

Saana Kaidesoja

LEHMÄNMAIDON SOLUTASOON VAIKUTTAVIEN TEKIJÖIDEN TUTKIMINEN LINEAARISELLA SEKAMALLILLA

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Toukokuu 2025

Tiivistelmä

Saana Kaidesoja: Lehmänmaidon solutasoon vaikuttavien tekijöiden tutkiminen
linearisella sekamallilla
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin tutkinto-ohjelma
Toukokuu 2025

Lehmänmaidon somaattisten solujen määrään vaikuttavat monet eri tekijät. Maidon solutasoa on tärkeää mitata ja seurata mahdollisten utaretulehdusten varalta. Utaretulehdukset ovat yksi maitotilojen suurimmista kulueristä, joten niiden ennaltaehkäisy on hyvin tärkeää. Somaattisten solujen määrä vaikuttaa myös maidosta maksettavaan hintaan.

Tässä tutkielmassa luodaan lineaarinen sekamalli R-ohjelmiston lmer-funktion avulla. Aineiston muuttujat jaetaan kiinteisiin vaikutuksiin ja satunnaisvaikutuksiin. Mallin tavoitteena on löytää solutasoon vaikuttavat tekijät. Mallin luomisessa hyödynnetään REML-menetelmää. Mallista lasketaan studentoituja jäännöksiä, joiden tarkoituksena on löytää solutason poikkeavia havaintoja. Aineistona käytetään Rengon maitotilan lehmille tehtyjä koelypsymittauksia.

Mallin avulla havaittiin, että lehmän iällä on vaikutusta maidon solutasoon. Holstein-rotuisten lehmien solutaso oli keskimäärin korkeampi kuin ayrshire-rotuisten. Mallin mukaan yksilöiden välillä solutasossa oli voimakasta vaihtelua. Keskilämpötilalla ja maidon määrällä ei ollut selvää vaikutusta solutasoon. Studentoitujen jäännösten avulla aineistosta löydettiin 15 poikkeavaa havaintoa, joista kuusi havaittiin samalla yksilöllä. Luotua sekamallia tulisi vielä kehittää tai aineistoon tulisi kerätä lisää havaintoja, jotta tilastollisesti solutasoon vaikuttavia tekijöitä voitaisiin mallin avulla tunnistaa enemmän.

Avainsanat: studentoitu jäännös, toistomittausmalli, kiinteä vaikutus, satunnaisvaikutus, BLUP, ML, REML

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

Tekoälyn käyttö opinnäytteessä

Opinnäytteessäni on käytetty tekoälysovelluksia:

Kyllä,

Ei.

Tekoälysovellusten nimi ja versio

ChatGPT, GPT-4

Käyttötarkoitus

Olen käyttänyt tekoälyä oikeinkirjoituksen tarkistamisessa sekä R-koodien virheilmoitusten selvittämisessä. Tekoälyä olen käyttänyt myös apuna englanninkielisen kirjallisuuden kääntämisessä suomen kielelle.

Osiot, joissa tekoälyä on käytetty

Oikeinkirjoituksen olen tarkistanut tekoälyllä osioissa: tiivistelmä, johdanto, tutkimusaineisto, aineiston mallintaminen ja johtopäätökset.

Olen tietoinen siitä, että olen täysin vastuussa koko opinnäytteeni sisällöstä, mukaan lukien tekoälyllä tuotetut osat, ja hyväksyn vastuun mahdollisista eettisten ohjeiden rikkomuksista.

Sisällys

1	Johdanto	1
2	Pitkittäisaineiston lineaarinen sekamalli	2
2.1	Toistomittausmallin perusteet	2
2.2	Ehdollisen odotusarvon estimointi	3
2.3	Varianssiparametrien estimointi	5
2.4	Jäännökset	7
3	Tutkimusaineisto	8
3.1	Tutkimusongelman tausta	8
3.2	Aineiston esittely	8
4	Aineiston mallintaminen	12
4.1	Kiinteä vaikutus vai satunnaisvaikutus	12
4.2	Mallin muodostaminen ja analysointi	12
4.3	Jäännösten tarkastelu	14
5	Johtopäätökset	17
	Lähteet	18

1 Johdanto

Lehmänmaidon somaattiset solut ovat keskeinen mittari maidon laadun ja utareterveyden arvioinnissa. Somaattisten solujen mittaaminen ja seuraaminen on tärkeää, jotta utaretulehduksia voidaan ennaltaehkäistä ja näin säästyä suurilta kulueriltä maitotilalla. Maidon somaattisten solujen määrään vaikuttavat niin yksilölliset ominaisuudet kuin ulkopuoliset tekijät. Ulkopuolisia tekijöitä voivat olla esimerkiksi mittaushetkenä vallinnut ulkolämpötila.

Tutkielman aineisto koostuu Rengon maitotilan lypsylehmille tehdyistä koelypsymittauksista, jotka on kerätty joka toinen kuukausi. Yhteensä aineistossa on havaintoja 35:lta eri lehmältä. Havaintojen kokonaismäärä on 627. Koelypsymittaukset sisältävät jokaiselta yksilöltä mittauspäivänä tuotetun maidon määrän kiloina (Mkg) sekä maidon somaattisten solujen määrän (10^3 kpl/ml). Lehmät on jaettu viiteen eri ikäryhmään sekä kahteen rotuun. Koelypsypäiviltä on haettu päivän keskilämpötila Ilmatieteenlaitoksen sivuilta.

Työn tavoitteena on tutkia solutasoon vaikuttavia tekijöitä lineaarisen sekamallin avulla. Lineaarinen sekamalli on malli, jossa yhdistyvät kiinteät vaikutukset ja satunnaisvaikutukset. Satunnaisvaikutusten avulla voidaan tarkastella yksilöiden välistä vaihtelua. Tavoitteena on löytää mallin avulla somaattisten solujen määrään tilastollisesti merkittävästi vaikuttavia tekijöitä. Sekamallin avulla voidaan laskea erilaisia jäännöksiä. Tässä tutkielmassa lasketaan studentoituja jäännöksiä, joiden avulla pyritään löytämään poikkeavia solutaso arvoja.

Tutkielman luvussa 2 tutustutaan pitkittäisaineiston lineaarisen sekamallin perusteisiin, ehdollisen odotusarvon ja varianssiparametrien estimointiin sekä jäännöksiin. Luvussa 3 esitellään tutkielman aineisto sekä tutkimusongelma tarkemmin. Luvussa 4 luodaan aineistosta sekamalli ja analysoidaan saatuja tuloksia. Luvussa tarkastellaan myös aineistossa esiintyviä poikkeavia havaintoja. Viimeisessä luvussa 5 kootaan työn keskeiset tulokset ja arvioidaan niiden merkitystä. Lisäksi luvussa pohditaan mallin toimivuutta sekä esitetään mahdollisia parannusehdotuksia.

2 Pitkittäisaineiston lineaarinen sekamalli

2.1 Toistomittausmallin perusteet

Lineaariseksi sekamalliksi kutsutaan mallia, jossa yhdistyvät kiinteät vaikutukset ja satunnaisvaikutukset. Mallia hyödynnetään kuvaamaan vastemuuttujan ja selittävien muuttujien suhteita sellaisessa aineistossa, joka on ryhmitelty yhden tai useamman tekijän mukaan. (Pinheiro & Bates 2000, s. 1.)

Mallia voidaan kutsua myös toistomittausmalliksi tai hierarkkiseksi malliksi. Sekamallia käytetään usein analysoimaan pitkittäisaineistoja. Yleisiä käyttökohteita ovat esimerkiksi lääketieteelliset tai biologiset aineistot. (Demidenko 2013, s. 1.) Tämän luvun merkinnät perustuvat teokseen *Parameter Estimation and Inference in the Linear Mixed Model* (Gumedze & Dunne 2011).

Pitkittäisaineistossa jokaisella havainnolla i on oma satunnaisvektori, joka on muotoa

$$\mathbf{y}_i = \{y_{it}\} = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix}, \text{ missä } t = 1, 2, \dots, T.$$

Jokaiselle havainnolle i voidaan muodostaa oma sekamalliyhtälö, joka esitetään muodossa

$$(2.1) \quad \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \dots, N,$$

missä $\mathbf{X}_i\boldsymbol{\beta}$ kuvaa mallin kiinteää osaa, $\mathbf{Z}_i\mathbf{u}_i$ satunnaisosaa ja $\boldsymbol{\varepsilon}_i$ satunnaisvirhettä. Kun kaikkien yksittäisten havaintojen mallit (2.1) yhdistetään, ne voidaan esittää muodossa

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{Z}_N \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{pmatrix}.$$

Nyt koko aineistolle muodostettu lineaarinen sekamalli on muotoa

$$(2.2) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

missä \mathbf{y} on $n \times 1$ selitettävä vektori. $\mathbf{X}\boldsymbol{\beta}$ kuvaa mallin kiinteää osaa, missä \mathbf{X} on $n \times p$ -suunnittelumatriisi ja $\boldsymbol{\beta}$ on $p \times 1$ parametrivektori. Vastaavasti $\mathbf{Z}\mathbf{u}$ kuvaa mallin satunnaisosaa, missä \mathbf{Z} on $n \times q$ -suunnittelumatriisi ja \mathbf{u} on $q \times 1$ satunnaisvaikutusten vektori. Mallin satunnaisvirheitä kuvaa $\boldsymbol{\varepsilon}$, joka on $n \times 1$ vektori.

Lineariselle sekamallille pätee oletus, että

$$E(\mathbf{u}) = \mathbf{0} \text{ ja } E(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

Lisäksi voidaan olettaa, että \mathbf{u} ja $\boldsymbol{\varepsilon}$ noudattavat moniulotteista normaalijakaumaa

$$\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{G}(\boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}(\boldsymbol{\rho}) \end{pmatrix} \right),$$

missä $\boldsymbol{\gamma}$ ja $\boldsymbol{\rho}$ ovat vektoreita, jotka sisältävät varianssiparametreja liittyen satunnaisiin tekijöihin ja satunnaisvirheisiin. Oletetaan, että \mathbf{u} ja $\boldsymbol{\varepsilon}$ ovat toisistaan riippumattomia, jolloin

$$\text{Cov}(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}.$$

Käytetään kovarianssimatriiseista seuraavia merkintöjä

$$\text{Var}(\mathbf{u}) = \sigma^2 \mathbf{G} \text{ ja } \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{R}.$$

Havaintojen \mathbf{y} kovarianssimatriisi voidaan nyt kirjoittaa muodossa

$$(2.3) \quad \text{Var}(\mathbf{y}) = \sigma^2 (\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}) = \sigma^2 \mathbf{H},$$

missä $\mathbf{H} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.

2.2 Ehdollisen odotusarvon estimointi

Kiinteiden vaikutusten $\boldsymbol{\beta}$ ja satunnaisvaikutusten \mathbf{u} estimoimiseen samanaikaisesti on olemassa monia erilaisia menetelmiä. Esimerkiksi Hendersonin sekamalliyhtälöt, Golbergin menetelmä sekä Bayes-estimointi (Gumedze & Dunne 2011). Tässä luvussa käsitellään kiinteiden vaikutusten ja satunnaisvaikutusten estimointia Hendersonin sekamalliyhtälöiden (engl. mixed model equations, MMEs) avulla. Hendersonin yhtälöillä on yhteys varianssiparametrien estimointiin suurimman uskottavuuden menetelmällä, jota käsitellään luvussa 2.3.

Gumedze & Dunne (2011) johtaa Hendersonin sekamalliyhtälöt maksimoimalla (\mathbf{y}, \mathbf{u}) logaritmisen yhteisjakauman (engl. log-joint distribution). Logaritminen yh-

teisjakauma voidaan esittää muodossa

$$\begin{aligned}
\log f(\mathbf{y}, \mathbf{u}) &= \log f(\mathbf{y}|\mathbf{u}) + \log f(\mathbf{u}) \\
&= -\frac{1}{2}(n \log \sigma^2 + \log \mathbf{R} + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})/\sigma^2) \\
&\quad - \frac{1}{2}(q \log \sigma^2 + \log \mathbf{G} + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u}/\sigma^2) \\
(2.4) \quad &= -\frac{1}{2}((n+q) \log \mathbf{R} + \log \mathbf{G} + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2) \\
&\quad - \frac{1}{2\sigma^2}(\mathbf{u}'(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\mathbf{u} - 2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{R}^{-1}\mathbf{Z}\mathbf{u}).
\end{aligned}$$

Osittaisderivoidaan yhteisjakauma (2.4) erikseen parametrien $\boldsymbol{\beta}$ ja \mathbf{u} suhteen, jolloin ratkaisuiksi saadaan

$$\begin{aligned}
\frac{\partial \log f(\mathbf{y}, \mathbf{u})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2}(-\mathbf{y}'\mathbf{R}^{-1}\mathbf{X} - \mathbf{y}'\mathbf{R}^{-1}\mathbf{X} + 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + 2(\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}\mathbf{X}) \\
&= -\frac{1}{2\sigma^2}(-2\mathbf{y}'\mathbf{R}^{-1}\mathbf{X} + 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + 2(\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}\mathbf{X}) \\
(2.5) \quad &= -\frac{1}{\sigma^2}(-\mathbf{y}'\mathbf{R}^{-1}\mathbf{X} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + (\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}\mathbf{X})
\end{aligned}$$

ja

$$\begin{aligned}
\frac{\partial \log f(\mathbf{y}, \mathbf{u})}{\partial \mathbf{u}} &= -\frac{1}{2\sigma^2}(2\mathbf{u}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + 2\mathbf{u}'\mathbf{G}^{-1} - 2\mathbf{y}'\mathbf{R}^{-1}\mathbf{Z} + 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}) \\
(2.6) \quad &= -\frac{1}{\sigma^2}(\mathbf{u}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{u}'\mathbf{G}^{-1} - \mathbf{y}'\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}).
\end{aligned}$$

Asetetaan saadut osittaisderivaatat (2.5) ja (2.6) nolliksi. Ratkaisuiksi saadaan pistehtälöt

$$\begin{aligned}
-\frac{1}{\sigma^2}(-\mathbf{y}'\mathbf{R}^{-1}\mathbf{X} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + (\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}\mathbf{X}) &= 0 \\
-\mathbf{y}'\mathbf{R}^{-1}\mathbf{X} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + (\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}\mathbf{X} &= 0 \\
-\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{u} &= 0 \\
(2.7) \quad \mathbf{X}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{u} &= 0
\end{aligned}$$

ja

$$\begin{aligned}
-\frac{1}{\sigma^2}(\mathbf{u}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{u}'\mathbf{G}^{-1} - \mathbf{y}'\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}) &= 0 \\
\mathbf{u}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{u}'\mathbf{G}^{-1} - \mathbf{y}'\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} &= 0 \\
\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{u} + \mathbf{G}^{-1}\mathbf{u} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} &= 0 \\
\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{u} - \mathbf{G}^{-1}\mathbf{u} &= 0 \\
(2.8) \quad \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\mathbf{u} &= 0.
\end{aligned}$$

Saadut yhtälöt (2.7) ja (2.8) voidaan esittää yksinkertaisemmassa matriisimuodossa seuraavasti

$$(2.9) \quad \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

Jos matriisit \mathbf{G} ja \mathbf{R} tiedetään, saadaan estimoitua parametrit $\boldsymbol{\beta}$ ja \mathbf{u} . Ratkaisuksi sekamalliyhtälöstä (2.9) saadaan paras lineaarinen harhaton estimaattori (engl. best linear unbiased estimator, BLUE)

$$(2.10) \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}^{-1}\mathbf{y},$$

sekä paras lineaarinen harhaton ennustin (engl. best linear unbiased predictor, BLUP)

$$(2.11) \quad \hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

BLUP-ennustin on yleinen tapa estimoida satunnaisvaikutuksia. Ennustinta käytetään paljon esimerkiksi eläinten jalostuksessa. BLUP-ennustin on nimensä mukaan lineaarinen, sillä satunnaismuuttujien \mathbf{u} arvot riippuvat havainnoista \mathbf{y} . Ennustin on harhaton, koska estimaatin odotusarvo vastaa estimoitavan arvon odotusarvoa. Lisäksi ennustimet ovat parhaita, sillä niiden keskineliövirhe on pienin mahdollinen. (Robinson 1991.) Voimme myös ratkaista varianssimatriisit estimaattoreille $\hat{\boldsymbol{\beta}}$ ja $\hat{\mathbf{u}}$ seuraavasti

$$(2.12) \quad \begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= \sigma^2 \left((\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}\mathbf{H}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1} \right) \\ &= \sigma^2 \left((\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1} \right) \end{aligned}$$

ja

$$(2.13) \quad \begin{aligned} \text{Var}(\hat{\mathbf{u}}) &= \sigma^2\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{H}\mathbf{P}\mathbf{Z}\mathbf{G} \\ &= \sigma^2\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}, \end{aligned}$$

missä $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}^{-1}$.

2.3 Varianssiparametrien estimointi

On olemassa lukuisia menetelmiä varianssiparametrien estimointiin lineaarisessa sekamallissa. Menetelmiä ovat esimerkiksi ANOVA, MINIQUE sekä Hendersonin menetit I, II ja III. Nykyään useimmin käytettyjä menetelmiä ovat suurimman uskottavuuden menetelmä (engl. Maximum likelihood, ML) sekä rajoitettu suurimman uskottavuuden menetelmä (engl. Restricted maximum likelihood, REML).

ML- ja REML-menetelmiä voidaan hyödyntää niin tasapainoisen kuin epätasapainoisen aineiston varianssiparametrien estimointiin. ML-menetelmä ei ota huomioon kiinteiden vaikutusten estimoinnissa menetettyjä vapausasteita, joten estimaattorit varianssiparametreille ovat usein alaspäin harhaisia. REML-menetelmä huomioi nämä menetetyt vapausasteet, joten sen avulla saadaan yleensä vähemmän harhaisia estimaattoreita. (Gumedze & Dunne 2011.)

Suurimman uskottavuuden estimoinnin tavoitteena on muodostaa uskottavuusfunktio havaittua aineistoa vastaavan todennäköisyysjakauman perusteella (Galecki & Burzykowski 2013, s. 256). Voidaan olettaa, että \mathbf{u} ja \mathbf{y} ovat yhteisesti normaali-jakautuneita (engl. jointly Gaussian distributed)

$$(2.14) \quad \begin{pmatrix} \mathbf{u} \\ \mathbf{y} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{G}(\boldsymbol{\gamma}) & \mathbf{G}(\boldsymbol{\gamma})\mathbf{Z}' \\ \mathbf{Z}\mathbf{G}(\boldsymbol{\gamma}) & \mathbf{H}(\boldsymbol{\gamma}, \boldsymbol{\rho}) \end{pmatrix} \right).$$

Näin ollen vektorin \mathbf{y} marginaalinen log-uskottavuusfunktio (engl. marginal log-likelihood funktion) on muotoa

$$(2.15) \quad l_{ML}(\boldsymbol{\beta}, \boldsymbol{\phi}; \mathbf{y}) = -\frac{1}{2} \left(n \log(2\pi) + n \log \sigma^2 + \log |\mathbf{H}| + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \right),$$

missä $\boldsymbol{\phi} = (\mathbf{k}', \sigma^2)'$ ja $\mathbf{k} = (\boldsymbol{\gamma}', \boldsymbol{\rho}')'$.

REML-menetelmässä maksimoidaan uskottavuusfunktiota, jossa selitettävälle vektorille \mathbf{y} tehdään muunnos, jolloin jakauma ei ole riippuvainen kiinteistä vaikutuksista $\boldsymbol{\beta}$. Uutta lineaarikombinaatiota merkataan $\mathbf{K}'\mathbf{y}$. \mathbf{K} on valittava niin, että se on täysiasteinen, sekä toteuttaa ehdon $E(\mathbf{K}'\mathbf{y}) = 0$. Yksinkertaisin tapa toteuttaa mainitut ehdot on varmistaa, että $\mathbf{K}'\mathbf{X} = 0$. \mathbf{K} :lla saa olla enintään $n - p$ määrää riippumattomia sarakkeita, missä p on mallin riippumattomien parametrien lukumäärä. (Robinson & Hamann 2011, s. 263-264.) REML-menetelmässä maksimoidaan muunnosta $\mathbf{K}'\mathbf{y}$ alkuperäisen \mathbf{y} :n sijaan. Nyt $\mathbf{K}'\mathbf{y}$ noudattaa normaalijakaumaa seuraavin parametrein

$$\mathbf{K}'\mathbf{y} \sim N(0, \sigma^2 \mathbf{K}'\mathbf{H}\mathbf{K}).$$

REML-uskottavuusfunktio voidaan esittää muodossa

$$\begin{aligned} l_R(\boldsymbol{\phi}; \mathbf{y}) &= -\frac{1}{2} \left((n - p) \log \sigma^2 + \log |\mathbf{H}| + \log |\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}| + \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\sigma^2} \right) \\ &= -\frac{1}{2} \left((n - p) \log \sigma^2 + \log |\mathbf{H}| + \log |\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}| + \frac{\mathbf{y}'\mathbf{P}\mathbf{y}}{\sigma^2} \right), \end{aligned}$$

missä $\hat{\boldsymbol{\beta}}$ on sama kuin kaavassa (2.10) ja \mathbf{P} on sama kuin kaavassa (2.13).

2.4 Jäännökset

Jäännösten (engl. residual) avulla voidaan tarkastella mallin oletuksia esimerkiksi normaalisuutta tai mahdollisia poikkeavia havaintoja. Lineariselle sekamallille voidaan muodostaa erilaisia jäännöksiä. Marginaaliset jäännökset (engl. marginal residual) ovat hyödyllisiä, kun halutaan tarkastella vaikutusten lineaarisuutta ja havaintojen kovarianssirakennetta. Ehdollisia jäännöksiä (engl. conditional residual) käytetään, kun halutaan tarkastella poikkeavia havaintoja. (Schützenmeister & Piepho 2012.) Ehdollinen jäännös on havaittujen arvojen ja ennustettujen arvojen erotus. Ehdollinen jäännös ottaa huomioon myös mallin satunnaisvaikutukset. (West et al. 2014, s. 41.) Ehdollinen jäännös voidaan esittää muodossa

$$(2.16) \quad \hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\tilde{\mathbf{u}},$$

missä $\hat{\boldsymbol{\beta}}$ ja $\tilde{\mathbf{u}}$ ovat samat kuin kaavoissa (2.10) ja (2.11). Nyt varianssi kaavasta (2.16) saadaan muotoon

$$(2.17) \quad \text{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2 \mathbf{T} = \sigma^2 \mathbf{RPR},$$

missä $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}^{-1}$ eli sama kuin aiemmin kaavassa (2.13).

Yleensä ehdolliset jäännökset äsken esitetystä muodosta (2.16) eivät sovellu kovinkaan hyvin mallin oletusten tai poikkeavien arvojen analysoimiseen, sillä ehdolliset jäännökset ovat usein korreloituneita ja niiden varianssit voivat olla eri. Tämä ongelma voidaan ratkaista jäännösten skaalaamisella eli jakamalla jäännökset keskihajonnalla tai arvioidulla keskihajonnalla. Harvoin tiedetään todellisia keskihajontoja, joten joudutaan tyytymään arvioituihin keskihajontoihin. Näin saatuja jäännöksiä kutsutaan studentoiduiksi jäännöksiksi (engl. studentized residual). (West et al. 2014, s. 42.) Studentoitu jäännös voidaan esittää muodossa

$$(2.18) \quad \hat{\varepsilon}_k^* = \frac{\hat{\varepsilon}_k}{\sqrt{\hat{t}_{kk}}},$$

missä \hat{t}_{kk} on estimaatti t_{kk} :sta, joka on k :nnes alkio \mathbf{T} diagonaalimatriisista. Matriisin \mathbf{T} alkiot kuvaavat kiinteiden vaikutusten ja satunnaisvaikutusten yhteisvaikutusta. (Schützenmeister & Piepho 2012.) Studentoitujen jäännösten tarkastelussa asetetaan raja, jonka ylittäessään havainto luokitellaan poikkeavaksi havainnoksi. Yleensä rajaksi asetetaan esimerkiksi |2| tai |3|.

3 Tutkimusaineisto

3.1 Tutkimusongelman tausta

Maidon somaattisia soluja pidetään maidon laadun ja lehmän utareterveyden mittarina. Somaattisia soluja käytetään yhtenä maidon luokitteluperusteista, jonka mukaan maidon laatu- ja hintaluokka muodostuvat. Taulukon 3.1 mukaan paras hinta maidosta maksetaan silloin, kun solutason kolmen kuukauden liukuva geometrinen keskiarvo on alle 250 000 solua/ml. Suomessa parhaaseen hintaluokkaan pääsee 97 % maidosta ja huonoimpaan II-luokkaan jää alle 0,05 %. Suomessa tuotetun maidon laatu on kansainvälisissä vertailuissa huippulaadukasta. (Aho et al. 2020.)

Taulukko 3.1. Tuottajamaidon laatuluokat ja hinnoitteluperuste

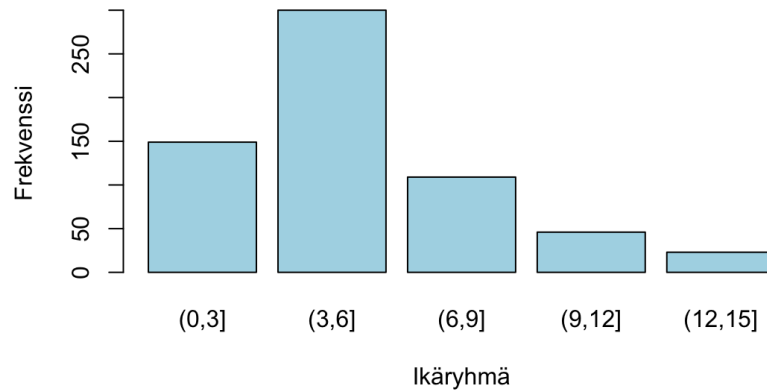
Luokka	Somaattisten solujen määrä/ml (geometrinen keskiarvo)
E	< 250 000
I	250 000 - 400 000
II	> 400 000

Mikäli maidon solutaso kohoaa yli 200 000 kpl/ml, kyseessä voi olla utaretulehdus. Utaretulehduksen lisäksi somaattisten solujen määrään voi vaikuttaa myös esimerkiksi poikimakerta, perinnölliset tekijät sekä vuodenaika. Tutkimusten mukaan maidon solutaso on korkeampi kesällä kuin talvella. (Isokallio 2015.) Maidon solutason mittaaminen ja seuranta on tärkeää, jotta eläimen utareterveyttä ja sen kehitystä voidaan seurata. Utaretulehdukset ovat iso kuluerä maitotiloille, joten niiden ennaltaehkäisyllä on suuri rooli. Utaretulehdukset ovat myös yksi suurimmista lehmien poistosityistä tiloilta, joten myös tämän kannalta solutason seuranta on hyvin tärkeää ja hyödyllistä. (Kervinen & Kervinen 2020.)

3.2 Aineiston esittely

Tutkielman aineisto sisältää jokaisesta 35 lehmästä yksilöllisiä tietoja, nimen ja syntymäajan. Syntymäajan avulla saadaan kullekin lehmälle ikä mittaushetkellä. Jaetaan aineisto viiteen ikäryhmään käyttämällä R-ohjelman cut-funktiota. Sama lehmä voi esiintyä samassa ikäryhmässä useammin kuin kerran, jos siltä on useita

mittauksia kyseisen ikäryhmän sisällä. Histogrammista 3.1 havaitaan, että suurin osa mittauksista on tehty lehmien ollessa 3-6 vuoden ikäisiä. Lehmä alkaa tuottamaan maitoa yleensä noin 2 vuoden ikäisenä ensimmäisen poikimäkerran jälkeen.



Kuva 3.1. Ikäryhmien jakauma. Frekvenssi kuvaa havaintojen kokonaismäärää, ei yksilöiden lukumäärää.

Tilan lehmät ovat joko ayrshire-rotuisia (AY=1) tai holstein-rotuisia (HOL=2). Aineistossa on 15 ayrshire-rotuista ja 20 holstein-rotuista lehmää. Ahon et al. (2020) mukaan holstein-rotuiset lehmät tuottavat keskimäärin 10 000 kg maitoa vuodessa kun taas ayrshire-rotuiset 9 000 kg. Taulukon 3.2 mukaan tutkielman aineisto tukee Ahon et al. (2020) väitettä, jonka mukaan holstein-rotuiset tuottavat enemmän maitoa kuin ayrshire-rotuiset. Taulukosta 3.2 havaitaan myös, että holstein-rotuisilla solutason keskiarvo on korkeampi kuin ayrshire-rotuisilla. Tämä ero keskiarvoissa voi kuitenkin selittyä esimerkiksi sillä, että holstein-rotuisista on aineistossa enemmän havaintoja, joten poikkeamien määrä voi olla myös suurempi.

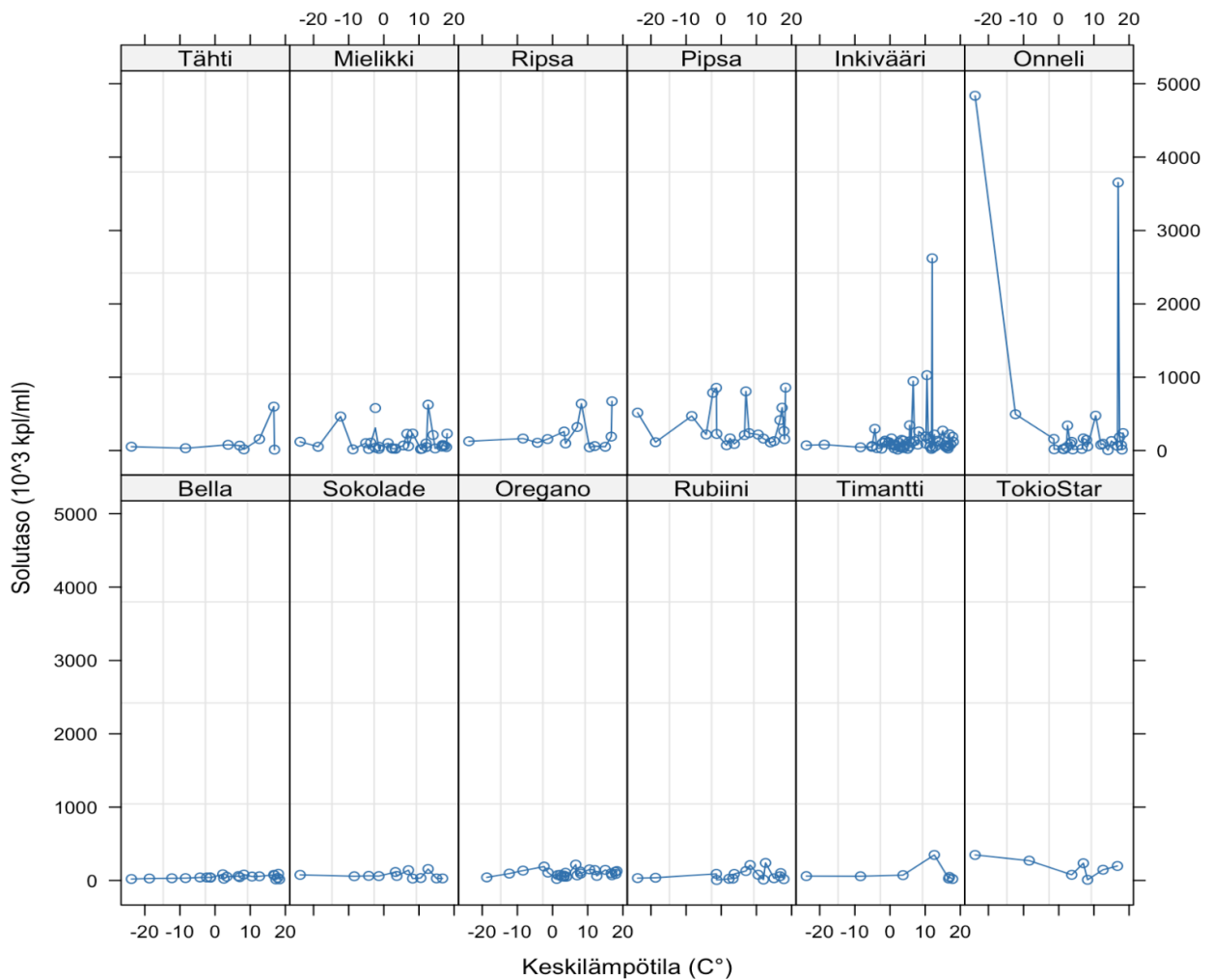
Taulukko 3.2. Maidontuotannon ja somaattisten solujen keskiarvot rotuittain

	Rodut 1 ja 2	Ayrshire (1)	Holstein (2)
Maidon tuotanto (Mkg)	30,6	28,4	32,3
Somaattiset solut (10^3 /ml)	186,7	170,9	198,4

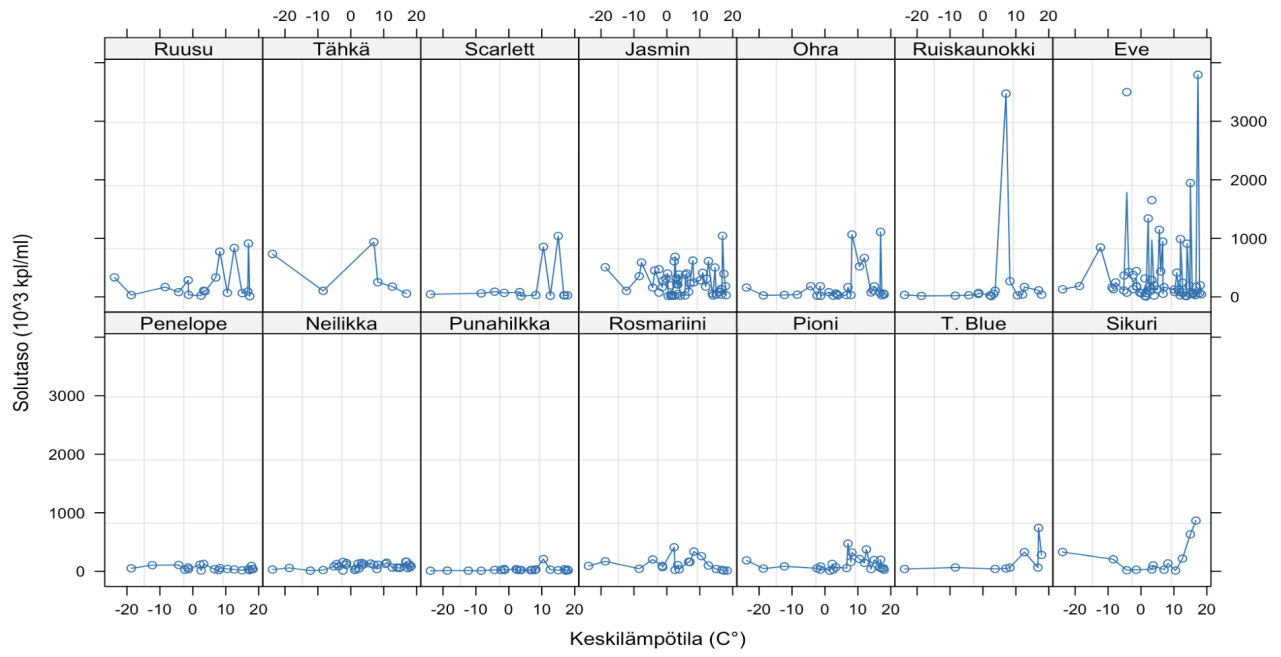
Aineisto sisältää myös jokaiselta koelypsypäivältä päivän keskilämpötilan. Lämpötilat on haettu Ilmatieteenlaitoksen nettisivulta (Ilmatieteenlaitos, Havaintojen la-

taus, n.d.). Havaintoasemana on käytetty Kauhavan lentokentän mittausasemaa.

Kuvien 3.2 ja 3.3 perusteella voidaan havaita, että solutaso on saanut yleensä – muutamia poikkeuksia lukuunottamatta – korkeampia arvoja, mitä korkeampi lämpötila on ollut. Kuvista huomataan myös, että yksilöillä on paljon eroavaisuuksia keskenään. Esimerkiksi Onnellilla, Inkiväärillä ja Evellä on yksittäisiä suuria poikkeavia havaintoja, kun taas Bellalla ja Penelopella poikkeavia havaintoja ei näy lainkaan. Sikurin kuvaajasta huomataan, että hänellä solutaso kohoaa lineaarisesti silloin, kun lämpötila nousee yli 10 asteen.



Kuva 3.2. Rodun 1 lehmien solutasojen vaihtelu lämpötilan mukaan (valikoitu otos)



Kuva 3.3. Rodun 2 lehmien solutasojen vaihtelu lämpötilan mukaan (valikoitu otos)

4 Aineiston mallintaminen

4.1 Kiinteä vaikutus vai satunnaisvaikutus

Ennen mallin rakentamista täytyy valita, mitä muuttujia käsitellään kiinteinä vaikutuksina ja mitä satunnaisvaikutuksina. Valinnat vaikuttavat mallin analyysiin ja siitä tehtäviin johtopäätöksiin. Osa analyytikoista ovat sitä mieltä, että luokittelu riippuu halutusta päättelystä, kun taas osa katsoo sen riippuvan ainoastaan tutkimusasetelmasta. Selkeää linjaa ei siis ole, milloin vaikutus on kiinteä ja milloin satunnainen. Osa vaikutuksista voi olla sekä kiinteitä että satunnaisia. (Robinson & Hamann 2011, s. 220.)

Kiinteät vaikutukset ovat usein sellaisia, jotka valitaan tarkoituksellisesti ja niiden tasot edustavat ainoastaan itseään. Yleinen esimerkki kiinteästä vaikutuksesta voi olla muuttuja sukupuoli. Tällöin kerätyt havainnot edustavat ainoastaan kyseistä populaatiota. (Robinson & Hamann 2011, s. 221.) Valitaan tässä tutkielmassa rotu yhdeksi kiinteäksi muuttujaksi, jolloin havainnot edustavat joko rotua 1 tai 2. Valitaan myös muuttuja ikäryhmä kiinteäksi vaikutukseksi, jolloin eri tasoja tulee viisi.

Satunnaisvaikutusten avulla pyritään mallintamaan populaation sisäistä vaihtelua yksilöiden tai ryhmien välillä. Satunnaisvaikutukset mahdollistavat tulosten yleistämisen myös suurempaan populaatioon, sillä ne edustavat satunnaista hajontaa eri tasoilla. (West et al. 2014, s. 9.) Valitaan muuttujat keskilämpötila sekä maidontuotanto niin, että ne vaikuttavat kiinteinä vaikutuksina sekä satunnaisvaikutuksina. Muodostetaan malli niin, että satunnaisvaikutukset vaihtelevat yksilöittäin. Näin pystytään tarkastelemaan satunnaisvaikutusten merkitystä yksilötasolla.

4.2 Mallin muodostaminen ja analysointi

Muodostetaan sekamalli käyttämällä R-ohjelman funktiota `lmer`, joka on osa `lme4` kirjastoja. Funktio `lmer` käyttää mallin sovittamiseen oletusarvoisesti REML-menetelmää. Jos malli haluttaisiin sovittaa käyttäen ML-menetelmää, täytyisi koodiin muuttaa argumentti `REML = FALSE`.

```
malli <- lmer(Solut ~ Keskilämpötila + Mkg + Rotu + ikäryhmä +  
             (Keskilämpötila + Mkg | Nimi),  
             data = koelypsyt, REML = TRUE)
```

Käyttämällä ImerTest-kirjastoa saadaan komennon summary avulla laskettua tunnuslukujen lisäksi myös kiinteille vaikutuksille p-arvot. P-arvojen avulla voidaan arvioida mitkä muuttujat ovat olleet tilastollisesti merkitseviä ja mitkä eivät.

Taulukko 4.1. Skaalatut jäännökset

Minimi	1Q	Mediaani	3Q	Maksimi
-1,54	-0,30	-0,15	0,04	9,56

Taulukon 4.1 tuloksista huomataan, että jäännösten jakauma ei ole symmetrinen. Suurin jäännösarvo 9,56 on suuri, joten mallissa todennäköisesti esiintyy poikkeavia havaintoja. Mallin jäännöksiä tarkastellaan tarkemmin luvussa 4.3.

Taulukko 4.2. Satunnaisvaikutukset

	Muuttuja	Varianssi	Keskihajonta	Korrelaatio	
Nimi	Vakiotermi	17666,8	132,9		
	Keskilämpötila	40,6	6,4	-0,60	
	Mkg	14,4	3,8	-0,55	-0,34
Jäännös		145872,1	381,9		

Taulukon 4.2 mukaan satunnaisvaikutuksista vakiotermin keskihajonta on 132,9, mikä on melko suuri. Tämä viittaa siihen, että yksilöiden välillä on suurta satunnaista vaihtelua. Myös keskilämpötila ja maidon määrä vaihtelee yksilöittäin, mutta niiden vaikutus ei ole niin merkittävää. Vakiotermin ja keskilämpötilan korrelaatio on -0,60. Yksilöt, joiden solutasot ovat valmiiksi korkeammat eivät ole niin herkkiä lämpötilan vaikutukseen. Lämpötilan muutokset vaikuttavat voimakkaammin niihin yksilöihin, joiden perustaso on alhaisempi.

Taulukko 4.3. Kiinteät vaikutukset

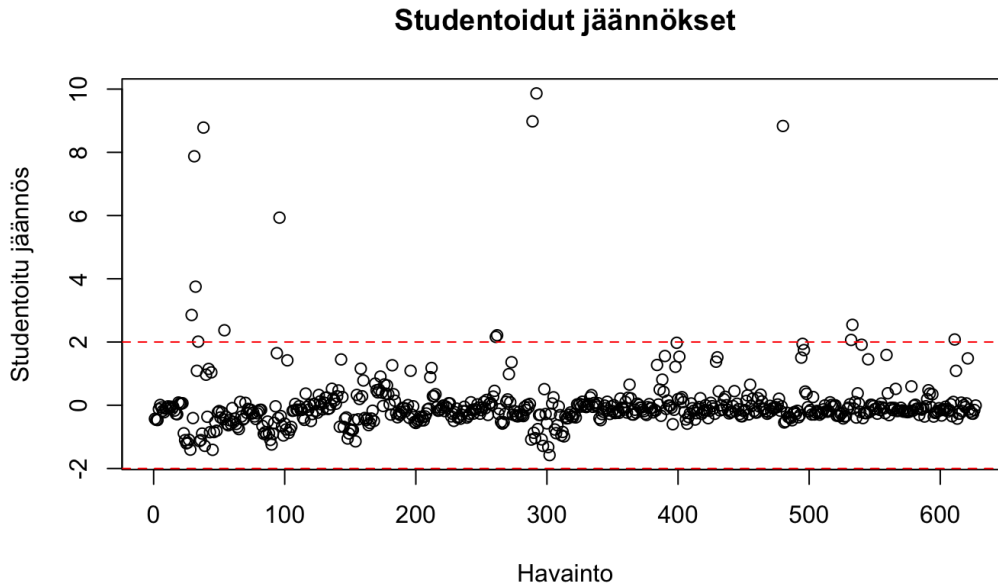
Kiinteä vaikutus	Estimaatti	Keskivirhe	Vapausasteet	t-arvo	p-arvo
Vakiotermi	60,55	88,04	56,40	0,69	0,494
Keskilämpötila	-0,01	1,88	30,60	-0,01	0,994
Mkg	-1,68	2,19	17,74	-0,77	0,452
Rotu	46,85	41,57	51,84	1,13	0,265
ikäryhmä(3,6]	68,58	40,41	554,63	1,70	0,090
ikäryhmä(6,9]	186,99	53,00	483,93	3,53	0,0005
ikäryhmä(9,12]	397,52	70,67	368,27	5,63	< 0,0001
ikäryhmä(12,15]	482,00	93,02	311,99	5,19	< 0,0001

Taulukosta 4.3 nähdään kiinteiden vaikutusten estimaatit, keskivirheet, vapausasteet, t-arvot ja p-arvot. Valitaan merkitsevyytasoksi yleinen 0,05. Jos saatu p-arvo on pienempi kuin valittu merkitsevyytaso, voidaan katsoa muuttujan olevan tilastollisesti merkitsevä. Vakiotermi, keskilämpötila, maidontuotanto (Mkg) tai rotu eivät ole tilastollisesti merkitseviä p-arvon perusteella. Estimaattien perusteella voidaan kuitenkin havaita, että rodulla on vaikutusta, vaikka vaikutus ei ole tilastollisesti merkitsevää. Tulosten mukaan rodulla 2 solutaso on 46,85 yksikköä korkeampi kuin rodulla 1. Saaduista tuloksista havaitaan, että ikäryhmien (6,9], (9,12] sekä (12,15] p-arvot ovat alle merkitsevyytason, joten ne ovat tilastollisesti merkitseviä. Jos merkitsevyytasoksi valittaisiin 0,1, myös ikäryhmä (3,6] olisi tilastollisesti merkitsevä. Vakiotermi sisältää tässä tapauksessa ikäryhmän (0,3], joka toimii vertailuryhmänä. Myös estimaattien arvoista huomataan, että mitä suurempi ikäryhmä, sitä suurempi estimaatti. Tulokset viittaavat siihen, että lehmän iällä on vaikutusta solutasoon. Tulosten mukaan mitä vanhempi lehmä, sitä korkeampi solutaso.

4.3 Jäännösten tarkastelu

Studentoidut jäännökset on laskettu käyttäen R-ohjelman kirjaston `redres` funktiota `compute_redres`. Tyyppinä on käytetty `std_cond`, joka laskee studentoidut jäännökset. Mallina käytetään luvussa 4.2 esiteltyä mallia. Asetetaan rajaksi $|2|$. Jos laskettu jäännös on suurempi kuin $|2|$, havainto voidaan laskea poikkeavaksi havainnoksi.

Kuvassa 4.1 x-akselilla on jokainen havainto sekä y-akselilla studentoitu jäännös. Punaiset katkoviivat kuvaavat rajoja, joiden ulkopuolella olevat havainnot ovat poikkeavia.



Kuva 4.1. Poikkeavien havaintojen tarkastelu studentoiduilla jäännöksillä

Kuvasta 4.1 voidaan silmämääräisesti havaita, että poikkeavia havaintoja ei ole kovinkaan paljoa suhteessa koko otoksen kokoon. Kuvasta voidaan huomata, että suurimman osan havaintojen studentoidut jäännökset ovat lähellä nollaa. Poimitaan kuvasta kaikki itseisarvon kaksi ylittävät havainnot ja taulukoidaan ne. Taulukosta 4.4 havaitaan, että eniten poikkeavia havaintoja on esiintynyt Evellä. Tämä voi osin selittyä sillä, että hänestä on eniten mittauskertoja. Kuitenkin voidaan huomata, että suurin osa poikkeavista havainnoista on tapahtunut iän ollessa yli 12 vuotta. Samoin huomataan, että muidenkin poikkeavat havainnot ovat lähtökohtaisesti tapahtunut vanhemmassa iässä. Silmämääräisesti taulukon mukaan keskilämpötilalla ei ole ollut vaikutusta solutason poikkeaviin havaintoihin, sillä lämpötilat vaihtelevat $-23,8$ asteesta aina $16,9$ asteeseen.

Taulukosta havaitaan myös, että poikkeavia havaintoja ei ole yhdenkään lehmän kohdalla peräkkäisillä koelypsykerroilla. Tämä voisi viitata siihen, että solutason muutokset ovat hetkellisiä, eivätkä ne vaikuta enää kahden kuukauden päästä. Studentoitujen jäännösten mukaan myöskään maidontuotannon määrällä kyseisenä päivänä ei ole ollut vaikutusta poikkeaviin havaintoihin muutamaa poikkeusta lu-

kuunottamtta. Ohralla ja Onnelilla on havaittu poikkeava havainto samana päivänä 13.08.2024. Todennäköisesti poikkeavien havaintojen osuminen samalle päivälle on ainoastaan sattumaa. On myös mahdollista, että kyseisenä ajankohtana solutasoon on vaikuttanut jokin ulkopuolinen tekijä esimerkiksi kiertävä tauti, jota malli ei ota huomioon.

Kun studentoituja jäännöksiä vertaa kuvaan 3.2, voidaan huomata, että Onnelilla on selkeästi kaksi poikkeavaa havaintoa ja Inkiväärillä yksi selvästi muita suurempi havaittu solutaso arvo. Kuvasta 3.3 voidaan havaita, että Evellä on selkeästi paljon vaihtelua solutasojen arvoissa. Tähkällä ja Scarletilla on myös poikkeavia havaintoja, vaikka solutaso arvo ei kohoakaan läheskään yhtä suuriin arvoihin kuin esimerkiksi Evellä tai Ruiskaunokilla. Silti arvot ovat yksilön tasoon nähden muista havainnoista poikkeavia, minkä vuoksi studentoiduissa jäännöksissä arvot nousevat esiin. Kuvan 3.3 mukaan esimerkiksi Jasminilla on vaihtelua solutasossa, mutta vaihtelu on jatkuva, joten arvot eivät muutu tarpeeksi paljon, jotta ne laskettaisiin poikkeaviksi havainnoiksi.

Taulukko 4.4. Poikkeavat havainnot studentoitujen jäännösten perusteella

Nimi	Rotu	Päivä	Mkg	Solutaso	Keskilämpötila	Ikäryhmä
Eve	2	17.04.2023	31	1652	3,3	(12,15]
Eve	2	12.05.2022	45,2	3498	-4,3	(12,15]
Eve	2	02.08.2022	7	1944	15	(12,15]
Eve	2	08.04.2022	31,6	1339	2,2	(12,15]
Eve	2	09.08.2021	43	3794	17,3	(9,12]
Eve	2	24.04.2018	5	1145	5,6	(6,9]
Inkivääri	1	03.06.2022	33,2	2631	12	(9,12]
Ohra	2	11.10.2024	37,4	1065	8,2	(6,9]
Ohra	2	13.08.2024	33,4	1109	16,9	(6,9]
Onneli	1	13.08.2024	36,6	3654	16,9	(6,9]
Onneli	1	09.02.2024	35,2	4837	-23,8	(6,9]
Ruiskaunokki	2	13.04.2024	35,8	3472	7	(3,6]
Scarlett	2	07.20.2022	33	854	10,5	(0,3]
Scarlett	2	02.08.2022	31,2	1039	15	(0,3]
Tähkä	2	13.04.2024	22	937	7	(0,3]

5 Johtopäätökset

Laadittu sekamalli selittää lehmänmaidon solutasen muutosta kiinteiden vaikutusten ja satunnaisvaikutusten avulla. Kiinteiden vaikutusten analyysin p-arvojen perusteella tilastollisesti merkitseviä muuttujia ovat ikäryhmät (6,9], (9,12] ja (12,15]. Muilla kiinteillä vaikutuksilla havaittiin olevan vaikutusta solutasen muutoksiin, mutta ne eivät olleet tilastollisesti merkitseviä valitulla merkitsevyystasolla 0,05. Kiinteiden vaikutusten perusteella rodulla 2 on keskimäärin korkeampi solutaso kuin rodulla 1. Tätä tulosta tukevat myös taulukossa 3.2 esitetyt solutasen keskiarvot roduittain.

Satunnaisvaikutusten vakiotermin keskihajonta oli 132,9, mikä viittaa yksilöiden väliseen voimakkaaseen vaihteluun. Keskilämpötilan ja maidontuotannon vaikutus yksilöittäin ei vaihtele merkittävästi. Satunnaisvaikutusten korrelaatioiden perusteella korkeampi solutasen vakio liittyy heikompaan keskilämpötilan vaikutukseen. Vastaavasti matalampi solutaso liittyy voimakkaampaan keskilämpötilan vaikutukseen yksilöittäin. Vastaava yhteys on havaittavissa myös maidon määrän suhteen.

Mallista laskettujen studentoitujen jäännösten perusteella aineistossa havaittiin 15 poikkeavaa havaintoa. Suurin osa studentoiduista jäännöksistä pysyi valitun |2| rajan sisäpuolella. Studentoitujen jäännösten perusteella aineistossa eniten poikkeavia havaintoja esiintyi Evellä. Yksi mahdollinen syy Eveen liittyvien poikkeavien havaintojen suureen määrään on se, että Evestä oli aineistossa eniten havaintoja. Tällöin myös poikkeavuuksia mahtuu enemmän. Taulukosta 4.4 käy ilmi, että suurin osa poikkeavista havainnoista on esiintynyt vanhemmalla iällä. Tätä väitettä tukevat myös mallin analyysissä saadut tulokset tilastollisesti merkitsevistä muuttujista.

Tehdyn sekamallin perusteella voidaan siis todeta, että keskilämpötilalla, maidontuotannon määrällä tai rodulla ei ole tilastollisesti merkitsevää vaikutusta maidon solutasoon. Vanhemmissa ikäryhmissä havaittiin tilastollinen merkitsevyys, mikä viittaa siihen, että iän myötä myös solutaso saa korkeampia arvoja. Mallin avulla löydettiin aineistosta poikkeavia havaintoja. Iän lisäksi ei havaittu muita tekijöitä, joilla olisi merkittävää vaikutusta solutasoon. Mallia olisi syytä kehittää tai aineistoon kerätä lisää havaintoja, jotta solutasoon vaikuttavia tekijöitä voitaisiin tunnistaa tarkemmin.

Lähteet

- Aho, J., Koponen, M., Pasto, M-P. & Stalder, S. (2020). *Monipuolinen elintarvikeala: elintarvikkeiden valmistus ja tuotanto*. Opetushallitus, luku 4, Oppikirjan lisämateriaali. <https://www.oph.fi/fi/oppimateriaali/monipuolinen-elintarvikeala/4-maidon-jalostus>, viitattu 27.1.2025.
- Demidenko, E. (2013). *Mixed Models: Theory and Applications with R: Second Edition*, Wiley.
- Galecki, A. & Burzykowski, T. (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. 1st ed., Springer Nature.
- Gumedze, F. N. & T. T. Dunne. (2011). “Parameter Estimation and Inference in the Linear Mixed Model.” *Linear Algebra and Its Applications*, vol. 435, no. 8, pp. 1920–44. <https://doi.org/10.1016/j.laa.2011.04.015>.
- Ilmatieteenlaitos. Havaintojen lataus. (n.d.). Viitattu 22.10.2024. <https://www.ilmatieteenlaitos.fi/havaintojen-lataus>.
- Isokallio, J. (2015). *Utaretulehduksen diagnostiikka: maitonäytteenottomenetelmän vaikutus bakteriologisiin tuloksiin*. Helsingin yliopisto.
- Kervinen, E. & Kervinen, O. (2020). *Karjakohtaisen utareterveysraportoinnin kehittäminen*. Oulun ammattikorkeakoulu.
- Pinheiro, J. C. & Bates, D.M. (2000). *Mixed-Effects Models in S and S-Plus*. Springer.
- Robinson, A. P. & Hamann, J.D. (2011). *Forest Analytics with R: An Introduction*. 1st ed., Springer Nature.
- Robinson, G. K. (1991). “That BLUP Is a Good Thing: The Estimation of Random Effects.” *Statistical Science*, vol. 6, no. 1, pp. 15–32. <https://www.jstor.org/stable/2245695>.
- Schützenmeister, A. & Piepho, H.-P. (2012). “Residual Analysis of Linear Mixed Models Using a Simulation Approach.” *Computational Statistics & Data Analysis*, vol. 56, no. 6, pp. 1405–1416. <https://doi.org/10.1016/j.csda.2011.11.006>.

West, B. T., Welch, K. B. & Galecki, A. T. (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software*. 2nd ed., CRC Press.