Usporedba različitih statističkih modela na predviđanje smrtnosti od COVID-a

Labaš, Maja

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Maja Labaš

USPOREDBA RAZLIČITIH STATISTIČKIH MODELA NA PREDVIĐANJE SMRTNOSTI OD COVID-a

Diplomski rad

Voditelj rada:
prof. dr. sc. Siniša Slijepčević

Zagreb, 2024.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

COVID - 19 pandemija započela je krajem 2019. godine, u kineskom gradu Wuhanu. Nekoliko mjeseci nakon proširila se na gotovo sve zemlje svijeta i razvila u najznačajniju krizu u novijoj povijesti, utječući negativno na svjetsku ekonomiju, gospodarstvo i zdravstvo. Brojni izazovi koje je svijetu donijela ova pandemija zahtijevali su hitan razvoj prediktivnih modela koji mogu učinkovito predvidjeti buduće trendove zaraze i tako osigurati pravovremene intervencije. Jedan od korisnih načina za generiranje traženih modela svakako je bila linearna regresija.

Linearna regresija statistička je metoda koja se koristi za modeliranje odnosa jedne ili više nezavisnih varijabli i jedne zavisne varijable. Primarni je cilj metode pronaći najprikladniju linearnu jednadžbu koja će minimalizirati razliku između stvarnih promatranih točaka i predviđenih vrijednosti koje je generirao model. Linearna se regresija naširoko koristi u različitim područjima, od ekonomije i financija do društvenih znanosti i epidemiologije. Služi kao temeljni alat za promatranje i razumijevanje odnosa unutar skupova podataka, što je čini bitnom komponentom statističke analize.

U ovom radu promatrat ćemo povezanost broja zaraženih i umrlih od virusa s nekim socio - ekonomskim, zdravstvenim i demografskim čimbenicima u različitim zemljama svijeta. Prvo ćemo postaviti teorijsku osnovu o linearnoj regresiji neophodnu za razumijevanje razvoja modela. Zatim ćemo obraditi metode odabira podskupa te metode sažimanja. Na kraju ćemo dobivene teorijske rezultate primijeniti u stvaranju modela koristeći podatke o COVID-u preuzete sa [9].

Poglavlje 1

Osnovni pojmovi vjerojatnosti i statistike

Da bismo lakše i detaljnije mogli razumijeti teoriju na kojoj će se temeljiti naš regresijski model, na početku ćemo se prisjetiti nekih osnovnih matematičkih pojmova iz područja vjerojatnosti i statistike, prateći definicije iz [3] i [6].

1.1 Osnovni pojmovi vjerojatnosti

Fiksirajmo vjerojatnosni prostor $(\Omega, \mathcal{F}, \mathbb{P})$, pri čemu je Ω neprazan skup elementarnih događaja, $\mathcal{F}$ σ-algebra na Ω te $\mathbb{P}$ vjerojatnost na izmjerivom prostoru $(\Omega, \mathcal{F})$. 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. *Kažemo da je $X : \Omega \to \mathbb{R}^k$ k-**dimenzionalna slučajna veličina** ukoliko je X izmjerivo preslikavanje u paru σ-algebri $(\mathcal{F}, \mathcal{B}(\mathbb{R}^k))$, tj. ako vrijedi da je*

$$(\forall B \in \mathcal{B}(\mathbb{R}^k))\{X \in B\} \in \mathcal{F}.$$

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

Definicija 1.1.2. *Neka je $X : \Omega \to \mathbb{R}^k$ k-dimenzionalna slučajna veličina definirana na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Induciranu vjerojatnost $\mathbb{P}_X$, definiranu na izmjerivom prostoru $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ relacijom*

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)), B \in \mathcal{B}((R^k)),$$

*zovemo **zakon razdiobe od** X.*

Definicija 1.1.3. *Neka je X k-dimenzionalna slučajna veličina sa zakonom razdiobe $\mathbb{P}_X$. Funkciju $F_X : \mathbb{R}^k \to [0,1]$ definiranu s*

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

*zovemo **funkcija razdiobe** ili **distribucije** od X.*

Neka je λ Lebesgueova mjera definirana na izmjerivom prostoru $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$.

Definicija 1.1.4. *Kažemo da je k-dimenzionalna slučajna veličina X **apsolutno neprekidna** ili, kraće, **neprekidna** ako postoji nenegativna Borelova funkcija f_X definirana na $\mathbb{R}^k$ takva da se funkcija razdiobe F_X može prikazati na sljedeći način:*

$$F_X(x) = \int_{\langle -\infty, x]} f_X(y)\, d\lambda(y), x \in \mathbb{R}^k.$$

*Funkciju f_X zovemo **funkcija gustoće razdiobe od** X ili, kraće, **gustoća od X**.*

Definicija 1.1.5. *Slučajna veličina X dimenzije k je **diskretna** ako postoji prebrojiv skup $D \subseteq \mathbb{R}^k$ takav da je $\mathbb{P}_X(D) = 1$. Nadalje, funkcija gustoće diskretne k-dimenzionalne slučajne veličine X je funkcija $f_X : \mathbb{R}^k \to \mathbb{R}$ definirana s*

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

Definicija 1.1.6. *Niz $(X_n)_{n\in I}$ je niz **nezavisnih** slučajnih veličina ako za svaki konačan podskup $I' = \{n_1, n_2, ..., n_l\}$ skup indeksa I i sve $B_1 \in \mathcal{B}(\mathbb{R}^{k_{n_1}})$, $B_2 \in \mathcal{B}(\mathbb{R}^{k_{n_2}})$, ..., $B_l \in \mathcal{B}(\mathbb{R}^{k_{n_l}})$ vrijedi*

$$\mathbb{P}(X_{n_1} \in B_1, X_{n_2} \in B_2, ..., X_{n_l} \in B_l) = \mathbb{P}(x_{n_1} \in B_1) \cdot \mathbb{P}(x_{n_2} \in B_2) \cdot ... \cdot \mathbb{P}(x_{n_l} \in B_l).$$

*Kažemo da je $(X_n)_{n\in I}$ niz **zavisnih** slučajnih veličina ako nije niz nezavisnih slučajnih veličina.*

Definicija 1.1.7. *Neka je $X : \Omega \to \mathbb{R}$ slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Kažemo da X ima matematičko očekivanje ako vrijedi*

$$\mathbb{E}|X| = \int_\Omega |X|\, d\mathbb{P} = \int_\Omega |X|(\omega)\, d\mathbb{P}(\omega) < \infty.$$

Tada matematičko očekivanje od X definiramo kao

$$\mathbb{E}[X] = \int_\Omega X\, d\mathbb{P} = \int_\Omega X(\omega)\, d\mathbb{P}(\omega).$$

Definicija 1.1.8. *Neka je $X = (X_1, X_2, ..., X_k)$ slučajni vektor. Kažemo da X ima matematičko očekivanje ako svaka komponenta $X_1, X_2, ..., X_k$ tog vektora ima matematičko očekivanje te u tom slučaju definiramo matematičko očekivanje od X kao vektor*

$$\mathbb{E}[X] = (\mathbb{E}X_1, \mathbb{E}X_2, ..., \mathbb{E}X_k).$$

Definicija 1.1.9. *Neka je X neka slučajna varijabla takva da je $\mathbb{E}[X^2] < \infty$. Tada varijancu od X definiramo s*

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

1.2 Osnovni pojmovi statistike

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

Primijetimo da statističku strukturu $(\Omega, \mathcal{F}, \mathcal{P})$ s jednočlanom množinom $\mathcal{P} = \{\mathbb{P}\}$ možemo poistovjetiti s vjerojatnosnim prostorom $(\Omega, \mathcal{F}, \mathbb{P})$.
Nadalje, množina $\mathcal{P}$ je često parametrizirana konačnodimenzijskim parametrom θ:

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

Definicija 1.2.2. ***Slučajan uzorak duljine*** *n na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je niz $X_1, X_2, ..., X_n$ slučajnih veličina na $(\Omega, \mathcal{F})$ takvih da su nezavisne i jednako distribuirane u odnosu na svaku vjerojatnost $\mathbb{P} \in \mathcal{P}$.*

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

Posebno, slučajnu veličinu T konačne dimenzije k ($k \geq 1$) zovemo statistikom ako postoje brojevi $n \in \mathbb{N}$, $d \in \mathbb{N}$, izmjerivo preslikavanje $t : \mathbb{R}^{dn} \to \mathbb{R}^k$ i n-dimenzionalni slučajni uzorak $(X_1, X_2, ..., X_n)$ veličina dimenzije d ($d \geq 1$), takvi da je

$$T = t(X_1, X_2, ..., X_n).$$

Primjer 1.2.4. *Neka je $X_1, X_2, ..., X_n$ slučajan uzorak varijabli, duljine n. Tada su*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{(aritmetička sredina)},$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \textit{(uzoračka varijanca)}$$

statistike. Obje navedene statistike su dimenzije 1. *Za aritmetičku sredinu je* $\bar{X}_n = t(X_1, X_2, ..., X_n)$ *gdje je*

$$t(x_1, x_2, ..., x_n) = \bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n), \quad (x_1, x_2, ..., x_n) = \mathbf{x} \in \mathbb{R}^n,$$

izmjeriva funkcija $t : \mathbb{R}^n \to \mathbb{R}$*, a za uzoračku varijancu je* $S_n^2 = v(X_1, X_2, ..., X_n)$ *gdje je*

$$v(x_1, x_2, ..., x_n) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad (x_1, x_2, ..., x_n) = \mathbf{x} \in \mathbb{R}^n,$$

izmjeriva funkcija $v : \mathbb{R}^n \to \mathbb{R}$*.*

Neka je $\mathbf{X} = (X_1, X_2, ..., X_n)$ slučajni uzorak duljine n ($n \geq 1$) iz parametarskog modela

$$\mathcal{P} = \{f(\cdot; \theta) \quad : \quad \theta \in \Theta\}$$

gdje su $f(\cdot; \theta)$ gustoće dimenzija d ($d \geq 1$) slučajnih veličina $X_1, X_2, ..., X_n$ parametrizirane parametrom θ dimenzije m ($m \geq 1$). Želimo procijeniti vrijednost od $\tau(\theta)$ gdje je $\tau : \Theta \to \mathbb{R}^k$ izmjeriva funkcija ($k \geq 1$). Tada je **procjenitelj** od $\tau(\theta)$ statistika $T = t(\mathbf{X})$ iste dimenzije kao i funkcija τ. Statistiku T još zovemo *točkovni procjenitelj* od $\tau(\theta)$. Zaključujemo da procjenitelj može biti bilo koja statistika iste dimenzije kao funkcija parametra koju želimo procijeniti.

Nećemo svaku statistiku dimenzije k koristiti za procjenu od $\tau(\theta)$ već ćemo među takvim statistikama za procjenitelja odabrati onu statistiku $T = t(\mathbf{X})$ koja će, u nekom smislu, biti optimalan odabir [3, poglavlje 4]. Poželjna svojstva procjenitelja su da je lociran blizu prave vrijednosti parametra te da ima malu raspršenost.

Definicija 1.2.5. *Procjenitelj* T *je* ***nepristran*** *za* τ *ako vrijedi*

$$\mathbb{E}[T] = \tau.$$

Za procjenitelj koji nije nepristran kažemo da je ***pristran****.*

Nepristranost procjenitelja dobro je svojstvo, no u nekim situacijama pristrani procjenitelj biti će bolji od nepristranog jer, od nepristranosti, bitnije je da procjenitelj ima malu *srednjekvadratnu grešku*.

Definicija 1.2.6. *Neka je* $T = t(\mathbf{X})$ *procjenitelj za* $\tau(\theta)$ *kojemu komponente imaju konačnu varijancu za svaki* $\theta \in \Theta$*. Tada je* ***srednjekvadratna pogreška procjene*** *od* $\tau(\theta)$ *s* T *u odnosu na* $\mathbb{P}_\theta$ *broj*

$$MSE_\theta(T) = \mathbb{E}_\theta[|T - \tau(\theta)|^2].$$

Neka je X neko statističko obilježje koje promatramo. Svaka pretpostavka o populacijskoj razdiobi od X naziva se *statističkom hipotezom*. Ukoliko ona jednoznačno određuje razdiobu od X zovemo je *jednostavna statistička hipoteza*, a u suprotnom kažemo da je *složena*.

Osnovna hipoteza koja se testira zove se *nul-hipoteza* i označava se sa H_0. Uz nul-hipotezu, postavlja se i njoj alternativna hipoteza koju označavamo s H_1. Na osnovi realizacije slučajnog uzorka za X želimo donijeti odluku hoćemo li *odbaciti* ili *ne odbaciti* H_0 u korist H_1. Postupak donošenja odluke o odbacivanju statističke hipoteze zove se *testiranje statističkih hipoteza*.

Definicija 1.2.7. *Test za testiranje statističke hipoteze H_0 u odnosu na H_1 je preslikavanje* $\tau : \mathbb{R}^n \to \{0, 1\}$.
Ako je za realizaciju $\mathbf{x}$ *uzorka* $\mathbf{X}$ $\tau(\mathbf{x}) = 1$*, tada odbacujemo H_0 u korist H_1, a ako je* $\tau(\mathbf{x}) = 0$*, tada ne odbacujemo H_0 u korist H_1.*

Razina značajnosti testa α je vjerojatnost odbacivanja H_0 ako je H_0 istinita hipoteza. Ako odbacimo H_0, a ona je istinita, činimo *pogrešku prve vrste*. S druge strane, ako ne odbacujemo H_0, a istinita je H_1 kažemo da smo napravili *pogrešku druge vrste*.

Poglavlje 2

Linearna regresija

Linearna regresija jedna je od najčešće korištenih statističkih metoda. Njezin cilj je uspostaviti vezu između varijable odziva (zavisna varijabla) i varijable poticaja (nezavisna varijabla). Zavisna varijabla je uvijek jedna, dok nezavisnih varijabli može biti više. Ukoliko promatramo zavisnost varijable odziva o samo jednoj nezavisnoj varijabli koristimo model jednostavne ili jednostruke linearne regresije. Ako nas zanima povezanost zavisne varijable s više nezavisnih varijabli koristimo model višestruke linearne regresije [8].

2.1 Oblikovanje modela

Jednostruka linearna regresija

Jednostruka linearna regresija najjednostavniji je slučaj regresije i čini bazu na kojoj se grade kompliciraniji regresijski modeli. Ako nas zanima utjecaj nezavisne varijable x na zavisnu varijablu y dobivamo model sljedećeg oblika:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.1}$$

pri čemu su β_0 i β_1 nepoznati parametri koje trebamo procijeniti, a ϵ je slučajna greška. β_0 još nazivamo parametrom presjeka, dok je β_1 parametar nagiba regresijskog pravca.

Najčešće analizu neke pojave temeljimo na više od jednom mjerenju. Stoga pretpostavimo da imamo niz sparenih mjerenja duljine n, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, pri čemu su $x_i, i = 1, 2, ..., n$ vrijednosti nezavisne, a $y_i, i = 1, 2, ..., n$ odgovarajuće vrijednosti zavisne varijable. Tada se, u skladu s (2.1), dani uzorak opisuje preko sljedećih n jednadžbi:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

Bitno je za naglasiti da x_i-evi nisu slučajne varijable već samo brojevi. S druge strane, ϵ_i-evi su nezavisne slučajne varijable s očekivanjem $\mathbb{E}[\epsilon_i] = 0$ i varijancom $\text{Var}(\epsilon_i) = \sigma^2 > 0$. y_i-evi su također slučajne varijable s obzirom da oni ovise o ϵ_i-evima.

Višestruka linearna regresija

U prethodnom odjeljku vidjeli smo kako izgleda linearni regresijski model kada varijabla odziva ovisi o jednoj nezavisnoj varijabli. U stvarnosti taj slučaj je zapravo dosta rijedak jer na skoro svaku pojavu utječe barem nekoliko različitih čimbenika. Motivirani tom činjenicom, u ostatku poglavlja bavit ćemo se višestrukim linearnim regresijskim modelom.

Model višestruke linearne regresije dan je s:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon, \tag{2.2}$$

pri čemu $x_1, ..., x_k$ označavaju k nezavisnih varijabli.

Analogno kao i kod jednostavne linearne regresije, kada imamo n mjerenja oblika $(x_{i1}, x_{i2}, ..., x_{ik}, y_i)$, $i = 1, 2, ..., n$, model poprima oblik

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + \epsilon_n. \end{aligned} \tag{2.3}$$

Često je (2.3) lakše zapisati matrično. U tu svrhu definiramo sljedeće matrice:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Dakle, X je matrica dimenzija $n \times (k+1)$ koja u i-tom retku na prvoj poziciji sadrži jedinicu, a ostali elementi su vrijednosti nezavisnih varijabli iz i-tog mjerenja. Nadalje, $\mathbf{y}$ je vektor stupac duljine n s opaženim vrijednostima zavisne varijable, a $\boldsymbol{\beta}$ i $\boldsymbol{\epsilon}$ su redom vektor stupac duljine $k+1$ i vektor stupac duljine n koji sadrže nepoznate parametre, odnosno slučajne greške.

Sada model (2.3) poprima matrični oblik:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.4}$$

2.2 Procjena parametara

Sljedeći cilj je odrediti nepoznate parametre $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_k]^T$ kako bi utvrdili značajnost utjecaja svake varijable na varijablu odziva. Najpoznatija metoda za procjenu tih parametara je metoda najmanjih kvadrata koja se zasniva na minimiziranju sume kvadrata reziduala dane sa:

$$\begin{aligned} S &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2 \\ &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 . \end{aligned} \tag{2.5}$$

Da bi minimizirali izraz u (2.5) možemo ga derivirati po svakom β_i-u i izjednačiti derivacije s 0. Dobivamo

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) = 0 \\ \frac{\partial S}{\partial \beta_j} &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) x_{ij} = 0, \ j = 1, 2, ..., k \end{aligned} \tag{2.6}$$

što je ekvivalentno

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i \\
&\vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i.
\end{aligned}
\tag{2.7}
$$

S obzirom na veličinu zapisa (2.7) nekad je praktičnije koristiti matrični prikaz. U matričnom prikazu suma kvadrata reziduala jednaka je

$$
\begin{aligned}
S =& (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) \\
=& \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T X^T\mathbf{y} - \mathbf{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} \\
=& \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T(X^T\mathbf{y}) + \boldsymbol{\beta}^T(X^T X)\boldsymbol{\beta},
\end{aligned}
\tag{2.8}
$$

pri čemu smo koristili da je $\mathbf{y}^T X\boldsymbol{\beta}$ skalar jednak $\boldsymbol{\beta}^T X^T\mathbf{y}$.
Da bi minimizirali sumu kvadrata reziduala u ovakvom prikazu, koristit ćemo matrične derivacije. U tom slučaju imamo:

$$
\frac{\partial S}{\partial \boldsymbol{\beta}} = -2X^T\mathbf{y} + 2X^T X\boldsymbol{\beta}. \tag{2.9}
$$

Kada izjednačimo (2.9) s 0, dobivamo da procjenitelj $\mathbf{b}$ od $\boldsymbol{\beta}$ zadovoljava sljedeću jednakost:

$$
(X^T X)\mathbf{b} = X^T\boldsymbol{\beta}^T X^T\mathbf{y}. \tag{2.10}
$$

Ukoliko je matrica $X^T X$ regularna, procjenitelj metodom najmanjih kvadrata je jedinstven i dan s

$$
\mathbf{b} = (X^T X)^{-1}X^T\mathbf{y}. \tag{2.11}
$$

Time smo dobili procijenjene parametre regresije i sada možemo izračunati predviđene vrijednosti $\hat{\boldsymbol{y}}$ zavisne varijable $\mathbf{y}$. Ako uzmemo u obzir n mjerenja dobivamo

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = X\mathbf{b} = X(X^TX)^{-1}X^T\mathbf{y}. \tag{2.12}$$

Definirajmo $H = X(X^TX)^{-1}X^T$ te $M = I - H$. Tada je

$$\hat{\boldsymbol{y}} = H\boldsymbol{y}, \tag{2.13}$$

$$MX = (I - H)X = X - X(X^TX)^{-1}X^TX = X - X = 0. \tag{2.14}$$

Matricu H nazivamo kapa matricom. Ona preslikava vektor opaženih vrijednosti u vektor predviđenih vrijednosti. Sada ako reziduale definiramo na način

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}, \tag{2.15}$$

koristeći (2.13), definiciju od M, (2.4) te (2.14) dobivamo da je

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} \overset{(2.13)}{=} \boldsymbol{y} - H\boldsymbol{y} = (I - H)\boldsymbol{y} = M\boldsymbol{y} \overset{(2.4)}{=} M(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = MX\boldsymbol{\beta} + M\boldsymbol{\epsilon} \overset{(2.14)}{=} M\boldsymbol{\epsilon}. \tag{2.16}$$

Sada lako možemo dokazati tvrdnju teorema koji slijedi.

Teorem 2.2.1. *Vektor reziduala $\boldsymbol{e}$ ortogonalan je na matricu nezavisnih varijabli X te na vektor predviđenih vrijednosti zavisne varijable $\hat{\mathbf{y}}$.*

Dokaz. Zbog (2.16) imamo da je

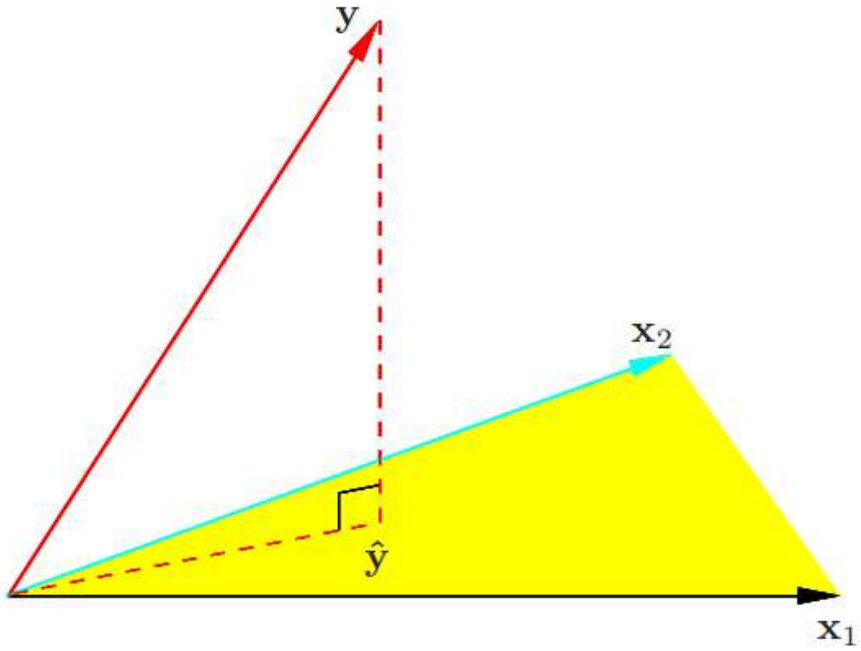
$$X^T\boldsymbol{e} = X^TM\boldsymbol{\epsilon} \overset{(2.14)}{=} 0\boldsymbol{\epsilon} = \mathbf{0}, \tag{2.17}$$

pri čemu smo s $\mathbf{0}$ označili nul vektor. Sada slijedi

$$\hat{\mathbf{y}}^T\boldsymbol{e} = \boldsymbol{b}^TX^T\boldsymbol{e} = 0, \tag{2.18}$$

čime je teorem dokazan. □

Prethodni teorem nam govori da minimizaciju $S = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})$ vršimo tako što odabiremo $\boldsymbol{\beta}$ takav da je vektor reziduala $\boldsymbol{e} = \boldsymbol{y} - \hat{\mathbf{y}}$ okomit na potprostor određen vektorima stupcima matrice X. Tada $\hat{\mathbf{y}}$ predstavlja ortogonalnu projekciju vektora $\boldsymbol{y}$ na taj potprostor. Ovime je dana geometrijska interpretacija najmanjih kvadrata što ju čini intuitivno prihvatljivijom. Slika 2.1 prikazuje prethodno opisano u slučaju dvije nezavisne varijable.



Slika 2.1: *Prikaz geometrijske interpretacije najmanjih kvadrata.* Izvor [2]

Rekli smo da je s (2.11) dan jedinstveni procjenitelj ukoliko je X^TX regularna matrica, odnosno ukoliko su nezavisne varijable međusobno nekolinearne. No što ako je X^TX singularna? Tada će stupci od X biti linearno zavisni i koeficijenti $\boldsymbol{b} = [\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k]$ dobiveni metodom najmanjih kvadrata neće biti jedinstveni. Unatoč tome, procijenjene vrijednosti $\hat{\boldsymbol{y}} = X\boldsymbol{b}$ biti će i dalje ortogonalne projekcije od $\boldsymbol{y}$ na potprostor razapet sa stupcima matrice X. Jedina je razlika da će u ovom slučaju postojati više različitih mogućnosti za prikaz te projekcije kao kombinacije vektora stupaca.

2.3 Gauss - Markovljevi uvjeti

U ovom odjeljku pitamo se koliko je dobra metoda najmanjih kvadrata i da li ona uvijek dobro procjenjuje parametre $\boldsymbol{\beta} = [\beta_0, ..., \beta_k]$? Da bi dali odgovore na ta pitanja, prvo ćemo spomenuti da postoje određeni uvjeti koji jamče kvalitetu procjene najmanjih kvadrata. Također ćemo dokazati jedan od najpoznatijih rezultata u statistici, Gauss - Markovljev teorem, koji govori da procjenitelj od $\boldsymbol{\beta}$ dobiven metodom najmanjih kvadrata ima najmanju varijancu od svih linearno nepristranih procjenitelja.

Definicija 2.3.1. *Pretpostavimo da za slučajne greške* $\boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_n]$ *vrijedi:*

- $\mathbb{E}[\epsilon_i] = 0$
- $\mathbb{E}[\epsilon_i^2] = \sigma^2$
- $\mathbb{E}[\epsilon_i \epsilon_j] = 0$, $i \neq j$.

Navedene uvjete nazivamo ***Gauss - Markovljevi uvjeti****.*

Uvjete također možemo zapisati u matričnoj formi

$$\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0},\ \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I. \tag{2.19}$$

Slijedi

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[X\boldsymbol{\beta} + \boldsymbol{\epsilon}] = X\boldsymbol{\beta} \tag{2.20}$$

te

$$\operatorname{cov}(\boldsymbol{y}) = \mathbb{E}[(\boldsymbol{y} - X\boldsymbol{\beta})(\boldsymbol{y} - X\boldsymbol{\beta})^T] = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I. \tag{2.21}$$

Teorem 2.3.2. *Pretpostavimo da vrijede Gauss - Markovljevi uvjeti te da je* $\boldsymbol{b}$ *procjenitelj od* $\boldsymbol{\beta}$ *dobiven metodom najmanjih kvadrata. Tada je* $\boldsymbol{b}$ *nepristrani procjenitelj od* $\boldsymbol{\beta}$ *te vrijedi*

$$cov(\boldsymbol{b}) = \sigma^2 (X^T X)^{-1}. \tag{2.22}$$

Dokaz. Zbog (2.20) imamo

$$\mathbb{E}[\boldsymbol{b}] = \mathbb{E}[(X^T X)^{-1} X^T \boldsymbol{y}] = (X^T X)^{-1} X^T \mathbb{E}[\boldsymbol{y}] = (X^T X)^{-1} X^T X\boldsymbol{\beta} = \boldsymbol{\beta}$$

odakle direktno slijedi nepristranost procjenitelja $\boldsymbol{b}$. Neka je $A = (X^T X)^{-1} X^T$. Tada je $\boldsymbol{b} = A\boldsymbol{y}$ pa koristeći (2.21) dobivamo

$$\begin{aligned}\operatorname{cov}(\boldsymbol{b}) =& A \operatorname{cov}(\boldsymbol{y}) A^T = \sigma^2 AIA^T = \sigma^2 AA^T \\ =& \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1},\end{aligned}$$

čime je teorem dokazan.

□

Procjena σ^2

Kako bi mogli koristiti dobivene formule iz ovog odjeljka, trebati će nam σ^2 koja je u pravilu nepoznata te ju treba procijeniti. U tome će nam pomoći reziduali. S obzirom da je $M = (m_{ij})$ simetrična, idempotentna matrica vrijedi

$$\sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^T\boldsymbol{e} = \boldsymbol{\epsilon}^T M^T M\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^T M\boldsymbol{\epsilon} = \sum_{i=1}^{n} m_{ii}\epsilon_i^2 + \sum_{i,j=1,i\neq j}^{n} m_{ij}\epsilon_i\epsilon_j. \tag{2.23}$$

Slijedi da je

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} e_i^2\right] &= \sum_{i=1}^{n} m_{ii}\mathbb{E}[\epsilon_i^2] + \sum_{i,j=1,i\neq j}^{n} m_{ij}\mathbb{E}[\epsilon_i\epsilon_j] \\ &= \sigma^2 \sum_{i=1}^{n} m_{ii} = \sigma^2\mathrm{tr}(M) \\ &= (n-k-1)\sigma^2\end{aligned}$$

kada je slobodni član β_0 različit od nule i postoji k nezavisnih varijabli jer tada je $\mathrm{tr}(M) = \mathrm{tr}(I) - \mathrm{tr}(H) = n - k - 1$. Ako sada definiramo

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}, \tag{2.24}$$

vidimo da je s^2 nepristrani procjenitelj od σ^2. U slučaju da je $\beta_0 = 0$ dobivamo da je

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k} \tag{2.25}$$

nepristrani procjenitelj od σ^2.

Koeficijent determinacije R^2

Vidjeli smo da su reziduali korisni kod utvrđivanja jesu li zadovoljeni Gauss - Markovljevi uvjeti. Osim toga, biti će korisni u određivanju kvalitete modela. Pomoću reziduala ćemo definirati mjeru koja će nam označavati jačinu linearne veze između zavisne i nezavisnih varijabli. Ta mjera naziva se koeficijent determinacije, u oznaci R^2, a definiramo ga na sljedeći način:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.26}$$

kada je $\beta_0 \neq 0$, odnosno

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n y_i^2} \tag{2.27}$$

kada je $\beta_0 = 0$. S $\bar{y}$ označavamo $\frac{1}{n}\sum_{i=1}^n y_i$. Dodatno, vrijedi da je korijenovana vrijednost u slučaju postojanja slobodnog člana dana s

$$R = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2\right]^{\frac{1}{2}}}. \tag{2.28}$$

Da bi se uvjerili da (2.28) zaista vrijedi koristit će nam sljedeće dvije tvrdnje.

Teorem 2.3.3. *Neka je* $\bar{y} = n^{-1}\sum_{i=1}^n y_i$. *Tada je*

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i{}^2 = \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right) - \left(\sum_{i=1}^n \hat{y_i^2} - n\bar{y}^2\right).$$

Dokaz. Iz teorema 2.2.1 slijedi $\hat{\boldsymbol{y}}^T \boldsymbol{e} = \sum_{i=1}^n e_i \hat{y}_i = 0$ pa imamo

$$\begin{aligned}
\sum_{i=1}^n y_i^2 &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i)^2 \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n \hat{y}_i{}^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)\hat{y}_i \\
&= \sum_{i=1}^n e_i^2 + \sum_{i=1}^n \hat{y}_i{}^2 + 2\sum_{i=1}^n e_i \hat{y}_i \\
&= \sum_{i=1}^n e_i^2 + \sum_{i=1}^n \hat{y}_i{}^2,
\end{aligned}$$

odakle slijedi tvrdnja. □

Korolar 2.3.4. *Ako u modelu postoji slobodan član* $\beta_0 \neq 0$, *vrijedi*

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

Dokaz. Kako je $\beta_0 \neq 0$, teorem 2.2.1 povlači $\mathbf{1}^T \boldsymbol{e} = \boldsymbol{e}^T \mathbf{1} = \sum_{i=1}^n e_i = 0$ zbog čega je i $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$. U ovom slučaju je i očekivanje opaženih vrijednosti isto kao očekivanje predviđenih vrijednosti. Na kraju, iz teorema 2.3.3 slijedi

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) - \left(\sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2\right) \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,\end{aligned}$$

čime je tvrdnja dokazana. □

Pokažimo da je kvadrat izraza (2.28) jednak izrazu u (2.26). S obzirom da je $\sum_{i=1}^{n} e_i\hat{y}_i = 0$, imamo

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y}) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} e_i\hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.\end{aligned}$$

Zaključujemo da je

$$R = \frac{\left[\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right]^{\frac{1}{2}}}{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}} \left(\Longleftrightarrow R^2 = \frac{\left[\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right]}{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}\right),$$

odakle sada korištenjem korolara 2.3.4 slijedi

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

Može se zaključiti da koeficijent determinacije leži između 0 i 1 te vrijedi da je prilagodba modelu bolja što je R^2 bliži 1.

Dodavanjem varijabli u model smanjuje se suma kvadrata reziduala čime se povećava R^2. Navedena promjena mogla bi rezultirati donošenjem krivih zaključaka. Kako bi se to izbjeglo, često se, umjesto R^2, koristi prilagođeni koeficijent R_a^2. Njega definiramo na sljedeći način:

$$R_a^2 = 1 - \frac{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-k-1}}{\frac{\sum_{i=1}^n (y_i - \bar{y}_i)^2}{n-1}} = 1 - \frac{s^2}{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)}. \tag{2.29}$$

Obje sume podijelili smo odgovarajućim stupnjem slobode. Na taj način R_a^2, za razliku od R^2, plaća cijenu za uključivanje nepotrebnih varijabli u model. Uz to, možemo primijetiti da je maksimiziranje prilagođenog R_a^2 ekvivalentno minimiziranju s^2.

Gauss - Markovljev teorem

Prije nego počnemo s dokazom Gauss - Markovljevog teorema, uvest ćemo još neke definicije.

Definicija 2.3.5. *Neka je $\boldsymbol{y}$ vektor opaženih vrijednosti zavisne varijable te neka je $L : \mathbb{R}^{k+1} \to \mathbb{R}$ linearna funkcija takva da je $L(\boldsymbol{\beta}) = l^T\boldsymbol{\beta}$, za neki vektor $l \in \mathbb{R}^n$. Za statistiku T kažemo da je:*

- *linearni procjenitelj za $L(\boldsymbol{\beta})$ ako je*

$$T = c^T\boldsymbol{y}, \quad \text{za neki neslučajni vektor } \ c \in \mathbb{R}^n.$$

- *najbolji linearni nepristrani procjenitelj (BLUE) za $L(\boldsymbol{\beta})$ ako on za $L(\boldsymbol{\beta})$:*
  - *je linearni procjenitelj*
  - *je nepristrani procjenitelj*
  - *u klasi svih nepristranih linearnih procjenitelja za $L(\boldsymbol{\beta})$ ima najmanju varijancu.*

Ako promatramo procjene proizvoljne linearne kombinacije parametara oblika $l^T\boldsymbol{\beta}$, možemo zaključiti da je procjenitelj za $l^T\boldsymbol{\beta}$ dobiven metodom najmanjih kvadrata dan s $l^T\boldsymbol{b} = l^T(X^TX)^{-1}X^T\boldsymbol{y}$. S obzirom da je matrica X fiksirana, $l^T\boldsymbol{b}$ je linearna funkcija oblika $c^T\boldsymbol{y}$ koja ovisi o varijabli odziva $\boldsymbol{y}$. Sada je

$$\mathbb{E}[l^T\boldsymbol{b}] = \mathbb{E}[l^T(X^TX)^{-1}X^T\boldsymbol{y}] = l^TT(X^TX)^{-1}X^TX\boldsymbol{\beta} = l^T\boldsymbol{\beta} \tag{2.30}$$

pa zaključujemo da je $l^T\boldsymbol{b}$ nepristrani procjenitelj za $l^T\boldsymbol{\beta}$. Gauss - Markovljev teorem će nam reći da, ukoliko postoji bilo koji drugi linearni nepristrani procjenitelj $c^T\boldsymbol{y}$ za $l^T\boldsymbol{\beta}$, vrijedi

$$\text{var}(l^T\boldsymbol{b}) \leq \text{var}(c^T\boldsymbol{y}). \tag{2.31}$$

Teorem 2.3.6. *Neka je* $\boldsymbol{b} = (X^T X)^{-1} X^T \boldsymbol{y}$ *i* $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$*. Tada je, pod Gauss - Markovljevim uvjetima, procjenitelj* $l^T \boldsymbol{b}$ *funkcije* $l^T \boldsymbol{\beta}$ *najbolji linearni nepristrani procjenitelj (BLUE).*

Dokaz. Neka je $c^T \boldsymbol{y}$ neki drugi linearni nepristrani procjenitelj za $l^T \boldsymbol{\beta}$. S obzirom da je $c^T \boldsymbol{y}$ nepristrani procjenitelj od $l^T \boldsymbol{\beta}$, vrijedi $l^T \boldsymbol{\beta} = \mathbb{E}[c^T \boldsymbol{y}] = c^T X \boldsymbol{\beta}$ za sve $\boldsymbol{\beta}$ pa dobivamo

$$c^T X = l^T. \tag{2.32}$$

Sada je

$$\operatorname{var}(c^T \boldsymbol{y}) = c^T \operatorname{cov}(\boldsymbol{y}) c \overset{(2.21)}{=} c^T (\sigma^2 I) c = \sigma^2 c^T c \tag{2.33}$$

i

$$\operatorname{var}(l^T \boldsymbol{b}) = l^T \operatorname{cov}(\boldsymbol{b}) l \overset{(2.22)}{=} \sigma^2 l^T (X^T X)^{-1} l \overset{(2.32)}{=} \sigma^2 c^T X (X^T X)^{-1} X^T c. \tag{2.34}$$

Konačno slijedi

$$\begin{aligned} \operatorname{var}(c^T \boldsymbol{y}) - \operatorname{var}(l^T \boldsymbol{b}) =& \sigma^2 [c^T c - c^T X (X^T X)^{-1} X^T c] \\ =& \sigma^2 c^T [I - X (X^T X)^{-1} X^T] c \geq 0, \end{aligned}$$

pri čemu smo u zadnjem koraku koristili činjenicu da je matrica $I - X(X^T X)^{-1} X^T = M$ semi - definitna.

□

Za kraj ovog odjeljka pogledajmo srednje kvadratnu pogrešku (MSE) procjenitelja $\tilde{\theta}$ kod procjene za θ:

$$\begin{aligned} MSE(\tilde{\theta}) =& \mathbb{E}[\tilde{\theta} - \theta]^2 \\ =& \operatorname{var}(\tilde{\theta}) + (\mathbb{E}[\tilde{\theta}] - \theta)^2. \end{aligned}$$

Vidimo da je srednje kvadratna pogreška jednaka zbroju varijance procjenitelja i njegove kvadratne pristranosti. Gauss - Markovljev teorem implicira da procjenitelj dobiven metodom najmanjih kvadrata ima najmanju srednje kvadratnu pogrešku među svim linearnim nepristranim procjeniteljima. Međutim, može se dogoditi da postoji neki procjenitelj s još manjom srednje kvadratnom pogreškom koji je pristran. Takav bi procjenitelj zamijenio malo pristranosti za veće smanjenje varijance. Bilo koja metoda koja neke koeficijente dobivene metodom najmanjih kvadrata smanji ili ih izjednači s nulom može dati pristrane procjene [2, poglavlje 3]. S obzirom da nam je u interesu dobiti procjenitelja sa što manjom srednje kvadratnom pogreškom, u poglavljima koja slijede dotaknut ćemo se nekih metoda koje će rezultirati pristranim procjenama.

2.4 Test hipoteza i intervali pouzdanosti

Ako se želimo uvjeriti da je model dobar, nameće nam se još nekoliko pitanja. Pomažu li sve varijable poticaja objasniti zavisnu varijablu ili je samo dio njih koristan? Ili još općenitije pitanje, postoji li uopće ijedna varijabla poticaja koja je značajna u predviđanju vrijednosti varijable odziva? Kako bi dobili odgovore na spomenuta pitanja koristit ćemo test hipoteza, odnosno test značajnosti linearnog regresijskog modela te ćemo na kraju definirati intervale pouzdanosti za parametre $\beta_0, \beta_1, ..., \beta_k$. Za njihovo oblikovanje potrebno je donijeti neke dodatne pretpostavke. Pretpostavimo da uz Gauss - Markovljeve uvjete vrijedi $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, ..., n$. Iz teorema 2.3.2 slijedi da je

$$\mathbb{E}[\boldsymbol{b}] = \boldsymbol{\beta}, \;\; \text{cov}(\boldsymbol{b}) = \sigma^2 (X^T X)^{-1}.$$

Stoga je

$$\boldsymbol{b} \sim N(\boldsymbol{\beta}, (X^T X)^{-1} \sigma^2)$$

te za svaki $j \in 0, 1, .., k$ vrijedi

$$\hat{\beta}_j \sim N(\beta_j, v_j \sigma^2), \tag{2.35}$$

pri čemu je v_j j-ti dijagonalni element matrice $(X^T X)^{-1}$. Već smo ranije procijenili varijancu σ^2 s

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}.$$

Može se pokazati da je

$$(n - k - 1)\frac{s^2}{\sigma^2} \sim \chi^2(n - k - 1).$$

Dodatno, $\boldsymbol{b}$ i s^2 su nezavisne slučajne varijable.

Da bi testirali hipotezu da je određeni koeficijent β_j jednak nuli, formiramo standardizirani koeficijent koji se još naziva Z - *score*:

$$z_j = \frac{\hat{\beta}_j}{s\sqrt{v_j}}. \tag{2.36}$$

U slučaju da je nulta hipoteza $\beta_j = 0$ prihvaćena, z_j ima t distribuciju s $n - k - 1$ stupnjeva slobode i zato će velika vrijednost od z_j dovesti do odbacivanja navedene nul - hipoteze.

Ako s zamijenimo s poznatom vrijednošću σ, z_j će imati standardnu normalnu razdiobu. Porastom veličine uzorka razlike između repnih kvantila t distribucije i standardne

normalne distribucije postaju zanemarive te se zbog toga uglavnom koriste kvantili normalne distribucije [2, poglavlje 3].

Ponekad testove provodimo ne samo za jedan koeficijent već za određenu grupu koeficijenata. Možemo izdvojiti podskup od l varijabli za koje smo uvjereni da su značajno povezane sa zavisnom varijablom i tako od potpunog modela veličine k dobiti reducirani model te provjeriti da li je on dovoljan. Provodimo test tako da na prvih l mjesta stavimo l izdvojenih prediktora. Tada ako je m takav da je $l + m = k$ postavljamo sljedeće hipoteze:

$$H_0 : \beta_{l+1} = \beta_{l+2} = ... = \beta_{l+m} = 0 \text{ (Reducirani model je dovoljan)}$$
$$H_a : \beta_j \neq 0, j \in \{l + 1, l + 2, ..., l + m\} \text{ (Potreban je potpuni model).}$$

Označimo sa SSE sumu kvadrata reziduala potpunog modela, a sa SSE_r sumu kvadrata reziduala reduciranog modela. Odluku o prihvaćanju, odnosno odbacivanju hipoteze donosimo na temelju testne statistike F koja je dana s

$$F = \frac{\frac{SSE_r - SSE}{k-l}}{\frac{SSE}{n-k-1}}.$$

F - statistika mjeri promjenu u sumi kvadrata reziduala po dodatnim parametrima u većem modelu te je normalizirana procjenom od σ^2. Također, može se pokazati da je F - statisika za izbacivanje samo jednog koeficijenta β_j jednaka kvadratu odgovarajućeg Z - *score*-a z_j definiranog s (2.36). Pod Gauss - Markovljevim uvjetima i nul - hipotezom da je manji model dovoljan F - statistika imat će F distribuciju:

$$F = \frac{\frac{SSE_r - SSE}{k-l}}{\frac{SSE}{n-k-1}} \overset{H_0}{\sim} F(k - l, n - k - 1).$$

U nekim slučajevima možemo posumnjati u postojanje linearne veze između zavisne i nezavisnih varijabli. Tada provjeravamo postoji li barem jedna varijabla poticaja koja je značajna u predviđanju vrijednosti varijable odziva. Test kojim radimo takvu provjeru naziva se test značajnosti linearnog regresijskog modela. U višestrukom regresijskom modelu s k nezavisnih varijabli zanima nas da li su svi regresijski koeficijenti jednaki nuli, odnosno da li je $\beta_1 = \beta_2 = ... = \beta_k = 0$. U tu svrhu postavljamo test hipoteza u kojem nul - hipotezom tvrdimo da su svi koeficijenti nula, dok alternativna hipoteza tvrdi da je barem jedan od koeficijenata različit od nule, što bi značilo da postoji značajna veza između varijable odziva i barem jedne nezavisne varijable. Opisani test hipoteza prikazujemo na sljedeći način:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_a : \beta_j \neq 0, j \in \{1, 2, ..., k\}.$$

Odluku o prihvaćanju, odnosno odbacivanju hipoteze donosimo na temelju testne statistike F koja je u ovom slučaju dana s

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}},$$

pri čemu je $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, a $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Primjetimo da je MSE zapravo jednak s^2. Dakle, MSE je nepristrani procjenitelj za σ^2 te vrijedi $\mathbb{E}[MSE] = \sigma^2$. Ako vrijedi nul - hipoteza H_0, može se pokazati da je i $\mathbb{E}[MSR] = \sigma^2$. Sada možemo zaključiti da u slučaju nepostojanja veze između odzivne varijable i varijable poticaja opažena vrijednost F - statistike biti će blizu 1. S druge strane, ako odbacujemo nul - hipotezu u korist alternativne hipoteze H_a vrijedit će $\mathbb{E}\left[\frac{SSR}{k}\right] > \sigma^2$ pa će F poprimiti vrijednost veću od 1. Ako je n velik tada je dovoljno da je vrijednost F - statistike malo veća od 1 da bi sa sigurnošću mogli odbaciti H_0. Obrnuto, ako je n malen tada vrijednost F - statistike mora biti značajno veća od 1 kako bi odbacili nul - hipotezu [5, poglavlje 3]. Dodatno, ako je je hipoteza H_0 istinita, a ϵ_i-evi su normalno distribuirani, F - statistika dolazi iz F distribucije:

$$F = \frac{MSR}{MSE} \overset{\text{H}_0}{\sim} F(k, n-k-1).$$

Za proizvoljne vrijednosti n i k statistički softveri, osim što izračunavaju F - statistiku, pomoću F distribucije mogu izračunati p - vrijednosti za svaku od k nezavisnih varijabli. Tako pomoću p - vrijednosti možemo saznati koje varijable poticaja su statistički značajne za naš model. Ako je p - vrijednost uz i-tu varijablu dovoljno blizu nule, odbaciti ćemo nultu hipotezu i zaključiti da postoji povezanost između i-te varijable i varijable odziva.

Odredimo sada intervale pouzdanosti za parametre β_j, $j \in (0, 1, ..., k)$. Pomoću (2.35) možemo zaključiti da je $(1-\alpha)100\%$ pouzdani interval za β_j dan s:

$$\left\langle \hat{\beta}_j - t_{\frac{\alpha}{2}}\sqrt{v_j}s,\ \ \hat{\beta}_j + t_{\frac{\alpha}{2}}\sqrt{v_j}s \right\rangle.$$

Korištenje F - statistike za testiranje bilo kakve povezanosti između zavisne i nezavisnih varijabli funkcionira kad je broj nezavisnih varijabli malen, odnosno kada je k manji od n. Međutim, često imamo jako velik broj varijabli. Ako je $k > n$ tada postoji više koeficijenata β_j za procjenu nego mjerenja iz kojih ih treba procijeniti. U ovakvom slučaju ne možemo provesti metodu najmanjih kvadrata kako bi dobili odgovarajući višestruki regresijski model, a samim tim ne možemo koristiti F - statistiku [5, poglavlje 3]. Unatoč tome, postoje neke metode (poput *stepwise unaprijed* koja će biti obrađena u sljedećem poglavlju) koje se mogu koristiti kod velikih k-ova.

Poglavlje 3

Metode s pristranim procjenama

U ovom poglavlju razrađujemo neke načine kojima se višestruki linearni regresijski model može poboljšati, zamjenom metode najmanjih kvadrata nekim alternativnim metodama. Zašto bi uopće koristili drugu metodu pored metode najmanjih kvadrata? Vidjet ćemo da će nam te metode dati veću *točnost procjene* i bolju *interpretaciju modela* [5, poglavlje 6].

- *Točnost procjene*: Pod uvjetom da je stvarni odnos varijable odziva i nezavisnih varijabli približno linearan, procjenitelji dobiveni metodom najmanjih kvadrata imati će malu pristranost. Ako je broj mjerenja n puno veći od broja varijabli k, procjenitelji dobiveni metodom najmanjih kvadrata imati će i malu varijancu što će sveukupno dati dobre rezultate u opažanjima. Međutim, ako n nije toliko veći od k, tada može postojati varijabilnost u prilagodbi najmanjih kvadrata, što će rezultirati pretjeranom prilagodbom (eng. *overfitting*) i lošim predviđanjima budućih opažanja. Na kraju, ako je k veći od n, varijanca će ići u beskonačnost i metoda najmanjih kvadrata neće se uopće moći koristiti. Često ćemo, ograničavanjem ili smanjivanjem nekih koeficijenata, moći značajno smanjiti varijancu pod cijenu zanemarivog povećanja pristranosti što nas može dovesti do značajnih poboljšanja u preciznosti predviđanja.
- *Interpretacija modela*: Čest je slučaj u kojem neke od varijabli u višestrukom regresijskom modelu zapravo nisu povezane sa zavisnom varijablom. Uključivanje takvih irelevantnih varijabli dovodi do nepotrebne složenosti modela. Uklanjanjem tih varijabli, odnosno postavljanjem određenih procijenjenih koeficijenata na nulu, možemo dobiti model koji je lakše interpretirati. Vrlo je mala vjerojatnost da će metoda najmanjih kvadrata bilo koji koeficijent procijeniti točno s nulom. Stoga ćemo u ovom poglavlju vidjeti neke druge pristupe koji će služiti za isključivanje irelevantnih varijabli iz modela višestruke linearne regresije.

Dotaknut ćemo se dviju kategorija pristranih metoda. Prvu kategoriju čine metode odabira podskupa (eng. *subset selection*) među kojima ćemo istaknuti odabir najboljeg pod-

skupa (eng. *best subset selection*), stepenastu selekciju unaprijed (eng. *forward stepwise selection*), stepenastu selekciju unazad (eng. *backward stepwise selection*) te etapnu selekciju unaprijed (eng. *forward stagewise selection*). Druga kategorija su metode sažimanja (eng. *shrinkage methods*) kod kojih ćemo spomenuti ridge i lasso regresiju.

3.1 Metode odabira podskupa

Odabir najboljeg podskupa

Prva od metoda odabira podskupa koju ćemo opisati je metoda odabira najboljeg podskupa. Kroz tu metodu oblikujemo k različitih modela koji sadrže točno jednu nezavisnu varijablu, zatim $\binom{k}{2} = \frac{k(k-1)}{2}$ modela koji sadrže točno dvije nezavisne varijable, i tako dalje. Za svaki $l \in \{0, 1, ..., k\}$ će se među $\binom{k}{l}$ l-članih podskupova koji sadrže nezavisne varijable odabrati onaj koji će imati najmanju sumu kvadrata reziduala. Time ćemo problem s 2^k mogućih modela reducirati na $k+1$ modela od kojih moramo odabrati jedan najbolji. Kako bi odredili koji je model najbolji, možemo koristiti različite statistike za procjenu kvalitete: prilagođeni R_a^2, Mallowljev C_p, AIC (*Akaike information criterion*) ili BIC (*Bayesian information criterion*) (Detaljnije definicije i rezultati su u [5]). Opisano možemo sažeti u sljedeći algoritam:

1. Označi s $\mathcal{M}_0$ nul - model koji ne sadrži nijednu nezavisnu varijablu.
2. Za $l = 1, 2, ..., k$
   a) Oblikuj svih $\binom{k}{l}$ modela koji sadrže točno l prediktora.
   b) Izaberi najbolji od $\binom{k}{l}$ modela i nazovi ga $\mathcal{M}_l$ pri čemu je *najbolji* definiran kao onaj s najmanjim SSE ili ekvivalentno onaj s najvećim R^2.
3. Izaberi najbolji model među $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_k$ koristeći prilagođeni R_a^2, C_p, AIC ili BIC.

Iako se odabir najboljeg podskupa čini kao jednostavan pristup, on ima neka ograničenja. Naime, broj mogućih modela koje treba razmotriti raste eksponencijalno s porastom broja nezavisnih varijabli. Tako na primjer za $k = 10$ radimo s $2^{10} = 1024$ modela, a za $k = 20$ s $2^{20} = 1048576$ modela. Zbog toga, za k-ove veće od 40 odabir najboljeg podskupa postaje neizvodljiv čak i uz iznimno brza moderna računala [5, poglavlje 6]. Iz tog razloga, stepwise metode, koje istražuju ograničeniji skup modela, dobre su alternative odabiru najboljeg podskupa.

Stepenasta selekcija unaprijed

Stepenasta selekcija unaprijed započinje s modelom koji sadrži samo slobodan član, a zatim dodaje jednu po jednu varijablu sve dok svi prediktori ne budu u modelu. U svakom koraku modelu se dodaje onaj prediktor koji daje najbolje poboljšanje prilagodbi. Postupak opet možemo prikazati algoritmom:

1. Označi s $\mathcal{M}_0$ nul - model bez ijednog prediktora.
2. Za $l = 0, 1, ..., k-1$:
   a) Razmotri svih $k-l$ modela koji proširuju model $\mathcal{M}_l$ za jedan dodatni prediktor.
   b) Izaberi najbolji od $k-l$ modela i nazovi ga $\mathcal{M}_{l+1}$ pri čemu je *najbolji* definiran kao onaj s najmanjim SSE ili najvećim R^2.
3. Izaberi najbolji model među $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_k$ koristeći prilagođeni R_a^2, C_p, AIC ili BIC.

Za razliku od odabira najboljeg podskupa, koji je uključivao oblikovanje 2^k modela, stepenasta selekcija unaprijed uključuje stvaranje nul - modela zajedno s još $k-l$ modela u l-toj iteraciji, za $l = 0, 1, ..., k-1$. To ukupno daje $1 + \sum_{l=0}^{k-1}(k-l) = 1 + \frac{k(k+1)}{2}$ modela što čini popriličnu razliku. Na primjer, za $k = 20$, odabir najboljeg podskupa zahtjeva prlagodbu 1048576 modela, dok se stepenasta selekcija unaprijed svodi samo na njih 211.

U koraku 2.b) moramo identificirati najbolji model od njih $k-l$ koji proširuju $\mathcal{M}_l$ s jednom dodatnom varijablom. To jednostavno možemo učiniti odabirom modela s najmanjom sumom kvadrata reziduala ili najvišim koeficijentom R^2. Međutim, u trećem koraku moramo identificirati najbolji model među skupom modela s različitim brojem varijabli. Taj problem je izazovniji te se u tom koraku možemo osloniti na izračunavanje neke od statistika kao što je prilagođeni R_a^2.

Iako smo sada pokazali da je stepenasta selekcija brža i praktičnija od metode odabira najboljeg podskupa, ona nam ipak ne garantira uvijek pronalazak najboljeg od svih 2^k modela. Naime, pretpostavimo da nam je dan set podataka s $k = 3$ nezavisne varijable te da je najbolji model s jednom varijablom onaj koji sadrži x_1, a najbolji model s dvije varijable onaj koji sadrži varijable x_2 i x_3. U tom slučaju stepenasta selekcija unaprijed neće uspjeti pronaći najbolji mogući model s dvije varijable jer će, zbog toga što $\mathcal{M}_1$ sadrži x_1, $\mathcal{M}_2$ također morati sadržavati x_1 i još jednu varijablu.

Stepenasta selekcija unazad

Obrnuto od selekcije unaprijed, stepenasta selekcija unazad započinje s punim regresijskim modelom koji sadrži svih k nezavisnih varijabli. Zatim uklanja najmanje korisnu varijablu na način prikazan u sljedećem algoritmu:

1. Označi s $\mathcal{M}_k$ puni model sa svih k prediktora.
2. Za $l = k, k-1, ..., 1$:
   a) Razmotri svih l modela koji sadrže sve osim jednog prediktora iz $\mathcal{M}_l$ (modeli s ukupno $l-1$ prediktora).
   b) Izaberi najbolji od l modela i nazovi ga $\mathcal{M}_{l-1}$ pri čemu je *najbolji* definiran kao onaj s najmanjim SSE (ii najvećim R^2).
3. Izaberi najbolji model među $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_k$ koristeći prilagođeni R_a^2, C_p, AIC ili BIC.

Stepenasta selekcija unazad rješenje također traži u skupu od $1 + \frac{k(k+1)}{2}$ modela pa se i ona može primjenjivati u slučajevima kada je k prevelik da bi se proveo odabir najboljeg podskupa. Međutim, kao i stepenasta selekcija unaprijed, ne jamči pronalazak najboljeg rješenja. Uz to, selekcija unazad zahtjeva da je broj mjerenja n veći od broja nezavisnih varijabli k, dok se selekcija unaprijed može koristiti kada je $n < k$, što ju čini jedinom održivom metodom odabira podskupa u slučaju jako velikih k-ova.

Hibridni pristup

Odabir najboljeg podskupa, stepenasta selekcija unaprijed i stepenasta selekcija unazad općenito daju slične, ali ne nužno identične modele. Tako se javlja još jedna alternativa u obliku hibridne verzije stepenaste selekcije unaprijed i unazad. Kao i selekcija unaprijed, počinje bez nezavisnih varijabli i odabire jednu po jednu varijablu za ulazak u model otprilike na isti način. No nakon ulaska svake nove varijable ova metoda ispituje svaku od varijabli koja je već uključena u model kako bi se provjerilo treba li koju od njih izbrisati, kao u selekciji unazad. Na taj način uklanjaju se sve varijable koje više ne pružaju poboljšanje modela. Tako ovaj pristup pokušava oponašati odabir najboljeg podskupa zadržavajući prednosti koje donose stepenaste selekcije.

Etapna regresija unaprijed

Posljednja metoda odabira podskupa koju ćemo spomenuti je etapna selekcija unaprijed. Ona započinje kao i stepenasta selekcija unaprijed, sa slobodnim članom jednakim $\bar{y}$ i centriranim prediktorima čiji su koeficijenti u početku jednaki nula. Algoritam etapne selekcije unaprijed u svakom koraku identificira onu varijablu koja je najviše korelirana s trenutnim rezidualom. Nakon toga izračunava koeficijent jednostavne linearne regresije od reziduala na toj varijabli, a zatim ga dodaje trenutnom koeficijentu za tu varijablu. Postupak se nastavlja sve dok nijedna od varijabli nije u korelaciji s rezidualima.

Nakon svih obrađenih metoda, treba spomenuti nekoliko upozorenja na kraju. Unatoč njihovim prednostima, pri korištenju ovih procedura treba paziti da se ne dogodi ispuštanje važne varijable. Na primjer, ako je svrha nekog istraživanja utvrditi vezu između cijene i prodaje nekog proizvoda, bilo bi besmisleno ispustiti varijablu cijene samo zato što neka od metoda to preporučuje (takva loša preporuka može biti posljedica loše formirane matrice X ako je stupac male duljine ili je jako povezan s nekim drugim stupcem) [8, poglavje 11]. Također, istraživači čiji je primarni cilj predviđanje trebali bi se pobrinuti da ne odustaju od nezavisne varijable koju je lako moći predvidjeti u korist neke koja je problematičnija. U konačnici, istraživač i njegova intuicija trebali bi biti na prvom mjestu kod donošenja odluka o odabiru varijabli, ispred metoda koje provode računala.

3.2 Metode sažimanja

Metode opisane u prethodnim odjeljcima zadržavaju odabrani podskup prediktora te odbacuju preostale. Time dobivamo diskretan proces kojim se svaka varijabla ili zadržava ili odbacuje što često rezultira visokom varijancom. Zbog toga ponekad želimo posegnuti za drugom vrstom metoda, metodama sažimanja. Možemo oblikovati model koji sadrži svih k varijabli koristeći tehnike koje ograničavaju i reguliraju procjene koeficijenata, odnosno koje sažimaju procjene koeficijenata prema nuli [5, poglavlje 6.1]. Takvi pristupi su više neprekidni i mogu značajno smanjiti varijancu stoga u nastavku donosimo dva najpoznatija načina sažimanja: ridge i lasso regresiju.

Ridge regresija

Prisjetimo se, metoda najmanjih kvadrata donosi procjene parametra $\beta_0, \beta_1, ..., \beta_k$ minimiziranjem sume kvadrata reziduala

$$SSE = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 .$$

Ridge regresija vrlo je slična metodi najmanjih kvadrata. Procjena koeficijenata se također zasniva na minimiziranju, samo što je izraz koji se minimizira malo drugačije forme. Preciznije, procjene koeficijenata $\boldsymbol{b}^r$ dobivene ridge regresijom su vrijednosti dobivene minimiziranjem

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 = SSE + \lambda \sum_{j=1}^{k} \beta_j^2, \tag{3.1}$$

pri čemu je $\lambda \geq 0$ parametar složenosti. Izraz (3.1) sadrži dva kriterija. Kao i kod najmanjih kvadrata, ridge regresija želi pronaći procjene koeficijenata koje dobro odgovaraju podacima, čineći SSE malim. Međutim, drugi član, $\lambda \sum_{j=1}^{k} \beta_j^2$, kojeg još nazivamo penaliziranjem, je malen kada su $\beta_1, ..., \beta_k$ blizu nule pa on ima učinak sažimanja procjena β_j prema nuli. Parametar λ služi kako bi kontrolirao relativan učinak ova dva člana na procjene koeficijenata regresije. Kada je $\lambda = 0$ penaliziranje nema nikakvog učinka i ridge regresija dat će istu procjenu kao i procjena metodom najmanjih kvadrata. No kada $\lambda \to \infty$, utjecaj penaliziranja raste i procjene koeficijenata ridge regresije približit će se nuli.

Primijetimo da se u (3.1) penaliziranje primjenjuje na $\beta_1, ..., \beta_k$, ali ne i na koeficijent β_0. Želimo smanjiti procijenjenu povezanost svake varijable s odgovorom, no ne želimo smanjiti i slobodni član. Ridge regresija često se primjenjuje na centriranom modelu u kojem se svaki x_{ij} zamjenjuje s $x_{ij} - \bar{x}_j$, pri čemu je $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$. Tada β_0 procjenjujemo s $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$. Matrica X u tom slučaju ima k, umjesto $k+1$ stupaca. Sada je matrična forma od (3.1) dana s

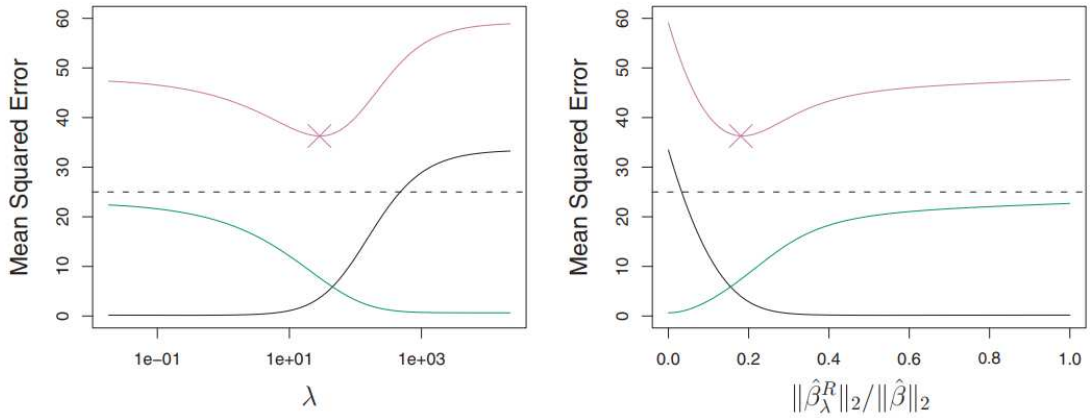
$$(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \tag{3.2}$$

te rješenje ridge regresije ima oblik

$$\boldsymbol{b}^r = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}, \tag{3.3}$$

pri čemu je I $k \times k$ jedinična matrica. Zaključujemo da je procjena $\boldsymbol{b}^r$, kao i $\boldsymbol{b}$, linearna funkcija od $\boldsymbol{y}$. U (3.3) vidimo da se, prije invertiranja, dijagonali matrice $X^T X$ dodaje pozitivna konstanta. To čini problem nesingularnim, čak i ako $X^T X$ nije punog ranga, što je bila glavna motivacija za ridge regresiju kada se je prvi put uvela u statistiku 1970. godine [2, poglavlje 3].

Prednost ridge regresije nad najmanjim kvadratima ukorijenjena je u kompromisu između varijance i pristranosti. Kako se λ povećava, fleksibilnost prilagodbe ridge regresije se smanjuje, tražeći smanjenje varijance, ali povećanje pristranosti. Spomenuto je prikazano na lijevom dijelu slike 3.1, koristeći set podataka od $n = 50$ mjerenja s $k = 45$ nezavisnih varijabli.



Slika 3.1: *kvadratna pristranost (crno), varijanca (zeleno) i MSE (ljubičasto) za predviđanja ridge regresije nad danim setom podataka prikazani kao funkcije od* λ *i* $\frac{\|b^r\|_2}{\|b\|_2}$*. Horizontalna isprekidana linija označava najmanji mogući MSE, ljubičasti križić označava ridge regresijski model za koji je MSE najmanji.* Izvor [5]

Zelena krivulja na lijevom dijelu slike 3.1 predstavlja varijancu predviđanja ridge regresije kao funkciju od λ. Kod procjene koeficijenata najmanjih kvadrata, koji odgovaraju ridge regresiji s parametrom $\lambda = 0$, varijanca je visoka, ali nema pristranosti. No, kako se λ povećava, smanjenje procjene ridge koeficijenata dovodi do značajnog smanjenja varijance, na račun blagog povećanja pristranosti. Za vrijednosti λ do oko 10, varijanca se brzo smanjuje, s vrlo malim porastom pristranosti koju prikazujemo crnom krivuljom. Prisjetimo se da je srednje kvadratna pogreška MSE, prikazana ljubičastom krivuljom, zbroj varijance i kvadratne pristranosti. Ona porastom λ od 0 do otprilike 10 značajno pada. Nakon što λ prijeđe taj interval, smanjenje varijance usporava, a sažimanje nad koeficijentima uzrokuje njihovo značajno podcjenjivanje, što rezultira velikim povećanjem u pristranosti. Minimalni MSE postiže se otprilike za $\lambda = 30$ i označen je ljubičastim križićem. Zanimljivo je da je, zbog svoje visoke varijance, MSE povezan s metodom najmanjih kvadrata (kada je $\lambda = 0$), gotovo jednako visok kao i kod nul - modela kod kojeg su procjene svih koeficijenata jednake nuli, u slučaju kad $\lambda \to \infty$.

Desni dio slike 3.1 prikazuje kvadratnu pristranost, varijancu i srednje kvadratnu pogrešku u odnosu na l_2 normu procjene koeficijenata ridge regresije podijeljenu s l_2 normom procjene koeficijenata najmanjih kvadrata. U ovom slučaju, pomicanjem slijeva nadesno, prilagodba postaje fleksibilnija te pristranost pada, a varijanca raste.

Općenito, u situacijama u kojima je odnos između varijable odziva i prediktora blizu linearnom, procjene metodom najmanjih kvadrata imat će nisku pristranost, ali mogu imati visoku varijancu. To znači da bi mala promjena u podacima koje smo koristili za stvaranje

modela mogla uzrokovati veliku promjenu u procjenama koeficijenata najmanjih kvadrata. Konkretno, kada je broj nezavisnih varijabli k, gotovo jednak kao i broj opažanja n, kao i u primjeru vezanom za sliku 3.1, procjene najmanjih kvadrata biti će vrlo varijabilne. Ako je $k > n$ znamo da procjene najmanjih kvadrata neće biti jedinstvene, dok ridge regresija i tada može dobro raditi, mijenjajući malo povećanje u pristranosti za veliko smanjenje u varijanci.

Lasso regresija

Ridge regresija ima jedan očiti nedostatak. Za razliku od metoda odabira podskupa, koje će odabrati modele koji uključuju samo podskup varijabli, ridge regresija će uključivati svih k prediktora u konačnom modelu. Penaliziranje dano s $\lambda \sum_{j=1}^{k} \beta_j^2$ u (3.1) će smanjivati sve koeficijente prema nuli, ali neće postaviti niti jednog od njih točno na nulu (osim ako je $\lambda = \infty$). Povećanje vrijednosti parametra λ težit će smanjenju veličine koeficijenata, ali neće rezultirati isključivanjem nijedne od varijabli. To možda neće biti problem u točnosti predviđanja, no može stvoriti izazov u interpretaciji modela u slučajevima u kojima je broj varijabli velik.

Lasso regresija je relativno nova alternativa ridge regresiji koja nadilazi prethodno opisani nedostatak. Lasso koeficijenti, $\boldsymbol{b}^l$, minimiziraju sljedeći izraz:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{k} \left| \beta_j \right| = SSE + \lambda \sum_{j=1}^{k} \left| \beta_j \right|. \tag{3.4}$$

Uspoređujući (3.1) s (3.4), vidimo da lasso i ridge regresija imaju slične formulacije. Jedina razlika je u tome da smo član β_j^2 kod penaliziranja u ridge regresiji zamijenili s $\left|\beta_j\right|$ u lasso penaliziranju.

Kao i ridge regresija, lasso regresija smanjuje procjene koeficijenata prema nuli. Međutim, u lasso slučaju, penaliziranje može imati učinak prisiljavanja nekih procjena da budu točno nula, kada je parametar λ dovoljno velik. Stoga možemo reći da lasso regresija vrši neku vrstu kontinuiranog odabira podskupa [2, poglavlje 3]. Kao rezultat toga, modele generirane lasso regresijom možemo lakše interpretirati.

Alternativna formulacija Ridge i Lasso regresije

Može se pokazati da procjene koeficijenata lasso i ridge regresije rješavaju sljedeće probleme:

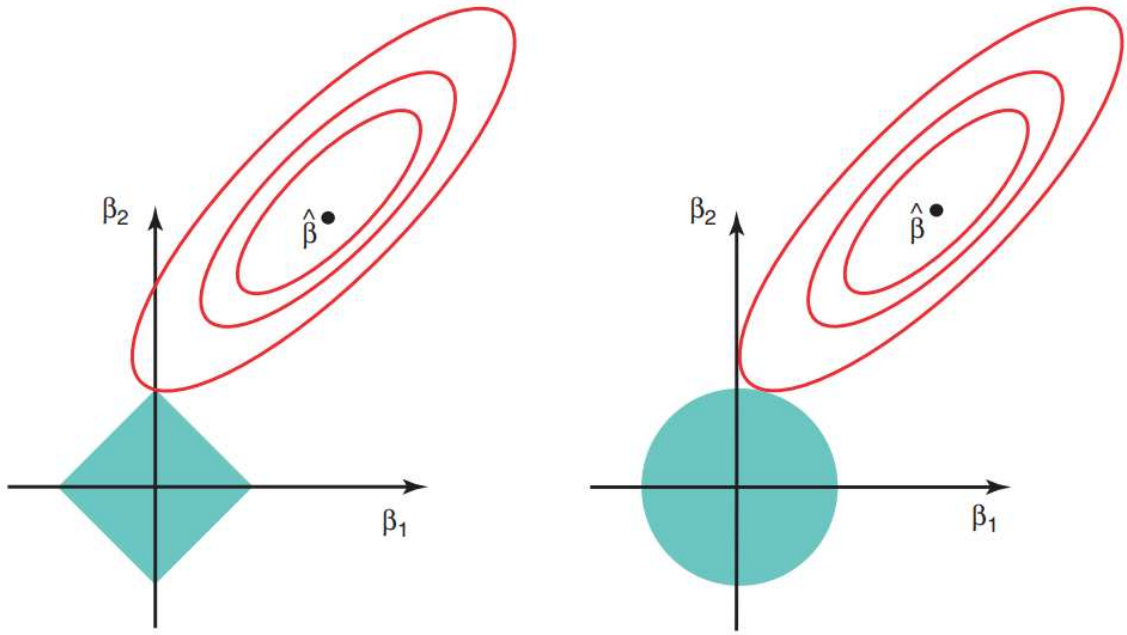
$$\underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right\}, \quad \text{pod uvjetom} \quad \sum_{j=1}^{k} \left|\beta_j\right| \leq t \tag{3.5}$$

$$\underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right\}, \quad \text{pod uvjetom} \quad \sum_{j=1}^{k} \beta_j^2 \leq t. \tag{3.6}$$

Drugim riječima, za svaku vrijednost λ postoji neki t takav da problem minimiziranja (3.4) i problem dan s (3.5) kao svoje rješenje daju iste procjene lasso koeficijenata. Analogno, za svaki λ postoji odgovarajući t takav da problem minimiziranja (3.1) i problem dan s (3.6) kao svoje rješenje daju iste procjene ridge koeficijenata.

O problemu (3.5) možemo razmišljati na sljedeći način: provodimo lasso regresiju, pokušavajući pronaći skup procjena koeficijenata koji vode do najmanje sume kvadrata reziduala, pod uvjetom da postoji ograničenje t u tome koliko velika smije biti $\sum_{j=1}^{k} \left|\beta_j\right|$. Kada je t iznimno velik, ograničenje nije jako restriktivno pa procjene koeficijenata mogu biti velike. Uz to, ako je t dovoljno velik da rješenje dobiveno metodom najmanjih kvadrata spada unutar ograničenja, tada će (3.5) jednostavno dati rješenje metode najmanjih kvadrata. Suprotno tome, ako je t mali, $\sum_{j=1}^{k} \left|\beta_j\right|$ također mora biti mala kako bi se izbjeglo kršenje ograničenja. Slično, (3.6) ukazuje na to da izvođenjem ridge regresije tražimo skup procjena koeficijenata tako da je suma kvadrata reziduala najmanja moguća, zahtjevajući da $\sum_{j=1}^{k} \beta_j^2$ ne prelazi ograničenje t.

Ako je $k = 2$, lasso regresija zapisana pomoću (3.5) implicira da, među svim točkama koje leže unutar kvadrata definiranog s $|\beta_1| + |\beta_2| \leq t$, koeficijenti dobiveni lasso regresijom daju najmanji SSE. Analogno, od svih točaka koje leže unutar kruga definiranog s $\beta_1^2 + \beta_2^2 \leq t$, koeficijenti procijenjeni ridge regresijom rezultiraju s najmanjim SSE. Opisano možemo vidjeti na slici 3.2.



Slika 3.2: *Područja određena ograničenjem (plavo) i konture od SSE (crveno) za lasso (lijevo) i ridge (desno) regresiju s dvije nezavisne varijable.* Izvor [5]

Plavi kvadrat i krug označavaju lasso, odnosno ridge ograničena područja. S $\hat{\beta}$ označeno je rješenje dobiveno metodom najmanjih kvadrata. Na ovoj slici $\hat{\beta}$ leži van plavih područja pa procjene dobivene lasso i ridge regresijom neće biti jednake procjeni dobivenoj metodom najmanjih kvadrata, no, u slučaju velikih vrijednosti t, $\hat{\beta}$ može se nalaziti unutar ograničenih područja. Svaka od crvenih elipsi sa središtem u $\hat{\beta}$ predstavlja skup točaka koje daju isti SSE. Kako se elipse proširuju, SSE raste te će u jednom trenutku neka elipsa ući u ograničeno područje. (3.5) i (3.6) govore nam da su procjene koeficijenata lasso i ridge regresije dane prvom točkom u kojoj se elipsa i ograničeno područje dodiruju. Budući da ridge regresija ima područje u obliku kruga (za $k = 3$ oblik sfere itd.), bez oštrih krajeva, dodirna točka općenito se neće nalaziti na osi pa će procjene koeficijenata biti različite od nula. S druge strane, ograničenje lasso regresije ima oštrih krajeva koji se nalaze na osima te će elipse često dodirivati ograničeno područje baš u njima. U takvom slučaju, neki od koeficijenta biti će nula. Na slici 3.2 elipsa i kvadrat dodiruju se na vertikalnoj osi pa će β_1 biti 0, a konačni model sadržavati će samo β_2.

Poseban slučaj

U ovom odjeljku detaljnije razmatramo jednostavan slučaj regresije u kojem je $n = k$ te je X jedinična matrica, odnosno ulazni podaci su ortonormirani. S ovim pretpostavkama, problem pronalaska koeficijenata metodom najmanjih kvadrata pojednostavljuje se na minimiziranje sume dane s

$$\sum_{j=1}^{k}(y_j - \beta_j)^2.$$

Tada za rješenje $\boldsymbol{b} = [\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k]$ metode najmanjih kvadrata vrijedi

$$\hat{\beta}_j = y_j, \quad j = 1, 2, ..., k. \tag{3.7}$$

Nadalje, ridge regresijom želimo pronaći $\beta_1, \beta_2, ..., \beta_k$ takve da minimiziraju

$$\sum_{j=1}^{k}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{k}\beta_j^2.$$

Analogno, lasso regresijom želimo pronaći koeficijente koji minimiziraju

$$\sum_{j=1}^{k}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{k}\left|\beta_j\right|.$$

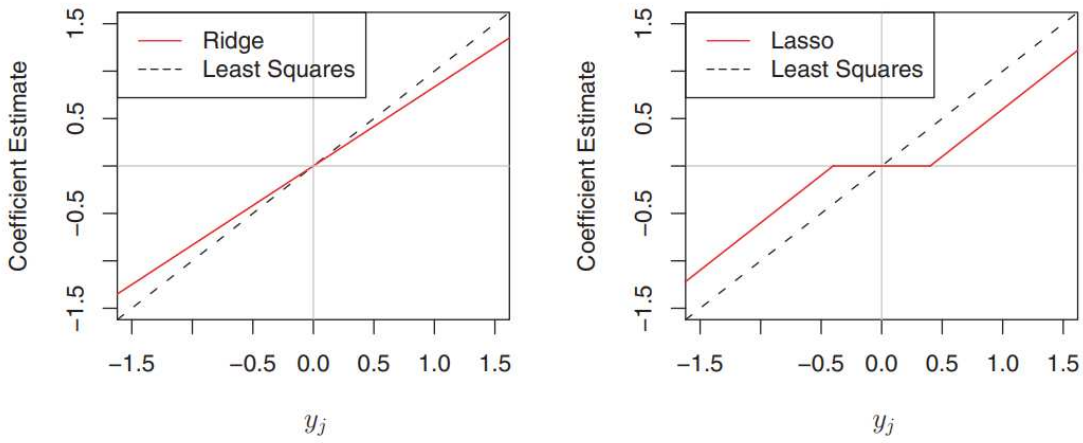
Može se pokazati da su u ovom slučaju procjene koeficijenta ridge regresijom jednake skaliranoj procjeni koeficijenata najmanjih kvadrata. Točnije, vrijedi

$$\hat{\beta}_j^{\,r} = \frac{y_j}{1 + \lambda}, \quad j = 1, 2, ..., k. \tag{3.8}$$

Također, za lasso procjene vrijedi

$$\hat{\beta}_j^{\,l} = \begin{cases} y_j - \frac{\lambda}{2}, & \text{ako } \; y_j > \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2} & \text{ako } \; y_j < -\frac{\lambda}{2} \\ 0 & \text{ako } \; \left|y_j\right| \leq \frac{\lambda}{2} \end{cases}, \quad j = 1, 2, ..., k. \tag{3.9}$$

Navedeno možemo vidjeti na slici 3.3.



Slika 3.3: *Prikaz ridge i lasso procjena koeficijenata u slučaju $m = k$ i $X = I$.* Izvor [5]

Ovdje se mogu primijetiti razlike u sažimanju ridge i lasso regresijom. U ridge regresiji se svaka procjena koeficijenata najmanjih kvadrata smanjuje za isti omjer, Nasuprot tome, lasso regresija smanjuje svaki koeficijent najmanjih kvadrata za konstantan iznos, $\frac{\lambda}{2}$. Dodatno, koeficijenti najmanjih kvadrata koji su po apsolutnoj vrijednosti manji od $\frac{\lambda}{2}$ smanjeni su točno na nulu.

U slučaju da je matrica X općenitije zadana, priča postaje kompliciranija, ali glavne ideje prikazane ovim primjerom ostaju zadržane: ridge regresija smanjuje svaki oblik podataka u jednakom omjeru, dok lasso skuplja koeficijente prema nuli za sličan iznos, a dovoljno male koeficijente izjednačava s nulom.

Poglavlje 4

Primjena linearne regresije na podacima o COVID-u

COVID - 19 smatra se jednom od najvećih pandemija modernog doba. Utjecala je na brojne sfere ljudskog života poput zdravlja, obrazovanja, zapošljavanja, gospodarstva, turizma. Trebat će proći puno vremena da se ljudi oporave od svih negativnih učinaka prouzročenih ovom pandemijom te da se svijet vrati u normalu. Ono što se nažalost ne može vratiti su brojni ljudski životi izgubljeni tijekom širenja pandemije. Upravo zbog toga, važno je razumjeti kako različiti socio - ekonomski, demografski i zdravstveni faktori mogu utjecati na porast ili smanjenje broja preminulih od ovog virusa. U tu svrhu ćemo u ovom poglavlju provesti analizu seta podataka preuzetih sa [9] te na temelju njih razvijati modele za predviđanje broja oboljelih i umrlih. Navedeni postupak biti će sproveden u programskom jeziku R prateći teorijske rezultate iz prethodnih poglavlja.

Opis varijabli

Navedeni podaci sadržavaju brojne varijable od kojih ćemo mi izdvojiti nekoliko najzanimljivijih i pratiti njihov utjecaj na zavisnu varijablu. Uz to, odbacujemo mjerenja kod kojih nije zabilježena vrijednost neke od varijabli. U konačnici dobivamo 163 mjerenja, pri čemu svako mjerenje predstavlja jednu državu. Za zavisnu varijablu ćemo prvo uzeti broj zaraženih, a nakon toga broj umrlih te ćemo pokušati uspostaviti linearnu vezu sa sljedećih 8 nezavisnih varijabli:

- x_1 =gustoća populacije
- x_2 =prosječne godine
- x_3 =udio stanovnika starijih od 65

- x_4 =GDP
- x_5 =indeks razvijenosti
- x_6 =srčani bolesnici
- x_7 =dijabetičari
- x_8 =indeks strogoće mjera

Potrebno je napraviti neke transformacije na zavisnim varijablama kako bi podaci bili normalizirani. Točnije, obje varijable, i broj zaraženih i broj umrlih, dijelimo s ukupnim brojem stanovnika u državama kako bi zapravo dobili udio umrlog, odnosno zaraženog stanovništva. Na ovaj način dobivamo smislenije modele jer bi inače regresija upućivala na linearnu zavisnost u kojoj države s većim brojem stanovnika imaju veći broj umrlih i zaraženih. Detaljniji opisi svih varijabli i potrebne transformacije sažeti su u tablici prikazanoj na slici 4.1.

POČETNA VARIJABLA	OPIS VARIJABLE	TRANSFORMACIJA
broj umrlih	ukupan broj umrlih od COVID-a u nekoj državi	broj umrlih/broj stanovnika
broj zaraženih	ukupan broj zaraženih COVID-om u nekoj državi	broj zaraženih/broj stanovnika
gustoća populacije	ukupan broj stanovnika neke zemlje podijeljen njezinom površinom (mjerenom u km2)	nepromijenjeno
prosječne godine	prosjek godina stanovnika neke države	nepromijenjeno
stariji od 65	udio stanovnika neke države stariji od 65 godina	nepromijenjeno
GDP	GDP po stanovniku neke države (GDP per capita)	nepromijenjeno
indeks razvijenosti	mjeri postignuća u tri dimenzije ljudskog razvoja: dug i zdrav život, znanje, kvalitetan životni standard, iskazano na skali od 0 (nerazvijeno) do 1 (najrazvijenije)	nepromijenjeno
srčani bolesnici	broj umrlih od srčanih bolesti na 100 000 osoba u nekoj državi	nepromijenjeno
dijabetičari	udio dijabetičara u nekoj državi	nepromijenjeno
indeks strogoće mjera	mjera temeljena na 9 pokazatelja uključujući npr. zatvaranje škola, zatvaranje radnih mjesta i zabrane putovanja, iskazano na skali od 0 (najmanje strogo) do 100 (najstrože)	nepromijenjeno

Slika 4.1: Opis i transformacije varijabli

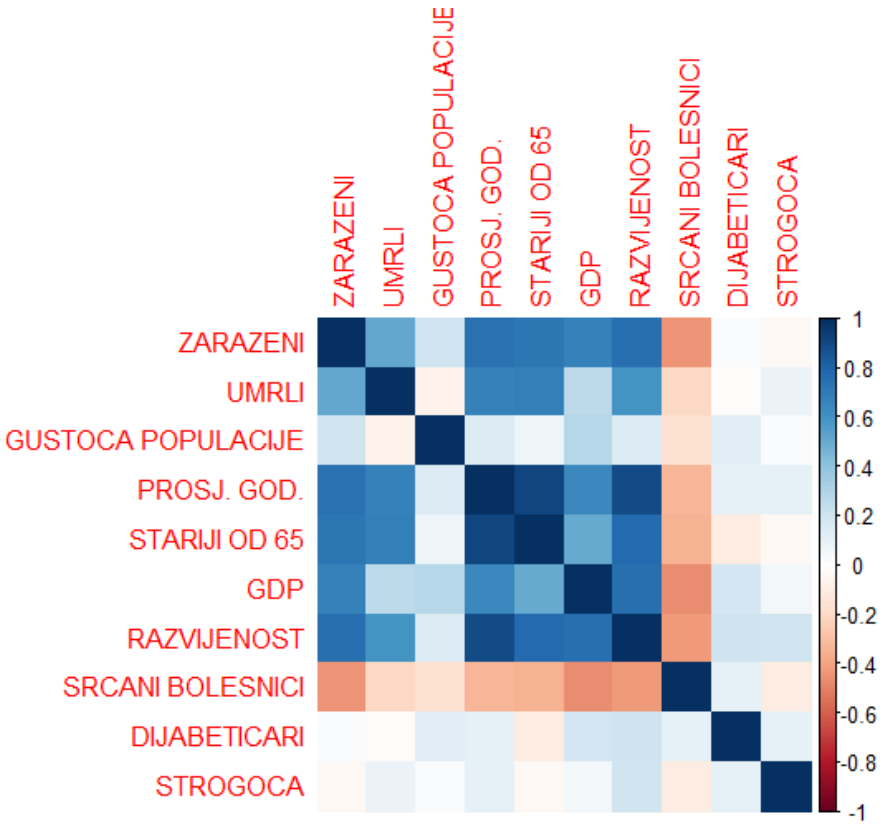
Gradimo modele oblika

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \epsilon$$

procjenjujući koeficijente β_i, $i = 0, 1, ..., 8$, pri čemu su nam zavisne varijable:

- y = udio zaraženih = $\frac{\text{broj zaraženih u državi}}{\text{ukupan broj stanovnika u državi}}$
- y = udio umrlih = $\frac{\text{broj umrlih u državi}}{\text{ukupan broj stanovnika u državi}}$.

Za početak možemo pogledati stupanj linearne povezanosti među varijablama računajući Pearsonov koeficijent korelacije. Rezultate prikazujemo vizualiziranjem korelacijske matrice na način prikazan slikom 4.2. Sliku možemo tumačiti u tri smjera: kvadratići obojani nijansama plave upućuju da je koeficijent blizu 1 te je tada veza između varijabli pozitivna, kvadratići u crvenim nijansama upućuju na koeficijent blizu -1 te je tada veza među varijablama negativna i na kraju kvadratići u svijetlim nijansama približno bijele boje ukazuju da je koeficijent blizu 0 te da je linearna povezanost među varijablama slaba. Ako promatramo odnos zavisne varijable udio umrlih s ostalim varijablama, možemo uočiti da je ona negativno korelirana s gustoćom populacije, brojem umrlih od srčanih bolesti i postotkom dijabetičara. Osim toga, vidimo da udio umrlih ima najveći stupanj povezanosti s udjelom zaraženih, prosječnim godinama, udjelom starijih od 65 i indeksom razvijenosti. Promatrajući zavisnu varijablu udio zaraženih zaključujemo da je ona negativno korelirana s brojem umrlih od srčanih bolesti i indeksom strogoće mjera. Najveći stupanj povezanosti ima s prosječnim godinama, udjelom starijih od 65, GDP-om i indeksom razvijenosti. Nadalje, uočavamo da su nezavisne varijable indeks razvijenosti i prosječne godine visoko pozitivno korelirane (0.899), kao i prosječne godine i udio starijih od 65 (0.915, što smo mogli i pretpostaviti bez računanja koeficijenta). Kod modeliranja ćemo pripaziti da nam se jako korelirane nezavisne varijable ne nalaze zajedno u istom modelu kako bi izbjegli problem multikoliearnosti.



Slika 4.2: Korelacije među varijablama

Udio zaraženih kao zavisna varijabla

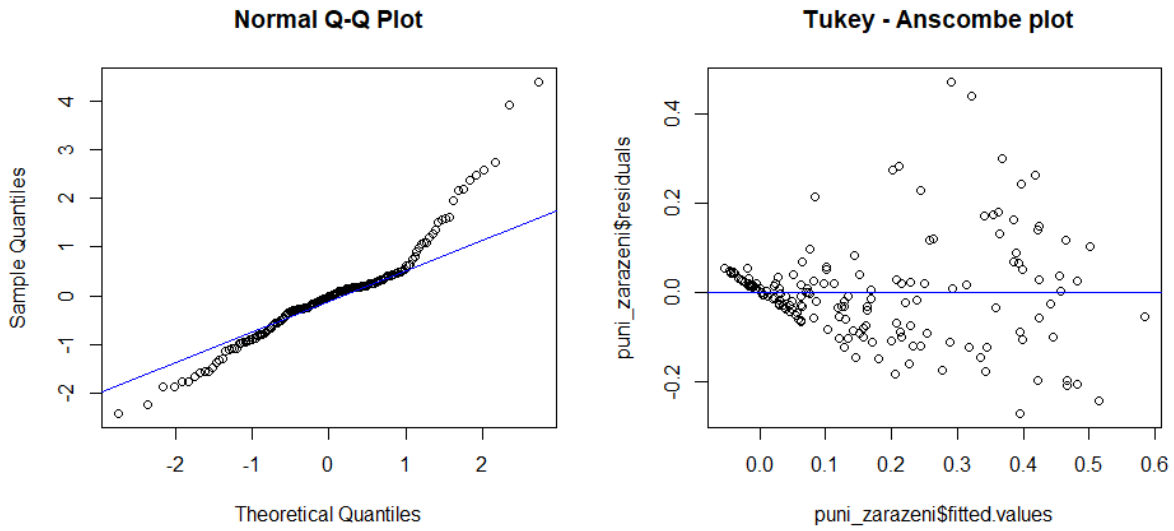
Prvo gradimo model za predviđanje udjela zaraženih ovisno o svih osam nezavisnih varijabli. Rezultate prikazane na slici 4.3 dobili smo korištenjem funkcije *summary()* u koju smo kao argument stavili rezultat koji nam je prethodno dala funkcija *lm()* koja služi za generiranje traženog modela.

```
> puni_zarazeni<-lm(zarazeni/populacija~gustoca_pop+prosjek_god+stariji_od_65+gdp+
+                     razvijenost+srcani+dijabeticari+strogoca)
> summary(puni_zarazeni)

Call:
lm(formula = zarazeni/populacija ~ gustoca_pop + prosjek_god +
    stariji_od_65 + gdp + razvijenost + srcani + dijabeticari +
    strogoca)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27095 -0.05767 -0.00224  0.03449  0.47342

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.233e-01  7.927e-02  -1.555  0.12199
gustoca_pop    1.646e-05  1.447e-05   1.137  0.25710
prosjek_god   -7.767e-04  4.003e-03  -0.194  0.84642
stariji_od_65  1.084e-02  4.596e-03   2.359  0.01957 *
gdp            2.198e-06  7.809e-07   2.814  0.00553 **
razvijenost    4.184e-01  1.711e-01   2.446  0.01558 *
srcani        -1.353e-04  8.755e-05  -1.546  0.12427
dijabeticari  -1.571e-03  2.576e-03  -0.610  0.54279
strogoca      -1.845e-03  9.885e-04  -1.866  0.06394 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1133 on 154 degrees of freedom
Multiple R-squared:  0.6795,    Adjusted R-squared:  0.6628
F-statistic: 40.81 on 8 and 154 DF,  p-value: < 2.2e-16
```

Slika 4.3: Rezultati za potpuni model sa zavisnom varijablom udio zaraženih

Na danoj slici možemo iščitati procijenjene koeficijente uz svaku varijablu. Osim toga, vidimo da je koeficijent determinacije R^2 jednak 0.6795, a prilagođeni R^2 je 0.6628 što je zadovoljvajuće. Nadalje, promatrajući stupac s danim p - vrijednostima zaključujemo da su najznačajnije varijable udio starijih od 65, GDP, indeks razvijenosti i indeks strogoće mjera.

Da bi mogli prihvatiti model, moramo provjeriti da li on zadovoljava pretpostavke obuhvaćene Gauss-Markovljevim uvjetima. U tu svrhu ćemo uz svaki model crtati pripadni QQ graf i rezidual-fit graf koji se još naziva Tukey - Anscombe graf. Prvi od njih dobit ćemo crtanjem standardiziranih reziduala u odnosu na kvantile standardne normalne distribucije i tako ispitati jesu li greške normalno distribuirane. Rezidual-fit graf prikazivati će odnos reziduala i predviđenih vrijednosti. Na temelju njega moći ćemo ispitati da li je $\mathbb{E}[\epsilon] = 0$, odnosno da li je varijanca konstantna (homoskedastičnost).



Slika 4.4: QQ graf i rezidual-fit graf za potpuni model sa zavisnom varijablom udio zaraženih

Na slici 4.4 prikazani su navedeni grafovi za prvi model.
Lijevo vidimo da se točke u sredini grafa približno grupiraju u pravac, dok na lijevom i desnom rubu imamo odstupanja od pravca. Navedena opažanja ukazuju da reziduali pripadaju distribuciji težeg repa od normalne pa ne možemo zaključiti da pretpostavka o normalnosti reziduala vrijedi.

Desno uočavamo da se točke otprilike grupiraju oko x - osi iz čega slijedi da je $\mathbb{E}[\epsilon] \approx 0$. Međutim, na grafu uočavamo da kod predviđanja većeg broja zaraženih dolazi i do većih odstupanja pa je kod reziduala prisutna heteroskedastičnost.

Ovi rezultati mogli bi se korigirati daljnim transformacijama zavisne varijable, na primjer logaritmiranjem, no za potrebe ovog rada dovoljan je model bez transformiranih varijabli. Dodatni razlog ovakvih grafova moglo bi biti postojanje nekog čimbenika koji je nama nepoznat i samim time nije uključen u gradnju modela, a u stvarnosti je važan za promatrani problem.

Već smo komentirali da su neke varijable u modelu značajnije od drugih pa nas zanima da li ćemo izbacivanjem varijabli koje su manje značajne dobiti kvalitetniji model. S tim ciljem prvo provodimo stepenastu selekciju unazad. Koristit ćemo funkciju *regsubsets()* koja je dio paketa *leaps*. Ona funkcionira na temelju algoritama iz 3. poglavlja. Unutar funkcije možemo odabrati metodu koju želimo koristiti: 'backward', 'forward' ili 'seqrep' (za hibridni pristup). Također možemo sami odrediti maksimalan broj varijabli koje pri-

hvaćamo u model te prisilno ubaciti ili izbaciti neke od varijabli.

```
> #stepenasta selekcija unazad
> covid_podaci<-covid_podaci[, -3]
> back_zarazeni<- regsubsets(zarazeni/populacija~., data=covid_podaci[, -2], nvmax = 9,
+                            method = "backward")
> sum_back_zarazeni<-summary(back_zarazeni)
> sum_back_zarazeni
Subset selection object
Call: regsubsets.formula(zarazeni/populacija ~ ., data = covid_podaci[,
    -2], nvmax = 9, method = "backward")
8 Variables  (and intercept)
                     Forced in Forced out
`GUSTOCA POPULACIJE`     FALSE      FALSE
`PROSJ. GOD.`            FALSE      FALSE
`STARIJI OD 65`          FALSE      FALSE
GDP                      FALSE      FALSE
RAZVIJENOST              FALSE      FALSE
`SRCANI BOLESNICI`       FALSE      FALSE
DIJABETICARI             FALSE      FALSE
STROGOCA                 FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: backward
         `GUSTOCA POPULACIJE` `PROSJ. GOD.` `STARIJI OD 65` GDP RAZVIJENOST `SRCANI BOLESNICI` DIJABETICARI STROGOCA
1  ( 1 ) " "                  " "           "*"             " " " "          " "                " "          " "
2  ( 1 ) " "                  " "           "*"             "*" " "          " "                " "          " "
3  ( 1 ) " "                  " "           "*"             "*" "*"          " "                " "          " "
4  ( 1 ) " "                  " "           "*"             "*" "*"          "*"                " "          " "
5  ( 1 ) " "                  " "           "*"             "*" "*"          "*"                " "          "*"
6  ( 1 ) "*"                  " "           "*"             "*" "*"          "*"                " "          "*"
7  ( 1 ) "*"                  " "           "*"             "*" "*"          "*"                "*"          "*"
8  ( 1 ) "*"                  "*"           "*"             "*" "*"          "*"                "*"          "*"
> sum_back_zarazeni$adjr2
[1] 0.5252596 0.6494202 0.6559031 0.6603138 0.6658359 0.6660996 0.6649155 0.6628220
> max(sum_back_zarazeni$adjr2) #najveci prilagodeni R^2 je kod modela sa 6 varijabli
[1] 0.6660996
```

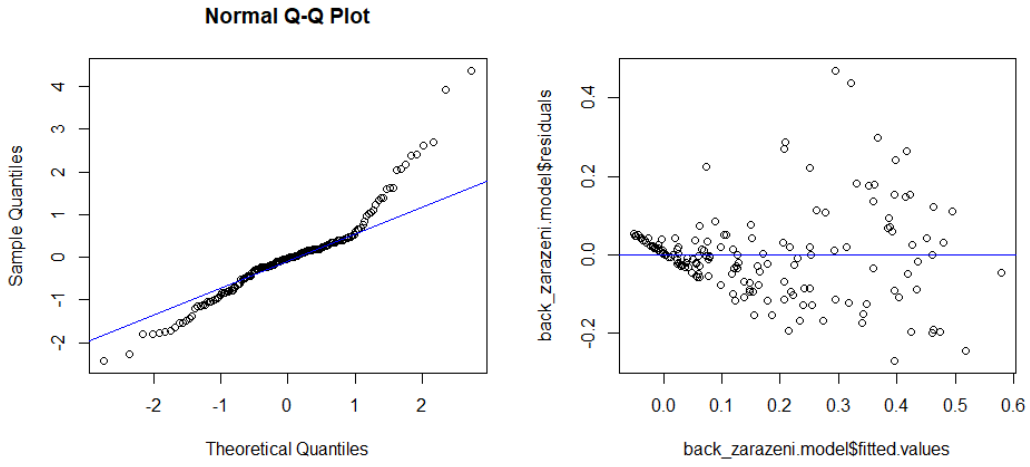
Slika 4.5: Rezultati stepenaste selekcije unazad za zavisnu varijablu udio zaraženih

Na slici 4.5 vidimo dobivene rezultate. Oznaka "*" znači da je varijabla uključena u određeni model. Na primjer, u najbolji model sa dvije varijable uključene su udio starijih od 65 i GDP. Ispisivanjem prilagođenih R^2 za sve najbolje modele s određenim brojem varijabli vidimo da je najveći 0.6660996. On se nalazi na šestoj poziciji pa zaključujemo da je vezan za model sa šest varijabli koji uključuje gustoću populacije, udio starijih od 65, GDP, indeks razvijenosti, broj umrlih od srčanih bolesti i indeks strogoće mjera.

Sada generiramo model s navedenim varijablama. Već smo spomenuli da je prilagođeni $R^2 = 0.6661$ te vidimo da je $R^2 = 0.6785$. Uspoređujući slike 4.3 i 4.6 uočavamo da su koeficijenti uz nezavisne varijable slični kao i kod potpunog modela, kao i značajne varijable. Crtajući QQ graf ne možemo zaključiti da su greške normalno distribuirane s obzirom da točke na krajevima odstupaju od pravca. Kod Tukey - Anscombe grafa primjećujemo da se točke ponovno jednoliko grupiraju oko pravca $y = 0$, ali udaljenost točaka od x - osi raste povećanjem predviđenih vrijednosti pa je heteroskedastičnost prisutna.

```
> back_zarazeni.model<-lm(zarazeni/populacija~gustoca_pop+stariji_od_65+gdp+razvijenost
+                         +srcani+strogoca)
> summary(back_zarazeni.model)

Call:
lm(formula = zarazeni/populacija ~ gustoca_pop + stariji_od_65 +
    gdp + razvijenost + srcani + strogoca)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2704 -0.0550 -0.0033  0.0371  0.4699

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.138e-01  7.709e-02  -1.476  0.14194
gustoca_pop    1.510e-05  1.424e-05   1.060  0.29070
stariji_od_65  1.088e-02  2.443e-03   4.454  1.6e-05 ***
gdp            2.186e-06  7.623e-07   2.868  0.00470 **
razvijenost    3.625e-01  1.370e-01   2.646  0.00898 **
srcani        -1.528e-04  8.143e-05  -1.876  0.06248 .
strogoca      -1.845e-03  9.532e-04  -1.936  0.05467 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1127 on 156 degrees of freedom
Multiple R-squared:  0.6785,    Adjusted R-squared:  0.6661
F-statistic: 54.86 on 6 and 156 DF,  p-value: < 2.2e-16
```

Slika 4.6: Rezultati za model dobiven stepenastom selekcijom unazad



Slika 4.7: QQ graf i rezidual-fit graf za reducirani model sa zavisnom varijablom broj zaraženih

Ako gledamo prilagođeni R^2 potpunog i reduciranog modela vidimo da reducirani mo-

del ima malo bolju prilagodbu. No da li je reducirani model zaista dovoljan ili je ipak potreban potpuni model možemo provjeriti izračunavanjem F - statistike. U skladu s oznakama varijabli s početka ovog poglavlja, postavljamo sljedeće hipoteze:

$$H_0 : \beta_2 = \beta_7 = 0 \text{ (Reducirani model je dovoljan)}$$
$$H_a : \beta_j \neq 0, j \in \{2, 7\} \text{ (Potreban je potpuni model).}$$

Dobivamo da je testna statistika jednaka $F = 0.2418$ te da je p - vrijednost 0.7855 što je veće od svih standardnih razina značajnosti pa ne odbacujemo nultu hipotezu da je reducirani model dovoljan.

Želimo li pogledati rezultate selekcije unaprijed potrebno je kod pokretanja funkcije *regsubsets()* uključiti argument *method='forward'*. U ovom slučaju dobivamo da je najveći prilagođeni R^2 isti onaj kao i kod selekcije unazad iako postoje razlike u određivanju najboljih modela s drugim brojem varijabli. Točnije, najbolji model s jednom varijablom određen ovom metodom je onaj koji sadrži samo indeks razvijenosti, a najbolji model s dvije varijable sadrži indeks razvijenosti i udio starijih od 65. U ostatku rezultata ove dvije metode se podudaraju.

Na kraju provodimo hibridnu selekciju. Njezini rezultati prikazani su slikom 4.8. Vidimo da je u ovom slučaju maksimalan prilagođeni R^2 0.665839 te da se on nalazi na petoj poziciji. Dobivamo da je najbolji model onaj s pet varijabli uključujući udio starijih od 65, GDP, indeks razvijenosti, broj umrlih od srčanih bolesti i indeks strogoće mjera.

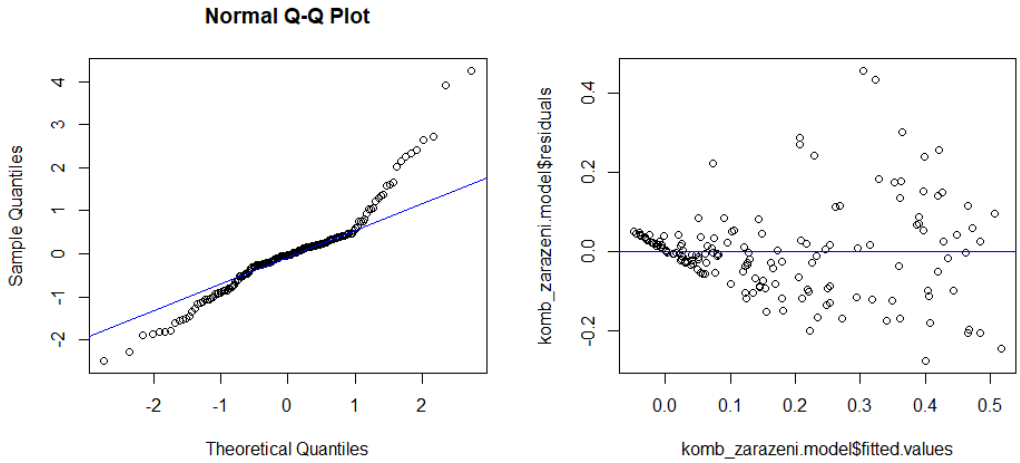
```
> #kombinacija forward i bacward selekcije
> komb_zarazeni<- regsubsets(zarazeni/populacija~., data=covid_podaci[, -2], nvmax = 9,
+                            method = "seqrep")
> sum_komb_zarazeni<-summary(komb_zarazeni)
> sum_komb_zarazeni
Subset selection object
Call: regsubsets.formula(zarazeni/populacija ~ ., data = covid_podaci[,
    -2], nvmax = 9, method = "seqrep")
8 Variables  (and intercept)
                     Forced in Forced out
`GUSTOCA POPULACIJE`     FALSE      FALSE
`PROSJ. GOD.`            FALSE      FALSE
`STARIJI OD 65`          FALSE      FALSE
GDP                      FALSE      FALSE
RAZVIJENOST              FALSE      FALSE
`SRCANI BOLESNICI`       FALSE      FALSE
DIJABETICARI             FALSE      FALSE
STROGOCA                 FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: 'sequential replacement'
         `GUSTOCA POPULACIJE` `PROSJ. GOD.` `STARIJI OD 65` GDP RAZVIJENOST `SRCANI BOLESNICI` DIJABETICARI STROGOCA
1  ( 1 ) " "                  " "           " "             " " "*"          " "                " "           " "
2  ( 1 ) " "                  " "           "*"             "*" " "          " "                " "           " "
3  ( 1 ) "*"                  "*"           "*"             " " " "          " "                " "           " "
4  ( 1 ) " "                  " "           "*"             "*" "*"          "*"                " "           " "
5  ( 1 ) " "                  " "           "*"             "*" "*"          "*"                " "           "*"
6  ( 1 ) "*"                  "*"           "*"             "*" "*"          "*"                " "           " "
7  ( 1 ) "*"                  " "           "*"             "*" "*"          "*"                "*"           "*"
8  ( 1 ) "*"                  "*"           "*"             "*" "*"          "*"                "*"           "*"
> sum_komb_zarazeni$adjr2
[1] 0.5724089 0.6494202 0.5741378 0.6603138 0.6658359 0.6592838 0.6649155 0.6628220
> max(sum_komb_zarazeni$adjr2)#najveci prilagodeni R^2 je kod modela sa 5 varijabli
[1] 0.6658359
```

Slika 4.8: Rezultati hibridnog pristupa za zavisnu varijablu udio zaraženih

Generiramo model s pet varijabli. Kao što smo već rekli prilagođeni R^2 je 0.6658 te vidimo da je $R^2 = 0.6761$. Sada je značajnost svake od preostalih varijabli u modelu visoka. Promatrajući grafove za ovaj model dolazimo do istih zaključaka kao i ranije - ne možemo pretpostaviti da su reziduali normalno distribuirani i heteroskedastičnost je prisutna.

```
> komb_zarazeni.model<-lm(zarazeni/populacija~stariji_od_65+gdp+razvijenost
+                         +srcani+strogoca)
> summary(komb_zarazeni.model)

Call:
lm(formula = zarazeni/populacija ~ stariji_od_65 + gdp + razvijenost +
    srcani + strogoca)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27676 -0.05483 -0.00278  0.03758  0.45887

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.062e-01  7.679e-02  -1.383  0.16851
stariji_od_65  1.087e-02  2.444e-03   4.448 1.63e-05 ***
gdp            2.383e-06  7.396e-07   3.222  0.00155 **
razvijenost    3.507e-01  1.366e-01   2.567  0.01118 *
srcani        -1.574e-04  8.135e-05  -1.934  0.05485 .
strogoca      -1.811e-03  9.530e-04  -1.900  0.05923 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1128 on 157 degrees of freedom
Multiple R-squared:  0.6761,    Adjusted R-squared:  0.6658
F-statistic: 65.56 on 5 and 157 DF,  p-value: < 2.2e-16
```

Slika 4.9: Rezultati za model dobiven hibridnim pristupom



Slika 4.10: QQ graf i rezidual-fit graf za reducirani model sa zavisnom varijablom udio zaraženih

Za kraj možemo pogledati da li je model s pet varijabli, dobiven korištenjem hibridnog pristupa, dovoljan ili je potreban model sa šest varijabli koji je dobiven stepenastom selekcijom unazad (i unaprijed). Ponovno računamo F - statistiku kako bi donjeli odluku. Postavljamo sljedeće hipoteze:

$$H_0 : \text{Model s pet varijabli je dovoljan.}$$
$$H_a : \text{Potreban je model sa šest varijabli.}$$

Dobivamo da je testna statistika jednaka $F = 1.124$ te da je p - vrijednost 0.2907 što je veće od svih standardnih razina značajnosti pa ne odbacujemo nultu hipotezu da je model s jednom varijablom manje dovoljan.

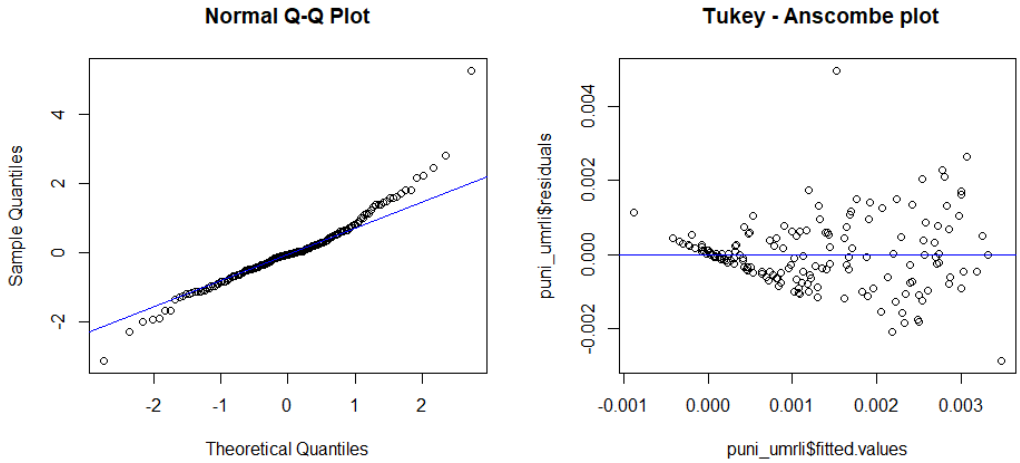
Udio umrlih kao zavisna varijabla

Neka je sada zavisna varijabla udio umrlih. Generiramo prvo potpuni model.

```
> ###UMRLI ZAVISNA VARIJABLA###
> puni_umrli<-lm(umrli/populacija~gustoca_pop+prosjek_god+stariji_od_65+gdp+
+                  razvijenost+srcani+dijabeticari+strogoca)
> summary(puni_umrli)

Call:
lm(formula = umrli/populacija ~ gustoca_pop + prosjek_god + stariji_od_65 +
    gdp + razvijenost + srcani + dijabeticari + strogoca)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0028751 -0.0005253 -0.0000566  0.0004473  0.0049783

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.461e-03  6.737e-04  -3.653 0.000355 ***
gustoca_pop   -2.035e-07  1.230e-07  -1.655 0.100056
prosjek_god    6.697e-05  3.403e-05   1.968 0.050836 .
stariji_od_65  5.062e-05  3.906e-05   1.296 0.196953
gdp           -2.216e-08  6.637e-09  -3.338 0.001058 **
razvijenost    2.450e-03  1.454e-03   1.685 0.094066 .
srcani        -3.041e-07  7.441e-07  -0.409 0.683372
dijabeticari  -7.866e-06  2.189e-05  -0.359 0.719844
strogoca       2.326e-06  8.402e-06   0.277 0.782310
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0009628 on 154 degrees of freedom
Multiple R-squared:  0.5444,    Adjusted R-squared:  0.5207
F-statistic:    23 on 8 and 154 DF,  p-value: < 2.2e-16
```

Slika 4.11: Rezultati za potpuni model sa zavisnom varijablom udio umrlih

Vidimo da su najznačajnije varijable u ovom modelu gustoća populacije, prosječne godine, GDP i indeks razvijenosti. Najmanje značajne su broj umrlih od srčanih bolesti, postotak dijabetičara te indeks strogoće mjera. Koeficijent determinacije je 0.5444, a prilagođeni R^2 iznosi 0.5207.



Slika 4.12: QQ graf i rezidual-fit graf za potpuni model sa zavisnom varijablom udio umrlih

Na slici 4.12 oba grafa imaju ista svojstva kao i već proučeni grafovi pa slijede i slični zaključci.

Provedimo sada ponovno stepenastu selekciju unazad. Ispisujemo prilagođeni R^2 za svaki od najboljih modela s određenim brojem varijabli. Zapažamo da je najveći među njima $R^2 = 0.5193169$ koji se nalazi na četvrtoj poziciji pa promatramo model sa četiri varijable. U taj model uključeni su gustoća populacije, udio starijih od 65, GDP i indeks razvijenosti. Isključeni su broj umrlih od srčanih bolesti, postotak dijabetičara i indeks strogoće mjera koje smo i naveli kao najmanje značajne varijable, te prosječne godine. Spomenimo da smo, u ovom slučaju, pozivanjem naredbe *regsubsets()*, iskoristili argument *force.out=2* kojim smo prisilno izbacili varijablu prosječne godine. Naime, funkcija bez korištenja tog argumenta kao model s najboljim prilagođenim R^2 izbacuje onaj koji istovremeno sadrži prosječne godine i razvijenost, koje su jako korelirane što bi moglo narušiti kvalitetu procjene.

Primjenom selekcije unaprijed i hibridnog pristupa opet dobivamo da je model sa iste četiri varijable cjelokupno najbolji iako postoje razlike u određivanju najboljih modela s drugim brojem varijabli.

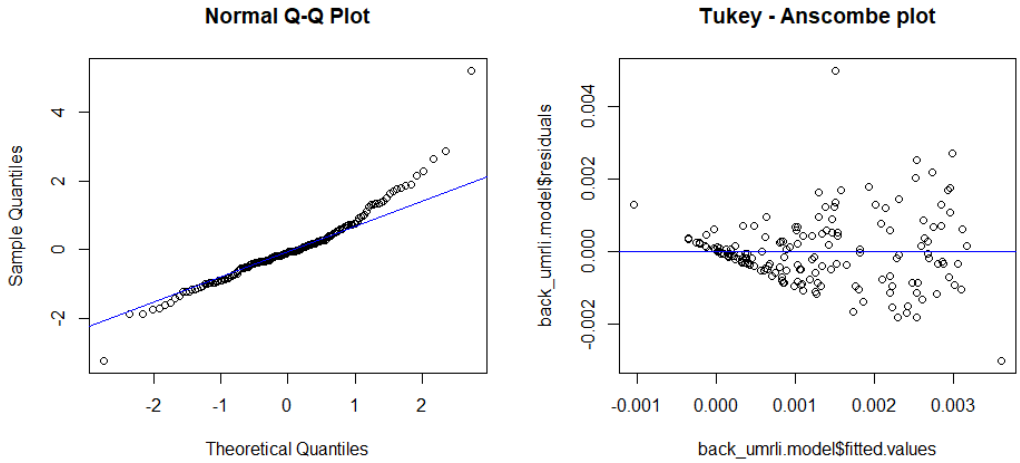
```
> #stepenasta selekcija unazad
> back_umrli<- regsubsets(umrli/populacija~., data=covid_podaci[, -1], nvmax = 9, force.out=2,
+                         method = "backward")
> sum_back_umrli<-summary(back_umrli)
> sum_back_umrli
Subset selection object
Call: regsubsets.formula(umrli/populacija ~ ., data = covid_podaci[,
    -1], nvmax = 9, force.out = 2, method = "backward")
8 Variables  (and intercept)
                    Forced in Forced out
`GUSTOCA POPULACIJE`     FALSE      FALSE
`STARIJI OD 65`          FALSE       TRUE
GDP                      FALSE      FALSE
RAZVIJENOST              FALSE      FALSE
`SRCANI BOLESNICI`       FALSE      FALSE
DIJABETICARI             FALSE      FALSE
STROGOCA                 FALSE      FALSE
`PROSJ. GOD.`            FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: backward
         `GUSTOCA POPULACIJE` `STARIJI OD 65` GDP RAZVIJENOST `SRCANI BOLESNICI` DIJABETICARI STROGOCA `PROSJ. GOD.`
1  ( 1 ) " "                  "*"             " " " "          " "                " "           " "       " "
2  ( 1 ) " "                  "*"             " " "*"          " "                " "           " "       " "
3  ( 1 ) " "                  "*"             "*" "*"          " "                " "           " "       " "
4  ( 1 ) "*"                  "*"             "*" "*"          " "                " "           " "       " "
5  ( 1 ) "*"                  "*"             "*" "*"          " "                " "           "*"       " "
6  ( 1 ) "*"                  "*"             "*" "*"          "*"                " "           "*"       " "
7  ( 1 ) "*"                  "*"             "*" "*"          "*"                "*"           "*"       " "
> sum_back_umrli$adjr2
[1] 0.4654317 0.4727157 0.5159990 0.5193169 0.5179752 0.5149763 0.5118617
> max(sum_back_umrli$adjr2) #najveci prilagodeni R^2 je kod modela sa 4 varijable
[1] 0.5193169
```

Slika 4.13: Rezultati stepenaste selekcije unazad za zavisnu varijablu udio umrlih

```
> #udio starijih od 65, GDP i razvijenost
> back_umrli.model<-lm(umrli/populacija~gustoca_pop+stariji_od_65+gdp+razvijenost)
> summary(back_umrli.model)

Call:
lm(formula = umrli/populacija ~ gustoca_pop + stariji_od_65 +
    gdp + razvijenost)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0030047 -0.0005121 -0.0000776  0.0004349  0.0049949

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.258e-03  5.959e-04  -3.790 0.000214 ***
gustoca_pop   -1.760e-07  1.215e-07  -1.448 0.149525
stariji_od_65  1.063e-04  1.948e-05   5.458 1.83e-07 ***
gdp           -2.101e-08  6.164e-09  -3.408 0.000829 ***
razvijenost    4.212e-03  1.072e-03   3.929 0.000127 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0009643 on 158 degrees of freedom
Multiple R-squared:  0.5312,    Adjusted R-squared:  0.5193
F-statistic: 44.76 on 4 and 158 DF,  p-value: < 2.2e-16
```

Slika 4.14: Rezultati za model dobiven stepenastom selekcijom unazad

Stvaranjem reduciranog modela dobivamo koeficijente prikazane na slici 4.14. R^2 za ovaj model je 0.5312, a prilagođeni R^2 iznosi 0.5193 što je nešto manje od prilagođenog R^2 kod potpunog modela. Uočavamo da su u ovom slučaju sve varijable značajne i crtamo pripadni QQ graf i rezidual-fit graf.



Slika 4.15: QQ graf i rezidual-fit graf za reducirani model sa zavisnm varijablom udio umrlih

Rezultati na slici 4.15 slični su kao i za sve modele do sada.
Na kraju provjeravamo da li je reducirani model zaista dovoljan. Postavljamo sljedeće hipoteze:

$$H_0 : \beta_2 = \beta_6 = \beta_7 = \beta_8 = 0 \text{ (Reducirani model je dovoljan)}$$
$$H_a : \beta_j \neq 0, j \in \{2, 6, 7, 8\} \text{ (Potreban je potpuni model).}$$

Računanjem testne statistike dobivamo da je $F = 1.1179$. Nadalje, p - vrijednost iznosi 0.3501 što je veće od svih standardnih razina značajnosti stoga ne odbacujemo nul - hipotezu o dovoljnosti reduciranog modela u odnosu na potpuni.

Poglavlje 5

Dodatak

U ovom dodatku nalazi se kod pomoću kojeg su dobiveni svi opisani rezultati.

Kod je napravljen u programskom jeziku R te su u njemu aktivirani paketi *leaps* (za korištenje funkcije *regsubsets()*) i *corrplot* (za izradu korelacijske matrice).

Prije samog pokretanja koda u program je importirana .csv datoteka pod imenom "covid_podaci" koja je sačinjena od 11 stupaca (to su redom "umrli", "zarazeni", "populacija", "gustoca_pop", "prosjek_god", "stariji_od_65", "gdp", "razvijenost", "srcani", "dijabeticari", "strogoca") koji sadrže podatke za 163 zemlje.

```

###ZARAZENI ZAVISNA VARIJABLA###
puni_zarazeni<-lm(zarazeni/populacija~gustoca_pop+prosjek_god+stariji_od_65+gdp+
                    razvijenost+srcani+dijabeticari+strogoca)
summary(puni_zarazeni)
par(mfrow=c(1, 2))
qqnorm(sort(rstandard(puni_zarazeni)))
qqline(sort(rstandard(puni_zarazeni)), col='blue')
plot(puni_zarazeni$fitted.values, puni_zarazeni$residuals, main=' Tukey - Anscombe plot')
abline(h=0, col='blue')


covid_podaci<-covid_podaci[, -3] #bri emo stupac populacija iz podataka kako bi funkcija regsubsets dobro radila

#####funkcija regsubsets()######
install.packages('leaps')
library(leaps)
#stepenasta selekcija unazad
back_zarazeni<- regsubsets(zarazeni/populacija~., data=covid_podaci[, -2], nvmax = 9,
```

```
                            method = "backward")
sum_back_zarazeni<-summary(back_zarazeni)
sum_back_zarazeni
sum_back_zarazeni$adjr2
max(sum_back_zarazeni$adjr2) #najveci prilagodeni R^2 je kod modela sa 6
    varijabli koji uklju uje:gusto u populacije, udio starijih od 65,
#GDP, indeks razvijenosti, broj umrlih od sr anih bolesti i indeks
    strogo e mjera
back_zarazeni.model<-lm(zarazeni/populacija~gustoca_pop+stariji_od_65+
    gdp+razvijenost
                        +srcani+strogoca)
summary(back_zarazeni.model)
qqnorm(sort(rstandard(back_zarazeni.model)))
qqline(sort(rstandard(back_zarazeni.model)), col='blue')
plot(back_zarazeni.model$fitted.values, back_zarazeni.model$residuals)
abline(h=0, col='blue')

anova(back_zarazeni.model, puni_zarazeni, test='F')
#p-vrijednost je 0.7855 sto je vece od svih standardnih razina
    znacajnosti (1, 5, 10%) pa
#ne odbacujemo nultu hipotezu da je reducirani model dovoljan

#stepenasta selekcija unaprijed
for_zarazeni<- regsubsets(zarazeni/populacija~., data=covid_podaci[,
    -2], nvmax = 9,
                          method = "forward")
sum_for__zarazeni<-summary(for_zarazeni)
sum_for__zarazeni
sum_for__zarazeni$adjr2
max(sum_for__zarazeni$adjr2)#najveci prilagodeni R^2 je kod modela sa 6
    varijabli koji uklju uje:gusto u populacije, udio starijih od 65,
#GDP, indeks razvijenosti, broj umrlih od sr anih bolesti i indeks
    strogo e mjera(isto kao i za backward)

#kombinacija forward i bacward selekcije
komb_zarazeni<- regsubsets(zarazeni/populacija~., data=covid_podaci[,
    -2], nvmax = 9,
                           method = "seqrep")
sum_komb_zarazeni<-summary(komb_zarazeni)
sum_komb_zarazeni
sum_komb_zarazeni$adjr2
max(sum_komb_zarazeni$adjr2)#najveci prilagodeni R^2 je kod modela sa 5
    varijabli koji uklju uje:udio starijih od 65,
#GDP, indeks razvijenosti, broj umrlih od sr anih bolesti i indeks
    strogo e mjera
komb_zarazeni.model<-lm(zarazeni/populacija~stariji_od_65+gdp+
    razvijenost
```

```
                          +srcani+strogoca)
summary(komb_zarazeni.model)
qqnorm(sort(rstandard(komb_zarazeni.model)))
qqline(sort(rstandard(komb_zarazeni.model)), col='blue')
plot(komb_zarazeni.model$fitted.values, komb_zarazeni.model$residuals)
abline(h=0, col='blue')

anova(komb_zarazeni.model, back_zarazeni.model, test='F')
#p-vrijednost je 0.2907 sto je vece od svih standardnih razina znacajnosti (1, 5, 10%) pa
#ne odbacujemo nultu hipotezu da je reducirani model dovoljan


###UMRLI ZAVISNA VARIJABLA###
puni_umrli<-lm(umrli/populacija~gustoca_pop+prosjek_god+stariji_od_65+gdp+
                 razvijenost+srcani+dijabeticari+strogoca)
summary(puni_umrli)
qqnorm(sort(rstandard(puni_umrli)))
qqline(sort(rstandard(puni_umrli)), col='blue')
plot(puni_umrli$fitted.values, puni_umrli$residuals, main=' Tukey - Anscombe plot')
abline(h=0, col='blue')

#stepenasta selekcija unazad
back_umrli<- regsubsets(umrli/populacija~., data=covid_podaci[, -1], nvmax = 9, force.out=2,
                          method = "backward")
sum_back_umrli<-summary(back_umrli)
sum_back_umrli
sum_back_umrli$adjr2
max(sum_back_umrli$adjr2) #najveci prilagodeni R^2 je kod modela sa 4 varijable koji uklju uje:gusto u populacije, udio starijih od 65, GDP i indeks
#razvijenosti
back_umrli.model<-lm(umrli/populacija~gustoca_pop+stariji_od_65+gdp+razvijenost)
summary(back_umrli.model)
qqnorm(sort(rstandard(back_umrli.model)))
qqline(sort(rstandard(back_umrli.model)), col='blue')
plot(back_umrli.model$fitted.values, back_umrli.model$residuals, main='Tukey - Anscombe plot')
abline(h=0, col='blue')

#stepenasta selekcija unaprijed
for_umrli<- regsubsets(umrli/populacija~., data=covid_podaci[, -1], nvmax = 9, force.out=2,
```

```
                        method = "forward")
sum_for_umrli<-summary(for_umrli)
sum_for_umrli
sum_for_umrli$adjr2
max(sum_for_umrli$adjr2) #najveci prilagodeni R^2 je kod modela sa 4
    varijable koji uklju uje:gusto u populacije, udio starijih od 65,
    GDP i indeks
#razvijenosti(isto kao i za backward)


#kombinacija forward i bacward selekcije
komb_umrli<- regsubsets(umrli/populacija~., data=covid_podaci[, -1],
    nvmax = 9, force.out=2,
                         method = "seqrep")
sum_komb_umrli<-summary(komb_umrli)
sum_komb_umrli
sum_komb_umrli$adjr2
max(sum_komb_umrli$adjr2) #najveci prilagodeni R^2 je kod modela sa 4
    varijable koji uklju uje:gusto u populacije, udio starijih od 65,
    GDP i indeks
#razvijenosti(isto kao i za backward)

anova(back_umrli.model, puni_umrli, test='F')
#p-vrijednost je 0.3501 sto je vece od svih standardnih razina
    znacajnosti (1, 5, 10%) pa
#ne odbacujemo nultu hipotezu da je reducirani model dovoljan


#korelacijska matrica
matrica<-as.data.frame(covid_podaci)
matrica[, 1]<-zarazeni/populacija
colnames(matrica)[1]<-'ZARAZENI'
matrica[, 2]<-umrli/populacija
colnames(matrica)[2]<-'UMRLI'
korelacije<-cor(matrica, method=c('pearson'))
korelacije
install.packages("corrplot")
library(corrplot)
par(mfrow=c(1, 1))
corrplot(korelacije, method=c('color'))
```

Bibliografija

[1] A. Almalki et al., Regression Analysis for COVID-19 Infections and Deaths Based on Food Access and Health Issues, Healthcare 2022, 10, 324

[2] J. Friedman, T. Hastie i R. Tibshirani, The Elements of Statistical Learning, Second Edition, Springer, 2009.

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

[4] M. Huzak, Vjerojatnost i matematička statistika, Prirodoslovno-matematički fakultet, Zagreb, 2006.

[5] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2021.

[6] N. Sandrić, Z. Vondraček, Vjerojatnost, Prirodoslovno-matematički fakultet, Zagreb, 2019.

[7] R. W. Keener, Theoretical Statistics, Springer, 2010.

[8] A. Sen, M. Srivastava, Regression Analysis: Theory, Methods and Application, Springer, 1990.

[9] `https://ourworldindata.org/coronavirus` (veljača 2024.)

Sažetak

Ovaj rad sadržava najbitnije teorijske rezultate vezane za višeparametarsku linearnu regresiju te primjenu tih rezultata na predviđanje smrtnosti od COVID-a. U prvom poglavlju naveli smo neke osnovne definicije iz vjerojatnosti i statistike koji su korisni za ostatak rada. Drugo poglavlje započeli smo s definicijom jednostavne linearne regresije koju smo odmah proširili na složenu linearnu regresiju. Objasnili smo metodu najmanjih kvadrata i dali njezinu geometrijsku interpretaciju. Dokazali smo Gauss - Markovljev teorem i spomenuli test hipoteza. U trećem poglavlju obradili smo metode odabira podskupa uključujući odabir najboljeg podskupa, stepenastu selekciju unaprijed i unazad te etapnu regresiju unaprijed. Osim toga, dotaknuli smo se metoda sažimanja, točnije ridge i lasso regresije. U zadnjem, četvrtom poglavlju počinjemo graditi modele višestruke linearne regresije promatrajući prvo udio zaraženih kao zavisnu varijablu, a nakon toga i udio umrlih od virusa.

Summary

This paper contains the most important theoretical results related to multiple linear regression and the application of these results in predicting COVID mortality. In the first chapter we mentioned some basic probability and statistics definitions that are useful for further work. We started the second chapter with the definition of simple linear regression, which we immediately expanded to multiple linear regression. We explained the method of least squares and gave its geometric interpretation. We proved the Gauss-Markov theorem and mentioned hypothesis testing. In the third chapter, we covered subset selection methods including best subset selection, forward stepwise selection, backward stepwise selection and forward stagewise regression. Additionally, we touched on shrinkage methods, specifically ridge and lasso regression. In the final, fourth chapter, we begin to build multiple linear regression models, first considering the share of infected as a dependent variable, and then the share of deaths from the virus.

Životopis

Rođena sam 4. ožujka 2000. godine u Zagrebu. Završila sam Osnovnu školu Matije Gupca u Gornjoj Stubici 2014. godine i nakon toga upisala opći smjer u Gimnaziji Sesvete. Nakon završetka srednjoškolskog obrazovanja, 2018. godine upisujem preddiplomski studij matematike nastavničkog smjera na Prirodoslovno - matematičkom fakultetu u Zagrebu. Preddiplomski studij završavam 2021. godine te upisujem diplomski studij, smjer Financijska i poslovna matematika na istom fakultetu. Na završnoj godini diplomskog studija započinjem svoje radno iskustvo zapošljavanjem u sektoru za aktuarske poslove u Allianz osiguranju.