Tampereen yliopisto

Markus Linjamäki

Bootstrap-menetelmän käyttö lineaarisessa regressiossa

Informaatioteknologian ja viestinnän tiedekunta
Tilastollisen data-analyysin kandidaattitutkielma
Toukokuu 2021

Tiivistelmä

Markus Linjamäki: Bootstrap-menetelmän käyttö lineaarisessa regressiossa
Tilastollisen data-analyysin kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastotieteen tutkinto-ohjelma
Toukokuu 2021

Tietokoneiden kehittyessä ja niiden laskentatehon kasvaessa erilaiset simulaatiomenetelmät ovat yleistyneet tilastotieteessä osana tilastollista päättelyä. Yksi näistä simulaatiomenetelmistä on nimeltään bootstrap, joka muodostaa palauttaen tehtyjä otoksia satunnaisotoksen estimoidusta kertymäfunktiosta. Bootstrap-menetelmän avulla voidaan muodostaa parametrin estimaatin bootstrap-jakauma, jonka avulla voidaan laskea esimerkiksi estimaatin harha tai varianssi. Tutkimuksessa esitellään aluksi bootstrap-menetelmän perusteita ja johdetaan muutama tärkeä tulos bootstrap-jakaumasta.

Regressioanalyysi tarkastelee vastemuuttujan ja selittävien muuttujien välistä suhdetta. Lineaarinen regressioanalyysi olettaa tämän suhteen olevan lineaarinen. Tässä tutkimuksessa esitellään yksinkertainen lineaarinen regressiomalli, joka soveltuu vastemuuttujan ja yhden selittävän muuttujan väliseen tarkasteluun. Mallin parametrien estimoimiseen esitellään pienimmän neliösumman menetelmä.

Tutkimuksessa esitellään aluksi kaksi eri tapaa bootstrap-menetelmän hyödyntämiseen lineaarisessa regressiossa. Menetelmät ovat bootstrap pareille ja bootstrap residuaaleille. Menetelmien välisessä vertailussa hyödynnetään aluksi havaintoaineistoa, joka todetaan soveltuvan erittäin hyvin lineaariseen regressioanalyysiin. Sen jälkeen tarkastelua jatketaan simuloidun havaintoaineiston avulla, jossa yksi regressiomalliin liittyvä oletus ei ole kunnossa. Oletus koskee regressiomallin virhetermien varianssia, ja tulosten tarkastelussa painottuukin estimaattien varianssien vertailu. Tutkimuksessa havaitaan, että bootstrap residuaaleille -menetelmä toimii toista menetelmää paremmin tilanteessa, jossa kaikki regressiomallille asetetut oletukset ovat kunnossa. Bootstrap pareille -menetelmän havaitaan taas toimivan toista menetelmää paremmin tilanteessa, jossa kaikki regressiomalliin liittyvät oletukset eivät ole kunnossa. Menetelmien tuottamat erot eri havaintoaineistoissa todetaan johtuvan niiden tavasta muodostaa vastemuuttujat ennen varsinaista bootstrap-jakauman muodostamista. Lopuksi tutkimuksessa esitellään wild-bootstrap -menetelmä ja havaitaan sen tuottavan parhaimman tuloksen aiemmin esitellyssä simuloidussa havaintoaineistossa.

Sisällys

1 Johdanto **4**

2 Bootstrap-menetelmä **5**
2.1 Esittely 5
2.2 Epäparametrinen bootstrap 5
2.3 Parametrinen bootstrap 6
2.4 Tuloksia 6
2.5 Katsaus R-ohjelmiston tarjoamiin pakkauksiin 6
2.5.1 Ohjelmakirjasto bootstrap 7
2.5.2 Ohjelmakirjasto boot 7
2.5.3 Muita ohjelmakirjastoja 7

3 Lineaarinen regressiomalli **8**
3.1 Yksinkertainen lineaarinen regressiomalli 8
3.2 Pienimmän neliösumman menetelmä 9
3.3 Regressiomalli matriisi notaatiolla 12
3.4 Bootstrap-menetelmä lineaarisessa regressiossa 12
3.4.1 Bootstrap residuaaleille 13
3.4.2 Bootstrap pareille 14
3.4.3 Menetelmien vertailua 15
3.5 Mallin sovitus heteroskedastisessa aineistoissa 15
3.5.1 Wild-bootstrap 17

4 Yhteenveto **18**

Lähteet **19**

Liite A: Kukkien verholehtien korkeudet ja leveydet **20**

Liite B: R-koodi **21**

Liite C: Simuloitu data **23**

1 Johdanto

Tämän tutkimuksen tarkoituksena on esitellä bootstrap-menetelmän hyödyntämistä lineaarisessa regressiossa. Bootstrap on suhteellisen uusi tilastollinen menetelmä, joka hyödyntää uudelleenotantaa ja koneellista laskentaa parametrin estimaatin jakauman muodostamisessa. Se kehitettiin alun perin toisen samantyylisen simulaatiomenetelmän, jackknife-menetelmän, yksinkertaistetuksi sovellukseksi, mutta monissa tilanteissa bootstrap-menetelmä tuottaa samanlaisia tai jopa tarkempia tuloksia jackknife-menetelmään verrattuna (Chernick & LaBudde 2011, s. 1). Työn alussa esitellään bootstrap-menetelmä, ja muutamia bootstrap-jakaumasta johdettuja tuloksia.

Regressioanalyysin avulla tutkitaan selittävien muuttujien vaikutusta selitettävään muuttujaan. Lineaarisessa regressiomallissa selitettävän muuttujan oletetaan muodostuvan selittävien muuttujien lineaarikombinaatioista. Tässä työssä esitellään yksinkertainen lineaarinen regressiomalli, jonka parametrit ovat vakiotermi ja kulmakerroin. Malli soveltuu tilanteisiin, jossa tarkastellaan selitettävän muuttujan riippuvuutta yhdestä selittävästä muuttujasta. Kun tarkastellaan kahden muuttujan välistä riippuvuutta, on niiden välinen suhde harvoin täysin lineaarinen. Tämän vuoksi regressiomalliin kuuluu myös virhetermi, joka kuvaa havainnon arvon ja regressiomallin ennustaman arvon erotusta. Virhetermien estimaatteina toimivat residuaalit, jotka ovat avuksi tarkasteltaessa regressiomallille asetettujen oletusten toteutumista. Regressiomallin parametrien estimoimiseen esitellään pienimmän neliösumman menetelmä, joka minimoi mallin virhetermien neliösumman.

Tutkielmassa esitellään kolme eri tapaa bootstrap-menetelmän hyödyntämiseen lineaarisessa regressiossa. Menetelmät ovat bootstrap pareille, bootstrap residuaaleille ja wild-bootstrap. Niiden eroavaisuudet perustuvat tapaan, millä ne muodostavat boostrap-vastemuuttujat. Menetelmiä käytetään erilaisissa aineistossa ja vertaillaan keskenään niiden tuottamia tuloksia.

2 Bootstrap-menetelmä

Bootstrap kuuluu niiden tilastollisten menetelmien joukkoon, jotka hyödyntävät satunnaisotoksen uudelleenotantaan parametrin estimaatin jakauman muodostamisessa. Bootstrap-menetelmän tarkoituksena on muodostaa palauttaen tehtyjä otoksia estimoidusta kertymäfunktiosta. Tässä luvussa esitellään bootstrap-menetelmä ja siihen liittyviä perustuloksia. Lähteinä on käytetty teoksia Bootstrap methods and their application (Davison & Hinkley, 1997) ja An introduction to the bootstrap (Efron & Tibshirani, 1993).

2.1 Esittely

Olkoon $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ yksittäinen satunnaisotos riippumattomista ja samoin jakautuneista satunnaismuuttujista $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, joiden todennäköisyysjakauma on f ja kertymäfunktio F. Bootstrap-menetelmässä saatua satunnaisotosta hyödynnetään populaation parametrin θ estimoinnissa, merkitään

$$\theta = t(F),$$

missä $t(\cdot)$ on funktio, jolla parametri lasketaan kertymäfunktiosta F.

Parametrin estimoimiseen bootstrap-menetelmän avulla on kaksi lähestymistapaa, riippuen populaation kertymäfunktioon F liittyvistä oletuksista. Jos populaation kertymäfunktio ei ole tiedossa, hyödynnetään sen estimoinnissa sen empiiristä jakaumaa

$$\hat{F}(x) = \frac{\#\{x_i \leq x\}}{n}, \tag{2.1}$$

missä $\#\{A\}$ on tapahtumien A lukumäärä. Empiirisessä jakaumassa jokaisen otoshavainnon todennäköisyys tulla valituksi on n^{-1}. Nyt populaation parametrin estimaattina voidaan käyttää empiirisestä jakaumasta laskettua estimaattia

$$\hat{\theta} = t(\hat{F}),$$

mikä lasketaan samalla tavalla kuin θ kertymäfunktiosta F.

Toisessa lähestymistavassa satunnaismuuttujien kertymäfunktio F on määriteltävissä jollain vakioilla tai parametreillä $\boldsymbol{\psi}$. Tällöin empiirisen jakauman sijaan bootstrap-menetelmässä hyödynnetään kertymäfunktiota $F_{\hat{\boldsymbol{\psi}}}(x)$, jossa $\hat{\boldsymbol{\psi}}$ ovat kertymäfunktion estimoidut parametrit.

2.2 Epäparametrinen bootstrap

Oletetaan saaduksi yksittäinen satunnaisotos $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, jonka satunnaismuuttujat ovat riippumattomia, samoin jakautuneita ja peräisin tuntemattomasta jakaumasta F. Olkoon $\hat{\theta} = s(\boldsymbol{x})$ satunnaisotoksesta $\boldsymbol{x}$ laskettu populaation parametrin θ estimaatti. Määritellään bootstrap-otos

$$\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*), \tag{2.2}$$

joka on palauttaen tehty n-alkion kokoinen satunnaisotos empiirisestä jakaumasta $\hat{F}$. Saadusta bootstrap-otoksesta voidaan laskea parametrin θ estimaatti $\hat{\theta}^* = s(\boldsymbol{x}^*)$ samalla tavalla, kuin satunnaisotoksesta $\boldsymbol{x}$ laskettu estimaatti $\hat{\theta}$. Simuloidaan suuri määrä, B kappaletta bootstrap-otoksia $\boldsymbol{x}^{*1}, \boldsymbol{x}^{*2}, \ldots, \boldsymbol{x}^{*B}$, ja lasketaan jokaisesta otoksesta estimaattorit $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$. Kyseiset bootstrap-estimaattorit muodostavat estimaattorin $\hat{\theta}$ bootstrap-jakauman, jonka avulla voidaan määritellä esimerkiksi estimaattorin luottamusvälit, harha tai keskihajonta.

2.3 Parametrinen bootstrap

Parametrinen bootstrap tekee jonkin jakaumaoletuksen saadulle satunnaisotokselle. Olkoon $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ satunnaisotos, jonka satunnaismuuttujat ovat riippumattomia ja peräisin jakaumasta, joka koostuu tietystä parametriperheestä $\boldsymbol{\psi}$. Merkitään satunnaisotoksen tiheys- ja kertymäfunktioita $f_{\boldsymbol{\psi}}(x)$ ja $F_{\boldsymbol{\psi}}(x)$. Parametrien $\boldsymbol{\psi}$ estimoimiseen voidaan hyödyntää esimerkiksi suurimman uskottavuuden menetelmää, merkitään estimoituja parametreja $\hat{\boldsymbol{\psi}}$. Nyt satunnaisotoksen kertymäfunktiota voidaan estimoida kertymäfunktiolla $F_{\hat{\boldsymbol{\psi}}}(x)$, ja menetellä muuten samoin kuin epäparametrisessa bootstrapissa.

2.4 Tuloksia

Estimaattorin $\hat{\theta}$ harha määräytyy sen odotusarvon poikkeamana estimoitavan parametrin arvosta. Bootstrap-jakaumasta muodostettu harhan estimaatti on

$$E(\hat{\theta}^*) - \hat{\theta}, \tag{2.3}$$

missä bootstrap-estimaatin odotusarvoa puolestaan estimoidaan bootstrap-jakaumasta lasketulla keskiarvolla

$$E(\hat{\theta}^*) = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^* = \bar{\theta}^*. \tag{2.4}$$

Estimaattori bootstrap-jakauman varianssille on

$$\hat{\sigma}^2(\hat{\theta}^*) = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_i^* - \bar{\theta}^*)^2. \tag{2.5}$$

2.5 Katsaus R-ohjelmiston tarjoamiin pakkauksiin

R-ohjelmisto tarjoaa useita eri ohjelmakirjastoja bootstrap-menetelmää varten. Jotkin sen pakkaukset pohjautuvat julkaistuihin kirjoihin bootstrap-menetelmästä, toiset taas monimutkaisempien tilastollisten menetelmien sovelluksiin kyseisestä aiheesta.

2.5.1 Ohjelmakirjasto bootstrap

R-ohjelmiston bootstrap kirjasto tarjoaa funktioita bootstrap-menetelmää varten. Pakkauksen kuvauksessa sitä ei suositella uusien projektien hyödyntämisessä, vaan ensisijaisesti avuksi teoksen (Efron & Tibshirani, 1997) lukemiseen, johon kirjaston funktiot muutenkin pohjautuvat. Kirjaston bootstrap funktion avulla voidaan toteutaa epäparametrinen bootstrap halutulle aineistolle, parametrista vaihtoehtoa ei pakkauksesta löydy. Bootstrap-jakaumasta laskettuja luottamusvälejä voidaan muodostaa funktioiden abcnon ja bcanon avulla.

2.5.2 Ohjelmakirjasto boot

Boot on hyvin monipuolinen R-ohjelmiston tarjoama paketti bootstrap-menetelmän käyttöön. Sen funktiot ja aineistot pohjautuvat teokseen (Davison & Hinkley, 1997). Verrattuna ohjelmistokirjastoon bootstrap, boot tarjoaa huomattavasti enemmän funktioita bootstrap-menetelmän eri sovelluksiin. Funktion boot avulla voidaan toteuttaa niin parametrinen kuin epäparametrinen bootstrap halutulle aineistolle. Pakkauksen boot.ci funktion avulla voidaan muodostaa epäparametrisiä luottamusvälejä bootstrap-jakaumasta yhteensä viidellä eri tavalla. Pakkaus sisältää myös useita funktioita kuvaajien piirtoon, esimerkiksi funktio plot.boot muodostaa kuvaajan saadusta bootstrap-jakaumasta.

2.5.3 Muita ohjelmakirjastoja

Ohjelmakirjasto lmboot mahdollistaa bootstrap-menetelmän käytön lineaaristen mallien muodostamisessa. Mallien parametrien estimoinnissa hyödynnetään pienimmän neliösumman menetelmää. Paketin sisältämien funktioiden avulla pystytään muodostamaan boostrap-jakaumia mallin parametreille ja testaamaan parametreihin liittyviä hypoteesejä. Muita hyödyllisiä ohjelmistokirjastoja bootstrap-menetelmän sovelluksiin ovat esimerkiksi klusterointiin tarkoitettu paketti ClusBoot ja Bayes-tilastotiedettä hyödyntävä ohjelmakirjasto bayesboot.

3 Lineaarinen regressiomalli

Tässä luvussa esitellään aluksi yksinkertainen lineaarinen regressiomalli, sekä mallin parametrien estimoinnissa hyödynnettävää pienimmän neliösumman menetelmää. Jäljempänä esitellään bootstrap-menetelmän hyödyntämistä lineaarisessa regressiossa. Teorian lähteinä kappaleissa (3.1), (3.2) ja (3.3) on käytetty teoksia Regression Analysis: Theory, Methods, and Applications (Anish & Srivastava, 1990) ja An Introduction to the Bootstrap (Efron & Tibshirani, 1993).

3.1 Yksinkertainen lineaarinen regressiomalli

Regressioanalyysissä tarkastellaan selittävien muuttujien vaikutusta vastemuuttujaan. Lineaarisessa regressiossa vastemuuttujan ja selittävän muuttujan välillä oletetaan olevan lineaarinen yhteys. Olkoon $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ vektori, jossa jokainen alkio on itsessään pari

$$x_i = (\boldsymbol{c}_i, y_i), \tag{3.1}$$

missä $\boldsymbol{c}_i = (c_{i1}, c_{i2}, \ldots, c_{ip})$ on vektori, joka koostuu vakiotermistä ja riippumattomasta selittävästä muuttujasta ja y_i riippuva vastemuuttuja. Määritellään yksinkertainen lineaarinen regressiomalli

$$y_i = \boldsymbol{c}_i\boldsymbol{\beta} + \varepsilon_i = \beta_0 + \beta_1 z_i + \varepsilon_i, \tag{3.2}$$

missä $\boldsymbol{c}_i = (1, z_i)$ ja $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. Parametri β_0 on mallin vakiotermi, β_1 kulmakerroin, z_i selittävä muuttuja ja ε_i virhetermi. Regressiomallissa virhetermit ε_i määritellään havaintojen y_i poikkeamana ehdollisesta odotusarvosta $E(y|\boldsymbol{c}_i)$, jolloin siis

$$\varepsilon_i = y_i - E(y|\boldsymbol{c}_i). \tag{3.3}$$

Regressiomallissa selittävä muuttuja $\boldsymbol{c}_i$ ajatellaan kiinteäksi. Virhetermit oletetaan olevan toisistaan riippumattomia ja identtisesti jakautuneita noudattaen tuntematonta jakaumaa F

$$\varepsilon_i \sim F, \tag{3.4}$$

missä virhetermien odotusarvon oletetaan olevan 0 ja varianssin jokin tuntematon vakio σ^2. Lineaarisessa regressiomallissa vastemuuttujan y_i ehdollinen odotusarvo havaituilla selitettävän muuttujan $\boldsymbol{c}_i$ arvoilla on

$$\begin{aligned} E(y_i|\boldsymbol{c}_i) &= E(\boldsymbol{c}_i\boldsymbol{\beta} + \varepsilon_i|\boldsymbol{c}_i) = E(\boldsymbol{c}_i\boldsymbol{\beta}|\boldsymbol{c}_i) + E(\varepsilon_i|\boldsymbol{c}_i) \\ &= \boldsymbol{c}_i\boldsymbol{\beta}, \end{aligned} \tag{3.5}$$

mikä on seurausta kohtien (3.2) ja (3.4) tuloksista. Samojen kohtien avulla voidaan johtaa vastemuuttujan y varianssi

$$var(y_i|\boldsymbol{c}_i) = var(\beta_0 + \beta_1 z_i + \varepsilon_i) = var(\varepsilon_i) = \sigma^2. \tag{3.6}$$

Koska regressiomallin parametrit ajatellaan kiinteiksi, muodostuu selittävän muuttujan ehdollinen varianssi virhetermien varianssista.

3.2 Pienimmän neliösumman menetelmä

Halutaan estimoida regressiomallin parametrejä $\boldsymbol{\beta}$ havaituilla vastemuuttujan ja selittävien muuttujien arvoilla. Pienimmän neliösumman menetelmä minimoi virhetermien summan neliön

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 z_{i1})^2, \tag{3.7}$$

jonka derivoimalla ja laskemalla nollakohdat saadaan helposti muodostettua parametrien β_0 ja β_1 estimaatit $\hat{\beta}_0$ sekä $\hat{\beta}_1$ (ks. Sen & Srivastava 1990, s. 8).

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{z}, \tag{3.8}$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} (y_i - \bar{y})(z_i - \bar{z}) / \sum_{i=1}^{n} (z_i - \bar{z})^2. \tag{3.9}$$

Nyt saadaan jokaisen y_i arvolle sen ennustettu arvo $\hat{y}_i$

$$\hat{y}_i = \boldsymbol{c}_i \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + z_i \hat{\beta}_1. \tag{3.10}$$

Määritellään seuraavaksi residuaalit e_i, jotka ovat havaintojen y_i poikkeama ennustetusta vastemuuttujan arvosta $\hat{y}_i$

$$e_i = y_i - \hat{y}_i. \tag{3.11}$$

Residuaalien avulla voidaan tutkia mallin oikeellisuutta tarkastelemalla toteuttavatko residuaalit virhetermeille asetettuja oletuksia, esimerkiksi onko virhetermien varianssi vakiomuotoinen. Tavanomainen estimaatti virhetermien varianssille σ^2 on residuaalinen jäännösvarianssi

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2. \tag{3.12}$$

Mallin (3.2) estimoitujen parametrien $\hat{\beta}_0$ ja $\hat{\beta}_1$ perusominaisuudet saadaan helposti johdettua (ks. Sen & Srivastava 1990, s. 21).

$$E(\hat{\beta}_0) = \beta_0, \quad var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{z}^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}\right), \tag{3.13}$$

$$E(\hat{\beta}_1) = \beta_1, \quad var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}. \tag{3.14}$$

Koska virhetermien varianssin σ^2 on usein tuntematon, käytetään sen tilalla residuaalien jäännövarianssia $\hat{\sigma}^2$. Saaduista tuloksista huomataan, että jos regressiomallin virhetermeille asetetut oletukset ovat kunnossa, ovat pienimmän neliösumman menetelmän avulla saadut regressiomallin parametrien estimaatit harhattomia.

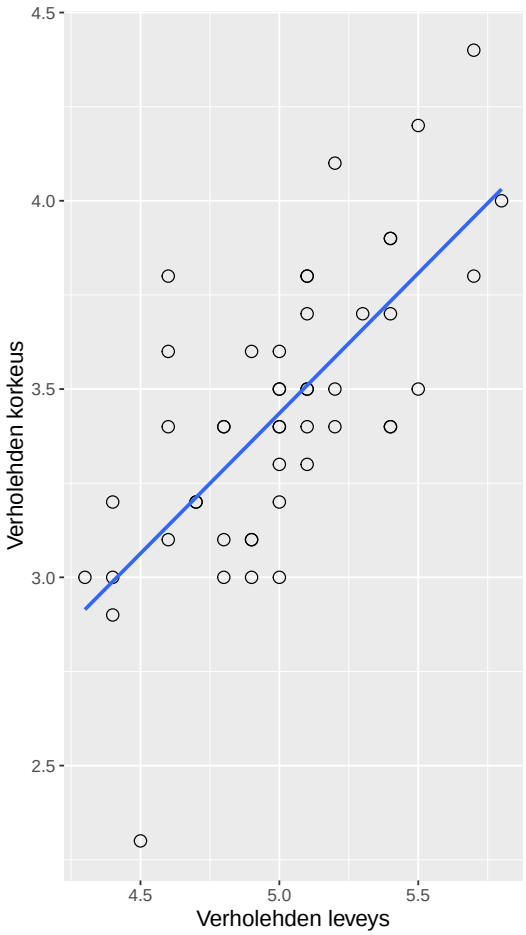
Esimerkki 3.1. Esimerkkiaineistona toimii alunperin Fisherin (1963) käyttämä aineisto, joka koostuu kukkien verholehdistä. Aineisto sisältää tiedot yhteensä 48 kukan verholehtien pituuksista ja leveyksistä, jotka löytyvät taulukoituna liitteestä A.

Estimoidaan verholehden korkeutta yksinkertaisen lineaarisen regressiomallin avulla, jossa selitettävä muuttuja on verholehden korkeus ja selittävä muuttuja verholehden leveys. Käytetään mallin muodostamisessa R-ohjelmiston lm-funktiota, jonka avulla saadaan muodostettua mallin parametrien estimaatit sekä keskihajonnat hyödyntäen pienimmän neliösumman menetelmää. Taulukkoon (3.1) on raportoitu saadut tulokset.

Taulukko 3.1. Estimoidut parametrien arvot sekä keskihajonnat

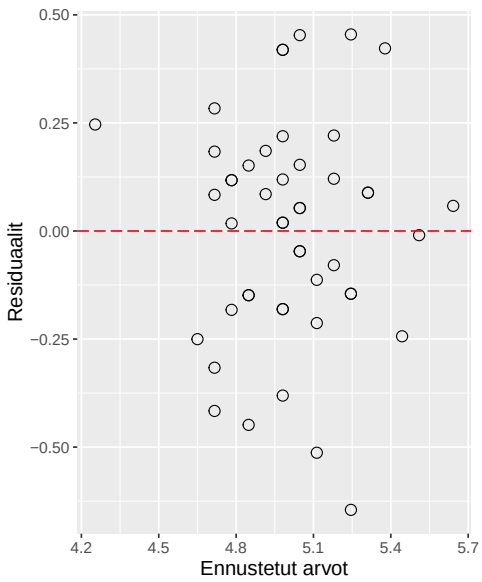
	Estimaatti	Keskihajonta
$\hat{\beta}_0$	2.733	0.343
$\hat{\beta}_1$	0.661	0.099

Kuvassa (3.1) on aineistoon sovitettu regressiomalli pienimmän neliösumman menetelmän tuottamilla arvoilla. Havaitaan, että verholehtien korkeudella ja leveydellä on selvästi lineaarinen yhteys, ja että estimoitu regressiosuora asettuu hyvin havaintopisteiden lomaan.



Kuva 3.1. Regressiosuora sovitettuna aineistoon

Voidaan myös tutkia virhetermeihin liittyvien oletuksien toteutumista. Tarkastellaan virhetermien varianssin vakiomuoto oletusta muodostamalla kuvaaja residuaaleista ja vastemuuttujan estimaateista. Havaintopisteiden pitäisi olla keskittyneenä vaakasuoran suoran ympärille. Kuvasta (3.2) nähdään, että pisteet eivät leviä tai suppene ennustettujen arvojen kasvaessa, joten oletus näyttäisi olevan kunnossa.



Kuva 3.2. Mallin residuaalit sovitettuna ennustettuihin arvoihin

3.3 Regressiomalli matriisi notaatiolla

Kohdan (3.2) yksinkertainen lineaarinen regressiomalli voidaan esittää helposti myös matriisimuodossa, merkitään

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.15}$$

missä $\boldsymbol{y}$ on $n \times 1$ vastemuuttujavektori, $\boldsymbol{C} = (\mathbf{1}, \boldsymbol{Z})$ on $n \times 2$ havaintomatriisi, $\boldsymbol{\varepsilon}$ on $n \times 1$ virhetermeistä koostuva vektori ja $\boldsymbol{\beta}$ on 2×1 parametrivektori, jonka estimaatit pienimmän neliösumman menetelmällä saadaan kaavasta (ks. Sen ja Srivastava 1990, s. 30)

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{y}. \tag{3.16}$$

Oletuksena tälle kaavalle on, että $\boldsymbol{C}$ on täysiasteinen, jolloin sen käänteismatriisi on olemassa. Myös parametrien varianssit saadaan helposti laskettua matriiseilla

$$var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T var(\boldsymbol{y})\boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{C})^{-1} = \sigma^2(\boldsymbol{C}^T\boldsymbol{C})^{-1}, \tag{3.17}$$

missä

$$var(\boldsymbol{y}) = var(\boldsymbol{C}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{I}\sigma^2.$$

Matriisi $\boldsymbol{I}$ on $n \times n$ identiteettimatriisi.

3.4 Bootstrap-menetelmä lineaarisessa regressiossa

Tässä kappaleessa tarkastellaan kahta eri bootstrap-menetelmää lineaarisessa regressiossa. Molemmat menetelmät hyödyntävät havaintoaineiston empiiristä jakaumaa bootstrap-otosten muodostamisessa. Mallin parametrien estimoinnissa käytetään pienimmän neliösumman menetelmää.

3.4.1 Bootstrap residuaaleille

Bootstrap residuaaleille hyödyntää regressiomallin residuaaleja bootstrap-otoksen muodostamisessa. Olkoon $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ kohdan (3.1) mukainen riippumaton satunnaisotos jakaumasta P

$$P = (\boldsymbol{\beta}, F), \tag{3.18}$$

missä $\boldsymbol{\beta}$ on vektori, joka koostuu regressiomallin parametreista ja F virhetermien jakauma. Halutaan estimoida jakaumaa P. Regressiomallin parametrien estimaatit $\hat{\boldsymbol{\beta}}$ saadaan pienimmän neliösumman menetelmän avulla. Virhetermien jakaumaa voidaan estimoida residuaalien empiirisellä jakaumalla $\hat{F}$. Yleensä empiirisen jakauman odotusarvon oletetaan olevan 0, kuten vaadittiin virhetermien odotusarvolle (3.4). Muodostetaan seuraavaksi palauttaen tehty satunnaisotos $\boldsymbol{e}^* = (e_1^*, e_2^*, \ldots, e_n^*)$ residuaalinen empiirisestä jakaumasta $\hat{F}$. Tämän jälkeen muodostettaan bootstrap-vastemuuttujat kohdan (3.2) mukaan

$$y_i^* = \boldsymbol{c}_i\hat{\boldsymbol{\beta}} + e_i^* = \hat{\beta}_0 + \hat{\beta}_1 z_i + e_i^*, \qquad i = 1, 2, \ldots, n. \tag{3.19}$$

Saatu Bootstrap-otos $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_n^*)$ on siis muotoa

$$\boldsymbol{x}^* = \{(\boldsymbol{c}_1, \boldsymbol{c}_1\hat{\boldsymbol{\beta}} + e_{i_1}), (\boldsymbol{c}_2, \boldsymbol{c}_2\hat{\boldsymbol{\beta}} + e_{i_2}), \ldots, (\boldsymbol{c}_n, \boldsymbol{c}_n\hat{\boldsymbol{\beta}} + e_{i_n})\}, \tag{3.20}$$

missä $i_1, i_2, \ldots, i_n$ on palauttaen tehty satunnaisotos joukosta $\{1, 2, \ldots, n\}$. Koska lineaarisessa regressiossa selittävät muuttujat $\boldsymbol{c}_i$ ajatellaan kiinteiksi, pysyvät ne muuttumattomina myös bootstrap-otoksessa. Bootstrap-menetelmällä saadut pienimmän neliösumman estimaatit $\hat{\boldsymbol{\beta}}^*$ saadaan minimoimalla summa

$$\sum_{i=1}^{n} (y_i^* - \boldsymbol{c}_i\hat{\boldsymbol{\beta}}^*)^2, \tag{3.21}$$

jotka mukaillen kaavaa (3.16) saadaan muodossa

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{y}^*. \tag{3.22}$$

(Efron & Tibshirani 1993, s. 111–112).

Simuloimalla useita bootstrap-otoksia saadaan muodostettua regressiomallin parametri estimaattien bootstrap-jakauma, josta voidaan laskea esimerkiksi keskiarvot ja keskihajonnat mallin parametreille.

Esimerkki 3.2. Hyödynnetään samaa aineistoa, kuin esimerkissä (3.1). Estimoidaan verholehden korkeutta regressiomallin avulla hyödyntäen bootstrap residuaaleille-menetelmää. Käytetään mallin muodostamisessa R-ohjelmistossa itse toteutettua funktiota resBoot (Liite B). Simuloidaan 1000 boostrap-otosta, ja lasketaan jokaisesta otoksesta estimaatit regressiomallin parametreille. Taulukkoon (3.2) on laskettu muodostetusta bootstrap-jakaumasta estimaattien odotusarvo, harha ja keskihajonta (lauseet 2.3, 2.4 ja 2.5). Harhan laskemisessa hyödynnetään pienimmän neliösumman estimaatteja, jotka laskettiin esimerkissä (3.1).

Taulukko 3.2. Estimoidut parametrien keskiarvot, harhat ja keskihajonnat

	Estimaatti	Harha	Keskihajonta
$\hat{\beta}_0^*$	2.742	0.009	0.346
$\hat{\beta}_1^*$	0.659	-0.002	0.100

3.4.2 Bootstrap pareille

Bootstrap pareille muodostaa bootstrap-otoksia vastemuuttujan ja selittävän muuttujan muodostamasta kahden muuttujan jakaumasta. Määritellään kohdassa (3.1) esitellyt vastemuuttujat $\boldsymbol{y}$ sekä vektori $\boldsymbol{c}$ olevan riippumattomia ja identtisesti jakautuneita satunnaisotoksia jakaumasta F. Estimoidaan kyseistä jakaumaa sen empiirisellä jakaumalla $\hat{F}$, joka koostuu pareista $(y_i, \boldsymbol{c}_i)$, missä jokaisen parin todennäköisyys tulla valituksi on n^{-1}. Muodostetaan bootstrap-otos kyseisestä jakaumasta, mikä on muotoa

$$\boldsymbol{x}^* = (\boldsymbol{c}_{i_1}, y_{i_1}), (\boldsymbol{c}_{i_2}, y_{i_2}), \ldots, (\boldsymbol{c}_{i_n}, y_{i_n}), \tag{3.23}$$

missä $i_1, i_2, \ldots, i_n$ on palauttaen tehty satunnaisotos joukosta $\{1, 2, \ldots, n\}$. Pienimmän neliösumman avulla voidaan muodostaa taas parametrien bootstrap-estimaatit minimoimalla summa

$$\sum_{i=1}^{n} (y_i - \boldsymbol{c}_i \hat{\boldsymbol{\beta}})^2, \tag{3.24}$$

jotka kaavan (3.16) avulla saadaan muodossa

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{C}^{*T} \boldsymbol{C}^*)^{-1} \boldsymbol{C}^{*T} \boldsymbol{y}^*. \tag{3.25}$$

(Davison ja Hinkley 1997, s. 264–265)

Koska myös selittävät muuttujat muodostetaan satunnaisotoksena empiirisestä jakaumasta, omaa matriisi $\boldsymbol{C}^*$ nyt eri muodon kuin alkuperäisessä mallissa (3.15). Ehtona edelleen on, että matriisi on täysiasteinen.

Esimerkki 3.3. Hyödynnetään edellisissä esimerkeissä käytettyä aineistoa ja sovitetaan sille lineaarinen regressiomalli hyödyntäen bootstrap pareille -menetelmää. Käytetään mallin muodostamisessa R-ohjelmistossa itse toteutettua funktiota pairsBoot (Liite B). Muodostetaan regressiomallin estimaattien bootstrap-jakauma simuloimalla 1000 bootstrap-otosta ja lasketaan jakaumasta odotusarvo, harha ja keskihajonta mallin parametreille. Tulokset löytyvät taulukosta (3.3).

Taulukko 3.3. Estimoidut parametrien keskiarvot sekä keskihajonnat

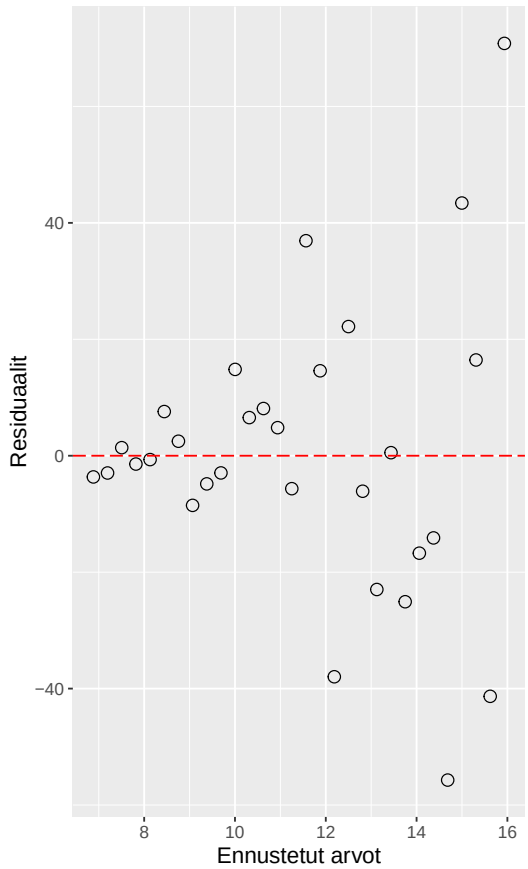
	Estimaatti	Harha	Keskihajonta
$\hat{\beta}_0^*$	2.748	0.015	0.329
$\hat{\beta}_1^*$	0.657	-0.004	0.096

3.4.3 Menetelmien vertailua

Edellä esitellyt menetelmät ovat hieman toisistaan poikkeavia. Vertailtaessa menetelmien muodostamia bootstrap-otoksia havaitaan, että bootstrap pareille ei tee lineaarisuusoletusta bootstrap-vastemuuttujalle, toisin kuin bootstrap residuaaleille -menetelmässä. Tämän vuoksi bootstrap pareille voi tuottaa kohtuullisia tuloksia vaikka regressiomalliin liittyvät oletukset eivät olisikaan kunnossa (Chernick & LaBudde 2011, s. 59). Toisaalta mallin oletusten ollessa kunnossa bootstrap residuaaleille -menetelmä pääsee hyödyntämään kyseistä tilannetta paremmin. Tämä tulee osittain ilmi myös esimerkeissä (3.2) ja (3.3), missä regressiomalliin liittyvät oletukset todettiin olevan kunnossa: bootstrap residuaaleille tuotti vähemmän harhaisia estimaatteja verrattuna bootstrap pareille -menetelmään.

3.5 Mallin sovitus heteroskedastisessa aineistoissa

Aiemmissa esimerkeissä käytetty aineisto oli varsin sopiva regressiomallin sovittamiseen. Tarkastellaan seuraavaksi bootstrap-menetelmällä saatuja tuloksia regressiomallissa simuloidulla havaintoaineistolla, jossa ilmenee vahvaa heteroskedastisuutta, eli virhetermien varianssi ei ole vakiomuotoinen. Heteroskedastisuus ei sinänsä vaikuta estimoitujen parametrien harhaan, vaan varianssin suuruuteen, mikä puolestaan vaikuttaa esimerkiksi estimaattien merkitsevyyden testaamisessa (Sen & Srivastava 1990, s. 111). Simuloitu data löytyy liitteestä (C). Residuaaleista ja estimoiduista vastemuuttujista muodostetusta kuvaajasta (3.3) nähdään selvästi, että havaintopisteet leviävät selvästi estimoitujen vastemuuttujan arvojen kasvaessa, joten virhetermien varianssi ei selvästikään ole vakiomuotoinen.



Kuva 3.3. Mallin residuaalit sovitettuna ennustettuihin arvoihin

Taulukkoon 3.4 on koottu regressiomallin parametrien estimaatit hyödyntäen kumpaakin bootstrap-menetelmää lineaarisessa regressiossa. Simuloidun datan vastemuuttujat muodostettiin yksinkertaisen lineaarisen regressiomallin avulla asettamalla vakiotermin arvoksi 1.25 ja kulmakertoimen arvoksi 2.55. Havaitaan, että bootstrap pareille -menetelmän tuottamat estimaattien arvot ovat lähempänä alkuperäisen mallin parametrien arvoja. Kuitenkin parametrien keskihajonnat ovat varsin suuria, varsinkin ensimmäisen parametrin, joten on varsin todennäköistä, että simulointi tuottaisi jollain toisella kerralla myös toisenlaisia tuloksia. Vertailemalla keskihajontoja havaitaan, että bootstrap pareille tuotti yhteensä pienemmän keskihajonnan mallin parametreille, mikä on tässä tilanteessa tärkeämpi tulos.

Taulukko 3.4. Estimoidut parametrien keskiarvot sekä keskihajonnat

Menetelmä	Estimoitava parametri	Estimoitu arvo	Keskihajonta
Bootstrap residuaaleille	β_0	6.174	10.044
	β_1	0.974	1.622
Bootstrap pareille	β_0	5.780	8.365
	β_1	1.057	2.159

3.5.1 Wild-bootstrap

Virhetermien heteroskedastisuuden mallintamiselle on monia vaihtoehtoja lineaarisessa regressiossa. Joskus heteroskedastisuutta voidaan mallintaa jollain varianssifunktiolla $V(\cdot)$, jolloin pienimmän neliösumman estimoinnin sijaan voidaan käyttää painotettua pienimmän neliösumman menetelmää regressiomallin parametrien estimoinnissa (ks. Davison & Hinkley 1997, s. 271).

Joissain tilanteissa virhetermien varianssia σ_i^2 ei pystytä mallintamaan, jolloin voidaan hyödyntää wild-boostrap -menetelmää, jossa varianssit estimoidaan yksittäisille residuaaleille. Menetelmä on samankaltainen kuin bootstrap residuaaleille, mutta nyt virhetermeistä muodostettu boostrap-otos muodostetaan kahden pisteen jakaumasta

$$P\{e_i^* = e_i(1 - \sqrt{5}/2\} = \pi, \quad P\{e_i^* = e_i(1 + \sqrt{5}/2\} = 1 - \pi, \tag{3.26}$$

missä $\pi = (5 + \sqrt{5}/10)$ ja e_i regressiomallin residuaalit. (Davison & Hinkley 1997, s. 272.)

Esimerkki 3.4. Tarkastellaan wild bootstrap -menetelmän tuottamia tuloksia heteroskedastisessa aineistossa, jota käytettiin kappaleessa (3.5). Käytetään parametrien estimoimisessa R-ohjelmistossa itse toteutettua funktiota resBoot (Liite B). Saadut estimoidut regressiomallin parametrit ja keskihajonnat ovat taulukossa (3.5). Vertailtaessa estimoituja arvoja havaitaan, että wild-bootstrap ei tuottanut tarkempia estimaatteja verrattuna muihin menetelmiin. Menetelmä tuottaa kuitenkin yhteensä pienimmän keskihajonnan mallin parametreille verrattuna muihin menetelmiin.

Taulukko 3.5. Estimoidut parametrien keskiarvot sekä keskihajonnat

	Estimaatti	Keskihajonta
β_0^*	6.335	7.860
β_1^*	0.873	2.078

4 Yhteenveto

Tutkielman alussa esiteltiin lyhyesti bootstrap-menetelmä sekä muutama perustulos koskien bootstrap-jakaumaa. Lineaarisesta regressiomallista esiteltiin sen yleistä teoriaa sekä tyypillisintä menetelmää mallin parametrien estimoinnissa, pienimmän neliösumman menetelmää.

Bootstrapin hyödyntämisessä lineaarisessa regressiossa esiteltiin aluksi kaksi menetelmää, bootstrap residuaaleille ja bootstrap pareille. Menetelmien erot perustuvat tapaan, millä ne muodostavat bootstrap-vastemuuttujat. Bootstrap residuaaleille -menetelmä hyödyntää mallin lineaarisuusoletusta bootstrap-vastemuuttujan muodostamisessa, kun taas bootstrap pareille -menetelmä ei tee tällaista oletusta. Menetelmät toteutettiin R-ohjelmistolla itse tehdyillä funktioilla. Aluksi menetelmiä tarkasteltiin aineistossa, missä regressiomallille asetetut oletukset toteutuivat. Vertailtaessa menetelmien tuottamia estimaattien arvoja harhattomiin pienimmän neliösumman menetelmän estimaatteihin havaittiin, että bootstrap residuaaleille tuotti vähemmän harhaiset estimaattien arvot kummallekkin parametrille.

Vaikka pienimmän neliösumman menetelmä tarjoaa joissain tilanteissa erittäin toimivan ratkaisun regressiomallin parametrien estimoimiseen, on se herkkä tilanteille, joissa regerssiomalliin liittyvät oletukset eivät toteudu. Yksi näistä tilanteista on virhetermien heteroskedastisuus, joka aiheuttaa suuren varianssin estimoiduille regressiomallin parametreille. Viimeisessä kappaleessa tarkasteltiin bootstrap residuaaleille ja bootstrap pareille -menetelmien tuottamia tuloksia simuloidulla havaintoaineistolla, missä esiintyi heteroskedastisuutta virhetemien osalta. Havaittiin, että bootstrap pareille tuotti yhteensä pienemmän varianssin estimoiduille parametreille. Tutkielman lopussa esiteltiin vielä wild-bootstrap -menetelmä, joka soveltuu parhaiten tilanteisiin, jossa virhetermien varianssia ei pystytä kunnolla mallintamaan, johtuen esimerkiksi aineiston pienestä koosta. Menetelmää testattiin samalla simuloidulla aineistolla, ja havaittiin sen tuottavan pienimmän varianssin mallin parametreille verrattuna kahteen muuhun menetelmään.

Lähteet

[1] Chernick, M. R. & LaBudde, R. A. (2011). *An introduction to bootstrap methods with applications to R.* 1st edition. Hoboken, New Jersey: Wiley.

[2] Davison, A. C. & Hinkley, D. V (1997). *Bootstrap Methods and Their Application.* New York: Cambridge University Press.

[3] Efron, B. & Tibshirani, R. J (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

[4] Fisher, R. A (1963). *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, 7, s. 179–188.

[5] Sen, A & Srivastava, M. (1990). *Regression Analysis: Theory, Methods, and Applications.* New York: Springer.

Liite A: Kukkien verholehtien korkeudet ja leveydet

Kasvi	Korkeus	Leveys	Kasvi	Korkeus	Leveys
1.	3.5	5.1	25.	3.0	5.0
2.	3.0	4.9	26.	3.4	5.0
3.	3.2	4.7	27.	3.5	5.2
4.	3.1	4.6	28.	3.4	5.2
5.	3.6	5.0	29.	3.2	4.7
6.	3.9	5.4	30.	3.1	4.8
7.	3.4	4.6	31.	3.4	5.4
8.	3.4	5.0	32.	4.1	5.2
9.	2.9	4.4	33.	4.2	5.5
10.	3.1	4.9	34.	3.1	4.9
11.	3.7	5.4	35.	3.2	5.0
12.	3.4	4.8	36.	3.5	5.5
13.	3.0	4.3	37.	3.6	4.9
14.	4.0	5.8	38.	3.0	4.4
15.	4.4	5.7	39.	3.4	5.1
16.	3.9	5.4	40.	3.5	5.0
17.	3.5	5.1	41.	2.3	4.5
18.	3.8	5.7	42.	3.2	4.4
19.	3.8	5.1	43.	3.5	5.0
20.	3.4	5.4	44.	3.8	5.1
21.	3.7	5.1	45.	3.0	4.8
22.	3.6	4.6	46.	3.8	4.6
23.	3.3	5.1	47.	3.7	5.3
24.	3.4	4.8	48.	3.3	5.0

Liite B: R-koodi

```
# Markus Linjamäki
# 18.3.2021

## Bootsrap residuals in regression model
# @ param model: regressio model lm.function
# @ param iterations: number of iterations
# @ param wild: wild-bootstrap
##
# return: dataframe including simulated beta values

resBoot <- function(model,iterations, wild = FALSE){
  betas <- as.numeric(coefficients(model))
  residuals <- resid(model)
  pi <- (5 + sqrt(5)) / 10
  len <- length(residuals)

  C <- model.matrix(model) # model in matrix form
  # simulated beta values
  boot_Betas <- matrix(nrow = iterations, ncol = length(betas))
  for(i in 1:iterations){
    # residual sample with replacement
    residualSample <- sample(residuals,len,replace = T)
    if(wild){
      wildSample <- sample(c((1-sqrt(5)) / 2, (1+sqrt(5)) / 2),
                           size = len, replace = T, prob = c(pi,1-pi))
      residualSample <- residuals * wildSample
    }
    # Bootstrap responses
    y_star <- C %*% betas + residualSample

    # bootstrap least-square estimates
    beta_star <- solve(t(C) %*% C) %*% t(C) %*% y_star

    for(a in 1:length(betas)){
      boot_Betas[i,a] <- beta_star[a]
    }
  }
  return(as.data.frame(boot_Betas))
}

## Bootsrap pairs in regression model
# @ param model: regressio model lm.function
```

```r
# @ param iterations: number of iterations
##
# return: dataframe including simulated beta values

pairsBoot <- function(model, iterations){
  betas <- as.numeric(coefficients(model))
  n <- length(model$residuals)
  i_set <- seq(1,n,1)
  boot_Betas <- matrix(nrow = iterations, ncol = length(betas))

  ## not sampled values
  y_values <- as.matrix(model.frame(model)[1])
  c_values <- as.matrix(model.frame(model)[,-1])

  # sampled values
  c_vec <- matrix(nrow = length(c_values), ncol = length(c_values[1]))
  y_vec <- matrix(nrow = n)
  # for the C matrix
  ones <- rep(1,n)

  for(a in 1:iterations){
    # for every iteration new sample
    i_bootSample <- sample(i_set,n,replace = T)
    for(i in 1:n){
      index <- i_bootSample[i]
      y_vec[i] <- y_values[index]
      c_vec[i,] <- c_values[index,]
    }
    C <- cbind(ones,c_vec)
    beta_star <- solve(t(C) %*% C) %*% t(C) %*% y_vec

    for(o in 1:length(betas)){
      boot_Betas[a,o] <- beta_star[o]
    }

  }
  return(as.data.frame(boot_Betas))
}
```

Liite C: Simuloitu data

```
x <- seq(1,10,length = 30)
beta_0 <- 1.250
beta_1 <- 2.550
sigma <- x^1.68
err <- rnorm(30, 0, sigma)
y <- beta_0 + beta_1 * x + err
```