Linearna regresija s vremenskim nizovima

Buljubašić, Ana

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike

Ana Buljubašić

Linearna regresija s vremenskim nizovima

Diplomski rad

Osijek, 2019.

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike

Ana Buljubašić

Linearna regresija s vremenskim nizovima

Diplomski rad

Mentor: izv. prof. dr. sc. Nenad Šuvak

Osijek, 2019.

Sadržaj

1 Uvod

U mnogim disciplinama društvenih znanosti često se susrećemo s pojavama koje zasebno možemo promatrati tijekom vremena, a među njima naslućujemo postojanje neke vrste veze. Kako bismo tu vezu mogli razumijeti, opisati i donijeti valjane statističke zaključke, potrebno je provesti, primjerice, regresijsku analizu vremenskih nizova, čijim ću se teorijskim osnovama i primjenom na stvarnim podacima baviti u ovom radu.

U prvom dijelu rada obradit ću teorijsku osnovu o vremenskim nizovima. Svrha je promatranja pojava kao vremenskih nizova pokušaj predviđanja i opisivanja budućih vrijednosti na temelju onih zabilježenih.

Idući bitan pojam koji će uz vremenske nizove biti temelj ovog rada jest linearna regresija. Ideja linearne regresije je za promatrane pojave definirati linearnu funkcijsku vezu te na temelju mjerenja jedne ili više pojava procijeniti vrijednost druge pojave, odnosno ustanoviti prirodu ovisnosti među njima.

Nakon uvodnog teorijskog pregleda, u nastavku rada bavit ću se linearnom regresijom vremenskih nizova navodeći primjere takvih modela te njihovih svojstava, kao i metodama procjene parametara i njihovim svojstvima. Nakon prvog dijela rada, sama ideja ovakvog modeliranja postaje jasnija, a temelji se na promatranju kronološki zabilježenih pojava koje u danom vremenskom trenutku tvore linearnu vezu. Najčešće primjer možemo vidjeti u analizi ekonomskih veličina koje zasebno promatramo u obliku vremenskog niza, a između više njih može postojati linearna zavisnost. Ovakav način modeliranja tada omogućuje promatranje efekta jedne ili više pojava na drugu u danom trenutku, ali i predikciju tog efekta u budućim trenucima.

Treći dio rada je praktični dio koji za glavni cilj ima primjenu obrađene teorije, metoda i zaključaka na stvarnim podacima. Glavna ideja je pokušati prediktirati koncentraciju peludi ambrozije na temelju meteoroloških podataka poput temperature, vlažnosti zraka, vjetra i padalina zabilježenih u jednakim vremenskim intervalima.

2 Osnovno o analizi vremenskih nizova

2.1 Vremenski niz

Često imamo potrebu neku pojavu promatrati tijekom vremena – bilježiti njezine vrijednosti u određenim vremenskim trenucima te pokušati shvatiti promjene koje se događaju. Također, na temelju tih promjena želimo zaključiti postoji li neki obrazac ponašanja pomoću kojeg možemo opisati buduće vrijednosti. Odgovore na takva, ali i mnoga druga pitanja možemo dobiti provodeći analizu vremenskih nizova.

Definicija 2.1.1. Vremenski niz jest niz podataka $x_{t_1}, ..., x_{t_n}$ prikupljenih u uzastopnim vremenskim trenucima $t_1, ..., t_n$, takvih da su $x_{t_1}, ..., x_{t_n} \in \mathbb{R}, t_1 < ... < t_n$.

Dakle, vremenski niz promatramo kao niz izmjerenih vrijednosti neke pojave tijekom vremena. Kako prošlost može utjecati na budućnost, ali ne i obrnuto, ključno svojstvo takvog niza podataka upravo je njegov kronološki poredak koji nosi potencijalno važne informacije. Ako su pripadni vremenski trenuci raspoređeni ekvidistantno, bitna nam je frekvencija mjerenja. Tako možemo analizirati podatke mjerene na, primjerice, dnevnoj, mjesečnoj ili godišnjoj bazi, što može biti vrlo važan faktor u analizi.
Cilj je analize vremenskih nizova odabrati odgovarajući model koji dobro opisuje zabilježene podatke. U tu svrhu potrebno je razumijevanje stohastičkog mehanizma koji vodi do određenih realizacija. Koncept slučajnosti u analizi vremenskih nizova dolazi iz činjenice da ne znamo koja će biti vrijednost promatrane varijable u budućnosti, posebno ako ona ovisi o vrijednostima nekih drugih varijabli (npr. koliko će iznositi BDP na kraju godine). Stoga takve pojave promatramo kao slučajne varijable, a model za vremenski niz kao slučajni proces.

Definicija 2.1.2. Slučajni proces jest familija $\{X_t; t \in T\}, T \subseteq \mathbb{R}$ slučajnih varijabli definiranih na istom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathcal{P})$.

Za tako definiran slučajni proces vrijedi sljedeće:

- $\forall t \in T$, X_t je slučajna varijabla,
- $\forall \omega \in \Omega$, $t \mapsto X_t(\omega)$ je trajektorija slučajnog procesa.

U nastavku rada bavit ćemo se procesima u diskretnom vremenu, a za njih vrijedi $T \subseteq \mathbb{N}$.
Jedan zabilježeni podatak x_{t_i} promatramo kao jednu realizaciju slučajne varijable X_t u trenutku t_i. Analogno tome, vremenski niz $x_{t_1}, ..., x_{t_n}$ tada čini jednu realizaciju slučajnog procesa $\{X_t; t \in T\}$ u trenucima $t_1, ...t_n \in T$. Kada modeliramo vremenski niz, želimo u distribucijskom smislu okarakterizirati slučajni proces $\{X_t; t \in T\}$ koji će ga moći opisati u svakom vremenskom trenutku. S obzirom na to da uvijek raspolažemo samo s dijelom trajektorije slučajnog procesa, teško je samo na temelju toga definirati sam proces. Stoga je potrebno postaviti uvjete na strukturu procesa.

2.2 Stacionarni procesi

Iz prethodnog poglavlja znamo da modelirati vremenski niz slučajnim procesom znači odrediti slučajni proces u distribucijskom smislu.
Ako se translacijom vremenskog parametra konačnodimenzionalne distribucije slučajnog procesa ne mijenjaju, tada kažemo da je proces stacionaran u užem smislu ili strogo stacionaran.

Definicija 2.2.1. Slučajni proces $\{X_t; t \in T\}$ jest strogo stacionaran ako za svaki $h > 0$ vrijedi:

$$(X_{t_1}, ...X_{t_n}) \overset{\mathrm{d}}{=} (X_{h+t_1}, ...X_{h+t_n}), \quad \forall t_1, ..., t_n \in T, \quad h \in \mathbb{N}.$$

Posebno, $X_t \overset{\mathrm{d}}{=} X_s, \forall t, s \in T$.

S obzirom na to da je često teško postići strogu stacionarnost, nekad će nam biti dovoljno, umjesto distribucije, poznavati samo neka svojstva procesa kao što su očekivanje, varijanca i kovarijanca te će nam biti važno da ta svojstva budu nepromijenjena tijekom vremena [5]. Iz tog razloga uvodi se koncept slabe stacionarnosti.

Definicija 2.2.2. Za slučajni proces $\{X_t, t \in T\}$ t.d. je $E[X_t^2] < \infty, \forall t \in T$ definiramo sljedeće funkcije:

- funkcija očekivanja procesa $\{X_t\}$ jest funkcija $\mu : T \to \mathbb{R}$ definirana s

$$\mu(t) = E[X_t]$$

- funkcija autokovarijanci procesa $\{X_t\}$ jest funkcija $\gamma : T^2 \to \mathbb{R}$ definirana s

$$\gamma(t, s) = Cov(X_t, X_s) = E[(X_t - E[X_t])(X_s - E[X_s])] = E[X_t X_s] - E[X_t]E[X_s]$$

- autokorelacijska funkcija procesa $\{X_t\}$ jest funkcija $\rho : T^2 \to [-1, 1]$ definirana s

$$\rho(t, s) = Corr(X_t, X_s) = \frac{Cov(X_t, X_s)}{\sqrt{\gamma(t, t)\gamma(s, s)}}.$$

Definicija 2.2.3. Slučajni proces $\{X_t, t \in T\}$ slabo je stacionaran ako vrijedi:

(i) $E[X_t^2] < \infty, \forall t \in T$

(ii) $\mu(t) = c, \forall t \in T$

(iii) $\gamma(t, s) = \gamma(t + h, s + h), \forall t, s, h \in T$.

Za razliku od stroge stacionarnosti, koja zahtijeva neosjetljivost konačnodimenzionalnih distribucija procesa na vremenske pomake, za slabu stacionarnost dovoljno je da očekivanje i varijanca postoje i da su konstante te da funkcija autokovarijanci ovisi samo o vremenskoj razlici između dva mjerenja. U nastavku rada pojam stacionaran odnosit će se na slabu stacionarnost. Stacionarnost nam je važna jer ćemo se u idućim poglavljima baviti modeliranjem vremenskih nizova slučajnim procesima – kako bismo ih bolje razumjeli i kako bismo mogli donositi zaključke o budućim vrijednostima, potreban nam je određeni koncept stabilnosti tijekom vremena.

Definicija 2.2.4. Za stacionaran slučajni proces $\{X_t, t \in \mathbb{Z}\}$ kažemo da spada u kategoriju procesa sa slabom zavisnošću ako $Corr(X_t, X_{t+h}) \to 0$ kad $h \to \infty$.

Za takav proces još možemo reći i da je asimptotski nekoreliran.

2.2.1 Primjeri stacionarnih procesa

Neki od najvažnijih primjera stacionarnih procesa jesu nezavisni jednako distribuirani šum i bijeli šum.

Definicija 2.2.5 (NJD šum)**.** Neka je $\{X_t, t \in \mathbb{Z}\}$ niz nezavisnih jednako distribuiranih slučajnih varijabli takvih da $\forall t \in \mathbb{Z}$ vrijedi:

(i) $E[X_t^2] < \infty$

(ii) $E[X_t] = 0$

(iii) $Var(X_t) = \sigma^2$.

Tada za $\{X_t\}$ kažemo da je nezavisni jednako distribuirani (NJD) šum uz oznaku $\{X_t\} \sim IID(0, \sigma^2)$.

Definicija 2.2.6 (Bijeli šum)**.** Neka je $\{X_t, t \in \mathbb{Z}\}$ niz slučajnih varijabli takvih da $\forall t \in \mathbb{Z}$ vrijedi:

(i) $E[X_t^2] < \infty$

(ii) $E[X_t] = 0$

(iii) $Var(X_t) = \sigma^2$

(iv) $Cov(X_t, X_s) = 0, \forall t \neq s$.

Tada za $\{X_t\}$ kažemo da je bijeli šum uz oznaku $\{X_t\} \sim WN(0, \sigma^2)$.

Primijetimo da je svaki NJD šum ujedno i bijeli šum, no obratno ne vrijedi.

Definicija 2.2.7 (AR(1) proces)**.** Autoregresivni proces reda 1, oznaka AR(1), jest proces $\{X_t, t \in \mathbb{Z}\}$ zadan s

$$X_t = \phi X_{t-1} + Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2), \quad |\phi| < 1.$$

Definicija 2.2.8. Slučajni proces $\{X_t, t \in \mathbb{Z}\}$ jest autoregresivni proces reda $p \in \mathbb{N}$, ako je stacionaran i ako je

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + Z_t, \quad t \in \mathbb{Z},$$

gdje je

$$\{Z_t\} \sim WN(0, \sigma^2), \quad \phi_1, ..., \phi_p \in \mathbb{R}, \quad \phi_p \neq 0,$$

uz oznaku $\{X_t\} \sim AR(p)$.

Ideja autoregresivnih procesa je da se sadašnja vrijednost procesa prikaže kao linearna kombinacija p prošlih vrijednosti uz dodani šum.

2.3 Trendovi i sezonalnost

U praksi se često susrećemo s procesima koji nisu stacionarni. Uzroci nestacionarnosti mogu biti različiti, a jedan je od najčešćih deterministički trend.

Definicija 2.3.1. Kažemo da proces $\{X_t\}$ ima deterministički trend ako ga možemo prikazati kao $X_t = \mu_t + Y_t$, gdje je $\{Y_t\}$ stacionaran, $E[Y_t] = 0$, a $\mu : T \to \mathbb{R}$ funkcija koja nije konstanta.

Tada od nestacionarnog procesa $\{X_t\}$ možemo jednostavnom transformacijom dobiti stacionarni proces, a to je proces $\{X_t - \mu_t\}$. Ako je μ linearna funkcija, kažemo da proces ima linearni trend. S obzirom na to da radimo s podacima koji dolaze iz vremenskih nizova, često se suočavamo s pojavama trenda i sezonalnosti. U praksi često veličine koje želimo modelirati imaju tendenciju rasta tijekom vremena – bitno je uočiti to ponašanje te u skladu s time donositi zaključke. Jednostavan primjer koji naglašava važnost detektiranja trenda jest sljedeći: ako promatrana pojava ima tendenciju rasta vrijednosti tijekom vremena, a mi tu činjenicu zanemarimo, možemo se dovesti do situacije da zaključimo da je rast uzrokovan promjenama varijabli kojima modeliramo tu pojavu, što ne mora biti točno. U većini slučajeva, ako dva procesa vremenskih nizova imaju tendenciju rasta (ili pada) koji je uzrokovan nekim neizmjerenim faktorom (ono što smatramo greškom modela), oni mogu izgledati korelirano. Jedan od načina koje Wooldrige navodi u knjizi [9] kako uključiti u obzir i trend jest prikazivanje slučajnog procesa $\{Y_t\}$ na sljedeći način:

$$Y_t = \alpha_0 + \alpha_1 t + E_t, \quad t = 1, 2, ...$$

gdje je, u najjednostavnijem slučaju, $\{E_t\}$ n.j.d. niz slučajnih varijabli s $E[E_t] = 0$ i $Var(E_t) = \sigma_E^2$. Ako za svaka dva uzastopna trenutka $(t-1)$ i t definiramo razliku $\Delta E_t = E_t - E_{t-1}$, tada

uz pretpostavku $\Delta E_t = 0$ vrijedi

$$\Delta Y_t = Y_t - Y_{t-1} = \alpha_1.$$

U zapisu jasno vidimo postojanje linearnog trenda koji uzrokuje parametar α_1. Također, na taj način i u sam model, uz ostale neovisne varijable, možemo dodati vremensku komponentu kao varijablu.
Uvedimo oznaku $X_{tj}, j \in \mathbb{N}, t \in T$ za slučajnu varijablu X_j u trenutku t.

Primjer 2.3.1. Neka je dan model

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + U_t$$

za koji pretpostavljamo da postoje neizmjereni faktori koji utječu na povećanje ili smanjenje ovisne varijable tijekom vremena. Tada uključivanjem trenda dobivamo sljedeći model:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 t + U_t$$

gdje vremensku komponentu t promatramo kao treću neovisnu varijablu X_{t3}, a njoj odgovarajući parametar β_3 upućuje na smjer kretanja trenda.

Drugi način za promatranje vremenskih nizova koji imaju trend jest da očekivana vrijednost bude linearna funkcija vremena, odnosno

$$E[Y_t] = \alpha_0 + \alpha_1 t.$$

Zbog svog slučajnog karaktera realizacije slučajnih varijabli $Y_t, t \in \{t_1, \ldots, t_n\}$ nikad se ne nalaze točno na pravcu $y = \alpha_0 + \alpha_1 t$, no u kontekstu gornjeg zahtjeva očekivane vrijednosti u različitim vremenskim trenucima jesu na pravcu. Za razliku od očekivanja varijanca je konstantna $Var(Y_t) = Var(E_t) = \sigma^2$. Neke veličine ne možemo dobro opisati linearnim trendom, nego je prikladnije koristiti se, primjerice, eksponencijalnim ili nekim drugim trendom.
Još jedna prirodna stvar koja se može pojaviti u vremenskim nizovima jest sezonalnost – uočavamo ju ako se određeno ponašanje ponavlja periodički. Primjerice, kod bilježenja broja noćenja u nekom turističkom području u određenim mjesecima različitih godina možemo uočiti sličnost – npr. veći broj noćenja ljeti ili za vrijeme novogodišnjih praznika. Primijetimo da u tom slučaju moramo dopustiti u modelu da se očekivana vrijednost ovisne varijable razlikuje po, npr., mjesecima. Kao i kod trenda, sezonalnost je važno uočiti u podacima te ju uzeti u obzir prilikom modeliranja i interpretacije. Ako to ne učinimo, to nas može dovesti do donošenja pogrešnih zaključaka o efektima promjene neovisnih varijabli na promjene ovisne varijable, kao i do pojave koreliranosti između greške modela i neovisnih varijabli. Sljedeći primjer iz [9] navodi jedan od načina kako uključiti sezonalnost u izgradnju modela.

Primjer 2.3.2. Prepostavimo da podatke bilježene mjesečno možemo opisati sljedećim modelom:

$$y_t = \beta_0 + \beta_1 x_{t_1} + \beta_2 x_{t_2} + ... + \beta_k x_{t_k} + u_t.$$

Ako uočimo pojavu sezonalnosti na mjesečnoj razini, u model možemo uključiti indikator varijable za svaki mjesec. Vrijednost svake od njih ovisit će o tome kojem mjesecu odgovara trenutak t. Stoga će, primjerice, varijabla dec_t koja se odnosi na prosinac imati sljedeće vrijednosti:

$$dec_t = \begin{cases} 1, t = 12 \\ 0, t = 1, ..., 11. \end{cases}$$

Model zapisujemo na sljedeći način:

$$y_t = \delta_1 feb_t + \delta_2 mar_t + ... + \delta_{11} dec_t + \beta_0 + \beta_1 x_{t_1} + \beta_2 x_{t_2} + ... + \beta_k x_{t_k} + u_t,$$

što znači da će u određenom mjesecu konstantni član β_0 biti uvećan za vrijednost odgovarajućeg parametra tog mjeseca, tj. u prosincu će, primjerice, iznositi $\beta_0 + \delta_{11}$. Primijetimo da je u tom primjeru konstantni član za mjesec siječanj upravo β_0 jer tada sve indikator varijable imaju vrijednost 0.

3 Osnovno o linearnoj regresiji

U ovom poglavlju objasnit ću osnovne koncepte regresijske analize. Pretpostavimo da promatramo ponašanje dviju (ili više) veličina te na temelju zabilježenih podataka želimo ustanoviti postoji li ovisnost jedne veličine o drugima, odnosno htjeli bismo uspostaviti funkcijsku vezu ako je to moguće. Htjeli bismo na temelju tih mjerenja opisati vrijednosti jedne veličine, ovisne varijable, uz pomoć ostalih neovisnih varijabli, do na dodanu grešku. Na taj način, dobivanjem modela koji definira funkcijsku ovisnost, mogli bismo vidjeti utjecaj promjene vrijednosti neke od neovisnih varijabli na promjenu vrijednosti ovisne varijable ili pak procijenjivati vrijednost ovisne varijable za neke nezabilježene vrijednosti neovisnih varijabli.
Promatrane pojave opisujemo slučajnim varijablama, a zabilježena mjerenja tada su realizacije slučajnih varijabli.
Dakle, u najjednostavnijem slučaju imamo sparena mjerenja, tj. parove podataka (x_1, y_1),...,(x_n, y_n) koji su nezavisne realizacije slučajnog vektora (X, Y). Tada pretpostavljamo da postoji linearna funkcija f t.d.

$$y_i = f(x_i) + u_i,$$

odnosno možemo zapisati

$$Y = \beta_0 + \beta_1 X + U \tag{3.1}$$

pri čemu je X neovisna varijabla koju nazivamo regresorom, Y ovisna varijabla koju nazivamo outputom modela, a U slučajna varijabla koju nazivamo greškom modela. Pod greškom modela smatramo sve ostale neizmjerene faktore koji uz X utječu na ovisnu varijablu. Tako opisan model nazivamo jednostavni linearni regresijski model. Linearnost se u tom slučaju odnosi na odnos ovisne varijable Y i neovisne varijable X te znači da jedinična promjena vrijednosti neovisne varijable uvijek ima isti efekt na varijablu Y, do na realizaciju greške, bez obzira na njezinu početnu vrijednost. No, kako je navedeno u [9], najveći problem prilikom postavljanja regresijskog modela jest pitanje smijemo li zaista donositi zaključke o utjecaju X na Y uz tvrdnju da su ostali uvjeti nepromijenjeni. Kako bismo mogli to tvrditi, potrebno je definirati odnos neovisne varijable i greške modela. S obzirom na to da su i X i U slučajne varijable, pretpostavke su sljedeće:

(1) $E[U] = 0$

(2) $E[U|X] = E[U]$.

Pretpostavka (1) govori samo o očekivanju greške, dok (2) govori o tome da očekivana vrijednost greške treba biti jednaka za sve vrijednosti neovisne varijable. Posljedica (1) i (2) jest pretpostavka:

(3) $E[U|X] = 0$
koju zovemo još i stroga egzogenost.

Posljednja pretpostavka (3) govori nam i o tome kako neovisna varijabla utječe na očekivanje ovisne varijable. Uvjetno očekivanje od Y uz dani X možemo, uz pretpostavku (2), zapisati na sljedeći način:

$$E[Y|X] = E[\beta_0] + \beta_1 E[X|X] + E[U].$$

Uz pretpostavku (1) vrijedi

$$E[Y|X] = \beta_0 + \beta_1 X, \tag{3.2}$$

iz čega je jasno da je slučajna varijabla $E[Y|X]$ linearna funkcija od X. Još jedna posljedica tih pretpostavki jest tvrdnja koja govori da bilo koja funkcija transformacija regresora ne korelira s greškom modela (no to ne znači nužno i njihovu nezavisnost).
Neka je h funkcija t.d. $E[|h(X) \cdot U|] < \infty$. Također, uočimo da iz (3.1) te (3.2) vrijedi

$$U = Y - (\beta_0 + \beta_1 X) = Y - E[Y|X].$$

Tada, uz primjenu klasičnih svojstava uvjetnog očekivanja, vrijedi:

$$\begin{aligned} E[h(X) \cdot U] &= E[h(X) \cdot (Y - E[Y|X])] = \\ &= E[h(X) \cdot Y] - E[h(X) \cdot E[Y|X]] = \\ &= E[E[h(X) \cdot Y|X]] - E[E[h(X) \cdot E[Y|X]|X]] = \\ &= E[h(X) \cdot E[Y|X]] - E[h(X)E[Y|X] \cdot E[1|X]] = \\ &= E[h(X)E[Y|X]] - E[h(X)E[Y|X]] = 0. \end{aligned} \tag{3.3}$$

Često u praksi uporaba linearne regresije nije opravdana – upravo zbog toga što je teško očekivati da svi ostali neizmjereni faktori (objedinjeni u grešci modela) koji utječu na ovisnu varijablu ne utječu i na neovisnu, odnosno da X i U nisu korelirani. Ako je moguće još neki od njih mjeriti i uključiti u analizu, tada regresijski model više nije jednostavni, već višestruki ili multivarijatni te očekujemo da ćemo uključivanjem ostalih faktora postići bolji učinak u tumačenju ovisne varijable. Multivarijatni linearni regresijski model možemo zapisati u sljedećem obliku:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + U.$$

U tom slučaju imamo k neovisnih varijabli kojima želimo opisati ovisnu varijablu. Kao i u jednostavnom modelu, ključna pretpostavka potrebna za dobro definiranje regresijskog modela jest nekoreliranost greške s neovisnim varijablama, tj.

$$E[U|X_1, X_2, ..., X_k] = 0.$$

Problem određivanja linearnog regresijskog modela svodi se na procjenu nepoznatih parametara $\beta_1, ..., \beta_k$. U nastavku ćemo obraditi jednu od najpoznatijih metoda za procjenu parametara u regresijskim modelima.

3.1 Metoda najmanjih kvadrata

Radi jednostavnijeg zapisa metoda će biti objašnjena na primjeru jednostavne linearne regresije, no primjenjiva je i za višestruku. Pretpostavimo da imamo izmjerene parove podataka (x_1, y_1),...,(x_n, y_n) na temelju kojih želimo izgraditi regresijski model

$$Y = \beta_0 + \beta_1 X + U,$$

odnosno za parove mjerenja (x_i, y_i) želimo odrediti parametre β_0 i β_1 takve da je

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, ..., n.$$

Za svaku izmjerenu vrijednost x_i možemo odrediti teorijsku vrijednost $y_i^{'} = \beta_0 + \beta_1 x_i$. Uz pretpostavku postojanja greške modela, teorijska i stvarna, izmjerena vrijednost y_i razlikuju se zbog čega točke (x_i, y_i) ne leže na regresijskom pravcu $y = \beta_0 + \beta_1 x$. Ideja metode najmanjih kvadrata jest odrediti parametre β_0 i β_1 tako da razlika $y_i - y_i^{'}$ bude što manja, a to je moguće postići minimizacijom sume kvadrata odstupanja teorijskih od izmjerenih vrijednosti. Dakle, želimo odrediti $\hat{\beta}_0$ i $\hat{\beta}_1$ koji minimiziraju

$$\sum_{i=1}^{n} (y_i - y_i^{'})^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

po β_0 i β_1. Minimizacijski problem svodi se na rješavanje sljedećeg sustava jednadžbi:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \tag{3.4}$$

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{3.5}$$

Prije izvoda rješenja, spomenimo još jedan pristup određivanju regresijskih parametara koji se temelji na pretpostavkama spomenutim u prethodnom poglavlju, a to je metoda momenata. Glavne pretpostavke od kojih ćemo krenuti jesu pretpostavke (1) te posljedica (3.3) da su greška modela i neovisna varijabla nekorelirane.

Ako (3.1) zapišemo kao

$$U = Y - \beta_0 - \beta_1 X$$

te se prisjetimo svojstva $E[U] = 0$, dobijemo sljedeće:

$$E[Y - \beta_0 - \beta_1 X] = 0. \tag{3.6}$$

Nadalje, uz $Cov(U, X) = 0$, odnosno

$$Cov(U, X) = E[U \cdot X] - E[U] \cdot E[X] = E[U \cdot X] = 0$$

možemo računati:

$$E[X(Y - \beta_0 - \beta_1 X)] = 0. \tag{3.7}$$

Za dani skup podataka želimo odrediti $\hat{\beta}_0$ i $\hat{\beta}_1$ t.d. su iz (3.6) i (3.7)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \tag{3.8}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{3.9}$$

Primijetimo da je dobiveni sustav jednadžbi ekvivalentan sustavu (3.4) i (3.5).
Ako s $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ označimo aritmetičku sredinu podataka $y_1, ..., y_n$, tj. procjenu očekivanja od Y, a s $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ aritmetičku sredinu podataka $x_1, ..., x_n$, tj. procjenu očekivanja od X, tada iz (3.8) dobijemo

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}.$$

Nadalje, ako

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.10}$$

uvrstimo u (3.9), slijedi

$$\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) = 0,$$

odakle slijedi da je

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}). \tag{3.11}$$

Primjenom osnovnih operacija nad sumama dobijemo sljedeće:

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

i

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

a uvrštavanjem u (3.11) dobivamo izraz za procjenu $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{3.12}$$

Na taj način dobili smo procjene $\hat{\beta}_0$ i $\hat{\beta}_1$ za parametre β_0 i β_1. Procjene su realizacije procjenitelja koje ćemo za potrebe rada i radi jednostavnosti označavati istom oznakom. Ako primijetimo da

u brojniku stoji izraz za uzoračku kovarijancu X i Y, a u nazivniku za uzoračku varijancu od X, odnosno 3.12 zapišemo na sljedeći način:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)},$$

tada imamo izraz za procjenitelja $\hat{\beta}_1$ parametra β_1. Procjenitelje dobivene na ovaj način u nastavku ćemo prema nazivu metode zvati OLS (*Ordinary Least Squares*) procjeniteljima, a procjene OLS procjenama. Za izračunate procjene možemo definirati i procijenjenu vrijednost mjerenja y_i

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

te rezidual

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

koji označava razliku između stvarne vrijednosti y_i i procijenjene vrijednosti $\hat{y}_i$. Rezidual također možemo promatrati kao procjenu greške modela.

Dakle, na temelju uzorka možemo dobiti procjene $\hat{\beta}_0$ i $\hat{\beta}_1$ koje su potrebne da bismo mogli procijeniti vrijednost ovisne varijable $\hat{y}_i$ za svako mjerenje. Svi parovi vrijednosti neovisne varijable i procijenjene vrijednosti ovisne varijable nalaze se na procjeni regresijskog pravca.

Sada možemo definirati još neke veličine koje će nam u nastavku biti potrebne.

Definicija 3.1.1. Za linearni regresijski model $Y = \beta_0 + \beta_1 X + U$ definiramo

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

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

te

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

gdje je $\hat{y}_i$ procijenjena vrijednost za y_i, a $\hat{u}_i$ rezidual.

SST (*Total Sum of Squares*) predstavlja sumu kvadrata odstupanja izmjerenih vrijednosti od aritmetičke sredine $\bar{y}$, tj. mjeri koliko su raspršene vrijednosti y_i u uzorku, dok SSE (*Explained Sum of Squares*) predstavlja sumu kvadrata odstupanja procijenjenih vrijednosti od aritmetičke sredine $\bar{y}$. Zajedno u zbroju predstavljaju SSR (*Residual Sum of Squares*) koji označava sumu kvadrata reziduala, odnosno vrijedi:

$$SSR = SST + SSE. \tag{3.13}$$

Te mjere korisne su ako želimo znati koliko je dobar naš model, odnosno koliko dobro, u nekom smislu, neovisna varijabla opisuje ovisnu varijablu. Uz pretpostavku da SST nije 0 (odnosno da sve vrijednosti y_i nisu jednake), možemo (3.13) podijeliti s SST te dobiti sljedeće:

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}$$

te definiramo

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

veličinu koju nazivamo koeficijent determinacije. On predstavlja udio objašnjene varijance u ukupnoj varijanci ovisne varijable Y. Jasno je da, ako svi podaci leže na regresijskom pravcu, R^2 iznosi 1 što znači da model savršeno opisuje podatke. U generalnom slučaju linearnog modela s k nezavisnih varijabli tražimo procjenitelje za $\beta_0, \beta_1, ..., \beta_k$ u jednadžbi

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + U \tag{3.14}$$

koji minimiziraju sumu kvadrata reziduala. Analogno definiramo i prediktiranu vrijednost i-tog mjerenja

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}$$

te rezidual

$$\hat{u}_i = y_i - \hat{y}_i,$$

kao i koeficijent determinacije R^2. U slučaju multivarijatnih modela treba biti oprezan s interpretacijom R^2. S obzirom na to da se definira kao suma kvadrata, dodavanjem nove varijable u model može se samo povećati, stoga nije uvijek dovoljan za odluku treba li nova varijabla biti dodana u model. U nastavku ćemo navesti osnovne pretpostavke multivarijatne linearne regresije.

3.1.1 Klasične linearne pretpostavke

Pretpostavka 3.1.1 (Linearnost u parametrima)**.** *Linearni regresijski model*

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + U$$

jest linearan u parametrima, pri čemu su $\beta_0, \beta_1, ..., \beta_k$ parametri koje želimo procijeniti, a U je slučajna greška modela.

Pretpostavka 3.1.2 (Slučajni uzorak)**.** *Podaci kojima se koristimo za procjenu regresijskog modela iz Pretpostavke 3.1.1 dolaze iz nezavisnog slučajnog uzorka s n mjerenja $\{(x_{i1}, ..., x_{ik}, y_i), i = 1, ..., n\}$.*

Pretpostavka 3.1.3 (Multikolinearnost)**.** *U regresijskom modelu nijedna varijabla nije konstanta, niti postoji linearna veza među neovisnim varijablama.*

Multikolinearnost znači da između dviju ili više neovisnih varijabli u multivarijatnom modelu postoji jaka korelacija, odnosno vrijednost neke neovisne varijable ima smisla predviđati koristeći se linearnom vezom s nekima ili svim preostalim neovisnim varijablama, što će poslije uzrokovati veliku varijancu procijenjenog parametra.

Pretpostavka 3.1.4. *Uvjetno očekivanje greške modela na regresore je 0, tj.*

$$E[U|X_1, ..., X_k] = 0.$$

Ako je pretpostavka ispunjena, za neovisne varijable kažemo da su egzogene. U suprotnom, ako su greška i regresori korelirani, tada kažemo da su endogene.

Uz navedene pretpostavke možemo iskazati sljedeći teorem.

Teorem 3.1.1 (Nepristranost OLS procjenitelja)**.** *Uz Pretpostavke 3.1.1 – 3.1.4, OLS procjenitelj $\hat{\beta}_j, j = 1, ..., k$ jest nepristran procjenitelj za parametar β_j, tj.*

$$E[\hat{\beta}_j] = \beta_j.$$

Dokaz. Radi jednostavnijeg izvođenja dokaz ćemo provesti za slučaj jednostavne linearne regresije i procjene parametra β_1.

$$Y = \beta_0 + \beta_1 X + U. \tag{3.15}$$

Pokažimo da vrijedi $E[\hat{\beta}_1] = \beta_1$. Pokazali smo da izraz za procjenitelja za β_1 možemo prikazati na sljedeći način:

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)}. \tag{3.16}$$

Ako u izraz u brojniku uvrstimo izraz (3.15) te primijenimo osnovna svojstva kovarijance, dobijemo sljedeće:

$$Cov(X, Y) = Cov(X, \beta_0 + \beta_1 X + U) = Cov(X, \beta_0) + \beta_1 Cov(X, X) + Cov(X, U) =$$

$$= 0 + \beta_1 Var(X) + 0 = \beta_1 Var(X).$$

S dobivenim izrazom vratimo se u (3.16) te dobijemo sljedeće:

$$\hat{\beta}_1 = \frac{\beta_1 Var(X)}{Var(X)} = \beta_1. \tag{3.17}$$

Sada, ako djelujemo s očekivanjem na izraz (3.17), vrijedi

$$E[\hat{\beta}_1] = E[\beta_1],$$

a s obzirom na to da je s desne strane konstanta, slijedi

$$E[\hat{\beta}_1] = \beta_1.$$

□

Pretpostavka 3.1.5 (Homoskedastičnost). $Var(U|X_1, ..., X_k) = \sigma^2$.

Dakle, varijanca greške uz dane neovisne varijable uvijek je ista. U slučaju da se to ne ispuni, tada za model kažemo da je heteroskedastičan. Pretpostavke 3.1.1 – 3.1.5 zajedno čine Gauss-Markovljeve pretpostavke. Ako s $\boldsymbol{X}$ označimo vektor nezavisnih varijabli $(X_1, ..., X_k)$, tada Pretpostavke 3.1.1 – 3.1.4 možemo zapisati na sljedeći način:

$$E[Y|\boldsymbol{X}] = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \tag{3.18}$$

a Pretpostavku 3.1.5:

$$Var(Y|\boldsymbol{X}) = \sigma^2. \tag{3.19}$$

Izraz (3.18) govori nam da je očekivanje od Y uz dani $\boldsymbol{X}$ linearno te da ovisi o $X_1, ..., X_k$, a (3.19) da varijanca od Y ne ovisi o $X_1, ..., X_k$.

Uz zadovoljene navedene pretpostavke možemo izračunati i varijancu OLS procjenitelja, o čemu govori sljedeći teorem.

Teorem 3.1.2. *Uz Pretpostavke 3.1.1 – 3.1.5,*

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, j = 1, 2, ..., k$$

gdje je $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j^2)$, *a* R_j^2 *jest koeficijent determinacije prediktora* X_j *i ostalih neovisnih varijabli.*

Iz Teorema 3.1.1 znamo da OLS metoda daje nepristrane procjenitelje, no dalje nas zanima jesu li to najbolji procjenitelji koje možemo dobiti ili postoje u nekom smislu bolji. Sljedeći teorem govori upravo o tome.

Teorem 3.1.3 (Gauss-Markovljev teorem). *Uz Pretpostavke 3.1.1 – 3.1.5, procjenitelji* $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ *najbolji su linearni nepristrani procjenitelji (Best Linear Unbiased Estimator – BLUE) za* $\beta_0, \beta_1, ..., \beta_k$.

Pritom nepristranost znači $E[\hat{\beta}_j] = \beta_j$, a linearnost znači da se procjena može prikazati kao linearna funkcija vrijednosti ovisne varijable, odnosno,

$$\hat{\beta}_j = \sum_{i=1}^{n} \omega_{ij} y_i$$

gdje je ω_{ij} težinska funkcija uzoračkih vrijednosti neovisne varijable, tj.

$$\omega_{ij} = \frac{x_{ij} - \bar{x}_j}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}, j = 1, ..., k.$$

Najbolji procjenitelj (BLUE) znači procjenitelj s najmanjom varijancom. Gauss-Markovljev teorem nam dakle govori da, ako vrijede Gauss-Markovljeve pretpostavke, OLS procjenitelj najbolji je mogući procjenitelj te nema potrebe da tražimo boljeg.

Iduća pretpostavka nije nužna za dobivanje najboljeg procjenitelja, no važna je u provođenju statističkih testova.

Pretpostavka 3.1.6. *Greška modela U jest nezavisna s varijablama $X_1, ..., X_k$ te je normalno distribuirana s očekivanjem 0 i varijancom σ^2, tj. $U \sim \mathcal{N}(0, \sigma^2)$.*

Pretpostavke 3.1.1 – 3.1.6 zajedno čine pretpostavke klasičnog linearnog modela, pa stoga i model koji ih zadovoljava zovemo klasični linearni model.

Teorem 3.1.4. *Uz klasične pretpostavke linearnog modela OLS procjenitelj $\hat{\beta}_j$ ima normalnu distribuciju s očekivanjem β_j i varijancom definiranom u teoremu 3.1.2, tj.*

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, Var(\hat{\beta}_j)).$$

Sva navedena svojstva procjenitelja odnose se na konačne uzorke. U slučaju kad je dimenzija uzorka dovoljno velika, možemo primjenjivati i asimptotska svojstva ili svojstva velikih uzoraka. Dosad smo nepristranost navodili kao vrlo važno svojstvo, no ono ne može uvijek biti zadovoljeno. Često se događa da procjenitelj nije nepristran, ali je konzistentan, što je po mnogima važniji uvjet. Prisjetimo se, ako je $\hat{\beta}_j$ procjenitelj za β_j, nepristranost nam govori da distribucija procjenitelja $\hat{\beta}_j$ ima očekivanje jednako β_j, dok za konzistentnost vrijedi da kad veličina uzorka raste, niz procjenitelja za parametar β_j po vjerojatnosti konvergira prema stvarnoj vrijednosti tog parametra. Oba svojstva zahtijevaju zadovoljavanje Pretpostavki 3.1.1 – 3.1.4.

Teorem 3.1.5. *Uz Pretpostavke 3.1.1 – 3.1.4, OLS procjenitelj $\hat{\beta}_j, j = 1, ..., k$ je konzistentan procjenitelj za parametar β_j, tj.*

$$\lim_{n\to\infty} P(|\hat{\beta}_j - \beta_j| > \varepsilon) = 0, \forall \varepsilon > 0. \tag{3.20}$$

Dokaz. Vidi [9], str. 169, Theorem 5.1. □

Izraz (3.20) može biti zapisan i s oznakom *plim* za limes u konvergenciji po vjerojatnosti na način da je

$$\underset{n\to\infty}{plim}\, \hat{\beta}_j = \beta_j.$$

Taj teorem vrijedi i ako postavimo slabiju verziju Pretpostavke 3.1.4, odnosno dovoljno je pretpostaviti da je svaka neovisna varijabla nekorelirana s greškom, tj.

$$E[U|X_i] = 0, i = 1, ..., k.$$

Stoga, koreliranost greške i bilo koje regresorske varijable uzrokuje, osim pristranosti, i nekonzistentnost procjenitelja. Kako bismo mogli donositi valjane statističke zaključke, za konstruiranje pouzdanih intervala i provođenje statističkih testova potrebno je znati i distribuciju procjenitelja. U slučaju kad Pretpostavka 3.1.6 nije zadovoljena, uvjeti za teorem 3.1.4 nisu ispunjeni. No idući teorem govori nam o asimptotskoj distribuciji procjenitelja za koju nije nužna normalna distribucija greške modela.

Teorem 3.1.6. *Uz Gauss-Markovljeve pretpostavke 3.1.1 – 3.1.5 vrijedi sljedeće:*

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} \mathcal{N}(0, \sigma^2/a_j^2)$,
gdje je $\sigma^2/a_j^2 > 0$ *asimptotska varijanca od* $\sqrt{n}(\hat{\beta}_j - \beta_j)$, $a_j^2 = plim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n} \hat{r}_{ij}^2)$, *gdje su* $\hat{r}_{ij}$ *reziduali koji nastaju regresijom varijable* x_j *na ostale neovisne varijable. Kažemo da* $\hat{\beta}_j$ *ima asimptotski normalnu distribuciju.*

(ii) $\hat{\sigma}^2$ *je konzistentan procjenitelj za* $\sigma^2 = Var(U)$

(iii) za svaki j,

$$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \overset{a}{\sim} \mathcal{N}(0,1)$$

i

$$(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \overset{a}{\sim} \mathcal{N}(0,1),$$

gdje je $sd(\hat{\beta}_j)$ *standardna devijacija, a* $se(\hat{\beta}_j)$ *standardna greška OLS procjenitelja regresijskog parametra* β_j.

4 Linearna regresija s vremenskim nizovima

U ovom poglavlju bavit ću se linearnom regresijom vremenskih nizova koja se temelji na teorijskim konceptima obrađenim u prethodna dva poglavlja. Ono što je istovremeno prednost i nedostatak u takvom skupu podataka jest činjenica da trenutačna vrijednost promatrane veličine najčešće ima (vrlo) jaku vezu s vrijednostima u prošlosti. Prednost je u tome što nam olakšava predviđanje budućih vrijednosti, a nedostatak jer postojanje zavisnosti može stvarati problem pri donošenju statističkih zaključaka. Ideja samog modeliranja ista je, opisati ponašanje jedne (ovisne) varijable pomoću ostalih (neovisnih) varijabli, uz pretpostavku da među njima postoji linearna funkcijska veza. Dodatno je moguće kao regresore uključiti i prošle vrijednosti ovisne varijable. Glavna je razlika u tome što nemamo više jednostavni slučajni uzorak, odnosno različita mjerenja više jedinki iz populacije, već mjerenja nastaju za istu pojavu u različitim vremenskim trenucima, pri čemu je važan njihov redoslijed. Ono što možemo dobiti regresijskim modeliranjem vremenskih nizova jest procjena efekta uzrokovanog promjenom neke varijable tijekom vremena, kao i predikcija budućih vrijednosti vremenskog niza.

4.1 Primjeri regresijskih modela vremenskog niza

4.1.1 Statički model

Definicija 4.1.1. Neka su $y_1, ..., y_n$ i $x_1, ..., x_n, n \in T \subseteq \mathbb{N}$, dva vremenska niza, tj. realizacije slučajnih procesa $\{X_t, t \in T\}$ i $\{Y_t, t \in T\}$, takvi da su y_t i x_t realizacije u istom vremenskom trenutku, $\forall t \in \{1, ...n\}$. Tada model

$$Y_t = \beta_0 + \beta_1 X_t + U_t, \quad t = 1, ..., n$$

zovemo statički model.

Naziv statički dolazi od činjenice da opisuje zavisnost dviju varijabli samo u danom vremenskom trenutku. Dakle, očekujemo da bi promjena varijable X_t u trenutku $t \in T$ imala trenutni učinak na varijablu Y_t, ali ne i na varijable $Y_{t+1}, Y_{t+2}, ...$ Točnije, uz sve ostalo nepromijenjeno, tj. $\Delta U_t = 0$, vrijedi

$$\Delta Y_t = \beta_1 \Delta X_t.$$

Primjer 4.1.1. Primjer statičkog modela jest statička Phillipsova krivulja koja u ekonomiji definira odnos stope inflacije i stope nezaposlenosti.

$$i_t = \beta_0 + \beta_1 n_t + u_t,$$

pri čemu je i_t oznaka za stopu inflacije u trenutku t, a n_t oznaka za stopu nezaposlenosti u trenutku t.

Također, u statičkom modelu možemo imati i više od jednog regresora.

Definicija 4.1.2. Neka su $y_1, ..., y_n$ $x_{1,j}, ..., x_{n,j}, n \in T \subseteq \mathbb{N}, j = 1, ..., k$ realizacije slučajnih procesa $\{Y_t, t \in T\}$, $\{X_{t,j}, t \in T\}, j = 1, ..., k$. Tada linearni regresijski model

$$Y_t = \beta_0 + \beta_1 X_{t,1} + ... + \beta_k X_{t,k} + U_t, \quad t = 1, ..., n$$

zovemo statički model s regresorima $X_1, ..., X_k$.

U notaciji $X_{t,j}$ t označava vremenski trenutak, a j jednu od k neovisnih varijabli.

4.1.2 Finite distributed lag model

Modele u kojima se efekt promjene regresora ne odražava na ovisnu varijablu samo u danom trenutku već i u budućim trenucima zovemo dinamičkim modelima, a jedan od njih jest Finite Distributed Lag (FDL) model.

Definicija 4.1.3. Neka su $y_1, ..., y_n$ i $x_1, ..., x_n, n \in T \subseteq \mathbb{N}$ dva vremenska niza, tj. realizacije slučajnih procesa $\{Y_t, t \in T\}$, i $\{X_t, t \in T\}$. Tada linearni regresijski model

$$Y_t = \alpha + \delta_0 X_t + \delta_1 X_{t-1} + ... + \delta_q X_{t-q} + U_t, \quad t = 1, ..., n, \quad q \in \mathbb{N}$$

zovemo finite distributed lag model reda q.

U FDL modelu na trenutačnu vrijednost ovisne varijable utječu sadašnje, ali i prošle vrijednosti regresora. Isto tako, sadašnja vrijednost regresora utjecat će i na buduće vrijednosti ovisne varijable.

Primjer 4.1.2. Neka je

$$Y_t = \alpha + \delta_0 X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2} + U_t.$$

Pretpostavimo da do trenutka t X_u ima konstantnu vrijednost $c \in \mathbb{R}$, da se u trenutku t dogodi njezino jedinično povećanje te da se nakon toga opet vrati na prethodnu, konstantnu vrijednost c. Dakle,

$$...X_{t-2} = c, X_{t-1} = c, X_t = c + 1, X_{t+1} = c, ...$$

Nadalje, pretpostavimo da je greška U_t konstantna $\forall t \in T$, tj. $\Delta U_t = 0$. Pogledajmo sada vrijednosti varijable Y u trenutcima $t-1, t, t+1, t+2$ i $t+3$:

$$Y_{t-1} = \alpha + \delta_0 c + \delta_1 c + \delta_2 c + U_t,$$

$$Y_t = \alpha + \delta_0 (c + 1) + \delta_1 c + \delta_2 c + U_t,$$

$$Y_{t+1} = \alpha + \delta_0 c + \delta_1 (c + 1) + \delta_2 c + U_t,$$

$$Y_{t+2} = \alpha + \delta_0 c + \delta_1 c + \delta_2 (c + 1) + U_t,$$

$$Y_{t+3} = \alpha + \delta_0 c + \delta_1 c + \delta_2 c + U_t,$$
$$Y_t - Y_{t-1} = \delta_0,$$
$$Y_{t+1} - Y_{t-1} = \delta_1,$$
$$Y_{t+2} - Y_{t-1} = \delta_2,$$
$$Y_{t+3} - Y_{t-1} = 0.$$

Koeficijent δ_0 daje informaciju o intenzitetu promjene ovisne varijable uzrokovane promjenom regresora te ga zovemo trenutni multiplikator učinka, dok δ_1 i δ_2 opisuju promjene budućih vrijednosti ovisne varijable. S obzirom na to da u Primjeru 4.1.2 u danom modelu postoje dvije prošle vrijednosti regresora, trenutna promjena bit će vidljiva u najviše dva iduća trenutka. U modelu

$$Y_t = \alpha + \delta_0 X_t + \delta_1 X_{t-1} + ... + \delta_k X_{t-k} + U_t$$

jedinična promjena regresora u trenutku t reflektirat će se na vrijednost ovisne varijable u idućih k trenutaka. Niz koeficijenata $\delta_0, ..., \delta_k$ opisuju dinamički efekt trenutnog jediničnog povećanja regresora. Osim trenutne promjene, zanima nas i kako bi trajna promjena regresora utjecala na trenutnu te buduće vrijednosti ovisne varijable.

Primjer 4.1.3. Kao u prethodnom primjeru, neka je

$$Y_t = \alpha + \delta_0 X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2} + U_t.$$

Neka je vrijednost regresora konstantna do nekog trenutka $t \in T$, a zatim se poveća za jedan te se u budućim trenucima ne mijenja, tj.

$$..., X_{t-2} = c, X_{t-1} = c, X_t = c + 1, X_{t+1} = c + 1, X_{t+2} = c + 1, ...$$

Nadalje, pretpostavimo da je greška konstantna, tj. $\Delta U_t = 0, \forall t \in T$. Tada vrijedi:

$$Y_{t-1} = \alpha + \delta_0 c + \delta_1 c + \delta_2 c + U_t,$$
$$Y_t = \alpha + \delta_0 (c + 1) + \delta_1 c + \delta_2 c + U_t,$$
$$Y_{t+1} = \alpha + \delta_0 (c + 1) + \delta_1 (c + 1) + \delta_2 c + U_t,$$
$$Y_{t+2} = \alpha + \delta_0 (c + 1) + \delta_1 (c + 1) + \delta_2 (c + 1) + U_t,$$
$$Y_{t+3} = \alpha + \delta_0 (c + 1) + \delta_1 (c + 1) + \delta_2 (c + 1) + U_t,$$
$$Y_t - Y_{t-1} = \delta_0,$$
$$Y_{t+1} - Y_{t-1} = \delta_0 + \delta_1,$$
$$Y_{t+2} - Y_{t-1} = \delta_0 + \delta_1 + \delta_2,$$
$$Y_{t+3} - Y_{t-1} = \delta_0 + \delta_1 + \delta_2.$$

Koeficijent δ_0 opisuje trenutnu promjenu ovisne varijable. Promjena u prvom idućem trenutku iznosit će $\delta_0 + \delta_1$, a nakon 2 trenutka $\delta_0 + \delta_1 + \delta_2$ te se više neće mijenjati.

Možemo reći da je suma koeficijenata modela $\delta_0+\delta_1+...+\delta_q$ tada dugoročna promjena ovisne varijable uzrokovana jediničnim povećanjem u trenutku t, a zbog toga ju zovemo dugoročni multiplikator učinka. Također, uvijek možemo promatrati sumu $\delta_0 + \delta_1 + ... + \delta_h, h \leq q$ koja predstavlja promjenu ovisne varijable h trenutaka nakon jediničnog povećanja regresora X. Ako vrijedi $\delta_1 = \delta_2 = ... = \delta_q = 0$, FDL model zapravo promatramo kao statički model u kojem varijabla Y ovisi o varijabli X u danom trenutku.

4.2 Svojstva OLS procjenitelja

Kako bismo mogli prediktirati buduće vrijednosti ovisne varijable, potrebno je definirati model – što se i u ovom slučaju svodi na problem procjene parametara. Metoda najmanjih kvadrata koja je navedena u prethodnom poglavlju i s podacima iz vremenskih nizova daje najbolje procjenitelje za parametre, no u uvjetima pretpostavki koje navodimo u nastavku.

Pretpostavka 4.2.1 (Linearnost u parametrima)**.**
(k+1)-dimenzionalni stohastički proces $\{(X_{t1}, ..., X_{tk}, Y_t) : t = 1, ..., n\}$ *zadovoljava linearni model*

$$Y_t = \beta_0 + \beta_1 X_{t1} + ... + \beta_k X_{tk} + U_t,$$

pri čemu je $\{U_t : t = 1, ..., n\}$ *proces grešaka modela, a n broj mjerenja (vremenskih trenutaka).*

Uočimo da je pretpostavka analogna Pretpostavci 3.1.1, uz razliku što ovdje regresore čine vremenski nizovi. Nadalje, označimo s $X = (X_{t1}, ..., X_{tk})$ vektor neovisnih varijabli u trenutku t. S $\boldsymbol{X}$ tada označavamo slučajnu matricu svih neovisnih varijabli u svim vremenskim trenucima.

Pretpostavka 4.2.2 (Multikolinearnost)**.** *U uzorku, kao ni u stohastičkom procesu, nijedna neovisna varijabla nije konstanta niti je linearna kombinacija svih ostalih neovisnih varijabli.*

Pretpostavka 4.2.3. *Za svaki t, očekivana vrijednost greške* U_t *uvjetno na neovisne varijable u svim vremenskim trenucima jest* 0*, tj.*

$$E[U_t|\boldsymbol{X}] = 0, \quad t = 1, ..., n.$$

Pretpostavka nam govori da je greška U_t nekorelirana sa svakom neovisnom varijablom u svakom vremenskom trenutku ([9], str. 350). Ako ju promatramo samo za dani trenutak t,

$$E[U_t|X_{t1}, ..., X_{tk}] = E[U_t|\boldsymbol{X}_t] = 0, \tag{4.1}$$

jasno je da je to direktna posljedica Pretpostavke 3.1.4. Stoga i ovdje kažemo, ako je (4.1) zadovoljeno, da su varijable $X_{tj}, j = 1, ..., k$ istovremeno egzogene, odnosno da su greška U_t i neovisne varijable trenutno nekorelirane, tj.

$$Corr(X_{tj}, U_t) = 0, \quad \forall j.$$

Osim toga, Pretpostavka 4.2.3 zahtijeva i više od istovremene egzogenosti te govori da U_t mora biti nekorelirana s X_{sj}, za $s \neq t$. U tom slučaju kažemo da su neovisne varijable strogo egzogene. Posljedice te pretpostavke možemo prikazati i na sljedeći način:

$$E[U_t] = E[E[U_t|\boldsymbol{X}]] = E[0] = 0, \tag{4.2}$$

$$E[X_{tj}U_t] = E[E[X_{tj}U_t|X_{tj}]] = E[X_{tj}E[U_t|X_{tj}]] = 0. \tag{4.3}$$

U podacima koji dolaze iz jednostavnog slučajnog uzorka nismo imali potrebe razmišljati o tom problemu, no kako u vremenskim nizovima nemamo slučajnog uzorka, moramo eksplicitno pretpostaviti da greška nije korelirana s neovisnim varijablama ni u jednom trenutku. Također, u [9] je naveden sljedeći primjer:

Primjer 4.2.1. Imamo jednostavni statički regresijski model

$$Y_t = \beta_0 + \beta_1 Z_t + U_t.$$

Pretpostavka 4.2.3 zahtijeva ne samo koreliranost U_t i Z_t, već i U_t sa svim prošlim i budućim vrijednostima od Z_t.

To znači da stroga egzogenost isključuje mogućnost da promjena greške danas uzrokuje promjene varijable Z_t u budućnosti, što je u [9] pojašnjeno još jednim primjerom:

Primjer 4.2.2. Imamo jednostavni statički model

$$mrdrte_t = \beta_0 + \beta_1 polpc_t + u_t,$$

gdje je *mrdrte* godišnja stopa ubojstava u nekom gradu, a *polpc* broj policajaca po stanovniku. Razumno je očekivati da je greška u_t nekorelirana s $polpc_t$, čak i s njezinim prošlim vrijednostima. No promotrimo sljedeću situaciju, ako grad odluči povećati broj policajaca zbog prošlih vrijednosti stope ubojstava *mrdrte*, to znači da $polpc_{t+1}$ može biti korelirano s u_t iz jednostavnog razloga - veći u_t vodi ka većem $mrmdrte_t$. U tom slučaju dolazi do kršenja Pretpostavke 4.2.3.

Obično se ne brinemo o tome utječe li greška na prošle vrijednosti, no utjecaj na buduće vrijednosti ovisne varijable itekako može biti problem. Dakle, iako Pretpostavka 4.2.3 može ponekad biti nerealna, moramo ju pokušati zadovoljiti. Također, još jedna direktna posljedica Pretpostavke 4.2.3 jest da je očekivanje ovisne varijable uvjetno na $\boldsymbol{X}$ linearna funkcija regresora u danom trenutku t, tj. uz uvjet $E[Y_t|\boldsymbol{X}] = 0$ vrijedi

$$E[Y_t|\boldsymbol{X}] = \beta_0 + \beta_1 X_{t2} + ... + \beta_k X_{tk},$$

pa regresijski model možemo zapisati i kao

$$Y_t = E[Y_t|\boldsymbol{X}] + U_t,$$

a grešku modela kao

$$U_t = Y_t - E[Y_t|\boldsymbol{X}].$$

Navedene pretpostavke vode do teorema analognog Teoremu 3.1.1.

Teorem 4.2.1 (Nepristranost OLS procjenitelja). *Uz Pretpostavke 4.2.1 – 4.2.3, OLS procjenitelji $\hat{\beta}_0, ..., \hat{\beta}_k$ nepristrani su procjenitelji za parametre $\beta_0, ..., \beta_k$, tj.*

$$E[\hat{\beta}_j] = \beta_j, j = 0, 1, ..., k.$$

Pretpostavka 4.2.4 (Homoskedastičnost). *Varijanca od U_t, uvjetno na $\boldsymbol{X}$, konstantna je za svaki trenutak t, tj.*

$$Var(U_t|\boldsymbol{X}) = Var(U_t) = \sigma^2.$$

U slučaju kada ta pretpostavka nije zadovoljena, kažemo da je model heteroskedastičan.

Pretpostavka 4.2.5. *Greške u različitim vremenskim trenucima, uvjetno na $\boldsymbol{X}$, nekorelirane su, tj.*

$$Corr(U_t, U_s|\boldsymbol{X}) = 0, \quad \forall t, s, \quad t \neq s.$$

Ako ta pretpostavka nije zadovoljena, kažemo da su greške serijski korelirane (autokorelirane). Pretpostavke 4.2.1 – 4.2.5 zajedno čine Gauss-Markovljeve pretpostavke u primjeni regresijske analize s podacima iz vremenskih nizova.

Teorem 4.2.2 (Gauss-Markovljev teorem). *Uz Pretpostavke 4.2.1 – 4.2.5, OLS procjenitelji $\hat{\beta}_0, ..., \hat{\beta}_k$ su najbolji procjenitelji za parametre $\beta_0, ..., \beta_k$ uvjetno na $\boldsymbol{X}$.*

Zaključujemo da se koeficijenti u modelu s vremenskim nizovima mogu procijeniti OLS metodom, no multikolinearnost definitivno može predstavljati velik problem, pogotovo u slučaju kad su regresori varijable iz istog stohastičkog procesa, a različitih vremenskih trenutaka.

Pretpostavka 4.2.6. *Greške U_t nezavisne su s $\boldsymbol{X}$, međusobno nezavisne te normalno distribuirane, $U_t \sim \mathcal{N}(0, \sigma^2)$.*

Pretpostavka implicira Pretpostavke 4.2.3, 4.2.4 i 4.2.5, no jača je zbog nezavisnosti i normalnosti. Nije nužna za zadovoljavanje Gauss-Markovljeva teorema, odnosno i bez nje možemo tvrditi da su OLS procjenitelji najbolji procjenitelji s najmanjom varijancom (BLUE), no korisna je zbog jednostavnijeg konstruiranja intervala pouzdanosti. Pretpostavke 4.2.1 – 4.2.6 zajedno čine klasične linearne pretpostavke za modeliranje s vremenskim nizovima.

Teorem 4.2.3 ([9], Theorem 10.5). *Uz Pretpostavke 4.2.1 – 4.2.6 OLS procjenitelji normalno su distribuirani uvjetno na $\boldsymbol{X}$. U uvjetima nulte hipoteze svaka t statistika ima t distribuciju, a svaka F statistika F distribuciju. Intervali pouzdanosti konstruiraju se na isti način.*

Uz te pretpostavke procjene i statističko zaključivanje s podacima iz vremenskih nizova možemo izvoditi jednako kao i kod podataka iz slučajnog uzorka. Važno je znati da je svako statističko zaključivanje dobro kao i pretpostavke na kojima se temelji [9], a one su u ovom slučaju mnogo restriktivnije (primjerice, stroga egzogenost i izostanak serijske korelacije često u praksi mogu biti nerealni zahtjevi što će biti vidljivo i u praktičnom dijelu rada).

Kao što je slučaj i u analizi podataka iz jednostavnog slučajnog uzorka, nekad ne možemo pokazati da je procjenitelj nepristran, niti možemo odrediti distribuciju procjenitelja kako bismo mogli izvoditi statističke zaključke. Također, često se događa i da greška modela nije normalno distribuirana. Kako bismo se u takvim situacijama i dalje mogli pouzdati u OLS procjenitelje, moramo proučiti njihova asimptotska svojstva. S obzirom na to da se asimptotska svojstva temelje na centralnom graničnom teoremu i zakonu velikih brojeva, ovdje ćemo koristiti varijante tih teorema u kojima je klasična pretpostavka o nezavisnosti zamijenjena pretpostavkom o stacionarnosti i slaboj zavisnosti procesa (vidi [6] i [7]). U nastavku navodimo pretpostavke uz koje vrijede asimptotska svojstva.

4.3 Asimptotska svojstva OLS procjenitelja

Pretpostavka 4.3.1 (Linearnost i slaba zavisnost)**.** *Stohastički proces* $\{(X_{t1}, ..., X_{tk}, Y_t) : t = 1, ..., n\}$ *koji zadovoljava linearni model*

$$Y_t = \beta_0 + \beta_1 X_{t1} + ... + \beta_k X_{tk} + U_t \tag{4.4}$$

jest stacionaran i pripada klasi slabo zavisnih procesa.

Ova pretpostavka analogna je dosadašnjim klasičnim pretpostavkama o linearnosti, uz dodatak stacionarnosti i slabe zavisnosti koje nam omogućuju da primjenjujemo centralni granični teorem i zakon velikih brojeva (vidi [6] i [7]). U ovom zapisu modela (4.4) parametri $\beta_0, ..., \beta_k$ su ti koje želimo procjeniti OLS metodom.

Pretpostavka 4.3.2. *U uzorku, kao ni u stohastičkom procesu, nijedna neovisna varijabla nije konstanta niti je linearna kombinacija svih ostalih neovisnih varijabli.*

Pretpostavka o odsutnosti kolinearnosti u potpunosti je jednaka onoj u klasičnim linearnim pretpostavkama.

Pretpostavka 4.3.3. *Neovisne varijable* $\boldsymbol{X_t} = (X_{t1}, ..., X_{tk})$ *istovremeno su egzogene, odnosno*

$$E[U_t|\boldsymbol{X}_t] = 0, \quad \forall t = 1, ..., n.$$

Pretpostavka je slabija u odnosu na Pretpostavku 4.2.3 koja govori o nekoreliranosti greške s neovisnim varijablama u svim vremenskim trenucima – ovdje je dovoljna u istom vremenskom trenutku. Primijetimo da ako stacionaran proces zadovoljava pretpostavku, dovoljno je pokazati nekoreliranost u jednom vremenskom trenutku – zbog stacionarnosti znamo da vrijedi i za sve ostale.

Teorem 4.3.1. *Uz Pretpostavke 4.3.1 – 4.3.3, OLS procjenitelj $\hat{\beta}_j$ konzistentan je procjenitelj za parametar β_j, odnosno*

$$\underset{n\to\infty}{plim}\,\hat{\beta}_j = \beta_j, j = 0, 1, ..., k.$$

Teorem o konzistentnosti procjenitelja moguće je izvesti i za slabije pretpostavke. Dovoljno je da greška modela U_t ima očekivanje 0 te da je nekorelirana sa svakim X_{tj}, odnosno $E[U_t] = 0, Cov(X_{tj}, U_t) = 0, j = 1, ..., k$.

Pretpostavka 4.3.4. *Greška je modela u danom trenutku t homoskedastična, tj.*

$$Var(U_t|\boldsymbol{X}_t) = \sigma^2,$$

pri čemu je $\boldsymbol{X} = (X_{t1}, \ldots, X_{tk})$.

Pretpostavka 4.3.5. *Za svaki $t \neq s, E[U_t U_s|\boldsymbol{X}_t, \boldsymbol{X}_s] = 0$.*

Pretpostavku 4.3.5 možemo promatrati i na drugi način – možemo pretpostaviti da su greške U_t i U_s nekorelirane za svaki $t \neq s$. Pretpostavke 4.3.4 i 4.3.5 bitne su nam kako bismo mogli donositi statističke zaključke. Primijetimo kako su obje pretpostavke slabije verzije klasičnih linearnih pretpostavki. Dakle, homoskedastičnost i nekoreliranost greške više nisu nužne uvjetno na neovisne varijable u svim vremenskim trenucima, već samo u istom trenutku. Serijsku koreliranost često je teško izbjeći, a detaljnije ćemo se njome baviti poslije. Ovih pet pretpostavki nužan su uvjet sljedećeg teorema.

Teorem 4.3.2 ([9], Theorem 10.1)**.** *Uz Pretpostavke 4.3.1 – 4.3.5 OLS procjenitelji asimptotski su normalno distribuirani.*

Čak i ako model ne zadovoljava klasične linearne pretpostavke, ako zadovolji uvjete teorema, možemo tvrditi da su procjenitelji konzistentni te provoditi statističko zaključivanje.

Idućim primjerom pokazat ćemo jedan slučaj u kojem se ne možemo pouzdati u klasične pretpostavke, već se moramo koristiti asimptotskim.

Primjer 4.3.1. Pretpostavimo da imamo model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t, \tag{4.5}$$

takav da vrijedi

$$E[U_t|Y_{t-1}, Y_{t-2}, ...] = 0. \tag{4.6}$$

Koristeći se izrazima (4.5) i (4.6) dobijemo sljedeće:

$$E[Y_t|Y_{t-1}, Y_{t-2}, ...] = E[Y_t|Y_{t-1}] = \beta_0 + \beta_1 Y_{t-1}. \tag{4.7}$$

Iz dobivenog izraza (4.7) možemo zaključiti da na očekivanu vrijednost varijable Y_t utječe samo prošla vrijednost na koraku 1, Y_{t-1}.

S obzirom na to da se vektor ovisnih varijabli $\boldsymbol{X}_t$ sastoji samo od varijable Y_{t-1} te uz pretpostavku modela (4.6) vrijedi sljedeće:

$$E[U_t|\boldsymbol{X}_t] = E[U_t|Y_{t-1}] = 0,$$

što zadovoljava Pretpostavku 4.3.3 o istovremenoj egzogenosti. Da bismo zadovoljili i klasičnu Pretpostavku 4.2.3 moralo bi vrijediti da je greška modela U_t u trenutku t nekorelirana sa svim regresorima u svim vremenskim trenucima. Iz (4.6) jasno je da je nekorelirana sa svima do trenutka $(t-1)$, no da bi zadovoljila strogu egzogenost, mora biti nekorelirana i s Y_t, odnosno moralo bi vrijediti $Cov(U_t, Y_t) = 0$, a to u ovom modelu ne može biti istinito, što možemo i pokazati primjenom osnovnih svojstava kovarijance:

$$Cov(U_t, Y_t) = Cov(U_t, \beta_0 + \beta_1 Y_{t-1} + U_t) =$$

$$= Cov(U_t, \beta_0) + \beta_1 Cov(U_t, Y_{t-1}) + Cov(U_t, U_t) = Cov(U_t, U_t) =$$

$$= Var(U_t) = \sigma^2 \neq 0.$$

Nadalje, možemo pokazati i kako ovaj model zadovoljava i Pretpostavku 4.3.5 o izostanku serijske korelacije. Prisjetimo se, ona zahtijeva sljedeće:

$$E[U_t U_s|\boldsymbol{X}_t, \boldsymbol{X}_s] = 0, \quad \forall s \neq t.$$

U tom slučaju, dovoljno je pokazati da je $E[U_t U_s|Y_{t-1}, Y_{s-1}] = 0, \forall t \neq s$.
Pretpostavimo $s < t$.
Kako je $U_s = Y_s - \beta_0 - \beta_1 Y_{s-1}$ funkcija slučajnih varijabli koje u vremenu prethode Y_t, uz (4.6) vrijedi

$$E[U_t|U_s, Y_{t-1}, Y_{s-1}] = 0,$$

pa uz svojstvo očekivanja[1]

$$E[U_t U_s|U_s, Y_{t-1}, Y_{s-1}] = 0.$$

Ako dobiveni izraz iskoristimo zajedno sa svojsvom dvostrukog očekivanja[2], dobijemo sljedeće:

$$E[U_t U_s|Y_{t-1}, Y_{s-1}] = E[E[U_t U_s|U_s, Y_{t-1}, Y_{s-1}]|Y_{t-1}, Y_{s-1}] = 0.$$

Dakle, greške su serijski nekorelirane.

Navedeni primjer pokazuje nam da modeli koji za neovisnu varijablu imaju prošlu vrijednost ovisne varijable na nekom koraku, uz uvjet (4.7), ne mogu zadovoljiti strogu egzogenost, ali greške modela moraju biti serijski nekorelirane. Također, možemo reći da je pretpostavka da su greške modela serijski nekorelirane jednaka pretpostavci da na očekivanje $E[Y_t|Y_{t-1}, Y_{t-2}, ...]$

[1] $E[a(X)b(Y)|X] = a(X)E[b(Y)|X]$
[2] $E[Y|X] = E[E[Y|X, Z]|X]$

utječe samo prošla vrijednost varijable Y na koraku 1, Y_{t-1}.
Zaključak o serijskoj nekoreliranosti ne odnosi se samo na specifične modele kao u tom primjeru, već se u određenim uvjetima može i generalizirati.
Primjerice, u jednostavnom statičkom regresijskom modelu

$$Y_t = \beta_0 + \beta_1 Z_t + U_t,$$

uz pretpostavku

$$E[U_t|Z_t, Y_{t-1}, Z_{t-1}, ...] = 0, \tag{4.8}$$

možemo tvrditi da je Pretpostavka 4.3.5 o serijskoj nekoreliranosti greške modela zadovoljena, što se može pokazati na isti način kao u prethodnom primjeru. Također, na isti način pokazuje se i da je zadovoljena pretpostavka o istovremenoj egzogenosti, tj. $E[U_t|Z_t] = 0$.
Nadalje, i za FDL model

$$Y_t = \beta_0 + \beta_1 Z_t + \beta_2 Z_{t-1} + \beta_3 Z_{t-2} + U_t$$

želimo pretpostaviti

$$E[Y_t|Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, ...] = E[Y_t|Z_t, Z_{t-1}, Z_{t-2}]$$

kako bismo mogli opisati učinak varijable Z u različitim vremenskim trenucima na trenutnu vrijednost ovisne varijable Y. Zajedno s pretpostavkom (4.8) dobijemo sljedeće:

$$E[Y_t|Z_t, Y_{t-1}, Z_{t-1}, ...] = E[Y_t|Z_t, Z_{t-1}, Z_{t-2}],$$

tj. zaključujemo da nijedan drugi korak od Y niti Z, osim onih uključenih u model, nemaju efekt na trenutnu vrijednost ovisne varijable Y.
Generalno, za model

$$Y_t = \beta_0 + \beta_1 X_{t1} + ... + \beta_k X_{tk} + U_t$$

u kojem vektor ovisnih varijabli $\boldsymbol{X}_t$ može, ali i ne mora sadržavati prošle vrijednosti ovisne ili neke od neovisnih varijabli, pretpostavljamo

$$E[U_t|\boldsymbol{X}_t, Y_{t-1}, \boldsymbol{X}_{t-1}, ...] = 0. \tag{4.9}$$

Tada uvjetno očekivanje na ovisnu varijablu u trenutku t možemo zapisati na sljedeći način:

$$E[Y_t|\boldsymbol{X}_t, X_{t-1}, \boldsymbol{X}_{t-1}, ...] = E[Y_t|\boldsymbol{X}_t].$$

To bi značilo da, bez obzira na to kakve su neovisne varijable uključene u model, one su dovoljne za opisivanje trenutne vrijednosti ovisne varijable te poznavanje i njihovih ostalih prošlih vrijednosti neće doprinijeti boljem opisivanju trenutne vrijednosti ovisne varijable. Za model koji zadovoljava pretpostavku (4.9) kažemo da je dinamički potpun model te za takav model

vrijedi da uvijek zadovoljava Pretpostavku 4.3.5 o serijskoj nekoreliranosti (što je dokazano u prethodnom primjeru).

U praksi nije moguće uvijek dobiti dinamički potpun model, kao ni izbjeći serijsku koreliranost grešaka modela, stoga je nužno poznavati kakve posljedice nosi serijska koreliranost te na koji način tretirati takve modele. U nastavku ćemo se detaljnije baviti upravo modelima s koreliranim greškama.

4.4 Serijska korelacija

Prisjetimo se, serijska korelacija, ili autokorelacija, pojavljuje se kad su greške modela u različitim vremenskim trenucima korelirane. Do zaključka o koreliranosti dolazimo analizom reziduala. Najčešći oblik serijske korelacije u modelima jest koreliranost u dva uzastopna trenutka, odnosno na koraku 1. Ako greške u trenucima t i $(t-1)$ možemo prikazati na sljedeći način:

$$U_t = \rho U_{t-1} + E_t, \quad -1 < \rho < 1,$$

tada kažemo da je prisutna serijska korelacija prvog reda, a parametar ρ zovemo autokorelacijskim koeficijentom prvog reda. On nam govori o vrsti i jačini veze između dva uzastopna reziduala – za $\rho > 0$ možemo reći da ukazuje na pozitivnu serijsku korelaciju te očekujemo da će dva reziduala biti istog predznaka. Analogno, za $\rho < 0$ očekujemo da će idući rezidual biti suprotnog predznaka od onog trenutačnog. Također, serijska korelacija se isto tako može dogoditi na nekom većem koraku, što se najčešće događa u podacima u kojima postoji sezonalnost. Prisutnost serijske korelacije u podacima predstavlja problem jer ukazuje na to da možda postoji bitna informacija koja nije uključena u model te da model iznova ponavlja sličnu grešku pokušavajući opisati ovisnu varijablu [4].

Prisutnost serijske korelacije u modelu utječe i na OLS procjenitelje. Prisjetimo se, teoremi koji govore o nepristranosti i konzistentnosti OLS procjenitelja zahtijevali su strogu egzogenost, no ne i nekoreliranost grešaka što bi značilo da postojanje serijske koreliranosti neće utjecati na nepristranost i konzistentnost procjenitelja. No ako se prisjetimo i Gauss-Markovljeva teorema, OLS procjenitelji neće biti najbolji procjenitelji (BLUE), što ćemo pokazati sljedećim teoremom. Teorem je radi jednostavnosti iskazan u slučaju jednostavne linearne regresije.

Teorem 4.4.1 ([9], str. 413)**.** *Pretpostavimo da je u linearnom regresijskom modelu*

$$Y_t = \beta_0 + \beta_1 X_t + U_t$$

prisutna serijska korelacija, odnosno grešku modela možemo modelirati AR(1) modelom na sljedeći način:

$$U_t = \rho U_{t-1} + E_t, |\rho| < 1, E_t \sim WN(0, \sigma_t^2).$$

Tada za procjenitelj $\hat{\beta}_1$ vrijedi:

$$Var(\hat{\beta}_1 | \boldsymbol{X}) = \frac{\sigma_U^2}{\sum_{t=1}^{T} X_t^2} + 2\sigma_U^2 \frac{\sum_{t=1}^{T-1} \sum_{s=1}^{T-t} \rho^s X_t X_{t+s}}{(\sum_{t=1}^{T} X_t^2)^2} \tag{4.10}$$

Izraz (4.10) govori nam da varijanca OLS procjenitelja ovisi i o koeficijentu korelacije greške ρ. Ako je $\rho = 0$, odnosno greške nisu korelirane, izraz (4.10) jednak je izrazu za varijancu OLS procjenitelja uz Gauss-Markovljeve pretpostavke. U suprotnom, ako korelacija postoji i $\rho \neq 0$, varijanca može biti veća (ili manja). Primjerice, ako postoji pozitivna korelacija, drugi pribrojnik u izrazu (4.10) bit će pozitivan te će varijanca biti veća od one koju bismo dobili zanemarivanjem serijske korelacije – dakle mogli bismo zaključiti da je OLS procjenitelj $\hat{\beta}_1$ precizniji nego što zaista jest. U drugom slučaju, ako se dogodi da je drugi pribrojnik u izrazu (4.10) negativan, stvarna varijanca bit će manja nego ona koju bismo dobili pretpostavljajući da nema serijske korelacije. S obzirom na to da se standardna greška računa kao procjena standardne devijacije procjenitelja koja ne uključuje drugi dio izraza (4.10), uz prisutnost serijske korelacije neće biti konzistentna, stoga ni statistički zaključci koji se oslanjaju na standardnu grešku neće biti valjani. U nastavku ćemo se detaljnije osvrnuti na neke testove koji se upotrebljavaju za provjeru serijske korelacije u regresijskim modelima.

4.4.1 Testiranje hipoteze o nekoreliranosti grešaka modela

Kako smo spomenuli, serijsku korelaciju u podacima možemo uočiti analizom reziduala. Dosad smo spomenuli koje probleme s OLS procjeniteljima serijska korelacija može uzrokovati te smo izveli izraz za varijancu OLS procjenitelja u jednostavnom regresijskom modelu u kojem je greška modelirana AR(1) modelom uz prisutnost korelacije. U nastavku ćemo pokazati kako testirati prisutnost AR(1) korelacije.
Prisjetimo se, grešku modela možemo opisati AR(1) modelom na sljedeći način:

$$U_t = \rho U_{t-1} + E_t, t = 1, ..., n, \tag{4.11}$$

$$|\rho| < 1. \tag{4.12}$$

Za početak, pretpostavimo da su u linearnom regresijskom modelu

$$Y_t = \beta_0 + \beta_1 X_{t1} + ... + \beta_k X_{tk} + U_t \tag{4.13}$$

neovisne varijable strogo egzogene. Uz (4.11) moramo pretpostaviti i

$$E[E_t | U_{t-1}, U_{t-2}, ...] = 0 \tag{4.14}$$

te

$$Var(E_t | U_{t-1}) = Var(E_t) = \sigma_E^2. \tag{4.15}$$

Uz pretpostavke (4.14) i (4.15) možemo upotrebljavati asimptotske rezultate velikih uzoraka. Ako želimo testirati hipotezu o nekoreliranosti, nul-hipoteza je sljedeća:

$$H_0 : \rho = 0.$$

Kad bismo grešku modela mogli mjeriti i bilježiti, problem procjene koeficijenta korelacije sveo bi se na procjenu parametra jednostavnog linearnog regresijskog modela definiranog u (4.11). No, s obzirom na to da greška nije mjerljiva, moramo raditi s njezinom procjenom, odnosno rezidualima. Testiranje uz pretpostavku stroge egzogenosti i greške modelirane AR(1) modelom jednostavno provodimo u sljedeća tri koraka ([9]):

(i) Provođenjem OLS regresije na modelu (4.13) dobijemo procjenitelje $\hat{\beta}_0, ..., \hat{\beta}_k$, a na temelju njih i reziduale $\hat{u}_t, \forall t = 1, 2, ..., n$.

(ii) S dobivenim rezidualima provesti OLS regresiju $\hat{u}_t$ na $\hat{u}_{t-1}, \forall t = 2, ..., n$, tj. za model

$$\hat{u}_t = \rho \hat{u}_{t-1} + e_t.$$

(iii) t statistiku $t_{\hat{\rho}}$ dobivenu na temelju procijenjenog parametra $\hat{\rho}$ ([9], str. 122) iskoristiti za testiranje nul-hipoteze, uz alternativnu hipotezu

$$H_1 : \rho \neq 0 \quad (\rho > 0).$$

Na temelju potonjeg testa u slučaju odbacivanja nul-hipoteze možemo tvrditi da u početnom modelu s navedenim pretpostavkama postoji serijska korelacija na koraku 1. Primijetimo da se testom može detektirati serijska korelacija i na nekom većem koraku - no uz uvjet da postoji i na koraku 1.
Još jedan test koji je moguće upotrebljavati za provjeru postojanja serijske korelacije, uz uvjet da se greška modela može modelirati AR(1) modelom, jest Durbin-Watsonov test. Test statistika DW testa također upotrebljava OLS reziduale, a temelji se na sumi kvadrata razlike reziduala u uzastopnim trenucima. Dakle, na početku promatramo dva susjedna reziduala, $\hat{u}_s$ i $\hat{u}_{s-1}, s = 2, ..., n$, odnosno njihovu razliku $\hat{u}_s - \hat{u}_{s-1}$. Kvadriranjem tih razlika te sumiranjem po svim trenucima $s = 2, ..., n$ te dijeljenjem sumom kvadrata svih reziduala dobivamo DW test statistiku:

$$DW = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}.$$

Usporedbom DW statistike s izrazom za procjenu autokorelacijskog koeficijenta prvog reda $\hat{\rho}$ može se pokazati sljedeći odnos (vidi [1], str. 25):

$$DW \approx 2 - 2\hat{\rho}. \tag{4.16}$$

Stoga, hipoteza o nepostojanju serijske korelacije

$$H_0 : \rho = 0$$

jest ekvivalentna hipotezi

$$H_0 : DW = 2.$$

Također, ako u alternativnoj hipotezi postavimo $\rho > 0$, u kontekstu DW statistike možemo ju prikazati kao

$$H_1 : DW < 2.$$

Iz (4.16) možemo vidjeti da što je DW statistika bliža 0, to je $\hat{\rho}$ bliže 1, odnosno možemo pretpostaviti da postoji jaka pozitivna serijska korelacija. Također, što je DW statistika bliža 4, pretpostavljamo da postoji jaka negativna serijska korelacija. Primijetimo da na taj način možemo testirati samo serijsku korelaciju na koraku 1, što je glavni nedostatak DW testa. Također, još su neki nedostaci to što kritične vrijednosti ovise o broju neovisnih varijabli, stoga se ne može generalizirati te se može pokazati da se ne može upotrebljavati za testiranje u modelima s autoregresivnom strukturom reda većeg od 1.
U slučaju kad model ne može zadovoljiti strogu egzogenost, odnosno ako postoji korelacija između neovisne varijable i greške u nekom od trenutaka, nijedan od navedenih testova neće biti valjan. Primjerice, ako imamo model koji u neovisnim varijablama uključuje i prošle vrijednosti ovisne varijable, očito je da će postojati autokorelacija. U tom slučaju test možemo provesti u sljedećim koracima [9]:

(i) Provesti OLS regresiju y_t na $x_{t1}, ..., x_{tk}$ te dobiti vrijednosti reziduala $\hat{u}_t, t = 1, ..., n$.

(ii) Provesti OLS regresiju $\hat{u}_t$ na $x_{t1}, ..., x_{tk}, \hat{u}_{t-1}, t = 2, ..., n$. Na taj način dobije se procjena koeficijenta $\hat{\rho}$ te njegova t statistika $t_{\hat{\rho}}$.

(iii) Dobivenu statistiku $t_{\hat{\rho}}$ iskoristiti za testiranje hipoteze

$$H_0 : \rho = 0,$$

uz alternativnu hipotezu

$$H_1 : \rho \neq 0.$$

Na taj način dopuštamo da bilo koja od varijabli bude korelirana s u_{t-1}.
Nadalje, potonji test može se provesti i za modele čija je greška opisana nekim AR modelom reda većeg od 1.

Generalno, za AR(q) model u kojem vrijedi

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + ... + \rho_q U_{t-q} + E_t$$

i za nul-hipotezu

$$H_0 : \rho_1 = \rho_2 = ... = \rho_q = 0,$$

testiranje provodimo u sljedećim koracima [9]:

(i) Provesti OLS regresiju y_t na $x_{t1}, ..., x_{tk}$ te dobiti vrijednosti reziduala $\hat{u}_t, t = 1, ..., n$.

(ii) Provesti OLS regresiju $\hat{u}_t$ na $x_{t1}, x_{t2}, ..., x_{tk}, \hat{u}_{t-1}, \hat{u}_{t-2}, ..., \hat{u}_{t-q}, t = (q+1), ..., n$.

(iii) Provesti F test za zajedničku značajnost $\hat{u}_{t-1}, \hat{u}_{t-2}, ..., \hat{u}_{t-q}$ na temelju kojeg donosimo zaključak o istinitosti nul-hipoteze, odnosno postojanju autokorelacije.

Test ne zahtijeva strogu egzogenost, a u slučaju da ona ipak postoji, tada varijable X_{tj} koje su nekorelirane s $U_{t-1}, U_{t-2}, ..., U_{t-q}$ mogu biti izostavljene iz drugog koraka. Ono što zahtijeva jest pretpostavka homoskedastičnosti, tj.

$$Var(U_t|\boldsymbol{X}_t, U_{t-1}, ..., U_{t-q}) = \sigma^2.$$

U slučaju kad u modelu uočavamo pojavu sezonalnosti, odnosno pretpostavljamo postojanje korelacije, npr., na koraku 4 za kvartalne podatke,

$$U_t = \rho_4 U_{t-4} + E_t,$$

možemo postupati kao i kod AR(1) testova za serijsku korelaciju i upotrebljavati t statistiku za regresiju $\hat{U}_t$ na $\hat{U}_{t-4}, t = 5, ..., n$.
Još jedna generalizacija DW testa jest Breusch-Godfreyev test koji se koristi za detektiranje autokorelacije višeg reda. Procedura ima korake jednake prethodno navedenim, dakle temelji se na regresiji reziduala na svim ostalim rezidualima te procjenom pripadnih koeficijenata autokorelacije, ρ_i. Pritom je nul-hipoteza

$$H_0 : \rho_1 = \rho_2 = ... = \rho_q = 0,$$

a ako ju odbacimo, zaključujemo da postoji autokorelacija na nekom koraku.

4.4.2 Korekcije modela sa serijskom korelacijom

Nakon provođenja testova, ako zaključimo da postoji serijska korelacija grešaka u modelu, potrebno je nešto poduzeti kako bismo ju pokušali ukloniti ili kako bismo mogli donijeti valjane statističke zaključke. Jedan od načina elimininiranja serijske korelacije iz procesa grešaka modela jest diferenciranje podataka. Za početak, pretpostavimo da imamo model koji zadovoljava Gauss-Markovljeve pretpostavke 4.2.1 – 4.2.4 čije greške možemo modelirati AR(1) modelom,

$$U_t = \rho U_{t-1} + E_t, \quad t = 1, 2, ... \tag{4.17}$$

Varijancu greške U_t možemo prikazati na sljedeći način:

$$Var(U_t) = \frac{\sigma_E^2}{1 - \rho^2}.$$

Nadalje, radi jednostavnosti izračuna pretpostavimo da je model definiran na sljedeći način:

$$Y_t = \beta_0 + \beta_1 X_t + U_t, \quad t = 1, 2, ..., n. \tag{4.18}$$

Za $t > 1$ možemo ga prikazati i u terminima $(t-1)$, tj.

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + U_{t-1}. \tag{4.19}$$

Ako izraz (4.19) pomnožimo s ρ, oduzmemo od izraza (4.18) te iskoristimo (4.17), dobijemo sljedeće:

$$Y_t - \rho Y_{t-1} = (1-\rho)\beta_0 + \beta_1(X_t - \rho X_{t-1}) + E_t, \quad t \geq 2.$$

Ako uvedemo sljedeće oznake $\widetilde{Y}_t = Y_t - \rho Y_{t-1}$ i $\widetilde{X}_t = X_t - \rho X_{t-1}$, prethodni izraz jednak je

$$\widetilde{Y}_t = (1-\rho)\beta_0 + \beta_1 \widetilde{X}_t, \quad t \geq 2. \tag{4.20}$$

Tako definirane $\widetilde{Y}_t$ i $\widetilde{X}_t$ zovemo kvazidiferenciranim nizovima. U specifičnom slučaju kad je $\rho = 1$ zovemo ih samo diferenciranim nizovima. Primijetimo da su u modelu (4.20) greške serijski nekorelirane te da takav model zadovoljava sve Gauss-Markovljeve pretpostavke, stoga bismo, ako bismo znali vrijednost parametra ρ, mogli procijeniti parametre β_0 i β_1. Problem je što tako dobiveni procjenitelji ne mogu biti najbolji linearni procjenitelji s najmanjom varijancom jer nisu definirani za $t = 1$. Ako početni model (4.18) zapišemo u trenutku $t = 1$,

$$Y_1 = \beta_0 + \beta_1 Y_1 + U_1, \tag{4.21}$$

s obzirom na to da su E_t i U_1 nekorelirani za svaki t, uključivši (4.21) u (4.20) možemo i dalje imati serijski nekorelirane greške, no varijanca greške više neće biti konstantna, $Var(U_1) = \frac{\sigma_E^2}{1-\rho^2} > \sigma_E^2 = Var(E_t)$. Kako bismo postigli istu varijancu, izraz (4.21) moramo pomnožiti s $\sqrt{(1-\rho^2)}$. Tada dobijemo

$$\widetilde{Y}_1 = \sqrt{(1-\rho^2)}\beta_0 + \beta_1 \widetilde{X}_1 + \widetilde{U}_1, \tag{4.22}$$

gdje su $\widetilde{U}_1 = \sqrt{(1-\rho^2)}U_1$, $\widetilde{Y}_1 = \sqrt{(1-\rho^2)}Y_1$ i $\widetilde{X}_1 = \sqrt{(1-\rho^2)}X_1$. Primijetimo da je sada $Var(\widetilde{U}_1) = (1-\rho^2)Var(U_1) = \sigma_E^2$, odnosno da (4.20) i (4.22) možemo upotrebljavati zajedno, uz jednostavne transformacije uz poznati ρ, kako bismo dobili najbolje linearne procjenitelje s najmanjom varijancom za parametre β_0 i β_1.
Generalno, za $t \geq 2$, model možemo zapisati na sljedeći način:

$$\widetilde{Y}_t = (1-\rho)\beta_0 + \beta_1 \widetilde{X}_{t1} + ... + \beta_k \widetilde{X}_{tk} + E_t, \tag{4.23}$$

gdje je $\widetilde{X}_{tj} = X_{tj} - \rho X_{(t-1)j}$. Za $t = 1$ imamo $\widetilde{Y}_1 = \sqrt{(1-\rho^2)}Y_1, \widetilde{X}_{1j} = \sqrt{(1-\rho^2)}X_1 j$, dok je konstantni član $\sqrt{(1-\rho^2)}\beta_0$. Takvim transformiranjem nizova možemo ukloniti serijsku korelaciju te postići da procjenitelji budu najbolji linearni procjenitelji s najmanjom varijancom, što omogućava da daljnji statistički zaključci budu valjani. Problem je to što u praksi najčešće ne znamo vrijednost parametra ρ, nego ga moramo procijeniti. Problem određivanja procjene $\hat{\rho}$ za parametar ρ već nam je poznat, a možemo ga dobiti regresijom reziduala $\hat{u}_t$ na $\hat{u}_{t-1}$, u slučaju kad se greška može modelirati AR(1) modelom. U tom slučaju, (4.23) postaje

$$\widetilde{Y}_t = \beta_0 \widetilde{X}_{t0} + \beta_1 \widetilde{X}_{t1} + ... + \beta_k \widetilde{X}_{tk} + E_t, \tag{4.24}$$

gdje je $\widetilde{X}_{t0} = (1 - \hat{\rho})$ za $t \geq 2$ i $\widetilde{X}_{10} = (1 - \hat{\rho}^2)^{1/2}$. Procjenitelji za parametre iz modela zovu se dostižni GLS procjenitelji (FGLS). Postoji više metoda za dobivanje FGLS procjenitelja, a najpoznatije su Cochrane-Orcuttova (CO) procedura i Prais-Winstenova (PW) procedura ([9], str. 425). Razlikuju se u tome što se u CO proceduri jednostavno izostavlja prvo mjerenje, dok PW procedura upotrebljava prvo mjerenje na način koji je dobiven u (4.24). U nastavku ćemo opisati korake CO procedure:

(i) Provesti OLS regresiju y_t na $x_{t1}, ..., x_{tk}$ te izračunati reziduale $\hat{u}_t, t = 1, 2, ..., n$.

(ii) Provesti regresiju $\hat{u}_t$ na $\hat{u}_{t-1}$ te procijeniti koeficijent prvog reda serijske korelacije $\hat{\rho}$.

(iii) Transformirati varijable tako da dobijemo $\tilde{y}_t$ i $\tilde{x}_{tk}$.

(iv) S dobivenom procjenom $\hat{\rho}$ provesti OLS regresiju za (4.24) te dobiti procjene za $\beta_0, ..., \beta_k$.

(v) S dobivenim procjenama ponovno izračunati reziduale i vratiti se na drugi korak i ponoviti proceduru.

Iterativni proces staje kad se procjena parametra ρ stabilizira. S obzirom na to da umjesto prave vrijednosti parametra ρ upotrebljavamo procjenu koja sadrži neku grešku, FGLS procjentitelji više nisu nepristrani, ali su konzistentni. Stoga sve statističke zaključke moramo donositi u kontekstu asimptotske teorije. OLS i FGLS procjenitelji dobiveni su različitim procedurama, stoga je za očekivati i da ne daju iste procjene. U slučaju kad imamo prisutnost serijske korelacije, sugerira se upotreba FGLS procjenitelja s obzirom na to da su u tom slučaju test statistike asimptotski valjane. Uz male izmjene potonje procedure mogu se provesti i za modele sa serijskom korelacijom višeg reda.

5 Modeliranje koncentracije peludi

U ovom dijelu rada bit će opisan praktični dio koji se temelji na teorijskim osnovama navedenim u prethodnim poglavljima. Na temelju stvarnih podataka koji su zabilježeni tijekom vremena, odnosno u formi vremenskih nizova, pokušat ću primijeniti modeliranje linearnom regresijom te zatim ispitati pretpostavke modela. Nakon toga bit će donesen zaključak o valjanosti i upotrebi modela.
Za izgradnju modela upotrebljavani su podaci o koncentraciji peludi ambrozije te meteorološki podaci izmjereni u Novom Sadu u Republici Srbiji. Razdoblje mjerenja je od 1. siječnja 2000. do 30. lipnja 2017. godine. Mjerenja su bilježena svaka dva sata, a podaci s kojima raspolažemo prikazuju prosječne, minimalne i maksimalne vrijednosti mjerenja na dnevnoj razini. Skup podataka podijeljen je u dva dijela – podaci do kraja 2015. godine upotrebljavani su za modeliranje, a podaci iz 2016. i 2017. ostavljeni su za validaciju modela.
Ovisna varijabla koju ćemo pokušati opisati na temelju ostalih odnosit će se na prosječnu dnevnu koncentraciju peludi.

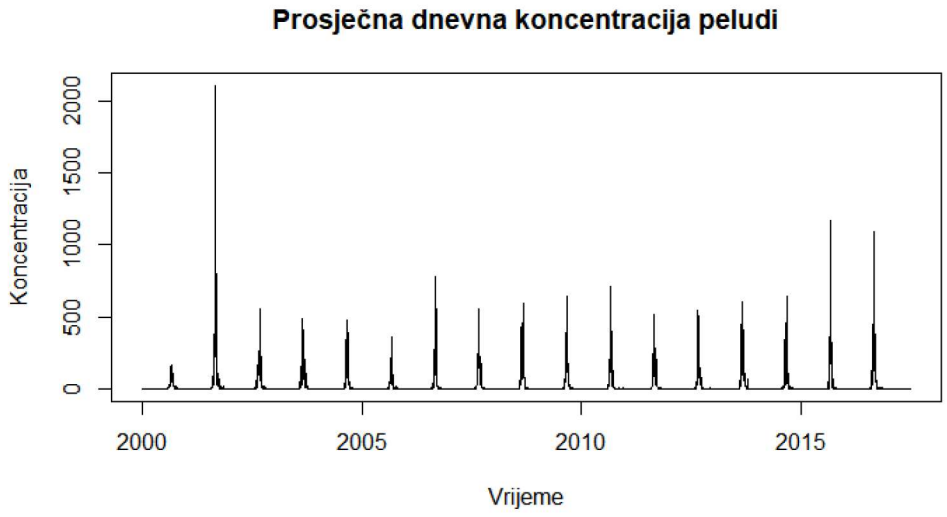
5.1 Opis varijabli

5.1.1 Pelud

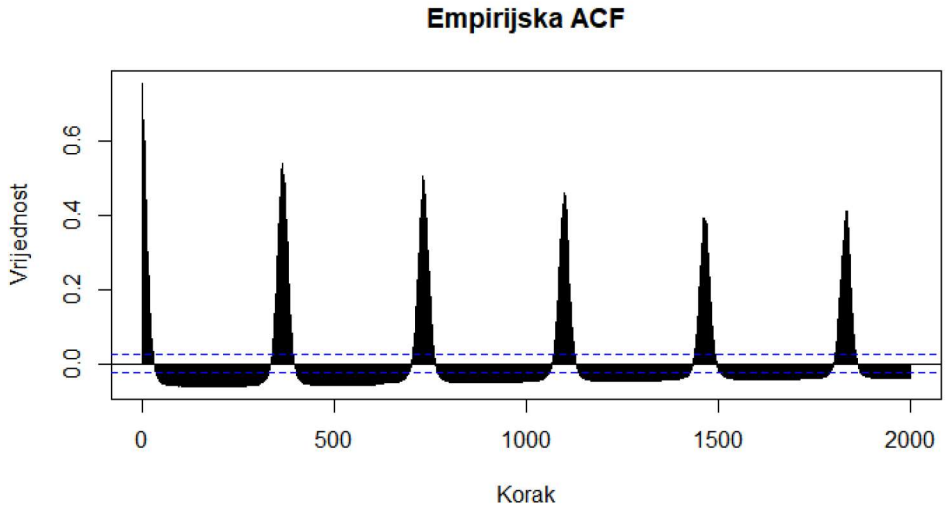
Koncentracija peludi mjeri se brojem zrnaca po kubnom metru zraka. U bazi imamo zabilježene dnevne prosječne vrijednosti, dnevne maksimalne vrijednosti te neke transformacije kao što je, npr., logaritamska. Sezona cvjetanja ambrozije traje od lipnja do listopada. S obzirom na to da nas zanima dnevna prosječna koncentracija peludi, prikazat ćemo osnovne karakteristike te varijable tijekom cijele godine te sezone.

	Min	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Max
Cijela godina	0	0	0	18.67	1.298	2104.58
Sezona	0	0	1.167	45.09	30.50	2104.58

Tablica 1: Numeričke karakteristike prosječne dnevne koncentracije peludi (2000.-2017.)

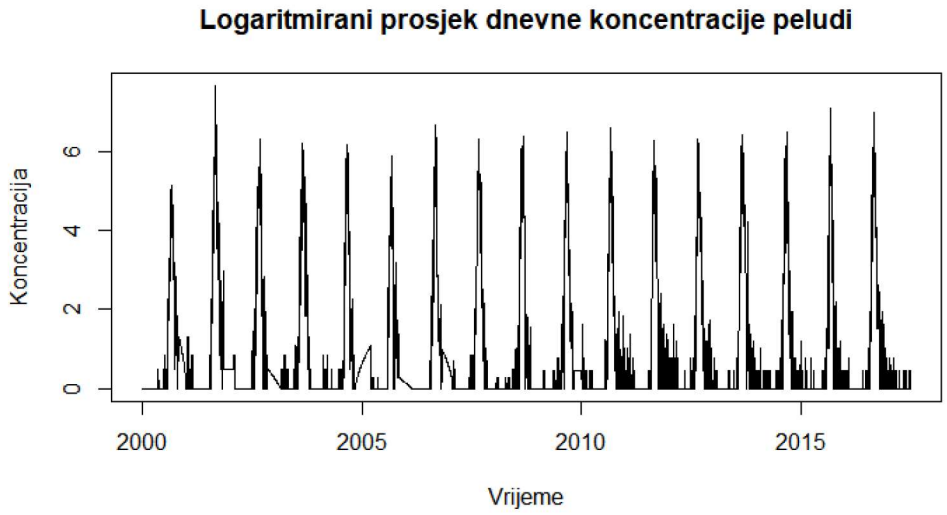


Slika 1: Trajektorija prosječne dnevne koncentracije peludi (2000.-2017.)



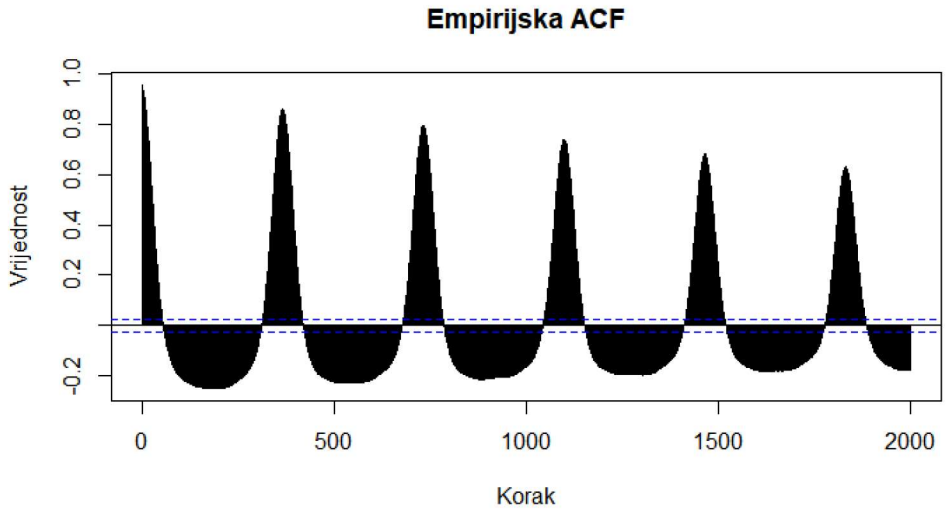
Slika 2: Empirijska ACF prosječne dnevne koncentracije peludi (2000.-2017.)

Trajektorija na Slici 1 te graf uzoračke autokorelacijske funkcije na Slici 2 sugeriraju nestacionarnost prosječnih koncentracija peludi, stoga ih logaritamski transformiramo funkcijom $x \to ln(1+x)$. Pogledajmo trajektoriju transformirane varijable:



Slika 3: Trajektorija transformirane koncentracije peludi (2000.-2017.)

Na Slici 3 jasno možemo uočiti postojanje sezonalnosti, pri čemu maksimalne vrijednosti trajektorije u svakoj godini pripadaju sezoni cvjetanja ambrozije.
Nadalje, u početnom procesu prosječne dnevne koncentracije peludi iz Slike 2 možemo naslutiti postojanje serijske koreliranosti koncentracija. Za transformiranu varijablu iz sljedeće slike ne možemo naslutiti ništa bolje.



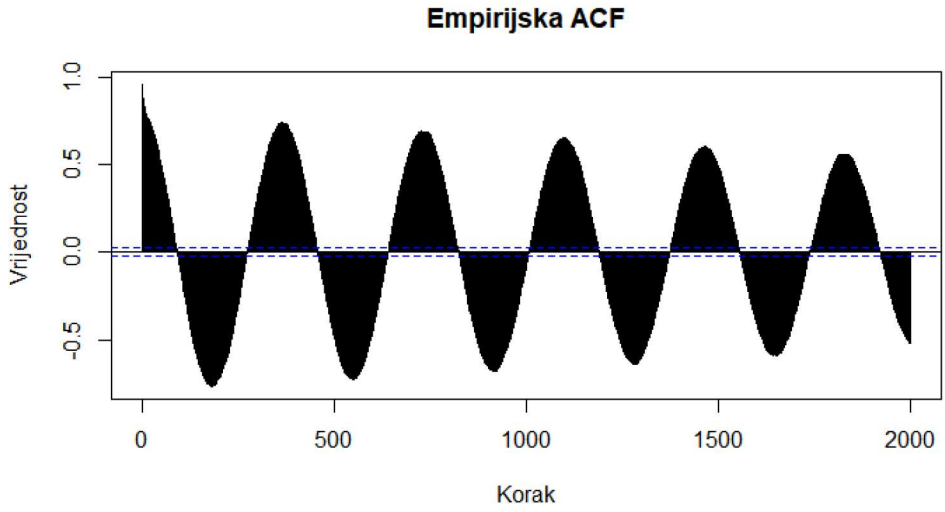
Slika 4: Empirijska ACF logaritmirane koncentracije peludi (2000.-2017.)

5.1.2 Temperatura

Temperatura zraka mjeri se u Celzijevim stupnjevima. Za mjerenja u kojima nije navedena temperatura uzet je prosjek prethodnog i idućeg dana. U bazi podataka imamo zabilježene prosječne, minimalne i maksimalne vrijednosti dnevne temperature. Pogledajmo osnovne karakteristike prosječne temperature:

	Min	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Max
Cijela godina	-19.6	5.225	12.7	12.085	19.4	32.4
Sezona	1.1	16.20	20	19.45	23.10	32.4

Tablica 2: Numeričke karakteristike prosječne dnevne temperature (2000.-2017.)



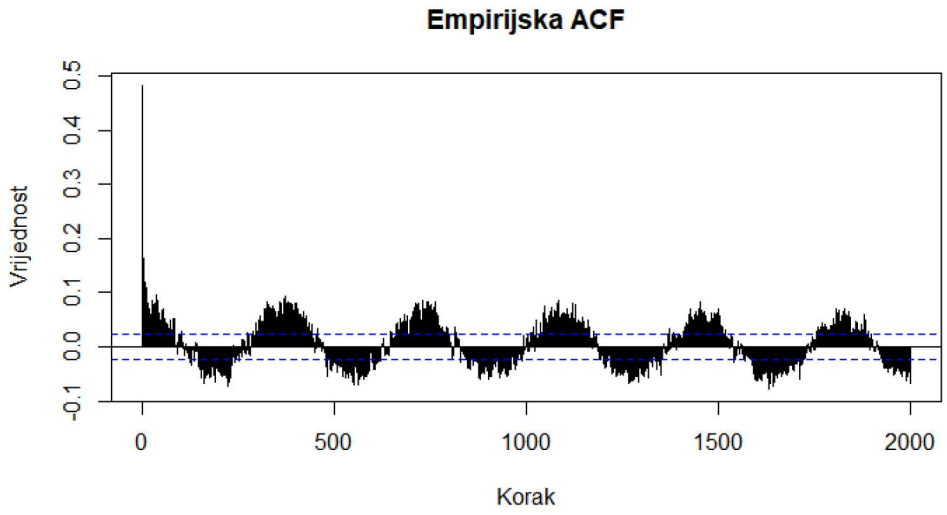
Slika 5: Empirijska ACF prosječne dnevne temperature (2000.-2017.)

5.1.3 Vjetar

Brzina vjetra mjerena je u m/s. U bazi imamo zabilježene prosječnu dnevnu brzinu vjetra za svaki dan. Pogledajmo osnovne karakteristike:

	Min	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Max
Cijela godina	0.4	5.9	7.6	8.834	10.7	41.3
Sezona	0.9	5.4	6.7	7.537	9.1	31.9

Tablica 3: Numeričke karakteristike prosječne dnevne brzine vjetra (2000.-2017.)



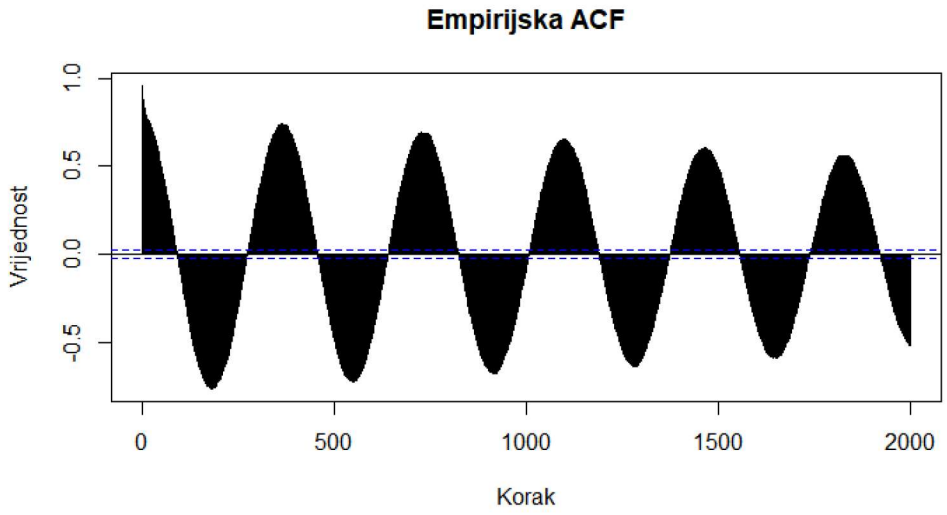
Slika 6: Empirijska ACF prosječne dnevne brzine vjetra (2000.-2017.)

5.1.4 Vlažnost zraka

Prosječna dnevna vlažnost zraka iskazuje se u postocima od 1 do 100%. Pogledajmo osnovne karakteristike:

	Min	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Max
Cijela godina	26	64	75	74	85	100
Sezona	34	61	69	69.45	78	98

Tablica 4: Numeričke karakteristike prosječne dnevne vlažnosti zraka (2000.-2017.)



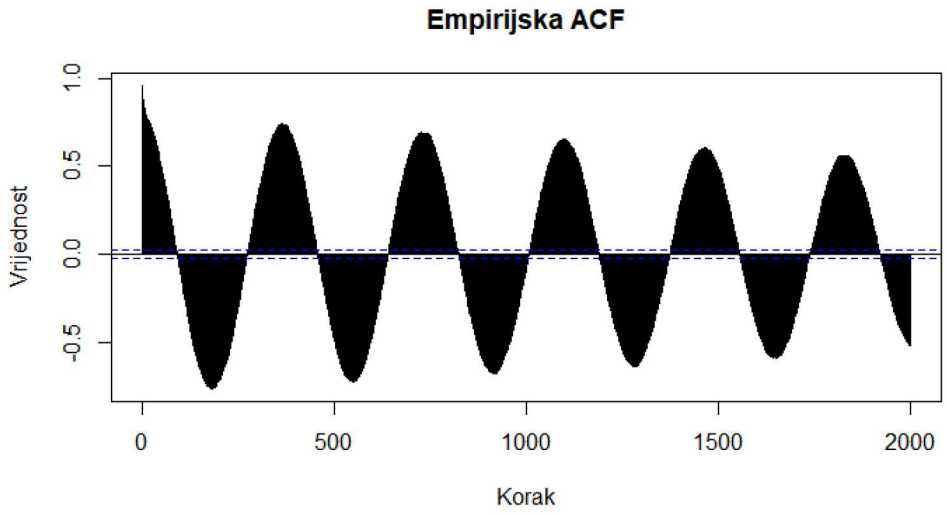
Slika 7: Empirijska ACF prosječne dnevne vlažnosti zraka (2000.-2017.)

5.1.5 Padaline

Zabilježene vrijednosti padalina izmjerene su u mm što predstavlja njihovu količinu u l/m^2. Pogledajmo osnovne karakteristike:

	Min	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Max
Cijela godina	0	0	0	1.808	0.768	121.92
Sezona	0	0	0	2.241	0.6	121.92

Tablica 5: Numeričke karakteristike dnevne količine padalina (2000.-2017.)

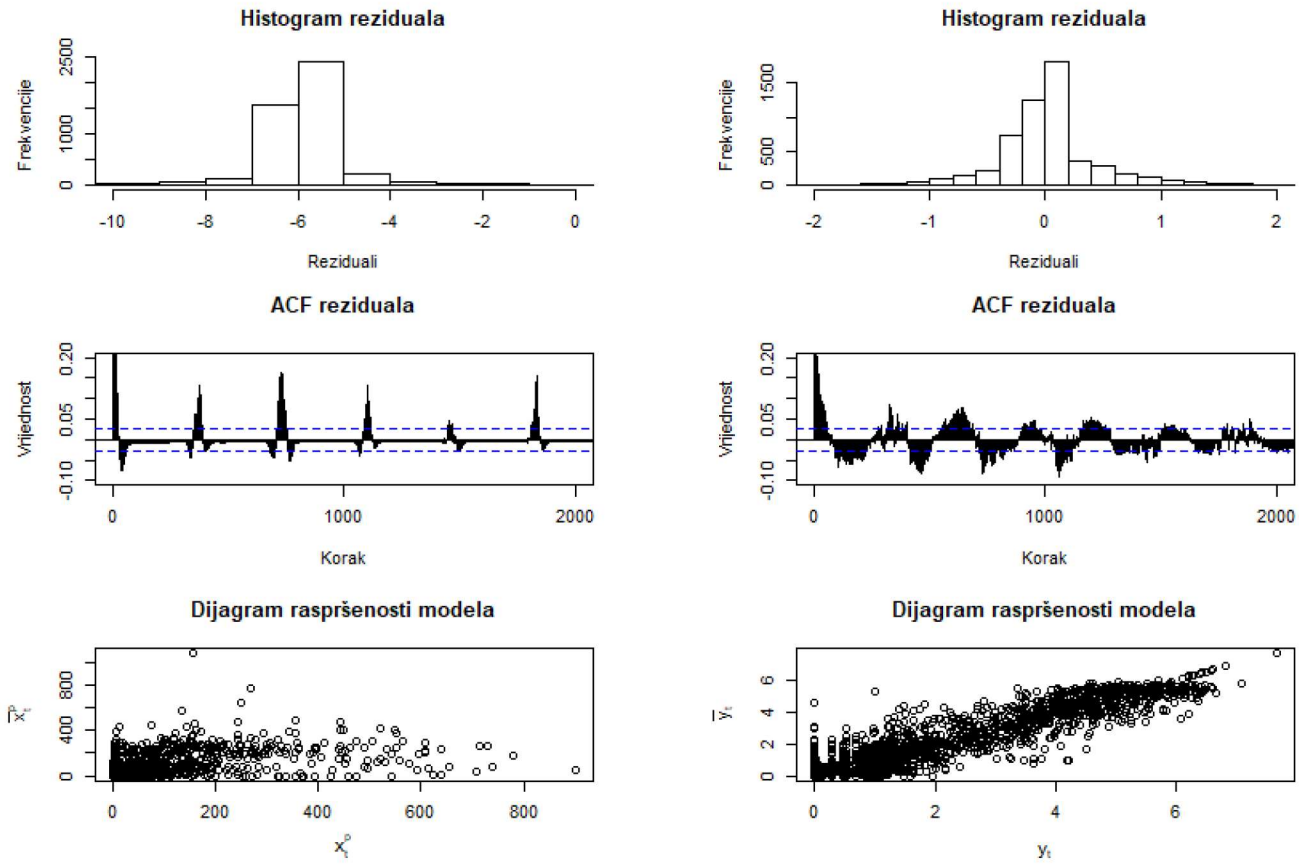


Slika 8: Empirijska ACF prosječne dnevne količine padalina (2000.-2017.)

Primijetimo da je u svim navedenim varijablama prisutna serijska koreliranost i to na velikom broju koraka, što je u skladu s uočenom sezonalnošću. Takvo je ponašanje očekivano za meteorološke varijable.
Osim osnovnih prikazanih vrijednosti, u bazi se nalaze još neke transformacije varijabli kao što su prosjeci posljednjih 7 dana, vrijednost prethodnog dana itd. S obzirom na to da ih ima ukupno 67, nećemo analizirati svaku posebno, već ćemo analizirati one koje su odabrane za model.

5.2 Odabir prediktora za linearni regresijski model

Prvo je potrebno odlučiti koju varijablu uzeti za ovisnu. Već smo vidjeli da log-transformacija $x \mapsto ln(1+x)$ primjenjena na prosječnu dnevnu koncentraciju peludi generira varijablu s prihvatljivijim statističim svojstvima za modeliranje. Dodatno, definirajmo dva jednostavna linearna modela. Prvi, čija je ovisna varijabla x_t^p prosječna dnevna koncentracija peludi, a neovisna $\bar{x}_t^p$ prosjek srednjih koncentracija peludi na isti dan u svim dostupnim prethodnim godinama, te drugi, čija je ovisna varijabla $y_t = ln(1 + x_t^p)$ logaritmirana prosječna dnevna koncentracija peludi, a neovisna $\bar{y}_t$ prosjek log-transformacija prosječnih koncentracija peludi na isti dan u svim dostupnim prethodnim godinama.



Slika 9: Usporedba originalne (lijevo) i log-transformirane (desno) prosječne dnevne koncentracije peludi

Na Slici 9 jasno vidimo da je transformirana varijabla bolji odabir – i reziduali i autokorelacijska funkcija bliže su željenoj strukturi reziduala u linearnim regresijskim modelima, a dijagram raspršenosti sugerira prikladnost linearne veze za opisivanje ovisnosti logaritmirane prosječne dnevne koncentracije o prosjeku log-transformacija koncentracija peludi na isti dan u svim dostupnim prethodnim godinama. Iz tog razloga odlučujemo se za transformiranu koncentraciju peludi, a varijabla $\bar{y}_t$ bit će glavni prediktor u modelu.
Nadalje, za odabir ovisnih varijabli modela upotrebljavane su najprije *splitwise* procedure u R-u koje sugeriraju najadekvatniji model. Kako je sugerirani model sadržavao vrlo velik broj varijabli, isprobano je više različitih mogućnosti redukcije tog modela te su u konačnici odabrane one koje su intuitivno najlogičnije, uz napomenu da nismo mnogo izgubili tim odabirom, odnosno model je i dalje dovoljno dobar u smislu značajnosti varijabli i mjera kao što je, npr., koeficijent determinacije R^2, ali i u smislu analize pretpostavki koje su nam potrebne za daljni rad s modelom.
Odabrane su varijable sljedeće:

1. prosjek log-transformacija prosječnih koncentracija peludi na isti dan u svim dostupnim prethodnim godinama : $\overline{y}_t$

2. prosjek log-transformacija prosječnih koncentracija peludi za 7 dana koji prethode danu t:
$$\overline{y}_{(t-7):(t-1)} = \frac{y_{t-7} + y_{t-6} + y_{t-5} + y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1}}{7}$$

3. prosječna temperatura u danu $(t-1)$: $x_{t-1}^{(tmp)}$

4. maksimum prosječnih temperatura za 6 dana koji prethode danu $(t-1)$:
$$\widehat{x}_{(t-7):(t-2)}^{(tmp)} = \max\left\{x_{t-7}^{(t)}, x_{t-6}^{(t)}, x_{t-5}^{(t)}, x_{t-4}^{(t)}, x_{t-3}^{(t)}, x_{t-2}^{(t)}\right\}$$

5. prosječna vlažnost zraka u danu $(t-1)$: $x_{t-1}^{(vl)}$

6. ukupna količina padalina u danu $(t-1)$: $x_{t-1}^{(pad)}$

7. prosječna brzina vjetra u danu $(t-1)$: $x_{t-1}^{(vj)}$

8. maksimum prosječnih brzina vjetra za 6 dana koji prethode danu $(t-1)$:
$$\widehat{x}_{(t-7):(t-2)}^{(vj)} = \max\left\{x_{t-7}^{(vj)}, x_{t-6}^{(vj)}, x_{t-5}^{(vj)}, x_{t-4}^{(vj)}, x_{t-3}^{(vj)}, x_{t-2}^{(vj)}\right\}$$

Grafički prikazi nizova vrijednosti prediktora (koji nisu uvršteni u ovaj rad) očekivano sugeriraju nestacionarnost. Stacionarnost možemo postići diferenciranjem niza, a to nam može suregirati provođenje ADF testa o postojanju jediničnog korijena. Za sve varijable posebno je proveden test, a dobivene p vrijednosti jednake su i iznose $p = 0.01 < 0.05$, što znači da je na razini značajnosti 0.05 u svakom slučaju odbačena nul-hipoteza o postojanju jediničnog korijena iz čega zaključujemo da nema potrebe diferencirati.
Sada kad smo odabrali varijable, provođenjem linearne regresije dobivamo sljedeću jednadžbu i procjene koeficijenata:

$$\begin{aligned} y_t = \beta_0 + \beta_1 \cdot \overline{y}_t + \beta_2 \cdot \overline{y}_{(t-7):(t-1)} + \beta_3 \cdot x_{t-1}^{(tmp)} + \beta_4 \cdot \widehat{x}_{(t-7):(t-2)}^{(tmp)} + \\ + \beta_5 \cdot x_{t-1}^{(vl)} + \beta_6 \cdot x_{t-1}^{(pad)} + \beta_7 \cdot x_{t-1}^{(vj)} + \beta_8 \cdot \widehat{x}_{(t-7):(t-2)}^{(vj)} + u_t \end{aligned} \tag{5.1}$$

$\hat{\beta}_0 = 0.1697581$ **			
$\hat{\beta}_1$	0.4860608 ***	$\hat{\beta}_5$	-0.0022595 ***
$\hat{\beta}_2$	0.5098022 ***	$\hat{\beta}_6$	-0.0032560 **
$\hat{\beta}_3$	0.0129544 ***	$\hat{\beta}_7$	0.0045537 **
$\hat{\beta}_4$	-0.0108482 ***	$\hat{\beta}_8$	-0.0032560 **

Tablica 6: OLS procjenitelji

** značajan na razini 0.01
*** značajan na razini 0.001

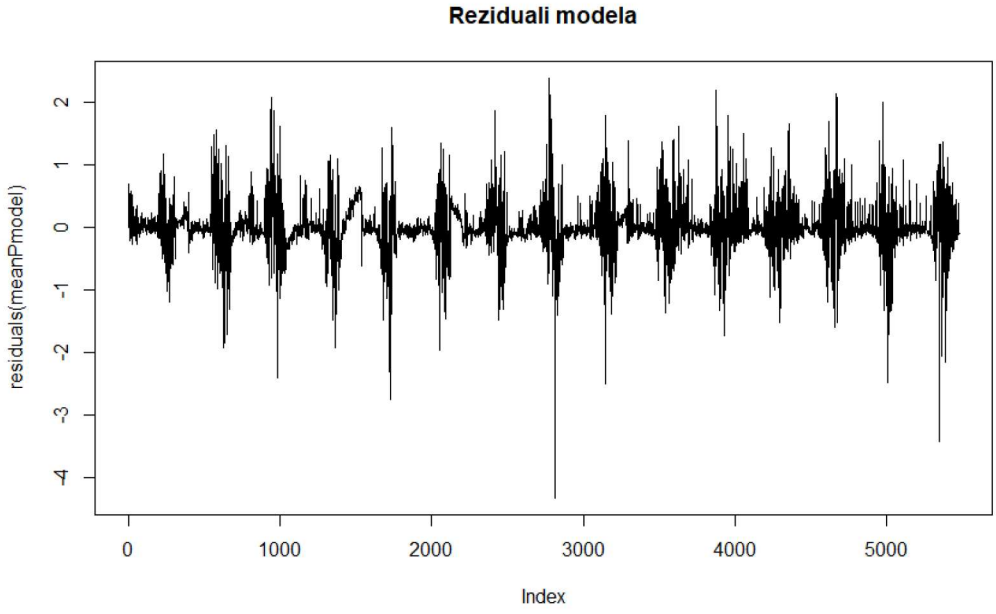
U prikazanom modelu koeficijent determinacije iznosi $R^2 = 0.9247$.

5.3 Pretpostavke modela

Kako bismo na temelju modela mogli donositi valjane zaključke, model bi trebao zadovoljavati klasične pretpostavke linearnih regresijskih modela s vremenskim nizovima.

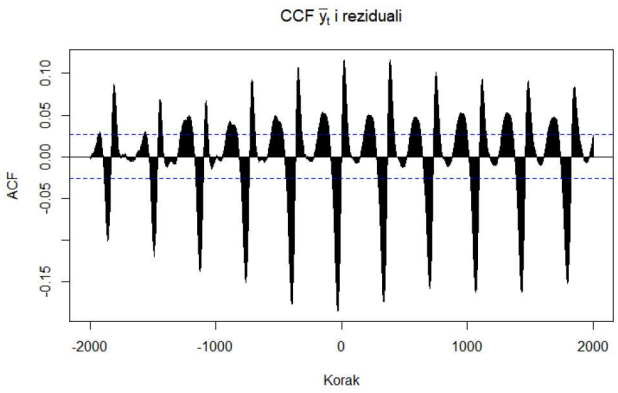
5.3.1 Analiza reziduala

Prisjetimo se, klasične pretpostavke linearnog regresijskog modela s vremenskim nizovima, sažete u Pretpostavci 4.2.6, govore nam da bi greške i prediktori trebali biti nezavisni te da bi greške trebale dolaziti iz niza nezavisnih normalnih slučajnih varijabli s očekivanjem 0 i varijancom σ^2, $\sigma > 0$. Stoga ćemo u nastavku provesti odgovarajuće testove. Kako je greška nemjerljiva veličina, o ispunjenosti tih pretpostavki zaključujemo na temelju niza procjena grešaka, tj. reziduala.

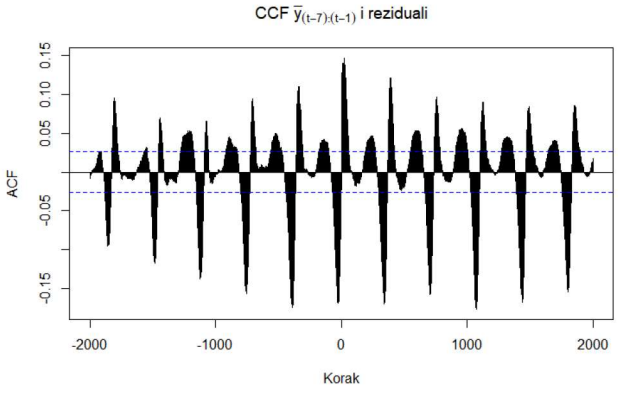


Slika 10: Trajektorija reziduala modela

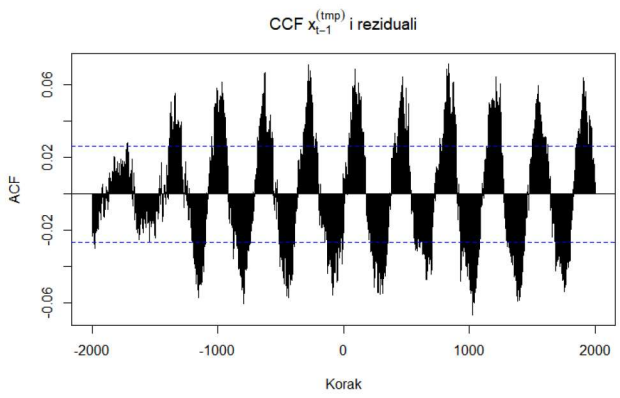
Već iz Slike 10 naslućujemo nestacionarnost procesa grešaka modela. U svrhu analize pretpostavki na greške modela pogledajmo prvo uzoračke kroskorelacijske funkcije reziduala i pojedinih ovisnih varijabli,



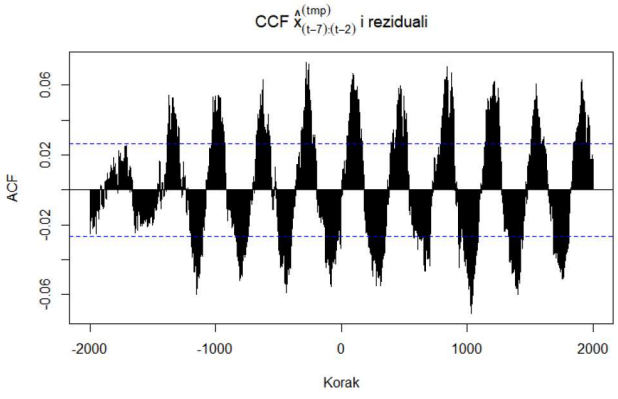
Slika 11: CCF reziduala i $\bar{y}_t$



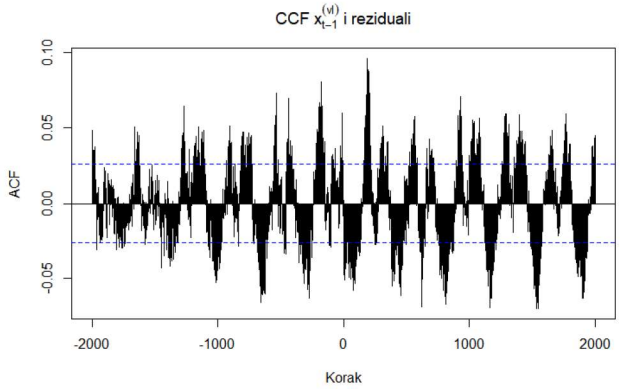
Slika 12: CCF reziduala i $\bar{y}_{(t-7):(t-1)}$



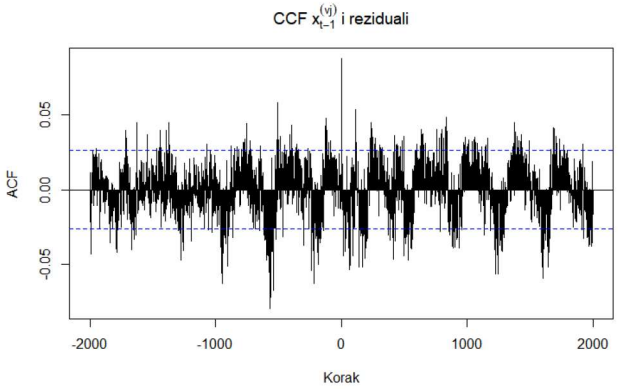
Slika 13: CCF reziduala i $x_{t-1}^{(tmp)}$



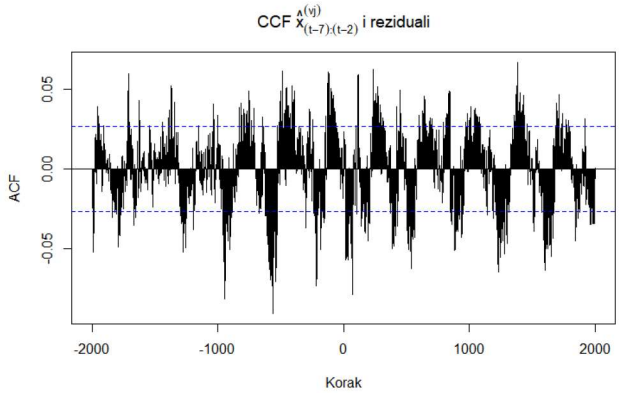
Slika 14: CCF reziduala i $\widehat{x}_{(t-7):(t-2)}^{(tmp)}$



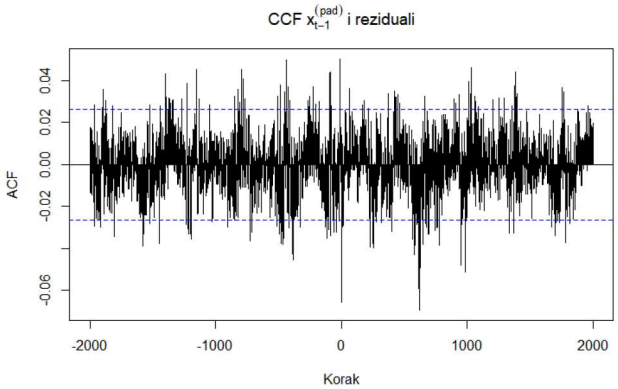
Slika 15: CCF reziduala i $x_{t-1}^{(vl)}$



Slika 16: CCF reziduala i $x_{t-1}^{(vj)}$

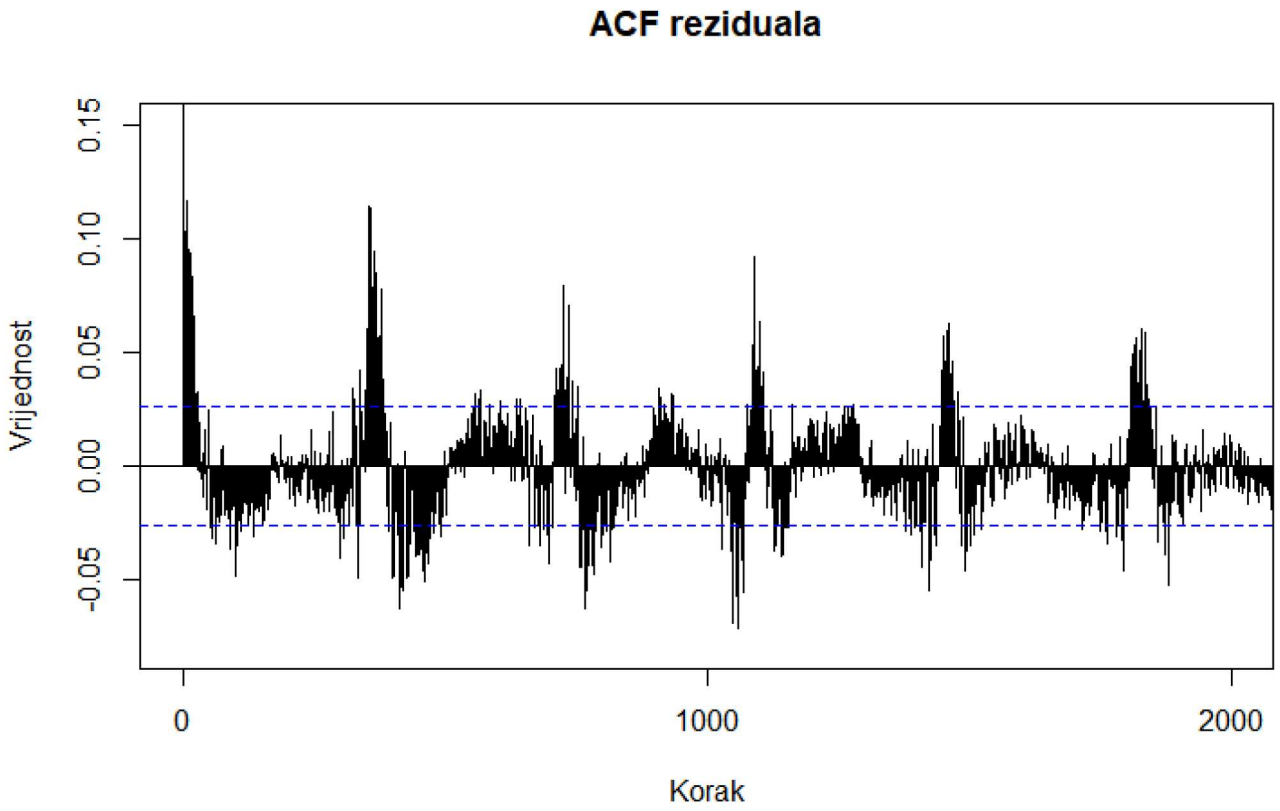


Slika 17: CCF reziduala i $\widehat{x}_{(t-7):(t-2)}^{(vj)}$



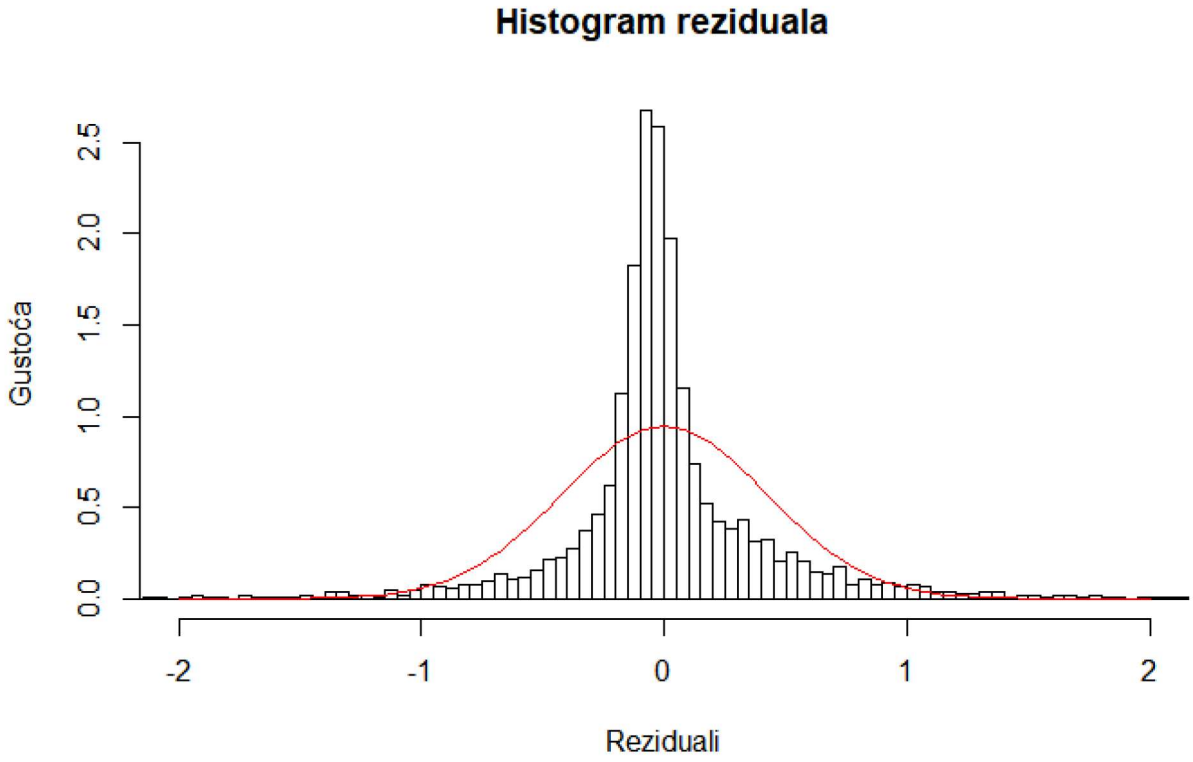
Slika 18: CCF reziduala i $x_{t-1}^{(pad)}$

Svi grafički prikazi iz Slika 11 – 18 sugeriraju da pretpostavka o nekoreliranosti greške s prediktorima nije zadovoljena. Nadalje, pogledajmo i autokorelacijsku funkciju reziduala.



Slika 19: Empirijska autokorelacijska funkcija reziduala

Također naslućujemo da pretpostavka o nekoreliranosti niza grešaka nije zadovoljena, a to ćemo i pokazati Breusch-Godfreyjevim (BG) testom. Dobivena p-vrijednost BG testa jest $p = 2.2 \times 10^{-16} < 0.05$, što znači da na razini značajnosti 0.05 odbacujemo nul-hipotezu i potvrđujemo serijsku koreliranost grešaka modela. Još ćemo dodatno provjeriti zadovoljavaju li niz grešaka pretpostavku o normalnosti distribucije.

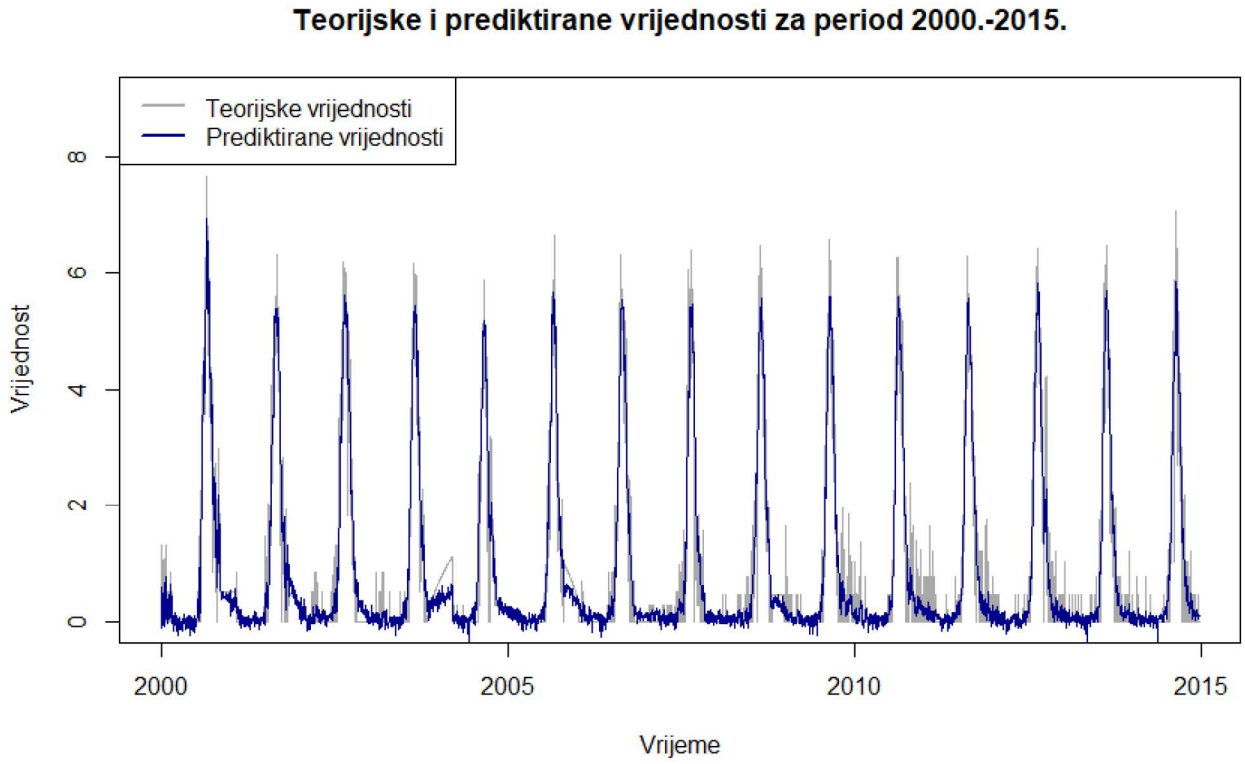


Slika 20: Histogram reziduala modela

Provođenjem Shapiro-Wilkovog testa [8] dobivamo p-vrijednost $p = 2.2 \times 10^{-16} < 0.05$, stoga na razini značajnosti 0.05 odbacujemo nul-hipotezu o normalnosti distribucije greške. I posljednje, htjeli bismo provjeriti zadovoljavaju li reziduali pretpostavku homoskedastičnosti. U tu svrhu koristimo Breusch-Paganov test [2] koji u nul-hipotezi podrazumijeva homoskedastičnost. Dobivena p-vrijednost toga testa iznosi $p = 2.2 \times 10^{-16} < 0.05$, stoga na razini značajnosti 0.05 odbacujemo nul-hipotezu i zaključujemo da su greške modela heteroskedastične.
Dakle, konačno možemo zaključiti da reziduali ne zadovoljavaju nijednu klasičnu pretpostavku linearnog modela zbog čega ovakav model ne bismo trebali upotrebljavati za predikcije. U ovom slučaju poboljšanje bi moglo biti u smislu modeliranja reziduala nekim stacionarnim procesom. Spomenuli smo neke od načina korekcije modela, među ostalim i Cochrane-Orcuttovu proceduru, no sve te korekcije uključuju modeliranje reziduala AR(1) procesom. S obzirom na to da reziduali našeg modela pokazuju autokoreliranost i na većem broju koraka, pokušali smo pronaći odgovarajući AR(q) proces kojim bi bilo moguće opisati reziduale. Procjenom AR(q) dobiven je red $q = 22$ što bi značilo da bismo reziduale mogli modelirati AR(22) procesom, no korekcije za takve slučaje nisu teoretski obrađene u ovom radu.

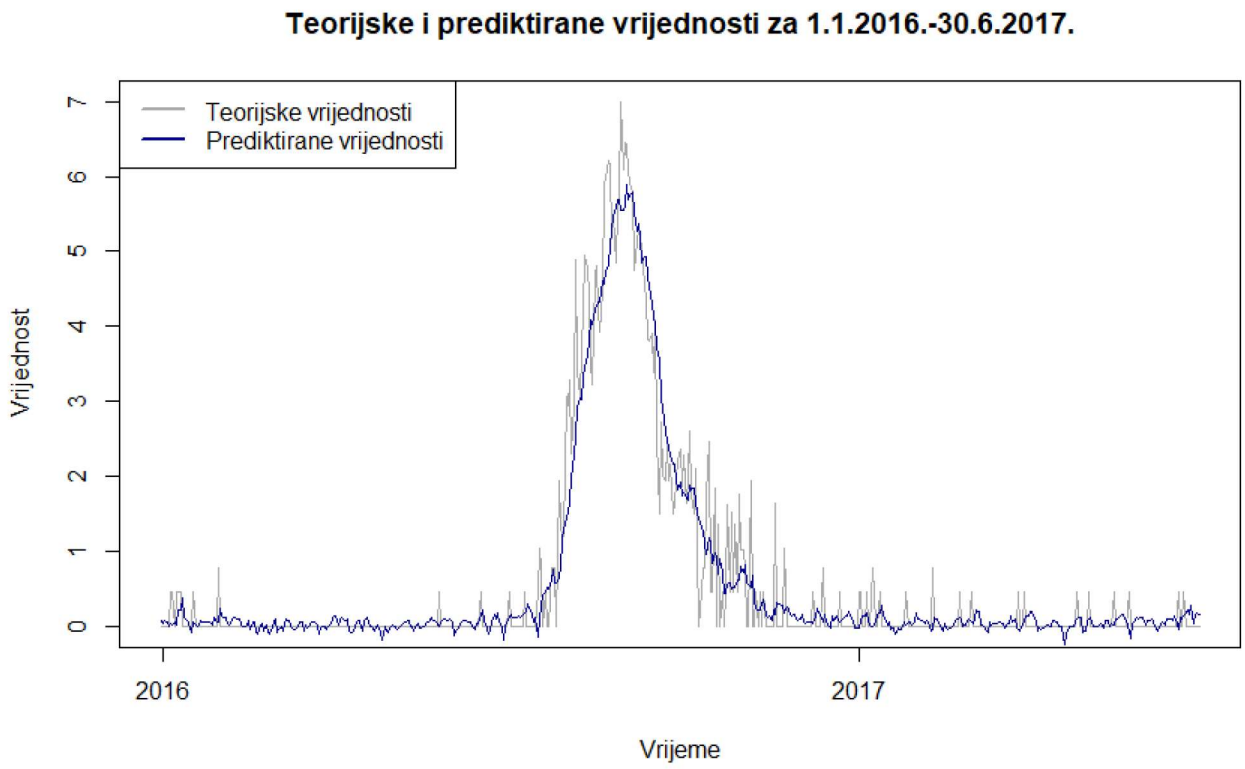
5.4 Predikcije modela

Prediktirane vrijednosti možemo dobiti uvrštavanjem dobivenih OLS procjenitelja u jednadžbu modela (5.1). Te prediktirane vrijednosti možemo usporediti sa stvarnim vrijednostima te vidjeti koliko dobro naš model pogađa log-transformirane prosječne dnevne koncentracije peludi na temelju danih meteoroloških podataka.



Slika 21: Trajektorije teorijskih i prediktiranih vrijednosti logaritmirane prosječne dnevne koncentracije peludi (2000.-2015.)

Na Slici 21 prikazane su teorijske i prediktirane vrijednosti u razdoblju koje smo odabrali za treniranje modela. Vidimo da se teorijske i prediktirane vrijednosti ne podudaraju najbolje, posebno u maksimumima. Pogledajmo sada i kako izgleda u razdoblju koje smo odvojili za validaciju:



Slika 22: Trajektorije teorijskih i prediktiranih vrijednosti logaritmirane prosječne dnevne koncentracije peludi (2016. i 2017.)

Iz Slike 22 također zaključujemo da model ne pogađa dobro visoke vrijednosti, odnosno u velikoj mjeri podcjenjuje stvarne vrijednosti. S obzirom na to da smo već pokazali da ne zadovoljava klasične linearne pretpostavke, zaključujemo da se nećemo pouzdati u predikcije, no promotrit ćemo klasifikacijsku moć modela za unaprijed definirane kategorije koncentracije peludi.

5.5 Kategoriziranje vrijednosti

Pokušajmo sada kategorizirati koncentracije peludi i pogledati kako se model ponaša u tom smislu.

5.5.1 Kategorizacija Nastavnog zavoda za javno zdravstvo "Dr. Andrija Štampar"

Za početak, pogledajmo kategorizaciju Nastavnog zavoda za javno zdravstvo "Dr. Andrija Štampar" [10]:

Razina peludi	Koncentracija peludi
Niska	$[0, 10]$
Umjerena	$(10, 50]$
Visoka	$(50, 500]$
Vrlo visoka	> 500

Tablica 7: Kategorizacija koncentracije peludi

U sljedećoj tablici prikazat ćemo koliko stvarnih mjerenja te prediktiranih vrijednosti pripadaju određenoj kategoriji, pri čemu su izostavljeni oni dani gdje je stvarna vrijednost mjerenja 0, odnosno dani izvan sezone.

Kategorija	% stvarnih	% prediktiranih
Niska	86.09	86.24
Umjerena	5.06	4.98
Visoka	8.27	8.67
Vrlo visoka	0.58	0.11

Tablica 8: Kategorizacija stvarnih i prediktiranih vrijednosti prosječne dnevne koncentracije peludi

Stvarne i prediktirane vrijednosti koncentracija za 95.12% mjerenja pripadaju istoj kategoriji. Zatim, za 4.71% mjerenja prediktirane vrijednosti od stvarnih se razlikuju samo za jednu kategoriju, dok se za dvije kategorije razlikuje 0.17% mjerenja. Na temelju toga možemo zaključiti da model u ovoj kategorizaciji jako dobro klasificira prediktirane koncentracije peludi.

5.5.2 Kategorizacija na temelju centila empirijske distribucije dugoročnih predikcija

Još jedna kategorizacija koju ćemo ispitati jest kategorizacija na temelju centila empirijske distribucije dugoročnih predikcija (doc.dr.sc. Nataša Krklec Jerinkić, PMF, Novi Sad). Ta podjela ima 10 kategorija, a postotke pripadnosti pojedinoj kategoriji za stvarne i prediktirane podatke možemo vidjeti u sljedećoj tablici:

Kategorija	% stvarnih	% prediktiranih
$(0, 4]$	60.69	60
$(4, 7]$	5.83	5.74
$(7, 8]$	1.06	1.28
$(8, 16]$	4.24	4.91
$(16, 29]$	3.51	4.13
$(29, 47]$	4.40	3.59
$(47, 79]$	4.69	4.99
$(79, 137]$	5.34	4.24
$(137, 274]$	5.87	9.17
> 274	4.37	0.96

Tablica 9: Kategorizacija stvarnih i prediktiranih vrijednosti prosječne dnevne koncentracije peludi

I u ovom slučaju zanima nas koliko dana (u postotku) smo predikcijom iz modela pogodili stvarnu kategoriju, a koliko smo dana pogriješili za jednu ili više kategorija. Kako bismo to dobili, izračunali smo razlike prediktirane i stvarne kategorije za svako mjerenje. Razlika može imati vrijednosti od -9 do 9. Pritom, -9 i 9 odnose se na slučajeve kad je prediktirana najveća (> 274), a u stvarnosti je najmanja kategorija ($(0, 4]$) i obrnuto, dakle pogriješili smo za 9 kategorija. Vrijednost 0 imaju ona mjerenja u kojima su i prediktirane i stvarne vrijednosti u istoj kategoriji.

razlika	%		razlika	%
-7	0.041		0	68.169
-5	0.206		1	11.148
-4	0.371		2	4.542
-3	1.404		3	1.197
-2	3.055		4	0.743
-1	9.042		5	0.082

Tablica 10: Razlike broja prediktiranih i broja stvarnih kategorija za prosječne dnevne koncentracije peludi

Iz Tablice 10 možemo vidjeti da je u 68.169% mjerenja prediktirana kategorija jednaka stvarnoj, odnosno možemo reći da je kategorija dobro pogođena. Ostale razlike možemo interpretirati na način da one s negativnim predznakom prikazuju postotak u kojem prediktirane vrijednosti precjenjuju stvarne, a one s pozitivnim predznakom prikazuju postotak onih u kojima prediktirane vrijednosti podcjenjuju stvarne. Pritom možemo reći da za ukupno 14.119% mjerenja prediktirane vrijednosti precjenjuju stvarne koncentracije, dok za 17.712 % mjerenja prediktirane vrijednosti podcjenjuju stvarne koncentracije.

Naš je model malo manje uspješan za tu kategorizaciju, no i dalje sa zadovoljavajućim postotkom pogađa kategoriju.
U konačnici možemo zaključiti da, unatoč tome što model ne zadovoljava pretpostavke i nije dobar za prediktiranje budućih vrijednosti, mogao bi se upotrebljavati za prediktiranje kategorije jer u slučaju kategorizacije iz [10] s vrlo velikim postotkom pogađa točno.

Literatura

[1] B. Bercu, F. Proia, *A sharp analysis on the asymptotic behavior of the Durbin-Watson statistic for the first-order autoregressive process*, ESAIM: Probability and Statistics, EDP Sciences, 2013, 17, pp.500-530. ff10.1051/ps/2012005ff. ffhal-00642634

[2] T. S. Breusch and A. R. Pagan, *A Simple Test for Heteroscedasticity and Random Coefficient Variation*, Econometrica Vol. 47, No. 5 (Sep., 1979), pp. 1287-1294

[3] P. J. Brockwell, R. A. Davis, *Introduction to Time Series and Forecasting*, Second Edition, Springer, Springer-Verlag, New York, 2002.

[4] D. Halcoussis, *Understanding Econometrics*, South-Western, Cengage Learning, Mason, OH, 2005.

[5] M. Verbeek, *A Guide to Modern Econometrics*, 2nd edition, John Wiley & Sons,Ltd, Chichester 2004.

[6] S. Karlin, H. M. Taylor, *A first course in stochastic processes*, Second Edition, Academic Press, Inc., New York, 1975.

[7] S. Karlin, H. M. Taylor, *A second course in stochastic processes*, Academic Press, Inc., New York, 1981.

[8] S. S. Shapiro, M. Wilk, *"An analysis of variance test for normality (complete samples)"*, Biometrika, Vol. 52, No. 3/4. (Dec., 1965), pp. 591-611.

[9] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 5th edition, South-Western, Cengage Learning, Mason, OH, 2012.

[10] *Nastavni zavod za javno zdravstvo "Dr. Andrija Štampar"*,
URL: `http://www.stampar.hr/hr/alergijski-semafor-peludna-prognoza-i-peludni-kalendar`

[11] J. Parker, Reed College,
URL: `https://www.reed.edu/economics/parker/312/tschapters/S13_Ch_2.pdf`

[12] Tony E. Smith,
URL: `https://www.seas.upenn.edu/~ese302/extra_mtls/Autocorrelation_Notes.pdf`

Sažetak i ključne riječi

Sažetak. U ovom radu predstavljeni su osnovni pojmovi iz teorije vremenskih nizova te multivarijatne regresijske analize. Nakon toga dana je teorijska osnova linearne regresije s vremenskim nizovima. Prvo su predstavljeni primjeri regresijskih modela s vremenskim nizovima, a zatim svojstva OLS procjenitelja i klasične pretpostavke linearnih modela s vremenskim nizovima. Nakon toga detaljnije je opisan jedan od glavnih problema u regresijskoj analizi vremenskih nizova, a to je serijska koreliranost grešaka modela. Navedeni su neki od načina testiranja serijske koreliranosti kao i moguće korekcije modela čije je greške moguće opisati nekim autoregresivnim modelom.

Primjena teorije i praktični dio rada proveden je na temelju podataka o koncentraciji peludi u Novom Sadu koji je za cilj imao modeliranje prosječne dnevne koncentracije peludi na temelju izmjerenih meteoroloških podataka kao što su temperatura, količina padalina, brzina vjetra i udio vlažnosti u zraku u formi vremenskih nizova. Napravljen je multivarijatni linearni regresijski model u kojem su ovisna varijabla i regresori vremenski nizovi. Na temelju analize reziduala pokazano je da model ne zadovoljava klasične pretpostavke linearnog modela te se ne treba koristiti za interpretaciju i predikciju budućih dnevnih vrijednosti prosječne dnevne koncentracije peludi. Ipak, ukoliko se vrijednosti koncentracije peludi kategoriziraju, model se pokazao kao vrlo uspješan u pogađanju kategorije.

Ključne riječi: vremenski niz, stacionarnost, AR model, linearna regresija, metoda najmanjih kvadrata, OLS procjenitelj, serijska korelacija, statički model, pelud, ambrozija, koncentracija

Summary and keywords

Summary. This master's thesis presents basic terms of time series analysis and regression analysis. With the theoretical basis of time series regression analysis, the examples of time series regression models are given as well as properties of OLS estimators and classical linear model assumptions for the time series applications. Afterwards, a serial correlation in the error terms, one of the main problems in time series regression, has been described in more detail. Different tests and correction methods for serial correlation are described for models, errors of which can be described using autoregressive process.

Practical application based on a data of pollen concentration has been performed with the aim of modeling average daily pollen concentration for the given meteorological data such as temperature, precipitation, wind speed and humidity in form of a time series. A multivariate linear regression model with time series dependent and regressor variables has been developed. It was shown based on the analysis of residuals that given model does not satisfy classical linear model assumptions, therefore it should not be used for interpretation and prediction purpose. However, if one categorizes pollen concentration levels, the model proved itself to be very successful in predicting a category.

Keywords: time series, stationarity, AR model, linear regression, ordinary least squares, OLS estimator, serial correlation, static model, pollen, ambrosia, concentration

Životopis

Rođena sam 10.05.1993. u Osijeku u Republici Hrvatskoj. U istom gradu pohađala sam Osnovnu školu Vladimira Becića, a zatim sam upisala i III. gimnaziju Osijek. Tijekom osnovnoškolskog i srednjoškolskog obrazovanja sudjelovala sam na brojnim općinskim i županijskim natjecanjima iz matematike.

2011. godine upisala sam Preddiplomski sveučilišni studij matematike na Odjelu za matematiku u Osijeku te sam 2014. godine stekla naziv prvostupnice matematike uz završni rad *Teorija kodiranja i linearni kodovi* pod mentorstvom izv. prof. dr. sc. Ivana Matića. Iste godine upisala sam Diplomski sveučilišni studij matematike, smjer Financijska matematika i statistika na Odjelu za matematiku.

Tijekom studija odradila sam dvije stručne prakse - u Hrvatskoj agenciji za hranu gdje sam radila statističku analizu prehrambenih navika hrvatskih građana te u sklopu Span Academy ljetne prakse tvrtke Span d.o.o. gdje sam radila na razvoju baze podataka. U trenutku pisanja rada zaposlena sam u Spanu na mjestu mlađeg razvojnog inženjera.