Linearna regresija i primjene

Obšivač, Veronika

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Veronika Obšivač

LINEARNA REGRESIJA I PRIMJENE

Diplomski rad

Voditelj rada:
doc. dr. sc. Marija Galić

Zagreb, 2024.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Svoju diplomu posvećujem svojim roditeljima, bratu, sestri, djedovima i bakama ♡
Zahvalna sam vam na svemu što ste mi pružili, posebno na neizmjernoj ljubavi i podršci koje su me uvijek gurale naprijed. Zbog vas sam danas tu gdje jesam.

Hvala prijateljima s kojima sam na ovom putu dijelila iste radosti i brige. Zbog vas ću se ovog perioda života uvijek rado prisjećati.

Na kraju, najveće hvala dragom Bogu koji mi je podario sve ove ljude, koji je uvijek znao pravi trenutak i koji me vodio do kraja ovog puta.
Iz 43, 1-4

Sadržaj

Uvod

Linearna regresija statistička je metoda koja omogućava istraživanje i modeliranje odnosa između jedne zavisne (odazovne) i jedne ili više nezavisnih (eksplanatornih) varijabli. Cilj ove metode je izgraditi regresijski model koji najbolje opisuje dani skup podataka. Proučavajući odnos između nezavisne varijable X i zavisne varijable Y, nastojimo donijeti zaključke koji nam omogućuju predviđanje nepoznatih ili budućih vrijednosti zavisne varijable.

Postoje dva tipa linearne regresije: jednostavna i višestruka. Jednostavna linearna regresija proučava utjecaj jedne nezavisne varijable na zavisnu varijablu, dok višestruka regresija proširuje ovaj pristup na više nezavisnih varijabli te model postaje složeniji jer uzima u obzir različite čimbenike koji mogu utjecati na zavisnu varijablu.

Primjene linearne regresije izuzetno su široke i prisutne su u mnogim disciplinama, uključujući inženjerstvo, ekonomiju, biologiju, medicinu, društvene znanosti i mnoge druge. Ova tehnika omogućava istraživačima i analitičarima dublje razumijevanje međusobnih odnosa među varijablama, čime pomaže u donošenju informiranih odluka temeljenih na podacima.

U prvom poglavlju definirat ćemo ključne pojmove koji su nužni za daljnje razumijevanje i razvoj modela linearne regresije. U ovom dijelu upoznat ćemo se s osnovnim statističkim konceptima, kao što su mjere srednjih vrijednosti i mjere raspršenosti, te s osnovnim pojmovima vezanim uz slučajne varijable, koji čine temelj za razumijevanje regresijskih modela.

U drugom poglavlju detaljno ćemo razraditi model linearne regresije. Objasnit ćemo kako se model gradi te kako se procjenjuju njegovi parametri, s posebnim naglaskom na metodu najmanjih kvadrata. Također, razmotrit ćemo ključne pretpostavke koje ovaj model podrazumijeva, te ćemo prikazati primjer linearizacije funkcije koja u svom izvornom obliku nije linearna.

U trećem poglavlju predstavit ćemo matričnu formulaciju problema traženja regresijskog pravca, koja nam omogućava efikasno rješavanje regresijskog modela.

Na kraju, u četvrtom poglavlju, izložit ćemo primjene linearne regresije u stvarnom životu. Razmotrit ćemo konkretne primjere njezine primjene u različitim područjima, kako bismo prikazali njezinu širinu i korisnost u analizi podataka i donošenju odluka.

Poglavlje 1

Osnovni pojmovi

1.1 Mjere srednje vrijednosti

Ako promatramo niz podataka koji se sastoji od manjeg broja elemenata, statistička analiza neće biti teška. No, što je veći broj članova niza, to je teže shvatiti odnos među članovima te se stvara potreba da se umjesto velikog broja članova niza uzme jedan jedini izraz koji će karakterizirati varijacije obilježja članova tog niza. Na primjer, želimo li usporediti primanja zaposlenih u dvije ili više firmi, takva bi usporedba pomoću distribucije zaposlenih prema njihovim primanjima bila teška, ali ako bismo za svaku firmu izračunali prosječna primanja zaposlenih, onda bi ta usporedba bila znatno jednostavnija.

Ponekad je srednja vrijednost dovoljna informacija da zaključimo ono što nas zanima o našem nizu podataka. Srednja se vrijednost može utvrditi na više načina pa postoji više vrsta srednjih vrijednosti od kojih svaka ima određena svojstva i određen sadržaj. U nastavku ćemo prikazati neke od njih.

Definicija 1.1.1. ***Aritmetička sredina*** *uzorka $x_1, x_2, \ldots, x_n$ definira se kao količnik sume podataka i ukupnog broja podataka odnosno*

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{1.1}$$

Aritmetička sredina je najčešće korištena srednja vrijednost.

Definicija 1.1.2. ***Medijan*** *je položajna srednja vrijednost koja numerički niz podataka uređen po veličini dijeli na dva jednakobrojna dijela.*

Ako je broj podataka neparan, medijan je vrijednost središnjeg podatka, a ako je broj podataka paran medijan predstavlja srednju vrijednost dva središnja podatka. Odnosno,

ako medijan označimo s m, za uzorak $x_1, x_2, \ldots, x_n$ tada vrijedi:

$$m = \begin{cases} \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & \text{za } n \text{ paran}, \\ x_{\frac{n+1}{2}}, & \text{za } n \text{ neparan}. \end{cases} \tag{1.2}$$

U oba slučaja vidimo da je u našem uzorku n podataka veće od medijana i n podataka manje od medijana.

1.2 Mjere raspršenosti

Ako želimo znati koliko dobro prosjek opisuje naš niz podataka, odredit ćemo raspršenost uzorka koji promatramo. Kao i kod prosjeka, postoji više načina kako odrediti raspršenost pa ćemo navesti neke od njih koje će nam koristiti u ovom radu.

Da bi se lakše shvatio smisao i potreba definiranja mjera raspršenosti, prvo ćemo pokazati dva jednostavna slučaja. Zamislimo da su mjerenjem veličine X dobiveni rezultati:

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5,$$

a mjerenjem veličine Y rezultati:

$$y_1 = y_2 = y_3 = y_4 = y_5 = 3.$$

Iako ova dva niza imaju jednake aritmetičke sredine ($\bar{x} = \bar{y} = 3$) i jednake medijane ($m_x = m_y = 3$), radi se o dva različita niza te vidimo da je u prvom nizu očita raspršenost podataka, dok u drugom nizu uopće nema raspršenosti podataka. Stoga je potrebno definirati parametar koji će mjeriti raspršenost podataka. Te se mjere zovu *mjere raspršenosti* ili *disperzije*. Navedimo neke od njih.

Definicija 1.2.1. ***Varijanca*** *uzorka $x_1, \ldots, x_n$ definirana je kao*

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.3}$$

Vidimo da je varijanca zapravo aritmetička sredina kvadrata odstupanja vrijednosti obilježja od aritmetičke sredine.

Definicija 1.2.2. ***Standardna devijacija*** *je srednje kvadratno odstupanje podataka od njihove aritmetičke sredine tj.*

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{1.4}$$

Vidimo da je standardna devijacija definirana kao drugi korijen iz varijance.

1.3 Slučajna varijabla

U ovom ćemo poglavlju definirati osnovne pojmove teorije vjerojatnosti i detaljno se upoznati s njima. Ovi pojmovi predstavljaju temelj za matematičku analizu podataka s kojima radimo jer nam omogućuju da ih precizno obradimo i interpretiramo. Razumijevanje vjerojatnosnog prostora bit će ključno za daljnje modeliranje podataka, posebno u kontekstu primjene linearne regresije, kojom ćemo se baviti nešto kasnije. Sljedeće su definicije preuzete iz ([13]).

Skup svih ishoda nekog slučajnog pokusa označavamo s Ω i zovemo ga prostor elementarnih događaja. Podskup $A \subseteq \Omega$ zovemo događaj.

Definicija 1.3.1. *Familija $\mathcal{F}$ podskupova od Ω je σ-algebra na Ω ako je:*

(F1) $\emptyset \in \mathcal{F}$,

(F1) $A \in \mathcal{F} \Rightarrow A^c = \Omega \setminus A \in \mathcal{F}$,

(F1) $A_n \in \mathcal{F}, n \in \mathbb{N} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Definicija 1.3.2. *Vjerojatnost je funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ sa sljedećim svojstvima:*

(P1) $0 \leq \mathbb{P}(A) \leq 1$, *za svaki* $A \in \mathcal{F}$,

(P2) $\mathbb{P}(\Omega) = 1$,

(P3) Ako su $A_n \in \mathcal{F}, n \in \mathbb{N}$ međusobno disjunktni, tada je

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

*Uređenu trojku $(\Omega, \mathcal{F}, \mathbb{P})$ zovemo **vjerojatnosni prostor**.*

Definicija 1.3.3. ***Slučajna varijabla** je funkcija koja svakom elementarnom događaju pridružuje realan broj. Označavamo ih velikim slovima abecede: $X, Y, Z, \ldots$*

Svaka slučajna varijabla može biti diskretna ili neprekidna (kontinuirana), ovisno o tome može li poprimiti konačno ili beskonačno mnogo vrijednosti. U nastavku ćemo ih preciznije definirati.

Definicija 1.3.4. *Funkciju $X : \Omega \to \mathbb{R}$ koja poprima najviše prebrojivo mnogo različitih vrijednosti zovemo **diskretna slučajna varijabla**.*

Za $a \in \mathbb{R}$, promatrat ćemo vjerojatnost da slučajna varijabla X poprimi vjerojatnost a, te ćemo to zapisivati $\{X = a\}$.

Definicija 1.3.5. *Ako diskretna slučajna varijabla X može poprimiti vrijednosti $a_1, a_2, \ldots$ i to s vjerojatnostima $p_i = \mathbb{P}(X = a_i)$, $i = 1, 2, \ldots$, brojeve $a_1, a_2, \ldots$ i pripadne vjerojatnosti $p_1, p_2, \ldots$ zovemo* ***distribucijom (ili razdiobom) slučajne varijable*** *X i pišemo*

$$X \sim \begin{pmatrix} a_1 & a_2 & \ldots \\ p_1 & p_2 & \ldots \end{pmatrix}.$$

Uočimo, nužno je $p_i \geq 0$ za sve $i = 1, 2, \ldots$ i $\sum_{i=1}^{\infty} p_i = 1$.

Definicija 1.3.6. *Neka je X slučajna varijabla s distribucijom*

$$X \sim \begin{pmatrix} a_1 & a_2 & \ldots \\ p_1 & p_2 & \ldots \end{pmatrix}.$$

Matematičko očekivanje *od X je broj*

$$\mathbb{E}[X] := \sum_{n=1}^{\infty} a_n p_n$$

ako taj red apsolutno konvergira (tj. ako $\sum_{n=1}^{\infty} |a_n| p_n < \infty$). U suprotnom kažemo da X nema očekivanje.

Definicija 1.3.7. *Ako slučajna varijabla X ima očekivanje, varijanca od X je*

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Varijanca mjeri prosječno odstupanje slučajne varijable od svoje očekivane vrijednosti. Napomena: Varijancu najčešće računamo kao $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Definicija 1.3.8. *Neka su X i Y diskretne slučajne varijable (definirane na istom vjerojatnosnom prostoru) koje poprimaju vrijednosti $(a_i)_{i \in \mathbb{N}}$, odnosno $(b_j)_{j \in \mathbb{N}}$. Kažemo da su X i Y* ***nezavisne*** *ako vrijedi*

$$\mathbb{P}(X = a_i, Y = b_j) = \mathbb{P}(X = a_i)\mathbb{P}(Y = b_j), \quad \text{za sve } i, j \in \mathbb{N},$$

pri čemu je $\{X = a_i, Y = b_j\}$ oznaka za događaj $\{X = a_i\} \cap \{Y = b_j\}$.

Definicija 1.3.9. *Slučajna varijabla X na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ je* ***neprekidna slučajna varijabla*** *ako postoji funkcija $f : \mathbb{R} \to [0, \infty)$ takva da je*

$$\mathbb{P}(X \leq a) = \int_{-\infty}^{a} f(t)\, dt, \quad \forall a \in \mathbb{R}.$$

Funkciju f zovemo funkcijom gustoće slučajne varijable X.

Definicija 1.3.10. *Slučajna varijabla X ima* ***normalnu distribuciju*** *s parametrima $\mu \in \mathbb{R}$ i $\sigma^2 > 0$ ako joj je funkcija gustoće dana sa*

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \quad \forall t \in \mathbb{R}.$$

Tada pišemo $X \sim N(\mu, \sigma^2)$.

Napomena: Ako je $X \sim N(\mu, \sigma^2)$, onda je $\mathbb{E}[X] = \mu$ i $\mathrm{Var}(X) = \sigma^2$. Broj $\sigma = \sqrt{\mathrm{Var}(X)}$ nazivamo standardna devijacija.

Definirajmo još mjeru zajedničke varijacije dviju slučajnih varijabli.

Definicija 1.3.11. *Neka su X i Y slučajne varijable takve da je* $\mathrm{Var}(X)$, $\mathrm{Var}(Y) < \infty$. *Kovarijanca od X i Y je* $\mathrm{Cov}(X, Y) := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Drugim riječima, kovarijanca pokazuje kako se dvije slučajne varijable mijenjaju u odnosu jedna na drugu. Ako je kovarijanca pozitivna, to znači da ako raste X, tada raste i Y. Ako je kovarijanca negativna, to znači da kada jedna varijabla raste, druga pada i obrnuto. Kada je kovarijanca jednaka 0, to znači da nema linearnog odnosa između varijabli.

Poglavlje 2

Linearna regresija

2.1 Model jednostavne linearne regresije

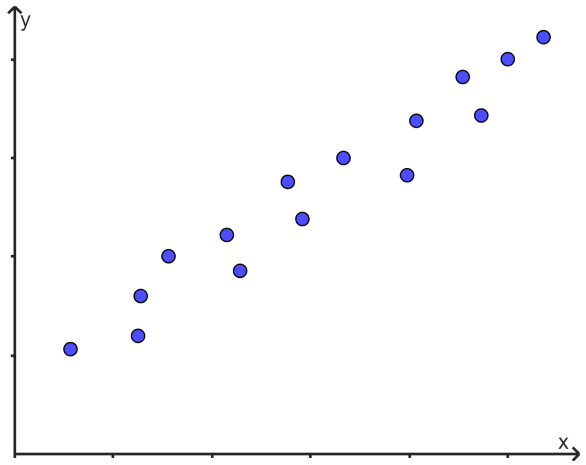
Deterministička veza između dviju varijabli je veza zadana pravilom

$$y = f(x)$$

gdje je x nezavisna varijabla, y zavisna varijabla, a $f : \mathbb{R} \longrightarrow \mathbb{R}$ zadana funkcija.

Pravila $y = 2x + 3$, $y = 5x^2 - 2x$, $y = \log_4(x + 4)$ su primjeri determinističke veze među varijablama x i y jer za svaku dopuštenu vrijednost nezavisne varijable x možemo izračunati točnu vrijednost zavisne varijable y.

U statističkim istraživanjima najčešće ne možemo očekivati determinističke veze jer na zavisnu varijablu mogu utjecati i mnogi drugi čimbenici, stoga tu pričamo o *statističkoj vezi*. To ćemo najbolje uočiti promatrajući dijagram svojstven statistici, a to je **dijagram raspršenosti** podataka koji u koordinatnom sustavu prikazuje uređene parove podataka iz dviju slučajnih varijabli. Takav tip grafa nam omogućuje vizualni prikaz iz kojeg možemo uočiti kako se vrijednosti jedne varijable mijenjaju u odnosu na vrijednosti druge varijable. Iz rasporeda točaka u dijagramu raspršenosti donosimo prve zaključke o vezi između varijabli. Na primjer, veza među varijablama može biti linearna (točke se formiraju oko pravca) ili krivolinijska (točke se formiraju oko krivulje). Nama će biti zanimljivo proučavati linearnu vezu među varijablama te ćemo se na to bazirati u ovome radu (Slika 2.1).



Slika 2.1: Primjer dijagrama raspršenosti

Regresijska metoda modeliranja koju ćemo opisati u ovom poglavlju pretpostavlja da možemo uspostaviti funkcijsku vezu do na dodanu grešku, odnosno da će veza između nezavisne varijable x i zavisne slučajne varijable $Y(x)$ biti oblika

$$Y(x) = f(x) + \varepsilon \tag{2.1}$$

gdje pretpostavljamo da je ε slučajna varijabla koja opisuje grešku u modeliranju.

Cilj je da za dani niz mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$ dvaju obilježja ustanovimo prirodu ovisnosti slučajne varijable Y (čije su realne vrijednosti brojevi $y_1, \ldots, y_n$) o nezavisnoj varijabli x (čije su izmjerene vrijednosti $x_1, \ldots, x_n$). Promatrat ćemo matematički model oblika

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.2}$$

gdje je funkcija f realna funkcija jedne varijable, a $\varepsilon_1, \ldots, \varepsilon_n$ međusobno nezavisne slučajne varijable. Takav model zovemo regresijski model. Ako pretpostavimo da je graf funkcije $f(x)$ pravac, onda dobivamo linearni regresijski model oblika:

$$y_i = \beta_1 x_i + \beta_0 + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.3}$$

Cilj je pronaći model koji najbolje opisuje dane podatke, odnosno tražimo onaj model koji najmanje odstupa od zadanih podataka. Postoje razni načini za mjerenje odstupanja ([1]):

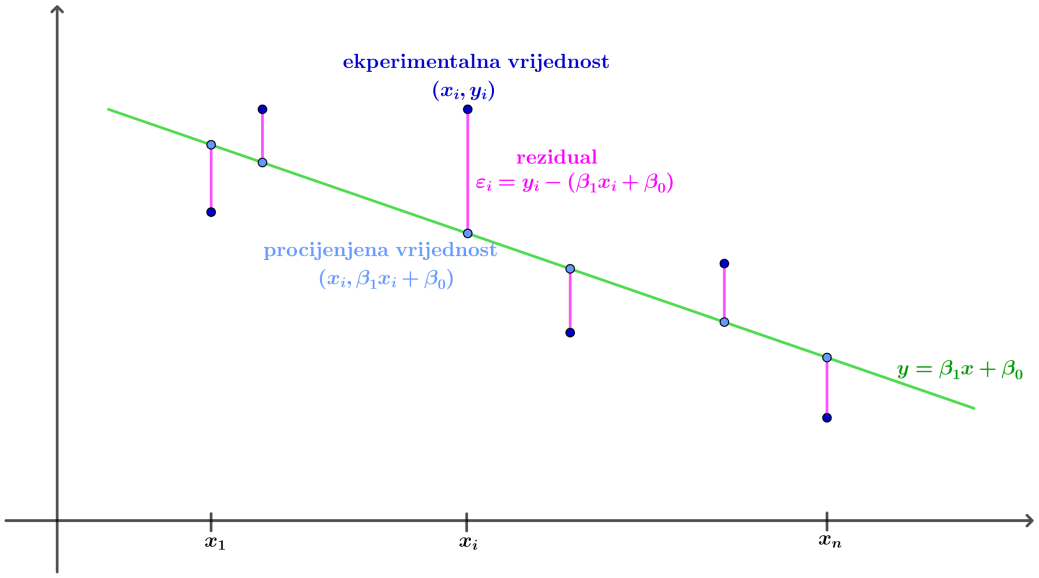
- zbrajanjem razlika y-koordinata ili x-koordinata
- zbrajanjem euklidskih udaljenosti od pravca za svaku točku

- zbrajanjem apsolutnih vrijednosti razlike y-koordinata ili x-koordinata
- zbrajanjem kvadrata razlika y-koordinata.

Proučavanjem različitih metoda, matematičari su došli do nekoliko ključnih zaključaka. Zbrajanje razlika između eksperimentalnih i teorijskih vrijednosti može uzrokovati probleme, jer se pozitivna i negativna odstupanja mogu međusobno poništiti. Sljedeći korak bio je razmatranje apsolutne vrijednosti odstupanja, no budući da funkcija apsolutne vrijednosti nije neprekidno derivabilna, nije moguće precizno odrediti minimum. Stoga je uvedena metoda kvadriranja odstupanja, jer kvadriranjem uvijek dobivamo pozitivne vrijednosti, bez obzira na predznak odstupanja. Dodatno, kvadriranjem odstupanja dobivamo kvadratnu funkciju koja je glatka, što omogućuje diferenciranje i pronalaženje jednoznačno određenog minimuma. Posljednja metoda, poznatija kao metoda najmanjih kvadrata se zbog svoje učinkovitosti i praktičnosti pokazala kao najbolja te ćemo je iz tog razloga u nastavku detaljnije objasniti.

2.2 Metoda najmanjih kvadrata

Neka je dan sljedeći dijagram raspršenosti te pretpostavimo da je dodan graf pravca $y = \beta_1 x + \beta_0$.



Slika 2.2: Metoda najmanjih kvadrata

Pogledamo li gornji dijagram raspršenosti, možemo uočiti da se dane točke mjerenja formiraju oko nekakvog pravca pa možemo pretprostaviti da između varijabli x i Y postoji linearna zavisnost. Želimo naći pravac koji najbolje modelira dane podatke. Takav pravac nazivamo **pravac regresije** te će biti oblika:

$$y = \beta_1 x + \beta_0 \tag{2.4}$$

gdje slobodni koeficijent β_0 predstavlja odsječak na y-osi, a koeficijent β_1 nagib pravca. Također, parametar β_1 pokazuje onu veličinu za koju će se u prosjeku promijeniti zavisna varijabla ako se nezavisna varijabla promijeni za jednu jedinicu.

Možemo uočiti da procijenjeni pravac regresije ne sadrži sve točke, odnosno neke se točke nalaze iznad, a neke ispod procijenjenog pravca. Neka je eksperimentalna (stvarna) vrijednost označena s Y, a procijenjena (teorijska) vrijednost s $\hat{Y}$. Vidljivo je da postoji razlika između stvarne i procijenjene vrijednosti te ćemo tu razliku označiti s $\varepsilon_i = y_i - \hat{y}_i$ za i-to mjerenje. Tu razliku još zoveno **greška** ili **rezidual** za i-to mjerenje (Slika 2.2). Pravac $\hat{Y} = \beta_1 x + \beta_0$ koji najbolje opisuje stvarne vrijednosti mora biti položen između točaka u ravnini tako da zbroj kvadrata odstupanja stvarnih vrijednosti Y od teorijskih vrijednosti $\hat{Y}$ bude minimalan, odnosno mora vrijediti:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \longrightarrow \min . \tag{2.5}$$

Uz taj uvjet jednadžba traženog pravca može se odrediti **metodom najmanjih kvadrata** te ima svojstvo da prolazi kroz aritmetičke sredine podataka. Uočimo da je $\hat{y}_i = \beta_1 x_i + \beta_0$ pa je pogreška individualne točke opisana izrazom:

$$\varepsilon_i = y_i - (\beta_1 x_i + \beta_0) \tag{2.6}$$

gdje je (x_i, y_i) koordinata i-te točke, a β_1 i β_0 koeficijenti pravca pa se problem minimizacije sume kvadrata svodi na minimizaciju funkcije $S(\beta_1, \beta_0)$:

$$S(\beta_1, \beta_0) = \sum_{i=1}^{n} (y_i - (\beta_1 x_i + \beta_0))^2 \longrightarrow \min . \tag{2.7}$$

S obzirom da je S glatka funkcija (jer je kvadratna funkcija), za minimum je nužno da vrijedi:

$$\frac{\partial S}{\partial \beta_1} = 0 \quad \text{i} \quad \frac{\partial S}{\partial \beta_0} = 0. \tag{2.8}$$

Parcijalnom derivacijom funkcije S po β_1 dobivamo:

$$\begin{aligned}\frac{\partial S}{\partial \beta_1} &= \sum_{i=1}^{n} 2[y_i - (\beta_1 x_i + \beta_0)] \cdot (-x_i) \\ &= \sum_{i=1}^{n} 2(-x_i y_i + \beta_1 x_i^2 + \beta_0 x_i) \\ &= -2\sum_{i=1}^{n} x_i y_i + 2\beta_1 \sum_{i=1}^{n} x_i^2 + 2\beta_0 \sum_{i=1}^{n} x_i\end{aligned}$$

te izjednačavanjem s 0 dobivamo:

$$\begin{aligned}&-2\sum_{i=1}^{n} x_i y_i + 2\beta_1 \sum_{i=1}^{n} x_i^2 + 2\beta_0 \sum_{i=1}^{n} x_i = 0 \\ \iff &\beta_1 \sum_{i=1}^{n} x_i^2 + \beta_0 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0.\end{aligned}$$

Parcijalnom derivacijom funkcije S po β_0 dobivamo:

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} &= \sum_{i=1}^{n} 2[y_i - (\beta_1 x_i + \beta_0)] \\ &= 2\sum_{i=1}^{n} y_i - 2\beta_1 \sum_{i=1}^{n} x_i - 2\beta_0 \sum_{i=1}^{n} 1\end{aligned}$$

te izjednačavanjem s 0 dobivamo:

$$\begin{aligned}&2\sum_{i=1}^{n} y_i - 2\beta_1 \sum_{i=1}^{n} x_i - 2\beta_0 \sum_{i=1}^{n} 1 = 0 \\ \iff &\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - \beta_0 n = 0.\end{aligned}$$

Ovime smo dobili sustav jednadžbi:

$$\begin{aligned}\beta_1 \sum_{i=1}^{n} x_i^2 + \beta_0 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0 \\ \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - \beta_0 n = 0.\end{aligned} \tag{2.9}$$

Sjetimo se da smo rekli da pravcu regresije pripadaju aritmetičke vrijednosti podataka

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{i} \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2.10}$$

te iz toga slijedi:

$$\sum_{i=1}^{n} x_i = n\bar{x} \quad \text{i} \quad \sum_{i=1}^{n} y_i = n\bar{y}. \tag{2.11}$$

Sada ćemo supstituirati te izraze u (2.9) i riješiti sustav. Iz druge jednadžbe sustava imamo:

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n}$$

odnosno

$$\beta_0 = \bar{y} - \beta_1 \bar{x}. \tag{2.12}$$

Uvrštavanjem (2.11) i (2.12) u prvu jednadžbu sustava dobivamo:

$$\begin{aligned}
\beta_1 \sum_{i=1}^{n} x_i^2 + (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i &= 0 \\
\beta_1 \sum_{i=1}^{n} x_i^2 + \bar{y} \sum_{i=1}^{n} x_i - \beta_1 \bar{x} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i &= 0 \\
\beta_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) + n\bar{x}\bar{y} - \sum_{i=1}^{n} x_i y_i &= 0.
\end{aligned}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{2.13}$$

Jednadžba pravca linearne regresije je

$$\hat{y} = \beta_1 x + \beta_0, \tag{2.14}$$

gdje su parametri β_1 i β_0 dani s (2.12) i (2.13).

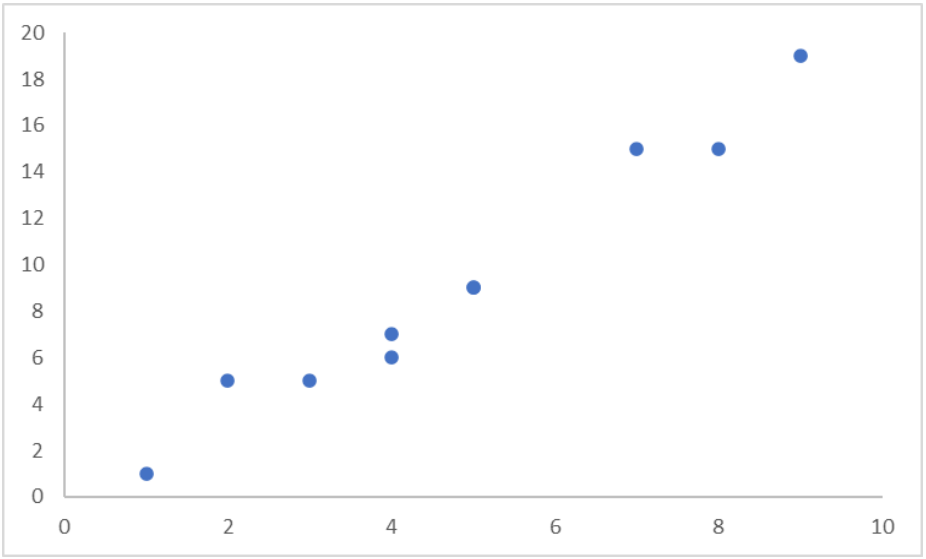
Pokažimo izvedeno na sljedećem primjeru.

Primjer 2.2.1. *Metodom najmanjih kvadrata pronađite funkcijsku vezu koja najbolje opisuje dane podatke:*

x	5	2	4	3	1	5	8	4	7	9
y	9	5	7	5	1	9	15	6	15	19

.

Nacrtajmo najprije pripadajući dijagram raspršenosti.



Iz dijagrama je vidljivo da se točke nalaze približno na nekom pravcu, stoga tražimo funkciju oblika $\hat{y} = \beta_1 x + \beta_0$. Prema gore navedenom, koeficijente β_1 i β_0 ćemo izračunati kao:

$$\begin{aligned} \beta_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \\ \beta_0 &= \bar{y} - \beta_1 \bar{x}. \end{aligned}$$

Izračunajmo potrebne podatke:

x_i	5	2	4	3	1	5	8	4	7	9
y_i	9	5	7	5	1	9	15	6	15	19
x_i^2	25	4	16	9	1	25	64	16	49	81
$x_i y_i$	45	10	28	15	1	45	120	24	105	171

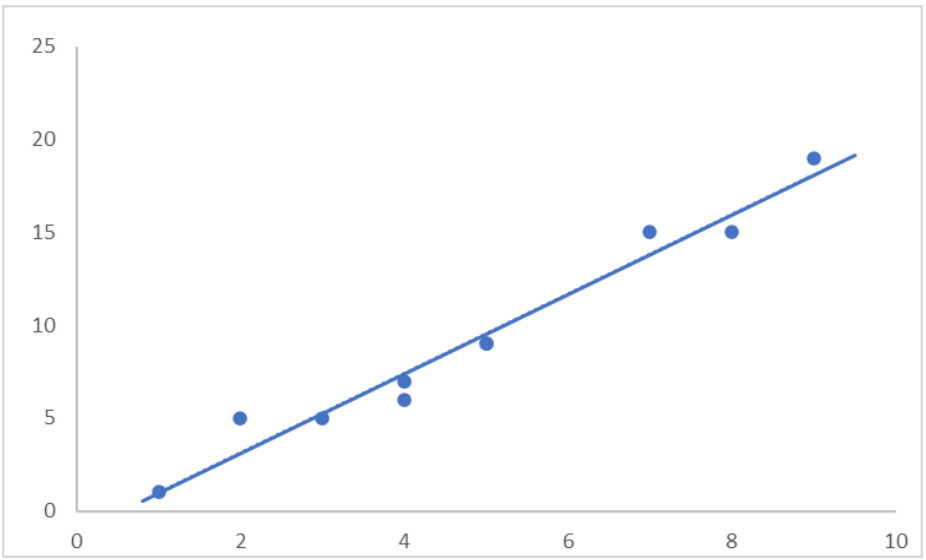
$$\begin{aligned}
\sum_{i=1}^{10} x_i &= 48 \\
\bar{x} &= 4.8 \\
\sum_{i=1}^{10} y_i &= 91 \\
\bar{y} &= 9.1 \\
\sum_{i=1}^{10} x_i^2 &= 290 \\
\sum_{i=1}^{10} x_i y_i &= 564.
\end{aligned}$$

Sada imamo:

$$\begin{aligned}
\beta_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{564 - 10 \cdot 4.8 \cdot 9.1}{290 - 10 \cdot 4.8^2} = 2.13, \\
\beta_0 &= \bar{y} - \beta_1 \bar{x} = 9.1 - 2.13 \cdot 4.8 = -1.12.
\end{aligned}$$

Tražena funkcija je:

$$\hat{y} = 2.13x - 1.12$$



2.3 Analiza modela

Pretpostavke modela

Polazne pretpostavke u analizi modela jednostavne linearne regresije su:

(A1) Veza između zavisne varijable y i nezavisne varijable x je linearna.

(A2) Varijabla x je deterministička varijabla i vrijedi $\mathbb{E}[\varepsilon_i|x_i] = 0, \forall i$.

Model može biti linearan u varijablama ili linearan u parametrima. Linearnost modela linearne regresije odnosi se na linearnost među parametrima što znači da se koriste prve potencije parametara bez dodatnih transformacija. Ukoliko početni zapis modela nije linearan, ponekad ga je moguće jednostavnim transformacijama svesti na linearan, što ćemo detaljnije opisati nešto kasnije.

Želimo da nezavisna varijabla bude deterministička varijabla odnosno da pri svakom novom mjerenju podatci nezavisne varijable ostaju isti te se prikupljaju novi podatci samo za zavisnu i slučajnu varijablu greški. U stvarnosti to najčešće nije slučaj, no bitno je uvesti ovu pretpostavku kako bi fokus bio isključivo na proučavanju veze između zavisne i nezavisne varijable, a ne i na izvoru varijacija nezavisne varijable. Dakle, $\mathbb{E}[\varepsilon_i|x_i] = 0$ znači da opažanja vezana za nezavisnu varijablu x ne utječu na očekivanu vrijednost slučajne varijable ε. Ako je nezavisna varijabla deterministička, uvodimo dodatne pretpostavke koje se odnose na slučajnu varijablu ε.

Definicija 2.3.1. *([6]) Neka su $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ i-te vrijednosti regresijskog pravca. Za slučajne greške ε_i, $\forall i = 1, \ldots, n$ vrijedi:*

(A3) *centriranost:* $\mathbb{E}[\varepsilon_i] = 0, \forall i,$

(A4) *jednakost varijanci:* $\mathrm{Var}[\varepsilon_i] = \sigma^2, \forall i,$

(A5) *nekoreliranost:* $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0, \forall i \neq j.$

Uvjete $(A3) - (A5)$ *zovemo Gauss–Markovljevim uvjetima.*

U nastavku ćemo objasniti navedene pretpostavke.

(A3) Očekivana vrijednost slučajne varijable ε iznosi 0, što znači da će veza između zavisne i nezavisne varijable doista biti linearna. Iz toga proizlazi da će uvjetno očekivanje zavisne varijable biti jednako determinističkom dijelu modela tj. $\mathbb{E}[y_i|x_i] = \beta_1 x_i + \beta_0, \forall i$.

(A4) Stalnost varijanci je važna pretpostavka kako bi raspršenost podataka zavisne varijable bila jednaka za sve podatke nezavisne varijable.

(A5) Svake dvije slučajne varijable grešaka su međusobno nezavisne.

Možemo reći da metoda najmanjih kvadrata daje dobre procjenitelje parametara regresijskog pravca ako su zadovoljeni Gauss–Markovljevi uvjeti.

Dodatno se još može pretpostaviti da za slučajnu varijablu greške ε_i vrijedi:

$$\varepsilon_i \sim N(0, \sigma^2), \forall i.$$

Reprezentativnost modela

Nakon pronalaženja regresijskih koeficijenata i samog pravca linearne regresije, možemo se pitati koliko dobro taj pravac reprezentira odnos među varijablama. Osnovna mjera reprezentativnosti modela je raspršenost podataka oko pravca regresije koja se mjeri prema odstupanjima eksperimentalnih vrijednosti od procijenjenih vrijednosti pravcem regresije. To su rezidualna odstupanja odnosno greške procjene te ih ima koliko i vrijednosti varijabli odnosno n. Ako je rezidualno odstupanje jednako nuli, stvarne vrijednosti ovisne varijable bit će na pravcu regresije što znači da taj pravac precizno procjenjuje ovisnu varijablu. Kako se rezidualno odstupanje povećava, stvarne vrijednosti ovisne varijable će sve više odstupati od pravca regresije, što znači da taj pravac daje slabu procjenu ovisne varijable. Prema tome, rezidualna odstupanja su dobar pokazatelj preciznosti modela. Definirajmo mjere koje najbolje opisuju prosječno odstupanje podataka i reprezentativnost modela.

Definicija 2.3.2. *Varijanca regresije, u oznaci σ^2, je aritmetička sredina kvadrata rezidualnih odstupanja tj.*

$$\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}.$$

Definicija 2.3.3. *Standardna devijacija regresije je pozitivni drugi korijen iz varijance regresije tj.*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}.$$

Standardna devijacija pokazuje koliko je prosječno odstupanje eksperimentalnih vrijednosti zavisne varijable od regresijskih vrijednosti te je izražena istom mjernom jedinicom kao i zavisna varijabla.

Definicija 2.3.4. *Koeficijent varijacije, u oznaci V, je definiran kao*

$$V = \frac{\sigma}{\bar{y}} \cdot 100$$

te pokazuje koliko je prosječno odstupanje eksperimentalnih vrijednosti zavisne varijable od regresijskih vrijednosti izraženo u relativnom odnosu (u %).

Test jakosti modela

Definicija 2.3.5. *Koeficijent determinacije, u oznaci R^2, je mjera raspršenosti podataka koja opisuje jakost veze između nezavisne i zavisne varijable te je definirana kao*

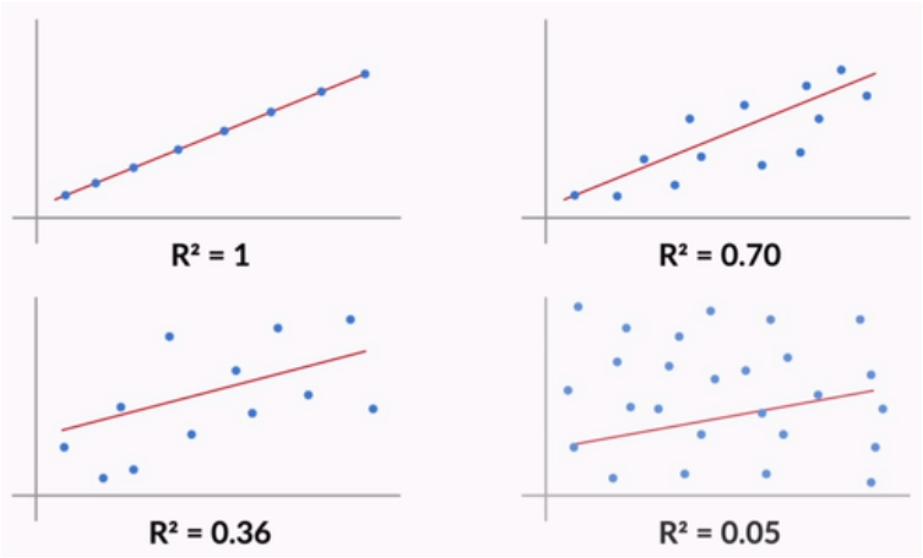
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad 0 \leq R^2 \leq 1.$$

Vrijednost koeficijenta determinacije pripada intervalu $[0, 1]$. Kada je koeficijent determinacije $R^2 = 0$, to znači da regresijski model ne objašnjava nikakvu varijaciju zavisne varijable u odnosu na nezavisnu. U tom slučaju, nijedna od točaka podataka ne leži na pravcu linearne regresije. U drugom krajnjem slučaju, kada je koeficijent determinacije $R^2 = 1$, regresijski model u potpunosti objašnjava varijaciju zavisne varijable te sve točke podataka, koje promatramo u dijagramu raspršenosti, leže na regresijskom pravcu. Možemo zaključiti, što je vrijednost koeficijenta determinacije bliža broju 1, to su podatci bliži pravcu te je veza između varijabli jača tj. dobiveni model dobro reprezentira dane podatke. Suprotno, što je vrijednost koeficijenta determinacije bliža broju 0, to su podatci raspršeniji te je veza između varijabli slabija. Tablica 2.1 prikazuje ovisnost koeficijenta determinacije i jakosti linearne veze između zavisne i nezavisne varijable.

koeficijent determinacije R^2	**značenje**
0.00	odsutnost veze
0.00 - 0.25	slaba veza
0.25 - 0.64	veza srednje jakosti
0.64 - 1.00	čvrsta veza
1.00	potpuna veza

Tablica 2.1: Chadockova ljestvica

Na Slici 2.3 možemo vidjeti prikaz dijagrama raspršenosti za različite vrijednosti koeficijenta determinacije.



Slika 2.3: Koeficijent determinacije, preuzeto sa ([10])

Primjer 2.3.6. *Provjerimo koliko je prosječno odstupanje podataka te jakost veze među varijablama u Primjeru 2.2.1.*

y_i	9	5	7	5	1	9	15	6	15	19
$\hat{y}_i$	9.53	3.14	7.4	5.27	1.01	9.53	15.92	7.4	13.79	18.05

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{9.4274}{10}} = 0.97$$

$$V = \frac{\sigma}{\bar{y}} \cdot 100 = \frac{0.97}{9.1} \cdot 100 = 10.66\%$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{270.3994}{7660.09} = 0.96$$

Dobivena rješenja pokazuju da je prosječno odstupanje relativno malo. Koeficijent determinacije pokazuje da je 96% varijacija zavisne varijable objašnjeno ovim modelom, dok je samo 4% ostalo neobjašnjeno te je veza između zavisne i nezavisne varijable čvrsta. Prema tome, možemo zaključiti da je dobiveni pravac regresije dobro reprezentira i precizno procjenjuje dane podatke.

2.4 Višestruka linearna regresija

Jednostavna linearna regresija je koristan pristup u predviđanju varijable odaziva (zavisne varijable) na temelju jednog prediktora (nezavisne narijable). Međutim, u praksi često imamo više od jednog prediktora. Na primjer, možemo analizirati kako unos hrane i tjelesna aktivnost utječu na gubitak tjelesne težine. U ovom slučaju imamo dva prediktora odnosno dvije nezavisne varijable (unos hrane i tjelesna aktivnost) i jednu varijablu odaziva odnosno zavisnu varijablu (gubitak tjelesne težine). Stoga, želimo proširiti jednostavni linearni regresijski model tako da izravno možemo obraditi više prediktora.

Općenito, pretpostavimo da zavisnu varijablu Y želimo predviđati na osnovu p različitih prediktora, $X_1, \ldots, X_p$. Tada višestruki linearni regresijski model ima oblik:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.15}$$

gdje je:

- y_i predviđena vrijednost zavisne varijable za i-to mjerenje
- x_{ip} vrijednost varijabli prediktora p za i-to mjerenje
- β_0 slobodni koeficijent
- β_p regresijski koeficijenti koji objašnjavaju povezanost između p-tog prediktora i zavisne varijable Y u smislu prosječne promjene zavisne varijable kada se vrijednost x_p promijeni za jednu jedinicu dok vrijednosti ostalih prediktora ostaju nepromijenjeni
- ε_i rezidual za i-to mjerenje.

Grafički prikaz podataka u višestrukoj linearnoj regresiji nije toliko jednostavan kao u jednostavnoj linearnoj regresiji. U slučaju višestruke linearne regresije bi se prikazala hiperravnina u p-dimenzionalnom koordinatnom sustavu. Međutim, ideja traženja koeficijenata višestruke linearne regresije je ista kao i ranije. Cilj je odrediti koeficijente tako da je suma kvadrata reziduala minimalna pa koeficijente procjenjujemo metodom najmanjih kvadrata. Međutim, već za $p = 3$ ta metoda postaje dovoljno složena da se u efektivnim izračunima moraju koristiti računala što nećemo razmatrati u ovom radu.

2.5 Linearizacija

Ako niz točaka želimo aproksimirati funkcijom φ koja nelinearno ovisi o parametrima, dobili bismo nelinearni sustav jednadžbi koji se teško rješava te to neće biti tema ovog rada. Međutim, u određenim slučajevima se jednostavnim transformacijama problem može svesti u linearni problem najmanjih kvadrata. Takav postupak se naziva **linearizacija**. Navedimo primjer linearizacije funkcije u problem najmanjih kvadrata.

Primjer 2.5.1. *Zadane su točke* $(x_0, y_0), \ldots, (x_n, y_n)$ *koje pomoću diskretne metode najmanjih kvadrata aproksimiramo funkcijom oblika* $\varphi(x) = a_0 x^{a_1}$.

Greška aproksimacije, odnosno funkcija koju minimiziramo jednaka je

$$S = S(a_0, a_1) = \sum_{k=0}^{n} (y_k - \varphi(x_k))^2 = \sum_{k=0}^{n} (y_k - a_0 x_k^{a_1})^2 \to \min .$$

S obzirom da tražimo minimum, gornju funkciju ćemo parcijalno derivirati po parametrima a_0 i a_1 te izjednačiti s nulom:

$$\begin{aligned} \frac{\partial S}{\partial a_0} &= -2 \sum_{k=0}^{n} (y_k - a_0 x_k^{a_1}) x_k^{a_1} = 0 \\ \frac{\partial S}{\partial a_1} &= -2 \sum_{k=0}^{n} (y_k - a_0 x_k^{a_1}) a_0 x_k^{a_1} \ln x_k = 0. \end{aligned}$$

što je nelinearan sustav jednadžbi. Međutim, logaritmiramo li funkciju $\varphi(x) = a_0 x^{a_1}$ dobit ćemo:

$$\log \varphi(x) = \log a_0 + a_1 \log x.$$

Moramo još logaritmirati i vrijednosti funkcije y u točkama x_k pa uz supstitucije

$$h(x) = \log y(x), \quad h_k = h(x_k) = \log y_k, \quad k = 0, \ldots, n$$

i

$$\psi(x) = \log \varphi(x) = b_0 + b_1 \log x$$

gdje je

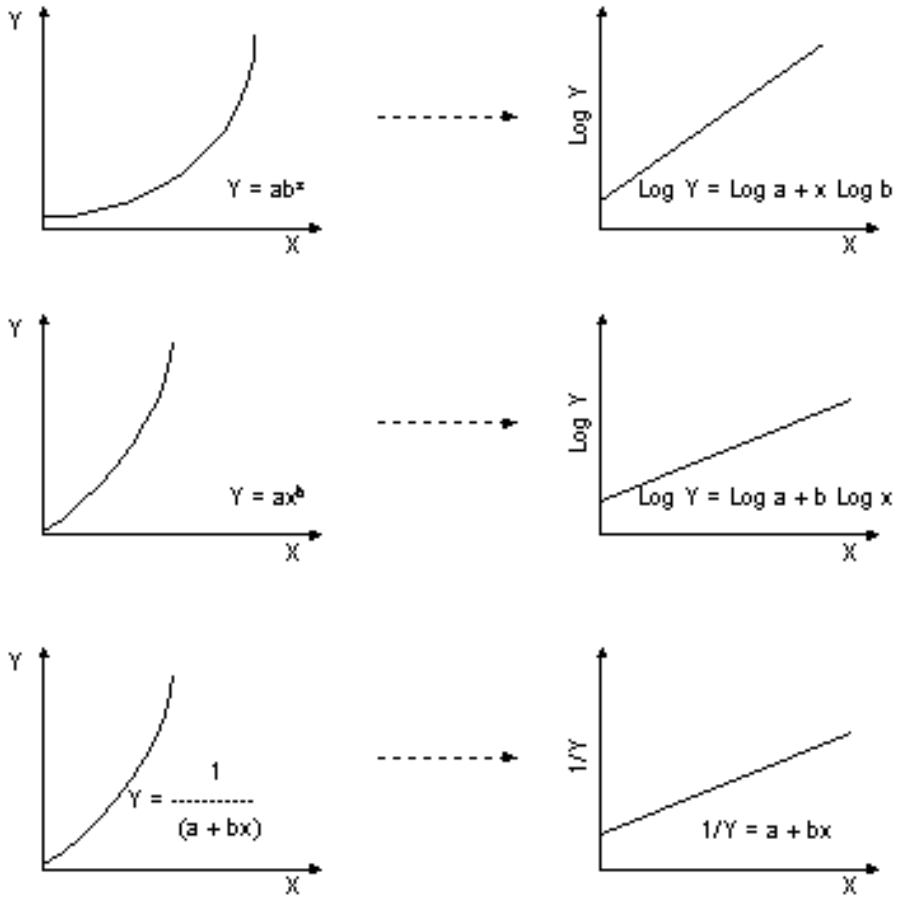
$$b_0 = \log a_0, \quad b_1 = a_1$$

dobivamo linearni problem najmanjih kvadrata

$$\tilde{S} = \tilde{S}(b_0, b_1) = \sum_{k=0}^{n} (h_k - \psi(x_k))^2 = \sum_{k=0}^{n} (h_k - b_0 - b_1 \log x_k)^2 \to min.$$

Da bismo proveli linearizaciju u ovom slučaju, mora vrijediti $x_k > 0$ i $y_k > 0$.

Na Slici 2.4 možemo vidjeti primjere linearizacije funkcija.



Slika 2.4: Primjeri linearizacije, preuzeto iz ([15])

Poglavlje 3

Matrična formulacija problema traženja regresijskog pravca

U ovom poglavlju obradit ćemo matričnu formulaciju linearne regresije. Nakon što objasnimo teorijsku pozadinu, primijenit ćemo je na konkretan primjer kako bismo pokazali praktičnu primjenu ovog pristupa.

3.1 Problem najmanjih kvadrata

Neka su točke $(x_1, y_1), \ldots, (x_n, y_n)$ dobiveni podatci nekog eksperimenta. Kada bi sve točke ležale na istom pravcu $y = \beta_1 x + \beta_0$, tada bi vrijedilo:

$$\begin{aligned} \beta_1 x_1 + \beta_0 &= y_1 \\ \beta_1 x_2 + \beta_0 &= y_2 \\ &\vdots \\ \beta_1 x_n + \beta_0 &= y_n. \end{aligned}$$

Na taj način dobivamo sustav od n jednadžbi s dvije nepoznanice β_0 i β_1 te ga možemo zapisati matrično kao

$$Ax = b$$

gdje je

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad x = \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix}, \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

S obzirom da znamo da sve te točke ne leže na istom pravcu odnosno da sustav nije rješiv, najbliže što možemo učiniti je pronaći x takav da je Ax što bliži vektoru b. Zapravo zamišljamo Ax kao aproksimaciju b. Što je manja udaljenost između b i Ax, to je bolja aproksimacija. Tako se problem svodi na problem minimizacije

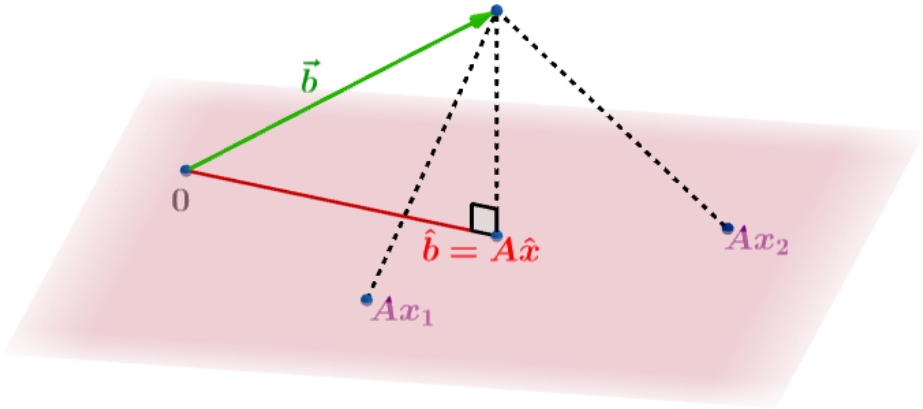
$$\|b - Ax\| \longrightarrow \min . \tag{3.1}$$

Definicija 3.1.1. *Neka je $A \in M_{m,n}$ i $b \in M_{m,1}$. Rješenje problema najmanjih kvadrata $Ax = b$ je $\hat{x} \in M_{n,1}$ takav da je*

$$\|b - A\hat{x}\| \leq \|b - Ax\| \tag{3.2}$$

za svaki $x \in M_{n,1}$.

Važno je napomenuti da, bez obzira koji x odabrali, Ax nužno mora pripadati slici od A. Dakle, tražimo x takav da je Ax u slici od A i da je najbliži vektoru b. Te uvjete zadovoljava upravo vektor koji je ortogonalna projekcija vektora b te ćemo ga označiti s $\hat{b}$.



Slika 3.1: Vektor b je bliži $A\hat{x}$ nego Ax za neki drugi x

Budući da je $\hat{b}$ u slici od Ax, jednadžba $Ax = \hat{b}$ je konzistentna te postoji rješenje te jednadžbe koje ćemo označiti s $\hat{x}$.

Nadalje, možemo uočiti da je vektor $b - A\hat{x}$ ortogonalan na svaki vektor od A. Ako je a_j bilo koji stupac od A, tada vrijedi $a_j(b - A\hat{x}) = 0$ odnosno $a_j^T(b - A\hat{x}) = 0$. S obzirom da je svaki a_j^T redak od A^T dalje slijedi:

$$\begin{aligned} A^T(b - A\hat{x}) &= 0 \\ A^Tb - A^TA\hat{x} &= 0 \\ A^TA\hat{x} &= A^Tb. \end{aligned}$$

Iz ovoga slijedi da svako rješenje najmanjih kvadrata jednadžbe $Ax = b$ zadovoljava jednadžbu:

$$A^T Ax = A^T b \tag{3.3}$$

koja se zove **normalna jednadžba** za $Ax = b$.

Teorem 3.1.2. *Skup rješenja problema najmanjih kvadrata je neprazan i zadovoljava jednadžbu $Ax = b$. Tada zadovoljava i normalnu jednadžbu $A^T Ax = A^T b$.*

Dokaz. Neka je skup rješenja problema najmanjih kvadrata neprazan skup i svako rješenje $\hat{x}$ problema najmanjih kvadrata zadovoljava normalnu jednadžbu. Obratno, pretpostavimo da $\hat{x}$ zadovoljava jednadžbu

$$A^T A\hat{x} = A^T b.$$

Tada $\hat{x}$ zadovoljava jednadžbu

$$A^T(b - A\hat{x}) = 0,$$

što pokazuje da je $b - A\hat{x}$ ortogonalan na sve retke matrice A^T, pa je prema tome ortogonalan na sve stupce matrice A. Budući da stupci matrice A razapinju $\mathrm{Im}(A)$, vektor $b - A\hat{x}$ je ortogonalan na cijelu $\mathrm{Im}(A)$. Stoga je jednadžba

$$b = A\hat{x} + (b - A\hat{x})$$

zapis vektora b kao linearne kombinacije vektora koji pripadaju $\mathrm{Im}(A)$ i vektora koji su ortogonalni na $\mathrm{Im}(A)$. Zbog jedinstvenosti ortogonalnog rastava, $A\hat{x}$ mora biti ortogonalna projekcija vektora b na $\mathrm{Im}(A)$. Odnosno, $A\hat{x} = \hat{b}$ i $\hat{x}$ je rješenje problema najmanjih kvadrata. □

Dokaz prethodnog teorema preuzet je iz [7].

Teorem 3.1.3. *Neka je $A \in M_{m,n}$. Sljedeće tvrdnje su ekvivalentne:*

1. *Jednadnžba $Ax = b$ ima jedinstveno rješenje problema najmanjih kvadra za svaki $b \in \mathbb{R}^m$.*
2. *Stupci matrice A su linearno nezavisni.*
3. *Matrica $A^T A$ je invertibilna.*

Kada vrijedi istinitost ovih tvrdnji, tada je $\hat{x}$ rješenje problema najmanjih kvadrata dano kao

$$\hat{x} = (A^T A)^{-1} A^T b.$$

3.2 Primjena matrične formulacije

Primjer 3.2.1. *Metodom najmanjih kvadrata pronađite jednadžbu pravca* $y = \beta_1 x + \beta_0$ *koji najbolje opisuje sljedeće podatke:* $(1, 0), (2, 1), (4, 2)$ *i* $(5, 3)$.

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}$$

Za rješenje najmanjih kvadrata $Ax = b$, normalna jednadžba glasi

$$A^T Ax = A^T b$$

pa računamo:

$$A^T A = \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix} = \begin{bmatrix} 46 & 12 \\ 12 & 4 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 25 \\ 6 \end{bmatrix}$$

Normalna jednadžba sada glasi:

$$\begin{bmatrix} 46 & 12 \\ 12 & 4 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix} = \begin{bmatrix} 25 \\ 6 \end{bmatrix}$$

Iz toga je

$$\begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix} = \begin{bmatrix} 46 & 12 \\ 12 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 25 \\ 6 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 2 & -6 \\ -6 & 23 \end{bmatrix} \begin{bmatrix} 25 \\ 6 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 14 \\ 12 \end{bmatrix} = \begin{bmatrix} \frac{7}{10} \\ \frac{3}{5} \end{bmatrix}$$

Iz ovoga slijedi da je tražena jednadžba:

$$y = \frac{7}{10} x + \frac{3}{5}.$$

Poglavlje 4

Primjene

Linearna regresija je najčešće korištena statistička tehnika za istraživanje i modeliranje odnosa između varijabli.. Primjene regresije su brojne i pojavljuju se u gotovo svim područjima, uključujući inženjerstvo, ekonomiju, biologiju, medicinu, društvene znanosti i brojne druge. Navedimo nekoliko primjera.

Ekonomija i financije: Linearna regresija se koristi u ekonomiji za analizu odnosa između ekonomskih varijabli, poput utjecaja kamatnih stopa na potrošnju ili odnosa između inflacije i nezaposlenosti.

Marketing i prodaja: Tvrtke koriste jednostavnu linearnu regresiju za predviđanje prodaje. Analiziranjem povijesnih podataka o prodaji i čimbenika poput potrošnje na oglašavanje ili promjena cijena, tvrtke mogu predvidjeti buduću prodaju i prilagoditi svoje strategije.

Zdravlje okoliša: Stručnjaci u ovom području koriste ovaj regresijski model za procjenu odnosa između prirodnih elemenata, poput tla, vode i zraka. Primjer je utjecaj temperature ili količine oborina na rast biljaka. Također, čest je primjer analiza kvalitete tla kao što su pH vrijednost, vlaga ili prisutnost hranjivih tvari na uspijevanje poljoprivrednih kultura. Regresijska analiza od velike je koristi ekolozima u predviđanju učinaka zagađenja zraka ili vode na zdravlje okoliša.

Medicina: Linearna regresija može se primijeniti u zdravstvu za proučavanje odnosa između varijabli poput dobi pacijenta i medicinskih troškova, doze lijekova i ishoda liječenja ili zadovoljstva pacijenata i vremena čekanja u bolnici. Na primjer, koristi se za analizu utjecaja čimbenika poput prehrambenih navika, tjelesne aktivnosti, pušenja i konzumacije alkohola na zdravstvene pokazatelje kao što su krvni tlak, razina kolesterola ili tjelesna masa. Također možemo procijeniti kako povećanje tjelesne aktivnosti ili smanjenje konzumacije alkohola može utjecati na smanjenje krvnog tlaka.

Sportska analitika: U sportskoj analitici, može se koristiti za analizu mjernih podataka o performansama igrača (npr. prosječan broj udaraca u bejzbolu ili postotak golova

u nogometu) i njihovih odnosa s čimbenicima poput intenziteta treninga, umora igrača ili strategija trenera. Također može se koristiti za predviđanje posjećenosti utakmica na temelju statusa timova koji igraju i veličine tržišta, kako bi savjetovali menadžere timova o mjestima održavanja utakmica i cijenama karata koje mogu maksimizirati profit.

Energija i komunalne usluge: Energetske tvrtke mogu koristiti linearnu regresiju za predviđanje potrošnje energije na temelju povijesnih podataka i vremenskih uvjeta. To pomaže u planiranju resursa i optimizaciji distribucije energije.

Navedimo detaljnije par primjera postavljanja modela linearne regresije ([3]):

Primjer 4.0.1. Tvrtke često koriste linearnu regresiju kako bi analizirale odnos između troškova oglašavanja i prihoda. Tako na primjer, mogu postaviti jednostavni linearni regresijski model koristeći troškove oglašavanja kao nezavisnu varijablu i prihode kao zavisnu varijablu. Regresijski model bi tada imao sljedeći oblik:

$$\text{prihodi} = \beta_0 + \beta_1 \cdot \text{trošak oglašavanja}.$$

- Koeficijent β_0 bi predstavljao ukupne očekivane prihode kada su troškovi oglašavanja nula.
- Koeficijent β_1 bi predstavljao prosječnu promjenu ukupnih prihoda kada se trošak oglašavanja poveća za jednu jedinicu (npr. jedan euro).
- Ako je β_1 negativan, to bi značilo da veći troškovi oglašavanja dovode do manjih prihoda.
- Ako je β_1 blizu nule, to bi značilo da trošak oglašavanja ima mali utjecaj na prihode.
- Ako je β_1 pozitivan, to bi značilo da veći trošak oglašavanja dovodi do većih prihoda.

Ovisno o vrijednosti β_1, tvrtka se može odlučiti na smanjivanje ili povećavanje svojih troškova oglašavanja.

Primjer 4.0.2. Istraživači mogu davati različite doze određenog lijeka pacijentima i promatrati kako pritom reagira njihov krvni tlak. Mogu postaviti jednostavni linearni regresijski model koristeći dozu lijeka kao nezavisnu varijablu i krvni tlak kao zavisnu varijablu. Regresijski model bi tada imao sljedeći oblik:

$$\text{krvni tlak} = \beta_0 + \beta_1 \cdot \text{doza lijeka}.$$

- Koeficijent β_0 predstavljao bi očekivani krvni tlak kada je doza lijeka nula.
- Koeficijent β_1 predstavljao bi prosječnu promjenu u krvnom tlaku kada se doza lijeka poveća za jednu jedinicu.

- Ako je β_1 negativan, to bi značilo da povećanje doze lijeka dovodi do smanjenja krvnog tlaka.
- Ako je β_1 blizu nule, to bi značilo da povećanje doze lijeka ne uzrokuje promjenu u krvnom tlaku.
- Ako je β_1 pozitivan, to bi značilo da povećanje doze lijeka dovodi do povećanja krvnog tlaka.

Ovisno o vrijednosti β_1, istraživači mogu odlučiti o promjeni doze lijeka koju pacijent prima.

Primjer 4.0.3. Poljoprivredni znanstvenici mogu primijeniti različite količine gnojiva i vode na različitim poljima i vidjeti kako to utječe na prinos usjeva. Mogu postaviti višestruki linearni regresijski model koristeći gnojivo i vodu kao varijable prediktore, a prinos usjeva kao zavisnu varijablu. Regresijski model bi imao sljedeći oblik:

$$\text{prinos usjeva} = \beta_0 + \beta_1 \cdot \text{količina gnojiva} + \beta_2 \cdot \text{količina vode}.$$

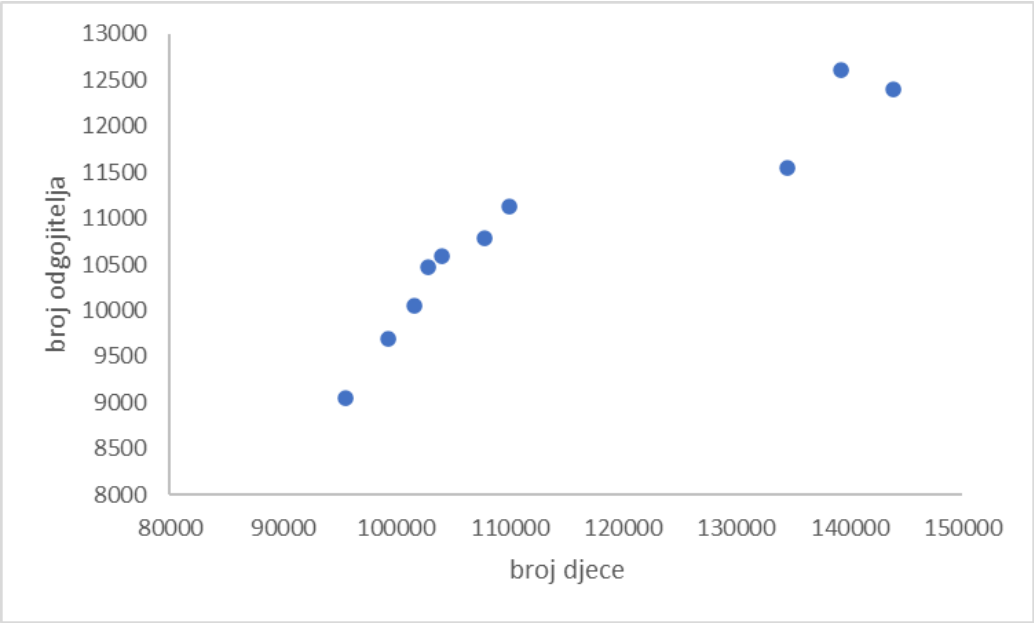
- Koeficijent β_0 predstavljao bi očekivani prinos usjeva bez gnojiva ili vode.
- Koeficijent β_1 predstavljao bi prosječnu promjenu u prinosu usjeva kada se količina gnojiva poveća za jednu jedinicu, pod uvjetom da količina vode ostane nepromijenjena.
- Koeficijent β_2 predstavljao bi prosječnu promjenu u prinosu usjeva kada se količina vode poveća za jednu jedinicu, pod uvjetom da količina gnojiva ostane nepromijenjena.

Ovisno o vrijednostima β_1 i β_2, znanstvenici mogu odlučiti o promijeni količine gnojiva i vode koje se koriste kako bi maksimizirali prinos usjeva.

Primjer 4.0.4. *U danoj tablici su podatci preuzeti od Državnog zavoda za statistiku koji prikazuju broj djece u predškolskom obrazovanju i broj odgojitelja u Hrvatskoj u razdoblju od 2008. do 2017. godine.*

Godina upisa	Broj djece	Broj odgojitelja
2008./2009.	95516	9054
2009./2010.	99317	9699
2010./2011.	101638	10046
2011./2012.	102857	10464
2012./2013.	104080	10591
2013./2014.	107823	10785
2014./2015.	109963	11125
2015./2016.	134573	11538
2016./2017.	143878	12396
2017./2018.	139228	12601

Neka je nezavisna varijabla x broj djece upisane u ustanovu predškolskog obrazovanja, a zavisna varijabla y broj odgojitelja. Nacrtajmo najprije dijagram raspršenosti da vidimo u kojem su odnosu varijable.



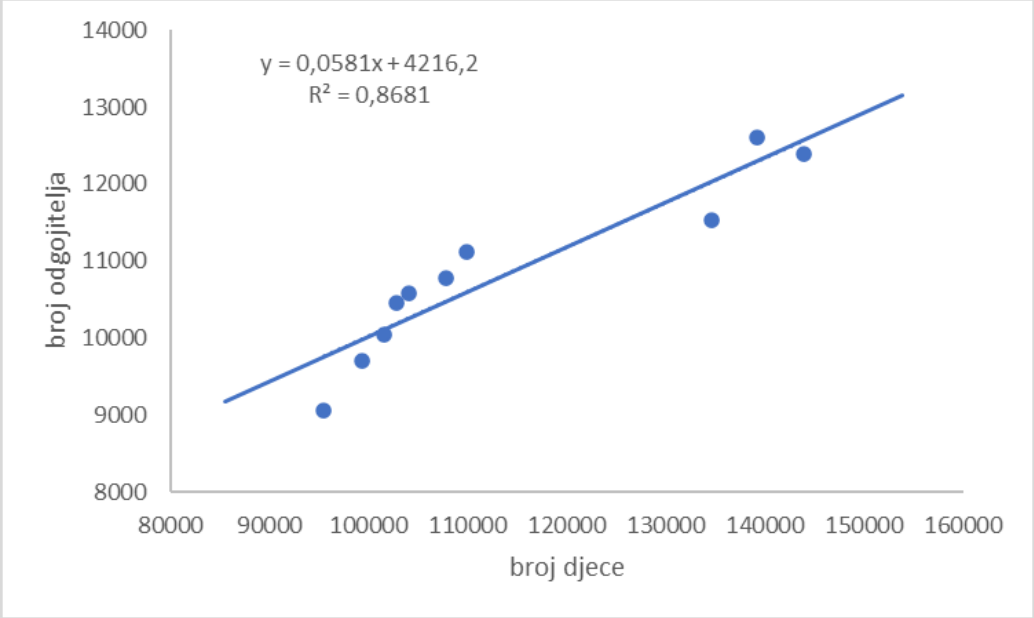
Iz dijagrama možemo pretpostaviti da, iako dani podatci nisu savršeno linearni, među varijablama postoji linearna ovisnost. Linearnom regresijom modelirajmo pravac koji najbolje opisuje dane podatke. Tražimo $y = \beta_1 x + \beta_0$ gdje je:

- y = broj odgojitelja (zavisna varijabla),
- x = broj djece (nezavisna varijabla),

- koeficijent β_1 pokazuje prosječnu promjenu broja odgojitelja kada se broj djece poveća za 1,
- koeficijent β_0 teorijski predstavlja broj odgojitelja kada je broj djece 0, što u stvarnom svijetu nema smisla promatrati.

Ako je β_1 pozitivan, povećanjem broja djece se povećava i broj odgojitelja, što je u skladu s očekivanjima u stvarnom svijetu. Ovdje iz grafa možemo očitati da će vodeći koeficijent pravca biti pozitivan te da će veza među varijablama biti pozitivna.

Da bismo odredili pravac linearne regresije, potrebno je izračunati (2.14). Međutim, ako imamo više podataka mjerenja, taj proces računanja je iscrpljujuć te lako možemo pogriješiti. Zbog toga se češće koriste tehnička pomagala ili računalni programi koji olakšavaju i ubrzavaju ovaj postupak. Mi ćemo pri analiziranju ovog slučaja koristiti program Excel koji će nam olakšati traženje jednadžbe regresijskog pravca i vrijednosti koeficijenta determinacije.



Jednažba pravca linearne regresije je $y = 0.0581x + 4216.2$, a koeficijent determinacije R^2 iznosi 0.8681. Koeficijent determinacije pokazuje da je 86.81% varijacija zavisne varijable objašnjeno dobivenim modelom, dok je samo 13.19% ostalo neobjašnjeno. Dakle, model dobro reprezentira dane podatke i veza među varijablama je čvrsta.

Na temelju ovog modela možemo predvidjeti broj odgojitelja koji će biti potreban za određeno povećanje broja djece za neku od narednih godina. Takav model predstavlja koristan alat za istraživače, jer im omogućuje da precizno odrede koliko zaposlenika je

potrebno u skladu s promjenama u broju djece. Na primjer, želimo znati koliko će nam biti potrebno odgojitelja kada bi bilo upisano 160000 djece. Jednostavno ćemo uvrstiti broj djece u jednadžbu regresijskog pravca te na taj način predvidjeti potreban broj odgojitelja. S obzirom da se povećao broj djece, možemo očekivati i povećanje broja odgojitelja. S druge strane, ako se s godinama smanji broj djece, bit će potrebno i manje odgojitelja.

Bibliografija

[1] M. Bašić, Ž. Milin Šipuš, predavanja kolegija *Metodika nastave matematike*, Prirodoslovno–matematički fakultet Sveučilišta u Zagrebu, Zagreb, 2023.

[2] M. Benšić, N. Šuvak, *Primijenjena statistika*, Osijek, 2013.

[3] Z. Bobbitt, *4 Examples of Using Linear Regression in Real Life*, dostupno na `https://www.statology.org/linear-regression-real-life-examples/` (pristupljeno 6.11.2024.)

[4] R. B. Darlington, A. F. Hayes, *Regression Analysis and Linear Models: Concepts, Applications and Implementation*, Guilford Press, 1990.

[5] Z. Drmač, V. Hari, M. Marušić, M. Rogina, S. Singer, S. Singer, *Numerička analiza*, Zagreb, 2013.

[6] M. Huzak, skripta s predavanja *Vjerojatnost i matematička statistika*, Prirodoslovno–matematički fakultet Sveučilišta u Zagrebu, Zagreb, 2006.

[7] D. C. Lay, S. R. Lay, J. J. McDonald, *Linear Algebra and Its Applications*, Pearson, 2016.

[8] E. A. Pack, D. C. Montgomery, G. G. Vining, *Introduction to Linear Regression Analysis (5th Edition)*, Wiley, 2012.

[9] J. Perkov, *Regresija i korelacija*, dostupno na `https://www.unizd.hr/portals/4/nastavni_mat/2_godina/statistika/10_predavanje.pdf` (pristupljeno 5.11.2024.)

[10] B. Singh, *Model Evaluation Metrics Used For Regression*, dostupno na `https://www.linkedin.com/pulse/model-evaluation-metrics-used-regression-brijesh-singh` (pristupljeno 5.11.2024.)

[11] T. Škrinjarić, *Linearni regresijski model*, Hrvatska narodna banka, Zagreb, 2023. dostupno na https://www.hnb.hr/repec/hnb/druge/pdf/d-001.pdf (pristupljeno 5.11.2024.)

[12] I. Šošić, *Primijenjena statistika*, Školska knjiga, Zagreb, 2004.

[13] Materijali kolegija *Vjerojatnost i statistika*, Prirodoslovno–matematički fakultet Sveučilišta u Zagrebu, Zagreb, dostupni na https://www.pmf.unizg.hr/images/50025978/VIS_vjezbe.pdf

[14] https://360digitmg.com/blog/simple-linear-regression (pristupljeno 6.11.2024.)

[15] https://devopedia.org/regression-modelling (pristupljeno 4.11.2024.)

Sažetak

Ovaj diplomski rad istražuje jednu od najčešće korištenih metoda u statistici — linearnu regresiju. Nakon uvođenja osnovnih pojmova koji su ključni za razumijevanje ove tehnike napravljena je detaljna analiza modela jednostavne linearne regresije i najefikasnije metode procjene parametara, metode najmanjih kvadrata. Zatim je predstavljen model višestruke linearne regresije i primjer linearizacije funkcije. Daljnje istraživanje usmjereno je na matričnu formulaciju metode najmanjih kvadrata. Na samom kraju, dani su primjeri primjene ove metode u stvarnom životu.

Summary

This thesis delves into one of the most commonly used methods in statistics – the linear regression model. After introducing the basic concepts essential for understanding this technique, a detailed analysis of the simple linear regression model and the most efficient parameter estimation method, the least squares method, is provided. Next, a multiple linear regression model and an example of function linearization are presented. Further investigation focuses on the matrix formulation of the least squares method. Finally, examples of the method's application in real-life scenarios are presented.

Životopis

Veronika Obšivač rođena je 13. studenoga 1998. godine u Metkoviću. Osnovnoškolsko obrazovanje završila je 2013. godine u Osnovnoj školi Stjepana Radića u Metkoviću te potom upisuje prirodoslovno–matematički smjer u Gimnaziji Metković. Godine 2017. na Prirodoslovno–matematičkom fakultetu Sveučilišta u Zagrebu upisuje preddiplomski sveučilišni studij Matematika; smjer nastavnički. Godine 2022. stječe akademski naziv sveučilišne prvostupnice edukacije matematike. Iste godine upisuje diplomski sveučilišni studij Matematika; smjer nastavnički.