Tampereen yliopisto

Jasmine Hakala

Kyselyaineiston analysointi pääkomponentti- ja regressioanalyysia hyödyntäen

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Huhtikuu 2022

Tiivistelmä

Jasmine Hakala: Kyselyaineiston analysointi pääkomponentti- ja regressioanalyysia hyödyntäen
Kandidaattitutkielma
Tampereen yliopisto
Tilastollisen data-analyysin tutkinto-ohjelma
Huhtikuu 2022

Tässä tutkielmassa tutkitaan koronaepidemiaan liittyvien huolien yhteyttä harrastuksiin, terveydentilaan ja luontokokemuksiin. Aineistona käytetään valmista kyselytutkimusta, jolla on kartoitettu luonnossa ulkoilua ja sen merkitystä epidemian aikana. Kyselyyn on vastattu viisiportaisen Likertin asteikon avulla. Aineistossa on useita kysymyksiä huoliin, harrastuksiin ja luontokokemuksiin liittyen. Työn kannalta on oleellista pyrkiä yksinkertaistamaan alkuperäisten muuttujien joukkoa. Tutkimuskysymyksenä on, miten koronaepidemiaan liittyvät huolet ovat olleet yhteydessä harrastuksiin, terveydentilaan ja luontokokemuksiin.

Pääkomponenttianalyysi on tilastotieteellinen analyysi, jonka tarkoituksena on selittää uusien korreloimattomien muuttujien avulla yhteyksiä alkuperäisten korreloituneiden muuttujien välillä. Analyysia käytetään mallin yksinkertaistamiseen. Regressioanalyysi on yksi tunnetuimmista tilastollisista menetelmistä, joka estimoi muuttujien väliset yhteydet. Analyysilla voidaan kuvailla aineistoa ja testata muuttujien välisten yhteyksien hypoteeseja. Poistovalinnan avulla täydellisestä regressiomallista poistetaan vaiheittain vähiten merkitsevimmät muuttujat. Tutkielman aiheena on pääkomponentti- ja regressioanalyysin keinoin etsiä yhteyksiä muuttujien ja vasteen väliltä eli työn tavoitteena on tuottaa yksi mahdollisimman yksinkertainen tilastollisesti merkitsevä malli.

Aluksi tutkielmassa esitellään analyysien teoriataustat. Tutkielman kannalta oleelliseksi muodostunutta poistovalintaa esitellään regressioanalyysin yhteydessä. Tämän jälkeen tutkielmassa perehdytään Likertin asteikon ongelmiin. Tutkielma päättyy tulosten esittelyyn ja analysointiin sekä yhteenvetoon tutkielmasta.

Menetelmät sopivat hyvin aineistoon ja tutkielman tärkeimpänä tuloksena saadaan regressiomalli. Mallissa vastetta eli koronaepidemiaan liittyviä huolia selitetään tilastollisesti merkitsevien muuttujien avulla. Merkitseviksi muuttujiksi saadaan terveydentila, harrastukset, kuten ohjelmien katselu, syöminen ja puhelut läheisten kanssa sekä erilaiset luontokokemukset.

Avainsanat: pääkomponenttianalyysi, regressioanalyysi, poistovalinta

Sisällys

1 Johdanto

Pääkomponenttianalyysi on yleisesti käytetty tilastollinen menetelmä, jonka tavoitteena on pienentää dimensiota etsimällä keskeisimmät pääkomponentit niin, ettei merkitsevää tietoa menetetä (Suryanarayana ja Mistry 2016). Regressioanalyysi on yksi tunnetuimmista analyyseista tilastolliseen mallintamiseen. Analyysin avulla voidaan estimoida muuttujien väliset yhteydet. Sen avulla tutkitaan, ovatko muuttujat merkitsevästi yhteydessä vasteeseen ja etsitään yhteyksiä muuttujien välillä. (Kim 2021)

Viime aikoina on tutkittu koronaepidemian yhteyttä mielenterveyteen. Tutkimuksissa on havaittu, että koronaepidemia on aiheuttanut huomattavia riskejä mielenterveydelle (Kilkku 2020). Arnoutin et al. (2020) tutkimus puoltaa myös väitettä, että koronaepidemia on aiheuttanut huomattavia riskejä mielenterveydelle. Mielenterveyden muutoksia on aiheuttanut lisääntynyt stressi, johon liittyy vahvasti pelko, turhautuminen, suru, ahdistus, häpeän tunne ja kateus (Arnout et al. 2020). Kilkku (2020) korostaa, että koronaepidemian pitkäaikaiset vaikutukset tulevat näkymään vasta myöhemmin.

Sainio et al. (2021) ovat tutkineet huolien lisääntymistä koronaepidemian aikana. Tutkimuksessa havaittiin, että toimintarajoitteisilla henkilöillä korostui huolet, yksinäisyys, univaikeudet sekä heikentynyt taloudellinen tilanne (Sainio et al. 2021). Ruotsissa tehdyssä tutkimuksessa on havaittu, että huoli koronapandemiasta on levinnyt laajalle (Kulin et al. 2021). Aikaisemmissa tutkimuksissa korostuu demografisten tekijöiden yhteys huoliin. Tässä tutkielmassa keskitytään tutkimaan muiden kuin demografisten muuttujien yhteyttä huoliin.

”Arki ja ulkoilu koronaepidemian aikana” -kyselyn tarkoituksena on ollut kerätä tietoa arjesta ja ulkoilusta koronaepidemian aikana Suomessa vuonna 2020 (Korpela, Salonen ja Hyvönen 2020). Tämän työn tavoitteena on kyseisen aineiston pohjalta etsiä koronaepidemiaan liittyvien huolien yhteyttä harrastuksiin, luontokokemuksiin ja terveydentilaan koronaepidemian aikana.

Huolet, harrastukset ja luontokokemukset perustuvat useampaan kysymykseen. Koska aineisto on suuri ja muuttujia on paljon, on tarkoituksenmukaista vähentää dimensiota. Pääkomponenttianalyysin avulla saadaan peruste käyttää useista muuttujista niiden summamuuttujaa. Tämän käyttäminen vähentää dimensiota huomattavasti. Regressioanalyysia käytetään muuttujien välisten yhteyksien löytämiseksi. Erityisesti ollaan kiinnostuneita luontokokemuksien ja huolien yhteydestä ja siitä, miten yhteys on muuttunut koronaepidemian edetessä.

Regressioanalyysin avulla saatujen yhteyksien kausaliteetti ei ole yksiselitteinen, eikä voida olla varmoja syy-seuraussuhteesta. Tässä tutkielmassa keskitytään yhteyksien löytämiseen, eikä syy-seuraussuhteen tutkimiseen. Aluksi perehdytään analyysien teoriataustaan. Tämän jälkeen käsitellään Likertin asteikon käyttöön liittyviä ongelmia. Lopuksi sovelletaan menetelmiä aineistoon ja tulkitaan tulokset.

2 Pääkomponenttianalyysi

Pääkomponenttianalyysi on yleinen apuväline tilastollisessa mallintamisessa. Analyysin tarkoituksena on selittää yhteyksiä korreloituneiden muuttujien välillä uusien korreloimattomien muuttujien eli pääkomponenttien avulla. (Nummi 2021) Sen tärkein käyttökohde on mallin yksinkertaistaminen ilman, että hyödyllistä tietoa menetetään (Suryanarayana ja Mistry 2016).

Analyysin kehitti Karl Pearson vuonna 1901 ja sitä hyödynnetään monilla eri tieteenaloilla, kuten luonnontieteissä, terveystieteissä ja kauppatieteissä. Tutkimuksissa kerätään usein paljon erilaisia muuttujia. Tällöin on helppo löytää tilastollisesti merkitsevät muuttujat pääkomponenttianalyysin avulla. Analyysin avulla identifioidaan joukkoja aineistosta. Identifioinnissa korostuu yhteneväisyydet ja erilaisuudet muuttujien välillä. (Suryanarayana ja Mistry 2016)

Joskus regressioanalyysissa selittäviä muuttujia on liian monta tai muuttujien välillä on suurta korrelaatiota (Nummi 2021), eikä tällöin graafinen tarkastelu tai tärkeiden muuttujien identifiointi onnistu helposti. Yksinkertainen graafinen tarkastelu pääkomponenttianalyysin avulla mahdollistaa ulkopuolisten havaintojen ja ryhmittymien havaitsemisen muita analyyseja helpommin. Pääkomponenttiregressiolla tarkoitetaan sitä, kun monen selittäjän regressioanalyysissa käytetään apuna pääkomponenttianalyysia. Sen avulla voidaan luoda esimerkiksi ennustemalleja. (Suryanarayana ja Mistry 2016)

2.1 Pääkomponenttipisteiden laskeminen

Pääkomponenttipisteet lasketaan matemaattisin keinoin niin, että ensimmäinen komponentti selittää vaihtelusta enimmäisosuutta (Suryanarayana ja Mistry 2016). Olkoon $x_1, x_2, x_3, \ldots, x_p$ aineiston korreloituneita muuttujia. Tällöin uudet korreloimattomat muuttujat merkitään $y_1, y_2, y_3, \ldots, y_q$ $(q < p)$. Ensimmäinen pääkomponentti voidaan kirjoittaa muodossa

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = \boldsymbol{a}_1'\boldsymbol{x},$$

mikä toteuttaa ehdon $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$ ja maksimoi varianssin $\mathrm{Var}(y_1)$.
Toinen pääkomponentti voidaan kirjoittaa muodossa

$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p = \boldsymbol{a}_2'\boldsymbol{x},$$

mikä toteuttaa ehdon $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$ ja maksimoi varianssin $\mathrm{Var}(y_2)$. Toinen pääkomponentti y_2 ei korreloi ensimmäisen pääkomponentin y_1 kanssa.
Yleisemmin j. pääkomponentti voidaan kirjoittaa muodossa

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p = \boldsymbol{a}_j'\boldsymbol{x},$$

mikä toteuttaa ehdon $\boldsymbol{a}_j'\boldsymbol{a}_j = 1$ ja maksimoi varianssin $\mathrm{Var}(y_j)$. Saatu pääkomponentti ei korreloi muiden pääkomponenttien kanssa.
Pääkomponentit voidaan laskea esimerkiksi maksimoimalla y-muuttujan varianssi. Tällöin siis maksimoidaan

$$\text{Var}(y_1) = \text{Var}(\boldsymbol{a}_1'\boldsymbol{x}) = \boldsymbol{a}_1'\text{Var}(\boldsymbol{x})\boldsymbol{a}_1 = \boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_1,$$

mikä toteuttaa ehdon $\boldsymbol{a}_j'\boldsymbol{a}_j = 1$. Merkitään kovarianssimatriisin Σ ominaisarvohajotelmaa

$$\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}',$$

missä $\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_p)$ on ominaisvektoreista koostuva matriisi ja $\boldsymbol{\Lambda}$ on ominaisarvoista koostuva diagonaalimatriisi. Merkitään $\boldsymbol{b}_1 = \boldsymbol{T}\boldsymbol{a}_1$. Tällöin $\boldsymbol{b}_1'$-muuttujan neliösumma on yhtä kuin yksi, koska

$$\boldsymbol{b}_1'\boldsymbol{b}_1 = \boldsymbol{a}_1'\boldsymbol{T}'\boldsymbol{T}\boldsymbol{a}_1 = \boldsymbol{a}_1'\boldsymbol{a}_1 = 1.$$

Ensimmäisen pääkomponentin varianssi on nyt muotoa

$$\text{Var}(y_1) = \boldsymbol{a}_1'\boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}'\boldsymbol{a}_1 = \boldsymbol{b}_1'\boldsymbol{\Lambda}\boldsymbol{b}_1.$$

Koska $\boldsymbol{b}_1'\boldsymbol{b}_1 = b_{11}^2 + \cdots + b_{1p}^2 = 1$, niin

$$\lambda_1 b_{11}^2 + \lambda_2 b_{12}^2 + \cdots + \lambda_p b_{1p}^2 \leq \lambda_1(b_{11}^2 + \cdots + \lambda_p b_{1p}^2) = \lambda_1.$$

Joten pääkomponentin suurin ominaisarvo λ_1 on maksimi ensimmäisen pääkomponentin varianssille $\text{Var}(y_1)$. Oletetaan, että $\boldsymbol{a}_1 = \boldsymbol{t}_1$. Tällöin

$$\text{Var}(\boldsymbol{t}_1\boldsymbol{x}) = \boldsymbol{t}_1'\boldsymbol{\Sigma}\boldsymbol{t}_1 = \boldsymbol{t}_1'(\lambda_1\boldsymbol{t}_1) = \lambda_1.$$

Joten ensimmäinen pääkomponentti voidaan määritellä ensimmäisen ominaisvektorin $\boldsymbol{t}_1$ avulla. Ja näin yleinen pääkomponentti voidaan määritellä yleisen termin ominaisvektorilla $\boldsymbol{t}_j$, jolloin $\text{Var}(x_j) = \lambda_j$.

2.2 Pääkomponenttien valinta

Kun pääkomponenttipisteet on laskettu, päätetään, kuinka monta komponenttia valitaan edustamaan alkuperäisiä muuttujia. Analyysi tuottaa yhtä monta pääkomponenttia kuin mitattuja muuttujia on. Usein suurin osa vaihtelusta pystytään selittämään muutamalla pääkomponentilla, jolloin näitä komponentteja voidaan käyttää myöhemmässä vaiheessa analyysia alkuperäisten muuttujien tilalla. (Suryanarayana ja Mistry 2016)

Pääkomponentit valitaan siten, että peräkkäiset komponentit selittävät suurimman osan mahdollisesta jäljellä olevasta vaihtelusta (Suryanarayana ja Mistry 2016). Valittujen komponenttien tulisi selittää vaihtelusta noin 70–90 %. Joskus käytetään sääntöä, että valittujen komponenttien ominaisarvojen tulisi olla suurempia kuin ominaisarvojen keskiarvo $\bar{\lambda}$. (Nummi 2021) Joskus valitaan ne pääkomponentit, joiden ominaisarvo on suurempi kuin yksi (Suryanarayana ja Mistry 2016). Rajana voidaan käyttää myös lukua 0.7. Pääkomponenteista luotu kuvaaja auttaa kynnysarvon löytämistä. (Nummi 2021)

On helppo osoittaa, että ominaisarvojen summa on yhtä kuin varianssien summa,

$$\textstyle\sum_{i=1}^{p} \lambda_i = \text{Var}(x_1) + \text{Var}(x_2) + \cdots + \text{Var}(x_p).$$

Tällöin yleisen termin pääkomponentin selitysosuus P_j saadaan laskettua seuraavasti:

$$P_j = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} = \frac{\lambda_j}{tr(\boldsymbol{\Sigma})}.$$

Selitysosuuden avulla voidaan valita pääkomponentit.

3 Regressioanalyysi

Regressioanalyysi on yksi tunnetuimmista tilastollisista menetelmistä, joka estimoi muuttujien väliset yhteydet. Analyysi tarjoaa joustavan keinon kuvailla aineistoa ja testata muuttujien välisten yhteyksien hypoteeseja. Tyypillisesti sitä käytetään estimoimiseen, ennustamiseen ja hypoteesitestaukseen. Analyysin on kehittänyt Sir Francis Galton 1800-luvulla, kun hän huomasi lasten pituuksien lähestyvän väestön keskipituutta. (Kim 2021)

3.0.1 Yhden selittäjän regressio

Regressioanalyysin pohjalla on lineaarinen malli. Yhden selittäjän regression avulla voidaan tutkia kahden jatkuvan muuttujan välistä suhdetta. (Kim 2021) Vastetta Y selitetään muuttujalla X. Yhden selittäjän regressio on tällöin muotoa

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

missä jäännöstermit ϵ_i oletetaan riippumattomiksi.

3.0.2 Monen selittäjän regressio

Monen selittäjän regressiossa oletetaan n kappaletta havaintoja ja p kappaletta mahdollisia selittäjiä $\{X_{1i}, X_{2i}, \ldots, X_{pi}, Y_i\},\ i = 1, \ldots, n$ (Draper ja Smith 1998). Oletetaan, että havainnot ovat satunnaisotos populaatiosta, ja että niillä on lineaarinen yhteys (Kim 2021). Monen selittäjän regressio on muotoa

$$\begin{gathered} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i, \\ i = 1, \ldots, n, \quad n > p, \quad \epsilon_i \sim N(0, \sigma^2), \end{gathered}$$

missä β-kertoimet ovat tuntemattomia parametreja ja ϵ_i-termit ovat tuntemattomia jäännöstermejä, jotka oletetaan riippumattomiksi. Tällöin malli voidaan esittää myös matriisimuotoisena

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

missä $\boldsymbol{\epsilon} \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Mallissa $\boldsymbol{X}$ on ns. suunnittelu- tai mallimatriisi, $\boldsymbol{\beta}$-termi on parametreista koostuva vektori ja $\boldsymbol{\epsilon}$-termi on virhetermeistä koostuva vektori. Oletetaan, että vektorin $\boldsymbol{\epsilon}$ alkiot ovat keskenään riippumattomia.

3.1 Estimointi

Oletetaan, että tulolla $\boldsymbol{X'X}$ on täysi sarakeaste, tällöin voidaan pienimmän neliösumman avulla estimoida parametrien $\beta_0, \beta_1, \ldots, \beta_p$ arvot. Yhtälöstä (3.1) voidaan muotoilla jäännöstermien vektori $\boldsymbol{\epsilon}$ muotoon

$$\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X\beta}.$$

Jäännöstermien neliösumma on muotoa $\boldsymbol{\epsilon'\epsilon}$,

$$\begin{aligned}\boldsymbol{\epsilon'\epsilon} &= (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y'Y} - \boldsymbol{\beta'X'Y} - \boldsymbol{Y'X\beta} + \boldsymbol{\beta'X'X\beta} \\ &= \boldsymbol{Y'Y} - 2\boldsymbol{\beta'X'Y} + \boldsymbol{\beta'X'X\beta}.\end{aligned}$$

Derivoidaan neliösumma $\boldsymbol{\beta}$-termin suhteen,

$$\tfrac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{Y'Y} - 2\boldsymbol{\beta'X'Y} + \boldsymbol{\beta'X'X\beta}) = -2\boldsymbol{X'Y} + 2\boldsymbol{X'X\beta}.$$

Asetetaan neliösumman derivaatta yhtä suureksi kuin nolla,

$$\begin{aligned}-2\boldsymbol{X'Y} + 2\boldsymbol{X'X\beta} &= 0 \\ 2\boldsymbol{X'X\beta} &= 2\boldsymbol{X'Y} \\ \boldsymbol{X'X\beta} &= \boldsymbol{X'Y} \\ (\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta} &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}.\end{aligned}$$

Koska $(\boldsymbol{X'X})^{-1}\boldsymbol{X'X}$ on identiteettimatriisi niin saadaan $\boldsymbol{\beta}$-vektorin estimaatiksi

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}. \tag{3.2}$$

Voidaan todistaa, että yhtälön (3.2) estimaatit ovat harhattomia oletetulle mallille. Tämä voidaan osoittaa seuraavasti:

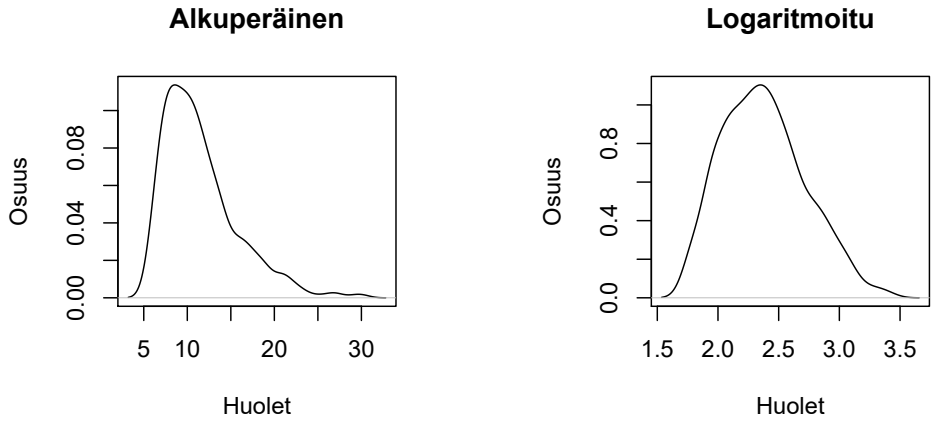
$$\begin{aligned}E(\hat{\boldsymbol{\beta}}) &= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}) \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}E(\boldsymbol{Y}) \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta} \\ &= \boldsymbol{\beta}.\end{aligned}$$

3.2 Regressiomallin muunnos

Lineaarisessa regressiossa oletetaan, että havainnot ovat riippumattomia sekä normaalisti jakautuneita, ja niiden satunnaisvirheiden varianssien olevan yhtä suuria. Jos malli ei toteuta kaikkia oletuksia, voidaan havainnoille tehdä muunnos, kuten BOX-COX-muunnos. (Kim 2021) Muunnokset eivät vaikuta mallin muuttujien hyvyyteen, mutta niiden avulla saadaan käyttökelpoisempia muuttujia. BOX-COX-muunnos on muotoa $u = y^{\alpha}$. Tarkoituksena on etsiä sopivaa α-muuttujan arvoa. Muunnos tuottaa paremmin normaalijakaumaa noudattavan mallin. (Nyblom 2015) Muunnos tuottaa logaritmimuodon, kun

$$\lim_{\alpha \to 0} \frac{y^{\alpha} - 1}{\alpha} = log(y),$$

missä muuttujat saavat suurempia tai yhtä suuria arvoja kuin nolla. Kuvasta 3.1 havaitaan, että huolien summamuuttujan alkuperäinen jakauma ei täytä normaalijakauman oletuksia. Tästä syystä havainnoille tehdään logaritmimuunnos, joka näkyy kuvassa 3.1 logaritmoituna jakaumana. Todetaan, että muunnettu jakauma on lähempänä normaalijakaumaa kuin alkuperäinen jakauma. Jatkoanalyysissa voidaan tämän perusteella käyttää logaritmoitua muotoa.



Kuva 3.1. Alkuperäisten havaintojen jakauma ja logaritmoitu jakauma.

3.3 Mallin rakentaminen ja muuttujien valinta

Käytännön tutkimuksissa voi olla käytettävissä useita mahdollisia muuttujia, joilla voitaisiin selittää vastetta. Ylimääräiset muuttujat saattavat aiheuttaa kollineaarisuutta ja kohinaa muihin estimaatteihin. Näissä tilanteissa pyritään selittämään aineistoa mahdollisimman yksinkertaisesti. (Kim 2021) Askeltavan regression avulla voidaan etsiä tärkeimmät muuttujat muuttujien joukosta (Sahay 2016). Menetelmiin kuuluu etenevä valinta ja poistovalinta. Muita keinoja ovat esimerkiksi parhaan osajoukon regressio tai LASSO-regressio. (Kim 2021) Tässä työssä keskitytään poistovalinnan menetelmään.

3.3.1 Poistovalinta

Askeltavassa regressiossa rakennetaan malli mahdollisista muuttujista joko lisäämällä tai poistamalla muuttujia vaiheittain. Jäljelle jääneistä muuttujista muodostetaan uusi malli. Jokaisessa välivaiheessa tutkitaan p-arvoja. (Kim 2021) Huomioitavaa on, että suuri määrä käsiteltyjä muuttujia saattaa johtaa sekä hyväksymis- että hylkäysvirheeseen (Sahay 2016). Poistovalinnassa aloitetaan täyden mallin selittäjien p-arvojen tutkimisella. Selittäjistä poistetaan muuttuja, jolla on suurin p-arvo, ja jonka arvo on suurempaa kuin α_R (Kim 2021). Tutkijan valitseman merkitsevyystason α_R perusteella hyväksytään tai hylätään hypoteesi. Usein käytetty α_R on 0.05. (Lin, Lucas ja Shmueli 2013)

1. Tutkitaan täydellistä mallia, jossa on mukana kaikki selittäjät.
2. Mallista poistetaan selittäjä, jolla on suurin p-arvo, ja jonka arvo on suurempi kuin α_R.
3. Jäljelle jääneistä selittäjistä muodostetaan uusi malli ja palataan kohtaan 2.
4. Poistovalinta lopetetaan, kun kaikkien selittäjien p-arvot ovat pienempiä kuin α_R.

3.4 Selitysaste

Tilastollisesti merkitsevät muuttujat, joiden p-arvot ovat sopivan pieniä, eivät välttämättä kuitenkaan selitä vastetta tarpeeksi, jotta malli olisi käyttökelpoinen. Vaikka regressiomalli tuottaa tilastollisesti merkitsevän muuttujan, se ei välttämättä selitä hajontaa muuttujien välillä. Tällöin ennusteiden tekeminen ei ole kovin luotettavaa. Selitysaste R^2 on tunnetuin lineaaristen mallien sopivuuden mitta. (Draper ja Smith 1998) Se mittaa vaihtelua Y:ssä, joka selittyy regressiolla. Sopiva malli tuottaa lähellä yhtä olevan R^2 luvun. (Kim 2021) R^2 on määritelty seuraavasti:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

R^2-luvun käyttöön liittyy ongelmia. Esimerkiksi selittäjien määrän lisääntyessä R^2-luvun arvo kasvaa, vaikka parametrit eivät olisi merkitseviä muuttujia mallin kannalta. (Draper ja Smith 1998) Tällöin muokattu selitysaste R_m^2-luku on parempi arvio mallin selitysasteesta, koska se kompensoi ylöspäin suuntautuvaa harhaa (Kim 2021). Sen avulla aineistojen selitysasteiden vertailu on luotettavampaa (Draper ja Smith 1998). Muokattu selitysaste R_m^2 on määritelty seuraavasti:

$$R_m^2 = R^2 - \frac{1}{n-2}(1 - R^2).$$

R^2-luku on useimmissa tapauksissa hyödyllinen sopivuuden mitta, mutta tulee kuitenkin muistaa, että luvun käyttöön liittyy ongelmia. Pelkkään sen tarkasteluun ei voi täysin luottaa.

4 Likertin asteikko

Tutkimuksissa yleinen tapa kerätä vastaajien mielipide on Likertin asteikko (Bishop ja Herron 2015), jota käytetään erityisesti sosiaalitieteissä, psykologiassa sekä kasvatustieteissä. Asteikon kehitti Rensis Likert 1932. Sen käyttö on usein käyttökelpoisempaa kuin muiden mittaustyökalujen. Keskiössä on asteikon portaiden lukumäärän valitseminen, koska määrä vaikuttaa keskiarvoon, varianssiin sekä jakauman muodostumiseen. Portaiden määrä on myös tutkimuksen luotettavuuden ja pätevyyden keskiössä. Riippuu tutkijasta ja tutkimuskohteesta, kuinka monta porrasta asteikkoon halutaan. Usein käytetään viisiportaista asteikkoa. (Ilhan ja Güler 2017)

4.1 Ongelmat

Likertin asteikon perusteella saatuja havaintoja pidetään usein järjestysasteikollisina muuttujina. Ne rikkovat kuitenkin normaalijakauman oletuksia, mikä luo ongelmia analyysin tekemiseen. Jos havainnot ovat järjestysasteikollisia muuttujia, parametrittomat menetelmät sopivat aineistoon paremmin. Kun taas intervalliasteikollisille muuttujille parametriset menetelmät ovat käyttökelpoisempia. Parametriset menetelmät ovat voimakkaampia parametrittomiin menetelmiin verrattuna, ja harhan suuruus kasvaa parametrisia menetelmiä käytettäessä. Tilastotieteellisessä tutkimuksessa keskeisenä tavoitteena on pätevien tulosten löytäminen. (Bishop ja Herron 2015) Tästä syystä on oleellista tietää, ovatko muuttujat järjestysasteikollisia vai intervalliasteikollisia muuttujia.

Ongelmallista on se, kuinka Likertin asteikosta saadut havainnot vaikuttavat tutkimuksen pätevyyteen ja luotettavuuteen. Likertin asteikko on ongelmallinen, sillä se on riippuvainen siitä, mitä kategoriaa vastaajat mieluiten vastaavat. (Ilhan ja Güler 2017) On havaittu, että vastaukset keskittyvät todennäköisimmin keskimmäisiin lukuihin. Asteikon ääriluvut jäävät usein harvinaisemmiksi vastauksiksi. Tämä puoltaa väitettä, että havainnot olisivat järjestysasteikollisia muuttujia. Portaiden välien epätasainen jakautuminen lisää analyysissa harhaa. Myös se vaikuttaa, mitkä sanat on valittu kuvaamaan asteikon numeroita. Nämä tekijät saattavat aiheuttaa jakauman vinoutumista, ja tällöin havainnot ovat todennäköisemmin järjestysasteikollisia muuttujia. (Bishop ja Herron 2015) Likertin asteikosta saadut havainnot voitaisiin näin ollen olettaa järjestysasteikollisiksi muuttujiksi, mutta tämä ei kuitenkaan ole analyysin kannalta järkevää, koska järjestysasteikolliset muuttujat rikkovat normaalijakauman oletuksia.

5 Menetelmien sovellus aineistoon

Koronaepidemian aikana mielenkiinnon kohteeksi on noussut epidemian yhteys huoliin. On havaittu, että koronaepidemia aiheuttaa huomattavia riskejä mielenterveydelle ja psykososiaalisen tuen lisääntymiselle. (Kilkku 2020) Tässä tutkielmassa tutkitaan, kuinka koronaepidemiaan liittyvät huolet ovat olleet yhteydessä harrastuksiin, luontokokemuksiin ja terveydentilaan epidemian aikana keväällä 2020.

5.1 Aineisto ja menetelmät

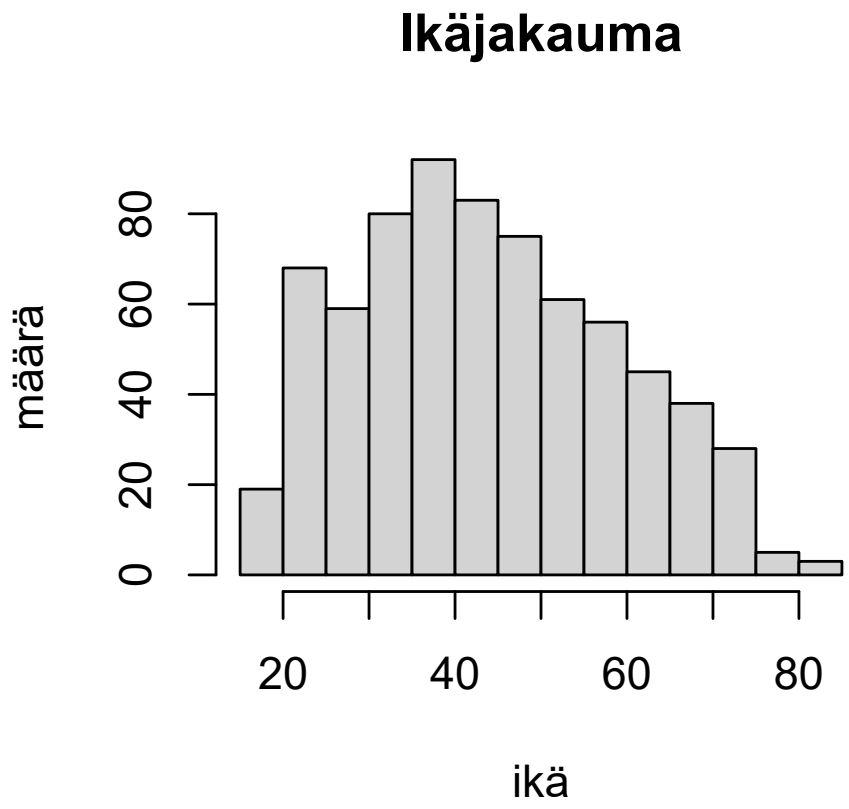
Aineistona käytetään Tampereen ja Jyväskylän yliopistojen psykologian oppiaineissa kerättyä aineistoa, jolla on kartoitettu luonnossa ulkoilua ja sen merkitystä epidemian aikana keväällä 2020 (Korpela, Salonen ja Hyvönen 2020). Tässä tutkimuksessa tutkitaan, miten epidemiaan liittyvät huolet ovat olleet yhteydessä harrastuksiin, luontokokemuksiin ja terveydentilaan. Tutkitaan myös, ovatko yhteydet muuttuneet ajan kuluessa. Erityisesti ollaan kiinnostuneita, onko huolien yhteys luontokokemuksiin muuttunut ajan kuluessa. Tutkimuksessa käytetään R-ohjelmistoa menetelmien soveltamiseen kyseiseen aineistoon.

Kyselyyn on vastannut 714 henkilöä kevään 2020 aikana. Aineistossa on puuttuvaa tietoa, jota joudutaan käsittelemään analyysin onnistumisen kannalta. Päädytään poistamaan kaikki tyhjiä kohtia sisältävät rivit. Poistettujen rivien jälkeen havaintoja jää 616. Puuttuvalla tiedolla voi olla vaikutusta saatuihin tuloksiin ja lopulliseen malliin. Analyysissa oletetaan kuitenkin puuttuvan tiedon olevan satunnaista, ja se ei näin ollen vaikuta merkitsevästi analyysiin lopputulokseen. Vastaukset on annettu viisiportaisen Likertin asteikon avulla. Tässä tutkimuksessa taustamuuttujina on henkilön sukupuoli, ikä ja koulutustausta. Sukupuolivaihtoehdot on kategoriset vaihtoehdot mies, nainen, muu tai ei halua kertoa. Taulukosta 5.1 havaitaan, että vastaajien sukupuolet ovat epätasaisesti jakautuneet. Vastaajista 607 on naisia, 105 miehiä, yksi muu kuin nainen tai mies ja yksi ei ole halunnut kertoa sukupuoltaan.

Taulukko 5.1. Sukupuolten jakautuminen aineistossa.

Sukupuoli	n
Nainen	607
Mies	105
Muu	1
Ei halua kertoa	1

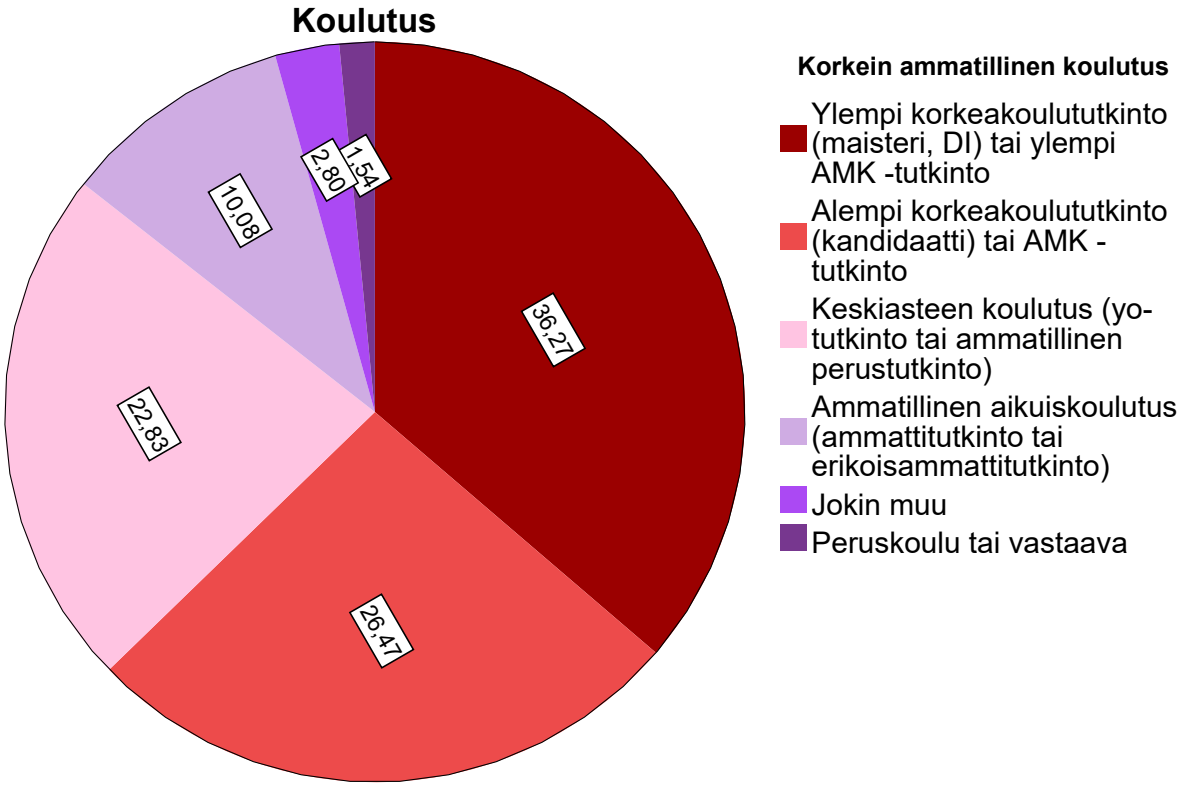
Ikä on vastattu numeroina ja sitä käsitellään jatkuvana muuttujana. Kuvasta 5.1 havaitaan, että ikäjakauma on likimain normaalisti jakautunut, jonka huippu on 35–40-vuotiaiden keskuudessa. Havaitaan, että suurin osa vastaajista on ollut keski-ikäisiä.



Kuva 5.1. Histogrammi ikäjakaumasta.

Koulutustausta on vastattu numerona sen mukaan, mikä on ollut vastaajan viimeisin tutkinto. Numero yksi vastasaa peruskoulua, kaksi toisen asteen koulutusta, kolme ammatillista aikuiskoulutusta, neljä alempaa korkeakoulututkintoa ja viisi ylempää korkeakoulututkintoa. Kuusi on jokin muu tutkinto, johon on saanut antaa tarkentavan vastauksen. Tarkentavista vastauksista käy ilmi, että suurin osa on tohtorin tutkinnon omaavia, eli kyseessä on korkein mahdollinen tutkinto. Yksi vastaajista on insinööri ja tämän vastaajan oletetaan kuuluvan ryhmään neljä, ammattikorkeakoulun käyneet. Koulutustausta käsitellään analyysissa jatkuvana muuttujana. Kuvan 5.2 ympyrädiagrammista havaitaan, että koulutustausta on epätasaisesti jakautunut. Suurin osa vastaajista (36.3 %) on ollut ylemmän korkeakoulututkinnon omaavia, 26.5 % on alemman korkeakoulututkinnon omaavia ja 22.9 % toisen asteen käyneitä. Kolme suurinta ryhmää ovat tasaisemmin jakautuneita kuin loput ryhmät. Ammatillisen aikuiskoulutuksen on käynyt 10.1 % vastaajista. Jonkin muun tutkinnon käyneitä on ollut 2.8 %. Ja pelkästään peruskoulun käyneitä on ollut vain 1.5 %. Havaitaan, että taustamuuttujat ovat epätasaisesti jakautuneet. Se ei kuitenkaan oleellisesti vaikuta analyysiin, koska taustamuuttujat käsitellään vakioina analyysissa. Kyselyssä on ollut kuusi kysymystä liittyen huoliin, 12 harrastuksiin, 13 luontokokemuksiin ja yksi terveydentilaan.

Pääkomponenttianalyysin perusteella havaitaan, että huolien kaksi ensimmäistä

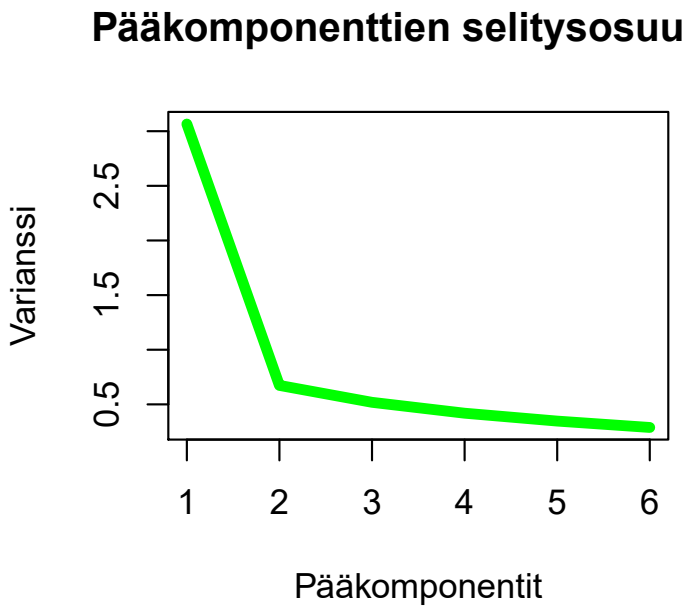


Kuva 5.2. Ympyrädiagrammi koulutuksen jakautumisesta.

pääkomponenttia selittävät vaihtelusta yli 70 %. Kuva 5.3 vahvistaa saatua tulosta. Siitä havaitaan, että ensimmäinen pääkomponentti selittää huomattavan osuuden varianssista. Ja kaksi ensimmäistä pääkomponenttia selittävät yhdessä merkitsevän osuuden varianssista. Taulukosta 5.2 havaitaan, että ensimmäisen pääkomponentin muuttujat ovat suurin piirtein yhtä latautuneita. Tästä syystä jatkoanalyysissa voidaan käyttää huolista summamuuttujaa. Huolien summamuuttujan jakauman huomataan olevan vinoutunut. Huomataan, että logaritmoitu summamuuttuja noudattaa paremmin normaalijakaumaa ja tätä käytetään jatkossa vasteena. Havaitaan, että luontokokemuksiin ja harrastuksiin liittyvissä kysymyksissä pääkomponenteista 5–6 ensimmäistä selittää 70 % vaihtelusta. Tällöin summamuuttujia ei voda hyödyntää näissä muuttujissa. Poistovalinnan avulla etsitään näistä muuttujista tilastollisesti merkitsevät muuttujat. Kaikkien merkitsevien muuttujien kesken tehdään yksi regressiomalli, jonka muuttujien merkitsevyyttä tutkitaan vielä uudestaan poistovalinnan avulla.

5.2 Tutkimustulokset ja niiden tulkinta

Lopulliseen malliin saadaan kolme merkitsevää muuttujaa harrastuksiin liittyen, yksi terveydentilaan ja neljä luontokokemuksiin. Taustamuuttujat ikä, sukupuoli ja koulutus käsitellään vakioina ja ne kulkevat läpi poistovalinnan muuttumattomina. Havaitaan, että mallin selitysaste R^2 on 0.44 ja muokattu selitysaste R^2_m on 0.43. Selitysaste on suhteellisen hyvä, ja valitut muuttujat selittävät hyvin vastetta.



Kuva 5.3. Pääkomponenttien selitysosuudet.

Taulukko 5.2. Ensimmäisen pääkomponentin muuttujien lataukset.

Muuttujat	Lataukset
Pelkotiloja	-0.42
Vaikeuksia saada unta tai keskeytynyttä unta	-0.29
Vahvaa epätoivoa	-0.42
Toimintakyvyn laskua	-0.40
Ärtyneisyyttä ja hermostuneisuutta	-0.48
Huolta	-0.42

Taulukosta 5.3 havaitaan lopullisen regressiomallin merkitsevät muuttujat, niiden estimaatit sekä p-arvot. Taulukosta havaitaan, että estimaatit ovat pieniä, mutta tilastollisesti merkitseviä 5 % merkitsevyystasolla. Havaitaan myös, että ohjelmien ja elokuvien katselulla, syömisellä ja puheluilla ystävien ja läheisten kanssa on tilastollisesti merkitsevä positiivinen yhteys huoliin. Tämä osoittaa, että henkilöt, joilla on enemmän huolia, käyttävät enemmän aikaa ohjelmien katseluun, syömiseen ja puheluihin läheisten kanssa. Huolet joko lisäävät kyseisiin harrastuksiin käytettyä aikaa tai harrastukset lisäävät huolia. Terveydentilaan liittyvä kysymys on ollut oma arvio terveydentilasta viisiportaisen Likertin asteikon avulla. Taulukosta havaitaan, että terveydentila on tilastollisesti merkitsevä muuttuja ja se heikentyy huolien kasvaessa. Tuloksesta voidaan tulkita, että huolet heikentävät terveyttä tai hyvä terveydentila vähentää huolia. Luontokokemuksiin liittyvistä kysymyksistä löytyy kaksi yksinään merkitsevää muuttujaa ”Luonnossa ulkoilun merkitys lisääntyy korona-aikana” ja ”Huolet (myös huoli koronasta) häiritsee luonnossa”. Taulukosta havaitaan, että nä-

mä muuttujat ovat positiivisessa yhteydessä huoliin. Henkilöillä, joilla ulkoilun merkitys on lisääntynyt korona-aikana, on ollut enemmän huolia. Huolet, myös huoli koronasta, häiritsevät luonnossa sitä enemmän, mitä enemmän on huolia.

Tutkitaan, miten huolet ovat muuttuneet ajan kuluessa keväällä 2020. Havaitaan, että aika ei ole yksinään merkitsevä muuttuja, eli ajan kuluessa huolien määrä ei ole tilastollisesti merkitsevästi muuttunut. Ollaan erityisesti kiinnostuneita ajan ja tilastollisesti merkitsevien luontokokemuksien yhdysvaikutuksesta. Havaitaan, että ajan yhdysvaikutus ”Luonnossa on tilaa olla ja hengittää” -muuttujan kanssa on tilastollisesti merkitsevä. Taulukosta 5.3 havaitaan, että yhteys on negatiivinen. Ajan edetessä huolia on vähemmän, kun henkilö kokee, että luonnossa on tilaa olla ja hengittää. Ajan yhdysvaikutus ”Koronaan liittyvä huoli tai ahdistus helpottuu luontokäyntien avulla” -muuttujan kanssa osoittautuu myös tilastollisesti merkitseväksi. Taulukosta havaitaan, että yhteys on positiivinen. Koronaepidemian edetessä huolet helpottuvat luontokäyntien avulla silloin, kun henkilöllä on enemmän huolia.

Taulukko 5.3. Regressiomallin merkitsevät muuttujat.

Muuttuja	Estimaatti	p-arvo
Ohjelmien / elokuvien katselu	0.042	< 0.001
Syöminen	0.047	< 0.001
Puhelut ystävien ja läheisten kanssa	0.027	0.005
Terveydentila	-0.060	< 0.001
Luonnossa ulkoilun merkitys lisääntynyt	0.030	0.013
Huolet häiritsevät luonnossa	0.099	< 0.001
Aika & luonnossa on tilaa hengittää ja olla	-0.000004	< 0.001
Aika & koronaan liittyvä huoli / ahdistus helpottuu luonnossa	0.000004	< 0.001

6 Johtopäätökset ja pohdinta

Aikaisemmassa tutkimuksissa havaittiin, että huolilla on yhteys toimintarajoitteisten henkilöiden elämään koronaepidemian aikana (Sainio et al. 2021). Ruotsissa havaittiin, että huoli koronapandemiasta on levinnyt laajalle. Vanhemmat henkilöt, naiset, pienituloiset sekä naimisissa tai avoliitossa olevat kokivat eniten huolta pandemiasta. Huoli liittyi etenkin sairastumisen pelkoon ja talouden heikkenemiseen. (Kulin et al. 2021) Aikaisemmissa tutkimuksissa korostui demografisten tekijöiden yhteys huoliin. Tämän tutkielman tavoitteena oli löytää huolien yhteyksiä terveydentilaan, harrastuksiin ja luontokokemuksiin.

Työn tavoitteessa onnistuttiin ja löydettiin huolien yhteyksiä harrastuksiin, luontokokemuksiin ja terveydentilaan. Huoliin, harrastuksiin ja luontokokemuksiin liittyi useampia kysymyksiä ja tutkimuksessa saatiin valittua lopulliseen malliin tilastollisesti merkitsevät muuttujat. Havaittiin, että keväällä 2020 huolien määrä ei ollut muuttunut tilastollisesti merkittävästi. Ajan yhdysvaikutus luontokokemuksien kanssa nosti esille kaksi tilastollisesti merkitsevää muuttujaa.

Havaittiin, että huolet ovat tilastollisesti merkitsevästi yhteydessä harrastuksiin, kuten ohjelmien ja elokuvien katseluun, syömiseen ja puheluihin. Henkilöt, joilla on huolia käyttävät enemmän aikaa kyseisiin harrastuksiin. Huolien todettiin olevan positiivisessa yhteydessä ”Luonnossa ulkoilun merkitys lisääntyy korona-aikana” -muuttujan, ”Huolet (myös huoli koronasta) häiritsee luonnossa” -muuttujan ja ajan yhdysvaikutuksen ”Koronaan liittyvä huoli tai ahdistus helpottuu luontokäyntien avulla” -muuttujan kanssa. Kyseiset muuttujat joko lisäävät huolia tai huolet lisäävät muuttujiin käytettyä aikaa. Todettiin myös, että huolet ovat negatiivisessa yhteydessä terveydentilan ja ajan yhdysvaikutuksen ”Luonnossa on tilaa hengittää ja olla” -muuttujan kanssa. Huolet joko vähentävät kyseisiä muuttujia tai muuttujat vähentävät huolia.

Työssä käytetyt menetelmät sopivat hyvin aineistoon. Pääkomponenttianalyysilla todettiin, että huolista voitiin käyttää analyysissa summamuuttujaa. Harrastuksien ja luontokokemuksien pääkomponenttianalyysilla ei löytynyt perustetta käyttää näistä muuttujista summamuuttujaa. Regressioanalyysin poistovalinnan avulla löydettiin kuitenkin merkitsevimmät muuttujia harrastuksista ja luontokokemuksista, jotka selittivät tilastollisesti merkitsevästi huolia. Mahdollisia muuttujia oli useita ja dimensio oli alkuperäisessä mallissa suuri. Pääkomponenttianalyysin ja regressioanalyysin avulla löydettiin tilastollisesti merkitsevät muuttujat ja saatiin pienennettyä dimensiota 31 muuttujasta kahdeksaan muuttujaan, jos taustamuuttujia, kuten ikää, sukupuolta ja koulutustaustaa ei oteta huomioon.

Tulevaisuudessa olisi mielenkiintoista perehtyä lisää siihen, korostuuko saaduissa tuloksissa demografiset tekijät, kuten aikaisemmin mainitussa Ruotsissa tehdyssä tutkimuksessa. Korostuuko joissakin tilastollisesti merkitsevissä muuttujissa erilaiset demografiset tekijät enemmän kuin toisissa muuttujissa. Korostuuko esimerkiksi sukupuolierot harrastuksien yhteydestä huoliin.

Koronaepidemian pitkäaikaiset vaikutukset näkyvät vasta tulevaisuudessa (Kilkku 2020). Tästä syystä tulevaisuudessa olisi mielenkiintoista tutkia pidemmällä aika-

välillä koronaepidemiaan liittyvien huolien yhteyksiä harrastuksiin, luontokokemuksiin sekä terveydentilaan. Tutkielmassa tutkittiin yhteyksiä keväällä 2020 kerättyyn aineistoon. Aikaväli oli suhteellisen lyhyt, eikä aikamuuttuja ollut tilastollisesti merkitsevä selittämään huolia. Jatkotutkimuksena voitaisiin tutkia, onko huolilla nähtävissä pitkäaikaisia vaikutuksia.

Lähteet

Arnout, B. et al. (2020). ”The Effects of Corona Virus (COVID-19) Outbreak on the Individuals' Mental Health and on the Decision Makers: A Comparative Epidemiological Study”. *International Journal of Medical Research Health Sciences*, s. 26–47.

Bishop, P. ja R. Herron (2015). ”Use and Misuse of the Likert Item Responses and Other Ordinal Measures”. *International journal of exercise science*, s. 297–302.

Draper, N. ja H. Smith (1998). *Applied regression analysis*. 3. painos. Wiley Series in Probability and Stataistics. John Wiley Sons, Incorporated, s. 243–247. URL: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=1775203 (viitattu 01. 02. 2022).

Ilhan, M. ja N. Güler (2017). ”The Number of Response Categories and the Reverse Scored Item Problem in Likert-Type Scales: A Study with the Rasch Model”. *Journal of Measurement and Evaluation in Education and Psychology*, s. 321–343.

Kilkku, N. (2020). ”Koronapandemia nosti esiin mielenterveyden merkityksen yhteiskunnalle”. *Alusta!*

Kim, H.-J. (2021). ”DATA.STAT.460 Regression Analysis”. *Luentomoniste, Tampereen yliopisto*, s. 3–60.

Korpela, K., K. Salonen ja K. Hyvönen (2020). ”Arki ja ulkoilu koronaepidemian aikana”. *Tutkimusaineisto*.

Kulin, J. et al. (2021). ”Oro över coronapandemin i det svenska samhället”. *Sociologisk forskning*, s. 77–102.

Lin, M., H. Lucas ja G. Shmueli (2013). ”Too Big to Fail: Large Samples and the p-Value Problem”. *Information System Research*, s. 906–917.

Nummi, T. (2021). ”Multivariate Analysis”. *Luentomoniste, Tampereen yliopisto*, s. 40–60.

Nyblom, J. (2015). ”Yleistetyt lineaariset mallit”. *luentomoniste, Jyväskylän yliopisto*, s. 23–24.

Sahay, A. (2016). *Applied Regression and Modeling. A Computer Integrated Approach*. Business Expert Press. URL: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=4560113 (viitattu 20. 02. 2022).

Sainio, P. et al. (2021). ”Multivariate Analysis”. *Sosiaalilääketieteellinen aikakausilehti*, s. 235–252.

Suryanarayana, T. ja P. Mistry (2016). *Principal Component Regression for Crop Yield Estimation. Principal Component Analysis in Transfer Function*. SpringerBriefs in Applied Sciences and Technology. Springer. URL: https://link-springer-com.libproxy.tuni.fi/book/10.1007/978-981-10-0663-0 (viitattu 11. 02. 2022).