M-estimaattorit

Kristian Vähäkuopus

Kandidaatintutkielma
Toukokuu 2025

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Kristian Vähäkuopus: M-estimaattorit
Kandidaatin-tutkielma, 22 s., 1 liites.
Tilastotiede
Toukokuu 2025

Tämä tutkielma pohjautuu Rand R. Wilcoxin teokseen Introduction to Robust Estimation and Hypothesis Testing, joka esittää eri tapoja mallintaa tilastoaineistoa ja käsitellä poikkeuksellisia havaintoja. Tämän tutkielman tarkoitus on tutustua tavallisiin havaintoaineston estimaatteihin ja laajentaa tarkastelua robusteihin estimaatteihin eli estimaatteihin, jotka eivät ole kohtuuttoman herkkiä ääriarvoille ja poikkeuksellisille havainnoille.

M-estimaattorit ovat suurimman uskottavuuden estimaattoreita, jotka lasketaan havaintoaineistosta, ja niiden tehokkuus erityisesti normaalijakaumaoletuksen vallitessa voidaan osoittaa hyväksi. Tässä tutkielmassa tutustutaan yleisiin M-estimaattoreihin ja niiden käyttöön. M-estimaattorit sopivat hyvin lineaarisen regression mallintamiseen eli ongelmiin jotka kuvaavat jonkin selittävän muuttujan kasvaesssa sen lineaarista vaikutusta vastemuuttuujaan. Yksinkertainen esimerkki tällaisesta mallista olisi työvuosien vaikutus jonkin mielenkiinnon kohteena olevan populaation vuosituloihin.

Tilastotieteissä merkittävä ongelma havaintoestimaatteja laskettaessa ovat poikkeukselliset havaintoarvot ja havaintoaineistot, joihin normaalijakaumaoletukset eivät päde. Paksuhäntäiset ja vinot jakaumat aiheuttavat ongelmia perinteisille ei-robusteille estimaateille, joissa poikkeukselliset arvot voivat siirtää estimaattia pois populaation todellisesta arvosta. Robustit estimaatit pyrkivät vastaamaan tähän ongelmaan asettamalla rajoja poikkeuksellisten arvojen vaikutukselle poistamatta kuitenkaan näitä arvoja aineistos-

ta. Näillä estimaateilla vältytään menettämästä mahdollisesti tärkeätä tietoa arvojen poistamisen vuoksi, sekä vältytään käyttämästä aikaa aineiston poikkeuksellisten arvojen analysointiin.

Sisällys

1 Peruskäsitteitä

Määritelmä 1.1. Lokaatio-estimaattori on määrre, joka kuvaa populaatiota tai jakaumaa. Sillä on vähintään neljä ehtoa, joiden täytyy toteutua. Nämä ehdot ovat

$$f(X+a) = f(X) + a \tag{1}$$

$$f(-X) = -f(X) \tag{2}$$

$$X \geq 0 \implies f(X) \geq 0 \tag{3}$$

$$f(aX) = af(X). \tag{4}$$

Ehdot (1),(2),(3) määrittävät, että lokaatioestimaatin arvot kuuluvat joukkoon [X] eli estimaattori ei voi saada arvoa, joka ei kuulu aineistoon. Ehto (4) määrittää skaalan suhteiden säilyvyyden. Skaalaus siis takaa, että ei ole merkitystä mitataanko etäisyyttä metreissä taikka kilometreissä.[2]

Määritelmä 1.2. M-estimaattoriksi kutsutaan jokaista estimaattoria θ , joka määritellään minimoimalla sopivasti valittua kaavaa

$$min \sum \rho(x_i; \theta)$$

tai implisiittisellä yhtälöllä

$$\sum \phi(x_i; \theta) = 0,$$

missä $\phi(x_i; \theta) = \frac{\partial}{\partial \theta}\rho(x_i; \theta)$. Näitä estimaattoreita kutsutaan suurimman uskottavuuden estimaattoreiksi.[1] Erityisesti lokaatioestimaatit $\rho(x, \theta)$, jotka ovat muotoa

$$\sum (x_i - \theta) = \sum r_i, \tag{5}$$

eli yhtälöt, jotka minimoivat estimaatin ja havaintojen residuaaleja, ovat tämän tutkielman pääpainona. Kun kirjoitetaan kaava (5) muotoon

$$\sum w_1(x_1 - \theta),$$

missä

$$w_i = \frac{(x_i - \theta)}{\phi(x_i - \theta)},$$

saadaan painotettu kohdefunktio. Esimerkiksi pienimmän neliösumman kaavassa paino $w = 1$. Painofunktio tulee merkittäväksi erityisesti luvussa 2.4.1, kun tarkastellaan robusteja lokaatioestimaattoreita ja ne valitaan esimerkiksi satunnaisotoksesta.

Tässä luvussa määrittellyistä M-estimaattoreista saadaan skaalan suhteen ekvivalentteja lisäämällä sopiva τ, joka toimii skaala-estimaattina.

Esimerkki 1.3. Määritelmässä 1.2 mainittuun pienimmän neliön estimaattiin lisätään otosvarianssimuuttuja τ, jolloin funktio on muotoa $\rho(x, \theta, \tau)$ ja eksplisiittinen kaava saa muodon $\sum \rho(\frac{x-\theta}{\tau})$. Yleisesti käytetty otosvarianssi on MAD (Median Absolute Deviation) eli $median(|X_i - median(X_i))|$.

Esimerkki 1.4. Tavallisimpia otosestimaatteja on otoskeskiarvo. Olkoon $\rho(x_i; \theta)$ funktio, joka on muotoa $\rho(.) = (x_i - \theta)^2$, mikä antaa tavallisen pienimmän neliösumman M-estimaatin. Tällöin estimaatti, joka minimoi $\sum \rho(.)$ on otoskeskiarvo. Olkoon $\phi = \frac{\partial}{\partial \theta}\rho$. Nyt asettamalla summan nollaksi saadaan yhtälöt

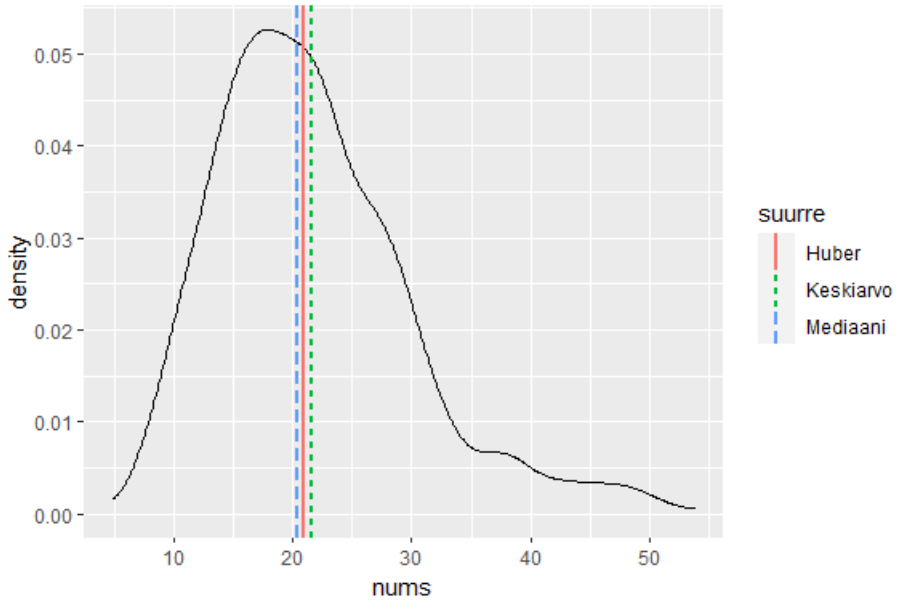
$$\begin{aligned} \sum \phi(x_i; \theta) &= 0 \\ -2\sum (x_i - \theta) &= 0 \\ \sum x_i - \sum \theta &= 0 \\ n\bar{x} - n\theta &= 0 \\ \implies \bar{x} &= \theta. \end{aligned}$$

Toinen yleinen tilastomatematiikassa käytetty M-estimaatti on logaritminen uskottavuusfunktio, joka on muotoa $\sum -logf(x; \theta)$ ja saadaan uskottavuusfunktiosta $L(\theta) = \prod f(x; \theta)$.

2 Robustit M-estimaatit

Luvussa 1 M-estimaattien ongelma on niiden herkkyys ääriarvoille. Tämän välttämiseksi M-estimaateille on kehitetty robusteja estimointitapoja.

Yksinkertaisuuden vuoksi esitellään lyhyesti robustisuus. M-estimaatit voidaan osoittaa kvantitaviisesti ja kvalitatiivisesti robusteiksi[1]. Kvalitatiivinen robustisuus vaatii, että estimaattorin funktio $\rho(x, \theta)$ ei kasva suhteettomasti pienten arvojen x muutoksien kanssa. Helppo tapa saavuttaa tämä on vaatia funktion olevan jatkuva. Kvantitatiivinen robustisuus yksinkertaisuuden vuoksi määritellään vain funktion hajoamispisteenä.[2] Käytännössä jos mielivaltaisen suuren havaintoarvon x lisääminen kasvattaa M-estimaatin mielivaltaisen suureksi, sanotaan hajoamispisteen olevan 0. Esimerkiksi mediaanin hajoamispiste on 0,5, eli puolet arvoista voidaan muuttaa mielivaltaisen suuriksi estimaatin arvon muuttumatta mielivaltaisen suureksi.



Kuva 1: Paksuhäntäisen todennäköisyysjakauman estimoituja arvoja eri metodein.

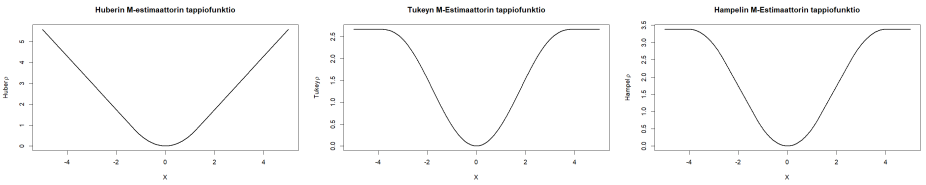
Kuva 1 näyttää, että Huberin robusti M-estimaattori antaa vähemmän painoarvoa ääriarvoille kuin tavallinen keskiarvo ja on täten tehokkaampi estimaattori. Huberin estimaatti ei myöskään ole yhtä kriittinen kuin mediaani, joka ei huomioi ääriarvoja laisinkaan.

2.1 Robusteja M-estimaattoreita

Taulukossa 1 on mainittuna ehdotettuja robusteja M-estimaattoreita, joita ovat Huberin M-estimaattori, Tukeyn Biweight ja Hampelin M-estimaattori joka on laajennus Huberin estimaattorista. Kuvassa 2 on esitettyinä taulukon 1 tappiofunktioiden kuvaajat.

Taulukko 1: Taulukko tavallisesti ehdotetuista M-estimaattoreista.

Funktio	ρ	ϕ	raja
Huber	$\frac{1}{2}x^2$	x	$\|x\| \leq K$
	$\|x\| K - \frac{1}{2}K^2$	$Ksign(x)$	$\|x\| > K$
Tukey		$x(1-x^2)^2$	$\|x\| < 1$
		0	$\|x\| \geq 1$
Hampel	$\frac{1}{2}x^2$	x	$\|x\| \leq a$
	$a\|x\| - \frac{1}{2}a^2$	$asign(x)$	$a < \|x\|$
	$\frac{a(c\|x\|-\frac{1}{2}x^2)}{c-b} - \frac{7}{6}a^2$	$\frac{asign(x)(c-\|x\|)}{c-b}$	$b < \|x\|$
	$a(b+c-a)$	0	$\|x\| > c$



(a) Huber ρ kuvaaja (b) Tukey ρ kuvaaja (c) Hampel ρ kuvaaja

Kuva 2: M-estimaattorien ρ kuvaajat.

Taulukon 1 kaavoista on nähtävissä, että arvon x_i etäisyyden kasvaessa estimaatista θ yli vakioiden K,a,b,c, selittävän muuttujan x_i vaikutus estimaattiin vähenee.

2.1.1 Huber M-estimaatti

Huberin M-estimaatti noudattaa pienimmän neliösumman kaavaa $min \sum(x-\theta)^2$ johonkin valittuun vakion arvoon K, $|x| \leq K$ saakka,

ja kohdefunktio $\rho(.)$ kasvaa polynomisen yhtälön mukaisesti. Arvon $x - \theta$ kasvaessa yli jonkin vakion K, kohdefunktio $\rho(.)$ kasvaa lineaarisesti. K valitaan sopivasti riippuen halutuista ominaisuuksista, mutta $K = 1,28$ on yksi vaihtoehto, joka on normaalijakauman 0,9 kvantiili. Hampel noudattaa Huberin kaavaa, kunnes poikkeukselliset arvot alkavat menettää vaikutustaan jonkin vakion b jälkeen ja jäävät huomiotta jonkin vakion c jälkeen. Hampelin estimaatin etu Huberin estimaattiin on, että kun $x \to \infty$, Hampelin estimaatti ei kasva äärettömästi. Toisin sanoen Hampelin estimaatilla on korkeampi hajoamispiste. Hampelin ja Huberin estimaattien kaavoista voidaan nähdä, että niiden käyttö ei ole täysin suoraviivaista. Tarvitaan tietoa estimaatista ennen kuin voidaan määritellä rajat, jotka määrittävät milloin havaintoarvo on poikkeuksellinen. Tämän vuoksi M-estimaattorit vaativat iteratiivisen menetelmän todellisen estimaatin ratkaisemiseksi.

Esimerkki 2.1. Olkoon $k = 0$ iteraatioaskel ja $\hat{\theta_k} = M$, missä M on havaintoaineiston mediaani sekä $K = 1.28$, joka on yleisesti valittu arvo. Newton-Rhapson menetelmää käyttäen M-estimaatti ratkaistaan iteratiivisesti käyttämällä algoritmia:

- Askel 1. Olkoon
$$C = \sum \rho(\frac{X_i - \hat{\theta}_k}{MADN}),$$
missä $\rho(.)$ on Huberin M-estimaattorin yhtälö.

- Askel 2.Olkoon
$$H = \sum \phi(\frac{X_i - \hat{\theta}_k}{MADN}),$$
missä $\phi(.)$ on Huberin estimaatin influenssifunktio eli $\frac{\partial \rho}{\partial \theta}$.

- Askel 3. Asetetaan
$$\hat{\theta}_{k+1} = \hat{\theta}_k + \frac{MADN * C}{H}$$

- Askel 4. Kasvatetaan askelta k yhdellä ja toistetaan askeleet 1-3 kunnes arvot konvergoivat. Toisin sanoen $\left|\hat{\theta}_{k+1} - \hat{\theta}_k\right| < \epsilon$, jollakin mielivaltaisen pienellä arvolla ϵ.

2.1.2 Tukeyn Biweight-estimaatti

Tukeyn estimaatti kasvaa myös nopeasti arvon x_i ollessa estimaatin lähellä, mutta ylisuuret tai poikkeukselliset havainnot menettävät vaikutustaan estimaattiin havaintojen kasvaessa erityisen suuriksi. Kuten Hampelin estimaattorin tapauksessa, Tukeyn Biweight jättää huomiotta poikkeuksellisen suuret arvot.

2.2 Influenssifunktio

Influenssifunktio mittaa lisätyn arvon x vaikutusta estimaattiin. Olkoon estimaatti T(F) jokin kuvaus jakaumasta F, ja F todennnäköisyysjakauma joka riippuu muuttujasta X. Influenssifunktio on toisin sanoen estimaatin T derivaatta ja mittaa muutosta estimaattiin jakauman F muuttuessa.

Määritelmä 2.2. Olkoon x jostain jakaumasta δ_x, jossa todennäköisyys saada arvo x on 1. Olkoon F jokin todennäköisyysjakauma ja $F_{x,\epsilon}$ jakauma, jossa todennäköisyydellä ϵ havainto tulee jakamausta δ_x. Yhtälö saa muodon

$$F_{x,\epsilon} = (1-\epsilon)F + \epsilon\delta_x.$$

Suhteellinen vaikutus funktioon T(F), kun arvo x ilmenee, on

$$\frac{T(F_{x,\epsilon}) - T(F)}{\epsilon}$$

ja influenssi funktio

$$IF(x) = \lim_{\epsilon^+ \to \inf} \frac{T(F_{x,\epsilon}) - T(F)}{\epsilon}. \tag{6}$$

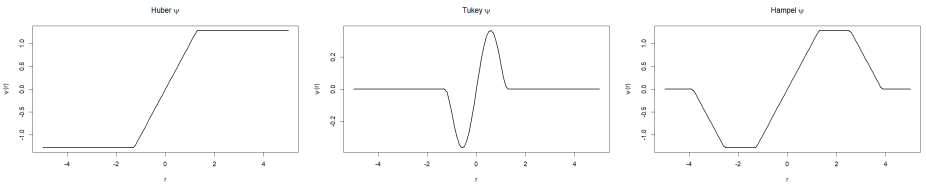
Yleisesti IF(X) on suhteellinen vaikutus jollekin suureelle, joka määrittelee funktiota T, kun todennäköisyys havainnolle x on mielivaltaisen lähellä arvoa 0.

Esimerkki 2.3. Olkoon $T(F) = \mathbf{E}[X] = \mu$ eli populaation odotusarvo. Nyt $T(F_{x,\epsilon}) - T(F) = (1-\epsilon)\mu + \epsilon x - \mu = \epsilon(x-\mu)$, eli

$$\begin{aligned} IF(x) &= \frac{\epsilon(x-\mu)}{\epsilon} \\ &= x - \mu, \end{aligned}$$

missä x ei ole rajattu.

Robusteja M-estimaatteja määrittäessä täytyy lisätä ehto, että käytettävä influenssifunktio on rajattu. Ylisuuret havainnot eivät siis vaikuta rajattomasti estimaatin arvoon. Lisäksi ϕ funktio rajataan kuvaajiin, jotka ovat symmetrisiä $-\phi(x) = \phi(-x)$.



(a) Huber ϕ kuvaaja (b) Tukey ϕ kuvaaja (c) Hampel ϕ kuvaaja

Kuva 3: M-estimaattorien ϕ, eli influenssifunktioiden kuvaajat.

Kuva 3 esittää robustien M-estimaattoreiden rajattuja influenssifunktioita. Huberin estimaatilla arvon $|x_i|$ kasvaessa sen vaikutus muuttuu lineaariseksi, kun Tukeyn ja Hampelin estimaateilla influenssifunktio painuu nollaan eli ääriarvot eivät vaikuta estimaattiin. Kuvan 3 kolme estimaattia ovat myös influenssifunktion suhteen lineaarisia nollan ympäristössä, eli estimaatin läheisyydessä havainnoilla on suurempi paino itse estimaattiin.

2.3 Skaala

Yksi ehdoista robusteille estimaattoreille on skaalaekvivalenssi. Taulukossa 1 mainituilla kaavoilla ei ole tätä ominaisuutta. Tätä varten tarvitaan skaalaussuurre.

Määritelmä 2.4. Olkoon laskettavan Huberin M-estimaatin funktio $\phi(x, \theta, \tau)$, missä τ on skaalaparametri. Nyt Huberin M-estimaatin kaava on muotoa

$$\phi(\frac{x-\theta}{\tau}) = \begin{cases} -K, & \text{jos} \quad (x-\theta)/\tau < -K \\ \frac{x-\theta}{\tau}, & \text{jos} \quad -K \leq (x-\theta)/\tau \leq K \\ K, & \text{jos} \quad (x-\theta)/\tau > K. \end{cases} \tag{7}$$

Kertomalla yhtälön 7 rajaehdot parametrilla τ saadaan rajat

$$(x-\theta) < -K\tau \tag{8}$$

$$-K\tau \leq (x-\theta) \leq K\tau \tag{9}$$

$$(x-\theta) > K\tau. \tag{10}$$

Yhtälöistä (8), (9), (10) on nähtävissä, että skaalaparametri τ määrittää, milloin arvo x on erityisen kaukana estimaatista. Arvo x ei ole erityisen suuri tai pieni, jos $|x-\theta| \leq K\tau$. Havainto x arvo on poikkeuksellisen suuri, jos $|x-\theta| > K\tau$.[2]
Tyypillinen skaalauskerroin on niin kutsuttu MADN, joka on keskihajontaestimaatti σ kun otetaan havaintoja normaalijakaumasta. Yhtälö

$$MADN = \frac{MAD}{z_{0.75}} \approx \frac{MAD}{0.6745}$$

missä MAD (Median Absolute Deviation) on mediaani hajonnan itseisarvoista ja $z_{0,75}$ on 0,75 kvantiili normaalijakaumsta. $\mathbf{MAD} = \mathbf{MED}(|\mathbf{X_i}| - \mathbf{M}$, $i = 1, 2, 3...n$ ja $\mathbf{M}$ on havaintojen mediaani.

2.4 Regressio

Määritelmä 2.5. Olkoon $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, missä selittävä muuttuja $\mathbf{X}$ on $n \times p$ matriisi, regressiokerroinmatriisi $\boldsymbol{\beta}$ $p \times 1$, vektori $\boldsymbol{\epsilon}$ virhetermi ja vektori $\mathbf{Y}$ on $n \times 1$ vastemuuttuja. Lineaarinen regressio on $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$ missä $y_i = \sum_{j=0}^{p} x_{ij}\beta_j + \epsilon_i$.

Mallinnettaessa lineaarista regressiota mielenkiinnon kohteena on kertoimen $\boldsymbol{\beta}$ arviointi. Olkoon $S(\boldsymbol{\beta}) = \sum(y_i - \mathbf{x_i}\boldsymbol{\beta})^2$. Pienimmän neliösumman menetelmällä ratkaisu saadaan derivoimalla regressiokertoimien suhteen ja asettamalla kaikki yhtälöt nollaksi. $\frac{\partial S(\beta)}{\partial \beta_j} = 0$, $j = 0, 1, 2...p$. Nyt kun minimoidaan residuaaleja saadaan ratkaisukaavaksi

$$\begin{aligned} minS(\boldsymbol{\beta}) &= min \sum (y_i - \mathbf{x_i}\boldsymbol{\beta})^2 \\ \frac{\partial S}{\partial \beta_j} &= \sum (y_i - \mathbf{x_i}\boldsymbol{\beta}) x_{ij} = 0, \quad j = 0, 1, 2...p. \end{aligned} \tag{11}$$

Voidaan osoittaa, että $\hat{\boldsymbol{\beta}}$ saadaan normaaliyhtälöillä

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Pienimmän neliösumman menetelmä on kuitenkin altis ääriarvoille, mikä on helposti nähtävissä yhtälöstä (11), jos annamme $x \to \inf$. Tarvitaan robustimpia menetelmiä rajoittamaan ääriarvojen vaikutusta.

2.4.1 Painotettu pienimmän neliösumman regressio

Pienimmän neliösumman metodista saadaan robustimpi painottamalla jokaista arvoa x_i painolla w_i, jolloin arvot jotka ovat kauempana estimaatin arvosta $\boldsymbol{\beta}$ saavat pienemmän painonarvon.

Esimerkki 2.6. Olkoon minimoitava funktio lineaarinen regressio

$$S(\boldsymbol{\beta}) = \sum w_i \rho(\mathbf{X}, \boldsymbol{\beta}, \tau) = \sum w_i (\frac{y_i - \mathbf{x_i}\boldsymbol{\beta})}{\tau}). \tag{12}$$

Paino $w_i = \sqrt{1 - h_{ii}}$, missä h_{ii} on hattumatriisin $\mathbf{H}$ diagonaalialkio tai toisin kutsuttuna vipupiste. Hattumatriisi saadaan kaavasta $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$. Nyt derivoituna kaava (12) saa muodon

$$\sum \sqrt{1 - h_{ii}}(y_i - \mathbf{x_i}\boldsymbol{\beta})x_{ij} = 0, \quad i = 0, 1, 2...n,$$

joka saa matriisimuodon

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}'\mathbf{Y}$$

missä $\mathbf{W}$ on diagonaalimatriisi, jonka arvona $w_{ii} = [\mathbf{H}]_{\mathbf{ii}}$.

2.4.2 Iteratiivinen painotettu pienin neliösumma

Olkoon jälleen $\mathbf{y_i} = \mathbf{x_i}\boldsymbol{\beta}$, missä $i = 0, 1, 2...n$ ja $\boldsymbol{\beta}$ on $p \times 1$ vektori. Laskettaessa M-regressioestimaattori Schweppen painoilla, missä paino w_i

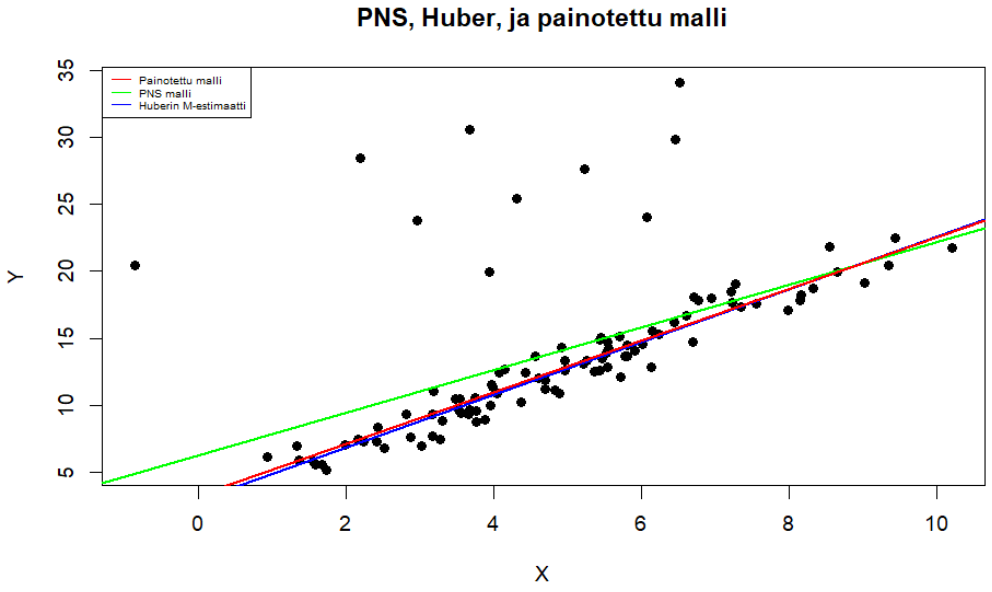
$$w_{i,k} = \frac{\sqrt{1 - h_{ii}}\hat{\tau}}{r_i^{(t)}} \phi(\frac{r_i^{(t)}}{\sqrt{1 - h_{ii}}\hat{\tau}})$$

ja

$$\phi(x) = max[-K, min(K, x)].$$

$K = 2\sqrt{(p+1)/n}$ eli Huberin M-estimaatti. Aloitetaan iteraatioaskel $t = 0$, lasketaan pienimmän neliösumman estimaatti kertoimille $\boldsymbol{\beta}$ esim normaaliyhtälöillä ja hattumatriisi olkoon $\mathbf{H}$.

- Lasketaan residuaalit $r_i^{(t)} = y_i - \mathbf{x_i}\hat{\boldsymbol{\beta}}$. Olkoon $\mathbf{M^{(t)}}$ mediaani suurimmista n-p residuaaleista $\left|r_{i,k}^{(t)}\right|$, ja $\hat{\tau}^{(t)} = 1.48\mathbf{M^{(t)}}$.
- Muodostetaan painot w_i.
- Käytetään painoja w_i laskeaksemme painotettu pienimmän neliösumman estimaatti $\hat{\boldsymbol{\beta}}$. Kasvatetaan iteraatioaskelta yhdellä.
- Toistetaan askeleet, kunnes estimaatit konvergoivat eli $\left|\hat{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}}^{(t+1)}\right| \leq \epsilon$, missä ϵ jokin valittu pieni raja-arvo.



Kuva 4: Regressiot PNS-mallilla, Huber M-regressiolla ja painotetulla mallilla

Esimerkki 2.7. R-esimerkki iteratiivisesta painotetusta regressiosta Schweppen painoilla ja tavallisesta pienimmän neliösumman regressiosta. Olkoon $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ja $y = 5 + 2x_1 - 3x_2 - x_3 + \epsilon$, missä $\epsilon \sim \mathbf{N}(0, 1)$. Vektoriin $\mathbf{Y}$ on lisätty satunnaisia ääriarvoja ja satunnaismuuttuja $\mathbf{X}$ on 100×3 matriisi, jonka arvot arvottu normaalijakaumasta, $\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$.

Taulukko 2: Esimerkki 2.7 tulokset

	Vastemuuttuja:	
	y	
	Painotettu Schweppe	PNS
X1	2.033***	1.799***
	(0.075)	(0.322)
X2	-3.090***	-3.098***
	(0.064)	(0.283)
X3	-1.063***	-1.180***
	(0.078)	(0.319)
Constant	5.154***	6.596***
	(0.069)	(0.292)
Observations	300	300
R^2	0.919	0.359
Adjusted R^2	0.918	0.353
Residual Std. Error (df = 296)	0.791	5.043
F Statistic (df = 3; 296)	1,123.387***	55.282***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Robustien M-estimaattoreiden käyttö on hyödyllistä lineaarisissa regressioissa. Taulukosta 2 nähdään, että ei-robusteilla menetelmillä ääriarvot siirtävät pienimmän neliösumman mallia. Painotetun Schweppen painoilla tuotetun mallin selitysvoima on lähes kolminkertainen verrattuna pienimmän neliösumman malliin, Painotettu malli ($R^2 = 0.919$), PNS-malli($R^2 = 0.359$). Tavalliset ei-robustit menetelmät johtavat tilanteeseen, joissa estimaatteja varten joudutaan tarkastelemaan aineistoa ja Cookin mittaa määrittämään poistettavat ääriarvot mallin sopivuuden parantamiseksi. Robusteilla M-estimaattoreilla ei tarvitse poistaa ääriarvoja, jolloin vältetään menettämästä mahdollisesti merkittävää tietoa.

3 Luottamusväli ja hypoteesitestaus

Tärkeä osa tilastollista päättelyä estimaatteja laskettaessa on arvioida estimaattien uskottavuutta ja sitä, kuinka hyvin ne tukevat asetettuja hypoteeseja. Koska tämän tutkielman pääpaino on M-estimaattoreissa, on tarkastelu jätetty erään esimerkki metodin esittämiseen. Käytetty estimointimenetelmä voi antaa yksiselitteisen arvon aineistosta, mutta aineistoon sisältyy vääjäämättömästi epävarmuutta, joka välittyy estimaattiin. Jopa sellaiset aineistot jotka kuvaavat kokonaista populaatiota, voivat sisältää epävarmuutta. Tämä epävarmuus voi syntyä esimerkiksi mittausvirheistä tai tavallisista inhimillisistä virheistä mittaustuloksia kirjatessa.

3.1 Luottamusväli yhden otoksen tapauksessa

Tehokkain tapa arvoida luottamusvälejä M-estimoinnissa on prosentuaalinen saapasremmimenetelmä [2].

Määritelmä 3.1. Olkoon satunnaisotanta $\mathbf{X}(n \times 1)$ vektori. Saapasremmimetodissa valitaan satunnaisotanta palauttamalla aineistosta $\mathbf{X}$, jolloin saadaan uusi satunnaisotanta $\mathbf{X_i^*}$. Olkoon $\hat{\theta}_i^*$ M-estimaatti, joka on laskettu otannasta $\mathbf{X_i^*}$. Toistetaan prosessi $\mathbf{M}$ kertaa, jolloin saadaan estimaattivektori $\hat{\theta}^*$. Nyt olkoot $\gamma = \alpha\mathbf{M}/2$ pyöristettynä lähimpään kokonaislukuun ja

$\delta = \mathbf{M} - \gamma$. Järjestämällä arvot $\hat{\theta}_i^*$ nousevaan arvojärjestykseen, $\hat{\theta}_1^* \leq ... \leq \hat{\theta}_M^*$ saamme $1 - \alpha$ luottamusvälin estimaatille θ, joka on $(\hat{\theta}_{\gamma+1}^*, \hat{\theta}_\delta^*)$.

Määritelmä 3.2. Olkoon nollahypoteesi

$$H_0 : \theta = \theta_0$$

missä θ_0 on jokin vakio. Olkoon nyt $p^* = \mathbf{P} = (\hat{\theta}^* < \theta_0)$. Arvo p^* voidaan estimoida yhtälöllä

$$\hat{p^*} = \frac{A}{M},$$

missä $A = \sum_{i=1}^{j} 1$ ja j on vektorin $\hat{\theta}^*$ indeksin arvo, jossa $\hat{\theta}_j^* < \theta_0$. Nollahypoteesi hylätään kun $\hat{p}^* \leq \alpha/2$ tai $\hat{p}^* \geq 1 - \alpha/2$.

Ehdotettu minimiraja saapasremmiestimaattien määrälle M=500 [2].

3.2 Regressiopäättely

Saapasremmimenetelmä toimii myös regressiota mallintaessa. Kuten luvussa 3.1, saapasremmimenetelmällä kootaan uusi satunnaisotanta alkuperäisestä havaintoaineistosta palauttamalla ja kootaan saapasremmimenetelmän tuottamat regressiokertoimet omaan vektoriin. Toisin kuin tavanomaisessa regressiopäättelyssä, saapasremmimenetelmä sallii virhetermin heteroskedastisuutta.

3.2.1 Menetelmä hypoteesitestaukseen mallintaessa regressiota

Olkoon regressiomalli $y_i = x_{i0} + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$, missä $i = 1, 2, 3, ..., n$. Nyt lineaarinen malli on muotoa $\hat{y}_i = x_{i0} + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + ... + x_{ip}\hat{\beta}_p + \epsilon$, missä ϵ noudattaa jotain satunnaisjakaumaa. Olkoon tarkasteltava nollahypoteesi

$$H_0 = \beta_1 = \beta_2 = ... = \beta_p = 0.$$

Olkoon havaintoaineisto

$$\begin{pmatrix} y_1 & x_{11}, & \dots, & x_{1p} \\ & & \vdots & \\ y_n & x_{n1}, & \dots, & x_{np} \end{pmatrix}.$$

Lasketaan regressiokertoimet $\hat{\beta}_j$ ja luodaan saapasremmimenetelmällä uusi matriisi $\mathbf{X}^*$, missä y_i korvataan arvolla y_i^* ja x_{ij} korvataan arvoilla x_{ij}^*. Näistä lasketaan uudet regressiokertoimet $\hat{\beta_{ij}^*}$, jotka lisätään $(p \times M)$ matriisiin $\hat{\beta}^*$. $\hat{\beta}^*$ matriisin alkioiden $\hat{\beta_{ij}^*}$ indeksit i liittyvät tarkasteltaviin regressiokertoimiin ja indeksi j liittyy saapasremmi-iteraatioon. Nyt estimoitu kovarianssi kertoimien $\hat{\beta}_c$ ja $\hat{\beta}_h$, $c \neq h$ on

$$v_{ch} = \frac{1}{M-1} \sum_{j=1}^{M} (\hat{\beta_{cj}^*} - \overline{\beta}_c^*)(\hat{\beta}_{hj}^* - \overline{\beta}_h^*),$$

missä $\bar{\beta}_c^* = \sum \hat{\beta}_{cj}^*/M$. Kootaan arvot v_{ch}, $(p \times p)$ matriisiin $\mathbf{V}$. Etäisyys alkuperäisten regressiokertoimien ja saapasremmimenetelmän tuottamien regressiokertoimien välillä on

$$d_m^2 = (\hat{\beta}_{1m}^* - \hat{\beta}_1, ..., \hat{\beta}_{pm}^* - \hat{\beta}_p)\mathbf{V^{-1}}(\hat{\beta}_{1m}^* - \hat{\beta}_1, ..., \hat{\beta}_{pm}^* - \hat{\beta}_p)'.$$

Hypoteesin testaus perustuu siis estimoitujen regressiokertoimien etäisyyteen nollahypoteesista. Olkoon

$$\mathbf{D} = \sqrt{(\hat{\beta}_1, ..., \hat{\beta}_p)\mathbf{V^{-1}}(\hat{\beta}_1, ..., \hat{\beta}_p)'}$$

eli alkuperäisesta aineistosta laskettu arvo. Järjestämällä etäisyydet $d_i, i = 1, 2, 3, ..., M$, asettamalla $k = (1 - \alpha)M$ ja olkoot $K = k$ pyöristettynä lähimpään kokonaislukuun. Nollahypoteesi voidaan hylätä jos

$$D > d_K.$$

Edellä mainitut hypoteesitestaukset, joissa on useita estimaatteja voidaan yleistää koskemaan myös tapausta

$$H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0$$

, missä $q < p$.

4 Yhteenveto

Tässä tutkielmassa on esitelty eräitä M-estimaattoreita ja niiden käyttöä datan mallintamiseen. Ei-robustit tavat, kuten pienimmän neliösumman metodi, toimivat hyvin suurissa aineistoissa. Pienissä aineistoissa, tai aineistoissa joissa ilmenee myös ääriarvoja, M-estimaattorit on helposti muokattavissa

robustimmaksi. Estimaattien käyttö hypoteesien testaukseen ja luottamusvälien laskemiseen ei poikkea muista metodeista. Tässä tutkielmassa on esitetty eräs tapa robustiin hypoteesin testaukseen ja luottamusvälin laskentaan, jotka ovat laajennettavissa useamman muuttujan tapauksiin.

Tiivistettynä M-estimaattori on Huberin määrittelemä summafunktion minimoiva estimaattori tai summafunktion derivaatan nollakohta. Robustimmat estimaattorit käsittelevät ääriarvoja ilman poistoa ja rajoittavat niiden vaikutusta estimaattiin kuten Tukeyn Biweight ja Hampelin M-estimaattori. Robusteja estimaatteja laskettaessa on käytettävä iteratiivisia prosesseja, joissa lasketaan ensin esimerkiksi ei-robusti alkuarvo estimaatille. Iteratiivista prosessia käyttämällä robustilla estimaattorilla voidaan määritellä ääriarvot uudelleen ja saadaan tarkempia tuloksia.

5 Liitteet

5.1 R koodit

```
# Maaritellaan funktiot
#Huber
huberM <- function(x, K) {
  ifelse(abs(x) <= K, 0.5 * x^2, K * (abs(x) - 0.5 * K))
}
# Tukeyn Biweight

tukeyM <- function(x, c) {
  ifelse(abs(x) <= c, (c^2 / 6) * (1 - (1 - (x / c)^2)^3), c^2 / 6)
}
# Hampelin

hampelM <- function(x, hampel_a, hampel_b, hampel_c) {
  ifelse(abs(x) <= hampel_a, 0.5 * x^2,
         ifelse(abs(x) <= hampel_b, hampel_a * (abs(x) - 0.5 * hampel_
                ifelse(abs(x) <= hampel_c, (hampel_a * (abs(x)- hampel
                       hampel_a * (hampel_b + hampel_c - hampel_a)/2))
}
# luodaan aineisto ja m    ritell    n rajat
x_arvot <- seq(-5, 5, length.out = 500)
K <- 1.28
c <- 4
hampel_a <- 1.28
hampel_b <- 2.56
hampel_c <- 4

# Lasketaan funktion arvot
y_arvot <- sapply(x_arvot, huberM, K = K)
tukey_y_arvot <- sapply(x_arvot, tukeyM, c = c)
```

```
hampel_y_arvot <- sapply(x_arvot, hampelM, hampel_a, hampel_b, hampel_

# Luodaan kuvaajat
plot(x_arvot, tukey_y_arvot, type = "l", lwd = 2,
     xlab = "X",
     ylab = expression(paste("Tukey ", rho)),
     main = "Tukeyn M-Estimaattorin tappiofunktio")

plot(x_arvot, hampel_y_arvot, type = "l", lwd = 2,
     xlab = "X",
     ylab = expression(paste("Hampel ", rho)),
     main = "Hampelin M-Estimaattorin tappiofunktio")

plot(x_arvot, y_arvot, type = "l", lwd = 2,
     xlab = "X",
     ylab = expression(paste("Huber ", rho)),
     main = "Huberin M-estimaattorin tappiofunktio")




# Derivaatta funktiot

huber_fii <- function(r, K) {
  ifelse (abs(r) <= K, r,
          K *sign(r))
}
tukey_fii <- function(r, K) {
  ifelse (abs(r) <= K, r * ( 1 - (r / K)^2)^2,
          0)
}
hampel_fii <- function(r, a, b, c) {
  ifelse (abs(r) <= a, r,
```

```
        ifelse(abs(r)<= b, sign(r) * a,
               ifelse(abs(r) <= c, (sign(r) *a * (c - abs(r)))/(c -
        ))
}
#Luodaan rajat ja aineisto
K <- 1.28
a <- 1.28
b <- 2.56
c <- 3.84

r_arvot <- seq(-5, 5, by = 0.1)

# Lasketaan funktion arvot

hub_fii_arvot <- sapply(r_arvot, huber_fii, K)
tuk_fii_arvot <- sapply(r_arvot, tukey_fii, K)
hampel_fii_arvot <- sapply(r_arvot, hampel_fii, a , b , c)

#Luodaan kuvaajat
plot(r_arvot, hub_fii_arvot, type = "l",
     main = expression(paste("Huber ",psi)),
     xlab = "r",
     ylab = expression(paste(psi,"(r)")),
     lwd = 2)
plot(r_arvot, tuk_fii_arvot, type = "l", lwd = 2,
     main = expression(paste("Tukey ",psi)),
     xlab = "r",
     ylab = expression(paste(psi,"(r)")))

plot(r_arvot, hampel_fii_arvot, type = "l",
     main = expression(paste("Hampel ",psi)),
     xlab = "r",
     ylab = expression(paste(psi,"(r)")),
```

```
      lwd = 2)
```

```
# Luodaan aineisto
set.seed(30)
x1 <- rnorm(100, mean = 5, sd = 2)
#Luodaan uudelleen painotettu mallin funktiot
huber_paino <- function(residuals, k = 1.28) {
  s <- mad(residuals)
  r <- residuals / s
  w <- ifelse(abs(r) <= k, 1, k / abs(r))
  return(w)
}

uudelleenpain <- function(X, y, max_iter = 50, tolerance = 1e-6) {
  PNS_malli <- lm(y ~ X)
  for (i in 1:max_iter) {
    residuals <- PNS_malli$residuals
    w <- huber_paino(residuals)
    paiv_malli <- lm(y ~ X, weights = w)
    if (sum(abs(paiv_malli$coefficients - PNS_malli$coefficients)) < t
      break
    }
    PNS_malli <- paiv_malli
  }
  return(PNS_malli)
}
```

```
#Luodaan vastemuuttuja ja lis t    n variaatiota
y <- 3 + 2 * x1  + rnorm(100, mean = 0, sd = 1 )
aariarvo <- sample(1:40, 10)
y[aariarvo] <- y[aariarvo] + c(10,15, 20, 15, 20)
# Create a data frame
data <- data.frame( x1 = x1, y = y )

# Fit the multiple linear regression model
PNS_malli <- lm(y ~ x1, data = data)
robusti_malli <- rlm(y ~ x1, data = data, psi = psi.hampel)
painotettu_malli <- uudelleenpain(x1 , y)
# Summary of the model
summary(PNS_malli)
summary(robusti_malli)
summary(painotettu_malli)
plot(data$x, data$y, type = "p", pch = 16,
     main = "PNS, Huber, ja painotettu malli",
     xlab = "X",
     ylab = "Y")

abline(PNS_malli,lwd =2, col= "green")
abline(robusti_malli, lwd = 2, col = "blue")
abline(painotettu_malli, lwd = 2, col = "red")
legend("topleft",cex = 0.5, legend = c("Painotettu malli","PNS malli",



# Schweppen painoilla iteroitu funktio
Psi <- function(x, K) {
  pmax(-K, pmin(K, x))
}
irew_schwep <- function(X, y, p, max_iter = 500, tol = 1e-6) {
```

```
  #Alustetaan funktio PNS mallilla ja alkuarvoilla
  piennel_malli <- lm(y ~ X)
  residuals <- piennel_malli$residuals
  M_k <- median(abs(residuals))
  tau_k <- 1.48 * M_k
  e_k <- residuals / tau_k
  h_ii <- hatvalues(piennel_malli)
  K <- 2 * sqrt((p + 1) / n)

  # Iteraation alustus
  iter <- 0
  konverg <- FALSE

  while (!konverg && iter < max_iter) {
    iter <- iter + 1
    w_k <- sqrt(1 - h_ii) / e_k * Psi(e_k / sqrt(1 - h_ii), K)
    painotettu_malli <- lm(y ~ X, weights = w_k)
    #konvergoitumisen tarkistus
    if (sum(abs(painotettu_malli$coefficients - piennel_malli$coefficie
      converged <- TRUE
    }

    # P ivitet n arvot
    piennel_malli <- painotettu_malli
    residuals <- painotettu_malli$residuals
    M_k <- median(abs(residuals))
    tau_k <- 1.48 * M_k
    e_k <- residuals / tau_k
  }

  # Palautetaan malli
  return(painotettu_malli)
}
```

```
#Luodaan monimuuttuja aineisto
n <- 300
p <- 3
X <- matrix(rnorm(n*p), nrow = n)
y <- 5 + 2 * X[, 1] - 3 * X[, 2]- X[, 3] + rnorm(n)
aariarvo <- sample(1:40, 30)
y[aariarvo] <- y[aariarvo] + sample(c(10, 15, 20), 10, replace = TRUE)


ilws_schweppe <- irew_schwep(X, y , p)
multiPNS <- lm(y ~ X)
install.packages("stargazer")
library(stargazer)
stargazer(ilws_schweppe , multiPNS, title = "Tulokset", align = TRUE)
```

Kirjallisuutta

[1] Peter J. Huber, Elvezio M. Ronchetti: *Robust Statistics*

[2] R. R. Wilcorx: *Introduction to Robust Estimation and Hypothesis Testing.*