# Bayesovsko zaključivanje za linearnu regresiju

**Kolar, Agata**

**SVEUČILIŠTE U ZAGREBU**

**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**

**MATEMATIČKI ODSJEK**

Agata Kolar

# BAYESOVSKO ZAKLJUČIVANJE ZA LINEARNU REGRESIJU

Diplomski rad

Voditelj rada:
doc. dr. sc. Snježana Lubura Strunjak

Zagreb, 2023

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. _____

*Mojoj obitelji, posebno mami i tati*

# Sadržaj

# Uvod

Statistika je znanstvena disciplina koja povezuje prikupljanje, registriranje i opisivanje podataka te njihovu analizu i tumačenje kako bismo u konačnici mogli odgovoriti na određena pitanja vezana uz područje istraživanja od našeg interesa. To uključuje odabir metoda za prikupljanje podataka relevantnih za postojeća pitanja, izbor metoda za prikaz i interpretaciju podataka te izbor metoda koje nam omogućavaju donošenje zaključaka koji su potkrijepljeni podacima. Podaci gotovo uvijek sadrže nesigurnost. Nesigurnost može proizaći iz odabira karakteristike koja se mjeri, ali i iz varijabilnosti procesa mjerenja. Izvođenje općih zaključaka iz podataka osnova je za svako racionalno znanstveno istraživanje. Statističko zaključivanje daje nam metode i alate kako bismo unatoč prisutnosti nesigurnosti podataka ipak mogli protumačiti informacije sadržane u podacima i ocijeniti nova saznanja dobivena iz tih podataka. Metode koje se koriste za analizu ovise o načinu na koji su podaci prikupljeni. Od ključne je važnosti da postoji vjerojatnosni model koji objašnjava komponentu nesigurnosti u podacima.

Postoje dva glavna pristupa statističkoj analizi: frekvencionistički i Bayesovski. Osnovna razlika je u tumačenju vjerojatnosti. Frekvencionistički, klasični pristup, vjerojatnost shvaća kao limes relativnih frekvencija ponovljenog pokusa. Parametri, to jest brojčane karakteristike populacije, su prema ovom pristupu fiksne, ali nepoznate konstante, te se stoga o njihovoj vrijednosti ne mogu donositi vjerojatnosni iskazi. Umjesto toga, uzorak se vadi iz populacije i računa se statistika uzorka. Distribucija vjerojatnosti statistike određuje se pomoću svih mogućih nasumičnih uzoraka iz populacije i naziva se distribucija uzorkovanja (engl. *sampling distribution*) statistike. Parametar populacije će biti parametar distribucije uzorkovanja, a izjava o vjerojatnosti, koja se može donijeti o statistici na temelju njezine distribucije uzorkovanja, pretvara se u izjavu o pouzdanosti parametra.

Alternativni pristup statističkoj analizi je Bayesov pristup. Prema Bayesovskom pristupu zakoni vjerojatnosti se primjenjuju izravno na problem, a vjerojatnost je subjektivna i ovisi o našim prethodnim uvjerenjima. Frekvencija se interpretira kao stupanj vjerovanja ili stupanj uvjerenja koji se temelji na dostupnim informacijama o događaju. Za razliku od frekvencionističkog pristupa, nepoznate parametre u statističkom modelu smatramo slučajnim varijablama. Subjektivna vjerovanja o nepoznatom parametru su sadržana

u apriornoj distribuciji parametra. Nakon što dobijemo podatke, svoja uvjerenja o parametrima revidiramo pomoću Bayesovog teorema.

Bayesov teorem dobio je naziv prema engleskom matematičaru i teologu Thomasu Bayesu (1702.-1761.) koji je prvi dokazao specijalni slučaj teorema u svome radu *An Essay Towards Solving a Problem in the Doctrine of Chances*. Rad je posthumno dao objaviti njegov prijatelj Richard Price 1763. godine. Neovisno o Bayesu, francuski matematičar i astronom Pierre-Simon Laplace 1774. godine dolazi do istog rezultata, ali u općenitijem obliku. Bayesove metode su u 19. stoljeću prihvatili Laplace i drugi znanstvenici, međutim te metode su početkom 20. stoljeća uglavnom pale u zaborav. Do sredine 20. stoljeća interes za Bayesove metode je ponovno porastao i danas se Bayesovska statistika naširoko primjenjuje u raznim područjima kao što su ekonomija, biomedicina, sociologija i inženjerstvo.

Cilj ovoga rada je opisati metode Bayesovskog zaključivanja za linearnu regresiju. Linearna regresija je statistička metoda koja se koristi za modeliranje odnosa između varijable odaziva i jedne ili više eksplanatornih varijabli. Razvijena je u 19. stoljeću i smatra se jednom od osnovnih i najkorištenijih statističkih metoda za analizu podataka. Bayesovska linearna regresija je statistička metoda koja kombinira Bayesovske statističke principe s tradicionalnom linearnom regresijom u svrhu izvođenja zaključaka o nepoznatim parametrima modela i predviđanja budućih vrijednosti odaziva za zadane vrijednosti eksplanatornih varijabli.

# Poglavlje 1

# Osnovni pojmovi

Za početak uvodimo ključne definicije i rezultate iz teorije vjerojatnosti i matematičke statistike koji su bitni za razumijevanje temeljnih koncepata iz Bayesovske statistike. Definicije, teoremi i ostali rezultati su preuzeti iz [2] i [7].

## 1.1 Vjerojatnosni prostor

*Pokus* je svaka dobro definirana procedura. Rezultati pokusa nazivaju se *ishodi* ili *elementarni događaji*, a skup svih ishoda pokusa zove se *prostor elementarnih događaja* (engl. sample space). Prostor elementarnih događaja tradicionalno se označava s $\Omega$, dok se elementarni događaji označuju s $\omega$, sa ili bez indeksa.

*Slučajni pokus* je pokus koji ima više mogućih ishoda. Intuitivno, slučajni pokus je svaki pokus (proces, opažanje, mjerenje) kojemu se ishod ne može sa sigurnošću predvidjeti. *Događaj* (u $\Omega$) je podskup prostora elementarnih događaja $\Omega$. Događaje najčešće označavamo velikim početnim slovima abecede, sa ili bez indeksa. Specijalno, $\Omega$ i prazan skup $\emptyset$ su događaji. Ako su $A$ i $B$ događaji, tada su i

$$A \cup B, A \cap B, A \setminus B, A^c = \Omega \setminus A$$

događaji. Uz navedeno, prebrojive unije i prebrojivi presjeci događaja su također događaji.

Označimo s $\mathcal{F}$ familiju svih događaja. Vrijedi $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, gdje $\mathcal{P}(\Omega)$ označava partitivni skup od $\Omega$. Familija događaja mora biti zatvorena na komplementiranje, prebrojive unije i presjeke te mora sadržavati $\Omega$. U nastavku iznosimo formalnu definiciju familije događaja.

**Definicija 1.1.1.** *Neka je $\Omega$ neprazan skup. Familija podskupova $\mathcal{F}$ od $\Omega$ zove se $\sigma$-algebra ako vrijede sljedeća tri svojstva:*

*(i)* $\Omega \in \mathcal{F}$;

*(ii) (zatvorenost na komplement) Ako je* $A \in \mathcal{F}$*, onda je i* $A^c \in \mathcal{F}$*;*

*(iii) (zatvorenost na prebrojive unije) Ako su* $A_j \in \mathcal{F}$*,* $j \in \mathbb{N}$*, onda je i* $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$*.*

*Uređen par* $(\Omega, \mathcal{F})$ *zove se izmjeriv prostor.*

U sljedećoj propoziciji ćemo navesti osnovna svojstva $\sigma$-algebra.

**Propozicija 1.1.2.** *Neka je* $\mathcal{F}$ *$\sigma$-algebra na nepraznom skupu* $\Omega$*. Tada vrijedi:*

*(i)* $\emptyset \in \mathcal{F}$*;*

*(ii) (zatvorenost na prebrojive presjeke) Ako su* $A_j \in \mathcal{F}$*,* $j \in \mathbb{N}$*, onda je i* $\bigcap_{j=1}^{\infty} A_j \in \mathcal{F}$*;*

*(iii) (zatvorenost na konačne unije) Za svaki* $n \in \mathbb{N}$*, ako su* $A_1, A_2, \ldots, A_n \in \mathcal{F}$*, onda je i* $\bigcup_{j=1}^{n} A_j \in \mathcal{F}$*.*

Nakon što smo definirali familiju događaja i nabrojali neka osnovna svojstva takve familije, spremni smo navesti definiciju vjerojatnosti.

**Definicija 1.1.3.** *Neka je* $\Omega$ *neprazan skup i* $\mathcal{F}$ *$\sigma$-algebra događaja. Vjerojatnost na izmjerivom prostoru* $(\Omega, \mathcal{F})$ *je funkcija* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *koja zadovoljava sljedeća tri aksioma*

*(A1) (nenegativnost) Za sve* $A \in \mathcal{F}$*,* $\mathbb{P}(A) \geq 0$*;*

*(A2) (normiranost)* $\mathbb{P}(\Omega) = 1$*;*

*(A3) (*$\sigma$*-aditivnost) Za svaki niz* $(A_j)_{j \in \mathbb{N}}$ *po parovima disjunktnih događaja* $A_j \in \mathcal{F}$ *(* $A_i \cap A_j = \emptyset$ *za* $i \neq j$*) vrijedi*

$$\mathbb{P}\Big( \bigcup_{j=1}^{\infty} \Big) = \sum_{j=1}^{\infty} \mathbb{P}(A_j).$$

*Uređena trojka* $(\Omega, \mathcal{F}, \mathbb{P})$ *zove se vjerojatnosni prostor.*

U idućoj propoziciji navodimo neka od svojstava vjerojatnosti koja slijede iz aksioma $(A1) - (A3)$.

**Propozicija 1.1.4.** *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor. Tada vrijedi:*

*(i)* $\mathbb{P}(\emptyset) = 0$*;*

*(ii) (konačna aditivnost) Za svaki* $n \in \mathbb{N}$ *i svaki konačan niz* $A_1, A_2, \ldots, A_n$ *po parovima disjunktnih događaja iz* $\mathcal{F}$ *vrijedi*

$$\mathbb{P}\Big(\bigcup_{j=1}^{n}\Big) = \sum_{j=1}^{n} \mathbb{P}(A_j);$$

*(iii)* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;

*(iv)* $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$;

*(v) (monotonost) Ako su* $A, B \in \mathcal{F}$ *i* $A \subseteq B$*,onda je* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Za iskaz i razumijevanje Bayesovog teorema, koji je temelj za Bayesovsku statistiku, važnu ulogu će imati pojam uvjetne vjerojatnosti.

**Definicija 1.1.5.** *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor te* $B \in \mathcal{F}$ *događaj takav da je* $\mathbb{P}(B) > 0$*. Uvjetna vjerojatnost događaja A uz dano B definira se formulom*

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Uočimo da iz definicije uvjetne vjerojatnosti odmah slijedi da je

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B).$$

Ako je i $\mathbb{P}(A) > 0$, onda je dobro definirana uvjetna vjerojatnost $\mathbb{P}(B \mid A)$ te vrijedi

$$\mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A). \tag{1.1}$$

Za događaje $A$ i $B$ pozitivne vjerojatnosti može vrijediti da događanje jednog ni na koji način ne utječe na vjerojatnost drugog događaja, i obratno. To znači da je $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, odnosno $\mathbb{P}(B \mid A) = \mathbb{P}(B)$. Svaka od tih jednakosti je ekvivalentna s $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ te ćemo u tom slučaju reći da su događaji $A$ i $B$ nezavisni.

**Definicija 1.1.6.** *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor. Događaji* $A, B \in \mathcal{F}$ *su nezavisni ako vrijedi*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

## 1.2 Slučajne varijable

**Definicija 1.2.1.** *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor. Funkcija* $X : \Omega \to \mathbb{R}$ *zove se diskretna slučajna varijabla ako postoji prebrojiv skup* $D = \{a_1, a_2, \ldots\} \subset \mathbb{R}$ *takav da je*

*(i)* $X(\omega) \in D$ *za sve* $\omega \in \Omega$;

*(ii)* $\{X = a_j\} = \{\omega \in \Omega : X(\omega) = a_j\} \in \mathcal{F}$ *za sve* $j \in \mathbb{N}$.

**Definicija 1.2.2.** *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor. Slučajna varijabla na* $(\Omega, \mathcal{F}, \mathbb{P})$ *je svaka funkcija* $X : \Omega \to \mathbb{R}$ *takva da vrijedi*

$$\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}, \quad \text{za sve } x \in \mathbb{R}.$$

**Napomena 1.2.3.** *Neka je* $\Omega$ *konačan ili prebrojiv te* $\mathcal{F} = \mathcal{P}(\Omega)$*. Tada je svaka funkcija* $X : \Omega \to \mathbb{R}$ *diskretna slučajana varijabla.*

**Definicija 1.2.4.** *Neka je X slučajna varijabla definirana na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$*. Funkcija distribucije (ili funkcija raspodjele) od X je funkcija* $F : \mathbb{R} \to [0, 1]$ *definirana sa*

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

**Teorem 1.2.5.** *Neka je X slučajna varijabla definirana na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *te neka je F pripadna funkcija distribucije. Tada vrijedi*

*(i) F je neopadajuća;*

*(ii) F je neprekidna zdesna u svakoj točki* $x \in \mathbb{R}$*;*

*(iii) F ima limes s lijeva u svakoj točki* $x \in \mathbb{R}$*;*

*(iv)* $F(-\infty) := \lim_{x \to -\infty} F(x) = 0$ *i* $F(\infty) := \lim_{x \to \infty} F(x) = 1$.

**Definicija 1.2.6.** *Neka je* $X : \Omega \to D = \{a_1, a_2, \ldots\} \subset \mathbb{R}$ *diskretna slučajna varijabla definirana na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$*. Funkcija* $f : \mathbb{R} \to [0, \infty\rangle$ *definirana sa*

$$f(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

*zove se diskretna (vjerojatnosna) funkcija gustoće slučajne varijable X.*

**Napomena 1.2.7.** *Ako je* $f : \mathbb{R} \to [0, \infty\rangle$ *funkcija gustoće neke diskretne slučajne varijable, tada vrijedi sljedeće:*

*(i)* $f(x) = 0$ *za sve* $x \notin D$*;*

*(ii)* $\sum_{x \in \mathbb{R}} f(x) = \sum_{x \in D} f(x) = 1$.

**Teorem 1.2.8.** *Neka je X diskretna slučajna varijabla definirana na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *te neka su F i f, redom, pripadne funkcije distribucije i gustoće. Tada vrijedi* $F(x) = \sum_{y \leq x} f(x)$.

**Definicija 1.2.9.** *Slučajna varijabla $X : \Omega \to \mathbb{R}$ je apsolutno neprekidna ako postoji $f : \mathbb{R} \to [0, \infty\rangle$ takva da za sve $x \in \mathbb{R}$ vrijedi*

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

*Funkcija f zove se funkcija gustoće od X.*

**Napomena 1.2.10.** *Ako je $f : \mathbb{R} \to [0, \infty\rangle$ funkcija gustoće neke neprekidne slučajne varijable vrijedi sljedeće:*

*(i)* $f(x) = 0$ *za sve* $x \in \mathbb{R}$*;*

*(ii)* $\int_{-\infty}^{\infty} f(x)dx = 1.$

## Zajednička razdioba slučajnih varijabli

**Definicija 1.2.11.** *Neka su X i Y dvije slučajne varijable definirane na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Uređeni par $(X, Y)$ zove se dvodimenzionalni slučajni vektor.*

Ako su $X_1, \ldots, X_n$ slučajne varijable definirane na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$, na analogan način definiramo $n$-dimenzionalni slučajni vektor $(X_1, \ldots, X_n)$.

**Definicija 1.2.12.** *Neka je $(X, Y)$ dvodimenzionalni diskretni slučajni vektor definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Funkcija distribucije slučajnog vektora $(X, Y)$ (ili zajednička razdioba dviju slučajnih varijabli X i Y) je funkcija $F : \mathbb{R}^2 \to [0, 1]$ definirana s*

$$F(x, y) = \mathbb{P}((X, Y) \leq (x, y)), \quad (x, y) \in \mathbb{R}^2.$$

**Definicija 1.2.13.** *Dvodimenzionalni slučajni vektor $(X, Y) : \Omega \to \mathbb{R}^2$ je neprekidan slučajan vektor ako postoji funkcija $f : \mathbb{R}^2 \to [0, \infty\rangle$ takva da za sve $(x, y) \in \mathbb{R}^2$ vrijedi*

$$F(x, y) = \mathbb{P}((X, Y) \leq (x, y)) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)dudv$$

*Funkciju f zovemo gustoćom (ili funkcijom gustoće) neprekidnog slučajnog vektora $(X, Y)$*

Ako želimo naglasiti da je riječ o funkciji distribucije i funkciji gustoće slučajnog vektora $(X, Y)$, možemo pisati $F_{X,Y}$ i $f_{X,Y}$, redom.

**Napomena 1.2.14.** *Svojstva zajedničke gustoće neprekidnih varijabli iz prethodne definicije su:*

*(i)* $f(x, y) \geq 0$ *za sve* $x, y$*;*

*(ii)* $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1.$

Ako je $f_X$ funkcija gustoće neprekidne slučajne varijable $X$ i $f_Y$ funkcije gustoće neprekidne slučajne varijable $Y$, onda za $x, y \in \mathbb{R}$ vrijedi

$$\int_{-\infty}^{\infty} f(x, y)dy = f_X(x) \quad \text{i} \quad \int_{-\infty}^{\infty} f(x, y)dx = f_Y(y)$$

te

$$\lim_{y\to\infty} F(x, y) = F_X(x) \quad \text{i} \quad \lim_{x\to\infty} F(x, y) = F_Y(y).$$

Funkcije $f_X$ i $f_Y$ se zovu marginalne funkcije gustoće slučajnog vektora $(X, Y)$, a funkcije $F_X$ i $F_Y$ marginalne raspodjele (distribucije) od $(X, Y)$.

**Definicija 1.2.15.** *Neka su X i Y neprekidne slučajne varijable definirane na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$*, s funkcijama gustoće* $f_X$ *i* $f_Y$ *. Nadalje, neka je f funkcija gustoće od* $(X, Y)$*. Kažemo da su X i Y nezavisne ako vrijedi* $f(x, y) = f_X(x)f_Y(y)$ *za sve* $x, y \in \mathbb{R}$*.*

**Definicija 1.2.16.** *Neka je* $(X, Y)$ *dvodimenzionalni slučajni vektor definiran na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$*, s funkcijom gustoće f te neka su* $f_X$ *i* $f_Y$ *pripadne marginalne gustoće. Uvjetna funkcija gustoće slučajne varijable X uz dano* $Y = y$ *definira se s*

$$f_{X|Y}(x|y) := \frac{f(x, y)}{f_Y(y)}, \quad x \in \mathbb{R},$$

*za sve* $y \in \mathbb{R}$ *za koje je* $f_Y(y) > 0$*.*

Uočimo, ako su $X$ i $Y$ nezavisne, onda je $f_{X|Y} = f_X$. Zaista

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x),$$

gdje smo u drugoj jednakosti iskoristili nezavisnost.

## Matematičko očekivanje

**Definicija 1.2.17.** *Neka je $X$ slučajna varijabla s gustoćom $f$. Ako je $\sum_{x\in\mathbb{R}} |x|f(x) < \infty$ kada je $X$ diskretna slučajna varijabla, odnosno $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$ kada je $X$ neprekidna slučajna varijabla, tada kažemo da $X$ ima matematičko očekivanje koje je definirano kao*

$$\mathbb{E}(X) := \sum_{x\in\mathbb{R}} xf(x) \qquad \textit{(ako je X diskretna)},$$

$$\mathbb{E}(X) := \int_{-\infty}^{\infty} xf(x)dx \qquad \textit{(ako je X neprekidna)}.$$

Svojstva matematičkog očekivanja su:

(i) $\mathbb{E}(\lambda X) = \lambda\mathbb{E}(X), \quad \lambda \in \mathbb{R}$;

(ii) $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

## Varijanca i standardna devijacija

**Definicija 1.2.18.** *Neka je $X$ slučajna varijabla s očekivanjem $\mathbb{E}[X]$. Varijanca od $X$ se definira kao srednje kvadratno odstupanje $X$ od $\mathbb{E}[X]$, to jest*

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

*Standardna devijacija od $X$, u oznaci $\sigma(X)$, je drugi korijen varijance:*

$$\sigma(X) := \sqrt{Var(X)}$$

Korištenjem linearnosti matematičkog očekivanja može se pokazati da je

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

Svojstva varijance jesu:

(i) $Var(\lambda X) = \lambda^2 Var(X), \quad \lambda \in \mathbb{R}$;

(ii) $Var(X + \lambda) = Var(X), \quad \lambda \in \mathbb{R}$.

**Definicija 1.2.19.** *Neka su $X$ i $Y$ dvije slučajne varijable takve da je $\mathbb{E}(X^2) < \infty$ i $\mathbb{E}(Y^2) < \infty$. Kovarijanca od $X$ i $Y$ definira se kao:*

$$Cov(X, Y) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

**Primjer 1.2.20.** *Normalna slučajna varijabla $X$ s parametrima $\mu \in \mathbb{R}$ i $\sigma^2 > 0$, u oznaci $X \sim N(\mu, \sigma^2)$, je neprekidna slučajna varijabla s funkcijom gustoće*

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

*Može se pokazati da vrijedi*

$$\mathbb{E}(X) = \mu \quad i \quad Var(X) = \sigma^2.$$

Sljedeći primjer je preuzet iz [5].

**Primjer 1.2.21.** *a) Za slučajan vektor $(X_1, X_2)$ kažemo da ima bivarijatnu normalnu razdiobu ako je $(X_1, X_2)$ dvodimenzionalan slučajan vektor s funkcijom gustoće*

$$f(x_1, x_2|\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \\ \cdot exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\},$$

*za $x_1, x_2 \in \mathbb{R}$, gdje je $\sigma_1, \sigma_2 > 0$, te $-1 < \rho < 1$ definiran s*

$$\rho = \frac{\mathbb{E}\left[(X_1-\mu_1)(X_2-\mu_2)\right]}{\sigma_1 \cdot \sigma_2}$$

*i zovemo ga koeficijentom korelacije. Uočimo da je $Cov(X_1, X_2) = \rho\sigma_1\sigma_2$. Zapišimo sada gustoću bivarijatne normalne slučajne varijable u matričnoj formi. Vektor opažanja slučajnih varijabli, pripadni vektor očekivanja i kovarijacijska matrica su:*

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

*Inverz kovarijacijske matrice u bivarijatnom normalnom slučaju je dan s*

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho)^2}\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

*Determinanta bivarijatne normalne kovarijacijske matrice je $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho)^2$.*

*Sada je zajednička funkcija gustoće bivarijatne normalne slučajne varijable u matričnom obliku dana s*

$$f(x_1, x_2|\mu, \Sigma) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}. \tag{1.2}$$

*b) Dimenzije vektora opaženih vrijednosti* $\mathbf{x}$ *i vektora očekivanja* $\mu$ *mogu biti* $k \geq 2$*. U tom slučaju možemo poopćiti distribuciju iz podzadatka a) koja je zadana gustoćom u* (1.2)*.*

*Za slučajni vektor* $(X_1, X_2, \ldots, X_k)$*,* $k \geq 2$ *kažemo da ima multivarijatnu normalnu razdiobu ako je* $(X_1, X_2, \ldots, X_k)$ *k-dimenzionalan slučajan vektor s funkcijom gustoće*

$$f(x_1, x_2, \ldots, x_k|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)},$$

*gdje su parametri distribucije, vektor očekivanja* $\mu$ *i kovarijacijska matrica* $\Sigma$*, redom dani s*

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \ldots & \rho_{1k}\sigma_1\sigma_k \\ \vdots & \ddots & \vdots \\ \rho_{k1}\sigma_k\sigma_1 & \ldots & \sigma_k^2 \end{bmatrix}.$$

## 1.3 Osnovni pojmovi matematičke statistike

Za kraj ovog poglavlja donosimo nekoliko osnovnih definicija iz područja matematičke statistike. Definicije su preuzete iz [3] i [4].

**Definicija 1.3.1.** *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor te* $X$ *slučajna varijabla s funkcijom distribucije* $F$*. Slučajne varijable* $X_1, X_2, \ldots, X_n$ *čine slučajan uzorak duljine* $n$ *iz distribucije* $F$ *ako su* $X_1, X_2, \ldots, X_n$ *nezavisne, jednako distribuirane slučajne varijable s funkcijom distribucije* $F$*. Realizaciju slučajnog uzorka, odnosno opažene vrijednosti* $x_i$ *od* $X_i$*,* $i = 1, 2, \ldots, n$ *zovemo uzorkom.*

**Definicija 1.3.2.** *Neka je* $(\Omega, \mathcal{F})$ *izmjeriv prostor i* $\mathcal{P}$ *neka familija vjerojatnosti na* $(\Omega, \mathcal{F})$*. Uređena trojka* $(\Omega, \mathcal{F}, \mathcal{P})$ *naziva se statistička struktura. Familija* $\mathcal{P}$ *često je parametrizirana konačnodimenzionalnim parametrom* $\theta$ *i zapisuje se u obliku* $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$*, pri čemu je* $\Theta$ *podskup od* $\mathbb{R}^k$ $(k \geq 1)$ *koji zovemo parametarski prostor.*

Neka je $X : \Omega \to \mathbb{R}^d$ neprekidna ili diskretna slučajna veličina (varijabla ili vektor) dimenzije $d$ $(d \geq 1)$ definirana na parametriziranoj statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ s gustoćom $f(\cdot|\theta) \equiv f_\theta : \mathbb{R}^d \to [0, \infty\rangle$ u odnosu na vjerojatnost $\mathbb{P}_\theta$. Kako je navedeno u [3], $\mathcal{P}$ tada možemo poistovjetiti s množinom gustoća

$$\mathcal{P} \equiv \{f(\cdot|\theta) : \theta \in \Theta\}.$$

Neka je $X_1, X_2, \ldots, X_n$ slučajan uzorak iz razdiobe $F$ te neka je $f$ pripadna funkcija gustoće. Funkcijom $f(\mathbf{x}|\theta) = f(x_1, x_2, ..., x_n|\theta)$ označavamo zajedničku gustoću slučajnog uzorka $X_1, X_2, \ldots, X_n$ s parametrom $\theta \in \Theta$, gdje je $\Theta$ parametarski prostor.

**Definicija 1.3.3.** *Statistika na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *je svaka slučajna varijabla (ili slučajni vektor za* $k \geq 2$*)* $T : \Omega \to \mathbb{R}^k$ *takva da za neki* $n \in \mathbb{N}$ *postoji n-dimenzionalni slučajni vektor* $(X_1, X_2, \ldots, X_n)$ *na* $(\Omega, \mathcal{F}, \mathcal{P})$ *te izmjeriva funkcija* $t : \mathbb{R}^n \to \mathbb{R}^k$ *takva da je* $T = t(X_1, X_2, \ldots, X_n)$ *.*

**Definicija 1.3.4.** *Statistika* $T = t(X_1, X_2, \ldots, X_n)$ *dimenzije* $k$ $(k \geq 1)$ *je dovoljna za* $\theta$ *ako za svako* $y \in \mathbb{R}^k$ *za koje postoji uvjetna razdioba slučajnog uzorka* $(X_1, X_2, \ldots, X_n)$ *uz uvjet* $T = y$*, ta uvjetna razdioba ne ovisi o parametru* $\theta$*.*

**Definicija 1.3.5.** *Neka je* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$*,* $\mathbf{x} \in \mathbb{R}^n$ *jedna realizacija slučajnog uzorka* $(X_1, X_2, \ldots, X_n)$ *sa zajedničkom funkcijom gustoće* $f(\mathbf{x}|\theta)$ *te neka je* $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k$ *nepoznati parametar. Funkcija vjerodostojnosti je funkcija* $L : \Theta \to \mathbb{R}$ *definirana s*

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta), \quad \theta \in \Theta.$$

**Primjer 1.3.6.** *Odredimo funkciju vjerodostojnosti normalne slučajne varijable s parametrima* $\mu$ *i* $\sigma^2$*. Neka je* $x_1, x_2, \ldots, x_n$ *opaženi uzorak iz* $N(\mu, \sigma^2)$*. Tada je*

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(\sqrt{\sigma^2}\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \quad .$$

# Poglavlje 2

# Bayesov teorem

Bayesov teorem je matematička formula koja se koristi u vjerojatnosti i statistici za računanje uvjetnih vjerojatnosti događaja. Ovaj teorem nam omogućava revidiranje vjerojatnosti nekog događaja nakon što smo dobili nove podatke ili saznanja o tome događaju. Bayesov teorem ima širok spektar primjena u različitim područjima, međutim, jedna od najvažnijih je primjena u Bayesovskoj statistici.

U nastavku navodimo dvije verzije Bayesovog teorema: diskretnu i neprekidnu verziju. Zatim ćemo se upoznati s osnovnim idejama Bayesovskog zaključivanja gdje ćemo pokazati kako pomoću Bayesovog teorema ažuriramo prethodna uvjerenja ili znanja na temelju novih informacija. Kao izvor za ovo poglavlje korišteni su [5], [7], [8] i [9].

## 2.1 Diskretna verzija Bayesovog teorema

Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A, B \in \mathcal{F}$ događaji takvi da je $\mathbb{P}(A) > 0$. Uvjetna vjerojatnost od $B$ uz dano $A$ definira se kao

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{2.1}$$

Iz jednakosti $A = (A \cap B) \cup (A \cap B^c)$ i konačne aditivnosti slijedi

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c).$$

Uvrstimo dobiveni izraz u (2.1) za $\mathbb{P}(A)$ i dobijemo sljedeće:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)}.$$

Sada iskoristimo (1.1) i imamo

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c)}.$$

Ovdje je važno primijetiti da su skupovi $B$ i $B^c$ disjunktni te da u uniji čine cijeli skup $\Omega$. Poopćimo sada prethodnu formulu na proizvoljnu familiju disjunktnih događaja.

**Definicija 2.1.1.** *Neka je $(H_i)_{i\in I}$ konačna ili prebrojiva familija događaja iz $\mathcal{F}$ takva da je $\mathbb{P}(H_i) > 0$ za sve $i \in I$, $H_i \cap H_j = \emptyset$ za $i \neq j$, te $\cup_{i=1} H_i = \Omega$. Familiju $(H_i)_{i\in I}$ zovemo potpun sustav događaja.*

**Propozicija 2.1.2.** *(Formula potpune vjerojatnosti) Neka je $(H_i)_{i\in I}$ potpun sustav događaja. Tada za svaki $A \in \mathcal{F}$ vrijedi*

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(H_i)\mathbb{P}(A \mid H_i).$$

*Dokaz.*

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega) = \mathbb{P}(A \cap (\cup_{i=1}^{\infty} H_i)) = \mathbb{P}(\cup_{i\in I}(A \cap H_i)) \\ &= (\sigma - aditivnost) = \sum_{i\in I} \mathbb{P}(A \cap H_i) \overset{1.1}{=} \sum_{i\in I} \mathbb{P}(H_i)\mathbb{P}(A \mid H_i).\end{aligned}$$

□

**Teorem 2.1.3.** *(Bayesov teorem) Neka je $(H_i)_{i\in I}$ potpun sustav događaja na vjerojatnosno prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Tada za svaki $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$ vrijedi*

$$\mathbb{P}(H_j \mid A) = \frac{\mathbb{P}(H_j)\mathbb{P}(A \mid H_j)}{\sum_{i\in I} \mathbb{P}(H_i)\mathbb{P}(A \mid H_i)}.$$

*Dokaz.* Korištenjem definicije uvjetne vjerojatnosti u prvoj jednakosti, (1.1) u drugoj i formule potpune vjerojatnosti, dobivamo

$$\begin{aligned}\mathbb{P}(H_j \mid A) &= \frac{\mathbb{P}(H_j \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(H_j)\mathbb{P}(A \mid H_j)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(H_j)\mathbb{P}(A \mid H_j)}{\sum_{i\in I} \mathbb{P}(H_i)\mathbb{P}(A \mid H_i)}.\end{aligned}$$

□

Originalne vjerojatnosti hipoteza $\mathbb{P}(H_j)$ zovu se *apriorne vjerojatnosti*. To su vjerojatnosti koje pridružujemo hipotezama u nedostatku drugih informacija. Nakon što prikupimo dodatne informacije, apriorne vjerojatnosti modificiramo tako da uključimo novu informaciju. Dobivene vjerojatnosti $\mathbb{P}(H_j \mid A)$ zovu se *aposteriorne vjerojatnosti*.

**Primjer 2.1.4.** *Medicinski dijagnostički test za otkrivanje konkretne bolesti ima osjetljivost jednaku* 95%. *To znači da ako osoba ima tu bolest, vjerojatnost da će test dati pozitivan odgovor je* 0.95. *Specifičnost testa je* 0.90. *To znači da ako osoba nema tu bolest, vjerojatnost da će test dati negativan odgovor je* 0.90, *odnosno da je stopa lažno pozitivnih rezultata testa* 0.10. *Poznato je da u populaciji* 1% *ljudi ima bolest. Koja je vjerojatnost da testirana (slučajno odabrana) osoba ima tu bolest ako je rezultat testa pozitivan?*

*Neka je D događaj "osoba ima bolest" i neka je T događaj "test daje pozitivan rezultat". Prema Bayesovoj formuli slijedi*

$$\begin{aligned}\mathbb{P}(D|T) &= \frac{\mathbb{P}(T|D)\mathbb{P}(D)}{\mathbb{P}(T)} = \frac{\mathbb{P}(T|D)\mathbb{P}(D)}{\mathbb{P}(T|D)\mathbb{P}(D) + \mathbb{P}(T|D^c)\mathbb{P}(D^c)} \\ &= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.1 \cdot 0.99} = 0.08756.\end{aligned}$$

*Dakle, iako je osoba bila pozitivna na bolest, vjerojatnost da ona zaista ima bolest je oko* 8.756%. *Razlog tome je što smo uzeli u obzir visoke stope lažno pozitivnih rezultata uz činjenicu da je bolest relativno rijetka.*

## 2.2 Neprekidna verzija Bayesovog teorema

**Propozicija 2.2.1.** *(Neprekidna verzija formule potpune vjerojatnosti) Neka su X i Y slučajne varijable na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *s funkcijama gustoće* $f_X(x)$ *i* $f_Y(y)$ *te neka je A proizvoljan Borelov skup. Vrijedi*

$$\mathbb{P}(X \in A) = \int_{\mathbb{R}} \left( \int_A f_{X|Y}(x|y)dx \right) f_Y(y)dy.$$

*Dokaz.*

$$\begin{aligned}\mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in \mathbb{R}) &= \int_A \int_{\mathbb{R}} f_{X,Y}(x,y)dydx \\ &= \int_A \int_{\mathbb{R}} f_{X|Y}(x|y) f_Y(y)dydx = \int_{\mathbb{R}} \left( \int_A f_{X|Y}(x|y)dx \right) f_Y(y)dy.\end{aligned}$$

□

**Teorem 2.2.2.** *(Neprekidna verzija Bayesovog teorema) Neka su X i Y slučajne varijable takve da* $f_Y(y) > 0$. *Tada vrijedi*

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{\mathbb{R}} f_{Y|X}(y|x)f_X(x)dx}.$$

*Dokaz.* Prema definiciji uvjetne funkcije gustoće dobijemo da

$$f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x).$$

Sada slijedi

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

Druga jednakost iz iskaza teorema slijedi direktno iz propozicije 2.2.1. □

## 2.3 Bayesovsko zaključivanje

U ovom odjeljku ćemo pružiti temeljnu okosnicu za Bayesovsko statističko zaključivanje. Pretpostavimo da nas zanimaju vrijednosti $k$ ($k \geq 1$) nepoznatih parametara

$$\theta = (\theta_1, \ldots, \theta_k),$$

gdje je $\theta \in \Theta \subset \mathbb{R}^k$, te da imamo neka apriorna uvjerenja o njihovim vrijednostima koja možemo izraziti pomoću funkcije gustoće $g(\theta)$. Uz navedeno, pretpostavimo da smo dobili neke podatke relevantne za vrijednosti nepoznatih parametara. Točnije, pretpostavimo da imamo $n$ opažanja $\mathbf{x} = (x_1, \ldots, x_n)$. Podatke $\mathbf{x} = (x_1, \ldots, x_n)$ možemo shvatiti kao jednu realizaciju slučajnog vektora $\mathbf{X} = (X_1, \ldots, X_n)$ koji ima distribuciju vjerojatnosti, pri čemu potonja distribucija ovisi o $k$ nepoznatih parametara $\theta$. Dakle funkcija gustoće vektora $\mathbf{X}$ (neprekidna ili diskretna), u oznaci $f(\mathbf{x}|\theta)$, ovisi o vektoru nepoznatih parametara $\theta$ na poznati način.

**Napomena 2.3.1.** *Primijetimo da smo uveli neznatnu promjenu u notaciji. Koristimo $f()$ kako bismo označili distribuciju vjerojatnosti (uvjetnu ili bezuvjetnu) slučajnog vektora $\mathbf{X}$, dok $g()$ koristimo kako bismo označili distribuciju vjerojatnosti (uvjetnu ili bezuvjetnu) slučajnog vekora nepoznatih parametara $\theta$. Ovime zapravo želimo naglasiti razliku između slučajnog vektora čije vrijednosti opažamo i slučajnog vektora nepoznatih parametara o kojima želimo izvesti zaključke.*

Želimo pronaći način na koji bismo izrazili svoja vjerovanja o $\theta$ uzimajući u obzir naša prijašnja uvjerenja o nepoznatim parametrima, ali i dobivene podatke. Naravno, moguće je da se nečija prethodna uvjerenja o parametru $\theta$ razlikuju od naših, no vrlo često se svi mogu složiti oko načina na koji su podaci povezani s $\theta$, to jest oko oblika funkcije $f(\mathbf{x}|\theta)$. U tom slučaju postojat će razlika u dobivenim aposteriornim uvjerenjima. Međutim, može se pokazati da, ako uspijemo prikupiti dovoljnu količinu podataka, oba aposteriorna uvjerenja će uglavnom postati vrlo bliska.

Osnovni alat koji nam treba za revidiranje naših uvjerenja na temelju dobivenih podataka je Bayesov teorem. Prema teoremu vrijedi

$$g(\theta \mid \mathbf{x}) = \frac{g(\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x})}, \tag{2.2}$$

pri čemu je

$$f(\mathbf{x}) = \begin{cases} \sum f(\mathbf{x}|\theta) g(\theta), & \text{u diskretnom slučaju,} \\ \int f(\mathbf{x}|\theta) g(\theta) d\theta, & \text{u neprekidnom slučaju.} \end{cases}$$

Bayesov teorem često pišemo u njegovom proporcionalnom obliku kao

$$g(\theta|\mathbf{x}) \propto g(\theta) f(\mathbf{x}|\theta), \tag{2.3}$$

gdje je $1/f(\mathbf{x})$ konstanta proporcionalnosti koja je u (2.2) potrebna kako bi se aposteriorne uvjetne vjerojatnosti zbrojile, odnosno integrirale do jedinice.

Uz dane podatke $\mathbf{x}$, gustoću $f(\mathbf{x}|\theta)$ možemo promatrati kao funkciju parametara $\theta$, a ne kao funkciju uzorka $\mathbf{x}$. U tom slučaju funkciju $f(\mathbf{x}|\theta)$ nazivamo funkcijom vjerodostojnosti parametara $\theta$ za dani uzorak $\mathbf{x}$ i pišemo $L(\theta)$ te tada umjesto (2.3) možemo pisati

$$g(\theta|\mathbf{x}) \propto g(\theta) L(\theta).$$

S ovakvom interpretacijom funkcije $f(\mathbf{x}|\theta)$, ako $g(\theta)$ definiramo kao apriornu funkciju gustoće od $\theta$ i $g(\theta|\mathbf{x})$ kao aposteriornu funkciju gustoće za $\theta$ uz dano $\mathbf{x}$, o Bayesovom teoremu možemo razmišljati u pamtljivijem obliku kao

$$\text{aposteriorna gustoća} \propto \text{apriorna gustoća} \times \text{vjerodostojnost.}$$

Gornji odnos opisuje način prema kojem bismo trebali modificirati svoja uvjerenja uzimajući u obzir podatke koje imamo na raspolaganju. Član $g(\theta)$ predstavlja stupanj vjerovanja ili znanja o parametrima $\theta$ prije nego što smo prikupili podatke. Funkcija vjerodostojnosti $L(\theta)$ nam omogućava da izmijenimo apriorne pretpostavke o $\theta$ na temelju dobivenih podataka. Drugim riječima, vjerodostojnost predstavlja izmijenjena prethodna uvjerenja o nepoznatim parametrima (opisana aposteriornom funkcijom gustoće) u svjetlu relevantnih podataka koje smo prikupili. Na taj način dolazimo do aposteriornih uvjerenja, odnosno možemo izračunati funkciju gustoću $g(\theta|\mathbf{x})$ parametara $\theta$ uz dani uzorak $\mathbf{x}$.

Uočimo da zbog načina na koji pišemo Bayesov teorem sa znakom proporcionalnosti, rezultat se ne mijenja ako $L(\theta)$ pomnožimo s bilo kojom konstantom ili općenito bilo kojom funkcijom od $\mathbf{x}$. Sukladno tome, vjerodostojnost možemo poistovjetiti s bilo kojom funkcijom koja je jednaka $f(\mathbf{x}|\theta)$ do na multiplikativnu konstantu, a ne nužno s funkcijom koja je upravo $f(\mathbf{x}|\theta)$.

Također, dodatna mogućnost ove metode je što se može primijeniti sekvencijalno. To znači da nam Bayesov teorem omogućava revidiranje informacije o $\theta$ kako naš uzorak raste. Dakle, ako imamo početni uzorak opažanja $\mathbf{x}_1$ iz distribucije slučajne varijable $X$, prema Bayesovoj formuli vrijedi

$$g(\theta|\mathbf{x}_1) \propto g(\theta)L_1(\theta),$$

gdje je $L_1(\theta)$ funkcija vjerodostojnosti parametara $\theta$ za dani uzorak $\mathbf{x}_1$. Sada pretpostavimo da imamo drugi skup opažanja $\mathbf{x}_2$ iz iste distribucije slučajne varijable $X$, koji je nezavisan o prvom uzorku $\mathbf{x}_1$. Tada

$$g(\theta|\mathbf{x}_1, \mathbf{x}_2) \propto g(\theta)L_{1,2}(\theta), \tag{2.4}$$

pri čemu je $L_{1,2}(\theta)$ funkcija vjerodostojnosti parametara $\theta$ za dano $\mathbf{x}_1$ i $\mathbf{x}_2$. Međutim, nezavisnost podrazumijeva

$$f(\mathbf{x}_1, \mathbf{x}_2|\theta) = f(\mathbf{x}_1|\theta)f(\mathbf{x}_2|\theta)$$

iz čega je očito

$$L_{1,2}(\theta) \propto L_1(\theta)L_2(\theta).$$

Uvrstimo prethodni izraz u (2.4) i dobijemo slijedeće:

$$\begin{aligned} g(\theta|\mathbf{x}_1, \mathbf{x}_2) &\propto g(\theta)L_1(\theta)L_2(\theta) \\ &\propto g(\theta|\mathbf{x}_1)L_2(\theta). \end{aligned}$$

Dakle, možemo pronaći aposteriornu funkciju gustoće za $\theta$ uz dane podatke $\mathbf{x}_1$ i $\mathbf{x}_2$ tretirajući aposteriornu gustoću od $\theta$ za dano $\mathbf{x}_1$ kao apriornu funkciju gustoće za opažanje $\mathbf{x}_2$. Gornja formula će funkcionirati bez obzira na vremenski poredak u kojem se promatraju $\mathbf{x}_1$ i $\mathbf{x}_2$. Ova je činjenica jedna od očitih prednosti Bayesovog pristupa.

# Poglavlje 3

# Model linearne regresije

Regresijska analiza je jedna od statističkih metoda koja se koristi za modeliranje odnosa između varijable odaziva (koja se naziva i zavisna varijabla) i jedne ili više eksplanatornih varijabli (također poznatih kao nezavisne varijable ili prediktori). Neka je $k$ broj eksplanatornih varijabli u modelu. Regresijska metoda pretpostavlja da možemo uspostaviti funkcijsku vezu između varijable odaziva $Y$ i eksplanatornih varijabli $x_1, \ldots, x_k$, odnosno

$$Y = f(x_1, \ldots, x_k)$$

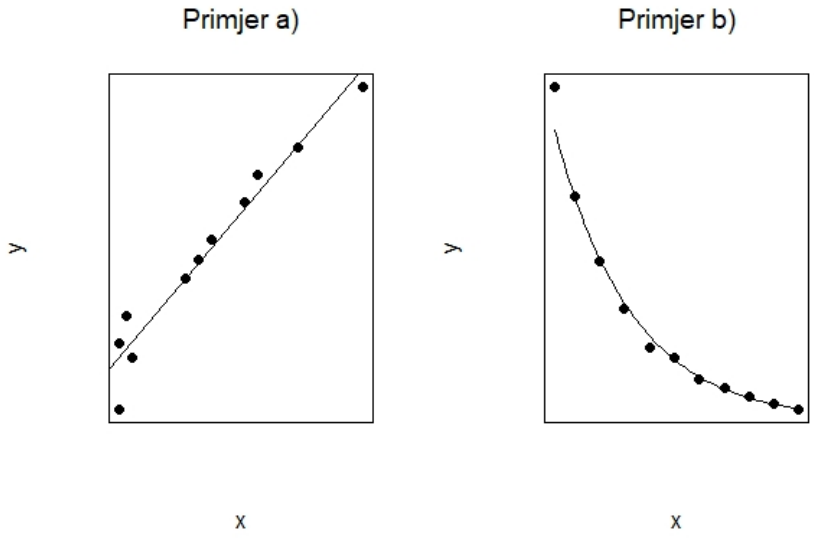
za neku funkciju $f : \mathbb{R}^k \to \mathbb{R}$. Međutim, uočimo da to nema previše smisla jer je $Y$ u pravilu slučajna varijabla. Stoga, pretpostavljamo vezu oblika

$$Y = f(x_1, \ldots, x_k) + \varepsilon,$$

gdje je $f : \mathbb{R}^k \to \mathbb{R}$ neka funkcija, a varijabla $\varepsilon$ je *slučajna pogreška* koja nastaje zbog utjecaja ostalih faktora na varijablu $Y$. Primjerice, to mogu biti greške koje su se dogodile prilikom mjerenja. Drugim riječima, slučajna greška u modelu predstavlja odstupanje varijable odaziva koje se ne može objasniti funkcijom eksplanatornih varijabli.

Zaključci o odnosu između varijable odaziva i eksplanatornih varijabli donose se na temelju uzorka dobivenog opažanjem ili mjerenjem neke pojave. Kao prvi korak regresijske analize za određivanje prikladnog oblika regresije koristimo dijagram rasipanja. Dijagram rasipanja je jednostavan alat pomoću kojeg odlučujemo o vrsti funkcije koja "najbolje" opisuje naše podatke. Na slici 3.1 prikazana su dva primjera dijagrama rasipanja. Promatramo odnos varijable odaziva i jedne eksplanatorne varijable. Na $x$-osi prikazane su vrijednosti eksplanatorne varijable, a na $y$-osi vrijednosti varijable odaziva. Vidimo da se na temelju rasporeda točaka jasno može naslutiti o tipu funkcije koji opisuje odnos između varijabli. Riječ je o linearnoj funkciji u primjeru a) i eksponencijalnoj funkciji u primjeru b).

Linearna regresija je specifična vrsta regresijske analize gdje pretpostavljamo da je zakonitost koja povezuje varijablu odaziva i eksplanatorne varijable linearnog tipa. Dakle,




Slika 3.1: Dijagram rasipanja za dva različita slučaja

pretpostavljamo da je funkcija $f : \mathbb{R}^k \to \mathbb{R}$ oblika

$$f(x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

gdje su $\beta_0, \beta_1, \ldots, \beta_k$ nepoznati parametri regresijskog modela. Riječ "linearna" označava da je model linearan u parametrima. Parametre je potrebno procijeniti tako da model "najbolje odgovara" našim podacima, a procjenjuju se na temelju unaprijed definiranog kriterija. Najčešće korišteni kriterij je metoda najmanjih kvadrata. Međutim, postoje i drugi kriteriji koji će rezultirati različitim procjenama regresijskih parametara čija će se statistička svojstva razlikovati od svojstava procjenitelja dobivenog metodom najmanjih kvadrata. U nastavku ovog rada za izvođenje procjena regresijskih parametara koristit će se metoda najmanjih kvadrata.

U regresijskom modelu $x_1, x_2, \ldots, x_k$ su kontrolirane (neslučajne) varijable, a slučajna varijabla $Y$ je mjerena u ovisnosti o $x = (x_1, x_2, \ldots, x_k)$. Linearna regresijska veza između varijable odaziva $Y$ i skupa eksplanatornih varijabli $x_1, x_2, \ldots, x_k$ utvrđuje se na osnovu uzorka veličine $n$. Radi se o $n$ opservacija ili mjerenja koja se mogu prikazati kao nizovi uređenih (k+1)-torki realnih brojeva oblika

$$(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i), \quad i = 1, 2, \ldots, n,$$

pri čemu su $x_{1i}, x_{2i}, \ldots, x_{ki}$ vrijednosti eksplanatornih varijabli $x_1, x_2, \ldots, x_k$ respektivno za $i$-tu opservaciju, a $y_i$ odgovarajuća vrijednost slučajne varijable $Y_i$. Ako regresijski mo-

del adekvatno odražava odnos između varijable odaziva i eksplanatornih varijabli, ovaj se model može koristiti za predviđanje budućih vrijednosti varijable $Y$ na temelju odabranog skupa vrijednosti eksplanatornih varijabli.

Ovisno o broju eksplanatornih varijabli govorimo ili o jednostavnoj ili o višestrukoj linearnoj regresiji. Poglavlje ćemo započeti s jednostavnom linearnom regresijom gdje se u modelu koristi samo jedna eksplanatorna varijabla ($k = 1$) kako bi se opisalo ponašanje varijable odaziva. Nakon toga proučit ćemo općenitiji model, model višestruke linearne regresije, s jednom varijablom odaziva i više od jedne eksplanatorne varijable ($k > 1$). Kao glavni izvori za ovo poglavlje korišteni su [4], [5] i [6].

## 3.1 Jednostavna linearna regresija

Jednostavna linearna regresija je statistička metoda koju koristimo za modeliranje odnosa jedne varijable odaziva i jedne eksplanatorne varijable. Neka je $x$ eksplanatorna varijabla te pretpostavimo da slučajna varijabla $Y$ na linearan način ovisi o $x$.

### Model jednostavne linearne regresije

Neka je $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ reprezentacija uzorka duljine $n$ iz linearnog regresijskog modela. Opći oblik modela jednostavne linearne regresije je dan s:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

gdje su

$y_i$ = vrijednost varijable odaziva za opservaciju i,
$x_i$ = vrijednost eksplanatorne varijable za opservaciju i,
$\beta_0, \beta_1$ = nepoznati parametri modela,
$\varepsilon_i$ = vrijednost slučajne greške za opservaciju i.

Slučajne greške $\varepsilon_i$, $i = 1, \ldots, n$ sadrže čimbenike koji utječu na vrijednost varijable odaziva, a nisu uključeni u model. Greške nisu opservabilne i njihovim uključivanjem u model (3.1) dobili smo vjerojatnosni model za podatke. Parametar $\beta_0$ nazivamo odsječak na osi ordinata (slobodni koeficijent), a parametar $\beta_1$ je koeficijent nagiba koji predstavlja utjecaj varijable $x_i$ na promjenu u varijabli $Y_i$ ako se $x_i$ poveća za jednu jedinicu.

## Metoda najmanjih kvadrata

Model jednostavne linearne regresije je dan s (3.1) uz pretpostavku da su odsječak na y-osi i koeficijent smjera nepoznati parametri. Sljedeći korak je pronaći "dobre" procjene za $\beta_0$ i $\beta_1$ tako da model najbolje opisuje podatke iz uzorka.

Postoji nekoliko kriterija za određivanje procjenitelja nepoznatih parametara $\beta_0$ i $\beta_1$. Najčešće korišten kriterij je metoda najmanjih kvadrata (engl. least square estimation) prema kojoj se procjene parametara određuju minimiziranjem sume kvadrata udaljenosti između opaženih podataka i teoretskih vrijednosti.

Dakle, procjene $\hat{\beta}_0$ i $\hat{\beta}_1$ regresijskih parametara određujemo minimiziranjem funkcije SSE (engl. sum of squared errors):

$$SSE = SSE(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2. \tag{3.2}$$

Definiramo reziduale kao

$$e_i := y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1 \ldots, n.$$

Zbog gornje definicije se ponekad u literaturi funkcija SSE označava i kao RSS (engl. residual sum of squares).

Znamo da su u točki u kojoj funkcija (3.2) postiže minimum njene parcijalne derivacije obzirom na $\hat{\beta}_0$ i $\hat{\beta}_1$ jednake nuli. Problem svodimo na rješavanje sustava jednadžbi:

$$\frac{\partial SSE}{\partial \hat{\beta}_0}(\hat{\beta}_0, \hat{\beta}_1) = 0 \tag{3.3}$$

$$\frac{\partial SSE}{\partial \hat{\beta}_1}(\hat{\beta}_0, \hat{\beta}_1) = 0. \tag{3.4}$$

Za početak uvodimo oznake koje ćemo koristiti u nastavku.

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2 \\
S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2 \\
S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}
\end{aligned}$$

Raspišimo prvo lijevu stranu u (3.3):

$$\frac{\partial SSE}{\partial \hat{\beta}_0}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] \cdot (-1)$$

Iz toga slijedi:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0,$$

odnosno

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0. \tag{3.5}$$

Pogledajmo sada lijevu stranu jednadžbe (3.4). Vrijedi:

$$\frac{\partial SSE}{\partial \hat{\beta}_1}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] \cdot (-x_i).$$

Uvrstimo prethodno dobiven izraz natrag u (3.3) i dobijemo

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0.$$

Iz (3.5) izrazimo $\beta_0$ i uvrstimo u gornju jednadžbu. Nakon sređivanja i uz oznake koje smo uveli, dobijemo izraz za procjenu koeficijenta smjera regresijskog pravca:

$$\hat{\beta_1} = \frac{S_{xy}}{S_{xx}}\,. \tag{3.6}$$

Još preostaje uvrstiti $\hat{\beta_1}$ u (3.5) kako bismo dobili procjenu za odječak na y-osi. Vrijedi:

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}. \tag{3.7}$$

Zaključno, jednadžba pravca prilagođenog metodom najmanjih kvadrata, odnosno jednadžba procjene regresijskog pravca zadana je s

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1}x, \tag{3.8}$$

gdje za procjene parametara regresije vrijedi (3.6) i (3.7).

**Primjer 3.1.1.** *Izmjereni su broj otkucaja srca (x) i unos kisika (y) za jednu osobu pod različitim uvjetima vježbanja. Želi se utvrditi može li se broj otkucaja srca, koji je lakše izmjeriv, koristiti za predviđanje unosa kisika. Ako je to tako, procijenjeni se unos kisika na temelju izmjerenog otkucaja srca tada može koristiti umjesto izmjerenog unosa kisika za kasnija ispitivanja na pojedincu. Podaci su dani tablicom:*

| *x (min)* | *94* | *96* | *94* | *95* | *104* | *106* | *108* | *113* | *115* | *121* | *131* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *y (L/min)* | *0.47* | *0.75* | *0.83* | *0.98* | *1.18* | *1.29* | *1.40* | *1.60* | *1.75* | *1.90* | *2.23* |

*Procijenimo parametre pravca prilagođenog metodom najmanjih kvadrata i nacrtajmo dijagram rasipanja s podacima zajedno s regresijskim pravcem. Iz* $\bar{x} = 107$, $\bar{y} = 1.30727$ *i*

$$S_{xx} = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2 = 1486$$

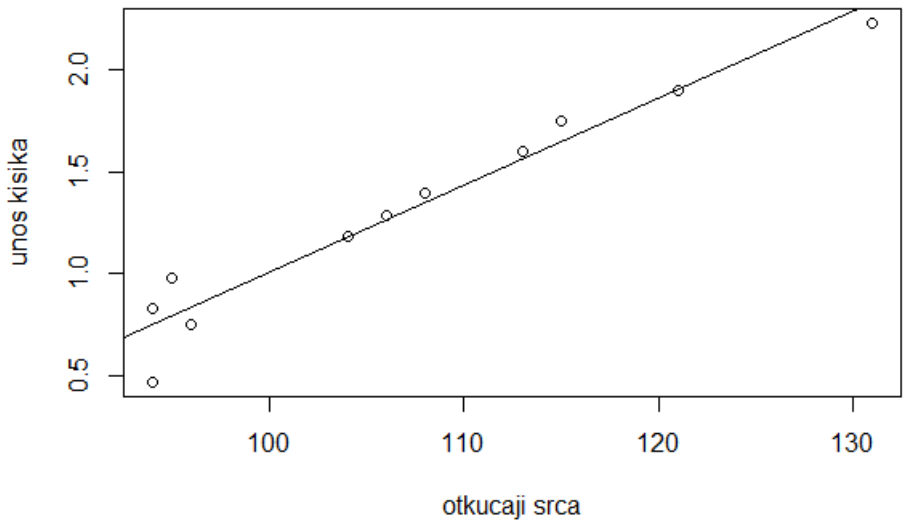
$$S_{yy} = \sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2 = 2.85602$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 63.38$$

*prema* (3.6) *i* (3.7) *dobijemo procjene parametara linearne regresije.*

$$\hat{\beta_1} = \frac{S_{xy}}{S_{xx}} = 0.042651, \quad \hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} = -3.256428.$$

*Jednadžba regresijskog pravca dana je s* $\hat{y} = \hat{\beta_0} + \hat{\beta_1}x = -3.256428 + 0.042651x$.




Slika 3.2: Dijagram rasipanja i regresijski pravac

## Alternativni oblik pravca prilagođenog metodom najmanjih kvadrata

Uz koeficijent smjera, bilo koja točka na pravcu koja nije odsječak na y-osi, također određuje pravac. Označimo s $\hat{\beta}_{\bar{x}}$ vrijednost od (3.8) ukoliko uzmemo $x = \bar{x}$, to jest

$$\hat{\beta}_{\bar{x}} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \overset{(3.5)}{=} \bar{y}. \tag{3.9}$$

Dakle, pravac prilagođen metodom najmanjih kvadrata prolazi točkom $(\bar{x}, \bar{y})$ te zbog toga možemo izvesti drugi oblik jednadžbe spomenutog pravca koji će nam biti osobito koristan. Alternativan oblik jednadžbe (3.8) glasi:

$$\hat{y} = \hat{\beta}_{\bar{x}} + \hat{\beta}_1(x - \bar{x}) = \bar{y} + \hat{\beta}_1(x - \bar{x}). \tag{3.10}$$

## Pretpostavke modela jednostavne linearne regresije

Za donošenje zaključaka o nepoznatim parametrima modela jednostavne linearne regresije moramo postaviti neke pretpostavke o vjerojatnosnoj distribuciji slučajnih grešaka u modelu. Pretpostavke na slučajne greške su sljedeće:

1. $\mathbb{E}[\varepsilon_i] = 0,\ i = 1, \ldots, n;$

2. $\text{Var}[\varepsilon_i] = \sigma^2,\ i = 1, \ldots, n$;

3. $Cov(\varepsilon_i, \varepsilon_j) = 0,$ za sve $i \neq j$;

4. $\varepsilon_1, ..., \varepsilon_n$ su normalno distribuirane (pišemo: $\varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n$).

Korištenjem alternativne parametrizacije dobivamo

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i = \beta_{\bar{x}} + \beta_1(x_i - \bar{x}) + \varepsilon_i,$$

gdje je $\beta_{\bar{x}}$ srednja vrijednost od $y$ uz dano $x = \bar{x}$, dok je $\beta_1$ koeficijent smjera. Svaki $\varepsilon_i$ je normalno distribuiran s očekivanjem 0 i poznatom varijancom $\sigma^2$. Nadalje, greške $\varepsilon_i$ su međusobno nezavisne jedna od druge. Uz ove pretpostavke vrijedi

$$\mathbb{E}[Y|x = x_i] = \beta_0 + \beta_1 x_i = \beta_{\bar{x}} + \beta_1(x_i - \bar{x}).$$

Tada varijabla odaziva $Y$ uvjetno na $x = x_i$ ima normalnu distribuciju

$$Y \sim N(\beta_{\bar{x}} + \beta_1(x_i - \bar{x}), \sigma^2), \tag{3.11}$$

te su $Y_i$ uz dano $x = x_i$ međusobno nezavisne jedna od druge za $i = 1, \ldots, n$.

## Procjena varijance

Nepristrana procjena varijance slučajnih grešaka oko pravca prilagođenog metodom najmanjih kvadrata je dana s

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n [y_i - (\hat{\beta}_{\bar{x}} + \hat{\beta}_1(x_i - \bar{x}))]^2}{n-2} = \frac{\sum_{i=1}^n [y_i - \hat{y}_i]^2}{n-2}, \tag{3.12}$$

što je zapravo jednako sumi kvadrata reziduala podijeljenoj s brojem stupnjeva slobode. Kako u izrazu (3.12) za nepristrani procjenitelj varijance koristimo dvije procjene za parametre, $\hat{\beta}_0$ i $\hat{\beta}_1$, broj stupnjeva slobode je $n - 2$.

**Primjer 3.1.2.** *(**nastavak na Primjer 2.1.1**) Odredimo procjenu varijance oko pravca prilagođenog metodom najmanjeg kvadrata za podatke o broju otkucaja srca i unosu kisika.*

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n [y_i - \hat{y}_i]^2}{n-2} = \frac{0.153}{9} = 0.01697$$

*Procjena standardne devijacije jednaka je*

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 0.13029.$$

### Transformacija modela

Osim linearnih modela, postoje i nelinearni regresijski modeli koji nisu linearni u parametrima. Međutim, ponekad je moguće nelinearne modele transformirati tako da veza između transformiranih varijabli bude linearna. Ako smo nakon transformacije dobili linearnu vezu između parametara, možemo koristiti model linearne regresije. U suprotnom, model je nelinearan.

Promotrimo u ovu svrhu eksponencijalni model koji je dan s:

$$u = e^{\beta_0+\beta_1 x} \tag{3.13}$$

gdje su $\beta_0$ i $\beta_1$ parametri modela. Logaritmiramo jednadžbu (3.13) i dobijemo:

$$y = \beta_0 + \beta_1 x, \tag{3.14}$$

pri čemu je $y = \ln u$. Primijetimo da smo dobili model koji je linearan u parametrima i sada možemo procijeniti parametre metodom najmanjih kvadrata koristeći y kao varijablu odaziva. Prilagođeni eksponencijalni model je oblika:

$$u = e^{\hat{\beta}_0+\hat{\beta}_1 x}, \tag{3.15}$$

gdje su $\hat{\beta}_0$ i $\hat{\beta}_1$ dobivene procjene izračunate iz logaritmiranih podataka.

## 3.2 Višestruka linearna regresija

U prošlom potpoglavlju obradili smo temu linearne regresije s jednom eksplanatornom varijablom. Međutim, u praksi se često želi utvrditi postoji li linearan odnos između varijable odaziva i više od jedne eksplanatorne varijable. U tom slučaju koristit ćemo model višestruke linearne regresije.

### Model višestruke linearne regresije

Neka je $Y$ varijabla odaziva, a $x_1, x_2, \ldots, x_k$ eksplanatorne varijable, pri čemu je $k \geq 2$. Pretpostavljamo da je veza između $Y$ i $(x_1, x_2, \ldots, x_k)$ linearna. Nadalje, pretpostavimo da se $i$-ti podatak iz uzorka, od ukupno $n$ opažanja ili mjerenja neslučajnih varijabli, može prikazati kao niz numeričkih varijabli $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$. Opći model višestruke linearne regresije je:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.16}$$

gdje su:

$y_i$ = vrijednost varijable odaziva za opservaciju $i$,
$x_{i1}, x_{i2}, \ldots, x_{ik}$ = vrijednosti eksplanatornih varijabli za opservaciju $i$,
$\beta_0, \beta_1, \ldots, \beta_k$ = nepoznati parametri modela,
$\varepsilon_i$ = vrijednost slučajne greške za opservaciju $i$.

Linearna funkcija u (3.16) određuje hiperravninu u (k+1)-dimenzionalnom prostoru. U toj jednadžbi parametar $\beta_i$, za $i = 1, \ldots, k$, predstavlja promjenu u varijabli $y_i$ kada se $x_i$ poveća za jednu jedinicu, a ostale vrijednosti eksplanatornih varijabli ostaju nepromijenjene. Kažemo da je $\beta_i$ *parametar nagiba* u smjeru $x_i$, za $i = 1, \ldots, k$. Parametar $\beta_0$ je konstantni dio varijable odaziva kada su sve vrijednosti eksplanatornih varijabli jednake nuli i nazivamo ga *parametrom presjeka*.

Zapišimo (3.16) u matričnom obliku kao:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.17}$$

ili jednostavnije :

$$\mathbf{y} = X\beta + \varepsilon, \tag{3.18}$$

pri čemu su:

$\mathbf{y}$ = vektor stupac s $n$ vrijednosti varijable odaziva,
X =matrica tipa $n \times (k+1)$ čiji prvi stupac sadrži jedinice, a ostali stupci vrijednosti eksplanatornih varijabli $x_1, x_2, \ldots, x_k$,
$\beta$ = vektor stupac s $k+1$ nepoznatih parametara,
$\varepsilon$ = vektor stupac s $n$ grešaka.

Matrična forma višestrukog linearnog modela nam omogućava praktičnije i učinkovitije donošenje zaključaka o svojstvima regresijskog modela.

## Procjenitelji parametara modela

Procjenu vektora nepoznatih parametara, u oznaci $\mathbf{b}_{LS} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T$, određujemo metodom najmanjih kvadrata. Definiramo vektor reziduala:

$$\hat{\varepsilon} = \mathbf{y} - X\mathbf{b}_{LS}.$$

Dakle, tražimo $\mathbf{b}_{LS}$ koji minimizira sumu rezidualnih odstupanja:

$$\min_{\mathbf{b}_{LS}}[\hat{\varepsilon}^T\hat{\varepsilon}] = \min_{\mathbf{b}_{LS}}[(\mathbf{y} - X\mathbf{b}_{LS})^T(\mathbf{y} - X\mathbf{b}_{LS})] = \min_{\mathbf{b}_{LS}}[SSE(\mathbf{b}_{LS})]. \tag{3.19}$$

Budući da su u točki u kojoj funkcija dostiže minimum njene parcijalne derivacije jednake nuli, zahtjev (3.19) se svodi na rješavanje sustava jednadžbi:

$$\frac{\partial}{\partial \mathbf{b}_{LS}}[(\mathbf{y} - X\mathbf{b}_{LS})^T(\mathbf{y} - X\mathbf{b}_{LS})] = 0,$$

ili ekvivalentno

$$\frac{\partial}{\partial \mathbf{b}_{LS}}[(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T X\mathbf{b}_{LS} + \mathbf{b}_{LS}^T X^T X\mathbf{b}_{LS})] = 0.$$

Uzimanjem parcijalne derivacije u odnosu na svaku komponentu vektora $\hat{\beta}$ dobivamo sljedeću normalnu jednadžbu modela višestruke linearne regresije:

$$X^T X\mathbf{b}_{LS} = X^T\mathbf{y}.$$

Uz pretpostavku da je $X$ matrica punog ranga ($r(X) = k + 1$), $X^T X$ je invertibilna matrica i vrijedi

$$\mathbf{b}_{LS} = (X^T X)^{-1} X^T\mathbf{y}, \tag{3.20}$$

što daje procjenu za $\beta$.

## Pretpostavke modela višestruke linearne regresije

Pretpostavke za model višestruke linearne regresije su jednake pretpostavkama u slučaju jednostavne linearne regresije uz jedan dodatan uvjet. Dakle, pretpostavke na greške su:

1. $\mathbb{E}[\varepsilon_i] = 0,\ i = 1, \ldots, n$;
2. $\mathrm{Var}[\varepsilon_i] = \sigma^2,\ i = 1, \ldots, n$;
3. $Cov(\varepsilon_i, \varepsilon_j) = 0,$ za sve $i \neq j$;
4. $\varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n.$

Dodatna pretpostavka na model koju smo iz numeričkih razloga uveli u prošlom potpoglavlju je:

5. $X$ je matrica punog ranga, to jest vrijedi $r(X) = k + 1$, što povlači da je $X^T X$ invertibilna matrica.

Pretpostavke (1.) - (4.) možemo ekvivalentno zapisati na sljedeći način:

$$\varepsilon \sim N(0, \sigma^2 I_n),$$

gdje je $I_n$ jedinična matrica tipa $n \times n$.

Uz ove pretpostavke za distribuciju slučajnog vektora $\mathbf{Y}$ vrijedi:

$$\mathbf{Y} \sim N(X\beta, \sigma^2 I_n), \tag{3.21}$$

odnosno odaziv $\mathbf{Y}$ se generira iz multivarijatne normalne distribucije s očekivanjem koje je jednako umnošku vektora parametara i matrice eksplanatornih varijabli, a varijanca je predstavljena kao kvadrat standardne devijacije pomnožene matricom identiteta $I_n$.

Pretpostavljamo da opaženi podaci dolaze iz postavljenog linearnog regresijskog modela. Budući da je vektor $\mathbf{b}_{LS}$ dobiven metodom najmanjih kvadrata zapravo linearna funkcija vektora opaženih vrijednosti za varijablu odaziva, pod pretpostavkama modela njegova kovarijacijska matrica je

$$\begin{aligned} V_{LS} &= (X^T X)^{-1} X^T (\sigma^2 I_n) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}. \end{aligned}$$

## Procjena varijance

Vrijednost $\sigma^2$ je često nepoznata te ju je potrebno procijeniti. Uz pretpostavke modela, nepristrani procjenitelj varijance dan je s

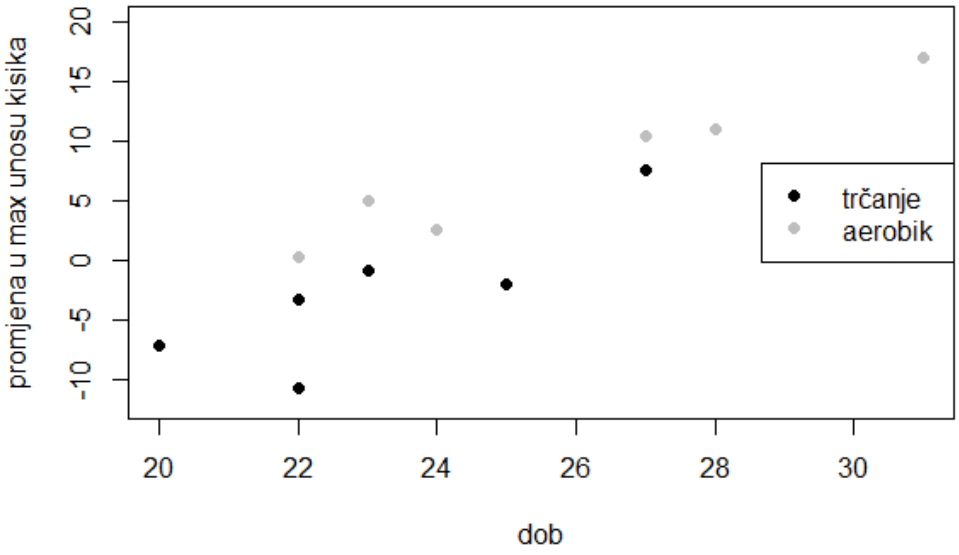
$$\hat{\sigma}^2 = \frac{SSE}{n - (k+1)} = \frac{(\mathbf{y} - X\mathbf{b}_{LS})^T (\mathbf{y} - X\mathbf{b}_{LS})}{n - (k+1)}. \tag{3.22}$$

**Primjer 3.2.1.** [1] *U istraživanju učinaka dva različita režima vježbanja na unos kisika, provedeno je ispitivanje na dvanaest zdravih muškaraca koji nisu redovito vježbali. Nasumičnim odabirom šest muškaraca raspoređeno je u program trčanja po ravnom terenu, a preostalim šest dodijeljen je program aerobika. Svaki od programa traje dvanaest tjedana. Mjeren je maksimalni unos kisika svakog ispitanika (u litrama po minuti) tijekom trčanja na kosoj traci za trčanje, prije i nakon dvanaestotjednog programa. Očekuje se da će promjena u maksimalnom unosu kisika kod ispitanika ovisiti o tome kojem su programu dodijeljeni, ali i o njihovoj dobi. Podaci su dani u tablici ispod.*

| $x_1(dob)$ | 23 | 22 | 22 | 25 | 27 | 20 | 31 | 23 | 27 | 28 | 22 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_2(program)$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| $y$ (L/min) | 0.87 | 10.74 | 3.27 | 1.97 | 7.50 | 7.25 | 17.05 | 4.96 | 10.40 | 11.05 | 0.26 | 2.51 |

[1] Primjer je preuzet iz [1].

*Eksplanatorna varijabla $x_2$ označava u kojem je od dvaju programa ispitanik sudjelovao. Ako je $x_2 = 0$, riječ je o programu trčanja po ravnom terenu, a $x_2 = 1$ naznačuje da se radi o programu aerobika.*




Slika 3.3: *Podaci o dobi ispitanika i promjeni u maksimalnom unosu kisika*

*Kako bismo opisali ovisnost maksimalnog unosa kisika o dobi ispitanika i tipu programa, odabrali smo linearni model oblika:*

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

*gdje su*

$x_{i1} =$ *dob ispitanika i,*

$x_{i2} =$ *0 ako je subjekt i u programu trčanja, 1 ako je na aerobiku,*

$x_{i3} = x_{i1} \times x_{i2}$.

*Pronađimo procjenitelje parametara $\beta_0, \beta_1, \beta_2$ i $\beta_3$ metodom najmanjih kvadrata. Nakon što konstruiramo matricu tipa $12 \times 4$ čiji su prvi stupac jedinice, a ostali stupci vektori $x_1, x_2$ i $x_3$, kao što smo to napravili u* (3.17)*, možemo izračunati matrice $X^T X$ i $X^T \mathbf{y}$:*

$$X^T X = \begin{bmatrix} 12 & 294 & 6 & 155 \\ 294 & 7314 & 155 & 4063 \\ 6 & 155 & 6 & 155 \\ 155 & 4063 & 155 & 4063 \end{bmatrix}, \qquad X^T y = \begin{bmatrix} 29.63 \\ 978.81 \\ 46.23 \\ 1298.79 \end{bmatrix}.$$

*Prema* (3.20)*, za određivanje procjena parametara još je potrebno invertirati matricu $X^T X$ i rezultat pomnožiti matricom $X^T \mathbf{y}$. Dobijemo:*

$$\mathbf{b}_{LS} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^T = (-51.29, 2.09, 13.11, -0.32)^T.$$

*Dakle, procijenjeni regresijski model glasi:*

$$y = -51.29 + 2.09x_1 + 13.11x_2 - 0.32x_3.$$

*Uz $n = 12$ i $k = 3$, nepristrani procjenitelj varijance može se dobiti korištenjem izraza* (3.22)*:*

$$\hat{\sigma}^2 = \frac{SSE}{n - (k+1)} = \frac{68.33981}{8} = 8.54.$$

# Poglavlje 4

# Bayesovsko zaključivanje za linearnu regresiju

Bayesovsko zaključivanje za linearnu regresiju je statistički pristup koji kombinira regresijsku analizu s principima Bayesovskog zaključivanja. U ovom pristupu pretpostavljamo da su parametri modela linearne regresije slučajne varijable s apriornim distribucijama vjerojatnosti. Zatim koristimo Bayesov teorem za korigiranje naših uvjerenja o parametrima modela uzimajući u obzir promatrane podatke što rezultira aposteriornim distribucijama vjerojatnosti. Nakon što imamo aposteriornu distribuciju, možemo je koristiti za izvođenje zaključaka o parametrima modela, kao na primjer za određivanje intervala vjerodostojnosti za koeficijent smjera. Također možemo je koristiti za predviđanje varijable odaziva za nove vrijednosti eksplanatornih varijabli.

Postoji nekoliko prednosti Bayesovskog zaključivanja za linearnu regresiju u odnosu na tradicionalne frekvencionističke metode, kao što je sposobnost uključivanja prethodnih informacija ili uvjerenja o parametrima modela te mogućnost računanja aposteriornih vjerojatnosti hipoteza. Međutim, treba napomenuti da ovakav pristup često zahtijeva složenije izračune, posebno za veće skupove podataka. Poglavlje ćemo započeti s primjenom Bayesovske statistike na jednostavne modele linearne regresije, a zatim ćemo generalizirati rezultate na modele višestruke linearne regresije. Kao izvor korišteni su [5] i [8].

## 4.1 Bayesovsko zaključivanje za jednostavnu linearnu regresiju

U ovom odjeljku bavimo se Bayesovskim zaključivanjem za jednostavnu linearnu regresiju. U klasičnoj linearnoj regresiji razmatrali smo frekvencionistički pristup procjene nepoznatih parametara modela. Sada ćemo pokazati kako koristiti Bayesov teorem, od-

nosno Bayesovsko zaključivanje za revidiranje prethodnih uvjerenja o parametrima modela na temelju promatranih podataka.

Ponovno ćemo razmotriti normalan model jednostavne linearne regresije, to jest pretpostavljamo da su slučajne greške u modelu nezavisne i jednako distribuirane s raspodjelom $N(0, \sigma^2)$. Pokazali smo da uz pretpostavke na model jednostavne linearne regresije varijabla odaziva $Y$ uvjetno na $x = x_i$ dolazi iz normalne razdiobe s nepoznatim parametrom očekivanja $\beta_{\bar{x}} + \beta_1(x_i - \bar{x})$ i parametrom varijance $\sigma^2$. Stoga je funkcija gustoće slučajne varijable $Y$ dana s

$$f(y \mid \beta_{\bar{x}}, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}[y-(\beta_{\bar{x}}+\beta_1(x_i-\bar{x}))]^2}.$$

Bayesov teorem možemo iskazati na sljedeći način:

$$\textit{aposteriorna gustoća} \propto \textit{apriorna gustoća} \times \textit{funkcija vjerodostojnosti}.$$

Dakle, za početak moramo odrediti funkciju vjerodostojnosti i odabrati prikladnu apriornu distribuciju za ovaj model. Primijetimo da nam ova formula ne daje egzaktnu aposteriornu gustoću, već samo dobivamo uvid u njen oblik.

## Funkcija vjerodostojnosti parametara modela

Neka je $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ reprezentacija uzorka duljine $n$ iz linearnog regresijskog modela. Funkcija vjerodostojnosti $i$-te opservacije iz uzorka modela jednaka je njezinoj funkciji gustoće, ali izražena kao funkcija triju parametara $\beta_1, \beta_{\bar{x}}$ i $\sigma^2$ (ili kao funkcija dvaju parametara $\beta_1$ i $\beta_{\bar{x}}$ ako nam je varijanca poznata), gdje su varijable $x_i$ i $y_i$ fiksirane na opaženim vrijednostima. Prvo ćemo odrediti funkciju vjerodostojnosti u slučaju kada nam je varijanca modela nepoznata. Vjerodostojnost za $i$-tu opservaciju varijable odaziva je oblika

$$L_i(\beta_{\bar{x}}, \beta_1, \sigma^2) = f(y_i \mid x_i, \beta_{\bar{x}}, \beta_1, \sigma^2) \propto \frac{1}{\sqrt{\sigma^2}} e^{-\frac{1}{2\sigma^2}[y_i-(\beta_{\bar{x}}+\beta_1(x_i-\bar{x}))]^2}, i = 1, \ldots, n,$$

budući da možemo zanemariti konstantu $\frac{1}{\sqrt{2\pi}}$ koja ne sadrži parametre.

Opažanja $x_1, \ldots, x_n$ su međusobno nezavisna pa je vjerodostojnost cijelog uzorka jed-

naka produktu individualnih vjerodostojnosti.

$$\begin{aligned} L(\beta_{\bar{x}}, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f(y_i \mid x_i, \beta_{\bar{x}}, \beta_1, \sigma^2) \\ &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}} e^{-\frac{1}{2\sigma^2}[y_i - (\beta_{\bar{x}} + \beta_1(x_i - \bar{x}))]^2} \\ &\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (\beta_{\bar{x}} + \beta_1(x_i - \bar{x}))]^2} \\ &\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - \bar{y} + \bar{y} - (\beta_{\bar{x}} + \beta_1(x_i - \bar{x}))]^2}. \end{aligned}$$

Promotrimo izraz koji se nalazi unutar sume u eksponentu. Nakon grupiranja, rastavimo taj izraz na tri sume te tako dobijemo

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \bar{y})(\bar{y} - (\beta_{\bar{x}} + \beta_1(x_i - \bar{x}))) + \sum_{i=1}^{n}(\bar{y} - (\beta_{\bar{x}} + \beta_1(x_i - \bar{x})))^2$$

ili jednostavnije zapisano kao:

$$S_{yy} - 2\beta_1 S_{xy} + {\beta_1}^2 S_{xx} + n(\beta_{\bar{x}} - \bar{y})^2.$$

Sada za funkciju vjerodostojnosti vrijedi

$$L(\beta_{\bar{x}}, \beta_1, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}[S_{yy} - 2\beta_1 S_{xy} + {\beta_1}^2 S_{xx} + n(\beta_{\bar{x}} - \bar{y})^2]}. \tag{4.1}$$

Ako nam je parametar varijance poznat, vrijednost $\sigma^2$ je konstanta i više se ne smatra jednim od parametara u ovom modelu. Analognim računom kao za slučaj nepoznate varijance dolazimo do idućeg izraza za funkciju vjerodostojnosti

$$L(\beta_{\bar{x}}, \beta_1) \propto e^{-\frac{1}{2\sigma^2}[S_{yy} - 2\beta_1 S_{xy} + {\beta_1}^2 S_{xx} + n(\beta_{\bar{x}} - \bar{y})^2]}, \tag{4.2}$$

gdje smo dio koji ne ovisi o nepoznatim parametrima zanemarili jer ga možemo smatrati proporcionalnom konstantom. Zapišimo sada gornji rezultat kao produkt dviju eksponencijalnih funkcija. Vrijedi

$$L(\beta_{\bar{x}}, \beta_1) \propto e^{-\frac{1}{2\sigma^2}[S_{yy} - 2\beta_1 S_{xy} + {\beta_1}^2 S_{xx}]} \times e^{-\frac{1}{2\sigma^2}[n(\beta_{\bar{x}} - \bar{y})^2]}.$$

Nakon što u prvom članu u posljednjem izrazu izlučimo $S_{xx}$ ispred zagrade, nadopunimo do kvadrata razlike i apsorbiramo dio koji ne ovisi o parametrima u konstantu proporcionalnosti, imamo sljedeće:

$$L(\beta_{\bar{x}}, \beta_1) \propto e^{-\frac{1}{2\sigma^2/S_{xx}}[(\beta_1 - \frac{S_{xy}}{S_{xx}})^2]} \times e^{-\frac{1}{2\sigma^2/n}[(\beta_{\bar{x}} - \bar{y})^2]}. \tag{4.3}$$

Primijetimo da je prema (3.6) i (3.9) $\hat{\beta_1} = \frac{S_{xy}}{S_{xx}}$ i $\hat{\beta}_{\bar{x}} = \bar{y}$ te da smo funkciju vjerodostojnosti zapravo zapisali kao produkt dviju pojedinačnih vjerodostojnosti:

$$L(\beta_{\bar{x}}, \beta_1) \propto L(\beta_{\bar{x}}) \times L(\beta_1), \tag{4.4}$$

gdje je

$$L(\beta_{\bar{x}}) \propto e^{-\frac{1}{2\sigma^2/S_{xx}}(\beta_1 - \hat{\beta_1})^2} \tag{4.5}$$

i

$$L(\beta_1) \propto e^{-\frac{1}{2\sigma^2/n}(\beta_{\bar{x}} - \hat{\beta}_{\bar{x}})^2}. \tag{4.6}$$

S obzirom na to da je funkcija vjerodostojnosti nepoznatih parametara modela jednaka produktu pojedinačnih funkcija vjerodostojnosti, zaključujemo da su vjerodostojnost parametara $\beta_{\bar{x}}$ i vjerodostojnost parametara $\beta_1$ nezavisne. Dodatno, prepoznajemo da je vjerodostojnost parametra $\beta_1$ proporcionalna funkciji gustoće normalne slučajne varijable s očekivanjem $\hat{\beta_1}$ i varijancom $\sigma^2/S_{xx}$. Slično, vjerodostojnost parametra $\beta_{\bar{x}}$ je proporcionalna funkciji gustoće normalne slučajne varijable s očekivanjem $\hat{\beta}_{\bar{x}}$ i varijancom $\sigma^2/n$.

## Apriorna i aposteriorna distribucija

U Bayesovskoj statistici apriorne distribucije predstavljaju naša uvjerenja ili znanja o nepoznatim parametrima prije nego što su nam dani podaci. Ovisno o količini informacija s kojima raspolažemo ili o stupnju vjerovanja kojeg imamo o nepoznatom parametru, koristimo ili informativne ili neinformativne apriorne distribucije. Neinformativne su one distribucije koje označavaju da posjedujemo minimalno apriorno znanja o nepoznatom parametru, dok informativne apriorne distribucije sadrže neke korisne informacije o nepoznatom parametru. Odabir između informativnih i neinformativnih apriornih distribucija ovisi i o stupnju utjecaja koji se želi pripisati apriornim informacijama u aposteriornoj distribuciji.

Neinformativne apriorne distribucije biramo kada nemamo dovoljno podataka o nepoznatom parametru. One minimalno utječu na aposteriornu distribuciju i dopuštaju podacima da govore sami za sebe. Primjeri neinformativnih apriornih distribucija uključuju uniformnu distribuciju, Jeffreyjevu apriornu distribuciju i referentne apriorne distribucije.

S druge strane, ako imamo čvrsto prethodno znanje ili uvjerenje o nepoznatom parametru, odabrat ćemo informativnu apriornu distribuciju kako bismo te informacije uključili u analizu. Zbog toga informativne apriorne distribucije imaju snažan utjecaj na aposteriornu distribuciju. Jedan primjer informativne apriorne distribucije je normalna distribucija.

Od sada pa nadalje pretpostavljamo da su parametri modela jednostavne linearne regresije i varijanca modela međusobno nezavisni. Ako je parametar varijance nepoznat, tada za zajedničku apriornu funkciju gustoće nepoznatih parametara modela vrijedi iduće:

$$g(\beta_{\bar{x}}, \beta_1, \sigma^2) = g(\beta_{\bar{x}}, \beta_1) \times g(\sigma^2). \tag{4.7}$$

Pretpostavimo još i nezavisnost između $\beta_{\bar{x}}$ i $\beta_1$. Zajednička apriorna funkcija gustoće parametara regresije jednaka je umnošku marginalnih apriornih gustoća, to jest

$$g(\beta_{\bar{x}}, \beta_1) = g(\beta_{\bar{x}}) \times g(\beta_1). \tag{4.8}$$

U nastavku ćemo pogledati par posebnih slučajeva apriornih distribucija.

**Referentna apriorna distribucija**

Ukoliko imamo minimalno informacija o nepoznatim parametrima modela možemo odabrati referentnu apriornu distribuciju za koju vrijedi:

$$g(\beta_{\bar{x}}, \beta_1, \sigma^2) \propto \frac{1}{\sigma^2}. \tag{4.9}$$

Uz pretpostavku o nezavisnosti parametara modela, izraz (4.9) ekvivalentan je tome da za parametre $\beta_{\bar{x}}$ i $\beta_1$ koristimo uniformnu apriornu distribuciju te apriornu distribuciju za $\sigma^2$ koja je proporcionalna $\frac{1}{\sigma^2}$, odnosno

$$g(\beta_{\bar{x}}) \propto 1 \qquad g(\beta_1) \propto 1 \qquad g(\sigma^2) \propto \frac{1}{\sigma^2}. \tag{4.10}$$

Ovdje je riječ o standardnom odabiru za neinformativnu apriornu distribuciju čiji je smisao da se zaključci suštinski formiraju na temelju podataka.

Raspišimo sada izraz za aposteriornu funkciju gustoće. Prema Bayesovom teoremu zajednička aposteriorna funkcija gustoća nepoznatih parametara modela je proporcionalna umnošku zajedničke apriorne gustoće i vjerodostojnosti, odnosno

$$g(\beta_{\bar{x}}, \beta_1, \sigma^2 \mid \mathbf{y}, \mathbf{x}) \propto g(\beta_{\bar{x}}, \beta_1, \sigma^2) \times L(\beta_{\bar{x}}, \beta_1, \sigma^2),$$

gdje su $\mathbf{x} = (x_1, \ldots, x_n)$ i $\mathbf{y} = (y_1, \ldots, y_n)$ $n$-dimenzionalni vektori čije komponente dolaze iz skupa uređenih parova točaka $(x_1, y_1), \ldots, (x_n, y_n)$ koje predstavljaju podatke iz uzorka za jednostavnu linearnu regresiju. Iskoristimo (4.1) i (4.9) i dobijemo

$$\begin{aligned}
g(\beta_{\bar{x}}, \beta_1, \sigma^2 \mid \mathbf{y}, \mathbf{x}) &\propto (\sigma^2)^{-\frac{(n+2)}{2}} e^{-\frac{1}{2\sigma^2}[S_{yy} - 2\beta_1 S_{xy} + {\beta_1}^2 S_{xx} + n(\beta_{\bar{x}} - \bar{y})^2]} \\
&\propto (\sigma^2)^{-\frac{(n+2)}{2}} e^{-\frac{1}{2\sigma^2}[S_{yy} - \frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} - 2\beta_1 S_{xy} + {\beta_1}^2 S_{xx} + n(\beta_{\bar{x}} - \bar{y})^2]} \\
&\propto (\sigma^2)^{-\frac{(n+2)}{2}} e^{-\frac{1}{2\sigma^2}[S_{ee} + S_{xx}(\beta_1 - \frac{S_{xy}}{S_{xx}})^2 + n(\beta_{\bar{x}} - \bar{y})^2]} \\
&\propto (\sigma^2)^{-\frac{(n+2)}{2}} e^{-\frac{1}{2\sigma^2}[S_{ee} + S_{xx}(\beta_1 - \hat{\beta}_1)^2 + n(\beta_{\bar{x}} - \hat{\beta}_{\bar{x}})^2]},
\end{aligned}$$

gdje smo u trećem retku uveli supstituciju $S_{ee} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$, a u četvrtome smo upotrijebili izraze za procjene regresijskih parametara. Zapišimo prethodni rezultat kao umnožak dviju eksponencijalnih funkcija:

$$g(\beta_{\bar{x}}, \beta_1, \sigma^2 \mid \mathbf{y}, \mathbf{x}) \propto (\sigma^2)^{-\frac{(n+1)}{2}} e^{-\frac{1}{2\sigma^2}[S_{ee}+n(\beta_{\bar{x}}-\hat{\beta}_{\bar{x}})^2]} \times (\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{\sigma^2/S_{xx}}(\beta_1-\hat{\beta}_1)^2}. \tag{4.11}$$

Lako se vidi da je aposteriorna distribucija parametra $\beta_1$ uz dane $\hat{\beta}_1$ i $\sigma^2$ normalna s očekivanjem $\hat{\beta}_1$ i varijancom $\sigma^2/S_{xx}$. Sada integriranjem po $\beta_1$ iz zajedničke aposteriorne distribucije dobijemo

$$\begin{aligned} g(\beta_{\bar{x}}, \sigma^2 \mid \mathbf{y}, \mathbf{x}) &= \int g(\beta_{\bar{x}}, \beta_1, \sigma^2 \mid \mathbf{y}, \mathbf{x}) d\beta_1 \\ &\propto (\sigma^2)^{-\frac{(n+1)}{2}} e^{-\frac{1}{2\sigma^2}[S_{ee}+n(\beta_{\bar{x}}-\hat{\beta}_{\bar{x}})^2]}. \end{aligned}$$

U [8] je pokazano da ako

$$g(\theta, \sigma^2) \propto (\sigma^2)^{-(\nu+1)/2-1} e^{-\frac{1}{2\sigma^2}[S+n(\theta-\bar{x})^2]}$$

i $s^2 = S/\nu$, tada

$$t = \frac{(\theta - \bar{x})}{s/\sqrt{n}} \sim t_\nu \qquad i \qquad \sigma^2 \sim S/Z,$$

gdje Z ima takozvanu log $\chi^2$ distribuciju s $\nu$ stupnjeva slobode, to jest vrijedi da je $Z = logW$ slučajna varijabla takva da W ima $\chi^2$ distribuciju s $\nu$ stupnjeva slobode. Iz navedenoga proizlazi da ako za aposteriornu distribuciju parametra $\beta_{\bar{x}}$ uz dane **x** i **y** vrijedi $s^2 = S_{ee}/(n-2)$, onda

$$\frac{(\beta_{\bar{x}} - \hat{\beta}_{\bar{x}})}{s/\sqrt{n}} \sim t_{n-2}. \tag{4.12}$$

Vratimo se sada na izraz (4.11) i zapišimo ga u drugačijem obliku:

$$g(\beta_{\bar{x}}, \beta_1, \sigma^2 \mid \mathbf{y}, \mathbf{x}) \propto (\sigma^2)^{-\frac{(n+1)}{2}} e^{-\frac{1}{2\sigma^2}[S_{ee}+S_{xx}(\beta_1-\hat{\beta}_1)^2]} \times (\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{\sigma^2/n}(\beta_{\bar{x}}-\hat{\beta}_{\bar{x}})^2}.$$

Prepoznajemo da je aposteriorna distribucija parametra $\beta_{\bar{x}}$ uz dane $\hat{\beta}_{\bar{x}}$ i $\sigma^2$ normalna s očekivanjem $\hat{\beta}_{\bar{x}}$ i varijancom $\sigma^2/n$. Slično kao prije, aposteriorna distribucija od $\beta_1$ može se pronaći integriranjem po $\beta_{\bar{x}}$ iz zajedničke aposteriorne distribucije i pokaže se da je tada

$$\frac{(\beta_1 - \hat{\beta}_1)}{s/\sqrt{S_{xx}}} \sim t_{n-2}. \tag{4.13}$$

Naposljetku, primijetimo da je

$$\sigma^2 \sim S_{ee}/Z, \tag{4.14}$$

gdje je $Z$ slučajna varijabla takva da je $Z = logW$, pri čemu slučajna varijabla $W$ ima $\chi^2$ distribuciju s $n-2$ stupnja slobode. Također, treba napomenuti da se iz dobivenih aposteriornih funkcija gustoće može pokazati da slučajne varijable $\beta_{\bar{x}}$ i $\beta_1$ nisu nezavisne, ali su nezavisne uz dano $\sigma^2$. Uočimo još da je očekivanje aposteriorne distribucije od $\beta_1$ jednako $\hat{\beta}_1$, a očekivanje aposteriorne distribucije od $\beta_{\bar{x}}$ je upravo $\hat{\beta}_{\bar{x}}$.

U situaciji kada znamo vrijednost varijance $\sigma^2$, problem određivanja aposteriorne distribucije je nešto jednostavniji, budući da se varijanca više ne smatra jednim od parametara modela. Pomoću izraza (4.8) za zajedničku funkciju gustoće od $\beta_{\bar{x}}$ i $\beta_1$ te uz pretpostavku (4.10), dobijemo da vrijedi $g(\beta_{\bar{x}}, \beta_1) \propto 1$. Prema Bayesovom teoremu imamo

$$\begin{aligned} g(\beta_{\bar{x}}, \beta_1 \mid \mathbf{y}, \mathbf{x}) &\propto g(\beta_{\bar{x}}, \beta_1) \times L(\beta_{\bar{x}}, \beta_1) \\ &\propto e^{-\frac{1}{2\sigma^2}[S_{yy} - 2\beta_1 S_{xy} + {\beta_1}^2 S_{xx} + n(\beta_{\bar{x}} - \bar{y})^2]} \\ &\propto e^{-\frac{1}{2\sigma^2/S_{xx}}(\beta_1 - \frac{S_{xy}}{S_{xx}})^2} \times e^{-\frac{1}{2\sigma^2/n}(\beta_{\bar{x}} - \bar{y})^2}, \end{aligned}$$

gdje smo u posljednja dva retka koristili izvode (4.2) i (4.3). Očito je

$$\beta_1 \sim N(\hat{\beta}_1, \sigma^2/n) \quad \text{i} \quad \beta_{\bar{x}} \sim N(\hat{\beta}_{\bar{x}}, \sigma^2/S_{xx}), \tag{4.15}$$

pri čemu smo koristili $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ te $\hat{\beta}_{\bar{x}} = \bar{y}$. Vidimo da vrijedi i nezavisnost.

### Normalna apriorna distribucija

Osim neinformativne referentne apriorne distribucije nepoznatih parametara modela, možemo razmotriti korištenje informativne normalne apriorne distribucije ako su nam dostupne određene informacije o parametrima. U ovom odjeljku pretpostavljamo da je varijanca poznata te da su $\beta_{\bar{x}}$ i $\beta_1$ jedini nepoznati parametri u modelu.

Pokazali smo da, uz pretpostavke na greške u modelu jednostavne linearne regresije, za funkcije vjerodostojnosti nepoznatih parametra modela vrijedi (4.5) i (4.6). Zbog toga je ovdje normalna distribucija prirodan izbor za apriornu distribuciju od $\beta_{\bar{x}}$ i $\beta_1$. Prije nego što krenemo na određivanje aposteriorne distribucije, potrebno je odrediti prikladne parametre očekivanja i standardne devijacije za spomenute normalne distribucije.

Prednost korištenja parametrizacije (3.10) modela jednostavne linearne regresije je u tome što imamo intuitivnije apriorno znanje o $\beta_{\bar{x}}$, nego o $\beta_0$. Na temelju naših vjerovanja o tome koja je srednja vrijednost varijable odaziva $Y$, odredimo parametar očekivanja apriorne distribucije od $\beta_{\bar{x}}$ i označimo ga s $m_{\beta_{\bar{x}}}$. Odredimo sada parametar standardne devijacije apriorne distribucije od $\beta_{\bar{x}}$, u oznaci $s_{\beta_{\bar{x}}}$. Prema metodi koja je opisana u [5], potrebno je razmotriti točke iz uzorka te odrediti gornju i donju granicu mogućih vrijednosti slučajne varijable Y. Parametar standardne devijacije $s_{\beta_{\bar{x}}}$ dobijemo tako da duljinu intervala s tim granicama podijelimo sa 6. Na taj smo način dobili razumnu vjerojatnosnu distribuciju na cijelom rasponu vrijednosti varijable odaziva za koje vjerujemo da su moguće.

Obično nas više zanima parametar $\beta_1$ jer ponekad želimo utvrditi može li poprimiti vrijednost 0. Stoga, možemo odabrati $m_{\beta_1} = 0$ kao parametar očekivanja apriorne distribucije od $\beta_1$. Zatim razmislimo o gornjoj i donjoj granici učinka povećanja od x za jednu jedinicu na vrijednost varijable Y. Razliku između gornje i donje granice podijelimo sa 6 i dobijemo $s_{\beta_1}$, parametar standardne devijacije apriorne distribucije od $\beta_1$. U drugim slučajevima, ako imamo prethodno uvjerenje o koeficijentu smjera iz prijašnjih podataka, tada parametre $m_{\beta_1}$ i $s_{\beta_1}$ odabiremo upravo na temelju tih podataka te bismo koristili normalnu apriornu distribuciju $N(m_{\beta_1}, s_{\beta_1})$.

Primijenimo sada Bayesov teorem. Zajednička aposteriorna funkcija gustoća od $\beta_{\bar{x}}$ i $\beta_1$ je proporcionalna umnošku vjerodostojnosti i zajedničke apriorne gustoće, odnosno

$$g(\beta_{\bar{x}}, \beta_1 | y, x) \propto g(\beta_{\bar{x}}, \beta_1) \times L(\beta_{\bar{x}}, \beta_1).$$

Zajednička apriorna gustoća i funkcija vjerodostojnosti se obje mogu rastaviti na dijelove koji ovise samo o $\beta_{\bar{x}}$ i dijelove koji ovise samo o $\beta_1$. Njihova raspodjela rezultira zajedničkom aposteriornom funkcijom gustoće koja se može zapisati kao umnožak marginalnih aposteriornih gustoća:

$$\begin{aligned} g(\beta_{\bar{x}}, \beta_1 | y, x) &\propto g(\beta_{\bar{x}}) \times g(\beta_1) \times L(\beta_{\bar{x}}) \times L(\beta_1) \\ &\propto g(\beta_{\bar{x}} | y, x) \times g(\beta_1 | y, x), \end{aligned}$$

gdje smo u prvom koraku iskoristili (4.4) i (4.8).

Ažurirajuće pravilo. Budući da je zajednička aposteriorna funkcija gustoće produkt marginalnih, marginalne aposteriorne funkcije gustoće na jedinstven način određuju zajedničku aposteriornu gustoću. Marginalne aposteriorne gustoće možemo pronaći koristeći jednostavno ažurirajuće pravilo[1] prema kojem se revidiraju parametri očekivanja i varijance za normalnu distribuciju. Spomenuto pravilo se može primijeniti ako koristimo normalnu ili uniformnu apriornu distribuciju. Na primjer, ako koristimo $N(m_{\beta_1}, {s_{\beta_1}}^2)$ apriornu distribuciju za $\beta_1$, dobijemo $N(m'_{\beta_1}, (s'_{\beta_1})^2)$ aposteriornu distribuciju gdje je

$$\frac{1}{(s'_{\beta_1})^2} = \frac{1}{s^2_{\beta_1}} + \frac{S_{xx}}{\sigma^2} \tag{4.16}$$

i

$$m'_{\beta_1} = \frac{\frac{1}{s^2_{\beta_1}}}{\frac{1}{s^2_{\beta_1}} + \frac{S_{xx}}{\sigma^2}} m_{\beta_1} + \frac{\frac{S_{xx}}{\sigma^2}}{\frac{1}{s^2_{\beta_1}} + \frac{S_{xx}}{\sigma^2}} \hat{\beta}_1 = \frac{\frac{1}{s^2_{\beta_1}}}{\frac{1}{(s'_{\beta_1})^2}} m_{\beta_1} + \frac{\frac{S_{xx}}{\sigma^2}}{\frac{1}{(s'_{\beta_1})^2}} \hat{\beta}_1. \tag{4.17}$$

Preciznost aposteriorne distribucije jednaka je zbroju preciznosti apriorne distribucije i preciznosti funkcije vjerodostojnosti parametra $\beta_1$, gdje je *preciznost* distribucije definirana kao recipročna vrijednost varijance. Očekivanje aposteriorne distribucije jednako je

[1] Pravilo je detaljnije objašnjeno u [5].

težinskom prosjeku očekivanja apriorne distribucije i očekivanja funkcije vjerojatnosti parametra $\beta_1$, pri čemu su težine jednake pribrojnicima iz rastava aposteriorne preciznosti. Aposteriorna distribucija je također normalna.

Slično, ako koristimo $N(m_{\beta_{\bar{x}}}, s^2_{\beta_{\bar{x}}})$ apriornu distribuciju za $\beta_{\bar{x}}$, dobijemo normalnu aposteriornu distribuciju $N(m'_{\beta_{\bar{x}}}, (s'_{\beta_{\bar{x}}})^2)$ gdje je

$$\frac{1}{(s'_{\beta_{\bar{x}}})^2} = \frac{1}{s^2_{\beta_{\bar{x}}}} + \frac{n}{\sigma^2}$$

i

$$m'_{\beta_{\bar{x}}} = \frac{\frac{1}{s^2_{\beta_{\bar{x}}}}}{\frac{1}{(s'_{\beta_{\bar{x}}})^2}} m_{\beta_{\bar{x}}} + \frac{\frac{n}{\sigma^2}}{\frac{1}{(s'_{\beta_{\bar{x}}})^2}} \hat{\beta}_{\bar{x}}.$$

Zaključujemo da ako krenemo s normalnom apriornom distribucijom, dobivamo normalnu aposteriornu distribuciju. Ovo je svojstvo koje karakterizira takozvane *konjugirane apriorne distribucije*.

**Definicija 4.1.1.** *Za familiju $\mathcal{F}$ vjerojatnosnih distribucija na $\Theta$ kažemo da je konjugirana za statističku strukturu $\mathcal{P}$ ako za svaki $f \in \mathcal{F}$ također vrijedi $f(\theta|x) \in \mathcal{F}$, za svaki $x$. Tada apriornu i aposteriornu distribuciju nazivamo konjugiranim distribucijama, a apriornu funkciju gustoće nazivamo konjugiranim apriorom za funkciju vjerodostojnosti danu pomoću modela $\mathcal{P}$.*

## Problem revidiranja regresijske linije

Vidjeli smo da u slučaju kada odabiremo referentnu apriornu distribuciju za nepoznate parametre modela linearne regresije (varijanca nepoznata), dobijemo da za aposteriornu distribuciju parametara vrijedi

$$g(\beta, \alpha, \sigma^2|y, x) \propto (\sigma^2)^{-\frac{(n+2)}{2}} e^{-\frac{1}{2\sigma^2}[S_{yy} - 2\beta_1 S_{xy} + \beta_1{}^2 S_{xx} + n(\beta_{\bar{x}} - \bar{y})^2]}.$$

Prema [8] moguće je izvesti formule za rekurzivnu konstrukciju regresijskog pravca u slučaju kada uz postojeće podatke (na temelju kojih smo dobili procjene parametara i izračunali njihovu aposteriornu distribuciju) prikupimo dodatne informacije. Zbog razloga koji će se uskoro pojaviti, uvodimo male promjene u notaciji i vrijednosti dobivene iz podataka označimo sa $n'$, $\overline{x}'$, $\overline{y}'$, $\beta'_{\bar{x}}$, $\beta'_1$, $S'_{xx}$, $S'_{ee}$, i tako dalje.

Pretpostavimo sada da smo, uz postojeće podatke za model, prikupili dodatnih $n''$ podataka. Ako na trenutak zaboravimo da imamo prijašnje podatke i da su ovo jedini dostupni

podaci, konstruirali bismo regresijsku liniju na temelju podataka iz uzorka veličine $n''$. Ostale vrijednosti dobivene na temelju tog uzorka označili bismo sa $\overline{x}'', \overline{y}'', \beta''_{\bar{x}}, \beta''_1, S''_{xx}, S''_{ee}$, i tako dalje.

Međutim, ako bismo u obzir uzeli svih $(n' + n'')$ podataka, tada bi se regresijska linija temeljila na podacima za $n = n' + n''$, što bi rezultiralo sljedećim vrijednostima: $\overline{x}$, $\overline{y}$, $\beta_{\bar{x}}$, $\beta_1$, $S_{xx}$, $S_{ee}$, i tako dalje.

### Formule za rekurzivnu konstrukciju regresijske linije

Prije nego što krenemo na raspis formula za rekurzivnu konstrukciju regresijske linije, potreban nam je iskaz sljedećeg teorema kojeg navodimo bez dokaza.[2]

**Teorem 4.1.2.** *(Načelo dovoljnosti) Statistika $T = t(\mathbf{X})$ slučajnog uzorka $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je dovoljna za $\theta \in \Theta$ ako i samo ako*

$$L(\theta) \propto L_t(\theta),$$

*pri čemu je $L : \Theta \to \mathbb{R}$ funkciju vjerodostojnosti parametra $\theta$ uz dani uzorak $\mathbf{x}$, dok je $L_t : \Theta \to \mathbb{R}$ funkcija vjerodostojnosti parametra $\theta$ uz dano $t(\mathbf{x})$.*

Prema načelu dovoljnosti moguće je pronaći $\bar{x}, \bar{y}, \beta, \ldots$ iz vrijednosti $\bar{x}', \bar{y}', \beta', \ldots$ i $\bar{x}'', \bar{y}'', \beta'', \ldots$. Može se pokazati da su $n, \bar{x}$ i $\beta_{\bar{x}} = \bar{y}$ dani sljedećim izrazima

$$\begin{aligned} n &= n' + n'' \\ \bar{x} &= (n'\bar{x}' + n''\bar{x}'')/n \\ \bar{y} &= (n'\bar{y}' + n''\bar{y}'')/n, \end{aligned}$$

te ako definiramo

$$\begin{aligned} n^h &= (n'^{-1} + n''^{-1})^{-1} \\ S^c_{xx} &= n^h(\bar{x}' - \bar{x}'')^2 \\ S^c_{xy} &= n^h(\bar{x}' - \bar{x}'')(\bar{y}' - \bar{y}'') \\ S^c_{yy} &= n^h(\bar{y}' - \bar{y}'')^2 \\ \beta^c &= S^c_{xy}/S^c_{xx}, \end{aligned}$$

[2]Dokaz teorema i detaljnije pojašnjenje idućeg rezultata, kao i primjer koji slijedi u nastavku, možete pronaći u [8].

tada su $S_{xx}, \beta$ i $S_{ee}$ zadani sa

$$\begin{aligned}
S_{xx} &= S'_{xx} + S''_{xx} + S^c_{xx} \\
S_{yy} &= S'_{yy} + S''_{yy} + S^c_{yy} \\
\beta &= (\beta' S'_{xx} + \beta'' S''_{xx} + \beta^c S^c_{xx})/S_{xx} \\
S_{ee} &= S'_{ee} + S''_{ee} + \left[(\beta' + \beta'')^2 S'_{xx} S''_{xx} + (\beta'' + \beta^c)^2 S''_{xx} S^c_{xx} + (\beta^c + \beta')^2 S^c_{xx} S'_{xx}\right]/S_{xx}.
\end{aligned}$$

Između navedenih formula, jedina koju je kompliciranije izvesti je posljednja. Pokažimo stoga kako smo došli do formule za $S_{ee}$.

$$\begin{aligned}
S_{ee} &= S_{yy} - \beta^2 S_{xx} \\
&= S'_{yy} + S''_{yy} + S^c_{yy} - \beta^2 S_{xx} \\
&= S'_{ee} + S''_{ee} + \beta'^2 S'_{xx} + \beta''^2 S''_{xx} + \beta^{c2} S^c_{xx} - \beta^2 S_{xx}
\end{aligned}$$

Nadalje,

$$\begin{aligned}
&S_{xx}(\beta'^2 S'_{xx} + \beta''^2 S''_{xx} + \beta^{c2} S^c_{xx} - \beta^2 S_{xx}) \\
&\quad = (S'_{xx} + S''_{xx} + S^c_{xx})(\beta'^2 S'_{xx} + \beta''^2 S''_{xx} + \beta^{c2} S^c_{xx}) - (\beta' S'_{xx} + \beta'' S''_{xx} + \beta^c S^c_{xx})^2 \\
&\quad = \beta'^2 S'_{xx} S''_{xx} + \beta''^2 S''_{xx} - 2\beta'\beta'' S'_{xx} S''_{xx} \\
&\quad + \beta''^2 S''_{xx} S^c_{xx} + \beta^{c2} S''_{xx} S^c_{xx} - 2\beta''\beta^c S''_{xx} S^c_{xx} \\
&\quad + \beta^{c2} S^c_{xx} S'_{xx} + \beta'^2 S^c_{xx} S'_{xx} - 2\beta^c\beta' S^c_{xx} S'_{xx} \\
&\quad = (\beta' + \beta'')^2 S'_{xx} S''_{xx} + (\beta'' + \beta^c)^2 S''_{xx} S^c_{xx} + (\beta^c + \beta')^2 S^c_{xx} S'_{xx}
\end{aligned}$$

što daje traženi rezultat.

Pokažimo na konkretnom primjeru kako bismo iskoristili formule za rekurzivnu konstrukciju regresijske linije.

**Primjer 4.1.3.** *Na temelju izmjerenih podataka o količini oborina u studenome za grad New York, želimo predvidjeti količina oborina za mjesec prosinac. Podaci (dani u mm) su prikazani u tablici ispod.*

| *godina (i)* | *1971* | *1972* | *1973* | *1974* | *1975* | *1976* | *1977* | *1978* | *1979* | *1980* |
|---|---|---|---|---|---|---|---|---|---|---|
| *studeni* ($x'_i$) | *23.9* | *43.3* | *36.3* | *40.6* | *57.0* | *52.5* | *46.1* | *142.0* | *112.6* | *23.7* |
| *prosinac* ($y'_i$) | *41.0* | *52.0* | *18.7* | *55.0* | *40.0* | *29.2* | *51.0* | *17.6* | *46.6* | *57.0* |

*Iz* $n' = 10$, $\bar{x}' = 57.8$, $\bar{y}' = 40.81$, $S'_{xx} = 13538.66$, $S'_{yy} = 1889.089$ *i* $S'_{xy} = -2183.51$ *te pomoću formula* (3.6) *i* (3.7)*, dobijemo procjene parametara regresijskog pravca. Dakle,*

$$\hat{\beta}'_0 = 50.132 \quad i \quad \hat{\beta}'_1 = -0.1613.$$

*Procjena regresijskog pravca je* $\hat{y}' = 50.132 - 0.1613x$. *Ako zapisujemo u alternativnom obliku, uz* $\hat{\beta}'_{\bar{x}} = 40.81$, *jednadžba glasi:* $\hat{y}' = 40.81 - 0.1613(x - 57.8)$. *Nadalje,*

$$S'_{ee} = S'_{yy} - (S'_{xy})^2/S'_{xx} = 1536.933$$

*iz čega lako dobijemo procjenu varijance oko regresijskog pravca:* $\hat{\sigma}' = \sqrt{S'_{ee}/(n'-2)} = 13.861$. *Zajednička aposteriorna distribucija nepoznatih parametara modela je dana s:*

$$g(\beta_{\bar{x}}, \beta_1, \sigma^2 \mid \mathbf{y}', \mathbf{x}') \propto (\sigma^2)^{-(n'+2)/2} e^{-\frac{1}{2\sigma^2}[S'_{ee} + S'_{xx}(\beta_1 + \hat{\beta}'_1)^2 + n'(\beta_{\bar{x}} - \hat{\beta}'_{\bar{x}})^2]},$$

*a Bayesovske intervale pouzdanosti možemo odrediti pomoću tvrdnji* (4.12), (4.13) *i* (4.14).

*Pretpostavimo sada da smo prikupili dodatne podatke:*

| *godina (i)* | *1981* | *1982* | *1983* | *1984* | *1985* | *1986* |
|---|---|---|---|---|---|---|
| *studeni* ($x''_i$) | *34.1* | *62.0* | *106.9* | *34.1* | *68.3* | *81.0* |
| *prosinac* ($y''_i$) | *12.3* | *90.4* | *28.8* | *106.2* | *62.3* | *50.5* |

*Ako bismo računali procjene nepoznatih parametara samo na temelju ovih podataka za* $n'' = 6$, *dobili bismo sljedeće:*

$$\bar{x}'' = 64.4,\ \hat{\beta}'_{\bar{x}} = \bar{y}'' = 58.42,\ \hat{\beta}'_1 = -0.381,\ S''_{xx} = 3938.96 \ \ i \ \ S''_{ee} = 5814.9.$$

*Međutim, ako imamo podatke o količini oborina za svih 16 godina, procjena regresijskog pravca bi se računala uzimajući u obzir sve podatke i u tom slučaju bismo dobili:*

$$\bar{x} = 60.28,\ \hat{\beta}_{\bar{x}} = \bar{y} = 47.41,\ \hat{\beta}_1 = -0.184,\ S_{xx} = 17640.97 \ \ i \ \ S_{ee} = 8840.66.$$

*Formule za rekurzivnu konstrukciju regresijske linije nam omogućavaju da direktno, i bez dodatnog ponavljanja cijelog postupka, dođemo do prethodnog rezultata korištenjem vrijednosti dobivenih iz podataka za prvih 10 godina i vrijednosti iz podataka za dodatnih 6 godina. Dakle, vrijedi:*

$$\begin{aligned} n &= n' + n'' = 16 \\ \bar{x} &= (n'\bar{x}' + n''\bar{x}'')/n = 60.28 \\ \bar{y} &= (n'\bar{y}' + n''\bar{y}'')/n = 47.41 \end{aligned}$$

*Zatim, dobijemo da je*

$$n^h = 3.75,\ S^c_{xx} = 163.35,\ S^c_{xy} = 435.6,\ S^c_{xx} = 2.667,$$

*stoga je*

$$S_{xx} = 17640.97,\ \hat{\beta}_{\bar{x}} = -0.184\ S_{ee} = 8840.66,$$

*što je u skladu s ranije prikazanim rezultatima dobivenim razmatranjem podataka za svih 16 godina zajedno.*

## Bayesovski interval vjerodostojnosti za koeficijent smjera

Aposteriorna distribucija parametra $\beta_1$ rezimira naša ukupna vjerovanja o tom parametru na temelju podataka iz uzorka za model jednostavne linearne regresije. Cjelokupna vjerovanja o koeficijentu smjera regresijskog pravca možemo sažeti u $(1-\alpha)\cdot 100\%$ Bayesovskom intervalu vjerodostojnosti za $\beta_1$.

$(1-\alpha)\cdot 100\%$ Bayesovski interval vjerodostojnosti parametra označava interval za koji je uvjetna vjerojatnost da sadrži stvarnu vrijednost parametra jednaka $(1-\alpha)\cdot 100\%$, uz dani slučajni uzorak. Uočavamo da postoji razlika u interpretaciji frekvencionističkog pouzdanog intervala i Bayesovskog intervala vjerodostojnosti, budući da pouzdani interval opisujemo kao interval za kojeg smo $(1-\alpha)\cdot 100\%$ sigurni da sadrži pravu, ali nepoznatu vrijednost parametra. Zbog toga možemo reći da Bayesovski interval vjerodostojnost daje razumljiviji i korisniji odgovor od intervala pouzdanosti.

**Definicija 4.1.4.** *Neka je $X_1, \ldots, X_n$ slučajan uzorak iz distribucije s nepoznatim parametrom $\theta$. Neka je $\alpha \in [0,1]$ te neka su $a$ i $b$ realni brojevi takvi da*

$$\int_{-\infty}^{a} f(\theta \mid x_1, \ldots, x_n)d\theta = \frac{\alpha}{2} = \int_{b}^{\infty} f(\theta \mid x_1, \ldots, x_n)d\theta.$$

*Tada vrijedi*

$$\mathbb{P}(a \leq \theta \leq b \mid x_1, \ldots, x_n) = \int_{a}^{b} f(\theta \mid x_1, \ldots, x_n)d\theta = 1-\alpha.$$

*Interval $[a,b]$ nazivamo $(1-\alpha)\cdot 100\%$ Bayesovski interval vjerodostojnosti za nepoznati parametar $\theta$.*

Ako smo za nepoznate parametre izabrali referentnu apriornu distribuciju, prilikom konstrukcije intervala vjerodostojnosti za koeficijent smjera koristit će se tvrdnja (4.13) u slučaju poznate varijance, odnosno tvrdnja u (4.15) kada je varijanca nepoznata. U nastavku ćemo se ograničiti na slučaj kada biramo normalnu apriornu distribuciju za nepoznate parametre modela i prikazat ćemo kako u toj situaciji glasi Bayesovski interval vjerodostojnosti za $\beta_1$. Međutim, vrijedi napomenuti da će u oba slučaja postupak određivanja intervala vjerodostojnosti biti vrlo sličan.

Pokazali smo da ako za $\beta_1$ koristimo $N(m_{\beta_1}, {s_{\beta_1}}^2)$ normalnu apriornu distribuciju, dobijemo normalnu aposteriornu distribuciju $N(m'_{\beta_1}, (s'_{\beta_1})^2)$, pri čemu parametre $m'_{\beta_1}$ i ${s'_{\beta_1}}^2$ određujemo prema pravilu koje je opisano s (4.16) i (4.17). Pridružimo sada slučajnoj varijabli $\beta_1$ njenu *standardiziranu verziju*

$$Z = \frac{\beta_1 - m'_{\beta_1}}{s'_{\beta_1}} \sim N(0,1).$$

Budući da slučajna varijabla $Z$ ima standardnu jediničnu normalnu razdiobu, slijedi da je za $\alpha \in [0, 1]$

$$\mathbb{P}\Big( - z_{\frac{\alpha}{2}} \leq \frac{\beta_1 - m'_{\beta_1}}{s'_{\beta_1}} \leq z_{\frac{\alpha}{2}} \Big) = 1 - \alpha,$$

odnosno

$$\mathbb{P}\Big( m'_{\beta_1} - z_{\frac{\alpha}{2}} \cdot s'_{\beta_1} \leq \beta_1 \leq m'_{\beta_1} + z_{\frac{\alpha}{2}} \cdot s'_{\beta_1} \Big) = 1 - \alpha,$$

gdje je vrijednosti $z_{\frac{\alpha}{2}}$ $\frac{\alpha}{2}$-kvantil standardne normalne distribucije koju određujemo iz tablice standardne normalne distribucije dane u Dodatku B. Dakle, $(1 - \alpha) \cdot 100\%$ Bayesovski interval vjerodostojnosti za koeficijent smjera dan je s

$$\Big[ m'_{\beta_1} - z_{\frac{\alpha}{2}} \cdot s'_{\beta_1}, m'_{\beta_1} + z_{\frac{\alpha}{2}} \cdot s'_{\beta_1} \Big]. \tag{4.18}$$

U stvarnosti često ne znamo vrijednost varijance uzorka i u tom je slučaju razuman pristup korištenje $\hat{\sigma}^2$, procjene za $\sigma^2$ izračunate iz reziduala (vidi (3.12)). Međutim, tada moramo proširiti interval pouzdanosti kako bismo uzeli u obzir povećanu nesigurnost zbog nepoznavanja vrijednosti $\sigma^2$. To činimo tako da umjesto standardizirane verzije od $\beta_1$ koristimo *studentiziranu* verziju $T = \frac{\beta_1 - m'_{\beta_1}}{\hat{\sigma}}$. $T$ ima Studentovu $t$-razdiobu s $n - 2$ stupnjem slobode[3] i vrijedi

$$\mathbb{P}\Big( - t_{\frac{\alpha}{2}} \leq \frac{\beta_1 - m'_{\beta_1}}{\hat{\sigma}} \leq t_{\frac{\alpha}{2}} \Big) = \mathbb{P}\Big( m'_{\beta_1} - t_{\frac{\alpha}{2}} \cdot s'_{\beta_1} \leq \beta_1 \leq m'_{\beta_1} + t_{\frac{\alpha}{2}} \cdot s'_{\beta_1} \Big) = 1 - \alpha,$$

pri čemu je $t_{\frac{\alpha}{2}}$ $\frac{\alpha}{2}$-kvantil Studentove $t$-distribucije s $n - 2$ stupnjem slobode. $(1 - \alpha) \cdot 100\%$ interval vjerodostojnosti je tada dan s

$$\Big[ m'_{\beta_1} - t_{\frac{\alpha}{2}} \cdot s'_{\beta_1}, m'_{\beta_1} + t_{\frac{\alpha}{2}} \cdot s'_{\beta_1} \Big],$$

gdje vrijednosti $t_{\frac{\alpha}{2}}$ čitamo iz tablice Studentove $t$-distribucije.

### Frekvencionistički interval pouzdanosti za koeficijent smjera

**Definicija 4.1.5.** *Neka su $L_n = l_n(X_1, \ldots, X_n)$ i $D_n = d_n(X_1, \ldots, X_n)$ statistike slučajnog uzorka $X_1, \ldots, X_n$. Za $[L_n, D_n]$ kažemo da je $(1 - \alpha)100\%$ pouzdani interval za parametar $\theta$ ako vrijedi*

$$\mathbb{P}(L_n \leq \theta \leq D_n) \geq 1 - \alpha, \quad \alpha \in \langle 0, 1 \rangle.$$

[3] Kako je navedeno u [5], zapravo za nepoznati parametar $\sigma^2$ koristimo apriornu funkciju gustoće za koju vrijedi $f(\sigma^2) \propto (\sigma^2)^{-1}$. Marginalna aposteriorna gustoća od $\beta_1$ se tada može pronaći integriranjem po $\sigma^2$ iz zajedničke aposteriorne funkcije gustoće.

Kada je varijanca $\sigma^2$ nepoznata, $(1-\alpha)\cdot 100\%$ interval pouzdanosti za koeficijent smjera $\beta_1$ je oblika

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right],$$

gdje je $\hat{\sigma}^2$ procjena varijance izračunata iz reziduala pravca prilagođenog metodom najmanjih kvadrata. Ako koristimo uniformnu apriornu distribuciju za $\beta_1$ i $\beta_{\bar{x}}$, interval pouzdanosti je istog oblika kao Bayesovski vjerodostojni interval. Naravno interpretacija je drugačija. Pod frekvencionističkim pretpostavkama mi smo sigurni da za barem $(1-\alpha)\cdot 100\%$ realizacija slučajnog uzorka $(X_1, \ldots, X_n)$, interval pouzdanosti sadrži stvarnu vrijednost nepoznatog parametra. S druge strane, Bayesovski $(1-\alpha)\cdot 100\%$ interval vjerodostojnosti predstavlja interval za koji je uvjetna vjerojatnost da sadrži pravu vrijednost parametra uz dane podatke jednaka $(1-\alpha)$.

## Testiranje hipoteza za koeficijent smjera

*Statistička hipoteza* je pretpostavka o populacijskoj razdiobi promatrane varijable. Osnovna hipoteza koja se testira zove se *nulhipoteza* i označava se sa $H_0$. Nulhipoteza često predstavlja aktualno znanje o vrijednostima parametara ili neutralnu izjavu. Uz nulhipotezu, postavlja se i njoj *alternativna hipoteza* koju označavamo sa $H_1$.

*Statistički test* je pravilo podjele prostora vrijednosti uzoraka na dva podskupa: na područje vrijednosti uzoraka koji su konzistentni sa $H_0$, i na njegov komplement u kojem se nalaze vrijednosti nekonzistentne sa $H_0$.

Odluka o odbacivanju ili ne odbacivanju nulhipoteze donosi se na osnovi vrijednosti testne statistike. Područje vrijednosti koje testna statistika poprima dijeli se na područje vrijednosti koje su konzistentne sa $H_0$ i na područje nekonzistentno sa $H_0$. Područje testne statistike koje je nekonzistentno sa $H_0$ zove se *kritično područje*. Dakle, ako se opažena vrijednost testne statistike nalazi u kritičnom području, $H_0$ se odbacuje (u korist $H_1$).

*Razina značajnosti testa* $\alpha$ je vjerojatnost odbacivanja $H_0$ ako je $H_0$ istinita hipoteza. Pojmovi su preuzeti iz [2].

### Jednostrane hipoteze

Često želimo utvrditi ako je utjecaj varijable $x$ na promjenu u varijabli $Y$, kada se $x$ poveća za jednu jedinicu, veći od neke vrijednosti koju ćemo označiti s $\beta$. Pretpostavimo da je varijanca poznata. Testiramo:

$$H_0 : \beta_1 \leq \beta$$

$$H_1 : \beta_1 > \beta$$

na razini značajnosti $\alpha$ na Bayesovski način. Za provođenje Bayesovskog statističkog testa, računamo aposteriornu vjerojatnost nulte hipoteze. Dakle,

$$\mathbb{P}(\beta_1 \leq \beta \mid \mathbf{x}, \mathbf{y}) = \int_{-\infty}^{\beta} g(\beta_1 \mid \mathbf{x}, \mathbf{y}) d\beta_1 = \mathbb{P}\left(Z \leq \frac{\beta - m'_{\beta_1}}{s'_{\beta_1}}\right).$$

Ako je vjerojatnost manja od $\alpha$, odbacujemo $H_0$ i možemo zaključiti da je koeficijent smjera $\beta_1$ zaista veći od $\beta$.

**Napomena 4.1.6.** *Ako nam nije poznata vrijednost varijance i koristimo njenu procjenu, tada bismo koristili Studentovu t-razdiobu s $n-2$ stupnjeva slobode umjesto standardne normalne slučajne varijable Z.*

### Dvostrane hipoteze

Ako je $\beta_1 = 0$, to bi značilo da očekivanje slučajne varijable $Y$ uopće ne ovisi o vrijednosti $x$. Želimo na Bayesovski način testirati

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

na razini značajnosti $\alpha$, prije nego što upotrijebimo regresijski model za predikcije. Kako bismo sproveli Bayesovski statistički test, dovoljno je pogledati gdje se 0 nalazi u odnosu na vjerodostojni interval. Ako se nalazi izvan intervala, odbacujemo $H_0$. U suprotnom, ne možemo odbaciti nultu hipotezu i ne bismo trebali koristiti regresijski model kao pomoć u predviđanjima vrijednosti varijable odaziva za dane vrijednosti eksplanatorne varijable $x$.

**Primjer 4.1.7.** ***(nastavak na Primjer 2.1.1)*** *Pretpostavimo da unos kisika s obzirom na broj otkucaja srca dolazi iz normalne razdiobe, gdje je varijanca $\sigma^2 = 0.13^2$ poznata. Odredimo aposteriornu distribuciju od $\beta_1$ ako za taj parametar koristimo $N(0, 1^2)$ apriornu distribuciju. Iskoristimo u tu svrhu* (4.16) *i* (4.17).
*Aposteriorna preciznost od $\beta_1$ jednaka je:*

$$\frac{1}{(s'_{\beta_1})^2} = \frac{1}{1^2} + \frac{1486}{0.13^2} = 87929.99408$$

*pa je aposteriorna standardna devijacija jednaka:*

$$s'_{\beta_1} = 0.00337$$

*Izračunajmo još aposteriorno očekivanje od* $\beta_1$.

$$m'_{\beta_1} = \frac{1/1^2}{1/(0.00337)^2} \times 0 + \frac{1486/0.13^2}{1/(0.00337)^2} \times 0.042651 = 0.0427$$

*Dakle, aposteriorna distribucija od* $\beta_1$ *je normalna s očekivanjem* $0.0427$ *i varijancom* $0.00337^2$.

*Nadalje, pronađimo procjenu 95%-tnog Bayesovskog intervala vjerodostojnost za* $\beta_1$. *Bayesovski* $(1-\alpha)\cdot 100\%$ *interval vjerodostojnosti je oblika*

$$\left[m'_{\beta_1} - z_{\frac{\alpha}{2}} \cdot s'_{\beta_1}, m'_{\beta_1} + z_{\frac{\alpha}{2}} \cdot s'_{\beta_1}\right],$$

*pri čemu su* $m'_{\beta_1}$ *i* $s'_{\beta_1}$ *očekivanje i standardna devijacija aposteriorne distribucije. Dakle, procjena 95%-tnog Bayesovskog intervala vjerodostojnosti je*

$$[0.0427 - 1.96 \cdot 0.00337, 0.0427 + 1.96 \cdot 0.00337] = [0.03609, 0.04931]\,.$$

*Provedimo sada Bayesovski statistički test hipoteza*

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

*na razini značajnosti od 5%. Riječ je o testiranju dvostranih hipoteza pa za testiranje koristimo 95%-tni interval vjerodostojnosti. Primjećujemo da se vrijednost* $\beta_1 = 0$ *nalazi izvan 95%-tnog intervala vjerodostojnosti što znači da odbacujemo nultu hipotezu na razini značajnosti od 5%.*

## Prediktivna distribucija za buduće opservacije

Procjena predviđanja budućih vrijednosti odaziva za određenu vrijednost eksplanatorne varijable $x$ jedna je od glavnih svrha linearne regresije. Često, nakon što smo iz podataka ustanovili da postoji linearan odnos između eksplanatorne varijable x i varijable odaziva $Y$, želimo iskoristiti taj odnos za izradu predviđanja sljedeće vrijednosti $y_{n+1}$ za fiksnu sljedeću vrijednost eksplanatorne varijable $x_{n+1}$. Dakle, nas zanima predviđanje za $Y$ uz dano $x = x_{n+1}$. U tu svrhu uvodimo oznaku $Y_{n+1}$. Najbolje predviđanje za $Y_{n+1}$ bit će

$$\tilde{Y}_{n+1} = \hat{\beta}_{\bar{x}} + \hat{\beta}_1(x_{n+1} - \bar{x}),$$

gdje je $\hat{\beta}_1$ procjena koeficijenta smjera, a $\hat{\beta}_{\bar{x}}$ procjena regresijskog pravca u točki $x = \bar{x}$. Postavlja se pitanje koliko je dobro to predviđanje. Dva su izvora neizvjesnosti. Prvo, koristimo procijenjene vrijednosti parametara u predviđanju, a ne prave vrijednosti koje su

nepoznate. Smatramo da su parametri slučajne varijable i u jednom od prethodnih odjeljaka smo pronašli njihovu aposteriornu distribuciju. Drugo, nova opsevacija $Y_{n+1}$ sadrži vlastitu pogrešku opažanja $\varepsilon_{n+1}$ koja će biti neovisna o svim prethodnim pogreškama opažanja. *Prediktivna distribucija* sljedeće opservacije $Y_{n+1}$ uz dano $x = x_{n+1}$ i uz dane podatke, u oznaci $f(y_{n+1}|x_{n+1}, \mathbf{x}, \mathbf{y})$, uzima u obzir oba izvora nesigurnosti i pronalazi se pomoću Bayesovog teorema.

**Pronalazak prediktivne distribucije**

Prediktivna distribucija pronalazi se integracijom po parametrima $\beta_{\bar{x}}$ i $\beta_1$ iz zajedničke aposteriorne distribucije varijable $Y_{n+1}$ te parametara modela uz danu vrijednost eksplanatorne varijable $x = x_{n+1}$ i uz prethodna opažanja iz modela, $(x_1, y_1), \ldots, (x_n, y_n)$. Vrijedi

$$f(y_{n+1} \mid x_{n+1}, \mathbf{x}, \mathbf{y}) = \int\int f(y_{n+1}, \beta_{\bar{x}}, \beta_1 | x_{n+1}, \mathbf{x}, \mathbf{y}) d\beta_{\bar{x}} d\beta_1.$$

Gornji postupak integracije po parametrima smetnje iz zajedničke aposteriorne funkcije gustoće poznat je pod nazivom *marginalizacija*. Ova metoda, koja uvijek funkcionira, je jedna od očitih prednosti Bayesovske statistike jer nam omogućava da se uspješno nosimo s parametrima smetnje. Kada pronađemo prediktivnu distribuciju, sve parametre smatramo parametrima smetnje.

Prvo je potrebno odrediti zajedničku aposteriornu gustoću parametara i sljedeće opservacije za varijablu odaziva uz danu vrijednost $x_{n+1}$ i podatke:

$$f(y_{n+1}, \beta_{\bar{x}}, \beta_1 \mid x_{n+1}, \mathbf{x}, \mathbf{y}) = f(y_{n+1} \mid \beta_{\bar{x}}, \beta_1, x_{n+1}, \mathbf{x}, \mathbf{y}) \times g(\beta_{\bar{x}}, \beta_1 \mid x_{n+1}, \mathbf{x}, \mathbf{y}).$$

Odaziv $Y_{n+1}$ je uz dane $\beta_{\bar{x}}, \beta_1$ i $x_{n+1}$ još jedna slučajna varijabla iz regresijskog modela te je nezavisna od ostalih varijabli $Y_1, \ldots, Y_n$. Zbog toga, uz dane parametre modela, varijabla odaziva $Y_{n+1}$ ne ovisi o prethodnim podacima. Zajednička aposteriorna funkcija gustoće za $\beta_{\bar{x}}$ i $\beta_1$ koju smo izračunali na temelju podataka ne ovisi o sljedećoj vrijednosti eksplanatorne varijable $x_{n+1}$. Stoga možemo pojednostaviti zajedničku aposteriornu gustoću parametara i varijable odaziva $Y_{n+1}$ kao:

$$f(y_{n+1}, \beta_{\bar{x}}, \beta_1 \mid x_{n+1}, \mathbf{x}, \mathbf{y}) = f(y_{n+1} \mid \beta_{\bar{x}}, \beta_1, x_{n+1}) \times g(\beta_{\bar{x}}, \beta_1 \mid \mathbf{x}, \mathbf{y}),$$

što je zapravo umnožak gustoće od $Y_{n+1}$ uz dane parametre i aposteriorne funkcije gustoće parametara uz dane prethodne podatke. Prema pretpostavkama modela, $Y_{n+1}$ je normalno distribuirana s očekivanjem $\mu_{n+1} = \beta_{\bar{x}} + \beta_1(x_{n+1} - \bar{x})$ i poznatom varijancom $\sigma^2$.

Aposteriorne distribucije parametara $\beta_{\bar{x}}, \beta_1$ su, uz odabir normalnih apriornih distribucija i s obzirom na prethodne podatke, nezavisne i u jednom od prethodnih odjeljaka smo ih pronašli pomoću ažurirajućih pravila. Dobili smo da je

$$\beta_{\bar{x}} \sim N(m'_{\beta_{\bar{x}}}, (s'_{\beta_{\bar{x}}})^2) \quad \text{i} \quad \beta_1 \sim N(m'_{\beta_1}, (s'_{\beta_1})^2).$$

Budući da varijabla $Y_{n+1}$ ovisi o parametrima modela samo kroz linearnu funkciju, uvodimo supstituciju

$$\mu_{n+1} = \beta_{\bar{x}} + \beta_1(x_{n+1} - \bar{x})$$

kako bismo mogli pojednostaviti problem i pretpostaviti da je $\mu_{n+1}$ jedini nepoznati parametar. Parametri $\beta_{\bar{x}}$ i $\beta_1$ su nezavisni te će stoga aposteriorna distribucija od $\mu_{n+1}$ biti normalna s očekivanjem $m'_\mu = m'_{\beta_{\bar{x}}} + (x_{n+1} - \bar{x}) \cdot m'_{\beta_1}$ i varijancom $(s'_\mu)^2 = (s'_{\beta_{\bar{x}}})^2 + (x_{n+1} - \bar{x})^2 \cdot (s'_{\beta_1})^2$.

Prediktivnu distribuciju ćemo pronaći marginalizacijom parametra $\mu_{n+1}$ iz zajedničke aposteriorne funkcije gustoće od $Y_{n+1}$ i $\mu_{n+1}$.

$$\begin{aligned}
f(y_{n+1}|x_{n+1}, \mathbf{x}, \mathbf{y}) &= \int f(y_{n+1}, \mu_{n+1}|x_{n+1}, \mathbf{x}, \mathbf{y}) d\mu_{n+1} \\
&= \int f(y_{n+1}|\mu_{n+1}, x_{n+1}, \mathbf{x}, \mathbf{y}) \cdot g(\mu_{n+1}|x_{n+1}, \mathbf{x}, \mathbf{y}) d\mu_{n+1} \\
&= \int f(y_{n+1}|\mu_{n+1}) \cdot g(\mu_{n+1}|x_{n+1}, \mathbf{x}, \mathbf{y}) d\mu_{n+1} \\
&\propto \int e^{-\frac{1}{2\sigma^2}(y_{n+1}-\mu_{n+1})^2} \cdot e^{-\frac{1}{2(s'_\mu)^2}(\mu_{n+1}-m'_\mu)^2} d\mu_{n+1} \\
&\propto \int e^{-\frac{1}{2\sigma^2(s'_\mu)^2/(\sigma^2+(s'_\mu)^2)}\left[\mu_{n+1}-\frac{y_{n+1}(s'_\mu)^2+m'_\mu\sigma^2}{(s'_\mu)^2+\sigma^2}\right]^2} \cdot e^{-\frac{1}{2((s'_\mu)^2+\sigma^2)}(y_{n+1}-m'_\mu)^2} d\mu_{n+1}.
\end{aligned}$$

Drugi faktor ne ovisi o $\mu_{n+1}$ i možemo ga izlučiti ispred integrala, a prvi izraz se može integrirati i tako na kraju ostaje :

$$f(y_{n+1}|x_{n+1}, \mathbf{x}, \mathbf{y}) \propto e^{-\frac{1}{2((s'_\mu)^2+\sigma^2)}(y_{n+1}-m'_\mu)^2}. \tag{4.19}$$

Prepoznajemo da je ovdje riječ o normalnoj distribuciji $N(m'_y, (s'_y)^2)$, gdje je $m'_y = m'_\mu$ i $(s'_y)^2 = (s'_\mu)^2 + \sigma^2$. Stoga je prediktivno očekivanje varijable $Y_{n+1}$ jednako aposteriornom očekivanju od $\mu_{n+1} = \beta_{\bar{x}} + \beta_1(x_{n+1} - \bar{x})$, a prediktivna varijanca od $Y_{n+1}$ je zbroj aposteriorne varijance od $\mu_{n+1} = \beta_{\bar{x}} + \beta_1(x_{n+1} - \bar{x})$ i varijance opaženog uzorka $\sigma^2$. Dakle, u prediktivnoj distribuciji su oba izvora nesigurnosti uzeta u obzir .

### Bayesovski interval vjerodostojnosti za predikcije

Često želimo pronaći interval čija je aposteriorna vjerojatnost da sadrži vrijednost sljedeće opservacije varijable odaziva $y_{n+1}$ za danu vrijednost $x_{n+1}$ jednaka $1 - \alpha$. To će biti $(1-\alpha)\cdot 100\%$ Bayesovski interval vjerodostojnosti za predviđanje varijable odaziva uz danu vrijednost eksplanatorne varijable. Od prije znamo da su očekivanje i varijanca prediktivne distribucije redom jednaki $m'_y$ i $(s'_y)^2$. Bayesovski interval vjerodostojnosti za predviđanje

je zadan pomoću svojih granica

$$\begin{aligned} m'_y \pm z_{\frac{\alpha}{2}} \cdot (s'_y)^2 &= m'_\mu \pm z_{\frac{\alpha}{2}} \cdot \sqrt{(s'_\mu)^2 + \sigma^2} \\ &= m'_{\beta_{\bar{x}}} + m'_{\beta_1} \cdot (x_{n+1} - \bar{x}) \pm z_{\frac{\alpha}{2}} \cdot \sqrt{(s'_{\beta_{\bar{x}}})^2 + (s'_{\beta_1})^2 \cdot (x_{n+1} - \bar{x})^2 + \sigma^2}, \end{aligned}$$

u slučaju kada je varijanca poznata. Ako ne znamo vrijednost varijance uzorka, koristimo procjenu varijance izračunatu iz reziduala, a Bayesovski vjerodostojni interval je tada određen s granicama:

$$\begin{aligned} m'_y \pm t_{\frac{\alpha}{2}} \cdot (s'_y)^2 &= m'_\mu \pm t_{\frac{\alpha}{2}} \cdot \sqrt{(s'_\mu)^2 + \hat{\sigma^2}} \\ &= m'_{\beta_{\bar{x}}} + m'_{\beta_1} \cdot (x_{n+1} - \bar{x}) \pm t_{\frac{\alpha}{2}} \cdot \sqrt{(s'_{\beta_{\bar{x}}})^2 + (s'_{\beta_1})^2 \cdot (x_{n+1} - \bar{x})^2 + \hat{\sigma^2}}, \end{aligned}$$

pri čemu smo kritičnu vrijednost dobili iz Studentove $t$-distribucije s $n-2$ stupnjem slobode. Ovi intervali vjerodostojnosti za predviđanje su Bayesovski analogoni frekvencionističkih intervala pouzdanosti, budući da dopuštaju i pogrešku procjene i pogrešku opažanja. Bayesovski intervali vjerodostojnosti za predviđanje će općenito biti kraći od odgovarajućih frekvencionističkih pouzdanih intervala predviđanja jer Bayesovski intervali koriste informacije iz apriorne distribucije kao i informacije iz podataka. Oba pristupa daju potpuno iste rezultate kada se koriste uniformne apriorne distribucije za koeficijent smjera i za presjek.

## 4.2 Bayesovsko zaključivanje za višestruku linearnu regresiju

Prelazimo na raspravu o Bayesovskom zaključivanju za višestruku linearnu regresiju. Za pronalazak zajedničke funkcije vjerodostojnosti vektora nepoznatih parametara $\beta$ koristit ćemo pretpostavke modela višestruke linearne regresije koje smo naveli u prošlom poglavlju. Zatim ćemo primijeniti Bayesov teorem kako bismo pronašli zajedničku aposteriornu distribuciju. U općem slučaju, to zahtijeva procjenu $(k+1)$-dimenzionalnog integrala ($k$ predstavlja broj eksplanatornih varijabli u modelu) koji se obično računa pomoću raznih numeričkih metoda integriranja. Međutim, pogledat ćemo dva slučaja u kojima možemo pronaći egzaktnu aposteriornu funkciju gustoće bez potrebe za numeričkom integracijom. U prvom slučaju upotrijebit ćemo nezavisne uniformne apriorne distribucije za sve komponente nepoznatog vektora parametara, dok ćemo u drugom slučaju koristiti konjugiranu apriornu distribuciju.

Nakon toga ćemo pokazati kako izvesti Bayesovske zaključke o parametrima modela višestruke linearne regresije. Odredit ćemo Bayesovske intervale vjerodostojnosti za pojedinačne parametre, a onda i Bayesovske regije vjerodostojnosti za vektor nepoznatih

parametara. Naposljetku ćemo se baviti problemom određivanja prediktivne distribucije za buduće opservacije. Kao izvor ideja za ovo poglavlje poslužio je [5].

Kao što smo spomenuli, naš model mora zadovoljavati neke pretpostavke kako bismo kasnije mogli donositi određene zaključke. Ovdje pretpostavljamo da su slučajne greške u modelu nezavisne i da za vektor stupaca grešaka vrijedi

$$\varepsilon \sim N(0, \sigma^2 I_n).$$

Slijedom navedenih pretpostavki, za distribuciju slučajnog vektora $\mathbf{Y}$ vrijedi:

$$\mathbf{Y} \sim N(X\beta, \sigma^2 I_n), \tag{4.20}$$

odnosno odaziv $\mathbf{Y}$ ima multivarijatnu normalnu razdiobu s očekivanjem koje je jednako umnošku vektora nepoznatih parametara i matrice eksplanatornih varijabli, a varijanca je zadana dijagonalnom kvadratnom matricom reda $n$, čiji su svi elementi na dijagonali jednaki $\sigma^2$. Funkcija gustoće vektora $\mathbf{Y}$ je dana s

$$f(\mathbf{y} \mid \beta, \sigma^2) = (2\pi\sigma^2)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)}.$$

**Napomena 4.2.1.** *Tvrdnju* (4.20) *možemo zapisati na drugačiji način kao*

$$\mathbf{Y} \sim MNV(X\beta, \sigma^2 I_n),$$

*gdje "MNV" označava da je riječ o multivarijatnoj normalnoj distribuciji.*

## Funkcija vjerodostojnosti za jednu opaženu vrijednost

Pod pretpostavkama modela, slučajna varijabla odaziva $Y_i$, $i = 1, \ldots, n$ je za dane vrijednosti eksplanatornih varijabli $x_{i1}, \ldots, x_{ik}$ normalna s očekivanjem $\mu_i$ i varijancom $\sigma^2$, gdje je očekivanje jednako

$$\mu_i = \sum_{j=0}^{k} x_{ij}\beta_j = \mathbf{x}_i\beta,$$

pri čemu je $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$ $i$-ti redak matrice $X$ u modelu (3.18). Primijetimo da je $x_{i0} = 1$. Dakle, funkcija vjerodostojnosti za jednu opaženu vrijednost je dana s

$$L_i(\beta) = f(y_i \mid \beta) \propto e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\beta)^2}.$$

## Funkcija vjerodostojnosti slučajnog uzorka opažanja

Može se pokazati da su, uz pretpostavke na slučajne greške i uz dane prediktorske varijable, varijable odaziva $Y_1, \ldots, Y_n$ međusobno nezavisne. Zbog toga je vjerodostojnost slučajnog uzorka jednaka umnošku vjerodostojnosti individualnih opservacija i dana je s

$$\begin{aligned} L(\beta) &= f(\mathbf{y} \mid \beta) \\ &= \prod_{i=1}^{n} f(y_i \mid \beta) \\ &\propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\beta)^2}. \end{aligned}$$

Zapišimo funkciju vjerodostojnosti za slučajni uzorak u matričnom obliku kao

$$f(\mathbf{y} \mid \beta) \propto e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)}.$$

Dodamo i oduzmemo $X\mathbf{b}_{LS}$ u izrazu u eksponentu te nakon raspisivanja dobijemo iduće

$$\begin{aligned} (\mathbf{y} - \mathrm{X}\beta)^T(\mathbf{y} - X\beta) =& (\mathbf{y} - X\mathbf{b}_{LS} + X\mathbf{b}_{LS} - X\beta)^T(\mathbf{y} - X\mathbf{b}_{LS} + X\mathbf{b}_{LS} - X\beta) \\ =& (\mathbf{y} - X\mathbf{b}_{LS})^T(\mathbf{y} - X\mathbf{b}_{LS}) + (\mathbf{y} - X\mathbf{b}_{LS})^T(X\mathbf{b}_{LS} - X\beta) \\ &+ (X\mathbf{b}_{LS} - X\beta)^T(\mathbf{y} - X\mathbf{b}_{LS}) + (X\mathbf{b}_{LS} - X\beta)^T(X\mathbf{b}_{LS} - X\beta). \end{aligned}$$

Pogledajmo pobliže drugi pribrojnik u gornjem izrazu:

$$\begin{aligned} (\mathbf{y} - X\mathbf{b}_{LS})^T(X\mathbf{b}_{LS} - X\beta) =& (\mathbf{y} - X(X^TX)^{-1}X^T\mathbf{y})^T(X(\mathbf{b}_{LS} - \beta)) \\ =& \mathbf{y}^T(I_n - X(X^TX)^{-1}X^T)(X(\mathbf{b}_{LS} - \beta)) \\ =& 0. \end{aligned}$$

Slično se pokaže da je i treći pribrojnik jednak 0. Sada za vjerodostojnost slučajnog uzorka vrijedi sljedeće

$$f(\mathbf{y} \mid \beta) \propto e^{-\frac{1}{2\sigma^2}\left[(\mathbf{y}-X\mathbf{b}_{LS})^T(\mathbf{y}-X\mathbf{b}_{LS}) + (\mathrm{X}\mathbf{b}_{LS}-\mathrm{X}\beta)^T(\mathrm{X}\mathbf{b}_{LS}-\mathrm{X}\beta)\right]}.$$

Budući da prvi dio unutar eksponenta u prethodnom izrazu ne sadrži vektor nepoznatih parametara, on se može apsorbirati u konstantu proporcionalnosti te se vjerodostojnost može pojednostaviti na idući oblik:

$$f(\mathbf{y} \mid \beta) \propto e^{-\frac{1}{2\sigma^2}(\mathbf{b_{LS}}-\beta)^{\mathbf{T}}(\mathrm{X}^{\mathbf{T}}\mathrm{X})(\mathbf{b_{LS}}-\beta)}.$$

Sada uočimo da vjerodostojnost ima oblik $MVN(\mathbf{b}_{LS}, V_{LS})$, gdje je $b_{LS} = (X^TX)^{-1}X^T\mathbf{y}$ vektor očekivanja i $V_{LS} = \sigma^2(\mathrm{X}^{\mathbf{T}}\mathrm{X})^{-1}$ kovarijacijska matrica.

## Aposteriorna distribucija

### Korištenje neprekidne multivarijatne apriorne distribucije

Pretpostavimo da koristimo neprekidnu multivarijatnu apriornu distribuciju $g(\beta)$ za vektor nepoznatih parametara $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$. U matričnoj formi, zajednička aposteriorna funkcija gustoće će biti proporcionalna umnošku zajedničke apriorne gustoće i zajedničke funkcije vjerodostojnosti, to jest

$$g(\beta \mid \mathbf{y}) \propto g(\beta) \times f(\mathbf{y} \mid \beta)$$

ili zapisano po komponentama kao

$$g(\beta_0, \ldots, \beta_k \mid y_1, \ldots, y_n) \propto g(\beta_0, \ldots, \beta_k) \times f(y_1, \ldots, y_n \mid \beta_0, \ldots, \beta_k).$$

Kako bi pronašli egzaktnu aposteriornu funkciju gustoće, dijelimo proporcionalnu aposteriornu distribuciju $g(\beta_0, \ldots, \beta_k) \times f(y_1, \ldots, y_n \mid \beta_0, \ldots, \beta_k)$ s njezinim integralom po svim parametarskim vrijednostima. To daje sljedeće

$$g(\beta \mid \mathbf{y}) = \frac{g(\beta_0, \ldots, \beta_k) \times f(y_1, \ldots, y_n \mid \beta_0, \ldots, \beta_k)}{\int \ldots \int g(\beta_0, \ldots, \beta_k) \times f(y_1, \ldots, y_n \mid \beta_0, \ldots, \beta_k) d\beta_0 \ldots d\beta_k}.$$

Za mnoge apriorne distribucije ovaj integral se mora računati korištenjem raznih numeričkih metoda integriranja, što često otežava cijeli proces računanja. U nastavku ćemo se upoznati s dva tipa apriornih distribucija gdje ćemo egzaktnu aposteriornu distribuciju moći izračunati bez potrebe za numeričkom integracijom.

### Korištenje multivarijatne uniformne apriorne distribucije

Ako koristimo multivarijatnu uniformnu apriornu distribuciju, odnosno uzimamo

$$g(\beta_0, \ldots, \beta_k) = 1 \qquad \text{za} \quad -\infty < \beta_j < \infty,\ j = 0, \ldots, k,$$

tada će zajednička aposteriorna gustoća biti proporcionalna zajedničkoj funkciji vjerodostojnosti. Dakle,

$$g(\beta \mid \mathbf{y}) \propto e^{-\frac{1}{2\sigma^2}(\mathbf{b}_{LS} - \beta)^T (X^T X)(\mathbf{b}_{LS} - \beta)}.$$

Prepoznajemo da se radi o $MVN(\mathbf{b}_{LS}, V_{LS})$ distribuciji. Stoga je očekivanje aposteriorne distribucije jednako vektoru procjene za $\beta$ dobivenog metodom najmanjih kvadrata, to jest

$$\mathbf{b}_1 = \mathbf{b}_{LS} = (X^T X)^{-1} X^T \mathbf{y}.$$

Kovarijacijska matrica aposteriorne distribucije jednaka je

$$V_1 = V_{LS} = \sigma^2 (X^T X)^{-1}.$$

**Korištenje multivarijatne normalne apriorne distribucije**

Ustanovili smo da vjerodostojnost ima oblik $MVN(\mathbf{b}_{LS}, V_{LS})$ distribucije. Konjugirana apriorna distribucija će u tom slučaju također biti multivarijatna normalna istih dimenzija. Pokazat ćemo da se, uz odabir $MVN(\mathbf{b}_0, V_0)$ apriorne distribucije za $\beta$, aposteriorna distribucija može izračunati korištenjem jednostavnih pravila za revidiranje parametara multivarijatne normalne distribucije pri čemu neće biti potrebno koristiti metode za numeričko rješavanje integrirala. Prema Bayesovom teoremu zajednička aposteriorna gustoća je proporcionalna umnošku apriorne gustoće i vjerodostojnosti. Slijedom navedenog, raspisujemo izraz za aposteriornu gustoću:

$$\begin{aligned} g(\beta \mid \mathbf{y}) &\propto g(\beta) \times f(\mathbf{y} \mid \beta) \\ &\propto e^{-\frac{1}{2}\left[(\beta-\mathbf{b}_0)^T V_0^{-1}(\beta-\mathbf{b}_0)\right]} \times e^{-\frac{1}{2}\left[(\beta-\mathbf{b}_{LS})^T V_{LS}^{-1}(\beta-\mathbf{b}_{LS})\right]} \\ &\propto e^{-\frac{1}{2}\left[(\beta-\mathbf{b}_0)^T V_0^{-1}(\beta-\mathbf{b}_0)+(\beta-\mathbf{b}_{LS})^T V_{LS}^{-1}(\beta-\mathbf{b}_{LS})\right]} \\ &\propto e^{-\frac{1}{2}\left[\beta^T(V_0^{-1}+V_{LS}^{-1})\beta-\beta^T(V_{LS}^{-1}\mathbf{b}_{LS}+V_0^{-1}\mathbf{b}_0)-(\mathbf{b}_{LS}^T V_{LS}^{-1}+\mathbf{b}_0^T V_0^{-1})\beta+(\mathbf{b}_{LS}^T V_{LS}^{-1}+\mathbf{b}_0^T V_0^{-1})(V_{LS}^{-1}\mathbf{b}_{LS}+V_0^{-1}\mathbf{b}_0)\right]}. \end{aligned}$$

Zadnji pribrojnik u eksponentu ne sadrži $\beta$ pa on neće utjecati na oblik aposteriorne gustoće te može biti apsorbiran u konstantu proporcionalnosti. Ako s $V_1^{-1}$ označimo $V_1^{-1} = V_0^{-1} + V_{LS}^{-1}$, aposteriornu funkciju gustoće možemo zapisati kao

$$g(\beta \mid \mathbf{y}) \propto e^{-\frac{1}{2}\left[\beta^T V_1^{-1}\beta-\beta^T(V_{LS}^{-1}\mathbf{b}_{LS}+V_0^{-1}\mathbf{b}_0)-(\mathbf{b}_{LS}^T V_{LS}^{-1}+\mathbf{b}_0^T V_0^{-1})\beta\right]}.$$

**Definicija 4.2.2.** *Kažemo da je realna kvadratna matrica A ortogonalna ako vrijedi*

$$AA^T = A^T A = I.$$

Neka je $U^T U = V_1^{-1}$ gdje je $U$ ortogonalna matrica. Pretpostavljamo da je $V_1^{-1}$ punog ranga. Tada su obje matrice $U$ i $U^T$ također punog ranga te njihovi inverzi postoje. Nadopunimo do kvadrata razlike izraz u eksponentu dodajući i oduzimajući sljedeći izraz

$$(\mathbf{b_{LS}}^T V_{LS}^{-1} + \mathbf{b}_0^T V_0^{-1})U(U^T)^{-1}(V_{LS}^{-1}\mathbf{b_{LS}} + V_0^{-1}\mathbf{b}_0).$$

Dijelovi koji ne sadrže parametar $\beta$ tretiramo kao konstantu. Stoga za aposteriornu gustoću vrijedi da je proporcionalna donjoj funkciji

$$\propto e^{-\frac{1}{2}\left[\beta^T U^T U\beta-\beta^T U^T(U^T)^{-1}(V_{LS}^{-1}\mathbf{b}_{LS}+V_0^{-1}\mathbf{b}_0)-(\mathbf{b}_{LS}^T V_{LS}^{-1}+\mathbf{b}_0^T V_0^{-1})U^{-1}U\beta+(\mathbf{b}_0^T V_0^{-1}+\mathbf{b}_{LS}V_{LS}^{-1})U^{-1}(U^T)^{-1}(V_0^{-1}\mathbf{b}_0+V_{LS}^{-1}\mathbf{b}_{LS})\right]}.$$

Nakon što faktoriziramo izraz u eksponentu, imamo sljedeće

$$\begin{aligned} g(\beta \mid \mathbf{y}) &\propto e^{-\frac{1}{2}\left[\beta^T U^T-(V_{LS}^{-1}\mathbf{b}_{LS}+V_0^{-1}\mathbf{b}_0)U^{-1}\right]\left[U\beta-(U^T)^{-1}(V_0^{-1}\mathbf{b}_0+V_{LS}^{-1}\mathbf{b_{LS}})\right]} \\ &\propto e^{-\frac{1}{2}\left[\beta-(\mathbf{b}_0^T V_0^{-1}+\mathbf{b}_{LS}^T V_{LS}^{-1})U^{-1}(U^T)^{-1}\right]^T(U^T U)\left[\beta-U^{-1}(U^T)^{-1}(V_0^{-1}\mathbf{b}_0+V_{LS}^{-1}\mathbf{b}_{LS})\right]}, \end{aligned}$$

gdje smo u zadnjem retku izlučili $U^T$ iz prvog faktora unutar eksponenta te $U$ iz drugog faktora. Budući da je $U^T U = V_1^{-1}$ i sve su matrice punog ranga, postoji $(U^T)^{-1}U^{-1} = V_1$. Nakon supstitucije dobijemo

$$g(\beta \mid \mathbf{y}) \propto e^{-\frac{1}{2}(\beta-\mathbf{b}_1)^T V_1^{-1}(\beta-\mathbf{b}_1)},$$

pri čemu je $\mathbf{b}_1 = V_1 V_0^{-1}\mathbf{b}_0 + V_1 V_{LS}^{-1}\mathbf{b_{LS}}$. Vidimo da će aposteriorna distribucija za $\beta$ uz dano $\mathbf{y}$ biti $MVN(\mathbf{b}_1, V_1)$.

Ažurirajuće formule. Kada su pretpostavke modela višestruke linearne regresije zadovoljene, i koristimo $MVN(\mathbf{b}_0, V_0)$ apriornu distribuciju, aposteriorna distribucija će biti $MVN(\mathbf{b}_1, V_1)$. gdje se parametri aposteriorne multivarijatne normalne distribucije pronalaze pomoću ažurirajućih formula. Prema tim formulama, aposteriorna matrica preciznosti jednaka je zbroju apriorne matrice preciznosti i matrice preciznosti funkcije vjerodostojnosti (*matrica preciznosti* se definira kao inverz kovarijacijske matrice), to jest

$$V_1^{-1} = V_0^{-1} + V_{LS}^{-1}.$$

Aposteriorni vektor očekivanja je težinska sredina vektora očekivanja apriorne distribucije i vektora procjena za nepoznate parametre dobivenog metodom najmanjih kvadrata, pri čemu su težine jednake pribrojnicima iz rastava aposteriorne matrice preciznosti. Dakle,

$$\mathbf{b}_1 = V_1 V_0^{-1}\mathbf{b}_0 + V_1 V_{LS}^{-1}\mathbf{b}_{LS}.$$

## Zaključivanje za višestruki normalni model linearne regresije

Donošenje zaključaka o parametrima u višestrukom normalnom modelu linearne regresije uključuje testiranje hipoteza o nepoznatim parametrima i određivanje njihovih Bayesovskih intervala vjerodostojnosti. Kažemo *normalni* model jer pretpostavljamo da su zadovoljene pretpostavke o normalnosti i nezavisnosti slučajnih grešaka. Razmotrit ćemo donošenje zaključaka o jednoj komponenti vektora nepoznatih parametara. Time zapravo pokušavamo odrediti učinak pojedinačne eksplanatorne varijable na varijablu odaziva. Kasnije ćemo razmotriti donošenje zaključaka o cijelom vektoru parametara $\beta$ jer želimo odrediti učinak istovremene promjene u svim vrijednostima eksplanatornih varijabli na vrijednost varijable odaziva.

### Zaključivanje o jednom parametru nagiba

Donošenje zaključka o samo jednom parametru nagiba u modelu višestruke linearne regresije zahtjeva da se ostali nepoznati parametri nagiba, uključujući $\beta_0$, smatraju parametrima smetnje. Zaključke o jednom parametru nagiba donosimo na temelju marginalne

aposteriorne gustoće toga parametra. Aposteriorna distribucija vektora nepoznatih parametara je $MVN(\mathbf{b}_1, V_1)$. Pretpostavimo da je $\beta_j$ parametar od našeg interesa. Marginalna aposteriorna distribucija od $\beta_j$ je $N(m'_{\beta_j}, s_j^2(s'_{\beta_j}))$, gdje je parametar očekivanja $m'_{\beta_j}$ zapravo $j$-ta komponenta vektora očekivanja aposteriorne distribucije $\mathbf{b}_1$, a varijanca $s_j^2(s'_{\beta_j})$ je $j$-ti dijagonalni element kovarijancijske matrice $V_1$ aposteriorne distribucije.

Interval pouzdanosti za jedan parametar nagiba. $(1-\alpha)\cdot 100\%$ Bayesovski interval vjerodostojnosti za parametar nagiba $\beta_j$ je svaki interval koji ima aposteriornu vjerojatnost jednaku $(1-\alpha)$. Kada nam je standardna devijacija poznata, $(1-\alpha)\cdot 100\%$ interval vjerodostojnosti je dan s

$$\left[m'_{\beta_j} - z_{\frac{\alpha}{2}} s_j s'_{\beta_j}, m'_{\beta_j} + z_{\frac{\alpha}{2}} s_j s'_{\beta_j}\right].$$

Ako je prava standardna devijacija nepoznata i koristimo se procjenom izračunatom iz uzorka, tada kritične vrijednosti nalazimo pomoću Studentove $t$-razdiobe s $n-k-1$ stupnjem slobode umjesto standardne normalne razdiobe. Na taj ćemo način dobiti aproksimativni Bayesovski interval vjerodostojnosti za $\beta_j$.

Dvostrane hipoteze za jedan parametar nagiba. Testiramo hipoteze:

$$H_0 : \beta_j = \beta_{j0}$$
$$H_1 : \beta_j \neq \beta_{j0}.$$

Vjerodostojnost nulte hipoteze možemo testirati pomoću intervala vjerodostojnosti. Nazovimo $\beta_{j0}$ nultom vrijednosti za $\beta_j$. Ako vrijednost $\beta_{j0}$ leži izvan $(1-\alpha)\cdot 100\%$ Bayesovskog intervala vjerodostojnosti za $\beta_j$ , tada možemo odbaciti nultu hipotezu na $\alpha$ razini značajnosti. Međutim, ako se nulta vrijednost za $\beta_j$ nalazi unutar intervala, vrijednost $\beta_{j0}$ ostaje vjerodostojna i ne možemo odbaciti nultu hipotezu.

Jednostrane hipoteze za jedan parametar nagiba. Testiramo jednostranu hipotezu

$$H_0 : \beta_j \leq \beta_{j0}$$
$$H_1 : \beta_j > \beta_{j0}$$

za parametar $\beta_j$. Računamo aposteriornu vjerojatnost nulte hipoteze koristeći njenu marginalnu aposteriornu distribuciju. Ako je ta vjerojatnost manja od razine značajnosti $\alpha$, tada odbacujemo nultu hipotezu $H_0$ i zaključujemo da je alternativna hipoteza $H_1$ istinita.

### Zaključivanje o vektoru parametara nagiba

Sada želimo izvesti zaključke o svim parametrima nagiba. U ovom slučaju je jedini parametar smetnje presjek $\beta_0$. Koristit ćemo marginalnu aposteriornu gustoću svih parametara nagiba kako bismo dobili zaključke. Vektor parametara nagiba $\beta' = (\beta_1, \ldots, \beta_k)^T$

je $MVN(\mathbf{b}_{\beta'}, V_{\beta'})$, gdje komponente vektora očekivanja i kovarijacijske matrice dolaze od vektora očekivanja $b_1$ i kovarijacijske matrice $V_1$ aposteriorne distribucije cijelog vektora parametra (uključujući parametar presjeka). Pretpostavit ćemo da je kovarijacijska matrica $V_{\beta'}$ punog ranga. U suprotnom, smanjit ćemo broj parametara nagiba dok ne bude punog ranga.

### Bayesovski interval vjerodostojnosti za vektor parametara nagiba

Komponenta $\beta_j$ vektora parametara nagiba ima $N(m'_{\beta_j}, s^2_j(s'_{\beta_j}))$ distribuciju gdje je parametar očekivanja $m'_{\beta_j}$ zapravo $j$-ta komponenta vektora očekivanja aposteriorne distribucije $\mathbf{b}_1$, a varijanca $s^2_j(s'_{\beta_j})$ je $j$-ti dijagonalni element kovarijancijske matrice $V_1$ aposteriorne distribucije. Ustanovili smo da je $(1-\alpha)\cdot 100\%$ Bayesovski interval vjerodostojnosti za $\beta_j$ zadan granicama $m'_{\beta_j} \pm z_{\frac{\alpha}{2}} s_j s'_{\beta_j}$. Mogli bismo pronaći pojedinačni $(1-\alpha)100\%$ Bayesovski interval vjerodostojnosti za svaku komponentu vektora parametara nagiba, a njihov presjek bi tvorio $k$-dimenzionalno područje vjerodostojnosti za $\beta'$. Kada kažemo da je vektor parametara nagiba $\beta'$ u području vjerodostojnosti, zapravo tvrdimo da su sve komponente vektora $\beta'$ istodobno unutar svojih odgovarajućih intervala vjerodostojnosti. Međutim, ako kombiniramo pojedinačne intervale vjerodostojnosti na ovaj način, izgubit ćemo kontrolu nad ukupnom razinom vjerodostojnosti. Aposteriorna vjerojatnost da su sve komponente vektora $\beta'$ istovremeno sadržane u svojim pojedinačnim intervalima vjerodostojnosti bi u tom slučaju bila mnogo manja od željene razine $(1-\alpha)$. Dakle, moramo pronaći područje vjerodostojnosti u $k$-dimenzionalnom prostoru za sve komponente vektora parametara nagiba čija je vjerojatnost da su sve komponente istovremeno u tome području jednaka $(1-\alpha)$.

Može se pokazati da slučajna varijabla

$$U = (\beta' - \mathbf{b}_{\beta'})V^{-1}_{\beta'}(\beta' - \mathbf{b}_{\beta'})$$

ima $\chi^2$ distribuciju s $k$ stupnjeva slobode te da vrijedi

$$\mathbb{P}(U \leq \chi_\alpha(k)) = 1 - \alpha,$$

gdje je $\chi_\alpha(k)$ $\alpha$ - kvantil $\chi^2$-razdiobe s $k$ stupnjeva slobode. Stoga je $(1-\alpha)100\%$ Bayesovsko područje vjerodostojnosti za vektor parametara nagiba zapravo skup svih točaka $\beta'$ takvih da je

$$(\beta' - \mathbf{b}_{\beta'})V^{-1}_{\beta'}(\beta' - \mathbf{b}_{\beta'}) < \chi_\alpha(k),$$

pri čemu vrijednost $\chi_\alpha(k)$ čitamo iz tablice $\chi^2$ distribucije.[4]

[4]Ovo područje vjerodostojnosti sadrži sve točke koje su "blizu" vektora očekivanja aposteriorne distribucije pri čemu se blizina mjeri aposteriornom distribucijom vektora parametara.

**Točkovne hipoteze za vektor parametara nagiba**

Želimo testirati hipoteze

$$H_0 : \beta' = \beta'_0$$
$$H_1 : \beta' \neq \beta'_0.$$

Prema pretpostavci nulte hipoteze, svaki parametar nagiba $\beta_j$ jednak je svojoj nultoj vrijednosti $\beta_{j0}$ za $j = 1, \ldots, k$. Ako bilo koji od nagiba nije jednak svojoj nultoj vrijednosti, alternativna hipoteza je istinita. Stoga u $k$-dimenzionalnom prostoru postoji samo jedna točka u kojoj je nulta hipoteza istinita. Vjerodostojnost nulte hipoteze možemo testirati pomoću Bayesovskog područja vjerodostojnosti. Ako $\beta'_0$ leži izvan Bayesovskog područja vjerodostojnosti, tada možemo odbaciti nultu hipotezu na $\alpha$ razini značajnosti. S druge strane, ako se nulta vrijednost $\beta'_0$ nalazi unutar Bayesovskog područja vjerodostojnosti, tada ne možemo odbaciti nultu hipotezu jer ona ostaje vjerodostojna na razini $\alpha$.

Najčešće želimo znati jesu li svi parametri nagiba jednaki nula. Ako jesu, ni jedna od eksplanatornih varijabli nije od koristi u modelu. U idućem odjeljku ćemo testirati nultu vrijednost $\beta'_0 = 0$, odnosno testiramo jesu li sve komponente vektora parametara jednake 0, u odnosu na alternativu, gdje je barem jedan od parametara nagiba različit od 0.

## Problemi modeliranja: Uklanjanje nepotrebnih varijabli

Često se prilikom odabira modela višestruke linearne regresije uključe sve moguće eksplanatorne varijable za koje imamo podatke. Neke od ovih eksplanatornih varijable mogu utjecati vrlo malo na odaziv ili uopće ne utječu na njega. Pripadajući parametar nagiba takve varijable bio bi vrlo blizu nule. Takve nepotrebne eksplanatorne varijabli u modelu mogu znatno zakomplicirati određivanje učinaka preostalih eksplanatornih varijabli na odaziv ako postoji korelacija među samim eksplanatornim varijablama u skupu podataka. Uklanjanje ovih nepotrebnih eksplanatornih varijabli rezultirat će boljim modelom za predikcije budućih vrijednosti varijable odaziva.

Željeli bismo ukloniti sve eksplanatorne varijable $x_l$ gdje je stvarni parametar $\beta_l$ jednak 0. Ovo nije tako lako kao što zvuči jer ne znamo koji su parametri nagiba doista jednaki nuli. Imamo nasumični uzorak iz zajedničke aposteriorne distribucije $\beta_1, \ldots, \beta_L$ . Kada su eksplanatorne varijable $x_1, \ldots, x_L$ u korelaciji, neke od tih varijabli mogu ili pojačavati ili prikrivati učinak drugih eksplanatornih varijabli. To znači da vrijednost parametra procijenjena iz aposteriornog uzorka može biti vrlo blizu nule, ali učinak njegove eksplanatorne varijable zapravo može biti veći. Ostale eksplanatorne varijable prikrivaju njegov učinak. Ponekad cijeli skup eksplanatornih varijabli može prikrivati učinak drugih tako da svaka pojedinačna eksplanatorna varijabla izgleda nepotrebno (neznačajno), ali je skup kao cjelina vrlo značajan.

Ne bismo trebali redom testirati hipotezu za svaki parametar nagiba pojedinačno. Pojedinačni test za $H_0 : \beta_l = 0$ u odnosu na $H_0 : \beta_l \neq 0$ se temelji na dodatnom učinku varijable $x_l$ uzimajući u obzir da su druge eksplanatorne varijable već obuhvaćene modelom. Stoga za svaku eksplanatornu varijablu vrijedi da njezin učinak može biti skriven drugim varijablama koje su već zastupljene u modelu.

Umjesto toga, trebali bismo ispitati aposteriornu distribuciju svih parametara nagiba i identificirati sve one s očekivanjem čija je vrijednost blizu 0. Tako ćemo dobiti eksplanatorne varijable koje su kandidati za uklanjanje iz modela. Neka je $x_{j1}, \ldots, x_{jq}$ skup od $q$ eksplanatornih varijabli koje su kandidati za uklanjanje. Neka je $\beta'' = (\beta_{j1}, \ldots, \beta_{jq})^T$ vektor tih nagiba. $\beta''$ ima marginalnu aposteriornu distribuciju $MVN(\mathbf{b}_{\beta''}, V_{\beta''})$, gdje su komponente vektora očekivanja i kovarijacijske matrice dane odgovarajućim komponentama vektora očekivanja $\mathbf{b}_1$ i kovarijacijske matrice $V_1$. $\mathbf{b}_1$ je vektor očekivanja, a $V_1$ kovarijacijska matrica aposteriorne distribucije vektora svih parametara (uključujući parametar presjeka). Određujemo $(1-\alpha)\cdot 100\%$ područje vjerodostojnosti za reducirani vektor parametara nagiba $\beta''$. To će biti područje sastavljeno od svih točaka $\beta''$ takvih da

$$(\beta'' - \mathbf{b}_{\beta''})V_{\beta''}^{-1}(\beta'' - \mathbf{b}_{\beta''}) < \chi_\alpha(k),$$

gdje je $\chi_\alpha(k)$ $\alpha$ - kvantil $\chi^2$-razdiobe s $k$ stupnjeva slobode. Testiramo nultu hipotezu $H_0 : \beta'' = 0$ u odnosu na $H_0 : \beta'' \neq 0$ na razini značajnosti $\alpha$ pomoću Bayesovskog područja vjerodostojnosti. Ako 0 leži unutar Bayesovskog područja vjerodostojnosti, tada ne možemo odbaciti nultu hipotezu i vjerodostojno je da su pripadajući parametri nagiba svih tih eksplanatornih varijabli istovremeno jednaki 0. Ako je to slučaj, uklanjamo varijable $x_{j1}, \ldots, x_{jq}$ iz modela i ponavljamo analizu s preostalim eksplanatornim varijablama.

## Prediktivna distribucija za buduće opservacije

U ovom odjeljku razmatramo Bayesovsko predviđanje vrijednosti odaziva za dane vrijednosti eksplanatornih varijabli pomoću modela višestruke linearne regresije. Kao u slučaju jednostavne linearne regresije, imamo novo opažanje i želimo predvidjeti odaziv, $Y_{n+1}$. Međutim, u ovoj situaciji naše novo opažanje $\mathbf{x}_{n+1}$ je vektor redak duljine $k+1$ čiji je prvi element jednak 1, a $(i+1)$-ti element odgovara novoj vrijednosti $i$-te eksplanatorne varijable.

Kada bi vektor parametara $\beta$ i varijanca $\sigma^2$ slučajnih grešaka bili poznati, i ako bi bile zadovoljene standardne pretpostavke nezavisnosti, normalnosti i jednakosti varijanci slučajnih grešaka, tada bi slučajna varijabla $Y_{n+1}$ imala $N(\mathbf{x}_{n+1}\beta, \sigma^2)$ distribuciju. Međutim, mi ne znamo vrijednosti za $\beta$ i $\sigma^2$. Znamo samo da su njihove aposteriorne distribucije procijenjene iz podataka.

Stoga, kao i kod jednostavne linearne regresije, trebamo pronaći zajedničku gustoću slučajne varijable $Y_{n+1}$ i parametara modela uz dane vrijednosti $\mathbf{x}_{n+1}$ i prethodne podatke,

a zatim integrirati po $\beta$ i varijanci $\sigma^2$ iz dobivenog izraza. Dakle, aposteriorna prediktivna distribucija za $Y_{n+1}$ je zadana sa

$$f(y_{n+1} \mid \mathbf{x}_{n+1}, X, \mathbf{y}) = \int f(y_{n+1} \mid \mathbf{x}_{n+1}, X, \mathbf{y}, \beta, \sigma^2) g(\beta, \sigma^2 \mid \mathbf{x}_{n+1}, X, \mathbf{y}) d\beta d\sigma^2.$$

Odaziv $Y_{n+1}$ je uz dane $\beta_{\bar{x}}$, $\beta_1$ i $x_{n+1}$ još jedna slučajna varijabla iz regresijskog modela te je nezavisna od ostalih varijabli $Y_1, \ldots, Y_n$. Zbog toga, uz dane parametre modela, varijabla odaziva $Y_{n+1}$ ne ovisi o prethodnim podacima. Zajednička aposteriorna funkcija gustoće za $\beta_{\bar{x}}$ i $\beta_1$, koju smo izračunali na temelju podataka, ne ovisi o sljedećoj vrijednosti eksplanatorne varijable $x_{n+1}$.

Prvo analiziramo slučaj kada je $\sigma^2$ poznat. Odaziv $Y_{n+1}$ uz dane $\beta$ i $\mathbf{x}_{n+1}$ je još jedna slučajna varijabla iz regresijskog modela te je nezavisna od ostalih varijabli $Y_1, \ldots, Y_n$, stoga distribucija varijable $Y_{n+1}$ ne ovisi o prethodnim podacima. Dodatno, zajednička aposteriorna funkcija gustoće za vektor parametara $\beta$, koju smo izračunali na temelju podataka, ne ovisi o sljedećim vrijednostima eksplanatornih varijabli $\mathbf{x}_{n+1}$. Zbog toga, distribucija od $\beta$ ne ovisi o $\mathbf{x}_{n+1}$. Slijedom navedenih tvrdnji, za prediktivnu funkciju gustoće varijable $Y_{n+1}$ imamo

$$f(y_{n+1} \mid \mathbf{x}_{n+1}, X, \mathbf{y}) = \int f(y_{n+1} \mid \mathbf{x}_{n+1}, \beta) g(\beta \mid X, \mathbf{y}) d\beta.$$

Ako su $\beta$ i $\sigma^2$ poznati, tada je $Y_{n+1} \sim N(\mathbf{x}_{n+1}\beta, \sigma^2)$ i aposteriorna distribucija od $\beta$ je $MVN(\mathbf{b}_1, V_1)$, dakle

$$\begin{aligned} f(y_{n+1} \mid \mathbf{x}_{n+1}, X, \mathbf{y}) &\propto \int e^{-\frac{1}{2\sigma^2}(y_{n+1} - \mathbf{x}_{n+1}\beta)^2} e^{-\frac{1}{2}(\beta - \mathbf{b}_1)^T V_1^{-1} (\beta - \mathbf{b}_1)} d\beta \\ &= \int e^{-\frac{1}{2\sigma^2}\left[y_{n+1}^2 - 2\mathbf{x}_{n+1}\beta y_{n+1} + (\mathbf{x}_{n+1}\beta)^2\right] - \frac{1}{2}\left[\beta^T V_1^{-1}\beta - 2\beta^T V_1^{-1}\mathbf{b}_1 + \mathbf{b}_1^T V_1^{-1}\mathbf{b}_1\right]} d\beta. \end{aligned}$$

Izraz $\mathbf{b}_1^T V_1^{-1} \mathbf{b}_1$ ne ovisi o $\beta$ pa ga tretiramo kao konstantu. Također je u ovom trenutku prikladno promatrati $\mathbf{x}_{n+1}$ kao vektor stupac i od sada pa nadalje ćemo pisati $\beta^T \mathbf{x}_{n+1}$ umjesto $\mathbf{x}_{n+1}\beta$. Pogledajmo detaljnije izraz u eksponentu. Uz supstituciju $\lambda = \frac{1}{\sigma^2}$, imamo sljedeće

$$\begin{aligned} &-\frac{1}{2\sigma^2}\left[y_{n+1}^2 - 2\beta^T \mathbf{x}_{n+1}\beta y_{n+1} + (\beta^T \mathbf{x}_{n+1})^2\right] - \frac{1}{2}\left[\beta^T V_1^{-1}\beta - 2\beta^T V_1^{-1}\mathbf{b}_1\right] \\ =& -\frac{1}{2}(\lambda y_{n+1}^2 - 2\lambda\beta^T \mathbf{x}_{n+1} y_{n+1} + \lambda\beta^T \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\beta + \beta^T V_1^{-1}\beta - 2\beta^T V_1^{-1}\mathbf{b}_1) \\ =& -\frac{1}{2}(\beta^T(\lambda\mathbf{x}_{n+1}\mathbf{x}_{n+1}^T + V_1^{-1})\beta - 2\beta^T(\lambda y_{n+1}\mathbf{x}_{n+1} + V_1^{-1}\mathbf{b}_1) + \lambda y_{n+1}^2). \end{aligned}$$

Ako stavimo $V = \lambda\mathbf{x}_{n+1}\mathbf{x}_{n+1}^T + V_1^{-1}$ i $m = V^{-1}(\lambda y_{n+1}\mathbf{x}_{n+1} + V_1^{-1}\mathbf{b}_1)$, pod pretpostavkom da je $V$ invertibilna, tada možemo nadopuniti do razlike kvadrata i dobijemo da je izraz u

eksponentu idućeg oblika:

$$(\beta - m)^T V(\beta - m) - m^T Vm + \lambda y_{n+1}^2.$$

Sada za prediktivnu aposteriornu funkciju gustoće vrijedi da je

$$f(y_{n+1} \mid \mathbf{x}_{n+1}, X, \mathbf{y}) \propto \int e^{-\frac{1}{2}(\beta-m)^T V(\beta-m)} \times e^{\frac{1}{2}(m^T Vm - \lambda y_{n+1}^2)} d\beta.$$

Drugi dio u gornjem izrazu ne ovisi o $\beta$ i stoga se može izlučiti van ispred integrala:

$$f(y_{n+1} \mid \mathbf{x}_{n+1}, X, \mathbf{y}) \propto e^{\frac{1}{2}(m^T Vm - \lambda y_{n+1}^2)} \int e^{-\frac{1}{2}(\beta-m)^T V(\beta-m)} d\beta.$$

Integral je proporcionalan gustoći multivarijatne normalne distribucije i zbog toga se integrira u konstantu koja se može apsorbirati u konstantu proporcionalnosti. Preostaje preurediti $e^{\frac{1}{2}(m^T Vm - \lambda y_{n+1}^2)}$ u oblik koji nam je poznat. Ako se još jednom usredotočimo na izraz u eksponentu, uočimo da se $(\lambda y_{n+1}^2 - m^T Vm)$ može zapisati u kvadratnoj formi. Nakon toga ponovno nadopunimo izraz do razlike kvadrata i dobijemo

$$f(y_{n+1} \mid \mathbf{x}_{n+1}, X, \mathbf{y}) \propto e^{-\frac{1}{2(\sigma^2 + \mathbf{x}_{n+1}^T V_1 \mathbf{x}_{n+1})}(y - \mathbf{b}_1^T \mathbf{x}_{n+1})^2}.$$

Zadnja tvrdnja nije trivijalna i zahtjeva korištenje Sherman-Morrison formule.

**Teorem 4.2.3.** *(Sherman-Morrison formula) Neka je $A$ invertibilna kvadratna matrica, i neka su $\mathbf{u}, \mathbf{v}$ vektor stupci. Ako je $1 + \mathbf{v}^T A^{-1}\mathbf{u} \neq 0$, tada vrijedi*

$$(A + \mathbf{u}\mathbf{v}^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}}{1 + \mathbf{v}^T A^{-1}\mathbf{u}}.$$

Znamo da je zadovoljen uvjet za član u nazivniku jer je $V_1$ kovarijacijska matrica, stoga je invertibilna i pozitivno–semidefinitna što osigurava da je kvadratna forma $\mathbf{x}_{n+1}^T V_1 \mathbf{x}_{n+1}$ uvijek veća ili jednaka nuli.

Dakle, aposteriorna prediktivna distribucija proporcionalna normalnoj distribuciji s očekivanjem $\mathbf{b}_1^T \mathbf{x}_{n+1}$ i varijancom $\sigma^2 + \mathbf{x}_{n+1}^T V_1 \mathbf{x}_{n+1}$. Varijanca ima dvije komponente: $\sigma^2$ predstavlja nesigurnost uzorkovanja, a $\mathbf{x}_{n+1}^T V_1 \mathbf{x}_{n+1}$ predstavlja nesigurnost vezanu uz $\beta$. Ako koristimo uniformnu apriornu distribuciju za $\beta$, tada će vektor očekivanja i kovarijacijska matrica aposteriorne distribucije za $\beta$ biti jednaki procjenama dobivenih metodom maksimalne vjerodostojnosti, koje su u ovom slučaju ekvivalentne procjenama dobivenih metodom najmanjih kvadrata. Drugim riječima, ako koristimo uniformnu apriornu distribuciju

za $\beta$, tada je aposteriorna distribucija od $\beta$ multivarijatna normalna s parametrima $\mathbf{b}_1 = \mathbf{b}_{LS}$ i $V_1 = V_{LS}$. Varijanca aposteriorne prediktivne distribucije se tada pojednostavljuje na $\sigma^2(1 + \mathbf{x}_{n+1}^T(X^TX)^{-1}\mathbf{x}_{n+1})$.

Do sada smo pretpostavljali da je $\sigma^2$ poznata, što je nerealno u praksi. U slučaju kada je $\sigma^2$ nepoznat, može se pokazati da aposteriorna prediktivna gustoća ima oblik Studentove t-distribucije s očekivanjem $\beta_1^T\mathbf{x}_{n+1}$, varijancom $s^2+\mathbf{x}_{n+1}^T(X^TX)^{-1}\mathbf{x}_{n+1}$ i $n-k$ stupnjeva slobode gdje je $s^2$ očekivanje kvadrata reziduala, to jest

$$s^2 = \frac{1}{n-k}(y - X\mathbf{b}_{LS})^T(y - X\mathbf{b}_{LS}).$$

Imajte na umu da ovaj rezultat vrijedi točno kada koristimo uniformnu apriornu distribuciju za nepoznate parametre modela ($\beta$ i $\sigma^2$), a približno ako je apriorna distribucija vrlo neinformativna.

**Napomena 4.2.4.** *Ovime smo naveli osnovne rezultate određivanja prediktivne distribucije za buduće opservacije u modelu višestruke linearne regresije. Izvođenje istih zahtijevalo bi dugotrajnu algebarsku manipulaciju i stoga nije prikazano u ovome radu.*

# Poglavlje 5

# Razni primjeri

U ovom poglavlju prikazujemo primjenu obrađene teorije i metoda Bayesovskog statističkog zaključivanja za jednostavnu i višestruku linearnu regresiju na nekoliko odabranih primjera. Zadaci su riješeni u programskom jeziku R, a pripadajući kodovi su priloženi u Dodatku A na kraju rada. Zadaci su preuzeti iz [5] i [8].

**Primjer 5.0.1.** *Istraživač proučava odnos između potrošnje goriva i brzine vožnje. Šest puta je napravio vožnju na testnoj stazi, svaku različitom brzinom, i izmjerio je prijeđene kilometre s jednom litrom goriva. Brzine u kilometrima na sat (x) i udaljenosti u kilometrima (y) zabilježene su u tablici u nastavku.*

| *x (km/h)* | *80* | *90* | *100* | *110* | *120* | *130* |
|---|---|---|---|---|---|---|
| *y (km)* | *55.7* | *55.4* | *52.5* | *52.1* | *50.5* | *49.2* |

*a) Izračunajmo procjene nepoznatih parametara regresijskog pravca.*
*Korištenjem ugrađene funkcije* lm *u R-u dobijemo procjene parametara linearne regresije:*

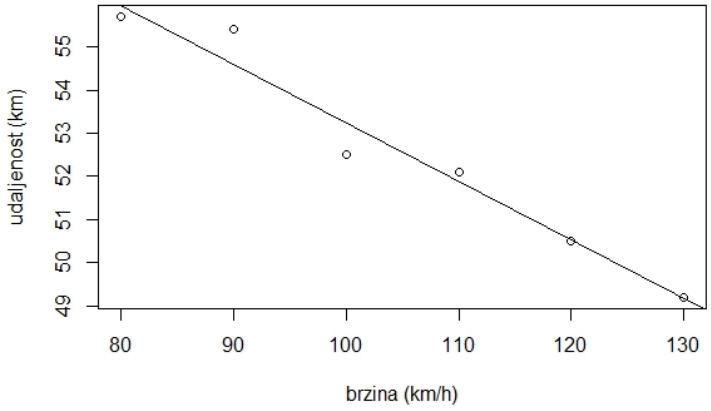
$$\hat{\beta}_1 = -0.136 \quad i \quad \hat{\beta}_0 = 66.847.$$

*Dakle, jednadžba regresijskog pravca je dana sa:* $\hat{y} = 66.847 - 0.136x$.
*Na slici 5.1 nalazi se prikaz dobivene procjene regresijskog pravca na dijagramu rasipanja prijeđene udaljenosti u odnosu na brzinu.*

*b) Odredimo procjenu varijance oko regresijskog pravca za podatke o brzini i prijeđenoj udaljenosti.*

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} [y_i - \hat{y}_i]^2}{n-2} = 0.32633$$

*Procjena standardne devijacije jednaka je* $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 0.57126$.




Slika 5.1: Prikaz krivulje regresijskog pravca na dijagramu rasipanja

*c) Pretpostavimo da prijeđena udaljenost uz danu brzinu vožnje dolazi iz* $N(\beta_0 + \beta_1 x, \sigma^2)$ *razdiobe, gdje je varijanca* $\sigma^2 = 0.57^2$ *poznata. Odredimo aposteriornu distribuciju od* $\beta_1$ *ako za taj parametar koristimo* $N(0, 1^2)$ *apriornu distribuciju.*
*Iskoristimo u tu svrhu* (4.16) *i* (4.17)*. Aposteriorna preciznost od* $\beta_1$ *jednaka je:*

$$\frac{1}{(s'_{\beta_1})^2} = \frac{1}{1^2} + \frac{1750}{0.57^2} = 5387.273$$

*pa je aposteriorna standardna devijacija jednaka* $s'_{\beta_1} = 0.01362$*. Izračunajmo još aposteriorno očekivanje od* $\beta_1$*.*

$$m'_{\beta_1} = \frac{1/1^2}{5387.273} \cdot 0 + \frac{1750/0.57^2}{5387.273} \cdot (-0.136) = -0.13597$$

*Dakle, aposteriorna distribucija od* $\beta_1$ *je* $N(-0.13597, 0.01362^2)$*.*

*d) Pronađimo sada procjenu 95%-tnog Bayesovskog intervala vjerodostojnost za* $\beta_1$*.*
*Bayesovski* $(1 - \alpha) \cdot 100\%$ *interval vjerodostojnosti je oblika*

$$\left[m'_{\beta_1} - z_{\frac{\alpha}{2}} \cdot s'_{\beta_1}, m'_{\beta_1} + z_{\frac{\alpha}{2}} \cdot s'_{\beta_1}\right],$$

*pri čemu su* $m'_{\beta_1}$ *i* $s'_{\beta_1}$ *očekivanje i standardna devijacija aposteriorne distribucije. Vrijednost* $z_{\frac{\alpha}{2}}$ *čitamo iz tablice za standardnu normalnu distribuciju koja se nalazi u Do-*

*datku B ili je računamo u R-u pomoću funkcije* qnorm. *Dobijemo da je procjena 95%-tnog Bayesovskog intervala vjerodostojnosti jednaka*

$$[-0.163, -0.109]\,.$$

*e) Provedimo sada Bayesovski statistički test hipoteza*

$$H_0 : \beta_1 \geq 0$$
$$H_1 : \beta_1 < 0$$

*na razini značajnosti od 5%. Testiramo jednostrane hipoteze, stoga je potrebno izračunati aposteriornu vjerojatnost nulte hipoteze.*

$$\mathbb{P}(\beta_1 \geq 0) = \mathbb{P}\left(Z \geq \frac{0 - (-0.13597)}{0.01362}\right) = \mathbb{P}(Z \geq 9.98028) = 1 - \mathbb{P}(Z \leq 9.98028) = 0.$$

*Vjerojatnost je manja od razine značajnosti testa i odbacujemo $H_0$ te možemo zaključiti da je koeficijent smjera $\beta_1$ zaista manji od* 0.

**Primjer 5.0.2.** *Sljedeći podaci prikazuju procijenjenu dob x (u tjednima) i težinu y (u gramima) za dvanaest beba ženskog spola:*

| $x$ | 40 | 36 | 40 | 38 | 42 | 39 | 40 | 37 | 36 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 3317 | 2729 | 2935 | 2754 | 3210 | 2817 | 3126 | 2539 | 2412 | 2991 | 2875 | 3231 |

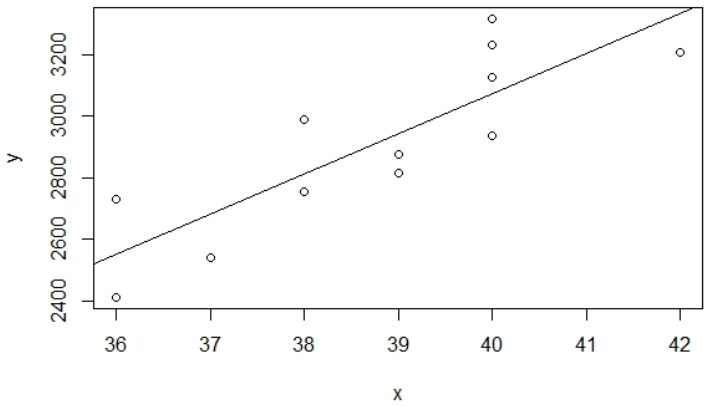
*Izračunajmo procjene nepoznatih parametara regresijskog pravca. Iskoristimo funkciju* lm *u R-u i dobijemo:*

$$\hat{\beta}_0 = -2141.7, \quad \hat{\beta}_1 = 130.4.$$

*Jednadžba regresijskog pravca je dana sa:* $\hat{y} = -2141.7 - 130.4x$.
*Možemo odrediti i alternativni oblik prethodne jednadžbe. Prema* (3.10), *uz* $\hat{\beta}_{\bar{x}} = 2911.33$ *i* $\bar{x} = 38.75$, *vrijedi:* $\hat{y} = 2911.33 - 130.4(x - 38.5)$.
*Procjena varijance oko regresijskog pravca je* $\hat{\sigma} = 157.7105$

*Odredimo procjenu 90%-tnog Bayesovskog intervala vjerodostojnosti za nepoznate parametre i varijancu modela uz pretpostavku da koristimo referentnu apriornu distribuciju. Iz vrijednosti* $S_{xx} = 36.5$, $S_{yy} = 865126.67$ *i* $S_{xy} = 4727$ *dobijemo da je* $S_{ee} = S_{yy} - S_{xy}^2/S_{xx} = 24872.59$. *Budući da je varijanca nepoznata, tražene intervale ćemo odrediti pomoću tvrdnji* (4.12), (4.13) *i* (4.14).
*Vrijednost* $t_{10}(0.95)$ *čitamo iz tablice kvantila Studentove t-distribucije ili računamo u R-u pomoću funkcije* qt. *Procjena 90%-tnog Bayesovskog intervala vjerodostojnosti za parametar* $\beta_{\bar{x}}$ *jednaka je*

$$\left[\bar{y} - t_{10}(0.95) \cdot \hat{\sigma}/\sqrt{n}, \bar{y} + t_{10}(0.95) \cdot \hat{\sigma}/\sqrt{n}\right] = [2828.817, 2993.849]\,.$$




Slika 5.2: Dijagramu rasipanja podataka s krivuljom regresijskog pravca

*Nadalje, procjena 90%-tnog Bayesovskog intervala vjerodostojnosti za parametar* $\beta_1$ *je*

$$\left[\hat{\beta}_1 - t_{10}(0.95) \cdot \hat{\sigma}/\sqrt{S_{xx}}, \hat{\beta}_1 + t_{10}(0.95) \cdot \hat{\sigma}/\sqrt{S_{xx}}\right] = [82.924, 177.876]\,.$$

*Za procjenu 90%-tnog Bayesovskog intervala vjerodostojnosti parametra* $\sigma^2$ *koristimo vrijednosti iz tablice za log* $\chi^2$ *distribuciju koja se nalazi u Dodatku B. Dobijemo:*

$$[S_{ee}/19.447, S_{ee}/4.258] = [12789.935, 58413.778]\,.$$

*Sada imamo da je procjena 90%-tnog Bayesovskog intervala vjerodostojnosti za parametar* $\sigma$ *jednaka*

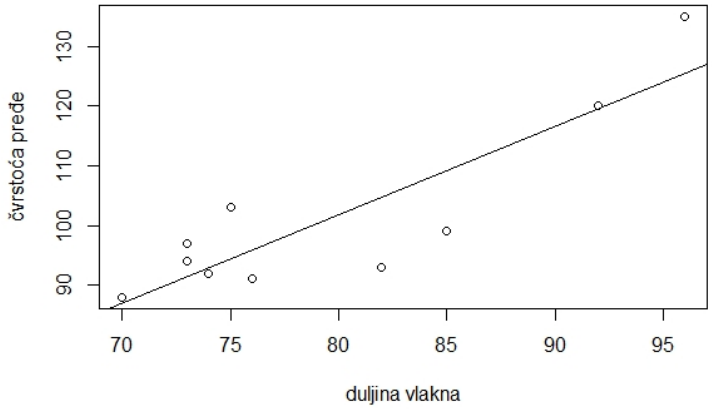
$$[113.093, 241.689].$$

**Primjer 5.0.3.** *Proizvođač tekstila je zabrinut zbog čvrstoće pamučne pređe. Kako bi se utvrdilo je li duljina vlakna važan čimbenik u određivanju čvrstoće pređe, voditelj kontrole kvalitete provjerio je duljinu vlakna (x) i čvrstoću (y) na uzorku od 10 segmenata pređe. Dobiveni su sljedeći rezultati:*

| $x$ | 85 | 82 | 75 | 73 | 76 | 73 | 96 | 92 | 70 | 74 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 99 | 93 | 103 | 97 | 91 | 94 | 135 | 120 | 88 | 92 |

*a) Izračunajmo procjene nepoznatih parametara regresijskog pravca.*
*Korištenjem funkcije* `lm` *dobijemo procjene parametara linearne regresije*

$$\hat{\beta}_1 = 1.478 \quad i \quad \hat{\beta}_0 = -16.409.$$




Slika 5.3: Prikaz regresijskog pravca na dijagramu rasipanja

*Jednadžba regresijskog pravca je dana sa:* $\hat{y} = -16.409 - 1.478x$.

b) *Odredimo procjenu varijance oko regresijskog pravca za podatke o čvrstoći pređe i duljini vlakna. Iskoristimo formulu* (3.12) *i dobijemo sljedeće:*

$$\hat{\sigma}^2 = 58.78057 \quad \Rightarrow \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2} = 7.66685.$$

c) *Pretpostavimo da je čvrstoća pređe uz duljinu vlakna* $N(\beta_0 + \beta_1 x, \sigma^2)$*, gdje je varijanca* $\sigma^2 = 7.7^2$ *poznata. Odredimo aposteriornu distribuciju od* $\beta_1$ *ako za taj parametar koristimo* $N(0, 10^2)$ *apriornu distribuciju.*
*Pomoću izraza* (4.16) *i* (4.17) *dobije se aposteriorna preciznost od* $\beta_1$*:*

$$\frac{1}{(s'_{\beta_1})^2} = 11.85685$$

*Aposteriorna standardna devijacija jednaka je* $s'_{\beta_1} = 0.29041$*, dok je aposteriorno očekivanje od* $\beta_1$ *jednako* $m'_{\beta_1} = 1.47626$*. Dakle, aposteriorna distribucija od* $\beta_1$ *je* $N(1.47626, 0.29041^2)$.

d) *Pronađimo sada procjenu 95%-tnog Bayesovskog intervala vjerodostojnost za* $\beta_1$.
*Iskoristimo* (4.18) *i dobijemo da je procjena 95%-tnog Bayesovskog intervala vjerodostojnosti za parametar nagiba jednaka*

$$[0.907, 2.045]\,.$$

*e) Provedimo Bayesovski statistički test hipoteza*

$$H_0 : \beta_1 \leq 0$$
$$H_1 : \beta_1 > 0$$

*na razini značajnosti od 5%. Testiramo jednostrane hipoteze. Zbog toga je potrebno izračunati aposteriornu vjerojatnost nulte hipoteze.*

$$\mathbb{P}(\beta_1 \leq 0) = \mathbb{P}\left(\frac{\beta_1 - 1.47626}{0.29041} \leq \frac{0 - 1.47626}{0.29041}\right) = \mathbb{P}(Z \leq -5.08336) = 1.85 \cdot 10^{-7}.$$

*Vjerojatnost je manja od razine značajnosti testa i odbacujemo $H_0$ te možemo zaključiti da je koeficijent smjera $\beta_1$ manji od* 0.

*f) Pronađimo prediktivnu distribuciju za $y_{11}$, čvrstoću sljedećeg komada pređe čija je duljina vlakna $x_{11} = 90$.*
*Najprije moramo odrediti aposteriornu distribuciju parametra $\beta_{\bar{x}}$ ako za njega koristimo $N(101.2, 7.83^2)$ (ovu apriornu distribuciju dobili smo koristeći pravilo za određivanje razumne apriorne distribucije na temelju podataka koje je opisano u Poglavlju 4). Aposteriorna preciznost od $\beta_{\bar{x}}$ je*

$$\frac{1}{(s'_{\beta_{\bar{x}}})^2} = 0.185,$$

*stoga je aposteriorna standardna devijacija jednaka $s'_{\beta_{\bar{x}}} = 2.3252$. Nadalje, aposteriorno očekivanje od $\beta_{\bar{x}}$ jednako je $m'_{\beta_{\bar{x}}} = 101.2$ pa je aposteriorna distribucija od $\beta_1$ upravo $N(101.2, 2.3252^2)$. Sada računamo:*

$$m'_y = 101.2 + (90 - 79.6) \cdot 1.47626 = 116.5531;$$

$$(s'_y)^2 = 2.3252^2 + (90 - 79.6)^2 \cdot 0.29041^2 + 7.7^2 = 73.8187.$$

*Prediktivna distribucija za čvrstoću sljedećeg komada pamučne pređe $y_{11}$ čija je duljina vlakna $x_{11} = 90$ je $N(116.5531, 8.5918^2)$.*

*g) Pronađimo 95% Bayesovski interval vjerodostojnosti za predviđanje varijable odaziva iz podzadatka h). Računamo:*

$$\begin{aligned}\left[m'_y - 1.96 \cdot s'_y, m'_y - 1.96 \cdot s'_y\right] &= [116.5531 - 1.96 \cdot 8.5918, 116.5531 - 1.96 \cdot 8.5918] \\ &= [99.714, 133.393]\,.\end{aligned}$$

*Time smo odredili traženi Bayesovski interval vjerodostojnosti.*

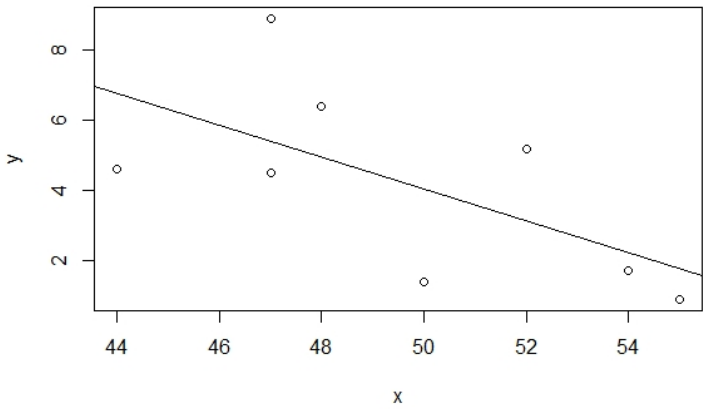
**Primjer 5.0.4.** *Pretpostavimo da sljedećih* 8 *opažanja varijabli* $(x, Y)$ *dolazi iz jednostavnog linearnog regresijskog modela zadanog sa*

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

*gdje je varijanca slučajnih grešaka modela* $\sigma^2 = 2^2$ *poznata.*

| $x$ | 54 | 47 | 44 | 47 | 55 | 50 | 52 | 48 |
|---|---|---|---|---|---|---|---|---|
| $y$ | 1.7 | 4.5 | 4.6 | 8.9 | 0.9 | 1.4 | 5.2 | 6.4 |

*a) Nacrtajmo dijagram rasipanja izmjerenih podataka i izračunajmo procjene nepoznatih parametara regresijskog pravca.*




Slika 5.4: Dijagramu rasipanja podataka s ucrtanom krivuljom regresijskog pravca

*Procjene parametara linearne regresije dobivene funkcijom* lm *su:*

$$\hat{\beta}_0 = 26.7535, \quad \hat{\beta}_1 = -0.4545.$$

*Jednadžba regresijskog pravca je dana sa:* $\hat{y} = 26.7535 - 0.4545x$.

*b) Odredimo aposteriornu distribuciju parametra* $\beta_1$ *ako za* $\beta_1$ *koristimo* $N(0, 3^2)$ *apriornu distribuciju te apriornu distribuciju* $N(4, 2^2)$ *za* $\beta_{\bar{x}}$.
*Pomoću funkcije* Bayes.lin.reg *u R-u pronalazimo aposteriornu distribuciju parametra nagiba:*

$$\beta_1 \sim N(-0.4525, 0.19772^2).$$

*Na slici 5.5 prikazan je odnos apriorne i aposteriorne distribucije parametra nagiba.*

*c) Pronađimo procjenu 95%-tnog Bayesovskog intervala vjerodostojnost za* $\beta_1$.
*Iskoristimo funkciju* qnorm *i parametre aposteriorne distribucije dobivene pomoću* Bayes.lin.reg. *Procjena 95%-tnog Bayesovskog intervala vjerodostojnosti je*

$$[-0.840, -0.065]\,.$$

*d) Provedimo Bayesovski statistički test hipoteza*

$$H_0 : \beta_1 \geq 1$$
$$H_1 : \beta_1 < 1$$

*na razini značajnosti od 5%. Testiramo jednostrane hipoteze, stoga je potrebno izračunati aposteriornu vjerojatnost nulte hipoteze.*

$$\mathbb{P}(\beta_1 \geq 0) = \mathbb{P}\left(Z \geq \frac{0 - (-0.4525)}{0.19772}\right) = \mathbb{P}(Z \geq 2.28861) = 1 - \mathbb{P}(Z \leq 2.28861) = 0.01105.$$

*Vjerojatnost je manja od razine značajnosti testa pa odbacujemo nultu hipotezu. Možemo zaključiti da je koeficijent smjera* $\beta_1$ *manji od* 1.

*e) Pronađimo prediktivnu distribuciju za* $y_9$ *uz danu vrijednost eksplanatorne varijable* $x_9 = 51$.
*Najprije moramo odrediti aposteriornu distribuciju parametra* $\beta_{\bar{x}}$. *Korištenjem funkcije* Bayes.lin.reg *dobijemo da je aposteriorna distribucija od* $\beta_{\bar{x}}$ *jednaka* $N(4.178, 0.66667^2)$.
*Odnos apriorne i aposteriorne distribucije prikazan je na slici 5.6.*

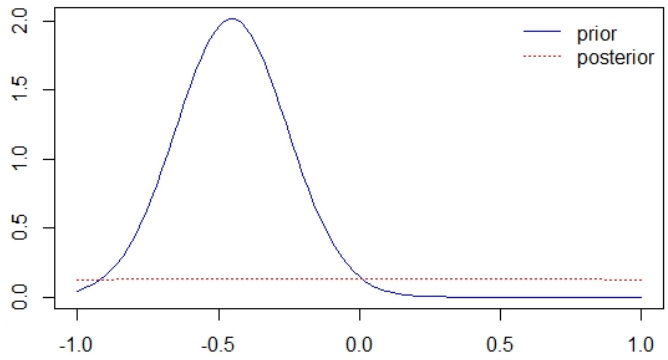
*Sada računamo:*

$$m_y' = 4.178 + (51 - 49.625) \cdot (-0.4525) = 3.5556;$$
$$(s_y')^2 = 0.66667^2 + (51 - 49.625) \cdot 1.978^2 + 2^2 = 2.1209.$$

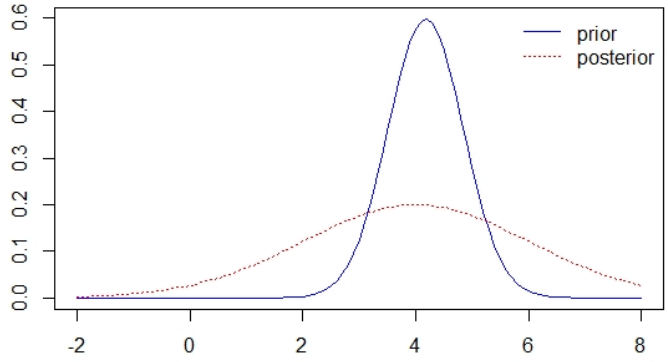
*Prediktivna distribucija za sljedeću vrijednost varijable odaziva* $y_9$ *za dano* $x_9 = 51$ *je* $N(3.5556, 2.1209^2)$.

*f) Pronađimo 95% Bayesovski interval vjerodostojnosti za predviđanje varijable odaziva iz podzadatka h).*
*Rješenje dobijemo ponovno pomoću funkcije* Bayes.lin.reg, *ali uz dodatak argumenta* pred.x *koji označava da tražimo prediktivnu vrijednost varijable odaziva. 95% Bayesovski interval vjerodostojnosti za predviđanje vrijednosti varijable odaziva* $y_9$ *za dano* $x_9 = 51$ *je*

$$[-0.6106, 7.7218].$$




Slika 5.5: Odnos apriorne i aposteriorne distribucije za parametar $\beta_1$



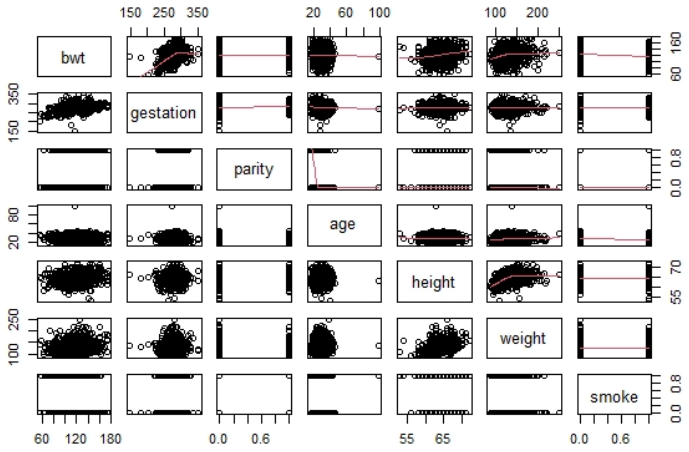

Slika 5.6: Odnos apriorne i aposteriorne distribucije za parametar $\beta_{\bar{x}}$

**Primjer 5.0.5.** *Podaci za ovaj zadatak se mogu preuzeti na* `http://www.stat.berkeley.edu/~statlabs/data/babies.data`. *Varijable u skupu podataka su sljedeće:*

| *Varijabla* | *Opis* |
|---|---|
| bwt | *Težina pri rođenju u uncama (999 = nepoznato)* |
| gestation | *Trajanje trudnoće u danima (999 = nepoznato)* |
| parity | *Red rođenja (0 = prvorođeni, 9 = nepoznato)* |
| age | *Majčina dob u godinama* |
| height | *Majčina visina u inčima(99 = nepoznato)* |
| weight | *Majčina težina prije trudnoće u funtama (999 = nepoznato)* |
| smoke | *Pušački status majke (0 = ne sada, 1 = da sada, 9 = nepoznato)* |

*Ispitivanjem se želi utvrditi utječu li trajanje trudnoće, red rođenja, majčina dob, majčina visina, majčina težina prije trudnoće i pušački status na težinu djeteta prilikom rođenja. Nakon što smo u R-u učitali podatke, važno je osigurati da radimo analizu samo s potpunim podacima. U protivnom bismo trebali imati model i za vrijednosti koje nedostaju. Nakon "čišćenja" podataka, funkcija* nrow *nam pokazuje da je ostalo 1175 (potpunih) slučajeva u podacima.*

*Uvijek je korisno nacrtati podatke prije razmatranja modela. Na taj način ponekad možemo otkriti značajke koje možda ne bismo primijetili u samim podacima te možemo dobiti uvid u moguće probleme. Matrica dijagrama raspršenja dobar je prvi korak kod modela višestruke linearne regresije (koristimo funkciju* pairs).




Slika 5.7: Matrica dijagrama rasipanja podataka

*Primjećujemo da se pojavljuje neuobičajena vrijednost za dob od 99 koju uzimamo kao zadanu vrijednost kada podatak o dobi nedostaje, iako to nije spomenuto u opisu podataka. Zbog toga ćemo ukloniti ovu točku iz naše analize.*

*Sada iskoristimo R funkciju* Bayes.lm *i multivarijatnu normalu apriornu distribuciju kako bismo uklopili model višestruke linearne regresije u ovaj skup podataka. Početni izbor apriorne distribucije je* $MNV(\mathbf{b}_0, V_0)$ *uz* $\mathbf{b}_0 = (0,0,0,0,0,0,0)^T$ *i* $V_0 = 10^6 I_7$*. Dobiveni rezultat za vektor očekivanja aposteriorne distribucije je*

$$\mathbf{b}_1 = \begin{pmatrix} 119.462 & 0.444 & -3.327 & -0.009 & 1.154 & 0.050 & -8.401 \end{pmatrix}^T.$$

*Četvrta i šesta komponenta vektora* $\mathbf{b}_1$ *odnose se na parametre nagiba za eksplanatorne varijable age i weight, redom. Uočavamo da su vrijednosti tih parametara relativno blizu nule. Stoga bismo željeli ispitati ako su varijable age i weight potrebne u modelu, to jest, utječu li one uopće na težinu prilikom rođenja. Testiramo hipoteze:*

$$H_0 : \begin{pmatrix} \beta_{age} \\ \beta_{weight} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad H_1 : \begin{pmatrix} \beta_{age} \\ \beta_{weight} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

*Ako s* $\beta''$ *označimo vektor* $(\beta_{age}, \beta_{weight})^T$*, znamo da* $\beta''$ *ima marginalnu aposteriornu distribuciju* $MVN(\mathbf{b}_{\beta''}, V_{\beta''})$*, gdje su komponente vektora očekivanja i kovarijacijske matrice dane odgovarajućim komponentama vektora očekivanja* $\mathbf{b}_1$ *i kovarijacijske matrice* $V_1$*. Dakle,* $\beta'' \sim MVN(\mathbf{b}_{\beta''}, V_{\beta''})$*, pri čemu su*

$$\mathbf{b}_{\beta''} = \begin{pmatrix} -0.009 \\ 0.050 \end{pmatrix} \quad i \quad V_0 = \begin{pmatrix} 7.365 \cdot 10^{-3} & -2.728 \cdot 10^{-4} \\ -2.728 \cdot 10^{-4} & 6.371 \cdot 10^{-4} \end{pmatrix}.$$

*Uz* $\alpha = 0.05$*, određujemo* 95%*-tno područje vjerodostojnosti za reducirani vektor parametara nagiba* $\beta''$*. To je područje čine sve točke* $\beta''$ *takve da je*

$$(\beta'' - \mathbf{b}_{\beta''}) V_{\beta''}^{-1} (\beta'' - \mathbf{b}_{\beta''}) < \chi_{0.05}(2).$$

*Testiramo hipoteze na razini značajnosti* $\alpha$ *pomoću Bayesovskog područja vjerodostojnosti. Pokazuje se da je za* $\beta'' = (0,0)^T$ *vrijednost* $\mathbf{b}_{\beta''} V_{\beta''}^{-1} \mathbf{b}_{\beta''} = 3.9714$ *što je manje od* $\chi_{0.05}(2) = 5.9915$ *(ovu vrijednost čitamo iz tablice kvantila* $\chi^2$ *razdiobe). Dakle, 0 leži unutar Bayesovskog područja vjerodostojnosti i zbog toga ne možemo odbaciti nultu hipotezu. Vjerodostojno je da su parametri nagiba uz eksplanatorne varijable age i weight jednaki nula i možemo ih ukloniti iz modela.*

# Bibliografija


[1] P. D. Hoff, *A First Course in Bayesian Statistical Methods*, Springer, 2009.

[2] M. Huzak, *Vjerojatnost i matematička statistika*, 2006.

[3] _____, *Matematička statistika, predavanja*, 2020.

[4] _____, *Statistika, predavanja*, Prirodoslovno - matematički fakultet, Matematički odjel, Sveučilište u Zagrebu.

[5] W. M. Bolstad i J. M. Curran, *Introduction to Bayesian Statistics, Second Edition*, A John Wiley & Sons, 2017.

[6] X. Yan i X. G. Su, *Linear regression analysis : Theory and computing*, World Scientific Publishing Co. Pte. Ltd., 2009.

[7] Nikola Sandrić i Zoran Vondraček, *Vjerojatnost, predavanja*, 2019.

[8] P. M. Lee, *Bayesian Statistics An Introduction*, Wiley, 2012.

[9] Nikola Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2002.

# Dodatak A

## R kod za Primjer 5.0.1.

```
#podaci:
x = c(80, 90, 100, 110,120,130)
y = c(55.7, 55.4, 52.5, 52.1, 50.5, 49.2)
n = 6
#a)
plot(x, y, xlab = "brzina (km/h)", ylab = "udaljenost (km)")
regr = lm(y ~ x)
regr
abline(regr$coefficients)
#b)
temp = (y - (mean(y) + regr$coefficients[2]*(x - mean(x))))^2
se = sqrt(sum(temp)/(n - 2))
#c)
sigma = 0.57
Sxx = sum(x^2) - n*mean(x)^2
preciz = (1/1) + (Sxx/sigma^2)
se_beta = sqrt(1/preciz)
m_beta = (1/preciz)*0 + (Sxx/sigma^2)/preciz*regr$coefficients[2]
#d)
z=qnorm(0.975)
i_lower = m_beta - z*se_beta
i_upper = m_beta + z*se_beta
#e)
1 - pnorm(-(m_beta)/se_beta, 0, 1)
```

## R kod za Primjer 5.0.2.

```
#podaci:
x = c(40, 36, 40, 38, 42, 39, 40, 37, 36, 38, 39, 40)
y = c(3317, 2729, 2935, 2754, 3210, 2817, 3126, 2539, 2412, 2991, 2875,
    3231)
n = 12

```

```r
plot(x, y)
regr = lm(y ~ x)
regr
abline(regr$coefficients)
mean(y)
mean(x)

Sxy = sum(x*y) - n*mean(x)*mean(y)
Sxx = sum(x^2) - n*mean(x)^2
Syy = sum(y^2) - n*mean(y)^2
See = Syy - (Sxy^2/Sxx)
se = sqrt(See/(n-2))

t=qt(0.95, 10)
i_lower_x=mean(y) - t*(se/sqrt(n))
i_upper_x= mean(y) + t*(se/sqrt(n))
i_lower_1=regr$coefficients[2] - t*(se/sqrt(Sxx))
i_upper_1= regr$coefficients[2] + t*(se/sqrt(Sxx))

#vrijednosti log-chi^2 uzimamo iz tablice u dodatku
q_1 = 19.447
q_2 = 4.258
i_lower_se_sq = See/q_1
i_upper_se_sq = See/q_2
i_lower_se = sqrt(i_lower_se_sq)
i_upper_se = sqrt(i_upper_se_sq)
```

# R kod za Primjer 5.0.3.

```r
#podaci:
x = c(85, 82, 75, 73, 76, 73, 96, 92, 70, 74)
y = c(99, 93, 103, 97, 91, 94, 135, 120, 88, 92)
n = 10
#a)
regr = lm(y ~ x)
regr
#b)
abline(regr$coefficients)
#c)
temp = (y - (mean(y) + regr$coefficients[2]*(x - mean(x))))^2
se = sqrt(sum(temp)/(n - 2))
#d)
sigma = 7.7
Sxx = sum(x^2) - n*mean(x)^2
preciz = (1/10^2) + (Sxx/sigma^2)
se_beta = sqrt(1/preciz)
m_beta = (1/preciz)*0 + (Sxx/sigma^2)/preciz*regr$coefficients[2]
```

```r
#e)
i_lower = m_beta - 1.96*se_beta
i_upper = m_beta + 1.96*se_beta
#f)
pnorm(-(m_beta)/se_beta, 0, 1)
#g)
m_x = mean(y)
se_x = (max(y) - min(y))/6
preciz_x = (1/se_x^2) + (n/sigma^2)
se_beta_x = sqrt(1/preciz_x)
m_beta_x = ((1/se_x^2)/preciz_x)*m_x + (n/sigma^2)/preciz_x*mean(y)

m_y = m_beta_x + (90 - mean(x))*m_beta
se_y_sq = se_beta_x^2 + (90 - mean(x))^2*se_beta^2 + sigma^2
se_y = sqrt(se_y_sq)
#h)
z = qnorm(0.975)
lower = m_y - z*se_y
upper = m_y + z*se_y
```

# R kod za Primjer 5.0.4.

```r
library("Bolstad")
#podaci
x = c(54, 47, 44, 47, 55, 50, 52, 48)
y = c(1.7, 4.5, 4.6, 8.9, 0.9, 1.4, 5.2, 6.4)
n = 8
sigma = 2
#a)
plot(x, y)
regr = lm(y ~ x)
regr
abline(regr$coefficients)
#b)
rezultati = bayes.lin.reg(y, x, "n", "n", 0, 3, 4, 2, sigma, ret = T)

curve(dnorm(x, rezultati$post.coef[2], rezultati$post.coef.sd[2]),
      from = -1, to = 1, xlab="", ylab="", col = "blue")
curve(dnorm(x, 0, 3), from = -1, to = 1, lty = 3, col = "red",add = T)
legend(x = "topright", legend = c("prior", "posterior"), bty = "n",
       lty = c(1, 3), col = c("blue", "red"))
#c)
inter_vjer = qnorm(c(0.025, 0.975), rezultati$post.coef[2],
                   rezultati$post.coef.sd[2])
#d)
q = (0-rezultati$post.coef[2])/rezultati$post.coef.sd[2]
1-pnorm(q, 0, 1)
#e)
```

```r
curve(dnorm(x, rezultati$post.coef[1], rezultati$post.coef.sd[1]),
      from = -2, to = 8, xlab="",  ylab="", col = "blue")
curve(dnorm(x, 4, 2), from = -2, to = 8, lty = 3, col = "red",add = T)
legend(x = "topright", legend = c("prior", "posterior"), bty = "n",
       lty = c(1, 3), col = c("blue", "red"))

m_y = rezultati$post.coef[1] + (51-mean(x))*rezultati$post.coef[2]
se_y_sq = rezultati$post.coef.sd[1]^2 +
  (51 - mean(x))*rezultati$post.coef.sd[2]^2 + sigma^2
se_y =  sqrt(se_y_sq)
#alternativno rjesenje
rezultati = bayes.lin.reg(y, x, "n", "n", 0, 3, 4, 2, sigma,
                          pred.x = c(51), ret = T)
#f)
z = qnorm(0.975)
lower = rezultati$pred.y - z*rezultati$pred.se
upper = rezultati$pred.y + z*rezultati$pred.se
```

# R kod za Primjer 5.0.5.

```r
library("Bolstad")
url = "http://www.stat.berkeley.edu/~statlabs/data/babies.data"
bw.df = read.table(url, head = TRUE)
bw.df = subset(bw.df, bwt != 999 & gestation != 999
               & parity != 9 & height != 99
               & weight != 999 & smoke != 9)
nrow(bw.df)

pairs(bw.df, upper.panel = panel.smooth)
bw.df = subset(bw.df, age != 99)
b0 = rep(0, 7)
V0 = 10^6*diag(7)

fit = bayes.lm(bwt ~ gestation + parity + age + height + weight + smoke,
               data = bw.df, prior = list(b0 = b0, V0 = V0))
summary(fit)
b1 = fit$post.mean
V1 = fit$post.var

b = b1[c(4, 6)]
V = matrix(c(V1[4, 4], V1[4, 6], V1[6, 4], V1[6, 6 ]), 2, 2)

t(b)%*%solve(V)%*%b
```

# Dodatak B

TABLICA VRIJEDNOSTI FUNKCIJE DISTRIBUCIJE STANDARDNE NORMALNE DISTRIBUCIJE

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7703 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| 2.8 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9979 | 0.9980 | 0.9981 |
| 2.9 | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9984 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |
| 3.1 | 0.9990 | 0.9991 | 0.9991 | 0.9991 | 0.9992 | 0.9992 | 0.9992 | 0.9992 | 0.9993 | 0.9993 |
| 3.2 | 0.9993 | 0.9993 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9995 | 0.9995 | 0.9995 |
| 3.3 | 0.9995 | 0.9995 | 0.9995 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9997 |
| 3.4 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9998 |

TABLICA KVANTILA STUDENTOVE t-RAZDIOBE

| $\alpha$ / $df$ | 0.1 | 0.05 | 0.025 | 0.0125 | 0.01 | 0.005 | 0.0025 | 0.0015 | 0.001 | 0.0005 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.0777 | 6.3138 | 12.7062 | 25.4517 | 31.8205 | 63.6567 | 127.3213 | 212.2050 | 318.3088 | 636.6192 |
| 2 | 1.8856 | 2.9200 | 4.3027 | 6.2053 | 6.9646 | 9.9248 | 14.0890 | 18.2163 | 22.3271 | 31.5991 |
| 3 | 1.6377 | 2.3534 | 3.1824 | 4.1765 | 4.5407 | 5.8409 | 7.4533 | 8.8915 | 10.2145 | 12.9240 |
| 4 | 1.5332 | 2.1318 | 2.7764 | 3.4954 | 3.7469 | 4.6041 | 5.5976 | 6.4348 | 7.1732 | 8.6103 |
| 5 | 1.4759 | 2.0150 | 2.5706 | 3.1634 | 3.3649 | 4.0321 | 4.7733 | 5.3760 | 5.89343 | 6.8688 |
| 6 | 1.4398 | 1.9432 | 2.4469 | 2.9687 | 3.1427 | 3.7074 | 4.3168 | 4.8002 | 5.2076 | 5.9588 |
| 7 | 1.4149 | 1.8946 | 2.3646 | 2.8412 | 2.9980 | 3.4995 | 4.0293 | 4.4421 | 4.7853 | 5.4079 |
| 8 | 1.3968 | 1.8595 | 2.3060 | 2.7515 | 2.8965 | 3.3554 | 3.8325 | 4.1991 | 4.5008 | 5.0413 |
| 9 | 1.3830 | 1.8331 | 2.2622 | 2.6850 | 2.8214 | 3.2498 | 3.6897 | 4.0240 | 4.2968 | 4.7809 |
| 10 | 1.3722 | 1.8125 | 2.2281 | 2.6338 | 2.7638 | 3.1693 | 3.5814 | 3.8920 | 4.1437 | 4.5869 |
| 11 | 1.3634 | 1.7959 | 2.2010 | 2.5931 | 2.7181 | 3.1058 | 3.4966 | 3.7890 | 4.0247 | 4.4370 |
| 12 | 1.3562 | 1.7823 | 2.1788 | 2.5600 | 2.6810 | 3.0545 | 3.4284 | 3.7065 | 3.9296 | 4.3178 |
| 13 | 1.3502 | 1.7709 | 2.1604 | 2.5326 | 2.6503 | 3.0123 | 3.3725 | 3.6389 | 3.8520 | 4.2208 |
| 14 | 1.3450 | 1.7613 | 2.1448 | 2.5096 | 2.6245 | 2.9768 | 3.3257 | 3.5827 | 3.7874 | 4.1405 |
| 15 | 1.3406 | 1.7531 | 2.1314 | 2.4899 | 2.6025 | 2.9467 | 3.2860 | 3.5350 | 3.7328 | 4.0728 |
| 16 | 1.3368 | 1.7459 | 2.1199 | 2.4729 | 2.5835 | 2.9208 | 3.2520 | 3.4942 | 3.6862 | 4.0150 |
| 17 | 1.3334 | 1.7396 | 2.1098 | 2.4581 | 2.5669 | 2.8982 | 3.2224 | 3.4589 | 3.6458 | 3.9651 |
| 18 | 1.3304 | 1.7341 | 2.1009 | 2.4450 | 2.5524 | 2.8784 | 3.1966 | 3.4279 | 3.6105 | 3.9216 |
| 19 | 1.3277 | 1.7291 | 2.0930 | 2.4334 | 2.5395 | 2.8609 | 3.1737 | 3.4007 | 3.5794 | 3.8834 |
| 20 | 1.3253 | 1.7247 | 2.0860 | 2.4231 | 2.5280 | 2.8453 | 3.1534 | 3.3764 | 3.5518 | 3.8495 |
| 21 | 1.3232 | 1.7207 | 2.0796 | 2.4138 | 2.5176 | 2.8314 | 3.1352 | 3.3548 | 3.5272 | 3.8193 |
| 22 | 1.3212 | 1.7171 | 2.0739 | 2.4055 | 2.5083 | 2.8188 | 3.1188 | 3.3353 | 3.5050 | 3.7921 |
| 23 | 1.3195 | 1.7139 | 2.0687 | 2.3979 | 2.4999 | 2.8073 | 3.1040 | 3.3176 | 3.4850 | 3.7676 |
| 24 | 1.3178 | 1.7109 | 2.0639 | 2.3909 | 2.4922 | 2.7969 | 3.0905 | 3.3016 | 3.4668 | 3.7454 |
| 25 | 1.3163 | 1.7081 | 2.0595 | 2.3846 | 2.4851 | 2.7874 | 3.0782 | 3.2870 | 3.4502 | 3.7251 |
| 26 | 1.3150 | 1.7056 | 2.0555 | 2.3788 | 2.4786 | 2.7787 | 3.0669 | 3.2736 | 3.4350 | 3.7066 |
| 27 | 1.3137 | 1.7033 | 2.0518 | 2.3734 | 2.4727 | 2.7707 | 3.0565 | 3.2613 | 3.4210 | 3.6896 |
| 28 | 1.3125 | 1.7011 | 2.0484 | 2.3685 | 2.4671 | 2.7633 | 3.0469 | 3.2499 | 3.4082 | 3.6739 |
| 29 | 1.3114 | 1.6991 | 2.0452 | 2.3638 | 2.4620 | 2.7564 | 3.0380 | 3.2394 | 3.3962 | 3.6594 |
| 30 | 1.3104 | 1.6973 | 2.0423 | 2.3596 | 2.4573 | 2.7500 | 3.0298 | 3.2296 | 3.3852 | 3.6460 |
| $\infty$ | 1.2816 | 1.6449 | 1.9510 | 2.2414 | 2.3264 | 2.5758 | 2.8070 | 2.9677 | 3.0902 | 3.2905 |

# TABLICA KVANTILA $\chi^2$-RAZDIOBE

| $\alpha$ / $df$ | 0.99 | 0.98 | 0.975 | 0.95 | 0.90 | 0.80 | 0.70 | 0.50 | 0.30 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0002 | 0.0006 | 0.0010 | 0.0039 | 0.0158 | 0.0642 | 0.1485 | 0.4549 | 1.0742 |
| 2 | 0.0201 | 0.0404 | 0.0506 | 0.1026 | 0.2107 | 0.4463 | 0.7133 | 1.3863 | 2.4079 |
| 3 | 0.1148 | 0.1848 | 0.2158 | 0.3518 | 0.5844 | 1.0052 | 1.4237 | 2.3660 | 3.6649 |
| 4 | 0.2971 | 0.4294 | 0.4844 | 0.7107 | 1.0636 | 1.6488 | 2.1947 | 3.3567 | 4.8784 |
| 5 | 0.5543 | 0.7519 | 0.8312 | 1.1455 | 1.6103 | 2.3425 | 2.9999 | 4.3515 | 6.0644 |
| 6 | 0.8721 | 1.1344 | 1.2373 | 1.6354 | 2.2041 | 3.0701 | 3.8276 | 5.3481 | 7.2311 |
| 7 | 1.2390 | 1.5643 | 1.6899 | 2.1673 | 2.8331 | 3.8223 | 4.6713 | 6.3458 | 8.3834 |
| 8 | 1.6465 | 2.0325 | 2.1797 | 2.7326 | 3.4895 | 4.5936 | 5.5274 | 7.3441 | 9.5245 |
| 9 | 2.0879 | 2.5324 | 2.7004 | 3.3251 | 4.1682 | 5.3801 | 6.3933 | 8.3428 | 10.6564 |
| 10 | 2.5582 | 3.0591 | 3.2470 | 3.9403 | 4.8652 | 6.1791 | 7.2672 | 9.3418 | 11.7807 |
| 11 | 3.0535 | 3.6087 | 3.8157 | 4.5748 | 5.5778 | 6.9887 | 8.1479 | 10.3410 | 12.8987 |
| 12 | 3.5706 | 4.1783 | 4.4038 | 5.2260 | 6.3038 | 7.8073 | 9.0343 | 11.3403 | 14.0111 |
| 13 | 4.1069 | 4.7654 | 5.0088 | 5.8919 | 7.0415 | 8.6339 | 9.9257 | 12.3398 | 15.1187 |
| 14 | 4.6604 | 5.3682 | 5.6287 | 6.5706 | 7.7895 | 9.4673 | 10.8215 | 13.3393 | 16.2221 |
| 15 | 5.2293 | 5.9849 | 6.2621 | 7.2609 | 8.5468 | 10.3070 | 11.7212 | 14.3389 | 17.3217 |
| 16 | 5.8122 | 6.6142 | 6.9077 | 7.9616 | 9.3122 | 11.1521 | 12.6243 | 15.3385 | 18.4179 |
| 17 | 6.4078 | 7.2550 | 7.5642 | 8.6718 | 10.0852 | 12.0023 | 13.5307 | 16.3382 | 19.5110 |
| 18 | 7.0149 | 7.9062 | 8.2307 | 9.3905 | 10.8649 | 12.8570 | 14.4399 | 17.3379 | 20.6014 |
| 19 | 7.6327 | 8.5670 | 8.9065 | 10.1170 | 11.6509 | 13.7158 | 15.3517 | 18.3377 | 21.6891 |
| 20 | 8.2604 | 9.2367 | 9.5908 | 10.8508 | 12.4426 | 14.5784 | 16.2659 | 19.3374 | 22.7745 |
| 21 | 8.8972 | 9.9146 | 10.2829 | 11.5913 | 13.2396 | 15.4446 | 17.1823 | 20.3372 | 23.8578 |
| 22 | 9.5425 | 10.6000 | 10.9823 | 12.3380 | 14.0415 | 16.3140 | 18.1007 | 21.3370 | 24.9390 |
| 23 | 10.1957 | 11.2926 | 11.6886 | 13.0905 | 14.8480 | 17.1865 | 19.0211 | 22.3369 | 26.0184 |
| 24 | 10.8564 | 11.9918 | 12.4012 | 13.8484 | 15.6587 | 18.0618 | 19.9432 | 23.3367 | 27.0960 |
| 25 | 11.5240 | 12.6973 | 13.1197 | 14.6114 | 16.4734 | 18.9398 | 20.8670 | 24.3366 | 28.1719 |
| 26 | 12.1981 | 13.4086 | 13.8439 | 15.3792 | 17.2919 | 19.8202 | 21.7924 | 25.3365 | 29.2463 |
| 27 | 12.8785 | 14.1254 | 14.5734 | 16.1514 | 18.1139 | 20.7030 | 22.7192 | 26.3363 | 30.3193 |
| 28 | 13.5647 | 14.8475 | 15.3079 | 16.9279 | 18.9392 | 21.5880 | 23.6475 | 27.3362 | 31.3909 |
| 29 | 14.2565 | 15.5745 | 16.0471 | 17.7084 | 19.7677 | 22.4751 | 24.5770 | 28.3361 | 32.4612 |
| 30 | 14.9535 | 16.3062 | 16.7908 | 18.4927 | 20.5992 | 23.3641 | 25.5078 | 29.3360 | 33.5302 |

# TABLICA KVANTILA $\chi^2$-RAZDIOBE

| $\alpha$ / $df$ | 0.20 | 0.10 | 0.05 | 0.025 | 0.02 | 0.01 | 0.005 | 0.002 | 0.001 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.6424 | 2.7055 | 3.8415 | 5.0239 | 5.4119 | 6.6349 | 7.8794 | 9.5495 | 10.8276 |
| 2 | 3.2189 | 4.6052 | 5.9915 | 7.3778 | 7.8240 | 9.2103 | 10.5966 | 12.4292 | 13.8155 |
| 3 | 4.6416 | 6.2514 | 7.8147 | 9.3484 | 9.8374 | 11.3449 | 12.8382 | 14.7955 | 16.2662 |
| 4 | 5.9886 | 7.7794 | 9.4877 | 11.1433 | 11.6678 | 13.2767 | 14.8603 | 16.9238 | 18.4668 |
| 5 | 7.2893 | 9.2364 | 11.0705 | 12.8325 | 13.3882 | 15.0863 | 16.7496 | 18.9074 | 20.5150 |
| 6 | 8.5581 | 10.6446 | 12.5916 | 14.4494 | 15.0332 | 16.8119 | 18.5476 | 20.7912 | 22.4577 |
| 7 | 9.8032 | 12.0170 | 14.0671 | 16.0128 | 16.6224 | 18.4753 | 20.2777 | 22.6007 | 24.3219 |
| 8 | 11.0301 | 13.3616 | 15.5073 | 17.5345 | 18.1682 | 20.0902 | 21.9550 | 24.3521 | 26.1245 |
| 9 | 12.2421 | 14.6837 | 16.9190 | 19.0228 | 19.6790 | 21.6660 | 23.5894 | 26.0564 | 27.8772 |
| 10 | 13.4420 | 15.9872 | 18.3070 | 20.4832 | 21.1608 | 23.2093 | 25.1882 | 27.7216 | 29.5883 |
| 11 | 14.6314 | 17.2750 | 19.6751 | 21.9200 | 22.6179 | 24.7250 | 26.7568 | 29.3536 | 31.2641 |
| 12 | 15.8120 | 18.5493 | 21.0261 | 23.3367 | 24.0540 | 26.2170 | 28.2995 | 30.9570 | 32.9095 |
| 13 | 16.9848 | 19.8119 | 22.3620 | 24.7356 | 25.4715 | 27.6882 | 29.8195 | 32.5352 | 34.5282 |
| 14 | 18.1508 | 21.0641 | 23.6848 | 26.1189 | 26.8728 | 29.1412 | 31.3193 | 34.0913 | 36.1233 |
| 15 | 19.3107 | 22.3071 | 24.9958 | 27.4884 | 28.2595 | 30.5779 | 32.8013 | 35.6276 | 37.6973 |
| 16 | 20.4651 | 23.5418 | 26.2962 | 28.8454 | 29.6332 | 31.9999 | 34.2672 | 37.1461 | 39.2524 |
| 17 | 21.6146 | 24.7690 | 27.5871 | 30.1910 | 30.9950 | 33.4087 | 35.7185 | 38.6485 | 40.7902 |
| 18 | 22.7595 | 25.9894 | 28.8693 | 31.5264 | 32.3462 | 34.8053 | 37.1565 | 40.1361 | 42.3124 |
| 19 | 23.9004 | 27.2036 | 30.1435 | 32.8523 | 33.6874 | 36.1909 | 38.5823 | 41.6103 | 43.8202 |
| 20 | 25.0375 | 28.4120 | 31.4104 | 34.1696 | 35.0196 | 37.5662 | 39.9968 | 43.0720 | 45.3147 |
| 21 | 26.1711 | 29.6151 | 32.6706 | 35.4789 | 36.3434 | 38.9322 | 41.4011 | 44.5222 | 46.7970 |
| 22 | 27.3015 | 30.8133 | 33.9244 | 36.7807 | 37.6595 | 40.2894 | 42.7957 | 45.9618 | 48.2679 |
| 23 | 28.4288 | 32.0069 | 35.1725 | 38.0756 | 38.9683 | 41.6384 | 44.1813 | 47.3915 | 49.7282 |
| 24 | 29.5533 | 33.1962 | 36.4150 | 39.3641 | 40.2704 | 42.9798 | 45.5585 | 48.8118 | 51.1786 |
| 25 | 30.6752 | 34.3816 | 37.6525 | 40.6465 | 41.5661 | 44.3141 | 46.9279 | 50.2234 | 52.6197 |
| 26 | 31.7946 | 35.5632 | 38.8851 | 41.9232 | 42.8558 | 45.6417 | 48.2899 | 51.6269 | 54.0520 |
| 27 | 32.9117 | 36.7412 | 40.1133 | 43.1945 | 44.1400 | 46.9629 | 49.6449 | 53.0226 | 55.4760 |
| 28 | 34.0266 | 37.9159 | 41.3371 | 44.4608 | 45.4188 | 48.2782 | 50.9934 | 54.4110 | 56.8923 |
| 29 | 35.1394 | 39.0875 | 42.5570 | 45.7223 | 46.6927 | 49.5879 | 52.3356 | 55.7925 | 58.3012 |
| 30 | 36.2502 | 40.2560 | 43.7730 | 46.9792 | 47.9618 | 50.8922 | 53.6720 | 57.1674 | 59.7031 |

GRANICE BayesOVSKIH INTERVALA VJERODOSTOJNOSTI ZA log $\chi^2$-RAZDIOBU

| $\alpha$ | 0.80 | | 0.90 | | 0.95 | | 0.99 | | 0.995 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $df$ | | | | | | | | | | |
| 3 | 0.779 | 7.622 | 0.476 | 9.434 | 0.296 | 11.191 | 0.101 | 15.128 | 0.064 | 16.771 |
| 4 | 1.308 | 9.042 | 0.883 | 10.958 | 0.607 | 12.802 | 0.264 | 16.903 | 0.186 | 18.612 |
| 5 | 1.891 | 10.427 | 1.355 | 12.442 | 0.989 | 14.369 | 0.496 | 18.619 | 0.372 | 20.390 |
| 6 | 2.513 | 11.784 | 1.875 | 13.892 | 1.425 | 15.897 | 0.786 | 20.295 | 0.614 | 22.116 |
| 7 | 3.165 | 13.117 | 2.431 | 15.314 | 1.903 | 17.393 | 1.122 | 21.931 | 0.904 | 23.802 |
| 8 | 3.841 | 14.430 | 3.017 | 16.711 | 2.414 | 18.860 | 1.498 | 23.532 | 1.233 | 25.450 |
| 9 | 4.535 | 15.727 | 3.628 | 18.087 | 2.953 | 20.305 | 1.906 | 25.108 | 1.596 | 27.073 |
| 10 | 5.246 | 17.009 | 4.258 | 19.447 | 3.516 | 21.729 | 2.344 | 26.654 | 1.991 | 28.659 |
| 11 | 5.970 | 18.279 | 4.906 | 20.789 | 4.099 | 23.135 | 2.807 | 28.176 | 2.410 | 30.231 |
| 12 | 6.707 | 19.537 | 5.570 | 22.119 | 4.700 | 24.525 | 3.291 | 29.685 | 2.853 | 31.777 |
| 13 | 7.454 | 20.786 | 6.246 | 23.437 | 5.317 | 25.901 | 3.795 | 31.171 | 3.317 | 33.305 |
| 14 | 8.210 | 22.026 | 6.935 | 24.743 | 5.948 | 27.263 | 4.315 | 32.644 | 3.797 | 34.821 |
| 15 | 8.975 | 23.258 | 7.634 | 26.039 | 6.591 | 28.614 | 4.853 | 34.099 | 4.296 | 36.315 |
| 16 | 9.747 | 24.483 | 8.343 | 27.325 | 7.245 | 29.955 | 5.404 | 35.539 | 4.811 | 37.788 |
| 17 | 10.527 | 25.701 | 9.060 | 28.604 | 7.910 | 31.285 | 5.968 | 36.972 | 5.339 | 39.253 |
| 18 | 11.312 | 26.913 | 9.786 | 29.876 | 8.584 | 32.608 | 6.545 | 38.388 | 5.879 | 40.711 |
| 19 | 12.104 | 28.120 | 10.519 | 31.140 | 9.267 | 33.921 | 7.132 | 39.796 | 6.430 | 42.160 |
| 20 | 12.900 | 29.322 | 11.259 | 32.398 | 9.958 | 35.227 | 7.730 | 41.194 | 6.995 | 43.585 |
| 21 | 13.702 | 30.519 | 12.005 | 33.649 | 10.656 | 36.526 | 8.337 | 42.583 | 7.566 | 45.016 |
| 22 | 14.508 | 31.711 | 12.756 | 34.896 | 11.362 | 37.817 | 8.951 | 43.969 | 8.152 | 46.421 |
| 23 | 15.319 | 32.899 | 13.514 | 36.136 | 12.073 | 39.103 | 9.574 | 45.344 | 8.742 | 47.832 |
| 24 | 16.134 | 34.083 | 14.277 | 37.372 | 12.791 | 40.384 | 10.206 | 46.708 | 9.341 | 49.232 |
| 25 | 16.952 | 35.264 | 15.044 | 38.603 | 13.515 | 41.657 | 10.847 | 48.062 | 9.949 | 50.621 |
| 26 | 17.774 | 36.441 | 15.815 | 39.830 | 14.243 | 42.927 | 11.491 | 49.419 | 10.566 | 52.000 |
| 27 | 18.599 | 37.615 | 16.591 | 41.052 | 14.977 | 44.191 | 12.143 | 50.764 | 11.186 | 53.381 |
| 28 | 19.427 | 38.786 | 17.372 | 42.271 | 15.715 | 45.452 | 12.804 | 52.099 | 11.815 | 54.752 |
| 29 | 20.259 | 39.953 | 18.156 | 43.486 | 16.459 | 46.706 | 13.467 | 53.436 | 12.451 | 56.111 |
| 30 | 21.093 | 41.119 | 18.944 | 44.696 | 17.206 | 47.958 | 14.138 | 54.761 | 13.091 | 57.473 |
| 35 | 25.303 | 46.906 | 22.931 | 50.705 | 21.002 | 54.156 | 17.563 | 61.330 | 16.384 | 64.165 |
| 40 | 29.566 | 52.640 | 26.987 | 56.645 | 24.879 | 60.275 | 21.094 | 67.792 | 19.782 | 70.766 |
| 45 | 33.874 | 58.330 | 31.100 | 62.530 | 28.823 | 66.326 | 24.711 | 74.172 | 23.277 | 77.269 |
| 50 | 38.220 | 63.983 | 35.260 | 68.366 | 32.824 | 72.324 | 28.401 | 80.480 | 26.857 | 83.681 |
| 55 | 42.597 | 69.605 | 39.461 | 74.164 | 36.873 | 78.272 | 32.158 | 86.717 | 30.499 | 90.039 |
| 60 | 47.001 | 75.200 | 43.698 | 79.926 | 40.965 | 84.178 | 35.966 | 92.908 | 34.207 | 96.324 |

# Sažetak

U ovom radu opisuju se metode Bayesovskog zaključivanja za linearnu regresiju. U prvom poglavlju uvodimo osnovne definicije i rezultate iz teorije vjerojatnosti i matematičke statistike koji su bitni za razumijevanje Bayesovog teorema i temeljnih koncepata iz Bayesovske statistike.

U drugom poglavlju donosimo iskaze diskretne i neprekidne verzije Bayesovog teorema, zajedno s dokazima i primjerima. Navedeni teoremi čine temeljnu okosnicu za Bayesovsko statističko zaključivanje. Uvode se ključni pojmovi u Bayesovskoj statistici: apriorna i aposteriorna distribucija.

U trećem poglavlju uvodimo model jednostavne i višestruke linearne regresije te korištenjem metode najmanjih kvadrata određujemo procjene parametara za modele. Također, navodimo i osnovne pretpostavke o vjerojatnosnoj distribuciji slučajnih grešaka u modelu bez kojih ne bismo mogli donositi validne zaključke o nepoznatim parametrima modela linearne regresije.

U četvrtome poglavlju opisujemo Bayesovsko zaključivanje za linearnu regresiju, ili drugim riječima, opisujemo statistički pristup koji kombinira regresijsku analizu s principima Bayesovskog zaključivanja. U ovom pristupu pretpostavljamo da su parametri modela linearne regresije slučajne varijable s apriornim distribucijama vjerojatnosti. Posebno obrađujemo temu jednostavne linearne regresije, a posebno temu višestruke linearne regresije. Opisani su načini biranja apriornih distribucija, a za određene slučajeve prikazano je kako koristimo Bayesov teorem za korigiranje naših uvjerenja o parametrima modela uzimajući u obzir dostupne podatke. Tako dolazimo do aposteriorne distribucije na temelju koje izvodimo zaključke o parametrima modela. Zatim uvodimo pojam Bayesovskog intervala vjerodostojnosti i donosimo njegovu interpretaciju u usporedbi s frekvencionističkim intervalom pouzdanosti. Opisujemo Bayesovske metode intervalne procjene i razmatramo postupak testiranja hipoteza. Na kraju ovog poglavlja koristimo Bayesov teorem za predviđanje vrijednosti odaziva za nove vrijednosti eksplanatornih varijabli. Prikazujemo postupak pronalaženja prediktivne distribucije i pomoću nje određujemo Bayesovske intervale vjerodostojnosti za predikcije.

U posljednjem, petom poglavlju, predstavljamo primjenu obrađenih metoda Bayesovskog zaključivanja kroz nekoliko primjera. Zadaci su riješeni u programskom jeziku R.

# Summary

In this paper, we describe Bayesian inference methods for linear regression. In the first chapter, we introduce basic definitions and results from probability theory and mathematical statistics that are essential for understanding Bayes' theorem and fundamental concepts of Bayesian statistics.

In the second chapter, the discrete and continuous versions of Bayes' theorem are presented, along with proofs and examples. These theorems form the basic framework for Bayesian statistical inference. Key concepts in Bayesian statistics are introduced: prior and posterior distribution.

In the third chapter, we introduce the model of simple and multiple linear regression, and, using the least squares method, we determine estimates of the parameters for the models. We also state basic assumptions about the probability distribution of random errors in the regression model, without which we would not be able to make valid conclusions about the unknown parameters of the linear regression model.

In the fourth chapter, we describe Bayesian inference for linear regression, or in other words, we describe a statistical approach that combines regression analysis with the principles of Bayesian inference. In this approach, we assume that the parameters of the linear regression model are random variables with prior probability distributions. We specifically discuss the topic of simple linear regression and multiple linear regression. Methods of choosing prior distributions are described, and for certain cases, we show how to use Bayes' theorem to update our beliefs about the parameters of the model based on available data. This leads to posterior distributions from which we draw conclusions about the parameters of the model. Next, we introduce the concept of the Bayesian credible interval and provide its interpretation in comparison to the frequentist confidence interval. We describe Bayesian methods of interval estimation and discuss hypothesis testing procedures. Finally, in this chapter, we use Bayes' theorem for predicting response values for new values of explanatory variables. We present the procedure for finding the predictive distribution and using it to determine the Bayesian credible intervals for predictions.

In the last chapter, we present the application of the discussed methods of Bayesian inference through several examples. The examples are solved in the programming language R.

# Životopis

Rođena sam 24. rujna 1995. godine u Zaboku. Odrasla sam u Pregradi gdje sam pohađala Osnovnu školu Janka Leskovara. Nakon završetka osnovne škole svoje obrazovanje nastavljam u Srednjoj školi Pregrada, smjer opća gimnazija. Godine 2014. selim se u Zagreb i upisujem preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. Potom 2020. godine na istom fakultetu upisujem diplomski sveučilišni studij Financijska i poslovna matematika.