Jednostavna linearna regresija

Koso, Asiyah Noemi

Sveučilište Jurja Dobrile u Puli
Fakultet ekonomije i turizma
«Dr. Mijo Mirković»

ASIYAH NOEMI KOSO

JEDNOSTAVNA LINEARNA REGRESIJA

Završni rad

Pula, 2023.

Sveučilište Jurja Dobrile u Puli
Fakultet ekonomije i turizma
«Dr. Mijo Mirković»

ASIYAH NOEMI KOSO

JEDNOSTAVNA LINEARNA REGRESIJA

Završni rad

JMBAG: 0303087603, redoviti student
Studijski smjer: Management i poduzetništvo

Predmet: Statistika u ekonomiji
Znanstveno područje: Društvene znanosti
Znanstveno polje: Ekonomija
Znanstvena grana: Opća ekonomija

Mentor: doc. dr. sc. Katarina Kostelić

Pula, 2023.

IZJAVA O AKADEMSKOJ ČESTITOSTI

Ja, dolje potpisana, Asiyah Noemi Koso, kandidat za prvostupnicu managementa i poduzetništva ovime izjavljujem da je ovaj Završni rad rezultat isključivo mojeg vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na objavljenu literaturu kao što to pokazuju korištene bilješke i bibliografija. Izjavljujem da niti jedan dio Završnog rada nije napisan na nedozvoljen način, odnosno da je prepisan iz kojega necitiranog rada, te da ikoji dio rada krši bilo čija autorska prava. Izjavljujem, također, da nijedan dio rada nije iskorišten za koji drugi rad pri bilo kojoj drugoj visokoškolskoj, znanstvenoj ili radnoj ustanovi.

Student

_____

U Puli, _____, _____ godine

IZJAVA O KORIŠTENJU AUTORSKOG DIJELA

Ja, Asiyah Noemi Koso, dajem odobrenje Sveučilištu Jurja Dobrile u Puli, kao nositelju prava iskorištavanja, da moj završni rad pod nazivom „Jednostavna linearna regresija“ koristi na način da gore navedeno autorsko djelo, kao cjeloviti tekst trajno objavi u javnoj internetskoj bazi Sveučilišne knjižnice Sveučilišta Jurja Dobrile u Puli te kopira u javnu internetsku bazu završnih radova Nacionalne i sveučilišne knjižnice (stavljanje na raspolaganje javnosti), sve u skladu sa Zakonom o autorskom pravu i drugim srodnim pravima i dobrom akademskom praksom, a radi promicanja otvorenoga, slobodnoga pristupa znanstvenim informacijama.

Za korištenje autorskog djela na gore navedeni način ne potražujem naknadu.

Student

_____

U Puli, _____, _____ godine

Sadržaj

1. Uvod

Jednostavna linearna regresija je statistička metoda kojom se promatra utjecaj jedne nezavisne varijable (x) na zavisnu varijablu (y) te se njihov odnos analitički zapisuje regresijskim modelom. Ova metoda se često koristi u analizi podataka i predviđanju.

U jednostavnoj linearnoj regresiji, promatra se utjecaj samo jedne nezavisne varijable na zavisnu varijablu te se pretpostavlja da postoji linearna veza između nezavisne varijable x i zavisne varijable y. Cilj je odrediti liniju koja najbolje opisuje tu vezu. Ta linija se naziva regresijska linija ili linija najmanjih kvadrata.

Kako bi se postavio regresijski model važno je odabrati nezavisnu varijablu za koju znanstvenik ili istraživač smatra najbitnijom, odnosno onu kojom se opisuje utjecaj na zavisnu varijablu.

Postupak jednostavne linearne regresije uključuje pronalaženje koeficijenta nagiba i odsječka na y-osi regresijske linije. Koeficijent nagiba predstavlja promjenu zavisne varijable y za svaku jediničnu promjenu nezavisne varijable x, dok odsječak na y-osi označava vrijednost zavisne varijable y kada je nezavisna varijabla x jednaka nuli. Nakon određivanja regresijske linije, može se koristiti ta linija za predviđanje vrijednosti zavisne varijable y na temelju poznatih vrijednosti nezavisne varijable x.

Istraživačko pitanje na kojem se traži odgovor u ovom završnom radu glasi:

- Kako se može primijeniti jednostavna linearna regresija za analizu i predviđanje odnosa između dviju varijabli?

Navedeno pitanje povlači dodatna potpitanja na koja će se ponuditi odgovor u radu:

- Koje su pretpostavke jednostavne linearne regresije?
- Kako se evaluira model jednostavne linearne regresije?
- Koji su prednosti i nedostaci jednostavne linearne regresije?

Cilj ovog završnog rada je istražiti i analizirati koncept jednostavne linearne regresije te ga primijeniti na stvarnim podacima s ciljem razumijevanja i procjene linearne veze između dviju varijabli. Konkretno, u radu se daje pregled teorijskih osnova jednostavne linearne regresije te prikaz istog na primjeru u kojem će se koristiti stvarni podaci. Uz

navedeno, kako bi odgovorili na istraživačka pitanja, razmatrat će se upotreba linearne regresije u radovima u području društvenih znanosti objavljenih na hrvatskom jeziku u otvorenom pristupu.

Svrha ovog završnog rada je prvenstveno pružiti pregled o konceptu jednostavne linearne regresije i njezinoj primjeni u analizi podataka. Također, rad ima svrhu razotkriti kako se ova statistička metoda može koristiti za predviđanje i modeliranje promjena u zavisnoj varijabli na temelju promjena u nezavisnoj varijabli. Kroz praktične primjere i analize, pruža se uvid u praktičnu upotrebu jednostavne linearne regresije u stvarnim situacijama i potencijalno doprinosi boljem razumijevanju i interpretaciji odnosa između varijabli u različitim kontekstima.

Struktura rada sačinjena je od pet poglavlja. U uvodu se predstavlja tema istraživanja i ističe njezina važnost, postavljaju se istraživačka pitanja i ciljevi istraživanja. U drugom poglavlju se opisuju teorijske osnove modela jednostavne linearne regresije i kako se primjenjuju u analizi podataka te opisuje kako se izračunava model jednostavne linearne regresije. Treće poglavlje sadrži konkretan primjer kreiranja modela jednostavne linearne regresije koristeći stvarne podatke. U četvrtom poglavlju analizira se primjena jednostavne linearne regresije u društvenim istraživanjima i uspoređuje se s višestrukom regresijom. U zaključku se sumiraju glavni odgovori na postavljena istraživačka pitanja i praktične implikacije.

U ovom završnom radu korištene su sljedeće znanstvene metode: metoda analize i sinteze, apstrakcije i deskripcije, povijesna metoda, te analiza prednosti i nedostataka.

2. Teorijska osnova jednostavne linearne regresije

Jednostavna linearna regresija je samo jedan oblik regresijske analize, a postoje i druge tehnike, poput višestruke linearne regresije, koje uključuju više regresora za predviđanje regresanda.

Jednostavna linearna regresija ima veliku važnost u području statistike i istraživanja:

- Modeliranje linearnih odnosa: Jednostavna linearna regresija omogućuje modeliranje linearnog odnosa između nezavisne varijable (regresor) i zavisne varijable (regresand). To pomaže u razumijevanju kako se jedna varijabla mijenja s promjenom druge varijable.
- Predviđanje vrijednosti: Regresijska analiza omogućuje predviđanje vrijednosti regresanda na temelju poznatih vrijednosti zavisne varijable, uz prisustvo određenih uvjeta odnosno predviđati je moguće jedino i isključivo na osnovu intervala nezavisne varijable koji se koristio pri izračunu modela.
    - To može biti korisno u mnogim područjima, poput ekonomije, financija, marketinga i zdravstva, gdje je predviđanje budućih događaja ili vrijednosti od velike važnosti.
- Identifikacija povezanosti: Jednostavna linearna regresija omogućuje identifikaciju i kvantifikaciju povezanosti između varijabli. Koeficijent determinacije (R^2) mjeri koliko dobro regresijski model objašnjava varijabilnost regresanda.
- Testiranje hipoteza: Jednostavna linearna regresija omogućuje testiranje statističkih hipoteza o povezanosti između regresora i regresanda. Ovo je važno u istraživačkom radu kako bi se odbacile ili ne odbacile postavljene hipoteze i donijele zaključke temeljene na empirijskim podacima.

Jednostavna linearna regresija je važan alat za modeliranje, predviđanje i razumijevanje povezanosti između varijabli.

Kako bi praćenje i razumijevanje rada bilo jednostavno i razumljivo, potrebno je prvo objasniti osnovne pojmove na temelju koje se ova analiza oslanja.

Prosjek, u kontekstu statistike, predstavlja srednju vrijednost niza brojeva ili podataka. To je osnovna mjera centralne tendencije koja se koristi kako bi se dobila ideja o tome kako se vrijednosti podataka raspoređuju oko sredine (Belullo, 2011.). Matematički, aritmetički prosjek se izražava formulom (Belullo, 2011):

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Varijacija podataka u statistici odnosi se na mjeru raznolikosti ili raspršenosti vrijednosti unutar skupa podataka. To je važna statistička mjera koja pomaže u razumijevanju koliko su podaci raznoliki ili kako se razlikuju jedni od drugih. Varijacija je ključna u mnogim statističkim analizama jer može pružiti informacije o rasponu, disperziji ili konzistenciji podataka. Formula za izračun varijacije podataka koristi se kako bi se mjerila raspršenost ili varijacija podataka u skupu.

U kontekstu varijacije podataka, najmjerodavnija je formula varijance. Varijanca je statistička mjera koja se koristi za kvantificiranje raspršenosti ili varijabilnosti vrijednosti u skupu podataka (Horvat, 2014.). To je osnovna mjera disperzije podataka koja daje informaciju o tome kako se pojedinačne vrijednosti razlikuju od srednje vrijednosti (aritmetičkog prosjeka) skupa podataka. Varijanca se često koristi u statističkim analizama kako bi se razumjela stabilnost i varijabilnost podataka. Formula za izračun varijance glasi (Horvat, 2014.):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Gdje:

- σ^2 predstavlja varijancu
- n je broj opažanja u skupu podataka
- x_i su pojedinačne vrijednosti u skupu podataka
- μ je aritmetički prosjek (srednja vrijednost) svih vrijednosti u skupu podataka

Kovarijanca je statistička mjera koja se koristi za kvantificiranje stupnja međusobne promjene između dviju varijabli u skupu podataka. Ova mjera pruža informacije o tome kako se dvije varijable povezuju i da li se njihove vrijednosti zajedno mijenjaju. Kovarijanca može biti pozitivna, negativna ili blizu nule, što ukazuje na različite oblike odnosa između varijabli. Formula za izračun kovarijanca između dvije varijable x i y u skupu podataka od n opažanja je (Belullo, 2014.):

$$Kovarijanca\ (cov(x,y)) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Normalna distribucija, također poznata kao Gaussova distribucija ili zvonasta krivulja, je statistički model raspodjele podataka koji se često pojavljuje u prirodi i mnogim aspektima ljudskog života (Horvat, 2014.). Normalna distribucija se često koristi u statistici zbog svoje predvidljive prirode i brojnih matematičkih svojstava. Mnogi statistički testovi, poput testova hipoteza i intervali pouzdanosti, pretpostavljaju normalnu distribuciju podataka. Međutim, važno je napomenuti da stvarni podaci često nisu savršeno normalno distribuirani, pa se u praksi koriste različite metode za procjenu i modeliranje raspodjele podataka.

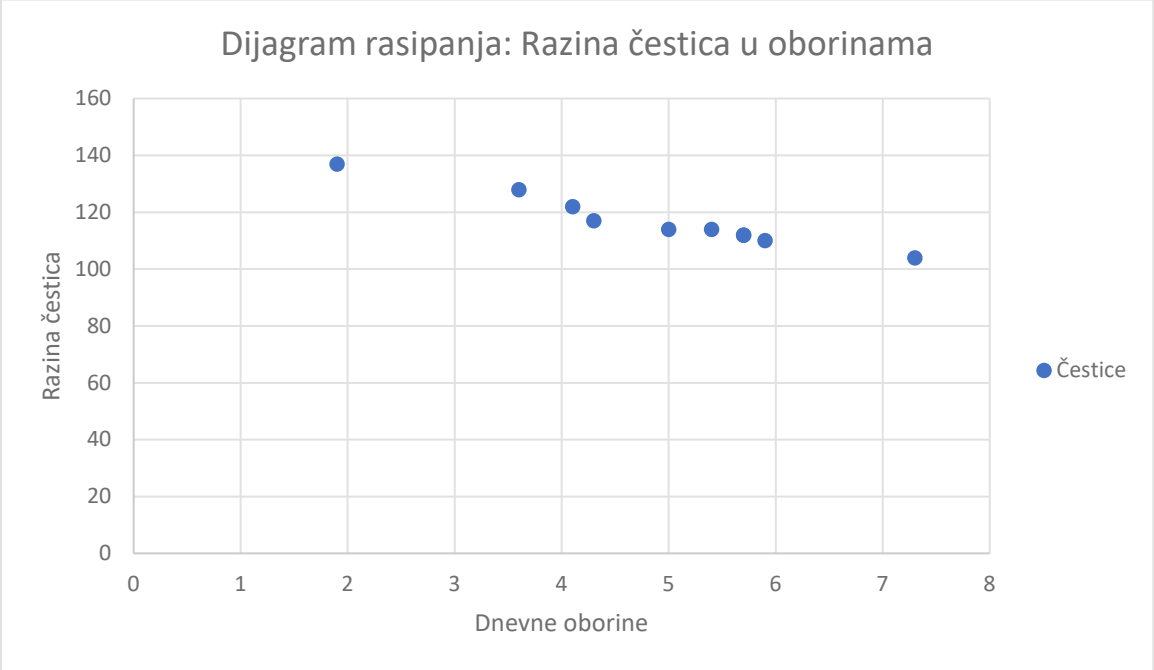
Dijagram rasipanja, također poznat kao scatter plot ili scattergram, je grafički prikaz podataka koji se koristi za prikazivanje odnosa između dvije ili više varijabli (Belullo, 2011.). Ovaj tip grafa omogućuje vizualno prikazivanje kako se vrijednosti jedne varijable mijenjaju u odnosu na vrijednosti druge varijable. Dijagrami rasipanja su korisni za identificiranje uzoraka, trendova i korelacija između varijabli.

Jednostavna linearna regresija zahtijeva dva ključna skupa podataka: zavisnu varijablu (y) i nezavisnu varijablu (x). Odnos između ovih varijabli modelira se kako bi se bolje razumjelo ili predvidjelo ponašanje zavisne varijable na temelju nezavisne varijable. Na primjer, možemo istraživati odnos između godine proizvodnje automobila (x) i njegove cijene (y), koristeći teoretsko uporište koje sugerira da stariji automobili obično imaju niže cijene. Pomoću dovoljno podataka, o cijenama i godinama proizvodnje različitih automobila, može se izgraditi regresijski model koji pomaže predvidjeti cijenu automobila na temelju godine proizvodnje. Iako se ovdje spominju godine, u pitanju su godine starosti automobila u trenutku provođenja istraživanja. Takvi podaci se zovu presječni podaci.

Slučajno uzorkovanje igra ključnu ulogu u važnosti rezultata u kontekstu danog primjera jednostavne linearne regresije. Kada se prikupljaju podaci o godinama proizvodnje i cijenama automobila, bitno je da uzorak bude slučajan te da broj opažanja bude veći od 250 (Schönbrodt i Perugini, 2013) kako bi rezultati bili reprezentativni za širu populaciju automobila. Slučajno uzorkovanje osigurava da svaki automobil u populaciji ima jednaku vjerojatnost da bude uključen u uzorak, čime se smanjuje pristranost. Ako se ne koristi slučajno uzorkovanje i, na primjer, odabrani su samo najstariji automobili ili samo oni određenih marki, rezultati bi bili pristrani prema tim kriterijima i ne bi odražavali stvarni odnos između godina proizvodnje i cijena automobila u populaciji. Slučajno uzorkovanje povećava pouzdanost zaključaka i, ukoliko su i ostale relevantne pretpostavke modela ispoštovane, omogućava generalizaciju rezultata na cijelu populaciju. To je temeljna praksa u istraživačkim studijama kako bi se osigurala valjanost i relevantnost dobivenih informacija.

Slično tome, u drugim primjerima, kao što su brzina automobila i potrošnja goriva ili temperatura i prodaja sladoleda, jednostavna linearna regresija omogućuje istraživanje i modeliranje odnosa između varijabli da bi se bolje razumjelo ili predvidjelo ponašanje zavisnih varijabli na temelju nezavisnih varijabli, koristeći relevantna teorijska načela kao polazište.

Graf 1. Primjer dijagrama rasipanja



Izvor: Microsoft (2023): Prikazivanje podataka na raspršenom ili linijskom grafikonu, dostupno na https://support.microsoft.com/hr-hr/topic/prikazivanje-podataka-na-raspr%C5%A1enom-ili-linijskom-grafikonu-4570a80f-599a-4d6b-a155-104a9018b86e, pristupljeno 20.09.2023.

Korelacija je statistička mjera koja se koristi za kvantificiranje stupnja međusobnog odnosa između dviju ili više varijabli. Ova mjera pomaže u određivanju kako se promjene u jednoj varijabli povezuju s promjenama u drugoj varijabli. Korelacija se često koristi kako bi se utvrdila jačina i smjer veze između varijabli u skupu podataka. Ključne značajke korelacije uključuju (Belullo, 2011.):

- Vrijednosti korelacije: Korelacija ima vrijednosti između -1 i 1. Korelacija blizu 1 ukazuje na snažnu pozitivnu korelaciju, što znači da kada jedna varijabla raste, druga također često raste. Korelacija blizu -1 ukazuje na snažnu negativnu korelaciju, što znači da kada jedna varijabla raste, druga često opada. Korelacija blizu 0 ukazuje na slabo ili nikakvo linearno povezivanje između varijabli.
- Smjer korelacije: Pozitivna korelacija znači da varijable rastu zajedno, dok negativna korelacija znači da varijable rastu suprotno jedna drugoj. Korelacija blizu 0 ukazuje na slabu ili nikakvu vezu.
- Korelacija ne implicira uzročnost: Važno je napomenuti da korelacija ne implicira uzročnost. Samo zato što dvije varijable pokazuju određenu razinu korelacije ne znači nužno da jedna uzrokuje drugu. U mnogim slučajevima, postoji mogućnost da postoji treća, skrivena varijabla koja utječe na obje varijable.
- Korelacijski koeficijent: Za kvantificiranje korelacije koristi se korelacijski koeficijent, najčešće Pearsonov korelacijski koeficijent. Drugi korelacijski koeficijenti uključuju Spearmanov i Kendallov koeficijent. Pearsonov korelacijski koeficijent mjeri linearnu korelaciju između dvije varijable, dok Spearmanov i Kendallov koeficijent mjere rangiranu ili nemonotonsku korelaciju.

Korelacija se često koristi u mnogim disciplinama, uključujući statistiku, ekonomiju, znanost, inženjering, medicinu i druge, kako bi se istraživali odnosi između varijabli i donosili zaključci o njihovoj povezanosti.

2.1. Jednadžba modela jednostavne linearne regresije

Točna, regresijska jednadžba jednostavne linearne regresije može se zapisati kao:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

U ovoj jednadžbi:

- y predstavlja zavisnu varijablu (regresand) koja se pokušava predvidjeti ili objasniti.
- x predstavlja nezavisnu varijablu (regresor) koja se koristi za predviđanje ili objašnjavanje vrijednosti y.
- β_0 jest konstantni član te predstavlja konstantnu vrijednost i označava odsječak na osi y, odnosno koeficijent smjera. Konstantni član označava prosječnu vrijednost zavisne varijable y kada je nezavisna varijabla x jednaka nuli ($x = 0$). Također, predstavlja i koeficijent smjera te je to početna vrijednost regresijske linije.
- β_1 predstavlja koeficijent nagiba te objašnjava promjenu zavisne varijable y u situaciji kada se poveća nezavisna varijabla x za jednu jedinicu.
- ε predstavlja slučajnu pogrešku ili rezidual, što su nepredvidive i nesistematske komponente koje ostaju nakon što se objasne varijabilnosti nezavisne varijable x.

Razlozi javljanja slučajne pogreške ε (Belullo, 2011):

- Neodređenost teorije
- Nedostupnost podataka
- Javljanje manje važnih varijabli
- Slučajnosti koje su svojstvena ljudskom ponašanju
- Loše zamjenske varijable
- Krive funkcionalne forme

Regresijska jednadžba omogućuje predviđanje vrijednosti y na temelju poznatih vrijednosti x, pri čemu β_0 predstavlja početnu vrijednost regresijske linije, a β_1 mjeri koliko se zavisna varijabla y mijenja za jediničnu vrijednost porasta nezavisne varijable x.

Primjer: Ako se primijeni jednostavnu linearnu regresiju na podacima BDP - a (y) i izvoza (x), $\beta 0$ će predstavljati prosječan BDP $(X = 0)$, dok će $\beta 1$ mjeriti koliko se prosječni BDP mijenja za svaku dodatnu jedinicu izvoza. Slučajna greška (ε) odražava sve ostale faktore koji mogu utjecati na BDP osim izvoza i može biti posljedica različitih unutarnjih ili vanjskih čimbenika u zemlji

Važno je napomenuti da vrijednosti $\beta_0 \, i \, \beta_1$ se procjenjuju na temelju dostupnih podataka i metodom najmanjih kvadrata kako bi se pronašla najbolja regresijska linija koja odgovara tim podacima.

2.2. Koncept linearnog odnosa i pretpostavke jednostavne linearne regresije

Koncept linearnog odnosa u jednostavnoj linearnoj regresiji se odnosi na pretpostavku da postoji linearna veza između nezavisne varijable (regresora) i zavisne varijable (regresanda) y. To znači da se promjene u vrijednostima regresora x očekuju da uzrokuju proporcionalne promjene u vrijednostima regresanda y.

Pretpostavke jednostavne linearne regresije su sljedeće:

1. Linearnost: Pretpostavka je da postoji linearni odnos između x i y, što znači da se očekuje da regresijska linija bude ravna.

2. Nezavisnost: Podaci koje se koriste za regresijsku analizu trebaju biti nezavisni. To znači da nema sustavnih veza ili uzoraka među podacima.

3. Sredina slučajne pogreške ε jest jednaka nuli. Simbolički bi to izgledalo na slijedeći način $E(\varepsilon_i|x_i) = 0$

4. Homoskedastičnost: Ako varijanca slučajne greške je ista za sva opažanja. Suprotno ovoj tezi bi bilo da ako varijanca slučajne greške nije ista za sva opažanja, nego ovisi o nekoj od nezavisnih varijabli tada se govori o heteroskedastičnosti, odnosno varijanca slučajnog odstupanja ε_i nije konstantna. Drugim riječima, varijabilnost reziduala trebala bi biti konstantna za sve vrijednosti x.

5. Normalnost reziduala: Pretpostavlja se da su reziduali distribuirani normalno, odnosno da su simetrično raspoređeni oko nule.

6. Broj opažanja mora biti veći od broja parametara koji se procjenjuju iz razloga što na temelju samo jednog opažanja (jedna točka u dvodimenzionalnom prostoru) ne može se procijeniti linija na tom prostoru, odnosno potrebne su barem dvije točke kako bi se mogli odrediti parametri pravca $\beta_0\ i\ \beta_1$. Također, da bi se postigla stabilna razina predikcije za generalizaciju zaključka zahtijeva se uzorak veličine n=250 (Schönbrodt i Perugini, 2013).

7. Pravilno specificiran regresijski model: pravilno specificirana funkcionalna forma, pravilno specificirana nezavisna varijabla i pravilno specificirana pretpostavka o vjerojatnosti $y, x\ i\ \varepsilon$.

Važno je provjeriti ove pretpostavke prije zaključivanja temeljem modela jednostavne linearne regresije kako bi se osiguralo da model daje pouzdane rezultate. Kršenje ovih pretpostavki može utjecati na točnost i interpretaciju rezultata regresije. U nekim slučajevima, može biti potrebno primijeniti transformacije podataka ili koristiti alternativne modele ako pretpostavke nisu zadovoljavajuće.

2.3. Metoda najmanjih kvadrata (OLS)

Metoda najmanjih kvadrata (OLS - Ordinary Least Squares) je statistička tehnika koja se koristi za procjenu parametara regresijskog modela kako bi se pronašla najbolja regresijska linija koja odgovara dostupnim podacima. Glavna ideja metode najmanjih kvadrata je minimizacija sume kvadrata reziduala (SSR) (razlika između stvarnih vrijednosti zavisne varijable i predviđenih vrijednosti) kako bi se pronašli najbolji parametri regresijskog modela. OLS traži parametre modela (koeficijente) koji čine ovu sumu kvadrata razlika što bližom nuli. Što su kvadrati razlika između stvarnih vrijednosti zavisne varijable (y) i predviđenih vrijednosti $(\bar{y})$ bliže nuli to znači da model pokušava minimizirati greške predikcije.

Prednosti OLS metode su:

- Jednostavnost primjene i interpretacije.
- Robusnost u brojnim situacijama.
- Procjenjuje optimalne parametre koji minimiziraju sumu kvadrata reziduala.

OLS metoda koristi sljedeće korake:

1. Formuliranje regresijskog modela:

Prvo se vrši priprema podataka te se definira regresijske jednadžba oblika

$$y = \beta_0 + \beta_1 x + \varepsilon$$

2. Procjena parametara:

Cilj je procijeniti vrijednosti parametara $\beta_0 \, i \, \beta_1$ koji minimiziraju SSR. To se postiže odabirom takvih $\beta_0 \, i \, \beta_1$ koji daju najmanju sumu kvadrata reziduala.

OLS metoda koristi matematičke metode za pronalaženje optimalnih procjena parametara. Matematički, procjene parametara $\beta_0 \, i \, \beta_1$ se dobivaju rješavanjem sustava jednadžbi ili korištenjem izraza koji minimiziraju SSR.

3. Evaluacija modela:

Procijenjeni parametri $\beta_0 \, i \, \beta_1$ se koriste za konstrukciju regresijske linije.

Model se evaluira koristeći različite statističke metrike kao što su koeficijent determinacije (R^2), standardna pogreška regresije, F-test i t-test za parametre.

Ove metrike pomažu u ocjeni koliko dobro regresijski model objašnjava varijabilnost zavisne varijable y.

4. Interpretacija rezultata:

Procijenjeni parametri $\beta_0 \, i \, \beta_1$ se tumače kako bi se razumjela povezanost između nezavisne varijable x i zavisne varijable y.

Koeficijent nagiba β_1 ukazuje na promjenu u vrijednosti y za svaku jedinicu promjene x, dok odsječak na y-osi β_0 predstavlja vrijednost y kada je x jednako nuli.

Važno je napomenuti da OLS metoda ima pretpostavke koje je potrebno provjeriti prije primjene, kao što su linearnost, normalnost reziduala.

2.4. Izračunavanje koeficijenata nagiba (β_1) i odsječka na y-osi (β_0)

Da bi se izračunali koeficijent nagiba (β_1) i odsječak na y-osi (β_0) u jednostavnoj linearnoj regresiji, potrebno je koristiti metodu najmanjih kvadrata (OLS) na dostupnim podacima. Ovdje je opisan postupak za izračunavanje tih koeficijenata:

1. Priprema podataka:

Prisutan je skup podataka koji sadrži vrijednosti nezavisne varijable x i zavisne varijable y.

2. Izračunavanje srednjih vrijednosti – aritmetička sredina:

Izračunavanje srednje vrijednosti za x i y, označene kao $\hat{x}\ i\ \hat{y}$.

3. Izračunavanje razlika:

Za svaki par vrijednosti (x, y), izračunava se razlika:

$$\Delta x = x - \bar{x}$$

$$\Delta y = y - \bar{y}$$

4. Izračunavanje kvadrata razlika:

Za svaku razliku Δx, izračunava se kvadrat: $(\Delta x)^2$

Za svaku razliku Δy, izračunava se kvadrat: $(\Delta y)^2$

5. Izračunavanje umnoška razlika:

Za svaki par razlika (Δx, Δy), izračunava se umnožak: $\Delta x \cdot \Delta y$

6. Izračunavanje suma:

Zbrajaju se sve vrijednosti $(\Delta x)^2$, $(\Delta y)^2$ i $\Delta x \cdot \Delta y$. Označavaju se kao $\sum(\Delta x)^2$, $\sum(\Delta y)^2$ i $\sum(\Delta x \cdot \Delta y)$

7. Izračunavanje koeficijenta nagiba (β_1):

Koristeći izračunate vrijednosti, koeficijent nagiba može se izračunati pomoću formule:

$$\beta_1 = \frac{SS_{yy}}{SS_{xx}} = \frac{\sum(\Delta x \cdot \Delta y)}{\sum(\Delta x)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

8. Izračunavanje odsječka na y-osi (β_0):

Odsječak na y-osi se može izračunati pomoću formule:

$$\beta_o = \bar{y} - \beta_1 \cdot \bar{x}$$

Nakon izračunavanja β_1 i β_0, dobiva se regresijska jednadžba $y = \beta_0 + \beta_1 x$ koja omogućuje predviđanje vrijednosti y na temelju poznatih vrijednosti x.

Da bi se situacija olakšala, mnogi statistički softveri i programski jezici automatski izračunavaju ove koeficijente prilikom izvođenja regresijske analize.

2.5. Tumačenje koeficijenata nagiba (β_1) i odsječka na y-osi (β_0)

Koeficijent nagiba (β_1) i odsječak na y-osi (β_0) su važni parametri u jednostavnoj linearnoj regresiji. Tumačenje koeficijenata:

1. Koeficijent nagiba (β_1):

β_1 mjeri koliko se prosječna vrijednost zavisne varijable (y) mijenja za svaku jedinicu promjene nezavisne varijable (x).

Pozitivna vrijednost β_1 ukazuje na pozitivnu povezanost između x i y, što znači da se očekuje da će y rasti s povećanjem x.

Negativna vrijednost β_1 ukazuje na negativnu povezanost između x i y, što znači da se očekuje da će y opadati s povećanjem x

Primjer tumačenja β1 u kontekstu BDP-a i stope nezaposlenosti: Ako je β1 jednako 0.5, to znači da se prosječni BDP povećava za 0.5 za svaki postotni poen stope nezaposlenosti. Dakle, veći BDP može biti povezan s većom stopom nezaposlenosti.

2. Odsječak na y-osi (β_0):

β_0 predstavlja vrijednost zavisne varijable (y) kada je nezavisna varijabla (x) jednaka nuli.

Odsječak na y-osi odražava početnu vrijednost regresijske linije, bez obzira na vrijednost x.

Interpretacija β_0 ovisi o kontekstu i značenju varijabli koje se proučavaju.

Primjer tumačenja β_0 u kontekstu BDP – a i uvoza: Ako je β0 jednako 55,7 milijardi $, to znači da je prosječan BDP zemlje $(x = 0)$ 55,7 milijardi $. To je početna vrijednost regresijske linije, odnosno BDP kada na njega uvoz nema utjecaj.

Važno je napomenuti da tumačenje koeficijenata nagiba i odsječka na y-osi treba biti usklađeno s kontekstom istraživanja i interpretacijom varijabli koje se proučavaju. Također, treba uzeti u obzir i statističku značajnost koeficijenata kako bi se potvrdila njihova relevantnost i pouzdanost.

2.6. Evaluacija modela

2.6.1. Pretpostavke modela

Evaluacija pretpostavki modela jednostavne linearne regresije ključna je faza analize podataka i statističkog modeliranja. Ova evaluacija pomaže osigurati valjanost i pouzdanost regresijskog modela koji se koristi za predviđanje i objašnjavanje odnosa između dvije varijable, često označenih kao x (nezavisna varijabla) i y (zavisna varijabla) (Neuman, 2014.). Pretpostavke modela jednostavne linearne regresije obuhvaćaju niz uvjeta i zahtjeva koji bi trebali biti zadovoljeni kako bi se model smatrao valjanim. Ključne pretpostavke i postupci za njihovu evaluaciju su (Neuman, 2014.):

- Linearnost odnosa:

1. Pretpostavka: Odnos između nezavisne (x) i ovisne (y) varijable treba biti linearan.
2. Evaluacija: Provjerava se grafikonom. Scatter plot (dijagram rasipanja) x i y varijabli trebao bi ukazivati na linearnu vezu. Ako postoje znakovi krivulje ili ne-linearnog odnosa, model jednostavne linearne regresije možda nije prikladan.

- Normalna distribucija reziduala:

1. Pretpostavka: Razlika između stvarnih i predviđenih vrijednosti y (ostatci) trebaju biti normalno distribuirani.
2. Evaluacija: Može se provjeriti pomoću grafikona kvantil-kvantil (QQ plot) i testova normalnosti poput Kolmogorov-Smirnov testa ili Shapiro-Wilk testa. Ako

su razlika između stvarnih i predviđenih vrijednosti y (ostataka) značajno odstupaju od normalne distribucije, to može utjecati na predikciju modela.

- Homoskedastičnost:

1. Pretpostavka: Varijabilnost razlike između stvarnih i predviđenih vrijednosti y trebala bi ostati konstantna kroz različite razine x varijable. Ovo znači da bi varijabilnost reziduala trebala biti uniformna.
2. Evaluacija: Scatter plot ostataka u odnosu na predviđene vrijednosti y može pokazati je li varijabilnost konstantna. Testovi poput Breusch-Pagan testa ili White testa mogu također pomoći u ocjeni homoskedastičnosti. U slučaju narušene homoskedastičnosti, zaključuje se da odabrani regresijski model nije primjeren za analizu promatranih pojava.

- Neovisnost reziduala između stvarnih i predviđenih vrijednosti y:

1. Pretpostavka: Razlike između stvarnih i predviđenih vrijednosti y bi trebale biti neovisna jedna o drugima – reziduali pojedinih opažanja nemaju efekata na reziduale drugih opažanja.
2. Evaluacija: Ako izostaje neovisnost reziduala može upućivati na to da podaci nisu presječni, nego su prikupljani u različitim trenucima. Također, to može upućivati na to da su podaci prije analize bili sortirani po veličini, jer takav postupak može rezultirati uočavanjem obrazaca u rezidualima.

Evaluacija ovih pretpostavki važan je korak u analizi podataka jer pomaže osigurati valjanost i pouzdanost regresijskog modela. Ako su pretpostavke prekršene, to može dovesti do netočnih zaključaka i interpretacija. U takvim slučajevima, mogu se razmotriti alternativni modeli ili korekcije kako bi se ispravili problemi i poboljšala valjanost modela regresije.

2.6.2. Koeficijent determinacije (R^2)

Koeficijent determinacije, često se označava kao R^2, se koristi za procjenu koliko dobro odabrani model odgovara stvarnim podacima. R^2 poprima vrijednosti između 0 i 1 te što je R^2 bliže 1, to model bolje objašnjava varijacije u podacima, a što je bliže 0, to model lošije objašnjava varijacije u zavisnoj varijabli. Regresijski model je reprezentativniji što je koeficijent determinacije bliži jedinici. Koeficijent determinacije

predstavlja statističku mjeru koja prikazuje koliko varijance uzorka y jest objašnjeno modelom te se matematički koeficijent determinacije može definirati kao:

$$R^2 = \frac{SSR}{SS_{yy}}$$

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = \frac{\sum(\hat{y_i} - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

R^2 se računa kao proporcija objašnjene varijacije (varijacije koja se može pripisati modelu) i ukupne varijacije u podacima, odnosno koeficijent determinacije regresijskog modela prikazuje dio ukupne varijance y koja je obuhvaćena regresijskim modelom.

SSR (engl. sum of squared regression) – suma kvadrata razlike između $\hat{y}\ i\ \bar{y}$, odnosno protumačena odstupanja koja predstavljaju one varijacije koje su objašnjene regresijskim modelom (nezavisnom varijablom).

SS_{yy} – predstavlja sumu kvadrata razlike između $y - \bar{y}$, odnosno ukupno odstupanje mjeri sve varijacije u zavisnoj varijabli.

U nastavku jest prikazano za određenu vrijednost x (Belullo, 2011):

- Ukupno odstupanje y od njegovog prosjeka: $(y - \bar{y})$
- Regresijskim pravcem objašnjeno odstupanje: $y(\hat{y} - \bar{y}\ )$
- Regresijskim pravcem neobjašnjeno odstupanje $y(y - \hat{y}\ )$ koje se do sada objašnjavao kao rezidual ε, odnosno statistička pogreška

$$SS_{yy} = SSR + SSE$$

$$\sum(y - \bar{y}\ )^2 = \sum(\hat{y} - \bar{y}\ )^2 + \sum(y - \hat{y}\ )^2$$

SSE – predstavlja neprotumačena odstupanja, odnosno sva odstupanja koja se nalaze iznad linije regresije ($\hat{y}$) ne mogu se objasniti nezavisnom varijablom

2.6.3. p – vrijednost

P – value, odnosno p – vrijednost predstavlja razinu značajnosti testa. P – vrijednost je ključna statistička mjera koja se koristi za procjenu statističke značajnosti rezultata u mnogim znanstvenim i statističkim istraživanjima. S obzirom da se p-vrijednost

vezuje uz pogrešku tipa I (odbacivanje točne nul hipoteze, može se tumačiti i kao empirijska vjerojatnost pogreške tipa I ili vjerojatnost da će se učiniti pogrešku ako se odbaci nul hipoteza).

Statističke hipoteze regresijskog modela su sljedeće:

- Nul hipoteza (H_0): Ne postoji statistički značajna linearna veza između nezavisne varijable x i zavisne varijable y

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

- Alternativna hipoteza (H_1): Postoji statistički značajna linearan veza između nezavisne varijable x i ovisne varijable y

$$H_1: \beta_0 \neq \beta_1 \neq 0$$

Odluka o nul hipotezi regresijskog modela utvrđuje se temeljem F-testa i pripadajuće p-vrijednosti. No, detaljniji uvid omogućuje zasebno testiranje konstante i koeficijenta smjera u regresijskom modelu, što se provodi t-testom. U tom slučaju, postoje dva skupa hipoteza koje glase:

$$H_0: \beta_0 = 0 \qquad H_0: \beta_1 = 0$$
$$H_1: \beta_0 \neq 0 \qquad H_1: \beta_1 \neq 0$$

Uobičajeno je da se postavlja prag, često nazivanog "razina značajnosti" i označava se s α (najčešće 0.05 ili 0.01). Ako je p-vrijednost manja od tog praga, obično se odbacuje nulta hipoteza u korist alternativne hipoteze.

Interpretacija: Ako je p-vrijednost mala (npr. manja od 0.05), to bi značilo da su rezultati vjerojatno rezultat nekog efekta ili razlike u stvarnosti, a ne slučajnosti. Ako je p-vrijednost velika, to bi značilo da rezultati mogu biti posljedica slučajnosti i nulta hipoteza se ne odbacuje. Istovremeno, to znači da se model odbacuje.

Također, p-vrijednost nije garancija da je nulta hipoteza točna ili netočna. Ona samo pruža informaciju o vjerojatnosti rezultata s obzirom na nultu hipotezu. Interpretacija p-vrijednosti zahtjeva pažljivo razmatranje konteksta i dizajna istraživanja.

Hipoteza u kontekstu statističkog koeficijenta p-vrijednosti i razine značajnosti obično se formulira kako bi se testirala statistička značajnost u istraživanju. Postoje dvije osnovne hipoteze koje se često postavljaju (Belullo, 2011.):

- Nulta hipoteza (H_0): Nulta hipoteza obično tvrdi da nema značajne razlike ili efekta između varijabli ili skupova podataka. To znači da bilo koji promatrani učinak ili razlika nije statistički značajan.
- Alternativna hipoteza (H_1 ili H): Alternativna hipoteza izražava suprotno stajalište od nulte hipoteze i tvrdi da postoji značajna razlika, efekt ili veza između varijabli.

Testiranje statističkih hipoteza je postupak donošenja odluke odbacivanju ili neodbacivanju H_0 na osnovu informacije dobivene iz opažanja slučajnog uzorka.

U modelu jednostavne linearne regresije koriste se različiti statistički testovi kako bi se procijenila statistička značajnost i adekvatnost regresijskog modela. Važni testovi koji se ovdje često primjenjuju su (Belullo, 2011.):

- Testiranje koeficijenta nagiba: Ovaj test se koristi za ispitivanje je li koeficijent nagiba statistički različit od nule. Ako je koeficijent nagiba značajan, to ukazuje na postojanje linearnog odnosa između nezavisne i zavisne varijable.
- Testiranje koeficijenta odsječka: Ovaj test se primjenjuje kako bi se utvrdilo je li koeficijent odsječka statistički različit od nule. Ako je koeficijent odsječka značajan, to znači da model nije prisiljen prolaziti kroz ishodište.
- Testiranje ukupne statističke značajnosti modela: Ovaj test, poznat kao F-test, ocjenjuje ukupnu statističku značajnost regresijskog modela. Provjerava je li barem jedan od koeficijenata regresije (nagib ili odsječak) statistički značajan.
- Testiranje normalnosti ostataka: Normalnost reziduala podrazumijeva da su ostaci regresije normalno distribuirani. To se obično provjerava pomoću grafika kvantil-kvantil (QQ plot) i testova normalnosti poput Kolmogorov-Smirnov testa ili Shapiro-Wilk testa.
- Testiranje homoskedastičnosti: Ovaj test provjerava jednakost varijabilnosti ostataka kroz različite razine prediktora. Ako su ostaci homoskedastični, to znači da varijabilnost ostataka ostaje konstantna kroz različite vrijednosti nezavisne varijable.

Ovi testovi su važan dio analize jednostavne linearne regresije i pomažu u procjeni valjanosti regresijskog modela i statističke značajnosti njegovih komponenti.

Pri statističkom testiranju hipoteza mogu se napraviti dvije vrste grešaka: može se odbaciti hipoteza koja je istinita ili se može ne odbaciti hipoteza koja je kriva. Odbacivanjem istinite hipoteze zove se greška tipa I. tipa, dok se neodbacivanje krive hipoteze zove greška II. tipa. Vjerojatnost nastanka greške I. tipa neposredno kontrolira istraživač proizvoljnim odabirom razine značajnosti koja se označava sa α. Uobičajeno je testirati na razini značajnosti od 5%, odnosno s određenom vjerojatnošću greške I. tipa od 5%.

2.6.4. Tumačenje koeficijenta determinacije i njegova povezanost s reprezentativnosti regresijskog modela

Koeficijent determinacije igra ključnu ulogu u tumačenju i procjeni reprezentativnosti regresijskog modela. R^2 se koristi kako bi se razumjelo koliko dobro model objašnjava varijabilnost zavisne promjenjive varijable i kako bi se procijenila adekvatnost modela za analizirane podatke. Tumačenje koeficijenta determinacije u kontekstu kvalitete regresijskog modela:

1. R^2 se kreće od 0 do 1: R^2 je rezultat koeficijenta determinacije oscilira između 0 i 1, gdje 0 označava da model ne objašnjava varijabilnost u podacima, a 1 označava da model savršeno objašnjava varijabilnost. Dakle, što je veći R^2, to je bolji model u objašnjavanju varijacije u ciljnoj varijabli.
2. Visok R^2: Ako R^2ima visoku vrijednost, recimo blizu 1, to ukazuje na to da veći dio varijabilnosti ciljne varijable može biti objašnjen modelom. To je dobar znak i sugerira da model dobro odgovara podacima.
3. Nizak R^2: Ako je R^2 nizak, blizu 0, to ukazuje na to da model ne uspijeva dobro objasniti varijabilnost ciljne promjenjive varijable. Ovo može biti znak da model nije odgovarajući za analizirane podatke ili da nedostaju bitni regresori.
4. R^2od 0.5 ili manje: U nekim slučajevima, R^2 vrijednost od 0.5 ili manje može se smatrati prihvatljivim, ali to zavisi od konteksta. Na primjer, u društvenim i humanističkim znanostima, gdje varijabilnost može biti jako visoka, niže vrijednosti R^2 mogu biti očekivane.

Važno je napomenuti da R^2 ima svoja ograničenja i ne daje potpunu sliku o kvaliteti modela. Koeficijent determinacije prate 3 problema (Belullo, 2011.):

1. Ne vrijedi za regresijske modele koji su izračunati metodom najmanjih kvadrata bez konstantnog člana β_0 (kada u modelu nema konstantnog člana koristi se necentrirani koeficijent determinacije)
2. Ne vrijedi za regresijske modele koji nisu izračunati pomoću metode najmanjih kvadrata (OLS)
3. Koeficijent determinacije nikada ne opada s povećanjem broja nezavisnih varijabli čak i kada dodatne varijable ne objašnjavaju kretanje zavisne varijable (rješavanje ovog problema jest uključivanje u izračun koeficijenta determinacije stupnjeve slobode što onda daje korigirani koeficijent determinacije $\bar{R}^2$)

Stvari koje treba uzeti u obzir prilikom tumačenja R^2:

- Korelacija i uzročnost: Visok R^2 ne implicira uzročnu vezu između regresora i ciljne promjenjive. Važno je razmotriti korelaciju i kontekstualnu relevantnost regresora.
- Heteroskedastičnost i pretjerano prilagođavanje: R^2 može biti zavaravajuće ako model pati od heteroskedastičnosti (neravnomjerne varijabilnosti) ili pretjeranog prilagođavanja podacima. U takvim slučajevima, model može imati visok R^2, ali lošu sposobnost generalizacije na nove podatke.

R^2 je koristan alat za procjenu kvaliteta regresijskog modela, ali treba ga koristiti zajedno s drugim metrikama i kontekstualnim razmatranjima kako biste dobili potpuniju sliku o tome koliko dobro model odgovara podacima i da li je adekvatan za analizu.

2.7. Razlika između predikcije, korelacije i kauzalnosti

Predikcija u statistici se odnosi na proces procjene ili predviđanja budućih vrijednosti, događaja ili rezultata na temelju analize postojećih podataka i statističkih modela. Osnovna svrha predikcije u statistici je razumjeti uzorke u podacima i koristiti te uzorke kako bi se donosile informirane prognoze ili predviđanja.

Ključne komponente predikcije u statistici uključuju (Neuman, 2014.):

- Podaci: Prvo, potrebni su relevantni i kvalitetni podaci koji opisuju prošle ili trenutne događaje ili varijable. Ovi podaci služe kao temelj za izgradnju statističkog modela.
- Statistički model: Nakon prikupljanja podataka, koristi se odgovarajući statistički model kako bi se opisala veza između varijabli. To može uključivati linearnu regresiju, logističku regresiju, vremenske serije ili druge statističke metode, ovisno o prirodi problema.
- Analiza i procjena modela: Nakon što je model kreiran, provodi se analiza kako bi se procijenila njegova točnost i adekvatnost. To uključuje korištenje različitih metrika i testova kako bi se ocijenila prediktivna sposobnost modela.
- Predviđanja: Nakon što je model ocijenjen i verificiran, može se koristiti za predviđanje budućih vrijednosti ili događaja na temelju novih ulaznih podataka ili scenarija. To su predviđanja ili prognoze koje model generira na temelju svoje strukture i parametara.

Terminologija korelacije u kontekstu ovoga rada je pojašnjena u prvom poglavlju, međutim, važno je istaknuti njezinu važnost u statistici te u primjeni jednostavne linearne regresije. Korelacija je ključna statistička mjera koja igra izuzetno važnu ulogu u statistici i analizi podataka. Ona nam omogućuje da razumijemo i kvantificiramo međusobne odnose između različitih varijabli, što je ključno za donošenje informiranih odluka u mnogim disciplinama (Belullo, 2011.). Korelacija također igra ključnu ulogu u primjeni jednostavne linearne regresije, jednog od osnovnih alata za modeliranje i predviđanje.

Korelacija omogućuje da se kvantificira stupanj povezanosti između dviju ili više varijabli. To znači da možemo odgovoriti na pitanje koliko su varijable međusobno ovisne i u kojem smjeru (pozitivno ili negativno) (Neuman, 2014.). Korelacija pomaže u identifikaciji uzoraka u podacima. Ako dvije varijable pokazuju snažnu pozitivnu korelaciju, to znači da kada jedna varijabla raste, i druga obično raste. Ovo znanje može biti od velike koristi za razumijevanje prirode podataka. Na primjer, ako postoji jaka korelacija između dviju varijabli, možemo kreirati model u kojem ćemo koristiti jednu varijablu kao pokazatelj za predviđanje vrijednosti druge.

Kauzalnost u statistici se odnosi na odnos uzroka i posljedice između dviju varijabli, gdje se tvrdi da jedna varijabla uzrokuje promjene u drugoj varijabli. Drugim riječima, kauzalnost implicira da postoji uzročna veza između dvije varijable, gdje promjene u jednoj varijabli izazivaju promjene u drugoj varijabli (Neuman, 2014.).

Kauzalnost je ključni koncept u mnogim područjima, uključujući znanost, ekonomiju, medicinu, sociologiju i druge discipline, jer pomaže u razumijevanju uzroka i učinaka te omogućava donošenje informiranih odluka (Belullo, 2011.). Međutim, utvrđivanje stvarne kauzalne veze između varijabli može biti izazovno i često zahtijeva dublje istraživanje i analizu.

Važno je napomenuti da statistika sama po sebi ne može utvrditi uzročnu vezu između varijabli. Statističke metode mogu identificirati korelacije između varijabli, ali korelacija ne implicira nužno uzročnost. Postoji mnogo čimbenika koji mogu utjecati na promjene u varijablama, uključujući i treće, skrivene varijable koje nisu uključene u analizu.

Razlika između predikcije, korelacije i kauzalnosti ključna je u statistici i analizi podataka, jer svaki od ovih pojmova ima svoju specifičnu svrhu i značajku (Neuman, 2014.):

- Predikcija se odnosi na procjenu budućih vrijednosti, događaja ili rezultata na temelju analize postojećih podataka i statističkih modela.
- Korelacija se odnosi na statističku mjeru koja kvantificira stupanj međusobnog odnosa između dviju ili više varijabli.
- Kauzalnost se odnosi na odnos uzroka i posljedice između dviju varijabli, gdje se tvrdi da jedna varijabla uzrokuje promjene u drugoj varijabli.

Ključna razlika između ovih pojmova je u tome što se predikcija bavi budućim događajima, korelacija se bavi povezanošću između varijabli bez impliciranja uzroka, dok se kauzalnost bavi stvarnom uzročnom vezom između varijabli. Važno je napomenuti da kauzalnost često zahtijeva pažljivu analizu i kontrolu potencijalnih zbunjujućih (*engl. confounding*) faktora kako bi se pouzdano utvrdila uzročna veza.

Za utvrđivanje kauzalnih veza između varijabli često je potrebno provesti kontrolirane eksperimente ili koristiti naprednije statističke metode poput eksperimentalnog dizajna, analize prekidača (*engl. interrupted time series analysis*), ili metode za procjenu uzročnih efekata u okviru uzročne inferencijalne analize (*engl. causal inference*). Ovi

pristupi pomažu istražiteljima da identificiraju uzročne veze uzimajući u obzir različite faktore i kontrolirajući potencijalne zbunjujuće (*engl. confounding*) varijable.

3. Primjer kreiranja modela jednostavne linearne regresije

3.1. Teorijsko opravdanje traženja veze između promatranih varijabli

Promatraju se dvije varijable: BDP i HCI (*engl. Human Capital Index*).

BDP je makroekonomski indikator koji pokazuje vrijednost finalnih dobara i usluga proizvedenih u zemlji tijekom dane godine, izraženo u novčanim jedinicama. HCI izračunava doprinose zdravlja i obrazovanja produktivnosti radnika. U radu (Gulcemal, 2020) promatrao se utjecaj raznih varijabli, od kojih je jedna i HCI, na BDP te se pronašla statistički značajna pozitivna veza između BDP-a i HCI-a. Jedan argument zašto se očekuje pozitivna veza između te dvije varijable je i taj da povećanjem produktivnosti radnika putem obrazovanja i boljeg zdravlja raste i vrijednost finalnih dobara i usluga proizvedenih u toj zemlji.

3.2. Model jednostavne linearne regresije

U ovom dijelu kreira se model jednostavne linearne regresije koji promatra vezu između varijabli HCI i BDP-a. Smislenost i motivacija za promatranjem ove veze prethodno su komentirane.

Slika 1. Ispis statističkog software-a Gretl - Model jednostavne linearne regresije

	coefficient	std. error	t-ratio	p-value	
const	-0.919701	0.346348	-2.655	0.0087	***
Humancapitalinde~	1.71138	0.596794	2.868	0.0047	***

Mean dependent var	0.045273	S.D. dependent var	1.097584
Sum squared resid	196.4971	S.E. of regression	1.075112
R-squared	0.046140	Adjusted R-squared	0.040529
F(1, 170)	8.223262	P-value(F)	0.004660
Log-likelihood	-255.5086	Akaike criterion	515.0172
Schwarz criterion	521.3122	Hannan-Quinn	517.5712

Slika 2. Ispis statističkog software-a Gretl - Deskriptivne statistike promatranih varijabli

	Mean	Median	Minimum	Maximum
GDPcurrentUSNYGD~	4.0288e+011	2.4930e+010	5.1747e+007	2.1060e+013
Humancapitalinde~	0.56128	0.56383	0.29163	0.87913
	Std. Dev.	**C.V.**	**Skewness**	**Ex. kurtosis**
GDPcurrentUSNYGD~	1.8520e+012	4.5970	9.0368	89.167
Humancapitalinde~	0.13909	0.24781	0.11652	-0.95913
	5% perc.	**95% perc.**	**IQ range**	**Missing obs.**
GDPcurrentUSNYGD~	7.8700e+008	1.5687e+012	1.6986e+011	8
Humancapitalinde~	0.36233	0.79329	0.22046	43

Tablica 1. Tablica prikazuje aritmetičke sredine i standardne devijacije promatranih varijabli.

Varijabla	**Aritmetička sredina**	**Standardna devijacija**
BDP	$4.029 \cdot 10^{11}$	$1.852 \cdot 10^{12}$
HCI	0.5613	0.1391

Prije kreiranja i evaluacije modela linearne regresije moraju se pripremiti podaci. Varijabla BDP se standardizira te iz skupa podataka se izbacuju one države koje nemaju oba opažanja. Za analizu podataka koristi se statistički software Gretl.

Tablica 2. Tablica prikazuje rezultate regresijske analize.

	Coefficient	**Std. error**	**t-ratio**	**p-value**
Konstanta	−0.919701	0.346348	−2.655	0.0087
HCI	1.71138	0.596794	2.868	0.0047

Dobiveni model jednostavne linearne regresije glasi:

$$BDP_s = -0.919701 + 1.71138 \cdot HCI$$

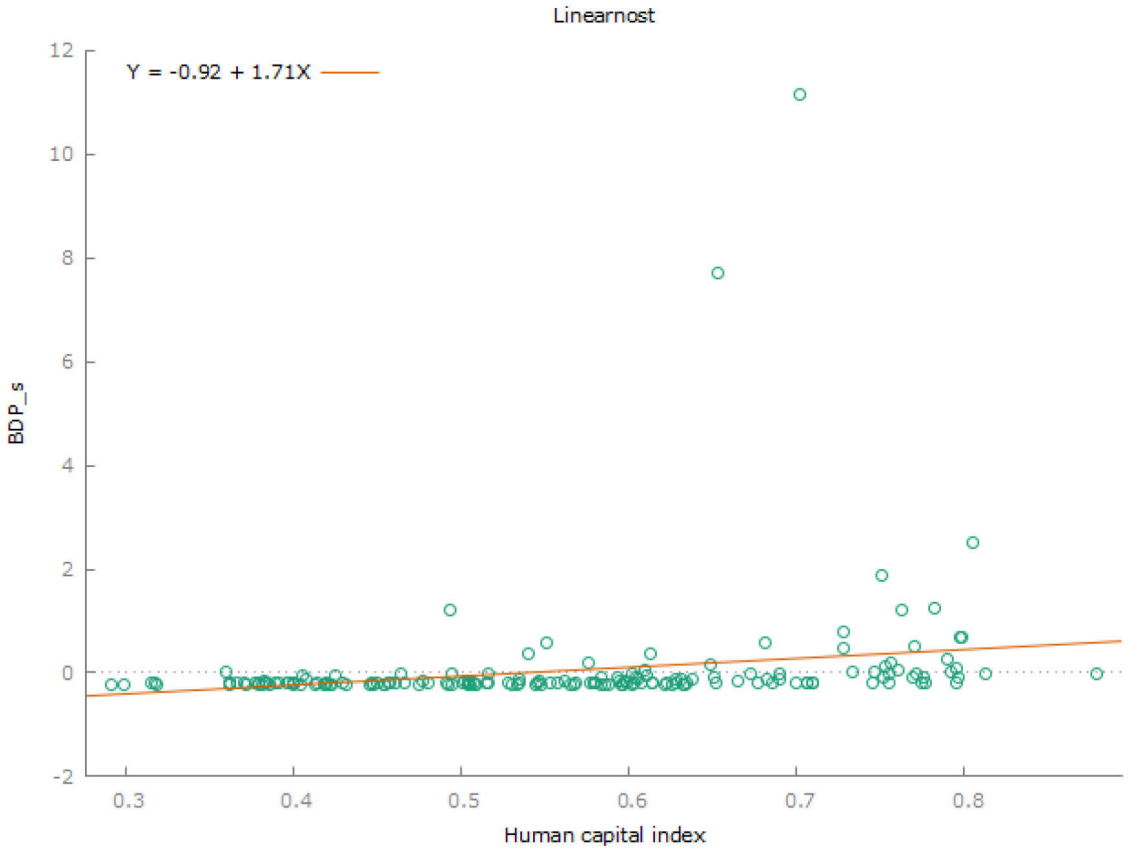
Koeficijenti regresije su statistički značajni te koeficijent uz nezavisnu varijablu (HCI) može se interpretirati na način da se jediničnim povećanjem nezavisne varijable zavisna varijabla (BDP_s) poveća za 1.71138. Odnosno, u terminima originalne varijable to znači da se BDP poveća za 1.71138 standardnih devijacija. U nastavku se evaluira dobiveni regresijski model.

3.3. Evaluacija regresijskog modela

3.3.1. Linearnost

Pretpostavka linearnosti zahtijeva da je veza između promatranih varijabli približno linearna. Na slici ispod prikazan je odnos promatranih varijabli i dodan je prethodno izračunati pravac linearne regresije. Iz slike se vidi da je veza između promatranih varijabli približno linearna. Dakle, pretpostavka linearnosti je zadovoljena. Također, može se uočiti da dvije vrijednosti značajno odstupaju od ostatka podataka, u takvim slučajevima treba provjeriti radi li se možda o grešci u podacima ili su to uistinu validni podaci. U ovom slučaju to su validni podaci i radi se o Kini i Sjedinjenim Američkim Državama koje imaju značajno veći BDP od ostatka država u promatranom skupu podataka.

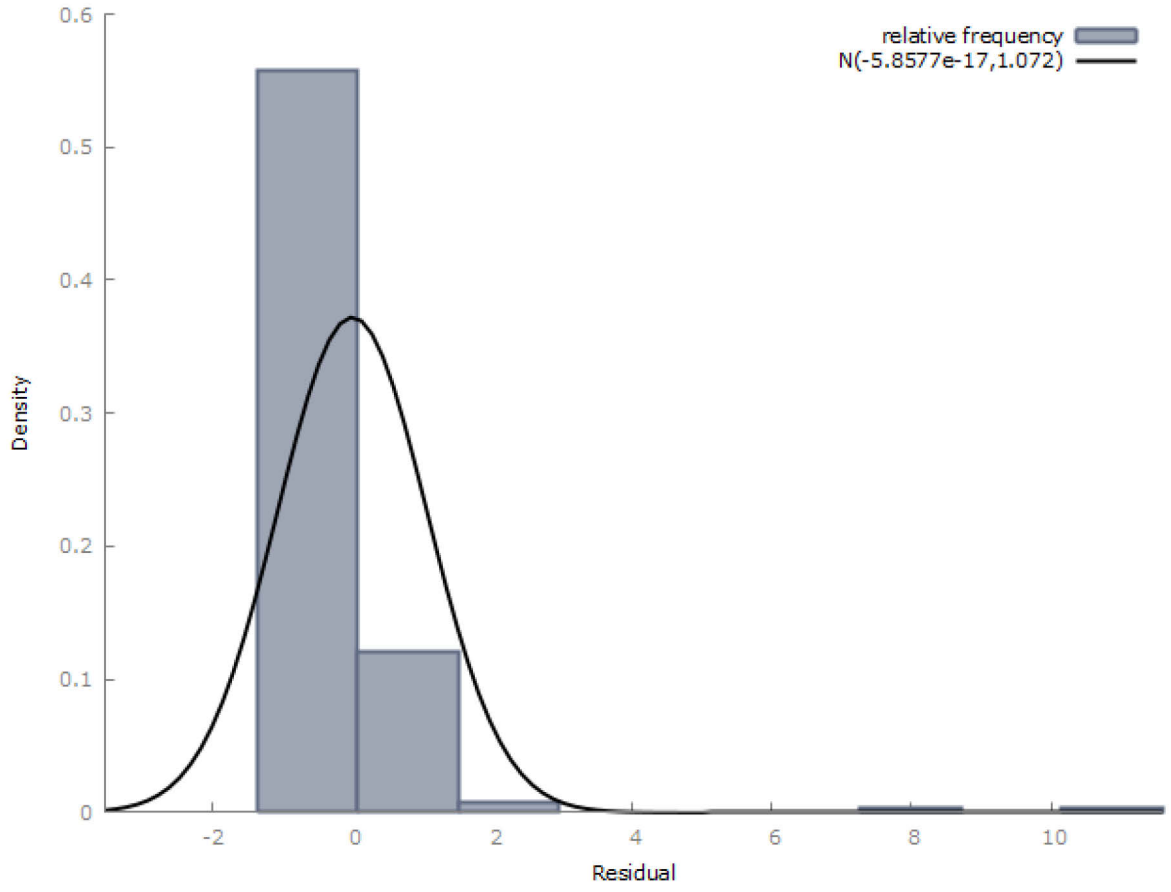
Graf 2. Graf disperzije promatranih varijabli: BDP i HCI (Human capital index).



3.3.2. Normalnost reziduala

Distribucija reziduala bi trebala biti približno normalna. Normalnost reziduala provjerava se histogramom i Q-Q grafom. Na slici ispod je prikazan histogram reziduala i za usporedbu je dodana gustoća normalne distribucije.

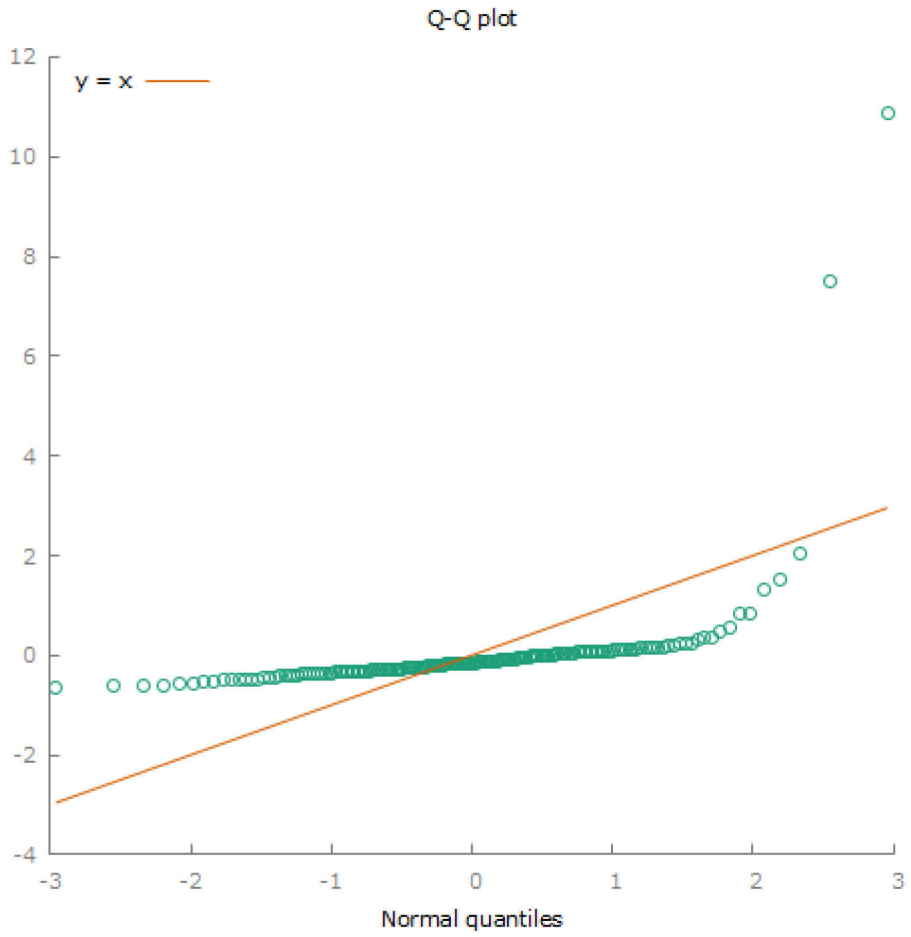
Graf 3. Histogram reziduala



Distribucija reziduala značajno odstupa od normalne distribucije jer nije simetrična te ima previše jako velikih vrijednosti a premalo malih u odnosu na što bi bilo očekivano u normalnoj distribuciji.

Odstupanje od normalnosti se jasno vidi i iz Q-Q grafa na kojemu bi u slučaju normalne distribucije podaci bili grupirani oko pravca, a u ovom slučaju značajno odstupaju od istog kao što je vidljivo iz slike ispod.

Graf 4. Q-Q graf

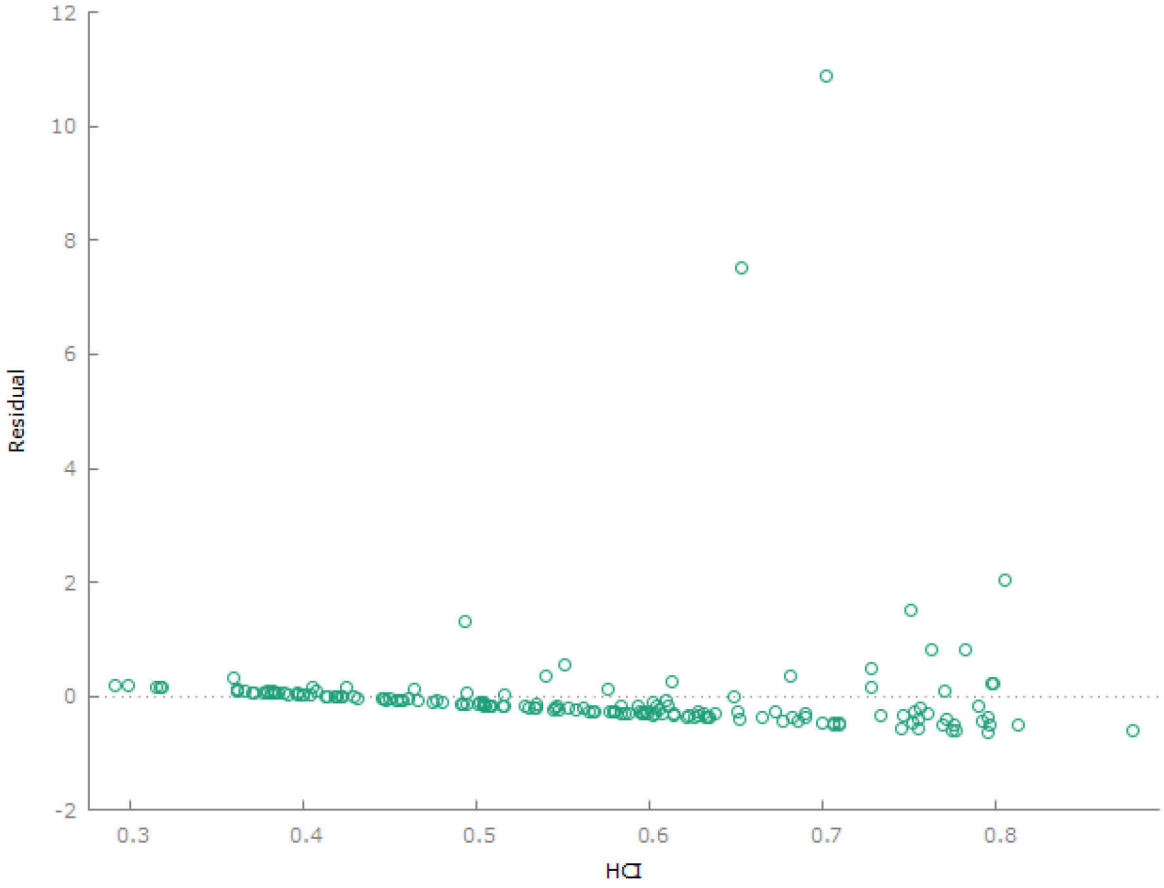


Dakle, prema prikazanom Q-Q grafu pretpostavka normanosti reziduala je narušena jer distribucija reziduala značajno odstupa od normalne što znači da ovaj model nije adekvatan i trebalo bi koristiti metode koje nemaju za pretpostavku normalnost reziduala.

3.3.3. Homogenost varijance

Varijabilnost reziduala bi trebala biti približno jednaka neovisno o vrijednosti nezavisne varijable. Na slici ispod se vidi da je varijabilnost reziduala veća za veće vrijednosti od HCI. Dakle, homogenost varijance nije zadovoljena.

Graf 5. Graf disperzije



3.4. Zaključak istraživanja

Model jednostavne linearne regresije pokazuje da postoji statistički značajna pozitivna veza između dviju promatranih varijabli. Koeficijent smjera iznosi 1.71138 i statistički je značajan ($t = 2.868,\ p = 0.0047$), a konstanta iznosi -0.919701 i statistički je značajna ($t = -2.655,\ p = 0.0087$). F-test značajnosti modela pokazuje da je model statistički značajan ($F(1{,}170) = 8.223262,\ p = 0.00466$).

Rezultati evaluacije modela ukazuju na to da ovaj model nije dobar jer su narušene pretpostavke o normalnosti i homogenosti varijance reziduala. Također, vrijednost statistike $R^2 = 4.61\%$ pokazuje da ovaj model objašnjava svega 4.61% varijabilnosti što znači da sam HCI nije dobar prediktor BDP-a već bi trebalo razmotriti i druge varijable koje rezultiraju promjenama u BDP-u.

4. Primjena jednostavne linearne regresije u ekonomiji i poslovnoj ekonomiji

Uvođenje i primjena statističkih metoda, poput jednostavne linearne regresije, ima ključnu ulogu u ekonomiji i poslovnoj ekonomiji. Ovdje se analiziraju odabrani radovi objavljeni na hrvatskom jeziku u otvorenom pristupu putem Portala hrvatskih znanstvenih i stručnih časopisa Hrčak iz područja društvenih znanosti, koji u ključnim riječima navode linearnu regresiju. Pretraga po ključnom pojmu „linearna regresija“ na Hrčku rezultira s 49 radova na dan 14.9.2023., od kojih je 12 iz područja društvenih znanosti. Ovdje su odabrani i predstavljeni radovi koji su tematski bliski području ekonomije i poslovne ekonomije.

4.1. Primjena u ekonomiji – komparativna analiza jednostavne i višestruke linearne regresije

U ekonomskoj analizi, statističke metode igraju ključnu ulogu jer omogućavaju preciznu analizu podataka i donošenje informiranih ekonomskih odluka. Jedna od najčešće korištenih tehnika u analizi ekonomske povezanosti između različitih faktora jest linearna regresija. U ovom poglavlju, autorica će se usredotočiti na primjenu linearne regresije u ekonomiji, posebno istražujući dvije ključne varijante: jednostavnu i višestruku linearnu regresiju.

Jednostavna i višestruka linearna regresija su temeljni alati za modeliranje odnosa između ekonomskih varijabli. Ova poglavlja će analizirati njihove karakteristike, prednosti i ograničenja kako bi čitatelj bolje razumio njihovu primjenu u kontekstu ekonomske analize. Autorica će se posebno posvetiti usporedbi između ove dvije statističke tehnike kako bi istaknula kada je prikladno koristiti jednostavnu ili višestruku linearnu regresiju u ekonomskom istraživanju.

U znanstvenim istraživanjima u kojima se implementira jednostavna i višestruka linearna regresija brojni autori su odabrali i primijenili ove statističke metode da bi dokazali određene hipoteze od znanstvenog značaja za akademsku zajednicu u ekonomiji.

Pejić Bach i Demonja (2008) istražili su kako se može otkriti porezna utaja iz informacijskog sustava porezne uprave metodom otkrivanja znanja iz baza podataka. U svome istraživanju njihov je cilj bio razviti model za otkrivanje porezne evazije, koristeći podatke iz informacijskog sustava Porezne uprave. Primjenom metode linearne regresije i odabirom slučajnog uzorka od 200 poduzeća odredili su ključnu zavisnu varijablu koja se odnosila na iznos utajenog poreza koji je identificiran tijekom poreznih nadzora provedenih u razdoblju od 2001. do 2005. godine. U kontekstu učinkovitosti jednostavne linearne regresije definirali su i nezavisne varijable (Pejić Bach i Demonja, 2008): dohodak vlasnika, ukupan prihod poduzeća te informacije o poslovnim i stambenim nekretninama koje su poduzeća kupila. U provedenoj analizi podijelili su istraživanje na pet modela koji su obuhvaćali nezavisne varijable te njihov učinak na zavisne. Koristili su test značajnosti nezavisnih varijabli modela na razini značajnosti p = 0,05, test tolerancije za isključenu varijablu regresijskog modela te u cijelom kontekstu deskriptivnu statistiku. Rezultati primjene jednostavne linearne regresije u kontekstu otkrivanja porezne utaje iz informacijskog sustava porezne uprave metodom otkrivanja znanja iz baza podataka autori su otkrili kako je porezna utaja u korelaciji s kupnjom stambenih nekretnina te su dali preporuku kako treba pratiti ovaj segment nezavisne varijable u unapređenju redukcije poreznih utaja.

Raos (2017) je istražio odnose nacionalnih i podnacionalnih (regionalnih i lokalnih) izbora. Analizom odnosa nacionalnih i podnacionalnih izbora u slučaju Hrvatske u lipnju 2017. autor je analizirao izlaznost građana na lokalne izbore i postotak glasača za HDZ, gdje je primjenom statističkih izračuna koristio metode deskriptivne statistike i njezine temeljne kvantitativne pokazatelje (mod, medijan, standardnu devijaciju, aritmetičku sredinu). Osim toga, iskorištena je i metoda višestruke linearne regresije s izračunom regresora, konstantom β_0, koeficijentom determinacije i korigiranim koeficijentom determinacije, rezidualnom pogreškom, F – test te Durbin – Watson test. Raos (2017) je proveo istraživanje u domeni političke znanosti, gdje je primjenom višestruke linearne regresije ukazao na utjecaj lokalnih izbora na uspjeh HDZ – a kao vodeće hrvatske stranke i neizravno utjecaja njihove vlasti na unapređenje lokalnog razvoja jedinica lokalne samouprave u Hrvatskoj. Po istom principu analiziran je postotak glasača za MOST, gdje je autor koristio i grafičke ilustracije – dijagrame raspršenosti i distribuciju glasova za obje stranke. Analiza detaljnog razmatranja predloženih modela linearne regresije otkrila je da se u slučaju HDZ-a može smatrati

da je konstruirani prediktivni model, koji se temelji na rezultatima parlamentarnih izbora i koristi za predviđanje ishoda lokalnih izbora, valjan.

Guzić (2014) je istražio zašto se nematerijalna imovina poistovjećuje s knjigovodstvenim pojmom onog dijela nedodirljive imovine koja se iskazuje u financijskim izvješćima. S obzirom da računovodstvo tretira nematerijalnu imovinu kao puno širi pojam od navedenog, autor je istražio analize međuovisnosti poslovnih rezultata i vrijednosti te nematerijalne imovine. U svojoj analitici koristio je sekundarne podatke s naglaskom na jednostavnu linearnu regresiju. U analizi financijskih indikatora odabranih poduzeća koristio je deskriptivnu statistiku i njezine temeljne kvantitativne pokazatelje (mod, medijan, standardnu devijaciju, aritmetičku sredinu). U analizi rasta nematerijalne imovine, prihoda, dobiti i imovine koristio je pokazatelje jednostavne linearne regresije – koeficijent korelacije i determinacije, analizu trenda i razinu značajnosti p = 0,00 s grafičkim prikazima dijagrama raspršenosti. Primjenom jednostavne linearne regresije došao je do zaključka kako postoji veza između nematerijalne imovine i testiranih vrijednosti (profita, prihoda i neto dobiti) kao zavisne varijable.

Antić (2004) je istražio utjecaj političkog sistema na porast životnog vijeka, gdje je u svome istraživanju primijenio metodu višestruke linearne regresije. U kontekstu utvrđivanja nezavisne i zavisne varijable, zavisna varijabla je povećanje životnog vijeka, koja ovisi o političkom režimu, rastu BDP – a per capita i duljini životnog vijeka građana kao nezavsnim varijablama. Metodom višestruke linearne regresije testirao je tri hipoteze (Antić, 2004):

1. Demokratski režimi imaju brži porast životnog vijeka nego diktatorski režimi.
2. Što je viši porast BDP per capita, to je veće povećanje životnog vijeka.
3. Što je viši postojeći nivo životnog vijeka, to je manje povećanje životnog vijeka.

Autor je višestrukom regresijom testirao nezavisne varijable na razini značajnosti 0,05, 0,01 i 0,001 te je prikazao dobivene rezultate pomoću regresijskih koeficijenata. Rezultati njegova istraživanja pokazali su da iako je prosječni životni vijek dulji u demokracijskim društvima, zanimljivo je primijetiti da je povećanje životnog vijeka bilo veće u diktatorskim režimima tijekom perioda od 1960. do 1999. godine. Dodatno, analiza linearnom regresijom pokazala je da u demokracijama, čak i uz uzimanje

kontrolnih varijabli u obzir, porast životnog vijeka bio je manji nego u državama pod diktatorskim režimom. Nadalje, regresijski model je sugerirao da se u zemljama s već visokim životnim vijekom, kao i u državama s brzim rastom BDP per capita, životni vijek povećava brže.

Kontuš i Šarlija (2019) su istražile utjecaj strukture kapitala na likvidnost dioničkih društava, čiji su financijski instrumenti uvršteni na tržište kapitala, i primjenu teorije hijerarhije financijskih izbora. U svome empirijskom istraživanju primijenile su panel višestruku linearnu regresiju i analizu korelacije. Autorice su definirale zavisne varijable (Kontuš i Šarlija, 2019): pokazatelji performanse likvidnosti poslovanja: koeficijent trenutne likvidnosti i koeficijent tekuće likvidnosti, kao i nezavisne: udjeli pojedinačnih komponenti strukture kapitala u ukupnom kapitalu i obvezama. U empirijskom istraživanju su primijenjene različite statističke metode kako bi se analizirali podaci. Korištene su panel višestruke linearne regresije i analiza korelacije, pri čemu je za mjerenje korelacije primijenjen Pearsonov koeficijent korelacije. U panel-analizi su se koristili različiti statistički modeli. Prvo, korišten je model s fiksnim efektima, koji predstavlja linearni model u kojem se konstantni član mijenja za svaku jedinicu promatranja, ali ostaje konstantan tijekom vremena. Osim toga, primijenjeni su i modeli sa slučajnim efektima te modeli s konstantnim regresijskim parametrima kako bi se bolje razumjeli i analizirali podaci i njihove promjene tijekom vremena.

Rezultati istraživanja provedenog u ovom radu ukazuju na pozitivnu korelaciju između udjela zadržane dobiti u ukupnom kapitalu i obveza te pokazatelja tekuće likvidnosti dioničkih društava čiji su financijski instrumenti kotirani na tržištu kapitala u Republici Hrvatskoj. Ova korelacija je statistički značajna posebno u 2009. i 2010. godini. Važno je napomenuti da rezultati analize veze između udjela zadržane dobiti u ukupnom kapitalu, obveza i pokazatelja tekuće likvidnosti hrvatskih dioničkih društava nisu dosljedni s rezultatima istraživanja Šarlija i Harz (2012). Naime, dokazano je da postoji pozitivna povezanost između udjela zadržane dobiti u ukupnom kapitalu i obveza, kao i pokazatelja tekuće likvidnosti, i to u određenim godinama promatranog razdoblja.

Šućur (2021) je u svome istraživanju analizirao vezu između dohodovnih nejednakosti i redistributivnih preferencija u Hrvatskoj i zemljama EU-a. Autor je koristio sekundarne izvore podataka: istraživanja Eurobarometra (2010. i 2018.) te makrostatističke

pokazatelje iz Eurostatove baze podataka. U ovom istraživanju su primijenjene različite statističke metode kako bi se analizirali podaci. Konkretno, korištene su bivarijatne korelacijske analize kako bi se istražila veza između dvije varijable. Također, primijenjena je linearna regresija kako bi se modelirali odnosi između varijabli te klaster analiza kako bi se grupirali slični podaci u određene klastere ili skupine. Ove različite metode analize omogućile su detaljno istraživanje različitih aspekata podataka i njihovih međusobnih odnosa. Osim toga, autor je deskriptivnom statistikom prikazao stanje dohodovnih nejednakosti i redistributivnih preferencija u Hrvatskoj i EU, a u primjeni višestruke linearne regresije je koristio grafičke prikaze raspršenosti, koeficijent korelacije, F – test, standardnu pogrešku i razine značajnosti. Rezultati njegova istraživanja pokazali su kako se uočavaju visoke redistributivne preferencije u gotovo svim zemljama EU-a. Istraživanje ukazuje na to da rast dohodovnih nejednakosti nije ključni faktor za visoku razinu redistributivnih preferencija. Umjesto toga, ključnim se čini percepcija dohodovnih nejednakosti i osjetljivost građana na ekonomske nejednakosti. Građani EU-a često imaju nerealnu percepciju razine nejednakosti u društvu i svojeg položaja na dohodovnoj ljestvici. Posebno je zanimljivo da ispitanici iz postsocijalističkih zemalja pokazuju veću "averziju" prema dohodovnim nejednakostima i iskazuju želju za većom ulogom države u redistribuciji i društvenom životu. Stanovnici EU-a podržavaju sve ključne mehanizme dohodovne redistribucije, uključujući porezni sustav, obrazovanje, socijalnu zaštitu i minimalnu plaću. Međutim, najveću podršku dobiva porezni sustav i progresivno oporezivanje bogatijih građana, dok se najviše raspravlja o potpuno besplatnom obrazovanju.

Šandrk Nukić i Šuvak (2013) istražili su odnos relevantnih aktivnosti upravljanja ljudskim potencijalima s odabranim kvalitativnim pokazateljem – percepcijom organizacijske uspješnosti stručnjaka za ljudske potencijale. U provedbi empirijskog istraživanja koristili su deskriptivnu statistiku i višestruku linearnu regresiju, s tabličnim prikazima pripadajućih pokazatelja višestruke regresije i dijagramima raspršenosti. Njihovi rezultati istraživanja pokazali su da percepcija organizacijske uspješnosti najviše ovisi o adekvatnosti, kvaliteti i obimu obuke za zaposlenike, načinu pohrane poslovnih informacija i dostupnosti istih te o efikasnosti procesa usvajanja znanja od strane partnera.

Starc (2016) je istražio koji su to čimbenici koji utječu na proces profesionalizacije u lancu zdravstvene zaštite. U istraživanju su primijenjene kvantitativne i kvalitativne

metodologije za analizu podataka. U kvantitativnom dijelu analize koristili su se deskriptivna statistika, analiza kontingencije, metoda najmanjih kvadrata i multivarijatna linearna regresija, s i bez kontrolnih varijabli. Sve te metode temeljile su se na indeksaciji. Kvalitativni dio analize obuhvatio je pregled podataka prikupljenih putem otvorenih pitanja i polu-strukturiranog intervjua. Za potrebe istraživanja izrađen je posebno prilagođen upitnik. Rezultati istraživanja ukazali su na nekoliko ključnih saznanja. Proces profesionalizacije bio je izraženiji kod sestrinskih profesionalaca starijih od 51 godine s više od 26 godina radnog iskustva, posebno onih zaposlenih na primarnoj razini zdravstvene zaštite. Stjecanje novih znanja doprinijelo je njihovom ljudskom kapitalu i podizanju razine stručnog znanja. Cjeloživotno učenje, autonomija sestrinskih profesionalaca i specifična znanja u sestrinstvu, kao endogeni i egzogeni čimbenici, imali su statistički značajan pozitivan utjecaj na proces profesionalizacije sestrinstva. Međutim, etika u sestrinstvu imala je samo marginalan statistički značajan pozitivan utjecaj na taj proces.

Harsono et al. (2021) istražili su učinak uslužno orijentiranog organizacijskog ponašanja prema građanima (SOCB) na kvalitetu usluge i usporediti SOCB, kvalitetu usluge i ponašanje korisnika kao dobrih građana (CCB) između dvije banke.

U analizi znanstvenih i stručnih članaka, gdje su autori u različitim ekonomskim domenama primjenjivali metode jednostavne i višestruke linearne regresije došlo se do spoznaja da kvantitativni pokazatelji i jednostavne i višestruke linearne regresije doprinose zaključcima o postojanju ili nepostojanju korelacije i statistički značajnih ili neznačajnih razlika između odabranih zavisnih i nezavisnih varijabli. Značaj jednostavne linearne regresije u pojedinim člancima ukazao je na temeljne spoznaje u osnovama korelacije odabranih determiniranih varijabli te njihova utjecaja jedne na drugu. U suprotnom, značaj višestruke linearne regresije u pojedinim člancima ukazao je na opsežne spoznaje i preciziranoj snazi korelacije (slabe ili jake te koliko slabe ili jake), kao i postojanje statistički značajnih ili neznačajnih razlika odabranih determiniranih varijabli, s naglaskom na preciziranje tih razlika.

Svi istraženi znanstveni radovi su proveli empirijska istraživanja koristeći model jednostavne linearne regresije u različitim domenama ekonomije. Isti su primarno u teorijskoj razradi svojih tema članaka prikazivali primarno tematiku koju su obrađivali (primjerice utjecaj organizacijskog ponašanja prema građanima, utjecaj upravljanja ljudskim potencijalima na uspjeh takvih stručnjaka i slično), dok se implementacija

jednostavne ili u pojedinim člancima višestruke linearne regresije provlačila kroz istraživanje kao praktična statistička metoda kojom su oni dokazivali svoje hipoteze. Dakle, u njihovim istraživanjima je primarni naglasak bio na samoj problematici istraživanja i predmetu istraživanja, dok je obrada teorijske osnove jednostavne linearne regresije bila naznačena samo sporadično.

4.2. Jednostavna vs. višestruka linearna regresija – prednosti i nedostaci

Višestruka linearna regresija i jednostavna linearna regresija su dvije osnovne tehnike analize podataka koje se koriste za istraživanje odnosa između varijabli. Oba pristupa imaju svoje prednosti i ograničenja, ali višestruka linearna regresija često nudi veću preciznost i sposobnost modeliranja složenijih odnosa između varijabli u usporedbi s jednostavnom linearnom regresijom.

Tablica 3. Prednosti i nedostaci jednostavne linearne regresije

Prednosti	Nedostaci
• Jednostavna linearna regresija je koristan alat za istraživanje odnosa između dviju varijabli (zavisne i nezavisne)	• Jednostavna linearna regresija zahtijeva veći broj opažanja kako bi pružila pouzdane rezultate. Ako je prisutna mala količina opažanja nedovoljno dobro se objašnjava model.
• Jednostavnost modela, odnosno jednostavnost razumijevanja i primjene modela	• Nedostatak znanja onih koji koriste metodu jednostavne linearne regresije
• Jednostavnost interpretacije modela. Kod jednostavne linearne regresije lakša je interpretacija koeficijenata, budući da su uključene samo dvije varijable u model.	• Jednostavna linearna regresija često nije dovoljna za modeliranje stvarnih situacija gdje je više čimbenika uključeno.

• Jednostavna linearna regresija jednostavno se može prikazati vizualno, odnosno na dvodimenzionalnom grafu što je vrlo važno za stjecanje uvida u podatke.	• Jednostavna linearna regresija, temelji se na određenim pretpostavkama, uključujući pretpostavku o normalnoj distribuciji pogrešaka i homoskedastičnosti. Ako ove pretpostavke nisu ispunjene, rezultati mogu biti nepouzdani. Također, ako reziduali nisu normalno distribuirani, model gubi svoja prognostička svojstva. Ali ako je pretpostavka o homoskedastičnosti narušena, model treba u potpunosti odbaciti. Najveći je problem u tome što se modeli često kreiraju i interpretiraju, ali se preskače provjera pretpostavki.

Tablica 4. Prednosti i nedostaci višestruke linearne regresije

Prednosti	Nedostaci
• Višestruka linearna regresija omogućuje istraživanje odnosa između ovisne varijable i više nezavisnih varijabli istodobno. Ovo je posebno korisno u stvarnim situacijama gdje su učinci više čimbenika istovremeno prisutni.	• Jedan od glavnih nedostataka višestruke linearne regresije je povećana složenost modela. Kako se dodaju dodatne nezavisne varijable, model postaje sve teži za interpretaciju. To može otežati donošenje jasnih zaključaka o učinak pojedinih varijabli na ovisnu varijablu.

• Višestruka linearna regresija omogućuje modeliranje složenijih odnosa između varijabli. Primjerice, može se uzeti u obzir interakcija između nezavisnih varijabli, što omogućuje preciznije predviđanje rezultata.	• Overfitting: Overfitting je pojava kada model previše dobro prilagođava podatke za treniranje, ali ne generalizira dobro na nove, neviđene podatke. Višestruka regresija ima veću tendenciju prema overfittingu u usporedbi s jednostavnom regresijom, posebno kada se koristi veliki broj nezavisnih varijabli. To može rezultirati lošim prediktivnim sposobnostima.
• U nekim slučajevima, višestruka regresija može smanjiti pristranost procjena. Višestruka regresija može razjasniti doprinos svake nezavisne varijable u odnosu na ostale.	• Potreba za većim uzorcima podataka: Višestruka regresija zahtijeva veći broj podataka kako bi pružila pouzdane rezultate. Ako je prisutna mala količina podataka u odnosu na broj varijabli, može se javiti nedostatak statističke snage za otkrivanje stvarnih učinaka.
• U mnogim istraživanjima postoji potreba za kontroliranjem za druge varijable koje mogu utjecati na odnos između nezavisne i ovisne varijable. Višestruka regresija omogućuje uključivanje kontrolnih varijabli u model kako bi se izolirao učinak varijable od interesa.	• Izbor odgovarajućih nezavisnih varijabli ključan je u višestrukoj regresiji. Loš odabir varijabli može rezultirati nepreciznim modelima i nepouzdanim rezultatima. To zahtijeva pažljivu analizu i poznavanje domenskog područja.

• Bolje prilagodbe podacima: Višestruka regresija često pruža bolje prilagodbe podacima nego jednostavna regresija. To znači da će model bolje odražavati stvarne podatke i bolje predviđati buduće vrijednosti zavisne varijable.	• Kao i kod jednostavne linearne regresije, višestruka regresija temelji se na određenim pretpostavkama, uključujući pretpostavku o normalnoj distribuciji pogrešaka i homoskedastičnosti. Ako ove pretpostavke nisu ispunjene, rezultati mogu biti nepouzdani.
• Višestruka regresija može pomoći u identifikaciji ključnih faktora koji najviše utječu na ovisnu varijablu. Ovaj uvid može biti od velike važnosti u donošenju odluka i oblikovanju strategija.	• Analiza višestruke regresije može biti računalno zahtjevna, posebno kada se radi s velikim brojem varijabli ili uzoraka. Ovo može rezultirati produženim vremenom analize i potrebom za snažnijim računalnim resursima.

Izvor: izrada autorice temeljem analize znanstvenih članaka i dostupne literature statističkih udžbenika

Unatoč ovim prednostima, važno je napomenuti da višestruka linearna regresija ima svoje zahtjeve i izazove, uključujući potrebu za većim uzorcima podataka, pažljiv odabir varijabli i oprezno tumačenje rezultata. Također, modeliranje višestrukom regresijom može postati kompleksno, posebno ako uključuje mnogo varijabli ili interakcija između njih.

Iako višestruka linearna regresija omogućuje modeliranje složenijih odnosa između varijabli i pruža mogućnost uključivanja više faktora u analizu, nosi sa sobom niz izazova i potencijalnih problema. Važno je pažljivo razmotriti potrebu za višestrukom regresijom i provesti analizu s obzirom na specifične karakteristike podataka i ciljeve istraživanja kako bi se donijela informirana odluka o odabiru između jednostavne i višestruke linearne regresije.

Uvjet linearnosti odnosa postoji i kod jednostavne i višestruke linearne regresije te u situacijama u kojima pojave nemaju linearan odnos, potrebno je posegnuti za drugačijim pristupom modeliranju. Slična je situacija i po pitanju pretpostavki. Jednostavna linearna regresija je jednostavna, te kao takva, predstavlja idealan početak za učenje o modelima. No, u praksi postoje broji odnosi pojava koji nisu linearni i neće zadovoljavati pretpostavke. To iziskuje daljnje učenje o različitim metodama i tehnikama objašnjavanja odnosa među pojavama.

4.3. Ograničenja jednostavne u odnosu na višestruku linearnu regresiju

Jednostavna linearna regresija, iako je moćan alat za analizu odnosa između dviju varijabli, ima svoja ograničenja koja su važna za razumijevanje u kontekstu modeliranja odnosa. Jednostavna linearna regresija pretpostavlja linearnu vezu između nezavisne i zavisne varijable. Ovo ograničenje znači da će model biti neprecizan ako stvarni odnos između varijabli nije linearan. Složeniji odnosi zahtijevaju upotrebu drugih regresijskih tehnika kao što su polinomijalna regresija ili nelinearna regresija.

Ova regresija analizira samo odnos između dviju varijabli. To znači da se u analizu može uključiti samo jedna nezavisna varijabla. Ako istraživanje uključuje više faktora ili varijabli koje utječu na ovisnu varijablu, tada će jednostavna regresija biti nedostatna, a potrebno je koristiti višestruku regresiju. Izbor odgovarajuće nezavisne varijable ključan je u jednostavnoj linearnoj regresiji. Loš odabir varijable može dovesti do nepreciznih rezultata. Ovisnost o izboru varijabli znači da različite varijable mogu rezultirati različitim modelima i interpretacijama.

Učinkovita primjena jednostavne linearne regresije zahtijeva velik uzorak podataka prikupljenih na slučajan način kako bi se osigurala statistička snaga. Mali uzorak može rezultirati nepouzdanim procjenama parametara i niskom preciznošću modela. Jednostavna linearna regresija temelji se na određenim pretpostavkama, uključujući pretpostavku o normalnoj distribuciji pogrešaka i homoskedastičnosti. Ako ove pretpostavke nisu ispunjene, rezultati će biti nepouzdani.

Jednostavna linearna regresija ne može modelirati interakcije između varijabli. Interakcije se javljaju kada utjecaj jedne varijable ovisi o vrijednosti druge varijable, što se često događa u stvarnim situacijama. Za modeliranje interakcija potrebno je koristiti višestruku regresiju. Jednostavna linearna regresija pruža samo ograničenu sliku stvarnosti jer se usredotočuje samo na odnos između dviju varijabli. Stvarnost je često složenija s mnogo faktora koji utječu na ovisnu varijablu.

Jednostavna linearna regresija je koristan alat za istraživanje odnosa između dviju varijabli, ali ima svoja ograničenja. Pri odabiru regresijskog modela važno je pažljivo razmotriti prirodu podataka, broj varijabli i kompleksnost odnosa između njih kako bi se odabrao odgovarajući statistički pristup koji će odgovarati istraživačkim pitanjima i ciljevima.

Višestruka linearna regresija je moćan statistički alat koji omogućava analizu odnosa između zavisne varijable i više nezavisnih varijabli. Međutim, kao i svaka statistička metoda, višestruka regresija ima svoja ograničenja i izazove u znanstvenim istraživanjima. Nekoliko ključnih ograničenja višestruke linearne regresije: Multikolinearnost se pojavljuje kada su nezavisne varijable u modelu visoko korelirane međusobno. Ovo ograničava sposobnost modela da razdvoji učinak svake varijable na ovisnu varijablu. Multikolinearnost otežava tumačenje koja varijabla ima značajniji utjecaj i može dovesti do nepreciznih procjena koeficijenata.

Višestruka regresija zahtijeva veći broj podataka kako bi dala pouzdane rezultate. Ako imate mali uzorak u odnosu na broj nezavisnih varijabli, može se javiti nedostatak statističke snage za otkrivanje stvarnih učinaka. Ovo je posebno izazovno u kliničkim istraživanjima ili studijama s ograničenim budžetima. Overfitting je pojava kada model previše dobro prilagođava podatke za treniranje, ali ne generalizira dobro na nove, neviđene podatke. U višestrukoj regresiji, dodavanje previše nezavisnih varijabli može povećati sklonost overfittingu, što rezultira modelom koji je previše složen za stvarne podatke.

Izbor odgovarajućih nezavisnih varijabli ključan je u višestrukoj regresiji. Loš odabir varijabli može rezultirati nepreciznim modelima i nepouzdanim rezultatima. Postoji potreba za pažljivim odabirom varijabli kako bi se eliminirale one koje nemaju stvarni utjecaj. Višestruka regresija temelji se na pretpostavkama kao što su normalna distribucija pogrešaka, homoskedastičnost i neovisnost pogrešaka. Ako ove

pretpostavke nisu ispunjene, rezultati mogu biti nepouzdani. Također, pretpostavke o neovisnosti varijabli često nisu ispunjene u stvarnim podacima.

Višestruka regresija može rezultirati složenim modelima s mnogo koeficijenata. Interpretacija ovih koeficijenata može biti izazovna, posebno kada se koristi veliki broj nezavisnih varijabli. Ovo može otežati donošenje jasnih zaključaka o utjecaju varijabli na ovisnu varijablu. Višestruka regresija pretpostavlja linearni odnos između varijabli. Ako stvarni odnosi nisu linearni, model može biti neprecizan i nepouzdan. Za analizu nelinearnih odnosa potrebno je koristiti druge statističke tehnike kao što su nelinearna regresija ili generalizirani linearni modeli.

Višestruka linearna regresija je koristan alat u znanstvenim istraživanjima, ali nosi sa sobom niz ograničenja koja istraživači trebaju uzeti u obzir. Važno je pažljivo razmotriti prirodu podataka, broj varijabli i pretpostavke modela kako bi se donijela pravilna odluka o primjeni višestruke regresije u istraživanju.

U budućim analizama, preporuka je svih obrađenih autora da se u statističkoj analitici bilo koje ekonomske teme primijeni izgradnja složenijih (multivarijantnih) regresijskih modela kako bi se pružila veća eksplanatorna snaga i postigla veća statistička pouzdanost pri predviđanju ishoda utjecaja nezavisnih varijabli na zavisne.

5. Zaključak

Jednostavna linearna regresija omogućuje analizu i kvantifikaciju povezanosti između dvije varijable. To je posebno korisno za istraživače i analitičare koji žele razumjeti kako dvije varijable međusobno utječu. Ovaj tip regresije omogućuje predviđanje vrijednosti zavisne varijable na temelju vrijednosti nezavisne varijable. To je korisno u mnogim područjima, uključujući financije (predviđanje cijena dionica), ekonomiju (predviđanje potrošačkog ponašanja), medicinu (predviđanje ishoda liječenja) i druge.

Jednostavna linearna regresija omogućuje ocjenjivanje koliko neovisna varijabla utječe na ovisnu varijablu. To je posebno važno u situacijama kada se želi razumjeti važnost određenih faktora ili varijabli. Analiza regresije omogućuje organizacijama i pojedincima donošenje pravilnih odluka. Na primjer, tvrtka može koristiti regresijski model za planiranje budućih zaliha ili prilagodbu marketinške strategije na temelju rezultata. Regresija može pomoći u identifikaciji outliersa (ekstremnih vrijednosti) u podacima, što može ukazivati na potrebu dodatnih istraživanja ili ispravaka u podacima.

Implementacija jednostavne linearne regresije predstavlja koristan alat u analizi statističkih podataka i pruža brojne prednosti, ali isto tako nosi i određena ograničenja koja je važno uzeti u obzir prilikom interpretacije rezultata istraživanja.

Koncept jednostavne linearne regresije relativno je jednostavan za razumjeti i primijeniti, što ga čini pristupačnim istraživačima različitih profila. Regresijska analiza omogućuje identifikaciju i kvantifikaciju linearnog odnosa između dviju varijabli, što pomaže u razumijevanju prirode njihove povezanosti.

Model jednostavne linearne regresije može se koristiti za predviđanje vrijednosti zavisne varijable na temelju vrijednosti nezavisne varijable. Ovo je korisno u mnogim situacijama, poput predviđanja budućih prodaja na temelju prošlih podataka. Regresijska analiza omogućuje procjenu utjecaja nezavisne varijable na zavisnu varijablu, što pomaže u ocjeni važnosti tih varijabli u kontekstu istraživanja. Grafički prikazi regresijskog modela, kao što su regresijske linije i dijagrami raspršenosti, vizualno prikazuju odnos između varijabli.

U odnosu na prednosti implementacije jednostavne linearne regresije u statističkim istraživanjima, postoje i njezina ograničenja. Model jednostavne linearne regresije

pretpostavlja linearni odnos između varijabli, što može biti ograničavajuće u situacijama kada odnos nije linearan. Regresijski modeli osjetljivi su na outliers, odnosno ekstremne vrijednosti, koje mogu značajno utjecati na rezultate analize. Jednostavna linearna regresija radi samo s dvije varijable - nezavisnom i zavisnom. U složenijim analizama može biti potrebno razmotriti više faktora. Korištenje linearne regresije zahtijeva zadovoljenje određenih pretpostavki, uključujući normalnu distribuciju pogreške i homoskedastičnost. Važno je razumjeti da regresija može pokazati korelaciju između varijabli, ali ne nužno i uzročni odnos.

Kako bi se ispravno primijenila jednostavna linearna regresija, važno je pažljivo razmotriti svoje istraživačke ciljeve, podatke i pretpostavke modela te provesti odgovarajuće analize kako bi se dobili pouzdani rezultati. U nekim slučajevima, složeniji modeli, kao što su višestruke regresije, mogu bolje odražavati kompleksne odnose između varijabli.

6. Popis tablica

7. Popis grafova

8. Popis slika

9. Literatura

Anon., n.d. *OpenStax.* [Mrežno] Available at: https://openstax.org/books/introductory-business-statistics/pages/13-introduction

Antić, M., 2004. Utjecaj političkog sistema na porast životnog vijeka. *Revija za sociologiju,* 35(3-4).

Belullo, A., 2011. *Uvod u ekonometriju.* s.l.:Pula: Sveučilište Jurja Dobrile u Puli, Odjel za ekonomiju i turizam "Dr. Mijo Mirković".

Bonner, A., n.d. *Simple linear regression in four lines of code.* [Mrežno] Available at: https://contentsimplicity.com/machine-learning-simple-linear-regression/

Dalpiaz, D., n.d. *Applied Statistics with R.* [Mrežno] Available at: https://book.stat420.org/index.html [Pokušaj pristupa 2023].

Đorđe Dobrota, B. L. M. O., 2010.. EKSPERIMENTALNO MODELIRANJE VOLUMETRIJSKE KORISNOSTI VISOKOTLAČNE ZUPČASTE PUMPE S VANJSKIM OZUBLJENJEM. *NAŠE MORE : znanstveni časopis za more i pomorstvo,* 57(5-6).

Eleonora Kontuš, N. Š., 2019.. UTJECAJ STRUKTURE KAPITALA I TEORIJE HIJERARHIJE FINANCIJSKIH IZBORA NA LIKVIDNOST DIONIČKIH DRUŠTAVA. *Ekonomska misao i praksa,* 28(2).

Eleonora Kontuš, N. Š., 2019.. UTJECAJ STRUKTURE KAPITALA I TEORIJE HIJERARHIJE FINANCIJSKIH IZBORA NA LIKVIDNOST DIONIČKIH DRUŠTAVA. *Ekonomska misao i praksa,* 28(2).

Gulcemal, T., 2020.. Effect of human development index on GDP for developing countries: a panel dataanaysis. *Journal of Economics, Finance and Accounting (JEFA),* 7(4), pp. 338.-345..

Guzić, Š., 2014.. Nematerijalna imovina kao bitan element uspješnosti poslovanja trgovačkih društava. *Journal of Accounting and Management,* IV(1).

Jasna Horvat, J. M., 2014. . *Osnove statistike.* drugo dopunjno izdanje ur. Zagreb: an.

Kontuš, E., 2021.. UTJECAJ STRUKTURE KAPITALA NA PROFITABILNOST POSLOVANJA: ANALIZA SLUČAJA HRVATSKIH, SLOVENSKIH I ČEŠKIH DIONIČKIH DRUŠTAVA. *Poslovna izvrsnost,* 15(1).

Microsoft (2023): Prikazivanje podataka na raspršenom ili linijskom grafikonu , dostupno na https://support.microsoft.com/hr-hr/topic/prikazivanje-podataka-na-raspr%C5%A1enom-ili-linijskom-grafikonu-4570a80f-599a-4d6b-a155-104a9018b86e , pristupljeno 20.09.2023.

Pejić Bech, M. D. M., 2008. Otkrivanje porezne utaje iz informacijskog sustava porezne uprave metodom otkrivanja znanja iz baza podataka. *Zbornik Ekonomskog fakulteta u Zagrebu,* 6(1).

Raos, V., 2017.. Lokalni izbori kao međuizbori?. *Političke analize : tromjesečnik za hrvatsku i međunarodnu politiku,* 8(30).

Schaalje, A. C. R. a. G. B., n.d. *LINEAR MODELS IN STATISTICS.* [Mrežno] Available at: https://www.utstat.toronto.edu/~brunner/books/LinearModelsInStatistics.pdf

Schönbrodt, F. D. &. P. M., 2013.. t what sample size do correlations stabilize?. *Journal of Research in Personality,* 47(5), pp. 609-612.

Soni Harsono, H. W. T. P. B. R., 2021.. Uslužno orijentirano organizacijsko ponašanje prema građanima, kvaliteta usluge i ponašanje korisnika kao dobrih građana: usporedba primjene i procjena iz perspektive korisnika banke. *Market-Tržište,* 33(1).

Starc, A., 2016.. Profesionalizacija u lancu zdravstvene zaštite. *Journal of Applied Health Sciences = Časopis za primijenjene zdravstvene znanosti,* 2(2).

Šućur, Z., 2021.. Dohodovne nejednakosti i redistributivne preferencije u Hrvatskoj i zemljama EU-a: makroanaliza. *Revija za socijalnu politiku,* 28(2).