TILA311 Harjoitustyö 1

Aleksi Pekkala (alvianpe@student.jyu.fi)

15.10.2017

1 Aineiston kuvaus

Tarkasteltava aineisto käsittelee vuoden 2009 matematiikan PISA-koetta. Aineisto sisältää 500 satunnaisesti valitun 9-luokkalaisen matematiikan pistemäärän sekä useita taustamuuttujia:

mpist PISA-kokeen matematiikan pistemäärä

id koulun tunnus

sukup sukupuoli

HISEI vanhempien ammatillinen status

SES perheen sosioekonominen status

koulusij koulun sijainti (maaseudulla/kaupungissa)

koulualue lääni, jossa koulu sijaitsee

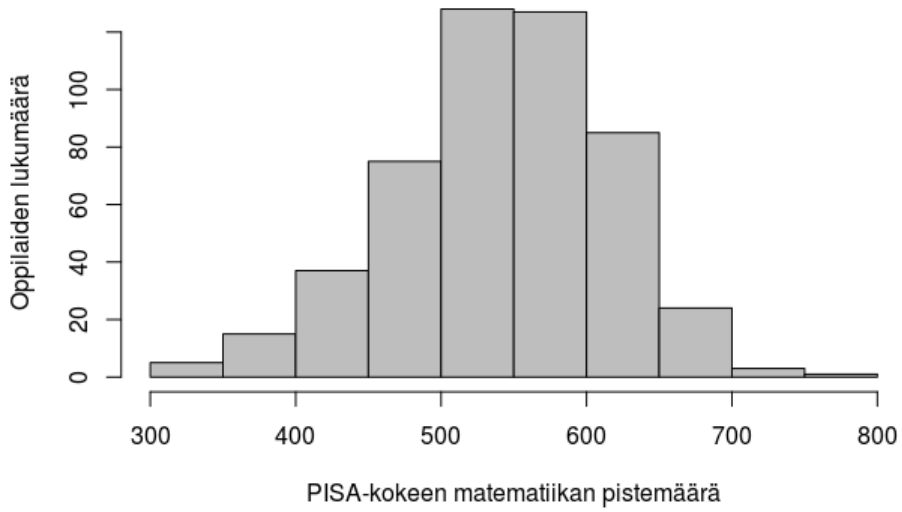
motiv 1, jos oppilas kokee olevansa motivoitunut opiskelemaan, muuten 0

matem matematiikan viimeisin arvosana

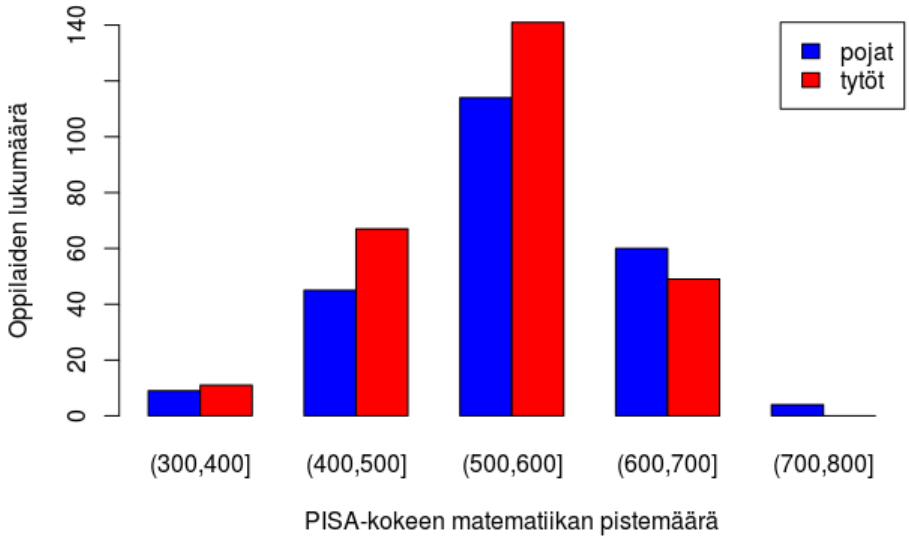
aidink äidinkielen viimeisin arvosana

Matematiikan pistemäärä on normaalisti jakautunut (kuva 1) ja sen keskiarvo on noin 542,4. Kuva 2 esittää pistemäärän jakaumaa sukupuolittain. Aineistossa on 268 tyttöä sekä 232 poikaa, ja vastaavat pistemäärän keskiarvot ovat 535,1 sekä 551,0. Kuva 3 esittää pistemäärän jaoteltuna sen suhteen, kokeeko oppilas olevansa motivoitunut opiskelemaan. Odotetusti oppilaan motivaatio näyttää korreloivan positiivisesti pistemäärän suhteen. Toisaalta aineiston oppilaista vain 161 ei koe olevansa motivoitunut opiskele-

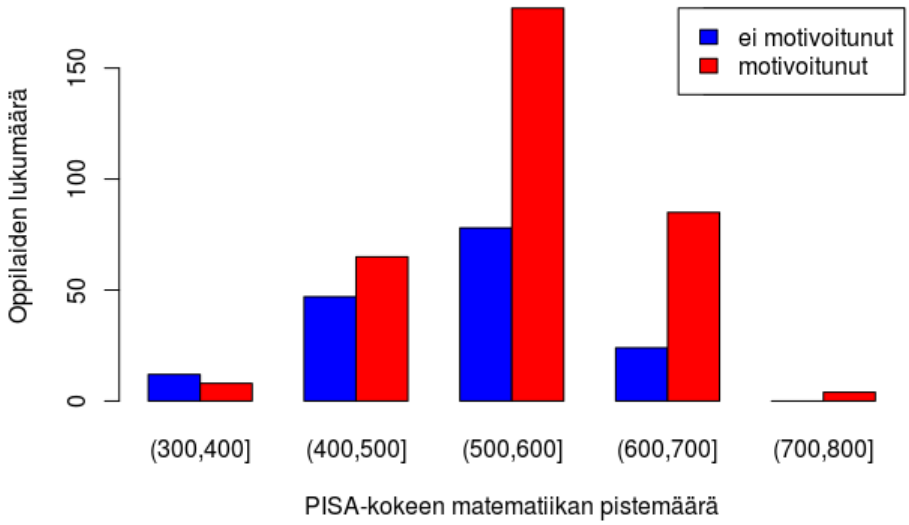
maan. Lisäksi kuva 4 esittää pistemäärän jakaumaa alueittain. Kuvasta ilmenee että Itä-Suomen pistemäärät ovat hieman keskimääräistä korkeampia mutta muuten vaihtelu on vähäistä.



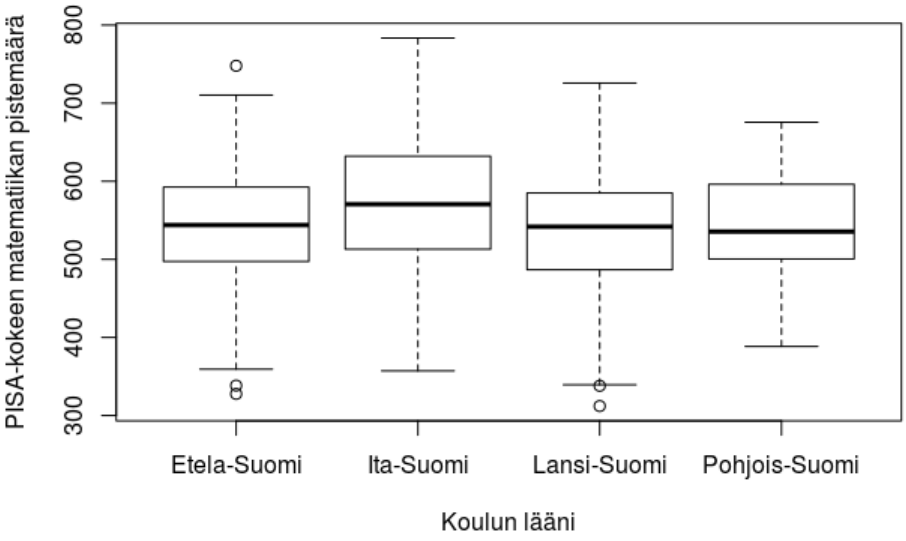
Kuva 1: Matematiikan pistemäärän histogrammi



Kuva 2: Matematiikan pistemäärä sukupuolittain



Kuva 3: Matematiikan pistemäärä motivaatiotasoittain



Kuva 4: Matematiikan pistemäärä eri lääneissä

2 Menetelmät

Mallin formuloinnissa on lähdetty liikkeelle käyttämällä kaikkia potentiaalisia muuttujia. Muuttujista on sitten pyritty karsimaan pois ne, jotka ovat merkitsevyydeltään vähäisimpiä. Lisäksi kaikista vahvimpien selittäjien osalta on kokeiltu erilaisia yhdistelmiä interaktiotermeinä. Tämän tarkoituksena on muodostaa matematiikan pistemäärää mahdollisimman hyvn selittävä malli.

Ensinnäkin mallista on poistettu koulun id-muuttuja. Aineistossa esiintyy yli 100 eri koulua, jolloin niiden vertailu tällä otoskoolla ei ole johdonmukaista. Koulun sijainti (maaseutu/kaupunki) ei osoittautunut merkitseväksi joten se on jätetty mallista pois. Samoin interaktiotermejä ei sisällytetty malliin matalan selitysarvon vuoksi. Koulualueista vain Itä-Suomi on matematiikan pistemäärän kannalta merkitsevä, joten *koulualue*-muuttuja on korvattu *ita*-muuttujalla joka saa arvon 1 tai 0 riippuen siitä onko koulun lääni Itä-Suomi. Myös motivaatiomuuttuja on jätetty mallista pois, koska sen vaikutus on vastoin odotuksia negatiivinen ja merkitsevyys melko matala. Jäljelle jäävistä muuttujista on muodostettu kaava 1.

$$mpist = \beta_0 + \beta_1 \text{HISEI} + \beta_2 \text{SES} + \beta_3 \text{sukup} + \beta_4 \text{matem} + \beta_5 \text{aidink} + \beta_6 \text{ita} + \epsilon \quad (1)$$

3 Tulokset

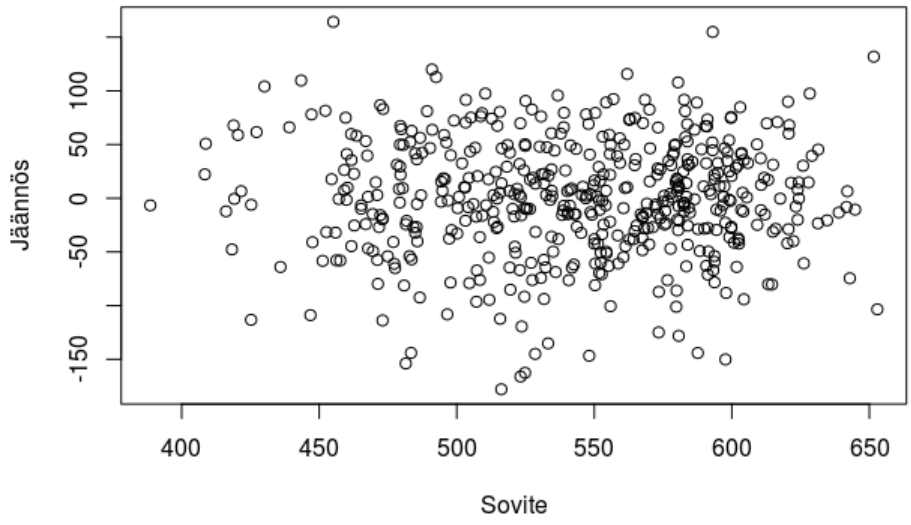
Mallin kertoimien estimaatit on koottu taulukkoon 1. Ennakko-oletuksia vastaavasti viimeisimmällä matematiikan arvosanalla on vahva vaikutus PISA-kokeen pistemäärään. Hieman yllättäen myös äidinkielen arvosanalla on samankaltainen vaikutus. Suurin negatiivinen vaikutus on sukupuolella tyttö. Vanhempien sosioekonomisella taustalla näyttää olevan merkittävä positiivinen vaikutus pistemääriin. *HISEI*-muuttuja vaikuttaa *SES*-muuttujaa vähemmän, mikä voi johtua siitä että *HISEI* on vain yksi laajemman *SES*-muuttujan osatekijöistä. Mallin R^2-arvo on 0,48, joten sen selitysastetta on kohtalaisen hyvä. Toisaalta mallin keskivirhe on melko korkea 54,7.

Kuvassa 5 on piirretty mallin jäännökset sovitteen suhteen. Kuvassa ei näy minkäänlaista systemaattista kuviota, jolloin mallin lineaarisuusoletuksen voidaan olettaa pitävän paikkansa. Kuvassa 6 on piirretty otoskvantiilit

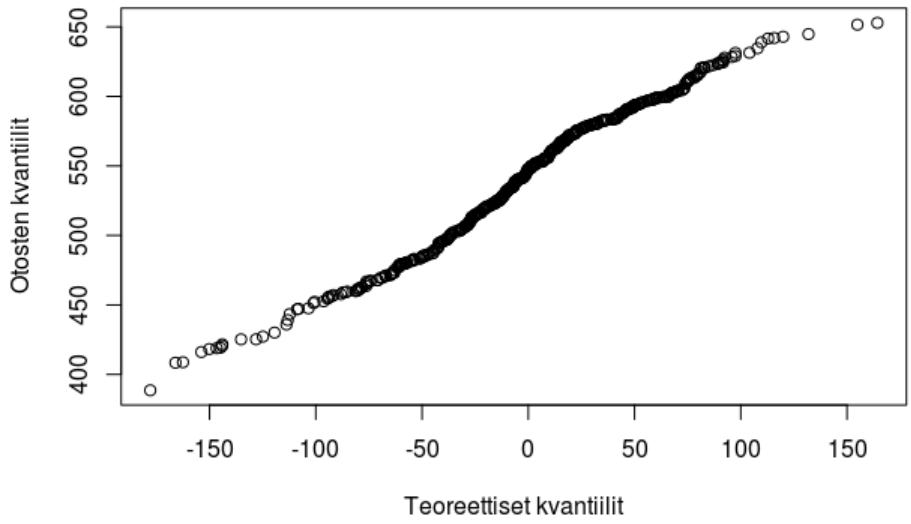
	estimaatti	keskivirhe	t-testi	p-arvo
vakiotermi	296,8	20,9	14,2	$< 0,001$
HISEI	-0,73	0,2	-3,5	$< 0,001$
SES	26,3	4,3	6,1	$< 0,001$
sukup	-24.5	5,3	-4,6	$< 0,001$
matem	23,7	2,2	10,6	$< 0,001$
aidink	13,8	3,0	4,6	$< 0,001$
ita	15,8	7,0	2,2	$< 0,01$

Taulukko 1: Ennustemallin muuttujat

teoreettisten kvantiilien suhteen. Kuvasta ilmenee että myös mallin normaalisuusoletus on voimassa.



Kuva 5: Mallin jäännökset sovitteen suhteen



Kuva 6: Mallin jäännösten normaalisuus

4 Johtopäätökset

Muodostetun mallin perusteella PISA-kokeen matematiikan pistemäärää selittävät eniten vanhempien sosioekonominen status, matematiikan ja äidinkielen viimeisimmät arvosanat sekä sukupuoli. Käytössä olevista muuttujista koulun sijainti tai oppilaiden motivaatio eivät osoittautuneet merkityksellisiksi.

Suuremman otoksen lisäksi mallin tarkkuutta voidaan parantaa lisäämällä muuttujia. Mielenkiintoisia selittäjiä voisivat olla esimerkiksi oppilaan arvio omasta osaamisestaan, opiskeluun käytetty aika, kielitausta, koulun ja opetusryhmien koko, erityisopetuksen tarve sekä oppilaan asenteet matematiikkaa kohtaan. Näistä erityisesti jälkimmäinen voisi selittää muodostettua mallia paremmin sukupuolten välisiä eroja. Lisäksi on syytä tarkastella uudestaan motivaation merkitystä, sillä mallia muodostaessa ilmennyt motivaation yllättävä negatiivinen vaikutus voi viestiä jonkinlaisesta virheestä. Otoskokoa kasvattaessa on myös syytä ottaa huomioon se, että osa oppilaista käy samaa koulua jolloin oppilaita ei voida pitää toisistaan riippumattomina.