LAPPEENRANNAN TEKNILLINEN YLIOPISTO

School of Engineering Science

Laskennallisen tekniikan koulutusohjelma

Kandidaatintyö

Essi Rautasalo

Tilastollista analyysiä DIA-valintapisteistä ja opintopistekertymistä

Ohjaaja Yliopisto-opettaja, TkT Jouni Sampo

TIIVISTELMÄ

Lappeenrannan teknillinen yliopisto

School of Engineering Science

Laskennallisen tekniikan koulutusohjelma

Essi Rautasalo

Tilastollista analyysiä DIA-valintapisteistä ja opintopistekertymistä

Kandidaatintyö

2018

30 sivua, 3 kuvaa, 10 taulukkoa, 4 liitettä

Ohjaaja Yliopisto-opettaja, TkT Jouni Sampo

Avainsanat: χ^2-homogeenisuustesti; Fisherin nelikenttätesti; Kruskal-Wallis -testi; regressioanalyysi; sisäänpääsypisteet; opintopistekertymä;

Tämän kandidaatintyön tavoitteena oli tutkia, löytyykö kandidaatintutkintoon yhteisvalinnassa valittujen opiskelijoiden sisäänpääsytavan ja opintojen etenemisen väliltä yhteyttä Lappeenrannan teknillisessä yliopistossa. Hyvin etenevät opinnot ovat tärkeä asia yliopistolle, koska sen rahoitus perustuu muun muassa opiskelijoiden suorittamiin opintopisteisiin.

Tutkimuksessa tilastollisina menetelminä käytettiin χ^2-homogeenisuustestiä, Fisherin nelikenttätestiä, Kruskal-Wallis -testiä sekä regressioanalyysiä. Tutkimusten lisäksi työssä esitellään siinä käytettyihin menetelmiin liittyvää teoriaa.

Opintopistekertymäjakaumien välillä ei ollut eroja eri valintaryhmissä. Myös kirjoiltapoistettujen ja kirjoilla olevien jakaumat olivat samanlaisia eri valintaryhmissä. Lisäksi selvisi, että valintapisteiden määrä selittää keskimäärin vain muutaman prosentin opintopistekertymän vaihtelusta. Näin ollen yliopiston ei kannata keskittyä erityisesti johonkin kolmesta valintaryhmästä tavoitellessaan nopeasti opintojaan suorittavia opiskelijoita.

ABSTRACT

Lappeenranta University of Technology

School of Engineering Science

Computational Engineering and Technical Physics

Essi Rautasalo

Statistical analysis of DIA selection scores and accumulation of ECTS credits

Bachelor's Thesis

2018

30 pages, 3 figures, 10 tables, 4 attachments

Supervisor University lecturer, D.Sc. (Tech.) Jouni Sampo

Keywords: Chi-square test of homogeneity; Fisher's exact test; Kruskal-Wallis test; regression analysis; selection scores; accumulation of ECTS credits;

The objective of this bachelor's thesis was to study is there an association between accumulation of ECTS credits and the way students are selected to university. Progressive studies are important to university because it gets its funding partly based on the amount of ECTS credits that students have completed.

Statistical methods used in this thesis were chi-square test of homogeneity, Fisher's exact test, Kruskal-Wallis test and regression analysis. In addition to the studies also theory behind the used methods is explained in this thesis.

It wasn't found differences in accumulation of ECTS credit distributions between different selection groups. Also distributions of those students who were removed from registers and who were still enrolled were similar in each selection group. In addition it was found out that the amount of selection scores explains on average only a few percent of accumulation of ECTS credits fluctuation. Hence it's not useful for university to concentrate on especially one selection group when their aim is to select students who complete ECTS credits fast.

Sisältö

Symboli- ja lyhenneluettelo

α	Riskitaso
χ^2-homogeenisuustesti	Khiin neliö -homogeenisuustesti
c_j	Sarakesumma
e_i	Residuaali
e_{ij}	Odotettu frekvenssi χ^2-homogeenisuustestissä
H_0	Nollahypoteesi
H_1	Vastahypoteesi
p-arvo	Merkitsevyystaso
r	Pearsonin tulomomenttikorrelaatiokerroin eli otoskorrelaatiokerroin
r_i	Rivisumma
R^2	Selitysaste

DIA	Diplomi-insinööri- ja arkkitehtikoulutus
LBM	LUT School of Business and Management
LENS	LUT School of Engineering Science
LES	LUT School of Energy Systems
LUT	Lappeenranta University of Technology, Lappeenrannan teknillinen yliopisto
op	Opintopiste
PNS-menetelmä	Pienimmän neliösumman menetelmä

1 JOHDANTO

Yliopistojen rahoituksen kannalta on tärkeää, että opiskelijat suorittavat tutkintoaan tavoiteaikataulussa. Yliopistojen koulutukseen perustuvassa rahoitusosuudessa huomioidaan muun muassa tutkintoon valmistuneiden ja lukuvuodessa vähintään 55 opintopistettä suorittaneiden opiskelijoiden lukumäärät [1]. Yliopistot pyrkivätkin kehittämään opiskelijavalintaansa ja toimintatapojaan, jotta mahdollisimman moni opiskelijoista suorittaisi lukuvuodessa vähintään 55 opintopistettä.

Teknilliset yliopistot tarjoavat kandidaatintutkinnolla alkavaa koulutusta, jossa opiskelijalla on opiskelupaikan saatuaan oikeus suorittaa diplomi-insinöörin tutkinto, sekä pelkästään ylempään korkeakoulututkintoon johtavia tekniikan maisteriohjelmia. Yliopistot saavat itse päättää opiskelijavalinnan valintaperusteista tietyin reunaehdoin. Paikkoja on varattava esimerkiksi sellaisille opiskelijoille, jotka eivät ole suorittaneet aikaisemmin korkeakoulututkintoa tai vastaanottaneet sellaista opiskelupaikkaa, joka johtaa korkeakoulututkintoon [2].

1.1 Työn taustaa

Opiskelijat valitaan teknillisien yliopistojen kandidaattiohjelmiin diplomi-insinööri- ja arkitehtikoulutuksen (DIA) yhteisvalinnassa kolmessa eri valintaryhmässä: ylioppilastutkintotodistuksen tai muun vastaavan mukaan laskettujen alkupisteiden perusteella, valintakokeesta saatujen pisteiden ja ylioppilastutkintotodistuksen mukaan laskettujen alkupisteiden yhteismäärän perusteella sekä ainoastaan valintakokeesta saatujen pisteiden perusteella. Jokaisessa valintaryhmässä on erilaiset maksimipisteet ja tällöin myös erilaiset sisäänpääsyrajat. Yliopistot voivat vaikuttaa eri valintaryhmien kokoihin varaamalla jokaiselle valintaryhmälle tietyn verran aloituspaikkoja kullekin koulutusalalle. Yhteishaun lisäksi yliopistot voivat valita opiskelijoita tiettyihin koulutuksiin myös erillisvalinnalla. Erillisvalinnalla valitaan opiskelijoita kandidaattiohjelmiin valtakunnallisten alaan liittyvien kilpailuiden menestyksen perusteella ja maisteriohjelmiin koulutusohjelmakohtaisilla valintaperusteilla.

Yliopistolla on käytettävissään tiedot vuosittain opiskelemaan valittujen opiskelijoiden sisäänpääsytavoista sekä hakupisteistä. Lisäksi yliopistolla on käytössä dataa opiskelijoiden opiskeluvauhdista eli tieto siitä, kuinka paljon opintopisteitä he ovat suorittaneet lukukausittain. Näitä tietoja analysoimalla voidaan saada selville valintapisteiden ja -tavan sekä opiskeluvauhdin yhteydet toisiinsa.

1.2 Työn tavoitteet

Työn tavoitteena on tutkia tilastollisin menetelmin, onko DIA-yhteisvalinnan sisäänpääsytavan ja opintojen etenemisen välillä yhteyksiä tekniikan aloilla Lappeenrannan teknillisessä yliopistossa (Lappeenranta University of Technology, LUT). Työssä tarkastellaan vain kandidaatintutkintoon DIA-yhteisvalinnalla valittujen opiskelijoiden sisäänpääsy- eli valintapisteitä ja heidän opintojensa etenemistä, eli erillisvalintaa ei käsitellä tässä kandidaatintyössä. Koska yliopistot voivat itse vaikuttaa osittain valintaperusteisiinsa, heitä kiinnostaa varmasti, mihin valintaryhmään heidän kannattaisi keskittyä opiskelijoiden opiskeluvauhdin ja siten oman rahoituksensa vuoksi. Työssä tutkitaan kahden vuoden aikana yhteishaussa kandidaattiohjelmiin valittuja opiskelijoita. Mikäli tulokset antavat viitteitä, että opiskeluvauhdissa on eroja eri valintatavalla opiskelemaan valittujen välillä, voidaan tutkimusta jatkaa laajempaa opiskelijajoukkoa koskevalla datamäärällä. Tällöin nähtäisiin, ovatko havaitut erot voimassa vuodesta toiseen. Lisäksi kiinnostaa, keskeyttävätkö opiskelijat opintojaan enemmän jossain tietyssä valintaryhmässä. Tätä tutkitaan kirjoiltapoistettujen ja kirjoilla olevien opiskelijoiden lukumäärien perusteella.

1.3 Työn toteutus

Työssä analysoidaan yliopistolla olevaa dataa opiskelijoiden sisäänpääsypisteistä, -tavasta ja opintojen etenemisestä tilastollisin menetelmin. Työssä tutkitaan yleisiä riippuvuuksia valintamenettelyn ja opiskeluvauhdin välillä eikä siinä tarkastella tai tutkimuksen tuloksista käy ilmi yksittäistä opiskelijaa koskevia tietoja.

Matemaattisia menetelmiä käytettäessä voidaan hyödyntää laskentaohjelmistoja, mutta joskus on käytännöllisempää ohjelmoida tarvitsemansa menetelmät itse. Tällöin koodin voi kirjoittaa omaan tutkimukseen sopivaksi esimerkiksi syötteiden ja tulostuksen muotoilun osalta. Tässä kandidaatintyössä päädyttiin toteuttamaan käytettävät menetelmät itse ohjelmoiden, jotta laskennan toteuttaminen käytännössä tulee tutuksi ohjelmoinnin kautta.

Tilastollisina menetelminä käytetään χ^2-homogeenisuustestiä, Fisherin nelikenttätestiä, Kruskal-Wallis -testiä ja regressioanalyysiä. Työssä esitellään ensin näihin menetelmiin liittyvää teoriaa ja käydään läpi niiden rajoitteita. Koska kandidaatintyön toteutukseen kuuluu myös valittujen tilastollisten menetelmien ohjelmoiminen Matlab-ohjelmistolla, luvussa kolme esitellään menetelmien ohjelmoimiseen liittyviä huomioita ja määrittelyjä. Luvussa neljä kuvataan datan käsittelyä sekä datan analysoimista itse kirjoitettuja ohjelmia käyttämällä. Lopuksi käsitellään tutkimuksen perusteella tehdyt johtopäätökset, pohditaan tutkimuksen jatkoa sekä tehdään yhteenveto tästä kandidaatintyöstä.

2 TILASTOLLISET MENETELMÄT

2.1 Khiin neliö -homogeenisuustesti

χ^2-homogeenisuustestillä tutkitaan, onko kahta eri satunnaismuuttujaa X ja Y kuvaavan luokitellun aineiston ryhmien välillä eroa eli onko taulukoidun aineiston vaaka- tai pystyrivijakaumissa eroja. Tutkittava aineisto esitetään ristiintaulukoimalla havaintoaineisto taulukon 1 mukaisesti. Taulukossa muuttujan X luokkia kuvaavat merkinnät $E_1, \ldots, E_k$ ja muuttujan Y luokkia kuvaavat merkinnät $F_1, \ldots, F_m$. Taulukon alkiot n_{ij} kuvaavat kyseiseen soluun kuuluvien havaintojen lukumäärää eli niitä havaintoja, joissa satunnaismuuttujaa X kuvaava arvo kuuluu luokkaan E_i ja satunnaismuuttujaa Y kuvaava arvo kuuluu luokkaan F_j.

Taulukko 1: Ristiintaulukoitu havaintoaineisto

$X \setminus Y$	F_1	F_2	$\ldots$	F_m	$\sum$
E_1	n_{11}	n_{12}	$\ldots$	n_{1m}	r_1
E_2	n_{21}	n_{22}	$\ldots$	n_{2m}	r_2
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
E_k	n_{k1}	n_{k2}	$\ldots$	n_{km}	r_k
$\sum$	c_1	c_2	$\ldots$	c_m	

Summat r_i ovat rivisummia ja summat c_j ovat sarakesummia, jotka saadaan laskemalla rivin tai sarakkeen havaintojen lukumäärät yhteen eli

$$r_i = \sum_{j=1}^{m} n_{ij} \qquad i = 1, \ldots, k$$

$$c_j = \sum_{i=1}^{k} n_{ij} \qquad j = 1, \ldots, m$$

Kaikkien havaintoarvojen lukumäärä n saadaan laskemalla joko rivi- tai sarakesummien summa eli

$$n = \sum_{i=1}^{k} r_i = \sum_{j=1}^{m} c_j$$

Testin hypoteesit voidaan kirjoittaa muodossa

H_0**:** Y:n vaakarivijakaumat ovat samanlaiset muuttujan X eri luokissa

H_1**:** Y:n vaakarivijakaumissa on eroja

tai

H_0**:** X:n pystyrivijakaumat ovat samanlaiset muuttujan Y eri luokissa

H_1**:** X:n pystyrivijakaumissa on eroja

Testiä varten lasketaan odotettujen frekvenssien arvot e_{ij} rivi- ja sarakesummien ja havaintoarvojen lukumäärän avulla seuraavasti

$$e_{ij} = \frac{r_i c_j}{n} \tag{1}$$

Testisuure lasketaan odotettujen frekvenssien avulla kaavalla

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{2}$$

χ^2-homogeenisuustestiä voidaan käyttää, mikäli korkeintaan 20 % odotetuista frekvensseistä on pienempiä kuin 5 ja kaikki odotetut frekvenssit ovat suurempia kuin 1 [3]. Tällöin testisuure noudattaa likimain jakaumaa

$$\chi^2 \sim_a \chi^2\big((k-1)(m-1)\big)$$

jossa k on havaintoaineistosta tehdyn taulukon rivien lukumäärä ja m on taulukon sarakkeiden lukumäärä. Kun testin riskitasoksi valitaan α, hylkäysehdoksi saadaan

$$\chi^2 > \chi^2_{1-\alpha}\big((k-1)(m-1)\big)$$

eli nollahypoteesi hylätään, jos testisuuren arvo on suurempi kuin kohdassa $1-\alpha$ laskettu χ^2-jakauman kertymäfunktion arvo $(k-1)(m-1)$ vapausasteella. Arvoa $\chi^2_{1-\alpha}\big((k-1)(m-1)\big)$ kutsutaan kriittiseksi arvoksi. Testisuuren arvon olleessa kriittistä arvoa pienempi nollahypoteesi jää voimaan.

Testin tulos voidaan laskea myös p-arvon eli merkitsevyystason avulla. Määritetään p-arvo kaavalla

$$p\text{-arvo} = P(X > \chi^2) \tag{3}$$

jossa χ^2 on testisuureen arvo ja satunnaismuuttujan X noudattaa χ^2-jakaumaa. Nollahypoteesi H_0 hylätään, jos testin p-arvo on pienempi kuin valittu riskitaso.

2.2 Fisherin nelikenttätesti

Tarkasteltavan havaintoaineiston ollessa hyvin pieni χ^2-homogeenisuustesti ei sovi taulukoidun aineiston pysty- tai vaakarivijakaumien samankaltaisuuden tutkimiseen odotettuihin frekvensseihin liittyvien ehtojen takia. Tällöin voidaan käyttää Fisherin nelikenttätestiä samankaltaisuuksien tutkimiseen, sillä Fisherin nelikenttätestissä ei ole ehtoja tutkittavien havaintoarvojen suuruudelle [3]. Fisherin nelikenttätesti sopii hyvin 2×2 -kokoiselle taulukolle, sillä laajemmissa taulukoissa laskeminen käy huomattavasti työläämmäksi [4].

Tutkittava aineisto esitetään ristiintaulukoimalla ja taulukon reunoille lasketaan rivisummat r_i ja sarakesummat c_j. Taulukossa 2 on esitetty tilanne 2×2 -kokoisessa taulukossa.

Taulukko 2: Fisherin nelikenttätestin havaintoaineisto taulukoituna

$X \setminus Y$	F_1	F_2	$\sum$
E_1	n_{11}	n_{12}	r_1
E_2	n_{21}	n_{22}	r_2
$\sum$	c_1	c_2	

Kaikkien havaintojen lukumäärä n saadaan joko rivi- tai sarakesummien summana

$$n = r_1 + r_2 = c_1 + c_2$$

Seuraavaksi lasketaan rajatodennäköisyys p_{cutoff}, jonka avulla määritetään testin lopputulos. Arvo p_{cutoff}:lle saadaan laskettua rivi- ja sarakesummien sekä taulukon alkioiden n_{ij} perusteella seuraavasti

$$p_{cutoff} = \frac{r_1!\, r_2!\, c_1!\, c_2!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} \tag{4}$$

Tapa on hypergeometrisen todennäköisyysfunktion yleistys usealle muuttujalle [4]. Tämän jälkeen muokataan havaintotaulukkoa siten, että etsitään kaikki muut mahdolliset taulukot, joissa taulukon alkiot n_{ij} ovat positiivisia kokonaislukuja sekä rivisummat r_i ja sarakesummat c_j pysyvät samoina kuin alkuperäisessä taulukossa. Jokaiselle näin muodostetulle taulukolle lasketaan p-arvo samalla tavalla kuin p_{cutoff} laskettiin alkuperäiselle taulukolle. Kun kaikkien eri taulukoiden p-arvot lasketaan yhteen, summaksi saadaan 1.

Testin kannalta merkityksellisiä ovat ne p-arvot, jotka ovat pienempiä tai yhtä suuria kuin p_{cutoff}. Laskemalla näiden p-arvojen summa

$$p_{sum} = \sum p_t \qquad p_t \leq p_{cutoff} \tag{5}$$

ja vertaamalla sitä testin riskitasoon α voidaan tehdä päätelmiä havaintoaineiston rivien tai sarakkeiden samankaltaisuudesta. Myös alkuperäisen matriisin p-arvo p_{cutoff} otetaan mukaan p_{sum}-arvoa laskettaessa. Jos p_{sum} on suurempi kuin riskitasoksi valittu arvo, ovat taulukon pysty- tai vaakarivijakaumat samanlaisia. Vastaavasti, jos p_{sum} on pienempi kuin testin riskitaso, pysty- tai vaakarivijakaumat eivät ole samanlaisia.

2.3 Kruskal-Wallis -testi

Kruskal-Wallis -testi soveltuu kolmen tai useamman jakauman samanlaisuuden tutkimiseen. Testi soveltuu myös sellaisten jakaumien tutkimiseen, joissa jäännöstermit eivät noudata normaalijakaumaa; riittää, että jäännöstermit noudattavat keskenään samaa jakaumaa [5].

Testattava aineisto sisältää k kappaletta keskenään vertailtavia ryhmiä. Yhteensä havaintoarvoja kaikissa k:ssa ryhmässä on n_T kappaletta. Näitä havaintoja x_{ij}, jossa $1 \leq i \leq k$ ja $1 \leq j \leq n_i$, on yhdessä ryhmässä n_i kappaletta. Kaikissa vertailtavissa ryhmissä ei tarvitse olla yhtä paljon havaintoja. Testin havaintoaineistoa voidaan havainnollistaa taulukolla, jonka ensimmäiseen sarakkeeseen järjestetään kaikki havaintoarvot pienimmästä suurimpaan. Taulukon toiseen sarakkeeseen kirjataan tieto siitä, mihin ryhmään kyseinen havainto kuuluu. Taulukon kolmanteen sarakkeeseen kirjataan havaintoarvojen järjestys, eli ne numeroidaan järjestyksessä $1, 2, \ldots, n_T$. Mikäli taulukossa on kahdella rivillä sama havaintoarvo, näiden kohdalla havaintoarvojen järjestys -sarakkeeseen kirjataan havaintoarvojen sijoitusten keskiarvo. Havaintotaulukkoa on havainnollistettu taulukossa 3, jossa havaintoarvoja on kuvattu kirjainsymboleilla.

Taulukko 3: Kruskal-Wallis -testiä varten taulukoitu havaintoaineisto

Havaintoarvot x_{ij}	Havaintoarvon ryhmä	Havaintoarvojen järjestys r_{ij}
a	k	1
b	$k-1$	2
e	$k-2$	3.5
e	$k-1$	3.5
$\vdots$	$\vdots$	$\vdots$
s	k	n_T

Testin hypoteesit voidaan kirjoittaa seuraavasti

H_0**:** Kaikki k tutkittavaa ryhmää tulevat samasta jakaumasta eli ryhmien välillä ei ole merkittävää eroa

H_1**:** Ainakin kaksi tutkittavista ryhmistä eroavat toisistaan

Testiä varten lasketaan jokaiselle ryhmälle havaintojen järjestyksien keskiarvo $\bar{r}_{1\cdot}, \ldots, \bar{r}_{k\cdot}$ kaavalla

$$\bar{r}_{i\cdot} = \frac{r_{i1} + \cdots + r_{in_i}}{n_i}$$

Testisuure h lasketaan keskiarvojen $\bar{r}_{i\cdot}$ ja havaintojen kokonaismäärän n_T avulla seuraavasti

$$h = \frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} n_i \bar{r}_{i\cdot}^2 - 3(n_T+1) \tag{6}$$

Testin johtopäätökset tehdään testin p-arvon ja riskitason avulla. Laskettaessa p-arvoa satunnaismuuttuja X noudattaa χ^2-jakaumaa $k-1$ vapausasteella. Tällöin

$$p\text{-arvo} = P(X > h) \tag{7}$$

Nollahypoteesi hylätään, eli ainakin kaksi tutkittavista ryhmistä eroaa toisistaan, jos p-arvo on pienempi kuin testin riskitaso. Muulloin nollahypoteesi jää voimaan, eli tutkittavien ryhmien välillä ei ole merkittävää eroa.

2.4 Regressioanalyysi

Regressioanalyysissä on tavoitteena löytää muuttujien välinen yhteys siten, että selittävien muuttujien x_k avulla voidaan kuvata selitettävää muuttujaa y. Selitettävä muuttuja y voi riippua joko yhdestä tai useammasta selittävästä muuttujasta ja riippuvuus voi olla lineaarista tai epälineaarista, kuten polynomiaalista tai eksponentiaalista.

Regressiota käsiteltäessä on hyvä muistaa Ayyub'n ja McCuenin teoksessaan esille nostama ero regression ja korrelaation välillä. Regressio on mallin muodostamisessa käytetty menetelmä ja siinä määritetään ennustavan yhtälön tuntemattomat kertoimet. Korrelaation avulla voidaan puolestaan arvioida muodostetun sovitteen hyvyyttä ja sitä voidaan käyttää muun muassa mallin muotoilussa. Regressiota käytettäessä on lisäksi tiedettävä, mikä muuttujista on selitettävä muuttuja ja minkä muuttujan avulla sitä selitetään. Määritettävät regressiokertoimet nimittäin eroavat toisistaan vaihdettaessa selitettävä muuttuja selittäväksi muuttujaksi, ellei korrelaatiokerroin ole tasan 1. Korrelaatiota laskettaessa tällainen erottelu selittävän muuttujan ja selitettävän muuttujan välillä ei ole tarpeellista. [6]

2.4.1 Yhden selittävän muuttujan lineaarinen regressio

Yhden selittävän muuttujan lineaarisessa regressiossa selitettävä muuttuja y riippuu vain yhdestä selittävästä muuttujasta x. Datapisteisiin $(x_1, y_1), \ldots, (x_n, y_n)$ sovitettava regressiosuora on muotoa

$$y = \beta_0 + \beta_1 x$$

Suoran sovitetta määritettäessä on tavoitteena löytää datapisteitä jollain tapaa lähinnä oleva suora. Yleisin tapa on Hayter'n mukaan minimoida datapisteiden ja suoran välistä pystysuuntaista eroa. Useimmiten minimoidaan pystysuuntaisten erojen neliöllistä summaa

$$q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{8}$$

Tämä voidaan perustella tutkimalla tarkemmin datapisteisiin (x_i, y_i) sovitettua regressiomallia

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{9}$$

jossa ϵ_i:t ovat jäännöstermejä. Datapisteen y_i arvo muodostuu siis sovitteen avulla lasketusta arvosta sekä jäännöstermistä, joka kuvaa todellisen ja mallin avulla lasketun arvon eroa pisteessä x_i. Seuraavassa esitetty päättely edellyttää, että jäännöstermit ovat toisistaan riippumattomat ja noudattavat normaalijakaumaa $N(0, \sigma^2)$ jäännösvarianssilla σ^2. Tällöin arvot $y_1, \ldots, y_n$ ovat havaintoja satunnaismuuttujasta

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

joka noudattaa jakaumaa

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Jäännöstermin E_i tiheysfunktio on

$$\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\epsilon_i^2/2\sigma^2}$$

ja mallin jäännöstermien $\epsilon_1, \ldots, \epsilon_n$ tiheysfunktio

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n \epsilon_i^2/2\sigma^2}$$

Tämä todennäköisyys halutaan mahdollisimman suureksi, koska suurimman todennäköisyyden kohdassa mallin parametreille saadaan parhaimmat estimaattien arvot. Todennäköisyys maksimoituu, kun minimoidaan jäännöstermien ϵ_i neliöiden summa

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = q$$

Käytettäessä pienimmän neliösumman menetelmää (PNS-menetelmä) jäännöstermien neliöllinen summa saadaan minimoitua. [5]

Parametrien estimaatit $\hat{\beta}_0$ ja $\hat{\beta}_1$, joita kutsutaan myös suurimman uskottavuuden estimaateiksi (maximum likelihood estimates), ovat ne arvot, jotka minimoivat q:n lausekkeen. Ne saadaan määritettyä laskemalla q:n osittaisderivaatat ja merkitsemällä ne nolliksi. Tällöin

$$\begin{cases} \frac{\partial q}{\partial \beta_0} &= \sum_{i=1}^n -2(y_i - (\beta_0 + \beta_1 x_i)) = 0 \\ \frac{\partial q}{\partial \beta_1} &= \sum_{i=1}^n -2x_i(y_i - (\beta_0 + \beta_1 x_i)) = 0 \end{cases}$$

josta saadaan

$$\begin{cases} \sum_{i=1}^n y_i &= \beta_0 n + \beta_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i &= \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 \end{cases}$$

Yllä olevia yhtälöitä kutsutaan normaaliyhtälöiksi [5]. Ratkaisemalla normaaliyhtälöiden ylemmästä yhtälöstä β_0 ja sijoittamalla se alempaan yhtälöön saadaan estimaatille $\hat{\beta}_1$ kaava

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \tag{10}$$

Tämän jälkeen $\hat{\beta}_0$ voidaan laskea $\hat{\beta}_1$:n avulla seuraavasti

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \tag{11}$$

Sovitetun regressiosuoran yhtälö on siis

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Jäännösvarianssin estimaatti $\hat{\sigma}^2$ voidaan laskea kaavalla

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{n-2} \tag{12}$$

2.4.2 Usean selittävän muuttujan lineaarinen regressio

Usean selittävän muuttujan lineaarisessa regressiossa selitettävä muuttuja y riippuu useasta selittävästä muuttujasta $x_1, \ldots, x_k$. Datajoukkoon

$$\begin{gathered}(y_1, x_{11}, x_{21}, \ldots, x_{k1}) \\ \vdots \\ (y_n, x_{1n}, x_{2n}, \ldots, x_{kn})\end{gathered}$$

sovitetulla regressiomallilla

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i \tag{13}$$

voidaan kuvata datapisteen arvoa y_i sovitteen arvon ja jäännöstermin summana. Tässäkin tapauksessa jäännöstermit ϵ_i ovat toisistaan riippumattomat ja noudattavat $N(0, \sigma^2)$-jakaumaa kuten yhden selittävän muuttujan lineaarisessa regressiossa. Arvot $y_1, \ldots, y_n$ viittaavat havaintoihin satunnaismuuttujasta Y, jonka odotusarvo on

$$E(Y|_{\mathbf{x}}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

selittävien muuttujien ollessa $\mathbf{x} = (x_1, \ldots, x_k)$ [5]. Tällöin datajoukkoon sovitettava hypertaso avaruudessa $\mathbb{R}^{k+1}$ on muotoa

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

jossa k kuvaa selittävien muuttujien lukumäärää. Yhden selittävän muuttujan lineaarinen regressio on siis erikoistapaus, jossa $k = 1$.

Jäännöstermien normaalijakautuneisuuden perusteella myös useamman selittävän muuttujan tilanteessa mallin kertoimien $\beta_0, \ldots, \beta_k$ suurimman uskottavuuden estimaatit $\hat{\beta}_0, \ldots, \hat{\beta}_k$ ovat ne parametrien arvot, jotka minimoivat lausekkeen

$$q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}))^2 \tag{14}$$

Estimaatit saadaan laskettua osittaisderivaattojen nollakohtien avulla lausekkeesta

$$\frac{\partial q}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})) = 0$$

sekä lausekkeesta

$$\frac{\partial q}{\partial \beta_j} = \sum_{i=1}^{n} -2x_{ji}(y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})) = 0$$

joka lasketaan jokaiselle j:lle väliltä $1 \leq j \leq k$. Näin saadaan $k + 1$ yhtälöä

$$\begin{cases} \sum_{i=1}^{n} y_i & = \beta_0 n + \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} + \cdots + \beta_k \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} y_i & = \beta_0 \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{1i} x_{2i} + \cdots + \beta_k \sum_{i=1}^{n} x_{1i} x_{ki} \\ & \vdots \\ \sum_{i=1}^{n} x_{ki} y_i & = \beta_0 \sum_{i=1}^{n} x_{ki} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{ki} + \beta_2 \sum_{i=1}^{n} x_{2i} x_{ki} + \cdots + \beta_k \sum_{i=1}^{n} x_{ki}^2 \end{cases}$$

joita kutsutaan myös normaaliyhtälöiksi [5]. Yksi tapa ratkaista tällainen yhtälö on kirjoittaa se matriisimuotoon ja ratkaista parametrit matriisilaskennan avulla. Tällöin lineaarinen malli voidaan kirjoittaa muodossa

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{15}$$

jossa vektori $\mathbf{Y}$ on pystyvektori, joka sisältää selitettävän muuttujan arvot $y_1, \ldots, y_n$, matriisi $\mathbf{X}$ on selittävien muuttujien arvot sisältävä matriisi

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

vektori $\boldsymbol{\beta}$ on estimoitavat parametrit $\beta_0, \ldots, \beta_k$ sisältävä pystyvektori ja vektori $\boldsymbol{\epsilon}$ on pystyvektori, joka sisältää jäännöstermit $\epsilon_1, \ldots, \epsilon_n$. Matriisimuotoa käytettäessä normaaliyhtälöt voidaan esittää muodossa

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \tag{16}$$

josta voidaan ratkaista PNS-estimaatit parametreille $\beta_0, \ldots, \beta_k$ [5]. Tällöin matriisi $\mathbf{X}'\mathbf{X}$ kirjoitetaan muodossa

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \cdots & \sum_{i=1}^n x_{ki} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \cdots & \sum_{i=1}^n x_{1i}x_{ki} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 & \cdots & \sum_{i=1}^n x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^n x_{ki} & \sum_{i=1}^n x_{1i}x_{ki} & \sum_{i=1}^n x_{2i}x_{ki} & \cdots & \sum_{i=1}^n x_{ki}^2 \end{bmatrix}$$

ja matriisi $\mathbf{X}'\mathbf{Y}$ saa muodon

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{1i}y_i \\ \vdots \\ \sum_{i=1}^n x_{ki}y_i \end{bmatrix}$$

Parametrien estimaateiksi saadaan

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{17}$$

mikäli matriisi $(\mathbf{X}'\mathbf{X})^{-1}$ on olemassa.

Lineaarisia malleja ovat myös sellaiset polynomiaaliset regressiomallit, joissa selittävä muuttuja on muotoa $x_i = x_1^i$ [5]. Tällöin malli saa muodon

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \cdots + \beta_k x_1^k$$

Malli on kertoimien suhteen lineaarinen ja ratkaistavissa edellä mainitulla PNS-menetelmällä. Tällöin normaaliyhtälöt voidaan esittää kaavan 16 mukaisessa matriisimuodossa

$$\underbrace{\begin{bmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \cdots & \sum_{i=1}^n x_i^k \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \cdots & \sum_{i=1}^n x_i^{k+1} \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 & \cdots & \sum_{i=1}^n x_i^{k+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^n x_i^k & \sum_{i=1}^n x_i^{k+1} & \sum_{i=1}^n x_i^{k+2} & \cdots & \sum_{i=1}^n x_i^{2k} \end{bmatrix}}_{\mathbf{X}'\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{\hat{\boldsymbol{\beta}}} = \underbrace{\begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \\ \sum_{i=1}^n y_i x_i^2 \\ \vdots \\ \sum_{i=1}^n y_i x_i^k \end{bmatrix}}_{\mathbf{X}'\mathbf{Y}}$$

Erikoistapaus polynomiaalisesta mallista on neliöllinen malli, jossa $k = 2$ eli pistejoukkoon sovitettava malli on muotoa

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Toinen yleisesti käytetty kertoimien suhteen lineaarinen regressiomalli on pintaa kuvaava malli

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

jossa viimeinen termi on muuttujien x_1 ja x_2 vuorovaikutustermiksi kutsuttu tulo x_1x_2 [5]. Koska polynomit ovat funktioina melko yksinkertaisia ja polynomiaalinen malli on kertoimiensa suhteen lineaarinen, polynomien avulla on helppo kuvata muuttujien välillä olevaa epälineaarista riippuvuutta. Pintaa kuvaavan mallin normaaliyhtälöt voidaan esittää kaavan 16 mukaisessa matriisimuodossa, jossa

$$\mathbf{X'X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}^3 & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i}^2 & \sum_{i=1}^{n} x_{1i}^2x_{2i} \\ \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}^3 & \sum_{i=1}^{n} x_{1i}^4 & \sum_{i=1}^{n} x_{1i}^2x_{2i} & \sum_{i=1}^{n} x_{1i}^2x_{2i}^2 & \sum_{i=1}^{n} x_{1i}^3x_{2i} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{1i}^2x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \sum_{i=1}^{n} x_{2i}^3 & \sum_{i=1}^{n} x_{1i}x_{2i}^2 \\ \sum_{i=1}^{n} x_{2i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i}^2 & \sum_{i=1}^{n} x_{1i}^2x_{2i}^2 & \sum_{i=1}^{n} x_{2i}^3 & \sum_{i=1}^{n} x_{2i}^4 & \sum_{i=1}^{n} x_{1i}x_{2i}^3 \\ \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{1i}^2x_{2i} & \sum_{i=1}^{n} x_{1i}^3x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i}^3 & \sum_{i=1}^{n} x_{1i}^2x_{2i}^2 \end{bmatrix}$$

ja

$$\mathbf{X'Y} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_{1i} \\ \sum_{i=1}^{n} y_i x_{1i}^2 \\ \sum_{i=1}^{n} y_i x_{2i} \\ \sum_{i=1}^{n} y_i x_{2i}^2 \\ \sum_{i=1}^{n} y_i x_{1i} x_{2i} \end{bmatrix}$$

Lisäksi osa regressiomalleista on muutettavissa lineaariseen muotoon. Esimerkiksi eksponentiaalinen malli

$$y = a_0 e^{a_1 x}$$

on muutettavissa lineaariseen muotoon

$$\ln(y) = \ln(a_0) + a_1 x$$

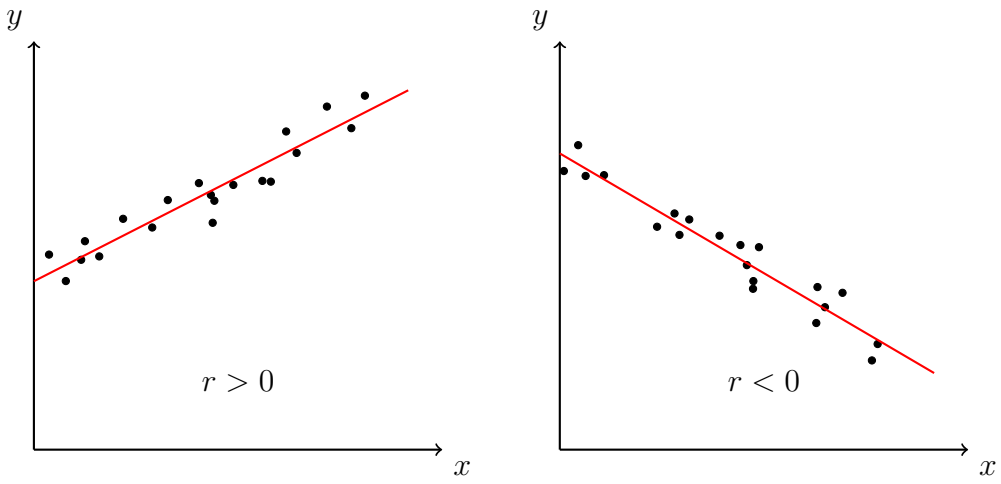
ottamalla alkuperäisestä yhtälöstä luonnollinen logaritmin puolittain. Malli sovitetaan datapisteisiin $(x_i, \ln(y_i))$ ja estimoitavat parametrit ovat vakiotermi $\ln(a_0)$ ja suoran kulmakerroin a_1. Kuten Ayyub ja McCuen toteavat, on tärkeää muistaa, että malli on sovitettu alkuperäisestä poikkeavaan, muunnettuun avaruuteen. Tällöin PNS-menetelmällä määritetyt mallin parametrit minimoivat datapisteiden ja sovitetun mallin eroa vain muunnetussa koordinaatistossa. Esimerkiksi lineaarisen mallin korrelaatiokerroin kuvastaa tilannetta vain muunnetussa koordinaatistossa, vaikka alkuperäisessä koordinaatistossa esitetyn eksponentiaalisen mallin korrelaatiokerroin olisikin usein käytännön kannalta kiinnostavampi. [6] Eksponentiaaliset mallit ovat kuitenkin yleisiä esimerkiksi monissa fysiikan ilmiöissä, joten niille on käyttöä muun muassa fysiikkaan liittyvissä sovelluskohteissa.

2.4.3 Korrelaatio ja residuaalit

Selittävän ja selitettävän muuttujan välistä lineaarista riippuvuutta voidaan mitata korrelaatiokertoimella. Pearsonin tulomomenttikorrelaatiokerroin eli otoskorrelaatiokerroin määritellään yhden selittävän muuttujan tapauksessa kaavalla

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}} \tag{18}$$

ja se saa arvoja väliltä $-1 \leq r \leq 1$. Jos korrelaatiokerroin saa positiivisen arvon, muuttujien välillä on positiivinen riippuvuus. Tämä tarkoittaa, että pieniin selittävän muuttujan arvoihin liittyy pieni selitettävän muuttujan arvo ja suuriin selittävän muuttujan arvoihin suuri selitettävän muuttujan arvo. Vastaavasti korrelaatiokertoimen ollessa negatiivinen myös riippuvuus on negatiivista. Tällöin pieniin selittävän muuttujan arvoihin liittyy suuri selitettävän muuttujan arvo ja suuriin selittävän muuttujan arvoihin pieni selitettävän muuttujan arvo. Kuvassa 1 on havainnollistettu positiivista ja negatiivista korrelaatiota. Jos $r = \pm 1$, havaintopisteet asettuvat samalle suoralle. Korrelaatiokertoimen arvo $r \approx 0$ tarkoittaa, että muuttujien välillä ei ole lineaarista riippuvuutta.



Kuva 1: Vasemmalla esimerkki positiivisesta korrelaatiosta ja oikealla negatiivisesta

Usein sovitettua mallia tutkitaan sen selitysasteen R^2 avulla. Selitysaste saa arvoja väliltä $0 \leq R^2 \leq 1$. Mitä lähempänä selitysaste on arvoa 1, sitä enemmän mallin selittävä muuttuja kuvaa selitettävän muuttujan arvoja. Selitysasteen saadessa arvon 0 sovitettu regressiosuora on vaakasuora, eikä selittävä muuttuja selitä selitettävän muuttujan arvoja [7]. Selitysaste lasketaan kaavalla

$$R^2 = \frac{SSD}{SST} = 1 - \frac{SSE}{SST} \tag{19}$$

jossa

$$SSD = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

on mallineliösumma,

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

on jäännösneliösumma ja

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 = SSD + SSE$$

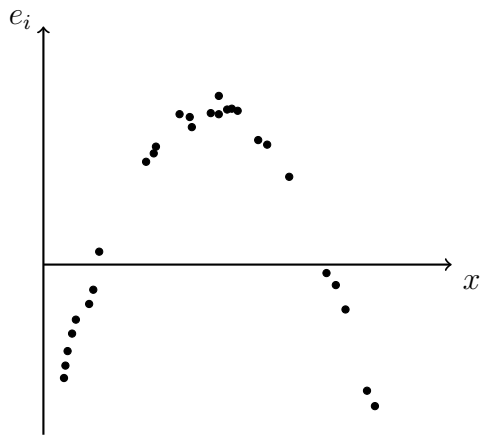
on kokonaisneliösumma. Ylläolevissa kaavoissa $\hat{y}_i$ on sovitteen arvo, y_i datapisteen arvo ja $\overline{y}$ on datapisteiden keskiarvo. Mikäli datapisteen arvon ja sovitteen arvon erotuksen neliö eli jäännösneliösumma on hyvin suuri verrattuna sovitteen arvon ja datapisteiden keskiarvon erotuksen neliöön eli mallineliösummaan, mallin selitysaste on pieni. Vastaavasti mallineliösumman ollessa huomattavasti jäännösneliösummaa suurempi mallin selitysaste on parempi.

Residuaalit e_i määritellään selitettävän muuttujan havaitun arvon ja sovitteen arvon $\hat{y}_i$ erotuksena eli

$$e_i = y_i - \hat{y}_i$$

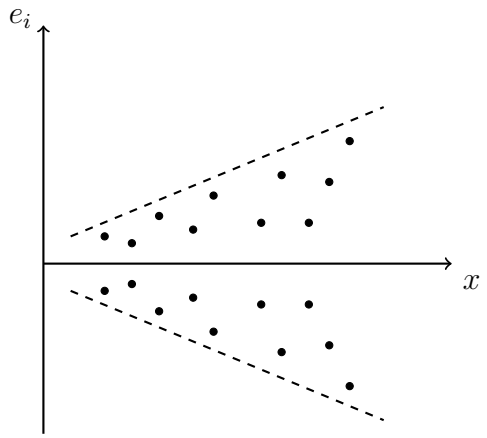
Hayter'n mukaan residuaaleja tutkimalla voidaan tunnistaa käytetystä datasta poikkeavia havaintoja (outlier), varmistaa käytetyn regressiomallin sopivuus kyseiseen tilanteeseen, tutkia, onko jäännösvarianssi vakio sekä selvittää, ovatko jäännöstermit normaalijakautuneita. Yhden selittävän muuttujan tapauksessa residuaalit kannattaa piirtää selittävän muuttujan x funktiona xe_i -koordinaatistoon. Poikkeavia havaintoja tutkittaessa kannattaa keskittyä itseisarvoltaan suuriin residuaaleihin. Niitä vastaavat datapisteet ovat kaukana sovitetusta mallista, joten on syytä pohtia, ovatko kyseiset datapisteet niin poikkeavia havaintoja, että ne kannattaa jättää mallia sovitettaessa datasta kokonaan pois. Tarkemmin poikkeavia havaintoja voisi tutkia jakamalla residuaalit jäännöshajonnalla $\hat{\sigma}$ ja piirtämällä näin saadut arvot muuttujan x funktiona.[5]

Muita Hayter'n esille nostamia residuaalikuvaajiin liittyviä mielenkiinnon kohteita ovat kuvaajiin muodostuvat kuviot. Jos residuaalit ovat ryhmittyneet positiivisiin ja negatiivisiin arvoihin esimerkiksi alaspäin aukeavan paraabelin muotoon kuten kuvassa 2, lineaarinen malli ei ole kyseiseen dataan sopiva. Tällöin regressiomalliksi on valittava jokin epälineaarinen malli.



Kuva 2: Positiivisiin ja negatiivisiin arvoihin ryhmittyneet residuaalit

Mikäli residuaaleja piirrettäessä muodostuu vaakatasossa oleva suppilo kuten kuvassa 3, riippuu residuaalin arvo selittävän muuttujan arvosta. Tällöin oletus, että jäännösvarianssi on vakio, ei pidä paikkaansa. Jäännöstermien normaalijakautuneisuutta voidaan tutkia normaalijakaumakuvion avulla, jossa residuaalit ja niistä lasketut normalisoidut residuaalit esitetään pistepareina koordinaatistossa. Mikäli pisteet muodostavat suoran, jäännöstermit ovat normaalijakautuneita. [5]



Kuva 3: Vaakatasossa olevan suppilon muotoon ryhmittyneet residuaalit

Residuaalien analysoimista sovelletaan myös usean selittävän muuttujan lineaarisessa regressiossa, jossa residuaalit ovat Hayter'n mukaan tärkeä analyysityökalu graafisen arvioimisen ollessa vaikeampaa. Residuaalit piirretään selitettävän muuttujan sovitteen arvon $\hat{y}_i$ funktiona $\hat{y}_i e_i$ -koordinaatistoon sekä jokaisen selittävän muuttujan funktiona $x_{ki} e_i$ - koordinaatistoihin. Näistä kuvaajista tutkitaan residuaalien käyttäytymistä kuten yhden selittävän muuttujan tapauksessa. [5]

3 MENETELMIEN OHJELMOIMINEN MATLABILLA

Datan analysointiin käytettiin Mathworksin Matlab-laskentaohjelmistoa (MATLAB R2016b). Matlabissa ei ole suoraan omaa funktiota χ^2-homogeenisuustestille, joten menetelmä ohjelmoitiin Matlabilla itse. Toteutus oli tekstipohjainen eli siinä ei ollut erillistä käyttöliittymää. Syötteet ja tulosteet toteutettiin komentoikkunan kautta. Käyttäjä syöttää testin riskitason sekä käytettävän havaintoaineiston valmiiksi ristiintaulukoituna matriisina, kuitenkin ilman rivi- tai sarakesummia. Ohjelma laskee annetusta matriisista rivi- ja sarakesummat, odotetut frekvenssit, testisuureen arvon sekä testin p-arvon. Tämän jälkeen testataan nollahypoteesia ja tulostetaan komentoikkunaan testin tulos sekä p-arvo. Lopuksi tehdään tarkastus testin pätevyydestä eli tarkistetaan, että korkeintaan 20% odotetuista frekvensseistä on alle 5 ja että yksikään odotetuista frekvensseistä ei ole alle 1. Mikäli odotetuista frekvensseissä löytyy liian pieniä arvoja, tulostaa ohjelma komentoikkunaan huomautuksen asiasta.

Myös Kruskal-Wallis -testi ohjelmoitiin Matlabilla itse. Matlabissa on olemassa valmis *kruskalwallis*-niminen funktio, jolla voidaan tehdä testi syötteenä annetulle matriisille. Testi antaa tuloksena mm. testin p-arvon sekä ANOVA-taulukon (analysis of variance, varianssianalyysi). Kruskal-Wallis -testin oma toteutus oli tekstipohjainen kuten χ^2- homogeenisuustestikin. Ohjelma kysyy käyttäjältä testin riskitason sekä sen tiedoston nimen, missä olevalle datalle testi tehdään. Testi tutkii taulukoidun datan sarakkeiden jakaumien samanlaisuutta, joten tämä tulee huomioida muokattaessa dataa testiä varten. Sarakkeissa ei tarvitse olla yhtä paljon alkioita. Ohjelma järjestelee datan testin tarvitsemaan muotoon, laskee järjestyslukujen mukaiset keskiarvot jokaiselle ryhmälle ja määrittää testisuureen arvon. Lopuksi ohjelma laskee testin p-arvon, vertaa sitä annettuun riskitasoon ja tulostaa sekä p-arvon että testin tuloksen komentoikkunaan.

Fisherin nelikenttätestille löytyy myös valmis funktio Matlabissa. Sillä voi tehdä testin 2×2-kokoiselle matriisille. Oletusarvona riskitasolle käytetään arvoa 0.05, mutta käyttäjä voi vaihtaa sitä halutessaan. Testin tulos on joko 0 tai 1, jotka viittaavat nollahypoteesin hyväksymiseen tai hylkäämiseen. Testi kertoo lisäksi käyttäjälle mm. laskemansa p-arvon, mutta käyttäjän täytyy itse määritellä se tulosteeksi. Tässä työssä Fisherin nelikenttätestikin toteutettiin itse Matlabilla ohjelmoiden. Testistä tehtiin kahden aiemman testin kanssa samantyylinen eli se on komentoikkunapohjainen, käyttäjän täytyy syöttää itse testin riskitaso sekä tutkittava data 2×2-kokoisessa matriisissa ja testin tulos tulostetaan komentoikkunaan. Testissä lasketaan ensin rivi- ja sarakesummat. Sen jälkeen lasketaan tarvittavat kertomat ja niiden avulla määritetään p_{cutoff}-arvo. Mikäli matriisin alkioiden arvot ovat suuria, niiden kertomat ja kertomien tulot ovat isoja lukuja, ja p_{cutoff}-arvon laskeminen ei onnistu suoraan yhdellä lausekkeella. Tulomuotoisen lausekkeen takia p-arvot voidaan laskea osissa jakamalla lause-

ke useaan pienempään jakolaskuun ja kertomalla niiden tulokset keskenään. Näin vältetään suurista luvuista aiheutuvat ongelmat.

Etsittäessä muita matriiseja, joilla on samat rivi- ja sarakesummat kuin alkuperäisellä matriisilla, huomattiin testin monimutkaisuus matriisin ollessa 2×2-kokoista suurempi. 2×2 -kokoisessa matriisissa kasvattamalla vaakarivillä toisen alkion arvoa yhdellä ja vähentämällä toisen alkion arvoa yhdellä rivisumma pysyy samana ja muuttamalla vastaavasti pystyrivin alkioiden arvoja saadaan sarakesummat pysymään vakioina. Tätä suuremmissa matriiseissa mahdollisia vaihtoehtoja olisi jo huomattavasti enemmän ja kaikkien mahdollisten matriisien löytäminen vaatisi paljon enemmän työtä. Suuremmille matriisille p-arvojen laskemisen pitäisi kuitenkin onnistua jakamalla lauseke pienempiin osiin kuten 2×2 -kokoisessa matriisissa. Tällöin matriisin alkioiden arvojen on kuitenkin oltava sen verran pieniä, että käytettävän laskentaohjelmiston laskutarkkuus riittää p-arvon laskemisessa tarvittavien kertomien laskemiseen. Fisherin nelikenttätesti toteutettiin siis nimensä mukaisesti vain 2×2-kokoiselle matriisille. Kun p-arvot on saatu laskettua muillekin kuin alkuperäiselle matriisille, lasketaan p_{sum}-arvo ja tehdään päätelmä testin lopputuloksesta. Lopuksi tieto nollahypoteesin hylkäämisestä tai hyväksymisestä sekä testin p-arvo eli p_{sum}-arvo tulostetaan komentoikkunaan käyttäjän nähtäville.

Regressioanalyysiä varten ohjelmoitiin Matlabilla yksinkertainen ohjelma, jolla voi sovittaa dataan joko lineaarisen yhden selittävän muuttujan mallin tai neliöllisen mallin $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Tämäkin ohjelma on tekstipohjainen ja tulokset ilmoitetaan sekä kuvaajien että komentoikkunan avulla. Ensin käyttäjä syöttää käytettävän datan sisältävän tiedoston nimen ja valitsee, kumpaa mallia haluaa käyttää. Tämän jälkeen ohjelma laskee mallin parametrien arvot datan perusteella, määrittää mallin selitysasteen sekä laskee residuaalit. Ohjelma piirtää kuvaajat sekä datapisteistä ja niihin sovitetusta mallista että residuaaleista. Selitysaste ja mallin parametrien arvot tulostetaan käyttäjän näkyville komentoikkunaan.

4 KÄYTETTÄVÄ DATA JA SEN ANALYSOIMINEN

4.1 Datan kuvaus

Työssä oli käytettävissä dataa kahtena peräkkäisenä vuonna DIA-yhteisvalinnassa LUT:iin opiskelemaan valittujen henkilöiden sisäänpääsytavasta, -pisteistä sekä opintopisteiden kertymisestä. Näistä kahdesta vuodesta käytetään tässä työssä nimiä sisäänpääsyvuosi 1 ja sisäänpääsyvuosi 2. Suoritetuista opintopisteistä kertova data kuvasi Weboodiin kirjattua opintopistemäärää keväällä neljä vuotta sisäänpääsyvuoden 1 jälkeen. Käytettävä data oli alun perin taulukkomuodossa ja sen muokkaamiseen laskentaan soveltuvaan muotoon käytettiin Microsoft Exceliä.

Opiskelijoiden opintopistekertymää tarkasteltiin kolmen opiskeluvuoden jälkeen. Datassa näkyneet ensimmäisen tai toisen vuosikurssin opiskelijat rajattiin siis pois. Tutkimuksessa ei huomioitu kirjoilta poistettujen opiskelijoiden opintopistekertymiä. Lisäksi huomiotta jätettiin sellaiset opiskelijat, joilla opintopistekertymä oli usean opiskeluvuoden jälkeen nolla Weboodissa.

Opintopistekertymät on esitetty valintaryhmittäin liitteessä 1 olevissa kuvaajissa. Koepisteiden perusteella opiskelemaan valituista on noin 30 havaintopistettä kumpanakin tutkittavana vuonna, yhteispisteillä valituista vuosittaisia havaintoja on noin 60 ja todistusvalinnalla valituista on myös noin 60 havaintoa vuodessa.

4.2 Datan analysoiminen ja tulokset

Työssä tutkittiin χ^2- homogeenisuustestillä, löytyykö eri valintatavoilla opiskelemaan valittujen välillä eroa opiskelujen etenemisessä. Kumpaakin sisäänpääsyvuotta tarkasteltiin erikseen ja datasta muodostettiin vuosikohtaiset taulukot, joissa opintopistekertymä jaettiin neljään eri kategoriaan ja valintatavat muodostivat kolme eri kategoriaa. Ristiintaulukoimalla saadut taulukot löytyvät liitteestä 2.

Testeistä saadut p-arvot on esitetty taulukossa 4. Käytettäessä riskitasona $\alpha = 0.05$, nollahypoteesi jäi voimaan kummankin tarkasteltavan vuoden kohdalla.

Taulukko 4: χ^2-homogeenisuustestistä saadut p-arvot

	p-arvo
Sisäänpääsyvuosi 1	0.50976
Sisäänpääsyvuosi 2	0.91514

Opintopistekertymäjakaumien samanlaisuutta eri valintaryhmissä tutkittiin myös Kruskal-Wallis -testillä. Testi käytti syötteenä Excel-tiedostoa, johon opintopistemäärät oli ryhmitelty valintaryhmittäin ja vuosittain. Testistä saadut p-arvot on esitetty taulukossa 5. Valittaessa riskitasoksi $\alpha = 0.05$ nollahypoteesi jäi voimaan kummankin vuoden kohdalla. Tämän vahvistaa χ^2-homogeenisuustestillä saatua tulosta, jonka mukaan kertyneiden opintopisteiden jakaumat ovat samanlaisia DIA-yhteisvalinnassa todistusvalinnalla, yhteispisteiden perusteella ja pelkkien koepisteiden perusteella opiskelemaan valittujen keskuudessa.

Taulukko 5: Kruskal-Wallis -testistä saadut p-arvot

	p-arvo
Sisäänpääsyvuosi 1	0.49011
Sisäänpääsyvuosi 2	0.52822

Opintopistekertymiä tutkittiin koko opiskelijajoukon lisäksi schooleittain Kruskal-Wallis -testillä. Schoolit ovat LUT:n yksiköitä, jotka keksittyvät omaan osaamisalueeseensa ja tarjoavat siihen liittyvää opetusta. LUT:n schoolit ovat LUT School of Business and Management (LBM), LUT School of Engineering Science (LENS) ja LUT School of Energy Systems. Jokainen schooli tekee tutkimusta omalla osaamisalueellaan ja tarjoaa kandidaattivaiheen opetusta 2-4 koulutusohjelmassa. Testistä saadut p-arvot on esitetty taulukossa 6. Testin perusteella opintopistekertymien jakaumissa ei ollut tilastollisesti merkittäviä eroja riskitasolla $\alpha = 0.05$ kumpanakaan tarkasteltavana vuonna missään LUT:n kolmesta schoolissa. Tämän perusteella myös schoolitasolla tarkasteltuna opinnot etenevät samaa vauhtia kaikissa valintaryhmissä.

Taulukko 6: Schoolikohtaisesta Kruskal-Wallis -testistä saadut p-arvot

	LBM	LENS	LES
Sisäänpääsyvuosi 1	0.80184	0.29812	0.31536
Sisäänpääsyvuosi 2	0.52668	0.90903	0.66195

Eroja kirjoiltapoistettujen ja kirjoilla olevien opiskelijoiden määrissä tutkittiin Fisherin nelikenttätestillä. Testiä varten laskettiin kirjoiltapoistettujen sekä kirjoilla olevien eli joko läsnä- tai poissaoleviksi merkittyjen opiskelijoiden lukumäärät. Luvut laskettiin erikseen todistusvalinnalla, yhteispisteillä ja koepisteillä opiskelemaan valittujen keskuudessa. Jakaumien samankaltaisuutta ei voitu tutkia χ^2- homogeenisuustestillä, sillä osa odotetuista frekvensseistä oli liian pieniä. Sen sijaan aineisto päätettiin jakaa useaan 2×2-kokoiseen matriisiin, joista voitiin tutkia Fisherin nelikenttätestillä suoraan, onko esimerkiksi todistuspisteillä ja koepisteillä opiskelemaan valittujen keskuudessa eroja kirjoilla olevien ja kirjoiltapoistettujen jakaumissa. Taulukoitu aineisto löytyy liitteestä 3. Testistä saadut p-arvot on esitetty taulukossa 7. Testin perusteella jakaumissa ei ole eroja riskitasolla $\alpha = 0.05$.

Taulukko 7: Fisherin nelikenttätestin p-arvot

	p-arvo
Sisäänpääsyvuosi 1, todistusvalinta ja yhteispisteet	1.00000
Sisäänpääsyvuosi 1, todistusvalinta ja koepisteet	0.36802
Sisäänpääsyvuosi 1, yhteispisteet ja koepisteet	0.33772
Sisäänpääsyvuosi 2, todistusvalinta ja yhteispisteet	0.31609
Sisäänpääsyvuosi 2, todistusvalinta ja koepisteet	0.35026
Sisäänpääsyvuosi 2, yhteispisteet ja koepisteet	1.00000

Opintopistekertymän riippuvuutta sisäänpääsypisteistä eli todistus- ja valintakoepisteistä tutkittiin regressioanalyysillä. Liitteen 1 mukaisiin datajoukkoihin sovitettiin ensin yhden selittävän muuttujan malli $y = \beta_0 + \beta_1 x$, jossa opintopistekertymä on selitettävä muuttuja ja sisäänpääsypisteet selittävä muuttuja. Sovitettujen mallien yhtälöt sekä sovitteiden selitysasteet ja otoskorrelaatiokertoimet on esitetty taulukossa 8. Lisäksi tutkittiin residuaaleja, jotka on kuvattu datajoukkokohtaisesti xe_i-koordinaatistossa liitteessä 4.

Taulukko 8: Regressioanalyysin tulokset mallille $y = \beta_0 + \beta_1 x$

	Sovitetun mallin yhtälö	R^2	r
Sisäänpääsyvuosi 1, todistusvalinta	$y = 196.229 - 1.032x$	0.0050	- 0.0709
Sisäänpääsyvuosi 1, yhteispisteet	$y = 43.662 + 4.008x$	0.2166	0.4654
Sisäänpääsyvuosi 1, koepisteet	$y = 154.485 + 1.495x$	0.0174	0.1318
Sisäänpääsyvuosi 2, todistusvalinta	$y = 72.810 + 4.213x$	0.0382	0.1955
Sisäänpääsyvuosi 2, yhteispisteet	$y = 100.816 + 1.655x$	0.0459	0.2141
Sisäänpääsyvuosi 2, koepisteet	$y = 120.043 + 1.3995x$	0.0074	0.0859

Datajoukkoihin sovitettiin testiksi myös neliöllinen malli $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Sovitettujen mallien yhtälöt ja sovitteiden selitysasteet on esitetty taulukossa 9.

Taulukko 9: Regressioanalyysin tulokset mallille $y = \beta_0 + \beta_1 x + \beta_2 x^2$

	Sovitetun mallin yhtälö	R^2
Sisäänpääsyvuosi 1, todistusvalinta	$y = 216.051 - 2.982x + 0.047x^2$	0.0051
Sisäänpääsyvuosi 1, yhteispisteet	$y = -233.338 + 20.926x - 0.254x^2$	0.2316
Sisäänpääsyvuosi 1, koepisteet	$y = 22.296 + 15.900x - 0.377x^2$	0.0276
Sisäänpääsyvuosi 2, todistusvalinta	$y = -18.554 + 13.202x - 0.216x^2$	0.0397
Sisäänpääsyvuosi 2, yhteispisteet	$y = 192.583 - 4.155x + 0.088x^2$	0.0548
Sisäänpääsyvuosi 2, koepisteet	$y = 287.444 - 21.591x + 0.747x^2$	0.0423

Tutkimuksen perusteella x^2-termin lisääminen ei paranna mallin selitysastetta kovin paljoa, joten pelkkä yhden selittävän muuttujan mallin riittää kyseessä oleville datajoukoille. Kuten PennState Eberly College of Sciencen selitysastetta käsittelevällä sivulla sanotaan, selitysasteen avulla voidaan sanoa, kuinka monta prosenttia y:n vaihtelusta voidaan selittää x:n avulla. Tämä ei kuitenkaan tarkoita sitä, että x aiheuttaisin y:n vaihtelun. Lisäksi se, mitä selitysasteen arvoa voidaan pitää suurena, riippuu tutkittavasta asiasta. Ihmisen käyttäytymistä tutkittaessa 30% on jo suuri selitysaste, kun taas insinööritieteiden puolella 30% on varsin pieni arvo selitysasteelle. [7] Tarkasteltaessa yhden selittävän muuttujan mallia vain yhdessä tutkittavista tapauksista selitysaste on yli 20%. Muissa tapauksissa sisäänpääsypisteet selittävät korkeintaan muutaman prosentin opintopistekertymän vaihtelusta.

Yhden selittävän muuttujan mallin residuaalikuvaajia tutkittaessa kuvaajista ei erottunut mitään huomiota herättäviä muotoja, vaan havainnot keskittyvät tasapaksulle alueelle. Kuvaajista havaittiin kuitenkin muutama itseisarvoltaan suuri residuaalin arvo. Sisäänpääsyvuoden 1 todistusvalinnan datan kuvaajasta löytyy yksi selvästi muita suurempi opintopistekertymä ja yksi selvästi muita pienempi opintopistekertymä. Sisäänpääsyvuoden 2 vastaavasta kuvaajasta löytyy yksi selvästi muita suurempi opintopistekertymä. Nämä havaintoarvot poistettiin tutkittavasta datasta ja yhden selittävän muuttujan regressiomalli sovitettiin näin saatuun dataan. Mallien yhtälöt, sovitteiden selitysasteet ja otoskorrelaatiokertoimet on esitetty taulukossa 10. Sovitettujen suorien kulmakertoimet ja selitysasteet eivät muutu kovin paljoa, vaikka yksittäiset selvästi muista eroavat datapisteet poistetaan datasta. Kuvaajia kannattaa siis tutkia residuaalianalyysin lisäksi muilla keinoilla.

Taulukko 10: Regressioanalyysin tulokset muokatulle datalle

	Sovitetun mallin yhtälö	R^2	r
Sisäänpääsyvuosi 1, todistusvalinta	$y = 190.491 - 0.744x$	0.0048	- 0.0696
Sisäänpääsyvuosi 2, todistusvalinta	$y = 69.425 + 4.160x$	0.0517	0.2275

Liitteen 1 kuvaajia tarkasteltaessa huomataan, että sisäänpääsyvuonna 1 pieniä opintopistekertymiä löytyy niin pieniltä kuin suuriltakin valintapisteiltä. Sisäänpääsyvuonna 2 pieniä opintopistekertymiä löytyy kuvaajien perusteella enemmän pienemmiltä valintakoepisteiltä. Toisaalta myös suurin osa datapisteistä löytyy näistä kuvaajista pienempien valintapisteiden päästä. Jos vähän opintopisteitä suorittaneiden opiskelijoiden määrää pyrittäisiin karsimaan nostamalla valintapisteiden rajaa, myös suuri osa paljon opintopisteitä suorittaneista opiskelijoista rajautuisi valinnan ulkopuolelle. Jos sisäänpääsyvuonna 1 yhteispisteillä opiskelemaan valittuja esittävästä kuvaajasta jätetään muutama alhaista opintopistekertymää kuvaava datapiste huomioimatta poikkeavina havaintoina, vaikuttaa siltä, että paremmilla valintapisteillä myös opintopistekertymä on suurempi. Muissa valintaryhmissä vastaavaa yhteyttä ei näytä olevan, vaan opintopistekertymät ovat melko samanlaisia sisäänpääsypisteistä riippumatta.

5 JOHTOPÄÄTÖKSET JA POHDINTA

DIA-yhteisvalinnassa opiskelemaan valittujen opintopistekertymiä tutkittiin ensin eri sisäänpääsytapojen eli todistusvalinnan, yhteispisteiden sekä koepisteiden perusteella valittujen välillä. Opintopistekertymissä ei löytynyt eroja valintatapojen väliltä. Vertailtaessa opintopistekertymiä schoolikohtaisesti eri sisäänpääsytapojen välillä ei jakaumista löytynyt eroja. Tämän perusteella yliopiston ei kannata keskittyä erityisesti johonkin tiettyyn valintaryhmään tavoitellessaan mahdollisimman nopeasti opintojaan suorittavia opiskelijoita.

Kirjoiltapoistettujen ja kirjoilla olevien jakaumia tutkittaessa ei ilmennyt eroja valintatapojen välillä. Tämän perusteella myöskään opintonsa keskeyttäneiden määrä ei kannusta keskittymään erityisesti johonkin kolmesta valintaryhmästä.

Tutkittaessa selittääkö valintapisteiden määrä opiskelijan opintopistekertymää todettiin, että valintapisteet selittävät keskimäärin vain muutaman prosentin opintopistekertymän vaihtelusta. Tämän perusteella paremmat sisäänpääsypisteet eivät selitä opintojen nopeaa etenemistä. Tällöin sisäänpääsyrajojen nostaminen vähentäisi huomattavasti myös nopeasti opintojaan suorittavien opiskelijoiden lukumäärää.

Tässä tutkimuksessa ei otettu huomioon opintopisteiden suoritustapaa. Esimerkiksi muualla suoritettuja opintopisteitä, jotka on hyväksiluettu LUT:ssa suoritettavaan tutkintoon, käsiteltiin samalla tavalla kuin LUT:ssa suoritettuja opintopisteitä. Käytössä olleen datan perusteella ei voitu tutkia opiskelijoiden koko opiskeluajalta vain LUT:ssa suorittamia opintopisteitä, sillä hyväksiluettujen kurssien määrä oli saatavissa vain kahdelta viimeiseltä lukukaudelta. Lisäksi tässä työssä ei huomioitu sitä, millaisilla arvosanoilla opintopisteitä oli suoritettu.

Mikäli itse ohjelmoituja menetelmiä käytetään jatkossa enemmän, voisi ohjelmia kehittää ulkoasun, virheentarkastuksen, menetelmien yksityiskohtien ja tehokkuuden osalta. Nyt ohjelmoidessa pääpaino oli itse menetelmässä ja ohjelman soveltuvuudessa käytössä olevalle datalle. Lisäksi erillisten funktioiden toteuttaminen useasti käytetyille samankaltaisille rakenteille tekisi koodista selkeämpää.

Muita aiheeseen liittyviä asioita on esimerkiksi se, löytyykö yksittäisten kurssikokonaisuuksien, kuten matematiikan tai fysiikan peruskurssien, suorittaneiden ja sisäänpääsypisteiden väliltä yhteyttä. Lisäksi jatkossa voitaisiin tutkia, onko jossain valintaryhmässä selvästi enemmän sellaisia opiskelupaikan saaneita opiskelijoita, jotka eivät koskaan opintojaan aloita. Myös eri sisäänpääsyvuosien laajempi vertailu keskenään ja todistusvalinnan valintapisteiden vertailu muiden yliopistojen valintapisteisiin voisi tuoda lisää tietoa asiasta.

6 YHTEENVETO

Kandidaatintyön tavoitteena oli tutkia tilastollisin menetelmin, löytyykö DIA- yhteisvalinnalla opiskelemaan valittujen opiskelijoiden sisäänpääsytavan ja opintojen etenemisen väliltä yhteyttä Lappeenrannan teknillisessä yliopistossa. Koska yliopiston rahoitus perustuu osittain opiskelijoiden suorittamien opintopisteiden lukumäärään, yliopistolle on taloudellisestikin tärkeää, että opiskelijoiden opinnot etenevät hyvin.

Työssä tutkittiin kahden vuoden sisäänpääsypisteitä ja kyseisten opiskelijoiden opintopistekertymää kolmen vuoden opiskelun jälkeen. Opintopistekertymäjakaumien samanlaisuutta eri valintaryhmissä tutkittiin χ^2-homogeenisuustestillä sekä Kruskal-Wallis -testillä. Opintopistekertymäjakaumia tutkittiin myös schoolikohtaisesti. Testien perusteella opintojen etenemisessä ei ole eroja eri valintatapojen välillä, joten yliopiston ei kannata keskittyä erityisesti johonkin kolmesta valintaryhmästä hyvin etenevien opintojen takia.

Eri valintaryhmissä kirjoiltapoistettujen ja kirjoilla olevien jakaumista ei löytynyt eroja Fisherin nelikenttätestillä. Valintapisteiden ja opintopistekertymän välistä yhteyttä selvitettiin puolestaan regressioanalyysillä. Yhden selittävän muuttujan lineaarisen mallin mukaan valintapisteet selittävät pääsääntöisesti vain muutaman prosentin opintopistekertymän vaihtelusta.

Testit toteutettiin Matlabilla itse ohjelmoidulla ohjelmilla. Ohjelmoidessa keskityttiin itse menetelmiin sekä ohjelman soveltuvuuteen käytössä olevan datan analysointiin. Jatkokäyttöä ajatellen ohjelmia voisi kehittää muun muassa paremmilla virheentarkastuksilla ja tehokkaammilla laskentamenetelmillä.

Etsittäessä opintopistekertymään vaikuttavia asioita kannattaa jatkossa kiinnittää huomiota muuhunkin kuin valintatapaan ja sisäänpääsypisteisiin. Lisäksi olisi mielenkiintoista selvittää, onko valintatavalla ja esimerkiksi matematiikan peruskurssien suorituksilla jotain yhteyttä.

LÄHDELUETTELO

[1] Opetus- ja kulttuuriministeriön asetus yliopistojen perusrahoituksen laskentakriteereistä 331/2016. Viitattu 12.10.2017. Saatavissa https://www.finlex.fi/fi/laki/alkup/2016/20160331

[2] Yliopistolaki 558/2009 § 36. Ajantasainen lainsäädäntö. Viitattu 12.10.2017. Saatavissa https://www.finlex.fi/fi/laki/ajantasa/2009/20090558#L5P36

[3] Bland M. 2001. An Introduction to Medical Statistics 3 Edition. UK: Oxford University Press. Viitattu 11.2.2018. Saatavissa https://books.google.fi/books?hl=fi&lr=&id=fKgXCgAAQBAJ&oi=fnd&pg=PP1&ots=Ew1UqKJZio&sig=xwW8E1mg9T8tX-zgj9IjSoYCcAk&redir_esc=y#v=onepage&q&f=false

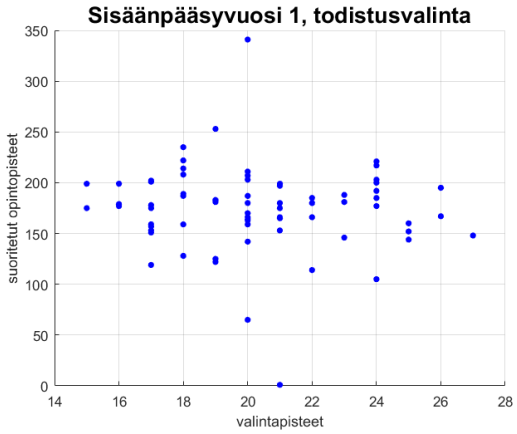
[4] Weisstein, E. W. Fisher's Exact Test. MathWorld – A Wolfram Web Resource. Viitattu 8.2.2018. Saatavissa http://mathworld.wolfram.com/FishersExactTest.html

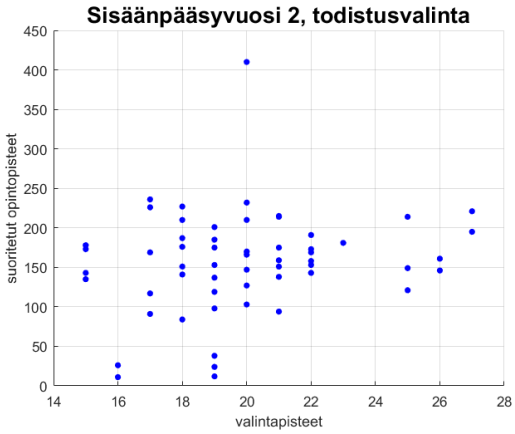
[5] Hayter A. 2013. Probability and Statistics for Engineers and Scientists, 4th edition, international edition. Canada: Brooks/Cole, Cengage Learning

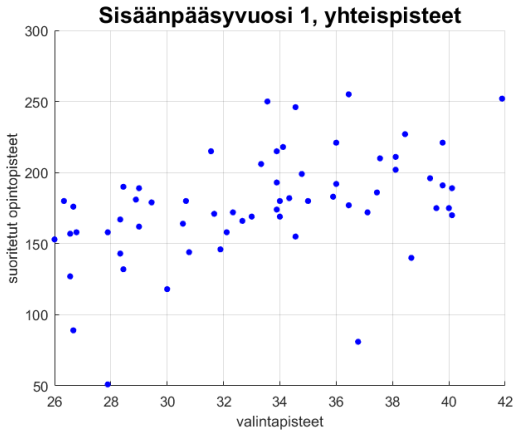
[6] Ayyub, B. M. & McCuen R. H. 1997. Probability, Statistics, & Reliability for Engineers. United States: CRC Press

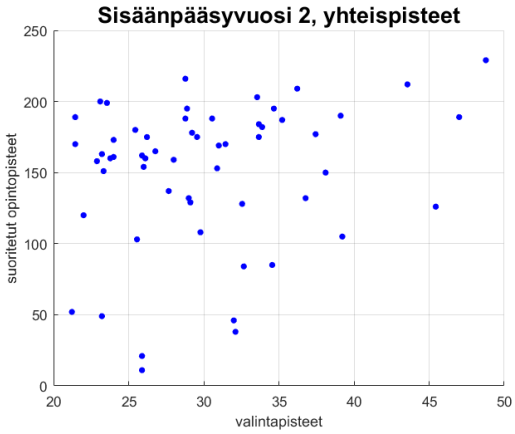
[7] PennState, Eberly College of Science, STAT 501. 1.5 The Coefficient of Determination, r-squared, 2018. Viitattu 17.3.2018. Saatavissa https://onlinecourses.science.psu.edu/stat501/node/255

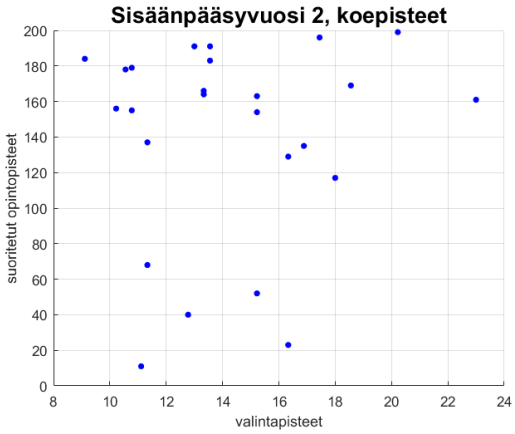
Liite 1: Kuvaajat opintopisteistä valintaryhmittäin esitettynä











Liite 2: Opintopistekertymät ristiintaulukoituna χ^2 -homogeenisuustestiin

Sisäänpääsyvuosi 1	1-140 op	141-160 op	161-180 op	181 - op
Todistusvalinta	8	13	19	34
Yhteispisteet	7	9	20	26
Koepisteet	6	4	4	16

Sisäänpääsyvuosi 2	1-140 op	141-160 op	161-180 op	181 - op
Todistusvalinta	17	12	12	17
Yhteispisteet	18	8	14	17
Koepisteet	9	3	7	6

Liite 3: Taulukot kirjoilla olevien ja kirjoiltapoistettujen lukumääristä

Sisäänpääsyvuosi 1	Kirjoilla olevia	Kirjoiltapoistettuja
Todistusvalinta	74	3
Yhteispisteet	62	2

Sisäänpääsyvuosi 1	Kirjoilla olevia	Kirjoiltapoistettuja
Todistusvalinta	74	3
Koepisteet	31	3

Sisäänpääsyvuosi 1	Kirjoilla olevia	Kirjoiltapoistettuja
Yhteispisteet	62	2
Koepisteet	31	3

Sisäänpääsyvuosi 2	Kirjoilla olevia	Kirjoiltapoistettuja
Todistusvalinta	69	3
Yhteispisteet	62	6

Sisäänpääsyvuosi 2	Kirjoilla olevia	Kirjoiltapoistettuja
Todistusvalinta	69	3
Koepisteet	26	3

Sisäänpääsyvuosi 2	Kirjoilla olevia	Kirjoiltapoistettuja
Yhteispisteet	62	6
Koepisteet	26	3

Liite 4: Kuvaajat yhden selittävän muuttujan regression residuaaleista