UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
DEPARTMAN ZA FIZIKU

Uticaj i tretman grešaka prilikom fitovanja podataka sa linearnom zavisnošću

- završni rad -

Mentor:
prof. dr Tijana Prodanović

Kandidat:
Srđan Šibalić

Novi Sad, 2016

Sadržaj

1. Uvod

Ovaj rad predstavlja pregled uobičajenih metoda fitovanja podataka sa linearnom zavisnošću: metod najmanjih kvadrata, Demingova regresija i metod ukupnih najmanjih kvadrata. Posebno se osvrćemo na uticaj grešaka na rezultate fita.

Uvodno poglavlje se bavi kratkim istorijatom problematike analize podataka, osnovnim pojmovima koji će biti korišćeni u daljem radu, generisanju nasumičnih podataka za potrebe demonstracije fitovanja različitih tipova podataka i osnovnim parametrima koji opisuju određeni set podataka.

U poglavljima 2, 3 i 4 su detaljno opisane početne pretpostavke i izvođenja pojedinačnih metoda fitovanja podataka sa linearnom zavisnošću. U navedenim poglavljima se razmatraju prednosti i nedostaci opisanih metoda. U poglavlju 5 se ukratko opisuje problem analize podataka sa asimetričnim neodređenostima.

Na kraju rada su predstavljeni i upoređeni dobijeni parametri svih opisanih metoda za generisani set podataka.

1.1. Istorijski pregled

S obzirom da je najveći deo ovog rada posvećen upravo metodi najmanjih kvadrata, prikladno je priložiti kratak istorijat otkrića i razvoja dotične metode. Ovaj metod je nastao krajem 18. i početkom 19. veka u vreme velikih geografskih otkrića, kao odgovor na potrebu astronoma i geodeta da precizno opišu kretanje nebeskih tela da bi omogućili bolju navigaciju brodovima na otvorenom moru.

Adrien-Marie Legendre (Adrijen-Mari Ležandr, 1752-1833) je bio francuski matematičar koji je odgovoran za popularizaciju metoda najmanjih kvadrata. Veoma brzo prihvatanje metoda kao standardnog metoda od strane naučnika Francuske, Italije, Prusije, a kasnije i Engleske, se desilo zahvaljujući radu iz 1805. godine u kojem Legendre veoma pristupačno opisuje svoj metod i primenjuje ga na podacima iz 1795. o francuskom meridijanskom luku.

Slika 1. 1820. karikatura Adrien-Marie Legendre-a francuskog umetnika Julien-Leopold Boilly.

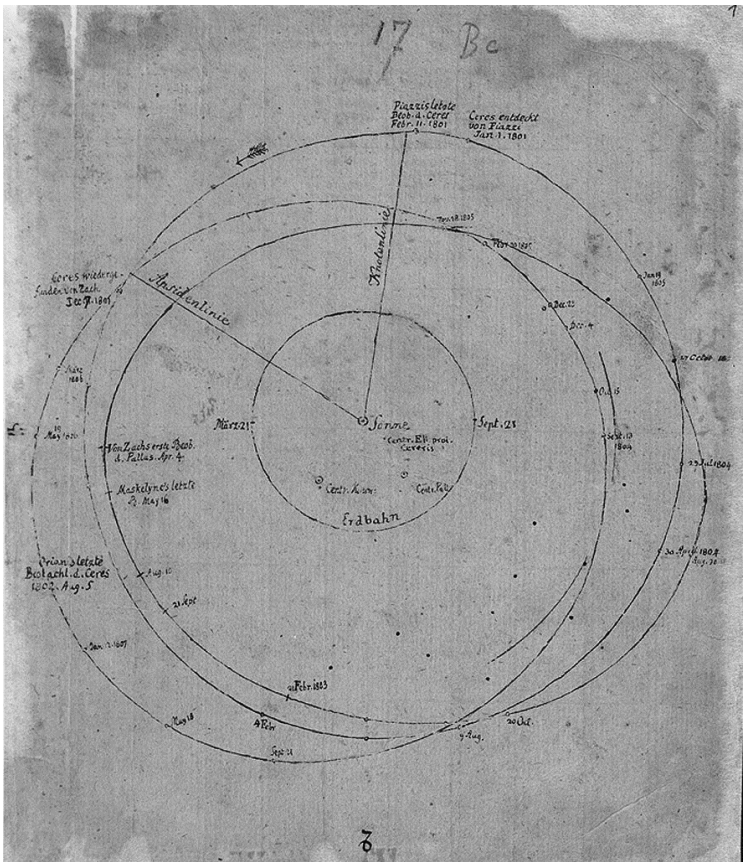
Legendre u svom radu iz 1805. (strane 72-73) o biranju sume kvadrata grešaka kaže sledeće:

„Od svih principa koji se mogu predložiti u ovu svrhu, moje mišljenje je da ne postoji uopšteniji, tačniji, i jednostavniji za primenu, od onog koji smo koristili u ovom radu; sastoji se od *minimalizacije* sume kvadrata grešaka. Ovim metodom, uspostavlja se neka vrsta ravnoteže među greškama koja, pošto sprečava ekstreme da dominiraju, je prikladna za otkrivanje stanja sistema koje se najviše približava istini.“

1809. godine Carl Friedrich Gauss (Karl Fridrih Gaus) (1777.-1855.) objavljuje rad o izračunavanju orbita nebeskih tela u kom tvrdi da koristi metod najmanjih kvadrata od 1795. Iako je nemoguće potvrditi ili opovrgnuti njegovu tvrdnju, svakako je nesumnjivo da je otišao korak dalje od Legendre-a i povezao metod najmanjih kvadrata sa principima verovatnoće i normalnom raspodelom. Uspeo je da upotpuni Laplace-ov pokušaj određivanja matematičkog oblika gustine verovatnoće posmatranja, koja zavisi od konačnog broja nepoznatih parametara, i definiše metod procene koji minimalizuje grešku procene. Gauss je pokazao da je aritmetička sredina zaista najbolja procena parametra lokacije menjajući i gustinu verovatnoće i metod procene. Okrenuvši problem, pitao se koji metod procene treba da se koristi da bi se dobila aritmetička sredina kao procena parametra lokacije. Na ovaj način je došao do normalne raspodele.

Slika 2. Carl Friedrich Gauss.

Jedna od prvih potvrda metoda najmanjih kvadrata se našla u predviđanju pozicije novootkrivenog asteroida Ceresa. 1. januara 1801. godine italijanski astronom Giuseppe Piazzi (Đuzepe Pjaci) je otkrio Ceres i uspeo da prati njegovu putanju 40 dana pre nego što je nestao u svetlosti Sunca. Na osnovu ovih posmatranja astronomi su pokušali da odrede poziciju Ceresa i jedina predviđanja koja su urodila plodom su bila Gauss-ova koristeći metod najmanjih kvadrata.



Slika 3. Originalna skica orbita asteroida Ceres, Palas i Vesta iz Gauss-ovog rada („Astronomische Untersuchungen und Rechnungen vornehmlich über die Ceres, Ferdinandea“. 1802.).

U naredna dva veka naučnici u oblasti teorije grešaka i statistike su uspeli pronaći mnoštvo načina primene metode najmanjih kvadrata.

1.2. Osnovni pojmovi

Uvešćemo neke osnovne pojmove iz statistike na primeru Hablovog zakona. Edwin Hubble (Edvin Habl, 1889-1953) je bio američki astronom koji je poznat po tome što je pokazao da se brzina udaljavanja galaksija povećava sa udaljenošću od Zemlje, što dovodi do zaključka da se svemir širi. Ova veza je poznata kao Hablov zakon, iako je prethodno otkrivena od strane Georges Lemaître-a (Žorž Lemetr, 1894-1966).

Dakle, želimo da odredimo brzinu širenja svemira na osnovu brzine udaljavanja galaksija. S obzirom da postoji više galaksija nego što smo u stanju da izbrojimo, nismo u mogućnosti da ih sve uzmemo u obzir. Ono što možemo da uradimo je da posmatramo određen broj galaksija, i pretpostavimo da zakonitost utvrđena posmatranjem nekih drugih galaksija važi i za njih.

Sve galaksije u svemiru čine **populaciju** a posmatrane galaksije čine **uzorak**. Dakle, populacija je set objekata ili događaja slične prirode od značaja za neki eksperiment. Shodno tome, uzorak je podskup objekata ili događaja izabran iz jedne populacije.

Izvođenje zaključaka o parametrima populacije na osnovu statistike uzorka naziva se **statističko zaključivanje**. Da bi ovako izvedeni zaključci imali težinu (tj. da bi opisivali stvarno stanje populacije) uzorak mora dobro predstavljati populaciju. Raspodela uzorka mora biti slična raspodeli populacije. Mora se izbeći pristrasnost pri uzorkovanju, tj. sistematska težnja da uzorkujemo podatke koji ne predstavljaju populaciju dovoljno dobro. U tom slučaju bi se raspodele uzorka i populacije razlikovale što bi dovelo do loših zaključaka.

Da bismo izbegli pristrasnost u uzorkovanju koristimo nasumično biranje. Nasumično izabran uzorak daje bolju predstavu o populaciji od bilo kog nenasumičnog metoda biranja. Nasumično izabran uzorak minimalizuje grešku pri zaključivanju i dozvoljava procenu preostale greške.

Posmatramo set podataka koji se sastoji od niza merenja x_i. Set podataka se najčešće opisuje, pre svega, **srednjom vrednošću**. Srednja vrednost ili očekivana vrednost je prosta aritmetička sredina svih vrednosti u uzorku. Dakle, sumiramo sve članove x_i i podelimo sa brojem članova N:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.1.1}$$

Pored srednje vrednosti korisno je podatke opisati i **varijansom**. Varijansa je prosečno kvadrirano odstupanje podataka od srednje vrednosti:

$$s_{xx} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{1.1.2}$$

S obzirom da se kvadrira odstupanje podataka od srednje vrednosti, pridaje se mnogo veći značaj podacima koji više odstupaju što znači da je varijansa veoma osetljiva na ekstremne podatke. Pored

ovog nedostatka, još jedna mana varijanse je što se ne može porediti sa srednjom vrednošću pošto je izražena u kvadriranim jedinicama.

Izraz (1.1.2) predstavlja varijansu uzorka. Ako se srednja vrednost uzorka $\bar{x}$ poklapa sa srednjom vrednošću populacije (u literaturi se obično označava sa μ) pomenuti izraz će važiti i za varijansu populacije. Za češći slučaj kada se srednje vrednosti uzorka i populacije razlikuju, izraz (1.1.2) će uvek potcenjivati varijansu populacije, tj. unosiće sistematsku grešku u procenu varijanse populacije. Za korigovanu procenu varijanse populacije koristi se sledeći izraz:

$$s_{xx} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad (1.1.3)$$

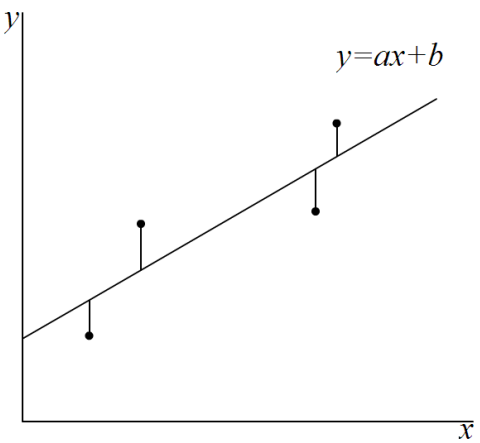
U gornjem izrazu faktor $1/(N-1)$ se naziva Beselova korekcija, po Friedrich Wilhelm Bessel-u (Fridrih Vilhelm Besel, 1784 - 1846).

Standardna devijacija je definisana kao kvadratni koren varijanse

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad (1.1.4)$$

Prednost u odnosu na varijansu je što je izražena u istim jedinicama kao i srednja vrednost (kao i podaci). Podaci koji znatno odstupaju od srednje vrednosti manje utiču na standardnu devijaciju nego kod varijanse ali i dalje značajno utiču.

Pretpostavimo da smo za neki nasumično izabran uzorak galaksija izmerili udaljenosti i brzine. Sada želimo da pronađemo kako su ove dve veličine povezane. Tim problemom se bavi **regresiona analiza**. Dakle, regresiona analiza pronalazi izraz koji najbolje opisuje vezu dve ili više promenljivih. Ta veza se naziva linija najboljeg fita. U našem slučaju Hablovog zakona, postoje dve promenljive, a podaci pokazuju linearnu zavisnost te se koristi **prosta linearna regresija**. Linearna regresija se bavi problemom nalaženja parametara funkcije linearne zavisnosti između zavisne promenljive i jedne ili više nezavisnih promenljivih. Najčešće korišćeni metod je **metod najmanjih kvadrata** (MNK). Sastoji se u smanjivanju sume kvadriranih vertikalnih rastojanja podataka od linije najboljeg fita (**Slika 4**):

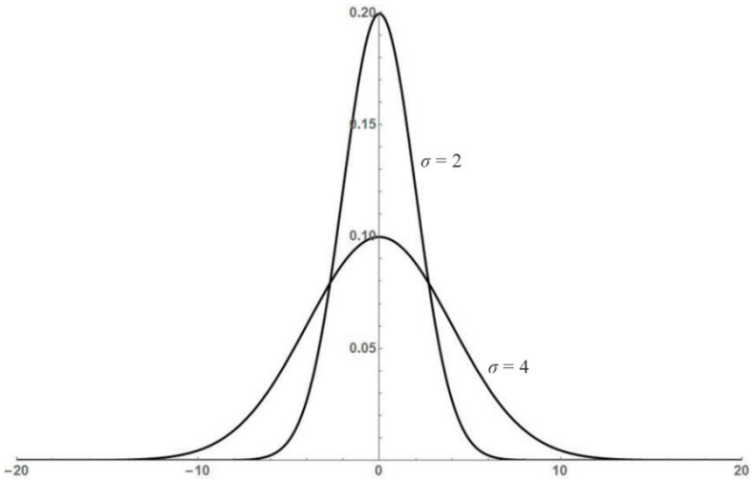


Slika 4. Prikaz vertikalnih rastojanja podataka (oznacenih tačkama) od linije linearnog fita $y=ax+b$,

MNK pretpostavlja da postoji jedna nezavisna promenljiva koja je određena tačno (bez greške) i jedna zavisna promenljiva čije su neodređenosti opisane Gausovom (normalnom) raspodelom koja ima oblik:

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1.1.5)$$

gde je σ standardna devijacija a μ je srednja vrednost ili očekivana vrednost raspodele. Gausova raspodela je prikazana na **Slici 5**.

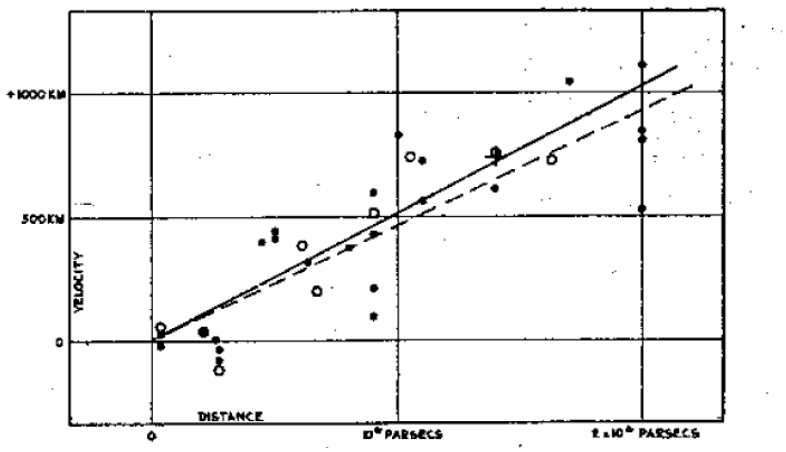


Slika 5. Gausove raspodele za $\mu = 0$ i $\sigma = 2$, $\sigma = 4$.

Tražimo izraz $y = ax + b$ takav da je suma kvadrata vertikalnih rastojanja podataka od linije najboljeg fita (χ^2) minimalna[1].

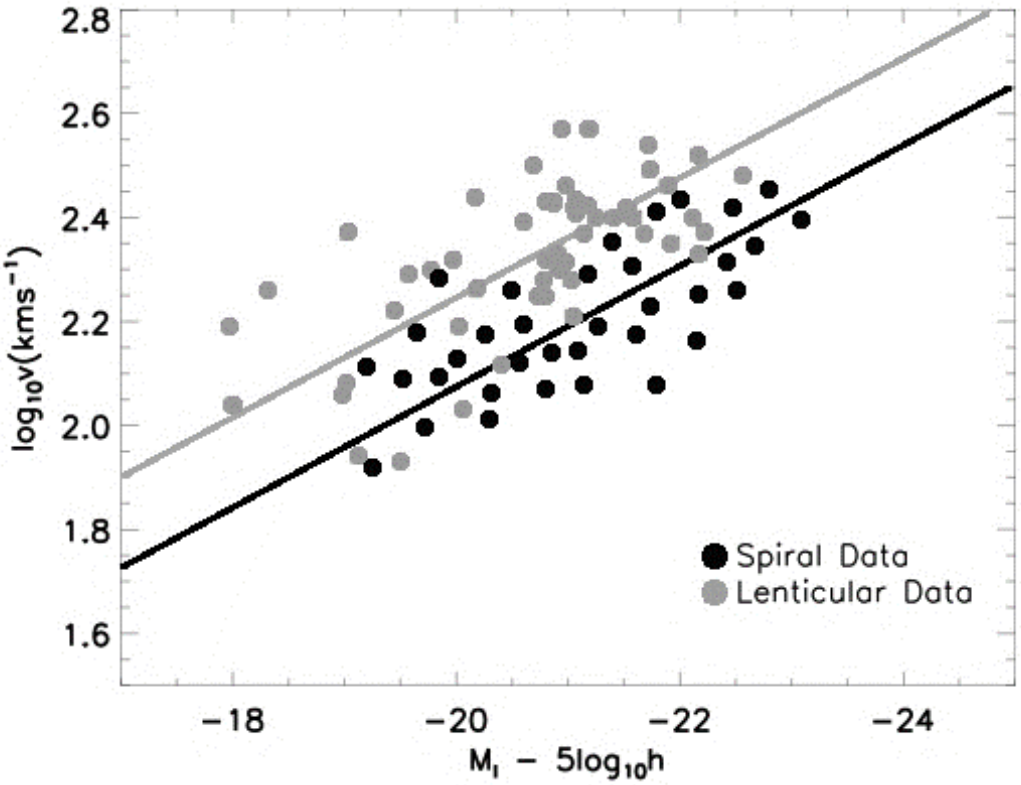
$$\chi^2 = \sum_{i=1}^{N} (y_i - (ax_i + b))^2 \qquad (1.1.6)$$

Parametar a se naziva **nagib** (eng. slope) prave, a parametar b označava presek sa ordinatom (eng. intercept). Na **Slici 6.** su prikazani Hablovi podaci o brzinama i udaljenostima relativno malog uzorka galaksija. Pošto se posmatra zavisnost brzine udaljavanja galaksija od njihove udaljenosti, parametar b je nula (tj. brzina udaljavanja u tački posmatrača je nula) tako da se nalaženje funkcije najboljeg fita svodi na određivanje parametra a, tj. nagiba prave.



Slika 6. Originalni grafik iz Hablovog rada 1929. godine. Na apscisi (x) su udaljenosti galaksija a na ordinati (y) brzine udaljavanja galaksija.

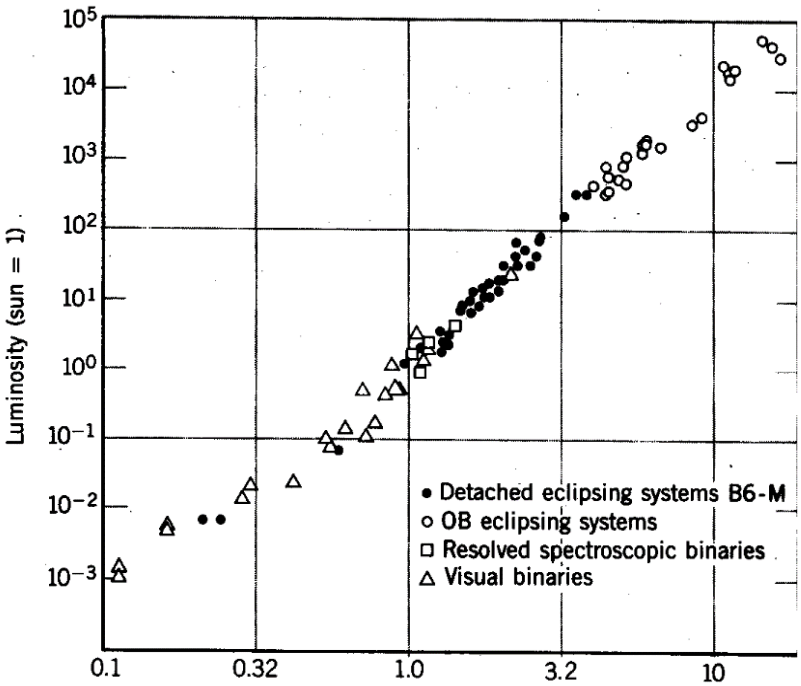
[1] Naravno, ovo nije jedini metod procene najboljeg fita. Pored najmanjih kvadrata može se koristiti najmanje apsolutno odstupanje gde se umesto kvadrata razlike u sumi koristi apsolutna vrednost razlike. Međutim, ovaj metod je nestabilan s obzirom da za određeni set podataka može da postoji beskonačno mnogo rešenja. Takođe, pored ovog i brojnih drugih metoda, postoji i Theil-Sen metod koji nalazi nagib najboljeg fita kao medijanu nagiba svih parova tačaka iz posmatranog uzorka. Prednost ovog metoda je što je veoma neosetljiv na tačke koje odstupaju.



Slika 7. Tali-Fišerova relacija za spiralne i sočivaste galaksije – veza mase ili luminoznosti spiralnih galaksija i ugaone brzine ili širine emisionih linija. Preuzeto sa https://en.wikipedia.org/wiki/Tully%E2%80%93Fisher_relation

Često se u astronomiji pojavljuju eksponencijalne zavisnosti koje se mogu svesti na linearnu zavisnost ako se prikažu u logaritamskoj skali. Primeri linearnih zavisnosti u astronomiji su dati na **Slici 7** i **Slici 8**.

Same vrednosti parametara *a* i *b* nisu od velike koristi jer one predstavljaju samo najverovatniju vrednost raspodele (moda) svih parametara *a* i *b* koji opisuju dati uzorak. Da bismo znali kako izgleda raspodela ovih parametara potrebne su nam i njihove neodređenosti. Na primer, kod gausove raspodele je to parametar σ – standardna devijacija, koja opisuje širinu raspodele. To znači da će se, kod gausove raspodele centrirane oko nule (μ = 0), 68.3% svih vrednosti nalaziti u intervalu [-σ, σ]. Za slučaj da je raspodela složenija uz dobijene parametre se moraju objaviti svi dodatni parametri potrebni za rekonstrukciju raspodele verovatnoće.



Slika 8. Veza mase i luminoznosti zvezda. Preuzeto sa http://chinook.kpc.alaska.edu/~ifafv/lecture/lecmeas.htm

U astronomiji nismo u mogućnosti da obavljamo eksperimente u kontrolisanim uslovima laboratorije, pa jedino što možemo je da posmatramo objekte i događaje u prirodi što dovodi do činjenice da su neodređenosti koje se javljaju pri posmatranju obično veće nego kod drugih prirodnih nauka. Neodređenosti (greške) pri merenju se mogu podeliti u dve kategorije: nasumične i sistematske neodređenosti. Nasumične neodređenosti potiču iz nemogućnosti da kontrolišemo sve uslove posmatranja. Uvek će postojati neki efekat koji nam je nepoznat a koji unosi šum u posmatranje. Sistematske greške imaju tendenciju da pomere vrednosti merenja u jednu stranu. Sem procesa koji nam nisu poznati uzrok može biti i loše kalibrisan instrument ili greška u modelu koji koristimo za opisivanje podataka. Nama su od interesa nasumične neodređenosti i kako one propagiraju kroz rezultate.

1.3. Podaci

Za potrebe demonstracije različitih metoda fitovanja generisani su podaci prikazani u **Tabeli 1**. Ovi podaci simuliraju nezavisno posmatranje linearno zavisnih promenljivih x i y. Kolone x i y su posmatrane vrednosti promenljivih koje sadrže određenu grešku merenja dok kolone $\hat{x}$ i $\hat{y}$ odgovaraju stvarnim vrednostima posmatranih promenljivih. Poslednje dve kolone su neodređenosti za obe promenljive. Tačna procedura dobijanja podataka je opisana na kraju poglavlja.

Tabela 1. Nasumično generisani podaci

$\hat{x}$	$\hat{y}$	x	y	σ_x	σ_y
7.47	13.74	6.77	10.69	1.14	1.09
52.74	36.37	53.56	37.2	1.68	1.67
43.68	31.84	42.84	29.35	1.21	3.51
75.79	47.89	74.68	46.00	1.20	3.83
64.30	42.15	62.90	43.50	1.56	3.18
60.92	40.46	61.43	40.11	1.08	1.67
96.31	58.15	99.66	54.00	1.75	3.96
62.57	41.29	62.74	41.76	1.52	1.39
73.53	46.77	73.47	51.57	1.97	3.44
27.68	23.84	27.31	25.74	1.84	3.16
22.35	21.17	20.68	24.60	1.10	2.16
51.66	35.83	53.33	39.39	1.73	3.79
70.83	45.42	65.86	47.44	1.46	3.51
89.25	54.62	86.64	56.29	1.47	2.59
74.55	47.28	76.14	47.400	1.35	1.10
54.77	37.38	55.13	40.03	1.66	2.24
79.25	49.62	80.76	49.38	1.39	3.02
21.25	20.62	20.18	20.46	1.25	3.86
5.10	12.55	7.62	21.05	1.39	3.59
33.00	26.50	32.14	25.36	1.10	1.22

Srednje vrednosti posmatranih promenljivih x i y su definisane jednačinama

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.2.1}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1.2.2}$$

Varijanse uzorka s_{xx} i s_{yy} opisuju odstupanje podataka od srednje vrednosti i predstavljene su jednačinama:

$$s_{xx} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{1.2.3}$$

$$s_{yy} = \frac{1}{N}\sum_{i=1}^{N} (\mathrm{y}_i - \bar{y})^2 \tag{1.2.4}$$

Kovarijansa s_{xy} opisuje koliko se dve promenljive zajedno menjaju:

$$s_{xy} = \frac{1}{N}\sum_{i=1}^{N} (\mathrm{x}_i - \bar{x})(\mathrm{y}_i - \bar{y}) \tag{1.2.5}$$

Standarde devijacije su date izrazima:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{1.2.6}$$

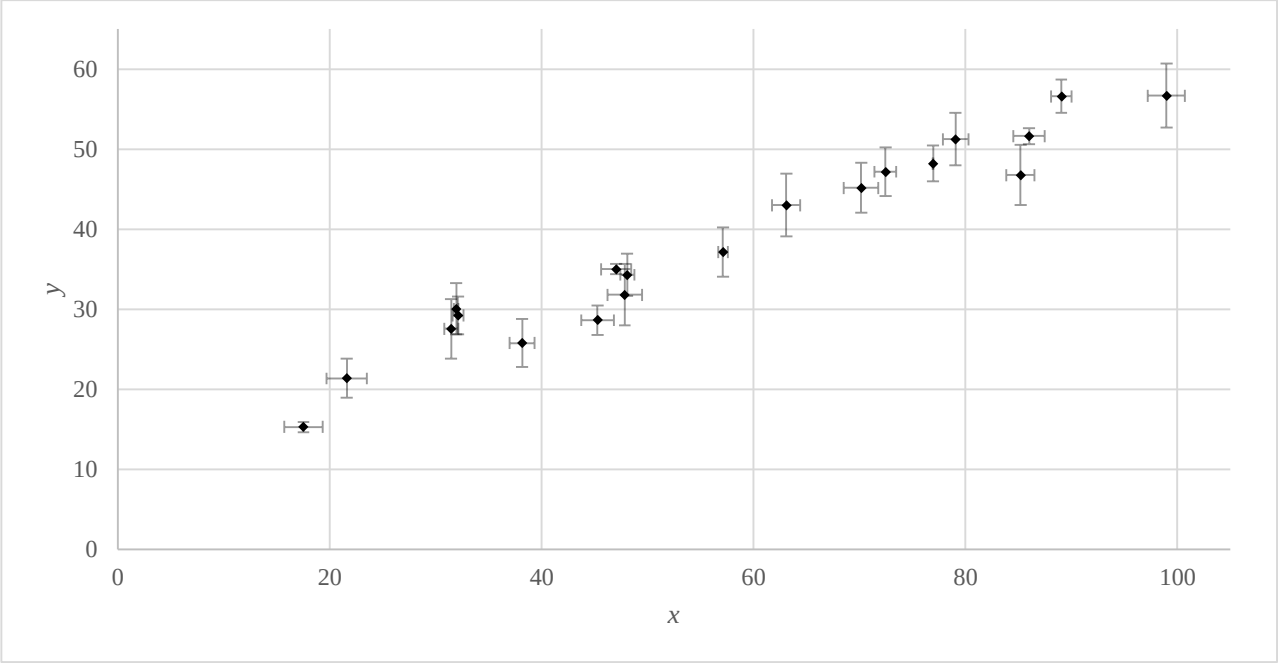
$$\sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{1.2.7}$$

Ako se izrazi (1.2.3) - (1.2.7) primenjuju na populaciju upotrebićemo Beselovu korekciju, tj umesto $1/N$ ćemo koristiti $1/(N-1)$.

Kao meru linearne zavisnosti između dve promenljive koristi se Pearson-ov korelacioni koeficijent *r*. Razvio ga je Karl Pearson (Karl Pirson, 1857-1936) na osnovu prethodne ideje Francis Galton-a (Fransis Galton, 1822-1911).

$$r = \frac{s_{xy}}{\sigma_x \sigma_y} \tag{1.2.8}$$

Korelacioni koeficijent r uzima vrednosti od -1 do 1, pri čemu vrednost 1 označava potpunu pozitivnu korelaciju (ako x raste, y raste), -1 potpunu negativnu korelaciju (ako x raste, y opada), dok nula označava nepostojanje korelacije.

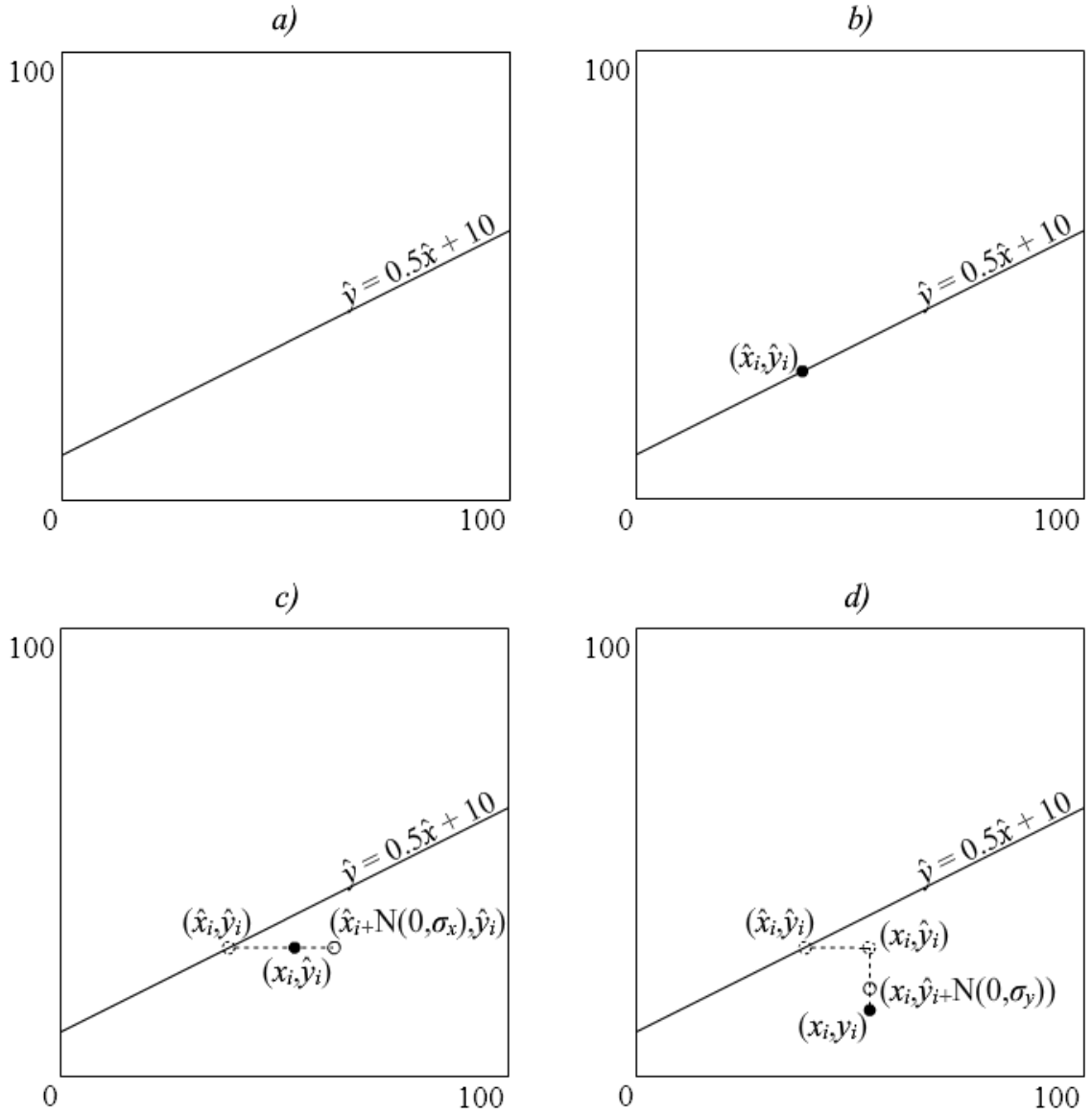


Slika 9. Grafički prikaz nasumično generisanih podataka iz **Tabele 1**.

Podaci su dobijeni na sledeći način: za n = [1 .. 20] za nasumično izabranu stvarnu vrednost $\hat{x}$ u rasponu [0 ... 100] se izračuna stvarna vrednost $\hat{y}$ na osnovu proizvoljno izabrane zavisnosti $\hat{y} = 0.5\hat{x}$ + 10 (prikazana na **Slici 9.**). Zatim, da bismo simulirali posmatranje, na obe vrednosti se dodaje nasumično odstupanje iz normalne raspodele sa srednjom vrednošću μ = 0 i nasumičnom standardnom devijacijom u opsegu σ_x = [1 .. 2] i σ_y = [1 .. 4] plus dodatno nasumično odstupanje iz normalne raspodele koje simulira nepoznati parametar posmatranja. Korišćene jednačine su date u izrazu (1.2.9). Nasumične greške će simulirati izvore nepoznatog šuma u podacima, tj. ako neka tačka odstupa značajnije od prave linearne zavisnosti a „izračunata“ neodređenost joj daje veću težinu, uticaće negativno na liniju najboljeg fita.

$$\begin{aligned} x_i &= \hat{x}_i + \mathrm{N}\left(0, \sigma_x^2\right) + \mathrm{N}\left(0,1\right) \\ y_i &= \hat{y}_i + \mathrm{N}\left(0, \sigma_y^2\right) + \mathrm{N}(0,1) \end{aligned} \qquad (1.2.9)$$

$\mathrm{N}(\mu, \sigma^2)$ je nasumična vrednost iz normalne raspodele sa srednjom vrednošću μ i standardnom devijacijom σ.



Slika 10. Grafički prikaz dobijanja podataka iz **Tabele 1.** a) proizvoljno izabrana linearna zavisnost, b) nasumično izabran uređen par koji leži na izabranoj pravoj, c) i d) dodaje se nasumično odstupanje iz normalne raspodele sa nasumičnim standardnim devijacijama σ_x na x i σ_y na y koordinatu kao i nasumično odstupanje iz normalne raspodele sa $\sigma = 1$.

2. Podaci bez poznatih neodređenosti

U ovom poglavlju će biti opisane dve standardne metode fitovanja podataka kod kojih nisu prijavljene neodređenosti ili su neodređenosti identične za sve tačke: metod najmanjih kvadrata i Demingova regresija.

2.1. Metod najmanjih kvadrata

Metod najmanjih kvadrata se sastoji u nalaženju funkcije oblika $y = ax + b$, za koju je zbir kvadrata vertikalnih rastojanja svake tačke iz datog seta podataka od linije fita minimalan.

Ovaj metod je opravdan za slučaj kada je x nezavisna a y zavisna promenljiva, a neodređenosti za x su zanemarljive u odnosu na neodređenosti za y, tj. pretpostavlja se da je x izmereno tačno. U praksi je ovaj uslov veoma retko ispunjen pa se za nezavisnu promenljivu bira ona kod koje su neodređenosti manje.

Za uređene parove (x_i, y_i) gde je $i = (1 \dots N)$ možemo definisati ukupnu grešku (odstupanje fita od podataka) za $y = ax + b$ kao

$$\chi^2(a,b) = \sum_{i=1}^{N} (y_i - (ax_i + b))^2 \tag{2.1.1}$$

Metod najmanjih kvadrata se sastoji u nalaženju vrednosti a i b koje minimalizuju ukupnu grešku. Dakle, tražimo vrednosti a i b takve da su parcijalni izvodi jednačine (2.1.1) jednaki nuli: $\partial\chi^2 / \partial a = 0$, $\partial\chi^2 / \partial b = 0$.

$$\frac{\partial \chi^2}{\partial b} = -2\sum_{i=1}^{N} (y_i - (ax_i + b)) = 0 \tag{2.1.2}$$

$$\begin{aligned} \sum_{i=1}^{N} y_i &= a\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} b \\ \frac{1}{N}\sum_{i=1}^{N} y_i &= a\frac{1}{N}\sum_{i=1}^{N} x_i + b \end{aligned} \tag{2.1.3}$$

$$\bar{y} = a\bar{x} + b \tag{2.1.4}$$

Pre nego što izraz (2.1.1) diferenciramo po a, izrazićemo b iz (2.1.4) i uvrstiti u (2.1.1)

$$\chi^2 = \sum_{i=1}^{N} (y_i - ax_i - \bar{y} + a\bar{x})^2 \tag{2.1.5}$$

Sada diferenciramo po a

$$\frac{\partial \chi^2}{\partial a} = -2\sum_{i=1}^{N}(y_i - \bar{y} - a(x_i - \bar{x}))(x_i - \bar{x}) = 0$$
$$a\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x}) \tag{2.1.6}$$

$$a = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2.1.7}$$

Ako iskoristimo definicije varijanse i kovarijanse dobijamo

$$a = \frac{s_{xy}}{s_{xx}} \tag{2.1.8}$$

$$b = \bar{y} - a\bar{x} \tag{2.1.9}$$

Ako želimo da fitujemo podatke tako da posmatramo x kao zavisnu a y kao nezavisnu promenljivu preuredićemo jednačinu $y = ax + b$ na sledeći način

$$x = \frac{1}{a}y - \frac{b}{a} \tag{2.1.10}$$

i uvesti smenu

$$m = \frac{1}{a} \quad \text{i} \quad n = -\frac{b}{a} \tag{2.1.11}$$

$$\bar{x} = m\bar{y} + n \tag{2.1.12}$$

te možemo postupiti na isti način kao u prethodnom slučaju pa dobijamo

$$m = \frac{s_{xy}}{s_{yy}}$$
$$n = \bar{x} - m\bar{y} \tag{2.1.13}$$

Sledi

$$a = \frac{s_{yy}}{s_{xy}}$$
$$b = \bar{y} - a\bar{x} \tag{2.1.14}$$

Varijanse procenjenih parametara su date izrazima

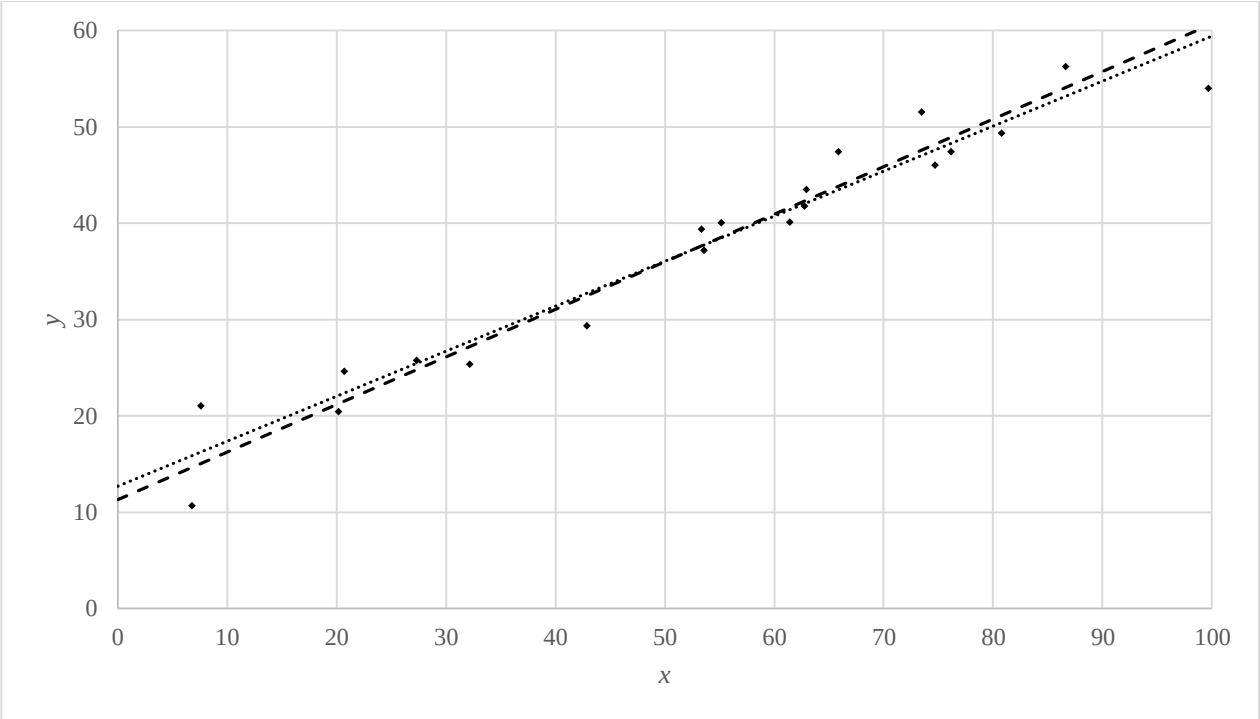
$$s_a^2 = \frac{s_\varepsilon^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}$$
$$s_b^2 = s_\varepsilon^2 \frac{\sum_{i=1}^{N} x_i^2}{N\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \tag{2.1.15}$$

gde je

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2}{N-2} \tag{2.1.16}$$

varijansa uzorka ili srednji kvadrat greške (eng. mean square error – MSE).

Za podatke iz **Tabele 1** kada fitujemo y = f(x) dobijamo a = 0.4675 ± 0.0259 i b = 12.699 ± 1.532, dok za obrnut slučaj x = f(y) dobijamo a = 0.49329 ± 0.02659 i b = 11.3269 ± 1.5734.



Slika 11. Grafički prikaz procenjenih fitova: ····· y = f(x); – – x = f(y).

2.2. Demingova regresija

Kod metode najmanjih kvadrata se posmatra zavisnost y od nezavisne promenljive x. Pretpostavlja se da je greška pri određivanju x zanemarljiva u odnosu na grešku za y. Ako posmatramo dve promenljive nezavisno jednu od druge (na primer, rektascenzija i deklinacija nekog objekta), obe su podložne eksperimentalnim i instrumentalnim greškama pa metod najmanjih kvadrata nije pogodan za fitovanje ovakvih podataka. U ovom slučaju se može koristiti Demingova regresija koja je poseban slučaj metode najmanjih ukupnih kvadrata.

U opštem slučaju imamo izmerene promenljive x i y koje se od stvarnih vrednosti razlikuju za ε i η:

$$y_i = \hat{y}_i + \varepsilon_i \tag{2.2.1}$$

$$x_i = \hat{x}_i + \eta_i \tag{2.2.2}$$

Greške ε i η su nezavisne jedna od druge i pretpostavlja se da je poznat odnos njihovih varijansi:

$$\mu = \frac{\sigma_\varepsilon^2}{\sigma_\eta^2} \tag{2.2.3}$$

Tražimo funkciju najboljeg fita

$$\hat{y} = a\hat{x} + b \tag{2.2.4}$$

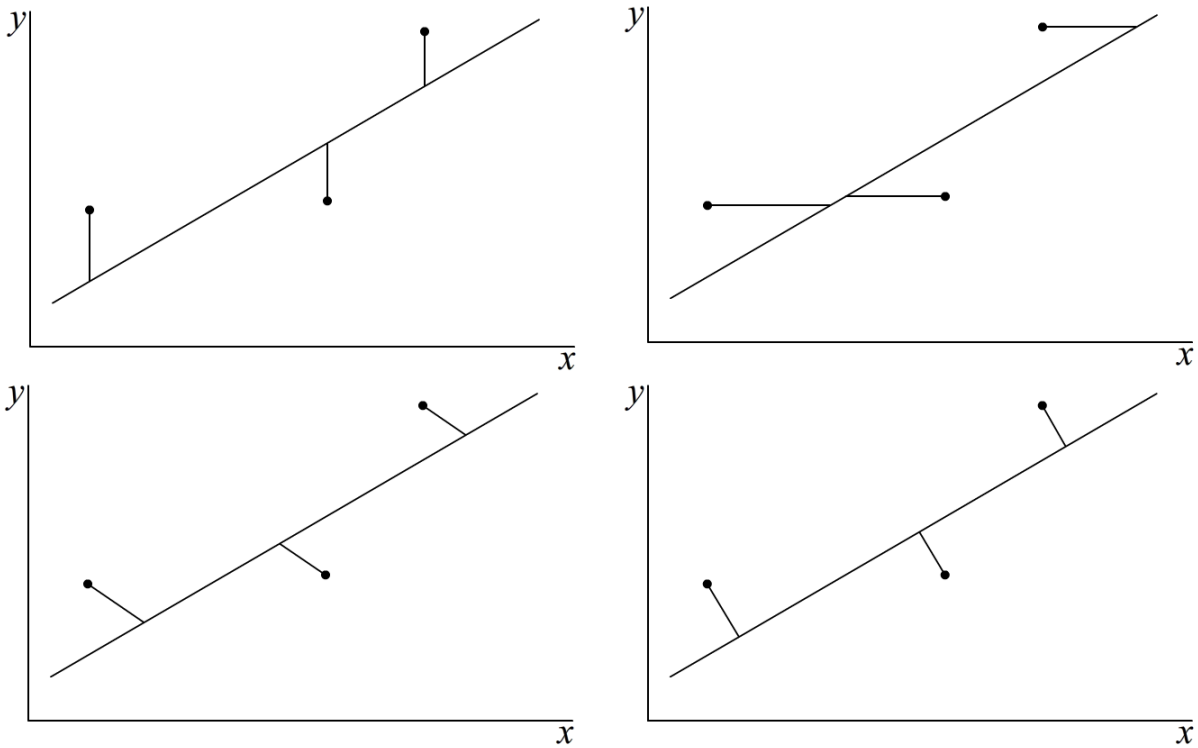
takvu da je suma kvadrata rastojanja tačaka od fita (χ^2) minimalna:

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{\varepsilon_i^2}{\sigma_\varepsilon^2} + \frac{\eta_i^2}{\sigma_\eta^2}\right) = \sum_{i=1}^{N}\left(\frac{\left(y_i - \hat{y}_i\right)^2}{\sigma_\varepsilon^2} + \frac{\left(x_i - \hat{x}_i\right)^2}{\sigma_\eta^2}\right) \tag{2.2.5}$$

$$\chi^2 = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{N}\left(\left(y_i - \hat{y}_i\right)^2 + \mu\left(x_i - \hat{x}_i\right)^2\right) \tag{2.2.6}$$

$$\chi^2 = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{N}\left(\left(y_i - a\hat{x}_i - b\right)^2 + \mu\left(x_i - \hat{x}_i\right)^2\right) \tag{2.2.7}$$

Primećuje se da ako je σ_ε nula (x je određeno bez greške), μ je nula pa se izraz (2.2.7) svodi na MNK za slučaj $y = f(x)$. S druge strane ako je σ_η nula izraz (2.2.7) se svodi na MNK za slučaj $x = f(y)$.



Slika 12. Merenje udaljenosti tačaka od linije najboljeg fita. *Levo gore*: metod najmanjih kvadrata u slučaju $y = f(x)$, *desno gore*: metod najmanjih kvadrata u slučaju $x = f(y)$, *levo dole*: Demingova regresija za slučaj kada je $\mu \neq 1$, *desno dole*: Demingova regresija za slučaj kada je $\mu = 1$ (ortogonalne udaljenosti).

Tačka (x_i, y_i) je statistički najbliža (pravac po kojem se meri udaljenost neke tačke od fita zavisi od μ (**Slika 12**)) tački $(\hat{x}_i, \hat{y}_i)$ na pravoj $\hat{y} = a\hat{x} + b$. Nagib prave $\hat{y} = a\hat{x} + b$ je a, a nagib prave od tačke (x_i, y_i) do statistički najbliže tačke na pravoj $\hat{y} = a\hat{x} + b$ je $\left(\varepsilon_i/\sigma_\varepsilon^2\right)/\left(\eta_i/\sigma_\eta^2\right)$. Pošto su ove dve prave normalne jedna na drugu proizvod nagiba je -1.

$$a\frac{\frac{\varepsilon_i}{\sigma_\varepsilon^2}}{\frac{\eta_i}{\sigma_\eta^2}} = a\frac{\sigma_\eta^2}{\sigma_\varepsilon^2}\frac{\varepsilon_i}{\eta_i} = a\frac{\sigma_\eta^2}{\sigma_\varepsilon^2}\frac{y_i - \hat{y}_i}{x_i - \hat{x}_i} = -1 \qquad (2.2.8)$$

$$a\frac{y_i - \hat{y}_i}{x_i - \hat{x}_i} = -\frac{\sigma_\varepsilon^2}{\sigma_\eta^2} = -\mu \qquad (2.2.9)$$

Ako kombinujemo jednačine (2.2.4) i (2.2.9) dobijamo

$$a\frac{y_i - a\hat{x}_i - b}{x_i - \hat{x}_i} = -\mu \qquad (2.2.10)$$

i odavde izrazimo $\hat{x}_i$

$$\hat{x}_i = \frac{1}{\mu + a^2}\left(a y_i - ab + \mu x_i\right) \tag{2.2.11}$$

i uvrstimo u (2.2.7) i dobijamo, posle sređivanja,

$$\chi^2 = \frac{1}{\sigma_\varepsilon^2}\frac{\mu}{\mu + a^2}\sum_{i=1}^{N}\left(y_i - a\hat{x}_i - b\right)^2 \tag{2.2.12}$$

Sada preuredimo

$$\sigma_\varepsilon^2 \frac{\mu + a^2}{\mu}\chi^2 = \sum_{i=1}^{N}\left(y_i - a\hat{x}_i - b\right)^2 \tag{2.2.13}$$

i diferenciramo po *b* pri čemu uzimamo $\partial\chi^2/\partial b = 0$ jer tražimo minimum

$$0 = -2\sum_{i=1}^{N}\left(y_i - a\hat{x}_i - b\right) \tag{2.2.14}$$

što se može napisati kao

$$\overline{y} = a\overline{x} + b \tag{2.2.15}$$

Ako izrazimo *b* iz (2.2.15) i uvrstimo u (2.2.12) i iskoristimo definicije za varijanse s_{xx}, s_{yy} i s_{xy} dobijamo

$$\sigma_\varepsilon^2 \frac{\mu + a^2}{\mu}\chi^2 = N\left(s_{yy} + a^2 s_{xx} - 2a s_{xy}\right) \tag{2.2.16}$$

Sada diferenciramo po *a* i postavimo $\partial\chi^2/\partial a = 0$ i iskoristimo pravilo $\left(u\cdot v\right)' = u'\cdot v + u\cdot v'$

$$\sigma_\varepsilon^2 \frac{\mu + a^2}{\mu}\cdot 0 + \frac{\partial\left(\sigma_\varepsilon^2 a^2/\mu\right)}{\partial a}\chi^2 = N\left(2a s_{xx} - 2s_{xy}\right) \tag{2.2.17}$$

$$2\frac{\sigma_\varepsilon^2}{\mu}a\chi^2 = N\left(2as_{xx} - 2s_{xy}\right) \tag{2.2.18}$$

Iz (2.2.16) i (2.2.18) eliminišemo χ^2

$$a(s_{yy} + a^2 s_{xx} - 2as_{xy}) = (\mu + \mathrm{a}^2)(as_{xx} - s_{xy}) \tag{2.2.19}$$

i posle kratkog sređivanja dobijamo

$$a^2 s_{xy} + a(\mu s_{xx} - s_{yy}) - \mu s_{xy} = 0 \tag{2.2.20}$$

Ako pretpostavimo da je $\mu s_{xy} \geq 0$ tražimo pozitivni koren jednačine (2.2.20):

$$\begin{aligned} a &= \frac{s_{yy} - \mu s_{xx} + \sqrt{\left(s_{yy} - \mu s_{xx}\right)^2 - 4\mu s_{xy}^2}}{2s_{xy}} \\ b &= \bar{y} - a\bar{x} \end{aligned} \tag{2.2.21}$$

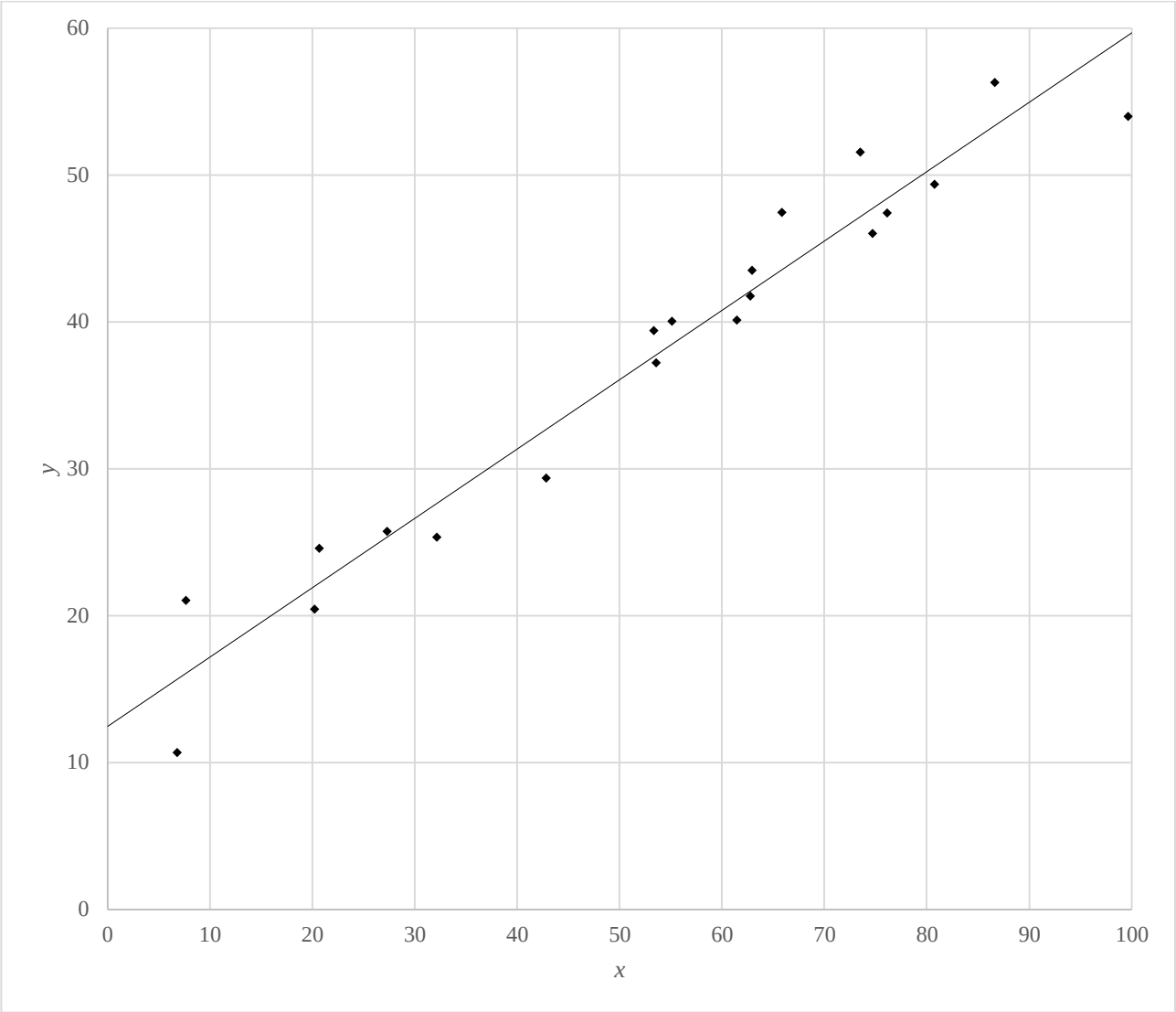
U specijalnom slučaju kada je $\mu = 1$, tj. kada su varijanse σ_ε i σ_η identične izrazi (2.2.21) će postati

$$\begin{aligned} a &= \frac{s_{yy} - s_{xx} + \sqrt{\left(s_{yy} - s_{xx}\right)^2 - 4s_{xy}^2}}{2s_{xy}} \\ b &= \bar{y} - a\bar{x} \end{aligned} \tag{2.2.22}$$

i odnosiće se na ortogonalne devijacije tačaka u odnosu na liniju fita.

Primenjeno na podatke iz **Tabele 1** dobijamo

$$a = 0.47216 \pm 0.02625 \quad b = 12.45088 \pm 1.54939 \tag{2.2.23}$$



Slika 13. Grafički prikaz fita dobijenog Demingovom regresijom

3. Podaci sa poznatim neodređenostima za jednu promenljivu

Ako prilikom fitovanja želimo da uzmemo u obzir neodređenosti po jednoj osi, to ćemo postići tako što ćemo svaku tačku otežati njenom neodređenošću (σ_y kolona u **Tabeli 1**) pri sumiranju vertikalnih udaljenosti

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - ax_i - b}{\sigma_i} \right)^2 \tag{3.1.1}$$

Dakle, što je neodređenost za neku tačku veća, to će ta tačka imati manji doprinos sumi. To znači da preciznije određene tačke igraju veću ulogu u nalaženju fita.

Parcijalni izvodi po a i b u minimumu iznose 0.

$$\begin{aligned} \frac{\partial \chi^2}{\partial a} &= 0 = 2\sum_{i=1}^{N} \frac{x_i \left(y_i - ax_i - b \right)}{\sigma_i^2} \\ \frac{\partial \chi^2}{\partial b} &= 0 = 2\sum_{i=1}^{N} \frac{y_i - ax_i - b}{\sigma_i^2} \end{aligned} \tag{3.1.2}$$

Ako definišemo sledeće sume

$$\begin{aligned} S &= \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \qquad S_x = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \qquad S_y = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \\ S_{xx} &= \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \qquad S_{xy} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \end{aligned} \tag{3.1.3}$$

možemo napisati izraze (3.1.2) na sledeći način

$$\begin{aligned} S_{xy} &= aS_{xx} + bS_x \\ S_y &= aS_x + bS \end{aligned} \tag{3.1.4}$$

uvodimo smenu

$$D = SS_{xx} - \left(S_x \right)^2 \tag{3.1.5}$$

pa je rešenje ovog sistema jednačina

$$a = \frac{SS_{xy} - S_x S_y}{D} \tag{3.1.6}$$

$$b = \frac{S_{xx} S_y - S_x S_{xy}}{D} \tag{3.1.7}$$

Neodređenosti za *a* i *b* možemo odrediti na sledeći način: Ako su podaci nezavisni, onda svaki pojedinačan podatak doprinosi neodređenosti parametara *a* i *b*. Iz propagacije grešaka imamo

$$\sigma_a^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a}{\partial y_i} \right)^2 \tag{3.1.8}$$

$$\sigma_b^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial b}{\partial y_i} \right)^2 \tag{3.1.9}$$

Izvodi *a* i *b* po y_i su

$$\frac{\partial a}{\partial y_i} = \frac{Sx_i - S_x}{\sigma_i^2 D} \tag{3.1.10}$$

$$\frac{\partial b}{\partial y_i} = \frac{S_{xx} - S_x x_i}{\sigma_i^2 D} \tag{3.1.11}$$

Ako uvrstimo izvode (3.1.10) i (3.1.11) u jednačine (3.1.8) i (3.1.9) dobijamo

$$\sigma_a^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{Sx_i - S_x}{\sigma_i^2 D} \right)^2 \tag{3.1.12}$$

$$\sigma_b^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{S_{xx} - S_x x_i}{\sigma_i^2 D} \right)^2 \tag{3.1.13}$$

Posle kratkog sređivanja dobijamo

$$\sigma_a^2 = \sum_{i=1}^{N} \frac{S^2 x_i^2 - 2Sx_i S_x + \left(S_x\right)^2}{\sigma_i^2 D^2} = \frac{S^2 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - 2SS_x \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + \left(S_x\right)^2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2}}{D^2} \tag{3.1.14}$$

$$\sigma_b^2 = \sum_{i=1}^{N} \frac{\left(S_{xx}\right)^2 - 2S_{xx}S_x x_i + \left(S_x\right)^2 x_i^2}{\sigma_i^2 D^2} = \frac{\left(S_{xx}\right)^2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} - 2S_{xx}S_x \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + \left(S_x\right)^2 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}}{D^2} \tag{3.1.15}$$

$$\sigma_a^2 = \frac{S^2 S_{xx} - 2S\left(S_x\right)^2 + S\left(S_x\right)^2}{D^2} = \frac{SD}{D^2} \tag{3.1.16}$$

$$\sigma_b^2 = \frac{\left(S_{xx}\right)^2 S - 2S_{xx}\left(S_x\right)^2 + S_{xx}\left(S_x\right)^2}{D^2} = \frac{S_{xx}D}{D^2} \tag{3.1.17}$$

$$\begin{aligned} \sigma_a^2 &= \frac{S}{D} \\ \sigma_b^2 &= \frac{S_{xx}}{D} \end{aligned} \tag{3.1.18}$$

Jednačine (3.1.18) predstavljaju varijanse procenjenih parametara *a* i *b*. Za korelacioni koeficijent *r* potrebna je još i kovarijansa Cov(*a*,*b*) koja se dobija iz propagacije grešaka na sličan način kao i varijanse.

$$\mathrm{Cov}\left(a,b\right) = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a}{\partial y_i}\right)\left(\frac{\partial b}{\partial y_i}\right) \tag{3.1.19}$$

Ako uvrstimo (3.1.10) i (3.1.11) u (3.1.19) dobijamo

$$\mathrm{Cov}\left(a,b\right) = -\frac{S_x}{D} \tag{3.1.20}$$

Sada možemo izraziti korelacioni koeficijent kao

$$r = \frac{\mathrm{Cov}\,(a,b)}{\sigma_a \sigma_b} \tag{3.1.21}$$

$$r = -\frac{S_x}{\sqrt{SS_{xx}}} \tag{3.1.22}$$

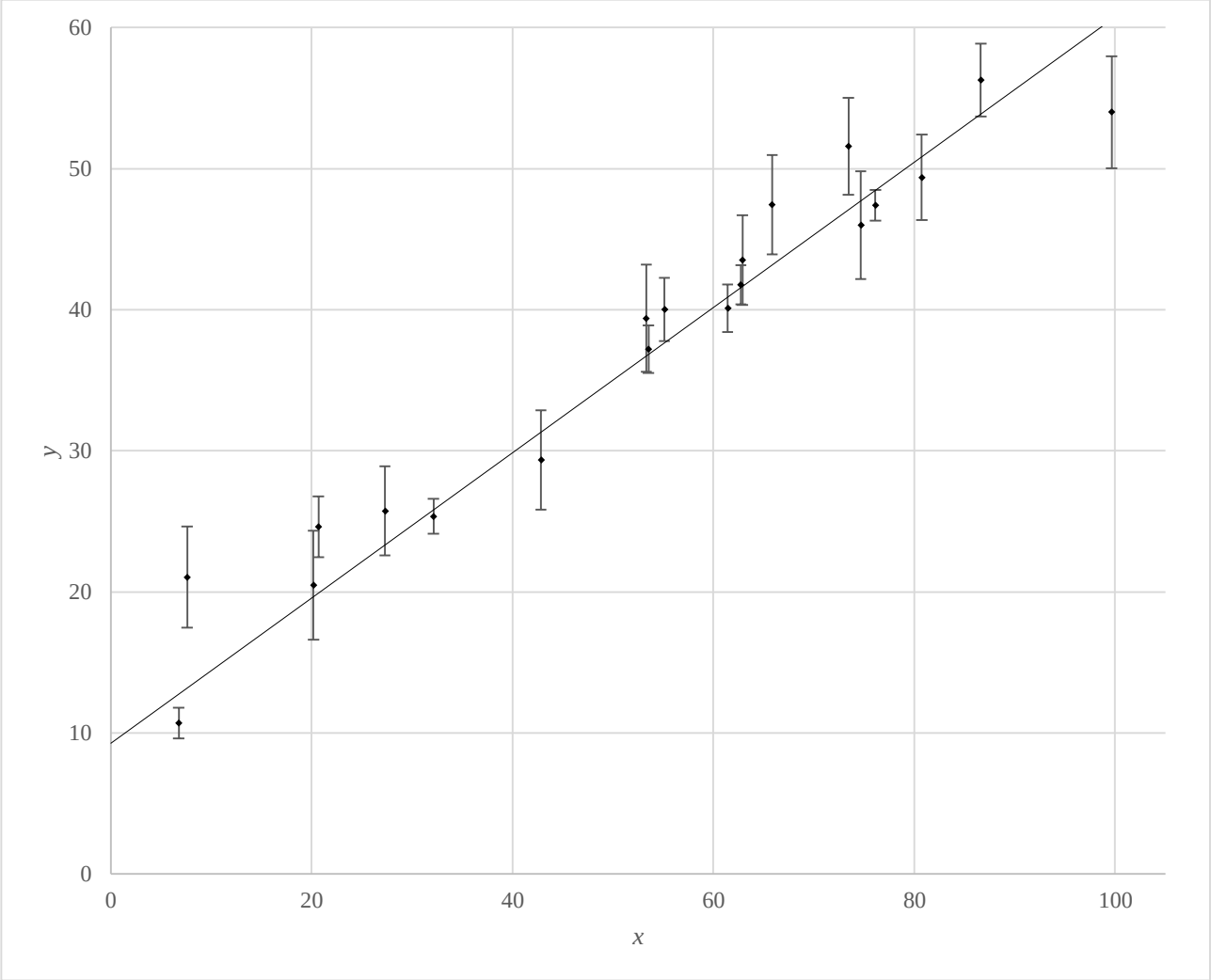
U našem slučaju imamo

$$S = 5.039 \quad S_x = 242.203 \quad S_y = 171.398 \quad S_{xx} = 15140.817 \quad S_{xy} = 10038.873$$

$$\begin{aligned} a &= 0.51457 \pm 0.01976 \\ b &= 9.28087 \pm 1.08289 \end{aligned} \tag{3.1.23}$$

$$r_{ab} = -0.8769$$

Negativna vrednost r_{ab} ukazuje da greške kod *a* i *b* verovatno imaju suprotan predznak (ako *a* raste, *b* opada i obrnuto).



Slika 14. Grafički prikaz procenjenog fita kada podatke otežamo greškama po zavisnoj promenljivoj

4. Podaci sa poznatim neodređenostima za obe promenljive

Iako je najčešće korišćena regresija *y* od *x*, u praksi su veoma retko *x* vrednosti bez grešaka. Najčešće su i *x* i *y* greške značajne i variraju od tačke do tačke. U ovom poglavlju će biti opisan metod ukupnih najmanjih kvadrata koji uzima u obzir greške u obe promenljive. U ovom slučaju suma koju treba minimalizovati ima sledeći oblik:

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{(y_i - ax_i - b)^2}{\sigma_{yi}^2 + a^2\sigma_{xi}^2} \tag{4.1.1}$$

Izvod po b je linearan ali izvod po a nije, tako da se ne može rešiti analitički kao u prethodnim metodama. Rešenje se nalazi iterativno numeričkim metodama. S obzirom na kompleksnost nalaženja optimalnih parametara iterativnim numeričkim metodama, postala je uobičajena praksa ili da se neodređenosti potpuno zanemare ili da se uzmu u obzir neodređenosti jedne promenljive (tipično promenljive sa većim neodređenostima).

Ovde ćemo dati jednu od mnoštva postojećih procedura za nalaženje parametara najboljeg fita za podatke sa neodređenostima u obe promenljive. Postupak koji sledi je opisan u York et al (2004).

Izabere se početno *a* (npr. iz metoda najmanjih kvadrata iz Poglavlja 3.)

Odrede se tegovi (eng. weight) za svaku tačku ω_{xi} i ω_{yi}:

$$\begin{aligned} \omega_{xi} &= \frac{1}{\sigma_{xi}^2} \\ \omega_{yi} &= \frac{1}{\sigma_{yi}^2} \end{aligned} \tag{4.1.2}$$

Sada, pomoću korelacionog koeficijenta r_i (ako je poznat), tegova ω_{xi} i ω_{yi} i početne pretpostavke za *a* se oforme izrazi

$$W_i = \frac{\omega_{xi}\omega_{yi}}{\omega_{xi} + a^2\omega_{yi} - 2ar_i\alpha_i} \tag{4.1.3}$$

za svaku tačku, gde je

$$\alpha_i = \sqrt{\omega_{xi}\omega_{yi}} \tag{4.1.4}$$

Sada se može izračunati

$$\bar{x} = \sum_{i=1}^{N} \frac{W_i x_i}{W_i}$$
$$\bar{y} = \sum_{i=1}^{N} \frac{W_i y_i}{W_i} \tag{4.1.5}$$

Odavde se odredi u_i i v_i

$$u_i = x_i - \bar{x}$$
$$v_i = y_i - \bar{y} \tag{4.1.6}$$

Na kraju se odredi β_i iz

$$\beta_i = W_i \left(\frac{u_i}{\omega_{yi}} + \frac{a v_i}{\omega_{xi}} - \left(a u_i + v_i \right) \frac{r_i}{\alpha_i} \right) \tag{4.1.7}$$

Koristeći W_i, β_i, u_i i v_i dobija se poboljšan parametar a

$$a = \frac{\sum_{i=1}^{N} W_i \beta_i v_i}{\sum_{i=1}^{N} W_i \beta_i u_i} \tag{4.1.8}$$

Ovako dobijen parametar *a* se ponovo uvrsti u jednačine (4.1.3) - (4.1.8) dok ne postignemo sukcesivne rezultate za *a* unutar željene tolerancije.

Presek sa *y* osom se dobija iz sledećeg izraza:

$$b = \bar{y} - a\bar{x} \tag{4.1.9}$$

Za određivanje neodređenosti dobijenih parametara *a* i *b* koristi se sledeća procedura:

Izračunaju se popravljene vrednosti X_i

$$X_i = \bar{x} + \beta_i \tag{4.1.10}$$

Pomoću (4.1.10) izrazimo

$$\bar{X} = \frac{\sum_{i=1}^{N} W_i X_i}{\sum_{i=1}^{N} W_i} \tag{4.1.11}$$

$$U_i = X_i - \bar{X} \tag{4.1.12}$$

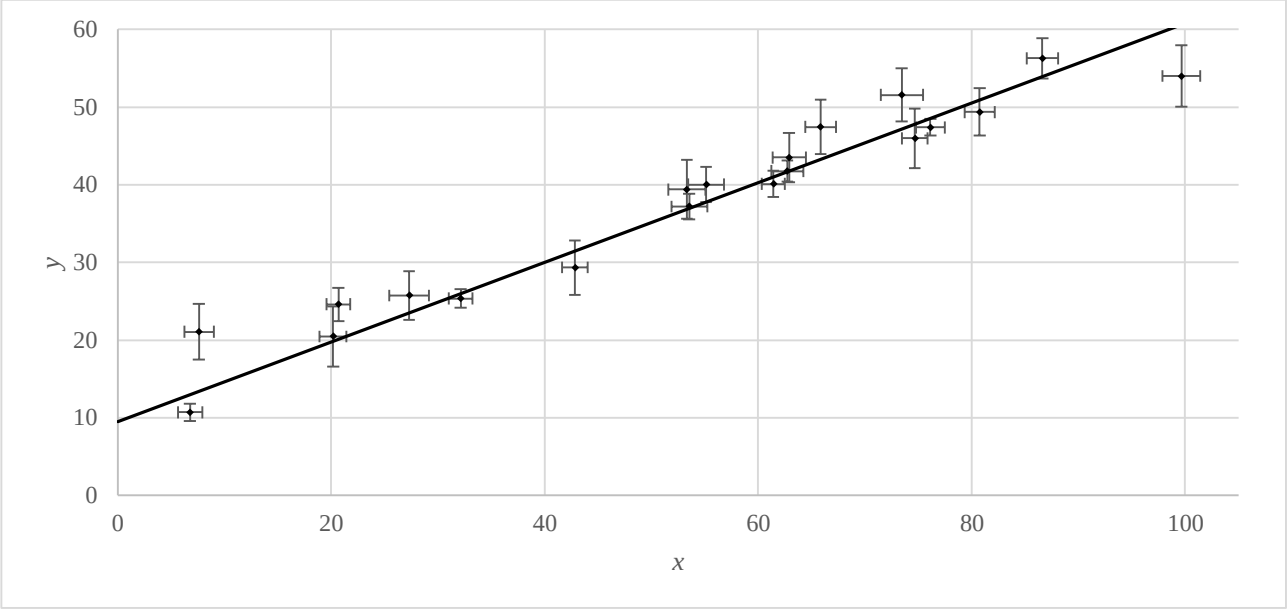
Konačno se neodređenosti dobijaju iz izraza

$$\begin{aligned} \sigma_a^2 &= \frac{1}{\sum_{i=1}^{N} W_i U_i^2} \\ \sigma_b^2 &= \frac{1}{\sum_{i=1}^{N} W_i} + \bar{X}^2 \sigma_a^2 \end{aligned} \tag{4.1.13}$$

Opisani algoritam ne garantuje konvergenciju za svaki set podataka, ali u praksi se pokazalo da u većini slučajeva iterativni postupak konvergira prilično brzo (oko 10 iteracija). Ovaj postupak je simetričan, tj. daće isti rezultat ako zamenimo *x* i *y*.

Procedura opisana u ovom poglavlju primenjena na podatke iz **Tabele 1.** daje sledeće parametre

$$\begin{aligned} a &= 0.51304 \pm 0.02065 \\ b &= 9.46078 \pm 1.12773 \end{aligned} \tag{4.1.14}$$



Slika 15. Grafički prikaz fita dobijenog procedurom koja uzima u obzir neodređenosti kod obe promenljive.

5. Podaci sa asimetričnim neodređenostima

U slučaju asimetričnih neodređenosti ne postoji ustaljen metod fitovanja. Rešenja na koja se nailazi u literaturi se obično svode na simetrizovanje neodređenosti uzimanjem veće od dve neodređenosti ili računanjem srednje vrednosti neodređenosti.

Čak i kod običnog sabiranja dva rezultata sa asimetričnim greškama koriste se procedure bez osnova u teoriji (pozitivne i negativne neodređenosti se zasebno dodaju u kvadraturi).

Postoji nekoliko uzroka asimetričnih grešaka. One mogu nastati kada χ^2 oko minimuma ili kriva maksimuma verovatnoće nisu simetrične, tj ne mogu se dobro aproksimirati parabolom, kao i kada y zavisi nelinearno od x unutar nekoliko standardnih devijacija oko očekivane vrednosti.

Najveći problem kod ovakvih podataka je što nije jasno šta predstavljeni rezultati znače. Ako pretpostavimo da greške predstavljaju standardni interval pouzdanosti, tj da 68.3% vrednosti se nalazi unutar ovog intervala, i dalje nije jasno da li prijavljena vrednost predstavlja srednju vrednost, modu (najverovatnija vrednost), ili pak nešto treće.

Roger Barlow (2003) predlaže nekoliko modela za aproksimaciju raspodele sa asimetričnim neodređenostima kao pokušaj pružanja doslednog metoda za analizu ovakvih podataka. Ovakvo rešenje se čini podjednako neosnovano kao i simetrizacija ili sabiranje u kvadraturi.

G. D. Agostini (2004) pruža probabilističko rešenje problema u tačnoj i aproksimiranoj formi.

6. Rezultati

U **Tabeli 2** su predstavljeni izračunati parametri fitova prethodno opisanih procedura primenjenih na podatke iz **Tabele 1**. Prva kolona se odnosi na MNK za slučaj kada je x nezavisna promenljiva a svi y_i imaju identične nepoznate neodređenosti iz normalne raspodele. Druga kolona se odnosi na suprotan slučaj, tj. sada je y nezavisna promenljiva a x zavisna promenljiva. Treća kolona je Demingova regresija za slučaj kada su varijanse promenljivih identične. Četvrta kolona je MNK sa poznatim neodređenostima σ_y u y koordinati, i peta kolona je procedura opisana u poglavlju 4 za podatke sa neodređenostima u obema koordinatama. Prva dva reda su parametri a i b, tj. nagib i presek sa y osom. treći i četvrti red su standardne greške procenjenih parametara. Peti red su srednje vrednosti 10^5 izračunatih odnosa vrednosti y očitanih sa fita i pravih vrednosti $\hat{y}$ za nasumično odabrano $\hat{x}$.

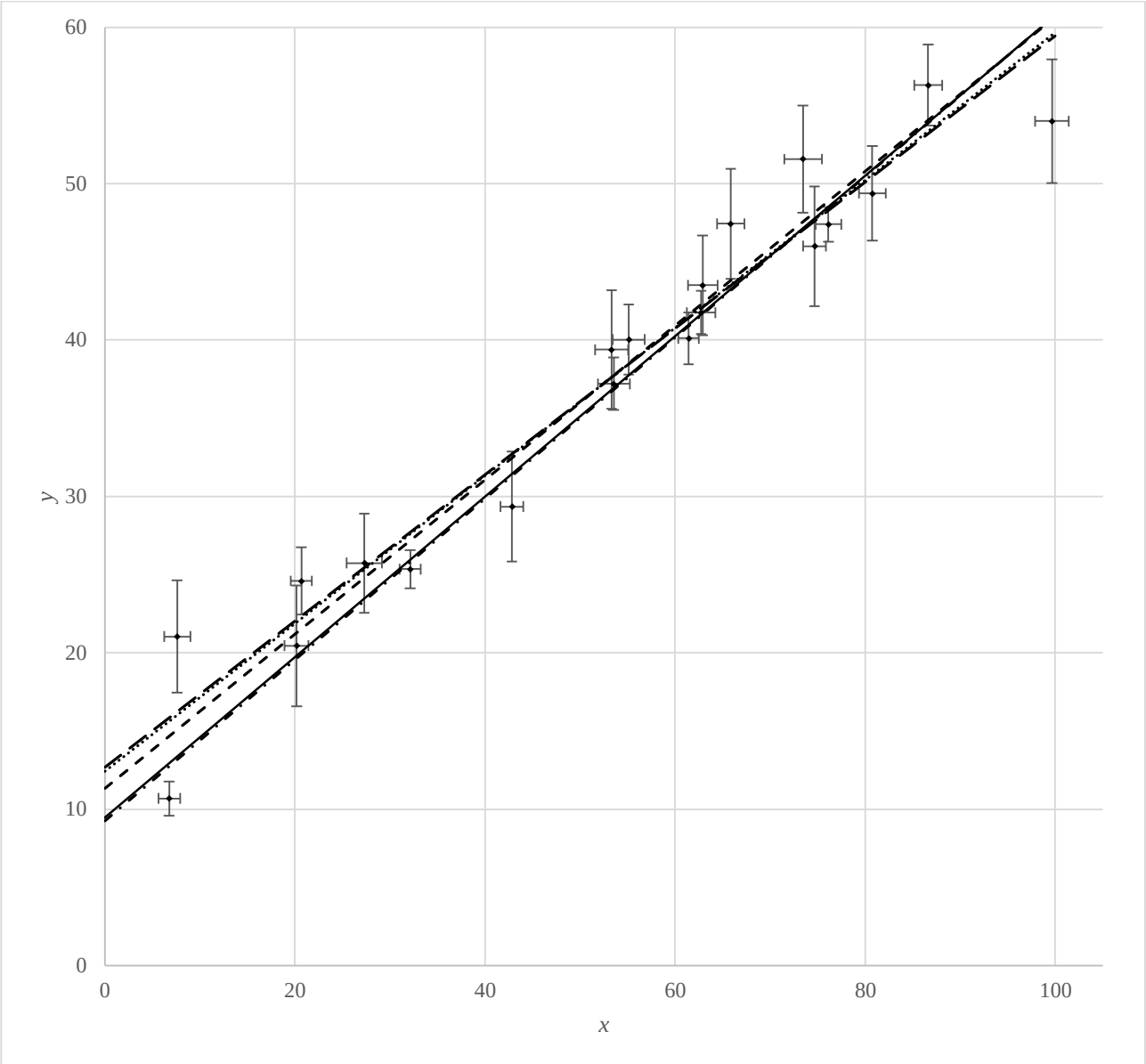
Tabela 2. Sumirani rezultati prethodno opisanih metoda primenjenih na podatke iz **Tabele 1.**

	MNK $y = f(x)$	MNK $x = f(y)$	Deming $\mu = 1$	MNK (σ_y)	MNK (σ_x, σ_y)
a	0.4675	0.4933	0.4721	0.51457	0.51304
b	12.699	11.327	12.451	9.28087	9.46078
σ_a	0.0259	0.0266	0.0263	0.01976	0.02065
σ_b	1.532	1.573	1.550	1.08289	1.12173
$y / \hat{y}$	1.055	1.039	1.052	0.993	0.997

Na **Slici 16** su grafički predstavljeni rezultati prethodno opisanih metoda linearnog fitovanja. Ako uporedimo fitove sa našim generativnim modelom, tj. ako se setimo da je linearna zavisnost koja stvara naše podatke $y = 0.5x + 10$, možemo zaključiti da metode koje ne uzimaju u obzir neodređenosti precenjuju vrednost parametra b za dati set podataka**.** Metode koje uzimaju u obzir neodređenosti za jednu ili obe promenljive bolje procenjuju presek sa y osom ali za dati set podataka precenjuju nagib. Razlog je što uzimanjem u obzir neodređenosti više otežavamo tačke koje manje odstupaju od izabrane zavisnosti. Prve dve kolone možemo posmatrati kao granice mogućih parametara a i b za Demingovu regresiju sa proizvoljnim μ. S obzirom da ne uzimaju greške u obzir, vidimo da potcenjuju nagib izabrane zavisnosti, gde prvi metod najviše odstupa.

Naravno ovakvi zaključci se nipošto ne mogu uopštiti već samo pomažu da bolje shvatimo uticaj grešaka na procenu parametara.

Metode iz trećeg i četvrtog poglavlja daju veoma slične parametre a i b. Uzrok leži u činjenici da su neodređenosti za x manje od neodređenosti za y, pa se metoda iz četvrtog poglavlja ponaša slično metodi koja uzima u obzir samo greške za jednu promenljivu. Ovo može poslužiti kao opravdanje za uzimanje u obzir grešaka za promenljivu sa većim neodređenostima. Ipak, svakako je poželjnije iskoristiti sva znanja koja imamo o podacima prilikom analize.



Slika 16. Grafički prikaz procenjenih parametara svih prethodno opisanih metoda. – – MNK y = f(x); — — MNK x = f(y); ····· Demingova regresija (ortogonalni fit); – · – MNK (σ_y); —— MNK (σ_x,σ_y).

7. Zaključak

U uvodnom poglavlju smo se upoznali sa istorijom nastanka MNK, kao i sa osnovnim pojmovima koji se tiču analize setova podataka. U drugom, trećem i četvrtom poglavlju su opisane sledeće metode fitovanja različitih tipova podataka:

- Metod najmanjih kvadrata za podatke bez poznatih neodređenosti
- Demingova regresija (ortogonalni fit) za podatke bez poznatih neodređenosti
- Otežani metod najmanjih kvadrata za podatke sa poznatim neodređenostima za jednu promenljivu
- Metod ukupnih najmanjih kvadrata za podatke sa poznatim neodređenostima za obe promenljive

Data su izvođenja za jednostavnije metode kao i izrazi za standardne greške prilikom procene parametara fita. Radi demonstracije, sve metode su primenjene na nasumično generisanom setu podataka.

Peto poglavlje se ukratko osvrće na problem podataka sa asimetričnim neodređenostima i upućuje čitaoca na detaljniju literaturu.

Treba napomenuti da se svođenjem nekog seta podataka na samo dva parametra, nagib i presek sa ordinatom, gube informacije o datom setu podataka. Alternativni pristup analizi slabo korelisanih podataka može se naći u Vukotić et al (2014), gde se ponavljanjem nasumičnog uzorkovanja rekonstruiše funkcija gustine verovatnoće, pa se za datu vrednost jedne promenljive, umesto čitanja jedne vrednosti sa fita, dobija raspodela koja se dalje može analizirati. U planu je proširenje pomenutog metoda u vidu uzimanja neodređenosti podataka u obzir.

Predlog za nastavak rada bi bio proširivanje modela za generisanje podataka tako da sa određenom učestalošću generiše podatke koji znatno odstupaju od izabrane relacije (eng. outlier points) radi provere osetljivosti različitih metoda fitovanja.

8. Literatura

1. William M. Bolstad (2007): *Introduction to bayesian statistics – second edition.* New Jersey: John Wiley & Sons
2. S. M. Stigler (1981): *Gauss and the invention of least squares.* The annals of Statistics, Vol. 9, No. 3, 465-474
3. S. M. Stigler (1986): *The history of statistics: The measurement of uncertainty before 1900*, The Belknap press of Harvard University Press, Cambridge, Massachusetts, and London, England
4. D. York, N. M. Evensen, M. Lopez Martinez, J. De Basabe Delgado (2004): *Unified equations for the slope, intercept, and standard errors of the best straight line*, Am. J. Phys. 72, 367
5. R. Barlow (2003): *Asymetric systematic errors*, arXiv:physics/0306138v1
6. William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery (2007): *Numerical recipes: The art of scientific computing – Third edition*. Cambridge University Press
7. T. Isobe, E. D. Feigelson, M. G. Akritas, G. J. Babu (1990): *Linear regression in astronomy*, The Astrophysical Journal, 364: 104-113
8. G. D. Agostini, M. Raso (2000): *Uncertainties due to imperfect knowledge of systematic effects: general considerations and approximate formulae*, arXiv:hep-ex/0002056v1
9. S. R. Scuro (2004): *Introduction to error theory*, Visual Physics Laboratory, Texas A&M University, College Station, TX 77843
10. D. W. Hogg (2010): *Data analysis recipes: Fitting a model to data*, arXiv:1008.4686v1
11. B. Vukotic, M. Jurkovic, D. Uroševic, B. Arbutina (2014): *On calibration of some distance scales in astrophysics*, MNRAS 440, 2026-2035
12. S. J. Miller (2006): *The method of least squares*, Mathematics Department, Brown University, Providence
13. L. Lyons (1991): *A practical guide to Data analysis for Physical Science Students*, Cambridge University Press
14. P. Glaister (March 2001): *Least squares revisited*, The Mathematical Gazette 85: 104-107

9. Biografija

Srđan Šibalić, rođen 10.09.1986. godine u Vrbasu, gde je završio osnovnu školu „Petar I Petrović Njegoš“ , a zatim gimnaziju „Žarko Zrenjanin“. Prirodno-matematički fakultet u Novom Sadu upisuje 2005. godine, smer astronomija sa astrofizikom.

U toku svojih studija imao je priliku da učestvuje na studentskim praksama u Češkoj (Ondřejov opservatorija, Prag, 2010) i Srbiji (Vidojevica, Astronomska opservatorija Beograd, 2014). Aktivan član astronomskog društva „Novi Sad“ od 2006. godine.

UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj: **RBR**	
Identifikacioni broj: **IBR**	
Tip dokumentacije: **TD**	Monografska dokumentacija
Tip zapisa: **TZ**	Tekstualni štampani materijal
Vrsta rada: **VR**	Završni rad
Autor: **AU**	Srđan Šibalić
Mentor: **MN**	Prof. dr Tijana Prodanović
Naslov rada: **NR**	Uticaj i tretman grešaka prilikom fitovanja podataka sa linearnom zavisnošću
Jezik publikacije: **JP**	srpski (latinica)
Jezik izvoda: **JI**	srpski/engleski
Zemlja publikovanja: **ZP**	Srbija
Uže geografsko područje: **UGP**	Vojvodina
Godina: **GO**	2016
Izdavač: **IZ**	Autorski reprint
Mesto i adresa: **MA**	Prirodno-matematički fakultet, Trg Dositeja Obradovića 4, Novi Sad
Fizički opis rada: **FO**	Broj poglavlja/strana/lit. citata/tabela/slika/ 7/35/14/2/16
Naučna oblast: **NO**	Fizika
Naučna disciplina: **ND**	Astronomija
Predmetna odrednica/ ključne reči: **PO** **UDK**	Fitovanje, linearna regresija, metod najmanjih kvadrata, Demingova regresija, ortogonalni fit, metod ukupnih najmanjih kvadrata
Čuva se: **ČU**	Biblioteka departmana za fiziku, PMF-a u Novom Sadu
Važna napomena: **VN**	nema
Izvod: **IZ**	Dat je pregled uobičajenih tehnika fitovanja podataka sa i bez poznatih neodređenosti kao i uticaja grešaka posmatranja na liniju najboljeg fita
Datum prihvatanja teme od NN veća: **DP**	05.09.2016.
Datum odbrane: **DO**	26.09.2016
Članovi komisije: **KO**	Prof. dr Dušan Mrđa, prof dr Milan Pantić, dr Branislav Vukotić, prof. dr Tijana Prodanović
Predsednik:	Prof. dr Dušan Mrđa
član:	Prof. dr Milan Pantić
član:	dr Branislav Vukotić, naučni saradnik AOB
član:	Prof. dr Tijana Prodanović

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE AND MATHEMATICS

KEY WORDS DOCUMENTATION

Accession number: **ANO**	
Identification number: **INO**	
Document type: **DT**	Monograph publication
Type of record: **TR**	Textual printed material
Content code: **CC**	Final paper
Author: **AU**	Srđan Šibalić
Mentor/comentor: **MN**	Prof. dr Tijana Prodanović
Title: **TI**	Impact and treatment of errors in fitting linearly dependent data
Language of text: **LT**	Serbian (Latin)
Language of abstract: **LA**	English
Country of publication: **CP**	Serbia
Locality of publication: **LP**	Vojvodina
Publication year: **PY**	2016
Publisher: **PU**	Author's reprint
Publication place: **PP**	Faculty of Science and Mathematics, Trg Dositeja Obradovića 4, Novi Sad
Physical description: ***PD***	No of chapters/pages/references & quotes/tables/pictures/ 7/35/15/2/16
Scientific field: **SF**	Physics,
Scientific discipline: **SD**	Astrofizika
Subject/ Key words: **SKW**	Fitting, linear regression, least squares method, Deming regression, orthogonal fitting, total least squares
UC	
Holding data: **HD**	Library of Department of Physics, Trg Dositeja Obradovića 4
Note: **N**	none
Abstract: **AB**	We give an overview of common data fitting techniques with or without known uncertainties as well as impact of observational errors on the best fit line
Accepted by the Scientific Board: **ASB**	05.09.2016.
Defended on: **DE**	26.09.2016.
Thesis defend board: **DB**	Prof. dr Dušan Mrđa, prof dr Milan Pantić, dr Branislav Vukotić, prof. dr Tijana Prodanović
President:	Prof. dr Dušan Mrđa
Member:	Prof. dr Milan Pantić
Member:	dr Branislav Vukotić, RA at Belgrade Astronomical Observatory
Member:	Prof. dr Tijana Prodanović