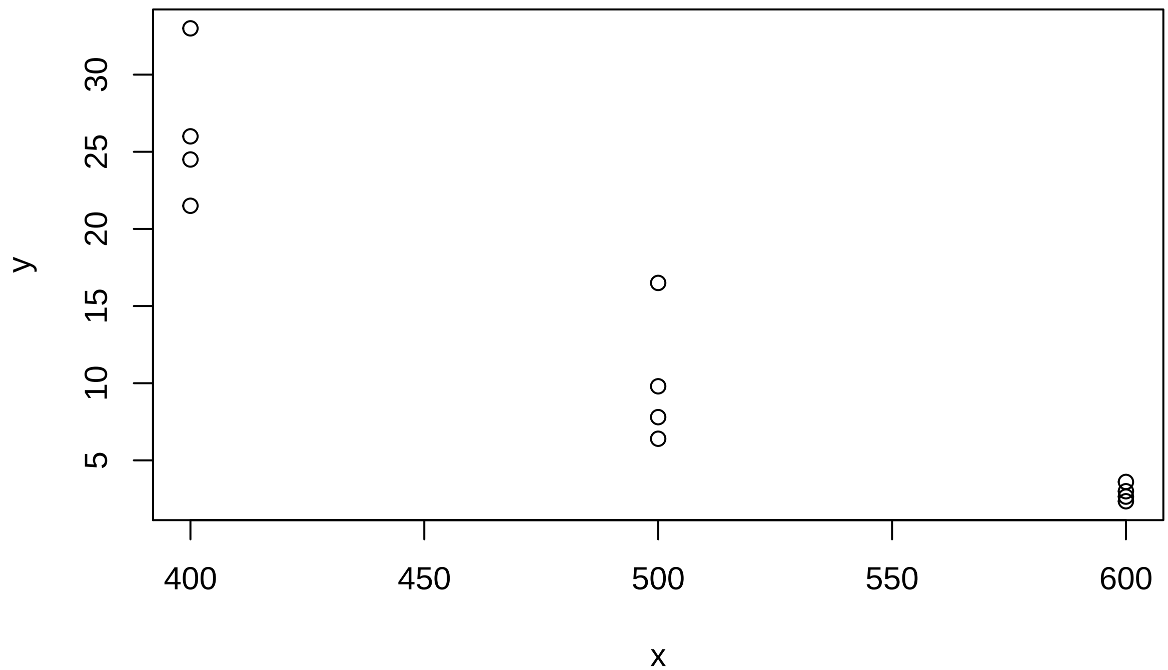
Dominik Stipić

May 4, 2019

```
data = c(600, 2.35
        ,600, 2.65
        ,600, 3.00
        ,600, 3.60
        ,500, 6.40
        ,500, 7.80
        ,500, 9.80
        ,500 ,16.50
        ,400, 21.50
        ,400, 24.50
        ,400, 26.00
        ,400, 33.00)
i = 1:length(data)
y.index = 0 == i%%2
x.index = 1 == i%%2

x = data[x.index]
y = data[y.index]
```

(a) Prikaz podataka u Kartezijevom koordinatnom sustavu

```
plot(x,y)
```



Linearna regresija

Linearne regresija je tehnika namjenjena pronalaženju linerane funkcije koja najbolje aproksimira povezanost između danih podataka. Jednostavni linearni regresijski model glasi:

$$y = \sum_{i=1}^{n} \beta_i x_i + \varepsilon$$

U gornjoj formuli varijabla y naziva se reakcija i to je varijabla koja se pokušava predvidjeti na temelju varijabla x, takozvanih regresora. U stvarnosti, ne mora postojati linearna ovisnost između regresora i reakcije, pa stoga skučajna varijabla ε predstavlja pogrešku koja nastaje zbog nedovoljno informacija o sustava, a i zbog nasumičnosti koje nastaju uzorkovanjem. Slučajne varijable β predstavljaju parametre koje je potrebno procijeniti. Pretpostavimo linearnu ovisnost između regresora i rekacije, tada je procijena reakcije:

$$\hat{y} = \sum_{i=1}^{n} b_i x_i$$

U našem slučaju linearni model glasi: $\hat{y} = b_0 + b_1 x$

Jednostavna linearna regresija ima 3 pretpostavke o podacima:

- očekivanje $E(Y_i)$ je linearna funkcija od x_i
- šumovi ϵ_i su nezavisni
- šumovi ϵ_i u točki x_i imaju $N(0, \sigma^2)$ distribuciju

Potrebno je pronaći parametre b_i koji najbolje opisuju stvarne parametre β_i. Za dani set podataka (x_i, y_i) i za model $\hat{y} = \sum_{i=1}^{n} b_i x_i$, definiramo i-ti residual kao:

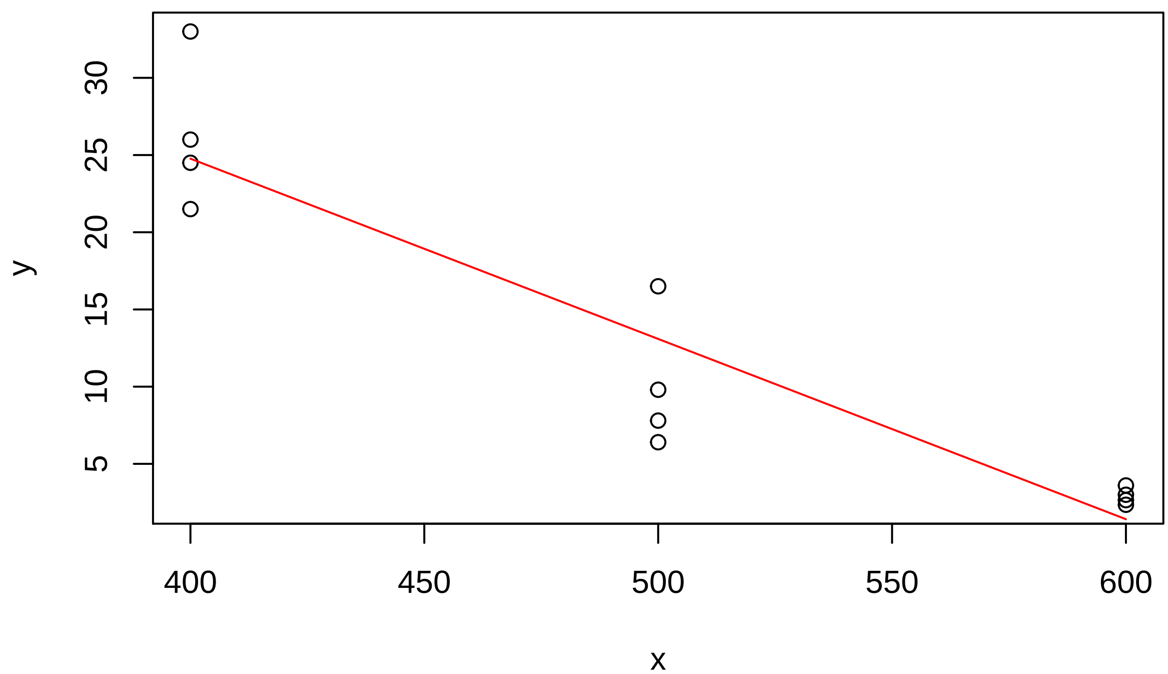
$$e_i = y_i - \hat{y_i}$$

. Tada metodom najmanjih kvadrata možemo pronaći parametre b_0, b_1 i dobiti regresijsku krivulju. Parametre računamo na način da minimiziramo residualnu sumu kvadrata:

$$SSE = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (yi - \hat{y})^2$$

.

```
plot(x,y)
fit = lm(y~x)
coeff = fit$coefficients
lines(x,fit$fitted.values,col='red')
```



Mjere kvalitete prilagodbe - Koeficijent determinacije

Koeficijent determinancije mjeri proporciju varijabilnosti koje se može objasniti pomoću prilagođenog modela. Koristi se za određivanje kvalitete kojom se model prilagođava podacima.

Računa se kao :

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

.

$$SSR = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SST = SSE + SSR$$

.

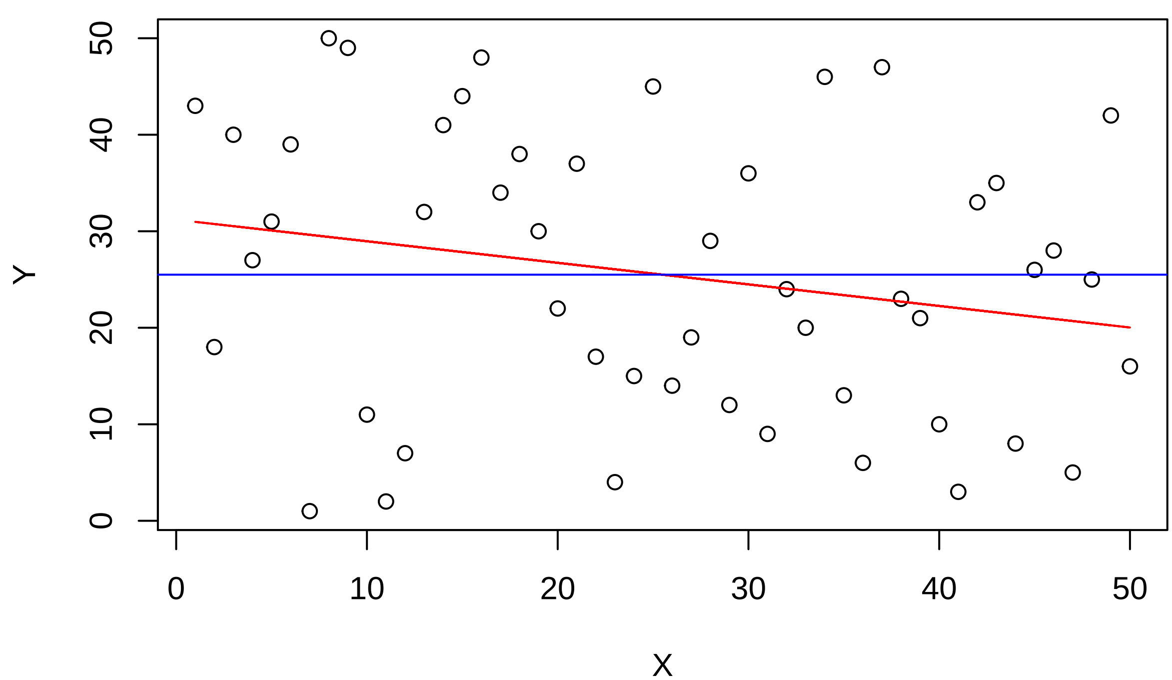
$$R^2 = 1 - \frac{SSE}{SST}$$

.

SST odstupanje može se razdvojiti u dvije sume: SSE i SSR. SSR predstavlja varijancu u podacima koju linearni model dobro opisuje, dok SSE predstavlja varijabilnost podataka oko regresijske linije. Za $R^2 \approx 1$ prilagodba je savršena, to jest ne postoji nikakav šum u podacima i imamo determinističan model. Ako $R^2 \approx 0$ tada imamo potpunu nepovezanost između regresora i reakcija.

Primjer

```r
# Uzorkujemo dvije potpuno nepovezane varijable
X = sample(50)
Y = sample(50)
plot(X,Y)
rand.fit = lm(Y~X)
lines(X,rand.fit$fitted.values,col='red')

SSE = sum(rand.fit$residuals**2)
SSR = sum((mean(Y) - rand.fit$fitted.values)**2)
SST = SSE + SSR
abline(h=mean(Y),col='blue')
```



```r
R = 1 - SSE/SST
cat("R-squared value:", R)
```

```
## R-squared value: 0.04992247
```

```r
summary(rand.fit)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.6335 -12.0422   0.6063  11.0961  24.0695
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.1976     4.1222   7.568 9.95e-10 ***
## X            -0.2234     0.1407  -1.588    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.36 on 48 degrees of freedom
## Multiple R-squared:  0.04992,    Adjusted R-squared:  0.03013
## F-statistic: 2.522 on 1 and 48 DF,  p-value: 0.1188
```

U primjeru su generirana 2 nasumična uzorka i izračunata je R-kvadrat vrijednost. Plava linija predstavlja srednju vrijednost uzorka Y i vidimo da se regresijska linija ne razlikuje puno od nje. Jasno se vidi da je R-kvadrat blizu nula i može se zaključiti da varijable nisu međusobno ovisne.

Izračun R^2 vrijednosti za dani skup podataka

```r
SSE = sum(fit$residuals**2)
SSR = sum((mean(y) - fit$fitted.values)**2)
SST = SSE + SSR
R = 1 - SSE/SST
cat("R-squared value:", R)
```

```
## R-squared value: 0.8549867
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.692 -3.273  1.083  1.733  8.233
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.4667     7.7031   9.278 3.15e-06 ***
## x            -0.1168     0.0152  -7.678 1.68e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.301 on 10 degrees of freedom
## Multiple R-squared:  0.855,  Adjusted R-squared:  0.8405
## F-statistic: 58.96 on 1 and 10 DF,  p-value: 1.683e-05
```

Naši podaci daju veliku R-kvadrat vrijednost stoga vidimo da je prilagodba regresijskog pravca podacima vrlo dobra.

Kvaliteta prilagodbe - ANOVA(Analysis of Variance)

Pomoću ANOVA analize možemo saznati kvalitetu prilagodbe našeg linearnog modela. ANOVA pristup temelji se na rastavljanju totalne varijance u podacima na različite značajne komponente koje se tada promatraju. .U linearnom modelu $y = \beta_0 + \beta_1 x$ koeficijent β_1 pokazuje kolika je povezanost između reakcije i regresora. Ako je $\beta_1 = 0$ tada na temelju danih podataka ne možemo zaključiti o postojanju veze između X i Y. Postavimo dvije hipoteze:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

. Varijancu zavisne varijable Y, σ_y^2 možemo rastavisiti na dvije komponente: $\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2$, odnosno nakon množenja stupnjevima slobode na $SST = SSR + SSE$. Drugačije kazano, totalnu varijancu podataka SST možemo razdvojiti na varijancu SSR koju je naš linearni model opisao i varijancu koja nije objašnjena našim modelom SSE. Testna statistika u ANOVA testu jest:

$$f = \frac{SSR}{s^2}$$

koja prati F distribuciju $f(1, n-2)$.

Test linearnosti s ponovljenim observacijama - Lack-of-Fit Test

U našem skupu podataka imamo 3 različite vrijednosti slučajne varijable X i u svake od te 3 grupe imamo 4 mjerenja za varijablu Y. Pretpostavimo da imamo k različitih grupa i u svakoj grupi po n mjerenja. Suma kvadrata SSE raspada se u dvije komponente:

- Suma odstupanja od srednje vrijednosti grupe:

  $$SSE_{pure} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2$$

  . Ova komponenta mjeri varijancu oko srednje vrijednosti grupe i predstavlja čistu eksperimentalnu grešku.

- Varijacija SSE' koja je nastala zbog šuma u podacima koje donose nelinearni članovi i koji utječu na Y, a nisu dio modela.

Statistika pomoću koje provodimo test je:

$$f = \frac{SSE - SSE_{pure}}{s^2(k-2)}$$

, koja prati f distribuciju: $f(k-2, n-k)$

```r
#Lack of fit test
library(alr3)
```

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

```
## Loading required package: carData
```

```r
pureErrorAnova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## x             1 1090.44 1090.44 74.2683 1.216e-05 ***
## Residuals    10  184.95   18.49
##  Lack of fit  1   52.81   52.81  3.5966    0.0904 .
##  Pure Error   9  132.14   14.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("P value: 0.0904")
```

```
## P value: 0.0904
```

Ispis ANOVA analize pokazuje nam da je P vrijednost za Lack of fit test 9.4%. Uz nivo signifikantnosti od 5 % ne možemo odbaciti hipotezu H0, koja govori da ne postoji linearna veza između varijabla Y i X. Ali ova P vrijednost je jako blizu kritične vrijednosti pa možemo sumnjati u linearnu povezanost između y i x.

(b) Transformacije podataka

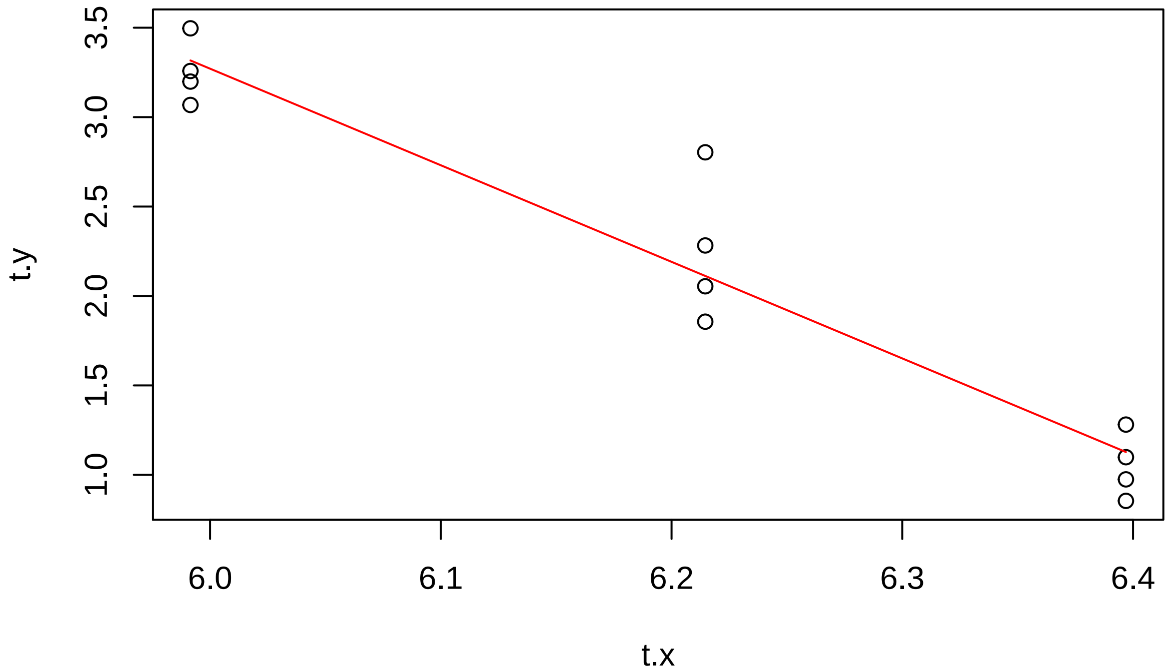
Često je praktično raditi s modelima u kojima X i Y ovise nelinearno. Primjena različitih transformacija može pobošljati prilagodbu regresijske linije podatcima i dati bolje predikcije. Stvarna ovisnost između y i x je: $y = \alpha x^{\beta}$. Uvođenjem zamijene:

$$y' = ln(y)$$

$$x' = ln(x)$$

. Dobivamo linearni model $\ln(y) = \beta \ln(x) + \ln(\alpha)$ i radimo regresiju za varijable x' i y'.

```r
#Lack of fit test
t.x = log(x)
t.y = log(y)
plot(t.x,t.y)
t.fit = lm(t.y~t.x)
lines(t.x,t.fit$fitted.values,col='red')
```




```r
sse = sum(t.fit$residuals**2)
ssr = sum((mean(t.y) - t.fit$fitted.values)**2)
sst = sse + ssr
R.t = 1 - sse/sst
cat("R-squared value:", R.t)
```

```
## R-squared value: 0.9223598
```

```r
print(" ")
```

```
## [1] " "
```

```r
pureErrorAnova(t.fit)
```

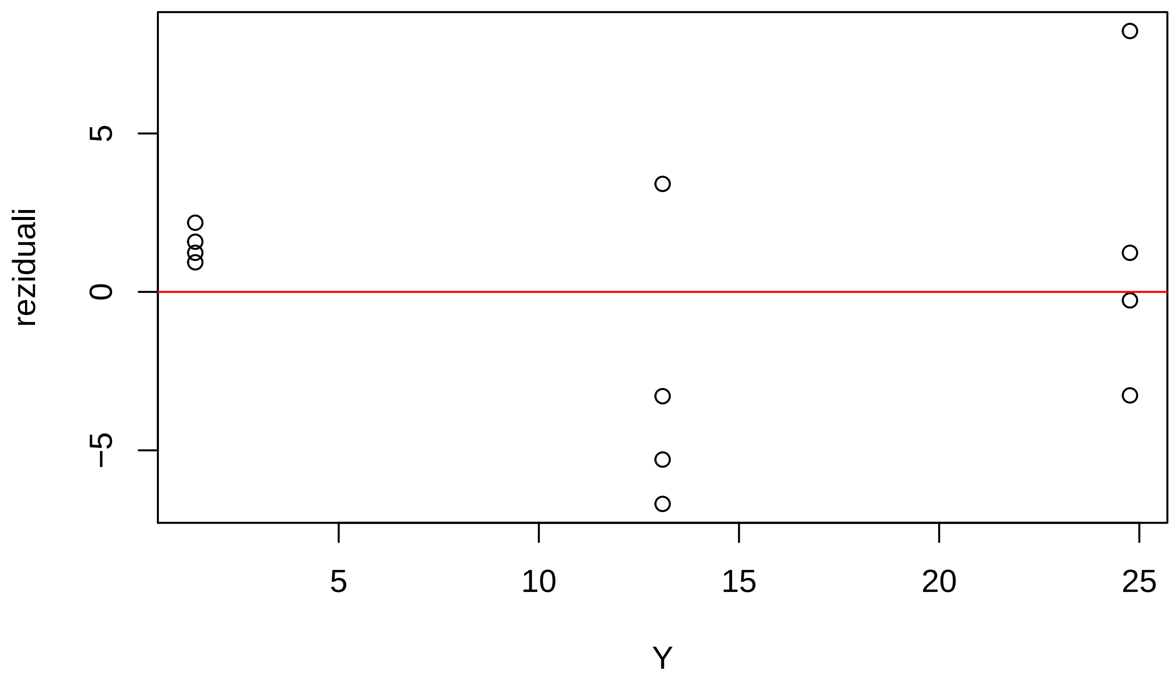
```
## Analysis of Variance Table
##
## Response: t.y
##              Df Sum Sq Mean Sq F value    Pr(>F)
## t.x           1 9.6194  9.6194  124.26 1.439e-06 ***
## Residuals    10 0.8097  0.0810
##  Lack of fit  1 0.1130  0.1130    1.46    0.2577
##  Pure Error   9 0.6967  0.0774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) Analiza reziduala

Pretpostavljamo da šum slijedi $N(0, \sigma)$ distibuciju i da su reziduali nezavisni.

Residual vs. Fit plot

```
plot(fit$fitted.values,xlab="Y",fit$residuals,ylab = "reziduali")
abline(h=0,col='red')
```



Standardizirani reziduali

Zbog definicije reziduala, suma reziduala unutar uzorka biti će jednaka 0, što implicira međusobnu zavisnost reziduala. Reziduali, za razliku od šuma, nemaju jednake varijance. Zbog toga je potrebno standardizirati reziduale da bi ih sveli na iste varijance. Standardizirani rezidual je:

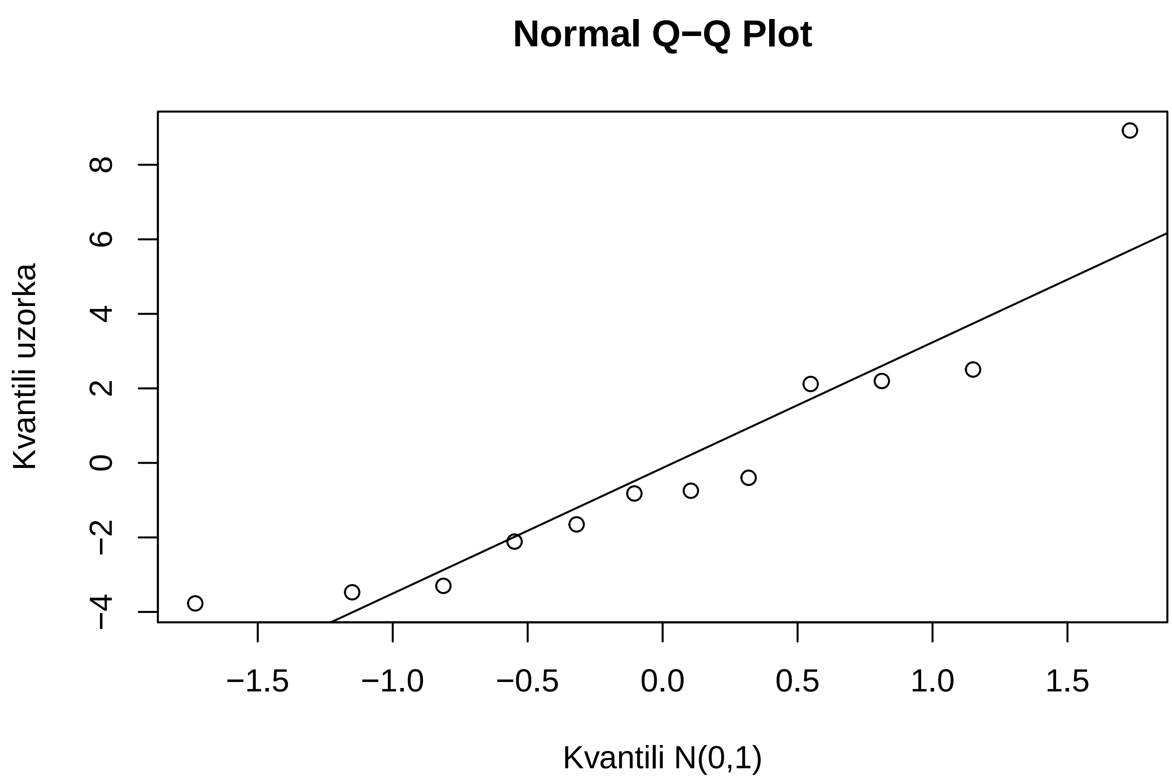
$$t_i = \frac{e_i}{s\sqrt{1 - hii}}$$

, gdje je $h_{ii} = \frac{1}{n} + \frac{(x_i - \hat{x})^2}{\sum_{j=1}^{n}(x_j - \hat{x})^2}$ Provjera dali standardizirani reziduali dolaze iz $N(0, \sigma^2)$ može se izvršiti pomoću 2 kriterija: grafa normalnih vjerovatnosti(Q-Q plot) i Kolmogorov-Smirnovljev testa

Normal Q-Q Plot

Provjera dali standardizirani podaci dolaze iz jedinične normalne distibucije N(0,1) putem QQ plota.

```
n = length(t.x)
hii = 1/n + (t.x-mean(t.x))**2/(sum((t.x-mean(t.x))**2))
s = 1/(n-2) * sum((t.y-t.fit$fitted.values)**2)
ti = t.fit$residuals/(s*sqrt(1-hii))

qqnorm(ti,xlab = "Kvantili N(0,1)",ylab = "Kvantili uzorka")
qqline(ti)
```



Kolmogorov Smirnov Test - KS test

KS test uspoređuje testnu i referentnu distribuciju i provjerva dali su jednake. H0: uzorak dolazi iz referentne distribucije. U našem slučaju referentna distribucija je normalana razdioba N(0,1)

```
ks.test(ti,"pnorm",mean=0,sd=1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ti
## D = 0.3672, p-value = 0.05866
## alternative hypothesis: two-sided
```

(d) Procijena intervala povjerenja parametra linearne regresije transformiranih podataka

Interval povjerenja reda p za parametar θ jest interval $[\theta_L, \theta_U]$ za koji vrijedi: $P(\theta_L < \theta < \theta_U) = p$. Vjerovatnost da se parametar θ nalazi u tom intervalu jest p.

```r
confint(t.fit)
```

```
##                2.5 %    97.5 %
## (Intercept) 28.82192 42.516712
## t.x         -6.50359 -4.295897
```

(e) Model za originalne podatke

Teorijska povezaost između x i y glasi: $y = \alpha x^{\beta}$

```r
# transformirani podaci : tx,ty
# originalni podaci
summary(t.fit)
```

```
##
## Call:
## lm(formula = t.y ~ t.x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27312 -0.17695 -0.05837  0.15764  0.69134
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.6693     3.0731   11.61 3.99e-07 ***
## t.x          -5.3997     0.4954  -10.90 7.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2846 on 10 degrees of freedom
## Multiple R-squared:  0.9224, Adjusted R-squared:  0.9146
## F-statistic: 118.8 on 1 and 10 DF,  p-value: 7.179e-07
```

Linearni model za transformirane podatke iznosi: $\hat{y}' = 35.6693 - 5.3997x'$. Potrebno je napraviti invrznu transformaciju i izračunati parametre α i β. Originalna transformacija glasi ovako:

$$y' = ln(y)$$

$$x' = ln(x)$$

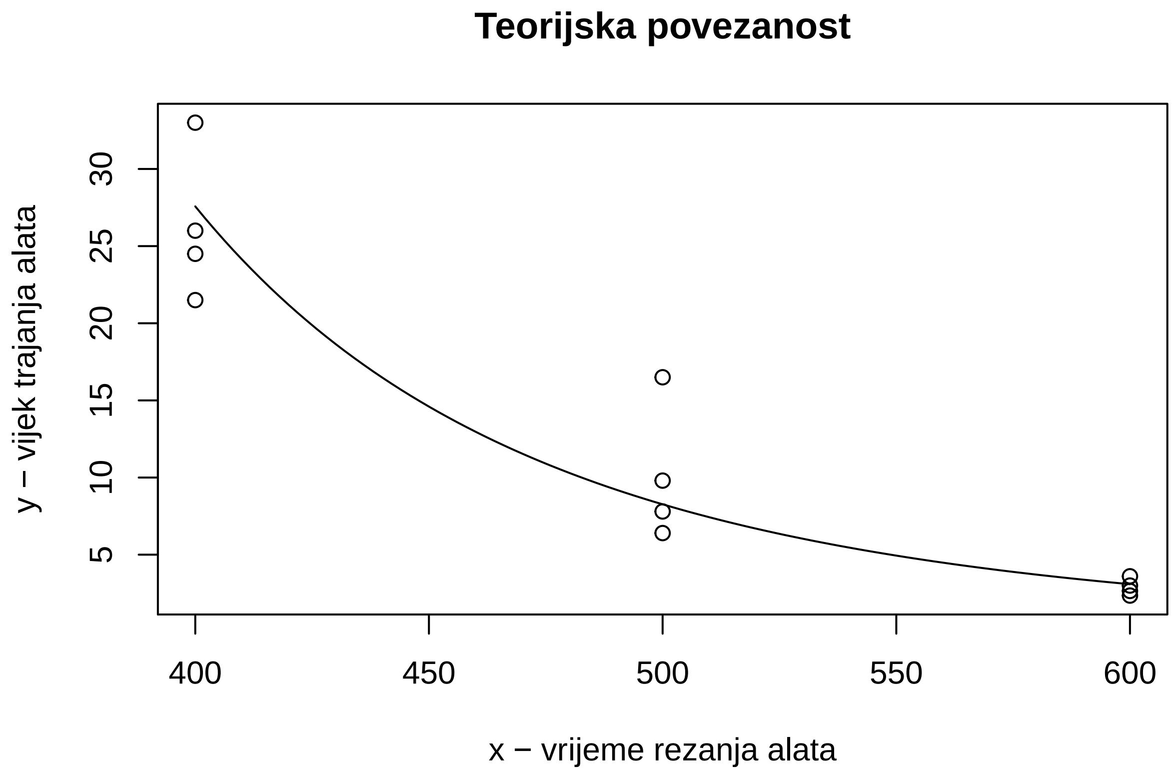
. Transformirana teorijska povezanost glasi: $ln(y) = \beta ln(x) + ln(\alpha)$, iz ove jednadžbe slijedi vrijednosti parametara:

$$\beta = -5.3997435$$
$$\alpha = e^{35.6693149} = 3.0973238 \times 10^{15}$$

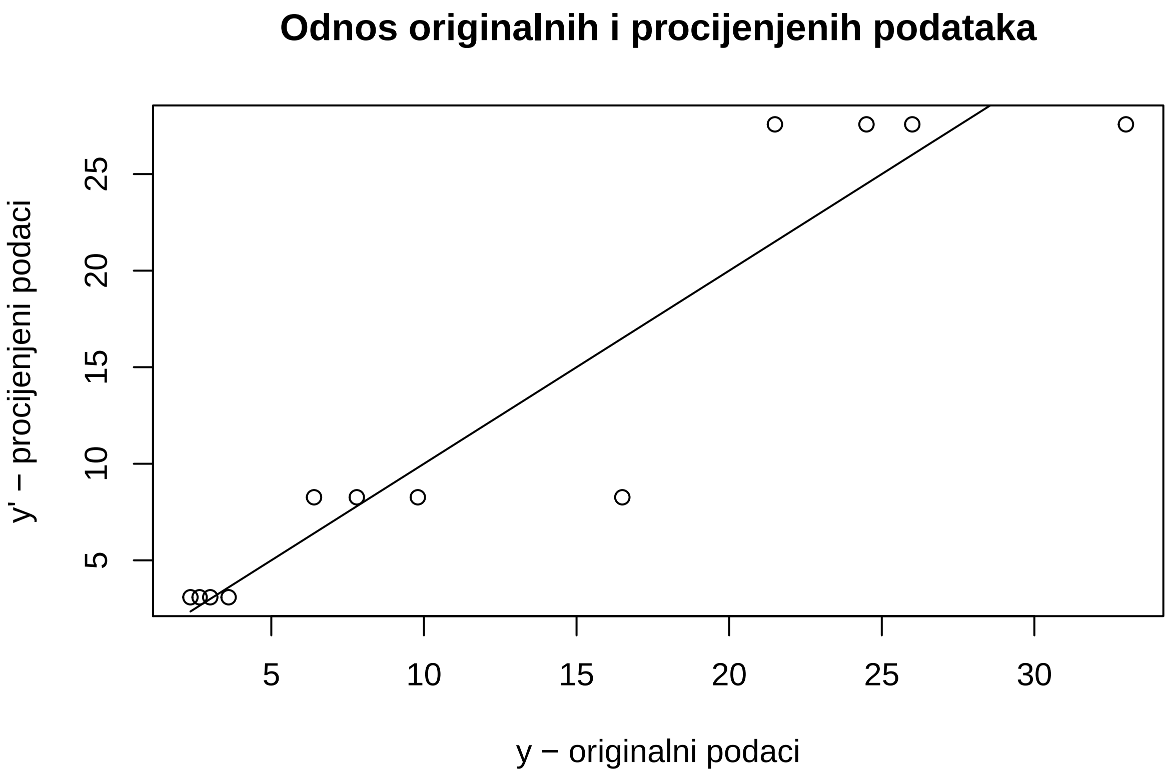
. Konačno teorijska povezanost je:

$$y = 3.0973238 \times 10^{15} x^{-5.3997435}.$$

```
alfa = exp(t.fit$coefficients[1])
beta = t.fit$coefficients[2]
fun = function(x){alfa*x**(beta)}
plot(x, y,xlab="x - vrijeme rezanja alata",ylab = "y - vijek trajanja alata",main = "Teorijska povezanos
curve(fun,add=TRUE)
```




```
plot(y, exp(t.fit$fitted.values), xlab = "y - originalni podaci", ylab = "y' - procijenjeni podaci", ma
curve(identity, add=TRUE)
```



Graf modela zadovoljavajuće prolazi kroz podatke. Doduše, samo iz grafa procijenjene vrijednosti ne vidi se varijabilnost zavisne varijable, što je također koristan podatak.

(f) Grafički prikaz intervala pouzdanosti

Interval povjerenja za očekivanu vrijednost od Y

Formula $\hat{y} = b_0 + b_1 x$ može se koristiti za predviđanje srednje vrijednosti $\mu_{Y|x0}$ ili za predviđanje ($Y_0 = y_0 | x = x_0$). Koristimo procjenitelj $\hat{Y_0} = B_0 + B_1 x_0$ za procijenu $\mu_{Y|x_0} = \beta_0 + \beta_1 x$. Distribucija uzorkovanja procjenitelja $\hat{Y_0}$ je normalna gdje je:

- $\mu_{Y|x_0} = E(\hat{Y}) = \beta_0 + \beta_1 x_0$
- $\sigma^2_{\hat{Y}} = \sigma^2(\frac{1}{n} + \frac{(x_0 - \overline{x})}{S_{xx}})$

Statistika koju koristimo za određivanje intervala povjerenja od $\mu_{Y|x_0}$ je:

$$T = \frac{\hat{Y} - \mu_{Y|x_0}}{S\sqrt{1/n + (x_0 - \overline{x})^2 / S_{xx}}}$$

, te imamo interval pouzdanosti:

$$\hat{y_0} - t_{\alpha/2} SE(\mu_{Y|x_0}) < \mu_{Y|x_0} < \hat{y_0} + t_{\alpha/2} SE(\mu_{Y|x_0})$$

Interval povjerenja za buduću vrijednosti Y

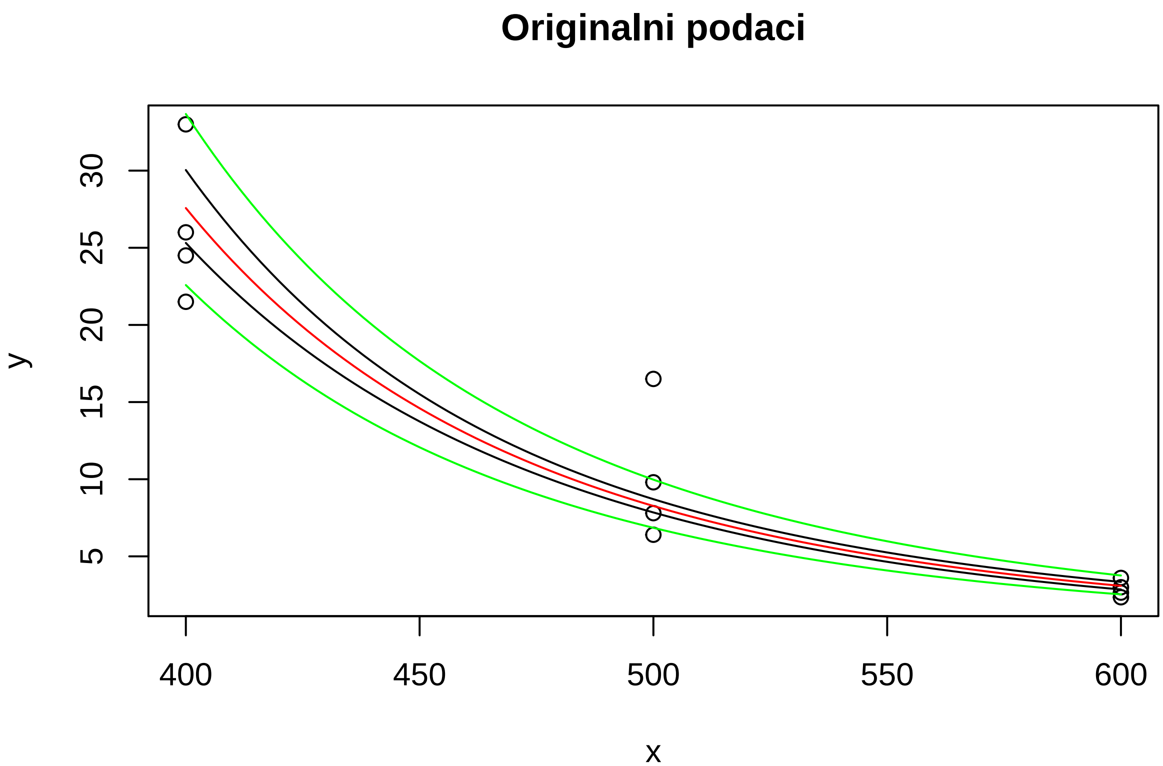
Distribucija uzorkovanja procjenitelja $\hat{Y_0} - Y_0$ je normalna gdje je:

- $\mu \hat{Y}_0 - Y_0 = E(\hat{Y}_0 - Y_0) = 0$
- $\sigma^2_{\hat{Y}_0 - Y_0} = \sigma^2(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}})$

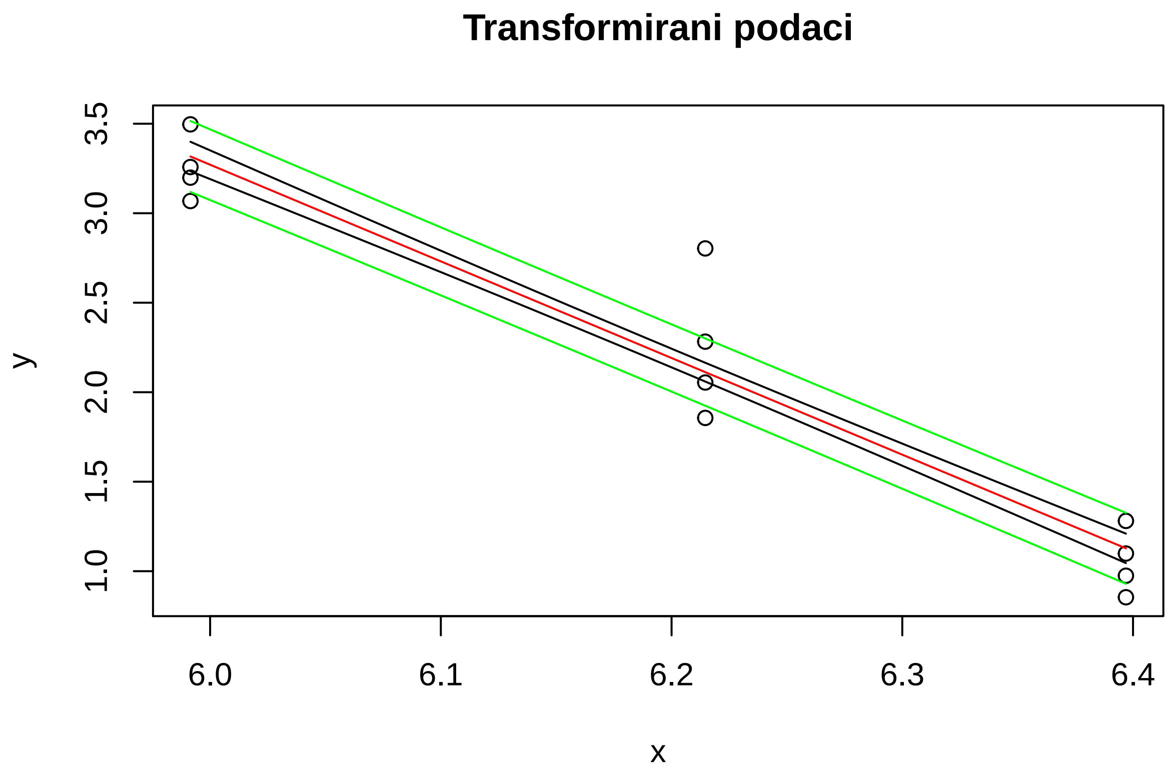
Interval pouzdanosti za $\hat{Y}_0 - Y_0$ je:

$$\hat{y_0} - t_{\alpha/2} SE(\hat{Y}_0 - Y_0) < \hat{y_0} < \hat{y_0} + t_{\alpha/2} SE(\hat{Y}_0 - Y_0)$$

```
conf.plot = function (x,y,fit,alpha = 0.05,title,transform_x=NULL,transform_y_back=NULL){
  n = length(x)
  s = sum(fit$residuals**2)/(n-2)
  t = qt(1-alpha/2,df = n-2)
  tx = if (is.null(transform_x)) {x} else {transform_x(x)}
  sxx = sum((tx-mean(tx))**2)
  b0 = fit$coefficients[1]
  b1 = fit$coefficients[2]

  y.hat = function(x) {b0 + b1*x}

  is.mean = TRUE
  se = function(x)  {t*s*sqrt((!is.mean)+1/n+(x-mean(x))**2/sxx)}

  bound = function(x, factor) {
    if (!is.null(transform_x)) {
      # x -> x'
      x = transform_x(x)
    }
    y_bound = y.hat(x) + se(x) * factor
    if (!is.null(transform_y_back)) {
      # y' -> y
      y_bound = transform_y_back(y_bound)
    }
    return(y_bound)
  }

  plot(x,y,main = title)
  curve(bound(x, +1), add=TRUE)
  curve(bound(x, 0), col = "red", add=TRUE)
  curve(bound(x, -1), add=TRUE)

  is.mean = FALSE
  curve(bound(x, +1),col = "green",add=TRUE)
  curve(bound(x, -1),col = "green",add=TRUE)
}


conf.plot(x,y,t.fit,title = "Originalni podaci", transform_x = log, transform_y_back = exp)
```




```r
conf.plot(t.x,t.y,t.fit,title = "Transformirani podaci")
```



Crvena krivulja - Regresijski pravac $\hat{y} = b_0 + b_1 x$
Crna krivulja - Interval povjerenja za $\mu_{y_i|x_i}$
Zelena krivulja - Interval povjerenja za $y_i|x_i$