STANDARDIKÄYRÄN SOVITUS AIKAEROTTEISESSA FLUORESENSSI-IMMUNOMÄÄRITYKSESSÄ JA PASSING–BABLOK -MENETELMÄVERTAILU

Valto Kuusisto

Pro gradu -tutkielma
28. kesäkuuta 2021

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

KUUSISTO, VALTO: STANDARDIKÄYRÄN SOVITUS AIKAEROTTEISESSA FLUORESENSSI-IMMUNOMÄÄRITYKSESSÄ JA PASSING–BABLOK -MENETELMÄVERTAILU
Pro gradu -tutkielma
Tilastotiede
28. kesäkuuta 2021

Aikaerotteisessa fluoresenssi-immunomäärityksessä on usein tavoitteena selvittää tietyn proteiinin pitoisuus testiliuoksesta. Pitoisuuden selvittämiseen käytetään ennalta määriteltyä standardikäyrää, joka muuntaa mitatun vasteen arvon tietyn proteiinin pitoisuudeksi. Standardikäyrän määrittämiseksi käytettävä asetelma pitää suunnitella huolellisesti, mukaan lukien aineiston kerääminen ja parametrien estimointiin käytettävät menetelmät.

Tässä työssä standardikäyrän määrittämisessä käytetään simuloitua aineistoa, joka vastaa todellisia mittaustuloksia. Ennen mallien sovittamista sekä vasteet että pitoisuudet muunnetaan tasavälisemmiksi logaritmimuunnoksella. Standardikäyrä on usein epälineaarista muotoa ja etenkin sigmoidikäyrän muotoiset epälineaariset mallit sopivat standardikäyräksi. Lisäksi tarkastellaan silotetun splinimallin soveltamista standardikäyräksi. Muunnettujen vasteiden variaatio ei ole homogeenista, jolloin mallien parametrien estimointi suoritetaan painotetun pienimmän neliösumman menetelmällä. Mallien välistä paremmuutta vertaillaan Passing–Bablok menetelmävertailulla, jossa jokaista eri menetelmällä sovitettua standardikäyrää verrataan referenssimenetelmään.

Standardikäyrän sovittamisessa lineaarinen malli ei kykene kuvaamaan tarkasti proteiinin pitoisuutta, kun taas epälineaariset mallit pystyvät. Sigmoidikäyrän muotoisista malleista symmetriset sigmoidimallit eivät myöskään kykene ennustamaan pitoisuuksia tarkasti, vaan epäsymmetriset sigmoidikäyrän muotoiset mallit tai silotettu splinimalli ennustavat tarkemmin proteiinin pitoisuutta. Passing–Bablok -menetelmävertailun sekä keskimääräistä ennustevirhettä tarkasteltaessa sigmoidikäyrän muotoiset mallit suoriutuivat silotettua splinimallia heikommin.

Silotettu splinimalli tuottaa pienemmän keskimääräisen ennustevirheen verrattuna sigmoidikäyrän muotoisiin malleihin. Lisäksi otoskoon kasvaessa keskimääräisen ennustevirheen suuruus pienenee ja sovitettu malli tulee tarkemmaksi. Passing–Bablok -menetelmävertailun perusteella mallien ennustamien pitoisuuksien ja todellisten pitoisuuksien välinen suhde ei ole lineaarista. Otoskoon ollessa suuri saattaa lineaarisuustarkastelun tulos olla virheellinen sekä hypoteesien testausmenetelmät olla liiankin kireitä. Silotettu splinimalli soveltuu parhaiten standardikäyräksi, mutta muita tarkasteluita on syytä tehdä ennen standardikäyrän käyttöönottoa.

Asiasanat: Standardikäyrä, silotettu splini, logistinen regressiomalli, painotettu pienimmän neliösumman menetelmä, Passing–Bablok -menetelmävertailu

Sisällys

1 Johdanto

Tässä työssä tarkastellaan ns. biologisessa määrityksessä käytettäviä tilastollisia menetelmiä, erityisesti proteiinien pitoisuuksien arviointiin käytetyn standardikäyrän sovittamista aikaerotteisessa fluoresenssi-immunomäärityksessä. Aikaerotteisen fluorerenssi-immunomäärityksen avulla tutkija kykenee määrittämään testiliuoksesta tietyn proteiinin pitoisuuden tai lääkäri pystyy määrityttämään potilaan verinäytteestä halutun antigeenin pitoisuuden, kuten sydäninfarktin-, raskaushormonin- tai koronaviruksen antigeenin pitoisuuden. Standardikäyrän avulla saadaan tarkasti määritettyä pitoisuus, mikäli standardikäyrä on luotu huolellisesti. Epätarkasti määritetty standardikäyrä voi johtaa virheellisesti arvioituihin antigeenin pitoisuulukemiin, jolloin tutkija saattaa ajautua vääriin johtopäätöksiin tai lääkäri saattaa tehdä virheellisiä tulkintoja potilaan terveydentilasta.

Luvussa 2 esitellään aikaerotteisen fluoresenssi-immunomäärityksen perusidea, jolloin selvitetään kyseiseen määritykseen liittyviä oletuksia. Lisäksi luvussa esitellään biologisen määrityksen sellaisia erityispiirteitä, joita tulee ottaa mallintaessa huomioon. Luvussa kerrotaan standardikäyrän tarkoitus sekä selvitetään, millainen on standardikäyrän luomisessa käytetty otosaineisto.

Luvussa 3 tarkastellaan biologisen määrityksen mallintamista. Aluksi selitetään yksinkertaisimman lineaarisen mallin sovittaminen aineistoon. Samalla esitellään pienimmän neliösumman menetelmä lineaarisen mallin tapauksessa, kun voidaan olettaa havaintojen olevan normaalisti jakautuneita. Pienimmän neliösumman menetelmän ratkaisemiseen esitellään myös numeerinen Gauss–Newton -menetelmä.

Neljännessä luvussa lineaarinen malli yleistetään epälineaarisiin parametrisiin malleihin. Epälineaarisista malleista esitellään sigmoidikäyrän muotoiset mallit, neli- ja viisiparametrinen logistinen regressiomalli sekä keskustellaan mallien erityispiirteistä. Samalla tarkastellaan pienimmän neliösumman menetelmän toimivuutta epälineaarisen mallin tapauksessa. Luvussa tutkitaan heterogeenisen aineiston parametrien estimointia painotetulla pienimmän neliösumman menetelmällä.

Viidennessä luvussa tarkastellaan epäparametrista mallintamista biologisessa määrityksessä. Luvussa tutkitaan silotetun splinimallin sovittamista sekä splinimallin solmukohtien asettamista havaintoaineistoon. Silotetun splinimallin sakkoparametrin sopiva arvo optimoidaan ristiinvalidointimenetelmällä, josta esitellään kaksi tapaa.

Kuudennessa luvussa esitellään Passing–Bablok -regressiomenetelmä sekä sovelletaan tätä mentelmää biologisen aineiston pitoisuuksien vertailuun. Passing–Bablok -regressiomenetelmässä estimoidaan parametrit, tarkastellaan menetelmien lineaarisuutta sekä tutkitaan kahden menetelmän samankaltaisuus.

Seitsemännessä luvussa tarkastellaan ja sovelletaan luvuissa 2−6 esiteltyjä menetelmiä simulointiaineistoon. Simulointiaineiston avulla saadaan myös vertailtua eri mallien soveltuvuutta generoituun aineistoon sekä suoritettua menetelmävertailua Passing–Bablok -regressiomenetelmällä

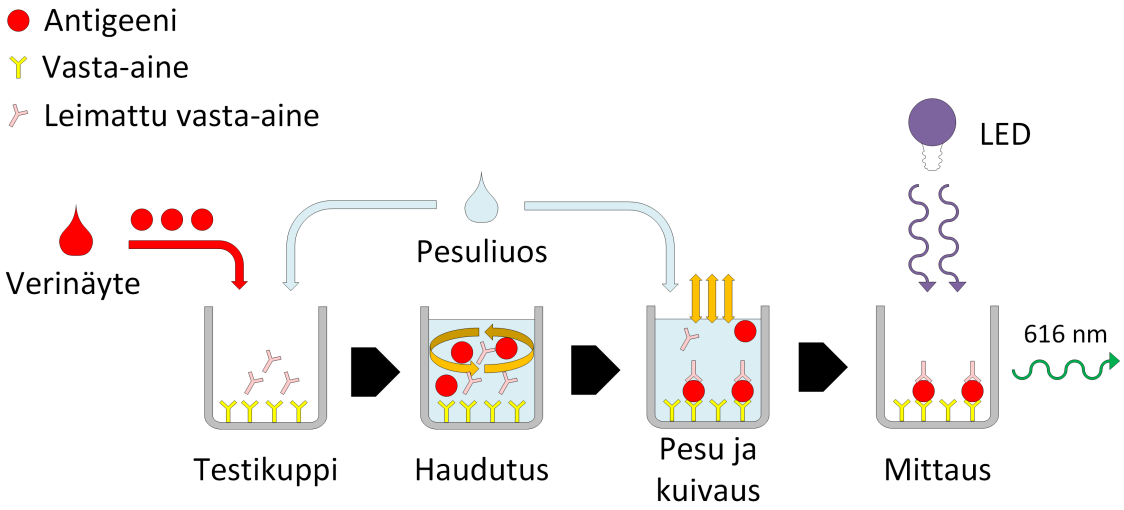
Työssä on käytetty pääosin D.J. Finneyn lähdeteosta [2], johon myös monet muut tämän työn lähdeteoksista viittaavat. Kappaleen 3 alkupuolella taas on seurattu T.V. Mzolon artikkelia [1] ja täydennetty esitystä sopivilta osin H. Huangin [4] ja D.J. Finneyn toisella teoksella [3]. Epälineaaristen menetelmien sovittamisessa on

seurattu P. Gotschalk et. al. [5] sekä H. Huangin [4] artikkeleita. Epäparametristen menetelmien yhteydessä on käytetty lähdemateriaaleina T. Hastien ja R. Tibshiranin teoksia [8] sekä [9], joita on täydennetty enemmän biologiseen määritykseen soveltavalla lähteellä [10]. Passing–Bablok -regressiomenetelmän lähteinä on käytetty H. Passingin ja W. Bablokin teosta [11] sekä L. Bilic-Zullen menetelmävertailuartikkelia [12].

2 Biologisesta määrityksestä

Tieteellistä koetta, jossa pyritään arvioimaan proteiinien biologista aktiivisuutta kutsutaan *biologiseksi määritykseksi* [1]. Biologisessa määrityksessä ja etenkin kemian tuotannossa on usein tavoitteena määrittää tiettyjen proteiinien pitoisuus liuoksesta. Liuoksena toimii tyypillisesti verinäyte, johon liukenee eri sairauksien tai lääkkeiden vaikutuksesta proteiineja. Näiden proteiinien määrittämiseen on kehitetty useita menetelmiä, joista tässä työssä perehdytään *aikaerotteisen fluoresenssi-immunomäärityksen* (time resolved fluorescence immunoassay, TR-FIA) mallintamiseen.

Liuoksesta selvitetyn pitoisuuden perusteella tutkija kykenee tekemään päätelmiä kemiallisesta reaktiosta tai vastaavasti lääkäri kykenee tekemään arvioita potilaan terveydentilasta. Väärin määritetty pitoisuus saattaa johtaa tutkijaa harhaan tai vastaavasti lääkäriä ohjaamaan potilaan vääriin jatkotutkimuksiin. Tässä työssä tarkastellaan ensin yleisellä tasolla ns. standardikäyrän sovittamista biologisessa määrityksessä. Erityisesti selvitetään standardikäyrän sovittamista aikaerotteiseen fluoresenssi-immunomääritykseen simulointiaineiston avulla.

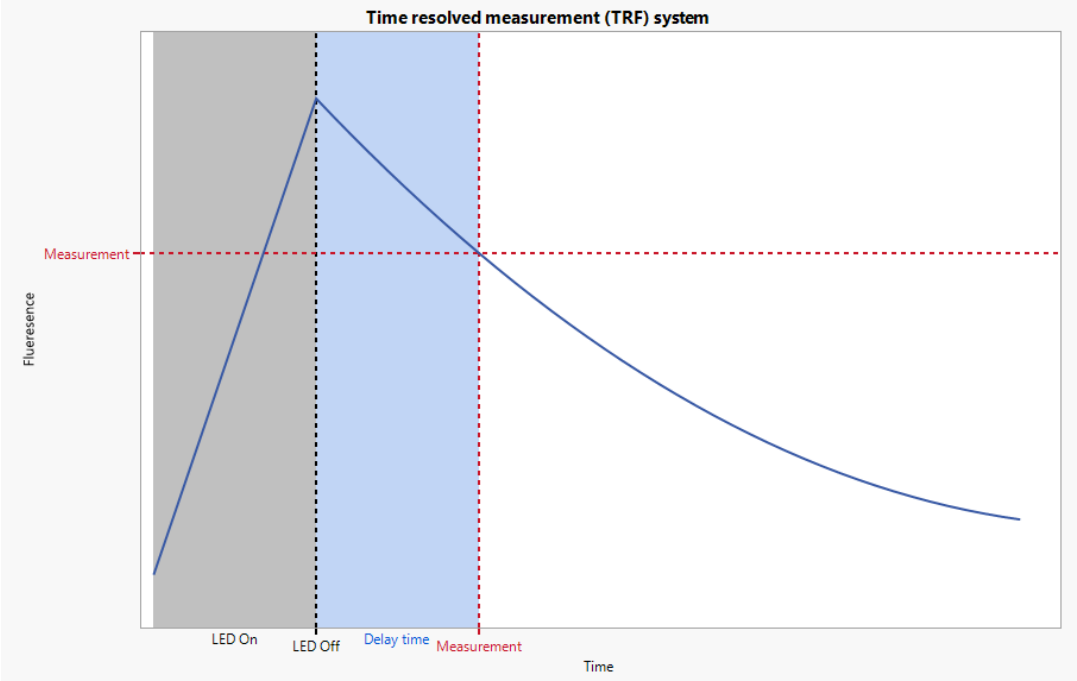


Kuva 1: Aikaerotteisen fluoresenssi-immunomäärityksen reaktiokuvaus.

Tilastollisesta näkökulmasta tarkasteltuna biologinen määritys on ennalta suunniteltu koeasetelma, jossa tutkitaan yhden tai useamman ärsykkeen biologista aktiivisuutta näytteessä [3]. Tässä ärsyke on yleinen nimitys lääkkeelle, myrkylle tai näytteelle, jota voidaan lisätä liuokseen tutkijan määrittämillä pitoisuuksilla. Tässä työssä liuos on verinäyte ja standardierä on verta vastaava liuos, johon on lisätty tunnettu määrä antigeenejä. Kun antigeenien pitoisuus tunnetaan stadardierässä, saadaan määritettyä ns. standardikäyrä kyseiselle standardiliuoserälle tarkasti. Standardikäyrä on siis ärsykkeen vaste kuvattuna pitoisuuden funktiona, jonka avulla saadaan arvioitua myöhemmin saadun ärsykkeen vasteesta pitoisuus. Standardikäyrä luodaan laitteen valmistajalla, joka toimitetaan tutkijalle tai lääkärille laitteen yhteydessä. Tällöin tutkija tai lääkäri ei itse sovita ja luo standardikäyrää, vaan tämän tekee asiaan perehtynyt asiantuntija.

2.1 Aikaerotteinen fluoresenssi-immunomääritys

Tässä tutkimuksessa tarkasteltu proteiini on jokin antigeeni, jonka pitoisuutta määritetään aikaerotteisella fluoresenssi-immunomäärityksellä. Tähän käytetään testialustaan kiinnitettyä vasta-ainetta sekä fluoresoivaa leimattua vasta-ainetta. Aikaerotteisen fluoresenssi-immunomäärityksen prosessi havainnollistetaan kuvassa 1. Testialustaan lisätään näyte, joka pestään hauduttamisen jälkeen pesuliuoksella. Haudutuksen aikana vasta-aine muodostaa näytteessä kiinnostuksen kohteena olevien antigeeni- tai proteiinimolekyylien kanssa sidoksia, johon liittyy vielä leimattu vasta-aine. Nämä vasta-aine-antigeenikompleksit jäävät testialustaan kiinni, kun muu sitoutumaton materiaali pestään pois testialustasta. Testialustaa valaistaan vielä fluoresoivalla valolla, jolloin valon avulla saadaan määritettyä leimattujen vasta-aine-antigeenisidosten *signaalien* lukumäärä. Mitä enemmän antigeeniä on näytteessä, sitä enemmän myös sidoksia syntyy ja näin ollen määrityksessä saadaan suurempi signaalivasteen lukumäärä. Tämä oletus on voimassa kohtuullisilla antigeenipitoisuuksilla, koska liuoksen *saturoitumisen* jälkeen antigeeni-vasta-ainesidoksia ei enää kykene syntymään, kun kaikki vasta-ainemolekyylien sitoutumispaikat on käytetty.



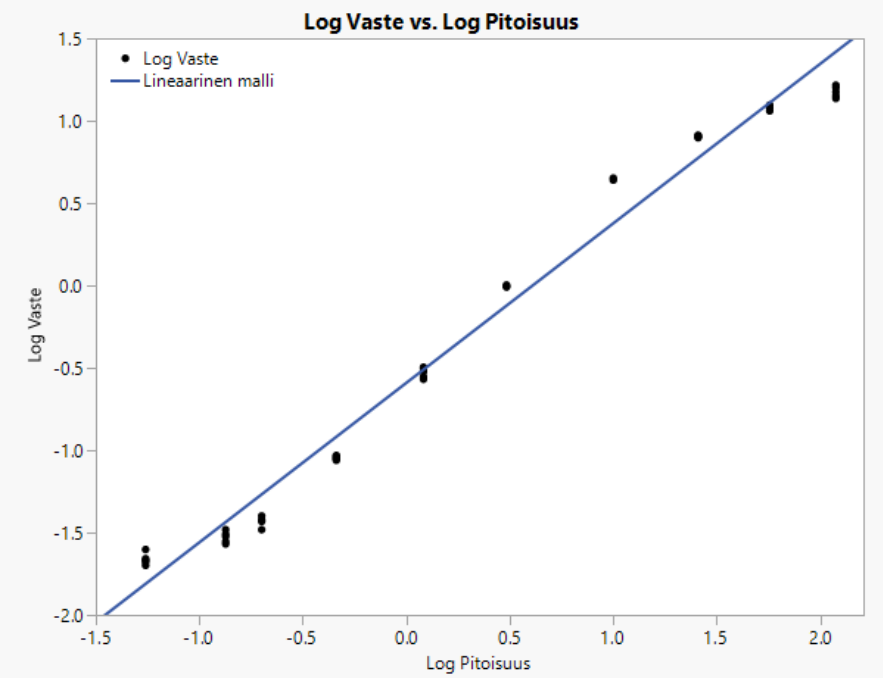
Kuva 2: Aikaerotteisen määrityksen taustaidea. Kun LED-valo on päällä, kasvaa fluoresenssin signaali. LED-valon sammuttua flueresenssin signaali laskee, kun antigeeni-vasta-ainemolekyylit fluoresoivat valoa. Luotettavan testituloksen saamiseksi viivästetään mittausta LED-valon sammuttamisesta.

Näytettä valaistaan LED-valolla, jolloin leimatut vasta-aineet keräävät LED-valon energiaa itseensä ja heijastavat tätä valoenergiaa ulospäin, eli vasta-aineet *fluoresoivat* valoa. LED-valoa pidetään päällä, kunnes vasta-aineet ovat keränneet tarpeeksi energiaa itseensä, minkä jälkeen valo sammutetaan. Mittausta ei voida kuitenkaan aloittaa heti valon sammuttamisesta, koska tällöin saataisiin mahdollisesti virheellinen mittaustulos. Valo myös fluoresoi ympärillä olevia materiaaleja, joten

mittauksen aloitusta viivästetään hieman ympäristön vaikutuksen tasaantumiseksi. Tällöin fluoresoivien vasta-aineiden lukumäärä saadaan laskettua tarkemmin ilman ympäristön tuomaa mittausvirhettä. Tätä prosessia on havainnollistettu kuvassa 2, joka esittää fluoresenssin määrää ajan funktiona.

Mittauslaitteen laskeman signaalivasteen lukumäärän lukemisen jälkeen tutkijan tai lääkärin ongelmaksi tulisi se, että signaalien lukumäärää ei itsessään voida pitää pitoisuutena. Tästä syystä antigeenin pitoisuus näytteessä arvioidaan laitteelle syötetyllä *standardikäyrällä*. Standardikäyrä määritetään aina yhdelle standardiliuoserälle kerrallaan, koska eri standardiliuoserien välillä saattaa olla eroavaisuuksia ja näin saadaan tarkempi määritys aikaiseksi. Standardikäyrää voidaan kutsua myös kalibrointikäyräksi tai vaihtoehtoisesti voidaan puhua myös standardisuorasta, mikäli standardikäyrä on lineaarinen. Tältä standardikäyrältä laite kykenee kääntämään signaalivasteesta vastaavan antigeenin pitoisuuden arvion. Standardikäyrän on tarkoitus kuvata mahdollisimman tarkasti vasta-aine-antigeenikompleksien lukumäärää eri antigeenien pitoisuuksilla. Tämän vuoksi epätarkasti määritetty standardikäyrä tuottaa epätarkan pitoisuuden arvon todelliseen pitoisuuteen verrattuna.

2.2 Standardikäyrä

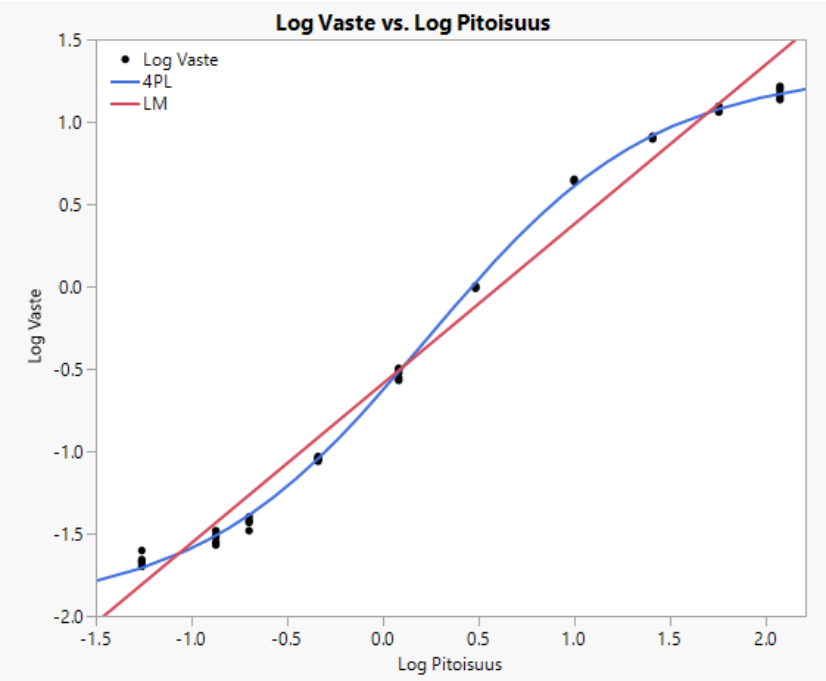


Kuva 3: Illustraatio lineaarisen mallin sovittamisesta signaalivasteisiin.

Standardikäyrän estimoinnin taustalla on käsite annosvasteen regressiofunktiosta. Oletetaan, että y_i on mitattu realisoitunut signaalivaste satunnaismuuttujasta Y_i standardierän S mittausta vastaavalla pitoisuudella x_i. Tämä voidaan kuvata regressiofunktiolla $F(\cdot)$:

$$y_i = F(x_i) + \varepsilon_i,\ \varepsilon_j \sim N(0, \sigma^2). \tag{1}$$

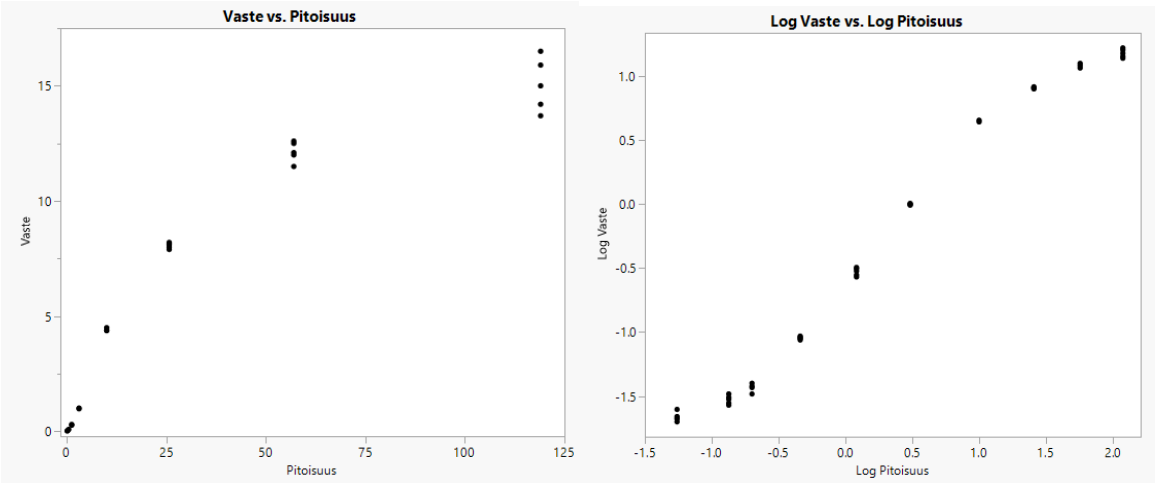
Standardikäyrän määritys käsittää useita erillisiä mittauksia kullakin standardierän S pitoisuudella x_i. Samalla pitoisuudella tehtyjä mittauksia kutsutaan toistomittauksiksi. Tavallisesti tällaisia toistomittauksia tehdään vähintään kolme [3] jokaisesta erän S pitoisuudesta x_i luotettavamman tuloksen saamiseksi. Yksinkertaisimpaan ja vähiten oletuksia vaativaan standardikäyrän määritykseen mitataan toistomittauksia vain kahdelta eri pitoisuudelta x_1 ja x_2. Tällöin valmistetun tuote-erän S standardikäyrä on lineaarinen suora kahden mittauspisteen välillä, kun taas todellisuudessa käyrän muoto saattaisikin olla epälineaarista (kuva 3). Kuvasta 4 sen sijaan voidaan nähdä, ettei tämä kuitenkaan aina kuvaa ilmiön todellista luonnetta, ja regressiosuoran käyttäminen standardikäyränä saattaa johtaa harhaiseen antigeenin pitoisuuden estimointiin.



Kuva 4: Annosvaste -ilmiön kuvaaja.Havaintoaineistoon on sovitettu neliparametrinen logistinen regressiofunktio sekä lineaarinen regressiomalli. Neliparametrinen logistinen regressio kykenee selittämään hyvin ilmiön todellisen luonteen. Sen sijaan lineaarinen malli ei kykene selittämään ilmiön S-mallista käyttäytymistä. Lineaarisella mallilla voitaisiin selittää ilmiön käyttäytymistä välillä (-0.5, 1.0), jossa muutos on hyvinkin lineaarista.

Kuvista 5 ja 4 voidaan myös havaita, että kun sekä pitoisuus että signaalivaste muutetaan logaritmiseen asteikkoon, saadaan ilmiö kuvattua selkeämmin verrattuna lineaariseen asteikkoon. Lineaarisessa asteikossa huomataan, että pienillä pitoisuuksilla olevat arvot näkyvät kuvaajassa yhtenä havaintopisteenä. Tämän lisäksi tasavälisellä asteikolla suuremmilla pitoisuuksilla on suuri vipuvoima. Log-log -asteikkoon muuntamalla saadaan havaintopisteiden horisontaaliset välit tasaisemmiksi ja näin ollen myös kolmen pienen pitoisuuden pisteet tulevat näkyviin ja paremmin mallinnettaviksi. Lisäksi sopivan mallin löytäminen lineaarisen asteikon havainnoille voisi olla ongelmallista tai pahimmassa tapauksessa saatettaisiin sovittaa kokonaan vää-

ränlainen mallityyppi aineistoon.



(a) Lineaarisilla asteikoilla (b) Log-Log asteikoilla

Kuva 5: Kuvassa (a) on kuvattu annosvaste -ilmö lineaarisilla asteikoilla, jolloin kolme alinta pitoisuuspisteen mittausta näkyvät kuvassa yhtenä pisteenä. Muuntamalla molemmat muuttujat logaritmiseen asteikkoon, saadaan tasavälisempi ja helpommin mallinnettava annosvaste -ilmiö.

2.3 Aineiston määrittäminen

Standardikäyrän määrityksessä havaintoaineiston pitoisuuspisteiden paikat tulee määrittää huolellisesti estimointivirheen minimoimiseksi ja tarkan standardikäyrän luomiseksi. Havaintoaineiston pitoisuuspisteiden paikat otetaan tässä työssä annettuina. Ennen havaintoaineiston mittaamista selvitetään, millä pitoisuusvälillä standardikäyrän tulee olla määritetty ja tarkka. Tältä pitoisuusväliltä mitataan havaintoaineiston signaalivasteet y_i eri standardipitoisuuksilla x_i, jotta standardikäyrä saadaan määritettyä tälle välille sopivaksi. Välin ulkopuolelle jäävällä alueella käyrän arvot joudutaan ekstrapoloimaan. Pitoisuusvälin ulkopuolella pitoisuudet ovat kuitenkin joko todella matalia tai korkeita, jolloin standardikäyrän estimoiman pitoisuuden harha ei ole joko tutkimuksen lopputuloksen tai potilaan terveyden tilan kannalta merkittävää.

Mittausvälin määrittämisen jälkeen pitoisuusvälille sijoitetaan n_c kappaletta pitoisuuspisteitä, joilta kultakin määritetään k kappaletta toistomittauksia luotettavamman mittaustuloksen takaamiseksi. Pisteet on jaoteltu välille niin, että ne ovat tasavälisesti logaritmisella asteikolla, jolloin suurien pitoisuuksien signaalivasteilla ei ole vipuvoimaa pienien pitoisuuksien signaalivasteisiin verrattuna. Vipuvoimalla tarkoitetaan, että suurien pitoisuuksien mahdolliset virheet vaikuttaisivat enemmän standardikäyrän sovitukseen ja estimointitulokseen kuin pienempien pitoisuuksien signaalivasteiden arvot. Otoskooksi saadaan $n = n_c \times k$ mittausta ja tällöin saadaan mitattua signaalivasteet y_i, $i = 1, \ldots, n$.

Standardikäyrän määrityksessä käytetään d mittauslaitetta mittausepävarmuuden minimoimiseksi ja suuremman otoskoon tuottamiseksi. Tällöin saadaan otosko-

koa kasvatettua d-kertaiseksi yhden laitteen mittauksiin verrattuna. Tällä on myös käytännön ajankäytöllinen etu, koska useammalla laitteella kyetään mittaamaan d mittausta rinnakkain, kun yhdellä laitteella saadaan samassa ajassa vain yksi mittaus. Lisäksi saadaan useammalta laitteelta mittauksia, jolloin yhden laitteen mahdollisesti väärin mittaamat arvot saadaan suljettua pois määrityksestä. Tässä työssä valitaan laitteiden lukumääräksi $d = 6$. Laitteen lukumäärän ollessa $d = 6$ saadaan siis jokaiselle pitoisuudelle toistomittausten lukumääräksi $6k$. Aineiston määrittämisen ja signaalivasteiden mittauksien jälkeen sovitetaan standardikäyrä aineistoon.

3 Lineaarinen malli biologisessa määrityksessä

Varhaisissa eläinkokeissa biologiset vasteet olivat usein binäärisiä, kuten kuolema vs. selviytyminen tai lääke auttoi vs. lääke ei auttanut hoidossa. Näiden binääristen vasteiden mallintamiseen kehitettiin 1900-luvun alkupuolella erilaisia normaalisia sigmoidikäyrän muotoisia logistisia malleja log-log asteikolle (probit-mallit). Tässä työssä vasteet ovat kuitenkin kvantitatiivisia eli jatkuvia ja määrällisiä. Niiden mallintamiseen käytettiin 1900-luvun alkupuolella lineaarista regressiomallia. Lineaarinen malli saatiin sovitettua, kun ilmiö linearisoitiin muuntamalla pitoisuudet ja vasteet logaritmisille asteikoille.

Oletetaan koeasetelma, jossa tehdään n:n suuruinen otos toisistaan riippumattomia mittauksia. Erän mittausvälille on asetettu ennalta määrätty määrä pitoisuuspisteitä, joista mitataan toistomittauksia. Tällöin saadaan signaalivasteet y_i pitoisuudella $x = (c_1, \ldots, c_1, c_2, \ldots, c_2, \ldots, c_{n_c})$, missä kukin pitoisuustaso toistuu $6k$ kertaa. Tällöin voidaan havaintoaineistoon sovittaa lineaarinen regressiomalli,

$$y_i = F(x_i; \alpha, \beta) + \varepsilon_i = \alpha + \beta \log(x_i) + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2), \tag{2}$$

jossa α on vakiotermi, β on kulmakerroin ja ε_i virhetermi, jonka oletetaan olevan toisista virhetermeistä riippumaton ja normaalisti jakautunut nollaodotusarvolla ja vakiovarianssilla σ^2.

Mallin parametrien estimoimiseksi lineaarinen malli (2) sovitetaan aineistoon *pienimmän neliösumman menetelmällä* (ordinary least squares). Pienimmän neliösumman menetelmää voidaan käyttää, koska pienimmän neliösumman menetelmä on ekvivalentti suurimman uskottavuuden menetelmän kanssa, kun vasteiden y_i jakaumat jokaisella pitoisuuspisteen x_i arvolla ovat normaalisia [7]. Pienimmän neliösumman menetelmässä pyritään löytämään parametrien $\hat{\alpha}$ ja $\hat{\beta}$ arvot, jotka minimoivat neliösummalausekkeen todellisten vasteiden ja mallin $F(\cdot)$ antamien ennusteiden välillä:

$$\sum_{i=1}^{n} (y_i - F(x_i; \alpha, \beta))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,\ i = 1, \ldots, n, \tag{3}$$

jossa n on havaintojen lukumäärä ja $F(x_i; \alpha, \beta) = \hat{y}_i$ mallin mukainen vasteen odotusarvo pisteessä x_i.

3.1 Parametrien estimointi lineaarisessa mallissa

Lineaarisen mallin tapauksessa estimointi tapahtuu pienimmän neliösuman menetelmällä logaritmisille signaalivaste-pitoisuuspareille (x_i, y_i), $i = 1, \ldots, n$. Tällöin empiirinen malli saadaan kirjoitettua aiemmin mainittuun muotoon (2). Merkitsemällä $F(\mathbf{X}; \boldsymbol{\theta}) = (F(x_1; \theta), F(x_2; \theta), \ldots F(x_n; \theta))$ saadaan malli kirjoitettua matriisimuodossa:

$$\mathbf{Y} = F(\mathbf{X}; \boldsymbol{\theta}) + \boldsymbol{\varepsilon}, \tag{4}$$

jossa $\boldsymbol{\theta}$ viittaa yleisesti p-ulotteiseen parametrivektoriin $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$. Parametrivektorin $\boldsymbol{\theta}$ parametriavaruus $\boldsymbol{\Theta}$ on osajoukko p-ulotteisesta reaaliavaruudesta

$\boldsymbol{\Theta} \subset \mathbb{R}^p$ ja

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}',$$

jossa merkintä x' viittaa matriisin transponointiin. Lisäksi

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}, \ \boldsymbol{\varepsilon}_1 = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}.$$

Merkitään lisäksi $F(\boldsymbol{\theta}) = F(\mathbf{X}; \boldsymbol{\theta})$ merkintöjen selventämiseksi.

Pienimmän neliösumman menetelmän estimaattoria parametrivektorille $\boldsymbol{\theta}$ merkitään $\hat{\boldsymbol{\theta}}$, joka on piste parametriavaruudessa. Tällä pisteellä funktion $F(\hat{\boldsymbol{\theta}})$ arvo on lähimpänä havaittua arvoa $\mathbf{y}$. Pienimmän neliösumman estimaattori saadaan johdettua minimoimalla *residuaalineliösumma (RSS)*:

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - F(x_i; \boldsymbol{\theta}))^2, \ \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p. \tag{5}$$

Oletetaan, että funktio $F(\boldsymbol{\theta})$ on differentioituva parametrin $\boldsymbol{\theta}$ suhteen. Etsitään parametrien pienimmän neliösumman estimaatit $\hat{\theta}$ seuraavan yhtälöryhmän ratkaisuna:

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_v} = 0, \ v = 1, \cdots, p.$$

Tästä yhtälöryhmästä saadaan johdettua *normaaliyhtälöt*

$$\sum_{i=1}^{n} \frac{\partial F(x_i; \boldsymbol{\theta})}{\partial \theta_v} (y_i - F(x_i; \boldsymbol{\theta})) = 0, \ v = 1, \cdots, p. \tag{6}$$

Kirjoittamalla $\mathbf{V}_{i,v}(\boldsymbol{\theta}) = \partial F(x_i; \boldsymbol{\theta}) / \partial \theta_v$ ja $\boldsymbol{\varepsilon} = \mathbf{y} - F(\boldsymbol{\theta})$ saadaan normaaliyhtälöt kirjoitettua matriisimuodossa

$$\mathbf{V}(\boldsymbol{\theta})' \boldsymbol{\varepsilon} = 0, \tag{7}$$

jossa matriisia $\mathbf{V}$ kutsutaan *vauhtimatriisiksi (velocity matrix)*. Lineaarisessa regressiomallissa matriisia $\mathbf{V}$ kutsutaan *mallimatriisiksi* (design matrix), jonka arvot ovat riippumattomia parametrien arvoista $\boldsymbol{\theta}$. Normaaliyhtälöt kyetään lineaarisen mallin tapauksessa ratkaisemaan, mutta epälineaaristen regressiofunktioiden kohdalla tarvitaan numeerisia ratkaisumenetelmiä. Yksi näistä on *Gauss–Newton* -menetelmä, joka etsii askeltavasti parhaimman estimaatin $\hat{\boldsymbol{\theta}}$ arvon.

3.2 Gauss–Newton -menetelmä

Normaaliyhtälöitä ratkaistaessa ajaudutaan usein tilanteeseen, jossa ei analyyttistä ratkaisua löydetä vaan täytyy turvautua numeerisiin menetelmiin. Näihin tilanteisiin ajaudutaan, kun regressiofunktio on epälineaarista muotoa, jolloin tavallinen pienimmän neliösumman estimointi on harhaista. Gauss–Newton -menetelmä

ratkaisee epälineaariset regressiofunktiot askeltavasti hyödyntämällä Taylorin sarjakehitelmää. Sarjakehitelmässä tarkastellaan pientä lähialuetta $\boldsymbol{\theta}^*$ parametrin $\boldsymbol{\theta}$ läheisyydessä, jolloin voidaan lineaarinen Taylorin sarjakehitelmä voidaan kirjoittaa muotoon

$$F(\mathbf{X};\boldsymbol{\theta}) \approx F(\mathbf{X};\boldsymbol{\theta}^*) + \nabla_\theta F(\mathbf{X};\boldsymbol{\theta}^*)'(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{8}$$

jossa

$$\nabla_\theta F(\mathbf{X};\boldsymbol{\theta})' = \left(\frac{\partial F(X;\boldsymbol{\theta}^*)}{\partial \theta_1}, \cdots, \frac{\partial F(X;\boldsymbol{\theta}^*)}{\partial \theta_p} \right).$$

Approksimatiivinen RSS saadaan funktion $F(x_i;\boldsymbol{\theta})$ lineaarisesta approksimaatiosta lähialueen $\boldsymbol{\theta}^*$ ympäristössä, kun

$$S(\boldsymbol{\theta}) \approx \sum_{i=1}^{n} (y_i - F(x_i;\boldsymbol{\theta}^*))^2 - \sum_{v=1}^{p} \mathbf{V}_{i,v}(\boldsymbol{\theta}^*)(\theta_v - \theta_v^*)^2.$$

Approksimaatio on kuitenkin pätevä vain lokaalisti valitun lähialueen $\boldsymbol{\theta}_t$ sisällä kyseiselle parametriestimaatille askeleella t. Lähialueen koko riippuu regressiofunktion kaarevuudesta; mitä kaarevampi (lineaarisempi) funktio, sitä pienemmässä (suuremmassa) lähialueessa lineaarinen approksimaatio pätee. Käyttämällä lineaarista Taylorin sarjakehitelmää askeltavaan päivittämiseen, saadaan vastaavat normaaliyhtälöt (6) kuin lineaarisen pienimmän neliösumman tapauksessa t:nnen iteraatiokierroksen kohdalla:

$$\mathbf{V}(\boldsymbol{\theta}_t)'\mathbf{V}(\boldsymbol{\theta}_t)(\boldsymbol{\theta} - \boldsymbol{\theta}_t) = \mathbf{V}(\boldsymbol{\theta}_t)'(\mathbf{y} - F(\boldsymbol{\theta}_t)). \tag{9}$$

Gauss–Newton -menetelmän askeltava lähestymistapa saadaan yhtälöstä

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \lambda_t \delta_t, \tag{10}$$

jossa λ_t on askeleen koko ja δ_t on päivityksen uuden arvon suunta:

$$\delta = (\mathbf{V}(\boldsymbol{\theta}_t)'\mathbf{V}(\boldsymbol{\theta}_t))^{-1}\mathbf{V}(\boldsymbol{\theta}_t)'(\mathbf{y} - F(\boldsymbol{\theta}_t)).$$

Sopivan askelluksen koon ja suunnan löydyttyä menetelmä siirtyy seuraavaan askellukseen.

Pienimmän neliösumman estimaattorin $\hat{\boldsymbol{\theta}}$ asymptoottinen kovarianssi saadaan kirjoitettua

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{V}(\hat{\boldsymbol{\theta}})'\mathbf{V}(\hat{\boldsymbol{\theta}}))^{-1}. \tag{11}$$

Lineaarisessa mallissa vauhtimatriisi vastaa mallimatriisia, $\mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{X}$, eikä riipu parametrien estimaateista. Lisäksi lineaarisen mallin kovarianssiyhtälö (11) ei tarvitse asymptoottista estimointia, jolloin se on luotettava otoskoosta riippumatta.

Lineaarinen malli kykenee monissa tapauksissa selittämään haluttua ilmiötä tarkasti ja usein logaritmisella muunnoksella saadaan myös epälineaariset tapaukset kuvattua lineaarisella mallilla. Kuitenkin aidosti epälineaarisia tapauksia ilmenee, jolloin lineaarisen mallin antamat estimointitulokset ovat harhaisia. Tällöin tulee pohtia epälineaaristen menetelmien käyttöä ilmiön kuvaamiseen. Usein vasteen ja pitoisuuden välinen suhde ei ole lineaarista vaan vasta logaritmiseen asteikkoon

muuntamalla saadaan suhde lineaariseksi ja näin voidaan käyttää pitoisuuden kertoimen estimointiin samaa mallia kuin kuvattiin yhtälössä (2). Kuten aiemmin jo mainittiinkin luvussa 2.2, vasteen ja pitoisuuden välinen suhde on usein sigmoidikäyrän muotoinen. Tällöin vasteen riippuvuutta konsentraatiosta ei saada logartimisellakaan muunnoksella lineaariseksi, vaan tällöin täytyy käyttää sigmoidifunktiota suhteen kuvaamiseen.

Ilmiön lineaarisuuden lisäksi lineaarinen malli vaatii useita oletuksia annosvaste -ilmiöstä, jotta standardierän kertoimen estimointitulos olisi luotettava ja harhaton. Näitä oletuksia ovat ainakin vasteen ja pitoisuuden välinen lineaarisuus sekä residuaalien normaalisuus ja homogeenisuus yli pitoisuuksien [4]. Yhdenkin oletuksen rikkoutuessa tulee tehdä korjaavia toimenpiteitä, ja etenkin varianssin homoskedastisuusoletus rikkoutuu usein, jolloin tämä tulee huomioida estimoinnissa. Tämän lisäksi myös epälineaarisen mallin tapauksessa estimointitulos voi riippua Gauss–Newton -menetelmälle annetuista alkuarvoista. Mikäli menetelmän alkuarvot on asetettu kauas todellisista arvoista, voi menetelmä pysähtyä mielestään parhaimpaan tulokseen, joka ei kuitenkaan vastaa todellista parasta arvoa.

Virhetermien $\boldsymbol{\varepsilon}$ ollessa normaalisti jakautuneita ja riippumattomia nollaodotusarvolla, pienimmän neliösumman estimaattori on harhaton, normaalisti jakautunut ja sillä on pienin varianssi kaikista *harhattomista lineaarisista estimaattoreista (Best Linear Unbiased Estimator, BLUE)*. Biologisessa määrityksessä epälineaaristen regressiofunktioiden tapauksessa estimaattorilla ei kuitenkaan ole näitä samoja ominaisuuksia, koska epälineaarisen regressiofunktion tapauksessa estimaattori on BLUE vain asymptoottisesti. Kasvattamalla otoskokoa suureksi, tulee Gauss–Newton estimaattori aymptoottisesti numeerisesti vakaaksi. Regressiofunktion ollessa hyvinkin epälineaarinen, jolloin vauhtimatriisin $\mathbf{V}(\boldsymbol{\theta})$ ensimmäisten derivaattojen arvot vaihtelevat rajusti, saattaa pienimmän neliösumman estimaattori olla numeerisesti epävakaa.

4 Epälineaarinen mallintaminen biologisessa määrityksessä

Lineaarisen mallin sovituksessa oletuksena on, että vaste kasvaa aina samalla kulmakertoimella pitoisuudesta riippumatta. Biologisessa määrityksessä liuoksen saturoituminen tai testin rakenne saattavat kuitenkin johtaa siihen, että vasteen arvon kasvu ei ole lineaarista, edes logaritmisella asteikolla. Saturoitumista kuvaa *massavaikutuksen laki (law of mass action)*, jonka mukaan kemiallisen reaktion tapahtuminen (vasta-aineiden ja antigeenin yhdistyminen) on lähes lineaarista ja nopeaa, kunnes reaktio alkaa lähestymään tasapainotilaansa, jolloin reaktio hidastuu. Tätä havainnollistetaan myös kuvassa 4, jossa havaitaan vasteen kasvavan kiihtyvästi, kunnes reaktio lähestyy tasapainotilaansa. Lähellä reaktion tasapainotilaa pitoisuuden kasvattaminen ei enää kasvata vasteen arvoa. Tällaista epälineaarista kasvamista voidaan kuvata erilaisilla sigmoidikäyrän muotoisilla malleilla, joiden yhtälöt ovat yleisesti samaa muotoa kuin yhtälössä (2).

Yleisessä tapauksessa yhtälössä (2) $F(\cdot)$ on jokin monotoninen funktio, $\boldsymbol{\theta}$ parametrivektori ja ε_i mallin virhetermit pitoisuudella x_i. Virhetermien ε_i oletetaan olevan toisistaan riippumattomia, normaalisti jakautuneita nollaodotusarvolla ja heteroskedastisella varianssilla σ_i^2, $N(0, \sigma_i^2)$. Homoskedastisuutta ei voida olettaa, sillä logaritmisen vasteen hajonta pienenee pitoisuuden x_i mukana ja homoskedastisuusoletus saattaisi tuottaa harhaista estimointia. Tähän kuitenkin on löydetty ratkaisu, joka ottaa huomioon heteroskedastisuuden, ja tähän palataan luvussa 4.2.

4.1 Sigmoidikäyrämallit

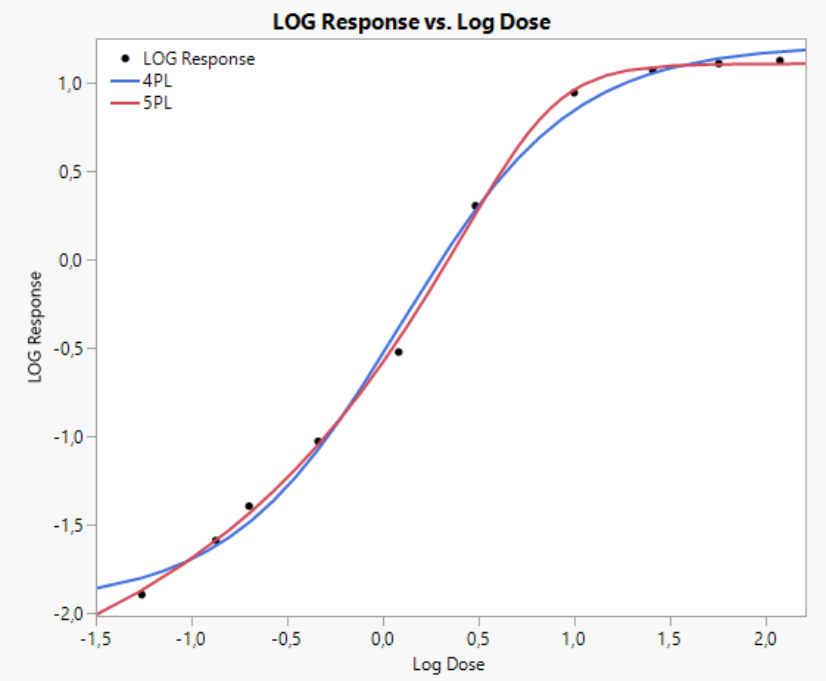
Kaksi perinteistä sigmoidifunktiomallia ovat symmetrinen *neliparametrinen logistinen regressiomalli* sekä laajennettu versio *viisiparametrinen logistinen regressiomalli*. Viisiparametrinen logistinen regressiomalli ottaa huomioon myös mahdollisen epäsymmetrian vasteen ja pitoisuuden välillä. Viisiparametrinen logistinen funktio on muotoa

$$F(x_i; b, c, d, e, f) = c + \frac{d - c}{(1 + \exp(-b(\log(x_i) - \log(e)))^f}, \tag{12}$$

jossa x_i pitoisuus, e on ns. EC50 (half maximal effective concentration) on se pitoisuuden x_i arvo, joka antaa puolet vasteen ylärajan d ja alarajan c väliltä $(d - c)/2$. Alaraja c on funktion alempi asymptootti, joka vastaa keskimääräisen vasteen suuruutta nollapitoisuudella, kun funktio on monotonisesti kasvava. Vastaavasti monotonisesti kasvavalle funktiolle yläraja d kuvaa keskimääräisen vasteen ylempää asymptoottia. Lisäksi parametri b kuvaa kulmakertoimen pisteessä e. Parametri f säätelee funktion epäsymmetristä käyttäytymistä parametrin e ympärillä. Funktiot, jotka ovat muotoa 12 ovat aina joko kasvavasti tai laskevasti monotonisia parametrien b, c ja d arvoista riippuen. Taulukkoon 1 on kerätty eri parametrien vaihtoehtojen vaikutukset siihen, onko funktio kasvava vai vähenevä. Epäsymmetriaparametrin ollessa $f = 1$, viisiparametrinen logistinen regressiomalli supistuu neliparametriseksi logistiseksi malliksi

$$F(x_i; b, c, d, e) = c + \frac{d - c}{1 + \exp(-b(\log(x_i) - \log(e)))}, \tag{13}$$

jossa muut parametrit ovat samat kuin viisiparametrisessa logistisessa regressiomallissa. Kuten aiemmin jo mainittiin, on neliparametrinen logistinen regressiomalli symmetrinen pisteen e suhteen, jolloin funktio kasvaa samalla nopeudella kohti yläarajaa d kuin se laskee kohti alarajaa c pisteen e suhteen. Mikäli alaraja $d = 0$, on neliparametrinen logistinen regressiomalli ekvivalentti *kolmiparametrisen logistisen regressiomallin* kanssa, jota ei käsitellä mainintaa enempää tässä työssä. Neli-



Kuva 6: Viisi- ja neliparametristen logististen regressiomallien sovitukset simuloituun aineistoon, joka on epäsymmetrinen pisteen $(e, \frac{a+d}{2})$ suhteen. Neliparametrinen logistinen regressiomalli ei kykene kuvaamaan ilmiötä yhtä hyvin kuin epäsymmetrinen viisiparametrinen logistinen regressiomalli.

tai viisiparametrisen regressiomallin sovituksen jälkeen halutaan arvioida havaituista signaalivasteista y_i tuntemattomat pitoisuudet x_i, mikä onnistuu ratkaisemalla yhtälö (12) pitoisuuden x_i suhteen

$$\log(x_i) = F^{-1}(y_i; b, c, d, e, f) = \log(e) - \frac{\ln\left(\left(\frac{d-c}{\log(y_i)-c}\right)^{1/f} - 1\right)}{b}, \; i = 1, \ldots, n, \quad (14)$$

ja vastaavasti neliparametriselle logistiselle regressiomallille, kun $f = 1$

$$\log(x_i) = F^{-1}(y_i; b, c, d, e, f) = \log(e) - \frac{\ln\left(\frac{d-c}{\log(y_i)-c} - 1\right)}{b}. \quad (15)$$

Neliparametrinen logistinen malli on symmetrinen funktion keskipisteen $(e, \frac{c+d}{2})$ suhteen, jolloin havaitaan taulukon 1 tapauksien 1 ja 4 olevan ekvivalentit keskenään, kun vain merkitään alarajan c arvo ylärajan d arvoksi ja toisinpäin sekä kun parametri $b \to -b$. Tämän välttämiseksi voidaan joko kiinnittää merkinnöissä $c > d$,

Taulukko 1: Parametrien c, b ja d vaikutukset monotonisten neli- ja viisiparametristen logististen regressiomallien derivaattoihin, eli kulmakertoimiin. Taulukossa $(+)$ tarkoittaa kasvavaa derivaatta- ja $(-)$ vähenevää derivaattafunktiota.

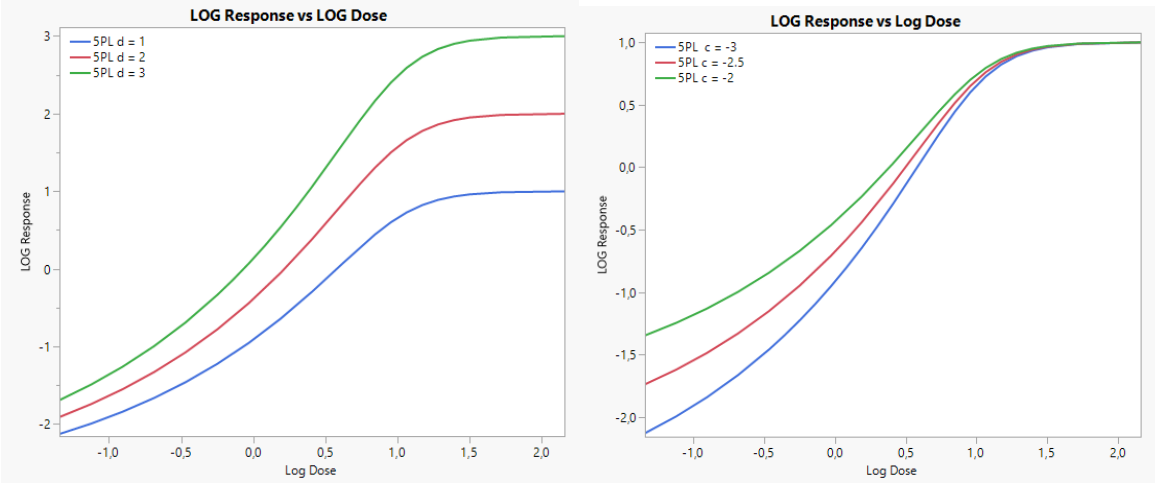
Tapaus	c:n ja d:n järjestys	b:n merkki	Kulmakerroin
1	$c > d$	$b > 0$	-
2	$c > d$	$b < 0$	+
3	$c < d$	$b > 0$	+
4	$c < d$	$b < 0$	-

jolloin parametrin b merkki määrää kulmakertoimen neliparametriselle logistiselle regressiomallille tai voidaan kiinnittää parametri $b > 0$, jolloin alarajan c ja ylärajan d järjestys määrää kulmakertoimen merkin. Kumpikaan näistä merkinnöistä ei vaikuta tai rajoita neliparametrisen logistisen regressiomallin toimintaa.

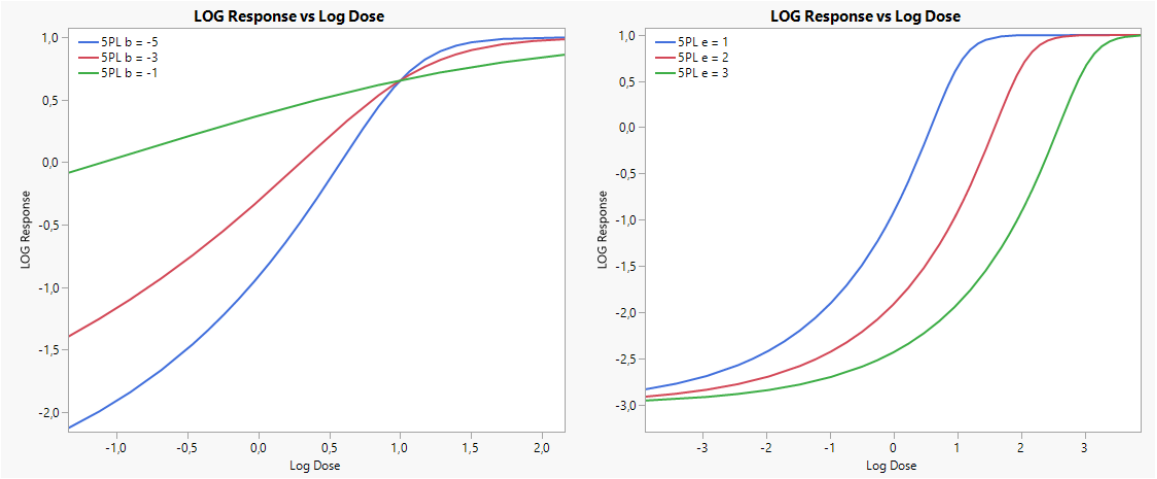
Sen sijaan viisiparametrinen logistinen regressiomalli ($f \neq 1$) ei ole symmetrinen pisteen e suhteen, jolloin taulukon 1 jokainen tapaus kuvaa eri tapausta. Tapaukset 1 ja 4 sopivat kuvaamaan laskevaa annosvastetta, kun taas tapaukset 2 ja 3 sopivat kasvavien annosvastefunktioiden kuvaamiseen. Kuten kuvasta 6 havaitaan, neliparametrinen logistinen regressiomalli ei kykene kuvaamaan epäsymmetrisiä annosvasteilmiöitä yhtä tarkasti kuin viisiparametrinen logistinen regressiomalli. Hyvin pienillä pitoisuuksilla malli näyttäisi yliarvioivan riippuvuutta ja ennustavan liian suuria vasteita näillä pitoisuuksilla. Samanlainen havainto tehdään myös suurilla pitoisuuksilla, joissa ensin pitoisuudella $LOGPitoisuus = 1.0$ neliparametrinen logistinen regressiomalli aliarvioi vasteen arvoa, kun taas pitoisuudesta $LOGPitoisuus = 2.0$ eteenpäin malli yliarvioi todellisen ilmiön vasteen arvoja. 5PL sen sijaan näyttäisi taipuvan hyvin ensin loivemmin kaartuvaan alapäähän ja jyrkemmin kaartuvaan yläpäähän epäsymmetrisen funktiotyypin vuoksi.

Kuvassa 7 esitetään parametrien $b - f$ arvojen vaihtelun vaikutukset viisiparametriseen logistiseen regressiomalliin, kun mallin muut parametrit ovat kiinnitettyinä. Siniset käyrät ovat keskenään ekvivalentit eri kuvaajien välillä. Kuvassa 7 (a) kasvatetaan ylärajan d arvoa, vastaavasti kuvassa 7 (b) kasvatetaan alarajan c arvoa. Kuvassa 7 (c) kasvatetaan kulmakerroinparametrin b arvoa, jolloin käyrän kasvu hidastuu. Kuvassa 7 (d) kasvatetaan keskipisteparametrin e arvoa, jolloin käyrän sijainti vain muuttuu x-akselilla mitattuna. Viimeisessä kuvassa 7 (e) kasvatetaan taas epäsymmetrisyysparametrin f arvoa, jolloin havaitaan, että malli tulee symmetrisemmäksi, kun $f \to 1$.

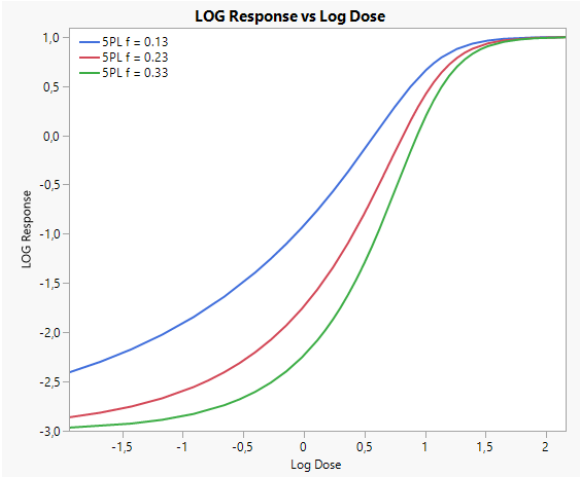
Myös neli- ja viisiparametristen logististen regressiomallien parametrit saadaan estimoitua käyttämällä pienimmän neliösumman menetelmää. Yksinkertaistettuna parametrien estimointi on residuaalineliösumman minimointia, kuten yhtälössä (3). Tässä tapauksessa mallit eivät ole lineaarisia, jolloin tavanomainen pienimmän neliösumman estimointi saattaa olla harhaista. Tästä syystä pienimmän neliösumman estimointi täytyy suorittaa numeerisia menetelmiä käyttäen, kuten luvun 3.2 Gauss–Newton -menetelmää. Tätä menetelmää täytyy kuitenkin laajentaa yleisemmäksi, jotta se toimisi myös epälineaaristen mallien tapauksessa.



(a) Parametrin d vaikutus (b) Parametrin c vaikutus



(c) Parametrin b vaikutus (d) Parametrin e vaikutus



(e) Parametrin f vaikutus

Kuva 7: (a) - (e) parametrien $b - f$ eri arvojen vaikutus viisiparametriseen logistiseen regressiomalliin. Siniset käyrät ovat keskenään yhtäpitävät eli ne on tehty samoilla parametrien arvoilla, jolloin kuvat ovat keskenään vertailukelpoisia. Näissä kuvaajissa $c > d$ ja $b < 0$, kuten kuvasta (c) havaitaan.

4.2 Painotettu pienimmän neliösumman menetelmä

Pienimmän neliösumman menetelmä epälineaarisissa regressiomenetelmissä on harhaton, jos havaintojen variaatio on homogeenistä eri pitoisuuksilla. Näin ei kuitenkaan usein ole, vaan logaritmisen pitoisuuden kasvaessa logaritmisten havaintojen variaatio pienenee. Tällöin epälineaarisen mallin pienimmän neliösumman estimoinnissa tulee huomioida myös heteroskedastisuus. Tämä huomioidaan *painotetulla pienimmän neliösumman estimoinnilla* (weighted least squares), jota laskennan helppoudesta johtuen käytetään suurimman uskottavuuden menetelmän sijasta [6]. Tilastollisen regressioteorian mukaan [7] painotettu pienimmän neliösumman menetelmä tuottaa suurimman uskottavuuden menetelmän kanssa täsmälleen samat, harhattomat parametrien estimaattorit, kun painotus on valittu oikein.

Painotettu pienimmän neliösumman estimointi olettaa jokaisen pitoisuuspisteen x_i havaintojen jakauman olevan suunnilleen normaalista [5]. Painotetussa pienimmän neliösumman menetelmässä tavoitteena on löytää parametrit, jotka minimoivat painotetun neliösumman havaittujen signaalivasteiden y_i ja sovitusten $F(x_i;\boldsymbol{\theta})$ välillä painokertoimella w_i kerrottuna

$$WRSS = \sum_{j=1}^{n_c} w_i(y_i - F(x_i;\boldsymbol{\theta})). \tag{16}$$

Painokertoimella w_i saadaan näin ollen tarkempi estimointi siihen osaan mallikäyrää, johon halutaan tarkempaa estimointia. Painokertoimen valintaan on esitetty eri vaihtoehtoja, mutta yleisesti käytetty painokerroin $w_i = \mathrm{Var}(y_i) = \sigma_j^2$ on havaintojen z_i otosvarianssi kyseisessä pitoisuuspisteessä x_i. Tällä menetelmällä saadaan painotettua sovitusmenetelmä niin, että käyrää sovitetaan tiukemmin pienille pitoisuuksille, joilla on suurempi varianssi ja löyhemmin suurille havaintojen arvoille, joilla on pienempi varianssi. Näin ollen saadaan estimointituloksena tarkempi tulos verrattuna tavalliseen pienimmän neliösumman estimointiin.

Painotetussa estimointimenetelmässä korvataan residuaalineliösumma $S(\boldsymbol{\theta})$ painotetulla residuaalineliösummalla WRSS (16), jolloin saadaan Taylorin laajennetulla sarjakehitelmällä (9) kirjoitettua normaaliyhtälöt muotoon:

$$\mathbf{V}(\boldsymbol{\theta}_t)'\mathbf{W}\mathbf{V}(\boldsymbol{\theta}_t)(\boldsymbol{\theta} - \boldsymbol{\theta}_t) = \mathbf{V}(\boldsymbol{\theta}_t)'\mathbf{W}(\mathbf{y} - F(\boldsymbol{\theta}_t)), \tag{17}$$

jossa $\mathbf{V}(\boldsymbol{\theta})$ on mallin parametrien $\boldsymbol{\theta}$ vauhtimatriisi tai toiselta tulkinnaltaan Jacobin matriisi ja $\mathbf{W}$ on painokertoimien w_i diagonaalimatriisi, jossa diagonaalialkioina ovat painokertoimet w_i. Painokerroinmatriisin $\mathbf{W}$ diagonaalialkoioiden w_i arvojen ollessa yksi on kyseessä tavallinen pienimmän neliösumman estimointi. Painotettu varianssi-kovarianssimatriisi saadaan, kun sijoitetaan painomatriisi $\mathbf{W}$ ja ratkaistaan varianssi-kovarianssimatsiisi kuten yhtälössä (11), jolloin

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{V}(\hat{\boldsymbol{\theta}})'\mathbf{W}\mathbf{V}(\hat{\boldsymbol{\theta}}))^{-1}.$$

Painotettu pienimmän neliösumman estimointi tuottaa tarkemman estimointituloksen, kun ei voida olettaa variaation olevan homoskedastista. Painotettu pienimmän neliösumman estimointikin vaatii muutamia oletuksia, kuten havaintojen approksimatiivisen normaalisuuden jokaisella pitoisuuspisteellä x_i. Lisäksi mallinnettavat

epälineaariset neli- ja viisiparametriset logistiset regressiomallit vaativat ilmiön monotonisuuden. Näitä oletuksia ei kuitenkaan aina voida todentaa, jolloin yhtenä vaihtoehtona on käyttää epäparametrisia mallinnusmenetelmiä, kuten *splinejä*.

5 Epäparametrinen mallintaminen biologisessa määrityksessä

Epäparametriset regressiomenetelmät sopivat useimpiin mallinnusongelmiin, kuten myös annosvasteilmiön mallintamiseen. Näistä käydään tässä työssä läpi splinifunktioiden *silotettu splini (smoothing spline)*, joka on kolmannen asteen paloittain määritelty polynomifunktio. Muita mahdollisia vaihtoehtoja ovat yksinkertaisin, ensimmäisen asteen splini ja hieman mutkikkaampi toisen asteen splini. Ensimmäisen asteen splini on paloittain määritelty lineaarinen funktio pitoisuuspisteiden t_j ($t = (t_0, t_1, \ldots, t_{n_c})$, $t_j \leq t_{i+1} \leq \ldots t_{n_c}$) välillä. Näitä kutsutaan *solmuiksi (knots)*, kun taas toisen asteen splinifunktio on paloittain kvadraattinen funktio välillä $T = [t_0,\ t_{n_c}]$. Tässä työssä käytetään kolmannen asteen silotettua spliniä.

5.1 Silotettu splini

Silotetun splinin tarkoituksena on löytää funktio $g(z)$, joka kuvaa annosvasteen ilmiötä tarkasti muttei ylisovita mallia otosaineistoon. Tässä työssä rajoitetaan kaikkien mahdollisten splinifunktioiden joukkoa sisältämään ainakin kahdesti jatkuvasti derivoituvat splinifunktiot välillä T.

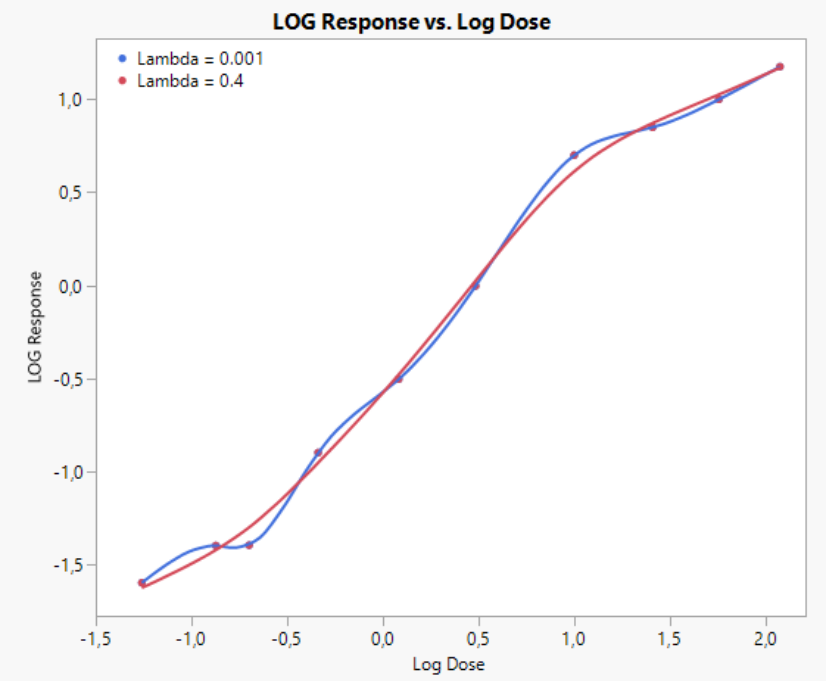
Halutaan siis löytää funktio, joka minimoi harhan ja variaation summaa. Harhan minimoiminen tarkoittaa residuaalineliösumman RSS minimointia, kun $RSS = \sum_i \sum_j (y_{ij} - g(z_{ij}))^2$. Harhan minimoiminen yksinään johtaisi havaintojen suoraan interpolointiin, jolloin saataisiin $RSS = 0$. Havaintojen interpolointi johtaisi siis aineiston ylisovittamiseen, joka kasvattaa funktiosovituksen variaatiota.

Asetetaan välin T pisteet vastaamaan standardien pitoisuuksia ($t_0 = x_0, t_1 = x_1, \ldots, t_{n_c} = x_{n_c}$). Tällöin tarkastellaan tilannetta, jossa kaikista kahdesti jatkuvasti derivoituvista funktioista $g(x_i)$ pyritään löytämään se, joka minimoi *sakotetun residuaalineliösumman (penalized residual sum of squares PRSS)*:

$$\mathrm{PRSS}(g(x_i), \lambda) = \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(v)^2 dv,\ \lambda \in (0, \infty). \tag{18}$$

$PRSS$ minimointi suoritetaan ennalta määrätyn *sakkoparametrin* $\lambda > 0$ avulla. Yhtälössä (18) ensimmäinen summalauseketermi *tappiofunktio* mittaa sopivuutta aineistoon, kun taas toinen termi *sakkofunktio* toisen derivaatan arvon suuruutta ja sakottaa sitä enemmän, mitä suurempi sakkoparametri λ on. Sakkoparametri λ on ennalta valittu, ja sen suuruus vaikuttaa ratkaisuna saatavan kolmannen asteen splinifunktion sileyteen, jota havainnoidaan kuvassa 8. Kun sakkoparametri $\lambda \to 0$ saadaan (rajalla) interpoloiva kolmannen asteen splinifunktio $g(\cdot)$. Kun taas $\lambda \to \infty$, täytyy $\int g''(v)^2 dv \to 0$, jotta minimointilauseke saisi pienen arvon. Näin ollen suurella sakkoparametrin λ arvolla sovitetaan lineaarinen malli aineistoon, koska sen toinen derivaatta on nolla. Sakkoparametrin λ avulla säädellään harhan ja variaation summan suuruutta ja vaikutusta.

Funktiolla $g(\cdot)$, joka minimoi yhtälön (18) on tiettyjä erityispiirteitä. Funktio $g(\cdot)$ on paloittain määritelty kolmannen asteen polynomifunktio yksikäsitteisillä solmujen t_j arvoilla $t_1, t_2, \ldots, t_{n_c}$. Lisäksi funktiolla $g(\cdot)$ on jatkuvasti derivoituvat ensimmäisen ja toisen asteen derivaatat $g'(\cdot)$ ja $g''(\cdot)$ jokaisessa solmussa t_j. Tarpeeksi



Kuva 8: Annosvaste -ilmiön kuvaajaan sovitettu kaksi Silotettua splinifunktiota eri sakkoparametrin $\lambda = 0.001$ ja $\lambda = 0.04$ arvoilla. Pienemmällä sakkoparametrin arvolla splinifunktio on hyvin joustava ja käy jokaisen havaintopisteen läpi. Suuremmalla λ arvolla splinifunktio on jäykempi ja muistuttaa enemmän sigmoidikäyrää.

tarkan ja sileän splinin löytämisen ongelma on siis käytännössä löytää sopiva sakkoparametrin λ arvo. Tämän löytämiseksi jaetaan osa otosaineistosta opetusaineistoon ja loput havainnot validointiaineistoon. Opetusaineiston avulla sovitetaan haluttu silotettu splinifunktio ennalta määrätyllä sakkoparametrin λ arvolla ja tällä mallilla ennustetaan validointiaineiston havaintojen arvoja ja lasketaan ennusteiden ja todellisten havaintojen etäisyyksien neliösumma. Tätä menetelmää kutsutaan *ristiinvalidoinniksi (Cross-Validation)* ja tässä työssä käsitellään kahta eri ristiinvalidointimenetelmää, jotka eroavat opetus- ja validointiaineistoon jaon suhteen.

5.2 Ristiinvalidointi

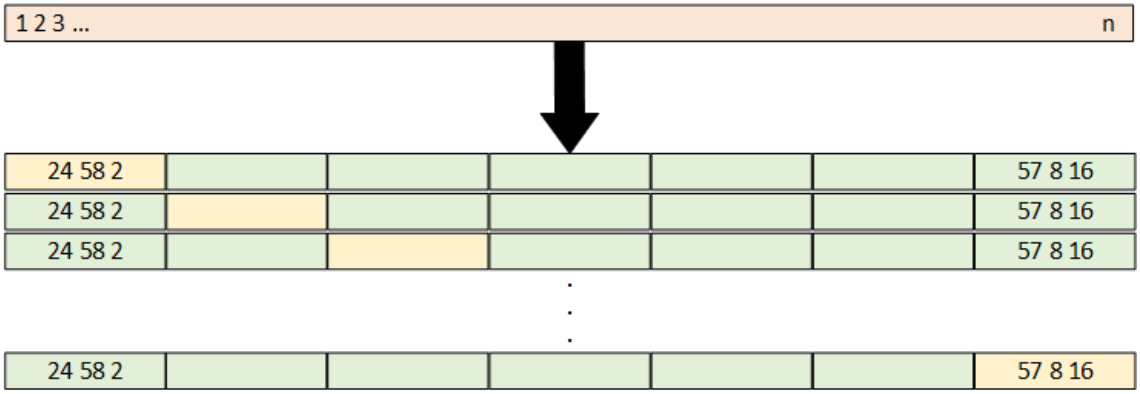
Ristiinvalidoinnissa pyritään jakamaan otosaineisto $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ satunnaisesti H kpl samansuuruiseen osajoukkoon. Tätä kutsutaan *H-kertaiseksi ristiinvalidoinniksi (h-fold Cross-Validation)*. Ensin valitaan osajoukko $h = 1$ validointiaineistoksi ja loput $\{2, 3, \ldots, H\}$ osajoukkoa opetusaineistoon. Tällöin validointiaineiston koko on n/H havaintoa ja opetusaineiston koko taas $n-n/H$. Opetusaineistoon sovitetaan paras silotettu splini ennalta määrätyllä sakkoparametrin λ_p $(p = 1, \ldots, P)$ arvolla tehtävän (18) mukaisesti. Tästä lasketaan sovitetun silotetun splinimallin avulla ennusteet validointiaineiston $(h = 1)$ havainnoille ja lasketaan ennusteneliövirheiden summa

$$MSE_{h=1} = \sum_{l=1}^{n/H} (y_l - \hat{y}_l)^2. \tag{19}$$

Tämän jälkeen asetetaan osajoukko $h = 2$ validointiaineistoksi ja loput $\{1, 3, \ldots, H\}$ osajoukkoa opetusaineistoksi ja lasketaan ennusteneliövirheiden summa $MSE_{h=2}$. Tätä jatketaan asettamalla jokainen osajoukko vuorollaan validointiaineistoksi, jolloin saadaan ennusteneliövirheiden summat jokaiselle osajoukolle $MSE_{h=1}, MSE_{h=2}, \ldots, MSE_{h=H}$ samalla sakkoparametrin λ_p arvolla. Näistä ennusteneliövirheiden summista lasketaan keskiennusteneliövirhe

$$CV_{(\lambda_p)} = \frac{1}{h} \sum_{l=1}^{h} MSE_{h=l},\ p = 1, 2, \ldots, P, \tag{20}$$

jolloin saadaan keskimääräinen ennusteneliövirheen estimaatti sakkoparametrin λ_p arvolle. Tätä toistetaan käyttämällä useita eri sakkoparametrin λ_p arvoja, joista valitaan keskiennusteneliövirheen mielessä paras sakkoparametrin λ_p arvo. Keskiennusteneliövirheen mielessä paras λ_p arvo on se, joka tuottaa pienimmän keskiennusteneliövirheen $Min(CV_{(\lambda_1)}, CV_{(\lambda_2)}, \ldots, CV_{(\lambda_P)})$.



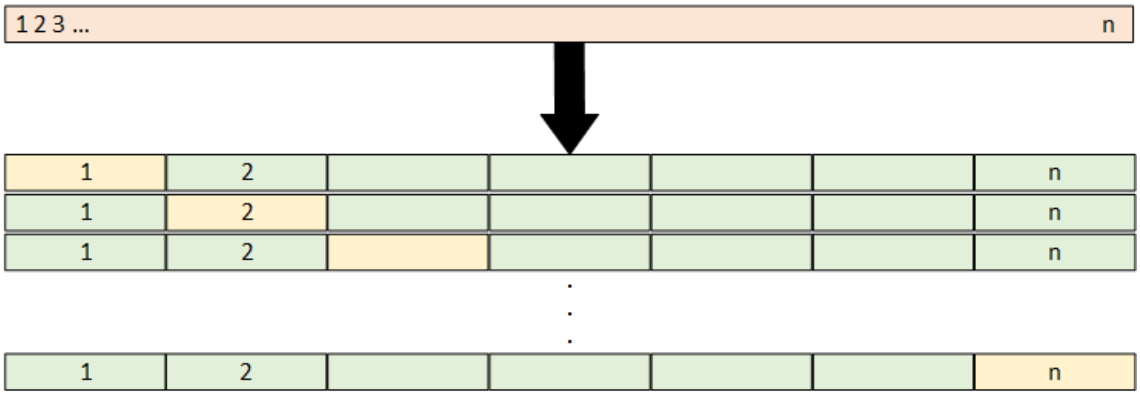
Kuva 9: Illustraatio H-kertaisen ristiinvalidoinnin toimintaperiaatteesta. Otosaineisto jaetaan satunnaisesti kuvan tapauksessa seitsemään osajoukkoon ($H = 7$), jolloin aina jokaista osajoukkoa käytetään vuorotellen validointiaineistona (keltainen laatikko) ja loppuja opetusaineistona (vihreät laatikot). Opetusaineiston avulla sovitetaan mahdollisimman sopiva silotettu splinimalli, jonka avulla ennustetaan validointiaineiston havaintojen arvoja. Jokaisen rivin kohdalla lasketaan ennusteneliövirheet $MSE_{h=1}$ validointiaineiston ja sovitetun mallin välille. Seuraavalla kierroksella valitaan seuraava osajoukkolaatikko validointiaineistoksi ja loput taas opetusaineistoksi, jolloin saadaan laskettua $MSE_{h=2}$. Tätä jatketaan, kunnes on kaikki osajoukot H käyty lävitse ja saatu laskettua $MSE_{h=1}, \ldots, MSE_{h=H}$.

Yllä on kuvattu H-kertainen ristiinvalidointimenetelmä, joka on varsin suosittu sen nopeuden ansiosta, osajoukkojen lukumäärän ollessa kohtuullinen esim. $H = 10$ tai $H = 5$. Tällöin voidaan käydä läpi tehokkaasti useitakin eri sakkoparametrin λ_p $(1, \ldots, P)$ arvoja eli voidaan asettaa P suureksikin. Ristiinvalidoinnin etuja on myös sen sisäänrakennettu testaus validointiaineistoa hyödyntämällä. Voidaan todeta, että opetusaineiston neliövirheen arvo aliarvioi todellista neliövirheen arvoa, jolloin siis opetusaineiston neliövirhe $<$ validointiaineiston neliövirhe. Vain opetusaineiston perusteella tehdyt päätelmät voisivat olla harhaisempia ja optimistisempia kuin

validointiaineiston kanssa tehdyt päätelmät. Kuitenkin opetus- ja validointiaineistoon jaossa saattaa tapahtua joitain virheitä, jolloin tämän jaon vuoksi tapahtuvaa harhaa saattaa syntyä. Tällöin voidaan todeta, että mitä useampaan osajoukkoon otosaineisto jaetaan, sitä useampia sovituksia tehdään ja voidaan olla varmempia valitun sakkoparametrin arvosta.

Äärimmäisinä erikoistapauksina ristiinvalidoinnista ovat otosaineiston jako vain kahteen osajoukkoon ($H = 2$), jolloin suoritetaan vain kaksi sovitusta jokaiselle sakkoparametrille λ_p. Toisena ääripäänä sen sijaan on otosaineiston jako niin useaan osajoukkoon kuin havaintojen lukumäärä on, jolloin siis osajoukkojen lukumäärä vastaa otosaineiston kokoa $H = n$. Näin ollen jokaiseen osajoukkoon h tulee vain yksi havaintopari (x_h, y_h), $h = 1, \ldots, H$. Tätä erikoistapausta kutsutaan *yksi-pois -ristiinvalidoinniksi (Leave-one-Out Cross-Validation LOOCV)* ja menetelmä toimii samoin kuin H-kertainen ristiinvalidointikin ja lasketaan yhtälön (20) mukaisesti. LOOCV-tapauksessa ($H = n$) yhtälö (20) voidaan kirjoittaa muotoon

$$CV_{(\lambda_p)} = \frac{1}{n_c} \sum_{j=1}^{n_c} MSE_{h=j},\ p = 1, 2, \ldots, P. \tag{21}$$



Kuva 10: Illustraatio yksi-pois -ristiinvalidoinnin (LOOCV) toimintaperiaatteesta. Otosaineisto jaetaan $H = n$ osajoukkoon, joihin kuuluu vain yksi havainto. Jokaista osajoukkoa käytetään vuorotellen validointiaineistona (keltainen laatikko) ja loppuja opetusaineistona (vihreät laatikot).

LOOCV-estimaatilla on huomattavia etuja verrattuna suurempikokoisiin osajoukkoihin eli tapauksiin, jossa $H = 2$ tai $H = 5$. Ensinnäkin LOOCV-menetelmällä lasketulla ennustekeskineliövirheen estimaatti on huomattavasti vähemmän harhainen, koska lähes koko otosaineisto käytetään mallin opettamiseen. Muissa tapauksissa käytetään vähemmän havaintoja mallin opettamiseen ja enemmän testaukseen, jolloin saatetaan yliarvioida validointiaineiston ennustevirhettä. Toisena etuna LOOCV-menetelmällä on sen toistettavuus. Koska jokainen havainto on vuorollaan validointiaineistossa, ei osajoukkoihin jaolla ole merkitystä ja saadaan aina sama keskineliöennustevirheen arvo sakkoparametrilla λ_p. LOOCV-menetelmän haittapuolena sen sijaan on sen hitaus, kun verrataan H-kertaiseen ristiinvalidointiin ja esimerkiksi $H = 5$. Otosaineiston ollessa hyvinkin suuri, esimerkiksi $n = 100000$, täytyy LOOCV-menetelmän tehdä silotetun splinin sovitus $n = 100000$ kertaa, kun

taas H-kertaisessa ristiinvalidoinnissa suoritetaan sovitus vain esimerkiksi $H = 5$ kertaa. Mikäli sakkoparametrin λ_p arvoja on vain muutamia ei tämä välttämättä hidasta merkittävästi, mutta jos $P = 100$, LOOCV suorittaa sovituksen $n \times P = 100000 \times 100 = 10000000$ kertaa ja H-kertainen ristiinvalidointi vain viidesosan tästä.

6 Passing–Bablok -menetelmävertailu

Biologisessa määrityksessä on tavoitteena löytää malli, joka kuvaisi mahdollisimman tarkasti halutun ilmiön luonnetta. Tässä työssä halutaan, että sovitettu malli kykenee tuottamaan signaalivasteesta käännetyn pitoisuuden arvon mahdollisimman tarkasti. Empiirisessä esimerkissä luvussa 7 tarkastellaan epälineaaristen mallien ja silotetun splinin sopivuutta vertaamalla simulointiaineiston kontrollipisteiden pitoisuuksien arvoja keskenään. Tätä kutsutaan *menetelmävertailuksi* (method comparison), jossa vertaillaan kahta menetelmää keskenään (käyrä, jolla simuloitu aineisto on luotu vs. sovitettu malli). Menetelmävertailu voidaan suorittaa mm. pääkomponenttianalyysillä, lineaarisella regressiomallilla tai vaihtoehtoisesti Blandt-Alman -kontrollikuvaajalla. Biologisessa määrityksessä on kuitenkin vakiintunut menetelmävertailuksi *Passing–Bablok -regressiomenetelmä* [11], [12], joka olettaa kahden menetelmän välille lineaarisen rakenteellisen suhteen.

Merkitään koeasetelmaa, jossa mitataan $n = n_c \times k$ suuruinen otos toisistaan riippumattomia mittauksia populaatiosta, kuten luvussa 2.3 on kuvattu. Merkitään menetelmän 1 pitoisuuksia satunnaismuuttujalla X ja vastaavasti menetelmän 2 pitoisuuksia satunnaismuuttujalla Z. Näin saadaan i:nnelle mittaukselle arvot x_i ja z_i, jotka ovat realisaatioita satunnaismuuttujista X ja Z, jossa X on tunnetun käyrän arvot ja Z taas mallin sovitteiden arvot. Tällöin voidaan kuvata jokainen satunnainen muuttuja kahden komponentin summana, jossa ensimmäinen muuttuja edustaa populaation kaikkien mahdollisten näytteiden odotusarvon vaihtelua ja toinen muuttuja kuvaa mittausvirheen variaatiota kyseiselle otokselle. Näin ollen voidaan kuvata mittauksen i arvot:

$$x_i = x_i^* + \zeta_i \quad \text{ja} \quad z_i = z_i^* + \eta_i, \tag{22}$$

jossa x_i^* ja z_i^* kuvaavat kyseisen mittauksen odotusarvoa ja ζ_i ja η_i mittausvirhettä. Kun kahden menetelmän välillä on lineaarinen rakenteellinen suhde, voidaan tämä kuvata lineaarisella mallilla:

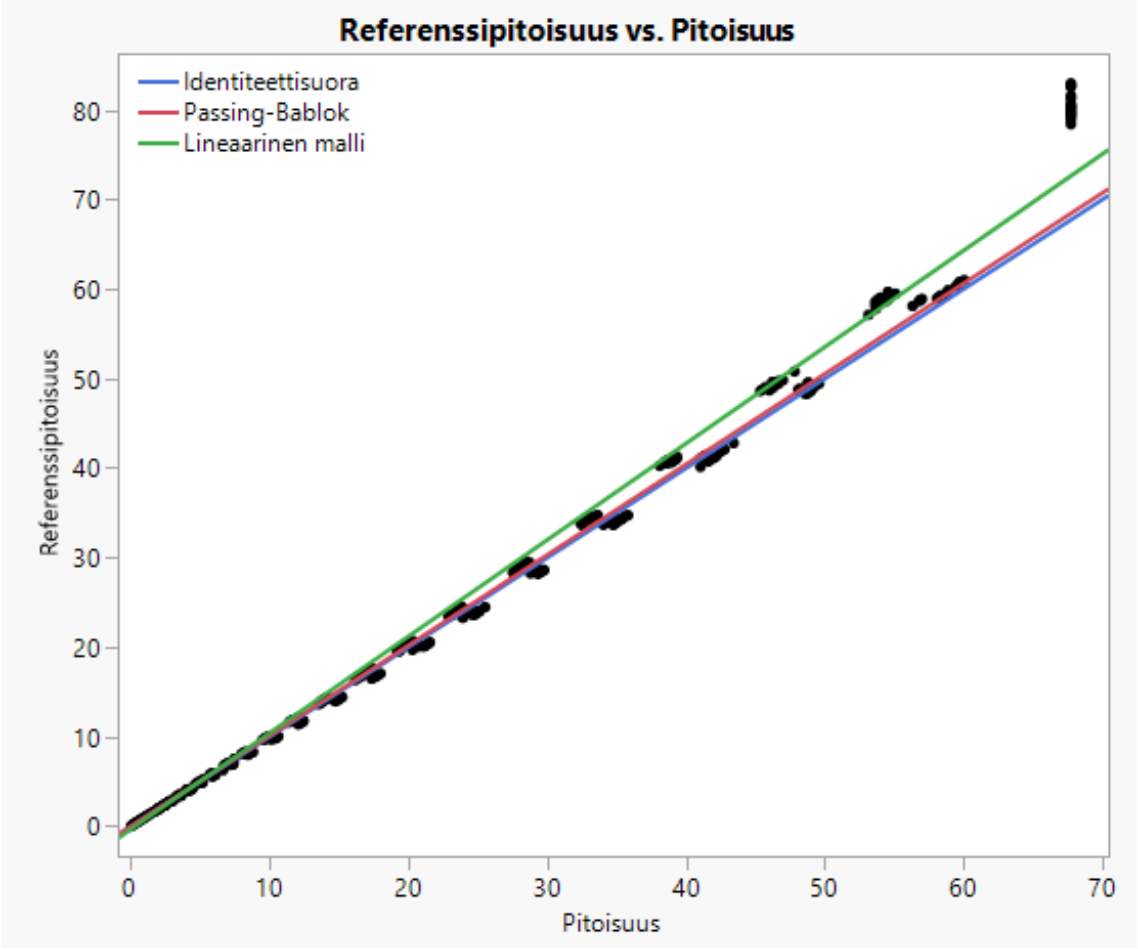
$$z_i^* = \alpha + \beta x_i^*,$$

jossa α on mallin vakiotermi ja β kulmakerroin. Tavoitteena on tällöin suorittaa n:stä mittausparista (x_i, z_i) menetelmävertailu seuraavasti:

1. estimoidaan parametrit α ja β;
2. testataan mallin lineaarisuus;
3. testataan hypoteesi $\beta = 1$;
4. testataan hypoteesi $\alpha = 0$.

Mikäli kaikki testit hyväksytään, voidaan sanoa kahden menetelmän olevan samankaltaisia keskenään eli $z_i^* = x_i^*$ ja näin ollen molemmat menetelmät tuottavat samat pitoisuudet tarkastellulla välillä.

Kuten yllä todettiin, mallinnetaan menetelmävertailututkimuksissa kahta menetelmää lineaarisella mallilla. Biologisessa määrityksessä menetelmävertailua suoritettaessa havaitaan, että:



Kuva 11: Identiteettisuora, Passing–Bablok -regressiosuora sekä lineaarinen malli sovitettuna kahden menetelmän pitoisuuksiin.

- Menetelmät 1 ja 2 sisältävät molemmat satunnaisvaihtelua.
- Mittausvirheiden jakaumat ovat harvoin normaalisti jakautuneita.
- Mittauksien odotusarvot x_i^* ja z_i^* eivät ole normaalisti jakautuneita, koska mittaukset mitataan monelta eri pitoisuuspisteeltä.
- Äärimmäiset arvot (oudokit) eivät välttämättä ole mittausvirheitä, vaan ne saattavat olla eri menetelmien ominaisuuksia. Tästä syystä näitä arvoja ei saa poistaa ilman painavaa kokeellista syytä.
- Mittausvirheiden variaatio ei ole vakio yli pitoisuuspisteiden. Variaatio kasvaa suuremmaksi, kun pitoisuus kasvaa, kuten luvussa 2.2 todettiin.

Tähän H. Passing ja W. Bablok kehittivät menetelmän, joka pystyy suorittamaan menetelmävertailun tarkasti, vaikka yllä mainitut ominaisuudet ovat läsnä. Vaikka menetelmä on lineaarinen regressiomalli, se eroaa parametrien α ja β estimoinnin suhteen tavallisiin lineaarisiin menetelmiin verrattuna. Tätä eroa havainnollistetaan kuvassa 11, jossa viimeisimmät havainnot vääntävät lineaarisen mallin kauemmas

identiteettisuorasta, kun taas Passing–Bablok -regressiosuora ei juurikaan eroa identiteettisuorasta.

6.1 Parametrien estimointi

Ennen Passing–Bablok regressiomenetelmän sovittamista oletetaan, että x_i^* ja z_i^* ovat satunnaismuuttujien X ja Z odotusarvoja mielivaltaisesta jatkuvasta jakaumasta. Oletetaan myös, että ζ_i ja η_i ovat mittausvirheiden realisaatioita, joiden jakaumat ovat samantyyppisiä. Lisäksi mittausvirheiden ζ_i ja η_i varianssien σ_ζ^2 ja σ_η^2 ei tarvitse olla vakioita pitoisuusvälillä, mutta niiden tulisi pysyä suhteellisina keskenään:

$$\frac{\sigma_\eta^2}{\sigma_\zeta^2} = \beta^{*2}.$$

Nämä oletukset takaavat tarkan ja luotettavan parametrien estimoinnin ja hypoteesintestauksen, kun $\beta \sim 1$ [11]. Tällöin saadaan estimoitua parametrien α ja β estimaatit.

Aluksi tulee estimoida vakiotermin α ja kulmakertoimen β arvot. Kahden pisteen välisen suoran kulmakerrointa voidaan hyödyntää, kun estimoidaan kulmakerroinparametrin β arvoa. Tällöin siis kahden pisteen välisen suoran kulmakerroin saadaan:

$$W_{ij} = \frac{z_i - z_j}{x_i - x_j}, \quad \text{kun } 1 \leq i < j \leq n, \tag{23}$$

jossa i ja j kuvaavat kahden eri pisteen arvoja. Tällöin, kun tarkastellaan koko aineiston läpi kaikki mahdolliset kulmakertoimet saadaan yhteensä lukumääräksi $\binom{n}{2}$. Näistä kaikista kulmakertoimista W_{ij} lasketaan korjattu mediaani, jonka laskemisessa on muutamia rajoitteita:

- Jos $z_i = z_j$ ja $x_i = x_j$, saa $W_{ij} = \frac{0}{0}$, joka on määrittelemätön. Nämä parit poistetaan estimoinnista.

- Jos $z_i > z_j$ ja $x_i = x_j$, annetaan W_{ij} arvoksi jokin hyvin suuri positiivinen luku. Koska kulmakertoimista lasketaan mediaani, ei suuri arvo ole ongelmallinen mediaanin määrityksessä.

- Jos $z_i < z_j$ ja $x_i = x_j$, annetaan W_{ij} arvoksi jokin hyvin suuri negatiivinen luku. Koska kulmakertoimista lasketaan mediaani, ei suuri negatiivinen arvo ole ongelmallinen mediaanin määrityksessä.

- Jos $W_{ij} = -1$, poistetaan tämä pari kulmakertoimen laskusta symmetriasyiden nojalla [11]. Tämä korjaa harhaa, joka johtuu riippumattomuusoletuksen puuttumisesta arvojen W_{ij} välillä.

- Jos $W_{ij} < -1$, käytetään tätä havaintoparia estimointiin, mutta lisätään yksi laskuriin L. Tätä laskurin L arvoa käytetään mediaanin korjaamiseen.

Kun kaikki yllä olevat askeleet on huomioitu järjestetään arvot W_{ij} kasvavaan suuruusjärjestykseen $W_{(1)} \leq W_{(2)} \leq \ldots \leq W_{(N)}$. Arvot W_{ij}:n laskemisessa eivät ole

keskenään riippumattomia, jolloin niiden mediaanin käyttäminen suoraan kulmakerroinparametrin β estimaattorina tuottaa harhaisen tuloksen. Tästä syystä käytetään laskurin L arvoa korjaamaan tätä harhaa, jolloin kulmakertoimen β arvo saadaan laskettua käyttämällä siirrettyä mediaania arvoista $W_{(i)}$:

$$\hat{\beta} = \begin{cases} W_{\frac{N+1}{2}+L}, & \text{kun } N \text{ pariton} \\ \frac{1}{2} \cdot (W_{\frac{N}{2}+L} + W_{\frac{N}{2}+1+L}), & \text{kun } N \text{ parillinen.} \end{cases}$$

Kaksisuuntainen luottamusväli kulmakerroinestimaattorille $\hat{\beta}$ saadaan luottamustasolla γ koottua, kun merkitään standardinormaalijakauman $(1-\gamma/2)$:ttä kvantiilia $q_{\gamma/2}$. Tällöin saadaan laskettua luottamusvälin indeksit C:

$$C_\gamma = q_{\gamma/2}\sqrt{\frac{n(n-1)(2n+5)}{18}},$$

jota siirretään,

$$M_1 = \frac{N - C_\gamma}{2}, \quad M_2 = N - M_1 + 1.$$

Tässä M_1 on pyöristetty lähimpään kokonaislukuun, jolloin myös M_2 on kokonaisluku ja N on havaintoparien kulmakertoimien todellinen lukumäärä. Tällöin saadaan luottamusväli laskettua parametrille $\hat{\beta}$ käyttämällä järjestettyjä kulmakertoimien W_{ij} arvoja:

$$b_L = W_{(M_1+L)} \leq \hat{\beta} \leq W_{(M_2+L)} = b_U. \tag{24}$$

Siirtolaskurin L arvo on siis havaintojen $W_{ij} < -1$ lukumäärä, joka vastaa nollahypoteesia $H_0 : \beta = 1$ [11]. Tällöin tässä tapauksessa $\hat{\beta}$ on hyvä ja luotettava estimaattori kulmakerroinparametrista β.

Kulmakerroinparametrin estimaattorin $\hat{\beta}$ avulla lasketaan vakiotermin estimaattori $\hat{\alpha}$ lineaarisesta mallista, kuten yhtälössä (2). Ratkaisemalla yhtälö vakiotermin α suhteen ja asettamalla kulmakertoimen tilalle estimoitu parametri $\hat{\beta}$, saadaan jokaiselle havainnolle erikseen oma vakiotermin arvo:

$$\alpha_i = z_i - \hat{\beta}x_i.$$

Vakiotermit järjestellään myös suuruusjärjestykseen $\alpha_{(1)} \leq \alpha_{(2)} \leq \ldots \leq \alpha_{(N)}$, joista lasketaan mediaani. Tämä mediaani on tällöin vakiotermin $\hat{\alpha}$ estimaatti. Kun ollaan saatu estimoitua parametreille $\hat{\alpha}$ ja $\hat{\beta}$ arvot, voidaan laskea näille luottamusvälit, joita käytetään menetelmien sopivuuden tarkasteluun.

Vastaavasti taas saadaan laskettua luottamusväli vakiotermin estimaattorille $\hat{\alpha}$ samalla luottamustasolla γ käyttämällä kulmakerroinparametrin luottamusvälin ala- ja ylärajan arvoja $(b_L; b_U)$.

Vakiotermin $\hat{\alpha}$ luottamusvälin $(a_L; a_U)$ estimointi vaatii, että vähintään puolet havaintopisteistä sijaitsee regressiosuoran yläpuolella tai suoralla. Tällöin, koska (X, Z) on jatkuva kaksiarvoinen muuttuja, niin myös sama lukumäärä havaintoja sijaitsee regressiosuoran alapuolella tai suoralla todennäköisyydellä 1. Havaintopiste (x_i, z_i) sijaitsee suoran yläpuolella vain, kun $a < z_i - \hat{\beta}x_i$, jolloin tulee todistettua, että:

$$\hat{\alpha} = \text{mediaani}(z_i - \hat{\beta}x_i)$$

on vakiotermiparametrin α estimaattori. Tätä tietoa hyödyntämällä saadaan vakiotermiparametrin luottamusväli johdettua:

$$\begin{aligned} a_L &= \text{mediaani}(z_i - b_U x_i) \\ a_U &= \text{mediaani}(z_i - b_L x_i), \end{aligned} \tag{25}$$

jossa a_L kuvaa luottamusvälin alarajaa ja vastaavasti a_U ylärajaa. Näitä luottamusvälejä $(b_L; b_U)$ ja $(a_L; a_U)$ käyttämällä voidaan testata nollahypoteeseja $H_0 : \beta = 1$ ja $H_0 : \alpha = 0$, kun ollaan ensin varmistuttu, että menetelmien X ja Z välinen ilmiö on todellisuudessa lineaarista. Lisäksi voidaan havaita, että estimointimenetelmä ei ole riippuvainen menetelmien X ja Z järjestyksestä, koska estimaatit voidaan laskea kaikille n havaintoparille (x_i, z_i), kun merkitään:

$$z^* = \hat{\alpha} + \hat{\beta} x^* \quad x^* = A + B z^*.$$

Yllä on laskettu estimaattorit parametreille α ja β, joilla on seuraava ominaisuus:

$$B = \frac{1}{\hat{\beta}} \quad A = -\frac{\hat{\alpha}}{\hat{\beta}}.$$

Tästä voidaan päätellä, että on estimoinnin kannalta samantekevää kumpaa menetelmää merkitään X:llä ja kumpaa Z:llä. Kun parametrit on estimoitu, täytyy varmistua, että menetelmien X ja Z välinen suhde on todellisuudessa lineaarista. Tämä tarkastelu on tarpeellista, koska menetelmiä vertaillaan lineaarisella mallilla, jolloin epälineaaristen suhteiden tarkastelu on ongelmallista. Lineaarisuutta testattaessa tulee tarkastella regressiosuoran sopivuutta aineistoon tai tarkastella, kuinka satunnaisesti aineisto on hajautunut sovituksen $z = \hat{\alpha} + \hat{\beta} x$ ympärillä.

6.2 Lineaarisuustarkastelu

Lineaarisuutta voidaan tarkastella *muokatulla Cusum-testillä*, jossa tarkastellaan regressiosuoran kummallakin puolella peräkkäisten havaintojen etäisyyden kumulatiivista summaa. Jos menetelmien X ja Z välillä on epälineaarista käyttäytymistä, voidaan havaita liian monta peräkkäistä mittausta joko sovitetun suoran ylä- tai alapuolella. Merkitään n_{pos}:llä havaintojen lukumäärää, jotka ovat sovitetun mallin yläpuolella $(z_i > \hat{\alpha} + \hat{\beta} x_i)$ ja n_{neg}:llä merkitään vastaavasti mallin alapuolelle jääviä havaintoja $(z_i < \hat{\alpha} + \hat{\beta} x_i)$. Jokaiselle havaintopisteparille (x_i, y_i) lasketaan pistemäärä r_i:

$$\begin{aligned} r_i &= \quad \sqrt{\frac{n_{neg}}{n_{pos}}}, \quad \text{kun } y_i > \hat{\alpha} + \hat{\beta} x_i, \\ r_i &= -\sqrt{\frac{n_{pos}}{n_{neg}}}, \quad \text{kun } y_i < \hat{\alpha} + \hat{\beta} x_i, \\ r_i &= \quad 0, \quad \text{kun } y_i = \hat{\alpha} + \hat{\beta} x_i. \end{aligned}$$

Kuitenkin pistemäärien r_i avulla laskettu tulos riippuu havaintopisteiden (x_i, y_i) järjestyksestä. Jos havaintopisteet järjestetään kasvavaan järjestykseen menetelmän X

mukaan saadaan eri tulos kuin menetelmän Z järjestyksellä. Tästä syystä järjestäminen on parempi suorittaa havaintopisteiden (x_i, y_i) ja sovitetun suoran $z = \hat{\alpha}+\hat{\beta}x$ välisen etäisyyden mukaan kasvavaan järjestykseen. Etäisyys D_i saadaan laskettua projisoimalla:

$$D_i = \frac{y_i + \frac{1}{\hat{\beta}}x_i - \hat{\alpha}}{\sqrt{1 + \frac{1}{\hat{\beta}^2}}}.$$

Tällöin voidaan pistemäärät r_i järjestää etäisyyksien mukaan kasvavaan järjestykseen $D_{(1)} \leq D_{(2)} \leq \ldots \leq D_{(N)}$. Tämä menetelmä takaa, että lineaarisuustarkastelu ei ole riippuvainen kummankaan menetelmän järjestyksestä, vaan ainoastaan etäisyydestä. Lineaarisuustarkastelun cusum-testisuure saadaan laskettua kumulatiivisen summan mukaan. Tarkastellaan koordinaatistojärjestelmää, jossa x-akseli kuvaa etäisyyksien D_i järjestystä $1 - N$ ja y-akseli taas kuvaa pistemäärien r_i kumulatiivista summaa. Kumulatiivinen summa saadaan laskettua indeksille i:

$$\text{cusum}(i) = \sum_{l=1}^{i} r_{(l)},$$

joka kertoo kyseisen indeksin positiivista tai negatiivista pistemäärän summaa pisteiden $1 - i$ välillä. Satunnainen pistemäärien järjestäminen johtaisi kohtuullisia $|\text{cusum}(i)|$ arvoja, kun taas peräkkäiset negatiiviset tai positiiviset pistemäärien r_i arvot johtavat suurempiin $|\text{cusum}(i)|$ arvoihin.

Cusum-testisuureen laskemisen jälkeen etsitään maksimaalinen peräkkäinen poikkeaminen regressiosuoralta $\max(|\text{cusum}(i)|)$. Mikäli valittu testisuureen arvo $\text{cusum}(i) < 0$, verrataan tätä niiden pistemäärien r_i:n osajoukon jakaumaan, jossa $r_i < 0$ ja vastaavasti verrataan $\text{cusum}(i) > 0$ osajoukon $r_i > 0$ jakaumaan. Tätä testisuureen arvoa verrataan Kolmogorov-Smirnov -jakauman kriittiseen pisteeseen h_γ luottamustasolla γ, taulukko 2. Mikäli testisuureen ja kriittisen pisteen välillä, jollekin indeksille i, $(i = 1, \ldots, n)$,

$$|\text{cusum}(i)| \geq h_\gamma \sqrt{n_{neg} + 1},$$

voidaan todeta ettei satunnaista hajontaa regressiosuoran ympärillä ole ja lineaarisuusoletus hylätään havaittujen menetelmien x^* ja z^* välillä kyseisen otoksen perusteella. Jos taas lineaarisuusoletus on jää voimaan, voidaan testata parametrien estimaattorien $\hat{\beta}$ ja $\hat{\alpha}$ arvoja. Suurilla otoksilla saattaa kuitenkin olla ongelmia cusum-lineaarisuustestin kanssa, koska suuressa otoksessa kaikki minimaalisetkin erot tulevat tilastollisesti merkitseviksi ja johtavat hypoteesin hylkäämiseen.. Suuret testisuureen arvot menevät yli Kolmogorov-Smirnov -jakauman kriittisen pisteen ja näin ollen lineaarisuusoletus tulee hylätyksi, vaikka menetelmien välillä olisikin todellisuudessa lineaarinen suhde.

6.3 Hypoteesintestaus Passing–Bablok -regressiosuoralle

Passing–Bablok regressiosuoran nollahypoteesina kulmakerroinparametrin estimaattorin $\hat{\beta}$ oletetaan saavan arvon yksi, $H_0 : \hat{\beta} = 1$, jolloin kahden menetelmän väliset arvot (x_i, y_i) ovat keskenään samat. Tätä saadaan testattua tarkastelemalla laskettua luottamusväliä 24 halutulla luottamustasolla. Luottamusvälin sisältäessä arvon

Taulukko 2: Kolmogorov-Smirnov -jakauman kriittisten pisteiden arvot tietyllä luottamustasolla γ.

γ	h_γ
1 %	1.63
5 %	1.36
10 %	1.22

yksi, $b_L \leq 1 \leq b_U$, voidaan olettaa menetelmien välisen suhteen olevan samankaltaista. Jos taas luottamusväli ei sisällä arvoa yksi, hylätään nollahypoteesi ja voidaan todeta, että menetelmien välillä on ainakin suhteellinen ero. Tämä testaus ei kuitenkaan ole täysin riippumaton alla piilevien jakaumien suhteen, mutta testaus tuottaa yleisesti tarpeeksi luotettavia tuloksia [11].

Kulmakerroinparametrin luottamusvälin sisältäessä arvon yksi, voidaan testata vakiotermiparametrin arvoa. Nollahypoteesina vakiotermiparametrin estimaattorille $H_0 : \hat{\alpha} = 0$, joka varmistaa, ettei kahden menetelmän välillä ole vakiotermin suuruista harhaa. Tästä voidaan varmistua, mikäli vakiotermiparametrin luottamusväli $a_L \leq 0 \leq a_U$ sisältää arvon nolla. Jos taas luottamusväliin ei kuulu arvoa nolla, hylätään nollahypoteesi ja todetaan menetelmien eroavan ainakin vakion verran ja tuottavan harhaa menetelmien välille.

Mikäli hyväksytään molemmat hypoteesit $\hat{\beta} = 1$ ja $\hat{\alpha} = 0$, päätellään $z^* = x^*$ ja menetelmien olevan identtiset.

7 Simulointiesimerkki

Tässä työssä suoritettiin esimerkki simuloimalla oikeasta mittausdatasta johdettu aineisto, johon oli lisätty kohinaa ja harhaa. Aineiston luonti ja analysointi suoritettiin SAS JMP ohjelmistolla. Simulointiesimerkin tavoitteena oli, illustroida kuinka sigmoidikäyrän muotoiset mallit sekä silotettu splinimalli sopivat standardikäyräksi. Standardikäyrän sovitukseen käytettiin standardipitoisuuspisteitä, joista jokaisesta mitattiin haluttu määrä toistomittauksia. Tämä toimi opetusaineistona, kuten luvussa 5.2. Tämän lisäksi simuloitiin validointiaineistoksi 50 kappaletta ns. kontrollipisteitä tai -mittauksia tasavälisesti logaritmisella asteikolla sijoitettuja pitoisuuspisteitä. Nämä kontrollipisteet oli sijoitettu siten, että pienimmän kontrollipisteen pitoisuus oli sama kuin pienimmän standardimittauksen pitoisuuden, sekä vastaavasti suurimman kontrollipisteen pitoisuus oli sama kuin suurimman standardimittauksen pitoisuus.

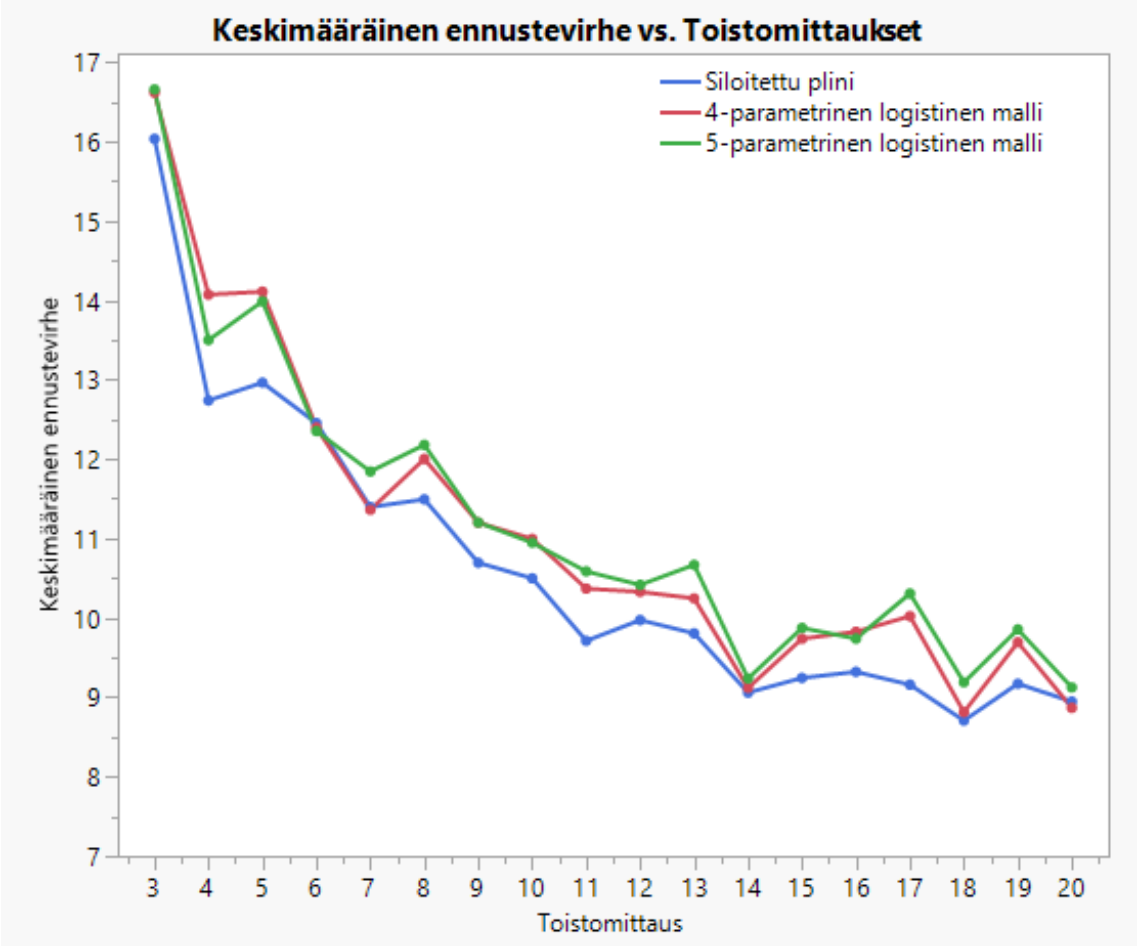
Taulukko 3: Simulointiaineiston 14 toistomittauksen otosvarianssit yhteensä kahdeksassa standardipitoisuuspisteestä.

Log-standardipitoisuus	Log-signaalivasteen otosvarianssi
-2.046	0.25455
-1.538	0.01888
-1.102	0.00662
-0.770	0.00085
-0.215	0.00548
0.461	0.00006
1.204	0.00010
1.830	0.00008

Parhaimman mallin löytämisen lisäksi tarkasteltiin toistomittauksien lukumäärien vaikutusta ja valittiin sopiva toistomittauksien lukumäärä. Toistomittauksien sopiva lukumäärä tuottaa riittävän pienen harhan, mutta ei ole liian suuri kustannuksien kannalta.

Sopivan lukumäärän löydyttyä simuloitiin vielä yksi aineisto käyttäen valittua toistomittauksien lukumäärää. Tämän jälkeen sovitettiin sigmoidikäyrämallit ja silotettu splini sekä arvioitiin kontrollipisteiden pitoisuudet. Näitä kontrollipisteiden arvioituja pitoisuuksia vertailtiin Passing–Bablok menetelmävertailulla referenssipitoisuuksiin verrattuna, jotka saatiin suoraan simulointiaineiston simulointiin käytetystä käyrästä.

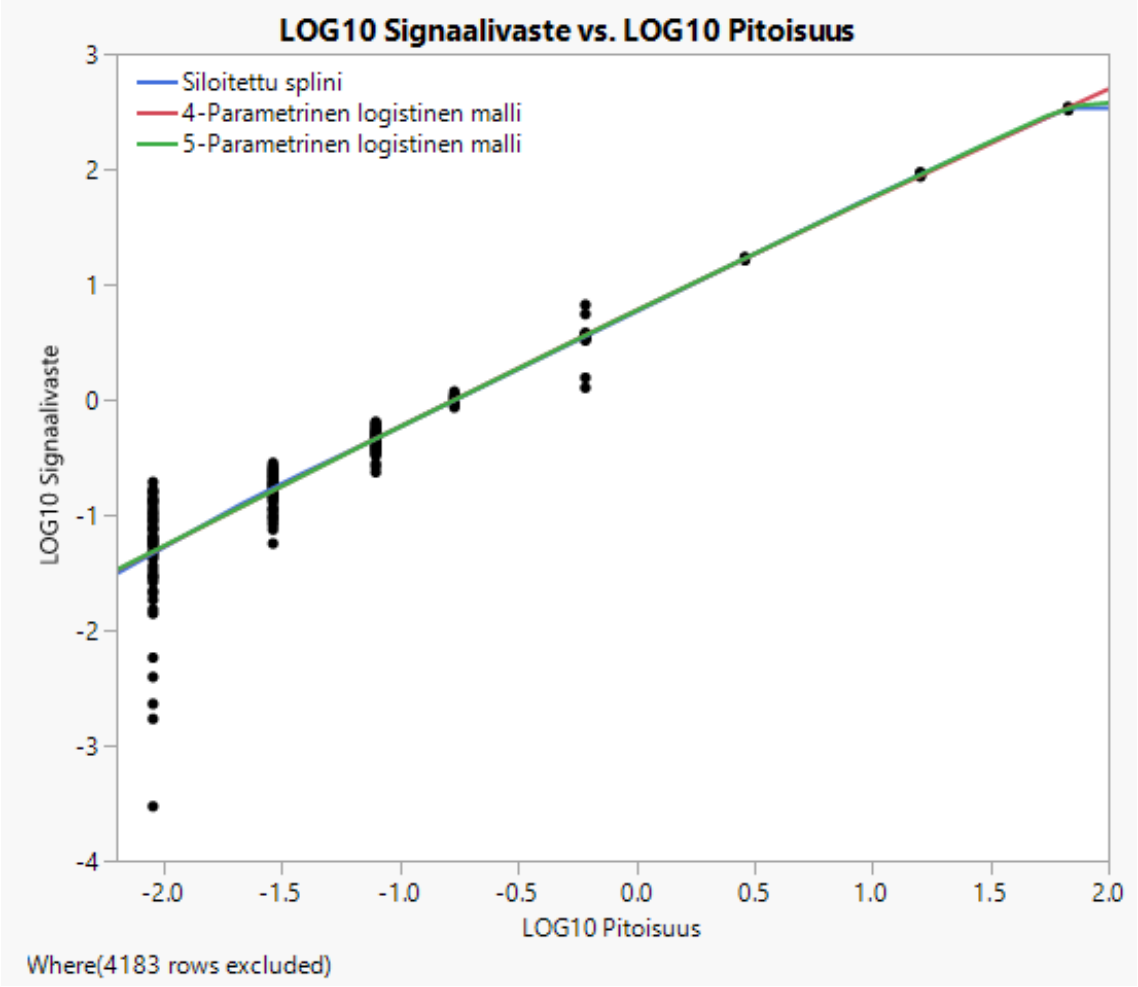
Aineisto simuloitiin käyttämällä kahdeksaa standardipitoisuuspistettä, 50 kontrollipistettä ja kuutta mittauslaitetta. Toistomittauksien lukumäärä vaihteli kolmesta mittauksesta kahteenkymmeneen, jolloin yhden simuloinnin tuottama testiaineiston koko vaihteli $8 \times 3 \times 6 = 144$ ja $8 \times 20 \times 6 = 960$ välillä ja vastaavasti validointiaineiston koko $900 - 6000$ välillä. Standardi- ja kontrollipisteiden signaalit saatiin toisen asteen polynomifunktiosta, johon oli lisätty normaalijakautunutta kohinaa ja jonka variaatio kasvoi pitoisuuden kasvaessa. Tällöin saatiin heterogeenisia normaalisti jakautuneita mittaustuloksia jokaiselta pitoisuuspisteeltä. Lisäksi jokaisella toistomittauksen lukumäärällä simuloitiin 150 aineistoa.



Kuva 12: Keskimääräisen ennustevirheen muutos toistomittauksien funktiona, kun validointiaineistona on 50 kontrollipitoisuuspistettä.

Logaritmisten signaalivasteiden otosvarianssi $\mathrm{Var}(y_i)$ laskettiin jokaiselta standardipitoisuuspisteeltä. Otosvarianssien avulla laskettiin painokertoimet pienimmän neliösumman estimointia varten. Logaritmisten signaalivasteiden otosvarianssit on kuvattu taulukossa 3. Logaritmisten signaalivasteiden otosvariansseista havaitaan, että ne pienenevät logaritmisen pitoisuuden kasvaessa. Tällöin, kun otosvarianssit sijoitetaan painotetun pienimmän neliösumman menetelmän kaavaan 16, niin pienien pitoisuuksien logaritmisilla signaalivasteilla on suurempi painokerroin kuin suuremmilla arvoilla.

Jokaiselta toistomittauksen lukumäärältä $(3 - 20)$ simuloitiin 150 eri aineistoa ja näihin jokaisen aineiston opetusaineistoksi valittiin standardipisteiden toistomittaukset. Opetusaineistoon sovitettiin painotetulla pienimmän neliösumman menetelmällä neliparametrinen-, viisiparametrinen- sekä silotettu splinimalli. Silotetulle splinimallille etsittiin jokaiselle aineistolle oma sakkoparametrin λ arvo käyttäen H-kertaista ristiinvalidointia. Ristiinvalidoinnin osajoukkoihin jako tapahtui standardipitoisuuspisteiden mukaisesti, jolloin jokaisen standardipitoisuuspisteen toistomittaukset olivat vuorotellen ristiinvalidoinnin validointiaineistona ja loput pisteet opetusaineistona. Sovitettujen mallien avulla ennustettiin viidenkymmenen kontrolli-

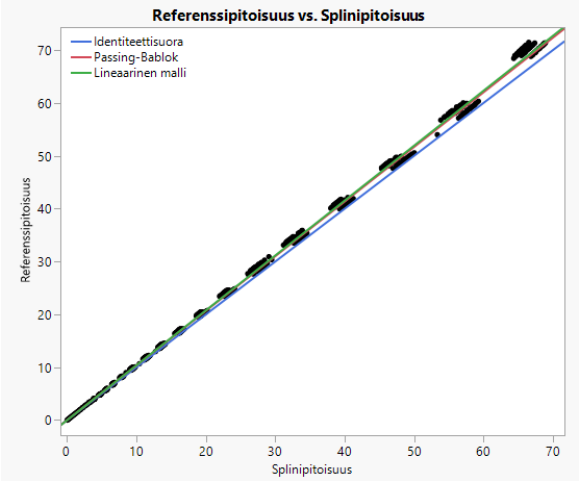


Kuva 13: silotettu splini, neliparametrinen logistinen malli sekä viisiparametrinen logistinen malli sovitettuna simulointiaineiston opetusaineistoon, kun toistomittauksien lukumäärällä 14.

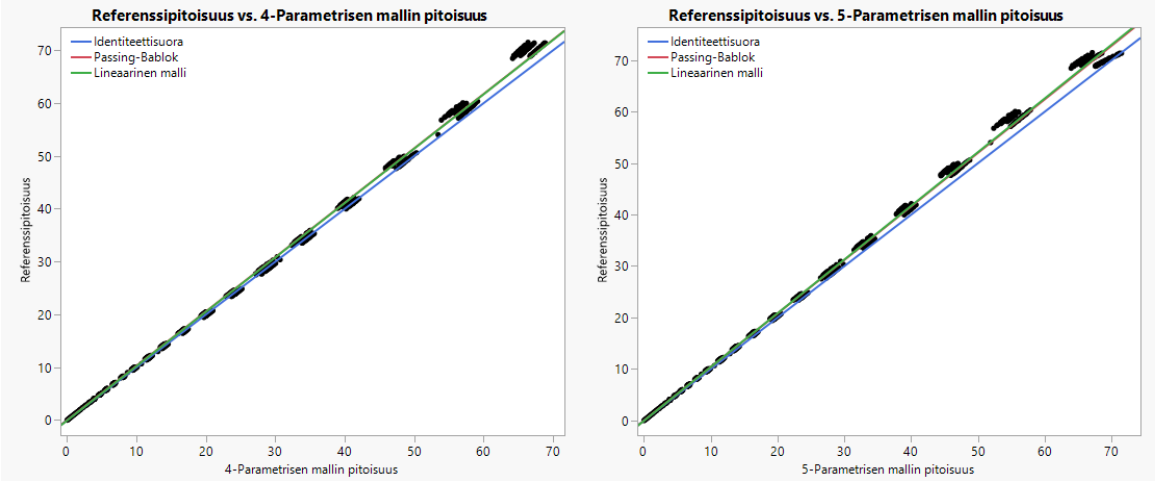
pisteen pitoisuudet, joita verrattiin simuloituihin referenssipitoisuuksiin. Näin ollen saatiin laskettua jokaiselle toistomittaukselle keskimääräiset ennustevirheet, jotka esitetään kuvassa 12.

Kuvasta 12 havaitaan, että silotetulle splinille sekä neliparametriselle logistiselle mallille saadaan pienin keskimääräinen ennusteneliövirhe 18 toistomittauksella, kun taas viisiparametriselle logistiselle mallille minimi saatiin 20 toistomittauksella. Havaittiin myös, että silotettu splinimalli tuottaa jatkuvasti pienimmän ennusteneliövirheen, muutamaa poikkeusta lukuun ottamatta. Kuvasta 12 nähdään myös, ettei keskimääräinen ennustevirhe laske juurikaan 14 toistomittauksen jälkeen, joten toistomittauksien lukumääräksi valittiin 14.

Simuloitiin vielä yksi aineisto asettamalla toistomittausten lukumääräksi 14. Aineistosta laskettiin otosvarianssit jokaiselle standardipitoisuuspisteelle toistomittauksien logaritmisista signaalivasteista. Nämä lasketut otosvarianssit asetettiin painotetun pienimmän neliösumman painokertoimiksi. Aineistolle sovitettiin neli- ja viisiparametrinen logistinen regressio malli sekä siloitettu splinimalli, joita esitetään kuvassa 13. Mallit sovitettiin kuten yllä käyttämällä standardimittauspisteiden lo-



(a) Silotetun splinimallin pitoisuuksien vertailu



(b) Neliparametrisen mallin pitoisuuksien vertailu(c) Viisiparametrisen mallin pitoisuuksien vertailu

Kuva 14: Menetelmävertailu simulointiaineiston referenssipitoisuuden, (a) silotetusta splinistä, (b) neliparametrisesta- ja (c) viisiparametrisesta mallista arvioitujen pitoisuuksien välillä. Menetelmävertailussa kuvattuna identiteettisuora, Passing–Bablok regressiosuora sekä lineaarinen malli.

garitmisia signaalivaste ja -pitoisuuspareja opetusaineistona. Opetusaineistoon sovitetuilla malleilla arvioitiin kontrollimittauksille pitoisuudet. Kuvasta 13 nähdään, että jokainen malli sopii hyvin opetusaineiston standardipitoisuuspisteiden toistomittauksiin. Validointiaineiston kontrollimittausten ennustevirheen perusteella silotettu splinimalli tuotti pienimmän ennustevirheen 7.3. Suurin keskimääräinen ennustevirhe saatiin neliparametrinella logistinella mallilla 10.4 ja viisiparametrinen malli lähes saman kuin silotettu splinimalli 7.8.

Jokaisesta sovitetusta mallista arvioiduille pitoisuuksille suoritettiin Passing–Bablok menetelmävertailu (kuviat 14 (a) - (c)). Lisäksi kuviin 14 (a) - (c) on lisätty Passing–Bablok regressiomalli sekä identiteettisuora, jossa vakiotermi on 0 ja kulmakerroin 1. Identiteettisuora kuvaa täydellistä riippuvuutta menetelmien välillä, jolloin voidaan menetelmien sanoa olevan identtisiä keskenään.

Passing–Bablok -regressiomenetelmällä estimoitujen vakiotermien α ja kulmakertoimien β arvot sekä 95 % luottamusvälit on annettu taulukossa 4. Näiden parametrien avulla varmistettiin mallin lineaarisuusoletus laskemalla Cusum-testisuureen arvot 5. Jokaiselle mallille havaittiin, että Cusum-testisuureen arvo oli huomattavasti kriittistä pistettä suurempi. Tällöin Cusum-testisuureen mukaan hylätään lineaarisuusoletus referenssipitoisuuksien ja arvioitujen pitoisuuksien välillä. Otoskoko $n = 4158$ oli hyvinkin suuri, jolloin Cusum-testisuureen lineaarisuustarkastelu saattoi olla harhaanjohtavaa.

Taulukko 4: Passing–Bablok -regressiomallin parametriestimaatit referenssipitoisuuden ja arvioitujen pitoisuuksien välillä sekä parametrien α ja β 95 % luottamusvälit.

Sovitus	α	Luottamusväli	β	Luottamusväli
Splini	0.0009	$(0.0007; 0.00011)$	1.0334	$(1.0321; 1.0349)$
Neliparametrinen malli	0	$(-0.0002; 0.0003)$	1.0289	$(1.0282; 1.0297)$
Viisiparametrinen malli	-0.0014	$(-0.0016; -0.0013)$	1.0399	$(1.0389; 1.0409)$

Passing–Bablok -regressiomenetelmän hypoteesintestaus suoritettiin, kuten luvussa 6.3 on esitetty. Mikäli kulmakerroinparametrin β luottamusväli sisältää arvon yksi ja vakiotermiparametrin α luottamusväli sisältää arvon nolla, voidaan hyväksyä nollahypoteesi menetelmien samankaltaisuudesta. Taulukosta 4 havaitaan, että yhdenkään mallin kulmakerroinparametrin β 95 % luottamusväli eivät sisällä arvoa yksi. Silotetun splinimallin ja viisiparametrisen logistisen mallin vakiotermiparametrien α luottamusvälit eivät sisältäneet arvoa nolla, mutta neliparametrisen logistisen mallin luottamusväli sisälsi. Tällöin hylättiin nollahypoteesi menetelmien samankaltaisuudesta 5 % luottamustasolla jokaisella sovitusmenetelmällä. Kulmakerroinparametrien β arvot olivat kuitenkin hyvinkin lähellä arvoa yksi ja vakiotermien α arvot olivat käytännössä nolla. Molempien parametrien luottamusvälit olivat erittäin tiukkoja suuren otoskoon takia.

Taulukko 5: Cusum-testisuureen arvo ja Kolmogorov-Smirnov -jakauman 5 % luottamustason kriittinen piste.

Sovitus	Cusum-testisuure	Kriittinen piste
Splini	880.44	51.21
Neliparametrinen malli	1457.45	40.18
Viisiparametrinen malli	1347.37	41.61

8 Päätelmät

Biologisessa määrityksessä standardikäyrän tarkka sovittaminen on olennainen osa päätelmien tekoa ja pitoisuuksien arviointia liuoksesta. Biologisia määrityksiä on useita erilaisia. Niistä etenkin aikaerotteinen fluoresenssi-immunomääritys hyödyntää ns. standardikäyrää halutun aineen pitoisuuden arviointiin mittalaitteen antaman signaalivasteen avulla. Standardikäyrä tulee määrittää mahdollisimman tarkasti, jotta arvioidut pitoisuudet olisivat tarkkoja. Standardikäyrä määritetään erillisellä määritysaineistolla etukäteen, ennen määritettävän näytteen pitoisuuden arviointia.

Määritysaineisto on kuitenkin harvoin homogeenista ja täysin satunnaista, mikä täytyy huomioida mallinnuksessa. Heterogeenisen variaation lisäksi pitoisuuden ja vastesignaalin riippuvuus on harvoin lineaarista. Aineisto saatetaan saada logaritmimuunnoksilla lineaariseksi tai ainakin tasavälisemmäksi. Muuntamattomalla määritysaineistolla erot pienimpien mittausten ja suurimpien mittausten välillä ovat suuria. Tällöin suurilla mittauksilla on huomattavan suuri vipuvoima mallintamisessa, mutta tämä saadaan logaritmimuunnoksella pienemmäksi.

Määritysaineisto on harvoin kuitenkaan täysin lineaarista, kun toistomittauksia mitataan usealta eri pitoisuuspisteeltä. Tällöin täytyy turvautua epälineaarisiin menetelmiin. Epälineaarisista menetelmistä etenkin sigmoidikäyrän muotoiset mallit, kuten neli- ja viisiparametriset logistiset regressiomallit ovat yleisesti käytettyjä biologisessa määrityksessä. Mallien parametrit saadaan estimoitua painotetulla pienimmän neliösumman menetelmällä, joka huomioi määritysaineiston heterogenian. Lisäksi painottamalla saadaan painotettua pienien pitoisuuksien arvoja mallin sovituksessa, jossa halutaan standardikäyrän olevan mahdollisimman tarkka. Epälineaaristen mallien tapauksessa parametriestimointi tapahtuu myös numeeristen menetelmien avulla iteratiivisesti ja tarkemmin sanottuna Gauss–Newton mentelmällä.

Sigmoidikäyrän muotoisten mallien lisäksi silotettuja splinimalleja voidaan käyttää standardikäyrän luomiseen. Tällöin ongelmaksi käytännössä tulee sopivan sakkoparametrin arvon löytäminen, jonka etsimiseen käytetään ristiinvalidointimenetelmiä. Tässä työssä käytettiin 8-kertaista ristiinvalidointia, jossa määritysaineisto jaetaan pienempiin osajoukkoihin pitoisuuspisteittäin. Osajoukkojen toistomittausten arvoja pyritään vuorotellen ennustamaan, kun silotettu splinimalli opetetaan käyttämällä loppujen pitoisuuspisteiden toistomittauksien arvoja.

Sopivan mallin löydyttyä halutaan varmistua mallin sopivuudesta. Tämä tapahtuu käytännössä vertailemalla sovitetusta standardikäyrästä arvioituja pitoisuuksia referenssipitoisuuksiin, jotka on saatu esimerkiksi vanhemmasta standardikäyrästä arvioimalla. Tähän menetelmävertailuun on kehitetty useita tapoja, joista Passing–Bablok -regressiomenetelmä on vakiinnuttanut paikkansa biologisessa määrityksessä. Passing–Bablok -regressiomenetelmä on lineaarista mallia luotettavampi regressiosuora, koska se ei ole herkkä mahdollisesti suuren vipuvoiman omaaville havainnoille. Passing–Bablok -regressiomenetelmä kuitenkin olettaa lineaarisen suhteen menetelmien välille, jota saadaan testattua Cusum-lineaarisuustestillä. Cusum-lineaarisuustesti kuitenkin hylkää herkästi suurella otoskoolla lineaarisuusoletuksen, vaikka menetelmien välinen suhde olisikin todellisuudessa lineaarista.

Simulointiaineiston perusteella havaittiin, että toistomittauksien lukumäärän kas-

vaessa saadaan tarkempi sovitus keskimääräisellä ennustevirheellä mitattuna. Kuitenkin kymmenen toistomittauksen jälkeen ennustevirheen ero alkaa pienenemään ja 14 toistomittauksen jälkeen ero on minimaalista. Lisäksi havaittiin, että silotettu splinimalli tuottaa pienimmän keskimääräisen ennustevirheen ja sigmoidikäyrän muotoisista malleista neliparametrinen logistinen malli vaikuttaisi olevan yleisempää viisiparametrisempaa logistista mallia keskimäärin hieman parempi.

Sovitetuista malleista arvioituja pitoisuuksia vertailtaessa referenssipitoisuuksiin Passing–Bablok -regressiomenetelmällä havaitaan, ettei yksikään sovitus ole Cusum-lineaarisuustestin mukaan lineaarinen. Kuvaa katsellessa kuitenkin suhde vaikuttaisi olevan lineaarista. Kun tiedetään Cusum-lineaarisuustestin hylkäävän herkästi lineaarisuusoletuksen suurella otoskoolla, voidaan suhteen olettaa olevan lineaarista. Passing–Bablok -regressiomenetelmän hypoteesintestaus mallien samankaltaisuudesta hylätään myös, koska parametrien luottamusvälit eivät sisällä arvoa yksi ja nolla. Parametrien estimaatteja lähemmin tarkasteltaessa tarkemmin voitiin kuitenkin vakiotermien olevan käytännössä nolla jokaisella sovituksella sekä kulmakertoimien olevan lähellä yhtä. Luottamusvälit eivät sisällä nollahypoteesin mukaisia arvoja, koska otoskoon ollessa suuri ovat luottamusvälit erittäin kapeat.

Simulointiaineiston perusteella todettiin, että jokainen sovitettu malli sopii aineistoon, mutta silotettu splinimalli tuottaa huomattavasti pienimmän keskimääräisen ennusteneliövirheen. Kuitenkin yksinkertaisemmat sigmoidikäyrän muotoiset mallit sopivat aineistoon hyvin, mutta kumpikaan näistä malleista ei vaikuttaisi sopivan toista mallia paremmin. Passing–Bablok -regressiomenetelmän mukaan neliparametrinen logistinen malli toimii kuitenkin parhaiten, kun taas silotetusta splinimallista arvioitujen pitoisuuksien suhde referenssipitoisuuksiin on eniten lineaarista. Kaiken kaikkiaan silotettu splinimalli näyttäisi sopivan parhaiten standardikäyrän sovittamiseen, mutta muita menetelmiä ja malliperheitä on syytä tarkastella ennen standardikäyrän käyttöönottoa.

Viitteet

[1] Mzolo T. V. *Statistical methods for the analysis of bioassay data*, Technische Universiteit Eindhoven, 2016

[2] Finney D. J., *Statistical Method in Biological Assay*, 2nd ed, 2nd impression, Charles Griffin & Company, 1971

[3] Finney D.J. *Bioassay and the Practice of Statistical Inference*, International Statistical Institute (ISI), Vol 47, No 1, 1979

[4] Huang H. *Nonlinear Regression Analysis*, International Encyclopedia of Education, 3rd ed., Oxford: Elsevier, 2010

[5] Paul G. Gottschalk, John R. Dunn *The five-parameter logistic: A characterization and comparison with the four-parameter logistic*, Analytical Biochemistry 343 (2005) 54–65, Brendan Technologies, Inc., 2236 Rutherford Road, Suite 107, Carlsbad, CA 92008, USA, 2005

[6] R. J. Carrol , D. Ruppert *Transformation and weighting in regression*, New York, NY: Chapman and Hall, 1988

[7] J. Berkson, *Estimation by least squares and by maximum likelihood*, Berkeley Symposium on Mathematical Statistics and Probability, pp. 1–11, 1956

[8] T. Hastie, R. Tibshirani, D. Witten, G. James *An Introduction to Statistical Learning with Applications in R*, 7th printing, Springer Texts in Statistics, 2013

[9] T. Hastie, R. Tibshirani, J. Friedman *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, 2nd ed.,Springer Series in Statistics, 2009

[10] V. Guardabasso, D. Rodbard, P. J. Munson *A model-free approach to estimation of relative potency dose-response curve analysis*, the American Physiological Society, Vol 252 Issue 3, 357-364, 1987

[11] H. Passing, W. Bablok *A new biometrical procedure for testing the equality of measurements from two different analytical methods*, J. Clin. Chem. Clin. Biochem. Vol. 21, pp. 709-720, 1983

[12] L. Bilic-Zulle, *Comparison of methods: Passing and Bablok regression*, Biochemia Medica 21(1):49-52, 2011