Višeparametarska linearna regresija i smrtnost od tumora

Varkaš, Matea

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Matea Varkaš

VIŠEPARAMETARSKA LINEARNA REGRESIJA I SMRTNOST OD TUMORA

Diplomski rad

Voditelj rada:
prof. dr. sc. Siniša Slijepčević

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. _____

Ovaj rad posvećujem svojim roditeljima čija je neizmjerna podrška i ljubav omogućila ostvarenje ovog putovanja. Hvala Goranu, mojem vjernom suputniku kroz sve uspone i padove te dragim prijateljima za sve trenutke koje smo proveli zajedno.

Sadržaj

Uvod

U stvarnom životu, pri donošenju odluka, često se susrećemo s pitanjem koji svi faktori utječu na konačni rezultat i na koji način. Regresijska analiza predstavlja jednu od najčešće korištenih statističkih metoda za modeliranje odnosa između varijabli odziva i varijabli poticaja. Njezin cilj je odrediti snagu i karakter odnosa između jedne ili više zavisnih varijabli (varijabli odziva, označenih s Y) te jedne ili više nezavisnih varijabli (varijabli poticaja, označenih s X), odnosno omogućiti predviđanje ili utvrđivanje uzročnih odnosa između nezavisnih i zavisnih varijabli. Prvi oblik regresijske analize koji je temeljito istražen je linearna regresija, koja modelira odnos između neprekidnih varijabli odziva i varijabli poticaja tako da je model linearan u parametrima. Nakon što se odabere statistički model, parametri modela obično se procjenjuju metodom najmanjih kvadrata. Jednostavna linearna regresija koristi se za analizu odnosa između jedne nezavisne i jedne zavisne varijable, dok višeparametarska linearna regresija proučava odnos između jedne nezavisne varijable i više zavisnih varijabli.

Prvi oblik regresije, poznat kao postupak najmanjih kvadrata, bio je rezultat rada Legendrea [11] i Gaussa [2] na početku 19. stoljeća. Obojica su primijenili ovu metodu kako bi predvidjeli putanje nebeskih tijela oko Sunca putem astronomskih promatranja. Gauss [3] je kasnije objavio daljnji razvoj teorije najmanjih kvadrata koji uključuje verziju teorema danas poznatog kao Gauss-Markovljev teorem.

Osnovna svrha ovog rada bila je analizirati višeparametarsku ili višestruku linearnu regresiju, često korištenu u različitim znanstvenim područjima poput ekonomije, medicine, marketinga i društvenih znanosti. Primjeri primjene u ekonomiji uključuju analizu javne i privatne potrošnje [8] te financijske uspješnosti privatnih komercijalnih banaka [13], dok se u marketingu primjenjuje za istraživanje utjecaja emocija i kulturološke orijentacije na podržavanje marketinških kampanja [9]. U ovom radu opisana je primjena

višeparametarske linearne regresije u istraživanju utjecaja epoetina alfa na kvalitetu života bolesnika koji su u aktivnom onkološkom liječenju i primaju kemoterapiju bez platine [1].

U prvom poglavlju uvest ćemo osnovne pojmove iz vjerojatnosti i statistike koje ćemo poslije koristiti u razvijanju modela višeparametarske linearne regresije. U drugom poglavlju dajemo pregled modela višeparametarske linearne regresije, procjenjujemo parametre modela metodom najmanjih kvadrata i metodom maksimalne vjerodostojnosti te proučavamo statistička svojstva dobivenih procjenitelja. Zatim provodimo statističke testove kako bi ocijenili njihovu kvalitetu. Opisujemo testove koje se najčešće koriste: test značajnosti svih parametara u modelu, test o značajnosti jednog parametra i test o značajnosti podskupa parametara (vidi [17]). Na kraju računamo pouzdane intervale za procjenitelje regresijskih koeficijenta i za očekivanje ciljne varijable. Teorijske rezultate o višeparametarskoj linearnog regresiji primijenit ćemo u trećem poglavlju u ispitivanju dobrobiti djelatne tvari epoetin alfa na kvalitetu života pacijenata koji boluju od malignih bolesti. Predstavit ćemo rezultate regresijske analize te detaljnije istražiti vezu između razine hemoglobina i kvalitete života.

Poglavlje 1

Vjerojatnost i statistika

Smatra se da je primarna zadaća matematičke statistike donošenje zaključaka o promatranom statističkom fenomenu na temelju konačnog broja statističkih podataka . Inferencijalna statistika može se shvatiti kao znanost o učenju o nepoznatom parametru vjerojatnosne razdiobe iz danog opažanja. Smatra se da je primarna zadaća matematičke statistike donošenje zaključaka o promatranom statističkom fenomenu na temelju konačnog broja statističkih podataka. Grana statistika koja se time bavi naziva se inferencijalna statistika, koja koristi metode kao što su izračun intervala pouzdanosti i testiranje hipoteza. U ovom poglavlju uvodimo temeljne koncepte iz područja vjerojatnosti i statistike.

1.1 Teorija vjerojatnosti

Kako bi mogli iskazati pojmove iz matematičke statistike, potrebno je prisjetiti se osnovnih pojmova i rezultata iz teorije vjerojatnosti. Osnovna pretpostavka za formiranje teorije je postojanje statističkih zakonitosti za promatrano obilježje. U okviru teorije vjerojatnosti, promatrano obilježje naziva se slučajna varijabla, a pripadajuću razdiobu vjerojatnosti nazivamo distribucija slučajne varijable. Kada istovremeno promatramo više statističkih obilježja, govorimo o slučajnom vektoru, čije su komponente slučajne varijable.

Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor, gdje je Ω neprazan skup elementarnih događaja, $\mathcal{F}$ je σ-algebra na Ω i $\mathbb{P}$ je vjerojatnost na izmjerivom prostoru $(\Omega, \mathcal{F})$. Nadalje, neka je $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ izmjeriv prostor sa σ-algebrom Borelovih skupova u $\mathbb{R}^k$ za $k \geq 1$, $k \in \mathbb{N}$.

Definicija 1.1.1. *Funkcija $X : \Omega \to \mathbb{R}^k$ je k-dimenzionalna slučajna veličina ako je X izmjerivo preslikavanje u paru σ-algebri $(\mathcal{F}, \mathcal{B}(\mathbb{R}^k))$ tj. ako vrijedi*

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} \text{ za sve } B \in \mathcal{B}(\mathbb{R}^k)\,.$$

Ako je $k = 1$, X zovemo slučajna varijabla, a ako je $k \geq 2$, X zovemo slučajni vektor.

Definicija 1.1.2. *Neka je X k-dimenzionalna slučajna veličina na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Vjerojatnosna mjera generirana s X je $\mathbb{P}_X : \mathcal{B}(\mathbb{R}^k) \to [0, 1]$ definirana relacijom*

$$\mathbb{P}_X(B) := \mathbb{P}(X \in B),\ B \in \mathcal{B}(\mathbb{R}^k)\,.$$

Funkcija distribucije slučajne veličine X je $F_X : \mathbb{R}^k \to [0, 1]$ definirana relacijom

$$F_X(x) := \mathbb{P}_X(\langle -\infty, x]),\ x \in \mathbb{R}^k\,.$$

Napomenimo da iz prethodne definicije slijedi $F_X(x) = \mathbb{P}(X \leq x)$ za sve $x \in \mathbb{R}^k$.

Razlikujemo diskretne i neprekidne slučajne varijable.

Definicija 1.1.3. *Slučajna veličina X dimenzije k je diskretna ako postoji skup $D \subseteq \mathbb{R}^k$ koji je prebrojiv i takav da je $\mathbb{P}_X(D) = 1$. Funkcija gustoće diskretne k-dimenzionalne slučajne veličine X je $f_X : \mathbb{R}^k \to \mathbb{R}$ definirana formulom*

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}_X(\{x\})\,.$$

Definicija 1.1.4. *Slučajna veličina X dimenzije k je neprekidna ako postoji nenegativna Borelova funkcija f_X definirana na $\mathbb{R}^k$ takva da za sve $x \in \mathbb{R}^k$ vrijedi*

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{\langle -\infty, x]} f_X(y)\, d\lambda(y)\,.$$

pri čemu je λ Lebesgueova mjera definirana na izmjerivom prostoru $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. Funkciju f_X zovemo funkcija gustoće od X.

Za primjere diskretnih i neprekidnih slučajnih varijabli pogledati u [15].

Temeljni koncepti vjerojatnosti i statistike koji su ključni za analizu podataka jesu matematičko očekivanje i varijanca. Matematičko očekivanje ili očekivana vrijednost je generalizacija pojma srednje vrijednosti, a varijanca opisuje raspršenost podataka, odnosno koliko se vrijednosti razlikuju od očekivane vrijednosti.

Definicija 1.1.5. *Za slučajnu varijablu definiranu na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *kažemo da ima matematičko očekivanje ako* $\int_\Omega |X(\omega)|\, d\mathbb{P}(\omega) < \infty$. *U tom slučanju matematičko očekivanje je*

$$\mathrm{E}X := \int_\Omega X d\mathbb{P}\,.$$

Definicija 1.1.6. *Za slučajan vektor* $X = (X_1, X_2, \ldots, X_k)$ *definiran na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *kažemo da ima matematičko očekivanje ako svaka komponenta tog vektora ima matematičko očekivanje. U tom slučanju matematičko očekivanje je*

$$\mathrm{E}X := (\mathrm{E}X_1,\ \mathrm{E}X_2, \ldots,\ \mathrm{E}X_k)\,.$$

Za kraj ovog potpoglavlja o vjerojatnosti definirat ćemo još neke veličine koje se često koriste.

Definicija 1.1.7. *Neka su* X *i* Y *slučajne varijable za koje vrijedi* $\mathrm{E}[X^2] < +\infty$ *i* $\mathrm{E}[Y^2] < +\infty$. *Varijanca od* X *je*

$$\mathrm{Var}X := \mathrm{E}[(X - \mathrm{E}X)^2],$$

standardna devijacija od X *je*

$$\mathrm{std}(X) := \sqrt{\mathrm{Var}X},$$

kovarijanca od X *i* Y *je*

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

i koeficijent korelacije od X *i* Y *(ako su* $\mathrm{std}(X) > 0$ *i* $\mathrm{std}(Y) > 0$*) je*

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\mathrm{std}(X)\mathrm{std}(Y)}.$$

1.2 Osnovni pojmovi matematičke statistike

Osnovnu ulogu u izgradnji statističkog modela ima definiranje skupa svih dopuštenih distribucija za slučajnu veličinu X. U vezi s tim, korisno je poopćiti pojam vjerojatnosnog prostora.

Definicija 1.2.1. *Neka je* $(\Omega, \mathcal{F})$ *izmjeriv prostor i* $\mathcal{P}$ *množina vjerojatnosnih mjera definiranih na* $(\Omega, \mathcal{F})$. *Tada je uređena trojka* $(\Omega, \mathcal{F}, \mathcal{P})$ *statistička struktura.*

Množina $\mathcal{P}$ je najčešće parametrizirana konačnodimenzionalnim parametrom θ

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\},$$

pri čemu je $\Theta \subseteq \mathbb{R}^m$ ($m \geq 1$) skup svih mogućih vrijednosti parametra θ koji zovemo parametarski prostor. Dakle, problem pronalaženja stvarne distribucije postaje problem pronalaženja vrijednosti nepoznatog parametra θ. Od sada nadalje pretpostavljamo da je model parametarski. Za $m = 1$ imamo jednoparametarski model, a za $m > 1$ višeparametarski model.

U statističkom zaključivanju, koristimo podatke koji su prikupljeni na određenom dijelu populacije poznatom kao uzorak.

Definicija 1.2.2. *Slučajni uzorak duljine n je niz nezavisnih jednakodistribiranih slučajnih veličina* $X_1, X_2, \ldots, X_n$.

Intuitivno, slučajni uzorak je niz opažanja vrijednosti veličine X na članovima odabranim u uzorak tako da svaka jedinka populacije ima jednaku šansu biti izabrana. Slučajni uzorak najčešće zapisujemo u obliku n- torke $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. U slučaju kada su elementi niza slučajne varijable, $\mathbf{X}$ je zapravo slučajni vektor koji se sastoji od nezavisnih i jednakodistribuiranih komponenti.

Funkcija slučajnog uzorka zove se statistika. Navodimo formalnu definiciju.

Definicija 1.2.3. *Statistika na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *je svaka slučajna veličina koja je izmjeriva funkcija slučajnog uzorka na toj statističkoj strukturi.*

S obzirom da izučavamo diskretne ili neprekidne slučajne veličine, množinu vjerojatnosnih mjera $\mathcal{P}$ zbog jednoznačnosti možemo poistovjetiti s množinom gustoća $\{f(\cdot;\theta) : \theta \in \Theta\}$ [6]. Neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak iz modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$. Kako bi odredili pravu razdiobu na osnovi opažanja, umjesto korištenja cijelog uzorka, želimo koristiti vrijednost statistike i pritom zadržati sve informacije o nepoznatom parametru. Kažemo da je statistika dovoljna ako predstavlja skup informacija iz uzorka koji sadrži sve bitne informacije o parametrima populacije koja se promatra. Dovoljna statistika ima važnu ulogu u procjeni parametara.

Definicija 1.2.4. *Statistika $T = t(X_1, X_2, \ldots, X_n)$ dimenzije k ($k \geq 1$) je dovoljna za θ ako uvjetna razdioba slučajnog uzorka $(X_1, X_2, \ldots, X_n)$ uz uvjet $T = y$ ne ovisi o parametru θ za svako $y \in \mathbf{R}^k$ za koje postoji ta uvjetna razdioba.*

Definicija dovoljne statistike ponekad nije korisna za pronalaženje dovoljne statistike ili za provjeravanje je li određena statistika dovoljna. Sljedeći teorem daje važnu karakterizaciju dovoljnosti pomoću koje se može provjeriti dovoljnost statistike. Dokaz navedenog teorema može se pronaći u [10].

Napomenimo da je funkcija gustoće za n-dimenzionalni slučajni uzorak $\mathbf{X}$ sa slučajnim veličinama dimenzije d ($d \geq 1$) iz modela $\mathcal{P}$ dana s

$$f_{\mathbf{X}}(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta),\ \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{dn}\,.$$

Teorem 1.2.5 (Neyman-Fisherov teorem o faktorizaciji)**.** *Neka je $T = t(\mathbf{X})$ statistika dimenzije k ($k \geq 1$) i $\mathbf{X}$ slučajni uzorak iz modela $\mathcal{P}$. Nužan i dovoljan uvjet da statistika T bude dovoljna je postojanje nenegativnih funkcija h, g_θ, za sve $\theta \in \Theta$, takvih da se gustoća slučajnog uzorka može faktorizirati na sljedeći način*

$$f_{\mathbf{X}}(\boldsymbol{x}; \theta) = g_\theta(t(\boldsymbol{x}))\, h(\boldsymbol{x}),\ x \in \mathbb{R}^{dn}\,.$$

Bijektivna i izmjeriva transformacija dovoljne statistike je također dovoljna statistika, što proizlazi iz prethodno navedenog teorema [6]. Uočimo da je i cijeli uzorak $\mathbf{X}$ dovoljna statistika. Cilj je od svih dovoljnih statistika izabrati u nekom smislu optimalnu (onu koja najbolje reducira podatke), što sugerira uvođenje pojmova minimalne dovoljne i potpune statistike.

Definicija 1.2.6. *Dovoljna statistika T je minimalna dovoljna statistika za θ ako za svaku drugu dovoljnu statistiku S za θ postoji izmjeriva funkcija g takva da je $T = g(S)$.*

Definicija 1.2.7. *Statistika T za θ je potpuna statistika ako za svaku Borelovu funkciju g za koju vrijedi $(\forall \theta \in \Theta)\, \mathbb{E}_\theta\,[g(T)] = 0$, slijedi da je $(\forall \theta \in \Theta)\, \mathbb{P}_\theta[g(T) = 0] = 1$.*

Nije svaka dovoljna statistika ujedno i potpuna. No, ako je statistika dovoljna i potpuna, pokazuje se da je onda ujedno i minimalna dovoljna. O tome govori sljedeći teorem, a dokaz teorema može se pronaći u [10].

Teorem 1.2.8. *Neka je T dovoljna i potpuna statistika za θ. Tada vrijedi da je T minimalna dovoljna statistika za θ.*

1.3 Procjena parametara

U ovom poglavlju predstavljamo metode za rješavanje problema procjene parametara, odnosno za pronalaženje najboljeg točkovnog procjenitelja parametara populacijske razdiobe. Želimo pronaći statistiku koja će najbolje procijeniti vrijednost nepoznatog parametra.

Neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak iz modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$, gdje su slučajne veličine X_i, $i = 1, 2, \ldots, n$, dimenzije d ($d \geq 1$), a funkcije gustoće su parametrizirane parametrom θ dimenzije m ($m \geq 1$). Neka je $\tau : \Theta \to \mathbb{R}^k$ funkcija čiju vrijednost $\tau(\theta)$ želimo procijeniti na osnovi informacije sadržane u uzorku. Procjenitelj nepoznatog parametra $\tau(\theta)$ je slučajna veličina dimenzije k definirana kao funkcija slučajnog uzorka, odnosno bilo koja statistika $T = t(\mathbf{X})$ dimenzije k. Procjenitelj od $\tau(\theta)$ i realizaciju tog procjenitelja najčešće označavamo s istom oznakom $\hat{\tau}(\theta)$, a značenje oznake ovisi o kontekstu.

Procjenitelja ima mnogo, a mi želimo od svih procjenitelja izabrati onaj optimalan. Neka je $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{dn}$ jedna realizacija slučajnog uzorka $\mathbf{X}$. Uobičajeni pristup kojim prosuđujemo koji je procjenitelj $T = t(\mathbf{X})$, u određenom smislu, bolji procjenitelj za parametar $\tau(\theta)$ je definiranje funkcije gubitka $(t(\boldsymbol{x}), \tau(\theta)) \to L(t(\boldsymbol{x}), \tau(\theta))$. Prema [14], vrijednost funkcije gubitka $L(t(\boldsymbol{x}), \tau(\theta))$ interpretiramo kao gubitak procjenjivanja parametra $\tau(\theta)$ vrijednošću procjenitelja T za danu realizaciju uzorka $\boldsymbol{x} \in \mathbb{R}^{dn}$. Funkcija gubitka je slučajna veličina pa možemo definirati funkciju rizika $\tau(\theta) \to R(\tau(\theta))$ s $R(\tau(\theta)) = \mathrm{E}_\theta[L(T, \tau(\theta)]$. Prema [14], vrijednost funkcije rizika označava očekivani gubitak kada nepoznati parametar $\tau(\theta)$ procjenjujemo vrijednošću procjenitelja T.

Za funkciju gubitka pri procjeni za $\tau(\theta)$ vrijednošću procjenitelja T za danu realizaciju uzorka $\boldsymbol{x}$ najčešće se koristi kvadratna greška

$$L(t(\boldsymbol{x}), \tau(\theta)) = |t(\boldsymbol{x}) - \tau(\theta)|^2,$$

gdje je $|\cdot|$ oznaka za euklidsku normu. Tada je pripadna funkcija rizika dana s

$$R(\tau(\theta)) = \mathrm{E}_\theta[|T - \tau(\theta)|^2],\ \theta \in \Theta.$$

Funkcija rizika mjeri koliko je dobar procjenitelj T za $\tau(\theta)$ u slučaju kada je stvarna vrijednost parametra jednaka θ. S obzirom da ne znamo koja je stvarna vrijednost parametra modela, želimo pronaći onaj procjenitelj koji ima najmanju funkciju rizika za sve $\theta \in \Theta$. Uspoređujući procjenitelje koristeći funkcije rizika možemo doći do problema jer takav najbolji procjenitelj možda ne postoji. Zbog toga uvodimo sljedeću definiciju.

Definicija 1.3.1. *Ako procjenitelj $T = t(\mathbf{X})$ za $\tau(\theta)$ zadovoljava uvjet*

$$(\forall \theta \in \Theta)\ \mathbb{E}_\theta[T] = \tau(\theta)$$

onda se kaže da je T nepristran procjenitelj. Za procjenitelj kažemo da je pristran ako nije nepristran.

Može se pokazati da u klasi nepristranih procjenitelja postoji najbolji procjenitelj u smislu minimalne srednjekvadratne greške. Prije formalne definicije takvog procjenitelja, uvodimo pojam procjenjive funkcije s obzirom da nepristrani procjenitelj za funkciju $\tau(\theta)$ ne mora postojati.

Definicija 1.3.2. *Neka je statistički model parametriziran parametrom θ. Funkcija $\tau(\theta)$ je procjenjiva ako postoji barem jedan nepristran procjenitelj od $\tau(\theta)$.*

U traženju najboljeg procjenitelja, klasu nepristranih procjenitelja dodatno ćemo filtrirati na procjenitelje konačne varijance. Statistika T definirana na statističkom modelu $\mathcal{P}$ je konačne varijance ako je varijanca od T konačna za svaki $\theta \in \Theta$. Neka je $\mathcal{W}_\tau$ množina svih nepristranih procjenitelja za $\tau(\theta)$ konačne varijance.

Definicija 1.3.3. *Neka je $\tau(\theta)$ procjenjiva funkcija. Statistika T je nepristrani procjenitelj uniformno minimalne varijance ili UMVUE procjenitelj od $\tau(\theta)$ ako vrijedi da je $T \in \mathcal{W}_\tau$ i*

$$(\forall S \in \mathcal{W}_\tau)\,(\forall \theta \in \Theta) \quad Var_\theta T \leq Var_\theta S\ .$$

Postoji nekoliko načina traženja procjenitelja uniformno minimalne varijance. Rao-Blackwell teorem, predstavljen u knjizi [10], sugerira da procjenitelj uniformno minimalne varijance treba biti funkcija dovoljne statistike. Iako prethodni teorem ukazuje na način poboljšanja procjenitelja, odnosno smanjenja varijance procjenitelja, još nije sasvim jasno kako pronaći procjenitelj uniformno minimalne varijance. Lehmann-Scheffe teorem,

obrađen u radu [6], govori kako odabrati dovoljnu statistiku u Rao-Blackwell teoremu kako bismo dobili najbolji procjenitelj. Postoje i drugi načini traženja procjenitelja minimalne varijance koji su korisni kada ne postoje potpune i dovoljne statistike ili ih je teško pronaći. Više informacije o tim metodama može se pronaći u [6].

Spomenimo još procjenu metodom najveće vjerodostojnosti. Neka je slučajan uzorak iz statističkog modela $\mathcal{P}$ dan s $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Ako je $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{dn}$ jedna realizacija od $\mathbf{X}$, tada je funkcija vjerodostojnosti $L : \Theta \to \mathbb{R}$ definirana s

$$L(\theta) \equiv L(\theta|\boldsymbol{x}) := f_{\mathbf{X}}(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta), \ \theta \in \Theta .$$

Definicija 1.3.4. *Statistika $\hat{\theta} \equiv \hat{\theta}(\mathbf{X})$ je procjenitelj maksimalne vjerodostojnosti ako vrijedi*

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta|\mathbf{X}) .$$

1.4 Testiranje statističkih hipoteza

Vrlo važna vrsta metode analize koju ukratko uvodimo u ovom odjeljku je testiranje statističkih hipoteza. Pretpostavljamo da promatrano obilježje ima slučajni karakter pa na temelju niza mjerenja želimo donijeti odluku o odbacivanju ili ne odbacivanju određene tvrdnje o promatranom obilježju. Tu tvrdnju zovemo statistička hipoteza, a postupak donošenja odluke zovemo testiranje statističkih hipoteza. U slučaju kada ne odbacujemo statističku hipotezu, ne govorimo o prihvaćanju te hipoteze jer nismo sigurni da je ona istinita, već samo da primijenjeni test nije uspio otkriti značajno odstupanje od nje. Osnovna ideja testiranja statističkih hipoteza jest donošenje odluke o tome je li slučajni uzorak tipičan u usporedbi s populacijom, pretpostavljajući da je hipoteza koju smo formirali o populaciji istinita.

Uz oznake koje smo uveli u prethodnim potpoglavljima, neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak iz parametarskog modela $\mathcal{P} = \{f(\cdot; \theta) : \theta \in \Theta\}$, gdje su slučajne veličine dimenzije d ($d \geq 1$), a parametar θ dimenzije m ($m \geq 1$). Neka su Θ_0 i Θ_1 neprazni disjunktni skupovi koji čine jednu particiju skupa Θ. Pretpostavimo da želimo testirati sljedeće hipoteze:

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1 . \tag{1.1}$$

Postavljene hipoteze zovemo nulta hipoteza, odnosno alternativna hipoteza. Uočimo da su hipoteze zapravo tvrdnje o distribuciji statističkog obilježja kojeg izučavamo. Kako bi donijeli odluku o odbacivanju ili neodbacivanju nulte hipoteze, potrebno je kontruirati statistički test. Statistički test možemo promatrati kao funkciju koja svakoj realizaciji slučajnog uzorka pridružuje vrijednost 1 ili 0, gdje vrijednost 1 označava odbacivanje nulte hipoteze u korist alternativne hipoteze, dok vrijednost 0 označava neodbacivanje nulte hipoteze.

Definicija 1.4.1. *Statistički test hipoteze H_0 u odnosu na H_1 je funkcija $\tau : \mathbb{R}^{dn} \to \{0, 1\}$.*

Odluka donesena statističkim testom temelji se na uzorku iz populacije pa ne može biti u potpunosti pouzdana. U slučaju kada donesena odluka nije ispravna, razlikujemo pogrešku prve vrste i pogrešku druge vrste. Odbacivanje nulte hipoteze kada je ona istinita naziva se pogreška prve vrste. Pogreška druge vrste je neodbacivanje nulte hipoteze kada je alternativna hipoteza istinita.

Kritično područje testa $C_\tau \subseteq \mathbb{R}^{dn}$ definiramo kao skup svih realizacija uzorka za koje se nulta hipoteza odbacuje u korist alternativne tj. $C_\tau := \tau^{-1}(1) = \{\boldsymbol{x} \in \mathbb{R}^{dn} : \tau(\boldsymbol{x}) = 1\}$. To su upravo one točke u kojima se događa značajno odstupanje od pretpostavljene hipoteze.

Definicija 1.4.2. *Jakost testa τ je funkcija $\gamma : \Theta \to [0, 1]$ definirana s*

$$\gamma_\tau(\theta) := \mathbb{E}_\theta[\tau(\mathbf{X})] = \mathbb{P}_\theta(\mathbf{X} \in C_\tau),\ \theta \in \Theta\,.$$

Jakost testa interpretiramo kao vjerojatnost odbacivanja nulte hipoteze u korist alternativne kada je stvarni parametar jednak θ. Za test hipoteza 1.1, ako uzememo $\theta \in \Theta_0$, funkcija jakosti testa je zapravo vjerojatnost pogreške prve vrste koju želimo ograničiti.

Definicija 1.4.3. *Značajnost testa τ je broj*

$$\alpha_\tau := \sup_{\theta \in \Theta_0} \gamma_\tau(\theta)\,.$$

Uobičajeno unaprijed zadajemo razinu značajnosti testa s oznakom $\alpha \in \langle 0, 1\rangle$ pomoću koje ograničavamo vjerojatnosti pogreške prve vrste. Kažemo da test τ ima (zadanu) razinu značajnosti α ako mu je značajnost α_τ manja ili jednaka od α. Među svim testovima koji imaju razinu značajnosti α, želimo pronaći test s najmanjom vjerojatnosti pogreške druge vrste. Takve testove zovemo uniformno najjačima za zadanu razinu značajnosti.

Definicija 1.4.4. *Za statistički test τ nul hipoteze $H_0 : \theta \in \Theta_0$ naprema alternativi $H_1 : \theta \in \Theta_1$ s značajnosti α_τ kažemo da je uniformno najjači na razini značajnosti α ako je $\alpha_\tau \leq \alpha$ i za svaki drugi test τ' istih hipoteza takav da je $\alpha_{\tau'} \leq \alpha$ vrijedi:*

$$\gamma_{\tau'}(\theta) \leq \gamma_\tau(\theta) \text{ za sve } \theta \in \Theta_1 .$$

Primijetimo da za sve $\theta \in \Theta_1$ vrijedi da je $1 - \gamma_\tau(\theta) = \mathbb{P}_\theta(\mathbf{X} \notin C_\tau)$ vjerojatnost pogreške druge vrste. Dakle, uniformno najjači test je optimalan u smislu da među svim testovima koji imaju razinu značajnosti ima najmanju vjerojatnost pogreške druge vrste. Odgovor na pitanje kako pronaći uniformno najjači test, ukoliko postoji, daje nam Neyman-Pearsonova lema (za detalje vidi [6]).

Umjesto na temelju kritičnog područja, odluke o odbacivanju ili neodbacivanju nulte hipoteze možemo donijeti pomoću p-vrijednosti. P-vrijednost je vjerojatnost da se dogodi dobiveni uzorak ili ekstremniji u slučaju da je nulta hipoteza istinita. Koristeći p-vrijednost možemo donijeti odluku: u slučaju kada je p-vrijednost manja ili jednaka od prethodno postavljene razine značajnosti α, odbacujemo nultu hipotezu u korist alternativne na razini značajnosti α, inače, nultu hipotezu ne odbacujemo.

Poglavlje 2

Višeparametarska linearna regresija

Višeparametarska linearna regresija je statistička metoda koja omogućava modeliranje odnosa između dvije ili više nezavisnih varijabli i jedne zavisne varijable. Cilj ove metode je pronaći linearan model koji najbolje odgovara dostupnim podacima, odnosno odrediti nepoznate koeficijente u modelu. Osim toga, važno je procijeniti i kvalitetu modela. Ukoliko je dobar, dobiveni model omogućava predviđanje vrijednosti varijable odziva (zavisne varijable) varijable na temelju vrijednosti varijabli poticaja (nezavisnih varijabli). Napomenimo da se prema [12] model višeparametarske linearne regresije najčešće koristi kao empirijski model: stvarna funkcijska veza između varijable odziva i varijabli poticaja je nepoznata, ali model pruža dobru aproksimaciju nepoznate funkcije koja je prilagođena danim podacima. Međutim, Verbeek [18] tvrdi da specifikacija modela nije jednostavna jer ne postoji jednostavno pravilo koje propisuje kako odabrati prikladnu specifikaciju.

2.1 Model višeparametarske linearne regresije

Radi jasnijeg definiranja modela višedimenzionalne linarne regresije uvedimo prikladne oznake. Općenito, varijabla odziva Y može biti povezana s k ($k \in \mathbb{N}$) ulaznih varijabli. Neka su $x_1, x_2, \ldots, x_k$ neslučajne ulazne varijable ili varijable poticaja. Najčešće se ulazne varijable zadaju, a Y opaža (mjeri). Neka su $\beta_0, \beta_1, \ldots, \beta_k$ konstantni realni brojevi i neka je ε slučajna varijabla s očekivanjem $E[\varepsilon] = 0$ i varijancom $V[\varepsilon] = \sigma^2 > 0$.

Linearni višeparametarski regresijski model s k prediktora definiramo relacijom

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \tag{2.1}$$

gdje realni brojevi $\beta_0, \beta_1, \ldots, \beta_k$ predstavljaju parametre modela, a ε interpretiramo kao slučajnu grešku ili šum.

Relacijom (2.1) opisana je situacija gdje se smatra da je vrijednost odzivne varijable Y posljedica postojanja linearne zavisnosti odzivne varijable o varijablama poticaja uz slučajnu grešku. Slučajna greška u sebi uključuje druge faktore koji također utječu na varijablu odziva, a nisu uključeni u model. Parametar β_j označava očekivanu promjenu u odzivnoj varijabli po jedinici promjene u varijabli x_j kada su sve preostale varijable poticaja ($x_i \neq x_j$) ostale nepromijenjene.

Nadalje , neka je n ($n \in \mathbb{N}$) broj opažanja i neka $x_{i1}, x_{i2} \ldots, x_{ik}$ ($i = 1, \ldots, n$) označavaju i-tu vrijednost ulaznih varijabli za koje je y_i pripadna vrijednost izlazne slučajne varijable Y_i. Također, neka su ε_i ($i = 1, \ldots, n$) nezavisne slučajne varijable s očekivanjem $E[\varepsilon_i] = 0$ i varijancom $V[\varepsilon_i] = \sigma^2 > 0$. Tada opći linearni regresijski model možemo zapisati u obliku

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n\,. \tag{2.2}$$

Uvedimo dodatne oznake kako bi model (2.2) zapisali u matričnom obliku. Neka je

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Općenito, $\mathbf{Y}$ i $\boldsymbol{\varepsilon}$ su n-dimenzionalni slučajni vektori stupci koje zovemo vektor izlaznih podataka, odnosno vektor greške. $\mathbf{X}$ je $n \times (k+1)$ matrica ulaznih podataka, a $\boldsymbol{\beta}$ je vektor stupac regresijskih koeficijenata. Pretpostavljamo da vrijedi $n > k$.

Primjenom uvedenih oznaka k-dimenzionalni linearni regresijski model može se zapisati u matričnom obliku

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,. \tag{2.3}$$

2.2 Procjena parametara

Prema Montgomeryju [12], u većini problema iz stvarnog svijeta vrijednosti parametara (koeficijenata regresije β_i) i varijance greške (σ^2) neće biti poznati i moraju biti procijenjeni iz uzorka. Glavni izazov višeparametarske linearne regresije je pronalaženje dobrih procjenitelja. Zatim se vrijednost procjenitelja računa, kao i inače, pomoću danog niza podataka.

U opisanom modelu, slučajni uzorak je niz $(x_{i1}, x_{i2}, \ldots, x_{ik}, Y_i)$, $i = 1, \ldots, n$, gdje su x_{ij}, $j = 1, \ldots, k$ vrijednosti j-te ulazne varijable u i-tom mjerenju, a $Y_1, \ldots, Y_n$ međusobno nezavisne slučajne varijable za koje vrijedi

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}, \quad V[Y_i] = \sigma^2, \quad i = 1, \ldots, n\,.$$

Metoda najmanjih kvadrata

Za pronalaženje najbolje procjene $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ nepoznatih koeficijenata u linearnoj regresijskoj analizi najčešće se koristi metoda najmanjih kvadrata. Dakle, funkciju najmanjih kvadrata definiranu s

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{2.4}$$

treba minimizirati obzirom na $\beta_0, \beta_1, ..., \beta_k$.
U modelu višeparametarske regresije prikladnije je funkciju definiranu s (2.4) izraziti u matričnom zapisu. Vrijedi

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^\tau \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\tau (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\,.$$

Označimo s $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^\tau$ procjenu za nepoznati vektorski parametar $\boldsymbol{\beta}$. Dakle, želimo odrediti $\hat{\boldsymbol{\beta}}$ tako da vrijedi

$$\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}}).$$

Primijetimo da se $S(\boldsymbol{\beta})$ može izraziti kao

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{Y}^\tau\mathbf{Y} - \boldsymbol{\beta}^\tau\mathbf{X}^\tau\mathbf{Y} - \mathbf{Y}^\tau\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\tau\mathbf{X}^\tau\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}^\tau\mathbf{Y} - 2\boldsymbol{\beta}^\tau\mathbf{X}^\tau\mathbf{Y} + \boldsymbol{\beta}^\tau\mathbf{X}^\tau\mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

gdje druga jednakost slijedi zbog toga što je $\boldsymbol{\beta}^\tau\mathbf{X}^\tau\mathbf{Y}$ skalar pa njegovim transponiranjem $(\boldsymbol{\beta}^\tau\mathbf{X}^\tau\mathbf{Y})^\tau = \mathbf{Y}^\tau\mathbf{X}\boldsymbol{\beta}$ dobijemo isti skalar. Procjene regresijskih koeficijenata dobivene metodom najmanjih kvadrata moraju biti stacionarne točke funkcije najmanjih kvadrata $S(\boldsymbol{\beta})$

$$\frac{\partial S}{\partial \boldsymbol{\beta}}\Big|_{\hat{\beta}} = -2\mathbf{X}^\tau\mathbf{Y} + 2\mathbf{X}^\tau\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0},$$

što možemo pojednostaviti

$$\mathbf{X}^\tau\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\tau\mathbf{Y}. \tag{2.5}$$

$\mathbf{X}$ je neslučajna matrica dimenzije $n \times (k+1)$ kojoj su stupci linearno nezavisni. S obirom da je matrica X punog ranga, kvadratna matrica $\mathbf{X}^\tau\mathbf{X}$ dimenzije $(k+1)$ je punog ranga pa je regularna, odnosno postoji inverz. Kako bismo riješili jednadžbu (2.5), pomnožimo obe strane jednadžbe inverzom od $\mathbf{X}^\tau\mathbf{X}$. Dakle, vrijedi

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}. \tag{2.6}$$

Preostalo je pokazati da je jedinstvena stacionarna točka ujedno i jedinstvena točka minimuma od funkcije $S(\boldsymbol{\beta})$. Druga derivacija od S jednaka je

$$\frac{\partial S}{\partial^2 \boldsymbol{\beta}} = 2\mathbf{X}^\tau\mathbf{X}$$

što je pozitivno definitna matrica neovisno o $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$. Dakle, S je konveksna funkcija na skupu $\mathbb{R}^{k+1}$ pa je jedinstvena stacionarna točka $\hat{\boldsymbol{\beta}}$ ujedno i točka minimuma. Iz toga slijedi da je procjenitelj metodom najmanjih kvadrata za nepoznate koeficijente linerarne regresije jednak (2.6).

Geometrijska interpretacija metode najmanjih kvadrata

Ponekad je korisno geometrijski interpretirati procjenu metodom najmanjih kvadrata. Napomenimo da je za ulazne varijable $x_1, x_2, \ldots, x_k$ regresijska funkcija oblika $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. Niz vrijednosti regresijske funkcije $(\mathbf{X}\boldsymbol{\beta})^\tau$ može se shvatiti kao vektor iz prostora $\mathbb{R}^n$, a budući da je i $\mathbf{Y}^\tau \in \mathbb{R}^n$, njihova apstrakstna udaljenost u prostoru $\mathbb{R}^n$ može se izraziti pomoću

$$\left|\mathbf{Y}^\tau - (\mathbf{X}\boldsymbol{\beta})^\tau\right| = \sqrt{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2}.$$

Traženje procjenitelja metodom najmanjih kvadrata može se geometrijski interpretirati kao traženje parametra $\hat{\boldsymbol{\beta}}$ za koji su vektori $\mathbf{Y}^\tau$ i $(\mathbf{X}\hat{\boldsymbol{\beta}})^\tau$ međusobno najbliži tj. njihova udaljenost u prostoru $\mathbb{R}^n$ je najmanja.
Matrica X sastoji se od $(k+1)$ vektor-stupaca. Ako označimo stupce od X s $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k$, može se pisati

$$\mathbf{X}\boldsymbol{\beta} = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \ldots + \beta_k\mathbf{x}_k\,.$$

Ovih $k+1$ linearno nezavisnih stupaca razapinju $k+1$-dimenzionalni potprostor $\mathcal{M}$ u prostoru $\mathbb{R}^n$. Dakle, problem traženja najbolje procjene $\hat{\boldsymbol{\beta}}$ može se shvatiti kao traženje onog vektora u potprostoru $\mathcal{M}$ koji je najbliži vektoru izlaznih podataka $\mathbf{Y}$. Znamo da je to onaj vektor $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ koji se dobije kao ortogonalna projekcija vektora $\mathbf{Y}$ na potprostor $\mathcal{M}$ (vidi [14]). U tom slučaju je vektor $\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ ortogonalan na svaki vektor koji razapinje potprostor $\mathcal{M}$ pa vrijedi

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\tau\mathbf{1} = 0, \quad (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\tau\mathbf{x}_j = 0, \quad j = 1, \ldots, k\,,$$

odnosno u matričnom zapisu

$$\mathbf{X}^\tau(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0 \quad \text{ili} \quad \mathbf{X}^\tau\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\tau\mathbf{Y}\,.$$

Uočimo da smo dobili isto rješenje za procjenu $\hat{\boldsymbol{\beta}}$ vektorskog parametra $\boldsymbol{\beta}$ kao u (2.5).

Svojstva procjenitelja metodom najmanjih kvadrata

Sada ćemo istaknuti neka statistička svojstva metode najmanjih kvadrata za procjenu nepoznatih parametara. Posebno je važan Gauss-Markovljev teorem iz kojeg slijedi da je

procjenitelj metodom najmanjih kvadrata za parametre linearnog regresijskog modela najbolji linearni nepristrani procjenitelj.

Pretpostavili smo da su slučajne greške $\varepsilon_i, i = 1, \ldots, n$, nezavisne slučajne varijable s očekivanjem $E[\varepsilon_i] = 0$ i varijancom $V[\varepsilon_i] = \sigma^2 > 0$. Dakle, za slučajni vektor $\boldsymbol{\varepsilon}$ vrijedi

$$\mathrm{E}\boldsymbol{\varepsilon} = 0\,, \tag{2.7}$$

$$\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 I\,. \tag{2.8}$$

Pokažimo da je vektor $\hat{\boldsymbol{\beta}}$ iz (2.6) linearan procjenitelj. Za slučajnu varijablu $\hat{\beta}_i$ onda vrijedi

$$\hat{\beta}_i = \sum_{j=1}^{n} c_{i,j} Y_j \quad i = 1, \ldots, p\,, \tag{2.9}$$

pri čemu je $c_{i,j}$ element u i-tom retku i j-tom stupcu matrice $(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau$. Dakle, slučajni vektor $\hat{\boldsymbol{\beta}}$ je linearan procjenitelj.
Promotrimo sada pristranost.

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}[\,(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}\,] \\ &= \mathrm{E}[\,(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\,] \\ &= \mathrm{E}[\,(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\boldsymbol{\varepsilon}\,] = \boldsymbol{\beta}\,, \end{aligned} \tag{2.10}$$

gdje zadnja jednakost slijedi zbog toga što je $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ i $(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{X} = \mathbf{I}$. Dakle, slučajni vektor $\hat{\boldsymbol{\beta}}$ nepristran je procjenitelj za $\boldsymbol{\beta}$.
Kovarijacijska matrica od slučajnog vektora $\hat{\boldsymbol{\beta}}$ je $(k+1)\times(k+1)$ simetrična matrica čiji i-ti dijagonalni element sadrži varijancu od $\hat{\beta}_i$, a nedijagonalni element na mjestu (i,j) sadrži kovarijancu od $\hat{\beta}_i$ i $\hat{\beta}_j$. Računamo

$$\begin{aligned} \mathrm{cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{cov}((\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}) = \mathrm{cov}((\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) \\ &= \mathrm{cov}(\boldsymbol{\beta} + (\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\boldsymbol{\varepsilon}) = (\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathrm{cov}(\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\tau\mathbf{X})^{-1}\,. \end{aligned} \tag{2.11}$$

Primijetimo da smo četvrtoj jednakosti koristili tvrdnju da za m-dimenzionalni slučajni vektor $X = AY + \mu$, gdje je Y n-dimenzionalnu slučajni vektor, A neslučajna matrica dimenzije $m \times n$ i $\mu \in \mathbb{R}^n$ zadani vektor, vrijedi

$$\mathrm{cov}(X) = A\mathrm{cov}(Y)A^\tau.$$

Iz prethodnog rapisa slijedi da je varijanca od $\hat{\beta}_j$ jednaka $\sigma^2 d_{jj}$ i kovarijanca od $\hat{\beta}_i$ i $\hat{\beta}_j$ je $\sigma^2 d_{ij}$, gdje je d_{ij} element na mjestu (i, j) matrice $(\mathbf{X}^\tau\mathbf{X})^{-1}$. Iz raspisa također slijedi

$$\mathrm{E}\mathbf{Y} = \mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + \mathrm{E}\boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} \tag{2.12}$$

$$\mathrm{cov}(\mathbf{Y}) = \mathrm{cov}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 I\,. \tag{2.13}$$

Sada kada smo pokazali da je procjenitelj metodom najmanjih kvadrata linearan i nepristran, zanima nas je li procjenitelj (2.9) najbolji među svim linearnim nepristranim procjeniteljima za nepoznati parametar β_i, odnosno ima li najmanju varijancu. Sljedeći teorem daje nam odgovor na to pitanje.

Teorem 2.2.1 (Gauss-Markov)**.** *Neka je $\hat{\boldsymbol{\beta}}$ procjenitelj metodom najmanjih kvadrata za parametre linearnog regresijskog modela i neka je $L : \mathbb{R}^{k+1} \to \mathbb{R}$ linearni funkcional parametara $L(\boldsymbol{\beta}) = l^\tau\boldsymbol{\beta}$. Pretpostavimo da za slučajne greške ε_i, $i = 1, 2, \ldots, n$, vrijede Gauss-Markovljevi uvjeti:*

(i) $E[\varepsilon_i] = 0$ *za sve* $i = 1, 2, \ldots, n$,

(ii) $Var[\varepsilon_i] = \sigma^2$ *za sve* $i = 1, 2, \ldots, n$,

(iii) $Cov[\varepsilon_i, \varepsilon_j] = 0$ *za sve* $i \neq j$. *Tada je statistika*

$$T = l^\tau\hat{\boldsymbol{\beta}}$$

najbolji linearni nepristrani procjenitelj za $L(\boldsymbol{\beta})$.

Dokaz. Statistika T je najbolji linearni nepristrani procjenitelj za $L(\boldsymbol{\beta})$ ako je linearan, nepristran i u klasi svih nepristranih linearnih procjenitelja za $L(\boldsymbol{\beta})$ ima najmanju varijancu. Pokažimo prvo da je T linearan procjenitelj. Vrijedi

$$T = l^\tau\hat{\boldsymbol{\beta}} = l^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y} = c_0^\tau\mathbf{Y},$$

gdje je $c_0 = \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}l \in \mathbb{R}^n$. Dakle, T je linearan procjenitelj.
Zatim pokažimo da je T nepristran procjenitelj.

$$\mathrm{E}_\theta(T) = \mathrm{E}_\theta(l^\tau\hat{\boldsymbol{\beta}}) = l^\tau\mathrm{E}_\theta(\hat{\boldsymbol{\beta}}) = l^\tau\boldsymbol{\beta}$$

povlači da je T nepristran procjenitelj.
Preostalo je pokazati da T ima najmanju varijancu u klasi svih nepristranih linearnih procjenitelja za $L(\boldsymbol{\beta})$. Neka je $U = c^\tau \mathbf{Y}$ neki drugi nepristrani linearni procjenitelj za $L(\boldsymbol{\beta})$. Dakle, zbog nepristranosti od U, vrijedi

$$l^\tau \boldsymbol{\beta} = \mathrm{E}_\theta(U) = \mathrm{E}_\theta(c^\tau \mathbf{Y}) = c^\tau \mathrm{E}_\theta(\mathbf{Y}) = c^\tau \mathbf{X}\boldsymbol{\beta}.$$

Slijedi da je U nepristran ako i samo ako je $l^\tau \boldsymbol{\beta} = c^\tau \mathbf{X}\boldsymbol{\beta}$, odnosno $l = \mathbf{X}^\tau c$.
Računamo

$$\begin{aligned}\mathrm{Var}U - \mathrm{Var}T &= \mathrm{Var}(c^\tau \mathbf{Y}) - \mathrm{Var}(l^\tau \hat{\boldsymbol{\beta}}) = c^\tau \mathrm{cov}(\mathbf{Y})c - l^\tau \mathrm{cov}(\hat{\boldsymbol{\beta}})l \\ &= \sigma^2(c^\tau c - l^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}l) = \sigma^2(c^\tau I c - c^\tau \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau c) \\ &= \sigma^2 c^\tau(I - \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau)c = \sigma^2 c^\tau M c \geq 0,\end{aligned}$$

pri čemu je $H = \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau$ ortogonalni projektor na $\mathcal{M}$, a $M = I - H$ ortogonalni projektor na ortogonalni komplement od $\mathcal{M}$. M je pozitivno semidefinitan operator pa vrijedi da je $c^\tau M c \geq 0$. Dakle, vrijedi $\mathrm{Var}T \leq \mathrm{Var}U$. □

Napomena 2.2.2. *Uzme li se $l = (1, 0, \ldots, 0)$, teorem* (2.2.1) *tvrdi da je statistika $T = l^\tau \hat{\boldsymbol{\beta}} = \hat{\beta}_0$ najbolji linearni nepristrani procjenitelj za $L(\boldsymbol{\beta}) = \beta_0$. Slično se može pokazati i za procjenitelje $\hat{\beta}_j$, $j = 1, \ldots, k$. Dakle, procjenitelj $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ je nepristrani linearni procjenitelj za $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ s najmanjom varijancom, odnosno najbolji linearni nepristrani procjenitelj.*

Procjena varijance greške (σ^2)

Procjenitelj za nepoznati parametar σ^2 možemo izvesti iz sume kvadrata razlike između izmjerene i procijenjene vrijednosti izlazne varijable Y.
Uvedimo prvo oznake koje ćemo koristiti. Neka su $x_1, x_2, \ldots, x_k$ neslučajne ulazne varijable. S $\hat{Y}$ označava se procjena izlazne varijable pa se prilagođeni regresijski model može pisati kao

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{2.14}$$

Dakle, za vektor procijenjih vrijednosti izlazne varijable $\hat{\mathbf{Y}}$ vrijedi

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y} = H\mathbf{Y}, \tag{2.15}$$

gdje je matrica H dimenzije $n \times n$ definirana s $H := \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau$ ortogonalni projektor na potprostor od $\mathbb{R}^n$ razapet stupcima od X. Nadalje, s e_i označava se razlika između stvarne vrijednosti Y i procijenjene vrijednosti $\hat{Y}$ izlazne varijable. Dakle, vrijedi $e = Y - \hat{Y}$. Vektor $\mathbf{e} = (e_1, e_2, \ldots, e_n)$ zovemo vektor reziudala i pišemo

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}\,. \tag{2.16}$$

Koristeći relaciju (2.15), vektor reziduala $\mathbf{e}$ može se izraziti i na način koji će se pokazati korisnim

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - H\mathbf{Y} = (I - H)\mathbf{Y} = M\mathbf{Y},$$

gdje je matrica $M = I - H$ ortogonalni projektor na ortogonalni komplement potprostora od $\mathbb{R}^n$ razapetim stupcima od X.

Definirajmo statistiku

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^\tau M\mathbf{Y}}{n - k - 1}. \tag{2.17}$$

Propozicija 2.2.3. *Statistika $\hat{\sigma}^2$ definirana s* (2.17) *je nepristran procjenitelj za nepoznati parametar zajedničke varijance σ^2.*

Dokaz. Vrijedi

$$\begin{aligned}\mathbf{Y}^\tau M\mathbf{Y} &= (M(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}), \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (M\boldsymbol{\varepsilon}, \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (M\boldsymbol{\varepsilon}, \mathbf{X}\boldsymbol{\beta}) + (M\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}) = (\boldsymbol{\varepsilon}, M\mathbf{X}\boldsymbol{\beta}) + (M\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}) \\ &= (M\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^\tau M\boldsymbol{\varepsilon}.\end{aligned} \tag{2.18}$$

Druga i predzadnja jednakost slijede iz činjenice da je $M\mathbf{X} = 0$, a u drugom redu smo iskoristili da je $M = M^\tau$. Neka je m_{ij} (i, j)-ti član matrice M. Slijedi

$$\mathbf{Y}^\tau M\mathbf{Y} = \boldsymbol{\varepsilon}^\tau M\boldsymbol{\varepsilon} = \sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + 2\sum_{1\le i\le j\le n} m_{ij}\varepsilon_i\varepsilon_j\,.$$

Sada možemo računati matematičko očekivanje

$$\begin{aligned}\mathrm{E}(\mathbf{Y}^\tau M\mathbf{Y}) &= \sum_{i=1}^{n} m_{ii}\mathrm{E}(\varepsilon_i^2) + 2\sum_{1\le i\le j\le n} m_{ij}\mathrm{E}(\varepsilon_i\varepsilon_j) \\ &= \sigma^2\sum_{i=1}^{n} m_{ii} = \sigma^2 tr(M) = \sigma^2 r(M) = \sigma^2(n - k - 1).\end{aligned}$$

U drugoj jednakosti iskoristili smo (2.7), odnosno pretpostavke o slučajnim greškama. Predzadnja jednakost vrijedi jer su rang i trag jednaki za ortogonalni projektor. Iz linearnosti matematičkog očekivanja slijedi tvrdnja propozicije:

$$\mathrm{E}\left(\frac{\mathbf{Y}^\tau M\mathbf{Y}}{n-k-1}\right) = \frac{1}{n-k-1}\mathrm{E}(\mathbf{Y}^\tau M\mathbf{Y}) = \sigma^2.$$

□

Metoda maksimalne vjerodostojnosti

Ukoliko u višedimenzionalnom linearnom regresijskom modelu uvedemo dodatnu pretpostavku o normalnoj distribuciji slučajnih grešaka, pokazuje se da se primjenom metode maksimalne vjerodostojnosti dobije isti procjenitelj za parametre modela kao i metodom najmanjih kvadrata. Procjene za nepoznati parametar σ^2 razlikuju se samo u faktoru.

Dakle, za višeparametarski linearni regresijski model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

pretpostavljamo da vrijedi $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, odnosno da su greške ε_i normalno distribuirane nezavisne slučajne varijable s očekivanjem $E[\varepsilon_i] = 0$ i konstantnom varijancom $V[\varepsilon_i] = \sigma^2$. Funkcija gustoće normalne slučajne varijable s očekivanjem μ i varijancom σ^2 je

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

Funkcija vjerodostojnosti dana je sljedećim izrazom

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} f(y_i) \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x_i}^\tau\boldsymbol{\beta})^2\right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}. \end{aligned}$$

Uobičajeno računamo s funkcijom vjerodostojnosti u logaritamskom obliku

$$ln\, L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\, ln(2\pi) - n\, ln(\sigma) - \frac{1}{2\sigma^2}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,. \tag{2.19}$$

Vidimo da za fiksnu vrijednost σ funkcija log-vjerodostojnosti (2.19) postiže maksimum kada član $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\tau(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ postiže minimum. Dakle, procjene nepoznatih parametara $\beta_0, \beta_1, \ldots, \beta_k$ dobivene metodom maksimalne vjerodostojnosti i metodom najmanjih kvadrata su ekvivalentne. Parcijalnim deriviranjem (2.19) po parametru σ^2 dobije se procjena za nepoznati parametar σ^2 dobivena metodom maksimalne vjerodostojnosti

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\tau(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n}\,. \tag{2.20}$$

Dobivena procjena za varijancu slučajne greške σ^2 razlikuje se od procjene dobivene metodom najmanjih kvadrata po faktoru, iz čega slijedi da ML-procjenitelj nije nepristran.

2.3 Testiranje hipoteza

U statističkoj analizi višeparametarskog regresijskog modela prvo smo se bavili problemom procjene parametara. Nakon procjene parametara, bitno je testirati kvalitetu procjene. Želimo odgovoriti na sljedeća pitanja:

1. Kakva je općenito prikladnost modela?
2. Koji regresori su značajni?

U ovom potpoglavlju navodimo nekoliko procedura testiranja hipoteza koje su korisne za obradu navedenih pitanja.
Pretpostavili smo da su slučajne greške ε_i nezavisne slučajne varijable s očekivanjem $E[\varepsilon_i] = 0$ i varijancom $Var[\varepsilon_i] = \sigma^2$. Da bi se dobili određeni rezultati, potrebno je uvesti dodatnu pretpostavku o normalnoj distribuciji slučajne greške. Dakle, pretpostavljamo da slučajni vektorprati normalnu distribuciju kojoj je n-dimenzionalni nul vektor $\mathbf{0}$ vektor očekivanja i $\sigma^2 I$ kovarijacijska matrica, odnosno vrijedi

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)\,. \tag{2.21}$$

Iz gornje pretpostavke i definicije modela (2.1) odmah slijedi da su slučajne varijable Y normalno distribuirane s očekivanjem $\beta_0 + \sum_{j=1}^{k} \beta_j x_j$ i varijancom σ^2. Dakle, slučajnom

vektoru $\mathbf{Y}$ pripada normalna razdioba

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)\,. \tag{2.22}$$

Nadalje, iz (2.9) vidimo da je vektor procjenitelja $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ za regresijske koeficijente linearno ovisan o slučajnom vektoru Y. Uz (2.10) i (2.11), slijedi da je slučajni vektor $\hat{\boldsymbol{\beta}}$ normalno distribuiran s vektorom očekivanja $\boldsymbol{\beta}$ i kovarijacijskom matricom $\sigma^2(\mathbf{X}^\tau\mathbf{X})^{-1}$, to jest

$$\hat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\tau\mathbf{X})^{-1})\,. \tag{2.23}$$

Uz dodatnu pretpostavku o normalnosti slučajnih grešaka možemo također pokazati da za procjenitelj definiran s (2.17) od varijance greške σ^2 vrijedi

$$(n-k-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k-1)\,. \tag{2.24}$$

Naime, definirajmo Z kao standardni normalni slučajni vektor:

$$\mathbf{Z} := \frac{1}{\sigma}\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, I)\,.$$

Uz prethodno uvedene oznake i jednakost (2.18) vrijedi

$$(n-k-1)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\mathbf{Y}^\tau M\mathbf{Y}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}^\tau M\boldsymbol{\varepsilon}}{\sigma^2} = \mathbf{Z}^\tau M\mathbf{Z} \sim \chi^2(n-k-1)$$

prema teoremu 2.7 iz [16] i zbog toga što je $r(M) = n-k-1$. Dodatno, može se pokazati da su $\hat{\boldsymbol{\beta}}$ i $\hat{\sigma}^2$ nezavisne slučajne veličine (vidi [7]).

Koeficijent determinacije $\mathbf{R^2}$

Prikladnost modela možemo ispitati koristeći koeficijent determinacije R^2 i korigirani koeficijent determinacije $\bar{R}^2$. Uvodimo sljedeće oznake

$$SSE := \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^\tau\mathbf{e} = \mathbf{Y}^\tau M\mathbf{Y}, \tag{2.25}$$

$$SSR := \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2, \tag{2.26}$$

$$SST := \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{2.27}$$

pri čemu je $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Veličina SSE zove se zbroj kvadrata greškaka, veličina SSR je regresijski zbroj kvadrata, a veličina SST ukupni zbroj kvadrata.

Teorem 2.3.1 (Osnovna jednakost analize varijance za regresijski model)**.** *Ukupan zbroj kvadrata SST podijeljen je na regresijski zbroj kvadrata SSR i zbroj kvadrata grešaka SSE, odnosno vrijedi*

$$SST = SSR + SSE\,. \tag{2.28}$$

Dokaz. Neka je $\mathcal{N}$ potprostor od $\mathbb{R}^n$ razapet vektorom-stupcem jedinica $\mathbf{1} = (1, 1, \ldots, 1)^\tau \in \mathbb{R}^n$. Uz prethodno uvedeni potprostor $\mathcal{M}$ razapet stupcima matrice $\mathbf{X}$, vrijedi $\mathcal{N} \leq \mathcal{M}$ tj. $\mathcal{N}$ je potprostor od $\mathcal{M}$. Označimo s N ortogonalni projektor na potprostor $\mathcal{N}$, dakle $N\mathbf{Y} = \bar{\mathbf{Y}}\mathbf{1}$. Iz teorije o projektorima slijedi da je $N = N^\tau$, $N^2 = N$, $NM = MN = \mathbf{0}$, $NH = HN = N$. Veličinu SST sada možemo izraziti kao $SST = |\mathbf{Y} - \bar{Y}\mathbf{1}|^2$ Sada možemo primijeniti Pitagorin poučak na vektor $\mathbf{Y} - \bar{Y}\mathbf{1}$, definiciju projektora N i navedena svojstva projektora. Vrijedi

$$\begin{aligned}
SST &= |\mathbf{Y} - \bar{Y}\mathbf{1}|^2 \\
&= |H(\mathbf{Y} - \bar{Y}\mathbf{1})|^2 + |M(\mathbf{Y} - \bar{Y}\mathbf{1})|^2 \\
&= |H\mathbf{Y} - HN\mathbf{Y}|^2 + |M\mathbf{Y} - MN\mathbf{Y}|^2 \\
&= |H\mathbf{Y} - N\mathbf{Y}|^2 + |M\mathbf{Y} - \mathbf{0}|^2 \\
&= |H\mathbf{Y} - \bar{Y}\mathbf{1}|^2 + |M\mathbf{Y}|^2 \\
&= |\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}|^2 + \mathbf{Y}^\tau M\mathbf{Y} \\
&= SSR + SSE,
\end{aligned}$$

gdje zadnja jednakost slijedi iz definicija veličina. □

Sada se može definirati koeficijent determinacije

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}\,. \tag{2.29}$$

S obzirom da ukupan zbroj kvadrata SST mjeri ukupnu varijabilnost izlaznih podataka, a zbroj kvadrata grešaka SSE mjeri preostalu varijabilnost izlaznih podataka nakon što se

uzme u obzir utjecaj regresora, koeficijent R^2 može se shvatiti kao udio rasipanja izlaznih podataka koji se može objasniti funkcijskom vezom ulaznih i izlaznih podataka. Zbog $0 \leq SSE \leq SST$, vrijedi $0 \leq R^2 \leq 1$. Prema [12], vrijednosti koeficijenta R^2 blizu 1 impliciraju da je veliki dio varijabilnosti u y objašnjen regresijskim modelom. Ako dobijemo malu vrijednost za R^2, to znači da velik dio rasipanja izlaznih podataka otpada na rezidualno rasipanje koje nije objašnjeno. U tom slučaju treba razmisliti o promjeni regresijskog modela, iako razlog može biti i slaba koreliranost između ulaznih i izlaznih podataka.
Međutim, velika vrijednost R^2 ne znači nužno da je regresijski model dobar procjenitelj, posebno kada imamo mali broj podataka u odnosu na dimenziju regresijskog modela. Zato se definira korigirani koeficijent determinacije $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}\,. \tag{2.30}$$

Kada se uvede dodatna regresorska varijabla u model, koeficijent determinacije R^2 nikada se ne smanji, bez obzira doprinosi li dodatna varijabla modelu. Zbog toga je teško procijeniti možemo li nešto bitno zaključiti iz povećanja vrijednosti R^2. S druge strane, korigirani koeficijent determinacije $\bar{R}^2$ poveća se uvođenjem dodatne regresorke varijable samo u slučaju kada to uvođenje smanjuje procjenu varijance slučajne greške.

Test značajnosti linearnog regresijskog modela

Test značajnosti regresije je drugi način na koji se može ispitati prikladnost modela. Zapravo želimo odrediti postoji li linearna povezanost izlazne varijable Y i barem neke od ulaznih regresorskih varijabli $x_1, x_2, \ldots, x_k$ (vidi [12]). Dakle, potrebno je testirati hipotezu da regresori nemaju utjecaj na izlaznu varijablu. Postavljamo hipoteze

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \text{ za bar neki j}\,.$$

Prema [14], u regresijskoj analizi uobičajeno je da se ulazne varijable zovu faktori. Nulta hipoteza je da su svi regresijski koeficijenti 0 tj. da ne postoji linearna povezanost izlazne varijable ni s jednim faktorom. Alternativna hipoteza je da je barem jedan regresijski koeficijent različit od 0. Prema tome, odbijanje nulte hipoteze ukazuje da barem neki od

regresora značajno doprinosi modelu.
Kako bi definirali testnu statistiku, uvodimo dodatne oznake. Neka su

$$MSR := \frac{1}{k} SSR, \tag{2.31}$$

$$MSE := \frac{1}{n-k-1} SSE. \tag{2.32}$$

Za testiranje hipoteza koristi se testna statistika F definirana s

$$F := \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}.$$

Sljedeći teorem govori da slučajnoj varijabli F pripada F-razdioba s $(k, n-k-1)$ stupnjeva slobode.

Teorem 2.3.2. *Ako vrijedi*

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_k = 0$$

i $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$*, tada su statistike* SSR *i* SSE *nezavisne i vrijedi*

$$SSR/\sigma^2 \sim \chi^2(k), \tag{2.33}$$

$$SSE/\sigma^2 \sim \chi^2(n-k-1). \tag{2.34}$$

Nadalje,

$$F = \frac{MSR}{MSE} \sim F(k, n-k-1). \tag{2.35}$$

Dokaz teorema može se pronaći u [12].

Cjelokupna procedura testa može se prikazati u tablici analize varijance 2.1 koja se naziva ANOVA tablica.
Pomoću rezultata ovog testa može se odlučiti ima li smisla provođenje daljnjeg testiranja kvalitete regresijskog modela. Naime, ukoliko ne postoji linearna povezanost ciljne varijable i barem neke od prediktora, daljnje istraživanje kvalitete procjene i značajnosti pojedinih regresora nema smisla.

Izvor varijabilnosti	Broj stupnjeva slobode	Zbroj kvadrata odstupanja	Srednje kvadratno odstupanje	F-statistika
model	k	SSR	MSR	F
slučajna pogreška	n-k-1	SSE	MSE	
ukupna varijabilnost	n-1	SST		

Tablica 2.1: ANOVA-tablica

Test značajnosti jedne kovarijate

Nakon što se testom značajnosti linearnog regresijskog modela utvrdi da postoji linearna zavisnost izlazne varijable Y i barem neke od ulaznih regresorskih varijabli, logično je odrediti koji su to regresori koji su značajni u modelu. Dakle, želimo testirati hipotezu da je pojedini regresijski koeficijent jednak nuli, što bi značilo da pripadajuća ulazna varijabla ne utječe na vrijednost izlazne varijable. Hipoteze za provjeru značajnosti pojedinog koeficijenta regresije su

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0\,.$$

U nultoj hipotezi sadržana je tvrdnja da ciljna varijabla ne ovisi o regresoru x_j ($j = 1, \ldots, k$). Stoga, ako nulta hipoteza nije odbijena, regresor x_j može se izbrisati iz regresijskog modela. Odbijanje nulte hipoteze, uz određenu razinu značajnosti, sugerira nam da je regresor x_j značajan u modelu.

Iz jednakosti (2.23) slijedi da $\hat{\beta}_j$ ima normalnu distribuciju $N(\beta_j, \sigma_j^2)$, gdje je $\sigma_j^2 = \sigma^2 a_{jj}$, a a_{jj} j-ti dijagonalni element matrice $(\mathbf{X}^\tau \mathbf{X})^{-1}$. Iz toga slijedi da slučajnoj varijabli $U = \dfrac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}}$ pripada standardna normalna razdioba N(0,1). Iz (2.24) vidimo da slučajnoj varijabli $V = (n-k-1)\dfrac{\hat{\sigma}^2}{\sigma^2}$ pripada $\chi^2(n-k-1)$ distribucija. U i V su nezavisne slučajne varijble pa statistici T definiranoj s

$$T = \frac{U}{\sqrt{V/(n-k-1)}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{a_{jj}}} \tag{2.36}$$

pripada Studentova razdioba ili t-razdioba s $n-k-1$ stupnjeva slobode prema teoremu 5.5-3 iz [4].

Pokazali smo da u slučaju kada je nulta hipoteza istinita, testna statistika definirana s

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{a_{jj}}} \tag{2.37}$$

ima Studentovu razdiobu s $n-k-1$ stupnjeva slobode. Dakle, za danu razinu značajnosti α nultu hipotezu odbacujemo ako vrijedi $|t| > t_{\alpha/2,n-k-1}$ tj. apsolutna vrijednost testne statistike je veća od $(1-\alpha/2)$ kvantila t(n-k-1) distribucije.
Važno je napomenuti da se radi o djelomičnom testu jer regresijski koeficijent β_j ovisi o svim drugim regresorskim varijablama $x_i\,(i \neq j)$ koje su uključene u model (vidi [12]). Stoga, ovo je zapravo test doprinosa varijable x_j s obzirom na druge varijable u modelu.

Test značajnosti podskupa parametara

Osim testom značajnosti jedne kovarijate, doprinos pojedine varijable x_j s obzirom na druge varijable u modelu može se ispitati i koristeći test značajnosti podskupa parametara. Ipak, test značajnosti podskupa parametara najčešće se koristi kako bi se istražio doprinos proizvoljnog podskupa ulaznih varijabli modelu. Prema [17], skup od k regresorskih varijabli treba se podijeliti na dva podskupa od kojih je jedan veličine m, $m < k$, čija je prisutnost u modelu nedvojbena, a jedan veličine $k-m$, za koji se provjerava doprinosi li značajno regresijskom modelu. Potpuni model s k regresora je proširenje manjeg modela s m regresora pa se zapravo provjerava dobije li se bolji model dodavanjem nekih od preostalih $k-m$ varijabli u m-parametarski model. Prikladne hipoteze su

$$H_0 : \beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \text{ za barem neki } j = m+1, \ldots, k\,.$$

Uočimo da nulta hipoteza tvrdi da je model s m regresora dovoljan, a alternativna hipoteza da je potreban potpuni model. Nulta hipoteza može se testirati koristeći statistiku

$$F = \frac{(SSR_{(m)} - SSR_{(k)})/(k-m)}{(SSR_{(k)})/(n-(k+1))}\,. \tag{2.38}$$

$SSR_{(m)}$ je regresijski zbroj kvadrata za reducirani model s m regresora $x_1, x_2, \ldots, x_m$, a $SSR_{(k)}$ je regresijski zbroj kvadrata za potpuni model s k regresora. Razlika $SSR_{(m)} - SSR_{(k)}$ je pozitivna jer regresijski zbroj kvadrata opada s povećanjem broja regresijskih varijabli.

Testna statistika, tj. empirijski F-omjer ima F(k-m, n-(k+1)) distribuciju. Dakle, za danu razinu značajnosti α, nulta se hipoteza odbacuje ako vrijedi $F > F_{\alpha,k-m,n-(k+1)}$, što implicira da je barem jedna od varijabli $x_{m+1}, x_{m+2}, \ldots, x_k$ značajna u modelu. Ovaj test uobičajeno se naziva parcijalni F-test jer mjeri doprinos podskupa regresora kada su ostali regresori već uključeni u model.

U slučaju $m = k - 1$, istražujemo doprinos modelu jedne ulazne varijable x_j kada je preostalih $k - 1$ varijabli $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ već u modelu. Korisno je razmišljati o tome kao ispitivanje doprinosa regresora x_j ako ga zadnjeg dodamo u model. Može se pokazati da je parcijalni F-test za jednu ulaznu varijablu ekvivalentan t-statistici definiranoj s (2.36) u prethodnom pododjeljku (za detalje pogledati u [12]). Ipak, parcijalni F-test je oćenitiji jer može ispitati značajnost proizvoljnog podskupa regresora.

Važna primjena parcijalnog F-testa je izgradnja najboljeg modela višeparametarske linerarne regresije, odnosno biranje najboljeg skupa regresora za model. Više o tome može se pronaći u [12].

2.4 Pouzdani intervali

U ovom odjeljku želimo odrediti pouzdane intervale za procjenitelje regresijskih koeficijenta i za očekivanje ciljne varijable. Širinu intervala pouzdanja možemo gledati i kao mjeru kvalitete regresijskog modela.

Vrijede pretpostavke kao u prethodnom odjeljku: slučajni vektor $\boldsymbol{\varepsilon}$ prati normalnu distribuciju kojoj je n-dimenzionalni nul vektor $\mathbf{0}$ vektor očekivanja i $\sigma^2\mathbf{I}_n$ kovarijacijska matrica.

Pouzdani intervali za nepoznate koeficijente

Nakon što smo odredili točkovne procjenitelje za regresijske koeficijente, želimo im pridružiti i pouzdane intervale za određenu razinu značajnosti α. Za konstrukciju pouzdanih intervala koristimo već definiranu statistiku (2.36) koja slijedi Studentovu distribuciju s $n - k - 1$ stupnjeva slobode.

Dakle, uz oznaku

$$\hat{\sigma}_j = \sqrt{\hat{\sigma}^2 a_{jj}}$$

za procjenu standardne devijacije regresijskog koeficijenta $\hat{\beta}_j$, $(1-\alpha)\cdot 100\%$ pouzdani interval za nepoznati parametar β_j, $j = 0, 1, \ldots, k$, je

$$\left[\hat{\beta}_j - t_{\alpha/2,n-p}\hat{\sigma}_j,\ \hat{\beta}_j + t_{\alpha/2,n-p}\hat{\sigma}_j\right].$$

Pouzdan interval za očekivanje ciljne varijable

Jedna od glavnih uloga linearne regresije je procjena srednje vrijednosti $E[Y]$ za konkretnu vrijednost prediktorskih varijabli $x_{01}, x_{02}, \ldots, x_{0k}$. Osim točkovne procjene, možemo konstruirati intervalnu procjenu za očekivanje ciljne varijable u određenoj točki. Definirajmo vektor vrijednosti prediktorskih varijabli $\mathbf{x_0}$ s

$$\mathbf{x_0} = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}.$$

Dakle, želimo procijeniti vrijednost $E[Y|\mathbf{x_0}]$. Procjena vrijednosti izlazne varijable u točki $\mathbf{x_0}$ jednaka je

$$\hat{Y}_0 = \mathbf{x_0}^\tau \hat{\boldsymbol{\beta}}.$$

$\hat{Y}_0$ je napristran procjenitelj za $E[Y|\mathbf{x_0}]$ jer vrijedi $E[\hat{Y}_0] = \mathbf{x_0}\boldsymbol{\beta} = E[Y|\mathbf{x_0}]$. Uočimo da je procjena očekivane vrijednosti od Y jednaka procjeni iznosa mjerena Y za dani $\mathbf{x} = \mathbf{x_0}$. Varijanca procjenitelja $\hat{Y}_0$ jednaka je $\mathrm{Var}(\hat{Y}_0) = \sigma^2\mathbf{x_0}^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{x_0}$ i procijenitelj je linear. Uz pretpostavku o normalnoj distribuciji slučajne greške, slijedi da je slučajna varijabla $\hat{Y}_0$ normalno distribuirana s očekivanjem $E[Y|\mathbf{x_0}]$ i varijancom $\sigma^2\mathbf{x_0}^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{x_0}$. Može se pokazati da statistika

$$\frac{\hat{Y}_0 - E[Y|\mathbf{x_0}]}{\sqrt{\hat{\sigma}^2\mathbf{x_0}^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{x_0}}} \tag{2.39}$$

slijedi t-distribuciju s $n-k-1$ stupnjeva slobode (vidi [7]) iz čega proizlazi da je $(1-\alpha)\cdot 100\%$ pouzdani interval za $E[Y|x_0]$, tj. za srednju vrijednost izlazne varijable u točki $x_{01}, x_{02}, \ldots, x_{0k}$ jednak

$$\left[\hat{Y}_0 - t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2\mathbf{x_0}^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{x_0}},\ \hat{Y}_0 + t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2\mathbf{x_0}^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{x_0}}\right].$$

Pouzdan interval za predviđanje novih opažanja

Linearna regresija najčešće se koristi kako bi se za određenu vrijednost prediktorskih varijabli predvidjela vrijednost izlazne varijable. S obzirom da na temelju slučajnog uzorka ne možemo znati koja bi bila stvarna vrijednost izlazne varijable za dane regresore, korisno je konstruirati pouzdane intervale. Napomenimo da je sada cilj razviti intervalnu procjenu za buduća opažanja, dok smo u prethodnom pododjeljku dobili intervalnu procjenu za srednju vrijednost budućih opažanja. Neka je $x_{01}, x_{02}, \ldots, x_{0k}$ konkretna vrijednost prediktorskih varijabli te $\hat{Y}_0$ točkovni procjenitelj za buduće opažanje Y_0 u točki $\mathbf{x_0}$ definiranoj s

$$\mathbf{x_0} = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}.$$

Točkovni procjenitelj $\hat{Y}_0$ jednak je točkovnom procjenitelju za očekivanu vrijednost od Y uz dano $x = x_0$: $\hat{Y}_0 = \mathbf{x_0}^\tau \hat{\boldsymbol{\beta}}$.

Neka je $Y_0 = \mathbf{x_0}^\tau \beta + \varepsilon_0 \sim N(\mathbf{x_0}^\tau \beta, \sigma^2)$ varijabla odziva koju želimo predvidjeti. Može se pokazati da statistika

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2(1 + \mathbf{x_0}^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{x_0})}} \tag{2.40}$$

slijedi t-distribuciju s $n - k - 1$ stupnjeva slobode (vidi [7]) iz čega proizlazi da je $(1 - \alpha) \cdot 100\%$ pouzdani interval za vrijednost izlazne varijable Y_0 u točki $x_{01}, x_{02}, \ldots, x_{0k}$ jednak

$$\left[\hat{Y}_0 - t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x_0}^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{x_0}\right)},\ \hat{Y}_0 + t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x_0}^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{x_0}\right)}\right].$$

2.5 Nelinearna regresija

Koncepti višeparametarske linearne regresije, koje smo obrađivali u prethodnim potpoglavljima, predstavljaju osnovu za shvaćanje složenijih modela koje ćemo ukratko razmotriti. Višeparametarska linearna regresija nosi naziv linearna jer je linearna u parametrima, to jest parametri u modelu imaju potenciju jednaku jedan. Do sada smo proučavali model

koji je linearan u parametrima i u ulaznim varijablama. Ponekad dobiveni niz podataka sugerira da izlazna varijabla ovisi o ulaznim varijablama, ali ne linearno. Prema [17], veliki broj modela linearnih u parametrima, a nelinearnih u ulaznim varijablama mogu se prikladnom transformacijom svesti na model linearan u parametrima i ulaznim varijablama te potom koristiti dobivene rezultate iz prethodnih potpoglavlja.

Opći multiplikativni regresijski model s k prediktora je

$$Y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} + \ldots + x_k^{\beta_k} \varepsilon \,. \tag{2.41}$$

Model je linearan u parametrima i nelinearan u ulaznim varijablama. Dakle, može se shvatiti kao model linearne regresije. Transformirajmo model u linearni logaritmiranjem:

$$lnY = ln\beta_0 + \beta_1 lnx_1 + \beta_2 lnx_2 + \ldots + \beta_k lnx_k + ln\varepsilon.$$

Na linearizirani model možemo primijeniti rezultate iz prethodnih potpoglavlja, s tim da umjesto originalnih vrijednosti koristimo logaritmirane vrijednosti zavisne i nezavisnih varijabli. U ovom modelu dobivene regresijske koeficijente $\hat{\beta}_j$ interpretiramo kao srednja vrijednost postotka promjene zavisne varijable Y ako se ulazna varijabla x_j poveća za 1%, a preostale varijable ostanu nepromijenjene.

Drugi primjer koji ćemo spomenuti je linearni regresijski polinom u kojem se pretpostavlja da je veza između zavisne varijable Y i jedne nezavisne varijable x polinom stupnja n u x:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon \,. \tag{2.42}$$

Ovaj model je također linear u parametrima i nelinearan u ulaznoj varijabli pa se uz supstituciju

$$x_1 = x \,,\; x_2 = x^2 \,, \ldots \,,\; x_n = x^n$$

može shvatiti i kao model n-dimenzionalne linearne regresije:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon. \tag{2.43}$$

Modeli koji nisu linearni u parametrima ili se ne mogu svesti na linearni regresijski model analiziraju se drugim metodama.

2.6 Indikatorske varijable

Pri razvijanju modela višeparametrske linearne regresije o varijablama modela najčešće razmišljamo kao o kvantitativnim varijablama. No, primjene regresije u stvarnim problemima često zahtijevaju prisutnost kvalitativnih varijabli u modelu. Kvalitativne varijable često se nazivaju i kategoričke varijable jer opisuju karakteristike koje su povezane s kategorijama, naprimjer spol, status zaposlenosti, obrazovanje i slično. Kvalitativne varijable se mogu pojaviti i kao nezavisne i kao zavisne varijable u modelu. U slučaju kada su kvalitativne varijable uključene u model kao nezavisne varijable, model analiziramo na isti način kao standardni regresijski model.

Kategoričke varijable uključujemo u model koristeći binarne (indikator) varijable koje poprimaju vrijednost 1 ili 0. Pretpostavimo da želimo otkriti vezu između promjene u ocjeni kvalitete života (Y), početne razine hemoglobina (x_1) i primanju terapije. U ovom primjeru primanje terapije je kategorička varijabla koja opisuje dvije mogućnosti: pacijent je primio terapiju i pacijent nije primio terapiju. Dakle, u regresijski model trebamo uključiti indikatorsku varijablu x_2 koja poprima sljedeće vrijednosti

$$x_2 = \begin{cases} 0 \text{ ako pacijent nije primio terapiju} \\ 1 \text{ ako je pacijent primio terapiju} \end{cases}$$

Svejedno je kojem modalitetu ćemo pridružiti vrijednost 0, odnosno 1. Imamo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon\,.$$

Zanima nas koja je interpretacija koeficijenta uz indikatorsku varijablu. U slučaju kada pacijent nije primio terapiju, regresijski model postaje

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon\,,$$

a u slučaju kada je pacijent primio terapiju, regresijski model postaje

$$Y = \beta_0 + \beta_2 + \beta_1 x_1 + \varepsilon\,.$$

Dakle, koeficijent β_2 predstavlja razliku u očekivanoj promjeni u ocjeni kvalitete života koja je proizašla iz toga što je pacijent primio terapiju.

Ovaj pristup se može poopćiti za bilo koji broj modaliteta kategoričke varijable. Ako kategorička varijabla koju želimo uključiti u model ima m modaliteta, broj indikator varijabli koji moramo uključiti u model jednak je $m - 1$. Naprimjer, u slučaju kada imamo jednu kvalitativnu varijablu A s 4 modaliteta i jednu kvalitativnu varijablu B s 2 modaliteta, uvodimo indikatorske varijable

$$A_i = \begin{cases} 1 \text{ ako varijabla A poprima } i\text{-ti modalitet} \\ 0 \text{ inače} \end{cases} \qquad i = 1, 2, 3, 4$$

$$B_i = \begin{cases} 1 \text{ ako varijabla B poprima } i\text{-ti modalitet} \\ 0 \text{ inače} \end{cases} \qquad i = 1, 2$$

i zapišimo model:

$$Y = \beta_0 + \beta_1 A_2 + \beta_2 A_3 + \beta_3 A_4 + \beta_4 B_2 + \varepsilon.$$

U ovom slučaju, naprimjer, koeficijent β_1 predstavlja efekt na Y koji doprinosi drugi modalitet u odnosu na prvi (bazni) modalitet od A.

Poglavlje 3

Utjecaj djelatne tvari epoetin alfa na kvalitetu života

U ovom poglavlju primijenit ćemo teorijske rezultate o višeparametarskoj linearnoj regresiji u istraživanju utjecaja epoetina alfa na kvalitetu života pacijenata s malignim bolestima. Posebno ćemo analizirati vezu između razine hemoglobina i kvalitete života. Opisano istraživanje i rezultati u ovom poglavlju temelje se na članku [1] pa to neće biti dalje naglašavano.

Veliki broj ljudi koji boluju od malignih tumora također boluje i od anemije. Anemija povezana s rakom uzrokuje razne simptome koji negativno utječu na kvalitetu života. Djelatna tvar epoetin alfa pokazala je učinkovitost u liječenju anemije kod onkoloških bolesnika. Kako bi procijenili utjecaj epoetina alfa na standard života, provedeno je ispitivanje u kojem je sudjelovalo 375 anemičnih pacijenata koji boluju od raka i primaju kemoterapiju bez platine. Cilj istraživanja je utvrditi kako epoetin alfa utječe na kvalitetu života kod pacijenata koji primaju kemoterapiju bez platine. Višeparametarskom linearnom regresijom podataka iz randomiziranog, dvostruko slijepog, placebom kontroliranog ispitivanja analizirana je dobrobit epoetina alfa na životni standard kod opisanih pacijenata. Apriorno planirana linearna regresijska analiza obuhvaćala je učinke progresije bolesti na kvalitetu života i nekoliko drugih mogućih zbunjujućih varijabli na kvalitetu života.

3.1 Motivacija za provođenje ispitivanja

Rak je veliki ekonomski i javnozdravstveni problem i očekuje se da će njegov teret još više rasti. Pored same bolesti, kvaliteta života predstavlja jedan od najznačajnijih izazova za osobe koje se suočavaju s onkološkim oboljenjem. Uz bol i mučninu, umor je najčešća i najviše zabrinjavajuća nuspojava primanja kemoterapije. Iako umor značajno utječe na obavljanje svakodnevnih aktivnosti, često je zanemaren i nedovoljno liječen kod onkoloških pacijenata. Ankete među oboljelim pacijentima izvješćuju o visokoj stopi prevalencije umora, pri čemu značajan postotak onih koji su završili liječenje i dalje pati od umora, što ukazuje na čestu pojavu umora u svim fazama bolesti, uključujući izliječene pacijente. Uzroci umora kod onkoloških pacijenata mogu biti različitog podrijetla, a najčešće su to nepovoljni učinci liječenja i osnovna bolest.

Anemija je stanje organizma uzrokovano smanjenim brojem crvenih krvnih stanica (eritrocita) ili smanjenom količinom hemoglobina u eritrocitima, koji prenosi kisik do tkiva. Pacijenti koji boluju od anemije najčešće se žale na dugotrajnu malaksalost, iscrpljenost, vrtoglavicu, glavobolju te brzo umaranje. Anemija se često javlja kod onkoloških pacijenata te je značajno povezana s kvalitetom života pa je često potrebna adekvatna terapija.

Prema [19], Epoetin alfa je djelatna tvar sadržana u lijeku za liječenje anemije u odraslih osoba koje primaju kemoterapiju za određene tipove raka i za smanjenje potrebe za transfuzijama krvi. Kopija je hormona naziva eritropoetin i djeluje na potpuno isti način kao prirodni hormon za stimulaciju proizvodnje crvenih krvnih stanica iz leđne moždine. Kod bolesnika koji primaju kemoterapiju anemiju može uzrokovati manjak eritropoetina ili organizam koji ne reagira adekvatno na eritropoetin koji se prirodno nalazi u tijelu. U tim slučajevima epoetin alfa koristi se za povećanje broja eritrocita. U nekoliko već objavljenih ispitivanja rekombinirani eritropoetin povećao je razinu hemoglobina u krvi i posljedično poboljšao anemiju kod oboljelih od raka. Štoviše, u studijima koji su uključivali procjenu kvalitete života, epoetin alfa pružio je poboljšanje energije, raspoloženja i cjelokupne kvalitete života. Također je pokazano da epoetin alfa pozitivno djeluje i na druge tjelesne funkcije, uključujući moždanu i kognitivnu funkciju.

Kako bi se precizno procijenili učinci liječenja epoetinom alfa na kvalitetu života uzete su u obzir potencijalne razlike između različitih skupina u čimbenicima koji bi mogli

utjecati na tu procjenu. Podaci o kvaliteti života ispitani su primjenom unaprijed planirane višestruke linearne regresijske analize, pri čemu su uzete u obzir moguće zbunjujuće varijable poput napredovanja bolesti, početnih kliničkih karakteristika i demografskih čimbenika. Ova analitička metoda omogućila je bolje razumijevanje stvarnih učinaka epoetina alfa na kvalitetu života, eliminirajući potencijalni utjecaj drugih faktora koji bi mogli iskriviti rezultate.

3.2 Pacijenti i metode

Općenito o pacijentima i ispitivanju

Randomizirano, dvostruko slijepo, placebom kontrolirano ispitivanje provedeno je u 15 zemalja na 73 mjesta s 375 pacijenata. Sudionici u ispitivanju dio su populacije za analizu prema namjeri za liječenje (ITT, eng. intent-to-treat) koja predstavlja skupinu u kliničkom istraživanju koja se analizira prema njihovoj prvotnoj randomizaciji i planiranim tretmanima, bez obzira na to jesu li zapravo primili određeni tretman ili napustili ispitivanje prije njegova završetka. Populacija prema namjeri za liječenje je važna kako bi se očuvala realnost kliničkog istraživanja jer u stvarnom svijetu pacijenti mogu napustiti tretman ili ne slijediti ga ispravno. U analiziranju kvalitete života sudjelovali su oni pacijenti iz populacije prema namjeri za liječenje koji su imali početnu procjenu i bar jednu naknadnu procjenu kvalitete života. Postojali su određeni zahtjevi koji su pacijenti morali ispunjavati kako bi bili uključeni u ovu analizu. Pacijenti su morali imati potvrđenu dijagnozu solidne zloćudne bolesti ili limfocitne leukemije za koju su primali ili nisu još primili kemoterapiju bez platine, ali bili su zakazani za primanje još 3 do 6 ciklusa takve terapije. Epoetin alfa je već u upotrebi širom svijeta kod pacijenata koji primaju kemoterapiju baziranu na platini pa oni nisu sudjelovali u analizi. Pacijenti oboljeli od akutne leukemije nisu sudjelovali u ispitivanju. Dodatno, pacijenti su otprilike tjedan dana prije randomizacije probrani prema razini hemoglobina u krvi. Samo oni pacijenti koji su imali razinu hemoglobina manju od 10,5 grama po decilitru krvi ili su imali razinu hemoglobina manju od 12 grama po decilitru krvi, a razina se smanjivala za 1,5 gram po decilitru krvi po mjesecu od početka primanja kemoterapije. Također su bili isključeni pacijenti koji su imali neliječeni nedostatak željeza, folata ili vitamina B12. Još neki od razloga zbog kojih određeni pacijenti nisu

sudjelovali u analizi su velika infekcija ili krvarenju u zadnjih mjesec dana, radioterapija ili alogena transfuzija krvi u zadnja dva tjedna, ili operacija ili teška bolest u zadnjih tjedan dana prije početka ispitivanja. Svi pacijenti dali su pismeni pristanak prije bilo kojeg postupka povezanog s ispitivanjem. Stratificirani slučajni uzorak čine skupine slučajno izabranih elemenata pojedinih podskupova osnovnoga skupa formiranih na temelju određenih faktora koji su značajni za istraživanje. Pacijenti su pri probiru stratificirani prema tipu tumora (solidni ili hematološki) te razini hemoglobina (manje od 10 g/dL ili između 10 g/dL i 12 g/dL). Randomizirani su koristeći permutirane blokove u omjeru 2:1 za primanje epoetina alfa 150-300 IU/kg triput tjedno ili placeba putem potkožne injekcije. Takvo liječenje primijenjeno je tijekom trajanja ispitivanja koje je moglo potrajati maksimalno 28 tjedana i uključivalo je 12 do 24 tjedna tijekom kojih su pacijenti primali kemoterapiju (3 do 6 ciklusa) te još 4 tjedna nakon posljednje doze kemoterapije.

Procjena kvalitete života

Hipoteza o kvaliteti života u ovom ispitivanju je da bi primjena primjena epoetina alfa poboljšala kvalitetu života u odnosu na placebo zbog poboljšanja razine hemoglobina uzrokovana liječenjem. Promjena razine hemoglobina od početne do zadnje dostupne vrijednosti je ključni krajnji rezultat koji se mjeri u ovom kliničkom istraživanju kako bi se utvrdila efikasnost tretmana. Promjena u kvaliteti života od početne do zadnje dostupne procjene je rezultat koji se također koristi za razumijevanje učinka tretmana epoetinom alfa.

Kvaliteta života onkoloških pacijenata procjenjivana je tijekom liječenja koristeći tri različita instrumenata za samoprocjenu pacijenta. Jedan od njih je upitnik Funkcionalna procjena terapije raka – anemija (engl. Functional Assessment of Cancer Therapy-Anaemia, kraticom FACT-An) koji je specifičan za osobe oboljele od raka. Sastoji se od 47 stavki, od kojih je 27 stavki iz općenite ljestvice Funkcionalna procjene terapije raka (engl. Functional Assessment of Cancer Therapy-General, kraticom FACT-G Total) te 20 stavki povezanih s anemijom, od kojih je 13 iz podljestvice koja procjenjuje utjecaj umora (FACT-An Fatigue Subscale) i 7 iz podljestvice koja procjenjuje utjecaj anemije (FACT-An Anemia Subscale). Drugi korišteni upitnik specifičan za procjenu kvalitete života onkoloških bolesnika je Linearna analogna ljestvica raka (engl. the Cancer Linear Analogue Scale, kraticom CLAS). Upitnik obuhvaća 3 ljestvice koje mjere energiju, sposobnost za obavljanje

dnevnih aktivnosti i svekupnu kvalitetu života. Treći upitnik koji su pacijenti u ispitivanju također ispunjavali je Upitnik SF 36 (engl. Short Form Health Survey-36) sastavljen od 36 općenitih pitanja za procjenu zdravljem uvjetovane kvalitete života. Navedeni instrument procjene daje i sažete rezultate za dvije izvedene mjere: sažetak fizičke komponente (engl. Physical Component Summary, kraticom PCS) i sažetak mentalne komponente (engl. Mental Component Summary, kraticom MCS). U svakom upitniku viši rezultati označavaju bolju kvalitetu života. Pacijenti su sami ispunili navedena 3 upitnika najviše 4 puta tijekom trajanja studija: prije početka primanja tretmana, odnosno placeba, prije početka drugog ciklusa kemoterapije ili otprilike u 4. tjednu, prije početka četvrtog ciklusa kemoterapije ili otprilike u 16. tjednu, te u roku od 5 dana od završetka studija. Podsjetimo se, u ispitivanje su bili uključeni oni pacijenti kojima je predviđeno liječenje uključivalo (dodatnih) 3 do 6 ciklusa kemoterapije. U slučaju promjene plana liječenja zbog zdravstvenih razloga, pacijenti su zamoljeni završiti procjenu predviđenu za završetak studija. Zbog različitog broja predviđenih ciklusa kemoterapije te različite duljine trajanja ciklusa vrijeme kada pacijenti ispunjavaju upitnik nije moglo biti predodređeno za točan dan tijekom trajanja studija.

Statističke metode

Promjena u ocjeni kvalitete života izračunata je oduzimanjem početnog rezultata od rezultata iz zadnje dostupne procjene. Dobiveni podaci ispitani su višestrukom linearnom regresijom. Kako bi dobili što jasniji utjecaj liječenja epoetinom alfa, u svakom modelu uzeli smo u obzir progresiju bolesti i druge moguće zbunjujuće čimbenike koji bi mogle utjecati na procjenu kvalitete života.

Za svaku od 7 skala za samoprocjenu kvalitete života navedenih u prethodnom pododjeljku provedena je zasebna višestruka linearna regresija. U svim modelima, zavisna varijabla bila je promjena u ocjeni kvalitete života. Nezavisne varijable uključivale su liječenje epoetinom alfa, napredovanje bolesti, interakcija liječenja i progresije bolesti, te nekoliko izabranih demografskih i kliničkih varijabli (dob, spol, rasa, početna razina hemoglobina, početni broj neutrofila, početni broj retikulocita, početna razina eritropoetina, zavisnost o transfuziji prije početka ispitivanja, tip tumora). U svaki model uključene su varijable koje predstavljaju liječenje epoetinom alfa, progresiju bolesti i interakciju liječenja i progresije

bolesti. Sve druge nezavisne varijable zadržane su u modelu ako su bile značajne na razini od 10% ili manje. Rezultati liječenja kemoterapijom prikupljeni su prilikom završetku ispitivanja ili prijevremenog povlačenja iz ispitivanja. Iz tih rezultata konstruirana je vrijednost varijable napredovanja bolesti: u slučaju kada se tumorska masa povećala za više od 25% ili se pojavila nova lezija, vrijednost varijable je 1, a inače je vrijednost 0. Za svaku skupinu za liječenje procijenjena je srednja vrijednost promjene kvalitete života metodom najmanjih kvadrata. Zatim su generirane T-statistike kako bi se testirale razlike u prosječnim promjenama između skupina za liječenje. Svi testovi bili su dvostrani, s razinom značajnosti od 0.05, prilagođavajući se za višestruke usporedbe korištenjem verzije Bonferroni postupka (vidi [5]).

Analiza procjene srednje vrijednosti metodom najmanjih kvadrata provedena je na cijeloj populaciji prilagođenoj za sve kovarijate, uključujući progresiju bolesti. S obzirom na to da je faktor napredovanje bolesti uključen u regresijski model, istu analizu možemo primijeniti za dvije grupe pacijenata koje se razliku po pokazatelju napredovanja bolesti. Treba naglasiti da je procjena srednje vrijednosti promjene kvalitete života kod pacijenata kojima bolest napreduje zapravo proizašla iz regresijskih parametara izračunatih iz cijele populacije. Time smo dobili predviđanje koliko epoetin alfa koristi u kvaliteti života ovisno o pokazatelju napretka bolesti, odnosno za pacijenta koji će doživjeti napredak bolesti i za pacijenta koji neće doživjeti napredak bolesti. U stvarnosti se ne može sa sigurnošću predvidjeti kako će bolest napredovati, no zbog toga što napredak bolesti znatno smanjuje kvalitetu života, eksplicitno smo uključili tu varijablu u model.

Zbog pretpostavke da povećanje razine hemoglobina uzrokovane primjenom djelatne tvari epoetin alfa pozitivno utječe na kvalitetu života, izračunati su Pearsonovi koeficijenti korelacije između početnih vrijednosti razine hemoglobina i procijenjene kvalitete života (poprečna korelacija), te između promjene u razini hemoglobina i promjene u ocjeni kvalitete života (longitudinalna korelacija). P vrijednosti povezane s ovim korelacijskim koeficijentima također su prilagođene za višestruke usporedbe korištenjem Bonferronijevog postupka.

Ljestvica	Postotak stavki koje nedostaju	Postotak ljestvica koje nedostaju
FACT-G Total	3.87	3.52
FACT-An Fatigue subscale	1.58	1.14
CLAS: Energy	0.00	0.00
CLAS: Daily activities	0.09	0.09
CLAS: Overall QoL	0.18	0.18
SF-36 PCS	1.86	5.30
SF-36 MCS	1.86	5.30
Ukupno	2.45	2.09

Tablica 3.1: Postotak nedostajućih stavki i ljestvica u procjenama kvalitete života

3.3 Rezultati

U ovom ispitivanju zabilježen je minimalan broj podataka koji su nedostajali zbog administrativnih ili logističkih razloga. Naprimjer, od očekivanih 1151 procjena kvalitete života za upitnik CLAS, bilo ih je dostupno 1076. Drugim riječima, boljom organizacijom ispitivanja moglo se spriječiti nedostajanje 6,5% očekivanih procjena koristeći CLAS upitnik. Ovo predstavlja vrlo nisku stopu izgubljenih podataka koji su mogli biti izbjegnuti u kontekstu kliničkih ispitivanja u onkologiji. Jedan od razloga izgubljenih podataka je što određeni upitnici nisu dostupni na nekim jezicima. Naprimjer, 54 pacijenta nisu imali ispunjene upitnike FACT ni SF-36 jer nisu bili dostupni na njihovom jeziku. Sveukupno je sudjelovalo 298 od 375 pacijenata početne populacije u analizi FACT rezultata, odnosno 336 pacijenata u analizi CLAS rezultata.

S obzirom na završetak procjene, postotak stavki koji nedostaju bio je malen. Podaci o nedostajućim stavkama navedeni su u tablici 3.1. Za sedam izabranih ljestvica za procjenu kvalitete života, prosječno je nedostajalo 2.45% stavki. U tablici 3.1 također je prikazan postotak ljestvica koje nedostaju. Razlog tome su ljestvice FACT i CLAS koje se sastoje od više stavki pa njihov rezultat zahtijeva potreban broj stavki, uobičajeno bar 50%.

Početne karakteristike

Početne karakteristike populacije u istraživanju po kovarijatama dane su u tablici 3.2. Populacija je podijeljena u dvije skupine od kojih jedna skupina prima tretman, a druga sku-

pina prima placebo. Iz tablice vidimo da su početne karakteristike za te dvije skupine slične.

Promjene u kvaliteti života

Zasebna analiza provedena je za svaki od sedam krajnjih ishoda koje proučavamo: razlike u kvaliteti života tijekom kemoterapije dobivene koristeći sedam različitih ljestvica za samoprocjenu. Dobivene vrijednosti parametara u modelu višeparametarske linearne regresije u kojem je zavisna varijabla promjena vrijednosti kvalitete života prikazane su u tablici 3.3. U tablici 3.4 prikazane su procjene prosječne promjene kvalitete života za cjelokupnu populaciju i po napredovanju bolesti.

Ova analiza potvrđuje da liječenje epoetinom alfa pozitivno djeluje na kvalitetu života tijekom primanja kemoterapije. Za grupu pacijenata koja je primala lijek, srednje vrijednosti promjene kvalitete života bile su pozitivne za sve korištene ljestvice. Suprotno tome, srednje vrijednosti promjene kvalitete života bile su negativne za grupu pacijenata koja je primala placebo za sve ljestvice osim SF-36 MCS i PCS. Te srednje vrijednosti za dvije grupe pacijenata uspoređene su na cjelokupnoj populaciji te su pronađene statistički značajne razlike koje podržavaju upotrebu epoetina alfa u svih pet ljestvica specifičnih za rak. Za preostale dvije ljestvice, izvedene iz SF-36 upitnika, koje nisu pokazivale značajne razlike u kvaliteti života između placebo grupe i grupe koja je primala lijek, nije pokazan ni negativan učinak lijeka. SF-36 upitnik je općeniti upitnik koji je uključen u ispitivanje dijelom zbog toga kako bismo osigurali da ne postoje očekivane nepoželjne posljedice koje utječu na ukupnu kvalitetu života.

Rezultati regresijskog modela govore nam da onkološki pacijent koji tijekom kemoterapije ne pokazuje napredovanje bolesti može očekivati poboljšanje kvalitete života mjerene prema FACT i CLAS upitniku ako se liječi epoetinom alfa. Također, u slučaju da pacijent pokazuje pogoršanje bolesti, prema ovom istraživanju nema razloga smatrati da bi liječenje epoetinom alfa poboljšalo njegovu kvalitetu života. Navedene rezultate možemo koristiti kao potvrdu za dobro izabrane instrumente procjene kvalitete života u ovom ispitivanju. Ako onkološki pacijent pokazuje značajno pogoršanje kvalitete života, očekivali bi da je to rezultat napretka bolesti, bez obzira na liječenje anemije povezane s karcinomom.

Za broj bodova koji predstavlja razliku u očekivanim promjenama kvalitete života

Kovarijata	**Epoetin alfa (n=238)**	**Placebo (n=111)**
Spol, n (%)		
žena	158 (66)	77 (69)
muškarac	80 (34)	34 (31)
Dob (godine)		
srednja vrijednost	58.1	59.2
interval	18.7-84.9	21.1-88.6
Rasa, n (%)		
bijelac	230 (97)	106 (96)
Indijac	1 (<1)	3 (3)
crnac	4 (2)	0 (0)
ostalo	3 (1)	2 (2)
Hemoglobin (g/dl)		
srednja vrijednost	9.9	9.8
medijan	10.0	9.7
interval	5.9-14.3	6.6-12.7
Vrsta tumora, n (%)		
solidni	127 (53)	60 (54)
leukemija	111 (47)	51 (46)
Transfuzija prije ispitivanja, n(%)	68 (29)	38 (34)
Broj neutrofila, %		
srednja vrijednost	61.5	62.1
medijan	66.0	67.4
interval	1.0-94.5	2.0-96.0
Broj retikulocita, %		
srednja vrijednost	2.2	2.4
medijan	2.0	2.1
interval	0.0-7.5	0.0-11.5
Hormon eritropoetin, mU/mL		
srednja vrijednost	102.9	94.1
medijan	49.0	50.0
interval	10-1890	10-597

Tablica 3.2: Početne karakteristike populacije po kovarijatama

	Procjene koeficijenata mjera kvalitete života (standardna devijacija)						
Kovarijate	**FACT-G Total**	**FACT-An Fatigue**	**Energy**	**Activities**	**Overall**	**PCS**	**MCS**
Veličina uzorka	n=253	n=266	n=312	n=284	n=311	n=237	n=225
Epoetin alfa	n=175	n=185	n=215	n=196	n=215	n=160	n=153
Placebo	n=78	n=81	n=97	n=88	n=96	n=77	n=72
R^2	0.23	0.19	0.15	0.20	0.17	0.25	0.11
Konstanta	6.37 (8.42)	7.31 (7.94)	-3.72 (16.10)	72.98^b (21.21)	-0.84 (15.76)	-0.32 (3.46)	20.79^b (7.176
Epoetin alfa	4.58^a (2.13)	6.64^b (1.82)	15.05^b (4.07)	13.95^b (4.65)	14.50^b (4.01)	2.92^a (1.27)	2.03 (1.87)
Progresija bolesti	-9.57^b (3.14)	-3.05 (2.68)	-7.68 (5.78)	-12.09 (6.47)	-4.95 (5.67)	-3.33 (1.82)	-4.13 (2.65)
Epoetin alfa * progresija bolesti	-0.30 (3.80)	-8.00^a (3.28)	-15.50^b (7.16)	-14.44 (8.03)	-20.61^b (7.02)	-5.97^b (2.27)	-1.77 (3.33)
Bijela rasa	12.88^b (4.00)	10.66^b (3.50)	23.68^b (8.02)	17.23 (8.97)	29.04^b (7.85)	9.84^b (2.35)	-
Dob	-	-0.09 (0.05)	-	-0.24 (0.12)	-	-0.10^b (0.04)	-
Žena	-	-	-	-	-	-1.89 (1.08)	-
Početni hemoglobin	-1.28 (0.71)	-1.29^a (0.62)	-2.05 (1.37)	-5.64^b (1.61)	-3.01^a (1.34)	-	-1.63^b (0.69)
Početni broj neutrofila	-	-	-	-0.26^b (0.09)	-	-	-
Početni broj retikulocita	-1.18^a (0.59)	-	-	-	-	-0.73^a (0.36)	-
Početni endogeni eritropoetin	-	-	-	-0.02^a (0.01)	-	-	-0.01^a (0.005)
Transfuzija prije ispitivanja	-	-	-	-	-	2.62^a (1.06)	-
Solidni tumor	-5.71^b (1.68)	-	-	-	-	-	-3.76^a (1.46)

Tablica 3.3: Procjene parametara višeparametarske linearne regresije za promjene vrijednosti kvalitete života

	Epoetin alfa				Placebo				Razlika			
Ljestvica	**Srednja vrijednost**	**Stand. devijacija**	**P vrijednost**	**Prilagođena P vrijednost**	**Srednja vrijednost**	**Stand. devijacija**	**P vrijednost**	**Prilagođena P vrijednost**	**Srednja vrijednost**	**Stand. devijacija**	**P vrijednost**	**Prilagođena P vrijednost**
Sveukupno												
FACT-G Total	2.01	0.98	0.04	0.04	-2.48	1.47	0.09	0.66	4.49	1.77	0.01	0.04
FACT-An Fatigue	2.70	0.84	<0.01	<0.01	-1.70	1.27	0.18	0.90	4.40	1.52	<0.01	0.02
CLAS: Energy	7.17	1.86	<0.01	<0.01	-3.40	2.80	0.22	0.90	10.57	3.36	<0.01	0.01
CLAS: Daily Activities	7.78	2.09	<0.01	<0.01	-1.96	3.16	0.54	0.96	9.74	3.80	0.01	0.04
CLAS: Overall	4.46	1.82	0.02	0.04	-4.10	2.76	0.14	0.83	8.56	3.31	0.01	0.04
SF-36: PCS	1.27	0.60	0.04	0.04	0.05	0.87	0.96	0.96	1.22	1.06	0.25	0.33
SF-36: MCS	1.88	0.87	0.03	0.04	0.35	1.28	0.78	0.96	1.53	1.55	0.33	0.33
Nema progresije bolesti												
FACT-G Total	4.83	1.14	<0.01	<0.01	0.25	1.81	0.89	0.89	4.58	2.13	0.03	0.06
FACT-An Fatigue	5.79	0.97	<0.01	<0.01	-0.85	1.54	0.58	0.89	6.64	1.82	<0.01	<0.01
CLAS: Energy	13.86	2.16	<0.01	<0.01	-1.19	3.45	0.73	0.89	15.05	4.07	<0.01	<0.01
CLAS: Daily Activities	15.53	2.41	<0.01	<0.01	1.58	3.98	0.69	0.89	13.95	4.65	<0.01	<0.01
CLAS: Overall	11.83	2.11	<0.01	<0.01	-2.67	3.41	0.43	0.89	14.50	4.01	<0.01	<0.01
SF-36: PCS	3.92	0.69	<0.01	<0.01	1-00	1.07	0.35	0.89	2.92	1.27	0.02	0.06
SF-36: MCS	3.55	0.99	<0.01	<0.01	1.52	1.60	0.34	0.89	2.03	1.88	0.28	0.28
Progresija bolesti												
FACT-G Total	-5.04	1.98	0.01	0.02	-9.31	2.53	<0.01	<0.01	4.27	3.16	0.18	0.94
FACT-An Fatigue	-5.26	1.64	<0.01	<0.01	-3.90	2.20	0.08	0.21	-1.36	2.74	0.62	0.94
CLAS: Energy	-9.32	3.67	0.01	0.02	-8.87	4.64	0.06	0.21	-0.45	5.91	0.94	0.94
CLAS: Daily Activities	-10.99	4.19	<0.01	0.02	-10.51	5.06	0.04	0.21	-0.48	6.49	0.94	0.94
CLAS: Overall	-13.73	3.59	<0.01	<0.01	-7.62	4.54	0.09	0.21	-6.11	5.79	0.29	0.94
SF-36: PCS	-5.38	1.20	<0.01	<0.01	-2.33	1.47	0.11	0.21	-3.05	1.89	0.11	0.75
SF-36: MCS	-2.36	1.82	0.20	0.20	-2.61	2.09	0.21	0.21	0.25	2.71	0.93	0.94

Tablica 3.4: Regresijska analiza: prosječni rezultat promjene kvalitete života

između grupa dobili smo da je statistički značajan. Zanima nas je li taj broj bodova i klinički relevantan. U zasebnoj analizi uspoređene su promjene ocjena između pacijenata koji su se poboljšali i onih koji su ostali stabilni. Razlike u odgovarajućim ocjenama smatrane su minimalno važnim razlikama. Na osnovu zaključaka te analize, rezultati za ljestvice specifične za rak nisu samo statistički značajni, već i klinički relevantni.

Rezultati opisanog ispitivanja potvrđuju pozitivan učinak liječenja epoetinom alfa u odnosu na placebo na kvalitetu života mjerenu svim korištenim upitnicima specifičnim za rak. Za općeniti upitnik SF-36 također su bile pozitivne promjene u korist epoetina alfa, ali nisu bile značajne. Takav rezultat nam sugerira da ne postoje nepredviđene nuspojave koje bi utjecale na općenitu kvalitetu života.

Korelacija hemoglobina i kvalitete života

S obzirom da se epoetin alfa koristi za povećanje razine hemoglobina u krvi, korelacijskom analizom proučena je veza između razine hemoglobina i kvalitete života. Izračunati su koeficijenti korelacije za svaku od sedam korištenih ljestvica samoprocjene kvalitete života koristeći dva različita pristupa.

Koeficijent korelacije izračunat je za početne vrijednosti razine hemoglobina u krvi i procjene kvalitete života. Dobiveni koeficijenti korelacije prikazani su u tablici 3.5. Osim

Ljestvica za samoprocjenu	**n**	**Koeficijent korelacije**	**P-vrijednost**	**Prilagođena p-vrijednost**
Početne vrijednosti				
FACT-G Total	288	0.26	<0.01	<0.01
FACT-An Fatigue subscale	293	0.16	<0.01	0.02
CLAS: Energy	336	0.14	0.01	0.03
CLAS: Daily Activities	336	0.17	<0.01	0.01
CLAS: Overall QoL	335	0.18	<0.01	0.01
SF-36: PCS	285	0.14	0.02	0.03
SF-36: MCS	285	0.09	0.15	0.15
Promjena kroz vrijeme				
FACT-G Total	266	0.26	<0.01	<0.01
FACT-An Fatigue subscale	273	0.29	<0.01	<0.01
CLAS: Energy	322	0.30	<0.01	<0.01
CLAS: Daily Activities	322	0.34	<0.01	<0.01
CLAS: Overall QoL	321	0.33	<0.01	<0.01
SF-36: PCS	250	0.26	<0.01	<0.01
SF-36: MCS	250	0.14	0.03	0.03

Tablica 3.5: Korelacija između razine hemoglobina i ocjene kvalitete života

za ljestvicu SF-36 MCS, pokazana je značajna korelacija između početne ocjene kvalitete života i početna razine hemoglobina. Za upitnike namijenjene pacijentima s onkološkim oboljenjima, zapažena je korelacija u rasponu od 0.14 do 0.26.

Koeficijent korelacije također je izračunat za promjene vrijednosti razine hemoglobina u krvi i promjene ocjene kvalitete života od prvog do zadnjeg dostupnog mjerenja, odnosno procjene. Ovim pristupom dobivena je nešto jača (pozitivna) korelacija, što je također prikazano u tablici 3.5. Vidimo da su u ovom slučaju utvrđeni koeficijenti korelacije u rasponu od 0.26 do 0.34 za upitnike FACT i CLAS.

Značajne korelacije sugeriraju da primjena djelatne tvari epoetin alfa doprinosi poboljšanju kvalitete života kroz povećanje razine hemoglobina u krvi.

3.4 Zaključak

U zadnjim desetljećima liječenje raka je uznapredovalo pa doktori i pacijenti ne razmišljaju samo o tome kako izliječiti bolest, već i brinu o kvaliteti života tijekom i nakon liječenja. Shodno tomu, kvaliteta života je sada uključena u sve više onkoloških istraživanja. Takva ispitivanja su bitna i zbog financiranja liječenja kako bi doktori donosili informirane odluke o izboru lijekova u terapiji.

Apriorno planirana višestruka linearna regresija, koja je obuhvaćala učinke progresije bolesti i nekoliko drugih moguće zbunjujućih varijabli na kvalitetu života, pokazala je značajnu prednost za epoetin alfa u odnosu na placebo za pet od sedam korištenih ljestvica. Analiza korelacije pokazala je značajnu pozitivnu vezu između promjene razine hemoglobina i promjene kvalitete života ocijenjene koristeći istih pet ljestvica koje su specifične za rak. Ovi nalazi ukazuju da epoetin alfa može poboljšati kvalitetu života kod anemičnih pacijenata oboljelih od raka koji prolaze kroz kemoterapiju, i da je ova promjena povezana s povećanjem razine hemoglobina.

Bibliografija

[1] L.Fallowfiel, D. Gagnond, M. Zagari, D. Cella, B. Bresnahan, T. J. Littlewood, P. McNulty, G. Gorzegno, M. Freund, *Multivariate regression analyses of data from a randomised, double-blind, placebo-controlled study confirm quality of life benefit of epoetin alfa in patients receiving non-platinum chemotherapy*, British Journal of Cancer, 87 (2002), 1341-1353

[2] C. F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, Hamburgi sumptibus Frid. Perthes et I.H.Besser, Hamburg, 1809

[3] C.F. Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae*, Apud Henricum Dieterich, Gottingen, 1823.

[4] R. V. Hogg, E. A. Tanis, D. L. Zimmerman, *Probability and Statistical Inference*, Pearson Education, Sjedinjene Američke Države, 2015.

[5] S. Holm, *A Simple Sequentially Rejective Multiple Test Procedure*, Scandinavian Journal of Statistics, 6 (1979), 65-70

[6] M. Huzak, *Matematička statistika*, Prirodoslovno-matematički fakultet, Zagreb, 2020.

[7] M. Huzak, *Statistika*, Prirodoslovno-matematički fakultet, Zagreb, 2021.

[8] I. Isa, B. Shyti, K.Spassov, *Multiple Regression Analysis used in Analysis of Private Consumption and Public Final Consumption Evolution, case of Albanian Economy*, European Journal of Marketing and Economics, 3 (2020), 63-70

[9] K. Jae-Eun, K. K. P. Johnson, *The Impact of Moral Emotions on Cause-Related Marketing Campaigns: A Cross-Cultural Examination*, Journal of Business Ethics, 112 (2013), 79–90

[10] R. W. Keener, *Theoretical Statistics*, Springer, New York, 2010.

[11] A.M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes*, Didot, Paris, 1805.

[12] D. C. Montgomery, E. A. Peck, G. Geoffrey, *Introduction to linear regression analysis*, John Wiley & Sons, New Jersey, 2012.

[13] N. S. Nataraja, N.R.Chilale, L. Ganesh, *Financial Performance of Private Commercial Banks in India: Multiple Regression Analysis*, Academy of Accounting and Financial Studies Journal, 22 (2018)

[14] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.

[15] N. Sandrić, Ž. Vondraček, *Vjerojatnost*, Prirodoslovno-matematički fakultet, Zagreb, 2019.

[16] G. A. F. Seber, A. J. Lee, *Linear Regression Analysis*, John Wiley & Sons, New Jersey, 2003.

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

[18] M. Verbeek, *Using linear regression to establish empirical relationships*, IZA World of Labor, 336 (2017)

[19] Epoetin Alfa Hexal, dostupno na `https://www.ema.europa.eu/en/documents/overview/epoetin-alfa-hexal-epar-summary-public_hr-0.pdf` (listopad 2023.)

Sažetak

Ovaj rad ima za cilj detaljno razraditi višeparametarsku linearnu regresiju te primijeniti dobivene rezultate u analizi učinkovitosti djelatne tvari epoetin alfa. U uvodnom poglavlju razmatraju se osnovni pojmovi iz vjerojatnosti i statistike, a zatim slijedi analiza modela višeparametarske linearne regresije kroz točkovnu i intervalnu procjenu parametara te testiranje hipoteza o modelu. Zaključno, rad primjenjuje teorijske spoznaje u analizi utjecaja epoetina alfa na kvalitetu života pacijenata s malignim bolestima, prezentira rezultate regresijske analize i istražuje povezanost između razine hemoglobina i kvalitete života. Analizom podataka iz randomiziranog, dvostruko slijepog, placebom kontroliranog ispitivanja koristeći višeparametarsku linearnu regresiju utvrđena je dobrobit epoetina alfa na životni standard kod bolesnika koji primaju kemoterapiju bez platine.

Summary

The aim of this work is to comprehensively elaborate on multiple linear regression and apply the obtained results in the analysis of the efficacy of the drug epoetin alfa. In the introductory chapter, the basic concepts of probability and statistics are considered, followed by the analysis of the multiple linear regression model through point and interval estimation of parameters and hypothesis testing about the model. In conclusion, the paper applies theoretical insights in analyzing the impact of epoetin alfa on the quality of life in patients with malignant diseases, presenting the results of regression analysis and exploring the correlation between hemoglobin levels and quality of life. Through the analysis of data from a randomized, double-blind, placebo-controlled trial using multiple linear regression, the benefit of epoetin alfa on the quality of life in patients receiving non-platinum chemotherapy was determined.

Životopis

Rođena sam 20.09.1997. godine u Splitu. Obrazovanje sam započela u Osnovnoj školi Pujanki, nakon čega sam upisala Treću gimnaziju u Splitu. Novo poglavlje u obrazovnom putu započela sam 2017. godine kada sam upisala preddiplomski studij Matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Stekla sam titulu sveučilišne prvostupnice matematike 2020. godine te sam iste godine upisala studij Financijske i poslovne matematike na istom fakultetu. Kroz svoje obrazovanje stekla sam ne samo teorijsko znanje, već i praktične vještine potrebne za snalaženje u svijetu matematike i financija. Na diplomskom studiju započela sam poslovnu karijeru u području financijske tehnologije. Očekujem s veseljem daljnje izazove koji će mi omogućiti stjecanje novih vještina i iskustava.