Aalto-yliopisto
Perustieteiden
korkeakoulu

Spatiaalisiin merkkeihin ja järjestyslukuihin perustuva moniulotteinen regressioanalyysi R-ohjelmistoa käyttäen (valmiin työn esittely)

Niko Lietzén

28.4.2014

Ohjaaja/valvoja: apulaisprof. Pauliina Ilmonen

Systeemianalyysin
laboratorio

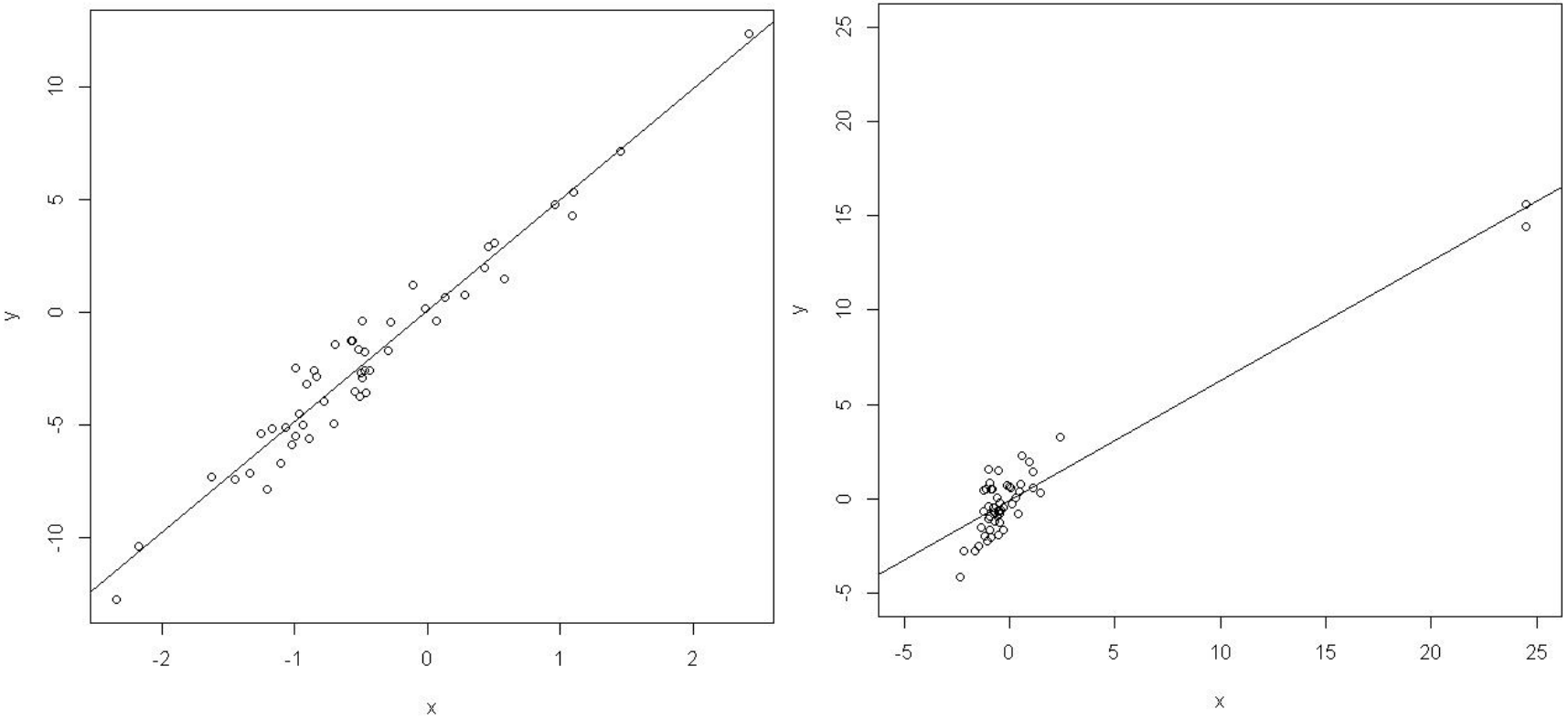
Esityksen Sisältö

- Tausta
- Tavoitteet
- Menetelmät
  - Pistemääräfunktio
  - Simuloiminen
- Tulokset
- Ongelmat
- Johtopäätökset
- Tietolähteet

Tausta

- Lineaarinen regressio erittäin laajasti käytetty esim. biologiassa, taloustieteessä ja rahoituksessa
- Harvoin täydellisesti tiettyä jakaumaa noudattavaa aineistoa, usein poikkeavia havaintoja
  - Vääristyneen regressioanalyysin mahdollisuus
- Robustit menetelmät eivät niin herkkiä poikkeaville havainnoille

Tausta

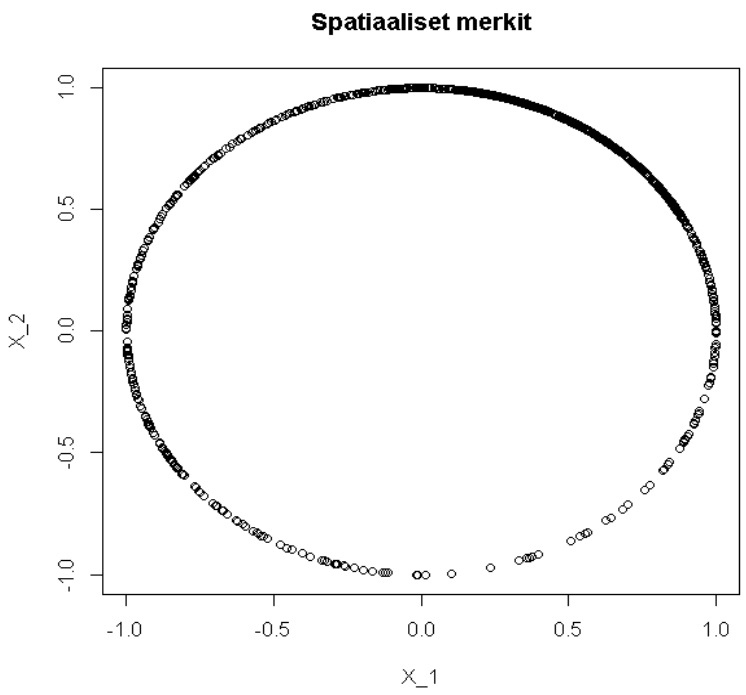


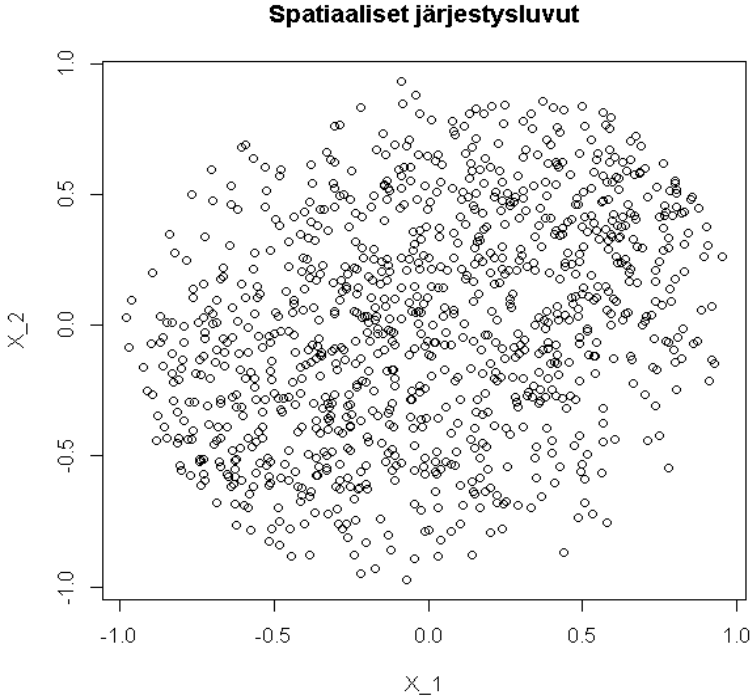
- Perinteinen L_2-regressio erittäin herkkä poikkeaville havainnoille

Tausta

- Yleinen lineaarinen malli: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

- Spatiaaliset merkit ja järjestysluvut





Tavoitteet

- Vertaillaan perinteisen L_2-regression, spatiaalisten merkkien ja spatiaalisten järjestyslukujen menetelmien toimivuutta simuloituun moniulotteiseen aineistoon
- Esitellään menetelmät ja niihin liittyvät oletukset ja algoritmit
- Tavallista yleisempi lähestymistapa estimoimiseen ja testaamiseen:
  - Pistemääräfunktio
  - Affiinisesti invariantit testisuureet
  - Täysin ekvivariantit estimaattorit

Pistemääräfunktio

- Suurimman uskottavuuden menetelmä, minimoidaan uskottavuusfunktio $\boldsymbol{L(\theta;x)}$
  - Logaritmin derivoiminen on usein yksinkertaisempaa
- Väljennetään oletuksia käyttämällä yleisempää pistemääräfunktiota
  - Uskottavuusfunktio on eräs mahdollinen pistemääräfunktio
- Korvataan alkuperäiset havainnot y_i spatiaalisilla merkeillä $\boldsymbol{U}(y_i)$ tai spatiaalisilla järjestysluvuilla $\boldsymbol{R}(y_i)$

Simuloiminen

1. Valitaan matriisi $\boldsymbol{\beta}$
2. Simuloidaan matriisit **X** ja $\boldsymbol{\varepsilon}$
3. Muodostetaan matriisi **Y** siten että $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
4. Estimoidaan $\mathbf{Y} \sim \mathbf{X}$
5. Vertaillaan esimaattia $\hat{\boldsymbol{\beta}}$ ja alkuperäistä $\boldsymbol{\beta}$

- Neljä erilaista aineistoa: normaali- ja elliptinen jakauma, lisäksi molempiin lisätty poikkeavia havaintoja (20%)
- $\mathbf{Y}, \boldsymbol{\varepsilon} \in R^{50x3}$, $\mathbf{X} \in R^{50x4}$ ja $\boldsymbol{\beta} \in R^{4x3}$
  - Moniulotteisuuden tarvittavat ominaisuudet näkyvät

Tulokset

- Keskimääräinen ero estimaatin $\hat{\boldsymbol{\beta}}$ ja alkuperäisen $\boldsymbol{\beta}$ eri alkioiden välillä:

	Norm. jak	Norm. jak + poikkeavat	Ell. jak.	Ell. jak. + poikkeavat
L2	**0.11**	0.69	0.86	0.40
Spat. Merkit	0.15	**0.09**	**0.10**	**0.27**
Spat. Järjestysluvut	0.12	0.18	0.12	0.29

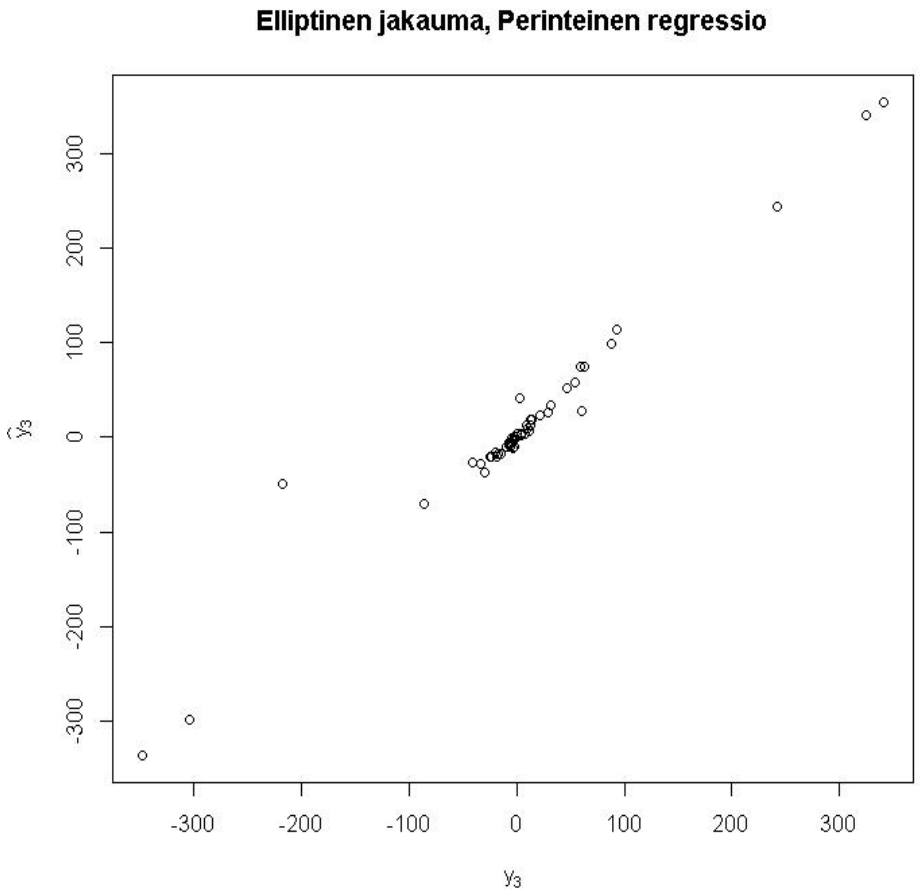
- Maksimaalinen ero estimaatin $\hat{\boldsymbol{\beta}}$ ja alkuperäisen $\boldsymbol{\beta}$ eri alkioiden välillä:

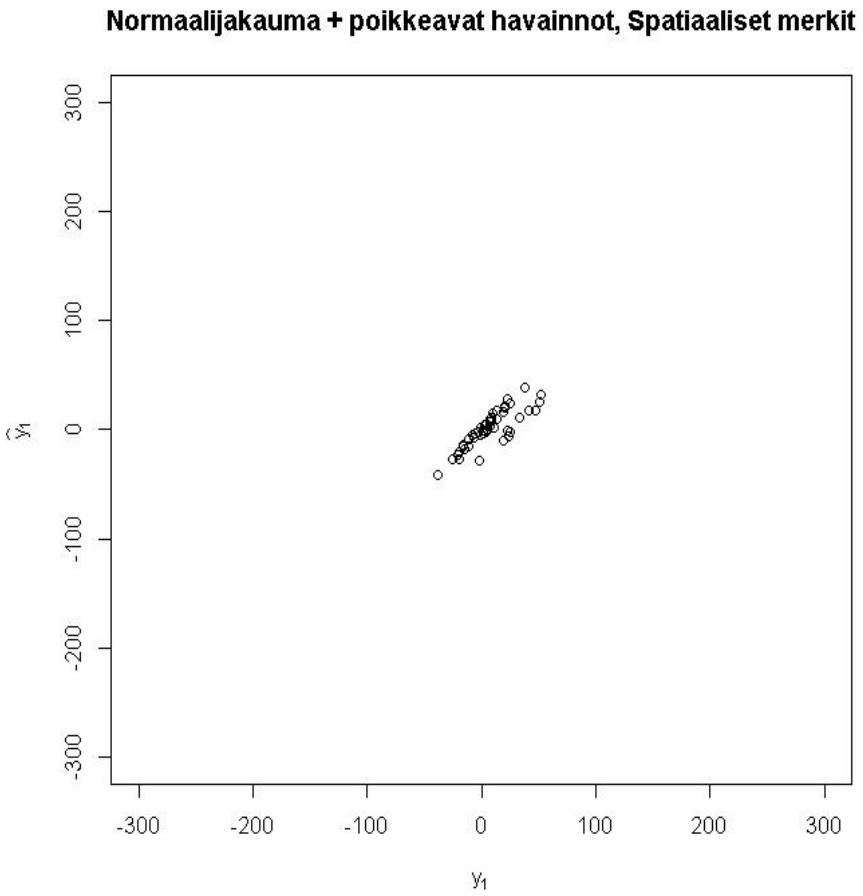
	Norm. jak	Norm. jak + poikkeavat	Ell. jak.	Ell. jak. + poikkeavat
L2	**0.20456**	1.662	2.6261	1.1249
Spat. Merkit	0.43	**0.26**	**0.25**	**0.55**
Spat. Järjestysluvut	0.293	0.37	0.363	0.61

- Simulointivaiheessa varmistettu haluttavien jakaumaoletusten voimassaolo

Tulokset

- Sovitteet: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\,\widehat{\boldsymbol{\beta}}$





Ongelmat

- Tavallisesti ei voida olla varmoja, mitkä jakaumaoletukset pätevät

- Moniulotteisen datan visualisointi
  - Poikkeavien havaintojen havaitseminen

Johtopäätökset

- Spatiaalisten merkkien ja järjestyslukujen robustinen luonne näkyy tuloksissa
- Perinteinen L_2-regressio toimii parhaiten normaalijakaumaoletusten vallitessa
- Spatiaalisten merkkien ja järjestyslukujen menetelmät parempia, jos elliptinen jakauma tai poikkeavia havaintoja

Tietolähteet

- Journal of Statistical Software: Multivariate L_1 Methods: The package MNM
  - Nordhausen & Oja, heinäkuu 2011
- MNM-paketin esimerkit ja ohjeet
  - http://cran.r-project.org/web/packages/MNM/MNM.pdf
- Multivariate Analysis
  - Mardia, Kent & Bibby, 2003
- Multivariate Nonparametric Mehtods with R
  - Oja, 2010