Aalto-yliopisto
Perustieteiden korkeakoulu
Teknillisen fysiikan ja matematiikan tutkinto-ohjelma

Spatiaalisiin merkkeihin ja järjestyslukuihin perustuva moniulotteinen regressioanalyysi R-ohjelmistoa käyttäen

kandidaatintyö
29.4.2014

Niko Lietzén

Aalto-yliopisto
Perustieteiden korkeakoulu

AALTO-YLIOPISTO PERUSTIETEIDEN KORKEAKOULU PL 11000, 00076 Aalto http://www.aalto.fi	KANDIDAATINTYÖN TIIVISTELMÄ

Tekijä: Niko Lietzén

Työn nimi: Spatiaalisiin merkkeihin ja järjestyslukuihin perustuva moniulotteinen regressioanalyysi R-ohjelmistoa käyttäen

Tutkinto-ohjelma: Teknillisen fysiikan ja matematiikan tutkinto-ohjelma

Pääaine: Systeemitieteet	Pääaineen koodi: F3010

Vastuuopettaja(t): Apulaisprof. Pauliina Ilmonen

Ohjaaja(t): Apulaisprof. Pauliina Ilmonen

Tiivistelmä:

Työssä tavoitteena on vertailla perinteistä lineaarista regressiota ja robustimpaa vaihtoehtoa. Spatiaalisiin merkkeihin ja järjestyslukuihin perustuva lineaarinen regressio on perinteistä L_2-regressiota robustimpi eli se ei ole niin herkkä poikkeaville havainnoille. Menetelmiä sovelletaan neljään eri simuloituun aineistoon. Aineistot simuloidaan normaalijakaumasta ja elliptisestä jakaumasta. Työssä määritellään lyhyesti elliptisen jakauman ominaisuudet. Aineistoissa on mukana myös poikkeavia havaintoja, jolloin menetelmien robustisuutta voidaan verrata. Simulointivaiheessa päätetään selitettävien ja selittävien muuttujien välinen todellinen lineaarinen riippuvuus. Vertailu regression onnistumisesta tehdään vertaamalla eri menetelmien tuottamia estimaatteja todelliseen lineaariseen riippuvuuteen.

Työssä otetaan tavallista yleisempi lähestymistapa estimointi- ja testiongelmaan. Spatiaaliset merkit ja järjestysluvut toimivat estimoimisessa pistemääräfunktioina. Spatiaalisiin merkkeihin ja järjestyslukuihin perustuvassa regressiossa estimaatteja ei voida ratkaista suljetussa muodossa. Estimaatit eivät sellaisenaan ole täysin ekvivariantteja mutta työssä esitellään algoritmi, jolla täysin ekvivariantit estimaatit saavutetaan. Perinteisen regression estimaatit saadaan käyttämällä pienimmän neliösumman menetelmää. Työssä esitetään lisäksi affiinisesti invariantit testisuureet, jotka saavutetaan sisäisellä standardoinnilla.

Monet moniulotteisuuteen liittyvät ongelmat tulevat työssä ilmi. Esimerkiksi regressiografiikan tuottaminen on erittäin haastavaa monessa ulottuvuudessa. Lisäksi tuloksista huomataan spatiaalisten merkkien ja järjestyslukujen robusti luonne, ne toimivat paremmin elliptisesti jakautuneelle aineistolle tai jos mukana on poikkeavia havaintoja. Perinteinen regressio antaa parhaan estimaatin, kun kaikki standardioletukset pätevät.

Päivämäärä: 29.4.2014	Kieli: suomi	Sivumäärä: 23

Avainsanat: spatiaaliset merkit, spatiaaliset järjestysluvut, R-ohjelmisto, moniulotteisuus, elliptinen jakauma, pistemääräfunktio, lineaarinen regressio, invarianttisuus, ekvivarianttisuus, robusti

Sisältö

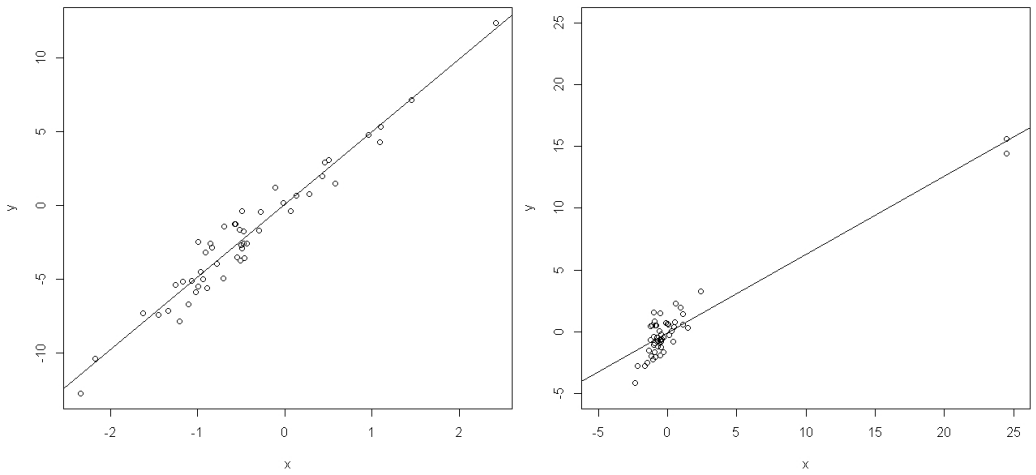
1 Johdanto

Spatiaaliset merkit ja järjestysluvut tarjoavat robustimman vaihtoehdon verrattuna perinteisiin regressiomenetelmiin, jotka hyödyntävät L_2-normia. Robustilla menetelmällä tarkoitetaan sellaista, joka ei ole herkkä poikkeaville havainnoille. L_2-normiin perustuvat regressiomenetelmät ovat optimaalisia, kun residuaalit noudattavat moniulotteista normaalijakaumaa. Vaikka L_2-regressio olisi mahdollinen, se ei kuitenkaan ole aina optimaalinen vaihtoehto, kun data on esimerkiksi paksuhäntäinen. Kuvassa 1 nähdään, että muutama poikkeava piste vääristää merkittävästi regressiosuoraa. Robustit menetelmät ovat erityisen tärkeitä moniulotteisessa tapauksessa, sillä poikkeavien havaintojen tunnistaminen ja jakaumaoletusten varmentaminen vaikeutuvat dimensioiden kasvaessa. Kaikki työssä esiintyvät kuvat on tuotettu R-ohjelmistolla.

Työssä esitellään perinteiseen L_2-regressioon liittyvät standarioletukset sekä lisäksi L_1-normiin perustuvien spatiaalisten merkkien ja järjestyslukujen vaadittavat oletukset. Työ on rajattu käsittelemään ainoastaan lineaarista regressiota, joten taustaoletuksena on lineaarinen malli. Spatiaalisten merkkien ja järjestyslukujen teoria on esitetty Hannu Ojan teoksessa Multivariate Nonparametric Methods with R (2010).

Spatiaalinen merkki on vektori, joka kertoo datapisteiden suunnan suhteessa moniulotteiseen origoon. Spatiaalinen järjestysluku taas on vektori, jonka suunta kertoo likimain datapisteen suunnan suhteessa dataparven moniulotteiseen keskipisteeseen ja pituus kertoo kuinka kaukana datapiste likimain sijaitsee dataparven moniulotteisesta keskipisteestä. Kumpaankaan menetelmään perustuvassa lineaarisessa regressiossa estimaattia ei voida ratkaista suljetussa muodossa. Perinteisesti L_1-normiin perustuvia menetelmiä on pidetty laskennallisesti raskaina, erityisesti moniulotteisessa tapauksessa. Spatiaalisiin merkkeihin ja järjestyslukuihin perustuvat estimaatit eivät välttämättä ole täysin ekvivariantteja. Täysin ekvivariantit estimaatit saavutetaan luvussa 3 esitetyllä sisäisellä standardoinnilla.

Aineistot luodaan työssä simuloimalla, mutta itse simulointiin liittyviin teknisiin yksityiskohtiin ei paneuduta. Simuloituja dataesimerkkejä on työssä neljä, moniulotteinen normaalijakauma sekä moniulotteinen elliptinen jakauma. Molempiin aineistoihin on lisätty lisäksi poikkeavia havaintoja, jotka on sijoiteltu epäsymmetrisesti. Moniulotteinen elliptinen jakauma generoidaan Studentin t-jakauman avulla. Simuloinnin työkaluna käytetään R-ohjelmistoa. R-ohjelmisto on työssä tärkein työkalu, ja ohjelmistossa on valmis paketti MNM, spatiaalisten merkkien ja järjestyslukujen hyödyntämiseen [7].



(a) L_2-regressio normaalijakautuneeseen aineistoon.

(b) L_2-regressio normaalijakautuneeseen aineistoon, missä poikkeavia havaintoja.

Kuva 1: L_2-regressio vääristyy merkittävästi jo pienilläkin määrillä poikkeavia havaintoja.

Tavoitteena on tarkastella kaikkien kolmen yllämainitun menetelmän toimivuutta eri aineistoihin. Standardioletusten toteutumista tutkivaan regressioanalyysiin paneudutaan lyhyesti, sillä simuloitu aineisto täyttää tarvittavat oletukset. Työssä siis vertaillaan eri regressiomenetelmien onnistumista lineaarisen mallin parametrien estimoimisessa.

2 Teoreettinen tausta

Tässä luvussa esitetään teoreettinen pohja menetelmien käytölle sekä tulosten tulkitsemiselle. Luvussa 2 esitellään työn kannalta tarvittavat oletukset, joihin tullaan viittaamaan myöhemmissä luvuissa. Luvussa lähdetään yleisestä moniulotteisesta lineaarisesta mallista ja paneudutaan eri regressiomenetelmien ominaisuuksiin sekä vaadittaviin oletuksiin. Lisäksi luvussa esitellään regressioanalyysissa ja estimoinnissa tarvittavan pistemääräfunktion käyttöä eri menetelmien vaatimalla tasolla. Viimeisessä alaluvussa on esitelty lyhyesti elliptinen jakauma. Spatiaalisten merkkien ja järjestyslukujen teorian lähteenä on Hannu Ojan kirja Multivariate Nonparametric Mehtods with R(2010) [9].

2.1 L_2-regressio

Työ on rajattu käsittelemään ainoastaan lineaarista mallia. Työssä jatkossa oletetaan moniulotteinen lineaarinen malli, johon sovelletaan lineaarista regressiota. Alaluvussa hyödynnetään kurssien Matematiikan peruskurssi L3 [2] sekä Ennustaminen ja aikasarja-analyysi luentomateriaaleja.

Määritelmä 2.1 (Lineaarinen malli)**.** Olkoon $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ selitettävät muuttujat sisältävä matriisi, $\boldsymbol{X} \in \mathbb{R}^{n\times q}$ selittävät muuttujat sisältävä matriisi, $\boldsymbol{\beta} \in \mathbb{R}^{q\times p}$ regressiokertoimet sisältävä matriisi ja $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times p}$ jäännöstermit sisältävä matriisi, jolloin moniulotteinen lineaarinen malli määritellään seuraavasti

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Lineaarisessa mallissa $\boldsymbol{Y}$ on selitettävien muuttujien havaittujen arvojen muodostama satunnainen matriisi, $\boldsymbol{X}$ on selittäjien havaittujen arvojen muodostama kiinteä eli ei-satunnainen matriisi, $\boldsymbol{\beta}$ on regressiokertoimien muodostama tuntematon ja kiinteä matriisi ja $\boldsymbol{\varepsilon}$ on jäännöstermien muodostama toisistaan riippumaton satunnaisotos origon suhteen keskitetystä jakaumasta. Origon suhteen keskitetyllä tarkoitetaan, että $\mathbf{E}\left(\boldsymbol{T}(\boldsymbol{x_i})\right) = \mathbf{0}$, jollakin valitulla pistemääräfunktiolla $\boldsymbol{T}$. Jäännöstermimatriisin saraketta i merkitään $\boldsymbol{\varepsilon_i}$. Tavoitteena lineaarisessa regressiossa on löytää estimaatti tuntemattomalle $\boldsymbol{\beta}$:lle.

Seuraavat oletukset tehdään matriisista $\boldsymbol{X}$, lähteenä kurssin Ennustaminen ja aikasarja-analyysin kalvot [4].

Oletus 2.2. Matriisin $\boldsymbol{X}$ on täysiasteinen matriisi: $rank\left(\boldsymbol{X}\right) = q$.

Kun selittävien muuttujien matriisi on täysiasteinen, minkään selittäjien välillä ei ole täydellistä lineaarista riippuvuutta. Oletus 2.2 takaa, että pienimmän neliösumman menetelmä perinteisessä L_2-regressiossa tuottaa yksikäsitteiset regressiokertoimet suljetussa muodossa. Jos selittäjä riippuu lineaarisesti muista selittäjistä, se ei sisällä regression kannalta uutta informaatiota ja se voidaan poistaa mallista. Täysiasteiteinen matriisi on epäsingulaarinen, jolloin käänteismatriisi on olemassa. Lisäksi $rank\left(\boldsymbol{X'X}\right)$=$rank\left(\boldsymbol{X}\right)$, jolloin myös matriisitulo on epäsingulaarinen. Tulos seuraa suoraan dimensiolauseesta [2].

Oletus 2.3. Matriisin $\boldsymbol{X}$ alkiot ovat kiinteitä eli ei-satunnaisia vakioita.

Työssä selittäjien arvot simuloidaan eli ne ovat kiinteitä. Simuloitaessa voidaan varmistaa, että selittävät muuttujat sisältävä matriisi $\boldsymbol{X}$ on täysiasteinen. Oletusten 2.2 ja 2.3 toteutumisen tutkiminen voidaan siis työssä sivuuttaa.

Lisäksi jäännöstermimatriisista $\boldsymbol{\varepsilon}$ oletetaan seuraavaa

Oletus 2.4. Jäännöstermimatriisin odotusarvo on $n \times p$ kokoinen nollamatriisi, $\mathbf{E}(\varepsilon) = \mathbf{0}$.

Oletuksen 2.4 perusteella kaikkien jäännöstermien odotusarvo on nolla. Oletuksen pätiessä, regressiokertoimien estimoinnissa ei ole tehty systemaattista virhettä.

Oletus 2.5. Jäännöstermit ovat homoskedastisia ja korreloimattomia, $\mathbf{Cov}(\varepsilon) = \boldsymbol{\sigma^2}\mathrm{I}$.

Homoskedastisuudella tarkoitetaan, että kaikilla jäännöstermimatriisin alkioilla on sama varianssi. Jos homoskedastisuus- tai korreloimattomuusoletus ei päde, tulee regressiokertoimien estimaattoreista tehottomia. Simulointivaiheessa varmistetaan, että oletukset 2.4 ja 2.5 pätevät. Aineiston simuloimisesta lisää luvussa 3.1.

Lisäksi oletetaan normaalijakautuneen aineiston yhteydessä seuraavaa jäännöstermimatriisista $\boldsymbol{\varepsilon}$

Oletus 2.6. Jäännöstermit ovat normaalijakautuneita, $\boldsymbol{\varepsilon_i \sim N_n(0, \sigma^2 I)}$, kaikille $i = 1, ..., p$.

Jäännöstermimatriisin kaikki sarakkeet noudattavat siis n-ulotteista normaalijakaumaa, jonka odotusarvo on n-ulotteinen nollamatriisi ja kovarianssimatriisi on σ^2 kerrottuna n-ulotteisella identiteettimatriisilla. Oletuksen 2.6 pätiessä, perinteisen L_2-regression muodostama estimaattori on paras eli tehokkain, kun tehokkuuden kriteerinä käytetään estimaattorin varianssia. Työssä oletetaan jatkossa oletusten $2.2 - 2.5$ pätevän, kun poikkeavia havaintoja ei huomioida. Jatkossa oletuksiin $2.2 - 2.6$ tullaan viittaamaan termillä standardioletukset ja oletukseen 2.6 termillä normaalisuusoletus.

Työssä oletetaan jatkossa seuraavaa

Oletus 2.7. Jollekin positiivisesti definiitille $q \times q$ matriisille $\boldsymbol{D}$ ja kaikille $p \times q$ matriiseille $\boldsymbol{C}$, joilla on positiivinen asteluku,
$\frac{1}{n}\boldsymbol{X'X} \to D$ ja $\frac{max_{1\leq i\leq n}\{\boldsymbol{x_i'C'Cx_i}\}}{\Sigma_{i=1}^{n}\{\boldsymbol{x_i'C'Cx_i}\}} \to 0$.

Oletusta 2.7 tarvitaan jatkossa regression testaamisen yhteydessä.

2.2 Spatiaaliset merkit

Spatiaalinen merkki on vektori, joka kertoo datapisteen suunnan suhteessa moniulotteiseen origoon. Spatiaalisen merkin pituus eli euklidinen normi on aina yksi. Spatiaaliset merkit siis sijaitsevat aina p-ulotteisen yksikköpallon pinnalla.

Määritelmä 2.8 (Spatiaaliset merkit)**.** Olkoon $\boldsymbol{x_1, x_2, ..., x_n}$ riippumaton p-ulotteinen satunnaisotos. Datapisteen x_i spatiaalinen merkki $\boldsymbol{U(x_i)}$ määritellään, $\boldsymbol{U(x_i)} = \begin{cases} |\boldsymbol{x_i}|^{-1}\boldsymbol{x_i} & \boldsymbol{x_i} \neq 0 \\ 0 & \boldsymbol{x_i} = 0, \end{cases}$

missä $|x_i| = \sqrt{x_1^2 + x_2^2 + ... + x_p^2}$ on euklidinen normi. Spatiaalisia merkkejä $\boldsymbol{U(x)}$ käytetään pistemäärfunktiona, jonka käyttöä esitellään luvussa 2.4.

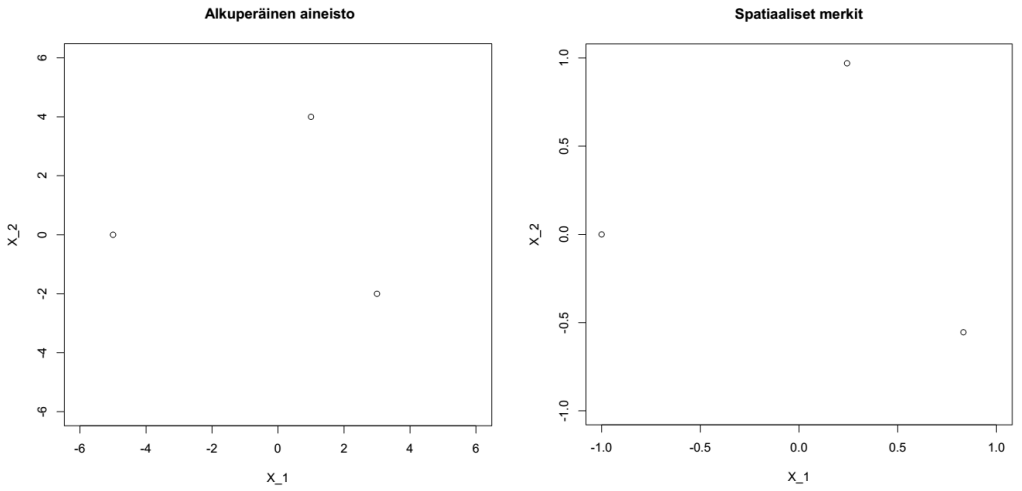
Spatiaalisten järjestyslukujen yhteydessä käytetään spatiaalista mediaania. Se on yksikäsitteinen, kun dimensioita on kaksi tai enemmän. Estimaattorina spatiaalinen mediaani on huomattavasti robustimpi kuin esimerkiksi keskiarvoestimaattori. Spatiaalinen mediaani $\hat{\mu}(\boldsymbol{Y})$ on ratkaisu minimoitavalle funktiolle

$$\mathbf{AVE}\{|\boldsymbol{y_i} - \boldsymbol{\mu}| - |\boldsymbol{y_i}|\}. \tag{2.2.1}$$

Spatiaalisten merkkien yhteydessä oletetaan spatiaalisen mediaanin olevan nolla.

Oletus 2.9. $\mathbf{E}(\boldsymbol{U}(\boldsymbol{\varepsilon_i})) = \mathbf{0}$

Oletus vaatii lisäoletuksen, että jäännöstermien $\boldsymbol{\varepsilon_i}$ tiheys on rajoitettu. Yleisessä lineaarisessa mallissa (2.1) määritetyn matriisin $\boldsymbol{X}$ sekä sen transpoosin



(a) Vektorit $\boldsymbol{a}, \boldsymbol{b}$ ja $\boldsymbol{c}$. (b) Vektoreiden spatiaaliset merkit

Kuva 2: Kuvasta nähdään spatiaalisten merkkien näyttävän suunnan suhteessa origoon. Pisteet sijaitsevat yksikköympyrällä.

tulee olla täysiasteisia, eli matriisin $\boldsymbol{X}$ sarakkeet sekä rivit oletetaan lineaarisesti riippumattomiksi. Riippumattomuudesta seuraa suoraan, että myös matriisitulo $\boldsymbol{X'X}$ antaa tuloksena täysiasteisen matriisin eli matriisitulon aste on q [6]. Spatiaaliset merkit ovat aina ortogonaalisest ekvivariantteja eli

$$\boldsymbol{U}(\boldsymbol{O}\boldsymbol{x_i}) = \boldsymbol{O}\boldsymbol{U}(\boldsymbol{x_i}), \tag{2.2.2}$$

kaikilla ortogonaalisilla matriiseilla $\boldsymbol{O}$ ja kaikilla $\boldsymbol{x_i}$. Spatiaaliset merkit ovat siis aina ortogonaalisesti ekvivariantteja mutta eivät välttämättä affiinisesti ekvivariantteja. Täysin ekvivariantit estimaatit ovat tärkeitä, sillä tuloksiin ei pitäisi vaikuttaa esimerkiksi aineiston skaalaus tai lokaation muutos.

Esimerkki 2.1. Olkoon $\boldsymbol{a}, \boldsymbol{b}$ ja $\boldsymbol{c}$ satunnaisesti generoituja vektoreita siten että $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^2$. Määritellään kullekin vektorille spatiaaliset merkit ja esitetään ne graafisesti.

$$a = \begin{pmatrix} 3 \\ -2 \end{pmatrix}, b = \begin{pmatrix} 1 \\ 4 \end{pmatrix}, c = \begin{pmatrix} -5 \\ 0 \end{pmatrix}$$

$$|a| = \sqrt{13} \ , \ \ |b| = \sqrt{17} \ , \ \ |c| = 5$$

$$\rightarrow \boldsymbol{U}(a) = \frac{1}{\sqrt{13}} \begin{pmatrix} 3 \\ -2 \end{pmatrix}, \ \boldsymbol{U}(b) = \frac{1}{\sqrt{17}} \begin{pmatrix} 1 \\ 4 \end{pmatrix}, \ \boldsymbol{U}(a) = \frac{1}{5} \begin{pmatrix} -5 \\ 0 \end{pmatrix}$$

2.3 Spatiaaliset järjestysluvut

Spatiaalinen järjestysluku on vektori, jonka suunta kertoo missä datapiste likimain sijaitsee suhteessa aineiston kuviteltuun keskipisteeseen. Spatiaalisen järjestysluvun pituus kertoo kuinka kaukana datapiste sijaitsee aineiston kuvitellusta keskipisteestä. Spatiaaliset järjestysluvut sijaitsevat siis aina p-ulotteisen yksikköpallon sisällä.

Määritelmä 2.10 (Spatiaaliset järjestysluvut)**.** Olkoon $\boldsymbol{x_1, x_2, ..., x_n}$ riippumaton p-ulotteinen satunnaisotos. Havaintopisteen $\boldsymbol{x_i}$ spatiaalinen järjestysluku $\boldsymbol{R(x_i)}$ määritellään hyödyntämällä spatiaalisien merkkien $\boldsymbol{U(x_i)}$ määritelmää 2.8, $\boldsymbol{R(x_i)} = AVE_j\{\boldsymbol{U(x_i - x_j)}\}$, $j = 1, 2, ..., i, ..., n$.

Merkinnällä AVE tarkoitetaan keskiarvoa. Spatiaalinen merkki siis on keskiarvo datapisteen ja muiden datapisteiden erotuksesta. Tästä seuraa, että $AVE_j\{\boldsymbol{R(x_j)}\}$=0.

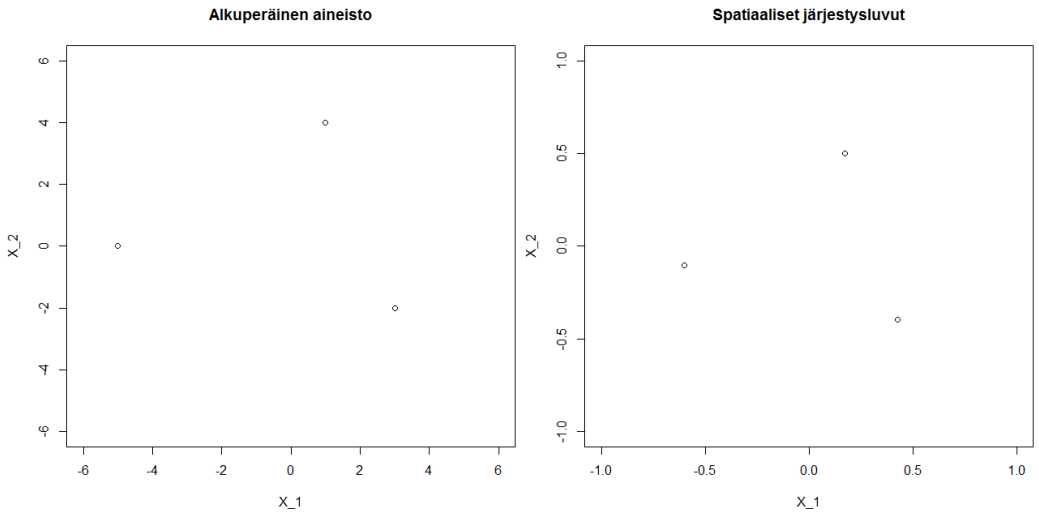
Spatiaaliset järjestysluvut ovat aina ortogonaalisesti ekvivariantteja siten että

$$\boldsymbol{R_{XO^T}(Ox_i) = OR_X(x_i)}, \tag{2.3.1}$$

kaikilla ortogonaalisilla matriiseilla $\boldsymbol{O}$ ja kaikilla $\boldsymbol{x_i}$. Spatiaalisten järjestyslukuihin perustuvan regression yhteydessä oletetaan oletuksen 2.9 pätevän. Kuten spatiaaliset merkit, myöskään spatiaaliset järjestysluvut eivät välttämättä ole täysin ekvivariantteja. Täysin ekvivariantit estimaatit saadaan luvun 3 algoritmilla.

Esimerkki 2.2 (Jatkoa esimerkille 2.1)**.** Olkoon $\boldsymbol{a, b}$ ja $\boldsymbol{c}$ satunnaisesti generoituja vektoreita siten että $\boldsymbol{a, b, c} \in \mathbb{R}^2$. Määritellään kullekin vektorille spatiaaliset järjestysluvut ja esitetään ne graafisesti kuvassa 3.

$$\begin{aligned}
a &= \begin{pmatrix} 3 \\ -2 \end{pmatrix}, b = \begin{pmatrix} 1 \\ 4 \end{pmatrix}, c = \begin{pmatrix} -5 \\ 0 \end{pmatrix} \\
U(a-b) &= \begin{pmatrix} 2 \\ -6 \end{pmatrix}, U(a-c) = \begin{pmatrix} 8 \\ -2 \end{pmatrix}, U(b-c) = \begin{pmatrix} 6 \\ 4 \end{pmatrix} \\
\rightarrow R(a) &= \frac{1}{3}\left(U(a-a) + U(a-b) + U(a-c)\right) \\
\rightarrow R(a) &\approx \begin{pmatrix} 0.429 \\ -0.397 \end{pmatrix}, R(b) \approx \begin{pmatrix} 0.172 \\ 0.501 \end{pmatrix}, R(c) \approx \begin{pmatrix} -0.601 \\ -0.104 \end{pmatrix}
\end{aligned}$$

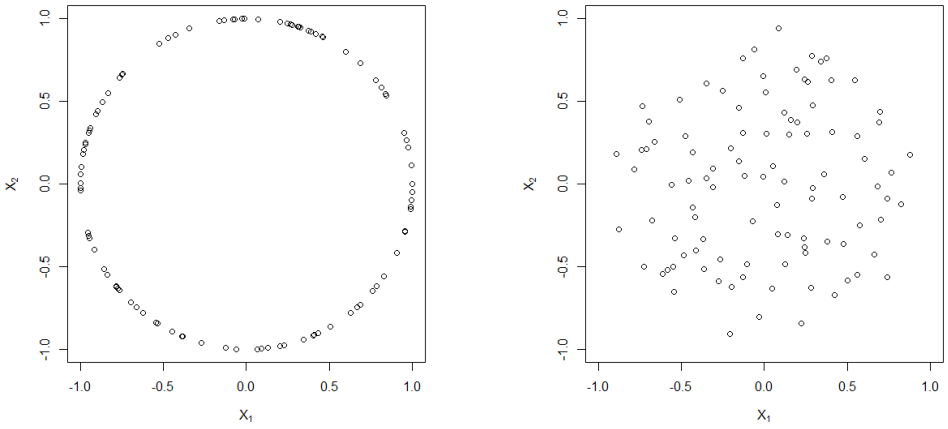


(a) Vektorit $\boldsymbol{a}, \boldsymbol{b}$ ja $\boldsymbol{c}$.

(b) Spatiaaliset järjestysluvut.

Kuva 3: Kuvasta nähdään spatiaalisten järjestyslukujen näyttävän havaintopisteiden suunnan ja suhteellisen etäisyyden dataparven keskipisteestä.

Kuvaan 4 on vielä havainnollistettu, miltä pistemääräfunktioina toimivat spatiaaliset merkit ja järjestysluvut näyttävät kahdessa dimensiossa normaalijakautuneelle aineistolle.



(a) Spatiaaliset merkit. (b) Spatiaaliset järjestysluvut.

Kuva 4: Spatiaaliset merkit ja järjestysluvut laskettu samasta normaalijakautuneesta aineistosta.

2.4 Pistemääräfunktio

Usein perinteisen regressioanalyysin työkaluna käytetään suurimman uskottavuuden menetelmää. Suurimman uskottavuuden estimaattorit saadaan maksimoimalla uskottavuusfunktiota $\boldsymbol{L(\theta; x)}$. Usein uskottavuusfunktion maksimia etsitään uskottavuusfunktion logaritmin avulla, sillä se on lähes aina derivoinnin kannalta yksinkertaisempaa, jolloin ääriarvojen etsiminen helpottuu. Työssä ei kuitenkaan oleteta, että aineisto noudattaisi mitään tunnettua jakaumaa. Työssä on myös mukana aineistoja, joihin on simuloitu merkittävä määrä poikkeavia havaintoja. Uskottavuusfunktiota ei voida siis työssä käyttää, sillä derivointia ei voida suorittaa ilman jakaumaoletuksia. Oletuksia voidaan väljentää käyttämällä uskottavuusfunktion logaritmin tilalla yleisempää pistemääräfunktiota.

Yleisesti käytetty tekniikka moniulotteisen datan analysoimiseksi on korvata alkuperäiset havoinnot $\boldsymbol{y_i}$ pistemääräfunktioilla $\boldsymbol{T_i = T(y_i)}$. Työssä pistemääräfunktioina käytetään perinteisen L_2-regression yhteydessä alkuperäisiä havaintoja. Lisäksi spatiaaliset merkit $\boldsymbol{U(y)}$ ja spatiaaliset järjestysluvut $\boldsymbol{R(y)}$ toimivat pistemääräfunktioina.

Estimoinnin ja testaamisen helpottamiseksi pistemääräfunktion arvot tulee standardoida ja keskittää jollain luonnollisella tavalla. Työssä käytetään si-

säistä standardointia, jota R-ohjelmiston MNM-paketti käyttää oletuksena spatiaalisten merkkien ja järjestyslukujen yhteydessä. Sisäisessa standardoinnissa etsitään aluksi symmetrinen kääntyvä matriisi $\boldsymbol{S}$ siten että

$$\text{jos } \hat{\boldsymbol{T}}_{\boldsymbol{i}} = \boldsymbol{T}(\boldsymbol{S}^{-1/2}\boldsymbol{y}_{\boldsymbol{i}}),\ \hat{\boldsymbol{T}} = (\hat{\boldsymbol{T}}_{\boldsymbol{1}}, ..., \hat{\boldsymbol{T}}_{\boldsymbol{n}}) \tag{2.4.1}$$

$$\rightarrow p\hat{\boldsymbol{T}}'\hat{\boldsymbol{T}} = tr(\hat{\boldsymbol{T}}'\hat{\boldsymbol{T}})\boldsymbol{I}_{\boldsymbol{p}}. \tag{2.4.2}$$

Merkinnällä tr tarkoitetaan matriisin jälkeä. Kun käytetään transformaatiota $\hat{\boldsymbol{T}}_{\boldsymbol{i}} = \boldsymbol{T}(\boldsymbol{S}^{-1/2}\boldsymbol{y}_{\boldsymbol{i}})$, testisuureeksi tulee

$$\boldsymbol{Q}^{\boldsymbol{2}} = n \cdot tr((\hat{\boldsymbol{T}}'\boldsymbol{P}_{\boldsymbol{x}}\hat{\boldsymbol{T}})(\hat{\boldsymbol{T}}'\hat{\boldsymbol{T}})^{-1}) = \frac{np}{tr(\hat{\boldsymbol{T}}'\hat{\boldsymbol{T}})}|\boldsymbol{P}_{\boldsymbol{x}}\hat{\boldsymbol{T}}|^2 \rightarrow_d \boldsymbol{\chi}^2_{pq}, \tag{2.4.3}$$

missä $\boldsymbol{P}_{\boldsymbol{X}}$ on

$$\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'. \tag{2.4.4}$$

Testisuure siis noudattaa usein asymptoottisesti jakaumaa $\boldsymbol{\chi}^2_{pq}$, missä p on muuttujien määrä ja q on vapausaste. Testisuure vertaa projisoidun ja transformaation kovarianssimatriisia standardoidun datan kovarianssimatriisiin.

Määritelmä 2.11 (Affiinisesti invariantti testisuure)**.** Testisuure on affiinisti invariantti, jos $\boldsymbol{Q}^{\boldsymbol{2}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Q}^{\boldsymbol{2}}(\boldsymbol{XV}, \boldsymbol{YW})$, missä $\boldsymbol{V} \in \mathbb{R}^{\boldsymbol{q}\times\boldsymbol{q}}$, $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{p}\times\boldsymbol{p}}$ ja $rank(\boldsymbol{V}) = q$, $rank(\boldsymbol{W}) = p$.

Affiinisesti invariatti testisuure antaa siis saman tuloksen lineaarisen transformaation jälkeen. Ominaisuus on tärkeä, jos aineistoa halutaan esimerkiksi skaalata tai tehdä lokaatiomuutoksia. Sisäisellä standardoinnilla saavutettu testisuure 2.4.3 on affiinisti invariantti.

2.5 Elliptinen jakauma

Monissa käytännön sovelluksissa normaalisuusoletukset eivät päde. Usein esimerkiksi rahoitukseen liittyvä aineisto on paksuhäntäinen suhteessa normaalijakaumaan. Elliptisesti jakautunut aineisto voi olla enemmän tai vähemmän paksuhäntäinen suhteessa normaalijakaumaan. Elliptisellä jakaumalla on kuitenkin samat symmetriaominaisuudet suhteessa normaalijakaumaan, usein muut standardioletukset paitsi normaalisuusoletus päteekin elliptisesti jakautuneelle aineistolle. Alaluvun teoria perustuu pääosin Davy Paindaveinen julkaisuun Elliptical symmetry [10].

Määritelmä 2.12 (Elliptinen jakauma)**.** d-ulotteinen vektori $\boldsymbol{x}$ on elliptisesti jakautunut, jos on olemassa vektori $\boldsymbol{\mu}$, täysiasteinen matriisi $\boldsymbol{A} \in \mathbb{R}^{d \times r}$ ja epänegatiivinen satunnaismuuttuja R, siten että

$$\boldsymbol{x} \overset{D}{=} \boldsymbol{\mu} + R\boldsymbol{A}\boldsymbol{u},$$

missä $\boldsymbol{u}$ on r-pituinen riippumaton vektori ja R on tasajakautunut muuttuja r-ulotteisen yksikköpallon pinnalta.

Merkinnällä $\overset{D}{=}$ tarkoitetaan yhtäsuuruutta jakauman suhteen. Eräs esimerkki elliptisestä jakaumasta on Studentin t-jakauma.

Määritelmä 2.13 (Moniulotteinen t-jakauma)**.** Olkoon $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ja $\boldsymbol{Y} \sim \boldsymbol{\chi^2_\nu}$ toisistaan riippumattomia. Tällöin sanotaan, että $\boldsymbol{U}$ noudattaa moniulotteista t-jakaumaa

$$\boldsymbol{U} \overset{D}{=} \frac{\boldsymbol{X}}{\sqrt{\frac{\boldsymbol{Y}}{\nu}}},$$

missä p-ulotteisen normaalijakauman odotusarvo on $\boldsymbol{\mu}$ ja kovarianssimatriisi $\boldsymbol{\Sigma}$ ja ν on $\boldsymbol{\chi^2_\nu}$-jakauman vapausaste [3].

Moniulotteinen t-jakauma täyttää aina elliptisen jakauman kriteerit. Todistus sivuutetaan työssä mutta se löytyy esimerkiksi Gabriel Frahmin väitöskirjasta [1].

3 Tutkimusongelma ja -menetelmät

Luvussa 3 esitellään menetelmät, joilla eri aineistot on simuloitu ja millainen on onnistunut regressio. Luvussa lisäksi käydään läpi, millaisilla eri menetelmillä ja algoritmeilla estimointi suoritetaan ja miten regression onnistumista testataan.

3.1 Aineiston simuloiminen

Aineiston simuloiminen toteutetaan R-ohjelmistolla. Työssä käytetään valmista pakettia MNM [7], johon sisältyy tarvittavat työkalut aineiston simuloimiseen sekä analysoimiseen. Työssä simuloidaan neljä erilaista aineistoa,

joita pyritään estimoimaan työssä aiemmin esitellyillä menetelmillä. Aineistoina ovat normaali- ja elliptinen jakauma. Molempiin lisäksi lisätään poikkeavia havaintoja, jolloin aineistojen kokonaismääräksi tulee neljä.

Poikkeavalla havainnolla tarkoitetaan havaintoa, joka eroaa merkittävästi muista havainnoista. Poikkavien havaintojen tunnistamiseen ei tarvitse työssä paneutua, sillä simulointivaiheessa ne luodaan tarkoituksenmukaisesti. Aineistoiden havaintojen kokonaismäärä on 50, joista poikkeavien havaintojen osuus on 20%. Poikkeavat havainnot on sijoitettu epäsymmetrisesti suhteessa muuhun aineistoon.

Aineisto sisältää kolme selitettävää muuttujaa, joita selittää neljä muuttujaa. Simuloinnissa noudatetaan seuraavia vaiheita

1. Valitaan matriisi $\boldsymbol{\beta} \in \mathbb{R}^{4\times 3}$.
2. Simuloidaan matriisit $\boldsymbol{X}$ ja $\boldsymbol{\varepsilon}$, $\boldsymbol{X} \in \mathbb{R}^{50\times 4}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^{50\times 3}$.
3. Muodostetaan matriisi $\boldsymbol{Y}$ siten että: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{Y} \in \mathbb{R}^{50\times 3}$.
4. Suoritetaan estimointi $\boldsymbol{Y} \sim \boldsymbol{X}$, eli selitetään matriisia $\boldsymbol{Y}$ matriisin $\boldsymbol{X}$ avulla.
5. Vertaillaan estimoitua $\hat{\boldsymbol{\beta}}$ todelliseen $\boldsymbol{\beta}$.

Estimoinnissa $\boldsymbol{X} \sim \boldsymbol{Y}$ saadaan estimaatti $\hat{\boldsymbol{\beta}}$. Estimaatti pyrkii selittämään lineaarisen riippuvuuden matriisien $\boldsymbol{X}$ ja $\boldsymbol{Y}$ välillä. Simuloinnilla voidaan havainnollistaa eri menetelmien toimivuutta erilaisten oletusten alla. Voidaan helposti vertailla eri menetelmien onnistumista, kun alkuperäinen $\boldsymbol{\beta}$ tunnetaan. Estimaatin avulla voidaan laskea sovitteet sisältävä matriisi $\hat{\boldsymbol{Y}}$,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{3.1.1}$$

Työssä käytetään kaikissa neljässä simuloinnissa samaa matriisia $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}. \tag{3.1.2}$$

Simuloinnissa vaiheessa 2. päätetään, millainen aineisto saadaan. Matriisit $\boldsymbol{X}$ ja $\boldsymbol{\varepsilon}$ simuloidaan, joko normaalijakaumasta tai t-jakaumasta. Normaalijakautunut aineisto saadaan simuloimalla molempien matriisien arvot moniulotteisesta normaalijakaumasta, jonka odotusarvo on nollavektori ja kovarianssimatriisi on identiteettimatriisi. Elliptinen aineisto taas saadaan moniulotteisesta Studentin t-jakaumasta, jonka vapausaste on 1 ja kovarianssimatriisi

on identiteettimatriisi. Poikkeavat havainnot sekä elliptiseen että normaalijakautuneeseen aineistoon saadaan simuloimalla jäännöstermimatriisiin arvoja moniulotteisesta normaalijakaumasta, jonka odotusarvo on kolmiulotteinen vektori $\boldsymbol{\mu} = (25, 25, 25)'$ ja kovarianssimatriisi on identiteettimatriisi. Simulointivaiheessa on varmistettu, että halutut oletukset pätevät. Oletukset eivät luonnollisesti päde poikkeaville havainnoille.

3.2 Perinteinen estimoiminen ja testaus

Alaluvun teoria perustuu pääpiirteiltään kurssin Ennustaminen ja aikasarja-analyysin luentokalvoihin [5], [4].

Lineaaristen regressiomallien estimoiminen edellyttää, että mallin rakenneosa on oikein spesifioitu. Onnistuneessa regressiossa jäännösneliösumma on pieni eli selitysaste on korkea. Perinteisessä L_2-regressiossa regressiomatriisi $\boldsymbol{\beta}$ estimoidaan pienimmän neliösumman (PNS) menetelmällä. PNS-menetelmässä regressiokertoimien estimaattori saadaan minimoimalla jäännöstermin neliösumma. Regressiokerroinmatriisin $\boldsymbol{\beta}$ PNS-estimaattori $\boldsymbol{B}$ ratkeaa suljetussa muodossa. Regressiokertoimet sisältävän matriisin $\boldsymbol{\beta}$ PNS estimaattori on,

$$\boldsymbol{B} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}. \tag{3.2.1}$$

Gaussin ja Markovin lauseen mukaan yleisen lineaarisen mallin paras estimaatti saadaan käyttämällä PNS-menetelmää, kun oletukset 2.2-2.6 ovat voimassa [4]. Paremmuuden kriteerinä käytetään varianssia, jolloin mahdollisimman pieni varianssi on toivottava ominaisuus. Estimaattori $\boldsymbol{B}$ tuottaa täysin ekvivariantteja estimaatteja $\hat{\boldsymbol{\beta}}$.

Määritelmä 3.1 (Täysin ekvivariantti estimaatti)**.**

Täysin ekvivariantilla estimaatilla $\hat{\boldsymbol{\beta}}$ on seuraavat ominaisuudet

1. Regressio ekvivarianttisuus: $\hat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{XH} + \boldsymbol{Y}) = \hat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{H}$,
   kaikilla $\boldsymbol{H} \in \mathbb{R}^{q \times p}$, $rank(\boldsymbol{H}) = min\{q, p\}$ eli $\boldsymbol{H}$ on täysiasteinen.
2. $\boldsymbol{Y}$ ekvivarianttisuus: $\hat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{YW}) = \hat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{W}$,
   kaikilla $\boldsymbol{W} \in \mathbb{R}^{p \times p}$, $rank(\boldsymbol{W}) = p$.
3. $\boldsymbol{X}$ ekvivarianttisuus: $\hat{\boldsymbol{\beta}}(\boldsymbol{XV}, \boldsymbol{Y}) = \boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y})$,
   kaikilla $\boldsymbol{V} \in \mathbb{R}^{q \times q}$, $rank(\boldsymbol{V}) = q$.

Täysin ekvivarianttiin estimaatin estimoimiseen ei siis vaikuta lineaarimuunnokset. Regressiokertoimien eli matriisin $\hat{\boldsymbol{\beta}}$ alkioiden merkitsevyyttä testataan tilastollisesti testaamalla nollahypoteesia

$$H_0 : \hat{\boldsymbol{\beta}}_{\boldsymbol{ji}} = 0. \tag{3.2.2}$$

Jos nollahypoteesi hylätään, selitettävän muuttujan ja selittävän muuttujan välillä oletetaan olevan lineaarinen riippuvuus $\hat{\boldsymbol{\beta}}_{\boldsymbol{ji}}$. Alaindeksillä i tarkoitetaan selitettävää muuttujaa ja indeksillä j selittävää muuttujaa. Yhden regressiokertoimen merkitsevyyttä voidaan testata t-testisuureella

$$t_{ji} = \frac{B_{ji}}{\hat{D}(B_{ji})}, \tag{3.2.3}$$

joka noudattaa t-jakaumaa vapausastein $(n - k - 1)$, n on havaintopisteiden lukumäärä ja k on selittäjien lukumäärä. PNS-estimaattorin B_{ji} varianssin harhaton estimaattori on $\hat{D}^2(B_{ji})$. Varianssin harhaton estimaattori voidaan laskea

$$\hat{D}^2(B_{ji}) = \frac{1}{n-k-1} \sum_{i=m}^{n} \boldsymbol{\varepsilon}_{\boldsymbol{m}}^{\boldsymbol{2}} \left((\boldsymbol{X'X})^{\boldsymbol{-1}} \right). \tag{3.2.4}$$

Testisuure mallin merkitsevyydelle, eli että vähintään yksi $\hat{\boldsymbol{\beta}}$ alkio on nollasta poikkeava, saadaan määrittämällä F-testisuure

$$\boldsymbol{F} = \frac{n-k-1}{k} \frac{R^2}{1-R^2}, \tag{3.2.5}$$

R^2 on mallin selitysaste. F-testisuure määritellään erikseen kaikille selitettäville muuttujille, joita on työssä kolme. F-testisuure noudattaa F-jakaumaa vapausastein k ja $(n - k - 1)$. Selitysaste määritellään seuraavasti

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{3.2.6}$$

n tarkoittaa havaintojen määrää, $\hat{y}$ sovitetta ja $\bar{y}$ selitettävän muuttujan havaintojen keskiarvoa. F- ja t-testisuureita vastaavat p-arvot voidaan katsoa tilastollisesta taulukosta, jonka perusteella hylätään tai hyväksytään nollahypoteesi. Erittäin pienet p-arvot tarkoittavat, että nollahypoteesi voidaan hylätä suuremmalla luottamustasolla. Jos p-arvo on pienempi kuin 0.01, pidetään riippuvuutta tilastollisesti merkitsevänä.

Huomioitavaa on, että F- ja t-testisuure eivät toimi halutulla tavalla, jos standardioletukset eivät ole voimassa. Työssä käsitellään myös aineistoja joissa standardioletukset eivät päde, joten F- ja t-testisuureita ei voida käyttää tilastollisen merkitsevyyden mittarina. Simuloidun aineiston analysoimisessa R-ohjelmisto tuottaa F-testit mallien merkitsevyydelle.

3.3 Spatiaalisiin merkkeihin perustuva estimoiminen ja testaaminen

Luvussa esitellään spatiaalisen merkkien estimointi- sekä testiongelma. Työssä paneudutaan ainoastaan affiinisesti invarianttiin testisuureeseen sekä täysin ekvivarianttiin estimaattoriin.

Estimaatti $\hat{\boldsymbol{\beta}}$ on ratkaisu minimoitavalle funktiolle 2.2.1

$$\boldsymbol{U}(\hat{\beta})'X = 0, \tag{3.3.1}$$

missä $\boldsymbol{U_i(\beta)} = \boldsymbol{U(y_i - \beta' x_i)}$, $i = 1, 2, ..., n$ ja $\boldsymbol{U(\beta)} = (\boldsymbol{U_1(\beta)}, ..., \boldsymbol{U_n(\beta)})'$. Estimaattoria ei pystytä yleensä ratkaisemaan suljetussa muodossa. Ongelma voidaan kuitenkin ratkaista erilaisilla algoritmeilla, kuten esimerkiksi kolmen askeleen transformaatio-retranformaatiotekniikalla

1. $\boldsymbol{\varepsilon_i = S^{-1/2}(y_i - \beta' x_i)}, \quad i = 1, ..., n.$
2. $\boldsymbol{\beta = \beta + \left(AVE(|\varepsilon_i|^{-1} x_i x_i')\right)^{-1} AVE(x_i U(\varepsilon_i)') S^{1/2}}.$
3. $\boldsymbol{S = pS^{1/2} AVE(U(\varepsilon_i) U(\varepsilon_i)') S^{1/2}}.$

Algoritmi päivittää ensin residuaalit, sitten matriisin $\boldsymbol{\beta}$ ja lopuksi residuaalien hajontamatriisin $\boldsymbol{S}$. Edellä $\boldsymbol{S}$ ja $\boldsymbol{S^{1/2}}$ ovat symmetrisiä matriiseja ja $\boldsymbol{S = S^{1/2} S^{1/2}}$. Algoritmin tuottamat estimaatit ovat täysin ekvivariantteja eli ne toteuttavat määritelmän 3.1 kriteerit.

Affiinisesti invariantti testisuure saadaan kaavasta 2.4.3 korvaamalla $\boldsymbol{T(y)}$ spatiaalisilla merkeillä. Affiinisesti invariantiksi testisuureeksi saadaan

$$\boldsymbol{Q^2(X, Y)} = p \cdot tr\{\boldsymbol{U(\hat{\beta}') P_X U(\hat{\beta})}\} = p|\boldsymbol{P_X U(\hat{\beta})}|^2 \rightarrow_d \boldsymbol{\chi^2_{pq}}. \tag{3.3.2}$$

Merkinnät ovat luvun 2.4 mukaiset.

3.4 Spatiaalisiin järjestyslukuihin perustuva estimoiminen ja testaaminen

Kuten spatiaalisten merkkien yhteydessä, alaluvussa esitellään ainoastaan täysin ekvivariantit estimaatit sekä affiinisesti invariantit testisuureet.

Estimaatti $\hat{\boldsymbol{\beta}}$ on ratkaisu funktiolle

$$\boldsymbol{R}(\hat{\boldsymbol{\beta}})'\boldsymbol{X} = \boldsymbol{0}, \tag{3.4.1}$$

missä $\boldsymbol{R}(\beta) = (\boldsymbol{R_1}(\boldsymbol{\beta}), ..., \boldsymbol{R_n}(\boldsymbol{\beta}))'$.

1. $\varepsilon_{ij} = \boldsymbol{S}^{-1/2}(y_{ij} - \boldsymbol{\beta}' x_{ij}), \quad i, j = 1, ..., n.$
2. $\boldsymbol{\beta} = \boldsymbol{\beta} + \left(AVE(|\varepsilon_{ij}|^{-1} x_{ij} x_{ij}')\right)^{-1} AVE(x_{ij}\boldsymbol{U}(\varepsilon_{ij})')\boldsymbol{S^{1/2}}.$
3. $\boldsymbol{S} = p\boldsymbol{S^{1/2}} AVE(U(\varepsilon_{ij})U(\varepsilon_{ik})')\boldsymbol{S^{1/2}}.$

Algoritmin toimintaperiaate on sama kuin spatiaalisten merkkien transformaatioretransformaatiotekniikalla ja saadaan täysin ekvivariantit estimaatit.

Affiinisesti invariantti testisuure saadaan kaavalla 2.4.3, nyt $\boldsymbol{T(y)}$ korvataan spatiaalisilla järjestysluvuilla. Affiinisesti invariantti testisuure on

$$\boldsymbol{Q^2(X, Y)} = \frac{np}{tr\{\boldsymbol{R}(\hat{\boldsymbol{\beta}})'\boldsymbol{R}(\hat{\boldsymbol{\beta}})\}} |\boldsymbol{P_X R}(\hat{\boldsymbol{\beta}})|^2 \rightarrow_d \boldsymbol{\chi}^2_{pq}. \tag{3.4.2}$$

Merkinnät ovat luvun 2.4 mukaiset.

4 Tulokset

Luvussa esitetään eri menetelmien estimaatit eri aineistoille. Menetelmien toimivuutta arvioidaan vertaamalla alkuperäistä $\boldsymbol{\beta}$ (3.1.2) matriisia estimaattiin $\hat{\boldsymbol{\beta}}$. Lisäksi sovitteita voidaan kuvata graafisesti selitettävien muuttujien funktiona – täysin onnistuneessa estimoinnissa kuvaan muodostuisi yhden tietyn suoran pisteitä. Tulosten yhteydessä ei tarvitse syvempää regressiodiagnostista tarkastelua, sillä aineisto on luotu simuloimalla. Kaikki työn testisuureet antavat p-arvot, jotka ovat pienempiä kuin 0.001. Testisuureen arvoihin ei voida kuitenkaan luottaa, sillä kaikki aineistot eivät toteuta tarvittavia oletuksia.

Taulukko 1: Normaalijakautunut aineisto, taulukossa vertailtu $\boldsymbol{\beta}$ alkioita estimaatin $\hat{\boldsymbol{\beta}}$ alkioihin. Poikkeamien keskiarvo on laskettu ottamalla estimaatin ja alkuperäisen matriisin alkioiden erotuksen itseisarvo, joista on laskettu keskiarvo.

	Poikkeamien keskiarvo	Max poikkeama
L_2	0.11	0.20
Spatiaaliset merkit	0.15	0.43
Spatiaaliset järjestysluvut	0.12	0.29

4.1 Normaalijautunut aineisto

Normaalijakautuneessa aineistossa selittävät muuttujat sisältävä matriisi $\boldsymbol{X}$ noudattaa neliulotteista normaalijakaumaa ja jäännöstermit sisältävä matriisi $\boldsymbol{\varepsilon}$ noudattaa kolmiulotteista normaalijakaumaa ja molemissa jakaumissa odotusarvo on moniulotteinen nollavektori ja kovarianssimatriisi identiteettimatriisi. Havaintoja on yhteensä 50. Saadaan seuraavanlaiset estimaatit käyttäen eri menetelmiä

$$\hat{\boldsymbol{\beta}_{\boldsymbol{L_2}}} = \begin{pmatrix} 0.99 & 2.17 & 3.05 \\ 3.97 & 4.99 & 6.12 \\ 6.81 & 8.17 & 9.07 \\ 10.20 & 10.84 & 11.86 \end{pmatrix}, \hat{\boldsymbol{\beta}_{\boldsymbol{SM}}} = \begin{pmatrix} 0.98 & 2.13 & 3.12 \\ 3.94 & 5.08 & 6.02 \\ 6.57 & 8.17 & 9.22 \\ 10.13 & 10.77 & 11.84 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}_{\boldsymbol{SJ}}} = \begin{pmatrix} 0.99 & 2.16 & 3.09 \\ 3.97 & 5.02 & 6.09 \\ 6.71 & 8.16 & 9.14 \\ 10.15 & 10.81 & 11.85 \end{pmatrix}.$$

Alaindeksillä L_2 tarkoitetaan perinteisen regression, SM spatiaalisten merkkien ja SJ spatiaalisten järjestyslukujen tuottamaa estimaattia. Taulukosta 1 nähdään, että perinteinen L_2-regressio tuottaa parhaan estimaatin kun aineisto on normaalijakautunut. Tulos on sopusoinnussa Gaussin ja Markovin lauseen kanssa [4], sillä oletukset 2.2-2.6 pätevät. Spatiaaliset järjestysluvut antavat tässä tapauksessa paremman estimaatin suhteessa spatiaalisiin merkkeihin.

Taulukko 2: Normaalijakautunut aineisto poikkeavilla havainnoilla, taulukossa vertailtu $\boldsymbol{\beta}$ alkioita estimaatin $\hat{\boldsymbol{\beta}}$ alkioihin. Poikkeamien keskiarvo on laskettu $AVE\{|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}|\}$.

	Poikkeamien keskiarvo	Max poikkeama
L_2	0.69	1.66
Spatiaaliset merkit	0.09	0.26
Spatiaaliset järjestysluvut	0.18	0.37

4.2 Normaalijakautunut aineisto poikkeavilla havainnoilla

Poiketen luvun 4.1 aineistosta, nyt 10 jäännöstermimatriisin alkiota on simuloitu normaalijakaumasta, jonka odotusarvo on vektori $\mu = (25, 25, 25)'$. Havaintojen kokonaismäärä on yhä 50. Muuten aineisto vastaa täysin luvun 4.1 aineistoa. Poikkeavien havaintojen seurauksena, oletukset 2.2-2.6 eivät ole enää kaikki voimassa, jolloin Gaussin ja Markovin lause ei päde. Saadaan seuraavanlaiset estimaatit

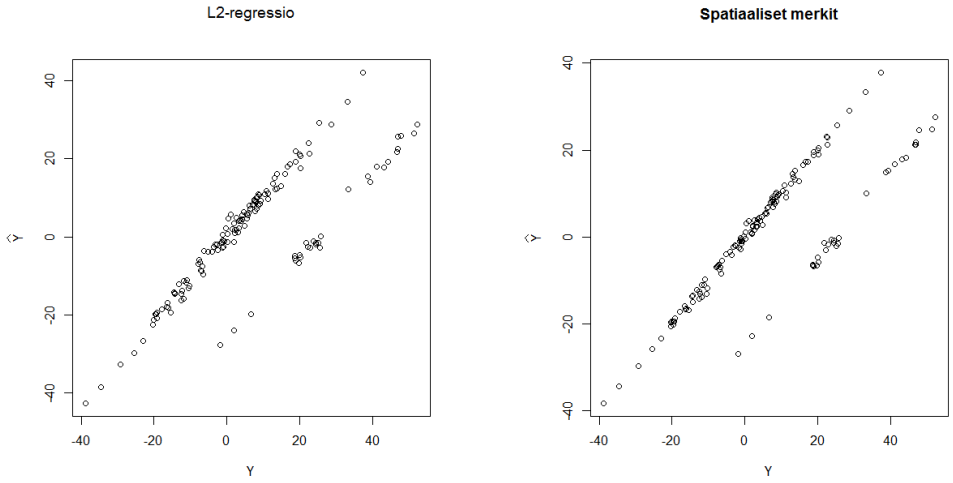
$$\hat{\boldsymbol{\beta}_{\boldsymbol{L_2}}} = \begin{pmatrix} 0.48 & 1.93 & 3.71 \\ 4.54 & 5.24 & 5.54 \\ 7.86 & 9.49 & 10.66 \\ 9.71 & 11.07 & 13.33 \end{pmatrix}, \hat{\boldsymbol{\beta}_{\boldsymbol{SM}}} = \begin{pmatrix} 1.02 & 1.91 & 3.09 \\ 3.93 & 5.02 & 5.98 \\ 7.20 & 8.26 & 9.20 \\ 9.99 & 11.06 & 12.07 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}_{\boldsymbol{SJ}}} = \begin{pmatrix} 1.11 & 2.01 & 3.37 \\ 3.82 & 4.86 & 5.83 \\ 7.21 & 8.30 & 9.31 \\ 9.98 & 11.12 & 12.24 \end{pmatrix}.$$

Kun mukana on poikkeavia havaintoja, perinteinen L_2-regressio tuottaa selvästi huonoimman estimaatin. Perinteinen regressio epäonnistuu suhteessa spatiaalisiin merkkeihin ja järjestyslukuihin, sillä se on huomattavasti vähemmän robusti menetelmä. Tulokset nähdään taulukosta 2. Kuvassa 5 on vertailtu perinteisen regression estimaattia spatiaalisten merkkien estimaattiin. Kuvia vertailemalla huomataan, että spatiaalisten merkkien sovitteet asettuvat suoralle paremmin.

4.3 Elliptisesti jakautunut aineisto

Studentin t-jakauma on esimerkki eräästä elliptisestä jakaumasta. Kuten luvussa 2.5 todettiin, t-jakauma on aina elliptinen. Matriisit $\boldsymbol{X}$ ja $\boldsymbol{\varepsilon}$ ovat si-



(a) L_2-regressio normaalijakautuneeseen aineistoon poikkeavilla havainnoilla.

(b) Spatiaaliset merkit samaan aineistoon.

Kuva 5: Sovitteet alkuperäisten selitettävien funktiona, kuvassa on kaikki matriisin $\boldsymbol{Y}$ dimensiot mukana.

muloitu moniulotteisesta t-jakaumasta, jonka vapausaste on 1 ja kovarianssimatriisi on identiteettimatriisi. Havaintoja on yhteensä 50. Jos t-jakauman vapausaste olisi ollut merkittävästi suurempi, olisi se muistuttanut enemmän normaalijakaumaa. Pienempi vapausaste tuo paremmin menetelmien väliset erot esiin. Estimaateiksi saadaan

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L_2}} = \begin{pmatrix} -0.84 & 1.65 & 0.37 \\ 4.66 & 5.22 & 6.96 \\ 7.49 & 8.21 & 9.46 \\ 10.90 & 11.24 & 13.35 \end{pmatrix}, \hat{\boldsymbol{\beta}}_{\boldsymbol{SM}} = \begin{pmatrix} 0.75 & 2.11 & 2.78 \\ 4.02 & 4.97 & 6.13 \\ 7.08 & 7.96 & 9.05 \\ 10.16 & 10.97 & 12.10 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{SJ}} = \begin{pmatrix} 0.64 & 2.08 & 2.77 \\ 4.09 & 4.98 & 6.14 \\ 7.13 & 7.98 & 9.06 \\ 10.17 & 10.97 & 12.11 \end{pmatrix}.$$

Poikkeamien keskiarvot on laskettu taulukkoon 3. Tuloksista nähdään, että perinteinen regressio antaa merkittävästi huonoimman estimaatin aineistolle. Syy estimaatin huonouteen johtuu t-jakauman paksuhäntäisyydestä. Vertailtaessa normaalijakaumaan, t-jakauma vapausasteella 1 on merkittävästi

Taulukko 3: Elliptisesti jakautunut aineisto, taulukossa vertailtu $\boldsymbol{\beta}$ alkioita estimaatin $\hat{\boldsymbol{\beta}}$ alkioihin. Poikkeamien keskiarvo on laskettu $AVE\{|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}|\}$.

	Poikkeamien keskiarvo	Max poikkeama
L_2	0.86	2.63
Spatiaaliset merkit	0.10	0.25
Spatiaaliset järjestysluvut	0.12	0.36

paksuhäntäisempi.

Spatiaalisten merkkien ja järjestyslukujen tuottamat estimaatit ovat hyvin lähellä toisiaan. Tulosten perusteella tulisi ehdottomasti käyttää spatiaalisia merkkejä tai järjestyslukuja elliptisesti jakautuneelle aineistolle.

4.4 Elliptisesti jakautunut aineisto poikkeavilla havainnoilla

Poiketen luvun 4.3 aineistosta, nyt 10 jäännöstermimatriisin alkiota on simuloitu normaalijakaumasta, jonka odotusarvo on vektori $\boldsymbol{\mu} = (25, 25, 25)'$. Havaintojen kokonaismäärä 50, kuten muissakin aineistoissa. Matriisit $\boldsymbol{X}$ ja $\boldsymbol{\varepsilon}$ noudattavat siis muuten t-jakaumaa. Saamme seuraavanlaiset estimaatit

$$\hat{\boldsymbol{\beta}_{\boldsymbol{L_2}}} = \begin{pmatrix} 1.27 & 2.17 & 3.04 \\ 4.75 & 5.53 & 6.36 \\ 5.88 & 7.20 & 8.49 \\ 9.83 & 10.92 & 12.05 \end{pmatrix}, \hat{\boldsymbol{\beta}_{\boldsymbol{SM}}} = \begin{pmatrix} 1.03 & 2.03 & 3.04 \\ 4.39 & 5.38 & 6.37 \\ 6.45 & 7.46 & 8.46 \\ 10.12 & 11.13 & 12.16 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}_{\boldsymbol{SJ}}} = \begin{pmatrix} 1.07 & 2.05 & 3.06 \\ 4.42 & 5.40 & 6.37 \\ 6.39 & 7.42 & 8.44 \\ 10.10 & 11.11 & 12.16 \end{pmatrix}.$$

Tulokset eri estimaattien onnistumisesta ovat taulukossa 4. Perinteinen regressio tuottaa paremman estimaatin elliptisesti jakautuneeseen aineistoon, kun mukana on poikkeavia havaintoja. Vaikka poikkeavat havainnot ovat sijoiteltu epäsymmetrisesti suhteessa muuhun aineistoon, korjaantuu perinteisen regression estimaatti suhteessa tilanteeseen, missä poikkeavia havaintoja ei ole. Parantumisen syy on perinteisen regression herkkyydessä poikkeaville havainnoille.

Taulukko 4: Elliptisesti jakautunut aineisto poikkeavilla havainnoilla, taulukossa vertailtu $\boldsymbol{\beta}$ alkioita estimaatin $\hat{\boldsymbol{\beta}}$ alkioihin. Poikkeamien keskiarvo on laskettu $AVE\{|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}|\}$.

	Poikkeamien keskiarvo	Max poikkeama
L_2	0.40	1.12
Spatiaaliset merkit	0.27	0.55
Spatiaaliset järjestysluvut	0.29	0.61

Spatiaaliset merkit ja järjestysluvut antavat huonommat estimaatit elliptisesti jakautuneelle aineistolle, kun mukana on poikkeavia havaintoja. Estimaatit ovat taas hyvin lähellä toisiaan, mutta spatiaalisten merkkien estimaatti on aavistuksen parempi.

5 Johtopäätökset

Työssä tavoitteena oli vertailla eri menetelmien onnistumista lineaarisen mallin regressiokertoimien estimoimisessa erilaisille simuloiduille aineistoille. Aiemmat luvut antavat lukijalle lisäksi valmiudet käyttää työssä esiteltyjä menetelmiä.

Taulukkoon 5 on kerätty eri menetelmät paremmuusjärjestyksessä. Poikkeavien havaintojen ja eri jakaumaoletusten seurauksena vertailua onnistumisesta tulisi tehdä ainoastaan aineistokohtaisesti.

Normaalijakaumaoletusten vallitessa L_2-regressio tuottaa parhaan estimaatin. Tulos on sopusoinnussa Gaussin ja Markovin lauseen kanssa. Toisaalta perinteinen regressio tuottaa selvästi huonoimman estimaatin muille aineistoille. Perinteinen regressio on siis erittäin herkkä poikkeaville havainnoille ja aineiston paksuhäntäisyydelle. Kuitenkin perinteistä L_2-regressiota tulisi ehdottomasti käyttää, jos voidaan varmistaa normaalisuusoletusten olevan voimassa.

Tulokset osoittavat spatiaalisten merkkien ja järjestyslukujen robustin luonteen. Spatiaaliset merkit tuottavat parhaat estimaatit muille paitsi normaalijakautuneelle aineistolle. Spatiaaliset järjestysluvut taas antavat kaikille aineistoille toiseksi parhaat estimaatit. Tulosten perusteella spatiaalisten merkkien ja järjestyslukujen paremmuutta on vaikea verrata. Spatiaaliset järjestysluvut määritellään spatiaalisten merkkien avulla, mikä selittää molempien menetelmien samankaltaiset tulokset. Spatiaaliset järjestysluvut antavat li-

Taulukko 5: Eri menetelmät paremmuusjärjestyksessä eri aineistoissa.

	Norm.	Norm. ja poikkeavat	Ell.	Ell. ja poikkeavat
L_2	1.	3.	3.	3.
Spatiaaliset merkit	3.	1.	1.	1.
Spatiaaliset järjestysluvut	2.	2.	2.	2.

säksi paremman estimaatin normaalijakautuneelle aineistolle suhteessa spatiaalisiin merkkeihin.

Kun dimensioiden määrä kasvaa ja aineiston jakaumaoletuksia ei tunneta, robustien menetelmien merkitys kasvaa. Kahdessa dimensiossa visualisointi on vielä mahdollista, jolloin regressiografiikan avulla voidaan etsiä poikkeavia havaintoja ja tutkia jakaumaoletusten paikkansapitävyyttä. Moniulotteiselle aineistolle pystytään harvoin tuottamaan regressiodiagnostiikkaa tukevaa regressiografiikkaa, jolloin eräs vaihtoehto ongelman sivuuttamiseksi on robustien menetelmien käyttö.

Viitteet

[1] G. Frahm. *Generalized Elliptical Distributions: Theory and Applications.* Inauguraldissertation zur Erlangung des Doktorgrades der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität zu Köln, 2004.

[2] J. Kinnunen ja K. Peltonen. *Matematiikan peruskurssi L3* kurssin luentomateriaali. https://noppa.aalto.fi/noppa/kurssi/mat-1.1030/luennot/ , luennoitu 5.11.2012, luettu 29.4.2014

[3] K.V. Mardia, J.T Kent ja J.M. Bibby *Multivariate Analysis.* Academic Press, 2003.

[4] I. Mellin. *Ennustaminen ja aikasarja-analyysi* kurssin luentomateriaali. https://noppa.aalto.fi/noppa/kurssi/ms-c2128/luennot, luento Yleinen lineaarinen malli, luennoitu 28.10.2013, luettu 29.4.2014

[5] I. Mellin. *Ennustaminen ja aikasarja-analyysi* kurssin luentomateriaali. https://noppa.aalto.fi/noppa/kurssi/ms-c2128/luennot, luento regressiodiagnostiikka, luennoitu 04.11.2013, luettu 29.4.2014

[6] I. Mellin. *Ennustaminen ja aikasarja-analyysi* kurssin lisämateriaali. https://noppa.aalto.fi/noppa/kurssi/ms-c2128/materiaali, B2 Matriisilaskentaa tilastotieteilijöille, osa 2 (moniste), luettu 29.4.2014

[7] K. Nordhausen, J. Mottonen ja H. Oja. *MNM: Multivariate Nonparametric Methods. An Approach Based on Spatial Signs and Ranks.* http://cran.r-project.org/web/packages/MNM/

[8] K. Nordhausen ja H. Oja *Multivariate L_1 Methods: The Package MNM* Journal of Statistical Software, heinäkuu 2011

[9] H. Oja. *Multivariate Nonparametric Methods with R.* Springer, 2010

[10] D. Paindaveine. *Elliptical symmetry.* Encyclopedia of Environmetrics, 2nd edition, A. H. El-Shaarawi and W. Piegorsch (eds). John Wiley & Sons Ltd, Chichester, UK, 802-807 (2012).