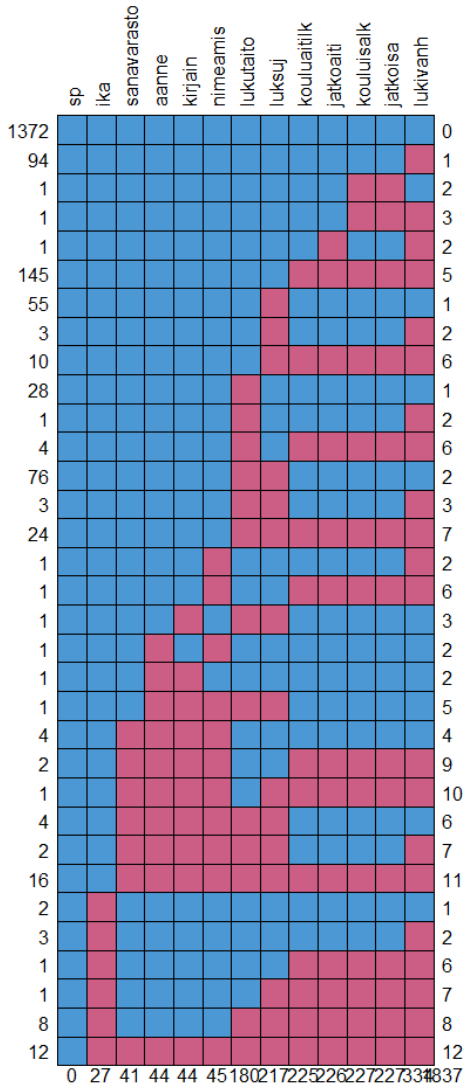
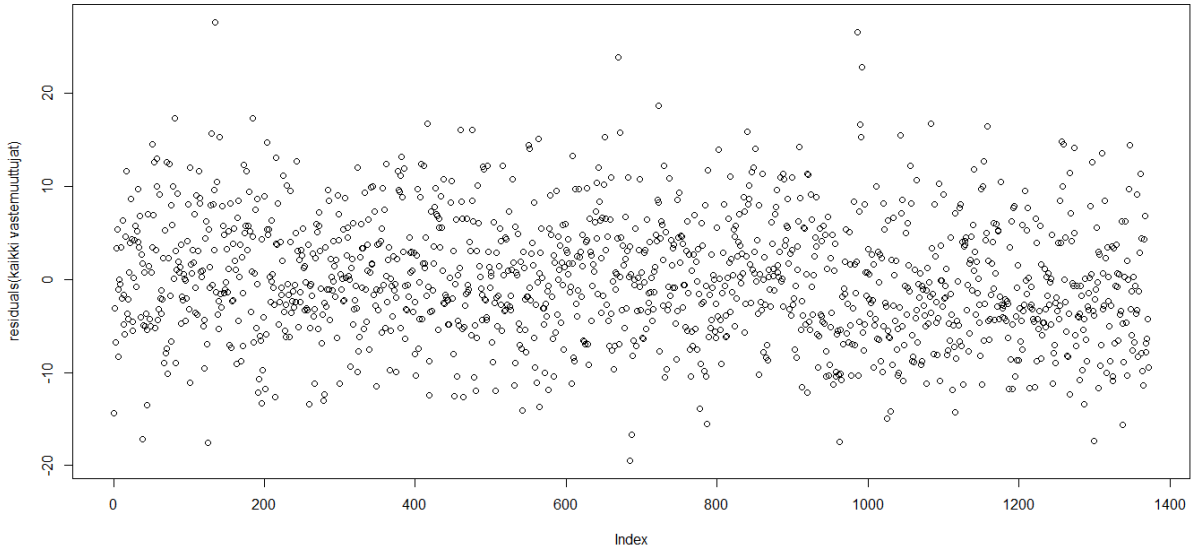
Tutkielman tarkoitus on selvittää, voidaanko joillakin esiopetusvuoden mittauksilla tai muilla selvitetyillä taustamuuttujilla ennustaa lukemisen sujuvuutta toisella luokalla. Luku sujuvuutta mitattiin kokeella, jossa oppilas yhdisti sanoja vastaaviin kuviin. Datassa on mitattu 1880 oppilaan lukemisen sujuvuutta, heidän sukupuolensa,ikä kuukausina, kirjainten tuntemus, äänteiden tuntemus, nimeämisnopeus, lukutaito, molempien vanhempien koulutustausta ja tieto vanhempien lukivaikeuksista. Datassa on 1372 täydellistä riviä eli noin 73 % aineistosta. Puuttuvuutta ilmenee enimmäkseen vanhempien tiedoissa, varsinkin lukivaikeuden kohdalla, joka puuttuu yksinään 94 rivillä ja yhteensä 334 rivillä. Vanhempien koulutuksessa puuttuu äidin tiedot 225 riviltä ja isän 227 riviltä. Vanhempien lukivaikeus muuttuja hiukan mietityttää. En nimittäin tiedä mitä merkitään, jos vain toinen vanhempi vastaa kysymykseen. miten asia on selvitetty, jos vaikka äiti vastaa, että hänellä on lukivaikeuksia mutta isä ei vastaa ollenkaan. Merkitäänkö dataan lukivaikeus yhdelle vanhemmalle tai merkitäänkö tieto puuttuvaksi. Olettaisin, että se merkitään puuttuvaksi mutta en voi asiasta olla varma.

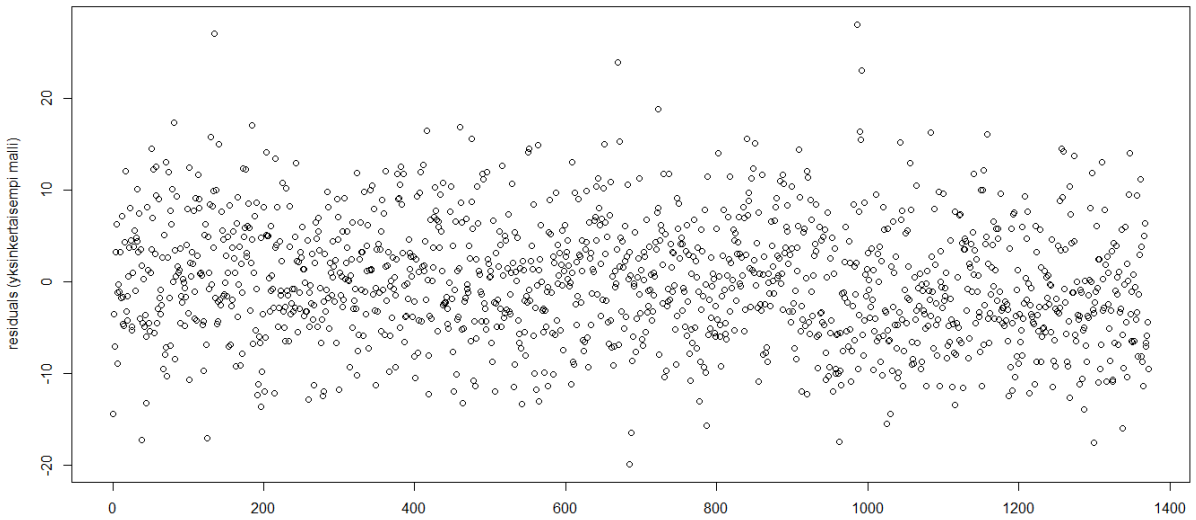


Kuvaaja puuttuvasta datasta(punainen kuvaa puuttuvaa tietoa)

Aloitin selvittämällä lineaarisella mallilla muuttujat, joilla on tilastollisesti merkittävä vaikutus lukusujuvuuteen ja tulos oli, että nimeämisnopeudella, kirjainten tuntemuksella, lukutaidolla, sanavarastolla ja vanhempien lukivaikeuksilla oli tilastollisesti merkittävä vaikutus. Tämän jälkeen tein mallin, jossa oli vain nämä muuttujat. Merkityksellisyys pysyi samana ja residuaalit eivät mallien välillä muuttuneet paljoa, joten koen, että yksinkertaisemmalla mallilla voidaan jatkaa.



Residuaalit kun kaikki vastemuuttujat ovat mallissa



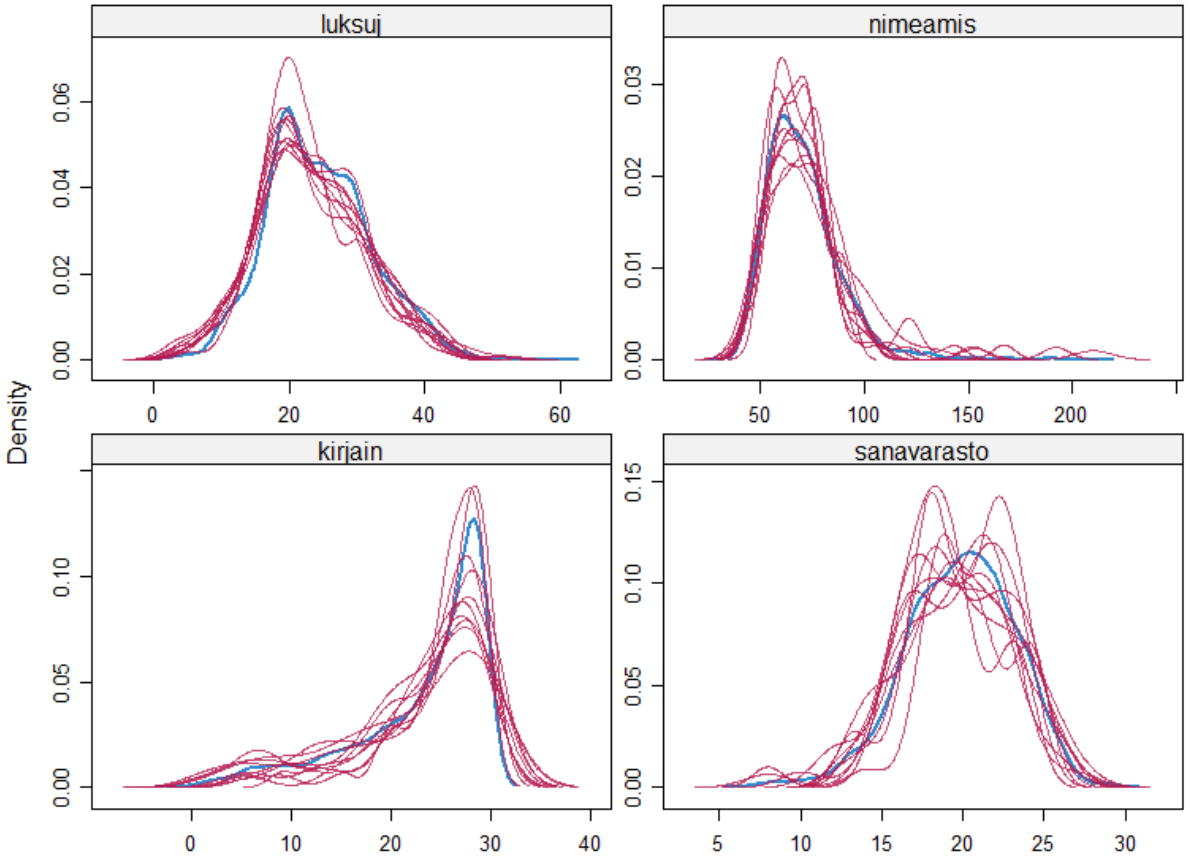
residuaalit yksinkertaisemmassa mallissa

Poistin mallista turhat muuttujat ja näistä jäljellä olevista muuttujista päätin tehdä lopullisen mallin täydellisten rivien analyysiä varten. Moni-imputointi mallia varten

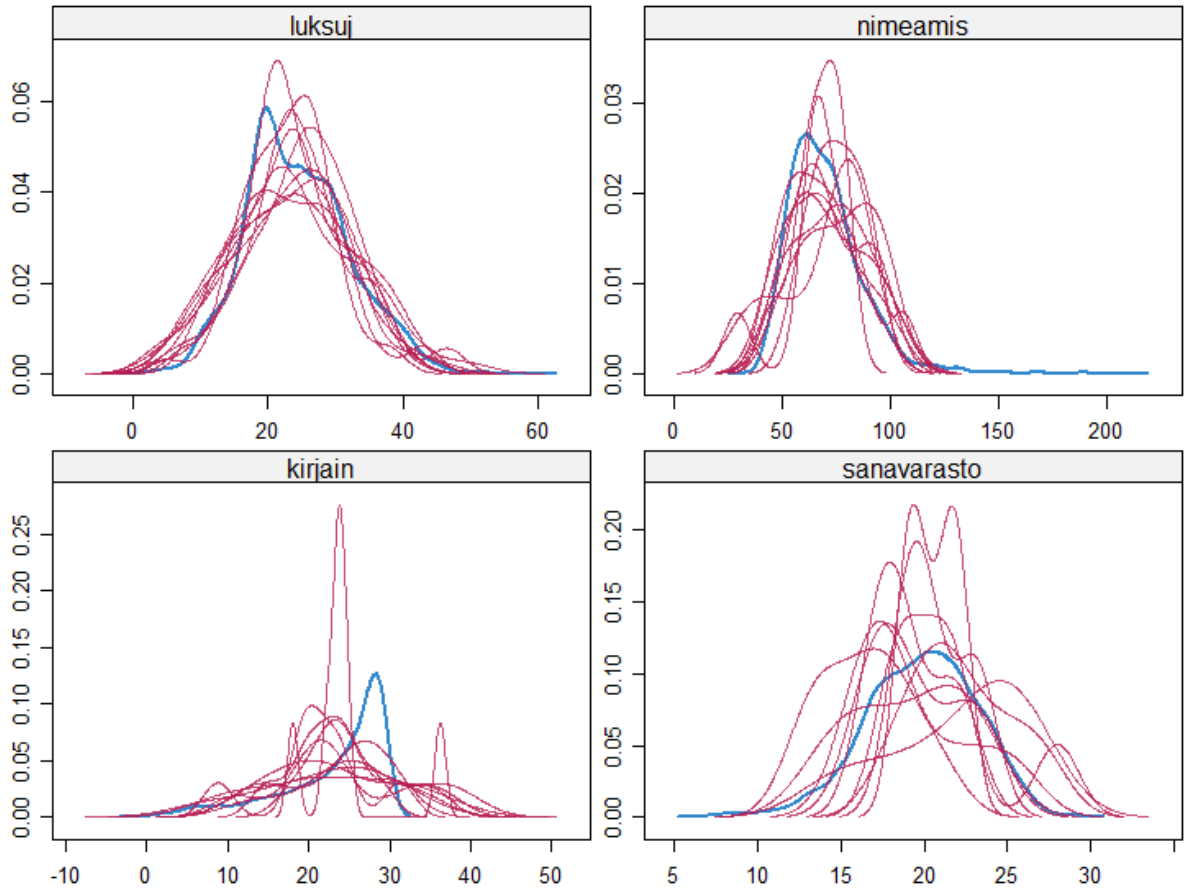
päädyin käyttämään samoja muuttujia kuin täydellisten rivien mallissa. Päädyin käyttämään itse luomaa ennustus matriisia puuttuvan tiedon paikkaamiseen mice (Multivariate Imputation by Chained Equations) algoritmilla.
Matriisi oli seuraavanlainen:

lukusujuvuus	(0,1,1,1,1,1),
nimeämisnopeus	(1,0,1,1,0,0),
kirjainten tuntemus	(0,1,0,1,0,0),
lukutaito	(0,0,1,0,1,0),
sanavarasto	(0,0,1,1,0,0),
lukivaikeus	(1,0,1,1,0,0)

0 tarkoittaa, että muuttujaa ei käytetty selittämään kyseistä muuttujaa.
1 tarkoittaa, että muuttujaa käytettiin selittämään kyseistä muuttujaa.
Imputointien määrä oli 10 ja iteraatioiden määrä oli 50. Uskon, että tämä on riittävä määrä tuottamaan luotettavia tuloksia. Metodina oli pmm (Predictive mean matching). Päädyin käyttämään pmm metodia, sillä norm (Bayesian linear regression) metodin tiheyskuvaajat eivät olleet omasta mielestä järkeviä.

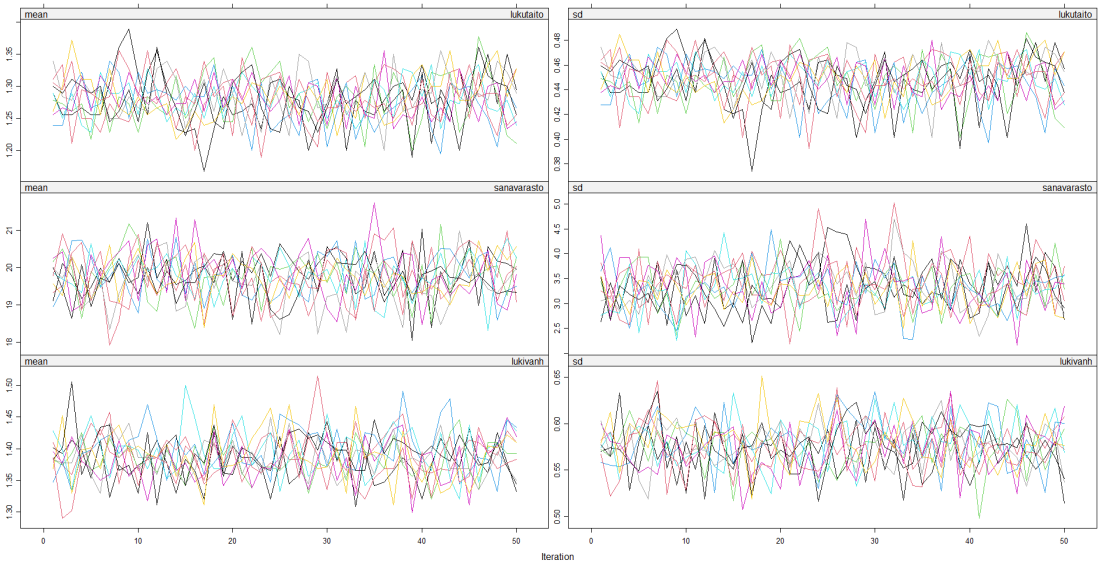


mice algoritmin pmm metodin tiheyskuvaajat



mice algoritmin norm metodin tiheyskuvaajat

Lisäksi moni-imputointi mallin estimaattien keskiarvojen ja keskihajontojen kuvaajissa ei näy minkäänlaisia trendejä tai kuvioita.



Moni-imputointi mallin estimaattien keskiarvojen ja keskihajontojen kuvaajat

Täydellisten rivien analyysillä saatiin seuraavanlaiset tulokset:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	20.20310	1.62607	12.425	< 2e-16	***
nimeämisnopeus	-0.07837	0.01165	-6.726	2.56e-11	***
kirjainten tuntemus	0.14956	0.03291	4.544	6.01e-06	***
lukutaito	4.99231	0.42114	11.854	< 2e-16	***
sanavarasto	0.23919	0.05707	4.191	2.95e-05	***
lukivaikeus toisella	-1.31962	0.40440	-3.263	0.00113	**
lukivaikeus molemmilla	-2.16778	0.91748	-2.363	0.01828	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Moni-inputoinnilla saatiin seuraavanlaiset tulokset:

term	estimate	std.error	statistic	df	p.value
(Intercept)	20.51887	1.50077	13.672	142.037	0.000000e+00
nimeämisnopeus	-0.07741	0.01027	-7.534	178.846	2.347011e-12
kirjainten tuntemus	0.17909	0.02859	6.262	489.355	8.294929e-10
lukutaito	4.39284	0.39170	11.215	326.770	0.000000e+00
sanavarasto	0.19936	0.05098	3.910	320.942	1.124581e-04
lukivaikeus toisella	-1.06263	0.39531	-2.688	122.927	8.183458e-03
lukivaikeus molem	-2.09202	0.88205	-2.371	101.811	1.958391e-02

Täydellisten rivien malli ja moni-imputointi malli tulivat suunnilleen samaan tulokseen siitä, että muuttujat ovat tilastollisesti merkittäviä ja estimaatitkin ovat erittäin lähellä toisiaan. Täydellisten rivien analyysiä ei suositella käyttämään, jos puuttuvuus ei ole tyyppiä MCAR ja tässä tapauksessa se on oletettu tyypiksi MAR. Voi olla mahdollista, että puuttuvuus olisikin tyyppiä MCAR, koska estimaatit ovat niin lähellä toisiaan, jolloin moni-imputointi ei välttämättä ollut tarpeellista. Sanavaraston tiheyskuvaaja ei välttämättä ole paras mahdollinen, ehkä olisi mahdollista saada siitä parempi muuttamalla sen selittäjiä, mutta muuten olen tyytyväinen moni-imputointi malliin. Tutkimuksen tulos on, että lukemisen sujuvuutta voidaan ennustaa joillakin esiopetusvuoden mittauksilla ja vanhempien lukivaikeudella.