View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UTUPub

Harjuregression ja lasson vertailu äidin rintamaidon oligosakkaridiaineistossa

Helena Ollila

Pro gradu -tutkielma
Toukokuu 2020

MATEMATIIKAN JA TILASTOTIETEEN LAITOS
TURUN YLIOPISTO

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun
alkuperäisyys on tarkastettu Turnitin OriginalityCheck -järjestelmällä.

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

OLLILA, HELENA: Harjuregression ja lasson vertailu äidin rintamaidon oligosakkaridiaineistossa
Pro gradu -tutkielma, 47 s., 31 liites.
Tilastotiede
Toukokuu 2020

Teknologian kehitys on yleistänyt suuret ja leveät aineistot myös lääketieteessä. Lineaarisen regression sovittaminen aineistoon ei aina ole mahdollista, kun havaintojen määrä on pieni muuttujien määrään verrattuna. Kutistamismenetelmillä pyritään korjaamaan PNS-estimoinnissa esiintyviä ongelmia kuten multikollineaarisuutta, poikkeavien havaintojen tuottamia ongelmia sekä estimaattorien varianssin kasvua, joka syntyy, kun muuttujien lukumäärä lähestyy havaintojen lukumäärää.

Harjuregressio ja lasso ovat kutistamismenetelmiä, jotka perustuvat regressiokertoimien kutistamiseen. Menetelmät rajoittavat regressiokertoimia valitulla sakkofunktiolla ja tasoitusparametrilla. Harjuregressiossa sakkofunktiona toimii parametrien neliöiden summa ja lassossa parametrien itseisarvojen summa. Tasoitusparametri kontrolloi kutistamisen määrää; mitä suurempi parametrin arvo on, sitä enemmän regressioparametreja kutistetaan kohti nollaa. Lasso kykenee myös mallin valintaan, koska se voi kutistaa parametrit täsmälleen nollaksi. Elastinen verkko yhdistää lasson ja harjuregression hyödyt. Siltaregressio on taas näiden kahden menetelmän yleistys.

Äidin rintamaidon oligosakkaridit ovat äidin rintamaidossa olevia hiilihydraatteja, joita lapsi ei käytä ravinnokseen. Tutkielmassa sovitetaan harjuregressio, lasso ja PNS-estimoitu lineaarinen regressio äidin rintamaidon oligosakkaridiaineistoon ja vertaillaan mallien tuottamia tuloksia ja niiden sopivuutta aineistoon. Tutkielmassa tutkitaan, miten nämä oligosakkaridit ennustavat lapsen painon z-scorea yhden vuoden iässä.

Asiasanat: kutistamismenetelmät, harjuregressio, lasso, äidin rintamaidon oligosakkaridit, HMO, lapsen kasvu.

Sisältö

1 Johdanto **2**

2 Lineaarinen regressio ja ennustaminen **4**
2.1 Yleinen lineaarinen malli 4
2.2 PNS-menetelmä 5
2.3 Ohjattu oppiminen ja ennustamisen näkökulma 6
2.4 Rajoitettu lineaarinen regressio 6
2.5 Tappiofunktio ja ennustevirhe 8

3 Harjuregressio **11**
3.1 Määritelmä 11
3.2 Mallin harha 13
3.3 Harjuregressio ja singlulaariarvohajotelma 14

4 Lasso **17**
4.1 Määritelmä 17
4.2 Mukautuva lasso 18
4.3 Ryhmitetty lasso 20

5 Harjuregression ja lasson vertailu **21**

6 Elastinen verkko ja siltaregressio **24**
6.1 Elastinen verkko 24
6.2 Siltaregressio 24

7 Multikollineaarisuus **26**

8 Soveltaminen aineistoon **27**
8.1 Aineisto 27
8.2 Alkutarkastelu 28
8.3 Mallien estimointi 33
8.4 Mallien tulokset ja niiden vertailu 39

9 Johtopäätökset **44**

Kirjallisuutta **46**

Liitteet **48**

1 Johdanto

Suuret aineistot yleistyvät yhä enemmän teknologian kehityksen tehdessä tiedon keruusta helpompaa. Tällaisten aineistojen yhteydessä ei voida soveltaa perinteisiä aineistonhallinnointitapoja ja aineiston analyysikeinoja. Tarvitaan suuriin ja monimutkaisiin aineistoihin sopivia uusia tilastollisia menetelmiä. Tilastollisella oppimisella pyritään eri menetelmien avulla poimimaan aineiston tärkeimmät rakenteet ja trendit ja ymmärtämään aineistoja. Tilastollinen oppiminen voidaan jakaa ohjattuun (supervised) ja ohjaamattomaan (unsupervised) oppimiseen. Ohjatussa oppimisessa on selvä vastemuuttuja (output), jota pyritään selittämään selittävillä muuttujilla (input) tilastollisen mallin avulla. Ohjaamattomassa oppimisessa havaitaan vain selittävät muuttujat ja tavoitteena on tutkia näiden muuttujien keskinäisiä suhteita ja aineiston rakennetta. Tilastollinen oppiminen hyödyntää usein perinteisien regressio- ja ennustemallien laajennuksia. [1]

Teknologian kehitys on mahdollistanut myös (esimerkiksi lääketieteen tutkimusaloilla kuten molekyylibiologiassa) yhä tarkempien mittauksien tekemisen ihmisen mikromittakaavan ilmiöiden tutkimuksissa. Lääketieteen tutkimuksissa löydetään usein paljon uusia mikrobeja ja yhdisteitä, joista ei ole olemassa aiempaa tietoa. Tällöin myös tilastollisen mallin muodostamisessa ei voida hyödyntää olemassa olevaa tietoa tämänkaltaisten yhdisteiden tärkeysjärjestyksestä tai yhteisvaikutuksista. Tilastollinen malli voi kasvaa suureksi, jos ei tiedetä, mitkä yhdisteet ovat tärkeitä tutkittavan ilmiön kannalta. Muun muassa tällaisissa tilanteissa tarvitaan tilastollisen oppimisen menetelmiä.

Yllä mainituissa tapauksissa törmätään usein tilanteeseen, jossa havaintojen määrä on pieni muuttujien määrään verrattuna. Tällaisessa tilanteessa lineaarisen regression sovittaminen aineistoon ei välttämättä ole teoreettisesti mahdollista. Jotta laajasta mallista päästään pienempään malliin, tarvitaan lineaarisen regression mallinvalintamenetelmiä. Näitä menetelmiä ovat muun muassa paras muuttujien osajoukko-algoritmi (best-subset selection), askeltavat muuttujien valitsijat (stepwise selection) ja kutistamismenetelmät (shrinkage methods) kuten harju- ja lassoregressiomenetelmät (ridge and lasso regression). Parhaan muuttujien osajoukon -algoritmissa käydään kaikki eri muuttujakombinaatiot läpi ja valitaan paras mahdollinen regressiomalli esimerkiksi mallinvalintakriteerien AIC ja BIC avulla tai pienimmän residuaalineliösumman avulla, kun vertaillaan malleja, joissa muuttujien lukumäärä on yhtä suuri. Perinteisiä askeltavia muuttujien valintamenetelmiä ovat forward, backward ja monisuuntainen askellus. Forward eli eteenpäin askeltava algoritmi lähtee mallista, jossa on vain vakiotermi, ja se lisää malliin yhden muuttujan kerrallaan. Backward eli taaksepäin askeltava algoritmi aloit-

taa täydestä mallista ja poistaa mallista kerrallaan yhden muuttujan, jolla on vähiten vaikutusta mallin sopivuuteen. Monisuuntainen askellus soveltaa näitä kahta menetelmää. Askeltavat algoritmit ovat rajoitetumpia ja tehokkaampia kuin osajoukon algoritmi, mutta ne eivät kuitenkaan usein päädy globaalisti optimaaliseen tulokseen. Mainitut diskreetit valintamenetelmät ovat usein laskennallisesti hyvin hitaita. Lisäksi, koska muuttujat joko jätetään tai poistetaan mallista, menetelmissä on suurta vaihtelua ja ne eivät välttämättä vähennä mallin ennustevirhettä. [1]

Tässä pro gradu -tutkielmassa käsitellään ohjatun oppimisen menetelmiin kuuluvia kutistamismenetelmiä lassoa ja harjuregressiota ja niiden erikoistapauksia. Näitä käsitellään luvuissa 3 ja 4. Lasso ja harjuregressio perustuvat regressiokertoimien kutistamiseen. Menetelmät lisäävät ja poistavat parametreja muokatun pienimmän neliösumman perusteella sekä rajoittavat samanaikaisesti regressiokertoimia valitulla sakkofunktiolla ja tasoitusparametrilla. Kutistamismenetelmät ovat jatkuvia mallinvalintamenetelmiä, joten ne eivät kärsi niin korkeasta vaihtelevuudesta kuin diskreetit menetelmät. [1] Kutistamismenetelmiä on edellämainituiden lisäksi monia, muun muassa elastinen verkko (elastic net), siltaregressio (bridge regression), mukautettu elastinen verkko (adaptive elastic net) ja bayesiläinen lasso (bayesian lasso). Menetelmiä kehitetään koko ajan lisää. [2]. Jotta kutistamismenetelmien teoriaa voidaan ymmärtää, on tutustuttava ensin yleiseen lineaariseen malliin ja pienimmän neliösumman menetelmään, joita käsitellään tutkielmassa luvussa 2.

Kutistamismenetelmiä sovelletaan tässä tutkielmassa havaintoaineistoon, jonka tutkimuskysymys on, miten äidin rintamaidosta mitatut oligosakkaridit (Human milk oligosaccharides = HMO) ovat yhteydessä lapsen kasvuun. Esimerkeissä käytetään Hyvän kasvun avaimet -seurantatutkimuksen (HKA) aineistoa äidin ominaisuuksista, äidin rintamaidon koostumuksesta ja lapsen kasvutiedoista. Tutkimuksen tarkempana tarkoituksena on selvittää, mitkä äidin ominaisuudet vaikuttavat rintamaidon oligosakkaridipitoisuuksiin ja miten nämä sakkaridipitoisuudet vaikuttavat lapsen kasvuun syntymästä viiden vuoden ikään. Tutkimuksen tulokset esitellään artikkelissa *"Associations between human milk oligosaccharides and growth in infancy and early childhood"* [3]. Pro gradun esimerkkiaineistossa on 19 oligosakkaridia, joiden yhteisvaikutuksia tai tärkeysjärjestystä lapsen kasvuun ei tunneta. Jotta usean sakkaridin mallista päästään yksinkertaisempaan malliin, mallin muuttujien valinnassa pitää käyttää jotain aineistoon perustuvaa mallinvalintamenetelmää kuten kutistamismenetelmiä. Tutkielman sovellusosiossa luvussa 8 vertaillaan näitä menetelmiä.

2 Lineaarinen regressio ja ennustaminen

Tässä luvussa esitetään lyhyesti yleisen lineaarisen regression peruskäsitteitä pohjustukseksi kutistamismenetelmien käsittelylle. Tutkielmassa merkitään lihavoinnilla matriiseja $\boldsymbol{X}$ ja vektoreita $\boldsymbol{x}$. Satunnaismuuttujat merkitään isoilla kirjaimilla ja havaittuja arvoja pienillä kirjaimilla. Lineaarisessa mallissa oletetaan, että regressiofunktio $E(Y \mid \boldsymbol{X})$ on lineaarinen selittävien tekijöiden $X_1, \ldots, X_p$ suhteen.

2.1 Yleinen lineaarinen malli

Yleisessä lineaarisessa mallissa (general linear model) pyritään selittämään tai ennustamaan vastemuuttujaa $Y_1, \ldots, Y_n$ kiinteiden ja ei-satunnaisten selittävien muuttujien $x_1, \ldots, x_p$ avulla. Mallissa on n kappaletta havaintoja ja p kappaletta selittäviä muuttujia. Muuttujan p ollessa yksi ($p = 1$) kyseessä on yksinkertainen lineaarinen malli (simple linear model). Muuttujan p saadessa suurempia arvoja kuin yksi kyseessä on yleinen lineaarinen malli.[4]

Määritelmä 2.1. Yleinen lineaarinen malli voidaan esittää muodossa

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Mallissa $Y_1, \ldots, Y_n$ ovat selitettäviä satunnaismuuttujia, x_{ij} kiinteitä selittävien muuttujien havaintoja, $\beta_0, \ldots, \beta_p$ tuntemattomia mallin parametreja ja $\epsilon_1, \ldots, \epsilon_n$ ei-havaittavia satunnaismuuttujia. Ei-havaittavia satunnaismuuttujia ϵ_i kutsutaan myös mallin virhetermeiksi. Parametria β_0 kutsutaan mallin vakiotermiksi. Yleinen lineaarinen malli voidaan esittää myös matriisimuodossa

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

jossa $\boldsymbol{Y}$ on n-dimensioinen sarakevektori, $\boldsymbol{X}$ $n \times (p+1)$ -dimensioinen mallimatriisi, $\boldsymbol{\beta}$ $(p+1)$-dimensioinen vektori ja $\boldsymbol{\epsilon}$ n-dimensioinen vektori. Yleisessä lineaarisessa mallissa oletetaan, että virhetermeille pätevät Gaussin ja Markovin ehdot eli $E(\epsilon_i) = 0$, homoskedastisuusoletus $Var(\epsilon_i) = \sigma^2$ ja virhetermien keskinäinen riippumattomuus. Virhetermien voidaan olettaa noudattavan myös jotain tunnettua tutkittavasta ilmiöstä riippuvaa jakaumaa. Klassisen yleisen normaalisen lineaarisen mallin tapauksessa virhetermit, ja täten myös korreloimattomat selitettävät satunnaismuuttujat, noudattavat normaalijakaumaa. Tällöin

$$\epsilon_1, \ldots, \epsilon_n \perp\!\!\!\perp, \quad \epsilon_i \sim N(0, \sigma^2), \quad \forall i = 1, \ldots, n$$

ja

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \quad \sigma^2 > 0.$$

[4]

2.2 PNS-menetelmä

Pienimmän neliösumman menetelmä eli PNS-menetelmä (Ordinary least squares, OLS) on yleisin käytetty estimointimenetelmä. Menetelmässä parametrien $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ estimaatit saadaan minimoimalla jäännösneliösummafunktiota (Residual sum of squares, RSS). Gauss-Markov lauseen mukaan PNS-menetelmän estimaattoreilla on pienin varianssi muiden harhattomien lineaarisien estimaattorien joukossa, jos Gaussin ja Markovin ehdot toteutuvat. Mikäli myös tietyt lisäoletukset toteutuvat, PNS-estimaattori on harhaton ja tarkentuva.[4]

Määritelmä 2.2. Yleisen lineaarisen mallin tapauksessa jäännösneliösummafunktio on

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ &= (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}), \end{aligned} \tag{2}$$

jossa satunnaismuuttujien $Y_1, \ldots, Y_n$ havaittuja arvoja merkitään vektorilla $\boldsymbol{y} = (y_1, \ldots, y_n)$.

Pienimmän neliösumman estimaatti saadaan, kun ensin derivoidaan jäännösneliösummafunktiota parametrivektorin $\boldsymbol{\beta}$ suhteen.

$$\frac{\partial RSS}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\beta})$$

Oletetaan, että mallimatriisi $\boldsymbol{X}$ on täysiasteinen ja $n \geq p$. Tällöin mallimatriisin $\boldsymbol{X}$ sarakkeet ovat lineaarisesti riippumattomia ja $\boldsymbol{X'X}$ on positiividefiniitti. Laskettu derivaatta asetetaan nollaksi, jolloin estimaatiksi saadaan

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{0}$$

$$\Leftrightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}.$$

Tällöin estimaattien kovarianssimatriisiksi saadaan

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'X})^{-1}\sigma^2,$$

jossa σ^2 on havaintojen y_i varianssi, joka usein estimoidaan

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Jakaja $n-p-1$ tekee estimaattorista $\hat{\sigma}^2$ harhattoman eli $E(\hat{\sigma}^2) = \sigma^2$. [1, 4]

PNS-estimoinnissa voi esiintyä ongelmia. Kovariaattien välisestä korkeasta korrelaatiosta voi seurata, että matriisi $\boldsymbol{X}$ ei ole täyttä astetta, jolloin $\boldsymbol{X'X}$ on singulaarinen. Singulaarinen matriisi on ei-kääntyvä, mikä tarkoittaa, ettei matriisille voida löytää käänteismatriisia. Tällaisen multikollineaarisuuden takia estimaatin ratkaisu ei ole enää yksikäsitteinen. PNS-estimointi on herkkä aineiston poikkeaville havainnoille. Muuttujien määrän p lähestyessä havaintojen määrää n PNS-estimaattien varianssi kasvaa ja luottamusvälit levenevät. Kun parametrien määrälle pätee $p > n$, PNS-estimaattorin varianssi voi kasvaa äärettömäksi. Tällöin PNS-estimoinnin tuottama malli sopii täydellisesti aineistoon. Tämän vuoksi mallin virhetermit, joita myös residuaaleiksi kutsutaan, ovat nollia, vaikka muuttujien välillä ei olisi yhteyttä. Tällöin kyseessä on mallin ylisovittaminen. [1, 5]

2.3 Ohjattu oppiminen ja ennustamisen näkökulma

Yleistä lineaarista mallia voidaan käyttää estimoimisen lisäksi myös ennustamiseen. Koneoppimisen näkökulmasta puhutaan tällöin opetusaineistosta (training set), joka koostuu selittäjä-vaste (input-output) pareista $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$. Ohjatun oppimisen tarkoituksena on oppia malli $\hat{f}$ esimerkistä opettajan avulla. Malli $\hat{f}$ ennustaa oppivan algoritmin avulla uudella aineistolla $\boldsymbol{x}$ uusien $\boldsymbol{y}$ arvon: $\hat{f}(\boldsymbol{x}) \approx \boldsymbol{y}$. Oppiva algoritmi pystyy muokkaamaan selittäjä-vaste suhdetta kuvaavaa mallia $\hat{f}$ erotuksen $y_i - \hat{f}(x_i)$ mukaisesti: mitä pienempi erotus oikean arvon ja ennusteen välillä, sitä parempi malli. Tätä prosessia kutsutaan esimerkistä oppimiseksi. Oppimisprosessin tarkoituksena on, että keinotekoiset ja oikeat arvot ovat tarpeeksi lähellä toisiaan, jotta uusille aineistoille saadaan mallin avulla käytännöllisiä ennusteita. [1]

2.4 Rajoitettu lineaarinen regressio

Luvussa 2.2 kerrottuja PNS-estimoinnin ongelmia pyritään korjaamaan yleistä lineaarista mallia täydentävillä menetelmillä. Vähentämällä selittävien muuttujien määrää pyritään parantamaan mallin ennusteiden tarkkuutta. Lisäksi mallin tulkintaa pyritään helpottamaan sisällyttämällä malliin vain vasteen kannalta merkityksellisiä tekijöitä. Näin vasteen ilmiöstä saadaan

parempi kokonaiskuva. Tutkielmassa esitetyt kutistusmenetelmät kuuluvat näihin täydentäviin menetelmiin johdannossa esitettyjen diskreettien mallinvalintamenetelmien lisäksi. [1]

Oikean mallin valinta on tasapainottelua harhan ja varianssin välillä. PNS-estimaatilla on usein pieni harha, mutta suuri varianssi. Ennusteiden tarkkuutta voidaan joskus parantaa kutistamalla jotkin parametrit nollaan. Tällöin PNS-estimaatin varianssi pienenee, mutta estimaatteihin tulee hieman harhaa lisää, kuitenkin yleinen estimaatin tarkkuus paranee kun keskineliövirhe pienenee. PNS-estimaattoria voidaan rajoittaa lisäämällä residuaalineliösummaan sakkofunktio (penalty function) $J(\boldsymbol{\beta})$. Sakkofunktio kerrotaan tasoitusparametrilla (smoothing parameter) λ. Kun tasoitusparametri λ saa arvon nolla, malli supistuu tavanomaiseen lineaariseen malliin. Parametrien alkuarvot voidaan estimoida pienimmän neliösumman menetelmällä.

Määritelmä 2.3. Parametrien estimointi rajoitetun lineaarisen regression tapauksessa on

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda J(\boldsymbol{\beta}) \right]. \tag{3}$$

Vakiotermi β_0 supistuu pois, kun muuttujat x_{ij} standardoidaan, jolloin $\sum_{i=1}^{n} x_{ij}/n = 0$ ja $\sum_{i=1}^{n} x_{ij}^2/n = 1$. Selittävät tekijät standardoidaan, jotta ne muuntuvat samalla tavalla tasoitusfunktiossa. Lisäksi vakiotermiä ei ole mielekästä kutistaa, koska sen estimaatti kuvaa vasteen keskimääräistä arvoa. [1, 5]

Penalisoiduissa menetelmissä tasoitusparametrin ja sakkofunktion yhdistelmää kutsutaan tasoitusfunktioksi. Yleinen L_q-normin tasoitusfunktio on

$$s(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|^q,$$

jossa $q \geq 0$ ja tasoitusparametri $\lambda \geq 0$. Tasoitusparametri λ määrää sakottamisen määrää ja sen kasvaessa estimaatit kutistuvat kohden nollaa. Tasoitusparametrin kasvaessa estimaattorin harha kasvaa, mutta varianssi pienenee. Tasoitusparametri voidaan valita muun muassa ristiinvalidoinnin avulla. K-ristiinvalidoinnissa aineisto jaetaan satunnaisesti k saman suuruiseen osaan, joista yhtä osaa käytetään vuorotellen testiaineistona ja muita jäljelle jääviä osia opetusaineistona. Ristiinvalidointi tehdään eri tasoitusparametrin λ arvoille ja valitaan se arvo, joka täyttää halutun ehdon. Ehtona voidaan käyttää esimerkiksi lineaarisen mallin keskineliövirheen (MSE, mean squared error) minimointia eli valitaan se tasoitusparametrin λ arvo, jonka mallin MSE on pienin. [5]

2.5 Tappiofunktio ja ennustevirhe

Mallinvalinnassa on kaksi vaihetta: mallin valinta ja valitun mallin arviointi. Kun aineistoon on sovitettu haluttu malli, pitää arvioida valitun mallin hyvyyttä ja sopivuutta aineistoon. Ennustamisen näkökulmasta arvioidaan, kuinka hyvin valittu malli ennustaa vastetta. Tässä on kyse tilastotieteen päätösteoriasta (Decision theory), jossa tutkitaan päätöksien arviointia erilaisilla arviointikriteereillä. Kriteeriä kutsutaan usein tappiofunktioksi (loss function) $L(\boldsymbol{y}, \hat{f}(\boldsymbol{x}))$, jossa $\hat{f}(\boldsymbol{x})$ on ennustusmalli, joka on estimoitu opetusaineiston avulla. [6]

Määritelmä 2.4. Päätösavaruus D kuvaa mahdollisten päätösten joukkoa. Teoreettiset esimerkit keskittyvät tapaukseen $D = \Theta$, jossa Θ kuvaa yleisiä vakio estimointikäytäntöjä.

$$\textit{Tappiofunktio on mikä tahansa funktio } L\textit{, jolle } \Theta \times D \in [0, \infty). \tag{4}$$

Tappiofunktio kuvaa siis mallin virhettä oikeisiin arvoihin verrattuna. Mallia voidaan arvioida myös hyötyfunktion (consequences or utility function) kautta. Hyötyfunktion määritelmä on päinvastainen kuin tappiofunktion eli $L(\boldsymbol{y}, \hat{f}(\boldsymbol{x})) = -U(\boldsymbol{y}, \hat{f}(\boldsymbol{x}))$ ja se voidaan usein määritellä oikean arvon ja estimoidun arvon etäisyydeksi. Hyötyfunktiota ei kuitenkaan usein voida muodostaa ajan ja tiedon puutteen vuoksi. Tällöin hyödynnetään usein klassisia tappiofunktioita, jotka ovat matemaattisesti mukautuvia ja hyvin dokumentoituja. [6]

Tappiofunktio valitaan aina käytetyn mallin mukaan. Yksinkertaisin tappiofunktio on luokitteluongelmissa käytetty nolla-yksi tappiofunktio (zero-one loss):

$$L(y_i, \hat{f}(x_i)) = \begin{cases} 1 \text{ jos } y_i \neq \hat{f}(x_i) \\ 0 \text{ muuten} \end{cases}$$

Tässä tappiofunktiota kutsutaan myös virheelliseksi luokitteluasteeksi (misclassification rate), jossa väärästä luokittelusta saa yhden virhepisteen ja oikeasta nolla. Lineaarisen regression tapaukseen tappiofunktioksi sopii neliövirhe tai itseisarvovirhe.

Määritelmä 2.5. Neliövirhe määritellään

$$L(\boldsymbol{y}, \hat{f}(\boldsymbol{x})) = (\hat{f}(\boldsymbol{x}) - \boldsymbol{y})^2 \tag{5}$$

ja itseisarvovirhe määritellään

$$L(\boldsymbol{y}, \hat{f}(\boldsymbol{x})) = |\hat{f}(\boldsymbol{x}) - \boldsymbol{y}|. \tag{6}$$

Näitä virheitä kutsutaan myös regressiovirheiksi. Ne saavat arvon nolla, jos ennustus on täysin oikein, muuten neliövirhe kasvaa neliöllisesti ja itseisarvovirhe lineaarisesti. Neliövirhe on yleisin arviointikriteeri. Sitä on kuitenkin kritisoitu siitä, että se rankaisee liikaa suuresta hajonnasta. Tällöin neliövirhe on myös herkkä poikkeaville arvoille, joten yksikin todella huono ennuste voi johtaa huonoon neliövirhearvoon. Itseisarvovirhe ei ole niin herkkä poikkeaville arvoille ja se ei rankaise liikaa isoja epätodennäköisiä virheitä, koska se kasvaa hitaammin. Itseisarvovirhe on kuitenkin vaikeampi optimoida ja analysoida kuin neliövirhe. Neliövirheen yleisyys johtuukin sen yksinkertaisuudesta. [6]

Tappiofunktion avulla voidaan kuvata ennustevirhettä (PE, prediction error) eli vastemuuttujan $\boldsymbol{y}$ oikean arvon ja mallin eli ennustettavan funktion $\hat{f}(\boldsymbol{x})$ tuottamien ennusteiden välistä eroa. Tappiofunktion odotusarvo kuvaa odotettua ennustevirhettä (EPE, expected prediction error). Odotettu ennustevirhe kuvaa ennustavan funktion ennusteiden odotettavissa olevaa virhettä.

Määritelmä 2.6. Odotettu ennustevirhe määritellään

$$EPE(\hat{f}) = E[L(\boldsymbol{y}, \hat{f}(\boldsymbol{x}))]. \tag{7}$$

Tämä kuvaa siis keskimääräistä ennustevirhettä satunnaisella uudella datalla. Oletetaan, että aineistoon on sovitettu malli $\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$ ja $E(\epsilon_i) = 0$ ja $Var(\epsilon_i) = \sigma_\epsilon^2$. Tilanteen yksinkertaistamiseksi oletamme, että muuttujien $\boldsymbol{x}_i$ arvot ovat kiinteitä. Selittävien muuttujien kiinnitetyllä arvolla x_0 odotettu ennustevirhe voidaan hajoittaa neliövirheen tapauksessa seuraavasti:

$$\begin{aligned} E[L(Y, \hat{f}(x_0))] &= E[(Y - \hat{f}(x_0))^2] \\ &= \sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\ &= \sigma_\epsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \\ &= \sigma_\epsilon^2 + MSE(\hat{f}(x_0)) \end{aligned} \tag{8}$$

Hajoituksen ensimmäinen termi σ_ϵ^2 on uuden testikohteen varianssi. Tätä ei voida kontrolloida vaikka todellinen $f(x_0)$ arvo olisi tiedossa. Toinen ja kolmas termi muodostavat funktion $\hat{f}(x_0)$ keskineliövirheen (MSE), kun estimoidaan arvoa $f(x_0)$. Keskineliövirhe voidaan hajoittaa harha- ja varianssikomponentteihin. Harhatermi on todellisen keskiarvon $f(x_0)$ ja estimaatin odotusarvon $E(\hat{f}(x_0))$ erotuksen neliö. Varianssitermi on estimaatin varianssi. [1]

Rajoitetussa lineaarisessa regressiossa tasoitusparametrin λ kasvaessa estimaattorin harha kasvaa, mutta varianssi pienenee. Tätä kutsutaan harhan

ja varianssin vaihtokaupaksi (bias-variance tradeoff). Yleisesti kun malli monimutkaistuu ja kasvaa, varianssi kasvaa ja harhan neliö yleensä pienenee. Koneoppimisen näkökulmasta mallin kompleksisuus valitaan usein harhan ja varianssin mukaan niin, että testivirhe (test error) eli ennustevirhe on mahdollisimman pieni. Ilmeinen testivirheen estimaatti on opetusvirhe (training error), joka neliövirheen tapauksessa on $\frac{1}{n}\sum_i L(y_i, \hat{f}(x_i)) = \frac{1}{n}\sum_i (y_i - \hat{f}(x_i))^2$. Tämä ei kuitenkaan ole testivirheen hyvä estimaatti, koska se ei kuvaa kunnolla mallin kompleksisuutta. Opetusvirhe pienenee, kun mallin kompleksisuus kasvaa. Kun malli on liian suuri ja tarkka, se oppii vain opetusaineiston ja ei osaa ennustaa hyvin uusilla aineistolla. Tällöin opetusvirhe on pieni, mutta testivirhe on suuri. Opetusvirhe on siis liian optimistinen estimaatti testivirheelle. Jos malli on taas liian yksinkertainen, se alisovittaa aineiston ja ennusteilla on suuri harha, joka johtaa taas huonoon ennustekykyyn uusilla aineistoilla. Käytännössä mallin valinta ja tasoitusparametrin valinta voidaan tehdä jakamalla aineisto kolmeen osaan: opetusaineistoon (training set), validointiaineistoon (validation set) ja testiaineistoon (test set). Opetusaineistolla sovitetaan malli, validointiaineistolla estimoidaan mallin ennustevirhe ja testiaineistolla lasketaan testivirhe ja arvioidaan lopullisen mallin yleistettävyyttä. Yleensä aineisto jaetaan niin, että 50% aineistosta kuuluu opetusaineistoon ja 25% validointi- ja testiaineistoon. [1]

3 Harjuregressio

Tässä luvussa käsitellään kutistamismenetelmiin kuuluvaa harjuregressiota. Lisäksi käsitellään harjuregression estimaattien harhaa ja vertaillaan sitä PNS-estimaattien harhaan. Lopuksi käsitellään harjuregression ja pääkomponenttianalyysin yhteyksiä.

3.1 Määritelmä

Harjuregressio (ridge regression) kutistaa regressiokertoimia sakottamalla niiden koosta. Harjuregression estimaatit saadaan minimoimalla sakotettua jäännösneliösummafunktiota. Harjuregressio käyttää L_2 normin mukaista sakkofunktiota ja se soveltuu parhaiten tilanteesen, jossa selittävien muuttujien välillä on voimakasta korrelaatiota [2].

Määritelmä 3.1. Parametrien estimointi harjuregression tapauksessa on Lagrangen muodossa esitettynä

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right], \tag{9}$$

jossa n on havaintojen lukumäärä ja p on parametrien eli muuttujien lukumäärä.

Sakkofunktiona harjuregressiossa on siis parametrien β_j neliöiden summa. Tasoitusparametri $\lambda \geq 0$ kontrolloi kutistamisen määrää. Mitä suurempi parametrin λ arvo on, sitä enemmän regressioparametreja kutistetaan kohti nollaa. Kertoimet eivät kuitenkaan voi saada täysin arvoa 0, joten harjuregressio ei kykene mallinvalintaan. Kun $\lambda = 0$, sakkofunktiolla ei ole vaikutusta lausekkeeseen ja harjuregressio muodostaa PNS-estimaatit. Harjuregressio voidaan kirjoittaa myös muodossa, jossa korostetaan parametrien rajoitusta.

Määritelmä 3.2. Parametrien estimointi harjuregression tapauksessa voidaan kirjoittaa

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2, \text{ehdolla} \quad \sum_{j=1}^{p} \beta_j^2 \leq t. \tag{10}$$

Määritelmässä (3.2) rajoitusparametri t on käänteisessä relaatiossa (3.1) parametrin λ kanssa. Jokaista parametria λ kohti on olemassa vain yksi parametrin t arvo. Toisin sanoen parametrilla t on sama tulkinta parametrin λ kanssa. [1]

Harjuregression antamat estimaatit eivät ole ekvivalentteja (equivariant) selittävien muuttujien skaalauksen suhteen. Harjuregressio antaa erilaiset estimaatit riippuen selittävien muuttujien skaalauksesta. Tämän vuoksi muuttujat yleensä standardoidaan ennen estimointia. Lisäksi vakiotermi β_0 jätetään usein pois sakkofunktiosta. Vakiotermin sakottaminen tekisi lausekkeen riippuvaiseksi alkuperäisestä vasteesta $\boldsymbol{Y}$. [1]

Kun keskistetään selittävät muuttujat, Lagrangen muodossa esitetyn harjuregression parametrien estimointiyhtälön ratkaisu voidaan jakaa kahteen osaan. Jokainen x_{ij} korvataan lausekkeella $x_{ij} - \bar{x}_j$ ja vakiotermiä β_0 estimoidaan vasteen keskiarvolla $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Tällä tavoin parametrit saadaan estimoitua harjuregression avulla ilman vakiotermiä. On huomion arvoista, että nyt mallimatriisissa $\boldsymbol{X}$ on p saraketta aikaisemman $p+1$ sarakkeen sijaan. Jäännösneliösumma voidaan tällöin kirjoittaa matriisimuodossa

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}.$$

Estimaattien ratkaisu saadaan derivoimalla jäännösneliösummaa seuraavasti:

$$\begin{aligned}
\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\left((\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}\right) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X\beta} - \boldsymbol{X}'\boldsymbol{\beta}'\boldsymbol{y} + \boldsymbol{X}'\boldsymbol{\beta}'\boldsymbol{X\beta} + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}\right) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} - 2\boldsymbol{y}'\boldsymbol{X\beta} + \boldsymbol{y}'\boldsymbol{y}\right) \\
&= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})'\boldsymbol{\beta} - 2\boldsymbol{X}'\boldsymbol{y} \\
&= 2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} - 2\boldsymbol{X}'\boldsymbol{y}
\end{aligned} \tag{11}$$

Derivoimisessa käytetään gradientin derivoinnin laskusääntöjä $\frac{\partial}{\partial \boldsymbol{w}}\boldsymbol{v}'\boldsymbol{w} = \boldsymbol{v}$ ja $\frac{\partial}{\partial \boldsymbol{w}}\boldsymbol{w}'\boldsymbol{M}\boldsymbol{w} = \boldsymbol{M}\boldsymbol{w} + \boldsymbol{M}'\boldsymbol{w}$. Seuraavaksi funktio minimoidaan asettamalla sen arvoksi $\mathbf{0}$.

$$\begin{aligned}
&2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} - 2\boldsymbol{X}'\boldsymbol{y} = \mathbf{0} \\
\Leftrightarrow &(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y} \\
\Leftrightarrow &\hat{\boldsymbol{\beta}}^{ridge} = (\boldsymbol{X'X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y},
\end{aligned} \tag{12}$$

jossa $\boldsymbol{I}$ on $p \times p$-kokoinen identiteettimatriisi. Harjuregression estimaattorin matriisimuoto on samankaltainen PNS-estimaatin matriisimuodon kanssa. Ainoa ero on, että PNS-estimaatin matriisin $\boldsymbol{X'X}$ diagonaalialkioihin on lisätty tasoitusparametri λ, jota kutsutaan myös harjanteeksi (ridge). [1]

Harjuregression edut PNS-menetelmään verrattuna pohjautuvat harhan ja varianssin vaihtokauppaan. Kun tasoitusparametri λ kasvaa, harjuregression mallin muodostamisen joustavuus laskee, jolloin estimaattien varianssi

pienenee, mutta harha nousee. Tilanteissa, jossa vasteen ja selittävien muuttujien välinen suhde on lähes lineaarinen, PNS-estimaateilla on matala harha, mutta niillä voi olla korkea varianssi. Tämä tarkoittaa, että pieni muutos harjoitusaineistossa voi aiheuttaa suuren muutoksen PNS-estimaatteihin. Varsinkin kun muuttujien määrä p on melkein yhtä suuri kuin havaintojen määrä n, PNS-estimaatit ovat todella vaihtelevia. Kun $p > n$, PNS-estimaattorilla ei ole yksiselitteistä ratkaisua, kun taas harjuregressio toimii tällaisissakin tilanteissa hyvin. Harjuregressio toimiikin parhaiten tilanteissa, joissa PNS-estimaateilla on suuri varianssi. Harjuregressiolla on myös merkittävä laskennallinen etu esimerkiksi parhaan osajoukon valintaan verrattuna. Parhaan muuttujien osajoukon -algoritmi vaatii 2^p mallin läpikäynnin. Pienilläkin arvoilla p tämä voi olla laskennallisesti raskasta. Harjuregressiossa muodostetaan vain yksi malli jokaiselle tasoitusparametrin λ arvolle. Harjuregression käyttäminen voi olla myös nopeaa ja se on lähes identtisesti yhtä nopeaa kuin lineaarisen mallin sovittaminen. [7]

3.2 Mallin harha

Sovitetulle lineaariselle mallille $\hat{f}(\boldsymbol{x}) = \boldsymbol{x}'\hat{\boldsymbol{\beta}}$ voidaan muodostaa odotettu ennustevirhe seuraavasti

$$\begin{aligned} E[L(\boldsymbol{Y}, \hat{f}(\boldsymbol{x_0}))] &= E[(\boldsymbol{Y} - \hat{f}(\boldsymbol{x_0}))^2] \\ &= \sigma_\epsilon^2 + [E\hat{f}(\boldsymbol{x_0}) - f(\boldsymbol{x_0})]^2 + ||\boldsymbol{h}(\boldsymbol{x_0})||^2\sigma_\epsilon^2, \end{aligned} \tag{13}$$

jossa p-pituinen parametrivektori $\boldsymbol{\beta}$ on estimoitu pienimmän neliösumman menetelmällä ja $\boldsymbol{x_0}$ kuvaa selittävien muuttujien kiinnitettyjä arvoja. Tässä hattumatriisi $\boldsymbol{h}(\boldsymbol{x_0}) = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}$ on n-pituinen vektori ja kuvaa lineaarisia painoja mallissa $\hat{f}(\boldsymbol{x_0}) = \boldsymbol{x_0}'(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$. Tällöin estimaatin varianssi on $Var[\hat{f}(\boldsymbol{x_0})] = ||\boldsymbol{h}(\boldsymbol{x_0})||^2\sigma_\epsilon^2$. Tässä mallin kompleksisuus on suoraan verrannollinen parametrien lukumäärään p.[1]

Harjuregressiolle odotettu ennustevirhe on muuten sama kuin lineaarisessa mallissa, mutta painot varianssitermissä ovat erilaiset $\boldsymbol{h}(\boldsymbol{x_0}) = \boldsymbol{X}(\boldsymbol{X'X} + \lambda\boldsymbol{I})'\boldsymbol{x_0}$. Olkoon $\boldsymbol{\beta_*}$ parhaan mallin f lineaarisen approksimaation parametrit eli

$$\boldsymbol{\beta_*} = \arg\min_{\beta} E(f(\boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{\beta})^2.$$

Tässä odotusarvo on muodostettu ottamalla huomioon satunnaismuuttujien $\boldsymbol{X}$ jakauma. Keskimääräinen harhan neliö on

$$E_{\boldsymbol{x_0}}[f(\boldsymbol{x_0}) - E\hat{f}_\lambda(\boldsymbol{x_0})]^2 = E_{\boldsymbol{x_0}}[f(\boldsymbol{x_0}) - \boldsymbol{x_0}'\boldsymbol{\beta_*}]^2 + E_{\boldsymbol{x_0}}[\boldsymbol{x_0}'\boldsymbol{\beta_*} - E(\boldsymbol{x_0}\hat{\boldsymbol{\beta}}_\lambda)]^2.$$

Tässä ensimmäinen termi on keskimääräinen mallin harhan neliö eli parhaiten sopivan lineaarisen mallin aproksimaation ja oikean funktion ero. Toinen

termi on keskimääräinen estimoidun harhan neliö. PNS-menetelmällä estimoidun lineaarisen mallin harha on nolla, kun taas harjuregressiolla se on positiivinen.[1]

3.3 Harjuregressio ja singlulaariarvohajotelma

Käsitellään seuraavaksi harjuregressiota singlulaariarvohajotelman (singular value decomposition, SVD) kautta. Ensin esitellään singulaariarvojen (singular values) määritelmä ja singlulaariarvohajotelma.

Määritelmä 3.3. Matriisin $\boldsymbol{A} \in \boldsymbol{M}_{m\times n}(\mathbb{R})$ singulaariarvoiksi kutsutaan lukuja

$$\sigma_i = \sqrt{\lambda_i} \quad (i = 1, \ldots, k), \tag{14}$$

missä $k = r(\boldsymbol{A})$ eli matriisin aste ja $\lambda_1, \ldots, \lambda_k$ ovat matriisin $\boldsymbol{A'A}$ positiiviset ominaisarvot [8].

Lause 3.4. Olkoon matriisin $\boldsymbol{A} \in \boldsymbol{M}_{m\times n}(\mathbb{R})$ singulaariarvot $\sigma_1 \geq \cdots \geq \sigma_k$. Merkitään $\boldsymbol{D} = diag(\sigma_1, \ldots, \sigma_k)$. Tällöin $\boldsymbol{A}$ voidaan esittää muodossa

$$\boldsymbol{A} = \boldsymbol{U\Sigma V'}, \tag{15}$$

jota kutsutaan matriisin $\boldsymbol{A}$ singulaariarvohajotelmaksi. $\boldsymbol{U}$ on ortonormaalinen $m \times m$ kokoinen matriisi, $\boldsymbol{V}$ on ortonormaalinen $n \times n$ kokoinen matriisi ja

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \in \boldsymbol{M}_{m\times n}(\mathbb{R}).$$

Matriisia $\boldsymbol{V}$ kutsutaan oikeaksi ja matriisia $\boldsymbol{U}$ vasemmaksi singulaarivektoriksi. [8]

Keskistetyn $n \times p$ havaintomatriisin $\boldsymbol{X}$ singlulaariarvohajotelma on

$$\boldsymbol{X} = \boldsymbol{UDV'}.$$

Tässä $\boldsymbol{U}$ on $n \times p$ kokoinen ortogonaalinen matriisi ja $\boldsymbol{V}$ on $p \times p$ kokoinen ortogonaalinen matriisi. (Neliömatriisi $\boldsymbol{A}$ on ortogonaalinen, jos se on kääntyvä ja $\boldsymbol{A}^{-1} = \boldsymbol{A'}$ [8].) Matriisin $\boldsymbol{U}$ sarakkeet muodostuvat symmetrisen matriisin $\boldsymbol{XX'}$ ominaisvektoreista ja vastaavasti matriisin $\boldsymbol{V'}$ sarakkeet matriisin $\boldsymbol{X'X}$ ominaisvektoreista. Hajotelmassa $\boldsymbol{D}$ on $p \times p$ kokoinen diagonaalimatriisi. Matriisin diagonaaliarvot $d_1 \geq d_2 \geq ... \geq d_p \geq 0$ ovat matriisin $\boldsymbol{X}$

singulaariarvoja eli siis matriisien $\boldsymbol{X'X}$ ja $\boldsymbol{XX'}$ ominaisarvojen neliöjuuria. Hajotelman avulla PNS-sovite voidaan kirjoittaa

$$\begin{aligned}
\boldsymbol{X}\hat{\boldsymbol{\beta}}^{PNS} &= \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \\
&= \boldsymbol{UDV'}((\boldsymbol{UDV'})'\boldsymbol{UDV'})^{-1}(\boldsymbol{UDV'})'\boldsymbol{y} \\
&= \boldsymbol{UDV'}(\boldsymbol{VD'U'UDV'})^{-1}(\boldsymbol{UDV'})'\boldsymbol{y} \\
&= \boldsymbol{UDV'}(\boldsymbol{VD'DV'})^{-1}(\boldsymbol{UDV'})'\boldsymbol{y} \\
&= \boldsymbol{UDV'}(\boldsymbol{V'})^{-1}\boldsymbol{D}^{-1}(\boldsymbol{D'})^{-1}\boldsymbol{V}^{-1}\boldsymbol{VD'U'y} \\
&= \boldsymbol{UDD}^{-1}(\boldsymbol{D'})^{-1}\boldsymbol{D'U'y} \\
&= \boldsymbol{UU'y}.
\end{aligned} \tag{16}$$

$\boldsymbol{U'y}$ ovat vektorin $\boldsymbol{y}$ koordinaatit, kun otetaan huomioon ortonormaalinen kanta $\boldsymbol{U}$. Harjuregression tapauksessa sovitteet ovat muotoa

$$\begin{aligned}
\boldsymbol{X}\hat{\boldsymbol{\beta}}^{ridge} &= \boldsymbol{X}(\boldsymbol{X'X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X'}y \\
&= \boldsymbol{UD}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{DU'y} \\
&= \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \lambda}\boldsymbol{u}_j'\boldsymbol{y},
\end{aligned} \tag{17}$$

jossa $\boldsymbol{u}_j$ ovat matriisin $\boldsymbol{U}$ sarakkeet. Lauseke $\frac{d_j^2}{d_j^2+\lambda} \leq 1$, koska $\lambda \geq 0$. Harjuregressio siis kutistaa vektorin $\boldsymbol{y}$ koordinaatteja tekijän $\frac{d_j^2}{d_j^2+\lambda}$ verran. Tällöin vektorin koordinaatteja, joilla on pieni singulaariarvon neliö d_j^2, kutistetaan eniten. [1]

Singulaariarvo on myös yhteydessä muuttujan varianssiin: havaintomatriisin $\boldsymbol{X}$ sarakkeella, jolla on pieni singulaariarvo d_j, on myös pieni varianssi, jolloin sillä on myös pieni ominaisarvo. Harjuregressio siis kutistaa näitä muuttujia eniten. Tämä nähdään suoraan myös pääkomponenttianalyysin teoriasta. Matriisin $\boldsymbol{X'X}$ ominaisarvohajotelma on

$$\boldsymbol{X'X} = \boldsymbol{VD}^2\boldsymbol{V'}.$$

Matriisin $\boldsymbol{V}$ sarakkeet $\boldsymbol{v}_j$ ovat ominaisvektoreita eli myös singulaarivektoreita ja havaintomatriisin $\boldsymbol{X}$ pääkomponenttisuuntia. Pääkomponentit saadaan ominaisarvojen avulla seuraavasti:

$$\boldsymbol{z}_j = \boldsymbol{Xv}_j = \boldsymbol{u}_j d_j.$$

Ensimmäisellä pääkomponentilla $\boldsymbol{z}_1$ on suurin otosvarianssi, joka saadaan

$$Var(\boldsymbol{z}_1) = Var(\boldsymbol{Xv}_1) = \frac{d_1^2}{n}.$$

Ensimmäinen pääkomponentti selittää eniten havaintomatriisin $\boldsymbol{X}$ variaatiosta. Toinen pääkomponentti pyrkii selittämään mahdollisimman paljon matriisin variaatiosta, mitä ensimmäinen komponentti ei selittänyt. Pääkomponentit ovat korreloimattomia keskenään ja ne ovat ortogonaalisia edelliseen verrattuna. Näin viimeisellä pääkomponentilla on pienin varianssi ja sen laskemiseen käytetään pienintä singulaariarvoa d_j. Harjuregressio siis suojelee muuttujia, joilla on suurin varianssi, koska oletuksena on, että vasteella on taipumus vaihdella eniten suuren varianssin omaavien muuttujien mukaan. [1]

Tarkastellaan vielä harjuregressiomallin vapausasteita. Harjuregressioon vaikuttavat vapausasteet saadaan kaavalla

$$\begin{aligned} df(\lambda) &= tr[\boldsymbol{X}(\boldsymbol{X'X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X'}] \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}. \end{aligned} \tag{18}$$

Yleensä tavallisen lineaarisen regression vapausasteiden määrä on vapaiden parametrien määrä p. Vaikka harjuregression parametrit eivät saa arvoa nolla, niitä kuitenkin kutistetaan tasoitusparametrin λ verran, mikä pitää ottaa huomioon myös vapausasteissa. Funktiosta (18) nähdään, että $df(\lambda) \to 0$ kun $\lambda \to \infty$. Vakiotermi lisää vapausasteiden määrää yhdellä. [1]

4 Lasso

Tässä luvussa käsitellään Lassoa, joka toimii hyvin mallinvalintamenetelmänä. Lisäksi käsitellään Lasson kahta erikoistapausta, mukautuvaa lassoa ja ryhmitettyä lassoa, jotka pyrkivät korjaamaan Lasson heikkoja puolia.

4.1 Määritelmä

Lasso (least absolute shrinkage and selection operator) on regressiokertoimien kutistamiseen perustuva rajoitettu lineaarinen regressiomenetelmä kuten harjuregressiokin. Lassomenetelmä siis lisää ja poistaa parametreja pienimmän neliösumman version perusteella. Samalla kuitenkin rajoitetaan regressiokertoimien itseisarvojen summaa. Regressiokertoimien itseisarvojen summan pitää olla pienempi kuin valittu rajoitusparametri $t \in [0, \infty]$, näin osa regressiokertoimista kutistuu kohti nollaa. Kun $\lambda \in [0, \infty]$ on suuri, monet kertoimet β_i saavat täsmälleen arvon 0 ja näin muuttuja tippuu pois mallista. Lasso toimii siis hyvin jatkuvana mallinvalintamenetelmänä. Kuten harjuregressiossa vakiotermi β_0 voidaan poistaa mallista standardisoimalla muuttujat ja estimoimalla $\hat{\beta}_0$ keskiarvolla $\bar{y}$.

Määritelmä 4.1. Parametrien estimointi lassoregression tapauksessa Lagrangen muodossa esitettynä on

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{19}$$

jossa n on havaintojen lukumäärä ja p on parametrien eli muuttujien lukumäärä.

Määritelmä 4.2. Parametrien estimointi lassoregression tapauksessa on

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2, \text{ehdolla} \quad \sum_{j=1}^{p} |\beta_j| \leq t. \tag{20}$$

Lasson sakkofunktio on L_1 normin mukainen sakko $\sum_1^p |\beta_j|$. Sakkofunktion takia parametrien estimointi ei ole lineaarinen vasteen y_i suhteen. Funktiolla ei ole siis samanlaista suljettua ratkaisumuotoa kuten harjuregressiolla. Jos mallimatriisi $\boldsymbol{X}$ on ortonormaalinen, niin lassolla on suljettu ratkaisu. Vaikka lasson ratkaisujen laskeminen on neliöllinen ohjelmointiongelma, on ratkaisupolkujen laskemiseen olemassa yhtä tehokkaita algoritmeja kuin harjuregression ratkaisemiselle. Jos rajoitusparametri t valitaan suuremmaksi kuin $\sum_1^p |\hat{\beta}_j|$, missä $\hat{\beta}_j = \hat{\beta}_j^{PNS}$ eli pienimmän neliösumman estimaatti,

niin lassoestimaatit ovat pienimmän neliösumman estimaatit $\hat{\beta}_j$. Vastaavasti sama tapahtuu kun valitaan $\lambda = 0$. Jos valitaan $t = \sum_1^p |\hat{\beta}_j|/2$ estimaatteja on kutistettu noin 50% keskimääräisistä PNS-estimaateista. Kuten harjuregressiossa ja muissa rajoitetuissa lineaarisissa regressioissa, parametri t tai λ tulisi valita niin, että odotetun ennustevirheen estimaatti minimoituu. [1]

4.2 Mukautuva lasso

Kuten tutkielmassa on aiemmin mainittu, tilastollisella oppimisella on kaksi tavoitetta: varmistaa korkea ennustetarkkuus ja löytää vasteen kannalta tärkeät selittävät muuttujat. Näillä perusteilla voidaan määrittää oracle-ominaisuudet (Oracle properties) määritelmän 4.3 mukaan.

Määritelmä 4.3. Olkoon $\boldsymbol{A} = \{j : \beta_j^* \neq 0\}$ mallin selittävien tekijöiden osajoukko, jolle pätee $|\boldsymbol{A}| = p_0 < p$. Olkoon $\hat{\boldsymbol{\beta}}(\delta)$ parametrien estimaattori kun sovitetaan malli δ. Malli δ on ennustava proseduuri, jos estimaattorilla $\hat{\boldsymbol{\beta}}(\delta)$ on seuraavat oracle-ominaisuudet:

1) *Proseduuri tunnistaa oikean osajoukkomallin* $\{j : \hat{\beta}_j \neq 0\} = \boldsymbol{A}$.
2) *Proseduurilla on optimaalinen estimointitarkkuus,*
$\sqrt{n}(\hat{\boldsymbol{\beta}}(\delta)_A - \boldsymbol{\beta}_A^*) \to_d N(\mathbf{0}, \boldsymbol{\Sigma}^*)$, *jossa* $\boldsymbol{\Sigma}^*$ *on oikean osajoukkomallin kovarianssimatriisi.*

(21)

[9]

Rajoitetun lineaarisen regression estimaattorilla on oracle-ominaisuudet silloin, kun se tunnistaa annetuista muuttujista oikean osajoukon ehdolla, että rajoitusparametri on oikein valittu. Lisäksi, kun oikea muuttujien osajoukko tiedetään etukäteen, rajoitettu regressio estimoi parametrit oikein eli samalla tavalla kuin lineaarinen regressio tällä muuttujien osajoukolla. Kutistamismenetelmä lasso vaikuttaa ennustavalta proseduurilta, mutta menetelmässä on kuitenkin muutama heikkous. Lasso pystyy suorittamaan mallinvalinnan pudottamalla osan muuttujista pois, koska L_1 normin mukainen sakko on lineaarinen. Kuitenkin lasso ylisakottaa suuria parametreja, jolloin se muodostaa näille parametreille harhaiset estimaatit verrattuna lineaariseen regressioon. Lisäksi on osoitettu, että optimaalinen tasoitusparametri λ voi tuottaa epäjohdonmukaisen muuttujavalinnan. Tämän takia oracle-ominaisuudet eivät päde lasson tapauksessa. Jotta lasson mallinvalinta olisi tarkempi, sen pitää täyttää tiettyjä ehtoja. Vaihdetaan λ_n arvoja otoskoon n mukaan. Tällöin lasson mallinvalinta on tarkentava jos ja vain jos $lim_n P(\boldsymbol{A}_n = \boldsymbol{A}) = 1$ eli todennäköisyydellä yksi, parametrin λ_n valitsema

osajoukko on oikea osajoukko. Kun havaintomatriisi $\boldsymbol{X}$ on ortonormaalinen tai kun muuttujien määrä $p = 2$, tarvittavat ehdot täyttyvät ja lasso toteuttaa oracle-ominaisuudet. Tarkemmat ehdot on esitelty artikkelissa Zou (2006). [9]

Mukautuva lasso (Adaptive lasso) on lasson versio, jonka avulla pyritään korjaamaan lasson epätarkentuvuutta mallinvalinnassa. Mukautuvassa lassossa asetetaan painot w_j, jotka säätelevät eri parametrien kutistamista. Se toteuttaa määritelmän 4.3 oracle-ominaisuudet. Mukautuva lasso on myös konveksi optimointiongelma L_1 normin sakkofunktiolla, joten se voidaan ratkaista samoilla tehokkailla algoritmeilla kuin lassokin. Sen globaali minimi voidaan siis ratkaista tehokkaasti.

Määritelmä 4.4. Parametrien estimointi mukautuvan lassoregression tapauksessa on

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{AL} &= \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right], \\ \textit{jossa } \hat{w}_j &= \frac{1}{|\hat{\beta}_j|^{\gamma}}, \quad \gamma > 0. \end{aligned} \tag{22}$$

Seuraava lause osoittaa, että L_1 sakko on yhtä hyvä kuin muutkin oracle-ominaisuudet täyttävät sakkofunktiot.

Lause 4.5. Oletetaan, että $\lambda_n/\sqrt{n} \to 0$ ja $\lambda_n n^{(\gamma-1)/2} \to \infty$. Tällöin mukautuvan lasson estimaatit toteuttavat seuraavat ehdot:

$$\begin{aligned} &1)\ \textit{Tarkentuvuus mallinvalinnassa}: \ lim_n P(\boldsymbol{A}_n^* = \boldsymbol{A}) = 1 \\ &2)\ \textit{Asymptoottisen normaalisuuden}: \\ &\sqrt{n}(\hat{\boldsymbol{\beta}}_A^{*(n)} - \boldsymbol{\beta}_A^*) \to_d N(\boldsymbol{0}, \sigma^2 \times \boldsymbol{C}_{11}^{-1}), \end{aligned} \tag{23}$$

jossa $\boldsymbol{A}$ on oikea muuttujien osajoukko ja $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix}$ positiivisesti definiitti matriisi ja $\frac{1}{n}\boldsymbol{X'X} \to \boldsymbol{C}$, jossa $\boldsymbol{C}_{11}$ on $p_0 \times p_0$ kokoinen matriisi. Matriisi $\boldsymbol{C}_{21}$ on $u \times p_0$ kokoinen matriisi, jolloin $\boldsymbol{C}_{12}$ on $b \times u$ ja $\boldsymbol{C}_{22}$ on $d \times u$ kokoisia matriiseja ja $[u, b, d] \in \mathbb{N}$ sekä $b + d = u + p_0$.

Aineistosta riippuvat painot $\hat{\boldsymbol{w}}$ johtavat lauseen 4.5 toteutumiseen. Kun otoskoko kasvaa, muuttujien, joilla on nolla-kerroin, painot lähestyvät ääretöntä, kun taas muuttujien, joilla on nollasta eriävä kerroin, painot konvergoituvat äärelliseen vakioon. Näin voidaan samanaikaisesti estimoida harhattomasti suuria ja pieniä parametreja. Tiedetään myös, että mukautuvan lasson ratkaisu on jatkuva. Mukautuva lasso siis toteuttaa oracle-ominaisuudet ja korjaa tavallisen lasson ratkaisuja. [9]

4.3 Ryhmitetty lasso

Selittävät tekijät voivat jossain tapauksissa kuulua ennalta määrättyihin ryhmiin. Esimerkiksi lääketieteessä suolistomikrobit voivat kuulua samaan biologiseen perheeseen kuten phylum ja family -tasoihin. Kategorisen muuttujan tasoja edustavat dummy-muuttujat kuuluvat myös samaan ryhmään, koska ne kuvaavat samaa muuttujaa. Tälläisessä tilanteessa voi olla mielekkäämpää kutistaa koko ryhmää yksittäisen muuttujan sijasta. Ryhmitetty lasso (grouped lasso) mahdollistaa tämän. Oletetaan, että p selittävää tekijää on jaettu L määrään ryhmiä. Ryhmässä l on p_l kappaletta muuttujia. Matriisi $\boldsymbol{X}_l$ kuvastaa ryhmän l selittäviä tekijöitä. Ryhmän parametrivektoria kuvastaa $\boldsymbol{\beta}_l$.

Määritelmä 4.6. Parametrien estimointi ryhmitetyn lasson tapauksessa on

$$\hat{\boldsymbol{\beta}}^{GL} = \min_{\beta \in \mathbb{R}^p} \left[||\boldsymbol{y} - \beta_0 \mathbf{1} - \sum_{l=1}^{L} \boldsymbol{X}_l \boldsymbol{\beta}_l||_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} ||\boldsymbol{\beta}_l||_2 \right], \tag{24}$$

jossa $\sqrt{p_l}$ kuvaa vaihtelevia ryhmäkokoja ja $|| \bullet ||_2$ on Euklidinen normi.

Euklidinen normi on nolla vain silloin kuin kaikki vektorin $\boldsymbol{\beta}_l$ komponentit ovat nollia. Näin kaava ottaa huomioon sekä ryhmä- ja yksittäisen muuttujan tason. Jollain λ arvolla, koko ryhmä voi tippua mallista. Ryhmitetyn lasson yleistetty muoto mahdollistaa myös ryhmien päällekkäisyyden. [1]

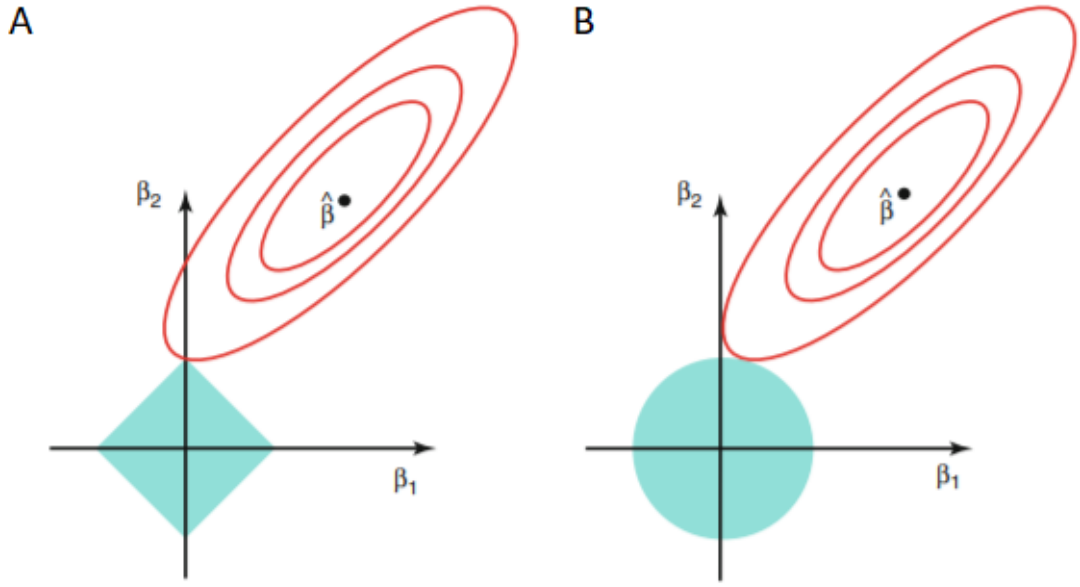
5 Harjuregression ja lasson vertailu

Verrataan seuraavaksi lassoa ja harjuregressiota mallinvalinnan näkökulmasta. Kun $p = 2$, lasson tuottamilla regressiokertoimilla on pienin jäännösneliösumma pisteillä, jotka osuvat yhtälön $|\beta_1| + |\beta_2| \leq t$ muodostamalle vinoneliön muotoiselle alueelle. Samanlaisesti harjuregression parametreilla on pienin jäännösneliösumma yhtälön $\beta_1^2 + \beta_2^2 \leq t$ muodostamalla ympyrällä. Tätä tilannetta havainnollistetaan kuvassa 1 [7]. Kun t on tarpeeksi suuri, kutistamismenetelmien rajoitteet sisältävät PNS-estimaatin $\hat{\boldsymbol{\beta}}$ ja näin menetelmien tuloksiksi tulee PNS-estimaatit. Ellipsit estimaatin $\hat{\boldsymbol{\beta}}$ ympärillä kuvaavat jäännösneliösummaa niin, että samalla ellipsillä olevilla pisteillä on sama jäännösneliösumman arvo. Jäännösneliösumman ellipsin ja regression rajoitteen ensimmäinen leikkauspiste antaa lasson ja harjuregression parametrien estimaatit. Koska harjuregression rajoite on ympyrä, leikkauspiste ei yleensä esiinny akselilla, joten harjuregressio ei voi saada estimaatiksi nollaa. Lasson rajoitteella on kulmat jokaisella akselilla, joten ellipsin ja rajoitteen leikkauspiste osuu usein akselille. Tällöin parametrin estimaatti saa arvokseen täysin nollan ja muuttuja tippuu pois mallista. Korkeissa dimensioissa moni parametri saa samanaikaisesti arvokseen nollan. Kuvassa 1 leikkauspiste osuu $\beta_1 = 0$ pisteeseen, joten malliin valikoituu vain parametria β_2 vastaava selittävä tekijä. Lasson etu harjuregressioon verrattuna on siis mallin yksinkertaistuminen. [7]

Lassolla ja harjuregressiolla on hyvin samanlainen ennustetarkkuus. Voisi kuvitella, että lasso muodostaisi paremman mallin tilanteessa, jossa vain muutamalla selittävällä tekijällä on merkittävä kerroin ja muiden parametrien arvo on nolla tai lähes nolla. Harjuregressio toimii paremmin, kun vaste on monen muuttujan funktio ja parametrit saavat hyvin samanarvoiset estimaatit. Mallien välinen paremmuus on kuitenkin aina aineistosta riippuvainen. Ristiinvalidoinnilla voidaan tutkia, kumpi lähestymistapa sopii aineistoon paremmin. Lisäksi molemmat kutistamismenetelmät ovat laskennaltaan yhtä tehokkaita. Molemmat tavat voidaan ratkaista samalla työmäärällä kuin PNS-estimoinnilla suoritettu malli. [7]

Vertaillaan vielä lasson ja harjuregression kutistamistapaa yksinkertaisessa erityistilanteessa. Oletetaan, että $n = p$ ja mallimatriisi $\boldsymbol{X}$ on diagonaalimatriisi, joka saa diagonaalilla arvon yksi ja muualla arvon nolla. Tällöin mallimatriisi $\boldsymbol{X}$ on myös ortonormaalinen. Lisäksi muodostetaan regressiomallit ilman vakiotermiä. PNS-menetelmä yksinkertaistuu lausekkeen

$$\sum_{j=1}^{p} (y_j - \beta_j)^2$$



Kuva 1: PNS-estimaatti on merkitty kuviin $\hat{\boldsymbol{\beta}}$. Punaiset ellipsit kuvaavat jäännösneliösumman tasa-arvokäyriä. Kuva A kuvastaa lasson tilannetta ja sininen vinoneliön alue kuvastaa lassorajoitetta $|\beta_1| + |\beta_2| \leq t$. Kuva B kuvastaa harjuregression tilannetta ja sininen ympyrä kuvastaa harjuregression rajoitetta $\beta_1^2 + \beta_2^2 \leq t$.

minimoitumiseen. Ratkaisuksi saadaan $\hat{\beta}_j = y_j$. Harjuregression estimaatit saadaan minimoimalla lauseketta

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

ja lasson estimaatit saadaan minimoimalla lauseketta

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

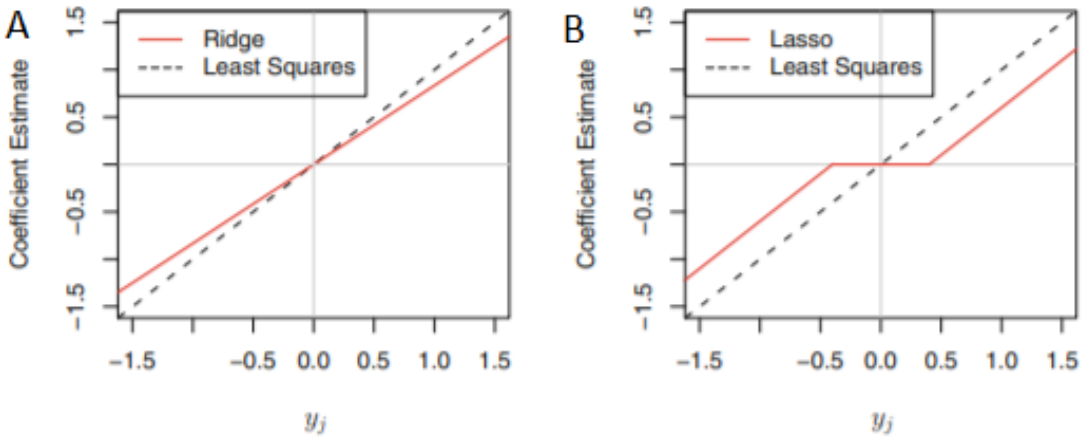
Harjuregression estimaateiksi saadaan

$$\hat{\beta}_j^{ridge} = \frac{y_j}{1+\lambda}$$

ja lasson ratkaisuksi saadaan

$$\hat{\beta}_j^{lasso} = \begin{cases} y_j - \lambda/2, & \text{jos } y_j > \lambda/2 \\ y_j + \lambda/2, & \text{jos } y_j < -\lambda/2 \\ 0, & \text{jos } |y_j| \leq \lambda/2. \end{cases}$$

Harjuregressio ja lasso kutistavat parametreja siis hyvin eri tavalla, mikä nähdään kuvasta 2 [7]. Harjuregression sakko kutistaa regressiokertoimien estimaatteja samassa suhteessa. Lasso taasen kutistaa jokaista parametria kohti nollaa saman vakion $\lambda/2$ verran. Kertoimet, joiden itseisarvo on pienempi kuin $\lambda/2$, kutistuvat nollaksi. Tätä tasaista kutistamista kutsutaan pehmeäksi kynnystämiseksi (soft-thresholding). Myös yleisen havaintomatriisin $\boldsymbol{X}$ tapauksessa havaittu ero kutistamisessa säilyy. Harjuregressio kutistaa jokaista aineiston muuttujaa samalla suhteella, kun taas lasso kutistaa parametreja saman verran ja suhteellisen pienet parametrit kutistuvat nollaksi.[7]



Kuva 2: Kuvassa A nähdään harjuregression ja kuvassa B nähdään lasson kutistamat estimaatit tilanteessa, kun $p = n$ ja mallimatriisi $\boldsymbol{X}$ on ortonormaalinen diagonaalimatriisi. Rajoittamaton PNS-estimaatti on piirretty katkoviivalla.

6 Elastinen verkko ja siltaregressio

Tarkastellaan seuraavaksi lyhyesti kahta kutistamismenetelmää, jotka yhdistävät lasson ja harjuregression sakkofunktiot ja hyödyt.

6.1 Elastinen verkko

Elastinen verkko yhdistää lasson ja harjuregression hyödyt ja se yhdistääkin sekä L_1 että L_2 normin mukaiset sakkofunktiot. Elastisella verkolla on myös ryhmittelevä vaikutus. Jos osan muuttujien välillä on korkeaa korrelaatiota, elastinen verkko ryhmittelee nämä korreloivat muuttujat yhteen ryhmään.[2]

Määritelmä 6.1. Elastinen verkko standardisoitujen muuttujien tapauksessa on

$$\begin{aligned} \min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^q} R_\lambda(\beta_0,\boldsymbol{\beta}) &= \min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^q} \left[\frac{1}{2n}\sum_{i=1}^{n}(y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta})^2+\lambda P_\alpha(\boldsymbol{\beta})\right], \\ \text{jossa} \quad P_\alpha(\boldsymbol{\beta}) &= \sum_{j=1}^{p}\left[(1-\alpha)\frac{1}{2}\beta_j^2+\alpha|\beta_j|\right]. \end{aligned} \tag{25}$$

Sakkofunktio P_α yhdistää siis harjuregression ja lasson sakkofunktion. Kun $\alpha = 0$ kyseessä on harjuregression estimointi ja kun $\alpha = 1$ kyseessä on lasson estimointi. Sakkofunktio P_α on erityisen hyödyllinen tilanteissa, joissa $p > n$ tai muuttujien välillä on voimakasta korrelaatiota. Elastinen verkko voidaan ratkaista laskevien koordinaattien menetelmällä (coordinate decent). [10]

6.2 Siltaregressio

Tässä luvussa esitellään lyhyesti siltaregressio (bridge regressio.) Siltaregressio käyttää sakkofunktiona luvussa 2.4 esiteltyä L_q $(q > 0)$ normin mukaista funktiota ja on näin harjuregression $(q = 2)$ ja lasson $(q = 1)$ yleistys. Siltaregressio sopii tilanteisiin, joissa tarvitaan muuttujan valintaa ja multikollineaarisuuden korjausta. Siltaregressio tekee muuttujan valinnan kun $0 < q \leq 1$ ja se kutistaa kertoimia kun $q > 1$. Parametri q valitaan ristiinvalidoinnin avulla. Sakkofunktio on konveksi kun $q > 1$ ja ei ole konveksi kun $q < 1$. Oletetaan, että vasteet y_i on keskistetty ja selittävät muuttujat $\boldsymbol{x}_i$ on standardisoitu.

Määritelmä 6.2. Parametrien estimointi siltaregression tapauksessa on

$$\hat{\boldsymbol{\beta}}^{bridge} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right], \tag{26}$$

jossa n on havaintojen lukumäärä ja p on parametrien eli muuttujien lukumäärä. [2]

Todetaan vielä lopuksi, että siltaregressio toteuttaa oracle-ominaisuudet (ks. Määritelmä 4.3) kun $0 < q < 1$, mutta kun $q < 1$ siltaregression ratkaisu ei ole jatkuva. Tämän takia L_1 mukaista sakkofunktiota suositaan enemmän kuin $L_q (q < 1)$ mukaista sakkofunktiota. [9]

7 Multikollineaarisuus

Muuttujien välinen korrelaatio pitää ottaa huomioon mallin valinnassa, koska erittäin vahva selittävien muuttujien korrelaatio eli multikollineaarisuus vaikuttaa mallin parametrien estimointiin ja parametrien varianssien kautta parametrien luottamusväleihin kuten todettiin kappaleessa 2.2. Parametrejä ei tällöin pystytä laskemaan yksikäsitteisesti ja niiden varianssi on suuri, jolloin luottamusväli levenee. Tällöin saadun mallin ennusteet ovat epävarmoja ja malli on heikko. Suuren positiivisen parametrin vaikutus voi poistua, jos samassa mallissa on vahvasti korreloitunut suuren negatiivisen vaikutuksen omaava parametri. Tätä ongelmaa voidaan lieventää rajoittamalla parametrien kokoa rajoitetun lineaarisen regression ja tasoitusparametrin avulla eli kutistamismenetelmien avulla.

Harjuregressio kutistaa korreloituneiden muuttujien kertoimia lähelle toisiaan ja sallii näin muuttujien lainata voimaa toisiltaan. Esimerkiksi jos mallissa on k kappaletta identtisiä muuttujia, muuttujat saavat identtiset kertoimet, jotka ovat $1/k$ osa parametrin koosta, jonka yksittäinen muuttuja saisi mallissa. Kaikki korreloituneet muuttujat kuitenkin jäävät malliin. [10]

Lasso käyttäytyy multikollineaarisuustilanteessa täysin eri tavalla. Lasso poimii korreloituneista muuttujista vain yhden ja kutistaa muiden muuttujien kertoimet nolliksi. Kun osan muuttujien regressiokertoimista tulee nollia, mallimatriisista tulee täysiasteinen vaikka alkuperäisessä havaintomatriisissa olisi identtisiä sarakkeita. Lasso toimiikin yhtenä multikollineaarisuuden korjausmenetelmänä, mutta siinä ei voi vaikuttaa, mikä korreloituneista muuttujista päätyy malliin. Lisäksi, jos korreloituneisuus pitäisi huomioida mallissa, lasso ei ole hyvä ratkaisu. Mukautuva lasso –menetelmässä voidaan painojen avulla soveltaa eri kutistusmääriä eri kertoimille. Näin saadaan asetettua jotkut muuttujat jäämään malliin ja huomioida korrelaatio käsin. Ryhmitetyn lasson avulla voidaan huomioida korreloituneiden muuttujien muodostamat ryhmät ja kutistaa parametreja joko yksittäin tai kutistaa koko muuttujien muodostamaa ryhmää. [10]

Elastinen verkko valinnalla $\alpha = 1-\epsilon$, kun $\epsilon > 0$ on jokin pieni luku, toimii samankaltaisesti kuin lasso ja se poistaa korrelaation aiheuttaman mallin heikkouden. Kun α nousee arvosta 0 arvoon 1, mallin muuttujien määrä lähenee monotonisesti lasson antamaa ratkaisua. [10]

8 Soveltaminen aineistoon

Tässä luvussa sovelletaan lassoa ja harjuregressiota lääketieteelliseen aineistoon. Luvun alussa esitellään aineisto ja tutkimuskysymys, jonka jälkeen tutkitaan aineistoa perinteisiä tilastollisia menetelmiä käyttämällä. Lopuksi vertaillaan PNS-estimoinnin, harjuregression ja lasson tuottamia malleja ja pohditaan mallien tuottamia tuloksia tutkimuskysymykseen. Soveltavan osion R-koodi on liitteenä E.

8.1 Aineisto

Aineisto on osa Turun yliopiston Hyvän kasvun avaimet -seurantatutkimusta (HKA). HKA -seurantatutkimus on monitieteinen tutkimus lasten hyvinvoinnista ja kehityksestä. Tutkimuksen syntymäkohortin (Southwest Finland Birth Cohort, SFBC) muodostavat Varsinais-Suomen sairaanhoitopiirin alueella vuosina 2008-2010 synnyttäneet äidit ($n = 13436$) ja heidän lapsensa ($n = 14946$). Näistä HKA-perusaineiston muodostavat 9811 äitiä ja 9946 lasta. Perusaineiston 1797 äitiä ja 1658 heidän puolisoaan lähti mukaan HKA:n seurantaryhmään. Lapsia seurantaryhmässä on 1827. Tutkimuksen äidit ja perheet on värvätty tutkimukseen raskauden aikana tai pian lapsen syntymän jälkeen.[11]

Tämän osatutkimuksen tarkoituksena on selvittää mitkä äidin ominaisuudet ovat yhteydessä äidin rintamaidon oligosakkaridipitoisuuksiin ja miten nämä sakkaridipitoisuudet ovat yhteydessä lapsen kasvuun ensimmäisien viiden elinvuosien aikana [3]. Tässä tutkielmassa keskitytään jälkimmäiseen tutkimuskysymykseen eli pyritään ennustamaan, mitkä rintamaidon oligosakkaridit ovat yhteydessä lapsen pituuteen ja painoon. Ennustamiseen käytetään lassoa ja harjuregressiota ja näiden tuloksia verrataan toisiinsa.

Rintamaitoruokinta tutkitusti parantaa vastasyntyneen terveyttä ja vaikuttaa lapsen terveyteen myöhemminkin elämässä muun muassa vähentämällä riskiä ylipainoisuuteen [12]. Äidin rintamaidon oligosakkaridit ovat glykaaneja eli hiilihydraattirakenteita, tarkemmin toisiinsa liittyneiden sokerimolekyylien muodostamia rakenteita, joita lapsi ei kykene käyttämään ravinnoksi. HMO -yhdisteitä esiintyy hyvin runsaasti äidin rintamaidossa ja niitä on rintamaidossa eniten verrattuna maitoproteiiniyhdisteisiin. HMO -yhdisteet ovat ravintoa suolistobakteereille, erityisesti terveyttä edistäville bifidobakteereille. Lisäksi on saatu selville, että HMO -molekyyleillä olisi vaikutusta infektioiden ehkäisyyn, immunijärjestelmän toimintaan ja niiden sisältämä siaalihappo olisi tärkeää aivojen kehitykselle. [13]

Tämän tutkimuksen aineistossa on 802 äitiä ja 810 lasta eli aineistossa esiintyy kahdeksan kaksosta. Kaikkiaan 812 äidiltä kerättiin rintamaitonäy-

te lapsen ollessa kolmen kuukauden ikäinen (lapsien iän keskiarvo näytteen ottohetkellä on 11.3 viikkoa ja keskihajonta 2.6 viikkoa). Kymmenen näytettä hylättiin teknisistä syistä. Näytteistä analysoitiin 19 eri oligosakkaridia (nmol/mL). Nämä oligosakkaridit esiintyvät rintamaidossa runsaimmin ja niiden massapitoisuus on yli 95% kaikista oligosakkarideista, joten ne edustavat kaikkia tiedettyjä HMO:n rakenteita. HMO-kokonaispitoisuus laskettiin 19 eri oligosakkaridien pitoisuuksien summana. Äidin geneettisesti määräytyvä kyky tehdä erilaisia HMO-yhdisteitä eli niin kutsuttu secretorstatus määriteltiin 2'-fucosyllactose (2FL) sakkaridin pitoisuudesta: korkean pitoisuuden omaavat äidit ovat secretor-tyyppiä ($n = 699$) kun taas matalan pitoisuuden omaavat äidit ovat nonsecretor-tyyppiä ($n = 103$). HMO-pitoisuudet ovat todella erilaisia secretorien ja nonsecretorien välillä, minkä takia aineisto päätettiin tutkia erikseen secretorstatuksen mukaan. Nonsecretor-äideillä kaikki HMO-pitoisuudet ovat yleisesti alhaisempia kuin secretoreilla.

Lapsen kasvuaineisto muodostettiin neuvoloista kerätystä aineistosta. Mittauspisteet 3, 6 ja 8 kuukautta sekä 1, 2, 3, 4 ja 5 vuotta poimittiin lähimmästä käyntikerrasta verrattuna lapsen ikään. Käyntikerta sai kuitenkin poiketa syntymäpäivästä korkeintaan yhden kuukauden verran. Pituudelle ja painolle laskettiin myös suomalaisten kasvukäyrien mukaiset z-score arvot. [14] Aineistossa on myös äidin ennen raskautta laskettu body mass index (BMI, kg/m^2), kolmiluokkainen lapsen synnytystapa (normaali vaginaalinen synnytys, avustettu synnytys, keisarileikkaus), lapsen sukupuoli (poika, tyttö) ja lapsen syntymäpainon z-score. Nämä muuttujat otetaan huomioon sekoittavina tekijöinä.

8.2 Alkutarkastelu

Poistetaan aineistosta aluksi kaksoset, koska kaksosten kasvu korreloi keskenään. Tarkastellaan seuraavaksi aineistoa ja tarkemmin oligosakkareiden ominaisuuksia. Kaikki käytettävät HMO-muuttujat ovat jatkuvia. Histogrammeista (liite A) nähdään, että oligosakkaridit eivät ole normaalijakautuneita. Lisäksi histogrammeista huomataan DFLac, DFLNT, LNFP I ja 2'FL - oligosakkaridien arvojen jakautuminen secretorstatuksen mukaan. Statuksen vaikutus nähdään myös HMO-muuttujien summan jakaumassa. Sakkaridit ovat myös hyvin erilaisia skaalaltaan. HMO-muuttujilla DFLNH, DSLNH, LNFP III ja LNH on poikkeavia arvoa. Poikkeavat arvot jätettiin aineistoon, koska tutkijoiden toimesta niiden arvot varmistettiin olevan oikeat.

Tilastolliset tunnusluvut (mediaani ja kvartiilit Q1 ja Q3) vahvistavat histogrammeista huomatut oligosakkaridien piirteet (taulukko 1). HMO - muuttujien tilastollisesti merkitsevä ero secretorstatuksen mukaan on tutkittu Wilcoxonin merkittyjen sijalukujen testillä, koska muuttujat poikkeavat

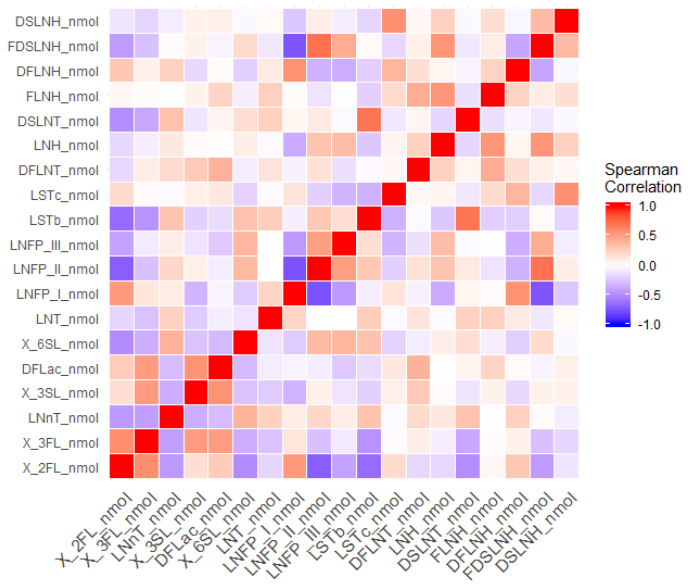
Muuttuja	Total	Secretor	Nonsecretor	P-arvo
HMO summa	16183 (15261, 17068)	16366 (15658, 17216)	9198 (8884, 9525)	<0.001
2'FL	6044 (4396, 7847)	6435 (4990, 8223)	46 (30, 109)	<0.001
3FL	346 (240, 487)	374 (274, 521)	120 (78, 169)	<0.001
LNnT	985 (780, 1286)	983 (781, 1272)	1071 (776, 1370)	0.32
3'SL	504 (386, 678)	523 (404, 705)	394 (320, 519)	<0.001
DFLac	539 (344, 680)	541 (404, 707)	9 (4, 15)	<0.001
6'SL	564 (396, 879)	527 (384, 777)	1187 (787, 1732)	<0.001
LNT	859 (609, 1161)	883 (631, 1172)	634 (369, 1016)	<0.001
LNFP I	1153 (637, 1743)	1254 (851, 1848)	77 (39, 110)	<0.001
LNFP II	1548 (1103, 2044)	1442 (1055, 1842)	3083 (2787, 3501)	<0.001
LNFP III	74 (54, 104)	71 (52, 93)	146 (104, 195)	<0.001
LSTb	112 (81, 154)	108 (78, 147)	162 (127, 210)	<0.001
LSTc	74 (51, 105)	77 (53, 110)	60 (41, 82)	<0.001
DFLNT	1489 (946, 1847)	1576 (1230, 1890)	586 (436, 788)	<0.001
LNH	58 (37, 84)	58 (37, 83)	59 (38, 98)	0.22
DSLNT	322 (227, 446)	318 (224, 444)	377 (261, 455)	0.018
FLNH	52 (29, 83)	56 (35, 88)	23 (14, 45)	<0.001
DFLNH	38 (25, 51)	41 (29, 53)	13 (9, 20)	<0.001
FDSLNH	238 (157, 368)	220 (147, 320)	574 (391, 740)	<0.001
DSLNH	68 (45, 102)	67 (44, 98)	87 (59, 119)	<0.001

Taulukko 1: Äidin rintamaidon oligosakkaridien (nmol/mL) mediaanit ja kvartiilit (Q1, Q3) on esitetty kaikki yhdessä ja erikseen äidin secretor-statuksen mukaan. Lisäksi esitetään Wilcoxonin testin p-arvo secretorien ja nonsecretorien väliselle erolle.

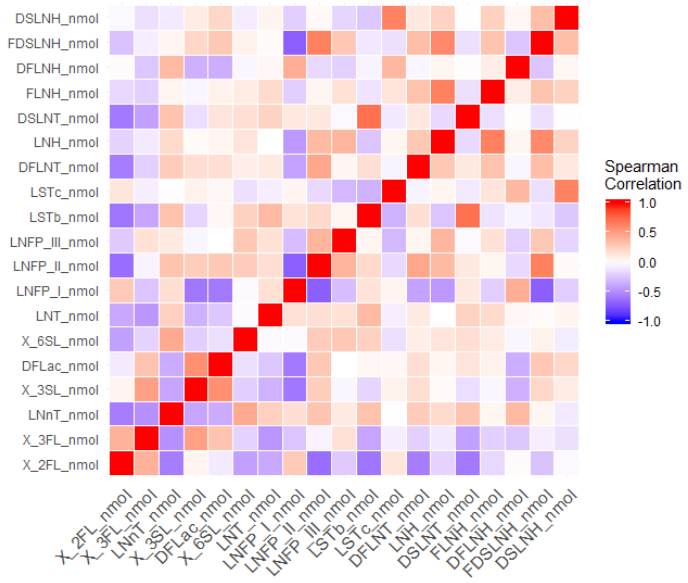
normaalijakaumasta. Nonsecretor-äideillä kokonais HMO -pitoisuus on pienempi kuin secretor-äideillä (p-arvo < 0.001). Useat HMO -pitoisuudet eroavat tilastollisesti merkitsevästi secretorstatuksen mukaan, minkä takia tulevat tilastolliset tarkastelut tehdään erikseen secretoreille ja nonsecretoreille. Vain oligosakkaridit LNnT (p-arvo $= 0.32$) ja LNH (p-arvo $= 0.22$) eivät eroa tilastollisesti merkitsevästi riippuen secretorstatuksesta. Tunnusluvuista huomataan myös muuttujien erilainen skaalautuminen. Osan oligosakkareiden pitoisuudet ovat tuhansia nmol/mL (esimerkiksi 2'FL ja LNFP II) kun taas muutaman pitoisuudet liikkuvat muutamassa kymmenessä nmol/mL (esimerkiksi LNH ja FLNH). Tämän vuoksi muuttujat standardoidaan ennen tilastollisien mallien sovittamista.

Tarkastellaan seuraavaksi oligosakkareiden välisiä korrelaatioita Spearmanin järjestysasteikkokorrelaation avulla. Korrelaatiot on kuvattuna lämpökarttoina (heatmap) kuvissa 3, 4 ja 5 ja lukuarvoina liittessä B. Koko aineistosta lasketuista Spearmanin korrelaatioista huomataan, että oligosakkaridien 2'FL, LNFP I, LNFP II ja FDSLNH välillä löytyy korkeaa korrelaatiota ($|\rho| > 0.7$). Secretor-äitien joukossa oligosakkaridien LSTb ja DSLNT välillä löytyy korkeaa korrelaatiota ($|\rho| > 0.7$). Secretoreilla korrelaatioita löytyy muidenkin oligosakkaridien välillä, mutta ne eivät ole kovinkaan korkeita. Nonsecretoreilla vahvoja korrelaatioita ei ole havaittavissa ja muutenkin korrelaatio on vähäistä. Multikollineaarisuutta ei ole havaittavissa oligosakkaridien välillä.

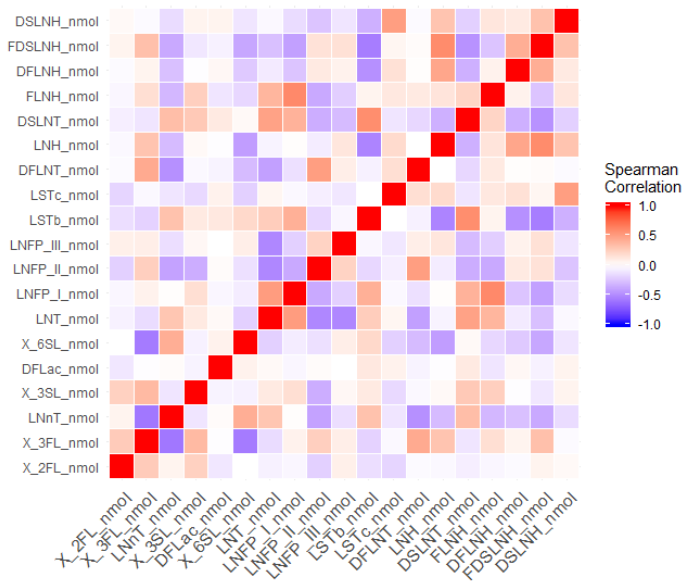
Sekoittavien tekijöiden eli lapsen sukupuolen, synnytystavan, syntymäpainon z-scoren ja äidin BMI:n välillä ei huomattu olevan eroa secretorien ja nonsecretorien välillä (taulukko 2). Lisäksi äidin secretorstatus ei vaikuta tilastollisesti merkitsevästi lapsen kokoon yhden vuoden iässä. Alkutarkastelun lopuksi tutkitaan lapsen yksivuotispituuden ja -painon z-scoren korrelaatiota oligosakkareihin (liite C). Kaikki havaittavat korrelaatiot ovat todella pieniä.



Kuva 3: Heatmap oligosakkaridien välisistä Spearmanin korrelaatiosta koko aineistolle. Negatiivinen korrelaatio on kuvattu sinisen sävyillä ja positiivinen korrelaatio punaisen sävyillä.



Kuva 4: Heatmap oligosakkaridien välisistä Spearmanin korrelaatiosta secretoreille. Negatiivinen korrelaatio on kuvattu sinisen sävyillä ja positiivinen korrelaatio punaisen sävyillä.



Kuva 5: Heatmap oligosakkaridien välisistä Spearmanin korrelaatiosta nonsecretoreille. Negatiivinen korrelaatio on kuvattu sinisen sävyillä ja positiivinen korrelaatio punaisen sävyillä.

Muuttuja	Total	Secretor	Nonsecretor	P-arvo
Sukupuoli (pojat, %)	54%	54%	57%	0.61
Synnytystapa (sektio, %)	12%	11%	14%	0.57
Syntymäpainon z-score	-0.014 (-0.669, 0.688)	-0.015 (-0.686, 0.695)	-0.002 (-0.633, 0.649)	0.45
Äidin BMI*	23 (21, 26)	23 (21, 26)	24 (21, 26)	0.94
Pituus z-score 1v	-0.081 (-0.717, 0.639)	-0.076 (-0.698, 0.657)	-0.167 (-0.865, 0.495)	0.21
Paino z-score 1v	-0.144 (-0.812, 0.520)	-0.135 (-0.812, 0.565)	-0.265 (-0.803, 0.227)	0.16

Taulukko 2: Aineiston sekoittavia tekijöitä ovat lapsen sukupuoli, synnytystapa, syntymäpainon z-score ja äidin ennen raskautta laskettu BMI. Lisäksi esitetään lapsen pituuden ja painon z-score yhden vuoden iässä. Jatkuvien muuttujien arvot on ilmoitettu taulukossa mediaaneina ja kvartiileina (Q1, Q3). Kategoriset muuttujat on ilmoitettu prosentteina. Jatkuvien muuttujien tilastollinen ero secretorstatuksen välillä on tutkittu Studentin t-testillä ja kategoristen muuttujien eroa on tutkittu khiin neliötestillä. *Äidin BMI:n tilastollinen ero on tutkittu Wilcoxonin testillä, koska muuttuja ei noudattanut normaalijakaumaa.

8.3 Mallien estimointi

Keskitytään tutkimaan lapsen painon z-scorea yhden vuoden iässä. Ennen mallien sovittamista standardoidaan jatkuvat muuttujat. Vasteena toimiva painon z-score pidetään alkuperäisessä muodossaan ja lisäksi syntymäpainon z-score pidetään alkuperäisessä muodossa. Kategoriset muuttujat muokataan dummy-muuttujiksi. Nämä tehdään R-ohjelman funktioiden *scale* ja *dummy_cols* avulla. Lisäksi poistetaan puuttuvat havainnot (n=54) *complete.cases* funktion avulla. Havaintojen puuttuminen johtuu liian aikaisesta tai myöhäisestä neuvolakäynnistä verrattuna lapsen syntymäpäivään. Tämän jälkeen jaetaan sattumanvaraisesti aineisto puoliksi opetus- ja testiaineistoon R-ohjelman *sample* -funktion avulla. Mallien muodostamisessa ei siis käytetä erillistä validointiaineistoa. Jako tehdään koko aineistolle, secretoreille ja nonsecretoreille erikseen. Koko aineistolle varmistetaan, että secretor ja nonsecretor -äidit ovat jakaantuneet tasaisesti opetus- ja testiaineiston kesken. Lisäksi varmistetaan, että kategoristen muuttujien frekvenssit jakautuvat tasaisesti opetus- ja testiaineiston kesken.

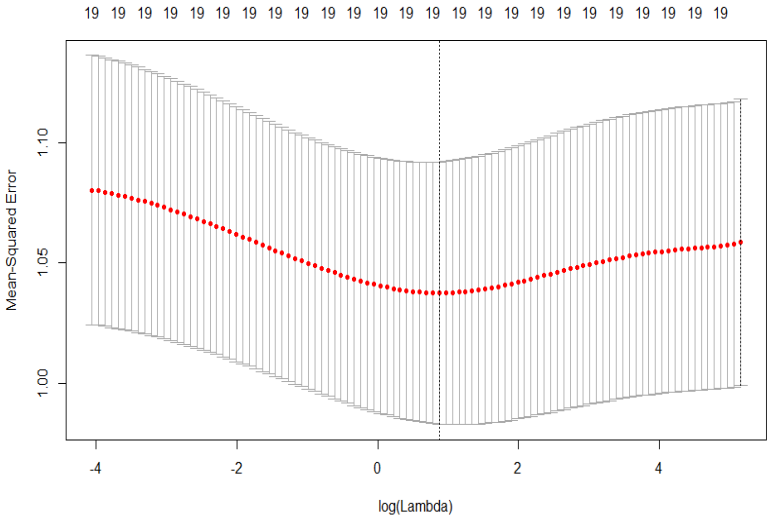
Lineaarinen malli PNS-estimoinnilla muodostetaan R-ohjelman *lm* -funktion avulla. Ennusteiden muodostamiseen käytetään *predict.lm* -funktiota. Malli muodostetaan opetusaineistolla, ennusteet tehdään testiaineistolla ja lopuksi lasketaan testiaineistolla keskineliövirhe, jota käytetään testi- ja ennustevirheenä. Lopullinen malli tehdään koko aineistolla ja secretoreille ja nonsecretoreille erikseen. Opetus- ja testiaineistolla valitaan siis aineistoon sopivin malli, jonka estimointitulokset lopulta esitetään käyttäen estimointiotoksena koko aineistoa. Mallit on muodostettu kaksi kertaa: kun selittävinä tekijöinä on vain oligosakkaridit (taulukot 3, 4 ja 5) ja kun muutkin selittävät tekijät eli lapsen sukupuoli, syntymäpainon z-score, synnytystapa ja äidin BMI ovat mukana (liite D).

Harjuregression ja lasson mallien muodostamisessa käytetään R-ohjelman *glmnet*-pakettia ja siitä funktiota *glmnet*. *Glmnet* -funktiossa on argumenttina α. Kun $\alpha = 1$ funktio muodostaa lassoregression ja kun $\alpha = 0$ funktio muodostaa harjuregression (vertaa elastiseen verkkoon luvussa 6). *Glmnet* -funktio standardisoi muuttujat automaattisesti, jotta ne ovat samalla skaalalla. Koska muuttujat standardisoitiin lineaarista mallia varten, asetetaan tämä ehto pois komennolla $standardize = FALSE$. Ennen mallin muodostamista selvitetään paras tasoitusparametrin λ arvo k-ristiinvalidoinnin avulla. Määritellään parhaaksi λ arvoksi se, jonka malli muodostaa pienimmän keskineliövirheen. Tässä käytetään opetusaineistoa ja funktiota *cv.glmnet*. Ristiinvalidoinnissa käytetään funktion oletusarvoa parametrille k eli $k = 10$. Tällöin aineisto jaetaan kymmeneen osaan ristiinvalidointia varten. Tämän jälkeen ensin muodostetaan malli opetusaineistolla, tehdään ennusteet tes-

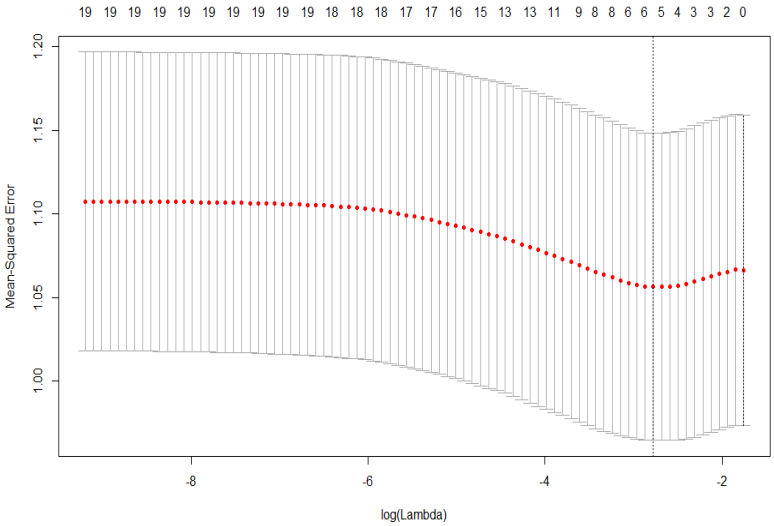
tiaineistolla ja lasketaan keskineliövirhe. Lopuksi vielä muodostetaan malli koko aineistolla ja käytetään ristiinvalidoinnilla saatua parasta tasoitusparametrin λ arvoa. Harjuregressiolla ja lassolla muodostetaann samat mallit kuin PNS-estimoinnilla. Tarkastelun lopuksi muodostetaan vielä PNS-estimoinnilla lineaariset mallit, joissa on lasson valitsemat muuttujat. [7, 15]

Tarkastellaan tarkemmin 10-kertaisen ristiinvalidoinnin tuloksia, jotka näkyvät kuvissa 6, 7 ja 8. Koko aineistolla harjuregression ristiinvalidointi tuottaa tasoitusparametrin λ arvoksi 2.40. Secretor-aineistolla λ saa arvon 3.83 ja nonsecretor-aineistolla 264. Vastaavat luvut lasson tapauksessa ovat 0.0623, 0.0325 ja 0.264. Lasson tasoitusparametrin arvot ovat siis paljon pienempiä kuin harjuregression. Koko aineiston ja secretor-aineiston ristiinvalidoinnin tulokset kehittyvät samankaltaisesti. Nonsecretor-aineistolla keskineliövirheen arvo lähtee nopeasti laskuun ja tasoitusparametrin λ arvo onkin isompi verrattuna koko aineiston tai secretor-aineiston tasoitusparametriin sekä lasson että harjuregression tapauksessa.

Tarkastellaan vielä kuvassa 9 harjuregression ja lasson mallien kehittymistä koko aineistolla. Harjuregression parametrien estimaatit kutistuvat tasaisesti kohti nollaa, kun tasoitusparametrin λ arvo kasvaa. Lasson mallissa jotkin parametrien estimaatit kutistuvat nopeasti nollaksi ja tippuvat mallista. Pystyviivalla on merkitty ristiinvalidoinnilla saatu paras tasoitusparametrin λ arvo. Myös kuvasta huomataan, että harjuregression mallissa ovat mukana kaikki 19 oligosakkaridia, kun taas lasson mallissa mukana on vain osa muuttujista.

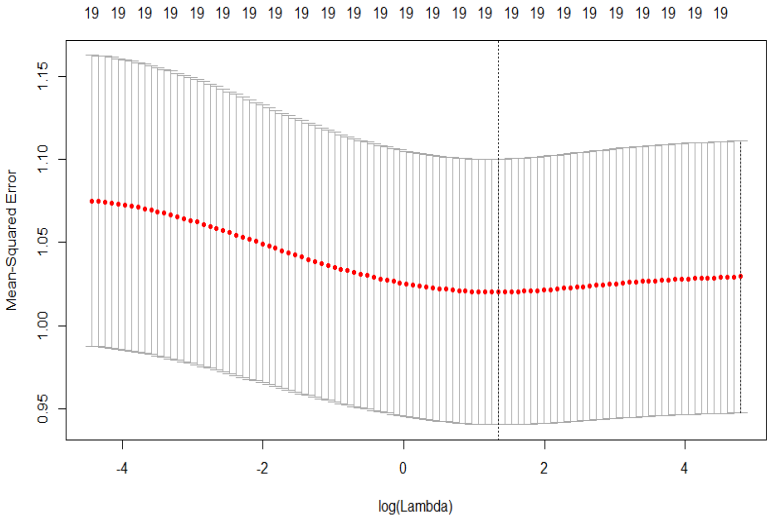


(a) Harjuregressio, koko aineisto

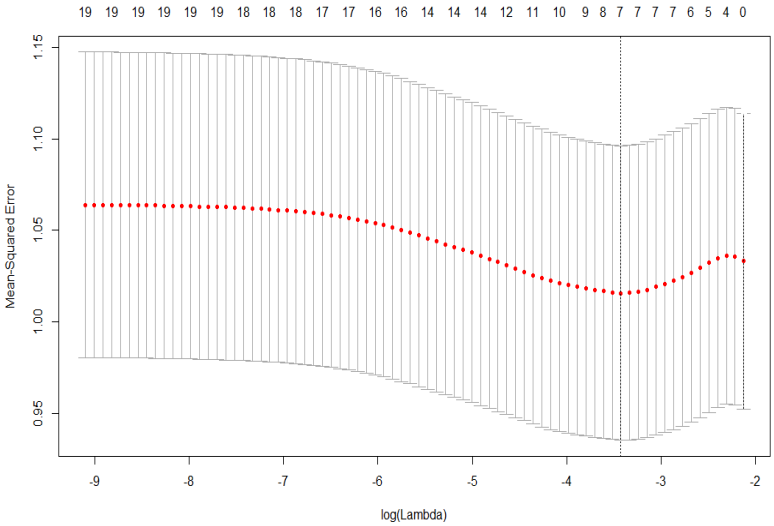


(b) Lasso, koko aineisto

Kuva 6: Kuvassa (a) on harjuregression ja (b) lasson 10-ristiinvalidoinnit parhaan tasoitusparametrin λ arvon löytämiselle koko aineiston opetusaineistossa. Y-akselilla on keskineliövirhe ja x-akselilla λ arvoja log-asteikolla. Pienimmän keskineliövirheen omaava λ on merkitty katkoviivalla. Mallin parametrien lukumäärä näkyy kuvan yläreunassa. Punainen käyrä kuvaa keskineliövirheen arvoja ja harmaat viivat kuvaavat arvon keskivirhettä.

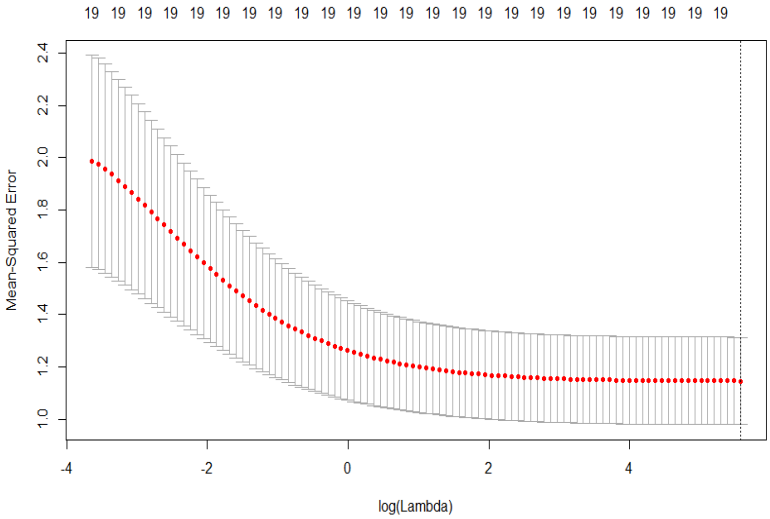


(a) Harjuregressio, secretor

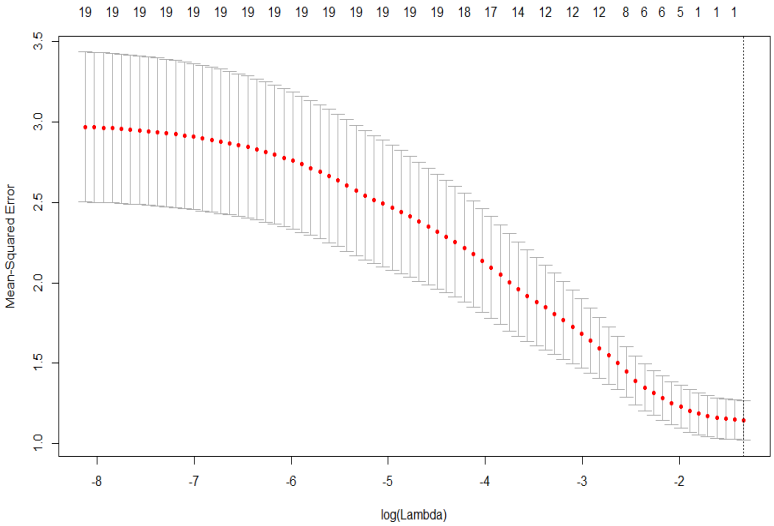


(b) Lasso, secretor

Kuva 7: Kuvassa (a) on harjuregression ja (b) lasson 10-ristiinvalidoinnit parhaan tasoitusparametrin λ arvon löytämiselle secretor-aineiston opetusaineistossa. Y-akselilla on keskineliövirhe ja x-akselilla λ arvoja log-asteikolla. Pienimmän keskineliövirheen omaava λ on merkitty katkoviivalla. Mallin parametrien lukumäärä näkyy kuvan yläreunassa. Punainen käyrä kuvaa keskineliövirheen arvoja ja harmaat viivat kuvaavat arvon keskivirhettä.

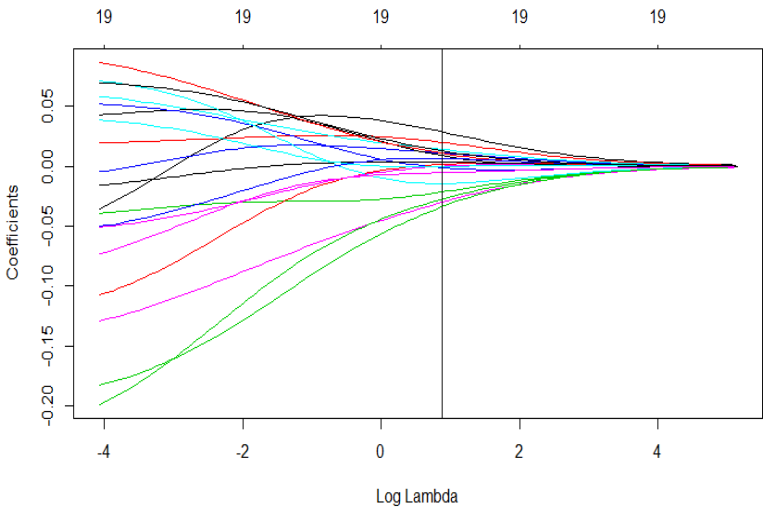


(a) Harjuregressio, nonsecretor

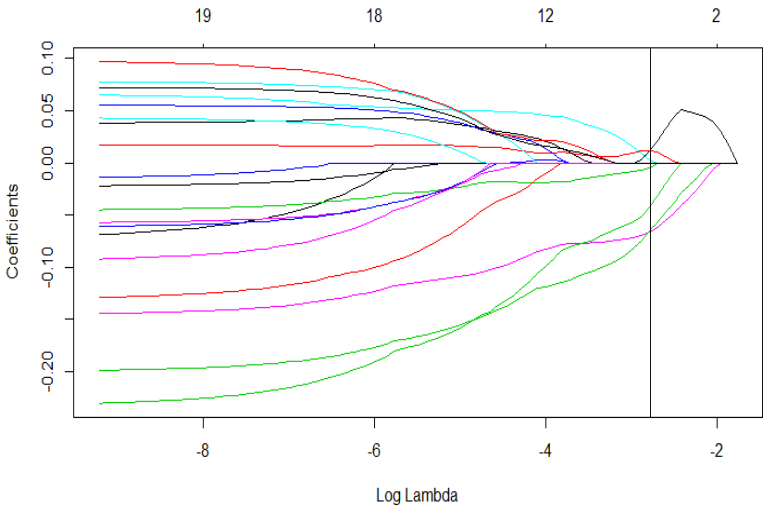


(b) Lasso, nonsecretor

Kuva 8: Kuvassa (a) on harjuregression ja (b) lasson 10-ristiinvalidoinnit parhaan tasoitusparametrin λ arvon löytämiselle nonsecretor-aineiston opetusaineistossa. Y-akselilla on keskineliövirhe ja x-akselilla λ arvoja log-asteikolla. Pienimmän keskineliövirheen omaava λ on merkitty katkoviivalla. Mallin parametrien lukumäärä näkyy kuvan yläreunassa. Punainen käyrä kuvaa keskineliövirheen arvoja ja harmaat viivat kuvaavat arvon keskivirhettä.



(a) Harjuregressio



(b) Lasso

Kuva 9: Kuvassa (a) on harjuregression ja (b) lasson mallien kehittyminen koko aineistolla kun tasoitusparametrin λ arvo muuttuu. X-akselilla on tasoitusparametrin λ luonnollisen logaritmin arvoja ja y-akselilla parametrien estimaattien saama arvo. Kuvan yläreunassa on pisteen λ muodostaman mallin parametrien lukumäärä. Viivalla on merkitty parhaan tasoitusparametrin λ arvo eli ristiinvalidoinnilla saatu arvo, jolla mallin keskineliövirhe minimoitui.

8.4 Mallien tulokset ja niiden vertailu

Tässä luvussa tarkastellaan lineaarisen mallin (PNS-malli), harjuregression ja lasson tuottamia malleja ja vertaillaan niiden tuloksia. Mallien vasteena toimii siis painon z-score lapsen yhden vuoden iässä ja selittävinä tekijöinä 19 eri oligosakkaridia. Mallit on muodostettu erikseen koko aineistolle, secretor äideille ja nonsecretor äideille. Mallien tulokset näkyvät taulukoissa 3, 4 ja 5. Lisäksi taulukossa on lineaarisen mallin, jonka mallin valinta on tehty lassolla, tulokset (PNS-Lasso).

PNS-estimoinnilla muodostetussa mallissa on mukana kaikki 19 oligosakkaridia. Näistä vain muutama on tilastollisesti merkitsevä painon z-scoren kannalta. Koko aineistolla vain kolme oligosakkaridia saa p-arvokseen $p < 0.1$, secretor aineistolla kaksi oligosakkaridia ja nonsecretor aineistolla vain yksi oligosakkaridi. Kun malliin lisätään muut selittävät tekijät, sukupuoli, syntymäpainon z-score, äidin BMI ja synnytystapa, oligosakkaridien merkitsevyys pienenee tai katoaa kokonaan. Hyvin moni mallissa oleva oligosakkaridi ei siis ole vasteen kannalta merkityksellinen ja voitaisiin näin poistaa mallista. Oligosakkareista DFLac, DSLNT, LNFP II, LSTc, 3FL ja DSLNH vaikuttavat aineiston kannalta merkittäviltä oligosakkareilta. Varsinki DSLNT tulee merkitseväksi monessa mallissa. Muuttujien estimaattien keskihajonnat eivät vaikuta kovinkaan suurilta.

Harjuregression muodostamassa mallissa on myös kaikki 19 oligosakkaridia mukana. Näin huomataan käytännössä, että harjuregressio ei kutista parametreja täysin nollaksi. Koko aineiston mallissa harjuregressio on kutistanut 16 parametrin estimaattia lähemmäksi nollaa verrattuna PNS-estimoituun malliin. Secretorien mallissa tasoitusparametrin λ arvo on hieman suurempi ja malli on kutistanut 18 parametrin estimaattia kohti nollaa. Nonsecretoreilla tasoitusparametrin arvo on hyvin paljon suurempi kuin muissa malleissa ja kaikkia parametrin estimaatteja onkin kutistettu hyvin paljon muihin malleihin verrattuna. Sama ilmiö on nähtävissä myös liitteen D malleissa.

Taulukoista huomataan myös, että lassolla muodostetussa mallissa on mukana paljon vähemmän muuttujia kuin PNS- tai harjuregressiomallissa. Koko aineiston mallissa lasso valitsee malliin oligosakkaridit 3FL, DFLac ja DSLNT, joista DFLac ja DSLNT tulivat PNS-mallissa tilastollisesti merkitseviksi. Secretor-mallissa lasson malliin on valikoitunut hieman enemmän muuttujia. Lasson muodostamassa mallissa ovat oligosakkaridit LNnT, DFLac, LNT, DFLNT, DSLNT ja DSLNH, joista DFLac ja DSLNT ovat tilastollisesti merkitseviä PNS-estimoidussa mallissa. Nonsecretoreilla lasso muodostaa mallin, joka sisältää pelkästään vakiotermin. Koska tasoitusparametrin λ arvo on ollut suuri, kaikki muuttujat ovat tippuneet pois mal-

Muuttuja	PNS-malli	Harjuregressio	Lasso	PNS-Lasso
λ		2.40	0.0623	
2'FL	-0.0721 (0.105)	0.0131		
3FL	-0.00293 (0.0626)	0.0163	0.00835	0.00235 (0.0501)
LNnT	-0.00970 (0.0533)	-0.0126		
3'SL	-0.00515 (0.0689)	0.0154		
DFLac	0.144 (0.0697)	0.0205	0.0424	0.129 (0.0505)
6'SL	-0.0361 (0.0591)	-0.00938		
LNT	-0.0141 (0.0467)	-0.00890		
LNFP I	-0.0665 (0.0819)	-0.00105		
LNFP II	-0.154 (0.0925)	-0.0138		
LNFP III	0.0285 (0.0533)	-0.00142		
LSTb	0.0710 (0.0633)	-0.00726		
LSTc	-0.0138 (0.0542)	0.00423		
DFLNT	-0.0611 (0.0494)	-0.00859		
LNH	0.00248 (0.0580)	0.000792		
DSLNT	-0.184 (0.0606)	-0.0247	-0.0470	-0.131 (0.0412)
FLNH	-0.0258 (0.0538)	0.000847		
DFLNH	0.0565 (0.0520)	0.00312		
FDSLNH	0.0113 (0.0838)	0.00270		
DSLNH	0.0313 (0.0558)	0.00918		
MSE	0.965	0.907	0.911	0.895

Taulukko 3: Estimoitujen mallien tulokset koko aineistolle, kun vain oligosakkaridit ovat mukana. Taulukossa on selittävien muuttujien regressiokertoimien estimaatit (keskivirheet), testiaineistojen keskineliöennustevirheet (MSE) ja tasoitusparametrin λ arvo harjuregressiolle ja lassolle. PNS-estimoidussa mallissa muuttujat DFLac ($p = 0.039$) ja DSLNT ($p = 0.002$) saivat tilastollisesti merkitsevät p-arvot. Oligosakkaridi LNFP II on merkitsevyyden rajalla ($p = 0.097$). Vähäparametrisen PNS-estimoidun mallin oligosakkareista DFLac ($p = 0.011$) ja DSLNT ($p = 0.002$) saivat tilastolliset merkitsevät p-arvot.

Muuttuja	PNS-malli	Harjuregressio	Lasso	PNS-Lasso
λ		3.83	0.0325	
2'FL	-0.0279 (0.193)	0.0101		
3FL	-0.0420 (0.0687)	0.0126		
LNnT	-0.0244 (0.0735)	-0.0134	-0.0105	-0.00107 (0.0540)
3'SL	-0.00592 (0.0745)	0.0122		
DFLac	0.174 (0.0883)	0.0153	0.0844	0.137 (0.0547)
6'SL	-0.0183 (0.0819)	-0.00497		
LNT	-0.0712 (0.0688)	-0.0150	-0.0281	-0.0441 (0.0465)
LNFP I	0.00545 (0.121)	-0.00150		
LNFP II	-0.0268 (0.138)	-0.00531		
LNFP III	0.0650 (0.0718)	0.000267		
LSTb	0.0639 (0.0724)	-0.00765		
LSTc	-0.0229 (0.0619)	0.00660		
DFLNT	-0.0340 (0.0900)	-0.00680	-0.00762	-0.0524 (0.0454)
LNH	-0.00973 (0.0713)	-0.00292		
DSLNT	-0.201 (0.0695)	-0.0200	-0.0931	-0.133 (0.0465)
FLNH	-0.0178 (0.0665)	-0.00187		
DFLNH	0.0478 (0.0588)	0.00109		
FDSLNH	-0.0687 (0.127)	-0.000290		
DSLNH	0.0767 (0.0664)	0.00955	0.00560	0.0378 (0.0410)
MSE	1.02	0.977	0.977	0.954

Taulukko 4: Estimoitujen mallien tulokset secretor-aineistolle, kun vain oligosakkaridit ovat mukana. Taulukossa on selittävien muuttujien regressiokertoimien estimaatit (keskivirheet), testiaineistojen keskineliöennustevirheet (MSE) ja tasoitusparametrin λ arvo harjuregressiolle ja lassolle. PNS-estimoidussa mallissa muuttujat DFLac ($p = 0.050$) ja DSLNT ($p = 0.004$) saivat tilastollisesti merkitsevät p-arvot. Vähäparametrisen PNS-estimoidun mallin oligosakkareista DFLac ($p = 0.012$) ja DSLNT ($p = 0.004$) saivat tilastolliset merkitsevät p-arvot.

Muuttuja	PNS-malli	Harjuregressio	Lasso
λ		264	0.264
2'FL	-3.40 (4.45)	-0.00000463	
3FL	-0.0156 (0.372)	0.0000886	
LNnT	-0.200 (0.211)	0.0000241	
3'SL	-0.233 (0.294)	0.0000298	
DFLac	5.31 (4.31)	0.00000642	
6'SL	-0.0864 (0.190)	-0.000363	
LNT	0.0449 (0.201)	0.000593	
LNFP I	-3.62 (2.82)	0.00000282	
LNFP II	-0.469 (0.396)	-0.000378	
LNFP III	-0.00760 (0.0911)	-0.0000476	
LSTb	0.213 (0.162)	0.0000177	
LSTc	-0.303 (0.178)	-0.000298	
DFLNT	-0.0876 (0.496)	-0.0000488	
LNH	0.122 (0.142)	0.000459	
DSLNT	-0.120 (0.216)	-0.0000444	
FLNH	0.0503 (0.135)	0.000325	
DFLNH	0.0310 (0.140)	0.0000686	
FDSLNH	0.00630 (0.235)	0.000437	
DSLNH	-0.0898 (0.137)	-0.000127	
MSE	1.15	0.476	0.484

Taulukko 5: Estimoitujen mallien tulokset nonsecretor-aineistolle, kun vain oligosakkaridit ovat mukana. Taulukossa on selittävien muuttujien regressiokertoimien estimaatit (keskivirheet), testiaineistojen keskineliöennustevirheet (MSE) ja tasoitusparametrin λ arvo harjuregressiolle ja lassolle. PNS-estimoidussa mallissa oligosakkaridi LSTc on merkitsevyyden rajalla ($p = 0.094$).

lista. PNS-estimoidussa mallissa ei kuitenkaan ollut kuin yksi tilastollisen merkitsevyyden rajalla oleva oligosakkaridi (LSTc), joten tulos tuntuu järkevältä. Lasson muodostavat parametrien estimaatit ovat myös lähempänä nollaa kuin PNS-estimoidun mallin estimaatit.

Lasson valitsemilla muuttujilla muodostettiin vielä mallit PNS-estimoinnilla, jotta voitaisiin verrata täyden mallin ja vähäparametrisen mallin estimaattien varianssia. Kaikkien parametrien estimaattien keskihajonnat ovat pienempiä vähäparametrisessa mallissa. Secretorien mallissa oligosakkaridin DFLNT estimaatin hajonta melkein puolittui.

Tarkastellaan seuraavaksi mallien keskineliövirhettä, jota käytetään testi- ja ennustevirheenä. PNS-estimoidun kaikki oligosakkarit sisältävän täyden mallin ennustevirhe on selvästi suurin kaikissa tapauksissa. Lisäksi virhe on suurempi secretor- ja nonsecretormalleissa, mutta tämä johtuu vain pienemmästä havaintojen määrästä. Koko aineiston ja secretorien tilanteessa vähäparametrisella PNS-mallilla on pienin ennustevirhe. Tämä malli siis sopii parhaiten näissä tapauksissa. Lisäksi huomataan, että harjuregression malleissa MSE on pienempi kuin lasson avulla muodostetuissa malleissa. Kuitenkin secretorien tapauksessa ero on hyvin pieni ja arvo on melkein sama. Lisäksi tilanne muuttuu, kun malliin lisätään muitakin selittäviä tekijöitä (katso liite D).

9 Johtopäätökset

Mallien tuloksista huomataan kuinka erilaisesti oligosakkaridit käyttäytyvät eri secretorstatuksen mukaan. Kuitenkin täytyy myös muistaa, että secretoräitejä on aineistossa paljon enemmän. Tilastollisesti merkitsevä yhteys lapsen yhden vuoden painon z-scoren kanssa löydettiin kuitenkin muutaman oligosakkaridin kanssa. Oligosakkaridit DFLac ja DSLNT voisivat vaikuttaa lapsen kokoon, mutta tämä pitäisi tutkia tarkemmin kliinisesti kontrolloidussa ympäristössä. Eettisistä ja käytännön syistä rintaruokintaa ei voida satunnaistaa tai rajoittaa ja toisaalta kaikki lapset eivät ole rintaruokittuja, vaikka tutkija niin haluaisikin. Liitteen D malleista huomataan, että sekoittavat tekijät kuten lapsen syntymäkoko ja äidin BMI vaikuttavat kuitenkin selkeämmin lapsen yksivuotispainoon.

Soveltavassa osiossa huomataan myös selkeästi teoriaosuudessa kuvatut lasson ja harjuregression ominaisuudet. Molemmat mallit kutistavat estimaatteja kohti nollaa. Harjuregressio ei kuitenkaan kutista estimaatteja täsmälleen nollaksi, vaikka tasoitusparametrin λ arvo olisi todella suuri. Lasson mallissa estimaatit voivat kuitenkin saada täsmälleen arvon nolla, joten lasso suorittaa samalla myös mallin valinnan. Mallia on helpompi tulkita, koska siinä on vähemmän muuttujia. Täysien PNS-estimoitujen mallien keskineliövirheet (MSE) ovat suuremmat kuin lasso- ja harjuregressiomallien, kun käytetään sopivaa tasoitusparametrin λ arvoa. Lasso- ja harjuregressiomalleilla on siis paremmat ominaisuudet ennustamisen näkökulmasta. Testiaineiston avulla laskettu keskineliövirhe on hyvä tapa arvioida eri regressiomalleja. Lasso vaikuttaa tässä tutkielman tilanteessa kuitenkin järkevämmältä mallilta kuin harjuregressio. Lasso-mallin on tulkinnaltaan paljon helpompi, koska se valitsee mukaan vain osan oligosakkareista. Jos tasoitusparametrin λ arvoja vaihdeltaisiin, voitaisiin vaikuttaa siihen, kuinka monta oligosakkaridia malliin halutaan jättää. Lisäksi ennustevirheen ero ei ole kovinkaan suuri harjuregression ja lasson välillä. Voidaan kuitenkin pohtia, tulivatko lasson malliin mukaan kaikki vasteen kannalta oleelliset oligosakkaridit vai alisovittaako lasso. Lopullisesta muuttujan valinnasta pitäisi aina päättää sovellusalan asiantuntijan ja tutkijan.

Artikkelissa [3] HMO-muuttujien yhteyttä painoon on tutkittu hierarkisella lineaarisella sekamallilla, joka on tehty toistuville mittauksille. Näissä malleissa vasteena on toiminut painon z-score ja selittävinä tekijöinä on ollut lapsen sukupuoli, synnytystapa, syntymäpainon z-score, äidin BMI, aika (vuosina), eri oligosakkaridit yksi kerrallaan ja oligosakkaridin ja ajan yhdysvaikutus. Aikaa on käsitelty kategorisena muuttujana ja toistuville mittauksille on käytetty rakenteetonta kovarianssimatriisia (unstructured covariance pattern). Tutkimuksessa huomattiin, että monella yksittäisellä oligosakka-

ridilla on korrelaatiota painon z-scoren kanssa ensimmäisen viiden elinvuoden aikana secretorien joukossa. Näitä oligosakkareita ovat mm. 2'FL, 3FL, LNnT, DFLac, LSTb ja 3SL. Lisäksi DSLNT, LNT ja DSLNH ovat merkitsevyyden rajalla. Joukossa on samoja oligosakkareita, kuin tämä tutkielman mallien merkittävät oligosakkaridit. Kuitenkin tutkimuksessa painotettua LNnT oligosakkaridia näiden mallien avulla ei löydetty. Artikkelin tutkimuksessa ja tässä tutkielmassa on kuitenkin aivan eri mallit ja aineistot, joten tulokset eivät ole aivan vertailukelpoisia. Mielenkiintoista olisi tutkia, pystyykö harjuregressiota tai lassoa soveltamaan toistomittaustilanteeseen. Tällöin tutkimuksessa olisi voitu tehdä toistomittausmalli kaikilla oligosakkareilla ja lasson avulla oltaisiin voitu tehdä mallinvalintaa ja tiputtaa osa oligosakkareista pois mallista. [3]

Tutkimuksesta ja tästä tutkielmasta voidaan päätellä, että HMO-yhdisteet voisivat olla yksi linkki äidin rintamaidon ja lapsen kasvun välillä. Tulevaisuudessa voitaisiin siis ehkä kehitellä uusia ravintotapoja, jotka joko voisivat muokata äidin rintamaidon koostumusta tai antaa lapselle hänen tarvitsemiaan HMO-yhdisteitä. Jotkin äidinmaidonkorvikkeet sisältävät jo nyt 2'FL ja LNnT oligosakkarideja. [3, 16]

Kirjallisuutta

[1] T. Hastie, R. Tibshirani & J. Friedman: *The Elements of Statistical Learning*. Springer, New York, 2011.

[2] Park C, Yoon Y.J.: Bridge regression: Adaptivity and group selection, Journal of Statistical Planning and Inference, 141:11,3506-3519, 2011.

[3] Lagström H, Rautava S, Ollila H, Kaljonen A, Turta O, Mäkelä J, Yonemitsu C, Gupta J, Bode L: Associations between human milk oligosaccharides and growth in infancy and early childhood, The American Journal of Clinical Nutrition, 2020.

[4] Saikkonen, Pentti: *Lineaarinen malli*, Luentomoniste, 2007, päivitetty 2013.

[5] Agresti, Alan: *Foundations of Linear and Generalized Linear Models*, Hoboken, New Jersey, 2015.

[6] Robert, Christian P.: *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer, New York, 2007.

[7] G. James, D. Witten, T. Hastie & R. Tibshirani: *An Introduction to Statistical Learning*, Springer, New York, 2013.

[8] Koppinen, Markku: *Matriisilaskenta*, Luentomoniste, 2012.

[9] Hui Zou: The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association, 101:476, 1418-1429, 2006.

[10] Friedman J, Hastie T, Tibshirani R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software, 33(1), 2010.

[11] Lagström H, Rautava P, Kaljonen A, Räihä H, Pihlaja P, Korpilahti P, Peltola V, Rautakoski P, Österbacka E, Niemi P, Simell O.: Cohort Profile: Steps to the Healthy Development and Well-being of Children (the STEPS Study), Int J Epidemiol 2013;42:1273-1284.

[12] Victora CG, Bahl R, Barros AJ, et al.: Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect, Lancet. 2016;387(10017):475-490

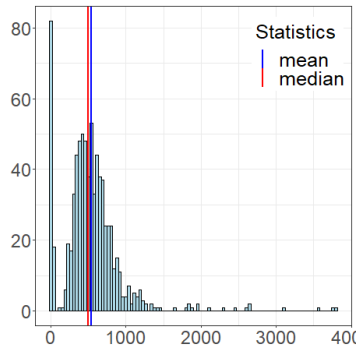
[13] Bode L.: Human milk oligosaccharides: every baby needs a sugar mama, Glycobiology. 2012;22(9):1147-1162.

[14] Saari A, Sankilampi U, Hannila ML, Kiviniemi V, Kesseli K, Dunkel L.:New Finnish growth references for children and adolescents aged 0 to 20 years: Length/height-for-age, weight-for-length/height, and body mass index-for-age, Ann Med. 2011;43(3):235-248

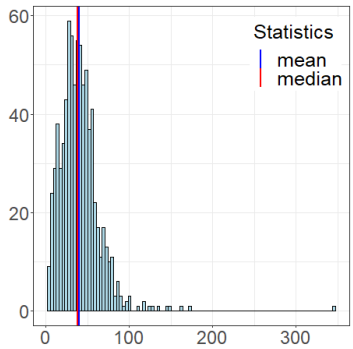
[15] Melkumova L.E. & Shatskikh S.Ya.: Comparing Ridge and LASSO estimators for data analysis, Procedia Engineering Volume 201, 2017, Pages 746-755.

[16] Vandenplas, Y., Berger, B., Carnielli, V. P., Ksiazyk, J., Lagström, H., Sanchez Luna, M., Migacheva, N., Mosselmans, J. M., Picaud, J. C., Possner, M., Singhal, A., & Wabitsch, M.: Human Milk Oligosaccharides: 2'-Fucosyllactose (2'-FL) and Lacto-N-Neotetraose (LNnT) in Infant Formula, Nutrients, 10(9),2018, 1161.

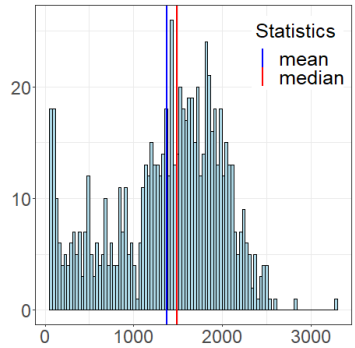
A. Oligosakkaridien histogrammit



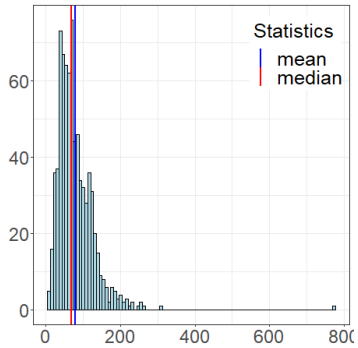
(a) DFLac



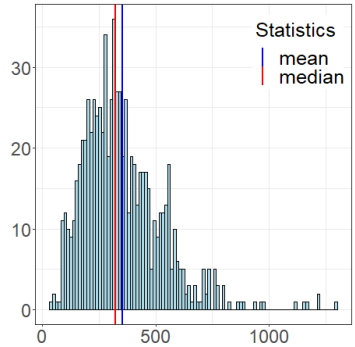
(b) DFLNH



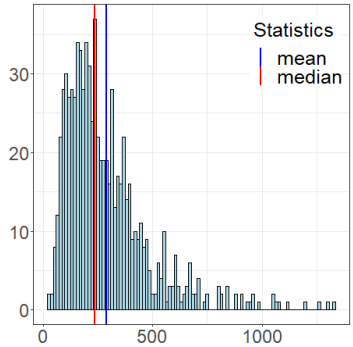
(c) DFLNT



(d) DSLNH

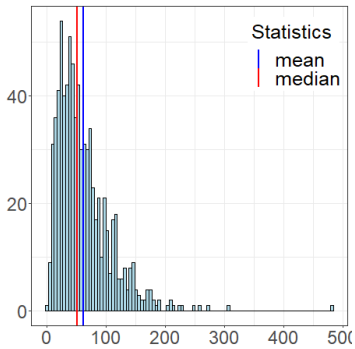


(e) DSLNT

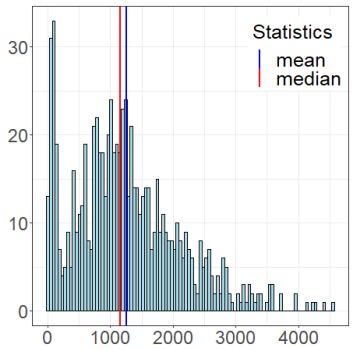


(f) FDSLNH

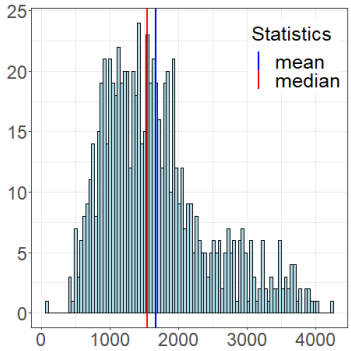
Kuva 10: Oligosakkaridien histogrammit. Sininen viiva kuvastaa keskiarvoa ja punainen viiva mediaania.



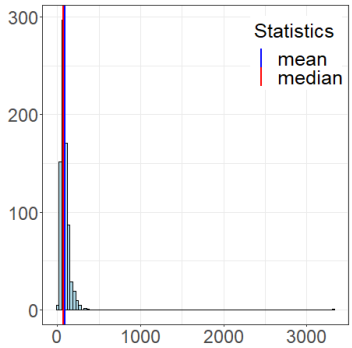
(a) FLNH



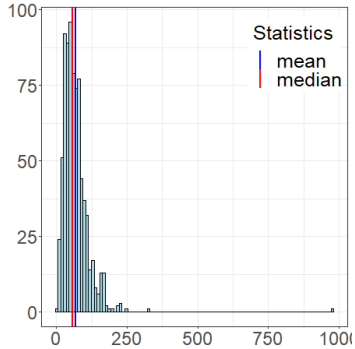
(b) LNFP I



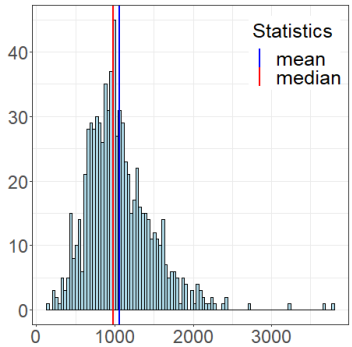
(c) LNFP II



(d) LNFP III

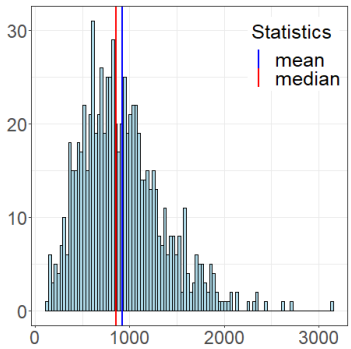


(e) LNH

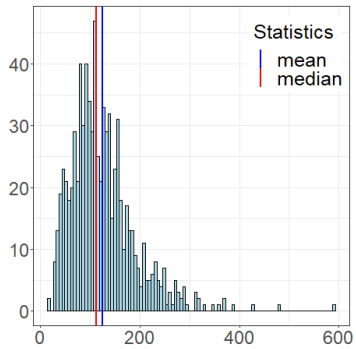


(f) LNnT

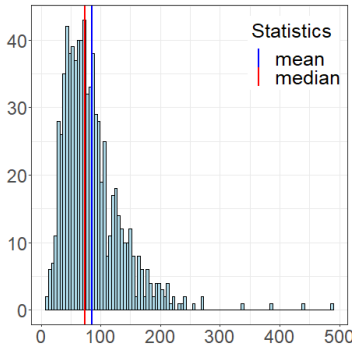
Kuva 11: Oligosakkaridien histogrammit. Sininen viiva kuvastaa keskiarvoa ja punainen viiva mediaania.



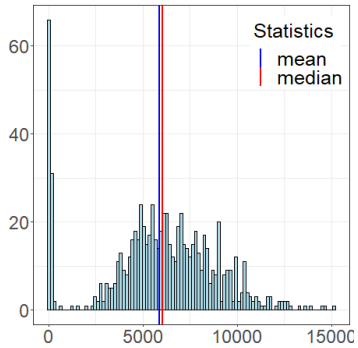
(a) LNT



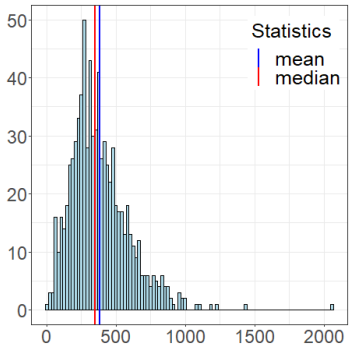
(b) LSTb



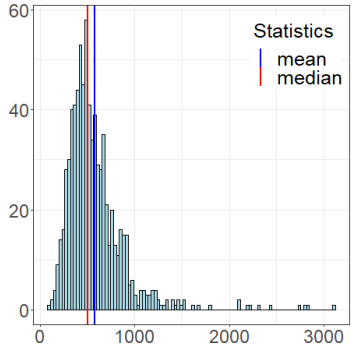
(c) LSTc



(d) 2FL

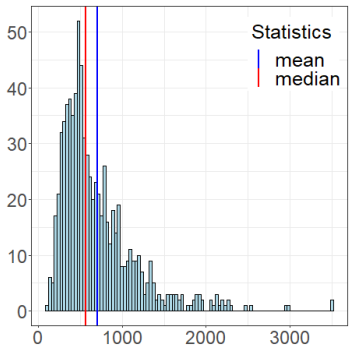


(e) 3FL

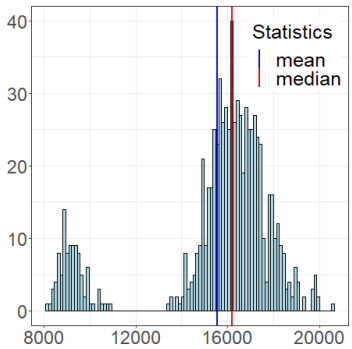


(f) 3SL

Kuva 12: Oligosakkaridien histogrammit. Sininen viiva kuvastaa keskiarvoa ja punainen viiva mediaania.



(a) 6SL



(b) HMO -muuttujien summa

Kuva 13: Oligosakkaridien histogrammit. Sininen viiva kuvastaa keskiarvoa ja punainen viiva mediaania.

B. Oligosakkaridien spearmanin korrelaatiot

	X_2FL_nmol	X_3FL_nmol	LNnT_nmol	X_3SL_nmol	DFLac_nmol	X_6SL_nmol	LNT_nmol	LNFP_I_nmol	LNFP_II_nmol	LNFP_III_nmol
X_2FL_nmol	1									
X_3FL_nmol	0.58	1								
LNnT_nmol	-0.44	-0.42	1							
X_3SL_nmol	0.17	0.51	-0.35	1						
DFLac_nmol	0.27	0.51	-0.29	0.55	1					
X_6SL_nmol	-0.5	-0.35	0.4	-0.26	-0.3	1				
LNT_nmol	-0.18	-0.26	0.24	-0.22	-0.08	-0.12	1			
LNFP_I_nmol	0.52	0.13	0.1	-0.32	-0.05	-0.22	0.22	1		
LNFP_II_nmol	-0.7	-0.27	0.21	0.08	-0.08	0.36	0	-0.75	1	
LNFP_III_nmol	-0.4	-0.09	0.09	-0.12	-0.23	0.38	0	-0.44	0.5	1
LSTb_nmol	-0.64	-0.47	0.31	-0.2	-0.15	0.32	0.26	-0.07	0.29	0.17
LSTc_nmol	0.19	0.02	-0.02	0.08	0.12	-0.19	-0.02	0.14	-0.2	-0.32
DFLNT_nmol	-0.17	0.09	0.19	0.28	0.4	-0.08	0.14	-0.06	0.16	-0.13
LNH_nmol	-0.17	-0.07	0.12	0.02	0.01	0.09	-0.02	-0.36	0.31	0.35
DSLNT_nmol	-0.49	-0.38	0.32	-0.11	0.06	0.17	0.24	0.05	0.11	-0.02
FLNH_nmol	0.04	0.02	-0.01	0.07	0.22	-0.07	0.25	0.02	-0.12	-0.01
DFLNH_nmol	0.29	0.07	0.24	-0.16	0.03	-0.21	0.11	0.56	-0.33	-0.35
FDSLNH_nmol	-0.43	-0.27	0.02	0.07	-0.05	0.19	-0.1	-0.74	0.71	0.43
DSLNH_nmol	-0.11	-0.18	-0.08	0.07	0.08	-0.03	0.03	-0.24	0.09	-0.08

	LSTb_nmol	LSTc_nmol	DFLNT_nmol	LNH_nmol	DSLNT_nmol	FLNH_nmol	DFLNH_nmol	FDSLNH_nmol	DSLNH_nmol
X_2FL_nmol									
X_3FL_nmol									
LNnT_nmol									
X_3SL_nmol									
DFLac_nmol									
X_6SL_nmol									
LNT_nmol									
LNFP_I_nmol									
LNFP_II_nmol									
LNFP_III_nmol									
LSTb_nmol	1								
LSTc_nmol	-0.33	1							
DFLNT_nmol	-0.02	0.04	1						
LNH_nmol	-0.24	0.06	0.24	1					
DSLNT_nmol	0.69	-0.1	0.05	-0.18	1				
FLNH_nmol	-0.2	0.19	0.43	0.54	-0.13	1			
DFLNH_nmol	-0.21	0.38	0.17	0.05	-0.05	0.23	1		
FDSLNH_nmol	0.03	-0.17	0.08	0.54	-0.1	0.1	-0.39	1	
DSLNH_nmol	-0.18	0.57	0.04	0.24	-0.03	0.17	-0.04	0.36	1

Taulukko 6: Oligosakkaridien väliset Spearmanin korrelaatiot koko aineistolle.

	X_2FL_nmol	X_3FL_nmol	LNnT_nmol	X_3SL_nmol	DFLac_nmol	X_6SL_nmol	LNT_nmol	LNFP_I_nmol	LNFP_II_nmol	LNFP_III_nmol
X_2FL_nmol	1									
X_3FL_nmol	0.4	1								
LNnT_nmol	-0.56	-0.48	1							
X_3SL_nmol	0.06	0.49	-0.39	1						
DFLac_nmol	-0.09	0.31	-0.36	0.57	1					
X_6SL_nmol	-0.41	-0.19	0.44	-0.21	-0.13	1				
LNT_nmol	-0.37	-0.45	0.25	-0.33	-0.24	-0.02	1			
LNFP_I_nmol	0.28	-0.25	0.17	-0.59	-0.58	-0.02	0.16	1		
LNFP_II_nmol	-0.63	-0.05	0.31	0.26	0.29	0.27	0.17	-0.69	1	
LNFP_III_nmol	-0.23	0.16	0.11	-0.04	0	0.29	0.16	-0.28	0.39	1
LSTb_nmol	-0.6	-0.39	0.32	-0.18	0.04	0.24	0.36	0.15	0.2	0.05
LSTc_nmol	0.13	-0.07	-0.01	0.07	0.04	-0.13	-0.07	0.06	-0.16	-0.31
DFLNT_nmol	-0.56	-0.2	0.27	0.18	0.17	0.09	0.11	-0.4	0.46	0.05
LNH_nmol	-0.19	-0.09	0.19	0.03	0.05	0.14	0	-0.44	0.36	0.38
DSLNT_nmol	-0.58	-0.41	0.31	-0.14	0.14	0.17	0.23	0.12	0.12	-0.03
FLNH_nmol	-0.17	-0.2	0.05	-0.04	0.07	0.1	0.19	-0.21	0.05	0.16
DFLNH_nmol	0.02	-0.24	0.36	-0.34	-0.35	-0.04	0.04	0.42	-0.16	-0.2
FDSLNH_nmol	-0.26	-0.07	0.05	0.21	0.29	0.07	0.02	-0.69	0.64	0.29
DSLNH_nmol	-0.03	-0.13	-0.09	0.1	0.21	-0.08	0.06	-0.21	0.03	-0.18

	LSTb_nmol	LSTc_nmol	DFLNT_nmol	LNH_nmol	DSLNT_nmol	FLNH_nmol	DFLNH_nmol	FDSLNH_nmol	DSLNH_nmol
X_2FL_nmol									
X_3FL_nmol									
LNnT_nmol									
X_3SL_nmol									
DFLac_nmol									
X_6SL_nmol									
LNT_nmol									
LNFP_I_nmol									
LNFP_II_nmol									
LNFP_III_nmol									
LSTb_nmol	1								
LSTc_nmol	-0.33	1							
DFLNT_nmol	0.17	-0.05	1						
LNH_nmol	-0.25	0.05	0.29	1					
DSLNT_nmol	0.71	-0.09	0.12	-0.16	1				
FLNH_nmol	-0.12	0.14	0.32	0.64	-0.13	1			
DFLNH_nmol	-0.05	0.37	-0.04	0.05	0.01	0.09	1		
FDSLNH_nmol	-0.1	-0.13	0.34	0.6	-0.13	0.31	-0.25	1	
DSLNH_nmol	-0.24	0.63	0.12	0.23	-0.01	0.24	0.04	0.34	1

Taulukko 7: Oligosakkaridien väliset Spearmanin korrelaatiot secretor-äideille.

	X_2FL_nmol	X_3FL_nmol	LNnT_nmol	X_3SL_nmol	DFLac_nmol	X_6SL_nmol	LNT_nmol	LNFP_I_nmol	LNFP_II_nmol	LNFP_III_nmol
X_2FL_nmol	1									
X_3FL_nmol	0.27	1								
LNnT_nmol	0.06	-0.59	1							
X_3SL_nmol	0.24	0.36	-0.11	1						
DFLac_nmol	-0.1	-0.01	0.02	-0.05	1					
X_6SL_nmol	0	-0.57	0.42	-0.06	0.07	1				
LNT_nmol	-0.06	-0.15	0.3	0.11	0.03	-0.2	1			
LNFP_I_nmol	-0.04	0.07	0.01	0.17	-0.04	-0.08	0.51	1		
LNFP_II_nmol	-0.2	0.25	-0.4	-0.35	0.02	-0.13	-0.52	-0.37	1	
LNFP_III_nmol	0.08	0.09	-0.14	0.04	0	0.09	-0.52	-0.2	0.23	1
LSTb_nmol	-0.14	-0.19	0.32	0.11	0.12	0.2	0.26	0.41	-0.17	-0.04
LSTc_nmol	-0.18	-0.03	-0.11	-0.16	0.07	-0.2	0.05	-0.03	-0.07	-0.1
DFLNT_nmol	-0.02	0.44	-0.48	-0.03	-0.05	-0.29	-0.41	-0.13	0.5	0.09
LNH_nmol	-0.03	0.31	-0.29	0.03	-0.01	-0.42	-0.05	0.01	-0.08	0.13
DSLNT_nmol	-0.07	-0.11	0.34	0.28	0.11	0.03	0.48	0.4	-0.35	-0.29
FLNH_nmol	-0.04	0.17	-0.31	0.25	-0.11	-0.17	0.38	0.6	-0.37	-0.21
DFLNH_nmol	-0.02	0.06	-0.27	-0.01	0.04	-0.23	-0.09	-0.25	0.11	0.07
FDSLNH_nmol	0.06	0.33	-0.37	-0.1	-0.06	-0.38	-0.27	-0.41	0.15	0.16
DSLNH_nmol	0.03	-0.02	-0.15	0.06	0.06	-0.11	-0.03	-0.15	-0.25	-0.11

	LSTb_nmol	LSTc_nmol	DFLNT_nmol	LNH_nmol	DSLNT_nmol	FLNH_nmol	DFLNH_nmol	FDSLNH_nmol	DSLNH_nmol
X_2FL_nmol									
X_3FL_nmol									
LNnT_nmol									
X_3SL_nmol									
DFLac_nmol									
X_6SL_nmol									
LNT_nmol									
LNFP_I_nmol									
LNFP_II_nmol									
LNFP_III_nmol									
LSTb_nmol	1								
LSTc_nmol	0	1							
DFLNT_nmol	-0.06	0.17	1						
LNH_nmol	-0.53	0.19	0	1					
DSLNT_nmol	0.58	-0.11	-0.17	-0.34	1				
FLNH_nmol	0.06	0.12	0.11	0.14	0.22	1			
DFLNH_nmol	-0.48	0.16	0.01	0.46	-0.34	0.07	1		
FDSLNH_nmol	-0.56	0.05	0.02	0.59	-0.47	-0.25	0.42	1	
DSLNH_nmol	-0.34	0.5	-0.02	0.31	-0.2	0.13	0.11	0.32	1

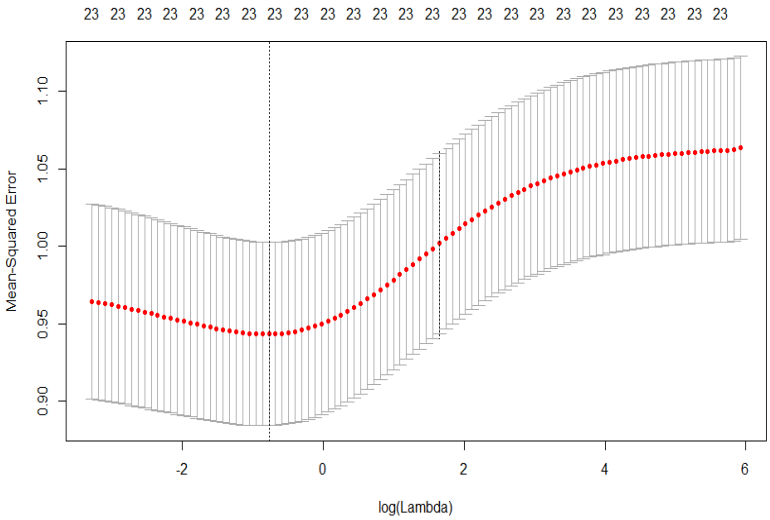
Taulukko 8: Oligosakkaridien väliset Spearmanin korrelaatiot nonsecretor-äideille.

C.Oligosakkaridien ja lapsen koon väliset korrelaatiot

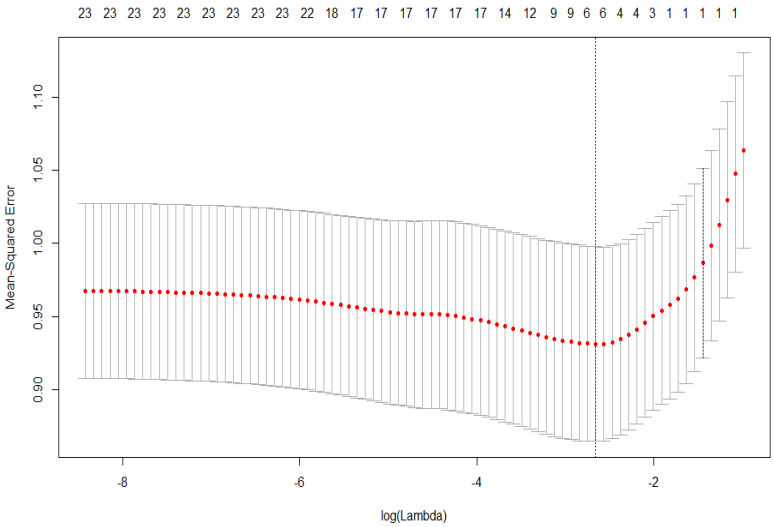
Muuttuja	Total		Secretor		Nonsecretor	
	Pituus SDS	Paino SDS	Pituus SDS	Paino SDS	Pituus SDS	Paino SDS
2'FL	0.096	0.098	0.098	0.096	-0.033	-0.022
3'FL	0.051	0.113	0.040	0.111	-0.121	0.065
LNnT	-0.059	-0.108	-0.084	-0.118	0.100	-0.047
3'SL	0.047	0.079	0.057	0.085	-0.058	0.033
DFLac	0.030	0.086	0.009	0.079	0.067	0.056
6'SL	-0.062	-0.057	-0.047	-0.034	0.003	-0.086
LNT	-0.057	-0.062	-0.084	-0.102	0.071	0.151
LNFP I	0.038	-0.002	0.021	-0.035	0.064	0.033
LNFP II	-0.121	-0.070	-0.116	-0.056	-0.146	-0.123
LNFP III	-0.075	-0.024	-0.070	-0.002	0.090	-0.049
LSTb	-0.087	-0.083	-0.089	-0.077	-0.002	-0.013
LSTc	0.032	0.029	0.035	0.040	-0.016	-0.098
DFLNT	0.034	-0.001	0.024	-0.011	-0.040	-0.072
LNH	-0.010	0.032	-0.029	0.013	0.115	0.181
DSLNT	-0.079	-0.094	-0.086	-0.100	0.002	-0.045
FLNH	0.031	0.021	0.021	0.000	0.045	0.059
DFLNH	0.029	0.014	-0.007	-0.015	0.129	0.061
FDSLNH	-0.081	-0.024	-0.068	-0.009	-0.055	0.117
DSLNH	0.023	0.039	0.039	0.059	-0.059	-0.087
HMO Summa	0.082	0.101	0.078	0.101	0.061	-0.061
Pituus SDS 1v	1.000	0.647	1.000	0.648	1.000	0.639
Paino SDS 1v	0.647	1.000	0.648	1.000	0.639	1.000

Taulukko 9: Spearmanin korrelaatiot äidin rintamaidon oligosakkaridien ja lapsen yksivuotispituuden ja -painon z-scoren (SDS) välillä.

D. Mallit sekoittavilla tekijöillä

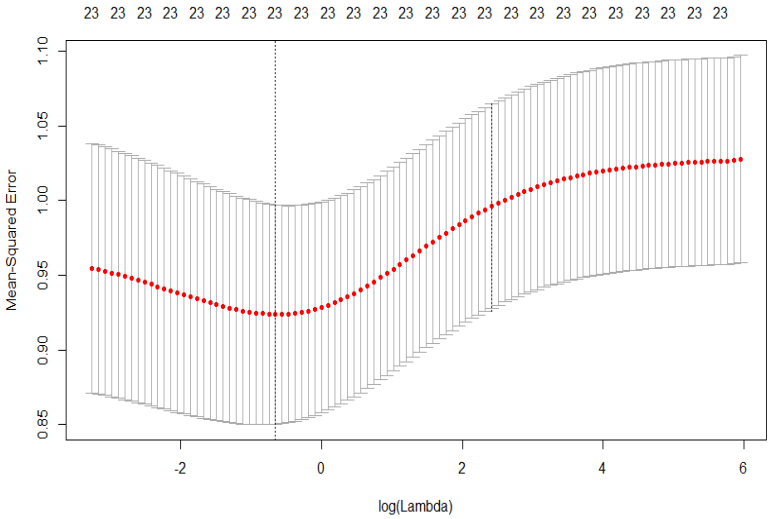


(a) Harjuregressio, koko aineisto

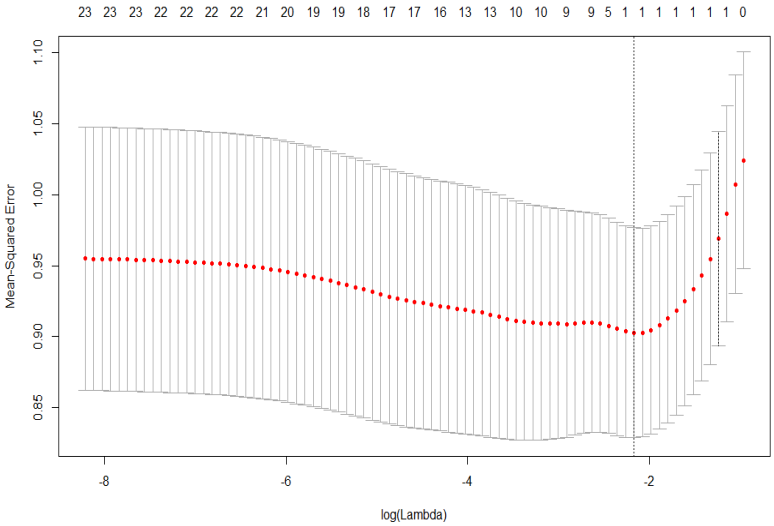


(b) Lasso, koko aineisto

Kuva 14: Kuvassa (a) on harjuregression ja (b) lasson 10-ristiinvalidoinnit parhaan tasoitusparametrin λ arvon löytämiselle koko aineiston opetusaineistossa. Y-akselilla on keskineliövirhe ja x-akselilla λ arvoja log-asteikolla. Pienimmän keskineliövirheen omaava λ on merkitty katkoviivalla. Mallin parametrien lukumäärä näkyy kuvan yläreunassa. Punainen käyrä kuvaa keskineliövirheen arvoja ja harmaat viivat kuvaavat arvon keskivirhettä.

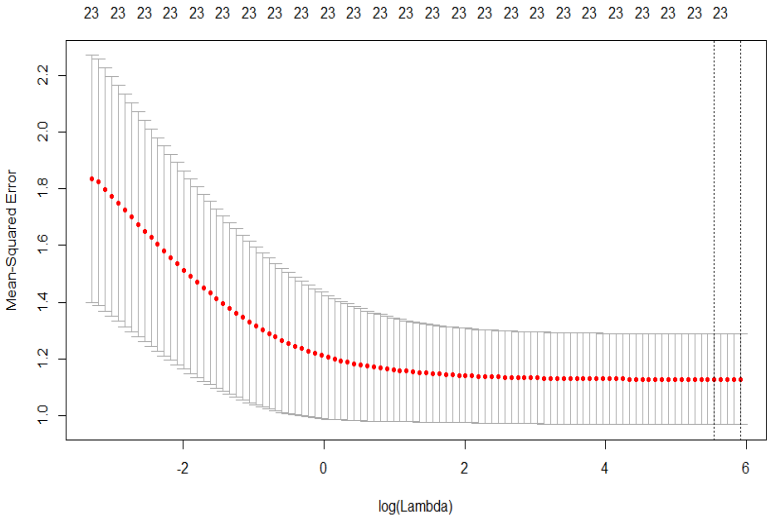


(a) Harjuregressio, secretor

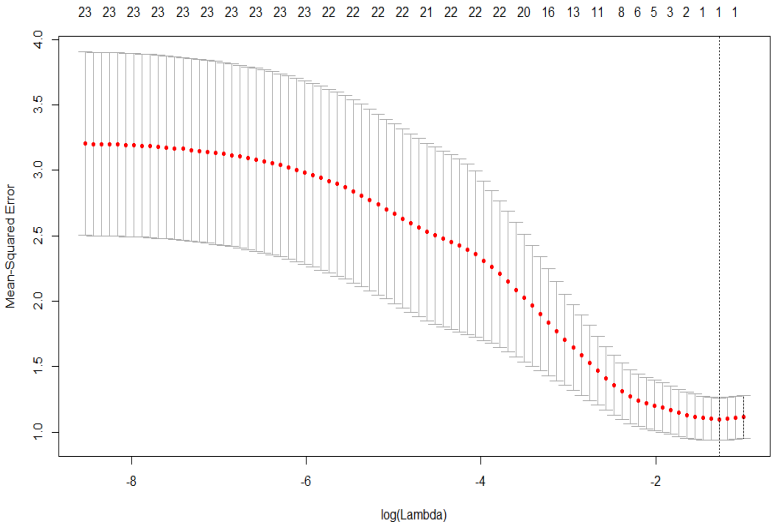


(b) Lasso, secretor

Kuva 15: Kuvassa (a) on harjuregression ja (b) lasson 10-ristiinvalidoinnit parhaan tasoitusparametrin λ arvon löytämiselle secretor-aineiston opetusaineistossa. Y-akselilla on keskineliövirhe ja x-akselilla λ arvoja log-asteikolla. Pienimmän keskineliövirheen omaava λ on merkitty katkoviivalla. Mallin parametrien lukumäärä näkyy kuvan yläreunassa. Punainen käyrä kuvaa keskineliövirheen arvoja ja harmaat viivat kuvaavat arvon keskivirhettä.

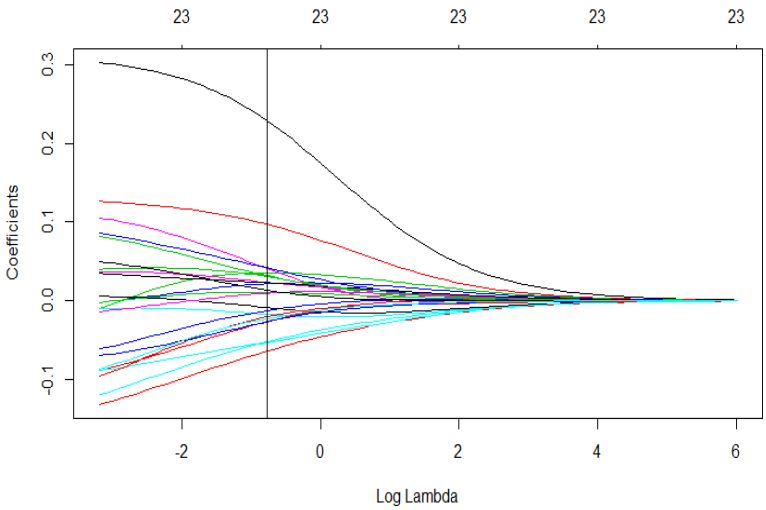


(a) Harjuregressio, nonsecretor

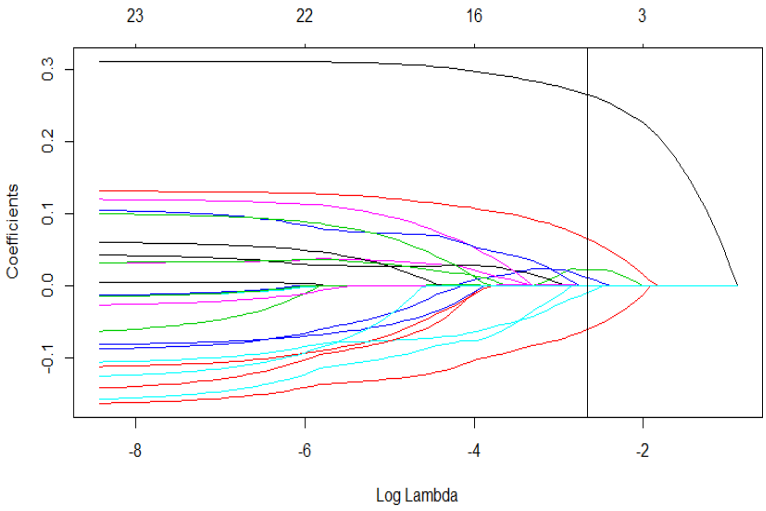


(b) Lasso, nonsecretor

Kuva 16: Kuvassa (a) on harjuregression ja (b) lasson 10-ristiinvalidoinnit parhaan tasoitusparametrin λ arvon löytämiselle nonsecretor-aineiston opetusaineistossa. Y-akselilla on keskineliövirhe ja x-akselilla λ arvoja logasteikolla. Pienimmän keskineliövirheen omaava λ on merkitty katkoviivalla. Mallin parametrien lukumäärä näkyy kuvan yläreunassa. Punainen käyrä kuvaa keskineliövirheen arvoja ja harmaat viivat kuvaavat arvon keskivirhettä.



(a) Harjuregressio



(b) Lasso

Kuva 17: Kuvassa (a) on harjuregression ja (b) lasson mallien kehittyminen koko aineistolla kun tasoitusparametrin λ arvo muuttuu. X-akselilla on tasoitusparametrin λ luonnollisen logaritmin arvoja ja y-akselilla parametrien estimmaattien saama arvo. Kuvan yläreunassa on pisteen λ muodostaman mallin parametrien lukumäärä. Viivalla on merkitty parhaan tasoitusparametrin λ arvo eli ristiinvalidoinnilla saatu arvo, jolla mallin keskineliövirhe minimoitui.

Muuttuja	PNS-malli	Harjuregressio	Lasso	PNS-Lasso
λ		0.462	0.0702	
2'FL	0.00184 (0.0991)	0.0162	0.0221	-0.00327 (0.0476)
3FL	-0.000137 (0.0588)	0.0196	0.0110	0.0710 (0.0405)
LNnT	0.00777 (0.0501)	-0.00769		
3'SL	0.0361 (0.0648)	0.0299		
DFLac	0.0935 (0.0656)	0.0440		
6'SL	-0.0425 (0.0555)	-0.0211	-0.0614	-0.0481 (0.0416)
LNT	-0.00292 (0.0438)	-0.00792		
LNFP I	-0.0322 (0.0771)	-0.00545		
LNFP II	-0.0824 (0.0871)	-0.0246		
LNFP III	0.0549 (0.0504)	0.0147		
LSTb	0.0390 (0.0596)	0.00193		
LSTc	-0.0301 (0.0511)	-0.00534		
DFLNT	-0.0219 (0.0469)	-0.0137		
LNH	0.0113 (0.0544)	0.00670		
DSLNT	-0.102 (0.0576)	-0.0463	-0.00993	-0.0461 (0.0374)
FLNH	-0.0326 (0.0504)	-0.00705		
DFLNH	0.0637 (0.0491)	0.0214		
FDSLNH	0.0243 (0.0786)	0.0139		
DSLNH	0.0405 (0.0523)	0.0197		
Sukupuoli	-0.0267 (0.0697)	-0.0119		
Syntymäpaino SDS	0.294 (0.0322)	0.215	0.265	0.296 (0.0315)
Synnytystapa	0.0506 (0.110)	0.00621		
Äidin BMI	0.0919 (0.0351)	0.0743	0.0651	0.0972 (0.0344)
MSE	0.881	0.822	0.795	0.809

Taulukko 10: Estimoitujen mallien tulokset koko aineistolle, kun mukana ovat oligosakkaridit ja muut selittävät tekijät. Taulukossa on selittävien muuttujien regressiokertoimien estimaatit (keskivirheet), testiaineistojen keskineliöennustevirheet (MSE) ja tasoitusparametrin λ arvo harjuregressiolle ja lassolle. PNS-estimoidussa mallissa syntymäpainon z-score ($p < 0.001$) ja äidin BMI ($p = 0.009$) saivat tilastollisesti merkitsevät p-arvot. Oligosakkaridi DSLNT ($p = 0.076$) on merkitsevyyden rajalla. Vähäparametrisen PNS-estimoidun mallin muuttujista syntymäpainon z-score ($p < 0.001$) ja äidin BMI ($p = 0.005$) saivat tilastolliset merkitsevät p-arvot ja oligosakkaridi 3FL ($p = 0.080$) on merkitsevyyden rajalla.

Muuttuja	PNS-malli	Harjuregressio	Lasso	PNS-Lasso
λ		0.521	0.115	
2'FL	-0.0211 (0.180)	0.0114		
3FL	-0.0470 (0.0641)	0.0109		
LNnT	-0.0357 (0.0686)	-0.0211		
3'SL	0.0529 (0.0697)	0.0288		
DFLac	0.0895 (0.0828)	0.0408		
6'SL	-0.00934 (0.0763)	-0.00181		
LNT	-0.0725 (0.0641)	-0.0360		
LNFP I	0.0185 (0.112)	-0.000815		
LNFP II	0.0188 (0.129)	-0.00347		
LNFP III	0.0868 (0.0671)	0.0220		
LSTb	0.0168 (0.0677)	-0.00174		
LSTc	-0.0535 (0.0580)	0.000740		
DFLNT	-0.0230 (0.0839)	-0.0124		
LNH	0.0106 (0.0665)	-0.00191		
DSLNT	-0.111 (0.0657)	-0.0472		
FLNH	-0.0329 (0.0620)	-0.0116		
DFLNH	0.0590 (0.0550)	0.0187		
FDSLNH	-0.0812 (0.119)	-0.000284		
DSLNH	0.102 (0.0619)	0.0336		
Sukupuoli	-0.0359 (0.0752)	-0.0127		
Syntymäpaino SDS	0.321 (0.0351)	0.226	0.247	0.342 (0.0333)
Synnytystapa	0.0235 (0.121)	-0.00193		
Äidin BMI	0.0692 (0.0387)	0.0618		
MSE	0.909	0.863	0.860	0.849

Taulukko 11: Estimoitujen mallien tulokset secretoraineistolle, kun mukana ovat oligosakkaridit ja muut selittävät tekijät. Taulukossa on selittävien muuttujien regressiokertoimien estimaatit (keskivirheet), testiaineistojen keskineliöennustevirheet (MSE) ja tasoitusparametrin λ arvo harjuregressiolle ja lassolle. PNS-estimoidussa mallissa syntymäpainon z-score ($p < 0.001$) sai tilastollisesti merkitsevän p-arvon. Oligosakkaridit DSLNT ($p = 0.092$) ja DSLNH ($p = 0.099$) ja äidin BMI ($p = 0.074$) ovat merkitsevyyden rajalla. Vähäparametrisen PNS-estimoidun mallin muuttujista syntymäpainon z-score ($p < 0.001$) sai tilastollisen merkitsevän p-arvon.

Muuttuja	PNS-malli	Harjuregressio	Lasso
λ		256	0.281
2'FL	-3.44 (4.37)	-0.00000477	
3FL	0.0771 (0.370)	0.0000915	
LNnT	-0.205 (0.213)	0.0000246	
3'SL	-0.310 (0.290)	0.0000306	
DFLac	3.89 (4.26)	0.00000661	
6'SL	-0.120 (0.190)	-0.000375	
LNT	0.0115 (0.204)	0.000611	
LNFP I	-4.47 (2.74)	0.00000287	
LNFP II	-0.508 (0.399)	-0.000390	
LNFP III	0.0156 (0.0906	-0.0000485	
LSTb	0.241 (0.159)	0.0000184	
LSTc	-0.252 (0.174)	-0.000307	
DFLNT	-0.190 (0.510)	-0.0000501	
LNH	0.064 (0.139)	0.000473	
DSLNT	-0.122 (0.215)	-0.0000459	
FLNH	0.0715 (0.133)	0.000335	
DFLNH	0.00325 (0.144)	0.0000709	
FDSLNH	-0.00726 (0.236)	0.000451	
DSLNH	-0.131 (0.136)	-0.000132	
Sukupuoli	0.0238 (0.208)	-0.0000147	
Syntymäpaino SDS	0.130 (0.0946)	0.000566	
Synnytystapa	0.553 (0.301)	0.000171	
Äidin BMI	0.185 (0.0904)	0.000652	
MSE	2.05	0.476	0.493

Taulukko 12: Estimoitujen mallien tulokset nonsecretoraineistolle, kun mukana ovat oligosakkaridit ja muut selittävät tekijät. Taulukossa on selittävien muuttujien regressiokertoimien estimaatit (keskivirheet), testiaineistojen keskineliöennustevirheet (MSE) ja tasoitusparametrin λ arvo harjuregressiolle ja lassolle. PNS-estimoidussa mallissa äidin BMI ($p < 0.044$) sai tilastollisesti merkitsevän p-arvon. Synnytystapa ($p = 0.070$) on merkitsevyyden rajalla.

E. Soveltavan osion R-koodi

```
########## PRO GRADU  -TUTKIELMA ##########
########## HELENA OLLILA ##################
########## 23.3.2020 ######################

########## TARVITTAVAT PAKETIT ##########

install.packages('lars')
install.packages('glmnet')
install.packages('leaps')
install.packages("sas7bdat")
install.packages("ggplot2")
install.packages("corrplot")
install.packages("gplots")
install.packages("reshape2")
install.packages("xtable")
install.packages("fBasics")
install.packages("fastDummies")

library(lars)
library(glmnet)
library(leaps)
library(sas7bdat)
library(ggplot2)
library(corrplot)
library(gplots)
library(reshape2)
library(xtable)
library(fBasics)
library(fastDummies)

########## DATA IMPORT ##########

setwd("C:/Users/hekoll/Desktop/Siirretty␣verkkolevylle␣paivityksen
ajaksi/Gradu")

Data <- read.sas7bdat("lapsidata.sas7bdat")

# Kaksoset poistettu (16), keskoset mukana
# Valitaan vain ikapiste 1

Data_1 <- Data[Data[, "aika"] == 1,]
```

```
# Kaikki datan henkilot

Data_kaikki <- Data[(Data[, "aika"] == 1 |
is.na(Data[, "aika"]) |
(Data[, "aika"] == 2 & Data[, "nro"] == 1435)),]

length(Data_kaikki[,1])
length(unique(Data$nro))
length(unique(Data$perhe))

########## ALKUTARKASTELU ##########

# Datan tarkistusta: na arvot, tyypit

Data_oli <- Data_kaikki[,c(18, 21:40)]

colSums(is.na(Data_oli))
Data_oli <- Data_oli[is.na(Data_oli[,"Secretor"]) != TRUE,]

for (i in 1:20){
  print(typeof(Data_oli[, i]))
}

# Histogrammit

for (i in 2:21){
  figure <- ggplot(data=Data_oli, aes(Data_oli[, i])) +
    geom_histogram(bins=100,col="black", fill="lightblue") +
    ylab("") + xlab("") + theme_bw() +
    geom_vline(aes(xintercept = median(Data_oli[, i]),
    col='median'),size=1) + geom_vline(aes(xintercept =
    mean(Data_oli[, i]), col='mean'),size=1) +
    scale_color_manual(name = "Statistics",
    values = c(median = "red", mean = "blue")) +
    theme(legend.position = c(0.83, 0.85)) +
    theme(legend.text=element_text(size=25),
    legend.title=element_text(size=25)) +
    theme(text = element_text(size=30))
  print(figure)

  figurename <- paste(colnames(Data_oli)[i], ".png", sep = "")
```

```
  dev.copy(png,figurename)
  dev.off()
}

# Tunnusluvut oligosakkareista

options(digits = 20)
apply(Data_oli[Data_oli[,"Secretor"]==1,], 2, summary, digits = 10)
apply(Data_oli[Data_oli[,"Secretor"]==0,], 2, summary, digits = 10)
apply(Data_oli, 2, summary, digits = 10)

# Wilcoxon testi secretoreiden valilla

for (i in 2:21){

x <- Data_oli[Data_oli[,"Secretor"]==1,]
y <- Data_oli[Data_oli[,"Secretor"]==0,]
x <- x[,i]
y <- y[,i]
tulos <- wilcox.test(x, y,
            alternative = "two.sided",
            conf.level = 0.95)
print(colnames(Data_oli)[i])
print(tulos$statistic)
print(tulos$parameter)
print(tulos$p.value)
print(format(tulos$p.value, scientific = FALSE))
print(tulos$method)
print("##########")
}

# Korrelaatiot oligosakkaridien valilla

sec0 <- Data_oli[Data_oli[,"Secretor"]==0,]
sec1 <- Data_oli[Data_oli[,"Secretor"]==1,]

# total
cor(Data_oli[,2:20], use = "all.obs", method = "pearson")
Korrelaatiot <- cor(Data_oli[,2:20], use = "all.obs",
method = "spearman")

# secretor
cor(sec1[,2:20], use = "all.obs", method = "pearson")
```

```
Korrelaatiot <- cor(sec1[,2:20], use = "all.obs",
method = "spearman")

# nonsecretor
cor(sec0[,2:20], use = "all.obs", method = "pearson")
Korrelaatiot <- cor(sec0[,2:20], use = "all.obs",
method = "spearman")

Korrelaatiot <- round(Korrelaatiot,2)
Korrelaatiot_m <- melt(Korrelaatiot)

upper<-Korrelaatiot
upper[upper.tri(Korrelaatiot)]<-""
upper<-as.data.frame(upper)
print(xtable(upper[,1:10]), type="html")
print(xtable(upper[,11:19]), type="html")

ggplot(data = Korrelaatiot_m, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Spearman\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()+
  theme(axis.title.x = element_blank(),
  axis.title.y = element_blank())

# Sirontakuviot (ei tutkielmassa, tarkasteltu)

my_cols <- c("#00AFBB", "#E7B800")
pairs(Data_oli[,2:6], lower.panel = NULL, cex = 0.7,
      col = my_cols[Data_oli$Secretor])

# Korrelaatio pituus-SDS ja paino-SDS ja oligosakkaridien valilla

Data_1_kor <- Data_1[, c(18, 21:40, 83, 85)]
sec0 <- Data_1_kor[Data_1_kor[,"Secretor"]==0,]
sec1 <- Data_1_kor[Data_1_kor[,"Secretor"]==1,]

round(cor(Data_1_kor[,2:23], use = "na.or.complete",
          method = "spearman"), 3)[,21:22]
```

```
round(cor(sec0[,2:23], use = "na.or.complete",
          method = "spearman"), 3) [,21:22]
round(cor(sec1[,2:23], use = "na.or.complete",
          method = "spearman"), 3) [,21:22]

# Adjustoivat muuttujat
# syntymatapa, sukupuoli, syntymapainon Z-score ja aidin BMI

sekoittavat <- Data_kaikki[, c(3, 8, 18, 69, 65)]
koko <- Data_1[,c(18,83,85)]

options(digits = 20)
apply(sekoittavat[sekoittavat[,"Secretor"]==1,], 2, summary,
digits = 10)
apply(sekoittavat[sekoittavat[,"Secretor"]==0,], 2, summary,
digits = 10)
apply(sekoittavat, 2, summary, digits = 10)
apply(koko[koko[,"Secretor"]==1,], 2, summary, digits = 10)
apply(koko[koko[,"Secretor"]==0,], 2, summary, digits = 10)
apply(koko, 2, summary, digits = 10)

table(sekoittavat$SUKUP)
table(sekoittavat$SYNNYTYSTAPALUOKKA)
sp_sec <- table(sekoittavat$SUKUP, sekoittavat$Secretor)
syn_sec <- table(sekoittavat$SYNNYTYSTAPALUOKKA, sekoittavat$Secretor)
chisq.test(sp_sec)
chisq.test(syn_sec)

hist(sekoittavat$SYNTYMA_ikapaino_sds_v)
hist(sekoittavat$BMI_aiti)
hist(koko$pituus_sds)
hist(koko$paino_sds)

x <- sekoittavat[sekoittavat[,"Secretor"]==1,]$BMI_aiti
y <- sekoittavat[sekoittavat[,"Secretor"]==0,]$BMI_aiti
tulos <- wilcox.test(x, y, alternative = "two.sided",
conf.level = 0.95)
print(tulos$p.value)
t.test(sekoittavat$SYNTYMA_ikapaino_sds_v~sekoittavat$Secretor)
t.test(koko$pituus_sds~koko$Secretor)
t.test(koko$paino_sds~koko$Secretor)
```

```
########## MALLIEN ESTIMOINTI ##########

set.seed(11)

x <- Data_1[,c(3, 8, 18, 69, 65, 83, 85, 21:39)]

# Poistetaan puuttuvat havainnot

x <- x[complete.cases(x), ]

# Dummy-muuttujat

x[,1] <- factor(x[,1])
x[,3] <- factor(x[,3])
x[,4] <- factor(x[,4])
x <- dummy_cols(x)

# Kaytetaan kategorisista muuttujista: SUKUP_2,
# SYNNYTYSTAPALUOKKA_3 (sektio)

# Jatkuvien muuttujien standardointi

x[,c(5, 8:26)]<- scale(x[,c(5, 8:26)])
round(colMeans(x[,c(5, 8:26)], na.rm = TRUE),5)
colStdevs(x[,c(5, 8:26)])

# Jako secretor-muuttujan mukaan

x_sec0 <- x[x[,"Secretor"]==0,]
x_sec1 <- x[x[,"Secretor"]==1,]

# Vasteet

y_paino <- x[,7]
y_paino_1 <- x_sec1[,7]
y_paino_0 <- x_sec0[,7]

# Erilaiset selittavat tekija -matriisit

# total
x_paino <- x[,-c(1,3,4,6,27,29,30,31,32,34,36)]
x_paino = model.matrix(paino_sds~.,x_paino)[,-1]
```

```
x_paino_oli <- x[,-c(1,2,3,4,5,6,27,28,29,30,31,32,33,34,35,36)]
x_paino_oli = model.matrix(paino_sds~.,x_paino_oli)[,-1]

# secretor
x_paino_1 <- x_sec1[,-c(1,3,4,6,27,29,30,31,32,34,36)]
x_paino_1 = model.matrix(paino_sds~.,x_paino_1)[,-1]
x_paino_oli_1 <- x_sec1[,-c(1,2,3,4,5,6,27,28,29,30,31,32,
33,34,35,36)]
x_paino_oli_1 = model.matrix(paino_sds~.,x_paino_oli_1)[,-1]

# nonsecretor
x_paino_0 <- x_sec0[,-c(1,3,4,6,27,29,30,31,32,34,36)]
x_paino_0 = model.matrix(paino_sds~.,x_paino_0)[,-1]
x_paino_oli_0 <- x_sec0[,-c(1,2,3,4,5,6,27,28,29,30,31,32,
33,34,35,36)]
x_paino_oli_0 = model.matrix(paino_sds~.,x_paino_oli_0)[,-1]

# Valitaan opetusdataksi datasta 50% ja testidataksi loput 50%

# total
opetus <- sample(1:nrow(x_paino), nrow(x_paino)/2)
testi <- (-opetus)
y_paino_opetus <- y_paino[opetus]
y_paino_testi <- y_paino[testi]

# Tarkistetaan, etta kategoriset muuttujat jakaantuvat tasaisesti

table(x[testi,]$Secretor)
table(x[opetus,]$Secretor)
table(x_paino[testi,22])
table(x_paino[opetus,22])
table(x_paino[testi,23])
table(x_paino[opetus,23])

# secretor
opetus_1 <- sample(1:nrow(x_paino_1), nrow(x_paino_1)/2)
testi_1 <- (-opetus_1)
y_paino_opetus_1 <- y_paino_1[opetus_1]
y_paino_testi_1 <- y_paino_1[testi_1]

table(x_paino_1[testi_1,22])
table(x_paino_1[opetus_1,22])
table(x_paino_1[testi_1,23])
```

```
table(x_paino_1[opetus_1,23])

# nonsecretor
opetus_0 <- sample(1:nrow(x_paino_0), nrow(x_paino_0)/2)
testi_0 <- (-opetus_0)
y_paino_opetus_0 <- y_paino_0[opetus_0]
y_paino_testi_0 <- y_paino_0[testi_0]

table(x_paino_0[testi_0,22])
table(x_paino_0[opetus_0,22])
table(x_paino_0[testi_0,23])
table(x_paino_0[opetus_0,23])

# LINEAARINEN MALLI OPETUS- JA TESTIAINEISTOLLA
# VAIN OLIGOSAKKARIDIT

# total
taysi_malli <- lm(paino_sds ~ X_2FL_nmol + X_3FL_nmol + LNnT_nmol +
X_3SL_nmol + DFLac_nmol + X_6SL_nmol + LNT_nmol +
LNFP_I_nmol + LNFP_II_nmol + LNFP_III_nmol +
LSTb_nmol + LSTc_nmol + DFLNT_nmol + LNH_nmol +
DSLNT_nmol + FLNH_nmol + DFLNH_nmol + FDSLNH_nmol +
DSLNH_nmol , data = x)
# subset = opetus ennusteen ja MSE laskemiseen
summary(taysi_malli)
taysi_malli_ennuste <- predict.lm(taysi_malli, x[testi,])
taysi_malli_ennuste
mse <- mean((taysi_malli_ennuste-y_paino_testi)^2)
mse

# secretor
taysi_malli_1 <- lm(paino_sds ~ X_2FL_nmol + X_3FL_nmol + LNnT_nmol +
X_3SL_nmol + DFLac_nmol + X_6SL_nmol + LNT_nmol +
LNFP_I_nmol + LNFP_II_nmol + LNFP_III_nmol +
LSTb_nmol + LSTc_nmol + DFLNT_nmol + LNH_nmol +
DSLNT_nmol + FLNH_nmol + DFLNH_nmol + FDSLNH_nmol +
DSLNH_nmol, data = x_sec1)
# subset = opetus_1 ennusteen ja MSE laskemiseen
summary(taysi_malli_1)
taysi_malli_ennuste_1 <- predict.lm(taysi_malli_1, x_sec1[testi_1,])
mse_1 <- mean((taysi_malli_ennuste_1-y_paino_testi_1)^2)
mse_1
```

```
# nonsecretor
taysi_malli_0 <- lm(paino_sds ~ X_2FL_nmol + X_3FL_nmol + LNnT_nmol +
X_3SL_nmol + DFLac_nmol + X_6SL_nmol + LNT_nmol +
LNFP_I_nmol + LNFP_II_nmol + LNFP_III_nmol +
LSTb_nmol + LSTc_nmol + DFLNT_nmol + LNH_nmol +
DSLNT_nmol + FLNH_nmol + DFLNH_nmol + FDSLNH_nmol +
DSLNH_nmol, data = x_sec0)
# subset = opetus_0 ennusteen ja MSE laskemiseen
summary(taysi_malli_0)
taysi_malli_ennuste_0 <- predict.lm(taysi_malli_0, x_sec0[testi_0,])
mse_0 <- mean((taysi_malli_ennuste_0-y_paino_testi_0)^2)
mse_0

# LINEAARINEN MALLI OPETUS- JA TESTIAINEISTOLLA
# SELITTAVAT TEKIJAT MUKANA

# total
taysi_malli <- lm(paino_sds ~ X_2FL_nmol + X_3FL_nmol + LNnT_nmol +
X_3SL_nmol + DFLac_nmol + X_6SL_nmol + LNT_nmol + LNFP_I_nmol +
LNFP_II_nmol + LNFP_III_nmol + LSTb_nmol + LSTc_nmol + DFLNT_nmol +
LNH_nmol + DSLNT_nmol + FLNH_nmol + DFLNH_nmol + FDSLNH_nmol +
DSLNH_nmol + SUKUP_2 + SYNTYMA_ikapaino_sds_v + SYNNYTYSTAPALUOKKA_3 +
BMI_aiti, data = x)
# subset = opetus ennusteen ja MSE laskemiseen
summary(taysi_malli)
taysi_malli_ennuste <- predict.lm(taysi_malli, x[testi,])
mse <- mean((taysi_malli_ennuste-y_paino_testi)^2)
mse

# secretor
taysi_malli_1 <- lm(paino_sds ~ X_2FL_nmol + X_3FL_nmol + LNnT_nmol +
X_3SL_nmol +DFLac_nmol + X_6SL_nmol + LNT_nmol + LNFP_I_nmol +
LNFP_II_nmol + LNFP_III_nmol + LSTb_nmol + LSTc_nmol + DFLNT_nmol +
LNH_nmol + DSLNT_nmol + FLNH_nmol + DFLNH_nmol + FDSLNH_nmol +
DSLNH_nmol + SUKUP_2 + SYNTYMA_ikapaino_sds_v + SYNNYTYSTAPALUOKKA_3 +
BMI_aiti, data = x_sec1)
# subset = opetus_1 ennusteen ja MSE laskemiseen
summary(taysi_malli_1)
taysi_malli_ennuste_1 <- predict.lm(taysi_malli_1, x_sec1[testi_1,])
mse_1 <- mean((taysi_malli_ennuste_1-y_paino_testi_1)^2)
mse_1
```

```
# nonsecretor
taysi_malli_0 <- lm(paino_sds ~ X_2FL_nmol + X_3FL_nmol +
LNnT_nmol + X_3SL_nmol + DFLac_nmol + X_6SL_nmol + LNT_nmol +
LNFP_I_nmol + LNFP_II_nmol + LNFP_III_nmol + LSTb_nmol +
LSTc_nmol + DFLNT_nmol + LNH_nmol + DSLNT_nmol + FLNH_nmol +
DFLNH_nmol + FDSLNH_nmol + DSLNH_nmol + SUKUP_2 +
SYNTYMA_ikapaino_sds_v + SYNNYTYSTAPALUOKKA_3 +
BMI_aiti, data = x_sec0)
# subset = opetus_0 ennusteen ja MSE laskemiseen
summary(taysi_malli_0)
taysi_malli_ennuste_0 <- predict.lm(taysi_malli_0, x_sec0[testi_0,])
mse_0 <- mean((taysi_malli_ennuste_0-y_paino_testi_0)^2)
mse_0

# HARJUREGRESSIO TESTI JA OPETUSAINEISTOLLA
# VAIN OLIGOSAKKARIDIT

# total
k.risti <- cv.glmnet(x_paino_oli[opetus,], y_paino_opetus, alpha=0)
plot(k.risti)
paras_l <- k.risti$lambda.min
paras_l
harjuregressio.malli <- glmnet(x_paino_oli[opetus,], y_paino_opetus,
alpha = 0, lambda = paras_l, standardize = FALSE)
harjuregressio.ennuste <- predict(harjuregressio.malli, s=paras_l,
newx = x_paino_oli[testi,])
mean((harjuregressio.ennuste-y_paino_testi)^2)
harjuregressio.malli <- glmnet(x_paino_oli, y_paino, alpha = 0,
lambda = paras_l, standardize = FALSE)
predict(harjuregressio.malli, type = "coefficients", s = paras_l)
harjuregressio.kuva <- glmnet(x_paino_oli[opetus,], y_paino_opetus,
alpha = 0,standardize = FALSE)
plot.glmnet(harjuregressio.kuva, xvar = "lambda")
abline(v=log(paras_l))

# secretor
k.risti_1 <- cv.glmnet(x_paino_oli_1[opetus_1,], y_paino_opetus_1,
alpha=0)
plot(k.risti_1)
paras_l_1 <- k.risti_1$lambda.min
paras_l_1
harjuregressio.malli <- glmnet(x_paino_oli_1[opetus_1,],
```

```
y_paino_opetus_1, alpha = 0,
lambda = paras_l_1, standardize = FALSE)
harjuregressio.ennuste <- predict(harjuregressio.malli,
s=paras_l_1, newx = x_paino_oli_1[testi_1,])
mean((harjuregressio.ennuste-y_paino_testi_1)^2)
harjuregressio.malli <- glmnet(x_paino_oli_1, y_paino_1,
alpha = 0, lambda = paras_l_1, standardize = FALSE)
predict(harjuregressio.malli, type = "coefficients", s = paras_l_1)

# nonsecretor
k.risti_0 <- cv.glmnet(x_paino_oli_0[opetus_0,], y_paino_opetus_0,
alpha=0)
plot(k.risti_0)
paras_l_0 <- k.risti_0$lambda.min
paras_l_0
harjuregressio.malli <- glmnet(x_paino_oli_0[opetus_0,],
y_paino_opetus_0, alpha = 0,
lambda = paras_l_0, standardize = FALSE)
harjuregressio.ennuste <- predict(harjuregressio.malli,
s=paras_l_0, newx = x_paino_oli_0[testi_0,])
mean((harjuregressio.ennuste-y_paino_testi_0)^2)
harjuregressio.malli <- glmnet(x_paino_oli_0, y_paino_0,
alpha = 0, lambda = paras_l_0, standardize = FALSE)
predict(harjuregressio.malli, type = "coefficients", s = paras_l_0)

# HARJUREGRESSIO TESTI JA OPETUSAINEISTOLLA
# SELITTAVAT TEKIJAT MUKANA

# total
k.risti <- cv.glmnet(x_paino[opetus,], y_paino_opetus, alpha=0)
plot(k.risti)
paras_l <- k.risti$lambda.min
paras_l
harjuregressio.malli <- glmnet(x_paino[opetus,], y_paino_opetus,
alpha = 0,lambda = paras_l, standardize = FALSE)
harjuregressio.ennuste <- predict(harjuregressio.malli,
s=paras_l, newx = x_paino[testi,])
mean((harjuregressio.ennuste-y_paino_testi)^2)
harjuregressio.malli <- glmnet(x_paino, y_paino, alpha = 0,
lambda = paras_l, standardize = FALSE)
predict(harjuregressio.malli, type = "coefficients", s = paras_l)
harjuregressio.kuva <- glmnet(x_paino[opetus,], y_paino_opetus,
alpha = 0,  standardize = FALSE)
```

```
plot.glmnet(harjuregressio.kuva, xvar = "lambda")
abline(v=log(paras_l))

# secretor
k.risti_1 <- cv.glmnet(x_paino_1[opetus_1,], y_paino_opetus_1,
 alpha=0)
plot(k.risti_1)
paras_l_1 <- k.risti_1$lambda.min
paras_l_1
harjuregressio.malli <- glmnet(x_paino_1[opetus_1,],
y_paino_opetus_1, alpha = 0,
lambda = paras_l_1, standardize = FALSE)
harjuregressio.ennuste <- predict(harjuregressio.malli,
s=paras_l_1, newx = x_paino_1[testi_1,])
mean((harjuregressio.ennuste-y_paino_testi_1)^2)
harjuregressio.malli <- glmnet(x_paino_1, y_paino_1,
alpha = 0, lambda = paras_l_1, standardize = FALSE)
predict(harjuregressio.malli, type = "coefficients", s = paras_l_1)

# nonsecretor
k.risti_0 <- cv.glmnet(x_paino_0[opetus_0,], y_paino_opetus_0,
alpha=0)
plot(k.risti_0)
paras_l_0 <- k.risti_0$lambda.min
paras_l_0
harjuregressio.malli <- glmnet(x_paino_0[opetus_0,],
y_paino_opetus_0, alpha = 0,
lambda = paras_l_0, standardize = FALSE)
harjuregressio.ennuste <- predict(harjuregressio.malli,
s=paras_l_0, newx = x_paino_0[testi_0,])
mean((harjuregressio.ennuste-y_paino_testi_0)^2)
harjuregressio.malli <- glmnet(x_paino_0, y_paino_0, alpha = 0,
lambda = paras_l_0, standardize = FALSE)
predict(harjuregressio.malli, type = "coefficients", s = paras_l_0)

# LASSO TESTI JA OPETUSAINEISTOLLA
# VAIN OLIGOSAKKARIDIT

# total
k.risti <- cv.glmnet(x_paino_oli[opetus,], y_paino_opetus,
alpha=1)
plot(k.risti)
paras_l <- k.risti$lambda.min
```

```
paras_l
lasso.malli <- glmnet(x_paino_oli[opetus,], y_paino_opetus,
alpha = 1,lambda = paras_l, standardize = FALSE)
lasso.ennuste <- predict(lasso.malli, s=paras_l,
newx = x_paino_oli[testi,])
mean((lasso.ennuste-y_paino_testi)^2)
lasso.malli <- glmnet(x_paino_oli, y_paino, alpha = 1,
lambda = paras_l, standardize = FALSE)
predict(lasso.malli, type = "coefficients", s = paras_l)
lasso.kuva <- glmnet(x_paino_oli[opetus,], y_paino_opetus, alpha = 1,
standardize = FALSE)
plot.glmnet(lasso.kuva, xvar = "lambda")
abline(v=log(paras_l))

# secretor
k.risti_1 <- cv.glmnet(x_paino_oli_1[opetus_1,], y_paino_opetus_1,
alpha=1)
plot(k.risti_1)
paras_l_1 <- k.risti_1$lambda.min
paras_l_1
lasso.malli <- glmnet(x_paino_oli_1[opetus_1,],
y_paino_opetus_1, alpha = 1,
lambda = paras_l_1, standardize = FALSE)
lasso.ennuste <- predict(lasso.malli, s=paras_l_1,
newx = x_paino_oli_1[testi_1,])
mean((lasso.ennuste-y_paino_testi_1)^2)
lasso.malli <- glmnet(x_paino_oli_1, y_paino_1, alpha = 1,
lambda = paras_l_1, standardize = FALSE)
predict(lasso.malli, type = "coefficients", s = paras_l_1)

# nonsecretor
k.risti_0 <- cv.glmnet(x_paino_oli_0[opetus_0,], y_paino_opetus_0,
alpha=1)
plot(k.risti_0)
paras_l_0 <- k.risti_0$lambda.min
paras_l_0
lasso.malli <- glmnet(x_paino_oli_0[opetus_0,],
y_paino_opetus_0, alpha = 1,
lambda = paras_l_0, standardize = FALSE)
lasso.ennuste <- predict(lasso.malli, s=paras_l_0,
newx = x_paino_oli_0[testi_0,])
mean((lasso.ennuste-y_paino_testi_0)^2)
lasso.malli <- glmnet(x_paino_oli_0, y_paino_0, alpha = 1,
```

```
lambda = paras_l_0, standardize = FALSE)
predict(lasso.malli, type = "coefficients", s = paras_l_0)


# LASSO TESTI JA OPETUSAINEISTOLLA
# SELITTAVAT TEKIJAT MUKANA

# total
k.risti <- cv.glmnet(x_paino[opetus,], y_paino_opetus,
alpha=1)
plot(k.risti)
paras_l <- k.risti$lambda.min
paras_l
lasso.malli <- glmnet(x_paino[opetus,],
y_paino_opetus, alpha = 1,
lambda = paras_l, standardize = FALSE)
lasso.ennuste <- predict(lasso.malli, s=paras_l,
newx = x_paino[testi,])
mean((lasso.ennuste-y_paino_testi)^2)
lasso.malli <- glmnet(x_paino, y_paino, alpha = 1,
lambda = paras_l, standardize = FALSE)
predict(lasso.malli, type = "coefficients", s = paras_l)
lasso.kuva <- glmnet(x_paino[opetus,], y_paino_opetus, alpha = 1,
standardize = FALSE)
plot.glmnet(lasso.kuva, xvar = "lambda")
abline(v=log(paras_l))

# secretor
k.risti_1 <- cv.glmnet(x_paino_1[opetus_1,], y_paino_opetus_1,
alpha=1)
plot(k.risti_1)
paras_l_1 <- k.risti_1$lambda.min
paras_l_1
lasso.malli <- glmnet(x_paino_1[opetus_1,],
y_paino_opetus_1, alpha = 1,
lambda = paras_l_1, standardize = FALSE)
lasso.ennuste <- predict(lasso.malli, s=paras_l_1,
newx = x_paino_1[testi_1,])
mean((lasso.ennuste-y_paino_testi_1)^2)
lasso.malli <- glmnet(x_paino_1, y_paino_1, alpha = 1,
lambda = paras_l_1, standardize = FALSE)
predict(lasso.malli, type = "coefficients", s = paras_l_1)
```

```
# nonsecretor
k.risti_0 <- cv.glmnet(x_paino_0[opetus_0,], y_paino_opetus_0,
alpha=1)
plot(k.risti_0)
paras_l_0 <- k.risti_0$lambda.min
paras_l_0
lasso.malli <- glmnet(x_paino_0[opetus_0,],
y_paino_opetus_0, alpha = 1,
lambda = paras_l_0, standardize = FALSE)
lasso.ennuste <- predict(lasso.malli, s=paras_l_0,
newx = x_paino_0[testi_0,])
mean((lasso.ennuste-y_paino_testi_0)^2)
lasso.malli <- glmnet(x_paino_0, y_paino_0, alpha = 1,
lambda = paras_l_0, standardize = FALSE)
predict(lasso.malli, type = "coefficients", s = paras_l_0)

# LINEAARINEN MALLI LASSON MALLILLE
# VAIN OLIGOSAKKARIDIT

# total
taysi_malli <- lm(paino_sds ~ X_3FL_nmol + DFLac_nmol +
DSLNT_nmol, data = x, subset = opetus)
summary(taysi_malli)
taysi_malli_ennuste <- predict.lm(taysi_malli, x[testi,])
mse <- mean((taysi_malli_ennuste-y_paino_testi)^2)
mse

# secretor
taysi_malli_1 <- lm(paino_sds ~ LNnT_nmol + DFLac_nmol +
LNT_nmol + DFLNT_nmol + DSLNT_nmol + DSLNH_nmol, data = x_sec1,
subset = opetus_1)
summary(taysi_malli_1)
taysi_malli_ennuste_1 <- predict.lm(taysi_malli_1, x_sec1[testi_1,])
mse <- mean((taysi_malli_ennuste_1-y_paino_testi_1)^2)
mse

# LINEAARINEN MALLI LASSON MALLILLE
# SELITTAVAT TEKIJAT MUKANA

# total
taysi_malli <- lm(paino_sds ~ X_2FL_nmol + X_3FL_nmol +
X_6SL_nmol + DSLNT_nmol + SYNTYMA_ikapaino_sds_v +
```

```
BMI_aiti, data = x, subset = opetus)
summary(taysi_malli)
taysi_malli_ennuste <- predict.lm(taysi_malli, x[testi,])
mse <- mean((taysi_malli_ennuste-y_paino_testi)^2)
mse

# secretor
taysi_malli_1 <- lm(paino_sds ~ SYNTYMA_ikapaino_sds_v ,
data = x_sec1, subset = opetus_1)
summary(taysi_malli_1)
taysi_malli_ennuste_1 <- predict.lm(taysi_malli_1, x_sec1[testi_1,])
mse <- mean((taysi_malli_ennuste_1-y_paino_testi_1)^2)
mse
```