Valmistumisaikaan vaikuttavat tekijät - elinaika- ja ryhmittelyanalyysin sovellus

Heli Hästbacka

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
Kevät 2021

Heli Hästbacka: *Valmistumisaikaan vaikuttavat tekijät - elinaika- ja ryhmittelyanalyysin sovellus*
Tilastotieteen pro gradu -tutkielma, 41 s.
Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
Kevät 2021

Tiivistelmä

Tämän tutkielman tarkoituksena on mallintaa Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan opiskelijoiden valmistumisaikoja ryhmittelyanalyysin ja elinaika-analyysin keinoin. Valmistumisaikana tarkastellaan aikaa opintojen aloittamisesta luonnontieteiden kandidaatiksi valmistumiseen saakka. Valmistumisaikaa pyritään selittämään ylioppilasarvosanojen ja muiden hakuvaiheen muuttujien avulla.

Aineistona tarkastellaan Jyväskylän yliopiston matemaattis-luonnontieteellisessä tiedekunnassa vuosina 2015-2020 aloittaneiden opiskelijoiden tietoja. Aineisto koostuu kahdesta osa-aineistosta: hakuvaiheen aineistosta ja opiskeluvaiheen aineistosta. Hakuvaiheen aineistosta käy ilmi muiden muassa ylioppilasarvosanat, valintajono sekä hakukohde. Opiskeluvaiheen aineistossa on tietoja esimerkiksi opiskelijan laitoksesta, opintopistekertymästä, läsnäolosta sekä kandidaatiksi valmistumisen ajankohdasta. Edellä mainittujen tietojen perusteella muodostettiin uusi muuttuja, joka kertoo läsnäololukukausien määrän opintojen aloittamisesta kandidaatiksi valmistumiseen saakka. Tavoiteaika kandidaatin tutkinnolle matemaattis-luonnontieteellisessä tiedekunnassa on kolme lukuvuotta eli kuusi lukukautta.

Tutkimusmenetelminä ovat ryhmittelymenetelmistä k:n keskiarvon ryhmittely sekä elinaika-analyysin menetelmistä Coxin regressiomalli. Perinteisen Coxin mallin sovittaminen osoittautui haastavaksi ylioppilasarvosanoissa ilmenevän ilmiöstä johtuvan puuttuvan tiedon vuoksi. Näin ollen aineistoon sovellettiin ryhmittelyanalyysiä ylioppilasarvosanatiedon tiivistämiseksi. Ryhmittelyyn valikoitiin ne ylioppilasaineet, joissa kirjoittaneiden osuus aineistossa on vähintään 10 prosenttia. Muodostettuja ryhmiä käytettiin Coxin regressiomallissa valmistumisajan selittäjinä.

Tutkimustuloksista käy ilmi, että ylioppilasarvosanojen perusteella muodostettujen ryhmien välillä on eroja kandidaatiksi valmistumisessa. Ryhmä, jossa on korkeat ylioppilasarvosanojen keskiarvot äidinkielestä, biologiasta ja maantieteestä, valmistuu ryhmien välisessä vertailussa nopeimmin. Puolestaan ryhmä, jossa on matalin keskiarvo äidinkielestä ja korkea keskiarvo matematiikasta, valmistuu hitaimmin. Ryhmämallia parannettiin lisäämällä malliin tieto valintajonosta. Mallin tuloksista ilmeni, että suoravalinnan kautta tulleet opiskelijat valmistuvat hitaimmin. Yhteishaun ulkopuolelta tulleiden opiskelijoiden valmistuminen on puolestaan kaikista nopeinta suhteessa suoravalittuihin opiskelijoihin.

Avainsanat: Coxin regressiomalli, kandidaatiksi valmistuminen, k:n keskiarvon ryhmittely, läsnäololukukaudet, ylioppilasarvosanat, valintajono

Sisältö

1 Johdanto

Vuodesta 2020 alkaen suurin osa Suomen korkeakouluopiskelijoista on valittu pelkän ylioppilastodistuksen perusteella. Ylioppilastutkintoaineiden painotukset ja pisteytettävien aineiden määrä vaihtelee koulutuksittain.

Korkeakoulujen valintauudistuksen seurauksena myös Jyväskylän yliopiston matemaattis-luonnontieteellisessä tiedekunnassa todistuksella valittavien osuudet muuttuivat vuodesta 2020 alkaen (Opintopolku, 2019). Biologian koulutusohjelmaan valituista opiskelijoista 51 % valitaan jatkossa ylioppilastodistuksen perusteella, kun taas luonnonvarat ja ympäristö -koulutukseen todistuksella valittavien osuus on 85 %. Fysiikan koulutusohjelmaan (fyysikko tai fysiikan aineenopettaja) todistuksella valittavien osuus on vähintään 70 % hakijoista. Tämä osuus sisältää myös suoravalinnan tilanteessa, jossa hakija on suorittanut fysiikan ylioppilaskokeen vähintään arvosanalla *eximia cum laude approbatur* (E). Kemian koulutusohjelmaan (kemisti tai kemian aineenopettaja) todistuksella valittavien osuus on vähintään 70 %, sisältäen suoravalinnan kemian ylioppilaskokeen arvosanan ollessa vähintään E. Matemaattisissa tieteissä, joihin kuuluvat matematiikan ja tilastotieteen koulutusohjelma sekä matematiikan aineenopettajakoulutus, todistuksella valittavien osuus on vähintään 70 % hakijoista. Aineenopettajakoulutukseen hakevilla on todistusvalinnan lisäksi soveltuvuuskoe. Todistuksella valittavien osuus sisältää myös matemaattisten tieteiden kohdalla suoravalinnan, mikäli hakijalla on pitkän matematiikan ylioppilaskokeesta vähintään arvosana E. Kaikissa matemaattis-luonnontieteellisen tiedekunnan koulutuksissa annetaan pisteitä äidinkielestä, matematiikasta sekä yhdestä tai kahdesta hakijalle parhaat pisteet tuottavasta aineesta.

Jatkossa suurin osa yliopistoon tulevista opiskelijoista valitaan ylioppilastodistuksen perusteella. Sekä yliopistoille että opiskelijoille olisi eduksi, mikäli valitut opiskelijat opiskelisivat tehokkaasti ja valmistuisivat tavoiteajassa, ja sitä ennustaisi aikaisempi menestyminen ylioppilaskirjoituksissa. Kandidaatintutkinnolle tavoiteaika on kolme lukuvuotta ja maisterintutkinnolle kaksi lukuvuotta (Yliopistolaki 558/2009 §40). Yliopistolaissa määritellyssä tavoiteajassa tutkinnon suorittaa vain alle kolmannes kaikista tutkinnon suorittajista (Opetus- ja kulttuuriministeriö, 2018).

Opiskelijavalintaa, valmistumista sekä opintomenestystä on tutkittu jonkin verran sekä Suomessa että ulkomailla. Aikaisemmissa valmistumiseen liittyvissä tutkimuksissa on huomattu, että menestyminen aiemmissa opinnoissa ja ylioppilaskirjoituksissa on yhteydessä myös korkeakouluopintomenestykseen. Erityisesti kasvatusalalla aikaisempaan koulumenestykseen perustuva valintapisteytys ennustaa hyvin yliopistosta valmistumista (Kallio ym. 2018). Ylioppilastutkintoarvosanojen on havaittu ennustavan hyvin myös korkeakouluopintojen kurssiarvosanoja (Mieskonen, 2017). Niin ikään kasvatusalalla opintomenestys ylioppilaskirjoituksissa korreloi positiivisesti yliopisto-opintojen opintomenestyksen kanssa (Kallio ym., 2018). Kirjoitettavista ylioppilasaineista erityisesti matematiikka indikoi menestymistä kauppatieteellisissä opinnoissa (Mieskonen, 2017).

Myös muilla kuin aiempiin opintoihin liittyvillä muuttujilla on huomattu olevan yhteyttä korkeakouluopinnoissa menestymiseen ja valmistumiseen. Viitanen (2016) on havainnut tarkastellessaan muun muassa sosiaali- ja taloustieteitä, että arvosanojen keskiarvo ja opiskelijan sukupuoli ovat merkitseviä muuttujia arvioitaessa valmistumistodennäköisyyttä. Myös ylimääräisten opintopisteiden (3 % tutkinnon laajuudesta) suorittaminen ensimmäisenä lukuvuotena kasvattaa valmistumisen todennäköisyyttä enemmän kuin yhden numeron parannus keskiarvossa (asteikolla 1-5). Mankki ym. (2018) havaitsivat luokanopettajakoulutusta koskevassa tutkimuksessaan, että VAKAVA-kokeessa menestyminen

tai ylioppilasarvosanojen perusteella annetut valintapisteet eivät ennustaneet opiskelijan valmistumista tavoiteajassa. Sosiaali- ja liikuntatieteissä sekä insinöörialoilla pääsykokeet toimivat ylioppilaskokeeseen perustuvaa pisteytystä parempina valmistumisen ja suoritettujen opintopisteiden ennustajina (Häkkinen, 2004).

Aikaisemmissa tutkimuksissa ylioppilasaineiden vaikutusta valmistumiseen tai opintomenestykseen on tutkittu rajoitetusti. Vaikutuksia ollaan tutkittu pääsääntöisesti ainoastaan laajasti kirjoitettujen ylioppilasaineiden kuten matematiikan ja äidinkielen osalta (mm. Mieskonen 2017, Häkkinen 2004, Mankki ym. 2018). Näissä tutkimuksissa vähemmän kirjoitetut ylioppilasaineet on joko sivuutettu tai yhdistetty niin kutsutun ”reaalikorin” mukaisesti (mm. Mieskonen 2017 ja Mankki ym. 2018). Edellisistä tutkimuksista poiketen Kallio ym. (2018) käyttivät ylioppilastutkinnon koearvosanojen keskiarvoa opintomenetyksen selittämisessä. Erona aikaisempiin tutkimuksiin tässä tutkielmassa pyritään selittämään valmistumisaikaa mahdollisimman monella ylioppilaskirjoitusaineella. Näin ollen aineistoon sovelletaan k:n keskiarvon ryhmittelymenetelmää ylioppilasarvosanatietojen tiivistämiseksi. Tässä tutkimuksessa opiskelijat ryhmitellään ylioppilaskirjoituksissa saavuttamiensa arvosanojen perusteella. Ryhmittelyssä otetaan huomioon ne ylioppilaskirjoitusten oppiaineet, joissa on tarpeeksi havaittuja arvosanoja. Tutkittavien muuttujien valintaa käsitellään tarkemmin luvussa 2.

Aikaisemmat tutkimukset eivät ole käsitelleet valmistumisaikaa elinaika-analyysin ja ryhmittelyanalyysin keinoin. Mankki ym. (2018) ovat käsitelleet valmistumista ja tavoiteajassa valmistumista dikotomisena muuttujana. Tällöin valmistumista tarkastellaan kaksiarvoisena muuttujana, joka saa arvon riippuen siitä, onko opiskelija valmistunut tavoiteajassa vai ei. Viitanen (2016) puolestaan tutki valmistumistodennäköisyyksiä luokittelumenetelmien avulla ja vertasi tuloksiaan Coxin suhteellisen vaaran mallin tuottamiin tuloksiin. Elinaikamallia käytettiin tutkimuksessa valmistumistodennäköisyyksien laskentaan eikä niinkään valmistumisaikojen selittämiseen (Viitanen, 2016). Petman (2017) puolestaan ennusti Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan tutkintojen lukumääriä tilasiirtymämallilla.

Valintauudistuksen lisäksi myös yliopistojen ja ammattikorkeakoulujen rahoitusmalleja uudistettiin vuoden 2021 alusta alkaen (Opetus- ja kulttuuriministeriö, 2018). Uudet rahoitusmallit tulivat voimaan 1. tammikuuta 2021 ja ne otettiin käyttöön sopimuskaudelle 2021-2024. Yliopistojen perusrahoituksessa perustutkintojen perusteella kohdentuva rahoitusosuus kasvoi 19 prosentista 30 prosenttiin. Uudessa rahoitusmallissa tutkinnon suoritusaika vaikuttaa tutkintojen perusteella määräytyvään rahoituskriteeriin. Tavoiteajassa suoritettuja tutkintoja painotetaan laskennassa enemmän kuin tavoiteaikaa pidemmässä ajassa valmistuneita tutkintoja.

Tässä pro gradu -tutkielmassa tutkitaan, mitkä tekijät selittävät yliopisto-opiskelijan valmistumisaikaa Jyväskylän yliopiston matemaattis-luonnontieteellisessä tiedekunnassa. Valmistumisaikaa käsitellään elinaikana opintojen aloittamisesta kandidaatiksi valmistumiseen saakka. Elinaikamallinnuksessa voidaan ottaa sensuroinnin avulla huomioon myös ne opiskelijat, jotka eivät ole vielä ehtineet valmistua. Sovellettava elinaikamalli on Coxin (1972) regressiomalli, jossa on mukana ylioppilasarvosanojen avulla tuotettu ryhmätieto sekä muita hakuvaiheen muuttujia. Coxin regressiomallin avulla saadaan lisätietoa selittäjien vaikutuksesta valmistumiseen, tarkalleen ottaen valmistumisen vaaraan. Erityisen kiinnostavaa on, mitkä tekijät selittävät tavoiteajassa valmistumista ja millä tavalla selittäjät vaikuttavat valmistumiseen.

Luvussa 2 esitellään tutkielman aineisto sekä tutkimusongelma. Luvusta 3 eteenpäin kerrotaan aineistoon sovellettavista analyysimenetelmistä sekä niiden teoriasta. Luvussa 3

esitellään tutkielmassa käytetty k:n keskiarvon ryhmittelymenetelmä. Elinaikavasteeseen liittyvät peruskäsitteet käydään läpi luvussa 4, minkä jälkeen esitellään aineistoon sovelletun elinaika-analyysimenetelmän, Coxin regression, teoria luvussa 5. Luvussa 6 esitellään mallinvalinnan teoriaa. Menetelmillä saadut tulokset sekä tulosten diagnostiikka esitellään kirjallisesti ja graafisesti luvussa 7. Luku 8 sisältää yhteenvedon.

2 Aineiston ja tutkimusongelman kuvaus

Tässä luvussa esitellään käytettävä aineisto, sen koonti sekä aineiston muuttujat. Aineisto koostuu kahdesta osa-aineistosta, niin kutsutusta hakuvaiheen aineistosta sekä opiskeluvaiheen aineistosta. Hakuvaiheen aineisto on koottu Opintopolku.fi-palvelusta, ja se sisältää tiedot esimerkiksi hakijan ylioppilaskoetuloksista ja pääsykoepisteistä. Opiskeluvaiheen aineisto on puolestaan koottu Jyväskylän yliopiston tietovarastosta, joka sisältää opintorekisterin tietoja. Tästä aineistosta käy ilmi esimerkiksi opiskelijan opintojen aloitusajankohta, oppiaine tai tutkinto-ohjelma sekä valmistumisajankohta. Aineistot yhdistettiin opiskelijakohtaisen id:n perusteella. Yhdistetyssä aineistossa on yhteensä 1255 havaintoa ja 137 muuttujaa.

2.1 Hakuvaiheen muuttujat

Hakuvaiheen aineistossa on mukana vuosina 2015-2020 Jyväskylän yliopiston matemaattis-luonnontieteelliseen tiedekunnan kandidaattiohjelmiin hyväksytyt suomalaiset hakijat, jotka ovat myös aloittaneet opintonsa tiedekunnassa. Aineistossa on tieto hakijoiden sukupuolesta, syntymäajasta sekä haetusta hakukohteesta. Hakukohdetiedoissa on hakuvuodesta riippuvia eroja, sillä vuonna 2017 matemaattis-luonnontieteellisessä tiedekunnassa siirryttiin oppiaineista tutkinto-ohjelmiin. Näin ollen esimerkiksi akvaattisten tieteiden oppiaineeseen ei ole voitu hyväksyä opiskelijoita vuoden 2017 jälkeen, vaan opiskelijat ovat aloittaneet opintonsa vuodesta 2017 lähtien tutkinto-ohjelmissa.

Vuodesta 2017 alkaen voimassa olevat tutkinto-ohjelmat ovat biologian kandidaatti- ja maisteriohjelma, luonnonvarat ja ympäristö -kandidaatti- ja maisteriohjelma, fysiikan kandidaatti- ja maisteriohjelma, kemian kandidaatti- ja maisteriohjelma, matematiikan aineenopettajan kandidaatti- ja maisteriohjelma sekä matematiikan ja tilastotieteen kandidaatti- ja maisteriohjelma. Oppiaineisiin ja tutkinto-ohjelmiin hyväksyttyjen ja paikan vastaanottaneiden opiskelijoiden lukumäärät on esitetty taulukossa 1.

Hakuvaiheen aineistosta käy ilmi ylioppilaskirjoituksissa saadut arvosanat. Ylioppilaskirjoituksista saadut arvosanat on ilmoitettu asteikolla korkeimmasta matalimpaan: *laudatur* (L), *eximia cum laude approbatur* (E), *magna cum laude approbatur* (M), *cum laude ap-*

Taulukko 1: Vuosina 2015-2020 valittujen ja paikan vastaanottaneiden opiskelijoiden lukumäärät sukupuolittain ja hakukohteittain.

	Mies	Nainen	Yhteensä
Akvaattiset tieteet	9	18	27
Biologian ala	15	54	69
Biologian kandidaatti- ja maisteriohjelma	39	108	147
Fysiikan (fyysikko tai aineenopettaja) kandidaatti- ja maisteriohjelma	121	32	153
Fysiikan ala	83	33	116
Kemian (kemisti tai aineenopettaja) kandidaatti- ja maisteriohjelma	72	100	172
Kemian ala	53	50	103
Luonnonvarat ja ympäristö -kandidaatti- ja maisteriohjelma	37	93	130
Matematiikan aineenopettajan kandidaatti- ja maisteriohjelma	35	29	64
Matematiikan ja tilastotieteen kandidaatti- ja maisteriohjelma	84	45	129
Matematiikka ja tilastotiede	73	45	118
Ympäristötiede ja -teknologia	6	21	27
Yhteensä	627	628	1255

probatur (C), *lubenter approbatur* (B), *approbatur* (A) ja *improbatur* (I) (Valtioneuvoston asetus ylioppilastutkinnosta 612/2019b, 6 §). Aineistossa on arvosanatietoja yhteensä 58 eri ylioppilaskirjoitusaineesta.

Valtioneuvoston ylioppilastutkintoa koskevan asetuksen (612/2019b) mukaan ylioppilastutkintoon kuuluu vähintään neljä koetta, joista äidinkielen ja kirjallisuuden koe on kaikille pakollinen. Äidinkielen ja kirjallisuuden kokeen lisäksi opiskelija valitsee kolme muuta pakollista koetta. Pakollisten kokeiden lisäksi opiskelija voi osallistua yhteen tai useampaan valinnaiseen kokeeseen (Valtioneuvoston asetus ylioppilastutkinnosta 612/2019b). Koska opiskelija voi kirjoittaa saman oppiaineen vain joko pakollisena tai valinnaisena, nämä kaksi tilannetta yhdistettiin tutkimuksessa. Tutkimukseen valitaan ne ylioppilaskirjoitusaineet, joita on kirjoittanut yli kymmenen prosenttia aineiston opiskelijoista. Näin ollen tutkimukseen valitaan arvosanat äidinkielestä (suomi), pitkästä ja lyhyestä matematiikasta, biologiasta, fysiikasta, kemiasta, maantieteestä, terveystiedosta, psykologiasta sekä kielistä. Kielistä tarkastellaan englannin pitkää oppimäärää ja ruotsin keskipitkää oppimäärää. Oppiaineet, joissa havaintoja oli vähemmän kuin 10 prosenttia aineistosta, jätettiin tutkimuksen ulkopuolelle. Tutkimuksessa mukana olevien oppiaineiden havaitut osuudet aineistossa on esitelty taulukossa 2. Taulukosta nähdään, että 95 % opiskelijoista on suorittanut hyväksytysti äidinkielen ja kirjallisuuden suomenkielisen ylioppilaskokeen. Viimeinen tutkimukseen mukaanotettava ylioppilaskirjoitusaine on psykologia, jonka on suorittanut hyväksytysti 11 % aineiston opiskelijoista.

Ennen mallin sovittamista ylioppilasarvosanat muokattiin numeeriseen muotoon. Kirjainarvosanat muokattiin numeerisiksi noudattaen Ylioppilastutkintolautakunnan noudattamaa kompensaatiopistekäytäntöä (Ylioppilastutkintolautakunta, 2020): laudatur 7, eximia cum laude approbatur 6, magna cum laude approbatur 5, cum laude approbatur 4, lubenter approbatur 3 ja approbatur 2. Improbatureiden kompensaatiopistemäärä on 0. Lisäksi kirjoittamatta jättäminen tulkittiin nollana. Opiskelija ei voi kirjoittaa samaa oppiainetta useampana eri oppimääränä. Matematiikan osalta pitkän ja lyhyen oppimäärän arvosanat yhdistettiin siten, että pitkän matematiikan arvosanaa painotettiin kertoimella 1.25 (Väisänen ja Ylönen, 2004). Tutkittavien oppiaineiden arvosanojen mediaanit sekä kvartiilivälit on esitetty taulukossa 3.

Taulukko 2: Kirjoittajien osuudet ylioppilasaineittain niiden oppiaineiden osalta, jotka otettiin mukaan tutkimukseen.

	Kirjoittajia (osuus)
Äidinkieli (suomi)	0.95
Englanti, pitkä oppimäärä	0.89
Matematiikka, pitkä oppimäärä	0.79
Kemia	0.56
Biologia	0.54
Fysiikka	0.50
Ruotsi, keskipitkä oppimäärä	0.35
Matematiikka, lyhyt oppimäärä	0.15
Terveystieto	0.14
Maantiede	0.12
Psykologia	0.11

Oppimäärien yhdistämisen jälkeen tarkasteltiin tutkimukseen valittujen ylioppilaskirjoitusaineiden kombinaatioita. Huomattiin, että aineistossa on tällöin yhteensä 151 erilaista ylioppilasainekombinaatiota. Kombinaatioiden määrä olisi vielä huomattavasti suurempi, jos tarkasteluissa huomioitaisiin kaikki 58 ylioppilaskirjoitusainetta.

Yhteisvalinnassa valinta yliopistoon tapahtuu valintajonon kautta. Jyväskylän yliopiston matemaattis-luonnontieteellisessä tiedekunnassa mahdollisia valintajonoja ovat suoravalintajono, ylioppilaskoepistejono, valintakoejono ja yhteispistejono. Yhteisvalinnassa sama opiskelija voi tulla valituksi hakukohteeseen vain yhden valintajonon kautta. Hakukohde voi asettaa ehdot suoravalinnalle, jolloin ehdot täyttävä opiskelija tulee valituksi hakukohteeseen suoravalintajonon kautta. Esimerkiksi matematiikan ja tilastotieteen hakukohteessa suoravalinnan ehtona on, että hakija on saanut pitkän matematiikan ylioppilaskokeesta vähintään arvosanan E. Eri valintajonojen kautta valituksi tulleiden opiskelijoiden lukumäärät on esitelty taulukossa 4.

Aineistossa suurin osa, 62 % opiskelijoista, on tullut valituksi ylioppilaskoepistejonosta. Hakuvaiheessa opiskelijoille annetaan pisteitä ylioppilasarvosanojen perusteella. Pistemäärä riippuu kirjoitetuista oppiaineista, niiden oppimäärän pituudesta sekä arvosanasta. Opiskelija voidaan hyväksyä hakukohteeseen tämän pistemäärän perusteella ylioppilaspistejonosta tai pistemäärän ja valintakoepistemäärän summalla yhteispistejonosta. Aikaisemmin jokainen hakukohde on voinut pisteyttää ylioppilasarvosanat omalle alalleen sopivalla tavalla. Todistusvalintauudistuksen yhteydessä pisteytyskäytäntöä on pyritty yhdenmukaistamaan, mikä on ollut myös opetus- ja kulttuuriministeriön toive (Opetus- ja kulttuuriministeriö, 2016).

Ylioppilasarvosanojen lisäksi tiedetään, onko opiskelija osallistunut valintakokeeseen ja kuinka monta pistettä hän on saanut valintakokeesta. Valintakokeiden pisteytyskäytännöt vaihtelevat laitoksittain ja hakukohteittain. Osallistuessaan valintakokeeseen opiskelija voi tulla valituksi valintakoejonon kautta. Jyväskylän yliopiston matemaattis-luonnontieteellisessä tiedekunnassa pelkän valintakoemenestyksen perusteella valittavien opiskelijoiden osuus yhteishaussa on ylioppilaskoejonon kiintiöitä pienempi. Mikäli opiskelijalla on hakuvaiheessa ylioppilastutkintotodistus ja lisäksi hän on osallistunut valintakokeeseen, hänet voidaan hyväksyä hakukohteeseen myös yhteispistejonon kautta. Yhteispistejonossa hyväksymisjärjestykseen vaikuttaa sekä ylioppilastutkinnon arvosanat että valintakoemenestys.

Taulukko 3: Tutkimukseen valittujen ylioppilasaineiden arvosanojen kvantiilit tarkasteltaessa ainoastaan kyseisen ylioppilasaineen kirjoittaneita opiskelijoita.

	0 %	25 %	50 %	75 %	100 %
Äidinkieli (suomi)	0	4	5	6	7
Matematiikka	0	5	6	8	9
Biologia	0	4	5	6	7
Kemia	0	4	5	6	7
Fysiikka	0	4	5	6	7
Ruotsi (keskipitkä oppimäärä)	0	4	5	6	7
Englanti (pitkä oppimäärä)	0	4	5	6	7
Historia	2	4	5	5	7
Terveystieto	2	4	5	6	7
Maantiede	2	5	6	6	7
Psykologia	2	4	5	6	7

Taulukko 4: Valittujen opiskelijoiden lukumäärät valintajonoittain.

	Opiskelijoiden lukumäärä
Ylioppilaspistejono	783
Suoravalintajono	177
Valintakoejono	147
Yhteispistejono	101
Yhteishaun ulkopuolelta	47
Yhteensä	1255

Hakukohteen tiedoissa on määritelty etukäteen, kuinka monta aloituspaikkaa on kutakin valintajonoa kohti. Hakukohteissa aloituspaikat täytetään valintajonoittain. Ensin hyväksytään suoravalintakriteerit täyttävät hakijat. Suoravalinnan jälkeen aloituspaikkoja täytetään ylioppilaspistejonon perusteella. Lopuksi täytetään valintakoejonon aloituspaikat. Tällöin opiskelija, joka voisi tulla hyväksytyksi useamman eri valintajonon kautta, tulee valituksi vain yhden valintajonon kautta hakukohteeseen. Esimerkiksi opiskelija, joka voisi tulla hyväksytyksi sekä ylioppilaspistejonosta että valintakoejonosta, valitaan hakukohteeseen ylioppilaspistejonosta täyttöjärjestyksestä johtuen. Suoravalitut opiskelijat valitaan hakukohteeseen kaikkien kiintiöiden yli. On siis mahdollista, että hakukohteeseen hyväksytään enemmän opiskelijoita kuin mitä yhteishaussa on ilmoitettu.

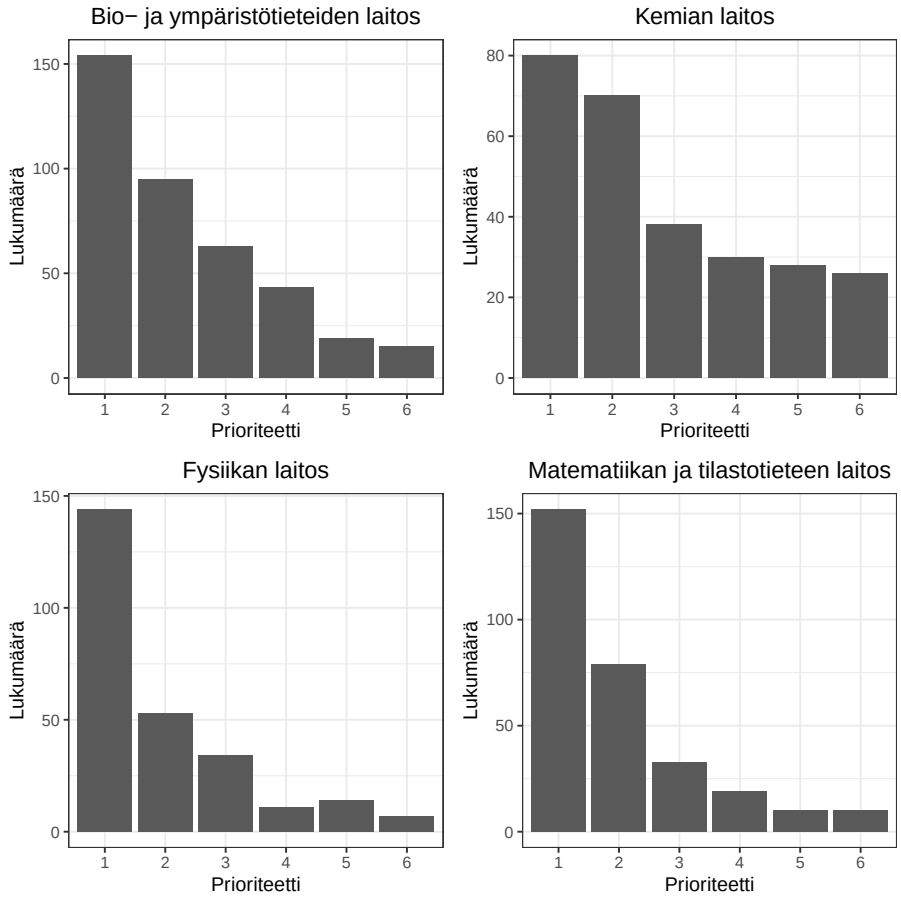
Yhteishaun lisäksi Jyväskylän yliopiston matemaattis-luonnontieteelliseen tiedekuntaan on mahdollista hakea opiskelemaan erillishaun kautta. Erillishaku on yhdistetty aineistossa osaksi isompaa, yhteishaun ulkopuolelta tulevien opiskelijoiden kokonaisuutta. Tässä kokonaisuudessa on mukana opiskelijat, jotka eivät olleet hakukelpoisia yhteishaussa sekä siirtohaun kautta tai kisamenestyksen perusteella valituksi tulleet opiskelijat.

Opiskelija voi suorittaa opintoja avoimessa yliopistossa jo ennen varsinaisen opiskeluoikeuden alkamista yliopistossa. Aineistossa on tieto siitä, onko opiskelija suorittanut opintoja Jyväskylän yliopiston avoimessa yliopistossa ja jos on, niin kuinka monta opintopistettä. Tämän tutkimuksen kannalta avoimessa yliopistossa suoritetut opintopisteet sekä avoimen väylä ovat erityisen kiinnostavia siksi, että avoimen väylää pyritään laajentamaan yliopistohauissa ja sen tehtävä on osaltaan varmistaa koulutukseen hakeutumismahdollisuuksia erilaisissa elämäntilanteissa oleville (Opetus- ja kulttuuriministeriö, 2016). Aineistossa yhteensä 152 opiskelijaa on suorittanut opintoja avoimessa yliopistossa. Eniten avoimessa yliopistossa opintoja suorittaneita opiskelijoita on bio- ja ympäristötieteiden laitoksella. Laitoskohtaiset avoimessa yliopistossa suoritettujen opintopisteiden keskiarvot on esitetty taulukossa 5.

Aineistosta käy ilmi opiskelijan prioriteetti, eli missä kohtaa hakutoivelistaa hakukohde oli. Mikäli prioriteetti on yksi, on hakukohde ollut opiskelijalla ensimmäinen hakutoivelis-

Taulukko 5: Avoimessa yliopistossa suoritettujen opintopisteiden keskiarvot laitoksittain niiden opiskelijoiden osalta, joiden opintopistemäärät on ilmoitettu aineistossa.

Laitos	Opintopisteiden keskiarvo
Bio- ja ympäristötieteiden laitos	19.97
Fysiikan laitos	8.33
Kemian laitos	8.63
Matematiikan ja tilastotieteen laitos	10.75



Kuvio 1: Opiskelijoiden hakuprioriteetit laitoksittain.

talla. Opiskelija voi ottaa samana lukuvuonna vastaan vain yhden korkeakoulututkintoon johtavan opiskelupaikan (Laki yliopistolain muuttamisesta 558/2009). Jos opiskelija tulee hyväksytyksi useampaan hakukohteeseen, niin hän voi ottaa vastaan vain korkeamman hakuprioriteetin kohteen (Valtioneuvoston asetus korkeakoulujen yhteishausta 289/2019a). Prioriteettijakaumat laitoksittain on esitelty tarkemmin kuviossa 1.

2.2 Opiskeluvaiheen muuttujat

Opiskeluvaiheen aineisto koostuu opiskelijoista, jotka ovat aloittaneet opintonsa Jyväskylän yliopiston matemaattis-luonnontieteellisessä tiedekunnassa vuosina 2015-2020. Hakuvaiheen aineiston tavoin myös opiskeluvaiheen aineisto rajattiin suomalaisiin opiskelijoihin. Aineistossa on tietoja opiskelijan laitoksesta sekä oppiaineesta, opintojen aloituspäivämäärästä, opintojen etenemisestä sekä mahdollisesta valmistumisesta kandidaatiksi ja/tai maisteriksi. Aineistossa on tieto lisäksi siitä, mikäli opiskelijalla on opinto-oikeus johonkin toiseen tutkinto-ohjelmaan Jyväskylän yliopistossa. Tämä opinto-oikeus voi olla myönnetty ennen hyväksymistä hakukohteeseen. Opiskelija on voinut myös vaihtaa pääainetta tai tutkinto-ohjelmaa matemaattis-luonnontieteelliseen tiedekuntaan hyväksymisen jälkeen.

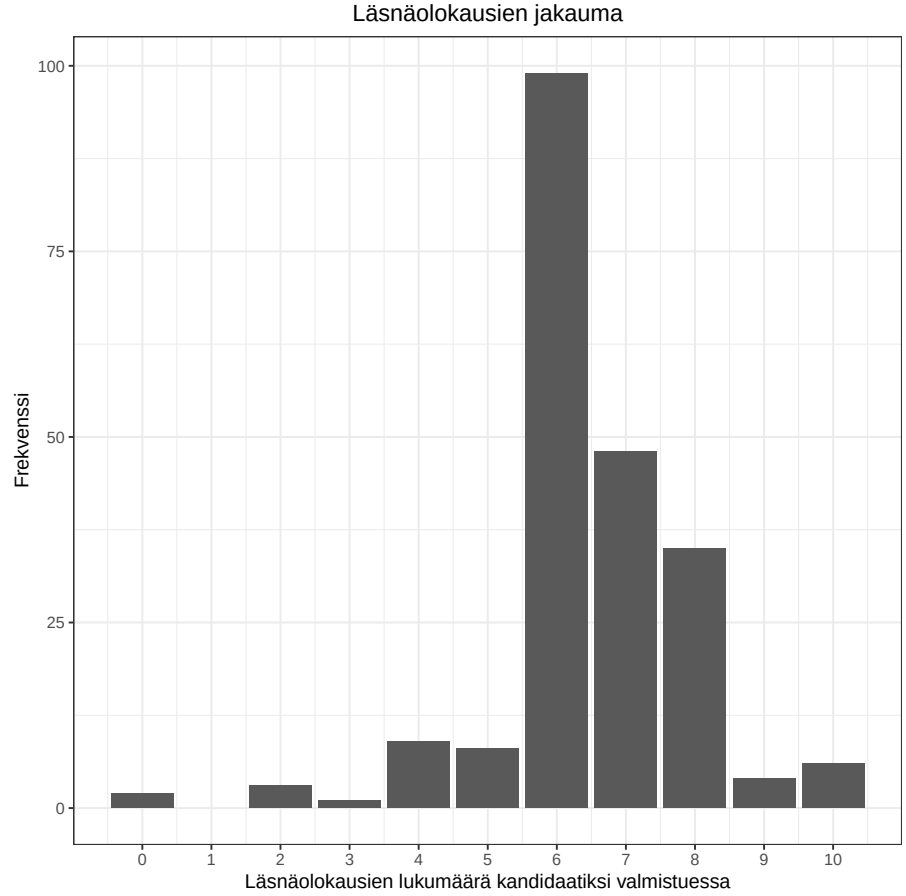
Opintojen etenemisestä kertovia muuttujia ovat opintopistekertymä ennen syksyä 2015, opintopistekertymät syksyltä ja keväältä vuosina 2015-2020 sekä kokonaisopintopistekertymän 31. heinäkuuta 2020 mennessä. Opintopistekertymän lisäksi aineistossa on tieto

siitä, onko opiskelija ilmoittautunut läsnä- tai poissaolevaksi kyseiselle lukukaudelle. Mikäli opiskelija ei ole ilmoittautunut läsnä- tai poissaolevaksi, on kyseessä puuttuva tieto. Puuttuvaa ilmoittautumistietoa käsitellään tässä tutkimuksessa poissaolon tavoin, sillä opiskelija voi suorittaa tutkintoon johtavia opintoja ja suorittaa tutkinnon ainoastaan läsnäolevaksi ilmoittautuneena (Yliopistolaki 558/2009). Aineistosta näkyy myös kandidaatintutkinnon ja/tai maisterintutkinnon myöntämispäivämäärä, mikäli opiskelija on ehtinyt valmistua aineiston poimintahetkeen 31. heinäkuuta 2020 mennessä. Aineistossa on myös tieto siitä, mistä oppiaineesta tai tutkinto-ohjelmasta opiskelija on valmistunut. Kandidaatiksi valmistuneiden lukumäärät laitoksittain on esitelty tarkemmin taulukossa 6.

Opintojen alkamisajankohdan ja valmistumisajankohdan perusteella pystytään laskemaan, kuinka monta läsnäololukukautta opiskelijalla on ehtinyt kertyä valmistumiseen mennessä. Tavoiteaika kandidaatintutkinnolle on kuusi läsnälukukautta eli kolme lukuvuotta (Yliopistolaki 558/2009 §40). Tämä uusi johdettu muuttuja nimettiin aineistossa nimellä *kandikaudet*. Muuttujan jakaumaa on tarkasteltu kuviossa 2. Kandidaatiksi valmistuneiden läsnäololukukausien keskiarvoja ja keskihajontoja laitoksittain on tarkasteltu taulukossa 7.

Taulukko 6: Opintonsa aloittaneiden ja kandidaatiksi valmistuneiden opiskelijoiden lukumäärät yhteensä laitoksittain 1.8.2015 - 31.7.2020.

Laitos	Aloittaneiden lkm	Valmistuneiden lkm
Bio- ja ympäristötieteiden laitos	399	99
Fysiikan laitos	269	39
Kemian laitos	276	34
Matematiikan ja tilastotieteen laitos	311	32



Kuvio 2: Opiskelijoiden kandidaatiksi valmistumiseen tarvittavien lukukausien jakauma koko tiedekunnassa niiden osalta, jotka ovat valmistuneet.

Taulukko 7: Kandidaatiksi valmistumiseen tarvittavien läsnäololukukausien lukumäärien keskiarvot ja keskihajonnat laitoksittain.

Laitos	Keskiarvo	Keskihajonta
Bio- ja ympäristötieteiden laitos	6.48	1.13
Fysiikan laitos	6.90	1.56
Kemian laitos	6.51	1.22
Matematiikan ja tilastotieteen laitos	6.11	2.01

Tutkimusongelmana on selvittää, mitkä hakuvaiheen muuttujat selittävät kandidaatiksi valmistumiseen tarvittavien lukukausien määrää. Erityisen kiinnostavaa on, mitkä tekijät ennustavat parhaiten opiskelijan valmistumista tavoiteajassa ja millä tavalla selittävät muuttujat vaikuttavat valmistumiseen.

3 K:n keskiarvon ryhmittely

Tässä tutkimuksessa opiskelijoiden ylioppilasarvosanatietoja pyritään tiivistämään ryhmittelemällä opiskelijat ylioppilasarvosanojen perusteella arvosanaryhmiin käyttäen k:n keskiarvon menetelmää. Ryhmittelyanalyysillä pyritään jakamaan aineisto samankaltaisiin ryhmiin. Tavoitteena on, että ryhmän sisällä opiskelijat ovat mahdollisimman samankaltaisia ja ryhmien välillä mahdollisimman erilaisia. Aineistossa olevia ryhmiä tai ryhmien lukumäärää ei tiedetä etukäteen. Seuraava k:n keskiarvon ryhmittelyn esitys mukailee Jamesin, Wittenin, Hastien ja Tibshiranin (2013) esitystä.

Oletetaan, että ryhmien lukumäärä k on valittu etukäteen. Kriteerejä parhaan k:n valitsemiseksi esitellään myöhemmin tässä luvussa. Merkitään i:nnettä havaintoa x_i ja l:nnettä ryhmää $K_l, l = 1, 2, ..., k$. Havaintojen samankaltaisuutta mitataan niin kutsutuilla etäisyysmitoilla. Havaintojen x_i ja x_j välisenä etäisyysmittana käytetään neliöityä euklidista etäisyyttä

$$d(x_i, x_j) = \|(x_i - x_j)\|^2. \tag{1}$$

K:n keskiarvon ryhmittelyn tavoitteena on muodostaa ryhmittely, jossa ryhmien sisäinen vaihtelu on mahdollisimman pieni. Ryhmän K_l sisäistä vaihtelua kuvataan tunnusluvulla

$$W(K_l) = \frac{1}{N_l} \sum_{i,j \in K_l} \sum_{p=1}^{P} (x_{ip} - x_{jp})^2, \tag{2}$$

missä N_l on havaintojen lukumäärä l:nnessä ryhmässä ja P on ryhmittelyssä käytettyjen muuttujien kokonaismäärä ja x_{ip} on i:nnen havainnon arvo p:nnessä sarakkeessa. Toisin sanoen sisäinen vaihtelu kertoo l:nnen ryhmän kaikkien havaintojen parittaisten neliöityjen euklidisten etäisyyksien keskiarvon.

Menetelmän tavoitteena on jakaa havainnot k:hon ryhmään siten, että yhtälö 2 minimoidaan jokaisen ryhmän K_l osalta. Tämä on kuitenkin haastavaa, sillä mahdollisia tapoja jakaa havainnot ryhmiin on valtava määrä. Minimointi toteutetaan suboptimaalisesti. Ratkaisualgoritmi esitetään heuristisesti Jamesin, Wittenin, Hastien ja Tibshiranin (2013) esitystä mukaillen algoritmissa 1.

Algoritmi 1: K:n keskiarvon ryhmittely

1. Arvotaan jokaiselle havainnolle x_i satunnaisesti kokonaisluku l väliltä [1,k], $k > 1$. Nämä luvut edustavat alustavia ryhmiä K_l.

2. Toistetaan vaihetta 2, kunnes ryhmät pysyvät muuttumattomina:

   (a) Jokaiselle ryhmälle K_l lasketaan klusterikeskus m_{K_l}. Klusterikeskus m_{K_l} on vektori, jossa on P muuttujakeskiarvoa ryhmän havainnoista.

   (b) Siirretään jokainen havainto x_i siihen ryhmään, jonka keskus on lähimpänä. Tässä tapauksessa lähin etäisyys on määritelty käyttäen euklidista etäisyyttä (kaava 1).

Koska ryhmittely löytää globaalin optimin sijaan lokaalin optimin, ryhmittelyn tulos riippuu aina ryhmittelyalgoritmin 1 vaiheessa 1 annetuista aloitusryhmistä. Tästä johtuen on

tärkeää toistaa ryhmittely useamman kerran ja valita ryhmittelyistä se, joka minimoi sisäisen vaihtelun 2. Tässä tutkimuksessa aineiston ryhmittely toteutettiin R-ohjelmiston (R Core Team, 2020) funktiolla `kmeans`. Funktio käyttää oletusarvoisesti ratkaisualgoritmina Hartigan-Wongin (1979) algoritmia. Funktiolle voi antaa parametriksi ryhmittelyn toistomäärän kullekin k:lle. Toistomäärä kertoo, kuinka monta kertaa funktio ryhmittelee aineiston k:hon ryhmään. Toistetuista ryhmittelyistä funktio valitsee sen ryhmittelyn, joka minimoi sisäisen hajonnan. Edellä mainittu menettely pienentää riskiä, että ryhmittelyn tulos riippuisi täysin iteraatiokerrasta.

Sopivimman ryhmien määrän k valitsemiseksi on olemassa useita menetelmiä. Tässä tutkielmassa sopivinta ryhmien lukumäärää tutkitaan sisäisen vaihtelun sekä keskimääräisen siluettikertoimen avulla. Sisäisen vaihtelun tarkastelussa aineiston ryhmittely toteutetaan ensin usealle eri ryhmien lukumäärälle, jonka jälkeen tarkastellaan ryhmien sisäisiä vaihteluja (kaava 2) kullakin ryhmien lukumäärällä k. Tarkastelussa etsitään niin kutsuttua nivelkohtaa, jossa sisäinen vaihtelu ei enää pienene merkittävästi. Tällöin ryhmien lukumäärän lisääminen ei paranna ryhmittelyä. Toinen tapa arvioida ryhmittelyn onnistumista on laskea keskimääräiset siluettikertoimet eri ryhmien määrälle k.

Olkoon etäisyysmitta d kuten edellä. Siluettikerroin havainnolle x_i on (Rousseeuw, 1987)

$$s_i(K, d) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{3}$$

missä

$$a(i) = \frac{1}{N_{K_i} - 1} \sum_{K_i = K_j, i \neq j} d(x_i, x_j)$$

ja

$$b(i) = \min_{r \neq K_i} \frac{1}{N_r} \sum_{K_j = r} d(x_i, x_j).$$

Tällöin funktio $a(i)$ kertoo havainnon i keskimääräisen etäisyyden oman ryhmänsä K_i muihin havaintoihin ja $b(i)$ kertoo pienimmän keskimääräisen etäisyyden muiden kuin oman ryhmän havaintoihin. Nähdään, että yhtälö 3 saa arvoja väliltä $[-1, 1]$. Siluettikerroin kertoo, kuinka hyvin opiskelija sopii omaan ryhmäänsä. Jos opiskelija on lähellä oman ryhmänsä keskusta ja kaukana muista ryhmistä, siluettikertoimen arvo on positiivinen. Mikäli opiskelijan etäisyys omaan ryhmään ja johonkin toiseen ryhmään ovat lähellä toisiaan, siluettikerroin on lähellä nollaa. Jos opiskelija on kaukana muista oman ryhmänsä havainnoista ja lähellä jotain toista ryhmää, on siluettikertoimen arvo negatiivinen.

Keskimääräinen siluettikerroin määritellään siluettikertoimien keskiarvona

$$\bar{S}(K, d) = \frac{1}{n} \sum_{i=1}^{n} s_i(K, d),$$

missä n on havaintojen lukumäärä. Ryhmien lukumäärän valinta tehdään siten, että keskimääräinen siluettikerroin on mahdollisimman suuri. Tällöin ryhmittelyn tuloksena saadaan ryhmät, joiden opiskelijat ovat lähellä omia keskuksiaan ja kaukana toisista ryhmistä. Siluettikertoimen laskemisessa käytettiin `cluster`-paketin funktiota `silhouette` (Maechler ym., 2019).

4 Elinaika-analyysin käsitteitä

Tässä tutkielmassa tarkastellaan aikaa opintojen aloittamisesta kandidaatiksi valmistumiseen saakka. Opintojen alkamishetkeksi on määritelty opiskelijan opintojen alkamispäivämäärä. Kiinnostava tapahtuma on kandidaatiksi valmistuminen. Valmistumisajankohta on päivämäärä, jolloin kandidaatin tutkinto on kirjattu Jyväskylän yliopiston opintorekisteriin. Koska valmistumisaika ilmoitetaan aineistossa kokonaisina lukukausina opintojen alkamishetkestä valmistumiseen, on kyseessä diskreetti elinaika. Tässä luvussa esitellään diskreetti elinaikavaste sekä elinaika-analyysin keskeisiä funktioita kuten vaarafunktio ja välttöfunktio. Lisäksi käydään läpi elinaika-aineiston sensuroituminen. Diskreetti elinaika ja siihen liittyvät funktiot esitellään alaluvussa 4.1. Elinaika-aineiston sensuroituminen esitellään alaluvussa 4.2.

4.1 Elinaikafunktiot diskreetille elinajalle

Elinaika T ilmoittaa kiinnostavan tapahtuman tapahtumishetken. Elinaika on aina positiivinen, ja sen aloitushetki ja loppumishetki tulee olla tarkasti määriteltyjä. Loppumishetken määrittävän tapahtuman tulee olla niin ikään tarkasti määritelty ja sen tulee voida sattua vain kerran. Seuraavaksi esitellään elinaika-analyysin keskeisiä funktioita Singeriä ja Willettiä (1993) mukaillen.

Olkoon valmistumisaika (elinaika) T diskreetti satunnaismuuttuja. Tällöin todennäköisyys sille, että valmistumisaika on t, voidaan esittää pistetodennäköisyytenä

$$p_t = P(T = t),\ t \geq 1, \tag{4}$$

missä t on kokonaisluku. Kertymäfunktio voidaan puolestaan esittää pistetodennäköisyyksien summana

$$F(t) = \sum_{u=1}^{t} p_u. \tag{5}$$

Kertymäfunktio ilmoittaa todennäköisyyden sille, että opiskelija on valmistunut ajanhetkeen t mennessä.

Välttöfunktio (*survival function*) kertoo puolestaan todennäköisyyden, että elinaika T on suurempi kuin tarkasteltu ajanhetki t. Tällöin opiskelija on "selviytynyt" valmistumatta ajanhetkeen t saakka. Välttöfunktio voidaan esittää kertymäfunktion avulla muodossa

$$S(t) = 1 - F(t) = \sum_{u=t+1}^{\infty} p_u. \tag{6}$$

Vaarafunktiolla (*hazard function*) kuvataan hetkellistä vaaraa ajanhetkellä t. Hetkellinen vaara tarkoittaa todennäköisyyttä, että valmistuminen tapahtuu ajanhetkellä t ehdolla, että tutkittava ei ole valmistunut ennen ajanhetkeä t. Matemaattisesti vaarafunktio on muotoa

$$h_t = P(t|T \geq t) = \frac{P(T = t)}{P(T \geq t)} = \frac{P(T = t)}{P(T > (t-1))} = \frac{p_t}{S(t-1)},\ t \geq 1. \tag{7}$$

Sekä välttöfunktio että pistetodennäköisyydet voidaan esittää vaihtoehtoisesti vaarafunktion avulla. Tällöin välttöfunktio $S(t)$ on muotoa

$$S(t) = \prod_{u=1}^{t}(1 - h_u) \tag{8}$$

ja pistetodennäköisyys p_t on muotoa

$$p_t = h_t S(t-1) = h_t \prod_{u=1}^{t-1}(1 - h_u). \tag{9}$$

4.2 Sensuroituminen

Elinaika-aineistoissa on tyypillistä, että osa havainnoista on sensuroituneita. Havainnon sensuroitumisella tarkoitetaan tilannetta, jossa kiinnostavaa tapahtumaa ei ole havaittu. Havainnon sensurointia on kolmenlaista: oikealta sensuroitumista, vasemmalta sensuroitumista tai välisensuroitumista.

Oikealta sensuroituneesta havainnosta tiedetään vain, että elinaika $T_i > c_i$, missä sensurointihetki c_i on tunnettu. Vasemmalta sensuroituneessa havainnossa elinaika $T_i < c_i$, missä niin ikään sensurointihetki c_i on tunnettu. Välisensuroinnissa sensurointia on tapahtunut sekä vasemmalta että oikealta. Tällöin tiedetään ainoastaan, että $c_{oi} < T_i < c_{vi}$, missä sensurointihetket c_{oi} ja c_{vi} ovat tunnettuja. Tämän tutkielman aineistossa esiintyy ainoastaan oikealta sensuroituneita havaintoja. Sensuroituneen havainnon tapauksessa opiskelijasta tiedetään ainoastaan opintojen alkamisajankohta, mutta valmistumista ei olla havaittu.

Oikealta sensuroituneet havainnot voidaan esittää havaintoparina (t_i, δ_i), missä

$$t_i = \min\{T_i, c_i\} \tag{10}$$

ja

$$\delta_i = \begin{cases} 1, & T_i \;\; = t_i \\ 0, & T_i \;\; > c_i. \end{cases} \tag{11}$$

Havaintoparissa (t_i, δ_i) ajanhetki t_i kertoo valmistumisajan ja δ_i sen, onko kyseessä valmistuminen ($\delta_i = 1$) vai oikealta sensuroituminen ($\delta_i = 0$). Sensuroituneelle aineistolle niin sanottu Kaplan-Meierin (1958) välttöfunktion estimaatti on

$$\hat{S}(t) = \prod_{i:\ t_i \le t} \left[1 - \frac{d_i}{n_i}\right]^{\delta_i}, \tag{12}$$

missä d_i on valmistuneiden lukumäärä ajanhetkellä t_i ja n_i on valmistumattomien tai sensuroimattomien opiskelijoiden lukumäärä ajanhetkellä t_i.

Tapahtumien keräämistä tutkimuksen aikana voidaan kontrolloida monella eri tavalla. Elandt-Johnsonin ja Johnsonin (1980) mukaan havainnoinnin lopettamiseen on kaksi yleisintä tutkimuksen toteutukseen liittyvää prosessia. Ensimmäisessä prosessissa havainnointi ja tapahtumien kerääminen lopetetaan ennaltamäärätyn ajankohdan jälkeen. Toisessa

prosessissa havainnointi ja tapahtumien kerääminen lopetetaan, kun saavutetaan ennaltamääritelty määrä tapahtumia. Huomioitavaa on, että edellä esitellyissä prosesseissa satunnaisuus kohdistuu eri tekijöihin. Ensimmäisen tyypin prosessissa tapahtumien lukumäärä on satunnainen, kun taas toisessa prosessissa havainnointiaika on satunnaistettu. Tutkielman aineiston tapauksessa kyse on ensimmäisen tyypin prosessista: tutkimusaineiston kerääminen lopetettiin 31. heinäkuuta 2020, jonka jälkeen mahdollisia valmistumisia ei ole havaittu.

5 Coxin regressiomalli

Tässä luvussa käydään läpi Coxin (1972) regressiomallin teoria. Coxin regressiomalli, joka tunnetaan myös nimellä Coxin suhteellisen vaaran malli, esitellään alaluvussa 5.1. Alaluvussa 5.2 johdetaan Coxin regressiomallin uskottavuusfunktio ja yleistetään se koskemaan sensuroitunutta elinaika-aineistoa. Lopuksi alaluvussa 5.3 tarkastellaan Coxin regressiomallin diagnostiikkaa.

5.1 Coxin regressiomalli yleisesti

Coxin regressiomalli on semiparametrinen malli. Mallissa oletetaan suhteelliset vaarat, mutta elinaikojen jakaumasta ei tehdä oletuksia. Coxin regressiomallin etuna on, että sillä saavutetaan tieto selittävien muuttujien vaikutuksesta vaarafunktioon. Coxin (1972) mallin vaarafunktio on muotoa

$$h_i(t) = \lambda_i h_0(t),\ i = 1, ..., n, \tag{13}$$

missä $\lambda_i = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})$, x_{ij} kuvaa j:nnen selittäjän arvoa i:nnellä opiskelijalla, $\beta_j, j = 1, 2, ...p$, kertoo estimoidun regressiokertoimen arvon j:nnelle selittäjälle ja $h_0(t)$ on referenssiluokkaan liittyvä vaarafunktio (perusvaarafunktio). Vaarojen suhteellisuuden näkee selkeämmin yhtälön (13) esitysmuodosta

$$\frac{h_i(t)}{h_0(t)} = \lambda_i,\ i = 1, ..., n,$$

missä λ_i on kuten edellä. Tällöin $\exp(\hat{\beta}_j)$ kertoo, moninkokertaiseksi vaara muuttuu, kun j:nnen selittäjän arvo muuttuu yhden mittayksikön verran. Satunnaismuuttujaan T_i liittyvä välttöfunktio on muotoa

$$S_i(t) = S_0(t)^{\lambda_i}.$$

Seuraavaksi johdetaan Coxin mallin uskottavuusfunktio. Vertailtaessa kahden satunnaismuuttujan T_1 ja T_2 ajankohtia voidaan osoittaa (Cox, 1984), että

$$P(T_1 < T_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Yleisemmin todennäköisyys sille, että T_1 on lyhin valmistumisaika, on

$$P(T_1 < \{T_2, ..., T_n\}) = \frac{\lambda_1}{\lambda_1 + ... + \lambda_n}. \tag{14}$$

Edellä olevat todennäköisyydet eivät siis riipu perusvaarafunktion muodosta.

Yhtälöstä 14 voidaan laajentaa todennäköisyys kaikkien elinaikojen järjestykselle. Mahdollisia eri järjestyksiä eli permutaatioita elinaikojen ja tapahtumien järjestykselle on yhteensä havaintojen lukumäärän n kertoman $n!$ verran. Todennäköisyys tapahtumien ajalliselle järjestykselle $T_{i_1} < T_{i_2} < ... < T_{i_n}$ i:nnessä permutaatiossa $(i_1, ..., i_n)$ on tällöin (Cox, 1984)

$$P(T_{i_1} < T_{i_2} < ... < T_{i_n}) = \frac{\lambda_{i_1}}{\lambda_{i_1} + \lambda_{i_2} + ... + \lambda_{i_n}} \cdot \frac{\lambda_{i_2}}{\lambda_{i_2} + \lambda_{i_3} + ... + \lambda_{i_n}} \cdot ... \cdot \frac{\lambda_{i_n}}{\lambda_{i_{n-1}} + \lambda_{i_n}}. \quad (15)$$

Merkitään järjestettyjä valmistumisaikoja $t_{(1)} < t_{(2)} < ... < t_{(n)}$, missä $t_{(i)}$ on järjestyksessään i:nnes valmistumisaika. Merkitään lisäksi $R(t_{(i)})$ tarkoittamaan riskijoukkoa eli niiden yksilöiden joukkoa, jotka ovat valmistumatta ja sensuroimattomia ajanhetkellä $t_{(i)}$.

Cox (1972) esitteli suhteellisen vaaran mallin uskottavuusfunktion tapahtumien ajallisten järjestysten todennäköisyyksien tulona, kuten kaavassa 15

$$L(\beta) = \prod_{i=1}^{n} \frac{\lambda_i}{\sum_{j \in R(t_{(i)})} \lambda_j}.$$

Uskottavuusfunktio riippuu ainoastaan valmistumisaikojen järjestyksestä.

5.2 Coxin regressiomalli sensuroituneelle aineistolle

Oletetaan, että elinaika-aineistossa on sekä havaittuja että sensuroituneita havaintoja. Edellisessä alaluvussa esitetty uskottavuusfunktio voidaan yleistää koskemaan sensuroitua aineistoa. Tällöin uskottavuusfunktio voidaan kirjoittaa muodossa (Collett, 2015)

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\lambda_i}{\sum_{j \in R(t_{(i)})} \lambda_j} \right\}^{\delta_i}, \quad (16)$$

missä δ_i saa arvon 1, mikäli tapahtuma on havaittu ja arvon 0, mikäli havainto on sensuroitunut. Mallinvalinnan yhteydessä käytetään logaritmoitua uskottavuusfunktiota, joka on muotoa

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \log \lambda_i - \log \sum_{j \in R(t_{(i)})} \lambda_j \right\}. \quad (17)$$

Uskottavuusfunktion 17 ratkaisemiseksi käytetään numeerisia menetelmiä, kuten Newton-Raphsonin menetelmää. Newton-Raphsonin menetelmän soveltaminen Coxin suhteellisen vaaran malliin on esitelty muun muassa Collettin (2015) teoksessa luvussa 3.3.3. Coxin regressiomallin sovittamisessa aineistoon käytettiin `survival`-paketin funktiota `coxph` (Therneau, 2020).

5.3 Coxin regressiomallin diagnostiikkaa

Coxin mallin sopivuutta aineistoon tarkastellaan jäännösten avulla. Jäännösten matemaattiset esitystavat on esitetty Collettin (2015) kirjaa mukaillen. Jäännöstarkasteluissa käytetyin jäännösesitys on Cox-Snellin jäännökset, jotka on nimetty Coxin ja Snellin (1968) mukaan. Cox-Snellin jäännökset esitetään muodossa

$$r_{C_i} = \exp(\lambda_i)\hat{H}_0(t_i), \quad (18)$$

missä $\lambda_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$. Lisäksi

$$\hat{H}_0(t_i) = \sum_{t_{(r)} \leq t_i} \frac{\delta_{(r)}}{\sum_{j \in R_{(r)}} \exp(\lambda_j)}$$

on aineistosta estimoitu kumulatiivinen vaarafunktio ajanhetkellä t_i, kun λ_j on kuten edellä. Mikäli malli on sopiva, Cox-Snellin jäännökset noudattavat eksponenttijakaumaa odotusarvolla yksi.

Odotusarvoltaan nollan olevan martingaalijäännöksen esittelivät Therneau, Grambsch sekä Fleming (1990). Martingaalijäännökset esitetään muodossa

$$r_{M_i} = \delta_i + r_{C_i}, \tag{19}$$

missä r_{C_i} ja δ_i ovat kuten edellä. Martingaalijäännökset korjaavat jäännökset nollakeskeisiksi, mutta symmetrisyys nollan suhteen puuttuu edelleen. Tähän ongelmaan Therneau ym. (1990) esittelivät devianssijäännökset, jotka ovat symmetrisempiä nollan suhteen. Devianssijäännös esitetään muodossa

$$r_{D_i} = \text{sgn}(r_{M_i}) \left[-2r_{M_i} + \log(\delta_i - r_{M_i}) \right]^{\frac{1}{2}}, \tag{20}$$

missä

$$\text{sgn}(r_{M_i}) = \begin{cases} -1, & r_{M_i} < 0 \\ 0, & r_{M_i} = 0 \\ 1, & r_{M_i} > 0. \end{cases}$$

Devianssijäännökset perustuvat nimensä mukaisesti devianssilaskennan periaatteisiin. Devianssijäännökset asettuvat symmetrisesti nollan molemmin puolin, mikäli malli on sopiva.

Edellä mainittuihin jäännöksiin liittyy kuitenkin ongelmia. Jäännökset riippuvat vahvasti havaitusta elinajasta. Lisäksi jäännösten laskeminen vaatii kumulatiivisen vaarafunktion estimoimista. Schoenfeld (1980) esitteli oman versionsa jäännöksistä. Schoenfeldin jäännökset selittäjälle j ovat muotoa

$$r_{P_{ji}} = \delta_i x_{ji} - \hat{a}_{ji}, \ j = 1, 2, ..., \tag{21}$$

missä x_{ji} on j:nnen selittäjän arvo i:nnelle henkilölle. Lisäksi

$$\hat{a} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp(\lambda_l)}{\sum_{l \in R(t_i)} \exp(\lambda_l)},$$

missä $\lambda_l = \hat{\beta}_1 x_{1l} + \hat{\beta}_2 x_{2l} + ... + \hat{\beta}_p x_{pl}$ ja $R(t_i)$ on riskijoukko ajanhetkellä t_i. Huomioitavaa on, että sensuroituneiden havaintojen jäännökset jäävät nolliksi.

6 Mallinvalinta

Tässä luvussa esitellään sisäkkäisten mallien vertailussa käytettävä testisuure sekä niin kutsuttu Akaiken infromaatiokriteeri. Näiden lisäksi mallinvalintamenetelmistä esitellään manuaalinen menetelmä sekä automatisoituja mallinvalintamenetelmiä. Sisäkkäisten ja ei-sisäkkäisten mallien vertailun teoria esitellään alaluvussa 6.1. Mallinvalintamenetelmiä esitellään alaluvussa 6.2.

6.1 Sisäkkäisten ja ei-sisäkkäisten mallien vertailu

Vertailtaessa sisäkkäisiä malleja M_1 ja M_2 voidaan testisuureena käyttää logaritmoitujen uskottavuusfunktion arvojen erotusta (Elandt-Johnson ja Johnson, 1980). Olkoon mallissa M_1 p parametria ja mallissa M_2 on $p+q$ parametria. Oletetaan, että molemmissa malleissa on numeerisesti maksimoitu logaritmoidun uskottavuusfunktion (17) arvo. Tällöin mallien vertailuun voidaan käyttää testisuuretta

$$\chi^2(q) = -2\log\hat{L}_1 + 2\log\hat{L}_2, \tag{22}$$

missä

$$\hat{L}_1 = L(\hat{\beta}_1, ..., \hat{\beta}_p)$$

ja

$$\hat{L}_2 = L(\hat{\beta}_1, ..., \hat{\beta}_{p+q}).$$

Testisuureen avulla testataan nollahypoteesiä H_0, jonka mukaan lisätyt selittäjät $p+1, p+2, p+3, \ldots, p+q$ eivät vaikuta valmistumisaikaan. Toisin sanoen selittäjien lisääminen ei paranna mallin sopivuutta aineistoon, ja suppeampi malli M_1 on riittävä. Testisuuretta verrataan χ^2-jakaumaan vapausasteella q. Sisäkkäisten mallien vertailu edellyttää, että mallit on sovitettu samaan aineistoon.

Ei-sisäkkäisiä malleja voidaan vertailla Akaiken (1998) informaatiokriteerin avulla. Akaiken informaatiokriteeri lasketaan

$$AIC = -2\log\hat{L} + 2p,$$

missä $\hat{L}$ on mallin uskottavuusfunktion arvo ja p on parametrien (selittäjien) lukumäärä. Mitä pienempi AIC, sitä paremmin malli sopii aineistoonsa.

6.2 Mallinvalintamenetelmistä

Seuraavaksi tarkastellaan erilaisia strategioita mallinvalinnan suorittamiseksi. Koska mahdollisia selittäviä muuttujia on paljon, on syytä harkita automatisoitujen valintarutiinien käyttämistä mallin selittäjien valinnassa. Kaikkien mahdollisten muuttujien ja mallien läpikäyminen manuaalisesti on haastavaa ja hidasta. Automatisoituihin mallinvalintamenetelmiin täytyy suhtautua kuitenkin kriittisesti, sillä niiden valitsemat selittäjät riippuvat usein käytetystä menetelmästä (Collett, 2015). Lisäksi menetelmä suhtautuu valittuun kriteeriin äärimmäisen tarkasti eikä ota huomioon mallien hierarkiaa: mikäli mallissa on

mukana kahden selittäjän välinen interaktiotermi, myös suorat selittäjät pitäisi sisällyttää malliin (Collett, 2015). Automatisoitu mallinvalinta tutkielman aineistolle toteutettiin `MASS`-paketin funktiolla `StepAIC` (Venables ja Ripley, 2002).

Yleisesti käytettyjä automatisoituja mallinvalintarutiineja on kolme: etenevä valinta (*forward selection*), poistovalinta (*backward elimination*) ja askeltava menettely (*stepwise procedure*). Seuraavat esitykset perustuvat Collettin (2015) teoksen esityksiin. Etenevässä valinnassa selittäjiä lisätään malliin yksi kerrallaan. Jokaisessa vaiheessa lisätty muuttuja on se, joka laskee arvoa $-2 \log \hat{L}$ eniten. Prosessi päättyy silloin, kun muuttujan lisääminen ei laske arvoa tarpeeksi. Tämä on niin kutsuttu pysäytyssääntö (*stopping rule*). Yleensä sääntö on ennaltamääritelty merkitsevyystaso p (esimerkiksi 0.05). Poistovalinnan alussa sovitetaan suurin mahdollinen malli, jossa on mukana kaikki selittäjät. Selittäjiä poistetaan mallista yksi kerrallaan. Jokaisessa vaiheessa mallista poistetaan se muuttuja, joka lisää $-2 \log \hat{L}$ -arvoa vähiten. Prosessi lakkaa, kun selittäjän poistaminen nostaa arvoa ennaltamääriteltyä, sallittua määrää enemmän. Askeltavassa menettelyssä yhdistetään kaksi edellistä menetelmää. Prosessi operoi samaan tapaan kuin etenevä valinta, mutta malliin lisätyt muuttujat voidaan poistaa prosessin edetessä. Selittäjän lisäämisen jälkeen prosessissa tarkistetaan, voidaanko jokin aiemmin lisätty selittäjä poistaa. Nämä valinnat perustuvat jälleen ennalta kiinnitettyihin pysäytyssääntöihin.

Kirjassaan Collett (2015) esittelee automatisoitujen mallivalintarutiinien vastineeksi oman mallinvalintastrategiansa, jossa on yhteensä neljä eri vaihetta. Collettin mukaan (2015) manuaalisen mallinvalinnan etuna on, että menettelyn avulla voidaan löytää useampi yhtä hyvä malli. Koska tässä mallinvalintamenettelyssä tutkija itse suorittaa mallien vertailun, ei tarkkaa merkitsevyystasoa tarvitse kiinnittää etukäteen (Collett, 2015). Collett (2015) suosittelee sopivaksi merkitsevyystasoksi p noin 0.10.

Ensimmäisessä vaiheessa sovitetaan kaikki mahdolliset yhden selittäjän mallit. Sen jälkeen arvoa $-2 \log \hat{L}$ verrataan niin kutsuttuun nollamalliin, jossa on pelkkä vakio selittäjänä. Jatkoon valitaan ne selittäjät, jotka laskevat arvoa $-2 \log \hat{L}$ tilastollisesti merkitsevästi. Toisessa vaiheessa kaikki ne selittäjät, jotka osoittautuivat tilastollisesti merkitseviksi ensimmäisessä vaiheessa, sovitetaan samaan malliin. Ne selittäjät, joiden poissaolo ei merkitsevästi nosta arvoa $-2 \log \hat{L}$, voidaan poistaa mallista. Lasketaan siis arvon muutos, kun jokainen muuttuja yksi kerrallaan poistetaan mallista. Ne muuttujat, joiden poissaolo nostaa tilastollisesti merkitsevästi arvoa $-2 \log \hat{L}$, jätetään malliin. Kun yksi muuttuja on pudotettu mallista pois, jäljellä oleville muuttujille toistetaan tämä vaihe uudelleen: poistetaan mallista, verrataan arvoa, ja niin edelleen. Kolmannessa vaiheessa ne muuttujat, jotka eivät olleet yksinään tilastollisesti merkitseviä ensimmäisessä vaiheessa, ja jotka eivät siten olleet mukana toisessa vaiheessa, lisätään nyt malliin yksi kerrallaan. Jos muuttujan lisääminen laskee tilastollisesti merkitsevästi arvoa $-2 \log \hat{L}$, muuttuja saa jäädä malliin. Tässä vaiheessa jokin muuttuja, joka on säilynyt mallissa toiseen vaiheen jälkeen, saattaa muuttua merkitsemättömäksi. Neljännessä vaiheessa suoritetaan viimeinen tarkistus. Tarkistetaan, että yhtäkään muuttujaa ei voida poistaa mallista ilman, että arvo $-2 \log \hat{L}$ nousee tilastollisesti merkitsevästi ja yhtäkään muuttujaa ei voida lisätä malliin siten, että arvo enää vähenisi.

7 Aineiston analyysi

Tässä luvussa opiskelijat ryhmitellään ylioppilasarvosanojen perusteella ja valmistumisaikaa selitetään hyödyntämällä ryhmittelyä Coxin regressiomallissa. Ryhmätiedon lisäksi mallia parannetaan lisäämällä malliin muita hakuvaiheen selittäjiä. Mallien vertailussa käytetään sisäkkäisten mallien testausta. Mallinvalintamenetelmistä käytetään manuaalista mallinvalintaa.

Ryhmittelyn tulokset esitellään tarkemmin seuraavassa alaluvussa 7.1. Coxin regressiomallien tulokset esitellään alaluvuissa 7.2 ja 7.3. Mallien hyvyystarkasteluja käsitellään alaluvussa 7.4.

7.1 *K*:n keskiarvon ryhmittelyn tulokset

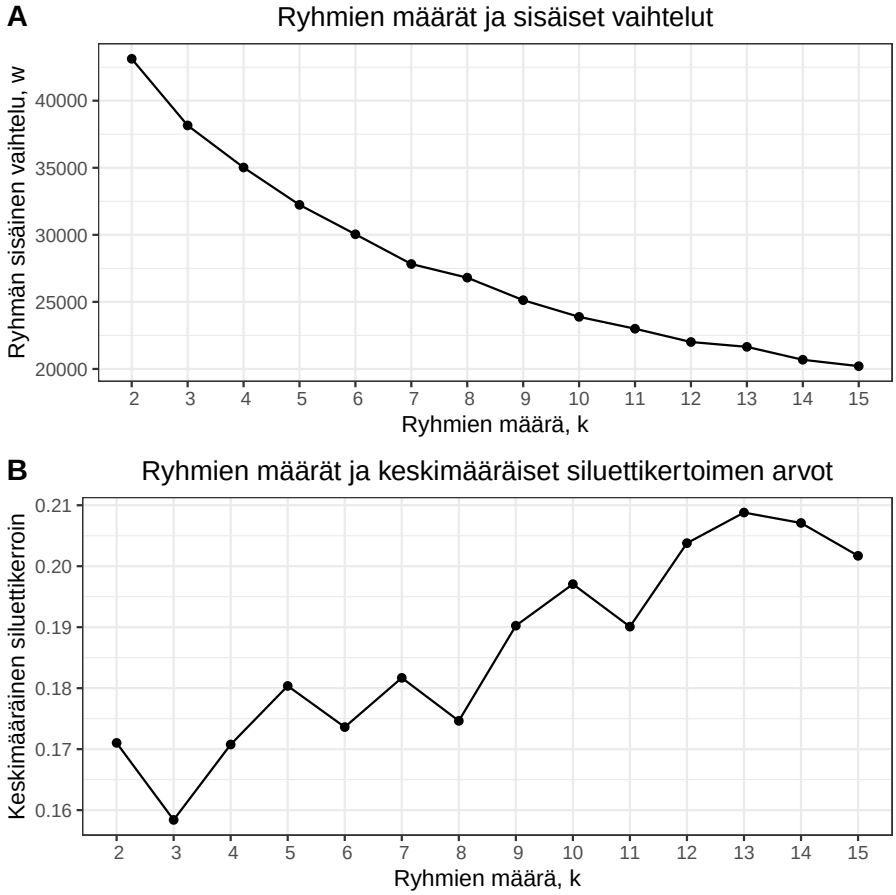
Aineisto ryhmiteltiin ylioppilasarvosanojen perusteella käyttäen k:n keskiarvon ryhmittelymenetelmää. Opiskelijat ryhmiteltiin äidinkielen, matematiikan, biologian, kemian, fysiikan, ruotsin keskipitkän oppimäärän, englannin pitkän oppimäärän, psykologian, terveystiedon ja maantieteen ylioppilaskirjoituksissa saavutettujen arvosanojen perusteella. Ryhmittelyssä oli mukana 96 % aineiston opiskelijoista, yhteensä 1209 havaintoriviä. Opiskelijoista 4 % karsiutui pois ryhmittelystä, sillä heiltä ei löytynyt arvosanaa yhdestäkään ryhmittelyssä käytetystä ylioppilaskirjoitusaineesta. Huomioitavaa on, että ryhmien järjestys ja tulokset riippuvat osittain numeerisesta ratkaisusta. Jokaisella k:n arvolla ryhmittely toistettiin viisi kertaa, ja jokaisesta toistosta valittiin sisäisen vaihtelun minimoiva ryhmittely. Ryhmien sisäisissä hajontatarkasteluissa päädyttiin $k = 7$ ryhmään. Ryhmien määrän ja sisäisen vaihtelun välisessä kuviossa on havaittavissa niin kutsuttu nivelkohta $k = 7$ kohdalla, jonka jälkeen sisäinen vaihtelu ei enää pienene merkittävästi (kuvio 3 (A)).

Sisäisen vaihtelun tarkastelun lisäksi ryhmien lukumäärän valintaa tutkittiin keskimääräisen siluettikertoimen avulla. Menetelmä on esitelty tarkemmin tutkielman luvussa 3. Myös keskimääräisen siluettikertoimen tarkastelun jälkeen päädyttiin $k = 7$ ryhmään. Keskimääräisen siluettikertoimen osalta tarkasteltiin, että kun $k = 7$, siluettikerroin saavuttaa lokaalin maksimin. Lisäksi sisäisen vaihtelun tarkastelu tuki ryhmälukumäärän valintaa. Keskimääräisen siluettikertoimen ja ryhmien määrän k välinen kuvaaja on esitelty kuviossa 3 (B).

Tarkastellaan ensimmäiseksi opiskelijoiden valmistumista ryhmissä. Ryhmien sisällä tarkastellaan ainoastaan niitä opiskelijoita, joiden elinaika on yhtä suuri tai enemmän kuin kuusi lukukautta. Tällöin voidaan laskea valmistuneiden osuus ryhmissä niiden opiskelijoiden kesken, jotka ovat saavuttaneet tavoiteajan. Opiskelijoiden lukumäärät ryhmittäin sekä ryhmien valmistumisprosentteja on esitelty taulukossa 8.

Kahdessa ryhmässä valmistuneiden osuus niiden kesken, jotka ovat opiskelleet tasan tai yli kuuden lukukauden ajan, ylittää 50 prosenttiyksikköä. Ryhmässä 6 valmistuneiden osuus on 0.58, ja ryhmässä 5 vastaava osuus on 0.50. Ryhmissä 4 ja 1 valmistuneiden osuudet ovat lähes 0.50. Ryhmien välisessä vertailussa matalin valmistumisprosentti on ryhmässä 2, jossa valmistuneiden osuus tavoiteajan saavuttaneista on 0.24. Jatkossa ryhmittelyn ja elinaikamallin tulokset esitellään taulukon 8 ryhmäjärjestyksessä.

Ryhmien ylioppilasarvosanojen keskiarvot on esitelty taulukossa 9 ja kuviossa 4. Keskiarvon laskennassa ovat mukana myös ne ryhmän opiskelijat, jotka eivät ole suorittaneet kyseisen ylioppilaskirjoitusaineen koetta hyväksytysti. Kirjoittamatta jättäminen merkittiin aineistossa nollana. Näin ollen taulukossa 9 esitetyt keskiarvot ovat hieman matalammat



Kuvio 3: Ryhmien määrät k ja niitä vastaavat ryhmien sisäiset vaihtelut (kuvio A) sekä keskimääräiset siluettikertoimet (kuvio B).

verrattuna tilanteeseen, jossa arvosanojen keskiarvot olisi laskettu vain kirjoittaneiden kesken. Ryhmässä paljon kirjoitettujen, korkeiden keskiarvojen oppiaineiden osalta vaikutus keskiarvoon oli kuitenkin vähäinen (ei esitetty). Taulukossa 10 on esitelty jokaisen tutkittavan ylioppilaskirjoitusaineen kirjoittaneiden osuudet ryhmittäin.

Tarkastellaan seuraavaksi ryhmiä valmistumisprosenttien mukaisessa järjestyksessä korkeimmasta matalimpaan. Ryhmässä 6 valmistuneiden osuus tavoiteajan saavuttaneissa on korkein. Ryhmässä 6 on opiskelijoita, joilla on korkeat keskiarvot äidinkielestä, biologiasta, englannista ja maantieteestä. Äidinkielen, biologian ja maantieteen keskiarvot ovat ryhmässä yli 5.0, mikä vastaa ylioppilaskirjoitusten kirjainarvosanaa *magna cum laude approbatur* (M). Ryhmässä on paljon äidinkielen, biologian, englannin pitkän oppimäärän ja maantieteen kirjoittaneita opiskelijoita (taulukko 10). Kaikki ryhmän 6 opiskelijat ovat suorittaneet äidinkielen ja maantieteen ylioppilaskokeen hyväksytysti.

Ryhmässä 5 on korkeat keskiarvot matematiikan ja englannin ylioppilaskirjoitusarvosanoissa. Ryhmän opiskelijoista vain 8 % on suorittanut kemian ylioppilaskokeen hyväksytyllä arvosanalla. Ryhmässä on lisäksi korkea keskiarvo ruotsin keskipitkässä oppimäärässä (keskiarvo 5.2). Ryhmässä on korkein psykologian keskiarvo (keskiarvo 1.5) ryhmien välisessä vertailussa. Ryhmän opiskelijoista 28 % on suorittanut psykologian ylioppilaskokeen hyväksytysti.

Ryhmät 4 ja 1 ovat menestyneet samantasoisesti matematiikan ylioppilaskirjoituksissa. Ryhmän 4 keskiarvo matematiikassa on 5.5 ja ryhmän 1 keskiarvo on 5.3. Sen sijaan ryh-

Taulukko 8: Opiskelijoiden lukumäärät yhteensä ryhmittäin, tavoiteajan saavuttaneiden lukumäärät ryhmittäin sekä valmistumisaikojen keskiarvot ja valmistumisprosentit ryhmittäin niiden opiskelijoiden osalta, jotka ovat saavuttaneet tavoiteajan (6 lukukautta) tai valmistuneet ennen tavoiteajan saavuttamista.

Ryhmä	Yhteensä	Tavoiteajan saavuttaneet	Valmistumisprosentti	Valmistumisajan keskiarvo
6	106	52	0.58	6.88
5	143	56	0.50	6.95
4	126	56	0.48	6.62
1	200	64	0.45	6.77
3	208	65	0.42	6.86
7	188	66	0.36	6.97
2	238	90	0.24	7.21

mässä 1 on korkeammat keskiarvot kemiassa ja englannissa. Ryhmän 4 keskiarvo englannin pitkässä oppimäärässä on ryhmien välisessä vertailussa matalin, 3.7, mikä vastaisi kirjainarvosanoissa korkeaa arvosanaa *lubenter approbatur* (B). Ryhmässä 4 on toisaalta korkein terveystiedon keskiarvo (keskiarvo 5.4) ja kaikki ryhmän opiskelijat ovat suorittaneet kyseisen ylioppilaskokeen hyväksytysti (taulukko 10). Ryhmässä 3 on taas korkeat keskiarvot matematiikassa, biologiassa, kemiassa, fysiikassa ja englannin pitkässä oppimäärässä verrattaessa muihin ryhmiin. Ryhmän 3 opiskelijoista kaikki ovat osallistuneet biologian sekä fysiikan ylioppilaskokeeseen ja suorittaneet sen hyväksytysti.

Ryhmä 7 erottuu perinteisten luonnontieteiden korkeilla keskiarvoilla. Matematiikan keskiarvo on 7.0, mikä kertoo siitä, että ryhmässä on paljon pitkän matematiikan korkealla arvosanalla kirjoittaneita opiskelijoita. Taulukosta 10 nähdään, että 99 % ryhmän 7 opiskelijoista on suorittanut matematiikan ylioppilaskokeen hyväksytysti. Lisäksi fysiikan ja kemian keskiarvot ovat molemmat yli 5.0. Myös englannin pitkän oppimäärän keskiarvo on ryhmien välisessä vertailussa korkein (keskiarvo 4.8). Ryhmässä on hyvin vähän psykologian, terveystiedon ja maantieteen kirjoittaneita opiskelijoita. Biologian hyväksytyllä arvosanalla kirjottaneita ei ole ryhmässä lainkaan (keskiarvo 0).

Ryhmässä 2 on matalin keskiarvo äidinkielessä. Ryhmän keskiarvosana äidinkielen ylioppilaskirjoituksista on 3.6, joka vastaa kirjainarvosanaa B. Ryhmän keskiarvo matematiikassa on suhteellisen korkea (keskiarvo 5.9). Ryhmässä on jonkin verran myös kemian ja fysiikan kirjoittaneita opiskelijoita, mutta näiden aineiden keskiarvot eivät nouse korkeiksi (kemian keskiarvo 1.1, fysiikan keskiarvo 2.9).

Taulukko 9: Ylioppilaskirjoitusten arvosanakeskiarvot ryhmittäin, kun mukana on myös ne ryhmän opiskelijat, jotka eivät ole kirjoittaneet kyseistä ylioppilasainetta.

Ryhmä	AI	MAT	BI	KE	FY	RUB	ENA	PSY	TT	MAAN
6	5.0	4.7	5.1	0.7	0.3	2.1	4.7	0.2	0.3	5.6
5	5.1	6.1	1.8	0.3	1.0	5.2	4.7	1.5	0.3	0.2
4	4.9	5.5	3.8	1.6	0.6	1.8	3.7	1.2	5.4	0.2
1	4.4	5.3	4.8	3.8	0.2	1.1	4.3	0.4	0.1	0.0
3	4.9	6.1	5.2	4.7	4.9	1.7	4.5	0.1	0.2	0.2
7	4.5	7.0	0.0	5.3	5.0	1.7	4.8	0.1	0.1	0.2
2	3.6	5.9	0.4	1.1	2.9	0.2	4.4	0.7	0.3	0.3

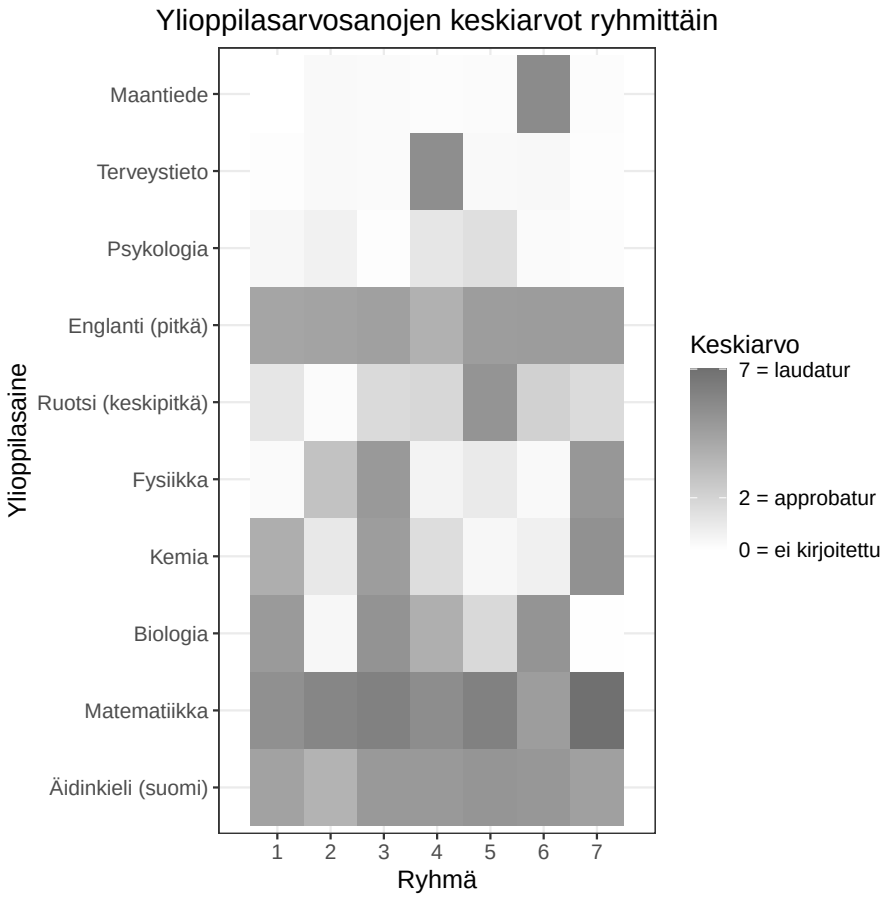
Taulukko 10: Kirjoittaneiden ja hyväksytyn arvosanan saavuttaneiden opiskelijoiden osuudet ylioppilasaineittain ja ryhmittäin. Taulukossa on laskettu osuudet ryhmittäin.

Ryhmä	AI	MAT	BI	KE	FY	RUB	ENA	PSY	TT	MAAN
6	1.00	0.88	0.95	0.17	0.07	0.42	0.98	0.05	0.07	1.00
5	0.99	0.94	0.34	0.08	0.20	1.00	0.92	0.28	0.07	0.04
4	0.98	0.96	0.71	0.35	0.13	0.40	0.87	0.23	1.00	0.03
1	0.99	0.97	0.99	0.84	0.08	0.24	0.91	0.07	0.02	0.00
3	0.99	0.99	1.00	0.93	1.00	0.34	0.93	0.02	0.05	0.04
7	1.00	0.99	0.02	0.99	0.96	0.34	0.95	0.03	0.02	0.03
2	0.95	0.98	0.11	0.32	0.71	0.07	0.92	0.15	0.07	0.06

Ryhmien laitosjakaumasta nähdään, että erityisesti ryhmissä 6, 4 ja 1 on paljon bio- ja ympäristötieteen laitoksen opiskelijoita (ks. taulukko 11). Ryhmässä 2 esiintyy eniten fysiikan sekä matematiikan ja tilastotieteen laitoksen opiskelijoita. Ryhmässä 5 on eniten matematiikan ja tilastotieteen laitoksen opiskelijoita. Ryhmässä 3 on eniten kemian laitoksen opiskelijoita, mutta myös bio- ja ympäristötieteen sekä fysiikan laitoksen opiskelijoita. Ryhmässä 7 on eniten fysiikan laitoksen opiskelijoita. Ryhmässä on tasaisesti sekä kemian että matematiikan ja tilastotieteen laitoksen opiskelijoita. Ryhmässä on vain muutama bio- ja ympäristötieteiden laitoksen opiskelija.

Taulukko 11: Laitosten opiskelijoiden jakautuminen ryhmittäin.

Ryhmä	Bio- ja ymp.	Fysiikka	Kemia	Mat ja til.
6	93	3	3	7
5	46	15	11	71
4	63	7	26	30
1	103	1	77	19
3	64	57	66	21
7	4	77	59	48
2	11	95	29	103



Kuvio 4: Ylioppilasarvosanojen keskiarvot ryhmittäin.

Tarkasteltaessa ryhmien hakutoivetta kuvaavan prioriteetin jakaumaa (taulukko 12) huomataan, että lähes kaikissa ryhmissä on eniten ensimmäisen prioriteetin opiskelijoita. Toisaalta esimerkiksi ryhmässä 2 eri hakutoivejärjestykset ovat jakautuneet laajalle. Ryhmissä 3 ja 6 on vain vähän matalampien prioriteettien opiskelijoita.

Tarkastellaan seuraavaksi valintajonojen jakaumaa ryhmissä (taulukko 13). Ryhmässä 6 on eniten ylioppilaspistejonon ja yhteispistejonon kautta valittuja opiskelijoita. Ryhmässä 5 on niin ikään eniten ylioppilaspistejonon kautta valittuja opiskelijoita. Ryhmissä 4 ja 1 on muihin ryhmiin nähden eniten valintakoejonon kautta tulleita opiskelijoita. Puolestaan ryhmissä 3 ja 7 on eniten suoravalintajonon kautta tulleita opiskelijoita verrattaessa muihin ryhmiin. Ryhmässä 7 suoravalittujen opiskelijoiden osuus on lähes yksi kolmannesta. Ryhmässä 2 on eniten ylioppilaspistejonon kautta valittuja opiskelijoita.

Taulukko 12: Opiskelijoiden hakutoiveiden jakautuminen ryhmittäin.

Ryhmä	1	2	3	4	5	6	NA
6	56	24	16	2	2	5	1
5	61	38	19	12	8	3	2
4	43	34	19	12	9	8	1
1	56	55	24	29	15	18	3
3	60	52	32	29	19	14	2
7	127	23	22	8	6	2	
2	117	61	31	11	9	7	2

Taulukko 13: Opiskelijoiden valintajonojen jakautuminen ryhmittäin.

Ryhmä	suoravalintajono	valintakoejono	yhteishaun ulkopuolelta	yhteispistejono	yopistejono
6	<5	17	<5	21	63
5	17	14	<5	<5	104
4	8	26	7	17	68
1	13	40	<5	25	118
3	44	16	<5	30	115
7	60	<5	<5	<5	121
2	29	19	5		185

7.2 Valmistumisajan selittäminen arvosanaryhmillä

Tässä alaluvussa tarkastellaan, miten ryhmätieto vaikuttaa opintojen aloittamisen ja kandidaatiksi valmistumisen välisen ajan pituuteen. Valmistumisen vaaraa verrataan referenssiluokkaan, joka on tässä tapauksessa ryhmä 6. Ryhmä valittiin referenssiryhmäksi sillä perusteella, että ryhmän sisäinen valmistumisprosentti on tavoiteajan saavuttaneiden opiskelijoiden keskuudessa suurin.

Merkitään arvosanaryhmää 1 merkinnällä $R1$, arvosanaryhmää 2 merkinnällä $R2$ ja niin edelleen. Tällöin Coxin suhteellisen vaaran malli on muotoa

$$h_i(t) = \lambda_i h_0(t), i = 1, ..., n, \tag{23}$$

missä

$$\lambda_i = \exp(\beta_1 \cdot R1_i + \beta_2 \cdot R2_i + \beta_3 \cdot R3_i + \beta_4 \cdot R4_i + \beta_5 \cdot R5_i + \beta_6 \cdot R7_i)$$

ja $h_0(t)$ on referenssiluokkaan liittyvä vaarafunktio (perusvaarafunktio). Tällöin $\exp(\beta_j)$ kertoo, kuinka moninkertaiseksi valmistumisen vaara muuttuu suhteessa ryhmään R_6, kun opiskelija kuuluu regressiokerrointa vastaavaan ryhmään R_j.

Mallin estimoidut β-kertoimet ryhmille on esitelty taulukossa 14. Tulostaulukossa ryhmät on järjestetty valmistumisprosentin mukaiseen järjestykseen. Tilastollisesti merkitseviä kertoimia ovat ainoastaan seitsemänteen arvosanaryhmään R7 liittyvä kerroin sekä toiseen arvosanaryhmään R2 liittyvä kerroin, sillä näiden ryhmien kertoimien 95 prosentin luottamusväli ei sisällä arvoa yksi. Molempien ryhmien tapauksessa valmistumisen vaara on pienempi kuin referenssiluokassa, ryhmässä 6.

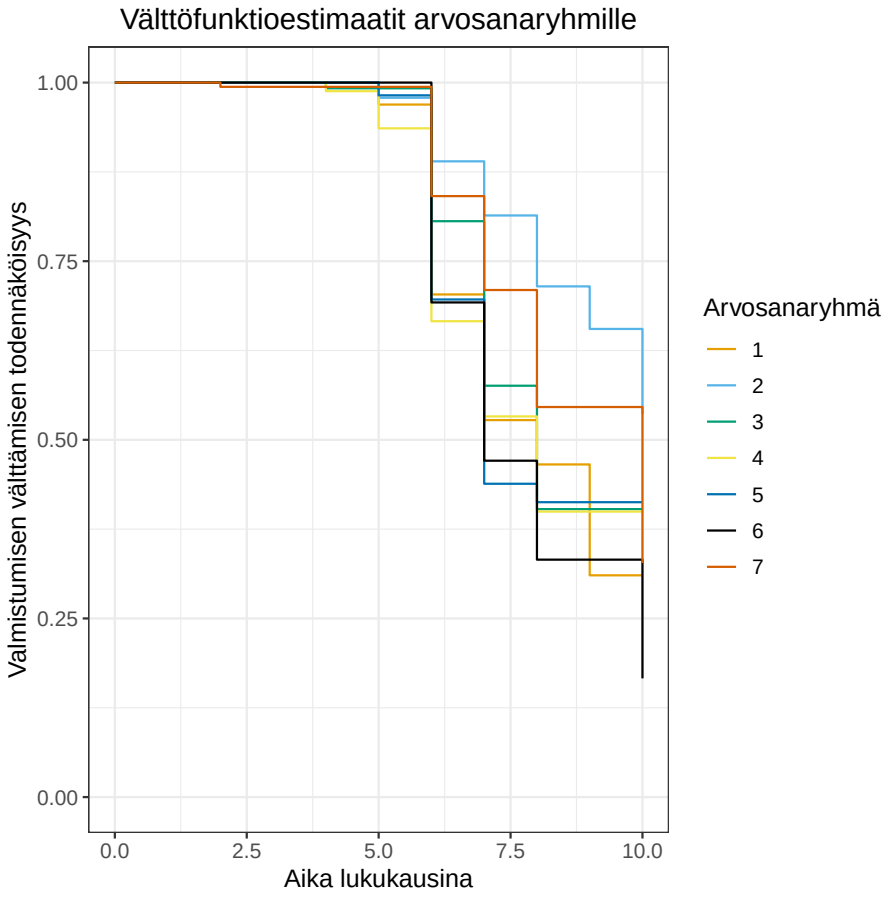
Ryhmän 4 valmistumisen vaara on lähes yhtäläinen referenssiryhmään 6 verrattuna. Ryhmien 5 ja 1 valmistumisen vaarat ovat keskimäärin 0.83-kertaiset referenssiryhmään 6 verrattuna. Edellä mainittujen ryhmien kertoimet eivät kuitenkaan ole tilastollisesti merkitseviä. Ryhmän 3 valmistumisen vaara on 0.70-kertainen ryhmään 6 verrattuna.

Taulukko 14: Ryhmien kertoimet sekä niitä vastaavat luottamusvälit, testisuureet ja p-arvot verrattaessa kuudenteen ryhmään.

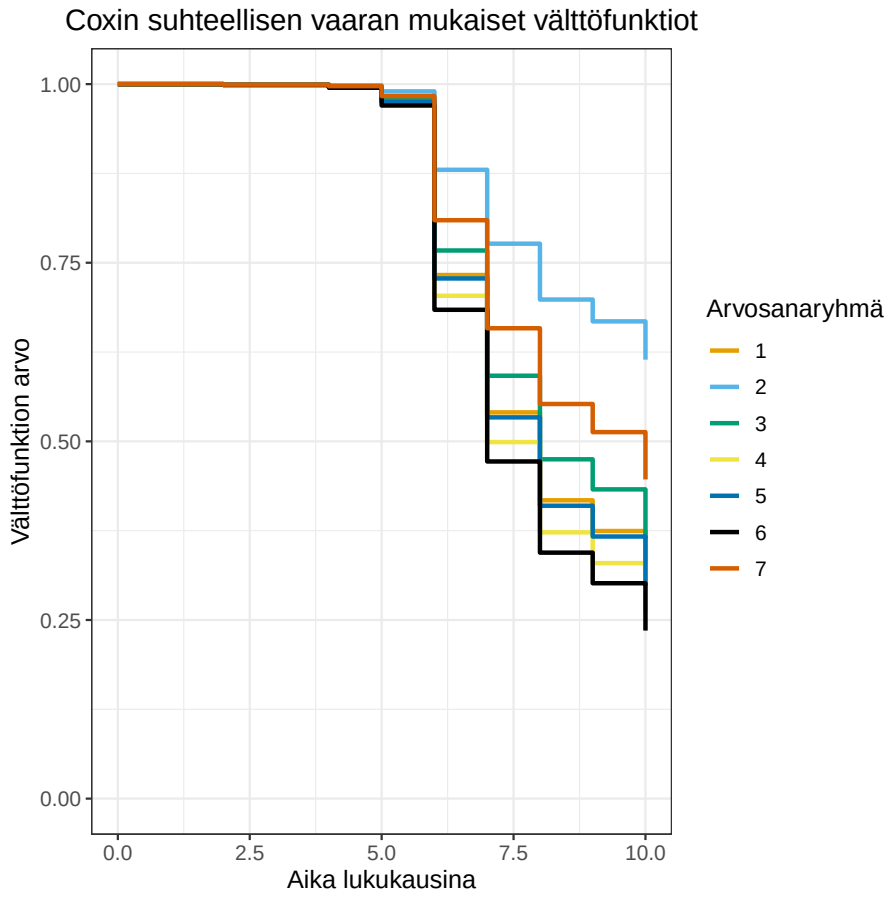
	exp(kerroin)	alaraja	yläraja	z-testisuure	p-arvo
R5	0.84	0.50	1.40	-0.68	0.50
R4	0.93	0.55	1.56	-0.29	0.77
R1	0.82	0.49	1.37	-0.77	0.44
R3	0.70	0.41	1.18	-1.35	0.18
R7	0.56	0.33	0.95	-2.14	0.03
R2	0.34	0.19	0.58	-3.87	0.00

Välttöfunktioiden Kaplan-Meier-estimaatit (12) eri arvosanaryhmille on esitetty kuviossa 5. Coxin mallin mukaiset välttöfunktiot on esitetty puolestaan kuviossa 6. Kuviosta 5 nähdään, että ryhmän 6 opiskelijat valmistuvat nopeimmin. Ryhmän 2 opiskelijat valmistuvat puolestaan hitaimmin. Kymmenennen lukukauden kohdalla ryhmän 6 valmistumisen välttämisen todennäköisyys on noin 15 % ja ryhmän 2 noin 50 %. Ryhmissä tapahtuu eroja valmistumisen suhteen jo ennen tavoiteaikaa eli kuudennetta lukukautta. Esimerkiksi ryhmissä 7 ja 4 on havaittavissa muutamia valmistumisia ennen tavoiteajan täyttymistä.

Coxin mallin mukaisten välttöfunktioiden kuviosta 6 nähdään mallin estimoimat välttöfunktion arvot kunkin lukukauden kohdalla. Mallin estimoimissa välttökäyrissä ryhmän 2 ero muihin arvosanaryhmiin korostuu. Kymmenennen lukukauden kohdalla Coxin mallin mukaiset välttöfunktion arvot ovat havaittuja välttötodennäköisyyksiä hieman korkeammat. Esimerkiksi ryhmän 6 välttöfunktion arvo kymmenennen lukukauden kohdalla on noin 25 % ja ryhmän 2 noin 63 %.



Kuvio 5: Välttöfunktioiden Kaplan-Meier-estimaatit eri ryhmille.



Kuvio 6: Coxin suhteellisen vaaran mallin mukaiset välttöfunktiot eri ryhmille.

7.3 Valmistumisajan selittäminen arvosanaryhmillä ja muilla muuttujilla

Ryhmämallia pyrittiin seuraavaksi parantamaan siten, että siihen lisätään tilastollisesti merkitsevästi mallia parantavia selittäjiä. Mallinvalintaa toteutettiin sisäkkäisten mallien vertailulla (ks. luku 6). Sisäkkäinen vertailu edellyttää, että mallit on sovitettu samaan aineistoon. Näin ollen esimerkiksi valintakoepisteitä ja hakutoivejärjestystä ei voitu huomioida mallinvalinnassa, sillä osalta opiskelijoista nämä tiedot puuttuivat. Lisäksi mallia pyrittiin parantamaan nimen omaan hakuvaiheen muuttujilla, joten opintovaiheen muuttujia ei huomioitu mallinvalinnassa. Laitostiedon osalta tulkittiin, että muuttuja on sekä haku- että opintovaiheen muuttuja. Hakukohde puolestaan tulkittiin hakuvaiheen muuttujana. Hakukohdetta ei kuitenkaan lisätty malliin, sillä hakukohde ja muut selittäjät, erityisesti arvosanaryhmä, voivat olla voimakkaasti riippuvaisia keskenään.

Mallinvalintaa toteutettiin lisäämällä ryhmämalliin suorat selittäjät kuten tieto valintajonosta sekä avoimessa yliopistossa suoritetuista opintopisteistä. Suorien selittäjien lisäksi malliin lisättiin uusien selittäjien ja arvosanaryhmien väliset interaktiotermit, joiden avulla voidaan tarkastella muuttujien yhteisvaikutusta valmistumisaikaan. Mallinvalinnan yhteydessä huomattiin, että ryhmämallia paransivat tieto valintajonosta sekä valintajonon ja arvosanaryhmien väliset interaktiotermit ($\chi^2 = 62.81$, df $= 27$, p < 0.001) tai tieto avoimessa yliopistossa suoritetuista opintopisteistä ja muuttujan interaktiotermit arvosanaryhmien kanssa ($\chi^2 = 18.74$, df $= 7$, p $= 0.009$). Mallin sovittamisen yhteydessä kuitenkin huo-

mattiin, että osassa interaktiotermejä kertoimien luottamusvälit olivat erittäin suuria tai kertoimet eivät estimoituneet ollenkaan. Ongelman pääteltiin johtuvan siitä, ettei tietyissä arvosanaryhmissä ole tarpeeksi eri valintajonojen kautta tulleita opiskelijoita tai tarpeeksi avoimessa yliopistossa opintopisteitä suorittaneita opiskelijoita. Estimointiongelma toistui siis molempien edellä mainittujen muuttujien yhteydessä. Näin ollen mallinvalintaa toteutettiin uudelleen ilman interaktiotermejä.

Kun mallinvalintaa jatkettiin lisäämällä ryhmämalliin pelkästään suorat selittäjät, merkitsevinä selittäjinä mallissa säilyivät tieto valintajonosta ($\chi^2 = 29.34$, df $= 4$, p < 0.001). Tieto avoimessa yliopistossa suoritetuista opintopisteistä ei puolestaan parantanut mallia enää tilastollisesti merkitsevästi ($\chi^2 = 1.34$, df $= 1$, p $= 0.25$). Mallien sisäisessä vertailussa havaittiin, että tieto avoimessa yliopistossa suoritetuista opintopisteistä ei parantanut mallia myöskään valintajonotiedon lisäämisen jälkeen ($\chi^2 = 0.25$, df $= 1$, p $= 0.62$). Lopuksi testattiin interaktiotermien puuttumisen vaikutus malliin. Interaktiotermien lisääminen ryhmä- ja valintajonomalliin ei parantanut mallia enää voimakkaasti tilastollisesti merkitsevästi ($\chi^2 = 33.47$, df $= 23$, p $= 0.07$). Näin ollen lopulliseksi malliksi saatiin malli, jossa on ryhmätiedon lisäksi mukana tieto valintajonosta. Uuden parannetun mallin avulla voidaan tarkastella valintajonojen välisiä tasoeroja. Mallin avulla ei voida kuitenkaan tarkastella valintajonojen ja arvosanaryhmien yhdysvaikutusta valmistumiseen, sillä interaktiotermin puuttumisen vuoksi malli ei salli erilaisia kertoimia eri valintajonon opiskelijalle saman arvosanaryhmän sisällä. Yhdysvaikutuksen tarkastelu on siten yksi mahdollinen jatkotutkimuksen kohde, kun aineistoa kertyy lisää.

Mallinvalinnan seurauksena muodostettu Coxin suhteellisen vaaran malli on muotoa

$$h_i(t) = \lambda_i h_0(t), i = 1, ..., n, \tag{24}$$

missä

$$\begin{aligned} \lambda_i = \quad & \exp(\beta_1 \cdot R5_i + \beta_2 \cdot R4_i + \beta_3 \cdot R1_i + \beta_4 \cdot R3_i + \beta_5 \cdot R7_i + \beta_6 \cdot R2_i + \\ & \beta_7 \cdot valintakoejono_i + \beta_8 \cdot yhteishaun\_ulkopuolelta_i + \\ & \beta_9 \cdot yhteispistejono_i + \beta_{10} \cdot yopistejono_i) \end{aligned} \tag{25}$$

ja $h_0(t)$ on referenssiluokkaan liittyvä vaarafunktio (perusvaarafunktio). Sekä ryhmämuuttujat R_j että valintajonomuuttujat ovat dikotomisia muuttujia, jotka saavat arvon yksi, mikäli opiskelija kuuluu kyseiseen ryhmään tai valintajonoon ja nolla, mikäli opiskelija ei kuulu kyseiseen ryhmään tai valintajonoon. Mallissa referenssiluokka on arvosanaryhmän 6 opiskelijat, jotka on valittu opiskelemaan suoravalintajonon kautta.

Taulukosta 15 huomataan, että ryhmässä 6 valmistumisen vaara on suoravalintajonoa korkeampi valintakoejonossa, yhteispistejonossa, ylioppilaspistejonossa sekä yhteishaun ulkopuolelta tulleilla opiskelijoilla. Tämä tarkoittaa sitä, että suhteessa suoravalintajonoon muiden valintajonojen kautta valitut opiskelijat valmistuvat nopeammin. Näistä tilastollisesti merkitseviä ovat valintakoejonon kerroin (luottamusväli (1.16, 4.88), p-arvo 0.02), yhteishaun ulkopuolelta tulleiden opiskelijoiden kerroin (luottamusväli (1.64, 9.34), p-arvo < 0.001) ja yhteispistejonon kerroin (luottamusväli (1.33, 5.88), p-arvo 0.01).

Ryhmävertailussa ryhmän 5 suoravalittujen opiskelijoiden valmistumisen vaara on 1.19-kertainen verrattuna ryhmän 6 suoravalittuihin opiskelijoihin. Ainoa tilastollisesti merkitsevä ryhmäkerroin on ryhmän 2 kerroin (luottamusväli (0.25, 0.79), p-arvo 0.01). Ryhmän 2 suoravalittujen valmistumisen vaara on 0.45-kertainen verrattuna ryhmän 6 suoravalittuihin opiskelijoihin. Ryhmämallin 23 tarkastelussa huomattiin, että ryhmien 7 ja 2 opiskelijoiden valmistumisen vaara erosi tilastollisesti merkitsevästi ryhmän 6 opiskelijoista.

Taulukko 15: Mallin kertoimet, niitä vastaavat luottamusvälit, testisuureet ja p-arvot verrattaessa referenssiluokkaan.

	exp(kerroin)	alaraja	yläraja	z-testisuure	p-arvo
R5	1.19	0.69	2.04	0.62	0.53
R4	0.99	0.59	1.68	-0.02	0.98
R1	0.82	0.49	1.38	-0.74	0.46
R3	0.86	0.51	1.46	-0.57	0.57
R7	0.79	0.45	1.40	-0.80	0.42
R2	0.45	0.25	0.79	-2.75	0.01
valintakoejono	2.38	1.16	4.88	2.36	0.02
yhteishaun ulkopuolelta	3.92	1.64	9.34	3.08	0.00
yhteispistejono	2.80	1.33	5.88	2.71	0.01
ylioppilaspistejono	1.16	0.60	2.25	0.43	0.66

Ryhmä- ja valintajonomallissa 24 puolestaan ryhmien välisessä vertailussa tilastollisesti merkitsevä ero on enää ryhmän 2 ja referenssiryhmän 6 opiskelijoiden välillä. Tämä tarkoittaa sitä, että tarkasteltaessa suoravalittuja opiskelijoita ryhmien 7 ja 6 valmistumisen vaarat eivät eroa enää tilastollisesti merkitsevästi. Mallin tulkinnoissa on huomioitava, että valintajonojen kertoimet kertovat vain valintajonon keskimääräisen vaikutuksen valmistumisen vaaraan.

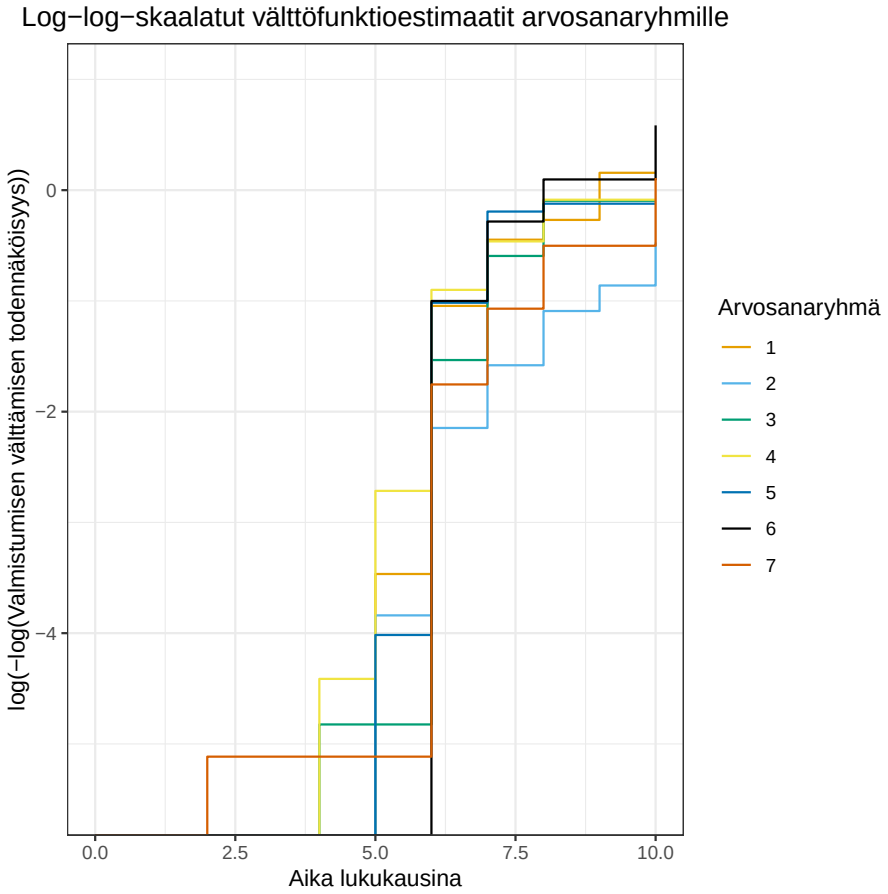
Mallista voidaan lisäksi tulkita, miten valmistumisen vaara muuttuu siirryttäessä ryhmästä ja valintajonosta toiseen. Esimerkiksi opiskelijan, joka on arvosanaryhmässä 5 ja hänet on valittu opiskelemaan yhteishaun ulkopuolelta, valmistumisen vaara verrattuna referenssiluokkaan on

$$\exp(\beta_{R5}+\beta_{yhteishaun\_ulkopuolelta}) = \exp(\beta_{R5})\cdot\exp(\beta_{yhteishaun\_ulkopuolelta}) = 1.19\cdot 3.92 = 4.67.$$

7.4 Elinaikamallien diagnostiset tarkastelut

Tässä alaluvussa käydään läpi tutkielmassa esiteltyjen Coxin suhteellisen vaaran mallien diagnostiset tarkastelut. Mallit, niiden yhtälöt ja tulokset on esitelty tarkemmin luvussa 7.2. Tässä luvussa mallien hyvyyttä tarkastellaan ensisijaisesti skaalattujen Schoenfeldin jäännösten, devianssijäännösten, martingaalijäännösten ja Cox-Snellin jäännösten avulla. Jäännöstarkastelut toteutettiin `survminer`-paketin funktiolla `ggcoxdiagnostics` (Kassambara ym. ,2020). Diagnostiikkaan liittyvää teoriaa on tarkasteltu tämän tutkielman luvun 5 alaluvussa 5.3.

Tarkastellaan aluksi suhteellisen vaaran oletuksen toteutumista. Testattava nollahypoteesi on, että vaarojen suhde pysyy vakiona ajasta riippumatta. Vaarojen suhteellisuutta tarkastellaan log-log-skaalattujen välttöfunktioiden avulla. Kuviossa 7 on esitelty opiskelijoiden välttöfunktiot arvosanaryhmittäin log-log-skaalassa. Huomioitavaa on, että ennen tavoiteajan täyttymistä log-log-skaalatuissa välttöfunktion arvoissa funktio estimoituu negatiiviseksi äärettömäksi. Näin ollen on tulkinnallisempaa katsoa välttöfunktioiden käyttäytymistä tavoiteajan, kuuden lukukauden, täyttymisen jälkeen. Log-log-skaalatuista välttöfunktiosta kaikki käyrät eivät ole samansuuntaisia. Erityisesti ennen tavoiteaikaa, kuudennetta lukukautta, välttökäyrät eivät näytä olevan suhteellisia. Kuitenkin kuudennesta lukukaudesta eteenpäin suhteellisen vaaran oletus näyttäisi olevan ainakin osittain voimassa. Esimerkiksi arvosanaryhmien 1 ja 2 sekä 6 ja 7 osalta vaarojen suhteellisuus toteutuu

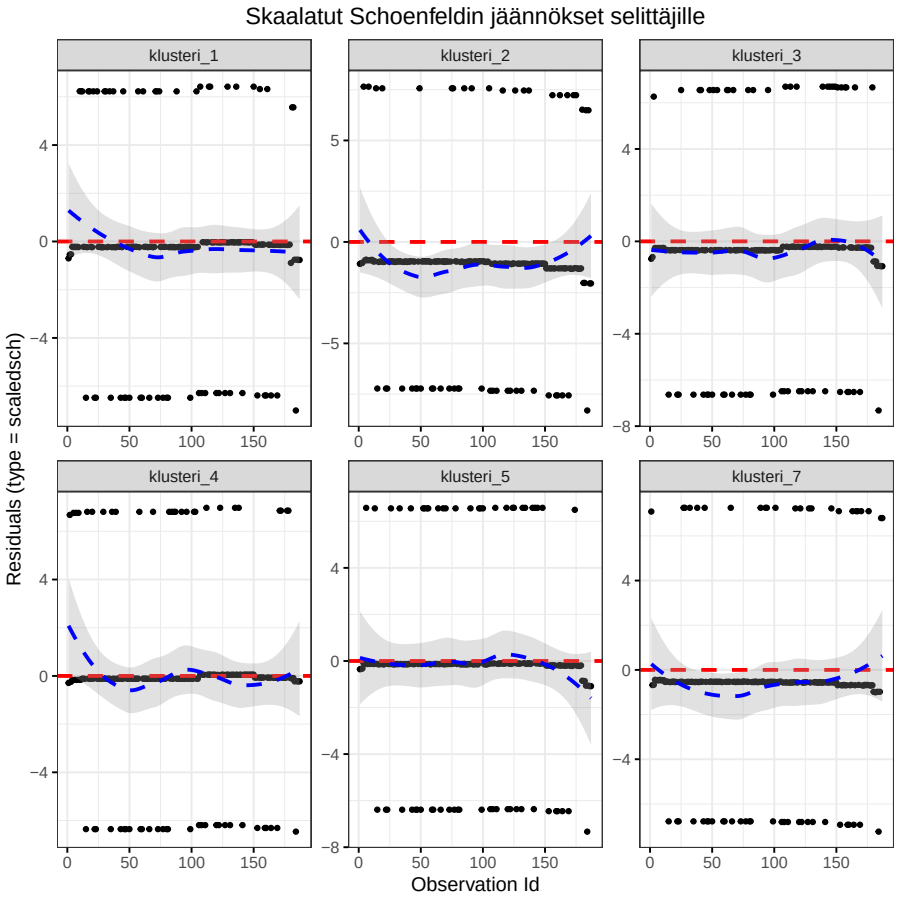


Kuvio 7: Log-log-välttöfunktiot eri arvosanaryhmille.

hyvin tavoiteajan jälkeen. Tarkasteltavia ryhmiä on kuitenkin suhteellisen paljon, minkä vuoksi suhteellisen vaaran oletusta ei voida täysin aukottomasti tulkita. Kuviossa on kuitenkin näyttöä sille, että nollahypoteesi jää voimaan.

Ryhmämallissa matemaattis-luonnontieteellisen tiedekunnan opiskelijat ensin ryhmiteltiin seitsemään ryhmään, minkä jälkeen muodostetuilla ryhmillä ennustettiin valmistumisen vaaraa. Arvosanaryhmämalli on esitelty tarkemmin alaluvussa 7.2. Devianssijäännöksistä huomataan, että ne ovat nollakeskisiä, kuten pitääkin.

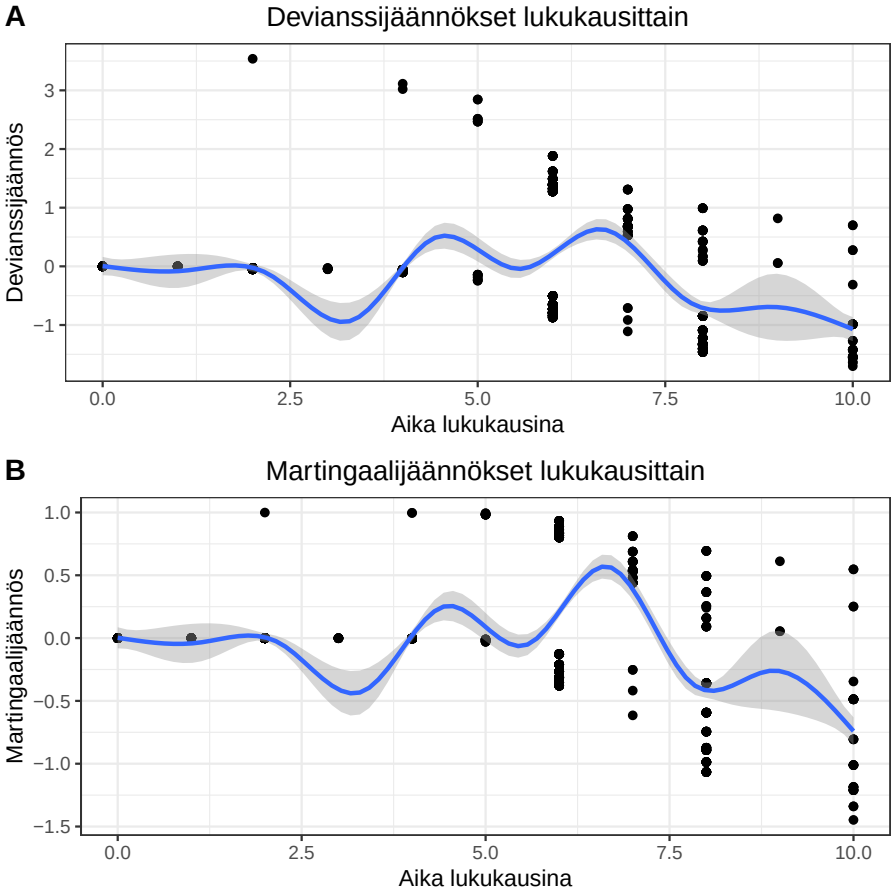
Skaalattujen Schoenfeldin jäännökset ryhmien osalta on esitelty kuviossa 8. Tarkastelussa huomataan, että kaikkien ryhmien osalta jäännökset ovat nollan molemmin puolin. Suhteellisen vaaran oletuksen ollessa voimassa jäännökset ovat odotusarvoisesti nollaa. Schoenfeldin jäännöstarkastelussa tämä oletus näyttää pitävän paikkansa.



Kuvio 8: Skaalatut Schoenfeldin jäännökset arvosanaryhmittäin.

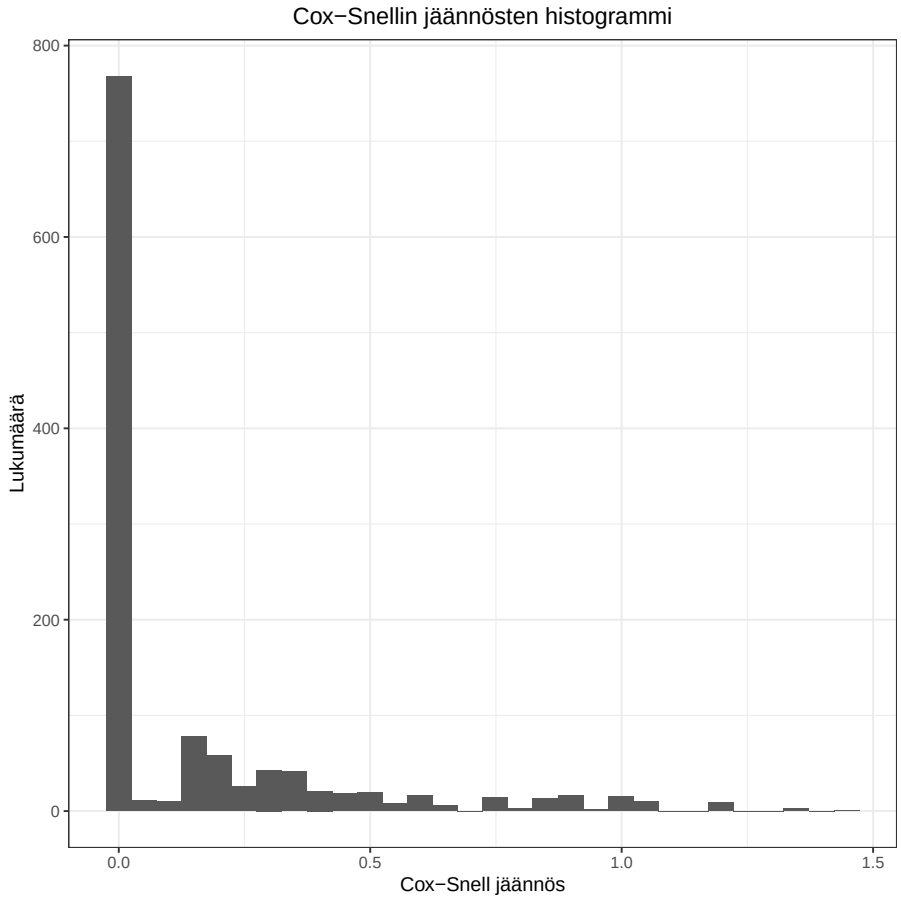
Devianssijäännökset ja martingaalijäännökset on esitelty kuviossa 9. Jäännöstarkastelussa on tärkeä huomioida, että valmistumisen tavoiteaika on kuusi lukukautta. Huomataan, että tavoiteaika vaikuttaa tarkasteltaviin jäännöksiin. Devianssijäännökset eivät ole nollakeskeisiä ennen tavoiteajan, kuuden lukukauden, täyttymistä. Sen sijaan kuudennen lukukauden jälkeen jäännökset ovat suhteellisen nollakeskeisiä. Pistekuvioon sovitetusta LOESS-käyrästä on havaittavissa, että lukukausien määrän kasvaessa jäännökset keskittyvät hieman nollan alapuolelle.

Martingaalijäännökset on esitelty kuviossa 9 (B). Martingaalijäännösten osalta tarkastellaan erityisesti nollan suhteen symmetrisyyttä. Symmetrisyys ei toteudu ennen tavoiteajan, kuuden lukukauden täyttymistä. Kun valmistumisaika ylittää tavoiteajan, jäännökset asettuvat suhteellisen hyvin nollan molemmin puolin.



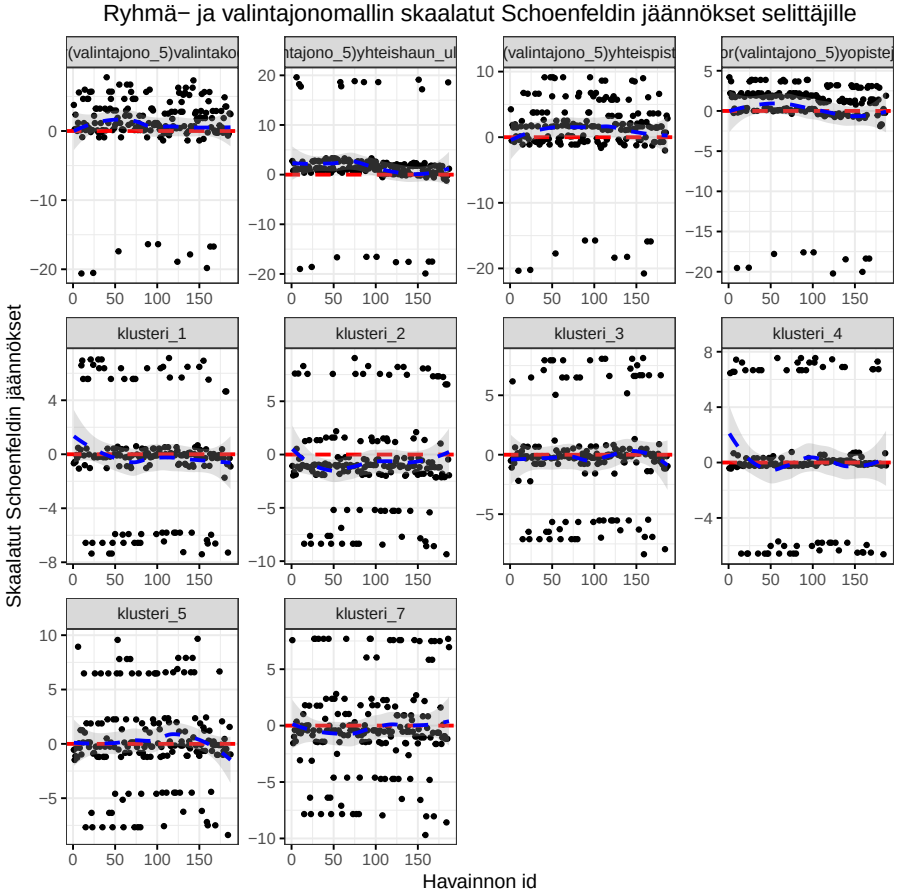
Kuvio 9: Devianssijäännökset (A) ja Martingaalijäännökset (B) lukukausittain. Kuvioihin on sovitettu LOESS-käyrä.

Tarkasteltaessa Cox-Snellin jäännöksiä oletukset eivät täysin toteudu. Nollahypoteesina on, että Cox-Snellin jäännökset noudattavat eksponenttijakaumaa. Kuviossa 10 on tarkasteltu Cox-Snellin jäännösten jakaumaa, joka näyttäisi silmämääräisesti olevan eksponenttijakauman kaltainen. Kuitenkin testattaessa jakaumaoletusta Kolmogorov-Smirnovin testin testisuureeksi saadaan $D = 0.59$ ja p-arvo < 0.001.



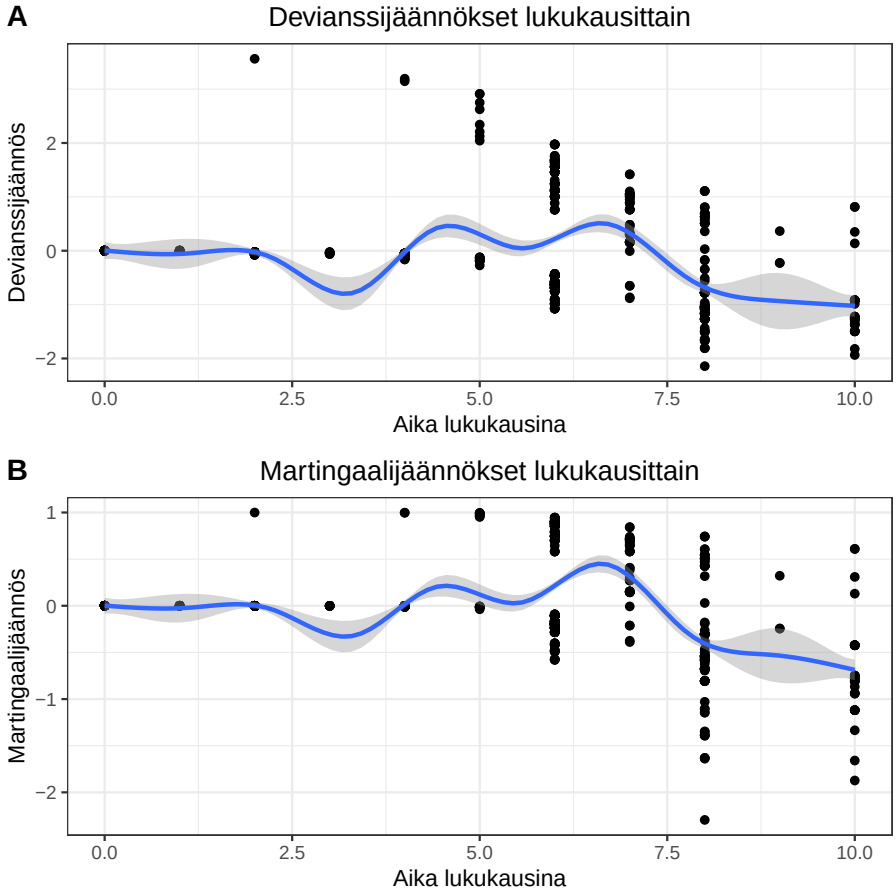
Kuvio 10: Ryhmämallin Cox-Snellin jäännösten jakauma.

Tarkastellaan vielä diagnostisesti mallia, jossa on ylioppilasarvosanoihin perustuvan ryhmätiedon lisäksi mukana myös valintajonosta kertovat muuttujat. Malli on esitelty tarkemmin alaluvussa 7.3. Jäännösten osalta tarkastellaan skaalattuja Schoenfeldin jäännöksiä (kuvio 11) sekä devianssi- ja marginaalijäännöksiä (kuvio 12). Schoenfeldin jäännösten kuviossa huomataan, että jäännökset ovat selkeämmin nollakeskeisiä kuin ryhmämallin tapauksessa.



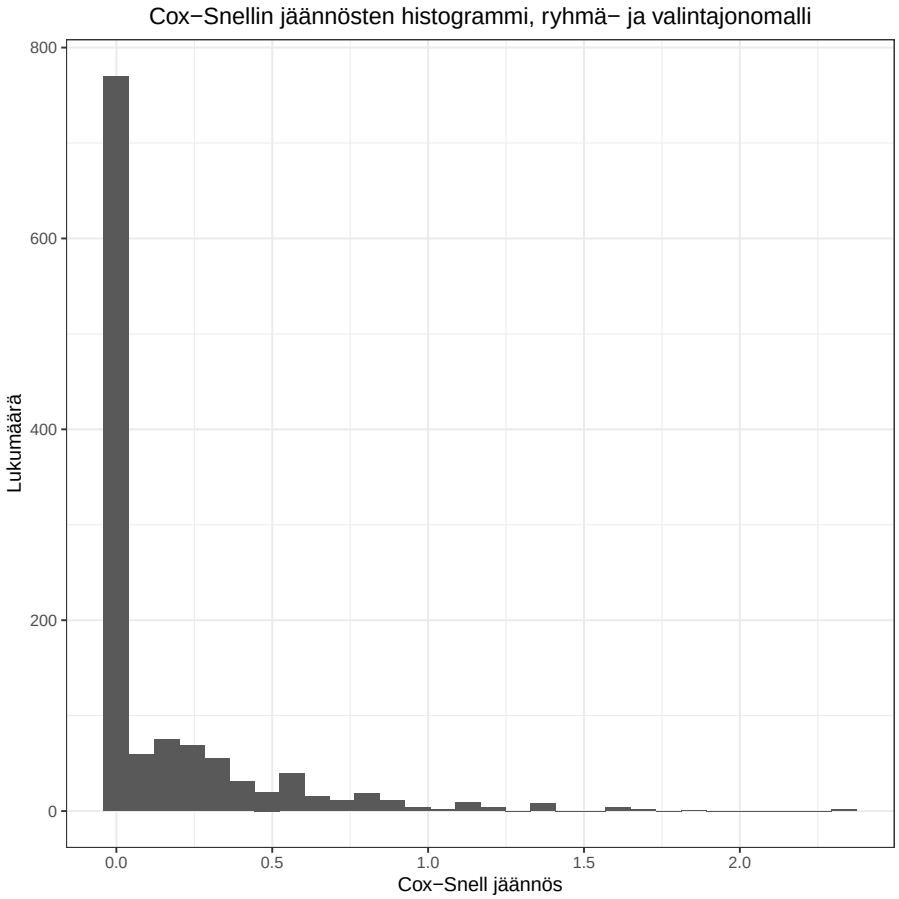
Kuvio 11: Skaalatut Schoenfeldin jäännökset lukukausittain arvosanaryhmittäin ja valintajonoittain.

Devianssi- ja martingaalijäännöksissä (kuvio 12) on havaittavissa samankaltaisia ilmiöitä kuin ryhmämallinkin tapauksessa. Jäännösten jakautumista nollan suhteen on mielekkäämpää tarkastella tavoiteajan, kuuden lukukauden, saavuttamisen jälkeen. Kuudennen lukukauden jälkeen sekä devianssi- että martingaalijäännökset ovat suhteellisen nollakeskeisiä. Lukukausien määrän kasvaessa sovitettu LOESS-käyrä osoittaa, että sekä devianssijäännökset (kuvio 12 (A)) ja martingaalijäännökset (kuvio 12 (B)) laskevat keskimääräisesti hieman nollan alapuolelle.



Kuvio 12: Ryhmä- ja valintajonomallin devianssijäännökset (A) ja martingaalijäännökset (B) lukukausittain. Kuvioihin sovitettu LOESS-käyrä.

Cox-Snellin jäännösten jakaumaa on tarkasteltu kuviossa 13. Kuvion perusteella jäännökset näyttää noudattavan silmämääräisesti eksponenttijakaumaa. Kuitenkin myös tämän mallin tapauksessa Kolmogorov-Smirnovin testin perusteella jakaumaoletus ei näytä toteutuvan ($D = 0.57$, $p < 0.001$).



Kuvio 13: Ryhmä- ja valintajonomallin Cox-Snellin jäännösten jakauma.

8 Yhteenveto

Tässä tutkielmassa pyrittiin mallintamaan Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan opiskelijoiden valmistumisaikoja. Opiskelijoiden valmistumisaikoja kandidaatiksi mallinnettiin elinaikamallinnuksen ja ryhmittelyanalyysin keinoin. Elinaikamallinnuksessa käytettiin Coxin suhteellisen vaaran mallia. Aineiston ryhmittelyyn käytettiin k:n keskiarvon ryhmittelymenetelmää.

Opiskelijat ryhmiteltiin äidinkielen, matematiikan, biologian, kemian, fysiikan, keskipitkän ruotsin, pitkän englannin, psykologian, terveystiedon ja maantieteen ylioppilasarvosanojen perusteella yhteensä seitsemään ryhmään. Ryhmittely toteutettiin näillä muuttujilla kaikille tiedekunnan opiskelijoille, joilta arvosanatieto löytyi edes yhdestä ryhmitellyyn käytetystä oppiaineesta. Ryhmittelyssä oli mukana 96 % aineiston opiskelijoista.

Aineiston ryhmittelyssä havaittiin, että ryhmien välillä on eroja ylioppilasarvosanoissa ja valmistumistodennäköisyyksissä kuuden lukukauden jälkeen. Ryhmässä, jossa oli korkein valmistumisprosentti kuuden lukukauden jälkeen, oli opiskelijoita, jotka olivat menestyneet erityisesti äidinkielen, biologian, pitkän englannin ja maantieteen ylioppilaskirjoituksissa. Heikoin valmistumisprosentti oli ryhmässä, jossa oli korkeahko ylioppilasarvosanojen keskiarvo matematiikasta sekä matalat ylioppilasarvosanojen keskiarvot biologiasta, fysiikasta ja kemiasta sekä ryhmien välisessä vertailussa matalin äidinkielen keskiarvo. Toiseksi heikoin valmistumisprosentti oli ryhmässä, jossa oli korkeat keskiarvot matematiikan, fysiikan ja kemian ylioppilaskirjoituksissa.

Arvosanojen lisäksi tarkasteltiin ryhmien muita ominaisuuksia kuten laitostietoa ja hakutoivejärjestystä. Korkean valmistumisprosentin ryhmässä oli enimmäkseen bio- ja ympäristötieteiden laitoksen opiskelijoita. Matalien valmistumisprosenttien ryhmissä oli enimmäkseen muiden kuin bio- ja ympäristötieteen laitoksen opiskelijoita. Matalimman valmistumisprosentin ryhmässä oli eniten fysiikan laitoksen sekä matematiikan ja tilastotieteen laitoksen opiskelijoita.

Ryhmittelyn tuloksena saatuja ryhmiä käytettiin niin ikään elinaikamallinnuksessa selittäjinä. Huomattiin, että ryhmien välillä oli myös tilastollisesti merkitseviä eroja valmistumisessa, tarkalleen ottaen valmistumisen vaarassa. Mallinnuksessa ryhmien valmistumisen vaaraa verrattiin niin kutsutun referenssiryhmän vaaraan. Referenssiryhmänä käytettiin ryhmää, jossa todettiin aiemmin korkein valmistumisprosentti. Tilastollisesti merkitsevä ero valmistumisen vaarassa suhteessa referenssiryhmään oli kahdella matalimman valmistumisprosentin ryhmällä. Välttöfunktion tarkastelussa huomattiin, että elinaikamallin tulokset olivat samankaltaisia valmistumisprosenttien kanssa. Esimerkiksi matalimman valmistumisprosentin ryhmässä myös valmistumisen välttäminen oli kaikista suurinta.

Tutkimuksessa huomattiin, että nopeimmin Jyväskylän yliopiston matemaattis-luonnontieteellisessä tiedekunnassa valmistuu sellaisen arvosanaryhmän opiskelija, jolla on korkeat keskiarvosanat äidinkielestä, biologiasta ja maantiedosta. Tuloksissa on nähtävissä tiedekunnan eri alojen suosio ja hakupaine. Ryhmittelyn tuottamien ryhmien tarkastelussa huomattiin, että kyseisen korkean valmistumisprosentin opiskelijat ovat lähes kaikki bio- ja ympäristötieteiden laitoksen opiskelijoita. Laitoksen alat ovat yhteishaussa suosittuja hakukohteita, joten valituiksi tulevat lähtökohtaisesti ne opiskelijat, joilla on korkeat arvosanat. Lisäksi motivaatio näillä aloilla voi olla korkeampi vaikeamman sisäänpääsyn vuoksi.

Heikointa valmistuminen oli ryhmässä, jossa oli korkea matematiikan arvosanojen keskiarvo ja ryhmien välisessä vertailussa matalin äidinkielen keskiarvo. Ryhmässä oli paljon

ylioppilaspistejonosta tulleita opiskelijoita. Suurin osa ryhmän opiskelijoista oli fysiikan laitoksen sekä matematiikan ja tilastotieteen laitoksen opiskelijoita. Tässäkin tapauksessa matalaa valmistumisprosenttia ja valmistumisen suhteellista vaaraa voi selittää moni muukin tekijä kuin ylioppilasarvosanat.

Ryhmämalliin pyrittiin lisäämään myös muita selittäjiä. Tilastollisesti merkitsevästi mallia paransi tieto valintajonosta. Valintajonojen osalta referenssiryhmänä käytettiin suoravalintajonoa. Kaikissa muissa valintajonoissa valmistumisen vaara oli suurempi kuin suoravalintajonossa. Suurin suhteellinen valmistumisen vaara oli yhteishaun ulkopuolelta tulleilla opiskelijoilla. Parannettuun malliin lisättiin valintajonotiedot ainoastaan suorina selittäjinä, sillä osa arvosanaryhmien ja valintajonojen välisistä interaktiotermeistä ei estimoitunut. Interaktiotermien puuttumisen vuoksi mallin avulla ei voida tarkastella arvosanaryhmän ja valintajonon yhdysvaikutusta valmistumiseen. Mallin avulla voidaan kuitenkin tarkastella valintajonojen välisiä tasoeroja valmistumisen vaarassa.

Ennen ryhmittelyanalyysin toteuttamista aineistoon sovitettiin niin kutsuttu klassinen malli, jossa opiskelijan valmistumisaikaa selitettiin valintakoepisteiden, avoimessa yliopistossa suoritettujen opintopisteiden sekä ylioppilasarvosanojen avulla. Automatisoidun mallinvalinnan jälkeen klassisen mallin selittäjiksi ylioppilasaineista jäivät ainoastaan arvosanat biologiasta, englannin pitkästä oppimäärästä ja äidinkielestä. Klassinen malli oli yksinkertainen ja helppo lähestymistapa monimutkaiseen ongelmaan. Sen tulokset olivat kuitenkin liian tasapäistetyt: esimerkiksi englannin arvosanan vaikutus valmistumisen vaaraan oli alle yhden, mikä tarkoittaisi sitä, että korkea arvosana englannin ylioppilaskirjoituksissa hidastaisi opiskelijan valmistumista. Malli ei ottanut huomioon ylioppilasaineiden kokonaisuuksia ja painottui aineistossa paljon kirjoitettuihin ylioppilasaineisiin. Näin ollen päätettiin, että ylioppilasarvosanojen tietoja tiivistetään ja samalla tuodaan ilmi arvosanoihin liittyviä yhteyksiä.

Aineiston ylioppilasarvosanatietoja pyrittiin tiivistämään usealla eri menetelmällä. Ylioppilasarvosana-aineistoon sovellettiin muun muassa ordinaatiomenetelmiä ja ryhmittelymenetelmiä. Ordinaatiomenetelmistä aineistoon sovellettiin pääkoordinaattianalyysiä, korrespondenssianalyysiä sekä ei-metristä moniulotteista skaalausta (ks. esim. Hill ja Gauch, 1980, Queen ym., 2002). Näitä menetelmiä käytetään yleensä biologian sovellusalan aineistoihin, joissa tavoitteena on tiivistää tutkimuspaikkojen tietoa lajimäärien perusteella muutamaan muuttujaan. Tässä tutkimuksessa ylioppilasarvosanat ajateltiin lajimäärinä, ja opiskelijat mittauspaikkoina. Menetelmillä onnistuttiin tiivistämään jonkin verran aineiston arvosanatietoja, mutta uusien muuttujien tulkinnallisuus osoittautui haasteelliseksi. Menetelmän sopivuuden ja tulkinnan kannalta ongelman ydin vaikuttaisi olevan ylioppilasaineiden valikoitumisessa ja nollien suuressa määrässä. Osa ylioppilasaineista on suosituimpia kuin toiset, ja erityisesti vähän kirjoitetut ylioppilasaineet korostuivat liian voimakkaasti esimerkiksi pääkoordinaattianalyysissä suhteessa kirjoittajien määrään.

Ryhmittelymenetelmistä aineistoon sovellettiin hierarkkista ryhmittelyä (ks. Rokach, 2005) ja k:n keskiarvon ryhmittelyä. Hierarkkinen ryhmittely ei kuitenkaan ryhmitellyt aineistoa tasaisesti, vaan ryhmittelyn tuloksena muodostui yksi suurempi ryhmä ja muutama pienempi ryhmä. Lisäksi ryhmien tarkastelussa kävi ilmi, ettei hierarkkinen ryhmittely ollut onnistunut muodostamaan toisistaan poikkeavia opiskelijaryhmiä. Tästä johtuen tutkimuksessa päädyttiin k:n keskiarvon ryhmittelyyn, sillä tämän ryhmittelymenetelmän tuloksena muodostuneissa ryhmissä oli toisistaan poikkeavia ominaisuuksia.

Tiedekuntatason lisäksi ryhmittelyä pyrittiin toteuttamaan myös laitostasolla. Ryhmittely ei tuottanut kuitenkaan toisistaan merkittävästi poikkeavia ryhmiä. Lisäksi elinaikamallissa ryhmien valmistumisen vaaroissa ei ilmennyt enää tilastollisesti merkitseviä eroja.

Tutkielmassa toteutettujen Coxin mallien hyvyyttä ja oletusten toteutumista tarkasteltiin muun muassa jäännöstarkastelujen avulla. Jäännöstarkasteluissa yleinen ilmiö oli, että ennen tavoiteaikaa jäännökset eivät jakautuneet tasaisesti nollan ympärille. Kuitenkin tavoiteajan saavuttamisen jälkeen jäännöstarkastelut näyttivät tukevan nollahypoteesia ja sitä, että malli sopii aineistoon. Jäännöstarkasteluihin voi osaltaan vaikuttaa se, että valmistumisajalla on odotettu tavoiteaika. Tavoiteaika ja opintojen mitoitus osaltaan vaikuttavat valmistumisaikojen todennäköisyyksiin tietyillä ajanhetkillä. Käytännössä valmistumisia ei odoteta tapahtuvan tasaisesti joka ajanhetkellä. Esimerkiksi tavoiteajan, kuudennen lukukauden, kohdalla valmistumisen todennäköisyys on huomattavasti suurempi kuin ennen tavoiteaikaa. Jatkotutkimuksissa tämä valmistumisaikaan liittyvä erikoislaatuisuus voitaisiin ottaa paremmin mallinnuksessa ja menetelmäkehityksessä huomioon. Myöskään mallien ennustuskykyä ei tarkasteltu erikseen tässä tutkimuksessa. Mikäli malleja haluttaisiin käyttää jatkotutkimuksissa valmistumisen ennustamiseen, tulisi niiden ennustuskykyä tarkastella esimerkiksi ristiinvalidoinnin avulla.

Tutkielma antaa uutta näkökulmaa valmistumisajan tarkasteluun. Sen sijaan, että tarkasteltiin yksittäisiä ylioppilaskirjoitusten aineita, muodostettiin arvosanaryhmiä. Tämä antaa esimerkiksi yhteisvalinnassa saatuihin ylioppilaskoepisteisiin verrattuna yksityiskohtaisempaa tietoa opiskelijoiden ylioppilaskoemenestyksestä.

Jatkossa vastaavaa ryhmittelyä ja valmistumisaikojen analyysiä voitaisiin toteuttaa myös muissa tiedekunnissa. Tämän tutkielman aineistossa ja tuloksissa korostuu korkea luonnontieteiden osaaminen sekä bio- ja ympäristötieteiden laitoksen hakukohteiden omalaatuisuus. Laitostiedon ja hakukohteen laajempi huomiointi esimerkiksi ryhmittelyssä voisi tarkentaa tuloksia.

Korkeat arvosanat eivät näyttäisi takaavan sitä, että opiskelija valmistuisi tavoiteajassa. Toisaalta valintajonotiedon lisääminen malliin ja kyseisen mallin tulokset kertovat siitä, että yliopisto-opinnoissa voi menestyä ja valmistua monenlaisista lähtökohdista.

Viitteet

Hirotugu Akaike (1998). Information theory and an extension of the maximum likelihood principle. Kirjassa *Selected Papers of Hirotugu Akaike*, ss. 199–213. Springer.

David Collett (2015). *Modelling Survival Data in Medical Research*. CRC press.

David R. Cox (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

David R. Cox ja David Oakes (1984). *Analysis of Survival Data*. Chapman and Hall/CRC.

David R. Cox ja E. Joyce Snell (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–265.

Regina C. Elandt-Johnson ja Norman Lloyd Johnson (1980). *Survival Models and Data Analysis*. Wiley Online Library.

Iida Häkkinen (2004). Do university entrance exams predict academic achievement? Työpaperi 2004:16, Uppsala.

John A Hartigan ja Manchek A Wong (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108.

Mark O. Hill ja Hugh G. Gauch (1980). Detrended correspondence analysis: An improved ordination technique. Kirjassa *Classification and Ordination*, ss. 47–58. Springer.

Gareth James, Daniela Witten, Trevor Hastie, ja Robert Tibshirani (2013). *An Introduction to Statistical Learning*. Springer.

Eeva Kallio, Jukka Utriainen, Mikko Niilo-Rämä, ja Eija Räikkönen (2018). Lukiomenestyksen ja yliopisto-opintojen aloitushetken iän yhteys yliopisto-opinnoissa menestymiseen ja opintojen etenemiseen: Seurantatutkimus. *Kasvatus*, 49(4).

Edward L Kaplan ja Paul Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Alboukadel Kassambara, Marcin Kosinski, ja Przemyslaw Biecek (2020). *survminer: Drawing Survival Curves using 'ggplot2'*. R-paketin versio 0.4.8.

Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, ja Kurt Hornik (2019). *cluster: Cluster Analysis Basics and Extensions*. R-paketin versio 2.1.0.

Ville Mankki, Pekka Räihä, ja Jorma Joutsenlahti (2018). Todistusvalinnan ennustevaliditeetti korkeakoulujen opiskelijavalinnassa: esimerkkinä luokanopettajakoulutus.

Emma Mieskonen (2017). Opintomenestyksen ennustaminen ja opiskelijavalinnan vaikutuksen arviointi Aalto-yliopiston kauppakorkeakoulussa. Pro gradu, Aalto-yliopiston kauppakorkeakoulu.

Opetus- ja kulttuuriministeriö (2016). Valmiina valintoihin. Ylioppilastutkinnon parempi hyödyntäminen korkeakoulujen opiskelijavalinnoissa. Sarjajulkaisu 2016:37.

Opetus- ja kulttuuriministeriö (2018). Luovuutta, dynamiikkaa ja toimintamahdollisuuksia: ehdotus ammattikorkeakoulujen ja yliopistojen rahoitusmalleiksi vuodesta 2021 alkaen. Sarjajulkaisu 2018:35.

Opetus- ja kulttuuriministeriö (2019a). Valtioneuvoston asetus korkeakoulujen yhteishausta 289/2019. Annettu Helsingissä 14. maaliskuuta 2019. https://www.finlex.fi/fi/laki/alkup/2019/20190289.

Opetus- ja kulttuuriministeriö (2019b). Valtioneuvoston asetus ylioppilastutkinnosta 612/2019 6. Annettu Helsingissä 9. toukokuuta 2019. https://www.finlex.fi/fi/laki/alkup/2019/20190612.

Opetusministeriö (2009). Yliopistolaki 558/2009. Annettu Naantalissa 24. heinäkuuta 2009. https://www.finlex.fi/fi/laki/alkup/2009/20090558.

Opintopolku (2019). Mikä korkeakoulujen opiskelijavalinnoissa muuttuu vuoteen 2020 mennessä? Verkkosivu. https://opintopolku.fi/wp/opo/korkeakoulujen-haku/mika-korkeakoulujen-opiskelijavalinnoissa-muuttuu-vuoteen-2020-menessa/yliopistojen-todistusvalinnat-2020/. Viitattu 30.4.2021.

Joni Petman (2017). Yliopistotutkintojen määrän ennustaminen bayes-mallilla. Pro gradu, Jyväskylän yliopisto.

Jerry P. Queen, Gerry P. Quinn, ja Michael J. Keough (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien, Itävalta.

Lior Rokach ja Oded Maimon (2005). Clustering methods. Kirjassa *Data Mining and Knowledge Discovery Handbook*, ss. 321–352. Springer.

Peter J. Rousseeuw (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

David Schoenfeld (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153.

Judith D. Singer ja John B. Willett (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18(2):155–195.

Terry M. Therneau (2020). *A Package for Survival Analysis in R*. R-paketin versio 3.2-7.

Terry M. Therneau, Patricia M. Grambsch, ja Thomas R. Fleming (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160.

Pertti Väisänen ja Sakari Ylönen (2004). Matemaattiset taidot ja matemaattinen minäkäsitys tilastollisten menetelmien oppimisessa. *Kasvatus: Suomen kasvatustieteellinen aikakauskirja 35 (2004): 4.*

W. N. Venables ja B. D. Ripley (2002). *Modern Applied Statistics with S*. Springer, New York. Neljäs painos.

Lauri Viitanen (2016). Identifying at-risk students at Metropolia UAS: Estimating graduation probability with survival models and statistical classifiers. Diplomityö, Aalto-yliopisto.