Primjena linearne regresije na predviđanje COVID-19 pandemije

Matić, Nikolina

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Nikolina Matić

PRIMJENA LINEARNE REGRESIJE NA PREDVIĐANJE COVID-19 PANDEMIJE

Diplomski rad

Voditelj rada:
prof. dr. sc. Siniša Slijepčević

Zagreb, svibanj, 2022.

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. _____

Zahvaljujem se svojoj obitelji koja je strpljivo bila uz mene, a posebno roditeljima bez kojih sva moja postignuća ne bi bila moguća.
Hvala mentoru prof. dr. sc. Siniši Slijepčeviću na pomoći pri pisanju ovoga rada.
Zahvaljujem svim prijateljima i kolegama koji su me pratili tijekom studiranja.
I na kraju, velika hvala dragome Bogu na podarenim talentima i koji mi je cijelo vrijeme šaptao:"Možeš ti to".

Sadržaj

Uvod

Prvi slučaj zaraze sa COVID-19 spominje se u kineskom gradu Wuhanu, u prosincu 2019. godine. Radi se o bolesti koju je potaknuo koronski virus SARS-CoV-2, a koja se u roku od nekoliko mjeseci proširila u ukupno 216 zemalja i time izazvala globalnu pandemiju. Ona je utjecala na globalno zdravlje, ekonomiju i gospodarstvo.

S obzirom na razorne posljedice, za sprječavanje širenja pandemije uvodile su se raznovrsne mjere pa se može reći da je u nekim zemljama "vrijeme stalo". Mediji su preplašili javnost objavom neutemeljenih informacija i time širili dezinformacije i strah. Primjenjivale su se neispitane metode i mjere (primjerice, nošenje maski na otvorenome, višednevna samoizolacija, razmak od 2 metra itd...) koje na kraju, ponekad, nisu pokazivale pozitivne rezultate. Upravo zbog svega navedenog, pojavila su se neka pitanja: "Što možemo učiniti da smanjimo broj oboljelih od COVID-19, a potom i broj preminulih?", "Koliko će narasti broj zaraženih za tjedan dana?" i slično. Svakako je zanimljiva situacija kako je određena mjera smanjila broj zaraženih u pojedinoj zemlji, dok je u drugoj zemlji povećala broj zaraženih. Pri odabiru statističkog modela, prikladna se činila linearna regresija. Radi se o statističkom alatu koji se može primijeniti u različitim područjima. Ona opisuje promjenu jedne varijable uvjetovanu promjenama drugih varijabli, stoga pomoću nje možemo predviđati vrijednosti, a možemo i kvantificirati jačinu relacije između dviju ili više varijabli.

Sve ove činjenice su motivirajuće za izradu odgovarajuće analize na temelju poznatih mjerenih podataka dostupnih na https://ourworldindata.org. U ovome radu, primjenom metode linearne regresije temeljeno na [2] izvoru, pokušat ćemo odgovoriti na neka od postavljenih pitanja te otkriti barem neke čimbenike koji utječu na širenje pandemije i one koji utječu na suzbijanje pandemije.

Poglavlje 1

Zakoni velikih brojeva i statistički testovi

Upoznati s teorijom vjerojatnosti i definicijom vjerojatnosnog prostora, želimo "matematizirati" vjerojatnosni prostor. To bi značilo definirati funkciju koja će preslikavati elementarne događaje u realne brojeve. U prvom dijelu poglavlja ćemo definirati pojmove koji su nam potrebni za matematizaciju stvarnih podataka kako bismo ih dalje mogli matematički sistematizirati i modelirati regresijom.

1.1 Slučajna varijabla

Definicija 1.1.1. *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor. Funkcija* $X : \Omega \to \mathbb{R}$ *je slučajna varijabla ako za sve* $a, b \in \mathbb{R}$:

$$\{\, a \leq X \leq b \,\} \in \mathcal{F}.$$

Funkciju $F = F_X : \mathbb{R} \to [0, 1]$ *definiranu sa*

$$F(a) := \mathbb{P}(X \leq a), a \in \mathbb{R}$$

zovemo funkcija distribucije slučajne varijable X.

Svaka slučajna varijabla može poprimiti beskonačno mnogo vrijednosti ili konačno mnogo vrijednosti. Ako je slučajna varijabla konačna, tada nju zovemo diskretnom slučajnom varijablom, a u suprotnome se naziva kontinuirana ili neprekidna slučajna varijabla. Preciznije:

Definicija 1.1.2. *Za slučajnu varijablu X kažemo da je neprekidna slučajna varijabla ako postoji nenegativna (izmjeriva) funkcija* $f : \mathbb{R} \to \mathbb{R}$ *takva da je*

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx,$$

za sve $a, b \in \mathbb{R}, a < b$. Funkciju f zovemo funkcija gustoće slučajne varijable X.

Kako bismo opisali promatranu slučajnu varijablu, koristimo funkciju gustoće koja je definirana u 1.1.2. Pomoću nje računamo vjerojatnost da se određeni X nalazi unutar intervala kojeg određuju granice integrala.

Definicija 1.1.3. *Neka je $(\Omega, \mathcal{P}(\Omega), P)$ diskretni vjerojatnosni prostor. Funkciju $X : \Omega \to \mathbb{R}$ zovemo diskretna slučajna varijabla ako je slika $Im(X)$ prebrojiv skup. Pri tome mora vrijediti da su skupovi*

$$X = x = w \in \Omega : X(w) = x$$

događaji za svaki $x \in Im(X)$. Funkcija gustoće vjerojatnosti diskretne varijable X je funkcija $f_x = f : \mathbb{R} \to \mathbb{R}^+$ definirana sa

$$f(x) = \mathbb{P}\{X = x\} = \begin{cases} 0 \ , x \neq a_i \\ p_i \ , x = a_i \end{cases}$$

Funkcija gustoće vjerojatnosti zadovoljava uvjet da je vjerojatnost sigurnog događaja jednaka 1, pa stoga znamo da će vrijednost slučajne varijable biti u području definicije varijable. Ukoliko imamo neprekidnu slučajnu varijablu, osim funkcije gustoće, možemo definirati matematičko očekivanje i varijancu varijable.

Definicija 1.1.4. *Neka je X neprekidna slučajna varijabla s funkcijom gustoće f_X. Matematičko očekivanje od X interpretira se kao srednja (očekivana) vrijednost od X u oznaci $\mathbb{E}[X]$ i vrijedi:*

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x \cdot f_X(x)dx, \text{ ako je } X \text{ neprekidna i } \int_{-\infty}^{\infty} x \cdot f_X(x)dx \text{ apsolutno konvergira.}$$
$$\mathbb{E}X := \sum_{x \in Im(X)} x \cdot f_X(x), \text{ ako je } X \text{ diskretna i red } \sum_{x \in Im(X)} x \cdot f_X(x) \text{ apsolutno konvergira.}$$

Neka je $g : \mathbb{R} \to \mathbb{R}$ realna, po dijelovima neprekidna funkcija i $X : \Omega \to \mathbb{R}$ slučajna varijabla. Tada je $g(X) = g \circ X : \Omega \to \mathbb{R}$ također slučajna varijabla i ona ima očekivanje $\mathbb{E}[g(X)]$:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx, \text{ ako je } X \text{ neprekidna i } \int_{-\infty}^{\infty} x \cdot f_X(x)dx \text{ apsolutno konvergira}$$
$$\mathbb{E}[g(X)] = \sum_{w_k \in \Omega} g(x) \cdot f_X(x), \text{ ako je } X \text{ diskretna i red } \sum_{x \in Im(X)} x \cdot f_X(x) \text{ apsolutno konvergira.}$$

Ove definicije su preuzete iz [12], poglavlje 2.

Kada smo odrediti očekivanje i varijancu slučajne varijable, zanima nas još prosječno kvadratno odstupanje numeričkih vrijednosti varijable od njihove aritmetičke sredine. Pomoću standardne devijacije komentiramo disperziju vrijednosti skupa podataka.

Definicija 1.1.5. *Za $r \in \mathbb{N}$ sa $M_r := \mathbb{E}[X^r]$ definiramo r-ti moment slučajne varijable X ukoliko očekivanje postoji. Ako postoji drugi moment slučajne varijable X, onda definiramo varijancu od X sa*

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

Broj $\sigma_X := \sqrt{\mathbb{V}ar(X)} \geq 0$ zovemo standardna devijacija slučajne varijable X. [8]

1.2 Razdiobe i statističko zaključivanje

Razdiobe slučajne varijable

Kod primjene, najvažnija nam je razdioba (distribucija) slučajnih varijabli koja se opisuje funkcijom gustoće vjerojatnosti. Postoje razne razdiobe koje se pojavljuju u područjima primjene, no one koje koristimo u ovom radu su normalna razdioba, (Studentova) t-razdioba, binomna i Poissonova razdioba. Dijelimo ih na diskretne (binomna i Poissonova) i neprekidne (normalna/Gaussova i Studentova/t).

Primjer 1.2.1. *Neka je X neprekidna slučajna varijabla vjerojatnosnog prostora $(\Omega, \mathcal{F}, \mathbb{P})$.*

a) Kažemo da slučajna varijabla X ima normalnu ili Gaussovu razdiobu s parametrima μ i $\sigma^2 > 0$ te pišemo $X \sim N(\mu, \sigma^2)$ ako je $ImX = \mathbb{R}$ i funkcija gustoće joj je

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

b) Kažemo da slučajna varijabla X ima (Studentovu) t-razdiobu s n stupnjeva slobode ako joj je funkcija gustoće

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \frac{1}{(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}$$

i pišemo $X \sim t(n)$.

Primjer 1.2.2. *Neka je X diskretna slučajna varijabla na diskretnom vjerojatnosnom prostoru $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$.*

a) Kažemo da slučajna varijabla X ima binomnu razdiobu s parametrima $n \in \mathbb{N}$ i $0 < p < 1$, ako joj je funkcija gustoće

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x \in 0, 1, ..., n.$$

i pišemo $X \sim B(n, p)$.
b) Kažemo da slučajna varijabla X ima Poissonovu razdiobu s parametrom $\lambda, \lambda > 0$ *ako joj je funkcija gustoće*

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, x \in \mathbb{N}_0.$$

i pišemo $X \sim P(\lambda)$.

Za više primjera vidi [5].

Statistika i statističko zaključivanje

Općenito, pri korištenju statističkih modela, koristimo poznate izmjerene parametre pa možemo reći da statistika sadrži samo poznate parametre. Cilj nam je pomoću statističkih modela, na temelju odabranih uzoraka, matematički modelirati niz mjerenja varijable X kako bismo mogli procijeniti i odrediti vrijednost za neko mjerenje koje nismo zapravo izvršili. U ovome dijelu definiramo statistiku i objašnjavamo test hipoteza.

Definicija 1.2.3. *Slučajna varijabla*

$$\Theta := g(X_1, ...X_n)$$

naziva se statistika. Statistikom nazivamo svaku funkciju koja ovisi o uzorku $X_1, ..., X_n$*, a ne ovisi (eksplicitno) o nepoznatom parametru.*

Definicija 1.2.4. *Slučajan uzorak je niz međusobno nezavisnih jednako distribuiranih slučajnih varijabli. Označavamo ga sa* $\underline{X}$*. Ukoliko se sastoji od n varijabli* $X_1, X_2, ..., X_n$*, tada je slučajan vektor* $\underline{X}$ *dan sa*

$$\underline{X} = (X_1, X_2, ..., X_n).$$

Uređenu n-torku brojeva $\underline{x} = (x_1, x_2, ..., x_n)$ *koja predstavlja realizaciju slučajnog uzorka* $\underline{X}$ *zovemo opaženi uzorak.*

Pri proučavanju praktičnih situacija u vezi sa slučajnim promjenama, donošenje odluka provodi s konačnim potvrdnim (DA) ili negacijskim (NE) odgovorom. U statistici nam pomažu statistički testovi pomoću kojih, između ostaloga, donosimo konačnu odluku. Dan nam je n-člani slučajan uzorak i želimo pokazati da određena tvrdnja (pretpostavka) vrijedi. Pretpostavka koju prihvaćamo ili odbacujemo na temelju vrijednosti slučajnog uzorka $X_1, ..., X_n$ naziva se statistička hipoteza, a postupak donošenja odluke testiranje. Statistički test je pravilo podjele prostora na vrijednosti uzoraka na područje koji su konzistentni sa H_0 i na njegov komplement.
Osnovna hipoteza koja se testira zove se nulhipoteza i označava se sa H_0. Ona najčešće označava neutralnu izjavu koja prezentira postojeće stanje slučajnog uzorka. Uz nju se

postavlja i njoj alternativna hipoteza H_1. U konačnici, statistički testovi nam pomažu odgovoriti na pitanje: "Daju li nam dani podaci dovoljno dokaza da odbacimo H_0?". Ukoliko se opažena vrijednost testne statistike nalazi u kritičnom području (područje nekonzistentno sa H_0), tada se H_0 hipoteza odbacuje u koristi H_1.
Dodatan parametar koji nam pomaže pri donošenju odluke je razina značajnosti testa α koju definiramo kao vjerojatnost odbacivanja H_0 u slučaju kada je H_0 istinita hipoteza. Ako odbacimo istinitu hipotezu H_0, tada kažemo da smo počinili pogrešku prve vrste. Obrnuto, ukoliko ne odbacimo H_0, a H_1 je istina hipoteza, tada smo počinili pogrešku druge vrste koju označavamo s β.
Idealan statistički test bi bio u slučaju da vjerojatnost spomenutih grešaka možemo učiniti proizvoljno malima, no takav test ne postoji. [11] i [13].

Jaki zakon velikih brojeva

Jaki zakon velikih brojeva je temeljni teorem teorije vjerojatnosti koji govori da ako beskonačno puta ponovimo eksperiment, tada je učestalost određenog događaja konstantna.

Teorem 1.2.5. ***Jaki zakon velikih brojeva***
Neka je $(X_n)_{n\in\mathbb{N}}$ niz nezavisnih jednako distribuiranih slučajnih varijabli takvih da je $\mathbb{E}(X_n) = \mu$. Tada

$$\frac{X_1 + ... + X_n}{n} \to \mu.$$

Dokaz. Za detaljan dokaz ovog teorema vidi [12], poglavlje 8. □

Poglavlje 2

Linearna regresija

Linearna regresija je statistički model koji objašnjava korelaciju između odzivne (slučajne) varijable Y i eksplanatorne (nezavisne) varijable X. Eksplanatornih varijabli može biti više, dok je odzivna varijabla uvijek samo jedna. U najjednostavnijem slučaju regresije proučavamo linearnu zavisnost jedne varijable Y o jednoj varijabli X. Kada modeliramo takav problem, koristimo princip jednostavne linearne regresije. S druge strane, ukoliko proučavamo zavisnost jedne varijable Y u odnosu na više eksplanatornih varijabli $X_1, X_2, ..., X_n$, tada modeliramo višestrukom linearnom regresijom. U poglavlju 2.1. ćemo objasniti kako modelirati najprije jednostavnom linearnom regresijom, a u poglavlju 2.2. kako modelirati višestrukom linearnom regresijom.

2.1 Jednostavna linearna regresija

Pravac regresije i metoda najmanjih kvadrata

Ovo poglavlje objašnjava pojam jednostavne linearne regresije te kako odrediti koeficijente regresije. Također, proučit ćemo i objasniti metode provjere koje nam pomažu pri procjeni odaziva.

Jednostavna linearna regresija je najjednostavniji oblik regresije pomoću koje možemo opisati ovisnost dviju kvantitativnih varijabli. Želimo donijeti zaključke o ponašanju jedne odzivne varijable Y proučavajući jednu eksplanatornu varijablu X. Opisana linearna veza dana je sa:

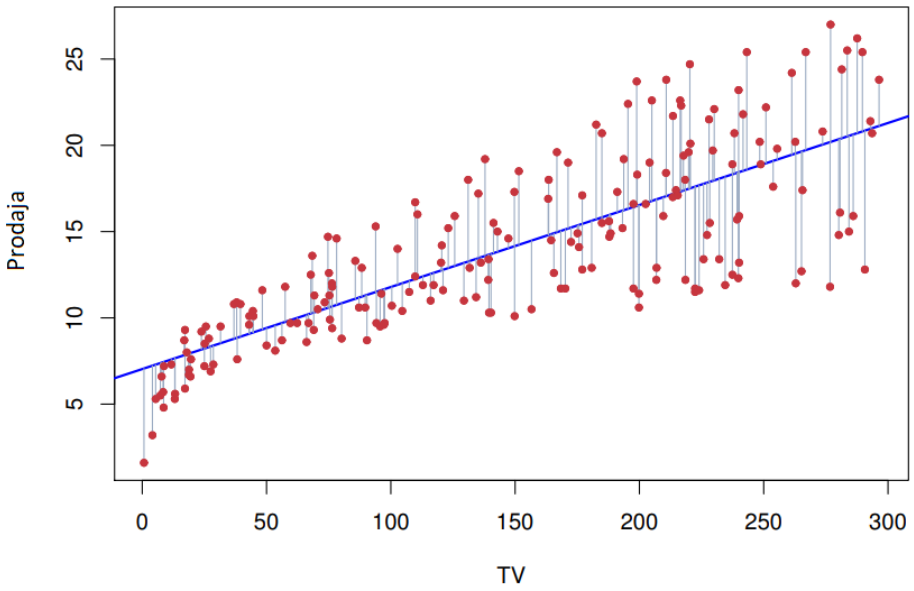
$$Y = \beta_0 + \beta_1 \cdot X. \tag{2.1}$$

Dan nam je niz sparenih, stvarnih mjerenja $(x_1, y_1), ..., (x_n, y_n)$ gdje su $x_1, ..., x_n$ vrijednosti nezavisne varijable X, a $y_1, ..., y_n$ odgovarajuće vrijednosti slučajne varijable Y. Pomoću modela 2.1 računamo srednju vrijednost odzivne varijable u ovisnosti o jednoj eksplana-

tornoj varijabli, a ukoliko želimo izračunati vrijednost za i-to mjerenje tada koristimo:

$$y_i = \beta_0 + \beta_1 X_i, \forall i. \tag{2.2}$$

Postupak određivanja regresijskih koeficijenata nije kompliciran. No, najprije dane podatke prikažemo **dijagramom raspršenosti** (točkama u koordinatnom sustavu) pomoću kojeg vidimo formiraju li se točke oko pravca ili neke druge krivulje. Ukoliko se formiraju oko pravca, možemo pretpostaviti da između odzivne i eksplanatorne varijable postoji linearna zavisnost, u suprotnom odbacujemo tu mogućnost. Na slici 2.1 prikazan je dija-



Slika 2.1: Dijagram raspršenja [9]

gram raspršenosti i pripadni procijenjeni pravac regresije (plavom bojom). Dani grafički prikaz prikazuje iznos proračuna za TV oglašavanje u ovisnosti o broju prodajnih mjesta. Detaljnije o provedenom istraživanju vidi [9], poglavlje 3. Crvene točke označuju stvarnu, izmjerenu vrijednost. Uočimo da regresijski pravac ne sadrži sve točke, primjerice, prva po redu točka se ne nalazi na pravcu regresije, već ispod njega. Uočimo da postoji razlika između stvarne i predviđene vrijednosti koja je na slici označena tankom crnom okomitom linijom te nju označimo s: $\epsilon_n = y_n - \hat{y_n}$. Grešku ϵ_n zovemo i rezidum za n-to mjerenje. U praksi gotovo nikad ne nailazimo na podatke koje je moguće savršeno opisati pravcem jer gotovo uvijek postoji barem jedno odstupanje, stoga će se stvarna vrijednost nalaziti iznad ili ispod predviđene vrijednosti na pravcu. Uzimajući u obzir sve rezidume, poboljšavamo

model 2.2 :

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \forall i. \tag{2.3}$$

gdje su:

- y_i: vrijednosti odzivne varijable za i-to mjerenje
- x_i: vrijednosti eksplanatorne varijable za i-to mjerenje
- ϵ_i: vrijednosti reziduma za i-to mjerenje
- β_0: parametar presjeka (nepoznat)
- β_1: parametar nagiba pravca regresije (nepoznat)

Kako bismo što kvalitetnije modelirali problem i odredili najprecizniji pravac regresije, želimo da suma reziduma bude najmanja moguća (idealan model bi bio onaj u kojem je suma reziduala jednaka 0).

Poznate su nam sve vrijednosti potrebne za određivanje regresijskog modela, osim parametara β_0, β_1, a oni određuju pravac regresije. Tražimo pravac koji najbolje modelira dane podatke, a to će biti u slučaju kada je suma kvadrata reziduma minimalna pa definiramo funkciju koja određuje sumu kvadratnih reziduala (odstupanje teoretskih od eksperimentalnih vrijednosti):

$$\begin{aligned} S(\beta_0, \beta_1) &:= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 \cdot x_i) \right]^2 . \end{aligned} \tag{2.4}$$

Uočimo, dani podaci $(x_1, y_1), ..., (x_n, y_n)$ su nepromjenjivi pa su i reziduali ϵ_i također nepromjenjivi. Procjenjujemo regresijske koeficijente u oznaci $\hat{\beta}_0, \hat{\beta}_1$ za koje funkcija poprima minimalnu vrijednost.

$$\begin{aligned} S_{(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2}(\hat{\beta}_0, \hat{\beta}_1) &= min \left[S(\beta_0, \beta_1) \right] \\ &= min \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 \cdot x_i)]^2. \end{aligned}$$

Prema nužnom uvjetu ekstrema, da bi točka $T(\hat{\beta}_0, \hat{\beta}_1)$ bila lokalni ekstrem diferencijabilne

funkcije $S(\beta_0, \beta_1)$ tada za prve parcijalne derivacije u toj točki vrijedi:

$$\frac{\partial S}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = \frac{\partial S}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 0.$$

Izračunamo prve parcijalne derivacije funkcije S:

$$\frac{\partial S}{\partial \beta_0}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 \cdot x_i) \right], \tag{2.5}$$

$$\frac{\partial S}{\partial \beta_1}(\beta_0, \beta_1) = -2 \cdot x_i \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 \cdot x_i) \right]. \tag{2.6}$$

Za pronalazak ekstrema izjednačimo 2.5 i 2.6 s 0. Iz 2.5 slijedi:

$$\begin{aligned} & -2 \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 \cdot x_i) \right] = 0 \\ \Leftrightarrow \; & \sum_{i=1}^{n} y_i - n \cdot \beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \end{aligned} \tag{2.7}$$

Uočimo, srednje vrijednosti od X i Y dane su sa: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ pa zamjenom u 2.7 i sređivanjem izraza slijedi:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.8}$$

Analogno, iz 2.6:

$$\begin{aligned} & -2 \cdot x_i \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 \cdot x_i) \right] = 0 \\ \Leftrightarrow \; & \sum_{i=1}^{n} y_i - n\beta_0 - n\beta_1 \sum_{i=1}^{n} x_i = 0 \end{aligned} \tag{2.9}$$

Uvrstimo 2.8 u 2.9 te izrazimo $\hat{\beta}_1$:

$$\begin{aligned} & \sum_{i=1}^{n} x_i y_i - n\hat{\beta}_1 \bar{x}(\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \\ \Leftrightarrow \hat{\beta}_1 = & \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}. \end{aligned} \tag{2.10}$$

Nismo sigurni postiže li funkcije $S(\beta_0, \beta_1)$ u dobivenom ekstremu maksimalnu ili minimalnu vrijednost, stoga provjeravamo dovoljan uvjet ekstrema koji kaže da diferencijabilna funkcija $S(\hat{\beta}_0, \hat{\beta}_1)$ postiže minimum u točki ekstrema T ako:

$$\begin{aligned} &1.\ \triangle_T > 0 \ \Leftrightarrow\ \frac{\partial S^2}{\partial^2 \beta_0} \cdot \frac{\partial S^2}{\partial^2 \beta_1} - \left(\frac{\partial S^2}{\partial \beta_0 \partial \beta_1}\right) > 0 \\ &2.\ \frac{\partial S^2}{\partial^2 \beta_0} > 0 \end{aligned} \tag{2.11}$$

Iz 2.5 i 2.6 izračunamo druge parcijalne derivacije i provjerimo dovoljan uvjet:

$$\frac{\partial S^2}{\partial^2 \beta_0}(\beta_0, \beta_1) = 2n \ \ , \ \ \frac{\partial S^2}{\partial^2 \beta_1}(\beta_0, \beta_1) = -2nx_i^2 \ \ , \ \ \frac{\partial S^2}{\partial \beta_0 \partial \beta_1}(\beta_0, \beta_1) = 0$$

Vidimo da u točki $T(\hat{\beta}_0, \hat{\beta}_1)$ funkcija $S(\beta_0, \beta_1)$ poprima minimalnu vrijednost te je pravac regresije dan s

$$Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot X + \epsilon \tag{2.12}$$

čije su vrijednosti dane s $y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i + \epsilon_i, \ \forall i.$

Analiza metode i statističko zaključivanje

Procijenili smo regresijske koeficijente $\hat{\beta}_0, \hat{\beta}_1$ i pripadni regresijski pravac za koji pretpostavljamo da najbolje povezuje dane podatke. No, tako možemo odrediti i pravac regresije za neke varijable između kojih inače ne postoji linearna zavisnost odredimo li, primjerice, za neka dva mjerenja, graf raspršenosti i pomoću njih odredimo jedinstveni pravac. Zbog potrebe provjere sposobnosti modela da objasni vrijednost opisne varijable u ovisnosti o eksplanatornoj varijabli, koristimo i provjeravamo statističke pokazatelje koji daju odgovor na pitanja:

a. Jesu li procijenjeni parametri približno jednaki stvarnim regresijskim parametrima i koliko su dobro procijenjeni?

b. Postoji li uopće linearna veza između eksplanatorne i odzivne varijable?

c. Koliko dobro procijenjeni regresijski pravac modelira podatke?

d. Mogu li procijeniti odziv koristeći regresijski pravac i koliko će procjena biti dobra? [9], poglavlje 3.

a. Nepristranost regresijskih koeficijenata

U ovom dijelu želimo provjeriti daju li parametri $\hat{\beta}_0, \hat{\beta}_1$ dovoljno dobru procjenu.
Pomoću metode najmanjih kvadrata odredili smo $\hat{\beta}_0, \hat{\beta}_1$ koeficijente za koje je suma pogrešaka minimalna, stoga smo danim podacima povukli linearni pravac koji bi najbolje opisao i povezao dane podatke. No, koliko dobro funkcija opisuje stvarne podatke? Hoće li buduće procijenjene vrijednosti biti dobro procijenjene? Tu nam uvelike pomažu Gauss-Markovljevi uvjeti koji nam daju sigurnost i pouzdanost da su dobiveni koeficijenti zaista dobri. Ukoliko su oni zadovoljeni, za dobivene procjenitelje, očekivano kvadratno odstupanje je najmanje o čemu nam preciznije govori Gauss - Markovljev teorem.

Kako su regresijski koeficijenti slučajne varijable, možemo odrediti njihova očekivanja. Time ćemo provjeriti jesu li procijenjeni regresijski koeficijenti blizu vrijednosti stvarnih koeficijenata.

Propozicija 2.1.1. *Neka je $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \epsilon$ pravac regresije dobiven metodom najmanjih kvadrata. Tada vrijedi:*

$$\mathbb{E}[\hat{\beta}_0] = \beta_0 \ , \ \mathbb{E}[\hat{\beta}_1] = \beta_1. \tag{2.13}$$

Dokaz. Iz pretpostavke, za vrijednosti eksplanatorne varijable vrijedi $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ pa definirajmo $c_i := \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Očito je:

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0, \tag{2.14}$$

$$\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} c_i (x_i - \bar{x}) = 1 \tag{2.15}$$

Sada, uz 2.10 i dobivene rezultate, pokažimo da vrijedi $\mathbb{E}[\hat{\beta}_1] = \beta_1$.

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_1\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n y_i(x_i - \bar{x}) - \bar{y}\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n c_i y_i\right] \\
&= \sum_{i=1}^n c_i \cdot \sum_{i=1}^n (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i \\
&= \beta_1.
\end{aligned} \tag{2.16}$$

Još je preostalo pokazati da je $\mathbb{E}\left[\hat{\beta}_0\right] = \beta_0$:

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0\right] &= \mathbb{E}\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n y_i}{n}\right] - \bar{x}\mathbb{E}\left[\hat{\beta}_1\right] \\
&= \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 \\
&= \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 \\
&= \beta_0.
\end{aligned} \tag{2.17}$$

□

Ovime smo dokazali da su dobiveni procijenjeni parametri β_0, β_1 nepristrani.

Teorem 2.1.2. *Neka su $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$ i-te vrijednosti regresijskog pravca. Za slučajne greške ϵ_i, $\forall i = 1, ..., n$ vrijedi:*

(1) centriranost: $\mathbb{E}[\epsilon_i] = 0, \forall i$

(2) jednakost varijanci: $\mathbb{V}ar[\epsilon_i] = \sigma^2, \forall i$

(3) nekolinearnost: $Cov(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$

Svojstva (1)-(3) jednim imenom zovemo ***Gauss-Markovljevim uvjetima****.*

Dokaz. Dokažimo teorem uz pomoć prethodne propozicije gdje je $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \epsilon$ regresijski pravac procijenjen metodom najmanjih kvadrata:
(1) Centriranost:

$$\begin{aligned}\mathbb{E}[\epsilon_i] &= \mathbb{E}[y_i - \hat{y}_i] \\ &= \mathbb{E}[\beta_0 + \beta_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] \\ &= \mathbb{E}[x_i(\beta_1 - \hat{\beta}_1) + \beta_0 - \hat{\beta}_0] \\ &= x_i(\mathbb{E}[\beta_1] - \mathbb{E}[\hat{\beta}_1] + \mathbb{E}[\beta_0] - \mathbb{E}[\hat{\beta}_0] \\ &= x_i(\beta_1 - \beta_1) + \beta_0 - \beta_0 = 0.\end{aligned}$$

(2) Jednakost varijanci:

$$\begin{aligned}\mathbb{V}ar[\epsilon_i] &= \mathbb{E}(\epsilon_i - \mathbb{E}(\epsilon_i)^2) \\ &= \mathbb{E}(\epsilon_i^2) \\ &= \sigma^2.\end{aligned}$$

(3) Nekolinearnost:

$$Cov(\epsilon_i, \epsilon_j) = \mathbb{E}(\epsilon_i \epsilon_j) - \mathbb{E}(\epsilon_i)\mathbb{E}(\epsilon_j) = 0.$$

□

Kao posljedica Gauss-Markovljevih svojstava slijedi da ϵ_i ima normalnu distribuciju $\epsilon_i \sim N(0, \sigma^2),\ \forall i$. [9]

Teorem 2.1.3. *Pretpostavimo da vrijede Gauss-Markovljevi uvjeti. Tada su procjenitelji $\hat{\beta}_0$, $\hat{\beta}_1$ nepristrani, što znači da među svim pristranim linearnim procjeniteljima β_0, β_1, upravo $\hat{\beta}_0$, $\hat{\beta}_1$ imaju najmanju varijancu (najmanje očekivano kvadratno odstupanje od stvarne vrijednosti).*

Kako su koeficijenti $\hat{\beta}_0$ i $\hat{\beta}_1$ slučajne varijable, za njih možemo odrediti varijance. Sljedeća propozicija iskazuje varijance regresijskih koeficijenata.

Propozicija 2.1.4. *Varijance od nepristranih regresijskih koeficijenata dane su sa:*

$$\mathbb{V}ar[\hat{\beta}_0] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\mathbb{V}ar[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Dokaz. Općenito,

$$\mathbb{V}ar(y_i) = \mathbb{V}ar(\beta_0 + \beta_1 x_i + \epsilon_i) = \mathbb{V}ar(\epsilon_i) = \sigma^2. \tag{2.18}$$

Koristeći 2.18 slijedi:

$$\begin{aligned}\mathbb{V}ar(\hat{\beta}_1) &= \sum_{i=1}^{n} c_i^2 \mathbb{V}ar(y_i) \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 \\ &= \sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2.\end{aligned}$$

Slično, za $\mathbb{V}ar(\hat{\beta}_0)$:,

$$\begin{aligned}\mathbb{V}ar(\hat{\beta}_0) &= \mathbb{V}ar(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= \mathbb{V}ar\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}\sum_{i=1}^{n} c_i y_i\right) \\ &= \mathbb{V}ar(y_i)\sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}c_i\right)^2 \\ &= \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{2\bar{x}c_i}{n} + \bar{x}^2 c_i^2\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).\end{aligned}$$

□

Uočimo da varijance $\hat{\beta}_0$ i $\hat{\beta}_1$ ovise o broju sparenih mjerenja n, danim vrijednostima x_i i varijanci reziduala. Pokazali smo da je $\hat{\beta}_0$ zaista nepristrani procjenitelj od β_0, a $\hat{\beta}_1$ nepristrani procjenitelj od β_1 što znači da smo od svih mogućih regresijskih procjenitelja, vjerojatno, odabrali one najbolje. Nepristranost je dobra karakteristika koeficijenta jer nam ona govori da je procijenjeni regresijski koeficijent veoma blizu stvarne vrijednosti regresijskog koeficijenta, no, ono što je bitnije od nepristranosti, je da procjenitelj ima malu srednjekvadratnu grešku.

Definicija 2.1.5. *Srednjekvadratna pogreška (u oznaci MSE) procjenitelja $\hat{\beta}_i$ za parametar β_i, $i = 1, 2$ je broj:*

$$MSE(\hat{\beta}_i) = \mathbb{E}\left[(\hat{\beta}_i - \beta_i)^2\right]$$

. U slučaju da je $\hat{\beta}_i$ nepristrani procjenitelj za β_i, tada je $MSE(\hat{\beta}_i) = \mathbb{V}ar(\hat{\beta}_i)$.

Definicija 2.1.6. *Kažemo da je procjenitelj konzistentan ili asimptotski nepristran ako njegova srednjekvadratna pogreška teži ka nuli kada veličina uzorka raste u beskonačnost:*

$$MSE(\hat{\beta}_i) \to 0,\ n \to \infty.$$

Kasnije u radu, kod višestruke linearne regresije, ovaj dio je ekvivalentan dokazivanju Gauss-Markovljevog teorema. Kako varijance procjenitelja ovise o broju mjerenja, regresijski pravac je precizniji. Ako je uzorak stvarnih mjerenja veći. Dodatno, varijance regresijskih koeficijenata ovise i o varijanci reziduala, stoga regresijski pravac ima manju varijancu ukoliko su eksplanatorne varijable dobro raspršene.

b. Standardna devijacija grešaka i intervali pouzdanosti

Napomenuli smo da vrijede Gauss-Markovljevi uvjeti, a jedan od njih je da su varijance reziduma konstantne: $\mathbb{V}ar[\epsilon_i] = \sigma^2,\ \forall i$ i međusobno nezavisne za svako i-to mjerenje. U 2.1.4 te G-M uvjeta pojavljuje se konstanta σ^2 koju zovemo standardna devijacija grešaka definirajući: $\mathbb{V}ar[\epsilon_i] = \sigma^2$. Njezinu vrijednost računamo koristeći rezultate sparnih mjerenja kao:

$$\begin{aligned} \sigma^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\epsilon_i^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned} \tag{2.19}$$

gdje vrijedi $\mathcal{X}^2$ distribucija sa $n-2$ stupnja slobode. U $\mathcal{X}^2$ distribuciji općenito je n stupnjeva slobode, no pri određivanju standardne devijacije gubimo dva stupnja slobode zbog procjene parametara $\hat{\beta}_0$ i $\hat{\beta}_1$. [10]

Napomena: S $\hat{\sigma}$ označavamo procijenjenu varijancu reziduala dobivenu na temelju sparnih mjerenja koju zovemo još i nepristrani procjenitelj. Nepristrana standardna devijacija grešaka se češće označava sa s.

U 2.1.4 smo dokazali da su procjenitelji $\hat{\beta}_0, \hat{\beta}_1$ nepristrani, odnosno da su blizu stvarne vrijednosti β_0 i β_1. No, kako oni ovise o izmjerenim podacima, mijenjajući podatke, mijenjaju se i procjenitelji. Primjerice, ukoliko dodamo ili oduzmemo jedno od mjerenja iz uzorka, koeficijenti se blago mijenjaju. Upravo zbog te labilnosti, odredit ćemo interval pouzdanosti za regresijske koeficijente te tako osigurati raspon vrijednosti za koje ćemo s 95% sigurnošću tvrditi da se unutra nalaze stvarni koeficijenti regresije.
Raspon određujemo tako da izračunamo gornju i donju granicu intervala pomoću stvarnih mjerenja. Za konstrukciju intervala pouzdanosti potrebne su nam vrijednosti standardne devijacije grešaka za koeficijente $\hat{\beta}_0, \hat{\beta}_1$ uz pomoću kojih dolazimo do odgovora na pitanje: Koliko daleko možemo odstupiti od nepristranog koeficijenta?

Definicija 2.1.7. *Neka su* $\mathbb{V}ar(\hat{\beta}_0), \mathbb{V}ar(\hat{\beta}_1)$ *procjenitelji nepristranih regresijskih koeficijenata i* $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ *nepristrani procjenitelj standardne devijacije grešaka. Uz uvjet da je* $\beta_0 \neq 0$ *definiramo standardne greške za* $\hat{\beta}_0$ *i* $\hat{\beta}_1$*:*

$$\begin{aligned} s.e(\hat{\beta}_0) &= s\sqrt{\frac{1}{n} + \frac{\hat{x}^2}{\sum_{i=1}^{n}(x_i - \hat{x})^2}} \\ s.e(\hat{\beta}_1) &= \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2}}. \end{aligned} \tag{2.20}$$

Uz pretpostavku da vrijede G-M uvjeti i da su reziduali normalno distribuirani, konstruirajmo intervale pouzdanosti uvrštavajući 2.20:

$$\begin{aligned} \beta_0 &\in \left[\hat{\beta}_0 - t_{n-2,\alpha/2} \cdot s.e.(\hat{\beta}_0),\ \hat{\beta}_0 + t_{n-2,\alpha/2} \cdot s.e.(\hat{\beta}_0)\right] \\ \beta_1 &\in \left[\hat{\beta}_1 - t_{n-2,\alpha/2} \cdot s.e.(\hat{\beta}_1),\ \hat{\beta}_1 + t_{n-2,\alpha/2} \cdot s.e.(\hat{\beta}_1)\right] \end{aligned} \tag{2.21}$$

gdje je t_{n-2} Studentova t-distribucija s $n-2$ stupnja slobode. [11]

Sada sa sigurnošću od 95% tvrdimo da se pravi koeficijenti regresije nalaze unutar konstruiranih intervala.

Postavlja se novo pitanje: Što ako intervali pouzdanosti sadrže 0? To bi značilo da bi ili β_0 ili β_1 bili jednaki 0. Ne predstavlja $\beta_0 = 0$, no u slučaju da je $\beta_1 = 0$ tada to znači da između eksplanatorne i odzivne varijable ne postoji ovisnost pa odzivna varijabla ne ovisi o eksplanatornoj:

$$Y = \beta_0 + \epsilon$$

stoga nema smisla modelirati jednostavnom linearnom regresijom. Kako bismo provjerili može li $\hat{\beta}_1$ biti nula, odnosno je li razumno tražiti regresijski pravac, provodimo t-testnu statistiku za regresijske koeficijente postavljajući hipoteze:

$$H_0 : \beta_1 = 0 \tag{2.22}$$

$$H_1 : \beta_1 \neq 0 \tag{2.23}$$

Želimo se uvjeriti da je $\hat{\beta}_1$ dovoljno daleko od 0 kako bismo sa sigurnošću mogli reći da je β_1 različit od 0. Za taj cilj, dodatno pogledajmo $\mathbb{V}ar(\beta_1)$. Ukoliko je vrijednost varijance mala, to znači da vrijednost koeficijenata $\hat{\beta}_1$ nije daleko od stvarne vrijednosti jer je njihova raspršenost mala, pa nam to daje sigurnost da je koeficijent β_1 definitivno različit od 0 i samim time postoji veza između X i Y varijabli. U suprotnom, ako je vrijednost $\mathbb{V}ar(\hat{\beta}_1)$ velika, to nam govori da je raspršenost velika pa tada i vrijednost $|\hat{\beta}_1|$ mora biti značajno velika kako bismo sa sigurnošću odbaciti nul-hipotezu. Nul-distribucija za t-test koja mjeri standardnu devijaciju za koju je $\hat{\beta}_1$ udaljena od 0, dana je sa:

$$t_{H_0} = \frac{\beta_1 - 0}{s.e(\hat{\beta}_1)}. \tag{2.24}$$

Grafički, t-distribucija ima zvonolik oblik pa vrijednosti za više mjerenja ($n > 30$) možemo promatrati analogno kao da gledamo standardnu normalnu distribuciju. Sukladno tome, računamo vjerojatnost promatranja broja takvog da je njegova vrijednost veća ili jednaka apsolutnoj vrijednosti od t uz početnu pretpostavku da je $\beta_1 = 0$. To je analogno kao da promatramo p-vrijednost i na temelju nje odlučujemo o odbacivanju ili prihvaćanju nul-hipoteze. Relativno mala p-vrijednost ($p < 0.1$) nam daje odgovor da postoji veza između X i Y pa možemo odbaciti nul-hipotezu. Dodatno možemo provjeriti nalazi li se p u intervalu pouzdanosti jer regija prihvaćanja je ekvivalentna intervalu pouzdanosti, pa ako se p-vrijednost nalazi unutra, također možemo odbaciti nul-hipotezu i zaključiti da veza između odzivne i eksplanatorne varijable postoji.

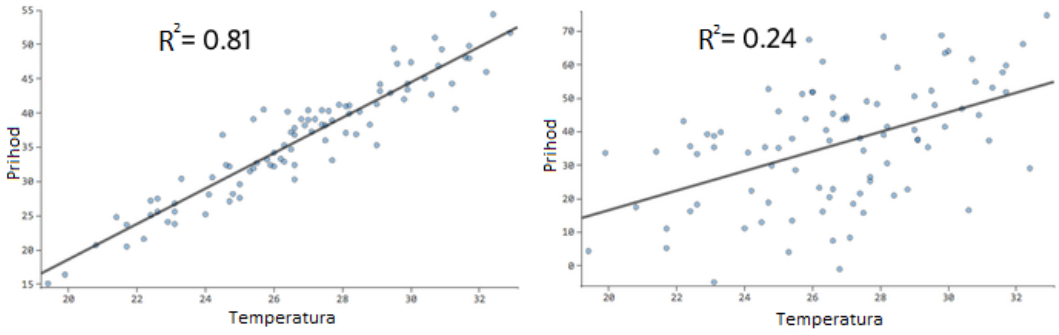
c. Test jakosti modela - Koeficijent determinacije

Procijenili smo koeficijente regresije, odbacili nul-hipotezu u korist alternativne hipoteze i time dokazali da veza između eskplanatorne i odzivne varijable uistinu postoji. Preostalo je još jedno pitanje: Koliko dobro procijenjeni regresijski pravac modelira i predviđa stvarne podatke? Odgovor na to pitanje daje nam koeficijent determinacije i rezidualna odstupanja.

Koeficijent determinacije, u oznaci R^2, govori o raspršenosti podataka te time opisuje jačinu linearne povezanosti eksplanatorne i odzivne varijable. Jačina povezanosti je ekvivalentna informaciji koliko rasipanja izlaznih podataka nastaje zbog funkcijske ovisnosti, a koliko otpada na rezidualno rasipanje. Dan je sa:

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{n}(\epsilon_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \ R^2 \in [0, 1]. \end{aligned} \tag{2.25}$$

Na slici 2.2, lijevi dijagram raspršenja prikazuje podatke čiji je koeficijent determinacije



Slika 2.2: Koeficijent determinacije, https://devopedia.org/regression-modelling

0.81, dok je za desni graf koeficijent determinacije 0.24. Uočimo da su na lijevoj slici stvarne vrijednosti bolje grupirane i zgusnute oko linearnog pravca, dok na desnoj slici su podaci više raspršeni i teže je proizvoljno nacrtati pravac. Prema Chadockovoj ljestvici koja je prikazana tablicom 2.1, između podataka koji su prikazani lijevim dijagramom raspršenja postoji čvrsta linearna zavisnost, dok iz desnog dijagrama i pripadnog koeficijenta determinacije možemo zaključiti da je linearna zavisnost slaba.

R^2	**Linearna zavisnost**
0.00	ne postoji
0.00 – 0.25	slaba
0.25 – 0.64	srednje jaka
0.64 – 1.00	čvrsta
1.00	potpuna

Tablica 2.1: Chadockova ljestvica

U tablici 2.1 prikazana je Chadockova ljestvica koja pojašnjava ovisnost koeficijenta korelacije i jakosti linearne veze između eksplanatorne i odzivne varijable. Što je R^2 bliže 1, to je i linearna ovisnost modela jača pa se veći postotak raspršenosti može objasniti pomoću eksplanatorne varijable. No, ako je R^2 bliže 0, to znači da je linearna veza slabija pa sve ako je $R^2 = 0$ linearna veza ne postoji.

d. Predviđanje odziva

Osnovni cilj linearne regresije je predviđanje vrijednosti za neku vrijednost eksplanatorne varijable X. Želimo procijeniti vrijednost odaziva u točki x_i ako smo odredili da je ovisnost varijabli:

$$\hat{y}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i, \ \forall i. \tag{2.26}$$

Odaberemo i-tu varijablu x_i iz domene i zanima nas kolika će otprilike biti njezina vrijednost? Očekivana vrijednost od Y za i-tu vrijednost x_i:

$$\mathbb{E}[Y|X = x_i] = \beta_0 + \beta_1 x_i, \ \forall i$$

na osnovi sparnih mjerenja. Koristimo li nepristrane procjenitelje, tada je očekivana vrijednost dana sa:

$$\hat{\mathbb{E}}[Y|X = x_i] := \hat{\beta}_0 + \hat{\beta}_1 x_i, \ \forall i. \tag{2.27}$$

Varijanca procjenitelja očekivane vrijednosti je dana sa:

$$\mathbb{V}ar[\hat{\mathbb{E}}[Y|X = x_i]] = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]. \tag{2.28}$$

Sada analogno kao i u 2.20 slijedi da je procijenjena standardna devijacija za procjenitelj $\hat{y}_i$:

$$s.e(\hat{y}_i) = \sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{2.29}$$

Uz 2.29 konstruiramo interval pouzdanosti za očekivanje srednjeg odaziva s obzirom na odabranu eksplanatornu vrijednost uz pomoć t-distribucije sa 95% sigurnošću da se stvarni odaziv nalazi u intervalu:

$$\hat{\mathbb{E}}[Y|X = x_i] : \hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{n-2,\alpha/2}\hat{\sigma} \cdot s.e(\hat{y}_i). \tag{2.30}$$

No, ovi intervali aproksimiraju očekivanu vrijednost zavisne varijable za odabranu nezavisnu varijablu. Takve intervale zovemo intervali pouzdanosti za očekivanje odaziva. Želimo odrediti interval u kojem se nalazi konkretna vrijednost zavisne varijable pa procjenjujemo iznos jedne konkretne vrijednosti Y za dani $X = x_i$. Time dolazimo do individualnog odziva $y - i$ za x_i-tu vrijednost. Tada je nepristrani procjenitelj za slučajnu vrijednost

$$\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i, \ \forall i,$$

a varijanca slučajne pogreške koja nastaje je jednaka:

$$\begin{aligned} \mathbb{V}ar[\hat{Y}_i - Y_i] &= \mathbb{V}ar[(\hat{\beta}_0 + \hat{\beta}_1 x_i) - (\beta_0 + \beta_1 x_i + \epsilon_i)] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_i - \hat{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \end{aligned} \tag{2.31}$$

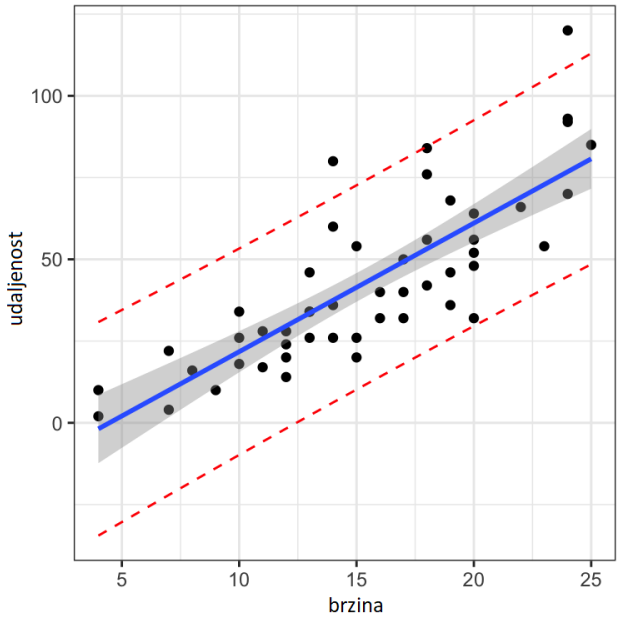
Uz 2.31 konstruiramo predikcijski interval - pouzdani interval za individualni odziv:

$$\hat{y}(x) \in \left[\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2,\alpha/2}\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right] \tag{2.32}$$

gdje za pogrešku procjene vrijedi:

$$t_{H_0} = \frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 + \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t(n-2).$$

Na slici 2.3 prikazan je dijagram raspršenja i označen je interval pouzdanosti crvenom isprekidanom linijom, dok je predikcijski interval obojen sivom bojom.



Slika 2.3: Predikcijski interval i interval pouzdanosti

2.2 Višestruka linearna regresija

Jednostavnom linearnom regresijom predviđali smo razne vrijednosti regresijske funkcije proučavajući vezu između jedne eksplanatorne i jedne odzivne varijable. U praksi, češće imamo više od samo jedne eksplanatorne varijable pa samim time imamo više faktora koji utječu na odaziv funkcije. U ovom poglavlju objasnit ćemo na koji način konstruirati model višestruke linearne regresije te što sve utječe na njegovu preciznost.

Najprije odredimo jedinstveni predikcijski model za dvije međusobno nezavisne eksplanatorne varijable. Dosadašnjim znanjem možemo konstruirati dva zasebna regresijska modela gledajući zasebno svaku eksplanatornu varijablu što znači da trebamo proučiti 2 različita regresijska pravca i jedinstveno ih spojiti. No, na koji način od dva različita regresijska pravca odrediti jedan koji će precizno povezati obje eksplanatorne varijable?

Umjesto da odredimo dva različita regresijska modela, možemo proširiti model jednostavne linearne regresije. Neka su X_1 i X_2 dvije nezavisne eksplanatorne varijable te njima dodajmo pripadne koeficijente β_1 i β_2. Regresijski model za određivanje vrijednosti i-tog mjerenja:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 x_{i2} + \epsilon_i. \tag{2.33}$$

Model možemo zapisati kao:

$$Y_i = \beta_0 + \sum_{k=1}^{2} \beta_k X_k + \epsilon_i. \tag{2.34}$$

- Y_i određuje vrijednost i-tog ispitivanja
- X_{i1} i X_{i2} reprezentiraju vrijednosti eksplanatornih varijabli za i-to ispitivanje
- ϵ_i je rezidual za i-to ispitivanje
- $\beta_0, \beta_1, \beta_2$ su regresijski koeficijenti

Nadalje, kako se reziduali mijenjaju u ovisnosti o regresijskim koeficijentima, tražimo one regresijske koeficijente za koje će suma reziduala biti najmanja moguća. Idealno bi bilo da je ona jednaka 0 pa pretpostavimo:

$$\mathbb{E}(\epsilon_i) = 0, \forall i.$$

Tada je očekivana funkcijska vrijednost ekvivalentna:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{2.35}$$

Parametre β_1, β_2 zovemo **parcijalnim regresijskim koeficijentima** jer objašnjavaju parcijalan efekt jedne eksplanatorne varijable kada su ostale varijable konstantne. β_1 objašnjava promjenu srednje vrijednosti odziva po jedinici povećavanja za varijablu X_1 kada je X_2 varijabla konstantne vrijednosti.

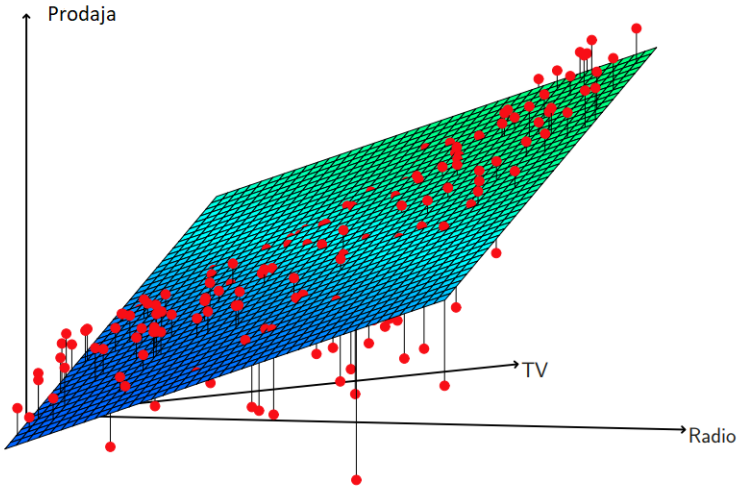
Kod jednostavne linearne regresije grafički prikaz modela je pravac i pomoću njega smo vizualno prikazali model, u ovom primjeru s dvije eksplanatorne varijable je ravnina kao što je prikazano na slici 2.4. Svaka točka na ravnini odgovara aritmetičkoj vrijednosti $\mathbb{E}(Y)$ na temelju danih vrijednosti varijabla X_1 i X_2. [10]

Možemo li dodatno proširiti model za p međusobno nezavisnih eksplanatornih varijabli? Svakoj varijabli $X_1, X_2, ..., X_p$ pridružimo njezin regresijski koeficijent:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon, \tag{2.36}$$

gdje X_p reprezentira p-tu eksplanatornu varijablu, a pripadni koeficijent smjera β_p povezanost između X_p i odzivne varijable Y u smislu prosječnog utjecaja X_p na Y povećavajući X_p za jednu jedinicu dok su sve ostale eksplanatorne varijable fiksne. Generalno, **opći linearan regresijski model** za izračun vrijednosti za i-to mjerenje:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i. \tag{2.37}$$



Slika 2.4: Model višestruke linearne regresije s dvije eksplanatorne varijable [9]

Dodatno, želimo li ispisati svih i promatranja, dobivamo:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_p x_{1p} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_p x_{2p} + \epsilon_2 \\ &\vdots \\ y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i. \end{aligned} \tag{2.38}$$

Uočimo da je ovakav zapis pomalo konfuzan te da ćemo teško provoditi daljnje izračune jer trebamo zapisati sustav od i jednadžbi. No, uočimo da ovakvu formu lako zapišemo u matričnoj formi:

$$Y = \beta X + \epsilon. \tag{2.39}$$

Y je vektor odgovora, ϵ vektor reziduala, X matrica dizajna, a β vektor parametara:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{i1} & x_{i2} & \dots & x_{ip} \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \tag{2.40}$$

Modeli 2.38 i 2.39 su ekvivalentni te ih zovemo **višestruki linearni regresijski model**. Što se tiče grafičkog prikaza, teško ćemo konstruirati dijagram raspršenosti i regresijski model jer bi to značilo prikaz hiperravnine u p-dimenzionalnom Koordinatnom sustavu.

Proširili smo model, no nismo odredili optimalne, točne regresijske koeficijente. Cilj je odredit regresijske koeficijente takve da je suma kvadrata reziduala minimalna pa pomoću metode najmanjih kvadrata, analogno kao i kod jednostavne linearne regresije, izračunamo regresijske koeficijente. [9]

Metoda najmanjih kvadrata u višestrukoj linearnoj regresiji

Želimo odrediti model koji najbolje povezuje dana mjerenja s ciljem predviđanja nekih nepoznatih mjerenja. Definirajmo funkciju $S(\beta_0, ..., \beta_p)$ koja reprezentira sumu kvadratnih reziduala:

$$\begin{aligned} S(\beta_0, ..., \beta_p) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) \right]. \end{aligned} \tag{2.41}$$

U daljnjem izračunu koristimo matrični zapis:

$$\begin{aligned} S(\beta_0, ..., \beta_p) &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - YX\beta - X^T \beta^T Y + X^T \beta^T X\beta \end{aligned}$$

Kako se radi o skalaru, vrijedi $Y^T X\beta = YX^T \beta^T$:

$$S(\beta_0, ..., \beta_p) = Y^T Y - 2\beta^T (X^T Y) + \beta^T (X^T X)\beta. \tag{2.42}$$

Traženjem minimuma funkcije 2.42 odrediti ćemo odgovarajući procjenitelj β za koji suma S postiže minimalnu vrijednost. Procijenjeni regresijski koeficijent za koji je suma minimalna zapišimo u oznaci $\mathbf{b}$: $\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$ gdje je b_p procijenjeni parametar za β_p.

Za minimalnu, koja je ujedno i ekstremna vrijednost funkcije više varijabli vrijedi, sumu vrijedi da je njezin gradijent, u ekstremnoj točki funkcije, jednak 0:

$$\begin{aligned} &\nabla S(\mathbf{b}) = 0 \\ \Leftrightarrow\ &\frac{\partial S}{\partial \beta_i}(\mathbf{b}) = 0, \ \forall i. \end{aligned}$$

Kako je funkcija zapisana matrično, gradijent funkcije odgovara diferencijaciji matrice:

$$\begin{aligned} \frac{\partial S}{\partial \beta}(\mathbf{b}) &= 0 \\ -2X^T Y + 2X^T X\mathbf{b} &= 0 \\ (2X^T X)\mathbf{b} &= 2X^T Y / \ (2X^T X)^{-1} \ sli\, jeva \\ \mathbf{b} &= (2X^T X)^{-1}(2X^T Y) \end{aligned} \tag{2.43}$$

Kako znamo da rješenje postoji? Da bi rješenje postojalo, matrica $X'X$ mora biti regularna jer za nju postoji inverzna matrica.

Definicija 2.2.1. *Kvadratna matrica $A \in M_n(\mathbb{F})$ je regularna ili invertibilna ako postoji matrica $B \in M_n(\mathbb{F})$ takva da je*

$$AB = BA = I.$$

Matricu B zovemo inverznom matricom od A i pišemo $B = A^{-1}$. U protivnom kažemo da je A singularna matrica.

U slučaju kada je matrica regularna, ona je punog ranga i tada postoji jedinstveno rješenje. Kako je matrica dimenzije $p \times p$, slijedi da su sve eksplanatorne varijable linearno nezavisne. U protivnom, kažemo da je matrica $X'X$ singularna. Singularnost matrice može biti uzrokovana:

(a) **Većim brojem prediktora nego izmjerenih podataka**
Ukoliko je broj eksplanatornih varijabli manji od broja mjerenja ($p < n$) tada modeliramo regresijom koja je preparametrizirana. Imamo previše prediktora, a premalo stvarnih vrijednosti pa ne možemo konstruirati regularnu matricu dimenzije pxp i rješenje nikako ne može biti jedinstveno.

(b) **Dupliciranim varijablama**
Pri korištenju stvarnih podataka, koristimo i mjerne jedinice. Može se dogoditi da se mjerenje duplicira promjenom mjerne jedinice. Primjer: izmjerena masa u kilogramima ili izmjerena masa u tonama.

(c) **Cirkularnim varijablama**

U tim slučajevima, regresijski procjenitelj neće biti jedinstven pa modificiramo model:

$$\mathbf{b} = (X^T X)^{-1} X^T Y = X^{-1} Y.$$

Napomenimo da u ovom slučaju $\mathbf{b}$ nije jedinstven regresijski koeficijent.

Analogno kao i kod jednostavne linearne regresije, želimo se uvjeriti da je dobiveni regresijski koeficijent dobar te da ne postoji drugi koji bi bolje opisao dane podatke. Ako je dobar koeficijent r, želimo se uvjeriti i da je sam model dobar i da ćemo njime moći procjenjivati vrijednosti pa tražimo odgovore na pitanja:

1. Postoji li "bolji" regresijski koeficijent od **b**?
   Odgovor: Nepristranost regresijskog koeficijenta.

2. Postoji li uopće veza između (barem jedne) eksplanatorne i odzivne varijable? Pomažu li sve eksplanatorne varijable pri provjeri odzivne varijable ili su samo neke od njih korisne?
   Odgovor: Test značajnosti procjenitelja.

3. Koliko dobro model povezuje stvarne vrijednosti?
   Odgovor: Koeficijent determinacije.

4. Ukoliko nam je dan skup eksplanatornih varijabli, koji odaziv možemo očekivati i koliko je dobra naša predikcija?
   Odogovor: Predikcija.

Nepristranost regresijskog koeficijenta

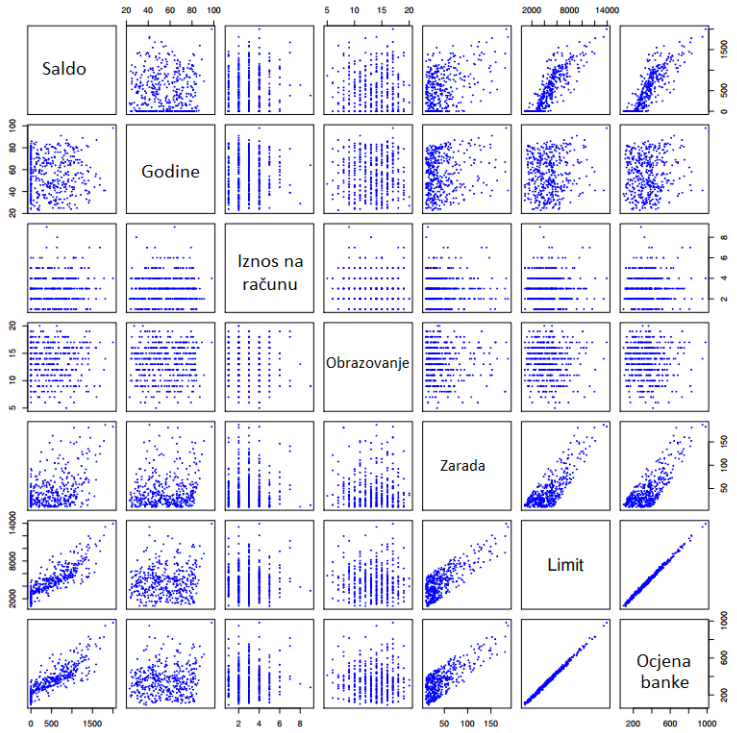
Kako bismo bili sigurni da je dobiveni regresijski koeficijent **b** dobar i da ne postoji bolji od njega, trebaju biti zadovoljeni Gauss-Markovljevi uvjeti. Kod jednostavne linearne regresije smo dokazali da Gauss-Markovljevi uvjeti vrijede za greške dobivene metodom najmanjih kvadrata. Pa se analogno može provjeriti da su uvjeti zadovoljeni i kod višestruke linearne regresije. Iskažimo 2.1.2 u matričnom obliku:

- centriranost: $\mathbb{E}(\epsilon) = 0$
- jednakost varijanci: $\mathbb{V}ar(\epsilon_i^2) = \sigma^2, \ \forall i$
- nekolinearnost: $\mathbb{E}(\epsilon_i \epsilon_j) = 0, \forall i.$

Vrlo česta je pojava pri konstrukciji regresijskog modela da nije zadovoljena nekolinearnost. Drugim riječima, neke od eksplanatornih varijabli su međusobno zavisne pa kažemo da postoji multikolinearnost. Najčešći pokazatelji da je ona prisutna u regresijskom modelu:

- znatno širi intervali pouzdanosti
- kod testa hipoteza, rezultati idu u prilog prihvaćanju hipoteze H_0

- ukupna F-vrijednost je značajna, no gotovo nijedan od individualnih p-vrijednosti pojedine eksplanatorne varijable nije značajan



Slika 2.5: Multikolinearnost [9]

Multikolinearnost se može uočiti na više načina. Konstrukcijom grafa i tablice korelacije između reziduala (analiza reziduala). Primjer takvog grafa prikazan je na idućoj slici 2.5 koja prikazuje graf korelacije između 7 eksplanatornih varijabli. Iz grafa lako zaključimo da postoji linearna veza između ocjene i ograničenja jer točne formiraju pravilan rastući pravac što je grafički prikaz linearne funkcije. Cilj je izbaciti sve eksplanatorne varijable koje su međusobno zavisne kako bi bili zadovoljeni Gauss-Markovljevi uvjeti te kako bismo konstruirali što precizniji regresijski model.

Postoji više metoda kako izbaciti eksplanatorne varijable koje su međusobno u korelaciji. Najpoznatija tri principa ”forward selection, backward selection i mixed selection” su detaljnije opisana u [9], poglavlje 3.3..

Propozicija 2.2.2. *Pretpostavimo da su zadovoljeni Gauss-Markovljevi uvjeti. Procjenitelj* $\mathbf{b}$ *je nepristrani od* β *ako vrijedi* $\mathbb{E}[\mathbf{b}] = \beta$.

Dokaz. Iz pretpostavke slijedi da je $\mathbb{E}[\epsilon_i] = 0$ pa uzimajući to u obzir zajedno s 2.43 vrijedi:

$$\begin{aligned}\mathbb{E}[\mathbf{b}] &= \mathbb{E}[(X'X)^{-1}X'Y] + 0 \\ &= (X^TX)^{-1}(X^TX)\beta \\ &= I\beta = \beta.\end{aligned}$$

□

Ovime smo zaključili da ne postoji drugi regresijski koeficijent za koji će biti suma kvadratnih reziduala manja pa zaključujemo da je $\mathbf{b}$ nepristran regresijski procjenitelj. Dodatno, u svrhu daljnje distribucije procjenitelja, uvodimo zahtjev: $\epsilon_i \sim N(0, \sigma^2)$ (koji proizlazi iz Gauss-Markovljevih uvjeta) što znači da su rezidualni normalno distribuirani. No ako je neka varijabla normalno distribuirana, tada možemo provoditi testiranja temeljena na distribucijama sa sigurnošću da su dodatna odstupanja u okviru prihvaćanja. Iz tog slijedi da su $\mathbf{b}$ i $\hat{Y}$ također normalno distribuirane varijable jer one ovise o rezidualima i mjerenjim vrijednostima.

Propozicija 2.2.3. *Pretpostavimo da vrijede Gauss-Markovljevi uvjeti i da je* ϵ_i *normalno distribuiran. Tada vrijedi:*

$$\begin{aligned}\mathbf{b} &\sim N(\beta, \sigma^2(X^TX)^{-1}) \\ \hat{y} &\sim N(X\beta, \sigma^2(X^TX)^{-1}X^T).\end{aligned}$$

Dokaz. Dokazali smo da je $\mathbb{E}[\mathbf{b}] = \beta$.

$$\begin{aligned}\mathbb{V}ar[\mathbf{b}] &= (X^TX)^{-1}X \cdot \mathbb{V}ar[Y] \cdot ((X^TX)^{-1}X)^T \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}.\end{aligned}$$

Kako je $\mathbf{b} \sim N(\mathbb{E}(\mathbf{b}), \mathbb{V}ar(\mathbf{b}))$, uvrštavanjem slijedi $\mathbf{b} \sim N(\beta, \sigma^2(X^TX)^{-1})$. Analogno,

$$\begin{aligned}\mathbb{E}[\hat{Y}] &= \mathbb{E}(X\mathbf{b} + \epsilon)X\beta \\ &= X\beta\end{aligned}$$

$$\begin{aligned}\mathbb{V}ar[\hat{Y}] &= \mathbb{V}ar[\mathbf{b}X + \epsilon] \\ &= X\mathbb{V}ar[\mathbf{b}]X^T \\ &= \sigma^2X(X^TX)^{-1}X^T.\end{aligned}$$

Zaista, $\hat{y} \sim N(X\beta, \sigma^2(X^TX)^{-1}X^T)$. □

Distribucija greške može odstupati od normalne distribucije. Sigurnost da će procjenjitelji i dalje imati normalnu distribuciju nam daje centralni granični teorem. Neka manja odstupanja se mogu tolerirati jer centralni granični teorem osigurava asimptotski normalnu distribuciju.

Teorem 2.2.4. ***Centralni granični teorem***
Neka je $X_1, X_2, ...$ niz nezavisnih jednako distribuiranih slučajni varijabli s konačnim matematičkim očekivanjem μ i konačnom varijancom $\sigma^2 > 0$. Nadalje, neka je $\bar{X}_n := \frac{X_1+X_2+...+X_n}{n}$ za sve prirodne brojeve n. Tada za sve $a < b$ vrijedi

$$\lim_{n\to+\infty} \mathbb{P}\left(a \leq \frac{\bar{X}_n - \Phi}{\sigma}\sqrt{n} \leq b\right) = \Phi(b) - \Phi(a), \tag{2.44}$$

gdje je $\Phi(x)$ funkcija distribucije jedinične normalne razdiobe.

Dokaz. Detaljan dokaz ovog teorema vidi u [12], poglavlje 8. □

Ovime smo dokazali da su koeficijenti **b** i $\hat{y}$ normalno distribuirani i da su oni nepristrani. Sada regresijski model s nepristranim procjenjiteljem možemo zapisati:

$$\hat{Y} = X\mathbf{b},$$

gdje je

$$\hat{\mathbf{Y}} = \begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_p} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_n} \end{pmatrix}.$$

Uz poznatu distribuciju, zanima nas vrijednost varijance reziduala kako bismo mogli vršiti testiranja i procjenjivati vrijednosti zavisne varijable konstruirajući intervale pouzdanosti. Varijanca greške se dobiva skaliranjem sume kvadrata reziduala. Kako smo procijenili regresijski koeficijent **b** te dokazali da je nepristran, tako je i procijenjena varijanca greške nepristrana. U obzir uzimamo prikladne stupnjeve slobode koji su ekvivalentni broju podataka umanjenima za broj procjenjenih parametara:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n-p} = \frac{e^T e}{n-p}. \tag{2.45}$$

Ponekad se u različitim literaturama koristi oznaka s_e^2.

Test značajnosti procjenjitelja

Kod jednostavne linearne regresije test značajnosti procjenjitelja provjeravali smo za samo jednu eksplanatornu varijablu: $\beta_1 \neq 0$ jer samo ona određuje postoji li ili ne linearna veza. Kod višestruke linearne regresije imamo p različitih eksplanatornih varijabli pa test moramo proširiti i prilagoditi za p različitih eksplanatornih varijabli. Prva opcija je provjeriti za svaku varijablu zasebno postoji li veza, no to bi značilo izradu p različitih testova. Ako je p velik, primjerice $p = 100$, trebalo bi nam puno vremena za analizu podataka. Svakako postoji efikasnije rješenje. U ovom poglavlju analizirat ćemo test značajnosti za p različitih procjenjitelja.

Postavljamo test hipoteza. Označimo nul-hipotezom tvrdnju da su svi regresijski koeficijenti jednaki 0. Dok alternativna hipoteza tvrdi da je barem jedan od regresijskih koeficijenata različit od 0, jer ukoliko je barem jedan od $\beta_1, ..., \beta_p$ koeficijenata različit od 0, to znači da postoji veza između eksplanatorne i odzivne varijable. Test hipoteza:

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \\ &H_a : \beta_j \neq 0, \ \ j \in \{1, ..., p\}. \end{aligned} \tag{2.46}$$

Kako bismo mogli sa sigurnošću donijeti odluku koju hipotezu prihvaćamo, a koju odbijamo, koristimo t-statistiku. F vrijednost t-statistike je dana sa:

$$f = \frac{MSR}{MSE} = \frac{\frac{SSR}{p}}{\frac{SSE}{n-(p+1)}} = \frac{\frac{\sum_{i=1}^{p}(y_i-\bar{y})^2-\sum_{i=1}^{p}(y_i-\hat{y}_i)^2}{p}}{\frac{\sum_{i=1}^{p}(y_i-\hat{y}_i)^2}{n-(p+1)}}. \tag{2.47}$$

Uspoređujemo 2.47 i t-distribuciju sa $n - (p + 1)$ stupnjeva slobode koju definiramo kao: $F^* = F_{\alpha,p,n-(p+1)}$ pa donosimo zaključak:

ako je $f <= F^*$ tada prihvaćamo nul-hipotezu H_0
ako je $f > F^*$ tada odbacujemo nul-hipotezu.

Dodatno, kada nema povezanosti između odzivne i eksplanatorne varijable, očekujemo da je vrijednost F-statistike blizu 1, jer ukoliko prihvatimo H_0 hipotezu tada tvrdimo da ne postoji veza između varijabli pa slijedi:

$$\mathbb{E}\left(\frac{RSS}{n-(p+1)}\right) = \mathbb{E}\left(\frac{TSS-RSS}{p}\right) = \sigma^2.$$

Obrnuto, kada odbacujemo H_0 hipotezu u korist H_a hipoteze, tada je $\mathbb{E}\left(\frac{TSS-RSS}{p}\right) > \sigma^2$ pa očekujemo da je i $F > 1$.

Stoga, ako je izračunata vrijednost F-statistike blizu 1, prihvaćamo hipotezu H_0 pa veza između eksplanatorne i odzivne varijable ne postoji. Obratno, ukoliko odbacujemo nul-hipotezu i prihvatimo alternativnu hipotezu, tada će vrijednost F-statistike biti veća od 1.

pa možemo zaključiti da postoji barem jedan β_j koji je u korelaciji s odzivnom varijablom Y.

Nameću se dodatna pitanja:

a. Koliko velika mora biti vrijednost F-statistike da bismo bili sigurni da možemo odbaciti hipotezu H_0?

   - Ako je n velik, tada je dovoljno da je $F > 1$ pa makar on bio samo neznatno malo veći od 1. To je dovoljan dokaz za odbacivanje H_0 hipoteze.
   - Ako je n malen, tada F vrijednost mora biti značajno veća od 1 kako bismo odbili H_0 hipotezu.

b. Odbijajući H_0 hipotezu znamo da postoji barem jedna eksplanatorna varijabla između koje postoji veza s odzivnom varijablom. Koja je to točno od p različitih varijabli? Koje varijable jesu, a koje nisu povezane s odzivnom varijablom?

   Potrebno je proučiti zasebne p-vrijednosti za svaku eksplanatornu varijablu. Statistički softver izračunava ukupnu F-vrijednost koja obuhvaća sve regresijske koeficijente te p-vrijednosti za pojedinu eksplanatornu varijablu. Vrijednost F-statistike može biti povoljna i ona nam potvrđuje da postoji veza između barem jedne eksplanatorne varijable, no pomoću p-vrijednosti možemo vidjeti koji prediktori nam nisu korisni u konstrukciji regresijskog modela. Ako je p-vrijednost i-tog prediktora blizu 0 to povlači da taj prediktor nije koristan u regresijskom modelu i ne doprinosi njegovoj preciznosti. Dodatnu provjeru za eksplicitnu eksplanatornu varijablu provodimo analogno kao kod jednostavne linearne regresije.

   S druge strane, ponekad ne želimo u test hipoteza uključiti svih β_p koeficijenata, već želimo provjeriti samo za određeni podskup koeficijenata. Zbog izdvajanja skupa eksplanatornih varijabli za koje smo sigurni da su povezane s odzivnom varijablom, razlikujemo potpuni model koji sadrži svih p prediktora i reducirani model. S obzirom da potpuni model sadrži prediktore reduciranog modela i dodatno sigurne prediktore, on pokriva podatke jednako kao i reducirani model. Provodimo F-test za grupu prediktora tako da pouzdan skup od l koeficijenata stavljamo na početak te pa postavljamo test hipoteza gdje je $l + k = p$:

$$\begin{aligned} &H_0 : \beta_{l+1} = \beta_{l+2} = ... = \beta_{l+k} = 0 \\ &H_a : \beta_j \neq 0, \ \ j \in \{l+1, ..., l+k\}. \end{aligned} \tag{2.48}$$

   Definiramo dvije zasebne sume kvadrata reziduala takve da je $SSE_p < SSE_r$, gdje je

$$\begin{aligned} SSE_p &- \text{suma kvadratnih reziduala za potpun model} \\ SSE_r &- \text{suma kvadratnih reziduala za reducirani model} \end{aligned}$$

Postupak odlučivanja koju hipotezu prihvaćamo ekvivalentan je, no različita je samo vrijednost testne statistike koju računamo kao

$$f = \frac{\frac{SSE_k - SSE_p}{p-k}}{\frac{SSE_p}{n-(p+1)}}.$$

Hipotezu H_0 odbacujemo ako je $f > F_{\alpha,p-k,n-(p+1)}$.

Koeficijent determinacije u višestrukoj linearnoj regresiji

Koeficijent determinacije R^2 govori o ukupnoj raspršenosti podataka koja nastaje uslijed modeliranja regresijom. Drugim riječima, opisuje jačinu linearne povezanosti. On mjeri proporcionalno smanjenje ukupnog odstupanja od stvarne vrijednosti Y u odnosu na broj eksplanatornih varijabli. Kod jednostavne linearne regresije on je jednak kvadratu korelacije između eksplanatorne i odzivne varijable, a kod višestruke linearne regresije dan je sa:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \end{aligned} \tag{2.49}$$

Ukoliko ne postoji korelacija između varijabli tada je vrijednost koeficijenta $R^2 = 0$, u suprotnom je $R^2 = 1$ i tada model pokriva veliki dio varijance u odgovoru.

Dodavanjem eksplanatornih varijabli povećava se vrijednost R^2 jer se povećava i suma kvadratnih reziduala. Kako bismo izbjegli donošenje krivih zaključaka zbog navedene promjene, koristimo modificirani koeficijent determinacije:

$$\begin{aligned} R_a^2 &= 1 - \frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \\ &= 1 - (\frac{n-1}{n-p})\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \end{aligned} \tag{2.50}$$

Svaka suma je podijeljena odgovarajućim stupnjevima slobode te ovakav prilagođen koeficijent determinacije može postati samo manji jer svako smanjenje sume kvadratnih reziduala se kompenzira smanjenjem stupnja slobode. Ukoliko je razlika $|R^2 - R_a^2|$ značajna to nam ukazuje na problem da model ima previše prediktora. No, svakako treba biti na oprezu pri donošenju zaključaka jer veliki koeficijent nas može dovesti do krivog zaključka. Zbog nevidljivih analitičkih grešaka, pomoću grafičkog prikaza lako možemo otkriti probleme krivog zaključka. [13]

Predikcija

Jednom kada smo odredili regresijski model i uvjerili se da je on dobar, želimo ga primijeniti kako bismo predvidjeli - odredili funkcijski odgovor na temelju danih izmjerenih podataka $X_1, ..., X_p$. Kako regresijski model procjenjuje očekivane funkcijske vrijednosti, postoje glavne 2 nesigurnosti pri toj procjeni.

1. Procijenjeni regresijski koeficijent **b** je samo procjena za stvarni koeficijent β. Proučavamo li grafički prikaz stvarnog modela i onog regresijskog, uspoređujemo dvije različite hiperravnine koje razlikuje greška procjene. Uključujući spomenutu srednju pogrešku (baziranu na svim mjerenjima), konstruiramo **interval pouzdanosti** i odredimo koliko blizu je srednja procijenjena vrijednost od srednje stvarne vrijednosti. Uz pretpostavku da je procijenjeni parametar nepristran i normalno distribuiran, konstruiramo intervale pouzdanosti za b:

2. Čak iako i da znamo stvarni odaziv i stvarni regresijski koeficijent, to ne znači da će naša predikcija biti potpuno točna. Zanima nas koliko smo točno procijenili pojedinu funkcijsku vrijednost. Zbog toga konstruiramo **intervale predikcije** koji su širi od intervala pouzdanosti jer oni uključuju dodatni rezidual koji reprezentira udaljenost određene točke na regresijskoj ravnini od stvarne vrijednosti. Dakle, zanima nas predikcijski interval za novi odaziv Y_k kada je vrijednost eksplanatorne varijable x_k. Formula pomoću koje određujemo granice:

Najvažnija razlika između intervala pouzdanosti i intervala predikcije je ta da širina intervala pouzdanosti je određena aritmetičkom rezidualnom sredinom, a interval predikcije dodatno sadrži i pojedini rezidual određene vrijednosti koju promatramo. Uzrok tome je razlika u standardnoj devijaciji.

Poglavlje 3

Analiza širenja COVID-19 pandemije

Pojavljivanje COVID-19 pandemije utjecalo je na gotovo sve države i uzrokovalo značajne gubitke ljudskih života te globalno ostavilo negativan trag u segmentima gospodarstva i ekonomije. Mnoge države su paralelno uvodile identične mjere restrikcija, poput socijalnog distanciranja, nošenja maski, samoizolacije i sl. No zanimljiva je činjenica da one bilježe drastične razlike u ukupnom broju pozitivnih na koronavirus i umrlih od istog. Upravo zbog te nelogičnosti, zanima nas koji su to drugi faktori koji utječu na povećanje ili smanjenje broja zaraženih virusom.

Proučavajući promjenu broja oboljelih i umrlih od COVID-19 pandemije u pojedinim državama, Chang i sur. [2] su identificirali 21 različitih faktora koji utječu na kretanje pandemije, kako u dobrom, tako i u lošem smjeru. Promatrali su gustoću naseljenosti pojedine zemlje, broj stanovnika, broj turista koji ulaze u državu pa čak i činjenicu je li na čelu zemlje žena ili muškarac. Tako je pokazano da političko-pravna slika države, pouzdanje stanovništva u vladu, državni BDP, tehnološki napredak i mnogi drugi faktori direktno utječu na porast ili smanjenje broja oboljelih od COVID-19 virusa, kao i umrlih. Svakako je zanimljiva činjenica kako tehnološki napredak povećava broj oboljelih, ali utječe na smanjenje broja umrlih. U ovom poglavlju proučiti ćemo i prezentirati raznolike faktore koji objašnjavaju zbog čega dolazi do promjena širenju COVID-19 pandemije u različitim zemljama, ako one donose identične mjere u identično vrijeme. Preuzeti podaci su iz perioda od 22. siječnja 2020. do 31. prosinca 2020. godine.

3.1 Faktori i njihov utjecaj na razvoj COVID-19 pandemije

U ovom dijelu opisat ćemo i nabrojiti faktore koji utječu na širenja ili suzbijanje COVID-19 pandemije. Sve promatrane varijable su odrednice država koje opisuju uređenje države

neovisno o pojavljivanju pandemije. Ukupno je promatrano 21 faktora koji u regresijskoj notaciji predstavljaju 21 eksplanatornu varijablu. Naknadno je dodana još jedna eksplanatorna varijabla koja je nastala kao posljedica pojavljivanja pandemije, a riječ je o broju testiranih osoba na virus. Sve zajedno, promatramo utjecaj 22 eksplanatorne varijable na promjenu broja pozitivnih i broja umrlih od virusa COVID-19 pandemije.

Opis i podjela faktora

Najprije zapišimo dvije glavne zavisne varijable:

- $Y_1 = ln(1+$ broj pozitivnih) - broj zaraženih na milijun osoba
- $Y_2 = ln(1+$ broj umrlih) - broj umrlih na milijun osoba.

Uočimo kao su varijable zadane pomoću prirodnog logaritma. Glavni razlog tome je što se SARS-CoV-2 virus, uzročnik COVID-19 pandemiju, širi eksponencijalno. Općenito, razmnožavanje virusa i bakterija modeliramo eksponencijalnom funkcijom.

Definirane zavisne varijable ne skaliramo pomoću broja stanovnika pojedine zemlje jer on nije prikladan za praćenje opsega kretanja pandemije. Iz [6] dodaju kako napredak pandemije je neovisan o ukupnom broju stanovnika zemlje pa je i broj oboljelih ili umrlih također neovisan o ukupnom broju stanovništva neke države, općenito. Iz tog razloga, broj stanovnika pojedine države gledamo kao eksplanatornu (nezavisnu) varijablu u regresijskom modelu.

Spomenuta je 21 varijabla koje utječe na COVID-19 pandemiju, a one su poredane u tablici 3.1. Možemo ih podijeliti u 4 velike skupine: s obzirom na demografsko-geografsku, političko-zakonsku, socijalno-ekonomsku sliku te skupina zdrastvene skrbi.

Demografsko – geografske varijable	**Političko – zakonske varijable**	**Socijalno – ekonomske varijable**	**Varijable zdravstvene skrbi**
• Stanovništvo • Gustoća stanovništva • Godine (srednja dob) • Muškarci (broj muškaraca na 100 žena) • Urbanizacija (razmjer stanovništva urbanog i ruralnog područja) • Temperatura (prosječna tjedna temperatura u °C) • Obrazovanje (stupanj obrazovanja) • Religijska opredijeljenost (religijska heterogenost)	• Demokracija (slobodni i pravedni politički izbori, odgovornost vlade prema stanovništvu) • Korupcija (indeks korupcije) • Sloboda medija (stupanj slobode medija) • Žena vođa (vrijednost 1 ako je na čelu države žene u protivnom je vrijednost 0) • Povjerenje vlastima (postotak stanovništva koji vjeruje nacionalnim vlastima) • Zakon (snaga i objektivnost u zakonski sustav)	• BDP (bruto društveno proizvod) • Nejednakost (Gini koeficijent) • Turizam (broj turista) • Tehnologija (ulaganje u razvoj tehnologije, pokrivenost 4G mrežom, uporaba društvenih mreža) • Zadovoljstvo (rezultat i mjera sreće, zadovoljstvo životom)	• SARS (broj SARS slučajeva evidentiranih u razdoblju od 1. studenog 2002. do 31. srpnja 2003.) • Bolnički kreveti (broj bolničkih kreveta spremnih za primitak hitnih pacijenata)

Slika 3.1: Eksplanatorne varijable

Spomenuli smo dodatnu, **kontrolnu varijablu**: broj testova.

Kada konstruiramo regresijske modele, pripadni regresijski koeficijenti mogu biti pozitivnog ili negativnog predznaka. Stoga, podijelimo ih u dvije skupine:

1. **Nepogodne varijable**
   - pripadni regresijski koeficijenti su pozitivnog predznaka
   - kontekst: varijabla pospješuje širenje pandemije
   - Stanovništvo, Gustoća stanovništva, Godine, Korupcija, BDP, Udio muškaraca, Urbanizacija, Demokracija, Tehnologija, Nejednakost, Turizam, Zadovoljstvo.

2. **Pogodne varijable**
   - pripadni regresijski koeficijenti su negativnog predznaka
   - kontekst: varijabla pospješuje suzbijanje pandemije
   - Temperatura, Obrazovanje, Religijska opredijeljenost, Sloboda medija, Žena vođa, Povjerenje vlastima, Zakon, SARS, Bolnički kreveti, Broj testova.

Deskriptivna statistika

Analiza se temelji na 99 različitih zemalja koje se grupiraju u 4 skupine na temelju ukupnog broja zaraženih i umrlih od virusa. Izdvojene države su SAD, Indija, Brazil, Rusija i Francuska kao države s najvećim komulativnim brojem zaraženih osoba u 2020. godini, te s druge strane izdvaja se SAD, Brazil, Indija, Meksiko i Italija kao države s najvećim brojem umrlih osoba u 2020. godini.

Iz [2] slijedi da je prosječna dob stanovništva uzorka 32.8 godina, a prosječna duljina trajanja školovanja 14.4 godine. Nadalje, 66.5% ljudi živi u urbanim područjima. Prosječna BDP vrijednost iznosi 26 850 dolara, a u 14% promatranih zemalja je na čelu žena. Nadalje, 30% ispitanika je bilo testirano.

Važno je istaknuti političke i kulturalne razlike. Standardna devijacija vjerske raznolikosti je 2.124, dok je standardna devijacija pri proučavanju nejednakosti među stanovništvom 7.395.

Empirijski rezultati

Pomoću Fama-MacBeth regresije vrše se testiranja i konstruira regresijski model za dane eksplanatorne varijable i zavisne varijable. Najprije se procjenjuju regresijski koeficijenti za pojedini tjedan, a potom prosječne vrijednosti na temelju tjednih procjena. Model u suštini uspoređuje varijacije različitih država gledajući njihove vrijednosti za određeni vremenski period (tjedan). On dopušta promjenu regresijskih koeficijenata pripadnih varijabli kroz promjenu promatranog tjedna. Procjenjujemo model koristeći Fama-MacBeth pristup:

$$ln(1 + Y_{i,t}) = \alpha + \beta X_{i,t} + \epsilon_{i,t}, \tag{3.1}$$

čiji koeficijenti reprezentiraju:

- $Y_{i,t}$ broj pozitivnih slučajeva/umrlih u i-toj državi u tjednu t
- X skup explanatornih varijabli iz 3.1
- $\epsilon_{i,t}$ rezidual za i-tu državu u tjednu t

Zavisne varijable	(1)	(2)	(3)	(4)
	Ln(1 + Broj pozitivnih)		Ln(1 + Broj umrlih)	
Stanovništvo	0.935*** (18.7)	0.963*** (18.6)	1.030*** (14.4)	1.052*** (14.8)
Gustoća stanovništva	0.314*** (6.3)	0.322*** (8.7)	0.053 (1.4)	0.065** (2.7)
Godine	0.049*** (11.1)	0.054*** (11.2)	0.087*** (21.4)	0.078*** (16.6)
Udio muškaraca	0.027*** (12.8)	0.023*** (5.8)	0.025*** (14.3)	0.032*** (8.9)
Urbanizacija	0.011*** (20.5)	0.012*** (17.4)	0.017*** (23.1)	0.016*** (17.9)
Temperatura	− 0.065*** (− 8.6)	− 0.062*** (− 8.9)	− 0.064*** (− 10.3)	− 0.072*** (− 11.5)
Obrazovanje	− 0.052*** (− 11.5)	− 0.077*** (− 15.6)	− 0.072*** (− 15.4)	− 0.089*** (− 12.0)
Religijska opredijeljenost	− 0.145*** (− 11.7)	− 0.133*** (− 12.8)	− 0.202*** (− 11.8)	− 0.194*** (− 13.4)
Demokracija	0.150*** (8.6)	0.108*** (6.1)	0.177*** (12.8)	0.174*** (13.9)
Korupcija	0.292*** (23.6)	0.266*** (20.8)	0.199*** (15.4)	0.153*** (14.3)
Medijska sloboda	− 0.001 (− 0.9)	− 0.004*** (− 3.2)	0.007*** (3.4)	0.004*** (3.5)
Ženski vođa	− 0.254*** (− 5.3)	− 0.105*** (− 3.7)	− 0.380*** (− 8.0)	− 0.399*** (− 12.0)
Povjerenje vlastima		− 0.006*** (− 6.2)		− 0.007*** (− 7.1)
Zakon	− 0.132*** (− 3.5)	− 0.001 (− 0.0)	− 0.218*** (− 8.8)	− 0.138*** (− 6.5)
BDP	0.016*** (16.7)	0.014*** (13.7)	0.017*** (12.0)	0.021*** (14.2)
Nejednakost	0.012*** (3.9)	0.018*** (9.1)	0.015*** (5.5)	0.017*** (10.8)
Turizam	0.009*** (10.9)	0.007*** (6.1)	0.014*** (8.9)	0.013*** (7.3)
Tehnologija	0.014*** (9.1)	0.012*** (7.0)	− 0.002** (− 2.3)	− 0.002* (− 1.7)
Zadovoljstvo	0.359*** (10.0)	0.328*** (8.6)	0.351*** (8.8)	0.362*** (8.8)
SARS	− 0.246*** (− 4.1)	− 0.211*** (− 5.1)	− 0.231*** (− 5.1)	− 0.332*** (− 11.4)
Bolnički kreveti	− 0.110*** (− 7.7)	− 0.117*** (− 7.7)	− 0.178*** (− 13.4)	− 0.188*** (− 13.5)
Broj testova		0.002*** (3.6)		− 0.002*** (− 6.7)
Konstanta	− 6.102*** (− 13.9)	− 5.959*** (− 8.3)	− 10.042*** (− 16.5)	− 10.472*** (− 16.1)
N	4851	4459	4,851	4459
R^2	0.790	0.790	0.750	0.721

Slika 3.2: Tablični prikaz Fama-MacBeth regresijskih rezultata na temelju broja pozitivnih i broja umrlih od COVID-19 pandemije. Podaci su preuzeti iz [2].

Na slici 3.2 je prikazana tablica s rezultatima Fama-MacBeth regresije. Zavisne varijable u prva dva stupca prikazuju rezultate regresijskog modela 3.1 gdje je $Y_{i,t}$ odzivna vrijednost

za broj oboljelih od COVID-19, dok preostala dva stupca, (3) i (4), prikazuju rezultate za broj smrtnih slučajeva uzrokovano COVID-19. U zagradama je izvješće t-statistike, dok ***, ** i * ukazuju značajnost na 1%, 5%, 10% razini, redom.

Razlika između (1) i (3), te (2) i (4) stupca jest ta da regresijski model u (2) i (4) stupcu sadrži varijable Povjerenje vlastima i Broj testiranih na COVID-19, dok one nisu sadržane za vrijednosti u (1) i (3) stupcu. Razlog tome je što podaci za navedene varijable nisu dostupni za sve promatrane države, pa su podaci tablično odvojeni i rezultati se nalaze u zasebnim stupcima.

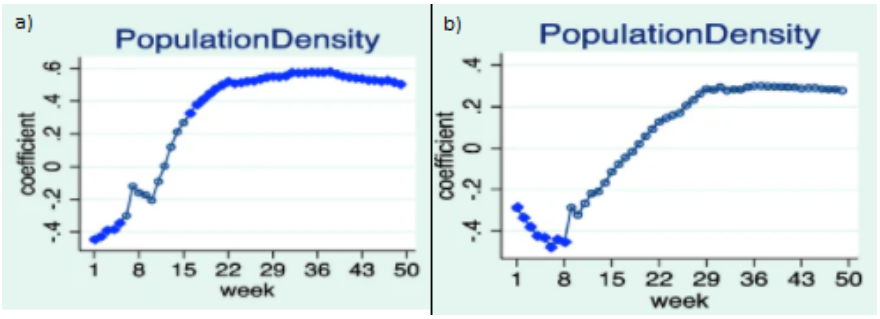
i. Rezultati nepogodnih varijabli

U ovom dijelu, na temelju rezultata iz 3.2 komentirat ćemo rezultate za proizvoljno odabrane 4 nepogodne varijable. Za analizu preostalih varijabli, vidi [2].

Gustoća stanovništva

Iz tablice na slici 3.2 možemo iščitati kako gustoća naseljenosti značajno utječe na porast broja zaraženih i umrlih što znači da se virus brže širi u gušće naseljenim zemljama. Pripadni koeficijent regresije za broj zaraženih virusom je 0.322, dok je za broj umrlih od istog 0.065. Standardna devijacija za taj faktor je 1.085 pa ukoliko se poveća gustoća naseljenosti stanovništva, tada se povećava i broj pozitivnih osoba za 34.94% od srednje vrijednosti broja pozitivnih osoba na COVID-19.

Na slici 3.3 je prikazana promjena regresijskog koeficijenta u odnosu na povećanje protek-

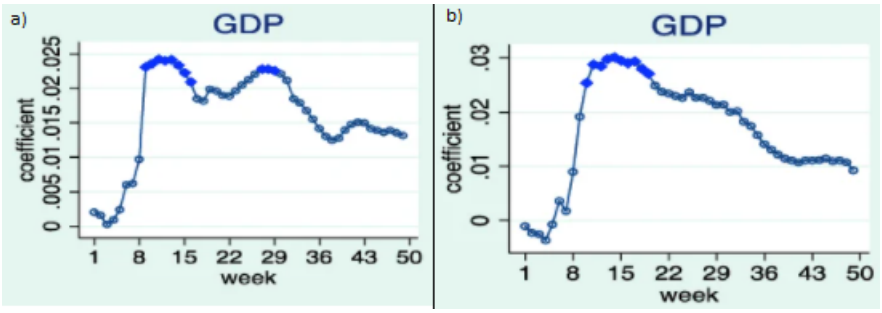


Slika 3.3: Grafički prikaz tjedne promjene regresijskog koeficijenta za varijablu Gustoća stanovništva. Lijevo:Y_1, desnoY_2. Izvor: [2]

lih tjedana. Na fotografiji, i za slučaj broja oboljelih i za slučaj broja umrlih od COVID-19 pandemije, vidimo kako je do 9. tjedna regresijski koeficijent bio negativnog predznaka, što znači da sve do 9. tjedna gustoća populacije nije utjecala na povećanje broja zaraženih ili broja umrlih. Nakon tog perioda, vrijednost koeficijenta raste što znači da veća gustoća naseljenosti stanovništva je u korelaciji s povećanjem broja zaraženih i broja umrlih.

BDP - bruto domaći proizvod

Iz [7] slijedi da BDP neke zemlje značajno utječe na promjenu broja oboljelih i umrlih tijekom COVID-19 pandemije. Kako razvijenije države imaju veći BDP od slabije razvijenih, logično je očekivati da će bogatije države imati više resursa za sprječavanje širenja pandemije. No, rezultati pokazuju suprotno. SAD, Švicarska, Ujedinjeno Kraljevstvo i Francuska imaju veliki BDP, no također imaju i visoku stopu oboljelih i umrlih od COVID-19. Kako razvijenije države imaju razvijenu internacionalnu trgovinu i ekonomiju slijedi da su interakcije među ljudima česte pa je i vjerojatnije da će doći do prijenosa, a potom i zaraze virusom. Nadalje, razvijenije države su liberalnije pa bi se njihove vlade mogle susresti s



Slika 3.4: Grafički prikaz tjedne promjene regresijskog koeficijenta za varijablu BDP. Lijevo:Y_1, desnoY_2. Izvor: [2]

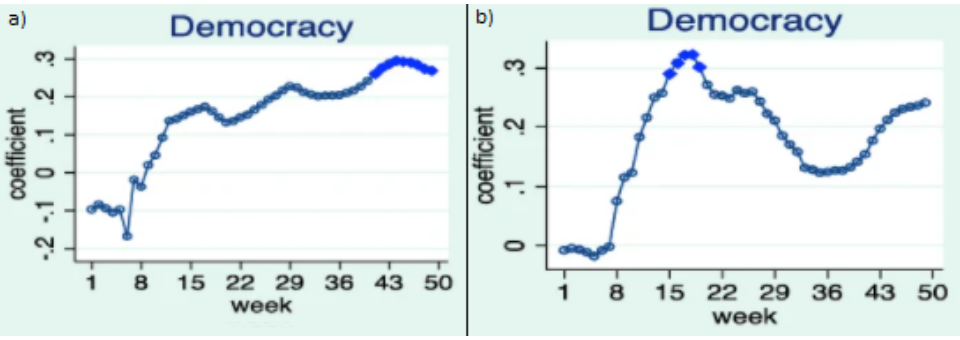
jačim otporom od građana prema mjerama koje propisuje vlada (kao što su socijalna distanca i samoizolacija) što može lako rezultirati napretkom pandemije.Sa slike 3.4 možemo vidjeti da je regresijski koeficijent vezan uz eksplanatornu varijablu BDP pozitivan za oba slučaja, utjecaj na broj oboljelih i na broj smrtnih slučajeva.

Udio muškaraca

Pozitivna vrijednost koeficijenta, definiran kao broj muškaraca na 100 žena, sugeriraj da su muškarci skloniji zarazi COVID-19 virusom. Što se tiče smrtnosti uzrokovane COVID-om, tu su žene sklonije. Moguće objašnjenje je da muškarci provode više vremena na poslu nego kod kuće pa su više u interakciji s drugim ljudima što povećava vjerojatnost zaraze virusom. Osim toga, Bwire [1], ističe da imunološke razlike temeljene na spolu također mogu uzrokovati veću stopu oboljenja i smrtnost kod muškaraca. Naime, oni imaju zastupljeniji enzim koji pretvara angiotenzin-2 (ACE 2) za razliku od žena. Taj enzim je receptor za SARS-CoV-2 virus. Dodatno, razlike u spolnom ponašanju i životnim stilovima također mogu igrati važnu ulogu. Primjerice, u usporedbi sa ženama, muškarci više puše i konzumiraju alkohol te su manje odgovorni pri poduzimanju preventivnih mjera (npr. često pranje ruku i nošenje maski za lice, naredbe o ostanku kod kuće), stoga su izloženiji

COVID-19 virusu.

Demokracija
Njezina mjera se određuje udjelom slobodnih i poštenih izbora te stupanjem vladinih intervencija u zemlji. Na samome početku, odnos između razine demokracije i uspjeha zemlje u borbi protiv koronavirusa pandemije nije odmah jasan. Teoretski, demokracija bi trebala potaknuti javno zdravlje jer u demokratskoj državi ljudi mogu glasati za ljude koji će biti u vladi i tako odabrati obećanja za poboljšanje zdravstvene slike države. Nasuprot tome, autoritativne vlade s lošom zdravstvenom slikom se ne suočavaju sa željama građana. Nadalje, demokratske zemlje obično usvajaju široku politiku i svoje resurse ravnomjernije raspoređuju, dok autoritarne vlade mogu favorizirati određene skupine. Koristeći podatke o svim epidemijama od 1960. do veljače 2020. godine, analiza The Economista pokazuje da, u prosjeku, demokratske zemlje imaju niže stope smrtnosti uzrokovane epidemijama nego što imaju one nedemokratske. Na slici 3.5 prikazan je grafički prikaz promjene regresijskog



Slika 3.5: Grafički prikaz tjedne promjene regresijskog koeficijenta za varijablu Demokracija. Lijevo: zavisna - ln(1+Broj pozitivnih) , desno: zavisna - ln(1+Broj umrlih) Izvor: [2]

koeficijenta za varijablu Demokracija. Uočimo da je koeficijent u oba slučaja povećavao svoju vrijednost, pa možemo zaključiti da je i broj pozitivnih i umrlih od COVID-19 se povećavao. U analizi se dalje ističe da „autoritarni režimi, iako sposobni koordinirati velike projekte, se ne snalaze u pitanjima koja zahtijevaju slobodan protok informacija i otvoreni dijalog između građana i vladara". Nadalje, Chang i sur. tvrde da osam od 10 najuspješnijih zemalja u borbi protiv COVID-19, uključujući Novi Zeland, Južnu Koreju, i skandinavske zemlje (npr. Finska, Norveška i Danska), su demokratske. Demokraciji u borbi protiv pandemije često otežava posao inherentna neučinkovitost i politička podijeljenost. Dodatno otkrivaju da su, za razliku od autokratskih zemalja, demokratske zemlje bile sporije u implementaciji mjere izolacije, koje se često smatraju suprotnim liberalnim pravima. Autoritarna vlada možda može učinkovito ublažiti i obuzdati pandemiju donošenjem

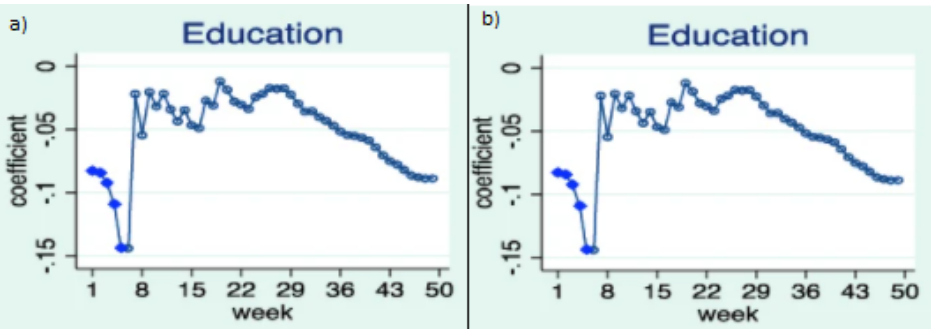
brzih odluka i mobilizacijom svih svojih privatnih i javnih ustanova, kao i uspostavljanje kontrole nad cjelokupnim stanovništvom. S druge strane,demokracija neizbježno rađa individualizam, za koju su [1] pokazali da pogoršava širenje COVID-19 pandemije smanjenjem učinkovitosti politike socijalnog distanciranja i ograničenja mobilnosti-uvođenjem samoizolacije.

ii. Rezultati ublažavajućih varijabli

U ovom dijelu prikazat ćemo i komentirati rezultate za 4 različite ublažavajuće varijabli i njihov utjecaj.

Obrazovanje

Uvjerili smo se da je obrazovna razina pojedine države, mjerena koristeći prosječan broj godina provedenih u školi, je značajan faktor koji pomaže pri smanjenju broja zaraženih i umrlih pandemijom. Iz tablice 3.2 možemo očitati da su koeficijenti u (1) i (3) stupcu negativnog predznaka. Pripada standardna devijacija za varijablu Obrazovanje iznosi 2.772 pa povećanje standardne devijacije u obrazovanju je povezano sa smanjenjem od 14.41% u broju oboljelih od njihove srednje vrijednosti. Na slici 3.6 vidimo kako je za zavisne

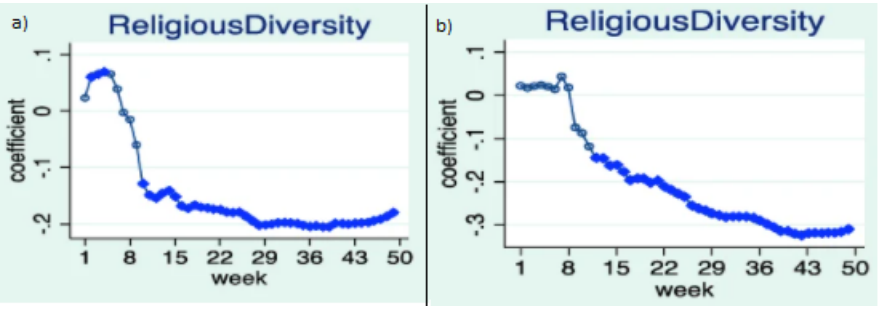


Slika 3.6: Grafički prikaz tjedne promjene regresijskog koeficijenta za varijablu Obrazovanje. Lijevo: zavisna - ln(1+Broj pozitivnih) , desno: zavisna - ln(1+Broj umrlih) Izvor: [2]

varijable Y_1 i Y_2 regresijski koeficijent negativan. Kako je vrijeme prolazilo, nakon 29. tjedna, apsolutna vrijednost regresijskog koeficijenta se povećavala pa možemo zaključiti da zemlje s razvijenijim obrazovanjem zaista doprinose suzbijanju COVID-19 pandemije. Ovaj zaključak podudara se s rezultatima Cutler i Lleras-Muney [4] koji su dokazali da je poveznica između obrazovanja i zdravlja pozitivna jer različite razina obrazovanja rezultiraju različitim razmišljanjem i donošenjem odluka. Konkretno, ljudi s višim stupnjem obrazovanja više vjeruju znanosti, više se informiraju i imaju visoko razvijeno kritičko razmišljanje. Uz to, dobra kognitivna sposobnost je povezana s višim obrazovnim stupnjem.

Religijska opredijeljenost

Vrijednost varijable Religijska opredijeljenost određena je veličinom religije u državi. Podaci pokazuju da je religijska opredijeljenost obrnuto proporcionalna s brojem zaraženih ili umrlih od COVID-19. Chang i sur. [2] istraživali su utjecaj jezične i nacionalne raznolikosti na kretanje broja zaraženih i umrlih, no nisu došli do značajnih rezultata. Razna istraživanja utvrdila su kako religija pomaže ljudima sa suočavanjem s tjeskobama i traumama, te da oni razviju otpornost pod utjecajem društvene podrške. Dodatno, ona može pomoći u discipliniranju pojedinaca i to posebice u zemljama gdje se stanovnici ne pridržavaju restriktivnih mjera. Na slici 3.7, za oba grafa, možemo uočiti da je u prvih 8



Slika 3.7: Grafički prikaz tjedne promjene regresijskog koeficijenta za varijablu Religijska opredijeljenost. Lijevo: zavisna - ln(1+Broj pozitivnih) , desno: zavisna - ln(1+Broj umrlih) Izvor: [2]

tjedana koeficijent bio pozitivan te je za to vrijeme postojala pozitivna korelacija sa zavisnim varijablama. No nakon toga, koeficijent poprima negativnu vrijednost, što znači da usporava širenje pandemije i doprinosi smanjenju broja zaraženih i umrlih.

Ženski vođa

Koeficijenti za zemlje u kojima je na vodećoj poziciji žena, u stupcima (1) i (3) u tablici 3.2 impliciraju da, u prosjeku, zemlje predvođene ženama imaju 25,4% manje pozitivnih slučajeva od zemalja koje vode muškarci. Analogno iz tablice možemo iščitati i razliku od 38% broja umrlih. Naime, iz [3] slijedi analiza podataka iz 34 razvijenih zemalja, uključujući i Kinu, pa su Coscieme i sur. zaključili da zemlje koje su vođene ženama bolje kontroliraju COVID-19 pandemiju. Njihovo proučavanje usredotočeno je na zemlje koje imaju visoko razvijeno gospodarstvo, visoki ljudski razvoj i vlada demokracija. Posebno su dodali Kinu jer je to bila prva zemlja koja je prijavila slučaj zaraze COVID-19. Nadalje, tvrde da nepredviđeni čimbenici, poput reagiranja u pravo vrijeme, odlučnost pri donošenju teških hitnih odluka te drugi čimbenici mogu objasniti njihove rezultate. Koristeći uzorak od 99 razvijenih zemalja i zemalja u razvoju, [2] potvrđuju kako žene kao vođe zemalja,

uspješnije u borbi s pandemijom. Naime, općenito su djelovale brže i odlučnije te su bile otvorenije za implementiranje inovativnih ideja.

Broj testiranih
Broj testiranih osoba na COVID-19 nije unaprijed određena karakteristika zemlje, ali može biti mehanički povezana s brojem prijavljenih zaraženih slučajeva. Stoga uključujemo broj testiranih slučajeva kao kontrolnu varijabli u stupcima (2) i (4) 3.2. Rezultati pokazuju da je testiranje na koronavirus u pozitivnoj korelaciji s brojem zaraženih. Jasno je da više provedenih testova otkriva veći broj pozitivnih slučajeva. Drugim riječima, testiranje na COVID-19 pomaže prebrojati oboljele ljudi. No, broj testiranih slučajeva negativno je povezan s brojem smrtnih slučajeva, što sugerira da točniji i pouzdaniji testovi značajno smanjuju smrtnost od COVID-19 ukoliko se zaraza otkrije ranije. Ukoliko uključimo varijablu broja testova COVID-19 u regresiju, varijabla ne utječe faktore koji su unaprijed određeni za pojedine države.

3.2 Diskusija i zaključak

Rezultati analiza 21 različitih faktora koji utječu na širenje COVID-19 pandemije su vidljivi. Dokazan je utjecaj nekoliko faktora na promjenu kretanja pandemije u pojedinoj državi i kako karakteristike pojedine države utječu na efikasnost u borbi s pandemijom. Svakako, valja istaknuti i da posljedične mjere pojedine države (nošenje maski, samoizolacija, socijalna distanca, itd.) također utječu na širenje pandemije, u ovom radu to se pokazalo za broj testiranih osoba.

Nastavno na istraživanje koje su proveli Chang i sur.[2] i na kojem se temelji ovo poglavlje, slijede rezultati promatranja različitih varijabli koje pospješuju širenje pandemije (pr. udio urbaniziranog stanovništva, veća gustoća naseljenosti). Nastavno na to, vladajuće grupacije zemalja trebale bi nastojati postići dobar omjer između urbaniziranog i ruralnog stanovništva pa provoditi decentralizaciju s ciljem učinkovitog korištenja urbanih i ruralnih područja zemlje. U kontekstu pandemije, takva preraspodjela svakako bi značajno utjecala kao preventivna mjera.

Nadalje, dotaknuli smo se razlike između autoritarne i demokratske vlade. Glavni problem demokratskog uređenja leži u osobama i organizacijama koje su na čelu vlasti jer bitna je značajka da tijekom pandemije stanovništvo polaže vjeru i nadu u vladajuće grupacije. One moraju biti u stanju uspostaviti ravnotežu između autoriteta i slobode prema pojedincima. Naime nužno je da vlada donosi hitne i efikasne odluke jer odgađanje donošenja odluka rezultira katastrofalnim posljedicama. Chang i sur. [2] nadodaju kako je prioritet države pri suočavanju s pandemijom upravo kvalitetno strukturirana vlada koja podupire medijsku slobodu jer samo tako može uspješno kontrolirati pandemiju na području svoje države i ostati u zdravom odnosu sa stanovništvom.

Činjenica je da je COVID-19 ostavio traga na razvijenijim zemljama u većoj mjeri nego kod nekih zemalja koje su u razvoju. Uspješnost kontrole pandemije ne određuje financijsko stanje države, već pravovremenost i učinkovitost intervencija pa se može reći da nije važan samo BDP, već i način preraspodjele iznosa državne blagajne. Vlade bi trebale nastojati uložiti više sredstava u razvoj medicine i javno zdravstvo što se može promatrati kao preventivna mjera za razvoj bolesti. [2] navode kako rezultati pokazuju veliku nejednakost financijskog dohotka kod stanovništva, stoga sugeriraju na dodatnu zaštitu siromašnih koji su osjetljiviji zbog osobnih financijskih poteškoća.

Obrazovanje omogućuje građanima donošenje kvalitetnih odluka koje su potkrijepljene raznolikim i kvalitetnim informacijama. Zemlje bi trebale nastavljati razvijati obrazovni sustav jer on utječe i na razvoj tehnologije koja se pokazala kao jedan od glavnih instrumenata pri borbi protiv pandemije. Što se tiče država na čijem su čelu žene, analiza je rezultirala činjenicom da one u pozitivnom tonu kontroliraju razvoj pandemije, no svakako je nerealno očekivati da su na čelu svih država žene. To znači da bi muškarci mogli štošta naučiti od kolegica pa bi bilo dobro da na njih i obrate pažnju.

Kao zaključak, sastavljeni faktori koji karakteriziraju razvijenost i osobnost pojedine zemlje mogu utjecati na brzinu proširenja ili suzbijanja pandemije. Stoga, svakako valja obratiti pozornost na uređenost države, glas i zadovoljstvo njezinih stanovnika jer kao što Alexander Graham Bell kaže: ”Prije svega, priprema je ključ uspjeha.”.

Bibliografija

[1] G.M. Bwire, *Coronavirus: Why Men are More Vulnerable to Covid-19 Than Women?*, SN Compr. Clin. Med., 847-876. str., 2020., dostupno na https://link.springer.com/article/10.1007/s42399-020-00341-w

[2] D. Chang, X. Chang , Y. He, K. K. Tan, *The determinants of COVID-19 morbidity and mortality across countries*, Scientific reports 12, 2022., 5888, dostupno na https://www.nature.com/articles/s41598-022-09783-9#article-info

[3] L. Coscieme i sur., *Women in power: Female leadership and public health outcomes during the COVID-19 pandemic*, 2020., dostupno na: https://www.medrxiv.org/content/10.1101/2020.07.13.20152397v2.full.pdf

[4] D. Cutler, A. Lleras-Muney, *Education and Health: Evaluating Theories and Evidence*, Working Paper 12352, National Bureau of Economic Research, 2006.

[5] V. Čuljak, *Vjerojatnost i statistika* - radni materijal, dostupno na http://www.grad.hr/vera/webnastava/vjerojatnostistatistika/vis-pdf.pdf

[6] K. Dietz, J. A. P. Heesterbeek, *Daniel Bernoulli's epidemiological model revisited*, Math. Biosci., 180(1–2), 1–21., 2002., dostupno na https://doi.org/10.1016/S0025-5564(02)00122-0

[7] Q. Feng, G.L. Wu, M. Yuan, S. Zhou, *What does Cross-Country Data Speak About COVID-19?* 2020., dostupno na https://personal.ntu.edu.sg/guiying.wu/FengWuYuanZhoufull20200815.pdf

[8] M. Huzak, *Vjerojatnost i matematička statistika* - predavanja, PMF-Matematički odjel, Zagreb, 2006.

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

[10] M. H. Kutner , C. J. Nachtsheim, J. Neter, W. Li , *Applied Linear Statistical Models*, McGraw-Hill Irwin, New York, 2005.

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

[12] N. Sandrić, Z. Vondraček, *Vjerojatnost - predavanja*, dostupno na https://www.pmf.unizg.hr/images/50023697/vjer_predavanja.pdf

[13] I. Šošic, V. Serdar, *Uvod u statistiku*, Školska knjiga, Zagreb, 1994.

Sažetak

U ovom radu, objašnjavamo teorijsku pozadinu linearne regresije te primjenjujemo statistički model regresije na stvarne podatke.

U prvom poglavlju iskazana je matematička teorija potrebna za razumijevanje sadržaja rada. Drugo poglavlje opisuje model jednostavne linearne regresije te model višestruke linearne regresije. Opisani su uvjeti koje proučavani podaci moraju zadovoljavati kako bi konstrukcija regresijskog modela bila smislena. Konačno, u trećem poglavlju, primijenjujemo linearnu regresiju na aktualnu tematiku današnjice, COVID-19 pandemiju.

Sama analiza uključivala je 22 različita faktora (nezavisne varijable) promatrana u 99 različitih zemalja, a promatrani podaci opisuju tjedni ukupan broj oboljelih i smrtnih slučajeva. Same varijable opisuju uređenost pojedine zemlje te one nisu posljedica pojavljivanja COVID-19 pandemije, već one definiraju demografsko-geografsku, političko-zakonsku, socijalno-ekonomsku i zdrastvenu sliku pojedine zemlje.

Summary

This work discusses the theoretical background of linear regression and application of statistical regression model with real data.

In the first chapter mathematical theory is stated, required for understanding the work itselfe. The second chapter describes the simple linear regression model as well as the multiple linear regression model. Various conditions are described which the examined data must satisfy in order for the construction to be meaningful. Finally, in the third chapter, linear regression is being applied on a trending topic of today, the COVID-19 pandemic.

The very analysis includes 22 different factors (independent variables) observed in 99 different states, whereas recorded data describes the total number of infected and death cases per week. The variables themselves describe the structure of the individual state and they are not the consequence of occurrences from COVID-19 pandemic, but they define demographic-geographical, political-legal, socio-economic and health aspects of the individual state.

Životopis

Rođena sam 3. svibnja 1996. godine u Čakovcu. Nakon završene Osnovne škole u Prelogu i Osnovne umjetničke škole Miroslav Magdalenić u Čakovcu, nastavljam srednjoškolsko obrazovanje u Čakovcu i Varaždinu. 2015. godine u Čakovcu završavam prirodoslovno-matematičku gimnaziju u Gimnaziji Josipa Slavenskog u Čakovcu, dok u Varaždinu polazim smjer klaviristice u Srednjoj glazbenoj školi u Varaždinu. Te iste godine upisujem Prirodoslovno-matematički fakultet u Zagrebu na matematičkom odsjeku: nastavnički smjer. Uz studiranje, organizacijske vještine razvijam radeći studentske poslove u sektoru menadžmenta. Nakon završetka preddiplomskog studija, 2019. upisujem diplomski studij Matematika: nastavnički smjer. Tijekom fakultetskog obrazovanje, dajem instrukcije te pripremam učenike za državnu maturu iz matematike te aktivno sudjelujem u organizaciji WISE-a na PMF-u 2019. i 2021. godine.