Linearna regresija

Luka Radivoj, Kristijan Rupić, Dominik Stipić

12.06.2019.

Zadatak A

```
X = c(41, 39, 53, 67, 61, 67, 46, 50, 55, 72, 63, 59,
      53, 62, 65, 48, 32, 64, 59, 54, 52, 64, 51, 62,
      56, 38, 52, 40, 65, 61, 64, 64, 53, 51, 58, 65)
Y = c(29, 19, 30, 27, 28, 27, 22, 29, 24, 33, 25, 20,
      28, 22, 27, 22, 27, 28, 30, 29, 21, 36, 20, 29,
      34, 21, 25, 24, 32, 29, 27, 26, 24, 25, 34, 28)

summary(X)
```

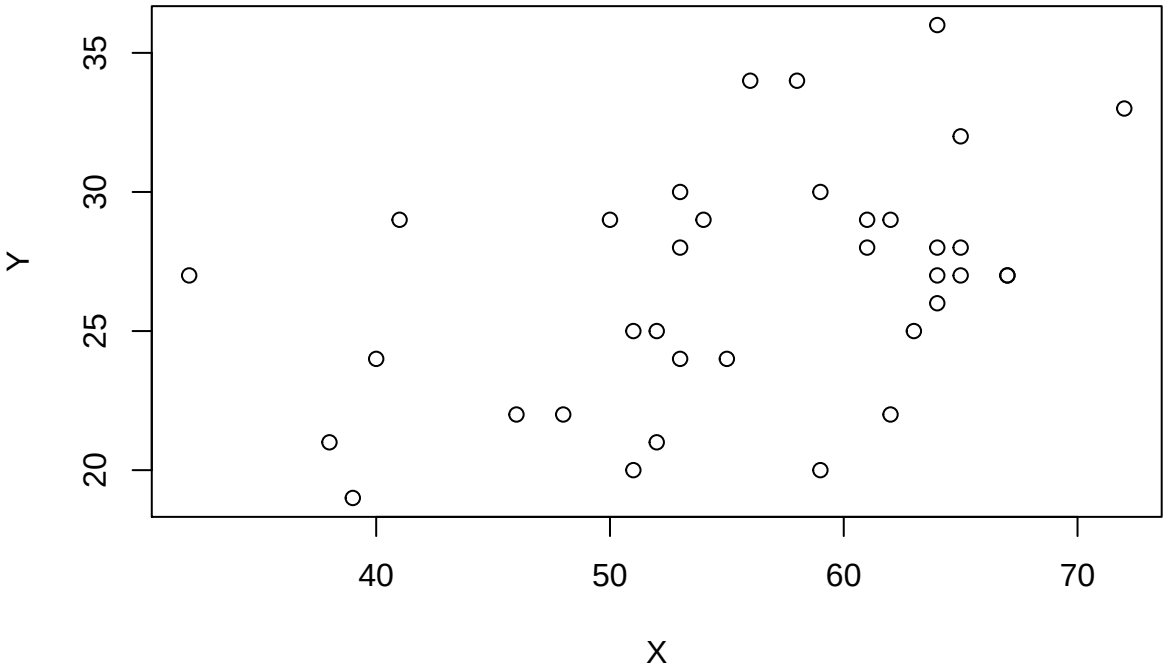
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32.00   51.00   57.00   55.72   64.00   72.00
```

```
summary(Y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   24.00   27.00   26.69   29.00   36.00
```

a)

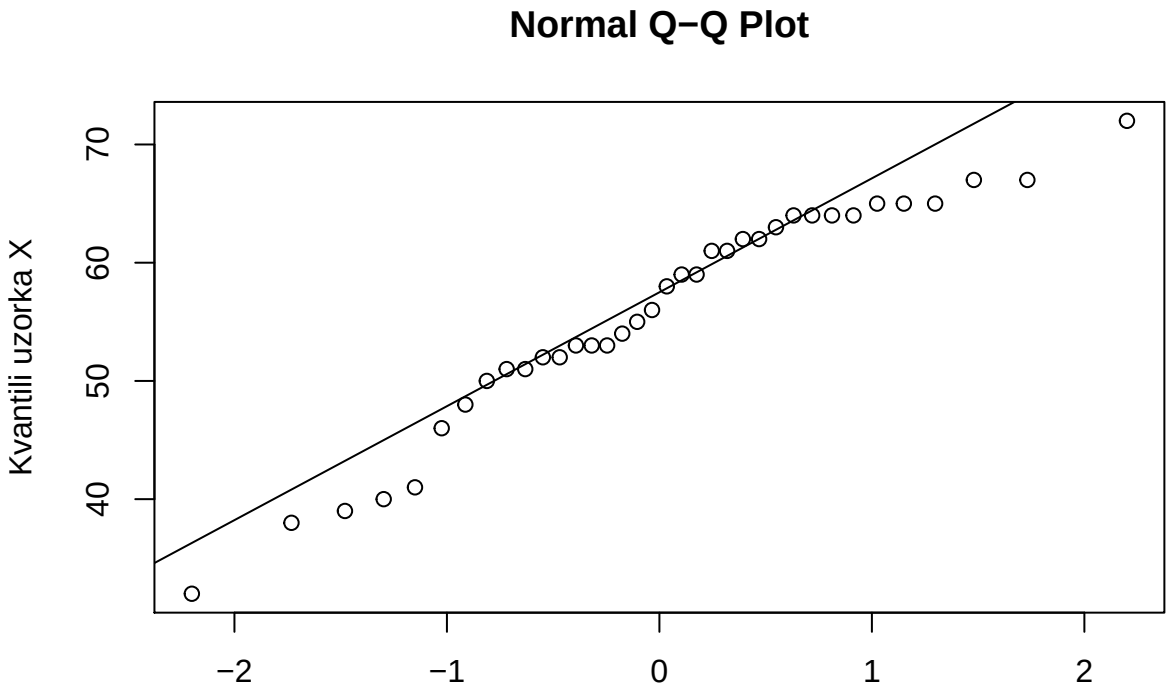
Prikaz podataka u Kartezijevom koordinatnom sustavu (scatter plot):

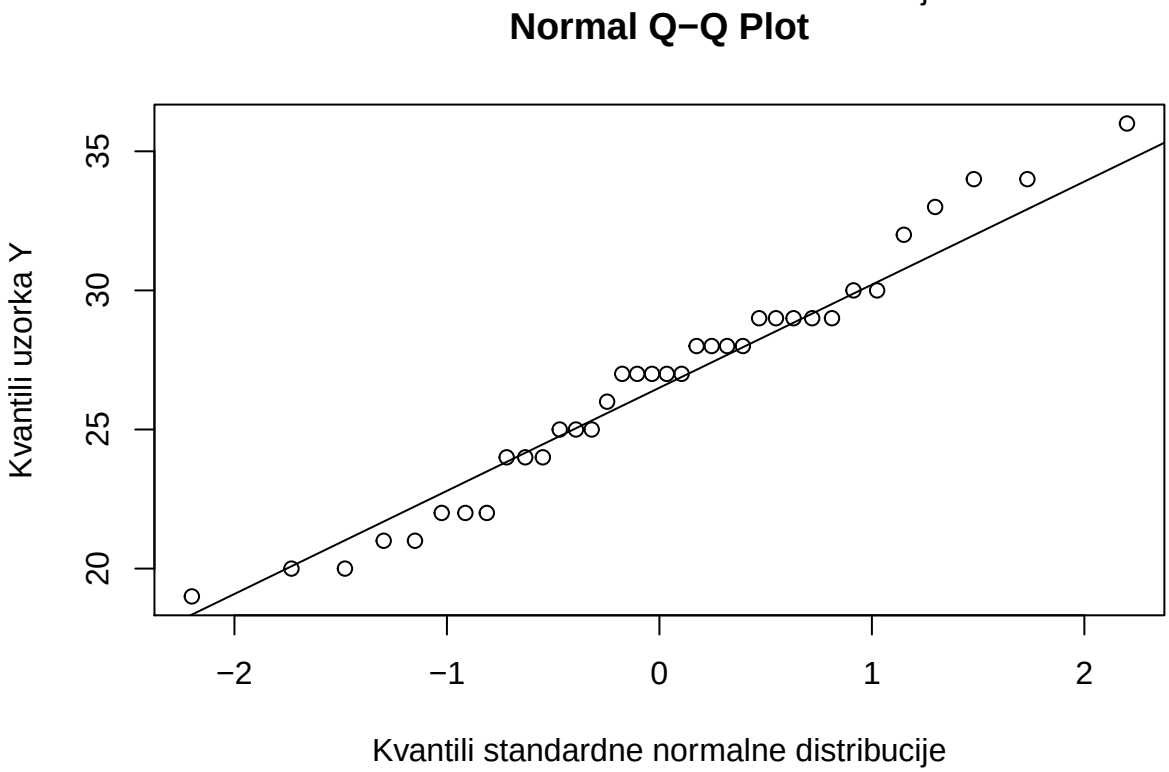


b)

Za obje varijable provjeravamo jesu li normalno distribuirane na 3 načina.

Q-Q Plot





Vidimo da se obje varijable dobro poklapaju s normalnom distribucijom u okolini medijana. Kvantili varijable X padaju ispod linije i razrjeđuju se u vanjskim kvartilima, što znači da ima tanje repove od normalne distribucije. Varijabla Y se puno bolje poklapa s normalnom distribucijom, no pokazuje blagu zakrivljenost

prema većim vrijednostima.

Lillieforsova inačica Kolmogorov-Smirnovljevog testa

Kolmogorov-Smirnovljev je neparametarski test koji služi za kvantificiranje razlike između empirijske funkcije distribucije iz uzorka i referentne funkcije distribucije. Nulta hipoteza tog testa jest da uzorak dolazi iz referentne distribucije iz čega slijedi da očekujemo poklapanje empirijske funkcije distribucije s referentnom. Prevelika razlika među njima dovodi u sumnju nultu hipotezu i navodi nas da ju odbacimo. Postoji i inačica testa koja testira jednakost distribucija dvaju uzoraka.

Empirijska funkcija distribucije iz uzorka n i.i.d. varijabli X_i jest:

$$F_n(x) = \sum_{i=1}^{n} I_{\langle -\infty, x]}(X_i)$$

a testna statistika, tj. maksimalno odstupanje od teoretske distribucije $F(x)$ je:

$$D_n = \sup_x |F_n(x) - F(x)|$$

Pod H_0 vrijedi:

$$\sqrt{n} D_n \xrightarrow{d} \sup_t |B(F(t))|$$

gdje je $B(t)$ jednični Brownov most, tj. Brownovo gibanje na intervalu $[0, 1]$ s početkom i završetkom u vrijednosti 0. Desna strana ima Kolmogorovljevu distrubuciju pa smatramo da i ima i lijeva za veće vrijednosti n; ovo je dakle aproksimativan test. H_0 se odbacuje na razini značajnosti α kada $\sqrt{n} D_n > K_\alpha$.

Problem s ovako formuliranim testom jest taj što zahtijeva točno specificiranu funkciju distribucije u nultoj hipotezi, tj. testira poklapanje s distribucijom s točno određenim parametrima. Kada ne bismo poznavali distribuciju uzorka ili njene parametre (a najčešće ne znamo), morali bismo ih procijeniti iz uzorka. Pokazuje se da KS test ne radi dobro s paramterima distribucije procijenjenim iz istog uzorka i da ga se ne smije koristiti na takav način. Ovo odgovara našoj situaciji jer i pod pretpostavkom normalnosti (s kojom su Q-Q plotovi uglavnom u skladu) i dalje ne znamo parametre naših distribucija.

Kao jedno rješenje nudi se Lillieforsov test, koji je inačica KS testa upravo za slučaj testiranja normalnosti ali bez poznatih parametara teoretske distribucije. Postupak je kao i u naivnoj primjeni KS testa - procjene se μ i σ iz uzorka te se računa maksimalno odstupanje od normalne distribucije $\mathcal{N}(\hat{\mu}, \hat{\sigma^2})$. Razlika od KS testa je ta da se ne radi usporedba s kritičnim vrijednostima Kolmogorovljeve distribucije jer se maksimalno odstupanje smanjilo izborom normalne distribucije s procijenjenim parametrima za teoretsku distribuciju pod nultom hipotezom, pa testna statistika sada ima Lillieforsovu distribuciju koja se računa Monte Carlo metodama radi svoje složenosti.

$$H_0 : \exists\mu \exists\sigma^2, F(x) = \Phi(x, \mu, \sigma^2)$$
$$H_1 : \forall\mu \forall\sigma^2, F(x) \neq \Phi(x, \mu, \sigma^2)$$

Nemamo zadane razine značajnosti, stoga samo računamo p-vrijednosti.

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X
## D = 0.12704, p-value = 0.1495
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Y
## D = 0.11209, p-value = 0.3021
```

Vidimo da su p-vrijednosti vrlo velike i sigurno ne bismo odbacili nultu hipotezu za uobičajene vrijednosti α. Da je p-vrijednost za X manja od Y slaže se s procjenom iz Q-Q plota, prema kojoj je Y bliže normalnosti od X.

Pearson χ^2 test

Pearsonov χ^2 test za prilagodbu distribuciji se sastoji od sljedećeg:

Neka su podaci iz uzorka $X_i, i \in \{1, \ldots, N\}$ i.i.d. grupirani u konačan broj k kategorija K_i koje čine particiju svih mogućih vrijednosti i neka su $o_i = |K_i|, i \in \{1, \ldots, k\}$ opažene frekvencije, tj. broj podataka u svakoj od kategorija. Pod pretpostavkom H_0 o distribuciji podataka $p_i = P(X \in K_i)$ su teoretske vjerojatnosti upadanja vrijednosti X u svaku od kategorija. Tada su $E_i = Np_i$ očekivane frekvencije svake od kategorija za uzorak veličine N.

Ovom transformacijom podatka vektor opaženih frekvencija poprima pod H_0 multinomijalnu distribuciju s gustoćom vjerojatnosti

$$f(o_1, \ldots, o_k; N; p_1, \ldots, p_k) = \binom{N}{o_1 \ldots o_k} \prod_{i=1}^{k} p_i^{o_i}$$

Tada vrijednost test statistike:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

ima asimptotski $\chi^2(\nu)$ distribuciju s $\nu = k - 1 - r$ stupnjeva slobode, gdje je r broj parametara teoretske distribucije procijenjen iz uzorka.

Velika vrijednost statistike ukazuje na preveliko odstupanje opaženih frekvencija u odnosu na očekivane pod hipotezom H_0. Stoga H_0 odbacujemo na razini značajnosti α akko $\chi^2 \geq \chi^2_{1-\alpha}(\nu)$.

Bitna pretpostavka radi kvalitete testa jest dovoljno velik N i e_i. Obično se kao kriterij uzima $\forall i, e_i \geq 5$. Nedostatak testa je potreba za grupiranjem podataka koji dolaze iz kontinuiranih distribucija u kategorije, čime se u test unosi arbitrarnost koja može utjecati na zaključak.

U našem slučaju, testiramo X i Y na normalnost, pa za obje moramo procjeniti $r = 2$ parametra iz uzorka. Procjenu parametara za nas vrši R funkcija `pearson.test`:

```
##
##  Pearson chi-square normality test
##
## data:  X
## P = 12.5, p-value = 0.0517

## Classes:  9

## p-value:  0.1302504

##
##  Pearson chi-square normality test
##
## data:  Y
## P = 2.5, p-value = 0.8685

## Classes:  9
```

```
## p-value:  0.9617309
```

Treba dodati nekoliko napomena. Funkcija sama procjenjuje broj kategorija u koje treba podijeliti podatke, a može se i podesiti ručno. Parametar `adjust` funkcije kontrolira radi li se korekcija oduzimanjem $r = 2$ od broja stupnjeva slobode; vidimo da dobivamo manje p-vrijednosti nakon korekcije. No dokumentacija funkcija dodaje opasku i da procjena parametara μ, σ^2 na uobičajen način nije ispravna te referira čitatelja na literaturu [Moore1986]. Objašnjenje je da naivnom procjenom parametara testna statistika nije distribuirana točno kao $\chi^2(k-3)$, već je za procjenu parametara potrebno riješiti sustav parcijalnih diferencijalnih jednadžbi za procjenitelje koji u slučaju normalne distribucije nemaju zatvorenu formu već se računaju numerički. Stvarna p-vrijednost se nalazi negdje između one sa i bez korekcije stupnjeva slobode pa ovdje računamo oboje.

Vidimo da za $\alpha = 0.05$ ne bismo mogli odbaciti H_0 ni za X ni za Y, a za $\alpha = 0.1$ bismo možda mogli odbaciti za X, ovisno o točnom iznosu stvarne p-vrijednosti. Vidimo da su ovi zaključci u skladu s onima dobivenim iz Q-Q plotova, gdje se može posumnjati u normalnost X, ali teže za Y.

c)

Bivarijatna normalna razdioba je dvodimenzionalna generalizacija normalne razdiobe. Pošto se uvođenjem dodatnih komponenti mogu pojaviti zavisnosti među njima, nešto je složenijeg oblika od jednodimenzionalne. Gustoća joj je:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right\}$$

Primjećujemo pojavljivanje novog parametra ρ koji odgovara koeficijentu korelacije X, Y: $\rho = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$. Za nekorelirane X, Y je $\rho = 0$ te se kvadratna forma u eksponentu pretvara u "Pitagorin poučak" nad standardiziranim varijablama. Za savršeno korelirane X, Y je $\rho = \pm 1$ te se kvadratna forma pretvara u kvadrat binoma.

Računanjem marginalnih razdioba pokazuje se da su komponente normalno distribuirane s odgovarajućim parametrima: $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, neovisno o ρ. Iz toga slijedi da je nužan uvjet da bi združena distribucija X, Y bila bivarijatna normalna da su obje normalno distribuirane. Obrat ne vrijedi, tj. moguće je da su X, Y obje normalno distribuirane, ali da nisu združeno bivarijatno normalno distribuirane. Pošto nismo odbacili normalnost X i Y pomoću prethodnih testova, još uvijek je moguće da imaju bivarijatnu normalnu razdiobu. Da bismo procijenili njene parametre $\mu_{X,Y}, \sigma^2_{X,Y}$ koristimo činjenicu da marginalne razdiobe imaju iste parametre koji se pojavljuju u združenoj. Tako su:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\hat{\sigma^2} &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \\
\hat{\rho} &= \frac{1}{(n-1)\hat{\sigma}_X\hat{\sigma}_Y}\sum_{i=1}^{n} (x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)
\end{aligned}$$

Tražimo ove vrijednosti i njihove 95% pouzdane intervale. $\hat{\mu}$ imaju pod pretpostavkom normalnosti egzaktno normalnu distribuciju uzorkovanja sa standaradnim devijacijama $\sigma/\sqrt{n}$. $\hat{\sigma^2}$ imaju opet pod pretpostavkom normalnosti egzaktno $\chi^2(n-1)$ distribucije. $\hat{\rho}$ je problematičan; egzaktna distribucija uzorkovanja za X, Y iz bivarijatne normalne razdiobe je poznata, ali daleko nepratkična za korištenje. Pošto nas ovdje zanima interval povjerenja za ρ, koristimo Fisherovu z-transformaciju:

$$z = \operatorname{artanh} \hat{\rho} = \frac{1}{2} \ln \frac{1+\hat{\rho}}{1-\hat{\rho}}$$

Vrijedi približno $Z \sim \mathcal{N}(\operatorname{artanh} \rho, 1/\sqrt{N-3})$. Time se dobiva:

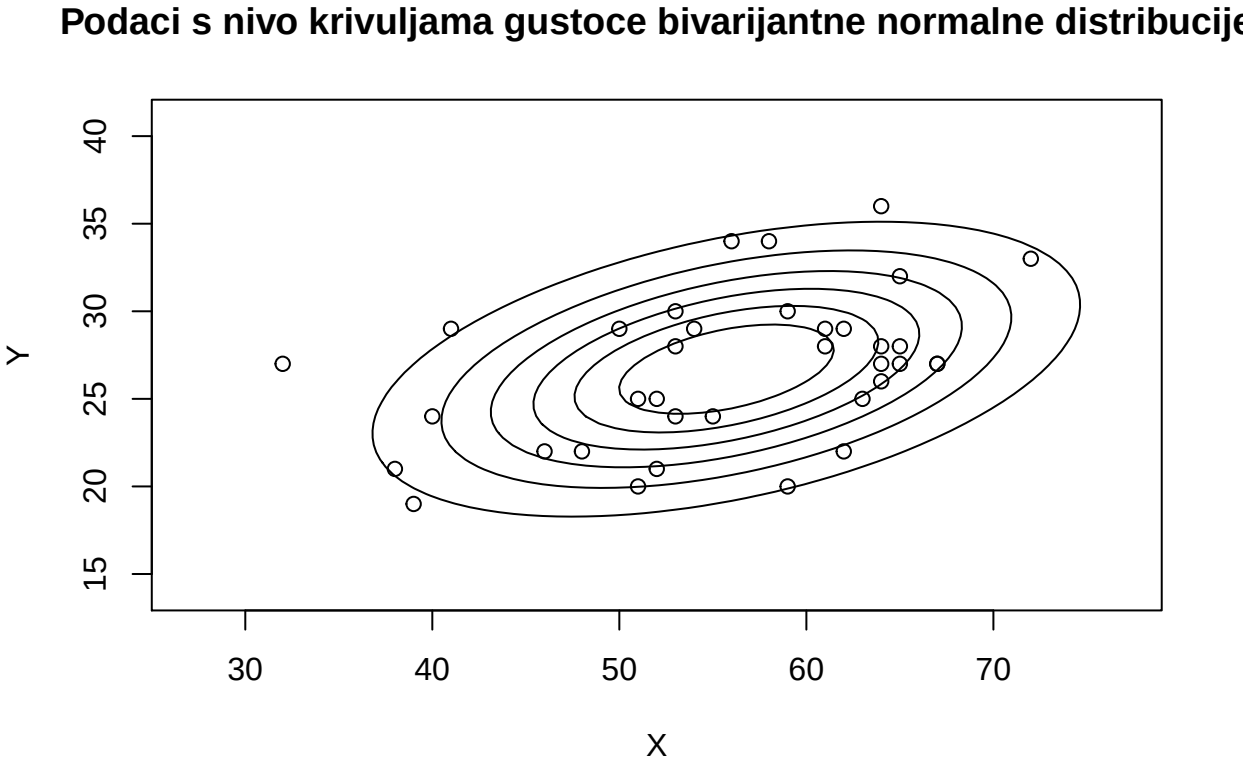
$$P\left(\rho \in \left[\tanh\left(\operatorname{artanh} \hat{\rho} - z_{\alpha/2}/\sqrt{N-3}\right), \tanh\left(\operatorname{artanh} \hat{\rho} + z_{\alpha/2}/\sqrt{N-3}\right)\right]\right) = 1-\alpha$$

za interval povjerenja za ρ.

Procijenjeni paramteri i intervali pouzadnosti dani su ispod.

$\hat{\mu}_X = 55.7222222,$	$[54.1357661, 57.3086783]$
$\hat{\mu}_Y = 26.6944444,$	$[25.988664, 27.4002249]$
$\hat{\sigma_X^2} = 90.6063492,$	$[59.6056885, 154.172014]$
$\hat{\sigma_Y^2} = 17.9325397,$	$[11.7969809, 30.5132674]$
$\hat{\rho} = 0.4372983,$	$[0.1270098, 0.6696296]$

d)



e)

$\hat{\rho} = 0.4372983.$

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0, \alpha = 0.05$$

```
##
##  Pearson's product-moment correlation
##
## data:  X and Y
## t = 2.8353, df = 34, p-value = 0.007654
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1270098 0.6696296
## sample estimates:
##       cor
## 0.4372983
```

Na razini značajnosti 0.95 odbacujemo nultu hipotezu o nekoreliranosti X, Y.

$$H_0 : \rho = 0.5$$
$$H_1 : \rho \neq 0.5, \alpha = 0.05$$

```
## rho:  0.4372983
```

```
## conf. interval:  0.2051665 0.7116367
```

```
## p-value:  0.6440943
```

Interval pouzdanosti za ρ jest [0.2051665, 0.7116367].

Pošto izračunata vrijenost $\hat{\rho}$ upada u interval povjerenja reda 0.95 pod H_0, ne odbacujemo nultu hipotezu na razini značajnosti 0.95. p-vrijednost je također vrlo velika.

f)

Provodimo χ^2 test za prilagodbu distribuciji koristeći kao razrede područja između izohipsi. Očekivane frekvencije znamo jer smo pomoću njih definirali izohipse. Sve očekivane frekvencije su ≥ 5, pa ne moramo združivati razrede. Nemamo zadanu razinu značajnosti pa samo računamo i komentiramo p-vrijednost.

$$H_0 : (X, Y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$H_1 : (X, Y) \not\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

```
## Expected:  6 5 5 5 5 5 5
```

```
## Observed:  4 6 7 5 3 7 4
```

```
## p-value:  0.7483991
```

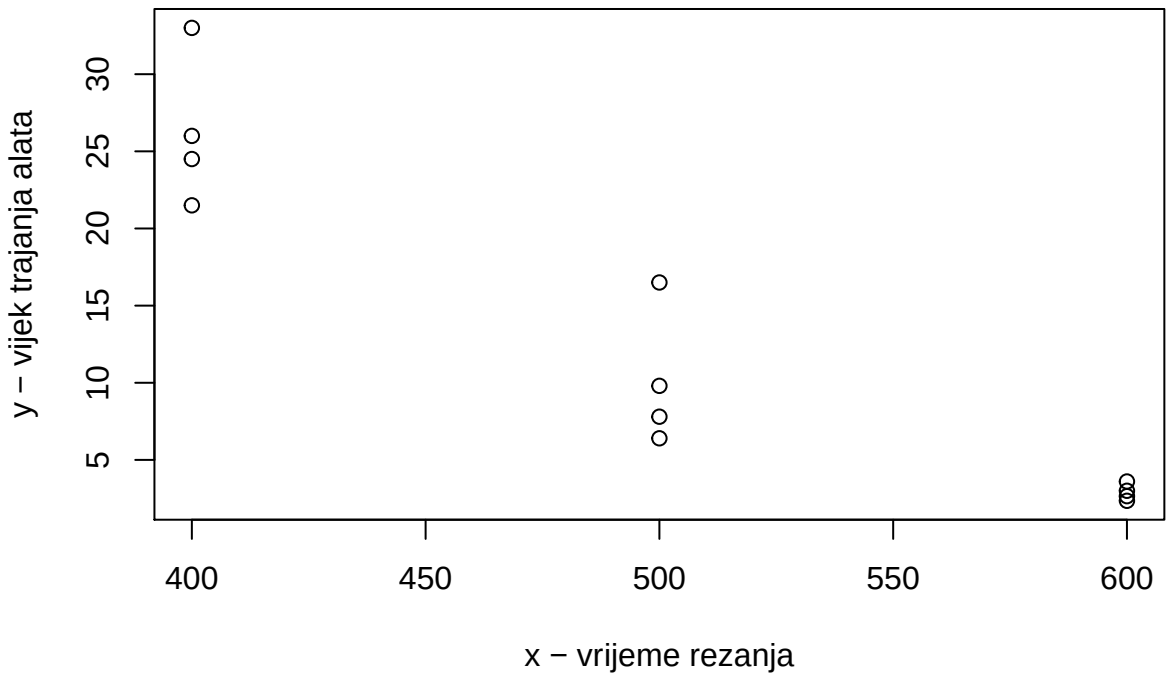
Dobivena p-vrijednost je vrlo velika i ne bismo odbacili H_0 na razumnim razinama značajnosti.

Zadatak B

```
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



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 procjena 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:

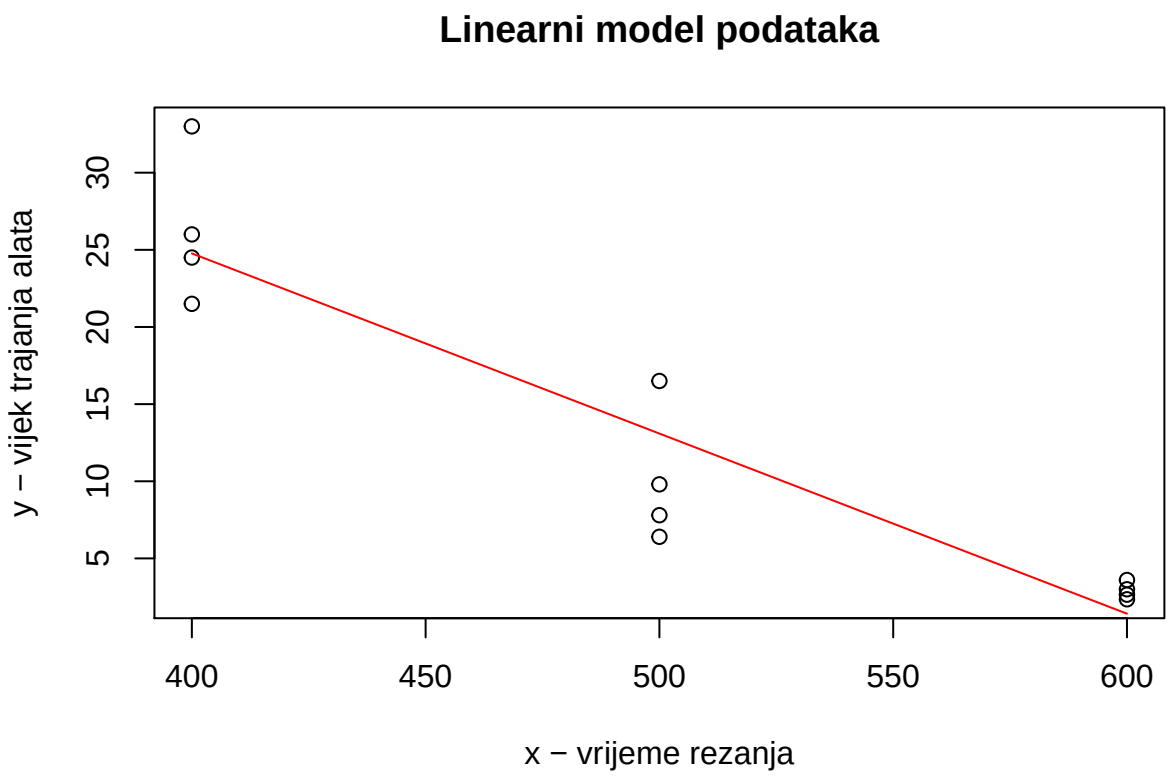
- očekivanje $E(Y_i)$ je linearna funkcija od x_i
- šumovi ϵ_i su nezavisni
- šumovi ϵ_i u točki x_i imaju $N(0, \sigma^2)$ distribuciju

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:

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



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 :

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

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

$$SST = SSE + SSR$$

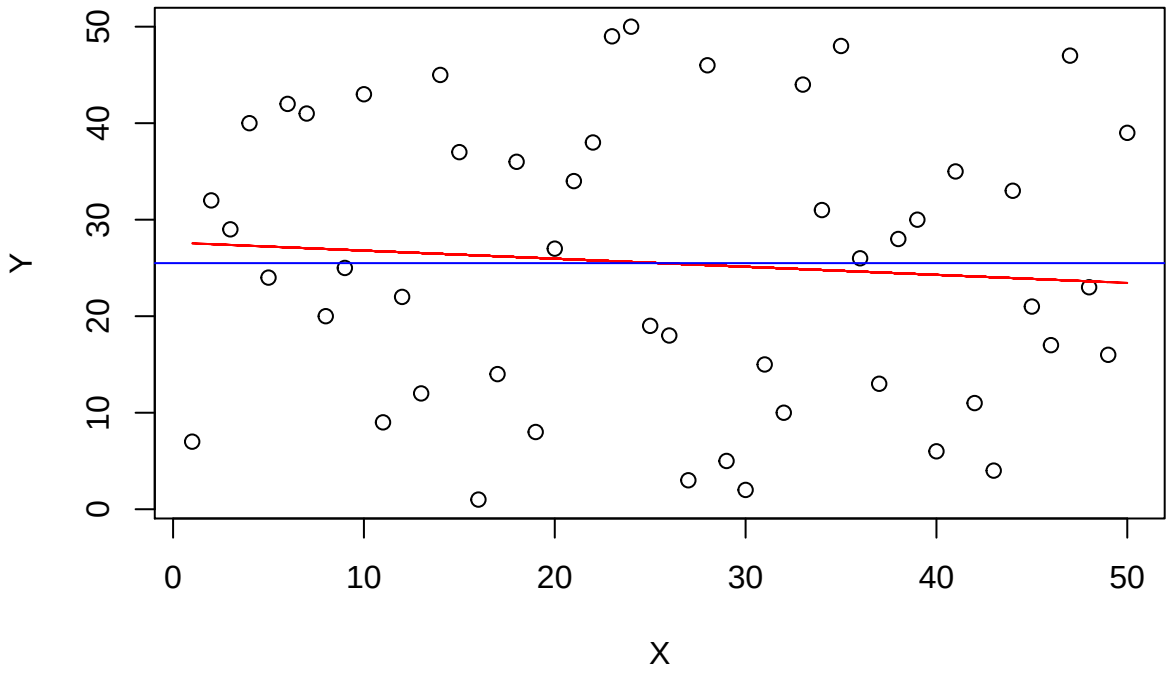
$$R^2 = 1 - \frac{SSE}{SST}$$

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 deterministčan model. Ako $R^2 \approx 0$ tada imamo potpunu nepovezanost između regresora i reakcija.

Primjer

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

```
## R-squared value: 0.007005268
```

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

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.2951 -12.0429   0.2114  11.8546  24.3745
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.6343     4.2143   6.557 3.51e-08 ***
## X            -0.0837     0.1438  -0.582    0.563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.68 on 48 degrees of freedom
## Multiple R-squared:  0.007005,   Adjusted R-squared:  -0.01368
## F-statistic: 0.3386 on 1 and 48 DF,  p-value: 0.5633
```

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

```
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
```

```
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 rastaviti 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:

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

  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: $f(k-2, n-k)$

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

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

```
## 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
```

```
## 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.

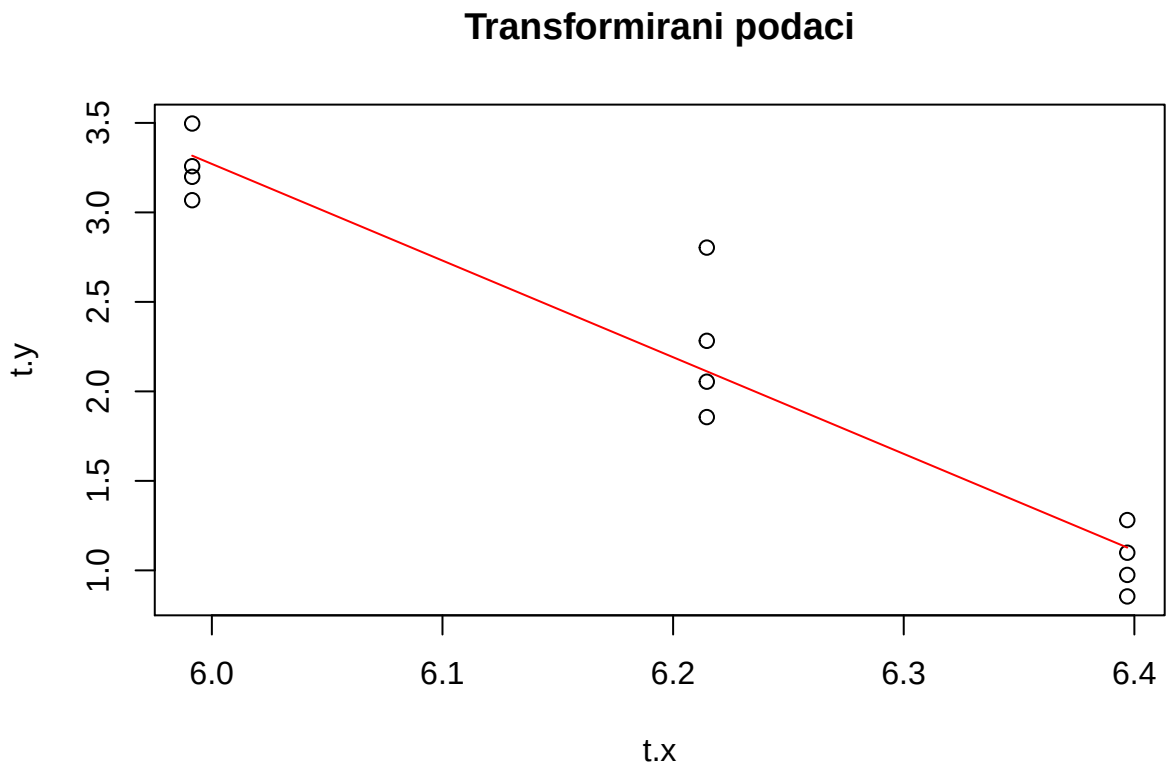
(b) Transformacije podataka

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 zamjene:

$$y' = \log(y)$$
$$x' = \log(x)$$

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,main="Transformirani podaci")
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, "\n")
```

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

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

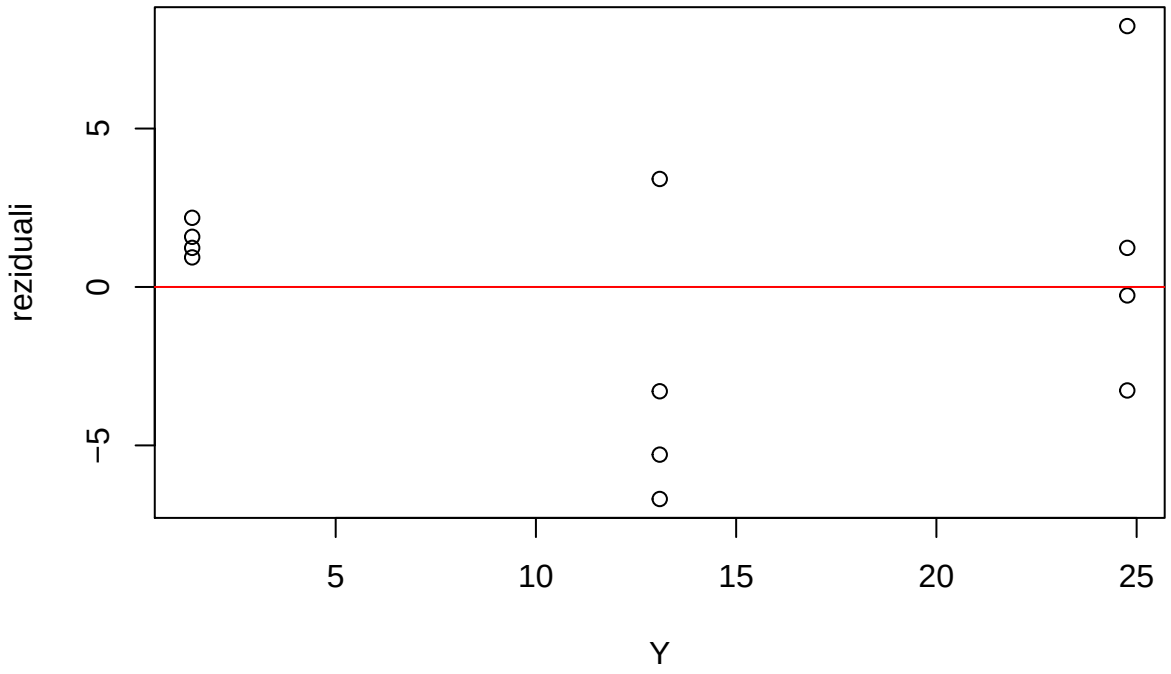
```
## Analysis of Variance Table
##
## Response: t.y
##              Df Sum Sq Mean Sq F value    Pr(>F)
## t.x           1 9.6194  9.6194  124.26 1.439e-06 ***
## Residuals    10 0.8097  0.0810
##  Lack of fit  1 0.1130  0.1130    1.46    0.2577
##  Pure Error   9 0.6967  0.0774
## ---
## Signif. codes:  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

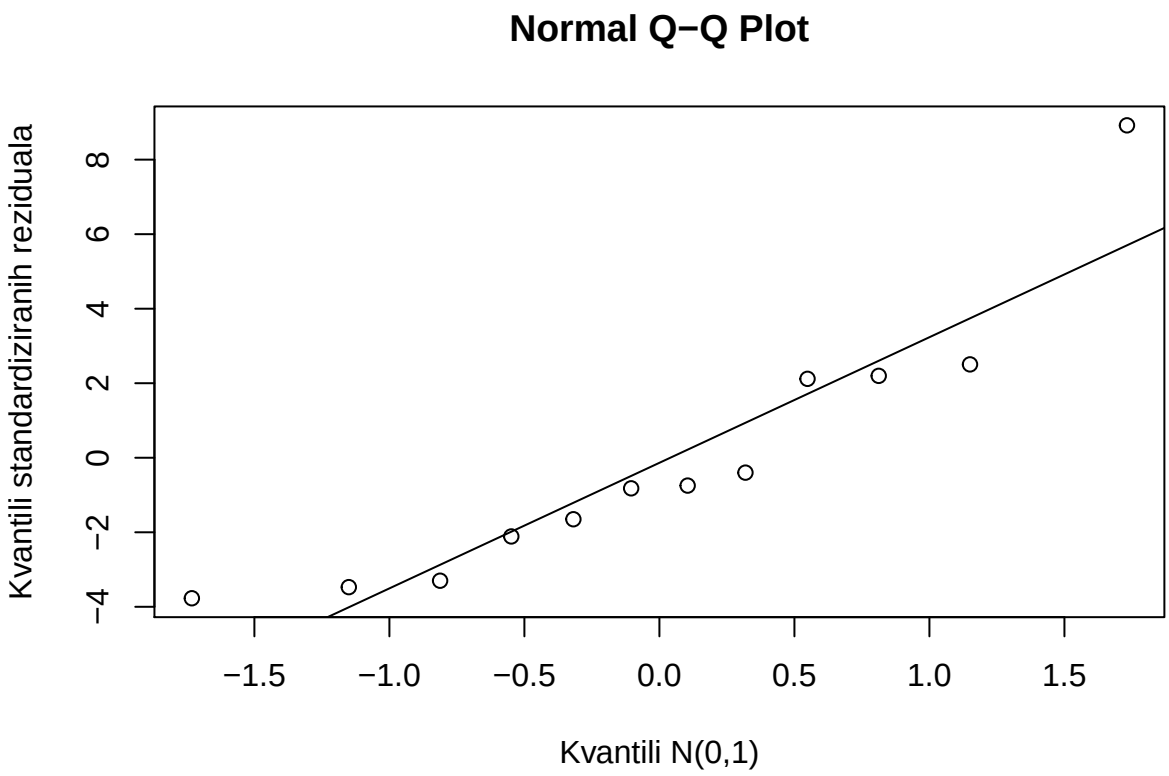
Zbog definicije reziduala, suma reziduala unutar uzorka biti će jednaka 0, što implicira međusobnu zavisnost reziduala. Reziduali, za razliku od šuma, nemaju jednake varijance. Zbog toga je potrebno standardizirati reziduale da bi ih sveli na iste varijance. Standardizirani rezidual je:

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

gdje je $h_{ii} = \frac{1}{n} + \frac{(x_i - \hat{x})^2}{\sum_{j=1}^{n} (x_j - \hat{x})^2}$. Provjera dolaze li standardizirani reziduali 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 dolaze li standardizirani podaci iz jedinične normalne distibucije N(0,1) putem QQ plota.



Kolmogorov Smirnov Test - KS test

KS test uspoređuje testnu i referentnu distribuciju i provjerva jesu li 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) Procjena intervala povjerenja parametra linearne regresije transformiranih podataka

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

```
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: $y = \alpha x^{\beta}$.

```
##
## 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:

$$y' = \log(y)$$

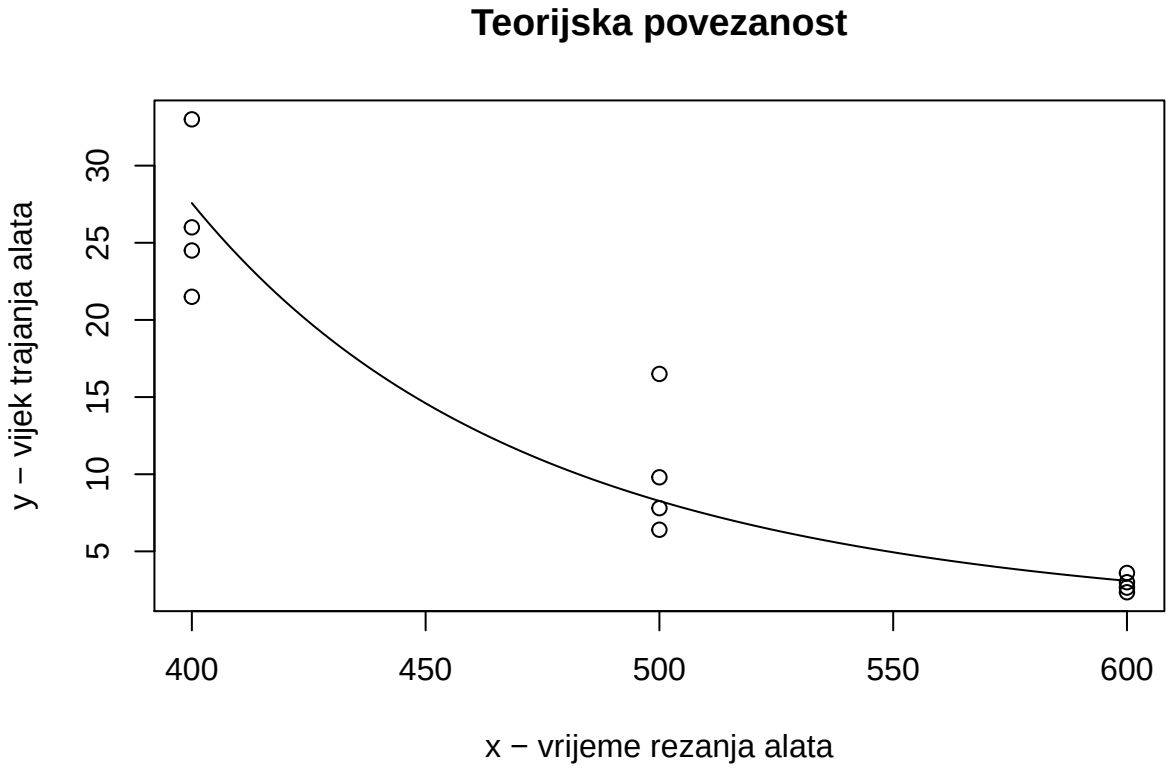
$$x' = \log(x).$$

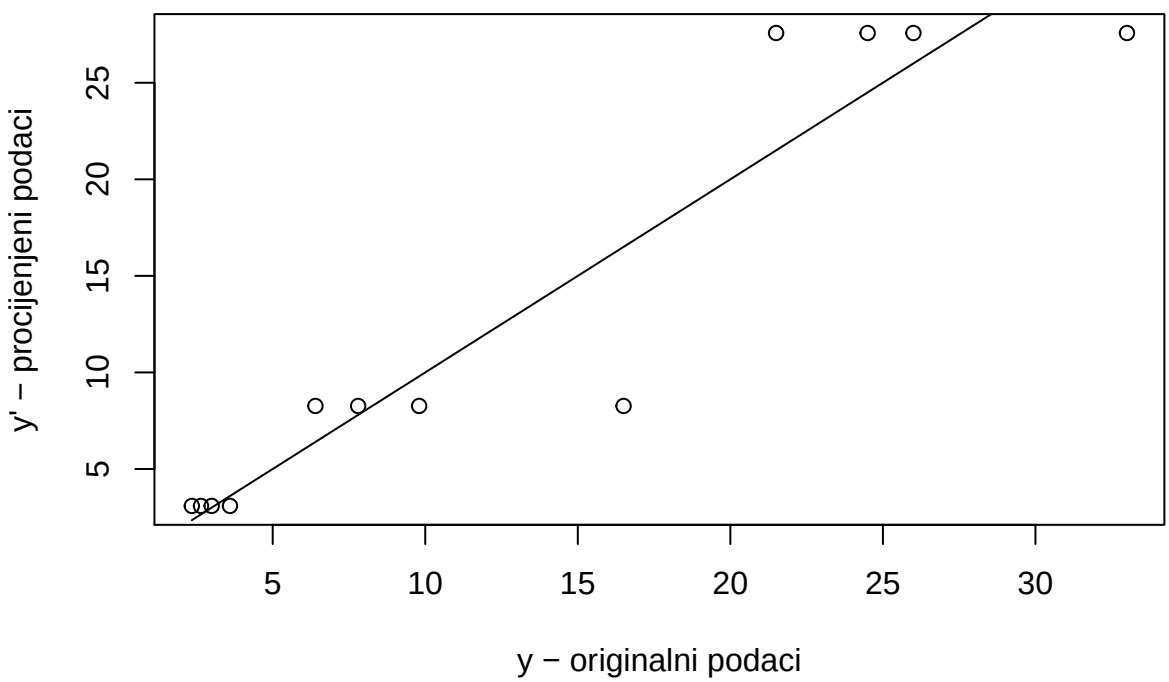
Transformirana teorijska povezanost glasi: $ln(y) = \beta ln(x) + ln(\alpha)$, iz ove jednadžbe slijedi vrijednosti parametara:

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

. Konačno teorijska povezanost je:

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





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:

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

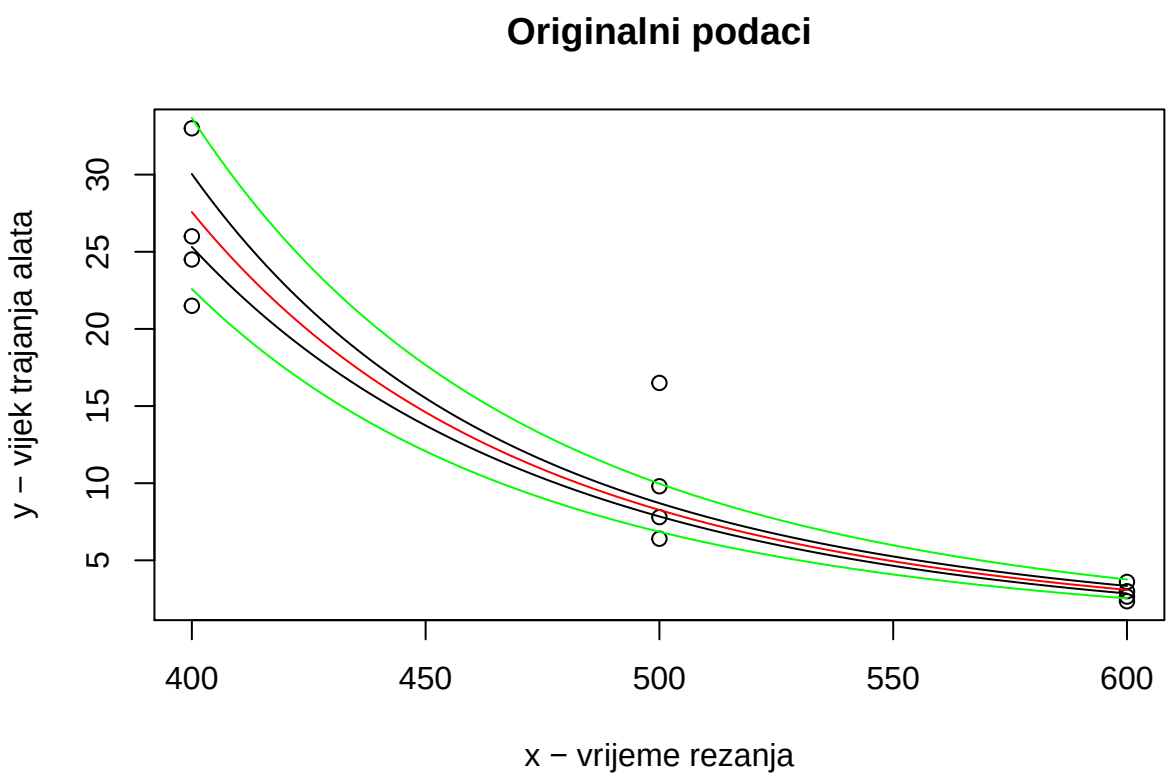
Interval povjerenja za buduću vrijednosti Y

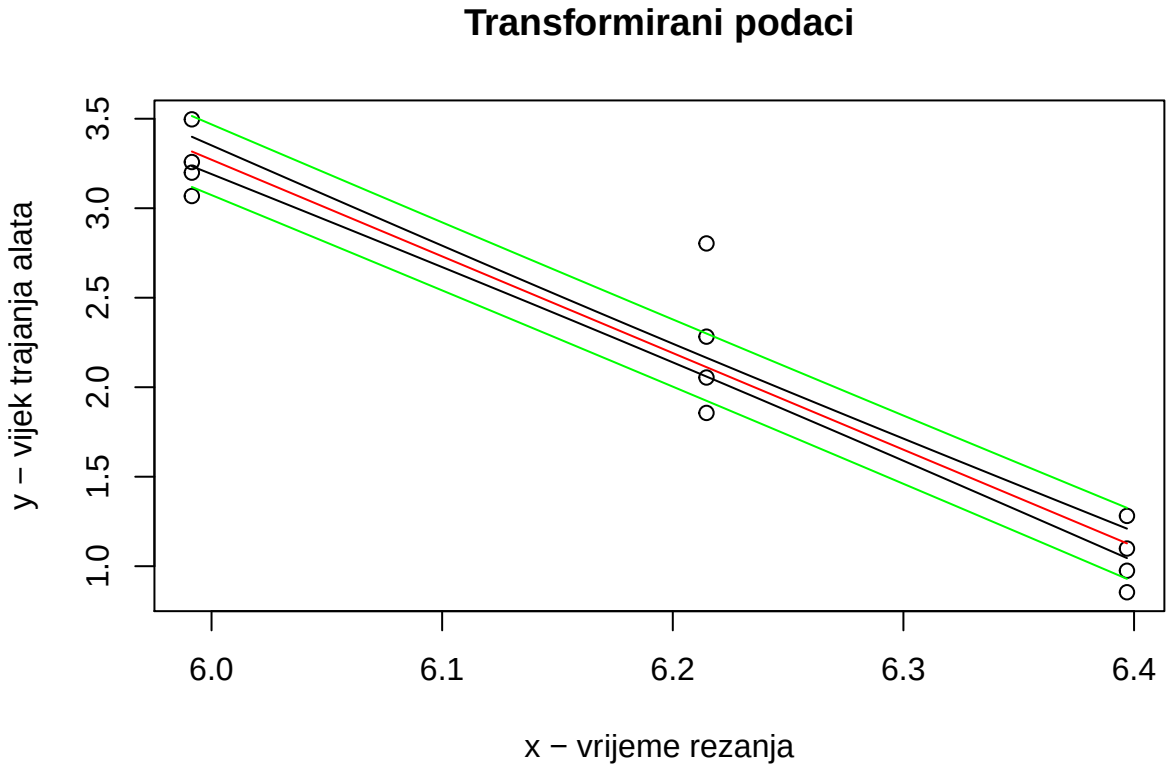
Distribucija uzorkovanja procjenitelja $\hat{Y_0} - Y_0$ je normalna gdje je:

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

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

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





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 $y_i|x_i$

Zadatak C

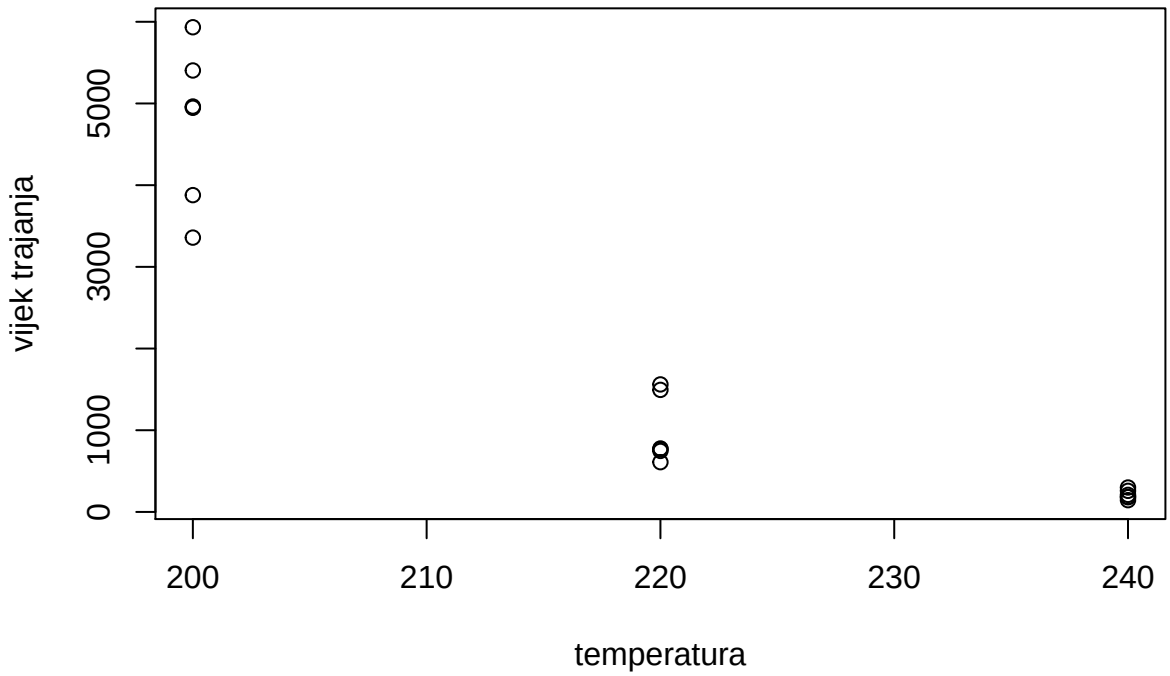
```
podaci=c(200, 5933,
         200, 5404,
         200, 4947,
         200, 4963,
         200, 3358,
         200, 3878,
         220, 1561,
         220, 1494,
         220, 747,
         220, 768,
         220, 609,
         220, 777,
         240, 258,
         240, 299,
         240, 209,
         240, 144,
         240, 180,
         240, 184)

i=1:length(podaci)
```

```
x.index=1==i%%2
y.index=0==i%%2

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

(a) Prikaz podataka u Kartezijevom koordinatnom sustavu



Graf ne sugerira linearnu vezu između podataka. Dok imamo samo tri vrijednosti ordinate, ipak je vidljiv nagli opad vijeka trajanja s porastom temperature, ali i nejednoliko rasipanje vijeka trajanja za različite vrijednosti temperature. Ovo upućuje na nelinearnu vezu.

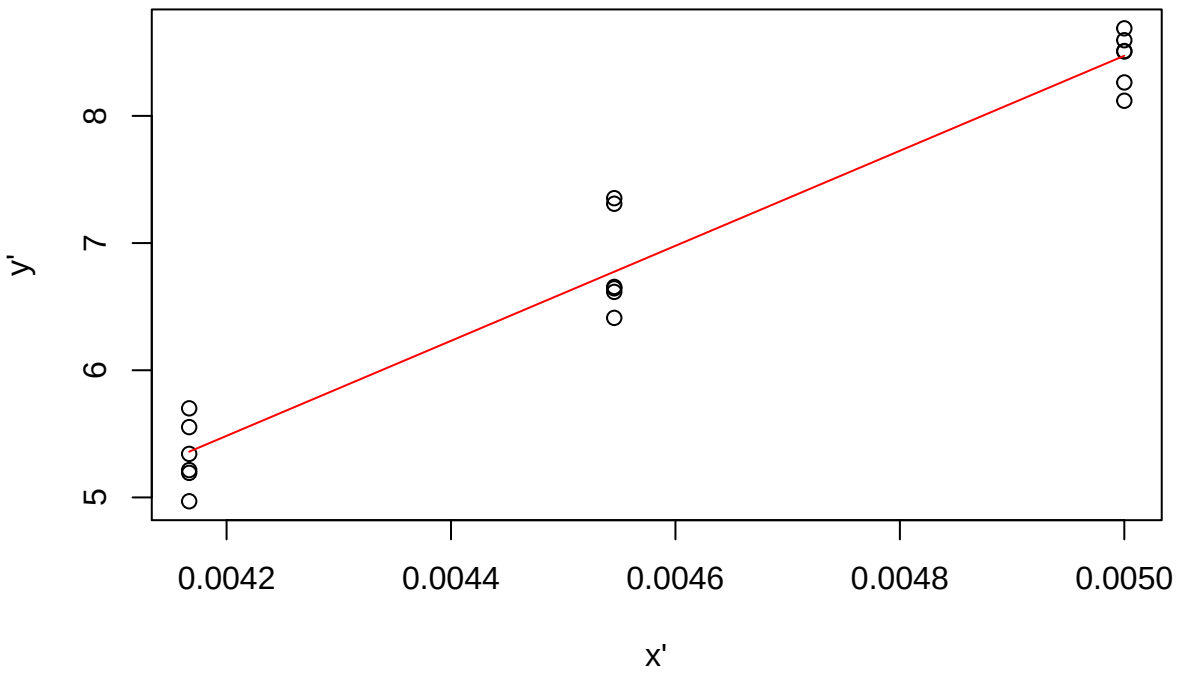
(b) Transformacija podataka

Koristimo transformaciju podataka

$$x' = \frac{1}{x}$$
$$y' = \log y$$

i primjenjujemo linearnu regresiju na parove (x', y').

```
t.x=1/x
t.y=log(y)
plot(t.x,t.y, main="Transformirane vrijednosti", xlab="x'", ylab="y'")
t.fit=lm(t.y~t.x)
lines(t.x,t.fit$fitted.values,col='red')
```




```
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: 0.9543027
```

```
cat("\n")
```

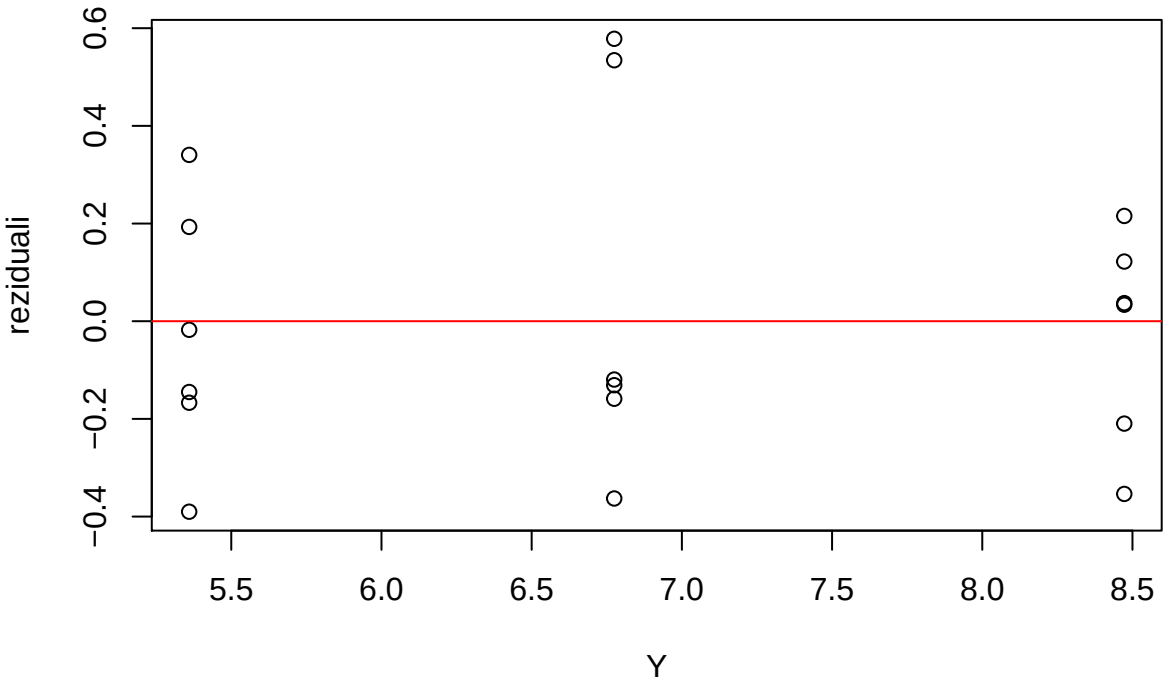
```
library(alr3)
pureErrorAnova(t.fit)
```

```
## Analysis of Variance Table
##
## Response: t.y
##              Df  Sum Sq Mean Sq  F value    Pr(>F)
## t.x           1 29.1501 29.1501 319.9097 1.582e-11 ***
## Residuals    16  1.3959  0.0872
##  Lack of fit  1  0.0291  0.0291   0.3191    0.5805
##  Pure Error  15  1.3668  0.0911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Za *lack-of-fit* je p-vrijednost vrlo visoka, stoga za razumne razine značajnosti ne možemo odbaciti hipotezu da je linearni model adekvatan. Već se i grafički vidi da transformirani podaci izgledaju linerano povezani, te je R^2 vrijednost vrlo blizu 1.

(c) Analiza reziduala

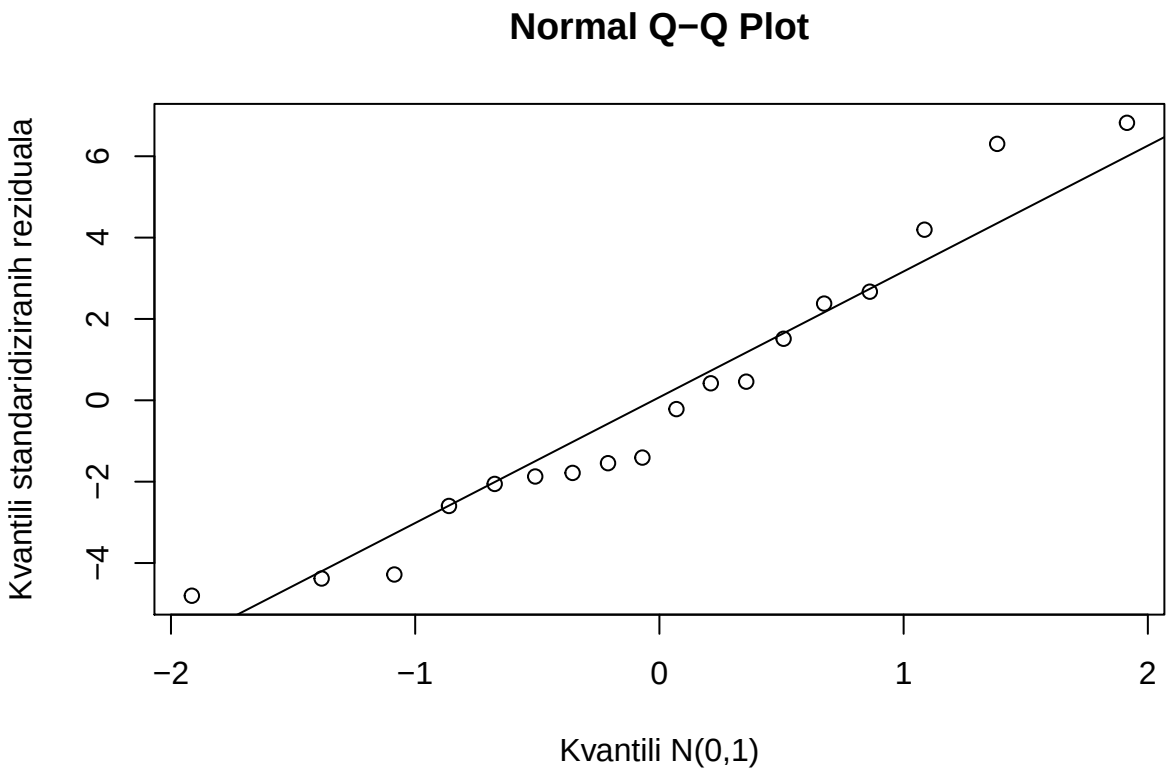
Residual vs. Fit plot



Sve tri ordinate imaju vrijednosti raspršene s obje strane pravca, s time da srednja izgleda neuobičajeno.

Standardizirani reziduali

Normal Q-Q Plot



Kvantili transformiranih reziduala zadovoljavajuće se poklapaju s kvantilima normalne distribucije.

Kolmogorov Smirnov Test

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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ti
## D = 0.42046, p-value = 0.002123
## alternative hypothesis: two-sided
```

Na temelju male p-vrijednosti bismo za uobičajene razine značajnosti odbacili nultu hipotezu da distribucija standardiziranih reziduala prati normalnu razdiobu $N(0,1)$

(d) Procjena intervala povjerenja parametra linearne regresije transformiranih podataka ($\alpha = 0.025$)

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

```
##                 2.5 %      97.5 %
## (Intercept)  -12.1901   -8.218929
## t.x         3302.2377 4168.664590
```

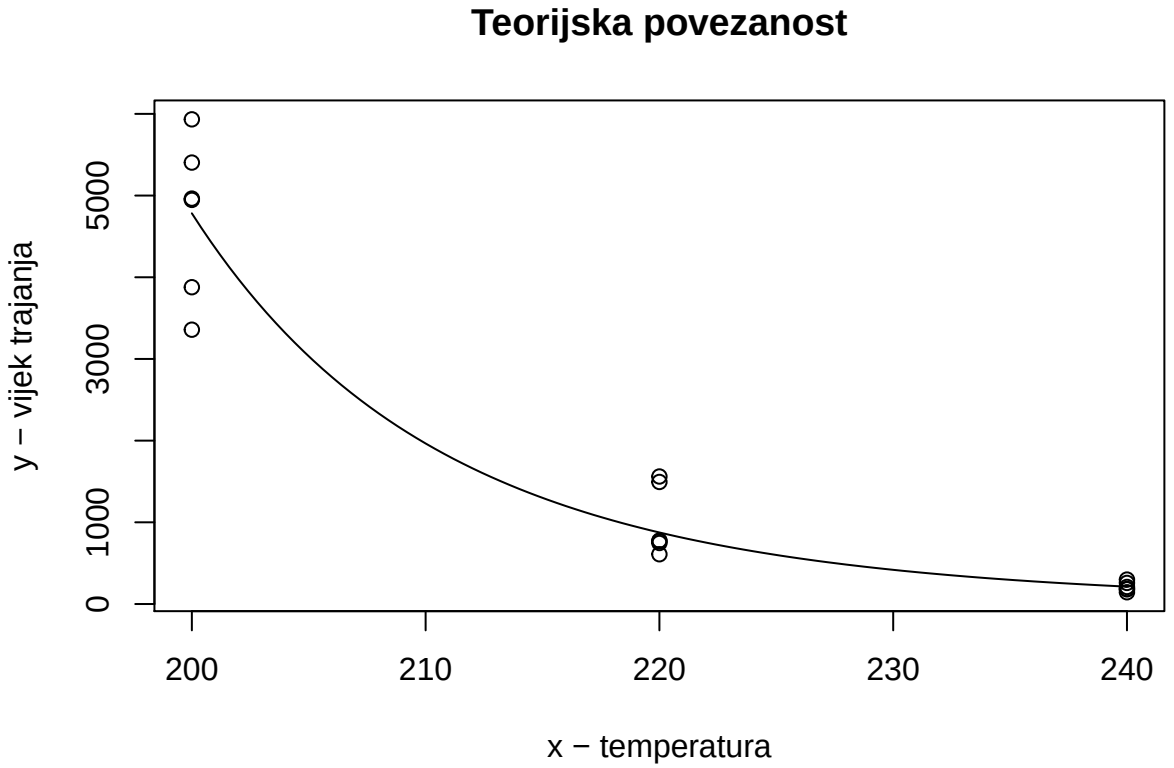
(e) Model za originalne podatke

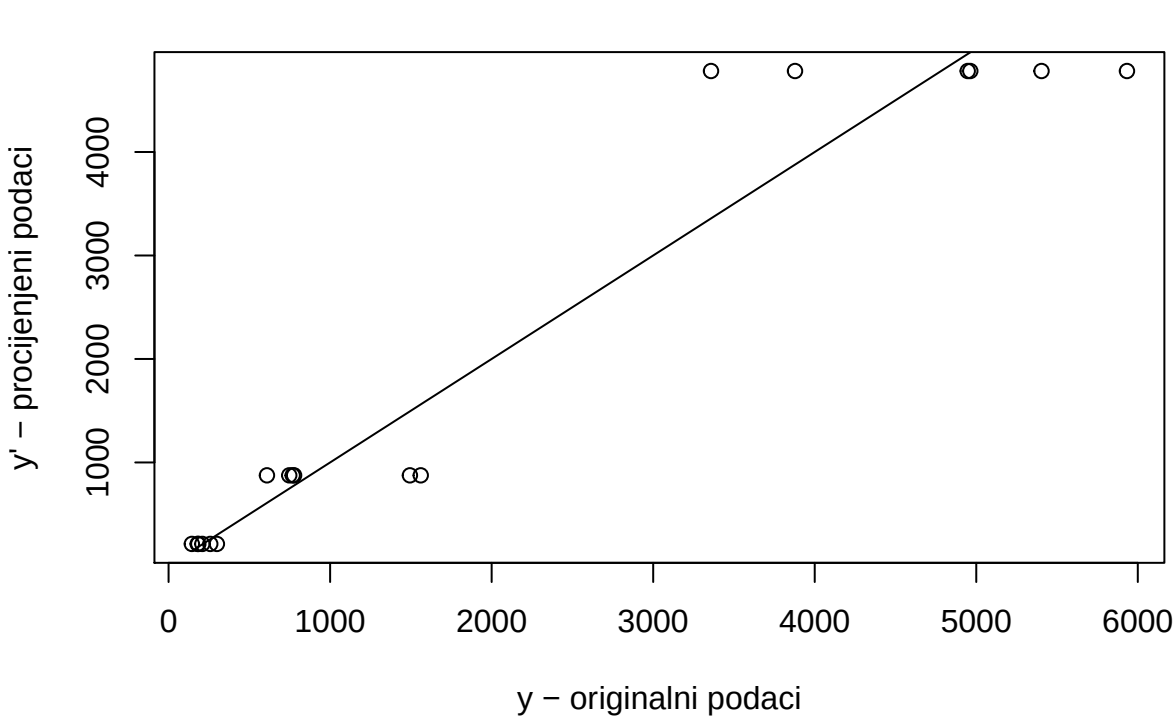
Ako model za transformirane podatke glasi:

$$y' = \beta_0 + \beta_1 x',$$

model za originalne podatke je tada:

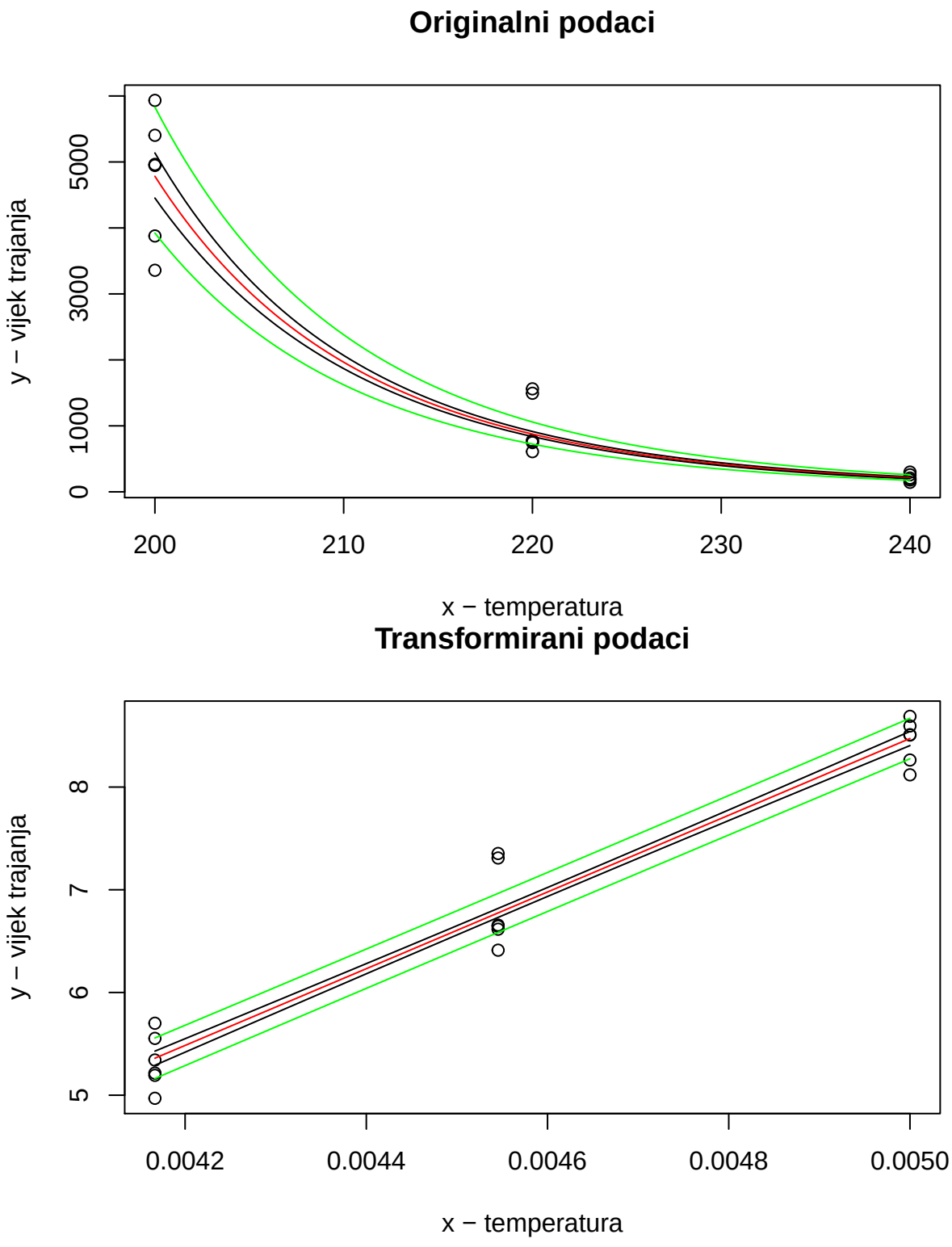
$$\begin{aligned} y &= e^{y'} = e^{\beta_0 + \beta_1 x'} = e^{\beta_0 + \frac{\beta_1}{x}} \\ y &= a b^{\frac{1}{x}} f \\ a &= 3.700293 \times 10^{-5}, b = 1.93089 \times 10^{1622}. \end{aligned}$$





Model daje zadovoljavajuće procjene za originalne podatke, ali varijabilnost podataka je jako velika za manje vrijednosti x pa je korisno pronaći i krivulje povjerenja za y.

(f) Grafički prikaz intervala pouzdanosti



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 $y_i|x_i$