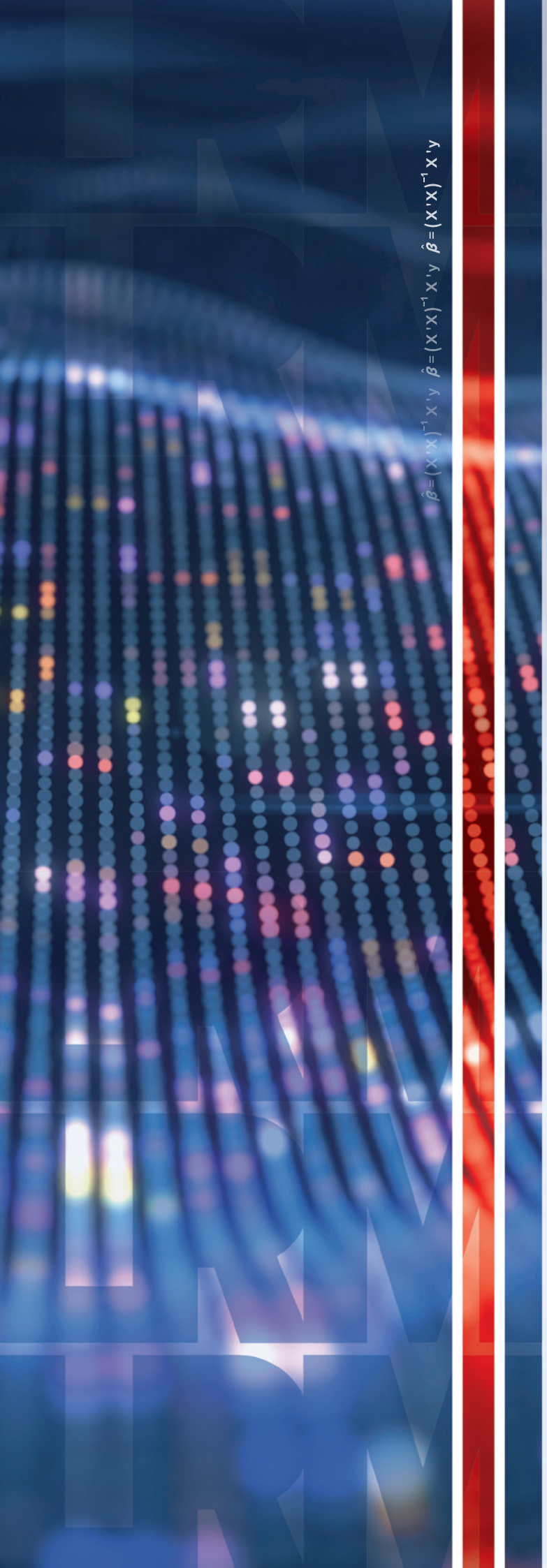


HNB
EUROSUSTAV

ODABRANE TEME PRIMIJENJENE EKONOMETRIJE

Linearni regresijski model

Tihana Škrinjarić

Zagreb, 2023.

HRVATSKA NARODNA BANKA
EUROSUSTAV

HNB
EUROSUSTAV

ODABRANE TEME
PRIMIJENJENE
EKONOMETRIJE

Linearni regresijski model

Tihana Škrinjarić

Zagreb, 2023.

HRVATSKA NARODNA BANKA
EUROSUSTAV

Komentare, prijedloge i uočene greške u kodovima molimo javiti na publikacije@hnb.hr

Hvala!

Predgovor urednika

"Odabrane teme primijenjene ekonometrije: linearni regresijski model" dio je serije edukativnih publikacija Hrvatske narodne banke. Riječ je o praktičnom pregledu osnovnih metoda ekonometrijske analize, koji kombinira istraživačke probleme s ekonometrijskom teorijom, a napisan je tako da čitatelja vodi korak po korak do rješenja.

"Odabrane teme" praktičan su uvod za početnike koji se tek spremaju otisnuti u empirijska istraživanja, kao što su polaznici ekonomskih ili drugih, uglavnom društvenih studija, koji primjenjuju metode ekonometrijske analize. Uz tu primarnu ciljnu skupinu, knjiga će dobro poslužiti kao konzultativni priručnik za iskusnije analitičare koji žele osvježiti znanje kako bi riješili novi konkretan problem s kojim se susreću, kao i za sve znatiželjnike koji će publikaciju samo prelistati kako bi dobili opći pregled izabranih tehnika ekonometrijske analize. Nadamo se da će knjiga doprijeti i do akademskih korisnika koji nisu ekonomisti, npr. iz STEM područja i zaintrigirati ih za teme koje istražuju ekonomisti.

Edukativne publikacije na hrvatskom jeziku koje obrađuju ekonometrijsku analizu, uključujući konkretne primjere analize domaćih makroekonomskih ili mikroekonomskih podataka kao i popratnu podršku u obliku računalnih kodova i riješenih problema nisu široko dostupne. Nadamo se da će stavljanje ove publikacije na slobodno raspolaganje istraživačima, ekonomistima, studentima i drugoj zainteresiranoj javnosti unaprijediti razumijevanje metoda kojima se služe ekonomisti i potaknuti buduće zanimanje za ekonomska istraživanja.

Predgovor recenzenta

Rukopis autorice dr. sc. Tihane Škrinjarić ambiciozno je djelo relevantne tematike koja je neophodna vrlo širokom krugu ekonomskih analitičara. Riječ je o knjizi koja na sustavan i pedagoški korektan način obrađuje tematiku osnovne ekonometrijske analize, te knjiga samim time može biti vrlo korisna studentima ekonomskih usmjerenja, istraživačima i središnjim bankarima, ali i praktičarima, poglavito iz financijskog sektora, čiji posao zahtijeva kvantitativnu analizu te preskriptivnu ili prediktivnu analitiku.

Autorica na minuciozan način uvodi čitatelja u složenu ekonometrijsku tematiku, sustavno obrađujući kompleksne metodološke pristupe na način da svaki sljedeći koncept logički nadograđuje onaj prethodni te time čitatelju pruža kvalitetan uvid u modernu statističku analizu ekonomskih podataka.

Iako knjiga u određenim dijelovima zahtijeva određeno predznanje matematike i statistike, autorica mnoštvom primjera dobro ilustrira praktičnu primjenu obrađenih metoda, pravilnu interpretaciju rezultata te se objektivno ističu i prednosti i nedostaci svake razmatrane ekonometrijske metode, što često nije slučaj kod knjiga ovog tipa.

Poseban je doprinos knjige to što autorica primjenu razmatranih metoda i tehnika ilustrira upotrebom programske podrške slobodnog dohvata R Studio. Riječ je o programu koji je besplatno dostupan svakom potencijalnom čitatelju knjige, ne zahtijeva plaćanje često vrlo skupih licencija, a vrlo je pouzdan i naširoko korišten i u poslovnom svijetu i u akademskoj zajednici. Samim time, čitateljima knjige pruža se dodatna vrijednost.

izv. prof. dr. sc. Petar Sorić

Predgovor autorice

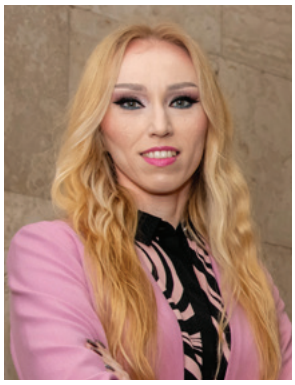
“Odabrane teme primijenjene ekonometrije: linearni regresijski model” nastale su kao rezultat mojega pedagoškog rada na Ekonomskom fakultetu u Zagrebu i istraživačkog rada u Hrvatskoj narodnoj banci. U radu sa studentima uočila sam potrebu za praktičnim vodičem kroz dinamično i stalno rastuće područje ekonometrijske analize, koji će ih na sustavan i dostupan način uvesti u materiju. Teorija, koju sam povezala s primjerima, na svrsishodan i dosljedan način utemeljuje primjere u široj ekonomskoj znanosti, a materijal je pripremljen na način da svatko zainteresiran može samostalno savladati osnove ekonometrijske analize.

“Odabrane teme” podijelila sam u pet cjelina. **Uvod u ekonometriju** definira terminologiju i vrste podataka, što prethodi svakom modeliranju. **Linearni regresijski model** i **Daljnja analiza regresijskog modela** donose procjenu linearnog modela, s interpretacijama koje slijede iz te procjene, testiranjem hipoteza i razmatranjem kvalitativnih regresorskih varijabli. Ova dva poglavlja pojašnjavaju osnovne metode procjene regresijskih parametara, uz izvode brojnih mjera koje se razmatraju pri toj procjeni.

Narušavanje pretpostavki regresijskog modela približava modele stvarnom životu i uobičajeno slijedi nakon osnovnih metoda u ekonometrijskim udžbenicima, tako da i ja iznosim građu tim redom. Posljednja cjelina, **Dodaci**, obuhvaća osnove matematike i statistike koje je potrebno znati kako bi se ostatak teksta lakše pratio. Uz materijale pripremila sam podatke i R-skripte dostupnu na internetskim stranicama HNB-a, kako bi čitatelj lakše pratio materiju i usporedno sâm krenuo u modeliranje.

Nadam se da će čitatelji u “Odabranim temama” pronaći praktične alate za vlastiti rad i poticaj za daljnja istraživanja, ili neki novi detalj o već poznatim stvarima. Zahvaljujem recenzentu na korisnim sugestijama koje su pridonijele kvaliteti konačnog materijala, kao i obitelji i bližnjima koji su u nekim trenucima bili zapostavljeni. Na kraju, zahvaljujem uredništvu HNB-a koje je uvidjelo korist ovakvog materijala i prihvatilo ga objaviti. Preostale pogreške moj su propust i pozivam čitatelje da me o njima obavijeste kako bih iduće izdanje unaprijedila.

Kodovi dostupni na:
https://www.hnb.hr/analize-i-publikacije/ostale-publikacije



O autorici

Tihana Škrinjarić savjetnica je u Direkciji za makrobonitetnu politiku i financijsku stabilnost u Hrvatskoj narodnoj banci. Bavi se temama povezanima s kvantitativnom podrškom u makrobonitetnoj politici, analizom cikličkih rizika, ocjenom karaktera makrobonitetne politike i kalibracijom njezinih instrumenata. Prije dolaska u HNB 2021. godine, Tihana je bila docentica na Ekonomskom fakultetu u Zagrebu na Katedri za matematiku, gdje je držala kvantitativne kolegije poput ekonometrije i matematike. Doktorat iz ekonomije obranila je 2018. godine, na Ekonomskom fakultetu – Zagreb, s temom ekonometrijskih modela promjene režima u menadžmentu portfelja. U razdoblju rada na fakultetu Tihanini interesi istraživanja bili su financijska tržišta i financijska ekonometrija te primjena kvantitativnih modela i metoda u području financija. Tihana ima više od 150 objavljenih radova, poglavlja u knjigama i istraživanja iz područja financija i makrobonitetne politike, bila je recenzentica nekoliko desetaka časopisa te je dobitnica više nagrada za znanstvene članke iz tih područja. Krajem 2022. godine, Tihana je rangirana u top 12,5% europskih autora iz područja ekonomije u bazi Ideas RePec.

Link na Google Scholar:
https://scholar.google.com/citations?user=C5yapgkAAAAJ&hl=hr&oi=ao

SADRŽAJ

OZNAKE

$\boldsymbol{x}$ – vektor stupac
$\boldsymbol{x}'$ – vektor redak, radi se o transponatu vektora stupca $\boldsymbol{x}$
x – varijabla
$\boldsymbol{X}$ – matrica
$E(\cdot)$ – operator očekivanja. Čitam: „očekivanje od…“
$E(y \mid x)$ – operator uvjetnog očekivanja. Čitam: „očekivana vrijednost y-a s obzirom na dani x“
$P(\cdot)$ – vjerojatnost nastupa nekog događaja. Čitam: „vjerojatnost nastupa događaja … je „
$P(y \mid x)$ – uvjetna vjerojatnost nastupa nekog događaja. Čitam: „vjerojatnost da nastupi y uz dane vrijednosti x iznosi… “
$\frac{dy(x)}{dx}$ – prva derivacija funkcije y, slučaj jedne nezavisne varijable
$\frac{\partial y(x_1,...,x_i,...,x_n)}{\partial x_i}$ – parcijalna derivacija prvog reda funkcije y po varijabli x_i, slučaj više nezavisnih varijabli
0 – broj nula
$\mathbf{0}$ – nul-vektor
I – jedinična matrica
y_i – i-to opažanje varijable y, empirijska vrijednost
$\hat{y}_i$ – i-ta procijenjena vrijednost varijable y
$\exists$ – postoji, npr. $\exists\, x$ – čitam „postoji (barem jedna) varijabla x“
$\neq$ – različito je
$\sim$ – slijedi, npr. $x \sim N$ – čitam „varijabla x slijedi normalnu distribuciju“
$\rightarrow$ – teži prema, npr. $x \rightarrow y$ – čitam „vrijednost x teži prema vrijednosti y“
Δ – delta, prva diferencija, promjena varijable, $\Delta y_t = y_t - y_{t-1}$
Δ^2 – druga diferencija, promjena promjene varijable,
$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = y_t - y_{t-1} - (y_{t-1} - y_{t-2})$

1.

TEMELJNI POJMOVI REGRESIJSKOG MODELA

1. TEMELJNI POJMOVI REGRESIJSKOG MODELA

1.1. Definiranje ekonometrije

Ekonometrija[1] (engl. *econometrics*) se može definirati kao društvena, interdisciplinarna znanost koja se koristi ekonomskom teorijom, matematikom i statistikom u svrhu analize ekonomskih fenomena (Golberger, 1964), sastoji se od primjene matematičke statistike nad ekonomskim podacima kako bi se empirijski poduprli modeli konstruirani u matematičkoj ekonomiji te kako bi se ostvarili numerički rezultati (Samuelson, Koopmans i Stone, 1954). Sam naziv ekonometrija odnosi se na „mjerenje u ekonomiji" (Brooks, 2014). Dakle, potrebna su određena znanja iz matematike i statistike, kao i iz ekonomske teorije, kako bi se analizirali empirijski, stvarni podaci i potkrijepila ili opovrgnula ekonomska teorija. Zato se i može reći da je ekonometrija **interdisciplinarna znanost** jer iziskuje znanja iz različitih područja, uključujući i računalne znanosti, s obzirom da danas nije moguće brojne analize raditi bez kvalitetne računalne podrške te poznavanja računalnih programa koji omogućavaju samu analizu. Maddala i Lahiri (2009:3) definiraju ekonometriju kao primjenu statističkih i matematičkih metoda za analizu ekonomskih podataka, s ciljem davanja empirijskog sadržaja ekonomskim teorijama, te njihovim potvrđivanjem ili opovrgavanjem. Za daljnju raspravu o definicijama ekonometrije, vidjeti Tintner (1953). Postoje dvije vrste ekonometrije: teorijska i aplikativna (primijenjena). Ako se usmjerimo na primjenu ekonometrije unutar ekonomskih znanosti, ona se može koristiti u mikro i makro analizama, u ovisnosti o podacima s kojima istraživač barata. Zato postoje različite grane ekonometrije koje analiziraju se specifičnosti podataka, poput makro-ekonometrije, mikro-ekonometrije, financijske ekonometrije, ekonometrije velikih podataka (engl. *big data econometrics*), itd.

Brojni su razlozi zašto izučavati ekonometriju, posebice u ekonomiji i poslovanju. Naime, za prognoziranje bilo koje makroekonomske varijable, ili pak potražnje i prodaje u poslovanju, nužno je koristiti ekonometriju. S jedne strane je moguće analizirati na koji način i koliko jedna pojava utječe na drugu i uključuje određeno vrednovanje (preskriptivna analiza), a s druge je moguće raditi isključivo prediktivnu analizu, u svrhe predviđanja budućih vrijednosti neke pojave, ili pak vrijednosti koje bi vrijedile ako se rezultati modela primjene na karakteristike nekih pojava. Proučavanjem veza između određenih varijabli od interesa u prošlosti može dati istraživaču dobre spoznaje o njihovim budućim kretanjima. Na taj način se samo poslovanje može bolje prilagoditi uvjetima na tržištu, kako bi poduzeće bilo konkurentnije i ostvarivalo bolje poslovne rezultate. Ili pak s druge strane, nositelji ekonomskih politika mogu bolje prilagoditi mjere ekonomske politike kojima se usmjerava buduće kretanje varijabli od interesa, poput BDP-a (bruto domaći proizvod), kamatnih stopa, nezaposlenosti i drugih značajnih varijabli.

Razlika između ekonometrije i srodnih disciplina je sljedeća. **Matematička ekonomija** proučava **deterministčke** modele, koji simbolima, formulama, jednadžbama i modelima predočavaju veze između ekonomskih varijabli. Ekonometrija s druge strane te modele pokušava testirati koristeći empirijske podatke i ekonometrijske metode, pri čemu je uz deterministički dio uključen i **stohastički, slučajni dio**. Primjerice, mogli bismo zapisati sljedeći model kao deterministički unutar matematičke ekonomije:

[1] Riječ ekonometrija prvi puta upotrebljava P. Ciompa 1910. godine, iako u današnjem smislu te riječi prvi ju definira R. Frisch 1936. godine (Pesaran, 1987), a odnosi se na spajanje riječi ekono(mija) i metrija. O povijesnom pregledu ekonometrije, vidjeti uz Pesaran (1987) i Ruggins (2015).

$$y = f(x), \tag{1.1}$$

što znači da je varijabla y poznata funkcija od varijable x. Dakle, kao u matematici, funkcija u relaciji (1.1) je deterministička, jer za određenu vrijednost varijable x koju bismo uvrstili u samu funkciju f, jednoznačno bismo odredili vrijednost varijable y. U ekonometriji se uz deterministički dio modela uključuje i stohastički dio. Stoga se radi o sljedećem modelu:

$$y = f(x) + \varepsilon, \tag{1.2}$$

gdje ε predstavlja stohastičku (slučajnu) komponentu. Pritom se model u (1.2) naziva aditivni model jer je slučajna komponenta dodana. Model može biti i multiplikativni, ako vrijedi:

$$y = f(x) \cdot \varepsilon, \tag{1.3}$$

dok sam funkcionalni oblik $f(\cdot)$ može biti bilo kakva funkcija, u ovisnosti o naravi podataka ili pak ekonomskoj teoriji koja može pretpostavljati određeni funkcionalni oblik. Nadalje, razlika između ekonometrije i **ekonomske statistike** je ta što se potonja bavi samo prikupljanjem, obradom i prikazom ekonomskih podataka u obliku dijagrama, tablica i sličnog ispisa, bez daljnjeg testiranja ekonomske teorije (Gujarati i Porter, 2010).

1.2. Metodologija ekonometrije

S obzirom na samu definiciju ekonometrije, potrebno je slijediti određen niz koraka koji čine suštinu same ekonometrije kao znanstvene discipline, ali i alata za primjenu u praksi. Kako se radi o empirijskom testiranju ekonomske teorije, kao **prvi korak** istraživač postavlja pitanje, hipotezu ili pak želi testirati određeni teorijski model. Nakon toga slijedi **drugi korak**: prikupljanje podataka. U **trećem koraku** zapisuje se matematički model koji opisuje odnose između varijabli. **Model** je pojednostavljen prikaz procesa koji se događaju u stvarnome svijetu. Može se sastojati od jedne ili više jednadžbi/nejednadžbi, koje opisuju odnose između varijabli. **Varijablama** nazivamo entitete koje razmatramo unutar ekonomskih modela, dok prikupljene podatke o tim varijablama nazivamo **opažanjima** (opservacijama). Primjerice, možemo razmatrati ekonomski model u kojemu funkcija potražnje q, mjerena potraživanom količinom, ovisi o cijeni proizvoda p. U tom slučaju model bismo zapisali na sljedeći način:

$$q(p) = \beta_0 + \beta_1 p, \tag{1.4}$$

gdje $q(p)$ predstavlja potraživanu količinu u ovisnosti o cijeni p, β_0 i β_1 predstavljaju parametre modela, pri čemu se pretpostavlja linearna veza između cijene i količine. **Parametre** modela bismo mogli nazvati konfiguracijskim varijablama ili karakteristikama unutar modela, jer oni određuju same karakteristike pojedinih varijabli ili njihovih međuodnosa. Varijabla koja se nalazi s lijeve strane jednakosti naziva se zavisna ili endogena, dok se varijable s desne strane jednakosti u modelu nazivaju nezavisnima ili egzogenima (vidjeti detalje u 2.1.1).

Funkcionalni oblik u (1.4) je linearna funkcija, gdje je β_0 odsječak na y-osi, a β_1 koeficijent smjera. Pritom se pretpostavlja da vrijedi: $\beta_0 \geq 0$ i $\beta_1 < 0$. Prva pretpostavka koja se odnosi na odsječak (konstantu) interpretira se na način da potraživana količina za nekim proizvodom kada je njegova cijena jednaka nuli iznosi 0 ili više (potraživana količina ne može biti negativna). Druga pretpostavka, $\beta_1 < 0$, odnosi se na pretpostavku da se radi o normalnom dobru, tj. povećanje cijene nekog proizvoda vodi smanjenju potražnje za tim proizvodom. Varijable u modelu su cijena i količina, dok bi prikupljeni podaci o cijeni i količini, primjerice kroz određen

broj dana ili mjeseci bila opažanja o te dvije varijable, a već je spomenuto da su β_0 i β_1 parametri modela jer određuju odnos između cijene i količine.
Dodatno, model u (1.4) nazivamo deterministčkim s obzirom na prethodnu diskusiju. Kako potraživana količina ovisi i o drugim varijablama, čija opažanja često možemo prikupiti, ali u nekim slučajevima i ne, sama ideja ekonometrijskih modela je „uloviti“ ponašanje u prosjeku, kako bi se dobila globalna slika o ponašanju i međuodnosima između varijabli, model (1.4) ćemo zapisati u stohastičkom obliku, gdje se dodaje slučajna komponenta ε:

$$q(p) = \beta_0 + \beta_1 p + \varepsilon. \qquad (1.5)$$

Nakon što se odredi model u trećem koraku, u **četvrtom** se procjenjuju nepoznati parametri nekom ekonometrijskom metodom. **Metoda** je način procjene nepoznatih parametara u modelu, gdje se koriste matematički i statistički alati prilagođeni vrsti podataka i pretpostavkama modela koji se razmatra. Postoje brojne ekonometrijske metode koje su upravo prilagođene specifičnostima određenih podataka kojima istraživač barata. U **petome koraku** ispituje se dijagnostika modela, njegova prikladnost za prikupljene podatke, testiraju se pretpostavke modela. Na taj način se dobiva informacija je li model dobro opisao stvarnost, koliko dobro i može li se dalje koristiti. U **šestome koraku** testiraju se hipoteze modela, vezane uz ekonomsku teoriju, dok se u **sedmome koraku** model koristi u prognostičke svrhe.

Dakle, osnovna je ideja da se temeljem određene teorije definira model koji se želi testirati, potrebno ga je zapisati u matematičkom obliku, potom prikupiti podatke o varijablama modela i nekom metodom procijeniti parametre modela. Kako bismo znali je li model dobra reprezentacija stvarnosti, potrebno je provjeriti dijagnostiku modela. Ako je model zadovoljavajući, testiraju se hipoteze koje istraživač postavlja, te se model koristi u prognostičke svrhe.

Primjer 1.1.
Razmatra se ekonomski model Keynesove funkcije potrošnje: $C(Y) = a + bY_d$, gdje su varijable modela C, koja predstavlja potrošnju, a Y_d raspoloživi dohodak. a i b su parametri modela, gdje a predstavlja autonomnu potrošnju, koja ne ovisi o raspoloživom dohotku, pri čemu je $a > 0$, b predstavlja graničnu sklonost potrošnji, za koju se u teoriji pretpostavlja da je $0 < b < 1$. Stohastički oblik modela je $C(Y) = a + bY_d + \varepsilon$, gdje je potrebno nekom ekonometrijskom metodom procijeniti parametre a i b, kako bi se temeljem stvarnih podataka testiralo vrijede li pretpostavke o tim parametrima.

1.3. Vrste ekonomskih podataka

U prethodnoj sekciji spomenuti su podaci i njihova opažanja. Prilikom prikupljanja podataka, važno je znati o kakvim se podacima radi. Ekonomski podaci mogu se odnositi na vrijednosti varijable u danome trenutku ili vremenskom razdoblju za pojedinačne entitete (ljudi, države, dionice, itd.). U tom slučaju radi se o **presječnim** podacima (engl. *cross section data*). Presječni podaci tako mogu biti BDP (bruto domaći proizvod) u 2020. godini za zemlje Europske unije. U tom slučaju koristi se indeks i prilikom zapisa samoga modela. Uzme li se slučaj prethodno obrađivanog modela potražnje u ovisnosti o cijeni, zapis modela je za svaki entitet sljedeći:

$$q_i(p_i) = \beta_0 + \beta_1 p_i + \varepsilon_i, \qquad (1.6)$$

gdje se uočava da je indeks i pridružen svakoj potražnji i cijeni, kao i slučajnoj komponenti. Većina ekonomskih podataka prikuplja se kroz vrijeme i tako nastaju **vremenski nizovi**. Ovdje

se radi o istome entitetu (čovjek, država, dionica) čija se opažanja bilježe na određene datume i odnose se na više vremenskih trenutaka ili razdoblje. Primjerice, ako se razmatra BDP Hrvatske u razdoblju od 2000. do 2020. godine, pri čemu se prikupljaju kvartalni podaci, radi se o vremenskom nizu. Prethodni model potražnje i cijene za slučaj vremenskih nizova, kada bismo promatrali nekog pojedinca kroz vrijeme, u tome slučaju bismo zapisali:

$$q_t(p_t) = \beta_0 + \beta_1 p_t + \varepsilon_t, \tag{1.7}$$

gdje se sada uočava da se pridodaje indeks t. Dakle, ako ćemo razmatrati presječne podatke, razmatrat ćemo neku (određenu) varijablu za više entiteta i, $i \in \{1, 2, \ldots, N\}$, gdje N predstavlja ukupni broj entiteta, dok za vremenske nizove razmatramo jedan entitet kroz vrijeme, $t \in \{1, 2, \ldots, T\}$, gdje je T posljednji vremenski trenutak ili razdoblje koje se razmatra.

Konačno, ako razmatramo više entiteta kroz više vremenskih razdoblja ili trenutaka, radi se o **panel podacima**. U tom slučaju koristimo indekse i i t. Kao primjer, može se razmatrati BDP svih zemalja članica Europske Unije za razdoblje od 2000. do 2020. godine. Ako bismo razmatrali ponovno model potražnje u ovisnosti o cijeni, tada bismo mogli razmatrati više pojedinaca u različitim gradovima, kroz svaki mjesec pa bismo model zapisali ovako:

$$q_{it}(p_{it}) = \beta_0 + \beta_1 p_{it} + \varepsilon_{it}. \tag{1.8}$$

Primjer 1.2.
Primjer presječnih podataka prikazan je u Tablici 1.1., gdje se nalaze vrijednosti BDP-a za odabrane zemlje u 2019. godini.

Tablica 1.1. BDP odabranih zemalja u 2019. godini, tekuće cijene, u milijunima eura

Godina/Država	Češka	Španjolska	Francuska	Hrvatska	Austrija
2019.	223.945	1.245.331	2.425.708	53.936,7	398.682,4

Izvor: Eurostat (2020)

Primjer 1.3.
Primjer vremenskih nizova prikazan je u Tablici 1.2, gdje se nalaze vrijednosti BDP-a za Hrvatsku, od 2010. do 2020. godine

Tablica 1.2. BDP Hrvatske od 2010. - 2019. godine, tekuće cijene, u milijunima eura

Godina	BDP
2010.	45111,8
2011.	44793
2012.	43940,8
2013.	43703,2
2014.	43401,3
2015.	44616,4
2016.	46615,5
2017.	49094,4
2018.	51625,1
2019.	53936,7

Izvor: Eurostat (2020)

Primjer 1.4.
Primjer panel podataka prikazan je u Tablici 1.3, gdje se nalaze vrijednosti BDP-a za odabrane zemlje, od 2010. do 2020. godine.

Tablica 1.3. BDP odabranih zemalja, 2010. - 2019. godine, tekuće cijene, u milijunima eura

Zemlja	Godina	BDP
Austrija	2010.	295.896,6
Austrija	2011.	310.128,7
Austrija	2012.	318.653,0
Austrija	2013.	323.910,2
Austrija	2014.	333.146,1
Austrija	2015.	344.269,2
Austrija	2016.	357.299,7
Austrija	2017.	370.295,8
Austrija	2018.	385.711,9
Austrija	2019.	398.682,4
Hrvatska	2010.	45.111,8
Hrvatska	2011.	44.793,0
Hrvatska	2012.	43.940,8
Hrvatska	2013.	43.703,2
Hrvatska	2014.	43.401,3
Hrvatska	2015.	44.616,4
Hrvatska	2016.	46.615,5
Hrvatska	2017.	49.094,4
Hrvatska	2018.	51.625,1
Hrvatska	2019.	53.936,7
…	…	…
…	…	…
…	…	…
Španjolska	2019.	1.245.331,0

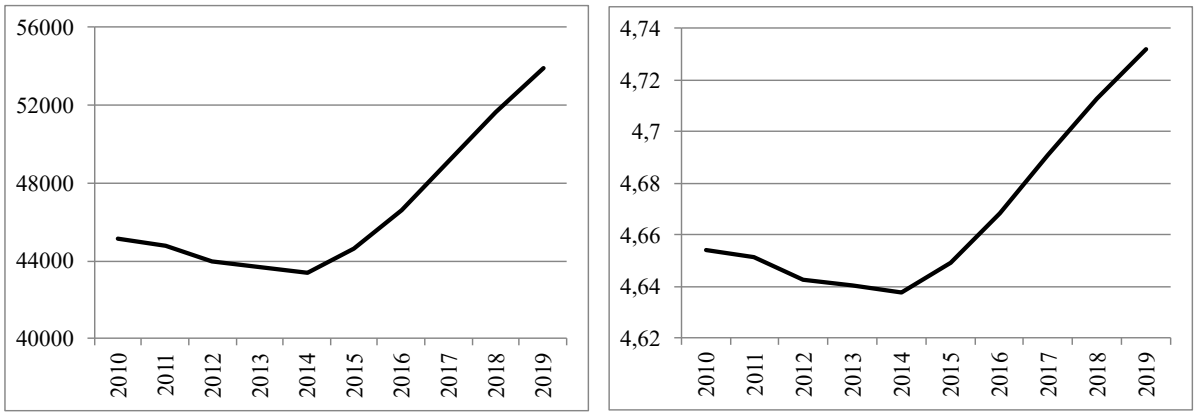
Izvor: Eurostat (2020)

1.4. Priprema podataka za ekonometrijsku analizu i grafičko predočavanje

U praksi se često događa da je prikupljene podatke prije same analize potrebno urediti. Ovdje se misli na tablično uređivanje, na način da programska podrška prepoznaje vrstu podataka, ali i na dodatne transformacije koje se mogu vršiti nad podacima poput logaritmiranja, diferenciranja, desezoniranja, deflacioniranja, itd. Ako ekonomska teorija pretpostavlja vezu između stopa rasta, a podaci su prikupljeni u razinama, potrebno je najprije izračunati stope rasta, te potom vršiti analizu. **Logaritmiranje** podataka je jedna od najčešćih transformacija nad podacima u empirijskim analizama, iz dva razloga. Prvi je što se stope rasta često računaju temeljem logaritmiranih vrijednosti varijabli, a drugi to što ova transformacija smanjuje varijancu novoga niza (važnost ovoga obrađuje se u poglavlju 4.3.). Slika 1.1 uspoređuje BDP Hrvatske u razinama, te logaritmirane vrijednosti, pri čemu se uočava da je dinamika kretanja obiju varijabli jednaka, no smanjena je varijanca logaritamskom transformacijom.

(a) BDP Hrvatske, od 2010. do 2019. godine, tekuće cijene, u milijunima Eura

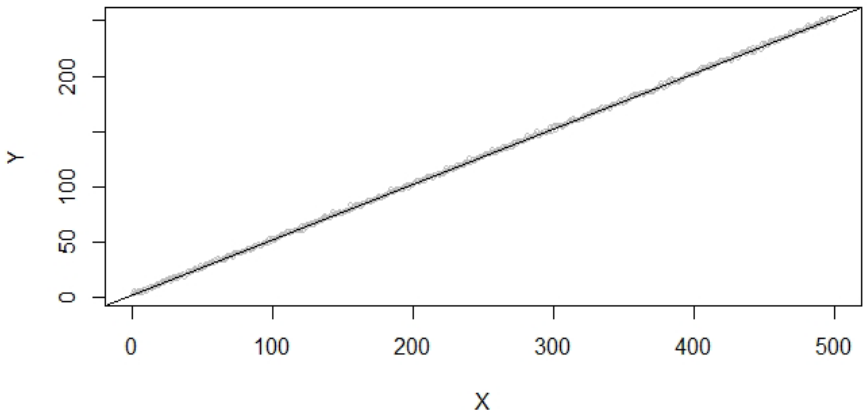
(b) logaritmirane vrijednosti BDP-a Hrvatske



Slika 1.1. Usporedba BDP-a Hrvatske u razinama, te logaritmirane vrijednosti

Ekonomske varijable se često transformiraju u **stope rasta** (o kojima će više riječi biti u drugoj publikaciji[2]). Žele li se razmotriti realne vremenske nizove, a ne nominalne, potrebno je podatke **deflacionirati**. **Desezoniranje** je postupak uklanjanja učinka sezonalnosti u vremenskim nizovima[3].

Prije formalne analize, često je korisno varijable predočiti grafički, posebice pomoću **dijagrama rasipanja**, koji predočava vezu između dvije varijable na način da se u koordinatnom sustavu predočavaju vrijednosti dviju varijabli pomoću točaka. Na taj način može se uočiti **postoji li određena veza** između varijabli ili ne, **jačina** te veze, kao i sam **smjer učinka** jedne varijable na drugu. Slika 1.2 predočava dijagram rasipanja između varijabli x i y , gdje se uočava da postoji veza, ujedno je jaka, pozitivna i linearna. Sive točke predočavaju kombinaciju vrijednosti obiju varijabli, te se može zamisliti da bismo mogli docrtati zamišljeni pravac (predočen punom crnom linijom), koji bi mogao opisivati vezu između x i y .

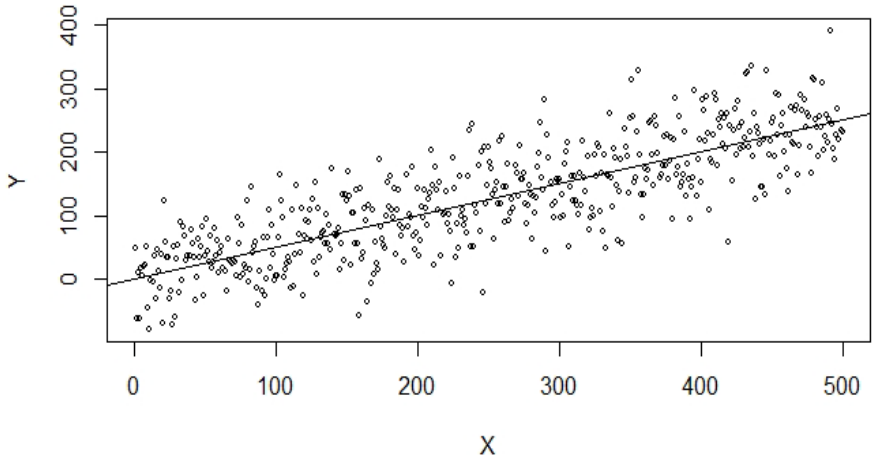


Slika 1.2. Dijagram rasipanja između varijabli x i y , jaka pozitivna linearna veza

S druge strane, u stvarnosti kada prikupimo podatke za varijable x i y , češće će se dogoditi da te točke neće gotovo savršeno pripadati pravcu, već će postojati određena raspršenost između

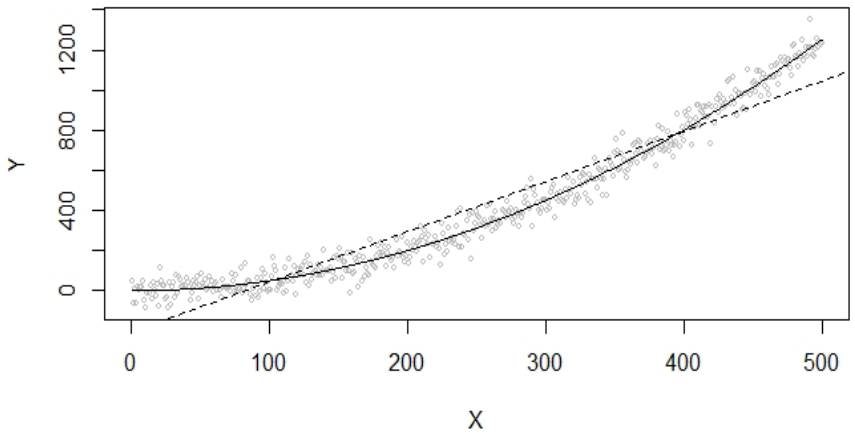
[2] Vidjeti "Odabrane teme primijenjene ekonometrije: Uvod u analizu vremenskih nizova".
[3] Više riječi o deflacioniranju i desezoniranju moguće je naći u "Odabrane teme primijenjene ekonometrije: Uvod u analizu vremenskih nizova".

njih, što je prikazano na slici 1.3. Uočava se i dalje da postoji pozitivna linearna povezanost između x i y, no raspršenost točaka je sada veća. No, i dalje bismo mogli koristiti pretpostavku linearne veze između razmatranih varijabli. Model koji bi se mogao zapisati za slučaj predočen slikom 1.3, bio bi sljedeći: $y = a + bx + \varepsilon$, uz pretpostavku da je $b > 0$ zbog pozitivnog nagiba pravca.



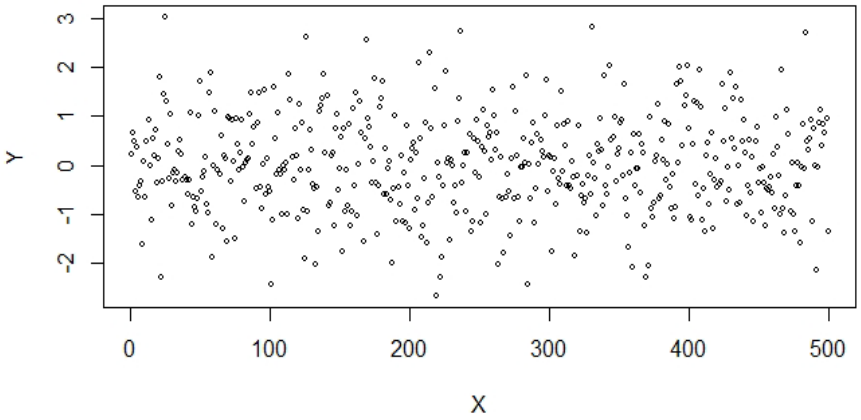
Slika 1.3. Dijagram rasipanja između varijabli x i y , umjerena pozitivna linearna veza

Nadalje, dijagram rasipanja može prikazivati i druge oblike povezanosti. Tako umjesto linearne veze, možemo razmatrati i druge funkcionalne oblike. Slika 1.4 predočava jednu takvu povezanost, gdje se uočava da pravac (predočen iscrtanom linijom) lošije opisuje podatke, u odnosu na punu krivulju (koja može predstavljati graf kvadratne funkcije ili pak graf eksponencijalne funkcije). Dakle, uočava se kako dijagram rasipanja može davati puno informacija o povezanosti između varijabli i prije nego se upuštamo u formalnu analizu, što može olakšati buduću analizu i adekvatan odabir modela koji se analizira. Tako bi se u slučaju slike 1.4 adekvatan model mogao zapisati na način: $y = a + bx + cx^2 + \varepsilon$, gdje vrijedi $c > 0$ jer je pripadna krivulja konveksnog oblika.



Slika 1.4. Dijagram rasipanja između varijabli x i y, parabolična veza

Konačno, dijagram rasipanja i može upućivati i na odsustvo bilo kakve povezanosti između dviju varijabli, kao što je prikazano na slici 1.5. U ovome slučaju nije potrebno raditi daljnje analize jer naprosto se ne uočava učinak x na y koji bi se mogao opisati nekom funkcijom.



Slika 1.5. Dijagram rasipanja između varijabli *x* i *y*, ne postoji veza

Primjer 1.5.
Učitajmo u RStudio[4] datoteku "**BDP_i_HICP.txt**" i nacrtajmo dijagram rasipanja između varijabli BDP i HICP (harmonizirani indeks cijena potrošnje). Najprije provjerimo koji je radni direktorij u RStudiju.

Nakon učitavanja RStudija, potrebno je u panelu u koji se unose naredbe upisati getwd() i pokrenuti naredba pritiskom na dugme Run. Naredba getwd(), kao većina u RStudiju, označava skraćenicu za izvršenje određene radnje. U panelu u kojem se pojavljuju ispisi izvršenih radnji će se potom ispisati da je izvršena naredba ("> getwd()" dio) i pisat će koji je radni direktorij (vidjeti sliku 1.6.). Definiranje radnog direktorija je bitno, s obzirom da će iz njega RStudio koristiti datoteke koje želimo učitati, kao i što će se u taj direktorij spremati određeni ispisi, slike i drugi korisniku važni objekti. Ako trenutni radni direktorij nije onaj u kojem želimo raditi, naredbom setwd(...), pri čemu se u zagradu u nazivnike navede naziv željenog direktorija će se upravo definirati novi direktorij.

```
> getwd()
[1] \\\\C/local/Users03$/Documents
```

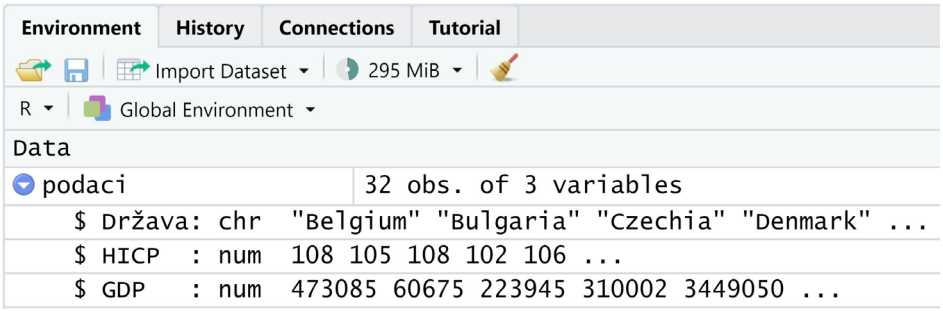
Slika 1.6. Ispis izvršene naredbe i radnog direktorija

Učitavanje datoteke u ovom primjeru vrši se pomoću naredbe read.table(), što je prikazano na slici 1.7. Na prvo mjesto navodimo naziv objekta kojeg definiramo, što je u ovom slučaju "podaci", potom se uvodi operator "definiram kao", što je znak "<-". Nakon tog operatora dolazi naredba kojom provodimo određenu radnju i uobičajeno je da sve naredbe imaju zagradu u kojoj se navode podaci nad kojima se provodi određena radnja, te pojedinačni uvjeti za izvršenje naredbe, u ovisnosti same radnje. U slučaju učitavanja podataka za ovaj primjer, na prvo mjesto u navodnike stavljamo ime datoteke koju učitavamo, potom naredba *header* = T označava da tablica koju učitavamo ima naziv varijabli u prvom retku (T označava *true*, engl. točno), i posljednja opcija sep="\t" označava da je tabulator razdjelnik podataka. Stoga je potrebno poznavati o kakvim podacima se radi, kako su definirane datoteke u kojima su spremljeni ti podaci (Excel tablica, CSV format, txt, itd.).

[4] Pretpostavlja se da je čitatelj instalirao RStudio, te da je upoznat s korištenjem tog editora. Dobar uvod je RStudio je *Using R for Introductory Econometrics,* dostupan na http://www.urfie.net/read/index.html.

Slika 1.7. Učitavanje datoteke BDP_i_HICP.txt

U panelu naziva "Environment" se nalazi spremljena tablica, koja sadrži informaciju o državama, te vrijednostima HICP i GDP (BDP). Na Slici 1.8. je vidljivo da se radi o 3 varijable, s ukupno 32 opažanja za svaku.

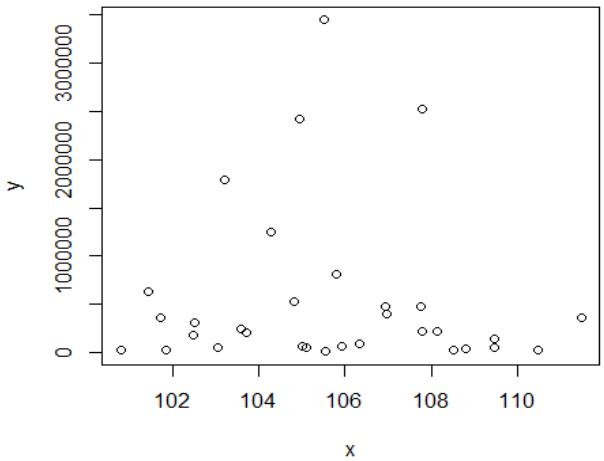


Slika 1.8. Učitani podaci

Ako želimo nacrtati dijagram raspianja, potrebno je posebno izdvojiti varijable HICP i GDP kao nove objekte. To ćemo postići tako da napišemo naredbe y<-podaci$GDP i x<-podaci$HICP, što znači da definiramo nazive y i x, potom naredbe "definira se kao", i iz postojeće tablice podaci naredbom "$" označavamo radnju da prikupljamo određen dio podataka, i posljednje, naziv varijable čije podatke prikupljamo. Kao rezultat, dobit ćemo dva nova objekta, niz *y* i *x*. Alternativno, moguće je učitati podatke pomoću naredbe "Import Dataset" koja se nalazi u panelu na slici 1.8. tako da se prate upiti koji se pojave u novom izborniku nakon što se odabere spomenuta naredba. Prednost naredbe "Import dataset" je da su nizovi odmah strukturiran kao tablica - data frame – u kojoj svaki stupac označava pojedini niz i olakšavamo daljnje analize tako da je taj niz odmah imenovan prema nazivu kojeg smo definirali u Excel, txt ili drugome formatu.

Konačno, za crtanje dijagrama rasipanja, koristi se naredba plot(x,y), pri čemu na prvo mjesto navodimo ime varijable čije će se vrijednosti predočavati na *x*-osi, a na drugo mjesto ime varijable čije će se vrijednosti predočavati na *y*-osi (vidjeti slike 1.9. za naredbe i 1.10. za dijagram rasipanja).

Slika 1.9. Naredbe za izdvajanje varijabli iz tablice "podaci" i crtanje dijagrama rasipanja

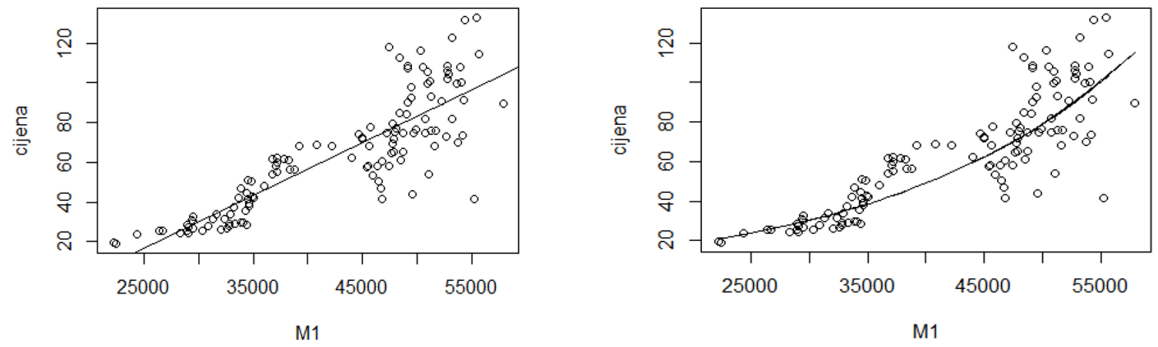


Slika 1.10. Dijagram rasipanja između HICP-a (x) i BDP-a (y)

1.5. Pitanja za ponavljanje

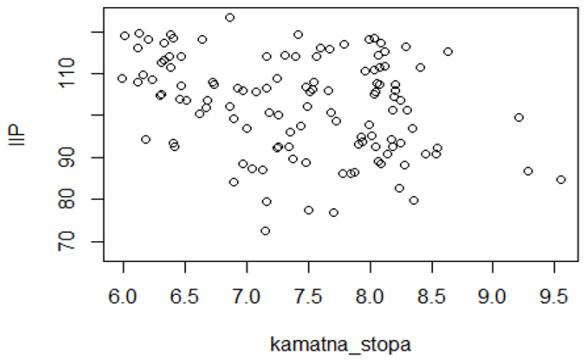
1) Što je ekonometrija?
2) Koja je razlika između ekonometrije i matematičke ekonomije?
3) Što je model, a što metoda?
4) Koja je razlika između parametra i varijable modela?
5) Koje su vrste ekonomskih podataka? Navedite nekoliko primjera za svaku.
6) Koje su česte transformacije podataka prije same analize?
7) Što je to dijagram rasipanja?
8) Dani su dijagrami rasipanja za odabrane ekonomske varijable u Hrvatskoj.

Dijagram rasipanja za monetarni agregat M1 i cijenu nafte:



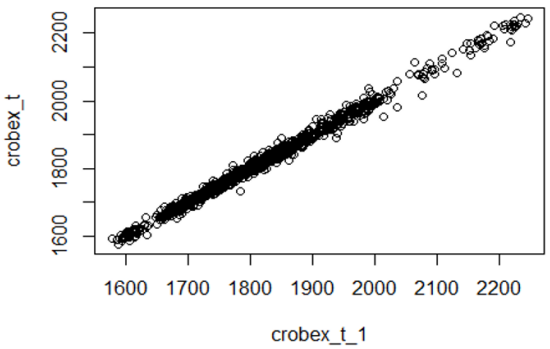
Kako biste zapisali model za procjenu za lijevi, a kako za desni dijagram?

Dijagram rasipanja za indeks industrijske proizvodnje (IP) i kamatnu stopu:



Što zaključujete o povezanosti IIP-a i kamatne stope?

Dijagram rasipanja za indeks CROBEX u mjesecu t i $t-1$:



9) Kakva je povezanost između vrijednosti indeksa CROBEX u mjesecu t i t–1 iz prethodnog pitanja?

10) Učitajte datoteku „**dijagram.txt**“ u RStudio (datoteka „dijagram.R“). Datoteka sadrži podatke o tri zavisne varijable, y_1, y_2 i y_3 te o jednoj nezavisnoj varijabli x. Predočite dijagram rasipanja između svake zavisne varijable i nezavisne varijable x. Komentirajte i zapišite kakav bi model najbolje opisivao vezu između nezavisne i svake zavisne varijable. Koja je jačina veze između nezavisne i svake od zavisnih varijabli?

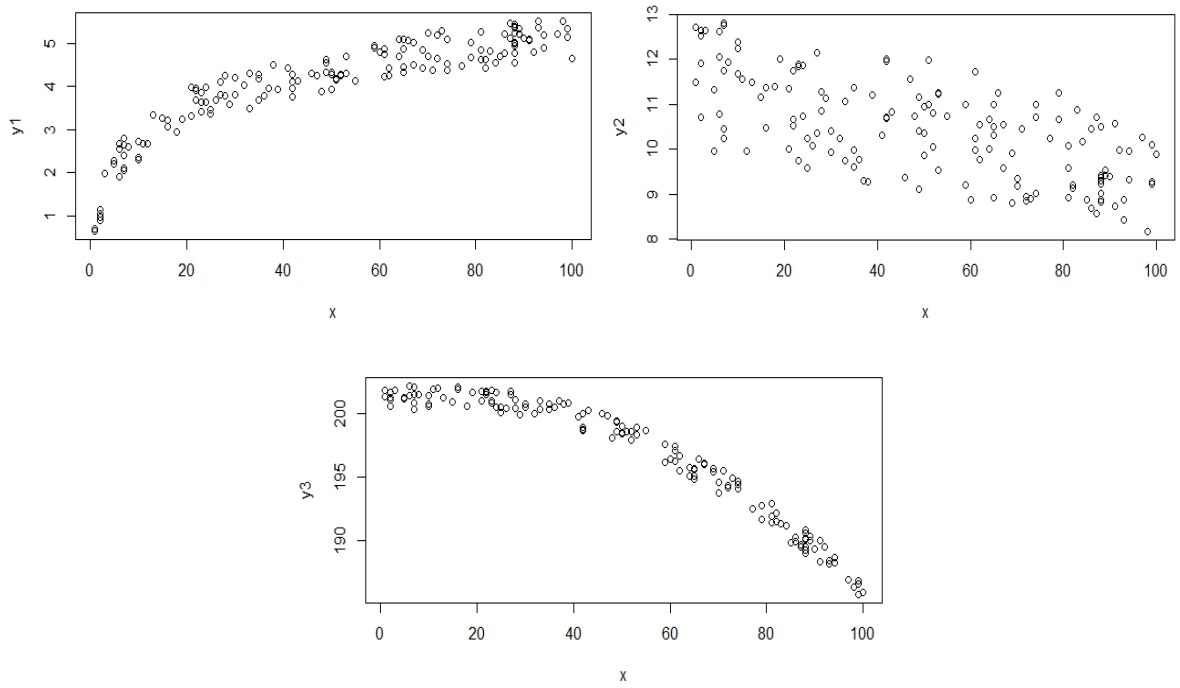
Rješenja

Zadatak 8)

$\text{Cijena}_t = a + b\text{M1}_t + u_t$

$\text{Cijena}_t = c + d\text{M1}_t + e\,\text{M1}^2_t + e_t$

Zadatak 10)



Model za y_1: $y_1 = a + b \log(x) + \varepsilon$
Model za y_2: $y_2 = c + d\,x + e$, $d < 0$
Model za y_3: $y_3 = b_0 + b_1\,x + b_2\,x^2 + u$, $b_2 < 0$

2.

LINEARNI REGRESIJSKI MODEL

2. LINEARNI REGRESIJSKI MODEL

2.1. Model jednostavne linearne regresije

U ovome poglavlju obrađuje se model u kojemu se pretpostavlja da jedna zavisna varijabla, y, ovisi o samo jednoj nezavisnoj varijabli, x. Regresijski model opisuje vezu između zavisne i nezavisne (ili više nezavisnih) varijabli. Općenito se zavisna varijabla označava s y, dok se nezavisna označava s x. Pritom se pretpostavlja određeni funkcionalni oblik između x i y, tako da je $y = f(x) + \varepsilon$. Kako se radi samo o jednoj nezavisnoj varijabli, ovaj oblik modela se naziva model jednostavne linearne regresije, bivarijatni linearni model.

2.1.1. Osnovna terminologija

Sama riječ „regresija“ prvi puta je korištena od strane F. Galtona (1822-1911) u Engleskoj, koji je razmatrao vezu između visine djece i njihovih roditelja. Kako je zaključio da visoki roditelji imaju visoku djecu, a niži roditelji nižu djecu, zaključio je da postoji „regresija dječje visine prema prosjeku“, odnosno „regresija prema mediokritetu“ (Maddala i Lahiri, 2009). Iako se danas regresija ne razmatra na način kako je Galton opisao, sam termin je ostao. Model jednostavne linearne regresije zapisuje se na način:

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

odnosno ako se zapisuje za svaki entitet u slučaju presječnih podataka:

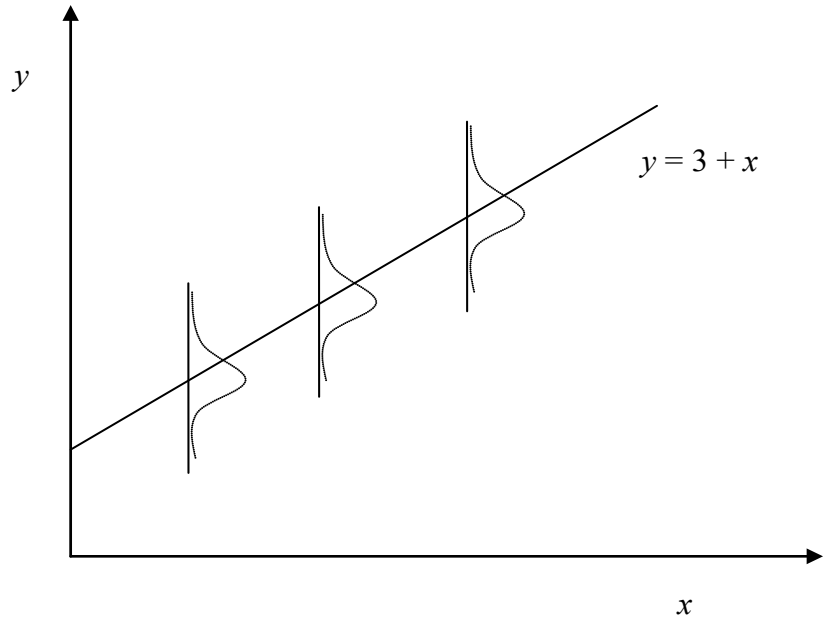
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{2.2}$$

Već je spomenuto kako su β_0 i β_1 nepoznati parametri koje je potrebno procijeniti, dok su varijable modela y i x, a ε je **slučajna** varijabla, stohastička varijabla ili greška relacije (engl. *error term*, *random variable*, *random disturbance*).

Tri glavna razloga zašto se slučajna varijabla uključuje u model jesu:

i) nepredvidivost slučajnog ponašanja varijabli koje se razmatraju,
ii) učinak izostavljanja drugih varijabli u model te
iii) greške u mjerenju zavisne varijable.

Zbog te slučajnosti, govori se o stohastičkoj vezi između varijable x i y, što je predočeno na slici 2.1. Na pravcu $y = 3 + x$ nalaze se deterministčke vrijednosti varijable y, dok je stohastička komponenta, slučajna varijabla uzrok tome da se stvarne vrijednosti varijable y nalaze negdje na pravcima koji su okomiti na os apscisa, tj. distribucija varijable y je centrirana oko očekivane vrijednosti varijable y uz dane vrijednosti varijable x, simbolički: $E(y \mid x)$. Vrijednosti varijable y koje su koordinate točaka koje su na pravcu $y = 3 + x$ tumače se koliko **u prosjeku** iznosi vrijednost varijable y s obzirom na vrijednosti varijable x, **a ne tumači se da je y točno jednak** vrijednosti $3 + x$ za sve jedinice iz populacije varijable x.



Slika 2.1. Stohastička veza između varijabli x i y

Drugi nazivi za varijable y, x i ε još su prikazani u tablici 2.1.

Tablica 2.1. Klasifikacija varijabli u regresijskoj analizi

$\boldsymbol{y}$	$\boldsymbol{x}$	$\boldsymbol{\varepsilon}$
Zavisna	Nezavisna	Slučajna
Endogena	Egzogena	Stohastička
Regresand	Regresor	Greška relacije
Prediktand	Prediktor	

Ako se model (2.2) želi procijeniti za sva opažanja u slučaju presječnih podataka, $i \in \{1, 2, \ldots, N\}$, tada se razmatra sustav od N jednadžbi:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ &\vdots \\ y_N &= \beta_0 + \beta_1 x_N + \varepsilon_N \end{aligned} \tag{2.3}$$

Korisnije je model zapisati u matričnoj formi:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.4}$$

gdje je $\boldsymbol{y} \in \mathbb{R}^N$ vektor stupac čiji su elementi opažanja zavisne varijable, $\boldsymbol{X} \in \mathcal{M}_{N,2}$ je matrica čiji prvi stupac čine jedinice, a drugi stupac vrijednosti opažanja nezavisne varijable, $\boldsymbol{\beta} \in \mathbb{R}^2$ je vektor stupac nepoznatih parametara, dok je $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ vektor stupac slučajne varijable:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}. \qquad (2.5)$$

Prvi stupac u matrici $\boldsymbol{X}$ čine jedinice, s obzirom na umnožak s vektorom $\boldsymbol{\beta}$ čiji je prvi element konstanta, kako bi ta konstanta bila uključena u model. Programska podrška za procjenu ekonometrijskih modela koristi matrični zapis modela temeljem kojeg se procjenjuju nepoznati parametri. Detaljnije se matrični zapis regresijskog modela obrađuje u poglavlju 2.1.2.3.

2.1.2. Pretpostavke modela jednostavne linearne regresije

Pretpostavke modela jednostavne linearne regresije su sljedeće (Greene, 2002):

1. Linearnost modela – pretpostavlja se linearna veza između zavisne i nezavisne varijable.
2. Egzogenost podataka u matrici $\boldsymbol{X}$, tj. egzogenost nezavisne varijable: $E(\varepsilon_i \mid x_i) = 0, \ \forall i$. To znači da očekivana vrijednost greške relacije ne ovisi o vrijednostima nezavisne varijable. Ova pretpostavka slijedi ako je nezavisna varijabla deterministička (u ponovljenim mjerenjima su vrijednosti nezavisne varijable fiksne).
3. Greška relacije u prosjeku ne utječe na zavisnu varijablu:
   $E(\varepsilon_i) = 0, \forall i$, tj. $E(y_i \mid x_i) = \beta_0 + \beta_1 x_i, \ \forall i$.
4. Varijanca greške relacije je konstantna (homoskedastična): $Var(\varepsilon_i) = Var(\varepsilon_i \mid x_i) = \sigma^2, \forall i$.
5. Nezavisnost slučajne varijable, tj. nekoreliranost:
   $E(\varepsilon_i, \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j \mid x_i) = 0$ za $i \neq j$.
6. Slučajna varijabla normalno je distribuirana: $\varepsilon_i \sim N(0, \sigma^2), \ \forall i$.

Linearnost modela odnosi se na linearnost u parametrima. Model može biti linearan u parametrima ili linearan u varijablama. Ako kažemo da je model linearan u parametrima, to znači da se parametri ne pojavljuju u eksponentima, nisu međusobno pomnoženi, ne dijele se međusobno, razmatra se prva potencija svakog parametra, itd. Linearnost u varijablama odnosi na to da su sve varijable u razinama, također bez transformacija, odnosno koriste se originalne vrijednosti varijabli, bez njihova logaritmiranja, potenciranja, ili neke druge nametnute transformacije u nekome funkcijskom obliku.

Tablica 2.2. prikazuje nekoliko modela koji su linearni u parametrima, odnosno u varijablama. Modeli linearni u parametrima u prvome stupcu tablice 2.2. uvijek sadrže samo vrijednosti β_0 i β_1. S druge strane, drugi stupac sadrži modele linearne u varijablama, s obzirom da se razmatraju samo x_i i y_i vrijednosti. Bilo kakva transformacija varijable ili parametra vodi do nelinearnosti modela. Tako je u tablici 2.2. jedino model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ linearan i u varijablama i u parametrima.

U prvome stupcu uočavamo da su ostali modeli nelinearni u varijablama, jer u drugome modelu $y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i$ uočavamo logaritmirane vrijednosti nezavisne varijable, u modelu $\sqrt{y_i} = \beta_0 + \beta_1 x_i + \varepsilon_i$ je korjenovana vrijednost zavisne, dok se u modelu $y_i^2 = \beta_0 + \beta_1 \frac{1}{x_i} + \varepsilon_i$ razmatra kvadrat zavisne varijable te recipročna vrijednost nezavisne. U drugome stupcu uočavamo nelinearnosti vezane uz parametre: tako je u prvome modelu konstanta korjenovana, u drugome

je logaritmiran β_1, u trećem modelu se razmatra kvadrat konstante te recipročna vrijednost β_1, a u posljednjem modelu je korjenovana β_1.

Tablica 2.2. Modeli koji su linearni u parametrima ili varijablama

Model linearan u parametrima	**Model linearan u varijablama**
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	$y_i = \sqrt{\beta_0} + \beta_1 x_i + \varepsilon_i$
$y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i$	$y_i = \beta_0 + \ln \beta_1 x_i + \varepsilon_i$
$\sqrt{y_i} = \beta_0 + \beta_1 x_i + \varepsilon_i$	$y_i = \beta_0^2 + \frac{1}{\beta_1} x_i + \varepsilon_i$
$y_i^2 = \beta_0 + \beta_1 \frac{1}{x_i} + \varepsilon_i$	$y_i = \beta_0 + \sqrt{\beta_1} x_i + \varepsilon_i$

Ako se radi o modelu koji u početnom zapisu nije linearan, ali ga je moguće transformirati u linearan (u parametrima), transformaciju je moguće napraviti na nekoliko osnovnih načina. Naime, ako je model dan u eksponencijalnom zapisu, tada se primjenjuje transformacija logaritmiranjem. Primjerice, model

$$y_i = e^{\beta_0 + \beta_1 x_i + \varepsilon_i} \tag{2.6}$$

je moguće transformirati na način da se promatra prirodni logaritam lijeve i desne strane jednakosti (2.6):

$$\ln(y_i) = \ln\left(e^{\beta_0 + \beta_1 x_i + \varepsilon_i}\right), \tag{2.7}$$

$$\ln(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i \ . \tag{2.8}$$

Sada je potrebno zavisnu varijablu logaritmirati, te se procijeni model koji je linearan u parametrima. Nadalje, ako se radi o modelu u kojemu postoji recipročna (inverzna) veza između zavisne i nezavisne varijable:

$$y_i = \frac{1}{\beta_0 + \beta_1 x_i + \varepsilon_i}, \tag{2.9}$$

potrebno je izračunati recipročne vrijednosti obiju strani jednakosti (2.9):

$$\frac{1}{y_i} = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{2.10}$$

pa je u ovome slučaju potrebno izračunati recipročnu vrijednost zavisne varijable i ostatak modela je linearan u parametrima.

Egzogenost nezavisne varijable, $E(\varepsilon_i \mid x_i) = 0 \ \forall i$, znači da opažanja vezana uz varijablu x ne sadrže informacije o očekivanoj vrijednosti slučajne varijable. Kada bi ova pretpostavka bila narušena, tada bi se vrijednosti nezavisne varijable mogle koristiti za predviđanje slučajne varijable, i ona više ne bi bila slučajna varijabla. Važno je napomenuti da se pretpostavlja da je nezavisna varijabla nestohastička (ili deterministička drugim riječima). To znači da u ponovljenim mjerenjima bismo uvijek prikupili jednake podatke za nezavisnu varijablu, pri čemu bi se prikupili novi podaci samo za zavisnu i slučajnu varijablu. U stvarnosti to najčešće

nije slučaj, već je i nezavisna varijabla slučajna varijabla. No ova pretpostavka se uvodi kako bi se u regresijskoj analizi razmatrala samo veza između zavisne i nezavisne varijable (kasnije i više nezavisnih, u modelu višestruke linearne regresije), a ne i izvori varijacija nezavisne varijable. Dakle, ako je nezavisna varijabla stohastička, uvode se daljnje pretpostavke o slučajnoj varijabli ε.

Pretpostavka $E(\varepsilon_i) = 0, \ \forall i$, tj. greška relacije u prosjeku ne utječe na zavisnu varijablu je implicirana prethodnom pretpostavkom jer vrijedi (Greene[5], 2002:14):

$$E(\varepsilon_i) = E_{\boldsymbol{X}}(E(\varepsilon_i \mid \boldsymbol{X})) = E_{\boldsymbol{X}}(0) = 0, \tag{2.11}$$

odnosno riječima, očekivanje u danoj matrici $\boldsymbol{X}$, ako se slučajna varijabla realizira uz uvjet dane matrice $\boldsymbol{X}$ jednako je nuli. Posljedica ove pretpostavke je ta što će uvjetno očekivanje zavisne varijable biti jednako determinističkom dijelu modela, tj.

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i, \forall i \ , \tag{2.12}$$

ili u matričnom zapisu:

$$E(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}. \tag{2.13}$$

Homoskedastičnost varijance greške relacije, $Var(\varepsilon_i) = \sigma^2, \ \forall i$, tj. nepromjenjivost varijance slučajne varijable važna je pretpostavka kako bi raspršenost podataka zavisne varijable y na slici 2.1. bila podjednaka za bilo koje opažanje nezavisne varijable x (raspršenost distribucija na slici 2.1. je podjednaka ako je zadovoljena pretpostavka homoskedastičnosti varijance greške relacije).

Nekoreliranost slučajne varijable, $E(\varepsilon_i,\varepsilon_j) = Cov(\varepsilon_i,\varepsilon_j) = Cov(\varepsilon_i,\varepsilon_j \mid x_i) = 0$ za $i \neq j$, znači da su dvije slučajne varijable međusobno nezavisne i ujedno i nekorelirane[6].

Pretpostavka normalne distribuiranosti slučajne varijable koristi se u svrhu provođenja inferencijalne analize regresijskog modela nakon njegove procjene, poput provođenja t-testa, intervalnih procjena parametara i testiranja hipoteza.

Za linearni regresijski model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, pretpostavke koje se odnose na slučajnu varijablu, nazivamo **Gauss-Markovljevim uvjetima**, i formalno se skraćeno mogu zapisati kao: $\varepsilon_i \sim N(0, \sigma^2), \ \forall i \ \ Cov(\varepsilon_i,\varepsilon_j) = 0$ za $i \neq j$ i $E(\varepsilon_i \mid x_i) = 0$.

Ako su ispunjene sve navedene pretpostavke, tada vrijedi sljedeće za zavisnu varijablu y:

---

[5] Relacija (2.11) vrijedi zbog zakona iterativnog očekivanja (engl. *law of iterated expectations*). Alternativno, naziva se i zakon ukupnog očekivanja (engl. *law of total explectations*): $E(E(\varepsilon_i \mid X)) = E(\varepsilon_i)$, vidjeti Hayashi (2000: 8). (2.11) znači da je očekivana vrijednost slučajne varijable jednaka uvjetnome očekivanju slučajne varijable uz dane vrijednosti druge varijable $\boldsymbol{X}$. Dokaz da ovo vrijedi je sljedeći. Pretpostavimo da je očekivanje slučajne varijable konačno, tj. $E(\varepsilon)<\infty$. Vrijedi: $E\left[E(\varepsilon \mid \boldsymbol{X})\right] = E\left[\sum_{\varepsilon} \varepsilon \cdot P(\varepsilon = \varepsilon \mid \boldsymbol{X})\right] =$

$$\sum_{x}\left[\sum_{\varepsilon} \varepsilon \cdot P(\varepsilon = \varepsilon \mid \boldsymbol{X} = \boldsymbol{x})\right] \cdot P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{x}\sum_{\varepsilon} \varepsilon \cdot P(\varepsilon = \varepsilon, \boldsymbol{X} = \boldsymbol{x}) = \sum_{x}\sum_{\varepsilon} P(\varepsilon = \varepsilon, \boldsymbol{X} = \boldsymbol{x}) = \sum_{x} \varepsilon \cdot P(\varepsilon = \varepsilon) = E(\varepsilon) \cdot$$

[6] Nezavisnost dviju varijabli znači da je njihova zajednička distribucija vjerojatnosti jednaka umnošku njihovih graničnih vjerojatnosti: $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$. Ako su dvije varijable nezavisne, tada su i nekorelirane. Ako su dvije varijable nekorelirane, tada je njihov koeficijent korelacije jednak nuli. Međutim, ako su nekorelirane, to ne implicira da su i nezavisne. Vidjeti Dodatak 5.3. za detalje.

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i \ , \ \forall i, \tag{2.14}$$

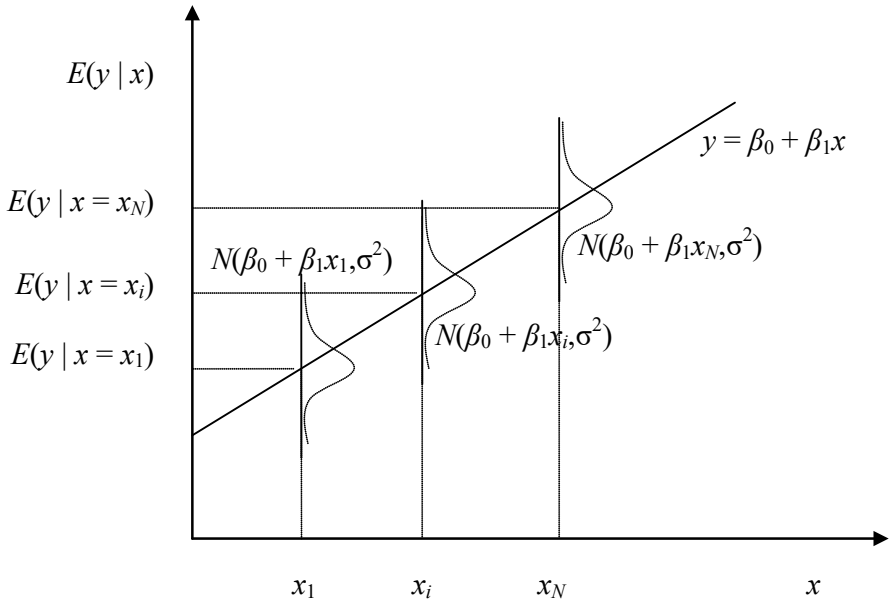
$$Cov(y_i, y_j) = Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ za } i \neq j, \tag{2.15}$$

$$Var(y_i) = Var(\varepsilon_i) = \sigma^2 \ , \ \forall i, \tag{2.16}$$

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \ , \ \forall i, \tag{2.17}$$

odnosno, zavisne varijable y_i imaju uvjetno očekivanje jednako $\beta_0 + \beta_1 x_i$, nekorelirane su slučajne varijable, s konstantnom varijancom i normalno su distribuirane.

Sve navedene pretpostavke jednostavnog linearnog regresijskog modela su sada predočene na slici 2.2., koja predstavlja proširenje slike 2.1.



Slika 2.2. Pretpostavke jednostavnog linearnog regresijskog modela

2.1.3. Metode procjene parametara

Kao prvi korak u regresijskoj analizi, potrebno je procijeniti nepoznate parametre modela koji se razmatra. Ako razmatramo model jednostavne linearne regresije $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, potrebno je procijeniti vrijednosti β_0 i β_1. Postoje različite metode za njihovu procjenu, pri čemu su najčešće metode koje se primjenjuju metoda najmanjih kvadrata, metoda najveće vjerodostojnosti i metoda momenata.

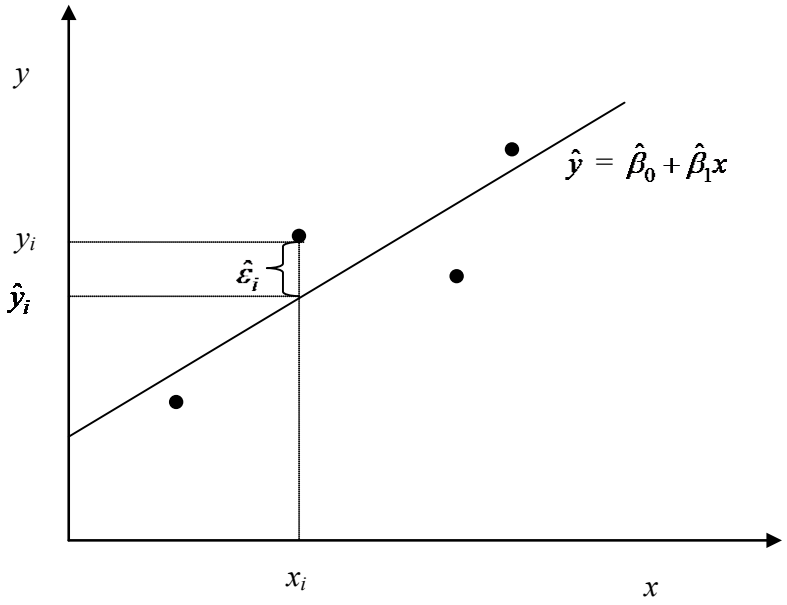
2.1.3.1. Metoda najmanjih kvadrata

Polazni model koji se razmatra, čiji se parametri trebaju procijeniti je $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Bez obzira na metodu procjene parametara, uvijek je ideja pronaći procjene parametara β_0 i β_1 koje su najbliže stvarnim vrijednostima po izabranom kriteriju. Ako označimo s $\hat{y}_i$ **procijenjenu vrijednost** zavisne varijable, te s $\hat{\beta}_0$ i $\hat{\beta}_1$ procijenjene vrijednosti nepoznatih parametara regresijskog modela, procijenjeni model zapisujemo u obliku:

$$E(y_i \mid x_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{2.18}$$

dok je $\hat{\varepsilon}_i$ procijenjena vrijednost slučajne varijable ili **rezidualno odstupanje** (rezidual), dana kao razlika stvarne i procijenjene vrijednosti zavisne varijable:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i). \tag{2.19}$$



Slika 2.3. Rezidualno odstupanje predočeno kao razlika između stvarne i procijenjene vrijednosti zavisne varijable

Grafički je rezidualno odstupanje predočeno na slici 2.3., gdje se uočava kako procijenjene vrijednosti zavisne varijable pripadaju regresijskom pravcu, dok stvarne vrijednosti odgovaraju ordinati točaka koje predstavljaju prikupljene podatke. Razlika između stvarne vrijednosti zavisne varijable, y_i, i procijenjene vrijednosti koja pripada regresijskom pravcu, $\hat{y}_i$, naziva se rezidualno odstupanje. Pozitivna vrijednost rezidualnog odstupanja označava da je regresijski model podcijenio stvarnu vrijednost zavisne varijable ($y_i > \hat{y}_i$), dok negativna vrijednost označava da smo modelom precijenili stvarnu vrijednost zavisne varijable ($y_i < \hat{y}_i$).

Ideja metode najmanjih kvadrata (engl. *least squares method*, LS, *ordinary least squares*, OLS) je **minimizirati sumu kvadrata odstupanja stvarnih vrijednosti zavisne varijable od procijenjenih**, pri čemu je potrebno odrediti vrijednosti $\hat{\beta}_0$ i $\hat{\beta}_1$ za koje će se ostvariti taj minimum. Funkcija cilja koja se minimizira je $S\left(\hat{\beta}_0, \hat{\beta}_1\right)$, dakle, suma kvadrata odstupanja:

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2, \tag{2.20}$$

pri čemu su nam nepoznate vrijednosti $\hat{\beta}_0$ i $\hat{\beta}_1$. U literaturi se uobičajeno funkcija cilja označava slovom *S*, a kako su nepoznate vrijednosti $\hat{\beta}_0$ i $\hat{\beta}_1$ označava se funkcija od nepoznatih parametara kao $S(\hat{\beta}_0, \hat{\beta}_1)$. Stoga se problem optimizacije zapisuje kao:

$$\arg\min_{\hat{\beta}_0,\hat{\beta}_1} S(\hat{\beta}_0,\hat{\beta}_1)=\sum_{i=1}^{N}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)^2. \tag{2.21}$$

Nužan uvjet za minimum funkcije više varijabli sastoji se u izjednačavanju svih prvih parcijalnih derivacija te funkcije s vrijednošću 0:

$$\frac{\partial S(\hat{\beta}_0,\hat{\beta}_1)}{\partial\hat{\beta}_0}=0,\ \frac{\partial S(\hat{\beta}_0,\hat{\beta}_1)}{\partial\hat{\beta}_1}=0, \tag{2.22}$$

koje iznose:

$$\begin{aligned}\frac{\partial S(\hat{\beta}_0,\hat{\beta}_1)}{\partial\hat{\beta}_0}&=2\sum_{i=1}^{N}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)(-1)=0\\ \frac{\partial S(\hat{\beta}_0,\hat{\beta}_1)}{\partial\hat{\beta}_1}&=2\sum_{i=1}^{N}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)(-x_i)=0\end{aligned}. \tag{2.23}$$

Sustav u (2.23) naziva se **sustav normalnih jednadžbi**. Sustav u (2.23) dalje rješavamo na način da se obje jednadžbe podijele s vrijednošću (−2):

$$\begin{aligned}&\sum_{i=1}^{N}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0\\ &\sum_{i=1}^{N}x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0\end{aligned}, \tag{2.24}$$

te potom operator sume primijenimo nad izrazima u zagradama:

$$\begin{aligned}&\sum_{i=1}^{N}y_i-N\hat{\beta}_0-\hat{\beta}_1\sum_{i=1}^{N}x_i=0\\ &\sum_{i=1}^{N}x_iy_i-\hat{\beta}_0\sum_{i=1}^{N}x_i-\hat{\beta}_1\sum_{i=1}^{N}x_i^2=0\end{aligned}, \tag{2.25}$$

gdje rješavanjem prve jednadžbe dobivamo izraz

$$\hat{\beta}_0=\frac{\sum_{i=1}^{N}y_i-\hat{\beta}_1\sum_{i=1}^{N}x_i}{N}=\underbrace{\frac{\sum_{i=1}^{N}y_i}{N}}_{\bar{y}}-\hat{\beta}_1\underbrace{\frac{\sum_{i=1}^{N}x_i}{N}}_{\bar{x}}=\bar{y}-\hat{\beta}_1\bar{x} \tag{2.26}$$

koji uvrstimo u drugu jednadžbu

$$\sum_{i=1}^{N}x_iy_i-\underbrace{\left(\bar{y}-\hat{\beta}_1\bar{x}\right)}_{\hat{\beta}_0}\sum_{i=1}^{N}x_i-\hat{\beta}_1\sum_{i=1}^{N}x_i^2=0, \tag{2.27}$$

tj.

$$\sum_{i=1}^{N} x_i y_i - \overline{y}\sum_{i=1}^{N} x_i + \hat{\beta}_1 \overline{x}\sum_{i=1}^{N} x_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i^2 = 0, \tag{2.28}$$

odakle nalazimo da je

$$\hat{\beta}_1 = \frac{\overline{y}\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i y_i}{\overline{x}\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2} = \frac{\overline{y}\underbrace{N\frac{\sum_{i=1}^{N} x_i}{N}}_{N\overline{x}} - \sum_{i=1}^{N} x_i y_i}{\underbrace{N\overline{x}^2}_{\overline{x}N\frac{\sum_{i=1}^{N} x_i}{N}} - \sum_{i=1}^{N} x_i^2} = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{xy}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}. \tag{2.29}$$

Dakle, izrazi (2.29) i (2.26) daju formule za procjenu parametara jednostavnog linearnog regresijskog modela, koje će osigurati minimalnu vrijednost funkcije cilja u (2.21). Da se doista radi o minimumu, pokazuje nam dovoljan uvjet: Hesseova matrica funkcije (2.20) je pozitivno definitna za vrijednosti procijenjenih parametara (2.29) i (2.26). Doista, ako se izračuna Hesseova matrica za vrijednosti (2.29) i (2.26), tj. matrica drugih parcijalnih derivacija funkcije (2.20):

$$H = \begin{bmatrix} 2N & 2\sum x_i \\ 2\sum x_i & 2\sum x_i^2 \end{bmatrix}, \tag{2.30}$$

očito je da je pozitivno definitna jer vrijedi $2N > 0$ i $4N\sum_i x_i^2 - 4\left(\sum_i x_i\right)^2 > 0$ s obzirom da je N pozitivan broj te vrijedi Cauchy-Schwartzova nejednakost[7].

Postavlja se pitanje zašto se razmatra suma kvadrata u funkciji cilja (2.21), a ne suma apsolutnih vrijednosti odstupanja, ili neki drugi funkcijski oblik. Za sumu apsolutnih vrijednosti odstupanja morali bismo koristiti neku od tehnika numeričke optimizacije, ne bismo mogli izravno zapisati analitički zapis rješenja kao u slučaju sume kvadrata odstupanja, te bi statistička teorija za procjenitelje bila kompliciranija u odnosu na procjenitelje u (2.29) i (2.26). Nadalje, ne može se razmatrati niti minimizacija sume odstupanja, tj. rezidualnog odstupanja[8], s obzirom da se oni u prosjeku poništavaju (zbog pretpostavke regresijskog modela koja se odnosi na očekivanje slučajne varijable, za detalje vidjeti Wooldridge, 2016), vidjeti detaljnije algebarska svojstva metode najmanjih kvadrata u naslovu 2.1.3.4.

---

[7] Cauchy-Schwartzova nejednakost u Euklidskom prostoru glasi: $\left(\sum_i u_i v_i\right)^2 \le \sum_i u_i^2 \sum_i v_i^2$ pa je primjena nejednakosti u slučaju $N\sum_i x_i^2 - \left(\sum_i x_i\right)^2$ sljedeća: $\left(\sum_i x_i\right)^2 = \sum_i x_i \sum_j x_j = \sum_i x_i^2 + \sum_i \sum_{j\neq i} x_i x_j \ge \sum_i x_i^2$. Odakle slijedi $\left(\sum_i x_i\right)^2 = \left(\sum_i 1 \cdot x_i\right)^2 \le N\sum_i x_i^2$ pa je $N\sum_i x_i^2 - \left(\sum_i x_i\right)^2 \ge 0$.

[8] Postoje i radovi s drugačijim pristupom, kao primjerice, Martić (1991 a, b).

Dodatno, izračun procjenitelja u (2.29) može se zapisati i u obliku:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{xy}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} = \frac{\widehat{Cov(x,y)}}{\widehat{Var(x)}}, \qquad (2.31)$$

gjde brojnik predstavlja procjenu kovarijance između nezavisne i zavisne varijable, dok nazivnik predstavlja procjenu varijance nezavisne varijable.

Primjer 2.1.
Dani su (simulirani) podaci o nezavisnoj i zavisnoj varijabli u tablici 2.3. Procijenimo model jednostavne linearne regresije koristeći formule (2.29) i (2.26). Izračunajmo rezidualna odstupanja i interpretirajmo ih.

Tablica 2.3. Opažanja nezavisne i zavisne varijable

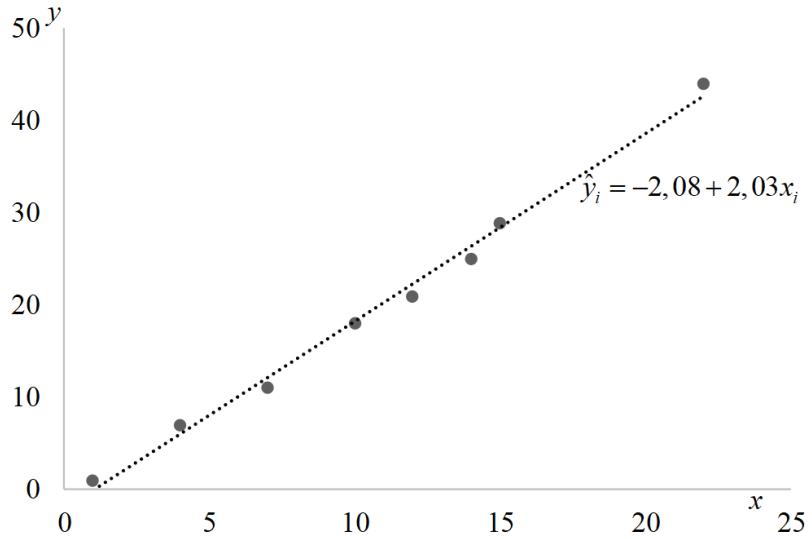
x	10	15	12	7	4	14	22	1
y	18	29	21	11	7	25	44	1

Najprije će se procijeniti parametar $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2} = \frac{(10\cdot 18 + \ldots + 1\cdot 1) - 8\cdot\frac{10+\ldots+1}{8}\cdot\frac{18+\ldots+1}{8}}{(10^2+\ldots+1^2) - 8\cdot\left(\frac{10+\ldots+1}{8}\right)^2} = 2{,}03,$$ te potom parametar

$\hat{\beta}_0$: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{18+\cdots+1}{8} - 2{,}03 \cdot \frac{10+\cdots+1}{8} = -2{,}08$. Dakle, procijenjeni model jest: $\hat{y}_i = -2{,}08 + 2{,}03x_i$.

Dijagram rasipanja zajedno s regresijskim pravcem predočen je na slici 2.4. Dodatno, uneseni su podaci u RStudio za varijable x i y, te su naredbama predočenim na slici 2.5. izračunati potrebni međurezultati kako bi se procijenili parametri modela. Ispis procijenjenih parametra je prikazan na slici 2.6. Nadalje, prikazan je dijagram rasipanja, zajedno s regresijskim pravcem na slici 2.8, temeljem naredbi prikazanih na slici 2.7.



Slika 2.4. Dijagram rasipanja za varijable *x* i *y* i regresijski pravac

```
x<-c(10,15,12,7,4,14,22,1)
y<-c(18,29,21,11,7,25,44,1)

x_potez<-mean(x)
y_potez<-mean(y)

umnozak<-t(x)%*%(y)
x_potez_2<-mean(x)^2
x_2<-sum(x^2)

beta_1<-(umnozak-8*x_potez*y_potez)/(x_2-8*x_potez_2)
beta_0<-y_potez-beta_1*x_potez
```

Slika 2.5. Unos potrebnih naredbi kako bi se procijenili parametri regresijskog modela u primjeru 2.1.

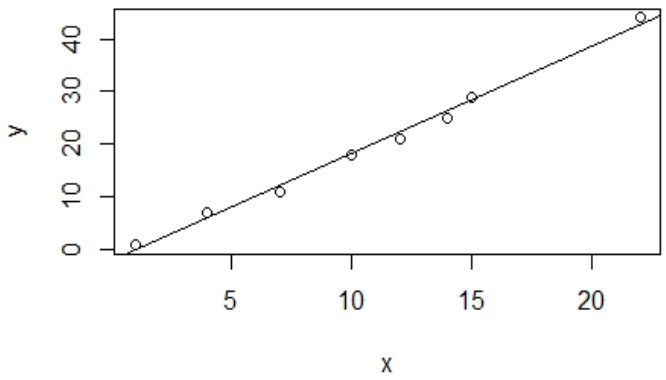
```
beta_1;beta_0

##          [,1]
## [1,] 2.031263

##           [,1]
## [1,] -2.082164
```

Slika 2.6. Ispis procijenjenih parametara modela u primjeru 2.1.

```
plot(x,y)
abline(a=beta_0,b=beta_1)
```

Slika 2.7. Naredbe potrebne za prikaz dijagrama rasipanja i regresijskog pravca

Nadalje, izračunata su rezidualna odstupanja kao razlika između stvarnih vrijednosti zavisne varijable i procijenjene i prikazana su u tablici 2.4. S obzirom da je vrijednost prvog opažanja zavisne varijable jednaka 18, a model je procijenio vrijednost 18,77, razlika iznosi –0,77, što znači da smo modelom precijenili vrijednost zavisne varijable. Modelom su peta, sedma i osma vrijednost zavisne varijable podcijenjene.



Slika 2.8. Dijagram rasipanja zajedno s regresijskim pravcem temeljem naredbi na slici 2.6

Tablica 2.4. Procjena rezidualnih odstupanja

y_i	18	29	21	11	7	25	44	1
$\hat{y}_i = -2{,}08 + 2{,}03x_i$	$-2{,}03+2{,}08\cdot 10 = 18{,}77$	29,17	22,93	12,53	6,29	27,09	43,73	0,05
$\hat{\varepsilon}_i = y_i - \hat{y}_i$	$18 - 18{,}77 = -0{,}77$	–0,17	–1,93	–1,53	0,71	–2,09	0,27	0,95

Slike 2.9. i 2.10. predočavaju potrebne naredbe za izračun procijenjenih vrijednosti i rezidualnih odstupanja te njihov ispis u RStudiju.

```
procjena<-beta_0[1,1]+beta_1[1,1]*x
rezidual<-y-procjena
procjena
```

Slika 2.9. Naredbe potrebne za izračun procijenjenih vrijednosti i rezidualnih odstupanja

```
procjena<-beta_0[1,1]+beta_1[1,1]*x
rezidual<-y-procjena

procjena

## [1] 18.2304609 28.3867735 22.2929860 12.1366733  6.0428858 26.3555110 42.6056112
## [8] -0.0509018

rezidual

## [1] -0.2304609  0.6132265 -1.2929860 -1.1366733  0.9571142 -1.3555110  1.3943888
## [8]  1.0509018
```

Slika 2.10. Ispis temeljem naredbi na slici 2.9

Ono što se može uočiti jest da će ovaj način procjene nepoznatih parametara biti veoma komplicirano, stoga je uobičajeno podatke prikazati u matričnom zapisu, kako bi izračun bio jednostavniji. Iduće potpoglavlje obrađuje metodu najmanjih kvadrata u matričnom zapisu.

2.1.3.2. Metoda najmanjih kvadrata u matričnom zapisu

Razmatra se model jednostavne linearne regresije u matričnom obliku:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.32}$$

gdje je $\boldsymbol{y} \in \mathbb{R}^N$ vektor stupac čiji su elementi opažanja zavisne varijable, $\boldsymbol{X} \in \mathcal{M}_{N,2}$ je matrica čiji prvi stupac čine jedinice, a drugi stupac vrijednosti opažanja nezavisne varijable, $\boldsymbol{\beta} \in \mathbb{R}^2$ je vektor stupac nepoznatih parametara, dok je $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ vektor stupac slučajne varijable:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}. \tag{2.33}$$

Procijenjeni model je:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{2.34}$$

gdje je $\hat{\boldsymbol{y}} \in \mathbb{R}^N$ vektor stupac s procijenjenim vrijednostima zavisne varijable, a $\hat{\boldsymbol{\beta}} \in \mathbb{R}^2$ je vektor stupac procijenjenih parametara. Sada je vektor rezidualnih odstupanja $\hat{\boldsymbol{\varepsilon}} \in \mathbb{R}^N$ vektor stupac:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{2.35}$$

$$\begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \tag{2.36}$$

Minimizira se suma kvadrata odstupanja stvarnih od procijenjenih vrijednosti zavisne varijable što se zapisuje kao:

$$\arg\min_{\hat{\boldsymbol{\beta}}} \left(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\right) = \arg\min_{\hat{\boldsymbol{\beta}}} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = S\left(\hat{\boldsymbol{\beta}}\right), \tag{2.37}$$

gdje je $S\left(\hat{\boldsymbol{\beta}}\right)$ oznaka za funkciju cilja. Izraz $\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$ može se zapisati kao:

$$\begin{aligned} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) &= \left(\boldsymbol{y}' - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\right)\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}. \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{2.38}$$

Nužni uvjet za postojanje minimuma implicira izjednačavanje svih parcijalnih derivacija prvog reda funkcije s nulom, tj. vektor kojemu su komponente parcijalne derivacije prvog reda izjednačavamo s nul-vektorom:

$$\frac{\partial S\left(\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}}} = \mathbf{0} \ , \tag{2.39}$$

odnosno

$$\frac{\partial\left(\boldsymbol{y}'\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}}} = \mathbf{0}, \tag{2.40}$$

$$-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{0}, \tag{2.41}$$

$$\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{0}, \tag{2.42}$$

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{2.43}$$

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \cdot / \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}, \tag{2.44}$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.45}$$

Dakle, formulom (2.45) definira se izračun nepoznatih parametara regresijskog modela u matričnom zapisu. Važno je uočiti u (2.45) kako je nužno uvesti pretpostavku da je matrica $\boldsymbol{X'X}$ punog ranga kako bi postojalo jedinstveno rješenje za $\hat{\boldsymbol{\beta}}$, odnosno da je sama matrica $\boldsymbol{X}$ punog ranga. O regularnosti matrice $\boldsymbol{X'X}$ bit će riječi i u 2.2.2. Dovoljan uvjet za minimum funkcije (2.37) jest da je Hesseova matrica $\frac{\partial^2 S\left(\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}}\,\partial \hat{\boldsymbol{\beta}}'} = 2\boldsymbol{X}'\boldsymbol{X}$ pozitivno definitna, s obzirom na skalarne umnoške odgovarajućih vektora redaka iz $\boldsymbol{X}'$ i vektora stupaca iz $\boldsymbol{X}$ koji nisu nul-vektori, glavni minori matrice $2\boldsymbol{X'X}$ su pozitivni (vidjeti matricu (2.30)).

Ako se uvede supstitucija $q = \hat{\boldsymbol{\varepsilon}}\boldsymbol{X'X}\hat{\boldsymbol{\varepsilon}}$, tada je $q = \boldsymbol{v'v}$, gdje je $\boldsymbol{v} = \boldsymbol{X}\hat{\boldsymbol{\varepsilon}}$, pri čemu svi elementi u q moraju biti pozitivni, osim ako je svaki element u $\boldsymbol{v}$ jednak 0. No u tom slučaju bi $\boldsymbol{v}$ bila linearna kombinacija stupaca iz $\boldsymbol{X}$ i što bi značilo da $\boldsymbol{X}$ nije punoga ranga, što je u suprotnosti s pretpostavkom da je matrica $\boldsymbol{X}$ punog ranga. Ali, uočeno se može zaključiti i temeljem oblika funkcije cilja (2.37), iz koje se vidi da se radi o strogom lokalnom minimumu.

Razmotrimo još pretpostavke jednostavnog linearnog regresijskog modela u matričnom zapisu.

1. Linearnost modela – pretpostavlja se linearna veza između zavisne i nezavisne varijable.
2. Egzogenost podataka u matrici $\boldsymbol{X}$, tj. egzogenost nezavisne varijable: $E(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \mathbf{0}$.
3. Greška relacije u prosjeku ne utječe na zavisnu varijablu: $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ tj. $E(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{\beta X}$.
4. Varijanca greške relacije je konstantna (homoskedastična).
5. Nezavisnost slučajne varijable, tj. nekoreliranost.
6. Slučajna varijabla normalno je distribuirana, $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\boldsymbol{\Omega})$, gdje je $\boldsymbol{\Omega}$ skalarna matrica čiji su elementi na glavnoj dijagonali jednaki σ^2 (vidjeti (2.46.)).

Pretpostavke 4 i 5 zapisuju se matrično na sljedeći način:

$$\Omega = E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I \tag{2.46}$$

gdje se uočava da se na glavnoj dijagonali matrice **Ω** nalazi varijanca slučajne varijable koja je nepromjenjiva (homoskedastična), dok se van glavne dijagonale nalaze nule, što znači da je slučajna varijabla nezavisna. Dakle, matrica **Ω** je skalarna matrica.

Primjer 2.2.
Dani su (simulirani) podaci o nezavisnoj i zavisnoj varijabli u tablici 2.5. Procijenimo model jednostavne linearne regresije koristeći formulu (2.45).

Tablica 2.5. Opažanja nezavisne i zavisne varijable

x	10	15	12	7	4	14	22	1
y	18	29	21	11	7	25	44	1

Najprije zapišimo matricu $\boldsymbol{X}$ i vektor $\boldsymbol{y}$: $\boldsymbol{X} = \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 12 \\ 1 & 7 \\ 1 & 4 \\ 1 & 14 \\ 1 & 22 \\ 1 & 1 \end{bmatrix}$, $\boldsymbol{y} = \begin{bmatrix} 18 \\ 29 \\ 21 \\ 11 \\ 7 \\ 25 \\ 44 \\ 1 \end{bmatrix}$.

Sada je

$$\hat{\boldsymbol{\beta}} = \left(\begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 12 \\ 1 & 7 \\ 1 & 4 \\ 1 & 14 \\ 1 & 22 \\ 1 & 1 \end{bmatrix}' \cdot \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 12 \\ 1 & 7 \\ 1 & 4 \\ 1 & 14 \\ 1 & 22 \\ 1 & 1 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 12 \\ 1 & 7 \\ 1 & 4 \\ 1 & 14 \\ 1 & 22 \\ 1 & 1 \end{bmatrix}' \cdot \begin{bmatrix} 18 \\ 29 \\ 21 \\ 11 \\ 7 \\ 25 \\ 44 \\ 1 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 15 & 12 & 7 & 4 & 14 & 22 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 10 \\ 1 & 15 \\ 1 & 12 \\ 1 & 7 \\ 1 & 4 \\ 1 & 14 \\ 1 & 22 \\ 1 & 1 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 15 & 12 & 7 & 4 & 14 & 22 & 1 \end{bmatrix} \cdot \begin{bmatrix} 18 \\ 29 \\ 21 \\ 11 \\ 7 \\ 25 \\ 44 \\ 1 \end{bmatrix} \right)$$

$$= \begin{bmatrix} -2,08 \\ 2,03 \end{bmatrix}$$

Dodatno, uneseni su podaci u RStudio za varijable *x* i *y*, te su naredbama, predočenim na slici 2.11. izračunati potrebni međurezultati kako bi se procijenili parametri modela.

```
x<-c(10,15,12,7,4,14,22,1)
y<-c(18,29,21,11,7,25,44,1)

y<-as.matrix(y)
jed<-c(rep(1,each=8))
jed<-as.matrix(jed)

x<-as.matrix(x)
X<-cbind(jed,x)

a<-t(X)%*%X
b<-solve(a)
c<-t(X)%*%y
beta<-b%*%c
```

Slika 2.11. Naredbe[9] za procjenu parametara regresijskog modela u primjeru 2.2.

Ispis procijenjenih parametra je prikazan na slici 2.12.

```
beta

##           [,1]
## [1,] -2.082164
## [2,]  2.031263
```

Slika 2.12. Ispis procijenjenih parametara modela u primjeru 2.2.

Procijenjene vrijednosti i rezidualna odstupanja izračunati su naredbama prikazanima na slici 2.13, a ispis je dan na slici 2.14.

```
procjena<-X%*%beta
rezidual<-y-procjena
procjena
```

Slika 2.13. Naredbe potrebne za izračun procijenjenih vrijednosti i rezidualnih odstupanja

```
procjena<-X%*%beta
rezidual<-y-procjena
```

```
procjena

##           [,1]
## [1,] 18.2304609
## [2,] 28.3867735
## [3,] 22.2929860
## [4,] 12.1366733
## [5,]  6.0428858
## [6,] 26.3555110
## [7,] 42.6056112
## [8,] -0.0509018
```

```
rezidual

##           [,1]
## [1,] -0.2304609
## [2,]  0.6132265
## [3,] -1.2929860
## [4,] -1.1366733
## [5,]  0.9571142
## [6,] -1.3555110
## [7,]  1.3943888
## [8,]  1.0509018
```

Slika 2.14. Ispis temeljem naredbi na slici 2.13.

[9] U ovome tekstu se koristi decimalni zarez, dok se u RStudiju zapisuje decimalna točka. Zarez u RStudiju označava razdvajanje brojeva ili izraza, pa zato postoji razlika između navođenja zareza kroz ovaj tekst, i decimalnih točaka na slikama. U tekstu je na pojedinim mjestima ostavljena decimalna točka ako se odnosi na uputu kako napisati neku naredbu u RStudiju.

2.1.3.3. Svojstva procjenitelja metode najmanjih kvadrata

Ako su zadovoljeni Gauss-Markovljevi uvjeti (vidjeti odjeljak 2.1.2), procjenitelj $\widehat{\boldsymbol{\beta}}$ ima određena poželjna svojstva. Prvo svojstvo naziva se **nepristranost** (engl. *unbiased estimator*), što znači da u ponovljenim mjerenjima i procjenama regresijskog modela, u prosjeku će očekivana vrijednost procjenitelja $\widehat{\boldsymbol{\beta}}$ biti jednaka stvarnoj vrijednosti nepoznatog $\boldsymbol{\beta}$ (prosjek ili očekivanje):

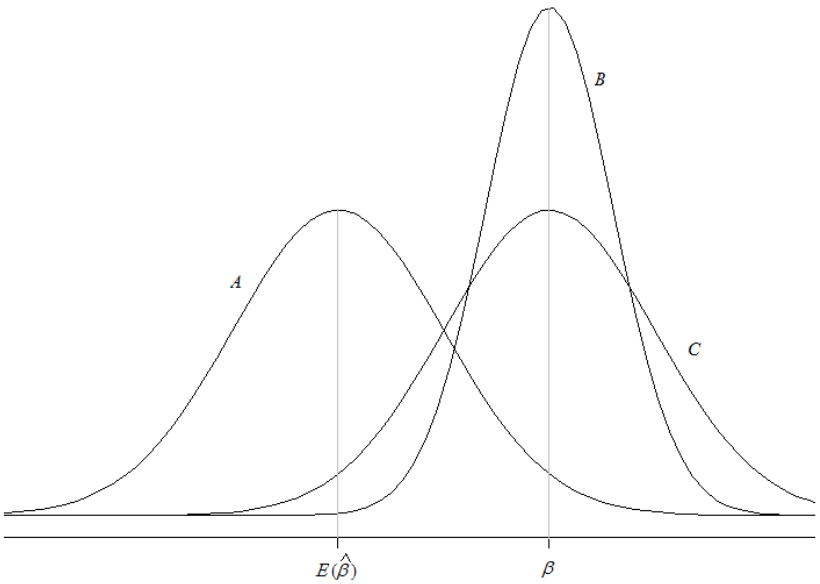
$$E\left(\hat{\boldsymbol{\beta}}\right)=E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\underbrace{\boldsymbol{y}}_{\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}}\right]=E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right)\right]$$

$$=E\left[\underbrace{\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}}_{I}\boldsymbol{\beta}+\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\right]$$

$$=E\left[\boldsymbol{\beta}+\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\right]$$

$$=\boldsymbol{\beta}+E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]\cdot\underbrace{E\left[\boldsymbol{\varepsilon}\right]}_{=0}=\boldsymbol{\beta} \tag{2.47}$$

Dakle, očekivana vrijednost procjenitelja jednaka je stvarnoj vrijednosti, ako je očekivana vrijednost slučajne varijable jednaka nuli i ako vrijedi pretpostavka egzogenosti nezavisne varijable (vidjeti pretpostavke 2 i 3 u naslovu 2.1.2). Procjenitelj za kojeg vrijedi $E\left(\hat{\boldsymbol{\beta}}\right)=\boldsymbol{\beta}$ nazivamo, dakle, **nepristranim**, dok onog procjenitelja za kojeg vrijedi $E\left(\hat{\boldsymbol{\beta}}\right)\neq\boldsymbol{\beta}$ nazivamo **pristranim** i razliku $E\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{\beta}$ nazivamo **pristranost** (engl. *bias*).



Slika 2.15. Usporedba procjenitelja parametra β

Slika 2.15. Uspoređuje procjenitelje za parametar β: očekivana vrijednost za distribuciju A, iznosi $E\left(\hat{\beta}\right) \neq \beta$, dok očekivana vrijednost za distribucije B i C iznosi β. Stoga su procjenitelji nepristrani u slučaju distribucija B i C, dok je pristran za distribuciju A.

Drugo svojstvo odnosi se na informaciju o tome koliko smo udaljeni od stvarne vrijednosti $\boldsymbol{\beta}$. Stoga se razmatra uvjetna varijanca procjenitelja, odnosno srednjekvadratna pogreška (engl. *mean squared error*) kako bi se razmotrila **efikasnost procjenitelja**, tj. odstupanje procjenitelja od stvarne vrijednosti parametara:

$$Var\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = E\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^2\right] \tag{2.48}$$

gdje vrijedi

$$E\left[\left(\underbrace{\hat{\boldsymbol{\beta}}}_{\substack{=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ =(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\varepsilon})}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right], \tag{2.49}$$

odnosno

$$E\left[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\varepsilon})-\boldsymbol{\beta})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\varepsilon})-\boldsymbol{\beta})'\right], \tag{2.50}$$

$$E\left[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})-\boldsymbol{\beta})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})-\boldsymbol{\beta})'\right], \tag{2.51}$$

$$E\left[\left(\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=I}\boldsymbol{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}-\boldsymbol{\beta}\right)\left(\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=I}\boldsymbol{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}-\boldsymbol{\beta}\right)'\right], \tag{2.52}$$

$$E\left[(\boldsymbol{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}-\boldsymbol{\beta})(\boldsymbol{\beta}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}-\boldsymbol{\beta})'\right], \tag{2.53}$$

$$E\left[(-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})(-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})'\right], \tag{2.54}$$

$$E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right], \tag{2.55}$$

$$E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')}_{=Var(\boldsymbol{\varepsilon})=\sigma^2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{2.56}$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=I}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{2.57}$$

$$\sigma^2\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=I}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2.58}$$

Dakle, varijanca procjenitelja jednaka je

$$Var\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2.59}$$

Za ovaj izvod koristili smo pretpostavke regresijskog modela 2 do 5. **Efikasan procjenitelj ima najmanju varijancu (2.59) u klasi svih linearnih nepristranih procjenitelja**.

Ako se ponovno razmotri slika 2.15., te se usporede distribucije *B* i *C*, oba procjenitelja su im nepristrana, ali je procjenitelj distribucije *B* efikasan jer mu je manja varijanca (usporedimo raspršenost obiju distribucija oko očekivane vrijednosti!)

Matrični zapis u (2.59) može se raspisati za slučaj modela jednostavne linearne regresije na sljedeći način:

$$\sigma^2 \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} = \begin{bmatrix} Var\left(\hat{\beta}_0\right) & Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) \\ Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) & Var\left(\hat{\beta}_1\right) \end{bmatrix}, \tag{2.60}$$

te su standardne pogreške procjenitelja jednake $SE(\hat{\beta}_0) = \sqrt{Var\left(\hat{\beta}_0\right)}$ i $SE(\hat{\beta}_1) = \sqrt{Var\left(\hat{\beta}_1\right)}$.

Kako vrijedi:

$$\left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix}^{-1} = \frac{1}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2} \begin{bmatrix} \sum_{i=1}^{N} x_i^2 & -\sum_{i=1}^{N} x_i \\ -\sum_{i=1}^{N} x_i & N \end{bmatrix}, \tag{2.61}$$

procjene varijanci $Var\left(\hat{\beta}_0\right)$ i $Var\left(\hat{\beta}_1\right)$ dane su kao:

$$Var\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{N} x_i^2}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2} \text{ i } Var\left(\hat{\beta}_1\right) = \frac{\sigma^2 N}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2}, \tag{2.62}$$

odnosno[10]

$$Var\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{N} + \frac{\overline{x}^2}{\hat{\sigma}_x^2} \right) \text{ i } Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\hat{\sigma}_x^2}. \tag{2.63}$$

Dakle, ako su ispunjene pretpostavke modela jednostavne linearne regresije, procjenitelj metodom najmanjih kvadrata je **najbolji linearni nepristrani procjenitelj**: nepristran je i ima najmanju varijancu – *BLUE* procjenitelj, tj. engl. *Best Linear Unbiased Estimator*.

---

[10] Jer vrijedi: $\frac{\sum_{i=1}^{N} x_i^2}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2} = \frac{\sum_{i=1}^{N} \left(x_i - \overline{x} \right)^2 + N\overline{x}^2}{N\left(\sum_{i=1}^{N} \left(x_i - \overline{x} \right)^2 \right)} = \frac{1}{N} + \frac{\overline{x}^2}{\sum_{i=1}^{N} \left(x_i - \overline{x} \right)^2} = \frac{1}{N} + \frac{\overline{x}^2}{\hat{\sigma}_x^2}$ i $\frac{N}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2} = \frac{N}{N\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2 \right)} = \frac{1}{\sum_{i=1}^{N} \left(x_i - \overline{x} \right)^2} = \frac{1}{\hat{\sigma}_x^2}$.

Procjena varijance σ^2 vrši se temeljem uzorka koji se razmatra. Naime, kako vrijedi[11]:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(I - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}, \qquad (2.64)$$

to je[12]

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \underbrace{\boldsymbol{M}\boldsymbol{X}}_{=0}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon} \quad . \qquad (2.65)$$

Stoga je varijanca rezidualnih odstupanja jednaka[13]

$$E\left[\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \mid \boldsymbol{X}\right] = E\left[\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right]. \qquad (2.66)$$

Matrica $\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}$ je formata (1,1), što znači da je upravo element koji se dobije umnoškom $\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}$ ujedno i trag te matrice. Stoga se lijeva strana jednakosti (2.66) može jednostavno izračunati pomoću traga desne strane te jednakosti. Kako su očekivane vrijednosti u (2.66) jednake, onda su im i tragovi jednaki, $E\left[tr\left(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \mid \boldsymbol{X}\right)\right] = E\left[tr\left(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right)\right]$.

Razmotrimo $E\left[tr\left(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right)\right]$[14]:

$$E\left[tr\left(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right)\right] = E\left[tr\left(\boldsymbol{M}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right)\right] = tr\left(\boldsymbol{M}E\left[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right]\right), \qquad (2.67)$$

pa je[15]

$$tr\left(\boldsymbol{M}E\left[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right]\right) = tr\left(\boldsymbol{M}\sigma^2 I\right), \qquad (2.68)$$

odnosno

$$tr\left(\boldsymbol{M}\sigma^2 I\right) = tr(\boldsymbol{M}) + tr\left(\sigma^2 I\right) = \left(N-2\right)\sigma^2, \qquad (2.69)$$

jer je iz (2.64) vidljivo da je

$$tr(\boldsymbol{M}) = tr(I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = tr(I) - tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = tr(I) - tr((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = N-2, \qquad (2.70)$$

jer je $\boldsymbol{X} \in \mathcal{M}_{N,2}$, a zbog formata $\hat{\boldsymbol{\varepsilon}}$ u (2.64) je format matrice I jednak formatu od $\hat{\boldsymbol{\varepsilon}}$.Stoga je

$$E\left[\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \mid \boldsymbol{X}\right] = (N-2)\sigma^2, \qquad (2.71)$$

te kako bi dobili nepristranu procjenu varijance metodom najmanjih kvadrata, potrebno je (2.71) podijeliti s izrazom (N–2):

$$E\left[\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \mid \boldsymbol{X}\right] = \sum_{i=1}^{N}\hat{\varepsilon}_i^2 = (N-2)\sigma^2 \;\big/ : (N-2), \qquad (2.72)$$

---

[11] Matrica $\boldsymbol{M}$ naziva se „hut“ matrica. To je simetrična i idempotentna matrica, odnosno simbolički zapisano: $\boldsymbol{M}'\boldsymbol{M}$ i $\boldsymbol{M}^2 = \boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

[12] U formuli (2.65) je $\boldsymbol{M}\boldsymbol{X} = \left(I - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}I = 0$.

[13] U formuli (2.66) je $\boldsymbol{M}'\boldsymbol{M} = \boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

[14] U formuli (2.67) je primijenjeno pravilo cikličkih permutacija u prvome koraku: $tr(AB) = tr(BA)$, dok je u drugome izračunata očekivana vrijednost od $\boldsymbol{M}$, s obzirom da je poznata iz (2.64).

[15] U formuli (2.68) je $E[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \mid \boldsymbol{X}] = \sigma^2 I$, vidjeti pretpostavke u naslovu 2.1.3.2.

pa je

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N-2}, \tag{2.73}$$

nepristran procjenitelj varijance metodom najmanjih kvadrata. Kako se radi o slučaju jednostavne linearne regresije, u nazivniku se pojavljuje izraz (N–2), što se može poopćiti na slučaj i višestruke linearne regresije, u kojemu će matrica $\boldsymbol{X} \in \mathcal{M}_{N,k+1}$, pa bi u nazivniku bio izraz (N–k–1).

Pretpostavlja se i da je **slučajna varijabla normalno distribuirana**, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$, jer je procjenitelj $\hat{\boldsymbol{\beta}}$ linearna funkcija slučajne varijable. Stoga se može pisati:

$$\hat{\beta}_j \mid \boldsymbol{X} \sim N\left(\beta_j, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}\right), \tag{2.74}$$

gdje *jj* predstavlja *j*-ti element na glavnoj dijagonali matrice $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Relacija (2.74), dakle, označava svojstvo normalne distribuiranosti procjenitelja $\hat{\boldsymbol{\beta}}$, što je važna pretpostavka kod inferencijalne analize regresijskog modela, u svrhu testiranja hipoteza, intervalne procjene parametara modela, itd. Oba elementa na glavnoj dijagonali matrice $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ su varijance procjenitelja linearnog regresijskog modela, dok njihov drugi pozitivni korijen predstavlja standardnu pogrešku procjenitelja.

Za slučaj jednostavnog linearnog modela to su $SE(\hat{\beta}_0) = \sqrt{Var\left(\hat{\beta}_0\right)}$ i $SE(\hat{\beta}_1) = \sqrt{Var\left(\hat{\beta}_1\right)}$. Tako se za svaki procjenitelj može pisati:

$$\hat{\beta}_0 \mid \boldsymbol{X} \sim N\left(\beta_0, Var\left(\hat{\beta}_0\right)\right) \tag{2.75}$$

i

$$\hat{\beta}_1 \mid \boldsymbol{X} \sim N\left(\beta_1, Var\left(\hat{\beta}_1\right)\right), \tag{2.76}$$

odnosno

$$\frac{\hat{\beta}_0 - \beta_0}{SE\left(\hat{\beta}_0\right)} \sim N(0,1) \text{ i } \frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)} \sim N(0,1). \tag{2.77}$$

No, kako se za standardne pogreške procjenitelja koristi procjena varijance u (2.73), slijedi:

$$\frac{\hat{\beta}_0 - \beta_0}{SE\left(\hat{\beta}_0\right)} \sim t(N-2) \text{ i } \frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)} \sim t(N-2), \tag{2.78}$$

odnosno standardizirane vrijednosti procjenitelja slijede Studentovu distribuciju (*t*-distribuciju) s N–2 stupnja slobode. U slučaju k nezavisnih varijabli, radilo bi se o N–k–1 stupnju slobode.

Primjer 2.3.
Za model jednostavne linearne regresije pokažimo da je procjenitelj dan izrazom (2.29) nepristran procjenitelj.

Kako je $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{N} x_i y_i - N\overline{xy}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}$, očekivana vrijednost iznosi:

$$E\left(\hat{\beta}_1\right) = E\left[\frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\,\overline{y}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}\right] = E\left[\frac{\sum_{i=1}^{N} x_i\left(\beta_0 + \beta_1 x_i + \varepsilon_i\right) - N\dfrac{\sum_{i=1}^{N} x_i}{N}\dfrac{\sum_{i=1}^{N}\left(\beta_0 + \beta_1 x_i + \varepsilon_i\right)}{N}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}\right]$$

$$= E\left[\frac{\beta_0\sum_{i=1}^{N} x_i + \beta_1\sum_{i=1}^{N} x_i^2 + \overbrace{\sum_{i=1}^{N} x_i\varepsilon_i}^{=0} - \sum_{i=1}^{N} x_i\left[\beta_0 + \left(\beta_1\overline{x}\right) + \overbrace{\dfrac{\sum_{i=1}^{N}\varepsilon_i}{N}}^{=0}\right]}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}\right] =$$

$$= \frac{\beta_1\sum_{i=1}^{N} x_i^2 - \beta_1\overline{x}\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} = \frac{\beta_1\left(\sum_{i=1}^{N} x_i^2 - \overline{x}\sum_{i=1}^{N} x_i\right)}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} = \beta_1\frac{\sum_{i=1}^{N} x_i^2 - \dfrac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} = \beta_1\frac{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} = \beta_1 .$$

Dakle, vrijedi $E\left(\hat{\beta}_1\right) = \beta_1$, što smo i trebali pokazati.

Primjer 2.4.

Zadan je procjenitelj za β_1 u obliku $\hat{\beta}_1 = \dfrac{y_n - y_1}{x_n - x_1}$. Pokažimo da se radi o nepristranom procjenitelju.

$E\left(\hat{\beta}_1\right) = E\left(\dfrac{y_n - y_1}{x_n - x_1}\right) = E\left(\dfrac{\beta_0 + \beta_1 x_n - \beta_0 - \beta_1 x_1}{x_n - x_1}\right) = \dfrac{\beta_1 x_n - \beta_1 x_1}{x_n - x_1} = \beta_1\dfrac{x_n - x_1}{x_n - x_1} = \beta_1$. Dakle, ponovno vrijedi $E\left(\hat{\beta}_1\right) = \beta_1$, pa se radi o nepristranom procjenitelju.

2.1.3.4. Algebarska svojstva metode najmanjih kvadrata

Kako vrijedi $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, nužan uvjet (2.42), $\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0}$, može se zapisati na sljedeći način:

$$\boldsymbol{X}'\left(\underbrace{\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}}_{\hat{\boldsymbol{\varepsilon}}}\right) = \boldsymbol{0} \tag{2.79}$$

$$\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}, \tag{2.80}$$

a kako je prvi stupac u matrici $\boldsymbol{X}$ stupac s jedinicama (vidjeti (2.33)), tada iz (2.80) slijede tri implikacije:

1) Suma rezidualnih odstupanja jednaka je nuli, tj. umnožak prvoga retka u matrici $\boldsymbol{X}'$ i stupca $\hat{\boldsymbol{\varepsilon}}$ jednak je $1\cdot\hat{\varepsilon}_1 + 1\cdot\hat{\varepsilon}_2 + ... + 1\cdot\hat{\varepsilon}_N = \sum_{i=1}^{N}\hat{\varepsilon}_i = 0$.
2) Regresijski pravac sadrži točku $(\overline{x}, \overline{y})$, što je vidljivo iz prve normalne jednadžbe (2.26). Možemo i prethodnu implikaciju zapisati na način: $\sum_{i=1}^{N}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$, odnosno $\sum_{i=1}^{N} y_i - N\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{N} x_i = 0 \,/ : N$, pa slijedi $\frac{\sum_{i=1}^{N} y_i}{N} - \hat{\beta}_0 - \hat{\beta}_1\frac{\sum_{i=1}^{N} x_i}{N} = 0$, tj. $\overline{y} - \hat{\beta}_0 - \hat{\beta}_1\overline{x} = 0$, što je upravo jednadžba (2.26) (prva jednadžba sustava normalnih jednadžbi (2.25)).
3) Prosječna vrijednost procijenjenih vrijednosti zavisne varijable jednaka je prosječnoj vrijednosti stvarnih vrijednosti zavisne varijable. Ako raspišemo prvu implikaciju: $\sum_{i=1}^{N}\hat{\varepsilon}_i = 0$ na sljedeći način: $\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right) = 0$ te izraz podijelimo s N: $\sum_{i=1}^{N} y_i - \sum_{i=1}^{N}\hat{y}_i = 0 \,/ : N$, slijedi $\frac{\sum_{i=1}^{N} y_i}{N} = \frac{\sum_{i=1}^{N}\hat{y}_i}{N}$, odnosno $\overline{y} = \overline{\hat{y}}$.

Primjer 2.5.
Provjerimo algebarska svojstva metode najmanjih kvadrata koristeći prethodni primjer.

Svojstvo 1) ispitamo tako da zbrojimo sva rezidualna odstupanja (slika 2.14 i 2.15.): $\sum_{i=1}^{8}\hat{\varepsilon}_i = -0,23 + 0,61 + ... + 1,05 = 0$

Svojstvo 2) ispitujemo tako da izračunamo prosječne vrijednosti zavisne i nezavisne varijable te ih potom uvrstimo u procijenjeni model:

$\frac{\sum_{i=1}^{8} y_i}{8} = \frac{18+\ldots+1}{8} = 19,5,\ \frac{\sum_{i=1}^{8} x_i}{8} = \frac{10+\ldots+1}{8} = 10,625,\ \hat{y}_i = -2,08 + 2,03x_i,\ \bar{y} = -2,08 + 2,03\bar{x}$
$= -2,08 + 2,03 \cdot 10,625 = 19,5.$

Svojstvo 3) ispitujemo tako da izračunamo prosjek procijenjenih vrijednosti zavisne varijable sa slika 2.16. i 2.17.:

$$\frac{\sum_{i=1}^{8} y_i}{8} = \frac{18+\ldots+1}{8} = 19,5$$

i usporedimo s prosjekom stvarnih vrijednosti zavisne varijable, koji također iznosi 19,5.

U RStudiju je potrebno napisati sljedeće naredbe na slici 2.16., kako bi se dobili rezultati na slici 2.17.

```
#svojstvo 1:
round(sum(rezidual),6)

#svojstvo 2:
mean(y)==beta_0+beta_1*mean(x)

#svojstvo 3:
mean(y);mean(procjena)
```

Slika 2.16. Naredbe za provjeru algebarskih svojstava metode najmanjih kvadrata

```
#svojstvo 1:
round(sum(rezidual),6)

## [1] 0

#svojstvo 2:
mean(y)==beta_0+beta_1*mean(x)

##      [,1]
## [1,] TRUE

#svojstvo 3:
mean(y);mean(procjena)

## [1] 19.5
## [1] 19.5
```

Slika 2.17. Rezultat temeljem naredbi zadanih na slici 2.16.

Mogu se koristiti i naredbe navedene na slici 2.18., kako bi se dobio rezultat dan na slici 2.19.

```
round(sum(rezidual),6)

mean(y); beta[1,]+beta[2,]*mean(x); mean(procjena)
```

Slika 2.18. Naredbe za provjeru algebarskih svojstava metode najmanjih kvadrata

```
round(sum(rezidual),6)

## [1] 0

mean(y); beta[1,]+beta[2,]*mean(x); mean(procjena)

## [1] 19.5

## [1] 19.5

## [1] 19.5
```

Slika 2.19. Rezultat temeljem naredbi zadanih na slici 2.18

2.1.3.5. Metoda najveće vjerodostojnosti

Metoda najveće vjerodostojnosti (ili vjerojatnosti, engl. *maximum likelihood method, ML*) je metoda procjene nepoznatih parametara koja se temelji na pretpostavci da je uvjetna distribucija promatranog pojma (endogene varijable) poznata uz izuzeće konačnog broja nepoznatih parametara. Ti parametri se procjenjuju na način da se odabiru one vrijednosti tih parametara za koje je vjerojatnost dobivanja vrijednosti iz uzorka maksimalna. Dakle, temeljem dostupnih podataka o nekoj varijabli, uz pretpostavku o distribuciji te varijable, određuje se vjerojatnost opažanja tog uzorka kao funkciju nepoznatih parametara koji karakteriziraju tu distribuciju. Potom se maksimizira ta vjerojatnost pri čemu su varijable odlučivanja nepoznati parametri distribucije.

Razmatra se slučajna varijabla y, sa slučajnim uzorkom opažanja ($y_1, y_2, \ldots, y_N$), pri čemu je funkcija gustoće vjerojatnosti, $f\ (y_i \mid \theta)$, uvjetovana skupom nepoznatih parametara θ. Zajednička funkcija gustoće vjerojatnosti N neovisnih identično distribuiranih (engl. *iid*, *independent identically distributed*) opažanja iz uzorka $f\ (y_1, y_2, \ldots, y_N \mid \theta)$, definira se kao umnožak pojedinačnih funkcija gustoće vjerojatnosti:

$$f(y_1, y_2, \ldots, y_N \mid \theta) = f(y_1 \mid \theta) \cdot \ldots \cdot f(y_N \mid \theta) = \prod_{i=1}^{N} f\left(y_i \mid \theta\right), \qquad (2.81)$$

pri čemu se ta funkcija $\prod_{i=1}^{N} f\left(y_i \mid \theta\right)$ definira kao funkcija nepoznatih parametara θ:

$$\prod_{i=1}^{N} f\left(y_i \mid \theta\right) = L\ (\theta \mid \boldsymbol{y}) \qquad (2.82)$$

za konkretni uzorak, jer se želi naglasiti da temeljem promatranog uzorka razmatramo nepoznate parametre θ koji će maksimizirati vrijednost $L\ (\theta \mid \boldsymbol{y})$, tj. da je vjerojatnost dobivanja vrijednosti iz uzorka maksimalna uz parametre θ. Stoga se funkcija $L\ (\theta \mid \boldsymbol{y})$ naziva **funkcija vjerodostojnosti**, izražena preko nepoznatih parametara θ za dana opažanja iz uzorka.

U empirijskim istraživanjima se češće razmatra logaritmirana vrijednost funkcije vjerodostojnosti:

$$\ln L\ (\theta \mid \boldsymbol{y}) = \ln \prod_{i=1}^{N} f\left(y_i \mid \theta\right) = \sum_{i=1}^{N} \ln f\left(y_i \mid \theta\right), \qquad (2.83)$$

te se potom traži maksimalna vrijednost od (2.83) pri čemu su varijable odlučivanja nepoznati parametri u θ:

$$\max_{\theta} \ln L(\theta \mid \boldsymbol{y}) = \max_{\theta} l(\theta). \tag{2.84}$$

Na taj način se lakše maksimizira funkcija (2.83), jer je aditivnog oblika, za razliku od multiplikativnog oblika u (2.82).

U slučaju jednostavnog linearnog regresijskog modela, ako su zadovoljene pretpostavke navedene u naslovu 2.1.2, y_i je normalno distribuiran uz danu vrijednost x_i, s očekivanjem $\mu_i = \beta_0 + \beta_1 x_i$ i varijancom σ^2, tj. $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Funkcija (2.84) u ovome slučaju bila bi:

$$l(\theta) = \sum_{i=1}^{N} \ln f\left(y_i \mid \beta_0, \beta_1, \sigma^2\right) = \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2}, \tag{2.85}$$

gdje je nužan uvjet za maksimizaciju vrijednosti funkcije $l(\theta)$ u (2.85) sljedeći:

$$\frac{\partial l(\theta)}{\partial \beta_0} = 0, \frac{\partial l(\theta)}{\partial \beta_1} = 0, \frac{\partial l(\theta)}{\partial \left(\sigma^2\right)} = 0, \tag{2.86}$$

odnosno[16]

$$\begin{aligned} \frac{\partial l(\theta)}{\partial \beta_0} &= \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta_1 x_i\right) = 0, \\ \frac{\partial l(\theta)}{\partial \beta_1} &= \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta_1 x_i\right) x_i = 0, \\ \frac{\partial l(\theta)}{\partial \left(\sigma^2\right)} &= -\frac{N}{2\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 = 0, \end{aligned} \tag{2.87}$$

gdje se uočava da iz prve i druge jednakosti slijede procjenitelji za β_0 i β_1 u slučaju metode najveće vjerodostojnosti jednaki procjeniteljima metodom najmanjih kvadrata. No, iz zadnje jednakosti, za procjenu varijance, dobiva se izraz:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N}, \tag{2.88}$$

gdje se radi o pristranom procjenitelju (usporediti s relacijom (2.73)). Može se zaključiti da ako su zadovoljene pretpostavke linearnog regresijskog modela, procjenitelji za β_0 i β_1 su jednaki u slučaju metode najmanjih kvadrata i metode najveće vjerodostojnosti, no u slučaju metode najmanjih kvadrata su procjenitelji nepristrani.

[16] Funkcija u (2.85) može se zapisati kao $N \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \frac{1}{\sigma^2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$.

2.1.3.6. Metoda momenata

Metoda momenata (engl. *method of moments*, *MM*) temelji se na procjeni momenata populacije pomoću momenata uzorka. Dakle, ako se želi procijeniti očekivana vrijednost neke varijable, potrebno je izračunati aritmetičku sredinu uzorka koji se prikupi. Slično tome se mogu računati ostali momenti. U slučaju jednostavne linearne regresije, jedna od pretpostavki bila je:

$$E(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = 0, \tag{2.89}$$

odnosno

$$E\left[\sum_{i=1}^{N} x_i \varepsilon_i\right] = 0, \tag{2.90}$$

te pretpostavka

$$E(\boldsymbol{\varepsilon}) = 0, \tag{2.91}$$

odnosno

$$E\left(\sum_{i=1}^{N} \varepsilon_i\right) = 0. \tag{2.92}$$

Temeljem uzorka, vrijedi:

$$E\left[\sum_{i=1}^{N} x_i \hat{\varepsilon}_i\right] = \sum_{i=1}^{N} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \tag{2.93}$$

i

$$E\left(\sum_{i=1}^{N} \hat{\varepsilon}_i\right) = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0. \tag{2.94}$$

Iz (2.94) slijedi da je $\sum_{i=1}^{N} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{N} x_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i^2 = 0$, dok iz (2.93) slijedi $\sum_{i=1}^{N} y_i - N\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{N} x_i = 0$, što su upravo jednadžbe sustava normalnih jednadžbi u (2.25) kod metode najmanjih kvadrata, čija su rješenja dana u (2.26) i (2.29).

Dakle, procjenitelji $\hat{\beta}_0$ i $\hat{\beta}_1$ dobiveni metodom momenata jednaki su procjeniteljima metodom najmanjih kvadrata. Ako se razmotri procjena varijance temeljem varijance uzorka:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N}, \tag{2.95}$$

zaključuje se kako je procjenitelj pristran jer vrijedi:

$$E\left(\hat{\sigma}^2\right) = E\left(\frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N}\right) = \frac{1}{N} E\left(\sum_{i=1}^{N} \hat{\varepsilon}_i^2\right) = \frac{(N-2)\sigma^2}{N}, \tag{2.96}$$

što se razlikuje od nepristranog procjenitelja (2.73). Međutim, kada bi se razmatrao uzorak veličine $N \to \infty$, tada bi izraz u (2.96) težio prema vrijednosti σ^2 i u tom slučaju procjenitelj metodom momenata postaje nepristran, pa se naziva **asimptotski nepristran procjenitelj**.

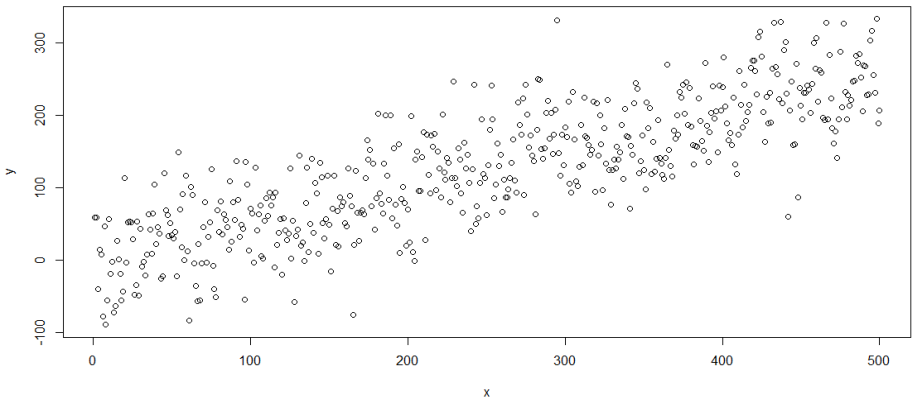
2.1.4. Sveobuhvatan primjer[17]

U RStudiju definirani su nizovi *x* i *y*, kako je prikazano na slici 2.20. Varijabla *x* generirana je kao niz vrijednosti 1 do 500, s razlikom između svake sukcesivne vrijednosti koja iznosi 1, te je slučajna varijabla (epsilon) generirana pomoću naredbe *rnorm*, koja generira slučajnu normalno distribuiranu varijablu, s očekivanjem 0 i varijancom 50. Potom je varijabla *y* generirana pomoću jednadžbe $y_i = 2 + 0.5x_i + \varepsilon_i$. U praski sa stvarnim podacima nećemo raspolagati s vrijednošću slučajne varijable, kao niti što ćemo znati stvarne vrijednosti parametara koji su ovdje definirani.

```
set.seed(1)
x<-seq(1,500,1)
epsilon<-rnorm(500,0,50)
y<-(2+0.5*x+epsilon)
```

Slika 2.20. Generiranje podataka za varijable *x* i *y*

Predočimo dijagram rasipanja između varijabli *x* i *y*, te komentirajmo njihovu vezu. Naredbom plot(x,y) dobiva se dijagram rasipanja na slici 2.21. Uočava se da postoji veza između varijabli *x* i *y*, pri čemu je to pozitivna umjerena veza.



Slika 2.21. Dijagram rasipanja temeljem generiranih podataka na slici 2.19.

Pomoću naredbe lm(…) (engl. *linear model*) procijenimo model jednostavne linearne regresije za zavisnu i nezavisnu varijablu, zapišimo procijenjeni model i za prvu procijenjenu vrijednost izračunajmo rezidualno odstupanje te interpretirajmo. Naredbom lm(y~x) procjenjuje se model u kojemu je s lijeve strane tilde zavisna varijabla, a s desne nezavisna. U okviru RStudija konstanta je uključena u procjenu kao automatski ugrađena naredba pa ju nije potrebno pisati. Dobiven je sljedeći rezultat, prikazan na slici 2.22.

Procijenjen model je $\hat{y}_i = 5{,}637 + 0{,}49x_i$. Temeljem naredbe *fitted*() možemo spremiti procijenjene vrijednosti zavisne varijable, te naredbom za spajanje u jednu tablicu predočiti

[17] Napomena: U primjeru se generiraju podaci o varijablama, stoga bi se vrijednosti u svakome ponavljanju razlikovale da se ne doda naredba set.seed(1). Čitatelj koji će provesti ove naredbe će u svojoj iteraciji dobiti različite vrijednosti od onih prdočenih u ovome i ostalim primjerima koji generiraju podatke, stoga valja paziti kod interpretacija, dok zaključci ostaju slični.

stvarne i procijenjene vrijednosti zavisne varijable, kao i rezidualna odstupanja, što je predočeno na slici 2.23, s rezultatima ispisa na slici 2.24.

```
lm(y~x)

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       5.637        0.490
```

Slika 2.22. Ispis temeljem naredbe lm(y~x)

```
procjena<-fitted(lm(y~x))
head(cbind(y,procjena,y-procjena))
```

Slika 2.23. Naredbe za izračun procijenjenih vrijednosti i tablice usporedbe vrijednosti

S obzirom da je stvarna vrijednost prvog opažanja zavisne varijable jednaka –28,82, dok je modelom procijenjena vrijednost 6,13, model je precijenio vrijednost zavisne varijable za 34,95 jedinica.

```
procjena<-fitted(lm(y~x))
head(cbind(y,procjena,y-procjena))

##           y procjena
## 1 -28.82269 6.126688 -34.949379
## 2  12.18217 6.616690   5.565476
## 3 -38.28143 7.106693 -45.388123
## 4  83.76404 7.596695  76.167346
## 5  20.97539 8.086697  12.888692
## 6 -36.02342 8.576699 -44.600118
```

Slika 2.24. Ispis stvarnih, procijenjenih vrijednosti zavisne varijable te rezidualnih odstupanja

2.1.5. Pitanja za ponavljanje

1) Što je regresijska analiza?
2) Što je model jednostavne linearne regresije?
3) Koja je razlika između zavisne i nezavisne varijable u regresijskom modelu?
4) Navedite nekoliko primjera jednostavnih linearnih regresijskih modela s ekonomskim varijablama.
5) Što se podrazumijeva pod linearnošću u linearnoj regresiji?
6) Što su greške relacije? Koja je njihova uloga u regresijskoj analizi?
7) Navedite pretpostavke jednostavnog linearnog regresijskog modela.
8) Zapišite Gauss-Markovljeve uvjete.
9) Koji od navedenih modela je linearan u parametrima, a koji u varijablama ?

$a)\ y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i$ _____

$b)\ y_i = \beta_0 + \log \beta_1 x_i + \varepsilon_i$ _____

$c)\ y_i = \beta_0^4 + \beta_1 x_i + \varepsilon_i$ _____

$d)\ y_i = \beta_0 + \beta_1 \sqrt[4]{x_i} + \varepsilon_i$ _____

$e)\ y_i = \beta_0 + \beta_1 \frac{2}{x_i} + \varepsilon_i$ _____

$f)\ y_i = \beta_0 + \sqrt{\beta_1 x_i} + \varepsilon_i$ _____

10) Klasificirajte sljedeće modele na linearne u parametrima ili varijablama:

$$\textit{(a)}\ y_i = \beta_0 + \beta_1 \sqrt[3]{x_i} + \varepsilon_i; \quad \textit{(b)}\ \sqrt{y_i} = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad \textit{(c)}\ y_i = \beta_0 + \beta_1 \frac{1}{x_i} + \varepsilon_i;$$

$$\textit{(d)}\ y_i = \beta_0 + \beta_1 e^{x_i} + \varepsilon_i; \quad \textit{(e)}\ y_i = \beta_0 x_i^{\beta_1} + \varepsilon_i; \quad \textit{(f)}\ \ln y_i = \ln \beta_0 + \beta_1 \ln x_i + \varepsilon_i;$$

$$\textit{(g)}\ y_i = \beta_0 + \frac{\beta_1}{x_i - \beta_2} + \varepsilon_i; \quad \textit{(h)}\ y_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} + \varepsilon_i; \quad \textit{(i)}\ y_i = \frac{x_i}{\beta_0 x_i + \beta_1} + \varepsilon_i.$$

11) Koje od modela u prethodnom zadatku možete linearizirati i kako?

12) Zadani su sljedeći nelinearni ekonometrijski modeli:

$$\textit{(a)}\ y_i = e^{\beta_0 + \beta_1 x_i + \varepsilon_i}; \quad \textit{(b)}\ y_i = \beta_0 + \beta_1^2 x_i + \varepsilon_i; \quad \textit{(c)}\ y_i = \left(\beta_0 + \beta_1 x_i + \varepsilon_i\right)^{-1}.$$

Koje je od navedenih modela moguće svesti na linearne modele u parametrima i zašto? Koje transformacije pri tome koristite?

13) Koja je razlika između sljedeća dva ekonometrijska modela?

$$\textit{(a)}\quad y_i = \beta_0 x_i^{\beta_1} \varepsilon_i; \qquad \textit{(b)}\quad y_i = \beta_0 x_i^{\beta_1} + \varepsilon_i$$

14) Popunite sljedeću tablicu kako biste procijenili parametre regresijskog modela $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Varijabla x odnosi se na cijenu (u kn), a varijabla y na potraživanu količinu (u kilogramima) kupusa u dućanu. Za procjenu parametara modela potom koristite formule (2.26) i (2.29). Izračunajte procijenjene vrijednosti zavisne varijable, te potom rezidualna odstupanja. Interpretirajte neko rezidualno odstupanje.

y_i	x_i	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$	$y_i - \bar{y}$	$\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$
5	1				
1	6				
3	3				
3,5	2				
4,2	4				
$\sum_{i=\_\_} y_i =$	$\sum_{i=\_\_} x_i =$	$\sum_{i=\_\_} \left(x_i - \bar{x}\right) =$	$\sum_{i=\_\_} \left(x_i - \bar{x}\right)^2 =$	$\sum_{i=\_\_} \left(y_i - \bar{y}\right) =$	$\sum_{i=\_\_} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) =$

15) U RStudio unesite podatke o nezavisnoj i zavisnoj varijabli iz prethodnog zadatka. Skicirajte dijagram rasipanja. Komentirajte ga. Kakva veza postoji između cijene i potraživane količine? Pretpostavimo da potraživana količina kupusa ovisi o cijeni na sljedeći način: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. O kojem modelu se radi s obzirom na (ne)linearnost varijabli i/ili parametara?

16) Opišite ukratko metodu najmanjih kvadrata. Koja je osnovna ideja te metode?

17) Zapišite vektor procijenjenih parametara $\hat{\boldsymbol{\beta}}$ dobiven iz nužnog uvjeta optimizacije funkcije (2.38).

18) Dani su (simulirani) podaci o nezavisnoj i zavisnoj varijabli u tablici ispod. Procijenimo model jednostavne linearne regresije koristeći formulu (2.45).

x	12	13	14	5	6	12	20	16
y	17	25	25	8	14	20	40	16

19) Za podatke dane u prethodnome zadatku, izračunajte procijenjene vrijednosti zavisne varijable te rezidualna odstupanja.
20) Provjerite algebarska svojstva metode najmanjih kvadrata temeljem podataka iz zadataka 18 i 19.
21) Koja su svojstva procjenitelja dobivenih metodom najmanjih kvadrata?
22) Opišite ukratko metodu najveće vjerodostojnosti. Kada su procjenitelji dobiveni metodom najveće vjerodostojnosti jednaki onima dobivenima metodom najmanjih kvadrata?
23) Opišite ukratko metodu momenata.

Rješenja

Zadatak 9):

a) P, b) V, c) V, d) P, e) P, f) V.

Zadatak 10):

a) P, b) P, c) P, d) P, e)V, f) P, g) P, h) ni u P ni u V, i) ni u P ni u V.

Zadatak 11):

Nelinearne modele u h) i i) nije moguće linearizirati.

Zadatak 12):

Modeli u a) i c) su nelinearni koje je moguće linearizirati, u a) pomoću log transformacije, a u c) pomoću recipročne/inverzne transformacije.

Zadatak 13):

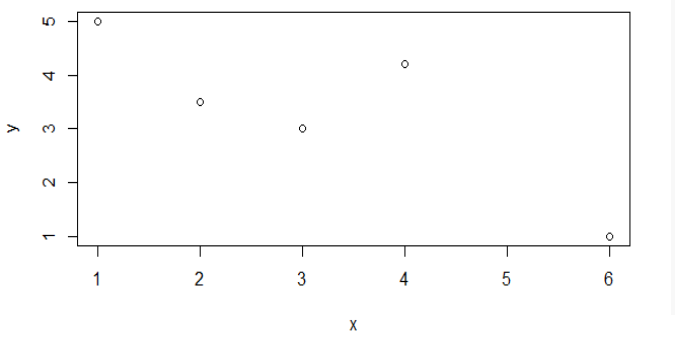
Model u a) je multiplikativni oblik modela, dok se u b) radi o aditivnom obliku modela. Model u a) možemo linearizirati.

Zadatak 14):

y_i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
5	1	-2,2	4,84	1,66	-3,65
1	6	2,8	7,84	-2,34	-6,55
3	3	-0,2	0,04	-0,34	0,07
3,5	2	-1,2	1,44	0,16	-0,19
4,2	4	0,8	0,64	0,86	0,74
16,7	16	0	14,8	0	-9,59

Zadatak 15):

```
y<-c(5,1,3,3.5,4.2)
x<-c(1,6,3,2,4)

plot(x,y)

lm(y~x)

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      5.4243      -0.6514
```



Zadatak 18):

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\begin{bmatrix}1 & 12\\ 1 & 13\\ 1 & 14\\ 1 & 5\\ 1 & 6\\ 1 & 12\\ 1 & 20\\ 1 & 16\end{bmatrix}\cdot\begin{bmatrix}1 & 12\\ 1 & 13\\ 1 & 14\\ 1 & 5\\ 1 & 6\\ 1 & 12\\ 1 & 20\\ 1 & 16\end{bmatrix}'\right)^{-1}\begin{bmatrix}1 & 12\\ 1 & 13\\ 1 & 14\\ 1 & 5\\ 1 & 6\\ 1 & 12\\ 1 & 20\\ 1 & 16\end{bmatrix}'\cdot\begin{bmatrix}17\\ 25\\ 25\\ 8\\ 14\\ 20\\ 40\\ 16\end{bmatrix} = \begin{bmatrix}0,55\\ 1,64\end{bmatrix}$$

Zadatak 19):

x	y	procjena y	rezidualno odstupanje
12	17	20,22	–3,22
13	25	21,85	3,15
14	25	23,49	1,51
5	8	8,74	–0,74
6	14	10,38	3,62
12	20	20,22	–0,22
20	40	33,32	6,68
16	16	26,77	–10,77

Zadatak 20):

Suma rezidualnih odstupanja jednaka je nuli, jer je zbroj vrijednosti u stupcu "rezidualno odstupanje" jednako 0.

Regresijski pravac je $\hat{y}_i = 0,55 + 1,44x_i$, te prosječne vrijednosti zavisne i nezavisne varijable (koje iznose 20,625 i 12,25) zadovoljavaju jednadžbu tog pravca. Prosječna vrijednost procijenjene vrijednosti zavisne varijable (stupac "procjena y") iznosi 20,625 i jednaka je prosječnoj vrijednosti zavisne varijable.

2.1.6. Interpretacija parametara u modelu jednostavne linearne regresije

2.1.6.1. Lin-lin model

Lin-lin model je onaj u kojemu su sve varijable (zavisna i nezavisna) u razinama. Nad varijablama nije izvršena nikakva transformacija, stoga se interpretacija vrši u mjernim jedinicama svake varijable. Ako se razmotri procijenjeni model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \,, \tag{2.97}$$

interpretacija $\hat{\beta}_1$ odnosi se na granični efekt utjecaja jedinične promjene nezavisne varijable na promjenu zavisne varijable (derivacija funkcije po toj varijabli). Dakle, ako se funkcija u (2.97) derivira po varijabli x_i, dobiva se $\frac{d\hat{y}_i}{dx_i} = \hat{\beta}_1$ i interpretacija $\hat{\beta}_1$ je sljedeća: ako se varijabla x_i poveća za jednu jedinicu, tada se u prosjeku varijabla y_i promjeni za $\hat{\beta}_1$ jedinica (poveća ili smanji, ovisi o predznaku). Sada se naglašava „u prosjeku“, s obzirom da se procijenjeni model (2.97) odnosi na sam regresijski pravac koji predstavlja očekivane vrijednosti zavisne varijable (očekivane ili prosječne).

Primjer 2.6.
Za procijenjeni model $\hat{y}_i = 5 + 200x_i$, varijabla x odnosi se na broj godina radnog staža, a varijabla y na dohodak zaposlenika u kn. Tumačenje koeficijenta uz varijablu x_i je sljedeće: ako se broj godina radnog staža zaposlenika poveća za 1 godinu, tada se dohodak zaposlenika poveća u prosjeku za 200 kn.

2.1.6.2. Lin-log model

Lin-log model je onaj kod kojeg je zavisna varijabla u razinama, dok je nezavisna logaritmirana. Ako se razmotri procijenjeni model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln x_i \,, \tag{2.98}$$

diferencijal jednadžbe (2.98) s obzirom na varijablu x_i je:

$$d\hat{y}_i = \hat{\beta}_1 \frac{1}{x_i} dx_i \,/\, \cdot \frac{x_i}{dx_i} \Rightarrow \frac{d\hat{y}_i}{\frac{dx_i}{x_i}} = \hat{\beta}_1 \,, \tag{2.99}$$

pri čemu se posljednja jednakost u (2.99) se može interpretirati kao koeficijent elastičnosti. Pritom promjenu nezavisne varijable mjerimo u postotcima, dok zavisne u mjernim jedinicama, i pritom se interpretacija vrši kao $\frac{\hat{\beta}_1}{100}$ jedinica. Ako se varijabla x_i poveća za 1%, tada se u prosjeku zavisna varijabla poveća/smanji za $\frac{\hat{\beta}_1}{100}$ jedinica.

Dijelimo sa 100 kako bismo dobili $\frac{d\hat{y}_i}{100\frac{dx_{1i}}{x_{1i}}} = \frac{\hat{\beta}_1}{100}$. U brojniku desne strane jednakosti nalazi se promjena zavisne varijable u mjernim jedinicama, dok je u nazivniku stopa rasta nezavisne

varijable. Zato desnu stranu jednakosti interpretiramo kao promjenu zavisne varijable za $\frac{\hat{\beta}_1}{100}$ jedinica ako se nezavisna varijabla poveća za 1%.

Primjer 2.7.
Razmatra se procijenjeni model $\hat{y}_i = 10 + 1400 \ln x_i$, varijabla x odnosi se na broj godina radnog staža, a varijabla y na dohodak zaposlenika u kn. Tumačenje koeficijenta uz varijablu x_i: ako se broj godina radnog staža zaposlenika poveća za 1%, tada se dohodak zaposlenika poveća u prosjeku za 14 kn.

2.1.6.3. Log-lin model

Log-lin model je onaj u kojemu je zavisna varijabla logaritmirana, a nezavisna je u razinama. Ako se razmotri procijenjeni model

$$\widehat{\ln y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad (2.100)$$

diferencijal jednadžbe (2.100) po varijabli x_i je

$$d(\widehat{\ln y_i}) = \hat{\beta}_1 d(x_{i1}) \Rightarrow \frac{\hat{dy}_i}{\hat{y}_i} = \hat{\beta}_1 dx_{i1} \Rightarrow \hat{\beta}_1 = \frac{\frac{\hat{dy}_i}{\hat{y}_i}}{dx_{i1}}. \qquad (2.101)$$

Posljednja jednakost se može interpretirati kao koeficijent elastičnosti, pri čemu promjenu nezavisne varijable mjerimo u njenim mjernim jedinicama, dok se promjena zavisne varijable mjeri u postotcima, pri čemu je $\hat{\beta}_1 \cdot 100\%$. Ako se x_i poveća za jednu mjernu jedinicu, tada se zavisna varijabla poveća/smanji u prosjeku za $\hat{\beta}_1 \cdot 100\%$.

U ovome slučaju parametar $\hat{\beta}_1$ množi se sa 100%, kako bismo dobili izraz $\frac{100\% \cdot \frac{d\hat{y}_i}{\hat{y}_i}}{dx_i} = 100\% \cdot \hat{\beta}_1$. U brojniku lijeve strane jednakosti nalazi se stopa rasta, dok je u nazivniku promjena izražena u mjernim jedinicama varijable x_i. Zato desnu stranu jednakosti interpretiramo kao postotnu promjenu zavisne varijable ako se nezavisna varijabla poveća za jednu jedinicu.

Primjer 2.8.
Procijenjen je model $\widehat{\ln y_i} = 0{,}2 + 0{,}02\, x_i$, varijabla x odnosi se na broj godina radnog staža, a varijabla y na dohodak zaposlenika u kn. Tumačenje koeficijenta uz varijablu x_i: ako se broj godina radnog staža zaposlenika poveća za 1 godinu, tada se dohodak zaposlenika poveća u prosjeku za 2%.

2.1.6.4. Log-log model

Log-log model je onaj u kojemu su sve varijable argumenti logaritamske funkcije. Ako se razmotri model

$$\widehat{\ln y_i} = \hat{\beta}_0 + \hat{\beta}_1 \ln x_i, \qquad (2.102)$$

diferencijal funkcije (2.102) po varijabli ($\ln x_i$) je

$$d\left(\widehat{\ln y_i}\right) = \hat{\beta}_1 d\left(\ln x_{1i}\right) \Rightarrow \frac{1}{\hat{y}_i} d\hat{y}_i = \hat{\beta}_1 \frac{1}{x_i} d\,x_i \Rightarrow \frac{d\hat{y}_i}{d\,x_i}\frac{x_i}{\hat{y}_i} = \hat{\beta}_1 \qquad (2.103)$$

U posljednjoj jednakosti prepoznajemo koeficijent elastičnosti $\left(E_{y,x} = \frac{dy}{dx}\cdot\frac{x}{y}\right)$, pa je u ovome slučaju interpretacija parametra $\hat{\beta}_1$: Ako se varijabla x_i poveća za 1%, tada se u prosjeku zavisna varijabla poveća/smanji za $\hat{\beta}_1$ %.

Primjer 2.9.
Procijenjen je model $\widehat{\ln y_i} = 0{,}2 + 0{,}8\ln x_i$, varijabla x odnosi se na broj godina radnog staža, a varijabla y na dohodak zaposlenika u kn. Tumačenje koeficijenta uz varijablu x_i: ako se broj godina školovanja zaposlenika poveća za 1%, tada se dohodak zaposlenika poveća u prosjeku za 0.8%.

Sva četiri slučaja interpretacije parametara sažeta su u Tablici 2.6.

Tablica 2.6. Interpretacija parametara u regresijskom modelu – sažetak

Model	**Interpretacija**
LIN-LIN	Mjerne jedinice svih varijabli
LOG-LOG	Postotci svih varijabli
LOG-LIN	y – postotak ($\hat{\beta}_1 \cdot 100\%$), x – mjerne jedinice
LIN-LOG	y – mjerna jedinica ($\frac{\hat{\beta}_1}{100}$ jedinica), x – postotak

Primjer 2.10.
Učitajmo datoteku „**BDP_i_HICP.txt**“ u RStudio. Datoteka sadrži podatke za 2019. godinu, za odabrane Europske zemlje o bruto domaćem proizvodu (BDP, GDP stupac, engl. *gross domestic product*, tekuće cijene, u milijunima eura), te varijabli opća razina cijena (HICP stupca, engl. *harmonized index of consumer prices*, indeks, 2015. godina bazna). Procijenimo lin-lin, log-log, log-lin te lin-log model, pri čemu je zavisna varijabla BDP i interpretirajmo koeficijent uz nezavisnu varijablu.

Slika 2.25. predočava naredbe potrebne da bi se podaci unijeli u RStudio te da bi se procijenila 4 modela, dok su rezultati predočeni na slici 2.26.

```
podaci<-read.table("BDP_i_HICP.txt",header=T, sep="\t")
m1<-lm(GDP~HICP,data=podaci)
m2<-lm(log(GDP)~log(HICP),data=podaci)
m3<-lm(log(GDP)~HICP,data=podaci)
m4<-lm(GDP~log(HICP),data=podaci)

m1$coefficients
m2$coefficients
m3$coefficients
m4$coefficients
```

Slika 2.25. Naredbe potrebne za procjene 4 modela

```
m1$coefficients

## (Intercept)        HICP
##  2216385.89   -15790.89

m2$coefficients

## (Intercept)   log(HICP)
##   28.929835   -3.590287

m3$coefficients

## (Intercept)        HICP
## 15.98955011 -0.03587012

m4$coefficients

## (Intercept)   log(HICP)
##     7769301    -1549691
```

Slika 2.26. Ispis koeficijenata sva 4 modela

Redom su procijenjeni modeli lin-lin, log-log, log-lin te lin-log model:

$$M1\colon \hat{y}_i = 2216385{,}89 - 15790{,}89\, x_i$$

$$M2\colon \widehat{\ln y}_i = 28{,}93 - 3{,}59 \ln x_i$$

$$M3\colon \widehat{\ln y}_i = 15{,}99 - 0{,}04 x_i$$

$$M4\colon \hat{y}_i = 7759301 - 1549691 \ln x_i$$

Te su interpretacije redom kako slijedi. M1: ako se vrijednost HICP indeksa poveća za jedan indeksni bod, vrijednost BDP-a će se smanjiti u prosjeku za 15790,89 milijuna eura. M2: ako se vrijednost HICP indeksa poveća za 1%, vrijednost BDP-a će se smanjiti u prosjeku za 3,59%. M3: ako se vrijednost HICP indeksa poveća za jedan indeksni bod, vrijednost BDP-a će se smanjiti u prosjeku za 4%. M4: ako se vrijednost HICP indeksa poveća za 1%, vrijednost BDP-a će se smanjiti u prosjeku za 15496,91 milijuna eura.

2.1.6.5. Napomena o konstanti u modelu jednostavne linearne regresije

Iako često konstanta u modelu jednostavne linearne regresije (ali i višestruke) nema ekonomsko značenje, ostavlja se u modelu prilikom procjene iz razloga što uklanjanje konstante može **rezultirati precijenjenim ili podcijenjenim vrijednostima ostalih parametara** u modelu. Slika 2.27. prikazuje što se događa ako se za dane podatke o zavisnoj i nezavisnoj varijabli za koje je potrebno procijeniti model uz uključenu konstantu isključi ta konstanta. Očito je da će regresijski pravac koji sadrži ishodište lošije opisivati vezu između obje varijable, pri čemu je u ovome slučaju nagib pravca predočenog iscrtanom linijom veći u odnosu na nagib pravca predočen punom crnom linijom kada je konstanta uključena. Kažemo da smo nagib pravca u slučaju iscrtane linije precijenili (veći je od nagiba pune crne linije). Model koji je predočen punom crnom linijom glasi: $\hat{y}_i = 9{,}985 + 2{,}029 x_i$, dok je procijenjen model za iscrtanu liniju sljedeći: $\hat{y}_i = 5{,}021 x_i$. Uočavamo kako je procijenjen parametar uz nezavisnu varijablu puno veći u odnosu na početnu vrijednost.

Sama interpretacija konstante ovisi o jednome od četiri prethodno obrađena slučaja modela. Ako se radi o lin-lin modelu, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, interpretacija $\hat{\beta}_0$ se odnosi na prosječnu vrijednost zavisne varijable kada bi vrijednost nezavisne varijable iznosila 0:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0, \tag{2.104}$$

dok je za slučaj log-log modela, $\widehat{\ln y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln x_i$, vrijedi:

$$\widehat{\ln y}_i = \hat{\beta}_0 + \hat{\beta}_1 \underbrace{lnx_i}_{=0}, \tag{2.105}$$

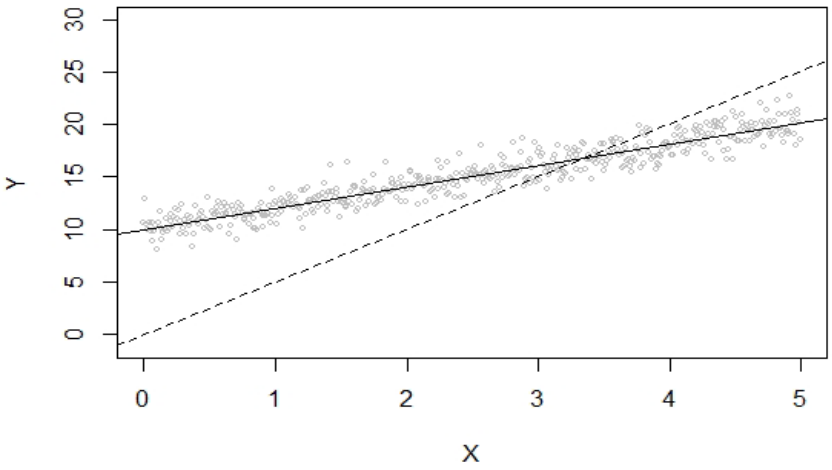
a to je kada je vrijednost $x_i = 1$. Dakle, za vrijednost nezavisne varijable jednaka 1, u prosjeku će vrijediti $\widehat{\ln y}_i = \hat{\beta}_0$. Stoga je $\hat{y}_i = e^{\hat{\beta}_0}$ u prosjeku vrijednost zavisne varijable ako je vrijednost nezavisne varijable jednaka 1. Nadalje, za slučaj log-lin, $\widehat{\ln y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, vrijedi:

$$\widehat{\ln y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0, \tag{2.106}$$

pa je vrijednost zavisne varijable u prosjeku jednaka $\hat{y}_i = e^{\hat{\beta}_0}$ ako je vrijednost nezavisne varijable jednaka 0. Konačno, za lin-log model, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln x_i$, vrijedi:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \underbrace{\ln x_i}_{=0} \tag{2.107}$$

pa je vrijednost zavisne varijable u prosjeku jednaka $\hat{\beta}_0$ ako je vrijednost nezavisne varijable jednaka 1.



Slika 2.27. Usporedba regresijskog pravca kroz ishodište (iscrtana linija) sa regresijskim pravcem uz konstantni član (puna linija)

Primjer 2.11.
Temeljem podataka iz prethodnog primjera (datoteka „**BDP_i_HICP.txt**“), interpretirajmo vrijednosti konstanti za sva 4 modela.

$$M1:\ \hat{y}_i = 2216385,89 - 15790,89x_i$$

$$M2:\ \widehat{\ln y_i} = 28,93 - 3,59 \ln x_i$$

$$M3:\ \widehat{\ln y_i} = 15,99 - 0,04x_i$$

$$M4:\ \hat{y}_i = 7759301 - 1549691 \ln x_i$$

U M1: Kada bi vrijednost HICP indeksa iznosila 0, u prosjeku bi BDP zemalja iznosio 2216385,89 milijuna eura. M2: kada bi vrijednost HICP indeksa iznosila 1 indeksni bod, u prosjeku bi BDP zemalja iznosio $e^{28.93}$ milijuna eura. M3: kada bi vrijednost HICP indeksa iznosila 0, u prosjeku bi BDP zemalja iznosio $e^{15.99}$ milijuna eura, dok M4: kada bi vrijednost HCIP indeksa iznosila 1, u prosjeku bi BDP zemalja iznosio 7759301 milijuna eura.

2.1.6.6. Interpretacija parametara u regresijskom modelu sa standardiziranom varijablom

Regresijskoj analizi moguće je pristupiti na način da se postigne veća usporedivost zavisne i nezavisne varijable, na način da se obje standardiziraju i da se potom rezultati tumače u terminima standardnih devijacija. U slučaju standardizacije varijabli, očekivana vrijednost obiju iznosi 0 pa se konstanta ne uključuje u model.

Dakle, najprije se standardiziraju zavisna i nezavisna varijabla:

$$y_i^* = \frac{y_i - \overline{y}}{\sigma_y}, x_i^* = \frac{x_i - \overline{x}}{\sigma_x}, \tag{2.108}$$

te se potom procijeni model:

$$\hat{y}_i^* = \hat{\beta}_1^* x_i^*. \tag{2.109}$$

U slučaju jednadžbe (2.109) se vrijednost parametra $\hat{\beta}_1^*$ interpretira na sljedeći način: ako se vrijednost nezavisne varijable poveća za 1 **standardnu devijaciju**, vrijednost zavisne varijable će se promijeniti u prosjeku za $\hat{\beta}_1^*$ **standardnih devijacija**.

Primjer 2.12.
Temeljem podataka iz prethodnog primjera o BDP-u i HICP-u odabranih Europskih zemalja, procijenimo standardizirani model u RStudiju i interpretirajmo rezultat.

U RStudiju naredba za linearni model je ponovno lm(…), no za standardizaciju varijabla koristi se naredba scale(…). Stoga se koristi naredba:

lm(scale(GDP)~0+scale(HICP),data=podaci),

gdje se u desnoj strani jednadžbe dodaje vrijednost 0 kako bi se procijenio model bez konstante (odnosno konstanta je jednaka 0). Rezultat je prikazan na slici 2.28. Uočava se da je procijenjeni model sljedeći: $\hat{y}_i^* = -0,053x_i^*$. Ako se vrijednost HICP indeksa poveća za jednu standardnu devijaciju, vrijednost BDP-a će se smanjiti u prosjeku za 0,053 standardnih devijacija.

```
lm(scale(GDP)~0+scale(HICP),data=podaci)

##
## Call:
## lm(formula = scale(GDP) ~ 0 + scale(HICP), data = podaci)
##
## Coefficients:
## scale(HICP)
##    -0.05293
```

Slika 2.28. Rezultat procjene standardiziranog modela

2.1.7. Intervalna procjena parametara jednostavne linearne regresije

Procjena parametara u jednostavnome linearnome regresijskom modelu dana formulama (2.26) i (2.29) naziva se **procjena jednim brojem** (engl. *point estimate*). Osim procjene jednim brojem, može se izvršiti **intervalna procjena parametara**. Radi se o procjeni donje i gornje granice intervala (engl. CI, *confidence intervals*), koji će za zadanu razinu pouzdanosti uključivati stvarnu vrijednost parametra koji se procjenjuje. Općenito se za intervalnu procjenu parametra θ razmatra pristup:

$$\hat{\theta} \pm \text{varijabilnost uzorkovanja}. \tag{2.110}$$

Kako temeljem pretpostavki regresijskog modela standardizirane vrijednosti (oznake t_0 i t_1) procijenjenih parametara $\hat{\beta}_0$ i $\hat{\beta}_1$ slijede Studentovu distribuciju s N–2 stupnja slobode:

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{SE\left(\hat{\beta}_0\right)} \sim t(N-2) \text{ i } t_1 = \frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)} \sim t(N-2), \tag{2.111}$$

čijom jednostavnom manipulacijom se konstruiraju intervalne procjene:

$$P\left(-t_{\gamma/2} < t_0 < t_{\gamma/2}\right) = 1-\gamma \tag{2.112}$$

i

$$P\left(-t_{\gamma/2} < t_1 < t_{\gamma/2}\right) = 1-\gamma \tag{2.113}$$

gdje je $1-\gamma$ **pouzdanost procjene**, dok $t_{\gamma/2}$ predstavlja koeficijent pouzdanosti, tj. vrijednost Studentove distribucije s N–2 stupnja slobode. Ta vrijednost se za zadanu razinu $1-\gamma$ i broj stupnjeva slobode iščitava iz tablice kritičnih vrijednosti Studentove distribucije. Uočimo da se radi o raspodjeli pouzdanosti na gornju i donju granicu, stoga se vrijednost γ dijeli s 2 (vidjeti sliku 2.29.).

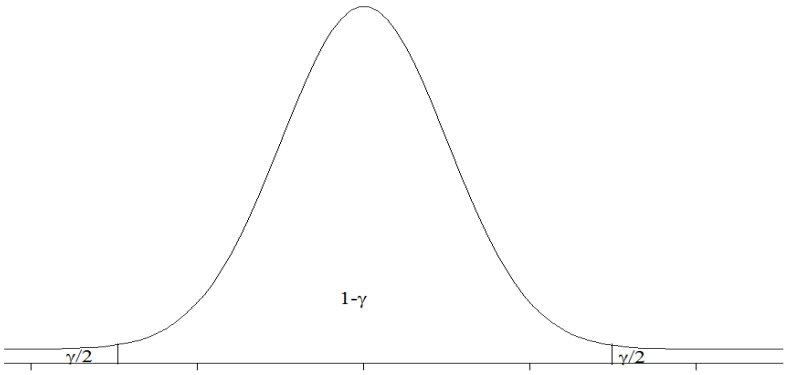
Relacije (2.112) i (2.113) mogu se zapisati na način da se u njih uvrste supstitucije iz (2.111):

$$P\left(\hat{\beta}_0 - t_{\gamma/2}SE\left(\hat{\beta}_0\right) < \beta_0 < \hat{\beta}_0 + t_{\gamma/2}SE\left(\hat{\beta}_0\right)\right) = 1-\gamma \tag{2.114}$$

i

$$P\left(\hat{\beta}_0 - t_{\gamma/2}SE\left(\hat{\beta}_1\right) < \beta_1 < \hat{\beta}_0 + t_{\gamma/2}SE\left(\hat{\beta}_1\right)\right) = 1-\gamma \tag{2.115}$$

Značenje intervala pouzdanosti je sljedeće (Wooldridge, 2016). Kada bi se slučajni uzorci prikupljali iz populacije, pri čemu bi se za svaki uzorak računala donja i gornja granica intervala (2.114), odnosno (2.115), tada bi nepoznate vrijednosti β_0 i β_1 pripadale intervalima u tim relacijama u slučaju $(1-\gamma)$% uzoraka. S obzirom da u stvarnosti prikupljamo samo jedan uzorak za procjenu parametara, ne znamo jesu li stvarne vrijednosti parametara unutar tih intervala. Uobičajeno se za vrijednost $(1-\gamma)$ uzima 90%, 95% ili 99%.



Slika 2.29. Studentova distribucija s granicama intervala procjene, uz pouzdanost $1-\gamma$

Sama interpretacija relacija (2.114) i (2.115) je sljedeća. Kada bi vrijednost nezavisne varijable u jednostavnom linearnom regresijskom modelu iznosila 0, tada bi u prosjeku vrijednost zavisne varijable iznosila između $\hat{\beta}_0 - t_{\gamma/2}SE\left(\hat{\beta}_0\right)$ i $\hat{\beta}_0 + t_{\gamma/2}SE\left(\hat{\beta}_0\right)$ mjernih jedinica; te kada bi se vrijednost nezavisne varijable povećala za jednu jedinicu, vrijednost zavisne varijable će se promijeniti u prosjeku između $\hat{\beta}_1 - t_{\gamma/2}SE\left(\hat{\beta}_1\right)$ i $\hat{\beta}_1 + t_{\gamma/2}SE\left(\hat{\beta}_1\right)$ jedinica. Ako su predznaci obje granice negativni, govorimo o smanjenju, dok pozitivni predznaci govore o povećanju. Međutim, može se dogoditi da se predznak donje i gornje granice razlikuje, stoga bi se govorilo da povećanje nezavisne varijable za jednu jedinicu vodi do u prosjeku smanjenja vrijednosti zavisne varijable za vrijednost $\hat{\beta}_1 - t_{\gamma/2}SE\left(\hat{\beta}_1\right)$ do povećanja za vrijednost $\hat{\beta}_1 + t_{\gamma/2}SE\left(\hat{\beta}_1\right)$ (može se, radi jednostavnosti, govoriti o promjeni zavisne varijable za između donje i gornje vrijednosti). U tom slučaju treba uočiti da je u tom intervalu uključena i vrijednost 0, što znači i da interpretacija da promjena nezavisne varijable nema učinka na zavisnu varijablu također ulazi u obzir. Često se u tom slučaju u analizi utvrdi kako nezavisna varijabla nije značajna u modelu[18].

Primjer 2.13.
Temeljem podataka iz prethodnog primjera o BDP-u i HICP-u odabranih Europskih zemalja, procijenimo sve četiri specifikacije modela iz naslova 2.1.6., te za procijenjene modele napišimo intervalne procjene parametra na razini pouzdanosti od 95% i interpretirajmo.

Temeljem naredbi prikazanih na slici 2.30., dobiveni su rezultati na slici 2.31., koje možemo zapisati na način:

[18] O testiranju značajnosti varijabli u modelu, vidjeti naslov 2.1.9.1.

$$M1: P(-126871,4 < \beta_1 < 95289,59) = 0,95$$
$$M2: P(-24,89 < \beta_1 < 17,71) = 0,95$$
$$M3: P(-0,24 < \beta_1 < 0,17) = 0,95$$
$$M4: P(-13303910 < \beta_1 < 10204528) = 0,95$$

Interpretacije su redom kako slijedi. M1: Uz razinu pouzdanosti od 95%, ako se vrijednost HICP indeksa poveća za jedan indeksni bod, vrijednost BDP-a će se promijeniti u prosjeku između – 12687,40 milijuna eura do 95289,59 milijuna eura.

```
confint(m1, level=0.95)
confint(m2, level=0.95)
confint(m3, level=0.95)
confint(m4, level=0.95)
```

Slika 2.30. Naredbe potrebne za intervalnu procjenu parametara u sva četiri modela

```
confint(m1,level=0.95)

##                   2.5 %      97.5 %
## (Intercept) -9524310.3 13957082.02
## HICP         -126871.4    95289.59

confint(m2,level=0.95)

##                  2.5 %    97.5 %
## (Intercept) -70.33952 128.19919
## log(HICP)   -24.89296  17.71239

confint(m3,level=0.95)

##                   2.5 %     97.5 %
## (Intercept) -5.2878607 37.2669610
## HICP        -0.2371789  0.1654386

confint(m4,level=0.95)

##                  2.5 %   97.5 %
## (Intercept) -47004746 62543347
## log(HICP)   -13303910 10204528
```

Slika 2.31. Intervalne procjene parametara u sva četiri modela, razina pouzdanosti 95%

M2: Uz razinu pouzdanosti od 95%, ako se vrijednost HICP indeksa poveća za 1%, vrijednost BDP-a će se promijeniti u prosjeku između –24,89% do a 17,71%. M3: Uz razinu pouzdanosti od 95%, ako se vrijednost HICP indeksa poveća za jedan indeksni bod, vrijednost BDP-a će se promijeniti u prosjeku između –23,72% do 16,54%.

M4: Uz razinu pouzdanosti od 95%, ako se vrijednost HICP indeksa poveća za 1%, vrijednost BDP-a će se promijeniti u prosjeku za –133039,10% do 102045,28%. Uočimo da su u ovome primjeru donja i gornja granica intervalne procjene suprotnog predznaka.

Ako se analizira interpretacija konstante u svakome modelu, govorimo o sljedećem. U M1: Kada bi vrijednost HICP indeksa iznosila 0, uz razinu pouzdanosti od 95%, u prosjeku bi BDP zemalja iznosio između -9524310,30 do 13957082,02 milijuna eura.

M2: kada bi vrijednost HICP indeksa iznosila 1 indeksni bod, uz razinu pouzdanosti od 95%, u prosjeku bi BDP zemalja iznosio između $e^{-70.33}$ i $e^{128.20}$ milijuna eura. M3: uz razinu pouzdanosti od 95%, kada bi vrijednost HICP indeksa iznosila 0, u prosjeku bi BDP zemalja iznosio između $e^{-5.29}$ i $e^{37.27}$ milijuna eura, dok za M4: uz razinu pouzdanosti od 95%, kada bi vrijednost HCIP indeksa iznosila 1, u prosjeku bi BDP zemalja iznosio između –47004746 i 62543347 milijuna eura.

Intervalne procjene konstanti možemo pisati kao:

$$M1: P(-9524310,3 < \beta_1 < 13957082,02) = 0,95$$
$$M2: P(-70,34 < \beta_1 < 128,20) = 0,95$$
$$M3: P(-5,29 < \beta_1 < 37,27) = 0,95$$
$$M4: P(-47004746 < \beta_1 < 62543347) = 0,95$$

Primjer 2.14.
Temeljem podataka iz prethodnog primjera o BDP-u i HICP-u odabranih Europskih zemalja, procijenimo sve četiri specifikacije modela iz naslova 2.1.6., te za procijenjene modele napišimo intervalne procjene parametra na razini pouzdanosti od 90% i 99% te usporedimo do kakvih promjena dolazi.

```
confint(m1,level=0.9)

##                   5 %       95 %
## (Intercept) -7540894 11973666.0
## HICP         -108106    76524.2

confint(m2,level=0.9)

##                    5 %      95 %
## (Intercept) -53.56944 111.42911
## log(HICP)   -21.29419  14.11362

confint(m3,level=0.9)

##                     5 %       95 %
## (Intercept) -1.6933585 33.6724587
## HICP        -0.2031708  0.1314305

confint(m4,level=0.9)

##                    5 %     95 %
## (Intercept) -37751485 53290087
## log(HICP)   -11318210  8218827

confint(m1,level=0.99)

##                    0.5 %     99.5 %
## (Intercept) -13592897.2 18025668.9
## HICP          -165364.9   133783.1

confint(m2,level=0.99)

##                   0.5 %    99.5 %
## (Intercept) -104.74004 162.59971
## log(HICP)    -32.27513  25.09456

confint(m3,level=0.99)
```

```
##                   0.5 %     99.5 %
## (Intercept) -12.6612732 44.6403735
## HICP         -0.3069398  0.2351996

confint(m4,level=0.99)

##                  0.5 %   99.5 %
## (Intercept) -65985986 81524587
## log(HICP)   -17377183 14277801
```

Slika 2.32. Intervalne procjene parametara u sva četiri modela, razina pouzdanosti 90% (lijevi panel) i 99% (desni panel)

Promjenom vrijednosti 0.95 u 0.90, odnosno u 0.99 na slici 2.31., dolazi se do intervalnih procjena uz 90% i 99% razina pouzdanosti, čiji su rezultati predočeni na slici 2.32. Uočava se da povećanje razine pouzdanosti vodi do šireg intervala procjene, odnosno da povećanjem razine pouzdanosti, dolazi do povećanje pogreške procjene. Stoga u praksi uvijek postoji određeno pravljenje ustupaka (engl. *trade off*) između razine pouzdanosti i pogreške procjene.

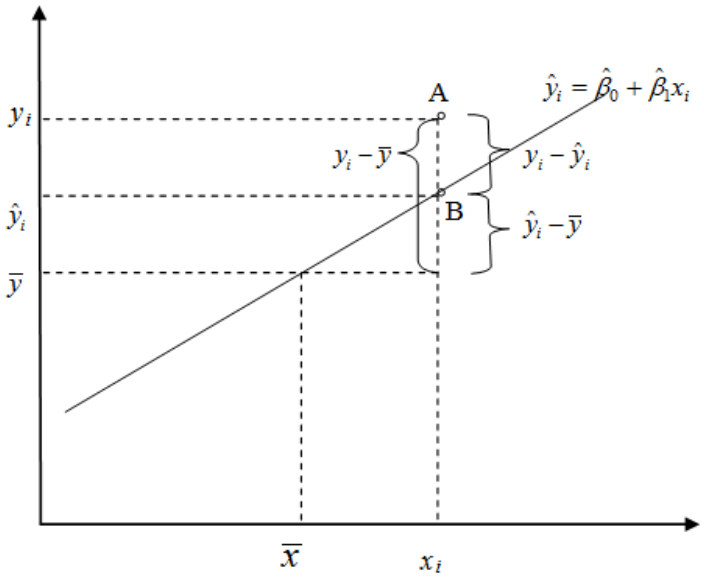
2.1.8. Analiza varijance u modelu jednostavne linearne regresije

Prilikom procjene regresijskog modela, postavlja se pitanje koliko uspješno nezavisna varijabla opisuje zavisnu varijablu (u stranoj literaturi se uobičajeno govori o *goodness of fit*). Stoga je ideja izračunati određene mjere koje nam govore koliko dobro regresijski pravac pojašnjava varijacija (varijancu) zavisne varijable. Uvriježeno je reći da se regresijski pravac dobro prilagođava podacima iz uzorka ako je velik udio proporcije varijance (varijacija) zavisne varijable protumačen regresijskim modelom.

Slika 2.33. predočava procijenjen model, predočen regresijskim pravcem $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, pri čemu se fokusiramo na točku A, čije su koordinate (x_i,y_i), dakle stvarne vrijednosti nezavisne i zavisne varijable. Modelom $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ se procjenjuje da za danu vrijednost x_i, vrijednost zavisne varijable iznosi $\hat{y}_i$, što je predočeno točkom B. Za izračun varijance zavisne varijable, koriste se stvarne vrijednosti i uspoređuju se s prosječnom. Stoga se **odstupanje stvarnih (empirijskih) vrijednosti zavisne varijable od prosjeka** može raščlaniti na odstupanje **protumačeno** regresijskim modelom (odstupanje procijenjene vrijednosti od prosječne) i odstupanje **neprotumačeno** modelom (rezidualno odstupanje):

$$\underbrace{y_i - \overline{y}}_{\text{ukupno odstupanje}} = \underbrace{\hat{y}_i - \overline{y}}_{\substack{\text{odstupanje protumačeno} \\ \text{modelom}}} + \underbrace{y_i - \hat{y}_i}_{\text{rezidualno odstupanje, } \hat{\varepsilon}_i} \tag{2.116}$$

što kad se usporedi sa slikom 2.33, uočava se da je upravo ukupno odstupanje za *i*-to opažanje zavisne varijable, $y_i - \overline{y}$, raščlanjeno na zbroj odstupanja koje je protumačeno modelom ($\hat{y}_i - \overline{y}$) i rezidualnog odstupanja ($\hat{\varepsilon}_i$, koje nismo mogli protumačiti modelom, $y_i - \hat{y}_i$). Htjeli bismo da je odstupanje protumačeno modelom što veće, jer na taj način odabrana nezavisna varijabla dobro objašnjava zavisnu varijablu.



Slika 2.33. Analiza varijance predočena grafički

Da bismo razmotrili varijancu zavisne varijable, potrebno je računati kvadrate odstupanja u (2.116) i zbrojiti:

$$\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{N}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2, \qquad (2.117)$$

gdje je izraz s lijeve strane jednakosti nazvan **ukupna suma kvadrata** (engl. *total sum of squares*, *SST*), koja se dijeli na zbroj **sume kvadrata odstupanja regresijskih vrijednosti od prosjeka** (engl. *explained sum of squares*, *SSE*) i **sume kvadrata rezidualnih odstupanja** (engl. *residual sum of squares*, *SSR*):

$$SST = SSE + SSR. \qquad (2.118)$$

Ako se zbrojevi kvadrata podijele s odgovarajućim stupnjevima slobode, dobivaju se **sredine kvadrata** koje su nezavisne procjene udjela u ukupnoj varijanci. Taj postupak se može predočiti u **tablici analize varijance** (engl. *ANOVA, analysis of variance*), predočeno u tablici 2.7. U prvome stupcu, sume kvadrata *SSE* i *SSR* u zbroju daju *SST*. Stupanj slobode[19] za slučaj *SSE* iznosi 1 jer se radi o jednoj nezavisnoj varijabli u modelu, dok je za slučaj *SSR* broj stupnjeva slobode jednak $N-2$. Ako se svaka suma kvadrata podijeli odgovarajućim stupnjevima slobode, u trećem stupcu dobivaju se sredine kvadrata odstupanja, odnosno $SSE/1$ i $SSR/(N-2)$. Sredina $SSE/1$ trebala bi biti veća u odnosu na sredinu $SSR/(N-2)$, jer će to značiti da je odabranom nezavisnom varijablom protumačen veći udio ukupne sume kvadrata. Stoga omjer tih dviju sredina, predstavljen *F*-omjerom[20] treba biti velika vrijednost. *F*-omjer služi za testiranje skupne značajnosti u regresijskom modelu, o čemu će detaljnije biti u 2.1.9.2.

[19] O stupnjevima slobode vidjeti u poglavlju Dodatak 5.3.
[20] Vidjeti poglavlje 2.1.9.2.

Tablica 2.7. Tablica analize varijance, model jednostavne linearne regresije

Izvor varijacije	Sume kvadrata	Stupnjevi slobode (*ss*)	Sredina kvadrata odstupanja	*F*-omjer
Regresija (protumačeno modelom)	*SSE*	1	*SSE*/1	$\frac{SSE/1}{SSR/(N-2)}$
Rezidualna odstupanja (neprotumačeno modelom)	*SSR*	*N*–2	*SSR*/(*N*–2)	
Ukupno	*SST*	*N*–1		

Kako je procjena varijance regresije dana formulom (2.73), $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N-2}$, odnosno:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N-2} = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-2}, \tag{2.119}$$

uočava se da je brojnik $\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ upravo suma kvadrata rezidualnih odstupanja (*SSR*). Stoga je **procjena standardne devijacije regresije** (engl. RMSE, *root mean squared error*) jednaka

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-2}} = \sqrt{\frac{\mathrm{SSR}}{N-2}}, \tag{2.120}$$

čija je interpretacija sljedeća: prosječno odstupanje stvarnih (empirijskih) vrijednosti zavisne varijable od regresijskih (procijenjenih), izražena u mjernim jedinicama zavisne varijable. Stoga se naziva **apsolutna mjera disperzije**. Naravno, poželjno je da je ova mjera što manja Nedostatak procijenjene standardne devijacije regresije je taj što je dana u mjernim jedinicama zavisne varijable, stoga se ne može komentirati je li to odstupanje značajno ili ne. Zato se uz nju stoga koristi i **relativna mjera disperzije**, **procjena koeficijenta varijacije**:

$$\hat{V} = \frac{\hat{\sigma}}{\bar{y}} 100\%, \tag{2.121}$$

te se interpretira kao prosječno odstupanje stvarnih vrijednosti zavisne varijable od regresijskih, izraženo postotkom. Također je poželjna što manja vrijednost ove mjere.

Ako se ponovno promotri relacija (2.118)

$$SST = SSE + SSR, \tag{2.122}$$

te se podijeli ukupnom vrijednošću *SST*:

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}, \tag{2.123}$$

tada omjer $\frac{SSE}{SST}$ predstavlja udio protumačenih odstupanja u ukupnoj sumi kvadrata odstupanja, i naziva se **koeficijent determinacije**, oznaka R^2 (engl. *coefficient of determination, R-squared*):

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\hat{\sigma}^2 (N-2)}{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2}, \tag{2.124}$$

Udio protumačenih odstupanja u ukupnoj sumi kvadrata odstupanja kreće se između 0 i 1 (tj. između 0% i 100%), stoga je R^2 **bolji što mu je veća vrijednost, tj. bliža jediničnoj vrijednosti**[21]. Intuitivno je zaključiti da je u intervalu [0,1], jer ako razmatramo dva krajnja slučaja, prvi u kojemu odabrani model ne objašnjava uopće varijaciju zavisne varijable, tada je *SSE* jednaka 0, pa je u (2.124) brojnik jednak 0 i posljedično tome, R^2 jednak je 0. S druge strane, ako modelom u potpunosti objašnjavamo varijaciju zavisne varijable, tada je *SSE* jednaka 1 i posljedično tome u formuli (2.124) je zbog *SSR*-a koja tada mora biti jednaka 0 (zbog formule (2.123)), pa preostaje 1−0, što je jednako 1, čemu je upravo jednaka vrijednost R^2. U slučaju kada je koeficijent determinacije jednak 1, to bi značilo da bi **sve** točke koje se razmatraju na dijagramu rasipanja, čije su koordinate stvarne vrijednosti nezavisne i zavisne varijable, pripadale regresijskom pravcu.

Drugi krajnji slučaj, kada je koeficijent determinacije jednak 0, kako model ne objašnjava varijacije zavisne varijable, nezavisna varijabla uopće ne utječe na zavisnu varijablu, pa se u tom slučaju može zamisliti regresijski pravac koji je horizontalna linija, za koju vrijedi da promjene nezavisne varijable ne vode promjeni zavisne. Uz R^2, u analizama se obično razmatra i **korigirani koeficijent determinacije**, $\bar{R}^2$ (engl. *adjusted R-squared*), koji se računa kao[22]:

$$\bar{R}^2 = 1 - \frac{N-1}{N-2}(1 - R^2), \tag{2.125}$$

koji se koristi u slučaju uspoređivanja više modela koji imaju različit broj nezavisnih varijabli. Detaljnije o razlogu korištenja korigiranog koeficijenta determinacije će se razmotriti u naslovu 2.2.6.

Može se pokazati[23] da je koeficijent determinacije jednak kvadratu koeficijenta korelacije između empirijskih i procijenjenih vrijednosti zavisne varijable (od tuda naziv R-kvadrat, dok se slovo R koristi za oznaku koeficijenta korelacije populacije, pri čemu je u praksi ostala oznaka i R za regresijsku analizu, Wooldridge (2016: 35)).

[21] Valja napomenuti da iako veća vrijednost koeficijenta determinacije ima spomenutu interpretaciju da je veći udio varijacija zavisne varijable pojašnjen modelom, velike vrijednosti tog koeficijenta (blizu jedinične) u empirijskim istraživanjima upućuju da je narušena neka od pretpostavki regresijskog modela (vidjeti odjeljak 2.1.2). Često se radi o problemu autokorelacije koji rezultira s velikom vrijednošću R^2. Zato se u empirijskim istraživanjima više pažnje posvećuje testiranju pretpostavki samog modela, te se provjerava vrijednost tog koeficijenta, ali on ne predstavlja temelj za interpretaciju kvalitete modela.

[22] Za izvod vidjeti Wooldridge (2016: 181).

[23] Vidjeti Greene (2018: 44).

Nadalje, za slučaj jednostavnog linearnog regresijskog modela, definira se **koeficijent jednostavne linearne korelacije**, koji mjeri smjer i jakost linearne povezanosti između zavisne i nezavisne varijable. Računa se na sljedeći način:

$$R = \pm\sqrt{R^2},\ sign(R) = sign(\hat{\beta}_1), \tag{2.126}$$

što znači da se radi o drugome korijenu koeficijenta determinacije, pri čemu **predznak** koeficijenta jednostavne linearne korelacije ovisi o predznaku procijenjenog parametra uz nezavisnu varijablu. Napomenimo da se samo u slučaju jednostavne linearne regresije može govoriti o smjeru spomenute korelacije, s obzirom da se radi o korelaciji između dvije varijable (nezavisna i zavisna). Ovo neće vrijediti u slučaju višestrukog linearnog regresijskog modela (vidjeti naslov 2.2.6), gdje će se samo jačina korelacije moći interpretirati, ali ne i smjer. Vrijednost koeficijenta jednostavne linearne korelacije nalazi se u intervalu [−1,1]. Što je njegova apsolutna vrijednost bliža jediničnoj vrijednosti, jača je korelacija između razmatranih varijabli. Pozitivan predznak, naravno, upućuje na pozitivnu korelaciju, dok negativan predznak upućuje na negativnu korelaciju između razmatranih varijabli. Vrijednost koja je blizu 0 upućuje na odsustvo korelacije između razmatranih varijabli.

Valja napomenuti kako (korigirani) koeficijent determinacije može poprimiti i **negativnu vrijednost**, s obzirom da se njegov izračun može predočiti i kao:

$$\overline{R}^2 = 1 - \frac{N-1}{N-2}(1-R^2) = 1 - \frac{N-1}{N-2}\frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{N-1}{N-2}\frac{(N-2)\hat{\sigma}^2}{(N-1)\hat{\sigma}_y^2} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}, \tag{2.127}$$

gdje se uočava da procjena varijance zavisne varijable, $\hat{\sigma}_y^2$, ne ovisi o odabranom modelu, stoga korigirani koeficijent determinacije ovisi o procjeni varijance regresije ($\hat{\sigma}^2$). Veća vrijednost $\hat{\sigma}^2$ vodi smanjenju vrijednosti $\overline{R}^2$.

Primjer 2.15.
Procijenimo jednostavni linearni regresijski model između zavisne varijable BDP i nezavisne varijable HICP (datoteka „**BDP_i_HICP.txt**“). Sastavimo tablicu ANOVA, izračunajmo koeficijent determinacije, korigirani koeficijent determinacije, procjenu standardne devijacije regresije, procjenu koeficijenta varijacije regresije te koeficijent jednostavne linearne korelacije i potom interpretirajmo rezultat.

Ako se najprije pomoću naredbe m1<–lm(GDP~HICP, data=podaci) spremi linearni model, pomoću naredbe sazetak<–summary(m1) i ispisom spremljenog objekta „sazetak“ dobiva se detaljni ispis prikazan na slici 2.34., gdje se uočava da je koeficijent determinacije jednak 0,0028 (u ispisu *Multiple R-squared*), dok korigirani koeficijent determinacije iznosi –0,0304. Stoga je ovim modelom objašnjeno 0,28% varijacija varijable BDP. U ovome primjeru dobili smo negativnu vrijednost korigiranog koeficijenta determinacije, što je prethodno spomenuto kao mogući problem u regresijskoj analizi. Ako se fokusiramo i samo na 0,28%, može se zaključiti kako model nije reprezentativan jer je samo 0,28% varijacija zavisne varijable pojašnjeno ovim modelom.

```
##
## Call:
## lm(formula = GDP ~ HICP, data = podaci)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -603036 -457350 -315196  -39255 2898603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2216386    5748839   0.386    0.703
## HICP          -15791      54391  -0.290    0.774
##
## Residual standard error: 852700 on 30 degrees of freedom
## Multiple R-squared:  0.002802,   Adjusted R-squared:  -0.03044
## F-statistic: 0.08429 on 1 and 30 DF,  p-value: 0.7736
```

Slika 2.34. Ispis procijenjenog linearnog regresijskog modela

Nadalje, procjena standardne devijacije regresije iznosi 852701,50 (koristi se naredba sazetak$sigma) te se interpretira da je prosječno odstupanje empirijskih od regresijskih vrijednosti varijable BDP 852701,50 milijuna eura. Koeficijent varijacije regresije procjenjuje se izrazom prikazanim na slici 2.35., gdje su i ostale prethodno spomenute naredbe. Rezultat dobiven za procjenu koeficijenta varijacije regresije iznosi 155,62, te se interpretira da je prosječno odstupanje empirijskih od regresijskih vrijednosti varijable BDP 155,62%, čime se također može zaključiti kako model nije reprezentativan.

```
sazetak<-summary(m1)
sazetak$sigma

## [1] 852701.5

GDP<-podaci$GDP
koef_v<-(sazetak$sigma/mean(GDP))*100
koef_v

## [1] 155.6208
```

Slika 2.35. Procjena standardne devijacije i koeficijenta varijacije regresije

Konačno, tablica ANOVA dobije se naredbom anova(m1), pri čemu se dobiva ispis prikazan na slici 2.36. Uočava se kako je broj stupnjeva slobode (stupac DF, engl. *degrees of freedom*) 1 za jednu nezavisnu varijablu, dok je 30 (što je jednako N–2) stupnjeva slobode za rezidualna odstupanja. Stupac naziva *Sum Sq* prikazuje sume kvadrata odstupanja koji su protumačeni modelom i neprotumačeni[24], dok su vrijednosti u stupcu *Mean Sq* dobivene dijeljenjem sume kvadrata odstupanja s odgovarajućim stupnjevima slobode. Konačno, vrijednost 0.08 u stupcu *F value* dobivena je kao omjer vrijednosti u stupcu *Mean Sq*, dok pripadajuća *p*-vrijednost iznosi 0,77[25]. I iz tablice ANOVA se može zaključiti da model nije reprezentativan jer je sredina kvadrata odstupanja koja je protumačena modelom manja od sredine kvadrata rezidualnih odstupanja (neprotumačena modelom) manja!

[24] Vrijednost 6,13e+10 je znanstveni zapis programa, to je vrijednost $6{,}13 \cdot 10^{10}$
[25] Vidjeti detalje o *p*-vrijednosti u 2.1.9. i 2.1.9.7.

```
## Analysis of Variance Table
##
## Response: GDP
##           Df     Sum Sq    Mean Sq F value Pr(>F)
## HICP       1 6.1286e+10 6.1286e+10  0.0843 0.7736
## Residuals 30 2.1813e+13 7.2710e+11
```

Slika 2.36. Tablica ANOVA za razmatrani primjer

Konačno, vrijednost koeficijenta jednostavne linearne korelacije računa se kao drugi korijen iz koeficijenta determinacije, $\sqrt{0,0028}$ (može se izračunati naredbom sqrt(sazetak$r.squared), gdje sqrt označava drugi korijen) te iznosi –0,0529. Negativne je vrijednosti jer je predznak procijenjenog koeficijenta uz varijablu HICP negativan (vidjeti sliku 2.33). Dakle, između zavisne i nezavisne varijable postoji slaba negativna korelacija. S obzirom da se radi o slučaju jedne nezavisne varijable, pomoću naredbe cor(…) može se provjeriti procijenjeni koeficijent korelacije, što je prikazano na slici 2.37.

```
HICP<-podaci$HICP
cor(HICP, GDP)

## [1] -0.05293131
```

Slika 2.37. Koeficijent korelacije između varijabli BDP i HICP

Primjer 2.16.
Temeljem simuliranih podataka o varijabli *x* i *y*, procijenjeno je nekoliko jednostavnih linearnih regresijskih modela[26]:

$$M1: y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$M2: y_i = \alpha_0 + \alpha_1 \sqrt{x_i} + u_i \,,$$
$$M3: y_i = \lambda_0 + \lambda_1 \ln x_i + e_i$$

čiji je ispis dan u tablici 2.8. Temeljem odgovarajućih mjera reprezentativnosti modela, odaberimo najbolji model i objasnimo zašto je najbolji.

Može se uočiti kako je najreprezentativniji model M1, jer ima najveću vrijednost koeficijenta determinacije (ukupno je 85% varijacija zavisne varijable objašnjeno tim modelom), te ujedno ima i najveću vrijednost korigiranog koeficijenta determinacije (uz M3), najmanju procijenjenu varijancu regresije (sukladno tome i najmanju procijenjenu standardnu devijaciju regresije, koja iznosi 20, pa je prosječno odstupanje empirijskih od regresijskih vrijednosti zavisne varijable 20 jedinica), kao i što je najmanja vrijednost procijenjenog koeficijenta varijacije regresije, jer iznosi 22% (prosječno odstupanje empirijskih od regresijskih vrijednosti zavisne varijable iznosi 22%, što je malo do umjereno odstupanje).

Drugo mjesto zauzimaju M2 i M3 jer je M2 bolji u slučaju većih koeficijenta determinacije i korigiranog koeficijenta determinacije, dok je M3 bolji u slučaju manje procijenjene varijance regresije i koeficijenta varijacije regresije.

[26] Uobičajena oznaka za slučajnu varijablu jest ε, no kako se radi o tri različita modela, oznaka za slučajnu varijablu u drugome modelu je *u*, te u posljednjem *e*, kako bi se razlikovale sve tri slučajne varijable. Dodatno, u sva tri modela su oznake za parametre koje je potrebno procijeniti različite, s obzirom na promjenu funkcionalnog oblika modela, dolazi do promjene u procijenjenim parametrima.

Tablica 2.8. Ispis procijenjenih modela M1, M2 i M3

Parametar ili mjera / Model:	**M1**	**M2**	**M3**
R^2	85%	80%	82%
$\overline{R}^2$	79%	77%	79%
$\hat{\sigma}^2$	400	450	420
$\hat{V}$	22%	30%	24%

2.1.9. Testiranje hipoteza u modelu jednostavne linearne regresije

2.1.9.1. *t*-test

Ako su zadovoljene pretpostavke linearnog regresijskog modela, procijenjeni parametri dobiveni metodom najmanjih kvadrata su nepristrani, stoga se može provoditi testiranje hipoteza temeljem *t*-testa ili *F*-testa. Ako se žele testirati hipoteze o vrijednosti nepoznatih parametara β_0 i β_1, može se provesti **jednosmjerni** ili **dvosmjerni** *t*-test. Napomenimo da vrijednosti parametara β_0 i β_1 ne znamo, ali u hipotezama simbolički pišemo pretpostavke upravo o tim parametrima koji nisu poznati. Vrijednosti $\hat{\beta}_0$ i $\hat{\beta}_1$ su poznate, stoga se testiranje hipoteza ne odnosi na njih!

Ako se želi testirati **značajnost nezavisne varijable** u modelu $y_i = \beta_0 + \beta_0 x_i + \varepsilon_i$, , to znači da se testira je li odabrana nezavisna varijabla značajna za objašnjavanje varijacija zavisne varijable. Prilikom provođenja testa, najprije se formiraju nulta (H_0) i alternativna hipoteza (H_1). Ako se razmatra dvosmjerni test, postupak je sljedeći. Nulta hipoteza pretpostavljat će da varijabla *x* (nezavisna) nije značajna u modelu (ili da je suvišna), dok će alternativna hipoteza pretpostavljati suprotno, odnosno varijabla *x* nije značajna, što simbolički pišemo na način:

$$\begin{aligned} H_0 &: y_i = \beta_0 + \varepsilon_i \\ H_1 &: y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \end{aligned}, \tag{2.128}$$

odnosno

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}, \tag{2.129}$$

pri čemu je najčešći pristup pisanja hipoteza dan u (2.129). Dakle, ako varijabla *x* nije značajna u modelu, njezin učinak na varijablu *y* jednak je 0 (nema učinka), što piše u H_0 u (2.129). Test se naziva dvosmjeran jer u alternativnoj hipotezi pretpostavljamo da vrijednosti mogu biti pozitivne ili negativne.

U slučaju **jednosmjernog** testa, u alternativnoj hipotezi se koristi znak strogo veće ili strogo manje. Ako se u H_1 piše znak strogo veće („>“), radi se o testu na **gornju** granicu, dok za znak strogo manje („<“) govorimo o testu na **donju** granicu. Dva su načina određivanja hoće li se raditi o testu na gornju ili donju granicu. Prvi način je da temeljem ekonomske teorije koja će pretpostavljati pozitivnu ili negativnu vezu između zavisne i nezavisne varijable odredimo sukladno tome odgovarajući test. Drugi način je promotriti procijenjenu vrijednost parametra $\hat{\beta}_1$, te sukladno njegovom predznaku odrediti tip testa. Test na gornju granicu piše se ovako:

$$\begin{aligned} H_0 &: \beta_1 \leq 0 \\ H_1 &: \beta_1 > 0 \end{aligned}, \tag{2.130}$$

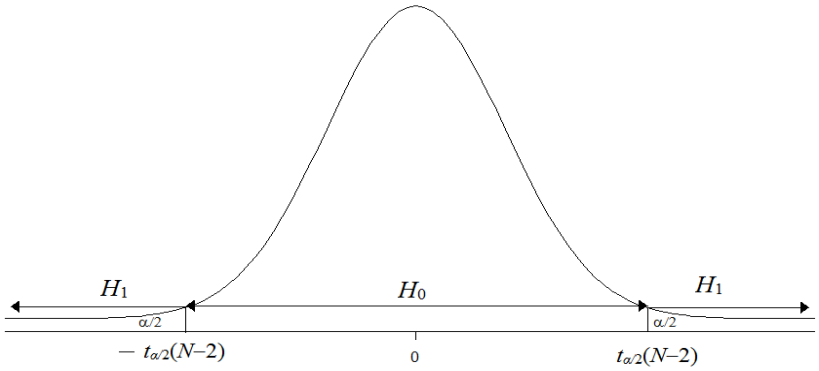
dok se test na donju granicu piše na način:

$$\begin{aligned} H_0 &: \beta_1 \geq 0 \\ H_1 &: \beta_1 < 0 \end{aligned}. \tag{2.131}$$

Dodatno, u nultim hipotezama u (2.130) i (2.131) može se pisati i znak jednakosti „=“. Nakon formiranja hipoteza, računa se **test veličina** (ili empirijska veličina, jer se procjenjuje temeljem empirijskih podataka). U slučaju *t*-testa, radi se o veličini:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)} \sim t(N-2), \tag{2.132}$$

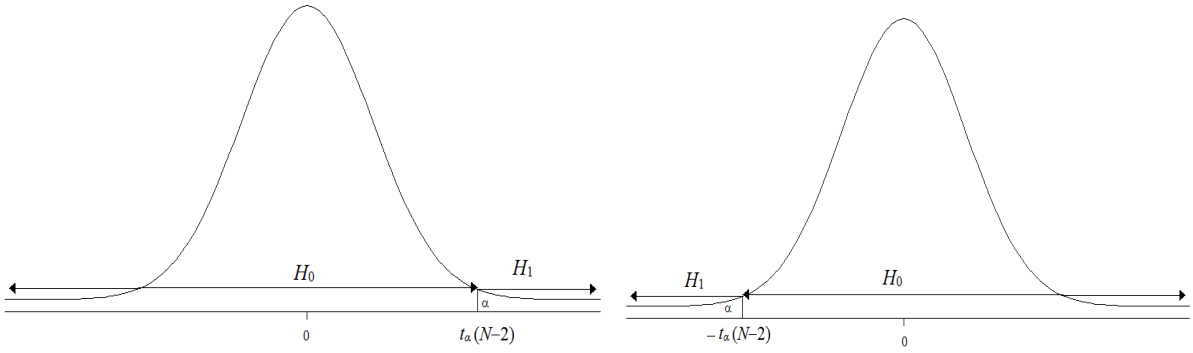
jer ako su, kako je već spomenuto, zadovoljene pretpostavke linearnog regresijskog modela, omjer $\frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)}$ slijedi *t*-distribuciju s N–2 stupnja slobode (vidjeti naslov 2.1.7). Test veličina (2.132) se potom uspoređuje s **teorijskom veličinom**, $t_\alpha(N-2)$, koja se određuje iz tablice kritičnih vrijednosti Studentove distribucije, temeljem određene **razine značajnosti** (signifikantnosti) α i broj stupnjeva slobode N–2. Razina značajnosti α je zadana vjerojatnost pogreške odbacivanja istinite nulte hipoteze. Drugim riječima, radi se o vjerojatnosti greške tipa I (vidjeti Dodatak 10.3), tj. $\alpha = P$ (odbaciti $H_0 \mid H_0$) (čitamo: vjerojatnost odbacivanja nulte hipoteze uz danu nultu hipotezu koja je istinita).



Slika 2.38. Dvosmjerni *t*-test o značajnosti nezavisne varijable *x*

U slučaju dvosmjernog testa se α dijeli s 2. Ako se razmotri slika 2.38., uočava se da se u slučaju $|t_1| > t_\alpha(N-2)$ odbacuje nulta hipoteza. Drugi naziv za teorijsku veličinu $t_\alpha(N-2)$ je i odgovarajući percentil *t*-distribucije s N–2 stupnja slobode. U slučaju jednosmjernog testa, za test na gornju granicu uspoređuju se vrijednosti t_1 i $t_\alpha(N-2)$, dok se za slučaj testa na donju granicu uspoređuju vrijednosti t_1 i $-t_\alpha(N-2)$, vidjeti sliku 2.39. (lijevi panel test na gornju granicu, desni panel test na donju granicu). U slučaju testa na gornju granicu, ako je $t_1 > t_\alpha(N-2)$, odbacuje se nulta hipoteza, dok u slučaju testa na donju granicu, ako vrijedi $t_1 < t_\alpha(N-2)$, tada se odbacuje nulta

hipoteza tog testa. Uočimo da se u slučaju jednosmjernih testova, α ne dijeli, već se razmatra u lijevome ili desnome repu distribucije.



Slika 2.39. Jednosmjerni *t*-testovi o značajnosti nezavisne varijable *x*

Ako se pak želi testirati značajnost konstante u modelu, tada se razmatraju sljedeće hipoteze za slučaj dvosmjernog testa:

$$\begin{aligned} H_0 &: \beta_0 = 0 \\ H_1 &: \beta_0 \neq 0 \end{aligned}, \tag{2.133}$$

sa sljedećim empirijskim *t*-omjerom:

$$t_0 = \frac{\hat{\beta}_0 - 0}{SE\left(\hat{\beta}_0\right)} = \frac{\hat{\beta}_0}{SE\left(\hat{\beta}_0\right)} \sim t(N-2). \tag{2.134}$$

Osim uspoređivanja empirijskog *t*-omjera s teorijskim, mogu se uspoređivati **empirijska razina značajnosti**, tj. *p*-vrijednost, i zadana teorijska razina značajnosti α. *p*-vrijednost je, dakle, vjerojatnost da slučajna varijabla poprimi vrijednost jednaku ili veću od empirijske veličine (za slučaj bilo koje varijable označit ćemo empirijski *t*-omjer kao t_{emp}):

$$p\text{-vrijednost} = P\left(t(N-2) \geq |\, t_{emp}\,|\right), \tag{2.135}$$

pri čemu se *p*-vrijednost u (2.135) uspoređuje s teorijskom razinom značajnosti α. Odluka se donosi na sljedeći način:

$$\begin{aligned} p - \text{vrijednost} > \alpha &\Rightarrow \text{ne odbacujem } H_0 \\ p - \text{vrijednost} < \alpha &\Rightarrow \text{odbacujem } H_0 \end{aligned}, \tag{2.136}$$

što znači da ako je nulta hipoteza istinita, empirijski *t*-omjer je slučajna varijabla koja slijedi *t*-distribuciju s *N*–2 stupnja slobode. Tada je ***p*-vrijednost vjerojatnost** da uz istinitu nultu hipotezu varijabla *t*(*N*–2) poprimi vrijednost jednaku ili veću od empirijske veličine $|t_{emp}|$. Ako je izračunata vjerojatnost u (2.135) manja od teorijske razine značajnosti, tada se nulta hipoteza odbacuje kao neistinita u (2.136). Obrnuto vrijedi ako je veća od α.

Za slučaj dvosmjernog testa, relacija (2.135) piše se na sljedeći način:

$$p\text{-vrijednost} = 2P\left(t(N-2) \geq |\, t_{emp}\,|\right), \tag{2.137}$$

i upravo se u većini programske podrške u ispisima prikazuje *p*-vrijednost izračunata formulom (2.137). Stoga se u provođenju dvosmjernog testa koristi izravno vrijednost dana u ispisu, dok se za slučaj jednosmjernih testova ta vrijednost podijeli s vrijednošću 2. Detalji o *p*-vrijednosti dani su u naslovu 2.1.9.7.

Želi li se testirati hipoteza da je vrijednost parametra β_1 jednaka nekoj **specifičnoj vrijednosti** $\tilde{\beta}_1$, tada je test veličina sljedeća:

$$t_1 = \frac{\hat{\beta}_1 - \tilde{\beta}_1}{SE\left(\hat{\beta}_1\right)} \sim t(N-2), \qquad (2.138)$$

i samo provođenje testa je identično kao u slučaju testiranja hipoteze o (ne)značajnosti nezavisne varijable u modelu.

Tablica 2.9. Sažet prikaz *t*-testova u modelu jednostavne linearne regresije

Dvosmjerni test	Jednosmjerni na gornju granicu	Dvosmjerni na gornju granicu
$H_0 : \beta_j = 0$ $H_1 : \beta_j \neq 0$	$H_0 : \beta_j = 0$ $H_1 : \beta_j > 0$	$H_0 : \beta_j = 0$ $H_1 : \beta_j < 0$
$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$	$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$	$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$
$\lvert t_j \rvert > t_\alpha(N-2) \rightarrow$ odbacujem H_0 $\lvert t_j \rvert < t_\alpha(N-2) \rightarrow$ ne odbacujem H_0	$t_j > t_\alpha(N-2) \rightarrow$ odbacujem H_0 $t_j < t_\alpha(N-2) \rightarrow$ ne odbacujem H_0	$t_j > -t_\alpha(N-2) \rightarrow$ odbacujem H_0 $t_j < -t_\alpha(N-2) \rightarrow$ ne odbacujem H_0

Primjer 2.17.
Procijenjen je jednostavni linearni regresijski model između zavisne varijable BDP i nezavisne varijable HICP (datoteka „**BDP_i_HICP.txt**“). Provedimo *t*-test o značajnosti varijable HICP: dvosmjerni i jednosmjerni test. Odabrana razina značajnosti je α=5%. Mijenja li se zaključak za α=1%, odnosno α=10%?

Ako se razmotri ispis na slici 2.40., središnji dio nazvan „Coefficients:“ sadrži u prvome stupcu (naziv „Estimate“) procijenjene vrijednosti parametara u modelu ($\hat{\beta}_0$ i $\hat{\beta}_1$), drugi stupac (naziv „Std. Error) sadrži vrijednosti procijenjenih standardnih pogrešaka ($SE(\hat{\beta}_0)$ i $SE(\hat{\beta}_1)$). U trećem stupcu (naziv „t value“) sadrži empirijske *t*-omjere, izračunate prema formuli (2.132) za nezavisnu varijablu HICP, odnosno (2.134) za konstantu. Posljednji stupac (naziv „Pr (> |t|))“ sadrži *p*-vrijednosti za oba empirijska *t*-omjera, ali za slučaj dvosmjernog testa. Za provedbu jednosmjernog testa se dane *p*-vrijednosti dijele s vrijednošću 2. S druge strane, na slici 2.41 prikazan je izračun kritičnih granica za slučaj dvosmjernog, kao i jednosmjernog testa.

```
## Call:
## lm(formula = GDP ~ HICP, data = podaci)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -603036 -457350 -315196  -39255 2898603
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2216386    5748839   0.386    0.703
## HICP          -15791      54391  -0.290    0.774
##
## Residual standard error: 852700 on 30 degrees of freedom
## Multiple R-squared:  0.002802,   Adjusted R-squared:  -0.03044
## F-statistic: 0.08429 on 1 and 30 DF,  p-value: 0.7736
```

Slika 2.40. Ispis modela jednostavne linearne regresije

Ako se provodi test značajnosti varijable HICP u modelu, temeljem slika 2.40. i 2.41., možemo zapisati sljedeće hipoteze, test veličine i zaključak.

Dvosmjerni test:

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$, $t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-15791}{54391} = -0,29$, p-vrijednost = 0,774. Kritična granica za α = 5% i 30 stupnjeva slobode (N–2 = 32–2) iznosi $t_{0,05/2}(30) = t_{0,025}(30) = 2,042$.

```
#kritična granica za dvosmjerni test
kriticna<-abs(qt(0.05/2,32-2))
kriticna

## [1] 2.042272

#za jednosmjerni (na gornju granicu):
kriticna2<-qt(1-0.05,32-2)
kriticna2

## [1] 1.697261

#za jednosmjerni (na donju granicu):
kriticna2<-qt(0.05,32-2)
kriticna2

## [1] -1.697261
```

Slika 2.41. Naredbe potrebne za izračun teorijskih veličina (teorijskih t-omjera), α = 5%

Kako je $|t_1| < t_{0,05}(30)$, zaključujemo da ne odbacujemo nultu hipotezu. Riječima, uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla HICP nije značajna u modelu. Mogli smo usporediti i p-vrijednost s razinom α: kako vrijedi da je p-vrijednost > α, ponovno zaključujemo kako se odbacuje nulta hipoteza.

U slučaju jednosmjernog testa, uočavamo da je vrijednost procijenjenog parametra uz varijablu HICP negativna. Stoga se provodi test na donju granicu:

$H_0 : \beta_1 \geq 0$
$H_1 : \beta_1 < 0$, $t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-15791}{54391} = -0,29$, p-vrijednost = 0,774/2 =0,387. Kritična granica za α=5% i 30 stupnjeva slobode (N–2 = 32–2) iznosi $t_{0,05}(30) = -1,697$.

Kako je $t_1 > t_{0,05}(30)$, zaključujemo da ne odbacujemo nultu hipotezu. Riječima, uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla HICP nije značajna u modelu. Ako

uspoređujemo *p*-vrijednost s razinom α: kako vrijedi da je *p*-vrijednost > α, ponovno zaključujemo kako se odbacuje H_0. Uočimo kako smo sada *p*-vrijednost iz ispisa podijelili s vrijednošću 2, s obzirom da RStudio računa tu vjerojatnost formulom (2.137)!

Tablica 2.10. Rezultati testiranja značajnosti varijable HICP, različiti testovi

Veličina / Test	Dvosmjerni	Jednosmjerni	Ishod
Empirijski *t*-omjer	–0,290		–
Teorijski *t*-omjer, α=5%	2,042	–1,697	H_0
Teorijski *t*-omjer, α=1%	2,750	–2,457	H_0
Teorijski *t*-omjer, α=10%	1,697	–1,310	H_0

Ako se pak provode testovi za slučaj α = 1% i 10%, rezultati su sažeti u tablici 2.10., temeljem slika 2.40., 2.41. i 2.42. Uočava se da vrijednost empirijskog *t*-omjera (–0,290) pripada intervalu ne odbacivanja nulte hipoteze u svim slučajevima. Dakle, pri uobičajenim razinama značajnosti, ne odbacujemo hipotezu da varijabla HICP nije značajna u modelu.

```
#kritična granica za dvosmjerni test
kriticna<-abs(qt(0.01/2,32-2))
kriticna

## [1] 2.749996

#za jednosmjerni (na gornju granicu):
kriticna2<-qt(0.01,32-2)
kriticna2

## [1] -2.457262

#kritična granica za dvosmjerni test
kriticna<-abs(qt(0.1/2,32-2))
kriticna

## [1] 1.697261

#za jednosmjerni (na gornju granicu):
kriticna2<-qt(0.1,32-2)
kriticna2

## [1] -1.310415
```

Slika 2.42. Naredbe potrebne za izračun teorijskih *t*-omjera, α =1% i 10%

2.1.9.2. Napomena o terminologiji kod testiranja hipotezi

Prema Wooldridge (2016), valja napomenuti da kada ne odbacujemo nultu hipotezu, kažemo u interpretaciji da „ne odbacujemo nultu hipotezu pri određenoj razini značajnosti“. **Ne govorimo da prihvaćamo nultu hipotezu**. Zašto je to tako? Uzmimo primjer u kojemu se testira nulta hipoteza: β_1 = 2, čiji je empirijski *t*-omjer jednak 0,2, za zadani teorijski *t*-omjer 1,697 i zaključak da ne odbacujemo nultu hipotezu jer je empirijski *t*-omjer manji od teorijskog. Potom testiramo nultu hipotezu β_1 = 2,2, čiji je empirijski *t*-omjer jednak 0,3, za zadani teorijski *t*-omjer 1,697 i ponovni je zaključak da ne odbacujemo nultu hipotezu. Očito je da u obje nulte hipoteze pretpostavljamo različite vrijednosti (2 i 2,2), stoga se ne može reći da se prihvaćaju te hipoteze, odnosno da se prihvaća nulta hipoteza, jer istovremeno ne može vrijediti β_1 = 2 i β_1 = 2,2. Zato se samo interpretacija vrši na način da kažemo da temeljem empirijskih podataka

nemamo dovoljno dokaza za odbacivanje nulte hipoteze pri nekoj razini značajnosti. **Ova napomena vrijedi za bilo koji test** (ne samo *t*-test).

2.1.9.3. *F*-test

U slučaju modela jednostavne linearne regresije, razmotrila se tablica ANOVA u 2.1.8., vidjeti tablicu 2.11. Spomenuto je kako je ideja da *F*-omjer bude čim veći, s obzirom da sredina kvadrata odstupanja protumačena modelom (*SSE*/1) treba biti čim veća, a sredina kvadrata odstupanja neprotumačena modelom (*SSR*/(*N*–2)) čim manja.

Tablica 2.11. Tablica ANOVA, slučaj jednostavne linearne regresije

Izvor varijacije	Sume kvadrata	Stupnjevi slobode (*ss*)	Sredina kvadrata odstupanja	*F*-omjer	*p*-*v*
Regresija (protumačeno modelom)	*SSE*	1	*SSE*/1	$\frac{SSE/1}{SSR/(N-2)}$	…
Rezidualna odstupanja (neprotumačeno modelom)	*SSR*	*N*–2	*SSR*/(*N*–2)		
Ukupno	*SST*	*N*–1			

Kako se empirijski *F*-omjer računa formulom:

$$F = \frac{\frac{SSE}{1}}{\frac{SSR}{N-2}} \sim F(1; N-2), \tag{2.139}$$

ako vrijede pretpostavke regresijskog modela (naslov 2.1.2), tada je *SSE*/1 slučajna varijabla s jednim stupnjem slobode, što pišemo: $SSE/1 \sim \chi^2(1)$, dok je SSR/σ^2 slučajna varijabla koja slijedi hi-kvadrat distribuciju s *N*–2 stupnja slobode, što pišemo: $SSR/\sigma^2 \sim \chi^2 (N-2)$. Tada je omjer u (2.139) slučajna varijabla koja slijedi *F*-distribuciju s 1 stupnjem slobode u brojniku i *N*–2 stupnja slobode u nazivniku[27]. Ako se pretpostavi da je nezavisna varijabla suvišna u modelu jednostavne linearne regresije, tada je omjer (2.139) jednak 1, jer vrijedi:

$$E(SSE) = E\left(\sum_{i=1}^{N}\left(\hat{y}_i - \overline{y}\right)^2\right) = E\left(\sum_{i=1}^{N}\left(\underbrace{\hat{\beta}_0}_{\overline{y}-\hat{\beta}_1\overline{x}\ \text{(vidjeti formulu 1.36)}} + \hat{\beta}_1 x_i - \overline{y}\right)^2\right)$$

$$= E\left(\sum_{i=1}^{N}\left(\overline{y} - \hat{\beta}_1\overline{x} + \hat{\beta}_1 x_i - \overline{y}\right)^2\right) = E\left(\hat{\beta}_1^2\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2\right) = E\left(\hat{\beta}_1^2\hat{\sigma}_x^2\right), \tag{2.140}$$

tj.

[27] Vidjeti Dodatak 5.3.

$$E\left(\hat{\beta}_1^2\hat{\sigma}_x^2\right)=E\left(\hat{\beta}_1-\beta_1+\beta_1\right)^2\hat{\sigma}_x^2=\left(\underbrace{E\left(\hat{\beta}_1-\beta_1\right)^2}_{Var\left(\hat{\beta}_1\right)}+\underbrace{2E\left(\left(\hat{\beta}_1-\beta_1\right)\beta_1\right)}_{E\left(\hat{\beta}_1-\beta_1\right)=0}+E\left(\beta_1^2\right)\right)\hat{\sigma}_x^2$$

$$=\left(\underbrace{Var\left(\hat{\beta}_1\right)}_{=\text{relacija (1.73)}}+\underbrace{\beta_1^2}_{=0}\right)\hat{\sigma}_x^2=\frac{\sigma^2}{\hat{\sigma}_x^2}\hat{\sigma}_x^2=\sigma^2 \tag{2.141}$$

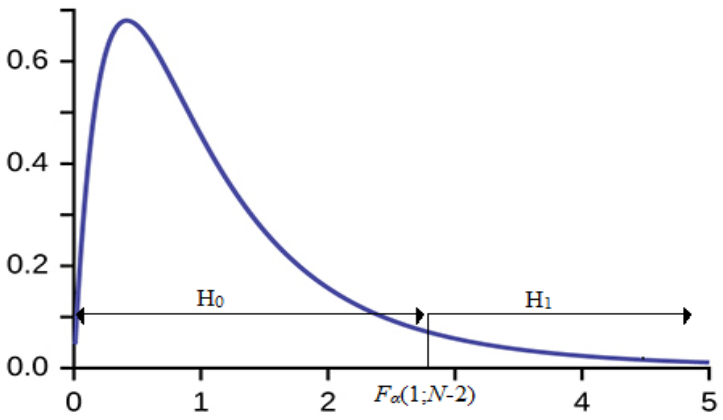
dok je

$$E(SSR)=E\left(\hat{\sigma}^2(N-2)\right)=\sigma^2(N-2). \tag{2.142}$$

Uvrštavanjem σ^2 iz (2.141) u brojnik od (2.139), te $\sigma^2(N-2)$ iz (2.142) u nazivnik od (2.139), dobiva se očekivana vrijednost od (2.139) jednaka 1 (uz ponavljamo, pretpostavku da je $\beta_1 = 0$). U suprotnome je očekivana vrijednost tog omjera veća od 1. ***F*-test se provodi na sljedeći način**. Formiraju se nulta i alternativna hipoteza:

$$\begin{aligned} H_0&:\beta_1=0\\ H_1&:\beta_1\neq 0\end{aligned}\ , \tag{2.143}$$

te se potom uz zadanu razinu značajnosti α izračuna teorijski *F*-omjer (F_{emp}), $F_\alpha(1;N-2)$, koji se uspoređuje s empirijskim *F*-omjerom iz tablice ANOVA, tj. iz (2.139), vidjeti sliku 2.43. Ako vrijedi $F_{emp} > F_\alpha(1;N-2)$, odbacuje se nulta hipoteza, dok se u slučaju $F_{emp} < F_\alpha(1;N-2)$ onda ne odbacuje.



Slika 2.43. *F*-test o značajnosti regresijskog parametra

Osim usporedbe *F* veličina, test se može provesti **usporedbom *p*-vrijednosti i teorijske razine značajnosti α**. Kako se *p*-vrijednost računa formulom:

$$p\text{-vrijednost}=P\left(F\left(1;N-2\right)\geq F_{emp}\right), \tag{2.144}$$

ako je *p*-vrijednost veća od α, nulta hipoteza se ne odbacuje, dok za slučaj *p*-vrijednosti < α se nulta hipoteza odbacuje.

Primjer 2.18.
Procijenjen je jednostavni linearni regresijski model između zavisne varijable BDP i nezavisne varijable HICP (datoteka „**BDP_i_HICP.txt**“). Provedimo *F*-test o značajnosti varijable HICP. Odabrana razina značajnosti je α=5%. Mijenja li se zaključak za α=1%, odnosno α=10%?

U okviru ispisa na slici 2.44., uočava se da je vrijednost empirijskog *F*-omjera jednaka 0,08429, s 1 stupnjem slobode u brojniku i 30 u nazivniku. Vrijednost je, dakle, izračunata kao (pratiti sliku 2.44.):

$$F = \frac{\frac{SSE}{1}}{\frac{SSR}{N-2}} = \frac{\frac{6{,}12\cdot 10^{10}}{1}}{\frac{2{,}18\cdot 10^{13}}{32-2}} = 0{,}0843$$, p-v = 0,7736. Hipoteze testa su: $\begin{matrix} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{matrix}$.

Teorijski *F*-omjer dan je temeljem tablice kritičnih vrijednosti *F*-distribucije uz α = 5%, te 1 stupanj slobode u brojniku i 30 u nazivniku (slika 2.46.) iznosi $F_{0,05}(1;30)$ = 4,171.

Stoga vrijedi: F_{emp} = 0,0843 < 4,171 = $F_{0,05}(1;30)$, pa se nulta hipoteza ne može odbaciti. Kako vrijedi i *p*-vrijednost = 0,7736 > 0,05 = α, dolazi se do istog zaključka. Riječima: uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla HICP nije značajna u modelu.

```
##
## Call:
## lm(formula = GDP ~ HICP, data = podaci)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -603036 -457350 -315196  -39255 2898603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2216386    5748839   0.386    0.703
## HICP          -15791      54391  -0.290    0.774
##
## Residual standard error: 852700 on 30 degrees of freedom
## Multiple R-squared:  0.002802,   Adjusted R-squared:  -0.03044
## F-statistic: 0.08429 on 1 and 30 DF,  p-value: 0.7736
```

Slika 2.44. Ispis modela jednostavne linearne regresije

```
anova(m1)

## Analysis of Variance Table
##
## Response: GDP
##           Df     Sum Sq    Mean Sq F value Pr(>F)
## HICP       1 6.1286e+10 6.1286e+10  0.0843 0.7736
## Residuals 30 2.1813e+13 7.2710e+11
```

Slika 2.45. Tablica ANOVA

```
#za F test:
qf(1-0.05,1,32-2)

## [1] 4.170877
```

Slika 2.46. Naredbe potrebne za izračun teorijskog *F*-omjera, α =5%

Ako se razina značajnosti promijeni u 1% ili 10%, *p*-vrijednost koja iznosi 0,7736 je veća i od 0,01 i od 0,1, stoga ne dolazi do promjene u zaključku testa. Usporedbu smo mogli izvršiti i uspoređujući empirijski *F*-omjer s teorijskima za 1% i 10%, koje bi u RStudiju izračunali naredbama qf(0.99,1,30) i qf(0.9,1,30), pri čemu bismo izračunali vrijednosti 7,562 i 2,881.

2.1.9.4. Waldov test

Waldov test, LR i LM test su testovi kojima se razmatraju ograničenja na parametre u regresijskom modelu. Kako Greene (2018) navodi, razmatrajući asimptotske distribucije te primjenom rezultata velikih uzoraka (engl. *large sample theory*) nad procjeniteljem iz metode najmanjih kvadrata, moguće je provesti inferencijalnu analizu regresijskog modela za provođenje određenih testova. U nastavku se pretpostavlja da se sva tri testa provode temeljem modela koji je procijenjen pomoću metode najveće vjerodostojnosti.

Općenito se u Waldovom testu pretpostavlja da jedan ili više parametara u regresijskom modelu zadovoljavaju određena linearna ograničenja. Ako se razmatra model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ u matričnom zapisu, J ograničenja na parametre $\boldsymbol{\beta}$ se u slučaju modela jednostavne linearne regresije mogu pisati kao:

$$\begin{aligned} r_{11}\beta_0 + r_{12}\beta_1 &= q_1 \\ r_{21}\beta_0 + r_{22}\beta_1 &= q_2 \\ &\vdots \\ r_{J1}\beta_0 + r_{J2}\beta_1 &= q_J \end{aligned}, \tag{2.145}$$

koja se općenito pišu u matričnoj formi:

$$\boldsymbol{R\beta} = \boldsymbol{q}, \tag{2.146}$$

gdje je $\boldsymbol{R} \in \mathcal{M}_{J,2}$, $\boldsymbol{\beta} \in \mathbb{R}^2$ i $\boldsymbol{q} \in \mathbb{R}^J$, te se formiraju nulta i alternativna hipoteza kao:

$$\begin{aligned} H_0&: \boldsymbol{R\beta} = \boldsymbol{q} \\ H_1&: \boldsymbol{R\beta} \neq \boldsymbol{q} \end{aligned}. \tag{2.147}$$

U slučaju modela jednostavne linearne regresije u kojem se **ispituje značajnost nezavisne varijable**, hipoteza se piše kao do sada H_0: $\beta_1 = 0$, odnosno matrično:

$$H_0: \boldsymbol{R\beta} = \boldsymbol{q}, \quad \boldsymbol{R} = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}, \tag{2.148}$$

što se lako provjeri: $\boldsymbol{R\beta} = [0\ 1] \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = [\ 0 \cdot \beta_0 + 1 \cdot \beta_1\ ] = [\ 0\ ] = \boldsymbol{q}$.

S obzirom da se test vrši temeljem procijenjenih vrijednosti u $\hat{\boldsymbol{\beta}}$, razmatra se $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{m}$, te se promatra koliko je $\boldsymbol{m}$ udaljen od $\boldsymbol{q}$, je li $\boldsymbol{m}$ različit od $\boldsymbol{q}$ temeljem varijabilnosti uzorkovanja ili se radi o statistički značajnoj razlici. Kako je $\hat{\boldsymbol{\beta}}$ normalno distribuiran, a $\boldsymbol{m}$ je linearna funkcija od $\hat{\boldsymbol{\beta}}$, vektor $\boldsymbol{m}$ je također normalno distribuiran. Ako je nulta hipoteza istina, tada vrijedi $\boldsymbol{R\beta} = \boldsymbol{q}$, te je očekivanje od $\boldsymbol{m}$ jednako:

$$E\left[\boldsymbol{m} \mid \boldsymbol{X}\right] = \boldsymbol{R}E\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] - \boldsymbol{q} = \boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{q} = 0\,, \tag{2.149}$$

dok je matrica varijanci-kovarijanci jednaka:

$$Var\left[\boldsymbol{m} \mid \boldsymbol{X}\right] = Var\left[\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q} \mid \boldsymbol{X}\right] = \boldsymbol{R}Var\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right]\boldsymbol{R}' = \boldsymbol{R}\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\,. \tag{2.150}$$

Test veličina temeljem empirijskih podataka (koristi se procjena varijance) dana je kao:

$$W = \boldsymbol{m}Var\left[\boldsymbol{m} \mid \boldsymbol{X}\right]^{-1}\boldsymbol{m} = \left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}\right)'\left(\boldsymbol{R}\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}\right) \sim \chi^2(J)\,, \tag{2.151}$$

stoga asimptotski slijedi hi-kvadrat distribuciju[28] s J stupnjeva slobode. U slučaju testa u (2.148) vrijedi:

$$W = \left(\hat{\beta}_1 - 0\right)'Var\left(\hat{\beta}_1 - 0\right)^{-1}\left(\hat{\beta}_1 - 0\right), \tag{2.152}$$

to je

$$W = \frac{\hat{\beta}_1^2}{Var\left(\hat{\beta}_1\right)} \sim \chi^2(1)\,. \tag{2.153}$$

Ako se za procjenu (2.153) koristi izraz (2.31) za $\hat{\beta}_1$ te procjena varijance u (2.88), tada je Wald test veličina dana kao[29]:

$$\begin{aligned} W = \frac{\hat{\beta}_1^2}{Var(\hat{\beta}_1)} &= \frac{\left(\frac{\widehat{Cov(x,y)}}{\widehat{Var(x)}}\right)^2}{\frac{\sum_{i=1}^N \hat{\varepsilon}_i^2}{N}} = \frac{\left(\frac{\widehat{Cov(x,y)}}{\widehat{Var(x)}}\right)^2}{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N\widehat{Var(x)}}} = \frac{\widehat{Cov(x,y)}^2 N}{\widehat{Var(x)}\sum_{i=1}^N (y_i - \hat{y}_i)^2} \\ &= \frac{\widehat{Cov(x,y)}^2}{\widehat{Var(x)}} \frac{N}{\sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N (\hat{y}_i - \bar{y})^2} \\ &= \frac{\widehat{Cov(x,y)}^2}{\widehat{Var(x)}} \frac{N}{\widehat{Var(y)}(1 - R^2)} = R^2 \frac{N}{1 - R^2} \end{aligned} \tag{2.154}$$

Dakle, Wald test veličina se može izračunati temeljem broja opažanja i koeficijenta determinacije originalnog modela. Originalni model je onaj koji procijenimo prije nametanja ograničenja koja se razmatraju u (2.147). Dodatno, test veličina u (2.154) asimptotski slijedi hi-kvadrat distribuciju (vidjeti detalje u Greene, 2018).

Nakon formiranja hipotezi i izračuna empirijske Wald test veličine u (2.153), tj. (2.154), uspoređuje se s teorijskom veličinom koja također slijedi hi-kvadrat distribuciju, s jednim

[28] Za izvod vidjeti Greene (2018: 135 i 1135).

[29] U posljednjem retku vrijedi $\frac{\widehat{Cov(x,y)}^2}{\widehat{Var(x)}\widehat{Var(y)}} = R^2$ jer se za slučaj jednostavne linearne regresije koeficijent korelacije između zavisne i nezavisne varijable može računati kao $R = \frac{\widehat{Cov(x,y)}^2}{\sqrt{\widehat{Var(x)}\widehat{Var(y)}}}$, a taj koeficijent drugi korijen od koeficijenta determinacije, vidjeti detalje u Gujarati i Porter (2010).

stupnjem slobode u slučaju testiranja (2.148), za zadanu razinu značajnosti α, χ^2_α (1), koja se iščitava iz tablice kritičnih granica hi-kvadrat distribucije. Ako je $W > \chi^2_\alpha$ (1), odbacuje se nulta hipoteza, dok se za slučaj $W < \chi^2_\alpha$ (1) ne odbacuje. Slično tome, ako je *p*-vrijednost < α, odbacuje se nulta hipoteza, dok se u slučaju *p*-vrijednost > α nulta hipoteza ne odbacuje.

Primjer 2.19.
Procijenjen je model jednostavne linearne regresije u kojemu BDP odabranih Europskih zemalja ovisi o HICP indeksu. Provedimo Waldov test o značajnosti varijable HICP u tom modelu, uz razinu značajnosti od 5%.

Test se mogao provesti temeljem rezultata na slici 2.44., gdje se nalazi ispis procijenjenog regresijskog modela (N = 32 i R^2 = 0,002802), ili pak korištenjem paketa „car“ u okviru RStudia, uz naredbu linearHypothesis(…) koja se koristi za Waldov test, parcijalni *F*-test i sl. Tako slika 2.47. prikazuje usporedbu dva modela: Model 1 je model s ograničenjem (engl. *restricted model*), u kojemu je nametnuto ograničenje $\beta_1 = 0$, stoga se radi o modelu u kojemu se nalazi samo konstanta, dok je Model 2 originalni model, u kojemu BDP ovisi o HICP varijabli. Ovdje je izračun test veličine („Chisq“) nešto drugačije predočen u odnosu na formulu (2.154), o čemu će više riječi biti u 2.2.7.3.

```
library(car)

ogranicenje<-"HICP=0"
linearHypothesis(m1,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## HICP = 0
##
## Model 1: restricted model
## Model 2: GDP ~ HICP
##
##   Res.Df        RSS Df  Sum of Sq  Chisq Pr(>Chisq)
## 1     31 2.1874e+13
## 2     30 2.1813e+13  1 6.1286e+10 0.0843     0.7716

qchisq(0.95,1)

## [1] 3.841459
```

Slika 2.47. Naredbe potrebne za provedbu Waldova testa, zajedno s rezultatom

No, dobivena empirijska veličina iznosi W = 0,0843, s pripadajućom *p*-vrijednosti („Pr(>Chisq)“) koja iznosi 0,7716, dok teorijska veličina dobivena naredbom qchisq(0.95,1) iznosi $\chi^2_{0,05}$ (1) = 3,841. Kako je $W < \chi^2_{0,05}$ (1), odnosno *p*-vrijednost > α, ne možemo odbaciti nultu hipotezu. Dakle, uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla HICP nije značajna u modelu.

2.1.9.5. LR test

LR test (engl. *likelihood ratio*) temelji se na procjeni parametara regresijskog modela metodom najveće vjerodostojnosti (vidjeti 2.1.3.5). Ideja je procijeniti originalni **model bez nametnutih ograničenja** na parametre, te potom **model s ograničenjima** te usporediti vrijednosti funkcija (2.82) – optimalne vrijednosti funkcije vjerodostojnosti modela bez ograničenja (oznaka $L(\boldsymbol{\theta})$)

i s ograničenjem (oznaka[30]: $L_R(\boldsymbol{\theta})$). Općenito se u nultoj hipotezi pretpostavljaju određena ograničenja:

$$H_0 : c(\boldsymbol{\theta}) = \boldsymbol{q}, \tag{2.155}$$

koja se mogu odnositi na parametre modela i varijancu. Primjerice, u nultoj hipotezi se može testirati $\beta_1 = 0$ i $\sigma^2 = 1$. Ako je nulta hipoteza istinita, tada se optimalne vrijednosti $L(\boldsymbol{\theta})$ i $L_R(\boldsymbol{\theta})$ neće puno razlikovati. Stoga se može formirati omjer:

$$\lambda = \frac{L_R(\boldsymbol{\theta})}{L(\boldsymbol{\theta})}, \tag{2.156}$$

čija je vrijednost između 0 i 1, s obzirom da su vrijednosti $L(\boldsymbol{\theta})$ i $L_R(\boldsymbol{\theta})$ pozitivne, a vrijednost optimuma s ograničenjem ne može biti veća od vrijednosti optimuma bez ograničenja.

Za provedbu samog testiranja, veličinu (2.156) je potrebno transformirati na način da dobivena *LR* veličina (test veličina) slijedi hi-kvadrat distribuciju s *J* stupnjeva slobode, gdje je *J* broj ograničenja u (2.155), (vidjeti Greene 2018: 554):

$$LR = -2\ln\lambda = -2\ln\frac{L_R(\boldsymbol{\theta})}{L(\boldsymbol{\theta})} = -2\left(\ln L_R(\boldsymbol{\theta}) - \ln L(\boldsymbol{\theta})\right) \sim \chi^2(J). \tag{2.157}$$

Ako se testira značajnost nezavisne varijable u jednostavnom linearnom regresijskom modelu, može se pokazati (vidjeti Maddala i Lahiri, 2009: 118-119) da je test veličinu (2.157) moguće računati kao:

$$LR = N\ln\frac{1}{1-R^2}. \tag{2.158}$$

Nakon formiranja nulte i alternativne hipoteze, te izračuna vrijednosti *LR*, ona se uspoređuje s teorijskom veličinom koja također slijedi hi-kvadrat distribuciju, s jednim stupnjem slobode u slučaju testiranja značajnosti jedne nezavisne varijable u modelu jednostavne linearne regresije, za zadanu razinu značajnosti α, $\chi^2_\alpha(1)$, koja se iščitava iz tablice kritičnih granica hi-kvadrat distribucije. Ako je $LR > \chi^2_\alpha(1)$, odbacuje se nulta hipoteza, dok se za slučaj $LR < \chi^2_\alpha(1)$ ne odbacuje. Slično tome, ako je *p*-vrijednost < α, odbacuje se nulta hipoteza, dok se u slučaju *p*-vrijednost > α nulta hipoteza ne odbacuje.

Primjer 2.20.
Procijenjen je model jednostavne linearne regresije u kojemu BDP odabranih Europskih zemalja ovisi o HICP indeksu. Provedimo LR test o značajnosti varijable HICP u tom modelu, uz razinu značajnosti od 5%.

I ovaj test se mogao provesti temeljem ispisa na slici 2.44. za slučaj testiranja $\beta_1 = 0$, s obzirom da je i za ovaj test mogu koristiti vrijednosti *N* i R^2. No, općenito se test provodi temeljem naredbe lrtest() u okviru RStudija, s obzirom na različita ograničenja koja je moguće testirati.

Najprije je potrebno procijeniti originalni model, gdje BDP ovisi o HICP (slika 2.48., model m1), te model s ograničenjem (slika 2.48., model m2, u kojem se BDP regresira samo na

[30] L_R je s engleskog, *R – restricted*.

konstantu). Koristeći se naredbom lrtest() u okviru paketa lmtest, dobiva se ispis na slici 51. Model 1 je originalni model, čija je vrijednost $\ln L(\hat{\boldsymbol{\theta}}) = -481{,}37$, dok je Model 2 onaj u kojemu je nametnuto ograničenje (HICP je suvišna u modelu), čija je vrijednost $\ln L_R(\hat{\boldsymbol{\theta}}) = -481{,}42$.

```
m1<-lm(GDP~HICP,data=podaci)
library(lmtest)

m2<-lm(GDP~1,data=podaci)
lrtest(m1,m2)

## Likelihood ratio test
##
## Model 1: GDP ~ HICP
## Model 2: GDP ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -481.37
## 2   2 -481.42 -1 0.0898     0.7645
```

Slika 2.48. Naredbe potrebne za provedbu LR testa, zajedno s rezultatom

Sada je $LR = -2(\ln L_R(\widehat{\boldsymbol{\theta}}) - \ln L(\widehat{\boldsymbol{\theta}})) = -2(-481{,}42 - (-481{,}37)) = 0{,}0898$, s pripadajućom *p*-vrijednosti („Pr>Chisq“) 0,7645. S druge strane, teorijska vrijednost $\chi^2_{0,05}$ (1) iznosi kao u prethodnom primjeru 3,841.

Kako vrijedi $LR = 0{,}0898 < 3{,}841 = \chi^2_{0,05}$ (1), odnosno *p*-vrijednost > α, ne možemo odbaciti nultu hipotezu. Dakle, uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla HICP nije značajna u modelu.

2.1.9.6. LM test

LM test (engl. *Lagrange multiplier*) **temelji se na procjeni modela s ograničenjem** (detaljan zapis i izvod procijenjenih parametara modela vidjeti u Greene, 2018: 126-127). S obzirom da se problem optimizacije vrši metodom Lagrangeovog množitelja, test se naziva LM test. Izvod test veličine u općenitom slučaju regresijskog modela, u slučaju takve optimizacije moguće je vidjeti u Greene (2018: 130).

Za slučaj testiranja značajnosti jedne nezavisne varijable u modelu jednostavne linearne regresije, može se pokazati[31] da je test veličinu moguće izračunati kao:

$$LM = NR^2 \sim \chi^2(1). \qquad (2.159)$$

Nakon formiranja nulte i alternativne hipoteze, te izračuna vrijednosti *LM* ona se uspoređuje s teorijskom veličinom koja također slijedi hi-kvadrat distribuciju, s jednim stupnjem slobode u slučaju testiranja značajnosti jedne nezavisne varijable u modelu jednostavne linearne regresije, za zadanu razinu značajnosti α, χ^2_α (1), koja se iščitava iz tablice kritičnih granica hi-kvadrat distribucije. Ako je $LM > \chi^2_\alpha$ (1), odbacuje se nulta hipoteza, dok se za slučaj $LM < \chi^2_\alpha$ (1) ne odbacuje. Slično tome, ako je *p*-vrijednost < α, odbacuje se nulta hipoteza, dok se u slučaju *p*-vrijednost > α nulta hipoteza ne odbacuje.

[31] Vidjeti Greene (2018: 130).

Primjer 2.21.
Procijenjen je model jednostavne linearne regresije u kojemu BDP odabranih Europskih zemalja ovisi o HICP indeksu. Provedimo LM test o značajnosti varijable HICP u tom modelu, uz razinu značajnosti od 5% temeljem ispisa na slici 2.43.

Kako je $N = 32$ i $R^2 = 0{,}002802$, to je $LM = NR^2 = 32 \cdot 0{,}002802 = 0{,}089664$. Pripadajuću p-vrijednost moguće je izračunati u RStudiju temeljem naredbe 1–pchisq(0.089664,1) koja iznosi 0,7646. S druge strane, teorijska vrijednost $\chi^2_{0,05}(1)$ iznosi kao u prethodnom primjeru 3,841.

Kako vrijedi $LM = 0{,}0897 < 3{,}841 = \chi^2_{0,05}(1)$, odnosno p-vrijednost > α, ne možemo odbaciti nultu hipotezu. Dakle, uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla HICP nije značajna u modelu.

Napomena. Za slučaj modela jednostavne linearne regresije, sljedeći je međuodnos između tri empirijske veličine: *W*, *LR* i *LM* (izvod vidjeti u Maddala i Lahiri, 2009: 121):

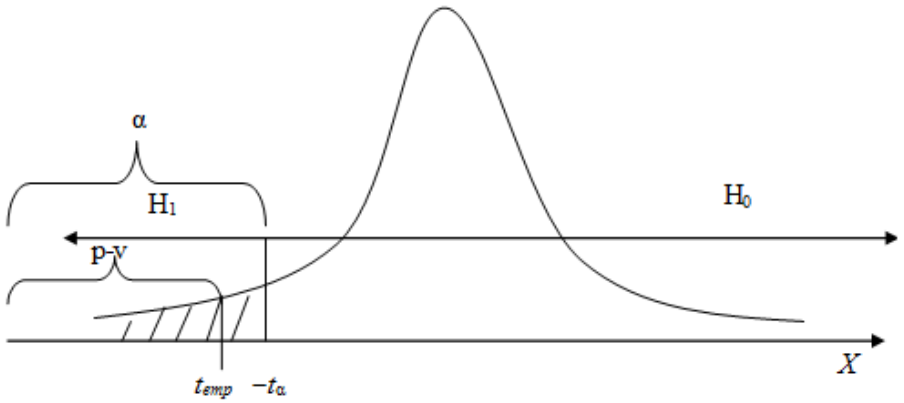
$$LM \leq LR \leq W. \tag{2.160}$$

2.1.9.7. Napomena o *p*-vrijednosti

Kod *t*-testa spomenuto je kako je *p*-vrijednost vjerojatnost da slučajna varijabla poprimi vrijednost jednaku ili veću od empirijske veličine (za slučaj *t*-testa empirijski *t*-omjer kao t_{emp} uspoređivao se s teorijskim $t(N–2)$):

$$p\text{-vrijednost} = P\left(t(N-2) \geq |t_{emp}|\right). \tag{2.161}$$

Ideja *p*-vrijednosti je da umjesto uspoređivanja empirijske test veličine s teorijskom pri različitim razinama značajnosti α, da se odgovori na pitanje: „uz danu vrijednost empirijske test veličine, koja je najmanja razina značajnosti pri kojoj će se nulta hipoteza odbaciti?“. Upravo je *p*-vrijednost ta razina značajnosti. Pitanje se može i ovako postaviti: „koja je najveća razina značajnosti uz koju možemo provesti test i odbaciti nultu hipotezu?“. Stoga male *p*-vrijednosti upućuju na odbacivanje nulte hipoteze jer upućuju da u slučaju istinite nulte hipoteze je mala vjerojatnost nastupa takvog ishoda temeljem danih podataka. Kako se radi o vjerojatnosti, *p*-vrijednost će uvijek biti u intervalu između 0 i 1. Pritom, geometrijski se radi o površini ispod grafa distribucije za koju se procjenjuje.



Slika 2.49. Usporedba *p*-vrijednosti i teorijske razine značajnosti α u slučaju *t*-testa na donju granicu

Za bolje razumijevanje tumačenja *p*-vrijednosti, razmotrimo slučaj *t*-testa na donju granicu (slika 2.49.). Teorijska razina značajnosti koja se određuje prilikom testa iznosi α. To je površina ispod grafa Studentove distribucije, koja se računa kao nepravi integral (donja granica teži prema $-\infty$, dok je gornja granica vrijednost $-t_\alpha$). S druge strane, temeljem empirijskog *t*-omjera t_{emp} računa se *p*-vrijednost kao površina ispod grafa Studentove distribucije, koja se također računa kao nepravi integral (donja granica teži prema $-\infty$, dok je gornja granica vrijednost t_{emp}). Ako je *p*-vrijednost < α, uočava se kako je površina predočena *p*-vrijednošću manja od površine predočenom vrijednošću α, odbacuje se nulta hipoteza. Drugim riječima, u slučaju istinite nulte hipoteze, mala je vjerojatnost nastupa takvog ishoda temeljem danih podataka. Suprotno bi vrijedilo za *p*-vrijednost > α.

2.1.10. Predviđanje modelom jednostavne linearne regresije

U naslovu 1.2. pojašnjavala se metodologija ekonometrije, te je spomenuto kako se procijenjeni model koristi za testiranje hipoteza vezanih uz ekonomsku teoriju. Kao posljednji korak spomenuto je korištenje modela u prognostičke svrhe. Potrebno je napraviti razliku između **predviđanja** (engl. *prediction*) i **prognoziranja** (engl. *forecasting*) modelom (Greene, 2018). Predviđanje regresijskim modelom znači primjenu regresijskog modela kako bi se izračunale procijenjene (engl. *fitted*) vrijednosti zavisne varijable temeljem novih podataka o nezavisne varijable (koji mogu biti stvarni ili pretpostavljeni). Previđanje se može vršiti za presječne podatke, vremenske nizove i panel podatke. Prognoziranje se koristi kod vremenskih nizova, u svrhu predviđanja budućih vrijednosti zavisne varijable.

Drugim riječima, u predviđanju se razmatra scenarij kojeg istraživač sam razvija, stoga pretpostavlja sam neke vrijednosti nezavisne varijable koje će primijeniti u procijenjenom modelu; dok se prognoziranje odnosi na previđanje budućih vrijednosti zavisne varijable u kontekstu vremenskih nizova, temeljem budućih podataka o nezavisnim varijablama. U slučaju poznatih budućih vrijednosti nezavisnih varijabli govori se o *ex post* prognoziranju, dok u slučaju najprije prognoziranja budućih vrijednosti nezavisne varijable govori se o *ex ante* prognoziranju.

Ako se razmatra model jednostavne linearne regresije $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, te se pretpostavi da takav model vrijedi i u okolini promotrenih točaka (presječni podaci) ili u budućnosti, tada se **predviđena vrijednost zavisne varijable** temeljem opaženih ili pretpostavljenih vrijednosti nezavisne varijable x_f može zapisati predviđena vrijednost zavisne varijable ovako:

$$y_f = \beta_0 + \beta_1 x_f + \varepsilon_f, \tag{2.162}$$

pri čemu vrijedi $\varepsilon_f \sim N(0, \sigma^2)$ i ako vrijede pretpostavke regresijskog modela, procijenjena (odnosno predviđena) vrijednost iznosi:

$$\hat{y}_f = \hat{\beta}_0 + \hat{\beta}_1 x_f\,, \tag{2.163}$$

za koju vrijedi $\hat{y}_f \sim N(\beta_0 + \beta_1 x_f, \sigma^2 x_f (\boldsymbol{X}'\boldsymbol{X})^{-1} x_f)$. **Pogreška predviđanja** (engl. *prediction error*) računa se ovako:

$$\begin{aligned} \hat{\varepsilon}_f &= y_f - \hat{y}_f = \beta_0 + \beta_1 x_f + \varepsilon_f - \hat{\beta}_0 - \hat{\beta}_1 x_f \\ &= \beta_0 - \hat{\beta}_0 + \left(\beta_1 - \hat{\beta}_1\right) x_f + \varepsilon_f \end{aligned}, \tag{2.164}$$

čija je očekivana vrijednost jednaka:

$$\begin{aligned} E\left(\hat{\varepsilon}_f\right) &= E\left(\beta_0 - \hat{\beta}_0 + \left(\beta_1 - \hat{\beta}_1\right)x_f + \varepsilon_f\right) \\ &= \beta_0 - \beta_0 + \left(\beta_1 - \beta_1\right)x_f + 0 = 0 \end{aligned}. \tag{2.165}$$

Varijanca predviđanja (engl. *prediction variance*) jednaka je (vidjeti Maddala, 1992: 86):

$$\begin{aligned} Var\left(\hat{\varepsilon}_f\right) &= Var\left(y_f - \hat{y}_f\right) = Var\left(\beta_0 - \hat{\beta}_0 + \left(\beta_1 - \hat{\beta}_1\right)x_f + \varepsilon_f\right) \\ &= Var\left(\beta_0 - \hat{\beta}_0\right) + x_f^2 Var\left(\beta_1 - \hat{\beta}_1\right) + 2x_f Cov\left(\beta_0 - \hat{\beta}_0, \beta_1 - \hat{\beta}_1\right) + Var\left(\varepsilon_f\right) \\ &= \sigma^2\left(\frac{1}{N} + \frac{\overline{x}^2}{\hat{Var}(x)}\right) + \sigma^2\frac{x_f^2}{\hat{Var}(x)} - 2x_f\sigma^2\frac{\overline{x}^2}{\hat{Var}(x)} + \sigma^2 \\ &= \sigma^2\left(1 + \frac{1}{N} + \frac{\left(x_f - \overline{x}\right)^2}{\hat{Var}(x)}\right) \end{aligned} \tag{2.166}$$

Standardna devijacija predviđanja ili prognostička pogreška, računa se formulom:

$$SE\left(\hat{\varepsilon}_f\right) = \sqrt{\sigma^2\left(1 + \frac{1}{N} + \frac{\left(x_f - \overline{x}\right)^2}{\hat{Var}(x)}\right)} = \sigma\sqrt{\left(1 + \frac{1}{N} + \frac{\left(x_f - \overline{x}\right)^2}{\hat{Var}(x)}\right)}, \tag{2.167}$$

gdje je potrebno standardnu devijaciju regresije procijeniti temeljem izraza (2.73), odnosno izrazom $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{N-2}$. Uočava se da će vrijednost $SE\left(\hat{\varepsilon}_f\right)$ biti manja što je vrijednost $\hat{\sigma}$ manja, za što veći broj podataka *N* i kada je razlika $x_f - \overline{x}$ manja.

Standardizirana devijacija regresije slijedi *t*-distribuciju s *N*–2 stupnja slobode:

$$\frac{\hat{\varepsilon}_f}{SE\left(\hat{\varepsilon}_f\right)} = \frac{y_f - \hat{y}_f}{SE\left(y_f - \hat{y}_f\right)} \sim t\left(N-2\right), \tag{2.168}$$

stoga se **interval predviđanja (prognostički interval)** može procijeniti formulom:

$$P\left(-t_{\gamma/2} < \frac{y_f - \hat{y}_f}{SE\left(y_f - \hat{y}_f\right)} < t_{\gamma/2}\right) = 1 - \gamma, \tag{2.169}$$

odnosno

$$P\left(\hat{y}_f - t_{\gamma/2}SE\left(y_f - \hat{y}_f\right) < y_f < \hat{y}_f + t_{\gamma/2}SE\left(y_f - \hat{y}_f\right)\right) = 1 - \gamma, \tag{2.170}$$

gdje 1–γ predstavlja pouzdanost procjene, a $t_{\gamma/2}$ predstavlja koeficijent pouzdanosti (kao u 2.1.7). Sličan komentar kao za (2.167) slijedi: što je manja vrijednost $\hat{\sigma}$, i za što manju razliku

$x_f - \bar{x}$ će interval (2.170) biti uži. **Interpretacija intervala** (2.170) glasi: u (1–γ)100% slučajeva će, uz danu vrijednost nezavisne varijable x_f, vrijednost zavisne varijable iznositi između $\hat{y}_f - t_{\gamma/2} SE\left(y_f - \hat{y}_f\right)$ i $\hat{y}_f + t_{\gamma/2} SE\left(y_f - \hat{y}_f\right)$ jedinica.

Primjer 2.22.
Temeljem procijenjenog modela u kojemu BDP ovisi o indeksu HICP, koliko iznosi predviđena vrijednost BDP-a ako se pretpostavlja vrijednost HICP=120? Koliko iznosi interval predviđanja uz 1–γ=0.95? Interpretirajmo dobivene rezultate.

S obzirom na sliku 2.44., već je u prethodnim primjerima spomenuto da je procijenjen model: $\hat{y}_i = 2216386 - 15791 x_i$. Najprije je pomoću naredbe m1<–lm(GDP~HICP,data=podaci) taj model spremljen u RStudiju, a potom je temeljem naredbi prikazanih na slici 2.50. dobiven sljedeći rezultat. Najprije je potrebno spremiti nove podatke, koji se odnose na x_f, što je napravljeno u prvome retku naredbi, gdje je zadano $x_f = 120$, a potom pomoću naredbe predict() za spremljeni model m1 predviđamo vrijednost BDP-a za danu vrijednost HICP-a=120. U istoj naredbi se može procijeniti predviđena vrijednost $\hat{y}_f$ (u ispisu „fit“), kao i interval, temeljem zadane razine 1–γ (u pravilu za 95%-tni interval nije potrebno zadavati razinu putem naredbe level=…, jer RStudio ima ugrađenu razinu od 95%, no moguće je mijenjati razine temeljem te naredbe) dobivaju se donja i gornja (u ispisu „lwr“ i „upr“) granica intervala u (2.170).

```
novo <- data.frame(HICP=120)
predict(m1, newdata = novo, interval = 'confidence',level = 0.95)

##        fit      lwr     upr
## 1 321479.1 -1300992 1943951
```

Slika 2.50. Naredbe potrebne za predviđanje vrijednosti zavisne varijable i interval predviđanja

Uočava se da za $x_f = 120$ vrijedi: $\hat{y}_f = 2216386 - 15791 x_f = 321479{,}1$, što znači da se ovim modelom predviđa da za razinu indeksa cijena HICP u vrijednosti od 120, očekivana razina BDP-a neke zemlje iznosi u prosjeku 321479,1 milijuna eura. U 95% slučajeva za pretpostavljenu vrijednost indeksa HICP u iznosu 120 bodova, stvarna vrijednost BDP-a neke Europske zemlje bit će između –1300992 i 1943951 milijuna eura. Naravno, kako smo već u prethodnim primjerima ustvrdili da odabrani model nije reprezentativan u opisivanju varijacija zavisne varijable, te varijabla HICP nije značajna u modelu, ne čudi što je donja granica intervala negativna. No, ako se razmotri ekonomska interpretacija rezultata, uočava se da negativna vrijednost BDP-a nema smisla, što upućuje na moguću krivu specifikaciju modela.

Ako se iz prethodnih primjera gdje se razmatrao log-log model za slučaj istih podataka razmotri predviđena vrijednost BDP-a za model M2 (vidjeti primjer 2.10.), pa se u naredbi "predict" umjesto m1, zapiše m2 na slici 2.50., dobiva se sljedeći rezultat. Za vrijednost fit, lwr i upr su redom vrijednosti 11,74, 8,97 i 14,52. Kako se radi o log-log modelu $\widehat{\ln \text{BDP}_i} = 28{,}93 - 3{,}59 \ln \text{HICP}_i$, uvrštavanjem vrijednosti 120 bodova za varijablu HICP i izračun intervalnog predviđanja sada se odnose na logaritmiranu vrijednost BPD-a. Zato interpretacije rezultata glase ovako: u 95% slučajeva za pretpostavljenu vrijednost indeksa HICP u iznosu 120 bodova, očekivana razina BDP-a neke zemlje iznosi u prosjeku 125492,3 milijuna eura ($e^{11{,}74}$), odnosno stvarna vrijednost BDP-a neke Europske zemlje bit će između 7863,6 i 2022814 milijuna eura.

2.1.11. Sveobuhvatan primjer

Učitani su podaci u RStudio o udjelu populacije starije od 65 godina (udio izražen između 0 i 1 kao decimalni broj) i ukupna državna potrošnja na zdravstvo (u milijunima Eura) za 31 Europsku zemlju (datoteka „**udio_65.txt**“). Procijenimo model jednostavne linearne regresije u kojemu potrošnja (nazvana health u datoteci) ovisi o udjelu populacije starije od 65 godina (nazvana Udio_65).

a) Interpretirajmo procijenjene parametre modela, provedimo dvosmjerni *t*-test za nezavisnu varijablu i konstantu u modelu, te potom odgovarajući jednosmjerni test za nezavisnu varijablu, pri razini značajnosti od 5%. Mijenja li se ishod jednosmjernog testa za razinu značajnosti od 10%?

```
podaci<-read.table("udio_65.txt",header=T, sep="\t")
model<-lm(health~Udio_65,data=podaci)
summary(model)

##
## Call:
## lm(formula = health ~ Udio_65, data = podaci)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73662 -58529 -16401  11478 288591
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -146700     134536   -1.09    0.285
## Udio_65      1052420     701823    1.50    0.145
##
## Residual standard error: 86530 on 29 degrees of freedom
## Multiple R-squared:  0.07196,    Adjusted R-squared:  0.03996
## F-statistic: 2.249 on 1 and 29 DF,  p-value: 0.1445
```

Slika 2.51. Ispis procijenjenog modela u primjeru

```
abs(qt(0.05/2,29))

## [1] 2.04523

qt(1-0.05,29)

## [1] 1.699127
```

Slika 2.52. Kritične granice za dvosmjerni i jednosmjerni *t*-test

Temeljem ispisa na slici 2.51. moguće je interpretirati sljedeće. Procijenjeni model je: $\hat{y}_i = -146700 + 1052420x_i$. Konstanta u ovome primjeru nema smisla jer bi tumačenje bilo: ako je udio populacije starije od 65 godina u nekoj zemlji jednak 0, državna potrošnja na zdravstvo iznosila bi u prosjeku −146700 milijuna eura. Interpretacija procijenjenog parametra uz nezavisnu varijablu glasi: ako se udio populacije starije od 65 godina poveća za 1 jedinicu[32], državna potrošnja na zdravstvo se poveća u prosjeku za 1.052.420 milijuna eura.

[32] Jedna jedinica u ovome slučaju predstavlja povećanje od 1, odnosno 100%, s obzirom da varijabla poprima vrijednosti između 0 i 1. Čitatelju ostaje za vježbu zadatak da tu varijablu pomnoži sa 100%, te da potom procijeni novi model s tako definiranom varijablom i onda interpretira procijenjeni parametar, u čijem se slučaju interpretacija mijenja na povećanje udjela populacije starije od 65 godina od 1 postotni bod.

Dvosmjerni *t*-test za Udio_65: $\begin{matrix} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{matrix}$, $t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{1052420}{701823} = 1{,}50$, p-v = 0,145.

Kritična granica za α=5% i 29 stupnjeva slobode (N–2=31–2) iznosi $t_{0,05/2}(29) = t_{0,025}(29)$ = 2,045 (slika 55).

Kako je t_1=1,50 < 2.045=$t_{0,025}(29)$, odnosno p-vrijednost > α (0,145 > 0,05), ne odbacuje se nulta hipoteza. Zaključujemo kako na razini značajnosti od 5% ne odbacujemo hipotezu da varijabla Udio_65 nije značajna u modelu.

Jednosmjerni test na gornju granicu za varijablu Udio_65:

Odabran je test na gornju granicu zbog pozitivnog predznaka procijenjenog parametra uz nezavisnu varijablu.

$\begin{matrix} H_0 : \beta_1 = 0 \\ H_0 : \beta_1 > 0 \end{matrix}$, $t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{1052420}{701823} = 1{,}50$, p-v = 0,145/2=0,0725.

Kritična granica za α = 5% i 29 stupnjeva slobode (N–2 = 31–2) iznosi $t_{0,05}(29)$ = 1,699 (naredba qt(1–0.005,29–2)). Kako je t_1=1,50 < 1,699 = $t_{0,05}(29)$, odnosno p-vrijednost > α (0,0725 > 0,05), ne odbacuje se nulta hipoteza. Zaključujemo kako na razini značajnosti od 5% ne odbacujemo hipotezu da varijabla Udio_65 nije značajna u modelu.

Uočimo kako je u slučaju jednosmjernog testa p-vrijednost 0,0725, koja kada se uspoređuje s razinom značajnosti od 10% manja, tj. 0,0725 < 0,1, što bi upućivalo na odbacivanje nulte hipoteze. U takvome slučaju, ili pak kada se radi o graničnim slučajevima, da je p-vrijednost veoma blizu vrijednosti 0,05, korisno je primijeniti druge testove kojima bi se potvrdilo odbacivanje ili ne nulte hipoteze (Wald test, LR ili LM test).

Možemo još komentirati značajnost konstante u modelu. S obzirom na mali empirijski *t*-omjer (–1,09), odnosno veliku p-vrijednost (0,285), zaključuje se kako se ne odbacuje nulta hipoteza da konstanta nije značajna u modelu.

b) Interpretirajmo koeficijent determinacije, koeficijent jednostavne linearne korelacije, procijenjenu standardnu devijaciju regresije i procijenjeni koeficijent varijacije regresije.

Koeficijent determinacije (vidjeti sliku 2.51.) iznosi R^2 = 0,07196, što znači da je vrlo malo varijacija zavisne varijable (varijable državna potrošnja na zdravstvo) protumačeno odabranim modelom. Koeficijent jednostavne linearne korelacije dobivamo kao drugi korijen od koeficijenta determinacije (slika 2.53.), te iznosi R = 0,2683, a s obzirom da je predznak procijenjenog parametra uz varijablu Udio_65 pozitivan, zaključujemo kako postoji pozitivna, slaba do umjerena linearna veza između zavisne i nezavisne varijable.

Procjena standardne devijacije regresije iznosi $\hat{\sigma}$ = 86527,65, dok je procjena koeficijenta varijacije regresije jednaka 161.15% (slika 2.53.). Dakle, prosječno odstupanje stvarnih od procijenjenih vrijednosti državne potrošnje na zdravstvo iznosi 86527,65 milijuna eura, što je

relativno 161,15%. Dakle, odstupanja su relativno velika i to upućuje na slabu reprezentativnost modela, zajedno s malom vrijednošću R^2.

```
sqrt(summary(model)$r.squared)

## [1] 0.2682537

summary(model)$sigma

## [1] 86527.65

potrosnja<-podaci$health
(summary(model)$sigma/mean(potrosnja))*100

## [1] 161.1532
```

Slika 2.53. Naredbe za izračun R, $\hat{\sigma}$ i $\hat{V}$

c) Procijenimo standardizirani oblik modela i interpretirajmo rezultat.

Temeljem naredbi danih na slici 2.54., sljedeći je oblik procijenjenog modela sa standardiziranim varijablama: $\hat{y}_i^* = 0,268x_i^*$. Ako se udio stanovništva starijeg od 65 godina u zemlji poveća za jednu standardnu devijaciju, državna potrošnja na zdravstvo poveća se u prosjeku za 0,268 standardnih devijacija.

```
lm(scale(health)~0+scale(Udio_65),data=podaci)

##
## Call:
## lm(formula = scale(health) ~ 0 + scale(Udio_65), data = podaci)
##
## Coefficients:
## scale(Udio_65)
##         0.2683
```

Slika 2.54. Procjena modela sa standardiziranim varijablama

d) Provedimo F-test o značajnosti nezavisne varijable u modelu, pri razini značajnosti od 5%. Zajedno s F-testom, razmotrimo odgovarajuću tablicu ANOVA i interpretirajmo rezultat.

Empirijski F-omjer izračunat je temeljem ispisa na slici 2.51.:

$$F = \frac{1,68 \cdot 10^{10} / 1}{2,17 \cdot 10^{11} / 29} = 2,2487 \text{ , } p\text{-v} = 0,1445.$$

Teorijski F-omjer dan je temeljem tablice kritičnih vrijednosti F-distribucije uz $\alpha = 5\%$, te 1 stupanj slobode u brojniku i 29 u nazivniku (slika 2.55.) iznosi $F_{0,05}(1;29) = 4,183$.
Stoga vrijedi: $F_{emp} = 0,1445 < 4,183 = F_{0,05}(1;29)$, pa se nulta hipoteza ne može odbaciti. Kako vrijedi i p-vrijednost $= 0,1445 > 0,05 = \alpha$, dolazi se do istog zaključka.

Riječima: uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla Udio_65 nije značajna u modelu.

```
qf(1-0.05,1,29)
## [1] 4.182964

anova(model)

## Analysis of Variance Table
##
## Response: health
##           Df     Sum Sq    Mean Sq F value Pr(>F)
## Udio_65    1 1.6836e+10 1.6836e+10  2.2487 0.1445
## Residuals 29 2.1712e+11 7.4870e+09
```

Slika 2.55. Tablica ANOVA i teorijski *F*-omjer

e) Interpretirajmo intervalnu procjenu parametara modela, uz razinu pouzdanosti od 90%.

Temeljem ispisa na slici 2.56., intervalne procjene parametara zapisujemo kao:

$P(-375293{,}4 < \beta_0 < 81893{,}38) = 0.90$ i $P(-140066{,}80 < \beta_1 < 2244905{,}84) = 0{,}90$, te je interpretacija sljedeća. Uz razinu pouzdanosti od 90%, ako je udio stanovništva starijeg od 65 godina u populaciji jednak 0, državna potrošnja na zdravstvo se kreće u prosjeku između –375293,40 i 81893,38 milijuna eura. Također, uz razinu pouzdanosti od 90% ako se udio stanovništva starijeg od 65 godina u populaciji poveća za jednu jedinicu, državna potrošnja na zdravstvo će se promijeniti u prosjeku za –140066,80 do 2244905,84 milijuna eura.

```
confint(model,level=0.90)

##                   5 %       95 %
## (Intercept) -375293.4   81893.38
## Udio_65     -140066.8 2244905.84
```

Slika 2.56. Intervalna procjena parametara modela

Uočimo kako je u intervalu $P(-140066{,}80 < \beta_1 < 2244905{,}84) = 0.90$ uključena vrijednost 0, što znači da je uključeno i tumačenje da varijabla Udio_65 nema učinka na zavisnu varijablu, što je u skladu sa zaključkom *t*-testa.

f) Provedimo Waldov test o značajnosti nezavisne varijable u modelu uz razinu značajnosti od 5%.

```
library(car)
ogranicenje<-"Udio_65=0"
linearHypothesis(model,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## Udio_65 = 0
##
## Model 1: restricted model
## Model 2: health ~ Udio_65
##
##   Res.Df        RSS Df  Sum of Sq  Chisq Pr(>Chisq)
## 1     30 2.3396e+11
## 2     29 2.1712e+11  1 1.6836e+10 2.2487     0.1337
```

Slika 2.57. Ispis Waldova testa

Temeljem ispisa na slici 2.57., proveden je Waldov test kako slijedi:

$$H_0: \boldsymbol{R\beta} = \boldsymbol{q} \\ H_1: \boldsymbol{R\beta} \neq \boldsymbol{q}, \quad \boldsymbol{R} = [0\ 1], \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{q} = [0], \quad W = 2{,}2487, p\text{-v} = 0{,}1337.$$

Kako je test veličina izračunata naredbom qchisq(0.95,1), i iznosi 3,841, zaključak je kako je $W = 2{,}2487 < 3{,}841$, odnosno da je p-vrijednost = 0,1337 > 0,05 = α, stoga se ne odbacuje nulta hipoteza. Uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla Udio_65 nije značajna u modelu.

g) Provedimo LR test o značajnosti nezavisne varijable u modelu uz razinu značajnosti od 5%.

Temeljem ispisa danog na slici 2.58., proveden je LR test kako slijedi:

$$H_0: \boldsymbol{R\beta} = \boldsymbol{q} \\ H_1: \boldsymbol{R\beta} \neq \boldsymbol{q}, \boldsymbol{R} = [0\ 1], \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{q} = [0], LR = -2(-396{,}53-(-395{,}37)) = 2{,}315, p\text{-v} = 0{,}128.$$

```
library(lmtest)

m2<-lm(health~1,data=podaci)
lrtest(model,m2)

## Likelihood ratio test
##
## Model 1: health ~ Udio_65
## Model 2: health ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -395.37
## 2   2 -396.53 -1 2.3151     0.1281
```

Slika 2.58. Ispis LR testa

Kako je test veličina izračunata naredbom qchisq(0.95,1), i iznosi 3.841, zaključak je kako je $LR = 2{,}3151 < 3{,}841$, odnosno da je p-vrijednost = 0,1281 > 0,05= α, stoga se ne odbacuje nulta hipoteza. Uz razinu značajnosti od 5%, ne odbacujemo hipotezu da varijabla Udio_65 nije značajna u modelu.

h) Uz prethodno procijenjen linearni model, procijenimo još drugi model, log-lin, te potom spojimo pomoću naredbe *stargazer* rezultate ispisa oba modela i potom ih usporedimo.

U slučaju potrebe procjene više modela koji se potom uspoređuju, korisna je naredba stargazer koja spaja rezultate u jednu preglednu tablicu. Tako je usporedba originalnog modela s modelom 2 iz ovog pitanja predočena na slici 2.59. Model 1 je već bio spremljen iz prethodnih postupaka (slika 2.51.), te je sada spremljen model 2 (slika 2.59.) i zajedno su prikazani pomoću naredbe stargazer. Prvi dio ispisa predočava procijenjene parametre zajedno s procijenjenim standardnim pogreškama procjenitelja u zagradama. Donji dio ispisa predočava nekoliko mjera prikladnosti svakog modela. Konačno, uz procijenjene parametre mogu se uočiti oznake *, ** ili *** ako je odabrana varijabla značajna u modelu (dvosmjerni test) uz 10%, 5% ili 1%.

Uočavamo kako se razlikuju procijenjeni parametri u oba modela, jer se u drugome modelu razmatra logaritmirana vrijednost zavisne varijable. Dok je prvi model: $\hat{y}_i = -146700 + 1052420x_i$, sada je drugi model: $\widehat{\ln y}_i = 5{,}67 + 21{,}08\, x_i$.

Interpretacija procijenjenih parametara u drugome modelu je sljedeća. Kada bi udio stanovništva starijeg od 65 godina u populaciji iznosio 0, u prosjeku bi državna potrošnja na zdravstvo iznosila $e^{5.67}$ milijuna eura. Povećanje udjela stanovništva starijeg od 65 godina u populaciji za jednu jedinicu vodi povećanju državne potrošnje na zdravstvo u prosjeku za 2108%.

Čitatelju se za vježbu ostavlja provedba dvosmjernog i jednosmjernog testa za varijablu Udio_65, te dvosmjerni test za konstantu u modelu. Za izračun empirijskih *t*-omjera se uz procijenjene parametre koriste procijenjene standardne greške u zagradama. Uočava se temeljem ispisa kako će ishod za konstantu biti odbacivanje nulte hipoteze pri razini značajnosti od 5%, dok će varijabla Udio_65 imati ishod ne odbacivanja nulte hipoteze pri uobičajenim razinama značajnosti.

```
model2<-lm(log(health)~Udio_65,data=podaci)

library(stargazer)

stargazer(list(model,model2),type="text")

##
## ============================================================
##                                    Dependent variable:
##                                ---------------------------
##                                    health      log(health)
##                                      (1)           (2)
## ------------------------------------------------------------
## Udio_65                          1,052,420.000     21.082
##                                  (701,822.900)    (13.387)
##
## Constant                         -146,700.000     5.670**
##                                  (134,535.800)    (2.566)
##
## ------------------------------------------------------------
## Observations                          31             31
## R2                                   0.072         0.079
## Adjusted R2                          0.040         0.047
## Residual Std. Error (df = 29)    86,527.650        1.650
## F Statistic (df = 1; 29)             2.249         2.480
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

Slika 2.59. Usporedba lin-lin i log-lin modela

Nadalje, uočava se da je koeficijent determinacije veći u slučaju drugog modela, kao i što je veći korigirani koeficijent determinacije. Stoga je veći udio varijacija zavisne varijable objašnjen drugim modelom. Čitatelju se ostavlja za vježbu izračun koeficijenta jednostavne linearne korelacije za drugi model, kao i njegova interpretacija.

Konačno, za provedbu *F*-testa, dan je empirijski *F*-omjer u drugome modelu koji iznosi 2,480, koji se uspoređuje s teorijskim omjerom iz postupka d), koji je iznosio 4,183 = $F_{0,05}(1;29)$. Kako je i u slučaju modela 2 empirijski *F*-omjer manji od teorijske razine, zaključak je kako na razini značajnosti od 5% ne odbacujemo hipotezu da varijabla Udio_65 nije značajna u modelu.

i) Temeljem procijenjenog modela 1 (u postupku a), koliko iznosi predviđena vrijednost državne potrošnje na zdravstvo ako se pretpostavlja vrijednost udjela stanovništva starijeg od 65 godina u populaciji u iznosu 0,2 (20%)? Koliko iznosi interval predviđanja uz $1-\gamma = 0{,}95$? Interpretirajmo dobivene rezultate.

```
novo <- data.frame(Udio_65=0.2)
predict(model, newdata = novo, interval = 'confidence',level = 0.95)

##        fit      lwr      upr
## 1 63783.91 29147.46 98420.36
```

Slika 2.60. Naredbe potrebne za predviđanje vrijednosti zavisne varijable i interval predviđanja

Uočava se da za $x_f = 0{,}2$ vrijedi: $\hat{y}_f = 63783{,}91$, što znači da se ovim modelom predviđa da za udio stanovništva starijeg od 65 u populaciji u iznosu od 0,2 očekivana razina državne potrošnje na zdravstvo neke zemlje iznosi u prosjeku 63783,91 milijuna eura.

U 95% slučajeva za pretpostavljenu vrijednost udjela stanovništva starijeg od 65 u populaciji u iznosu od 0,2, stvarna vrijednost državne potrošnje na zdravstvo neke zemlje će se kretati između 29147,46 i 98420,36 milijuna eura.

2.1.12. Pitanja za ponavljanje

1) Za sljedeće procijenjene modele, interpretirajte procijenjene parametre:

   a. $\hat{y}_i = 3 + x_i$ c. $\widehat{\ln y}_i = 3 + x_i$ e. $\widehat{\ln y}_i = 3 - x_i$ g. $\widehat{\ln y}_i = 3 + \ln x_i$
   b. $\hat{y}_i = 3 - x_i$ d. $\hat{y}_i = 3 + \ln x_i$ f. $\hat{y}_i = 3 - \ln x_i$ h. $\widehat{\ln y}_i = 3 - \ln x_i$

2) Zašto se ostavlja konstanta u modelu prilikom procjene parametara?
3) Što je to intervalna procjena parametara?
4) Interpretirajte sljedeću intervalnu procjenu parametra uz varijablu x, ako se razmatra ovisnost potrošnje (y) u kn o dohotku (x): $\hat{y}_i = 0{,}5 + 0{,}5x_i$, $P(0{,}1 < \beta_1 < 0{,}3) = 0{,}88$.
5) Iz prethodnog zadatka interpretirajte sljedeće: $P(0{,}01 < \beta_0 < 0{,}55) = 0{,}99$.
6) Što je to model sa standardiziranim regresijskim varijablama? Kako se vrši interpretacija procijenjenih parametara? Kada se koristi ovaj model?
7) Čemu služi tablica ANOVA? Objasnite njeno popunjavanje.
8) Što je procjena standardne devijacije regresije u modelu linearne regresije i kako ju interpretiramo? Koji je nedostatak ove mjere?
9) Koja relativna mjera reprezentativnosti modela se koristi uz procjenu standardne devijacije regresije i kako ju interpretiramo?
10) Što je koeficijent determinacije regresije? Kako ga interpretiramo?
11) Što je koeficijent jednostavne linearne korelacije? Kako ga interpretiramo?
12) Skicirajte dijagrame rasipanja za slučaj: a) jake pozitivne linearne korelacije; b) potpuno odsustvo korelacije; c) blage negativne linearne korelacije za slučaj dvije varijable.
13) Što je to korigirani koeficijent determinacije regresije? Kako ga interpretiramo?
14) Temeljem simuliranih podataka o varijabli x i y, procijenjeno je nekoliko jednostavnih linearnih regresijskih modela čiji je ispis dan u tablici. Temeljem odgovarajućih mjera reprezentativnosti modela, odaberimo najbolji model i objasnite zašto je najbolji.

Parametar ili mjera / Model:	**M1**	**M2**	**M3**
R^2	0,22	0,22	0,22
$\overline{R}^2$	0,21	0,18	0,20
$\hat{\sigma}^2$	20	22	20
$\hat{V}$	1%	12%	2%

15) Čemu služi t-test? Koje su hipoteze testa u modelu jednostavne linearne regresije? Kako donosimo odluku o ishodu testa za jednosmjerni, a kako za dvosmjerni test? Kako ćete odrediti provodite li jednosmjerni ili dvosmjerni test?
16) Čemu služi F-test? Koje su hipoteze testa u modelu jednostavne linearne regresije? Kako donosimo odluku o ishodu testa?
17) Što je to p-vrijednost i kako donosimo odluku o ishodu testa temeljem p-vrijednosti?
18) Čemu služe Waldov, LR i LM test?
19) Zapišite matrično hipoteze Waldova testa o značajnosti nezavisne varijable u slučaju jednostavne linearne regresije.
20) Na čemu se temelji LR test? Kakve je modele potrebno procijeniti za provedbu tog testa?
21) Na čemu se temelji LR test? Kakav je model potrebno procijeniti za provedbu tog testa?
22) Koja je razlika između predviđanja i prognoziranja?
23) Interpretirajte sljedeći rezultat predviđanja za model procijenjen u zadatku 4): $P(200 < y_f < 250) = 0{,}95$.

24) Učitajte datoteku „**phillips.txt**“ u RStudio. Datoteka sadrži simulirane podatke o stopi inflacije i stopi nezaposlenosti. Procijenite model u kojemu stopa inflacije ovisi o stopi nezaposlenosti, te potom provedite sve postupke a) – i) iz sveobuhvatnog primjera (bez postupka h) 2.1.11. U postupku pod i) se pretpostavlja da je zadana stopa nezaposlenosti 0,15.
25) Skicirajte dijagram rasipanja u RStudiju za podatke iz prethodnog zadatka. Kako se uočava da je Phillipsov model takav u kojemu postoji recipročna veza između stope inflacije i stope nezaposlenosti, procijenite takav model i ponovno provedite sve postupke a) – i) iz sveobuhvatnog primjera (bez postupka h) 2.1.11. U postupku pod i) se pretpostavlja da je zadana stopa nezaposlenosti 0,15.
26) Pomoću naredbe *stargazer* spojite rezultate procjena modela iz zadatka 24) i 25) te ih usporedite koji bolje opisuje podatke i zašto.

Rješenja

Zadatak 1):

a) Ako se vrijednost nezavisne varijable poveća za 1 jedinicu, vrijednost zavisne varijable će se povećati u prosjeku za jednu jedinicu.
b) Ako se vrijednost nezavisne varijable poveća za 1 jedinicu, vrijednost zavisne varijable će se smanjiti u prosjeku za jednu jedinicu.
c) Ako se vrijednost nezavisne varijable poveća za 1 jedinicu, vrijednost zavisne varijable će se povećati u prosjeku za 100%.
d) Ako se vrijednost nezavisne varijable poveća za 1%, vrijednost zavisne varijable će se povećati u prosjeku za 0,01 jedinica.
e) Ako se vrijednost nezavisne varijable poveća za 1 jedinicu, vrijednost zavisne varijable će se smanjiti u prosjeku za 100%.
f) Ako se vrijednost nezavisne varijable poveća za 1%, vrijednost zavisne varijable će se smanjiti u prosjeku za 0,01 jedinica.
g) Ako se vrijednost nezavisne varijable poveća za 1%, vrijednost zavisne varijable će se povećati u prosjeku za 1%.
h) Ako se vrijednost nezavisne varijable poveća za 1%, vrijednost zavisne varijable će se smanjiti u prosjeku za 0,01 jedinica.

Zadatak 4):

Uz razinu pouzdanosti od 88%, ako se dohodak poveća za 1 kn, stvarna vrijednost potrošnje će se povećati u prosjeku između 0,1 i 0,3 kn.

Zadatak 5):

Uz razinu pouzdanosti od 99%, ako bi dohodak iznosio 0 kn, stvarna vrijednost potrošnje iznosi između 0,01 kn i 0,55 kn.

Zadatak 14):

Najbolji je model M1 jer iako sva tri imaju jednaku vrijednost koeficijenta determinacije, ima najveću vrijednost korigiranog koeficijenta determinacije, te najmanju vrijednost koeficijenta varijacije regresije.

Zadatak 23):

Uz razinu pouzdanosti od 95%, modelom se predviđa da stvarna vrijednost potrošnje iznosi između 200kn i 250 kn.

Zadatak 24):

```
podaci<-read.table("phillips.txt",header=T, sep="\t")
model<-lm(s_inflacije~s_nezaposlenosti,data=podaci)
summary(model)

## Call:
## lm(formula = s_inflacije ~ s_nezaposlenosti, data = podaci)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.150 -3.199 -1.385  1.936 28.257
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        17.419      2.087   8.345 1.56e-09 ***
## s_nezaposlenosti  -23.286      5.202  -4.476 9.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.952 on 32 degrees of freedom
## Multiple R-squared:  0.385,  Adjusted R-squared:  0.3658
## F-statistic: 20.04 on 1 and 32 DF,  p-value: 9.04e-05
```

```
abs(qt(0.05/2,34-2))

## [1] 2.036933

qt(1-0.05,34-2)

## [1] 1.693889

sqrt(summary(model)$r.squared)

## [1] 0.6205206

summary(model)$sigma

## [1] 5.951779

s_inflacije<-podaci$s_inflacije
(summary(model)$sigma/mean(s_inflacije))*100

## [1] 64.21288
```

```
lm(scale(s_inflacije)~0+scale(s_nezaposlenosti),data=podaci)

## Call:
## lm(formula = scale(s_inflacije) ~ 0 + scale(s_nezaposlenosti),
##     data = podaci)
##
## Coefficients:
## scale(s_nezaposlenosti)
##                 -0.6205
qf(1-0.05,1,34-2)

## [1] 4.149097
```

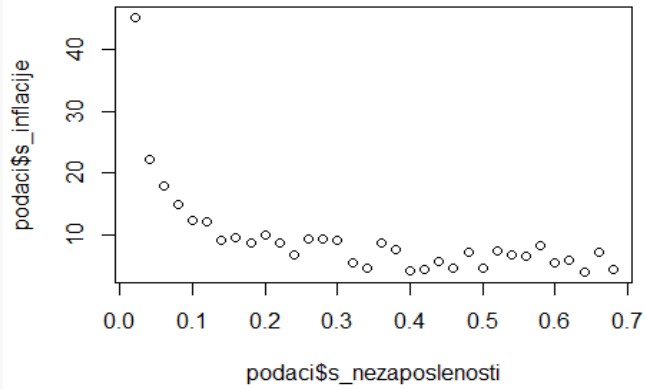
```
anova(model)

## Analysis of Variance Table
##
## Response: s_inflacije
##                  Df  Sum Sq Mean Sq F value   Pr(>F)
## s_nezaposlenosti  1  709.76  709.76  20.036 9.04e-05 ***
## Residuals        32 1133.56   35.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

```
confint(model,level=0.90)

##                         5 %      95 %
## (Intercept)       13.88308  20.95446
## s_nezaposlenosti -32.09730 -14.47382

library(car)
ogranicenje<-"s_nezaposlenosti=0"
linearHypothesis(model,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## s_nezaposlenosti = 0
##
## Model 1: restricted model
## Model 2: s_inflacije ~ s_nezaposlenosti
##
##   Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1     33 1843.3
## 2     32 1133.6  1    709.76 20.036  7.598e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(lmtest)
m2<-lm(s_inflacije~1,data=podaci)
lrtest(model,m2)

## Likelihood ratio test
##
## Model 1: s_inflacije ~ s_nezaposlenosti
## Model 2: s_inflacije ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -107.86
## 2   2 -116.12 -1 16.531  4.786e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#predviđanje:
novo <- data.frame(s_nezaposlenosti=0.15)
predict(model, newdata = novo, interval = 'confidence',level = 0.95)

##        fit      lwr      upr
## 1 13.92594 10.95708 16.89479
```

Zadatak 25):




```
model<-lm(s_inflacije~I(1/s_nezaposlenosti),data=podaci)
summary(model)

## Call:
## lm(formula = s_inflacije ~ I(1/s_nezaposlenosti), data = podaci)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.26387 -1.23447  0.08129  1.06495  2.44700
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.43759    0.29624   14.98 5.17e-16 ***
## I(1/s_nezaposlenosti)  0.79773    0.02718   29.35  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.436 on 32 degrees of freedom
## Multiple R-squared:  0.9642, Adjusted R-squared:  0.9631
## F-statistic: 861.6 on 1 and 32 DF,  p-value: < 2.2e-16
```

```
abs(qt(0.05/2,34-2))

## [1] 2.036933

qt(1-0.05,34-2)

## [1] 1.693889

sqrt(summary(model)$r.squared)

## [1] 0.9819316

summary(model)$sigma

## [1] 1.436249

s_inflacije<-podaci$s_inflacije
(summary(model)$sigma/mean(s_inflacije))*100

## [1] 15.49548

lm(scale(s_inflacije)~0+scale(I(1/s_nezaposlenosti)),data=podaci)
```

```
##
## Call:
## lm(formula = scale(s_inflacije) ~ 0 + scale(I(1/s_nezaposlenosti)),
##     data = podaci)
##
## Coefficients:
## scale(I(1/s_nezaposlenosti))
##                       0.9819

qf(1-0.05,1,34-2)

## [1] 4.149097

anova(model)

## Analysis of Variance Table
##
## Response: s_inflacije
##                        Df  Sum Sq Mean Sq F value    Pr(>F)
## I(1/s_nezaposlenosti)  1 1777.31 1777.31   861.6 < 2.2e-16 ***
## Residuals             32   66.01    2.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(model,level=0.90)

##                              5 %      95 %
## (Intercept)           3.9357854 4.9393975
## I(1/s_nezaposlenosti) 0.7516991 0.8437697

library(car)
ogranicenje<-"I(1/s_nezaposlenosti)=0"
linearHypothesis(model,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## I(1/s_nezaposlenosti) = 0
##
## Model 1: restricted model
## Model 2: s_inflacije ~ I(1/s_nezaposlenosti)
##
##   Res.Df     RSS Df Sum of Sq Chisq Pr(>Chisq)
## 1     33 1843.32
## 2     32   66.01  1    1777.3 861.6  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(lmtest)
m2<-lm(s_inflacije~1,data=podaci)
lrtest(model,m2)

## Likelihood ratio test
##
## Model 1: s_inflacije ~ I(1/s_nezaposlenosti)
## Model 2: s_inflacije ~ 1
##   #Df   LogLik Df Chisq Pr(>Chisq)
## 1   3  -59.522
## 2   2 -116.124 -1 113.2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#predviđanje:
novo <- data.frame(s_nezaposlenosti=0.15)
predict(model, newdata = novo, interval = 'confidence',level = 0.95)

##        fit      lwr      upr
## 1 9.755821 9.252957 10.25868
```

Zadatak 26):

```
m1<-lm(s_inflacije~s_nezaposlenosti,data=podaci)
m2<-lm(s_inflacije~I(1/s_nezaposlenosti),data=podaci)
library(stargazer)
stargazer(list(m1,m2),type="text")

## ============================================================
##                                   Dependent variable:
##                               ---------------------------
##                                       s_inflacije
##                                   (1)            (2)
## ------------------------------------------------------------
## s_nezaposlenosti                -23.286***
##                                  (5.202)
##
## I(1/s_nezaposlenosti)                           0.798***
##                                                  (0.027)
##
## Constant                        17.419***      4.438***
##                                  (2.087)        (0.296)
##
## ------------------------------------------------------------
## Observations                        34             34
## R2                                 0.385          0.964
## Adjusted R2                        0.366          0.963
## Residual Std. Error (df = 32)      5.952          1.436
## F Statistic (df = 1; 32)         20.036***     861.596***
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

2.2. Model višestruke linearne regresije

U ekonomiji rijetko jedna varijabla ovisi samo jednoj drugoj varijabli. Češće se događa da ovisi o mnogo drugih varijabli, što je korisno opisati modelom **višestruke linearne regresije**. U ovome poglavlju obrađuje se model u kojemu se pretpostavlja da jedna zavisna varijabla, y, ovisi o dvije ili više nezavisnih varijabli, x_1, x_2, x_3…, x_k . Već je uvedena oznaka za zavisnu varijablu, y, dok se nezavisne označavaju s x_1, x_2, ..., x_k. Općenito ćemo govoriti o i-toj regresijskoj varijabli, x_j, $j \in \{1, 2, \ldots, k\}$. Pritom se pretpostavlja određeni funkcionalni oblik između x_1, x_2, ..., x_k i y, tako da je $y = f(x_1, x_2, ..., x_k) + \varepsilon$. Kako se radi o dvije ili više nezavisnih varijabli, ovaj oblik modela se naziva model višestruke linearne regresije, odnosno višestruki linearni model.

2.2.1. Osnovna terminologija

Model višestruke linearne regresije zapisuje se ovako:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k + \varepsilon \tag{2.171}$$

odnosno ako se zapisuje za svaki entitet u slučaju presječnih podataka, na sljedeći način:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + .. + \beta_k x_{ik} + \varepsilon_i \,. \tag{2.172}$$

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ su **nepoznati parametri koje je potrebno procijeniti**. Ako se model (2.172) želi procijeniti za sva opažanja u slučaju presječnih podataka, $i \in \{1, 2, \ldots, N\}$, tada se razmatra sustav od N jednadžbi:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ y_N &= \beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} + ... + \beta_k x_{Nk} + \varepsilon_N \end{aligned} \,. \tag{2.173}$$

Model je moguće zapisati u matričnoj formi:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.174}$$

gdje je $\boldsymbol{y} \in \mathbb{R}^N$ vektor stupac čiji elementi su opažanja zavisne varijable, $\boldsymbol{X} \in \mathcal{M}_{N,k+1}$ je matrica čiji prvi stupac čine jedinice, a ostale stupce čine vrijednosti opažanja nezavisnih varijabli, $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ je vektor stupac nepoznatih parametara, dok je $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ vektor stupac slučajne varijable:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \tag{2.175}$$

Prvi stupac u matrici $\boldsymbol{X}$ čine jedinice, s obzirom na umnožak s vektorom $\boldsymbol{\beta}$ čiji je prvi element konstanta, kako bi ta konstanta bila uključena u model. Programska podrška za procjenu

ekonometrijskih modela koristi matrični zapis modela temeljem kojeg se procjenjuju nepoznati parametri.

2.2.2. Pretpostavke modela višestruke linearne regresije

Pretpostavke modela višestruke linearne regresije su sljedeće (Greene, 2002):

1. Linearnost modela – pretpostavlja se linearna veza između zavisne i nezavisnih varijabli. Drugim riječima se može reći da je y linearna kombinacija k nezavisnih varijabli.
2. Egzogenost podataka u matrici $\boldsymbol{X}$, tj. egzogenost nezavisne varijable: $E(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = 0$. To znači da očekivana vrijednost greške relacije ne ovisi o vrijednostima nezavisnih varijabli. Ova pretpostavka vrijedi ako su nezavisne varijable determinističke (u ponovljenim mjerenjima su vrijednosti nezavisnih varijabli fiksne).
3. Greška relacije u prosjeku ne utječe na zavisnu varijablu:
   $E(\varepsilon_i) = 0, \ \forall i$, tj. $E(y_i \mid x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}, \ \forall i$.
4. Varijanca greške relacije je konstantna (homoskedastična): $Var(\boldsymbol{\varepsilon}) = Var(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \sigma^2$.
5. Nezavisnost slučajne varijable, tj. nekoreliranost:
   $E(\varepsilon_i \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j \mid x_i) = 0$ za $i \neq j$.
6. Slučajna varijabla normalno je distribuirana: $\varepsilon_i \sim N(0, \sigma^2), \ \forall i$.
7. Varijable x_j su međusobno nezavisne, što znači da vrijedi $r(\boldsymbol{X}'\boldsymbol{X})^{-1} = r(\boldsymbol{X}) = k + 1$, tj. postoji inverz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Dakle, u odnosu na pretpostavke modela jednostavne linearne regresije, **dodaje se sedma pretpostavka** da su sve regresorske varijable međusobno nezavisne. Drugim riječima, stupci u matrici $\boldsymbol{X}$ linearno su neovisni, što znači da je za potrebe procjene parametara regresijskog modela u (2.172) moguće izračunati $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ (vidjeti naslove 2.2.3., kao i već obrađen 2.1.3.2). Ako su stupci u $\boldsymbol{X}$ linearno neovisni, tada je rang te matrice jednak $r(\boldsymbol{X}) = k + 1$, te je ujedno i rang matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ jednak $k + 1$, odnosno moguće je izračunati $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (tj. matrica $\boldsymbol{X}'\boldsymbol{X}$ je invertibilna).

2.2.3. Metoda najmanjih kvadrata i procjenitelj za slučaj višestruke linearne regresije

Razmatra se model višestruke linearne regresije u matričnom obliku:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.176}$$

gdje je $\boldsymbol{y} \in \mathbb{R}^N$ vektor stupac čiji elementi su opažanja zavisne varijable, $\boldsymbol{X} \in \mathcal{M}_{N,k+1}$ je matrica čiji prvi stupac čine jedinice, a ostale stupce čine vrijednosti opažanja nezavisnih varijabli, $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ je vektor stupac nepoznatih parametara, dok je $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ vektor stupac slučajne varijable:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \tag{2.177}$$

Procijenjeni model je:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{2.178}$$

gdje je $\hat{\boldsymbol{y}} \in \mathbb{R}^N$ vektor stupac s procijenjenim vrijednostima zavisne varijable, a $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{k+1}$ je vektor stupac procijenjenih parametara. Sada je vektor rezidualnih odstupanja definiran kao $\hat{\boldsymbol{\varepsilon}} \in \mathbb{R}^N$ vektor stupac:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{2.179}$$

$$\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}. \tag{2.180}$$

Minimizira se suma kvadrata odstupanja stvarnih od procijenjenih vrijednosti zavisne varijable, što se zapisuje kao:

$$\arg\min_{\hat{\boldsymbol{\beta}}} \left(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\right) = \arg\min_{\hat{\boldsymbol{\beta}}} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = S\left(\hat{\boldsymbol{\beta}}\right), \tag{2.181}$$

gdje je $S\left(\hat{\boldsymbol{\beta}}\right)$ oznaka za funkciju cilja. Izraz $\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$ može se zapisati kao: $\boldsymbol{y}'\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$ (vidjeti relaciju (2.38)). Nužni uvjet za postojanje minimuma implicira izjednačavanje svih parcijalnih derivacija prvog reda funkcije cilja s nulom, tj. vektor kojemu su komponente parcijalne derivacije prvog reda izjednačavamo s nul-vektorom (vidjeti postupke (2.39) - (2.44)) kako bi se izračunao vektor procijenjenih parametara:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.182}$$

S obzirom na uvedene oznake, pretpostavke iz naslova 2.2.2 se mogu zapisati u obliku:

1. Linearnost modela – pretpostavlja se linearna veza između zavisne i nezavisnih varijabli.
2. Egzogenost podataka u matrici $\boldsymbol{X}$, tj. egzogenost nezavisne varijable: $E(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \boldsymbol{0}$.
3. Greška relacije u prosjeku ne utječe na zavisnu varijablu: $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, tj. $E(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$.
4. Varijanca greške relacije je konstantna (homoskedastična).
5. Nezavisnost slučajne varijable, tj. nekoreliranost.
6. Slučajna varijabla normalno je distribuirana: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$.
7. Varijable x_i su međusobno nezavisne (ne postoji multikolinearnost između varijabli), što znači da vrijedi $r(\boldsymbol{X}'\boldsymbol{X})^{-1} = r(\boldsymbol{X}) = k + 1$, tj. postoji inverz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Pretpostavke 4 i 5 pišu se na sljedeći način zajedno:

$$\Omega = E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I \qquad (2.183)$$

Kako se radi o jediničnoj matrici I u (2.183), ponekad se u literaturi označava kojeg je formata, s obzirom na broj presječnih podataka, I_N, ili pak s obzirom na broj opažanja u slučaju vremenskih nizova, I_T.

Za linearni regresijski model (2.172), pretpostavke koje se odnose na slučajnu varijablu, nazivamo **Gauss-Markovljevim uvjetima**, i formalno se skraćeno mogu zapisati kao: $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ tj. $E(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X\beta}$, $Var(\boldsymbol{y}) = Var(\boldsymbol{\varepsilon}) = \sigma^2$, $Cov(y_i,y_j) = Cov(\varepsilon_i,\varepsilon_j) = 0$ za $i \neq j$, te se u odnosu na pretpostavke jednostavnog linearnog regresijskog modela upotpunjuju još uvjetom $r(\boldsymbol{X}'\boldsymbol{X})^{-1} = r(\boldsymbol{X}) = k + 1$.

Svojstva procjenitelja $\hat{\boldsymbol{\beta}}$ u (2.182) su identična onima u naslovu 2.1.3.3. Stoga će se samo kratko ponoviti. Radi se o **linearnom** procjenitelju jer je upravo linearna kombinacija dobivena izračunom u (2.182). **Nepristran** je procjenitelj jer vrijedi:

$$E(\hat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) = (\boldsymbol{Xq}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta}, \qquad (2.184)$$

s obzirom da je $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = I$. Matrica varijanci-kovarijanci procjenitelja jednaka je (vidjeti izvod u 2.1.3.3):

$$Var\left(\hat{\boldsymbol{\beta}}\right) = Var\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'Var\left(\boldsymbol{y}\right)\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}, \qquad (2.185)$$

te je **normalno distribuiran procjenitelj** s obzirom da je linearna kombinacija varijabli koje su normalno distribuirane, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1})$. Ponovno se radi o *BLUE* procjenitelju (najbolji linearni nepristrani procjenitelj). Može se pisati i $\hat{\beta}_i \mid \boldsymbol{X} \sim N\left(\beta_i, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{jj}\right)$ gdje jj predstavlja j-ti element na glavnoj dijagonali matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, $j \in \{1, 2, \ldots, k\}$.

Nepristran procjenitelj varijance za slučaj višestruke linearne regresije je (vidjeti izvod u 2.1.3.3):

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}'}{N-k-1} = \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{N-k-1} \sim \chi^2(N-k-1). \qquad (2.186)$$

Matrica varijanci-kovarijanci procjenitelja jednaka je

$$Var\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}, \qquad (2.187)$$

dok su **procijenjene standardne pogreške procjenitelja** dane izrazom

$$SE(\hat{\beta}_j) = \sqrt{Var\left(\hat{\beta}_j\right)} = \hat{\sigma}\sqrt{s_{jj}}\,, \tag{2.188}$$

gdje s_{jj} predstavlja j-ti element na glavnoj dijagonali matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. S obzirom da vrijedi $\hat{\beta}_i \mid \boldsymbol{X} \sim N\left(\beta_i, \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{jj}\right), \forall j$, standardizirana vrijednost svakog procjenitelja se može zapisati ovako:

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim N(0,1)\,. \tag{2.189}$$

No, kako se za standardne pogreške procjenitelja koristi procjena varijance u (2.186), slijedi:

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t(N-k-1)\,, \tag{2.190}$$

odnosno standardizirane vrijednosti procjenitelja slijede Studentovu distribuciju (t-distribuciju) s $N–k–1$ stupnjeva slobode. Za provođenje pojedinačnog testa o značajnosti neke varijable u modelu, u formuli (2.190) se pretpostavlja da je $\beta_j = 0$. Napomenimo da **algebarska svojstva** procjenitelja metodom najmanjih kvadrata vrijede i u slučaju višestruke linearne regresije.

Primjer 2.23.
Dani su (simulirani) podaci o zavisnoj i tri nezavisne varijable u tablici 2.12. Procijenimo model višestruke linearne regresije koristeći formulu (2.182).

Tablica 2.12. Opažanja nezavisne i zavisnih varijabli

y	18	29	21	11	7	25	44	1
x_1	10	15	12	7	4	14	22	1
x_2	13	12	16	8	10	11	25	5
x_3	9	3	7	14	18	2	1	22

Najprije zapišimo matricu $\boldsymbol{X}$ i vektor $\boldsymbol{y}$:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 10 & 13 & 9 \\ 1 & 15 & 12 & 3 \\ 1 & 12 & 16 & 7 \\ 1 & 7 & 8 & 14 \\ 1 & 4 & 10 & 18 \\ 1 & 14 & 11 & 2 \\ 1 & 22 & 25 & 1 \\ 1 & 1 & 5 & 22 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 18 \\ 29 \\ 21 \\ 11 \\ 7 \\ 25 \\ 44 \\ 1 \end{bmatrix}.$$

Sada je

$$\hat{\boldsymbol{\beta}} = \left(\begin{bmatrix} 1 & 10 & 13 & 9 \\ 1 & 15 & 12 & 3 \\ 1 & 12 & 16 & 7 \\ 1 & 7 & 8 & 14 \\ 1 & 4 & 10 & 18 \\ 1 & 14 & 11 & 2 \\ 1 & 22 & 25 & 1 \\ 1 & 1 & 5 & 22 \end{bmatrix}' \cdot \begin{bmatrix} 1 & 10 & 13 & 9 \\ 1 & 15 & 12 & 3 \\ 1 & 12 & 16 & 7 \\ 1 & 7 & 8 & 14 \\ 1 & 4 & 10 & 18 \\ 1 & 14 & 11 & 2 \\ 1 & 22 & 25 & 1 \\ 1 & 1 & 5 & 22 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 10 & 13 & 9 \\ 1 & 15 & 12 & 3 \\ 1 & 12 & 16 & 7 \\ 1 & 7 & 8 & 14 \\ 1 & 4 & 10 & 18 \\ 1 & 14 & 11 & 2 \\ 1 & 22 & 25 & 1 \\ 1 & 1 & 5 & 22 \end{bmatrix}' \cdot \begin{bmatrix} 18 \\ 29 \\ 21 \\ 11 \\ 7 \\ 25 \\ 44 \\ 1 \end{bmatrix} = \begin{bmatrix} -10,48 \\ 2,55 \\ -0,08 \\ 0,41 \end{bmatrix}$$

Dodatno, uneseni su podaci u RStudio za zadane vrijednosti svih varijabli, te su naredbama, predočenim na slici 2.61., izračunati potrebni međurezultati kako bi se procijenili parametri modela.

```
y<-c(18,29,21,11,7,25,44,1)
x1<-c(10,15,12,7,4,14,22,1)
x2<-c(13,12,16,8,10,11,25,5)
x3<-c(9,3,7,14,18,2,1,22)
y<-as.matrix(y)
jed<-c(rep(1,each=8))
jed<-as.matrix(jed)

x<-as.matrix(cbind(x1,x2,x3))
X<-cbind(jed,x)

a<-t(X)%*%X
b<-solve(a)
c<-t(X)%*%y
beta<-b%*%c
beta
```

Slika 2.61. Unos potrebnih naredbi kako bi se procijenili parametri regresijskog modela u primjeru

Ispis procijenjenih parametra je prikazan na slici 2.62., temeljem koje se može zapisati procijenjeni model: $\hat{y}_i = -10,48 + 2,55x_{1i} - 0,08x_{2i} + 0,41x_{3i}$.

```
##           [,1]
##    -10.4810160
## x1   2.5542314
## x2  -0.0848511
## x3   0.4108364
```

Slika 2.62. Ispis procijenjenih parametara modela u primjeru

Primjer 2.24.
Za simulirane podatke iz prethodnog primjera (2.23.), procijenimo model višestruke linearne regresije naredbom lm() u RStudiju, te ga potom zapišimo, zajedno sa standardnim pogreškama procjenitelja.

Temeljem naredbe i ispisa prikazanih na slici 2.63, možemo zapisati sljedeći procijenjeni model: $\hat{y}_i = -10,48 + 2,55x_{1i} - 0,08x_{2i} + 0,41x_{3i}$, dok su standardne pogreške procjenitelja sljedeće: $SE(\hat{\beta}_0) = 4,41, SE(\hat{\beta}_1) = 0,38, SE(\hat{\beta}_2) = 0,19$ i $SE(\hat{\beta}_3) = 0,23$. Stoga se mogu zapisati i standardizirane vrijednosti procjenitelja. Primjerice, za konstantu bismo izračunali na sljedeći

način: $t_0 = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} = \frac{-10,48}{4,42} = -2,374$. Čitatelju preostaje zadatak zapisati standardizirane vrijednosti procjenitelja za preostale nezavisne varijable u modelu.

```
summary(lm(y~x1+x2+x3))

## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##         1         2         3         4         5         6         7         8
##  0.344238  0.953249 -0.687998 -1.471505  0.717546 -0.166534 -0.001634  0.312639
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.48102    4.41491  -2.374  0.07649 .
## x1            2.55423    0.37535   6.805  0.00244 **
## x2           -0.08485    0.18530  -0.458  0.67078
## x3            0.41084    0.22836   1.799  0.14640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 4 degrees of freedom
## Multiple R-squared:  0.9967, Adjusted R-squared:  0.9942
## F-statistic: 399.9 on 3 and 4 DF,  p-value: 2.068e-05
```

Slika 2.63. Procjena modela višestruke linearne regresije

2.2.4. Interpretacija parametara u modelu višestruke linearne regresije

2.2.4.1. Lin-lin model

Lin-lin model je onaj u kojemu su sve varijable (zavisna i nezavisne) u razinama. Nad varijablama nije izvršena nikakva transformacija, stoga se interpretacija vrši u mjernim jedinicama svake varijable. Ako se razmotri procijenjeni model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}, \qquad (2.191)$$

interpretacija $\hat{\beta}_1$ odnosi se na granični efekt utjecaja jedinične promjene nezavisne varijable na promjenu zavisne varijable (**parcijalna derivacija funkcije** po toj varijabli, $\frac{\partial \hat{y}_i}{\partial x_{i1}} = \hat{\beta}_1$), uz uvjet da druga varijabla ostane nepromijenjena. Dakle, interpretacija $\hat{\beta}_1$ je sljedeća: ako se varijabla x_{i1} poveća za jednu jedinicu, **uz nepromijenjenu vrijednost varijable x_{i2}**, tada se u prosjeku varijabla y_i promjeni za $\hat{\beta}_1$ jedinica (poveća ili smanji, ovisi o predznaku $\hat{\beta}_1$).

Napomena. Ako su u modelu više od dvije varijabli, tada se u interpretaciji svih modela navodi „uz nepromijenjene ostale varijable“.

Primjer 2.25.
Za procijenjeni model $\hat{y}_i = 2 + 150x_{i1} + 20x_{i2}$, varijabla x_1 odnosi se na broj godina radnog staža, x_2 na broj godina školovanja, a varijabla y na dohodak zaposlenika u kn. Tumačenje koeficijenta uz varijablu x_{i1} je sljedeće: ako se broj godina radnog staža zaposlenika poveća za 1 godinu, uz nepromijenjen broj godina školovanja, tada se dohodak zaposlenika poveća u prosjeku za 150 kn.

2.2.4.2. Lin-log model

Lin-log model je onaj kod kojeg je zavisna varijabla u razinama, dok su nezavisne logaritmirane. Ako se razmotri procijenjeni model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln x_{i1} + \hat{\beta}_2 \ln x_{i2}, \tag{2.192}$$

parcijalni diferencijal jednadžbe (2.192) s obzirom na varijablu x_{i1} je:

$$\partial \hat{y}_i = \hat{\beta}_1 \frac{1}{x_{i1}} \partial x_{i1} \,/\cdot \frac{x_{i1}}{\partial x_{i1}} \Rightarrow \frac{\partial \hat{y}_i}{\frac{\partial x_{i1}}{x_{i1}}} = \hat{\beta}_1, \tag{2.193}$$

Pritom promjenu nezavisne varijable mjerimo u postotcima, dok zavisne u mjernim jedinicama, i pritom se interpretacija vrši kao $\frac{\hat{\beta}_1}{100}$ jedinica. Ako se varijabla x_{i1} poveća za 1%, uz nepromijenjenu vrijednost varijable x_{i2}, tada se u prosjeku zavisna varijabla poveća/smanji za $\frac{\hat{\beta}_1}{100}$ jedinica.

Dijelimo sa 100 kako bismo dobili $\frac{\partial \hat{y}_i}{100\frac{\partial x_{i1}}{x_{i1}}} = \frac{\hat{\beta}_1}{100}$. U brojniku desne strane jednakosti nalazi se promjena zavisne varijable u mjernim jedinicama, dok je u nazivniku stopa rasta nezavisne varijable. Zato desnu stranu jednakosti interpretiramo kao promjenu zavisne varijable za $\frac{\hat{\beta}_1}{100}$ jedinica ako se nezavisna varijabla x_{i1} poveća za 1% uz nepromijenjenu vrijednost varijable x_{i2}.

Primjer 2.26.
Razmatra se procijenjeni model $\hat{y}_i = 10 + 1200 \ln x_{i1} + 1000 \ln x_{i2}$, varijabla x_1 odnosi se na broj godina radnog staža, x_2 na broj godina školovanja, a varijabla y na dohodak zaposlenika u kn. Tumačenje koeficijenta uz varijablu x_{i1}: ako se broj godina radnog staža zaposlenika poveća za 1%, uz nepromijenjen broj godina školovanja, dohodak zaposlenika se poveća u prosjeku za 12 kn.

2.2.4.3. Log-lin model

Log-lin model je onaj u kojemu je zavisna varijabla logaritmirana, a nezavisne su u razinama.

Ako se razmotri procijenjeni model

$$\widehat{\ln y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_1 x_{i2}\,, \tag{2.194}$$

parcijalni diferencijal jednadžbe (2.194) po varijabli x_{i1} je

$$\partial(\widehat{\ln y}_i) = \hat{\beta}_1 \partial x_{1i} \Rightarrow \frac{\frac{\partial \hat{y}_i}{\hat{y}_i}}{\partial x_{i1}} = \hat{\beta}_1\,. \tag{2.195}$$

Posljednja jednakost se može interpretirati kao koeficijent elastičnosti, pri čemu promjenu nezavisne varijable mjerimo u njenim mjernim jedinicama, dok se promjena zavisne varijable mjeri u postotcima, pri čemu je $\hat{\beta}_1 \cdot 100\%$. Ako se x_{i1} poveća za jednu mjernu jedinicu, uz nepromijenjenu vrijednost varijable x_{i2}, tada se zavisna varijabla poveća/smanji u prosjeku za $\hat{\beta}_1 \cdot 100\%$. U ovome slučaju parametar $\hat{\beta}_i$ množi se sa 100%, kako bismo dobili izraz $\frac{\frac{\partial \hat{y}_i}{\hat{y}_i} 100\%}{\partial x_{i1}} = \hat{\beta}_1 100\%$. U brojniku lijeve strane jednakosti nalazi se stopa rasta, dok je u nazivniku promjena izražena u mjernim jedinicama varijable x_{i1}. Zato desnu stranu jednakosti interpretiramo kao postotnu promjenu zavisne varijable ako se nezavisna varijabla x_{i1} poveća za jednu jedinicu, uz nepromijenjenu vrijednost druge nezavisne varijable.

Primjer 2.27.
Procijenjen je model $\widehat{\ln y}_i = 0{,}03 + 0{,}02x_{i1} + 0{,}05x_{i2}$, varijabla x_1 odnosi se na broj godina radnog staža, x_2 na broj godina školovanja, a varijabla y na dohodak zaposlenika u kn. Tumačenje koeficijenta uz varijablu x_{i1}: ako se broj godina radnog staža zaposlenika poveća za 1 godinu, uz nepromijenjen broj godina školovanja, tada se dohodak zaposlenika poveća u prosjeku za 2%.

2.2.4.4. Log-log model

Log-log model je onaj u kojemu su argumenti logaritmirani. Ako se razmotri model

$$\widehat{\ln y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}, \tag{2.196}$$

diferencijal funkcije (2.196) po varijabli ($\ln x_{1i}$) je

$$\partial(\widehat{\ln y}_i) = \hat{\beta}_1 \partial(\ln x_{1i}) \Rightarrow \frac{\partial \hat{y}_i}{\hat{y}_i} = \frac{\partial x_{1i}}{x_{1i}} \hat{\beta}_1 \Rightarrow \frac{\partial \hat{y}_i}{\partial x_{1i}} \frac{x_{1i}}{\hat{y}_i} \hat{\beta}_1\,. \tag{2.197}$$

U posljednjoj jednakosti prepoznajemo koeficijent elastičnosti $\left(E_{y,x_j} = \frac{\partial y}{\partial x_j} \frac{x_j}{y} \right)$, pa je u ovome slučaju interpretacija parametra $\hat{\beta}_1$: Ako se varijabla x_{i1} poveća za 1%, tada se u prosjeku zavisna varijabla poveća/smanji za $\hat{\beta}_1$ %, uz nepromijenjenu vrijednost druge varijable.

Primjer 2.28.
Procijenjen je model $\widehat{\ln y_i} = 0{,}2 + 0{,}8 \ln x_{i1} + 1{,}4 \ln x_{i2}$, x_1 odnosi se na broj godina radnog staža, x_2 na broj godina školovanja, a varijabla y na dohodak zaposlenika u kn. Tumačenje koeficijenta uz varijablu x_{i1}: ako se broj godina školovanja zaposlenika poveća za 1%, uz nepromijenjen broj godina školovanja, tada se dohodak zaposlenika poveća u prosjeku za 0,8%.

2.2.4.5. Interpretacija parametara u modelu višestruke linearne regresije sa standardiziranim varijablama

U slučaju regresijskog modela s više nezavisnih varijabli, postavlja se pitanje koja nezavisna varijabla ima veći učinak na zavisnu varijablu. Kako su sve varijable iskazane u **vlastitim mjernim jedinicama**, **nije pogodno uspoređivati** vrijednosti procijenjenih parametara u linearnom regresijskom modelu. U tom slučaju standardiziraju se varijable, te se potom procijeni model sa standardiziranim vrijednostima svih varijabli. U slučaju standardizacije, očekivana vrijednost svake varijable je 0, stoga se konstanta ne uključuje u model. Parametri se interpretiraju u standardnim devijacijama.

Dakle, najprije se standardiziraju zavisna i nezavisne varijable:

$$y_i^* = \frac{y_i - \overline{y}}{\sigma_y},\ x_j^* = \frac{x_j - \overline{x}}{\sigma_{x_j}},\ j \in \{1,2,...,k\}, \tag{2.198}$$

te se potom procijeni model:

$$\hat{y}_i^* = \hat{\beta}_1^* x_{i1}^* + \hat{\beta}_2^* x_{i2}^* + ... + \hat{\beta}_k^* x_{ik}^*. \tag{2.199}$$

U slučaju jednadžbe (2.199) se vrijednost parametra $\hat{\beta}_1^*$ interpretira na sljedeći način: ako se vrijednost nezavisne varijable poveća za 1 **standardnu devijaciju**, **uz nepromijenjene vrijednosti ostalih nezavisnih varijabli**, vrijednost zavisne varijable će se promijeniti u prosjeku za $\hat{\beta}_1^*$ **standardnih devijacija**.
Valja napomenuti da prilikom uspoređivanja parametara u modelu (2.199), za jačinu učinka se uspoređuju apsolutne vrijednosti tih parametara.

Primjer 2.29.
Temeljem podataka u tablici 2.12. iz primjera 2.23, procijenimo standardizirani model u RStudiju i interpretirajmo rezultat.

Temeljem naredbi i ispisa prikazanih na slici 2.64., može se zapisati sljedeći procijenjeni model: $\hat{y}_i^* = 1{,}25x_{i1}^* - 0{,}04x_{i2}^* + 0{,}24x_{i3}^*$. Dakle, uočava se da najjači učinak na zavisnu varijablu ima prva nezavisna varijabla, te potom treća, a najslabiji učinak ima druga po redu varijabla.

```
lm(scale(y)~0+scale(x1)+scale(x2)+scale(x3))

##
## Call:
## lm(formula = scale(y) ~ 0 + scale(x1) + scale(x2) + scale(x3))
##
## Coefficients:
## scale(x1)  scale(x2)  scale(x3)
##   1.25299   -0.03756    0.23554
```

Slika 2.64. Rezultat procjene standardiziranog modela

Interpretacija jednog od procijenjenih parametra je sljedeća. Ako se druga nezavisna varijabla poveća za jednu standardnu devijaciju, uz nepromijenjene ostale varijable u modelu, vrijednost zavisne varijable će se smanjiti za približno 0,04 standardnih devijacija.

2.2.4.6. Napomena o konstanti u modelu višestruke linearne regresije

Kao što je već spomenuto za model jednostavne linearne regresije, konstantu je bolje ostaviti u samome modelu prilikom procjene parametara (vidjeti 2.1.6.5). Konstanta se u slučaju višestruke linearne regresije interpretira kao prosječna razina zavisne varijable, kada bi **sve nezavisne varijable iznosile 0 jedinica**. Naravno, treba paziti prilikom same interpretacije radi li se o lin-lin, lin-log, log-lin ili log-log modelu.

Primjer 2.30.
Procijenjen je model u primjeru 2.24.: $\hat{y}_i = -10{,}48 + 2{,}55x_{i1} - 0{,}08x_{i2} + 0{,}41x_{i3}^*$. Interpretirajmo konstantu u modelu.

Kada bi vrijednost svih nezavisnih varijabli u modelu iznosila 0, u prosjeku bi vrijednost zavisne varijable iznosila −10,48 jedinica. Čitatelju se ostavlja za vježbu interpretacija konstante u primjerima 2.25., 2.26., 2.27. i 2.28.

2.2.5. Intervalna procjena parametara višestruke linearne regresije

Temeljem pretpostavki regresijskog modela standardizirane vrijednosti (oznaka t_j) procijenjenih parametara $\hat{\beta}_j$ slijede Studentovu distribuciju s $N-k-1$ stupnjeva slobode:

$$t_j = \frac{\hat{\beta}_j - \beta_j}{SE\left(\hat{\beta}_j\right)} \sim t(N-k-1) \ , \qquad (2.200)$$

čijom jednostavnom manipulacijom se konstruiraju intervalne procjene:

$$P\left(-t_{\gamma/2} < t_j < t_{\gamma/2}\right) = 1-\gamma \ , \qquad (2.201)$$

gdje je $1-\gamma$ **pouzdanost procjene**, dok $t_{\gamma/2}$ predstavlja koeficijent pouzdanosti, tj. vrijednost Studentove distribucije s $N-k-1$ stupnjeva slobode. Kada se supstitucijom (2.200) u (2.201) zapiše intervalna procjena:

$$P\left(\hat{\beta}_j - t_{\gamma/2}SE\left(\hat{\beta}_j\right) < \beta_j < \hat{\beta}_j + t_{\gamma/2}SE\left(\hat{\beta}_j\right)\right) = 1-\gamma \ , \qquad (2.202)$$

dobiva se intervalna procjena za β_j. Kao i već spomenuto prije, uobičajeno se za vrijednost $(1-\gamma)$ uzima 90%, 95% ili 99%.

Interpretacija relacije (2.202) je sljedeća. Za slučaj konstante, kada bi vrijednost svih nezavisnih varijabli u modelu iznosila 0, tada bi u prosjeku vrijednost zavisne varijable iznosila između $\hat{\beta}_0 - t_{\gamma/2}SE\left(\hat{\beta}_0\right)$ i $\hat{\beta}_0 + t_{\gamma/2}SE\left(\hat{\beta}_0\right)$ mjernih jedinica uz pouzdanost procjene $1-\gamma$. Za

j-tu nezavisnu varijablu, interpretacija je: uz pouzdanost procjene $1-\gamma$, kada bi se vrijednost nezavisne varijable j povećala za jednu jedinicu, uz nepromijenjene vrijednosti ostalih nezavisnih varijabli, vrijednost zavisne varijable će se promijeniti u prosjeku između $\hat{\beta}_j - t_{\gamma/2} SE\left(\hat{\beta}_j\right)$ i $\hat{\beta}_j + t_{\gamma/2} SE\left(\hat{\beta}_j\right)$ jedinica. Ako su predznaci obje granice negativni, govorimo o smanjenju, dok pozitivni predznaci govore o povećanju zavisne varijable.

Primjer 2.31.
Temeljem podataka iz primjera 2.23 (tablica 2.12.), odredimo intervalne procjene svih parametara na razini pouzdanosti 92% i interpretirajmo ih.

Temeljem ispisa na slici 2.65., možemo pisati sljedeće intervalne procjene te ih interpretirati: Uz razinu pouzdanosti 92%, kada bi vrijednosti sve tri nezavisne varijable iznosile 0, u prosjeku bi vrijednost varijable y iznosila između –20,78 i –0,18 jedinica. Nadalje, uz razinu pouzdanosti 92%, ako se vrijednost varijable x_1 poveća za jednu jedinicu, uz ostale dvije varijable nepromijenjene, vrijednost varijable y bi se povećala u prosjeku između 1,69 i 3,43 jedinice.

$$P(-20{,}78 < \beta_0 < -0{,}18) = 0{,}92$$

$$P(1{,}69 < \beta_1 < 3{,}43) = 0{,}92$$

$$P(-0{,}52 < \beta_2 < -0{,}35) = 0{,}92$$

$$P(-0{,}12 < \beta_3 < -0{,}18) = 0{,}92$$

Čitatelju preostaje za vježbu interpretacija intervala za drugu i treću nezavisnu varijablu, pri čemu se uočava da se razlikuju predznaci donje i gornje granice intervala (stoga je interpretacija slična kao u primjeru 2.13.).

```
model<-lm(y~x1+x2+x3)
confint(model,level=0.92)

##                      4 %       96 %
## (Intercept) -20.7804435 -0.1815886
## x1            1.6785836  3.4298792
## x2           -0.5171283  0.3474261
## x3           -0.1219095  0.9435824
```

Slika 2.65. Intervalne procjene parametara, razina pouzdanosti 92%

2.2.6. Analiza varijance u modelu višestruke linearne regresije

Tablica ANOVA za slučaj modela višestruke linearne regresije je veoma slična tablici ANOVA u slučaju jednostruke linearne regresije.

Tablica 2.13. Tablica analize varijance, model višestruke linearne regresije

Izvor varijacije	Sume kvadrata	Stupnjevi slobode (*ss*)	Sredina kvadrata odstupanja	*F*-omjer	*p-v*
Regresija (protumačeno modelom)	*SSE*	*k*	*SSE*/*k*	$\frac{SSE/k}{SSR/(N-k-1)}$	…
Rezidualna odstupanja (neprotumačeno modelom)	*SSR*	*N*–*k*–1	*SSR*/(*N*–*k*–1)		
Ukupno	*SST*	*N*–1			

Ako se razmotri tablica 2.13. koja je za slučaj višestruke linearne regresije, uočava se da se sada stupac stupnjevi slobode sastoji od *k* stupnja za slučaj izvora varijacije koji je protumačen modelom jer se odnosi na *k* nezavisnih varijabli, dok se broj stupnjeva slobode za varijaciju koja nije protumačena modelom odnosi na *N*–*k*–1 (umanjeno s obzirom na broj nezavisnih varijabli *k* i konstantu) stupanj.

Jednadžba analize varijance i ovdje se zapisuje na sljedeći način (usporediti detaljan opis u 2.1.8):

$$\sum_{i=1}^{N}\left(y_i-\overline{y}\right)^2=\sum_{i=1}^{N}\left(\hat{y}_i-\overline{y}\right)^2+\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2, \qquad (2.203)$$

gdje se ukupna suma kvadrata (*SST*) dijeli na zbroj sume kvadrata odstupanja regresijskih vrijednosti od prosjeka (*SSE*) i sume kvadrata rezidualnih odstupanja (*SSR*):

$$SST = SSE + SSR. \qquad (2.204)$$

Sada kad se zbrojevi kvadrata podjele s odgovarajućim stupnjevima slobode, dobivaju se sredine kvadrata koje su nezavisne procjene udjela u ukupnoj varijanci, *SSE*/*k* i *SSR*/(*N*–*k*–1). **Procijenjena varijanca regresije** je dana izrazom *SSR*/(*N*–*k*–1):

$$\frac{SSR}{N-k-1}=\frac{\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2}{N-k-1}=\hat{\sigma}^2, \qquad (2.205)$$

dok je njezin pozitivni drugi korijen **procjena standardne devijacije regresije:**

$$\hat{\sigma}=\sqrt{\hat{\sigma}^2}, \qquad (2.206)$$

te se tumači kao i kod modela jednostruke regresije (prosječno odstupanje stvanih ili empirijskih vrijednosti zavisne varijable od njenih procijenjenih vrijednosti, izraženo apsolutno ili u mjernim jedinicama zavisne varijable). **Relativna mjera disperzije**, **procjena koeficijenta varijacije**:

$$\hat{V} = \frac{\hat{\sigma}}{\overline{y}} 100\%, \tag{2.207}$$

Interpretira se kao prosječno odstupanje stvarnih vrijednosti zavisne varijable od regresijskih, izraženo postotkom. Obje mjere disperzije trebaju biti što manje da bi model bio klasificiran kao što uspješniji. Udio protumačenih odstupanja u ukupnoj sumi kvadrata odstupanja, i naziva se **koeficijent determinacije**, oznaka R^2 (engl. *coefficient of determination, R-squared*):

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\hat{\sigma}^2 (N-k-1)}{\sum_{i=1}^{N} (y_i - \overline{y})^2}, \tag{2.208}$$

te je poželjno da je što bliži vrijednosti 1. No, **problem je što je pristran pokazatelj** jer ovisi o broju varijabli k koje uvrštavamo u model, unatoč tome što možda nema ekonomskog smisla njih uključivati u model. Zato se u tu svrhu najprije definira **nepristrana procjena varijance zavisne varijable:**

$$\hat{\sigma}_y^2 = \frac{\sum_{i=1}^{N} (y_i - \overline{y})^2}{N-1}, \tag{2.209}$$

koja je nepromjenjiva bez obzira na odabir nezavisnih varijabli, kako bi se definirao **korigirani koeficijent determinacije** $\overline{R}^2$:

$$\overline{R}^2 = 1 - \frac{N-1}{(N-k-1)}(1-R^2) = 1 - \frac{N-1}{(N-k-1)} \frac{SSR}{SST} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}. \tag{2.210}$$

Upravo se $\overline{R}^2$ koristi u slučaju uspoređivanja više modela koji imaju različit broj nezavisnih varijabli, kako bi modeli bili usporedivi, zbog problema s R^2. Što veća mjera $\overline{R}^2$ označava bolju protumačenost varijacija zavisne varijable odabranim modelom[33].

Nadalje, **koeficijent višestruke linearne korelacije** je standardizirana mjera jakosti linearne veze između zavisne varijable i skupa k nezavisnih varijabli, pri čemu se računa kao drugi pozitivni korijen iz koeficijenta determinacije:

$$R = \sqrt{R^2}. \tag{2.211}$$

Valja napomenuti da se ovdje **ne interpretira smjer** povezanosti jer povezanost između zavisne i nezavisnih varijabli može biti različitog predznaka! Vidjeti primjer 27, gdje se radi o različitim predznacima za tri nezavisne varijable u modelu. Za slučaj modela višestruke linearne regresije, koeficijent R poprima vrijednosti između [0, 1], za razliku od koeficijenta za jednostruku linearnu regresiju koji je poprimao vrijednosti između [−1, 1].

Konačno, F-omjer u tablici ANOVA odnosi se na F-test, odnosno test skupne značajnosti svih varijabli u modelu, koji će se detaljno pojasniti u 2.2.7.2.

[33] Međutim, vidjeti napomenu u fusnoti u poglavlju 2.1.8.!

Primjer 2.32.
Za podatke dane u tablici 2.12. u primjeru 2.23, sastavimo tablicu ANOVA, izračunajmo koeficijent determinacije, korigirani koeficijent determinacije, procjenu standardne devijacije regresije, procjenu koeficijenta varijacije regresije te koeficijent višestruke linearne korelacije i potom interpretirajmo rezultat.

Iz slike 2.66. uočava se da je koeficijent determinacije jednak 0,9967, dok korigirani koeficijent determinacije iznosi 0,9942. Stoga je ovim modelom objašnjen velik dio varijacija varijable *y*. Dakle, radi se o izvrsnom modelu, jer je njime objašnjeno gotovo cijela varijacija zavisne varijable. Nadalje, koeficijent višestruke linearne korelacije iznosi $\sqrt{0,9967}$= 0,9983, što znači da postoji jaka linearna povezanost između zavisne i nezavisnih varijabli u modelu.

Nadalje, na slici 2.67. dan je ispis procijenjene standardne devijacije regresije, kao i relativna mjera disperzije, procijenjeni koeficijent varijacije regresije. Prosječno odstupanje empirijskih od regresijskih vrijednosti zavisne varijable iznosi 1,04 jedinica, odnosno 5,32% relativno, što dodatno potvrđuje da je model jako dobar.

```
summary(lm(y~x1+x2+x3))

##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##         1         2         3         4         5         6         7         8
##  0.344238  0.953249 -0.687998 -1.471505  0.717546 -0.166534 -0.001634  0.312639
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.48102    4.41491  -2.374  0.07649 .
## x1            2.55423    0.37535   6.805  0.00244 **
## x2           -0.08485    0.18530  -0.458  0.67078
## x3            0.41084    0.22836   1.799  0.14640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 4 degrees of freedom
## Multiple R-squared:  0.9967, Adjusted R-squared:  0.9942
## F-statistic: 399.9 on 3 and 4 DF,  p-value: 2.068e-05
```

Slika 2.66. Ispis procijenjenog linearnog regresijskog modela

```
model<-lm(y~x1+x2+x3)
sazetak<-summary(model)

sazetak$sigma

## [1] 1.03757

sazetak$sigma/mean(y)

## [1] 0.0532087
```

Slika 2.67. Naredbe za procjenu greške regresije i koeficijenta varijacije regresije

Nadalje, za potrebe popunjavanja tablice ANOVA, potrebno je postojeći model usporediti s modelom u kojemu se zavisna varijabla procjenjuje samo na konstantu, kako bi se usporedila kvaliteta cijelog modela (svih nezavisnih varijabli), u odnosu na model bez svih varijabli

(vidjeti naslov 2.2.7.2 o *F*-testu za detalje). Stoga se uz prvotni model procijenio i drugi, naredbom model2<–lm(y~1), te je sada u naredbu anova() potrebno upisati nazive dva modela koji se uspoređuju (slika 2.68.). Ispis je nešto drugačiji u odnosu na tablicu ANOVA za model jednostruke linearne regresije, stoga je izračun sljedeći. Sada se u ispisu Model 1 odnosi na model sa sve tri nezavisne varijable, čija je suma kvadrata rezidualnih odstupanja jednaka 4,31 (RSS stupac), dok je u modelu samo s konstantom (Model 2) ta suma puno veća (1296).

```
model2<-lm(y~1)
anova(model,model2)

## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ 1
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1      4    4.31
## 2      7 1296.00 -3   -1291.7 399.95 2.068e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 2.68. Tablica ANOVA za razmatrani primjer

Upravo je suma kvadrata rezidualnih odstupanja u Modelu 2 vrijednost *SST* u tablici ANOVA, jer se bez zavisnih varijabli u modelu računa varijanca zavisne varijable. Sada se može popunjavati tablica ANOVA prikazana u tablici 2.14. na način da smo prvo popunili vrijednost 4,31 i 1296. Oduzimanjem te dvije vrijednosti dobiva se suma kvadrata 1291,69. Broj stupnjeva slobode za Model 1 je $k = 3$, dok je za slučaj neprotumačen modelom upravo za Model 2, tj. 7. Dalje se računaju sredine kvadrata odstupanja i empirijski *F*-omjer koji iznosi 399,95. Spomenuto je u naslovu 2.1.8 kako je ideja da je taj omjer što veći, jer bi upravo trebao biti što veći udio varijacije zavisne varijable protumačen modelom, odnosno nezavisnim varijablama.

Tablica 2.14. Tablica analize varijance u primjeru 2.32

Izvor varijacije	Sume kvadrata	Stupnjevi slobode (*ss*)	Sredina kvadrata odstupanja	*F*-omjer
Regresija (protumačeno modelom)	1291,69	3	430,9461	399,95
Rezidualna odstupanja (neprotumačeno modelom)	4,31	4	1,0775	
Ukupno	1296	7		

2.2.7. Testiranje hipoteza u modelu višestruke linearne regresije

2.2.7.1. *t*-test

Kao i u modelu jednostavne linearne regresije, i u slučaju višestruke linearne regresije može se provesti test o značajnosti pojedine nezavisne varijable u modelu. U tom slučaju govorimo o pojedinačnom *t*-testu, koji može ponovno biti dvosmjerni ili jednosmjerni test. Tablica 2.15. sažima izračun empirijskog *t*-omjera za *j*-tu varijablu u modelu (vidjeti formulu (2.200)), zajedno s hipotezama u slučaju dvosmjernog i jednosmjernog testa kao i odluku koja se donosi usporedbom empirijskog *t*-omjera s teorijskim.

Tablica 2.15. Provedba pojedinačnog *t*-testa

Izračun, zapis	Dvosmjerni test	Jednosmjerni, gornja granica	Jednosmjerni, donja granica
Empirijski *t*-omjer	$t_j = \dfrac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)}$		
Teorijski *t*-omjer	$t_{\alpha/2}(N-k-1)$	$t_{\alpha}(N-k-1)$	$-t_{\alpha}(N-k-1)$
Hipoteze testa	$H_0 : \beta_j = 0$ $H_1 : \beta_j \neq 0$	$H_0 : \beta_j = 0$ $H_1 : \beta_j > 0$	$H_0 : \beta_j = 0$ $H_1 : \beta_j < 0$
Odluka, ishod	$\lvert t_j \rvert > t_{\alpha/2} \Rightarrow$ odbacujem H_0 $\lvert t_j \rvert < t_{\alpha/2} \Rightarrow$ ne odbacujem H_0 ili p-$v/2 < \alpha \Rightarrow$ odbacujem H_0 p-$v/2 > \alpha \Rightarrow$ ne odbacujem H_0	$t_j > t_{\alpha} \Rightarrow$ odbacujem H_0 $t_j < t_{\alpha} \Rightarrow$ ne odbacujem H_0 ili p-$v < \alpha \Rightarrow$ odbacujem H_0 p-$v > \alpha \Rightarrow$ ne odbacujem H_0	$t_j > -t_{\alpha} \Rightarrow$ odbacujem H_0 $t_j < -t_{\alpha} \Rightarrow$ ne odbacujem H_0 ili p-$v < \alpha \Rightarrow$ odbacujem H_0 p-$v > \alpha \Rightarrow$ ne odbacujem H_0

Primjer 2.33.
Temeljem slike 2.64. iz prethodnog primjera (2.32.), provedimo dvosmjerne i jednosmjerne testove za sve tri nezavisne varijable u modelu, uz razinu značajnosti od 5%.

S obzirom da je potrebno provesti 6 testova, sažeto su prikazani izračuni u tablici 2.16. Primjerice, za dvosmjerni test o značajnosti prve nezavisne varijable empirijski *t*-omjer izračunat je na sljedeći način:

$t_1 = \dfrac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \dfrac{2{,}554}{0{,}375} = 6{,}805$, dok je pripadajući teorijski *t*-omjer (izračunat naredbom abs(qt(0.05/2,4))), te iznosi $t_{\alpha/2}(N-k-1) = t_{0{,}05/2}(8-3-1) = t_{0{,}025}(4) = 2{,}776$.

Hipoteze testa su sljedeće: $\begin{matrix} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{matrix}$.

Kako vrijedi $\lvert t_1 \rvert > t_{\alpha/2}$, odnosno |6,805| > 2,776, to je u području odbacivanja nulte hipoteze. Osim toga, kako je *p*-vrijednost u ispisu 0,002, što je manje od 0,05 (tj. *p*-vrijednost < α), također se donosi odluka o odbacivanju nulte hipoteze. Riječima: uz razinu značajnosti od 5%, odbacuje hipoteza da varijabla x_1 nije značajna u modelu.

S druge strane, ako se provodi odgovarajući jednosmjerni *t*-test, kako je predznak procijenjenog parametra pozitivan, radi se o testu na gornju granicu. U tom slučaju je teorijski *t*-omjer (izračunat naredbom abs(qt(0.05,4))) sljedeći: $t_{0{,}05}(4) = 2{,}132$.

Hipoteze testa su sljedeće: $\begin{matrix} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 > 0 \end{matrix}$.

U ovome slučaju je empirijski *t*-omjer veći od teorijskog, tj. vrijedi 6,805 > 2,132, odnosno pripadajuća *p*-vrijednost (0,002/2) koja iznosi 0,001 je ponovno manja od zadane razine značajnosti 0,05 pa se i u slučaju jednosmjernog testa donosi zaključak da uz razinu značajnosti od 5%, odbacuje hipoteza da varijabla x_1 nije značajna u modelu.

Tablica 2.16. Sažetak rezultata provedenih *t*-testova

Varijabla		Test:		
		Dvosmjerni	Jednosmjerni, gornja granica	Jednosmjerni, donja granica
Empirijski *t*-omjer:	x_1	6,805		
	x_2	–0,458		
	x_3	1,799		
Teorijski *t*-omjer		2,776	2,132	–2,132

Nadalje, za drugu nezavisnu varijablu, svi izračuni su sljedeći.

Dvosmjerni test: $H_0 : \beta_2 = 0$, $H_1 : \beta_2 \neq 0$, $t_2 = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{-0,085}{0,185} = -0,458$, *p*-vrijednost = 0,671.

Kako vrijedi $|t_2| < t_{\alpha/2}$, odnosno |–0,458| < 2,776, to je u području ne odbacivanja nulte hipoteze. Osim toga, kako je *p*-vrijednost u ispisu 0,671, što je veće od 0,05, također se donosi odluka o ne odbacivanju nulte hipoteze. Riječima: uz razinu značajnosti od 5%, ne odbacuje hipoteza da varijabla x_2 nije značajna u modelu.

Jednosmjerni test se provodi kao onaj na donju granicu, jer je predznak procijenjenog parametra negativan:

$H_0 : \beta_2 = 0$, $H_1 : \beta_2 < 0$, $t_2 = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{-0,085}{0,185} = -0,458$, *p*-vrijednost = 0,671/2 = 0,3355.

Kako vrijedi $t_2 > -t_{\alpha/2}$, odnosno |–0,458| > –2,776, to je u području ne odbacivanja nulte hipoteze. Osim toga, kako je *p*-vrijednost u ispisu 0,3355, što je veće od 0,05, također se donosi odluka o ne odbacivanju nulte hipoteze. Riječima: uz razinu značajnosti od 5%, ne odbacuje hipoteza da varijabla x_2 nije značajna u modelu. Čitatelju ostaje zadatak provesti oba testa za slučaj treće nezavisne varijable u modelu.

2.2.7.2. *F*-test

Test o značajnosti regresije ili *F*-test je onaj kojim se testira značajnost **svih** nezavisnih varijabli u modelu. Postoji razlika između ovoga i parcijalnog *F*-testa (vidjeti 2.2.7.4). Skupni *F*-test značajnosti nezavisnih varijabli u modelu pretpostavlja u **nultoj** hipotezi da **niti jedna** nezavisna varijabla **nije značajna** u modelu. Suprotno od toga, **alternativna** hipoteza pretpostavlja da postoji **barem jedna** varijabla koja **je značajna**. Simbolički:

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \\ &H_1 : \exists \beta_j \neq 0, j \in \{1, 2, ..., k\} \end{aligned} . \qquad (2.212)$$

Kako se empirijski *F*-omjer računa formulom:

$$F = \frac{SSE / k}{SSR / (N-k-1)} \sim F(k; N-k-1), \tag{2.213}$$

ako vrijede pretpostavke regresijskog modela (naslov 2.2.2), tada je *SSE*/*k* slučajna varijabla s *k* stupnja slobode, što pišemo: $SSE/k \sim \chi^2(k)$, dok je SSR/σ^2 slučajna varijabla koja slijedi hi-kvadrat distribuciju s *N*–*k*–1 stupnja slobode, što pišemo: $\frac{SSR}{\sigma^2} \sim \chi^2(N-k-1)$. Tada je omjer u (2.213) slučajna varijabla koja slijedi *F*-distribuciju s *k* stupnja slobode u brojniku i *N*–*k*–1 stupnja slobode u nazivniku[34].

Omjer (2.213) se može pisati i na sljedeći način:

$$F = \frac{SSE / k}{SSR / (N-k-1)} = \frac{R^2 / k}{(1-R^2) / (N-k-1)}. \tag{2.214}$$

F-test se provodi na sljedeći način. Formiraju se nulta i alternativna hipoteza (vidjeti (2.212)), te se potom uz zadanu razinu značajnosti α izračuna teorijski *F*-omjer (F_{emp}), $F_\alpha(k; N-k-1)$, koji se uspoređuje s empirijskim *F*-omjerom. Ako vrijedi $F_{emp} > F_\alpha(k; N-k-1)$, odbacuje se nulta hipoteza, dok se u slučaju $F_{emp} < F_\alpha(k; N-k-1)$ ona ne odbacuje. Također se mogu uspoređivati *p*-vrijednost te zadana razina značajnosti *α*: *p*-*v* < *α* → odbacujem H_0, odnosno *p*-*v* > *α* → ne odbacujem H_0.

Primjer 2.34.
Procijenjen je model višestruke linearne regresije u primjeru 2.32., slika 2.66. Provedimo skupni *F*-test o značajnosti varijabli u modelu. Odabrana razina značajnosti je α=5%. Mijenja li se zaključak za α=1%, odnosno α=10%?

Zapišimo najprije hipoteze testa: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
$H_1 : \exists \beta_j \neq 0, j \in \{1,2,3\}$

Koristeći ispis dan na slici 2.68., uočava se da je vrijednost empirijskog *F*-omjera jednaka 399,9, s 3 stupnja slobode u brojniku i 4 u nazivniku. Vrijednost je, dakle, izračunata kao:

$$F = \frac{R^2 / k}{(1-R^2) / (N-k-1)} = \frac{0,9967 / 3}{(1-0,9967) / (8-3-1)} = 399,9$$, *p*-vrijednost = $2,068 \cdot 10^{-5}$.

Teorijski *F*-omjer dan je temeljem tablice kritičnih vrijednosti *F*-distribucije uz α = 5%, te 3 stupa slobode u brojniku i 4 u nazivniku (naredba qf(1-0.05,3,4)) i iznosi $F_{0,05}(3;4) = 6,591$.

Stoga vrijedi: $F_{emp} = 399,9 > 6,591 = F_{0,05}(3;4)$, pa se nulta hipoteza odbacuje. Kako vrijedi i da je *p*-vrijednost jednaka $2,068 \cdot 10^{-5}$ što je manje od 0,05 = α, dolazi se do istog zaključka. Riječima: uz razinu značajnosti od 5%, odbacujemo hipotezu da niti jedna varijabla nije značajna u modelu. Ako se razina značajnosti promijeni u 1% ili 10%, *p*-vrijednost koja iznosi

[34] Vidjeti Dodatak 5.3.

$2{,}068 \cdot 10^{-5}$ je manja i od 0,01 i od 0,1, stoga ne dolazi do promjene u zaključku testa. Usporedbu smo mogli izvršiti i uspoređujući empirijski *F*-omjer s teorijskima za 1% i 10%, koje bi u RStudiju izračunali naredbama qf(0.99,3,4) i qf(0.9,3,4), pri čemu bismo izračunali vrijednosti 16,694 i 4,191.

2.2.7.3. Waldov test

Kao što je spomenuto u naslovu 2.1.9.4, općenito se u Waldovom testu pretpostavlja da jedan ili više parametara u regresijskom modelu zadovoljavaju određena linearna ograničenja. Ako se razmatra model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ u matričnom zapisu, *J* ograničenja na parametre $\boldsymbol{\beta}$ se u slučaju modela višestruke linearne regresije mogu pisati kao:

$$\begin{aligned} r_{11}\beta_0 + r_{12}\beta_1 + r_{13}\beta_2 + \ldots + r_{1k+1}\beta_k &= q_1 \\ r_{12}\beta_0 + r_{22}\beta_1 + r_{23}\beta_2 + \ldots + r_{2k+1}\beta_k &= q_2 \\ &\vdots \\ r_{J1}\beta_0 + r_{J2}\beta_1 + r_{J3}\beta_2 + \ldots + r_{Jk+1}\beta_k &= q_J \end{aligned}, \tag{2.215}$$

koja se općenito pišu u matričnoj formi:

$$\boldsymbol{R}\,\boldsymbol{\beta} = \boldsymbol{q}, \tag{2.216}$$

gdje je $\boldsymbol{R} \in \mathcal{M}_{J,k+1}$, $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ i $\boldsymbol{q} \in \mathbb{R}^{J}$, formiraju se nulta i alternativna hipoteza ovako:

$$\begin{aligned} H_0 &: \boldsymbol{R\beta} = \boldsymbol{q} \\ H_1 &: \boldsymbol{R\beta} \neq \boldsymbol{q} \end{aligned}. \tag{2.217}$$

Nekoliko primjera formiranja hipotezi u matričnom obliku za testiranje linearnih ograničenja na parametre modela su sljedeći.

i. Skupni test značajnosti svih nezavisnih varijabli u modelu:

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.218}$$

Provjerom zapisa (2.218) tako da se pomnože $\boldsymbol{R}$ i $\boldsymbol{\beta}$, dobiva se sustav jednadžbi:

$$\begin{aligned} 0 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2 + \ldots + 0 \cdot \beta_k &= 0 \\ 0 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2 + \ldots + 0 \cdot \beta_k &= 0 \\ &\vdots \\ 0 \cdot \beta_0 + 0 \cdot \beta_1 + 0 \cdot \beta_2 + \ldots + 1 \cdot \beta_k &= 0 \end{aligned}, \tag{2.219}$$

što je upravo pretpostavka nulte hipoteze skupnog testa značajnosti:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 . \tag{2.220}$$

ii. Pojedinačni test značajnosti *j*-te varijable u modelu:

$$\boldsymbol{R} = \begin{bmatrix} 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix} . \tag{2.221}$$

Provjerom zapisa (2.221) tako da se pomnože $\boldsymbol{R}$ i $\boldsymbol{\beta}$, dobiva se:

$$0 \cdot \beta_0 + 0 \cdot \beta_1 + ... + 1 \cdot \beta_j + ... + 0 \cdot \beta_k = 0 , \tag{2.222}$$

što je upravo pretpostavka nulte hipoteze testa značajnosti varijable *j*:

$$H_0 : \beta_j = 0 . \tag{2.223}$$

iii. Jednakost učinaka dviju varijabli u modelu, varijabla *j* i varijabla *l*:

$$\boldsymbol{R} = \begin{bmatrix} 0 & 0 & \cdots & 1 & \cdots & -1 & \cdots & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_l \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix} . \tag{2.224}$$

Provjerom zapisa (2.224) tako da se pomnože $\boldsymbol{R}$ i $\boldsymbol{\beta}$, dobiva se:

$$0 \cdot \beta_0 + 0 \cdot \beta_1 + ... + 1 \cdot \beta_j + ... - 1 \cdot \beta_l + ... + 0 \cdot \beta_k = 0 , \tag{2.225}$$

što je upravo pretpostavka nulte hipoteze testa:

$$H_0 : \beta_j = \beta_l \Rightarrow H_0 : \beta_j - \beta_l = 0 \Rightarrow H_0 : 1 \cdot \beta_j - 1 \cdot \beta_l = 0 . \tag{2.226}$$

iv. Zbroj učinaka varijabli *j*, *l* i *m* jednak je 3:

$$\boldsymbol{R}=\begin{bmatrix}0 & 0 & \cdots & 1 & \cdots & 1 & \cdots & 1 & \cdots & 0\end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_j\\ \vdots\\ \beta_l\\ \vdots\\ \beta_m\\ \vdots\\ \beta_k\end{bmatrix}, \boldsymbol{q}=\begin{bmatrix}3\end{bmatrix} \tag{2.227}$$

Provjerom zapisa (2.227) tako da se pomnože $\boldsymbol{R}$ i $\boldsymbol{\beta}$, dobiva se:

$$0\cdot\beta_0+0\cdot\beta_1+...+1\cdot\beta_j+...+1\cdot\beta_l+...+1\cdot\beta_m.+..+0\cdot\beta_k=0, \tag{2.228}$$

što je upravo pretpostavka nulte hipoteze testa:

$$H_0:\beta_j+\beta_l+\beta_m=3. \tag{2.229}$$

v. Sljedeća linearna ograničenja: $H_0:\beta_1+\beta_2=1,\ \beta_4=0$:

$$\boldsymbol{R}=\begin{bmatrix}0 & 1 & 1 & 0 & 0 & \dots & 0\\ 0 & 0 & 0 & 0 & 1 & \dots & 0\end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_k\end{bmatrix}, \boldsymbol{q}=\begin{bmatrix}1\\ 0\end{bmatrix} \tag{2.230}$$

Provjerom zapisa (2.230) tako da se pomnože $\boldsymbol{R}$ i $\boldsymbol{\beta}$, dobiva se sljedeći sustav jednadžbi:

$$\begin{aligned}&0\cdot\beta_0+1\cdot\beta_1+1\cdot\beta_2+0\cdot\beta_3+0\cdot\beta_4+...+0\cdot\beta_k=1\\ &0\cdot\beta_0+0\cdot\beta_1+0\cdot\beta_2+0\cdot\beta_3+1\cdot\beta_4.+..+0\cdot\beta_k=0\end{aligned} \tag{2.231}$$

što je upravo pretpostavka nulte hipoteze testa:

$$H_0:\beta_1+\beta_2=1,\ \beta_4=0. \tag{2.232}$$

Dakle, uočava se da je moguće u nultoj hipotezi pretpostaviti različita linearna ograničenja na parametre, što može biti temeljem ekonomske teorije ili temeljem iskustva istraživača.

U naslovu 2.1.9.4 je izvedena Waldova test veličina:

$$W=\boldsymbol{m}Var\left[\boldsymbol{m}\mid\boldsymbol{X}\right]^{-1}\boldsymbol{m}=\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{q}\right)'\left(\boldsymbol{R}\hat{\sigma}^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}'\right)\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{q}\right), \tag{2.233}$$

čija se procjena varijance $\hat{\sigma}^2$ vrši korištenjem formule (2.205) za slučaj višestruke linearne regresije. Ako je nulta hipoteza istinita, tada test veličina W u (2.233) asimptotski slijedi hi-kvadrat distribuciju s J stupnjeva slobode, $W \sim \chi^2(J)$.

Nakon formiranja hipotezi i izračuna empirijske Wald test veličina u (2.233) uspoređuje se s teorijskom veličinom koja također slijedi hi-kvadrat distribuciju, s J stupnjeva slobode, za zadanu razinu značajnosti α, $\chi^2_\alpha(J)$, koja se iščitava iz tablice kritičnih granica hi-kvadrat distribucije. Ako je $W > \chi^2_\alpha(J)$, odbacuje se nulta hipoteza, dok se za slučaj $W < \chi^2_\alpha(J)$ ne odbacuje. Slično tome, ako je p-vrijednost manja od α, odbacuje se nulta hipoteza, dok se u slučaju kada je p-vrijednost veća od α, nulta hipoteza ne odbacuje.

Dodatno, ako se test veličina u (2.233) podijeli s J stupnjeva slobode, W/J, tada se dobiva veličina koja slijedi F-distribuciju s J stupnjeva slobode u brojniku i $N–k–1$ stupnjeva slobode u nazivniku, i uz pretpostavku da su greške relacije nezavisne, identično normalno distribuirane[35]:

$$F = \frac{W}{J} = \frac{\left(\hat{\boldsymbol{\varepsilon}}_R^{'}\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}^{'}\hat{\boldsymbol{\varepsilon}}\right)/J}{\hat{\boldsymbol{\varepsilon}}^{'}\hat{\boldsymbol{\varepsilon}}\ /(N-k-1)} \sim F(J; N-k-1), \tag{2.234}$$

gdje je $\hat{\boldsymbol{\varepsilon}}_R^{'}\hat{\boldsymbol{\varepsilon}}_R$ suma kvadrata rezidualnih odstupanja u modelu s ograničenjima na parametre (što je pretpostavljeno u nultoj hipotezi), pa se može pisati i:

$$F = \frac{\left(SSR_R - SSR\right)/J}{SSR/(N-k-1)} \sim F(J; N-k-1). \tag{2.235}$$

Ono što se uočava u zapisu ove test veličine je sljedeće. Izraz $SSR_R - SSR$ je razlika sume kvadrata odstupanja za model u kojemu su nametnuta ograničenja u nultoj hipotezi i sume kvadrata modela bez ograničenja. Kada se ta razlika podijeli s brojem ograničenja J, dobije se sredina razlike kvadrata spomenutih odstupanja. S druge strane, nazivnik je sredina sume kvadrata odstupanja za originalni model bez ograničenja. Stoga se uspoređuje je li uvođenje ograničenja na parametre dovelo do značajnog smanjenja sume kvadrata rezidualnih odstupanja. Ako jest, tada je razlika $SSR_R - SSR$ značajna i u brojniku izraza (2.235) je velika vrijednost.

Primjer 2.35.

Za sljedeći procijenjeni regresijski model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4}$, zapišite sljedeće nulte hipoteze Waldova testa u matričnom obliku ($\boldsymbol{R\beta} = \boldsymbol{q}$):

a) Prosječni učinak svih nezavisnih varijabli u modelu iznosi 2.
b) Varijabla x_3 nije značajna u modelu.
c) Zbroj učinaka treće i četvrte nezavisne varijable jednak je 0,8.
d) Učinci prve i druge nezavisne varijable se poništavaju.

[35] Izvod vidjeti u Greene (2018).

a) $\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 8 \end{bmatrix}$, jer vrijedi $\frac{\beta_1 + \beta_2 + \beta_3 + \beta_4}{4} = 2$.

b) $\boldsymbol{R} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}$, jer se testira $\beta_3 = 0$.

c) $\boldsymbol{R} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0,8 \end{bmatrix}$, jer se testira $\beta_3 + \beta_4 = 0,8$.

d) $\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}$, jer se testira $\beta_1 = -\beta_2$.

Primjer 2.36.
Za procijenjen model u primjeru 2.23. (temeljem podataka u tablici 2.12.), provedite sljedeće Waldove testove:

a) Varijabla x_2 nije značajna u modelu.
b) Varijable x_1 i x_3 nisu značajne u modelu.
c) Učinak varijable $x3$ na nezavisnu varijablu iznosi 0,5.

a) Testira se sljedeća nulta hipoteza:

$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}$$

Stoga je u RStudiju potrebno u okviru paketa „car“ odrediti ograničenje „x2=0“ (naziv varijable pod navodnike izjednačiti s pretpostavljenom vrijednošću 0) te u okviru naredbe linearHypothesis(…) odrediti hi-kvadrat test jer Wald test veličina slijedi tu distribuciju. Ispis naredbi i rezultata dan je na slici 2.69. Kako je dobivena test veličina jednaka 0,2097 („Chisq“), s pripadajućom *p*-vrijednošću 0,647, što je veće od uobičajenih razina značajnosti, ne

odbacujemo nultu hipotezu. Interpretacija je sljedeća: uz razinu značajnosti od 5% (ili pri svim uobičajenim razinama značajnosti), ne odbacujemo hipotezu da varijabla x_2 nije značajna u modelu.

```
library(car)
ogranicenje<-"x2=0"
linearHypothesis(model,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## x2 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3
##
##   Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1      5 4.5319
## 2      4 4.3062  1   0.22574 0.2097      0.647
```

Slika 2.69. Naredbe i ispis Waldova testa

b) Testira se hipoteza:

$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Stoga je potrebno odrediti dva ograničenja naredbom ogranicenje<–c(„x1=0,“x3=0“). Ispis naredbi i rezultata dan je na slici 2.70. Kako je dobivena test veličina jednaka 263,51, s pripadajućom p-vrijednošću $2{,}2 \cdot 10^{-16}$, što je manje od uobičajenih razina značajnosti, odbacujemo nultu hipotezu. Interpretacija je sljedeća: uz razinu značajnosti od 5% (ili pri svim uobičajenim razinama značajnosti), odbacujemo hipotezu da varijable x_1 i x_3 nisu značajne u modelu.

```
library(car)
ogranicenje<-c("x1=0","x3=0")
linearHypothesis(model,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## x1 = 0
## x3 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3
##
##   Res.Df     RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1      6 287.984
## 2      4   4.306  2    283.68 263.51  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 2.70. Naredbe i ispis Waldova testa

c) Testira se hipoteza:

$$\boldsymbol{R\beta} = \boldsymbol{q}, \boldsymbol{R} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0,5 \end{bmatrix}$$

```
library(car)
ogranicenje<-c("x3=0.5")
linearHypothesis(model,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## x3 = 0.5
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3
##
##   Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1      5 4.4703
## 2      4 4.3062  1   0.16412 0.1524     0.6962
```

Slika 2.71. Naredbe i ispis Waldova testa

Stoga je potrebno odrediti ograničenje naredbom ogranicenje<–c(„x3=0.5“). Ispis naredbi i rezultata dan je na slici 2.71. Kako je dobivena test veličina jednaka 0,1524, s pripadajućom *p*-vrijednošću 0,6962, što je veće od uobičajenih razina značajnosti, ne odbacujemo nultu hipotezu. Interpretacija je sljedeća: uz razinu značajnosti od 5% (ili pri svim uobičajenim razinama značajnosti), ne odbacujemo hipotezu da učinak varijable x_3 na nezavisnu varijablu iznosi 0,5.

Primjer 2.37.
Provedimo odgovarajuće *F*-testove iz prethodnog primjera, koristeći formulu (2.234).

Hipoteze se zapisuju identično kao u primjeru 2.36., sada se samo odabir testa u RStudiju mijenja. U okviru naredbi za tip testa se bira „F“ (vidjeti slike 2.72., 2.73. i 2.74.). Tako je sada za odgovor u pitanju a) test veličina jednaka $F = 0,2097$, s *p*-vrijednošću 0,6708 (slika 2.72.) i ishod testa je isti kao u prethodnom primjeru pod a). Slično tome, u postupku b) (slika 2.73.), test veličina iznosi 131,75, s *p*-vrijednošću 0,00022, te je zaključak isti kao u prethodnom primjeru. Postupak c) (slika 2.74.) ostavlja se čitatelju za vježbu.

```
ogranicenje<-"x2=0"
linearHypothesis(model,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## x2 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      5 4.5319
## 2      4 4.3062  1   0.22574 0.2097 0.6708
```

Slika 2.72. Naredbe i ispis *F*-testa

```
linearHypothesis(model,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## x1 = 0
## x3 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1      6 287.984
## 2      4   4.306  2    283.68 131.75 0.0002236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 2.73. Naredbe i ispis *F*-testa

```
ogranicenje<-c("x3=0.5")
linearHypothesis(model,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## x3 = 0.5
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      5 4.4703
## 2      4 4.3062  1   0.16412 0.1524 0.7161
```

Slika 2.74. Naredbe i ispis *F*-testa

2.2.7.4. Parcijalni *F*-test

Skupni test značajnosti (*F*-test) koji se obrađivao u 2.2.7.2 odnosi se na ispitivanje značajnosti svih nezavisnih varijabli u regresijskom modelu. S druge strane, ako se žele testirati hipoteze koje se odnose na **dio varijabli** u modelu, u tom slučaju govorimo o **parcijalnom** *F*-testu. Ideja je usporediti sume kvadrata rezidualnih odstupanja originalnog modela, te onog u kojem su nametnuta ograničenja. Primjenjuje se u slučaju testiranja je li izostavljena značajna regresijska varijabla u modelu, ili je pak u model uključena nepotrebna regresijska varijabla. Općenito, empirijski *F*-omjer za provođenje parcijalnog *F*-testa može se zapisati u sljedećem obliku:

$$F = \frac{(SSR_R - SSR_*)/p}{SSR_*/(N-k-1)} \sim F(p; N-k-1), \qquad (2.236)$$

gdje SSR_R predstavlja sumu kvadrata rezidualnih odstupanja modela u kojemu je manji broj nezavisnih varijabli, dok SSR_* označava sumu kvadrata rezidualnih odstupanja modela u kojemu je veći broj nezavisnih varijabli. To će ovisiti o tome testira li se u nultoj hipotezi radi li se o izostavljenoj značajnoj regresijskoj varijabli (naslov 2.2.7.5) ili pak o uključenoj nepotrebnoj regresijskoj varijabli (naslov 2.2.7.6). Nadalje, p predstavlja broj varijabli koje se uključuju ili isključuju iz originalnog modela.

2.2.7.5. Napomena o izostavljenoj značajnoj regresijskoj varijabli

Problem izostavljene značajne regresijske varijable (engl. *omitted variable problem*) javlja se kada u modelu izostavimo neku varijablu koja je trebala biti uključena, no nismo mogli prikupiti podatke o toj varijabli i tako ju uključiti u sam model. Problem koji se pritom javlja jest to da je **procjenitelj** $\hat{\boldsymbol{\beta}}$ u tom slučaju **pristran** (vidjeti Dodatak 5.3 i naslov 2.1.3.3).

Testiranje je li značajna varijabla izostavljena iz modela može se provesti pomoću parcijalnog *F*-testa ili LR testa, na način da se usporede originalni modeli bez uključene varijable od interesa, s modelom gdje je uključena varijabla za koju se provodi test. Istovremeno se može testirati značajnost jedne ili više regresijskih varijabli, pri čemu nulta hipoteza testa pretpostavlja da nova varijabla ili skup novih varijabli nije značajan u modelu.

Za parcijalni *F*-test računa se sljedeći empirijski omjer:

$$F = \frac{(SSR - SSR_*)/J}{SSR_*/(N-k-1-J)} = \frac{(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} - \hat{\boldsymbol{\varepsilon}}_*'\hat{\boldsymbol{\varepsilon}}_*)/J}{\hat{\boldsymbol{\varepsilon}}_*'\hat{\boldsymbol{\varepsilon}}_*/(N-k-1-J)} \sim F(J; N-k-1-J), \quad (2.237)$$

gdje *SSR* predstavlja sumu kvadrata rezidualnih odstupanja modela u kojemu nismo uključili varijable za koje se provodi test, dok SSR_* označava sumu kvadrata rezidualnih odstupanja modela u kojemu smo ih uključili, *J* je broj novih uključenih varijabli. Ako je empirijski *F*-omjer veći od teorijskog, tj. $F_{emp} > F_\alpha(J; N–k–1–J)$, odbacuje se nulta hipoteza.

Za LR test, računa se sljedeća test veličina:

$$LR = -2(\ln L - \ln L_*) \sim \chi^2(J), \quad (2.238)$$

gdje ln*L* predstavlja maksimalnu vrijednost funkcije vjerodostojnosti za model bez dodanih regresijskih varijabli, dok lnL_* predstavlja maksimalnu vrijednost funkcije vjerodostojnosti za model u kojem je uključeno *J* novih regresijskih varijabli. Ako je test veličina *LR* veća od teorijske razine $\chi_\alpha^2(J)$, nulta hipoteza se odbacuje.

Primjer 2.38.
Temeljem podataka u tablici 15 iz primjera 2.23, procijenite model u kojemu varijabla *y* ovisi o varijablama x_2 i x_3, te potom provedite parcijalni *F*-test i LR test o izostavljenoj značajnoj regresijskoj varijabli x_1, uz razinu značajnosti od 5%.

Nulta hipoteza ovog testa glasi: varijabla x_1 je neznačajna u modelu.

Za provedbu parcijalnog *F*-testa procijene se 2 modela: prvi u kojemu su uključene sve tri varijable, te potom drugi u kojemu je isključena varijabla x_1. Kao ograničenje, tj. u nultoj hipotezi pretpostavlja se da varijabla x_1 nije značajna u modelu. Za svaki model računaju se SSR za potrebu izračuna test veličine u (2.237). Stoga je izračun sljedeći, temeljem ispisa na slici 2.75., gdje ssr_1 označava SSR_*, dok ssr_2 označava *SSR*:

$$F = \frac{(SSR - SSR_*)/J}{SSR_*/(N-k-1-J)} = \frac{(54{,}158 - 4{,}306)/1}{4{,}306/(8-2-1-1)} = 46{,}307$$

```
model1<-lm(y~x1+x2+x3)
ssr_1<-sum(resid(model1)^2)
model2<-lm(y~x2+x3)
ssr_2<-sum(resid(model2)^2)
ssr_1;ssr_2

## [1] 4.306203
## [1] 54.15769

emp<-((ssr_2-ssr_1)/1)/(ssr_1/(8-2-1-1))
emp
## [1] 46.30667
```

Slika 2.75. Parcijalni *F*-test o izostavljenoj značajnoj regresijskoj varijabli

Kako se radi o uključivanju jedne dodatne varijable u model, $J = 1$, dok u originalnom modelu bez te varijable broj nezavisnih varijabli iznosi $k = 2$. Pripadajuća *p*-vrijednost iznosi 0,002 (dobiveno naredbom 1-pf(emp,1,4)), što je manje od 0,05 pa time odbacujemo nultu hipotezu. Riječima, uz razinu značajnosti od 5% odbacujemo hipotezu da varijabla x_1 nije značajna u modelu. Test se mogao izvršiti i usporedbom empirijskog *F*-omjera s teorijskim (naredba qf(1-0.05,1,4)), koji iznosi 7,709. Uočava se kako je 46,307 > 7,709, pa je odluka da odbacujemo nultu hipotezu.

```
library(lmtest)
model2<-lm(y~x2+x3)
lrtest(model1,model2)

## Likelihood ratio test
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x2 + x3
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5  -8.874
## 2   4 -19.001 -1 20.255  6.779e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 2.76. LR test o izostavljenoj značajnoj regresijskoj varijabli

S druge strane, za provedbu LR testa uspoređuju se maksimalne vrijednosti funkcija vjerodostojnosti modela koji ima sve tri uključene varijable i onoga u kojemu je isključena samo prva (slika 2.76.). Sada se test veličina računa na sljedeći način:

$$LR = -2(\ln L - \ln L_*) = -2(-19{,}001-(-8{,}874)) = 20{,}255,$$

i pripadajuća *p*-vrijednost iznosi $6{,}8\cdot10^{-6}$, što je manje od 5% pa se ponovno odbacuje nulta hipoteza i identična je interpretacija riječima. Također, teorijska vrijednost hi-kvadrat distribucije iznosila bi (naredba qchisq(1-0.05,1)), 3,841, pa se ponovno uočava da je $LM > \chi^2_{0,05}(1)$, što upućuje na odbacivanje nulte hipoteze.

Primjer 2.39.
Usporedimo rezultate procijenjenih parametara u modelu iz prethodnog primjera kada su uključene sve tri nezavisne varijable, u odnosu na isključenu varijablu x_1.

Nulta hipoteza ovog testa glasi: varijabla x_1 je suvišna u modelu.

Ako se razmotri ispis na slici 2.77., uočava se da je sljedeći model za sve tri nezavisne varijable: $\hat{y}_i = -10{,}481 + 2{,}55x_{i1} - 0{,}08x_{i2} + 0{,}41x_{i3}$, dok u slučaju isključivanja varijable x_1 model glasi: $\hat{y}_i = 17{,}05 + 1{,}004x_{i2} - 1{,}06x_{i3}$. Uočimo kako je došlo do promjene predznaka i konstante i parametara uz varijable x_2 i x_3! Dakle, izostavljanjem značajne regresorske varijable iz modela može značajno izmijeniti procjene, i time rezultirati s modelom koji nije pouzdan za upotrebu.

```
model

##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Coefficients:
## (Intercept)           x1           x2           x3
##   -10.48102      2.55423     -0.08485      0.41084

model2

##
## Call:
## lm(formula = y ~ x2 + x3)
##
## Coefficients:
## (Intercept)           x2           x3
##      17.054        1.004       -1.064
```

Slika 2.77. Usporedba procijenjenih parametara modela iz primjera 2.38

2.2.7.6. Napomena o uključenoj nepotrebnoj regresijskoj varijabli

Ako se želi testirati je li jedna ili skup varijabli u modelu nepotrebna (suvišna, engl. *redundant variable test*), može se također provesti parcijalni *F*-test, kao i LR test. Ideja je usporediti dva modela: onaj u kojemu su uključene sve varijable od interesa, te drugi u kojemu su isključene varijable za koje se test provodi. Nulta hipoteza ovog testa glasi: „varijabla (odabran skup varijabli) je suvišna u modelu“.

Za parcijalni *F*-test računa se sljedeći empirijski omjer:

$$F = \frac{(SSR - SSR_*)/J}{SSR_*/(N-k-1)} = \frac{(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} - \hat{\boldsymbol{\varepsilon}}_*'\hat{\boldsymbol{\varepsilon}}_*)/J}{\hat{\boldsymbol{\varepsilon}}_*'\hat{\boldsymbol{\varepsilon}}_*/(N-k-1)} \sim F(J; N-k-1), \qquad (2.239)$$

gdje *SSR* predstavlja sumu kvadrata rezidualnih odstupanja modela u kojemu smo isključili *J* varijabli za koje se provodi test, dok SSR_* označava sumu kvadrata rezidualnih odstupanja modela u kojemu su te varijable ostavljene. Ako je empirijski *F*-omjer veći od teorijskog, tj. $F_{emp} > F_\alpha(J; N-k-1)$, odbacuje se nulta hipoteza.

Za LR test, računa se sljedeća test veličina:

$$LR = -2(\ln L - \ln L_*) \sim \chi^2(J), \qquad (2.240)$$

gdje $\ln L$ predstavlja maksimalnu vrijednost funkcije vjerodostojnosti za model gdje je isključeno *J* regresijskih varijabli, dok $\ln L_*$ predstavlja maksimalnu vrijednost funkcije vjerodostojnosti za model u kojem su te varijable ostavljene. Ako je test veličina *LR* veća od teorijske razine $\chi_\alpha^2(J)$, nulta hipoteza se odbacuje.

Primjer 2.40.
Temeljem podataka u tablici 15 iz primjera 2.23, procijenite model u kojemu varijabla y ovisi o varijablama x_1, x_2 i x_3, te potom provedite parcijalni F-test o uključenoj nepotrebnoj varijabli x_1, uz razinu značajnosti od 5%.

```
model<-lm(y~x1+x2+x3)
library(car)
ogranicenje<-"x1=0"
linearHypothesis(model,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## x1 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1      5 54.158
## 2      4  4.306  1    49.851 46.307 0.002437 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 2.78. Parcijalni F-test o uključenoj nepotrebnoj regresijskoj varijabli

Najprije je potrebno procijeniti i spremiti model u kojemu su uključene sve tri nezavisne varijable, te se provedbom naredbi prikazanima na slici 2.78. parcijalni F-test vrši na sljedeći način. Izračun empirijskog F-omjera je sljedeći:

$$F = \frac{(SSR - SSR_*)/J}{SSR_*/(N-k-1)} = \frac{(54{,}158 - 4{,}306)/1}{4{,}306/(8-3-1)} = 46{,}307\,.$$

Kako se radi o isključivanju jedne dodatne varijable u model, $J = 1$, dok u originalnom modelu s tom varijablom broj nezavisnih varijabli iznosi $k = 3$. Pripadajuća p-vrijednost iznosi 0,002, što je manje od 0,05 pa time odbacujemo nultu hipotezu. Riječima, uz razinu značajnosti od 5% odbacujemo hipotezu da je varijabla x_1 suvišna u modelu. Test se mogao izvršiti i usporedbom empirijskog F-omjera s teorijskim (naredba qf(1-0.05,1,4)), koji iznosi 7,709. Uočava se kako je 46,307 > 7,709, pa je odluka da odbacujemo nultu hipotezu.

2.2.7.7. Test o stabilnosti parametara

Ideja testa o stabilnosti parametara (engl. *Chow test, parameter stability test*) je ispitati jesu li procijenjeni parametri modela stabilni (nepromjenjivi) za dva poduzorka, pri čemu se češće koristi za vremenske nizove, u odnosu na presječne. Razvijen je u Chow (1960), te je postupak testiranja sljedeći. Uzorak veličine N podijeli se na dva poduzorka, veličine n_1 i n_2, pri čemu vrijedi $n_1 + n_2 = N$. Procijeni se regresijski model za cijeli uzorak N, te potom za oba poduzorka. Dakle, ako se pretpostavi da dolazi do promjene parametara u ovisnosti o poduzorku koji se razmatra, vrijede dvije regresijske jednadžbe.

Za poduzorak duljine n_1:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_k x_{ik} + \varepsilon_i, \qquad (2.241)$$

dok za poduzorak duljine n_2:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \qquad (2.242)$$

U nultoj hipotezi se testira H_0: $\alpha_j = \beta_j$, $j \in \{0, 1, \ldots, k\}$, odnosno da su svi parametri u modelu (2.241) jednaki ekvivalentnim parametrima u modelu (2.242).

Test veličina je sljedeći empirijski F-omjer:

$$F = \frac{(SSR - SSR_*)/(k+1)}{SSR_*/(N-2k-2)} \sim F(k+1; N-2k-2), \qquad (2.243)$$

gdje SSR predstavlja sumu kvadrata rezidualnih odstupanja u modelu koji je procijenjen nad cijelim uzorkom N, dok SSR_* predstavlja sumu kvadrata rezidualnih odstupanja kada se posebno procijene model (2.241) i (2.242), odnosno $SSR_* = SSR_{n1} + SSR_{n2}$. Ako je empirijski F-omjer veći od teorijskog, tj. $F_{emp} > F_\alpha(k+1; N-2k-2)$, odbacuje se nulta hipoteza.

Primjer 2.41.
Temeljem podataka u tablici 15 iz primjera 2.23., procijenite model jednostavne linearne regresije u kojemu varijabla y ovisi samo o varijabli x_1. Provedite test o stabilnosti parametara u modelu tako da uzorak podijelite na 4 opažanja za prvi i 4 opažanja za drugi model. Razina značajnosti je 5%.

Nulta hipoteza glasi: $\alpha_0 = \beta_0$ i $\alpha_1 = \beta_1$ (konstante i koeficijenti smjera jednaki su u oba modela). Temeljem naredbi prikazanih na slici 2.79., empirijski F-omjer iznosi 1,068, pri čemu je pripadajuća p-vrijednost 0,425. Kako je 0,425 > 0,05 ne odbacuje se nulta hipoteza o jednakosti parametara u modelu 1 i modelu 2.

```
y1<-y[1:4];x11<-x1[1:4]
y2<-y[5:8];x12<-x1[5:8]
library(gap)

chow.test(y1,x11,y2,x12)

##   F value     d.f.1     d.f.2   P value
## 1.0683583 2.0000000 4.0000000 0.4248619
```

Slika 2.79. Test o stabilnosti parametara

Ako se detaljnije razmotre oba modela na slici 2.80., uočava se da je procijenjeni koeficijent uz nezavisnu varijablu u oba modela veoma sličan (2,21 u prvome, te 2,02 u drugome modelu). No, empirijski F-omjer računat je na sljedeći način (vidjeti slike 2.81. i 2.82.):

$$F = \frac{(SSR - SSR_*)/(k+1)}{SSR_*/(N-2k-2)} = \frac{(9{,}195 - 1{,}309 - 4{,}685)/2}{(1{,}309 + 4{,}685)/(8 - 2\cdot 1 - 2)} = 1{,}608$$

```
summary(lm(y1~x11))

## Call:
## lm(formula = y1 ~ x11)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.5147     1.5788   -2.86  0.10363
```

```
## x11           2.2059     0.1387   15.90  0.00393 **
summary(lm(y1~x11))


## Call:
## lm(formula = y1 ~ x11)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.5147     1.5788   -2.86  0.10363
## x11           2.2059     0.1387   15.90  0.00393 **
```

Slika 2.80. Ispis modela za oba poduzorka

```
sum(resid(lm(y1~x11))^2)
## [1] 1.308824
sum(resid(lm(y2~x12))^2)
## [1] 4.684734
sum(resid(lm(y~x1))^2)
## [1] 9.19519
```

Slika 2.81. Sume kvadrata rezidualnih odstupanja potrebnih za izračun empirijskog *F*-omjera

((9.19519-1.308824-4.684734)/2)/((1.308824+4.684734)/(8-2-2))

Slika 2.82. Izračun empirijskog *F*-omjera temeljem slike 2.81.

2.2.7.8. Savjeti o određivanju podjele uzorka

Može se postaviti pitanje kako odrediti poduzorke za provođenje testa o stabilnosti parametara. Neki mogući pristupi su sljedeći:

- Vrijednosti zavisne varijable mogu se predočiti grafički, tako da se poredaju od najmanje do najveće vrijednosti ili u slučaju vremenskih nizova od početnog do krajnjeg datuma. Potom se razmotri postoji li veći skok u kretanju vrijednosti zavisne varijable i oko kojeg opažanja ili vremenskog razdoblja. To opažanje se potom može uzeti kao krajnje za prvi poduzorak, u odnosu na drugi.
- Ako se razmatraju vremenski nizovi, poduzorci se mogu podijeliti temeljem nekog značajnog datuma (slom dioničkog tržišta, promjena zakonodavstva, političkog vodstva, itd.).

2.2.7.9. RESET test

RESET (engl. *Ramsey regression specification test*, Ramsey 1969) test koristi se za testiranje je li odabrani funkcionalni oblik regresijskog modela pogrešan. Ako postoje određene nelinearnosti u podacima, onda linearni regresijski model neće dobro opisati povezanost zavisne i nezavisnih varijabli. Ideja RESET testa je da se najprije procijeni linearni regresijski model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \qquad (2.244)$$

prikupe se procijenjene vrijednosti zavisne varijable $\hat{y}_i$ uključe u novi model tako da se model (2.244) proširi za kvadrat od $\hat{y}_i$ i kub od $\hat{y}_i$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + u_i, \qquad (2.245)$$

te se testira nulta hipoteza H_0: $\gamma_1 = \gamma_2 = 0$, odnosno da je model (2.244) točno specificiran. Kako se radi o parcijalnom F-testu, jer se test provodi na dio parametara u modelu (2.245), empirijska test veličina slijedi F-distribuciju s 2 stupnja slobode u brojniku (jer nulta hipoteza sadrži 2 ograničenja), te $N–k–1–2$ stupnjeva slobode u nazivniku.

Primjer 2.42.
Temeljem podataka u tablici 15 iz primjera 2.23., procijenite model u kojemu varijabla y ovisi o varijablama x_1, x_2 i x_3, te potom provedite RESET test o adekvatnosti linearnog modela, tako da uključite kvadrat i kub procijenjenih vrijednosti zavisne varijable u originalnom modelu. Razina značajnosti je 5%.

Nulta hipoteza testa glasi: $\gamma_1 = \gamma_2 = 0$ za model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \varepsilon_i$. Za procijenjen model je na slici 2.83. proveden RESET test, tako da je za opciju potencije odabrano 2 i 3 (power=2:3). Test veličina iznosi $F = 0{,}44$, s 2 stupnja slobode u brojniku jer su dva ograničenja u nultoj hipotezi, dok su 2 stupnja slobode u nazivniku (8–3–1–2). Pripadajuća p-vrijednost iznosi 0,6935, koja je veća od razine značajnosti 0,05 stoga se ne odbacuje nulta hipoteza da su varijable $\hat{y}_i^2$ i $\hat{y}_i^3$ neznačajne u modelu (2.245).

```
model<-lm(y~x1+x2+x3)
library(lmtest)
resettest(model,power=2:3,type="fitted")

##
##  RESET test
##
## data:  model
## RESET = 0.44191, df1 = 2, df2 = 2, p-value = 0.6935
```

Slika 2.83. RESET test

```
model<-lm(y~x1+x2+x3)
y1<-fitted(model)
model_novo<-lm(y~x1+x2+x3+I(y1^2)+I(y1^3))
library(car)
ogranicenje<-c("I(y1^2)=0","I(y1^3)=0")
linearHypothesis(model_novo,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## I(y1^2) = 0
## I(y1^3) = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3 + I(y1^2) + I(y1^3)
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      4 4.3062
## 2      2 2.9864  2    1.3197 0.4419 0.6935
```

Slika 2.84. RESET test pomoću naredbi o linearnim ograničenjima na parametre

Alternativno, test se mogao provesti i u okviru paketa „car“ kao parcijalni F-test pomoću naredbi prikazanih na slici 2.84. Uočava se da je najprije procijenjen model, iz kojeg su

spremljene procijenjene vrijednosti zavisne varijable (y_1), te kao ograničenje navedena ograničenja da su kvadrat i kub od y_1 u novome modelu (model_novo) neznačajne. Ishod testa je identičan.

2.2.7.10. CUSUM test

Može se postaviti i pitanje stabilnosti procijenjenih parametara u regresijskom modelu. Ovdje se misli na to jesu li procijenjeni parametri nepromjenjivi ako bi se procijenio sam model na određenom podskupu razmatranog uzorka, pri čemu se ne zna unaprijed kada bi trebalo doći do strukturne promjene. Dakle, nije poznato opažanje i kod koje dolazi do značajne promjene parametara modela, ili pak ako se radi o vremenskim nizovima, u kojem trenutku t dolazi do tih promjena. U tom slučaju radi se o CUSUM testu (engl. *cumulative sum*), razvijenom u Page (1954). Ideja je procijeniti regresijski model nad određenim podskupom podataka iz uzorka, te pretpostaviti da će vrijednost procjenitelja $\hat{\boldsymbol{\beta}}$ koja je procijenjena nad tim podskupom i dalje ostati nepromijenjena za preostala opažanja iz uzorka. Najprije se procijeni model $\boldsymbol{y}_n = \boldsymbol{X}_n\boldsymbol{\beta}_n + \boldsymbol{\varepsilon}_n$ za podskup $i \in \{1, 2, \ldots, n\}$, gdje je $n < N$ (samo temeljem prvih n opažanja), te se u nultoj hipotezi pretpostavi H_0: $\boldsymbol{\beta}_n = \boldsymbol{\beta}$ za preostale opservacije, tj. od n+1 do N.

U drugom koraku se formira niz rezidualnih odstupanja koji se računaju na sljedeći način:

$$\hat{\varepsilon}_k = y_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{k-1}, \tag{2.246}$$

gdje $k \in \{n+1, n+2, \ldots, N\}$. Dakle, uzorak N se podijeli na n opažanja nad kojima se procijeni regresijski model, iz kojeg se koriste procijenjene vrijednosti $\hat{\boldsymbol{\beta}}_n$ da bi se izračunalo rezidualno odstupanje $\hat{\varepsilon}_{n+1}$ kao razlika $y_{n+1} - \boldsymbol{X}_{n+1}\hat{\boldsymbol{\beta}}_n$ za prvo opažanje nakon n-tog opažanja. Potom se postupak ponavlja za iduće opažanje, do posljednjeg, N-tog kako bi se na taj način formirao niz rezidualnih odstupanja u (2.246).

Nakon toga se formira kumulativna suma standardiziranih rezidualnih odstupanja:

$$W_k = \frac{\sum_{k=n+2}^{N} \hat{\varepsilon}_k}{\tilde{\sigma}\sqrt{N-n}}, \tag{2.247}$$

gdje je $\tilde{\sigma}\sqrt{N-n}$ procjena standardne devijacije rezidualnih odstupanja u poduzorku od n+1 do N-tog opažanja. Primijetimo da suma počinje od vrijednosti n+2 jer su potrebna najmanje dva opažanja kako bi se mogla računati standardna pogreška. Test veličina W_k slijedi CUSUM distribuciju, s očekivanjem 0 i povećanjem varijance kako se povećava uzorak[36]. Za fiksnu veličinu uzorka, mogu se izračunati gornja i donja granica kretanja vrijednosti W_k unutar kojih treba ostati test veličina ako nema strukturnih promjena u procijenjenim parametrima originalnog modela. Nulta hipoteza se odbacuje ako vrijednost W_k izlazi van tih granica.

Dodatno se može provesti i F-test, tj. Chow test o stabilnosti parametara, no potrebno je znati unaprijed za koje opažanje dolazi do promjene u parametrima modela. Nadalje, moguće je nadograditi test o stabilnosti parametara, na način da se ne određuje točno ono opažanje i kod koje dolazi do promjene u parametrima modela, već da se odrede donja i gornja granica između

[36] O konstrukciji granica, vidjeti Zeilesis (2000).

kojih opažanja dolazi do strukturnih promjena, na način da se u suštini računa pomični F-test od nekog opažanja $n+l$ kao donje granice, te do nekog drugog opažanja $n+m$ kao gornje granice, pri čemu vrijedi: $n < n + l < n + m < N$. Potom se empirijski F-omjer može izračunati kao prosječna vrijednost svih F-omjera izračunatih u podskupu između donje i gornje granice[37]. Nulta hipoteza testa se odbacuje ako je empirijski F-omjer veći od teorijskog, ili ako je p-vrijednost manja od zadane razine značajnosti. U praksi je za donju i granicu testa uobičajeno izdvojiti prvih 15% opažanja, kao i posljednjih 15%.

Primjer. 2.43.
Temeljem podataka u tablici 16, procijenite model u kojemu varijabla y ovisi o varijablama x_1, x_2 i x_3, te potom provedite CUSUM test o stabilnosti parametara u modelu. Dodatno provedite i odgovarajući F-test te donesite zaključak uz razinu značajnosti od 5%.

Tablica 16. Podaci o varijablama y, $x1$, $x2$ i $x3$

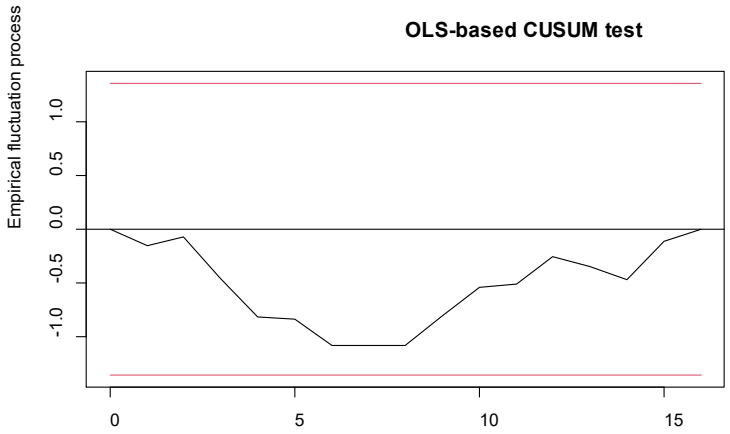
Varijabla	Vrijednosti															
y	18	29	21	11	7	25	44	1	0	43	24	6	10	20	28	17
x_1	10	15	12	7	4	14	22	1	0	21	13	3	6	11	14	9
x_2	13	12	16	8	10	11	25	5	4	24	10	9	7	15	11	12
x_3	9	3	7	14	18	2	1	22	21	0	1	17	13	6	2	8

Testiranje se provodi u okviru paketa strucchange u RStudiju. Kako je paket primjenjiv nad vremenskim nizovima, podaci u primjeru su privremeno definirani kao vremenski nizovi naredbama ts(…), vidjeti sliku 2.85. Potom, naredba efp(…) provodi CUSUM test za procijenjeni model, te je grafički pomoću naredbe plot(…) predočen niz W_k, predočen na slici 2.83. Nadalje, F-test je proveden također na slici 2.85., pomoću naredbe Fstats(…), gdje su određene kao donja i gornja granica treća, odnosno četvrto opažanje po redu. Kako u ovome primjeru raspolažemo samo s osam opažanja, odabrane su spomenute granice na način da postoji dovoljno opažanja za izračun pomičnih F-vrijednosti. U nultoj hipotezi se pretpostavlja da ne dolazi do strukturne promjene u parametrima regresijskog modela. Ako se razmotri grafički prikaz CUSUM testa na slici 2.86, uočava se da veličina W_k ne izlazi van granica unutar kojih je područje ne odbacivanja nulte hipoteze. Nadalje, ako se provodi F-test (slika 2.85), uočava se da je test veličina jednaka 1,896, uz pripadajuću p-vrijednost 0,773 > 0,05 pa se također ne odbacuje nulta hipoteza da su procijenjeni parametri stabilni za cijeli uzorak.

```
y<-ts(y,start=1,frequency = 1)
x1<-ts(x1,start=1,frequency = 1)
x2<-ts(x2,start=1,frequency = 1)
x3<-ts(x3,start=1,frequency = 1)
cusum <- efp(y~x1+x2+x3, type = "OLS-CUSUM")
plot(cusum,xlab=NA)

fs <- Fstats(y~x1+x2+x3, from = 10, to = 11)
sctest(fs, type="aveF")

##
##  aveF test
##
## data:  fs
## ave.F = 1.8958, p-value = 0.773
```

Slika 2.85. Naredbe potrebne za CUSUM test i odgovarajući F-test

[37] Mogući su i drugi načini izračuna, vidjeti Andrews (1993) i Andrews i Ploberger (1994).



Slika 2.86. CUSUM test predočen grafički

2.2.8. Predviđanje modelom višestruke linearne regresije

Ako se razmatra model višestruke linearne regresije $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_3 x_{i3} + \varepsilon_i$, te se pretpostavi da takav model vrijedi i u okolini promotrenih točaka (presječni podaci) ili u budućnosti, tada se **predviđena vrijednost zavisne varijable** temeljem opaženih ili pretpostavljenih vrijednosti nezavisne varijable x_f može zapisati predviđena vrijednost zavisne varijable ovako:

$$y_f = \beta_0 + \beta_1 x_{f1} + \beta_2 x_{f2} + \ldots + \beta_3 x_{f3} + \varepsilon_f, \tag{2.248}$$

pri čemu vrijedi $\varepsilon_f \sim N(0, \sigma^2)$ i ako vrijede pretpostavke regresijskog modela, procijenjena (odnosno predviđena) vrijednost iznosi:

$$\hat{y}_f = \hat{\beta}_0 + \hat{\beta}_1 x_{f1} + \hat{\beta}_2 x_{f2} + \ldots + \hat{\beta}_k x_{fk} = \boldsymbol{x}_f' \hat{\boldsymbol{\beta}}, \tag{2.249}$$

za koju vrijedi $\hat{y}_f \sim N(\boldsymbol{x}_f' \hat{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{x}_f' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_f')$. **Pogreška predviđanja** (engl. *prediction error*) računa se ovako:

$$\hat{\varepsilon}_f = \boldsymbol{y}_f - \boldsymbol{x}_f' \hat{\boldsymbol{\beta}} = \boldsymbol{x}_f' \boldsymbol{\beta} + \varepsilon_f - \boldsymbol{x}_f' \hat{\boldsymbol{\beta}} = \boldsymbol{x}_f' \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) + \varepsilon_f, \tag{2.250}$$

čija je očekivana vrijednost jednaka:

$$\boldsymbol{E}\left(\hat{\varepsilon}_f \right) = E\left[\boldsymbol{x}_f' \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) + \varepsilon_f \right] = \boldsymbol{x}_f' \boldsymbol{0} + 0 = 0. \tag{2.251}$$

Varijanca predviđanja (engl. *prediction variance*) jednaka je (vidjeti Greene, 2018 za izvod):

$$\mathrm{Var}\left(\hat{\varepsilon}_f \right) = \sigma^2 \left(1 + \boldsymbol{x}_f' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_f' \right). \tag{2.252}$$

Standardna devijacija predviđanja ili prognostička pogreška, računa se formulom:

$$SE(\hat{\varepsilon}_f) = \sqrt{\sigma^2 \left(1 + \boldsymbol{x}_f' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_f' \right)}, \tag{2.253}$$

gdje je potrebno standardnu pogrešku regresije procijeniti temeljem izraza (2.186), odnosno izrazom $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N-k-1} \sim \chi^2(N-k-1)$. **Standardizirana devijacija regresije** slijedi *t*-distribuciju s *N*–*k*–1 stupnjeva slobode:

$$\frac{\hat{\varepsilon}_f}{SE(\hat{\varepsilon}_f)} = \frac{y_f - \hat{y}_f}{SE(y_f - \hat{y}_f)} \sim t(N-k-1), \qquad (2.254)$$

stoga se **interval predviđanja (prognostički interval)** može procijeniti formulom:

$$P\left(\hat{y}_f - t_{\gamma/2} SE(y_f - \hat{y}_f) < y_f < \hat{y}_f + t_{\gamma/2} SE(y_f - \hat{y}_f)\right) = 1-\gamma, \qquad (2.255)$$

gdje 1–γ predstavlja pouzdanost procjene, a $t_{\gamma/2}$ predstavlja koeficijent pouzdanosti. **Interpretacija intervala** (2.255) glasi: u (1–γ)100% slučajeva će, uz dane vrijednosti nezavisnih varijabli $\boldsymbol{x}_f$, vrijednost zavisne varijable se nalaziti između $\hat{y}_f - t_{\gamma/2} SE\left(y_f - \hat{y}_f\right)$ i $\hat{y}_f + t_{\gamma/2} SE\left(y_f - \hat{y}_f\right)$ jedinica.

Primjer 2.44.
Temeljem procijenjenog modela iz primjera 2.23. (tablica 2.15), koliko iznosi predviđena vrijednost zavisne varijable ako se pretpostavlja vrijednost $x_1 = 10$, $x_2 = 5$ i $x_3 = 10$? Koliko iznosi interval predviđanja uz 1–γ=0,95? Interpretirajmo dobivene rezultate.

Najprije je procijenjen i spremljen model u kojemu zavisna varijabla ovisi o tri nezavisne varijable, a slika 2.87. predočava naredbe potrebne za predviđanje i intervalnu procjenu.

```
novo<-data.frame(x1=10,x2=5,x3=10)
predict(model,newdata = novo,interval = "confidence",level=.95)

##        fit      lwr      upr
## 1 18.74541 15.07897 22.41184
```

Slika 2.87. Naredbe potrebne za predviđanje vrijednosti zavisne varijable i interval predviđanja

Uočava se da za $x_1 = 10$, $x_2 = 5$ i $x_3 = 10$ vrijedi: $\hat{y}_f = 18{,}75$ što znači da se ovim modelom predviđa da za vrijednost varijable x_1 od 10 jedinica, varijable x_2 5 jedinica i varijable x_3 10 jedinica, očekivana razina zavisne varijable iznosi u prosjeku 18,75 jedinica. U 95% slučajeva za pretpostavljenu vrijednost varijable x_1 od 10 jedinica, varijable x_2 5 jedinica i varijable x_3 10 jedinica, stvarna vrijednost zavisne varijable bit će između 15,08 i 22,41 jedinica.

2.2.9. Sveobuhvatan primjer

S web stranice Svjetske banke (2020) preuzeti su podaci o ukupnoj potrošnji (milijarde konstantne LCU, engl. *local currency unit*), ukupnom BDP-u (milijarde konstantne LCU) te indeksu potrošačkih cijena (indeksni bodovi) za 140 zemalja svijeta u 2018. godini. Učitana je datoteka „**potrosnja.txt**“ u RStudio.

a) Procijenite model u kojemu potrošnja ovisi o BDP-u i indeksu potrošačkih cijena kao lin-lin model (m1), te kao log-log model (m2) i konačno kao model sa standardiziranim varijablama (m3). Pomoću naredbe stargazer spojite rezultate procjena u jednu tablicu. Interpretirajte procijenjene parametre uz nezavisne varijable u sva tri modela.

Na slici 2.88 prikazane su naredbe učitavanja datoteke u RStudio, zajedno s naredbama za spremanje tri modela te njihovim prikazom u jednoj zajedničkoj tablici. Ispis je prikazan na slici 2.89. Uočava se da su procijenjeni modeli sljedeći, gdje $\hat{y}_i$ predstavlja procijenjenu vrijednost varijable potrošnja, dok x_{1i} i x_{2i} predstavljaju BDP i indeks potrošačkih cijena:

$$\text{m1: } \hat{y}_i = 3951,83 + 0,645x_{i1} + 4,202x_{i2}$$

$$\text{m2: } \widehat{\ln y_i} = -0,458 + 0,956 \ln x_{i1} + 0,116 \ln x_{i2}$$

$$\text{m3: } \hat{y}_i^* = 0,998x_{i1}^* + 0,001x_{i2}^*$$

```
potrosnja<-read.table("potrosnja.txt",header=T,sep="\t")
m1<-lm(potrosnja~bdp+cijene,data=potrosnja)
m2<-lm(log(potrosnja)~log(bdp)+log(cijene),data=potrosnja)
m3<-lm(scale(potrosnja)~0+scale(bdp)+scale(cijene),data=potrosnja)
library(stargazer)
stargazer(list(m1,m2,m3),type="text")
```

Slika 2.88. Naredbe potrebne za ispis modela m1, m2 i m3

```
## ================================================================================================
##                                                Dependent variable:
## ------------------------------------------------------------------------------------------------
##                               potrosnja              log(potrosnja)           scale(potrosnja)
##                                  (1)                      (2)                       (3)
## ------------------------------------------------------------------------------------------------
## bdp                           0.645***
##                                (0.003)
##
## cijene                          4.202
##                               (38.637)
##
## log(bdp)                                                0.956***
##                                                         (0.016)
##
## log(cijene)                                              0.116
##                                                         (0.166)
##
## scale(bdp)                                                                        0.998***
##                                                                                   (0.005)
##
## scale(cijene)                                                                       0.001
##                                                                                   (0.005)
## Constant                     3,951.826                   -0.458
##                             (6,268.220)                 (0.816)
## ------------------------------------------------------------------------------------------------
## Observations                    140                       140                       140
## R2                             0.997                     0.965                     0.997
## Adjusted R2                    0.997                     0.965                     0.997
## Residual Std. Error    34,192.400 (df = 137)         0.556 (df = 137)          0.056 (df = 138)
## F Statistic         21,855.88*** (df = 2; 137) 1,906.151*** (df = 2; 137) 22,015.41*** (df = 2; 138)
## ================================================================================================
## Note:                                                                *p<0.1; **p<0.05; ***p<0.01
```

Slika 2.89. Ispis procijenjenih modela m1, m2 i m3

Interpretacija parametara je sljedeća:

- Model m1: $\hat{\beta}_1 = 0,645$ i $\hat{\beta}_2 = 4,202$. Ako se BDP zemlje poveća za 1 milijardu LCU, uz nepromijenjen indeks potrošačkih cijena, potrošnja te zemlje se poveća u prosjeku za 0,645 milijardi LCU. Ako se indeks potrošačkih cijena zemlje poveća za 1 indeksni bod, uz nepromijenjen BDP, potrošnja te zemlje se poveća u prosjeku za 4,202 milijardi LCU.
- Model m2: $\hat{\beta}_1 = 0,956$ i $\hat{\beta}_2 = 0,116$. Ako se BDP zemlje poveća za 1%, uz nepromijenjen indeks potrošačkih cijena, potrošnja te zemlje se poveća u prosjeku za 0,956%. Ako se indeks potrošačkih cijena zemlje poveća za 1%, uz nepromijenjen BDP, potrošnja te zemlje se poveća u prosjeku za 0,116%.
- Model m3: $\hat{\beta}_1^* = 0,998$ i $\hat{\beta}_1^* = 0,001$. Ako se BDP zemlje poveća za 1 standardnu devijaciju, uz nepromijenjen indeks potrošačkih cijena, potrošnja te zemlje se poveća u prosjeku za 0,998 standardnih devijacija. Ako se indeks potrošačkih cijena zemlje poveća za 1 standardnu devijaciju, uz nepromijenjen BDP, potrošnja te zemlje se poveća u prosjeku za 0,001 standardnih devijacija. Očito varijabla BDP ima puno jači učinak na zavisnu varijablu u odnosu na varijablu indeks potrošačkih cijena.

b) Usporedite modele m1 i m2 temeljem ispisa na slici 2.86. koji je reprezentativniji i objasnite zašto. Interpretirajte odgovarajuće mjere koje uspoređujete.

Oba modela imaju jednak broj nezavisnih varijabli, stoga se može koristiti koeficijent determinacije. Temeljem ispisa na slici 2.89. uočava se da je (korigirani)[38] koeficijent determinacije za model m1 jednak 0,997, dok je za m2 jednak 0,965. Bolji je model m1 jer je više varijacija zavisne varijable protumačeno tim modelom.

c) Provedite jednosmjerni *t*-test ($\alpha = 5\%$) za onaj model koji je bolji iz postupka b).

Kako je bolji model m1, promatra se njegov detaljniji ispis na slici 2.90. Kako su oba procijenjena parametra pozitivna (i za BDP i za varijablu indeks potrošačkih cijena), provode se *t*-testovi na gornju granicu, uz naredbu za teorijsku razinu *t*-omjera za $\alpha = 5\%$ i 137 stupnjeva slobode prikazanu na slici 2.91. Provedba jednosmjernih testova je sljedeća:

Varijabla BDP	**Varijabla indeks potrošačkih cijena**
$H_0 : \beta_1 = 0$ $H_1 : \beta_1 > 0$	$H_0 : \beta_2 = 0$ $H_1 : \beta_2 > 0$
$t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0,645}{0,003} = 209,073$	$t_2 = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{4,202}{38,64} = 0,109$
$t_{0,05}(N-k-1=137) = 1,656$	$t_{0,05}(N-k-1=137) = 1,656$
$t_1 > t_{0,05}(137) \rightarrow$ Odbacujem H_0 ili p-v $\approx 0 < 0,05 \rightarrow$ Odbacujem H_0	$t_2 < t_{0,05}(137) \rightarrow$ Ne odbacujem H_0 ili p-v $= 0,914 > 0,05 \rightarrow$ Ne odbacujem H_0

[38] Kako se radi o ispisu u kojemu su vrijednosti zaokružene na 3 decimale, ne uočava se razlika između koeficijenta determinacije i korigiranog koeficijenta determinacije. Međutim, korigirani koeficijent u slučaju svih modela ima manju vrijednost od originalnog.

```
summary(m1)

##
## Call:
## lm(formula = potrosnja ~ bdp + cijene, data = potrosnja)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -126036   -4467   -4373   -3738  329885
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.952e+03  6.268e+03   0.630    0.529
## bdp         6.450e-01  3.085e-03 209.073   <2e-16 ***
## cijene      4.202e+00  3.864e+01   0.109    0.914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34190 on 137 degrees of freedom
## Multiple R-squared:  0.9969, Adjusted R-squared:  0.9968
## F-statistic: 2.186e+04 on 2 and 137 DF,  p-value: < 2.2e-16
```

Slika 2.90. Ispis modela m1

Uz razinu značajnosti od 5%, odbacujemo hipotezu da varijabla BDP nije značajna u modelu, te ne odbacujemo hipotezu da varijabla indeks potrošačkih cijena nije značajna u modelu.

```
cbind(qt(0.95,137), qf(0.95,2,137))

##          [,1]     [,2]
## [1,] 1.656052 3.062204
```

Slika 2.91. Teorijske vrijednosti *t*-omjera i *F*-omjera

d) Provedite skupni test značajnosti za onaj model koji je bolji iz postupka b).

Kako se radi o modelu m1, temeljem slike 2.90, provodi se skupni test značajnosti kako slijedi: $H_0 : \beta_1 = \beta_2 = 0$, $H_1 : \exists \beta_j \neq 0, j \in \{1,2\}$, F = 21,860, p-vrijednost ≈ 0. Pritom vrijedi F = 21,860 > 3,06 = $F_{0,05}(2,137)$ (slika 2.91.), odnosno p-v < α, stoga se odbacuje nulta hipoteza. Uz razinu značajnosti od 5%, odbacujemo hipotezu da niti jedna varijabla nije značajna u modelu m1.

e) Procijenite dodatna dva modela: log-lin (m4) i lin-log (m5) između odabranog skupa varijabli, te spojite pomoću naredbe stargazer modele m1, m2, m4 i m5. Interpretirajte procijenjene parametre uz nezavisne varijable za modele m4 i m5.

Zajednički ispis modela m1, m2, m4 i m5 prikazan je na slici 2.92. Sada su još interpretacije za posljednja dva modela sljedeće[39]:

- Model m4: $\hat{\beta}_1 = 1,2 \cdot 10^{-6}$ i $\hat{\beta}_2 = -0,005$. Ako se BDP zemlje poveća za jednu milijardu LCU, uz nepromijenjen indeks potrošačkih cijena, potrošnja te zemlje se poveća u prosjeku za $1,2 \cdot 10^{-4}$ %. Ako se indeks potrošačkih cijena zemlje poveća za jedan indeksni bod, uz nepromijenjen BDP, potrošnja te zemlje se smanji u prosjeku za 0,05%.

[39] Napomena: naredbom summary(m4) uočava se da procijenjen parametar uz varijablu BDP ne iznosi 0, već $1,2 \cdot 10^{-6}$.

- Model m5: $\hat{\beta}_1 = 78786,46$ i $\hat{\beta}_2 = 2633,681$. Ako se BDP zemlje poveća za 1%, uz nepromijenjen indeks potrošačkih cijena, potrošnja te zemlje se poveća u 787,86 milijardi LCU. Ako se indeks potrošačkih cijena zemlje poveća za 1%, uz nepromijenjen BDP, potrošnja te zemlje se poveća u prosjeku za 26,33 milijardi LCU.

```
m4<-lm(log(potrosnja)~bdp+cijene,data=potrosnja)
m5<-lm(potrosnja~log(bdp)+log(cijene),data=potrosnja)
stargazer(list(m1,m2,m4,m5),type="text")
```

```
## ================================================================================
##                                          Dependent variable:
##                                ---------------------------------------------
##                                  potrosnja       log(potrosnja)        potrosnja
##                                     (1)          (2)         (3)          (4)
## --------------------------------------------------------------------------------
## bdp                              0.645***                 0.00000***
##                                   (0.003)                 (0.00000)
##
## cijene                             4.202                   -0.0005
##                                   (38.637)                 (0.003)
##
## log(bdp)                                        0.956***              78,786.460***
##                                                 (0.016)               (15,707.190)
##
## log(cijene)                                      0.116                  2,633.681
##                                                 (0.166)               (167,918.800)
##
## Constant                         3,951.826       -0.458     6.758***  -477,697.900
##                                 (6,268.220)     (0.816)     (0.506)   (825,124.000)
##
## --------------------------------------------------------------------------------
## Observations                        140          140         140          140
## R2                                 0.997        0.965       0.146        0.156
## Adjusted R2                        0.997        0.965       0.134        0.144
## Residual Std. Error (df = 137)  34,192.400      0.556       2.758     561,989.900
## F Statistic (df = 2; 137)      21,855.880*** 1,906.151*** 11.717***    12.658***
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

Slika 2.92. Ispis modela m1, m2, m4 i m5

f) Za najbolji model u postupku e) (pojasnite najprije zašto je najbolji) odredite i interpretirajte intervalne procjene parametara na razini pouzdanosti od 90%.

Uočava se da je na slici 2.92. između 4 procijenjena modela ponovno najbolji model m1, jer ima najveći koeficijent determinacije (jednak je broj nezavisnih varijabli u svim modelima) i najveći korigirani koeficijent determinacije. Sada se razmatra slika 2.93. intervalna procjena parametara, $P(0,64 < \beta_1 < 0,65) = 0,9$ i $P(-59,78 < \beta_2 < 68,19) = 0,9$ i čija je interpretacija sljedeća.

Uz razinu pouzdanosti od 90%, ako se vrijednost BDP-a poveća za milijardu LCU, uz nepromijenjenu vrijednost indeksa potrošačkih cijena, vrijednost potrošnje će se povećati u prosjeku između 0,64 i 0,65 milijardi LCU. Uz razinu pouzdanosti od 90%, ako se vrijednost indeksa potrošačkih cijena poveća za jedan bod, uz nepromijenjenu vrijednost BDP-a, vrijednost potrošnje će se promijeniti u prosjeku za –59,78 milijardi LCU 68,19 milijardi LCU. Uočava se kako je učinak 0 uključen u intervalnu procjenu za varijablu indeks potrošačkih cijena, što potvrđuje rezultate prethodnog *t*-testa za tu varijablu (da nije značajna u modelu).

```
confint(m1,level=.9)

##                         5 %         95 %
## (Intercept) -6428.6717521 1.433232e+04
## bdp             0.6399042 6.501225e-01
## cijene        -59.7832864 6.818696e+01
```

Slika 2.93. Intervalna procjena parametara za model m1

g) Izračunajte koeficijent determinacije, korigirani koeficijent determinacije, procjenu standardne devijacije regresije, procjenu koeficijenta varijacije regresije te koeficijent višestruke linearne korelacije i potom ih interpretirajte za model koji je bio najbolji u postupku e.

Kako je najbolji model m1, promatramo najprije sliku 2.92., kako bi temeljem ispisa zapisali sljedeće: $R^2 = 0{,}9969$, $\overline{R}^2 = 0{,}9968$, $R = \sqrt{0{,}9969}$ 0,9984, te ispis na slici 2.94., temeljem kojeg pišemo: $\hat{\sigma} =$ 34,192,4 i $\hat{V} = 37{,}67\%$.

```
y<-potrosnja$potrosnja
sazetak<-summary(m1)
cbind(sazetak$sigma,sazetak$sigma/mean(y))

##         [,1]      [,2]
## [1,] 34192.4 0.3766832
```

Slika 2.94. Procjena standardne devijacije regresije i koeficijenta varijacije regresije za m1

Interpretacije su sljedeće. Model m1 pojašnjava ukupno 99,69% varijacija varijable potrošnja. Postoji jaka linearna povezanost između varijable potrošnja i nezavisnih varijabli (jer je koeficijent višestruke linearne korelacije veoma blizu jedinične vrijednosti, iznosi 0,9984). Na temelju ove prve tri mjere, model je izvrstan. Prosječno odstupanje empirijskih od procijenjenih vrijednosti varijable potrošnja iznosi 34192,40 milijardi LCU, što je 37,67% relativno, pa je temeljem ove druge dvije mjere model umjereno dobar.

h) Za najbolji model u postupku b), provedite Waldov test je li učinak BDP-a na potrošnju dvostruko veći od učinka indeksa cijena. Zapišite matrično nultu hipotezu testa, te donesite odluku uz razinu značajnosti od 5%.

```
ogranicenje<-"bdp=2*cijene"
linearHypothesis(m1,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## bdp - 2 cijene = 0
##
## Model 1: restricted model
## Model 2: potrosnja ~ bdp + cijene
##
##   Res.Df        RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1    138 1.6018e+11
## 2    137 1.6017e+11  1  11785882 0.0101       0.92
```

Slika 2.95. Waldov test

Provodi se test na sljedeći način (slika 2.95.):

$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q}$$
$$H_0 : \boldsymbol{R\beta} \neq \boldsymbol{q}, \quad \boldsymbol{R} = \begin{bmatrix} 0 & 1 & -2 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}, \quad W = 0{,}0101, p\text{-v} = 0{,}92.$$

Kako je test veličina jednaka 0,0101, što je manje od teorijske razine 3,84 (dobivene naredbom qchisq(1-0.05,1)), uz pripadajuću *p*-vrijednost 0,92 što je veće od 0,05, ne odbacuje se nulta hipoteza. Dakle, uz razinu značajnosti od 5% ne odbacuje se hipoteza da je učinak varijable BDP dvostruko veći u odnosu na učinak varijable indeks potrošačkih cijena na zavisnu varijablu potrošnja.

i) Za najbolji model u postupku e, provedite Waldov test je li granična sklonost potrošnji jednaka 0.7 i indeks potrošačkih cijena suvišan u modelu. Zapišite matrično nultu hipotezu testa, te donesite odluku uz razinu značajnosti od 5%.

```
ogranicenje<-c("bdp=.7","cijene=0")
linearHypothesis(m1,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## bdp = 0.7
## cijene = 0
##
## Model 1: restricted model
## Model 2: potrosnja ~ bdp + cijene
##
##   Res.Df        RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1    139 5.3157e+11
## 2    137 1.6017e+11  2 3.714e+11 317.67  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 2.96. Waldov test

Provodi se test na sljedeći način (vidjeti sliku 2.96.):

$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q}$$
$$H_0 : \boldsymbol{R\beta} \neq \boldsymbol{q}, \quad \boldsymbol{R} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0{,}7 \\ 0 \end{bmatrix}, \quad W = 317{,}67, p\text{-v} \approx 0.$$

Kako je test veličina jednaka 317,67, što je manje od teorijske razine 5,99 (dobivene naredbom qchisq(1-0.05,2)), uz pripadajuću *p*-vrijednost veoma malu, što je manje od 0,05, odbacuje se nulta hipoteza. Dakle, uz razinu značajnosti od 5% odbacuje se hipoteza da je granična sklonost potrošnji jednaka 70% i indeks potrošačkih cijena suvišan u modelu.

j) Provedite odgovarajuće *F*-testove iz postupaka h) te i).

Hipoteze testa su identične onima zapisanima u postupcima h) i i), sada je prilikom izračuna potrebno odabrati opciju „F“ za provedbu ovog testa (vidjeti sliku 2.97., odnosno 2.98.).

```
ogranicenje<-"bdp=2*cijene"
linearHypothesis(m1,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## bdp - 2 cijene = 0
##
## Model 1: restricted model
## Model 2: potrosnja ~ bdp + cijene
##
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    138 1.6018e+11
## 2    137 1.6017e+11  1  11785882 0.0101 0.9202
```

Slika 2.97. *F*-test za postupak h)

```
ogranicenje<-c("bdp=.7","cijene=0")
linearHypothesis(m1,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## bdp = 0.7
## cijene = 0
##
## Model 1: restricted model
## Model 2: potrosnja ~ bdp + cijene
##
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    139 5.3157e+11
## 2    137 1.6017e+11  2 3.714e+11 158.84 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 2.98. F-test za postupak i)

Kako je empirijski *F*-omjer jednak 0,0101, uz pripadajuću *p*-vrijednost 0,9292, što je veće od 0.05, ne odbacuje se hipoteza da je učinak varijable BDP dvostruko veći u odnosu na učinak varijable indeks potrošačkih cijena na zavisnu varijablu potrošnja.

U slučaju slike 2.96, empirijski *F*-omjer jednak je 158,84, uz pripadajuću *p*-vrijednost ≈ 0, što je manje od 0,05, odbacuje se hipoteza da je granična sklonost potrošnji jednaka 70% i indeks potrošačkih cijena suvišan u modelu.

k) Provedite parcijalni *F*-test o uključenoj nepotrebnoj varijabli za varijablu BDP u modelu m1, uz razinu značajnosti od 5%.

Nulta hipoteza ovoga testa je: H_0: varijabla BDP je suvišna u modelu m1. Test veličina računa se kao: $F = \frac{(SSR - SSR_*)/J}{SSR_*/(N-k-1)} = \frac{(5{,}12\cdot 10^{13} - 1{,}6\cdot 10^{11})/1}{1{,}6\cdot 10^{11}/137} = 43{,}711$, uz *p*-vrijednost ≈ 0. Na razini značajnosti od 5%, odbacujemo hipotezu da je varijabla BDP suvišna u modelu m1.

```
ogranicenje<-c("bdp=0")
linearHypothesis(m1,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## bdp = 0
##
## Model 1: restricted model
## Model 2: potrosnja ~ bdp + cijene
##
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1    138 5.1264e+13
## 2    137 1.6017e+11  1 5.1104e+13 43711 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 2.99. Test o uključenoj nepotrebnoj varijabli

l) Provedite LR test o izostavljenoj značajnoj varijabli za varijablu indeks potrošačkih cijena u modelu m1, uz razinu značajnosti od 5%.

```
library(lmtest)
model2<-lm(potrosnja~bdp,data=potrosnja)
lrtest(m1,model2)

## Likelihood ratio test
##
## Model 1: potrosnja ~ bdp + cijene
## Model 2: potrosnja ~ bdp
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -1658.7
## 2   3 -1658.7 -1 0.0121     0.9125
```

Slika 2.100. Test o izostavljenoj značajnoj varijabli

Nulta hipoteza ovoga testa je: H_0: varijabla indeks potrošačkih cijena je neznačajna u modelu. Test veličina računa se kao[40] $LR = -2(-1658{,}707-(-1658{,}701)) = 0{,}0121$, uz p-vrijednost = 0,9125. Na razini značajnosti od 5%, ne odbacujemo hipotezu da je varijabla indeks potrošačkih cijena neznačajna u modelu.

m) Provedite test o stabilnosti parametara u modelu m1 tako da uzorak podijelite na dva jednaka dijela. Razina značajnosti je 5%.

Kako je ukupno 140 opažanja za sve varijable, prvi poduzorak je podijeljen tako da su uključene vrijednosti do 70. opažanja, a drugi poduzorak kreće sa 71. državom po redu (slika 2.98.). Testira se sljedeća hipoteza: H_0: $\alpha_0 = \beta_0$, $\alpha_1 = \beta_1$, $\alpha_2 = \beta_2$, za modele m11 (procjena modela m1 za prvi poduzorak) i m12 (procjena modela m2 za drugi poduzorak):

$$m1: y_{i1} = \alpha_0 + \alpha_1 x_{i11} + \alpha_2 x_{i12} + \varepsilon_i$$
$$m2: y_{i1} = \beta_0 + \beta_1 x_{i21} + \beta_2 x_{i22} + \varepsilon_i .$$

[40] U ispisu su vrijednosti zaokružene na jednu decimalu, no naredbom View(lrtest(m1,model2)) dobivaju se vrijednosti na tri decimale.

Test veličina (ispis na slici 2.101.) iznosi 0,35, s pripadajućom *p*-vrijednosti 0,71, što je veće od razine značajnosti 0,05. Uz razinu značajnosti od 5%, ne odbacujemo hipotezu da su korespondentni parametri u modelima m11 i m12 jednaki.

```
bdp<-potrosnja$bdp;cijene<-potrosnja$cijene
y1<-y[1:70];bdp1<-bdp[1:70];cijene1<-cijene[1:70]
y2<-y[71:140];bdp2<-bdp[71:140];cijene2<-cijene[71:140]
library(gap)
chow.test(y1,c(bdp1,cijene1),y2,c(bdp2,cijene2))

##     F value       d.f.1       d.f.2     P value
##   0.3498094   2.0000000 276.0000000   0.7051344
```

Slika 2.101. Test o stabilnosti parametara

n) Za model m1 provedite RESET test o adekvatnosti linearnog modela, tako da uključite kvadrat i kub procijenjenih vrijednosti zavisne varijable u originalnom modelu. Razina značajnosti je 5%.

```
library(lmtest)
resettest(m1,power=2:3,type="fitted")

##
##  RESET test
##
## data:  m1
## RESET = 218.19, df1 = 2, df2 = 135, p-value < 2.2e-16
```

Slika 2.102. RESET test o adekvatnosti linearnog modela

Pomoćna jednadžba nad kojom se provodi test je sljedeća za m1:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \varepsilon_i ,$$

gdje se testira nulta hipoteza: H_0: $\gamma_1 = \gamma_2 = 0$. Temeljem ispisa na slici 2.102, uočava se da je empirijska hi-kvadrat veličina jednaka 218,19, s pripadajućom *p*-vrijednosti ≈ 0, što je manje od 0,05. Stoga se uz razinu značajnosti od 5% odbacuje nulta hipoteza da su varijable $\hat{y}_i^2$ i $\hat{y}_i^3$ neznačajne u modelu. Postoje određene nelinearnosti u ponašanju zavisne varijable.

o) S obzirom na ishod testa u prethodnome postupku, procijenite novi model tako da uključite kao nezavisne varijable uključite kvadrat i kub procijenjenih vrijednosti zavisne varijable iz modela m1. Jesu li novo dodane varijable značajne u modelu?

U novome modelu dodani su kvadrat i kub procijenjenih vrijednosti varijable potrošnja iz modela m1 (naredba fitted(m1) koristi se za procijenjene vrijednosti zavisne varijable nekog modela), pri čemu je ispis modela prikazan na slici 2.103. Uočava se kako su varijable $\hat{y}_i^2$ i $\hat{y}_i^3$ značajne u modelu, jer su pripadajući *t*-omjeri jednaki 3,491 i –5,212, odnosno, odgovarajuće *p*-vrijednosti su veoma male, 0,0007 i $6{,}82 \cdot 10^{-7}$, što je manje od uobičajenih razina značajnosti i upućuje na zaključke o odbacivanju hipotezi da varijable $\hat{y}_i^2$ i $\hat{y}_i^3$ nisu značajne u modelu.

```
summary(lm(potrosnja~bdp+cijene+I(fitted(m1)^2)+I(fitted(m1)^3),data=potrosnja))

##
## Call:
## lm(formula = potrosnja ~ bdp + cijene + I(fitted(m1)^2) + I(fitted(m1)^3),
##     data = potrosnja)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -133688   -1404   -1305    -920  125675
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.265e+02  3.100e+03   0.267  0.79015
## bdp              6.788e-01  1.645e-02  41.263  < 2e-16 ***
## cijene           4.492e+00  1.893e+01   0.237  0.81281
## I(fitted(m1)^2)  5.908e-08  1.692e-08   3.491  0.00065 ***
## I(fitted(m1)^3) -1.034e-14  1.983e-15  -5.212 6.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16740 on 135 degrees of freedom
## Multiple R-squared:  0.9993, Adjusted R-squared:  0.9992
## F-statistic: 4.569e+04 on 4 and 135 DF,  p-value: < 2.2e-16
```

Slika 2.103. Ispis modela s uključenim kvadratom i kubom procijenjene vrijednosti varijable potrošnja

p) Koliko iznosi predviđena vrijednost potrošnje u modelu m1 ako se pretpostavlja vrijednost BDP-a 1000 (mlrd LCU) i indeks cijena 134 (bodova)? Koliko iznosi interval predviđanja uz 1–γ=0,90? Interpretirajte dobivene rezultate.

Uočava se (slika 2.104.) da za vrijednost BDP = 1000 i indeksa potrošačkih cijena = 134 vrijedi: $\hat{y}_f$ = 5159,89 što znači da se ovim modelom predviđa da za vrijednost BDP-a od 1000 milijardi LCU i vrijednost indeksa potrošačkih cijena u iznosu 134 bodova, očekivana razina potrošnje iznosi u prosjeku 5.159,89 milijardi LCU. U 90% slučajeva za pretpostavljenu vrijednost varijable BDP = 1000 i indeksa potrošačkih cijena = 134, stvarna vrijednost potrošnje će iznositi između 287,87 i 10031,90 milijardi LCU.

```
novo<-data.frame(bdp=1000,cijene=134)
predict(m1,newdata = novo,interval = "confidence",level=.90)

##        fit      lwr     upr
## 1 5159.886 287.8734 10031.9
```

Slika 2.104. Predviđena vrijednost potrošnje u modelu m1

2.2.10. Pitanja za ponavljanje

1) Što je to model višestruke linearne regresije?
2) Koje su pretpostavke modela višestruke linearne regresije? Kako se razlikuju u odnosu na pretpostavke modela jednostavne linearne regresije?
3) Zapišite procjenitelj metodom najmanjih kvadrata za model višestruke linearne regresije.
4) Matrično zapišite pretpostavke modela višestruke linearne regresije.
5) Iskažite Gauss-Markovljeve uvjete za slučaj modela višestruke linearne regresije.
6) Koja su svojstva procjenitelja metodom najmanjih kvadrata za model višestruke linearne regresije?
7) Za sljedeće procijenjene modele, interpretirajte procijenjene parametre:

   a. $\hat{y}_i = 2 + 3x_{1i} - 4x_{2i}$
   b. $\hat{y}_i = 2 + 3\ln x_{1i} - 4\ln x_{2i}$
   c. $\widehat{\ln y}_i = 2 + 3x_{i1} - 4x_{i2}$
   d. $\widehat{\ln y}_i = 2 + 3\, lnx_{i1} - 4lnx_{i2}$
   e. $\hat{y}_i = 2 + 3\ln x_{1i} - 4x_{2i}$
   f. $\widehat{\ln y}_i = 2 + 3x_{i1} - 4lnx_{i2}$

8) Što je to intervalna procjena parametara?
9) Interpretirajte sljedeću intervalnu procjenu parametra uz varijablu x_1, ako se razmatra ovisnost potrošnje (y) u kn o dohotku (x_1) i razini cijena (x_2): $\hat{y}_i = 2 + 3x_{1i} - 4x_{2i}$, $P(1 < \beta_1 < 4) = 0{,}99$.
10) Iz prethodnog zadatka interpretirajte sljedeće: $P(1 < \beta_0 < 3) = 0{,}95$.
11) Što je to model sa standardiziranim regresijskim varijablama? Kako se vrši interpretacija procijenjenih parametara? Kada se koristi ovaj model?
12) Opišite popunjavanje tablice ANOVA za slučaj višestruke linearne regresije.
13) Kako se interpretira procijenjena standardna devijacija regresije? Koji je njen nedostatak?
14) Koja relativna mjera reprezentativnosti modela se koristi uz procjenu standardne devijacije regresije i kako ju interpretiramo?
15) Što je koeficijent determinacije regresije? Kako ga interpretiramo?
16) Što je koeficijent višestruke linearne korelacije? Kako ga interpretiramo? Koja je razlika između ovog koeficijenta i koeficijenta jednostavne linearne korelacije?
17) Kada se koristi korigirani koeficijent determinacije?
18) Što je to *t*-test i kako se provodi? Ukratko opišite postupak.
19) Što je to *F*-test i kako se provodi? Ukratko opišite postupak.
20) Čemu služi Waldov test?
21) Što je to parcijalni *F*-test i po čemu se razlikuje od *F*-testa skupne značajnosti?
22) Koji problem se javlja ako govorimo o izostavljenoj značajnoj regresijskoj varijabli? Kako se provode odgovarajući testovi?
23) Opišite test o stabilnosti parametara. Što se njime testira?
24) Opišite RESET test. Što se njime testira?
25) Opišite CUSUM test. Što se njime testira?

26) Dan je model: $y_i = 3 + 2x_{i1} + 5x_{i2} - 4x_{i3} + \varepsilon_i$. Zapišite hipoteze Waldova testa u matričnom zapisu za sljedeće nulte hipoteze:
   a) Varijabla $x_{i,2}$ nije značajna u modelu.
   b) Varijable $x_{i,2}$ i $x_{i,3}$ nisu značajne u modelu.
   c) Utjecaj varijable $x_{i,1}$ je trostruko veći od utjecaja varijable $x_{i,3}$ na zavisnu varijablu.
   d) Prosječni utjecaj svih triju varijabli iznosi 3.
   e) Utjecaj varijable $x_{i,2}$ iznosi četvrtinu utjecaja varijable $x_{i,1}$ na zavisnu varijablu.
   f) Utjecaji varijable $x_{i,2}$ i $x_{i,3}$ se međusobno poništavaju.

g) Utjecaji varijable $x_{i,2}$ jednak je utjecaju varijable $x_{i,3}$.

27) Učitajte datoteku „**placa.txt**“ u RStudio. Datoteka sadrži podatke o 150 pojedinaca: iznos plaće (u kn), broj godina radnog staža (staz), te broj godina školovanja (obrazovanje). Provedite postupke a) – p) uz sve potrebne interpretacije (vidjeti primjer u naslovu 2.2.9).

a) Procijenite model u kojemu plaća ovisi o broju godina radnog staža i broju godina školovanja kao lin-lin model (m1), te kao log-log model (m2) i konačno kao model sa standardiziranim varijablama (m3). Pomoću naredbe stargazer spojite rezultate procjena u jednu tablicu. Interpretirajte procijenjene parametre uz nezavisne varijable u sva tri modela.
b) Usporedite modele m1 i m2 temeljem rezultata ispisa u postupku a) koji je reprezentativniji i objasnite zašto. Interpretirajte odgovarajuće mjere koje uspoređujete.
c) Provedite jednosmjerni *t*-test ($\alpha = 5\%$) za onaj model koji je bolji iz postupka b).
d) Provedite skupni test značajnosti za onaj model koji je bolji iz postupka b).
e) Procijenite dodatna dva modela: log-lin (m4) i lin-log (m5) između odabranog skupa varijabli, te spojite pomoću naredbe stargazer modele m1, m2, m4 i m5. Interpretirajte procijenjene parametre uz nezavisne varijable za modele m4 i m5.
f) Za najbolji model u postupku e) (pojasnite najprije zašto je najbolji) odredite i interpretirajte intervalne procjene parametara na razini pouzdanosti od 90%.
g) Izračunajte koeficijent determinacije, korigirani koeficijent determinacije, procjenu standardne devijacije regresije, procjenu koeficijenta varijacije regresije te koeficijent višestruke linearne korelacije i potom ih interpretirajte za model koji je bio najbolji u postupku e.
h) Za najbolji model u postupku 3, provedite Waldov test je li učinak povećanja radnog staža za jednu godinu za 500 veći od učinka povećanja jedne godine obrazovanja na plaću. Zapišite matrično nultu hipotezu testa, te donesite odluku uz razinu značajnosti od 5%.
i) Za najbolji model u postupku e, provedite Waldov test jesu li jednaki učinci nezavisnih varijabli na zavisnu. Zapišite matrično nultu hipotezu testa, te donesite odluku uz razinu značajnosti od 5%.
j) Provedite odgovarajuće *F*-testove iz postupaka h) te i).
k) Provedite parcijalni *F*-test o uključenoj nepotrebnoj varijabli za varijablu obrazovanje u modelu m1, uz razinu značajnosti od 5%.
l) Provedite LR test o izostavljenoj značajnoj varijabli za varijablu staž u modelu m1, uz razinu značajnosti od 5%.
m) Provedite test o stabilnosti parametara u modelu m1 tako da uzorak podijelite na dva jednaka dijela. Razina značajnosti je 5%.
n) Za model m1 provedite RESET test o adekvatnosti linearnog modela, tako da uključite kvadrat, kub i četvrtu potenciju procijenjenih vrijednosti zavisne varijable u originalnom modelu. Razina značajnosti je 5%.
o) Za model m1 provedite CUSUM test o stabilnosti parametara. Dodatno, provedite odgovarajući *F*-test gdje se sumnja da dolazi do promjene parametara između 80. i 120. opažanja. Razina značajnosti je 5%.
p) Koliko iznosi predviđena vrijednost plaće u modelu m1 ako se pretpostavlja vrijednost staža 19 godina i obrazovanja 15 godina? Koliko iznosi interval predviđanja uz $1-\gamma=0{,}99$? Interpretirajte dobivene rezultate.

Rješenja

Zadatak 7):

a) 2... kada bi vrijednosti obje nezavisne varijable iznosile 0, vrijednost zavisne varijable u prosjeku iznosi oko 2 jedinice
3... ako se vrijednost prve nezavisne varijable poveća za jednu jedinicu, uz nepromijenjenu vrijednost druge varijable, vrijednost zavisne varijable će se povećati u prosjeku za 3 jedinice
–4... ako se vrijednost druge nezavisne varijable poveća za jednu jedinicu, uz nepromijenjenu vrijednost prve varijable, vrijednost zavisne varijable će se smanjiti u prosjeku za 4 jedinice

b) 2... kada bi vrijednosti obje nezavisne varijable iznosile 1, vrijednost zavisne varijable u prosjeku iznosi oko 2 jedinice
3... ako se vrijednost prve nezavisne varijable poveća za 1%, uz nepromijenjenu vrijednost druge varijable, vrijednost zavisne varijable će se povećati u prosjeku za 0,03 jedinice
–4... ako se vrijednost druge nezavisne varijable poveća za 1%, uz nepromijenjenu vrijednost prve varijable, vrijednost zavisne varijable će se smanjiti u prosjeku za 0,04 jedinice

c) 2... kada bi vrijednosti obje nezavisne varijable iznosile 0, vrijednost zavisne varijable u prosjeku iznosi oko e^2 jedinica
3... ako se vrijednost prve nezavisne varijable poveća za jednu jedinicu, uz nepromijenjenu vrijednost druge varijable, vrijednost zavisne varijable će se povećati u prosjeku za 300%
–4... ako se vrijednost druge nezavisne varijable poveća za jednu jedinicu, uz nepromijenjenu vrijednost prve varijable, vrijednost zavisne varijable će se smanjiti u prosjeku za 400%

d) 2... kada bi vrijednosti obje nezavisne varijable iznosile 1, vrijednost zavisne varijable u prosjeku iznosi oko e^2 jedinica
3... ako se vrijednost prve nezavisne varijable poveća za 1%, uz nepromijenjenu vrijednost druge varijable, vrijednost zavisne varijable će se povećati u prosjeku za 3%
–4... ako se vrijednost druge nezavisne varijable poveća za 1%, uz nepromijenjenu vrijednost prve varijable, vrijednost zavisne varijable će se smanjiti u prosjeku za 4%

e) 2... kada bi vrijednosti prve nezavisne varijable iznosila 1, a druge 0, vrijednost zavisne varijable u prosjeku iznosi oko 2 jedinice
3... ako se vrijednost prve nezavisne varijable poveća za 1%, uz nepromijenjenu vrijednost druge varijable, vrijednost zavisne varijable će se povećati u prosjeku za 0,03 jedinice
–4... ako se vrijednost druge nezavisne varijable poveća za jednu jedinicu, uz nepromijenjenu vrijednost prve varijable, vrijednost zavisne varijable će se smanjiti u prosjeku za 4 jedinice

f) 2... kada bi vrijednosti prve nezavisne varijable iznosila 0, a druge 1, vrijednost zavisne varijable u prosjeku iznosi oko 2 jedinice
3... ako se vrijednost prve nezavisne varijable poveća za jednu jedinicu, uz nepromijenjenu vrijednost druge varijable, vrijednost zavisne varijable će se povećati u prosjeku za 300%
–4... ako se vrijednost druge nezavisne varijable poveća za 1%, uz nepromijenjenu vrijednost prve varijable, vrijednost zavisne varijable će se smanjiti u prosjeku za 4%

Zadatak 9):

Uz razinu pouzdanosti od 99%, ako se dohodak poveća za 1 kn, uz nepromijenjenu razinu cijena, stvarna potrošnja će se povećati u prosjeku između 1kn i 4 kn.

Zadatak 10):

Uz razinu pouzdanosti od 95%, kada bi dohodak iznosio 0 kn, te kad bi sve bilo besplatno (razina cijena jednaka 0), stvarna prosječna potrošnja bi iznosila između 1 kn i 3 kn.

Zadatak 26):

a)
$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}$$

b)
$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

c)
$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 1 & 0 & -3 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}$$

d)
$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 9 \end{bmatrix}$$

e)
$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 0,25 & -1 & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}$$

f)
$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}$$

$$H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \boldsymbol{q} = \begin{bmatrix} 0 \end{bmatrix}$$

g)

Zadatak 27):

```
## ================================================================================================
##                                            Dependent variable:
##                     ---------------------------------------------------------------------------
##                              placa                  log(placa)             scale(placa)
##                               (1)                      (2)                     (3)
## ------------------------------------------------------------------------------------------------
## staz                        1.390***
##                             (0.233)
##
## obrazovanje                  0.015
##                             (0.239)
##
## log(staz)                                             0.002***
##                                                       (0.0004)
##
## log(obrazovanje)                                      0.002***
##                                                       (0.0004)
##
## scale(staz)                                                                  0.938***
##                                                                              (0.157)
##
## scale(obrazovanje)                                                           0.010
##                                                                              (0.157)
##
## Constant                  4,008.723***               8.293***
##                             (0.995)                  (0.0005)
##
## ------------------------------------------------------------------------------------------------
## Observations                 150                      150                     150
## R2                          0.899                    0.791                   0.899
## Adjusted R2                 0.898                    0.789                   0.898
## Residual Std. Error     5.706 (df = 147)         0.002 (df = 147)         0.318 (df = 148)
## F Statistic         657.821*** (df = 2; 147) 278.873*** (df = 2; 147) 662.296*** (df = 2; 148)
## ================================================================================================
## Note:                                                                *p<0.1; **p<0.05; ***p<0.01
```

```
summary(m1)
```

```
## Call:
## lm(formula = placa ~ staz + obrazovanje, data = place)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6656  -4.0054   0.0725   4.0248  10.7120
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 4.009e+03  9.954e-01 4027.074  < 2e-16 ***
## staz        1.390e+00  2.330e-01    5.967 1.73e-08 ***
## obrazovanje 1.538e-02  2.394e-01    0.064    0.949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.706 on 147 degrees of freedom
## Multiple R-squared:  0.8995, Adjusted R-squared:  0.8981
## F-statistic: 657.8 on 2 and 147 DF,  p-value: < 2.2e-16
```

```
cbind(qt(0.95,147), qf(0.95,2,147))
```

```
##          [,1]     [,2]
## [1,] 1.655285 3.057621
```

```
m4<-lm(log(placa)~staz+obrazovanje,data=place)
m5<-lm(placa~log(staz)+log(obrazovanje),data=place)
stargazer(list(m1,m2,m4,m5),type="text")

## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------------
##                                 placa          log(placa)          placa
##                                  (1)         (2)        (3)         (4)
## --------------------------------------------------------------------------------
## staz                            1.390***              0.0003***
##                                 (0.233)                (0.0001)
##
## obrazovanje                      0.015                 0.00000
##                                 (0.239)                (0.0001)
##
## log(staz)                                    0.002***              7.780***
##                                              (0.0004)              (1.636)
##
## log(obrazovanje)                             0.002***              8.353***
##                                              (0.0004)              (1.614)
##
## Constant                      4,008.723***  8.293***   8.296***  3,995.443***
##                                 (0.995)     (0.0005)   (0.0002)    (1.954)
##
## --------------------------------------------------------------------------------
## Observations                      150         150        150         150
## R2                               0.899       0.791      0.899       0.791
## Adjusted R2                      0.898       0.789      0.898       0.788
## Residual Std. Error (df = 147)   5.706       0.002      0.001       8.237
## F Statistic (df = 2; 147)       657.821***  278.873*** 657.624***  277.371***
## ================================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

Napomena: kako su na slici iznad koeficijenti determinacije za m1 i m4 jednaki, možete odabrati bilo koji od ta dva. U nastavku je dan ispis za m1.

```
confint(m1,level=.9)

##                       5 %         95 %
## (Intercept) 4007.0755655 4010.3710507
## staz           1.0046636    1.7760848
## obrazovanje   -0.3809056    0.4116561

summary(m1)

##
## Call:
## lm(formula = placa ~ staz + obrazovanje, data = place)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6656  -4.0054   0.0725   4.0248  10.7120
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 4.009e+03  9.954e-01 4027.074  < 2e-16 ***
## staz        1.390e+00  2.330e-01    5.967 1.73e-08 ***
## obrazovanje 1.538e-02  2.394e-01    0.064    0.949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.706 on 147 degrees of freedom
## Multiple R-squared:  0.8995, Adjusted R-squared:  0.8981
## F-statistic: 657.8 on 2 and 147 DF,  p-value: < 2.2e-16
```

```
cbind(qt(0.95,147), qf(0.95,2,147))

##          [,1]     [,2]
## [1,] 1.655285 3.057621

y<-place$placa
sazetak<-summary(m1)
cbind(sazetak$sigma,sazetak$sigma/mean(y))

##          [,1]        [,2]
## [1,] 5.705662 0.001412991
```

```
ogranicenje<-"staz=500+obrazovanje"
linearHypothesis(m1,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## staz - obrazovanje = 500
##
## Model 1: restricted model
## Model 2: placa ~ staz + obrazovanje
##
##   Res.Df      RSS Df Sum of Sq   Chisq Pr(>Chisq)
## 1    148 36525046
## 2    147     4786  1  36520260 1121816  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ogranicenje<-"staz=obrazovanje"
linearHypothesis(m1,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## staz - obrazovanje = 0
##
## Model 1: restricted model
## Model 2: placa ~ staz + obrazovanje
##
##   Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1    148 5063.2
## 2    147 4785.5  1    277.71 8.5306   0.003492 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ogranicenje<-"staz=500+obrazovanje"
linearHypothesis(m1,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## staz - obrazovanje = 500
##
## Model 1: restricted model
## Model 2: placa ~ staz + obrazovanje
##
##   Res.Df      RSS Df Sum of Sq       F    Pr(>F)
## 1    148 36525046
## 2    147     4786  1  36520260 1121816 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ogranicenje<-"staz=obrazovanje"
linearHypothesis(m1,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## staz - obrazovanje = 0
##
## Model 1: restricted model
## Model 2: placa ~ staz + obrazovanje
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    148 5063.2
## 2    147 4785.5  1    277.71 8.5306 0.004043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

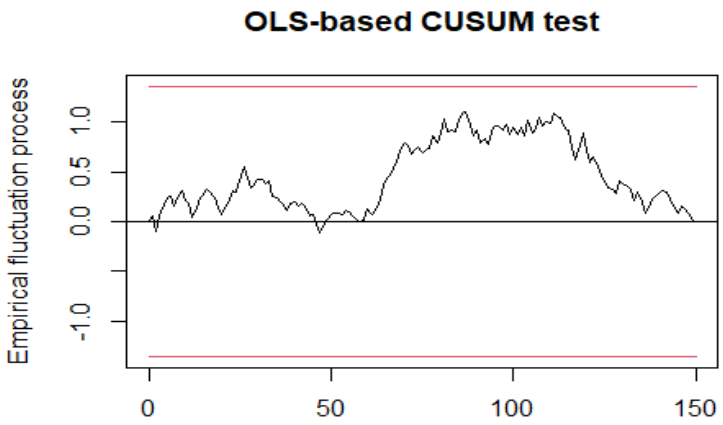
```
ogranicenje<-c("obrazovanje=0")
linearHypothesis(m1,ogranicenje,test="F")

## Linear hypothesis test
##
## Hypothesis:
## obrazovanje = 0
##
## Model 1: restricted model
## Model 2: placa ~ staz + obrazovanje
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    148 4785.7
## 2    147 4785.5  1   0.13428 0.0041 0.9489
```

```
library(lmtest)
model2<-lm(placa~obrazovanje,data=place)
lrtest(m1,model2)

## Likelihood ratio test
##
## Model 1: placa ~ staz + obrazovanje
## Model 2: placa ~ obrazovanje
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -472.54
## 2   3 -488.81 -1 32.532  1.172e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
placa<-place$placa;staz<-place$staz;obraz<-place$obrazovanje
placa1<-y[1:75];staz1<-staz[1:75];obraz1<-obraz[1:75]
placa2<-y[76:150];staz2<-staz[76:150];obraz2<-obraz[76:150]
library(gap)
chow.test(placa,c(obraz1,staz1),placa2,c(obraz2,staz2))

##      F value        d.f.1        d.f.2      P value
## 2.252555e+01 2.000000e+00 2.960000e+02 7.832376e-10
```

```
library(lmtest)
resettest(m1,power=2:4,type="fitted")

##
##  RESET test
##
```

```
## data:  m1
## RESET = 0.60545, df1 = 3, df2 = 144, p-value = 0.6125

library(strucchange)
placa<-ts(place$placa,start=1,frequency = 1)
staz<-ts(place$staz,start=1,frequency = 1)
obrazovanje<-ts(place$obrazovanje,start=1,frequency = 1)

cusum <- efp(placa~staz+obrazovanje, type = "OLS-CUSUM")
plot(cusum,xlab=NA)
```




```
fs <- Fstats(placa~staz+obrazovanje, from = 80, to = 120)
sctest(fs, type="aveF")

##
##  aveF test
##
## data:  fs
## ave.F = 5.9081, p-value = 0.08881

novo<-data.frame(staz=19,obrazovanje=15)
predict(m1,newdata = novo,interval = "confidence",level=.99)

##        fit      lwr     upr
## 1 4035.371 4033.712 4037.03
```

2.3. Asimptotska svojstva procjenitelja linearne regresije

U odjeljku 2.1.3.3. razmatrana su svojstva procjenitelja dobivenih metodom najmanjih kvadrata, nepristranost i efikasnost. Radilo se o svojstvima malih uzoraka. S druge strane, **asimptotska svojstva procjenitelja** su ona koja se odnose na velike uzorke. U asimptotska svojstva ubrajamo **konzistentnost** i **asimptotsku normalnost** (Greene, 2018; Hill et al., 2017).

Neka je $\hat{\theta}_N$ procjenitelj od θ temeljem uzorka veličine N. Kažemo da je $\hat{\theta}_N$ **konzistentan procjenitelj** ako za svaki $e > 0$ vrijedi:

$$P(|\hat{\theta}_N - \theta| < e) \to 1 \text{ za } N \to \infty. \tag{2.256}$$

Dakle, vjerojatnost da je udaljenost između $\hat{\theta}_N$ i θ manja od proizvoljno malog pozitivnog broja e teži prema 1 kada veličina uzorka neizmjerno raste, odnosno distribucija procjenitelja $\hat{\theta}_N$ se koncentrira oko vrijednosti θ pa se može reći da je za sve veće uzorke vjerojatnost da je $\hat{\theta}_N$ daleko od θ sve manja. Za konzistentne procjenitelje možemo pisati i:

$$\lim_{N\to\infty} B(\hat{\theta}_N) = \lim_{N\to\infty} E(\hat{\theta}_N - \theta) = 0, \tag{2.257}$$

što znači da pristranost B procjenitelja teži prema 0 za sve veći N, i to nazivamo **asimptotski nepristranim** procjeniteljem.

Asimptotska normalnost procjenitelja važna je za intervalne procjene. Neka je $\{Z_N : N \in \{1, 2, \dots\}\}$ niz slučajnih varijabli, takvih da za svaki broj z vrijedi:

$$P(Z_N \le z) \to \Phi(z) \text{ za } N \to \infty, \tag{2.258}$$

gdje je $\Phi(z)$ standardna normalna kumulativna funkcija distribucije. Tada kažemo da Z_N asimptotski slijedi standardnu normalnu distribuciju, što možemo pisati: $Z_N \overset{a}{\sim} N(0,1)$, gdje se „*a*“ odnosi na asimptotski. Interpretacija izraza (2.258) znači da se kumulativna funkcija distribucije varijable Z_N približava normalnoj distribuciji kada se N povećava, to jest neizmjerno raste.

Jedan od najsnažnijih rezultata u vjerojatnosti je **centralni granični teorem** (engl. CLT – *central limit theorem*): Neka je $\{y_1, y_2, \dots, y_N\}$ slučajni uzorak s očekivanjem μ i varijancom σ^2. Tada varijabla

$$Z_N = \frac{\overline{Y}_N - \mu}{\sigma / \sqrt{N}}, \tag{2.259}$$

ima asimptotski standardnu normalnu distribuciju. Pobliže o ovom jednom od najznačajnijih teorema teorije vjerojatnosti može se vidjeti, primjerice, u Sarapa (2002) ili White (2000).

3.

DALJNJA ANALIZA REGRESIJSKOG MODELA

3. DALJNJA ANALIZA REGRESIJSKOG MODELA

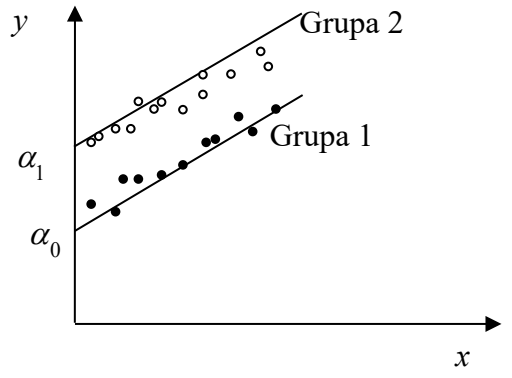
3.1. Kvalitativne regresorske varijable

Kvalitativne, indikatorske ili binarne varijable (engl. *binary, dummy*) koriste se kako bi se u ekonometrijske modele uključile kvalitativne informacije o različitim pojavama. U ekonomiji se najčešće primjenjuju kod vremenskih nizova za sezonske utjecaje, a koriste se kako bi i prikazale utjecaj određenih jednokratnih ili pak trajnih[41] šokova, koji se inače ne mogu drugačije inkorporirati u model osim na ovaj način, i slično. Također, kod presječnih podataka, kvalitativne varijable mogu opisivati prisutnost ili odsutnost neke karakteristike (spol, određena dobna skupina, platežna moć, itd.).

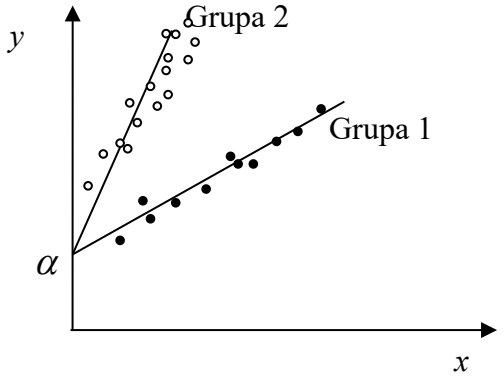
Kvalitativne varijable mogu se odnositi posebno **na konstantni član u modelu**, posebno za **promjene parametra** uz nezavisnu varijablu, kao i njihovom **kombinacijom**. Ako se razmatra model u kojemu se konstantni član mijenja s obzirom na binarnu varijablu, tada dolazi do promjene u **prosječnoj razini** zavisne varijable. Ako se radi o modelu u kojemu dolazi do promjene parametra uz nezavisnu varijablu, tada govorimo da dolazi do **promjene u učinku nezavisne varijable *x* na zavisnu varijablu *y*** u ovisnosti o binarnoj varijabli. Konačno, uključivanjem binarne varijable i za konstantni član i za nagib (parametar uz varijablu x), radi se o promjeni i prosječne razine zavisne varijable kao i promjene učinka nezavisne na zavisnu varijablu.

Slika 3.1 prikazuje jednostavan slučaj jedne zavisne i jedne nezavisne varijable kada se u (*a*) analizu uključuje kvalitativna varijabla za konstantni član te (*b*) za promjene u nagibu. Dakle, razmatra se učinak varijable x na varijablu y, pri čemu na slici (a) uočavamo da je odsječak na y-osi za Grupu 1 α_0, dok je za Grupu 2 odsječak jednak $\alpha_1 > \alpha_0$. S druge strane, na slici (b) uočavamo da obje grupe imaju jednak odsječak na y-osi, koji iznosi α, dok je učinak promjene varijable x za Grupu 1 manji u odnosu na Grupu 2, jer se nagibi pravaca razlikuju, pri čemu je veći nagib za Grupu 2.

(*a*) kvalitativna varijabla za konstantni član



(*b*) kvalitativna varijabla za promjene u nagibu



Slika 3.1. Regresijski pravci za model jedne zavisne i jedne nezavisne varijable

U slučaju (*a*) radi se o sljedećem modelu:

[41] Primjer jednokratnog šoka može biti iznenadan skok vrijednosti varijable koji se dogodio u samo jednome razdoblju, dok primjer trajnog šoka može biti ulazak u Europsku uniju.

$$y = \begin{cases} \alpha_1 + \beta x + \varepsilon, \text{ za grupu 1} \\ \alpha_2 + \beta x + \varepsilon, \text{ za grupu 2} \end{cases} \tag{3.1}$$

Umjesto (3.1), model koji uključuje kvalitativnu varijablu možemo pisati na način:

$$y = \alpha_1 + (\alpha_2 - \alpha_1)D + \beta x + \varepsilon, \text{ pri čemu je } D = \begin{cases} 0 \text{ za grupu 1} \\ 1 \text{ za grupu 2} \end{cases}. \tag{3.2}$$

Naime, ako je $D = 0$, radi se o modelu: $y = \alpha_1 + (\alpha_2 - \alpha_1)\cdot 0 + \beta x + \varepsilon = \alpha_1 + \beta x + \varepsilon$, dok za $D = 1$ o modelu: : $y = \alpha_1 + (\alpha_2 - \alpha_1)\cdot 1 + \beta x + \varepsilon = \alpha_2 + \beta x + \varepsilon$.

U slučaju (*b*) radi se o modelu:

$$y = \begin{cases} \alpha + \beta_1 x + \varepsilon, \text{ za grupu 1} \\ \alpha + \beta_2 x + \varepsilon, \text{ za grupu 2} \end{cases} \tag{3.3}$$

Umjesto (3.3), model možemo pisati i na način:

$$y = \alpha + \beta_1 x + (\beta_2 - \beta_1)xD + \varepsilon, \text{ pri čemu je } D = \begin{cases} 0 \text{ za grupu 1} \\ 1 \text{ za grupu 2} \end{cases}. \tag{3.4}$$

Naime, ako je $D = 0$, radi se o modelu: $y = \alpha + \beta_1 x + (\beta_2 - \beta_1)x\cdot 0 + \varepsilon = \alpha + \beta_1 x + \varepsilon$, dok za $D = 1$ o modelu: $y = \alpha + \beta_1 x + (\beta_2 - \beta_1)x\cdot 1 + \varepsilon = \alpha + \beta_2 x + \varepsilon$. Važno je uočiti da razlika $\beta_2 - \beta_1$ predstavlja razliku u koeficijentu smjera za grupu 1 i grupu 2.

Osim ovog najjednostavnijeg primjera, kvalitativne varijable koriste se i u vremenskim nizovima za uključivanje sezonskih utjecaja ili pak jednokratnih ili trajnih šokova. Postoji širok raspon primjene ovih varijabli.

Model u kojemu su uključene binarne varijable i za konstantni član i za koeficijent uz nezavisnu varijablu najjednostavnije se može zapisati na sljedeći način:

$$y = \beta_0 + \beta_1 x + \beta_2 D + \beta_3 xD + \varepsilon, \text{ pri čemu je } D = \begin{cases} 0 \text{ za grupu 1} \\ 1 \text{ za grupu 2} \end{cases}. \tag{3.5}$$

Naime, ako je $D = 0$, radi se o modelu: $y = \beta_0 + \beta_1 x + \beta_2\cdot 0 + \beta_3 x\cdot 0 + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$, dok se za slučaj $D = 1$ radi o modelu: $y = \beta_0 + \beta_1 x + \beta_2\cdot 1 + \beta_3 x\cdot 1 + \varepsilon = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x + \varepsilon$.

Napomena: broj binarnih varijabli u modelu uvijek je za 1 manji u odnosu na broj modaliteta koje ta varijabla može poprimiti! Razlog leži u činjenici da se u suštini varijabla D promatra kao jedna od nezavisnih varijabli, pa se prilikom formiranja matrice $\boldsymbol{X}$ mora u slučaju modela višestruke linearne regresije zadovoljiti pretpostavka 7) u naslovu 2.2.2. Interpretacija procijenjenih parametara uz binarne varijable se stoga vrši u odnosu na nedostajući modalitet. Primjerice, ako se radi o ovisnosti varijable y o binarnoj varijabli koja je 1 za grupu 1, a 0 za grupu 2, $y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$, tada se parametar uz binarnu varijablu interpretira na način da je za β_1 jedinica veća prosječna razina zavisne varijable za grupu 1, u odnosu na grupu 2.

Razmotrimo kako izgleda matrica $\boldsymbol{X}$ za slučaj modela (3.2):

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & 1 \\ 1 & x_1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_N & 1 \end{bmatrix} \tag{3.6}$$

pri čemu elementi u trećem stupcu u nekim slučajevima poprimaju vrijednost 1, dok za druge 0. Primjerice, prvo opažanje ima vrijednost 1, dok drugo 0.

Razmotrimo slučaj matrice u (3.6), kada bi imali 5 opažanja, te kada bi uključili dvije binarne varijable za slučaj odsječka na osi ordinate:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 20 & 1 & 0 \\ 1 & 22 & 1 & 0 \\ 1 & 18 & 0 & 1 \\ 1 & 25 & 0 & 1 \\ 1 & 15 & 0 & 1 \end{bmatrix}, \tag{3.7}$$

čiji rang iznosi $r(\boldsymbol{X}) = 3$, te umnožak $\boldsymbol{X}'\boldsymbol{X}$ također ima rang 3, tj. inverz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ nije moguće izračunati, a tako i procjenitelj $\hat{\boldsymbol{\beta}}$.

Ako se želi uključiti broj binarnih varijabli koliko je i modaliteta, u tom slučaju je potrebno isključiti konstantu (vidjeti Brooks, 2014) iz modela (3.2). Sada je model (3.2) moguće pisati na sljedeći način bez konstante:

$$y_i = \alpha_1 + \alpha_1 D_{i1} + \alpha_2 D_{i2} + \varepsilon_i, \text{ pri čemu je}$$

$$D_{1i} = \begin{cases} 0 \text{ za grupu } 1 \\ 1 \text{ za grupu } 2 \end{cases}, \; D_{2i} = \begin{cases} 1 \text{ za grupu } 1 \\ 0 \text{ za grupu } 2 \end{cases}. \tag{3.8}$$

Uočimo da je sada interpretacija modela (3.8) sljedeća: kada bi vrijednost nezavisne varijable iznosila 0, u prosjeku bi vrijednost zavisne varijable za prvu grupu iznosila α_1, dok bi za drugu grupu iznosila α_2. Dakle, za slučaj binarnih varijabli za konstantni član, sada se parametri uz binarne varijable interpretiraju kao prosječne razine pojave y (kada su sve nezavisne varijable u modelu jednake 0).

Primjer 3.1.
Učitajte datoteku „**binarne.txt**“ u RStudio. Datoteka sadrži podatke o 150 pojedinaca: iznos plaće (u kn), broj godina radnog staža (staz), broj godina školovanja (obrazovanje), te podatak radi li se o muškoj i osobi (m ili z). Generirajte binarnu varijablu koja je jednaka 1 u slučaju muškog spola, a 0 u slučaju ženskog. Procijenite model u kojemu plaća zaposlenika ovisi o binarnoj varijabli spol.

Slika 3.2. predočava naredbu kojom se temeljem učitanih podataka koji u stupcu spol sadrže kvalitativne podatke „m“ ili „z“, na način da se pomoću naredbe ifelse(…) definira nova

varijabla koja će sadržavati vrijednosti 1 ili 0. Nadalje, prikazan je procijenjeni model, koji se zapisuje na sljedeći način:

$$\hat{y}_i = 4.023,54 + 30,12D_i,\ D_i = \begin{cases} 1, & M \\ 0, & Z \end{cases}.$$

Dakle, ako se radi o ženskoj osobi, $D_i = 0$ pa je prosječna plaća žena jednaka 4023,54 kuna, dok je za muške osobe prosječna plaća jednaka 4023,54 + 30,12 = 4053,68 kuna, s obzirom da za muške osobe vrijedi $D_i = 1$. Primijetimo da je interpretacija parametara uz binarne varijable drugačija u odnosu na numeričke varijable x!

30,12 kuna je razlika između prosječne plaće muškaraca i žena. Nadalje, ako se promotri značajnost binarne varijable u modelu, uočava se da je na uobičajenim razinama značajnosti ta varijabla značajna u modelu.

```
bin<-read.table("binarne.txt",header=T,sep="\t")
binarna<-ifelse(bin$spol=="m",1,0)
summary(lm(placa~binarna,data=bin))

## Call:
## lm(formula = placa ~ binarna, data = bin)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.8617  -7.2568   0.2717   7.5248  16.7607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4023.544      1.087  3700.8   <2e-16 ***
## binarna       30.122      1.569    19.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.602 on 148 degrees of freedom
## Multiple R-squared:  0.7134, Adjusted R-squared:  0.7115
## F-statistic: 368.4 on 1 and 148 DF,  p-value: < 2.2e-16
```

Slika 3.2. Učitavanje podataka i definiranje binarne varijable te procjena modela

Primjer 3.2.
Proširite model iz prethodnog primjera u model u kojemu plaća zaposlenika ovisi o binarnoj varijabli spol, ali i o broju godina radnog staža i broju godina obrazovanja. Zapišite procijenjeni model, interpretirajte procijenjene parametre te komentirajte značajnost svih varijabli u modelu.

Sada se (slika 3.3.) uočava da je procijenjeni model:

$$\hat{y}_i = 4009,84 + 5,04D_i + 1,33x_{i1} - 0,11x_{i2},\ D_i = \begin{cases} 1, M \\ 0, Z \end{cases}.$$

Interpretacija parametara je sljedeća: 4009,84 kao konstanta nema ekonomsko značenje u modelu jer bi toliko iznosila prosječna plaća kada bi sve nezavisne varijable iznosile 0: dakle, kada bi se radilo o ženskoj osobi ($D_i = 0$), koja ima 0 godina radnog staža i 0 godina obrazovanja. Nadalje, 5,04 je razlika u prosječnoj plaći između muškaraca i žena, jer je prosječna plaća za muške osobe (kada je $D_i = 1$) za toliko kuna veća u odnosu na ženske osobe. Procijenjena vrijednost 1,33 ima sljedeću interpretaciju: ako se broj godina radnog staža poveća za 1 godinu, uz nepromijenjen broj godina obrazovanja, te **neovisno o tome radi li se o muškoj**

ili ženskoj osobi, plaća se poveća u prosjeku za 1,33 kune. Slično tome, vrijednost –0,11 se interpretira na sljedeći način: ako se broj godina obrazovanja poveća za 1 godinu, uz nepromijenjen broj godina radnog staža, te neovisno o tome radi li se o muškoj ili ženskoj osobi, plaća se smanji u prosjeku za 0,11 kuna.

Nadalje, ako se razmotri značajnost svih varijabli u modelu pojedinačno, jedino varijabla obrazovanje nije značajna pri uobičajenim razinama značajnosti.

```
summary(lm(placa~binarna+staz+obrazovanje,data=bin))

## Call:
## lm(formula = placa ~ binarna + staz + obrazovanje, data = bin)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9445  -4.0947   0.3051   4.0821  11.9770
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 4009.8377     1.0471 3829.499  < 2e-16 ***
## binarna        5.0384     1.7609    2.861  0.00484 **
## staz           1.3311     0.2285    5.826 3.48e-08 ***
## obrazovanje   -0.1084     0.2377   -0.456  0.64918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.571 on 146 degrees of freedom
## Multiple R-squared:  0.9048, Adjusted R-squared:  0.9029
## F-statistic: 462.7 on 3 and 146 DF,  p-value: < 2.2e-16
```

Slika 3.3. Procijenjeni model

Primjer 3.3.
U prethodnome primjeru procijenite ovisnost plaće pojedinaca o stažu i obrazovanju, ali pritom se pretpostavlja da spol utječe i na učinak staža na plaću i na učinak obrazovanja na plaću. Zapišite procijenjeni model, interpretirajte procijenjene parametre te komentirajte značajnost svih varijabli u modelu.

Sada je (slika 3.4.) procijenjeni model:

$$\hat{y}_i = 4009{,}99 + 1{,}45x_{i1} - 0{,}25x_{i2} - 0{,}13x_{i1}D_i + 0{,}32x_{i2}D_i,\ D_i = \begin{cases} 1, M \\ 0, Z \end{cases}.$$

Dakle, ako se radi o ženskoj osobi, model glasi: $\hat{y}_i = 4009{,}99 + 1{,}45x_{i1} - 0{,}25x_{i2}$, dok za mušku osobu model glasi: $\hat{y}_i = 4009{,}99 + 1{,}45x_{i1} - 0{,}25x_{i2} - 0{,}13x_{i1} + 0{,}32x_{i2} = 4009{,}99 + 1{,}32x_{i1} + 0.07x_{i2}$.

Interpretacije su sljedeće: za žensku osobu koja ima 0 godina radnog staža i 0 godina obrazovanja, prosječna plaća iznosi 4099,99 kuna. Povećanje godina radnog staža za 1 godinu, uz nepromijenjen broj godina obrazovanja za ženske osobe povećava plaću u prosjeku za 1,45 kuna, dok povećanje broja godina obrazovanja za 1 godinu, uz nepromijenjen broj godina radnog staža za ženske osobe smanjuje plaću u prosjeku za 0,25 kuna.

```
summary(lm(placa~staz+obrazovanje+I(staz*binarna)+I(obrazovanje*binarna),data=bin))

## Call:
## lm(formula = placa ~ staz + obrazovanje + I(staz * binarna) +
##     I(obrazovanje * binarna), data = bin)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9881  -4.2011   0.2059   3.9496  12.1290
##
## Coefficients:
##                            Estimate Std. Error  t value Pr(>|t|)
## (Intercept)              4009.9901     1.1464 3497.869  < 2e-16 ***
## staz                        1.4493     0.2915    4.972 1.85e-06 ***
## obrazovanje                -0.2500     0.3195   -0.782    0.435
## I(staz * binarna)          -0.1302     0.4386   -0.297    0.767
## I(obrazovanje * binarna)    0.3163     0.4879    0.648    0.518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.638 on 145 degrees of freedom
## Multiple R-squared:  0.9032, Adjusted R-squared:  0.9005
## F-statistic: 338.3 on 4 and 145 DF,  p-value: < 2.2e-16
```

Slika 3.4. Procijenjeni model

Za mušku osobu koja ima 0 godina radnog staža i 0 godina obrazovanja, prosječna plaća iznosi 4099,99 kuna. Povećanje godina radnog staža za 1 godinu, uz nepromijenjen broj godina obrazovanja za muške osobe povećava plaću u prosjeku za 1,32 kuna, dok povećanje broja godina obrazovanja za 1 godinu, uz nepromijenjen broj godina radnog staža za muške osobe povećava plaću u prosjeku za 0,07 kuna.

Stoga su vrijednosti kod članova interakcije $-0{,}13x_{i1}D_i + 0{,}32x_{i2}D_i$ razlike u učincima povećanja radnog staža, odnosno godina obrazovanja na plaću u ovisnosti o tome radi li se o muškoj ili ženskoj osobi. –0,13 se interpretira na način da je za 0,13 kuna za muške osobe manji učinak povećanja radnog staža za 1 godinu na povećanje plaće u odnosu na ženske osobe (uz nepromijenjen broj godina obrazovanja). S druge strane, 0,32 se interpretira na način da je za 0,32 kune za muške osobe veći učinak povećanja broja godina obrazovanja za 1 godinu na povećanje plaće u odnosu na ženske osobe (uz nepromijenjen broj godina radnog staža).

Primjer 3.4.
U istoj datoteci generirajte dvije nove binarne varijable. Prva neka je jednaka 1 u slučaju da osoba ima 15 i više godina radnog staža, a 0 inače. Druga binarna varijabla neka je jednaka 1 u slučaju da osoba ima 12 i više godina obrazovanja, a 0 inače. Procijenite model u kojemu plaća zaposlenika ovisi o dvije binarne varijable i njihovoj interakciji. Interpretirajte rezultate.

Najprije su generirane binarne varijable bin1 i bin2 (slika 3.5), te je potom procijenjen model u kojemu je za ovisnost zavisne varijable o nezavisnima i njihovom umnošku, tj. interakciji dovoljno pisati *x**y* (tj. samo umnožak, jer RStudio prepoznaje naredbu da se radi o zbroju pojedinačnih varijabli, te potom njihovih umnožaka).

Sada je model koji se razmatra sljedeći:

$$\hat{y}_i = 4017,41 + 13,66D_{i1} + 13,71D_{i2} + 3,73D_{i1}D_{i2},$$
$$D_{i1} = \begin{cases} 1, \text{ staž} \geq 15 \\ 0, \text{ inače} \end{cases}, \quad D_{i2} = \begin{cases} 1, \text{ obrazovanje} \geq 12 \\ 0, \text{ inače} \end{cases}$$

```
bin1<-ifelse(bin$staz>=15,1,0)
bin2<-ifelse(bin$obrazovanje>=12,1,0)
summary(lm(placa~bin1*bin2,data=bin))

## Call:
## lm(formula = placa ~ bin1 * bin2, data = bin)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.1027  -7.3073   0.0007   6.7974  21.9165
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 4017.410      1.490 2697.007   <2e-16 ***
## bin1          13.659     10.533    1.297   0.1967
## bin2          13.711      7.522    1.823   0.0704 .
## bin1:bin2      3.730     12.900    0.289   0.7729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.43 on 146 degrees of freedom
## Multiple R-squared:  0.6666, Adjusted R-squared:  0.6598
## F-statistic: 97.32 on 3 and 146 DF, p-value: < 2.2e-16
```

Slika 3.5. Generiranje binarnih varijabli i procijenjeni model

Stoga možemo popuniti sljedeću tablicu radi jednostavnijeg uočavanja interpretacije rezultata:

Obrazovanje / Staž	$D_{i1} = 0$	$D_{i1} = 1$
$D_{i2} = 0$	4017,41	4031,07
$D_{i2} = 1$	4031,12	4048,51

Interpretacije su sljedeće. Osobe koje imaju manje od 15 godina radnog staža i manje od 12 godina obrazovanja ($D_{i1} = D_{i2} = 0$) imaju plaću u prosjeku 4017,41 kuna. Osobe koje imaju 15 i više godina radnog staža i manje od 12 godina obrazovanja ($D_{i1} = 1$, $D_{i2} = 0$) imaju plaću u prosjeku 4031,07 kuna. Osobe koje imanju manje od 15 godina radnog staža 12 i više godina obrazovanja ($D_{i1} = 0$, $D_{i2} = 1$) imaju plaću u prosjeku 4031,12 kuna, dok osobe koje imaju 15 i više godina radnog staža te 12 i više godina obrazovanja imaju plaću u prosjeku 4048,51 kuna.

Stoga se procijenjeni parametri mogu interpretirati na sljedeći način: 13,66 kuna je razlika u prosječnoj plaći između osoba koje imaju manje od 15 godina radnog staža i onih koje imaju 15 i više godina radnog staža, a pritom imaju manje od 12 godina obrazovanja (uspoređuje se razlika u prvome retku tablice – fiksiramo da je $D_{i2} = 0$). 13,71 kuna je razlika u prosječnoj plaći između osoba koje imaju manje od 12 godina obrazovanja i onih koje imaju 12 i više godina obrazovanja, a pritom imaju manje od 15 godina radnoga staža (uspoređuje se razlika u prvome stupcu tablice, fiksiramo da je $D_{i1} = 0$). Konačno 3,73 kune nema svoju interpretaciju, već se radi o slučaju kada su obje binarne varijable jednake 1 pa je zbroj svih konstanti (13,66 + 13,71 + 3,73) razlika u prosječnoj plaći osoba koje imaju 15 i više godina radnog staža te 12

i više godina obrazovanja u odnosu na osobe koje imaju manje od 15 godina radnog staža te manje od 12 godina obrazovanja.

3.2. Nelinearni regresijski modeli

U poglavlju 2.1.2. navedena je pretpostavka linearnosti modela, pri čemu se naglašavala linearnost u parametrima. Postoje modeli kod kojih je moguće nekom transformacijom postići linearnost u parametrima, pa se mogu procijeniti metodom najmanjih kvadrata. S druge strane, kod onih modela kada nije moguće provesti linearizaciju, govorimo da se radi o pravim nelinearnim modelima. U tom slučaju se parametri procjenjuju nelinearnom metodom najmanjih kvadrata (engl. *NLS*, *nonlinear least squares*).

Neke od tipičnih transformacija modela u linearne u parametrima prikazane su u poglavlju 2.1.2. Ovdje će se detaljnije razmotriti specijalan slučaj, učestalo primjenjivan u mikro i makroekonomiji, Cobb-Douglasova proizvodna funkcija:

$$Q(L,K) = \beta_0 L^{\beta_1} K^{\beta_2} e^{\varepsilon}, \tag{3.9}$$

gdje Q predstavlja količinu proizvodnje, L i K rad i kapital (mjerne jedinice mogu varirati i biti izrađene u tonama, količinama proizvoda, novčano, satima rada, satima rada strojeva, itd.). Kako bi se model u (3.9) mogao procijeniti, najprije ga je potrebno logaritamskom transformacijom zapisati na način:

$$\ln Q(L,K) = \ln\left(\beta_0 L^{\beta_1} K^{\beta_2} e^{\varepsilon}\right) = \ln\beta_0 + \beta_1 \ln L + \beta_2 \ln K + \varepsilon, \tag{3.10}$$

odnosno

$$\ln Q_i = \underbrace{\ln\beta_0}_{=\alpha} + \beta_1 \ln L_i + \beta_2 \ln K_i + \varepsilon_i, \tag{3.11}$$

ako se radi o presječnim podacima, dok bi se koristio indeks t za vremenske nizove. Model u (3.11) se sada može procijeniti metodom najmanjih kvadrata, te je moguće vršiti interpretacije i testiranje kao i do sada, no treba paziti da se radi o log-log modelu.

Primjer 3.5.
Učitajte datoteku „**cobb-douglas.txt**“ u RStudio. Datoteka sadrži podatke o 40 poduzeća: ukupna količina proizvodnje (u kg), ukupno uložen rad (u satima), te ukupno uložen kapital (u satima rada strojeva). Procijenite model (3.9) i interpretirajte parametre. Provedite pojedinačne i skupni test značajnosti varijabli, uz razinu α = 5%. Provedite Waldov test jesu li prisutni konstantni prinosi na opseg.

Temeljem slike 3.6. možemo zapisati sljedeći model: $\widehat{\ln Q_i} = -0{,}05 + 0{,}37\ln L_i + 0{,}62\ln K_i$. Kako se radi o log-log modelu, interpretacije parametara su u postotcima (vidjeti naslov 2.2.4): ako se uložen rad poveća za 1%, uz nepromijenjen kapital, ukupna proizvodnja se poveća za približno 0,37%. Ako se ukupno uložen kapital poveća za 1%, uz nepromijenjen rad, ukupna proizvodnja se poveća za približno 0,62%. Ovdje se radi o **graničnom proizvodu rada i graničnom proizvodu kapitala**.

Nadalje, pojedinačni testovi značajnosti varijabli u modelu su sljedeći (provode se jednosmjerni testovi, s obzirom na ekonomsko tumačenje, ali i pozitivnih predznaka), pri čemu se donose

zaključci da uz razinu značajnosti od 5% odbacujemo hipotezu da varijabla ln$L$ nije značajna u modelu, kao i što odbacujemo hipotezu da varijabla ln$K$ nije značajna u modelu.

Varijabla ln(L)	**Varijabla ln(K)**
$H_0 : \beta_1 = 0$ $H_1 : \beta_1 > 0$	$H_0 : \beta_2 = 0$ $H_1 : \beta_2 > 0$
$t_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0,372}{0,003} = 114,546$	$t_2 = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{0,624}{0,003} = 181,667$
$t_{0,05}(N–k–1=37)=1,687$	$t_{0,05}(N–k–1=37)=1,687$
$t_1 > t_{0,05}(137) \rightarrow$ Odbacujem H_0 ili p-v $\approx 0 < 0,05 \rightarrow$ Odbacujem H_0	$t_2 > t_{0,05}(137) \rightarrow$ Odbacujem H_0 ili p- v $\approx 0 < 0,05 \rightarrow$ Odbacujem H_0

```
cd<-read.table("cobb-douglas.txt",sep="\t",header=T)
summary(lm(log(proizvodnja)~log(rad)+log(kapital),data=cd))
```

```
## Call:
## lm(formula = log(proizvodnja) ~ log(rad) + log(kapital), data = cd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0034333 -0.0015728 -0.0000916  0.0013707  0.0055202
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.047580   0.014192  -3.353  0.00186 **
## log(rad)      0.371616   0.003244 114.546  < 2e-16 ***
## log(kapital)  0.624127   0.003436 181.667  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002326 on 37 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 1.15e+05 on 2 and 37 DF,  p-value: < 2.2e-16
```

Slika 3.6. Procijenjen model

Nadalje, skupni test značajnosti provodi se temeljem ispisa na slici 3.6. na sljedeći način:

$$\begin{matrix} H_0 : \beta_1 = \beta_2 = 0 \\ H_1 : \exists \beta_j \neq 0, j \in \{1,2\} \end{matrix}, \quad F = 1,15 \cdot 10^5, p\text{-vrijednost} \approx 0 < 0,05,$$

stoga se odbacuje nulta hipoteza. Uz razinu značajnosti od 5%, odbacuje se hipoteza da niti jedna varijabla nije značajna u modelu.

Kako bi se Waldovim testom provelo testiranje hipoteze jesu li konstantni prinosi na opseg, ponovimo homogenost funkcija više varijabli. Kažemo da je funkcija $z = f(x_1, x_2, \ldots, x_n)$ od n varijabli definirana na skupu $A \subseteq \mathbb{R}^n$ homogena stupnja homogenosti α, ako vrijedi: $f(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) = \lambda^\alpha f(x_1, x_2, \ldots, x_n)$ za proizvoljno $\lambda \in \mathbb{R}$ i svako $(x_1, x_2, \ldots, x_n) \in A$ za koje je $(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) \in A$. Specijalno za $\alpha = 1$, kažemo da je funkcija linearno homogena.

Za slučaj funkcija proizvodnje, ako računamo:

$$Q(\lambda L, \lambda K) = \beta_0 \lambda^{\beta_1} L^{\beta_1} \lambda^{\beta_2} K^{\beta_2} e^{\varepsilon} = \lambda^{\beta_1+\beta_2} \beta_0 L^{\beta_1} K^{\beta_2} e^{\varepsilon},$$

uočava se kako je $\alpha = \beta_1+\beta_2$. Kako se α interpretira kao približna promjena vrijednosti funkcije ako se sve nezavisne varijable povećaju za 1%, u slučaju funkcije proizvodnje radi se interpretaciji prinosa na opseg proizvodnje. Ako vrijedi $\beta_1+\beta_2 = 1$, radi se o konstantnim prinosima na opseg.

Stoga se u zadatku testira u nultoj hipotezi: H_0: $\beta_1+\beta_2 = 1$. Matrično se nulta hipoteza Waldova testa stoga zapisuje kao:

$$H_0 : \boldsymbol{R\beta} = \boldsymbol{q},\ \boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix},\ \boldsymbol{q} = \begin{bmatrix} 1 \end{bmatrix}.$$

Temeljem slike 3.7., pišemo test veličinu $W = 4{,}20$, s pripadajućom p-vrijednošću 0,04, što je manje od 0,05, stoga se odbacuje nulta hipoteza. Uz razinu značajnosti od 5%, odbacujemo hipotezu da su prisutni konstantni prinosi na opseg.

```
model<-lm(log(proizvodnja)~log(rad)+log(kapital),data=cd)
library(car)
ogranicenje<-"log(rad)+log(kapital)=1"
linearHypothesis(model,ogranicenje,test="Chisq")

## Linear hypothesis test
##
## Hypothesis:
## log(rad)  + log(kapital) = 1
##
## Model 1: restricted model
## Model 2: log(proizvodnja) ~ log(rad) + log(kapital)
##
##   Res.Df        RSS Df  Sum of Sq  Chisq Pr(>Chisq)
## 1     38 0.00022289
## 2     37 0.00020014  1 2.2741e-05 4.2041    0.04033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 3.7 Waldov test

3.3. Pitanja za ponavljanje

1) Što su to binarne varijable? Čemu služe?
2) Kakva je to binarna varijabla za konstantni član u regresijskom modelu? Interpretirajte njeno značenje.
3) Kakva je to binarna varijabla za promjene u nagibu u regresijskom modelu? Interpretirajte njeno značenje.
4) Zapišite regresijski model u kojemu razmatramo ovisnost zavisne varijable y o dvije nezavisne varijable, x_1 i x_2, te binarnu varijablu za konstantni član.
5) Zapišite model u kojemu razmatramo kojemu razmatramo ovisnost zavisne varijable y o dvije nezavisne varijable, x_1 i x_2, te binarnu varijablu za promjene u nagibu za koeficijent smjera uz varijablu x_2.

6) Učitajte datoteku „**potrosnja.txt**“ u RStudio. Datoteka sadrži podatke o ukupnoj potrošnji (milijarde konstantne LCU, engl. *local currency unit*), ukupnom BDP-u (milijarde konstantne LCU) te indeksu potrošačkih cijena (indeksni bodovi) za 140 zemalja svijeta u 2018. godini.
   a) Generirajte binarnu varijablu koja poprima vrijednost 1 za one zemlje čiji je indeks cijena 125 i veći, a 0 inače. Procijenite 3 modela: u prvome neka potrošnja ovisi o BDP-u zemlje te binarnoj varijabli; u drugome neka binarna varijabla utječe na učinak BDP-a zemlje na potrošnju u ovisnosti je li indeks potrošačkih cijena 125 i veći ili ne; te u trećemu neka potrošnja ovisi samo o binarnoj varijabli. Spojite ispis sva tri modela pomoću naredbe stargarzer() i interpretirajte sve procijenjene parametre u sva tri modela.
   b) Potom generirajte drugu binarnu varijablu koja poprima vrijednost 1 ako zemlja ima BDP 1500 i veći (mlrd LCU). Procijenite 2 modela: u prvome neka potrošnja zemlje ovisi o dvjema binarnim varijablama (prva iz postupka a) te druga novoformirana); a u drugome neka potrošnja ovisi i o interakciji tih dviju binarnih varijabli. Spojite ispis oba modela pomoću naredbe stargarzer() i interpretirajte sve procijenjene parametre u njima.

Rješenja

Zadatak 4):

$$yi = a + bx_{i1} + cx_{i2} + dD_i + e_i,\ \ D_i = \begin{cases} 1, grupa\ 1 \\ 0, grupa\ 2 \end{cases}$$

Zadatak 5):

$$yi = a + bx_{i1} + cx_{i2} + dD_i x_{i2} + u_i,\ \ D_i = \begin{cases} 1, grupa\ 1 \\ 0, grupa\ 2 \end{cases}$$

Zadatak 6):

```
potrosnja<-read.table("potrosnja.txt",header=T,sep="\t")

bin<-ifelse(potrosnja$cijene>=125,1,0)

m1<-lm(potrosnja~bdp+bin,data=potrosnja)
m2<-lm(potrosnja~bdp+I(bin*bdp),data=potrosnja)
m3<-lm(potrosnja~bin,data=potrosnja)

library(stargazer)
stargazer(list(m1,m2,m3),type="text")

bin2<-ifelse(potrosnja$bdp>=1500,1,0)

m4<-lm(potrosnja~bin+bin2,data=potrosnja)
m5<-lm(potrosnja~bin*bin2,data=potrosnja)

stargazer(list(m4,m5),type="text")
```

```
## ================================================================================================
##                                          Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                potrosnja
##                                (1)                    (2)                      (3)
## ------------------------------------------------------------------------------------------------
## bdp                          0.645***               0.647***
##                               (0.003)                (0.018)
##
## bin                          7,458.418                                     135,676.000
##                             (5,780.006)                                   (102,416.000)
##
## I(bin * bdp)                                        -0.002
##                                                     (0.018)
##
## Constant                      988.872               4,520.722              24,872.530
##                              (4,007.148)           (2,933.061)            (71,377.000)
##
## ------------------------------------------------------------------------------------------------
## Observations                    140                    140                    140
## R2                             0.997                  0.997                  0.013
## Adjusted R2                    0.997                  0.997                  0.005
## Residual Std. Error    33,987.950 (df = 137)  34,192.080 (df = 137)  605,654.000 (df = 138)
## F Statistic         22,120.430*** (df = 2; 137) 21,856.290*** (df = 2; 137)  1.755 (df = 1; 138)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

```
## ===================================================================
##                                 Dependent variable:
##                     ----------------------------------------------
##                                       potrosnja
##                                (1)                    (2)
## -------------------------------------------------------------------
## bin                       116,490.000              34.242
##                          (102,469.500)          (145,249.100)
## bin2                      167,740.600             55,444.060
##                          (102,427.700)          (142,631.800)
## bin:bin2                                         231,379.100
##                                                 (204,736.500)
## Constant                  -49,678.860              230.725
##                           (84,295.390)           (95,087.870)
## -------------------------------------------------------------------
## Observations                  140                    140
## R2                           0.032                  0.041
## Adjusted R2                  0.017                  0.019
## Residual Std. Error 601,996.600 (df = 137) 601,388.500 (df = 136)
## F Statistic          2.229 (df = 2; 137)    1.915 (df = 3; 136)
## ===================================================================
## Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

4.

NARUŠAVANJE PRETPOSTAVKI REGRESIJSKOG MODELA

4. NARUŠAVANJE PRETPOSTAVKI REGRESIJSKOG MODELA

Pretpostavke linearnog regresijskog modela obrađene su u poglavljima 2.1.2. i 2.2.2. Kako se u praksi češće događa da je narušena neka od tih pretpostavki, u odnosu na to da su sve zadovoljene, potrebno je razmotriti do kakvih problema dolazi ako je narušena neka od pretpostavki, kako ćemo utvrditi postoji li neki problem, te kako ga ublažiti ili otkloniti. Problemi koji se javljaju su:

- Multikolinearnost nezavisnih varijabli
- Autokorelacija grešaka relacije
- Heteroskedastičnost grešaka relacije
- Nenormalnost distribucije grešaka relacije

4.1. Multikolinearnost nezavisnih varijabli

S obzirom da većina empirijskih modela uključuje više nezavisnih varijabli u modelu, potrebno je provjeriti postoji li multikolinearnost tih varijabli, što se opisuje u idućim potpoglavljima.

4.1.1. Definiranje problema multikolinearnosti nezavisnih varijabli

Jedna od pretpostavki linearnog regresijskog modela u slučaju 2 ili više nezavisnih varijabli bila je: varijable x_i su međusobno nezavisne, što znači da vrijedi $r(\boldsymbol{X}'\boldsymbol{X})^{-1} = r(\boldsymbol{X}) = k + 1$, tj. postoji inverz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Drugim riječima, stupci u matrici $\boldsymbol{X}$ bili su linearno neovisni, što znači da je za potrebe procjene parametara regresijskog modela u (2.172):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \tag{4.1}$$

moguće izračunati $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ (vidjeti naslove 2.2.3., kao i već obrađen 2.1.3.2). Ako su stupci u $\boldsymbol{X}$ linearno neovisni, tada je rang te matrice jednak $r(\boldsymbol{X}) = k + 1$, te je ujedno i rang matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ jednak $k + 1$, odnosno moguće je izračunati $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (tj. matrica $\boldsymbol{X}'\boldsymbol{X}$ je invertibilna).

Ako je **narušena pretpostavka o međusobnoj nezavisnosti regresorskih varijabli**, tada govorimo o **problemu multikolinearnosti**. Pritom se može raditi o **savršenoj** multikolinearnosti ili pak **približnoj**. Savršena multikolinearnost rjeđe se javlja u praksi sa stvarnim podacima, jer bi to značilo da su podaci za varijablu x_j dobiveni tako da je napravljena linearna transformacija nad podacima za neku varijablu x_l. Primjerice, da se jedan stupac u matrici $\boldsymbol{X}$ pomnoži s nekom konstantom $c \in \mathbb{R}$ da bi se formirao neki drugi stupac u toj matrici, u tom slučaju bismo rekli da postoji savršena multikolinearnost. Tada je determinanta matrice $\boldsymbol{X}'\boldsymbol{X}$ jednaka nuli. Stoga nije moguće procijeniti vektor parametara dan izrazom (2.182):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{4.2}$$

jer nije moguće izračunati inverz matrice $\boldsymbol{X}'\boldsymbol{X}$. Primijetimo da smo mogli formulu (4.2) zapisati i preko inverza matrice $\boldsymbol{X}'\boldsymbol{X}$:

$$\hat{\boldsymbol{\beta}} = \frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})} adj(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{X}'\boldsymbol{y}, \tag{4.3}$$

gdje $\det(\boldsymbol{X}'\boldsymbol{X})$ predstavlja determinantu i $adj(\boldsymbol{X}'\boldsymbol{X})$ adjunktu od $\boldsymbol{X}'\boldsymbol{X}$. Zato u slučaju savršene multikolinearnosti nije moguće izračunati procjenitelj $\hat{\boldsymbol{\beta}}$ budući da je vrijednost determinante u nazivniku izraza (4.3) jednaka nuli.

Češća je pojava približne multikolinearnosti između nezavisnih varijabli. U tom slučaju vrijedi $\det(\boldsymbol{X}'\boldsymbol{X}) \approx 0$ te je matrica $\boldsymbol{X}'\boldsymbol{X}$ loše uvjetovana (u literaturi se koriste nazivi *ill conditioned matrix, near singular matrix*). Tada je moguće procijeniti parametre regresijskog modela pomoću izraza (4.3), no problem se javlja kod procjene varijanci i kovarijanci procjenitelja u (2.185):

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \hat{\sigma}^2 \frac{1}{\det\left(\boldsymbol{X}'\boldsymbol{X}\right)} adj\left(\boldsymbol{X}'\boldsymbol{X}\right), \tag{4.4}$$

gdje su zbog $\det(\boldsymbol{X}'\boldsymbol{X}) \approx 0$ u nazivniku varijance i kovarijance u $Var\left(\hat{\boldsymbol{\beta}}\right)$ jako velike. **Stoga će provođenje *t*-testova biti nepouzdano** jer se empirijski *t*-omjeri računaju formulom (2.190) :

$$t_j = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)}, \tag{4.5}$$

što znači da će ti omjeri biti izuzetno mali jer je prema formuli (2.188):

$$SE(\hat{\beta}_j) = \sqrt{Var\left(\hat{\beta}_j\right)} = \hat{\sigma}\sqrt{s_{jj}}\,, \tag{4.6}$$

pa bi se prilikom provođenja *t*-testa zaključilo da neka varijabla nije značajna u modelu.

Nadalje, problem se javlja i kod intervalne procjene parametara, jer se koristi formula (2.202):

$$P\left(\hat{\beta}_j - t_{\gamma/2} SE\left(\hat{\beta}_j\right) < \beta_j < \hat{\beta}_j + t_{\gamma/2} SE\left(\hat{\beta}_j\right)\right) = 1-\gamma\,, \tag{4.7}$$

gdje se također koriste standardne pogreške procjenitelja $SE\left(\hat{\beta}_j\right)$ koje ovise o $\det(\boldsymbol{X}'\boldsymbol{X})$. Stoga su intervalne procjene parametara također nepouzdane. Uočimo da će u slučaju velikih vrijednosti $SE\left(\hat{\beta}_j\right)$ intervali procjene biti jako veliki.

4.1.2. Utvrđivanje postojanja problema multikolinearnosti nezavisnih varijabli

Prve indikacije da postoji problem multikolinearnosti nezavisnih varijabli su sljedeće (Greene, 2018):

- Male izmjene podataka uzrokuju velike promjene u procijenjenim vrijednostima $\hat{\boldsymbol{\beta}}$.
- Jako velike standardne pogreške procjenitelja koje upućuju na ne odbacivanje nulte hipoteze kod *t*-testova, dok istovremeno rezultat *F*-testa upućuje na skupnu značajnost varijabli u modelu (odbacivanje nulte hipoteze).
- Procijenjeni parametri imaju „pogrešan“ predznak ili pak nerealne jačine vrijednosti u odnosu na očekivane (u usporedbi s ekonomskom teorijom)

Formalno, razmatraju se sljedeći pokazatelji. Naime, kako je spomenuto da se varijance i kovarijance procjenitelja računaju formulom (4.4), pojedinačna varijanca računa se izrazom (vidjeti formulu (2.188)):

$$Var\left(\hat{\beta}_j\right) = \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj} = \hat{\sigma}^2 s_{jj}, \tag{4.8}$$

izraz (4.8) se može zapisati na sljedeći način (izvod vidjeti u Greene, 2018: 136):

$$\begin{aligned} Var\left(\hat{\beta}_j\right) = \hat{\sigma}^2 s_{jj} &= \frac{\hat{\sigma}^2}{(1-R_j^2)\sum_{i=1}^{N}\left(x_{ij}-\overline{x_j}\right)^2} \\ &= \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}\left(x_{ij}-\overline{x_j}\right)^2} \cdot \frac{1}{1-R_j^2} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}\left(x_{ij}-\overline{x_j}\right)^2} \cdot VIF_j \end{aligned}, \tag{4.9}$$

gdje R^2_j predstavlja koeficijent determinacije u modelu u kojemu j-ta regresorska varijabla postaje zavisna, a ostali regresori ostaju nezavisne. Naime, ako neka varijabla j ovisi o drugim varijablama, tj. linearno je povezana s njima, tada će koeficijent determinacije R^2_j u modelu u kojemu je ta varijabla j zavisna varijabla biti velik, stoga je vrijednost $1- R^2_j$ mala i vrijednost $Var\left(\hat{\beta}_j\right)$ je velika. U tu svrhu se koristi pokazatelj koji se naziva **faktor inflacije varijance** (engl. *VIF*, *variance inflation factor*), koji se računa na sljedeći način za svaku varijablu j:

$$VIF_j = \frac{1}{1-R_j^2}. \tag{4.10}$$

Obično se uzima da postoji problem multikolinearnosti varijabli u modelu ako je $R^2_j \geq 0.8$, odnosno ako je $VIF_j \geq 5$ (slijedi iz (4.10) i nejednakosti $R^2_j \geq 0.8$). Ponekad se koristi i pokazatelj *TOL* (engl. *tolerance*), koji je recipročna vrijednost *VIF*-u:

$$TOL_j = \frac{1}{VIF_j}, \tag{4.11}$$

pa se razmatra da postoji problem multikolinearnosti varijabli u modelu ako je $TOL_j \leq 0.2$. Neki autori (Wooldridge, 2012) razmatraju kao kritične granice sljedeće: $R^2_j \geq 0.9$, tj. $VIF_j \geq 10$, odnosno $TOL_j \leq 0.1$.

Ono što se može zaključiti temeljem izraza u (4.9) je da, uz ostalo fiksno, dolazi do povećanja $Var(\hat{\beta}_j)$ ako je veća korelacija j-te regresorske varijable s drugima (jer veća vrijednost R^2_j vodi smanjenju vrijednosti $1- R^2_j$ u nazivniku tog izraza i povećanju varijance procjenitelja); veća varijacija varijable j vodi manjoj vrijednosti $Var(\hat{\beta}_j)$ (jer veća vrijednost $\sum_{i=1}^{N}\left(x_{ij}-\overline{x_j}\right)^2$ u nazivniku uzrokuje smanjenje varijance procjenitelja); te što je regresijski model bolji u smislu da je procijenjena varijanca regresije manja, to će $Var(\hat{\beta}_j)$ biti manja (manja vrijednost procijenjene varijance $\hat{\sigma}^2$ u brojniku vodi smanjenju varijance procjenitelja).

Idući pristup testiranja multikolinearnosti jesu dva **Klienova pravila** (engl. *Klein's rules*, Klein, 1962). **Prvo pravilo** tvrdi da postoji problem multikolinearnosti varijabli u modelu ako je

barem jedan koeficijent korelacije između nezavisnih varijabli veći od koeficijenta korelacije regresije. Naime, ideja regresijskog modela je da nezavisne varijable dobro opišu varijaciju zavisne varijable. Stoga bi koeficijent korelacije regresije trebao biti što veći. S druge strane, korelacija između nezavisnih varijabli međusobno mora biti što manja. Stoga će postojati problem multikolinearnosti u modelu ako je po apsolutnoj vrijednosti koeficijent korelacije između nekih od regresorskih varijabli veći od koeficijenta korelacije regresije. Simbolički, neka je zadana korelacijska matrica za nezavisne varijable:

$$\tilde{\boldsymbol{R}} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{21} & 1 & r_{23} & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & \cdots & 1 \end{bmatrix} \tag{4.12}$$

gdje r_{ij} predstavlja koeficijent korelacije između varijable i i j. Koeficijent korelacije regresije već je prethodno uveden u poglavljima 2.1.8. i 2.2.6., oznaka R. Kažemo da postoji problem multikolinearnosti varijabli u modelu ako postoji barem jedan r_{ij} takav da je $| r_{ij} | > R$.

Drugo pravilo je već spomenuto na početku kao jedna od tri indikacije ovog problema: Jako velike standardne pogreške procjenitelja koje upućuju na ne odbacivanje nulte hipoteze kod t-testova, dok istovremeno rezultat F-testa upućuje na skupnu značajnost varijabli u modelu (odbacivanje nulte hipoteze). Formalno, kaže se da postoji problem multikolinearnosti varijabli u modelu ako je koeficijent determinacije modela, R^2, velik (veći od 0.7), a istovremeno su empirijski t-omjeri veoma mali. Naime, ako je velika vrijednost R^2, tada u izračunu empirijskog F-omjera, (vidjeti formulu (2.214)), dolazi do povećanja brojnika zbog izravnog povećanja R^2 i smanjenja nazivnika zbog smanjenja vrijednosti $(1- R^2)$:

$$F = \frac{R^2 / k}{(1-R^2)/(N-k-1)}. \tag{4.13}$$

Kako bi u tom slučaju empirijski F-omjer bio velik, upadao bi u područje odbacivanja nulte hipoteze da niti jedna varijabla nije značajna u modelu. S druge strane, ako su istovremeno empirijski t-omjeri veoma mali, to znači da upadaju u područje ne odbacivanja nulte hipoteze da varijable nisu značajne u modelu. Kako se radi o kontradiktornim ishodima, zaključuje se da postoji problem multikolinearnosti varijabli u modelu.

Drugi testovi otkrivanja multikolinearnosti su: Farraov hi-kvadrat test (Farrar i Glauber, 1967), kondicijski indeks (Maddala, 1988), Theilov indikator (Theil, 1971), Redov indikator (Kovác i dr., 2005), itd.

4.1.3. Ublažavanje/uklanjanje problema multikolinearnosti nezavisnih varijabli

Ako se u modelu utvrdi postojanje problema multikolinearnosti nezavisnih varijabli, sljedeći su postupci njegova ublažavanja/uklanjanja (Brooks, 2014):

- Izostavljanjem jedne ili više varijabli koja najviše doprinosi problemu multikolinearnosti. Međutim, tu može doći do problema izostavljene značajne varijable (spomenuti problem *omitted variable bias*, vidjeti poglavlje 2.2.7.5.).

- Transformacija onih varijabli koje su jako korelirane u omjere, te korištenje omjera u regresijskom modelu umjesto originalnih vrijednosti. Na taj način se uklanja eksponencijalni trend u varijabli koja doprinosi problemu.
- Povećanje uzorka, s obzirom da se može raditi o problemu uzorka, a ne populacije.
- Ignoriranje problema ako se model koristi isključivo u svrhe prognoziranja, a ostala dijagnostika modela je u redu.

4.1.4. Primjer

Učitajte datoteku „**stanovi.txt**“ u RStudio. Datoteka sadrži sljedeće podatke o 50 stanova: cijena kvadratnog metra pojedinog stana (cijena), broj kvadratnih metara (kvadrat), broj soba (sobe), starost stana (godine), udaljenost stana od centra grada (udaljenost). Procijenite linearni regresijski model u kojemu cijena stana ovisi o ostalim varijablama u modelu. Pomoću kriterija *VIF*, *TOL*, te pomoću oba Kleinova pravila ispitajte postoji li problem multikolinearnosti varijabli u modelu. Izračunajte koliko iznose koeficijenti determinacije za one regresijske varijable kod kojih se utvrdi problem multikolinearnosti.

Na slici 4.1. prikazan je postupak učitavanja podataka, procjene modela te provedbe testiranja pomoću *VIF* i *TOL* pokazatelja. Sada možemo pisati:

$$VIF_1 = 26{,}70,\ VIF_2 = 26{,}46,\ VIF_3 = 2{,}09,\ VIF_4 = 1{,}10,$$

gdje uočavamo da su čak dvije vrijednosti *VIF* veće od 5, stoga zaključujemo kako postoji problem multikolinearnosti varijabli u modelu. Ekvivalentno smo mogli temeljem pokazatelja *TOL*: $TOL_1 = 0{,}04$, $TOL_2 = 0{,}04$, $TOL_3 = 0{,}92$, $TOL_4 = 0{,}91$ doći do istog zaključka jer su prve dvije vrijednosti manje od 0.2. To znači da bi u regresijskim modelima u kojima bi prve dvije nezavisne varijable promatrali kao zavisne koeficijenti determinacije bili veći od 80%. To je prikazano naredbama i ispisom na slici 4.2, gdje se uočava kako je $R_1^2 = 0{,}963$ (model u kojemu je zavisna varijabla kvadrat) te $R_1^2 = 0{,}962$ (model u kojem je zavisna varijabla sobe).

```
stanovi<-read.table("stanovi.txt",sep="\t",header=T)
model<-lm(cijena~kvadrat+sobe+godine+udaljenost,data=stanovi)
library(car)
vif(model)

##    kvadrat       sobe     godine udaljenost
##  26.697603  26.456214   1.088931   1.103061

1/vif(model)
##    kvadrat       sobe     godine udaljenost
## 0.03745655 0.03779830 0.91833149 0.90656791
```

Slika 4.1. Ispitivanje multikolinearnosti u modelu

```
m2<-lm(sobe~kvadrat+godine+udaljenost,data=stanovi)
summary(m1)$r.squared
## [1] 0.9625435
summary(m2)$r.squared
## [1] 0.9622017
```

Slika 4.2. Koeficijenti determinacije R_j^2

Ako ispitujemo problem multikolinearnosti pomoću Kleinovih pravila, za prvo pravilo računamo korelacijsku matricu za razmatrane varijable, kao i koeficijent korelacije regresije za

početni model, što je prikazano na slici 4.3. Koeficijent višestruke linearne korelacije iznosi *R* = 0,982, dok su koeficijenti korelacija za svaki par varijabli dani u korelacijskoj matrici pomoću naredbe cor(). Sada se uočava da su svi koeficijenti korelacija po apsolutnoj vrijednosti manji od vrijednosti *R*, stoga se pomoću ovog prvog pravila dolazi do zaključka da nema problema multikolinearnosti varijabli u modelu.

```
cor(stanovi)

##                cijena     kvadrat         sobe      godine   udaljenost
## cijena      1.0000000  0.78350150  0.772192289 -0.52535133 -0.417120020
## kvadrat     0.7835015  1.00000000  0.978988399 -0.09359588  0.043776437
## sobe        0.7721923  0.97898840  1.000000000 -0.05742016 -0.001503659
## godine     -0.5253513 -0.09359588 -0.057420159  1.00000000  0.165709301
## udaljenost -0.4171200  0.04377644 -0.001503659  0.16570930  1.000000000

sqrt(summary(model)$r.squared)

## [1] 0.9820341
```

Slika 4.3. Korelacijska matrica i koeficijent korelacije regresije

```
summary(model)

## Call:
## lm(formula = cijena ~ kvadrat + sobe + godine + udaljenost, data = stanovi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2746  -5.0720  -0.2008   4.8410  11.9819
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2009.7457     2.9255 686.966  < 2e-16 ***
## kvadrat        0.6240     0.1246   5.010  8.9e-06 ***
## sobe           0.6457     2.5767   0.251    0.803
## godine        -1.4330     0.1075 -13.336  < 2e-16 ***
## udaljenost    -2.6140     0.2011 -13.000  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.186 on 45 degrees of freedom
## Multiple R-squared:  0.9644, Adjusted R-squared:  0.9612
## F-statistic: 304.7 on 4 and 45 DF,  p-value: < 2.2e-16
```

Slika 4.4. Ispis procijenjenog modela

```
m2<-lm(cijena~sobe+godine+udaljenost,data=stanovi)
m3<-lm(cijena~kvadrat+godine+udaljenost,data=stanovi)
library(stargazer)
stargazer(list(m2,m3),type="text")

## ===============================================================
##                                  Dependent variable:
##                              ---------------------------
##                                         cijena
##                                    (1)            (2)
## ---------------------------------------------------------------
## sobe                            13.308***
##                                  (0.619)
##
## kvadrat                                          0.655***
##                                                   (0.024)
##
```

```
## godine                            -1.556***      -1.427***
##                                    (0.129)        (0.104)
##
## udaljenost                         -2.352***      -2.627***
##                                    (0.240)        (0.193)
##
## Constant                         2,008.432***  2,009.915***
##                                    (3.597)        (2.818)
##
## ------------------------------------------------------------
## Observations                          50             50
## R2                                  0.945          0.964
## Adjusted R2                         0.941          0.962
## Residual Std. Error (df = 46)       7.637          6.123
## F Statistic (df = 3; 46)          261.092***     414.669***
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

Slika 4.5. Procijenjeni modeli uz isključivanje neke od nezavisnih varijabli

Ako se provodi drugo pravilo, tada se promatraju empirijski *t*-omjeri, *F*-omjer i koeficijent determinacije regresije (slika 4.4.). Ono što se uočava da je jedino empirijski *t*-omjer za varijablu sobe mali (0,251), dok ostali upućuju na značajnost ostalih varijabli u modelu. Nadalje, empirijski *F*-omjer upućuje na odbacivanje nulte hipoteze, jer je koeficijent determinacije jako velik (0.9644). Temeljem ispisa danog na slici 4.4., i prethodnih testova moglo bi se zaključiti da postoji problem multikolinearnosti varijabli u modelu, te treba razmotriti da se isključi varijabla kvadrat ili sobe iz modela.

Konačno, procijenjena su dva dodatna modela, u jednome je isključena varijabla kvadrat, a u drugome varijabla sobe, s obzirom na ishod *VIF* vrijednosti. Rezultati su prikazani na slici 4.5. Kako je varijabla sobe bila neznačajna na prethodnoj slici, uočava se da je i koeficijent determinacije u modelu m2 na slici 4.5. manji kada je ta varijabla uključena u model, u odnosu na varijablu kvadrat. Nadalje, u originalnom modelu je procijenjeni koeficijent uz varijablu kvadrat iznosio 0,62, dok u novome modelu m3 iznosi 0,66, što nije velika promjena. S druge strane, u originalnom modelu, procijenjeni koeficijent uz varijablu sobe iznosio je 0,65, dok sada u modelu m2 iznosi 13,31. Upravo to upućuje na problem multikolinearnosti kada je varijabla sobe uključena u modelu, jer se potvrđuje komentar na početku poglavlja: procijenjeni parametri imaju „pogrešan“ predznak ili pak nerealne jačine vrijednosti u odnosu na očekivane!

4.2. Autokorelacija grešaka relacije

Učestali problem u praksi na koji se nailazi jest autokorelacija grešaka relacije, posebice kada se rade analize nad vremenskim nizovima. U nastavku se detaljnije obrađuje ovaj problem.

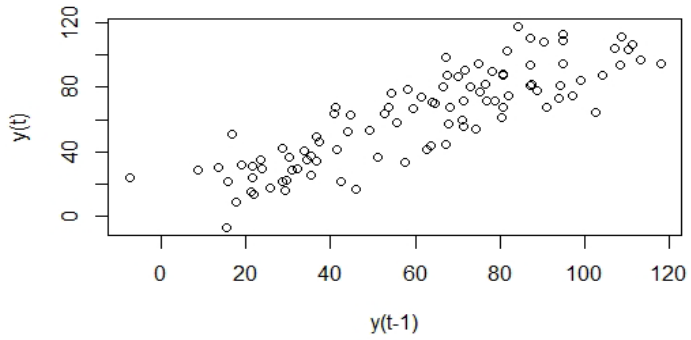
4.2.1. Definiranje problema autokorelacije grešaka relacije

Jedna od pretpostavki regresijskog modela bila je sljedeća: Nezavisnost slučajne varijable, tj. nekoreliranost: $E(\varepsilon_i,\varepsilon_j) = Cov(\varepsilon_i,\varepsilon_j) = Cov(\varepsilon_i,\varepsilon_j|\ x_i) = 0$ za $i \neq j$ (pogledati dodatak 10.3. o nezavisnosti). Ako je narušena ova pretpostavka, govori se o **autokorelaciji slučajne varijable**, pri čemu se prefiks auto odnosi na „sam sa sobom“, pa se radi o korelaciji unutar istog procesa. Autokorelacija se može javiti u presječnim podacima, iako rjeđe (npr. prelijevanje ekonomskih šokova između povezanih zemalja), a češće se javlja za vremenske nizove. Ako se govori o **autokorelaciji prvog reda**, za presječne podatke pišemo:

$$E(\varepsilon_i,\varepsilon_{i-1}) \neq 0, \qquad (4.14)$$

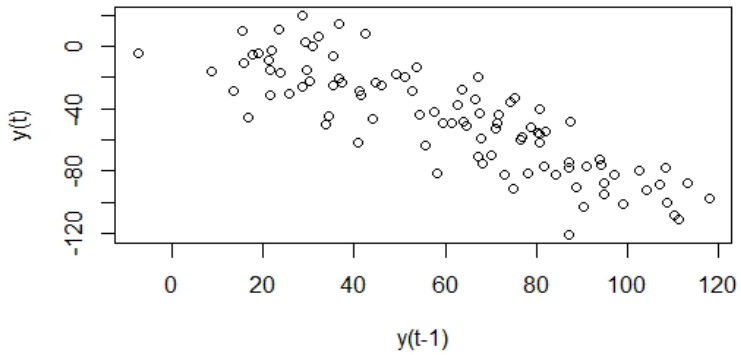
dok za vremenske nizove pišemo:

$$E(\varepsilon_t,\varepsilon_{t-1}) \neq 0. \qquad (4.15)$$

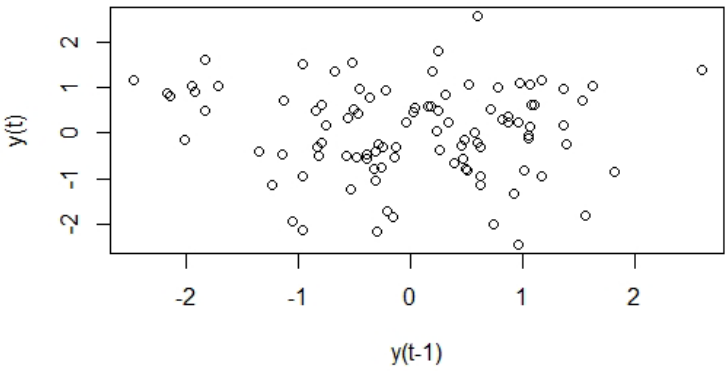


Slika 4.6. Pozitivna autokorelacija prvog reda

Kako koeficijent (auto)korelacije može biti pozitivan ili negativan, oba slučaja su predočena na slici 4.6., koja prikazuje pozitivnu autokorelaciju prvog reda, dok slika 4.7. prikazuje negativnu autokorelaciju prvog reda. Ako ne postoji autokorelacija prvoga reda, onda će to grafički izgledati kao na slici 4.8., odnosno podaci će biti raspršeni nasumično.



Slika 4.7. Negativna autokorelacija prvog reda



Slika 4.8. Nepostojanje autokorelacija prvog reda

Kada smo matrično pisali pretpostavku nezavisnosti slučajne varijable, radilo se o sljedećoj matrici (vidjeti formulu (2.183)):

$$\Omega = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I \tag{4.16}$$

gdje nule predstavljaju vrijednosti kovarijanci:

$$\begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1,\varepsilon_2) & Cov(\varepsilon_1,\varepsilon_3) & \cdots & Cov(\varepsilon_1,\varepsilon_N) \\ Cov(\varepsilon_2,\varepsilon_1) & Var(\varepsilon_2) & Cov(\varepsilon_2,\varepsilon_3) & \cdots & Cov(\varepsilon_2,\varepsilon_N) \\ Cov(\varepsilon_3,\varepsilon_1) & Cov(\varepsilon_3,\varepsilon_2) & \ddots & \cdots & Cov(\varepsilon_3,\varepsilon_N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(\varepsilon_N,\varepsilon_1) & Cov(\varepsilon_N,\varepsilon_2) & Cov(\varepsilon_N,\varepsilon_3) & \cdots & Var(\varepsilon_N) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \tag{4.17}$$

stoga se pretpostavlja ne samo da ne postoji autokorelacija slučajne varijable prvog reda, već ni drugog, trećeg, itd. do *N*-tog reda.

Međutim, ako matrica u (4.16), odnosno (4.17) nije skalarna, radi se o problemu autokorelacije grešaka relacije. U tom slučaju, ako se razmotri izračun varijance procjenitelja, kada ne vrijedi više $Var(\boldsymbol{\varepsilon}) = \sigma^2 I$, tada je $Var(\hat{\boldsymbol{\beta}})$ sljedeća matrica (izvod (2.48)-(2.56)):

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\ &= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')}_{Var(\varepsilon)\neq\sigma^2 I} \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Z\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned} \tag{4.18}$$

pa je matrica varijanci-kovarijanci u (4.18) različita od one u (2.59), tj. vrijedi:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Z\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{4.19}$$

odnosno u prisustvu autokorelacije grešaka relacije, matrica varijanci-kovarijanci procjenitelja **više nije jednaka** $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. **To znači da izračun empirijskih *t*-omjera više nije pouzdan, kao i što intervalne procjene također nisu pouzdane** (jer se temelje na standardnim pogreškama procjenitelja). Nadalje, kako se procjena varijance regresije računa na sljedeći način (vidjeti (2.205)):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-k-1} = \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{N-k-1} = \frac{SSR}{N-k-1}, \tag{4.20}$$

te se upravo ona koristi u izračunu empirijskog *F*-omjera (vidjeti (2.213)):

$$F = \frac{SSE/k}{SSR/(N-k-1)}, \tag{4.21}$$

tako se i u slučaju $Var(\boldsymbol{\varepsilon}) \neq \sigma^2 I$ dolazi do drugačije vrijednosti *SSR* u (4.21), pa time i **empirijski *F*-omjer postaje nepouzdan.** Slično tome, i izračun **koeficijenta determinacije** (vidjeti (2.208)):

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\hat{\sigma}^2 (N-k-1)}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \tag{4.22}$$

postaje nepouzdan zbog toga što se i ovdje koristi *SSR*. Dakle, *t*-test, *F*-test, intervalne procjene parametara, kao i koeficijent determinacije prestaju biti pouzdani. **Autokorelacija ne utječe na nepristranost procjenitelja.** Naime, kako se on računa izrazom $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$, uočava se da se u tome izrazu ne koristi matrica varijanci-kovarijanci procjenitelja.

Razlozi postojanja autokorelacije grešaka relacije su sljedeći:

- Učinak prenošenja (engl. *carryover effect*) kod ekonomskih vremenskih nizova (potrošnja u mjesecu *t* ovisi o potrošnji iz prethodnog mjeseca), te kod presječnih podataka[42] (zemlje koje su blizu jedna drugoj geografski mogu imati veoma slično kretanje ekonomskih varijabli).
- Pogrešna specifikacija modela. Primjerice, pretpostavlja se linearna veza između zavisne i nezavisnih varijabli, dok je zapravo nelinearna (logaritamska, eksponencijalna, itd.) i te nelinearnosti se prelijevaju u autokorelaciju.
- Isključenost određenih varijabli iz modela (za koje nismo mogli prikupiti podatke).

4.2.2. Utvrđivanje postojanja problema autokorelacije grešaka relacije

Kako je u prethodnom podnaslovu prikazano grafički da je moguće uočiti autokorelaciju prvog reda, jedan od inicijalnih indikatora jest grafičko predočavanje dijagrama rasipanja između rezidualnog odstupanja *i* i rezidualnog odstupanja *i*–1. Formalni testovi su sljedeći.

Durbin-Watson test (DW test, Durbin i Watson, 1950, 1951) je najjednostavniji test, s obzirom da se njime testira postojanje autokorelacije rezidualnih odstupanja prvoga reda. Dakle, greške relacije mogu se u tom slučaju razmotriti na sljedeći način:

$$\varepsilon_i = \rho \varepsilon_{i-1} + \nu_i, \tag{4.23}$$

tj. kao autoregresijski proces prvog reda (auto se odnosi na ovisnost o samome sebi, tj. o svojim vrijednostima s pomakom), a ν_i predstavlja slučajnu varijablu, za koju vrijedi $\nu_i \sim N(0,\sigma^2_\nu)$. U nultoj hipotezi DW testa se pretpostavlja da ne postoji autokorelacija prvog reda među članovima procesa ε_i. U alternativnoj hipotezi se može pretpostaviti pozitivna autokorelacija (test na gornju granicu), negativna autokorelacija (test na donju granicu), ili pak postojanje autokorelacije prvog reda koja može biti pozitivna ili negativna (dvosmjerni test). Sažetak mogućih testiranja prikazan je u tablici 4.1.

[42] Da može postojati autokorelacija i u presječnim podacima, vidjeti u Maddala i Lahiri (2009), Greene (2018), Brooks (2014), itd.

Tablica 4.1. Hipoteze Durbin-Watson testa

Hipoteza/test	Dvosmjerni	Na gornju granicu	Na donju granicu
H_0	$\rho = 0$	$\rho = 0$	$\rho = 0$
H_1	$\rho \neq 0$	$\rho > 0$	$\rho < 0$

DW test veličina se definira na sljedeći način:

$$DW = \frac{\sum_{i=2}^{N} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1} \right)^2}{\sum_{i=2}^{N} \hat{\varepsilon}_i^2}. \tag{4.24}$$

Veličina u (4.24) se može raspisati na sljedeći način:

$$DW = \frac{\sum_{i=2}^{N} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1} \right)^2}{\sum_{i=2}^{N} \hat{\varepsilon}_i^2} = \frac{\sum_{i=2}^{N} \hat{\varepsilon}_i^2 - 2\sum_{i=2}^{N} \hat{\varepsilon}_i \, \hat{\varepsilon}_{i-1} + \sum_{i=2}^{N} \hat{\varepsilon}_{i-1}^2}{\sum_{i=2}^{N} \hat{\varepsilon}_i^2}, \tag{4.25}$$

gdje se uočava da se sume $\sum_{i=2}^{N} \hat{\varepsilon}_i^2$ i $\sum_{i=2}^{N} \hat{\varepsilon}_{i-1}^2$ razlikuju jedino u prvom i posljednjem sumandu, stoga za $N \to \infty$ možemo pisati da $\sum_{i=2}^{N} \hat{\varepsilon}_i^2 \approx \sum_{i=2}^{N} \hat{\varepsilon}_{i-1}^2$, odnosno za dovoljno veliki N vrijedi aproksimacija:

$$DW \approx \frac{2\sum_{i=2}^{N} \hat{\varepsilon}_i^2 - 2\sum_{i=2}^{N} \hat{\varepsilon}_i \, \hat{\varepsilon}_{i-1}}{\sum_{i=2}^{N} \hat{\varepsilon}_i^2} = 2\left(\frac{\sum_{i=2}^{N} \hat{\varepsilon}_i^2 - \sum_{i=2}^{N} \hat{\varepsilon}_i \, \hat{\varepsilon}_{i-1}}{\sum_{i=2}^{N} \hat{\varepsilon}_i^2} \right) = 2\left(1 - \frac{\sum_{i=2}^{N} \hat{\varepsilon}_i \, \hat{\varepsilon}_{i-1}}{\sum_{i=2}^{N} \hat{\varepsilon}_i^2} \right), \tag{4.26}$$

gdje je $\sum_{i=2}^{N} \hat{\varepsilon}_i \, \hat{\varepsilon}_{i-1} = (N-1)\mathrm{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_{i-1})$, a $\sum_{i=2}^{N} \hat{\varepsilon}_i^2 = (N-1)\mathrm{Var}(\hat{\varepsilon}_i)$ pa pišemo:

$$DW \approx 2\left(1 - \frac{\mathrm{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_{i-1})}{\mathrm{Var}(\hat{\varepsilon}_i)} \right) = 2(1-\rho), \tag{4.27}$$

gdje je $\rho = \mathrm{Corr}(\hat{\varepsilon}_i, \hat{\varepsilon}_{i-1})$. Dakle, DW test veličina može se jednostavno izračunati kao približna vrijednost $2(1-\rho)$ pomoću koeficijenta autokorelacije prvog reda (uz navedenu pretpostavku na N). Naravno, koeficijent autokorelacije prvog reda se može procijeniti temeljem DW veličine:

$$\rho \approx 1 - \frac{DW}{2}. \tag{4.28}$$

Iz (4.27) slijedi da je DW približno 0 ako je koeficijent autokorelacije prvog reda jednak 1, tj. postoji jaka pozitivna autokorelacija rezidualnih odstupanja prvog reda:

$$DW \approx 2(1-1) = 0, \tag{4.29}$$

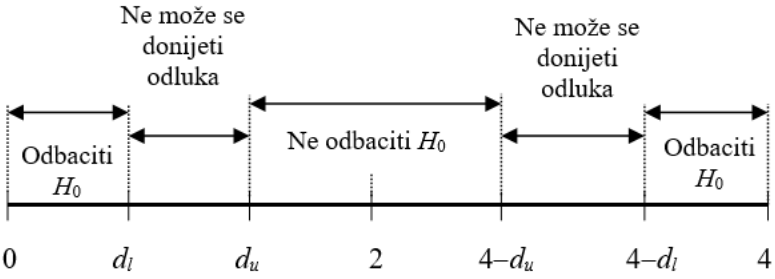
odnosno približno 4 ako je koeficijent autokorelacije prvog reda jednak −1, tj. postoji jaka negativna autokorelacija rezidualnih odstupanja prvog reda:

$$DW \approx 2(1+1) = 4, \tag{4.30}$$

dok u odsustvu autokorelacije (koeficijent autokorelacije prvog reda jednak je 0), DW približno iznosi 2:

$$DW \approx 2(1+0) = 2. \tag{4.31}$$

Stoga se test provodi na sljedeći način: ako je DW veličina u intervalu [0,2], tada se testira pozitivna autokorelacija prvog reda, pri čemu se veličina DW uspoređuje s vrijednostima d_l i d_u, za zadanu razinu značajnosti i veličinu uzorka, te broj regresijskih varijabli. Ako je pak DW veličina u intervalu [2,4], tada se provodi testiranje negativne autokorelacije prvog reda, pri čemu se veličina DW uspoređuje s vrijednostima $4–d_u$ i $4–d_l$: vidjeti sliku 4.9.



Slika 4.9. Durbin-Watson test

Naime, ako je DW vrijednost blizu 2, odnosno između d_u i $4–d_u$, tada ne odbacujemo nultu hipotezu da ne postoji autokorelacija rezidualnih odstupanja prvoga reda (ekvivalentno da smo rekli da je koeficijent autokorelacije prvog reda blizu vrijednosti 0). Nadalje, ako se testira pozitivna autokorelacija, tada ako je DW unutar intervala [0, d_l], odbacujemo nultu hipotezu testa na gornju granicu (jer je koeficijent autokorelacije pozitivan), dok ako je DW unutar intervala [$4–d_l$,4], odbacujemo nultu hipotezu testa na donju granicu (jer je koeficijent autokorelacije negativan). Međutim, uočimo da postoje određeni intervali za koje ne znamo ishod testa.

Stoga se uočava **problem** s ovim testom: postoje slučajevi kada se ne može donijeti odluka. Nadalje, testom se ispituje jedino postojanje autokorelacija prvoga reda, pa se može dogoditi slučaj da ne odbacujemo nultu hipotezu kod DW testa, ali da postoji autokorelacija višeg reda. Također, DW test se može koristiti jedino u slučaju da je obvezno uključena konstanta u regresijskom modelu, regresorske varijable ne smiju biti stohastičke (vidjeti drugu pretpostavku u naslovu 2.1.2. te 2.2.2.); te u modelu se ne smiju koristiti pomaci zavisne varijable (u tim slučajevima bi vrijednost DW bila pristrana prema vrijednosti 2, pa bi se pogrešno zaključivalo o neodbacivanju nulte hipoteze iako bi se trebala odbaciti, vidjeti Brooks, 2014).

Za testiranje autokorelacije **višeg reda**, koristi se **Breusch-Godfrey (1970) test**, koji se provodi u nekoliko koraka. Najprije se procijeni regresijski model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + + \beta_k x_{ik} + \varepsilon_i$ za koji će se testirati autokorelacija rezidualnih odstupanja. Potom se prikupe vrijednosti rezidualnih odstupanja $\hat{\varepsilon}_i$, te se formira **pomoćna regresijska jednadžba**:

$$\hat{\varepsilon}_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + + \alpha_k x_{ik} + \gamma_1 \hat{\varepsilon}_{i-1} + \gamma_2 \hat{\varepsilon}_{i-2} + \ldots + \gamma_m \hat{\varepsilon}_{i-m} + v_i, \quad (4.32)$$

gdje se testira značajnost za varijable $\hat{\varepsilon}_{i-1}$, $\hat{\varepsilon}_{i-2}$, … $\hat{\varepsilon}_{i-m}$. Dakle, testira se postojanje autokorelacije procesa $\hat{\varepsilon}_i$ do zaključno pomaka *m*. Nulta hipoteza testa glasi:

$$H_0: \gamma_1 = \gamma_2 = \ldots = \gamma_m = 0, \quad (4.33)$$

što čitamo da ne postoji autokorelacija među članovima procesa grešaka relacije do zaključno reda *m*. To znači da u suštini ispitujemo skupnu značajnost varijabli $\hat{\varepsilon}_{i-1}$, $\hat{\varepsilon}_{i-2}$, … $\hat{\varepsilon}_{i-m}$ u modelu (4.32). Ako sve varijable nisu značajne, ne odbacujemo nultu hipotezu. Alternativna hipoteza pretpostavlja suprotno: postoji barem jedan koeficijent γ_j različit od nule:

$$H_1: \exists\, \gamma_j \neq 0, j \in \{1, 2, \ldots, m\}. \quad (4.34)$$

Breusch (1970) i Godfrey (1970) su dokazali da koeficijent determinacije iz procijenjenog modela (4.32), R^2_{pom}, pomnožen s vrijednošću *N*–*m* (broj opažanja u pomoćnoj regresijskoj jednadžbi, dobiven da se od ukupnog broja opažanja umanji broj pomaka varijable $\hat{\varepsilon}_i$ u pomoćnoj jednadžbi jer toliko opažanja gubimo) slijedi hi-kvadrat distribuciju s *m* stupnjeva slobode. Stoga je empirijska test veličina sljedeća:

$$LM = (N-m)R^2_{pom} \sim \chi^2(m). \quad (4.35)$$

Ako je test veličina u (4.35) veća od teorijske razine uz zadanu razinu značajnosti i *m* stupnjeva slobode, nulta hipoteza se odbacuje (ekvivalentno, ako je *p*-vrijednost manja od zadane razine značajnosti, odbacuje se H_0).

Postavlja se pitanje odabira vrijednosti *m*. U praksi se za vremenske nizove koriste sljedeći pomaci: ako se radi o mjesečnim podacima, uzima se 6, 12 ili 24 mjeseca. Za kvartalne se mogu uzeti pomaci 4, 8 ili 12. Za dnevne se uzima 15 ili 30 dana.

Za vremenske nizove se često koristi i **Ljung-Box test (1987)**, koji su definirali sljedeću test veličinu:

$$Q = N(N+2)\sum_{i=1}^{m} \frac{\hat{\rho}_i^2}{(N-i)} \sim \chi^2(m), \quad (4.36)$$

jer su pokazali da slijedi hi-kvadrat distribuciju s *m* stupnjeva slobode, a $\hat{\rho}_i$ predstavlja procjenu koeficijenta autokorelacije reda *i*. Nulta i alternativna hipoteza testa su sljedeće:

$$H_0: \rho_1 = \rho_2 = \ldots = \rho_m = 0$$
$$H_1: \exists\ \rho_j \neq 0, j \in \{1, 2, \ldots, m\}, \quad (4.37)$$

odnosno nulta hipoteza pretpostavlja da su svi koeficijenti autokorelacije do zaključno reda *m* jednaki nula, dok alternativna pretpostavlja suprotno (postoji neki koeficijent autokorelacije reda *i* koji je različit od vrijednosti nula). Ako je test veličina u (4.37) veća od teorijske razine uz zadanu razinu značajnosti i *m* stupnjeva slobode, nulta hipoteza se odbacuje (ekvivalentno, ako je *p*-vrijednost manja od zadane razine značajnosti, odbacuje se H_0).

4.2.3. Ublažavanje/uklanjanje problema autokorelacije grešaka relacije

Ako se utvrdi problem autokorelacije grešaka relacije, sljedeći su postupci ublažavanja ili uklanjanja autokorelacije:

- U slučaju da postoji autokorelacija prvog reda, originalne varijable u modelu mogu se transformirati prvim diferencijama: $\Delta y_i = y_i - y_{i-1}$, $\Delta x_{ik} = x_{ik} - x_{i-1k}$ i model se procijeni nad diferenciranim varijablama. Međutim, ovdje može biti problem ako ekonomska teorija nalaže korištenje varijabli u razinama (y_i), a ne u diferencijama.
- Primjena generalizirane metode najmanjih kvadrata (vidjeti naslov 4.5.1.1.).
- Cochrane-Orcuttov (1949) postupak.

Opis Cochrane-Orcuttovog postupka je kako slijedi. Za regresijski model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \tag{4.38}$$

se pretpostavi određeno autokorelacijsko ponašanje za varijablu ε_i, primjerice: $\varepsilon_i = \rho\varepsilon_{i-1} + v_i$ kao što je pretpostavljeno u (4.23). Model vrijedi za bilo koje opažanje i, pa pišemo:

$$y_{i-1} = \beta_0 + \beta_1 x_{i-1,1} + \beta_2 x_{i-1,2} + \ldots + \beta_k x_{i-1,k} + \varepsilon_{i-1}, \tag{4.39}$$

te ako se sada u modelu (4.39) lijeva i desna strana pomnože s ρ:

$$\rho y_{i-1} = \rho\beta_0 + \rho\beta_1 x_{i-1,1} + \rho\beta_2 x_{i-1,2} + \ldots + \rho\beta_k x_{i-1,k} + \rho\varepsilon_{i-1}, \tag{4.40}$$

te se (4.40) oduzme od (4.38):

$$y_i - \rho y_{i-1} = \beta_0 - \rho\beta_0 + \beta_1 x_{i1} - \rho\beta_1 x_{i-1,1} + \beta_2 x_{i2} - \rho\beta_2 x_{i-1,2} + \ldots + \beta_k x_{ik} - \rho\beta_k x_{i-1,k} + \varepsilon_i - \rho\varepsilon_{i-1}, \tag{4.41}$$

$$y_i - \rho y_{i-1} = (1-\rho)\beta_0 + \beta_1(x_{i1} - \rho x_{i-1,1}) + \beta_2(x_{i2} - \rho x_{i-1,2}) + \ldots + \beta_k(x_{ik} - \rho x_{i-1,k}) + v_i, \tag{4.42}$$

pa je

$$y_i^* = \beta_0^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \ldots + \beta_k x_{ik}^* + v_i, \tag{4.43}$$

ako se uvedu supstitucije: $y_i^* = y_i - \rho y_{i-1}$, $x_{i1}^* = (x_{i1} - \rho x_{i-1,1})$, $x_{i2}^* = (x_{i2} - \rho x_{i-1,2})$, …, $x_{ik}^* = (x_{ik} - \rho x_{i-1,k})$. U modelu (4.43) sada varijabla v_i ne sadrži autokorelaciju, stoga se može procijeniti metodom najmanjih kvadrata. No, potrebno je znati vrijednost ρ. Stoga su sljedeći koraci: u prvom koraku se model (4.38) procjenjuje metodom najmanjih kvadrata, pri čemu se zanemaruje problem autokorelacije. Nakon toga se iz modela estrahiraju rezidualna odstupanja i procijeni se model: $\hat{\varepsilon}_i = \rho\,\hat{\varepsilon}_{i-1} + v_i$, kako bi se procijenila vrijednost $\hat{\rho}$. Tada se konstruiraju varijable $y_i^* = y_i - \hat{\rho} y_{i-1}$, $x_{i1}^* = (x_{i1} - \hat{\rho} x_{i-1,1})$, $x_{i2}^* = (x_{i2} - \hat{\rho} x_{i-1,2})$, …, $x_{ik}^* = (x_{ik} - \hat{\rho} x_{i-1,k})$ i konačno, procijeni se model (4.43). Međutim, postoje određene pretpostavke o parametrima u spomenutim modelima, koje moraju biti zadovoljene da bi se ova procedura provela (vidjeti Cochrane i Orcutt,1949; ili Brooks, 2014).

4.2.4. Primjer

Učitajte datoteku „**stanovi.txt**“ iz prethodnog primjera (naslov 4.1.4.). Za procijenjeni model iz tog primjera provedite testiranje autokorelacije prvog reda pomoću Durbin-Watson testa (dvosmjerni test te oba jednosmjerna), zapišite hipoteze svakog testa, test veličinu, približnu vrijednost koeficijenta autokorelacije prvog reda te donesite zaključak. Provedite testiranje autokorelacije do zaključno trećeg reda pomoću Breusch-Godfrey testa: zapišite pomoćnu regresijsku jednadžbu, hipoteze testa, test veličinu i donesite zaključak. Provedite testiranje autokorelacije do zaključno trećeg reda pomoću Ljung-Box testa: zapišite hipoteze testa, test veličinu i donesite zaključak. Razina značajnosti je 5%.

Sumarno su hipoteze Durbin-Watsonova testa prikazane u tablici 4.1., koju ćemo sada proširiti s test veličinom DW, kao i procjenom koeficijenta autokorelacije prvog reda:

Hipoteza/test	Dvosmjerni	Na gornju granicu	Na donju granicu
H_0	$\rho = 0$	$\rho = 0$	$\rho = 0$
H_1	$\rho \neq 0$	$\rho > 0$	$\rho < 0$
DW	1,97		
$\hat{\rho}$	0,006		

Vrijednosti u retcima DW i $\hat{\rho}$ su popunjene temeljem ispisa na slici 4.10. Za provedbu DW testa, potrebno je uključiti paket car (engl. *compantion to applied regression*) i koristiti naredbu durbinWatsonTest(). Ako se ne navodi pretpostavka alternativne hipoteze, tada se test provodi kao dvosmjerni. U tom slučaju u okviru ispisa piše „rho!=0“ što čitamo $\rho \neq 0$. Za test na gornju granicu pišemo kao alternativnu hipotezu „positive“, a za test na donju granicu pišemo „negative“ (pozitivna ili negativna autokorelacija prvog reda). Kako DW vrijednost iznosi 1,97, procjena koeficijenta autokorelacije prvog reda iznosi 0,006, što je veoma blizu vrijednosti 0. Stoga se u sva tri testa ne može odbaciti nulta hipoteza da ne postoji autokorelacija rezidualnih odstupanja prvog reda. Dodatno, u okviru testa se računa *p*-vrijednost temeljem koje sa također može donositi zaključak na uobičajenim razinama značajnosti.

```
library(car)
durbinWatsonTest(model)

##  lag Autocorrelation D-W Statistic p-value
##    1     0.006259401      1.970795   0.922
##  Alternative hypothesis: rho != 0

durbinWatsonTest(model,alternative = "positive")

##  lag Autocorrelation D-W Statistic p-value
##    1     0.006259401      1.970795   0.432
##  Alternative hypothesis: rho > 0

durbinWatsonTest(model,alternative = "negative")

##  lag Autocorrelation D-W Statistic p-value
##    1     0.006259401      1.970795   0.551
##  Alternative hypothesis: rho < 0
```

Slika 4.10. Durbin-Watson test

Nadalje, da bi se proveo Breusch-Godfreyev test, koristi se paket quantmod, kako bi se najprije definirali pomaci rezidualnih odstupanja iz originalnog modela (naredba Lag, vidjeti sliku 4.11.). Nakon definiranja prvog, drugog i trećeg pomaka rezidualnih odstupanja, procjenjuje se

model (4.32), gdje se rezidualno odstupanje $\hat{\varepsilon}_i$ regresira na četiri nezavisne varijable iz originalnog modela, te prethodna tri pomaka (ispis na slici 4.11.).

```
library(quantmod)
reziduali<-residuals(model)
rez1<-Lag(reziduali,1);rez2<-Lag(reziduali,2);rez3<-Lag(reziduali,3)
summary(pomocna<-lm(reziduali~kvadrat+sobe+godine+udaljenost+rez1+rez2+rez3,data=stanovi))

## Call:
## lm(formula = reziduali ~ kvadrat + sobe + godine + udaljenost +
##     rez1 + rez2 + rez3, data = stanovi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2809  -4.9404   0.4371   4.2712  12.3717
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.232455   3.156806  -0.074    0.942
## kvadrat     -0.031593   0.133236  -0.237    0.814
## sobe         0.587398   2.748807   0.214    0.832
## godine       0.010039   0.121023   0.083    0.934
## udaljenost  -0.003460   0.217382  -0.016    0.987
## rez1        -0.001735   0.162590  -0.011    0.992
## rez2         0.086990   0.166745   0.522    0.605
## rez3        -0.024371   0.164186  -0.148    0.883
##
## Residual standard error: 6.449 on 39 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.008903,   Adjusted R-squared:  -0.169
## F-statistic: 0.05005 on 7 and 39 DF,  p-value: 0.9998
```

Slika 4.11. Provedba Breusch-Godfrey testa

Pomoćna regresijska jednadžba je sljedeća:

$$\hat{\varepsilon}_i = -0.23 - 0{,}03x_{i1} + 0{,}59\,x_{i2} + 0{,}01x_{i3} - 0{,}03x_{i4} - 0{,}02\,\hat{\varepsilon}_{i-1} + 0{,}09\,\hat{\varepsilon}_{i-2} - 0{,}02\,\hat{\varepsilon}_{i-3}.$$

```
test_vel<-nobs(pomocna)*summary(pomocna)$r.squared
test_vel

## [1] 0.4184306

p_v<-1-pchisq(test_vel,3)
p_v

## [1] 0.9364094
```

4.12. Izračun empirijske veličine Breusch-Godfrey testa i pripadajuće *p*-vrijednosti

Hipoteze testa su sljedeće: $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$, $H_1 : \exists\ \gamma_j \neq 0, j \in \{1, 2, 3\}$. Test veličina računa se kao: $LM = (N-m)R^2_{pom} = 47 \cdot 0{,}0089 = 0{,}4184$, s *p*-vrijednošću 0,936 (izračun vidjeti na slici 4.12., *p*-vrijednost se računa temeljem *LM* vrijednosti koja slijedi hi-kvadrat distribuciju s 3 stupnja slobode), koja je veća od 0,05, stoga se ne odbacuje nulta hipoteza. Riječima: uz razinu značajnosti od 5%, ne odbacujemo hipotezu da ne postoji problem autokorelacije rezidualnih odstupanja do zaključno trećeg reda.

Konačno, može se provesti i Ljung-Boxov test, temeljem naredbi prikazanih na slici 4.13. Hipoteze testa su sljedeće: $H_0 : \rho_1 = \rho_2 = \rho_3 = 0$, $H_1 : \exists\ \rho_j \neq 0, j \in \{1, 2, 3\}$. Test veličina iznosi

Q = 0,312, koja slijedi hi-kvadrat distribuciju s 3 stupnja slobode, te pripadajućom p-vrijednošću 0,96. Kako je to veće od uobičajenih razina značajnosti, ne odbacujemo nultu hipotezu. Odnosno, uz razinu značajnosti od 5%, ne odbacujemo hipotezu da ne postoji problem autokorelacije rezidualnih odstupanja do zaključno trećeg reda.

```
Box.test(reziduali,lag=3,type="Ljung-Box")

##
##  Box-Ljung test
##
## data:  reziduali
## X-squared = 0.31157, df = 3, p-value = 0.9578
```

Slika 4.13. Ljung-Box test autokorelacije

4.3. Heteroskedastičnost grešaka relacije

Još jedan učestali problem na koji se nailazi u praksi jest promjenjivost varijance greške relacije ili heteroskedastičnost. Ako se ne želi modelirati ta promjenjivost, u nastavku su opisani načini ublažavanja ovog problema nakon obrade njegova testiranja.

4.3.1. Definiranje problema heteroskedastičnosti grešaka relacije

Kao četvrta pretpostavka u modelu linearne regresije (vidjeti naslove 2.1.2. i 2.2.2.) navedena je homoskedastičnost varijance greške relacije. Drugim riječima, nepromjenjiva je ili je konstantna : $Var(\varepsilon_i) = Var(\varepsilon_i \mid x_i) = \sigma^2, \forall i$. Pretpostavke odsustva autokorelacije i nepromjenjivosti varijance slučajne varijable matrično smo pisali na način:

$$\Omega = E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I \tag{4.44}$$

gdje se sada fokusiramo na elemente na glavnoj dijagonali. Međutim, ako matrica u (4.44) nije skalarna, nego dijagonalna, odnosno ako je:

$$E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix} \tag{4.45}$$

tj. $\boldsymbol{Var(\varepsilon_i) = \sigma_i^2}$**, kažemo da postoji problem heteroskedastičnosti grešaka relacije**, odnosno varijanca greške relacije je promjenjiva, stoga se dodaje indeks i. U tom slučaju, ako se razmotri izračun varijance procjenitelja, kada ne vrijedi više $Var(\boldsymbol{\varepsilon}) = \sigma^2 I$, nego $Var(\boldsymbol{\varepsilon}) = \sigma_i^2 I$, tada je $Var(\hat{\boldsymbol{\beta}})$ sljedeća (vidjeti izvod od (2.48) do (2.56)):

$$Var(\hat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')}_{Var(\varepsilon)\neq\sigma^2 I} \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Z}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{4.46}$$

pa je matrica varijanci-kovarijanci u (4.46) različita od one u (2.59), tj. vrijedi:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Z}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \neq \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{4.47}$$

odnosno u prisustvu heteroskedastičnosti grešaka relacije, matrica varijanci-kovarijanci procjenitelja **više nije jednaka** $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. To znači da **izračun empirijskih *t*-omjera više nije pouzdan, kao i što intervalne procjene također nisu pouzdane** (jer se temelje na standardnim pogreškama procjenitelja). Dakle, problemi koji se javljaju su veoma slični onima za problem autokorelacije grešaka relacije. Stoga **i *F*-test nije pouzdan**, jer se procjena varijance regresije računa na sljedeći način (vidjeti (2.205)):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N-k-1} = \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{N-k-1} = \frac{SSR}{N-k-1}, \tag{4.48}$$

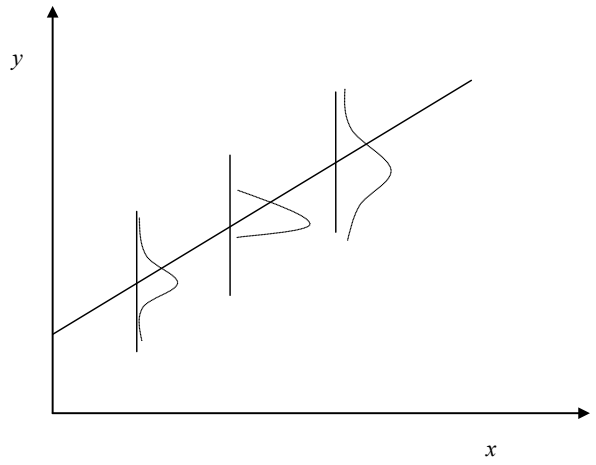
te se upravo ona koristi u izračunu empirijskog *F*-omjera (vidjeti (2.213)):

$$F = \frac{SSE/k}{SSR/(N-k-1)}, \tag{4.49}$$

tako se i u slučaju $Var(\boldsymbol{\varepsilon}) \neq \sigma^2 I$ dolazi do drugačije vrijednosti *SSR* u (4.49), pa time i empirijski *F*-omjer postaje nepouzdan.

Heteroskedastičnost ne utječe na nepristranost procjenitelja. Naime, kako se on računa kao $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, uočava se da se u tome izrazu ne koristi matrica varijanci-kovarijanci procjenitelja. Međutim, **utječe na efikasnost procjenitelja**, jer je narušeno svojstvo najmanje varijance procjenitelja u prisustvu ovoga problema.

Grafički bismo mogli problem heteroskedastičnosti predočiti na sljedeći način (vidjeti sliku 4.14). Promjenjivost varijance znači da će se varijabilnost podataka mijenjati u ovisnosti o opažanju *i* koje promatramo. Distribucija varijable *y* je centrirana oko očekivane vrijednosti varijable *y* uz dane vrijednosti varijable *x*, ali je varijabilnost te distribucije promjenjiva.



Slika 4.14. Problem heteroskedastičnosti greške relacije grafički

Mogući **razlozi postojanja heteroskedastičnosti** grešaka relacije su:

- Postojanje netipičnih vrijednosti u podacima (engl. *outlier*), koje su jako velike ili male naspram većine preostalih podataka, stoga utječu na procjenu varijance.
- Pogrešna specifikacija modela: neuključivanje značajnih varijabli u model, kriva transformacija varijabli prije korištenja u modeliranju ili pak krivo odabrana funkcijska forma modela.
- Asimetričnosti distribucija varijabli koje se koriste u modeliranju.

4.3.2. Utvrđivanje postojanja problema heteroskedastičnosti grešaka relacije

Nekoliko načina utvrđivanja postojanja ovoga problema su sljedeći: grafički način i formalno testiranje. Ako se razmatra **grafički način**, onda je potrebno iz procijenjenog modela procijeniti rezidualna odstupanja, tj. njihove kvadrate, kao i procijenjenu vrijednost zavisne varijable. Na os apscisa nanose se procijenjene vrijednosti zavisne varijable, a na os ordinata kvadrirane vrijednosti rezidualnih odstupanja (vidjeti sliku 4.15.). U slučaju postojanja heteroskedastičnosti, uočava se da povećanjem procijenjenih vrijednosti zavisne varijable dolazi do povećanja kvadrata rezidualnih odstupanja. Osim ovakvog ponašanja, moguće je uočiti i bilo kakvo sustavno ponašanje kvadrata rezidualnih odstupanja u ovisnosti o promjeni procijenjene vrijednosti zavisne varijable.

Formalni testovi heteroskedastičnosti su: Breusch-Paganov LM test, Whiteov test, Goldfeld-Quandtov test kao poznatiji, te potom RESET test Anscombea (1961), Glejserov (1969) test.

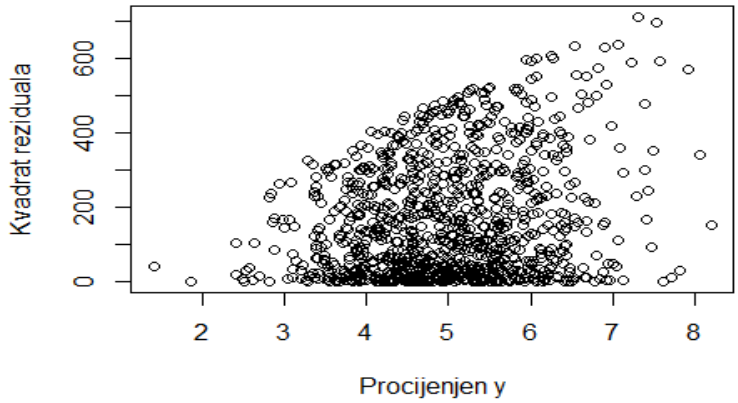
Breusch-Pagan LM test se temelji na pomoćnoj regresijskoj jednadžbi u kojoj se kvadrati rezidualnih odstupanja iz početnog modela procijene u ovisnosti o nekim varijablama z_j. Kako su općenito nepoznate varijable koje utječu na kvadrate rezidualnih odstupanja, koriste se nezavisne varijable x_j iz početnog modela:

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_k x_{ik} + \eta_i, \tag{4.50}$$

te se hipoteze formiraju na sljedeći način:

$$\begin{aligned} &H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0 \\ &H_1 : \exists \alpha_i \neq 0,\ i \in \{1, 2, ..., k\} \end{aligned}, \tag{4.51}$$

odnosno u nultoj hipotezi se pretpostavlja homoskedastičnost varijance rezidualnih odstupanja jer ako je H_0 istinita, tada procijenjeni model u (4.50) glasi: $\hat{\varepsilon}_i^2 = \hat{\alpha}_0$ i varijanca je nepromjenjiva, jer je $\hat{\alpha}_0$ konstanta. S druge strane, alternativna hipoteza pretpostavlja promjenjivost varijance koja ovisi o nekoj (ili svakoj) varijabli x_j. Test veličina temelji se na vrijednosti $\frac{\varepsilon_i^2}{\sigma^2}$. Ako je nulta hipoteza istinita, tada slučajna varijabala $\frac{\varepsilon_i^2}{\sigma^2}$ ima hi-kvadrat distribuciju s jednim stupnjem slobode, a njena varijanca $Var\left(\frac{\varepsilon_i^2}{\sigma^2}\right) = 2$ (jer je varijanca hi-kvadrat distribucije jednaka 2·broj stupnjeva slobode).



Slika 4.15. Grafičko utvrđivanje postojanja heteroskedastičnosti

Stoga vrijedi:

$$Var\left(\frac{\varepsilon_i^2}{\sigma^2}\right) = 2 \Rightarrow \frac{Var\left(\varepsilon_i^2\right)}{\sigma^4} = 2 \Rightarrow Var\left(\varepsilon_i^2\right) = 2\sigma^4 . \qquad (4.52)$$

Za velike uzorke vrijedi: $Var\left(\varepsilon_i^2\right) \approx Var\left(\hat{\varepsilon}_i^2\right)$ te $\sigma^2 \approx \hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N}$, pa je $Var\left(\hat{\varepsilon}_i^2\right) = 2\hat{\sigma}^4$, stoga je test veličina Breusch-Pagan LM testa jednaka:

$$LM = \frac{SSP_{pom}}{2\hat{\sigma}^4} = N \cdot R_{pom}^2 \sim \chi^2(k) , \qquad (4.53)$$

tj. to je omjer protumačene (SSP_{pom}) i neprotumačene sume kvadrata ($2\hat{\sigma}^4$) modela u (4.50), te je R^2_{pom} koeficijent determinacije pomoćne regresijske jednadžbe. Test veličina LM slijedi hi-kvadrat distribuciju s k stupnjeva slobode (broj nezavisnih varijabli uključenih u pomoćnu regresijsku jednadžbu). Ako je $LM > \chi_\alpha^2(k)$, nulta hipoteza se odbacuje (ili ekvivalentno, ako je pripadajuća p-vrijednost < α), dok se za slučaj $LM < \chi_\alpha^2(k)$ nulta hipoteza ne odbacuje (ili ako je p-vrijednost > α).

Nedostatak ovog testa je taj što se pretpostavlja da greške relacije slijede normalnu distribuciju, pretpostavlja se da su unaprijed poznati z_j, kao i oblik heteroskedastičnosti (linearni model (4.50)).

Whiteov test, s druge strane, nema spomenute nedostatke. Također se pretpostavlja u nultoj hipotezi da ne postoji problem heteroskedastičnosti rezidualnih odstupanja. No, sama pomoćna regresijska jednadžba se sada sastoji od kvadrata rezidualnih odstupanja iz originalnog modela kao zavisne varijable, no kao nezavisne varijable se uključuju nezavisne varijable iz početnog modela, njihovi kvadrati, kao i međusobni umnošci, jer se ne specificira točan oblik heteroskedastičnosti. Primjerice, pomoćna regresijska jednadžba za slučaj dvije nezavisne varijable iz početnog modela glasi:

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i1}^2 + \alpha_4 x_{i2}^2 + \alpha_5 x_{i1} x_{i2} + \eta_i, \qquad (4.54)$$

stoga u ovom slučaju nulta hipoteza glasila: H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_5 = 0$.

Test veličina je $W = N \cdot R^2_{pom} \sim \chi^2(r)$, gdje je r broj regresorskih varijabli u pomoćnoj regresijskoj jednadžbi (uključujući i kvadrate i i međusobne umnoške varijabli uz same regresijske varijable). Ako je $W > \chi_\alpha^2(r)$, nulta hipoteza se odbacuje (ili ekvivalentno, ako je pripadajuća p-vrijednost < α), dok se za slučaj $W < \chi_\alpha^2(r)$ nulta hipoteza ne odbacuje (ili ako je p-vrijednost > α).

Goldfeld-Quandtov test temelji se na podjeli uzorka na dva jednakobrojna dijela, pri čemu se za svaki poduzorak procijeni originalni regresijski model, te varijance regresije. Pritom se preporuča da se iz testa isključi c središnjih vrijednosti[43]. One koje preostaju se dijele na dva poduzorka koji imaju jednak broj opažanja[44]. Ako postoji razlika u disperziji oko regresijskog pravca (slika 4.14.) za poduzorke, tada će se varijance tih dviju regresija razlikovati. Stoga su nulta i alternativna hipoteza sljedeće[45]:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1; \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1, \tag{4.55}$$

gdje σ_1^2 predstavlja varijancu regresije prvog poduzorka, a σ_2^2 drugog. Nulta hipoteza pretpostavlja da im je omjer jednak 1 jer u tom slučaju nema heteroskedastičnosti (ne dolazi do promjene varijance s obzirom na poduzorak). S druge strane, ako dolazi do značajne promjene, taj omjer će biti veći od 1 (alternativna hipoteza, heteroskedastičnost). Test veličina ovoga testa je F-omjer:

$$F = \frac{SSR_1}{SSR_2} \sim F(ss_1, ss_2), \tag{4.56}$$

gdje je rezidualna suma kvadrata prvog poduzorka u brojniku, a drugog u nazivniku, pri čemu je broj stupnjeva slobode u brojniku (ss_1) jednak veličini prvog poduzorka umanjenog za broj parametara u regresijskom modelu: $n_1 - k - 1$, dok je broj stupnjeva slobode u nazivniku (ss_2) jednak veličini prvog poduzorka umanjenog za broj parametara u regresijskom modelu: $n_2 - k - 1$. Kako oba poduzorka imaju jednak broj opažanja, $n_1 = n_2$.

Nulta hipoteza se odbacuje ako je $F > F_\alpha(ss_1, ss_2)$, dok se ne odbacuje ako vrijedi $F < F_\alpha(ss_1, ss_2)$. Naravno, odluka se može donijeti i usporedbom p-vrijednosti s razinom značajnosti α.

Anscombeov RESET test se temelji na Ramseyevom testu koji je obrađivan u pododjeljku 2.2.7.8. Ideja je da se rezidualna odstupanja iz originalnog modela regresiraju na kvadrat procijenjene vrijednosti zavisne varijable, njezin kub, i više potencije te se provodi test značajnosti varijabli u modelu. **Glejserov test** se odnosi na procjenu modela u kojemu apsolutna vrijednost rezidualnih odstupanja iz originalnog modela ovisi o nezavisnoj varijabli x, pri čemu se može procijeniti linearni model, pretpostaviti recipročna veza između rezidualnih odstupanja i nezavisne varijable, itd. pri čemu se testira značajnost nezavisne varijable u tom modelu (Maddala i Lahiri, 2014).

[43] Obično se uzima da je $c = N/4$, odnosno da se od ukupnog broja opažanja ukupno četvrtina središnjih podataka izostavi.

[44] Iako nije moguće uvijek postići da je $N/4$ prirodan broj, moguće je dobiveni rezultat zaokružiti na prirodan broj kako bi se uzorak mogao podijeliti.

[45] U slučaju alternativne hipoteze u (4.55) pretpostavlja se da je varijanca regresije prvog poduzorka veća od varijance regresije drugog poduzorka. No, može se pretpostaviti i suprotno, pa će omjer u H_1 biti < 1, a može se pretpostaviti i općenito da je omjer različit od 1, pogotovo kada istraživač ne zna za koji poduzorak bi varijanca regresije mogla biti veća. Vidjeti Verbeek (2004) detaljnije.

4.3.3. Ublažavanje/uklanjanje problema heteroskedastičnosti grešaka relacije

U slučaju postojanja problema heteroskedastičnosti grešaka relacije postoji nekoliko pristupa ublažavanja ili uklanjanja ovog problema:

- Ako je poznat oblik heteroskedastičnosti, može se primijeniti vagana metoda najmanjih kvadrata (vidjeti pododjeljak 4.5.1.2.) ili metoda najveće vjerodostojnosti (jer se modelira poznat oblik promjenjivosti varijance)
- Ako je nepoznat oblik heteroskedastičnosti, u tom slučaju se varijable logaritmiraju, kako bi se smanjila varijabilnost podataka, varijable se mogu deflacionirati (pomoću neke varijable koja utječe na početnu varijabilnost)
- Dodatno, mogu se korigirati standardne pogreške procjenitelja kako bi postale robusne: koriste se Whiteova korekcija (kada je heteroskedastičnost nepoznatog oblika, ali ne postoji autokorelacija) ili Newey-West korekcija (općenitija, u slučaju nepoznatog oblika heteroskedastičnosti, ali i autokorelacije).

4.3.3.1. Whiteova korekcija standardnih pogrešaka procjenitelja

Whiteova (1980) korekcija sastoji se od dva koraka. U prvome se procijeni matrica varijanci-kovarijanci rezidualnih odstupanja:

$$\hat{\Omega} = \begin{bmatrix} \hat{\varepsilon}_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \hat{\varepsilon}_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \hat{\varepsilon}_3^2 & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{\varepsilon}_N^2 \end{bmatrix}, \tag{4.57}$$

te se potom matrica varijanci-kovarijanci procjenitelja procijeni tako da se procjena $\widehat{\boldsymbol{\Omega}}$ u (4.57) uvrsti u izraz (4.46):

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Z}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{4.58}$$

te se korigira za broj stupnjeva slobode da bi se procijenila matrica varijanci i kovarijanci procjenitelja:

$$Var(\widehat{\boldsymbol{\beta}}) = \frac{N}{N-k}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{\Omega}}\,\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{4.59}$$

te se upravo ta matrica koristi za provođenje *t*-testa (za standardne pogreške procjenitelja). Dakle, uočava se kako se Whiteova korekcija koristi u slučaju postojanja heteroskedastičnosti, ali odsustva autokorelacije grešaka relacije. White (1980) je dokazao da je $\boldsymbol{X}'\widehat{\boldsymbol{\Omega}}\boldsymbol{X} = \boldsymbol{X}'\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}'\boldsymbol{X}$ konzistentan procjenitelj od $\boldsymbol{X}'E(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}')\boldsymbol{X}$, ali je pristran, stoga je ova korekcija primjenjiva u slučaju velikih uzoraka. Stoga bi se matrica u (4.59) koristila dalje za provedbu *t*-testa u okviru regresijskog modela.

4.3.3.2. Newey-West korekcija standardnih pogrešaka procjenitelja

Newey-West (1987) korekcija koristi se u slučaju kada postoji i autokorelacija i heteroskedastičnost greške relacije u modelu (engl. *HAC – heteroskedasticity and autocorrealtion correction*). U okviru ove korekcije je potrebno odrediti red autokorelacije prije

same procjene matrice varijanci i kovarijanci rezidualnih odstupanja. U ovome slučaju se koristi matrica varijanci i kovarijanci rezidualnih odstupanja dugog roka (engl. *long run*) ili slučaju presječnih podataka velikog *N*, tako da se procijeni:

$$\hat{\Omega} = \hat{\Gamma}_0 + \sum_{j=1}^{m}\left(1 - \frac{j}{m+1}\right)\left[\hat{\Gamma}_j + \hat{\Gamma}_{-j}\right], \quad \hat{\Gamma}_j = \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_i\hat{\varepsilon}_{i-j}\boldsymbol{X}_i\boldsymbol{X}'_{i-j} \tag{4.60}$$

gdje *m* predstavlja red autokorelacije koja se pretpostavlja, a $\boldsymbol{X}_i$ redak *i* u matrici $\boldsymbol{X}$. Sada je moguće procijeniti matricu varijanci-kovarijanci procjenitelja na način da se izraz (4.60) uvrsti u (4.46) i korigira se faktorom *N*:

$$Var(\hat{\boldsymbol{\beta}}) = N(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{\Omega}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{4.61}$$

Nakon toga se matrica u (4.61) koristi dalje za provedbu *t*-testa u okviru regresijskog modela.

4.3.4. Primjer

Učitajte datoteku „stanovi.txt“ iz prethodnih primjera (odjeljci 4.1.4. i 4.2.4.). Za procijenjeni model iz tog primjera provedite testiranje heteroskedastičnosti greške relacije, pri razini značajnosti od 5%. Pritom provedite i Breusch-Paganov test, kao i Whiteov test. Zapišite pomoćne regresijske jednadžbe temeljem kojih se testovi provode, hipoteze svakog testa, kao i test veličine. Interpretirajte rezultate. Neovisno o ishodu testova, provedite Whiteovu korekciju standardnih pogrešaka procjenitelja, kao i Newey-Westovu korekciju, te potom usporedite rezultate *t*-testova za slučaj originalnog modela, kao i obiju korekcija.

Temeljem naredbi i ispisa danih na slici 4.16., može se provesti Breusch-Paganov test na sljedeći način. Pomoćna regresijska jednadžba je $\hat{\varepsilon}_i^2 = 12{,}39 - 0{,}79x_{i1} + 19{,}01x_{i2} + 0{,}21x_{i3} + 0.66x_{i4}$. Hipoteze testa su: H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, H_1: $\exists\, \alpha_j \neq 0, j \in \{1,2,3,4\}$, pri čemu je broj opservacija (naredba nobs(pomocna)) jednak 50, uz R^2_{pom} jednak 0,051. Stoga je test veličina jednaka $LM = N{\cdot}R^2_{pom} = 50{\cdot}0{,}051 = 2{,}57$. Pripadajuća *p*-vrijednost iznosi 0,63, izračunata temeljem 4 stupnja slobode. Kako vrijedi *p*-vrijednost > α, ne može se odbaciti nulta hipoteza da ne postoji problem heteroskedastičnosti rezidualnih odstupanja u modelu.

Nadalje, slika 4.17. predočava naredbe za provedbu i ispis Whiteova testa. Kako je u slučaju ovog testa uključen i kvadrat svake varijable, kao i međusobni umnošci, pomoćna regresijska jednadžba je sljedeća:

$$\begin{aligned}\hat{\varepsilon}_i^2 = {} & 30{,}60 + 2{,}17x_{i1} - 58{,}03x_{i2} + 5{,}44x_{i3} - 2{,}28x_{i4} + 0{,}02x_{i1}^2 + 10{,}07x_{i2}^2 - 0{,}22x_{i3}^2 + 0{,}14x_{i4}^2 \\ & -0{,}8x_{i1}x_{i2} - 0{,}2x_{i1}x_{i3} - 0{,}16x_{i1}x_{i4} + 3{,}64x_{i2}x_{i3} + 2{,}56x_{i2}x_{i4} + 0{,}26x_{i3}x_{i4}\end{aligned}$$

Hipoteze testa su: H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_{14} = 0$, H_1: $\exists\, \alpha_j \neq 0, j \in \{1,2,\ldots, 14\}$, pri čemu je broj opservacija jednak 50, uz R^2_{pom} jednak 0,0247. Stoga je test veličina jednaka $LM = N{\cdot}R^2_{pom} = 50{\cdot}0{,}0247 = 12{,}37$. Pripadajuća *p*-vrijednost iznosi 0,577, izračunata temeljem 14 stupnja slobode. Kako vrijedi *p*-vrijednost > α, ne može se odbaciti nulta hipoteza da ne postoji problem heteroskedastičnosti rezidualnih odstupanja u modelu. Dakle, temeljem oba testa ne odbacuje se nulta hipoteza o homoskedastičnosti varijance slučajne varijable. Unatoč tome, u nastavku

se prikazuje postupak provođenja Whiteove i Newey-West korekcija, u slučaju postojanja problema heteroskedastičnosti.

```
summary(bptest<-lm(residuals(model)^2~kvadrat+sobe+godine+udaljenost,data=stanovi))

## Call:
## lm(formula = residuals(model)^2 ~ kvadrat + sobe + godine + udaljenost,
##     data = stanovi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.832 -23.290  -8.723  14.187 121.484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.3898    18.1963   0.681    0.499
## kvadrat      -0.7793     0.7747  -1.006    0.320
## sobe         19.0070    16.0267   1.186    0.242
## godine        0.2025     0.6684   0.303    0.763
## udaljenost    0.6638     1.2507   0.531    0.598
##
## Residual standard error: 38.48 on 45 degrees of freedom
## Multiple R-squared:  0.0514, Adjusted R-squared:  -0.03292
## F-statistic: 0.6096 on 4 and 45 DF,  p-value: 0.6578

nobs(bptest);(summary(bptest))$r.squared

## [1] 50

## [1] 0.05140362

nobs(bptest)*(summary(bptest))$r.squared

## [1] 2.570181

p_vrijednost<-1-pchisq(nobs(bptest)*(summary(bptest))$r.squared,4)
p_vrijednost

## [1] 0.6321144
```

Slika 4.16. Breusch-Paganov test heteroskedastičnosti

Naredbe za provođenje obiju korekcija prikazane su na slikama 4.18. i 4.19. Kako bi se provela Whiteova korekcija pogrešaka, koristi se paket car te naredba hccm(), u okviru koje se koristi izraz „hc0“, što je naredba upravo za Whiteovu korekciju. Najprije je na slici 4.18. prikazana matrica varijanci i kovarijanci procjenitelja, dobivena temeljem formule (4.59), koja se sada može koristiti dalje u provođenju *t*-testova pojedinačne značajnosti svih varijabli u modelu. Test se provodi pomoću paketa lmtest i naredbe coeftest(…), gdje je potrebno za matricu varijanci-kovarijanci procjenitelja unijeti upravo dobivenu matricu (označena je s mat). U zadnjem dijelu ispisa (naziv t test of coefficients) se provode pojedinačni testovi značajnosti, ali sada se stupac pod nazivom „Std. Error“ razlikuje od onoga koji se dobije u ispisu samo uz procjenu regresijskog modela pomoću naredbe lm, jer su sada upravo pogreške procjenitelja korigirane odabranim postupkom. To znači da će se stupac s empirijskim *t*-omjerima (t value) razlikovati u odnosu na originalni ispis modela, i sukladno tome, pripadajuće *p*-vrijednosti. Ako se usporede vrijednosti dobivene sada uz korekciju na slici 4.18., s onima na slici 4.4., uočava se kako je došlo do promjena empirijskih *t*-omjera, međutim zaključci o (ne)odbacivanju nultih hipotezi ostaju nepromjenjivi. Jedino varijabla sobe i dalje ostaje neznačajna u modelu.

Nadalje, za provedbu Newey-West korekcije, koristi se paket sandwich te naredba NeweyWest(…), vidjeti sliku 4.19. Sada se ponovno sprema nova matrica varijanci-kovarijanci procjenitelja (u ispisu mat2), kako bi se upravo ona koristila u slučaju ove korekcije za provedbu *t*-testova. Ponovno dolazi do promjena u vrijednostima u stupcu „Std. Error“, „t value“ i „Pr(>|t|)“. Međutim, zaključci o (ne)odbacivanju nulte hipoteze i dalje ostaju nepromijenjeni. Napomenimo kako smo pretpostavili da je autokorelacija prisutna u modelu do zaključno 2. reda (u naredbi NeweyWest() je lag=2).

```
summary(white<-lm(residuals(model)^2~
                    kvadrat+sobe+godine+udaljenost
                  +I(kvadrat^2)+I(sobe^2)+I(godine^2)+I(udaljenost^2)
                  +I(kvadrat*sobe)+I(kvadrat*godine)+I(kvadrat*udaljenost)
                  +I(sobe*godine)+I(sobe*udaljenost)+I(godine*udaljenost)
                  ,data=stanovi))

## Call:
## lm(formula = residuals(model)^2 ~ kvadrat + sobe + godine + udaljenost +
##     I(kvadrat^2) + I(sobe^2) + I(godine^2) + I(udaljenost^2) +
##     I(kvadrat * sobe) + I(kvadrat * godine) + I(kvadrat * udaljenost) +
##     I(sobe * godine) + I(sobe * udaljenost) + I(godine * udaljenost),
##     data = stanovi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.691 -21.373  -3.651  15.247  85.288

##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               30.60218   55.66675   0.550   0.5860
## kvadrat                    2.17481    2.63676   0.825   0.4151
## sobe                     -58.03190   58.33160  -0.995   0.3266
## godine                     5.44016    3.69522   1.472   0.1499
## udaljenost                -2.28104    6.50207  -0.351   0.7278
## I(kvadrat^2)               0.02174    0.11802   0.184   0.8549
## I(sobe^2)                 10.06745   51.40239   0.196   0.8459
## I(godine^2)               -0.21858    0.11503  -1.900   0.0657 .
## I(udaljenost^2)            0.14422    0.36655   0.393   0.6964
## I(kvadrat * sobe)         -0.80112    4.90070  -0.163   0.8711
## I(kvadrat * godine)       -0.19547    0.11578  -1.688   0.1002
## I(kvadrat * udaljenost)   -0.15968    0.21289  -0.750   0.4582
## I(sobe * godine)           3.64219    2.43552   1.495   0.1438
## I(sobe * udaljenost)       2.56008    4.47629   0.572   0.5710
## I(godine * udaljenost)     0.26275    0.17227   1.525   0.1362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.86 on 35 degrees of freedom
## Multiple R-squared:  0.2473, Adjusted R-squared:  -0.05374
## F-statistic: 0.8215 on 14 and 35 DF,  p-value: 0.6421

nobs(white);(summary(white))$r.squared

## [1] 50
## 1] 0.2473302

nobs(white)*(summary(white))$r.squared
## [1] 12.36651
p_vrijednost<-1-pchisq(nobs(white)*(summary(white))$r.squared,14)
p_vrijednost
## [1] 0.5768947
```

Slika 4.17. Whiteov test heteroskedastičnosti

Konačno, u slučaju postojanja problema heteroskedastičnosti rezidualnih odstupanja u modelu, primjenjuje se Whiteova korekcija te se nakon toga provodi *t*-test. U slučaju postojanja i autokorelacije i heteroskedastičnosti rezidualnih odstupanja u modelu, tada se primjenjuje Newey-West korekcija prije same provedbe *t*-testa.

Napomena: Ako se koriste spomenute korekcije, potrebno ih je uvažiti i kod provođenja *F*-testa, Waldova testa, na način da se u okviru naredbe linearHypothesis(…) uključi i sintaksa vcov=mat, pri čemu je mat matrica dobivena jednom od spomenutih korekcija.

```
#White korekcija:
library(car)
mat<-hccm(model,type="hc0")
mat #ovo je matrica var-kovar procjenitelja uz White korekciju

##             (Intercept)       kvadrat         sobe        godine   udaljenost
## (Intercept)   7.2228921  0.1026964675 -3.128843180 -0.1022419712 -0.151749376
## kvadrat       0.1026965  0.0142705101 -0.298465323 -0.0001960663 -0.007060468
## sobe         -3.1288432 -0.2984653227  6.518833756  0.0026099914  0.134398624
## godine       -0.1022420 -0.0001960663  0.002609991  0.0079664313 -0.001587154
## udaljenost   -0.1517494 -0.0070604678  0.134398624 -0.0015871545  0.036076413

#t-test :
library(lmtest)
coeftest(model,vcov=mat)

##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept) 2009.745687    2.687544 747.8001 < 2.2e-16 ***
## kvadrat        0.624027    0.119459   5.2238 4.352e-06 ***
## sobe           0.645730    2.553201   0.2529    0.8015
## godine        -1.433028    0.089255 -16.0555 < 2.2e-16 ***
## udaljenost    -2.614038    0.189938 -13.7626 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 4.18. Whiteova korekcija standardnih pogrešaka procjenitelja

```
#Newey-West korekcija:
library(sandwich)
mat2<-NeweyWest(model,lag=2)
mat2 #ovo je matrica var-kovar procjenitelja uz Newey-West korekciju

##             (Intercept)      kvadrat        sobe       godine   udaljenost
## (Intercept)   9.5825437  0.130658529 -3.79284882 -0.210233655 -0.128148233
## kvadrat       0.1306585  0.010619590 -0.22636983 -0.002575908 -0.007324410
## sobe         -3.7928488 -0.226369830  5.09602924  0.065306161  0.118564355
## godine       -0.2102337 -0.002575908  0.06530616  0.007715074  0.003935095
## udaljenost   -0.1281482 -0.007324410  0.11856435  0.003935095  0.028147677

#t-test :
library(lmtest)
coeftest(model,vcov=mat2)

##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept) 2009.745687    3.095568 649.2332 < 2.2e-16 ***
## kvadrat        0.624027    0.103051   6.0555 2.583e-07 ***
## sobe           0.645730    2.257439   0.2860    0.7762
## godine        -1.433028    0.087835 -16.3149 < 2.2e-16 ***
```

```
## udaljenost    -2.614038    0.167773 -15.5808 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 4.19. Newey-West korekcija standardnih pogrešaka procjenitelja

4.4. Nenormalnost distribucije grešaka relacije

Velik broj testova koji se provode nakon procjene samog regresijskog modela ovise o pretpostavci normalnosti distribucije grešaka relacije. Zato se u nastavku obrađuje ovaj problem.

4.4.1. Definiranje problema nenormalnosti distribucije grešaka relacije

Jedna od pretpostavki regresijskog modela je bila sljedeća: $\varepsilon_i \sim N(0,\sigma^2),\ \forall i$, iz čega je slijedilo (vidjeti naslov 2.2.3.) da je $\hat{\beta}_j \mid \boldsymbol{X} \sim N(\beta_j, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj})$. Sama pretpostavka normalnosti slučajne varijable nije potrebna za procjenu parametara (jer se procjena vrši izrazom $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$), **no važna je kod testiranja hipoteza (*t*-test, *F*-test, hi-kvadrat test) i izračunavanja intervalnih procjena**, jer se temelje na normalnoj distribuiranosti slučajne varijable. S obzirom da se pri prikupljanju podataka u regresijskoj analizi prikupe podaci o zavisnoj i nezavisnoj varijabli, dok nam podaci o slučajnoj varijabli ε_i nisu dostupni, razmatra se je li zavisna varijabla normalno distribuirana ili ne. Ako se ne radi o normalnoj distribuiranosti varijable y, procjenitelji metodom najmanjih kvadrata mogu biti asimptotski normalno distribuirani prema središnjem graničnom teoremu (vidjeti poglavlje 2.3. i dodatak 10.3.), ako se radi o dovoljno velikom uzorku.

4.4.2. Utvrđivanje postojanja problema nenormalnosti distribucije grešaka relacije

Osnovni pristup utvrđivanja ovoga problema je **grafički**, putem histograma ili dijagrama vjerojatnosti. **Histogram** je grafički prikaz distribucije frekvencija numeričkih podataka, pri čemu se radi o površinskom grafikonu, koji se sastoji od pravokutnika. Osnovice pravokutnika predočuju vrijednosti numeričke varijable, tj. razrede u aritmetičkom mjerilu (Šošić, 2006), dok visina pravokutnika predočava frekvenciju podataka (koja može biti izražena apsolutno ili relativno). Pritom se histogram uspoređuje s prikazom teorijske normalne distribucije. Ako se utvrdi da postoje značajna odstupanja histograma od teorijske normalne distribucije, to će ukazivati da empirijska distribucija odudara od normalne. S druge strane, ako je histogram zvonolikog i simetričnog oblika, u tom slučaju će ukazivati na normalnu distribuciju. Slika 4.20. predočava usporedbu histograma cijena metara kvadratnog stana iz prethodnih primjera i uspoređuje s normalnom distribucijom, dok slika 4.21. uspoređuje histogram rezidualnih odstupanja iz modela procijenjenog u primjeru 4.1.4.

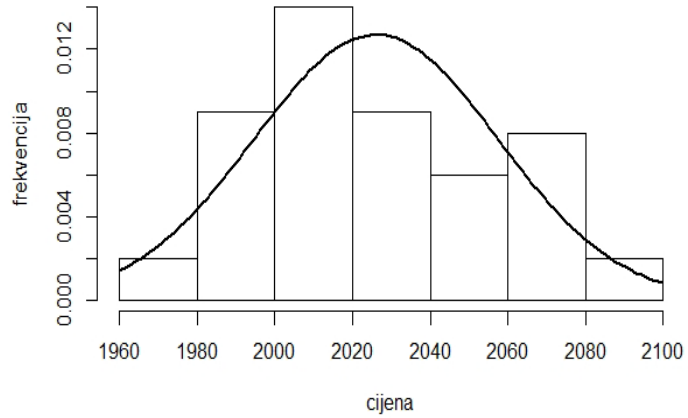
Nadalje, **dijagram vjerojatnosti** (engl. *probability plot*) je grafički prikaz na kojemu se uspoređuju očekivane vrijednosti rezidualnih odstupanja iz standardizirane normalne distribucije na osi apscisa s vrijednošću rezidualnih odstupanja na osi ordinata. Pritom se najprije rezidualna odstupanja poredaju po veličini od najmanjeg do najvećeg. Dakle, najprije se procijene rezidualna odstupanja $\hat{\varepsilon}_i$ iz procijenjenog regresijskog modela, potom se poredaju po veličini i zatim im se na dijagramu vjerojatnosti pridružuju očekivane vrijednosti standardne normalne distribucije. Ako rezidualna odstupanja slijede normalnu distribuciju, tada će točke na dijagramu vjerojatnosti biti raspoređene na način da će pripadati pravcu koji sadrži ishodište, a čiji je koeficijent smjera jednak standardnoj devijaciji regresije. Nakon što se vrijednosti rezidualnih odstupanja poredaju od najmanje do najveće vrijednosti, pri čemu se svakoj

vrijednosti pritom pridružuje očekivana vrijednost standardizirane normalne distribucije koja se računa izrazom:

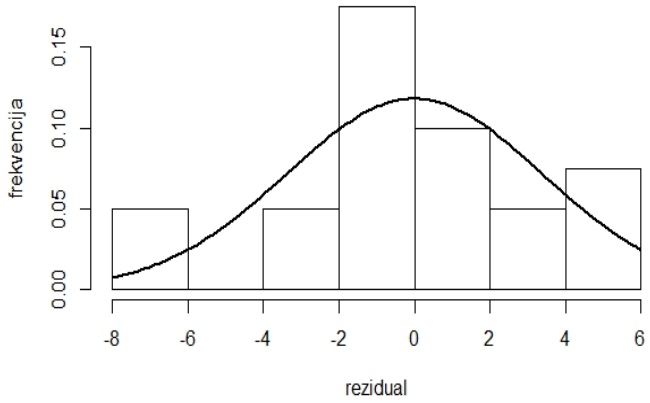
$$z_i = \phi^{-1}\left(\frac{3i-1}{3N+1}\right), \tag{4.62}$$

gdje ϕ^{-1} predstavlja inverz kumulativne distribucije standardizirane normalne varijable, i rang vrijednosti reziduala, a N veličinu uzorka kojeg analiziramo. Potom se na dijagramu istaknu točke čije koordinate predstavljaju vrijednosti (z_i, $\hat{\varepsilon}_i$). Slika 4.22. predočava dijagram vjerojatnosti za rezidualna odstupanja regresijskog modela iz primjera 4.1.4.

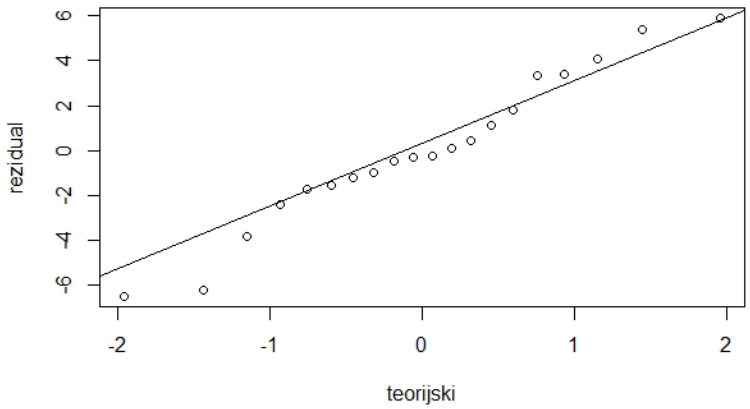
Ideja je da ako rezidualna odstupanja slijede normalnu distribuciju, točke će biti raspoređene oko spomenutog pravca, pri čemu se ne bi smjela uočavati prevelika odstupanja od tog pravca.



Slika 4.20. Usporedba histograma s normalnom distribucijom, cijena m^2 stana



Slika 4.21. Usporedba histograma s normalnom distribucijom, rezidualna odstupanja regresijskog modela



Slika 4.22. Dijagram vjerojatnosti rezidualnih odstupanja regresijskog modela

Formalno testiranje može se provesti putem nekoliko sljedećih testova.

Jarque-Bera test (1987) se temelji na usporedbi koeficijenta asimetrije i zaobljenosti neke empirijske distribucije s normalnom distribucijom. Kako su koeficijent asimetrije i koeficijent zaobljenosti za normalnu distribuciju jednaki $\alpha_3 = 0$ i $\alpha_4 = 3$, vrijednosti koeficijenata empirijske distribucije uspoređuju se s tim vrijednostima. Test veličina definirana je kao:

$$JB = N\left[\frac{\alpha_3^2}{6} + \frac{\left(\alpha_4 - 3\right)^2}{24}\right] \sim \chi^2(2)\,, \tag{4.63}$$

gdje su $\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{E\left(X-\mu\right)^3}{\sigma^3}$ i $\alpha_4 = \frac{\mu_4}{\sigma^4} = \frac{E\left(X-\mu\right)^4}{\sigma^4}$ treći i četvrti moment oko očekivane vrijednosti distribucije μ. Stoga α^2_3 u brojniku prvog razlomka u (4.63) predstavlja usporedbu kvadrata odstupanja koeficijenta asimetrije empirijske distribucije od vrijednosti za normalnu, tj. $(\alpha_3 - 0)^2 = \alpha^2_3$, dok $(\alpha_4 - 3)^2$ predstavlja kvadratno odstupanje koeficijenta zaobljenosti empirijske distribucije od vrijednosti koja vrijedi za normalnu, tj. 3. Jarque i Bera (1987) su izveli temeljem metode Lagrangeovih množitelja veličinu u (4.63), koja upravo u tom zapisu ako je zadovoljena pretpostavka normalne distribuiranosti varijable za koju se test provodi asimptotski slijedi hi-kvadrat distribuciju s 2 stupnja slobode. 2 su stupnja slobode jer se razmatraju dva momenta oko sredine. Stoga je test pogodan za velike uzorke.

Nulta i alternativna hipoteza ovoga testa su sljedeće:

$$H_0: \varepsilon \sim N(0,\sigma^2) \tag{4.64}$$
$$H_1: \varepsilon \not\sim N(0,\sigma^2)$$

ili

$$H_0: \text{greška relacije slijedi normalnu distribuciju} \tag{4.65}$$
$$H_1: \text{greška relacije ne slijedi normalnu distribuciju}$$

Empirijska test veličina uspoređuje se s teorijskom $\chi_\alpha^2(2)$. Ako je $LM > \chi_\alpha^2(2)$, tada se nulta hipoteza odbacuje, dok se za slučaj $LM < \chi_\alpha^2(2)$ ne odbacuje. Također, ishod se može zaključiti temeljem usporedbe p-vrijednosti s razinom značajnosti α.

Postoje još i Anderson-Darling test (1952), Kolmogorov–Smirnov test (1933, 1939), Shapiro-Wilkov test (1956), Lilliefors test (1967), Cramér-von Mises test (1928), itd. O prednostima i nedostacima pojedinih testova, zainteresirani čitatelji se upućuju na istraživanje Razali i Wah (2011), koje je pokazalo da najveću snagu testa ima Shapiro-Wilkov test; ili pak yap i Sim (2011). Nulte hipoteze ovih testova također pretpostavljaju normalnu distribuciju varijable koja se razmatra.

4.4.3. Ublažavanje problema nenormalnosti distribucije grešaka relacije

Mogući su sljedeći postupci ublažavanja ili uklanjanja problema nenormalnosti distribucije grešaka relacije:

- Transformacija zavisne varijable (korijen, logaritam) kako bi njena distribucija postala približno normalna.
- Transformacija nezavisnih varijabli ili odabir drugačijeg funkcijskog oblika regresijskog modela od početnog.

4.4.4. Primjer

Učitajte datoteku „**stanovi.txt**“ iz prethodnih primjera (naslovi 4.1.4., 4.2.4. i 4.3.4). Za procijenjeni model iz tog primjera provedite testiranje nenormalnosti grešaka relacije pomoću Jarque-Bera testa (skraćeno JB test) pri razini značajnosti 5%. Dodatno provjerite ishod Jarque-Bera testa pomoću Anderson-Darlingova, Cramér-von-Misesova i Lillieforsova testova.

Slika 4.23. predočava naredbe potrebne za izračun JB test veličine u (4.63), sama naredba jarque.test() daje ispis test veličine, no da bismo izračunali i koeficijente asimetrije i zaobljenosti, koriste se naredbe skewness() i kurtosis(). Stoga je test veličina sljedeća: $JB = N\left[\frac{\alpha_3^2}{6} + \frac{(\alpha_4-3)^2}{24}\right] = 50\left[\frac{0{,}0026^2}{6} + \frac{(2{,}184-3)^2}{24}\right] = 1{,}387$, s 2 stupnja slobode i s pripadajućom p-vrijednosti 0.4998, što je veće od 0.05 pa se nulta hipoteza H_0: $\varepsilon \sim N(0,\sigma^2)$ ne odbacuje. Uz razinu značajnosti od 5%, ne odbacujemo hipotezu da rezidualna odstupanja slijede normalnu distribuciju. Dodatno smo mogli usporediti test veličinu 1,387, s teorijskom razinom (naredba qchisq(.95,2), jer se razmatra značajnost od 5% i 2 su stupnja slobode). Slika 4.24. prikazuje naredbe i ispis preostalih testova, gdje se uočava da su sve p-vrijednosti odgovarajućih test veličina veće od zadane razine značajnosti od 5%, čime se potvrđuje prethodni zaključak.

```
#Jarque-Bera test:
reziduali<-resid(model)
library("moments")
jarque.test(reziduali)

##
##  Jarque-Bera Normality Test
##
## data:  reziduali
## JB = 1.3873, p-value = 0.4998
## alternative hypothesis: greater

#N, koef. asimetrije i zaobljenosti:
n<-length(reziduali); s<-skewness(reziduali); k<-kurtosis(reziduali)
n;s;k

## [1] 50

## [1] 0.002649004
```

```
## [1] 2.183994

JB<-n*(s^2/6+(k-3)^2/24)
JB

## [1] 1.387279
```

Slika 4.23. Jarque-Bera test normalnosti

```
#ostali testovi normalnost:

library(nortest)
ad.test(reziduali)

##
##  Anderson-Darling normality test
##
## data:  reziduali
## A = 0.25149, p-value = 0.7265

cvm.test(reziduali)

##
##  Cramer-von Mises normality test
##
## data:  reziduali
## W = 0.041388, p-value = 0.6485

lillie.test(reziduali)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  reziduali
## D = 0.078498, p-value = 0.6177
```

Slika 4.24. Anderson-Darlingov, Cramér-von-Misesov i Lilliefors testovi normalnosti

4.5. Alternativne metode procjene parametara

Spomenuto je kako u slučaju postojanja problema heteroskedastičnosti greške relacije postoje određene korekcije standardnih pogrešaka procjenitelja. Osim tih korekcija, moguće je primijeniti alternativne metode procjene parametara, koje se temelje na metodi najmanjih kvadrata. Ako postoji samo problem heteroskedastičnoti, tada se primjenjuje vagana metoda najmanjih kvadrata, dok u slučaju istovremenog postojanja i autokorelacije grešaka relacije primjenjujemo generaliziranu metodu najmanjih kvadrata. Njihov opis slijedi u idućim odjeljcima.

4.5.1. Generalizirana metoda najmanjih kvadrata

Generalizirana metoda najmanjih kvarata, engl. *Generalized least squares*, GLS metoda sastoji se od sljedećih koraka. Ako se radi o regresijskom modelu $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, $E[\boldsymbol{\varepsilon\varepsilon}'] = \widetilde{\Omega} \neq \Omega = \sigma^2 I$, dakle gdje je matrica varijanci-kovarijanci slučajne varijable $\widetilde{\Omega}$, pozitivno semidefinitna[46] matrica. Svaka pozitivno semidefinitna matrica može se zapisati na način (vidjeti dodatak 10.1.): $\widetilde{\Omega} = \boldsymbol{C\Lambda C}'$, gdje su stupci matrice $\boldsymbol{C}$ svojstveni vektori matrice $\widetilde{\Omega}$, a $\boldsymbol{\Lambda}$ je dijagonalna matrica sa svojstvenim vrijednostima matrice $\widetilde{\Omega}$ na glavnoj dijagonali. Neka je $\boldsymbol{\Lambda}^{1/2}$ dijagonalna

[46] Svaka matrica varijanci-kovarijanci je uvijek pozitivno semidefinitna, za dokaz vidjeti, npr. Sarapa (2002).

matrica, čiji su elementi na glavnoj dijagonali $\sqrt{\lambda_j}$, te neka je $\boldsymbol{T} = \boldsymbol{C}\boldsymbol{\Lambda}^{1/2}$. U tom slučaju je $\widetilde{\Omega}$ = $\boldsymbol{T}\boldsymbol{T}'$, te neka je $\boldsymbol{P}' = \boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}$ pa je $\widetilde{\Omega}^{-1} = \boldsymbol{P}'\boldsymbol{P}$.

Sada je model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, moguće zapisati na sljedeći način, tako da sve pomnožimo s $\boldsymbol{P}$ s lijeva:

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{\varepsilon}, \tag{4.66}$$

kojeg možemo pisati:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* , \tag{4.67}$$

gdje su $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{y}^*$, $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}^*$ i $\boldsymbol{P}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^*$. Sada je , $E[\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}] = \boldsymbol{P}\widetilde{\Omega}\boldsymbol{P}^{-1} = \sigma^2 I$, , gdje se pretpostavlja da je poznata matrica $\widetilde{\Omega}$. Sada je vektor procijenjenih parametara moguće dobiti metodom najmanjih kvadrata, na sljedeći način:

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{y} = (\boldsymbol{X}'\widetilde{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\widetilde{\Omega}\boldsymbol{y}, \tag{4.68}$$

te se radi o efikasnom procjenitelju dobivenim generaliziranom metodom najmanjih kvadrata (engl. *generalized squares estimator*).

Dakle, ideja je faktorizirati matricu $\widetilde{\Omega}$ na način da se transformacijom originalnog modela $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \widetilde{\Omega} \neq \Omega = \sigma^2 I$, dobije model $\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$ u kojemu će matrica varijanci-kovarijanci slučajne varijable zadovoljavati pretpostavke homoskedastičnosti i nezavisnosti slučajne varijable: $E[\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}] = \boldsymbol{P}\,\sigma^2\widetilde{\Omega}\boldsymbol{P}^{-1} = \sigma^2 I$.

Procjenitelj dobiven GLS metodom je **nepristran**, jer vrijedi:

$$E(\widehat{\boldsymbol{\beta}}_{GLS}) = E[(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{y}^*] = \boldsymbol{\beta} + E[(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{\varepsilon}^*] = \boldsymbol{\beta} , \tag{4.69}$$

te je **konzistentan** (izvod vidjeti u Greene, 2018), te je **asimptotski normalno distribuiran**, **s minimalnom varijancom (efikasan je):**

$$E(\widehat{\boldsymbol{\beta}}_{GLS}) = \sigma^2(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1} = \sigma^2(\boldsymbol{X}'\,\widetilde{\Omega}\boldsymbol{X})^{-1}. \tag{4.70}$$

Kako u stvarnosti nije poznata matrica $\widetilde{\Omega}$, ona se procjenjuje u prvome koraku, te se procjena uvrsti u (4.68), i to se potom naziva **ostvarljiva generalizirana metoda najmanjih kvadrata** (engl. *feasible generalized least squares*, FGLS ili *estimated generalized least squares*, EGLS). Potrebno je pretpostaviti kako se ponaša heteroskedastičnost slučajne varijable. Jedan pristup (Wooldridge, 2012) je sljedeći:

$$Var(\boldsymbol{\varepsilon}^*) = \sigma^2(\delta_0 + \delta_1 x_1 + \delta_2 x_3 + \ldots + \delta_k x_k) , \tag{4.71}$$

gdje se u suštini pretpostavlja samo problem heteroskedastičnosti slučajne varijable. Stoga se izraz (4.71) logaritmira kako bi se procijenila sljedeća jednadžba:

$$\ln(\sigma^2_i) = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + e_i. \tag{4.72}$$

Stoga se prvo procijeni originalni regresijski model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, prikupe se rezidualna odstupanja tog modela, izračunaju se njihovi kvadrati te se supstituiraju u (4.72):

$$\ln(\hat{\varepsilon}_i^2) = \delta_0 + \delta_1 x_{i1} + \delta_2 x_{i2} + \ldots + \delta_k x_{ik} + e_i. \tag{4.73}$$

Nakon što se procijeni model (4.73), prikupe se procijenjene vrijednosti $\widehat{\ln \hat{\varepsilon}_i^2}$, izračunaju se vrijednosti $\exp \widehat{\ln \hat{\varepsilon}_i^2} = \hat{h}_i$ te se uvrste u matricu $\widetilde{\Omega}$ kako bi se procijenio model $\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$.

Primjer 4.1.
Učitajte datoteku „**stanovi.txt**“ u RStudio. Datoteka sadrži sljedeće podatke o 50 stanova: cijena kvadratnog metra pojedinog stana (cijena), broj kvadrata stana (kvadrat), broj soba (sobe), starost stana (godine), udaljenost stana od centra grada (udaljenost). Procijenite linearni regresijski model u kojemu cijena stana ovisi o ostalim varijablama u modelu, pri čemu ćete primijeniti GLS metodu procjene, uz pretpostavku da se heteroskedastičnost grešaka relacije ponaša prema modelu (4.71). Usporedite rezultate procjena dobivenih pomoću GLS metode s metodom najmanjih kvadrata. Do kakvih je promjena došlo?

Najprije je originalni model u kojemu cijena kvadrata stana ovisi o nezavisnim varijablama sada spremljen pod nazivom OLS (vidjeti sliku 4.25.). Potom je u okviru paketa nlme pomoću naredbe GLS procijenjen isti model, pomoću generalizirane metode najmanjih kvadrata, gdje se uočava kako se dodajte sintaksa weights=varExp(), kojom se definira ponašanje heteroskedastičnosti greške relacije. Tablice predočene na slici 4.25. uspoređuju ispise modela koji se odnose na procijenjene parametre (Estimate), standardne pogreške procjenitelja (Std. Error), empirijske *t*-omjere i pripadajuće *p*-vrijednosti. Ono što valja uočiti kako, naravno, ne dolazi do promjene procijenjenih parametara, već njihovih standardnih pogrešaka, s obzirom da se GLS metoda primjenjuje da bi se korigirala matrica varijanci-kovarijanci procjenitelja, a sukladno tome dolazi do promjena empirijskih *t*-omjera i pripadajućih *p*-vrijednosti.

```
OLS<-lm(cijena~kvadrat+sobe+godine+udaljenost,data=stanovi)
library(nlme)
GLS<-gls(cijena~kvadrat+sobe+godine+udaljenost,
         data=stanovi,weights=varExp())
#usporedba procjena OLS i GLS metodom:
summary(OLS)$coefficients

##                 Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 2009.7456865  2.9255378 686.9662462 4.047604e-92
## kvadrat        0.6240274  0.1245594   5.0098793 8.901349e-06
## sobe           0.6457300  2.5767277   0.2506008 8.032635e-01
## godine        -1.4330282  0.1074557 -13.3359863 3.066576e-17
## udaljenost    -2.6140377  0.2010817 -12.9998786 7.671376e-17

summary(GLS)$tTable

##                   Value Std.Error    t-value      p-value
## (Intercept) 2009.6828429 2.9202770 688.182259 3.737988e-92
## kvadrat        0.6324484 0.1243709   5.085182 6.923700e-06
## sobe           0.4781664 2.5740122   0.185767 8.534624e-01
## godine        -1.4270009 0.1073192 -13.296785 3.410254e-17
## udaljenost    -2.6174081 0.2004430 -13.058115 6.537902e-17
```

Slika 4.25. Procijenjeni modeli metodom najmanjih kvadrata (OLS) i generaliziranom metodom najmanjih kvadrata (GLS)

Ono što se može uočiti da u konkretnom slučaju ne dolazi do promjena u ishodima *t*-testova značajnosti pojedinačnih varijabli u modelu. Ako postoji značajan problem heteroskedastičnosti, ishodi se mogu razlikovati, stoga u tom slučaju valja interpretirati rezultate dobivene GLS metodom.

4.5.2. Vagana metoda najmanjih kvadrata

Ako u regresijskom modelu ne postoji problem autokorelacije grešaka relacije, već samo problem heteroskedastičnosti, primjenjuje se vagana metoda najmanjih kvadrata (engl. *weighted least squares*, WLS). Pretpostavlja se sljedeći oblik matrice varijanci-kovarijanci grešaka relacije:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \begin{bmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & 0 \\ 0 & 0 & w_3 & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_N \end{bmatrix} = \sigma^2 \Omega \tag{4.74}$$

gdje w_i predstavlja ponder ili težinu i. Vrijedi:

$$\Omega^{-1} = \begin{bmatrix} w_1^{-1} & 0 & 0 & \cdots & 0 \\ 0 & w_2^{-1} & 0 & \cdots & 0 \\ 0 & 0 & w_3^{-1} & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_N^{-1} \end{bmatrix} \tag{4.75}$$

Kako vrijedi: $\Omega^{-1} = \boldsymbol{P}'\boldsymbol{P}$, gdje je $\boldsymbol{P}$ dijagonalna matrica čiji su dijagonalni elementi jednaki $\sqrt{w_i^{-1}}$, možemo model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ zapisati na način:

$$\boldsymbol{Py} = \boldsymbol{PX\beta} + \boldsymbol{P\varepsilon}, \tag{4.76}$$

tj.

$$\begin{bmatrix} y_1\sqrt{w_1^{-1}} \\ y_2\sqrt{w_2^{-1}} \\ \vdots \\ y_N\sqrt{w_N^{-1}} \end{bmatrix} = \begin{bmatrix} \sqrt{w_1^{-1}} & x_{11}\sqrt{w_1^{-1}} & \cdots & x_{1k}\sqrt{w_1^{-1}} \\ \sqrt{w_2^{-1}} & x_{21}\sqrt{w_2^{-1}} & \cdots & x_{2k}\sqrt{w_2^{-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{w_N^{-1}} & x_{N1}\sqrt{w_N^{-1}} & \cdots & x_{Nk}\sqrt{w_N^{-1}} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \varepsilon_1\sqrt{w_1^{-1}} \\ \varepsilon_2\sqrt{w_2^{-1}} \\ \vdots \\ \varepsilon_N\sqrt{w_N^{-1}} \end{bmatrix} \tag{4.77}$$

stoga se metoda najmanjih kvadrata može primijeniti na (4.77) kako bi se dobio procjenitelj $\widehat{\boldsymbol{\beta}}_{WLS}$ izrazom:

$$\widehat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wy}, \tag{4.78}$$

gdje $\boldsymbol{W}$ predstavlja matricu s težinama w_i. Naziv vagana metoda najmanjih kvadrata polazi od toga što svaka varijabla ima dodijeljenu težinu $\sqrt{w_i^{-1}}$. Postavlja se pitanje kako odrediti te težine.

Obično se u literaturi pretpostavlja da je varijanca greške relacije proporcionalna vrijednosti neke nezavisne varijable, stoga se može kao težina uzeti: $w_i = x_i^{-1}$, pa će matrica varijanci-kovarijanci $\boldsymbol{W}$ izgledati kao:

$$W = \begin{bmatrix} x_1^{-1} & 0 & 0 & \cdots & 0 \\ 0 & x_2^{-1} & 0 & \cdots & 0 \\ 0 & 0 & x_3^{-1} & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & x_N^{-1} \end{bmatrix} \tag{4.79}$$

Metoda pretpostavlja da je poznat oblik heteroskedastičnosti slučajne varijable, stoga se mora opravdati upravo korištenje određenog oblika ponašanja te heteroskedastičnosti.

Primjer 4.2.
Učitajte datoteku "**stanovi.txt**" u RStudio. Datoteka sadrži sljedeće podatke o 50 stanova: cijena kvadratnog metra pojedinog stana (cijena), broj kvadrata stana (kvadrat), broj soba (sobe), starost stana (godine), udaljenost stana od centra grada (udaljenost). Procijenite linearni regresijski model u kojemu cijena stana ovisi o ostalim varijablama u modelu, pri čemu ćete primijeniti WLS metodu procjene, uz pretpostavku da se heteroskedastičnost grešaka relacije mijenja u ovisnosti o broju soba obrnuto proporcionalno, tj. pretpostavite sljedeće težine: $w_i = x_i^{-1}$. Usporedite rezultate procjena dobivenih pomoću WLS metode s metodom najmanjih kvadrata. Do kakvih je promjena došlo?

Najprije je originalni model u kojemu cijena kvadrata stana ovisi o nezavisnim varijablama sada spremljen pod nazivom OLS (vidjeti sliku 4.26.), kao i u primjeru 4.1. Potom je dodanom sintaksom u okviru lm() naredbe (weights = 1/sobe) dodana opcija težina koje su obrnuto proporcionalne varijabli sobe.

```
OLS<-lm(cijena~kvadrat+sobe+godine+udaljenost,data=stanovi)
#WLS:
WLS<-lm(cijena~kvadrat+sobe+godine+udaljenost,data=stanovi,weights=1/sobe)
summary(OLS)$coefficients

##                 Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 2009.7456865  2.9255378 686.9662462 4.047604e-92
## kvadrat        0.6240274  0.1245594   5.0098793 8.901349e-06
## sobe           0.6457300  2.5767277   0.2506008 8.032635e-01
## godine        -1.4330282  0.1074557 -13.3359863 3.066576e-17
## udaljenost    -2.6140377  0.2010817 -12.9998786 7.671376e-17

summary(WLS)$coefficients

##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 2011.9939690  2.6863136 748.979545 8.284906e-94
## kvadrat        0.8166275  0.1220448   6.691213 2.928920e-08
## sobe          -3.7453205  2.5374195  -1.476035 1.468982e-01
## godine        -1.3830705  0.1044945 -13.235816 4.024384e-17
## udaljenost    -2.7022614  0.2003145 -13.490094 2.023451e-17
```

Slika 4.26. Procijenjeni modeli metodom najmanjih kvadrata (OLS) i ponderiranom metodom najmanjih kvadrata (WLS)

U ovome slučaju dolazi do promjena svih procijenjenih vrijednosti parametara, njihovih standardnih pogrešaka, i sukladno tome empirijskih *t*-omjera, kao i pripadajućih *p*-vrijednosti. Ishodi svih *t*-testova ostaju isti u slučaju obje metode procjene. Naravno, postoji značajan problem heteroskedastičnosti, ishodi se mogu razlikovati, stoga u tom slučaju valja interpretirati rezultate dobivene WLS metodom.

4.6. Sveobuhvatan primjer

Učitajte datoteku „**cobb-douglas.txt**“ u RStudio. Datoteka sadrži podatke o 40 poduzeća: ukupna količina proizvodnje (u kg), ukupno uložen rad (u satima), te ukupno uložen kapital (u satima rada strojeva). Procijenite log-log model u kojemu proizvodnja ovisi o radu i kapitalu.

a) Pomoću kriterija VIF, TOL, te pomoću prvog Kleinova pravila ispitajte postoji li problem multikolinearnosti varijabli u modelu. Zapišite koliko iznose koeficijenti determinacije za obje regresijske varijable i komentirajte rezultate.

```
cd<-read.table("cobb-douglas.txt",sep="\t",header=T)

#multikolinearnost
model<-lm(log(proizvodnja)~log(rad)+log(kapital),data=cd)
library(car)
vif(model)

##     log(rad) log(kapital)
##     2.884414     2.884414

1/vif(model)

##     log(rad) log(kapital)
##    0.3466909    0.3466909

m2<-lm(log(rad)~log(kapital),data=cd)
summary(m2)$r.squared

## [1] 0.6533091

cor(cd)

##             proizvodnja       rad   kapital
## proizvodnja   1.0000000 0.9244237 0.9706802
## rad           0.9244237 1.0000000 0.8057383
## kapital       0.9706802 0.8057383 1.0000000

sqrt(summary(model)$r.squared)

## [1] 0.9999196
```

Slika 4.27. Ispitivanje multikolinearnosti varijabli u modelu

Na slici 4.27. prikazan je postupak učitavanja podataka, procjene modela te provedbe testiranja pomoću *VIF* i *TOL* pokazatelja. Sada možemo pisati: $VIF_1 = 2{,}88 = VIF_2$, gdje uočavamo da su čak obje vrijednosti *VIF* manje od 5, stoga zaključujemo kako ne postoji problem multikolinearnosti varijabli u modelu.

Napomena. U slučaju dvije nezavisne varijable u modelu, obje *VIF* vrijednosti su jednake jer se u suštini procijeni regresijski model u kojemu najprije prva nezavisna varijabla ovisi o drugoj, a potom drugi u kojemu druga nezavisna ovisi o prvoj, te u tom slučaju je koeficijent determinacije u oba modela jednak.

Ekvivalentno smo mogli temeljem pokazatelja *TOL*: $TOL_1 = 0{,}35 = TOL_2$ doći do istog zaključka jer su obje vrijednosti veće od 0,2. To znači da bi u regresijskim modelima u kojima bi prve dvije nezavisne varijable promatrali kao zavisne koeficijenti determinacije bili manji od 80%. To je prikazano također na istoj slici (4.27.), gdje se uočava kako je $R_1^2 = 0{,}65 = R_2^2$.

Ako ispitujemo problem multikolinearnosti pomoću prvog Kleinovih pravila, na istoj slici 4.28. izračunata je korelacijska matrica za razmatrane varijable, kao i koeficijent korelacije regresije za početni model. Koeficijent višestruke linearne korelacije iznosi $R = 0{,}999$, dok su koeficijenti korelacija za svaki par varijabli dani u korelacijskoj matrici pomoću naredbe cor(). Sada se uočava da su svi koeficijenti korelacija po apsolutnoj vrijednosti manji od vrijednosti R, stoga se pomoću ovog prvog pravila također dolazi do zaključka da nema problema multikolinearnosti varijabli u modelu.

b) Za procijenjeni model iz tog primjera provedite testiranje autokorelacije prvog reda pomoću Durbin-Watsonova testa (dvosmjerni test te oba jednosmjerna), zapišite hipoteze svakog testa, test veličinu, približnu vrijednost koeficijenta autokorelacije prvog reda te donesite zaključak. Provedite testiranje autokorelacije do zaključno drugog reda pomoću Breusch-Godfreyeva testa: zapišite pomoćnu regresijsku jednadžbu, hipoteze testa, test veličinu i donesite zaključak. Provedite testiranje autokorelacije do zaključno drugog reda pomoću Ljung-Boxova testa: zapišite hipoteze testa, test veličinu i donesite zaključak. Razina značajnosti je 5%.

```
library(car)
durbinWatsonTest(model)

##  lag Autocorrelation D-W Statistic p-value
##    1        0.236226      1.481046   0.096
##  Alternative hypothesis: rho != 0

durbinWatsonTest(model,alternative = "positive")

##  lag Autocorrelation D-W Statistic p-value
##    1        0.236226      1.481046   0.064
##  Alternative hypothesis: rho > 0

durbinWatsonTest(model,alternative = "negative")

##  lag Autocorrelation D-W Statistic p-value
##    1        0.236226      1.481046   0.945
##  Alternative hypothesis: rho < 0
```

Slika 4.28. Durbin-Watsonov test autokorelacije

Sumarno su hipoteze Durbin-Watsonova testa prikazane u tablici kako slijedi:

Hipoteza/test	Dvosmjerni	Na gornju granicu	Na donju granicu
H_0	$\rho = 0$	$\rho = 0$	$\rho = 0$
H_1	$\rho \neq 0$	$\rho > 0$	$\rho < 0$
DW	1,48		
$\hat{\rho}$	0,236		

Kako DW vrijednost iznosi 1,48, procjena koeficijenta autokorelacije prvog reda iznosi 0,236. Ako se razmatra razina značajnosti od 5%, odbacuje se nulta hipoteza u slučaju dvosmjernog testa (p-v iznosi 0,096 > 0,05), kao i u slučaju testa na donju granicu (p-v iznosi 0,949 > 0,05). No, ako se provodi test na gornju granicu, u tom slučaju se odbacuje nulta hipoteza o nepostojanju autokorelacije rezidualnih odstupanja prvoga reda (p-v iznosi 0,047 < 0,05).

Nadalje, proveden je Breusch-Godfreyev test, vidjeti sliku 4.29. Nakon definiranja prvog i drugog pomaka rezidualnih odstupanja, procijenjena je pomoćna regresijska jednadžba, gdje se rezidualno odstupanje $\hat{\varepsilon}_i$ regresira na dvije nezavisne varijable iz originalnog modela, te svoja

prethodna dva pomaka. Pomoćna regresijska jednadžba je sljedeća: $\hat{\varepsilon}_i = -0{,}0004 + 0{,}003\ln x_{i1} - 0{,}003\ln x_{i2} + 0{,}346\hat{\varepsilon}_{i-1} - 0{,}111\hat{\varepsilon}_{i-2}$. Hipoteze testa su sljedeće: H_0: $\gamma_1 = \gamma_2 = 0$, H_1: $\exists\ \gamma_j \neq 0, j \in \{1, 2\}$. Test veličina računa se kao: $LM = (N-m)R^2_{pom} = 38 \cdot 0{,}1074 = 4{,}0821$, s p-v 0,1299, koja je veća od 0,05, stoga se ne odbacuje nulta hipoteza. Riječima: uz razinu značajnosti od 5%, ne odbacujemo hipotezu da ne postoji problem autokorelacije rezidualnih odstupanja do zaključno drugog reda.

Konačno, može se provesti i Ljung-Boxov test, temeljem naredbi prikazanih na slici 4.30. Hipoteze testa su sljedeće: H_0: $\rho_1 = \rho_2 = 0$, H_1: $\exists\ \rho_j \neq 0, j \in \{1, 2\}$. Test veličina iznosi $Q = 2{,}44$, koja slijedi hi-kvadrat distribuciju s 2 stupnja slobode, te pripadajućom p-vrijednošću 0,295. Kako je to veće od uobičajenih razina značajnosti, ne odbacujemo nultu hipotezu. Odnosno, uz razinu značajnosti od 5%, ne odbacujemo hipotezu da ne postoji problem autokorelacije rezidualnih odstupanja do zaključno drugog reda.

```
reziduali<-residuals(model)
rez1<-Lag(reziduali,1);rez2<-Lag(reziduali,2)
summary(pomocna<-lm(reziduali~log(rad)+log(kapital)+rez1+rez2,data=cd))

## Call:
## lm(formula = reziduali ~ log(rad) + log(kapital) + rez1 + rez2,
##     data = cd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0031862 -0.0014355  0.0000960  0.0009503  0.0057225
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0003672  0.0142200  -0.026   0.9796
## log(rad)      0.0030470  0.0033999   0.896   0.3766
## log(kapital) -0.0034138  0.0036401  -0.938   0.3551
## rez1          0.3456974  0.1790120   1.931   0.0621 .
## rez2         -0.1107569  0.1672579  -0.662   0.5124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002261 on 33 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.1074, Adjusted R-squared:  -0.0007681
## F-statistic: 0.9929 on 4 and 33 DF,  p-value: 0.4251

test_vel<-nobs(pomocna)*summary(pomocna)$r.squared
test_vel

## [1] 4.082075

p_v<-1-pchisq(test_vel,2)
p_v

## [1] 0.1298939
```

Slika 4.29. Breusch-Godfrey test autokorelacije

```
Box.test(reziduali,lag=2,type="Ljung-Box")
## Box-Ljung test
##
## data:  reziduali
## X-squared = 2.4395, df = 2, p-value = 0.2953
```

Slika 4.30. Ljung-Box test autokorelacije

c) Za procijenjeni model iz tog primjera provedite testiranje heteroskedastičnosti greške relacije, pri razini značajnosti od 5%. Pritom provedite i Breusch-Paganov test, kao i Whiteov test. Zapišite pomoćne regresijske jednadžbe temeljem kojih se testovi provode, hipoteze svakog testa, kao i test veličine. Interpretirajte rezultate. Neovisno o ishodu testova, provedite Whiteovu korekciju standardnih pogrešaka procjenitelja, kao i Newey-Westovu korekciju, te potom usporedite rezultate *t*-testova za slučaj originalnog modela, kao i obiju korekcija.

Temeljem naredbi i ispisa na slici 4.31. može se provesti Breusch-Paganov test na sljedeći način: Pomoćna regresijska jednadžba je $\hat{\varepsilon}_i^2 = 8{,}53 \cdot 10^{-5} - 1{,}9 \cdot 10^{-6} \ln x_{i1} + 9{,}09 \cdot 10^{-6} \ln x_{i2}$. Hipoteze testa su: H_0: $\alpha_1 = \alpha_2 = 0$, H_1: $\exists\ \alpha_j \neq 0, j \in \{1, 2\}$. pri čemu je broj opservacija jednak 40, uz R^2_{pom} jednak 0,128. Stoga je test veličina jednaka $LM = N \cdot R^2_{pom} = 5{,}12$. Pripadajuća *p*-vrijednost iznosi 0,077, izračunata temeljem 2 stupnja slobode. Kako vrijedi *p*-vrijednost > α, ne može se odbaciti nulta hipoteza da ne postoji problem heteroskedastičnosti rezidualnih odstupanja u modelu. Nadalje, slika 4.32. predočava naredbe za provedbu i ispis Whiteova testa.

```
summary(bptest<-lm(residuals(model)^2~log(rad)+log(kapital),data=cd))

## Call:
## lm(formula = residuals(model)^2 ~ log(rad) + log(kapital), data = cd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -8.994e-06 -4.282e-06 -1.471e-06  1.695e-06  2.271e-05
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.527e-05  4.176e-05   2.042   0.0483 *
## log(rad)     -1.903e-05  9.546e-06  -1.994   0.0536 .
## log(kapital)  9.093e-06  1.011e-05   0.899   0.3742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.844e-06 on 37 degrees of freedom
## Multiple R-squared:  0.1281, Adjusted R-squared:  0.08096
## F-statistic: 2.718 on 2 and 37 DF,  p-value: 0.0792

nobs(bptest);(summary(bptest))$r.squared

## [1] 40

## [1] 0.1280905

nobs(bptest)*(summary(bptest))$r.squared

## [1] 5.123619

p_vrijednost<-1-pchisq(nobs(bptest)*(summary(bptest))$r.squared,2)
p_vrijednost

## [1] 0.07716498
```

Slika 4.31. Breusch-Paganov test heteroskedastičnosti

```
summary(white<-lm(residuals(model)^2~log(rad)+log(kapital)
                  +I(log(rad)*log(kapital))
                  +I(log(rad)^2)+I(log(kapital)^2),data=cd))

##
## Call:
## lm(formula = residuals(model)^2 ~ log(rad) + log(kapital) + I(log(rad) *
##     log(kapital)) + I(log(rad)^2) + I(log(kapital)^2), data = cd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -8.799e-06 -3.395e-06 -5.574e-07  3.149e-06  1.396e-05
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.088e-04  1.292e-03  -0.162 0.872582
## log(rad)                    -9.164e-04  3.770e-04  -2.431 0.020494 *
## log(kapital)                 1.117e-03  3.069e-04   3.641 0.000894 ***
## I(log(rad) * log(kapital)) -5.738e-04  1.281e-04  -4.481 8.01e-05 ***
## I(log(rad)^2)                3.141e-04  6.083e-05   5.163 1.06e-05 ***
## I(log(kapital)^2)            2.397e-04  7.946e-05   3.017 0.004812 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.899e-06 on 34 degrees of freedom
## Multiple R-squared:  0.5895, Adjusted R-squared:  0.5291
## F-statistic: 9.765 on 5 and 34 DF,  p-value: 7.505e-06

nobs(white);(summary(white))$r.squared

## [1] 40

## [1] 0.5894874

nobs(white)*(summary(white))$r.squared

## [1] 23.5795

p_vrijednost<-1-pchisq(nobs(white)*(summary(white))$r.squared,5)
p_vrijednost

## [1] 0.0002614601
```

Slika 4.32. Whiteov test heteroskedastičnosti

Kako je u slučaju ovog testa uključen i kvadrat svake varijable, kao i međusobni umnošci, pomoćna regresijska jednažba je sljedeća:

$$\hat{\varepsilon}_i^2 = -2{,}09\cdot10^{-4} - 9{,}16\cdot10^{-4}\ln x_{i1} + 1{,}12\cdot10^{-3}\ln x_{i2} - 5{,}74\cdot10^{-4}\ln x_{i1}\ln x_{i2} + 3{,}14\cdot10^{-4}\left(\ln x_{i1}\right)^2 + 2{,}4\cdot10^{-4}\left(\ln x_{i2}\right)^2$$

Hipoteze testa su: H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_5 = 0$, H_1: $\exists\ \alpha_j \neq 0, j \in \{1, 2,.., 5\}$. , pri čemu je broj opservacija jednak 40, uz $R^2_{pom} = 0{,}589$. Stoga je test veličina jednaka $LM = N{\cdot}R^2_{pom} = 23{,}58$. Pripadajuća *p*-vrijednost iznosi 0,0003, izračunata temeljem 5 stupnja slobode. Kako vrijedi *p*-vrijednost $< \alpha$, odbacuje se nulta hipoteza da ne postoji problem heteroskedastičnosti rezidualnih odstupanja u modelu. Dakle, temeljem Whiteova testa se odbacuje nulta hipoteza o homoskedastičnosti varijance slučajne varijable.

Naredbe za provođenje obiju korekcija prikazane su na slici 4.33. Prva tablica s procjenama i provođenjem *t*-testova se odnosi na metodu najmanjih kvadrata, druga na Whiteovu korekciju, dok treća pri dnu slike 4.33. se odnosi na Newey-Westovu korekciju. Ishodi *t*-testova se nisu

promijenili u sva tri slučaja, s obzirom da su svi empirijski omjeri u svim slučajevima po apsolutnoj vrijednosti veći od odgovarajućih teorijskih, odnosno *p*-vrijednosti su manje od uobičajenih razina značajnosti. Međutim, kako je utvrđen problem heteroskedastičnosti rezidualnih odstupanja pomoću Whiteova testa, pouzdano je koristiti ispis koji se odnosi upravo na tu korekciju standardnih pogrešaka procjenitelja.

```
OLS<-lm(log(proizvodnja)~log(rad)+log(kapital),data=cd)
summary(OLS)$coefficients

##                 Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 2009.7456865  2.9255378 686.9662462 4.047604e-92
## kvadrat        0.6240274  0.1245594   5.0098793 8.901349e-06
## sobe           0.6457300  2.5767277   0.2506008 8.032635e-01
## godine        -1.4330282  0.1074557 -13.3359863 3.066576e-17
## udaljenost    -2.6140377  0.2010817 -12.9998786 7.671376e-17

#White korekcija:
library(car)
mat<-hccm(model,type="hc0")
#t-test :
library(lmtest)
coeftest(model,vcov=mat)

##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  -0.0475800  0.0134602  -3.5349  0.001116 **
## log(rad)      0.3716163  0.0045908  80.9485 < 2.2e-16 ***
## log(kapital)  0.6241268  0.0039069 159.7493 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Newey-West korekcija:
library(sandwich)
mat2<-NeweyWest(model,lag=2)
#t-test :
library(lmtest)
coeftest(model,vcov=mat2)

##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  -0.0475800  0.0144590  -3.2907  0.002202 **
## log(rad)      0.3716163  0.0042342  87.7656 < 2.2e-16 ***
## log(kapital)  0.6241268  0.0033445 186.6146 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Slika 4.33. Whiteova i Newey-Westova korekcija standardnih pogrešaka

d) Za procijenjeni model iz ovog primjera provedite testiranje nenormalnosti grešaka relacije pomoću Jarque-Bera testa pri razini značajnosti 5%. Dodatno provjerite ishod Jarque-Bera testa pomoću Anderson-Darlingova, Cramér-von-Misesova i Lillieforsova testova.

Ispisi svih testova prikazani su na slici 4.34. H_0: $\varepsilon \sim N(0,\sigma^2)$, H_1: $\varepsilon \not\sim N(0,\sigma^2)$. Test veličina je sljedeća: $JB = N\left[\frac{\alpha_3^2}{6} + \frac{(\alpha_4-3)^2}{24}\right] = 40\left[\frac{0{,}768^2}{6} + \frac{(2{,}985-3)^2}{24}\right] = 3{,}929$, s 2 stupnja slobode i s pripadajućom *p*-vrijednosti 0,140, što je veće od 0,05 pa se nulta hipoteza ne odbacuje. Uz

razinu značajnosti od 5%, ne odbacujemo hipotezu da rezidualna odstupanja slijede normalnu distribuciju.

Slika 4.34. prikazuje naredbe i ispis preostalih testova, gdje se uočava da su sve *p*-vrijednosti odgovarajućih test veličina manje od zadane razine značajnosti od 5%, što je u suprotnosti s prethodnim zaključkom. Dakle, potrebno je dodatno ispitati distribuciju rezidualnih odstupanja.

```
library("moments")
jarque.test(reziduali)

##
##  Jarque-Bera Normality Test
##
## data:  reziduali
## JB = 3.9295, p-value = 0.1402
## alternative hypothesis: greater

#N, koef. asimetrije i zaobljenosti:
n<-length(reziduali); s<-skewness(reziduali); k<-kurtosis(reziduali)
n;s;k

## [1] 40

## [1] 0.7676979

## [1] 2.984667

JB<-n*(s^2/6+(k-3)^2/24)
JB

## [1] 3.929459

#
library(nortest)
ad.test(reziduali)

##
##  Anderson-Darling normality test
##
## data:  reziduali
## A = 0.82998, p-value = 0.0294

cvm.test(reziduali)

##
##  Cramer-von Mises normality test
##
## data:  reziduali
## W = 0.13803, p-value = 0.03259

lillie.test(reziduali)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  reziduali
## D = 0.16657, p-value = 0.006836
```

Slika 4.34. Testiranje normalnosti distribucije rezidualnih odstupanja

e) Procijenite regresijski model pri čemu ćete primijeniti GLS metodu procjene, uz pretpostavku da se heteroskedastičnost grešaka relacije ponaša prema modelu (4.71). Usporedite rezultate procjena dobivenih pomoću GLS metode s metodom najmanjih kvadrata. Do kakvih je promjena došlo?

Sličan zaključak vrijedi kao i u postupku c); ishodi *t*-testova se ne razlikuju (slika 4.35.).

```
GLS<-gls(log(proizvodnja)~log(rad)+log(kapital),data=cd,weights=varExp())
summary(OLS)$coefficients

##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  -0.04758001 0.014192190  -3.352549 1.856875e-03
## log(rad)      0.37161629 0.003244253 114.546040 8.362369e-49
## log(kapital)  0.62412677 0.003435547 181.667369 3.346718e-56

summary(GLS)$tTable

##                    Value   Std.Error    t-value      p-value
## (Intercept)  -0.02643646 0.014688625  -1.799791 8.005225e-02
## log(rad)      0.37617018 0.002700396 139.301860 6.099844e-52
## log(kapital)  0.61568310 0.002680711 229.671581 5.756774e-60
```

Slika 4.35. Rezultati procjena OLS i GLS metodom

f) Procijenite regresijski model pri čemu ćete primijeniti WLS metodu procjene, uz pretpostavku da se heteroskedastičnost grešaka relacije mijenja u ovisnosti o varijabli rad obrnuto proporcionalno, tj. pretpostavite sljedeće težina: $w_i = x_i^{-1}$. Usporedite rezultate procjena dobivenih pomoću GLS metode s metodom najmanjih kvadrata. Do kakvih je promjena došlo?

```
WLS<-lm(log(proizvodnja)~log(rad)+log(kapital),data=cd,weights=1/log(rad))
summary(OLS)$coefficients

##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  -0.04758001 0.014192190  -3.352549 1.856875e-03
## log(rad)      0.37161629 0.003244253 114.546040 8.362369e-49
## log(kapital)  0.62412677 0.003435547 181.667369 3.346718e-56

summary(WLS)$coefficients

##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  -0.04762223 0.014208005  -3.351789 1.860777e-03
## log(rad)      0.37130233 0.003254334 114.094726 9.673678e-49
## log(kapital)  0.62449061 0.003449353 181.045712 3.798745e-56
```

Slika 4.36. Rezultati procjena OLS i WLS metodom

Sličan zaključak vrijedi kao i u postupku c) i e); ishodi *t*-testova se ne razlikuju (vidjeti sliku 4.36.).

4.7. Pitanja za ponavljanje

1) Nabrojite pretpostavke linearnog regresijskog modela.
2) Što je to multikolinearnost nezavisnih varijabli u regresijskom modelu?
3) Kakav je rang matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ u slučaju postojanja multikolinearnosti varijabli u regresijskom modelu?
4) Koja je razlika između savršene i približne multikolinearnosti?
5) Koje su posljedice postojanja problema multikolinearnosti?
6) Koje su prve indikacije postojanja problema multikolinearnosti?
7) Kako možemo utvrditi problem multikolinearnosti pomoću faktora inflacije varijance? Zašto ga je bitno uspoređivati s vrijednošću 5?
8) Objasnite oba Kleinova pravila za utvrđivanje multikolinearnosti.
9) Na koji način možemo ublažiti ili ukloniti problem multikolinearnosti varijabli u modelu?
10) Što je to autokorelacija grešaka relacije u regresijskom modelu? Zapišite simbolički te matrični zapis ovoga problema.
11) Koje su posljedice postojanja problema autokorelacije grešaka relacije?
12) Na koje svojstvo procjenitelja ne utječe postojanje problema autokorelacije grešaka relacije?
13) Koji su uzroci postojanja problema autokorelacije grešaka relacije?
14) Kako se može grafičkim putem utvrditi postojanje problema autokorelacije grešaka relacije?
15) Koji test se koristi za utvrđivanje problema autokorelacije grešaka relacije prvog reda? Zapišite hipoteze testa.
16) Koji je odnos između Durbin-Watson veličine i koeficijenta autokorelacije grešaka relacije prvog reda?
17) Koji su nedostaci Durbin-Watson testa?
18) Koji test se koristi u slučaju autokorelacije višeg reda? Zapišite pomoćnu regresijsku jednadžbu za slučaj dvije nezavisne varijable te testiranje autokorelacije grešaka relacije do zaključno 3. reda. Potom zapišite hipoteze testa.
19) Koji test za postojanje autokorelacije grešaka relacije se koristi u slučaju vremenskih nizova? Zapišite nultu i alternativnu hipotezu tog testa za slučaj testiranja autokorelacije do zaključno 5. reda.
20) Na koji način možemo ublažiti ili ukloniti problem autokorelacije grešaka relacije u modelu?
21) Što je to heteroskedastičnost grešaka relacije u regresijskom modelu? Zapišite simbolički te matrični zapis ovoga problema.
22) Koje su posljedice postojanja problema heteroskedastičnost grešaka relacije?
23) Na koje svojstvo procjenitelja ne utječe postojanje problema heteroskedastičnosti grešaka relacije?
24) Koji su uzroci postojanja problema heteroskedastičnosti grešaka relacije?
25) Kako se može grafičkim putem utvrditi postojanje problema heteroskedastičnosti grešaka relacije?
26) Koji su formalni testovi za utvrđivanje postojanja problema heteroskedastičnosti grešaka relacije?
27) Zapišite pomoćnu jednadžbu Breusch-Paganova testa za utvrđivanje postojanja problema heteroskedastičnosti grešaka relacije za slučaj 5. nezavisnih varijabli. Potom zapišite hipoteze testa.
28) Koji je nedostatak Breusch-Paganova testa? Koji test se može koristiti kao alternativan?

29) Zapišite pomoćnu jednadžbu Whiteova testa za utvrđivanje postojanja problema heteroskedastičnosti grešaka relacije za slučaj 2. nezavisne varijable. Potom zapišite hipoteze testa.
30) Opišite Goldfeld-Quandtov test za utvrđivanje postojanja problema heteroskedastičnosti grešaka relacije.
31) Na koji način možemo ublažiti ili ukloniti problem heteroskedastičnosti grešaka relacije u modelu?
32) Kada se koristi Whiteova korekcija standardnih pogrešaka procjenitelja, a kada Newey-West korekcija?
33) Što je to nenormalnost distribucije grešaka relacije u regresijskom modelu? Zapišite simbolički zapis ovoga problema.
34) Koje su posljedice postojanja problema nenormalnost distribucije grešaka relacije?
35) Na koji način možemo ublažiti ili ukloniti problem nenormalnosti distribucije grešaka relacije u modelu?
36) Koja je razlika između generalizirane metode najmanjih kvadrata i vagane metode najmanjih kvadrata? Kada ih koristimo?
37) Učitajte datoteku „**placa.txt**“ u RStudio. Datoteka sadrži podatke o 150 pojedinaca: iznos plaće (u kn), broj godina radnog staža (staz), te broj godina školovanja (obrazovanje). Procijenite linearni model u kojemu plaća pojedinaca ovisi o broju godina radnog staža te broju godina školovanja i spremite ga kao „model“.

a) Pomoću kriterija *VIF*, *TOL*, te pomoću oba Kleinova pravila ispitajte postoji li problem multikolinearnosti varijabli u modelu. Zapišite koliko iznose koeficijenti determinacije za one regresijske varijable kod kojih se utvrdi problem multikolinearnosti.
b) Za procijenjeni model provedite testiranje autokorelacije prvog reda pomoću Durbin-Watson testa (dvosmjerni test te oba jednosmjerna), zapišite hipoteze svakog testa, test veličinu, približnu vrijednost koeficijenta autokorelacije prvog reda te donesite zaključak. Provedite testiranje autokorelacije do zaključno trećeg reda pomoću Breusch-Godfreyeva testa: zapišite pomoćnu regresijsku jednadžbu, hipoteze testa, test veličinu i donesite zaključak. Provedite testiranje autokorelacije do zaključno trećeg reda pomoću Ljung-Boxova testa: zapišite hipoteze testa, test veličinu i donesite zaključak. Razina značajnosti je 5%.
c) Za procijenjeni model provedite testiranje heteroskedastičnosti greške relacije, pri razini značajnosti od 5%. Pritom provedite i Breusch-Paganov test, kao i Whiteov test. Zapišite pomoćne regresijske jednadžbe temeljem kojih se testovi provode, hipoteze svakog testa, kao i test veličine. Interpretirajte rezultate. Neovisno o ishodu testova, provedite Whiteovu korekciju standardnih pogrešaka procjenitelja, kao i Newey-Westovu korekciju (uz pretpostavku postojanja problema autokorelacije grešaka relacije do zaključno trećeg reda), te potom usporedite rezultate *t*-testova za slučaj originalnog modela, kao i obiju korekcija.
d) Za procijenjeni model iz tog primjera provedite testiranje nenormalnosti grešaka relacije pomoću Jarque-Bera testa pri razini značajnosti 5%.
e) Procijenite isti model koji ste na početku zadataka, ali pomoću GLS metode, kao i WLS metode (kod WLS metode pretpostavite da se heteroskedastičnost grešaka relacije mijenja u ovisnosti o broju godina školovanja obrnuto proporcionalno, tj. pretpostavite sljedeće težina: $w_i = x_i^{-1}$). Komentirajte do kakvih promjena dolazi u zaključcima pojedinačnih testova značajnosti.

Rješenja

Zadatak 37):

```
#multikolinearnost
model<-lm(placa~staz+obrazovanje,data=place)
library(car)
vif(model)

##        staz obrazovanje
##    36.18098    36.18098

1/vif(model)

##        staz obrazovanje
##  0.02763883  0.02763883

m1<-lm(staz~obrazovanje,data=place)
summary(m1)$r.squared

## [1] 0.9723612

cor(place)

##               pojedinac       placa        staz obrazovanje
## pojedinac    1.00000000 -0.07015636 -0.03102125 -0.01702086
## placa       -0.07015636  1.00000000  0.94841667  0.93549745
## staz        -0.03102125  0.94841667  1.00000000  0.98608375
## obrazovanje -0.01702086  0.93549745  0.98608375  1.00000000

sqrt(summary(model)$r.squared)

## [1] 0.9484182
```

```
#durbin watson
library(car)
durbinWatsonTest(model)

##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01298824      2.023118   0.834
##  Alternative hypothesis: rho != 0

durbinWatsonTest(model,alternative = "positive")

##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01298824      2.023118   0.555
##  Alternative hypothesis: rho > 0

durbinWatsonTest(model,alternative = "negative")

##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01298824      2.023118   0.449
##  Alternative hypothesis: rho < 0
```

```
#breusch-godfrey
library(quantmod)
reziduali<-residuals(model)
rez1<-Lag(reziduali,1);rez2<-Lag(reziduali,2);rez3<-Lag(reziduali,3)
summary(pomocna<-lm(reziduali~staz+obrazovanje+rez1+rez2+rez3,data=place))

##
## Call:
## lm(formula = reziduali ~ staz + obrazovanje + rez1 + rez2 + rez3,
##     data = place)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8499  -3.8661   0.0219   4.1305  10.6759
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.008743   1.014943   0.009    0.993
## staz         0.039127   0.233730   0.167    0.867
## obrazovanje -0.045741   0.239974  -0.191    0.849
## rez1         0.014716   0.083610   0.176    0.861
## rez2        -0.024190   0.084354  -0.287    0.775
## rez3        -0.016795   0.082939  -0.203    0.840
##
## Residual standard error: 5.687 on 141 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.001624,   Adjusted R-squared:  -0.03378
## F-statistic: 0.04588 on 5 and 141 DF,  p-value: 0.9987

test_vel<-nobs(pomocna)*summary(pomocna)$r.squared
test_vel

## [1] 0.2387691

p_v<-1-pchisq(test_vel,3)
p_v

## [1] 0.9711005
```

```
#ljung-box
Box.test(reziduali,lag=3,type="Ljung-Box")

##
##  Box-Ljung test
##
## data:  reziduali
## X-squared = 0.11924, df = 3, p-value = 0.9894
```

```
#BG test:
summary(bptest<-lm(residuals(model)^2~staz+obrazovanje,data=place))

##
## Call:
## lm(formula = residuals(model)^2 ~ staz + obrazovanje, data = place)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36.54 -24.65 -13.46  19.14 110.78
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.504      5.760   6.511 1.11e-09 ***
## staz          -2.440      1.348  -1.809   0.0724 .
## obrazovanje    2.440      1.385   1.761   0.0803 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.02 on 147 degrees of freedom
## Multiple R-squared:  0.02191,    Adjusted R-squared:  0.008607
## F-statistic: 1.647 on 2 and 147 DF,  p-value: 0.1962

nobs(bptest);(summary(bptest))$r.squared

## [1] 150
```

```
## [1] 0.02191474

nobs(bptest)*(summary(bptest))$r.squared

## [1] 3.287211

p_vrijednost<-1-pchisq(nobs(bptest)*(summary(bptest))$r.squared,2)
p_vrijednost

## [1] 0.1932819
```

```
#White test:
summary(white<-lm(residuals(model)^2~staz+obrazovanje
                  +I(staz*obrazovanje)
                  +I(staz^2)+I(obrazovanje^2),data=place))

##
## Call:
## lm(formula = residuals(model)^2 ~ staz + obrazovanje + I(staz *
##     obrazovanje) + I(staz^2) + I(obrazovanje^2), data = place)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.88 -24.77 -12.94  20.14 100.81
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            26.7511     8.4884   3.151  0.00198 **
## staz                   -6.1412     2.7691  -2.218  0.02814 *
## obrazovanje             7.1512     3.0573   2.339  0.02071 *
## I(staz * obrazovanje)  -3.3170     1.2807  -2.590  0.01059 *
## I(staz^2)               1.5660     0.6136   2.552  0.01175 *
## I(obrazovanje^2)        1.7388     0.6697   2.596  0.01040 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.59 on 144 degrees of freedom
## Multiple R-squared:  0.06631,    Adjusted R-squared:  0.03389
## F-statistic: 2.045 on 5 and 144 DF,  p-value: 0.07575

nobs(white);(summary(white))$r.squared

## [1] 150

## [1] 0.06630995

nobs(white)*(summary(white))$r.squared

## [1] 9.946492

p_vrijednost<-1-pchisq(nobs(white)*(summary(white))$r.squared,5)
p_vrijednost

## [1] 0.07676564
```

```
#White korekcija:
library(car)
mat<-hccm(model,type="hc0")
mat #ovo je matrica var-kovar procjenitelja uz White korekciju

##             (Intercept)        staz obrazovanje
## (Intercept)   1.1171687 -0.15555411  0.12805398
## staz         -0.1555541  0.06693509 -0.06770103
## obrazovanje   0.1280540 -0.06770103  0.07004318
```

```
#t-test:
library(lmtest)
coeftest(model,vcov=mat)

##
## t test of coefficients:
##
##               Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 4.0087e+03 1.0570e+00 3792.6843 < 2.2e-16 ***
## staz        1.3904e+00 2.5872e-01    5.3741 2.954e-07 ***
## obrazovanje 1.5375e-02 2.6466e-01    0.0581    0.9538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Newey-West korekcija:
library(sandwich)
mat2<-NeweyWest(model,lag=3)
mat2 #ovo je matrica var-kovar procjenitelja uz Newey-West korekciju

##             (Intercept)        staz obrazovanje
## (Intercept)   0.7735279 -0.13957166  0.12058891
## staz         -0.1395717  0.05786712 -0.05704447
## obrazovanje   0.1205889 -0.05704447  0.05737971

#t-test :
library(lmtest)
coeftest(model,vcov=mat2)

##
## t test of coefficients:
##
##               Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 4.0087e+03 8.7950e-01 4557.9345 < 2.2e-16 ***
## staz        1.3904e+00 2.4056e-01    5.7798 4.315e-08 ***
## obrazovanje 1.5375e-02 2.3954e-01    0.0642    0.9489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Jarque-Bera Normality Test
##
## data:  reziduali
## JB = 5.5857, p-value = 0.06125
## alternative hypothesis: greater
```

```
#N, koef. asimetrije i zaobljenosti:
n<-length(reziduali); s<-skewness(reziduali); k<-kurtosis(reziduali)
n;s;k

## [1] 150

## [1] -0.09295569

## [1] 2.073095

JB<-n*(s^2/6+(k-3)^2/24)
JB

## [1] 5.585722
```

```
library(nortest)
ad.test(reziduali)
```

```
##
##  Anderson-Darling normality test
##
## data:  reziduali
## A = 0.82084, p-value = 0.03325

cvm.test(reziduali)

##
##  Cramer-von Mises normality test
##
## data:  reziduali
## W = 0.12102, p-value = 0.0574

lillie.test(reziduali)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  reziduali
## D = 0.07755, p-value = 0.02776
```

```
OLS<-lm(placa~staz+obrazovanje,data=place)
library(nlme)
GLS<-gls(placa~staz+obrazovanje,data=place,weights=varExp())
summary(OLS)$coefficients

##                 Estimate Std. Error      t value     Pr(>|t|)
## (Intercept) 4.008723e+03  0.9954432 4.027074e+03 0.000000e+00
## staz        1.390374e+00  0.2330176 5.966821e+00 1.730807e-08
## obrazovanje 1.537528e-02  0.2394033 6.422334e-02 9.488796e-01

summary(GLS)$tTable

##                    Value Std.Error      t-value      p-value
## (Intercept) 4008.7624912 1.0155900 3947.2251424 0.000000e+00
## staz           1.3741493 0.2346451    5.8562886 2.976270e-08
## obrazovanje    0.0315033 0.2407076    0.1308779 8.960508e-01

WLS<-lm(placa~staz+obrazovanje,data=place,weights=1/obrazovanje)
summary(OLS)$coefficients

##                 Estimate Std. Error      t value     Pr(>|t|)
## (Intercept) 4.008723e+03  0.9954432 4.027074e+03 0.000000e+00
## staz        1.390374e+00  0.2330176 5.966821e+00 1.730807e-08
## obrazovanje 1.537528e-02  0.2394033 6.422334e-02 9.488796e-01

summary(WLS)$coefficients

##                 Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 4009.3466372  0.6337507 6326.378163 0.000000e+00
## staz           1.6478526  0.1929244    8.541441 1.540287e-14
## obrazovanje   -0.3075025  0.2050179   -1.499881 1.357899e-01
```

5.

DODACI

5. DODACI

5.1. Matrična algebra

5.1.1. Osnovni pojmovi

Matrica $\boldsymbol{A}$ formata (m,n) je pravokutna shema elemenata, pri čemu se m odnosi na ukupan broj redaka, a n na ukupan broj stupaca:

$$\boldsymbol{A} = \left[a_{ij} \right] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

a_{ij} je element u retku i i stupcu j u matrici $\boldsymbol{A}$, $i \in \{1, 2, \ldots, m\}, j \in \{1, 2, \ldots, n\}$.

Vrste matrica koje se najčešće primjenjuju u ekonometriji:

- Ako je $m \neq n$ – pravokutna matrica.
- Ako je $m = n$ – kvadratna matrica.
- Simetrična matrica: $a_{ij} = a_{ji}$ za svaki i i za svaki j.
- Vektor-redak – matrica s jednim retkom.
- Vektor-stupac – matrica s jednim stupcem.
- Dijagonalna matrica – kvadratna matrica čiji su svi nedijagonalni elementi jednaki 0, tj. $a_{ij} \neq 0$ za sve uređene parove (i,j) čim je $i \neq j$.
- Skalarna matrica – dijagonalna matrica, čiji su elementi na glavnoj dijagonali jednaki nekome skalaru s, tj. $a_{ij} \neq 0$ za sve uređene parove (i,j) čim je $i \neq j$ i $a_{ij} = s$ za sve i.
- Jedinična matrica – skalarna matrica, skalar je jednak 1. Oznaka: I.
- Donja trokutasta matrica – elementi a_{ij} za koje je $i > j$ su jednaki 0.
- Gornja trokutasta matrica – elementi a_{ij} za koje je $i < j$ su jednaki 0.
- Nul-matrica – matrica čiji su svi elementi jednaki 0.

5.1.2. Algebarske manipulacije s matricama

Matrice $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{M}_{m,n}$ su jednake ako i samo ako su istog formata i svi korespondentni elementi su im jednaki.

Transponirana matrica matrice $\boldsymbol{A}$, u oznaci $\boldsymbol{A}'$, je matrica dobivena tako da svaki redak i u matrici $\boldsymbol{A}$ postane stupac i, dok svaki stupac j u matrici $\boldsymbol{A}$ postane redak j.

Ako i samo ako je $\boldsymbol{A}$ simetrična matrica, vrijedi $\boldsymbol{A} = \boldsymbol{A}'$. Također, vrijedi i $(\boldsymbol{A}')' = \boldsymbol{A}$.

Matricu $\boldsymbol{A} \in \mathcal{M}_{m,n}$ množimo sa skalarom $a \in \mathbb{R}$ tako da svaki element u matrici $\boldsymbol{A}$ pomnožimo sa skalarom a.

Matrice $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{M}_{m,n}$ zbrajamo (oduzimamo) na način da im zbrojimo (oduzmemo) sve korespondentne elemente, tj. $\left[a_{ij} + b_{ij} \right] = \left[a_{ij} \right] + \left[b_{ij} \right]$, za sve uređene parove (i, j).

Nul-matrica neutralna je pri zbrajanju: $\boldsymbol{A} + \boldsymbol{0} = \boldsymbol{0} + \boldsymbol{A} = \boldsymbol{A}$.

Zbrajanje matrica je komutativno: $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$, i asocijativno: $(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C})$, a također vrijedi i $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$.
Skalarni umnožak vektora-retka $\boldsymbol{x}'$ i vektora-stupca $\boldsymbol{y}$ jest $\boldsymbol{x}'\boldsymbol{y} = \sum_{i=1}^{N} x_i y_i = c \in \mathbb{R}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$.
Dakle, c je neki realni broj, skalar.

Umnožak dviju matrica $\boldsymbol{A}$ i $\boldsymbol{B}$, $\boldsymbol{A} \in \mathcal{M}_{m,n}$, $\boldsymbol{B} \in \mathcal{M}_{n,p}$, koje su ulančane (matrica $\boldsymbol{B}$ ima redaka koliko matrica $\boldsymbol{A}$ ima stupaca) definira se kao matrica čiji su elementi definirani kao: $c_{ij} = \sum_{p=1}^{n} a_{ip} b_{pj}$ za svaki $i \in \{1, 2, \ldots, m\}$. Općenito, množenje matrica nije komutativno, tj. općenito je $\boldsymbol{AB} \neq \boldsymbol{BA}$.

Svojstva množenja matrica:

- Asocijativnost: $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$
- Distributivnost: $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$ (s lijeva) i $(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C}=\boldsymbol{AC}+\boldsymbol{BC}$, (s desna)
- Transponat umnoška: $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$
- Transponat umnožaka: $(\boldsymbol{ABC})' = \boldsymbol{C}'\,\boldsymbol{B}'\,\boldsymbol{A}'$

Neka je $\boldsymbol{e}$ vektor stupac jedinica, $\boldsymbol{e}'$ = [1 1 …. 1]. Tada možemo pisati: $\sum_{i=1}^{N} x_i = \boldsymbol{e}'\boldsymbol{x}$, gdje je $\boldsymbol{x}$ vektor stupac s komponentama $x_1, x_2, \ldots, x_N$. Za neku konstantu c i vektor $\boldsymbol{x}$ vrijedi: $\sum_{i=1}^{N} c x_i = \boldsymbol{ce}'\boldsymbol{x}$. Nadalje, vrijedi i $\sum_{i=1}^{N} x_i^2 = \boldsymbol{x}'\boldsymbol{x}$.

Idempotentna matrica je ona za koju vrijedi: $\boldsymbol{A}^2 = \boldsymbol{A}\cdot\boldsymbol{A} = \boldsymbol{A}$.

Inverzna matrica, $\boldsymbol{A}^{-1}$, matrice $\boldsymbol{A} \in \mathcal{M}_n$, je ona za koju vrijedi: $\boldsymbol{A}\,\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\,\boldsymbol{A} = I$. Ako postoji, inverzna matrica je jedinstvena. Matricu koja ima inverz nazivamo regularnom, u suprotnom ju nazivamo singularnom.

Svojstva inverza:

- $\boldsymbol{A} = (\boldsymbol{A}^{-1})^{-1}$
- $(\boldsymbol{A}^{-1})' = (\boldsymbol{A}')^{-1}$
- $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$
- $(\boldsymbol{ABC})^{-1} = \boldsymbol{C}^{-1}(\boldsymbol{AB})^{-1} = \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$

Trag matrice, $\mathrm{tr}(\boldsymbol{A})$, je zbroj elemenata na glavnoj dijagonali kvadratne matrice: $\sum_{i=1}^{n} a_{ii}$ za $\boldsymbol{A} \in \mathcal{M}_n$. Svojstva traga matrice su:

- $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}')$
- $\mathrm{tr}(\boldsymbol{A} + \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$
- $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$
- $\mathrm{tr}(c\boldsymbol{A}) = c\,\mathrm{tr}(\boldsymbol{A})$, $c \in \mathbb{R}$

- ciklička permutacija umnožaka: tr(***ABCD***) = tr(***BCDA***) = tr(***CDAB***) = tr(***DABC***)
- tr(*I*) = *k*, gdje je *k* red matrice *I*.

5.1.3. Linearna ovisnost vektora, rang matrice, determinanta kvadratne matrice

Vektor $x \in \mathbb{R}^N$ je uređena *N*-torka brojeva.

Linearna kombinacija vektora $\boldsymbol{y} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + ... + a_k\boldsymbol{x}_k$ je linearna kombinacija vektora $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_k$, dok su a_1, a_2, … , a_k koeficijenti linearne kombinacije.

Skup vektora $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ je linearno neovisan (nezavisan) ako je $a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + ... + a_k\boldsymbol{x}_k = \boldsymbol{0}$ (s desne strane jednakosti je nul-vektor) samo ako je $a_1 = a_2 = ... = a_k = 0$. S druge strane, ako postoji neki $a_i \neq 0$, $i \in \{1, 2, \ldots, k\}$, kažemo da je skup vektora $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ linearno ovisan (zavisan). Također, može se reći da je skup vektora linearno ovisan ako se jedan vektor iz tog skupa može prikazati kao linearna kombinacija preostalih vektora (ovo je teorem).

Dva vektora su okomita (ortogonalna) ako vrijedi $\boldsymbol{x}'\boldsymbol{y} = \sum_{i=1}^{N} x_i y_i = 0, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$.

Rang matrice, $r(\boldsymbol{A})$, je maksimalan broj linearno neovisnih redaka (stupaca) matrice. Za $\boldsymbol{A} \in \mathcal{M}_{m,n}$ vrijedi: $r(\boldsymbol{A}) = \min\{m, n\}$, dok za $\boldsymbol{A} \in \mathcal{M}_n$ vrijedi: $r(\boldsymbol{A}) \leq n$. Očito je $r(\boldsymbol{A}) \in \mathbb{N} \cup \{0\}$

Kvadratna matrica $\boldsymbol{A} \in \mathcal{M}_n$ za koju je $r(\boldsymbol{A}) = n$ (rang je maksimalan), je regularna.

Za rang matrice vrijedi: $r(\boldsymbol{A}) = r(\boldsymbol{A}'\boldsymbol{A}) = r(\boldsymbol{A}\boldsymbol{A}')$. Za kvadratne matrice vrijedi: $r(\boldsymbol{A}) = n$ ako i samo ako je $\boldsymbol{A}$ invertibilna matrica (tj. regularna, ima inverz).

Determinanta kvadratne matrice prvog reda ($\boldsymbol{A} = [a_{ij}]$) jednaka je $\det(\boldsymbol{A}) = a_{11}$.

Determinanta kvadratne matrice drugog reda $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ jednaka je $\det(\boldsymbol{A}) = a_{11}a_{22} - a_{12}a_{21}$.

Determinanta kvadratne matrice reda $n \geq 3$ računa se prema Laplaceovom razvoju. Za neki redak i, $\det(\boldsymbol{A}) = \sum_{j=1}^{n} a_{ij}A_{ij} = \sum_{j=1}^{n} a_{ij}(-1)^{i+j}M_{ij}$, gdje je A_{ij} kofaktor ili algebarski komplement matrice $\boldsymbol{A}$, $A_{ij} = (-1)^{i+j}M_{ij}$, a M_{ij} minor je determinanta matrice koja preostaje kada matrici $\boldsymbol{A}$ izostavimo redak i i stupac j.

Ako je determinanta matrice jednaka 0, matrica je singularna, u suprotnome je regularna.

Pravila o determinanti za matrice $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{M}_n$, I:

- $\det(c \cdot \boldsymbol{A}) = c^n \cdot \det(\boldsymbol{A})$
- $\det(I) = 1$
- $\det(\boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{A}) \cdot \det(\boldsymbol{B})$
- $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$

5.1.4. Sustavi linearnih jednadžbi, matrične jednadžbe

Za sustav $\boldsymbol{Ax} = \boldsymbol{b}$, $\boldsymbol{A} \in \mathcal{M}_n$, $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^N$ kažemo da:

- Ima jedinstveno rješenje ako je $\boldsymbol{A}$ regularna matrica.
- Ima jedinstveno rješenje, i to trivijalno ako je $\boldsymbol{A}$ regularna matrica i $\boldsymbol{b}$ je nul-vektor.
- Ako je $\boldsymbol{A}$ singularna matrica i $\boldsymbol{b}$ je ne-nul vektor, rješenje ne možemo zapisati u matričnom zapisu.
- Sustav ima barem trivijalno rješenje $\boldsymbol{x} = \boldsymbol{0}$ (nul-vektor) ako je $\boldsymbol{b}$ je nul-vektor. Ako je $\boldsymbol{A}$ singularna matrica i ako taj sustav ima trivijalno rješenje, onda ima beskonačno mnogo rješenja.

Za sustav $\boldsymbol{Ax} = \boldsymbol{b}$, $\boldsymbol{A} \in \mathcal{M}_{mn}$, $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^N$ kažemo da ima rješenje ako i samo ako je: $r(\boldsymbol{A}) = r(\boldsymbol{A}|\boldsymbol{b})$. U suprotnome (tj. ako je $r(\boldsymbol{A}) < r(\boldsymbol{A}|\boldsymbol{b})$), sustav nema rješenje.

Sustavi se mogu rješavati Gauss-Jordanovom metodom eliminacija, kojom se nizom elementarnih transformacija nad retcima dolazi do sustava koji je ekvivalentan početnome, a iz kojeg je moguće jednostavno iščitati rješenja sustava.

Za sustav $\boldsymbol{AX} = \boldsymbol{B}$, $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X} \in \mathcal{M}_n$ moguće je izračunati vrijednosti u matrici $\boldsymbol{X}$ tako da sustav izmnožimo s lijeva s inverznom matricom $\boldsymbol{A}^{-1}$: $\boldsymbol{A}^{-1} \cdot / \boldsymbol{AX} = \boldsymbol{B}$, pa vrijedi: $\boldsymbol{A}^{-1}\boldsymbol{AX} = \boldsymbol{A}^{-1}\boldsymbol{B}$, tj. $\boldsymbol{IX} = \boldsymbol{A}^{-1}\boldsymbol{B}$, odnosno: $\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B}$. Naravno, uz pretpostavku da je $\boldsymbol{A}$ regularna matrica.

Za sustav $\boldsymbol{XA} = \boldsymbol{B}$, $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X} \in \mathcal{M}_n$ moguće je izračunati matricu $\boldsymbol{X}$ tako da sustav pomnožimo s desna s inverznom matricom $\boldsymbol{A}^{-1}$: $\boldsymbol{XA} = \boldsymbol{B} \ / \cdot \boldsymbol{A}^{-1}$(s desna), pa vrijedi: $\boldsymbol{X}\,\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{B}\,\boldsymbol{A}^{-1}$, tj. $\boldsymbol{XI} = \boldsymbol{B}\boldsymbol{A}^{-1}$, odnosno: $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{A}^{-1}$. Naravno, uz pretpostavku da je $\boldsymbol{A}$ regularna matrica.

5.1.5. Svojstvene vrijednosti i svojstveni vektori

Neka je $\boldsymbol{A} \in \mathcal{M}_n$. Svojstvena (karakteristična, vlastita) vrijednost matrice $\boldsymbol{A}$ je realni broj λ ako postoji vektor $\boldsymbol{v}$ ™ P^n, $\boldsymbol{v} \neq \boldsymbol{0}$, takav da je $\boldsymbol{Av} = \lambda\boldsymbol{v}$. Vektor $\boldsymbol{v}$ je svojstveni (karakteristični, vlastiti) vektor matrice A.

Neka je $\boldsymbol{A} \in \mathcal{M}_n$, $\lambda_1, \lambda_2, ..., \lambda_n$ svojstvene vrijednosti, a $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_n$ korespondentni svojstveni vektori matrice $\boldsymbol{A}$. Neka su stupci matrice $\boldsymbol{P}$ upravo svojstveni vektori, to jest neka je

$$\boldsymbol{P} = [\, \boldsymbol{v}_1 \;\; \boldsymbol{v}_2 \;\; \ldots \;\; \boldsymbol{v}_n].$$

Ako je $\boldsymbol{P}$ regularna matrica, tada je

$$\boldsymbol{P}^{-1}\boldsymbol{AP} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \Lambda.$$

Posljednja matrica na glavnoj dijagonali ima svojstvene vrijednosti matrice $\boldsymbol{A}$; kažemo da smo matricu $\boldsymbol{A}$ dijagonalizirali.

Neka je $\boldsymbol{A}$ simetrična matrica. Ako su sve svojstvene vrijednosti matrice $\boldsymbol{A}$ pozitivne (negativne), tada kažemo da je matrica $\boldsymbol{A}$ pozitivno definitna (negativno definitna). Ako su neke vrijednosti jednake nuli, dok su ostale pozitivne (negativne), tada kažemo da je $\boldsymbol{A}$ nenegativno (nepozitivno) definitna. Ako $\boldsymbol{A}$ ima i negativne i pozitivne svojstvene vrijednosti, tada ju nazivamno indefinitnom.

5.2. Diferencijalni račun

5.2.1. Derivacije

Za funkciju jedne varijable $y = f(x)$, $f : A \rightarrow B$, označavamo prvu derivaciju kao: $y' = \frac{df}{dx}$, drugu derivaciju: $y'' = \frac{d^2 f}{dx^2}$.

Za funkciju više varijabli $y = f(x_1, x_2, \ldots, x_n)$ označavamo prvu parcijalnu derivaciju po varijabli i kao: $y_{x_i} = f_1 = \frac{\partial f}{\partial x_i}$, drugu parcijalnu derivaciju po varijabli i kao: $y_{x_i x_i} = f_{11} = \frac{\partial^2 f}{\partial x_i^2}$. Općenito drugu parcijalnu derivaciju možemo zapisati kao f_{ij}.

Za funkciju vektora $y = f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_n)$, $f : A \rightarrow B$, označavamo vektor parcijalnih derivacija kao vektor-gradijent, vektor čije su komponente parcijalne derivacije prvog reda: $\nabla f = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$. Matricu parcijalnih derivacija drugog reda nazivamo Hesseovom matricom:

$$H = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{22} & f_{21} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}.$$

Zbog Schwartz-youngova teorema o jednakosti mješovitih parcijalnih derivacija je Hesseova matrica simetrična.

5.2.2. Ekstremi

Neka je $f : (a, b) \rightarrow \mathrm{P}$ realna funkcija jedne realne varijable, definirana na intervalu $(a, b) \subseteq \mathrm{P}$. Kažemo da je $x_0 \in (a, b)$ točka lokalnog maksimuma funkcije f, ako postoji ε-okolina $N(x_0) = (x_0 - \varepsilon, x_0 + \varepsilon) \subset (a, b)$ točke x_0 tako da vrijedi $f(x_0) \geq f(x)$, $\forall x \in N(x_0)$. Za minimum bismo pisali umjesto znaka „≥“ znak „≤“. Nadalje, kada bismo umjesto ε-okoline razmatrali cijelu domenu funkcije f, tada bismo govorili o globalnom maksimumu ili minimumu.

Nužan uvjet za postojanje ekstrema funkcije f jest da postoji stacionarna točka x_0. Za stacionarnu točku vrijedi $f'(x_0) = 0$ i da je iz područja definicije funkcije. Dovoljan uvjet jest da je $f''(x_0) > 0$ za točku lokalnog minimuma, odnosno $f''(x_0) < 0$ za točku lokalnog maksimuma.

Neka je $f : A \rightarrow \mathrm{P}$, $A \subseteq \mathrm{P}^n$ neprekidna funkcija n varijabli, gdje je $(x_1, x_2, \ldots, x_n) \in A$, s neprekidnim parcijalnim derivacijama prvog i drugog reda. Kažemo da je $(x_1^0, x_2^0, \ldots, x_n^0) \in A$ točka globalnog (ili apsolutnog) minimuma funkcije f ako vrijedi $f(x_1^0, x_2^0, \ldots, x_n^0) \leq f(x_1, x_2, \ldots, x_n)$, $\forall (x_1, x_2, \ldots, x_n) \in A$. Za maksimum bismo pisali umjesto znaka „≤“ znak „≥“.Kada bismo razmatrali ε-okolinu točke $(x_1^0, x_2^0, \ldots, x_n^0)$, tada bismo govorili o lokalnom minimumu ili maksimumu funkcije f.

Nužan uvjet za postojanje lokalnog ekstrema funkcije više varijabli jest da su sve parcijalne derivacije prvog reda funkcije jednake nuli, tj. da je vektor-gradijent jednak nul-vektoru:

$$\nabla f = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Rješavanjem sustava $\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ nalazimo sve stacionarne točke.

Dovoljan uvjet za postojanje lokalnog ekstrema jest da je Hesseova matrica:

- Za lokalni minimum pozitivno definitna.
- Za lokalni maksimum negativno definitna.

Definitnost simetričnih matrica se može ispitivati i pomoću glavnih minora matrice (tj. determinanti na presjeku prvih i redaka i prvih i stupaca). Ako je su svi glavni minori pozitivni, tada je matrica pozitivno definitna, dok je matrica negativno definitna kada glavni minori alterniraju predznakom, s time da je prvi minor negativan.

5.3. Osnove teorije vjerojatnosti

5.3.1. Uvodno o vjerojatnosti

Slučajni eksperiment je postupak koji se ponavlja proizvoljan broj puta, pri čemu može završiti s dva ili više ishoda. Skup svih rezultata slučajnog eksperimenta naziva se prostor elementarnih događaja. Slučajni događaj je podskup prostora elementarnih događaja.

Vjerojatnost je brojčana mjera nastanka slučajnih događaja, ili drugim riječima, vjerojatnost slučajnog događaja je mogućnost ostvarenja tog događaja, izražen numerički. Tri definicije vjerojatnosti su:

- Klasična definicija, koja polazi od pretpostavke da slučajni pokus ima konačan broj mogućih ishoda, koji su svi jednako mogući. Tada je vjerojatnost nastupa događaja A jednaka $P(A) = m / n$, gdje je m broj povoljnih ishoda (za realizaciju događaja A) i n broj ukupnih ishoda.
- Subjektivna vjerojatnost je određena na temelju prosudbe istraživača.
- Statistička vjerojatnost je granična vrijednost relativne frekvencije povoljnog ishoda događaja A kada broj ponavljanja pokusa neizmjerno raste: $P(A) = \lim_{n\to\infty} \frac{m}{n}$.

Vjerojatnost je broj u intervalu [0,1]. Ako iznosi 0, tada kažemo da je nemoguć događaj, a ako je jednaka 1, tada ga nazivamo sigurnim događajem.

Uvjetna vjerojatnost je vjerojatnost pojave događaja B, uz uvjet da se ostvario događaj A: $P(B \mid A) = \frac{P(AB)}{P(A)}, P(A) > 0$, gdje je $P(AB)$ vjerojatnost nastupa i događaja A i događaja B.

5.3.2. Slučajna varijabla

Slučajna varijabla je ona čiji su mogući ishodi (realizacije) realni brojevi, tj. opaženi ishodi određenog procesa koji generira podatke (engl. *DGP*, *data generating process*). Slučajna varijabla može biti diskretna ili kontinuirana. Oznaka za slučajnu varijablu je obično veliko koso slovo, kao npr. varijabla X, dok će realizacije slučajne varijable biti malo koso slovo, x, ili pak pojedinačna realizacija i: x_i.

Diskretna slučajna varijabla je ona koja poprima prebrojivo mnogo vrijednosti s vjerojatnostima $P(X = x_i) = f(x_i), f(x_i) \geq 0, \sum_{i=1}^{N} f(x_i) = 1$.

Funkcija vjerojatnosti diskretne slučajne varijable, $f(x_i)$, je funkcija koja svakoj vrijednosti slučajne varijable pridružuje vjerojatnost nastupa (realizacije).

Distribucija vjerojatnosti diskretne slučajne varijable je skup uređenih parova $\{(x_i, f(x_i)), i \in \{1, 2,\ldots, N\}\}$.

Funkcija distribucije vjerojatnosti je definirana formulom $F(x) = P(X \leq x)$.

Za diskretnu slučajnu varijablu (X, Y) je zajednička funkcija gustoće vjerojatnosti (engl. *joint pdf*) jednaka $f(x, y) = P(X = x, Y = y)$. Vrijedi: $f(x, y) \geq 0, \sum_x \sum_y f(x, y) = 1$, te je zajednička funkcija distribucije dana kao $F(x, y) = P(X \leq x, Y \leq y)$.

Uvjetna funkcija gustoće vjerojatnosti dana je formulom $f(x \mid y) = P(X = x \mid Y = y) = \frac{f(x, y)}{f(y)}$ i čita se kao uvjetna vjerojatnost da varijabla X poprimi vrijednost x uz uvjet da je varijabla Y poprimila vrijednost y.

Zajednička funkcija gustoće $f(x, y)$ se sada može izraziti i kao: $f(x, y) = f(y) \cdot f(x \mid y)$. Za stohastički nezavisne slučajne varijable vrijedi da su zajedničke vjerojatnosti jednake umnošku pojedinačnih (graničnih) vjerojatnosti: $f(x, y) = f(x) \cdot f(y)$.

Kontinuirana slučajna varijabla je ona koja poprima neprebrojivo mnogo vrijednosti, čija su vjerojatnosna svojstva opisana funkcijom gustoće vjerojatnosti $f(x)$ (engl. pdf, *probability density function*).

Svojstva funkcije $f(x)$ su: $f(x) \geq 0$, $\int_{-\infty}^{\infty} f(x)dx = 1$, te je $F(x)$ funkcija distribucije vjerojatnosti takva da: $P(a < X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$, $a, b \in \mathbb{R}$.

Funkcija distribucije $F(x)$ može se računati izrazom $F(x) = P(X \leq x) = P(-\infty < X \leq x) = \int_{-\infty}^{x} f(t)dt$.

Za kontinuiranu slučajnu varijablu (X, Y) je zajednička funkcija gustoće vjerojatnosti sa svojstvima: $f(x, y) \geq 0$, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1$, te vjerojatnost da je $X \in (a, b]$ i $Y \in (c, d]$ iznosi: $P(a < x \leq b, c < y \leq d) = \int_{a}^{b}\int_{c}^{d} f(x, y)dxdy$.

Zajednička funkcija distribucije računa se kao: $P(X \leq a, Y \leq c) = \int_{-\infty}^{a}\int_{-\infty}^{c} f(x, y)dxdy$.

5.3.3. Distribucije vjerojatnosti

5.3.3.1. Uvodne oznake

Distribucije vjerojatnosti opisane su momentima slučajne varijable, pri čemu se najčešće koriste očekivana vrijednost i varijanca. Najprije navodimo osnovne oznake koje se koriste u zapisima distribucija vjerojatnosti.

Zbroj ili sumacija za diskretnu varijablu: $\sum_{i=1}^{N} x_i = x_1 + x_2 + ... + x_N$, oznaka za sumu veliko grčko slovo sigma: Σ. Svojstva operatora zbroja:

- $\sum_{i=1}^{N} k = Nk$, gdje je k konstanta
- $\sum_{i=1}^{N} kx_i = k\sum_{i=1}^{N} x_i$
- $\sum_{i=1}^{N} (x_i + y_i) = \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i$
- $\sum_{i=1}^{N} (a + bx_i) = Na + b\sum_{i=1}^{N} x_i$, gdje su a i b konstante

Zbroj ili sumacija za kontinuiranu varijablu: $\int_a^b f(x)dx = F(b) - F(a)$.

Očekivana vrijednost slučajne varijable je $E(X)$. Svojstvo[47] očekivane vrijednosti, za $a, b \in \mathrm{P}$ je $E(aX + b) = aE(X) + b$.

Varijanca diskretne slučajne varijable je $\mathrm{Var}(X) = E(X - E(X))^2 = E(X - \mu)^2$. Drugi pozitivni korijen od varijance je standardna devijacija σ_x. Svojstvo varijance, za $a, b \in \mathrm{P}$ je $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

Očekivana vrijednost linearne kombinacije varijabli X i Y za $a, b \in \mathrm{P}$ je $E(aX + bY) = aE(X) + bE(Y)$. Ako su X i Y nezavisne slučajne varijable, tada je $E(XY) = E(X) \cdot E(Y)$.

Kovarijanca slučajnih varijabli X i Y definira se kao: $\mathrm{Cov}(X, Y) = E[(X - E(X)) \cdot (Y - E(Y))] = E(XY) - E(X) \cdot E(Y)$.

Svojstva kovarijance su:

- $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$
- $\mathrm{Cov}(X, bY) = b\,\mathrm{Cov}(X, Y)$, $b \in \mathrm{P}$
- $\mathrm{Cov}(X, b) = 0$, $b \in \mathrm{P}$

Koeficijent linearne korelacije je relativna mjera jakosti i smjera linearne veze između varijabli X i Y, računa se kao: $\rho(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$. Vrijedi: $-1 \le \rho \le 1$.

Ako se računa $\mathrm{Var}(X + Y)$, izvod je sljedeći:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= E\left[(X+Y) - E(X+Y)\right]^2 = E\left[(X+Y) - E(X) - E(Y)\right]^2 \\
&= E\left[(X - E(X)) + (Y - E(Y))\right]^2 \\
&= E\left[(X - E(X))^2 + 2(X - E(X))E(Y - E(Y)) + (Y - E(Y))^2\right] \\
&= \mathrm{Var}(X) + 2\mathrm{Cov}(X,Y) + \mathrm{Var}(Y)
\end{aligned}$$

Ako se radi o nezavisnim varijablama (iznad navedeno $E(XY) = E(X) \cdot E(Y)$), tada za $\mathrm{Cov}(X,Y)$ vrijedi:

$$\begin{aligned}
\mathrm{Cov}(X+Y) &= E\left[(X - E(X))(Y - E(Y))\right] = E\left[XY - YE(X) - XE(Y) + E(X)E(Y)\right] \\
&= E(XY) - E(YE(X)) - E(XE(Y)) + E(E(X)E(Y)) \\
&= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y) \\
&= E(X)E(Y) - E(X)E(Y) = 0
\end{aligned}$$

[47] Dokaze svih svojstava vidjeti u Balakrishnan i dr. (2020).

$$\text{pa je } \rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y} = \frac{0}{\sigma_x \sigma_y} = 0 .$$

Stoga vrijedi: ako su varijable nezavisne, tada su i nekorelirane. S druge strane, varijable mogu biti nekorelirane, ali istovremeno i zavisne!

5.3.3.2. Normalna distribucija

Funkcija gustoće vjerojatnosti za slučajnu varijablu X, za prosječnu vrijednost μ i standardnu devijaciju σ je

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

te se označava da $X \sim N(\mu, \sigma^2)$. Normalna distribucija je simetrična oko prosječne vrijednosti ($\alpha_3 = 0$), te je koeficijent zaobljenosti jednak $\alpha_4 = 3$.

Ako se razmatra linearna transformacija $Z = \frac{X-\mu}{\sigma}$, radi se o standardiziranoj normalnoj varijabli, čije je očekivanje 0 i varijanca jednaka 1:

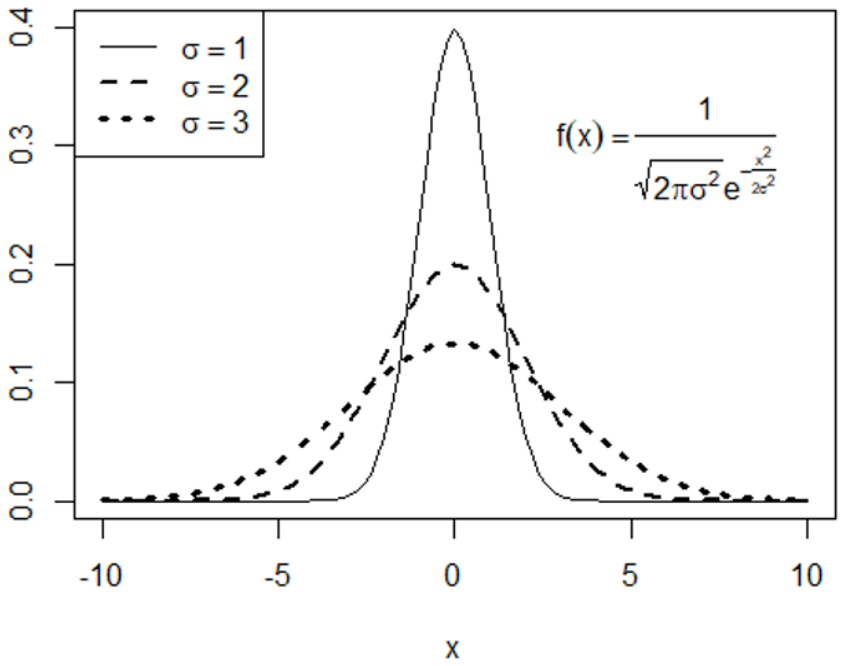
$$E\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma} E\left[X-\mu\right] = \frac{1}{\sigma} E(X) - \mu = 0 ,$$

$$\text{Var}\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma^2} \text{Var}\left[X-\mu\right] = \frac{1}{\sigma^2} \text{Var}\left[X\right] = 1 .$$

Stoga je funkcija gustoće vjerojatnosti sljedeća:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2},$$

te se označava da $Z \sim N(0, 1)$. Na slici A1 prikazana su tri primjera normalnih distribucija, pri čemu svaka ima očekivanu vrijednost 0, dok su standardne devijacije redom 1, 2 i 3.



Slika A1. Tri normalne distribucije

5.3.3.3. Hi-kvadrat distribucija

Ako je $Z \sim N(0,1)$, tada je $X = Z^2 \sim$ hi-kvadrat distribucija s jednim stupnjem slobode, u oznaci: $X \sim \chi^2(1)$.

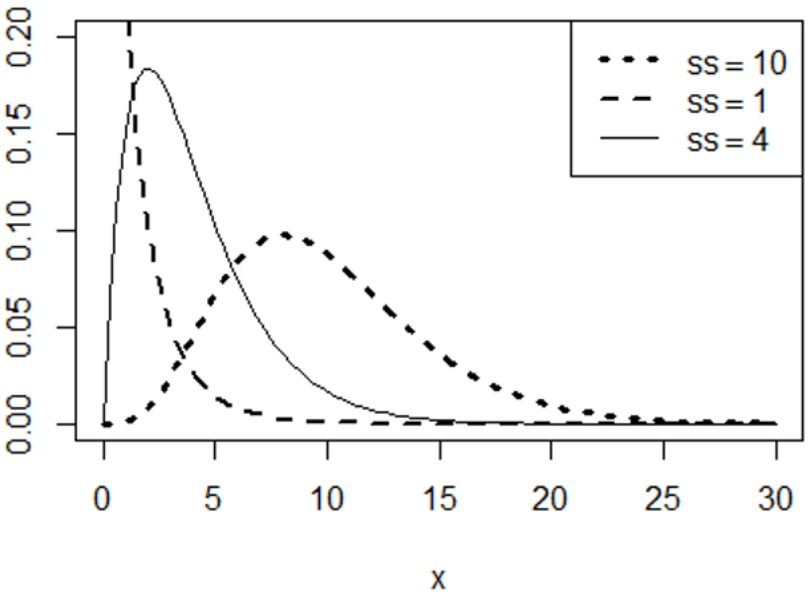
Ako su $(x_1, x_2, \ldots, x_n)$ nezavisne varijable koje slijede $\chi^2(1)$, tada je $\sum_{i=1}^{n} x_i \sim \chi^2(n)$.

Ako su $(Z_1, Z_2, \ldots, Z_n)$ nezavisne varijable koje slijede $N(0,1)$, tada je $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$.

Ako su $(Z_1, Z_2, \ldots, Z_n)$ nezavisne varijable koje slijede $N(0,\sigma^2)$, tada je $\sum_{i=1}^{n} \left(\frac{Z_i}{\sigma} \right)^2 \sim \chi^2(n)$.

Ako su x_1 i x_2 dvije nezavisne varijable koje slijede $x_1 \sim \chi^2(n_1)$ i $x_2 \sim \chi^2(n_2)$, tada je $x_1 + x_2$ varijabla koja $x_1 + x_2 \sim \chi^2(n_1 + n_2)$.

Na slici A2 prikazana su tri primjera hi-kvadrat distribucija, pri čemu se mijenjaju stupnjevi slobode (oznaka ss na slici).

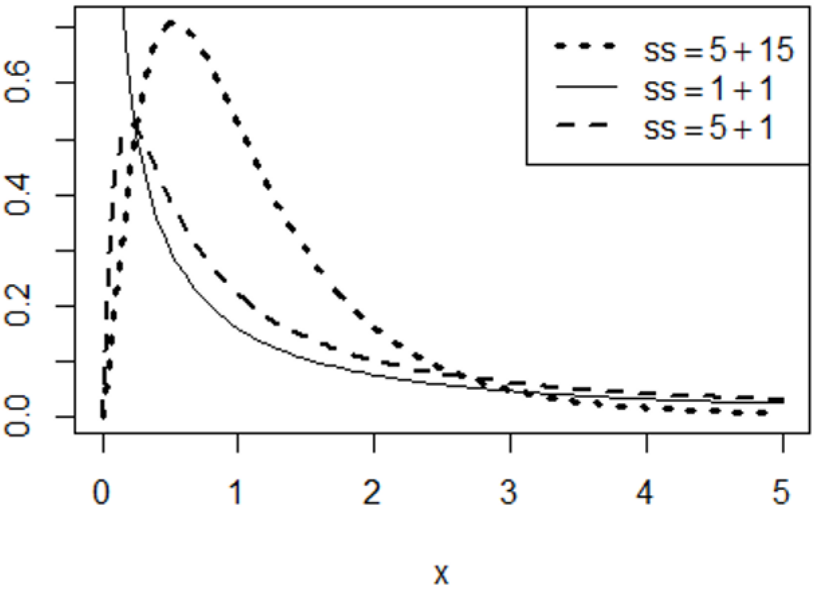


Slika A2. Tri hi-kvadrat distribucije

5.3.3.4. F-distribucija

Ako su x_1 i x_2 dvije nezavisne varijable koje slijede $x_1 \sim \chi^2(n_1)$ i $x_2 \sim \chi^2(n_2)$, tada omjer $F(n_1, n_2) = \frac{x_1 / n_1}{x_2 / n_2}$ slijedi F-distribuciju s n_1 stupnjeva slobode u brojniku i n_2 stupnjeva slobode u nazivniku.

Na slici A3 prikazana su tri primjera F-distribucija, pri čemu se mijenjaju stupnjevi slobode (oznaka ss na slici).



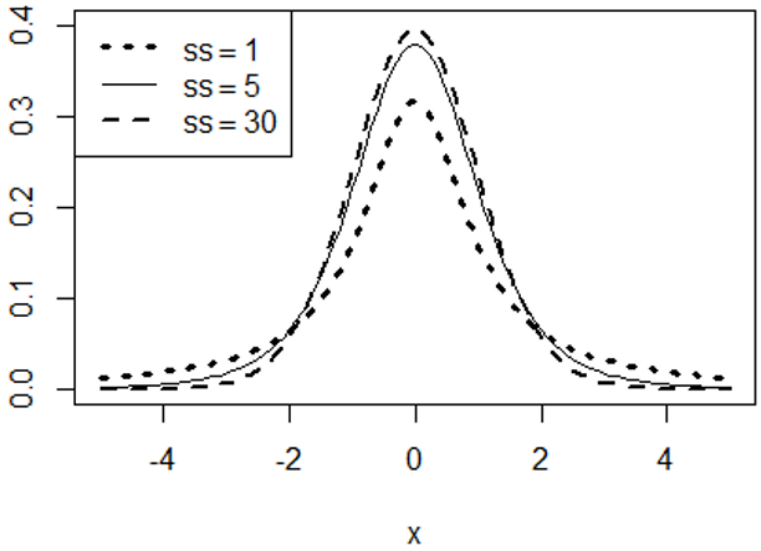
Slika A3. Tri F-kvadrat distribucije

5.3.3.5. Studentova distribucija

Ako je $Z \sim N(0,1)$, te je $X \sim \chi^2(n)$, te su Z i X međusobno nezavisne, tada omjer $t(n) = \frac{Z}{\sqrt{X/n}}$ slijedi *t*-distribuciju ili Studentovu distribuciju s n stupnjeva slobode.

Ako je $t \sim t(n)$, tada je $t^2 \sim F(1,n)$.

Na slici A4 prikazana su tri primjera *t*-distribucija, pri čemu se mijenjaju stupnjevi slobode (oznaka ss na slici).



Slika A4. Tri Studentove distribucije

5.3.3.6. Stupnjevi slobode

Stupnjevi slobode odnose se na broj vrijednosti (opservacija) koje se koriste za izračun neke mjere koji mogu slobodno varirati (mijenjati se), odnosno broj nezavisnih dijelova informacija koje se koriste za procjenu parametara.

Primjerice, varijanca uzorka obično se procjenjuje formulom $\hat{\sigma}_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$, gdje je nazivnik umanjen za vrijednost 1, što znači da je broj stupnjeva slobode jednak $N-1$. Ideja je da upravo vrijednost koju smo dobili $\hat{\sigma}_y^2$ temeljem odabranog uzorka je dobivena temeljem N vrijednosti koje smo prikupili za varijablu Y, umanjena za jedan parametar koji se procjenjuje prilikom procjene varijance, a to je prosječna ili očekivana vrijednost $\bar{y}$.

5.3.4. Testiranje hipoteza

Formalna procedura testiranja hipoteza sastoji se od sljedećih koraka. Najprije se formiraju nulta i alternativna hipoteza, H_0 i H_1. Potom se računa empirijska test veličina, temeljem prikupljenih podataka, za koju se pretpostavlja da je slučajna varijabla, jer je ideja da je uzorak koji je prikupljen također slučajan. Ako test veličina upada u područje ne odbacivanja nulte hipoteze, tada se ona ne odbacuje. U suprotnom se odbacuje.

Međutim, prilikom testiranja hipoteza, postoje dva tipa grešaka koje se mogu učiniti: greška tipa I, te greška tipa II:

- Greška tipa I: procedura testiranja rezultira odbacivanjem nulte hipoteze, a ona je istinita (engl. *false positive*)
- Greška tipa II: procedura testiranja rezultira ne odbacivanjem nulte hipoteze, a ona je lažna (engl. *false negative*)

Vjerojatnost greške tipa I se naziva veličina testa, označava se kao α, te se naziva razinom značajnosti (signifikantnosti). Vjerojatnost da točno odbacujemo lažnu nultu hipotezu je snaga testa, računa se kao $1-\beta = 1-P(\text{greška tipa II})$ (vidjeti tablicu A1).

Tablica A1. Tipovi grešaka

Donosi se odluka:	H_0 je zapravo:	
	Istinita	Lažna
Odbaciti H_0	Greška tipa I (vjerojatnost α)	Točno (vjerojatnost $1-\beta$)
Ne odbaciti H_0	Točno (vjerojatnost $1-\alpha$)	Greška tipa II (vjerojatnost β)

5.3.5. Procjenjivanje parametara i svojstva procjenitelja

Ideja inferencijalne statistike je primijeniti skup metoda kojima se temeljem prikupljenih podataka nad uzorkom donose zaključci o populaciji. Uzorak je podskup populacije, pri čemu se parametri od interesa računaju za uzorak kako bi se procijenio parametar populacije.

Za nepoznati populacijski parametar θ temeljem uzorka veličine N njegov procjenitelj $\hat{\theta}$ je slučajna varijabla. Procjena parametra može se vršiti jednim brojem (engl. *point estimate*) ili pak intervalnom (engl. *interval estimate*).

Poželjna svojstva procjenitelja su sljedeća: svojstva malog uzorka te asimptotska svojstva (svojstva velikog uzorka).

Svojstva malog uzorka su nepristranost i efikasnost, dok su asimptotska svojstva nepristranost i konzistentnost.

Nepristranost procjenitelja $\hat{\theta}$ znači da vrijedi:

$$E(\hat{\theta}) = \theta,$$

tj. procjena parametara je u prosjeku jednaka stvarnoj vrijednosti parametra. Za one procjenitelje za koje vrijedi $E(\hat{\theta}) \neq \theta$, procjenitelj se naziva pristranim.

Nepristran procjenitelj je efikasan ako i samo ako je njegova srednjekvadratna pogreška najmanja u skupu svih nepristranih procjenitelja. Srednjekvadratna pogreška (engl. MSE, *mean squared error*), računa se formulom:

$$\mathrm{MSE}(\hat{\theta}) = E((\hat{\theta} - \theta)^2) = (E(\hat{\theta}) - \theta)^2 + \mathrm{Var}(\hat{\theta}).$$

Između dva nepristrana procjenitelja, $\hat{\theta}_1$ i $\hat{\theta}_2$, bolji je $\hat{\theta}_1$ ako vrijedi $\mathrm{MSE}(\hat{\theta}_1) < \mathrm{MSE}(\hat{\theta}_2)$.

Nadalje, ako se promatra skup linearnih nepristranih procjenitelja, tada je procjenitelj koji je dobiven kao njihova linearna kombinacija najbolji linearni nepristrani procjenitelj (engl. BLUE, *best linear unbiased estimator*) ako su koeficijenti linearne kombinacije takvi da je linearni procjenitelj nepristran i efikasan.

Asimptotska nepristranost procjenitelja $\hat{\theta}$ znači da očekivana vrijednost tog procjenitelja teži pravoj vrijednosti parametra θ kada veličina uzorka neizmjerno raste:

$$\lim_{N\to\infty} E(\hat{\theta}_n) = \theta.$$

Konzistentnost procjenitelja $\hat{\theta}$ znači da je vjerojatnost da procjenitelj $\hat{\theta}_n$ s povećanjem uzorka bude dovoljno blizu stvarne vrijednosti θ:

$$\lim_{N\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) = 1, \varepsilon > 0.$$

Također se može reći da $\hat{\theta}_n$ po vjerojatnosti teži prema θ ($\hat{\theta}_n \xrightarrow{p} \theta$).

LITERATURA

1. Anderson, T. W., Darling, D. A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Annals of Mathematical Statistics, 23, str. 193–212.
2. Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown changepoint, Econometrica, 61, str. 821–856.
3. Andrews, D. W. K., Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative, Econometrica, 62, str. 1383–1414.
4. Anescombe, F. J. (1961). Examination of Reisudals, in: Proceedings oft he fourth Berkeley symposium on matehamtical statistics and probability, Berkeley, CA: University of California Press, str. 1-36.
5. Balakrishnan, N., Koutras, M. V., Politis, K. G. (2020). Introduction to Probability – Models and Applications, Wiley, New Jersey.
6. Breusch, T. S. (1978). Testing for Autocorrelation in Dynamic Linear Models. Australian Economic Papers, 17, str. 334–355.
7. Brooks, C. (2014). Introductory Econometrics for Finance, 3rd Edition, University Printing House, Cambridge, United Kingdom.
8. Chow, Gregory C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. Econometrica, 28(3), str. 591–605.
9. Cochrane, D., Orcutt, G. H. (1949). Application of Least Squares Regression to Relationships Containing Auto-Correlated Error Terms. Journal of the American Statistical Association, 44(245), str. 32–61.
10. Cramér, H. (1928). On the Composition of Elementary Errors. Scandinavian Actuarial Journal, 1928 (1), str. 13–74.
11. Durbin, J., Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression, I. Biometrika. 37(3–4), str. 409–428.
12. Durbin, J., Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression, II. Biometrika. 38(1–2), str. 159–179.
13. EC (2018), European Commission, Handbook on Seasonal Adjustmetn. Eurostat, Luxembourg. doi: 10.2785/941452.
14. Engle, R., Granger, C. W. J. (1991). Cointegration and Error Correction: Representation, Estimation, and Testing. In Engle and Granger (eds.), Long Run Economic Readings in Cointegration, Oxford University Press, New York, str. 81-113.
15. Engle, R., Granger, C. W. J., (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. Econometrica, 55(2), str. 251-276.
16. Eurostat (2020). https://ec.europa.eu/eurostat/data/database.
17. Farrar, D. E., Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisted. The Review of Economics and Statistics, 49(1), str. 92–107.
18. Glejser, H. (1969). A new test for heteroskedasticity. Journal oft he American Statistical Association, 64(325), str. 316-323.
19. Godfrey, L. G. (1978). Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables. Econometrica, 46, str. 1293–1301.
20. Goldberger, A. S. (1964). Econometric Theory. New York: John Wiley and Sons.
21. Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37, str. 424–438.
22. Greene, W. H. (2002). Econometric analysis, 5th edition, New Jersey: Prentice Hall, USA.
23. Greene, W. H. (2018). Econometric analysis, 8th edition, New York: Pearson.

24. Guidolin, M., Pedio, M. (2019). Essentials of Time Series for Financial Applications. UK: Academic Press.
25. Gujarati, D. N., Porter, D. C. (2010). ESSEntials of Econometrics, 4th Edition, The McGraw-Hill Companies, Inc., New York, USA.
26. Gujarati, D. N., Porter, D. C. (2010). ESSEntials of Econometrics, Fourth edition. New York: McGraw-Hill, USA.
27. Hayashi, F. (2002). Econometrics. Princeton University Press.
28. Jarque, C. M., Bera, A. K. (1987). A test for normality of observations and regression residuals. International Statistical Review, 55(2), str. 163–172.
29. Klein, L. R. (1962). An Introduction to Econometrics. Prentice-Hall, Englewood, Cliffs, N. J.
30. Kolmogorov A (1933). Sulla determinazione empirica di una legge di distribuzione. dell'Istituto Italiano degli Attuari, 4, str. 83–91.
31. Kovács, P., Petres, T., Tóth, L. (2005). A new measure of multicollinearity in linear regression models. International Statistical Review / Revue Internationale de Statistique, 73(3), str. 405–412.
32. Kovács, P., Petres, T., Tóth, L. (2005). A new measure of multicollinearity in linear regression models. International Statistical Review/Revue Internationale de Statistique, 73(3), str. 405–412.
33. Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association, 62(318), str. 399–402.
34. Ljung, G. M., Box, G. E. P. (1978). On a Measure of Lack of Fit in Time Series Models, Biometrika, 65(2), str. 297–303.
35. Maddala, G. S. (1988). Introduction to Econometrics. Macmillan Publications, New York.
36. Maddala, G. S. (1992). Introduction to Econometrics, 2nd edition, Macmillan Publications, Toronto.
37. Maddala, G. S., Lahiri, K. (2009). Introduction to Econometrics, 4th Edition, John Wiley and Sons.
38. Martić, Lj. (1991 a). Bikriterijalno programiranje u regresijskoj analizi, Zbornik radova KOI'91, str. 37-46, Zagreb.
39. Martić, Lj. (1991 b). Jednostavna regresija po kriterijima l1 i l2 norme, Zbornik radova KOI'92, str. 17-32, Rovinj.
40. Newey, W. K., West, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55(3), str. 703–708.
41. Page, E. S. (1954). Continuous Inspection Scheme. Biometrika. 41(1/2), str. 100–115.
42. Pesaran, M. H. (1987). Econometrics. The New Palgrave Dictionary of Economics, 1–25. doi:10.1057/978-1-349-95121-5_188-1.
43. Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. Journal of the Royal Statistical Society, Series B, 31(2), str. 350–371.
44. Razali, Nornadiah; Wah, Yap Bee (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. Journal of Statistical Modeling and Analytics, 2(1), str. 21–33.
45. Ruggins, S. M. (2015). A History of Econometrics: The Reformation from the 1970s. Journal of Cultural Economy, 9(2): str. 226-228.
46. Samuelson, P. A., Koopmans, T. C., Stone, J. R. N. (1954). Report of the Evaluative Committee for Econometrica. Econometrica, 22(2), str. 141–146.

47. Sarapa, N. (2002). Teorija vjerojatnosti. Zagreb: Školska knjiga.
48. Shapiro, S. S., Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), str. 591–611.
49. Smirnov, N.V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. Bull Moscow University, 2(2), str. 3–16.
50. Svjetska banka (2020). https://databank.worldbank.org/
51. Šego, B., Aljinović, Z., Marasović, B. (2011). Financijsko modeliranje. 2. izmjenjeno i dopunjeno izdanje. Split: Ekonomski fakultet – Split.
52. Šego, B., Škrinjarić, T. (2012). Modeliranje dnevne sezonalnosti prinosa na Zagrebačkoj burzi. U: Matematički modeli u analizi razvoja hrvatskog financijskog tržišta, Aljinović, Zdravka ; Marasović, Branka (ur.). Split: Ekonomski fakultet u Splitu, str. 159-172.
53. Šošić, I. (2006). Statistika, 2. izmijenjeno izdanje. Školska knjiga, Zagreb.
54. Theil, H. (1971). Principles of Econometrics. John Wiley & Sons, New York
55. Theil, H. (1971). Principles of Econometrics. John Wiley & Sons, New York.
56. Tintner, G. (1953). The Definition of Econometrics. Econometrica, 21(1), str. 31-40.
57. Verbeek, M. (2004). A Guide to Modern Econometrics, 2nd edition. John Wiley & Sons, New York.
58. von Mises, R. E. (1928). Wahrscheinlichkeit, Statistik und Wahrheit. Julius Springer.
59. Wei, W. S. (2005). Time Series Analysis - Univariate and Multivariate Methods, Second edition. Addison Wesley, USA.
60. White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity, Econometrica, 48, str. 817–838.
61. Wooldridge, J. F. (2016). Introductory Econometrics, A modern approach, 6th edition. Boston: Cengage Learning, USA.
62. Yap, B. W., Sim, C. H. (2011). Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation, 81(12), str. 2141-2155.
63. Zeileis, A. (2000). P-values and alternative boundaries for CUSUM tests. Working Paper 78, SFB Adaptive Information Systems and Modelling in Economics and Management Science.

POPIS POJMOVA

Izdavač:
Hrvatska narodna banka

Urednici:
Vedran Šošić
Davor Kunovac

Izvršni urednik:
Katja Gattin Turkalj

Recenzent:
izv. prof. dr. sc. Petar Sorić

Grafički urednik:
Slavko Križnjak

Dizajn naslovnice:
Vjekoslav Gjergja

Lektura predgovora:
Sanda Uzun-Ikić

Tehnički urednik:
Nevena Jerak Muravec

Fotografija:
Božidar Bengez

HRVATSKA NARODNA BANKA

ODABRANE TEME PRIMIJENJENE EKONOMETRIJE
LINEARNI REGRESIJSKI MODEL

ISBN 978-953-8013-10-2