Epälineaariset aikasarjamallit: STAR-malli

Juho Toivonen

LuK-tutkielma
Toukokuu 2025

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:
Väitöskirjatutkija Roope Rihtamo

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu
Turnitin OriginalityCheck-järjestelmällä

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

JUHO TOIVONEN: Epälineaariset aikasarjamallit: STAR-malli
LuK-tutkielma, 22 s.
Tilastotiede
Toukokuu 2025

Tutkielma esittelee aikasarja-analyysiin liittyvää teoriaa lineaaristen ja epälineaaristen aikasarjamallien osalta. Käsittely etenee lineaaristen mallien teorian kautta kohti epälineaarisen STAR-mallin esittelyä kohti.

Aluksi tutkielmassa esitellään lineaarisen aikasarjamallin perusoletuksia sekä AR(p)-malli. Tämän jälkeen siirrytään epälineaarisen aikasarjamallien teoriaan ja STAR(p)-mallin käsittelyyn. Lopuksi esitellään vielä mallinnustehtävä, jossa epälineaarisen aikasarjamallin käyttö on perusteltua. Tehtävässä valituilla valintakriteereillä parhaaksi malliksi osoittautuu LSTAR(3)-malli kynnysmuuttujan viiveellä yksi.

Käytännön sovelluksen tuloksista huomataan, että kriteerien valinnalla on suuri merkitys saatuun lopputulokseen.

Avainsanat: aikasarjamallit, AR-malli, STAR-malli.

Sisällys

1 Johdanto

Aineistoja, joissa havainnot ovat riippuvia pelkästään ajasta, kutsutaan aikasarjoiksi. Kuten regressiomallien kohdallakin, yksinkertaisimmat aikasarjojen mallinnuskeinot ovat lineaarisia. Niiden etuna on yksinkertaisuus. Lineaariset aikasarjamallit eivät sovellu kuitenkaan kaikkiin mallintamisongelmiin.

Lineaaristen mallintamiskeinojen jatkeeksi on kehitelty useita erilaisia epälineaarisia malleja. Näistä ehkä lähimpänä lineaarisia malleja ovat regiiminmuutosmallit, joissa aineiston sisäinen vaihtelu jaetaan kahteen tai useampaan lineaariseen vaiheeseen. Näistä malleista yksinkertaisimmat ovat TAR- ja STAR-mallit, jotka ovat lineaarisen AR-mallin ensimmäiset epälineaariset vaihtoehdot. Tämä tutkielma keskittyy STAR-mallin tarkasteluun. Lisäksi tutkielmassa esitellään käytännön sovellus, jossa lineaarisen mallinnuksen perusoletukset eivät toteudu, jolloin on perusteltua käyttää epälineaarista mallia.

Tämän tutkielman tavoitteena on esitellä aikasarjamallinnuksen yleistä teoriaa ja syventää tätä epälineaaristen mallien suuntaan. Lineaaristen aikasarjamallien käytön kannalta tärkeää teoriaa käsitellään luvussa 2. Luku 3 syventää teoriaa epälineaaristen mallien käytön mahdollistamiseksi. Luvussa 4 esitellään käytännön sovellus, jossa epälineaarinen STAR-malli tuottaa paremman ratkaisun verrattuna lineaariseen AR-malliin. Lukijalta oletetaan aikasarja-analyysin lineaarisen teorian perustason ymmärrystä.

Tutkielman laatimisessa lineaarisia aikasarjamalleja käsittelevän osuuden osalta seurataan Ruey S. Tsayn teosta *Analysis of Financial Times Series*. Lisäksi apuna käytetään Henri Nybergin Aikasarja-analyysin suomenkielistä monistetta. Epälineaarisia aikasarjamalleja käsittelevän osuuden kohdalla tärkeimpänä lähteenä käytetään Philip Hans Fransesin ja Dick van Dijkin teosta *Non-Linear Time Series Models in Empirical Finance*.

2 Lineaariset aikasarjamallit

Tässä luvussa esitellään lyhyesti lineaaristen aikasarjamallien perusoletukset sekä tarkastellaan tutkielman kokonaisuuden kannalta merkittävän lineaarisen aikasarjaprosessin yleistä tapausta. Luvussa esitellään myös erikoistapaus, joka on tämän yleisen aikasarjaprosessin yksinkertaisin muoto. Tutkielmassa rajoitutaan yksiulotteisten aikasarjojen tarkasteluun. Esitellyt oletukset toimivat teoreettisena pohjana lineaaristen mallien lisäksi myös epälineaaristen aikasarjamallien määrittelyssä. Luvussa esitellään lyhyesti stokastisen prosessin periaate, stationaarisuus ja autokorrelaatiofunktio. Lineaarisista prosesseista esitellään AR(p)-prosessi sekä sen yhden viiveen erikoistapaus AR(1).

2.1 Stokastinen prosessi

Aikasarja-analyysin kontekstissa stokastinen prosessi on joukko satunnaismuuttujia $\{y_t; t = 0, \pm1, \pm2, ...\}$, jotka ovat riippuvia aikaindeksistä t. Merkintää y_t käytetään myös jo havaituista aikasarjan arvoista (Nyberg, 2020). Satunnaismuuttujat voivat sijaita eri pituisilla aikaväleillä, mutta yksinkertaistuksena oletetaan, että tämä aikaväli on vakio. Yksiulotteisen aikasarjan havaittuja satunnaismuuttujia merkitään yksinkertaisesti y_t, kun t = 1,2,..., n, jossa n on havaittujen muuttujien lukumäärä. Tämä osuus on kuitenkin ainoastaan havaittu realisaatio aikasarjasta, jolloin vaaditaan tiettyjä rajoittavia oletuksia, jotta prosessin ominaisuuksia on mahdollista selvittää [4].

2.2 Stationaarisuus ja korrelaatiofunktiot

Lineaaristen aikasarjamallien käyttöä varten stokastiselta prosessilta vaaditaan tiettyjä oletuksia. Stationaarisuus on näistä oletuksista tärkein. Sen ehto voidaan todentaa kahdella eri tasolla. Näistä vahvempi *vahva stationaarisuus* on voimassa, kun satunnaisvektoreilla $(y_{t_1}, ..., y_{t_m})$ ja $(y_{t_{1+h}}, ..., y_{t_{m+h}})$ on identtinen yhteisjakauma, joka on riippumaton h:stä. Tällöin vahvan stationaarisen prosessin on oltava aikainvariantti kaikilla momenteilla. [8]

Useimmiten kuitenkin heikompi stationaarisuuden ehto on riittävä aikasarjan mallintamiselle. Prosessi on *heikosti stationaarinen*, mikäli se toteuttaa seuraavat ehdot. Ensimmäisenä ehtona on, että prosessin ensimmäinen momentti eli odotusarvo on äärellinen

$$\mathbb{E}(y_t) = \mu, \ \ \forall\, t = 0, \pm1, \pm2, ...$$

ja $\mu < \infty$. Toisena ehtona on, että *autokovarianssifunktio*

$$Cov(y_t, y_{t+h}) = \gamma_{t,t+h} = \gamma_h, \ \ \forall\, t = 0, \pm1, \pm2, ...$$

on riippuvainen ainoastaan aikasarjan ajankohtien välisestä etäisyydestä h. Lisäksi heikon stationaarisuuden voimassaolo vaatii, että stokastisen prosessin toinen momentti, varianssi, on äärellinen ja aikainvariantti eli $Var(y_t) < \infty$ [4].

Autokovarianssifunktion sijaan käytännössä hyödynnetään *autokorrelaatiofunktiota* (ACF). Heikon stationaarisuuden vallitessa autokorrelaatiofunktio määritellään

$$\rho_h = Cor(y_t, y_{t+h}) = \frac{Cov(y_t, y_{t+h})}{Var(y_t)} = \frac{\gamma_h}{\gamma_0}.$$

Autokorrelaatiokertoimilla ρ_h ovat seuraavat ominaisuudet:

$$\rho_0 = 1, \quad |\rho_h| \leq 1 \quad \text{ja} \quad \rho_h = \rho_{-h}.$$

Lisäksi voidaan yleisesti olettaa, että $\gamma_h \to 0$, kun $h \to \infty$. Toisin sanoen kun h on suuri, satunnaismuuttujat y_t ja y_{t+h} ovat ajallisesti kaukana toisistaan ja täten lähes korreloimattomat [4].

Autokorrelaatiofunktion lisäksi käytännön toteutuksissa hyödynnetään osittaisautokorrelaatiofunktiota. Funktion arvot saadaan hyödyntämällä Yule–Walkerin yhtälöitä. Funktio voidaan määritellä yhtälöllä

$$\alpha_h = \begin{cases} 1, & \text{kun } h = 0 \\ \Gamma_h^{-1}\gamma_h\text{:n viimeinen komponentti, kun } h \geq 1, \end{cases}$$

jossa $\gamma_h = [\gamma_1, ..., \gamma_h]'$ ja $\Gamma_h = [\gamma_{i-j}]_{i,j=1,...,p}$. Yleisellä tasolla voidaan osoittaa, että osittaisautokorrelaatiofunktio on identtinen osittaiskorrelaatiokertoimien kanssa, jolloin se mittaa muuttujien y_t ja y_{t-h} välisen korrelaation suuruuden, kun muuttujien $y_{t-1}, ..., y_{h+1}$ lineaarinen vaikutus on poistettu [4]. Otosautokorrelaatiofunktion ja osittaisautokorrelaatiofunktion kuvaajista on mahdollista määritellä mallintamistehtävään sopiva aikasarjaprosessi etsimällä kuvaajan korrelaatioarvoista katkoksia.

2.3 Valkoinen kohina ja yleinen lineaarinen malli

Valkoiseksi kohinaksi kutsutaan aikasarjaa y_t, jonka yksittäiset satunnaismuuttujat ovat riippumattomia ja identtisesti jakautuneita (*independent and indetically distributed*, iid-oletus) sekä tuottavat äärellisen odotusarvon ja varianssin. Tälle on ominaista, että autokorrelaatiofunktion jokainen arvo on nolla. Käytännössä mikäli aineiston ACF-arvot ovat lähellä nollaa, pidetään aikasarjaa valkoisena kohinana [8].

Aikasarjaa y_t voidaan pitää lineaarisena, mikäli se voidaan kirjoittaa muodossa

$$y_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \tag{1}$$

jossa μ on prosessin y_t keskiarvo, $\psi_0 = 1$ ja $\{\epsilon_t\}$ täyttää normaalisen valkoisen kohinan piirteet. Normaalinen valkoinen kohina toteuttaa iid-oletuksen ja noudattaa normaalijakaumaa odotusarvolla nolla ja varianssilla σ^2. Merkinnästä ϵ_t käytetään useimmiten nimitystä shokki tai virhetermi ja se on aikasarjan ei-havaittu osa [8]. Usein virhetermille asetetaan jakaumaoletus $\epsilon_t \sim nid(0, \sigma^2)$ eli virhetermi on normaalisti ja identtisesti jakautunut. Virhetermi ϵ_t sisältää kaiken selittämättömän satunnaisvaihtelun aikasarjassa.

Parametrien ψ_j, $j = 1, ..., \infty$ arvot toimivat kertoimina ja toimivat prosessin y_t painoina. Jotta yhtälöä voidaan käyttää, täytyy kertoimien ψ_j neliöiden olla äärellisiä. Toisin sanoen

$$\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty. \tag{2}$$

Heikon stationaarisuuden vallitessa on mahdollista määritellä prosessin y_t odotusarvo ja varianssi

$$\mathbb{E}(y_t) = \mu \quad \text{ja} \quad Var(y_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2, \tag{3}$$

missä σ^2 on virhetermin ϵ_t varianssi. Autokovarianssi viiveelle h on [8]

$$\gamma_h = Cov(y_t, y_{t+h}) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}.$$

Edellä esiteltyä yhtälöä (1) kutsutaan yleiseksi lineaariseksi prosessiksi tai malliksi. Siitä käytetään myös nimitystä kausaalinen lineaarinen prosessi, koska se ei riipu tulevien virhetermien ϵ_t arvoista. Tällaista yleistä lineaarista prosessia käytetään ainoastaan teoreettisena apuvälineenä. Kaikki käytännössä käytettävät lineaariset prosessit tai mallit ovat tämän yleisen version erikoistapauksia [4]. Kausaaliselle lineaariselle prosessille löytyy myös tietyin rajoituksin muodostettava ei-kausaalinen prosessi. Ohitetaan kuitenkin tämän prosessin tarkempi tarkastelu.

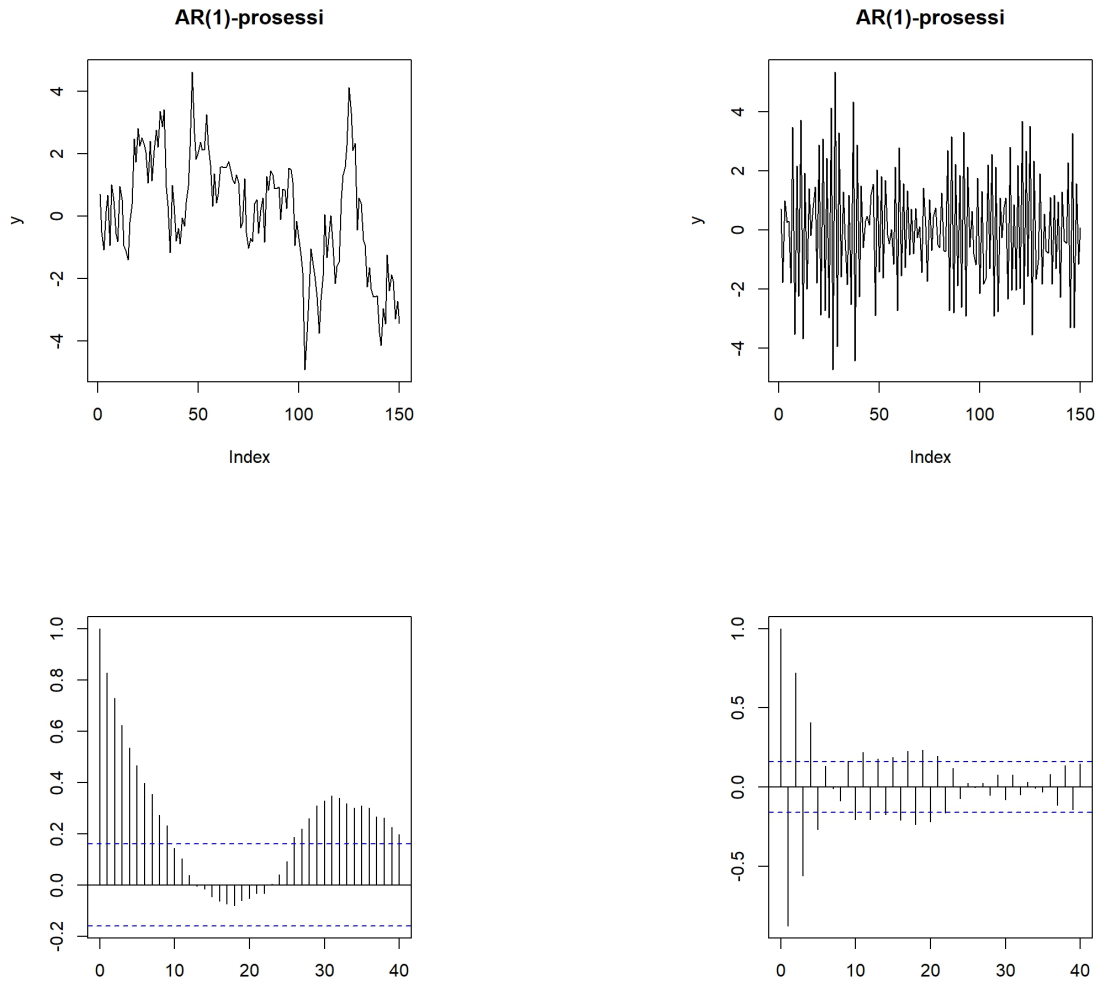
2.4 AR(1)-prosessi

Eräs yksinkertainen yleisen lineaarisen prosessin erikoistapaus on *ensimmäisen asteen autoregressiivinen prosessi*, AR(1). Prosessissa oletetaan, että $\psi_j = \phi^j$, jolloin oletuksen (2) täyttyminen vaatii rajoitusta $|\phi| < 1$. AR(1)-prosessi on sekä heikosti että vahvasti stationaarinen. Stationaarisuus on voimassa myös, kun $|\phi| > 1$, mutta tällöin prosessia ei ole mahdollista kirjoittaa yhtälön (1) muotoon. Tällöin kyse on luvussa 2.3 mainitusta ei-kausaalisesta prosessista.

Edellä tehdyn rajoituksen vallitessa prosessi voidaan kirjoittaa

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma^2). \tag{4}$$

Tulkinta on, että nykyinen arvo riippuu edellisestä havaitusta arvosta sekä virhetermistä [4]. Lisäksi mukana on vakiotermi ϕ_0, joka tässä yksinkertaisessa erikoistapauksessa oletetaan useimmiten nollaksi. Vakion arvo nolla saavutetaan keskistämällä tarkastelun alla oleva aikasarja. Tässä muodossaan AR(1)-prosessi muistuttaa lineaarista mallia. Vaaditut



Kuva 1: Kahden simuloidun AR(1)-prosessin kuvaajat ja alla niiden otosautokorrelaatiofunktiot (n = 150). Vasemmalla $\phi_1 = 0,9$, oikealla $\phi_1 = -0,9$.

momentit, odotusarvo ja varianssi, on mahdollista selvittää yleisten tulosten (3) perusteella. Odotusarvo on nolla ja varianssi

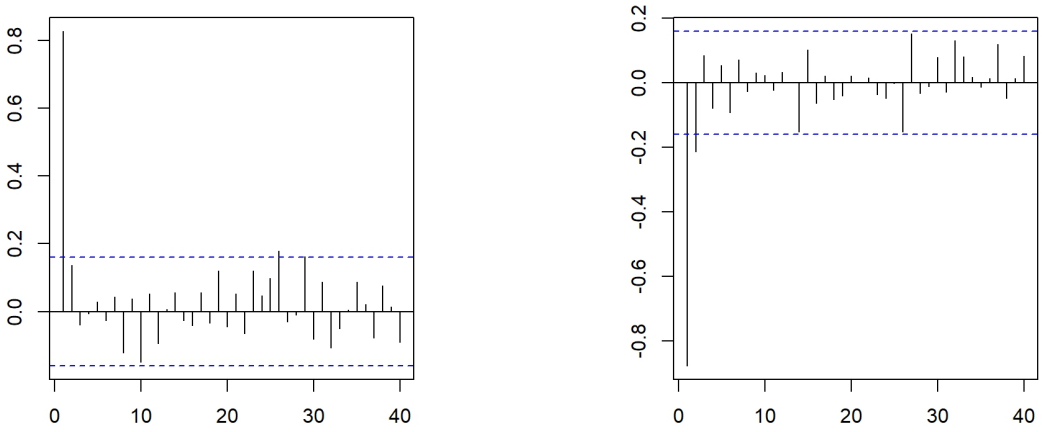
$$Var(y_t) = \sigma^2 \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma^2}{1-\phi^2}.$$

Kovarianssifunktion lauseke saadaan muotoon

$$Cov(y_t, y_{t+h}) = \sigma^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h} = \frac{\sigma^2 \phi^h}{1-\phi^2},$$

josta voidaan johtaa autokorrelaatiofunktio

$$\rho_h = \begin{cases} 1, & h = 0 \\ \phi^h, & h > 0. \end{cases}$$



Kuva 2: Kuvan 1 AR(1)-prosessien osittaisautokorrelaatiofunktiot.

Kuvassa 1 on kaksi simuloitua aineistoa AR(1)-prosessin pohjalta. Prosessi on muotoa (4). Vakion ϕ_0 arvo on nolla. Kuvassa vasemmalla ylhäällä on parametrin $\phi_1 = 0.9$ arvolla simuloidun aineiston kuvaaja ja tämän alapuolella tämän otosautokorrelaatiofunktio. Oikealla löytyvät vastaavat kuvaajat parametrin $\phi_1 = -0.9$ arvolla simuloituna. Kun parametrin arvo on lähellä yhtä tai miinus yhtä, aineisto on vahvasti korreloitunut. Positiivisella arvolla korrelaatio pienenee askelittain mitä isompi viiveiden määrä on. Korrelaatiokuvaajassa on kuitenkin havaittavissa aaltomaista liikettä, joka sekin vähitellen vaimenee kohti nollaa. Kun parametrilla on negatiivinen arvo perättäiset korrelaatiot vaihtavat etumerkkiä. Korrelaatio kuitenkin pienenee myös tässä tapauksessa viiveen suurentuessa.

2.5 AR(p)-prosessi

AR(p)-prosessi on edellisessä alaluvussa esitellyn AR(1)-prosessin yleistys.

Määritelmä 1. Määritellään AR(p)-prosessin yhtälö

$$y_t = \phi_0 + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t, \tag{5}$$

jossa ϕ_0 yhtälön vakiotermi, $\phi_1, \ldots, \phi_p$ on prosessin tuntemattomat parametrit, $y_{t-1}, \ldots, y_{t-p}$ aikasarjan edelliset havainnot sekä ϵ_t mallin virhetermi. Virhetermille pätee oletukset $\mathbb{E}(\epsilon_t) = 0$ ja $Var(\epsilon_t) = \sigma^2$. Lisäksi virhetermillä on iid-oletus.

Tätä prosessia kutsutaan *p:nnen asteen autoregressiiviseksi prosessiksi*, AR(p). Tämän tulkinta on, että havainto y_t on riippuvainen aikasarjan p:stä aiemmasta havainnosta sekä satunnaisuutta tuovasta ei-havaittavasta virhetermistä. Mikäli prosessia ei keskistetä, mukana on myös vakiotermi. AR(p)-prosessi on stationaarinen, mikäli polynomin $\phi(z)$ ($z \in \mathbb{C}$) juuret ovat kompleksitasoisen yksikköympyrän kehän ulkopuolella. Tämän tarkempi määrittely ohitetaan. (katso Nyberg, 2020) [4] . Stationaarisuusehdon toteutuminen ei kuitenkaan ole useimmiten tarpeellinen. Kausaalisten prosessien tarkastelussa ehto on kuitenkin välttämätön. Useimmissa tapauksissa luvussa 2.2 esitellyn heikon stationaarisuuden oletukset riittävät AR(p)-prosessin käyttöön aikasarjan mallinnustehtävissä.

Kuvassa 2 on esitelty kuvan 1 AR(1)-prosessista simuloitujen aineistojen osittaisautokorrelaatiofunktiot. Näistä kuvaajista on mahdollista tunnistaa aineiston mallintamiseen sopivan autoregressiivisen mallin viiveiden lukumäärä p. Kuten aiemmin on huomattu, tämä vaikuttaa siihen, miten montaa aikasarjan edellistä havaintoa hyödynnetään aineiston seuraavan askeleen mallintamiseen. Kuvaajasta pyritään etsimään katkoksia, joiden avulla on voidaan määritellä käyttökelpoinen prosessi mallintamiseen. Tämä ei kuitenkaan ole aina täysin yksiselittäistä. Kuvan 2 osittaiskorrelaatiofunktioiden kuvaajissa voidaan nähdä katkokset ensimmäisen viiveen kohdalla, kuten AR(1)-prosessista simuloitujen aineistojen kohdalla voidaan olettaakin.

Sekä otosautokorrelaatiofunktion että osittaisautokorrelaatiofunktion kuvaajissa on huomattavissa kaksi vaakasuoraa katkoviivaa. Nämä ovat etukäteen määriteltyjä kriittisiä pisteitä, joiden väliin mahtuu 95 prosenttia funktion estimaateista. Kriitiset rajat on määritelty kaavalla $\pm 1.96/\sqrt{n}$. Nämä kriittiset rajat voivat olla yksi apukeino katkosten löytämisessä.

3 Epälineaariset aikasarjamallit

Käytännön sovelluksissa löytyy kuitenkin useita sellaisia mallinnustehtäviä, joihin lineaarinen aikasarjamalli ei tuo toivotun tarkkaa tulosta. Tämä johtuu useimmiten siitä, että tutkittava aikasarja-aineisto ei toteuta lineaarisen mallin käytön mahdollistavia stationaarisuusoletuksia. Lineaarisen mallin heikkouksia korvaamaan on luotu useita epälineaarisia mallinnustapoja.

Tässä luvussa käsitellään epälineaarisen aikasarjan määritelmä ja käydään läpi regiimin-muutosmallien teoriaa. Malleista esitellään TAR- ja STAR-malli. Näistä tarkemman syvennytään STAR-malliin. Lineaarisen aikasarjateorian mukaisesti pitäydytään yksiulotteisten aikasarjojen tarkastelussa. Luvun päälähteenä käytetään Franses ja van Dijkin *Non-Linear Time Series Models in Empirical Finance* teosta [1] ellei lähdeviitteissä toisin mainita.

3.1 Epälineaarisen aikasarjan määrittely

Aiemmin mainittiin, kuinka stokastinen prosessi y_t on lineaarinen, jos se voidaan kirjoittaa muotoon (1) tiettyjen ehtojen ollessa voimassa. Kaikkia niitä stokastisia prosesseja, jotka eivät täytä lineaarisen stokastisen prosessin oletuksia, kutsutaan epälineaarisiksi. Mikä tahansa epälineaarisuus prosessissa johtaa epälineaariseen malliin. Yleisen epälineaarisen mallin suora käyttö on liiallisen parametrien määrän takia kuitenkin käytännössä

mahdotonta [8].

Jotta on mahdollista lähestyä epälineaarista mallia teoreettisesti, kirjoitetaan malli y_t:lle ehdollisten momenttien kautta. Olkoon F_{t-1} saatavilla olevan informaatiojoukko, joka koostuu satunnaismuuttujien $\{y_{t-1}, y_{t-2}, ...\}$ ja $\{\epsilon_{t-1}, \epsilon_{t-2}, ...\}$ lineaarikombinaatioista. Ehdollinen odotusarvo ja varianssi voidaan nyt kirjoittaa muotoon

$$\mu_t = \mathbb{E}(y_t \mid F_{t-1}) \equiv g(F_{t-1})$$
$$\sigma^2 = Var(y_t \mid F_{t-1}) \equiv h(F_{t-1}),$$

joissa g(.) ja h(.) on selkeästi määriteltyjä funktioita ja h(.) > 0. Täten mallia voidaan rajoittaa seuraavasti

$$y_t = g(F_{t-1}) + \sqrt{h(F_{t-1})}a_t,$$

missä $a_t = \epsilon_t/\sigma_t$ on standardoitu virhetermi. Lineaariselle prosessille y_t g(.) on F_{t-1} lineaarisen funktion elementtien osa ja h(.) = σ^2 [8].

3.2 Regiiminmuutosmallit

Regiiminmuutosmalleissa ajatellaan, että aikasarjassa on eri vaiheita eli regiimejä, joilla on toisistaan erilaiset ominaisuudet, kuten keskiarvo, varianssi ja/tai autokorrelaatio. Nämä voidaan jaotella kahteen kategoriaan: regiimit voidaan päätellä joko havaittujen muuttujien tai ei-havaittujen muuttujien avulla.

Tässä työssä tarkastellaan havaittuihin muuttujiin perustuvia malleja. Lisäksi rajoitetaan regiimien lukumäärää kahteen kuitenkin niin, että erillisissä regiimeissä olevat AR(p)-mallit saavat estimoinnissa parametriensa arvoikseen erilaiset arvot. Lisäksi odotusarvofunktiot ovat regiimikohtaisia. Tällä tavoin erottaudutaan lineaaristen mallien oletuksista.

3.2.1 TAR- ja STAR-malli

Ensimmäinen yksinkertaisesti muodostettava epälineaarinen malli, joka hyödyntää aiemmin esiteltyä AR(p)-prosessien teoriaa, on autoregressiivinen kynnysmalli (*Threshold Autoregressive model*, TAR). TAR-mallin perusajatuksena on, että aikasarjan aikana on yhtä useampi regiimi, joita on kuitenkin mahdollista mallintaa AR(p)-prosessien avulla. Sen esittelivät ensimmäisenä Tong(1978)[6] sekä Tong ja Lim(1980)[7]. TAR-malli olettaa, että regiimi on määritelty kynnysmuuttujan q_t ja kynnysarvon c välisen suhteen perusteella. Tämän erikoistapaus on, kun valitaan *kynnysmuuttujan* q_t arvoksi aikasarjan viive, jolloin $q_t = y_{t-d}$ mille tahansa kokonaisluvulle $d > 0$. Tässä tapauksessa regiimi on määritelty aikasarjan itsensä mukaan, jolloin mallia kutsutaan itsekehkeytyväksi autoregreesiiviseksi kynnysmalliksi (*Self-Exciting Threshold Autoregressive model, SETAR*).

Esitellään SETAR-malli yksinkertaisen esimerkin kautta. Olkoon $d = 1$, regiimien lukumäärä aiemmin rajattu kaksi sekä oletetaan molemmille regiimeille sama pohjamalli AR(1). Tällöin SETAR-malli voidaan kirjoittaa

$$y_t = \begin{cases} \phi_{0,1} + \phi_{1,1} y_{t-1} + \epsilon_t & \text{jos } y_{t-1} \leq c, \\ \phi_{0,2} + \phi_{1,2} y_{t-1} + \epsilon_t & \text{jos } y_{t-1} > c, \end{cases} \tag{6}$$

jossa virhetermille ϵ_t pätee iid-oletus, jonka informaatiojoukon suhteen ehdollinen odotusarvo on 0 ja ehdollinen varianssi σ^2. Vaihtoehtoinen kirjoitustapa SETAR-mallille on

$$y_t = (\phi_{0,1} + \phi_{1,1} y_{t\text{-}1})(1 - I[y_{t\text{-}1} > c]) \tag{7}$$

$$+(\phi_{0,2} + \phi_{1,2} y_{t\text{-}1}) I[y_{t\text{-}1} > c] + \epsilon_t \tag{8}$$

jossa I[A] on indikaattorifunktio arvoilla I[A] = 1, jos tapahtuma A esiintyy ja I[A] = 0 muutoin.

SETAR-malli olettaa, että regiimien välissä on jokin tietty kynnysmuuttujan arvo $y_{t\text{-}1}$. SETAR-mallissa regiimi vaihtuu, kun kynnysmuuttujan arvo ylittää kynnysparametrin c. Sen heikkous on, että ehdollisen odotusarvon yhtälö ei ole jatkuva. SETAR-malli ei myöskään ole välttämättä jatkuvaa koko aikasarjan ajan, vaan regiimin muututtua uuden regiimin mallintaminen ei ole sidottuna edellisen regiimin mallinnuksen päätepisteeseen. Mikäli mallinnuksen halutaan olevan jatkuva koko aikasarja-aineiston ajan tarvitaan vaihtoehtoinen indikaattorifunktion valinta.

Yksi vaihtoehto on valita sellainen indikaattorifunktio $G(y_{t\text{-}1}; \gamma, c)$, joka on jatkuva y_{t-1}:n suhteen. Kun indikaattorifunktio vaihdetaan, tuloksena saatavaa mallia kutsutaan tasaisen siirtymän autoregressiiviseksi malliksi (*Smooth Transition Autoregressive model*, STAR). Malli voidaan kirjoittaa

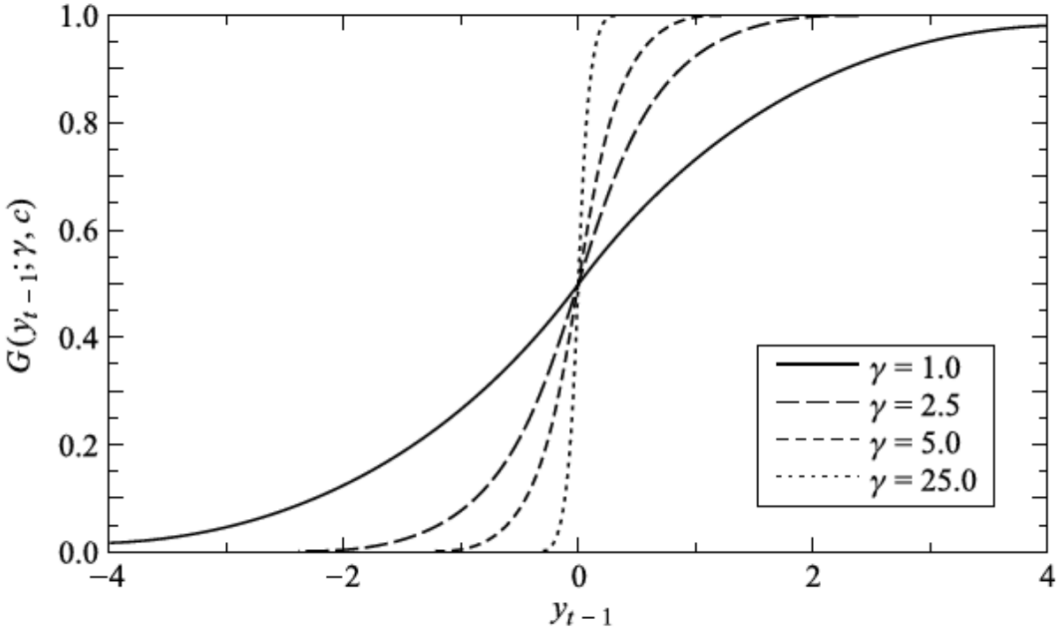
$$y_t = (\phi_{0,1} + \phi_{1,1} y_{t\text{-}1})(1 - G(y_{t\text{-}1}; \gamma, c)) \tag{9}$$

$$+(\phi_{0,2} + \phi_{1,2} y_{t\text{-}1}) G(y_{t\text{-}1}; \gamma, c) + \epsilon_t. \tag{10}$$

Suosittu valinta indikaattorifunktioksi $G(y_{t\text{-}1}; \gamma, c)$ on logistinen funktio

$$G(y_{t\text{-}1}; \gamma, c) = \frac{1}{1 + exp(-\gamma[y_{t\text{-}1} - c])} \tag{11}$$

ja tuloksena saatavaa mallia kutsutaan tällöin logistiseksi STAR-malliksi (LSTAR). Nyt indikaattorifunktio on jatkuva ja se saa arvoja nollan ja ykkösen välillä. LSTAR-mallissa regiimien vaihtuminen tapahtuu vähitellen. Tämä erottaa LSTAR-mallin SETAR-mallista ja malliin ei muodostu regiimien välille katkoksia. Mallin parametri γ määrittää, kuinka tasainen muutos siirtymäfunktion arvojen nolla ja yksi välillä on ja vaikuttaa myös regiimien vaihteluun. Kuvassa 3 on havainnollistettu parametrin γ eri arvoilla tapahtuvaa regiiminmuutoksen nopeutta. Suurilla γ:n arvoilla logistisen funktio alkaa muistuttaa SETAR-mallin indikaattorifunktiota I[A] ja kun $\gamma = 0$, STAR-malli redusoituu lineaariseksi malliksi.



Kuva 3: Esimerkkejä logistisen funktion $G(y_{t-1};\gamma,c)$ vaihteluvoimakkuudesta parametrin γ eri arvoilla, kun kynnysarvo c = 0. Kuva lainattu suoraan Franses ja van Dijkin kirjasta Non-Linear Time Series Models in Empirical Finance.

SETAR- ja STAR-malleilla on myös yleinen muoto, joka käsittää myös näiden korkeampiasteiset variaatiot. Ala- ja yläregiimin parametrien määrä saattaa olla samassa mallissa erisuuret. Merkitään tästä johtuen p1 ala- ja p2 yläregiimin parametrien määriksi. Nyt SETAR-malli voidaan esittää seuraavasti

$$y_t = \begin{cases} \phi_{0,1} + \phi_{1,1}y_{t-1} + ... + \phi_{p1,1}y_{t-p1} + \epsilon_t & \text{jos } y_{t-1} \leq c, \\ \phi_{0,2} + \phi_{1,2}y_{t-1} + ... + \phi_{p2,2}y_{t-p2} + \epsilon_t & \text{jos } y_{t-1} > c \end{cases}$$

ja STAR-malli seuraavasti

$$y_t = (\phi_{0,1} + \phi_{1,1}y_{t-1} + ... + \phi_{p1,1}y_{t-p1})(1 - G(y_{t-1};\gamma,c)) \quad (12)$$
$$+(\phi_{0,2} + \phi_{1,2}y_{t-1} + ... + \phi_{p2,2}y_{t-p1})G(y_{t-1};\gamma,c) + \epsilon_t. \quad (13)$$

STAR-mallin etu verrattuna SETAR-malliin on, että sen ehdollisen odotusarvon funktio on jatkuva. Tulevissa kappaleissa ohitetaan SETAR-mallin tarkempi tarkastelu ja keskitytään erityisesti STAR-mallin estimointiin, testaamiseen ja diagnostiikkaan.

3.2.2 STAR-mallin estimointi

STAR-mallin parametrien estimoinnissa on mahdollista käyttää epälineaarista pienimmän neliösumman (*nonlinear least squares*, NLS) menetelmää kohtuullisen suoraviivaisesti. Parametrit $\theta = (\phi_1', \phi_2', \gamma, c)$ voidaan estimoida

$$\hat{\theta} = argmin\, Qn\,(\theta) = argmin \sum_{t=1}^{n} [y_t - F(x_t;\theta)]^2, \tag{14}$$

jossa $F(x_t;\theta)$ on mallin runko, joka on

$$F(x_t;\theta) \equiv \phi_1' x_t (1 - G(y_{t-1};\gamma,c)) + \phi_2' x_t G(y_{t-1};\gamma;c), \tag{15}$$

missä $x_t = (1,\ y_{t-1}, ..., y_{t-p})$. Mikäli lisäoletuksena on, että virhetermi ϵ_t on normaalisti jakautunut, NLS vastaa suurimman uskottavuuden estimointia. Muussa tapauksessa NLS tulkitaan kvasi-suurimman uskottavuuden estimointina.

Estimoinnin voi suorittaa minkä tahansa tavanomaisen epälineaarisen optimoinnin menetelmän avulla. Tärkeimmät huomiota vaativat kiintopisteet ovat alkuarvojen valinta optimointialgoritmille, neliösummafunktion keskittäminen ja indikaattorifunktionfunktion tasaisuusparametrin γ estimointi.

Sopivien alkuarvojen löytäminen optimointia varten saattaa olla haastavaa. Kun parametreille γ ja c valitaan vakioarvot, STAR-malli on lineaarinen parametrien ϕ_1 ja ϕ_2 suhteen. Näin ollen kiinnittämällä γ ja c voidaan parametrit $\phi = (\phi_1', \phi_2')$ estimoida pienimmän neliösumman menetelmällä (*ordinary least squares*, OLS)

$$\hat{\phi}(\gamma,c) = \left(\sum_{t=1}^{n} x_t(\gamma,c) x_t(\gamma,c)' \right)^{-1} \left(\sum_{t=1}^{n} x_t(\gamma,c) y_t \right), \tag{16}$$

missä $x_t(\gamma,c) = (x_t'(1 - G(y_{t-1};\gamma,c)), x_t'(G(y_{t-1};\gamma,c))'$. Merkintää $\phi(\gamma,c)$ käytetään osoittamaan, että estimaatti ϕ on ehdollinen parametrien γ ja c suhteen. Residuaalit saadaan laskettua $\hat{\epsilon}_t = y_t - \hat{\phi}(\gamma,c)' c_t(\gamma,c)$ ja niiden varianssi $\hat{\sigma}^2(\gamma,c) = n^{-1} \sum_{t=1}^{n} \hat{\epsilon}_t^2(\gamma,c)$. Hyvä tapa valita järkevät alkuarvot epälineaariseen optimointialgoritmiin on suorittaa kaksiulotteinen ruudukkohaku parametrien arvoille γ ja c ja valita arvot, joilla tuloksena on pienin mahdollinen residuaalivarianssi.

Toinen tapa yksinkertaistaa estimointiongelmaa on tarkastella neliösummafunktiota [2]. Kiinnitetyillä parametrien γ ja c arvoilla STAR-malli on lineaarinen parametrien ϕ_1 ja ϕ_2 suhteen. Tällöin neliösummafunktio (14) voidaan kirjoittaa muodossa

$$Q_n(\gamma,c) = \sum_{t=1}^{n} (y_t - \phi(\gamma,c)' x_t(\gamma,c))^2.$$

Tämä vähentää NLS estimoinnin ulottuvuuksien määrää, koska neliösummafunktio minimoidaan vain parametrien γ ja c suhteen.

Parametrin γ tarkka estimointi on osoittautunut vaikeaksi. Yksi syy on se, että parametrin γ arvo määrittää logistisen funktion muodon ja suurilla arvoilla muoto muuttuu vain

vähän. Tästä johtuen parametrin γ tarkka estimointi vaatii useita havaintoja kynnysparametrin c ympärille. Tämä on kuitenkin harvinaista, jolloin parametrin γ estimointi on kohtuullisen epätarkkaa ja tästä johtuen gamman estimaatit eivät useinkaan ole tilastollisesti merkitseviä. Tätä ei kuitenkaan tule tulkita epälineaarisuuden puutteeksi STAR-mallin alaisuudessa, vaan tähän tarkoitukseen on kehitetty erilaisia diagnostisia menetelmiä.

3.2.3 Lineaarisuuden testaaminen STAR-mallissa

Regiiminmuutosmallien käytössä tärkein kysymys on, selittääkö useamman regiimin omaava malli tutkimuksen kohdetta paremmin kuin lineaarinen malli. Työn rajoitteiden mukaisesti keskitytään kahden regiimin malleihin, jolloin jäljelle jäävän lineaarisuuden testaaminen käy kohtuu kätevästi.

STAR-mallin kohdalla lineaarisuutta testatessa nollahypoteesin ollessa voimassa ongelmana on, että malliin sisältyy niin sanottuja kiusaparametreja. Nämä parametrit aiheuttavat sen, että tavanomaista tilastotieteen teoriaa ei voida hyödyntää testisuureen asymptoottisen jakauman selvittämiseen. Testisuureella onkin poikkeuksellinen jakauma, jonka kriittisten arvojen määrittäminen täytyy suorittaa simulaation keinoin.

STAR-mallin kohdalla jäljelle jäävää lineaarisuutta voidaan testata useammallakin nollahypoteesin valinnalla. Ensimmäinen ja luonteva nollahypoteesin valinta on verrata kahden regiimin parametrien samanlaisuutta $H_0 : \phi_1 = \phi_2$, jolloin vastahypoteesi on $H_1 : \phi_{i,1} \neq \phi_{i,2}$ vähintään yhdelle $i \in \{0, ..., p\}$. Tällöin mallin kiusaparametreina ovat γ ja c. Mikäli nollahypoteesi jää voimaan, malli redusoituu lineaariseksi AR-malliksi. Toinen vaihtoehto nollahypoteesin valinnalle on $H_0' : \gamma = 0$. Tällöin kiusaparametrejä ovat ϕ_1, ϕ_2 ja c. Mikäli $\gamma = 0$, niin logistinen funktio saa aina arvon 0.5 ja STAR-malli redusoituu lineaariseksi AR-malliksi parametrilla $(\phi_1 + \phi_2)/2$. Huomionarvoista on, että nollahypoteesilla H_0' parametrit ϕ_1 ja ϕ_2 voivat saada mitä tahansa arvoja, kunhan niiden keskiarvo pysyy samana.

Seuraavan tarkemman tarkastelun pohjana käytetään Luukkosen, Saikkosen ja Teräsvirran vuonna 1988 esittelemää analyysiä [3]. STAR-mallin lineaarisuuden testaamisessa on todettu mahdolliseksi käyttää Lagrange Multiplier (LM)-testiä, jolla on asymptoottinen χ^2-jakauma. Kirjoitetaan STAR-malli 13 uudelleen muotoon

$$y_t = \frac{1}{2}(\phi_1 + \phi_2)' x_t + (\phi_2 - \phi_1)' x_t G^*(y_{t-1}; \gamma, c) + \epsilon_t,$$

jossa $G^*(y_{t-1}; \gamma, c) = G(y_{t-1}; \gamma, c) - 1/2$. Nyt mikäli $\gamma = 0$, myös funktio $G^*(y_{t-1}; \gamma, c)$ saa arvon nolla. Olettamalla, että $\gamma = 0$ voidaan funktiota $G^*(y_{t-1}; \gamma, c)$ approksimoida ensimmäisen asteen Taylorin approksimaatiolla pisteen $\gamma = 0$ ympäristössä

$$T_1(y_{t-1}; \gamma, c) \approx G^*(y_{t-1}; 0, c) + \gamma \frac{\partial G^*(y_{t-1}; \gamma, c)}{\partial \gamma} = \frac{1}{4}\gamma(y_{t-1} - c),$$

jossa on käytetty aiempaa tulosta $G^*(y_{t-1}; 0, c) = 0$. Sijoitetaan Taylorin approksimaatio aiemman $G^*(.)$ funktion tilalle ja järjestelemällä termejä uudelleen saadaan apuregressiomalli

$$y_t = \beta_{0,0} + \beta_0^{'}\widetilde{x}_t + \beta_1^{'}\widetilde{x}_t y_{t-1} + \eta_t,$$

missä $\widetilde{x}_t = (y_{t-1}, ..., y_{t-p})$ ja $\beta_j = (\beta_{1,j}, ..., \beta_{p,j}) = 0, 1$. Tämän apuregressiomallin parametrien yhteys STAR-mallin parametrien voidaan esittää seuraavasti

$$\begin{aligned}
\beta_{0,0} &= \frac{\phi_{0,1} + \phi_{0,2}}{2} - \frac{1}{4}\gamma c(\phi_{0,2} - \phi_{0,1}),\\
\beta_{1,0} &= \frac{\phi_{1,1} + \phi_{1,2}}{2} - \frac{1}{4}\gamma(c(\phi_{1,2} - \phi_{1,1}) - (\phi_{0,2} - \phi_{0,1})),\\
\beta_{i,0} &= \frac{\phi_{i,1} + \phi_{i,2}}{2} - \frac{1}{4}\gamma c(\phi_{i,2} - \phi_{i,1}), i = 2, ..., p,\\
\beta_{i,1} &= \frac{1}{4}\gamma c(\phi_{i,2} - \phi_{i,1}), i = 1, ...p.
\end{aligned}$$

Yllä olevien yhtälöiden perusteella $\gamma = 0$ on yhtä kuin $\beta_{i,1} = 0,\ i = 1, ..., p$. Tällöin nollahypoteesi $H_0^{'} : \gamma = 0$ testaaminen vastaa täysin nollahypoteesin $H_0^{''} : \beta_1 = 0$ testaamista ja testisuure on asymptoottisesti χ^2-jakautunut p:llä vapausasteella. Tätä mallia käyttäessä kiusaparametreja ei tarvitse estimoida nollahypoteesin ollessa voimassa.

Tällä apuregressiomallilla ei ole kuitenkaan voimaa tilanteessa, jossa ainoastaan vakiotermit poikkeavat toisistaan eri regiimien välillä. Luukkonen, Saikkonen ja Teräsvirta ehdottivatkin, että funktio $G^*(y_{t-1}; \gamma, c)$ korvataan kolmannen asteen Taylorin approksimaatiolla

$$T_3(y_{t-1}; \gamma, c) \approx \frac{1}{4}\gamma(y_{t-1} - c) + \frac{1}{48}\gamma^3(y_{t-1} - c^3).$$

Approksimaation avulla saadaan apuregressiomalli

$$y_t = \beta_{0,0} + \beta_0^{'}\widetilde{x}_t + \beta_1^{'}\widetilde{x}_t y_{t-1} + \beta_2^{'}\widetilde{x}_t y_{t-1}^2 + \beta_3^{'}\widetilde{x}_t y_{t-1}^3 + \eta_t,$$

jossa $\beta_{0,0}$ ja β_j, j = 1,2,3, ovat parametrien ϕ_1, ϕ_2, γ ja c funktioita. Tällöin nollahypoteesi $H_0^{'} : \gamma = 0$ voidaan kirjoittaa muotoon $H_0^{''} : \beta_1 = \beta_2 = \beta_3 = 0$. Nyt voidaan käyttää LM-testiä ja nollahypoteesin ollessa voimassa testitulos noudattaa χ^2-jakaumaa 3p vapausasteella.

LM-testistä käytetään yleisesti kuitenkin vaihtoehtoista F-versiota, joka soveltuu erityisesti pienempien aineistojen testaukseen. F-version muodostamisessa käytetään pohjana yllä muodostettua apuregressiomallia. Ennen testisuureen muodostamista tarvitaan kaksi vaihetta. Ensin estimoidaan nollahypoteesin alainen malli, jossa vastetta y_t selitetään selittävillä muuttujilla x_t. Tästä lasketaan residuaalit $\tilde{\epsilon}_t$ ja residuaalien neliösumma $SSR_0 = \sum_{t=1}^{n} \tilde{\epsilon}_t^2$. Tämän jälkeen estimoidaan apuregressiomallin residuaalit $\hat{\epsilon}$ selittävillä muuttujilla x_t ja $\hat{x}_t y_{t-1}^j$, j = 1, 2, 3. Näiden avulla lasketaan residuaalien neliösummat SSR_1. Nyt testisuure voidaan muodostaa

$$LM = \frac{(SSR_0 - SSR_1)/3p}{SSR_1/(n - 4p - 1)},$$

joka on approksimatiivisesti F-jakautunut 3p ja n-4p-1 vapausasteilla nollahypoteesin ollessa voimassa.

3.2.4 STAR-mallin diagnostiikka

Lopuksi tarkastellaan vielä muutamia STAR-mallin diagnostisia testejä. Epälineaaristen mallien kohdalla on mahdollista käyttää tiettyjä samoja testejä, joita käytetään lineaaristen mallien tarkasteluun. Esimerkiksi residuaalien normaalisuutta ja LM-testiä on mahdollista edelleen käyttää, kun taas Ljung-Boxin testi ei anna luotettavia tuloksia ja sen käyttöä ei siksi suositella. STAR-mallin diagnostiikkaa tarkastellessa käytetään ainakin kolmea

testiä: autokorrelaatiotestiä, jäljelle jäävän epälineaarisuuden testaamista sekä parametrien uskottavuuden testaamista. Sivuutetaan viimeisenä mainitun testin esittely.

Autokorrelaatiotesti STAR-mallien tapauksessa on yleistys AR(p)-mallille tehtävästä sarjakorrelaation LM-testistä (katso tarkempi perustelu Franses ja van Dijk, 2000, s. 111)[1]. Esitellään testi yleisen p-asteisen epälineaarisen autoregressiivisen mallin

$$y_t = F(x_t; \theta) + \epsilon_t \tag{17}$$

kautta. Tässä $x_t = (1, y_{t-1}, ..., y_{t-p})'$ ja $F(x_t; \theta)$ on aiemmin esitelty mallin runko (15), joka on vähintään kahdesti differentioituva. LM-testin q:nnen kertaluvun sarjariippuvuus ϵ_t:n suhteen saadaan muotoon nR^2, missä R^2 on $\hat{\epsilon}_t$:n selityskerroin. Lopputuloksena saatava testitulos on asymptoottisesti χ^2-jakautunut q-vapausasteella.

Diagnostisessa tarkastelussa voidaan selvittää myös, kykeneekö epälineaarinen malli riittävällä tarkkuudella mallintamaan aikasarjan epälineaarisuutta. Tämän selvittämiseksi on mahdollista testata jäljelle jäävää epälineaarisuutta. Kahden regiimintapauksessa yleistä on testata nollahypoteesia, jonka mukaan kolmas regiimi on tarpeellinen. STAR-mallin etuna verrattuna SETAR-malliin on se, että LM-testin tekeminen on mahdollista, jolloin hypoteesin testaaminen on mahdollista ilman monimutkaisemman mallin parametrien estimointia. Eitrheim ja Teräsvirta (1996) ovat kehittäneet LM-testin kahden regiimin STAR-mallin testaamiselle kolmen regiimin mallia vastaan. Tämän tarkempi tarkastelu ohitetaan.

4 Epälineaarisen mallin käyttö käytännön sovelluksessa

Käytännön sovelluksena tarkastellaan mallintamistehtävää, jossa epälineaarisen mallin käyttäminen voi olla lineaarista mallia hyödyllisempää. Aluksi tarkastellaan aineistoa ja pyritään muunnosten avulla mahdollistamaan lineaarisen mallin käyttö. Valitaan sopiva AR(p)-malli käyttämällä soveltuvaa mallin valintakeinoa taulukoimalla mahdolliset vaihtoehdot. Seuraavaksi pyritään selvittämään sopiva epälineaarinen STAR(p)-malli käyttäen avuksi soveltuvia valintakeinoja ja niiden taulukointia.

Käytännön sovellus toteutetaan RStudio-ohjelmalla [5]. Sovelluksen toteutukseen tarvitaan seuraavia paketteja: `readxl`, `tidyverse`, `forecast` ja `tsDyn`. Parhaan mallin valinnassa hyödynnetään Akaiken informaatiokriteeriä sekä lineaarisissa että epälineaarisissa malleissa. Epälineaarisen mallin paremmuutta lineaariseen vastineeseen testataan luvussa 3.2.3 esitellyllä F-testillä ja lisäksi käytetään myös kahta muuta rajoittavaa kriteeriä.

4.1 Akaiken informaatiokriteeri

Akaiken informaatiokriteeri (AIC) on yksi käytetyimmistä mallinvalintakriteereistä. Kriteerin esitteli ensimmäisenä Hirotugu Akaike (1974). Tämä valintakriteeri vertailee mallin sopivuutta suhteessa mallin parametrien määrään eli se pyrkii valitsemaan mallin, joka sopii aineistoon hyvin välttäen kuitenkaan liikaa monimutkaisuutta. AR(p)-mallien tapauksessa Akaiken informaatiokriteeri kirjoitetaan

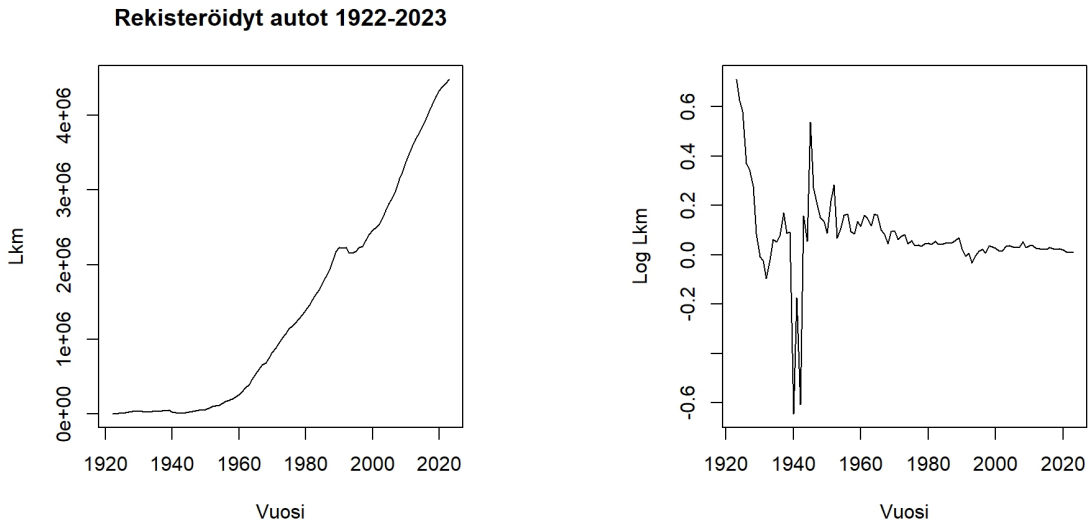
$$AIC = n\, ln(\hat{\sigma}^2) + 2k, \tag{18}$$

kun k = p + 1 ja $\hat{\sigma}^2 = 1/n \sum_{t=1}^{n} \hat{\epsilon}_t^2$, jossa AR-mallin residuaalit merkitään $\hat{\epsilon}_t$ [1]. 2-regiimisen STAR-mallin tapauksessa AIC voidaan laskea molempien regiimien AR-mallin mukaisesti ja lopullinen tulos on näiden kahden regiimin summa.

4.2 Aineisto ja autokorrelaatiokuvaajat

Aineistoksi käytännön tehtävään valikoitui Tilastokeskuksen ylläpitämä aikasarja-aineisto rekisteröityjen autojen lukumäärästä vuosina 1922-2023. Tämän aikasarjan kuvaajassa (kuva 4, vas) on huomattavissa kaksi toisistaan erottuvaa vaihetta. Kuvaajan alkupuolella vuosien 1920-1960 välillä rekisteröityjen autojen määrä vaikuttaa pysyvän kohtuullisen samanlaisena. 60-luvun vaihteessa trendi kuitenkin muuttuu vahvasti positiiviseksi. Aikasarjan mallintamisessa voisi olla luontevaa käyttää epälineaarista mallia.

Tehdään aineistolle kaksi muutosta. Logaritmisoidaan aikasarjan jokainen havainto ja poistetaan jäljelle jäävä positiivinen trendi ottamalla ensimmäisen viiveen differenssi. Muunnetun aineiston (kuva 4, oik) kuvaajasta todetaan, että varianssi on ajan kuluessa hyvin epätasaista. Jo pelkästään tämän perusteella voitaisiin todeta, että lineaarisen mallin käyttö tähän mallintamistehtävään ei ole perusteltua. Ohitetaan kuitenkin tämä havainto ja oletetaan heikon stationaarisuuden oletukset päteviksi.



Kuva 4: Alkuperäinen aikasarja (vas) ja aikasarja muunnosten jälkeen (oik)

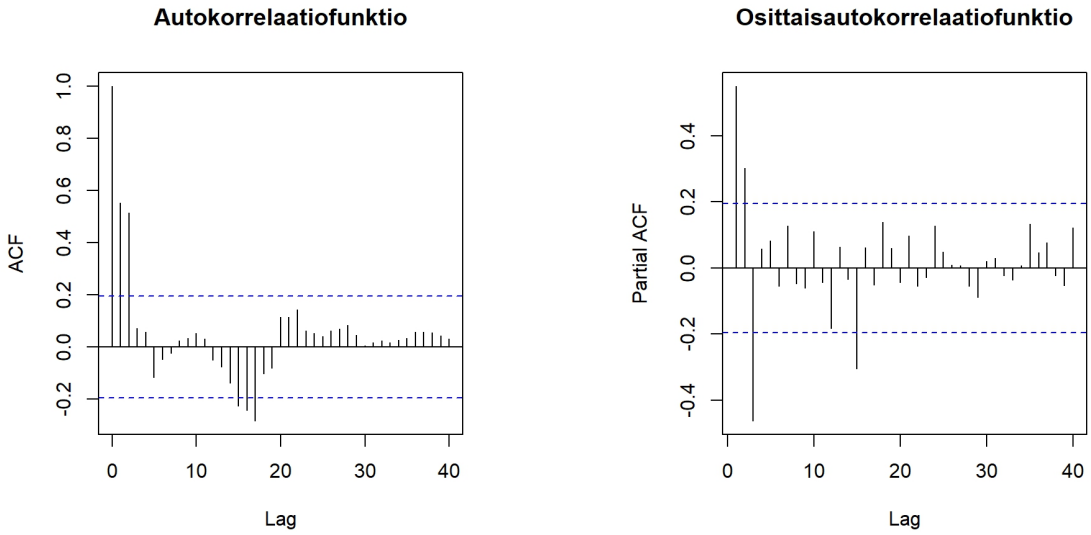
Autokorrelaatiota kuvaajan (kuva 5, vas) perusteella on paljon viipymillä 1 ja 2. Myös viiveen 15 kohdilla autokorrelaatiota on kohtuullisen paljon, mutta mitä pidempi viiveen pituus on sitä vähemmän korrelaatiota on. Osittaisautokorrelaatiokuvaajasta (kuva 5, oik) voidaan todeta, että ensimmäisillä viipymillä korrelaatiota on reilusti. Myös noin viiveen 15 kohdalla kriittinen raja ylittyy, kuten autokorrelaatiokuvaajan kohdallakin kävi. Osittaisautokorrelaatiokuvaajan perusteella parhaan mallin arvioiminen on hieman haastavaa. Mikäli rajoitetaan viiveiden määrää alle kymmeneen, tällöin malliksi voitaisiin valita AR(3).

4.3 AR(p)-mallin valinta ja parametrien estimointi

AR(p)-mallin valinnassa käytetään hyödyksi aiemmin mainitun `forecast`-paketin `auto.arima`-komentoa. Tämä komento käy läpi kaikki mallit käyttäjän rajoitusten mukaisesti. Paras malli valitaan käyttäen Akaiken informaatiokriteeriä, josta etsitään pienintä mahdollista arvoa.

p	AIC	ϕ_0	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
1	-113.37	0.09	0.64					
2	-129.23	0.12	0.41	0.43				
3	-155.41	0.08	0.61	0.57	-0.53			
4	-153.49	0.08	0.62	0.55	-0.55	0.03		
5	-155.59	0.09	0.62	0.66	-0.66	-0.07	0.22	
6	-153.62	0.09	0.62	0.66	-0.67	-0.06	0.22	0.02

Taulukko 1: Taulukossa nähdään mallin valintaan perustuva AIC sekä jokaisen mallin parametrien estimaatit.



Kuva 5: Autokorrelaatiofunktiot

Rajoitetaan etsintä 6 viiveeseen (p ≤ 6). Taulukkoon 1 on kerätty jokaisen mallin AIC sekä niiden parametrien estimaatit. Parhaimmaksi malliksi osoittautuu AR(5)-malli, jonka AIC on -155,59. Parametrit sijoitettuna (5) mallin mukaiseen muotoon antaa meille tulokseksi funktion

$$y_t = 0.09 + 0.62y_{t-1} + 0.66y_{t-2} - 0.66y_{t-3} - 0.07y_{t-4} + 0.22y_{t-5}.$$

4.4 STAR(p)-mallin valinta kynnysmuuttujan avulla ja parametrien estimointi

Parhaan epälineaarisen STAR(p)-mallin etsintä aloitetaan valitsemalla indikaattorifunktio. Indikaattorifunktioksi valitaan logistinen funktio (11). Parhaan mallin valintaa varten muodostetaan kaksi erillistä taulukkoa. Rajoitetaan etsintä 5 viiveeseen (p < 5). Taulukoiden toisena muuttuvana tekijänä on kynnynarvon suuruus. Kynnysmuuttujalle on asetettu rajoite $0 < d < p - 1$. Ensimmäinen taulukko 2 sisältää kaikkien rajoitteiden mukaisten mallien AIC:t. Toinen taulukko 3 sisältää näiden mallien F-testissä nollahypoteesi hylätään merkitsevyystasolla 0.05. Mallien määrittelyssä parametrin γ:a on rajoitettu välille 1-100. Joidenkin mallien kohdalla mallin muodostaminen pyytää γ:lle suurempaa vaihteluväliä, mutta nämä mallit voidaan hylätä valintavaiheessa. Perusteluna käytetään, että gamman arvo on liian suuri, jotta mallia voitaisiin pitää luotettavana. Usean mallin kohdalla myös mallin parametrien tilastolliseen merkitsevyyteen liittyvä testaaminen ei onnistu. Myös tällaiset mallit rajataan tarkastelun ulkopuolelle.

Parhaan mallin määrittelemiseksi käytetään useampaa kriteeriä. Käydään malleja yksitellen läpi niin, että aloitetaan pienimmän AIC:n saaneesta mallista. Tarkastellaan seuraavaksi F-testin p-arvoa. Mikäli lineaarisuuden nollahypoteesi hylätään tarkastellaan vielä

max(p)	d = 0	d = 1	d = 2	d = 3	d= 4
1	-428				
2	-444	-435			
3	-492	-478	-478		
4	-489	-475	-475	-484	
5	-489	-476	-476	-486	-477

Taulukko 2: LSTAR(p)-mallin AIC:t

max(p)	d = 0	d = 1	d = 2	d = 3	d= 4
1	<0.001				
2	0.057	0.687			
3	<0.001	<0.001	0.002		
4	<0.001	<0.001	0.003	0.018	
5	0.003	0.018	0.004	0.022	0.280

Taulukko 3: LSTAR(p)-mallien F-testin p-arvot, kun testataan nollahypoteesia $H_0 : \phi_1 = \phi_2$, jonka vastahypoteesi on $H_1 : \phi_{i,1} = \phi_{i,2}$. F-testin avulla todettu nollahypoteesin hylkääminen todetaan merkitsevyystasolla 0.05 eli kun p < 0.05. Kun p > 0.05, nollahypoteesi jää voimaan.

parametrin γ arvoa. Mikäli γ:n arvo on mielekkäiden rajojen puitteissa, voidaan tarkastella viimeisen vaiheen vaatimuksia. Viimeisenä varmistetaan vielä, että mallin parametrien tilastollisen merkitsevyyden testaaminen on mahdollista. Mikäli valitun mallin kohdalla kriteerit eivät täyty, siirrytään tarkastelemaan seuraavaksi parhaan AIC:n tuottaneen mallin tuloksia.

Ensimmäisenä tarkastelun läpi menevä malli on LSTAR(5)-malli kynnysmuuttujan viiveellä d = 0. Tämän mallin parametrin γ arvo on kuitenkin edelleen kohtuullisen korkea 28.65. Mallin käyttäminen olisi mahdollista, mutta tutkitaan vielä muita vaihtoehtoja. Seuraava kaikki kriteerit täyttävä malli on LSTAR(3)-malli kynnysmuuttujan viivellä d = 1. Tässä mallissa parametrin γ arvo on kohtuullinen 2.97. Kynnysarvo c on 0.27. Mallin parametrien estimoinnin jälkeen malli voidaan kirjoittaa funktion (10) mukaiseen muotoon

$$\begin{aligned} y_t = &(-0.44 + 0.93y_{t-1} + 0.05y_{t-2} - 1.49y_{t-3})(1 - G(y_{t-1} : \gamma, c) \\ &+(1.50 - 1.34y_{t-1} - 1.20y_{t-2} + 2.48y_{t-3})G(y_{t-1}; \gamma, c) + \epsilon_t. \end{aligned}$$

Mallin parametrien, eritysesti γ:n, arvot ovat nyt mielekkäämpiä tulkita. Valitaan malli mukaan vertailuun.

4.5 STAR(p)-mallin valinta muuttujan ν_t avulla

STAR-malli voidaan muodostaa aineiston useammasta edellisestä havainnosta. Käytetään nyt mallin valintaan muuttujaa $\nu_{t,j}$, joka määritellään j:nnen edellisen havainnon itseisarvojen keskiarvoina

$$\nu_{t,j} = \frac{1}{j}\sum_{i=0}^{j-1}|y_{t-i}|.$$

Tämän aineiston kohdalla mielekkäiden vertailtavien mallien etsintä vaatii hieman laajempaa tutkintaa. Taulukoidaan jälleen sekä AIC:t että F-testien p-arvot. Etsintäväli määritellään p:n mukaan välille p = 1,...,5 ja j:n mukaan välille j = 9,...,13, kun muuttuja on ν_{t-1}. Huomioitavaa on, että p on tässä tapauksessan p:nnen edellisen ν_t:n arvot ja j määrittää, montako edellistä havaintoa otetaan huomioon muuttujan ν_t laskemista varten.

max(p)	v = 9	v = 10	v = 11	v = 12	v = 13
1	-395	-393	-390	-387	-384
2	-412	-412	-408	-398	-390
3	-486	-474	-446	-439	-425
4	-458	-453	-444	-428	-412
5	-459	-451	-439	-425	-501

Taulukko 4: LSTAR(p)-mallien AIC:t

max(p)	v = 9	v = 10	v = 11	v = 12	v = 13
1	0.716	0.374	0.390	0,309	0.288
2	0.406	0.484	0.314	0,193	0.167
3	<0.001	<0.001	<0.001	<0,001	<0,001
4	<0.001	<0.001	<0.001	<0.001	0.005
5	<0.001	<0.001	0.001	0.008	0.091

Taulukko 5: LSTAR(p)-mallien F-testin p-arvot, kun testataan nollahypoteesia $H_0 : \phi_1 = \phi_2$, jonka vastahypoteesi on $H_1 : \phi_{i,1} = \phi_{i,2}$. F-testin avulla todettu nollahypoteesin hylkääminen todetaan merkitsevyystasolla 0.05 eli kun p < 0.05. Kun p > 0.05, nollahypoteesi jää voimaan.

Mallin valinnassa käytetään samoja kriteereitä kuin aiemmassa STAR(p)-mallin valinnassa. Tällä tavalla muodostettujen mallien kohdalla parametrin γ arvo on huomattavan usein epämiellyttävä ja tämä johtaa nopeasti useamman mallin käytön hylkäämiseen. Haarukoinnin avulla parhaaksi malliksi saadaan LSTAR(3)-malli, kun $\nu_{t-1,11}$ tai $\nu_{t-1,12}$. Molempien kohdalla γ:n arvo on kuitenkin edelleen kohtuullisen korkea ($\nu_{t-1,11} : \gamma = 15.07$, c = 0,27 ja $\nu_{t-1,12} : \gamma = 21,31$, c= 0,23). Kun muuttujana on $\nu_{t-1,11}$, mallin parametrien estimaateiksi saadaan

$$\begin{aligned} y_t =(& 0.02 + 0.48y_{t-1} + 0.88y_{t-2} - 1.15y_{t-3})(1 - G(y_{t-1} : \gamma, c) \\ & +0.36 - 0.32y_{t-1} - 3.33y_{t-2} + 2.85y_{t-3}G(y_{t-1} : \gamma, c) + \epsilon_t. \end{aligned}$$

4.6 Yhteenveto

STAR-malli oli tähän mallintamistehtävään käytännöllinen vaihtoehto. Lineaarisen mallin käytön ongelmaksi muodostui heikon stationaarisuuden toteamiseen vaadittavien oletusten toteutuminen. Selkeimmin nousi esiin log-differentioidun kuvaajan varianssi, joka oli selkeästi suurempaa aineiston alkuvaiheessa kuin loppuvaiheessa. Tästä huolimatta mukaan vertailuun valittiin AR(5)-malli, joka kuitenkin pärjäsi huonosti vertailtaessa AIC:tä niihin STAR-malleihin, jotka toteuttivat näille määritellyt kriteerit.

STAR-mallien valinnassa käytettiin kriteereinä AIC:ä sekä lineaarisuuden testaaminen F-testin avulla, jossa nollahypoteesi pyrittiin hylkäämään. Lisäkriteerinä tehtiin subjektiivinen tarkastelu parametrin γ arvolle. Lopuksi haluttiin vielä varmistaa, että mallin parametrien tilastollinen merkitsevyys on mahdollista testata. Näistä kriittisin vaihe oli lineaarisuuden testaaminen, mutta myös tilastollisen merkitsevyyden testaamisen epäonnistuminen toimi ehdottomana karsijana sopivien vaihtoehtojen ulkopuolelle. Parametrin γ subjektiivinen arviointi puolestaan perustui oletukseen, että suuren gamman estimoiva malli ei todennäköisesti tuota hyvää mallia. Parametrin suurella arvolla oli useimmiten myös suora yhteys tilastollisen merkitsevyyden testaamisen epäonnistumiseen. AIC:n käytöllä pyrittiin lähinnä tekemään haarukoinnista helpompaa käymällä läpi järjestyksessä pienimmät arvot saaneet mallit.

Parhaaksi malliksi valittujen kriteerien perusteella valikoituu LSTAR(3)-malli, kun kynnysmuuttuja d = 1. Tässä mallissa F-testin tuloksen perusteella nollahypoteesi voidaan hylätä, parametrien tilastollinen merkitsevyys on mahdollista laskea, parametrin γ arvo on valikoiduista malleista mielekkäin sekä AIC kaikkein pienin. On kuitenkin huomioitava, että tämän mallinnustehtävän parhaan mallin valinnassa käytettiin hyvin yksinkertaisia ja konkreettisia kriteereitä, joiden avulla laajemmassa tutkinnassa voitaisiin lähinnä määritellä sopivimmat mallit eri mallinvalintakeinojen kautta. Tämän takia kynnysmuuttujan avulla määritellyistä malleista paras vaikuttaa valittujen kriteerien valossa paremmalta kuin muuttujan $\nu_{t,j}$ avulla muodostettujen mallien parhaimmisto. Tarkastelussa ei kuitenkaan mennä syvemmälle esimerkiksi diagnostiseen tarkasteluun, jolloin muuttujan $\nu_{t,j}$ avulla muodostettujen mallien edut tulevat selkeämmin esille.

Tällä esimerkillä pyrittiin esittelemään vaihtoehtoja lineaaristen mallien käytölle, kun sen käyttöön liittyvät vaatimukset eivät toteudu. Vaikka tavoitteena oli löytää tälle paras epälineaarinen malli, tämä oli mahdollista lähinnä tehtyjen valintakriteerien ollessa hyvin konkreettisia. Kriteereiksi olisi voitu valita esimerkiksi diagnostiseen tarkasteluun liittyviä kriteereitä, jotka olisivat muokanneet lopputulosta. Sopivan mallin valintaan liittyy siis aina tulkinnallinen elementti.

5 Johtopäätökset

Tässä tutkielmassa syvennettiin aikasarja-analyysiin liittyvää lineaarista teoriaa kohti epälineaaristen mallien käyttöä. Lisäksi esiteltiin käytännön sovellus, jossa lineaarisen AR-mallin vaatimat perusoletukset eivät toteudu. Tähän vaihtoehdoksi esiteltiin STAR-mallin kaksi variaatiota: STAR(p)-mallit logistisella indikaattorifunktiolla kynnysmuuttujan viiveellä ja edellisten havaintojen itseisarvojen keskiarvolla.

Epälineaaristen mallien osalta ei ole yhtä selkeää linjausta parhaan mallinvalinnan kriteereistä. Tästä johtuen mallinvalintaa ohjaa vahvasti käytettävien kriteerien määrittely. Tässä tutkielmassa kriteereiksi valittiin malleja matemaattisesti rajaavia tekijöitä, joiden avulla valintaa oli mahdollista konkretisoida selkeämmin. Diagnostinen tarkastelu jätettiin verrattain pienelle huomiolle ja tällä on merkittävä vaikutus lopputulokseen. Näiden kriteerien perusteella parhaaksi malliksi valikoitui STAR(3)-malli kynnysmuuttujan viivellä d = 1.

Kriteerien erilaisella määrittelyllä esiin olisivat nousseet erilaiset mallit. Selkeän linjauksen puutteen takia tulkinnallisilla tekijöillä on suuri merkitys mallinvalintatilanteessa. Tarkempi erilaisilla kriteereillä valittujen mallien välinen vertailu jätettiin kuitenkin tämän tutkielman ulkopuolelle. Tutkielman laajennuksena voitaisiinkin tarkastella mallinnustehtävää laajentamalla valintakriteereiden käyttöä diagnostisten menetelmien kautta löytyviin kriteereihin. Tällä tavoin on mahdollista perustella mukana olleiden edellisten havaintojen itseisarvojen keskiarvon perusteella valittujen mallien käyttökelpoisuus mallintamistehtävässä.

Viitteet

[1] Franses Philip Hans, ja van Dijk Dick. *Non-Linear Time Series Models in Empirical Finance*, Cambridge University Press, 2000.

[2] Leybourne Stephen, Newbold Paul, Vougas Dimitrios. *Unit roots and smooth transitions*, Journal of Time Series Analysis 19, 1998.

[3] Luukkonen Ritva, Saikkonen Pentti ja Teräsvirta Timo. *Testing Linearity Against Smooth Transition Autoregressive Models*, Biometrika Vol 75. 1988. <https://doi.org/10.2307/2336599>

[4] Nyberg Henri. *Aikasarja-analyysi*, Turun Yliopisto, 2020.

[5] Toivonen Juho. Mallinnustehtävän R-koodit. Github, 28.4.2025. <https://github.com/toivjuho/Reg_vehicles_FIN.git>

[6] Tong Howell. *Pattern Recognition and Signal Processing*, Amsterdam: Sijthoff & Noordhoff, 1978.

[7] Tong Howell, Lim K. S. *Threshold Autoregression, Limit Cycles and Cyclical Data*, Journal of the Royal Statistical Society B 42, 1980.

[8] Tsay Ruey S. *Analysis of Financial Time Series*, John Wiley and Sons, Incorporated, 2010.

[9] Tsay Ruey S, ja Rong Chen. *Nonlinear Time Series Analysis*, Wiley Series in Probability and Statistics, 2018.