DOI: 10.15255/KUI.2016.004
KUI-7/2017
Stručni rad
Prispjelo 4. veljače 2016.
Prihvaćeno 30. svibnja 2016.

Simulacija jednostavne linearne regresije

S. Džalto[*] i *I. Gusić*

Fakultet kemijskog inženjerstva i tehnologije, Marulićev trg 19, 10 000 Zagreb

Sažetak

Cilj rada je računalno simuliranje uvjeta koji su pretpostavka modela jednostavne linearne regresije i računalno potvrđivanje temeljnih formula. Za tu svrhu opisan je model jednostavne linearne regresije i diskutirane su matematičke osnove na kojima se model zasniva. Navedene su formule za procjenu koeficijenata regresijskog pravca i intervala povjerenja (relacije (1.4), (3.5.1), (3.5.2), (3.8) i (3.9)) te za procjenu varijance pogreške (relacije (3.4), (3.6), (3.11) i (3.12)). U programskom paketu Matlab generirano je 10 000 nizova po sedam podataka kojima je simulirano ponavljanje 10 000 pokusa pri istim uvjetima u inženjerskoj praksi. Svaki od tih nizova simulira mjerenje vrijednosti zavisne veličine za sedam fiksiranih vrijednosti nezavisne veličine u okolnostima pri kojima su zadovoljene pretpostavke modela linearne regresije. Za slučajno odabrani niz od sedam podataka primijećeno je znatno odstupanje procjena parametara modela od stvarnih vrijednosti (tablica 1), što upućuje na to da relativno mali broj mjerenja u praksi može voditi do nepouzdanih procjena. S druge strane, pokazalo se da je aritmetička sredina pripadnih 10 000 izračunatih parametara gotovo identična stvarnoj vrijednosti parametara. Drugim riječima, potvrđeno je da su procjene dobivene uzastopnim ponavljanjem mjerenja pri istim uvjetima u prosjeku točne. Simulacija sa svih 10 000 generiranih nizova potvrdila je i druge spomenute formule. Tako računalna simulacija može poslužiti za bolje razumijevanje linearne regresije i uspješno zamijeniti zahtjevno dokazivanje matematičkih činjenica na kojima se ona zasniva.

Ključne riječi

Jednostavna linearna regresija, normalna razdioba, hi-kvadrat razdioba, Studentova razdioba, interval pouzdanosti, Matlab

1. Uvod

Linearna veza

$$y = ax + b \tag{1.1}$$

najjednostavnija je funkcijska veza među veličinama x, y. Pri analizi eksperimentalnih podataka nepoznati parametri a, b općenito se ne mogu točno odrediti, već samo procijeniti. To znači da se za svaku od vrijednosti $x_1, x_2 \ldots x_n$ veličine x pokusom odrede pripadne vrijednosti $x_1, x_2 \ldots x_n$ veličine y. Točke

$$(x_1, y_1),\ (x_2, y_2)\ \ldots\ (x_n, y_n) \tag{1.2}$$

u pravilu neće biti na jednom pravcu. Ako se izabere ciljna funkcija, onda ima smisla govoriti o optimalnoj procjeni parametara iz (1.1), na osnovi podataka (1.2), tj. o izboru pravca koji je najbolje prilagođen tim točkama. Prema metodi najmanjih kvadrata, koja je uobičajena u takvim okolnostima, ciljna funkcija ima oblik:[1]

$$F(a, b) = \sum_{i=1}^{n} (y_i - a x_i - b)^2 \tag{1.3}$$

Ovdje se nepoznati parametri a, b interpretiraju kao varijable o kojima ovisi ciljna funkcija F koja ima značenje zbroja kvadrata odstupanja eksperimentalnih vrijednosti y_i od pripadnih teorijskih vrijednosti $a x_i + b_i$. Traže se vrijednosti $\hat{a}, \hat{b}$ od a, b tako da F ima najmanju vrijednost. Rješenje je jedinstveno:[1]

$$\hat{a} = \sum_{i=1}^{n} (x_i - \bar{x}) y_i / SS_x, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}, \tag{1.4}$$

gdje je

$$\bar{x} = (x_1 + x_2 + \ldots + x_n)/n \tag{1.5}$$

aritmetička sredina podataka $x_1, x_2 \ldots x_n$ i slično za $\bar{y}$, a SS_x uobičajena oznaka za zbroj kvadrata odstupanja od aritmetičke sredine

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{1.6}$$

Nepoznata veza (1.1) procjenjuje se približnom linearnom vezom

$$y = \hat{a}x + \hat{b}. \tag{1.7}$$

Pravac zadan jednadžbom (1.7) zove se regresijski pravac. Druga jednakost u (1.4) govori da je točka $(\bar{x}, \bar{y})$ na regresijskom pravcu. Često takav linearni model nije pogodan. Jedan je od razloga to što mjerenje veličine y za pojedine zadane vrijednosti veličine x može biti neprecizno; pri istim vrijednostima veličine x, ponavljanjem pokusa kojim se određuje vrijednost od y, mogu se dobivati osjetno različiti rezultati. Drugi, za linearnu regresiju i važniji razlog, jest taj što veza među veličinama x, y ne mora biti funkcijska, tj. jednoj vrijednosti veličine x odgovara više različitih

[*] Autor za dopisivanje: Stjepan Džalto
e-pošta: sdzalto@fkit.hr

vrijednosti veličine y. Na primjer, ako je x visina, a y masa (ljudske populacije), onda su veličine x, y povezane, ali ta veza nije funkcijska. Ipak, može se govoriti o funkcijskoj vezi između x i pripadne prosječne vrijednosti veličine y. Linearnost takve veze može se modelirati modelom jednostavne linearne regresije.

Za razliku od determinističkog modela (1.1), taj je stohastički. I ovdje se polazi od skupa podataka (1.2), ali se za svaku vrijednost x_i veličine y, veličina y interpretira slučajnom (oznaka Y_i), tako da je

$$Y_i = a x_i + b + E_i, \quad i = 1, 2 \ldots n, \tag{1.8}$$

gdje je E_i slučajna varijabla koja intuitivno opisuje pogrešku pri mjerenju (odnosno odstupanje od prosjeka). Drugim riječima, vrijednost y_i interpretira se slučajnom vrijednošću veličine Y_i. U (R1)–(R3) naveden je popis pretpostavaka za model linearne regresije (1.8) koje će se upotrebljavati u ovome članku.

(R1.) Veličina x je obična (neslučajna) kojoj se vrijednosti x_i mogu egzaktno odrediti (makar su u praksi i one podložne pogreškama pri mjerenju).

(R2.) $E_i, i = 1, 2 \ldots n$ međusobno su nezavisne slučajne varijable.

(R3.) Sve su E_i normalno distribuirane s očekivanjem 0 i istom varijancom σ^2; kraće $E_i \sim N(0, \sigma^2)$. Odatle slijedi $Y_i \sim N(ax + b, \sigma^2)$, $i = 1, 2 \ldots n$.

Za razliku od determinističkog modela (1.1) gdje su vrijednosti x_i različite, u stohastičkom modelu (1.8) među podatcima x_1, $x_2 \ldots x_n$ može biti međusobno jednakih (što odgovara tome da je za neku konkretnu vrijednost od x pokus ponovljen više puta).

2. Matematičke osnove modela linearne regresije

Uobičajeno je da se obične varijable označavaju malim slovima, poput x, y, a slučajne velikima, poput X, Y. Također, vrijednosti slučajne varijable označavaju se malim slovima. Na primjer, za n slučajnih vrijednosti slučajne varijable Y uobičajena je oznaka $y_1, y_2 \ldots y_n$. Može se zamišljati da su se ponavljanjem nekog pokusa n puta i mjerenja u njemu pri istim uvjetima, dobili rezultati $y_1, y_2 \ldots y_n$. Da bi se podatci objektivno analizirali, interpretiraju se (nezavisnim) vrijednostima neke slučajne varijable. Pravilnost pojavljivanja slučajnih vrijednosti slučajne varijable opisuje se razdiobom te slučajne varijable, koja je jednoznačno određena funkcijom gustoće vjerojatnosti.[1–2] Temeljne numeričke karakteristike slučajne varijable X jesu očekivanje $E(X)$ i varijanca $V(X)$. Intuitivno, $E(X)$ je prosječna vrijednost koju X postiže, dok je varijanca mjera odstupanja slučajnih vrijednosti od očekivanja. U teoriji vjerojatnosti dokazuje se da za sve slučajne varijable X_1, X_2 vrijedi

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2), \tag{2.1}$$

a ako su X_1, X_2 nezavisne, onda vrijedi i

$$V(aX_1 + bX_2) = a^2 V(X_1) + b^2 V(X_2). \tag{2.2}$$

Za linearnu regresiju posebnu važnost imaju normalna razdioba, hi-kvadrat razdioba i Studentova razdioba. Normalna razdioba jednoznačno je određena svojim očekivanjem i varijancom.* Činjenicu da slučajna varijabla X ima normalnu razdiobu s očekivanjem μ i varijancom σ^2 (parametri normalne razdiobe) označava se kao $X \sim N(\mu, \sigma^2)$. Pokazuje se da vrijedi:

$$\begin{gathered}\text{ako je } X_1 \sim N(\mu_1, \sigma_1{}^2),\ X_2 \sim N(\mu_2, \sigma_2{}^2),\\ \text{onda je } aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1{}^2 + b^2\sigma_2{}^2).\end{gathered} \tag{2.3}$$

To znači da linearna kombinacija nezavisnih normalno distribuiranih slučajnih varijabla ima normalnu razdiobu.** Ako je očekivanje jednako 0, a varijanca 1 normalna razdioba naziva se jediničnom normalnom razdiobom i označava kao N(0,1). Za n slučajnih vrijednosti $x_1, x_2 \ldots x_n$ slučajne varijable X koja ima jediničnu normalnu razdiobu, zbroj kvadrata $x_1{}^2 + x_2{}^2 + \ldots + x_n{}^2$ može se interpretirati kao slučajna vrijednost slučajne varijable $X_1{}^2 + X_2{}^2 + \ldots + X_n{}^2$, gdje su $X_1, X_2 \ldots X_n$ nezavisne slučajne varijable koje imaju jediničnu normalnu razdiobu. Pokazuje se da ta slučajna varijabla ima hi-kvadrat razdiobu s n stupnjeva slobode (oznaka $\chi^2(n)$). Kraće,

$$\begin{gathered}\text{ako su } X_i, i = 1, 2 \ldots n, \text{ nezavisne varijable i } X_i \sim N(0,1),\\ \text{onda je } X_1{}^2 + X_2{}^2 + \ldots + X_n{}^2 \sim \chi^2(n).\end{gathered} \tag{2.4}$$

$$\chi^2(n) \text{ ima očekivanje } n \text{ i varijancu } 2n. \tag{2.5}$$

Slično se interpretira činjenica da slučajna varijabla $\dfrac{X_0}{\sqrt{(X_1{}^2 + X_2{}^2 + \ldots + X_n{}^2)/n}}$, gdje su $X_0, X_1 \ldots X_n$ nezavisne slučajne varijable s jediničnom normalnom razdiobom, ima Studentovu t-razdiobu s n stupnjeva slobode (oznaka $t(n)$).*** Kraće,

$$\begin{gathered}\text{ako su } X_i, i = 1, 2 \ldots n, \text{ nezavisne varijable i } X_i \sim N(0,1),\\ \text{onda je } \frac{X_0}{\sqrt{(X_1{}^2 + X_2{}^2 + \ldots + X_n{}^2)/n}} \sim t(n).\end{gathered} \tag{2.6}$$

[3]

3. Razvoj modela

Kao i u determinističkom modelu (1.1) i u stohastičkom modelu (1.8), nepoznati parametri a, b procjenjuju se s pomoću metode najmanjih kvadrata i dobiju se procjene $\hat{a}, \hat{b}$ kao u (1.4). Da bi se ispitala pouzdanost takve procjene, te se vrijednosti interpretiraju slučajnima. Intuitivno, zamišlja se da se svaki pokus kojim se pri vrijednošću $x = x_i$ određivala pripadna vrijednost y_i ponavlja više puta (za svaki konkretan i). Pri svakom ponavljanju svakog od tih pokusa, dobije se novi niz vrijednosti $y_i, i = 1, 2 \ldots n$, a

* To općenito ne vrijedi za svaku slučajnu varijablu.

** (2.1), (2.2) i (2.3) vrijedi za više slučajnih varijabla, a ne samo za dvije.

*** Graf funkcije gustoće t-razdiobe sliči onoj od jedinične razdiobe (vidi sl. 3),ali je spljoštеniji, jer je varijanca veća od 1, točnije, jednaka je $n/(n-2)$

time i nova procjena $\hat{a},\hat{b}$ nepoznatih parametara a,b. Da bi se taj postupak kontrolirao, uvodi se slučajna varijabla $\hat{A}$ kojoj je $\hat{a}$ slučajna vrijednost, a slično i slučajna varijabla $\hat{B}$. Za to je dovoljno u (1.4) svaku konkretnu vrijednost y_i zamijeniti slučajnom varijablom Y_i (pritom se $\bar{y}$ zamjenjuje s $\bar{Y} = \frac{Y_1 + Y_2 + \ldots + Y_n}{n}$). Dobije se

$$\hat{A} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{SS_x},\ \hat{B} = \bar{Y} - \hat{A}\bar{x}, \tag{3.1}$$

gdje je SS_x definiran u (1.6). Pokazuje se da $\hat{A},\hat{B}$ imaju normalnu razdiobu. Preciznije:

$$\hat{A} \sim N\left(a, \frac{\sigma^2}{SS_x}\right),\ \hat{B} \sim N\left(b, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)\right)^* \tag{3.2}$$

Primjena teorije vjerojatnosti u linearnoj regresiji ilustrirat će se dokazom činjenice $E(\hat{A}) = a$ koja je sastavni dio prve formule u (3.2). Dokaz se zasniva na (2.1)–(2.3) i jednostavnoj relaciji $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, koja se izravno provjeri, a iz koje se dobije i $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})x_i - 0 = \sum_{i=1}^{n}(x_i - \bar{x})x_i - \sum_{i=1}^{n}(x_i - \bar{x})\bar{x} = SS_x$. $\hat{A}$ je normalna jer je linearna kombinacija nezavisnih normalnih slučajnih varijabla. Kako je, prema (R3), $Y_i \sim N(ax_i + b, \sigma^2), i = 1, 2 \ldots n$,

primjenjujući gornje relacije i nezavisnost slučajnih varijabla Y_i, dobije se

$$E(\hat{A}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(Y_i)}{SS_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(ax_i + b)}{SS_x} =$$
$$= \frac{a\sum_{i=1}^{n}(x_i - \bar{x})\,x_i + b\sum_{i=1}^{n}(x_i - \bar{x})}{SS_x} = a$$

Formule (3.2) mogu se napisati kao

$$\frac{\hat{A} - a}{\sigma}\sqrt{SS_x} \sim N(0,1),\ \frac{\hat{B} - b}{\sigma\sqrt{1/n + \bar{x}^2/SS_x}} \sim N(0,1) \tag{3.3}$$

Procjena parametara *a*, *b*

Formule (3.3) mogle bi se iskoristiti za određivanje intervala povjerenja uz zadanu pouzdanost (intervala pouzdanosti) uz uvjet da je standardna devijacija σ poznata. Naime iz njih se iščitava da se brojevi $\frac{\hat{a} - a}{\sigma}\sqrt{SS_x}$ i $\frac{\hat{b} - b}{\sigma\sqrt{1/n + \bar{x}^2/SS_x}}$ mogu interpretirati kao slučajne vrijednosti slučajnih varijabli s jediničnom normalnom razdiobom. Kako je, uz parametre a,b, i parametar σ nepoznat, i njega treba procijeniti na osnovi podataka (1.2). Zamjenom σ procjenom $s = \sqrt{s^2}$, gdje je

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2}{n-2} \tag{3.4}$$

dobiju se brojevi $\frac{\hat{a} - a}{s}\sqrt{SS_x}$ i $\frac{\hat{b} - b}{s\sqrt{1/n + \bar{x}^2/SS_x}}$ koji se mogu interpretirati slučajnim vrijednostima slučajnih varijabla koje imaju Studentovu razdiobu s $n-2$ stupnjeva slobode. To omogućava procjenu nepoznatih parametara a,b uz unaprijed zadanu pouzdanost. To se zasniva na činjenici da se zamjenom u (3.3) parametra σ slučajnom varijablom S i jedinične normalne razdiobe Studentovom t-razdiobom $t(n-2)$ s $n-2$ stupnjeva slobode, dobiju korektne formule

$$\frac{\hat{A} - a}{S}\sqrt{SS_x} \sim t(n-2) \tag{3.5.1}$$

$$\frac{\hat{B} - b}{S\sqrt{1/n + \bar{x}^2/SS_x}} \sim t(n-2) \tag{3.5.2}$$

$$S^2 = \frac{SS_E}{n-2},\ \text{gdje je } SS_E = \sum_{i=1}^{n}(Y_i - \hat{A}x_i - \hat{B})^2. \tag{3.6}$$

Formule (3.5.1) i (3.5.2) bit će potvrđene simulacijom, a prije toga komentirat će se njihova primjena. Za broj p između 0 i 1, uobičajena je oznaka $t_p(k)$ za broj sa svojstvom

$$P(t(k) > t_p(k)) = p. \tag{3.7}$$

Intuitivno, za velik skup slučajnih vrijednosti Studentove razdiobe $t(k)$, njih oko $100(1-2p)$ nalazi se u intervalu $\langle -t_p(k), t_p(k)\rangle$. Činjenicu da se slučajna vrijednost slučajne varijable iz (3.5.1) nalazi u intervalu $\langle -t_p(n-2), t_p(n-2)\rangle$, može se zapisati kao

$$\hat{a} - t_p(n-2)\frac{s}{\sqrt{SS_x}} < a < \hat{a} + t_p(n-2)\frac{s}{\sqrt{SS_x}} \tag{3.8}$$

što daje **interval pouzdanosti** za nepoznati parametar a uz pouzdanost (**koeficijent pouzdanosti**, **razinu pouzdanosti**) $1-2p$.**

---

* Varijance u (3.2), koje opisuju rasap parametara $\hat{a}$, $\hat{b}$ oko a, odnosno b, ne ovise o rezultatima mjerenja već samo o varijanci σ^2 pogrešaka E_i i o vrijednostima $x_1, x_2 \ldots x_n$.
Formule (3.2) opisuju razdiobu koeficijenata $\hat{a},\hat{b}$ kao slučajnih vrijednosti normalnih slučajnih varijabla $\hat{A},\hat{B}$. Smisao je da je $\hat{a}$ "u prosjeku jednako" a, dok je $\hat{b}$ u prosjeku jednako b, tj. ponavljanjem mjerenja dovoljno mnogo puta i računanjem koeficijenata $\hat{a},\hat{b}$ svaki put, nepoznati koeficijent smjera a bio bi praktično jednak aritmetičkoj sredini svih takvih koeficijenata $\hat{a}$, a b aritmetičkoj sredini svih $\hat{b}$. Preciznije, $E(\hat{A}) = a$ i $E(\hat{B}) = b$. Drugim riječima, $\hat{A}$ odnosno $\hat{B}$ jesu nepristrani procjenitelji nepoznatih parametara a odnosno b,[1] dok se slučajne vrijednosti $\hat{a},\hat{b}$ nepristranih procjenitelja zovu nepristranim procjenama od a,b. Formulama (3.2) opisan je i rasap tih vrijednosti oko a, odnosno b.

** Iako se ovo često interpretira kako je vjerojatnost da se a nalazi u rečenom intervalu jednaka $1-2p$, to nije opravdano. Naprosto a jest ili nije u tom intervalu i nema smisla govoriti o vjerojatnosti. Intuitivno, ako su veličine x i Y zaista povezane linearnom vezom (1.8), uz relacije (R1)–(R3), onda bi ponavljanjem niza pokusa pri istim vrijednostima x_i i dobivanjem pripadnih vrijednosti y_i, otprilike u $100(1-2p)$ % izvođenja bila ispunjezina relacija (3.8). To će biti i pokazano simulacijom. Treba uočiti da bi se, pri svakom ponavljanju niza pokusa, vrijednosti $\hat{a}$ u (3.8) u pravilu mijenjale, pa tako i interval pouzdanosti (koji je, pri fiksiranom izboru vrijednosti x_i veličine x i razine pouzdanosti $1-2p$, jednoznačno određen rezultatima mjerenja y_i). U literaturi se koeficijent pouzdanosti često označava kao α ili γ. U slučaju oznake α, u formuli za interval pouzdanosti $t_p(n-2)$ treba zamijeniti s $t_{1-\alpha/2}(n-2)$. Interval pouzdanosti definira se u općenitijim okolnostima, a ovo je samo poseban slučaj.[1–3]

Slično se iz (3.5.2) dobije interval pouzdanosti za nepoznati parametar b uz razinu pouzdanosti $1-2p$:

$$\hat{b} - t_p(n-2)s\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{SS_x}} < b < \hat{b} + t_p(n-2)s\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{SS_x}} \quad (3.9)$$

Komentar relacija (3.4) i (3.5)

Iz (1.8) i (R1)–(R3) Uvoda slijedi da su $y_i - ax_i - b$, za $i = 1, 2 \ldots n$ slučajne vrijednosti normalne razdiobe s očekivanjem 0 i varijancom σ^2. Zato je

$$\frac{\sum_{i=1}^{n}(y_i - ax_i - b)^2}{n-1} \quad (3.10)$$

nepristrana procjena nepoznate varijance σ^2. Problem te procjene (zbog koje je ona neupotrebljiva) u tome je što ona ovisi o parametrima a, b koji su nepoznati, pa i njih treba procjenjivati. Ako se ti parametri u (3.9) zamijene njihovim procjenama $\hat{a}, \hat{b}$, u nazivniku umjesto $n-1$ stavi $n-2$, dobije se nepristrana procjena varijance σ^2 (formula (3.4)).* Naime, s^2 iz (3.4) slučajna je vrijednost slučajne varijable $S^2 = SS_E/(n-2)$, gdje je $SS_E = \sum_{i=1}^{n}(Y_i - \hat{A}x_i - \hat{B})^2$. Problem se svodi na opis razdiobe slučajne varijable SS_E. Intuitivno, pitanje je kako su raspoređeni brojevi oblika $\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2$ pri uzastopnom ponavljanju pokusa. Pokazuje se da su oni u prosjeku $(n-2)\sigma^2$, tj. da je

$$E(S^2) = \sigma^2, \quad (3.11)$$

odnosno da je s^2 iz (3.4) nepristrana procjena od σ^2. To proizlazi iz općenitijeg rezultata:

slučajna varijabla SS_E/σ^2 ima hi-kvadrat razdiobu $\chi^2(n-2)$ s $n-2$ stupnjeva slobode.[3] (3.12).

Formule (3.8), (3.9), (3.11) i (3.12) bit će podržane računalnom simulacijom.

4. Simulacija linearne regresije u Matlabu

Simulacija je provedena u programskom paketu Matlab 2010b. Program za simulaciju nalazi se u prilogu. Unutar programa moguće je mijenjati parametre (broj podataka u jednom nizu, broj nizova, parametre pravca, razine pouzdanosti). U modelu linearne regresije (1.8) izabrano je:

$$a = 2,\ b = 1;\ n = 7;\ x_i = i;\\ E_i \sim N(0, 2^2);\ i = 1, 2 \ldots 7 \quad (4.1)$$

To znači da su y_i iz (1.2) određeni formulom $y_i = 2i + 1 + e_i$, gdje su e_i generirani kao slučajne vrijednosti normalne razdiobe

$$N(0, 2^2);\ i = 1, 2 \ldots 7 \quad (4.2)$$

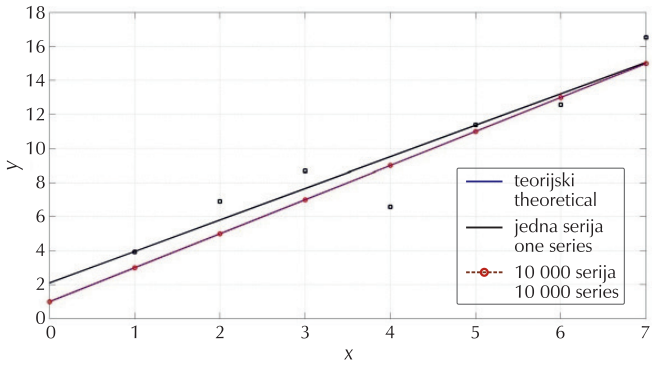
To je ponovljeno 10 000 puta i tako dobiveno 10 000 nizova od po 7 uređenih parova (x_i, y_i) prema pravilu iz (4.2). Naravno, u praksi se ne radi 10 000 mjerenja, ali je ovdje to učinjeno da se prikažu statističke zakonitosti bez dokazivanja. Za svaki niz od 7 uređenih parova, priloženi program računa parametre $\hat{a}$ i $\hat{b}$ prema formulama (1.4) i s^2 prema formuli (3.4). Na primjer, za prvi se niz dobije: $\hat{a} = 1{,}7255$, $\hat{b} = 2{,}8189$, $s^2 = 2{,}6370$. Vidi se znatno odstupanje tih procjena od stvarnih vrijednosti $a = 2$, $b = 1$, $\sigma^2 = 4$. Tipičnu okolnost u praksi (mali broj mjerenja) upravo ilustrira svaki od tih nizova od 7 točaka (x_i, y_i). Tablicom 1 predočen je jedan takav niz (svi su podatci na kraju zaokruženi na dvije decimale).

Tablica 1 – Slučajno odabran jedan od 10 000 nizova.
Table 1 – One randomly selected series out of 10 000

x_i	y_i	$\hat{a}x_i + \hat{b}$
1	3,93	3,96
2	6,91	5,81
3	8,68	7,66
4	6,58	9,51
5	11,38	11,36
6	12,57	13,21
7	16,53	15,07

Za taj niz dobiveno je $\hat{a} = 1{,}85$; $\hat{b} = 2{,}11$; $s^2 = 2.68$ (zaokruženo na dvije decimale).

Da bi se potvrdilo kako se formulama (1.4) u prosjeku dobiju prave vrijednosti parametara a, b, računaju se prosječne vrijednosti svih 10 000 izračunatih koeficijenata $\hat{a}$ (dobije se 1,9993) i prosječne vrijednosti svih izračunatih $\hat{b}$ (dobije se 0,9998), što su vrlo dobre procjene $a \approx 1{,}9993$ i $b \approx 0{,}9998$. To potvrđuje jednakosti $E(\hat{A}) = a$ i $E(\hat{B}) = b$ (koje su sadržane u formulama (3.2)). Na slici 1 predočen je regresijski pravac za niz iz tablice 1 i uspoređen sa zadanim pravcem $y = 2x + 1$.



Slika 1 – Regresijski pravac
Fig. 1 – Regression line

* U statistici za to ima heurističko objašnjenje zasnovano na stupnjevima slobode (analogno stupnjevima slobode u mehanici). Prijelaz od (3.10) do (3.4) nastaje zamjenom parametara a, b njihovim procjenama. Iako su to naizgled dvije procjene, u stvari je samo jedna. Naime, procjena $\hat{b}$ ovisi o procjeni $\hat{a}$ (druga formula u (1.4)), tako da, kad se procijeni a, automatski se procijenilo i b. Prema rečenoj heuristici, nazivnik treba umanjiti za broj međuprocjena, dakle za 1. Odatle $n-2$ u formuli (3.4).

Potvrđivanje formule $E(S^2) = \sigma^2$ iz (3.11)

Provodi se računanjem prosjeka svih izračunatih s^2 iz (3.4), čime se dobiva procjena varijance σ^2. Da se ilustrira kako se povećavanjem broja nizova dobiva bolja procjena, u tablici 2 predočeni su rezultati za 100, 1000 i svih 10 000 nizova. Također, uključen je podatak i samo za prvi niz generiranih vrijednosti. Vidi se da je najbolja procjena stvarne varijance koja je u simulaciji izabrana da bude 4, procjena $\sigma^2 \approx 3{,}99648$ dobivena iz svih 10 000 nizova.

Tablica 2 – Procjene varijance σ^2 formulom (3.4)
Table 2 – Evaluation of variance σ^2 with Equation (3.4)

Broj nizova, m Number of series, m	$\frac{\sum_{i=1}^{m} s_i^2}{m}$
1	2,63702
100	3,49633
1 000	3,92498
10 000	**3,99648**

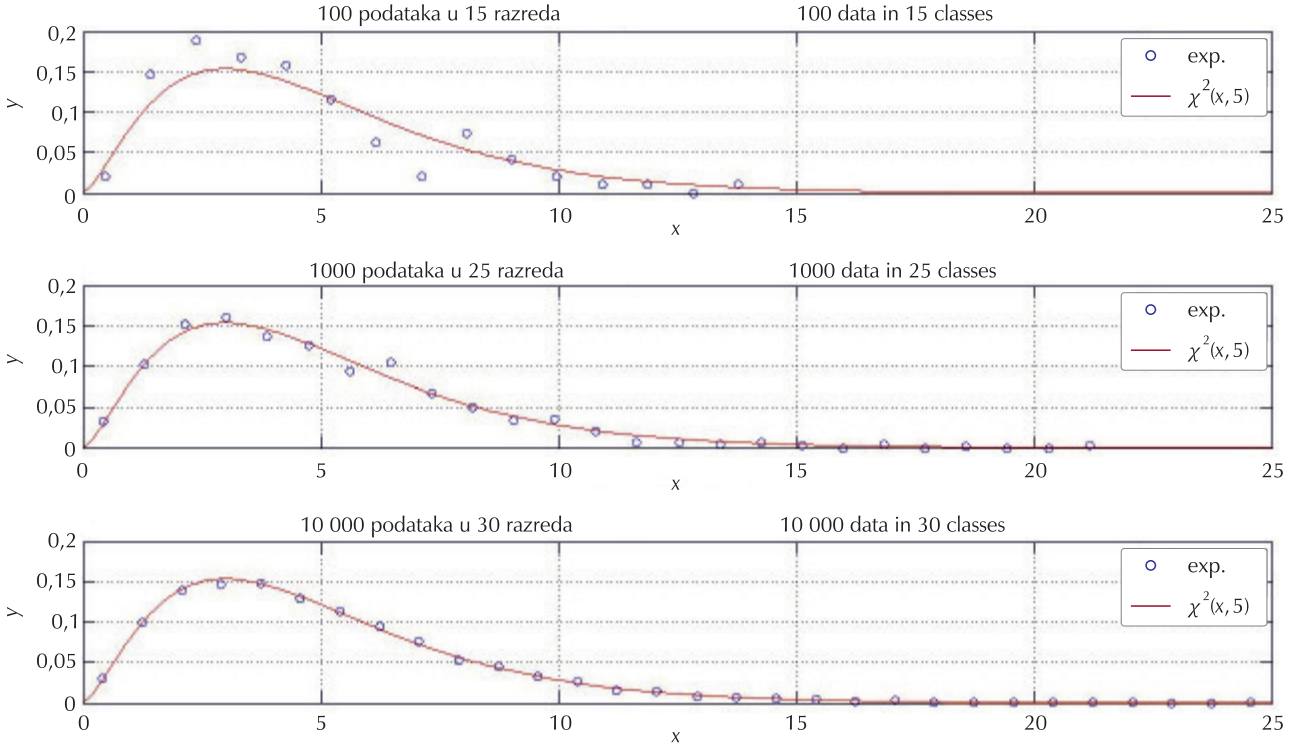
Simulacija tvrdnje (3.12): slučajna varijabla SS_E/σ^2 ima hi-kvadrat razdiobu $\chi^2(n-2)$ s $n-2$ stupnjeva slobode

Ilustriranje slaganja nekog skupa od m podataka s teorijskom razdiobom koja ima funkciju gustoće vjerojatnosti g, u ovom se članku provodi sljedećim koracima:

1. Odabere se interval unutar kojega se nalaze svi podatci (eventualno nakon odbacivanja nekoliko ekstremnih podataka).
2. Gornji se interval podijeli u L razreda jednake širine h.
3. Neka je f_j frekvencija j-tog razreda (broj podataka unutar tog razreda), $\frac{f_j}{m}$ relativna frekvencija j-tog razreda i $\overline{f}_j = \frac{f_j}{mh}$ relativna gustoća podataka u j-tom razredu.
4. U istom koordinatnom sustavu nacrta se graf funkcije gustoće vjerojatnosti g, i ucrtaju točke $(c_j, \overline{f}_j)$, za sve $j = 1, 2 \ldots L$, gdje je c_j sredina j-tog razreda. Gleda se koliko su te točke prilagođene teorijskom grafu.

U gornjem primjeru slučajne vrijednosti od SS_E računaju se tako da se za svaki niz od 7 generiranih vrijednosti y_i u (3.6) umjesto Y_i stavi y_i, umjesto A stavi se pripadni $\hat{a}$, a umjesto $\hat{B}$ pripadni $\hat{b}$ (dobiveni broj mjeri koliko izračunate vrijednosti $\hat{a}, \hat{b}$ odstupaju od zadanih vrijednosti a, b). Svaka od tih vrijednosti dijeli se sa σ^2 (što je u ovom primjeru 4). Te su vrijednosti pozitivne pa se u sljedećem koraku područje brojeva od 0 do najveće od tih vrijednosti podijeli u razrede jednake širine i nastavi se kako je opisano u koracima 1–4.

Da se prikaže ovisnost o ukupnom broju podataka, radi se, kao i prije, najprije sa 100 slučajnih vrijednosti (dobivenih od 100 generiranih nizova), potom s 1000 i na kraju s 10 000. Rezultati su predočeni na slici 2 i u tablici 3. Vidi se da je teorijska krivulja $\chi^2(5)$ dobro prilagođena eksperimentalnim podatcima, tim bolje što je broj podataka veći. Srednja vrijednost svih 10 000 podataka iznosi 4,9956 (što je u skladu s činjenicom da je očekivanje $\chi^2(5)$ razdiobe jednako 5), dok je varijanca svih podataka 10,0138 (što je, opet, u skladu s time da je varijanca $\chi^2(5)$ razdiobe jednaka 10).



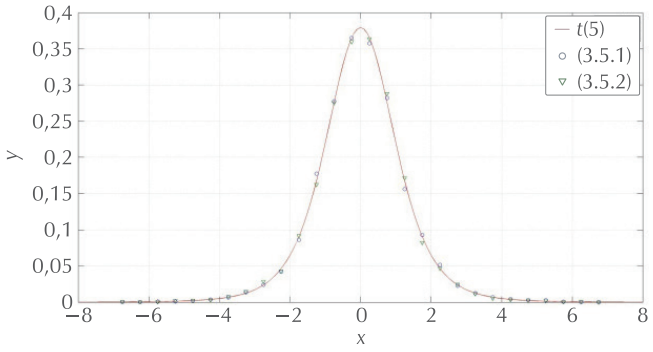
Slika 2 – Simulacija hi²-razdiobe
Fig. 2 – Simulation of chi² distribution

Tablica 3 – Podatci za sliku 2
Table 3 – Data for Fig. 2

100 podataka 100 data			1000 podataka 1000 data			10 000 podataka 10 000 data		
sredina razreda class mean	relativna gustoća relative density	frekvencija frequency	sredina razreda class mean	relativna gustoća relative density	frekvencija frequency	sredina razreda class mean	relativna gustoća relative density	frekvencija frequency
0,4751	0,0210	2	0,4318	0,0324	28	0,4162	0,0308	256
1,4252	0,1473	14	1,2954	0,1019	88	1,2487	0,0998	831
2,3754	0,1894	18	2,1590	0,1517	131	2,0811	0,1394	1160
3,3256	0,1684	16	3,0226	0,1610	139	2,9135	0,1473	1226
4,2757	0,1579	15	3,8862	0,1378	119	3,7460	0,1486	1237
5,2259	0,1158	11	4,7498	0,1262	109	4,5784	0,1299	1081
6,1760	0,0631	6	5,6134	0,0938	81	5,4108	0,1141	950
7,1262	0,0210	2	6,4771	0,1054	91	6,2433	0,0950	791
8,0764	0,0737	7	7,3407	0,0672	58	7,0757	0,0769	640
9,0265	0,0421	4	8,2043	0,0498	43	7,9081	0,0532	443
9,9767	0,0210	2	9,0679	0,0347	30	8,7406	0,0449	374
10,9268	0,0105	1	9,9315	0,0359	31	9,5730	0,0330	275
11,8770	0,0105	1	10,7951	0,0208	18	10,4054	0,0264	220
12,8272	0,0000	0	11,6587	0,0069	6	11,2379	0,0153	127
13,7773	0,0105	1	12,5223	0,0069	6	12,0703	0,0139	116
			13,3859	0,0046	4	12,9027	0,0079	66
			14,2495	0,0069	6	13,7352	0,0062	52
			15,1131	0,0035	3	14,5676	0,0056	47
			15,9767	0,0000	0	15,4000	0,0038	32
			16,8403	0,0046	4	16,2325	0,0020	17
			17,7039	0,0000	0	17,0649	0,0025	21
			18,5676	0,0023	2	17,8974	0,0012	10
			19,4312	0,0000	0	18,7298	0,0012	10
			20,2948	0,0000	0	19,5622	0,0007	6
			21,1584	0,0035	3	20,3947	0,0006	5
						21,2271	0,0004	3
						22,0595	0,0004	3
						22,8920	0,0000	0
						23,7244	0,0000	0
						24,5568	0,0001	1

Simulacija formula (3.5.1) i (3.5.2)

Na temelju 10 000 podataka (dobivenih od 10 000 nizova po 7 podataka) potvrđeno je da je u obje formule riječ o Studentovoj $t(5)$-razdiobi (slika 3 i tablica 4). Gotovo su svi dobiveni podatci u simetričnom intervalu od -7 do 7. Zato su, radi jednostavnosti, iz računa izbačeni svi brojevi manji od -7 i veći od 7. Provjera se vrši naredbom "sum(stA)" za prvu formulu (izbačeno je 16 od 10 000 ekstremnih vrijednosti), a za drugu naredbom "sum(stB)" (izbačeno je 15 od 10 000 vrijednosti). Segment $[-7,7]$ podijeljen je na 14 razreda širine 0,5. Ostalo je napravljeno kao u prethodnoj simulaciji: plavi kružići odnose se na (3.5.1), a zeleni trokutići na (3.5.2). Crvenom crtom prikazana je teorijska krivulja, tj. graf funkcije gustoće vjerojatnosti Studentove $t(5)$-razdiobe. Vidi se da se odgovarajući kružići i trokutići gotovo poklapaju i da prate teorijsku krivulju, što je i trebalo pokazati. Aritmetička sredina za prvu skupinu podataka iznosi 0,007, a za drugu $-0{,}0101$, što se dobro slaže s činjenicom da je očekivanje Studentove razdiobe 0. Varijanca prvog skupa podataka iznosi 1,7442, a drugog 1,7189, što je opet u skladu sa Studentovom $t(5)$-razdiobom, kojoj je očekivanje $5/3$ (rezultati se poklapaju na jednu decimalu).



Slika 3 – Simulacija formula (3.5.1) i (3.5.2)
Fig. 3 – Simulation of Equations (3.5.1) and (3.5.2)

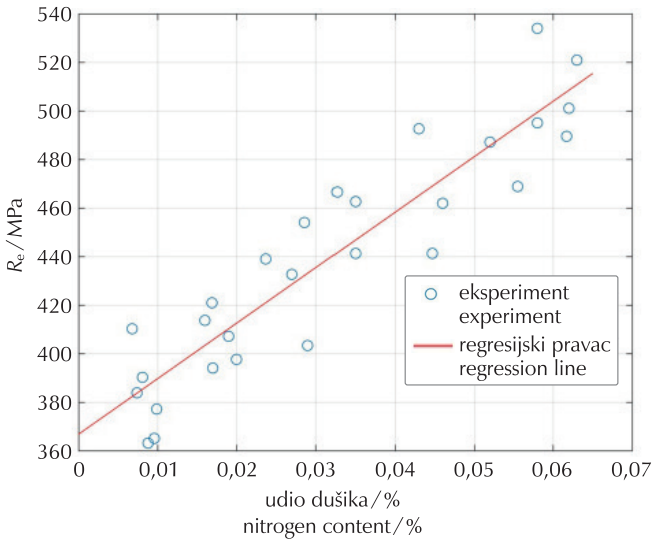
Tablica 4– Podatci za sliku 3
Table 4 – Data for Fig. 3

	Frekvencije Frequencies		Relativne gustoće Relative densities	
X	stA	stB	ystudentA	ystudentB
−6,75	2	4	0,0004	0,0008
−6,25	1	1	0,0002	0,0002
−5,75	5	2	0,001	0,0004
−5,25	5	11	0,001	0,0022
−4,75	9	12	0,0018	0,0024
−4,25	18	17	0,0036	0,0034
−3,75	33	38	0,0066	0,0076
−3,25	69	73	0,0138	0,0146
−2,75	121	143	0,0242	0,0286
−2,25	211	216	0,0422	0,0432
−1,75	432	459	0,0864	0,0918
−1,25	890	814	0,178	0,1628
−0,75	1388	1376	0,2776	0,2752
−0,25	1826	1799	0,3652	0,3598
0,25	1789	1819	0,3578	0,3638
0,75	1410	1441	0,282	0,2882
1,25	783	861	0,1566	0,1722
1,75	465	413	0,093	0,0826
2,25	258	234	0,0516	0,0468
2,75	114	123	0,0228	0,0246
3,25	62	58	0,0124	0,0116
3,75	35	27	0,007	0,0054
4,25	23	16	0,0046	0,0032
4,75	15	12	0,003	0,0024
5,25	14	9	0,0028	0,0018
5,75	4	2	0,0008	0,0004
6,25	1	5	0,0002	0,001
6,75	1	0	0,0002	0
Zbroj:	9984	9985		

Potvrda formula (3.8) i (3.9)

Ako je formula (3.8) točna, onda bi idealno u 9500 od 10 000 puta (95 %) parametar $a = 2$ trebao biti između $\hat{a} - t_{0,025}(5)\frac{s}{\sqrt{SS_x}}$ i $\hat{a} + t_{0,025}(5)\frac{s}{\sqrt{SS_x}}$, tj. $\hat{a} - 0{,}4848s < 2 < \hat{a} + 0{,}4845s$. Prebrojavanjem se dobije da se za te podatke to dogodilo u 9497 od 10 000 puta (94,97 %). Slično se za formulu (3.9) dobije da je $b = 1$ u 9487 od 10 000 puta (94,87 %) bio u intervalu $\langle \hat{b} - t_{0,025}(5)\, s\sqrt{1/7 + \bar{x}^2/SS_x}\,,\ \hat{b} + t_{0,025}(5)\, s\sqrt{1/7 + \bar{x}^2/SS_x}\,\rangle$, tj. da je bilo $\hat{b} - 2{,}1726s < 1 < \hat{b} + 2{,}1726s$. Treba uočiti da je rasap u drugoj formuli veći (koeficijent 2,1726, prema koeficijentu 0,4858) što je logično: uz malu promjenu nagiba pravca dolazi do relativno velike promjene odsječka na *y*-osi.

Primjer. Čvrstoća zavarenog titana ovisi o udjelu intersticijskog elementa. Jedna od karakteristika koje upućuju na čvrstoću jest granica razvlačenja, R_e. Ona se definira kao naprezanje pri vlačnom opterećenju pri kojem dolazi do znatnih plastičnih deformacija.[4] Neka je *x* udio dušika u postotcima i y granica razvlačenja u megapaskalima.



Slika 4 – Ovisnost granice razvlačenja o udjelu dušika
Fig. 4 – Dependency of yield strength versus nitrogen content

Tablica 5 i slika 4 upućuju na to da ima smisla razmatrati linearnu povezanost veličina *x* i *y*.[5–6] U deterministčkom modelu, optimalna linearna veza (1.7) na osnovi svih 28 rezultata mjerenja je $y = 2286{,}4x + 366{,}8$ (zaokruženo na jedno decimalno mjesto). Iako rezultati mjerenja sugeriraju da bi veličine mogle biti linearno povezane, to se može potvrditi (ili opovrgnuti) tek iz vrlo velikog broja mjerenja, ali ni tada s potpunom sigurnošću. To je u skladu s provedenom simulacijom. Analiza metodom jednostavne linearne regresije zasniva se na pretpostavci da se veličine zaista ponašaju prema (1.8), tj. da postoje parametri a, b i σ^2 takvi da bi ponavljanjem mjerenja za svaku odabranu vrijednost x_i od *x*, dobivene vrijednosti veličine *y* bile oblika $ax_i + b + e_i$, gdje bi "pogreške mjerenja" e_i bile slučajne vrijednosti normalne razdiobe $N(0, \sigma^2)$. Prave vrijednosti

(nepoznatih) parametara a, b i σ^2 ne mogu se točno odrediti, ali se mogu procijeniti na osnovi dostupnih rezultata mjerenja, uz određenu pouzdanost. Primjenjujući formulu (3.4), σ^2 se procjenjuje s pomoću s^2 što je 359,9 na jednu decimalu, odnosno $\sigma \approx 19{,}0$ (to je samo procjena čija bi se pouzdanost mogla ispitivati, ali u ovom članku to se ne radi). Parametri a i b procjenjuju se s pomoću (1.4) i dobije se $\hat{a} = 2286{,}4$ i $\hat{b} = 366{,}8$. Primjenjujući (3.8) i (3.9) uz razinu pouzdanosti 0,90, dobiju se intervali pouzdanosti $1963{,}2 < a < 2609{,}7$ i $354{,}8 < b < 378{,}8$ za a odnosno za b. Uz veću razinu pouzdanosti intervali su još širi. Tako se za razinu pouzdanosti 0,95 dobije $1896{,}9 < a < 2676{,}0$ i $352{,}3 < b < 381{,}2$.

Tablica 5 – Ovisnost granice razvlačenja o udjelu dušika
Table 5 – Dependency of yield strength versus nitrogen content

Udio N / % N content / %	R_e / Mpa	udio N / % N content / %	R_e / Mpa
0,0088	363,1	0,016	413,7
0,0169	420,9	0,027	432,6
0,02	397,6	0,0237	439,0
0,0286	454,0	0,035	462,6
0,035	441,3	0,0327	466,6
0,043	492,6	0,052	487,1
0,046	461,9	0,0555	468,8
0,058	534,0	0,058	495,0
0,0617	489,5	0,0447	441,3
0,063	520,9	0,029	403,3
0,062	501,0	0,0068	410,2
0,0099	377,1	0,017	394,0
0,0074	383,8	0,0081	390,2
0,019	407,1	0,0096	365,1

U praksi je često broj mjerenja n manji od 28, a u ovom članku simulacija se provodila za $n = 7$. Ako se iz tablice 5 uzme prvih 7 podataka i sve ponovi, dobiju se procjene $\sigma^2 \approx 411{,}4$ odnosno $\sigma \approx 20{,}3$, te $\hat{a} = 2812{,}5$ i $\hat{b} = 353{,}4$. Intervali pouzdanosti bitno se povećaju. Uz razinu pouzdanosti 0,90 dobije se $1609{,}0 < a < 4015{,}9$ i $316{,}0 < b < 390{,}8$, dok se za razinu pouzdanosti 0,95 dobije $1277{,}3 < a < 4347{,}6$ i $305{,}7 < b < 401{,}2$.

Uočava se osjetna ovisnost o dostupnim podatcima i izabranoj razini pouzdanosti, što zahtijeva oprez pri daljnjoj primjeni metode. Takva analiza odnosi se na dostupne podatke koji često ne moraju dobro ocrtavati stvarnu vezu. Isti podatci mogu se analizirati i s obzirom na neku općenitiju metodu linearne regresije, metodu nelinearne regresije ili neku drugu metodu. Analiza podataka dobivenih mjerenjem sa stanovišta metode jednostavne linearne regresije (ili neke druge metode) samo je jedan od faktora za razumijevanje veze među dvjema veličinama u inženjerstvu.

5. Zaključak

U praksi, kod primjene linearne regresije, polazi se od skupa podataka tipa (1.2) koji su dobiveni iz n mjerenja veličine y pri različitim vrijednostima veličine x. U simulaciji je $n = 7$, što je u skladu s činjenicom da broj mjerenja u pravilu nije velik. Simulirani podatci generirani su tako da zadovoljavaju linearnu vezu (1.8), ali iz slučajno odabranog niza od 7 podataka (koji odgovaraju situaciji koja se događa u inženjerskoj praksi) to se ne može zaključiti. Ako je i poznato da je veza linearna, ili se ima razloga prihvatiti da je takva, procjena nepoznatih parametara iz relativno malog broja mjerenja može biti nepouzdana. Na primjer, procjena parametra σ^2 iz 100 nizova po 7 mjerenja ispala je 3,49633 (tablica 2), što se osjetno razlikuje od prave vrijednosti koja je jednaka 4. Tek kad je broj mjerenja vrlo velik (u predočenoj simulaciji 10 000 puta po 7 mjerenja), rezultati mjerenja uvjerljivo ukazuju na linearnost veze (ako veza zaista jest linearna) i uspješno procjenjuju parametre. Računalnom simulacijom mogu se generirati podatci koji odgovaraju po volji velikom broju mjerenja i koji zadovoljavaju uvjete jednostavne linearne regresije. To može poslužiti za uvjerljivu demonstraciju statističkih zakonitosti i kao zamjena za njihovo dokazivanje koje je često matematički zahtjevno.

Popis kratica i simbola
List of abbreviations and symbols

$\bar{x}$	– aritmetička sredina $(x_1 + x_2 + \ldots + x_n)/n$ – arithmetic mean
SS_x	– suma kvadrata odstupanja od aritmetičke sredine – sum of squares of deviations from the mean
$\hat{a}$, $\hat{b}$	– procjene parametara a, b – estimations of parameters a, b
$\hat{A}$, $\hat{B}$	– procjenitelji parametara a, b – estimators of parameters a, b
$E(X)$	– očekivanje slučajne varijable X – expectation of random variable X
$V(X)$	– varijanca slučajne varijable X – variance of random variable X
$N(\mu, \sigma^2)$	– normalna razdioba s očekivanjem μ i varijancom σ^2 – normal distribution with expectation μ and variance σ^2
$N(0, 1)$	– jedinična normalna razdioba – standard normal distribution, unit normal distribution
$\chi^2(n)$	– hi-kvadrat razdioba s n stupnjeva slobode – chi-squared distribution with n degrees of freedom
$t(n)$	– Studentova t-razdioba s n stupnjeva slobode – Student's t-distribution with n degrees of freedom
$X \sim N(\mu, \sigma^2)$	– normalna slučajna varijabla X, slučajna varijabla X s normalnom razdiobom – normal random variable, normally distributed random variable X
s^2	– procjena od σ^2 – estimation of σ^2

S^2 – procjenitelj od σ^2
– estimator of σ^2

e_i – slučajna vrijednost normalne varijable $N(0, \sigma^2)$
– random value of normal variable $N(0, \sigma^2)$

$t_p(k)$ – vrijednost na osi x iza koje je površina ispod grafa $t(k)$ grafa jednaka p
– value on x axis after which the surface area equals p

Literatura
References

1. Željko Pauše, Uvod u matematičku statistiku, Školska knjiga, Zagreb 1993.
2. *F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, L. E. Meester,* A Modern Introduction to Probability and Statistics, Understanding Why and How, Springer-Verlag London Limited 2005., ISBN 1-85233-896-2, doi: https://doi.org/10.1007/1-84628-168-7.
3. *G. G. Roussas*, An Introduction to Probability and Statistical Inference, 2nd Ed., Elsevier, Academic Press, 2015., (ISBN: 978-0-12-800114-1).
4. *D. Jelaska*, Elementi strojeva, str. 38, URL: http://marjan.fesb.hr/~djelaska/documents/ES-skripta-760.pdf (1. 8. 2016.).
5. *D. D. Harwig, W. Ittiwattana, H. Castner*, Advances in Oxygen Equivalence Equations for Predicting the Properties of Titanium Welds, *The Welding Journal* **80** (2001) 126s–136s.
6. W. C. Navidi, Principles of Statistics for Engineers and Scientists, McGraw-Hill Company, 2009., (ISBN: 978–0–07–337634–9).

6. Prilozi

Program u Matlabu

```
clc
close all
clear all

a=2; %a i b su parametri pravca
b=1;
pom=7; %broj uređenih parova (x,y)
unutar=0;
unutar2=0;
broj=10000; %broj nizova od 'pom' brojeva

for i=1:broj
  for x=1:pom
     y(x)=a*x+b+normrnd(0,2); %normrnd(očekivanje,sigma)
  end

  %racunanje s kvadrat
  SSe=0;
  SSx=0;
  sum1=0; %za računanje akapa
  for x=1:pom
     SSx=SSx+(x-sum(1:pom)/pom)^2; %formula (1.6)
     sum1=sum1+(x-sum(1:7)/pom)*y(x);
  end
  akapa(i)=sum1/SSx; %procjena a (1.4)
  bkapa(i)=sum(y)/pom-akapa(i)*sum(1:pom)/pom; %procjena
b (1.4)

  %simulacija hi^2 razdiobe i računanje SSe
  for x=1:pom
     SSe=SSe+(y(x)-akapa(i)*x-bkapa(i))^2; %formula (3.6)
  end

  skvadrat2(i)=SSe/(pom-2); %procjena sigma^2 (formula (3.4))
  skvadrat1(i)=SSe/(pom-1); %informativno
  skvadrat0(i)=SSe/(pom);

  hikvadrat(i)=SSe/4; % (i) nakon formule (3.11)
  %za simulaciju studentove razdiobe
  studentA(i)=(akapa(i)-a)*sqrt(SSx)/sqrt(SSe/(pom-2));
%formula (3.5.1)
  studentB(i)=(bkapa(i)-b)/sqrt(skvadrat2(i))/sqrt(1/
pom+(sum(1:pom)/pom)^2/SSx); %formula (3.5.2)

  %razina pouzdanosti
  if (akapa(i)-tinv(0.975,5)*sqrt(skvadrat2(i))/sqrt(SSx))<a
%formula (3.8)
        if (akapa(i)+tinv(0.975,5)*sqrt(skvadrat2(i))/sqrt(SSx))>a
           unutar=unutar+1;
        end
  end

  if (bkapa(i)-tinv(0.975,5)*sqrt(skvadrat2(i))*sqrt(1/
pom+(sum(1:pom)/pom)^2/SSx))<b %formula (3.9)
     if (bkapa(i)+tinv(0.975,5)*sqrt(skvadrat2(i))*sqrt(1/
pom+(sum(1:pom)/pom)^2/SSx))>b
        unutar2=unutar2+1;
     end
  end
end

sirine=[max(hikvadrat(1:100))/15 max(hikvadrat(1:500))/20
max(hikvadrat(1:1000))/25 max(hikvadrat)/30];
granica=[100 500 1000 broj];

%hi^2 SIMULACIJA
for c=1:4
   for k=1:(10+c*5)
      zbroj(c,k)=0;
      donj=(k-1)*sirine(c);
      gornj=k*sirine(c);
      for i=1:granica(c)
         if hikvadrat(i)<=gornj
            if hikvadrat(i)>donj
               zbroj(c,k)=zbroj(c,k)+1;
            end
         end
      end
      yos(c,k)=zbroj(c,k)/granica(c)/sirine(c);
      xos(c,k)=(-0.5+k)*sirine(c);
   end

end

%STUDENT SIMULACIJA
stsirina=0.5;
donjagranica=-7;
for k=1:28
   stA(k)=0;
   stB(k)=0;
   donj=donjagranica+(k-1)*stsirina;
   gornj=donjagranica+k*stsirina;
   for i=1:broj
      if studentA(i)<=gornj
         if studentA(i)>=donj
            stA(k)=stA(k)+1;
         end
      end
      if studentB(i)<=gornj
         if studentB(i)>=donj
            stB(k)=stB(k)+1;
         end
      end
   end
```

```
    xstudent(k)=donjagranica+(-0.5+k)*stsirina; %gustoća razreda
se pridružuje sredini intervala
    ystudentA(k)=stA(k)/broj/stsirina; %prosječne gustoće razreda
    ystudentB(k)=stB(k)/broj/stsirina;
end

%graf hi^2 razdiobe
hi=[0:0.1:27];
yhi = chi2pdf(hi,5); %teorijska krivulja (hi,yhi)
figure
subplot(4,1,4), plot(xos(4, 1:30),yos(4, 1:30),'o', hi,yhi, '-r')
title('10000 podataka u 30 razreda'); xlabel('x'); ylabel('y');
grid on; legend('eksperiment', 'hi^2(x,5)'); xlim([0 27])
subplot(4,1,3), plot(xos(3, 1:25),yos(3, 1:25),'o', hi,yhi, '-r')
title('1000 podataka u 25 razreda'); xlabel('x'); ylabel('y'); grid
on; legend('eksperiment', 'hi^2(x,5)'); xlim([0 27])
subplot(4,1,2), plot(xos(2, 1:20),yos(2, 1:20),'o', hi,yhi, '-r')
title('500 podataka u 20 razreda'); xlabel('x'); ylabel('y'); grid
on; legend('eksperiment', 'hi^2(x,5)'); xlim([0 27])
subplot(4,1,1), plot(xos(1, 1:15),yos(1, 1:15),'o', hi,yhi, '-r')
title('100 podataka u 15 razreda'); xlabel('x'); ylabel('y'); grid
on; legend('eksperiment', 'hi^2(x,5)'); xlim([0 27])
grid on

%graf studentove razdiobe
student=[-8:0.1:8];
ystudent=tpdf(student,pom-2); %teorijska krivulja
(student,ystudent)
figure
plot(student,ystudent,'-r', xstudent, ystudentA, 'o', xstudent,
ystudentB, 'v'); grid on

%ISPIS PODATAKA
fprintf('JEDAN NIZ\n')
fprintf('nazivnik iz (3.4) je n-2\ns^2=%g\t\t s=%g\n',
skvadrat2(1),sqrt(skvadrat2(1)))
for c=1:4
    fprintf('\n%g NIZ\n',granica(c))
    fprintf('nazivnik iz (3.4) je n-2\ns^2=%g\t\
ts=%g\n',sum(skvadrat2(1:granica(c)))/
granica(c),sqrt(sum(skvadrat2(1:granica(c)))/granica(c)))
    fprintf('nazivnik iz (3.4) je n-1\ns^2=%g\t\
ts=%g\n', sum(skvadrat1(1:granica(c)))/
granica(c),sqrt(sum(skvadrat1(1:granica(c)))/granica(c)))
    fprintf('nazivnik iz (3.4) je n\ns^2=%g\t\
ts=%g\n', sum(skvadrat0(1:granica(c)))/granica(c),
sqrt(sum(skvadrat0(1:granica(c)))/granica(c)))
end
fprintf('\n\nUz razinu pouzdanosti 5%%,\na(kapa) je prihvaćen
prema formuli (3.8) u %g%% slučajeva \n(Prihvaćeno je %g/%g
podataka)\n',unutar/broj*100,unutar,broj)
fprintf('b(kapa) je prihvaćen prema formuli (3.9) u %g%%
slučajeva \n(Prihvaćeno je %g/%g podataka)\n',unutar2/
broj*100, unutar2, broj)
```

SUMMARY

Simulation of Simple Linear Regression

*Stjepan Džalto** and *Ivica Gusić*

The purpose of this paper is a computer simulation of conditions relevant for simple a linear regression model and computer confirmation of its basic equations. To that end, a simple linear regression model is described and mathematical foundations of the model are discussed. Listed are the equations for the objective function (Equation 1.3), regression line parameters (Equation 1.4), estimation of regression line parameters (Equation (3.2)), and confidence interval (Equations (3.8) and (3.9)). Estimation of variance (Equation (3.4)) is based on Equations (3.6) and (3.11), while (3.11) is based on (3.12). The conditions of the simple linear regression were simulated in Matlab. The model parameters were selected with Equations (4.1) and (4.2), and 10 000 series of 7 data were generated as a simulation of 10 000 experiments under the same conditions in engineering practice. Each series represented a measurement of a dependent variable for seven fixed independent variable values in circumstances in which the linear regression model assumptions had been satisfied. For a randomly chosen series of 7 data, the estimates of parameters can significantly deviate from true parameter values (Table 1), indicating that a relatively small number of measurements in practice can lead to unreliable estimates. The estimate can deviate from true value even if the number of measurements is relatively large (Table 2). On the other hand, it is shown that the arithmetic mean of 10 000 calculated parameters is almost identical to true parameter values. In other words, it is confirmed that the estimates from consecutive measurements under the same conditions are, in average, correct. Simulation of 10 000 series also confirmed other mentioned equations: distribution from Equation (3.12) (Table 3 and Fig. 2), *t*-distribution from (3.5.1) and (3.5.2) (Table 4 and Fig. 3), and confidence intervals for regression line parameters from Equations (3.8) and (3.9). The computer simulation can serve for the better understanding of the simple linear regression model and successfully replace proving the mathematical facts on which linear regression is based.

Keywords
Simple linear regression, normal distribution, chi-squared distribution, Student's distribution, confidence interval, Matlab

Faculty of Chemical Engineering and Technology
Marulićev trg 19,
10 000 Zagreb,
Croatia

Professional paper
Received February 4, 2016
Accepted May 30, 2016