Tampereen yliopisto

Roni Purolainen

BERKSONIN MITTAVIRHEMALLI JA SEN SOVELLUS SIMULOITUUN AINEISTOON

Informaatioteknologian ja viestinnän tiedekunta
Kandidaatintutkielma
Toukokuu 2022

Tiivistelmä

Roni Purolainen: Berksonin mittavirhemalli ja sen sovellus simuloituun aineistoon
Kandidaatintutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin kandidaattiohjelma
Toukokuu 2022

Tämän työn keskeisin menetelmä on Berksonin mittavirhemalli. Työ tarkastelee regressioanalyysin teoriaa, johon Berksonin mittavirhemalli perustuu, ja niin ikään Berksonin mallin teoriaa. Siinä käydään läpi myös yleisemmin mittavirhemallien teoriaa.

Lineaarinen malli on regressioanalyysissä käytetty malli, jossa selitettävää muuttujaa mallinnetaan yhdellä tai useammalla selittävällä muuttujalla. Yhden selittäjän malli on erikokistapaus lineaarisesta mallista. Mallin muodostamiseksi pitää laskea kaksi parametriarvoa β_0 ja β_1. Parametrien ja selittävän muuttujan avulla voidaan muodostaa malli, joka kuvaa selitettävää muuttujaa.

Mittavirhemallit voivat pohjautua lineaariseen malliin. Siinä tarkastellaankin sellaista selittävää muuttujaa, jossa on mukana mittavirhettä. Tällaisissa tapauksissa saadaan mittavirhemallilla parempia tuloksia, kuin tavallisella lineaarisella mallilla. Usein mittavirhemallin muodostamiseen vaaditaan oletus, jossa mittavirheen varianssin oletetaan tunnetuksi.

Berksonin mittavirhemalli on erikoistapaus mittavirhemalleista. Tässä mallissa ei vaadita oletusta tunnetusta mittavirheen varianssista, mutta mittavirhemallia muodostaessa täytyy selittävä muuttuja olettaa vakioksi. Nyt mallille voidaan laskea parametriestimaatit pienimmän neliösumman menetelmällä. Tässä työssä on muodostettu Berksonin mittavirhemalli simuloidulle aineistolle sekä vertailtu sitä malliin, jossa mittavirhettä ei ole.

Sisällys

1 Johdanto

Regressioanalyysin tarkoituksena on mallintaa tilastollisesti vastemuuttujaa (Freund, Wilson & Sa 2006). Malleista yksi yleisimmistä lienee lineaarinen malli. Regressioanalyysissä tarkoituksena on mallintaa vastemuuttujaa, joko yhtä tai useampaa selittävää muuttujaa käyttäen. Mallin yhteensopivuutta voidaan usein mitata esimerkiksi laskemalla R^2 arvo, joka kertoo kuinka monta prosenttia malli selittää vastemuuttujan satunnaisvaihtelusta. R^2 arvoa kutsutaan usein mallin selitysasteeksi tai selityskertoimeksi.

Tietyissä tapauksissa R^2 arvo saattaa olla niin huono, että mallin ennustamiskyky jää liian alhaiseksi. Tällaisissa tapauksissa voidaan turvautua erilaisiin keinoihin, joilla arvoa voidaan saada paremmaksi. Esimerkiksi muuttujan arvoja mitattaessa saattaa niihin tulla mittavirhettä. Tämä vaihtelu johtuu epäluotettavasta mittausmenetelmästä, ihmisen aiheuttamasta virheestä tai jokaisen mittauksen sisältämästä luonnollisesta epätarkkuudesta.

Jossain määrin kaikki tilastolliset analyysit sisältävät mittavirheitä. Tällaisten virheiden esiintymiseen on monia syitä, yleisimmät niistä ovat laitevirhe ja näytteenottovirhe (Buonaccorsi 2010). Buonaccorsi (2010) antaa myös esimerkkejä tilanteista, joissa esiintyy mittavirhettä, kuten: Harvard Six Cities -tutkimuksessa tarkasteltiin tietyille saasteille altistumisen vaikutusta lapsen hengitystilaan. Yksilön todellista altistumista oli vaikea mitata. Sen sijaan havainnot olivat valmistettu kotona, eri huoneissa eri vuodenaikoina. Nämä arvot toimivat altistumisen korvikkeina. Toimenpiteet validoitiin tutkimuksissa, joissa henkilöt käyttivät käännemonitoria, joka mittaa altistumista jatkuvasti.

Avuksi tällaisen datan analysointiin on kehitetty mittavirhemalleja. Stefanski (2000) kirjoittaa, että monilla sovellusalueilla tilastollisesti merkitykselliset mallit määritellään muuttujilla X, jotka eivät jostain syystä ole suoraan mitattavissa. Tällaisissa tapauksissa on tavallista havaita sen sijaan korvaavaa muuttujaa W. Tällaisia ongelmia usein kutsutaan mittavirheongelmiksi ja tilastollisia malleja ja menetelmiä kyseisen datan analysointiin kutsutaan mittavirhemalleiksi.

Mittavirhemalleilla saadaan muodostettua paremmin istuva malli muuttujalle, joka sisältää huomattavan määrän mittausvirhettä. Työni keskittyy pääasiassa mittavirhemalleihin ja kuinka mittavirhemalleja voidaan muodostaa ja millaisissa tilanteissa. Mittavirhemallit muodostetaan regressiomallin pohjalta. Asiaa havainnollistetaan si-

muloidulle aineistolle tehdylle Berksonin mittavirhemallin avulla.

2 Lineaarinen malli

Lineaarisella mallilla pyritään mallintamaan selitettävää muuttujaa selittävillä muuttujilla. Chatterjee & Hadi (2015) määrittelevät regressiomallin ja lineaarisen mallin seuraavasti: Merkitään selitettävä muuttuja Y ja selittävät muuttujat $X_1, X_2, ..., X_p$. Muuttujien yhteyttä voidaan approksimoida regressiomallin

$$Y = f(X_1, X_2, ..., X_p) + \varepsilon \tag{2.1}$$

avulla, missä ε oletetaan olevan satunnainen virhe, joka kuvaa approksimoinnin epätarkkuutta. Funktio $f(X_1, X_2, ..., X_p)$ kuvaa suhdetta Y:n ja $X_1, X_2, ..., X_p$:n välillä. Esimerkkinä on lineaarinen malli

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2, ..., +\beta_p X_p + \varepsilon, \tag{2.2}$$

missä vakiot $\beta_0, \beta_1, \beta_2, ..., \beta_p$, ovat aineiston pohjalta estimoitavia parametrejä.

2.1 Yhden selittävän muuttujan lineaarinen malli

Yhden selittävän muuttujan lineaarinen malli on erikoistapaus lineaarisesta mallista, jossa on vain yksi selittävä muuttuja. Malli on tällöin muotoa

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{2.3}$$

Parametreille β_0 ja β_1 voidaan estimoida arvot pienimmän neliösumman menetelmän avulla. Estimaattoreiksi saadaan

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X} \tag{2.4}$$

ja

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{2.5}$$

missä $\hat{\beta_0}$ ja $\hat{\beta_1}$ ovat estimaatteja parametreille β_0 ja β_1. Kun käytetään estimoituja parametriarvoja lineaarisen mallin muodostamiseen, saadaan

$$\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} X_i, \tag{2.6}$$

$\hat{Y_i}$ on estimaatti arvolle Y_i. Funktio $\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} X_i$ mallintaa siis Y_i:n käyttäytymistä eri arvoilla X_i.

3 Mittavirhemalli

Tarkastellaan lineaarista mallia, joka on muotoa

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \qquad t = 1, 2, \ldots, n, \tag{3.1}$$

jossa $(x_1, x_2, \ldots, x_n)$ ovat vakioita ja satunnaisvirheet ε_t ovat riippumattomia ja $(0, \sigma^2_{\varepsilon\varepsilon})$ normaalisti jakautuneita, jossa $\sigma^2_{\varepsilon\varepsilon}$ on ε:n varianssi.

Tutkitaan regressiomalleja, joissa x_t ei ole suoraan havaittavissa. Sen sijaan, että tarkkailtaisiin x_t:tä, voidaan tarkastella summaa

$$X_t = x_t + u_t, \tag{3.2}$$

jossa u_t on $N(0, \sigma^2_{uu})$ satunnaismuuttuja. Havaittua muuttujan arvoa X_t on usein kutsuttu $ilmaisinmuuttujaksi$ tai $osoittajamuuttujaksi$.(Fuller 1987)

Fuller (1987) antaa kirjassaan esimerkin tilanteesta, jossa x_t ei ole suoraan havaittavissa. Tutkitaan suhdetta maissisadon ja maaperän typpipitoisuuden välillä. Oletetaan, että (3.1) on sopiva approksimaatio sadon ja typen suhteesta. Kerroin β_1 tarkoittaa sadon kasvua, kun maaperän typpipitoisuus kasvaa yhden yksikön. Jotta maaperän typpipitoisuus saadaan arvioitua, pitää suorittaa laboratorioanalyysejä saaduille näytteille. Näytteiden ja laboratorioanalyysien vuoksi x_t:tä ei voida havainnoida suoraan, vaan havainnoidaan approksimaatiota x_t:stä. Niinpä voidaan esittää havaittu typpipitoisuus X_t:n avulla, missä X_t täyttää (3.2) vaatimukset ja u_t on mittauksesta ja tutkimuksista johtuva mittavirhe.

Oletetaan, että x_t on satunnaismuuttuja, jonka varianssi, $\sigma^2_{xx} > 0$. Oletetaan, että

$$(x_t, e_t, u_t)' \sim NI[(\mu_x, 0, 0)', diag(\sigma_{xx}, \sigma_{ee}, \sigma_{uu})], \tag{3.3}$$

missä NI tulee sanoista ”distributed normally and independently” eli normaalijakautuneet ja toisistaan riippumattomat. $diag(\sigma_{xx}, \sigma_{ee}, \sigma_{uu})$ tarkoittaa diagonaalimatriisia jossa annetut elementit ovat diagonalialkioita. (Fuller 1987)

Mallista (3.3) seuraa, että vektori $(Y_t, X_t)'$, missä Y_t on määritelty (3.1) ja X_t on määritelty (3.2), on kahden muuttujan normaalivektori, jonka odotusarvovektori on muotoa:

$$E\{(Y, X)\} = (\mu_y, \mu_x) = (\beta_0 + \beta_1 \mu_x, \mu_x) \tag{3.4}$$

ja kovarianssimatriisi

$$\begin{bmatrix} \sigma_{YY} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{XX} \end{bmatrix} = \begin{bmatrix} \beta_1^2\sigma_{xx} + \sigma_{ee} & \beta_1\sigma_{xx} \\ \beta_1\sigma_{xx} & \sigma_{xx} + \sigma_{uu} \end{bmatrix} \tag{3.5}$$

Nimetään

$$\hat{\gamma}_{1l} = \left[\sum_{t=1}^{n} (X_t - \bar{X})^2 \right]^{-1} \sum_{t=1}^{n} (X_t - \bar{X})(Y_t - \bar{Y}) \tag{3.6}$$

regressiokertoimeksi, joka on laskettu havaituista muuttujista. Kahden muuttujan normaalijakauman ominaisuuksien mukaan,

$$E\{\hat{\gamma}_{1l}\} = \sigma_{XX}^{-1}\sigma_{XY} = \beta_1(\sigma_{xx} + \sigma_{uu})^{-1}\sigma_{xx}. \tag{3.7}$$

On tärkeä muistaa, että kaava (3.7) johdettiin oletuksella, että mittavirhe X_t:ssä on riippumaton todellisista arvoista x_t ja virheistä ε_t. (Fuller 1987)

Fuller (1987) kertoo kirjassaan luotettavuussuhteen kuvaavan regressiokertoimen heikentymistä. Luotettavuussuhde on muotoa $\kappa_{xx} = \sigma_{XX}^{-1}\sigma_{xx}$. Parametrille β_1 saadaan estimaatti kaavalla

$$\hat{\beta}_1 = \kappa_{xx}^{-1}\hat{\gamma}_{1l}, \tag{3.8}$$

Parametrin β_0 estimaatti ei kuitenkaan ole muotoa

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X},$$

koska $\hat{\beta}_0$ ehdollinen odotusarvo on X:n funktio. Nyt $\hat{\beta}_0$ on muotoa

$$\hat{\beta}_0 = \beta_0 + \bar{v} - (\hat{\beta}_1 - \beta_1)\bar{X}, \tag{3.9}$$

missä $\bar{v} = n\sum_{t=1}^{n} v_t$ ja $v_t = e_t - u_t\beta_1$.(Fuller 1987)

3.1 Parametrien estimaatit, kun mittavirheen varianssi oletetaan tunnetuksi

Koska $\boldsymbol{Z}_t = (Y_t, X_t)$ on kaksiulotteinen normaalijakauma ja otos keskiarvo on $\bar{\boldsymbol{Z}} = (\bar{Y}, \bar{X})$ ja otos kovarianssit (m_{YY}, m_{XY}, m_{XX}), missä esimerkiksi,

$$m_{XY} = (n-1)^{-1} \sum_{t=1}^{n} (X_t - \bar{X})(Y_t - \bar{Y})$$

Suurimman uskottavuuden estimaatti $\boldsymbol{Z}$:n kovarianssimatriisille on

$$m_{ZZ} = (n-1)^{-1} \sum_{t=1}^{n} (\boldsymbol{Z}_t - \bar{\boldsymbol{Z}})'(\boldsymbol{Z}_t - \bar{\boldsymbol{Z}})$$

$\boldsymbol{Z}$ on 2*n matriisi ja $\bar{\boldsymbol{Z}}$ on siihen liittyvä odotusarvo matriisi. Parametreille β_0 ja β_1 saadaan estimaatit laskettua seuraavasti:

(3.10) $$\hat{\beta}_1 = (m_{XX} - \sigma_{uu})^{-1} m_{XY}$$

(3.11) $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

(Fuller 1987)

4 Berksonin mittavirhemalli

Johdannossa esiteltyjen menetelmien avulla on haastavaa tehdä analyysiä tosielämän aineistoon, koska useimmissa tapauksissa u_t ja σ_{uu} ei ole tiedossa. Mittavirhemallia voi kuitenkin käyttää tapauksissa, joissa voidaan olettaa X_t vakioksi. Esimerkkinä tilanteesta, jossa X_t voidaan olettaa vakioksi: Puhdistusainetta valmistaessa kone annostelee jokaiseen tuotteeseen salaista ainesosaa X aina saman verran. Nyt jokaisessa purkissa voidaan olettaa X_t:n olevan kiinteä. Todellinen ainesosan määrä kuitenkin vaihtelee, koska annostelukoneen annostelupaineessa on satunnaista vaihtelua. Tilanne on niin sanottu kiinteän X:n tapaus ja tähän saadaan sovellettua Berksonin mittavirhemallia.

Mittavirhe huomioiden malli on muotoa

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \qquad X_t = x_t + u_t, \qquad \boldsymbol{\varepsilon_t}' \sim NI(\mathbf{0}, \boldsymbol{\Sigma_{\varepsilon\varepsilon}}), \tag{4.1}$$

missä $\boldsymbol{\varepsilon_t} = (e_t, u_t)$ (Fuller 1987). Jos ylempänä esitetyn esimerkin tapauksessa annostelukone on kalibroitu oikein, niin u_t:n keskiarvo on nolla. Niinpä,

$$x_t = X_t - u_t, \tag{4.2}$$

missä u_t ovat $N(0, \sigma_{uu})$ jakautuneita satunnaismuuttujia (Fuller 1987). Tällaisessa tilanteessa mittausten tekijä kontrolloi mitattua arvoa X_t, koska on tärkeä, että arvo pysyy vakiona. Jos ylemmän esimerkin tapauksessa oletetaan vaihteluiden annostelupaineessa olevan toisistaan riippumattomat koneen annostelumäärän asetuksen kanssa, silloin u_t on riippumaton X_t:n kanssa. Berkson (1950) huomasi, että kun X_t on kontrolloitu, niin voidaan parametrit estimoida pienimmän neliösumman menetelmällä. (Fuller 1987).

4.1 Parametrien estimointi Berksonin mallissa

Olkoon

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \qquad x_t = X_t - u_t, \qquad t = 1, 2, \ldots, n,$$

missä (ε_t, u_t) ovat riippumattomia vektoreita odotusarvolla nolla ja kovarianssi

$$E\{(\varepsilon_t, u_t)'(\varepsilon_t, u_t)\} = diag(\sigma_{\varepsilon\varepsilon}, \sigma_{uu})$$

ja $\boldsymbol{X}' = (X_1, X_2, \ldots, X_n)$ on vektori jossa on kiinteitä vakioarvoja. Olkoon

$$\begin{aligned} \hat{\beta}_{1l} &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \\ \hat{\beta}_{0l} &= \bar{Y} - \hat{\beta}_{1l}\bar{X}, \end{aligned} \tag{4.3}$$

pienimmän neliösumman estimaattorit parametreille β_0 ja β_1. Tästä seuraa,

$$E\{(\hat{\beta}_{0l}, \hat{\beta}_{1l}\} = (\beta_0, \beta_1)$$

ja

$$\mathbf{V}\{(\hat{\beta}_{0l}, \hat{\beta}_{1l})\} = \begin{bmatrix} n^{-1} + \bar{X}^2 A_{XX}^{-1} & -\bar{X}A_{XX}^{-1} \\ -\bar{X}A_{XX}^{-1} & A_{XX}^{-1} \end{bmatrix} \sigma_{vv}, \tag{4.4}$$

missä $A_{XX} = (n-1)m_{XX}$, $v_t = e_t - u_t\beta_1$ ja $\sigma_{vv} = \sigma_{ee} + \beta_1^2\sigma_{uu}$.(Fuller 1987).

Todistus. Kun sijoitetaan x_t:n määritelmä malliin, niin saadaan

$$Y_t = \beta_0 + \beta_1 X_t + v_t, \tag{4.5}$$

missä $v_t = e_t - u_t\beta_1$. Oletuksista johtuen (e_t, u_t) on riippumaton X_t:n suhteen. Niinpä v_t on riippumaton X_t:stä. Sijoittamalla (4.5) kaavaan (4.3) saadaan

$$\begin{aligned} \hat{\beta}_{0l} - \beta_0 &= \bar{v} - \bar{X}(\hat{\beta}_{1l} - \beta_1) \\ \hat{\beta}_{1l} - \beta_1 &= \left[\sum_{i=1}^{n}(X_t - \bar{X})^2\right]^{-1} \sum_{i=1}^{n}(X_t - \bar{X})(v_t - \bar{v}) \end{aligned}$$

□

(Fuller 1987)

5 Aineiston simulointi

Päädyin käyttämään Berksonin mallilla simuloitua aineistoa, koska kunnon käytännön dataa en yrityksistä huoimatta onnistunut löytämään.

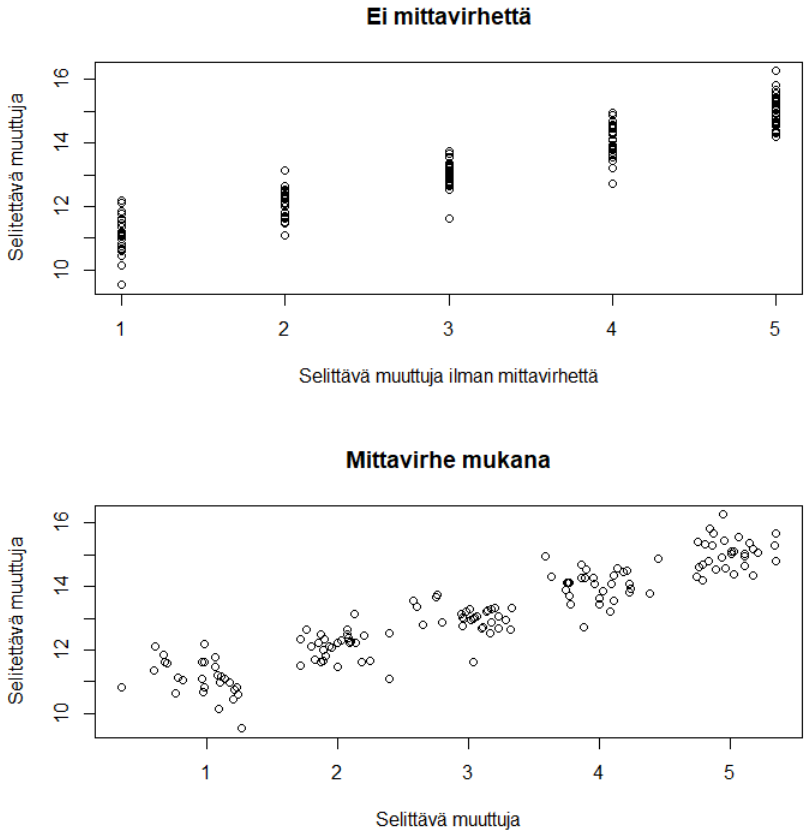
5.1 Simuloinnissa käytetty R-koodi

Aineiston generoinnin suoritin seuraavalla R-komennolla.

```
set.seed(2024)
#————————————————————————
# Muodostetaan havaintodata simuloimalla
>x <- rep(1:5, each=30)
>xt <- x + rnorm(150, sd=0.2)
>y <- 10 + 1*x + rnorm(150, sd=0.5)
```

5.2 Simuloidun aineiston graafinen tarkastelu

Simulointi tuottaa 150 tilastoyksikön aineiston, jossa on yksi selitettävä ja yksi selittävä muuttuja. Selittävä muuttuja saa kokonaislukuarvoja välillä [1,5]. Simuloidusta datasta on muodostettu kaksi mallia. Ensimmäinen malli on niin sanottu todellinen malli, jossa selittävänä muuttujana on käytetty todellisia arvoja. Toinen malli eli niin sanottu Berksonin mittavirhemalli on taas muodostettu kuvitteellisesti mitatuista arvoista, joissa on mukana mittavirhe. Selittävän muuttujan mittavirheen varianssi on 0.2. Myös selitettävälle muuttujalle on simuloitu satunnaisvaihtelua mukaan. Tämän satunnaisen vaihtelun varianssi on 0.5. Luonnollisesti molempien satunnaisvaihteluiden odotusarvot ovat nolla. Selitettävä muuttuja näyttää kuvaajalla seuraavalta.



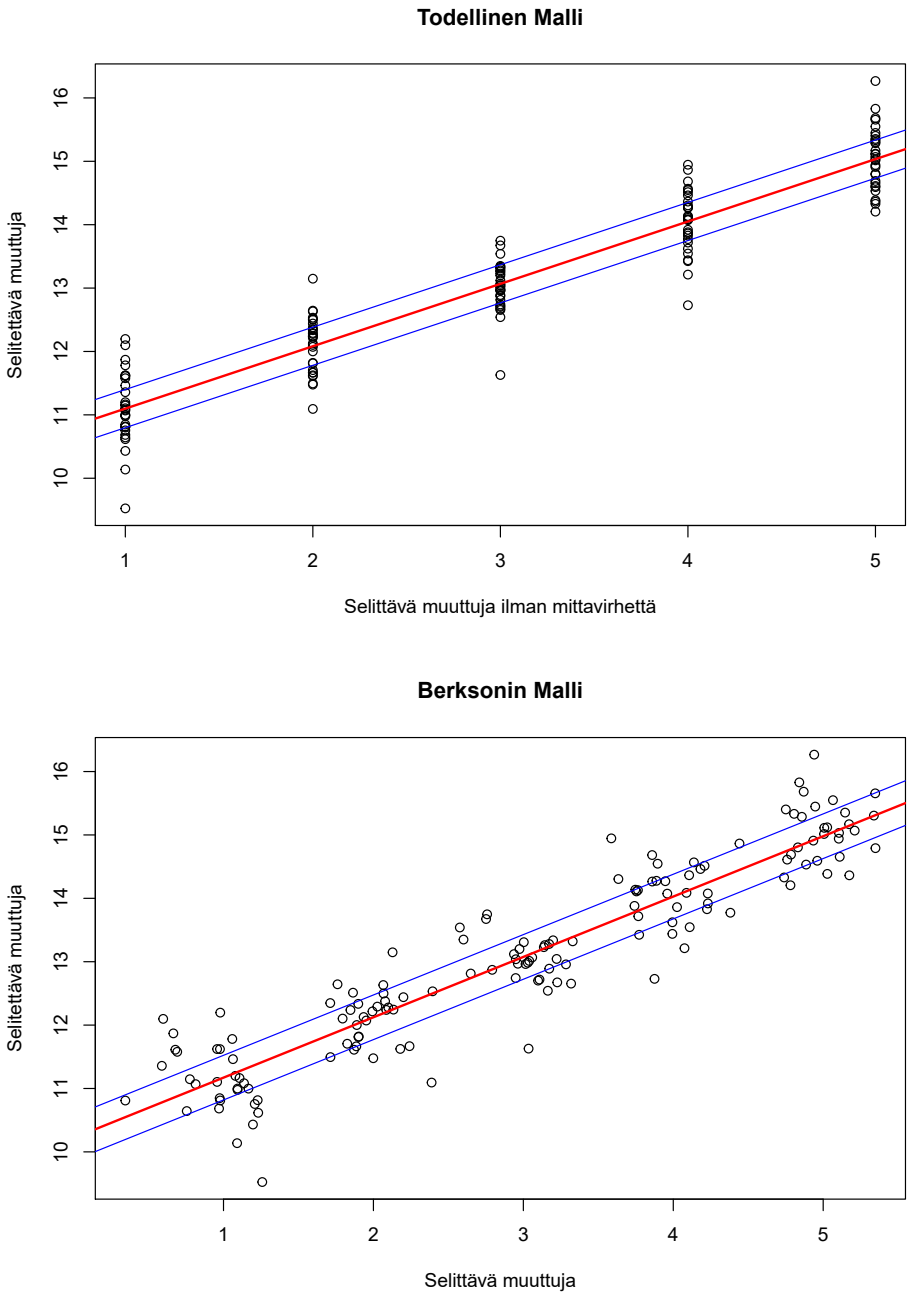
Kuva 5.1. Selitettävä muuttuja kuvattu selittävän muuttujan avulla tapauksessa, jossa ei ole mittavirhettä ja tapauksessa jossa on mittavirhe.

Kuvasta 5.2 huomaa, kuinka mittavirhe levittää havaintoja todellisen arvon ympärille. Ilman mittavirhettä jokainen arvo osuu samalle linjalle muiden samankokoisten arvojen kanssa. Mittavirheestä johtuen arvot muodostavat parven saman arvoisten mittausten kanssa. Kuvaajasta voi myös nähdä, että selittävällä ja selitettävällä muuttujalla on lineaarista riippuvuutta sekä virheettömässä, että virhettä sisältävässä kuvassa.

6 Simuloidun aineiston analysointi

Simulointiaineiston analysoinnissa käytetään jo aiemmin esiteltyä Berksonin mittavirhemallia. Muodostetaan lineaariset mallit tapauksesta, jossa mittavirhettä ei ole ja tapauksesta, jossa mittaamisesta aiheutuva virhe on mukana.

Kuvasta 6.1 huomataan, että lineaariset mallit näyttävät melko samanlaisilta. Niiden väliltä kuitenkin löytyy huomattavasti eroa. Kastotaan seuraavaksi yhteenvedot molemmista malleista. Lineaarimallista, jossa ei ole mittavirhettä ja mittavirhemallista jonka selittävä muuttuja sisältää mittavirheen.



Kuva 6.1. Lineaarinen malli tapauksesta jossa ei ole mittavirhettä ja Berksonin mittavirhemalli. Punainen viiva kuvaa muodostettua lineaarista mallia, ja siniset viivat kuvaavat mallin varianssia.

```
> summary(slm)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.57402 -0.31529  0.02194  0.28011  1.23210

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.11421    0.09142   110.6   <2e-16 ***
x            0.98406    0.02756    35.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4774 on 148 degrees of freedom
Multiple R-squared:  0.896,     Adjusted R-squared:  0.8953
F-statistic:  1275 on 1 and 148 DF,  p-value: < 2.2e-16

> summary(berksonlm)

Call:
lm(formula = y ~ xt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89301 -0.33943 -0.00114  0.31377  1.34384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.21895    0.10181  100.37   <2e-16 ***
xt           0.95218    0.03069   31.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5403 on 148 degrees of freedom
Multiple R-squared:  0.8668,    Adjusted R-squared:  0.8659
F-statistic: 962.9 on 1 and 148 DF,  p-value: < 2.2e-16
```

Kuva 6.2. R-ohjelmiston tulosteet yksinkertaiselle mittavirhemallille ja Berksonin mittavirhemallille.

Kuvan 6.2 esittämistä R-ohjelman ajotulosteista huomataan, että Berksonin mallin keskihajonta on suurempi, kuin virheettömässä tilanteessa. Virheettömän mallin jäännösten keskivirhe on 0.4774 ja Berksonin mittavirhemallissa se on 0.5403. Tämä tarkoittaa sitä, että Berksonin mallissa arvot eivät ole yhtä tiiviisti jakautuneita, vaan arvoilla on suurempaa vaihtelua. Mallien selitys osuuksissa, eli R^2 arvoissa, on myös eroja. R^2 kuvaa kuinka monta prosenttia malli kuvaa oikeiden arvojen satunnaisvaihtelusta. Molemmissa malleissa se on hyvä yli 80%. Virheettömässä tilanteessa se on kuitenkin parempi (89.6%) kuin virheen sisältämässä tilanteessa (86.68%).

7 Johtopäätökset

Molemmat lineaariset mallit antavat todella hyvät mallit simuloidulle datalle. Selitysosuudet molemmissa malleissa ovat korkeat. Mallit olisivat todellisuudessa hyvin käyttökelpoisia selitettävän muuttujan mallintamiseen. Selitysasteet eroavat toisistaan hieman, sillä virheettömässä mallissa selitysaste on 89.6% ja mittavirhemallissa 86.68%. Myös mallien jäännösten keskivirheillä on eroa. Virheettömän mallin jäännösten keskivirhe on 0.4774 ja mittavirhemallin 0.5403. Nämä erot malleissa selittyvät selittävän muuttujan sisältämällä mittavirheellä. Kyseinen mittavirheen varianssi on 0.2. Varsinkin simuloidun aineiston tilanteessa selitettävä muuttuja riippuu aidosti selittävästä muuttujasta. Jos muuttujassa on mittauksesta johtuvaa virhettä, niin silloin mallissakin on hieman enemmän varianssia ja selitysaste on heikompi.

Virheetön malli on kuitenkin parempi selitysasteen, sekä mallin varianssin suhteen. Herää kysymys, millaisissa tilanteissa mittavirhemalli on käytännöllinen? Simuloidun datan ansiosta voidaan olla varmoja mitattavan muuttujan todellisista arvoista, jolloin voidaan muodostaa virheetön malli. Tämä ei kuitenkaan luonnossa tapahtuvissa tilanteissa ole näin. Mittaamisessa tapahtuu aina virhettä, jolloin virheetöntä tilannetta ei voida muodostaa. Jos näissä tapauksissa halutaan muodostaa paras mahdollinen malli, voidaan käyttää siihen mittavirhemalleja kuten Berksonin malli ja muita malleja, joita voidaan muodostaa mittavirhe huomioiden.

Lähteet

[1] Berkson, J. (1950). *Are there two regressions?*. Journal of the american statistical association, 45(250), 164-180.

[2] Buonaccorsi, J.P. (2010). *Measurement Error: Models, Methods, and Applications (1st ed.)*. Chapman and Hall/CRC.

[3] Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

[4] Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis.* Elsevier.

[5] Fuller, W. A. (1987). *Measurement error models*. John Wiley & Sons.

[6] Stefanski, L. A. (2000). *Measurement error models*. Journal of the American Statistical Association.