JYVÄSKYLÄN YLIOPISTO
MATEMATIIKAN JA TILASTO-
TIETEEN LAITOS

Pro gradu-tutkielma

Oppivelvollisuuslain uudistuksen vaikutusten arvioiminen kausaalimallinnusta hyödyntäen

Jenni Väisänen

5. toukokuuta 2023

Tekijä
Jenni Väisänen

Otsikko
Oppivelvollisuuslain uudistuksen vaikutusten arvioiminen kausaalimallinnusta hyödyntäen

Tutkinto-ohjelma
Tilastotieteen maisteriohjema

Päivämäärä	**Sivumäärä**
5. toukokuuta 2023	44 + 11

Tiivistelmä

Tämän pro gradu -tutkielman tarkoituksena on selvittää, onko vuoden 2021 oppivelvollisuuslain uudistuksella ollut vaikutusta nuorten opintojen jatkamiseen heti yhdeksännen luokan päätyttyä. Tutkimusongelmaa lähestytään kausaalipäättelyn keinoin. Tarkastelu kohdistuu niihin 2021 peruskoulun päättäneisiin nuoriin, jotka kuuluvat uudistuneen oppivelvollisuuslain piiriin. Heistä kiinnostuksen kohteena ovat erityisesti ne nuoret, jotka historia-aineiston perusteella eivät todennäköisesti jatka opintoja heti 9. luokan päätyttyä.

Tässä tutkielmassa henkilön katsotaan jatkaneen opintoja, jos hän on ollut peruskoulun päättövuonna 20.9 kirjoilla lukiokoulutuksessa, ammatillisessa koulutuksessa, valmentavassa koulutuksessa tai lisäopetuksessa (10. luokka). Vuonna 2021 peruskoulun päättäneistä, oppivelvollisuuden piiriin kuuluvista nuorista, 1.7 % ei jatkanut opintoja heti yhdeksännen luokan jälkeen.

Tutkimusongelman tilanteessa kausaalimallinnuksessa tarvittavat vasteet ennustetaan logistisen regressioanalyysin avulla. Kun tarkastellaan kaikkia oppivelvollisuuden piiriin kuuluvia, peruskoulun vuonna 2021 päättäneitä 9. luokkalaisia, oppivelvollisuuslain uudistuksen keskimääräinen kausaalivaikutus koulutuksessa jatkamiseen on 0.6 prosenttiyksikköä. Eli opintojen jatkamisen todennäköisyys on noussut oppivelvollisuuslain uudistuksen myötä. Kun rajataan tarkastelu niihin henkilöihin, joiden todennäköisyys olla jatkamatta opintoja on historia-aineiston perusteella suurempi kuin 70 %, oppivelvollisuuslain uudistuksen keskimääräinen kausaalivaikutus on 22.1 prosenttiyksikköä.

Sisällys

1 Johdanto

Vuonna 1921 voimaan tullut oppivelvollisuuslaki täytti vastikään 100 vuotta. Pääministeri Marinin hallituskaudella tätä lakia on uudistettu ja laajennettu koskemaan toisen asteen koulutusta, samalla oppivelvollisuusikä nousi 18 ikävuoteen saakka. [Ahonen, 2021] Tutkielman tavoitteena on selvittää, onko oppivelvollisuuslain uudistuksella ollut vaikutusta niiden, vuonna 2021 peruskoulun päättäneiden, nuorten opintojen jatkamiseen, jotka kuuluvat laajentuneen oppivelvollisuuden piiriin. Kiinnostuksenkohteena ovat erityisesti ne 2021 peruskoulun päättäneet, oppivelvollisuuslain piiriin kuuluvat nuoret, jotka vuosien 2018–2020 peruskoulun päättäneiden aineiston perusteella eivät todennäköisimmin jatka opintoja heti 9. luokan päätyttyä.

Tässä tutkielmassa käsitellään luvussa 2 lyhyesti kouluttautumisen vaikutusta työllistymiseen. Nuorten kouluttautuminen sekä koulutusvalinnat ovat vahvasti kytköksissä taloudelliseen ja sosiaaliseen pääomaan, minkä vuoksi nuorten koulutusvalinnoilla on suuria vaikutuksia sekä lähivuosien, että heidän koko elämänsä osalta [Ruohola, 2012]. Lakiuudistuksen keskeisenä tavoitteena on, että kaikki peruskoulun päättäneet suorittavat toisen asteen tutkinnon [Oppivelvollisuuslaki, 1214/2020]. Luvussa 2 käydään läpi oppivelvollisuuslain taustaa, syitä lakiuudistukselle, lakiuudistuksen tavoitteita sekä kerrotaan keitä lakiuudistus konkreettisesti koskee.

Oppivelvollisuuslain uudistuksen esittelyn jälkeen luvussa 3 tarkastellaan tutkimuksessa käytettyä aineistoa. Tutkimusaineisto on koottu tätä työtä varten Opetushallituksen, koulutuksen järjestäjien ja Tilastokeskuksen ylläpitämistä rekisteriaineistoista. Aineistot ovat saatavilla Tilastokeskuksen tutkijapalveluiden FIONA-etäkäyttöjärjestelmän kautta erillisellä sopimuksella. Rekisteriaineistoista yhdistetyssä aineistossa on tieto niiden nuorten toisen asteen koulutukseen sijoittumisesta, jotka ovat päättäneet peruskoulun vuosina 2018–2021. Lopulliseen tutkimusaineistoon on tehty rajauksia vuoden 2021 oppivelvollisuuslain uudistuksen kohdejoukon perusteella.

Tutkimusaineiston henkilöt jakautuvat peruskoulun päättövuoden mukaan kahteen ryhmään oppivelvollisuuslain ja oppivelvollisuuden suhteen. Ryhmiin jakautumiseen ei ole voitu vaikuttaa tutkielmaa tehdessä, niinpä kyseessä on havainnoiva tutkimus, tarkemmin luonnollinen koe. Tämän tutkielman tilanteessa tapahtuma $do(X = 1)$, eli *oppivelvollisuuslain uudistus* koskee henkilöitä, jotka ovat päättäneet peruskoulun vuonna 2021 ja tapahtuma $do(X = 0)$, eli *oppivelvollisuuslaki ei uudistunut*, liittyy kontrafaktuaaliseen tilanteeseen. Tutkimusongelman tilanteessa on tarkoitus selvittää, miten intervention $do(X = 0)$ soveltaminen vuonna 2021 peruskoulun päättäneiden aineistoon vaikuttaa näiden nuorten opintojen jatkamiseen.

Tutkimusongelma on kausaalipäättelyn kontrafaktuaalitilanne. Kausaa-

lisuus eli syy-seuraussuhteet tarkoittavat tapahtumien välisiä yhteyksiä, joissa toinen tapahtuma (syy) aiheuttaa toisen tapahtuman (seuraus). Kausaalipäättelyllä pyritään tunnistamaan näitä tapahtumien välisiä syy-seuraussuhteita. [Pearl, 2009]. Tutkielmassa käytetyt menetelmät esitellään luvussa 4. Menetelmäosion alussa jäsennellään tutkimusongelman käsittely vaihe vaiheelta, minkä jälkeen edetään kausaalimallinnuksen teoriaan. Tässä tutkielmassa kausaalimallinnuksen teoriasta käsitellään graafiteoriaa, sekä määritellään d-separoituvuuden, takaovikriteerin ja takaovikorjauksen käsitteet. Kausaalipäättelyn teorian jälkeen esitellään logistinen regressioanalyysi, jolla tutkielman tilanteessa kausaalivaikutukset estimoidaan. Regressioanalyysin osalta keskitytään linkkifunktioihin, mallin muuttujien valintaan sekä mallin hyvyystarkasteluihin. Luvun 4 menetelmäosioiden pohjana on käytetty pääasiassa Pearlin (2009, 2016), Hollandin (1986, 1988) ja Hosmerin (2013) teoksia.

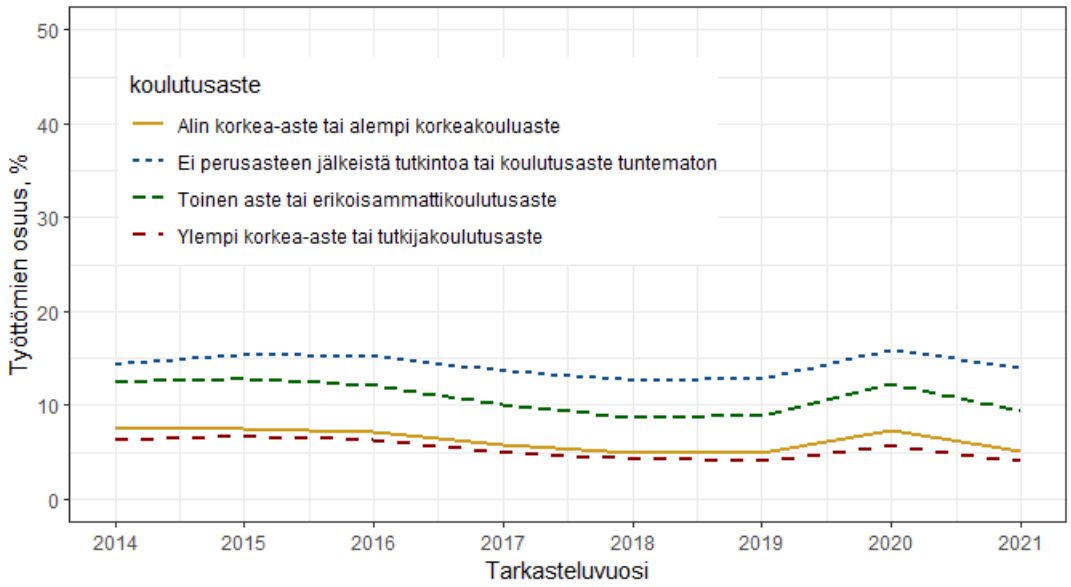
Menetelmäosion jälkeen luvussa 5 käsitellään tutkimustilanteen kausaalivaikutuksen estimointiin liittyvä mallin valinta. Tavoitteena on estimoida kausaalivaikutukset $E(Y \mid do(X = 1))$ ja $E(Y \mid do(X = 0))$ peruskoulun 2021 päättäneille. Kausaalivaikutus $E(Y \mid do(X = 1))$ estimoidaan logistisella regressiomallilla vuoden 2021 peruskoulun päättäneiden aineistosta ja kausaalivaikutus $E(Y \mid do(X = 0))$ estimoidaan sovittamalla malli vuonna 2018–2020 peruskoulun päättäneiden aineistoon ja ennustamalla kausaalivaikutus tällä mallilla vuonna 2021 peruskoulun päättäneiden aineistolle. Saadut tulokset tulkitaan luvussa 6 keskimääräistä kausaalivaikutusta hyödyntäen. Lopuksi lukuun 7 on koottu pohdintaa tutkielmaan ja tutkimusongelman käsittelyyn liittyen.

2 Oppivelvollisuuslaki

2.1 Taustaa

Suomen koulutustason nousu on pysähtynyt vuosina 1975–1979 syntyneisiin, mikä on viime vuosina herättänyt suurta koulutuspoliittista keskustelua [Witting, 2021]. Nuorten kouluttautumisella on selkeitä kytköksiä taloudelliseen ja sosiaaliseen pääomaan, minkä vuoksi nuorten peruskoulun jälkeiset koulutusvalinnat vaikuttavat nuorten elämään, niin heidän opiskeluajan kuin koko elämänsä osalta [Ruohola, 2012]. Wittingin [2021] mukaan vuosina 1975–1979 syntyneistä noin 88 prosenttia on suorittanut perusasteen jälkeisen tutkinnon ja korkea-asteen tutkinnon on suorittanut 47 prosenttia. Moni 1980-luvulla tai sen jälkeen syntyneistä on kuitenkin vailla perusasteen jälkeistä tutkintoa. Tämä on kyseenalaistanut Suomen asemaa koulutuksen

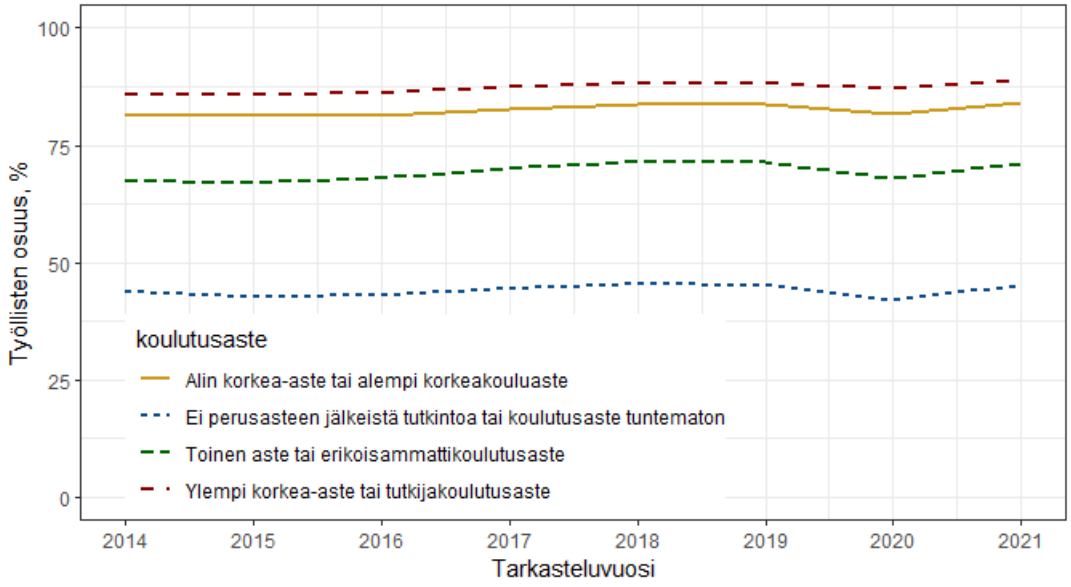
mallimaana.



Kuva 2.1: Työttömien osuus 18–64 vuotiaista koulutusasteen mukaan vuosittain 31.12. *Aineiston lähde: Työssäkäynti, Statfin, Tilastokeskus*

Muun muassa digitalisoitumisen myötä työmarkkinat ovat muuttuneet vuosien saatossa. Osaamisvaatimukset ovat kasvaneet ja työllistyminen nykyajan työelämässä edellyttää vähintään toisen asteen tutkinnon suorittamista (Kuvat 2.1 ja 2.2). Wittingin [2021] mukaan vuonna 2019 ainoastaan puolet 25–64-vuotiaista, pelkän perusasteen koulutuksen suorittaneista oli työllisiä. Toisen asteen koulutuksen suorittaneista hieman yli 70 prosenttia oli työllisiä ja korkea-asteen tutkinnon suorittaneista jopa 88 prosenttia oli työelämässä. Jos tarkastelu rajataan ainoastaan 25–34-vuotiaisiin, erot ovat vieläkin suuremmat.

Witting [2021] huomauttaa artikkelissaan, että eritutkimusten mukaan esimerkiksi vanhempien toimeentulo-ongelmat vaikuttavat keskeisesti lasten tulevaisuuteen, eli "huono-osaisuus periytyy". Ahosen [2021] mukaan peruskoululaisten keskinäiset osaamiserot ovat Suomessa muita maita pienemmät. Ahosen artikkelin tulokset perustuvat 2000-luvun alun kansainväliseen koulusaavutusmittaukseen. Ahosen mukaan samainen mittaus osoitti, että nuoren perhetausta ja sosioekonominen asema vaikuttivat edelleen nuoren koulutuspolkuihin ja sitä kautta tulevaisuuteen. Perhetausta näkyi muun muassa tutkintoon päätymisessä, opintomenestyksessä sekä koulutuksessa pysymisessä.



Kuva 2.2: Työllisten osuus 18–64 vuotiaista koulutusasteen mukaan vuosittain 31.12. *Aineiston lähde: Työssäkäynti, Statfin, Tilastokeskus*

2.2 Oppivelvollisuuslain historiaa

Vuonna 1921 voimaan tullut oppivelvollisuuslaki täytti 100 vuotta oppivelvollisuuslain uudistuksen voimaantulovuonna 2021. Pääministeri Marinin hallituskaudella oppivelvollisuus laajennettiin koskemaan toisen asteen koulutusta, samalla oppivelvollisuusikä nousi 18 ikävuoteen. Uudistus astui voimaan asteittain vuoden 2021 aikana.[Oppivelvollisuuslaki, 1214/2020]

Vuonna 1921 voimaan astuneella oppivelvollisuuslailla, haluttiin antaa kaikille 7–12 vuotiaille lapsille oikeus ja velvollisuus opetukseen ja siitä nauttimiseen, perhetaustasta riippumatta. Länsimaihin yleisen oppivelvollisuuden ajatus vakiintui Ranskan suuren vallankumouksen myötä. Valistusajattelijat korostivat, ettei oikeus sivistykseen saanut olla minkään ryhmän etuoikeus. [Ahonen, 2021] Ahosen mukaan Suomalainen oppivelvollisuuslaki on nähty koko kansan koulun mahdollistajana sen syntyajoilta lähtien. Ennen lakiuudistusta yksi viidesosa suomalaisista lapsista ei käynyt kansakoulua. Lakiuudistuksen myötä lasten ja nuorten tasa-arvo parani, sillä tämä viidennes kävi koulua lain vaatimat kuusi vuotta.

Uudistuksen myötä kuntien oli pakko rakentaa kouluverkkoja, mikä paransi maan eri osien yhdenvertaistumista. Vaikka vuonna 1921 voimaan astunut oppivelvollisuuslaki yhtenäisti lasten ja nuorten asemaa, oli syrjäseuduilla edelleen heikompi mahdollisuus opiskeluun. Oppositio asetti lakiin va-

rauksia, joiden myötä erityistarpeiset lapset, sekä syrjäseudulla asuvat voitiin jättää vaille koulunkäyntimahdollisuutta. Myös oppivelvollisuusiän jälkeisille kahdelle ikäluokalle säädetyt jatkoluokat jäivät maaseutukunnissa usein toteutumatta. [Ahonen, 2021]

Seuraava isompi muutos oppivelvollisuuden suhteen toteutui vuonna 1968, kun säädettiin peruskoululain ilmainen yhdeksänvuotinen kokeilu. Tämä muutos laajensi oppivelvollisuuden 16 ikävuoteen asti. Samalla lakiuudistus korjasi koulutuksenyhdenvertaisuuteen liittyviä puutteita, kuten jatko-opinto mahdollisuuksien yhdenvertaistamisen. Ahosen [2021] artikkelista käy ilmi, että jo yhden sukupolven aikana oppivelvollisuuskokeilu osoitti lupauksen yhtenäisestä mahdollisuudesta jatko-opintoihin pitävän paikkansa. Vuonna 1988 lukion tai ammatillisen oppilaitoksen päättötutkinnon suoritti 49 prosenttia, mikä on 14 prosenttiyksikköä enemmän kuin ennen uudistusta.

Niin teollisuus, palveluelinkeino, kuin maatalouskin ovat vuoden 1968 uudistuksen jälkeen kehittyneet ja digitalisoituneet, minkä myötä peruskoulutasoisesti koulutettujen on haastavaa löytää työpaikkoja. Kehityksen myötä koulutuspoliitikkojen katse on siirtynyt toisen asteen koulutukseen. Vuonna 2019–2020 hallitus laati eduskunnalle esityksen oppivelvollisuuden laajentamisesta 18-vuotiaisiin saakka. (Ahonen [2021], Kalenius [2014])

2.3 Muutokset uuden oppivelvollisuuslain myötä

Uusi oppivelvollisuuslaki astui voimaan 1.8.2021, hakeutumisvelvoitetta koskevat säännökset tulivat voimaan hieman aiemmin 1.1.2021. Yksi oppivelvollisuuden laajentamisen tavoitteista on, että jokainen nuori suorittaa toisen asteen koulutuksen, sillä osaamisvaatimukset työelämässä kasvavat ja työllistyminen edellyttää vähintään toisen asteen tutkinnon suorittamista. Uudistuksella pyritään muun muassa kaventamaan oppimiseroja, nostamaan koulutus- ja osaamistasoa kaikilla koulutusasteilla sekä kasvattamaan yhdenvertaisuutta, koulutuksellista tasa-arvoa ja lasten ja nuorten hyvinvointia. [OKM, 2021b]

Uudistuneen oppivelvollisuuslain myötä oppivelvollisuus laajeni 18 ikävuoteen saakka ja toisen asteen opinnoista tuli maksuttomia laajennetun oppivelvollisuuden piiriin kuuluville nuorille. Konkreettisesti velvoite hakeutua toisen asteen koulutukseen koskee nuoria, jotka keväällä 2021 olivat perusopetuksen 9. luokalla. Tästä alkaen oppivelvollisuuden laajennus tulee voimaan ikäluokka kerrallaan ja koskee lähes kaikkia perusopetuksen päättäneitä toiselle asteelle siirtyviä nuoria. [OKM, 2021a] Poikkeuksena tähän ryhmään ovat muun muassa ennen vuotta 2003 syntyneet. Heidän perusopetuslain mukainen oppivelvollisuutensa on päättynyt ennen vuotta 2021, vaikka he vielä olisivat peruskoulussa vuonna 2021 tai sen jälkeen. Uuden oppivelvollisuus-

lain piiriin eivät myöskään kuulu henkilöt, joilla ei ole kotikuntaa Suomessa. Tähän joukkoon sisältyvät muun muassa tilapäisesti Suomessa opiskelutarkoituksessa oleskelevat. Oppivelvollisuuden laajeneminen ei myöskään koske nuoria, joiden kotikunta on Ahvenanmaalla. [OKM, 2021b]

Uudistuneen oppivelvollisuuslain myötä toimijoiden velvollisuus ohjata nuoria jatkuu myös toisen asteen opintoihin. Perusopetuksen järjestäjän tulee ohjata ja valvoa, että nuori aloittaa opintonsa toisella asteella. Jos tämä velvollisuus nuoren osalta ei täyty, tulee perusopetuksen järjestäjän osoittaa nuorelle siirtymävaiheen koulutuksesta opiskelupaikka. Ennen uudistusta siirtymävaiheen opetuksia olivat perusopetuksen lisäopetus (10. luokka), lukioon valmentava koulutus (LUVA) ja ammatilliseen koulutukseen valmentava koulutus (VALMA). Syksystä 2022 alkaen näiden tilalla järjestetään tutkintoon valmentavaa koulutusta eli TUVA koulutusta. Työhön ja itsenäiseen elämään valmentavaa koulutusta järjestetään edelleen vaativaa erityistä tukea tarvitseville nuorille. (OKM [2021b], Oppivelvollisuuslaki [1214/2020])

Toisen asteen koulutuksen järjestäjillä on velvollisuus ohjata ja valvoa opiskelijoiden opintojen etenemistä. Oppivelvollisuuslaki velvoittaa myös huoltajia huolehtimaan nuoren opinnoista yhdessä toisen asteen järjestäjien kanssa. Koulutuksen järjestäjällä on velvollisuus ilmoittaa huoltajille, jos opinnot eivät etene suunnitellusti. Heillä on myös velvollisuus järjestää tarvittavia tukitoimia, jotta oppivelvollinen nuori kykenee suoriutumaan opinnoistaan. [Oppivelvollisuuslaki, 1214/2020] Tarvittaessa tukitoimia ja vaihtoehtoja selvitetään yhteistyössä toisen koulutuksen järjestäjän kanssa.

Oppivelvollisuus päättyy, kun opiskelija täyttää 18 vuotta tai suorittaa ammatillisen tutkinnon tai ylioppilastutkinnon [Opetushallitus, 2021]. Ensimmäisen toisen asteen tutkinnon suorittaminen on oppivelvolliselle maksutonta. Maksuttomuus jatkuu sen vuoden loppuun, kun opiskelija täyttää 20 vuotta. Oppivelvollisuutta voi suorittaa lukiokoulutuksen ja ammatillisen koulutuksen lisäksi joko kansanopistojen oppivelvollisille suunnatussa vapaan sivistystyön koulutuksessa, työhön tai itsenäiseen elämään valmentavassa koulutuksessa, aikuisen perusopetuksessa (tietyin ehdoin), tai saamelaisalueen koulutuskeskuksen saamen kielen ja kulttuurin koulutuksessa. Lisäksi oppivelvollisuutta voi suorittaa 10. luokalla, LUVA tai VALMA koulutuksessa lukuvuonna 2021–2022, sekä syksystä 2022 alkaen TUVA koulutuksessa. (OKM [2021b], Oppivelvollisuuslaki [1214/2020])

3 Aineisto

Tässä kappaleessa esitellään pro gradu -tutkielmassa käytetty aineisto ja aineiston kokoaminen. Vaikka tämän pro gradu -tutkielman perusjoukon poh-

jana käytetään samoja aineistoja, kuin Tilastokeskuksen Suomen virallisen tilaston aineistot, tutkielman aineiston luvut eroavat jonkin verran näistä luvuista. Syynä tähän on tarkasteltavan perusjoukon erilaisuus verrattuna Tilastokeskuksen julkaisemien tietojen perusjoukkoon. Lisäksi Tilastokeskuksen julkaisemissa tiedoissa on ikärajaus vasta vuodesta 2020 alkaen, tämän vuoksi tutkielman aineiston vuosien 2020 ja 2021 luvut vastaavat paremmin Tilastokeskuksen julkaisemia lukuja, kuin vuosien 2018 ja 2019 luvut (taulukko 1).

Vuosi	Osuus:Statfin	Osuus:aineisto
2018	3.1%	2.5%
2019	3.2%	2.3%
2020	1.6%	1.5%
2021	1.8%	1.7%

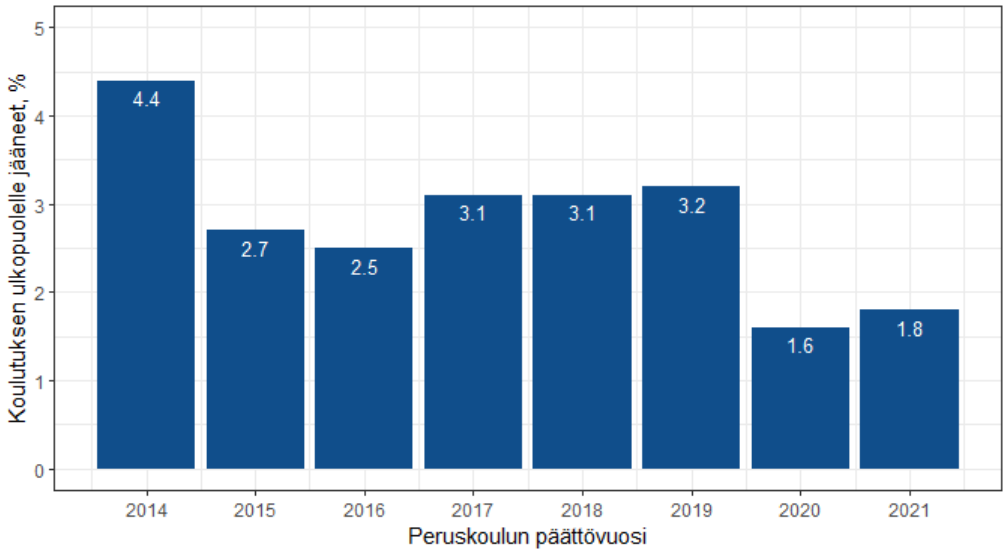
Taulukko 1: Vuosina 2018–2021 peruskoulun päättäneiden koulutuksessa jatkamattomien osuudet, Tilastokeskuksen Statfin tietokannan ja tutkimusaineiston mukaan.

Tilastokeskuksen julkaisemien tietojen mukaan vuosina 2014–2020 peruskoulun 9. luokan päättäneistä 1.6–4.4 % ei jatkanut tutkintotavoitteista eikä siirtymävaiheen opiskelua. Suurimmillaan tämä osuus oli vuonna 2014 ja pienimmillään vuonna 2020. (Kuva 3.1) [virallinen tilasto , SVT]. Kuvasta 3.1 täytyy huomata, että Tilastokeskuksen julkaisemissa luvuissa on ikärajaus vuodesta 2020 alkaen. Tutkimusaineiston perusteella vuosina 2018–2021 opinnoissa jatkamattomien osuudet vaihtelivat 1.5–2.5 % välillä, ollen pienimmillään vuonna 2020.

3.1 Aineiston kokoaminen

Kokosin tutkielman aineiston yhdistämällä SAS EG ohjelmalla tietoja eri rekisteriaineistoista. Rekisteri on informaatiojärjestelmä, joka tuottaa toistuvaa yksikkötasoista aineistoa jollekin määrätylle joukolle yksiköitä. Rekisterianeisto tai hallinnollinen aineisto on yleensä kaikkien Suomen kansalaisten tietoja sisältävä valtionhallinnon ylläpitämä tiedosto. Kun yksilön jokin tieto muuttuu, niin tieto rekisterissä päivittyy. [Tilastokeskus, 2023] Koska rekisteriaineisto sisältää tiedot kaikista perusjoukon henkilöistä, on kyseessä niin sanottu totaaliaineisto. Vaikka kyseessä on totaaliaineisto, voi aineistossa olla puutteita muun muassa tiedonkeruuongelmien ja inhimillisten virheiden vuoksi.

Tutkielmaa varten muokatun aineiston pohjana on vuosina 2018–2021 peruskoulun päättäneet 9. luokkalaiset. Tiedot peruskoulun päättäneistä pe-



Kuva 3.1: Vuosina 2014–2020 peruskoulun 9. luokan päättäneiden, tutkintotavoitteisessa tai siirtymävaiheen koulutuksessa jatkamattomien osuudet. *Aineiston lähde: Koulutukseen hakeutuminen, Statfin, Tilastokeskus*

rustuvat Opetushallituksen rekisteriin. Aineiston perusjoukkoon tehtiin aineiston koontivaiheessa lisärajauksia oppivelvollisuuden laajenemisen piiriin kuuluvaa ryhmää silmällä pitäen. Mukaan on otettu jokaiselta tilastovuodelta ainoastaan yhdeksännen luokan päättäneet. Mukana ei ole henkilöitä, joiden kotikunta on Ahvenanmaalla, eikä henkilöitä, joiden kotikunta ei ole tiedossa. Oppivelvollisuuden laajeneminen 2021 ei koske henkilöitä, jotka ovat syntyneet ennen vuotta 2003. Tämä on otettu aineistossa huomioon siten, että vuonna 2020 peruskoulun päättäneistä on mukana henkilöt, jotka ovat syntyneet vuonna 2002 tai sen jälkeen, peruskoulun 2019 päättäneistä mukana ovat vuonna 2001 tai sen jälkeen syntyneet ja vuonna 2018 peruskoulun päättäneistä mukana ovat vuonna 2000 tai sen jälkeen syntyneet.

Eri rekisteriaineistojen tietojen yhdistämisen vuoksi tässä pro gradu -tutkielmassa tarkastellaan niitä henkilöitä, joiden henkilötunnus on oikeellinen. Ennen perusjoukon rajausta aineistossa oli mukana 290 henkilöä, joiden henkilötunnus oli puutteellinen. Heistä 200 oli tilastovuoden 2019 aineistossa. Lähes kaikilla puutteellisen henkilötunnuksen omaavilla henkilöillä asuinkunta, kansalaisuus tai molemmat olivat tuntemattomia. Perusjoukon rajausten jälkeen jäljelle jäi yksi pois rajattava, puutteellisen henkilötunnuksen omaava henkilö, jonka asuinkunta on Suomessa. Kaikkien rajausten jälkeen perusjoukkoon kuuluu 225 980 henkilöä, noin 56 000 jokaista tilastovuotta kohden (taulukko 2).

Täydensin aineiston perusjoukon taustatietoja Tilastokeskuksen henkilötietovarannon (Herttua) pohjalta yksilöivän henkilötunnisteen perusteella. Jos perusjoukon henkilöllä oli puuttuvia taustatietoja, on tietovarannosta haettu näille henkilöille täydentäviä taustatietoja. Lisäksi tietovarannosta on yhdistetty tieto nuorten vanhemmista sekä vanhempien taustatietoja. Näitä taustatietoja ovat muun muassa sosioekonominen asema, syntyperä ja äidinkieli. Vuosien 2018–2019 peruskoulun päättäneiden aineistoissa oli valmiiksi mukana peruskoulun päättäneiden eri aineiden arvosanatiedot. Vuosien 2020 ja 2021 peruskoulun päättäneiden osalta nämä tiedot haettiin erikseen Tilastokeskuksen opetushallitukselta saamasta Koski-tietovarannosta ja yhdistettiin yksilöivän henkilötunnisteen avulla kohdehenkilöihin.

2018	2019	2020	2021
55 498	56 216	56 816	57 449

Taulukko 2: Tutkimusaineistosta lasketut peruskoulun päättäneiden lukumäärät vuosina 2018–2021.

Taustatietojen lisäksi aineistoon on yhdistetty tiedot toisen asteen koulutukseen hakeutumisesta sinä vuonna, jolloin kyseinen henkilö on suorittanut peruskoulun 9. luokan. Toisen asteen koulutukseen hakeneella tarkoitetaan tässä yhteydessä henkilöä, joka on peruskoulun jälkeisen toisen asteen ammatillisen koulutuksen tai lukiokoulutuksen yhteishaussa ensisijaisesti hakenut johonkin tilastovuonna alkavaan tutkintoon johtavaan koulutukseen tai hakenut valmentavaan koulutukseen, vapaan sivistystyön koulutukseen tai erityisopetuksena järjestettävään ammatilliseen koulutukseen. Tiedot koulutukseen hakeneista perustuvat Opetushallituksen toisen asteen hakurekistereihin.

Lopuksi lisäsin aineistoon tiedon, onko perusjoukkoon kuuluva henkilö ottanut opiskelupaikan vastaan kyseisenä tilastovuonna. Opiskelupaikan vastaanotaneiden tieto on aineistossa poikkileikkaustieto ajankohdan 20.9 tilanteesta. Henkilöt, jotka ovat kyseisenä päivänä kirjoilla toisen asteen oppilaitoksessa, valmentavassa koulutuksessa tai 10. luokalla, luokitellaan opintoja jatkaneeksi henkilöiksi ($Y = 0$). Muussa tapauksessa henkilö ei ole jatkanut opintoja ($Y = 1$). Tiedot opintojen jatkamisesta pohjautuvat koulutuksen järjestäjien/oppilaitosten ilmoittamiin tietoihin sekä Tilastokeskuksen koulutuksen järjestäjien kautta tai suoraan oppilaitoksilta keräämiin tietoihin,

Jokainen perusjoukon henkilö on aineistossa vain kerran kyseistä tilastovuotta kohden. Tietojen yhdistämisen jälkeen kaikki henkilötunnisteet poistettiin, eikä yksittäistä henkilöä ole mahdollista tunnistaa aineistosta. Eri aineistojen yhdistämisen jälkeen tutkielman aineistossa on useita kymmeniä taustamuuttujia, joita tarkastellaan seuraavassa osiossa.

3.2 Aineiston kuvailu

Lopullisessa yhdistetyssä aineistossa on yhteensä 225 980 riviä ja 80 eri muuttujaa. Näistä muuttujista on mallinnusvaiheessa johdettu muuttujia muun muassa erilaisilla ryhmittelyillä. Aineiston muuttujat voidaan karkeasti jaotella kohdehenkilöä kuvaaviin, lähtökoulua kuvaaviin, kohdehenkilön vanhempia kuvaaviin muuttujiin. Vaikka aineistossa on useita muuttujia, kaikki eivät ole esimerkiksi puuttuvan tiedon vuoksi käyttökelpoisia.

Käsittelen seuraavaksi osaa niistä muuttujista, jotka ennakkotiedon perusteella vaikuttavat nuoren koulutuspolkuihin ja jotka aineiston perusteella tukevat tätä tulkintaa. Tutkitun tiedon pohjalta muun muassa vanhempien sosioekonominen asema ja nuoren syntyperä vaikuttavat nuoren kouluttautumiseen [Witting, 2021]. Nuoret päätyvät herkästi vastaaville koulutusaloille vanhempiensa kanssa. Erityisesti äidin matala kouluttautuminen vaikuttaisi olevan kytköksissä lasten matalaan koluttautumiseen. Myös sukupuolittainen eriytyminen koulutttautumisen suhteen on edelleen nähtävissä nuorten keskuudessa. [Keski-Petäjä and Witting, 2016] Lisäksi nuoren peruskouluaikaisella koulumenestyksellä on vaikutusta niin opiskelupaikan valintaan, kuin haetun opiskelupaikan sisäänpääsyynkin.

3.2.1 Kohdehenkilöön kohdistuvat muuttujat

Kohdehenkilöön kohdistuvia muuttujia on aineistossa 35 kappaletta. Nämä muuttujat kuvaavat muun muassa kohdehenkilön äidinkieltä, ikää, syntymävuotta, kansalaisuutta, syntyperää sekä eri oppiaineiden yhdeksännen luokan arvosanatietoa. Eri muuttujien lyhenteet ja selitykset löytyvät liitteenä olevista taulukoista 13 ja 14.

Lopullisessa tutkimusaineistossa lähes kaikki (93.0 %) kohdehenkilöt olivat peruskoulun päättövuonna 15-vuotiaita ja hieman yli puolet (51.1 %) olivat poikia. Peruskoulun päättövuosittain tarkasteltuna koulutuksessa jatkamattomien tyttöjen osuudet olivat jokaisena vuonna suuremmat, mitä samana vuonna koulutuksessa jatkamattomien poikien osuudet. (taulukko 3). Vuosina 2018–2019 sekä tyttöjen että poikien koulutuksessa jatkamattomien osuudet olivat yli 2 % ja vuosina 2020–2021 alle 2 %.

Äidinkieleltään suomenkielisiä on aineistossa 91 %, heistä noin 2.0 % ei jatkanut toisen asteen opintoja heti 9. luokan päätyttyä. Aineistossa äidinkieli muuttujaa on karkeistettu siten, että ne kielet, joita äidinkielenään puhuu alle 50 henkilöä, on yhdistetty samaan luokkaan. Tähän luokkaan on yhdistetty myös 310 henkilöä, joiden äidinkielitieto puuttuu. Äidinkielen mukaan tarkasteltuna opinnoissa jatkamattomien osuudet vaihtelevat 1.5–4.8 % välillä.

Peruskoulun	Ei jatkanut opintoja, %	
päättövuosi	tytöt	pojat
2018	2.6	2.3
2019	2.5	2.1
2020	1.5	1.4
2021	1.8	1.6

Taulukko 3: Koulutuksessa jatkamattomien osuudet peruskoulun päättövuoden mukaan sukupuolittain

Aineistossa on syntyperä muuttuja, joka jakaa henkilöt syntyperän mukaan neljään luokkaan: suomalaistaustainen syntynyt suomessa, suomalaistaustainen syntynyt ulkomailla, ulkomaalaistaustainen syntynyt suomessa sekä ulkomaalaistaustainen syntynyt ulkomailla. Syntyperämuuttujan osalta aineistossa oli 35 henkilöä, joiden tieto oli tuntematon. Heidän kohdallaan syntyperä on päätelty äidinkielen sekä isän ja äidin syntymävaltion avulla.

Syntyperämuuttujan suhteen tarkasteltuna kaikista muissa ryhmissä, paitsi "suomalaistaustainen syntynyt suomessa"yli 2% ei jatkanut opintoja heti 9. luokan jälkeen. Kun tarkastellaan ulkomaalaistaustaisia, jotka ovat syntyneet ulkomailla, jopa 3 % ei jatkanut opintoja heti peruskoulun päätyttyä (taulukko 4).

syntyperä	Ei jatkanut opintoja
suomalaistaustainen syntynyt suomessa	1.94
suomalaistaustainen syntynyt ulkomailla	2.76
ulkomaalaistaustainen syntynyt suomessa	2.17
ulkomaalaistaustainen syntynyt ulkomailla	3.04

Taulukko 4: Koulutuksessa jatkamattomien osuudet syntyperän mukaan ryhmiteltynä, vuosina 2018–2020

Jokaisesta peruskoulun päättäneestä on aineistossa tieto peruskoulun päättöarvosanasta oppiaineittain. Päättöarvosana on joko numeerinen tai merkintänä S eli suoritettu tai H eli hylätty. Joistain oppiaineista, kuten opinto-ohjauksesta, arvosanatieto puuttuu lähes kaikilta. Tällaisia oppiaineita ei ole otettu mallintamiseen mukaan. Logistisessa mallinnuksessa malliin on harkittu vain oppiaineita, joissa arvosana puuttuu alle yhdeltä prosentilta kohdejoukon henkilöistä. Puuttuva tieto on koodattu samaan luokkaan hylätyn arvosanan kanssa.

Arvosanatietojen avulla jokaiselle perusjoukon henkilölle on laskettu arvosanojen keskiarvo. Mallinnuksessa keskiarvo on jatkuvana muuttujana ja voi

saada arvoja väliltä [4.5 , 10]. Jos jonkin oppiaineen arvosanatieto on puuttuva jollain kohdehenkilöistä, on tämä otettu huomioon keskiarvon laskennassa. Aineistossa on nuoren asuinpaikkaa kuvaavaa tietoa erilaisilla tarkkuusasteilla. Suuralueittain tarkasteltuna on huomattavissa pieniä eroja koulutuksessa jatkamisessa. Helsinki-Uudenmaan alueella 1.6 % ei jatkanut opintoja heti peruskoulun jälkeen, eniten koulutuksessa jatkamattomia oli Pohjois- ja Itä-Suomen alueella (2.6 %).

3.2.2 Vanhempien tietoja kuvaavat muuttujat

Aineiston koonnin aikana perusjoukon henkilöihin yhdistettiin muuttujia, jotka kuvaavat heidän vanhempiensa perustietoja. Näitä muuttujia on aineistossa 22 kappaletta, niistä 11 kuvaa äidin tietoja ja 11 isän tietoja (liitetaulukko 15). Aineistossa on muun muassa vanhempien kansalaisuutta, työllisyystilannetta, äidinkieltä ja sosioekonomista asemaa kuvaavat muuttujat. Tarkastellaan näistä tarkemmin vanhempien sosiekonomisen aseman vaikutusta nuorten koulutuksessa jatkamiseen.

Sosioekonominen asema on tutkimusaineistossa ja mallinnuksessa tarkemmalla tasolla, mutta muuttujan tiedot kuvaillaan tässä tutkielmassa karkeammalla tasolla. Tutkimusaineiston perusteella peruskoulun päättäneistä nuorista, joiden vanhemmista isä tai äiti on joko ylempi tai alempi toimihenkilö, yli 98 % jatkaa opintoja peruskoulun päättövuonna.

Koko perusjoukosta vuosittan koulutuksessa ei jatkanut 1.5 – 2.5 % (taulukko 1). Sosioekonomisen aseman mukaan tarkastellessa 2.9 – 4.0 % kohdehenkilöistä ei jatkanut opintoja, jos jomman kumman (tai molemman) vanhemman sosioekonominen asema on eläkeläinen, muu (työtön, varusmies- tai siviilipalvelus) tai tuntematon (taulukko 5). Näissä ryhmissä koulutuksessa jatkamattomien osuus on selvästi suurempi, kuin minkään yksittäisen vuoden koulutuksessa jatkamattomien osuus. Erityisen suuri koulutuksessa jatkamattomien osuus on silloin, kun isän tai äidin sosioekonominen asema on eläkeläinen. Jos jompikumpi vanhemmista on yrittäjä, työntekijä tai opiskelija koulutuksessa jatkamattomien osuudet vaihtelevat 1.8 % ja 2.9 % välillä.

3.2.3 Peruskoulun tietoja kuvaavat muuttujat

Aineistossa on jokaisen kohdehenkilön osalta mukana tieto siitä, missä oppilaitoksessa kohdehenkilö suoritti peruskoulun yhdeksännen luokan. Tätä oppilaitosta nimetään tässä työssä lähtökouluksi. Lähtökoulua kuvaavia muuttujia on aineistossa 8 kappaletta (liitetaulukko 16). Aineistossa on tietoa muun muassa oppilaitoksen opetuskielestä, maakunnasta, oppilaitoksen omistajasta (yksityinen, kunta) sekä oppilaitostyypistä.

Isän sosioekonominen asema	Ei jatkanut opintoja
Yrittäjä	2.14
Ylempi toimihenkilö	1.58
Alempi toimihenkilö.	1.53
Työntekijä	1.79
Opiskelija	2.93
Eläkeläinen	3.20
Muut	2.90
Tuntematon	3.00
Äidin sosioekonominen asema	Ei jatkanut opintoja
Yrittäjä	2.25
Ylempi toimihenkilö	1.26
Alempi toimihenkilö	1.63
Työntekijä	2.57
Opiskelija	2.69
Eläkeläinen	4.03
Muut	3.47
Tuntematon	3.63

Taulukko 5: Koulutuksessa jatkamattomien osuudet vanhempien sosioekonomisen aseman mukaan ryhmiteltynä, vuosina 2018–2020 .

Tässä aineistossa lähes kaikkien koulujen oppilaitostyyppi on peruskoulu (93.8 %). Toiseksi suurin oppilaitostyyppi on perus- ja lukioasteen koulut (5.2 %) ja loput ovat peruskouluasteen erityiskouluja (0.9 %).

Lähes kaikkien koulujen opetuskieli on suomi (93.8 %). Muita opetuskielijä ovat ruotsi (5.6%), englanti (0.34 %), suomi ja ruotsi (0.12 %) tai jokin muu (0.17 %). Maakunnittain tarkasteltuna oppilaitoksista suurin osa sijaitsee Uudenmaan alueella (28.8 %). Seuraavaksi eniten oppilaitoksia on Pohjois-Pohjanmaalla (9.6 %) sekä Pirkanmaalla (9.2 %).

4 Menetelmät

Tutkimusongelman tilanne on niin sanottu kontrafaktuaalinen (Counter factuals) eli havaitsemattoman vasteen tilanne. Kontrafaktuaalisessa päättelyssä toiminnan ajatellaan tapahtuvan menneisyydessä ja toteutuneille lopputuloksille etsitään eräänlaisia vastakohtia [Roese, 1997]. Pyrkimyksenä on selvittää, aiheuttiko X:n muutos Y:n ja mikä Y olisi, jos X ei olisi tapahtunut. Kontrafaktuaalipäättelyssä kausaalilaskentaa sovelletaan tarkasteluun, mikä kiinnostavan muuttujan Y jakauma olisi ollut, jos toimintoa $do(X = x)$ olisi (tai ei olisi) toteutunut [Pearl, 2009]. Toiminnon kausaalivaikutusta muuttujaan Y tarkastellaan jakaumana $P(Y = y \mid do(X = x))$ [Pearl, 2009].

Jaotellaan tutkimusongelman käsittely kolmeen osa-alueeseen, joita ovat kausaalirakenteen päätteleminen, kausaalivaikutuksen identifiointi sekä kausaalivaikutuksen estimointi. Esitetään tutkimusongelman käsittely näiden kolmen osa-alueen avulla:

1. Kausaalirakenteen päätteleminen:

    - Määritellään kausaalirakenne eli muuttujien väliset kausaalisuhteet graafin avulla. Peruskoulun päätövuosi vaikuttaa taustamuuttujien jakaumaan. Hyödynnetään tätä oletusta kausaalirakenteen päättelyssä ja klusteroidaan taustamuuttujat. (kuva 4.1)

2. Kausaalivaikutuksen identifiointi:

    - Kausaalirakenteen päättelyn jälkeen selvitetään, onko kausaalivaikutus mahdollista estimoida. Mikäli kausaalivaikutus on identifioituva kausaalirakenteen perusteella, vasteet saadaan sovittamalla tarvittaville jakaumille tilastollinen malli.
    - Takaovikriteerin (kaava 1) tilanteessa sovitettava malli on sellainen, jossa muuttujaa Y selitetään kiinnostavalla muuttujalla X

ja muuttujaparin (X, Y) d-separoimalla muuttujajoukolla $\mathbf{Z}$ (kaava 2). Tutkimusongelman tilanteessa $Y = 1$, kun kohdehenkilö ei jatka opintoja ja $Y = 0$ muutoin.

3. Kausaalivaikutuksen estimointi:

   - Takaovikorjauksen tilanteessa intervention $do(X = x)$ kausaalivaikutus dikotomiseen vasteeseen Y voidaan selvittää logistisella regressioanalyysillä.
   - Tavoitteena on estimoida $E(Y \mid do(X = 1))$ ja $E(Y \mid do(X = 0))$ vuonna 2021 peruskoulun päättäneillä oppivelvollisilla nuorilla. Estimointia varten jaetaan aineisto kahteen erilliseen aineistoon peruskoulun päättövuoden perusteella siten, että toisessa aineistossa on 2021 peruskoulun päättäneet ja toisessa aineistossa 2018–2020 peruskoulun päättäneet.
   - $E(Y \mid do(X = 1))$ estimoidaan sovittamalla logistinen regressiomalli vuoden 2021 aineistolle. $E(Y \mid do(X = 0))$ saadaan estimoitua sovittamalla logistinen regressiomalli vuosien 2018–2020 aineistolle ja soveltamalla sovitettu malli vuoden 2021 aineistolle.

4. Lopuksi tulkitaan tulokset keskimääräisen kausaalivaikutuksen avulla.

Käsitellään seuraavaksi tutkimusongelman kannalta oleellisia tilastollisia menetelmiä. Menetelmissä keskitytään kausaalipäättelyyn Pearlin kausaalimallinnuksen näkökulmasta. Ensimmäisessä osiossa käsitellään Pearlin kausaalimallinnukseen liittyviä käsitteitä, graafiteoriaa sekä kausaalivaikutuksen identifiointia. Tämän jälkeen käsitellään logistista regressioanalyysiä, jota käytetään kausaalivaikutuksen estimoimiseen.

4.1 Kausaalipäättely

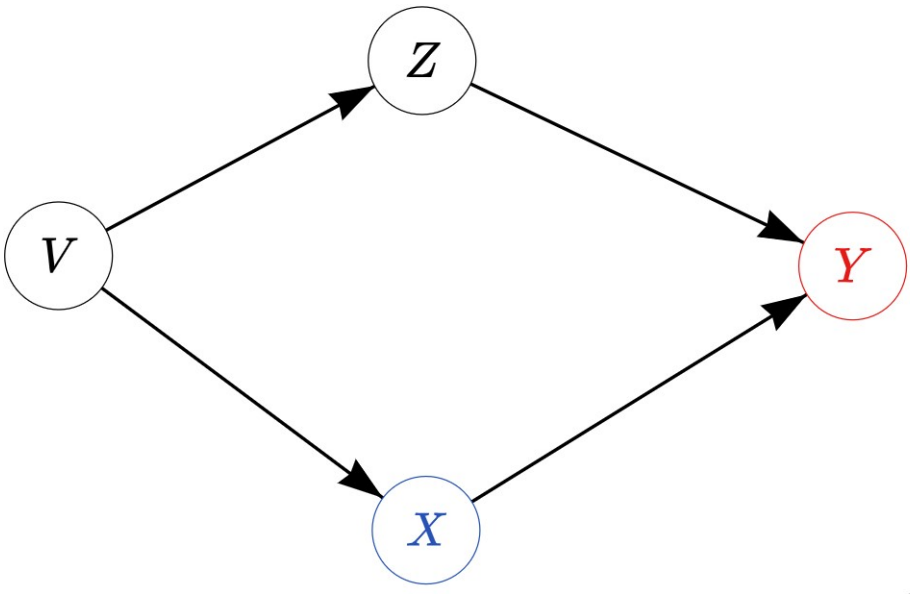
Kausaalipäättelyn teorian pohjana on käytetty pääasiassa Pearlin (Pearl [2009],Pearl et al. [2016]) ja Hollandin [Holland, 1986] teoksia. Kausaalisuus eli syy-seuraussuhteet tarkoittavat tapahtumien välisiä yhteyksiä, joissa toinen tapahtuma (syy) aiheuttaa toisen tapahtuman (seuraus). Kausaalisuhteita muuttujien välillä pyritään todentamaan asiantuntemuksen ja tilastollisten menetelmien avulla. Perinteisesti tilastotiede on keskittynyt tilastollisten riippuvuuksien eli korrelaatioiden tutkimiseen, mutta korrelaatio ei ole tae syy-seuraussuhteelle. (Holland [1988], Pearl et al. [2016])

Pearlin kausaalipäättelyssä pyritään selvittämään, onko kausaalivaikutus määriteltävissä eli identifioituva hyödyntäen interventiota eli kiinnostavaan

muuttujaan kohdistettua toimintaa. Interventiosta käytetään Pearlin teoksissa merkintää $do(X = x)$ ja intervention vaikutusta muuttujaan Y tarkastellaan jakaumana $P(Y = y \mid do(X = x))$.[Pearl, 2009] Pearlin kausaalimallinnuksessa $do(\cdot)$ -operaattorin avulla muodostetut todennäköisyysjakaumat kuvaavat kausaalivaikutusta.

Pearlin kausaalimalli pohjautuu graafiteoriaan ja kausaalipäättelyn ensimmäinen vaihe on kausaalirakenteen määrittely graafien avulla. Suunnattu graafi eli DAG (Directed acyclic graph) koostuu solmuista ja särmistä (kuva 4.1) [Pearl, 2009]. Yksi solmu kuvaa joko yhtä muuttujaa tai useamman muuttujan ryhmää eli klusteria [Tikka et al., 2021]. Klusteroiduista muuttujista on kerrottu tarkemmin Tikka et al. [2021] artikkelissa. Tietyin reunaehdoin klusteroitua muuttujaa voidaan käsitellä kuin yksittäistä muuttujaa ja soveltaa siihen myöhemmin esiteltävien d-separoituvuuden ja takaovikriteerin käsitteitä.

Särmä on nuoli kahden solmun välillä ja kuvaa näiden muuttujien välisiä kausaalisuhteita. DAG:n särmät perustuvat aiempaan tietoon ja kuvaavat muuttujien välisiä kausaalisia oletuksia [Pearl, 2009]. Esimerkiksi kuvan 4.1 DAG:ssa muuttuja V vaikuttaa muuttujiin Z ja X ja muuttujat Z ja X vaikuttavat muuttujaan Y. Reittejä muuttujista tai muuttujajoukoista toisiin kutsutaan poluiksi.



Kuva 4.1: Tutkimusongelman DAG. X on muuttuja johon kohdistetaan interventio, Y on kiinnostuksen kohteena oleva vaste, V on peruskoulun päättövuosi ja **Z** on klusteroitu taustamuuttujien joukko.

Kausaaligraafi esitetään yleensä parina $G = \langle \mathbf{V}, \mathbf{E} \rangle$, missä joukko $\mathbf{V}$ sisältää kaikki graafin solmut. Joukko $\mathbf{E}$ sisältää kaikki parit (X, Y), joille pätee

$$X, Y \subset \mathbf{V}, X \neq Y,$$

missä X kuvaa lähtösolmua ja Y tulosolmua. Näin ollen joukko $\mathbf{E}$ sisältää kaikki graafin G sisältämät särmät. [Pearl, 2009]

Kausaalirakenteen määrittämisen jälkeen selvitetään, onko kausaalivaikutus mahdollista estimoida, eli onko kausaalivaikutus identifioituva. Kiinnostuksenkohteena olevan kausaalivaikutuksen identifioituvuus voidaan selvittää kausaalilaskennan laskusääntöjä ja d-separoituvuutta hyödyntämällä [Pearl, 1995]. Kausaalilaskennan laskusäännöt on esitelty Pearlin [Pearl, 2009] teoksessa, mutta niihin ei syvennytä tässä tutkielmassa. Määritellään kuitenkin d-separoituvuus sekä tutkimusongelman kausaalirakenteen identifioimisen kannalta oleelliset käsitteet, joita ovat takaovikriteeri sekä takaovikorjaus:

Määritelmä 4.1. (d-separoituvuus, Pearl (2009) 1.2.3 ja 1.2.4) *Solmujoukko $\boldsymbol{Z}$ d-separoi eli sulkee polun p, jos ja vain jos*

a) p sisältää ketjun $i \rightarrow m \rightarrow j$ tai haarukan $i \leftarrow m \rightarrow j$ siten, että keskisolmu m kuuluu solmujoukkoon $\boldsymbol{Z}$ tai

b) p sisältää käänteisen haarukan $i \rightarrow m \leftarrow j$ siten, että solmujoukko $\boldsymbol{Z}$ ei sisällä keskisolmua m eikä yhtään keskisolmun m jälkeläistä.

Jos ja vain jos solmujoukko $\boldsymbol{Z}$ sulkee kaikki polut muuttujasta X muuttujaan Y, niin solmujoukko $\boldsymbol{Z}$ d-separoi muuttujaparin (X,Y). Tällöin muuttujat X ja Y ovat riippumattomia ehdolla $\boldsymbol{Z}$

Esimerkkigraafin 4.1 tilanteessa muuttujat X ja Y ovat riippumattomia ehdolla $\mathbf{Z}$, sillä solmu $\mathbf{Z}$ d-separoi polun $X \leftarrow V \rightarrow \mathbf{Z} \rightarrow Y$. Samoin muuttujat V ja Y ovat riippumattomia ehdolla $\mathbf{Z}$, sillä solmu $\mathbf{Z}$ d-separoi polun $V \rightarrow \mathbf{Z} \rightarrow Y$.

Määritelmä 4.2. (Takaovikriteeri, Pearl (2009) 3.3.1 ja 5.3.2) *Olkoon X ja Y kaksi solmua graafissa G. Muuttujajoukon $\boldsymbol{Z}$ sanotaan toteuttavan takaovikriteerin muuttujaparille (X,Y), jos*

a) muuttujajoukko $\boldsymbol{Z}$ ei sisällä muuttujan X jälkeläisiä ja

b) $\boldsymbol{Z}$ sulkee (d-separoi) kaikki sellaiset polut, muuttujien X ja Y välillä, jotka sisältävät nuolen muuttujaan X.

Esimerkkigraafissa 4.1 muuttuja $\mathbf{Z}$ ei ole muuttujan X jälkeläinen ja muuttuja $\mathbf{Z}$ d-separoi muuttujien X ja Y välisen polun, joten muuttuja $\mathbf{Z}$ toteuttaa takaovikriteerin muuttujaparille (X, Y).

Lause 4.3. *(Takaovikorjaus, Pearl (2009) 3.3.2) Jos muuttujajoukko $\boldsymbol{Z}$ toteuttaa takaovikriteerin muuttujaparille (X,Y), niin muuttujan X kausaalivaikutus muuttujaan Y on identifioituva ja saadaan takaovikorjaus kaavalla*

$$P(y \mid do(x)) = \sum_{\boldsymbol{z}} P(y \mid x, \boldsymbol{z}) P(\boldsymbol{z}). \tag{1}$$

Suurten lukujen lakia hyödyntämällä (Shalizi [2013], 21.9) suurella otoskoolla takaovikorjauksen kaava voidaan esittää muodossa:

$$P(y \mid do(x)) \approx \frac{1}{n} \sum_{i=1}^{n} P(y \mid x, \mathbf{z}_i), \tag{2}$$

missä n on otoskoko. Muuttujan (tai muuttujajoukon) $\mathbf{Z}$ jakaumaa ei tarvitse näin ollen estimoida, vaan sen estimaattina toimii itse aineisto. Vastaavasti odotusarvolle pätee:

$$E(y \mid do(x)) \approx \frac{1}{n} \sum_{i=1}^{n} E(y \mid x, \mathbf{z}_i), \tag{3}$$

Missä ehdollinen odotusarvo $E(y \mid x, \mathbf{z}_i)$ kuvaa sovitetun mallin selittäjien x ja z yhteyttä tapahtuman y todennäköisyyteen.

Eri lähteissä (Holland [1988], Pearl et al. [2016]) kausaalivaikutus esitetään hyödyntäen keskimääräistä kausaalivaikutusta ACE (Average causal effect). Määritellään ACE käyttäen Pearlin kausaalimallinnuksen merkintätapoja

$$ACE = E(Y \mid do(X = 1)) - E(Y \mid do(X = 0)) \tag{4}$$

Määritellään vielä tulosten tulkintaa varten populaation käsite. Olkoon populaatio U ja $u \in U$ yksikkö tässä populaatiossa. Yksiköitä ovat kiinnostuksen kohteena olevat henkilöt tai kohteet, joihin interventio kohdistetaan. On oleellista huomata, että kaava (4) mahdollistaa kausaalivaikutuksen arvioimisen keskimääräisellä kausaalivaikutuksella yli populaation U tai sen osapopulaation $U_i \in U$, missä $i = 1 \ldots n$ on osapopulaatioiden määrä.

Tutkimusongelman tilanteessa merkintä $do(X = 1)$ kuvaa interventiota *oppivelvollisuuslain uudistus* ja merkintä $do(X = 0)$ interventiota *oppivelvollisuuslaki ei uudistunut*. Tuloksia käsitellään kahdella eri populaatiolla, minkä vuoksi määritellään että keskimääräinen kausaalivaikutus populaatiossa U on

$$ACE(U) = E(Y \mid do(X = 1)) - E(Y \mid do(X = 0))$$

4.2 Kausaalivaikutuksen estimointi

Kun kausaalirakenne tunnetaan ja kausaalivaikutus on identifioituva, vasteet saadaan sovittamalla tarvittaville jakaumille tilastollinen malli. Takaovikriteerin (kaava 1) tilanteessa sovitettava malli on sellainen, jossa muuttujaa Y selitetään kiinnostavalla muuttujalla X ja muuttujaparin (X, Y) d-separoimalla muuttujajoukolla $\mathbf{Z}$ (kaava 2). Tutkimusongelman tilanteessa $Y = 1$, kun kohdehenkilö ei jatka opintoja ja $Y = 0$ muutoin.

Yhden tai useamman selittäjän vaikutusta vasteeseen voidaan tutkia monipuolisen ja joustavan regressioanalyysin avulla. Tämän osion teoria perustuu Hosmerin [2013] teokseen. Osa tämän tutkielman merkintätavoista saattaa kuitenkin poiketa teoksen merkintätavoista. Tässä tutkielmassa regressioanalyysin tarkastelu keskittyy erityisesti mallin muuttujien valintaan ja mallin hyvyyden tarkasteluun.

Lineaarisessa regressiossa jatkuvalle vasteelle Y mallinnetaan ehdollista odotusarvoa $E(Y \mid \mathbf{x})$, eikä odotusarvon mahdollisia arvoja ole periaatteessa rajoitettu [Hosmer et al., 2013]. Logistinen regressioanalyysi on perinteisen regressioanalyysin erityistyyppi, jolla pystytään mallintamaan dikotomisia, eli kaksiarvoisia vasteita. Kaksi arvoinen vaste tulee logistisen regressioanalyysin tilanteessa koodata arvoiksi yksi ja nolla. [Gelman and Hill, 2007] Logistisen regression tilanteessa ehdollisen odotusarvon $E(Y \mid \mathbf{x})$ mahdollisten arvojen tulee olla välillä $[0, 1]$. Logistisen regressioanalyysin tuloksena saadaan siis tietoa selittäjien yhteydestä tapahtuman todennäköisyyteen. Tällöin intervention $do(X = x)$ kausaalivaikutukselle pätee:

$$P(Y \mid do(X = x)) = E(Y \mid do(X = x))$$

Olkoon Y dikotominen vaste ja oletetaan, että muuttujalle y pätee todennäköisyydet $P(y = 1) = \pi$ ja $P(y = 0) = 1 - \pi$. Yksinkertaisesti ajateltuna, logistinen regressiomalli on tavallinen regressiomalli, jossa todennäköisyyttä π mallinnetaan linkkifunktion $g(\cdot)$ avulla seuraavasti:

$$g(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{5}$$

missä p on selittäjien määrä, x_{ij} on j:nnes selittäjä yksilölle i ja β_j on vastaavan selittäjän regressiokerroin.

Yksinkertaisin linkkifunktio $g(\cdot)$ on identtinen linkkifunktio $g(\pi) = \pi$, jolloin $\pi = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$. Tässä tilanteessa todennäköisyys (tai odotusarvo) π ei välttämättä sijoitu välille $[0, 1]$, eikä sitä voida käyttää dikotomisten vasteiden tilanteessa, vaan käytetään muita linkkifunktioita.

Logistisen regressiomallin tilanteessa käytetään yleisimmin joko logit linkkiä, probit linkkiä tai cloglog (complementary log-log) linkkiä. Näistä yleisimmin käytetty on logit linkki.

Logit linkki:

$$logit(\pi) = log\Big(\frac{\pi}{1-\pi}\Big) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{6}$$

Probit linkki:

$$\phi^{-1}(\pi) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{7}$$

joka saadaan hyödyntämällä normaalijakaumaa, jolloin.

$$\pi = \phi\Big(\frac{x-\mu}{\sigma}\Big)$$

cloglog linkki:

$$log(-log(1-\pi)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{8}$$

jossa π saadaan ääriarvojakaumasta (extreme value distribution) johtamalla

$$\pi = 1 - exp(-exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij},))$$

Kaikki kolme linkkifunktioa käyttäytyvät samankaltaisesti lähellä arvoa $\pi = 0.5$, mutta eroavat toisistaan lähellä arvoja nolla ja yksi. Logit ja probit linkit ovat symmetrisiä todennäköisyyden $\pi = 0.5$ suhteen, mutta Probit linkki vaihtelee vähemmän suhteessa logit linkkiin. Cloglog linkki on pienillä todennäköisyyden arvoilla lähellä Logit linkin arvoja, mutta suurilla todennäköisyyksillä poikkeaa selvästi.

4.2.1 Mallin muuttujien valinta

Kausaalipäättelyssä mallin muuttujat valikoituvat aikaisemman tiedon määrittelemän kausaalirakenteen perusteella. Malliin valikoituvat muuttujat määräytyvät kausaalilaskennan laskusääntöjä ja d-separoituvuutta hyödyntäen siten, että kausaalivaikutus on identifioituva. Tämän tutkielman tilanteessa kausaalirakenteen määrittämisessä havaittiin, että taustamuuttujat

voidaan esittää graafissa yhtenä klusterina. Peruskoulun päättövuosi vaikuttaa taustamuuttujien jakaumaan ja taustamuuttujat vaikuttavat opintojen jatkamiseen.

Klusterin sisältämien taustamuuttujien valintaan on hyödynnetty tutkittua tietoa muuttujien vaikutuksesta nuorten opinnoissa jatkamiseen. Tutkimusaineistossa on kuitenkin mukana muuttujia, jotka kuvaavat likimain samaa asiaa eri muodossa. Esimerkiksi vanhempien työllisyyttä kuvaa kolme eri muuttujaa, joita ovat ammattiasemaa, pääasiallista toimintaa sekä sosioekonomista asemaa kuvaavat muuttujat (liitetaulu 15). Tämän vuoksi mallin muuttujien valinnassa on hyödynnetty erilaisia logistisessa regressioanalyysissä käytettäviä muuttujien valintamenetelmiä. Näiden menetelmien avulla malliin valitaan opinnoissa jatkamiseen vaikuttavista muuttujista ne, joilla mallin selitysaste saadaan mahdollisimman hyväksi.

Logistisen regressioanalyysin selittäjien valinnassa voidaan yleisesti hyödyntää muun muassa eteenpäin valintaa (forward selection) ja taaksepäin valintaa (backward selection). Näistä kahdesta metodista on myös modifioitu niin sanottu askeltava valinta (stepwise selection). [Hosmer et al., 2013] Askeltavaa valintaa käytettäessä täytyy ensin asettaa tilastollinen merkitsevyystaso p_{in}, jolla muuttuja otetaan mukaan malliin ja toinen tilastollinen merkitsevyystaso p_{out}, jolla muuttuja poistetaan mallista. Sopiva taso p_{in}:lle ja p_{out}:lle on lähellä arvoa 0.05.

Askeltavassa valinnassa algoritmi ottaa selittävän muuttujan malliin, jos sen merkitsevyystaso on pienempi kuin asetettu merkitsevyystason raja p_{in}. Jos tällaisia muuttujia on useampia, ottaa algoritmi malliin muuttujan, joka toimii parhaana yksittäisenä selittäjänä. Seuraavassa vaiheessa algoritmi ottaa malliin mukaan toisen muuttujan, jonka merkitsevyystaso on pienempi kuin p_{in} ja joka toimii parhaana lisäselittäjänä edellisen lisätyn muuttujan kanssa. Mikäli muuttujan lisääminen nosti aiemmin lisätyn muuttujan merkitsevyyttä yli merkitsevyystason p_{out}, algoritmi poistaa tämän muuttujan mallista. Tästä eteenpäin algoritmi vuorotellen poistaa muuttujia mallista ja lisää muuttujia malliin asetettujen merkitsevyystasojen mukaisesti. Askeltavan valinnan algoritmi etenee niin kauan, kunnes mallissa on mukana ainoastaan muuttujia, jotka ovat merkitseviä asetetun merkitsevyystason p_{out} mukaan, eikä mikään mallin ulkopuolisista muuttujista ole merkitsevä asetetun merkitsevyystason p_{in} perusteella.

4.2.2 Kerrointen merkitsevyyden testaaminen

Logistisen regressiomallin parametrit estimoidaan suurimman uskottavuuden menetelmällä (method of maximum likelihood). Yleensä menetelmä toimii, kun tapauksia on riittävästi ja selittäjiä ei ole kovin paljoa. [Hosmer et al.,

2013] Logistisen regressiomallin sovittamisen yhteydessä yleensä tarkastellaan estimoitujen parametrien merkitsevyyttä ja arvioidaan ovatko kaikki selittäjät mallin kannalta merkitseviä. Tätä tarkastelua on hyödynnetty myös tässä työssä niiden muuttujien välillä, jotka kuvaavat keskenään samaa ilmiötä. Tällaisia muuttujia ovat esimerkiksi edellisessä osiossa esitetyt vanhempien työllisyystilannetta kuvaavat muuttujat.

Estimoitujen kertoimien merkitsevyyttä voidaan tarkastella Waldin testisuureen avulla. Waldin testisuureen arvo selittäjälle i saadaan jakamalla selittäjän i suurimman uskottavuuden estimaatti $\hat{\beta}_i$ sen estimoidulla keskivirheellä $\hat{SE}$.

$$W_i = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)},$$

Waldin testisuureen kaksisuuntainen p-arvo on $P(\mid Z \mid > W_i) = p$, missä z on standardinormaalijakaumasta saatu satunnaismuuttuja. [Hosmer et al., 2013] Yleisesti käytetään nyrkkisääntöä, jos $p < 0.05$, niin estimaatti $\hat{\beta}_i$ on tilastollisesti merkitsevä ja se kannattaa jättää malliin.

4.2.3 Mallin hyvyyden arviointi

Vaikka yksittäisten muuttujien tarkastelu on tärkeää mallia valittaessa, tulee mallin hyvyyttä tarkastella myös muiden ominaisuuksien perusteella. Hosmer et al. [2013] käsittelee teoksessaan erilaisia keinoja arvioida logistisen regressiomallin hyvyyttä. Esittelen kaksi yleisesti käytettyä menettelyä mallin hyvyyden arvioimiseen. Näitä menetelmmiä ovat havaintoaineiston luokittelu logistisen regressiomallin perusteella sekä yhteensopivuustestit.

Luokittelu tarkastelussa logistisella regressiomallilla ennustettujen todennäköisyyksien pohjalta jokaiselle yksilölle y_i asetetaan arvo 1 tai 0. Tarkastelua varten määritellään leikkauskohta (cutpoint), jonka mukaan jaottelu tehdään. Tavallisimmin tämä leikkauskohta on todennäköisyys 0.5. Jos mallin ennustama todennäköisyys yksilölle y_i on pienempi kuin 0.5, saa yksilö arvon 0 ja jos mallin ennustama todennäköisyys on suurempi kuin 0.5, saa yksilö arvon 1. Tämän luokittelun mukaan muodostetaan nelikenttä (taulukko 6), josta lasketaan kuinka suuri osuus yksilöistä on sijoittunut oikeaan luokkaan.

Mitä suurempi "OIKEIN"osuus on, sitä parempi malli on kyseessä. Luokitteluvirheen, eli väärin ennustettujen osuus, pitää olla aina alle 0.5. Jos luokitteluvirhe on suurempi kuin 0.5, ennustaa malli jossa on ainoastaan vakio, ilman selittäjiä, paremman tuloksen kuin sovitettu malli. Jos luokittelu tarkastelu tehdään mallinnusaineistolla, antaa se ylipositiivisen kuvan todellisesta tilanteesta, minkä vuoksi tarkastelu tulisi tehdä testiaineistolle.

	Ennustettu Y:n arvo 0	Ennustettu Y:n arvo 1	Yhteensä
Todellinen Y:n arvo 0	OIKEIN	VÄÄRIN	n_{1n}
Todellinen Y:n arvo 1	VÄÄRIN	OIKEIN	n_{2n}
Yhteensä	n_{m1}	n_{m2}	n_{mn}

Taulukko 6: Nelikenttä logistisen regressiomallin hyvyyden tarkasteluun

Luokittelu tarkastelussa voi hyödyntää myös niin sanottua linkkuveitsi (jackknife) -menetelmää [Wu, 1986], jossa todellista luokittelutilannetta simuloidaan seuraavalla tavalla:

1. Jätetään aina vuorollaan kukin luokiteltava yksilö pois
2. Lasketaan logistisen funktion kertoimet jäljelle jäävien $n-1$ yksilöiden perusteella
3. suoritetaan pois jätetyn yksilön luokittelu näin saaduilla kertoimilla.

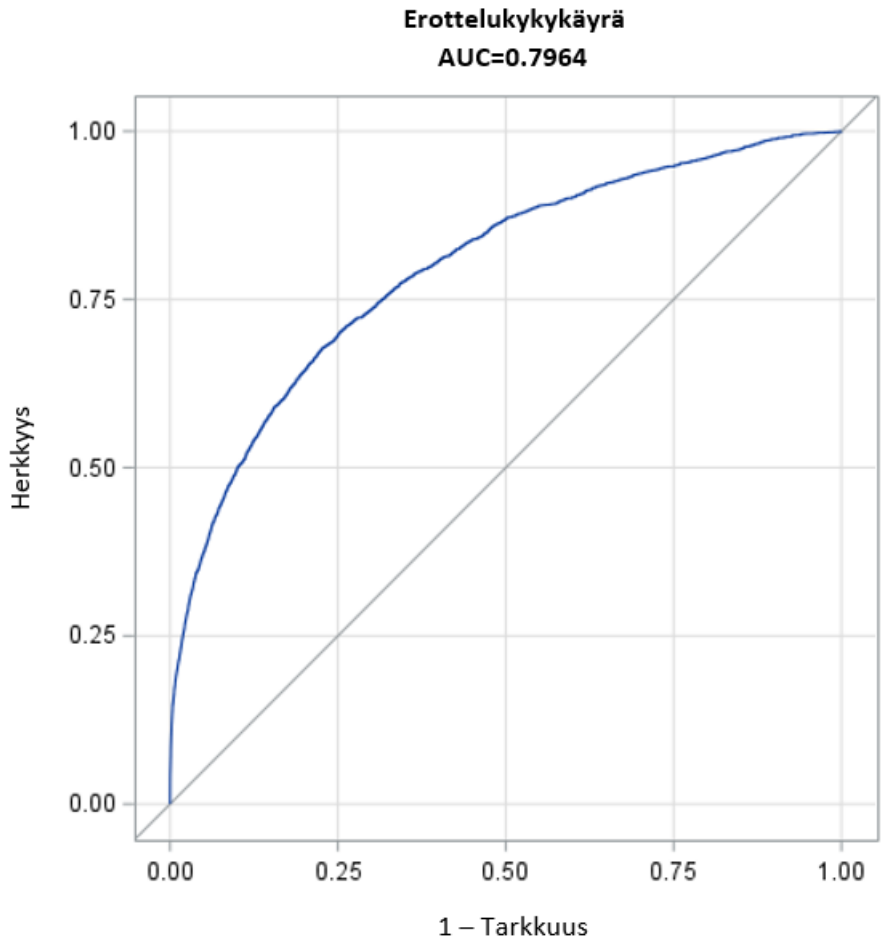
Luokittelua hyödyntävä, mutta "parempi"vaihtoehto on erottelukykykäyrä, eli ROC (receiver operation characteristic) käyrä (kuva 4.2). Tämä käyrä muodostuu mallin herkkyyden (Sensitivity) ja tarkkuuden (Specificity) avulla ja hyödyntää niitä mallin ennustekyvyn laskennassa. Herkkyys kuvaa oikein ennustettujen $y = 0$ osuutta kaikista todellisista $y = 0$ tilanteista

$$\text{herkkyys} = \frac{n_{11}}{n_{1n}}$$

ja tarkkuus kuvaa oikein ennustettujen $y = 1$ osuutta kaikista todellisista $y = 1$ tilanteista

$$\text{tarkkuus} = \frac{n_{22}}{n_{2n}}.$$

Erottelukykykäyrä muodostuu laskemalla herkkyys ja tarkkuus eri leikkauspisteillä välillä $[0, 1]$. Nämä lasketut arvot asetetaan xy-koordinaatistoon siten, että x-akselille sijoitetaan 1-spesifisyys arvo ja y-akselille sensitiivisyys arvo ja näistä pisteistä muodostuu erottelukykykäyrä. Erottelukykykäyrä tulkitaan siten, että mitä suurempi käyrän alle jäävä pinta-ala AUC (Area under the curve) on, sitä parempi on mallin ennustekyky. Erottelukykykäyrän ennustekyvyn tulkinnalle on esitetty seuraavanlaista yleistä ohjetta. [Hosmer et al., 2013]



Kuva 4.2: Sovitetun mallin erottelukykykäyrä. Käyrän alle jäävä pinta-ala AUC=0.7964

$$\text{Jos} = \begin{cases} AUC \leq 0.5 & \text{mallilla ei ole ennustekykyä} \\ 0.5 < AUC \leq 0.7 & \text{mallin ennustekyky on huono} \\ 0.7 < AUC \leq 0.8 & \text{mallin ennustekyky on välttävä} \\ 0.8 < AUC \leq 0.9 & \text{mallin ennustekyky on kiitettävä} \\ AUC \geq 0.9 & \text{mallin ennustekyky on erinomainen} \end{cases}$$

Erottelukykykäyrän alle jäävän pinta-alan tulkinta on hyvä apukeino mallia valittaessa. Kuten luokittelunkin tilanteessa, erottelukykykäyrän antaman informaation hyödyntäminen on haasteellista harvinaisen vasteen Y tilanteessa. Luokittelun ja erottelukykykäyrän ohella mallin hyvyyttä voidaan arvioida erilaisilla yhteensopivuustesteillä. Tässä työssä on hyödynnetty niin sanottua Hosmer-Lemeshovin testiä [Hosmer et al., 2013].

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i \cdot (1 - E_i/n_i)}, df = g - 2 \tag{9}$$

Hosmer-Lemeshovin testissä suuri testisuureen arvo ja p-arvo < 0.5 tarkoittavat mallin huonoa yhteensopivuutta datan kanssa. Kaavassa n_i kuvaa ryhmän i kokoa ja g on ryhmien lukumäärä (tavallisesti $g = 10$). Ryhmä $i = 1$ muodostuu henkilöistä, joiden tapahtuman $y = 1$ todennäköisyys on pienin ja ryhmä $i = 10$ niistä, joiden vastaavan tapahtuman todennäköisyys on suurin. Kunkin ryhmän havaituille tapahtumien määrille O_i lasketaan logistisen mallin perusteella odotetut tapahtumat E_i. [Hosmer et al., 2013]

5 Mallinnus

Tutkielman tavoitteena on selvittää, onko oppivelvollisuuslain uudistaminen vaikuttanut opintojen jatkamiseen niillä nuorilla, jotka kuuluvat laajennetun oppivelvollisuuden piiriin. Muotoillaan tutkimusongelma kausaalipäättelyn mukaiseksi kontrafaktuaalipäättelyksi.

Tutkimusongelman tilanteessa merkintä $do(X = 1)$ kuvaa tilannetta *oppivelvollisuuslain uudistus* ja merkintä $do(X = 0)$ tilannetta *oppivelvollisuuslaki ei uudistunut*. Tavoitteena on estimoida kausaalivaikutus

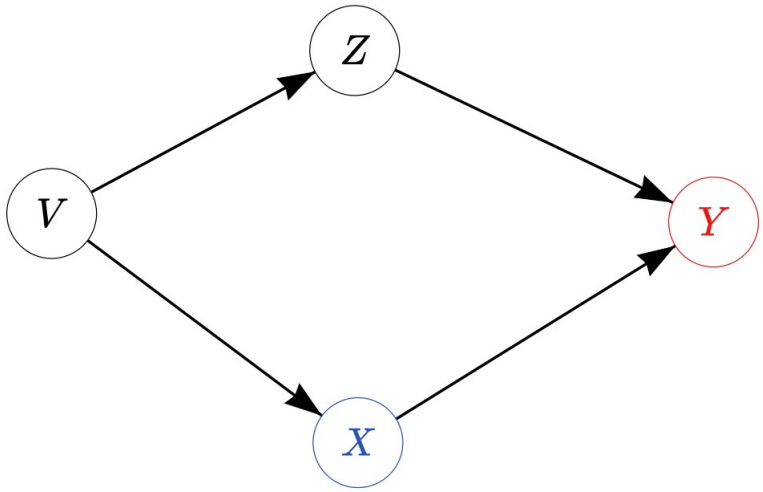
$$P(Y = 1 \mid do(X = 1)) = E(Y = 1 \mid do(X = 1))$$

eli todennäköisyys olla jatkamatta opintoja kun oppivelvollisuuslaki on uudistunut ja kausaalivaikutus

$$P(Y = 1 \mid do(X = 0)) = E(Y = 1 \mid do(X = 0))$$

eli todennäköisyys olla jatkamatta opintoja, jos oppivelvollisuuslaki ei olisi uudistunut.

Tarkastellaan muuttujien välisiä kausaalisuhteita tutkimusongelman tilanteessa kausaalirakenteen avulla. Tutkimusongleman graafissa G (kuva 5.1) on muuttujat X, Y, V, sekä klusteroitu muuttuja $\mathbf{Z}$.



Kuva 5.1: Tutkimusongelman graafi G. X on muuttuja johon kohdistetaan interventio, Y on kiinnostuksen kohteena oleva vaste, V on peruskoulun päättövuosi ja Z on klusteroitu taustamuuttujien joukko.

Graafissa G:

X : Oppivelvollisuuslain uudistus

1= Oppivelvollisuuslaki on uudistunut

0= Oppivelvollisuuslaki ei ole uudistunut

V : Peruskoulun päättövuosi

Voi saada arvot 2018–2021

Y : Opintojen jatkaminen peruskoulun päätyttyä

1= kohdehenkilö ei jatka opintoja

0= kohdehenkilö jatkaa opintoja.

Klusteroitu muuttuja $\mathbf{Z}$ kuvaa niitä taustamuuttujia, jotka on otettu lopulliseen malliin mukaan (taulu 7). Näiden muuttujien jakauma määräytyy peruskoulun päättövuoden perusteella (solmu V) ja nämä muuttujat yhdessä vaikuttavat vasteeseen Y, eli siihen jatkaako nuori opintoja.

Koska klusteroitu solmu $\mathbf{Z}$ d-separoi polun $X \leftarrow V \rightarrow Z \rightarrow Y$ muuttujien (X, Y) välillä ja klusteri $\mathbf{Z}$ ei sisällä muuttujan X jälkeläisiä, klusteri $\mathbf{Z}$ toteuttaa takaovikriteerin ja voidaan soveltaa takaovikorjausta (kaava 1). Näin ollen kausaalivaikutus $P(Y \mid do(X = x))$ saadaan sovittamalla malli, jossa vastetta Y selitetään muuttujilla X ja $\mathbf{Z}$.

Tutkimusongelman selvittämiseksi jaetaan aineisto kahteen erilliseen aineistoon peruskoulun päättövuoden mukaan. Toisessa aineistossa ovat vuonna 2021 peruskoulun päättäneet (57 449 riviä) ja toisessa vuosina 2018–2020 peruskoulun päättäneet (168 501 riviä).

Oppivelvollisuuslain uudistuminen koskee kaikkia vuonna 2021 peruskoulun päättäneiden aineistossa olevia nuoria. Vuonna 2021 tapahtuma $do(X = 1)$ toteutui ja interventio $do(X = 0)$ on kontrafaktuaalinen tilanne. Näin ollen vuonna 2021 peruskoulun päättäneiden aineistosta saadaan tietoa kausaalivaikutuksesta $P(Y = y \mid do(X = 1))$ sovittamalla aineistoon logistinen regressiomalli. Kontrafaktuaalisesta tilanteesta $do(X = 0)$ saadaan tietoa vuosina 2018–2020 peruskoulun päättäneiden aineistosta. Kausaalivaikutus $P(Y = y \mid do(X = 0))$ saadaan mallintamalla vuosina 2018–2020 peruskoulun päättäneiden aineistoa logistisella regressiomallilla ja soveltamalla sovitettu malli vuonna 2021 peruskoulun päättäneiden aineistoon.

5.1 Mallin valinta

Mallinnus aloitettiin käymällä läpi kaikkien kiinnostavien muuttujien tunnuslukuja ja hyödyntämällä olemassa olevaa tutkimustietoa muuttujien valinnassa. Aiemmin tehtyjen tutkimusten perusteella muun muassa sosioekonominen asema ja syntyperä vaikuttavat nuorten kouluttautumiseen [Witting, 2021].

Muuttujien tunnuslukuja tarkastellessa mahdollisia pieniä luokkia yhdistettiin ja joitain muuttujia karkeistettiin. Muuttujien tarkastelun yhteydessä rajattiin osa muuttujista pois puuttuvan tiedon vuoksi. Tällaisia muuttujia olivat muun muassa äidin ja isän pääasiallinen toiminta sekä opinto-ohjauksen arvosanatieto.

Kuten aineiston muokkaus, myös logistinen mallinnus toteutettiin SAS EG ohjelmistolla käyttäen PROC LOGISTIC proseduuria. Mallin hyvyyttä voidaan arvioida luokittelumenetelmällä, minkä vuoksi vuosien 2018–2020 aineisto jaettiin mallinnus- ja testiaineistoksi. Jako toteutettiin satunnaisotannalla PROC SURVEYSELECT proseduurilla. 80 % aineiston kohdehenkilöistä sijoittuivat mallinnusaineistoon ja loput 20 % testiaineistoon. Mallinnusaineistossa oli 134 825 riviä ja testiaineistossa 33 676 riviä.

Malliin harkittiin muuttujia, jotka tutkimustiedon ja aineiston tarkastelun perusteella todennäköisesti vaikuttavat nuorten opinnoissa jatkamiseen.

Aineistossa on muuttujia, jotka kuvaavat samaa asiaa eri muodossa. Tällaisia tietoja ovat esimerkiksi vanhempien työllisyyttä kuvaavaat muuttujat tai nuoren peruskoulun oppimenestystä kuvaavat oppiaineiden arvosanaa kuvaavat muuttujat. Koska tällaisia samankaltaista tietoa kuvaavia muuttujia on useita, eikä mallia haluta ylisovittaa, on mallin muuttujien valinnassa hyödynnetty muun muassa askeltavaa valintaa (stepwise selection).

Mahdollisia malliin sisällytettäviä muuttujia on paljon, joten asetin askeltavan valinnan merkitsevyystasojen p_{in} ja p_{out} rajaksi 0.05. Koska askeltava valinta on hidas usealla muuttujalla, toteutettiin askeltava valinta useamman kerran erilaisilla muuttujavariaatioilla. Menetelmää kokeiltiin pelkkien päävaikutustermien erilaisilla variaatioilla, sekä ottamalla mukaan päävaikutustermien interaktiotermejä. Myöskin erilaisia muuttujien karkeistuksia, luokitteluja ja muun muassa neliöllisten termien vaikutusta mallin hyvyyteen arvioitiin. Erilaisilla muuttuja variaatioilla sovitettuja logistisia regressiomalleja verrattiin keskenään eri linkkifunktioiden tilanteessa.

Lopulliseen malliin valikoitui kohdehenkilöä kuvaavista muuttujista sukupuoli, ikä, kuntaryhmä ja syntyperä. Lisäksi mukaan malliin valikoitui keskiarvo muuttuja, joka on aineistossa jatkuvana muuttujana. Arvosanatiedoista mukaan valikoitui niiden oppiaineiden arvosanatiedot, jotka parhaiten lisäsivät mallin selitysastetta. Näitä arvosanatietoja ovat: A1-kieli, fysiikka, terveystieto, yhteiskuntaoppi, kotitalous, musiikki, käsityö ja liikunta. Lähtökoulua kuvaavista muuttujista mukaan valikoituivat lähtökoulun oppilaitostyyppi, opetuskieli ja maakunta. Vanhempien tietoja kuvaavista muuttujista mukaan otettiin sekä äidin että isän sosioekonominen asema. Lisäksi malliin valikoitui kuntaryhmän ja lähtökoulun maakunnan interaktiotermi. Nämä malliin valikoituneet muuttujat muodostavat klusterin $\mathbf{Z}$ (taulu 7).

Malliin valikoituneista muuttujista kaikkien muiden merkitsevyystaso on reilusti alle 0.05, paitsi kuntaryhmän ja syntyperän. Syntyperä on jätetty malliin, koska tutkitun tiedon perusteella syntyperällä on vaikutusta nuorten koulutuspolkuihin [Witting, 2021]. Kuntaryhmän ja oppilaitoksen maakunnan interaktiotermi on merkitsevä, minkä vuoksi kuntaryhmä on jätetty malliin mukaan. Malliin valikoituneet muuttujat on taulukoitu vapausasteineen taulukkoon 7. Lopullisen mallin muuttujien valintaan vaikuttivat seuraavassa osiossa esiteltävät hyvyystarkastelut.

Vuonna 2021 peruskoulun päättäneiden aineistoon sovitettiin logistinen regressiomalli vastaavilla selittäjillä, kuin vuosina 2018–2020 peruskoulun päättäneiden aineistoon. Vuoden 2021 aineistolla kaikki muuttujat eivät ole yhtä merkitseviä, kuin vuosien 2018–2020 aineistolla (taulukko 8).

Molempien sovitettujen mallien regressiokertoimet ja keskivirheet on taulukoituna liitetauluissa 18 – 22. Liitetauluihin on laskettu regressiokertoimien erotukset mallien välillä. Lähes kaikkien regressiokertoimien erotus on välil-

Selittävä muuttuja	DF	Wald χ^2	Pr > ChiSq
sp	1	77.3952	<.0001
ika	3	257.3111	<.0001
lahtokoulu oltyp	2	54.0771	<.0001
lahtokoulu okieli	4	44.0481	<.0001
kuntaryh	2	0.0007	0.9997
lahtokoulu onuts3	18	316.7382	<.0001
sose isa	17	68.633	<.0001
sose aiti	17	84.6163	<.0001
syntyp2	3	2.5123	0.4731
A1	8	111.0281	<.0001
FY	8	43.5699	<.0001
TE	8	39.8795	<.0001
YH	8	224.9515	<.0001
KO	8	36.1955	<.0001
MU	8	53.2372	<.0001
KS	8	70.6029	<.0001
LI	8	253.8325	<.0001
ka	1	36.0683	<.0001
kuntaryh x lahtokoulu onuts3	36	88.5549	<.0001

Taulukko 7: Logistisen mallin selittävät muuttujat 2018–2020 peruskoulun päättäneiden aineistolla.

Selittävä muuttuja	DF	Wald χ^2	Pr > ChiSq
sp	1	24.5496	<.0001
ika	3	88.6662	<.0001
lahtokoulu oltyp	2	14.3218	0.0008
lahtokoulu okieli	4	29.8767	<.0001
kuntaryh	2	0.0001	0.9999
lahtokoulu onuts3	18	162.9844	<.0001
sose isa	17	41.1447	0.0009
sose aiti	17	49.5123	<.0001
syntyp2	3	10.2988	0.0162
A1	8	28.6009	0.0004
FY	8	22.1092	0.0047
TE	8	38.3899	<.0001
YH	8	271.9016	<.0001
KO	8	30.5751	0.0002
MU	8	22.3165	0.0044
KS	8	23.0847	0.0033
LI	8	89.9609	<.0001
ka	1	1.9844	0.1589
kuntaryh x lahtokoulu onuts3	36	66.8474	0.0013

Taulukko 8: Logistisen regressiomallin selittävät muuttujat vuonna 2021 peruskoulun päättäneiden aineistolla.

lä $[-2; 2]$. Taulukoista voidaan huomata, että regressiokertoimien erotuksista 13 on tämän välin ulkopuolella. Näistä erotuksista 10 on sellaisia, jotka koskeva joko lähtökoulun maakunta -muuttujaa tai interaktiotermiä, joka sisältää lähtökoulun maakunta -muuttujan. Näiden kahden muuttujan regressiokertoimien keskivirheet ovat vuoden 2021 aineistoon sovitetun mallin osalta poikkeuksellisen suuret. Palataan tähän pohdinta luvussa.

5.2 Mallin hyvyystarkastelut

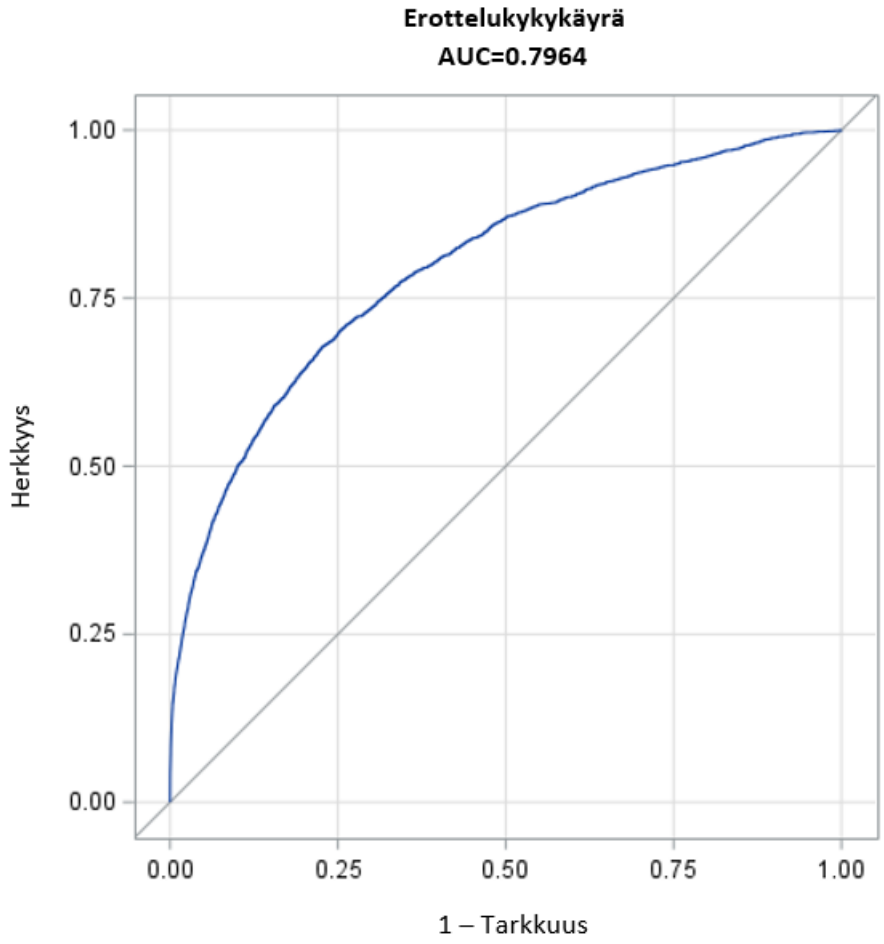
Mallin hyvyyttä on tarkasteltu luokittelun (taulukko 9), erottelukykykäyrän (kuva 5.2), Hosmer-Lemeshovin taulukon (taulu 10) ja Hosmer-Lemeshovin testiarvon avulla. Testiaineistolla tehdyn luokittelun ja malliaineistolla tehdyn erottelukykykäyrän tarkastelun perusteella valikoitunut malli on yhtä hyvä cloglog (oikein luokiteltu 98.0 %, AUC=0.794) ja logit (Oikein luokiteltu 98.0 %, AUC=0.796) linkillä. Sen sijaan probit linkillä oikein luokiteltujen osuus ja ennustuskyky hieman heikkenivät muihin linkki funktioihin verrattuna. Lopulliseen malliin valikoitui logit linkki, sillä logit linkillä Hosmer-Lemeshovin testiarvo (χ^2=15.3108, p=0.0534) oli parempi suhteessa cloglog linkillä muodostetun mallin testiarvoon (χ^2=25.0620, p=0.0015).

Luokittelun (taulukko 9) perusteella malli ennustaa oikein 98.0 % kohdehenkilöistä. Vaikka luokittelun perusteella malli on hyvä, hankaluuksia tuottaa koulutuksessa jatkamattomien ($Y = 1$) pieni joukko. Vaikka malli ennustaisi kaikki henkilöt jatkamaan opintoja ($Y = 0$), luokittelu antaisi silti korkean todennäköisyyden (99.1 %) olla oikeassa ryhmässä. Tarkastelemalla mallin hyvyyttä erottelukykykäyrän (kuva 5.2) avulla, mallin ennustuskyky jää välttävän puolelle, ollen 0.8.

	Ennustettu Y:n arvo 0	Ennustettu Y:n arvo 1	Yhteensä
Todellinen Y:n arvo 0	32951	29	32980
Todellinen Y:n arvo 1	649	47	696
Yhteensä	33600	76	33676

Taulukko 9: Nelikenttä logistisen regressiomallin ennustekyvystä vuosina 2018–2020 peruskoulun päättäneiden aineiston testiaineistolla. Y=0 tarkoittaa koulutuksessa jatkaneita ja Y=1 koulutuksessa jatkamattomia

SAS EG ohjelmassa PROC LOGISTIC proseduuriin on mahdollista lisätä komento LACKFIT, joka automaattisesti tuottaa Hosmer-Lemeshow taulukon ja yhteensopivuustestin tulokset. Taulukko muodostuu siten, että estimoidut todennäköisyydet lajitellaan nousevaan järjestykseen, minkä jälkeen



Kuva 5.2: Erottelukykykäyrä vuosina 2018–2020 peruskoulun päättäneiden aineiston mallinnusaineistolla, lopullisen mallin tilanteessa

ohjelma jakaa aineiston luokkiin. Taulukosta pystyy tarkistamaan kuinka moni kohdehenkilö missäkin luokassa sijoittuu oikeaan ryhmään. Hosmer-Lemeshow taulukosta 10 voidaan nähdä, että ryhmissä 1, 3-5 ja 7 havaittu koulutuksessa jatkamattomien määrä on pienempi kuin estimoitu koulutuksessa jatkamattomien määrä. Lopuissa ryhmissä havaittu koulutuksessa jatkamattomien määrä on suurempi, kuin estimoitu koulutuksessa jatkamattomien määrä. Ero on suurin luokissa 4 ja 5, muissa ryhmissä erot ovat kohtalaiset.

Hosmer-Lemeshovin testin testiarvo $\chi^2 = 15.3108$ ja p-arvo = 0.0534. Koska p-arvo > 0.05 malli soveltuu suhteellisen hyvin aineistoon. Mallinnettava vaste on suhteellisen harvinainen (1.5–2.5% aineiston perusjoukosta vuosittain (taulukko 1)), mikä aiheuttaa omat haasteensa sopivan mallin valinnassa. Tapahtuma $Y = 1$ sisältää todennäköisesti myös satunnaisuutta, jota ei pystytä mallinnuksessa ottamaan huomioon.

		Koulutuksessa jatkamattomat (Y=1)		Koulutuksessa jatkavat (Y=0)	
Luokka	yhteensä	Havaittu	Estimoitu	Havaittu	Estimoitu
1	13485	37	44.56	13448	13440.44
2	13480	73	70.89	13407	13409.11
3	13483	75	91.44	13408	13391.56
4	13487	97	112.47	13390	13374.53
5	13482	110	136.5	13372	13345.5
6	13480	169	166.16	13311	13313.84
7	13482	201	206.34	13281	13275.66
8	13483	279	268.94	13204	13214.06
9	13483	425	393.86	13058	13089.14
10	13480	1348	1322.78	12132	12157.22

Taulukko 10: Hosmer-Lemeshow taulukko vuosina 2018–2020 peruskoulun päättäneiden aineiston mallinnusaineistosta.

6 Tulokset

Määritellään tulosten tarkastelua varten kaksi erillistä populaatiota. Jatkossa vuonna 2021 peruskoulun päättäneiden aineston populaatioa merkitään tunnuksella A, tähän populaatioon kuuluu 57 449 henkilöä. Vuosina 2018–2020 peruskoulun päättäneiden aineiston avulla voitiin ennustaa vuonna 2021 peruskoulun päättäneille todennäköisyydet olla jatkamatta opintoja. Populaatioon B kuuluu 75 kohdehenkilöä, joilla tämä ennustettu todennäköisyys

on suurempi kuin 70 %.

Tarkastellaan ensin oppivelvollisuuslain uudistuksen vaikutusta näissä kahdessa populaatiossa A ja B. Hyödyntämällä kaavaa 3, saadaan populaatiossa A intervention $do(X = 1)$ kausaalivaikutukseksi

$$E(Y = 1 \mid do(X = 1)) = 0.017$$

ja intervention $do(X = 0)$ kausaalivaikutukseksi

$$E(Y = 1 \mid do(X = 0)) = 0.022$$

Keskimääräinen kausaalivaikutus populaatiossa A on

$$\begin{aligned} ACE(A) &= E(Y = 1 \mid do(X = 1)) - E(Y = 1 \mid do(X = 0)) \\ &= 0.0166931 - 0.0223156 = -0.0056225 \end{aligned}$$

Kun tarkastellaan populaatioa B, intervention $do(X = 1)$ kausaalivaikutus on

$$E(Y = 1 \mid do(X = 1)) = 0.658$$

ja intervention $do(X = 0)$ kausaalivaikutus

$$E(Y = 1 \mid do(X = 0)) = 0.880$$

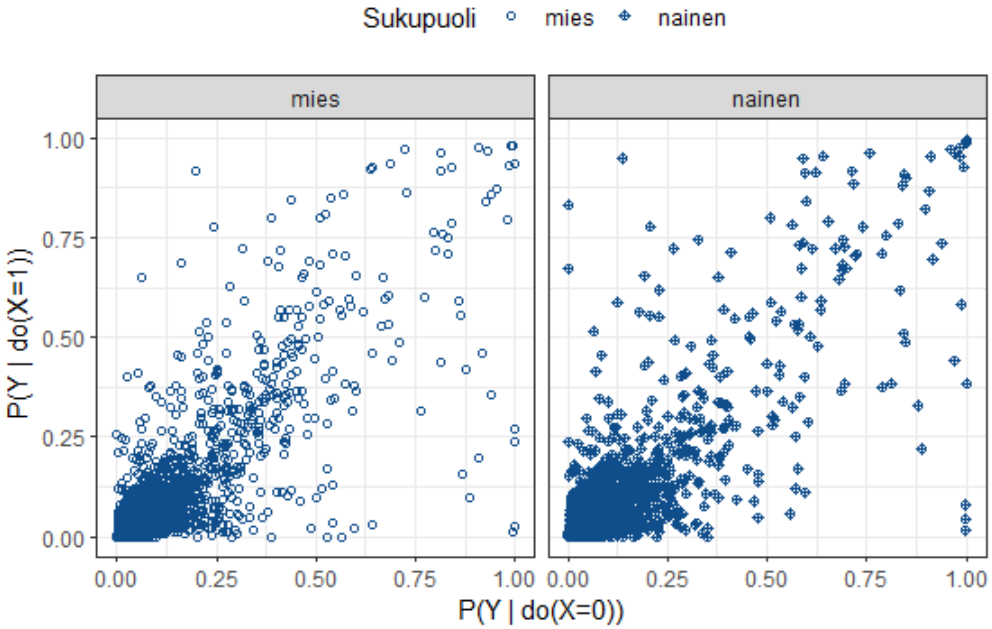
Tällöin keskimääräinen kausaalivaikutus tässä populaatiossa on

$$\begin{aligned} ACE(B) &= E(Y = 1 \mid do(X = 1)) - E(Y = 1 \mid do(X = 0)) \\ &= 0.6584648 - 0.8793603 = -0.2208955 \end{aligned}$$

Populaatiossa A oppivelvollisuuslain uudistamisen keskimmäräinen kausaalivaikutus on -0.006 ja populaatiossa B -0.221. Molemmissa populaatioissa oppivelvollisuuden laajentuminen on pienentänyt todennäköisyyttä olla jatkamatta opintoja. Toisin sanoen opintojen jatkamisen todennäköisyys populaatiossa A on noussut 0.6 prosenttiyksikkö (eli noin 345 henkilöä, 57 449 henkilöstä) ja populaatiossa B 22.1 prosenttiyksikköä (eli noin 15 henkilöä 75 henkilöstä).

Keskimääräistä kausaalivaikutusta voidaan tarkastella myös yli eri osapopulaatioiden. Tarkastellaan keskimääräistä kausaalivaikutusta sukupuolittain ryhmitellyissä osapopulaatioissa A_{nainen} ja A_{mies}. Kuvasta 6.1 voimme nähdä, että molemmissa osapopulaatioissa kausaalivaikutukset

$$P(Y \mid do(X = 1)) \text{ ja } P(Y \mid do(X = 0))$$



Kuva 6.1: Ennustetut kausaalivaikutukset sukupuolittain vuonna 2021 peruskoulun päättäneillä. $P(Y \mid do(X = 1))$ kuvaa kausaalivaikutusta oppivelvollisuuslain uudistuksen tilanteessa ja $P(Y \mid do(X = 0))$ kuvaa kausaalivaikutusta tilanteessa, jossa oppivelvollisuuslaki ei olisi uudistunut.

sijoittuvat likimain samalle $x = y$ suoralle. Lisäksi suurin osa vasteista sijoittuu lähelle arvoa nolla, eli suurimmalla osalla kohdehenkilöistä todennäköisyys olla jatkamatta opintoja on pieni.

Kuvasta 6.1 nähdään, että suuremmilla kausaalivaikutuksilla $P(Y \mid do(X = 0))$, kausaalivaikutuksen $P(Y \mid do(X = 1))$ hajontaa on suhteessa enemmän, kuin pienemmillä kausaalivaikutuksilla $P(Y \mid do(X = 0))$. Kuvasta 6.1 nähdään myös, että osalla kohdehenkilöistä

$$P(Y \mid do(X = 1)) > P(Y \mid do(X = 0)),$$

eli oppivelvollisuuslain uudistus on nostanut todennäköisyyttä olla jatkamatta opintoja. Kuvan perusteella näitä henkilöitä on suhteessa vähemmän, kuin henkilöitä, joilla oppivelvollisuuslain uudistus on laskenut todennäköisyyttä olla jatkamatta opintoja. Tarkastellaan oppivelvollisuuslain uudistuksen vaikutusta sukupuolittain keskimääräisten kausaalivaikutusten avulla.

Sukupuolittain lasketut keskimääräiset kausaalivaikutukset populaatiossa A ovat

$$\begin{aligned} ACE(A_{mies}) &= E(Y = 1 \mid do(X = 1)) - E(Y = 1 \mid do(X = 0)) \\ &= 0.0156058 - 0.0215862 = -0.0059804 \end{aligned}$$

$$\begin{aligned} ACE(A_{nainen}) &= E(Y = 1 \mid do(X = 1)) - E(Y = 1 \mid do(X = 0)) \\ &= 0.0178235 - 0.023074 = -0.0052505 \end{aligned}$$

Eli tarkasteltaessa ainoastaan miehiä, oppivelvollisuuslain uudistuksenn keskimääräinen kausaalivaikutus on 0.60 prosenttiyksikkä ja naisilla vastaavasti 0.53 prosenttiyksikköä. Opintojen jatkamisen todennäköisyys on siis noussut oppivelvollisuuslain myötä sekä naisilla, että miehillä populaatiossa A.

Sukupuolittain lasketut keskimääräiset kausaalivaikutukset populaatiossa B ovat

$$\begin{aligned} ACE(B_{mies}) &= E(Y = 1 \mid do(X = 1)) - E(Y = 1 \mid do(X = 0)) \\ &= 0.6350938 - 0.8814328 = -0.246339 \end{aligned}$$

$$\begin{aligned} ACE(B_{nainen}) &= E(Y = 1 \mid do(X = 1)) - E(Y = 1 \mid do(X = 0)) \\ &= 0.6812208 - 0.8773423 = -0.1961215 \end{aligned}$$

Tässä populaatiossa miehillä lakiuudistuksen keskimääräinen kausaalivaikutus on 24.6 prosenttiyksikköä ja naisilla 19.6 prosenttiyksikköä.

Taulukkossa 11 on keskimääräiset kausaalivaikutukset populaatiossa A vanhempien sosioekonomisen aseman mukaan ryhmiteltynä. Tarkasteltaessa kaikkia vuonna 2021 peruskoulun päättäneitä, oppivelvollisuuslain piiriin kuuluvia nuoria (populaatio A), on hyvä huomata, että kohdehenkilöiden määrä vaihtelee tarkasteluryhmittäin. Äidin sosioekonomisen aseman mukaan tarkasteltaessa suurimmassa ryhmässä on noin 24 000 henkilöä ja pienimmässä hieman alle 1 200 henkilöä.

Taulukossa 12 on keskimääräiset kausaalivaikutukset populaatiossa B. Tässä populaatiossa ryhmien välinen henkilömäärän vaihtelu ei ole yhtä iso kuin populaatiossa A. Jokaisessa ryhmässä on vähintää viisi ja korkeintaan 17 henkilöä. Taulukkoa 12 tarkasteltaessa voimme huomata, että jos äidin tai isän sosioekonominen asema on työntekijä, on oppivelvollisuuslain keskimääräinen kausaalivaikutus vajaa 39 prosenttiyksikköä. Mikä on selvästi suurempi, kuin missään muussa tarkasteluryhmässä.

Kausaalivaikutukset populaatiossa A			
Äidin sosioekonominen asema	$E(Y \mid do(X = 1))$	$E(Y \mid do(X = 0))$	ACE
Yrittäjät	0.0229	0.0247	0.0018
Ylemmät toimihenkilöt	0.0111	0.0142	0.0031
Alemmat toimihenkilöt	0.0134	0.0182	0.0048
Työntekijät	0.0228	0.0302	0.0074
Opiskelijat	0.0188	0.0307	0.0119
Eläkeläiset	0.028	0.0413	0.0133
Muut	0.026	0.0383	0.0123
Tuntematon	0.0329	0.0444	0.0115
Isän sosioekonominen asema	$E(Y \mid do(X = 1))$	$E(Y \mid do(X = 0))$	ACE
Yrittäjät	0.0166	0.0249	0.0083
Ylemmät toimihenkilöt	0.0149	0.0164	0.0015
Alemmat toimihenkilöt	0.013	0.0167	0.0037
Työntekijät	0.0136	0.0206	0.007
Opiskelijat	0.0196	0.0322	0.0126
Eläkeläiset	0.0274	0.0344	0.007
Muut	0.0237	0.0331	0.0094
Tuntematon	0.0278	0.0346	0.0068

Taulukko 11: Keskimääräiset kausaalivaikutukset äidin sosioekonomisen aseman mukaan. ($ACE = E(Y \mid do(X = 0)) - E(Y \mid do(X = 1))$). Populaatioon A kuuluu kaikki vuonna 2021 peruskoulun päättäneet, oppivelvollisuuslain piiriin kuuluvat nuoret.

Kausaalivaikutukset populaatiossa B			
Äidin sosioekonominen asema	$E(Y \mid do(X = 1))$	$E(Y \mid do(X = 0))$	ACE
Yrittäjät	0.7337	0.821	0.0873
Ylemmät toimihenkilöt	0.6987	0.858	0.1593
Alemmat toimihenkilöt	0.579	0.8879	0.3089
Työntekijät	0.5331	0.9225	0.3894
Opiskelijat	.	.	.
Eläkeläiset	.	.	.
Muut	0.7022	0.8812	0.179
Tuntematon	0.7208	0.7208	0
Isän sosioekonominen asema	$E(Y \mid do(X = 1))$	$E(Y \mid do(X = 0))$	ACE
Yrittäjät	0.6485	0.866	0.2175
Ylemmät toimihenkilöt	0.582	0.8297	0.2477
Alemmat toimihenkilöt	0.8779	0.9192	0.0413
Työntekijät	0.5263	0.9133	0.387
Opiskelijat	.	.	.
Eläkeläiset	.	.	.
Muut	0.5871	0.895	0.3079
Tuntematon	0.7953	0.8567	0.0614

Taulukko 12: Keskimääräiset kausaalivaikutukset vanhempien sosioekonomisen aseman mukaan. ($ACE = E(Y \mid do(X = 0)) - E(Y \mid do(X = 1))$). Populaatioon B kuuluu ne henkilöt, joilla todennäköisyys olla jatkamatta opintoja on suurempi kuin 70 %. Populaatiossa B on salattu pisteellä (.) ne ryhmät, joissa kohdehenkilöitä on alle 5.

7 Pohdintaa

Toteutettujen tarkastelujen perusteella vaikuttaisi sille, että oppivelvollisuuslain uudistuksella on ollut vaikutusta vuonna 2021 peruskoulun päättäneiden nuorten opinnoissa jatkamiseen. Uudistus näyttäisi vaikuttaneen etenkin nii-

den nuorten opinnoissa jatkamiseen, jotka historia-aineiston perusteella todennäköisimmin eivät jatka opintoja heti 9. luokan päätyttyä.

Tutkimustilanne on luonnollinen koe, niinpä havaintoihin ei ole pystytty vaikuttamaan. Vaikka käytössä on ollut rekisteriaineistoista koottu tutkimusaineisto, voi aineistossa siitä huolimatta olla puutteita. Lisäksi taustalla voi olla latentteja muuttujia, jotka eivät ole tiedossa, mutta ovat voineet vaikuttaa tuloksiin. Koulutuksessa jatkamattomien osuudet tutkimusaineistossa vaihtelevat tarkasteluvuosien välillä 0.2–1.0 prosenttiyksikköä (taulukko 1). Koulutuksessa jatkamattomien osuudet ovat likimain samat vuosina 2018–2019 peruskoulun päättäneiden osalta. Jatkamattomien osuus laski kuitenkin jo vuonna 2020, eli ennen oppivelvolisuuslain uudistuksen voimaan astumista.

Vuosina 2020–2021 on eletty poikkeuksellisia aikoja koronan vuoksi. Koronan tiedetään vaikuttaneen nuorten hyvinvointiin, mutta koronan vaikutusta nuorten opinnoissa jatkamiseen ei pystytä tässä tutkielmassa huomioimaan, vaan se on latenttina taustamuuttujana. Koronan vaikutus näkyy todennäköisesti sekä vuonna 2020 peruskoulun päättäneille, että vuonna 2021 peruskoulun päättäneillä. Niinpä koronan vaikutus ei kohdistu ainoastaan kiinnostuksen kohteena olevaan perusjoukkoon.

Tutkimusaineisto on neljän vuoden ajalta, eikä ajan vaikutusta tutkimustulokseen voida täysin pois sulkea. Ajan vaikutus on pyritty huomioimaan sovittamalla kaksi tilastollista mallia peruskoulun päättövuoden mukaan. Toinen lähestymistapa olisi ollut käsitellä yhtä aineistoa, jossa oppivelvollisuuslaki on omana kaksiarvoisena muuttujanaan. Tälle aineistolle olisi voitu sovittaa yksi, kausaalirakenteen mukainen tilastollinen malli.

Tutkimuksessa tarkastellaan suhteellisen harvinaista tapahtumaa. Ainoastaan 1.5–2.5 % nuorista vuosittain ei jatkanut opintoja heti yhdeksännen luokan päätyttyä. Tämä on tuonut omat haasteensa kausaalipäättelyyn, erityisesti kausaalivaikutuksen estimointiin. Mallin valinnnassa on pyritty hyödyntämään erilaisia menetelmiä, jotta kausaalivaikutukset saataisiin ennustettua mahdollisimman tarkasti. Käytössä on ollut laaja aineisto, mikä on mahdollistanut erilaisten muuttujavariaatioiden hyödyntämisen ja testaamisen.

Mallia valitessa erilaisten muuttujavariaatioiden testaaminen ja mallin hyvyystarkastelut toteutettiin vuosina 2018–2020 peruskoulun päättäneiden aineistolla. Näiden tarkastelujen pohjalta valittiin lopullisen mallin selittäjät ja vastaavilla selittäjillä sovitettiin logistinen regressiomalli vuonna 2021 peruskoulun päättäneiden aineistolla. Vuoden 2021 mallissa regressiokertoimien keskivirheet olivat muutamien muuttujien suhteen poikkeuksellisen suuret. Tällaisia muuttujia olivat esimerkiksi lähtökoulun maakunta -muuttuja sekä lähtökoulun maakunnan ja kuntaryhmän interaktiotermi. Vaikuttaisi sille,

ettei vuoden 2021 peruskoulun päättäneiden aineistolla sovitetussa mallissa kaikki kertoimet estimoidu kunnolla.

Vuoden 2021 aineistossa oli kohdehenkilöitä 1/3 vuosien 2018–2020 aineistoon verrattuna. Lisäksi mallinnettava tapahtuma "ei jatkanut opintoja"on harvinainen, niinpä joidenkin luokkien havaintojen määrät jäivät pieniksi. Tällaisia luokkia olivat esimerkiksi lähtökoulun maakunta -muuttujassa Kymenlaakso, Kainuu ja Keski-Pohjanmaa (lahtokoulu_onuts3 arvot 9, 16, 18). Jos vuoden 2021 aineistolle sovitetun mallin muuttujien tarkasteluja olisi tehty aikaisemmassa vaiheessa, olisi mallin muuttujien pieniä luokkia voitu yhdistää tai mallin muuttujia vaihtaa. Suuret keskivirheet huomattiin vasta työn loppuvaiheessa, eikä mallia ollut aikataulullisesti enää mahdollista muokata.

Tässä työssä tarkastellaan opintojen jatkamista siitä näkökulmasta, onko nuori toisen asteen oppilaitoksessa kirjoilla 20.9 peruskoulun päättövuonna. Vaikka nuori olisi ollut kirjoilla toisen asteen oppilaitoksessa 20.9, ei se tarkoita että nuori lopulta suorittaisi kyseisen koulutuksen. Tutkimusta olisikin mielenkiintoista laajentaa, tarkastelemalla vuonna 2021 peruskoulun päättäneiden tilannetta sinä vuonna, kun kohdehenkilöt täyttävät 18 vuotta. Kuinka moni 18-vuotta täyttäneistä, oppivelvollisuuden piiriin kuuluvista henkilöistä on kyseisenä vuonna edelleen suorittamassa toisen asteen tutkintoa?

Toisen asteen tutkinto ei useinkaan ole suoritettuna vielä 18-vuotiaana, joten voisi olla myös mielenkiintoista tutkia, kuinka monella 2021 peruskoulun päättäneistä oppivelvollisuuden piiriin kuuluvista nuorista on suoritettuna toisen asteen tutkinto 20-vuotiaana. 2021 peruskoulun päättäneistä nuorista suurin osa täyttää 18-vuotta vuonna 2024 ja 20-vuotta vuonna 2026, niinpä näitä asioita ei vielä käytettävissä olevalla aineistolla kyetty tutkimaan.

Viitteet

S Ahonen. Kiistelty oppivelvollisuus. *Koulu ja menneisyys*, 58:8–37, 2021.

A. Gelman and J. Hill. Data analysis using regression and multilevel/hierarchical models. *Canbridge Univercity*, 2007.

P. W. Holland. Statistics and causal inference. *Journal of American Statistical Association*, 81(396):945–960, 1986.

P. W. Holland. Causal inference, path analysis, and recursive structural equation models. *Sociological methodology*, 18:449–84, 1988.

W. Hosmer, Jr. David, Stanley Lemeshow, Rodney X Sturdivant, et al. *Applied logistic regression (Vol. 398)*. John Wiley & Sons Hoboken, NJ, 2013.

A Kalenius. Suomalaisten koulutusrakenteen kehitys 1970-2090. *Opetus- ja kulttuuriministeriön julkaisuja 2014:1*, 2014.

M. Keski-Petäjä and M. Witting. Vanhempien koulutus vaikuttaa lasten valintoihin. *Tieto & trendt. Tilastokeskus*, 2016.

OKM. Oppivelvollisuuden laajentaminen., 2021a. URL https://okm.fi/oppivelvollisuuden-laajentaminen. Haettu 1.1.2023.

OKM. Kysymyksiä ja vastauksia oppivelvollisuudesta, 2021b. URL https://okm.fi/kysymyksia-ja-vastauksia-oppivelvollisuudesta. Haettu 1.1.2023.

Opetushallitus. Perustietoa oppivelvollisuuden laajentamisesta„ 2021. URL https://www.oph.fi/fi/kehittaminen/perustietoa-oppivelvollisuuden-laajentamisesta. Haettu 1.1.2023.

Oppivelvollisuuslaki. Suomen valtio, 1214/2020. URL https://finlex.fi/fi/laki/alkup/2020/20201214. Annettu Helsingissä 30.12.2020, Haettu 5.4.2023.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, second edition, 2009.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Neal J Roese. Counterfactual thinking. *Psychological bulletin*, 121(1):133, 1997.

S. Ruohola. Äidiltä tyttärelle. *Koulutuskulttuurisia siirtymiä neljässä sukupolvessa. Turun yliopiston julkaisuja C342.*, 2012.

C. R. Shalizi. Advanced data analysis from an elementary point of view. *Citeseer*, 2013.

Santtu Tikka, Jouni Helske, and Juha Karvanen. Clustering and structural robustness in causal diagrams. *arXiv preprint arXiv:2111.04513*, 2021.

Tilastokeskus. Johdatus tilastotieteeseen: 2.2.5 rekisteriaineistot, [viitattu 13.2.2023], 2023.

Suomen virallinen tilasto (SVT). Koulutukseen hakeutuminen [verkkojulkaisu], 2020. URL http://www.stat.fi/til/khak/2019/khak_2019_2020-12-10_tie_001_fi.html.

M. Witting. Lukio, amis vai pelkkä peruskoulu? -perusopetuksen jälkeisillä valinnoilla on usein kauaskantoiset vaikutukset. *Lapset Suomessa. Tieto & trendt. Tilastokeskus*, 2021.

Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295, 1986.

Liitteet

Aineiston muuttujat ja niiden kuvailua

Mallien regressiokertoimet keskivirheineen

Muuttujan lyhenne	Muuttujan selite	Muuttujan luokat
aikieli	Äidinkieli	Kielet joita äidinkielenään puhuu alle 50 aineiston henkilöä karkeistettu samaan luokkaan. Muut kielet muodossa fi, sv, ru
askun	Asuinkunta	Tarkalla 3-numerotasolla
asuminen	Tieto siitä, asuvatko nuoren vanhemmat samassa kunnassa	1 ei tiedossa, 2 samassa kunnassa, 3 eri kunnassa
ika	Kohdehenkilön ikä peruskoulun päättövuonna	14–17
kansal	Kansalaisuus	Kaikilla koodilla 246, Suomen kansalainen
kuntaryh	Kunnan tyyppi:	Maaseutumaiset kunnat, Kaupunkimaiset kunnat Taajaan asutut kunnat
luokka	Vuosiluokka	Kaikilla 9. luokka
muoto	Tieto siitä, asuuko nuori kumman vanhemman kanssa	1 ei tiedossa, 2 asuu äidin kanssa, 3 asuu isän kanssa, 4 asuu molempien vanhempiensa kanssa
oppivelvollinen	Onko henkilö oppivelvollinen	1 jos kuuluu oppivelvollisuuslain uudistuksen piiriin, 0 muuten
sijainti	Vastaanotetun opiskelupaikan sijainti suhteessa nuoren asuinkuntaan peruskoulun päättyessä	1 sama, 0 eri
sp	Sukupuoli	1 nainen, 2 mies
suuralue	Asuinpaikan suuralue:	1 Helsinki, Uusimaa, 2 Etelä-Suomi 3 Länsi-Suomi 4 Pohjois- ja Itä-Suomi
svaltio	Syntymävaltio, karkeistettu	fi Suomi, sv Ruotsi, ru Venäjä, 98 muut

Taulukko 13: Kohdehenkilöä kuvaavia muuttujia

Muuttujan lyhenne	Muuttujan selite	Muuttujan luokat
syntv	Syntymävuosi	Tarkalla vuosiluku tasolla
syntyp2	Syntyperä:	11 Suomalaistaustainen, syntynyt Suomessa, 12 Suomalaistaustainen, syntynyt ulkomailla, 22 Ulkomaalaistaustainen, syntynyt ulkomailla, 21 Ulkomaalaistaustainen, syntynyt Suomessa
A1	A1-kielen arvosana	Numeerisena tai kirjainkoodina
AI	Äidinkielen arvosana	Kaikilla oppiaineilla mahdolliset arvot:
B1	B1-kielen arvosana	4-10
BI	Biologian arvosana	S eli suoritettu
FI	Filosofian arvosana	H eli hylätty
FY	Fysiikan arvosana	
GE	Maantiedon arvosana	
HI	Historian arvosana	
KE	Kemian arvosana	
KO	Kotitalouden arvosana	
KS	Käsitöiden arvosana	
KT	Uskonnon/elämänkatsomustiedon arvosana	
KU	Kuvaamataidon arvosana	
LI	Liikunnan arvosana	
MA	Matematiikan arvosana	
MU	Musiikin arvosana	
OP	Opinto-ohjauksen arvosana	
TE	Terveysteidon arvosana	
YH	Yhteiskuntaopin arvosana	
YL	Ympäristö- ja luonnontieto	

Taulukko 14: Kohdehenkilöä kuvaavia muuttujia

Muuttujan lyhenne	Muuttujan selite	Muuttujan luokat
amas1	Ammattiasema	1 palkansaaja, 2 yrittäjä
ika	Ikä	Karkeistettu 10 vuoden tarkkuudella
kan1sapv	Suomen kansalaisuuden saantivuosi	1 ennen vuotta 1990, 2 vuosina –2000, 3 vuosina 2001–2005, 4 vuosina 2006–2010 , 2011 alkaen tarkalla tasolla 0000 jos syntyjään Suomen kansalainen
kansa1	Kansalaisuus	fi Suomi, es Viro, sv Ruotsi, xx muut
kieli	Äidinkieli	fi Suomi, ru Venäjä, sv Ruotsi, xx muut
ptoim1	Pääasiallinen toiminta,	11 Työlliset, 12 Työttömät 21 0–14 -vuotiaat 22 Opiskelijat, koululaiset 24 Eläkeläiset (pl. työttömyyseläkeläiset) 25 Varusmiehet, siviilipalvelusmiehet 29 Työttömyyseläkeläiset 99 Muut työvoiman ulkopuolella olevat
sivs	Siviilisääty	1 Naimaton 2 Aviossa tai rekisteröidyssä parisuhteessa 4 Eronnut, 5 Leski
sose	Sosioekoniminen asema	1 Yrittäjät 3 Ylemmät toimihenkilöt 4 Alemmat toimihenkilöt 5 Työntekijät, 6 Opiskelijat 7 Eläkeläiset, 8 Muut, 9 Tuntematon
svaltio	Syntymävaltio	fi Suomi, sv Ruotsi, ru Venäjä, 98 muut
syntypy2	Syntyperä	11 Suomalaistaustainen, syntynyt Suomessa 12 Suomalaistaustainen, syntynyt ulkomailla 22 Ulkomaalaistaustainen, syntynyt ulkomailla 21 Ulkomaalaistaustainen, syntynyt Suomessa
tyol	Työllisyystilanna	1 Henkilö ollut työvoimapoliittisessa koulutuksessa, 0 muuten

Taulukko 15: Vanhempia kuvaavat muuttujat. Jokainen muuttuja aineistossa erikseen kummallekin vanhemmale.

Muuttujan lyhenne	Muuttujan selite	Muuttujan luokat
lahtokoulu jarj	Järjestäjä tunnus	Yksilöivä tunnus
lahtokoulu okieli	Opetuskieli	Suomi, Ruotsi, Englanti, Suomi ja Ruotsi, Jokin muu
lahtokoulu oltyp	Oppilaitostyyppi	Peruskoulu, Perus- ja lukioasteen koulu, Peruskouluasteen erityiskoulu
lahtokoulu omist	Lähtökoulun omistaja	Yksityinen, Kunta
lahtokoulu onimi	Oppilaitoksen nimi	Yksilöivä nimi
lahtokoulu onuts3	Sijainti maakunta	01 Uusimaa 02 Varsinais-Suomi 04 Satakunta 05 Kanta-Häme 06 Pirkanmaa 07 Päijät-Häme 08 Kymenlaakso 09 Etelä-Karjala 10 Etelä-Savo 11 Pohjois-Savo 12 Pohjois-Karjala 13 Keski-Suomi 14 Etelä-Pohjanmaa 15 Pohjanmaa 16 Keski-Pohjanmaa 17 Pohjois-Pohjanmaa 18 Kainuu 19 Lappi 99 Tuntematon
lahtokoulun oppilaitos koodi	Oppilaitoskoodi	Yksilöivä koodi
lahtokoulutkkun	Sijainti kunta	Tarkalla 3-numero tasolla

Taulukko 16: Lähtökoulua, eli nuoren 9-luokan suorituskoulua, kuvaavat muuttujat

Muuttujan lyhenne	Muuttujan selite
vuosi	Peruskoulun päättövuosi
sij	Numeerinen sijoittumistieto, johon yhdistetty alla kuvatut dummy muuttujat omina koodeinaan
amm	Opiskelee ammatillisessa oppilaitoksessa
lukio	Opiskelee lukiossa
kymp	Opiskelee 10. luokalla
valm	Opiskelee valmentavassa koulutuksesssa
eiopis	Ei opiskele missän yllä mainituista koulutuksista

Taulukko 17: Muut muuttujat: Peruskoulun päättövuosi, Yhdistetty sijoittumistieto muuttuja sekä dummy muuttujat, jotka kuvaavat nuoren tilannetta 20.9 peruskoulun päättövuonna

		Malli 2018–2020		Malli 2021		Erotus:
Selittäjä	Arvo	β_{1i}	SE_1	β_{2i}	SE_2	$\beta_{2i} - \beta_{1i}$
Intercept		-1.62	57.03	-4.17	52.95	**-2.55**
sp	1	-0.22	0.02	-0.21	0.04	0.00
ika	14	0.80	0.14	0.57	0.28	-0.23
ika	15	-0.97	0.06	-0.93	0.12	0.04
ika	16	-0.41	0.07	-0.17	0.13	0.25
lk oltyp	11	-0.18	0.05	-0.07	0.10	0.11
lk oltyp	12	0.56	0.08	0.55	0.14	-0.02
lk okieli	1	0.03	0.20	0.52	0.33	0.49
lk okieli	2	-0.54	0.22	0.36	0.37	0.90
lk okieli	3	0.02	0.60	1.91	0.64	1.89
lk okieli	4	1.25	0.28	1.76	0.43	0.51
kuntaryh	1	0.09	3.61	0.41	50.31	0.33
kuntaryh	2	-0.09	5.56	0.40	73.65	0.49
lk onuts3	1	0.19	3.61	0.77	50.14	0.58
lk onuts3	2	-0.23	3.61	1.08	50.14	1.31
lk onuts3	4	-0.08	3.61	0.14	50.14	0.22
lk onuts3	5	-0.68	3.61	0.60	50.14	1.28
lk onuts3	6	0.01	3.61	0.57	50.14	0.56
lk onuts3	7	0.27	3.61	1.35	50.14	1.07
lk onuts3	8	0.62	3.61	-1.88	87.11	**-2.50**
lk onuts3	9	0.59	3.62	-2.86	558.10	**-3.45**
lk onuts3	10	-0.03	3.61	1.62	50.14	1.66
lk onuts3	11	-0.11	3.61	0.50	50.14	0.61
lk onuts3	12	-0.34	3.61	0.64	50.14	0.98
lk onuts3	13	0.58	3.61	1.86	50.14	1.28
lk onuts3	14	0.23	3.61	0.93	50.14	0.70
lk onuts3	15	0.69	3.61	0.55	50.14	-0.14
lk onuts3	16	-2.20	44.12	2.73	50.14	**4.92**
lk onuts3	17	1.01	3.61	2.03	50.14	1.02
lk onuts3	18	-0.14	3.61	-10.62	121.40	**-10.48**
lk onuts3	19	0.49	3.61	1.03	50.14	0.53

Taulukko 18: Regressiokertoimet ja keskivirheet vuosien 2018–2020 ja vuoden 2021 logistisille regressiomalleille, sekä regressiokertoimien erotus mallien välillä. (lk=lähtökoulu)

		Malli 2018–2020		Malli 2021		Erotus:
Selittäjä	Arvo	β_{1i}	SE_1	β_{2i}	SE_2	$\beta_{2i} - \beta_{1i}$
sose isa	10	0.26	0.13	0.33	16.99	0.07
sose isa	20	0.17	0.07	0.62	16.99	0.45
sose isa	31	0.28	0.10	0.88	16.99	0.60
sose isa	32	0.11	0.10	0.72	16.99	0.61
sose isa	33	0.46	0.12	1.33	16.99	0.87
sose isa	34	0.10	0.10	0.90	16.99	0.80
sose isa	41	-0.11	0.15	0.61	16.99	0.71
sose isa	42	-0.08	0.11	0.56	16.99	0.64
sose isa	43	-0.82	0.70	-10.62	288.80	**-9.81**
sose isa	44	-0.03	0.09	0.58	16.99	0.61
sose isa	51	-0.32	0.22	0.54	16.99	0.85
sose isa	52	-0.16	0.07	0.38	16.99	0.54
sose isa	53	-0.26	0.11	0.36	16.99	0.62
sose isa	54	-0.20	0.08	0.13	16.99	0.33
sose isa	60	0.31	0.17	0.57	16.99	0.27
sose isa	70	0.10	0.09	0.75	16.99	0.66
sose isa	81	0.13	0.08	0.70	16.99	0.57
sose aiti	10	-0.19	0.18	0.44	0.32	0.64
sose aiti	20	0.22	0.08	0.57	0.13	0.35
sose aiti	31	0.10	0.14	-0.44	0.29	-0.53
sose aiti	32	-0.17	0.13	-0.32	0.23	-0.15
sose aiti	33	-0.14	0.09	0.11	0.14	0.24
sose aiti	34	-0.24	0.10	-0.44	0.18	-0.20
sose aiti	41	-0.33	0.16	-0.45	0.33	-0.13
sose aiti	42	-0.14	0.07	-0.12	0.14	0.01
sose aiti	43	0.05	0.21	-0.03	0.44	-0.09
sose aiti	44	-0.10	0.05	0.02	0.09	0.12
sose aiti	51	-0.13	0.29	-0.37	0.68	-0.24
sose aiti	52	-0.12	0.12	-0.02	0.25	0.09
sose aiti	53	-0.11	0.13	-0.16	0.25	-0.05
sose aiti	54	0.17	0.07	0.34	0.13	0.18
sose aiti	60	0.19	0.11	0.06	0.22	-0.12
sose aiti	70	0.31	0.10	0.23	0.20	-0.09
sose aiti	81	0.31	0.07	0.26	0.12	-0.04

Taulukko 19: Regressiokertoimet ja keskivirheet vuosien 2018–2020 ja vuoden 2021 logistisille regressiomalleille, sekä regressiokertoimien erotus mallien välillä.

Selittäjä	Arvo	Malli 2018–2020 β_{1i}	SE_1	Malli 2021 β_{2i}	SE_2	Erotus: $\beta_{2i} - \beta_{1i}$
syntyp2	11	0.10	0.07	-0.26	0.10	-0.10
syntyp2	12	-0.10	0.13	0.34	0.18	0.44
syntyp2	21	-0.05	0.11	-0.27	0.18	-0.22
A1	10	-0.92	0.14	-0.55	0.17	0.36
A1	4	0.99	0.79	0.41	0.38	-0.58
A1	5	0.17	0.14	0.36	0.16	0.19
A1	6	0.07	0.12	0.29	0.13	0.23
A1	7	-0.03	0.12	0.07	0.12	0.10
A1	8	-0.15	0.12	0.17	0.12	0.32
A1	9	-0.33	0.12	-0.08	0.13	0.25
A1	S	-0.68	0.33	-0.21	0.24	0.48
FY	10	-0.39	0.15	0.47	0.20	0.85
FY	4	1.21	0.54	0.93	0.36	-0.28
FY	5	0.14	0.12	0.25	0.16	0.11
FY	6	0.01	0.11	0.03	0.14	0.02
FY	7	-0.18	0.11	0.00	0.14	0.18
FY	8	-0.35	0.11	-0.03	0.14	0.32
FY	9	-0.38	0.12	0.21	0.16	0.60
FY	S	0.87	0.40	-0.70	0.32	-1.57
TE	10	-0.57	0.16	-0.86	0.19	-0.29
TE	4	2.88	0.83	1.58	0.38	-1.30
TE	5	-0.15	0.16	0.31	0.19	0.46
TE	6	-0.26	0.14	-0.15	0.15	0.11
TE	7	-0.26	0.14	-0.35	0.14	-0.09
TE	8	-0.41	0.14	-0.39	0.14	0.02
TE	9	-0.58	0.14	-0.46	0.15	0.12
TE	S	0.56	0.38	-0.02	0.31	-0.58
YH	10	-0.71	0.14	-0.95	0.17	-0.24
YH	4	1.35	0.64	2.03	0.40	0.67
YH	5	-0.30	0.13	-1.01	0.17	-0.71
YH	6	-0.49	0.11	-0.98	0.12	-0.49
YH	7	-0.63	0.11	-1.02	0.11	-0.39
YH	8	-0.69	0.11	-1.05	0.11	-0.36
YH	9	-0.75	0.12	-1.14	0.13	-0.40
YH	S	-0.30	0.34	0.74	0.21	1.03

Taulukko 20: Regressiokertoimet ja keskivirheet vuosien 2018–2020 ja vuoden 2021 logistisille regressiomalleille, sekä regressiokertoimien erotus mallien välillä.

		Malli 2018–2020		Malli 2021		Erotus:
Selittäjä	Arvo	β_{1i}	SE_1	β_{2i}	SE_2	$\beta_{2i} - \beta_{1i}$
KO	10	0.33	0.21	-0.12	0.17	-0.46
KO	4	-0.98	1.51	-0.13	0.57	0.85
KO	5	0.76	0.23	0.15	0.22	-0.62
KO	6	0.25	0.21	0.33	0.15	0.08
KO	7	0.20	0.20	-0.16	0.13	-0.36
KO	8	0.13	0.20	-0.35	0.12	-0.48
KO	9	0.04	0.20	-0.44	0.14	-0.48
KO	S	-0.25	0.28	-0.06	0.17	0.18
MU	10	-1.10	56.91	0.37	0.19	1.47
MU	4	8.91	455.30	-0.06	1.00	**-8.97**
MU	5	-1.39	56.91	0.05	0.24	1.44
MU	6	-1.33	56.91	-0.17	0.19	1.15
MU	7	-1.35	56.91	-0.25	0.16	1.10
MU	8	-1.39	56.91	-0.13	0.16	1.26
MU	9	-1.47	56.91	-0.02	0.17	1.45
MU	S	-0.53	56.91	0.34	0.21	0.87
KS	10	0.18	0.19	0.29	0.18	0.11
KS	4	-1.03	1.28	0.32	0.67	1.35
KS	5	0.49	0.21	0.26	0.21	-0.23
KS	6	0.28	0.18	-0.09	0.16	-0.37
KS	7	0.00	0.17	-0.29	0.14	-0.30
KS	8	-0.26	0.17	-0.27	0.13	-0.01
KS	9	-0.08	0.17	-0.08	0.14	0.00
KS	S	-0.01	0.27	0.16	0.17	0.17
LI	10	-0.47	0.10	-0.23	0.15	0.23
LI	4	0.62	0.53	0.92	0.37	0.30
LI	5	0.79	0.11	0.97	0.16	0.18
LI	6	0.17	0.10	0.42	0.14	0.25
LI	7	-0.20	0.09	-0.18	0.12	0.02
LI	8	-0.68	0.09	-0.49	0.12	0.19
LI	9	-0.55	0.09	-0.49	0.12	0.06
LI	S	-0.17	0.18	-0.53	0.17	-0.36
ka		0.19	0.03	0.10	0.07	-0.10
kuntaryh x lko	1x1	-0.31	3.61	-0.57	50.31	-0.26
kuntaryh x lko	1x2	0.09	3.61	-0.27	50.31	-0.36
kuntaryh x lko	1x4	-0.18	3.61	-0.05	50.31	0.13

Taulukko 21: Regressiokertoimet ja keskivirheet vuosien 2018–2020 ja vuoden 2021 logistisille regressiomalleille, sekä regressiokertoimien erotus mallien välillä. (lko=lähtökoulu onuts3 -muuttuja)

		Malli 2018–2020		Malli 2021		Erotus:
Selittäjä	Arvo	β_{1i}	SE_1	β_{2i}	SE_2	$\beta_{2i} - \beta_{1i}$
kuntaryh x lko	1x5	-0.12	3.61	-0.53	50.31	-0.41
kuntaryh x lko	1x6	-0.20	3.61	-0.54	50.31	-0.35
kuntaryh x lko	1x7	-0.32	3.61	-0.90	50.31	-0.58
kuntaryh x lko	1x8	-0.58	3.62	2.42	87.21	**3.00**
kuntaryh x lko	1x9	-0.69	3.62	3.54	558.10	**4.23**
kuntaryh x lko	1x10	0.10	3.61	-0.74	50.31	-0.84
kuntaryh x lko	1x11	-0.14	3.61	-0.63	50.31	-0.49
kuntaryh x lko	1x12	-0.37	3.61	-0.62	50.31	-0.25
kuntaryh x lko	1x13	-0.30	3.61	-0.82	50.31	-0.53
kuntaryh x lko	1x14	0.28	3.61	-0.48	50.31	-0.76
kuntaryh x lko	1x15	-0.33	3.61	-0.27	50.31	0.07
kuntaryh x lko	1x16	2.25	44.12	-1.88	50.31	**-4.13**
kuntaryh x lko	1x17	-0.47	3.61	-0.96	50.31	-0.49
kuntaryh x lko	1x18	-1.11	3.62	-0.42	142.40	0.69
kuntaryh x lko	1x19	-0.34	3.61	-0.16	50.31	0.18
kuntaryh x lko	2x1	0.03	5.57	-0.21	73.66	-0.24
kuntaryh x lko	2x2	-0.02	5.57	-0.58	73.66	-0.56
kuntaryh x lko	2x4	0.09	5.57	-0.99	73.66	-1.08
kuntaryh x lko	2x5	0.13	5.57	-0.06	73.66	-0.19
kuntaryh x lko	2x6	0.15	5.57	0.03	73.66	-0.12
kuntaryh x lko	2x7	0.27	5.57	0.15	73.66	-0.13
kuntaryh x lko	2x8	1.03	5.59	5.44	102.50	**4.41**
kuntaryh x lko	2x9	0.73	5.61	-7.26	1114.00	**-7.99**
kuntaryh x lko	2x10	-0.37	5.57	0.13	73.66	0.49
kuntaryh x lko	2x11	-0.18	5.57	-0.97	73.66	-0.79
kuntaryh x lko	2x12	-0.23	5.57	-0.16	73.66	0.07
kuntaryh x lko	2x13	0.28	5.57	0.03	73.66	-0.25
kuntaryh x lko	2x14	-0.19	5.57	-0.20	73.66	-0.02
kuntaryh x lko	2x15	0.06	5.57	-0.13	73.66	-0.19
kuntaryh x lko	2x16	-5.50	88.13	2.25	73.66	**7.75**
kuntaryh x lko	2x17	0.01	5.56	-0.46	73.65	-0.48
kuntaryh x lko	2x18	0.25	5.57	-0.21	176.30	-0.46
kuntaryh x lko	2x19	-0.54	5.57	-1.49	73.66	-0.96

Taulukko 22: Regressiokertoimet ja keskivirheet vuosien 2018–2020 ja vuoden 2021 logistisille regressiomalleille, sekä regressiokertoimien erotus mallien välillä. (lko=lähtökoulu onuts3 -muuttuja)