Prepoznavanje eksponencijalnog rasta

Novački, Antonija

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Antonija Novački

PREPOZNAVANJE EKSPONENCIJALNOG RASTA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Marušić

Zagreb, veljača, 2017.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Cilj je ovog diplomskog rada pokazati da suma kvadrata reziduala kod linearne regresije slijedi χ^2 distribuciju. Definirajmo stoga pojmove koji će biti ključni u daljnjem radu. Regresijska je analiza statistički proces za utvrđivanje odnosa između varijabli. Točnije, želimo naći vezu između zavisne i jedne ili više nezavisnih, predikcijskih varijabli. Cilj je pronaći funkciju nezavisnih varijabli koju zovemo regresijskom funkcijom, a koja najbolje opisuje dane podatke. Najjednostavnija je linearna regresija. Kod nje je zavisna varijabla linearna kombinacija nezavisnih varijabli. Često korištena metoda je metoda najmanjih kvadrata (više u narednim poglavljima). Definirajmo još pojam reziduala. Rezidual je pojam kojim definiramo razliku između promatrane vrijednosti zavisne varijable i njene procijenjene vrijednosti. U eksperimentalnom dijelu rada među generiranim podacima tražimo točku u kojoj linearnost prelazi u kvadratičnost.

Poglavlje 1

Linearna regresija

1.1 Općenito o linearnoj regresiji

Neka su $x_1, x_2, ..., x_k$ nezavisne varijable, odnosno varijable poticaja i Y slučajna varijabla mjerena u ovisnosti o $x = (x_1, x_2, ..., x_k)$, odnosno $Y = Y(x)$.
Linearni model ovisnosti varijable Y o varijabli x dan je s

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_k x_k + \epsilon \tag{1.1}$$

gdje je ϵ slučajna pogreška ili šum, a $\theta_0, \theta_1, ..., \theta_k$ su parametri modela.
Neka je

$$(x_{i1}, x_{i2}, ...x_{ik}, Y_i) i = 1, 2, ..., n$$

slučajni uzorak iz linearnog regresijskog modela.
Vektorski zapis:

$$Y = X\theta + \epsilon$$

gdje su

$$\begin{aligned} Y &= (Y_1, Y_2, ..., Y_n)^\tau \\ \epsilon &= (\epsilon_1, \epsilon_2, ..., \epsilon_n)^\tau \\ \theta &= (\theta_0, \theta_1, ..., \theta_k)^\tau \in \mathbb{R}^{k+1} \end{aligned}$$

vektori stupci a X je matrica:

$$X = (1, x_1, x_2, ...x_k) \in M_{n,k+1}(\mathbb{R})$$

kojoj su stupci:

$$1 = (1, 1, ..., 1)^\tau \in \mathbb{R}^n$$
$$x_j = (x_{1j}, x_{2j}, ..., x_{nj})^\tau, j = 1, 2, ..., k$$

Pretpostavlja se da je $n \geq k + 1$.
Pretpostavljamo da za slučajne pogreške vrijede Gauss-Markovljevi uvjeti:

1. $E[\epsilon_i] = 0 \qquad \forall$ i=1,2,...,n
2. $E[\epsilon_i \epsilon_j] = 0 \qquad \forall i, j = 1, 2, ..., n$ takve da je $i \neq j$
3. $Var[\epsilon_i] = \sigma^2 > 0 \qquad \forall i = 1, 2, ..., n$

1.2 Metoda najmanjih kvadrata

Parametre $\theta_0, ..., \theta_k$ dobivamo zakozvanom metodom najmanjih kvadrata. Ona se sastoji u tome da minimiziramo sljedeću funkciju:

$$L(\theta_0, \theta_1, ..., \theta_k) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_{i1} - ... - \theta_k x_{ik})^2 \tag{1.2}$$

Pri tome su $y_1, y_2, ..., y_n$ realizacije slučajne varijable Y.
U grubo govoreći, tražimo takve parametre θ da zbroj kvadrata udaljenosti svake točke y_i do njene procijenjene vrijednosti bude što manji.
Procijenitelj vektora parametara θ metodom najmanjih kvadrata je:

$$\hat{\theta} = (X^\tau X)^{-1} X^\tau Y$$

Pokažimo ovo. Dokaz je preuzet iz [3]

Dokaz. Želimo naći minimum funkcije

$$\begin{aligned} L(\theta) &= |Y - X\theta|^2 \\ &= (Y - X\theta, Y - X\theta) \\ &= |Y|^2 - 2(Y, X\theta) + (X\theta, X\theta) \\ &= |Y|^2 - 2(X^\tau Y, \theta) + ((X^\tau X)\theta, \theta) \end{aligned}$$

Pritom s $(\cdot, \cdot)$ označavamo standardni (euklidski) skalarni produkt, a s $|\cdot|$ standardnu (euklidsku) normu vektora.
Stacionarne točke funkcije L dobivamo iz:

$$0 = \nabla_\theta L(\theta) = -2X^\tau Y + 2X^\tau X\theta$$

Odavde imamo:

$$2X^\tau X\hat{\theta} = 2X^\tau Y,$$

to jest:

$$\hat{\theta} = (X^\tau X)^{-1}X^\tau Y$$

□

Procijenitelji za Y_i su

$$\hat{Y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i1} + ... + \hat{\theta}_k x_{ik}, \text{ za i=1,2,...n.}$$

Nama su od važnosti reziduali. Oni su slučajne varijable oblika:

$$e_i = Y_i - \hat{Y}_i.$$

Poglavlje 2

Distribucija reziduala

Ako imamo linearnu regresiju u kojoj smo procijenili $k+1$ parametara, tada suma kvadrata reziduala ima χ^2 distribuciju s $n-k-1$ stupnjeva slobode skalirano varijancom grešaka. Slijedi dokaz ove tvrdnje. Skica dokaza preuzeta je sa stranice [1]

2.1 Dokaz

Dokaz. Zapišimo prvo našu tvrdnju:

$$\sum \frac{(Y_i - \hat{\theta}_0 - \hat{\theta}_1 x_{i1} - ... - \hat{\theta}_k x_{ik})^2}{\sigma^2} \sim \chi^2_{(n-k-1)} \tag{2.1}$$

Pogledajmo prvo kako izgleda projekcija $\hat{Y}$ vektora Y na potprostor razapet stupcima matrice X. Prisjetimo se izgleda matrice X

$$X = \begin{vmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{vmatrix}$$

Dakle matrica ima n redaka i $k+1$ stupaca.
Projekcija $\hat{Y}$ vektora Y na potprostor razapet stupcima matrice X je

$$\hat{Y} = X\hat{\theta} = X(X^\tau X)^{-1} X^\tau Y \tag{2.2}$$

to jest

$$\hat{Y} = HY, \text{ gdje je } H := X(X^\tau X)^{-1} X^\tau \tag{2.3}$$

Vektor reziduala sada zapisujemo kao

$$e = Y - \hat{Y} = MY, \text{ gdje je } M := I - H \tag{2.4}$$

M i H su ortogonalni projektori u $\mathbb{R}^n$ takvi da vrijedi:
$I = H + M$, $r(H) = k + 1$, $r(M) = n - k - 1$.
Prema tome:

$$\hat{Y} \perp e$$

i

$$e \perp 1, x_1, x_2, ..., x_k$$

Odavde sada slijedi

$$e^\tau 1 = e_1 + e_2 + ... + e_n = 0$$

Iz 2.3 slijedi da sumu kvadrata reziduala, tj. 2.1 možemo pisati kao:

$$\begin{aligned}
\sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{\sigma^2} &= \frac{(Y - \hat{Y})^\tau (Y - \hat{Y})}{\sigma^2} \\
&= \frac{(Y - HY)^\tau (Y - HY)}{\sigma^2} \\
&= \frac{(Y^\tau - Y^\tau H^\tau)(Y - HY)}{\sigma^2} \\
&= \frac{Y^\tau Y - Y^\tau HY - Y^\tau H^\tau Y + Y^\tau H^\tau HY}{\sigma^2} \\
&= \frac{Y^\tau Y - Y^\tau HY - Y^\tau HY + Y^\tau HY}{\sigma^2} \\
&= \frac{Y^\tau Y - Y^\tau HY}{\sigma^2} \\
&= \frac{Y^\tau (I_n - H) Y}{\sigma^2}
\end{aligned} \tag{2.5}$$

Ovo vrijedi jer je matrica H ortogonalni projektor, tj. za nju vrijedi sljedeće: $H = H^\tau$ i $H = H^2$. Pokažimo to:

$$\begin{aligned}
H^\tau &= (X(X^\tau X)^{-1} X^\tau)^\tau \\
&= X(X^\tau X)^{-1} X^\tau \\
&= H
\end{aligned} \tag{2.6}$$

jer je matrica $X^\tau X$ kvadratna pa je transponiranje njenog inverza jednako invertiranju njoj transponirane matrice, tj.

$$((X^\tau X)^{-1})^\tau = ((X^\tau X)^\tau)^{-1} = (X^\tau X)^{-1}$$

Pokažimo sada i $H^2 = H$:

$$\begin{aligned} H^2 &= \Big(X(X^\tau X)^{-1}X^\tau\Big)\Big(X(X^\tau X)^{-1}X^\tau\Big) \\ &= X(X^\tau X)^{-1}(X^\tau X)(X^\tau X)^{-1}X^\tau \\ &= X(X^\tau X)^{-1}X^\tau \\ &= H \end{aligned} \tag{2.7}$$

Također, vrijedi:

$$\begin{aligned} HH^\tau &= \Big(X(X^\tau X)^{-1}X^\tau\Big)\Big(X(X^\tau X)^{-1}X^\tau\Big)^\tau \\ &= X(X^\tau X)^{-1}X^\tau X(X^\tau X)^{-1}X^\tau \\ &= X(X^\tau X)^{-1}X^\tau \\ &= H \end{aligned} \tag{2.8}$$

kao i $H^\tau H = H$.

Teorem 2.1.1. *Svojstvo ortogonalnih projektora je da su njihove svojstvene vrijednosti jednake 0 ili 1.*

Dokaz. Neka je λ svojstvena vrijednost matrice H koja je ortogonalni projektor te neka je v pripadni svojstveni vektor. To zapisujemo ovako:

$$Hv = \lambda v$$

Djelujemo li na tu jednakost projektorom H, slijedi:

$$H(Hv) = H(\lambda v)$$

Lijeva strana ove jednakosti:

$$H(Hv) = H^2 v = Hv$$

ovo slijedi iz 2.7.
Desna strana ove jednakosti:

$$H(\lambda v) = \lambda(Hv) = \lambda(\lambda v) = \lambda^2 v$$

Dakle, imamo $Hv = \lambda v$ i $Hv = \lambda^2 v$, što povlači:

$$\lambda v = \lambda^2 v$$

$\Rightarrow \lambda = 0$ ili $\lambda = 1$. □

Pokažimo sada da je $I_n - H$ također ortogonalni projektor.
Očito je da vrijedi $(I_n - H)^\tau = I_n - H$.
Pokažimo još $(I_n - H)^2 = (I_n - H)$.

$$\begin{aligned}(I_n - H)^2 &= (I_n - H)(I_n - H) \\ &= I_n^2 - I_n H - H I_n + H^2 \\ &= I_n - 2H + H \\ &= I_n - H\end{aligned} \tag{2.9}$$

Također vrijedi pravilo da je suma svojstvenih vrijednosti jednaka tragu matrice.

$$\begin{aligned}tr(I_n - H) &= tr(I_n) - tr(H) \\ &= n - tr(X(X^\tau X)^{-1}X^\tau) \\ &= n - tr((X^\tau X)^{-1}X^\tau X) \\ &= n - tr(I_{k+1}) \\ &= n - k - 1\end{aligned} \tag{2.10}$$

Ovdje smo u trećoj jednakosti koristili da je $tr(AB) = tr(BA)$ ukoliko je A $m \times n$ matrica, a B $n \times m$ matrica. U ovom slučaju X je matrica A i ona je reda $n \times k + 1$, a $(X^\tau X)^{-1}X^\tau$ je matrica B i ona je reda $k + 1 \times n$.
Budući da matrica $I_n - H$ mora imati n svojstvenih vrijednosti, zaključujemo da $I_n - H$ ima n-k-1 svojstvenih vrijednosti jednakih 1 i k+1 svojstvenih vrijednosti jednakih 0.
U nastavku ćemo koristiti sljedeći teorem:

Teorem 2.1.2 (Spektralna dekompozicija)**.** *Svaku simetričnu matricu $A \in \mathbb{R}^n$ možemo zapisati u obliku: $A = UDU^\tau$, gdje je $D = diag(d_1, d_2, ..., d_n)$ dijagonalna sa svojstvenim vrijednostima od A na dijagonali, a $U = [u_1, u_2, ..., u_n]$ je ortogonalna ($U^\tau U = UU^\tau = I_n$) i sadrži svojstvene vektore od A.*

Dakle, ovaj teorem nam govori da za simetričnu matricu $I_n - H$ postoje dijagonalna matrica D i ortogonalna matrica U takve da je da je $I_n - H = UDU^\tau$.
Pritom se na dijagonali od D nalaze svojstvene vrijednosti od $I_n - H$, to jest D je oblika:

$$D = \begin{pmatrix} I_{n-k-1} & 0_{[n-k-1]\times[k+1]} \\ 0_{[k+1]\times[n-k-1]} & 0_{[k+1]\times[k+1]} \end{pmatrix}$$

budući da znamo da ima n-k-1 svojstvenih vrijednosti jednakih 1 i k+1 svojstvenih vrijednosti jednakih 0.
Sljedeće korisno svojstvo je da je $HX = X(X^\tau X)^{-1}X^\tau X = X$. To nam pomaže odrediti

matricu U.
$HX = X \Rightarrow$

$$\begin{aligned}(I_n - H)X &= X - HX \\ &= X - X \\ &= 0\end{aligned}$$

to jest, zapisano pomoću gornje dekompozicije
$UDU^\tau X = 0 \Rightarrow DU^\tau X = 0 \Rightarrow$

$$(U^\tau X)_{ij} = 0 \text{ za i=1,2,...,n-k-1, j=1,2,...,k+1.} \tag{2.11}$$

Dakle, naš originalni problem $\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sigma^2} = \frac{Y^\tau(I_n - H)Y}{\sigma^2}$ možemo zapisati u obliku

$$\begin{aligned}\frac{Y^\tau UDU^\tau Y}{\sigma^2} &= \frac{(U^\tau Y)^\tau DU^\tau Y}{\sigma^2} \\ &= \frac{\sum_{i=1}^{n} D(U^\tau Y)_i^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^{n-k-1}(U^\tau Y)_i^2}{\sigma^2}\end{aligned}$$

Prema našem modelu $Y \sim N(X\theta, \sigma^2 I)$, što povlači da je

$$\begin{aligned}U^\tau Y &\sim N(U^\tau X\theta, U\sigma^2 IU^\tau) \\ &\sim N(U^\tau X\theta, \sigma^2 UU^\tau) \\ &\sim N(U^\tau X\theta, \sigma^2 I)\end{aligned}$$

Ovo pokazuje da su komponente od $U^\tau Y$ nezavisne.
Također, zbog 2.11 slijedi da je $(U^\tau Y)_i \sim N(0, \sigma^2)$ za i=1,2,..n-k-1

Nekoliko važnih tvrdnji:

1. $Z \sim N(0,1) \Rightarrow Z^2 \sim \chi^2(1)$
2. $Z_i \sim \chi^2(1)$ i Z_i su međusobno nezavisne (i=1,2,...,n) $\Rightarrow \sum_{i=1}^{n} Z_i \sim \chi^2(n)$

Pokažimo tvrdnju 1. Skica dokaza nalazi se na stranici [2]:

Dokaz. Označimo sa $V = Z^2$, gdje je Z standardna normalna $Z \sim N(0,1)$.
Moramo dokazati da V ima χ^2 distribuciju. χ^2 distribucija s n stupnjeva slobode je specijalni oblik gama distribucije ($\Gamma(\alpha,\beta)$) uz vrijednosti parametara $\alpha = \frac{n}{2}, \beta = \frac{1}{2}$.
U našem slučaju $n = 1$. Dakle, moramo pokazati sljedeće: $V \sim \Gamma(\frac{1}{2}, \frac{1}{2}) = \chi^2(1)$.
Funkcija gustoće vjerojatnosti gama funkcije izgleda ovako:

$$f(x) = \begin{cases} 0 & x \leq 0 \\ C \cdot x^{\alpha-1} e^{-\beta x} & x > 0 \end{cases} \tag{2.12}$$

gdje je $C = \dfrac{\beta^\alpha}{\Gamma(\alpha)}$, a $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ gama funkcija , $x > 0$.
Dakle funkcija gustoće vjerojatnosti χ^2 distribucije s jednim stupnjem slobode izgleda ovako:

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \dfrac{(\frac{1}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} \cdot x^{\frac{1}{2}-1} e^{-\frac{1}{2}x} & x > 0 \end{cases} \tag{2.13}$$

Znači mi moramo pokazati da je funkcija gustoće slučajne varijable $V = Z^2$ jednaka upravo:

$$g(v) = \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} \cdot v^{-\frac{1}{2}} e^{-\frac{1}{2}v} \tag{2.14}$$

za $v > 0$.
Označimo sada s

$$G(v) = P(V \leq v) = P(Z^2 \leq v)$$

kumulativnu funkciju distribucije od V. Što možemo zapisati kao:

$$G(v) = P(-\sqrt{v} \leq Z \leq \sqrt{v})$$

Sada integriramo funkciju gustoće standardne normalne varijable Z nad intervalom $[-\sqrt{v}, \sqrt{v}]$.

$$G(v) = \int_{-\sqrt{v}}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-z^2}{2}\right) dz \tag{2.15}$$

Zbog simetričnosti normalne distribucije, ovo možemo zapisati kao:

$$G(v) = 2\int_{0}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-z^2}{2}\right) dz \tag{2.16}$$

Sada radimo sljedeću zamjenu varijabli:

$$z = \sqrt{y} = y^{1/2} \Rightarrow$$
$$dz = \frac{1}{2}y^{-1/2}dy = \frac{1}{2\sqrt{y}}dy$$

pa je $z^2 = y$ i granice integrala su:

$$z = 0 \Rightarrow y = 0, z = \sqrt{v} \Rightarrow y = v$$

Sada imamo:

$$\begin{aligned} G(v) &= 2\int_0^v \frac{1}{\sqrt{2\pi}} exp\Big(\frac{-y}{2}\Big)\Big(\frac{1}{2\sqrt{y}}\Big)dy \\ &= \int_0^v \frac{1}{\sqrt{2}\sqrt{\pi}} y^{-1/2} exp\Big(\frac{-y}{2}\Big)dy \text{ ,za } v > 0. \end{aligned} \tag{2.17}$$

Iz Newton-Leibnizove formule:

$$\int_0^x f(t)dt = F(x) - F(0)$$

kao i :

$$\frac{d}{dx}\int_0^x f(t)dt = F'(x) = f(x)$$

možemo derivirati $G(v)$ da dobijemo funkciju gustoće $g(v)$:

$$g(v) = G'(v) = \frac{1}{\sqrt{2}\sqrt{\pi}} v^{-\frac{1}{2}} e^{-\frac{1}{2}v} \text{ za } 0 < v < \infty$$

Ako zadnju jednakost usporedimo s 2.14, uočavamo da su one jednake ukoliko vrijedi:

$$\Gamma(\frac{1}{2}) = \sqrt{\pi}$$

Budući da je $g(v)$ funkcija gustoće slučajne varijable, njen integral na području definicije mora biti jednak 1, to jest mora vrijediti sljedeće:

$$\int_0^\infty \frac{1}{\sqrt{2}\sqrt{\pi}} v^{-\frac{1}{2}} e^{-\frac{1}{2}v} dv = 1 \tag{2.18}$$

Sada uvodimo sljedeću zamjenu varijabli:

$$v = 2x$$
$$dv = 2dx$$

Sada dobivamo:

$$\frac{1}{\sqrt{\pi}} \int_0^\infty \frac{1}{\sqrt{2}} (2x)^{-\frac{1}{2}} e^{-x} 2dx = 1 \tag{2.19}$$

Što možemo ljepše zapisati kao:

$$\frac{1}{\sqrt{\pi}} \int_0^\infty \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} x^{-\frac{1}{2}} e^{-x} 2dx = 1 \tag{2.20}$$

ili kada skratimo:

$$\frac{1}{\sqrt{\pi}} \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = 1 \tag{2.21}$$

Sada je, iz definicije gama funkcije, jasno da imamo:

$$\frac{1}{\sqrt{\pi}} \Gamma(\frac{1}{2}) = 1 \tag{2.22}$$

odnosno, vrijedi $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
Ovime je tvrdnja 1. dokazana. □

Pokažimo sada i tvrdnju 2. da zbroj n nezavisnih slučajnih varijabli koje sve imaju $\chi^2(1)$ distribuciju ima χ^2 distribuciju s n stupnjeva slobode.

Dokaz. Dokazat ćemo ovu tvrdnju matematičkom indukcijom.
Baza: pokazujemo da tvrdnja vrijedi za 2 nezavisne slučajne varijable. Neka su X i Y dvije nezavisne slučajne varijable takve da vrijedi $X \sim \chi^2(1)$ i $Y \sim \chi^2(1)$. Tada su njihove funkcije gustoće jednake:

$$f_X(x) = \frac{1}{\sqrt{(2)}\Gamma(\frac{1}{2})} x^{\frac{-1}{2}} e^{\frac{-1}{2}x}$$
$$f_Y(y) = \frac{1}{\sqrt{(2)}\Gamma(\frac{1}{2})} y^{\frac{-1}{2}} e^{\frac{-1}{2}y}$$

Zbog nezavisnosti, vrijedi sljedeće:

$$f_{X+Y}(x, y) = f_X(x) f_Y(y)$$

odnosno, kada uvrstimo

$$f_{X+Y}(x, y) = \frac{1}{2\Gamma(\frac{1}{2})^2} (xy)^{-\frac{1}{2}} e^{-\frac{1}{2}(x+y)}$$

Budući da je u dokazu tvrdnje 1. pokazano da vrijedi $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, slijedi:

$$f_{X+Y}(x,y) = \frac{1}{2\pi}(xy)^{-\frac{1}{2}}e^{-\frac{1}{2}(x+y)}$$

Sada uvodimo zamjenu varijabli: $a = xy$, $b = x + y$. Izrazimo sada x i y preko a i b.
$a = xy = x(b - x) = bx - x^2$, to jest dobivamo kvadratnu jednadžbu:

$$x^2 - bx + a = 0$$

Iz čega slijedi da je

$$x_{1,2} = \frac{b \pm \sqrt{b^2 - 4a}}{2}$$

Za $x = \frac{b + \sqrt{b^2 - 4a}}{2}$, dobivamo $y = \frac{b - \sqrt{b^2 - 4a}}{2}$, a za $x = \frac{b - \sqrt{b^2 - 4a}}{2}$, dobivamo $y = \frac{b + \sqrt{b^2 - 4a}}{2}$.
Budući da su jednadžbe simetrične, možemo uzeti da vrijedi prva i na kraju pomnožiti Jakobijan s 2.
Izračunajmo sada Jakobijan:

$$J_{(A,B)}(a,b) = \begin{vmatrix} -(b^2 - 4a)^{-\frac{1}{2}} & \frac{1 + b(b^2 - 4a)^{-\frac{1}{2}}}{2} \\ (b^2 - 4a)^{-\frac{1}{2}} & \frac{1 - b(b^2 - 4a)^{-\frac{1}{2}}}{2} \end{vmatrix} = (b^2 - 4a)^{-\frac{1}{2}}$$

Sada možemo izračunati $f_{(A,B)}(a,b)$.

$$f_{(A,B)}(a,b) = 2 \cdot \frac{1}{2\pi}a^{-\frac{1}{2}}e^{-\frac{b}{2}}(b^2 - 4a)^{-\frac{1}{2}}$$

Mi želimo f_B pa gornju jednadžbu integriramo po a.

$$f_B(a,b) = 2 \cdot \frac{e^{-\frac{b}{2}}}{2\pi}\int_0^{\frac{b^2}{4}} a^{-\frac{1}{2}}(b^2 - 4a)^{-\frac{1}{2}}da$$

Sada uvodimo novu supstituciju: $a = \frac{b^2}{4}\sin^2(t) \Rightarrow da = \frac{b^2}{4}\sin(2t)dt$.
Iz ovoga slijedi:

$$\begin{aligned} f_B(b) &= 2 \cdot \frac{e^{-\frac{b}{2}}}{2\pi} \int_0^{\frac{\pi}{2}} \frac{2}{b\sin t}(b^2 - b^2\sin^2 t)^{-\frac{1}{2}} \frac{b^2}{4}\sin(2t)dt \\ &= \frac{e^{-b/2}}{\pi} \int_0^{\frac{\pi}{2}} \frac{1}{\sin t\sqrt{b^2\cos^2 t}} \frac{b}{2}\sin(2t)dt \\ &= \frac{e^{-b/2}}{\pi} \int_0^{\frac{\pi}{2}} \frac{b}{2b\sin t\cos t}\sin(2t)dt \\ &= \frac{e^{-b/2}}{\pi} \int_0^{\frac{\pi}{2}} dt \\ &= \frac{e^{-b/2}}{\pi} \end{aligned}$$

Dakle, dobili smo $f_{(X+Y)}(x,y) = \dfrac{e^{-(x+y)/2}}{\pi}$.
S druge strane, funkcija gustoće varijable $Z \sim \chi^2(2)$ odgovara funkciji gustoće slučajne varijable koja ima distribuciju $\Gamma(1, \frac{1}{2})$.

$$f_Z(z) = \frac{\frac{1}{2}}{\Gamma(1)} z^{1-1} e^{-\frac{1}{2}z} = \frac{1}{2\Gamma(1)} e^{-\frac{1}{2}z}$$

Budući da vrijedi $\Gamma(n) = (n-1)!$, slijedi da je $\Gamma(1) = 0! = 1$ pa je

$$f_Z(z) = \frac{e^{-\frac{1}{2}z}}{2}$$

Time je baza dokazana.

□

Vratimo se na sumu kvadrata reziduala. Za sada znamo:

1. $(U^\tau Y)_i$ su nezavisne slučajne varijable
2. $(U^\tau Y)_i \sim N(0, \sigma^2)$, to jest $\dfrac{(U^\tau Y)_i}{\sigma} \sim N(0,1)$, za i=1,2,...n-k-1

Iz prethodno dokazane tvrdnje 1 slijedi:

$$\frac{(U^\tau Y)_i^2}{\sigma^2} \sim \chi^2(1) \text{ i=1,2,...,n-k-1} \tag{2.23}$$

A sada iz tvrdnje 2 slijedi konačan rezultat:

$$\sum_{i=1}^{n-k-1} \frac{(U^\tau Y)_i^2}{\sigma^2} \sim \chi^2(n-k-1) \tag{2.24}$$

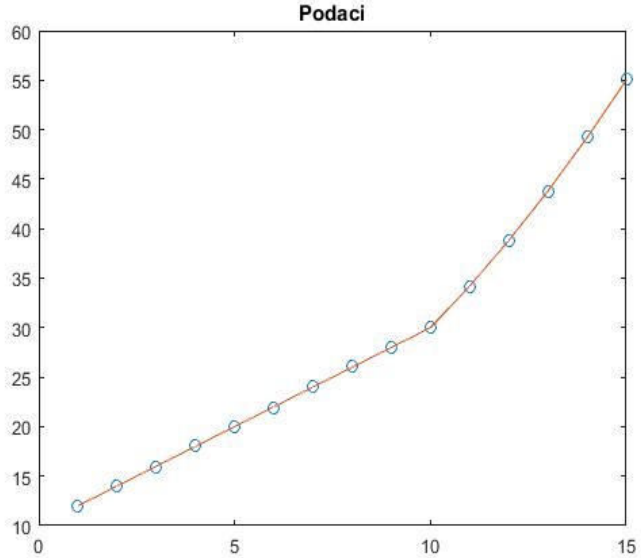
□

Poglavlje 3

Eksperimantalni dio

3.1 Simulirani podaci

Ovaj dio rada pokazuje kako se pomoću reziduala linearne regresije i određenih statističkih testova u podacima određuje eksponencijalna faza rasta. U programskom jeziku MATLAB generiraju se podaci, provode simulacije i testiraju dvije metode. Te metode ćemo kasnije testirati na realnim podacima koje ćemo prvo logaritmirati. Zato su podaci generirani tako da se prvih nekoliko točaka nalazi na pravcu, a ostale točke slijede neku kvadratnu funkciju. Također se dodaje šum. Cilj je pomoću dvije metode pronaći prvu točku koja nije na pravcu i usporediti učinkovitost tih metoda. Neka podaci izgledaju ovako:



Podaci su generirani na sljedeći način: prvih 10 točaka smješteno je na pravac $y = 2x + 10$, a sljedećih 5 na parabolu

$y = 0.2x^2 + 10$ te je dodan normalni šum (očekivanje=0, standardna devijacija=0.05). Pripadni kod u MATLAB-u:

```
a=2;
b=10;
x1=1:10;
x2=11:15;
x=[x1, x2];
y1=a*x1+b;
y2=0.2*x2.^2+b;
y=[y1 y2];
n=length(x);
e=0.05*randn(1,n);
y=y+e;
figure(1)
plot(x,y,'o');
hold on
plot(x,y-e);
title('Podaci');
```

Cilj je da obje metode izbace točku 11 kao prvu točku koja nije na pravcu.

Prva metoda

Prva metoda sastoji se u tome da se uspoređuje linearni i kvadratični model. Dakle, uzme se prvih n točaka i fitaju se pravcem, a potom parabolom. Potom se gleda koji model bolje opisuje podatke. Ukoliko je to pravac, postupak se ponavlja, s time da se sada uzima $n + 1$ točaka. Ukoliko je to parabola, nađena je točka koja nije na pravcu. Kako se uspoređuje je li bolji pravac ili parabola? Gledaju se reziduali, izračuna testna statistika i odredi p-vrijednost testa. Testiraju se sljedeće hipoteze:

$$H_0 = \text{linearni model bolje opisuje podatke}$$
$$H_1 = \text{kvadratični model bolje opisuje podatke}$$

Kod u MATLAB-u izgleda ovako:

```
koef1=polyfit(x(1:i),y(1:i),1);
y1=polyval(koef1,x(1:i));
koef2=polyfit(x(1:i),y(1:i),2);
y2=polyval(koef2,x(1:i));
```

```
%sume kvadrata reziduala
rez1=sum((y(1:i)-y1).^2);
rez2=sum((y(1:i)-y2).^2);
% za usporedbu modela koristimo F statistiku
%  F=(SS1-SS2)/(df1-df2) / SS2/df2

df1=i-2;
df2=i-3;
F=((rez1-rez2)/(df1-df2))/(rez2/df2);
pv=1-fcdf(F,df1,df2);
% p vrijednost~0 -> odbacujemo nultu hipotezu da je linearni
% model bolji
```

Dakle, koristi se statistika

$$F = \frac{\frac{rez1-rez2}{df1-df2}}{\frac{rez2}{df2}}$$

gdje su:

rez1 - suma kvadrata reziduala kod linearnog fita,

rez2 - suma kvadrata reziduala kod kvadratičnog fita,

df1 - broj stupnjeva slobode kod linearnog fita,

df2 - broj stupnjeva slobode kod kvadratičnog fita.

Kada se izračuna ova statistika, lako se dobije p-vrijednost pomoću naredbe $fcdf(F, df1, df2)$. Ukoliko je p-vrijednost 'velika', ne može se odbaciti nulta hipoteza, to jest bolji je linearni model, a ukoliko je 'mala', odbacuje se nulta hipoteza u korist alternative, odnosno, može se reći da je bolji kvadratični model. Rađena je analiza za razine značajnosti 1%, 3%, 5%, 7% i 10%. Vidjet će se da se rezultati podosta razlikuju za različite razine značajnosti.

Da bi se vidjelo koliko je ova metoda dobra za detektiranje prve točke koja nije na pravcu, provode se simulacije. Podaci su generirani 10000 puta pri čemu je mijenjan samo šum pomoću MATLAB funkcije *randn* (očekivanje=0, standardna devijacija=0.05). Kao što je već rečeno, uzima se 15 točaka, od kojih je prvih 10 na pravcu, a zadnjih 5 na paraboli. Dakle, točka koja se traži je 11. Slijedi tablica pogodaka s obzirom na različite razine značajnosti.

nivo značajnosti	broj pogodaka 11	srednja vrijednost pogodaka	standardna devijacija pogodaka
0.01	429	11.8361	0.7277
0.03	8774	10.7962	0.9247
0.05	8563	10.61	0.7277
0.07	8208	10.4893	1.3239
0.1	7785	10.3278	1.5166

Dakle, uzimajući nivo značajnosti 1%, dobra točka je pronađena samo 429 puta od 10000, kod nivoa značajnosti 3% dobra točka je pronađena 8774 puta. Može se uočiti da se broj pogodaka smanjuje kako raste razina značajnosti. Izuzetak je jedino razina značajnosti od 1% gdje je broj pogodaka neuobičajeno malen. Vidimo da je u tom slučaju srednja vrijednost pogodaka 11.8361, dakle metoda često puta pogađa točke veće od 11 što ćemo kasnije vidjeti i na histogramu.

Sljedeće što se provjerava je utječe li promjena standardne devijacije šuma na točnost podataka. Ova tablica je rađena za nivo značajnosti 5%.

standardna devijacija šuma	broj pogodaka 11	srednja vrijednost pogodaka	standardna devijacija pogodaka
0.05	8563	10.61	0.7277
0.075	8524	10.6263	1.1562
0.1	8280	10.6703	1.1799
0.25	5520	10.9665	1.3326
0.5	3263	11.2025	1.4116

Primjećuje se da kako raste standardna devijacija šuma, tako se broj pogodaka smanjuje. Međutim, razlika nije velika kod prvih triju vrijednosti, ali jest kod posljednje dvije. Kada je devijacija jednaka 0.075 u odnosu na devijaciju jednaku 0.05, razlika u broju pogodaka iznosi samo 39 ili 0.39%. Kada je devijacija jednaka 0.1 razlika je 2.83%. To su, dakle, male razlike pa se može reći da standardna devijacija, ukoliko je manja od 0.1, ne utječe znatno na točnost modela. Ipak, za veće vrijednosti ta razlika je sve veća i devijacija jako utječe na točnost.

Preostaje još provjeriti kako broj točaka utječe na točnost. Uzme se prvo duplo veći broj točaka. Razmak među točkama je sada 0.5, to jest sada su nam točke 1, 1.5, 2, 2.5, i tako dalje do 15. U ovom slučaju, prva točka koja nije na pravcu (točka 11) trebala bi se nalaziti na dvadeset i prvom mjestu.

razina značajnosti	broj pogodaka 21	srednja vrijednost pogodaka	standardna devijacija pogodaka
0.03	6986	18.8846	4.0067
0.05	6302	18.1719	4.588
0.07	5806	17.6060	4.9762

Sve su dobivene vrijednosti lošije, nego u prvom slučaju. Broj pogodaka i ovdje opada kako povećavamo razinu značajnosti.

Broj točaka može se i smanjiti, to jest uzme se da je razmak među točkama 1.5, s time da počinjemo od 0.5 i završavamo s 15.5. Prva točka koja nije na pravcu se sada nalazi na osmom mjestu.

razina značajnosti	broj pogodaka 8	srednja vrijednost pogodaka	standardna devijacija pogodaka
0.03	92	8.9305	0.4462
0.05	4720	8.3857	0.7250
0.07	8634	7.9148	0.5918

Može se uočiti da su vrijednosti puno lošije. Također, u ovom slučaju nema pravila da se točnost pogodaka s povećanjem razine značajnosti smanjuje. Zaključujemo da broj točaka znatno utječe na točnost metode.

Druga metoda

Druga metoda sastoji se u tome da se uspoređuju dva linearna modela. Pretpostavimo da imamo n točaka. Prvo pronađemo najbolji pravac kroz te točke (u smislu da je suma kvadrata reziduala najmanja). Na početnih n točaka dodamo još jednu točku, dakle, sada imamo $n + 1$ točku. Sada pronađemo najbolji pravac kroz te točke. Postupak je dalje sličan kao u prethodnoj metodi. Pomoću kvadrata reziduala izračuna se testna statistika i p-vrijednost. Testna statistika u ovom slučaju izgleda ovako

$$F = \frac{(rez2 - rez1)/(df2 - df1)}{rez1/df1}$$

uz oznake jednake kao i u prvoj metodi. Dio koda u MATLAB-u koji prikazuje kako dolazimo do testne statistike:

```
% prvo fitam pravcem na i tocaka
koef1=polyfit(x(1:i),y(1:i),1);
y1=polyval(koef1,x(1:i));
% zatim fitam pravcem na i+1 tocki
koef2=polyfit(x(1:i+1),y(1:i+1),1);
```

```
y2=polyval(koef2,x(1:i+1));

%sume kvadrata reziduala
rez1=sum((y(1:i)-y1).^2);
rez2=sum((y(1:i+1)-y2).^2);
% za usporedbu modela koristimo F statistiku
%  F=(SS1-SS2)/(df1-df2) / SS2/df2

df1=i-2;  % broj podataka je i umanjeno
                % za 2 procijenjena parametra
df2=i-1;  % broj podataka je i+1 umanjeno
              % za 2 procijenjena parametra
F=((rez2-rez1)/(df2-df1))/(rez1/df1);
pv=1-fcdf(F,df1,df2);
```

Hipoteze definiramo na sljedeći način:

H_0 = pravac kroz n točaka bolje opisuje podatke
H_1 = pravac kroz n+1 točku bolje opisuje podatke

Ukoliko se dobije 'mala' p-vrijednost, odbacuje se nulta hipoteza u korist alternative, to jest bolji 'fit' imamo u slučaju s jednom točkom više, što znači da smo pronašli točku koja nije na pravcu. Ponovo provodimo 10000 simulacija pri čemu imamo iste podatke kao kod prve metode i samo mijenjamo šum pomoću MATLAB funkcije randn (očekivanje šuma=0, standardna devijacija šuma=0.05).
U sljedećoj tablici prikazano je kako razina značajnosti utječe na broj pogodaka.

nivo značajnosti	broj pogodaka 11	srednja vrijednost pogodaka	standardna devijacija pogodaka
0.01	9532	10.9291	0.4475
0.03	9020	10.8044	0.8471
0.05	8578	10.6745	1.1153
0.07	8233	10.5559	1.3238
0.1	7718	10.3613	1.5989

Vidljivo je da se broj pogodaka smanjuje kako raste razina značajnosti. Također, ako usporedimo s identičnom tablicom kod prve metode, uočavamo da je kod svake razine značajnosti, osim 0.1, više pogodaka kod druge metode. Primijećujemo i da je srednja vrijednost pogodaka uvijek manja od 11.

Sada želimo provjeriti kako kod ove metode standardna devijacija šuma utječe na točnost pogodaka.

standardna devijacija šuma	broj pogodaka 11	srednja vrijednost pogodaka	standardna devijacija pogodaka
0.05	8578	10.6745	1.1153
0.075	8578	10.6745	1.1153
0.1	8578	10.6745	1.1153
0.25	8578	10.6745	1.1153
0.5	8208	10.7183	1.1434

Uočavamo da su sve vrijednosti jednake, osim posljednje, koja je manja, ali ne mnogo. Možemo zaključiti da standardna devijacija ne utječe na točnost podataka.

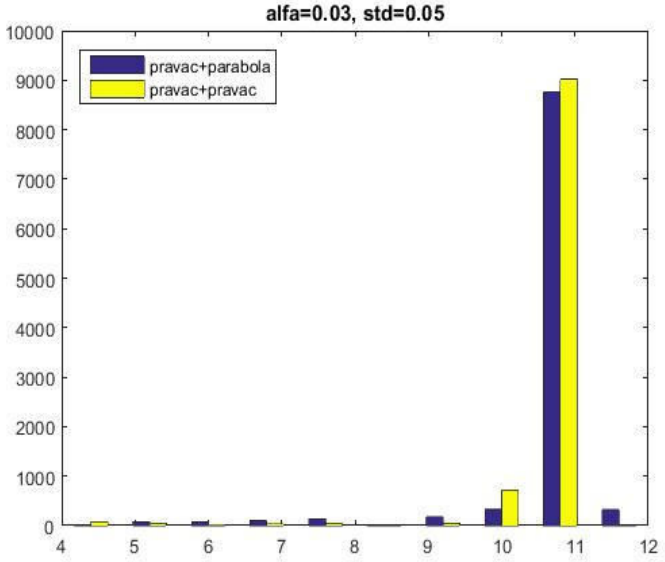
Kao i kod prve metode, i kod ove ćemo gledati utječe li broj točaka na točnost metode. Jednako kao kod prve metode prvo smanjimo, a zatim povećamo razmak među točkama.

razina značajnosti	broj pogodaka 21	srednja vrijednost pogodaka	standardna devijacija pogodaka
0.03	7689	19.9069	3.0119
0.05	6994	19.315	3.792
0.07	6426	18.7489	4.3677

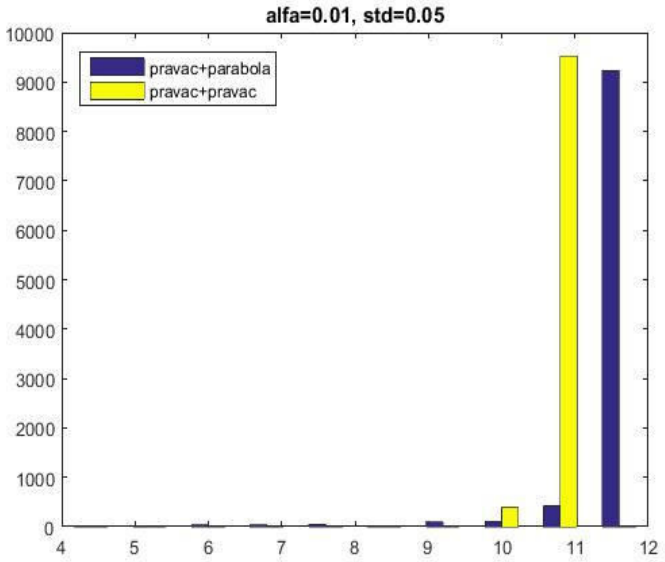
razina značajnosti	broj pogodaka 8	srednja vrijednost pogodaka	standardna devijacija pogodaka
0.03	9709	7.8999	0.4644
0.05	9011	7.8340	0.6167
0.07	8685	7.7708	0.7271

U oba slučaja broj pogodaka opada kako raste razina značajnosti baš kao i s originalnim podacima. Kada je razmak među točkama manji, imamo mnogo manje točnih pogodaka, a kada povećamo razmak među točkama, broj pogodaka se poveća u odnosu na originalne podatke. Možemo zaključiti da broj točaka značajno utječe na točnost metode.

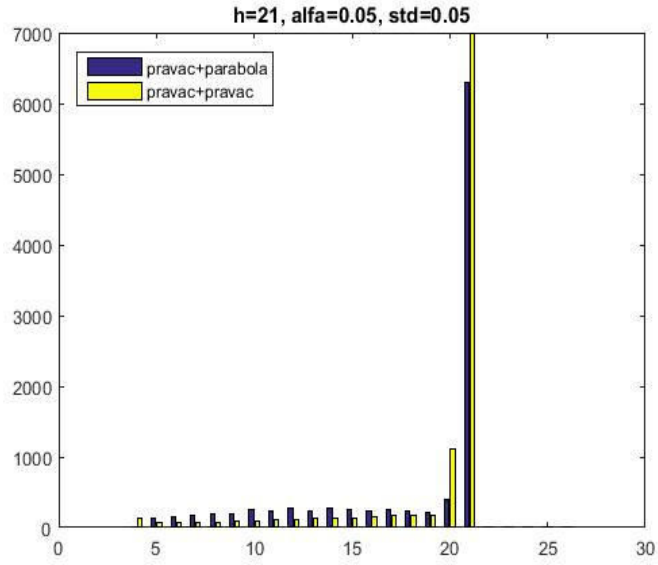
Pogledajmo kako izgledaju histogrami pogodaka.

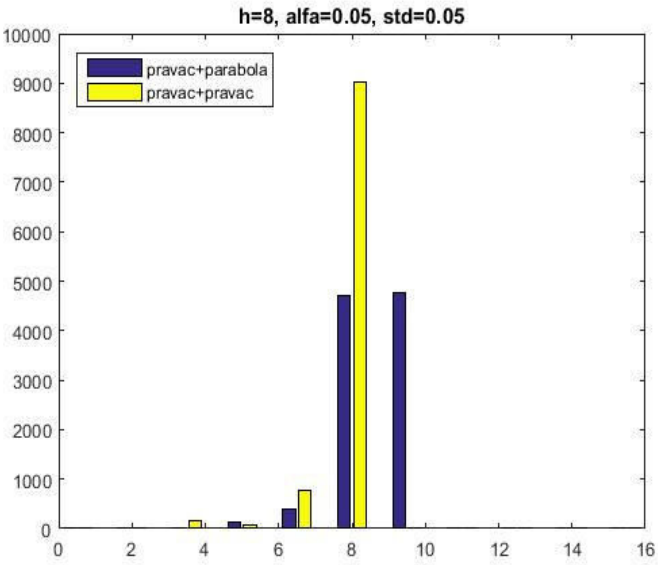


Ovo je slučaj u kojem imamo 15 točaka, nivo značajnosti je 3%, a standardna devijacija šuma iznosi 0.05. Iz histograma je vidljivo da je jedanaestica pogođena u većini slučajeva. Također je vidljivo da je više puta pogođena u drugoj metodi. Kod druge metode sve pogrešne točke su manje od 11, dok je kod prve metode nekoliko puta pogođena dvanaestica. Pogledajmo i histogram pogodaka za nivo značajnosti 0.01% budući da je za taj nivo prva metoda imala iznimno malo pogodaka.



Uočavamo da kod prve metode imamo najviše pogodaka dvanaestice, dakle metoda je pogriješila za jednu točku.
Pogledajmo također distribuciju pogodaka kad povećamo i smanjimo broj točaka.





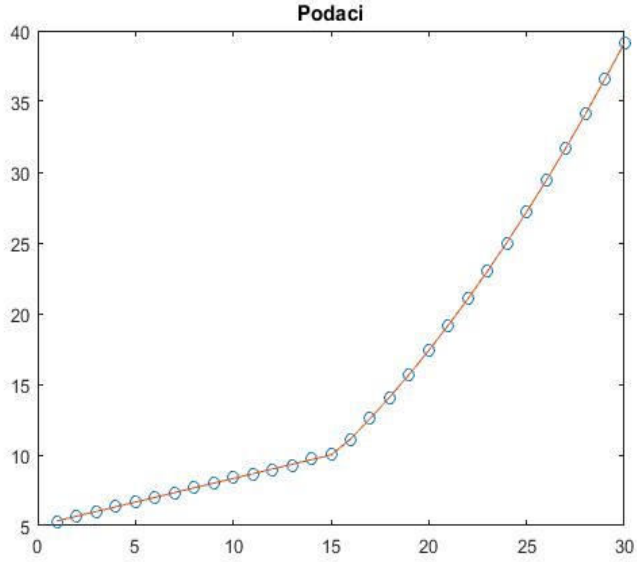
I na ovim grafovima uočavamo da je druga metoda bolja. Također je vidljivo da druga metoda nikad ne pogodi točku veću od ispravne, dok je kod prve to često slučaj.

Možemo zaključiti da je druga metoda bolja od prve. Kod svih razina značajnosti druga metoda daje više pogodaka. Promjena standardne devijacije šuma (za razumne vrijednosti) ne utječe na točnost metode. Broj točaka utječe, ali u mnogo manjoj mjeri, nego kod prve metode, čak štoviše, ukoliko smanjimo broj točaka, dobijemo bolji rezultat.

Da bismo se uvjerili u ispravnost ovog zaključka, pogledajmo i neke druge podatke. Uzmimo sada ovakav pravac: $y = \frac{1}{3}x + 5$ te neka prvih 15 točaka prati taj pravac. Sljedećih 15 točaka neka prati parabolu $y = \frac{1}{23}x^2$. Podacima također dodajmo normalni šum.

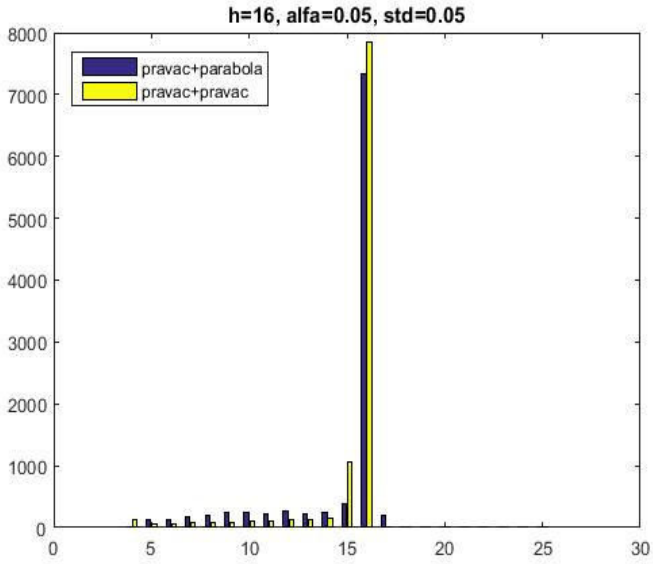
```
    a=1/3;
    b=5;
    x1=1:15;
    x2=16:30;
    x=[x1, x2];
    y1=a*x1+b;
    y2=1/23*x2.^2;
    y=[y1 y2];
    n=length(x);
e=0.05*randn(1,n);
y=y+e;
figure(1)
plot(x,y,'o');
hold on
plot(x,y-e);
```

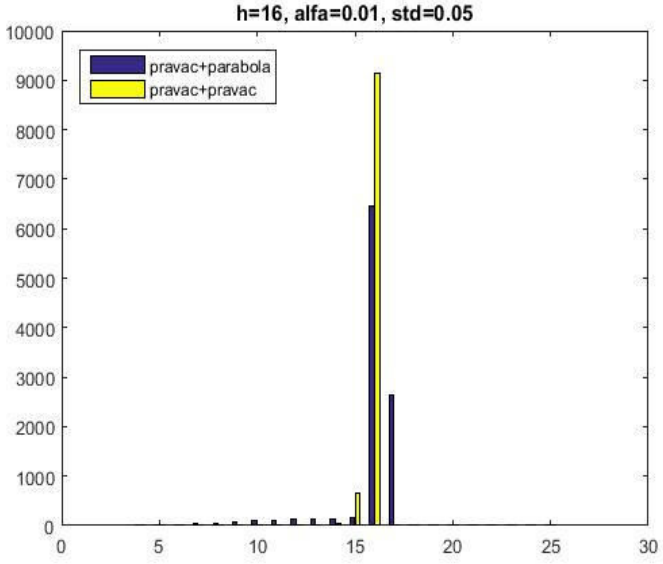
Graf podataka izgleda ovako:



razina značajnosti	broj pogodaka 16 u prvoj metodi	broj pogodaka 16 u drugoj metodi
0.01	6447	9154
0.05	7345	7840
0.1	6370	6703

Ovako izgledaju histogrami za razine značajnosti 5% i 1%:





Ponovo imamo istu stvar. Bolja je druga metoda. Kako raste nivo značajnosti, imamo manje pogodaka i u jednoj i u drugoj metodi (izuzetak je opet nivo značajnosti 1% kod prve metode). Histogrami također izgledaju slično. Prva metoda pogađa i točke veće od one koja se traži, dok druga pogađa samo manje.
Potpuni kod U MATLAB-u za prvu metodu (za standardnu devijaciju šuma 0.05 i razinu značajnosti 5%):

```
rng(543);
a=2;
b=10;
x1=1:10;
x2=11:15;
x=[x1, x2];
y1=a*x1+b;
y2=0.2*x2.^2+b;
y=[y1 y2];
n=length(x);
h=[];
for j=1:10000
    e=0.05*randn(1,n);
    % normalni sum sa standardnom devijacijom 0.05
    y=y+e;
    for i=5:n-1
```

```
koef1=polyfit(x(1:i),y(1:i),1);
y1=polyval(koef1,x(1:i));
koef2=polyfit(x(1:i),y(1:i),2);
y2=polyval(koef2,x(1:i));

%sume kvadrata reziduala
rez1=sum((y(1:i)-y1).^2);
rez2=sum((y(1:i)-y2).^2);

df1=i-2;
df2=i-3;
F=((rez1-rez2)/(df1-df2))/(rez2/df2);
pv=1-fcdf(F,df1,df2);

nasla=0;
if pv<=0.05  % ovo je za razinu znacajnosti 5%

   %ponavljamo postupak jer ignoriramo slucajeve
   % u kojima metoda nade neku tocku, ali tocke
   % nakon nje su i dalje na pravcu

    i=i+1;
    koef1=polyfit(x(1:i),y(1:i),1);
    y1=polyval(koef1,x(1:i));
    koef2=polyfit(x(1:i),y(1:i),2);
    y2=polyval(koef2,x(1:i));
    %sume kvadrata reziduala
    rez1=sum((y(1:i)-y1).^2);
    rez2=sum((y(1:i)-y2).^2);
    df1=i-2;
    df2=i-3;
    F=((rez1-rez2)/(df1-df2))/(rez2/df2);
    pv=1-fcdf(F,df1,df2);
    if pv<=0.05
        nasla=i-1;
        break;
    end
   i=i-1;
```

```
        end
    end
    h=[h nasla];
    y=y-e;
end
```

Kod u MATLAB-u za drugu metodu s istim podacima:

```
rng(543);
a=2;
b=10;
x1=1:10;
x2=11:15;
x=[x1, x2];
y1=a*x1+b;
y2=0.2*x2.^2+b;
y=[y1 y2];
n=length(x);
h2=[];
for j=1:10000
    e=0.05*randn(1,n);
    y=y+e;
    for i=3:n-1
    % prvo fitam pravcem na i tocaka
    koef1=polyfit(x(1:i),y(1:i),1);
    y1=polyval(koef1,x(1:i));
    % zatim fitam pravcem na i+1 tocki
    koef2=polyfit(x(1:i+1),y(1:i+1),1);
    y2=polyval(koef2,x(1:i+1));

    %sume kvadrata reziduala
    rez1=sum((y(1:i)-y1).^2);
    rez2=sum((y(1:i+1)-y2).^2);

    df1=i-2;  % broj podataka je i umanjeno
                % za 2 procijenjena parametra
    df2=i-1;  % broj podataka je i+1 umanjeno
                % za 2 procijenjena parametra
    F=((rez2-rez1)/(df2-df1))/(rez1/df1);
    pv=1-fcdf(F,df1,df2);
```

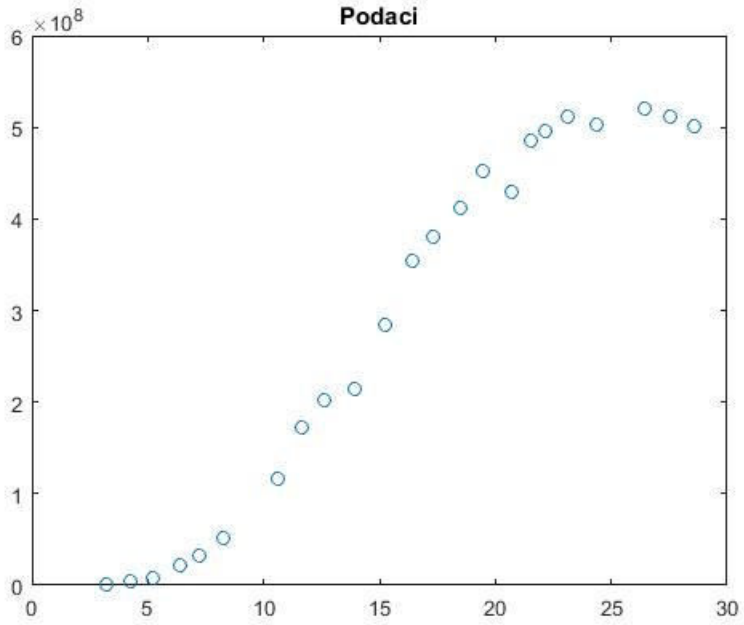
```
    if pv<=0.05  % nivo znacajnosti 5%
        i=i+1;
        koef1=polyfit(x(1:i),y(1:i),1);
        y1=polyval(koef1,x(1:i));
        koef2=polyfit(x(1:i+1),y(1:i+1),1);
        y2=polyval(koef2,x(1:i+1));

        %sume kvadrata reziduala
        rez1=sum((y(1:i)-y1).^2);
        rez2=sum((y(1:i+1)-y2).^2);

        df1=i-2;
        df2=i-1;
        F=((rez2-rez1)/(df2-df1))/(rez1/df1);
        pv=1-fcdf(F,df1,df2);

        if pv<=0.05
            nasla=i;
            break;
        end
        i=i-1;
    end
    end
    h2=[h2 nasla];
    y=y-e;
end
```

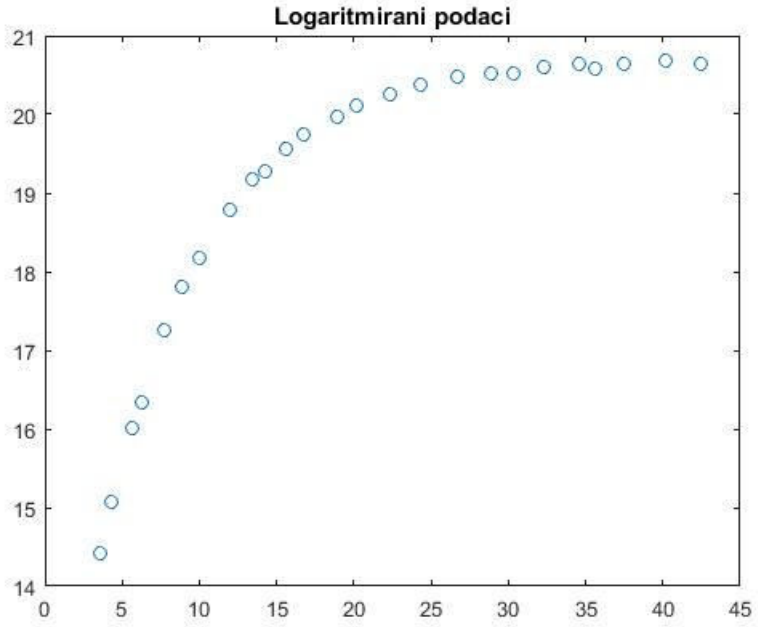
3.2 Realni podaci

Na stvarnim podacima vrši se usporedba ovih dviju metoda. Imamo 15 datoteka koje sadrže dva stupca, u prvom stupcu nalazi se vrijeme, a u drugom volumen koji se s vremenom povećava. Želimo naći fazu u kojoj volumen raste eksponencijalno. U tu svrhu podatke prvo trebamo logaritmirati, zatim primjenjujemo gornje metode. Metode trebaju dati točke u kojima rast prestaje biti ekponencijalan. Učitavanje i prikaz podataka:

```
pod1=importdata('in-01.data');
t=pod1(:,1);
v=pod1(:,2);
logv=log(v);
```

```
figure(1)
plot(t,v,'o');
title('Podaci');
figure(2)
plot(t,logv,'o');
title('Logaritmirani podaci');
```





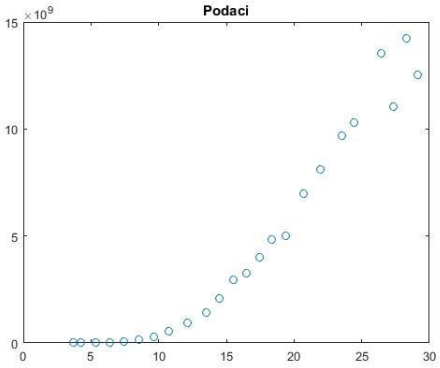
Rješenja koja dobijemo:

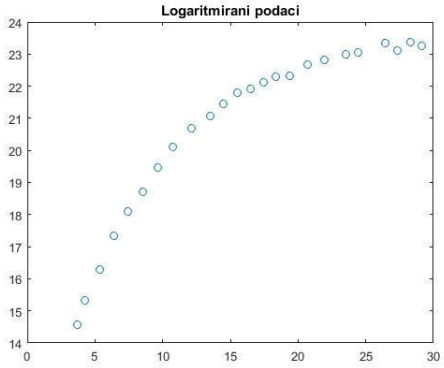
```
alfa =

    0.0100    0.0300    0.0500    0.0700    0.1000

tocka1 =

     7     6     6     6     6

tocka2 =

    15     7     7     6     6
```

Vektor alfa sadrži razine značajnosti testa, vektor tocka1 sadrži tocke koje je pronašla prva metoda redom za razine značajnosti iz vektora alfa i vektor tocka2 sadrži tocke koje je našla druga metoda za iste razine značajnosti. To su tocke u kojima prestaje eksponencijalan rast.

Prva metoda daje šestu točku za sve razine značajnosti, osim za 1%, dok druga metoda za 1% daje petnaestu točku, za 3% i 5% daje sedmu točku, a za 7% i 10% daje također šestu točku. Budući da su simulacije pokazale da je druga metoda bolja i da je najtočnija kod razine značajnosti 1% možemo zaključiti da eksponencijalan rast prestaje u petnaestoj

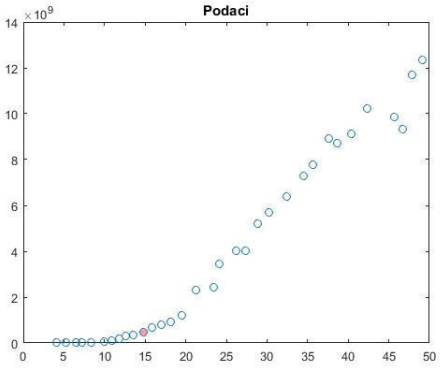
točki.
Slijede ostali podaci. Rješenja koja dobijemo redom su za razine značajnosti 1%,3%,5%,7% i 10%. Drugi set podataka:

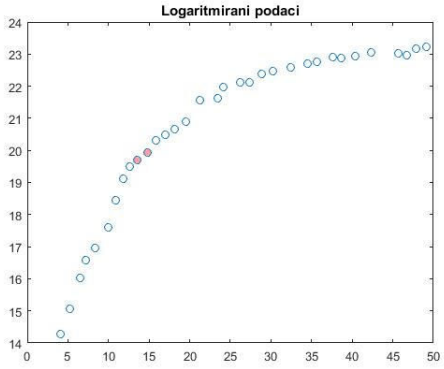





```
tocka1 =

     6     6     6     6     6

tocka2 =

    21     9     9     9     9
```

Treći set podataka:

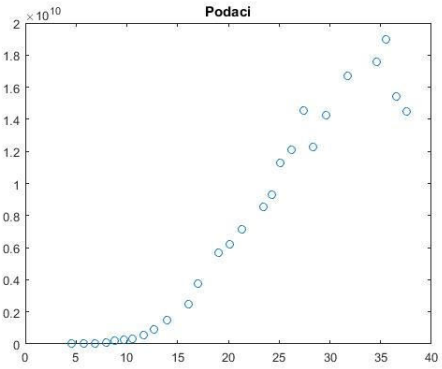


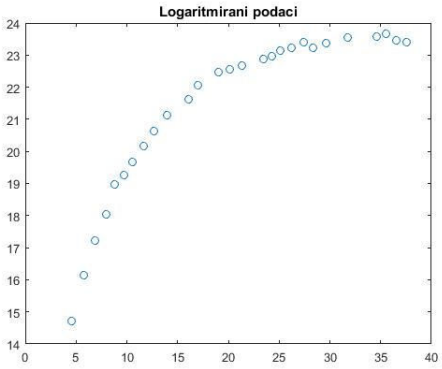



```
tocka1 =

    11    11    11    11    11
```

```
tocka2 =

    13    11    11    10     4
```

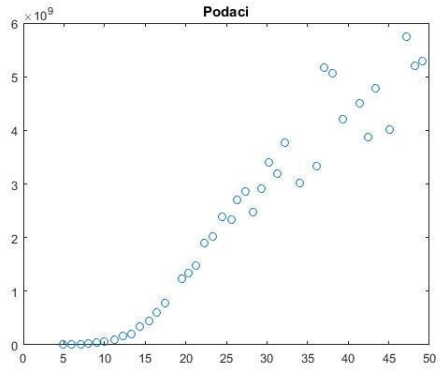
Četvrti set podataka:

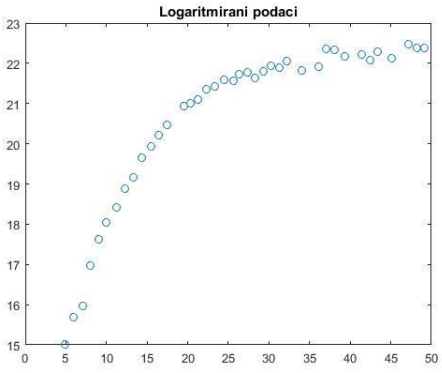





```
tocka1 =

     7     7     6     6     6

tocka2 =

     0    23    13    10    10
```

Peti set podataka:

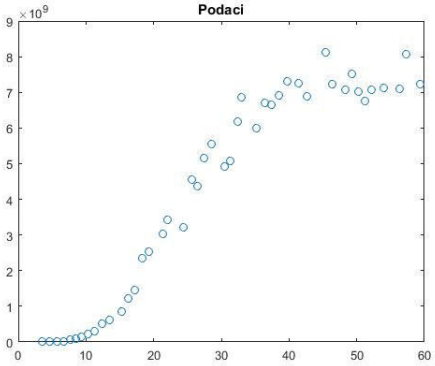


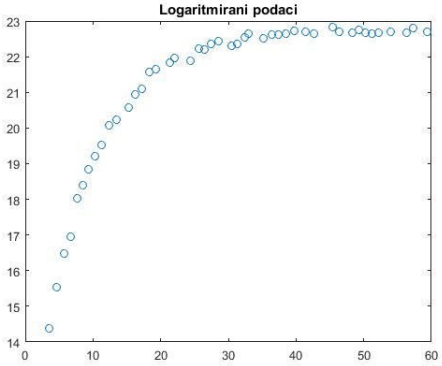



```
tocka1 =
```

```
    10     9     9     9     9

tocka2 =

    14    14    14    13    12
```

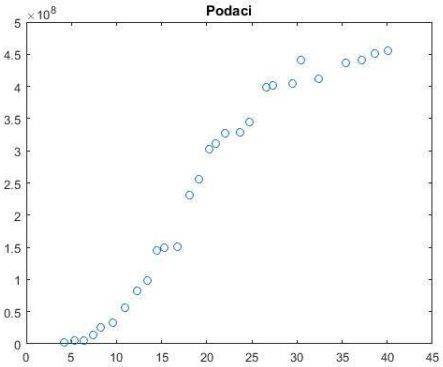
Šesti set podataka:

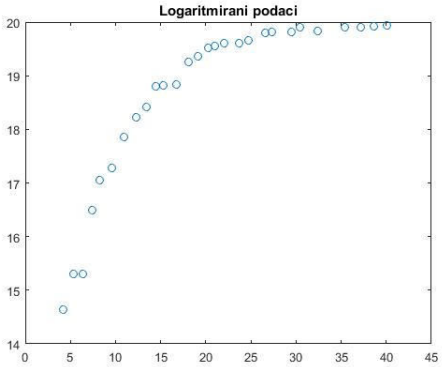





```
tocka1 =

     8     8     7     7     7

tocka2 =

    36    11     8     8     8
```

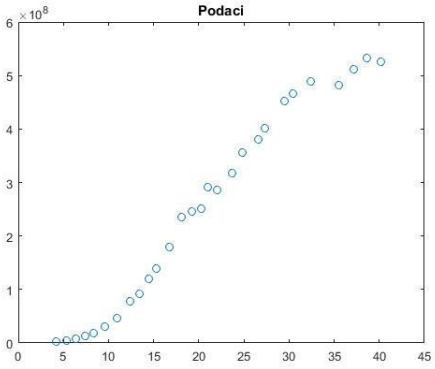
Sedmi set podataka:

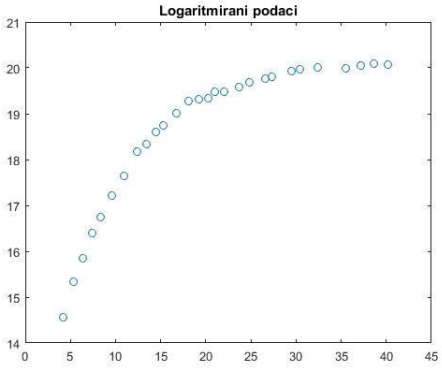




```
tocka1 =

    10     9     9     9     9

tocka2 =

    24    18    18    17    17
```

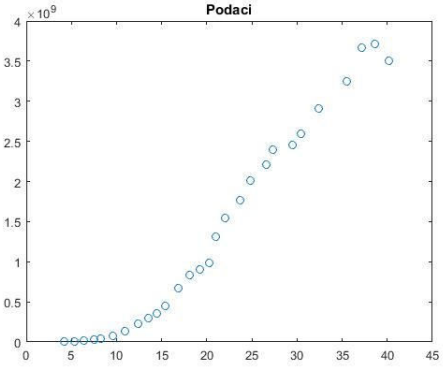
Osmi set podataka:

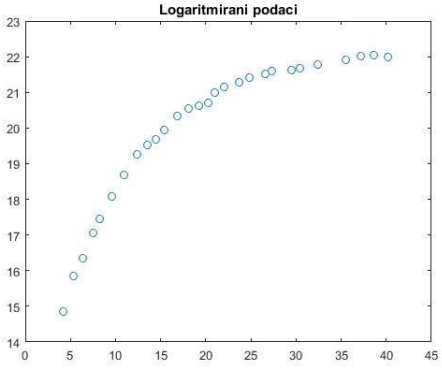





```
tocka1 =

     6     6     5     5     5

tocka2 =

    24    14    12    12    11
```

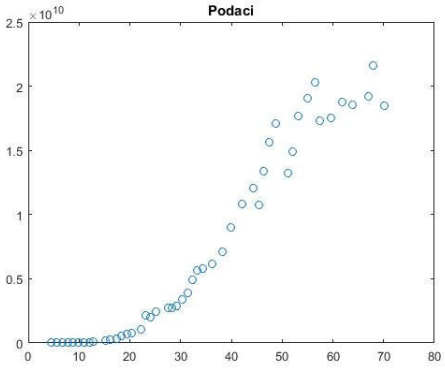
Deveti set podataka:

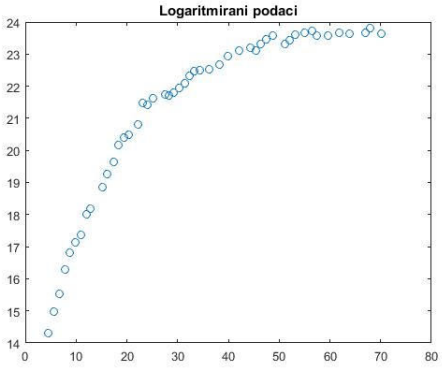




```
tocka1 =

     7     7     6     6     6

tocka2 =

    22     9     9     9     9
```

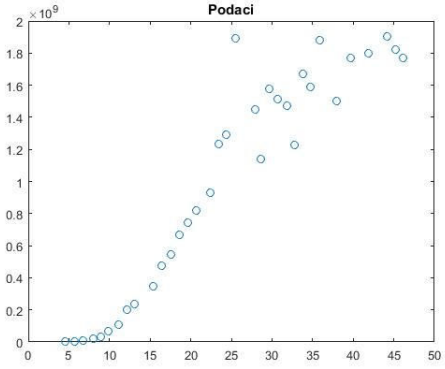
Deseti set podataka:

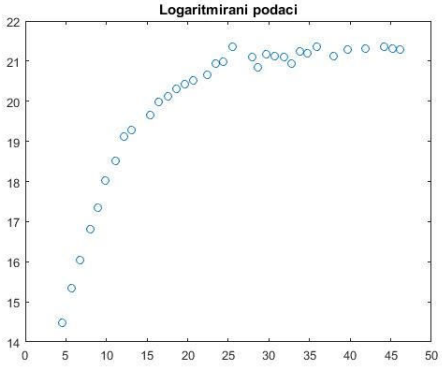





```
tocka1 =

     9     7     7     7     7

tocka2 =

     6     6     6     6     6
```

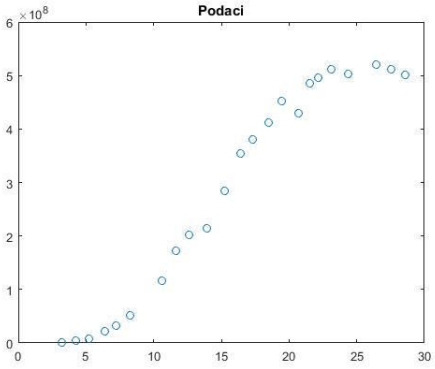
Jedanaesti set podataka:

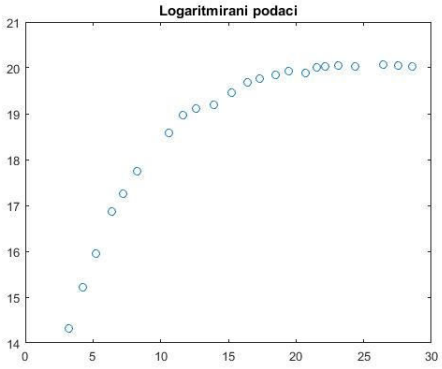




```
tocka1 =

     7     7     5     5     5

tocka2 =

     9     9     9     9     9
```

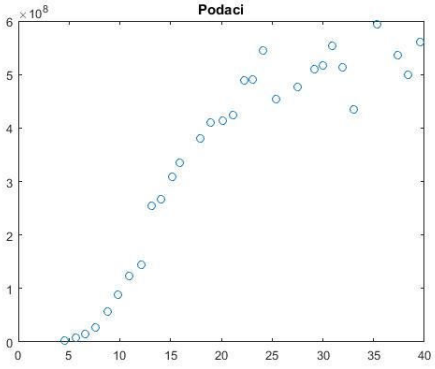
Dvanaesti set podataka:

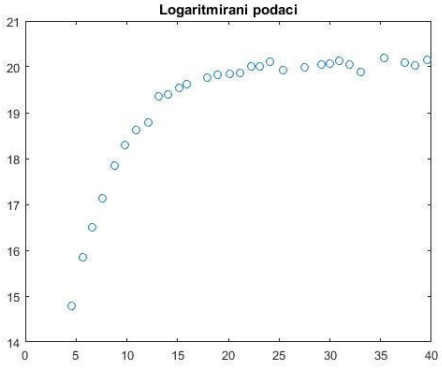





```
tocka1 =

     6     6     6     6     6

tocka2 =

     0    21    20     5     5
```

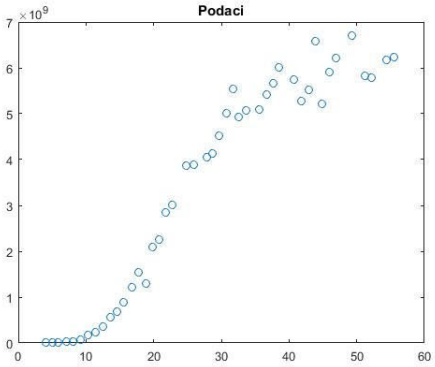
Trinaesti set podataka:

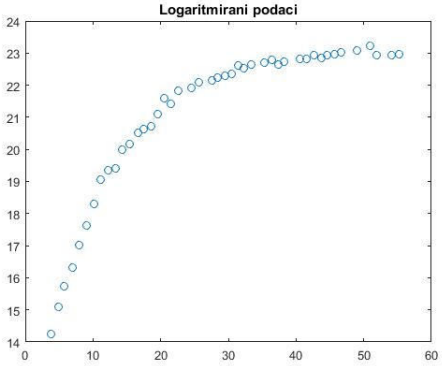




```
tocka1 =

     6     6     6     6     5

tocka2 =

     0    13     7     7     7
```

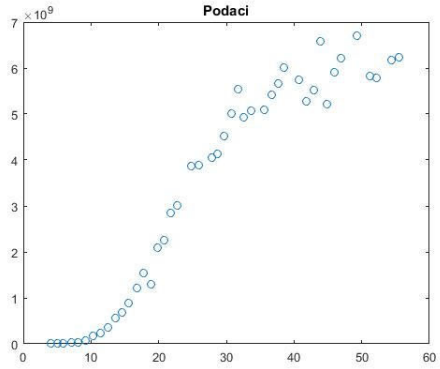
Četrnaesti set podataka:

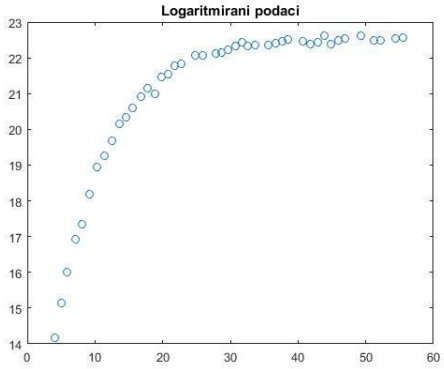





```
tocka1 =

    10    10     9     9     9

tocka2 =

     9     9     9     9     9
```

Petnaesti set podataka:






```
tocka1 =

     8     8     5     5     5

tocka2 =

    20     8     8     4     4
```

Na temelju grafičkih prikaza podataka i dobivenih točaka može se zaključiti da druga metoda daje dobre rezultate kod razine značajnosti 1%, dok prva metoda u većini slučajeva pronalazi premalene vrijednosti.

Bibliografija

[1] *Distribution of sum of squares error for linear regression*, http://stats.stackexchange.com/questions/24921/distribution-of-sum-of-squares-error-for-linear-regression.

[2] *The Standard Normal and The Chi-Square*, https://onlinecourses.science.psu.edu/stat414/node/154.

[3] Kristijan Kilassa Kvaternik, *Primijenjena statistika, bilješke s predavanja (prof. dr. sc. Miljenko Huzak) akademske godine 2014./2015.*, 2015.

Sažetak

Ukratko, ovaj rad sastoji se od dva dijela: teorijskog i eksperimentalnog. Prva dva poglavlja su teorijska i u njima je objašnjena linearna regresija i pojmovi vezani uz nju. Pokazano je kako se metodom najmanjih kvadrata dolazi do parametara linearne regresije. Od velike su važnosti reziduali i njihova distribucija pa je temeljito dokazana tvrdnja da suma kvadrata reziduala ima χ^2 distribuciju. Treće poglavlje odnosi se na eksperimentalni dio u kojem se među rastućim podacima, prvo simuliranim, a potom realnim, želi odrediti eksponencijalna faza rasta. Podaci su logaritmirani i na njima se provode dvije metode koje koriste linearnu regresiju. U prvoj metodi uspoređujemo linearni i kvadratični model, a u drugoj uspoređujemo dva linearna modela. Točnost metoda provjeravamo u ovisnosti o različitim razinama značajnosti testova, različitim standardnim devijacijama šuma te različitom duljinom podataka. Simulacije su pokazale da je druga metoda, sa dva linearna modela, u većini slučajeva bolja.

Summary

In this paper, we have two parts: theoretical and experimental part. First two chapters are theoretical and they explain linear regression and terms related to it. It is shown how one finds parameters of linear regression using the least square method. Reziduals and their distribution are very important, therefore is the claim that sum of squared residuals follows χ^2 distribution proofed in detail. Third chapter reffers to experimental part in which one between data, first simulated, then real, wants to find exponential phase of growth. Logarithm of the data is being put through two methods, which use linear regression. In first method linear and quadratic model are being compared, and in the second one, two linear models are being compared. The accuracy of methods are being checked depending on the different levels of significance, different standard deviation of the noise of the data and different length of the data. Simulations have shown that the second method is in most of the cases better.

Životopis

Osobni podaci:

- Ime i prezime: Antonija Novački
- Datum i mjesto rođenja: 15.03.1992., Zagreb
- [redacted]

[redacted]

Apsolventica sam diplomskog studija Matematička statistika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. Prethodne razine školovanja:

- PMF-Preddiplomski studij matematike (2011.-2014.)
- Srednja škola Krapina - prirodoslovno matematička gimnazija (2007.-2011.)
- Osnovna škola Side Košutić Radoboj (1999.-2007.)

Tijekom školovanja stekla sam neke računalne vještine:

- rad u MATLAB-U, R-u, SAS-u
- osnove programiranja u programskom jeziku C
- rad s bazama podataka
- izrada web stranica

Govorim engleski jezik. Kao studentica sam radila u Jadranskom osiguranju - prodaja polica auto osiguranja, kaska, putnih osiguranja. Također sam radila u Studilici dajući instrukcije iz matematike.