Box–Cox-muunnos

Siiri Pollari

LuK-tutkielma
Toukokuu 2026

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:
FT Pekka Nieminen

TURUN YLIOPISTO, Matematiikan ja tilastotieteen laitos

LuK-tutkielma
Pääaine: Tilastotiede
Tekijä: Siiri Pollari
Otsikko: Box–Cox-muunnos
Sivumäärä: 20 sivua
Ohjaaja: FT Pekka Nieminen
Aika: Toukokuu 2026

Tässä tutkielmassa käsitellään Box–Cox-muunnosta. Box–Cox-muunnoksen tavoitteena on löytää sopiva muunnos tilastollisen mallin vastemuuttujalle potenssimuunnosten perheestä, jotta malliin liitetyt oletukset toteutuisivat paremmin. Box–Cox-muunnosta hyödynnetään erityisesti virhetermien heteroskedastisuuden vähentämiseen, normaalisuusoletuksen parantamiseen sekä vastemuuttujan ja selittävien muuttujien välisen suhteen linearisoimiseen.

Tutkielman alussa esitellään Box–Cox-muunnoksen matemaattinen kehikko ja esitellään sen sovelluskohteita. Tämän jälkeen käsitellään erikoistapauksena lineaarista regressiomallia ja Box–Cox-muunnoksen hyödyntämistä lineaarisen regressiomallin yhteydessä. Lisäksi muunnoksen toimivuutta lineaarisessa regressiomallissa havainnollistetaan aineistoesimerkkien avulla. Tutkielman lopuksi esitellään muutamia Box–Cox-muunnoksen pohjalta kehitettyjä laajennuksia, jotka monipuolistavat muunnoksen käyttömahdollisuuksia.

Sisällys

1 Johdanto

Usein lineaarisiin malleihin liitetään oletus mallin lineaarisesta rakenteesta, virhetermien normaalisuudesta sekä virhetermien homoskedastisuudesta. Näiden oletusten pohjalta Box ja Cox esittelivät tutkimusartikkelissaan *An Analysis of Transformations* [2] (1964) Box–Cox-muunnoksen, jonka tavoitteena on löytää sopiva potenssimuunnos mallin vastemuuttujalle, mikäli edellä mainitut oletukset eivät alkuperäisen aineiston pohjalta rakennetussa mallissa toteudu. Box–Cox-muunnosta voidaan hyödyntää laaja-alaisesti erilaisissa tilastollisissa malleissa, kuten aikasarjamalleissa, yleistetyissä lineaarisissa malleissa tai lineaarisissa sekamalleissa. Lisäksi muunnosta voidaan hyödyntää selittävien muuttujien muuntamiseen.

Tämän tutkielman tavoitteena on ensin yleisesti perehtyä Box–Cox-muunnokseen. Menetelmän yleisin sovelluskohde on usein kuitenkin lineaarinen regressiomalli, jonka vuoksi tutkielmassa tarkastellaan tarkemmin lineaarisen regressiomallin matemaattista rakennetta sekä parametrien estimointia suurimman uskottavuuden menetelmällä. Tämän jälkeen käydään läpi, miten Box–Cox-muunnosta voidaan hyödyntää lineaarisen regressiomallin yhteydessä ja havainnollistetaan tätä kahdella aineistoesimerkillä.

Box–Cox-muunnoksella on kuitenkin rajoitteensa, koska menetelmä on alun perin määritelty vain positiivisille vastemuuttujan arvoille. Box–Cox-muunnoksen pohjalta on täten kehitetty lukuisia erilaisia laajennuksia, joissa mahdollistetaan esimerkiksi muunnosten tekeminen negatiivisille vastemuuttujan arvoille tai selittävien muuttujien ja vastemuuttujan yhtäaikainen muuntaminen. Muunnoksen eri laajennuksiin tutustutaan lyhyesti tutkielman viimeisessä kappaleessa.

Työn keskeisenä lähteenä on käytetty Neterin ym. teosta *Applied Linear Statistical Models* [6] (1996), jota hyödynnetään etenkin Box–Cox-muunnoksen ja lineaarisen regressiomallin määrittelyssä työn alkuvaiheessa. Tämän lisäksi työn tärkeimpiin lähteisiin kuuluvat Sheatherin teos *A Modern Approach to Regression with R* [8] (2009), jota hyödynnetään parametrien estimoinnissa luvussa 3 sekä Davisonin teos *Statistical Models* [4] (2003), jota seurataan etenkin parametrien estimoinnissa luvussa 4.

2 Box–Cox-muunnos

Box–Cox-muunnoksen tavoitteena on löytää sopiva muunnos tilastollisen mallin vastemuuttujalle *potenssimuunnosten perheestä* (engl. *the family of power transformations*), jotta kyseiseen malliin liitetyt oletukset toteutuvat paremmin. Box–Cox-muunnosta hyödynnetään etenkin malleissa, joiden oletuksena on vastemuuttujan odotusarvon rakenteen yksinkertaisuus, virhetermien normaalisuus sekä virhetermien homoskedastisuus. Vastemuuttujan odotusarvon rakenteen yksinkertaisuus voi tarkoittaa esimerkiksi sitä, että vastemuuttujan odotusarvo esitetään lineaarisena selittävien muuttujien funktiona tai muuna vastaavana riittävän yksinkertaisena funktiona. Tämän luvun pääasiallisina lähteinä on hyödynnetty alkuperäistä Boxin ja Coxin artikkelia [2] sekä kirjan [6] lukuja 3.8 ja 3.9.

Tarkastellaan seuraavaksi aineistoa, joka sisältää yhteensä n havaintoyksikköä. Merkitään havaintoyksiköitä indeksillä $i = 1, \ldots, n$ ja satunnaismuuttujien Y_i realisoituneita arvoja eli vastemuuttujan arvoja merkinnällä y_i. Määritellään potenssimuunnosten perheen muunnokset näiden merkintöjen avulla seuraavasti:

$$y_i^* = y_i^\lambda \tag{1}$$

jossa $y_i > 0$, λ on reaalinen parametri, joka määritellään aineiston perusteella ja y_i^* on muunnetun vastemuuttujan arvo havaintoyksikölle i. Perheeseen kuuluvia muunnoksia ovat esimerkiksi:

$\lambda = 2$	$y_i^* = y_i^2$	
$\lambda = 0.5$	$y_i^* = \sqrt{y_i}$	
$\lambda = 0$	$y_i^* = \log(y_i)$	(sopimuksen mukaisesti)
$\lambda = -0.5$	$y_i^* = \dfrac{1}{\sqrt{y_i}}$	
$\lambda = -1.0$	$y_i^* = \dfrac{1}{y_i}$	

Box–Cox-muunnoksen etuna on, että se muodostaa jatkuvan muunnosperheen, joka sisältää useita yleisesti käytettyjä muunnoksia erityistapauksinaan. Tällaisia ovat esimerkiksi logaritmi-, juuri- ja käänteislukumuunnokset. Box–Cox-muunnos määritellään positiivisille vastemuuttujan y_i arvoille seuraavasti:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & (\lambda \neq 0), \\ \log(y_i) & (\lambda = 0) \end{cases} \tag{2}$$

jossa $y_i > 0$. Box–Cox-muunnosta (2) on suositeltavampaa käyttää teoreettisen tarkastelun kannalta kuin tavallista potenssimuunnosten määritelmää (1), koska muunnos on jatkuva kohdassa $\lambda = 0$. Tämä voidaan todeta tarkastelemalla muunnoksen raja-arvoa, kun parametri λ lähestyy nollaa:

$$\begin{aligned} \lim_{\lambda \to 0} \frac{y_i^\lambda - 1}{\lambda} &= \lim_{\lambda \to 0} \frac{e^{\lambda \log(y_i)} - e^0}{\lambda} = \lim_{\lambda \to 0} \frac{e^{\lambda \log(y_i)} - e^{0 \cdot \log(y_i)}}{\lambda - 0} \\ &= D\left[e^{\lambda \log(y_i)}\right]_{\lambda = 0} = e^{0 \cdot \log(y_i)} \cdot \log(y_i) = \log(y_i). \end{aligned}$$

Lisäksi esimerkiksi lineaaristen mallien tapauksessa ykkösellä vähentämisellä ja parametrilla λ jakamisella ei ole merkittävää vaikutusta mallin tulkintaan, jolloin muunnokset [1] ja [2] ovat käytännössä ekvivalentit[3, s. 84]. Box–Cox-muunnos säilyttää myös vastamuuttujan ja selittävän muuttujan välisen suhteen suunnan kaikilla parametrin λ arvoilla, mikä helpottaa lineaarisissa malleissa parametrien tulkintaa. Tämä tarkoittaa, että Box–Cox-muunnos on kasvava funktio kaikilla parametrin λ arvoilla, kun taas tavallisessa potenssimuunnosten määritelmässä muunnos on vähenevä muuttujan y suhteen, kun $\lambda < 0$. [12, s. 151]

Box–Cox-muunnoksen hyödyntäminen ei kuitenkaan rajoitu lineaarisen regressiomallin tapaukseen. Muunnos keskittyy etenkin mallin virhetermien homoskedastisuuden ja normaalisuuden parantamiseen eikä niinkään itse mallin rakenteeseen. Tämän vuoksi Box–Cox-muunnosta voidaan hyödyntää myös muiden mallien, kuten aikasarjamallien, epälineaaristen mallien, sekamallien tai yleistettyjen lineaaristen mallien kontekstissa esimerkiksi virhetermien homoskedastisuuden ja normaalisuuden parantamiseen. Kirjallisuudessa Box–Cox-muunnos usein kuitenkin esitetään lineaarisen regressiomallin yhteydessä, minkä vuoksi luvuissa 3 ja 4 keskitytään Box–Cox-muunnoksen hyödyntämiseen lineaarisen regressiomallin tapauksessa.

3 Lineaarinen regressiomalli

Lineaarinen regressiomalli on laaja-alaisesti käytetty tilastotieteen menetelmä, jota hyödynnetään monilla eri tilastotieteen sovellusaloilla, kuten biologiassa, sosiaalitieteissä ja taloustieteessä. Lineaarisen regressiomallin avulla pyritään vastaamaan ongelmiin, joissa selvitetään kahden tai useamman muuttujan välisiä riippuvuuksia. Mallin selittävät muuttujat voivat olla jatkuvia tai luokka-asteikollisia sekä näiden yhdysvaikutuksia. Vastemuuttuja puolestaan on lähtökohtaisesti jatkuva numeerinen muuttuja. Määritellään seuraavaksi lineaarinen regressiomalli, mitä oletuksia siihen liittyy sekä miten mallin parametreja voidaan estimoida. Luvun pääasiallisina lähteinä on hyödynnetty kirjan [6] lukua 6. Lisäksi alaluvuissa 3.2 ja 3.3 on hyödynnetty kirjaa [8, s. 21, 90-91].

3.1 Lineaarisen regressiomallin määrittely

Tarkastellaan edelleen aineistoa, joka koostuu satunnaismuuttujien Y_i realisoituneista arvoista y_i sekä selittävien muuttujien havaituista arvoista $x_{i,0}, \ldots, x_{i,p}$. Aineistossa on siis yhteensä n havaintoyksikköä ja $p+1$ selittävää muuttujaa. Lineaarinen regressiomalli kertoo, miten satunnaismuuttujan Y_i jakauma riippuu selittävien muuttujien arvoista ja se määritellään havaintoyksikölle i seuraavasti:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

jossa $i = 1, \ldots, n$ on havaintoyksikön indeksi ja lisäksi:

- $\beta_0, \ldots, \beta_p$ ovat mallin kiinteitä, mutta tuntemattomia parametreja, joiden arvoja yritetään estimoida. Parametri β_0 on mallin vakiotermi, jota vastaava selittävä muuttuja saa arvon yksi kaikilla havaintoyksiköillä.

- $x_{i,0}, \ldots, x_{i,p}$ ovat mallin selittävien muuttujien arvoja havaintoyksikölle i, joiden oletetaan olevan tunnettuja kiinteitä vakioita. Vakiotermiä vastaavaa selittäjää $x_{i,0} = 1$ kutsutaan vakioselittäjäksi.

- ε_i on havaintoyksikön i virhetermi. Se kuvaa mallin satunnaisosaa eli vastemuuttujan havaitun arvon ja mallin ennustaman arvon erotusta.

. Vastemuuttujan odotusarvon ja mallin selittävien muuttujien suhde on muotoa

$$E[Y_i] = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}.$$

Huomionarvoista on, että mallin lineaarisuus liittyy nimenomaan vastemuuttujan ja parametrien β_j väliseen riippuvuuteen, jolloin esimerkiksi selittävien muuttujien korottaminen potenssiin mallin linearisoimiseksi on mahdollista. Lineaarinen regressiomallin voidaan lisäksi esittää matriisimuodossa seuraavasti:

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

jossa $\mathbf{Y}$ on $(n \times 1)$ vektori, joka koostuu satunnaismuuttujista Y_i, $\boldsymbol{X}$ on $(n \times (p+1))$ matriisi, joka koostuu selittävien muuttujien havaituista arvoista sekä vakioselittäjästä, $\boldsymbol{\beta}$ on $((p+1) \times 1)$ vektori, joka koostuu mallin parametreista, ja $\boldsymbol{\varepsilon}$ on $(n \times 1)$ vektori, joka koostuu mallin virhetermeistä. Esimerkiksi yhden selittävän muuttujan lineaarinen regressiomalli voidaan matriisimuodossa kirjoittaa seuraavasti:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

eli lyhyesti

$$\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon}.$$

3.2 Lineaarisen regressiomallin oletukset

Lineaariseen regressiomalliin liittyy tiettyjä oletuksia, joiden täytyy toteutua, jotta mallista tehtyjä päätelmiä voidaan pitää luotettavina. Tämän vuoksi oletusten toteutumista on syytä tarkastella aina uuden mallin muodostamisen jälkeen. Käydään nämä oletukset seuraavaksi tarkemmin läpi:

1) Lineaarisuus: Vastemuuttujan odotusarvon ja selittävien muuttujien välinen suhde on muotoa $E[Y_i] = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$.

2) Homoskedastisuus: Mallin virhetermeillä ε_i oletetaan olevan vakio varianssi σ^2 riippumatta havaintoyksiköstä. Tästä seuraa, että myös mallin vastemuuttujilla Y_i on sama varianssi σ^2.

3) Normaalisuus: Mallin virhetermit ε_i noudattavat normaalijakaumaa, jonka odotusarvo on nolla ja varianssi σ^2. Yhtäpitävästi $Y_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$.

4) Riippumattomuus: Mallin virhetermit ε_i ja yhtäpitävästi mallin vastemuuttujat Y_i ovat toisistaan riippumattomia.

Näiden oletusten lisäksi mallin selittävien muuttujien välillä ei sallita täydellisiä lineaarisia yhteyksiä [6, s. 285-290]. Toisin sanoen selittävien muuttujien matriisin $\boldsymbol{X}$ sarakkeiden tulee olla lineaarisesti riippumattomia eli matriisin täysiasteinen.

3.3 Mallin parametrien SU-estimaatit

Mallin parametrit täytyy estimoida, jotta löydetään aineistoon parhaiten sopiva malli. Lähestytään ongelmaa *suurimman uskottavuuden estimoinnin* eli *SU-estimoinnin* avulla. Koska vastemuuttujat oletetaan riippumattomiksi ja vastemuuttujan jakauma normaaliseksi, yhteistiheysfunktioksi saadaan

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right). \end{aligned}$$

Uskottavuusfunktio saadaan suoraan yhteistiheysfunktiosta, kun aineisto ajatellaan kiinteänä ja mallin parametrit puolestaan muuttujina:

$$L(\boldsymbol{\beta},\sigma^2|\mathbf{y}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right).$$

Muutetaan funktio vielä logaritmiseksi uskottavuusfunktioksi tarkastelun helpottamiseksi. Lisäksi uskottavuusfunktiosta voidaan jakaa pois ne termit, jotka eivät riipu mallin parametreista eli joilla ei ole vaikutusta parametrien estimointiin. Merkitään näiden termien pois jättämistä jatkossa merkinnällä $\propto$. Logaritminen uskottavuusfunktio saa nyt muodon:

$$\begin{aligned} \log L(\boldsymbol{\beta},\sigma^2|\mathbf{y}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}RSS(\beta) \\ &\propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}RSS(\beta) \end{aligned}$$

jossa

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

on *jäännösneliösummafunktio* (engl. *the residual sum of squares*). Uskottavuusfunktio maksimoituu parametrin $\boldsymbol{\beta}$ suhteen, kun jäännösneliösummafunktio saa pienimmän arvonsa. Kyseinen piste on parametrin $\boldsymbol{\beta}$ SU-estimaatti. Suurimman uskottavuuden estimaatti löydetään derivoimalla jäännösneliösummafunktio parametrin $\boldsymbol{\beta}$ suhteen ja asettamalla derivaatta nollaksi. Suurimman uskottavuuden estimaatiksi saadaan tällöin

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i y_i.$$

Nähdään, että kyseinen estimaatti yhtyy parametrin $\boldsymbol{\beta}$ *pienimmän neliösumman estimaattiin*. Tällöin sillä on myös samat hyödylliset ominaisuudet. Estimaattori $\hat{\boldsymbol{\beta}}$ on harhaton ja tarkentuva sekä sillä on kaikista harhattomista estimaattoreista pienin varianssi [6, s. 35, 227].

Korvaamalla logaritmisesta uskottavuusfunktiosta parametrit β_j niiden SU-estimaateilla $\hat{\beta}_j$, saadaan parametrin σ^2 logaritminen profiiliuskottavuusfunktio. Derivoimalla logaritminen profiiliuskottavuusfunktio parametrin σ^2 suhteen ja asettamalla derivaatta nollaksi saadaan myös parametrille σ^2 suurimman uskottavuuden estimaatti:

$$\hat{\sigma}^2 = \frac{1}{n} RSS(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}})^2.$$

Tästä estimaattorista saadaan harhaton versio korvaamalla luku n luvulla $n - p - 1$ [8, s. 20].

4 Box–Cox-muunnos lineaarisen regressiomallin yhteydessä

4.1 Mallin parametrien estimointi

Jos lineaarisen regressiomallin oletukset eivät toteudu, voidaan harkita muunnosten tekemistä mallin vastemuuttujalle tai selittäville muuttujille. Oikeanlaisen muunnoksen tunnistaminen voi kuitenkin olla hankalaa, mihin Box–Cox-muunnos tarjoaa keinon sopivan potenssimuunnoksen löytämiseksi. Käydään seuraavaksi läpi, miten Box–Cox-muunnosta voidaan hyödyntää lineaarisen regressiomallin tapauksessa ja miten mallin parametrit tällöin estimoidaan. Tämä luku perustuu kirjaan [4, s. 389-390].

Oletetaan nyt, että lineaarisen regressiomallin vastemuuttujille Y_i tehdään kaavan (2) mukainen muunnos. Tällöin lineaarinen regressiomalli saa muodon

$$\boldsymbol{Y}^{(\lambda)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Oletetaan lisäksi, että tällöin muunnetuille satunnaismuuttujille pätee

$$Y_i^{(\lambda)} \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2).$$

Tavoitteena seuraavaksi on muodostaa parametrille λ profiiliuskottavuusfunktio ja sen avulla löytää parametrille λ SU-estimaatti. Alkuperäisten satunnaismuuttujien Y_i tiheysfunktiot voidaan ratkaista tiheysfunktioiden muunnoskaavan avulla seuraavasti:

$$f_{Y_i}(y_i) = f_{Y_i^{(\lambda)}}(y_i^{(\lambda)}) \cdot |J| = f_{Y_i^{(\lambda)}}(y_i^{(\lambda)}) \cdot \frac{dy_i^{(\lambda)}}{dy_i} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right) \cdot y_i^{\lambda-1}$$

jossa $|J|$ on Jacobin determinantti Box–Cox-muunnokselle. Satunnaisvektorin $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ yhteystiheysfunktio voidaan nyt johtaa satunnaismuuttujien Y_i tiheysfunktioiden tulona, kun satunnaismuuttujat oletetaan riippumattomiksi:

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \lambda) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^{(\lambda)} - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right) \cdot \prod_{i=1}^{n} y_i^{\lambda-1}.$$

Tämä voidaan vastaavasti muuttaa uskottavuusfunktioksi ja edelleen logaritmiseksi uskottavuusfunktioksi:

$$\begin{aligned}\log L(\boldsymbol{\beta}, \sigma^2, \lambda|\boldsymbol{y}) &= \log L(\boldsymbol{\beta}, \sigma^2, \lambda) \\ &\propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^{(\lambda)} - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \sum_{i=1}^{n}(\lambda - 1)\log(y_i).\end{aligned}$$

Profiiliuskottavuusfunktion ideana on maksimoida uskottavuusfunktio ”kiusaparametrien” suhteen jokaisella kiinnostavan parametrin λ arvolla. Tämän vuoksi muodostetaan ensin logaritmisen uskottavuusfunktion avulla parametreille $\boldsymbol{\beta}$ ja σ^2 SU-estimaatit parametrin λ funktiona. Merkitään saatuja SU-estimaatteja

$$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}^{(\lambda)} \qquad \hat{\sigma}^2_\lambda = \frac{1}{n}RSS(\hat{\boldsymbol{\beta}}_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(\lambda)} - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\lambda\right)^2$$

jossa $RSS(\hat{\boldsymbol{\beta}}_\lambda)$ on muunnoksen $y^{(\lambda)}$ jäännösneliösummafunktio. Parametrin λ logaritminen profiiliuskottavuusfunktio saadaan nyt korvaamalla edellisestä logaritmisesta uskottavuusfunktiosta parametrit β ja σ^2 näiden SU-estimaateilla $\hat{\boldsymbol{\beta}}_\lambda$ ja $\hat{\sigma}^2_\lambda$:

$$\begin{aligned}\log L_P(\lambda) &= \log L(\hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}^2_\lambda, \lambda|\boldsymbol{y}) \\ &\propto -\frac{n}{2}\log(\hat{\sigma}^2_\lambda) - \frac{1}{2\hat{\sigma}^2_\lambda}\sum_{i=1}^{n}\left(y_i^{(\lambda)} - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\lambda\right)^2 + (\lambda - 1)\sum_{i=1}^{n}\log(y_i) \\ &\propto -\frac{n}{2}\log(\hat{\sigma}^2_\lambda) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i).\end{aligned}$$

Etsimällä se parametrin λ arvo, jossa logaritminen profiiliuskottavuusfunktio saa suurimman arvonsa, löydetään parametrille λ SU-estimaatti.

On myös huomioitavaa, että Box–Cox-muunnosta voidaan hyödyntää vastemuuttujan lisäksi myös selittäville muuttujille. Jos esimerkiksi lineaarisen regressiomallin lineaarisuusoletus ei toteudu, mutta mallin virhetermit ovat normaalijakautuneita ja homoskedastisia, muunnosten tekeminen yhdelle tai useammalle selittävälle muuttujalle on suositeltavaa, koska vastemuuttujan muuntaminen voi johtaa virhetermien jakauman epänormaalisuuteen tai varianssin heteroskedastisuuteen. Jos puolestaan lineaarisuusoletus toteutuu, mutta mallin virhetermit eivät ole normaalijakautuneita ja niiden varianssi ei ole homoskedastinen, tarvitaan lähtökohtaisesti muunnosta vastemuuttujalle, jotta mallin kaikki oletukset toteutuisivat. Tapauksissa, joissa sekä normaalisuus- ja homoskedastisuusoletusten lisäksi mallin lineaarisuusoletus ei toteudu, saattaa vastemuuttujan muuntaminen auttaa linearisoimaan mallia. Lisäksi selittävien muuttujien ja vastemuuttujan yhtäaikainen muuntaminen voi auttaa mallin linearisoimisessa tai lineaarisuuden ylläpitämisessä. [6, s. 124-126]

4.2 Box–Cox-muunnoksen geometrinen versio

Box ja Cox esittelivät tutkimusartikkelissaan muunnoksesta myös muokatun version, jossa muunnetut arvot kerrotaan vastemuuttujan y geometrisella keskiarvolla, joka korotetaan potenssiin $1-\lambda$. Määritellään Box–Cox-muunnoksen geometrinen versio seuraavasti:

$$y_{i*}^{(\lambda)} = y_i^{(\lambda)} \cdot \mathrm{gm}(y)^{1-\lambda} = \begin{cases} \mathrm{gm}(y)^{1-\lambda} \cdot (y_i^\lambda - 1)/\lambda & (\lambda \neq 0), \\ \mathrm{gm}(y) \cdot \log(y) & (\lambda = 0) \end{cases} \tag{3}$$

jossa $\mathrm{gm}(y) = \left(\prod_{i=1}^n y_i\right)^{1/n}$ on vastemuuttujan y geometrinen keskiarvo. Geometrisella keskiarvolla kertominen varmistaa sen, että muunnoksen $y_{i*}^{(\lambda)}$ yksiköt ovat samat kaikilla parametrin λ arvoilla. Tästä seuraa, että jäännösneliösummien suuruusluokka ei riipu parametrin λ arvosta. Tämä tekee jäännösneliösummien vertailusta helpompaa eri parametrin λ arvoilla. Voidaan lisäksi osoittaa, että Jacobin determinantti muunnokselle $y_*^{(\lambda)}$ saa arvon yksi kaikilla muuttujan λ arvoilla [12, s. 153, s. 289]. Tällöin logaritminen uskottavuusfunktio muunnokselle $y_*^{(\lambda)}$ saadaan yksinkertaisempaan muotoon:

$$\begin{aligned} \log L(\boldsymbol{\beta}, \sigma^2, \lambda | \boldsymbol{y}) &= \log L(\boldsymbol{\beta}, \sigma^2, \lambda) \\ &\propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_{i*}^{(\lambda)} - \boldsymbol{x}_i'\boldsymbol{\beta})^2. \end{aligned}$$

Tämän avulla voidaan jälleen muodostaa kiinnostuksen kohteena olevalle parametrille λ profiiliuskottavuusfunktio. Logaritminen uskottavuusfunktio maksimoidaan ensin parametrien β ja σ^2 suhteen jokaisella parametrin λ arvolla, mikä vastaa parametrien β ja σ^2 SU-estimaatteja, kun λ olisi jokin tunnettu vakio. Saadut SU-estimaatit ovat täten

$$\hat{\boldsymbol{\beta}}_{\lambda*} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_*^{(\lambda)} \qquad \hat{\sigma}^2_{\lambda*} = \frac{1}{n}RSS(\hat{\boldsymbol{\beta}}_{\lambda*}) = \frac{1}{n}\sum_{i=1}^n \left(y_{i*}^{(\lambda)} - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\lambda*}\right)^2$$

missä $\mathrm{RSS}(\hat{\boldsymbol{\beta}}_{\lambda*})$ on geometrisen Box–Cox-muunnoksen jäännösneliösummafunktio. Kun kyseiset SU-estimaatit sijoitetaan edelliseen logaritmiseen uskottavuusfunktioon, parametrin λ logaritmiseksi profiiliuskottavuusfunktioksi saadaan

$$\begin{aligned} \log L_P(\lambda) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2_{\lambda*}) - \frac{1}{2\hat{\sigma}^2_{\lambda*}}\sum_{i=1}^n \left(y_{i*}^{(\lambda)} - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\lambda*}\right)^2 \\ &\propto -\frac{n}{2}\log(\hat{\sigma}^2_{\lambda*}). \end{aligned}$$

Etsimällä tämän funktion maksimikohta löydetään edelleen parametrille λ SU-estimaatti $\hat{\lambda}$.

4.3 Luottamusvälin muodostaminen parametrille λ

Muodostetaan seuraavaksi parametrille λ approksimatiivinen luottamusväli luottamustasolla $1-\alpha$ perustuen kirjaan [4, s. 126] sekä tilastollisen päättelyn kurssien

luentomonisteeseen [7]. Luottamusvälin muodostaminen perustuu uskottavuusosamäärään otossuureeseen

$$W(\lambda) = -2r(\lambda; \boldsymbol{y}) = -2[\log L_P(\lambda) - \log L_P(\hat{\lambda})]$$

jossa $r(\lambda; \boldsymbol{y})$ on normitettu logaritminen profiiliuskottavuusfunktio parametrille λ. Jos tilastollinen malli toteuttaa oletukset SU-estimaattorin asymptoottisesta normaaliudesta, pätee

$$W(\lambda) \xrightarrow{D} \chi^2_1$$

kun $n \to \infty$. Tämä tarkoittaa, että uskottavuusosamäärän otossuure suppenee jakaumaltaan kohti khiin neliön jakaumaa yhdellä vapausasteella, kun n kasvaa rajatta. Käytännössä suurilla n arvoilla $W(\lambda)$ likimain noudattaa χ^2_1-jakaumaa. Tämän perusteella parametrille λ voidaan muodostaa nyt approksimatiivinen luottamusväli luottamustasolla $1 - \alpha$:

$$\log L_P(\hat{\lambda}) - \log L_P(\lambda) \leq \frac{1}{2}\chi^2_1(\alpha)$$

jossa $\chi^2_1(\alpha)$ on χ^2_1-jakauman α-yläkvantiili eli piste, jonka oikealle puolella on osuus α todennäköisyysmassasta. Tämä tarkoittaa, että parametrin λ arvoja, joille $W(\lambda) \leq \chi^2_1(\alpha)$ voidaan pitää uskottavina luottamustasolla $1 - \alpha$. Jos luottamustasoksi on valittu esimerkiksi 95 % eli $\alpha = 0.05$, niin $\chi^2_1(0.05) \approx 3.84$.

Parametrin λ arvo valitaan usein luottamusvälin sisältä SU-estimaatin arvon läheltä, koska nämä arvot tekevät mallista usein helpommin tulkittavan kuin tietty SU-estimaatin arvo. Tällöin SU-estimaatin arvon löytäminen ei ole välttämättä tarpeellista. Esimerkiksi arvon $\lambda = 0$ käyttäminen SU-estimaatin $\hat{\lambda} = 0.13$ sijasta voi olla helpommin ymmärrettävissä ilman, että muunnoksen tehokkuus kärsii. Jos parametrin λ SU-estimaatti on puolestaan lähellä lukua 1, muunnoksen tekeminen vastemuuttujalle voi olla kokonaan tarpeetonta. Sopivan parametrin λ arvon löytämiseksi kannattaa kuitenkin tutkia parametrin λ uskottavuusfunktion käyttäytymistä. [6, s. 133-134]

4.4 Esimerkkejä muunnoksen soveltamisesta aineistoon

Käydään seuraavaksi läpi kaksi esimerkkiä, joissa aineistoon sovitetaan lineaarinen regressiomalli ja Box–Cox-muunnosta hyödynnetään virhetermien heteroskedastisuuden vähentämiseen sekä normaalisuusoletuksen parantamiseen. Ensimmäisen esimerkin aineisto on yhdistetty Tilastokeskuksen tuottamasta aineistosta ”Asuntokuntien tulot ja tulojen rakenne alueittain, 1995-2024” [9] sekä Maanmittauslaitoksen tuottamasta aineistosta ”Asuinpientalokiinteistöt haja-asutusalueella / rakentamattomat kohteet” [5]. Tilastokeskuksen aineisto sisältää tiedot Suomen kuntien asuntoväestön määrästä sekä keskimääräisistä bruttotuloista asuntokunnittain vuonna 2024. Maanmittauslaitoksen aineisto sisältää puolestaan tiedot Suomen kuntien haja-asutusalueiden rakentamattomien pientalokiinteistöjen kiinteistökauppojen keskimääräisistä neliöhinnoista sekä kiinteistöjen keskimääräisistä neliöiden lukumääristä vuonna 2024. Yhdistetyn aineiston havaintoyksikköinä ovat ne Suomen

kunnat, joissa haja-asutusalueiden rakentamattomien pientalokiinteistöjen kiinteistökauppojen lukumäärä oli vähintään kolme vuonna 2024. Näitä kuntia on yhteensä 175 kappaletta.

Sovitetaan ensin yhdistettyyn aineistoon tavallinen lineaarinen regressiomalli, jonka vastemuuttujana on kiinteistökauppojen keskimääräinen neliöhinta ($€/m^2$) ja selittävinä muuttujina ovat kiinteistöjen keskimäärinen pinta-ala neliömetreinä, asuntoväestön lukumäärä sekä keskimääriset bruttotulot asuntokunnittain. Lineaarinen regressiomalli saa tällöin muodon

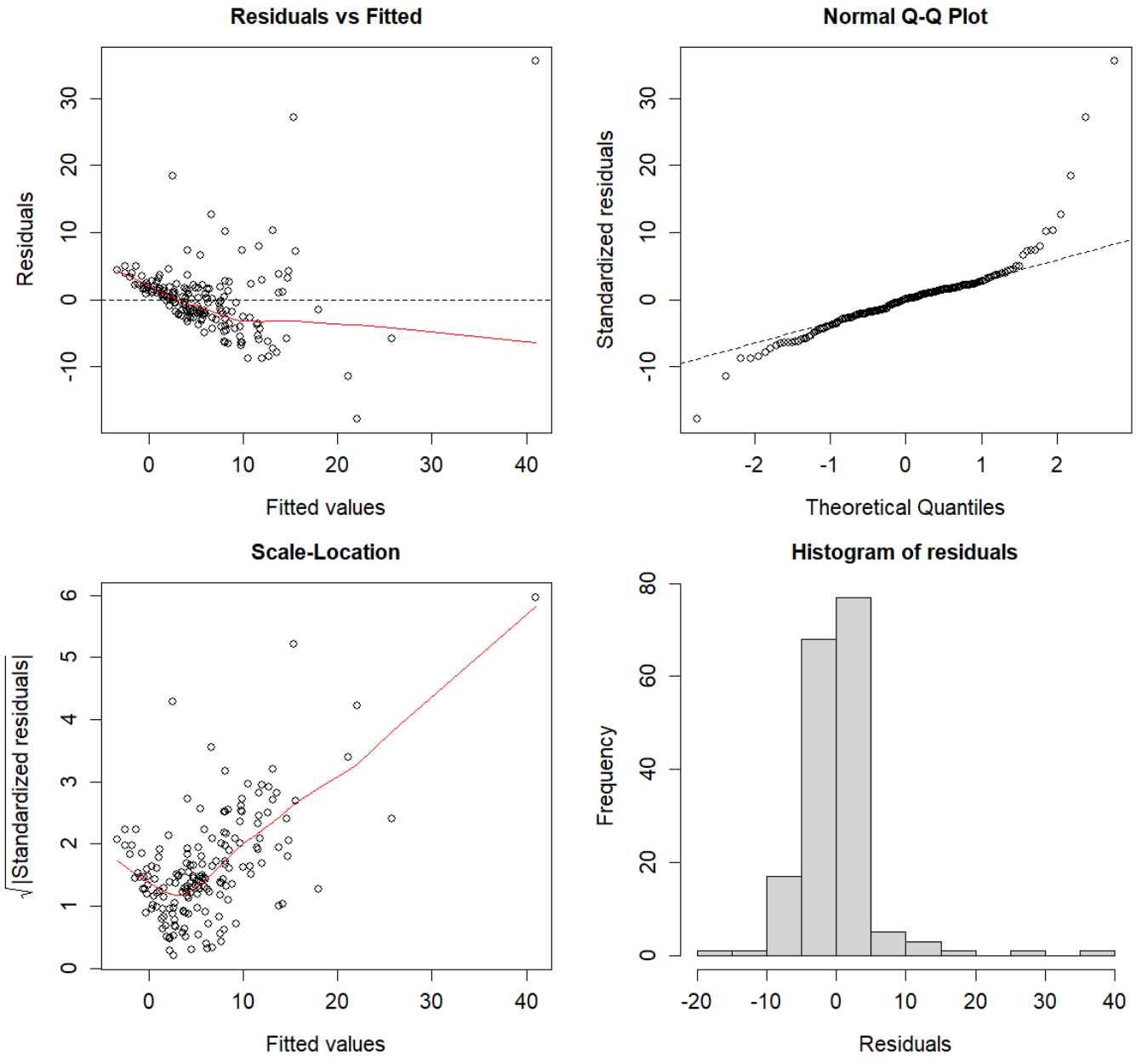
$$\texttt{hintakeskiarvo} = \beta_0 + \beta_1 \texttt{pinta-ala\_ka} + \beta_2 \texttt{asuntovaesto} + \beta_3 \texttt{bruttotulot} + \varepsilon$$

ja lisäksi R:n funktioiden `lm()` ja `summary()` avulla saadaan mallin parametreille seuraavat SU-estimaatit sekä testitulokset:

```
Call:
lm(formula = hintakeskiarvo ~ pinta_ala_ka + asuntovaesto +
    bruttotulot, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-17.797  -2.405   0.085   1.786  35.615

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.707e+01  3.081e+00  -5.542 1.11e-07 ***
pinta_ala_ka -4.830e-04  1.969e-04  -2.453   0.0152 *
asuntovaesto  8.420e-05  9.261e-06   9.092 2.40e-16 ***
bruttotulot   3.918e-04  4.526e-05   8.655 3.52e-15 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.42 on 171 degrees of freedom
Multiple R-squared:  0.5155,    Adjusted R-squared:  0.507
F-statistic: 60.66 on 3 and 171 DF,  p-value: < 2.2e-16
```

Tuloksista nähdään, että mallin korjattu selitysaste saa arvon 0.507. Lisäksi kaikki selittävät muuttujat ovat yksittäisten muuttujien t-testien perusteella tilastollisesti merkitseviä luottamustasolla 0.05 ja F-yleistestin p-arvo on merkitsevä, mikä viittaa siihen, että kokonaisuutena selittävillä muuttujilla on tilastollisesti merkittävä yhteys mallin vastemuuttujaan.

Kuvan 1 perusteella kuitenkin huomataan, että lineaarisen regressiomallin oletukset eivät kunnolla toteudu, jolloin mallin tuloksia ei voida pitää täysin luotettavina. Mallin residuaalien varianssi kasvaa voimakkaasti sovitteen arvon kasvaessa. Lisäksi residuaalien jakauma on voimakkaasti oikealle vino.



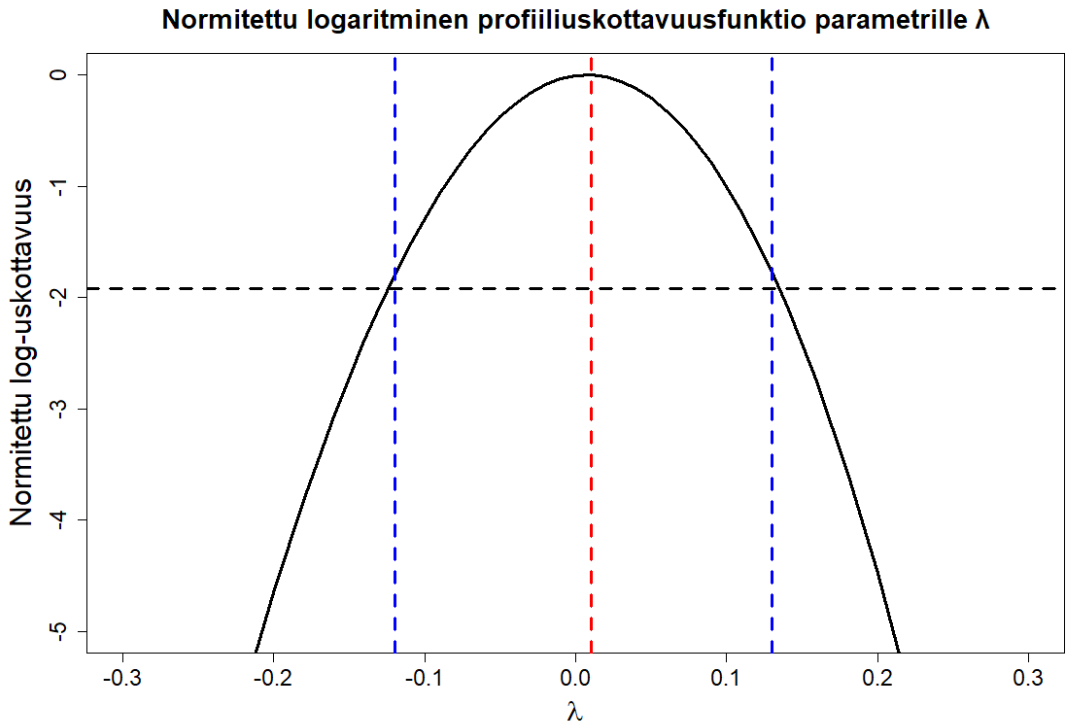
Kuva 1: Alkuperäisen mallin diagnostiikkapiirrokset.

Oletetaan seuraavaksi, että mallin vastemuuttujalle tehdään kaavan (2) mukainen Box–Cox-muunnos ja aineistoon sovitetaan uusi lineaarinen regressiomalli:

$$\texttt{hintakeskiarvo}^{(\lambda)} = \beta_0 + \beta_1 \texttt{pinta-ala\_ka} + \beta_2 \texttt{asuntovaesto} + \beta_3 \texttt{bruttotulot} + \varepsilon.$$

Parametrin λ estimointiin voidaan hyödyntää R:n `MASS`-paketista löytyvää funktiota `boxcox()`, joka laskee ja piirtää logaritmisen profiiliuskottavuusfunktion sekä esittää approksimatiivisen 95 % luottamusvälin Box–Cox-muunnoksen parametrille λ:

```
bc <- boxcox(
    lm(hintakeskiarvo ~ pinta_ala_ka + asuntovaesto +
    bruttotulot, data = data),
    lambda = seq(-4, 4, 0.01),
    plotit = FALSE).
```



Kuva 2: Parametrin λ normitettu logaritminen profiiliuskottavuusfunktio, SU-estimaatti sekä 95 % luottamusväli.

Käytännössä `boxcox()` laskee valitulta väliltä useille parametrin λ arvoille logaritmisen profiiliuskottavuusfunktion arvon. Näiden pisteiden perusteella muodostetaan logaritmisen profiiliuskottavuusfunktion kuvaaja, jonka sovittamisessa hyödynnetään oletusarvoisesti splini-interpolointia. Parametrin λ SU-estimaattia etsitään

puolestaan oletusarvoisesti väliltä $[-2, 2]$ frekvenssillä 0.1, mikä tarkoittaa, että uskottavuus lasketaan yhteensä 41 eri parametrin λ arvolle. Näistä arvoista se, jonka profiiliuskottavuus on suurin, on parametrin λ SU-estimaatti. [11, s. 171]

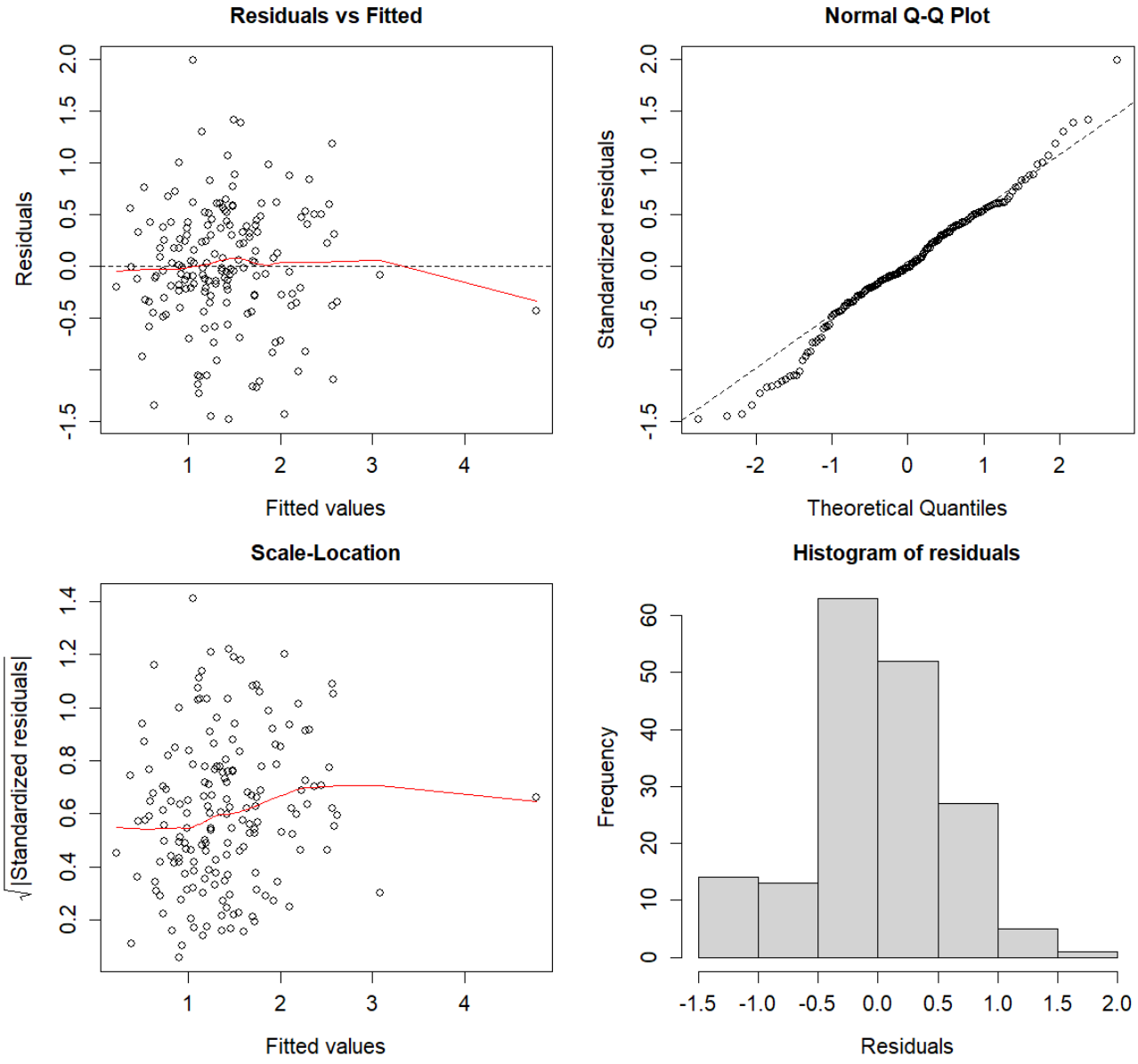
Tämän luvun esimerkeissä parametrin λ SU-estimaattia etsitään väliltä $[-4, 4]$ frekvenssillä 0.01. Parametrin λ SU-estimaatiksi saadaan täten $\hat{\lambda} = 0.01$ ja 95 % luottamusväliksi $[-0.12, 0.13]$. Valitaan parametrin λ arvoksi $\lambda = 0$, koska se on tulkinnallisesti helpompi kuin saatu SU-estimaatin arvo ja $\lambda = 0$ sisältyy SU-estimaatin 95 % luottamusväliin. Vastemuuttujalle tehtävä muunnos vastaa tällöin siis logaritmimuunnosta.

Kuvan 3 perusteella uuden mallin residuaalien varianssi ei enää kasva yhtä voimakkaasti sovitteen arvon kasvaessa sekä residuaalien jakauma muistuttaa enemmän normaalijakaumaa. Lineaarisen regressiomallin oletukset toteutuvat nyt siis huomattavasti paremmin. R:n funktioiden `lm()` ja `summary()` avulla saadaan jälleen uuden mallin parametreille seuraavat SU-estimaatit sekä testitulokset:

```
Call:
lm(formula = Hintakeskiarvo_bc ~ pinta_ala_ka + asuntovaesto +
    bruttotulot, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48304 -0.30208 -0.02021  0.39244  1.99159

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.398e+00  3.390e-01  -4.124 5.80e-05 ***
pinta_ala_ka -6.985e-05  2.167e-05  -3.223  0.00152 **
asuntovaesto  7.007e-06  1.019e-06   6.876 1.11e-10 ***
bruttotulot   4.978e-05  4.981e-06   9.995  < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5964 on 171 degrees of freedom
Multiple R-squared:  0.5057,    Adjusted R-squared:  0.497
F-statistic: 58.31 on 3 and 171 DF,  p-value: < 2.2e-16
```

Tuloksista huomataan, että uuden mallin korjattu selitysaste saa arvon 0.497 eli mallin selitysvoima on säilynyt lähes ennallaan. Lisäksi yksittäisten t-testien perusteella kaikki selittävät muuttujat ovat edelleen tilastollisesti merkitseviä luottamustasolla 0.01 ja F-yleistestin perusteella kokonaisuutena selittäjillä on tilastollisesti merkitsevä yhteys mallin vastemuuttujaan. Mallin selittäviä muuttujia `pinta-ala_ka`, `asuntovaesto` ja `bruttotulot` vastaavat parametrien estimaatit ovat $\hat{\beta}_1 = -6.985 \cdot 10^{-5}$, $\hat{\beta}_2 = 7.007 \cdot 10^{-6}$ ja $\hat{\beta}_3 = 4.978 \cdot 10^{-5}$. Tämä tarkoittaa, että yhden yksikön lisäys selittävään muuttujaan muuttaa vastemuuttujaa kertoimella $e^{\hat{\beta}_j}$ olettaen, että mallin muut selittäjät pysyvät vakioina. Esimerkiksi yhden neliömetrin lisäys

kiinteistön pinta-alaan pienentää kiinteistökaupan hinnan keskiarvoa noin 0.007%. Muuttujien `asuntovaesto` ja `bruttotulot` vaikutus vastemuuttujaan on puolestaan positiivinen.



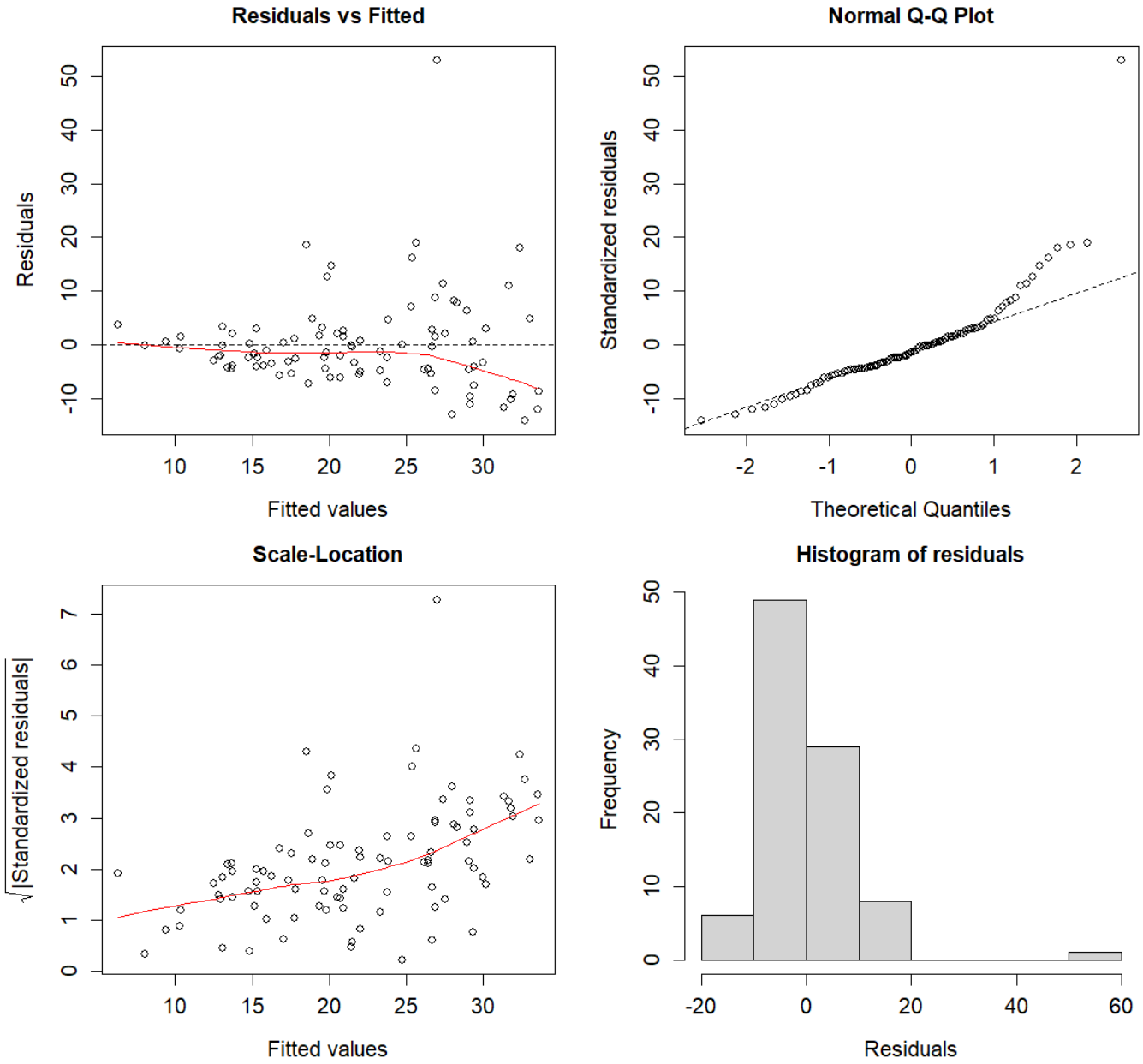
Kuva 3: Uuden mallin diagnostiikkapiirrokset.

Havainnollistaan seuraavaksi vielä Box–Cox-muunnoksen toimintaa R:n `MASS`-paketista löytyvän aineiston `Cars93` avulla, joka koostuu erilaisista autojen hintoihin ja ominaisuuksiin liittyvistä muuttujista. Sovitetaan aineistoon seuraava lineaarinen

regressiomalli:

$$\texttt{Max.Price} = \beta_0 + \beta_1 \texttt{Weight} + \varepsilon$$

jossa `Max.Price` on auton korkein arvioitu myyntihinta kerrotuna tuhannella dollarilla ja `Weight` on auton paino paunoina. Kuvasta 4 nähdään, että residuaalien varianssi kasvaa sovitteen arvojen kasvaessa ja lisäksi residuaalien jakauma on oikealle vino.

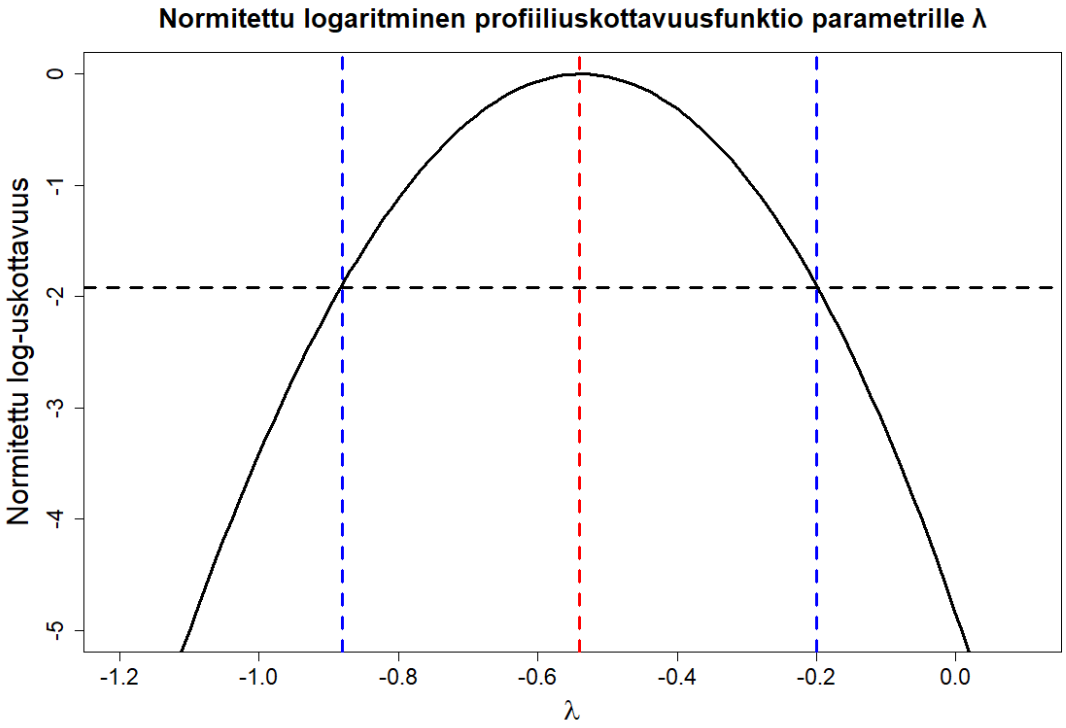


Kuva 4: Alkuperäisen mallin diagnostiikkapiirrokset.

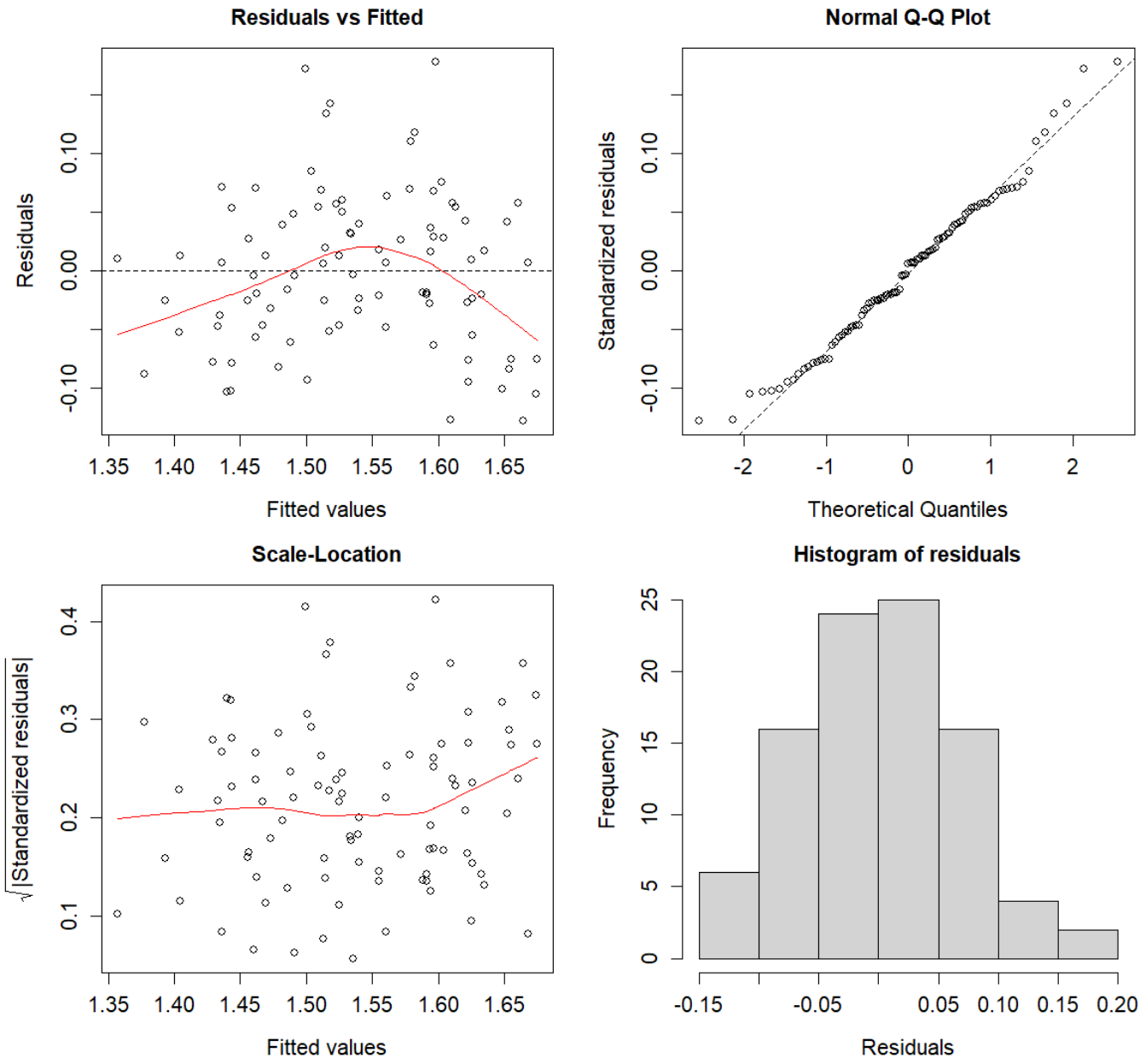
Oletetaan, että mallin vastemuuttujalle `Max.Price` tehdään kaavan 2 mukainen muunnos, jolloin lineaarinen regressiomalli saa muodon

$$\texttt{Max.Price}^{(\lambda)} = \beta_0 + \beta_1 \texttt{Weight} + \varepsilon.$$

Estimoidaan parametrin λ arvo jälleen SU-menetelmällä R:n `boxcox`-funktion avulla. Parametrin λ SU-estimaatiksi saadaan $\hat{\lambda} = -0.54$ ja 95 % luottamusväliksi $[-0.88, -0.2]$. Valitaan parametrin λ arvoksi selkeyden vuoksi nyt $\lambda = -0.5$. Uuden mallin diagnostiikkapiirroksista nähdään, että residuaalien varianssi ei kasva enää yhtä voimakkasti sovitteen arvojen kasvaessa ja lisäksi residuaalien jakauma muistuttaa normaalijakaumaa. Box–Cox-muunnoksen avulla onnistuttiin siis molemmissa esimerkeissä löytämään potenssimuunnosten perheestä vastemuuttujalle sopiva muunnos, jonka avulla saatiin mallia parannettua.



Kuva 5: Parametrin λ normitettu logaritminen profiiliuskottavuusfunktio, SU-estimaatti sekä 95 % luottamusväli.



Kuva 6: Uuden mallin diagnostiikkapiirrokset.

5 Box–Cox-muunnoksen laajennuksia ja sovelluskohteita

Box–Cox-muunnos määriteltiin edellisissä luvuissa ainoastaan positiivisille vastemuuttujan arvoille. Tämä kuitenkin rajaa muunnoksen käyttömahdollisuuksia, kos-

ka todellisuudessa vastemuuttujan arvot voivat usein olla myös negatiivisia. Tämän vuoksi Box–Cox-muunnoksen pohjalta on kehitetty lukuisia erilaisia laajennuksia, jotka sallivat esimerkiksi negatiiviset vastemuuttujan arvot tai mahdollistavat usean muuttujan muuntamisen samanaikaisesti. Käydään seuraavaksi läpi näistä laajennuksista muutamia.

5.1 Muunnokset negatiivisille vastemuuttujan arvoille

Alkuperäisessä Boxin ja Coxin tutkimusartikkelissa [2] määriteltiin myös negatiivisille vastemuuttujan arvoille sopiva muunnos seuraavasti:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & (\lambda_1 \neq 0) \\ \log(y_i + \lambda_2) & (\lambda_1 = 0) \end{cases}$$

jossa $y_i > -\lambda_2$.

Vastaavasti tämän normalisoitu versio on

$$z^{(\lambda)} = \frac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1 \cdot (\mathrm{gm}(y_i + \lambda_2))^{\lambda_1 - 1}}$$

missä $\mathrm{gm}(y_i + \lambda_2)$ on muuttujan $y_i + \lambda_2$ arvojen geometrinen keskiarvo.

Toinen tapa ottaa huomioon negatiiviset vastemuuttujan arvot on Yeon ja Johnsonin [1] (2000) kehittelemä versio Box–Cox-muunnoksesta:

$$y_{i,\mathrm{YJ}}^{(\lambda)} = \begin{cases} \dfrac{(y_i + 1)^{\lambda} - 1}{\lambda} & (y_i \geq 0, \quad \lambda \neq 0), \\ \log(y_i + 1) & (y_i \geq 0, \quad \lambda = 0), \\ -\dfrac{(-y_i + 1)^{2-\lambda} - 1}{2 - \lambda} & (y_i < 0, \quad \lambda \neq 2), \\ -\log(-y_i + 1) & (y_i < 0, \quad \lambda = 2). \end{cases}$$

Kun vastemuuttujan arvo y_i on positiivinen, *Yeo–Johnson-muunnos* vastaa aiemmin luvussa 2 esitettyä tavallista Box–Cox-muunnosta (2), jossa jokaiseen vastemuuttujan arvoon lisätään luku 1. Vastaavasti kun vastemuuttujan y_i arvo on negatiivinen, Box–Cox-muunnoksen tavalliseen versioon sijoitetaan vastemuuttujan paikalle arvo $-y_i + 1$ ja parametrin λ paikalle $2 - \lambda$. Verrattuna alkuperäiseen Box–Cox-muunnokseen Yeo–Johnson-muunnos käyttäytyy eri tavalla etenkin positiivisille lähellä nollaa oleville vastemuuttujan arvoille. Box–Cox-muunnoksessa tällaiset arvot muuttuvat enemmän kun taas Yeo–Johnson-muunnos muuntaa tällaisia arvoja maltillisemmin [12, s. 160-161]. Yeo–Johnson-muunnoksesta on vastaavasti olemassa myös geometrinen versio. Lisäksi muunnoksen pohjalta on kehitetty robusti versio, joka mahdollistaa eri parametrin λ arvot positiivisille ja negatiivisille vastemuuttujan arvoille [1].

5.2 Usean muuttujan samanaikainen muuntaminen

Joissakin tilanteissa voidaan hyötyä vastemuuttujan sekä yhden tai useamman selittävän muuttujan samanaikaisesta muuntamisesta. Esimerkiksi, jos vastemuuttujan sekä selittävän muuttujan jakaumat ovat vinoja, voidaan molempien muuttujien muuntamista kokeilla mallin linearisoimiseksi. Velilla [10] (1993) esitteli Box–Cox-muunnoksesta usean muuttujan version (engl. *Multivariate generalization of the Box–Cox transformations method*), jossa Box–Cox-muunnoksen geometrista versiota (3) sovelletaan kahdelle tai useammalle muuttujalle multinormaalijakauman saavuttamiseksi. Multinormaalijakauman käsitettä ja käyttötarkoituksia tarkastellaan esimerkiksi kirjassa [6, s. 645-650].

Käydään seuraavaksi lyhyesti läpi, mitä useamman muuttujan samanaikaisella muuntamisella käytännössä tarkoitetaan. Olkoon $\boldsymbol{y} = (y_1, \ldots, y_p)$ vektori muuttujista, joiden arvoja halutaan muuntaa. Oletetaan nyt, että jokaiselle muuttujalle y_j tehdään kaavan (3) mukainen geometrinen Box-Cox-muunnos ja merkitään näin saatuja muuttujia merkinnällä $\boldsymbol{\psi}_M(\boldsymbol{y}, \boldsymbol{\lambda}) = (\psi_M(y_1, \lambda_1), \ldots, \psi_M(y_p, \lambda_p))$. Oletetaan lisäksi, että muunnoksen jälkeen $\boldsymbol{\psi}_M(\boldsymbol{y}, \boldsymbol{\lambda})$ noudattaa multinormaalijakaumaa:

$$\boldsymbol{\psi}_M(\boldsymbol{y}, \boldsymbol{\lambda}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

jossa $\boldsymbol{\mu}$ on muuttujien odotusarvoista koostuva vektori ja $\boldsymbol{\Sigma}$ on jokin tuntematon positiivisesti definiitti symmetrinen matriisi, jonka arvoja yritetään estimoida. Multinormaalijakauman yhteistiheysfunktion avulla voidaan esimerkiksi muodostaa parametrivektorille $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ logaritminen profiiliuskottavuusfunktio, joka voidaan maksimoida parametrivektorin $\boldsymbol{\lambda}$ suhteen. Estimoinnin tarkempia yksityiskohtia käsitellään kirjassa [12, s. 290-291] ja hyvä esimerkki monimuuttujaversion hyödyntämisestä mallin linearisoimiseksi sekä parametrivektorin $\boldsymbol{\lambda}$ estimoimisesta R:n avulla löytyy esimerkiksi kirjasta [8, s. 95-98]. Lisäksi on huomioitavaa, että menetelmän yhteydessä voidaan hyödyntää Box–Cox-muunnoksen sijaan esimerkiksi Yeo–Johnson-muunnosta, mikä laajentaa menetelmän käyttömahdollisuuksia [12, s. 290].

Viitteet

[1] Atkinson, A. C., Riani, M. ja Corbellini, A. (2020) *The Analysis of Transformations for Profit-and-Loss Data*. Journal of the Royal Statistical Society Series C: Applied Statistics, 69(2), 251–275. https://doi.org/10.1111/rssc.12389

[2] Box, G. E. P. ja Cox, D. R. (1964) *An Analysis of Transformations. Journal of the Royal Statistical Society*. Series B (Methodological), 26(2), 211–252. http://www.jstor.org/stable/2984418

[3] Cook, R. D. ja Weisberg, S. (1999) *Applied Regression Including Computing and Graphics*, Wiley.

[4] Davison, A. C. (2003) *Statistical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815850

[5] Maanmittauslaitos. Tilastotietoa kiinteistökaupoista. *Asuinpientalokiinteistöt: Asuinpientalokiinteistöt haja-asutusalueella: rakentamattomat kohteet*. (2024) https://khr.maanmittauslaitos.fi/tilastopalvelu/rest/v2026.2/index.html?v=2025.2.9# Viitattu 20.4.2026.

[6] Neter, J., Kutner, M., Nachtsheim, C. ja Wasserman, W. (1996) *Applied Linear Statistical Models*. Irwin.

[7] Nieminen, P. ja Nyberg, H. (2024) *Tilastollinen Päättely I & II*-luentomoniste. Turun Yliopisto.

[8] Sheather, S. J. (2009) *A Modern Approach to Regression with R*. Springer.

[9] Tilastokeskus. Tulonjakotilasto. *Asuntokuntien tulot ja tulojen rakenne alueittain, 1995-2024*. https://statfin.stat.fi/PxWeb/pxweb/fi/StatFin/StatFin__tjt/statfin_tjt_pxt_118w.px/ Viitattu 20.4.2026.

[10] Velilla, S. (1993) *A note on the multivariate Box-Cox transformation to normality*. Statistics & Probability Letters, 17(4), 259-263. https://doi.org/10.1016/0167-7152(93)90200-3

[11] Venables, W.N. ja Ripley, B.D. (2002) *Modern Applied Statistics with S* (4. laitos). Springer.

[12] Weisberg, S. (2005) *Applied Linear Regression* (3. laitos). Wiley-interscience.

[13] Yeo, I.-K. ja Johnson, R. A. (2000) *A New Family of Power Transformations to Improve Normality or Symmetry*. Biometrika, 87(4), 954–959. http://www.jstor.org/stable/2673623