kukaheroko

PISA SCORE PREDICTION

KUVIOT

Kuvio 1 Matematiikan PISA-kokeen pistemäärien jakautuminen aineiston muuttujien joukossa 8
Kuvio 2 Ennustemallin jäännökset sovitteiden suhteen 13

TAULUKOT

TAULUKKO 1 Aineiston muuttujat tunnuslukuina: muuttujan nimi, havaintojen lukumäärä, keskiarvo, keskihajonta, mediaani, minimi, maksimi sekä dikotomisten muuttujien ja luokittelevien muuttujien luokkafrekvenssit .. 7
TAULUKKO 2 Sovitettu ennustetmalli PISA-kokeen matematiikan pistemäärälle 12

SISÄLLYS

KUVIOT 2

TAULUKOT 2

SISÄLLYS 3

1 HARJOITUSTYÖN AIHE: PISA-KOKEEN MATEMATIIKAN PISTEMÄÄRÄN ENNUSTAMINEN 4
- 1.1 Aineiston kuvailu 6
- 1.2 Lineaarinen regressio 10
- 1.3 Tulokset 11
- 1.4 Johtopäätökset 14

LÄHTEET 17

LIITE 1 JÄÄNNÖSTARKASTELUA 18

1 Harjoitustyön aihe: PISA-kokeen matematiikan pistemäärän ennustaminen

PISA-tutkimus on kolmen vuoden välein järjestettävä kansainvälinen tutkimus, jonka tavoitteena on arvioida eri maiden koulutusjärjestelmien suorituskykyä testaamalla 15-vuotiaiden koululaisten tietoja ja taitoja matematiikassa, luonnontieteissä, lukemisessa ja ongelmanratkaisussa (OECD 2017b). Tutkimustulosten hyödyntämisellä pyritään muun muassa tasa-arvoisten oppimismahdollisuuksien edistämiseen ja opetuksen kehittämiseen niin järjestelmien kuin menetelmien osalta. Vuoden 2009 PISA-tutkimus painottui testaamaan oppilaiden osaamista lukemisessa, matematiikassa ja luonnontieteissä. (OECD 2017a.) Tämän harjoitustyön tarkoituksena on rakentaa ennustemalli PISA-kokeen matematiikan pistemäärälle. Pääongelmana on selvittää, mitkä tekijät ovat merkitseviä, kun halutaan ennustaa PISA-kokeen matematiikan pistemäärää ja miten ne voi yhdistää yhdeksi mahdollisimman toimivaksi ennustemalliksi. Aineistona toimii vuoden 2009 PISA-aineiston osa-aineisto sisältäen 500 satunnaisesti valitun suomalaisen 9-luokkalaisen matematiikan PISA-kokeen tulokset sekä useita taustamuuttujia.

Aineistossa mukana olevien taustamuuttujien vaikutusta oppilaiden oppimiseen on tutkittu runsaasti. Esimerkiksi perheen sosioekonomisen aseman on havaittu vaikuttavan positiivisesti oppilaan oppimistuloksiin useissa tutkimuksissa (Sirin 2005; Hampden-Thompson & Johnston 2006). Chiu ja Xihua (2008) keskittyivät tarkastelemaan matematiikan oppimistuloksia ja havaitsivat perheen sosioekonomisen aseman vaikuttavan niihin positiivisesti. Perheen sosioekonomista statusta mittaava muuttuja koostuu kolmesta komponentista — vanhempien koulutus, ammatillinen status ja varallisuus — jotka kaikki ovat muuttujan toiminnan kannalta olennaisia (Hauser & Huang 1997). Toisaalta esimerkiksi Chopra (1967) havaitsi vanhempien ammatillisen statuksen vaikuttavan oppilaitten oppimistuloksiin tutkiessaan sen vaikutuksia erillisenä muuttujana.

Oppilaan motivaatio on sekin varsin paljon tutkimuksellista huomiota saanut muuttuja. Lin, Hung, Lin, Lin ja Lin (2009) havaitsivat oppilaiden tavoitteellisuuden vaikuttavan positiivisesti matematiikan oppimistuloksiin. Chiu ja

Xihua (2008) havaitsivat nimenomaisesti sisäisen motivaation oppimiselle olevan tärkeämpää kuin välineellisen motivaation. oppilaiden matematiikan oppimistulokset olivat parempia, jos he olivat kiinnostuneita matematiikasta ja sen oppimisesta (Chiu & Xihua 2009). Myös Singh, Granville ja Dika (2002) havaitsivat oppilaiden motivaation, asenteen ja kiinnostuksen parantavan oppimistuloksia matematiikassa.

Mahdolliset erot sukupuolten välillä eivät ole yhtä selkeitä. Esimerkiksi Hyden, Fennema ja Lamon (1990) eivät meta-analyysillaan löytäneet merkittävää eroa matematiikan hallinnassa sukupuolten välillä. Tytöt olivat maltillisesti parempia matematiikan taidoissa kuin pojat varsinkin ensimmäisellä asteella. Mielenkiintoista on kuitenkin havainto, jonka mukaan toisen asteen aikana poikien matemaattiset ongelmanratkaisutaidot nousevat tyttöjä paremmiksi. (Hyde ym. 1990.) Kimball (1989) ehdottaa ja esittelee näyttöä, että tytöt saavat parempia matematiikan arvosanoja tavallisista matematiikan kokeista, mutta asetelma kääntyy poikien eduksi esimerkiksi standardoiduissa kokeissa, kuten PISA-kokeet.

Tulokset koulun sijainnin vaikutuksesta oppilaiden oppimistuloksiin ovat ristiriitaisia. Silti suurin osa tutkimuksista ei ole löytänyt tilastollisesti merkitsevää eroa kaupunkikoulujen ja maaseutukoulujen opiskelijoiden oppimistulosten välillä (Fan & Chen 1998; Alspaugh 1992). Koulualue voi vaikuttaa oppimistuloksiin etunenässä resurssien allokoinnilla. Esimerkiksi Wenglinskyn (1997) havaitsi, että sellaisten koulualueiden, jotka jakoivat kouluille käytettäväksi enemmän resursseja oppilasta kohden, kouluissa matematiikan oppimistulokset olivat parempia kuin sellaisissa koulualueiden kouluissa, joissa oppilasta kohden käytettävissä olevat resurssit oli allokoitu vähäisemmäksi. Johtopäätös oli, runsaammat resurssit pienensivät luokkakokoja ja loivat positiivisemman opiskeluympäristön, jotka paransivat oppilaiden oppimistuloksia (Wenglinsky 1997). Mikäli resursseilla todella on oppimistuloksia parantavia vaikutuksia, voi näiden olettaa olevan voimassa myös koulutasolla. Toisaalta sosioekonominen status voi näkyä myös koulualueen kautta, sillä eittämättä toisilla alueilla asuu keskimäärin koulutetumpaa väestöä kuin toisilla, mikä saattaa näkyä alueellisina eroina oppimistuloksissa.

Edellä esitetyn suppean kirjallisuuskatsauksen valossa voidaan esittää aineiston muuttujia koskevat hypoteesit:

- H_1: Matematiikan viimeisin arvosana vaikuttaa positiivisesti PISA-kokeen matematiikan pistemäärän ennusteeseen.
- H_2: Äidinkielen viimeisin arvosana vaikuttaa positiivisesti PISA-kokeen matematiikan pistemäärän ennusteeseen.
- H_3: Vanhempien ammatillinen status vaikuttaa positiivisesti PISA-kokeen matematiikan pistemäärän ennusteeseen.
- H_4: Perheen sosioekonominen status vaikuttaa positiivisesti PISA-kokeen matematiikan pistemäärän ennusteeseen.
- H_5: Poikien PISA-kokeen matematiikan pistemäärän ennuste on korkeampi kuin tyttöjen.

- H_6: Opiskelijan motivaatio vaikuttaa positiivisesti PISA-kokeen matematiikan pistemääränsä ennusteeseen.
- H_7: Koulu vaikuttaa PISA-kokeen matematiikan pistemäärän ennusteeseen.
- H_8: Koulun sijainti ei vaikuta PISA-kokeen matematiikan pistemäärän ennusteeseen.
- H_9: Koulualue vaikuttaa PISA-kokeen matematiikan pistemäärän ennusteeseen.

Harjoitustyöraportti jatkuu aineiston kuvailulla, jonka jälkeen esitellään käytettävät menetelmät, mistä jatketaan tulosten esittelyyn ja edelleen lopuksi johtopäätösten esittelyyn.

1.1 Aineiston kuvailu

Harjoitustyön aineistona toimii vuoden 2009 PISA-tutkimuksen aineiston osa-aineisto sisältäen 500 satunnaisesti valitun suomalaisen 9-luokkalaisen matematiikan PISA-kokeen tulokset sekä yhdeksän taustamuuttujaa. Aineistossa ei esiinny lainkaan puuttuvia havaintoja, vaan kaikilta aineistoon kuuluvalta 500 oppilaalta on saatu kerättyä PISA-kokeen matematiikan pistemäärä ja tiedot kaikkien taustamuuttujien arvoista. Taustamuuttujat ovat koulun tunnus, sukupuoli, vanhempien ammatillinen status, perheen sosioekonominen status, koulun sijainti, koulualue, oppilaan motivaation taso, matematiikan viimeisin arvosana ja äidinkielen viimeisin arvosana. Aineiston kuvaamisen tueksi taulukkoon 1 on koottu kaikkien muuttujien tunnusluvut.

Harjoitustyön vastemuuttujana toimivan PISA-kokeen matematiikan pistemäärän pienin arvo aineistossa on 312.1 pistettä ja suurin 783.5 pistettä. Aineistosta laskettu keskiarvo PISA-kokeen matematiikan pistemäärälle on 542.4 pistettä keskihajonnan ollessa samalla 75.3 pistettä. Mediaani 547.2 pistettä on hyvin lähellä PISA-kokeen matematiikan pistemäärän keskiarvoa. Matematiikan viimeisimmän arvosanan keskiarvo on 7.7 ja keskihajonta 1.4. Mediaani asettuu sen sijaan tasan 8.0:aan. Aineiston pienin arvo matematiikan viimeisimmälle arvosanalle on 4.0 ja suurin puolestaan 10.0. Äidinkielen viimeisimmän arvosanan tunnusluvut poikkeavat hiukan matematiikan vastaavista. Esimerkiksi keskiarvo asettuu korkeammaksi 7.9:ään ja keskihajonta pienemmäksi 1.1:een. Mediaani on kuitenkin sama 8.0, kuten myös pienin ja suurin arvo jotka ovat vastaavasti 4.0 ja 10.0.

Vanhempien ammatillinen status on sitä parempi, mitä korkeampi indeksi-muuttujan arvo on. Muuttuja voi saada arvoja 16 ja 99 väliltä ja ne muodostavat myös aineiston vaihteluvälin ala- ja ylärajan. Vanhempien ammatillisen statusindeksin keskiarvo aineistossa on 54.9 ja keskihajonta puolestaan 16.1. Mediaani 53.0 on hyvin lähellä keskiarvoa. Perheen sosioekonominen status on yhdistetty muuttuja vanhempien ammatista, koulutuksesta ja varallisuudesta. Mitä korkeamman arvon muuttuja saa, sitä parempi on perheen sosioekonomi-

nen status. Sen pienin arvo aineistossa on -2.9 ja suurin arvo puolestaan 2.3. Muuttujan keskiarvo aineistossa asettuu 0.3:een, keskihajonta 0.8:aan ja mediaani 0.4:ään, joka on hyvin lähellä keskiarvoa. Koulun tunnus on luku väliltä 1001 ja 9001. Aineistossa on edustettuna kaikkiaan 135 eri koulua. Yleisin yhdestä koulusta mukana olevien oppilaiden lukumäärä on 3. Toisaalta suurin yhdestä koulusta mukana olevien oppilaiden lukumäärä on 9 ja pienin 1.

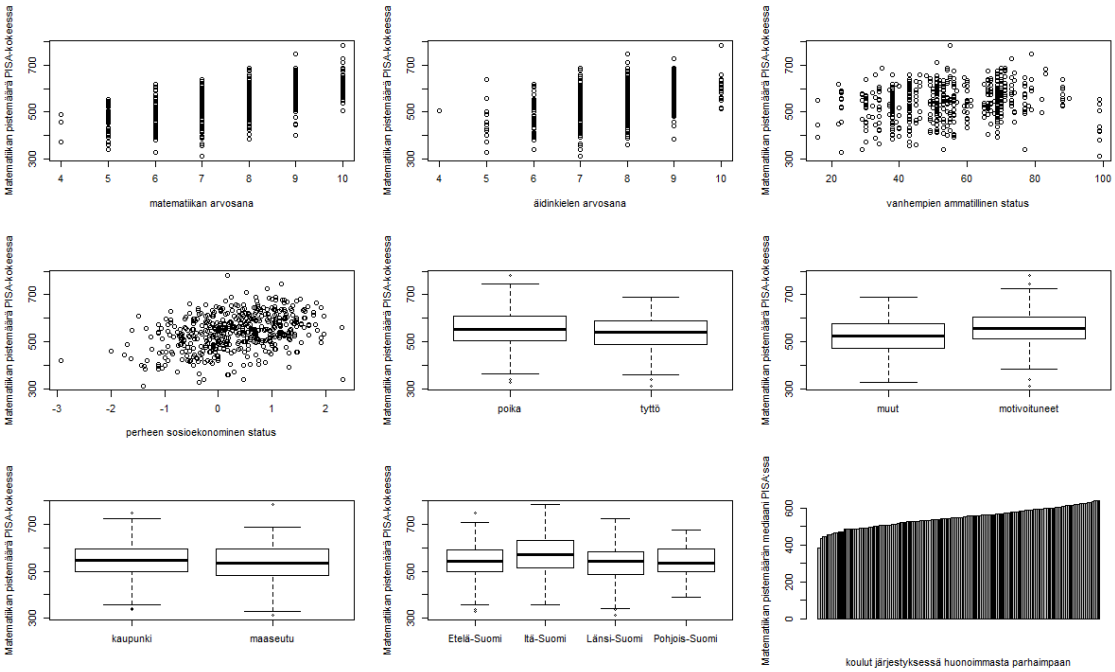
TAULUKKO 1 Aineiston muuttujat tunnuslukuina: muuttujan nimi, havaintojen lukumäärä, keskiarvo, keskihajonta, mediaani, minimi, maksimi sekä dikotomisten muuttujien ja luokittelevien muuttujien luokkafrekvenssit

Muuttuja	**n**	**ka**	**sd**	**md**	**min**	**max**
PISA-kokeen matematiikan pistemäärä	500	542.4	75.3	547.2	312.1	783.5
Matematiikan viimeisin arvosana	500	7.6	1.4	8.0	4.0	10.0
Äidinkielen viimeisin arvosana	500	7.9	1.1	8.0	4.0	10.0
Vanhempien ammatillinen status (indeksi välillä 16-99)	500	54.9	16.1	53.0	16.0	99.0
Perheen sosioekonominen status (yhdistetty muuttuja vanhempien ammatista, koulutuksesta ja varallisuudesta)	500	0.3	0.8	0.4	– 2.9	2.3
Koulun tunnus	500	–	–	–	1001	9001
	n	**poikia**	**tyttöjä**			
Sukupuoli	500	232 (46.4 %)	268 (53.6 %)			
	n	**motivoituneita**	**muita**			
Opiskelijan motivaatio	500	339 (67.8 %)	161 (32.2 %)			
	n	**kaupunki**	**maaseutu**			
Koulun sijainti	500	395 (79 %)	105 (21 %)			
	n	**Etelä-Suomi**	**Itä-Suomi**	**Länsi-Suomi**	**Pohjois-Suomi**	
Koulualue	500	233 (46.6 %)	73 (14.6 %)	129 (25.8 %)	65 (13 %)	

Sukupuolet ovat edustettuina aineistossa melko tasaisesti, sillä kaikkiaan 500 mukana olevasta oppilaasta 232 on poikia ja 268 tyttöjä. Vastaavat prosentuaaliset osuudet ovat 46.4 % ja 53.6 %. Aineistossa on kaikkiaan 339 itsensä motivoituneeksi kokenutta opiskelijaa, mikä vastaa 67.8 % osuutta kaikista aineiston opiskelijoista. Muita kuin motivoituneita opiskelijoita on 161, mikä vastaa 32.2 % osuutta kaikista aineiston opiskelijoista. Itsensä motivoituneiksi kokeneet opiskelijat muodostavat siten huomattavan enemmistön aineistossa. Koulun sijainti muuttujassa on tallennettuna tieto siitä, sijaitseeko oppilaan koulu kaupunki-

alueella vai maaseuduksi luokiteltavalla alueella. Huomattava enemmistö aineiston opiskelijoista käy kaupungeissa sijaitsevissa kouluissa, kaikkiaan 395, mikä vastaa 79 % osuutta kaikista aineiston opiskelijoista, kun taas maaseudulla sijaitsevissa kouluissa käy 105 opiskelijaa, mikä vastaa 21 % kaikista aineiston opiskelijoista.

Koulualue muuttujassa on tallennettuna tieto alueesta jolla opiskelijan koulu sijaitsee. Mahdollisia sijainteja ovat Etelä-, Itä-, Länsi- ja Pohjois-Suomi. Aineistossa on selvästi eniten opiskelijoita Etelä-Suomen kouluista, sillä kaikista 500 opiskelijasta 233 edustaa Etelä-Suomen kouluja, mikä vastaa 46.6 %:n kokonaisosuutta. Seuraavaksi eniten opiskelijoita tulee Länsi-Suomen kouluista, kaikkiaan 129, mikä vastaa 25.8 %:n kokonaisosuutta. Itä-Suomen kouluista tulee 73 opiskelijaa, mikä vastaa 14,6 %:n kokonaisosuutta ja Pohjois-Suomen kouluista 65 opiskelijaa, mikä vastaa 13 %:n kokonaisosuutta. Itä-Suomen ja Pohjois-Suomen kouluista tulevia opiskelijoita on siis aineistossa huomattavasti vähemmän kuin Etelä- tai Länsi-Suomesta tulevia opiskelijoita.



Kuvio 1 Matematiikan PISA-kokeen pistemäärien jakautuminen aineiston muuttujien joukossa

Kuvio 1 koostuu yhdeksästä pienkuviosta, joissa esitetään graafisesti PISA-kokeen matematiikan pistemäärien käyttäytyminen erikseen kunkin taustamuuttujan osalta. Kuvion 1 vasemmassa ylälaidassa on esitetty PISA-kokeen matematiikan pistemäärän ja matematiikan viimeisimmän arvosanan hajontakuvio. Kuviosta on selkeästi nähtävissä muuttujien välinen positiivinen lineaarinen yhteys. Tämä käy ilmi myös muuttujien välisen korrelaatiokertoimen arvosta 0.62. Ylärivissä keskellä on esitetty PISA-kokeen matematiikan pistemäärän ja äidinkielen viimeisimmän arvosanan hajontakuvio. Myös näiden kahden muuttujan välillä on nähtävissä melko selkeä positiivinen lineaarinen yhteys.

Korrelaatiokerroin 0.48 viestii sekin näiden muuttujien positiivisesta yhteydestä. Ylärivillä oikealla on esitetty PISA-kokeen matematiikan pistemäärän ja vanhempien ammatillisen statuksen hajontakuvio. Nyt yhteys ei ole lähellekään niin selvä kuin matematiikan arvosanan ja äidinkielen arvosanan tapauksessa ja vaikuttaa jopa hiukan epälineaariselta. Korrelaatiokerroin jää vaatimattomaksi 0.17 eikä viittaa kovin vahvaan yhteyteen.

Kuvion 1 toisen rivin vasemmalla puolella on esitetty PISA-kokeen matematiikan pistemäärän ja sosioekonomisen statuksen hajontakuvio. Kuviosta on jälleen nähtävissä, että muuttujien välillä on jonkin verran positiivista lineaarista yhteyttä. Myös korrelaatiokerroin 0.36 viittaisi positiivisen yhteyden olemassaoloon. Vanhempien koulutus ja varallisuus vaikuttaisivat tuovan lisäinformaatiota yhdistettynä ammatillisen statuksen kanssa verrattuna pelkkään ammatillista statusta mittaavaan muuttujaan. Keskimmäisen rivin keskimmäisessä kuviossa on esitetty PISA-kokeen matematiikan pistemäärän jakautuminen erikseen tytöillä ja pojilla. Kuviosta on nähtävissä, että poikien pistemäärien mediaani on hiukan korkeammalla kuin tyttöjen. Sen sijaan tyttöjen pistemäärien hajonta vaikuttaisi aavistuksen pienemmältä kuin poikien. Keskimmäisen rivin oikeassa laidassa on esitetty PISA-kokeen matematiikan pistemäärän jakautuminen erikseen itsensä motivoituneiksi tunteneiden ja muiden opiskelijoiden joukossa. Motivoituneiden opiskelijoiden pistemäärän mediaani on nyt selvästi korkeammalla kuin muiden opiskelijoiden. Pistemäärien hajonta vaikuttaa samansuuruiselta molemmissa ryhmissä.

Kuvion 1 alimman rivin vasemmalla puolella on esitetty PISA-kokeen matematiikan pistemäärän jakautuminen erikseen opiskelijoilla, jotka käyvät kaupunkialueella sijaitsevaa koulua ja opiskelijoilla, jotka käyvät maaseudulla sijaitsevaa koulua. Kuviosta on nähtävissä, että kaupunkikoulujen opiskelijoiden pistemäärien mediaani on aavistuksen korkeampi kuin maaseutukoulujen opiskelijoilla joiden ryhmässä pistemäärien hajonta sen sijaan on suurempaa. Alimman rivin keskellä on esitetty PISA-kokeen matematiikan pistemäärän jakautuminen erikseen eri koualualueilta tuleville opiskelijoille. Itä-Suomen kouluista tulevien opiskelijoiden pistemäärien mediaani on selvästi korkein. Seuraavina tulevat Etelä-Suomen ja Länsi-Suomen koulujen opiskelijat hyvin lähekkäin toisiaan. Alin mediaani on Pohjois-Suomen koulujen opiskelijoilla, mutta ero ei ole kovin suuri verrattuna Etelä- tai Länsi-Suomen koulujen opiskelijoihin. Pistemäärien hajonta on toisaalta suurinta Itä-Suomen koulujen opiskelijoiden joukossa ja Pohjois-Suomen opiskelijoiden joukossa se on pienintä. Alalaidassa oikealla on esitetty kaikkien aineiston koulujen PISA-kokeen matematiikan pistemäärän mediaanit järjestyksessä heikoimmasta parhaimpaan. Kärkipään koulujen opiskelijoiden ja heikomman pään koulujen opiskelijoiden pistemäärässä on nähtävissä huomattava ero. Tähän eroon on kuitenkin suhtauduttava mitä korkeimmalla varauksella, sillä aineistossa tyypillisin opiskelijamäärä yhdestä koulusta on kolme, mutta joukossa on runsaasti myös kouluja, joista on mukana vain yksi tai kaksi opiskelijaa.

1.2 Lineaarinen regressio

Harjoitustyön tavoitteena on rakentaa ennustemalli PISA-kokeen matematiikan pistemäärälle. Menetelmäksi mallin rakentamiseen soveltuu lineaarinen regressioanalyysi, jolla voidaan kuvata numeerisen vastemuuttujan keskiarvojen vaihtelua prediktorimuuttujien lineaaristen funktioiden määrittämissä osapopulaatioissa. Lineaarisella regressiomallilla on prediktorimuuttujien lineaarifunktio, jolla voidaan ennustaa vastemuuttujan arvo annetuilla prediktorimuuttujien arvoilla. Regressiokertoimet voidaan tulkita ennusteiden eroiksi tai tiettyjen keskiarvojen eroiksi aineistossa. (Gelman & Hill 2007, 31.) Lineaarinen regressiomalli voidaan kirjoittaa muodossa:

$$(1) \qquad y_i = X_i\beta + \varepsilon_i$$

$$= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \text{ havainnoille } i = 1, \ldots, n,$$

missä virheet ε_i ovat riippumattomia sekä toisistaan että selittäjistä ja noudattavat normaalijakaumaa $N\sim(0, \sigma^2)$. Lineaarinen regressiomalli voidaan edelleen esittää muodossa:

$$(2) \qquad y \sim N(X\beta, \sigma^2, I),$$

missä y on n:n pituinen vektori, X on n x k –dimensioinen prediktorimatriisi, β on k:n pituinen sarakevektori, ja I on n x n –dimensioinen identiteettimatriisi. (Gelman & Hill 2007, 38—39.) Regressiokertoimien estimoinnissa käytetään pienimmän neliösumman menetelmää. Tarkoituksena on saada aikaiseksi mahdollisimman hyvin aineistoon sovittuva regressiosuora minimoimalla mallin jäännösten neliösumma regressiokertoimien suhteen. Mikäli lineaarisen regression oletukset jäännösten riippumattomuudesta, homoskedastisuudesta ja normaalijakaantuneisuudesta pitävät paikkansa, ovat pienimmän neliösumman estimaatit ja suurimman uskottavuuden estimaatit yhtä suuret. (Gelman & Hill 2007, 39—40.)

Lineaariseen regressiomalliin liittyy useita olettamuksia. Tärkein niistä on mallin ja toisaalta sen estimoinnissa käytetyn aineiston validius. Vastemuuttujan tulee mitata ilmiötä jota halutaan tutkia ja prediktorimuuttujien tulee mitata ilmiötä jolla vastetta halutaan ennustaa tai selittää. Aineiston tulee mahdollistaa asetettuun tutkimusongelmaan vastaaminen. (Gelman & Hill 2007, 45.) Sekä ilmiön selittämisen että ennustamisen kannalta on olennaista, että malliin saadaan mukaan kaikki relevantit prediktorit. Ennustemallin validius syntyy kyvystä ennustaa vasteen arvoja mahdollisimman tarkasti. Mallin tulee olla yleistettävissä tapauksiin, joihin sitä on tarkoitus soveltaa. (Gelman & Hill 2007, 45.) Toisin sanoen mallin ennustustarkkuuden tulee säilyä myös estimointiaineiston ulkopuolella sellaisten tapausten ennustamisessa, joihin sen on tutkimusongelmaa asetettaessa rajattu soveltuvan.

Toinen olettama on regressiomallin lineaarisuus, mikä vaatii sen, että regressiomallin deterministinen komponentti on erillisten prediktorimuuttujien lineaarinen funktio. Mallin muuttujille voidaan suorittaa erilaisia muunnoksia tilanteissa, joissa lineaarisuusolettama ei muuten pätisi. Lineaarisen regressioanalyysin tavoitteena löytää lineaarinen malli joka on riittävä approksimaatio tutkittavalle ilmiölle. (Gelman & Hill 2007, 46.) Kolmas olettama on mallin virheiden riippumattomuus (Gelman & Hill 2007, 46). Virheiden tulisi olla satunnaisia havainnosta toiseen, mitä ei esimerkiksi aikasarja-aineistoissa useinkaan voida olettaa. Esimerkiksi raaka-aineiden hinnat vaihtelevat päivästä toiseen, mutta ei voida järkevästi olettaa, ettei edellisen päivän hinta vaikuttaisi seuraavan päivän hintaa millään tavalla. Neljäs lineaarista regressioanalyysia koskeva olettama on homoskedastisuus eli virheiden varianssien vakioisuus. Homoskedastisuuden tilanteessa mallin jäännösten vaihtelu pysyy samansuuruisena kaikilla sovitteen arvoilla. (Gelman & Hill 2007, 46.) Viides olettama on virheiden normaalijakaantuneisuus. Erityisesti ennustamisen kannalta tulisi huomioida poikkeavien havaintojen vaikutus regressiokertoimiin, sillä ne saattavat jo vähissä määrin vääristää regressiokertoimia huomattavasti, mikä puolestaan heikentää mallin ennustetarkkuutta estimointiotoksen ulkopuolella. (Gelman & Hill 2007, 46.)

1.3 Tulokset

Harjoitustyön tuloksena rakennettuun lopulliseen ennustemalliin muuttujiksi valikoituivat matematiikan viimeisin arvosana, äidinkielen viimeisin arvosana, perheen sosioekonominen status ja sukupuoli. Mallin tulkinnan helpottamiseksi kolme ensimmäistä muuttujaa päätettiin keskistää. Matematiikan viimeisimmän arvosana ja äidinkielen viimeisin arvosana keskistettiin mediaanilla, joka aineistosta laskettuna on 8. Perheen sosioekonominen status keskistettiin keskiarvolla 0,3. Sovitettu malli voidaan kirjoittaa seuraavaan muotoon:

$$\text{(3) mpist} = 567.9 + 23.8(\text{matem} - 8) + 14.1(\text{aidink} - 8) + 15.4(\text{SES} - 0.3) + 26.4(\text{sukup(poika)}),$$

missä mpist on yhtä kuin ennustettu PISA-kokeen matematiikan pistemäärä, matem on yhtä kuin matematiikan viimeisin arvosana, aidink on yhtä kuin äidinkielen viimeisin arvosana, SES on yhtä kuin perheen sosioekonominen status ja sukup(poika) saa arvon 1, mikäli oppilas, jolle pistemäärää ennustetaan, on poika. Taulukkossa 2 on esitetty PISA-kokeen matematiikan pistemäärän ennustamiseksi estimoidun lineaarisen regressiomallin kerroinestimaatit, niiden keskivirheet, 95 %:n luottamusvälit ja t-testin tulokset merkitsevyysasteineen.

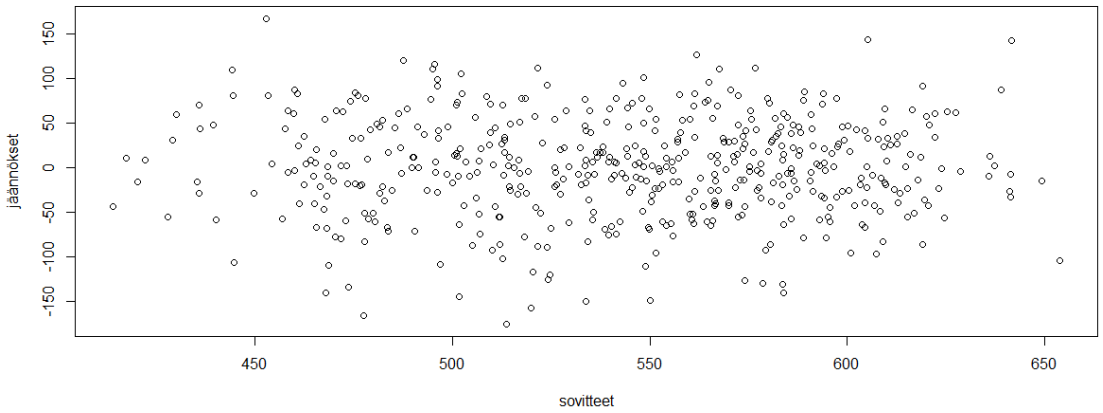
TAULUKKO 2 Sovitettu ennustetmalli PISA-kokeen matematiikan pistemäärälle

muuttuja	**β^**	**SE**	**95 % CI**	**t**	**p**
vakio	567.9	3.6	[534.4, 548.7]	149.3	0.001
matematiikan viimeisin arvosana - 8	23.8	2.3	[19.3, 28.3]	10.5	0.001
äidinkielen viimeisin arvosana – 8	14.1	3.1	[8.1, 20.1]	4.6	0.001
perheen sosioekonominen status – 0.3	15.4	3.3	[9.0, 21.9]	4.7	0.001
sukupuoli (poika)	26.4	5.4	[15.7, 37.0]	4.9	0.001
R^2	0.46				
F-testi	105.6		495 df		0.001
RSE	55.5		495 df		

Mallin mukaisesti naispuolinen opiskelija, jonka matematiikan ja äidinkielen viimeisimmät arvosanat ovat molemmat 8 ja jonka perheen sosioekonominen status vastaa 0.3:a, saa PISA-kokeen matematiikan pistemäärän ennusteekseen 567.9 pistettä. Muut muuttujat vakioituna, yhden numeron parannus matematiikan arvosanassa nostaa mallin mukaista PISA-kokeen matematiikan pistemäärän ennustetta 23.8 pisteellä. Yhden numeron parannus äidinkielen arvosanassa nostaa mallin mukaista PISA-kokeen matematiikan pistemäärän ennustetta 14.1 pisteellä, kun muut muuttujat oletetaan vakioiduiksi. Edelleen yhden yksikön parannus perheen sosioekonomisessa statuksessa nostaa mallin mukaista PISA-kokeen matematiikan pistemäärän ennustetta 15.4 pisteellä, kun muut muuttujat oletetaan vakioiduiksi. Poikien PISA-kokeen matematiikan pistemäärän ennuste on 26.4 pistettä korkeampi kuin tyttöjen. Kaikki muuttujat ovat tilastollisesti erittäin merkitseviä ja luottamusvälit ovat varsin kaukana nollasta. Näin muuttujien vaikutusten voi olettaa säilyvän myös estimointiotoksen ulkopuolella.

Mallin ennustuskyvyn vaikuttaa korkeintaan kohtalaiselta, sillä se kykenee selittämään noin 46 % PISA-kokeen matematiikan pistemäärän vaihtelusta. F-testin mukaan malli on tilastollisesti erittäin merkitsevä, mikä viestii mallin muuttujilla olevan huomattavasti enemmän kykyä selittää PISA-kokeen matematiikan pistemäärän vaihtelua verrattuna pelkästään vakiosta koostuvaan malliin. Jäännösten keskihajonnan perusteella mallin antama ennuste oppilaan PISA-kokeen matematiikan pistemäärälle poikkeaa havaitusta arvosta vähemmän kuin 55 pistettä kahdessa kolmesta tapauksesta, mitä voi ennustamisenkannalta pitää korkeintaan kohtalaisena tarkkuutena.

Kuviossa 2 on piirrettynä mallin jäännökset sovitteiden suhteen. Kuviosta ei ole erotettavissa erityistä systemaattista kuviota, minkä perusteella lineaarisuusolettama pysyy voimassa. Lineaarisuusolettamaa tukee aivan matalimpia ja korkeimpia vasteen arvoja lukuun ottamatta myös liitteen 1 alempi kuvio, jossa on piirrettynä vasteet mallin sovitteiden suhteen. Kuvion 2 perusteella jäännösten varianssi vaikuttaa pysyvän melko vakiona kaikilla sovitteen arvoilla, minkä perusteella virhevarianssien vakioisuuden olettama säilyy voimassa. Liitteen 1 ylemmästä kuviosta voidaan havaita, että myös jäännösten normaalisuusolettama vaikuttaa pysyvän voimassa.



Kuvio 2 Ennustemallin jäännökset sovitteiden suhteen

Tulokset puoltavat ensimmäisen, toisen, neljännen ja viidennen hypoteesien hyväksymistä. Matematiikan viimeisin arvosana, äidinkielen viimeisin arvosana, perheen sosioekonominen status vaikuttavat PISA-kokeen matematiikan pistemäärän ennusteeseen nostavasti tilastollisesti merkitsevällä tasolla. Poikien ennuste taas on korkeampi kuin tyttöjen. Kyseisten muuttujien ennustusvoima myös säilyi vahvana, kun ne yhdistettiin monimuuttujamalliin.

Aineiston kaikkien muuttujien ennustuskykyä testattiin ennen lopullisen mallin valintaa muodostamalla niistä yhden muuttujan regressiomallit. Koulua, koulun sijaintia ja koulun aluetta mitanneet muuttujat eivät olleet ennustuskykyisiä tilastollisesti merkitsevällä tasolla yksittäin eivätkä monimuuttujamallin osana. Yksi syy tälle voi olla aineiston suhteellisen pieni koko, jonka ongelmat korostuivat näiden muuttujien kohdalla. Kahdeksas hypoteesi voidaan varauksella hyväksyä. Sen sijaan seitsemäs ja yhdeksäs hypoteesi hylätään, mutta niiden osalta lisätarkastelut ovat paikallaan.

Oppilaan motivaatiota ja vanhempien ammatillista statusta mitanneet muuttujat osoittautuivat ennustuskykyisiksi yksittäin tarkasteltuna. Oppilaan tuntema motivaatio vaikutti nostavasti hänen ennustettuun PISA-kokeen matematiikan pistemääräänsä tilastollisesti merkitsevällä tasolla. Motivaatiomuuttuja kuitenkin menetti tilastollisen merkitsevyytensä, kun se yhdistettiin osaksi monimuuttujamallia. Todennäköinen syy tälle on se, että motivaatiomuuttujan informaatio sisältyy merkittäviltä osin matematiikan ja äidinkielen arvosanoja mittaaviin muuttujiin. Oppilaan motivaatiolla on kuitenkin selvästi nostava vaikutus PISA-kokeen matematiikan pistemäärään, minkä vuoksi kuudes hypoteesi voidaan varauksella hyväksyä.

Vanhempien ammatillista statusta mitannut muuttuja oli tilastollisesti merkitsevä prediktori sekä yksittäin että osana monimuuttujamallia. Muuttujan kerroin oli positiivinen, kun se muodosti itsenäisen mallin, mutta kerroin muuttui negatiiviseksi, kun se yhdistettiin osaksi monimuuttujamallia. Toisin sanoen vanhempien ammatillinen status vaikuttaisi itsenäisenä muuttujana nostavan ennustetta PISA-kokeen matematiikan pistemäärälle, mutta yhdistettynä mui-

den muuttujien informaation kanssa se vaikutus kääntyy negatiiviseksi. Muuttujan kertoimen epälooginen käyttäytyminen, sen suhteellinen pienuus ja vaatimaton selitysvoima olivat syyt, minkä vuoksi muuttuja päätettiin jättää pois lopullisesta ennustemallista. Muuttujalle suoritettiin myös esimerkiksi logistinen muunnos, mutta ennustekyky toimenpiteistä huolimatta parantunut. Vanhempien ammatillinen status on yksi komponentti perheen sosioekonomista statusta mittaavaa muuttujaa ja todennäköisesti suuri osa sen informaatiosta sisältyi jo tähän malliin otettuun muuttujaan. Kolmatta hypoteesia ei voida hyväksyä ilman lisätarkasteluja.

1.4 Johtopäätökset

Tässä harjoitustyössä tavoitteena oli rakentaa ennustemalli PISA-kokeen matematiikan pistemäärälle. Ongelmana oli löytää ennustuskykyiset muuttujat ja yhdistää ne yhteen ennustusmalliin. Harjoitustyön tuloksena syntyi vakion lisäksi kaikkiaan neljästä prediktorimuuttujasta koostuva lineaarinen regressiomalli PISA-kokeen matematiikan pistemäärän ennustamiseen. Malliin sisältyvät muuttujat ovat matematiikan viimeisin arvosana, äidinkielen viimeisin arvosana, perheen sosioekonominen status ja sukupuoli. Mallin sovittuu aineistoon vähintään kohtalaisen hyvin, kykenee selittämään noin 46 % PISA-kokeen matematiikan pistemäärän vaihtelusta ja on tilastollisesti erittäin merkitsevä.

Mallista voidaan johtaa useita johtopäätöksiä. Ensimmäinen näistä on, että matematiikan arvosanalla on selvästi positiivinen vaikutus ennustettuun PISA-kokeen matematiikan pistemäärään. Mitä parempi matematiikan arvosana, sitä korkeampi on ennuste PISA-kokeen matematiikan pistemäärälle. Koulujen matematiikan opetus luo siten hyvän pohjan PISA-kokeen matematiikan osuudelle. Myös äidinkielellä on selvästi ennustuskykyä PISA-kokeen matematiikan pistemäärälle. Vaikuttaisi siltä, että hyvä äidinkielen arvosana ennustaa hyvää tulosta PISA-kokeen matematiikan osuudesta. Sekä matematiikan arvosanan että äidinkielen arvosanan ennustuskykyä parantaa mitä todennäköisimmin se, että ne heijastavat subjektiosaamisen lisäksi informaatiota oppilaan motivaatiosta. Matematiikan arvosana heijastaa motivaatiota oppia matematiikkaa ja äidinkielen arvosana motivaatiota oppia yleensä. Tästä näkökulmasta tämän harjoitustyön tulokset ovat oppilaan motivaation selitys- ja ennustuskyvyn suhteen sopusoinnussa aiemman kirjallisuuden kanssa.

Perheen sosioekonominen status vaikuttaa positiivisesti PISA-kokeen matematiikan pistemäärän ennusteeseen. Toisin sanoen mitä parempi sosioekonominen asema perheellä on, sitä paremmat ennusteet sen lapset saavat PISA-kokeen matematiikan osuudelle. Tulokset perheen sosioekonomisen statuksen ennustuskyvystä ovat linjassa aiemman tutkimuksen kanssa. Sen sijaan vanhempien ammatillisen statuksen ennustuskyvystä ei saatu selkeää näyttöä, mikä on ristiriidassa aiempien tulosten kanssa. Johtopäätös voidaan tehdä sen suhteen, että mitä todennäköisimmin perheen sosioekonomista asemaa mittaava muuttuja sisältää suuren osan vanhempien ammatillista statusta mittaavan

muuttujan selitysvoimasta. Vanhempien ammatillisen statuksen itsenäisestä ennustuskyvyn selvittäminen vaatii lisätarkasteluja.

Sukupuolella vaikuttaisi todella olevan merkitystä, kun ennustetaan matematiikan oppimistuloksia. Tämän harjoitustyön tulosten perusteella 15-vuotiaiden poikien PISA-kokeen matematiikan pistemäärän ennuste on noin 26 pistettä korkeampi kuin samanikäisillä tytöillä. Harjoitustyön tulokset tuovat jatkoa niiden aiempien tutkimusten tuloksiin, joissa sukupuolella on löydetty olevan merkitystä matematiikan oppimistuloksiin. Jatkotutkimuksissa olisi hyvä pyrkiä selvittämään onko sukupuolten välillä todella eroa ja jos on, niin milloin tämä ero syntyy ja mistä syistä johtuen.

Tämän harjoitustyön tulosten perusteella koululla, sen sijainnilla tai alueella ei ole merkitystä ennustettaessa PISA-kokeen matematiikan pistemäärää. Tämä on osittain ristiriidassa aiemman tutkimuksen kanssa ja myös intuitiivisesti outoa, minkä vuoksi näiden muuttujien osalta suositellaan lisätarkasteluja. Yksi merkittävä tekijä, jonka vuoksi näiden muuttujien mahdollinen ennustuskyky ei tullut esille, oli harjoitustyön suppea aineisto. Esimerkiksi koulun vaikutuksen havaitseminen vaatisi huomattavasti enemmän havaintoja per koulu kuin mitä tässä harjoitustyössä oli käytettävissä. Sama pätee myös kahteen muuhun muuttujaan. Toisaalta koulun sijainnin merkityksettömyydestä matematiikan oppimistuloksiin voidaan jo tällä aineistolla tehdä varovainen johtopäätelmä.

Tämän harjoitustyön suurimmat rajoitukset liittyvät aineiston suppeuteen. Tulosten yleistämiseen tulee suhtautua suurella varauksella. Käytettävissä oli useita mielenkiintoisia muuttujia, mutta useita olennaisia ja ennustusvoimaisia muuttujia jäi tämän harjoitustyön ulkopuolelle. Harjoitustyössä estimoidun ennustusmallin selitysaste jäi korkeintaan kohtalaiseksi, minkä voidaan päätellä, että mallia voitaisiin kehittää huomattavasti täydentämällä sitä uusilla relevanteilla muuttujilla. Tulevaisuuden tutkimuksissa olisi tärkeää pyrkiä tunnistamaan näitä puuttuvia ennustuskykyisiä muuttujia esimerkiksi opetuksen kehittämisen tueksi. Ennustuskykyisiä muuttujia voisivat olla esimerkiksi oppilaan tuntemus terveys- ja vireystilastaan, fysiikan ja kemian arvosanat, liikunnan harrastamiseen käytetty aika, tuki- tai lisäopetuksen määrä, ohjelmointi-, shakki tai pokeriharrastus, oppilaan ajatus tulevasta ammatillisesta suuntautumisestaan ja vanhempien matemaattiset taidot.

Aineistosta ei haluttu lähteä poistamaan äärimmäisiä havaintoja, vaikka tällä olisi todennäköisesti saatu parannettua ennustemallin sovittuvuutta jonkin verran. Parantavan vaikutuksen arvioitiin olevan suhteessa liian pieni verrattuna mahdolliseen informaation menetykseen. Merkittävä rajoite harjoitustyön tulosten, etenkin ennustemallin laadun, yleistämiselle on testiotoksen puute. Mahdollisuus olisi ollut rajata osa estimointiotoksesta testaamisen tarpeisiin, mutta estimointiotoksen suhteellisen pienen koon vuoksi tästä vaihtoehdosta luovuttiin. Ottaen huomioon, että minkä tahansa ennustemallin toimivuuden tai hyvyyden mitta on viime kädessä se, miten hyvin malli kykenee ennustamaan tapauksia joihin se on tarkoitettu, olisi mallin relevanssia tärkeää testata uusissa tutkimuksissa. Myös esimerkiksi vanhempien ammatillisen statusta

mittaavan muuttujan poisjättäminen lopullisesta ennustemallista ei ollut itsestään selvä valinta ja sen ennustus- ja selityskyvyn tutkiminen oppimistuloksissa kaipaa lisäselvitystä.

LÄHTEET

Alspaugh, J. W. (1992). Socioeconomic Measures and Achievement: Urban vs. Rural. *Rural Educator*, *13*(3), 2-7.

Chiu, M. M., & Xihua, Z. (2008). Family and motivation effects on mathematics achievement: Analyses of students in 41 countries. *Learning and Instruction*, *18*(4), 321-336.

Chopra, S. L. (1967). Parental occupation and academic achievement of high school students in India. *The Journal of Educational Research*, *60*(8), 359-362.

Fan, X., & Chen, M. J. (1998). Academic Achievement of Rural School Students: A Multi-Year Comparison with Their Peers in Suburban and Urban Schools.

Gelman, A. & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press.

Hampden-Thompson, G., & Johnston, J. (2006). Variation in the Relationship Between Non school Factors and Student Achievement on International Assessments.

Hauser, R. M., & Huang, M. H. (1997). Verbal ability and socioeconomic success: A trend analysis. *Social Science Research*, *26*(3), 331-376.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin*, *105*(2), 198.

OECD. (2017a). PISA 2009 key findings. http://www.oecd.org/pisa/aboutpisa/pisa2009keyfindings.htm. Viitattu 9.10.2017.

OEDC. (2017b). What is PISA? http://www.oecd.org/pisa/aboutpisa/. Viitattu 9.10.2017.

Singh, K., Granville, M., & Dika, S. (2002). Mathematics and science achievement: Effects of motivation, interest, and academic engagement. *The Journal of Educational Research*, *95*(6), 323-332.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, *75*(3), 417-453.

LIITE 1 JÄÄNNÖSTARKASTELUA