Tilastotieteen pro gradu -tutkielma

Tilastollinen päättely kanonisessa korrelaatioanalyysissä

Tatiana Denisova

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos

Elokuu 2013

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos

Denisova, Tatiana: Tilastollinen päättely kanonisessa korrelaatioanalyysissä

Tilastotieteen pro-gradu tutkielma, 34 sivua, liite
elokuu 2013

Tiivistelmä

Liikunnalla on keskeinen rooli terveyden edistämisessä. Koska terveyden ja fyysisen aktiivisuuden välistä yhteyttä pidetään tärkeänä, on mielenkiintoista selvittää, ovatko terveys ja fyysinen aktiivisuus yhteydessä toisiinsa. Tässä työssä tutkitaan, vallitseeko terveyden ja fyysisen aktiivisuuden välillä todellakin riippuvuus, kun molemmat terveys- ja liikunta-aktiivisuusmuuttujajoukot koostuvat useista muuttujista. Tätä ongelmaa lähestytään käyttäen kanonista korrelaatioanalyysiä. Sovellusaineistona on Jyväskylän yliopistossa kerätty aineisto, jossa terveyttä kuvaavina muuttujina käytetään verenpainetta, sisäelinten ympärillä olevaa rasvaa ja triglyseridipitoisuutta. Liikunta-aktiivisuutta kuvaavat muuttujat ovat kyselylomakkeella saadut omat arvioit liikunta-aktiivisuudesta sekä reisilihasten päivittäinen aktiivisuus ja päivän pisin epäaktiivisuusaika. Tarkoituksena on löytää molemmista ryhmästä muuttujien sellaiset lineaarikombinaatiot, että lineaarikombinaatioiden väliset korrelaatiot ovat mahdollisimman suuria. Kanonisen korrelaatioanalyysin tuloksena saadaan kanoniset korrelaatiot ja vastaavat kanoniset muuttujat. Esitellään uskottavuusosamäärätesti korrelaatikertoimien merkitsevyyden testaamiseksi. Vaihtoehtoisena lähestymistapana esitellään permutaatiotesti, jonka käyttö ei edellytä minkälaista jakaumaoletusta. Osoittautuu, että testien tuomat tulokset ovat yhdenmukaisia. Tässä tutkielmassa lasketaan kanonisten muuttujien eli saatujen lineaarikombinaatioiden kertoimille luottamusvälejä ja -alueita bootstrap-menetelmän avulla. Lasketut luottamusvälit estimaateille ovat leveitä, mikä hankaloittaa luotettavien tulkintojen tekemistä. Aineiston perusteella päädytään tulokseen, että sekä miesten että naisten ryhmissä terveys- ja liikunta-aktiivisuusmuuttujien välillä ei ole todettu olevan riippuvuutta, kun iän vaikutus vakioidaan.

Avainsanat: Kanoninen korrelaatioanalyysi, kanoninen korrelaatio, kanoninen muuttuja, permutaatiotesti, bootstrap, luottamusväli.

Sisältö

LUKU 1

Johdanto

Fyysinen aktiivisuus ja terveys ovat olennaisia osia ihmisten elämää. Liikunta edistää terveyttä. Näin ollen ajatellaan, että liikunnan ja terveyden yhteys on olemassa ja kyseistä yhteyttä pidetään tärkeänä. Tällöin kysymykseksi nousee, mitkä tekijät ovat keskeisiä liikunnan ja terveyden välisessä suhteessa. On kiinnostava selvittää, vallitseeko terveyden ja fyysisen aktiivisuuden välillä todellakin riippuvuus. Työssä pyritään vastaamaan tähän kysymykseen tarkastelemalla useista muuttujista koostuvia muuttujajoukkoja yhtä aikaa. Koska muuttujajoukkojen väliset riippuvuudet ovat kiinnostuksen kohteena ja kysymys on terveys- ja liikunta-aktiivisuusmuuttujien samanaikaisesta tarkastelusta, niin ongelman ratkaisemiseen käytetään kanonista korrelaatioanalyysiä.

Kanoninen korrelaatioanalyysi (*Canonical Correlation Analysis, CCA*) kuuluu klassisiin monimuuttujamenetelmiin. Sillä tutkitaan kahden eri muuttujajoukon välistä riippuvuutta. Kun molemmissa ryhmissä on useita muuttujia, on kyse kanonisesta korrelaatioanalyysistä. Voidaan ajatella, että toinen ryhmä koostuu selittävistä ja toinen selitettävistä muuttujista. Tällöin voidaan puhua selittävien ja selitettävien muuttujien välisten riippuvuussuhteiden samanaikaisesta tarkastelusta. Kanonisen korrelaatioanalyysin ideana on löytää jokaisesta muuttujajoukosta alkuperäisten muuttujien sellainen lineaarikombinaatio, että muodostettujen lineaarikombinaatioiden välinen korrelaatio on suurin. Menetelmän keskeisen teorian kehitti H. Hotelling (1936). Menetelmää sovelletaan muun muassa talous- ja lääketieteessä.

Kanonisessa korrelaatioanalyysissä kuten monissa monimuuttujamenetelmissä kysymys on ominaisarvojen ja ominaisvektoreiden laskemisesta ja analysoinnista. Analyysi johtaa tietyn ominaisarvotehtävän ratkaisuun. Kun ominaisarvot ja -vektorit las-

ketaan otoksesta, niiden otosjakaumat ovat tuntemattomia. Bootstrap-menetelmä on eräs likimääräinen ratkaisu. Tutkielman tavoitteena on kehittää bootstrap-algoritmi, jolla voi laskea kanonisten korrelaatioiden ja vastaavien vektoreiden luottamusvälejä ja -alueita.

Sovellusaineistona käytetään Jyväskylän yliopiston liikuntabiologian laitokselta peräisin olevaa aineistoa, johon on kerätty ihmisten terveyttä koskevat arviot sekä fyysisen aktiivisuuden mittaukset. Tässä tutkielmassa tutkitaan terveys- ja liikunta-aktiivisuusmuuttujien välisiä relaatioita kanonisen korrelaatioanalyysin avulla ja estimoidaan kanoniset muuttujat. Bootstrap-algoritmi tuottaa luottamusvälejä kanonisille muuttujille eli lineaarikombinaatioiden kertoimille.

Tutkielma etenee siten, että ensin esitetään kanoniseen korrelaatioanalyysiin liittyvää teoriaa. Kahden muuttujajoukon riippuvuuden testaaminen permutaatiotestin avulla esitellään luvussa 3. Luvussa 4 käydään läpi algoritmi luottamusvälien ja -alueen laskemiseksi. Luvussa 5 esitellään aineisto ja saadut tulokset.

LUKU 2

Kanoninen korrelaatioanalyysi

Kanonisen korrelaatioanalyysin lähtökohtana ovat kaksi muuttujajoukkoa, joiden välillä vallitsevat riippuvuussuhteet ovat mielenkiinnon kohteena. Kanonisessa korrelaatioanalyysissä, kuten monissa monimuuttujamenetelmissä, tarkasteltavien muuttujien tulee noudattaa normaalijakaumaa. Käytettäessä menetelmää kuvailevaan analyysiin normaalisuusoletusta ei kuitenkaan tarvita.

2.1 Kanoninen korrelaatio ja kanoniset muuttujat

Tämän luvun teoria perustuu teoksiin Anderson (2003) ja Dillon et al. (1984).

Olkoot $\mathbf{X}^{\mathrm{T}} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_m)$ m-ulotteinen satunnaisvektori ja $\mathbf{Y}^{\mathrm{T}} = (\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_p)$ p-ulotteinen satunnaisvektori ($m \leq p$) sekä $\boldsymbol{\mu}_x$ ja $\boldsymbol{\mu}_y$ niiden odotusarvovektorit vastaavasti. Lisäksi merkitään

$$\begin{aligned}
\boldsymbol{\Sigma}_{xx} &= \mathrm{E}\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^{\mathrm{T}}\}, \\
\boldsymbol{\Sigma}_{yy} &= \mathrm{E}\{(\mathbf{Y} - \boldsymbol{\mu}_y)(\mathbf{Y} - \boldsymbol{\mu}_y)^{\mathrm{T}}\}, \\
\boldsymbol{\Sigma}_{xy} &= \mathrm{E}\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{Y} - \boldsymbol{\mu}_y)^{\mathrm{T}}\},
\end{aligned}$$

missä $\boldsymbol{\Sigma}_{xx}$ on $m \times m$ $\mathbf{X}$:n kovarianssimatriisi, $\boldsymbol{\Sigma}_{yy}$ on $p \times p$ $\mathbf{Y}$:n kovarianssimatriisi ja $\boldsymbol{\Sigma}_{xy}$ on $m \times p$ $\mathbf{X}$:n ja $\mathbf{Y}$:n välinen kovarianssimatriisi, jonka asteluku on $r \leq \min(m, p)$. Oletetaan yksinkertaisuuden vuoksi, että $\boldsymbol{\mu}_x = 0$ ja $\boldsymbol{\mu}_y = 0$. Ajatuksena on muodostaa vektoreista $\mathbf{X}$ ja $\mathbf{Y}$ lineaarikombinaatiot

$$\mathrm{X}^* = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m, \tag{1}$$

$$Y^* = \boldsymbol{\beta}^T \mathbf{Y} = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_p y_p \tag{2}$$

siten, että muuttujien X^* ja Y^* välinen korrelaatiokerroin on suurin. Vektorit $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$ valitaan siten, että X^*:n ja Y^*:n varianssit ovat ykkösiä, toisin sanoen

$$\begin{aligned} \mathrm{Var}(X^*) &= \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha} = 1, \\ \mathrm{Var}(Y^*) &= \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta} = 1. \end{aligned} \tag{3}$$

Todetaan, että

$$\begin{aligned} \mathrm{E}(X^*) &= \mathrm{E}(\boldsymbol{\alpha}^T \mathbf{X}) = \boldsymbol{\alpha}^T \mathrm{E}(\mathbf{X}) = 0, \\ \mathrm{E}(Y^*) &= \mathrm{E}(\boldsymbol{\beta}^T \mathbf{Y}) = \boldsymbol{\beta}^T \mathrm{E}(\mathbf{Y}) = 0. \end{aligned} \tag{4}$$

Näin ollen X^* ja Y^* ovat normeerattuja. Uusien muuttujien X^*:n ja Y^*:n välinen korrelaatiokerroin on

$$\rho(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta}}{\{(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha})(\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta})\}^{1/2}} = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta}. \tag{5}$$

Tarkoituksena nyt on siis etsiä sellaiset painovektorit $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$, että ne maksimoivat X^*:n ja Y^*:n välisen korrelaatiokertoimen. Toisin sanoen maksimoidaan korrelaatiokerroin (5) rajoitteilla (3). Sovelletaan edellä mainitun optimointiongelman ratkaisemiseen Lagrangen menetelmää, jolloin maksimoitava lauseke on

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta} - \frac{1}{2}\lambda(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha} - 1) - \frac{1}{2}\mu(\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta} - 1), \tag{6}$$

missä λ ja μ ovat Lagrangen kertoimia. Derivoimalla lauseke (6) vektoreiden $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$ suhteen ja asettamalla osittaisderivaatat nolliksi saadaan

$$\frac{\partial \psi(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} = \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta} - \lambda \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha} = 0, \tag{7}$$

$$\frac{\partial \psi(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{\Sigma}_{yx} \boldsymbol{\alpha} - \mu \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta} = 0. \tag{8}$$

Kun ensimmäinen yhtälö kerrotaan vektorilla $\boldsymbol{\alpha}^T$ ja toinen yhtälö vektorilla $\boldsymbol{\beta}^T$ vasemmalta, saadaan seuraavat yhtälöt

$$\begin{aligned} \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta} - \lambda \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha} = 0, \\ \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yx} \boldsymbol{\alpha} - \mu \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta} = 0. \end{aligned}$$

Nyt $\lambda = \mu = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\beta}$, mikä seuraa tehdyistä rajoitteista $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha} = 1$ ja $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}\boldsymbol{\beta} = 1$, joten yhtälöt (7) ja (8) voidaan kirjoittaa muotoon

$$\begin{aligned} -\lambda\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha} + \boldsymbol{\Sigma}_{xy}\boldsymbol{\beta} &= \mathbf{0}, \\ \boldsymbol{\Sigma}_{yx}\boldsymbol{\alpha} - \lambda\boldsymbol{\Sigma}_{yy}\boldsymbol{\beta} &= \mathbf{0}. \end{aligned} \tag{9}$$

Yhtälöryhmää (9)vastaavaa matriisimuoto on

$$\begin{pmatrix} -\lambda\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & -\lambda\boldsymbol{\Sigma}_{yy} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \mathbf{0}. \tag{10}$$

Epätriviaali ratkaisu, joka täyttää vaatimukset (3) ja (4), saadaan, kun kerroinmatriisi kaavassa (10) on singulaarinen, eli sen determinantti

$$\begin{vmatrix} -\lambda\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & -\lambda\boldsymbol{\Sigma}_{yy} \end{vmatrix} = 0.$$

Näin saadaan $(m+p)$:nnen asteen polynomi λ:n suhteen, jolla on $(m+p)$ juurta $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{m+p}$. Aiemmin on huomattu, että $\lambda = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\beta}$ on satunnaismuuttujien $\mathrm{X}^* = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}$ ja $\mathrm{Y}^* = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Y}$ välinen korrelaatiokerroin, missä $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$ toteuttavat yhtälön (10) jollakin arvolla λ. Kun valitaan $\lambda = \lambda_1$, korrelaatiokerroin saa suurimman arvon. Oletetaan, että $\boldsymbol{\alpha}^{(1)}$ ja $\boldsymbol{\beta}^{(1)}$ ovat yhtälön (10) ratkaisut kun $\lambda = \lambda_1$. Tällöin $\mathrm{X}_1^* = \boldsymbol{\alpha}^{(1)\mathrm{T}}\mathbf{X}$ ja $\mathrm{Y}_1^* = \boldsymbol{\beta}^{(1)\mathrm{T}}\mathbf{Y}$, ja niillä on maksimikorrelaatio. Nämä muodostavat ensimmäisen kanonisen muuttujaparin. Toinen pari $(\mathrm{X}_2^*, \mathrm{Y}_2^*)$ on sellainen, että sekä X_2^* että Y_2^* eivät korreloi X_1^*:n ja Y_1^*:n kanssa, ja lineaarikombinaatioiden $\mathrm{X}_2^* = \boldsymbol{\alpha}^{(2)\mathrm{T}}\mathbf{X}$ ja $\mathrm{Y}_2^* = \boldsymbol{\beta}^{(2)\mathrm{T}}\mathbf{Y}$ välinen korrelaatio maksimoituu. Menettelyä jatketaan samaan tapaan kunnes r:nnellä askeleella saadaan lineaarikombinaatiot $\mathrm{X}_r^* = \boldsymbol{\alpha}^{(r)\mathrm{T}}\mathbf{X}, \mathrm{Y}_r^* = \boldsymbol{\beta}^{(r)\mathrm{T}}\mathbf{Y}$, joiden välinen korrelaatiokerroin on λ_r. Muistetaan, että r on $\boldsymbol{\Sigma}_{xy}$:n asteluku.

Määritellään kanoninen muuttujapari seuraavasti. Muuttujaparia $\mathrm{X}_k^* = \boldsymbol{\alpha}^{(k)\mathrm{T}}\mathbf{X}$ ja $\mathrm{Y}_k^* = \boldsymbol{\beta}^{(k)\mathrm{T}}\mathbf{Y}$, $k = 1, \ldots, r$, missä $\mathbf{X}$ ja $\mathbf{Y}$ m- ja p- ulotteisia satunnaisvektoreita $(m \leq p)$, sanotaan k. kanoniseksi muuttujapariksi ja kanonisten muuttujien X_k^* ja Y_k^* välistä maksimikorrelaatiota k. kanoniseksi korrelaatioksi, jos lineaarikombinaatioiden $\mathrm{X}_k^* = \boldsymbol{\alpha}^{(k)\mathrm{T}}\mathbf{X}$ ja $\mathrm{Y}_k^* = \boldsymbol{\beta}^{(k)\mathrm{T}}\mathbf{Y}$ varianssit ovat ykkösiä ja ne eivät korreloi aikaisempien $(k-1)$:n muuttujaparien kanssa. (Anderson 2003, 495).

Kanoninen korrelaatio voidaan myös johtaa liittyen matriisien ominaisarvoihin. Tekemällä muunnoksia yhtälöryhmässä (9) päädytään seuraaviin yhtälöihin (Dillon et al. 1984, 341):

$$\begin{aligned} \left(\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} - \lambda^2\mathbf{I}\right)\boldsymbol{\alpha} &= \mathbf{0}, \\ \left(\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} - \lambda^2\mathbf{I}\right)\boldsymbol{\beta} &= \mathbf{0}. \end{aligned} \tag{11}$$

Sekä matriisin $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$ että matriisin $\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ aste on r ($r \leq m \leq p$). Tällöin matriiseilla $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$ ja $\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ on korkeintaan r kappaletta nollasta poikkeavia ominaisarvoja $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_r^2 > 0$ ja ominaisarvot ovat kyseisillä matriiseilla samat. Nämä ominaisarvot ovat kanonisten korrelaatioiden $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$ neliöt. Ominaisarvoon $\lambda_i^2, i = 1, \ldots, r$ liittyviä ominaisvektoreita on kaksi joukkoa, toinen liittyy matriisiin $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$ ja toinen matriisiin $\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$. Matriisin $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$ ominaisvektorit $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(r)}$ saadaan kaavasta (Dillon et al. (1984, 342))

$$\boldsymbol{\alpha}^{(i)} = \frac{\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\beta}^{(i)}}{\lambda_i}, i = 1, \ldots, r.$$

Vastaavasti matriisin $\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ ominaisvektorit $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(r)}$ saadaan kaavasta (Dillon et al. 1984, 342)

$$\boldsymbol{\beta}^{(i)} = \frac{\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{\alpha}^{(i)}}{\lambda_i}, i = 1, \ldots, r.$$

Edelliset kaavat saadaan, kun kerrotaan yhtälöryhmän (9) ensimmäinen yhtälö matriisilla $\boldsymbol{\Sigma}_{xx}^{-1}$ ja toinen yhtälö matriisilla $\boldsymbol{\Sigma}_{yy}^{-1}$ vasemmalta. Kertolaskujen suorittamisen jälkeen ratkaistaan ensimmäisestä yhtälöstä vektori $\boldsymbol{\alpha}$ ja toisesta yhtälöstä vektori $\boldsymbol{\beta}$. Käytetään jatkossa seuraavaa merkintää: λ_i on teoreettinen i. kanoninen korrelaatio ja $\hat{\lambda}_i$ on sen estimaatti.

Kun jokaisessa ryhmässä on vain yksi muuttuja $\mathbf{X} = \mathrm{X}$, $\mathbf{Y} = \mathrm{Y}$, saadaan yksi kanoninen korrelaatio, joka on sama kuin X:n ja Y:n välinen Pearsonin korrelaatiokerroin. Kanonisen korrelaatioanalyysin eräs erikoistapaus on usean selittävän muuttujan regressiomalli. Se tulee kyseeseen silloin, kun toisessa ryhmässä on yksi muuttuja (esim. $m = 1$). Tällöin kanoninen korrelaatiokerroin on satunnaismuuttujan $\mathbf{X} = \mathrm{X}_1$ ja $\mathbf{Y}$:n välinen yhteiskorrelaatiokerroin.

2.2 Kanonisten muuttujien ominaisuudet

Seuraavaksi esitetään kanonisten muuttujien ominaisuudet teoksen Rao (2002) mukaan.

Oletetaan, että $r = m \leq p$ ja λ_i, $i = 1, 2, \ldots, r$ on i. kanoninen korrelaatio. Olkoon $\mathrm{X}_i^* = \boldsymbol{\alpha}^{(i)\mathrm{T}}\mathbf{X}$ ja $\mathrm{Y}_i^* = \boldsymbol{\beta}^{(i)\mathrm{T}}\mathbf{Y}$ i. kanoninen muuttujapari.

(i) Muuttujajoukosta $\mathbf{X}$ muodostetut kanoniset muuttujat (lineaarikombinaatiot) ovat keskenään korreloimattomia. Sama pätee ryhmän $\mathbf{Y}$ kanonisille muuttujille.

(a) $$\mathrm{Cor}(\boldsymbol{\alpha}^{(i)\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}^{(j)\mathrm{T}}\mathbf{X}) = \begin{cases} 1, & \text{kun } i = j \\ 0, & \text{kun } i \neq j \end{cases}$$

(b) $$\mathrm{Cor}(\boldsymbol{\beta}^{(i)\mathrm{T}}\mathbf{Y}, \boldsymbol{\beta}^{(j)\mathrm{T}}\mathbf{Y}) = \begin{cases} 1, & \text{kun } i = j \\ 0, & \text{kun } i \neq j \end{cases}$$

(ii) Saman muuttujaparin muodostavat kanoniset muuttujat $\mathrm{X}_i^* = \boldsymbol{\alpha}^{(i)\mathrm{T}}\mathbf{X}$ ja $\mathrm{Y}_i^* = \boldsymbol{\beta}^{(i)\mathrm{T}}\mathbf{Y}$ korreloivat keskenään.

(a) $$\mathrm{Cor}(\boldsymbol{\alpha}^{(i)\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}^{(i)\mathrm{T}}\mathbf{Y}) = \begin{cases} \lambda_i > 0, & \text{kun } i = 1, \ldots, r \\ 0, & \text{kun } i > r \end{cases}$$

(b) $$\mathrm{Cor}(\boldsymbol{\alpha}^{(i)\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}^{(j)\mathrm{T}}\mathbf{Y}) = 0, \quad \text{kun } i \neq j.$$

2.3 Otokseen perustuva kanoninen korrelaatioanalyysi

Edellä esitelty teoria perustuu teoreettisiin kovarianssimatriiseihin $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{xy}$, $\boldsymbol{\Sigma}_{yy}$, jotka käytännössä eivät ole kuitenkaan tunnettuja. Tällöin kanonisia korrelaatioita ja muuttujia estimoidessa edellä mainitut kovarianssimatriisit korvataan otoksesta lasketuilla estimaateilla.

Oletetaan, että

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{bmatrix}, \ldots, \begin{bmatrix} \mathbf{x}_n \\ \mathbf{y}_n \end{bmatrix}$$

on n:n alkion otos $(m+p)$-ulotteisesta multinormaalijakaumasta

$$N\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{xy} \end{bmatrix} \right).$$

Kovarianssimatriisin $\boldsymbol{\Sigma}$ estimaattina käytetään otoskovarianssimatriisia

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{xy} \end{bmatrix},$$

missä

$$\mathbf{S}_{xx} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^{\mathrm{T}},$$

$$\mathbf{S}_{yy} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^{\mathrm{T}},$$

$$\mathbf{S}_{xy} = \mathbf{S}_{yx}^{\mathrm{T}} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{y}_j - \bar{\mathbf{y}})^{\mathrm{T}}.$$

Kanonisten korrelaatioiden estimaatit saadaan matriisin $\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}$ ominaisarvojen $\hat{\lambda}_1^2 \geq \hat{\lambda}_2^2 \geq \ldots \geq \hat{\lambda}_m^2 > 0$ neliöjuurina. Vastaavat ominaisvektorit ratkaistaan yhtälöistä

$$\begin{aligned}\left(\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx} - \hat{\lambda}_j^2\mathbf{I}\right)\mathbf{a}^{(j)} &= \mathbf{0}, \\ \left(\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} - \hat{\lambda}_j^2\mathbf{I}\right)\mathbf{b}^{(j)} &= \mathbf{0}, \quad j = 1, \ldots, m.\end{aligned} \tag{12}$$

Huomaa, että matriisien aste edellä olevissa yhtälöissä on m, toisin sanoen matriiseilla on m nollaa suurempaa ominaisarvoa. Vaikka matriisi $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$ olisi vajaaasteinen, eli matriisin aste $< m$, otoksesta laskettu vastaava matriisi on täysiasteinen todennäköisyydellä 1 (Rao 2002, 586). Jälkimmäisestä yhtälöstä (12) saadaan ominaisarvoa nolla vastaavat ominaisvektorit $\mathbf{b}^{(m+1)}, \ldots, \mathbf{b}^{(p)}$, joita tarvitaan jatkossa.

Muuttujat kuvaavat erilaisia ominaisuuksia ja usein eivät ole keskenään vertailukelpoisia, koska ne on mitattu eri skaaloissa. Tämän takia voidaan käyttää kovarianssimatriisin sijasta sen standardoitua muotoa eli korrelaatiomatriisia. Silloin otoskorrelaatiomatriiseista $\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$ ja $\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$ lasketut kanoniset korrelaatiot ovat samat kuin otoskovarianssimatriisien tapauksessa, sen sijaan ominaisvektorit eivät ole samat standardoinnin vuoksi.

2.4 Estimointi ja testaus

Tämän luvun teoria perustuu teokseen Dillon et al. (1984). Kanonisen korrelaatioanalyysin tuloksena saadaan useat kanoniset muuttujaparit ja niitä vastaavat kanoniset korrelaatiot. Jotta pystyttäisiin luotettavasti päättelemään, tuoko kyseinen muuttujapari lisäselitystä muuttujien väliseen yhteyteen ja mitkä kanoniset muuttujat nousevat oleellisiksi tulkintojen kannalta, on olennaista testata saatujen kanonisten muuttujaparien tilastollista merkitsevyyttä. Bartlettin testi on eräs menetelmä, jota käytetään kanonisten korrelaatioiden merkitsevyyden testaamiseksi. Bartlettin testillä voidaan testata kanonisten korrelaatioiden merkitsevyyttä sekä yksittäisen kanonisen muuttujaparin että kaikkien kanonisten muuttujaparien osalta.

Oletetaan edelleen,

$$\left[\begin{array}{c}\mathbf{x}_1\\ \mathbf{y}_1\end{array}\right], \left[\begin{array}{c}\mathbf{x}_2\\ \mathbf{y}_2\end{array}\right], \ldots, \left[\begin{array}{c}\mathbf{x}_n\\ \mathbf{y}_n\end{array}\right]$$

n:n alkion satunnaisotos $(m+p)$-ulotteisesta multinormaalijakaumasta

$$N\left(\left[\begin{array}{c}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{array}\right], \left[\begin{array}{cc}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{xy}\end{array}\right]\right).$$

Nollahypoteesi, jonka mukaan vektorit $\mathbf{X}$ ja $\mathbf{Y}$ ovat riippumattomia vastaa nollahypoteesia, että $\boldsymbol{\Sigma}_{xy} = \mathbf{0}$. Tätä hypoteesia voidaan testata käyttämällä Bartlettin χ^2-approksimaatiota Wilksin lambda -jakaumalle. Määritellään

$$q = -\left[n - \frac{1}{2}(m+p+1)\right] \ln \Lambda, \tag{13}$$

missä

$$\Lambda = \prod_{j=1}^{m}(1-\hat{\lambda}_j^2) = \frac{|\mathbf{S}_{xx} - \mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}|}{|\mathbf{S}_{xx}|} = |\mathbf{I} - \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}|.$$

Kaavan (13) testisuure noudattaa likimäärin χ^2-jakaumaa vapausastein mp, kun nollahypoteesi $\boldsymbol{\Sigma}_{xy} = \mathbf{0}$ pitää paikkansa. Yllä olevassa kaavassa n on otoskoko, m ja p ovat $\mathbf{X}$:n ja $\mathbf{Y}$:n komponenttien lukumäärät ja $\hat{\lambda}_j^2$ on otoksesta laskettu j. ominaisarvo. Jos testisuureen arvo on asetettua kriittistä arvoa suurempi, ainakin yksi kanoninen korrelaatio eroaa merkitsevästi nollasta. Tässä tapauksessa se on ensimmäinen ja suurin kanonisista korrelaatioista. Sen jälkeen testataan, eroavatko nollasta jäljellä olevat kanoniset korrelaatiot $\lambda_2, \ldots, \lambda_m$. Testisuure voidaan kirjoittaa seuraavassa muodossa

$$q = -\left[n - \frac{1}{2}(m+p+1)\right] \sum_{j=2}^{m} \ln(1-\hat{\lambda}_j^2),$$

joka noudattaa nollahypoteesin $\lambda_2 = \cdots = \lambda_m = 0$ ollessa voimassa χ^2-jakaumaa vapausastein $(m-1)(p-1)$. Testaamista voidaan jatkaa samaan tapaan käymällä läpi jäljellä olevat kanoniset korrelaatiot, kunnes ei enää löydy tilastollisesti merkitseviä korrelaatioita. Näin jäljellä olevien ($(k+1)$:nnestä m:hen) kanonisten korrelaatioiden nollasta poikkeavuuden testaamiseksi käytetään testisuuretta

$$q = -\left[n - \frac{1}{2}(m+p+1)\right] \sum_{j=k+1}^{m} \ln(1-\hat{\lambda}_j^2), \tag{14}$$

joka nollahypoteesin $\lambda_{k+1} = \cdots = \lambda_m = 0$ pätiessä noudattaa χ^2-jakaumaa vapausastein $(m-k)(p-k)$.

LUKU 3

Permutaatiotesti

Monella tilastollisella testillä on omat edellytyksensä, joiden on oltava voimassa, jotta testiä voidaan käyttää ja testin tulokset olisivat luotettavia. Oletukset yleensä koskevat jakaumaominaisuuksia kuten esimerkiksi perusjoukon normaalijakautuneisuutta. Kuitenkin käytännössä satunnaismuuttujan jakauma populaatiossa ei välttämättä ole tunnettu. On olemassa epäparametrisia tai parametrittomia testejä, jotka toimivat ja antavat luotettavia tuloksia jakaumaoletuksista riippumatta.

Fisher esitteli permutaatiotestin idean 1930-luvulla kahden riippumattoman otoksen t-testin parametrittomana vastineena. Testin idea kehittyi ja laajentui vuosien mittaan, mutta se yleistyi vasta viime aikoina tietokoneiden tehojen kehittymisen myötä, joten tarvittavat laskutoimitukset pystyttiin suorittamaan. Permutaatiotesteissä lähtökohtana on satunnainen otos, jonka perusteella muodostetaan testisuureen jakauma tämän otoksen permutaatioista. Sen takia, että permutaatiotesti toimii jakaumaoletuksista riippumatta, yleisenä menetelmänä sen tulokset ovat luotettavampia silloin, kun normaalijakautuneisuus ei pidä paikkansa. Permutaatiotesti on eksakti testi äärellisissä otoksissa ja se soveltuu pienille aineistoille.

3.1 Permutaatiotesti kanonisessa korrelaatioanalyysissä

Permutaatiotestiä voidaan soveltaa riippuvuuden merkitsevyyden testaamiseen silloin, kun normaalisuusoletus ei päde. Näin tutkittaessa kahden muuttujajoukon välistä yhteyttä permutaatiotesti on yksi mahdollinen lähestymistapa. Permutaatiotestin ajatuksena on verrata aineistosta laskettua testisuureen (13) arvoa siihen permutaatiojakaumaan, joka muodostetaan permutoimalla toisen muuttujajoukon havaintoja

rikkomatta muuttujien välistä riippuvuusrakennetta kyseisessä joukossa (Manly 1991, 218).

Olkoon meillä kaksi havaintomatriisia

$$\mathcal{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix} \quad \text{ja} \quad \mathcal{Y} = \begin{bmatrix} \mathbf{y}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{y}_n^{\mathrm{T}} \end{bmatrix},$$

joiden dimensiot ovat $n \times m$ ja $n \times p$. Olkoon $\pi_{(1)}, \pi_{(2)}, \ldots, \pi_{(n)}$ lukujen $1, 2, \ldots, n$ permutaatiota. Merkitään nyt

$$\mathcal{Y}_\pi = \begin{bmatrix} \mathbf{y}_{\pi_{(1)}}^{\mathrm{T}} \\ \vdots \\ \mathbf{y}_{\pi_{(n)}}^{\mathrm{T}} \end{bmatrix},$$

joka saadaan permutoimalla $\mathcal{Y}$:n rivien järjestystä.

Tutkittaessa joukkojen $\mathcal{X}$ ja $\mathcal{Y}$ välistä riippuvuutta, jolloin nollahypoteesi on, että $\mathcal{X}$ ja $\mathcal{Y}$ ovat riippumattomia, permutaatiotestin avulla voidaan testata tämän nollahypoteesin merkitsevyyttä. Kyseessä oleva nollahypoteesi tarkoittaa, että mikä tahansa matriisin $\mathcal{X}$ rivi i liittyy yhtä suurella todennäköisyydellä jokaiseen matriisin $\mathcal{Y}$ riviin j. Riittää, että permutoidaan vain toisen matriisin $\mathcal{X}$ tai $\mathcal{Y}$ rivejä (Manly 1991, 218). Tämä selittyy sillä, että rivien järjestyksellä ei ole merkitystä — jokaisen permutaation jälkeen toisen matriisin rivien järjestys muuttuu suhteessa toisen matriisin riveihin. Rivien erilaisia järjestyksiä on kaiken kaikkiaan $n!$. Silloin permutaatiojakauma saadaan permutoimalla toisen matriisin rivit toisen matriisin riveistä riippumatta, missä permutaatio on valittu siten, että nollahypoteesin vallitessa rivin jokaisella permutaatiolla on yhtä suuri todennäköisyys. Periaatteessa pitää käydä kaikki mahdolliset permutaatiot läpi, mutta permutaatioiden isosta määrästä johtuen otetaan vain satunnaisotos niistä. Ensin tehdään kanoninen korrelaatioanalyysi alkuperäisestä aineistosta

$$[\mathcal{X}, \mathcal{Y}] = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} & \mathbf{y}_1^{\mathrm{T}} \\ \vdots & \vdots \\ \mathbf{x}_n^{\mathrm{T}} & \mathbf{y}_n^{\mathrm{T}} \end{bmatrix}$$

ja lasketaan testisuure (13). Seuraavaksi tehdään analyysi ja samalla lasketaan testisuure (13) yhdistetystä matriisista

$$[\mathcal{X}, \mathcal{Y}_\pi] = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} & \mathbf{y}_{\pi_{(1)}}^{\mathrm{T}} \\ \vdots & \vdots \\ \mathbf{x}_n^{\mathrm{T}} & \mathbf{y}_{\pi_{(n)}}^{\mathrm{T}} \end{bmatrix}.$$

Permutoinnissa ryhmien välinen riippuvuusrakenne muuttuu, koska havaintojen järjestys toisessa matriisissa muuttuu. Sen jälkeen, kun kaikki permutaatiot on käyty läpi, saadaan testisuureelle permutaatiojakauma. Merkitsevyyden testaus tehdään permutaatioperiaatteella, toisin sanoen testisuureen alkuperäisen aineiston perusteella saatu arvo (havaittu arvo) vertaillaan permutoimalla muodostettuun testisuureen jakaumaan. Näin saadaan p-arvo.

3.2 Likimääräinen permutaatiotesti

Oletetaan edelleen, että meillä on kaksi matriisia $\mathcal{X}$ ja $\mathcal{Y}$, joiden dimensiot ovat $n \times m$ ja $n \times p$ vastaavasti. Ajatellaan nyt, että kanoninen korrelaatioanalyysi matriiseista $\mathcal{X}$ ja $\mathcal{Y}$ on tehty, toisin sanoen

$$\mathcal{U} = \mathcal{X}\mathbf{A}$$
$$\mathcal{V} = \mathcal{Y}\mathbf{B},$$

missä vektorit

$$\mathbf{u}_1 = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}}\mathbf{a}_1 \\ \mathbf{x}_2^{\mathrm{T}}\mathbf{a}_1 \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}}\mathbf{a}_1 \end{bmatrix}, \mathbf{u}_2 = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}}\mathbf{a}_2 \\ \mathbf{x}_2^{\mathrm{T}}\mathbf{a}_2 \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}}\mathbf{a}_2 \end{bmatrix}, \ldots, \mathbf{u}_m = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}}\mathbf{a}_m \\ \mathbf{x}_2^{\mathrm{T}}\mathbf{a}_m \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}}\mathbf{a}_m \end{bmatrix},$$

ja

$$\mathbf{v}_1 = \begin{bmatrix} \mathbf{y}_1^{\mathrm{T}}\mathbf{b}_1 \\ \mathbf{y}_2^{\mathrm{T}}\mathbf{b}_1 \\ \vdots \\ \mathbf{y}_n^{\mathrm{T}}\mathbf{b}_1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} \mathbf{y}_1^{\mathrm{T}}\mathbf{b}_2 \\ \mathbf{y}_2^{\mathrm{T}}\mathbf{b}_2 \\ \vdots \\ \mathbf{y}_n^{\mathrm{T}}\mathbf{b}_2 \end{bmatrix}, \ldots, \mathbf{v}_p = \begin{bmatrix} \mathbf{y}_1^{\mathrm{T}}\mathbf{b}_p \\ \mathbf{y}_2^{\mathrm{T}}\mathbf{b}_p \\ \vdots \\ \mathbf{y}_n^{\mathrm{T}}\mathbf{b}_p \end{bmatrix}.$$

muodostavat matriisien $\mathcal{U}$ ja $\mathcal{V}$ sarakkeet. Matriisien $\mathbf{A}$ ja $\mathbf{B}$ sarakkeet ovat vastaavasti $\mathbf{a}_1, \ldots, \mathbf{a}_m$ ja $\mathbf{b}_1, \ldots, \mathbf{b}_p$. Muistetaan, että $\mathbf{b}_{m+1}, \ldots, \mathbf{b}_p$ ovat nollaominaisarvoja vastaavat ominaisvektorit. Ositetaan matriisi $\mathcal{V}$ sarakkeittensa suhteen seuraavalla tavalla:

$$\mathcal{V} = [\mathbf{v}_1, \mathcal{V}_2],$$

missä $\mathcal{V}_2$ sisältää matriisin $\mathcal{V}$ kaikki sarakkeet toisesta lähtien.

Tarkastellaan nyt tilannetta, jossa oletetaan, että olisi ainakin yksi kanoninen pari, joka todella korreloi, toisin sanoen $\lambda_1 > 0$. Likimääräinen testi koskee sitä hypoteesia, että ensimmäistä korrelaatiota λ_1 lukuun ottamatta muut kanoniset korrelaatiot ovat nollia eli nollahypoteesi H_0 on $\lambda_2 = \lambda_3 = \cdots = \lambda_m = 0$. Korrelaatioiden merkitsevyyttä voidaan testata permutaatiotestin perusteella seuraavasti:

1. Alkuperäisestä aineistosta lasketaan testisuure

$$\hat{q} = -\left[n - \frac{1}{2}(m+p+1)\right] \sum_{j=2}^{m} \ln(1-\hat{\lambda}_j^2). \tag{15}$$

2. Permutoidaan matriisin $\mathcal{V}$ rivejä seuraavasti: ensimmäinen sarake eli $\mathbf{v}_1$ pysyy ennallaan ja matriisin $\mathcal{V}_2$ rivejä permutoidaan.

3. Lasketaan matriiseista $\mathcal{U}$ ja $\mathcal{V}_{\text{perm}} = [\mathbf{v}_1, \mathcal{V}_{2,\text{perm}}]$ testisuure kuten askeleessa 1.

4. Toistetaan askeleet 2 ja 3 K kertaa. Saadaan arvot $q_1, \ldots, q_{\text{K}}$, missä K on permutaatioiden lukumäärä.

5. Lasketaan p-arvo, joka on testisuureen $\hat{q}$ havaittua arvoa ylittävien osuus kaikista testisuureen saaduista arvoista eli

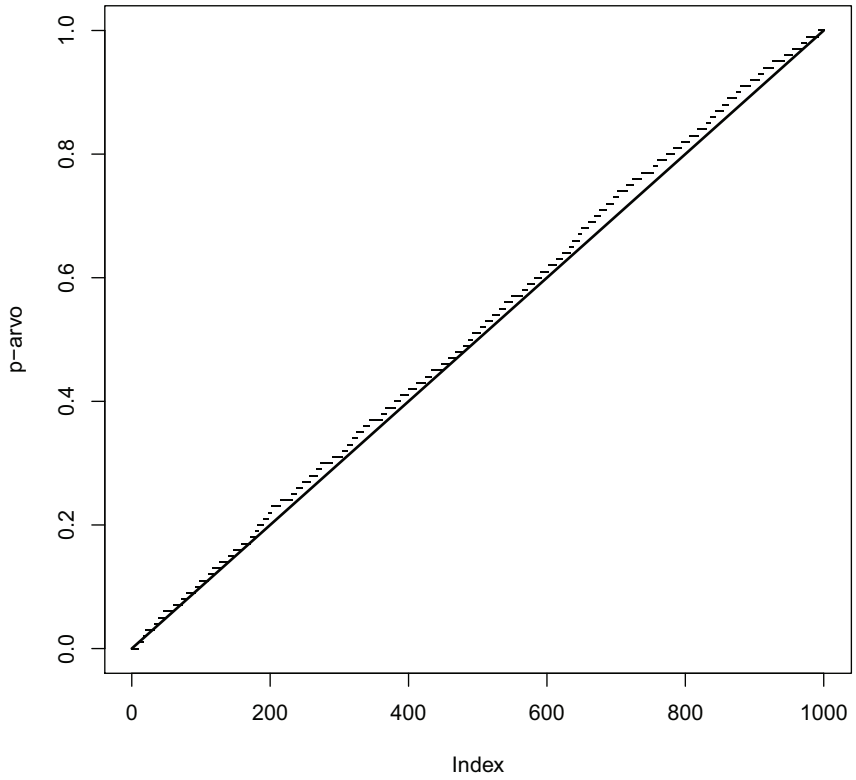
$$p = \frac{1}{\text{K}} \sum_{j=1}^{\text{K}} \text{I}(q_j > \hat{q}). \tag{16}$$

Testi olisi täsmälleen oikea permutaatiotesti, jossa ei ole virhettä, mikäli ei jouduttaisi estimoimaan kanonisia kertoimia. Tämän takia testi on likimääräinen.

Seuraavaksi tehdään simulointikoe likimääräiseen permutaatiotestiin liittyen. Tarkoituksena on tarkistaa edellä esitellyn menetelmän toimivuutta simulointikokeen avulla. Generoidaan kaksi matriisia $\mathcal{X}$ ja $\mathcal{Y}$ kokoa $n \times m$ seuraavasti:

- molempien matriisien ensimmäiset sarakkeet riippuvat toisistaan ja ovat peräisin $N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$-jakaumasta, missä $\rho > 0$, $\rho = \lambda_1$.
- Matriisien loput alkiot generoidaan riippumattomasti $N(0,1)$-jakaumasta.

Lasketaan generoidusta aineistosta testisuure (15) kuten askeleessa 1. Aineiston permutointi tapahtuu kohdan 2 vastaavalla tavalla. Permutoinnin jälkeen saadaan vastaavat arvot $q_1, \ldots, q_{\text{K}}$, joiden perusteella saadaan yksi p-arvo. Koko proseduuri toistetaan jokaiselle generoidulle aineistolle. Kunkin aineiston generointi tehdään tuhat kertaa (N = 1000) ja kullekin simuloidulle aineistolle tehdään sata permutaatiota (K = 100). Simulointikokeessa valitaan otoskooksi 50 havaintoa ($n = 50$). Näin ollen saadaan p-arvo jokaiselle toistolle erikseen. Simuloinnin tulokset on esitetty kuvassa 1, josta nähdään, että p-arvot ovat likipitäen tasajakautuneita välille $(0,1)$, mistä voidaan päätellä, että menettely toimii.



Kuva 1: Likimääräisellä permutaatiotestillä simuloimalla saatu p-arvon empiirinen jakauma. Tässä aineisto on generoitu 1000 kertaa ja kullekin aineistolle on tehty 100 permutaatiota. Jokaiselle generoidulle aineistolle p-arvo on laskettu permutaatiojakauman perusteella.

LUKU 4

Luottamusvälit kanonisten muuttujien kertoimille

Tässä luvussa esitetään algoritmi, jota käytetään bootstrap-luottamusvälien estimointiin kanonisten muuttujien kertoimille. Menetelmänä bootstrap soveltuu tunnusluvun otosjakauman approksimoimiseksi. Bootstrapin ideana on muodostaa tunnusluvun empiirinen jakauma yhden otoksen perusteella, jonka ajatellaan edustavan koko populaatiota. Alkuperäisestä otoksesta generoidaan uudet otokset, mutta toisin kun permutaatiotestissä, jossa alkiot poimitaan riippumattomasti ilman takaisinpanoa, bootstrapissa alkiot valitaan satunnaisesti palauttaen. Näin sama alkio voi esiintyä uudessa otoksessa useita kertoja. Kustakin bootstrap-otoksesta estimoidaan kiinnostuksen kohteena olevan parametrin arvo. Saadaan sen empiirinen jakauma, joka on luotettava arvio parametrin oikealle otantajakaumalle.

Tässä työssä lasketaan luottamusvälejä kanonisten muuttujien kertoimille eli ominaisvektoreille bootstrap-menetelmällä. Oletetaan nyt kaksi matriisia $\mathcal{X}$ ja $\mathcal{Y}$, joiden dimensiot ovat $n \times m$. Kanonisten korrelaatioiden ja vastaavien vektoreiden luottamusvälejä voi laskea seuraavalla algoritmilla:

1. Tehdään kanoninen korrelaatioanalyysi matriisista $[\mathcal{X}, \mathcal{Y}]$ ja otetaan talteen kanoniset korrelaatiot $\hat{\lambda}_i$ ja vastaavat kanoniset vektorit $\mathbf{a}_i$ ja $\mathbf{b}_i$, missä $i = 1, \ldots, m$.

2. Normeerataan vektorit $\mathbf{a}_i$ ja $\mathbf{b}_i$ ykkösen pituisiksi: vektorin kukin komponentti jaetaan vektorin pituudella.

3. Poimitaan n kappaletta rivejä palauttaen yhdistetystä matriisista $[\mathcal{X}, \mathcal{Y}]$.

4. Lasketaan ja otetaan talteen vastaavat kanoniset vektorit $\mathbf{a}_i^*$ ja $\mathbf{b}_i^*$, $\quad i = 1, \ldots, m$.

5. Normeerataan vektorit $\mathbf{a}_i^*$ ja $\mathbf{b}_i^*$ ykkösen pituisiksi. Valitaan vektoreiden $\mathbf{a}_i^*$ ja $\mathbf{b}_i^*$ etumerkki niin, että vektoreiden skalaaritulo

$$\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i^* > 0, \quad \text{ja} \quad \mathbf{b}_i^{\mathrm{T}}\mathbf{b}_i^* > 0,$$

missä $\mathbf{a}_i$ ja $\mathbf{b}_i$ ovat alkuperäisestä aineistosta laskettuja i:nnen kanonisen muuttujan kerroinvektoreita, ja $\mathbf{a}_i^*$ ja $\mathbf{b}_i^*$ ovat bootstrap-otoksesta laskettuja vastaavia vektoreita.

6. (a) Lasketaan kerroinvektoreiden $\mathbf{a}_i$ ja $\mathbf{b}_i$ koordinaattikohtaiset luottamusvälit prosenttipistemenetelmällä.

   (b) Estimoidaan luottamusalue kulman avulla kerroinvektorille $\mathbf{a}_i$ ja $\mathbf{b}_i$.

4.1 Prosenttipistemenetelmä

Menetelmä perustuu tarkasteltavan parametrin bootstrap-jakaumaan. Merkitään $\hat{\theta}$:lla parametrin θ estimaattia. Olkoon $\hat{\theta}$ kerroinvektorin $\mathbf{a}_i$ yksi koordinaatti. Muodostetaan B kappaletta riippumattomia bootstrap-otoksia ja lasketaan niistä arvot $\hat{\theta}_j^*, j = 1, \ldots, \mathrm{B}$. Järjestetään bootstrap-arvot suuruusjärjestykseen $\hat{\theta}_{(1)}^* \leq \hat{\theta}_{(2)}^* \leq \cdots \leq \hat{\theta}_{(\mathrm{B})}^*$. Silloin $100(1-2\alpha)$-prosentin luottamusvälin ala- ja yläraja ovat järjestetyn aineiston $\mathrm{B}\alpha$:s ja $(\mathrm{B}(1-\alpha)+1)$:s arvot. (Efron ja Tibshirani 1993).

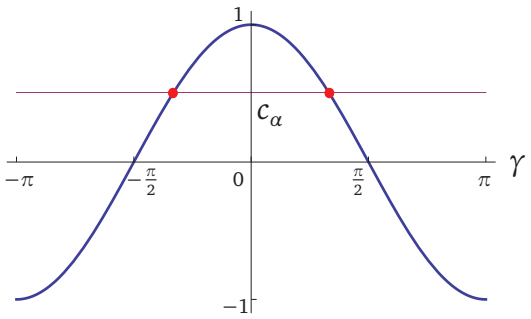
4.2 Luottamusalue kulman avulla

Oletetaan, että $\boldsymbol{\alpha}_i$ on ominaisarvoon λ_i liittyvä $m \times 1$ populaation kerroinvektori, jonka pituus on 1. Vastaavasti otoksesta lasketut estimaatit ovat $\mathbf{a}_i$ ja $\hat{\lambda}_i$. Koska kerroinvektoreiden otosjakaumia on vaikea johtaa, niin luottamusaluetta

$$\left\{\boldsymbol{\alpha}_i : \mathbf{a}_i^{\mathrm{T}}\boldsymbol{\alpha}_i \geq c_\alpha\right\}, \quad \mathrm{P}(\mathbf{a}_i^{\mathrm{T}}\boldsymbol{\alpha}_i \geq c_\alpha) = \alpha,$$

ei voi käytännössä laskea (Beran ja Srivastava 1985). Tässä c_α voidaan laskea bootstrap-jakaumasta.

Kanonisen korrelaatioanalyysin tuloksena saadulle kerroinvektorille $\mathbf{a}_i$ voidaan muodostaa estimoinnin tarkkuutta kuvaava luottamusalue, joka pyritään estimoimaan kulman avulla. Toisin kuin edellinen menetelmä, jolla lasketaan luottamusvälejä kanonisten muuttujien painokertoimille (vektoreiden koordinaateille), kyseisellä menetelmällä estimoidaan luottamusalue kanonisille muuttujille eli kerroinvektoreille. Se



Kuva 2: Kosini-funktion kuvaaja välillä $[-\pi, \pi]$. Havaitaan, että $\cos(\gamma) \geq c_\alpha$ kulman γ sellaisilla arvoilla, jotka sijoittuvat kahden punaisen pisteen välille.

on käyttökelpoinen menettely siinä mielessä, että saadaan kokonainen kuva vektorin estimoinnista eikä pelkästään vektorin koordinaateista.

Menetelmän ideana on laskea luottamusväli vektoreiden $\mathbf{a}_i$ ja $\mathbf{a}^*_{i,j}$ avulla. Tässä vektorit $\mathbf{a}_i$ ja $\mathbf{a}^*_{i,j}$ ovat alkuperäisestä aineistosta ja bootstrapilla lasketut kerroinvektorit. Alaindeksi j viittaa bootstrap-otokseen, joita on muodostettu B kappaletta. Olkoon näiden kahden vektorin välinen kulma $\gamma^*_{i,j}$ ja se on sellainen, että

$$c^*_{i,j} = \cos(\gamma^*_{i,j}) = \mathbf{a}_i^{\mathrm{T}} \mathbf{a}^*_{i,j}, \quad j = 1, \ldots, \mathrm{B},$$

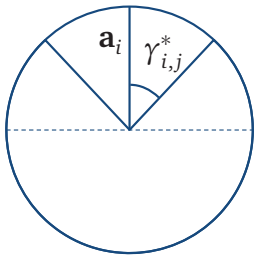
ja vektori $\mathbf{a}^*_{i,j}$ on valittu niin, että $\mathbf{a}_i^{\mathrm{T}} \mathbf{a}^*_{i,j} > 0$, mikä tarkoittaa, että kulman $\gamma^*_{i,j} = \arccos(c^*_{i,j})$ suuruus rajoittuu välille $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Tämä voidaan havainnollistaa kuvan 2 avulla. Kun arccos määritellään cos:n käänteisfunktioksi välillä $[0, \pi]$ ja koska $\cos(\gamma^*_{i,j}) \geq c_\alpha$, missä c_α on pystyakselilla oleva kriittinen piste, niin kulman $\gamma^*_{i,j}$ pitää olla välillä $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

Järjestetään $c^*_{i,j}$:t suuruusjärjestykseen $c^*_{i,(1)} < c^*_{i,(2)} < \cdots < c^*_{i,(\mathrm{B})}$. Siten $100(1 - 2\alpha)$-prosentin luottamusväli vektoreiden $\mathbf{a}_i$ ja $\mathbf{a}^*_{i,j}$ väliselle kulmalle on

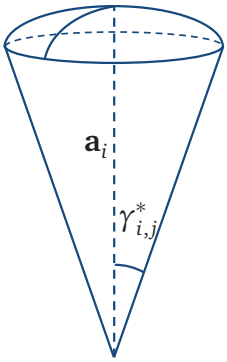
$$(-\arccos(c^*_{i,(\mathrm{B}\alpha)}), \arccos(c^*_{i,(\mathrm{B}\alpha)})) \qquad \text{kaikilla } i \text{ ja } j = 1, \ldots, \mathrm{B}.$$

Näin kulman suuruus sijoittuu välille $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Koska $\gamma^*_{i,j} = \arccos(c^*_{i,j})$, välin suuruudeksi saadaan $2\gamma^*_{i,(\mathrm{B}\alpha)}$.

Seuraavaksi selvitetään, minkä muotoinen luottamusalue on kaksiulotteisen sekä kolmiulotteisen avaruuden tapauksessa. Kaksiulotteisen vektorin $\mathbf{a}_i$ tilanteessa ($m = 2$) luottamusalue on rajoitettu kulmalla, jonka suuruus on $2\gamma^*_{i,j}$. Luottamusalue voidaan esittää kaaren segmentillä. Sen graafinen esitys on kuvassa 3. Kerroinvektorin $\mathbf{a}_i$ alkupää on ympyrän keskuskulmassa ja vektorin kärki osuu yksikköympyrälle. Kulman muodostama kaaren segmentti on luottamusalue kerroinvektorille $\mathbf{a}_i$. On mahdollista käyttää myös sektorin kaaren pituutta estimoinnin tarkkuuden kuvaamiseksi.



Kuva 3: Luottamusalue kaksiulotteisessa tilanteessa – kaaren segmentti. Kuvaan on merkitty aineistosta estimoitu vektori $\mathbf{a}_i$ ja kulma $\gamma^*_{i,j}$ määrittää luottamusalueen rajat. Tässä pystysuunnassa suunnattu napa-akseli on valittu siten, että se yhtyy vektoriin $\mathbf{a}_i$.



Kuva 4: Luottamusalue kolmiulotteisessa tilanteessa – kalotti. Kuvaan on merkitty aineistosta estimoitu vektori $\mathbf{a}_i$ ja $\gamma^*_{i,j}$ on kulma, joka määrittää kartion.

Kolmiulotteisessa tapauksessa ($m = 3$) luottamusalue kerroinvektorille on pinta, joka on muodostettu kartion ja yksikköpallopinnan leikkauksella, toisin sanoen luottamusalue on yksikköpallonkalotti. Luottamusalue on rajoitettu kartiolla (Kuva 4) niin, että vektori $\mathbf{a}_i$ määrittää kartion akselin, ja muodostajasuoran ja kartion akselin muodostama kulma on $\gamma^*_{i,j}$. Kuten kaksiulotteisessa tapauksessa estimoinnin tarkkuutta voidaan arvioida pallokalotin pinta-alan avulla.

LUKU 5

Sovelluksia

Tässä luvussa sovitetaan permutaatiotesti tutkimusaineistoon ja vertaillaan saatuja tuloksia Bartlettin testin tuloksiin. Tutkitaan, onko terveys- ja liikunta-aktiivisuusmuuttujien välillä riippuvuutta.

5.1 Tutkimusaineisto

Tässä työssä käytetty aineisto on peräisin Jyväskylän yliopiston liikuntabiologian laitokselta. Aineisto koostuu terveyskunnon ja liikunta-aktiivisuuden mittauksista. Osallistuminen tutkimukseen oli vapaaehtoista ja osallistujia haettiin mainonnan avulla (Tikkanen et al. 2013). Halukkaita osallistumaan tutkimukseen oli kaikkiaan 245, josta noin puolet täytti vaaditut terveyskriteerit ja heille on tehty mittaukset (Tikkanen et al. 2013). Lopullinen aineisto sisältää tiedot 84 osallistuvalta, josta on 44 naista ja 40 miestä ja joiden ikä vaihtelee $20-76$ vuoden välillä: $20-29$ -vuotiaat ($n=27$), $30-59$ -vuotiaat ($n=40$), $60-76$ -vuotiaat ($n=17$). Tutkimuksessa mitattiin terveiden henkilöiden lihasaktiivisuutta ja epäaktiivisuutta. Osa mittauksista suoritettiin laboratorio-olosuhteissa ja osa päivittäisten toimintojen yhteydessä. Kaikkiaan aineistossa on 16 muuttujaa, joista 6 terveyttä, 3 liikunta-aktiivisuutta ja epäaktiivisuutta kuvaavia muuttujia ja 7 taustamuuttujaa. Tutkielmassa tutkitaan terveys- ja liikunta-aktiivisuusmuuttujien välisiä riippuvuussuhteita. Tutkimuksen kannalta kiinnostavat terveysmuuttujat ovat ensisijaisesti verenpaine, sisäelinten ympärillä oleva rasva ja seerumin triglyseridipitoisuus. Triglyseridit ovat tärkeitä veren rasvoja, joiden lisääntyminen suurentaa sepelvaltimotaudin riskin. Liikunta-aktiivisuutta mittaavat muuttujat ovat kyselylomakkeella arvioitu liikunta-aktiivisuus, reisilihasten kes-

Taulukko 1: Estimoidut kanoniset korrelaatiot.

	Naiset			Miehet		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
Ikää ei otettu huomioon	0.637	0.408	0.023	0.610	0.122	0.014
Ikä vakioitu	0.469	0.210	0.014	0.280	0.131	0.011

kimääräinen päivittäinen aktiivisuus, joka on määritelty prosenttiosuutena mitatusta reisilihasten isometrisestä maksimista (vaihteluväli $0.9 - 16.9\%$ isometrisestä maksimista) ja reisilihasten päivän pisin epäaktiivisuusaika (vaihteluväli $2.5 - 38.3$ minuuttia). Taustamuuttujiksi valitaan sukupuoli ja ikä. Muuttujat seerumin triglyseridipitoisuus ja reisilihasten keskimääräinen päivittäinen aktiivisuus ja päivän pisin epäaktiivisuusaika on logaritmoitu analyysiä varten.

5.2 Tulokset

Aineisto on jaettu sukupuolen mukaan kahteen ryhmään. On tehty kaksi analyysiä, joista toisessa iästä johtuvaa vaihtelua ei ole otettu huomioon ja toisessa iän vaikutusta on eliminoitu sovittamalla lineaarinen regressiomalli. Jokainen terveys- ja liikunta-aktiivisuusmuuttuja on regressoitu iän suhteen. Kanoninen analyysi on tehty regressioanalyysin jäännöksistä.

5.2.1 Permutaatiotesti

Tässä luvussa tutkitaan, onko terveys ja liikunta-aktiivisuus muuttujaryhmien välillä riippuvuutta. Käytetään asian tutkimisessa luvussa 2.4 esitettyä Bartlettin testiä ja luvussa 3 esitettyä permutaatiotestiä. Kanonisen analyysin tuloksena saadaan kanoniset korrelaatiot, jotka on esitetty taulukossa 1. On huomattava, että kun ikä on otettu huomioon, ensimmäinen ja suurin kanoninen korrelaatio on alle 0.5 ja loput korrelaatiot ovat varsin pieniä. Taulukosta nähdään, että terveys- ja liikunta-aktiivisuusmuuttujien välinen korrelaatio $\hat{\lambda}_1$ sekä miehillä että naisilla on suunnilleen samaa luokkaa, kun ikä ei ole otettu huomioon. Kun iän vaikutusta on eliminoitu, ensimmäinen korrelaatio miesten osalta on selvästi naisia alhaisempi.

Seuraavaksi testataan kanonisten korrelaatiokertoimien merkitsevyyttä. Tällöin nollahypoteesi väittää ettei terveys- ja liikunta-aktiivisuusmuuttujien välillä ole riippuvuutta, toisin sanoen $\lambda_1 = \lambda_2 = \lambda_3 = 0$. Taulukossa 2 ovat permutaatio- ja Bartlet-

Taulukko 2: Bartlettin testillä ja permutaatiotestillä saadut p-arvot. Nollahypoteesi $\lambda_1 = \lambda_2 = \lambda_3 = 0$.

	Naiset		Miehet	
	Bartlett	Perm. testi	Bartlett	Perm. testi
Ikää ei otettu huomioon	0.001	0.002	0.048	0.052
Ikä vakioitu	0.237	0.231	0.940	0.946

tin testin tulokset. Kun niitä verrataan keskenään, huomataan että permutaatiotestillä saadut tulokset ovat samansuuntaisia Bartlettin testin kanssa. Sukupuolittain tarkasteltuna sekä naisilla että miehillä terveys ja liikunta-aktiivisuus ryhmien välillä vallitsee riippuvuus (p-arvo≤ 0.05). Näin ainakin ensimmäinen korrelaatio on merkitsevä. On kuitenkin huomattava, että tapauksessa, jolloin iän vaikutus on eliminoitu, muuttujien välillä ei enää esiinny merkitsevää riippuvuutta.

Taulukossa 3 esitetään permutaatio- ja Bartlettin testin p-arvot, kun testataan, ovatko loput kanonisista korrelaatioista λ_2 ja λ_3 nollasta eroavia. Tulosten perusteella voidaan todeta, että edellä mainitut kanoniset korrelaatiot eivät ole merkitseviä sen enempää naisilla kuin miehilläkään. Näin ollen, kyseessä on yksi tilastollisesti merkitsevä kanoninen korrelaatio.

5.2.2 Bootstrap-luottamusvälit

Kanonisten muuttujien kertoimet ovat kanonisten korrelaatioiden lisäksi kanonisen analyysin keskeisimmät tulokset. Seuraavaksi lasketaan kerroinvektoreiden koordinaattikohtaiset luottamusvälit kappaleessa 4.1 esitetyllä prosenttipistemenetelmällä.

Taulukko 3: Bartlettin testillä ja permutaatiotestillä saadut p-arvot. Nollahypoteesi $\lambda_2 = \lambda_3 = 0$.

	Naiset		Miehet	
	Bartlett	Perm. testi	Bartlett	Perm. testi
Ikää ei otettu huomioon	0.126	0.112	0.970	0.966
Ikä vakioitu	0.533	0.537	0.906	0.901

Taulukko 4: Muuttujien väliset korrelaatiot naisten aineistossa.

	Verenpaine	Sisäelint. rasva	Triglyseridi	Liikunta-aktiiv.	Päivit. aktiiv.	Pisin epäaktiiv. aika
Verenpaine	1	0.69	0.63	0.22	0.50	0.24
Sisäelinten rasva	0.69	1	0.60	0.08	0.58	−0.03
Triglyseridi	0.63	0.60	1	0.14	0.38	0.16
Liikunta-aktiivisuus	0.22	0.08	0.14	1	0.10	−0.04
Päivittäinen aktiivisuus	0.50	0.58	0.38	0.10	1	−0.15
Pisin epäaktiivisuusaika	0.24	−0.03	0.16	−0.04	−0.15	1

Taulukko 5: Muuttujien väliset korrelaatiot miesten aineistossa.

	Verenpaine	Sisäelin. rasva	Triglyseridi	Liikunta-aktiiv.	Päivitt. aktiiv.	Pisin epäaktiiv. aika
Verenpaine	1	0.30	0.46	−0.03	0.47	0.24
Sisäelinten rasva	0.30	1	0.50	−0.14	0.41	0.16
Triglyseridi	0.46	0.50	1	−0.02	0.30	0.17
Liikunta-aktiivisuus	−0.03	−0.14	−0.02	1	0.03	0.08
Päivittäinen aktiivisuus	0.47	0.41	0.30	0.03	1	0.05
Pisin epäaktiivisuusaika	0.24	0.16	0.17	0.08	0.05	1

Tarkastellaan ensin kaikkien muuttujien parittaiset korrelaatiot. Korrelaatiot on esitetty naisten osalta taulukossa 4 ja miesten osalta taulukossa 5. Havaitaan, että kaikki terveysmuuttujat korreloivat keskenään positiivisesti naisilla ja miehillä ja että miesten aineistossa kaikki liikunta-aktiivisuusmuuttujien väliset korrelaatiot ovat myös positiivisia. Nähdään, että muuttujien parittaiset korrelaatiot terveys- ja liikunta-aktiivisuusryhmien sisällä sekä ryhmien välillä ovat selvästi korkeammat naisilla kuin miehillä. Naisten ryhmässä terveysmuuttujat korreloivat voimakkaasti keskenään. Miehillä vastaavien muuttujien väliset korrelaatiot ovat kohtalaisia. Naisten aineistossa verenpaine korreloi vahvasti kaikkien terveysmuuttujien kanssa, sen sijaan miesten aineistossa muuttuja triglyseridi korreloi terveysmuuttujien kanssa eniten. Liikunta-aktiivisuusmuuttujat korreloivat keskenään heikosti (korrelaatiot ovat itseisarvoltaan alle 0.15) miehillä ja naisilla. Naisten ryhmässä muuttuja pisin epäaktivisuusaika korreloi negatiivisesti muuttujien liikunta-aktiivisuus ja päivittäinen aktiivisuus kanssa.

Kun tarkastellaan terveys- ja liikunta-aktiivisuusryhmien välisiä korrelaatioita, huomataan, että saadut korrelaatiot ovat epäjohdonmukaisia siinä mielessä, että korrelaatioiden etumerkit ovat erilaiset miehillä ja naisilla, mikä vaikeuttaa tulkintaa. Naisilla kaikki korrelaatiot ovat positiivisia muuttujien pisin epäaktiivisuusaika ja sisäelinten ympärillä oleva rasva välistä korrelaatiota lukuun ottamatta. Miehillä muuttuja liikunta-aktiivisuus korreloi negatiivisesti kaikkien terveysmuuttujien kanssa, mutta samalla muuttuja päivittäinen aktiivisuus on positiivisessa yhteydessä vastaavien terveysmuuttujien kanssa.

Koska ainoastaan ensimmäinen kanoninen korrelaatio on merkitsevä, vain ensimmäinen kanoninen muuttujapari on kiinnostava. Taulukoissa 6-9 ovat painokertoimien estimaatit ja niiden luottamusvälit ensimmäiselle kanoniselle parille naisten ja miesten tapauksessa. Estimaatit ja niiden luottamusvälit on saatu kovarianssimatriisista, joten se on otettava huomioon tulkinnoissa. Taulukoissa on myös esitetty standardoidut estimaatit, jotta pystytään vertailemaan saatuja painoja keskenään. Standardointi on tehty niin, että muuttujien keskiarvo on 0 ja varianssi on 1. Standardoituja painoja käytetään ainoastaan painokertoimien vertailussa. Ensin keskitytään tarkastelemaan kertoimien standardoituja estimaatteja ja sen jälkeen siirrytään painokertoimien standardoimattomiin estimaatteihin ja niiden luottamusväleihin.

Taulukoissa 6 ja 7 painokertoimien estimaatit on laskettu ottamatta iän vaikutusta huomioon. Taulukosta 6 nähdään, että terveysmuuttujista verenpaine saa suurimman painon (0.870), kun taas muuttujan triglyseridi paino on pienin, 0.074. Liikunta-aktiivisuutta kuvaavista muuttujista päivittäinen aktiivisuus saa suurimman

Taulukko 6: Naisten ensimmäisen kanonisen parin kertoimien estimaatit 95% luottamusvälineineen, kun ikää ei ole otettu huomioon.

	Muuttujat	Standardoitu estimaatti	Estimaatti ja sen luottamusväli		
			Estimaatti	$\alpha=0.025$	$\alpha=0.975$
Terveysmuuttujat	Verenpaine	0.870	0.307	−0.510	0.613
	Sisäelinten rasva	0.487	0.094	−0.234	0.362
	Triglyseridi	0.074	0.947	0.499	1.0
Liikunta-aktiivisuusmuuttujat	Liikunta-aktiivisuus	0.209	0.023	−0.046	0.099
	Päivittäinen aktiivisuus	0.904	0.956	0.304	0.999
	Pisin epäaktiivisuusaika	0.374	0.291	−0.521	0.945

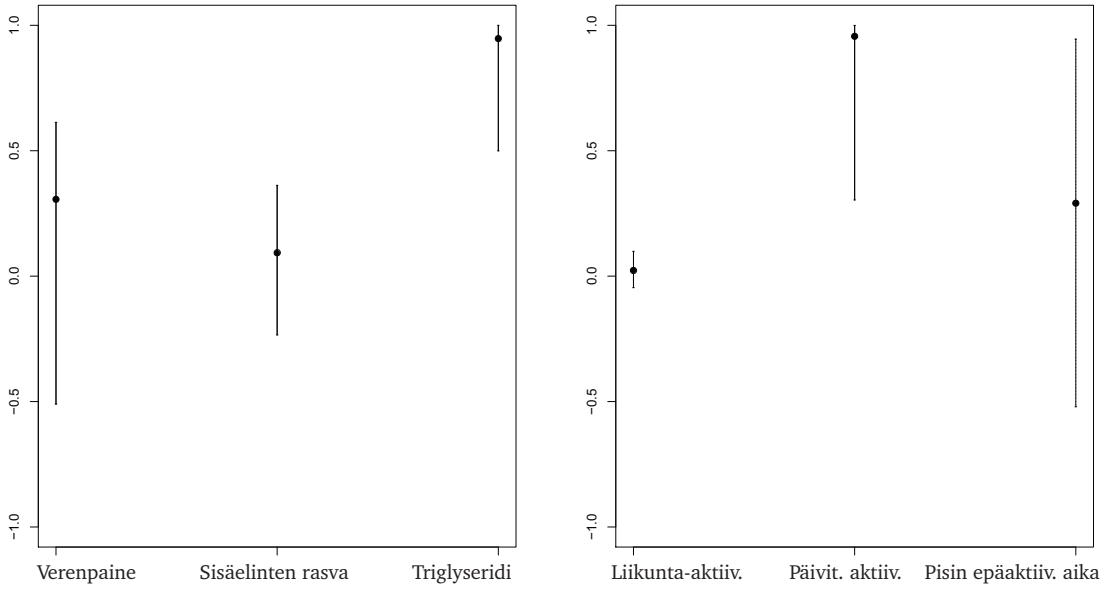
Taulukko 7: Miesten ensimmäisen kanonisen parin kertoimien estimaatit 95% luottamusvälineineen, kun ikää ei ole otettu huomioon.

	Muuttujat	Standardoitu estimaatti	Estimaatti ja sen luottamusväli		
			Estimaatti	$\alpha=0.025$	$\alpha=0.975$
Terveysmuuttujat	Verenpaine	0.774	0.310	−0.524	0.773
	Sisäelinten rasva	0.628	0.132	−0.242	0.373
	Triglyseridi	−0.079	−0.941	−0.999	−0.390
Liikunta-aktiivisuusmuuttujat	Liikunta-aktiivisuus	−0.227	−0.032	−0.191	0.077
	Päivittäinen aktiivisuus	0.895	0.943	0.651	1.0
	Pisin epäaktiivisuusaika	0.384	0.331	−0.130	0.756

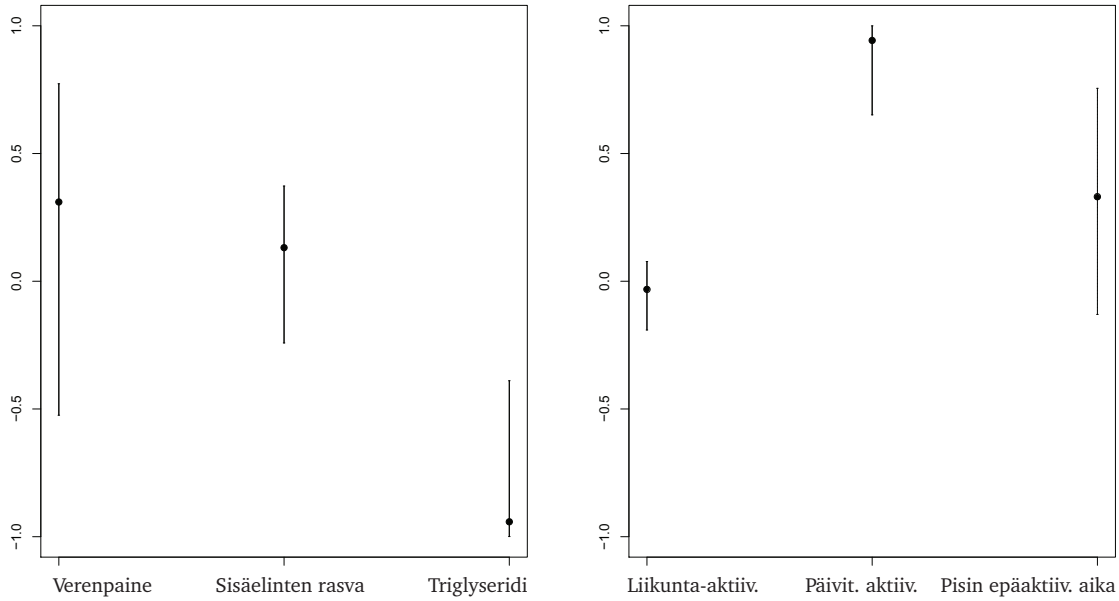
painokertoimen (0.904), pienin painokerroin on muuttujalla kyselylomakkeella arvioitu liikunta-aktiivisuus (0.209). Huomataan, että eniten painottuvien muuttujien, eli muuttujien verenpaine ja päivittäinen aktiivisuus parittainen korrelaatio on toiseksi suurin (taulukko 4). On vaikeaa löytää tulkinta sille, miksi verenpaineen ja päivittäisen aktiivisuuden välinen yhteys on positiivinen. Kuten huomataan, kaikki muuttujat saavat positiiviset kertoimet. Muistetaan, että ensimmäisten kanonisten muuttujien välinen korrelaatiokerroin on 0.637.

Miesten ensimmäisten kanonisten muuttujien painokertoimet ja niiden luottamusvälit on esitetty taulukossa 7. Korrelaatiomatriisista saatuja estimaatteja tarkastelemalla voidaan sanoa, että myös miesten tapauksessa muuttujat verenpaine ja päivittäinen aktiivisuus painottuvat voimakkaimmin. Vastaavat painokertoimet ovat 0.774 ja 0.895. Tarkasteltaessa kyseisten muuttujien välistä korrelaatiota (taulukko 5) havaitaan, että se on suurin (0.47). Kanonisen muuttujaparin välinen korrelaatio on 0.610. Liikunta-aktiivisuusmuuttujien painokertoimien estimaatteja vertailtaessa huomataan, että estimaatit itsearvoltaan ovat suunnilleen samat miehille ja naisille. Tässä miesten tulokset eroavat naisten tuloksista siinä, että muuttujien triglyseridi ja kyselylomakkeella arvioitu liikunta-aktiivisuus painokertoimet ovat negatiivisia. Saatujen tulosten perusteella voidaan kuitenkin sanoa, että miehillä ja naisilla muuttujat verenpaine ja päivittäinen aktiivisuus dominoivat. Näin verenpaine on yhteydessä päivittäiseen aktiivisuuteen.

Kanonisten muuttujien kertoimien estimaattien 95%:n luottamusvälit on esitetty graafisesti kuvissa 5 ja 6. Tässä iästä johtuvaa vaihtelua ei oteta huomioon. Muis-



Kuva 5: Kanonisten muuttujien kertoimien 95%:n luottamusvälit naisten aineistossa, kun ikää ei ole otettu huomioon.



Kuva 6: Kanonisten muuttujien kertoimien 95%:n luottamusvälit miesten aineistossa, kun ikää ei ole otettu huomioon.

tetaan, että kuvissa esitetyt estimaatit ja luottamusvälit on laskettu kovarianssimatriisista. Miehillä terveyttä kuvaavan kanonisen muuttujan kertoimien luottamusvälit ovat leveämpiä verrattuna naisiin. Nähdään, että estimaattien luottamusvälit ovat varsin leveitä verenpainetta mittaavan muuttujan osalta niin miehillä kuin naisillakin. Liikunta-aktiivisuusryhmään kuuluvista muuttujista muuttujien pisin epäaktiivisuusaika naisilla ja kyselylomakkeella arvioitu liikunta-aktiivisuus miehillä luottamusvälit ovat erittäin leveitä. Nämä luottamusvälit sisältävät nollan. Lisäksi luottamusväli estimaatille sisäelinten ympärillä oleva rasva sisältää myös nollan. Kaiken kaikkiaan voidaan sanoa, että estimaatit ovat epätarkkoja.

Vaikka kanonisessa analyysissä, jossa iän vaikutus vakioidaan ja joka on tehty jäännöksistä, jotka on saatu regressoimalla terveys- ja liikunta-aktiivisuusmuuttujat iän suhteen, ensimmäisten kanonisten muuttujien muodostaman parin välinen korrelaatio ei ole merkitsevä, jatketaan tarkastelua ja tulokset esitellään vertailua varten. Taulukoista 8 ja 9 nähdään naisten ja miesten ensimmäisen kanonisen muuttujaparin painokertoimien estimaatit, kun analyysissä iän vaikutus otetaan huomioon. Naisilla terveyttä kuvaavassa kanonisessa muuttujassa painottuu eniten muuttuja sisäelinten ympärillä oleva rasva kertoimella -0.803. Muuttujien kyselylomakkeella arvioitu liikunta-aktiivisuus, päivittäinen aktiivisuus ja pisin epäaktiivisuusaika muodostamassa kombinaatiossa epäaktiivisuusaika saa suurimman painon (0.868). Myös muuttuja kyselylomakkeella arvioitu aktiivisuus painottuu positiivisella painokertoimella 0.404. Tulosten mukaan muuttujien sisäelinten ympärillä oleva rasva ja pisin epäaktiivisuusaika välillä on negatiivinen yhteys, jolle on hankala löytää järkevää tulkintaa.

Taulukko 8: Naisten ensimmäisen kanonisen parin kertoimien estimaatit 95% luottamusvälineineen, kun ikä on vakioitu.

	Muuttujat	Standardoitu estimaatti	Estimaatti ja sen luottamusväli		
			Estimaatti	$\alpha=0.025$	$\alpha=0.975$
Terveysmuuttujat	Verenpaine	0.578	0.108	-0.598	0.744
	Sisäelinten rasva	-0.803	-0.082	-0.266	0.310
	Triglyseridi	0.146	0.991	0.295	1.0
Liikunta-aktiivisuusmuuttujat	Liikunta-aktiivisuus	0.404	0.059	-0.037	0.103
	Päivittäinen aktiivisuus	0.290	0.413	-0.570	1.0
	Pisin epäaktiivisuusaika	0.868	0.909	-0.246	0.918

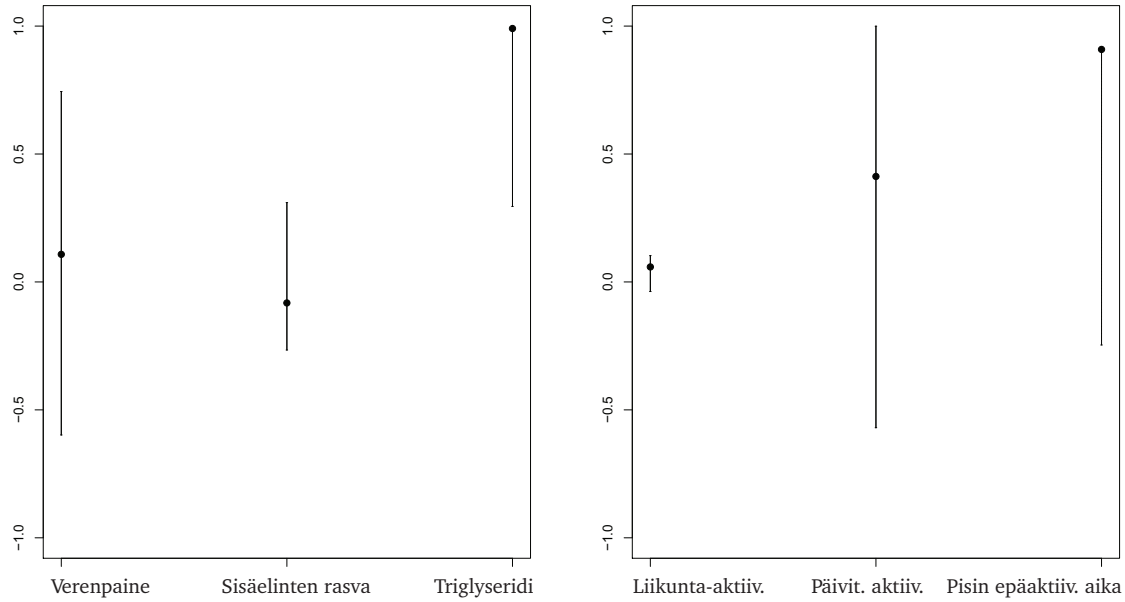
Taulukko 9: Miesten ensimmäisen kanonisen parin kertoimien estimaatit 95% luottamusvälineineen, kun ikä on vakioitu.

	Muuttujat	Standardoitu estimaatti	Estimaatti ja sen luottamusväli		
			Estimaatti	$\alpha=0.025$	$\alpha=0.975$
Terveysmuuttujat	Verenpaine	−0.767	−0.079	−0.733	0.688
	Sisäelinten rasva	−0.552	−0.030	−0.309	0.325
	Triglyseridi	0.324	0.996	0.336	1.0
Liikunta-aktiivisuusmuuttujat	Liikunta-aktiivisuus	0.325	0.047	−0.073	0.184
	Päivittäinen aktiivisuus	−0.910	−0.973	−1.0	−0.664
	Pisin epäaktiivisuusaika	−0.259	−0.227	−0.743	0.134

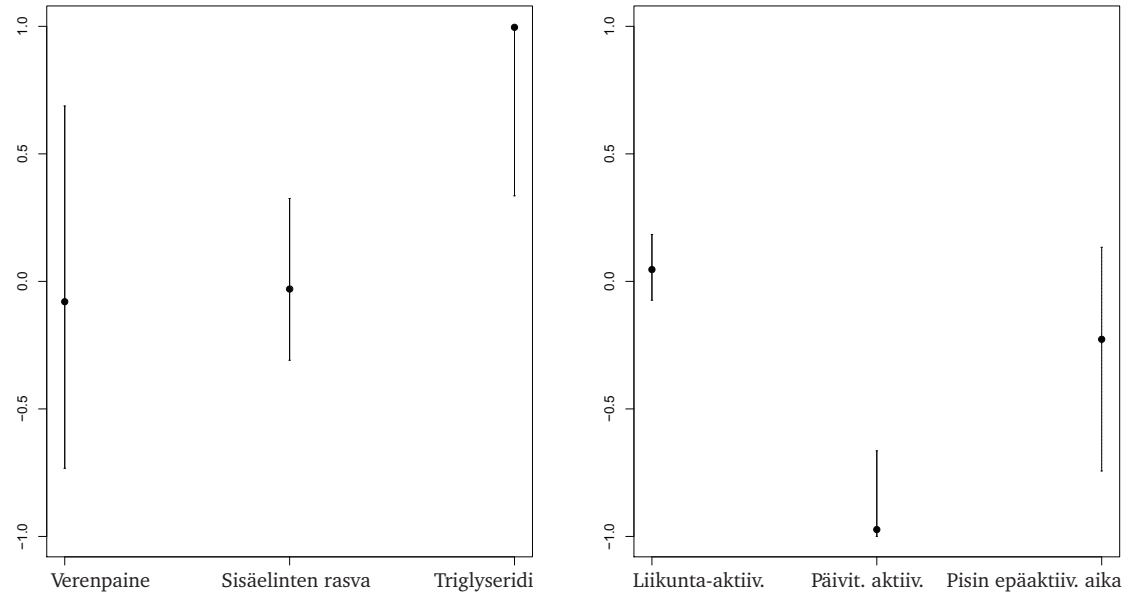
Korrelaatio ensimmäisten kanonisten muuttujien välillä on suuruudeltaan 0.469 ja se ei osoittaudu tilastollisesti merkitseväksi. Taulukoissa 6 ja 8 painokertoimien estimaatteja vertailtaessa todetaan, että kun iän vaikutus otetaan huomoon, kanoninen korrelaatioanalyysi johtaa erilaisiin tuloksiin.

Miesten ensimmäisten kanonisten muuttujien painokertoimien estimaatit on esitetty taulukossa 9. Nähdään, että muuttujat verenpaine ja päivittäinen aktiivisuus edelleen painottuvat eniten ensimmäisessä kanonisessa muuttujaparissa. Erityisesti liikunta-aktiivisuusmuuttujien painot etumerkkiä lukuun ottamatta eivät paljon muutu edellisestä analyysistä, jossa ikää ei ole otettu huomioon. Muuttujien verenpaine ja päivittäinen aktiivisuus samansuuntainen eli positiivinen yhteys säilyy. Miesten tapauksessa korrelaatio kanonisten muuttujien välillä on heikko (0.280) ja se ei ole tilastollisesti merkitsevä.

Estimaattien 95% luottamusvälit naisille ja miehille on esitetty kuvissa 7 ja 8. Kuvista nähdään, että kertoimien estimaattien luottamusväleissä on havaittavissa pieniä eroja sukupuolten välillä. Erityisesti terveysmuuttujien painokertoimien estimaattien luottamusvälit ovat leveydeltään hyvin lähellä toisiaan. Naisilla liikunta-aktiivisuusmuuttujien osalta estimaattien kaikki luottamusvälit sisältävät nollan. Muuttujan päivittäinen aktiivisuus luottamusväli on erittäin leveä. Naisten kertoimien luottamusvälit, kun ikä otetaan huomioon, ovat selvästi leveämmät verrattuna luottamusväleihin, joissa iästä johtuvaa vaihtelua ei ole otettu huomioon. Näyttää siltä, että iästä johtuvan vaihtelun huomioon ottaminen ei miesten osalta vaikuta saatuihin luottamusväleihin. Päädytään suunnilleen samoihin tuloksiin kuin aiemmin. Tulokset eroavat



Kuva 7: Kanonisten muuttujien kertoimien 95%:n luottamusvälit naisten aineistossa, kun ikä on vakioitu.



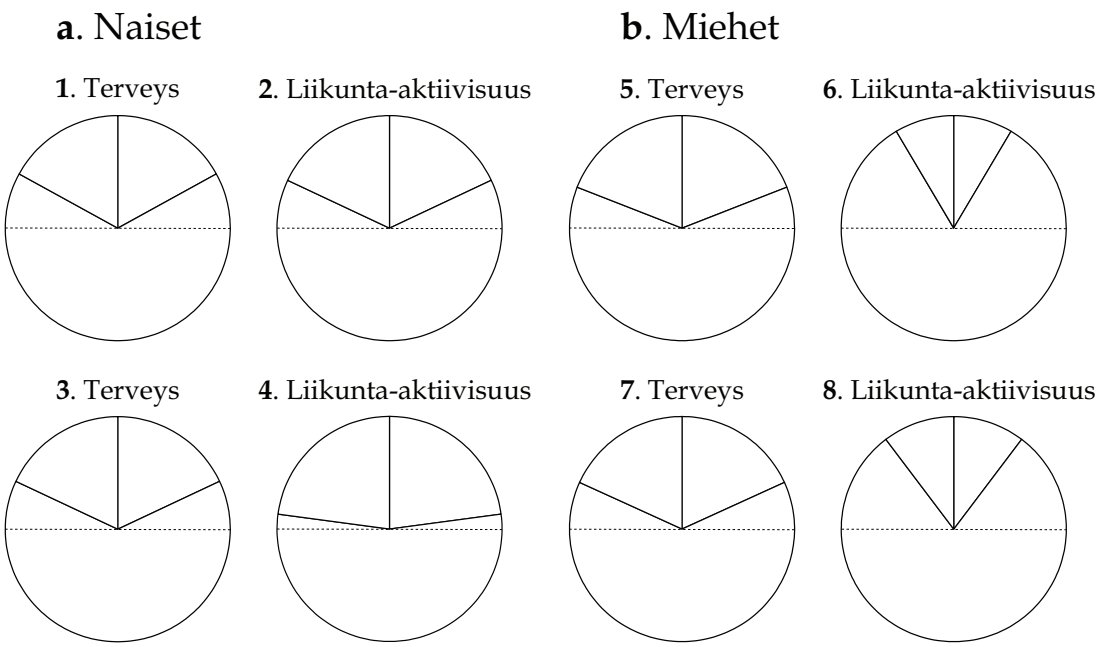
Kuva 8: Kanonisten muuttujien kertoimien 95%:n luottamusvälit miesten aineistossa, kun ikä on vakioitu.

Taulukko 10: Luottamusalueen ala- ja ylärajat (asteissa) vektorille, jonka koordinaatit muodostavat kertoimet ensimmäiselle kanoniselle muuttujalle.

		Terveys		Liikunta-aktiivisuus	
		alaraja	yläraja	alaraja	yläraja
Naiset					
	Ikää ei otettu huomioon	−62.4	62.4	−65.5	65.5
	Ikä vakioitu	−65.5	65.5	−82.6	82.6
Miehet					
	Ikää ei otettu huomioon	−68.7	68.7	−30.6	30.6
	Ikä vakioitu	−70.4	70.4	−37.0	37.0

edellisestä ainoastaan suurimpien painojen omaavien muuttujien estimaattien etumerkin suhteen, missä painokertoimien estimaatit on laskettu kovarianssimatriisin avulla. Estimaatit ovat epätarkkoja, ja luottamusvälit ovat leveitä.

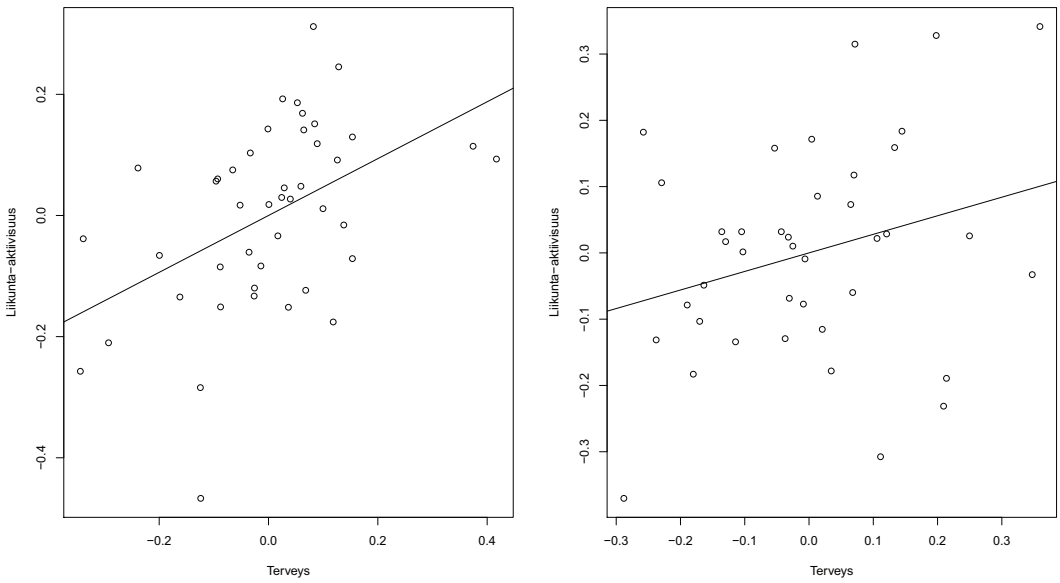
Taulukossa 10 ovat 95%:n luottamusalueet ensimmäisille kanonisille muuttujille, jotka on estimoitu käyttämällä luvussa 4.2 esitettyä menetelmää. Estimoinnissa käytetään alkuperäisiä eli standardoimattomia muuttujia. Taulukon 10 ensimmäisellä rivillä oleva luku (62.4) tarkoittaa aineistosta ja bootstrapilla laskettujen kerroinvektoreiden välistä kulmaa asteina mitattuna. Kuvassa 9 on saadut luottamusalueet graafisesti esitettyinä. Esitetään saadut luottamusalueet kaksiulotteisessa avaruudessa eli ympyröinä ymmärtämisen helpottamiseksi. Siinä on ympyrä, jossa alkuperäisestä aineistosta estimoitu vektori on valittu mielivaltaisesti siten, että se on suunnattu suoraan ylöspäin. Vektoreiden skalaaritulo on positiivinen, niin kulman suuruus sijoittuu välille $[-\frac{\pi}{2}, \frac{\pi}{2}]$, kuten luvussa 4.2 esitettiin. Esimerkiksi taulukon 10 luku 62.4 tarkoittaa bootstrapilla lasketun vektorin ja aineistosta estimoidun vektorin välistä kulmaa. Luku -62.4 vastaa vektoreiden välistä kulmaa, joka on laskettu vastapäivään. Naisten ja miesten tulokset luottamusalueiden osalta ovat pääasiassa samansuuntaisia kuin luottamusvälien tapauksessa. Luottamusalueet naisilla ja miehillä ovat isot ja sopusoinnussa sen kanssa, että myös luottamusvälit ovat leveitä. Sekä miehillä että naisilla on selviä eroja luottamusalueissa, kun ikä otetaan huomioon. Luottamusalueet ovat isompia verrattuna analyysin tuloksiin, missä ikää ei ole otettu huomioon. Tämä tulos on odotusten mukaista - kun ikä on otettu huomioon kanonisten muuttujien välille ei jää merkitsevää riippuvuutta ja luottamusalueet ovat isot. Naisilla liikunta-aktiivisuutta kuvaavan kanonisen muuttujan luottamusalue on sen verran iso, että mikä tahansa vektori käy. Miesten liikunta-aktiivisuuden kanonisen muuttujan luotta-



Kuva 9: Kanonisten muuttujien luottamusalue naisten (**a**, **1**-**4**) ja miesten (**b**,**5**-**8**) aineistoissa. Luottamusalue on esitetty ympyröinä, joissa ylöspäin suunnattu vektori on alkuperäisestä aineistosta estimoitu vektori ja luottamusalueen rajat on kuvattu ympyrän keskipisteestä lähtevällä jatkuvalla viivalla. Ylimpänä (**1**-**2**) ja (**5**-**6**) ikää ei ole otettu huomioon ja alimpana (**3**-**4**) ja (**7**-**8**) ikä on vakioitu.

musalue on huomattavasti kapeampi naisiin verrattuna.

Kuvassa 10 ovat hajontakuviot naisten ja miesten ensimmäisille kanonisille muuttujille tapauksessa, kun iän vaikutus otetaan huomioon. Sukupuolten välillä on nähtävissä selviä eroja. Miehillä on paljon enemmän vaihtelua. Nähdään, että pisteet hajontakuviossa ovat melko satunnaisesti jakautuneet.



Kuva 10: Naisten (vasemmalla) ja miesten (oikealla) hajontakuviot ensimmäiselle kanoniselle parille, joihin on lisätty regressiosuorat. Ikä on otettu huomioon. Korrelaatiokertoimet ovat 0.469 ja 0.280.

LUKU 6

Yhteenveto ja johtopäätökset

Tässä tutkielmassa tutkittiin terveyttä ja liikunta-aktiivisuutta kuvaavien muuttujien välisiä riippuvuussuhteita käyttämällä tutkimusmenetelmänä kanonista korrelaatioanalyysiä. Analyysin tuloksena syntyy kanonisia korrelaatioita ja niitä vastaavia vektoreita, joiden luottamusvälejä oli tarkoitus estimoida bootstrap-menetelmällä. Tehtiin kaksi analyysiä, joista ensimmäisessä iän vaikutus jätettiin huomiotta ja toisessa iän vaikutus otettiin huomioon. Tarkoituksena oli myös selvittää, mitkä muuttujista ovat ratkaisevia terveys- ja liikunta-aktiivisuusryhmien välisessä yhteydessä.

Kanonisen korrelaatioanalyysin perusteella niin naisilla kuin miehilläkin muuttujaparien välisistä kanonisista korrelaatioista ainoastaan ensimmäinen kanoninen korrelaatio osoittautui tilastollisesti merkitseväksi silloin, kun ikää ei ollut otettu huomioon. Terveyttä ja liikunta-aktiivisuutta kuvaavien kanonisten muuttujien painokertoimien estimaatit olivat melko samanlaisia etumerkkiä lukuun ottamatta molemmilla sukupuolilla. Näin ollen naisilla ja miehillä verenpaineen ja päivittäisen aktiivisuuden välillä oli positiivinen yhteys, jolle on vaikea löytää mielekästä tulkintaa. Taustamuuttujana ikä tuo vaihtelua aineistoon, joten analyysissä iästä johtuva vaihtelu on otettava huomioon. Kun ikä oli vakioitu, terveys- ja liikunta-aktiivisuusmuuttujien välille ei jäänyt merkitsevää riippuvuutta. Naisilla painokertoimien estimaateissa oli havaittavissa selviä eroja edellisen analyysin estimaatteihin verrattuna. Tulokset osoittivat, että sisäelinten ympärillä oleva rasva ja epäaktiivisuusaika dominoivat terveys- ja liikunta-aktiivisuutta kuvaavissa kanonisissa muuttujissa. Miesten ryhmässä ei sen sijaan ollut olennaisia eroja tulosten välillä. Tällaiset erot tuloksissa naisilla ja miehillä voivat johtua siitä, että naiset ja miehet ovat biologisesti erilaisia ja sukupuolen sisäiset riippuvuudet miehillä ja naisilla voivat olla erilaisia. Esimerkiksi hormonaaliset

tekijät voivat vaikuttaa terveyteen ja fyysiseen kuntoon. Kaiken kaikkiaan osoittautuu, että terveyden ja fyysisen aktiivisuuden välillä kyseisen tutkimusaineiston perusteella ei ole merkitsevää riippuvuutta.

Tutkielmassa laskettiin 95%:n luottamusvälit saaduille estimaateille bootstrap-algoritmilla. Jälkimmäisessä analyysissä, jossa iän vaikutus poistettiin, sekä miehille että naisille oli tyypillistä , että luottamusvälit estimaateille olivat leveämmät kuin luottamusvälit, jotka oli saatu, kun ikä oli jätetty huomiotta. Tosin estimaattien luottamusvälit kertoimille olivat myös varsin leveitä, kun ikää ei ollut otettu huomioon. Luottamusalueet jäivät liian suuriksi, mikä on yhdenmukaista sen kanssa, että luottamusvälit olivat leveitä.

On mahdollista, että otos ei ollut täysin satunnainen eli kyse on valikoitumisesta. Fyysisesti aktiiviset ja hyvässä fyysisessä kunnossa olevat henkilöt olivat kiinnostuneimpia osallistumaan fyysisen kunnon ja aktiivisuuden arviointia koskevaan tutkimukseen kuin vähän liikkuvat henkilöt. Otoskoot jäivät liian pieniksi luotettavien tulosten saamiseksi. On myös mahdollista, että suuremmalla aineistolla päädyttäisiin toisenlaisiin tuloksiin. On tärkeä ottaa ikä jotenkin huomioon poiminnassa.

Lähteet

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Wiley, New York.

Beran, R., Srivastava, M. S. (1985). Bootstrap Tests and Confidence Regions for Functions of a Covariance Matrix. *The Annals of Statistics* **13**, 95-115.

Dillon, W. R., Goldstein, M. (1984). *Multivariate analysis. Methods and applications*. Wiley, New York.

Efron, B., Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. Chapman and Hall, New York.

Manly, B. F. J. (1991). *Randomization and Monte Carlo Methods In Biology*. Chapman and Hall, London.

Rao, C. R. (2002). *Linear Statistical Inference and its Application*. 2nd ed. Wiley, New York.

Tikkanen, O., Haakana, P., Pesola, A. J., Häkkinen, K., Rantalainen, T., et al. (2013). Muscle Activity and Inactivity Periods during Normal Daily Life. *PLoS ONE* **8**, *e52228*. doi:10.1371/journal.pone.0052228.

Liite

R-koodit

```
## Simulointikoe likimääräiseen permutaatiotestiin liittyen ##

# n = havaintojen lkm
# m = komponenttien lkm
# k = 2*m, h, s, p = parametreja
# rho = korrelaatiokerroin
# ntoisto = toistojen lkm
# permutations = permutaatioiden lkm

n ← 50
m ← 2
k ← 4
rho ← 0.7
h ← s ← p ←(m+k)/2
ntoisto ← 1000
permutations ← 100

chi_2.vector ← matrix(0, nrow = ntoisto, ncol = 1)
p_value.matrix ← matrix(0, nrow = ntoisto, ncol = 1)

generate_data ← function(n, m, k, rho)
{
## aineiston generointi
   Z ← matrix(rnorm(n*m, mean = 0, sd = 1), ncol = 2)
   M ← array(rho, dim = c(m, m))
   diag(M) ← 1
   L ← chol(M)
   X ← Z%*%L

   X1 ← as.matrix(X[,1])
   Y1 ← as.matrix(X[,2])
   XY ← matrix(0, n, k)
   for (j in 1:k)
   {
      XY[,j] ← rnorm(n, mean = 0, sd = 1)
   }
   X_data ← as.matrix(cbind(X1, XY[,1:(k/2)]))
   Y_data ← as.matrix(cbind(Y1, XY[,((k/2)+1):k]))
   return(list(X_data = X_data, Y_data = Y_data))
}

canonical_variables ← function(X_data, Y_data, kankor, n, q, h)
{
## muodostetaan kanoniset muuttujat
   U ← matrix(0, ncol = q, nrow = n)
   V ← matrix(0, ncol = q, nrow = n)

   for (j in 1:q)
   {
      for (l in 1:h)
      {
         U[,j] ← U[,j] + (kankor$xcoef[l,j] * X_data[,l])
         V[,j] ← V[,j] + (kankor$ycoef[l,j] * Y_data[,l])
      }
```

```
   }
   return(list(U = U, V = V))
}

for (i in 1:ntoisto)
{
   XY_data ← generate_data(n, m, k, rho)
   X_data ← XY_data$X_data
   Y_data ← XY_data$Y_data

   kankor ← cancor(X_data,Y_data)
   q ← length(kankor$cor)

   UV_data ← canonical_variables(X_data, Y_data, kankor, n, q, h)
   U ← UV_data$U
   V ← UV_data$V
   U1 ← U[,1]
   V1 ← V[,1]
   U2 ← U[,2:q]
   V2 ← V[,2:q]

   ## lasketaan chi^2- testisuureen arvo
   lambda0 ← matrix(0, nrow = 1, ncol = s)
   correlation ← cor(U,V)
   lambda0 ← (diag(correlation))^2
   chi_2 ← -((n-1)-0.5*(s+p+1)) * sum(log(1-lambda0[2:s]))
   chi_2.vector[i,] ← chi_2

   lambda ← matrix(nrow=permutations, ncol=ncol(U))
   for (j in 1:permutations)
   {
   ## tehdään permutaatio ja tallennetaan kanoniset korrelaatiot
      permU2 ← U2
      permU2 ← permU2[sample(nrow(U2)),]
      permU ← as.matrix(cbind(U1, permU2))
      canonical_analysis ← cancor(permU, V)
      canonical_correlations ← canonical_analysis$cor
      lambda[j,] ← (canonical_correlations)^2
   }
## lasketaan chi^2- testisuure jokaiselle permutaatiolle
chi_2.trial ←  -((n-1)-0.5*(s+p+1)) * apply(log(1-lambda[, 2:s]), 1, sum)

## lasketaan p-arvo permutaatiojakauman perusteella
p_value ← mean(chi_2.trial > chi_2)
p_value.matrix[i,] ← p_value
}

X11()
plot(sort(p_value.matrix), ylab = "p-arvo", pch = ".")
lines(x = c(0,1000), y = c(0,1), lwd = 2)
```

```
## Luottamusvälit kanonisten muuttujien kertoimille ##
## Tässä R-koodi on esitetty naisten aineistolle. R-koodit miesten aineistolle ovat
     vastaavanlaisia.

aineisto ← read.table("EMG12.dat", header=TRUE)
attach(aineisto)

# Muuttujat
# Terveys muuttujat: BP15sys, Visceralfat, Strigly
# Liikunta-aktivisuus muuttujat: METh12kk, Average amplitude, Longest inactivity
    period
# n = otoskoko

n ← length(aineisto$id)
X ← aineisto[aineisto$sex == 1, 9:11]
Y ← aineisto[aineisto$sex == 1 , 15:17]

log_Strigly ← log(X[,3])
X ← cbind(X[,1:2], log_Strigly)

log_Average ← log(Y[, 2])
log_Longest ← log(Y[, 3])
Y ← as.matrix(cbind(Y[,1], log_Average, log_Longest))

X_Muuttuja ← ncol(X)
Y_Muuttuja ← ncol(Y)
n_cancor ← min(X_Muuttuja, Y_Muuttuja)
# nboot = bootstrap-otosten lkm
nboot ← 1000

cancorXY ← cancor(X,Y)
cancor_corrs ← cancorXY$cor[1:n_cancor]
cancor_x_coeffs ← cancorXY$xcoef[,1:n_cancor]
cancor_y_coeffs ← cancorXY$ycoef[,1:n_cancor]

## normeerataan ominaisvektorit ykkösen pituisiksi
norm_cancor_x_coeffs ← matrix(0, nrow = nrow(cancor_x_coeffs), ncol = n_cancor)
norm_cancor_y_coeffs ← matrix(0, nrow = nrow(cancor_y_coeffs), ncol = n_cancor)
for (k in 1:ncol(X))
{
   for (l in 1:nrow(cancor_x_coeffs))
   {
      norm_cancor_x_coeffs[l,k] ← cancor_x_coeffs[l,k]/sqrt(sum(cancor_x_coeffs[,k]
          ^2))
      norm_cancor_y_coeffs[l,k] ← cancor_y_coeffs[l,k]/sqrt(sum(cancor_y_coeffs[,k]
          ^2))
   }
}

corr ← matrix(0, nrow = nboot, ncol = n_cancor)
x_coeffs ← matrix(0, nrow = nboot, ncol = X_Muuttuja * n_cancor)
y_coeffs ← matrix(0, nrow = nboot, ncol = Y_Muuttuja * n_cancor)

for (j in 1:nboot)
{
## poimitaan bootstrap-otoksia, lasketaan ja tallennetaan kanoniset vektorit
  i ← sample(1:havainnot, replace=TRUE)
  X_boot ← X[i,]
  Y_boot ← Y[i,]
  cancor_boot ← cancor(X_boot,Y_boot)
```

```
  boot_corr ← cancor_boot$cor[1:n_cancor]
  boot_x_coeffs ← cancor_boot$xcoef[,1:n_cancor]
  boot_y_coeffs ← cancor_boot$ycoef[,1:n_cancor]

  norm_boot_x_coeffs ← matrix(0, nrow=nrow(boot_x_coeffs), ncol=ncol(boot_x_coeffs)
      )
  norm_boot_y_coeffs ← matrix(0, nrow=nrow(boot_y_coeffs), ncol=ncol(boot_y_coeffs)
      )

  for (k in 1:ncol(boot_x_coeffs))
  {
      for (l in 1:nrow(boot_x_coeffs))
      {
      norm_boot_x_coeffs[l,k] ← boot_x_coeffs[l,k]/sqrt(sum(boot_x_coeffs[,k]^2))
      norm_boot_y_coeffs[l,k] ← boot_y_coeffs[l,k]/sqrt(sum(boot_y_coeffs[,k]^2))
      }
   }
   corr[j,] ← boot_corr
   x_coeffs[j,] ← c(norm_boot_x_coeffs)
   y_coeffs[j,] ← c(norm_boot_y_coeffs)
}

## valitaan bootstrapilla saatujen vektoreiden etumerkki
X1_coeffs ← matrix(x_coeffs[,1:3], ncol=ncol(X), nrow=nboot)
X2_coeffs ← matrix(x_coeffs[,4:6], ncol=ncol(X), nrow=nboot)
X3_coeffs ← matrix(x_coeffs[,7:9], ncol=ncol(X), nrow=nboot)

Y1_coeffs ← matrix(y_coeffs[,1:3], ncol=ncol(Y), nrow=nboot)
Y2_coeffs ← matrix(y_coeffs[,4:6], ncol=ncol(Y), nrow=nboot)
Y3_coeffs ← matrix(y_coeffs[,7:9], ncol=ncol(Y), nrow=nboot)

for (i in 1:nboot)
{
   if (X1_coeffs[i,] %*% norm_cancor_x_coeffs[,1] < 0) X1_coeffs[i,] ← - X1_coeffs[
       i,]
   if (X2_coeffs[i,] %*% norm_cancor_x_coeffs[,2] < 0) X2_coeffs[i,] ← - X2_coeffs[
       i,]
   if (X3_coeffs[i,] %*% norm_cancor_x_coeffs[,3] < 0) X3_coeffs[i,] ← - X3_coeffs[
       i,]

   if (Y1_coeffs[i,] %*% norm_cancor_y_coeffs[,1] < 0) Y1_coeffs[i,] ← - Y1_coeffs[
       i,]
   if (Y2_coeffs[i,] %*% norm_cancor_y_coeffs[,2] < 0) Y2_coeffs[i,] ← - Y2_coeffs[
       i,]
   if (Y3_coeffs[i,] %*% norm_cancor_y_coeffs[,3] < 0) Y3_coeffs[i,] ← - Y3_coeffs[
       i,]
}

## Prosenttipistemenetelmä ##

## lasketaan kanonisten muuttujien kertoimien 95% luottamusvälit
    prosenttipistemenetelmällä
x_coeffsNEW ← as.matrix(cbind(X1_coeffs, X2_coeffs, X3_coeffs))
y_coeffsNEW ← as.matrix(cbind(Y1_coeffs, Y2_coeffs, Y3_coeffs))

sort_x_coeffsNEW ← matrix(0, nrow = nboot, ncol = ncol(x_coeffsNEW))
sort_y_coeffsNEW ← matrix(0, nrow = nboot, ncol = ncol(y_coeffsNEW))

for (i in 1:ncol(x_coeffsNEW))
```

```
{
   sort_x_coeffsNEW[,i] ← sort(x_coeffsNEW[,i])
   sort_y_coeffsNEW[,i] ← sort(y_coeffsNEW[,i])
}
x_coeffs_lower ← matrix(0, nrow=1, ncol=ncol(x_coeffsNEW))
x_coeffs_upper ← matrix(0, nrow=1, ncol=ncol(x_coeffsNEW))
y_coeffs_lower ← matrix(0, nrow=1, ncol=ncol(y_coeffsNEW))
y_coeffs_upper ← matrix(0, nrow=1, ncol=ncol(y_coeffsNEW))

## lasketaan luottamusvälien ala- ja ylärajat
for (i in 1:ncol(x_coeffsNEW))
{
   x_coeffs_lower[i] ← sort_x_coeffsNEW[25,i]
   x_coeffs_upper[i] ← sort_x_coeffsNEW[976,i]
   y_coeffs_lower[i] ← sort_y_coeffsNEW[25,i]
   y_coeffs_upper[i] ← sort_y_coeffsNEW[976,i]
}

## Luottamusalue kulman avulla ##

cos.matrix.x1.coeffs ← matrix(0, ncol=1, nrow=nboot)
cos.matrix.x2.coeffs ← matrix(0, ncol=1, nrow=nboot)
cos.matrix.x3.coeffs ← matrix(0, ncol=1, nrow=nboot)

cos.matrix.y1.coeffs ← matrix(0, ncol=1, nrow=nboot)
cos.matrix.y2.coeffs ← matrix(0, ncol=1, nrow=nboot)
cos.matrix.y3.coeffs ← matrix(0, ncol=1, nrow=nboot)

for (i in 1:nboot)
{
## lasketaan vektoreiden välinen kulma
   cos.matrix.x1.coeffs[i,] ← X1_coeffs[i,] %*% norm_cancor_x_coeffs[,1]
   cos.matrix.x2.coeffs[i,] ← X2_coeffs[i,] %*% norm_cancor_x_coeffs[,2]
   cos.matrix.x3.coeffs[i,] ← X3_coeffs[i,] %*% norm_cancor_x_coeffs[,3]

   cos.matrix.y1.coeffs[i,] ← Y1_coeffs[i,] %*% norm_cancor_y_coeffs[,1]
   cos.matrix.y2.coeffs[i,] ← Y2_coeffs[i,] %*% norm_cancor_y_coeffs[,2]
   cos.matrix.y3.coeffs[i,] ← Y3_coeffs[i,] %*% norm_cancor_y_coeffs[,3]
}

sort.cos.matrix.x ← as.matrix(cbind(cos.matrix.x1.coeffs, cos.matrix.x2.coeffs,
    cos.matrix.x3.coeffs))
sort.cos.matrix.y ← as.matrix(cbind(cos.matrix.y1.coeffs, cos.matrix.y2.coeffs,
    cos.matrix.y3.coeffs))

for (i in 1:n_cancor)
{
   sort.cos.matrix.x[,i] ← sort(sort.cos.matrix.x[,i])
   sort.cos.matrix.y[,i] ← sort(sort.cos.matrix.y[,i])
}

## 180*(angle radian/pi) = angle degrees

angle.degrees.lower.x ← matrix(0, nrow = 1, ncol = n_cancor)
angle.degrees.upper.x ← matrix(0, nrow = 1, ncol = n_cancor)
angle.degrees.lower.y ← matrix(0, nrow = 1, ncol = n_cancor)
angle.degrees.upper.y ← matrix(0, nrow = 1, ncol = n_cancor)

for (i in 1:n_cancor)
{
```

```
    angle.degrees.lower.x[i] ← -acos(sort.cos.matrix.x[25, i])/pi*180
    angle.degrees.upper.x[i] ← acos(sort.cos.matrix.x[25, i])/pi*180

    angle.degrees.lower.y[i] ← -acos(sort.cos.matrix.y[25, i])/pi*180
    angle.degrees.upper.y[i] ← acos(sort.cos.matrix.y[25, i])/pi*180

}
```