VATT-TUTKIMUKSIA

22

VATT-RESEARCH REPORTS

Marjo Pyy

NUORTEN TYÖLLISTYMISEN KUVAAMINEN ELINAIKA-ANALYYSIN MENETELMIN

VALTION TALOUDELLINEN TUTKIMUSKESKUS

Government Institute for Economic Research

Helsinki 1994

ISBN 951-561-106-7
ISSN 0788-5008

Valtion taloudellinen tutkimuskeskus
Government Institute for Economic Research
Hämeentie 3, 00530 Helsinki, Finland

J-Paino Ky
Helsinki 1994

Esipuhe

Työttömyyden yleisin mittari, työttömyysaste, riippuu sekä työttömäksi tulevien virrasta että työttömyysjaksojen pituudesta. Yksilön hyvinvoinnin kannalta on varmasti tärkeämpää se, kuinka kauan mahdollinen työttömyys kestää kuin se, joutuuko hän ylipäänsä työttömäksi vai ei. Nuorille tyypillisiä ovat lyhyet työttömyysjaksot, jotka liittyvät usein opiskelemasta työelämään siirtymiseen tai työpaikan vaihtamiseen. Tämänkaltainen työttömyys on normaalia yksilön elämäntilanteen muuttumiseen liittyvää kitkatyöttömyyttä. Työttömyyden pitkittyminen voi sen sijaan johtaa työelämästä syrjäytymiseen etenkin nuorilla, joilla ei ole vielä vakiintunutta asemaa työmarkkinoilla. Tämän vuoksi on tärkeää tutkia, millaiset henkilökohtaiset ominaisuudet ja ulkoiset olosuhteet johtavat työttömyyden pitkittymiseen nuorilla.

Työttömyyden kestoa voidaan analysoida elinaika-analyysin (duraatioanalyysi) menetelmin. Yleisin työttömyyden keston kuvaamiseen käytetty elinaikamalli on ns. täysin parametroitu regressiomalli. Viime aikoina mielenkiinnon kohteina ovat olleet ns. semiparametriset mallit, joissa ei tarvitse tehdä oletusta, toisin kuin täysin parametroiduissa malleissa, työttömyyden keston jakaumasta. Tutkimuksessa sovellettu semiparametrinen malli ei rajoita työllistymisintensiteetin riippuvuutta työttömyyden kestosta. Tämä on tärkeää, koska työllistymisintensiteetin ajallisella käyttäytymisellä on merkitystä mm. työttömyysturvajärjestelmän suunnittelussa.

Marjo Pyyn tutkimus liittyy Valtion taloudellisen tutkimuskeskuksen käynnistämään työmarkkinat muutoksessa tutkimusalueeseen. Nuorten saaminen mahdollisimman joustavasti työelämän piiriin on tärkeää, ei ainoastaan yksilön henkisen kehityksen, vaan myös talouden tuotantopotentiaalin kasvun näkökulmasta. Marjo Pyyn tutkimus on merkittävä sekä tutkimusaiheen keskeisyyden että uusien tutkimusmenetelmien soveltamisen näkökulmasta. Haluan kiittää tekijää aiheen ennakkoluulottomasta ja menetelmällisesti kestävästä käsittelystä.

Helsingissä marraskuussa1994

Seppo Leppänen

MARJO PYY: NUORTEN TYÖLLISTYMISEN KUVAAMINEN ELINAIKA-ANALYYSIN MENETELMIN. Helsinki, VATT, Valtion taloudellinen tutkimuskeskus, Government Institute for Economic Research, 1994. (B, ISSN 0788-5008, No 22) ISBN 951-106-7.

TIIVISTELMÄ: Tutkimuksessa perehdytään työttömyyden keston kuvaamisessa yleisesti käytettyyn menetelmään: elinaika- eli duraatioanalyysiin ja sovelletaan sitä nuorten työttömyyden keston mallintamiseen. Duraatioanalyysin menetelmin voidaan tarkastella erilaisten tekijöiden vaikutusta työllistymistodennäköisyyteen sekä tutkia työllistymistodennäköisyyden riippuvuutta työttömyyden kestosta. Tutkimuksen teoriaosassa perehdytään sekä täysin parametroituihin (jossa kesto on oletettu Weibull-jakautuneeksi) että ns. semiparametrisiin elinaikamalleihin, joissa työttömyyden keston jakaumaa ei tarvitse lainkaan spesifioida. Tutkimuksen empiirisessä osassa tarkastellaan suomalaisten alle 30-vuotiaiden nuorten työllistymistodennäköisyyteen vaikuttavia tekijöitä semiparametrisella Coxin mallilla. Estimoitu työttömyyden keston hasardifunktio on laskeva, eli työllistymistodennäköisyys pienenee työttömyyden keston kasvaessa. Nuorten työllistymistodennäköisyyteen vaikuttaa mm. sukupuoli, koulutusaste ja -ala, työnhaun alkamisajankohta sekä työnhakua edeltävä toiminta. Työllisyyskoulutuksessa olleiden nuorten työllistymistodennäköisyys ei ole tilastollisesti merkittävästi vertailuryhmän työllistymistodennäköisyyttä pienempi. Jos työllisyyskoulutukseen on pyritty valitsemaan nimenomaan heikosti työllistyviä työnhakijoita, voi tuloksen tulkita siten, että työllisyyskoulutus on pystynyt parantamaan koulutettujen asemaa työmarkkinoilla.

ASIASANAT: nuorten työttömyyden kesto, duraatioanalyysi, Coxin malli

MARJO PYY: NUORTEN TYÖLLISTYMISEN KUVAAMINEN ELINAIKA-ANALYYSIN MENETELMIN. Helsinki, VATT, Valtion taloudellinen tutkimuskeskus, Government Institute for Economic Research, 1994. (B, ISSN 0788-5008, No 22) ISBN 951-106-7.

ABSTRACT: This study concerns failure time or duration analysis, which is a relatively new technique in the analysis of the duration of unemployment. Duration analysis enables one to study the effects of different variables on the probability of getting a job and to study how this probability depends on the duration of unemployment. In the theoretical part of this study both fully parametric (where the duration of unemployment is Weibull-distributed) and semiparametric (which needs no assumption about the distribution of durations) regression models are examined. In the empirical part the semiparametric Cox regression model is applied to data of 1764 Finnish unemployed persons who are under 30 years old when becoming unemployed. The estimated employment intensity curve is downward-sloping, meaning that the probability of getting a job diminishes as the duration of unemployment is prolonged. The probability of getting a job depends on the sex, level and field of education and of the activities prior to unemployment. The probability of getting employed of individuals having got training for employment is not statistically significantly lower than that of the reference group. If individuals are selected to training on the basis of low probability of employment, this result could be taken as an indication of the benefits of training for employment.

KEY WORDS: duration of youth unemployment, duration analysis, Cox's model

Saatteeksi

Tutkimus on hyväksytty toukokuussa 1994 Helsingin yliopiston kansantaloustieteen laitoksen opinnäytetyöksi.

Kiitän lämpimästi ohjaajaani, professori Yrjö Vartiaa viisaista ohjeista sekä VTT Mervi Eerolaa asiantuntevasta ja kärsivällisestä opastuksesta.

Haluan esittää kiitokseni myös Teuvo Korvuolle VTKK:sta, Sirkka Kuokkaselle Haapaniemenkadun työvoimatoimistosta sekä Oiva Lönnbergille Työministeriöstä. Teuvolle kiitokset tutkimusaineiston muodostamisesta, Sirkan ja Oivan asiantuntemus puolestaan oli korvaamaton apu mm. työnhakijarekisteriin liittyvissä kysymyksissä. Erityiskiitoksen ansaitsevat samoin VTM Mika Kuismasen kommentit ja rohkaisu.

Valtion taloudellista tutkimuslaitosta kiitän aineiston kustantamisesta sekä työni julkaisusta.

Helsingissä 6.11.1994
Marjo Pyy

Sisältö

Luku 1

Johdanto

Elinaika-analyysi (survival analysis, failure time analysis) tarkastelee erilaisten tapahtumien sattumistodennäköisyyksiä ja sattumistodennäköisyyksien muuttumista ajassa. Sen varhaisimmat sovellukset löytyvät väestötieteestä ja vakuutusmatematiikasta. Eräs varhaisimmista elinaika-analyysin menetelmistä on väestötieteilijöiden 1600-luvulla kehittämä eloonjäämistaulu (life table), jota käytettiin alun perin eri ikäryhmien kuolleisuuden ja jäljellä olevien elinaikojen ennustamiseen ja kuvailuun. Väestötieteilijät tutkivat kuolleisuuden lisäksi myös mm. syntyvyyttä ja avioitumisia. Elinaika-analyysin menetelmiä on sovellettu jo pitkään lääketieteessä, missä tyypillinen sovelluskohde on potilaiden elinaikojen tutkiminen. Insinööritieteissä samoja menetelmiä käytetään erilaisten laitteiden kestävyystutkimuksissa.

Taloustieteessä elinaika-analyysin menetelmiä alettiin soveltaa 1970-luvun lopulla. Sovelluskohteet ovat löytyneet enimmäkseen työmarkkinoiden tutkimuksesta; eniten on tutkittu työttömyyden kestoa ja siihen vaikuttavia tekijöitä. Elinaika-analyysin menetelmin on pyritty selvittämään mm. työttömyyskorvausten tason vaikutusta työttömyyden kestoon sekä testaamaan erilaisten työn etsintäteorioiden paikkansapitävyyttä. Lancasterin (1979) ja Nickellin (1979) tutkimukset olivat ensimmäisiä tutkimuksia, joissa elinaika-analyysiä sovellettiin työttömyyden keston kuvaamiseen. Elinaika-analyysin menetelmiä on käytetty myös työsuhteen keston sekä työpaikan avoinnaolon keston tutkimukseen. Työttömyyden kestoa tarkasteltaessa kiinnostava tapahtuma on yleensä työllistyminen, vaikka muitakin työttömyyden päättymistapoja voidaan tutkia: työttömyys voi päättyä työllistymisen lisäksi myös työvoimasta poistumiseen esim. opintojen tai asepalveluksen aloittamisen vuoksi.

Tutkimuksen tavoitteena on perehtyä elinaika-analyysin teoriaan sekä soveltaa sitä nuorten työttömyyden keston tutkimiseen. Yksilön hyvinvoinnin kannalta on uskoakseni tärkeämpää, kuinka kauan mahdollinen työttömyys kestää, kuin se, jou-

tuuko tämä ylipäänsä työttömäksi vai ei. Lyhytaikaiset työttömyysjaksot liittyvät jo työelämässä olevilla nuorilla usein työpaikan vaihtamiseen ja työelämään ensi kertaa tulevilla ensimmäisen työpaikan etsimiseen (ns. kitkatyöttömyys). Moni ilmoittautuu työnhakijaksi odottaessaan opintojen tai asepalveluksen alkamista. Tämänkaltainen lyhytaikainen työttömyys on normaalia yksilön elämäntilanteen muuttumiseen liittyvää työttömyyttä. Työttömyyden pitkittyminen voi sen sijaan johtaa taloudellisiin tai sosiaalisiin ongelmiin tai työelämästä syrjäytymiseen. Nuoret, joilla ei vielä ole vakiintunutta asemaa työmarkkinoilla, ovat tässä suhteessa erityinen riskiryhmä. Tämän vuoksi on tärkeää selvittää, millaiset henkilökohtaiset ominaisuudet ja taustatekijät johtavat työttömyyden pitkittymiseen.

Käytettävissäni on yksilötason aineisto Työministeriön työnhakijarekisteristä. Aineisto on otos vuonna 1991 työttömäksi tulleista alle 30-vuotiaista työvoimatoimistoon työnhakijoiksi ilmoittautuneista nuorista. Nuoria seurataan työttömyyden keston ajan, kuitenkin enintään vuoden 1993 maaliskuun alkuun. Työnhakijarekisteri sisältää varsin yksityiskohtaisia tietoja työnhakijoista, sekä tiedot työttömyyttä edeltävästä toiminnasta ja työttömyyden päättymissyystä. Aineisto mahdollistaa työttömyyden keston mittaamisen päivän tarkkuudella[1] Rekisterin puutteena on seurannasta poistuneiden yksilöiden, ns. katotapausten suuri määrä.

Taloustieteessä työllistymistodennäköisyyttä on yleensä mallinnettu Weibull-jakaumaan perustuvalla täysin parametroidulla mallilla (täysin parametroidulla mallilla tarkoitetaan mallia, jossa sekä työttömyyden keston jakauma että selittävien tekijöiden vaikutustapa spesifioidaan parametreja vaille etukäteen). Tälle varsin rajoittavalle jakaumaoletukselle on kuitenkin esitetty harvoin perusteluja. Työttömyyden kestoon vaikuttaa mm. lainsäädäntö,[2] joka vaihtelee maittain. Muun muassa tämän vuoksi on epärealistista kuvitella, että on olemassa jokin yleispätevä työttömyyden keston todennäköisyysjakauma. Jos työttömyyden keston jakaumasta ei ole ennakkotietoja, on turvallista valita semiparametrinen malli, jossa jakaumaoletusta ei tarvitse tehdä lainkaan. Tutkimuksessa nuorten työllistymistä on kuvattu semiparametrisella Coxin mallilla. Coxin mallissa regressiokertoimien estimointi perustuu informaatioon yksilöiden keskinäisestä työllistymisjärjestyksestä. Tietoa työttömyysjaksojen tarkoista kestoista ei käytetä, koska kestojen jakaumaa ei tunneta. Semiparametriset mallit ovat yleistyneet viime vuosina taloustieteessä mm. täysin parametroitujen mallien spesifikaatio-ongelmien vuoksi.

[1]Suomalaisissa työttömyyden keston tutkimuksissa (mm. Kettunen 1991, Lilja 1992) käytetyt aineistot ovat yleensä peräisin joko Työministeriön työnhakijarekisteristä tai Tilastokeskuksen työvoimatutkimuksen vuosihaastatteluista. Työvoimatutkimuksen työttömyyden keston mittaustarkkuus on oleellisesti alhaisempi kuin työnhakijarekisterin: työttömyyden kestot tunnetaan ainoastaan kolmen kuukauden tarkkuudella.

[2]Esimerkiksi U.S.A:ssa työttömyyskorvauksia maksetaan normaalisti vain 26 työttömyysviikon ajan. Empiirisissä tutkimuksissa tämän on havaittu lisäävän selvästi työllistymistodennäköisyyttä hieman ennen korvausten loppumista.

Estimointitulosten mukaan työllistymistodennäköisyyteen vaikuttaa mm. sukupuoli, koulutusaste ja -ala, työnhakua edeltävä toiminta sekä työnhaun alkamisajankohta. Naisten työllistymistodennäköisyys on miesten työllistymistodennäköisyyttä suurempi. Koulutus parantaa työllistymismahdollisuuksia ja vaikutus kasvaa koulutusasteen myötä. Akateemisen koulutuksen omaamisen vaikutukset työlistymistodennäköisyyteen ovat jossain määrin ristiriitaisia. Koulutusaloista hoitoala kasvattaa ja kaupallinen ja tekninen ala puolestaan pienentävät työllistymistodennäköisyyttä. Kotityöstä työnhakijoiksi tulleiden työllistymistodennäköisyys on muita ryhmiä pienempi. Tammikuussa työttömäksi tulleet työllistyvät muina kuukausina työttömäksi tulleita nopeammin. Tutkimuksen seuranta-aika (1991-1992) oli alkavan taloudellisen taantuman aikaa, mikä heijastuu estimointituloksissa. Naisten miehiä korkeampi työllistymistodennäköisyys johtunee seuranta-ajan työpaikkojen tarjontatilanteesta: talouden taantuman vaikutukset alkoivat näkyä naisia runsaasti työllistävällä julkisella sektorilla myöhemmin kuin yksityisellä sektorilla. Niin sanottuja naisten työpaikkoja on siis todennäköisesti ollut seuranta-aikana runsaammin tarjolla. Työllisyyskoulutuksesta työnhakijoiksi tulleiden työllistymistodennäköisyys ei estimoitujen mallien perusteella poikkea merkitsevästi vertailuryhmän (armeijasta, opiskelemasta tai työelämästä työnhakijoiksi tulleet sekä ns. ongelmataustan omaavat) työllistymistodennäköisyydestä. Koska työllisyyskoulutukseen valinnan kriteerinä on nimenomaan heikko työllistyvyys (ks. kpl 9.1), voi tuloksen tulkita siten, että työllisyyskoulutus on pystynyt parantamaan koulutettujen asemaa työmarkkinoilla. Taloustieteellisissä tutkimuksissa perinteistä kiinnostuksen kohdetta; työttömyyskorvausten tason vaikutusta työllistymiseen ei ollut mahdollista käsitellä tässä tutkimuksessa.

Luvussa 2 käydään läpi elinaika-analyysin ominaispiirteitä ja peruskäsitteitä. Käsitteiden hallitseminen on tärkeää elinaikamallien ymmärtämiseksi. Luvussa 3 tarkastellaan kahta ei-parametrista eloonjäämisfunktion estimaattoria, Kaplan–Meier-estimaattoria ja eloonjäämistaulua. Ei-parametriset estimaattorit soveltuvat aineiston alustavaan tarkasteluun. Luvuissa 4 ja 5 käsitellään täysin parametroituja elinaikamalleja. Luvussa 5 otetaan populaation heterogeenisuus huomioon sallimalla työttömyyden keston jakauman riippua selittävistä tekijöistä. Luvussa 6 kerrotaan semiparametrisista Coxin ja luokiteltujen tapahtuma-aikojen malleista. Luvussa esitellään lisäksi joustava, täysin parametroitu elinaikamalli; paloittainen eksponenttimalli. Tarkimmin perehdytään Coxin malliin, jota sovelletaan tutkielman empiirisessä osassa. Luvussa 7 käsitellään erilaisten työttömyyden päättymissyiden huomioimista analyysissa. Luvussa 8 kerrotaan Coxin mallin spesifikaatiotestauksesta. Tutkimuksen empiirinen osa koostuu luvusta 9.

Luku 2

Elinaika-analyysin ominaispiirteitä ja peruskäsitteitä

Tässä luvussa esitetään elinaika-analyysin ominaispiirteitä ja peruskäsitteitä. Käsitteitä esitellään sovelluskohteen mukaisesti työttömyyden keston mallintamisen näkökulmasta. Elinaika-analyysin termistö on peräisin elinaikojen tutkimuksesta, jossa kiinnostava tapahtuma on kuolema. Kuolema mielletään yleensä ei-toivotuksi tapahtumaksi, työllistyminen puolestaan toivotuksi. Tästä johtuen elinaika-analyysin käsitteet herättävät helposti vääränlaisia mielleyhtymiä, kun niitä sovelletaan työttömyyden keston tutkimiseen. Elinaika-analyysin termistö on kuitenkin vakiinnuttanut asemansa ja sitä käytetään yleisesti erilaisten tapahtumien toteutumistodennäköisyyksien kuvaamiseen sovellutuskohteesta riippumatta.

Tässä vaiheessa ei vielä erotella erilaisia työttömyyden päättymistapoja, vaan oletetaan, että työttömyys voi päättyä ainoastaan työllistymiseen. Erilaisten työttömyyden päättymissyiden huomioimista analyysissä käsitellään luvussa 7. Kappaleissa 2.1-2.3 kerrotaan elinaika-analyysin ominaispiirteistä ja peruskäsitteistä. Kappale 2.4 käsittelee tarkemmin hasardifunktiota, joka on eräs elinaika-analyysin keskeisimmistä käsitteistä. Elinaikamallien parametrien estimointia käsitellään kappaleessa 2.5.

2.1 Elinaika-analyysin ominaispiirteitä

Elinaika-analyysin menetelmin pyritään selittämään tapahtuman toteutuma-aikaa eli sitä, **milloin muutos tapahtuu** (kuinka kauan työttömyys kestää; milloin työllistyminen tapahtuu). On huomattava, että "milloin" ei tässä yhteydessä viittaa tiettyyn päivämäärään eikä aikaa siten mitata kalenteriaikana, vaan aikana jostain mielekkäästä alkutilanteesta lähtien. Työttömyyden kestoa tutkittaessa aikaa on järkevää mitata kunkin seurattavan työttömäksituloajankohdasta lähtien. Kullakin seurattavalla on siis oma "aikalaskuri", joka lähtee käyntiin työttömäksituloajankohtana. Kutsutaan tapahtuman toteutuma-aikaa **tapahtuma-ajaksi**. Työttömyyden kestoa tutkittaessa tapahtuma-aika on työttömyyden kesto ja kiinnostava tapahtuma yleensä työllistyminen.

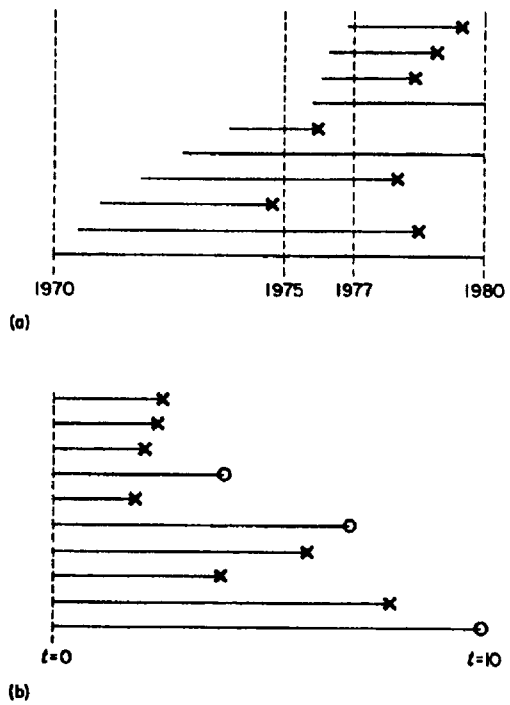
Elinaika-analyysi tarkastelee vain yhtä tapahtumaa kunkin seurattavan osalta. Yleisemmin, jos tarkastellaan useita ja mahdollisesti useanlaisia toisiaan seuraavia tapahtumia eli tapahtumasarjoja, puhutaan **tapahtumahistoria-analyysista**. Elinaika-analyysi on tapahtumahistoria-analyysin erikoistapaus.

Elinaika-analyysin keinoin tarkasteltavat tapahtumat ovat selkeitä, ajallisesti (yksilökohtainen aika) paikallistettavia muutoksia verrattuna aikaisempaan tilanteeseen. Allison (1984) määrittelee tapahtuman laadulliseksi muutokseksi tietyllä ajan hetkellä. Määritelmä sulkee pois vähitellen tapahtuvat muutokset tilasta toiseen kuten esimerkiksi vanhenemisen tai vuodenajan vaihtumisen. Tapahtuma-ajan määrittämiseksi on tärkeää voida spesifioida seurannan aloittamisajankohta sekä ajankohta, jolloin tapahtuma sattuu ja seuranta lopetetaan. Työttömyyden kestoa tutkittaessa seurannan aloittamis- ja lopettamisajankohta on helppo määrittää, mutta esimerkiksi laitteiden kestävyystutkimuksissa tapahtuma-aika (laitteen rikkoutuminen) voidaan määritellä useilla eri tavoilla: seuranta voidaan päättää, kun laitteen toimintakyky alittaa tietyn tason tai kun laite ei toimi enää lainkaan.

Käytännössä työttömyyden keston määrittämisessä on usein aineiston epätäydellisyydestä johtuvia ongelmia. Työttömyyden keston tutkimuksissa käytettävät seuranta-aineistot koostuvat tavallisesti melko harvoin toistuvista haastatteluista, eikä haastateltavien työmarkkina-asemasta haastatteluajankohtien väliaikoina ole useinkaan tietoa. Jos haastateltavan työmarkkina-asema on muuttunut edelliseen haastatteluun verrattuna, tiedetään usein ainoastaan, että muutos on tapahtunut haastattelujen välisenä aikana. Tällaisissa aineistoissa työttömyyden kestoista tiedetään siis ainoastaan, mille haastatteluajankohtien määräämälle aikavälille ne kuuluvat. Mittaustarkkuuden määrää haastattelujen välinen etäisyys. Tällaisille ns. luokitelluille tapahtuma-ajoille on omat analyysimenetelmänsä.

Seuranta-aineistoissa yksilöt tulevat seurannan piiriin yleensä eri aikoina (ns. vai-

heittainen sisääntulo). Usein valitaan ennalta jokin kiinteä seuranta-aika. Kuvan 2.1 a) tapauksessa seuranta-aika on 10 vuotta. Kutakin yksilöä seurataan työllistymiseen tai enintään seuranta-ajan päättymiseen saakka. Seuranta siis yleensä sekä alkaa että päättyy eri henkilöillä eri ajankohtina. Yleensä osalle seurattavista tapahtuma ei ehdi sattua seurannan aikana. Tutkittavat tapahtumat voivat olla luonteeltaan sellaisia, että ne eivät koskaan tapahdukaan kaikille seurattaville (esim. avioituminen). Osa seurattavista voi poistua tutkimuksesta ennen seurannan päättymistä (ns. katotapaukset). Tällaisia havaintoja sanotaan **oikealta sensuroiduiksi**. Sensuroidut havainnot ovat tavallisia seuranta-aineistoissa. Kuva 2.1 havainnollistaa seurannan alkamisajankohdan vaihtelevuutta ja sensurointia. (kuvan 2.1 sensuroidut havainnot ovat havaintoja, joille tapahtuma ei ole ehtinyt sattua seurannan aikana, ts. sensurointi ei tässä tapauksessa johdu kadosta).



Kuva 2.1: (a) seuranta-aika (reaaliaika), (b) seurannan pituus vuosissa
x = tapahtumat, o = sensuroinnit. Lähde: Cox & Oakes s. 3.

Sensuroinnit tekevät aineistosta epätäydellisen: sensuroitujen havaintojen osalta ei tapahtuma-aikaa tunneta. Sensuroidut havainnot sisältävät kuitenkin analyysin kannalta tärkeää informaatiota: tiedetään, että näiden havaintojen tapahtuma-aika on pidempi kuin seurannassaoloaika. Kutsutaan sensuroitujen havaintojen seurannassaoloaikaa **sensurointiajaksi**. Sensurointiaika tarkoittaa aikaväliä seurannan yksilökohtaisesta alkamisajankohdasta seurannasta poistumiseen, jos kyseessä on katotapaus. Havainnoille, joille tutkittava tapahtuma ei ole ehtinyt sattua seuranta-aikana, sensurointiaika tarkoittaa aikaväliä seurannan alkamisajankohdasta seuranta-ajan (esimerkin tapauksessa 10 vuotta) päättymiseen.

Sensuroinnin vuoksi kuhunkin havaintoon liittyvä seurantatieto esitetään yleensä muodossa (X_i, V_i), missä $X_i = \min(T_i, c_i)$. T_i on havaintoon i liittyvä satunnainen tapahtuma-aika, c_i sensurointiaika (oletetaan, että sensurointia tapahtuu vain seuranta-ajan päättymisen vuoksi. Tällöin sensurointiajat ovat ennalta tunnettuja, kiinteitä lukuja). V_i on ns. sensurointi-indikaattori: $V_i = 1$ jos $T_i \leq c_i$ (tapahtuma toteutuu), $V_i = 0$ jos $T_i > c_i$ (tapahtuma ei ehdi toteutua). Siis jos tapahtuma ehtii toteutua ennen sensurointiaikaa, on X_i havaintoon i liittyvä tapahtuma-aika. Muulloin X_i on havainnon i sensurointiaika.

I lajin sensuroinnissa sensurointiaika valitaan etukäteen ja on sama kaikille seurattaville eli $c_i = c \ \forall i = 1, \ldots, n$. n on seurattavien lukumäärä. **II lajin sensuroinnissa** seuranta päättyy, kun jokin ennalta sovittu määrä tapahtumia on sattunut. II lajin sensuroinnissa ei seuranta-ajan pituudella ole ylärajaa. Samoin kuin I lajin sensuroinnissa, on sensurointiaika II lajin sensuroinnissa kaikille sama. Sensurointiaikaa ei kuitenkaan nyt tiedetä ennalta, vaan c on satunnaismuuttujan C havaittu arvo, joka riippuu seurattavien historiasta (tapahtumista ja ei-tapahtumista) sensurointiajankohtaan saakka. II lajin sensurointia käytetään mm. laitteiden kestävyystutkimuksissa.

Sensuroinnin lisäksi toinen erityispiirre seuranta-aineistoissa on se, että tapahtuma-aikaa selittävien tekijöiden arvot voivat vaihdella seurannassaoloaikana. Työttömyyden kestoa tutkittaessa tällaisia muuttujia ovat esimerkiksi työttömyyskorvauksen taso (mikäli työttömyys kestää pitkään, putoaa ansiosidonnaista päivärahaa saava peruspäivärahaa saavien joukkoon) ja työvoiman kysyntä. Tämänkaltaista selittävien tekijöiden ajallista vaihtelevuutta on vaikea mallintaa tavallisen regressioanalyysin keinoin.

2.2 Peruskäsitteitä

Seuraavassa tarkastellaan elinaika-analyysin peruskäsitteitä. Olkoon T positiivisia arvoja saava jatkuva satunnaismuuttuja, joka kuvaa tarkasteltavan populaation

jäsenten (työttömien työnhakijoiden) työttömyyden kestoa. Käytännössä työttömyyden kesto on diskreetti muuttuja koska kestot tunnetaan enintään päivän tarkkuudella. Päivä on kuitenkin tarkin mielekäs työttömyyden keston mittayksikkö (kestojen mittaaminen esim. tuntien tarkkuudella ei tuo lisäinformaatiota tutkittavan ilmiön kannalta), ja työttömyyden kestot vaihtelevat yhdestä päivästä jopa yli tuhanteen päivään (Pitkiä työttömyyden kestoja on etenkin vanhemmilla työnhakijoilla). Työttömyyden kestoa voidaan siten pitää jatkuvana muuttujana. Oletetaan, että populaatio on homogeeninen. Tällöin työnhakijoiden työttömyyden kestot ovat realisaatioita samasta todennäköisyysjakaumasta. Määritellään ns. **eloonjäämisfunktio** seuraavalla tavalla:

$$S(t) = P(T \geq t). \tag{2.1}$$

Eloonjäämisfunktio kuvaa todennäköisyyttä, että työttömyys kestää vähintään t:n verran. Kun T on jatkuva, on eloonjäämisfunktio kertymäfunktion komplementti: $S(t) = 1 - F(t)$. $S(t)$ on vähenevä ja saa ääriarvot $S(0) = 1$ ja $\lim_{t\to\infty} S(t) = 0$. Toinen keskeinen käsite elinaikamalleissa on ns. **hasardifunktio**:

$$h(t) = \lim_{dt\to 0} \frac{P(t \leq T < t + dt \mid T \geq t)}{dt}. \tag{2.2}$$

$h(t)dt$ kuvaa ehdollista työttömyyden päättymistodennäköisyyttä pienellä dt:n pituisella aikavälillä, ehdolla, että työttömyys ei ole päättynyt ennen aikavälin alkua ($T \geq t$). Hasardifunktion englanninkielisiä nimityksiä ovat hazard function, rate, failure rate ja intensity. Sana hazard merkitsee vaaraa tai riskiä, minkä vuoksi on nurinkurista puhua työllistymishasardista. Jatkossa käytetään funktiosta $h(t)$ nimitystä **työllistymisintensiteetti** tai **työttömyyden keston hasardifunktio.**

Hasardifunktion ja tiheysfunktion erona on se, että tiheysfunktiossa ei ole ehdollistavaa tekijää: $f(t)dt$ kuvaa yksinkertaisesti työllistymistodennäköisyyttä aikavälillä dt. Toisin sanoen, $h(t)dt$ kuvaa niiden henkilöiden työllistymistodennäköisyyttä aikavälillä dt, jotka ovat hetkellä t yhä vailla työtä, kun taas $f(t)dt$ kuvaa kaikkien seurattavien työllistymistodennäköisyyttä kyseisellä aikavälillä riippumatta siitä, onko seurattavan työttömyys päättynyt hetkeen t mennessä vai ei. Samoin kuin tiheys- ja kertymäfunktio, kuvaavat sekä eloonjäämisfunktio että hasardifunktio satunnaismuuttujan jakauman yksikäsitteisesti. Ehdollisen todennäköisyyden kaavaa[1] soveltamalla saadaan hasardifunktiolle seuraavanlainen esitysmuoto:

$$h(t) = \lim_{dt\to 0} \frac{P(t \leq T < t + dt)}{P(T \geq t)\, dt} = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}, \tag{2.3}$$

jossa $f(t)$ on T:n tiheysfunktio ja $F(t)$ kertymäfunktio. Hasardifunktiolle pätee myös:

$$h(t) = \frac{-d\ln(1 - F(t))}{dt} = \frac{-d\ln S(t)}{dt}. \tag{2.4}$$

[1] $P(A \mid B) = P(AB)/P(B)$

Koska hasardifunktio voidaan esittää eloonjäämisfunktion logaritmin derivaattana, kuvaa hasardifunktio eloonjäämisfunktion **suhteellista muutosta**. Tiheysfunktiolle puolestaan pätee $f(t) = -dS(t)/dt$ eli tiheysfunktio kuvaa eloonjäämisfunktion **absoluuttista muutosta**. Integroimalla yhtälö 2.4 puolittain t:n suhteen ja huomioimalla, että $\ln S(0) = 0$, voidaan eloonjäämisfunktio esittää hasardifunktion avulla:

$$S(t) = \exp\{-\int_0^t h(u)du\} = \exp\{-H(t)\}. \tag{2.5}$$

Tällöin

$$f(t) = h(t)S(t) = h(t)\exp\{-\int_0^t h(u)du\} = h(t)\exp\{-H(t)\}. \tag{2.6}$$

Kaarisulkeissa olevaa termiä $H(t) = \int_0^t h(u)du$ kutsutaan kumulatiiviseksi hasardifunktioksi.

Jos työttömyyden päättymistodennäköisyyksiä tarkastellaan tietyillä aikaväleillä (kuten on usein laita haastattelututkimuksiin perustuvissa aineistoissa), on T diskreetti satunnaismuuttuja. Oletetaan, että T saa arvoja $t_1 < t_2 < \ldots$. T:n eloonjäämisfunktio on tällöin

$$S(t) = P(T \geq t) = \sum_{j|t_j \geq t} f(t_j), \tag{2.7}$$

jossa $f(t_j)$ on T:n pistetodennäköisyysfunktio. Hasardifunktion arvo ajankohtana t_j kuvaa diskreetissä tapauksessa todennäköisyyttä, että työnhakija työllistyy kyseisenä ajankohtana ehdolla, että hän on tällöin yhä työtön: [2]

$$h(t_j) = P(T = t_j \mid T \geq t_j) = \frac{f(t_j)}{f(t_j) + f(t_{j+1}) + \ldots} = \frac{f(t_j)}{S(t_j)}. \tag{2.8}$$

Eloonjäämisfunktio ja pistetodennäköisyysfunktio voidaan, kuten jatkuvassa tapauksessa, esittää hasardifunktion avulla. Todennäköisyys, että työnhakijan työttömyys kestää vähintään t:hen saakka voidaan laskea ehdollisten todennäköisyyksien ketjusäännön [3] avulla työttömänä pysymisen ehdollisten todennäköisyyksien $(1 - h(t_j))$, $t_j < t$ tulona:

$$S(t) = \prod_{j|t_j<t} (1 - h(t_j)). \tag{2.9}$$

Tiheysfunktio hasardifunktion avulla esitettynä on vastaavasti:

$$f(t_j) = h(t_j)S(t_j) = h(t_j)\prod_{l=1}^{j-1}(1 - h(t_l)). \tag{2.10}$$

[2] jatkuvassa tapauksessa todennäköisyystulkinta voidaan antaa $h(t)dt$:lle.

[3] $P(ABC) = P(A \mid BC) \times P(B \mid C) \times P(C)$

2.3 Sensuroinnin epäinformatiivisuus

Yleensä oletetaan analyysin yksinkertaistamiseksi, että sensurointi on tutkittavan tapahtuman kannalta **epäinformatiivista**. Tämä tarkoittaa sitä, että sensuroituminen ei vaikuta tapahtuman toteutumistodennäköisyyteen; tapahtuman oletetaan sattuvan sensuroiduille havainnoille yhtä suurella todennäköisyydellä kuin samanlaisille (samat selittävien tekijöiden arvot omaaville) sensuroimattomille havainnoille. Oletus ei pitäisi paikkaansa esim. tilanteessa, jossa yksilöt, joilla on tavallista pienempi todennäköisyys saada työtä, poistuvat seurannasta muita useammin. Tällöin sensurointi sisältäisi työllistymistodennäköisyyteen liittyvää informaatiota. Jos seurannasta poistuu systemaattisesti alhaisen työllistymistodenäköisyyden omaavia yksilöitä, on estimoitu hasardifunktio ylöspäin harhainen eli estimoitu työllistymisintensiteetti on todellista suurempi.

I Lajin sensuroinnissa sensurointiajat ovat ennaltamäärättyjä vakioita eikä sensuroituminen siten sisällä työllistymistodennäköisyyteen liittyvää informaatiota. Usein sensurointiaikoja ei kuitenkaan tunneta etukäteen, kuten on laita seurannasta poistuneiden havaintojen osalta (katotapaukset). Oletetaan, että havaitut sensurointiajat c_i ovat tällöin satunnaismuuttujien C_i, $i = 1, \ldots, n$ realisaatioita [4] ja että satunnaismuuttujien eloonjäämis- ja tiheysfunktiot ovat $G(c)$ ja $g(c)$. Tapahtuma-aikojen T_i, $i = 1, \ldots, N$ eloonjäämis- ja tiheysfunktiot ovat, kuten edellä, $f(t)$ ja $S(t)$. Sensurointi on epäinformatiivista, mikäli sensurointiajat ovat tapahtuma-ajoista riippumattomia. Kun kyseessä on tapahtuma-aika ($v_i = 1$), on X_i:n ($X_i = min(T_i, C_i)$) ja V_i:n yhteisjakauma

$$P(X_i \in (t, t+dt), V_i = 1) = P(T_i \in (t, t+dt), C_i > t) = f(t)dt\, G(t). \qquad (2.11)$$

Sensuroiduille havainnoille X_i:n ja V_i:n yhteisjakauma on puolestaan

$$P(X_i \in (t, t+dt), V_i = 0) = P(C_i \in (t, t+dt), T_i > t) = g(t)dt\, S(t). \qquad (2.12)$$

n:n riippumattoman havainnon (sensuroidun tai sensuroimattoman) yhteisjakauma on siis

$$\begin{aligned} f(t_1, \ldots, t_n) &= \prod_{i=1}^{n} \left[(f(t_i)G(t_i))^{v_i} (g(t_i)S(t_i))^{1-v_i} \right] \qquad (2.13) \\ &= \prod_{i=1}^{n} \left[(G(t_i))^{v_i} (g(t_i))^{1-v_i} \right] \prod_{i=1}^{n} \left[(f(t_i))^{v_i} (S(t_i)^{1-v_i} \right]. \end{aligned}$$

Epäinformatiivisuusoletuksen mukaisesti lausekkeen 2.14 ensimmäinen termi ei sisällä työllistymisen kannalta kiinnostavia parametreja ja voidaan siten jättää huo-

[4] Jos tapahtuma sattuu seuranta-aikana, jää sensurointiaika luonnollisesti tuntemattomaksi. Cox ja Oakes (1984) puhuvat tässä yhteydessä ns. potentiaalisista sensurointiajoista.

mioimatta analyysissä. Tällöin havaintojen uskottavuusfunktio

$$L \propto \prod_{i=1}^{n} \left[(f(t_i))^{v_i} (S(t_i))^{1-v_i} \right]. \tag{2.14}$$

Sensuroitujen havaintojen kontribuutio uskottavuusfunktioon on $S(t_i)$; todennäköisyys, että tapahtuma-aika on pidempi kuin sensurointiaika. Sensuroimattomien havaintojen kontribuutio on $f(t_i)$; todennäköisyys, että tapahtuma-aika on havaitun pituinen.

2.4 Hasardifunktion erityisasemasta

Tiheysfunktio ja hasardifunktio ovat vaihtoehtoisia tapoja kuvata tapahtuman sattumistodennäköisyyden muuttumista ajassa. $f(t)dt$ kuvaa kaikkien seurattavien työllistymistodennäköisyyttä dt:n pituisella aikavälillä, kun taas $h(t)dt$ kuvaa riskijoukon (hetkellä t yhä vailla työtä olevien) työllistymistodennäköisyyttä kyseisellä aikavälillä. Elinaika-analyysissä on perinteisesti kuvattu tapahtuman sattumistodennäköisyyden muutoksia hasardifunktion avulla. Tämä on luontevaa, sillä usein juuri hasardifunktiosta on olemassa ennakkotietoja. Näin on esimerkiksi mallinnettaessa ihmisen elinikää. Tiedetään, että eliniän hasardifunktio (kuolemisriski) on suunnilleen U-kirjaimen muotoinen: lasten ja vanhusten kuolleisuus on suurempi kuin muun ikäisten. Työttömyyden kestoa tarkasteltaessa näin ei valitettavasti ole. Työllistymistodennäköisyyteen vaikuttavat lukuisat tekijät, joiden yhteisvaikutusta on vaikea arvioida. Työn etsintäteorioiden implikaatiot hasardifunktion muodolle riippuvat tehdyistä oletuksista. Etsintäteorioiden mukaan hasardifunktio on kasvava eli ehdollinen työllistymistodennäköisyys kasvaa (positiivinen duraatioriippuvuus), mikäli työnhakija madaltaa reservaatiopalkkaansa työttömyyden pitkittyessä. Reservaatiopalkalla tarkoitetaan pienintä palkkaa, jolla työnhakija hyväksyy työtarjouksen. Reservaatiopalkkaa saava työntekijä on indifferentti työllisyyden ja työttömyyden välillä. Palkkavaatimuksen madaltaminen tuntuu luontevalta oletukselta, sillä useimmille työttömyyden pitkittyminen lienee ei-toivottua. Toisaalta, jos työnantajat katsovat työttömyyden heikentävän työnhakijan työtaitoja ja -motivaatiota, voi työttömyyden pitkittyminen pienentää työllistymistodennäköisyyttä. On vaikea tietää ennalta, millainen on näiden ja mahdollisesti monien muiden tekijöiden yhteisvaikutus työttömyyden keston hasardifunktioon.

Duraatioriippuvuudella tarkoitetaan sitä, että työllistymistodennäköisyys vaihtelee työttömyyden keston suhteen. Jos työllistymistodennäköisyys ei riipu siitä, kuinka kauan työttömyys on kestänyt, on hasardifunktio vakio eikä duraatioriippuvuutta esiinny. Eksponenttijakauman hasardifunktiolla on tällainen ns. unohtavaisuusominaisuus. **Positiivisella duraatioriippuvuudella** tarkoitetaan sitä, että

työllistymistodennäköisyys kasvaa työttömyyden pitkittyessä eli

$$\frac{dh(t)}{dt} > 0. \tag{2.15}$$

Negatiivinen duraatioriippuvuus merkitsee puolestaan sitä, että työttömyyden pitkittyminen pienentää työllistymistodennäköisyyttä:

$$\frac{dh(t)}{dt} < 0. \tag{2.16}$$

Vaikka ennakkotietoja hasardifunktion muodosta ei työttömyyden kestoa tarkasteltaessa olisi käytössä, on luultavaa, että sillä, kuinka kauan työttömyys on kestänyt, on vaikutusta työllistymistodennäköisyyteen. Tällöin on luontevaa tarkastella ehdollisia työllistymistodennäköisyyksiä ja mallintaa hasardifunktiota.

2.5 Elinaikamallien parametrien estimoinnista

Tavallinen pienimmän neliösumman estimointimenetelmään perustuva regressioanalyysi ei pysty hyödyntämään sensuroitujen havaintojen sisältämää informaatiota, koska selitettävän tekijän, tapahtuma-ajan (työttömyyden kesto) arvot ovat sensuroitujen havaintojen osalta tuntemattomia. Sensuroitujen havaintojen poisjättäminen olisi informaation tuhlausta, sensurointiaikojen käsittely tapahtuma-aikoina puolestaan aliarvioisi todellisia tapahtuma-aikoja ja tuottaisi harhaisia parametriestimaatteja.

Elinaikamallien parametrit on estimoitava suurimman uskottavuuden menetelmällä. Menetelmä perustuu havainnoista lasketun uskottavuusfunktion maksimointiin. Uskottavuusfunktio on satunnaismuuttujien $T_1, \ldots, T_n$ yhteisjakauma pisteessä $t_1, \ldots, t_n$ ja kuvaa todennäköisyyttä saada havaitunkaltaiset työttömyyden kestot. Uskottavuusfunktio on n:n riippumattoman, samoin jakautuneen havainnon tapauksessa

$$L = \prod_{i=1}^{n} f(t_i; \theta), \tag{2.17}$$

missä $t_1, \ldots, t_n$ ovat satunnaismuuttujien $T_1, \ldots, T_n$ realisaatioita ja $f(t; \theta)$ satunnaismuuttujien todennäköisyysjakauma. Uskottavuusfunktio on tuntemattoman jakaumaparametrin θ funktio. Periaatteena on etsiä se jakaumaparametrin arvo, joka maksimoi todennäköisyyden saada havaitunkaltainen otos (havaitut työttömyyden kestot). Tällaista parametrin arvoa kutsutaan suurimman uskottavuuden estimaatiksi.

Uskottavuusfunktio voidaan esittää myös hasardifunktion avulla. Sensuroiduista havainnoista käytetään hyväksi tieto, että työttömyyden kesto on pidempi kuin sensurointiaika (ks. kpl 2.3). Henkilön i toteutuneen seurannassaoloajan $x_i = \min(t_i, c_i)$ avulla esitettynä on uskottavuusfunktio

$$L = \prod_t f(x_i) \prod_s S(x_i), \tag{2.18}$$

missä t on työllistyneiden ja s sensuroitujen henkilöiden joukko ($t + s = n$). Kun T on jatkuva, on uskottavuusfunktio hasardifunktion avulla esitettynä

$$\begin{aligned} L &= \prod_t [h(x_i)S(x_i)] \prod_s S(x_i) \\ &= \prod_t h(x_i) \prod_n S(x_i) \\ &= \prod_n h(x_i)^{v_i} \exp(-H(x_i)). \end{aligned} \tag{2.19}$$

v_i on henkilön i sensurointi-indikaattorin havaittu arvo. Diskreetissä tapauksessa sensuroidun havainnon kontribuutio uskottavuusfunktioon on

$$P(T_i > c_i) = S(c_i + 0) = \prod_{j=1}^{i} (1 - h(t_j)), \tag{2.20}$$

missä $S(c_i + 0) = \lim_{x \to 0+} S(c_i + x)$. Jos ajankohtana t_i työllistyy d_i työnhakijaa ja m_i sensuroidaan, on uskottavuusfunktio seuraavanlainen:

$$\begin{aligned} L &= \prod_{i=1}^{k} \{[h(t_i) \prod_{j=1}^{i-1} (1 - h(t_j))]^{d_i} \prod_{j=1}^{i} (1 - h(t_j))^{m_i}\} \\ &= \prod_{i=1}^{k} \{h(t_i)^{d_i} \prod_{j=1}^{i} (1 - h(t_j))^{d_i + m_i} (1 - h(t_i))^{-d_i})\} \\ &= \prod_{i=1}^{k} \{h(t_i)^{d_i} (1 - h(t_i))^{r_i - d_i}\}. \end{aligned} \tag{2.21}$$

$r_i = (m_i + d_i) + \ldots + (m_k + d_k)$ on ns. **riskijoukko** ajankohdan t_i alussa: se koostuu henkilöistä, joka eivät ole työllistyneet tai joita ei ole sensuroitu kyseiseen ajankohtaan mennessä. Yleensä oletetaan, että ajankohtana t_i sensuroitu havainto on sensuroitu välittömästi kyseisen ajankohdan jälkeen, jolloin havainto kuuluu vielä ajankohdan riskijoukkoon. k on ajankohtien t_i lukumäärä. Viimeisestä lausekkeesta nähdään, että diskreetissä tapauksessa uskottavuusfunktio saadaan laskemalla tapahtumien ja ei-tapahtumien (työllistymisien ja ei-työllistymisien) ehdollisten todennäköisyyksien tulo kunakin ajankohtana.

Luku 3

Aineiston ei-parametriset kuvaustavat

Tässä luvussa käsitellään eloonjäämisfunktion ei-parametrisia estimaattoreita, Kaplan–Meier-estimaattoria ja eloonjäämistaulua (life table). Esitys seuraa Cox & Oakesia (1984) sekä Kalbfleisch & Prenticea (1980). Kaplan–Meier-estimaattoria sanotaan myös tulo-raja-estimaattoriksi. Ei-parametristen estimaattorien laskemiseksi ei tarvitse spesifioida työttömyyden keston jakaumaa. Niitä käytetäänkin yleensä aineiston alustavaan tarkasteluun ja apuna jakauman valinnassa. Kaplan–Meier-estimaattori ja eloonjäämistaulu ovat porrasfunktioita, jotka kuvaavat yhä työttöminä olevien henkilöiden osuutta tutkittavasta joukosta tietyllä ajan hetkellä tai aikavälillä. Osittamalla aineistoa työttömyyden kestoa selittävien tekijöiden suhteen ja estimoimalla eloonjäämisfunktiot erikseen ositteille voidaan tutkia alustavasti selittävien tekijöiden vaikutusta työttömyyden kestoon.

Estimaattorien laskemiseksi on oletettava, että tutkittava populaatio on homogeeninen, ts. havaitut työttömyyden kestot ovat realisaatioita samasta todennäköisyysjakaumasta. Aineistoa ositettaessa oletetaan, että osittava muuttuja on ainoa heterogeenisuutta aiheuttava tekijä ja että homogeenisuusoletus on voimassa ositteiden sisällä. Tämä on tietenkin varsin vahva oletus. Jos aineistoa osittavan muuttujan ja jonkin muun selittävän muuttujan välillä on interaktiota, voivat estimaatit antaa vääristyneen kuvan muuttujan vaikutuksesta työttömyyden kestoon.

Kaplan-Meier–estimaattoria laskettaessa oletetaan työttömyyden keston jakauman olevan diskreetti. Estimaatit lasketaan kullekin havaitulle ei-sensuroidulle työttömyyden kestolle erikseen. Eloonjäämistauluja laskettaessa oletetaan työttömyyden keston jakauma jatkuvaksi. Eloonjäämistaulu soveltuu tilanteisiin, joissa tiedetään ainoastaan, mille aikavälille $[t_{j-1}, t_j)$ työttömyyden kestot kuuluvat.

3.1 Kaplan–Meier-estimaattori

Kaplan-Meier–estimaattori on empiirisen eloonjäämisfunktion yleistys, joka ottaa huomioon seuranta-aineistoissa tavalliset sensuroidut havainnot. Empiirinen eloonjäämisfunktion estimaattori on

$$\hat{S}(t) = \frac{\text{t:n ylittävien kestojen lkm}}{n}, \tag{3.1}$$

jossa n on havaintojen lukumäärä. Tätä estimaattoria ei voida kuitenkaan käyttää aineistoihin, jotka sisältävät sensuroituja havaintoja, koska yli t:n pituisten kestojen lukumäärää ei tunneta kaikilla t:n arvoilla. Kaplan-Meier–estimaattori huomioi sensuroidut havainnot käyttämällä hyväksi riskijoukon käsitettä. Olkoon $t_1 < t_2 < \ldots < t_k$ diskreetin satunnaismuuttujan T saamat arvot; havaitut työttömyyden kestot n:n suuruisessa joukossa. Yleensä $n \neq k$, koska usealla henkilöllä saattaa olla samanpituinen työttömyysjakso. Eloonjäämisfunktion Kaplan-Meier–estimaattori on (vrt. yhtälö 2.9):

$$\hat{S}(t) = \prod_{i=1}^{t-1}(1 - \hat{h}(t_i)), \tag{3.2}$$

jossa $\hat{h}(t_i)$:t ovat uskottavuusfunktion (2.21) maksimoivia diskreetin hasardifunktion suurimman uskottavuuden estimaattoreita (suurimman uskottavuuden estimaattoreita on k kpl eli yksi kutakin sensuroimatonta työttömyyden kestoa kohti). Uskottavuusfunktio maksimoidaan siis funktion, ei kuten tavallisesti parametrien suhteen. Uskottavuusfunktio maksimoidaan derivoimalla uskottavuusfunktion logaritmi h_i :n suhteen $(i = 1, \ldots, k)$ ja asettamalla derivaatta nollaksi. Uskottavuusfunktion logaritmi on

$$l = \sum_{i=1}^{k} [d_i \ln h_i + (r_i - d_i) \ln(1 - h_i)]. \tag{3.3}$$

$\hat{h}(t_i)$ saadaan seuraavan yhtälön ratkaisuna:

$$\frac{\partial l}{\partial h_i} = \frac{d_i}{h_i} - \frac{r_i - d_i}{1 - h_i} = 0. \tag{3.4}$$

$\hat{h}_i = d_i/r_i$. Työllistymisen ehdollisen todennäköisyyden suurimman uskottavuuden estimaattori ajankohtana t_i on siis kyseisenä ajankohtana työllistyneiden suhde riskijoukkoon. Kaplan-Meier–estimaattori on siten

$$\hat{S}(t) = \prod_{i=1}^{t-1}(1 - d_i/r_i). \tag{3.5}$$

Estimaattorin arvo muuttuu jokaisen työllistymiseen päättyneen työttömyysjakson kohdalla eli aina, kun $d_i \neq 0$. Ajankohtana t_i sensuroidut kestot vaikuttavat vain

saman ja aiempien ajankohtien riskijoukkoon (oletetaan, että ajankohtana t_i sensuroidut havainnot on sensuroitu välittömästi kyseisen ajankohdan jälkeen). Jos aineisto ei sisällä sensuroituja havaintoja, vastaa Kaplan–Meier-estimaattori empiiristä eloonjäämisfunktiota 3.1.

3.2 Eloonjäämistaulu

Eloonjäämistaulu on eräs elinaika-analyysin varhaisimpia menetelmiä. Oletetaan, että todellisia päättyneiden tai sensuroitujen työttömyysjaksojen pituuksia ei tunneta, vaan tiedetään ainoastaan, että jaksot kuuluvat jollekin aikavälille $[t_{j-1}, t_j)$, $j = 1, \ldots, m$, $(t_0 = 0)$. $r_j = (t_j - t_{j-1})$ on aikavälin pituus. Jos d_j työllistyy ja m_j sensuroidaan aikavälillä $[t_{j-1}, t_j)$, on ehdollisen työllistymistodennäköisyyden (hasardifunktion) estimaattori kyseisellä aikavälillä

$$\hat{q}_j = \frac{d_j}{n_j - m_j/2}, \tag{3.6}$$

missä n_j on riskijoukko aikavälin alussa. Riskijoukko ei ole vakio aikavälin sisällä vaan pienenee sensurointien vuoksi. Tämä otetaan huomioon vähentämällä aikavälin alun riskijoukosta puolet kyseisellä aikavälillä sensuroiduista. Jos sensuroinnit ajoittuvat tasaisesti aikavälin sisällä, kuvaa $n_j - m_j/2$ riskijoukkoa aikavälin puolessa välissä. Hasardifunktiota approksimoidaan siis funktiolla, joka on vakio kullakin aikavälillä. Pienentämällä aikavälin pituutta saadaan yhä tarkempi hasardifunktion estimaattori. Paloittain vakio hasardifunktio merkitsee sitä, että työttömyysjaksojen pituudet ovat kullakin aikavälillä eksponentiaalisesti jakautuneita (ks. kpl 4.1). Eloonjäämisfunktion estimaattori j:nnen aikavälin lopussa on

$$\tilde{S}(t_j) = \prod_{i=1}^{j} (1 - \hat{q}_i). \tag{3.7}$$

Luku 4

Työttömyyden keston kuvaamisessa käytettyjä jakaumia

Tässä luvussa esitellään työttömyyden keston kuvaamisessa käytettyjä jakaumia. Weibull-jakauma on ollut empiirisissä työttömyyden keston tutkimuksissa selvästi suosituin. Eräs syy tälle lienee se, että Weibull-jakauman eri esitysmuodot ovat suhteellisen yksinkertaisia ja siten laskuteknisiltä hankaluuksilta vältytään mallin parametreja estimoitaessa. Weibull-jakauma on myös kaksiparametrisena jakaumana joustavampi kuin yksiparametrinen eksponenttijakauma. Eksponenttijakauma on Weibull- ja gammajakauman tärkeä erikoistapaus. Jos työttömyyden kesto on eksponenttijakautunut, on työllistymistodennäköisyys riippumaton työttömyyden kestosta. Gammajakaumaa on käytetty mallinnettaessa ns. havaitsematonta heterogeenisuutta (ks.kpl 5.4). Yleistetty gammajakauma sisältää erikoistapauksina kaikki edellämainitut jakaumat, samoin kuin log-normaalin jakauman. Esitellyt jakaumat ovat jatkuvia ja niiden todennäköisyysmassa keskittyy ei-negatiivisille muuttujan arvoille.

Jakauman valintaan vaikuttaa laskuteknisten seikkojen lisäksi myös hasardifunktion muoto. Useissa sovellutuksissa juuri hasardifunktiosta on olemassa ennakkotietoja, jolloin mahdollisten jakaumien joukko voidaan rajoittaa niihin jakaumiin, joiden hasardifunktio käyttäytyy oletetulla tavalla. Työttömyyden kestoa tutkittaessa ei tällaisia ennakkotietoja ole käytettävissä. Tällöin on turvallista valita jokin mahdollisimman joustava jakauma, joka rajoittaa hasardifunktion muotoa mahdollisimman vähän.

Luvussa oletetaan edelleen, että tutkittava populaatio on homogeeninen eli että populaation jäsenet ovat samanlaisia työttömyydeen vaikuttavien tekijöiden suhteen. Tämä merkitsee sitä, että populaation jäsenien havaitut työttömyyden kestot ovat

realisaatioita samasta todennäköisyysjakaumasta. Luvussa 5 lievennetään tätä epärealistista oletusta.

4.1 Eksponenttijakauma

Eksponenttijakauma on ensimmäisiä tapahtuma-aikojen kuvaamiseen käytettyjä jakaumia. Eksponenttijakaumaa on käytetty mm. erilaisten laitteiden käyttöikien mallintamiseen. Mikä tahansa seuraavista funktioista kuvaa jakauman täydellisesti:

$$\begin{aligned} f(t) &= \lambda \exp(-\lambda t) \\ F(t) &= 1 - \exp(-\lambda t) \\ S(t) &= \exp(-\lambda t) \\ h(t) &= \lambda. \end{aligned}$$

Eksponenttijakautuneen satunnaismuuttujan T (työttömyyden kesto) hasardifunktio on vakio. Tämä merkitsee sitä, että jakaumalla on ns. **unohtavaisuusominaisuus**: työllistymistodennäköisyys ei riipu siitä, kuinka kauan työttömyys on kestänyt. Työttömyyden keston suhteen vakio työllistymistodennäköisyys on voimakas ja usein epärealistinen oletus. Tämä rajoittaa eksponenttijakauman käyttöä käytännön sovelluksissa. Eloonjäämisfunktion luonnollinen logaritmi; $\ln S(t) = -\lambda t$ on työttömyyden keston lineaarinen funktio. Yksinkertainen tapa tarkistaa, onko työttömyyden kesto eksponenttijakautunut, on piirtää esim. Kaplan–Meier-menetelmällä estimoidun eloonjäämisfunktion luonnollinen logaritmi vastaan t. Jos oletus eksponenttijakaumasta pitää paikkansa, tulisi pisteiden sijaita likimain origon kautta kulkevalla suoralla [1]. Kappaleessa 2.2 osoitettiin, että jatkuvan satunnaismuuttujan T eloonjäämisfunktio voidaan esittää kumulatiivisen hasardifunktion avulla: $S(t) = \exp[-H(t)]$. Vertaamalla yo. yhtälöä ja eksponenttijakautuneen satunnaismuuttujan eloonjäämisfunktiota nähdään, että kumulatiivisella hasardifunktiolla muunnettu satunnaismuuttuja T; $H(T)$ noudattaa standardoitua eksponenttijakaumaa ($\lambda = 1$) T:n jakaumasta riippumatta. Tätä tulosta käytetään myöhemmin hyväksi esitettäessä suhteellisten hasardien malli lineaarisena regressiomallina.

4.2 Weibull-jakauma

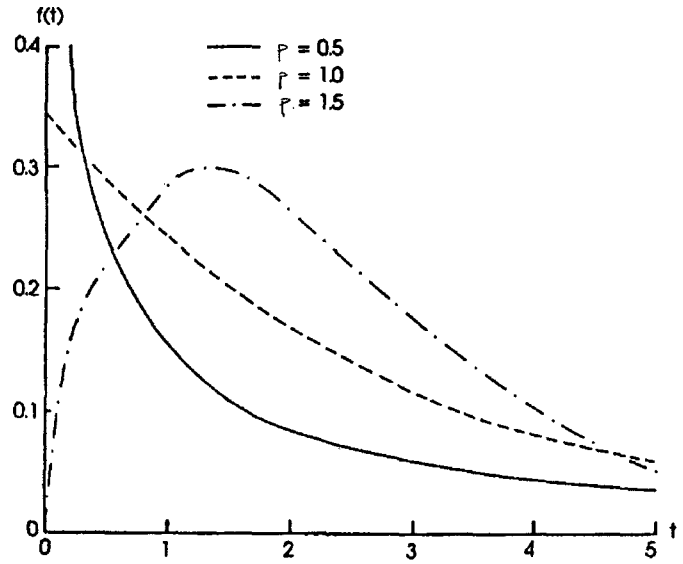
Weibull-jakauma on saanut nimensä ruotsalaisen fyysikon Waloddi Weibullin mukaan, joka käytti jakaumaa erilaisten materiaalien murtumislujuuksien kuvaamiseen.

[1] On muistettava, että taustalla on oletus siitä, että havaitut työttömyyden kestot ovat realisaatioita samasta todennäköisyysjakaumasta, mikä ei yleensä pidä paikkaansa.

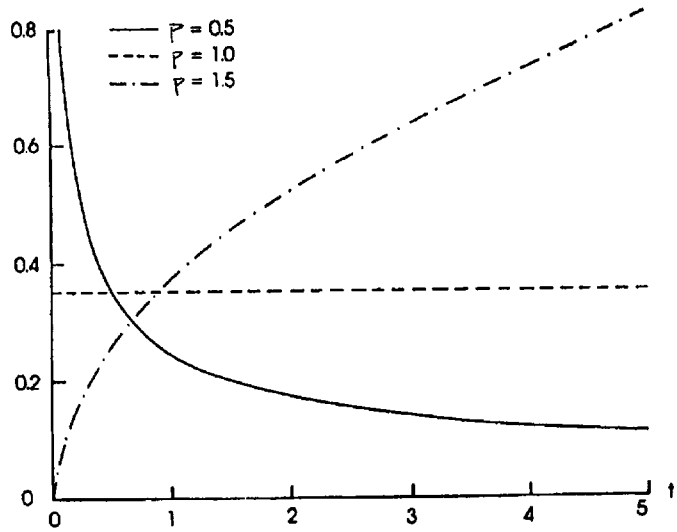
Weibull-jakauma on yleisimmin käytetty työttömyyden kestoa kuvaava jakauma. Jakauma on kaksiparametrisena eksponenttijakaumaa joustavampi ja sen esitysmuodot ovat silti yksinkertaisia. Weibull-jakautuneen satunnaismuuttujan tiheysfunktio, kertymäfunktio, eloonjäämisfunktio ja hasardifunktio ovat:

$$\begin{aligned} f(t) &= \lambda p(\lambda t)^{p-1} \exp[-(\lambda t)^p] \\ F(t) &= 1 - \exp[(\lambda t)^p] \\ S(t) &= \exp[-(\lambda t)^p] \\ h(t) &= \lambda p(\lambda t)^{p-1}. \end{aligned}$$

Hasardifunktion muoto riippuu parametrista p. Hasardifunktio on monotonisesti kasvava, kun $p > 1$ ja monotonisesti vähenevä, kun $p < 1$. Eksponenttijakauma saadaan Weibull-jakauman erikoistapauksena, kun $p = 1$. Parametri λ on ns. skaalaparametri, joka muuttaa aika-akselin mittayksikköä. $\ln(-\ln S(t)) = p(\ln t + \ln \lambda)$ on $\ln t$:n lineaarinen funktio, joten pisteiden $(\ln t_i, \ln(-\ln S(t_i)))$(missä t_i, $i = 1, \ldots, n$, ovat havaitut työttömyyden kestot), tulisi sijaita likimain suoralla, mikäli työttömyyden kesto on Weibull-jakautunut.



Kuva 4.1: Weibull-jakauman tiheysfunktioita p:n arvoilla 0.5, 1 ja 1.5. Lähde: Lancaster (1990) s.37.



Kuva 4.2: Weibull-jakauman hasardifunktioita p:n arvoilla 0.5, 1 ja 1.5. Lähde: Lancaster (1990) s. 36.

4.3 Gammajakauma

Gammajakauma on, kuten Weibull-jakauma, eksponenttijakauman kaksiparametrinen yleistys. Gammajakauman esitysmuodot ovat:

$$\begin{aligned} f(t) &= \frac{\lambda(\lambda t)^{k-1}\exp(-\lambda t)}{\Gamma(k)} \\ F(t) &= I(k,\lambda t) \\ S(t) &= 1 - I(k,\lambda t) \\ h(t) &= \frac{\lambda(\lambda t)^{k-1}\exp[-\lambda t](\Gamma(k))^{-1}}{1 - I(k,\lambda t)}, \end{aligned}$$

missä $I(k,\lambda t)$ on epätäydellinen gammafunktio.[2]

λ on jakauman skaalaparametri ja k muotoparametri. Eksponenttijakauma saadaan

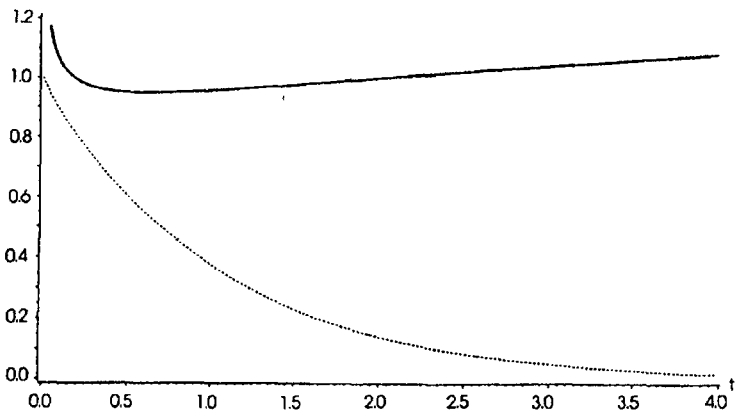
[2]Epätäydellinen gammafunktio on

$$I(k,x) = 1/\Gamma(k) \times \int_0^x u^{k-1}\exp(-u)\,du. \tag{4.1}$$

Tämä on yksiparametrisen gammajakauman ($\lambda = 1$) kertymäfunktio. Jos T noudattaa gammajakaumaa parametrein λ, k, noudattaa λT yksiparametrista gammajakaumaa parametrilla k.

gammajakauman erikoistapauksena, kun $k = 1$. Gammajakauman hasardifunktio kasvaa monotonisesti lähtien nollasta, kun $k > 1$ ja vähenee monotonisesti ääretömästä, kun $k < 1$. Molemmissa tapauksissa hasardifunktio lähestyy λ:aa, kun $t \to \infty$.

4.4 Yleistetty gammajakauma

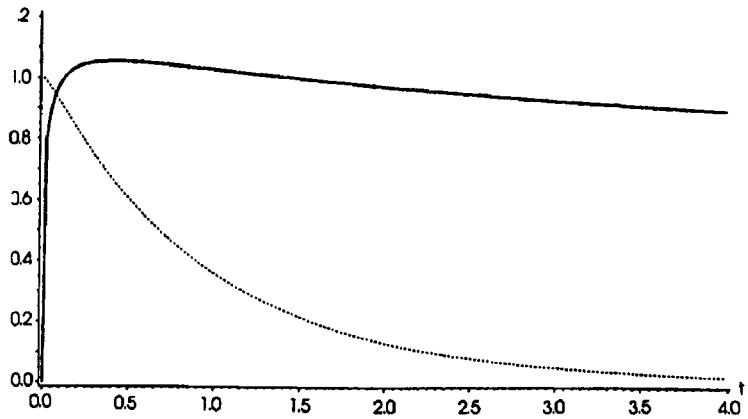


Kuva 4.3: Yleistetyn gammajakauman hasardifunktio ja eloonjäämisfunktio. $p = 1.2$ ja $k = 0.7$ Lähde: Lancaster (1990) s. 39.

Yleistetty gammajakauma on joustava kolmiparametrinen jakauma, joka sisältää erikoistapauksinaan kaikki edellä mainitut jakaumat. Jakauman esitysmuodot ovat:

$$\begin{aligned} f(t) &= \frac{\lambda p(\lambda t)^{kp-1}\exp[-(\lambda t)^p]}{\Gamma(k)} \\ F(t) &= I(k,(\lambda t)^p) \\ S(t) &= 1 - I(k,(\lambda t)^p) \\ h(t) &= \frac{\lambda p(\Gamma(k))^{-1}(\lambda t)^{kp-1}\exp[-(\lambda t)^p]}{1 - I(k,(\lambda t)^p)}. \end{aligned}$$

Eksponenttijakauma saadaan yleistetyn gammajakauman erikoistapauksena, kun $p = k = 1$. Weibull-jakaumaan tai gammajakaumaan päädytään, kun vastaavasti $k = 1$ tai $p = 1$. Yleistetyn gammajakauman voidaan osoittaa lähestyvän log-normaalia jakaumaa, kun $k \to \infty$. Jakauman hasardifunktio voi olla hyvin monimuotoinen. Yleistetty gammajakauma on hyödyllinen, kun tutkittavan satunnaismuuttujan jakaumasta ei ole ennakkotietoja ja lähtökohdaksi halutaan valita



Kuva 4.4: Yleistetyn gammajakauman hasardifunktio ja eloonjäämisfunktio. $p = 0.8$ ja $k = 1.5$ Lähde: Lancaster (1990) s. 40.

riittävän yleinen jakauma. Parametrirajoituksia testaamalla voidaan tutkia, voidaanko satunnaismuuttujan todennäköisyysjakaumaa kuvata jollain yleistetyn gammajakauman sisältämällä yksinkertaisemmalla jakaumalla.

Luku 5

Selittävät tekijät elinaikamalleissa

Edellä oletettiin, että tutkittava populaatio on homogeeninen, eli että havaitut populaation jäsenien työttömyyden kestot ovat samoin jakautuneiden satunnaismuuttujien realisaatioita. Oletus on useimmiten epärealistinen. Työnhakijat eroavat toisistaan työttömyyden kestoon vaikuttavien ominaisuuksien suhteen. Tällaisia ominaisuuksia voivat olla esimerkiksi ikä, sukupuoli, koulutus ja työkokemus. Havaitut työttömyyden kestot t_i, $i = 1, \ldots, n$ ovat siis yleensä realisaatioita eri todennäköisyysjakaumista. Työnhakijoiden heterogeenisuutta pyritään kontrolloimaan sallimalla työttömyyden keston jakauman riippua työttömyyden kestoa selittävistä tekijöistä. Havaitut työttömyyden kestot oletetaan realisaatioiksi T:n (satunnaismuuttuja työttömyyden kesto) ehdollisista jakaumista, ehdolla selittävät tekijät. Esimerkiksi $h(t; z)$ kuvaa tällöin niiden yksilöiden hasardifunktiota, jotka ovat homogeenisia selittävien tekijöiden z suhteen. Ehdolliset jakaumat poikkeavat toisistaan sen jakaumaparametrin (tai parametrien) osalta, jonka oletetaan riippuvan selittävistä tekijöistä.

Työttömyyden kestoa selittävät tekijät voidaan jakaa ajassa vaihteleviin ja vakioihin selittäviin tekijöihin. Yksinkertainen esimerkki ajassa vaihtelevasta selittävästä tekijästä on työnhakijan ikä. Toinen ajassa vaihteleva, työttömyyden kestoa selittävä tekijä on työvoiman kysyntä. Joidenkin muuttujien ajallinen vaihtelu liittyy nimenomaan työttömyyden kestoon eikä, kuten edellisissä esimerkeissä, kalenteriaikaan. Tällainen kestoriippuva muuttuja on esim. työttömyyskorvausten taso. Useiden selittävien tekijöiden ajallinen muuttuminen on kuitenkin niin hidasta, että niitä voidaan pitää vakioina, etenkin jos tarkasteltavat työttömyysjaksot ovat melko lyhyitä. Tässä tutkielmassa oletetaan, että selittävät tekijät ovat vakioita työttömyyden keston suhteen.

Elinaikamallit voidaan jakaa selittävien tekijöiden vaikutustavan perusteella kahteen

luokkaan: kiihdytetyn elinajan malleihin (accelerated failure time models) ja suhteellisten hasardien malleihin (proportional hazard models). Kiihdytetyn elinajan malleissa selittävät tekijät vaikuttavat multiplikatiivisesti työttömyyden kestoon, kun taas suhteellisten hasardien malleissa multiplikatiividen vaikutus kohdistuu työttömyyden keston hasardifunktioon; työllistymisintensiteettiin. Sekä kiihdytetyn elinajan että suhteellisten hasardien mallit voidaan esittää lineaarisina regressiomalleina. Kiihdytetyn elinajan malleissa työttömyyden keston luonnollinen logaritmi riippuu selittäjistä (parametrien suhteen) lineaarisen funktion kautta. Suhteellisten hasardien malleissa samanlainen riippuvuus voidaan esittää kumulatiivisen perushasardin luonnollisella logaritmilla muunnetulle työttömyyden kestolle. Lineaaristen regressiomallien estimointimenetelmä, pienimmän neliösumman menetelmä, ei ole kuitenkaan käyttökelpoinen sensuroitujen havaintojen ja mahdollisten työttömyysjakson aikana vaihtelevien selittävien tekijöiden vuoksi.

Usein ei kaikkia työttömyyden kestoon vaikuttavia tekijöitä tunneta eivätkä siten käytettävissä olevat selittävät tekijät pysty kontrolloimaan kaikkea populaation heterogeenisuutta. Populaatioon selittävien tekijöiden kontrolloiman vaihtelun jälkeen jäävää heterogeenisuutta sanotaan havaitsemattomaksi heterogeenisuudeksi (unobserved heterogeneity). Havaitsematonta heterogeenisuutta pyritään yleensä kontrolloimaan virhetermillä, jonka oletetaan kuvaavan kaikkien poisjääneiden selittävien tekijöiden vaikutusta työttömyyden kestoon. Havaitsemattoman heterogeenisuuden vaikutusta työttömyyden kestoa kuvaaviin malleihin tarkastellaan kappaleessa 5.4.

5.1 Kiihdytetyn elinajan mallit

Kiihdytetyn elinajan mallissa mielivaltaisen yksilön i työttömyyden kesto T_i riippuu ns. **referenssihenkilön** työttömyyden kestosta T_0 seuraavalla tavalla:

$$T_i = \frac{T_0}{\psi(z_i)}, \tag{5.1}$$

missä $\psi(z)$ on jokin työttömyyden kestoa selittävistä tekijöistä riippuva, positiivisia arvoja saava funktio, jolle pätee $\psi(0) = 1$. Referenssihenkilöllä tarkoitetaan henkilöä, jonka selittävien tekijöiden vektori on nollavektori, ts. kaikki selittävät tekijät saavat arvon nolla. Muuttujien arvot mitataan usein poikkeamina keskiarvostaan, jolloin referenssihenkilö kuvaa aineiston keskimääräiset ominaisuudet omaavaa henkilöä (indikaattorimuuttujien osalta referenssihenkilö kuuluu ns. vertailuryhmään). yksilön i työttömyyden kesto on siis jokin vakio kertaa T_0, jossa vakio määräytyy selittävien tekijöiden z_i perusteella. Jos h_0, f_0 jaS_0 kuvaavat referenssihenkilön työttömyyden keston jakaumaa, voidaan yksilön i työttömyyden keston jakauma esittää

seuraavalla tavalla [1]

$$h(t;z_i) = h_0(t\,\psi(z_i))\,\psi(z_i) \quad (5.2)$$
$$f(t;z_i) = f_0(t\,\psi(z_i))\,\psi(z_i) \quad (5.3)$$
$$S(t;z_i) = S_0(t\,\psi(z_i)). \quad (5.4)$$

Kiihdytetyn elinajan malleissa selittävien tekijöiden voidaan ajatella muuttavan aika-akselin mitta-asteikkoa siten, että ajan kuluminen hidastuu tai nopeutuu suhteessa referenssihenkilön aikaan. Jos $\psi(z_i) < 1$, on $T_i > T_0$ ja ajan kuluminen hidastuu suhteessa referenssihenkilöön. Vastaavasti jos $\psi(z_i) > 1$, ajan kuluminen nopeutuu. Toisin ilmaistuna, selittävät tekijät muuttavat yksilöiden etenemisnopeutta aika-akselilla. Otetaan yhtälöstä 5.1 puolittain luonnolliset logaritmit ja määritellään $\mu_0 = E(\ln T_0)$, jolloin $\ln T_0 = \mu_0 + e$. Tällöin

$$\ln T = \mu_0 - \ln \psi(z;\beta) + e. \quad (5.5)$$

e on z:stä riippumaton virhetermi, jonka odotusarvo on nolla. Jos valitaan $\psi(z) = \exp(z\prime\beta)$ (β on estimoitavien regressiokertoimien vektori), riipuu työttömyyden keston luonnollinen logaritmi selittävistä tekijöistä (parametrien suhteen) lineaarisen funktion kautta:

$$\ln T = \mu_0 - z\prime\beta + e. \quad (5.6)$$

Virhetermin e jakaumaa varioimalla saadaan työttömyyden kestolle erilaisia täysin parametroituja regressiomalleja. Kiihdytetyn elinajan mallit ovat käytännön sovellutuksissa harvinaisia. Tämä johtunee siitä, että on vähän tutkimusongelmia, joissa olisi luontevaa olettaa selittävien tekijöiden kiihdyttävän tai hidastavan ajan kulumisen nopeutta.

5.2 Suhteellisten hasardien mallit

Suhteellisten hasardien mallissa selittävien tekijöiden vaikutus työttömyyden kestoon spesifioidaan hasardifunktion kautta. Hasardifunktion oletetaan koostuvan kahdesta osasta, joista toinen kuvaa hasardifunktion riippuvuutta työttömyyden kestosta ja toinen riippuvuutta selittävistä tekijöistä:

$$h(t;z) = h_0(t)\psi(z). \quad (5.7)$$

$h_0(t)$ on ns. **perushasardifunktio**, joka on selittävistä tekijöistä riippumaton. Kuten kiihdytetyn elinajan malleissa, valitaan $\psi(z)$ yleensä siten, että $\psi(0) = 1$.

[1] Jos satunnaismuuttujan X tiheysfunktio on $f(x)$, on sen funktion $Y = y(X)$ (oletetaan, että funktio on differentioituva ja monotoninen) tiheysfunktio $g(y) = f(x(y))\frac{dx(y)}{dy}$.

Tällöin, koska referenssihenkilölle $z = 0$, kuvaa perushasardifunktio referenssihenkilön hasardifunktiota: $h(t;0) = h_0(t)$. $\psi(z)$ kuvaa ominaisuudet (selittäjävektorin) z omaavan henkilön työllistymisintensiteetin suhdetta referenssihenkilön työllistymisintensiteettiin:

$$\frac{h(t;z)}{h(t;0)} = \frac{h_0(t)\psi(z)}{h_0(t)} = \psi(z). \tag{5.8}$$

Suhteellisten hasardien mallissa selittävät tekijät kasvattavat tai pienentävät työllistymisintensiteettiä suhteessa referenssihenkilöön riippuen siitä, onko $\psi(z)$ suurempi vai pienempi kuin yksi.

Ominaisuudet z omaavan henkilön työttömyyden keston tiheysfunktio ja eloonjäämisfunktio voidaan suhteellisten hasardien mallissa esittää seuraavalla tavalla:

$$S(t;z) = \exp\left(-\int_0^t h(u;z)du\right) = \exp\left(-\int_0^t h_0(u)du\,\psi(z)\right) \tag{5.9}$$

$$= \exp\left(-\int_0^t h_0(u)du\right)^{\psi(z)} = S_0(t)^{\psi(z)} \tag{5.10}$$

$$f(t;z) = h(t;z)S(t;z) = \psi(z)h_0(t)S_0(t)^{\psi(z)}, \tag{5.11}$$

missä $S_0(t)$ on referenssihenkilön eloonjäämisfunktio.

Suhteellisten hasardien mallissa kahden mielivaltaisen yksilön i ja j hasardifunktioiden suhde on työttömyyden kestosta riippumaton:

$$\frac{h(t;z_i)}{h(t;z_j)} = k_{ij} \;\; \forall i, j = 1, \ldots, n. \tag{5.12}$$

Koska perushasardifunktio $h_0(t)$ on kaikille yksilöille sama, supistuu tämä työttömyyden kestosta riippuva osa hasardifunktioiden suhteesta pois. Jos esim. $k = 3$, on yksilön i työllistymisintensiteetti ja siten työllistymistodennäköisyys kolminkertainen yksilöön j verrattuna. k_{ij}:ta kutsutaan yksilöiden i ja j väliseksi **riskisuhteeksi**. Jos osa selittävistä tekijöistä on työttömyyden keston suhteen vaihtelevia, ei tämä ns. **proportionaalisuusoletus** pidä enää paikkaansa. Tästä huolimatta myös työttömyyden keston suhteen vaihtelevia muuttujia sisältäviä malleja, joissa hasardifunktio separoituu toisaalta työttömyyden kestosta ja toisaalta selittävistä tekijöistä riippuvaan osaan, kutsutaan suhteellisten hasardien malleiksi.

Funktioksi $\psi(z)$ valitaan tavallisesti, kuten kiihdytetyn elinajan malleissa, log-lineaarinen $\exp(z\prime\beta)$. Valinta takaa sen, että hasardifunktio on aina positiivinen. "Suuri" regressiokertoimen estimaatti merkitsee suhteellisten hasardien mallissa suurta työllistymisintensiteettiä ja siten nopeaa työllistymistä. Kiihdytetyn elinajan malleissa regressiokertoimet vaikuttavat päinvastaiseen suuntaan: suuri regressiokertoimen estimaatti merkitsee pitkää työttömyyden kestoa ja alhaista työllistymisintensiteettiä.

Kun $\psi(z) = \exp(z\prime\beta)$, voidaan suhteellisten hasardien malli esittää lineaarisena regressiomallina. Integroimalla ja ottamalla puolittain luonnollinen logaritmi yhtälöstä 5.7 huomataan, että kumulatiivisen perushasardifunktion luonnollisella logaritmilla muunnettu satunnaismuuttuja työttömyyden kesto; $\ln H_0(T)$ riippuu selittävistä tekijöistä (parametrien suhteen) lineaarisen funktion kautta:

$$-\ln \int_0^T h_0(u)du = z\prime\beta - \ln \int_0^T h(u;z)du \tag{5.13}$$

Merkitään

$$T^* = -\ln \int_0^T h_0(u)du = -\ln H_0(T) \tag{5.14}$$

$$e = -\ln \int_0^T h(u;z)du = -\ln H(T;z). \tag{5.15}$$

Tällöin yhtälö 5.13 voidaan esittää seuraavalla tavalla:

$$T^* = z\prime\beta + e, \tag{5.16}$$

missä e on ääriarvojakautunut virhetermi [2].

Cox (1972) osoitti, että suhteellisten hasardien mallissa regressiokertoimet β voidaan estimoida spesifioimatta perushasardifunktion muotoa. Tällöin ei myöskään tutkittavan satunnaismuuttujan jakaumasta tarvitse tehdä mitään oletuksia. Suhteellisten hasardien malleja, joissa hasardifunktion aikariippuvasta osasta ei tehdä mitään oletuksia, kutsutaan semiparametrisiksi malleiksi. Semiparametriset ovat suosittuja mm. lääketieteessä ja ovat viime vuosina yleistyneet myös taloustieteessä. Semiparametrisiin malleihin perehdytään seuraavassa luvussa.

5.3 Kiihdytetyn elinajan mallin ja suhteellisten hasardien mallin vertailua

Kiihdytetyn elinajan ja suhteellisten hasardien mallit eroavat toisistaan selittävien tekijöiden vaikutustavan suhteen. Kiihdytetyn elinajan malleissa selittävät tekijät vaikuttavat multiplikatiivisesti työttömyyden kestoon, kun taas suhteellisten hasardien malleissa multiplikatiivinen vaikutus kohdistuu työttömyyden keston hasardifunktioon; työllistymisintensiteettiin. Käsitteet työttömyyden kesto ja työllistymisintensiteetti ovat läheisessä yhteydessä toisiinsa: suuri työllistymisintensiteetti merkitsee todennäköistä *nopeaa* työllistymistä ja siis lyhyttä työttömyyden kestoa. Sekä

[2] Kappaleessa 2.1 todettiin, että kumulatiivisella hasardifunktiolla muunnettu satunnaismuuttuja T; $H(T)$ noudattaa standardoitua eksponenttijakaumaa T:n jakaumasta riippumatta. Tällöin $e = -\ln H(T)$ noudattaa standardoitua ääriarvojakaumaa.

kiihdytetyn elinajan että suhteellisten hasardien mallit voidaan esittää lineaarisina regressiomalleina. Kiihdytetyn elinajan mallissa työttömyyden keston logaritmi $\ln T$ voidaan esittää selittävien tekijöiden lineaarisena funktiona. Virhetermin jakaumaa varioimalla saadaan erilaisia työttömyyden kestoa kuvaavia täysin parametroituja kiihdytetyn elinajan malleja. Suhteellisten hasardien mallissa kumulatiivisen perushasardifunktion luonnollisella logaritmilla muunnettu työttömyyden kesto $\ln H_0(T)$ riippuu selittävistä tekijöistä lineaarisen funktion kautta. Virhetermin jakauma on aina standardoitu ääriarvojakauma, mutta kumulatiivinen hasardifunktio $H(T)$ voidaan valita vapaasti.

Weibull-jakaumaan perustuva työttömyyden keston regressiomalli spesifioidaan tavallisesti siten, että jakauman skaalaparametri λ riippuu selittävistä tekijöistä eli $\lambda = \lambda(z)$ (useimmiten $\lambda(z) = \exp(z'\beta)$. Hasardifunktio on tällöin

$$h(t;z) = pt^{p-1}\lambda(z)^p. \tag{5.17}$$

Koska

$$pt^{p-1}\lambda(z)^p = p\lambda(z)(t\lambda(z))^{p-1} = \lambda(z)h_0(t\lambda(z)), \tag{5.18}$$

missä $h_0(t) = pt^{p-1}$, kuuluu Weibull-jakaumaan perustuva malli kiihdytetyn elinajan malliluokkaan (vrt. yhtälö 5.2). Merkitään kiihdytetyn elinajan mallin selittävien tekijöiden funktiota $\lambda_{aft}(z)$:lla ja suhteellisten hasardien mallin vastaavaa funktiota $\lambda_{ph}(z)$:lla. Kun määritellään $\lambda_{ph}(z) = \lambda_{aft}(z)^p$, havaitaan, että Weibull-malli on myös suhteellisten hasardien muotoa, jossa hasardifunktio separoituu työttömyyden kestosta riippuvaan ja selittävistä tekijöistä riippuvaan osaan:

$$h(t;z) = pt^{p-1}\lambda_{ph}(z) = h_0(t)\lambda_{ph}(z), \tag{5.19}$$

missä $h_0(t) = pt^{p-1}$. Jos lisäksi $\lambda_{aft}(z) = \exp(\beta_{aft}'z)$, on $\lambda_{ph}(z) = \exp(\beta_{aft}'z)^p = \exp(p\beta_{aft}'z) = \exp(\beta_{ph}'z)$, missä $\beta_{ph} = p\beta_{aft}$. Voidaan osoittaa (ks. Cox & Oakes 1984 s. 71), että Weibull-jakauma on ainoa jakauma, johon perustuvat kiihdytetyn elinajan ja suhteellisten hasardien mallit ovat samat (kun parametri p on selittävistä tekijöistä riippumaton). Tämä pätee tietenkin myös Weibull-jakauman erikoistapaukselle eksponenttijakaumalle.

5.4 Havaitsematon heterogeenisuus

Usein tietoja kaikista niistä tekijöistä, joiden oletetaan vaikuttavan työttömyyden kestoon, ei ole saatavilla. Osaa tekijöistä voi olla vaikea mitata, kuten esim. motivoituneisuutta työnetsintään. Sitä populaation heterogeenisuutta, jota käytettävissä olevat selittävät tekijät eivät pysty kontrolloimaan, sanotaan **havaitsemattomaksi**

heterogeenisuudeksi. Havaitsematon heterogeenisuus johtuu siis yksilöiden eroavuudesta niiden työttömyyden kestoon vaikuttavien tekijöiden suhteen, jotka eivät sisälly selittäjävektoriin.

Poisjääneet työttömyyden kestoa selittävät tekijät aiheuttavat hasardifunktioon negatiivista duraatioriippuvuutta. Seuraava asiaa havainnollistava esimerkki on Kieferin (1988) artikkelista. Tarkastellaan poisjääneiden selittävien tekijöiden vaikutusta hasardifunktioon populaatiossa, jossa työttömyyden kestot ovat eksponenttijakautuneita. Oletetaan yksinkertaisuuden vuoksi, että populaation jäsenet eroavat toisistaan ainoastaan yhden työllistymiseen vaikuttavan tekijän suhteen. Olkoon tämä tekijä työnhakijan sukupuoli. Oletetaan, että osuus p populaatiosta on naisia (ryhmä 1) ja osuus $(1-p)$ miehiä (ryhmä 2). Työttömyyden kestot ryhmissä 1 ja 2 ovat peräisin eri todennäköisyysjakaumista. Ryhmien 1 ja 2 työttömyyden keston tiheysfunktiot ovat

$$f_1(t) = f(t \mid \text{ryhmä } 1) = \lambda_1 e^{-\lambda_1 t} \text{ ja} \tag{5.20}$$

$$f_2(t) = f(t \mid \text{ryhmä } 2) = \lambda_2 e^{-\lambda_2 t} \tag{5.21}$$

ja hasardifunktiot vastaavasti λ_1 ja λ_2. Hasardifunktio on molemmissa ryhmissä työttömyyden kestosta riippumaton. Kuvatkoon indikaattorimuuttuja z työnhakijan sukupuolta siten, että $z = 0$, jos työnhakija on nainen ja $z = 1$, jos työnhakija on mies. Mielivaltaisen populaation jäsenen i hasardifunktio voidaan tällöin esittää seuraavalla tavalla:

$$h(t \mid z_i) = \lambda_1 + z_i(\lambda_2 - \lambda_1) \tag{5.22}$$

Jos työnhakijoiden sukupuoli on tuntematon, ei tiedetä, kummasta todennäköisyysjakaumasta kukin havaittu työttömyysjakso on peräisin. Tällöin työttömyysjaksot ovat peräisin sekajakaumasta

$$f(t) = pf_1(t) + (1-p)f_2(t) \tag{5.23}$$

ja hasardifunktioiden λ_1 ja λ_2 sijaan on tyydyttävä estimoimaan sekajakauman hasardifunktio:

$$h(t) = \frac{f(t)}{S(t)} = \frac{p\lambda_1 e^{-\lambda_1 t} + (1-p)\lambda_2 e^{-\lambda_2 t}}{pe^{-\lambda_1 t} + (1-p)e^{-\lambda_2 t}}. \tag{5.24}$$

Lausekkeesta havaitaan, että sekajakauman hasardifunktio on itse asiassa naisten ja miesten hasardifunktioiden painotettu keskiarvo (painot $w_1 = p\exp^{-\lambda_1 t}$ ja $w_2 = (1-p)\exp^{-\lambda_2 t}$). Sekajakauman hasardifunktio riippuu työttömyyden kestosta, toisin kuin ryhmien 1 ja 2 hasardifunktiot. Lisäksi $\partial h(t)/\partial t < 0$ eli hasardifunktio on vähenevä. Yleisesti, havaitsematon heterogeenisuus populaatiossa vaimentaa positiivista ja vahvistaa negatiivista duraatioriippuvuutta T:n jakaumasta riippumatta, mikäli poisjääneet selittävät tekijät vaikuttavat hasardifunktioon multiplikatiivisesti. Havaitsematon heterogeenisuus vaikuttaa regressiokertoimien estimaatteihin siten, että ne ovat harhaisia kohti nollaa (kun selittävien tekijöiden vaikutus hasardifunktioon on multiplikatiivinen). (Lancaster 1990.)

Tulos on intuitiivisesti ymmärrettävä: oletetaan, että naisten työllistymisintensiteetti on pienempi kuin miesten. Tällöin $\lambda_1 < \lambda_2$ ja naisten työttömyysjaksot ovat keskimäärin pidempiä kuin miesten. Miehet työllistyvät siis naisia nopeammin ja ajan kuluessa naisten osuus työttömistä (riskijoukosta) kasvaa. Koska $\lambda_1 < \lambda_2$, näkyy naisten osuuden kasvu riskijoukossa hasardifunktion vähenemisenä eli negatiivisena duraatioriippuvuutena.

Havaitsematonta heterogeenisuutta mallinnetaan yleensä hasardifunktioon multiplikatiivisesti vaikuttavalla virhetermillä, jonka oletetaan kuvaavan kaikkien poisjääneiden selittävien tekijöiden yhteisvaikutusta hasardifunktioon. Tavallinen hasardifunktion spesifikaatio on

$$h(t; z, v) = v\, h_0(t) \exp(\beta' z_1), \tag{5.25}$$

missä $v = \exp(\beta' z_2)$ ja $z = z_1 + z_2$ on kaikki työttömyyden kestoa selittävät tekijät sisältävä vektori. z_1 sisältää työttömyyden kestoa selittävät tunnetut ja z_2 tuntemattomat tekijät. Tuntemattomien selittävien tekijöiden vaikutus spesifioidaan siis samalla tavalla kuin tunnettujen; selittävät tekijät siirtävät perushasardia vakiokertoimen $\exp(\beta' z)$ verran. Samoin kuin tunnetut työttömyyden kestoa selittävät tekijät, voivat myös havaitsematonta heterogeenisuutta aiheuttavat tekijät olla aikariipuvia. Eräs työttömyyden kestosta riippuva tuntematon tekijä voisi olla esim. motivoituneisuus työn etsintään: työttömyyden pitkittyessä voi työnhakijan motivaatio etsiä työtä vähentyä. Toisaalta juuri ennen ansiosidonnaisen päivärahan muuttumista peruspäivärahaksi voi työnetsintämotivaatio kasvaa, jos työnhakija haluaa välttää tulojen pienenemisen.

Hasardifunktioon multiplikatiivisesti vaikuttava virhetermi ei ole ainoa mahdollinen tapa kuvata poisjääneiden selittävien tekijöiden vaikutusta (sama pätee tietysti myös tunnetuille selittäville tekijöille). Coxin mukaan (ks. Lancaster & Nickell 1980) virhetermin vaikutus voitaisiin spesifioida esim. seuraavilla tavoilla:

$$h(t; z, v) = h(t; z_1) + v \quad \text{tai} \tag{5.26}$$

$$h(t; z, v) = (h(z_1) + v) h_0(t) \tag{5.27}$$

Jos havaitsemattoman heterogeenisuuden vaikutusmekanismi on yllä kuvatun kaltainen, ei sen mallintamatta jättäminen välttämättä aiheuta hasardifunktioon negatiivista duraatioriippuvuutta (Lancaster & Nickell 1979).

Poisjääneiden selittävien tekijöiden vaikutusta kuvaava virhetermi on yleensä oletettu gammajakautuneeksi. Heckmanin ja Singerin (1984) mukaan erilaiset virhetermin jakaumavalinnat saattavat tuottaa varsin erilaisia parametri- ja regressiokertoimien estimaatteja. He esittivät menetelmän, jolla havaitsemattoman heterogeenisuuden huomioivia malleja voidaan estimoida spesifioimatta virhetermin jakaumaa etukäteen. Menetelmässä virhetermille estimoidaan samanaikaisesti työttömyyden keston

jakaumaparametrien ja regressiokertoimien kanssa diskreetti jakauma, joka koostuu yleensä muutamasta todennäköisyysmassapisteestä. Menetelmää on kuitenkin sovellettu vähän luultavasti koska se on sekä teoreettisesti että laskennallisesti varsin vaativa.

Yleinen spesifikaatio havaitsemattoman heterogeenisuuden huomioivissa malleissa ollut Weibull-jakautunut työttömyyden kesto ja gammajakautunut virhetermi. Pudneyn (1989) mukaan virhetermin tehtävä tällaisissa malleissa on lähinnä liian rajoittavan jakaumaoletuksen lieventäminen (Weibull-jakauma sallii ainoastaan joko monotonisesti kasvavan tai vähenevän hasardifunktion). Vaihtoehtoinen tapa spesifioida riittävän yleinen malli on käyttää semiparametrisia eli osittain parametroituja menetelmiä, joissa työttömyyden keston jakaumaa ei tarvitse spesifioida. Ongelmat täysin parametroitujen mallien spesifioinnissa ovat johtaneet semiparametristen menetelmien yleistymiseen taloustieteessä viime vuosina. Semiparametriset mallit tuovat ratkaisun työttömyyden keston jakauman valintaongelmaan. Semiparametristen mallien käyttö ei kuitenkaan poista sitä seikkaa, että käytettävissä olevat aineistot ovat usein puutteellisia eivätkä sisällä tietoja kaikista tutkittavan työttömyyden kestoon vaikuttavista tekijöistä. Myös semiparametrisissa malleissa on kokeiltu havaitsemattoman heterogeenisuuden huomioimista virhetermillä (ks. esim. Meyer 1990, Sueyoshi 1992). Näissä malleissa virhetermi on oletettu gammajakautuneeksi.

Empiiriset tutkimustulokset tulokset viittaavat siihen, että mitä paremmin tutkittavan populaation heterogeenisuus on otettu huomioon, sitä vähemmän työttömyyden keston hasardifunktiossa esiintyy negatiivista duraatioriippuvuutta. Meyer (1990) käytti semiparametristä *luokiteltujen tapahtuma-aikojen mallia* (ks. kpl 6.3) ja 11 selittävää tekijää työttömyyden keston kuvaamiseen. Tällä spesifikaatiolla ei duraatioriippuvuutta löytynyt. Mahdollisten poisjääneiden selittävien tekijöiden huomioiminen gammajakautuneella virhetermillä johti selvään positiiviseen duraatioriippuvuuteen. Regressiokertoimien estimaatit olivat itseisarvoltaan suurempia mallissa, jossa havaitsematon heterogeenisuus otettiin huomioon. Liljan (1992) käyttämässä semiparametrisessa mallissa oli peräti 18 selittävää tekijää. Työnhaun erilaiset päättymissyyt huomioitiin ja mallit estimoitiin erikseen ansiosidonnaista päivärahaa ja peruspäivärahaa saaville sekä henkilöille, jotka eivät olleet oikeutettuja työttömyysturvaan. Selvää duraatioriippuvuutta ei havaittu millään ryhmällä. Selittävien tekijöiden oletettiin kontrolloivan täysin populaation heterogeenisuuden. Lancaster ja Nickell (1980) käyttivät mallintamisessa Weibull-jakaumaa ja vain kolmea selittävää tekijää. Gammajakautuneen virhetermin lisääminen malliin lievensi hasardifunktion negatiivista duraatioriippuvuutta. Virhetermin lisääminen malliin kasvatti myös regressiokertoimien estimaattien itseisarvoja.

Luku 6

Täysin parametroidut vs. semiparametriset menetelmät

Tapahtumien sattumistodennäköisyyksien analyysissä voidaan erottaa kolme lähestymistapaa: **ei-parametriset, täysin parametroidut** ja osittain parametroidut eli **semiparametriset** menetelmät. Lähestymistavat poikkeavat toisistaan tapahtuma-ajan jakaumasta tehtävien oletusten suhteen.

Ei-parametriset menetelmät, kuten eloonjäämisfunktion estimaattorit Kaplan–Meier-estimaattori ja eloonjäämistaulu soveltuvat lähinnä aineiston alustavaan tarkasteluun. Ei-parametristen estimaattien laskemiseksi ei tarvitse spesifioida työttömyyden keston jakaumaa etukäteen. Estimaattien laskemiseksi on kuitenkin oletettava, että tarkasteltava populaatio on homogeeninen.

Täysin parametroiduissa menetelmissä työttömyyden keston oletetaan noudattavan jotain tunnettua, muutamalla parametrilla kuvattavaa jakaumaa. Aineiston heterogeenisuus otetaan huomioon sallimalla työttömyyden keston jakauman riippua selittävistä tekijöistä. Täysin parametroidut regressiomallit voidaan jakaa selittävien tekijöiden vaikutustavan perusteella kiihdytetyn elinajan malleihin ja suhteellisten hasardien malleihin.

Semiparametrisia menetelmiä voidaan pitää edellisten lähestymistapojen välimuotona. Semiparametriset regressiomallit ovat suhteellisten hasardien malleja. Suhteellisten hasardien malleissa hasardifunktio koostuu kahdesta osasta, joista toinen kuvaa hasardifunktion työttömyyden kestoon liittyvää vaihtelua ja toinen aineiston heterogeenisuuden aiheuttamaa vaihtelua mitattuna yksilökohtaisilla selittävillä tekijöillä. Jos molemmat osat esitetään parametroidussa muodossa, on kyseessä täysin parametroitu malli. Semiparametrisissa malleissa jätetään hasardifunktion kesto-

riippuvuutta kuvaava perushasardifunkito mallintamatta.(Eerola 1990.) Tällöin ei myöskään työttömyyden keston jakaumaa tarvitse lainkaan spesifioida etukäteen.

Perushasardifunktion spesifioiminen täysin parametroiduissa malleissa merkitsee sitä, että työttömyyden keston jakauma spesifioidaan jakaumaparametreja vaille etukäteen. Esimerkiksi jos perushasardifunktion oletetaan olevan muotoa pt^{p-1}, on suhteellisten hasardien malli

$$h(t;z) = pt^{p-1}\exp(z\prime\beta) \tag{6.1}$$

ja työttömyyden keston ehdollinen jakauma Weibull-jakauma, jonka sijaintiparametri λ riippuu selittävistä tekijöistä log-lineaarisen funktion kautta: $\lambda^p = \exp(z\prime\beta)$. Parametri p on Weibull-jakauman muotoparametri, jonka arvosta riippuen jakauman hasardifunktio on monotonisesti kasvava ($p > 1$), monotonisesti vähenevä ($p < 1$) tai vakio ($p = 1$).

Useiden empiiristen tutkimusten mukaan, joissa perushasardifunktio on estimoitu ei-parametrisesti tai perustuen johonkin joustavaan, monimuotoista hasardifunktion käyttäytymistä sallivaan jakaumaan, on estimoitu perushasardifunktio ei-monotoninen (Lilja 1992, Meyer 1990, Moffitt 1985, Bergström & Edin 1991). Tässä valossa yleisesti käytetty mallispesifikaatio, jossa työttömyyden kestot oletetaan Weibull-jakautuneiksi, vaikuttaa liian yksinkertaistavalta työllistymisprosessin kuvaukselta. Toisaalta Bergströmin ja Edinin (1991) estimointikokeilujen mukaan suhteellisten hasardien mallien regressiokertoimien estimaatit ovat varsin robusteja työttömyyden keston jakaumaoletuksen suhteen. Jos kuitenkin ollaan kiinnostuneita myös työllistymistodennäköisyyden vaihtelusta työttömyyden keston suhteen, on syytä pyrkiä rajoittamaan etukäteen mahdollisimman vähän työttömyyden keston jakauman muotoa. Tämä voidaan toteuttaa valitsemalla työttömyyden keston jakaumaksi jokin riittävän yleinen jakauma (esim. yleistetty gammajakauma, ks. kpl 4.4), johon perustuvat mallit saattavat tosin olla laskennallisesti varsin hankalia[1], tai käyttämällä semiparametrisia menetelmiä. Semiparametrisia Coxin mallia ja luokiteltujen tapahtuma-aikojen mallia käsitellään kappaleissa 6.2 ja 6.3.[2]

Yksinkertaisin oletus työttömyyden keston jakaumasta on eksponenttijakauma. Eksponenttijakautuneen työttömyyden keston hasardifunktio on vakio eli työllistymistodennäköisyys on työttömyyden kestosta riippumaton. Tämä on varsin rajoittava oletus, mutta jakamalla aika-akseli väleihin $[c_{m-1}, c_m)$ ja olettamalla, että hasardifunktio on vakio kunkin aikavälin sisällä, päädytään joustavaan mallityyppiin, ns. **paloittaiseen eksponenttimalliin**. Pienentämällä aikavälien pituutta saadaan yhä tarkempi approksimaatio työttömyyden keston todelliselle jakaumalle. Aineis-

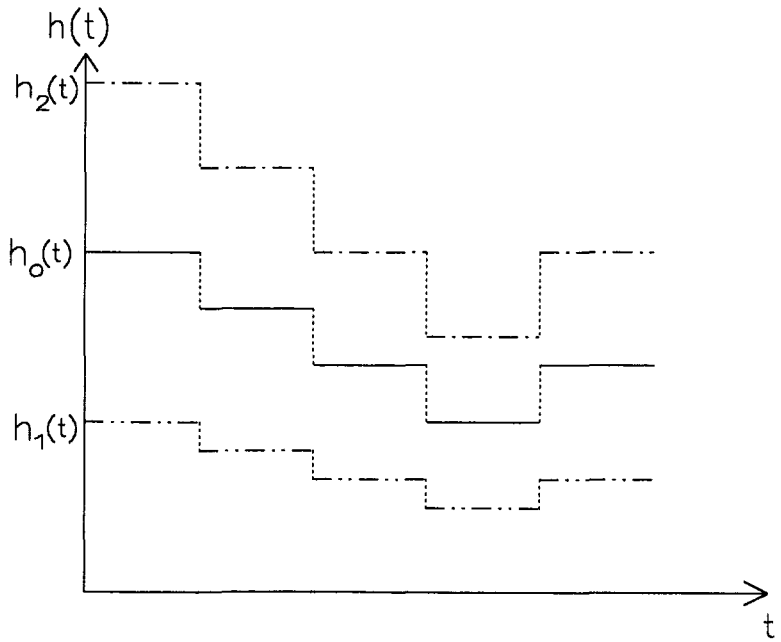
[1] Ainakaan GLIM:in, EGRET:in ja SAS:n elinaikaohjelmissa ei ole mahdollista estimoida yleistettyyn gammajakaumaan perustuvia malleja.

[2] Semiparametrisiin malleihin kuuluu lisäksi Coxin diskreettien tapahtuma-aikojen malli, ks. esim. Cox & Oakes (1984) s. 101.

ton koko asettaa kuitenkin rajat aikavälien lukumäärälle: jokainen uusi aikaväli lisää estimoitavien parametrien määrää yhdellä. Seuraavassa kappaleessa esitellään paloittainen eksponenttimalli tarkasti mitattujen työttömyyden kestojen tapauksessa. Esitys perustuu Lancasterin (1990) kirjaan.

6.1 Paloittainen eksponenttimalli

Tässä mallityypissä yksilöiden hasardifunktiot ovat paloittain vakioita ajan funktioita. Kuten suhteellisten hasardien malleissa yleensä, on perushasardifunktio $h_0(t)$ kaikille sama ja selittävät tekijät vaikuttavat hasardifunktioon multiplikatiivisesti siten, että kunkin yksilön hasardifunktion suhde perushasardiin on työttömyyden kestosta riippumaton vakio. Kuvassa 6.1 on yksilöiden 1 ja 2 paloittain vakiot hasardifunktiot $h_1(t)$ ja $h_2(t)$ sekä perushasardifunktio $h_0(t)$.



Kuva 6.1: Perushasardifunktio $h_0(t)$ sekä yksilöiden 1 ja 2 hasardifunktiot. $h_1(t) = 0.5 \times h_0(t)$, $h_2(t) = 1.5 \times h_0(t)$.

Kun työttömyysjaksojen pituudet on mitattu tarkasti, jää aikavälien määrittelemiminen tutkijan tehtäväksi. Jaetaan aika-akseli M:ään aikaväliin $[c_{m-1}, c_m), m =$

$1, \ldots, M, (c_0 = 0, c_M = \infty)$ siten, että 1. aikaväli on $[0, c_1)$ ja m:s $[c_{m-1}, c_m)$. Yksilön i hasardifunktio m:nnellä aikavälillä on:

$$h_i(t; z) = \exp(z_i\prime\beta)\exp(\lambda_m), \; i = 1, \ldots, N, \tag{6.2}$$

missä $\exp(\lambda_m)$ on m:nnen aikavälin vakio perushasardifunktio. Määritellään indikaattorimuuttujat $d_m(t)$ ja $D_m(t)$ seuraavalla tavalla:

$$d_m(t) = \begin{cases} 1, \text{ kun } c_{m-1} \leq t < c_m, \; m = 1, \ldots, M \\ 0 \text{ muulloin.} \end{cases} \tag{6.3}$$

$$D_m(t) = \prod_{j=1}^{m}[1 - d_j(t)], \; m = 1, \ldots, M-1, \; D_0(t) = 1, \; D_M(t) = 0. \tag{6.4}$$

$d_m(t)$ saa arvon yksi, jos työttömyys on päättynyt m:nnellä aikavälillä, $D_m(t)$ puolestaan saa arvon yksi, jos $t \geq c_m$ eli jos työttömyys on kestänyt vähintään ajankohtaan c_m saakka. Käyttäen määriteltyjä indikaattorimuuttujia apuna voidaan yksilön i hasardifunktio ja eloonjäämisfunktio esittää seuraavalla tavalla:

$$h_i(t) = \exp(z_i\prime\beta)\exp\left\{\sum_{m=1}^{M}\lambda_m d_m(t_i)\right\} \tag{6.5}$$

$$S_i(t) = \exp\left\{-\exp(z_i\prime\beta)\sum_{m=1}^{M} e^{\lambda_m}[(t - c_{m-1})d_m(t_i) + (c_m - c_{m-1})D_m(t_i)]\right\} \tag{6.6}$$

Yhtälössä 6.5 summalauseke $\sum_{m=1}^{M}$ käy läpi kaikki aikavälit ja indikaattori $d_m(t_i)$ osoittaa sen aikavälin, jolla t_i sijaitsee. Eloonjäämisfunktion lausekkeessa termi $\exp(z_i\prime\beta)\sum_{m=1}^{M} e^{\lambda_m}(c_m - c_{m-1})D_m(t_i)$ laskee yhteen aikavälien perushasardifunktiot eli "kumuloi" yksilön i ehdollista työllistymistodennäköisyyttä työttömäksi tulosta lähtien siihen aikaväliin saakka, jolla t_i sijaitsee; $\exp(z_i\prime\beta)e^{\lambda_m}(t - c_{m-1})$ kuvaa yksilön i työllistymistodennäköisyyttä kyseisellä aikavälillä, ehdolla, että yksilö on yhä työtön aikavälin alussa. $\sum_{m=1}^{M}[(t_i - c_{m-1})d_m + (c_m - c_{m-1})D_m(t_i)]$ on yksilön i seurannassaoloaika (työttömyyden kesto tai sensurointiaika). Paloittaiseen eksponenttimallin N:n havainnon työttömyysjaksoihin perustuva uskottavuusfunktio (kun työttömyysjaksojen pituudet tunnetaan tarkasti) on siten

$$L = \prod_{i=1}^{N}\left[[\exp(z_i\prime\beta)\exp(\sum_{m=1}^{M}\lambda_m d_m(t_i))]^{\nu_i}\right.$$
$$\left.\times \exp\left(-\exp(z_i\prime\beta)\sum_{m=1}^{M} e^{\lambda_m}[(t_i - c_{m-1})d_{im} + (c_m - c_{m-1})D_m(t_i)]\right)\right], \tag{6.7}$$

missä v_i on yksilön i sensurointi-indikaattori, jonka arvosta riippuen t_i on joko työttömyyden kesto ($v_i = 1$) tai sensurointiaika ($v_i = 0$). Estimoitavia parametreja ovat siis regressiokertoimet β sekä aikavälien vakiota perushasardifunktiota kuvaavat parametrit λ_m, $m = 1, \ldots, M$. Mitä pienempiin osiin aika-akseli jaetaan, sitä tarkemmin estimoitu perushasardifunktio approksimoi todellista perushasardifunktiota. Jos aika-akselin ositus on tiheä, voi työllistymisien lukumäärä jollain aikavälillä jäädä nollaksi, jolloin perushasardifunktion estimaatti kyseisellä aikavälillä on nolla. Valitsemalla aika-akselin osituskohdiksi työllistymisajankohdat t_j saadaan kullekin aikavälille positiivinen perushasardifunktion estimaatti. Tällaiseen malliin perustuvat regressiokertoimien suurimman uskottavuuden estimaatit ovat hyvin lähellä Coxin mallin estimaatteja (Aitkin, Anderson, Francis & Hinde 1989).

6.2 Coxin malli

Coxin v. 1972 esittämä malli kuuluu suhteellisen hasardin malleihin, joissa hasardifunktio separoituu kahteen osaan:

$$h(t; z) = h_0(t) \exp(z\prime\beta). \tag{6.8}$$

Toisin kuin täysin parametroiduissa malleissa, ei Coxin mallissa tarvitse lainkaan spesifioida hasardifunktion kestoriippuvuutta kuvaavaa perushasardifunktiota. Tällöin ei myöskään työttömyyden keston jakaumaa tarvitse spesifioida etukäteen. Coxin mallissa $h_0(t)$ on siis tuntematon työttömyyden keston funktio. Tämä tekee mallista joustavan verrattuna täysin parametroituihin malleihin, joissa perushasardifunktion muoto on aina jollain tavalla rajoitettu. Jos työttömyyden keston jakaumasta ei ole ennakkotietoja, on Coxin malli turvallinen lähtökohta, jonka avulla estimoitua perushasardifunktiota tutkimalla voidaan mahdollisesti löytää jokin työttömyyden kestoa riittävän hyvin kuvaava tunnettu jakauma. Coxin mallissa regressiokertoimien estimointi perustuu ns. **osittaisuskottavuusfunktion** (partial likelihood) maksimointiin. Osittaisuskottavuusfunktio käyttää työttömyysjaksoista hyväksi vain informaation siitä, miten selittävät tekijät vaikuttavat yksilöiden keskinäiseen työllistymisjärjestykseen, ei siis työttömyysjaksojen pituuksiin (Eerola 1990). Coxin mallissa jätetään siis osa aineiston sisältämästä informaatiosta käyttämättä. Jos työttömyyden kesto noudattaakin jotain tunnettua, muutamalla parametrilla kuvattavaa jakaumaa, on osittaisuskottavuusfunktioon perustuva estimointimenetelmä tehottomampi (parametriestimaattien varianssit suurempia) kuin tavanomaiseen uskottavuusfunktioon perustuva.

Seuraavassa kappaleessa esitellään Coxin mallin estimointimenetelmä, osittaisuskottavuusfunktio. Esitys perustuu Lancasterin (1990) kirjaan. Kappaleissa 6.2.2 ja 6.2.3 käsitellään osittaisuskottavuusfunktiota tapauksessa, jossa aineisto sisältää

sensuroituja havaintoja tai sidoksia. Coxin mallin estimointimenetelmällä jää perushasardifunktio $h_0(t)$ tuntemattomaksi. Perushasardifunktio voidaan kuitenkin estimoida jälkeenpäin käyttämällä hyväksi osittaisuskottavuusfunktion perusteella estimoituja regressiokertoimia. Kappale 6.2.4 käsittelee perushasardifunktion estimoimista Breslow:n menetelmällä.

6.2.1 Osittaisuskottavuusfunktio

Tarkastellaan N:n työnhakijan työttömyyden kestoa. Merkitään yksilön i havaittua työttömyyden kestoa t_i:llä. Työttömyyden kestoihin sisältyvä informaatio voidaan jakaa kahteen osaan: työttömyyden keston mukaan suuruusjärjestykseen järjestettyihin työnhakijoihin (kenen työttömyysjakso oli lyhyin, kenen suraavaksi lyhyin jne.) sekä suuruusjärjestykseen järjestettyihin työttömyysjaksojen pituuksiin ilman tietoa siitä, kenelle jaksot kuuluvat. Esimerkiksi tiedot kolmen työnhakijan työttömyyden kestoista

Taulu 1.
yksilön 1 työttömyys kesti 8 viikkoa
yksilön 2 työttömyys kesti 12 viikkoa
yksilön 3 työttömyys kesti 2 viikkoa

voidaan jakaa kahteen osaan seuraavalla tavalla:

Taulu 2.
yksilön 3 työttömyysjakso oli lyhyin
yksilön 1 työttömyysjakso oli seuraavaksi lyhyin
yksilön 2 työttömyysjakso oli pisin

ja

Taulu 3.
lyhyimmän työttömyysjakson pituus oli 2 viikkoa
toiseksi lyhyimmän työttömyysjakson pituus oli 8 viikkoa
pisimmän työttömyysjakson pituus oli 12 viikkoa

Taulujen 2 ja 3 tiedot vastaavat yhdessä taulun 1 tietoja; taulut 2 ja 3 tuntemalla saadaan selville esimerkin yksilöiden työttömyysjaksojen pituudet. Täysin parametroituja malleja estimoitaessa käytetään hyväksi kaikki työttömyysjaksoihin sisältyvä

tieto eli tiedot työttömyysjaksojen pituuksista sekä siitä, keille (millaisin selittävin tekijöin varustetulle yksilöille) jaksot kuuluvat. Osittaisuskottavuusfunktio käyttää nimensä mukaisesti vain osan työttömyysjaksoihin sisältyvästä informaatiosta. Parametrien estimointi perustuu pelkästään informaatioon siitä, miten selittävät tekijät vaikuttavat työnhakijoiden työttömyysjaksojen keskinäiseen järjestykseen (taulu 2), ei työttömyysjaksojen pituuksiin.

Satunnaismuuttujia $T_i, i = 1, \ldots, N$ vastaa siis N paria satunnaismuuttujia A_i, B_i. Satunnaismuuttuja A_i on i:nneksi lyhyimmän työttömyysjakson omaavan työnhakijan tunniste (esim. alkuperäisen, järjestämättömän aineiston havainnon järjestysnumero) ja B_i i:nneksi lyhyimmän työttömyysjakson pituus. Havaintojen t_i uskottavuusfunktio voidaan esittää lukuparien a_i, b_i (a_i ja b_i ovat satunnaismuuttujien A_i ja B_i, $i = 1, \ldots, N$ realisaatioita) avulla:

$$L = f(t_1, \ldots, t_N) = f(a_1, b_1, \ldots, a_N, b_N) \tag{6.9}$$

Vaikka satunnaismuuttujat T_i $i = 1, \ldots, N$ voidaan olettaa riippumattomiksi, ei riippumattomuusoletus ole voimassa pareille A_i, B_i: työnhakijan sijainti suuruusjärjestykseen järjestetyssä aineistossa riippuu muiden työnhakijoiden sijainnista. Riippuvuus tulee selvimmin esille kahden havainnon tapauksessa, jolloin ensimmäisen yksilön saama järjestysnumero määrää täysin toisen yksilön järjestysnumeron. Uskottavuusfunktio ei siten separoidu N:n lukuparin a_i, b_i tuloksi, vaan on esitettävä ehdollisten todennäköisyyksien ketjusäännön avulla:

$$\begin{aligned} L &= f(a_1, b_1) \times f(a_2, b_2 \mid a_1, b_1) \times f(a_3, b_3 \mid a_2, b_2, a_1, b_1) \times \ldots \\ &\times f(a_N, b_N \mid a_{N-1}, b_{N-1}, \ldots, a_1, b_1) \\ &= \prod_{j=1}^{N} f(a_j, b_j \mid a^{(j-1)}, b^{(j-1)}) \\ &= \prod_{j=1}^{N} f(b_j \mid b^{(j-1)}, a^{(j-1)}) \times \prod_{j=1}^{N} f(a_j \mid b^{(j)}, a^{(j-1)}), \end{aligned} \tag{6.10}$$

missä $a^{(j)} = (a_1, \ldots, a_j)$ ja $b^{(j)} = (b_1, \ldots, b_j)$. Termi

$$L_p = \prod_{j=1}^{N} f(a_j \mid b^{(j)}, a^{(j-1)}) \tag{6.11}$$

on työnhakijoiden havaittuihin järjestystunnuslukuihin $a_1, \ldots, a_N$ perustuva osittaisuskottavuusfunktio. Voidaan osoittaa, että osittaisuskottavuusfunktioon perustuvilla parametriestimaattoreilla on laadullisesti samat asymptoottiset ominaisuudet kuin tavallisilla suurimman uskottavuuden estimaattoreilla, toisin sanoen estimaattorit ovat tarkentuvia ja asymptoottisesti normaalisti jakautuneita (Lancaster 1990). Osittaisuskottavuusfunktioon perustuvat estimaattorit eivät kuitenkaan yleensä ole yhtä tehokkaita kuin tavalliset suurimman uskottavuuden estimaattorit. Osittaisuskottavuusfunktion j:s termi $f(a_j \mid b^{(j)}, a^{(j-1)})$ kuvaa todennäköisyyttä, että yksilö a_j

työllistyy ajankohtana b_j ehdolla, että joku kyseisen ajankohdan riskijoukkoon R_j kuuluvista työllistyy tällöin: [3]

$f(a_j \mid b^{(j)}, a^{(j-1)}) =$

$P(a_j$ työllistyy hetkellä $b_j \mid$ joku R_j:oon kuuluvista työllistyy tällöin$) =$

$$\frac{h_{a_j}(b_j; x_{a_j})}{\sum_{k \in R_j} h_k(b_j; x_k)} = \frac{h_0(b_j)\exp(x_{a_j}\prime\beta)}{\sum_{k \in R_j} h_0(b_j)\exp(x_k\beta)} = \frac{\exp(x_{a_j}\prime\beta)}{\sum_{k \in R_j}\exp(x_k\prime\beta)} \tag{6.12}$$

N:n riippumattoman havainnon osittaisuskottavuusfunktio on siten

$$\begin{aligned} L_p &= \prod_{j=1}^{N} f(a_j \mid b^j, a^{(j-1)}) \\ &= \prod_{j=1}^{N} \frac{\exp(z_{a_j}\prime\beta)}{\sum_{k \in R_j}\exp(z_k\beta)} \end{aligned} \tag{6.13}$$

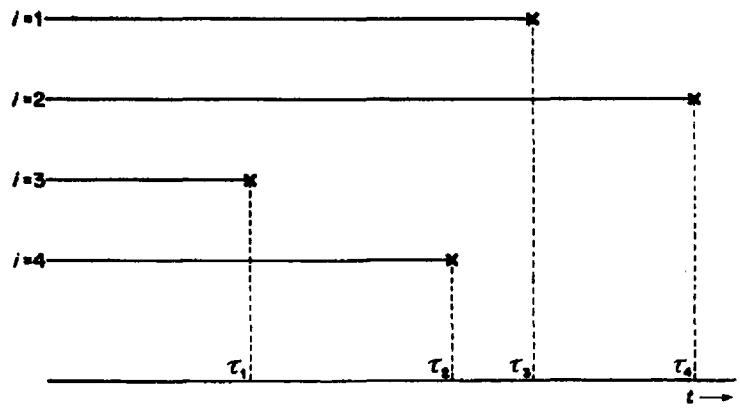
Nähdään, että kaikille yhteinen, hasardifunktion kestoriippuvuutta kuvaava perushasardifunktio supistuu osittaisuskottavuusfunktion lausekkeesta pois. Koska osittaisuskottavuusfunktio ei riipu $h_0(t)$:sta, ei parametrien β estimointia varten tarvitse lainkaan spesifioida perushasardifunktiota eikä siten työttömyyden keston jakaumaa. Osittaisuskottavuusfunktion osoittajassa ja nimittäjässä esiintyvät todennäköisyydet ovat ehdollisia todennäköisyyksiä, joissa ehdollistavana tekijänä on se, että työttömyys on kestänyt vähintään hetkeen b_j saakka. Tämä merkitsee sitä, että tarkastelu rajataan kunkin työllistymisajankohdan riskijoukkoon eli niihin, jotka eivät ole vielä työllistyneet ko. ajankohtaan mennessä, eikä siis tarkastella koko otosta.

Koska hasardifunktion työttömyyden kestosta riippuva osa supistuu osittaisuskottavuusfunktion lausekkeesta pois (ks. yhtälö 6.12), ei L_p riipu työttömyysjaksojen pituuksista:

$$L_p = \prod_{j=1}^{N} f(a_j \mid b_{(j)}, a_{(j-1)}) = \prod_{j=1}^{N} f(a_j \mid a_{(j-1)}) = f(a_1, \ldots, a_N), \tag{6.14}$$

mikä on järjestystunnuslukujen $A_1, \ldots, A_N$ yhteisjakauma. Osittaisuskottavuusfunktiolle voidaan antaa tulkinta järjestystunnuslukujen yhteisjakaumana vain, kun selittävät tekijät ovat vakioita työttömyyden keston suhteen. Esimerkkinä osittaisuskottavuusfunktion muodostamisesta tarkastellaan kuvan 6.2 tapausta, jossa on yksilöiden 1, 2, 3 ja 4 päättyneet työttömyyden kestot.

[3] Työllistymisajankohdan b_j riskijoukko $R_j = (a_j, \ldots, a_N)$. Riskijoukko koostuu siten yksilöistä, jotka eivät ole vielä työllistyneet ajankohtaan b_j mennessä.



Kuva 6.2: Yksilöiden 1, 2, 3 ja 4 päättyneet työttömyyden kestot. Lähde: Cox & Oakes s.92.

Kuvassa $A_1 = 3$, $A_2 = 4$, $A_3 = 1$ ja $A_4 = 2$. Riskijoukot työllistymisajankohtina ovat: $R_1 = 1,2,3,4$, $R_2 = 1,2,4$, $R_3 = 1,2$ ja $R_4 = 2$. Osittaisuskottavuusfunkio on tällöin

$$L_p = \frac{\exp(z_3'\beta)}{\exp(z_1'\beta) + \exp(z_2'\beta) + \exp(z_3'\beta) + \exp(z_4'\beta)}$$
$$\times \frac{\exp(z_4'\beta)}{\exp(z_1'\beta) + \exp(z_2'\beta) + \exp(z_4'\beta)} \times \frac{\exp(z_1'\beta)}{\exp(z_1'\beta) + \exp(z_2'\beta)} \times \frac{\exp(z_2'\beta)}{\exp(z_2'\beta)} \quad (6.15)$$

6.2.2 Osittaisuskottavuusfunktio sensuroitujen havaintojen tapauksessa

Yleensä oletetaan, että sensurointiajat ja tapahtuma-ajat ovat toisistaan riippumattomia (ks. kpl 2.3). Tällöin sensuroituminen ei ole informatiivista parametrien β estimoinnin kannalta: parametrit β kuvaavat, kuinka selittävät tekijät vaikuttavat työllistymiseen; sensuroitumisen oletetaan riippuvan muista, työllistymiseen vaikuttavista tekijöistä riippumattomista tekijöistä. Jos N:n yksilön aineistossa on d kpl sensuroituja havaintoja, koostuu osittaisuskottavuusfunktio (N-d):stä komponentista. Sensuroidut havainnot vaikuttavat ainoastaan riskijoukon kokoon. Jos jokin sensurointiaika on yhtä pitkä kuin jokin työttömyysjaksoista b_j, $j = 1, \ldots, N$, oletetaan sensuroinnin tapahtuneen välittömästi työllistymisen jälkeen, jolloin sensuroitu havainto vaikuttaa samana ajankohtana työllistyneen riskijoukkoon. Osittaisuskottavuusfunktio sensuroitujen havaintojen tapauksessa on siten

$$L_p = \prod_{j=1}^{N-d} \left[\frac{\exp(z_{a_j}'\beta)}{\sum_{l \in R_j} \exp(z_l'\beta)} \right] \tag{6.16}$$

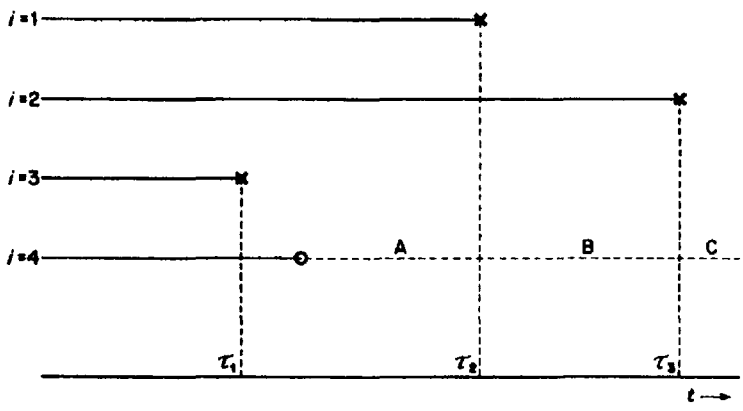
Yhtälöä 6.16 voidaan perustella myös ajattelemalla, että sensuroitujen havaintojen tapauksessa ei työnhakijoiden keskinäinen työllistymisjärjestys ole enää yksikäsitteinen: sensuroitujen havaintojen osalta tiedetään ainoastaan, että työttömyysjakso on vähintään sensuroitumisajan pituinen. Kuva 6.3 havainnollistaa asiaa: yksilön 4 sensuroidusta työttömyysjaksosta tiedetään ainoastaan, että jakso on kestänyt kauemmin kuin yksilön 3 työttömyysjakso. A, B, ja C ovat aikavälejä, joilla yksilön 4 työttömyys olisi voinut päättyä. Havaittua tilannetta vastaavia järjestystunnuslukuvektoreita on kolme kappaletta:

$$[3, 4, 1, 2], [3, 1, 4, 2], [3, 1, 2, 4].$$

Neljä havaintoa voidaan järjestää jonoon 4! = 24 eri tavalla; havaitun perusteella tiedetään, että todellinen järjestys on jokin kolmesta ylläolevasta. Osittaisuskottavuusfunktio muodostetaan käyttämällä hyväksi tätä tietoa: osittaisuskottavuusfunktio on mahdollisia järjestystunnuslukuvektoreita vastaavien osittaisuskottavuusfunktioiden summa: $L_p = L_p^A + L_p^B + L_p^C$. Jos yksilön 4 työttömyys päättyisi välillä A (mikä vastaa järjestystunnuslukuvektoria [3, 4, 1, 2]), olisi osittaisuskottavuusfunktio seuraavanlainen:

$$L_p^A = \frac{\exp(z_3'\beta)}{\exp(z_1'\beta) + \exp(z_2'\beta) + \exp(z_3'\beta) + \exp(z_4'\beta)}$$
$$\times \frac{\exp(z_4'\beta)}{\exp(z_1'\beta) + \exp(z_2'\beta) + \exp(z_4'\beta)} \times \frac{\exp(z_1'\beta)}{\exp(z_1'\beta) + \exp(z_2'\beta)} \times \frac{\exp(z_2'\beta)}{\exp(z_2'\beta)}.$$

Välillä B päättynyt työttömyys (järjestystunnuslukuvektori [3, 1, 4, 2]) tuottaisi



Kuva 6.3: Neljän työnhakijan työttömyysjaksot. x=päättynyt jakso, o=sensuroitu jakso. Lähde: Cox & Oakes, s. 94

osittaisuskottavuusfunktion:

$$L_p^B = \frac{\exp(z_3'\beta)}{\exp(z_1'\beta)+\exp(z_2'\beta)+\exp(z_3'\beta)+\exp(z_4'\beta)} \times \frac{\exp(z_1'\beta)}{\exp(z_1'\beta)+\exp(z_2'\beta)+\exp(z_4'\beta)} \times \frac{\exp(z_4'\beta)}{\exp(z_4'\beta)+\exp(z_2'\beta)} \times \frac{\exp(z_2'\beta)}{\exp(z_2'\beta)}.$$

Välillä C päättynyt työttömyys (järjestystunnuslukuvektori [3, 1, 2, 4]) tuottaisi osittaisuskottavuusfunktion:

$$L_p^C = \frac{\exp(z_3'\beta)}{\exp(z_1'\beta)+\exp(z_2'\beta)+\exp(z_3'\beta)+\exp(z_4'\beta)} \times \frac{\exp(z_1'\beta)}{\exp(z_1')+\exp(z_2'\beta)+\exp(z_4'\beta)} \times \frac{\exp(z_2'\beta)}{\exp(z_2'\beta)+\exp(z_4'\beta)} \times \frac{\exp(z_4'\beta)}{\exp(z_4'\beta)}.$$

Osittaisuskottavuusfunktio kuvan 6.3 neljälle työnhakijalle on siten

$$\begin{aligned} L_p &= L_p^A + L_p^B + L_p^C \\ &= \frac{\exp(z_3'\beta)}{\exp(z_1'\beta)+\exp(z_2'\beta)+\exp(z_3'\beta)+\exp(z_4'\beta)} \times \frac{\exp(z_1'\beta)}{\exp(z_1'\beta)+\exp(z_2'\beta)} \\ &\times \frac{\exp(z_2'\beta)}{\exp(z_2'\beta)}. \end{aligned}$$

6.2.3 Sidokset

Vaikka työttömyyden kestoa voidaan pitää jatkuvana satunnaismuuttujana, saattaa aineistossa esiintyä useita samanpituisia työttömyysjaksoja, ns. **sidoksia**. Työministeriön työnhakijarekisterin tietojen perusteella työttömyysjaksojen pituudet saadaan laskettua päivän tarkkuudella. On mahdollista, että otokseen valikoituu yksilöitä, joiden työttömyys kestää päivälleen yhtä kauan. Jos mittaustarkkuus on alhainen ja sidoksia paljon, on suositeltavaa käyttää esimerkiksi luokiteltujen tapahtuma-aikojen mallia (ks. kpl 6.3). Samoin kuin sensuroitujen havaintojen tapauksessa voidaan ajatella, että ongelmana on järjestystunnuslukuvektorin ei-yksikäsitteisyys: samanpituisen työttömyysjakson omaavien työnhakijoiden keskinäinen työllistymisjärjestys on tuntematon. Jos d_j:n työnhakijan työttömyysjakso on b_j:n pituinen, on kyseisten havaintojen kontribuutio osittaisuskottavuusfunktioon $d_j!$ (d_j:n kertoma) termin summa: d_j työnhakijaa voidaan järjestää $d_j!$:llä eri tavalla ja kutakin vastaa erilainen osittaisuskottavuusfunktio. Jos samanpituisia työttömyysjaksoja on paljon, on osittaisuskottavuusfunktio laskennallisesti raskas, koska mahdollisia työllistymisjärjestyksiä kuvaava summalauseke ei sievene kuten sensuroitujen havaintojen tapauksessa (yhtälö 6.17). Paljon käytetty ositaisuskottavuusfunktion approksimaatio, kun aineistossa on runsaasti sidoksia, on Peton (1972) approksimaatio:

$$L_p = \prod_{j=1}^{k} \frac{\exp(s_j'\beta)}{\left(\sum_{l \in R_j} \exp(z_l'\beta)\right)^{d_j}}, \tag{6.17}$$

missä k on eri pituisten työttömyysjaksojen lukumäärä ja s_j on b_j:n pituisen työttömyysjakson omaavien työnhakijoiden selittäjävektorien summa. Jos sidoksia ei ole, on $d_j = 1$, $j = 1, \ldots, k$ ja yhtälö 6.17 on sama kuin yhtälö 6.16. Peton approksimaatiossa ei samana ajankohtana työllistyneiden mahdollisia työllistymisjärjestyksiä oteta huomioon: riskijoukko on jokaiselle tiettynä ajankohtana työllistyneelle sama.

6.2.4 Perushasardifunktion estimointi

Paitsi työnhakijoiden ominaisuuksien ja taustatekijöiden vaikutuksesta työllistymistodennäköisyyteen, ollaan usein myös kiinnostuneita hasardifunktion duraatioriippuvuudesta l. siitä, kuinka työllistymistodennäköisyys vaihtelee työttömyyden keston suhteen. Täysin parametroiduissa malleissa regressiokertoimet ja perushasardifunktion muotoa kuvaavat parametrit estimoidaan samanaikaisesti suurimman uskottavuuden menetelmällä. Coxin mallin estimointimenetelmällä jää perushasardifunktio tuntemattomaksi. Perushasardifunktio voidaan kuitenkin estimoida jälkeenpäin osittaisuskottavuusfunktion avulla estimoituja regressiokertoimia hyväksi käyttäen. Seuraavassa esitetään Breslow:n (1974) estimaattori perushasardifunktiolle. Breslow:n estimaattori perustuu paloittaiseen eksponenttimalliin, jossa aika-akselin

osituskohdiksi on valittu työllistymisajankohdat b_j. Mallin diskreetin perushasardifunktionn suurimman uskottavuuden estimaattori on

$$\hat{h}_{0j} = \frac{d_j}{(b_j - b_{j-1}) \sum_{l \in R_j} \exp(z_l'\hat{\beta})} \quad j = 1, \ldots, k. \tag{6.18}$$

$\hat{h}_{0j}$ on välin $[b_{j-1}, b_j)$ vakio perushasardifunktion estimaattori. $\hat{h}_{0j}$ on porrasfunktio, jonka arvo muuttuu kunakin työllistymisajankohtana b_j. d_j on työllistyneiden lukumäärä j:nnellä aikavälillä ja k aikavälien lukumäärä. Estimaattoria voidaan käyttää perushasardifunktion estimoimiseen Coxin mallissa korvaamalla β:n suurimman uskottavuuden estimaattori $\hat{\beta}$ osittaisuskottavuusfunktioon perustuvalla estimaatilla $\tilde{\beta}$. Coxin mallin diskreetin kumulatiivisen perushasardifunktion estimaattori on

$$\tilde{H}_0(t) = \sum_{b_j \leq t} (b_j - b_{j-1}) \tilde{h}_{0j} = \sum_{b_j \leq t} \frac{d_j}{\sum_{l \in R_j} \exp(z_l'\tilde{\beta})}. \tag{6.19}$$

Kumulatiivisen perushasardifunktion estimaattoria voidaan käyttää proportionaalisuusoletuksen paikkansapitävyyden tutkimiseen Coxin mallissa. Coxin mallin spesifikaatiotestausta käsitellään luvussa 8.

6.3 Luokiteltujen tapahtuma-aikojen malli

Luokiteltujen tapahtuma-aikojen mallin esittivät ensimmäisinä Kalbfleisch ja Prentice (1973). Mallin estimointia ovat käsitelleet mm. Prentice ja Gloeckler (1978). Malli soveltuu tapauksiin, joissa työttömyysjaksojen pituuksia ei tunneta tarkasti. Usein tiedetään ainoastaan, millä aikavälillä (millä viikolla tai minä kuukautena) työttömyysjakso on päättynyt. Kalbfleisch ja Prentice suosittelevat mallia käytettäväksi myös tapauksissa, joissa työttömyysjaksojen pituudet tunnetaan varsin tarkasti, mutta aineistossa on havaintojen suuren määrän vuoksi paljon sidoksia. Jos sidoksia on paljon, tulee osittaisuskottavuusfunktiosta laskennallisesti raskas. Toisaalta esimerkiksi edellä esitetty Peton approksimaatio osittaisuskottavuusfunktiolle voi tuottaa pahasti harhaisia parametriestimaatteja, mikäli jollain ajankohdalla työllistyneiden osuus ko. ajankohdan riskijoukosta on suuri (Kalbfleisch & Prentice 1980). Kuten Coxin malli, on luokiteltujen tapahtuma-aikojen malli semiparametrinen; perushasardifunktiota ei tarvitse spesifioida. Mallissa oletetaan, että diskreettien havaintojen taustalla on jatkuvia satunnaismuuttujia T_i, $i = 1, \ldots, N$, joiden hasardifunktiot ovat suhteellisen hasardin muotoa

$$h_i(t; z) = h_0(t) \exp(z_i'\beta). \tag{6.20}$$

Työnhakijoiden toteutuneista työttömyysjaksojen pituuksista tiedetään siis ainoastaan, mille aikavälille $[c_{m-1}, c_m)$, $m = 1, \ldots, M$ ($c_0 = 0, c_M = \infty$) jaksot kuuluvat.

Todennäköisyys, että työnhakija ei työllisty aikavälillä m ehdolla, että hän on edelleen työttömänä ko. aikavälin alussa voidaan esittää jatkuvan hasardifunktion avulla seuraavalla tavalla:

$$P(\ T_i \geq c_m \mid T_i \geq c_{m-1}) = \frac{P(T_i \geq c_m)}{P(T_i \geq c_{m-1})} = \frac{S_i(c_m)}{S_i(c_{m-1})}$$
$$= \frac{\exp\left[-\int_0^{c_m} h_i(u;z)du\right]}{\exp\left[-\int_0^{c_{m-1}} h_i(u;z)du\right]} = \exp\left[-\int_{c_{m-1}}^{c_m} h_i(u;z)du\right]. \qquad (6.21)$$

Yhtälö 6.21 voidaan kirjoittaa myös seuraavalla tavalla:

$$P(T_i \geq c_m \mid T_i \geq c_{m-1}) = \exp[-\exp(z_i\prime\beta + \gamma_m)], \qquad (6.22)$$

missä $\gamma_m = \ln \int_{c_{m-1}}^{c_m} h_0(u)du$. Todennäköisyys, että työnhakija on yhä työttömänä aikavälin m alussa saadaan ehdollisten todennäköisyyksien ketjusäännön (ks. alaviite 3 s. 9) avulla:

$$P(T_i \geq c_{m-1}) = \prod_{l=1}^{m-1} \exp[-\exp(z_i\prime\beta + \gamma_l)]. \qquad (6.23)$$

Ehdollinen työllistymistodennäköisyys aikavälillä m on

$$P(\ c_{m-1} \leq T_i < c_m \mid T_i \geq c_{m-1}) = P(T_i < c_m \mid T_i \geq c_{m-1})$$
$$= 1 - P(T_i \geq c_m \mid T_i \geq c_{m-1}) = 1 - \exp[-\exp(z_i\prime\beta + \gamma_m)]. \qquad (6.24)$$

N:n havainnon uskottavuusfunktio on siten

$$L = \prod_{i=1}^{N} \left[[1 - \exp[-\exp(z_i\prime\beta + \gamma_{m_i})]]^{v_i} \times \prod_{l=1}^{m_i-1} \exp[-\exp(z_i\prime\beta + \gamma_l)] \right], \qquad (6.25)$$

missä m_i indikoi yksilön i työllistymis- tai sensuroitumisaikavälin. m_i:nnellä aikavälillä sensuroituneen yksilön i kontribuutio uskottavuusfunktioon on

$$P(T_i \geq c_{m_i-1}) = \prod_{l=1}^{m_i-1} \exp[-\exp(z_i\prime\beta + \gamma_l)] \qquad (6.26)$$

eli tietyllä aikavälillä sensuroitujen oletetaan sensuroituneen kyseisen aikavälin alussa. Sensuroidut havainnot eivät siten vaikuta aikavälin riskijoukkoon. Estimoitavia parametreja ovat regressiokertoimet β sekä kunkin aikavälin perusriskitasoa kuvaavat parametrit γ_m, $m = 1, \ldots, M$. Parametrit estimoidaan samanaikaisesti suurimman uskottavuuden menetelmällä.

Luokiteltujen tapahtuma-aikojen mallia ovat soveltaneet työllistymiseen vaikuttavien tekijöiden tutkimiseen mm. Meyer (1990), Narendranathan ja Stewart (1991) sekä Suomessa Lilja (1992). Liljan aineisto koostuu vuosien 1984-1987 Tilastokeskuksen työvoimatutkimuksen vuosihaastatteluista. Vuosihaastattelut koostuvat

seurattavien henkilöiden yleensä 3 kuukauden väliajoin toistuvista haastatteluista. Haastatteluja on kussakin tutkimuksessa 15 kuukauden aikana yhteensä 5. Lilja valitsi aineistoonsa kunkin tutkimuksen ensimmäisessä haastattelussa mukana olleet 15-54–vuotiaat työttömät. Työttömiä seurattiin enintään 15 kuukautta. Lilja valitsi mallinsa aikavälien pituudeksi haastatteluajankohtien etäisyyden mukaan 3 kuukautta.

Luku 7

Työttömyyden erilaisten päättymissyiden huomioiminen

Tähän asti on oletettu, että työttömyys voi päättyä ainoastaan työllistymiseen. Todellisuudessa työttömyys voi päättyä myös työmarkkinoilta poistumiseen esim. opiskelun, asepalvelun tai sairastumisen vuoksi. Jos halutaan tutkia nimenomaan työllistymiseen (ei vain työttömyyden päättymiseen) vaikuttavia tekijöitä, on tärkeää ottaa huomioon työttömyyden erilaiset päättymissyyt. Työllistyminen voidaan edelleen jakaa avoimille työmarkkinoille työllistymiseen ja työllisyystoimenpitein työllistymiseen. Tutkielman empiirisessä osassa tarkastellaan nuorten avoimille työmarkkinoille työllistymistä. Malleja, joissa otetaan huomioon se, että toteutuvia tapahtumia voi olla useanlaisia, kutsutaan **kilpailevien riskien malleiksi** (competing risks models).

Lilja (1992) jakoi työttömyyden päättymissyyt kolmeen luokkaan: työllistymiseen, vapaaehtoiseen työmarkkinoilta poistumiseen sekä pakottavista syistä johtuvaan työmarkkinoilta poistumiseen. Opintojen aloittaminen ja päätös jäädä tekemään kotityötä ovat esimerkkejä vapaaehtoisesta työmarkkinoilta poistumisesta. Pakottavia syitä työmarkkinoilta poistumiseen ovat mm. asepalvelu ja sairastuminen. Työllistyminen, vapaaehtoinen ja pakottavista syistä johtuva työmarkkinoilta poistuminen ovat tapahtumia, joiden taustalla on todennäköisesti varsin erilaisia tekijöitä. Tällöin ei johtopäätöksiä työllistymiseen vaikuttavista tekijöistä voida perustaa malliin, jossa erilaisia työttömyyden päättymissyitä ei ole otettu huomioon.

Liljan (1992) tulosten mukaan esim. sukupuoli vaikuttaa vapaaehtoiseen ja pakottavista syistä johtuvaan työmarkkinoilta poistumiseen päinvastaisella tavalla: naiset poistuvat vapaaehtoisesti työvoimasta todennäköisemmin kuin miehet, kun taas miesten todennäköisyys poistua työvoimasta pakottavista syistä johtuen on suurem-

pi kuin naisten. Kotitöiden tekeminen ja lasten hoitaminen näyttää yhä olevan pääosin naisten vastuulla. Naiset lienevät myös miehiä innokkaampia kouluttautumaan. Toisaalta miesten asevelvollisuus ja suurempi todennäköisyys joutua työkyvyttömäksi saavat aikaan sen, että miehet vetäytyvät työmarkkinoilta pakottavista syistä useammin kuin naiset. Myös Liljan estimoimat perushasardifunktiot eroavat toisistaan työttömyyden päättymissyyn mukaan. Aineiston keskimääräiset ominaisuudet omaavan henkilön (ns. referenssihenkilön) ehdollinen työllistymistodennäköisyys on koko seuranta-ajan suurempi kuin todennäköisyys vetäytyä työvoiman ulkopuolelle.

Kun työttömyden erilaiset päättymissyyt otetaan huomioon, liittyy kuhunkin työnhakijaan tiedot (t_i, v_i, c_i, z_i), kun $v_i = 1$ ja (t_i, v_i, z_i), kun $v_i = 0$. v_i on sensurointi-indikaattori ja c_i sisältää tiedon yksilön i työttömyyden päättymissyystä. Sensuroitujen havaintojen työttömyyden päättymissyy on tuntematon. t_i on yksilön i työttömyyden kesto tai sensurointiaika sensurointi-indikaattorin arvosta riippuen. Vektori z_i sisältää yksilön i työllistymistä selittävät tekijät.

Oletetaan, että työttömyyden kesto T on jatkuva satunnaismuuttuja. Määritellään ns. **tapahtumaspesifinen hasardifunktio**:

$$h_c(t;z) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt, C = c \mid z, T \geq t)}{dt}, \quad c = 1, \ldots, m. \tag{7.1}$$

$h_c(t;z)dt$ kuvaa (likimääräisesti) ehdollista todennäköisyyttä, että työttömyys päättyy syyhyn c pienellä aikavälillä $(t, t + dt)$. Ehdollistavana tekijänä on se, että työttömyys ei ole päättynyt mihinkään syyhyn ajankohtaan t mennessä. m on erilaisten työttömyyden päättymissyiden lukumäärä. $h_c(t;z)$ on satunnaismuuttujien T ja C yhteisjakauman hasardifunktio; todennäköisyys, että työttömyys päättyy hetkellä t syyhyn c, ehdolla, että työttömyys on jatkunut hetkeen t saakka (myös selittävien tekijöiden vektori z on ehdollistava tekijä). Koska kunkin työnhakijan työttömyys voi päättyä vain yhdellä tavalla, on työttömyyden (mihin syyhyn tahansa) päättymisen ehdollinen todennäköisyys tapahtumaspesifisten komponenttien summa:

$$h(t;z) = \sum_{c=1}^{m} h_c(t;z). \tag{7.2}$$

T:n eloonjäämisfunktio on siten

$$S(t;z) = \exp(-\int_0^t \sum_{c=1}^{m} h_c(u;z)du) = \prod_{c=1}^{m} \exp(-\int_0^t h_c(u;z)du). \tag{7.3}$$

T:n ja C:n yhteisjakauman tiheysfunktio on (vrt. yhtälö 2.6)

$$f_c(t;z) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt, C = c \mid z)}{dt} = h_c(t;z)S(t;z). \tag{7.4}$$

Yhtälöistä 7.2-7.4 nähdään, että havaintojen uskottavuusfunktio voidaan kirjoittaa tapahtumaspesifisten hasardifunktioiden avulla:

$$L = \prod_{i=1}^{N} \left((h_{c_i}(t_i; z_i))^{v_i} \prod_{c=1}^{m} \exp(-\int_0^{t_i} h_c(u; z)du) \right) \tag{7.5}$$

Kirjoittamalla uskottavuusfunktio seuraavalla tavalla:

$$L = \prod_{c=1}^{m} \prod_{i=1}^{N} \left((h_{c_i}(t_i; z_i))^{v_i^c} \exp(-\int_0^{t_i} h_{c_i}(u; z)du) \right) \tag{7.6}$$

missä $v_i^c = 1$, jos yksilön i työttömyyden päättymissyy c ja $v_i^c = 0$, jos työttömyys on päättynyt muulla tavalla tai jos havainto on sensuroitu, huomataan, että uskottavuusfunktio koostuu tapahtumaspesifisistä komponenteista L_c:

$$L = \prod_{c=1}^{m} L_c, \tag{7.7}$$

missä

$$L_c = \prod_{i=1}^{N} \left((h_{c_i}(t_i; z))^{v_i^c} \exp(-\int_0^{t_i} h_{c_i}(u; z)du) \right). \tag{7.8}$$

Niiden yksilöiden, joiden työttömyyden päättymissyy $\neq c$ tai joiden työttömyysjaksot ovat sensuroituja, kontribuutio uskottavuusfunktion komponenttiin L_c on $\exp(-\int_0^{t_i} h_{c_i}(u; z)du)$. Kussakin tapahtumaspesifisessä komponentissa L_c pidetään siis muutoin kuin syyhyn c päättyneitä työttömyysjaksoja sensuroituina. Uskottavuusfunktion separoituminen tapahtumaspesifisiin komponentteihin merkitsee sitä, että mikäli tyyppikohtaiset hasardifunktiot eivät sisällä yhteisiä parametreja (selittävien tekijöiden regressiokertoimien sekä hasardifunktion muotoa kuvaavien parametrien sallitaan vaihdella päättymissyittäin), voidaan syyhyn c päättyvien työttömyysjaksojen päättymistä selittävien tekijöiden regressiokertoimet β_c estimoida komponentin L_c perusteella. Estimoitaessa parametreja β_c ovat komponentit $L_k, k \neq c$ vakioita β_c:n suhteen ja voidaan siten jättää pois uskottavuusfunktion lausekkeesta.

Erilaiset työttömyyden päättymissyyt voidaan ottaa helposti huomioon myös Coxin mallissa. Kullekin työttömyyden päättymissyylle c, $c = 1, \ldots, m$ muodostetaan oma osittaisuskottavuusfunktio L_p^c, jossa muutoin kuin syihin c päättyneitä työttömyysjaksoja pidetään sensuroituina. Syihin $k \neq c$ päättyneet jaksot vaikuttavat siis vain riskijoukon kokoon komponentissa L_p^c. Osittaisuskottavuusfunktio koostuu siten komponenteista

$$L_p^c = \prod_{j=1}^{N_c} \left[\frac{\exp(z_{a_j}\prime\beta_c)}{\sum_{k \in R_j} \exp(z_k \beta_c)} \right], \; c = 1, \ldots, m, \tag{7.9}$$

missä N_c on syyhyn c päättyneen työttömyysjakson omaavien yksilöiden lukumäärä ja a_j on j:s yksilö, jonka työttömyys päättyi syyhyn c. R_j on riskijoukko yksilön a_j

työllistymisajankohtana. R_j:oon kuuluvat kaikki yksilöt, joiden työttömyys jatkuu yhä a_j:n työllistymisajankohtana; siis myös ne, joiden työttömyys päättyy muutoin kuin syyhyn c. Jos tapahtumaspesifiset hasardifunktiot eivät sisällä yhteisiä parametreja, voidaan päättymissyyn c parametrit β_c estimoida pelkästään komponentin L_p^c perusteella; osittaisuskottavuusfunktion muut komponentit ovat vakioita β_c:n suhteen.

Luku 8

Coxin mallin spesifikaatiotestauksesta

Luvussa käsitellään Coxin mallin proportionaalisuusoletuksen testaamista sekä mallin ennustekyvyn tutkimista. Seuraavassa kappaleessa esitettyä tapaa tutkia proportionaalisuusoletusta Kaplan–Meier-estimaattien avulla voidaan luonnollisesti soveltaa mallispesifikaatiosta riippumatta. Samoin kappaleessa 8.2. esitettyä menetelmää mallin ennustekyvyn tutkimiseksi voidaan käyttää myös muiden kuin Coxin mallin spesifikaatiotestaukseen.

8.1 Proportionaalisuusoletuksen testaaminen

Suhteellisten hasardien mallissa

$$h(t;z) = h_0(t)\exp(z\prime\beta) \tag{8.1}$$

selittävät tekijät vaikuttavat siirtämällä hasardifunktiota perushasardiin nähden vakiokertoimella $\exp(z\prime\beta)$. Tällöin yksilöiden hasardifunktioiden suhteet $h(t;z_i)/h(t;z_j)$ $\forall i,j = 1,\ldots,n$ ovat työttömyyden kestosta riippumattomia. Tämän ns. proportionaalisuusoletuksen paikkansapitävyyttä voidaan tutkia graafisesti eloonjäämisfunktioiden avulla tai formaalimmin sisällyttämällä malliin työttömyyden keston ja selittävien tekijöiden yhdysvaikutus- eli interaktiotermejä ja testaamalla termien regressiokertoimien merkitsevyyttä.

Graafisesti proportionaalisuusoletuksen realistisuutta voidaan tutkia seuraavalla tavalla. Työttömyyden keston kumulatiivinen hasardifunktio on suhteellisten hasar-

dien mallissa

$$H(t;z) = \int_0^t h(u;z)du = \exp(z\prime\beta)\int_0^t h_0(u)du = \exp(z\prime\beta)H_0(t). \qquad (8.2)$$

$H(t;z)$:n ja $H_0(t)$:n yhteys voidaan esittää seuraavalla tavalla:

$$\ln H(t;z)) = z\prime\beta + \ln H_0(t)). \qquad (8.3)$$

Koska $H_0(t)$ on kaikille sama, tulisi erilaiset selittävät tekijät omaavien yksilöiden kumulatiivisen hasardin luonnollisen logaritmin poiketa toisistaan vakiotermin verran, mikäli selittävät tekijät vaikuttavat hasardiin oletetulla tavalla. $\ln H_0(t)$:ien kuvaajien tulisi tällöin olla likimain samansuuntaisia ja kuvaajien vertikaalisen erotuksen vakio. Yhden tai muutaman muuttujan vaikutustapaa voidaan tutkia ennen varsinaista mallintamista jakamalla aineisto osiin muuttujien arvojen mukaan ja estimoimalla eloonjäämisfunktiot (Kaplan–Meier- tai eloonjäämistaulu-menetelmällä) ositteille erikseen. Ositteiden log-kumulatiivisten hasardien estimaattien vertikaalisen erotuksen tulisi olla likimain vakio, mikäli proportionaalisuusoletus on voimassa.

Mallintamisvaiheessa voidaan Coxin mallin proportionaalisuusoletusta testata samaan tapaan log-kumulatiivisia perushasardifunktioita vertaamalla. Ositetaan aineisto sen muuttujan (muuttujien) arvojen mukaan, jonka vaikutustapaa halutaan tutkia ja sovitetaan aineistoon malli, jossa sallitaan perushasardifunktion poiketa ositteiden välillä. Jos esimerkiksi halutaan tutkia, kuinka työttömyyskassan jäsenyys vaikuttaa hasardifunktioon, on estimoidaan seuraavanlainen malli:

$$h_i(t;z^-) = h_{0i}(t)\exp(z^-\beta^-),\ i = 1,2, \qquad (8.4)$$

missä h_1 on ryhmän 1 (ei-jäsenet) ja h_2 ryhmän 2 (jäsenet) hasardifunktio. z^- on selittäjävektori, josta puuttuu kassan jäsenyyttä kuvaava muuttuja ja β^- on vastaava regressiokertoimien vektori. Kassan jäsenyyden vaikutus näkyy nyt perushasardifunktiossa $h_{02}(t)$. Muuttujan vaikutustapaa ei ole rajoitettu millään tavalla: ryhmien 1 ja 2 perushasardifunktiot ovat toisistaan riippumattomia ja saavat määräytyä täysin vapaasti. Sen sijaan muiden kuin kassan jäsenyyttä kuvaavien muuttujien oletetaan vaikuttavan eri ryhmissä samalla tavalla (β^- on sama molemmissa ryhmissä). Regressiokertoimien β estimointi perustuu osittaisuskottavuusfunktioon, joka on nyt ositteiden osittaisuskottavuusfunktioiden tulo. Esimerkin tapauksessa muodostetaan siis kaksi osittaisuskottavuusfunktiota, toinen kassan jäsenille ja toinen kassaan kuulumattomille:

$$L_p = L_{p1} \times L_{p2}. \qquad (8.5)$$

Coxin mallin estimointimenetelmällä jää perushasardifunktio tuntemattomaksi. Kumulatiivinen perushasardifunktio onkin estimoitava regressiokertoimien estimoimisen jälkeen esim. Breslow:n menetelmällä (ks. kpl 6.2.4). Breslow:n menetelmällä

saadaan estimaatit kumulatiivisille perushasardifunktioille ($H_{0i}(t)$). Tällöin voidaan tarkastella aineistoa osittavan muuttujan vaikutusta log-kumulatiiviseen hasardifunktioon, kun muiden muuttujien vaikutus on eliminoitu (laskettaessa eloonjäämisfunktion estimaatteja Kaplan–Meier- tai eloonjäämistaulu-menetelmällä on oletettava, että yksilöt ositteiden sisällä ovat homogeenisia; menetelmät eivät kykene ottamaan huomioon havaintojen heterogeenisuutta). Jos log-kumulatiivisten perushasardifunktioiden vertikaalinen erotus on likimain vakio, voidaan aineistoa osittava muuttuja sisällyttää regressiokertoimien vektoriin. Jos proportionaalisuusoletus ei näytä pitävän paikkaansa, on ongelma jo ratkaistu sallimalla muuttujan vaikuttaa ei-proportionaalisesti perushasardifunktiossa.

Formaalimmin Coxin mallin proportionaalisuusoletusta voidaan testata sisällyttämällä malliin selittävien tekijöiden ja työttömyyden keston interaktiotermejä. Työttömyyskassan jäsenyyden vaikutusta hasardifunktioon voidaan testata seuraavalla mallilla:

$$h_i(t; z^-) = h_0(t) \exp(\beta_1 z_{1i} + \beta_2 z_{1i} \ln t + \beta^- z^-) \; i = 1, 2, \tag{8.6}$$

missä z_{1i} on työttömyyskassan jäsenyyttä kuvaava indikaattorimuuttuja: $z_{11} = 0$ (ei-jäsenet) ja $z_{12} = 1$ (jäsenet). Ryhmien 1 ja 2 hasardifunktiot ovat tällöin:

$$h_1(t; z^-) = h_0(t) \exp(\beta^- z^-) \text{ ja} \tag{8.7}$$

$$h_2(t; z^-) = h_0(t) \exp(\beta_1 + \beta_2 \ln t + \beta^- z^-) \tag{8.8}$$

ja hasardifunktioiden suhde

$$\frac{h_2(t; z^-)}{h_1(t; z^-)} = \exp(\beta_1 + \beta_2 \ln t). \tag{8.9}$$

Jos kerroin β_2 poikkeaa merkitsevästi nollasta, riippuu hasardifunktioiden suhde työttömyyden kestosta eikä proportionaalisuusoletus ole voimassa. Jos $\beta_2 > 0$, kasvaa jäsenyyden vaikutus hasardifunktioon työttömyyden keston myötä ja vastaavasti vähenee, jos $\beta_2 < 0$. $\ln t$:n sijaan voidaan tietenkin valita jokin muu työttömyyden keston funktio.

8.2 Mallin ennustekyvyn tutkiminen

Mallin ennustekykyä, ts. sitä, kuinka hyvin malli ennustaa tapahtuma-aikoja, voidaan tutkia residuaalien avulla. Seuraava suppea esitys perustuu pääosin Mervi Eerolan syksyllä 1993 pitämän tapahtumahistoria-analyysin kurssin luentomonisteisiin. Asiaa käsitellään perusteellisemmin tapahtumahistoria-analyysin moderneissa, laskuriprosessien teoriaan perustuvissa oppikirjoissa (esim. Fleming & Harrington 1991, Andersen, Borgan, Gill & Keiding 1993).

Kuten tavallisessa lineaarisessa regressioanalyysissä, on

residuaali = **havaittu** − **ennustettu**. Tässä yhteydessä

havaittu = yksilölle i seurannassaoloaikana X_i sattuneiden tapahtumien lkm

ennustettu = mallin antama ennuste yksilölle i ajankohtaan X_i mennessä sattuneiden tapahtumien lkm:stä

Seurannassaoloaika $X_i = min(T_i, C_i)$. Koska tutkimuksessa tarkastellaan vain yhtä työttömyysjaksoa kutakin yksilöä kohti, on havaittu tapahtumien lukumäärä korkeintaan yksi: työllistyneille **havaittu** = 1 ja sensuroiduille **havaittu** = 0.

Kumulatiivinen hasardifunktio $H_i(t)$ kuvaa yksilölle i ajankohtaan t mennessä sattuneiden tapahtumien odotettua lukumäärää. Elinaika-analyysin perusoppikirjoissa (esim. Cox & Oakes 1984, Kalbfleisch & Prentice 1980, Lawless 1982) ei tästä tulkinnasta juurikaan puhuta. Modernimmat tapahtumahistoria-analyysin oppikirjat esittävät tulkinnan nojautuen laskuriprosessien teoriaan, johon en ole vielä perehtynyt. Seuraavan perustelun kumulatiivisen hasardifunktion tulkinnalle tapahtumien odotettuna lukumääränä on esittänyt Anders Ekholm. Merkitään yksilölle i ajankohtaan t mennessä sattuneiden tapahtumien lukumäärää $N_i(t)$:llä. Määritellään

$$U(t) = E(N(t)) \text{ ja} \tag{8.10}$$

$$u(t) = \lim_{dt \to 0} \frac{U(t+dt) - U(t)}{dt}, \tag{8.11}$$

Tällöin

$$u(t)dt = E(N(t+dt) - N(t)) = P(N(t+dt) - N(t) = 1) \tag{8.12}$$

$$= P(t < T < t+dt \mid T > t) = \frac{f(t)dt}{S(t)} = h(t)dt. \tag{8.13}$$

Tästä seuraa, että

$$u(t) = h(t) \text{ ja} \tag{8.14}$$

$$U(t) = \int_0^t u(z)dz = \int_0^t h(z)dz = H(t). \tag{8.15}$$

Tällöin $H(t) = E(N(t))$ eli kumulatiivinen hasardifunktio on tapahtumien odotettu lukumäärä ja

$$N(t) = H(t) + e(t), \tag{8.16}$$

missä $e(t)$ on virhetermi, jonka odotusarvo on nolla. Koska virhetermejä ei kuitenkaan havaita (kuten ei tavallisessa lineaarisessa regressioanalyysissäkään), on tarkastelu perustettava residuaaleihin

$$\hat{e}(t) = N(t) - \hat{H}(t). \tag{8.17}$$

Työllistyneille residuaali

$$\hat{e}(t) = 1 - \hat{H}(t), \tag{8.18}$$

missä t on havaittu työttömyyden kesto. Sensuroiduille

$$\hat{e}(t) = 0 - \hat{H}(c), \tag{8.19}$$

missä c on sensurointiaika. Residuaalit kuvaavat siis havaittujen ja ennustettujen tapahtumien lukumäärän eroa seurannan yksilökohtaisella päättämishetkellä. Mikäli malli on oikein spesifioitu, tulisi estimoidun kumulatiivisen hasardifunktion olla lähellä ykköstä tapahtumahetkellä . Työllistyneiden residuaalien tulisi siis olla lähellä nollaa työttömyyden kestosta riippumatta. Koska pyrkimyksenä on selittää nimenomaan työllistymistä eikä sensurointia ja toisaalta sensuroinnin oletetaan olevan työllistymisen suhteen epäinformatiivista, on sensuroidun yksilön kumulatiivisen hasardifunktion arvo sensurointihetkellä sama kuin (selittävien tekijöiden suhteen) samanlaisen, kyseisenä ajankohtana työllistyneen yksilön. Koska kumulatiivinen hasardifunktio saa vain ei-negatiivisia arvoja, ovat sensuroitujen havaintojen residuaalit aina enintään nollan suuruisia.

Luku 9

Nuorten työllistymisen mallintamisesta

Tutkielman empiirisen osan tavoitteena on selvittää (Työministeriön työnhakijarekisterin tietojen puitteissa), mitkä tekijät vaikuttavat nuorten työllistymiseen ja kuinka työllistymistodennäköisyys vaihtelee työttömyyden keston suhteen. Nuorten työllistymistä on kuvattu osittain parametroidulla Coxin mallilla. Työllistymiseksi on määritelty työllistyminen avoimille työmarkkinoille joko omatoimisesti tai työvoimatoimiston välityksellä. Työllistämistoimenpitein työllistymistä säätelee lähinnä lainsäädäntö, eikä sitä ole mielekästä tarkastella samanlaisena tapahtumana kuin avoimille työmarkkinoille työllistymistä. Muita mahdollisia työttömyyden päättymissyitä ovat nuorilla esim. asepalveluksen tai opintojen aloittaminen. Kilpailevien riskien mallin (ks. luku 7) mukaisesti pidetään muutoin kuin kiinnostuksen kohteena olevaan tapahtumaan (avoimille työmarkkinoille työllistyminen) päättyneitä työttömyysjaksoja sensuroituina.

9.1 Nuoriin kohdistettu työvoimapolitiikka

1980-luvun alussa alettiin erityisongelmaisiin työttömiin kiinnittää enemmän huomiota. Nuorisotyöttömyyden ehkäisy muodostui tällöin erääksi työvoimapolitiikan painopistealueista. 1980-luvun puolivälissä omaksuttiin muista Pohjoismaista ajatus nuorten yhteiskuntatakuusta. Yhteiskuntatakuun ideana on edistää nuorten sijoittumista työmarkkinoille koulutuksen ja normaalien työvoimapalvelujen (työnvälitys avoimille työmarkkinoille) avulla. Viimeisenä vaihtoehtona on nuoren työllistäminen yhteiskunnan tuella. Yhteiskuntatakuu on suunnattu alle 25-vuotiaille siten, että se järjestetään 16-17-vuotiaille pääasiassa koulutustakuuna ja 18-24-vuotiaille sekä

koulutus- että työllistämistakuuna. Yhteiskuntatakuuseen kuuluvat mm. peruskoulun 10. luokka, oppisopimuskoulutus sekä erilaiset tukimuodot nuorten työelämään tutustuttamiseksi.

Vuoden 1987 työllisyyslaki vahvisti nuorten ja pitkäaikaistyöttömien asemaa työvoimapolitiikassa. Lain mukaan on alle 20-vuotiaan nuoren kotikunnan järjestettävä tälle kolmen kuukauden työttömyyden jälkeen työ- tai harjoittelupaikka kuudeksi kuukaudeksi, mikäli nuori ei ole päässyt koulutukseen tai saanut työtä tavanomaisten työvoimapalvelujen avulla[1]. Samoin on valtion tai kunnan järjestettävä pitkäaikaisja toistuvaistyöttömille työtä kuudeksi kuukaudeksi[2].

Niin sanottu marssijärjestys määrittelee työmahdollisuuksien järjestämisessä noudatettavat periaatteet sekä toimenpiteiden etenemisjärjestyksen. Sen mukaan työvoimatoimistoon työttömäksi ilmoittautunut henkilö pyritään ensi sijassa työllistämään työssäkäyntialueensa avoimiin työpaikkoihin [3] tai ohjaamaan koulutukseen. Seuraavana toimenpiteenä on työttömän työllistäminen ns. harkinnanvaraisten työllisyysmäärärahojen avulla. Viimeisenä keinona on velvoitetyöllistettävän työllistäminen valtion tai kunnan palvelukseen. Harkinnanvarainen työllistäminen ja velvoitetyöllistäminen ovat ns. palkkaperusteisia työllistämistoimenpiteitä. Työllistämistoimenpiteet pyritään suuntaamaan vaikeimmin työllistyvien työllistämiseen.

Nuorisotyöllisyysryhmän muistion (1992) mukaan kokemukset erilaisista työllistämistoimenpiteistä eivät ole kovinkaan rohkaisevia. Työryhmä toteaa, etteivät työllistämistoimenpiteet juuri paranna nuorten asemaa avoimilla työmarkkinoilla. Yhteiskuntatakuun periaatteiden mukaan nuorisotyöttömyyden hoidossa tulisi käyttää mahdollisimman normaaleja, ennaltaehkäiseviä toimenpiteitä, jotka parantavat pysyvästi nuoren asemaa työmarkkinoilla. Ammatillisen peruskoulutuksen hankkiminen yleisen koulujärjestelmän puitteissa tulisi olla työttömälle nuorelle aina ensisijainen vaihtoehto. [4] Käytännössä nuorisotyöttömyyttä on hoidettu pitkälti palkkaperusteisin työllistämistoimenpitein. Vuonna 1984 palkkaperusteisin toimenpitein työllistetyistä lähes 80 % oli alle 25-vuotiaita (Nio 1985). Nuorten osuus palkkaperusteisin työllistämistoimenpitein työllistetyistä on kuitenkin vähentynyt

[1] Velvoitetyötä edeltävää työttömyysaikaa pidennettiin säästöjen aikaansaamiseksi 1.8.1992 lähtien kuuteen kuukauteen. Nuorten ja pitkäaikaistyöttömien työllistämisvelvoite poistettiin laista vuoden 1993 alussa. Lainmuutoksella pyritään siirtämään määrärahoja velvoitejärjestelmästä harkinnanvaraiseen työllistämiseen ja koulutukseen.

[2] Pitkäaikaistyöttömäksi määritellään henkilö, joka on ollut yhtäjaksoisesti työttömänä 12 kk. Toistuvaistyöttömäksi määritellään henkilö, joka on ollut viimeisten kahden vuoden aikana työttömänä yhteensä 12 kk. Työllisyyslakia muutettiin 1.1.1992 lähtien siten, että pitkäaikaistyöttömäksi määritellään myös henkilö, joka on ollut viimeisten 13 kk aikana työttömänä 12 kk. Samalla poistettiin laista toistuvaistyöttömien työllistämisvelvoite.

[3] Työssäkäyntialueella tarkoitetaan henkilön asuinkuntaa sekä niitä kuntia, joissa asuinkunnasta käydään yleisesti työssä.

[4] Työvoimahallinnon järjestämä työllisyyskoulutus on tarkoitettu lähinnä aikuisten täydennys- ja uudelleenkouluttamiseksi.

työvoimapolitiikan painopisteen siirryttyä vuosikymmenen vaihteessa pitkäaikaistyöttömiin. Vuonna 1992 alle 25-vuotiaiden osuus palkkaperusteisin työllistämistoimenpitein työllistetyistä oli enää alle 40 %.

9.2 Tutkimusaineisto

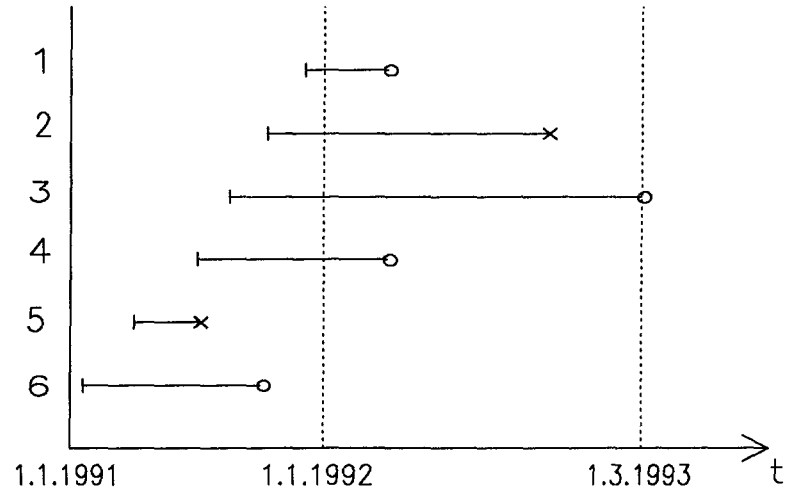
9.2.1 Otantatapa

Tutkimuksen perusjoukkona ovat vuonna 1991 työttömäksi tulleet, vuoden ensimmäisen työttömyysjakson alkaessa alle 30-vuotiaat työvoimatoimistoon työttömiksi työnhakijoiksi ilmoittautuneet nuoret. Tutkimusaineisto poimittiin työnhakijarekisteristä systemaattisella otannalla. Perusjoukko järjestettiin sukunimen 10 ensimmäisen merkin perusteella aakkosjärjestykseen. Saman sukunimen omaavat asetettiin järjestykseen syntymäpäivän päiväosan perusteella. Järjestetystä perusjoukosta poimittiin joka 50. Aineiston kooksi tuli tällöin 1794 havaintoa. Otokseen valittuja henkilöitä seurataan vuoden 1991 ensimmäisen työttömyysjakson ajan, kuitenkin enintään otoksen poiminta-ajankohtaan, vuoden 1993 maaliskuun alkuun saakka.

Yksilöt tulevat seurannan piiriin työttömäksi joutuessaan ja poistuvat seurannasta työttömyyden päättyessä tai viimeistään seuranta-ajan (1.1.1991-1.3.1993) päättyessä. Kuva 9.1 havainnollistaa tutkimuksen seuranta-asetelmaa. Kuvassa yksilöt 1, 4 ja 6 ovat katotapauksia ja yksilö 3 on sensuroitu seuranta-ajan päättymisen vuoksi. Yksilöt 2 ja 5 ovat työllistyneet seuranta-aikana. Otantatapaa, jossa yksilöt tulevat seurannan piiriin työttömyyden alkaessa, kutsutaan **virtaotannaksi. Varanto-otannassa** seurattavat poimitaan tiettynä ajankohtana poimintakriteerit täyttävien yksilöiden joukosta. Otokseen voitaisiin esim. poimia 1.1.1991 työttömänä olevia, työvoimatoimistoon ilmoittautuneita työnhakijoita. Varanto-otoksien ongelmana on se, että otokseen valikoituu liikaa pitkiä työttömyysjaksoja, ts. pitkät työttömyysjaksot ovat otoksessa yliedustettuina. Mitä pidempi työttömyysjakso on, sitä todennäköisemmin se osuu otoksen poiminta-ajankohdalle.

Osalle seurattavista on työttömyyden päättymissyyksi merkitty **muu syy tai ei tietoa**. Nämä ovat nuoria, jotka ovat lopettaneet yhteydenpidon työvoimatoimistoon (työnhaun jatkuminen edellyttää säännöllistä yhteydenpitoa sovituin väliajoin).[5] Osa näistä on luultavasti henkilöitä, joiden työnhaku on päättynyt ennen sovittua ilmoittautumispäivää ja jotka eivät ole viitsineet tai muistaneet ilmoittaa päättäneensä työnhaun. Nuorten elämäntilanteet muuttuvat nopeasti. Moni työnhakijaksi ilmoittautunut päättääkin aloittaa opinnot tai käydä ensin armeijan eikä välttämättä

[5] "Muu syy" tarkoittaa esim. vankilaan joutumista, mikä lienee harvinainen työttömyyden päättymissyy.



Kuva 9.1: Tutkimuksen seuranta-asetelma: esimerkkinä yksilöiden 1-6 havaitut työttömyysjaksot. x = tapahtuma, o = sensurointi

muista ilmoittaa työvoimatoimistoon lopettaneensa työnhaun. Osa yhteydenpidon katkaisseista on löytänyt työpaikan muutoin kuin työvoimatoimiston kautta. Suurin osa ryhmään kuuluvista on kuitenkin työvoimatoimiston virkailijoiden mukaan nuoria, jotka ovat yksinkertaisesti unohtaneet ilmoittautua työvoimatoimistoon ennalta sovittuna ajankohtana. Jos asiakas ei ilmoittaudu työvoimatoimistoon sovittuna ajankohtana, merkitään työnhaku (ja työttömyys) päättyneeksi. Työttömyyden päättymissyy jää tuntemattomaksi. Jos henkilö on unohtanut ilmoittautua työvoimatoimistoon ja ilmoittaa myöhemmin jatkavansa edelleen työnhakua, [6] merkitään uusi työnhaku alkaneeksi.

Lain mukaan tulee valtion tai kunnan tarjota kaikille pitkäaikaistyöttömille työntekomahdollisuus. Suurimmalla osalla aineiston yli vuoden työttömänä olleista nuorista työttömyys päättyykin työllistämistoimenpitein työllistymiseen. Vain kolmen yli vuoden työttömänä olleen nuoren työttömyys päättyy avoimille työmarkkinoille työllistymiseen, joten avoimille työmarkkinoille työllistymisen tutkiminen yli vuoden pituisilla työttömyysjaksoilla ei ole mielekästä. Tämän vuoksi yli 365 päivää kestävät työttömyysjaksot on sensuroitu 365 päivän kohdalla. Näiden havaintojen sisältämä informaatio on, kuten sensuroitujen havaintojen yleensä, että työttömyysjakso on pidempi kuin sensurointiaika (365 päivää). Tällaista sensurointia kutsutaan I lajin sensuroinniksi (ks. kpl 2.1).

[6] Työvoimatoimistosta menee tieto työnhaun päättymisestä Kelaan ja työttömyyskassoille, jotka lopettavat työttömyyskorvausten maksamisen; ilmoittautumisvelvollisuus muistetaan viimeistään tällöin.

Sensuroituja havaintoja ovat myös katotapaukset eli ennen seuranta-ajan päättymistä seurannasta poistuneet henkilöt, joiden työttömyyden päättymissyytä ei tunneta. Ryhmän "työttömyyden päättymissyy muu syy tai ei tietoa" havainnoista suurin osa on katotapauksia. Tapaukset, joissa henkilö on ollut ilmoittautumisajankohtana yhä työtön ja unohtanut ilmoittautua, ovat "aitoja" sensuroituja havaintoja, koska näiden henkilöiden työttömyysjakso on pidempi kuin sensurointiaika (aikaväli työttömyyden alusta ilmoittautumispäivään). Tapaukset, joissa henkilön työnhaku on päättynyt ennen ilmoittautumispäivää eikä työnhaun päättymisestä ole jostain syystä ilmoitettu työvoimatoimistoon, ovat ongelmallisempia, koska tällaisten havaintojen työttömyyden kesto on aina sensurointiaikaa lyhyempi eikä, kuten sensuroiduista kestoista oletetaan, sensurointiaikaa pidempi. Näitä ongelmallisia katotapauksia ei ole kuitenkaan mahdollista erottaa muista katotapauksista ja toisaalta suurimmalla osalla työttömyyden todellinen kesto on vähintään havaitun pituinen. Oletetaan, että katotapausten sensuroituminen on epäinformatiivista työllistymisen suhteen.

9.2.2 Aineiston muuttujat

Jokaiselle työnhakijaksi ilmoittautuvalle tehdään työvoimatoimistossa perushaastattelu, jonka tarkoituksena on työnhakijan elämäntilanteen kartoittaminen. Tässä yhteydessä tallennetaan työnhakijarekisteriin mm. työnhakijan koulutukseen, työhistoriaan ja työnhakuun liittyviä tietoja. Tämän jälkeen työnhakija ilmoittautuu työvoimatoimistoon sovituin väliajoin. Ilmoittautumiskerroilla päivitetään tarvittaessa työnhakijarekisterin tietoja. Ilmoittautumisväli saattaa vaihdella hyvinkin paljon työvoimatoimistoittain, jaostottain sekä työllisyystilanteesta riippuen. Työttömillä työnhakijoilla ilmoittautumisväli on tavallisesti muita työnhakijoita lyhyempi. Helsingin Haapaniemenkadun työvoimatoimiston alle 25-vuotiaiden ammattitaidottomien nuorten jaostossa työttömän työnhakijan ilmoittautumisväli oli syksyllä 1993 noin 5 kuukautta.

Kullekin tiedolle (muuttujalle) on varattu työnhakijarekisterissä oma ns. **yksilötaulu**, joka sisältää tiedon sekä päivämäärän, josta ilmenee, milloin tieto on tallennettu. Henkilötunnus liittää taulun oikeaan työnhakijaan. Taulut saattavat sisältää useita eri aikoina tallennettuja muuttujan arvoja: esimerkiksi työnhakijan muuttaessa päivitetään asuinkunta-tauluun uusi asuinkunta ja muuttopäivämäärä vanhan ohelle. Tutkimusaineiston muuttujien arvot on voitu valita päivityspäivämäärien ansiosta pääosin siten, että muuttujat kuvaavat työnhakijan tilannetta ennen työttömyyden alkaessa. Osa muuttujista on kuitenkin yhdistetty samaan tauluun: esimerkiksi työnhakuun liittyvät muuttujat (**hakal, työaikat, haked, työkok**) sijaitsevat kaikki työnhaku-taulussa, joka ei sisällä muuttujakohtaisia päivityspäivämääriä, vaan ainoastaan työnhaun alkamis- ja päättymispäivät. Työnhakuun liittyvät tiedot tallennetaan työnhakijaksi ilmoittauduttaessa. Jos työnhakija muuttaa esimerkiksi

työaikatoivettaan kesken työnhaun, päivitetään uusi tieto vanhan päälle. Työnhakutaulun muuttujat liittyvät siis ennen kaikkea siihen työnhakuun, jonka aikaista työttömyysjaksoa tutkitaan (vuoden 1991 ensimmäinen työttömyysjakso). On kuitenkin luultavaa, että esimerkiksi työaikatoive säilyy suurimmalla osalla työnhakijoista samana saman työnhaun aikana (osin jo pitkien ilmoittautumisvälien vuoksi) ja siten taulun muuttujat kuvaavat oleellisesti tilannetta ennen työttömyyden alkua. Toinen samantyyppinen, useita muuttujia ja vain yhden päivityspäivämäärän sisältävä taulu liittyy työnhakijan henkilötietoihin. Taulun tietoja on voitu periaatteessa päivittää otoksen poiminta-ajankohtaan, vuoden 1993 maaliskuuhun saakka. Suurin osa taulun tiedoista on kuitenkin luonteeltaan pysyviä. Käytännön merkitystä muuttujien päivitysajankohdalla työttömyyden kestoa selitettäessä lienee ainoastaan työttömyyskassan jäsenyyttä kuvaavan muuttujan (**jäsen**) osalta. Tämän muuttujan arvo on voinut muuttua työttömyysjakson alun tilanteesta, mikäli henkilö on ollut työvoimatoimiston asiakkaana työttömyyden päättymisen jälkeen. Tämä on syytä pitää mielessä tutkittaessa muuttujien vaikutusta työllistymistodennäköisyyteen.

Aineisto sisälsi melko paljon puuttuvia tietoja työnhakualuetta, kansalaisuutta ja työaikatoivetta kuvaavien muuttujien osalta. Raaka-aineiston 1794 havainnosta 399 havainnolta puuttui tieto työnhakualueesta, 365 tieto kansalaisuudesta ja 115 tieto työaikatoiveesta. Puuttuvien tietojen suuri määrä näiden muuttujien osalta oli havaittu myös Työministeriössä, jossa syyksi epäiltiin virhettä tietojen tallennusohjelmassa. Ymmärtääkseni ohjelman toimintaa ei ole kuitenkaan tutkittu ongelman ratkaisemiseksi ja kyseessä on vain oletus puuttuvien tietojen taustalla olevasta syystä. Tietojen tallennusohjelma käsittelee tietoja ainakin neljällä eri tavalla: osalla muuttujista on tallennusohjelmassa oletusarvo, toisin sanoen muuttujan arvoksi tulee automaattisesti sen oletusarvo, mikäli virkailija jättää tiedon tallentamatta. Oletusarvot ovat muuttujien tavallisimpia arvoja. Toisaalta muuttujat jakautuvat ns. pakollisiin ja ei-pakollisiin tietoihin. Pakollista tietoa ei ole mahdollista jättää tyhjäksi "enteriä" painamalla tai pyyhkimällä oletusarvo pois. Ei-pakollinen tieto voidaan sen sijaan jättää tyhjäksi. Työnhakijarekisterissä on siis ainakin neljän tyyppisiä tietoja:

1 pakolliset, oletusarvon omaavat tiedot,

2 ei-pakolliset, oletusarvon omaavat tiedot,

3 pakolliset tiedot, joilla ei oletusarvoa,

4 ei-pakolliset tiedot, joilla ei oletusarvoa.

Kansalaisuutta ja työnhakualuetta kuvaavat muuttujat kuuluvat luokkaan 2 [7]. On

[7]Työnhakijarekisterin kansalaisuutta kuvaavan muuttujan **kansah** oletusarvo on 246 (Suomen kansalainen) ja työnhakualuetta kuvaavan muuttujan **hakal** oletusarvo 1 (hakee työtä asuinpaikkakunnalta).

siis teoriassa mahdollista, että työvoimatoimiston virkailija on halunnut jättää kohdan tyhjäksi ja pyyhkinyt oletusarvon pois. Työnhakijarekisterin äidinkieltä kuvaavan muuttujan (**kiekd**) perusteella yhtä havaintoa lukuunottamatta kaikilla puuttuvan kansalaisuustiedon omaavilla havainnoilla oli äidinkielenä suomi tai ruotsi, joten luultavasti kyseessä on Suomen kansalaisia. On vaikea kuvitella, miksi virkailija haluaisi tallentaa puuttuvan tiedon tällaisten havaintojen osalta, etenkin, kun havaintoja on aineistossa melko paljon. Myös työnhakualue-tiedon tyhjäksi jättämiselle on vaikea löytää syytä. Eräs mahdollinen syy puuttuville tiedoille voisi olla se, että virkailija on vahingossa pyyhkinyt oletusarvon pois. Toisaalta syynä voi olla ohjelmavirhe, jonka seurauksena oletusarvo on jäänyt tallentumatta tyhjäksi jätettyyn kohtaan. Työaikatoivetta kuvaava muuttuja kuuluu luokkaan 1 eli on pakollinen, oletusarvon omaava tieto [8]. Tämän muuttujan osalta ei oletusarvon pois pyyhkiminen siten pitäisi olla edes mahdollista, joten ainoaksi selitykseksi puuttuville tiedoille jäänee tallennusohjelman virheellinen toiminta.

Syytä rekisterin puuttuville tiedoille ei siis tiedetä tarkkaan. On mahdollista, että puuttuvien tietojen taustalla on jokin muu syy kuin edellä on esitetty. Käytettävissä olevien tietojen perusteella on tutkimuksessa oletettu, että puuttuvat tiedot johtuvat tallennusohjelman virheestä tai siitä, että virkailija on vahingossa pyyhkinyt oletusarvon pois. Oletuksen perusteella puuttuvat tiedot täydennettiin oletusarvoillaan.

Havaintoja poistettiin puuttuvien tai virheellisten tietojen vuoksi yhteensä 30 kpl. Lopullisen aineiston koko on siten 1764 havaintoa. 671 havainnon työttömyys päättyy työllistymiseen avoimille työmarkkinoille ja 1093 (62 %) päättyy muulla tavalla tai sensuroituu. Sensuroituja havaintoja on 556 (32 %). Suurin osa sensuroiduista havainnoista, 394 kpl on katotapauksia. Seuranta-ajan päättymisen vuoksi sensuroituja havaintoja on 162 kpl. Siis lähes 10 % aineiston nuorista on ollut vähintään vuoden työttömänä. Näistä nuorista 101 työllistettiin työllistämistoimenpitein pitkäaikaistyöttömien työllistämisvelvoitteen mukaisesti. Liitteessä 1 kerrotaan tutkimusaineiston muuttujien konstruoinnista. Muuttujat on muodostettu työnhakijarekisterin tietojen perusteella. Suurin osa muuttujista on indikaattorimuuttujia (muuttujat **ikä**, **työkok** ja **aluekys** ovat jatkuvia). Liitteen 1 lopussa on luettelo lopullisista, estimoinneissa käytetyistä muuttujista.

Liitteessä 2 on muuttujien suoria jakaumia pylväskuvioina. Kuvassa 2.10 on työttömyyden keston jakauma. Nuorten työttömyysjaksot ovat tavallisesti melko lyhyitä, mikä näkyy myös kuvassa: työttömyysjaksoista noin 50% on enintään 12 viikon pituisia. Pitkäaikaistyöttömien työttömyysjaksot eivät näy kuvassa. Alle 20-vuotiaiden työllistämisvelvoite näkyy selkeästi kuvassa 2.11: työllistämistoimenpitein työllistyminen on suurin yksittäinen työttömyyden päättymissyy. 35 % alle

[8] Työnhakijarekisterin työaikatoivetta kuvaavan muuttujan **tatoive** oletusarvo on 1 (kokopäivätyö).

20-vuotiaiden enintään vuoden pituisista työttömyysjaksoista päättyy työllistämistoimenpitein työllistämiseen. Avoimille työmarkkinoille työllistymiseen päättyneitä jaksoja on 30 %. 20-24-vuotiaiden ja 25-29-vuotiaiden välillä ei ole juuri eroja työttömyyden päättymissyiden suhteen: vajaa 50 % työttömyysjaksoista päättyy työllistymiseen omatoimisesti tai työvoimatoimiston kautta, n. 14 % päättyy työllistämistoimenpitein työllistymiseen tai työllisyyskurssin aloittamiseen, ja runsas 10 % päättyy työvoimasta poistumiseen. Työvoimatoimiston rooli työnvälittäjänä vaikuttaa kuvan perusteella melko vähäiseltä: kaikissa ikäryhmissä suurin osa avoimille työmarkkinoille työllistyneistä on hankkinut työpaikkansa itse. Työvoimatoimisto on työllistänyt seuranta-aikana 20-29-vuotiaita suunnilleen yhtä paljon avoimille ja suljetuille (valtion ja kunnan tukemat työpaikat) työmarkkinoille. Alle 20-vuotiaita on – työllisyyslain velvoitteista johtuen – työllistetty lähinnä suljetuille työmarkkinoille. Katotapausten määrä on kaikissa ikäryhmissä melko suuri, noin neljännes havainnoista.

9.3 Muuttujien vaikutustavan alustava tarkastelu

Työnhakijarekisterin muuttujien vaikutusta työllistymistodennäköisyyteen tarkasteltiin ennen varsinaista mallintamista Kaplan–Meier-menetelmällä estimoitujen eloonjäämisfunktioiden avulla. Liitteessä 3 esitetään työttömyyden keston eloonjäämisfunktioita iän, sukupuolen ja työttömyyskassan jäsenyyden, koulutusasteen, työkokemuksen ja työnhakua edeltävän toiminnan mukaan ositetulle aineistolle. Eloonjäämisfunktio $S(t)$ kuvaa todenäköisyyttä, että työnhakijan työttömyysjakso on vähintään t:n pituinen. Log-log-eloonjäämisfunktioiden eli log-kumulatiivisten hasardifunktioiden ($\log(-\log(S(t))) = \log(H(t))$) avulla voidaan tutkia suhteellisten hasardien mallin proportionaalisuusoletuksen sopivuutta aineistoa osittavalle muuttujalle: mikäli oletus pitää paikkansa, tulisi käyrien vertikaalisen erotuksen olla likimain vakio.

Kuvissa 3.1 ja 3.2 on aineisto ositettu iän mukaan. Työllistymistodennäköisyydessä ei vaikuta olevan suuria eroja ikäryhmien välillä. Log-Rank- ja Wilcoxon-testien mukaan ikäryhmien eloonjäämiskäyrät eivät poikkea toisistaan, sen sijaan uskottavuusosamäärätestin mukaan erot ovat merkitseviä. Kuvissa 3.3 ja 3.4 tutkitaan samanaikaisesti työttömyyskassan jäsenyyden ja sukupuolen vaikutusta työllistymistodennäköisyyteen. Naisten työllistymistodennäköisyys on kuvan 3.3 perusteella selvästi suurempi kuin miesten. Työttömyyskassan jäsenyys näyttää vaikuttavan miehillä työllistymistodennäköisyyttä lisäävästi, kun taas naisilla jäsenyydellä ei ole lainkaan merkitystä. Testien mukaan ositteiden eloonjäämisfunktiot poikkeavat toisistaan merkitsevästi. Kuvan 3.4 log-kumulatiivisten hasardifunktioiden kuvaajien

vertikaalinen erotus säilyy noin neljännestä työttömyysviikosta lähtien likimain vakiona. Tätä lyhyemmillä työttömyyden kestoilla kuvaa on vaikea tulkita: käyrät nousevat jyrkästi ja kulkevat hyvin lähellä toisiaan. Proportionaalisuusoletus näyttää sopivan aineistoa osittaville muuttujille ainakin suurimmalla osalla tarkasteltavaa aikaväliä.

Kuvissa 3.5 ja 3.6 on aineisto ositettu koulutusasteen mukaan. Kuvien selkeyden säilyttämiseksi koulutusasteet on ryhmitelty luokkiin peruskoulu, lukio, keskiaste ja korkea-aste. Ainoastaan peruskoulun käyneiden työllistymistodennäköisyys on muiden ryhmien työllistymistodennäköisyyttä selvästi pienempi. Lukion, keskiasteen tai korkea-asteen koulutuksen vaikutuksella työllistymistodennäköisyyteen ei ole juuri eroa. Peruskoulun käyneiden log-kumulatiivisen hasardifunktion kuvaajan erotus muiden ryhmien kuvaajiin on lyhimpiä työttömyyden kestoja lukuunottamatta suunnilleen vakio. Kuvan 3.7 sekä testien perusteella ei työkokemuksella näyttäisi olevan vaikutusta työllistymiseen.

Kuvien 3.9 ja 3.10 mukaan työnhakua edeltävä toiminta vaikuttaa varsin paljon työllistymistodennäköisyyteen. Opiskelemasta tulleet ovat kuvan perusteella parhaassa asemassa työnsaannin suhteen, vaikka erot armeijasta tai työelämästä tulleisiin ovat pieniä. Testien perusteella eloonjäämiskäyrät poikkeavat toisistaan merkitsevästi. Proportionaalisuusoletus näyttää sopivan erityisen huonosti työllisyyskoulutuksesta tulleille: ryhmän log-kumulatiivisen hasardin kuvaaja kasvaa alussa nopeimmin ja leikkaa pahasti muiden ryhmien kuvaajia. Sama informaatio näkyy kuvassa 3.9: vaikka työllisyyskoulutuksesta tulleiden työllistymistodennäköisyys on enimmäkseen melko matala, työllistyvät he muutaman ensimmäisen työttömyysviikon ajan muita ryhmiä paremmin (ryhmän eloonjäämisfunktio laskee alussa jyrkimmin). Muille ryhmille proportionaalisuusoletus näyttää sopivan kohtuullisen hyvin, vaikka käyrät leikkaavat alussa jonkin verran toisiaan.

9.4 Estimointituloksista

Estimoinnit tehtiin semiparametrisella Coxin mallilla (ks. kpl 6.2). Coxin malli kuuluu suhteellisten hasardien malleihin, joissa selittävät tekijät vaikuttavat multiplikatiivisesti työttömyyden keston hasardifunktioon eli työllistymisintensiteettiin. Työllistymisintensiteetti pienellä dt:n pituisella aikavälillä; $h(t)dt$ kuvaa (likimain) ehdollista työllistymistodennäköisyyttä kyseisellä aikavälillä, ehdolla, että työttömyys on kestänyt aikavälin alkuun. Suuri regressiokertoimen estimaatti merkitsee suurta (ehdollista) työllistymistodennäköisyyttä ja siten nopeaa työllistymistä.

Liitteen 4 mallissa 1 (taulukko 4.1) on mukana kaikki käytettävissä olleet, työllistymistodennäköisyyttä mahdollisesti selittävät tekijät. Mallia on vaikea tulkita muut-

tujien suuren lukumäärän vuoksi. Suurin osa muuttujien regressiokertoimista on tilastollisesti ei-merkitseviä. Estimointien tavoitteena oli löytää muuttujajoukosta parhaiten työllistymistä selittävät tekijät. Estimointikokeiluja tehtiin sekä poistamalla kaikki muuttujat sisältävästä mallista ei-merkitseviä muuttujia että lähtemällä liikkeelle mahdollisimman yksinkertaisesta mallista ja laajentamalla sitä uusia muuttujia lisäämällä.

Lukuisten estimointikokeilujen[9] perusteella päädyttiin taulukossa 9.1 esitettyyn malliin (malli 2). Mallista on poistettu testien perusteella ei-merkitsevät (5 % merkitsevyystasolla) muuttujat. Muuttujat **työllkou**, **mjäsen** ja **ntyöraj** haluttiin kuitenkin säilyttää mallissa mielenkiinnon vuoksi, vaikka näiden muuttujien merkitsevyystaso ei aivan saavuta mainittua 5%:a.

Kuten alustavien tarkastelujen perusteella oli odotettavissa, ei iän vaikutus työllistymiseen ollut merkitsevä. Estimoinneissa ikää kokeiltiin sekä jatkuvana (muuttuja **ikä**) että indikaattorimuuttujana (muuttujat **ryhmä2** ja **ryhmä3**). Sen sijaan sukupuoli vaikuttaa työllistymiseen erittäin merkitsevästi. Mallin 1. mukaan naisten työllistymistodennäköisyys on miehiin verrattuna 1,5-kertainen (taulukon 9.1 sarake riskisuhde) [10]. Työkokemusta, työaikatoivetta ja työnhakualuetta kuvaavat muuttujat eivät estimoitujen mallien perusteella vaikuta työllistymiseen. Myöskään ruotsinkielisyys tai ulkomaalaisuus ei vaikuta työllistymiseen.

Kuten alustavien tarkastelujen perusteella nähtiin, vaikuttaa työttömyyskassan jäsenyys miehillä ja naisilla eri tavalla: jäsenyys lisää miesten työllistymistodennäköisyyttä 7 % merkitsevyystasolla (muuttuja **mjäsen**), kun taas naisille jäsenyydellä ei ole merkitystä (muuttuja **njäsen** poistettiin mallista tämän vuoksi). Ilmiötä kutsutaan muuttujien interaktioksi eli yhdysvaikutukseksi. Myös vajaakuntoisuus näyttäisi vaikuttavan eri tavalla miehillä ja naisilla: miehillä vajaakuntoisuus vaikuttaa voimakkaammin työllistymistodennäköisyyttä heikentävästi kuin naisilla (muuttujat **mtyöraj** ja **ntyöraj**). Raskasta ruumiillista työtä, jossa vakavien loukkaantumisien riskit ovat suuret, tekevät etupäässä miehet; ero muuttujan vaikutuksessa voi johtua miesten ja naisten eriasteisesta vajaakuntoisuudesta. Samoin alkoholismista kärsivät useammin miehet kuin naiset.

Estimoinneissa tutkittiin myös, vaikuttaako ikä tai työkokemus eri tavalla eri sukupuolilla. Kuten muuttujien päävaikutukset, eivät interaktiotermitkään olleet merkitseviä. Koulutusasteen vaikutusta työllistymiseen tutkittiin indikaattorimuuttujilla **lukio, alkeski, ylkeski, alkorkea, alkandi** ja **akateem**. Vertailuryhmänä olivat vain peruskoulun suorittaneet. Lukio-, keskiasteen tai alemman korkea-asteen tutkinnon omaaminen kasvattaa estimointitulosten mukaan merkitsevästi työllisty-

[9] Estimoitujen mallien kokonaismäärä on noin 200, joista tutkielmassa esitetään kolme

[10] Riskisuhde = $\exp(\hat{\beta}_{nainen})$ kuvaa naisten työllistymistodennäköisyyttä suhteessa vertailuryhmäänsä miehiin (ks. kpl 5.2)

mistodennäköisyyttä peruskoulun käyneisiin verrattuna. Sen sijaan alemman kandidaattiasteen tai akateemisen koulutuksen omaavat eivät näyttäisi poikkeavan peruskoulun käyneistä. Sallittaessa koulutusasteen vaikuttaa eri tavalla eri ikäryhmissä huomataan, että 20-24–vuotiailla akateeminen koulutus kasvattaa hyvin voimakkaasti työllistymistodennäköisyyttä. Sen sijaan 25-29–vuotiaat akateemiset eivät erotu peruskoulun käyneistä. Waldin testin mukaan muuttujien **r2akateem** ja **r3akateem** regressiokertoimet poikkeavat toisistaan 5 % merkitsevyystasolla (kyseistä testiä ei ole raportoitu tutkielmassa). Alle 20-vuotiaat eivät ole vielä ehtineet hankkia akateemista koulutusta. Koska aineiston 20-24–vuotiaista akateemisia on ainoastaan 3 kpl ja 25-29–vuotiaista 22 kpl, kuvannee tulos lähinnä tutkittua aineistoa eikä tulosta voida yleistää koskemaan kaikkia nuoria. Mallissa 2. koulutusasteiden luokitusta on karkeistettu siten, että muuttuja **lukkeski** kuvaa lukio- tai keskiasteen koulutusta ja **mkorkea** alemman korkea-asteen tai alemman kandidaattiasteen koulutusta. Koulutusasteen vaikutuksessa ei ole eroja sukupuolten välillä. Koulutusaloista kauppa ja tekniikka heikentävät mallin mukaan työllistymistodennäköisyyttä. Hoitoalan koulutus näyttäisi sitä vastoin parantavan asemia työmarkkinoilla. Tulos kuvannee lähinnä työvoiman kysyntätilannetta laskusuhdanteen alussa. Hoitoalan koulutuksen omaavat sijoittuvat yleensä julkiselle sektorille, jossa laman vaikutukset alkoivat näkyä myöhemmin kuin yksityisellä sektorilla. Muuttujien **kauppa** ja **tekniikk** regressiokertoimet ovat hyvin lähellä toisiaan, joten ne on yhdistetty muuttujaksi **kauptekn**. Työmarkkinat ovat eriytyneet sukupuolen mukaan siten, että esim. hoitoalan työntekijöistä pääosa on naisia ja tekniikan alalla puolestaan miehiä. Edellä esitetty tulos, jonka mukaan naisten työllistymistodennäköisyys on merkitsevästi suurempi kuin miesten, voi siten heijastaa koulutusalaa kuvaavien muuttujien tavoin työvoiman kysyntätilannetta seuranta-aikana; tyypillisiä naisten työpaikkoja on ollut kyseisenä aikana runsaammin tarjolla. Toisaalta, myös Kettusen (1991) estimointitulokset vuosien 1985-1986 työnhakijarekisteriin perustuvasta aineistosta ovat samansuuntaisia: naisten työllistymistodennäköisyys on suurempi kuin miesten. Muuttujan vaikutus ei ollut kuitenkaan Kettusen estimointitulosten mukaan merkitsevä.

Työnhakua edeltävistä toiminnoista äitiyslomalla oleminen tai lasten hoitaminen sekä "muu toiminta" pienentävät työllistymistodennäköisyyttä. Pienten lasten äidit eivät luultavasti mielellään ota vastaan lyhyitä "keikkatöitä" lasten hoitopaikan järjestelyvaikeuksien vuoksi; tämä saattaa olla syynä alhaiselle työllistymistodennäköisyydelle. Toinen syy voi olla työnantajien mahdolliset negatiiviset asenteet pitkään työelämästä poissa olleita naisia kohtaan. Työllisyyskoulutuksesta työnhakijoiksi tulleiden työllistymistodennäköisyys ei poikkea merkitsevästi vertailuryhmän (armeijasta, opiskelemasta tai työelämästä työnhakijoiksi tulleet sekä ns. ongelmataustan omaavat) työllistymistodennäköisyydestä. Koska työllisyyskoulutukseen valinnan kriteerinä on nimenomaan heikko työllistyvyys, voisi tuloksen tulkita siten, että työllisyyskoulutus on pystynyt parantamaan koulutettujen asemaa työmarkkinoilla.

Alueellinen kysyntä (työvoimapiirin työttömät työnhakijat / avoimet työpaikat) vaikuttaa työllistymiseen erittäin merkitsevästi: mitä enemmän työnhakijoita hakee kutakin työpaikkaa, sitä pienempi on työllistymistodennäköisyys. Estimoinneissa kokeiltiin myös työttömäksitulokuukautta indikoivia dummy-muuttujia. Tammikuussa työttömäksi tulleiden työllistymistodennäköisyys on merkitsevästi suurempi verrattuna muina kuukausina työttömäksitulleiden työllistymistodennäköisyyteen (muuttuja **tammi**). Tämä voi johtua suhdannetilanteen nopeasta heikkenemisestä vuoden 1991 aikana; vuoden alussa ei vielä osattu ennakoida talouden tilan vakavuutta ja työsuhteita solmittiin edellisten vuosien tapaan. Toisaalta moni opistotasoinen koulutus päättyy vuoden lopussa ja tämän vuoksi tammikuussa työmarkkinoille tulevien nuorten joukko on keskimäärin koulutetumpaa kuin muina aikoina työmarkkinoille tuleva (suurin osa aineiston nuorista on käynyt ainoastaan peruskoulun tai lukion). Muuttuja **tammi** saattaa siis osaksi kuvata (muuttujien **lukkeski, mkorkea** ja **r2akateem** lisäksi) myös koulutuserojen vaikutusta työllistymiseen.

Taulukko 9.1 Malli 2:
Työllistymisintensiteettiä selittävän mallin regressiokertoimien estimaatit

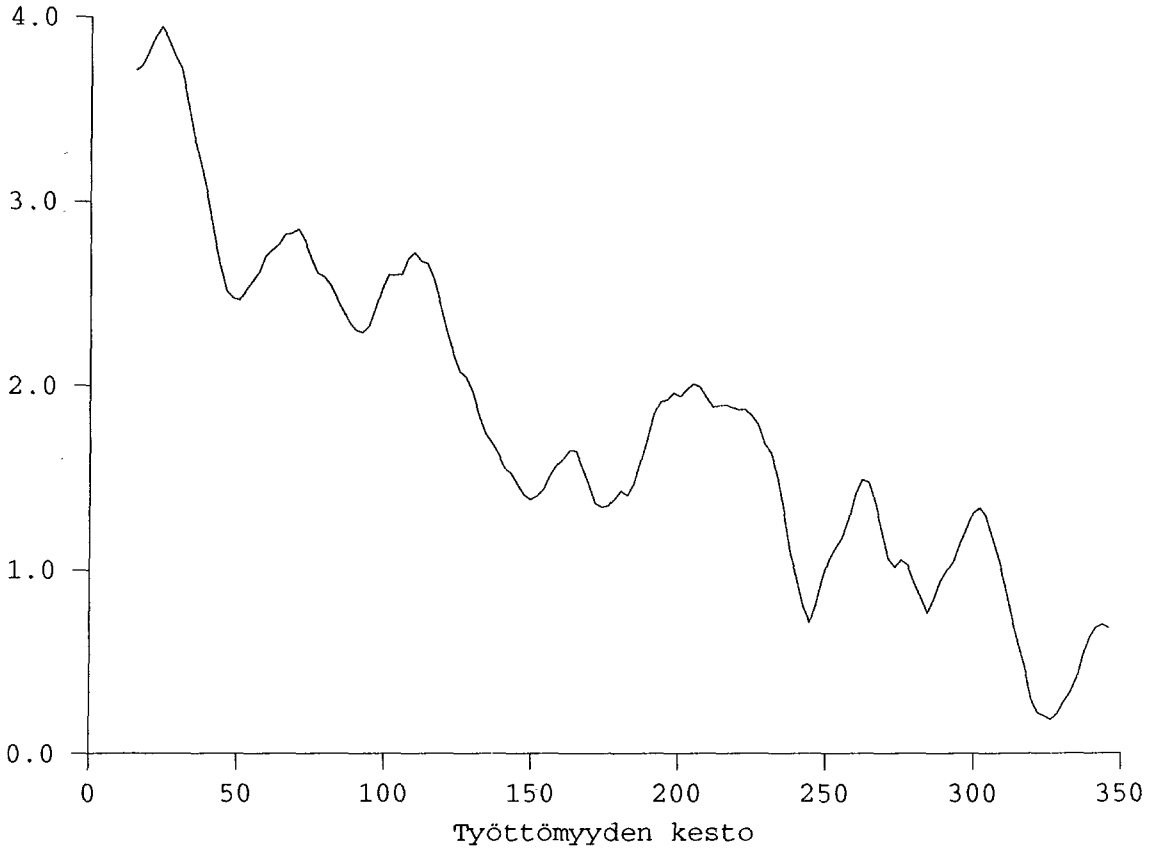
muuttuja	**estimaatti**	**keskivirhe**	**p-arvo**	**riskisuhde**
nainen	0.4440	0.0937	0.0001	1.559
mjäsen	0.2247	0.1187	0.0583	1.252
mtyöraj	-1.7333	0.7107	0.0147	0.177
ntyöraj	-0.6196	0.3388	0.0674	0.538
lukkeski	0.4813	0.1003	0.0001	1.618
mkorkea	0.7019	0.2082	0.0007	2.018
r2akateem	1.6209	0.7268	0.0257	5.057
kauptekn	-0.3237	0.0973	0.0009	0.723
hoito	0.6419	0.1753	0.0003	1.900
kotona	-0.6140	0.2163	0.0045	0.541
muutoim	-0.5435	0.1939	0.0051	0.581
työllkou	-0.4033	0.3059	0.1873	0.668
aluekys	-0.7770	0.2239	0.0005	0.460
tammi	0.4000	0.1153	0.0005	1.492
N	**työllistymiset**	**sensuroinnit**	**% sensuroituja**	
1764	671	1093	61.96	

Taulukko 9.2. Testejä nollahypoteesille $\beta_1 = \ldots = \beta_{14} = 0$

testi	$-2\log L(0)$	$-2\log L(\hat{\beta})$	**testisuureen arvo**	**p-arvo**
$-2\log(LR)$	7874.260	7725.112	149.148	0.0001
Score			154.805	0.0001
Wald			144.774	0.0001

Proportionaalisuusoletuksen paikkansapitävyyttä tutkittiin eräiden muuttujien (työnhakua edeltävää toimintaa, koulutusastetta, työttömäksituloajankohtaa, sukupuolta ja työttömyyskassan jäsenyyttä kuvaavien muuttujien) osalta vielä mallintamisvaiheessa osittamalla aineistoa muuttujien arvojen mukaan ja estimoimalla malleja, joissa ositteiden perushasardifunktioiden sallittiin poiketa toisistaan. Proportionaalisuusoletuksen sopivuutta voidaan tutkia vertailemalla Breslow:n menetelmällä estimoituja kumulatiivisia perushasardifunktioita. Näin voidaan tarkastella muuttujan vaikutustapaa hasardifunktioon, kun muiden muuttujien vaikutus on eliminoitu (ks. kpl 8.1). Tarkastelut tuottivat samankaltaisia tuloksia kuin ennen varsinaista mallintamista estimoidut Kaplan–Meier-käyrät: proportionaalisuusoletus vaikuttaa realistiselta suurimmalla osalla tarkasteltavaa aikaväliä. Työllisyyskoulutuksesta tulleiden käyrä leikkaa yhä muiden työnhakua edeltäviä toimintojen käyriä (Liite 4, kuva 4.1). Työllisyyskoulutuksesta tulleet ovat kuitenkin vain pieni osa tutkittavaa joukkoa (työllisyyskoulutuksesta tulleita on 39 kpl, joista 11 työllistyy seuranta-aikana), eikä ryhmälle ole tämän vuoksi järkevää estimoida omaa perushasardifunktiota.

Kuvassa 9.2 on mallin 2 perusteella (Breslow'n menetelmällä) estimoitu, tasoitettu perushasardifunktio. Perushasardifunktio on laskeva. Tämä merkitsee sitä, että työllistymistodennäköisyys pienenee työttömyyden keston kasvaessa; mitä kauemmin työttömyys kestää, sitä vähäisemmät ovat työllistymismahdollisuudet.

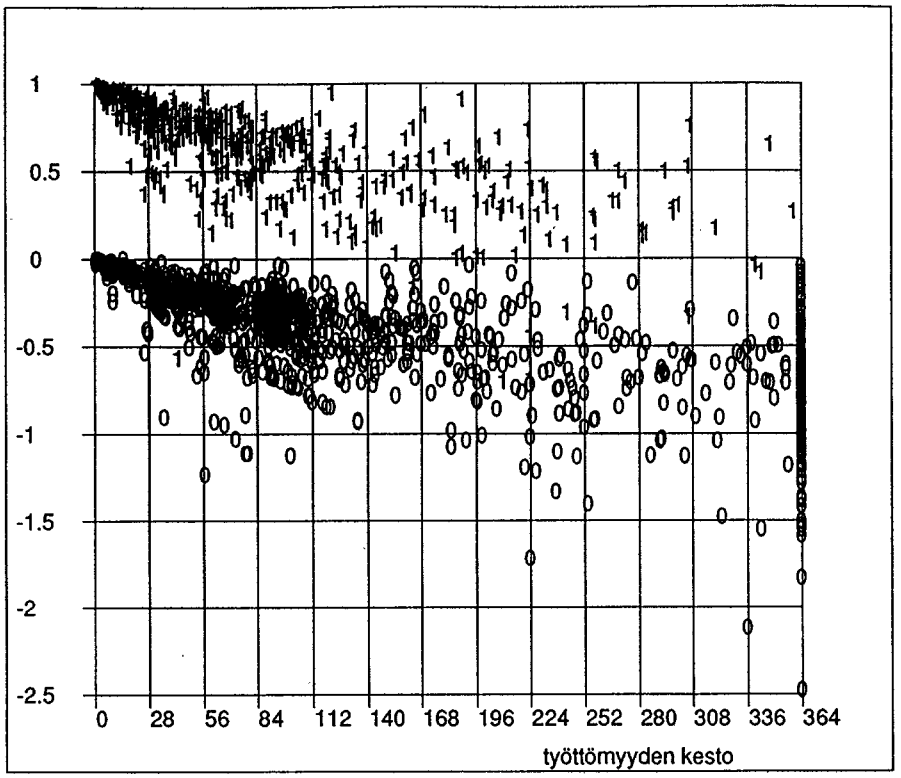


Kuva 9.2: Mallin 2. perusteella estimoitu, tasoitettu perushasardifunktio

Kuvassa 9.3 on mallin 2. perusteella lasketut residuaalit työttömyyden kestoa vastaan piirrettyinä. Residuaalit on merkitty ykkösellä tai nollalla sen mukaan, onko kyseessä työllistyminen vai sensurointi. Vaikuttaa siltä, että estimoitu malli ei pysty ottamaan huomioon sitä, että suuri osa työllistymisistä tapahtuu jo muutaman viikon kuluessa työttömyyden alkamisesta: estimoitu hasardifunktio ei kumuloidu alussa riittävän nopeasti ja on hyvin lähellä nollaa lyhyillä työttömyyden kestoilla. Tämän vuoksi residuaalit jakautuvat sen mukaan, onko kyseessä työllistyminen vai sensurointi: seuranta-aikana työllistyneiden residuaalit ($\hat{e}(t) = 1 - \hat{H}(t)$) sijoittuvat ykkösen ja sensuroitujen residuaalit ($\hat{e}(c) = 0 - \hat{H}(c)$) nollan ympärille. 162 yksilöä, joiden työttömyysjakso on katkaistu 365 päivän kohdalla, näkyvät kuvan oikeassa laidassa "pylväänä". Työllistymisiä tapahtuu varsin runsaasti jo alle viikossa: kuuden ensimmäisen työttömyyspäivän aikana työllistyy 64 yksilöä, mikä on 9,5 % seuranta-aikana työllistyneistä. Työvoimaviranomaisten mukaan nopeimmin työllistyy tyypillisesti vähän koulutettuja työnhakijoita, jotka eivät ole valikoivia työpaikan suhteen ja ottavat vastaan myös lyhytaikaisia töitä. Toisaalta osa työpaikan vaihtajista ja vastavalmistuneista ilmoittautuu työttömäksi varmuuden vuoksi, vaikka työpaikka olisi jo tiedossa. "Mitä tahansa" työtä hyväksyvien työllistymistä voisi mahdollisesti selittää haetun työpaikan tyyppiä kuvaavilla tiedoilla. Työnhakijarekisteri ei kuitenkaan valitettavasti sisällä tällaisia tietoja.

Estimoitu malli ei siis pysty selittämään nopeimmin tapahtumia työllistymisiä. Kuinka nämä havainnot sitten vaikuttavat estimointituloksiin? Alle viikon pituisen työttömyysjakson omaavien yksilöiden vaikutusta mallin regressiokertoimiin tutkittiin toistamalla estimointikokeiluja aineistolle, josta em. yksilöt oli poistettu (alle viikon pituisia työttömyysjaksoja oli 103 kpl, joista 64 päättyi työllistymiseen avoimille työmarkkinoille). Estimointitulokset eivät juurikaan muuttuneet: työllistymistodennäköisyyttä selittäviksi tekijöiksi valikoituivat samat tekijät kuin aiemmissa estimoinneissa. Estimoitu malli on liitteessä 4 (malli 3). Verrattaessa mallia 3 malliin 2 havaitaan, että eniten ovat muuttuneet koulutustasoa kuvaavien muuttujien **mkorkea** ja **r2akateem** sekä muuttujien **työllkou** ja **ntyöraj** regressiokertoimet. Syynä regressiokertoimien suurehkolle muuttumiselle lienee se, että alle viikossa tapahtuu suuri osa työrajoitteisten naisten sekä työllisyyskoulutuksesta tulleiden työllistymisistä ja toisaalta vain pieni osa korkea-asteen koulutuksen omaavien työllistymisistä. Keskimäärin kuitenkin työrajoitteiset ja työllisyyskoulutuksesta tulleet työllistyvät hitaasti ja korkeasti koulutetut nopeasti. Muuttujat vaikuttavat siis kuuden ensimmäiden työttömyyspäivän aikana päinvastaiseen suuntaan kuin seuranta-aikana keskimäärin.

Nopeimpien työllistymisien taustalla on siis tekijöitä, joita estimoitu malli (malli 2) ei pysty ottamaan huomioon. Toisaalta ainakaan alle viikon pituisen työttömyysjakson omaavat yksilöt eivät vaikuta paljoakaan estimointituloksiin. Vaikka mallin ennustekyky paranee hieman pidemmillä työttömyyden kestoilla, aliennustaa malli työllistymisiä myös pitkillä kestoilla. Vaikuttaa siltä, että mallista puuttuu tärkei-



Kuva 9.3: Residuaalit vs. työttömyyden kesto

tä työllistymistä selittäviä tekijöitä – työllistyminen on varmasti monimutkaisempi prosessi kuin mitä on mahdollista kuvata muutamalla rekisteritiedolla. Ongelmana on siis havaitsematon heterogeenisuus: käytettävissä olevat selittävät tekijät eivät pysty kontrolloimaan kaikkea populaation heterogeenisuutta. Kontrolloimaton heterogeenisuus vaikuttaa regressiokertoimien estimaatteihin siten, että ne ovat harhaisia kohti nollaa. Estimoitu hasardifunktio vähenee nopeammin (ks. kpl 5.4). On kuitenkin vaikeaa arvioida, kuinka paljon kontrolloimaton heterogeenisuus vaikuttaa regressiokertoimien ja hasardifunktion estimaatteihin. Kehnosta ennustekyvystään huolimatta malli antaa tietoa monien sekä yksilöön liittyvien että ulkoisten tekijöiden vaikutuksesta nuorten työllistymiseen.

Luku 10

Lopuksi

Tutkimuksessa perehdyttiin elinaika-analyysin teoriaan ja sovellettiin sitä nuorten työllistymiseen vaikuttavien tekijöiden tutkimiseen. Luvussa 2 käsiteltiin elinaika-analyysin ominaispiirteitä ja peruskäsitteitä. Elinaika-analyysin menetelmin tarkasteltaville aineistoille tyypillisiä ovat sensuroidut havainnot. Sensurointi tekee aineistosta epätäydellisen: sensuroitujen havaintojen osalta tapahtuma-aikoja ei havaita. Sensuroidut havainnot sisältävät kuitenkin analyysin kannalta tärkeää informaatiota: tiedetään, että näiden havaintojen tapahtuma-aika on pidempi kuin sensurointiaika.

Luvussa 3 esiteltiin kaksi aineiston alustavaan tarkasteluun soveltuvaa eloonjäämisfunktion ei-parametrista estimaattoria; Kaplan–Meier-estimaattori ja eloonjäämistaulu. Luvuissa 4 ja 5 käsiteltiin täysin parametroituja elinaikamalleja. Työttömyyden kestoon vaikuttavia tekijöitä on perinteisesti tutkittu täysin parametroidulla mallilla, jossa työttömyyden kesto on oletettu Weibull-jakautuneeksi. Weibull-jakaumaan perustuva malli kuuluu sekä kiihdytetyn elinajan että suhteellisten hasardien malliluokkaan. Kiihdytetyn elinajan malleissa selittävät tekijät vaikuttavat multiplikatiivisesti työttömyyden kestoon, suhteellisten hasardien malleissa puolestaan multiplikatiivinen vaikutus kohdistuu työttömyyden keston hasardifunktioon; työllistymisintensiteettiin. Käsitteet työttömyyden kesto ja työllistymisintensiteetti ovat läheisesti yhteydessä toisiinsa: suuri työllistymisintensiteetti merkitsee todennäköistä nopeaa työllistymistä.

Luvussa 6 esiteltiin semiparametriset Coxin malli ja luokiteltujen tapahtuma-aikojen malli sekä joustava, täysin parametroitu paloittainen eksponenttimalli. Tarkimmin käsiteltiin Coxin mallia. Coxin mallissa regressiokertoimien estimointi perustuu informaatioon yksilöiden keskinäisestä työllistymisjärjestyksestä. Työttömyyden keston analyysimenetelmien painopiste on viime vuosina siirtynyt semiparametrisiin

malleihin. Tähän ovat vaikuttaneet täysin parametroitujen mallien spesifioinnissa ilmenneet ongelmat.

Luvussa 7 tarkasteltiin erilaisten työttömyyden päättymisyiden huomioimista analyysissa. Kilpailevien riskien mallin mukaisesti pidetään työllistymistä tutkittaessa muihin syihin päättyneitä työttömyysjaksoja sensuroituina. Työllistyminen voidaan jakaa edelleen avoimille työmarkkinoille työllistymiseen ja työllistämistoimenpitein työllistymiseen. Työllistämistoimenpitein työllistymistä säätelee lähinnä lainsäädäntö eikä sitä ole mielekästä tarkastella samanlaisena tapahtumana kuin avoimille työmarkkinoille työllistymistä.

Tutkimuksen empiirisessä osassa tarkasteltiin nuorten työllistymistä avoimille työmarkkinoille semiparametrisella Coxin mallilla. Estimointitulosten mukaan työllistymistodennäköisyyteen vaikuttavia tekijöitä ovat mm. sukupuoli, koulutus ja työnhakua edeltävä toiminta. Estimoidun mallin mukaan ehdollista työllistymistodennäköisyyttä heikentävät fyysiset ja psyykkiset työrajoitteet, kaupan tai tekniikan alan koulutus, kotityöstä työnhakijaksi tuleminen sekä heikko työvoiman kysyntä mitattuna työttömien työnhakijoiden ja avoimien työpaikkojen suhteella työvoimapiireittäin. Nuoret miehet ovat nuoria naisia ja vähän koulutetut korkeasti koulutettuja heikommassa asemassa työmarkkinoilla. Estimointitulosten mukaan työllisyyskoulutuksesta tulleiden työllistymistodennäköisyys ei eroa vertailuryhmän työllistymistodennäköisyydestä. Tulos on mielenkiintoinen, koska työllisyyskoulutukseen pyritään valikoimaan nimenomaan heikoimmin työllistyviä työnhakijoita. Tuloksen voisi siis tulkita siten, että työllisyyskoulutus on täyttänyt tehtävänsä ja pystynyt kohentamaan koulutettujen asemaa työmarkkinoilla. Estimointituloksia tulkittaessa on pidettävä mielessä, että tutkimuksen seuranta-aika, 1991-1992 oli alkavan taloudellisen taantuman aikaa. Koulutusalaa kuvaavat muuttujat heijastavat työvoiman kysyntätilannetta laman alkuaikoina. On mahdollista, että hieman yllättävä estimointitulos, jonka mukaan nuorten naisten työllistymistodennäköisyys on yli 1,5-kertainen miehiin verrattuna, kuvaa ainakin osittain samaa asiaa. Perinteisiä naisten työpaikkoja, julkisen sektorin työpaikkoja on todennäköisesti ollut seuranta-aikana runsaammin tarjolla. Estimoitu työttömyyden keston hasardifunktio on laskeva. Tämä merkitsee sitä, että työllistymistodennäköisyys pienenee työttömyyden keston kasvaessa. Estimoidun hasardifunktion perusteella vaikuttaa siltä, että työttömyyden keston tutkimiseen yleisesti käytetty, Weibull-jakaumaan perustuva malli olisi sopinut kohtuullisen hyvin myös tarkasteltuun aineistoon.

Luku 11

Lähteet

Aitkin, Anderson, Francis & Hinde(1989): Statistical Modelling in GLIM. Clarendon Press, Oxford.

Allison, P.D.(1984):Event History Analysis.Sage university paper 46. Sage publications.

Andersen P.–Borgan, O.–Gill, R.–Keiding,N.(1993): Statistical Models Based on Counting Processes. Springer-Verlag.

Breslow, N.(1972): Contribution to the discussion of paper by D.R. Cox. J. of Royal Statistical Society A, 135, 216-217.

Bergström, R.–Edin, P.-A.(1991): Time Aggregation and the Distributional Shape of Unemployment Duration. Uppsala university working paper 1991:3.

Cox, D. R.(1972): Regression Models and Life Tables. Journal of the Royal Statistical Society B, 34, 187-220.

Cox, D.R.–Oakes, D.(1984):Analysis of Survival Data. Chapman and Hall, London.

Eerola, M.(1990):Tapahtumahistorioiden tilastollisesta analyysistä. Suomen tilastoseuran vuosikirja 1990, 55-71.

Fleming,T.–Harrington,D.(1991): Counting Processes & Survival Data. Wiley.

Heckman, J. J.—Singer, B.(1984): A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duraton Data. Econometrica, vol. 52, 271-320.

Kalbfleisch, J.D.–Prentice, R.L.(1973): Marginal Likelihoods Based on Cox's Regression and Life Model. Biometrika 60, 267-278.

Kalbfleisch, J.D.–Prentice, R.L.(1980): The Statistical Analysis of Failure Time Data. Wiley, New York.

Kiefer, N. M.(1988): Economic Duration Data and hazard Functions. Journal of Economic Literature, vol. XXVI, 64-67.

Kettunen, J. (1991): Heterogeneity in Unemployment Duration Models. ETLA discussion papers no. 352.

Lancaster, T. (1979): Econometric Methods for the Duration of Unemployment. Econometrica vol. 47, no. 4, 919-956.

Lancaster, T.—Nickell, S.(1980): The Analysis of Reemployment Probabilities for the Unemployed. Journal of the Royal Statistical Society A 143, 141-165.

Lancaster, T.(1990): The Econometric Analysis of Transition Data. Cambridge university press, Cambridge.

Lawless, J.F.(1982): Statistical Models and Methods for Lifetime Data. Wiley, New York.

Lilja, R. (1992): Modelling Unemployment Duration in Finland. TTT tutkimusselosteita 112.

Meyer, B.D.(1990):Unemployment Insurance And Unemployment Spells. Econometrica, vol.58, No.4, 757-782.

Moffitt, R.(1985): Unemployment Insurance and the Distribution of Unemployment Spells. Journal of Econometric Literature, 28, 85-101.

Narendranathan, W.–Stewart, M.B.(1991): How Does the benefit Effect Vary as Unemployment Spells Lengthen? University of Warwick, Economics Department.

Nickell, S. (1979): Estimating the Probability of Leaving Unemployment. Econometrica, vol. 47, 1249-1266.

Nio, I.(1985): Mitä tilastot kertovat nuorisotyöttömyydestä? Työpoliittinen aikakauskirja, Työministeriö.

Peto, R.(1972): Contribution to discussion of paper by D.R. Cox. J. of Royal Statistical Society, B 34, 205-207.

Prentice, R.L.–Gloeckler, L.A.(1978):Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. Biometrics 34, 57-67.

Pudney, S.(1989): Modelling Individual Choice: the Econometrics of Corners, Kinks and Holes. Basil Blackwell.

Santamäki-Vuori, T.—Sauramo, P.(1992): Nuoret työmarkkinoilla. TTT tutkimuksia 43.

Struthers, C.–Kalbfleisch, J.(1986): Misspecified Proportional Hazard Models. Biometrika 73, 2, 363-369.

Sueyoshi, G. T.(1992): Semiparametric Proportional Hazards Estimation of Competing Risks Models With Time-varying Covariates. Journal of Econometrics, no. 51, 25-58. North-Holland.

Työministeriö (1992): Nuorisotyöllisyystyöryhmän muistio.

1 Liite 1. Muuttujien konstruointi

Alueellista kysyntää kuvaava muuttuja **aluekys** on laskettu työvoimapiireittäin[1] työttömien työnhakijoiden lukumäärän suhteena avoimien työpaikkojen lukumäärään. Muuttuja **aluekys** kuvaa siis sitä, kuinka monta työnhakijaa keskimäärin hakee kutakin työpaikkaa. Tiedot kuukauden avoimista työpaikoista ja työttömistä työnhakijoista perustuvat kuukauden viimeisen arkipäivän tilanteeseen. Jos yksilö on joutunut työttömäksi kuukauden 15. päivänä tai ennen sitä, merkitään muuttujan **aluekys** arvoksi edellisen kuukauden työttömien työnhakijoiden lukumäärän suhde avoimien työpaikkojen määrään. Kuukauden 15. päivän jälkeen työttömäksi joutuneelle sijoitetaan muuttujan **aluekys** arvoksi työttömäksitulokuukauden työttömien työnhakijoiden ja avoimien työpaikkojen lukumäärän suhde.

Koulutusastetta kuvaavat indikaattorimuuttujat (muuttujat **lukio-akateem**) on muodostettu Tilastokeskuksen koulutusluokituksen mukaisesta koulutuskoodista. Koodin ensimmäinen numero kuvaa koulutusastetta, muut numerot mm. koulutusalaa. Osalla havainnoista (47 kpl) oli kuitenkin TK:n luokituksesta poikkeava koulutuskoodi. Työministeriön mukaan kyseessä on peruskoulun suorittaneita, vailla ammatillista koulutusta olevia henkilöitä, jotka ovat suorittaneet jonkin lyhyen työllisyyskurssin. Nämä havainnot sijoitettiin Työministeriön käytännön mukaisesti luokkaan "peruskoulun suorittaneet" (vertailuryhmä).

Kahdeksalla havainnolla oli koulutuskoodi "koulutusaste tuntematon". Kuudelle havainnolle arvioitiin koulutusaste pohjakoulutusta ja ammattia kuvaavien tietojen perusteella, kaksi havaintoa poistettiin puutteellisten tietojen vuoksi. Yksi havainto poistettiin tuntemattoman koulutusalan vuoksi (koulutusalaa ei voitu päätellä ammattinumeron perusteella). Yhdeltä havainnolta puuttui koulutuskoodin arvo kokonaan ja yhden havainnon koodi oli virheellinen. Näille havainnoille arvioitiin koulutusaste ja -ala kuten edellä ammatti- ja pohjakoulutustietojen perusteella.

Neljällä havainnolla oli työhakutietoihin merkitty "hakee työtä asuinpaikkakunnan ulkopuolelta". Ainoastaan asuinpaikkakuntansa ulkopuolelta työtä hakevaa ei määritellä työttömäksi, joten havainnot poistettiin aineistosta.

Tiedot työkokemuksesta puuttuivat 49 havainnolta. Työkokemuksen arvioimiseksi estimoitiin muiden havaintojen perusteella regressiomalli, jossa työkokemusta selitettiin työnhakijan iällä (muuttuja **ikä**), sekä indikaattorimuuttujilla **jäsen** (kuvaa työttömyyskassaan kuulumista), **ammatti** (muodostettu koulutustietojen perusteella) ja **työ** (muodostettu työnhakua edeltävää toimintaa kuvaavien tietojen perusteella). Muuttujat **ammatti** ja **työ** kuvaavat ammattiin tähtäävän koulutuksen omaamista sekä työnhakua edeltävää aikaa [2].

[1] Työvoimapiirejä on 13 kpl ja piirien rajat noudattavat pääsääntöisesti läänirajoja. Työvoimapiirit ovat: Uudenmaan läänin, Turun, Satakunnan, Hämeen, Kymen läänin, Mikkelin läänin, Vaasan läänin, Keski-Suomen läänin, Kuopion läänin, Pohjois-Karjalan läänin, Kainuun, Oulun ja Lapin läänin työvoimapiirit. Ahvenanmaa kuuluu Turun työvoimapiiriin.

[2]

ammatti=1 työnhakijalla on jokin ammattiin tähtäävä koulutus

työ=1 työnhakija ollut työssä tai yrittäjänä ennen työnhakua

Regressiokertoimien estimaatit ovat taulukossa 1.1. Vaikka regressiokertoimien estimaatit ovat p-arvojen perusteella merkitseviä, on mallin selitysaste varsin pieni eikä malli ennusta kovinkaan hyvin työnhakijoiden työkokemusta. Koska työkokemusta on muutoin vaikea arvioida aineiston tietojen perusteella, on mallia käytetty puutteistaan huolimatta ennustamaan työkokemus havainnoille, joilta muuttujan arvo puuttuu. Mallin antamat ennusteet pyöristettiin lähimmäksi kokonaisluvuksi.

Taulukko 1.1. Työkokemusta selittävä regressiomalli

muuttuja	estimaatti	keskivirhe	p-arvo
vakio	-1.3438	0.0976	0.0001
ikä	0.0795	0.0046	0.0001
jäsen	0.1514	0.0358	0.0001
ammatti	0.0835	0.0329	0.011
työ	0.3182	0.0329	0.0001
$R^2 = 0.2789$			

Aineistosta poistettiin lisäksi havaintoja, joiden puuttuvia tai virheellisiä tietoja ei voitu arvioida muiden tietojen perusteella: kaksi virheellisen kuntakoodin ja kahdeksan virheellisen työaikatoiveen omaavaa havaintoa sekä seitsemän puuttuvan työttömyyden päättymissyyn (muusta syystä kuin seuranta-ajan päättymisen vuoksi) omaavaa havaintoa. Lisäksi poistettiin kaksi ulkomailla asuvaa sekä neljä lomautettua työnhakijaa [3].

[3]Kolmen kuukauden lomautuksen jälkeen lomautetut katsotaan työttömiksi ja työllisyyskoodi muutetaan kakkoseksi. Aineistoon eksyneet lomautetut työnhakijat ovat ilmeisesti tällaisia tapauksia.

2 Tutkimusaineiston muuttujat

2.1 Työttömyyden ajankohtaa ja kestoa kuvaavat muuttujat

talku työttömyyden alkamispäivä (vvyvkkpp)

tloppu työttömyyden päättymispäivä (vvvvkkpp)

tkesto työttömuyyden kesto (tloppu-talku+1)[4]

2.2 Henkilötiedot

nainen

ruotsink äidinkieli ruotsi

työraj psyykkinen tai fyysinen työrajoite

ulkom ulkomaalainen

ikä ikä vuosina työttömyyden alkaessa

ikäryhmä (vertailuryhmä: alle 20 vuotta)

ryhmä2 20-24 vuotta

ryhmä3 24-29 vuotta

2.3 Työnhakua kuvaavat muuttujat

työaikat työaikatoive: 1=myös osa-aika- ja vuorotyö

hakal työnhakualue: 1=myös asuinpaikkakunnan ulkopuolelta

työnhakua edeltävä toiminta (vertailuryhmä: työelämästä tulleet)

opisk opiskelija

asevelv ase- tai siviilipalvelus

kotona kotityö

ongelmat ongelmatausta: vankila, huoltolaitos, sairaana

muutoim muu toiminta tai ei tietoa

työllkou työllisyyskoulutus

[4]Päivämäärämuuttujat **talku** ja **tloppu** muunnettiin työttömyyden keston laskemiseksi ns. SAS date value -arvoiksi, jotka kuvaavat 1.1.1960 jälkeen kuluneiden päivien lukumäärää.

työttömyyden päättymissyy

1= sijoitettu työllistämistoimenpitein

2= välitetty työhön yleisille työmarkkinoille

3= saanut itse työpaikan

4= aloittanut työllisyyskoulutuksen

5= siirtynyt työvoiman ulkopuolelle

6= muu syy tai ei tietoa

2.4 Työhistoriaa ja koulutusta kuvaavat muuttujat

jäsen työttömyyskassan jäsen

työkok työkokemus

0= ei työkokemusta

1= jonkin verran työkokemusta

2= täysi ammattitaito

koulutusaste (vertailuryhmä: peruskoulu)

lukio

alkeski alempi keskiaste (10-11 vuoden opinnot)

ylkeski ylempi keskiaste (12 vuoden opinnot)

alkorkea alin korkea-aste (13-14 vuoden opinnot)

alkandi alempi kandidaattiaste (15 vuoden opinnot)

akateem akateeminen loppututkinto (väh. 16 vuoden opinnot)

koulutusala (vertailuryhmä: yleissivistävä ja humanistinen koulutus sekä muu koulutus)

opetus

kauppa

tekniikka

liikenne

terveydenhuolto

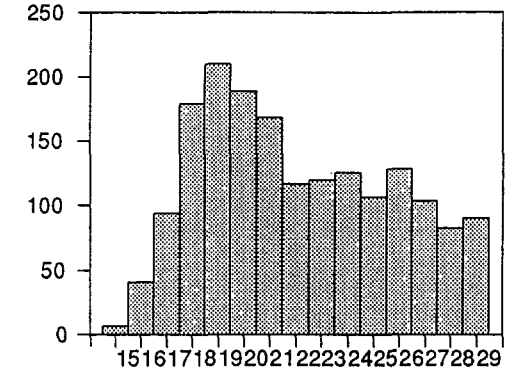
maatalous

2.5 Taustamuuttujat

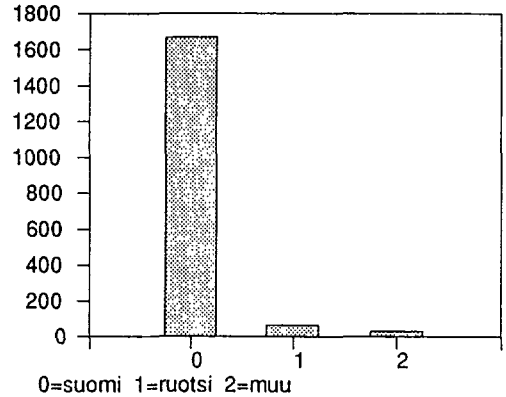
aluekys	työttömät työnhakijat/avoimet työpaikat työvoimapiireittäin
vuoalku	työttömyys alkanut tammi-kesäkuussa

Liite 2. Muuttujien suoria jakaumia

KUVA 2.1. IKÄ

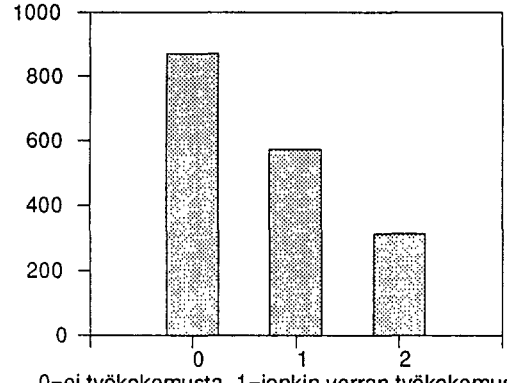


KUVA 2.2. ÄIDINKIELI



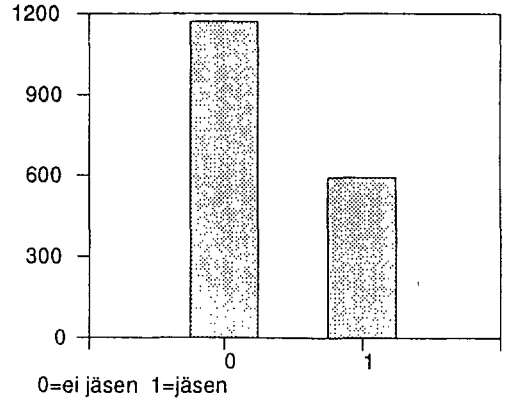
0=suomi 1=ruotsi 2=muu

KUVA 2.3. TYÖKOKEMUS



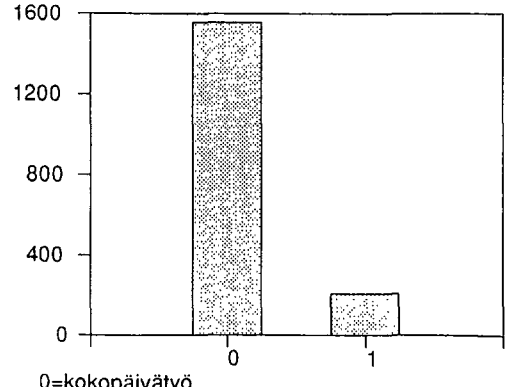
0=ei työkokemusta 1=jonkin verran työkokemusta
2=täysi ammattitaito

KUVA 2.4. TYÖTTÖMYYSKASSAN JÄSENYYS



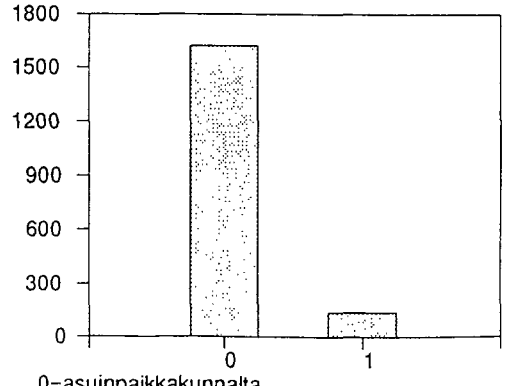
0=ei jäsen 1=jäsen

KUVA 2.5. TYÖAIKATOIVE



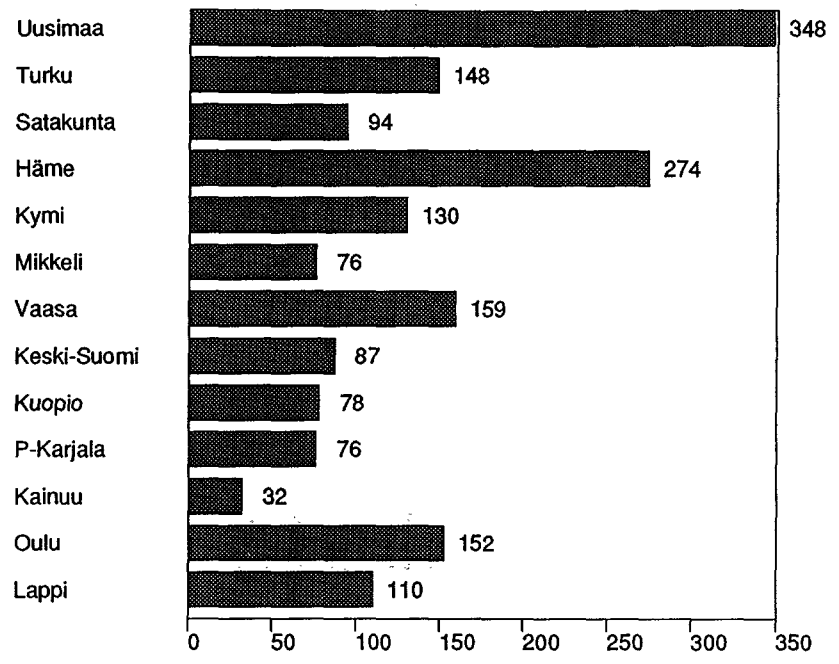
0=kokopäivätyö
1=myös vuorotyö t. osapäivä- ja osa-aikatyö

KUVA 2.6. TYÖNHAKUALUE

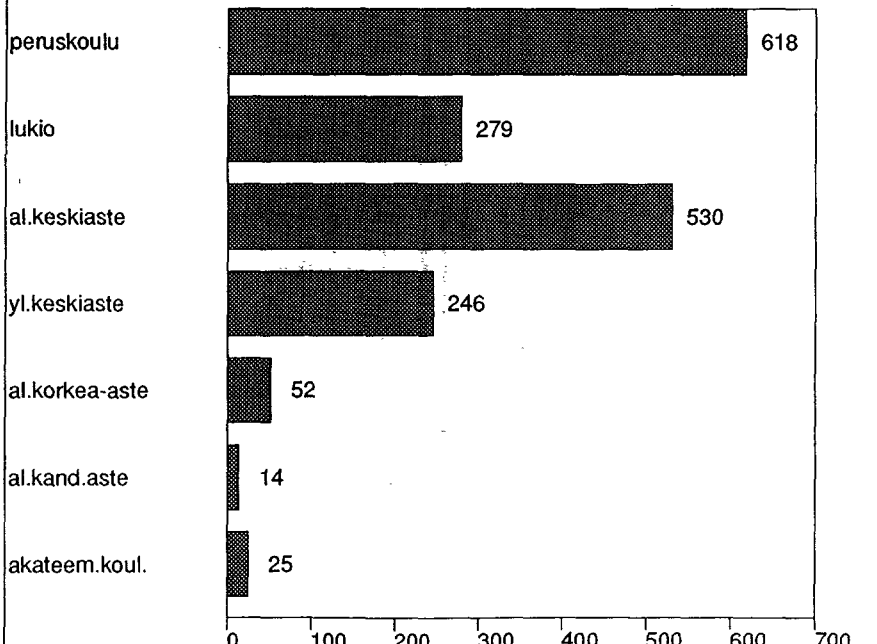


0=asuinpaikkakunnalta
1=myös asuinpaikkakunan ulkopuolelta

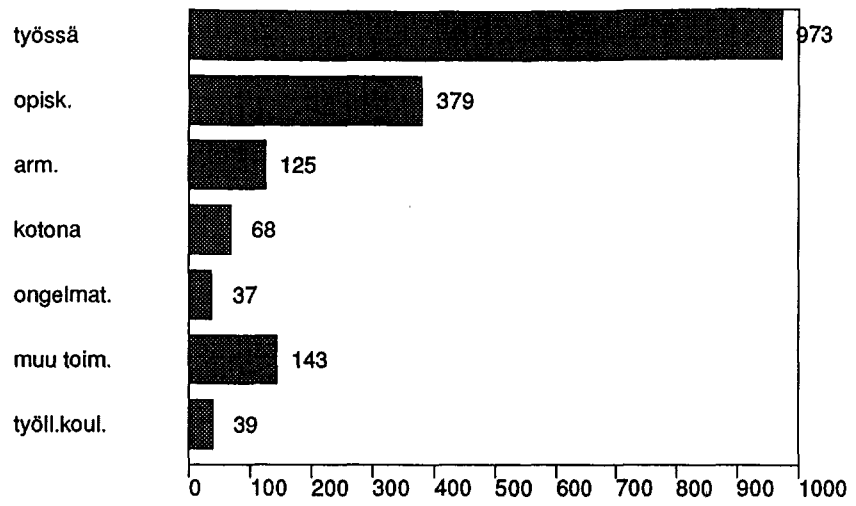
KUVA 2.7. TYÖVOIMAPIIRIT



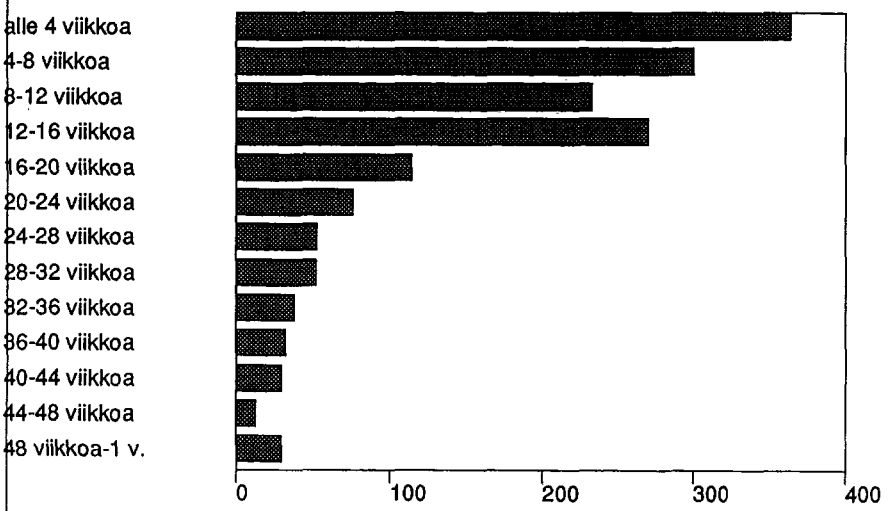
KUVA 2.8. KOULUTUSASTE



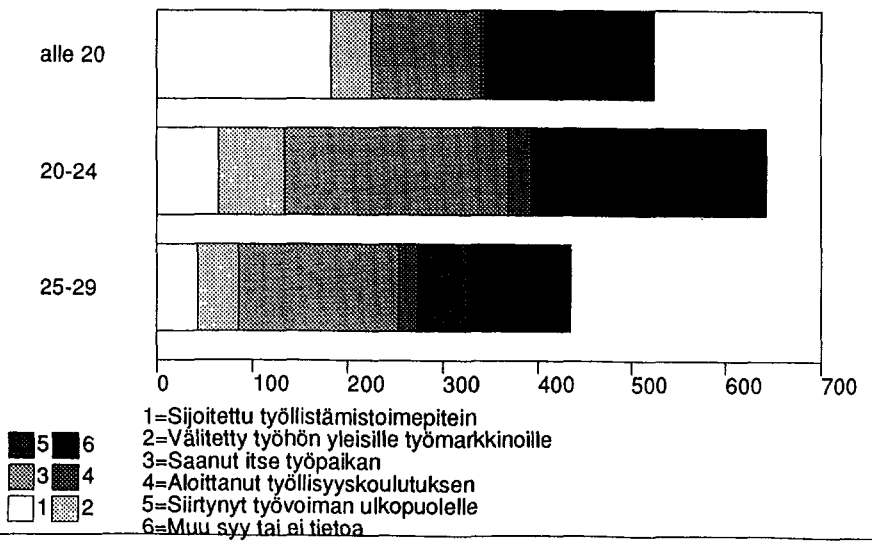
KUVA 2.9. TYÖNHAKUA EDELTÄVÄ AIKA



KUVA 2.10. TYÖTTÖMYYDEN KESTO *



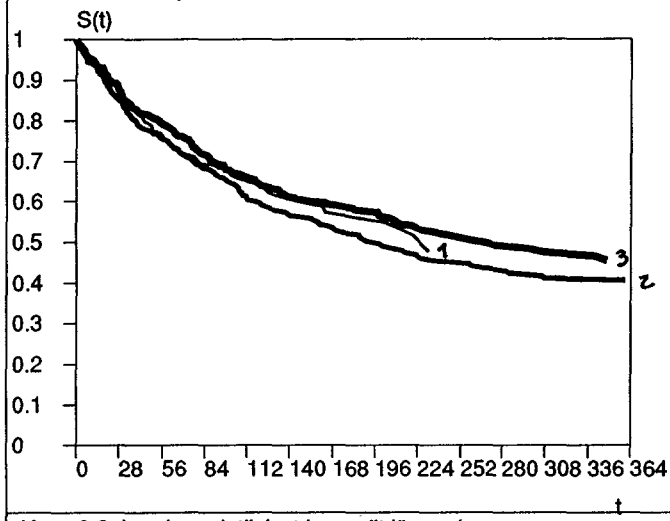
KUVA 2.11. TYÖTTÖMYYDEN PÄÄTTYMISSYYT IKÄRYHMITTÄIN *



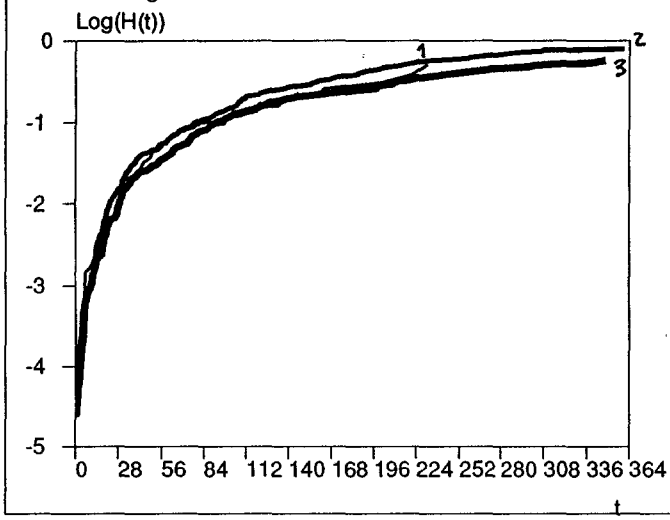
* yksilöt, joiden työttömyysjakso on pidempi kuin 365 päivää, on poistettu (N=1602)

Liite 3. Kaplan–Meier-estimaatit

Kuva 3.1. eloonjäämisfunktiot iän mukaan



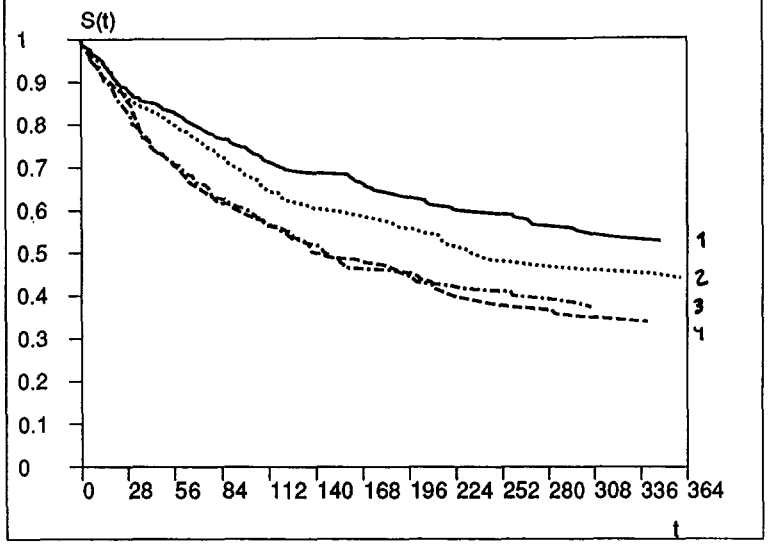
Kuva 3.2. Log-kumulatiiviset hasardit iän mukaan



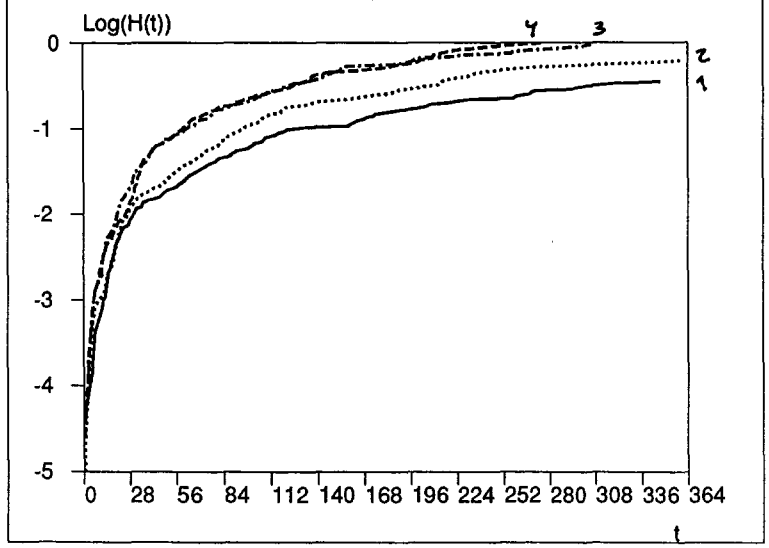
Taulukko 3.1. Työllistymiset ja sensuroinnit				
osite	N	työllistymiset	sensuroinnit	% sensuroituja
1: alle 20	531	158	373	70.24
2: 20-24	720	303	417	57.92
3: 25-29	513	210	303	59.06

Taulukko 3.2. Testejä nollahypoteesille $S_1(t) = S_2(t) = S_3(t)$		
Testi	testisuureen arvo	p-arvo
Log-Rank	3.2813	0.1939
Wilcoxon	2.6177	0.2701
-2 Log(LR)	11.1400	0.0038

Kuva 3.3. Eloonjäämisfunktiot sukupuolen ja jäsenyyden mukaan



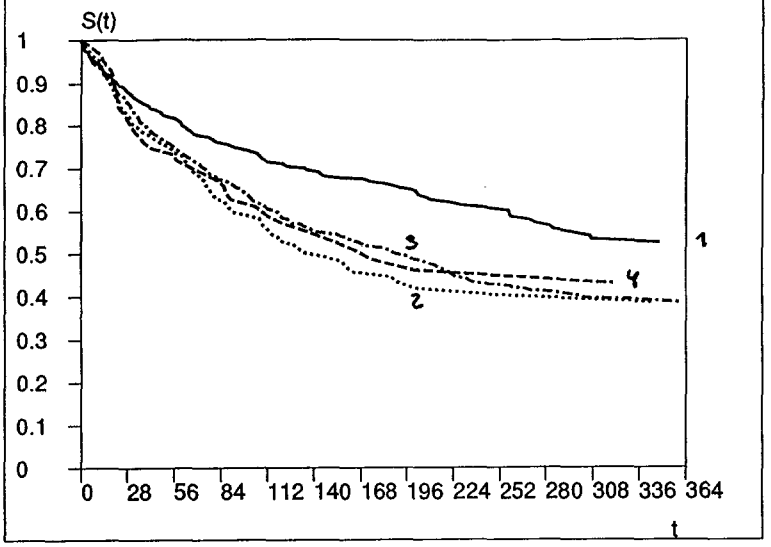
Kuva 3.4. Log kumul. hasardit sukupuolen ja jäsenyyden mukaan



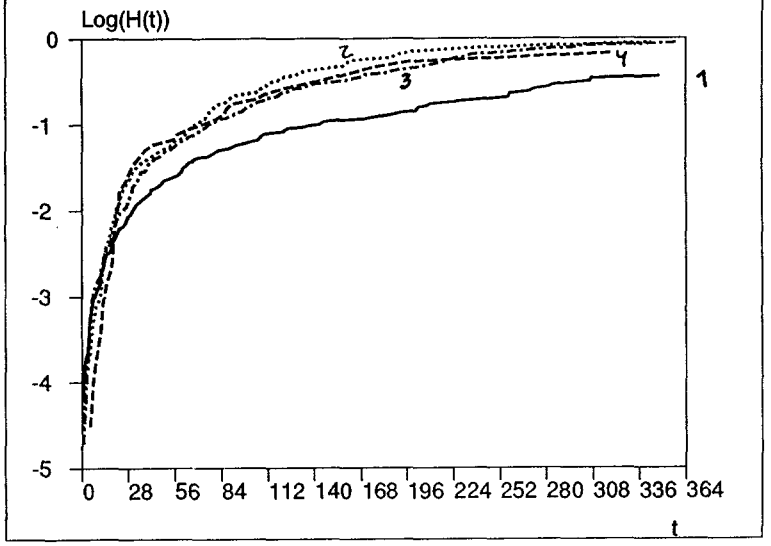
Taulukko 3.3. Työllistymiset ja sensuroinnit				
osite	N	työllistymiset	sensuroinnit	% sensuroituja
1: mies, ei jäsen	673	191	482	71.62
2: mies, jäsen	296	123	173	58.45
3: nainen, ei jäsen	498	209	289	58.03
4: nainen, jäsen	297	148	149	50.17

Taulukko 3.4. Testejä nollahypoteesille $S_1(t) = \ldots = S_4(t)$		
Testi	testisuureen arvo	p-arvo
Log-Rank	36.8818	0.0001
Wilcoxon	32.2951	0.0001
-2 Log(LR)	42.3954	0.0001

Kuva 3.5. Eloonjäämisfunktiot koulutusasteen mukaan



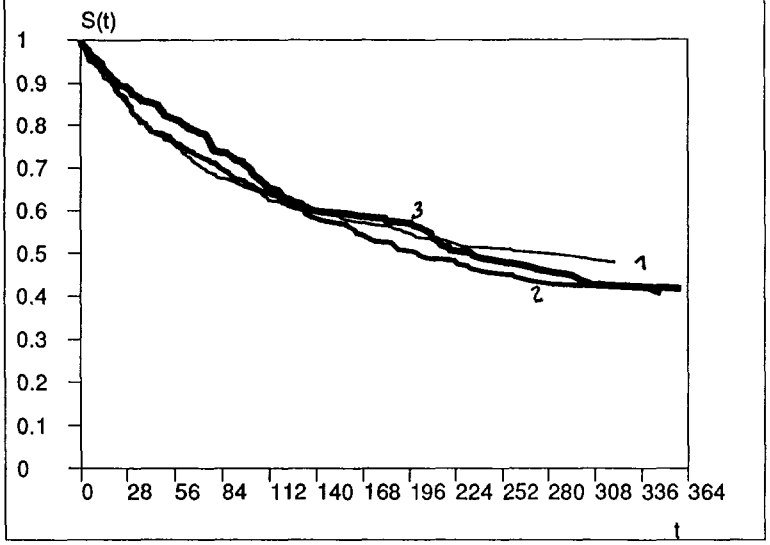
Kuva 3.6. Log kumulatiiviset hasardit koulutusasteen mukaan



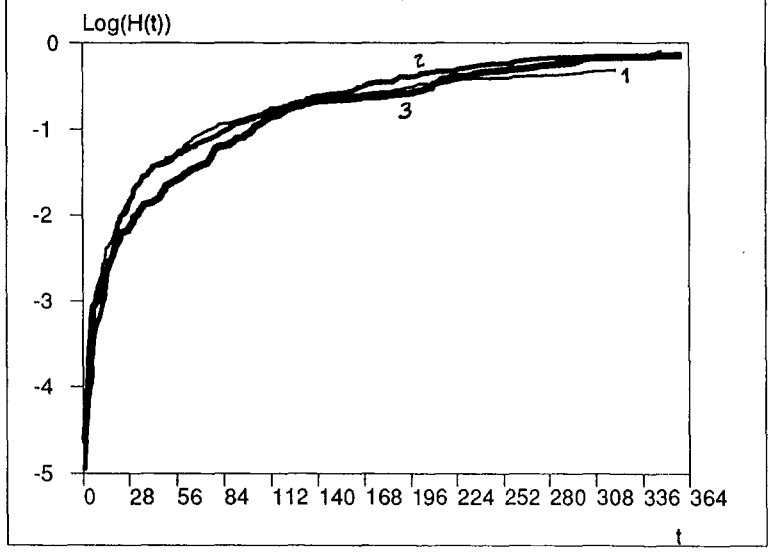
Taulukko 3.5. Työllistymiset ja sensuroinnit				
osite	**N**	**työllistymiset**	**sensuroinnit**	**% sensuroituja**
1: peruskoulu	618	176	442	71.52
2: lukio	279	113	166	59.50
3: keskiaste	776	343	433	55.80
4: korkea-aste	91	39	52	57.14

Taulukko 3.6. Testejä nollahypoteesille $S_1(t) = \ldots = S_4(t)$		
Testi	**testisuureen arvo**	**p-arvo**
Log-Rank	23.8566	0.0001
Wilcoxon	17.0628	0.0007
-2 Log(LR)	26.7388	0.0001

Kuva 3.7. Eloonjäämisfunktiot työkokemuksen mukaan



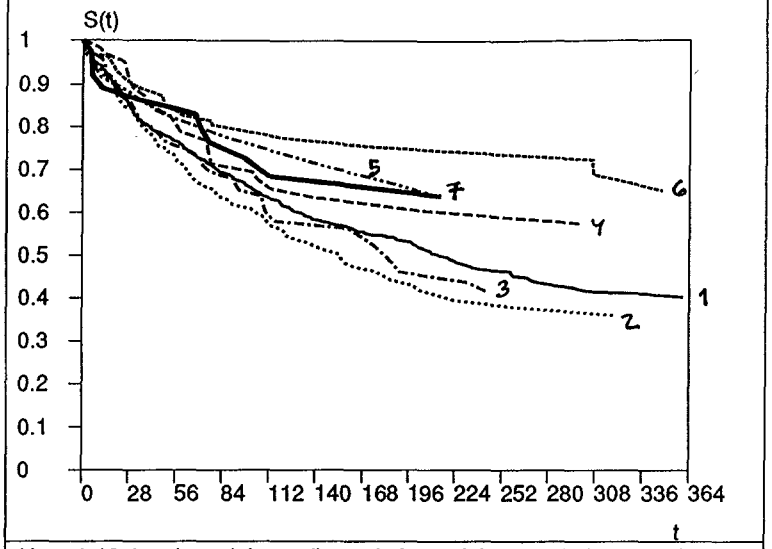
Kuva 3.8. Log-kumulatiiviset hasardit työkokemuksen mukaan



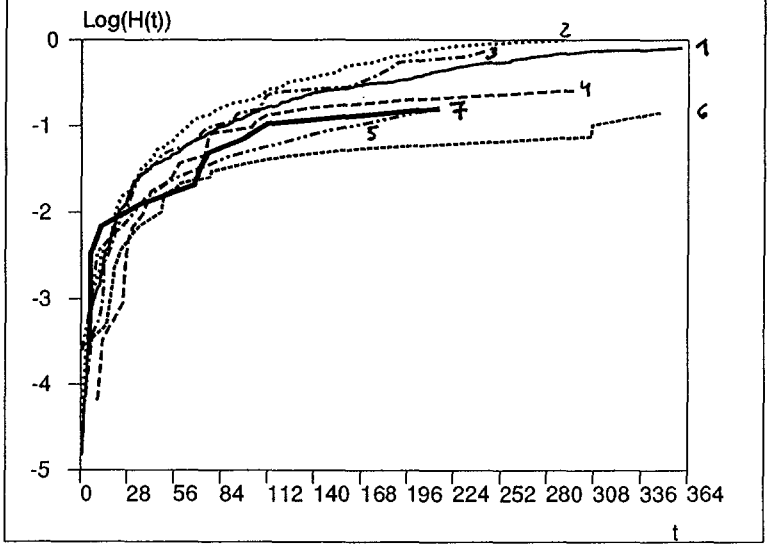
Taulukko 3.7. Työllistymiset ja sensuroinnit				
osite	**N**	**työllistymiset**	**sensuroinnit**	**% sensuroituja**
1: ei työkokemusta	873	301	572	65.52
2: jonkin verran työkokemusta	575	235	340	59.13
3: täysi ammattitaito	316	135	181	57.28

Taulukko 3.8. Testejä nollahypoteesille $S_1(t) = S_2(t) = S_3(t)$		
Testi	**testisuureen arvo**	**p-arvo**
Log-Rank	0.7382	0.6913
Wilcoxon	2.4904	0.2879
-2 Log(LR)	3.1136	0.2108

Kuva 3.9. Eloonjäämisfunktiot työnhakua edeltävän toiminnan mukaan



Kuva 3.10. Log kumul. hasardit työnhakua edeltävän toiminnan mukaan



Taulukko 3.9. Työllistymiset ja sensuroinnit				
osite	**N**	**työllistymiset**	**sensuroinnit**	**% sensuroituja**
1: työssä	973	397	576	59.20
2: opiskelemassa	379	152	227	59.89
3: armeijassa	125	50	75	60.00
4: kotona	68	23	45	66.18
5: ongelmatausta	37	9	28	75.68
6: muu toiminta	143	29	114	79.72
7: työllisyyskoulutuksessa	39	11	28	71.79

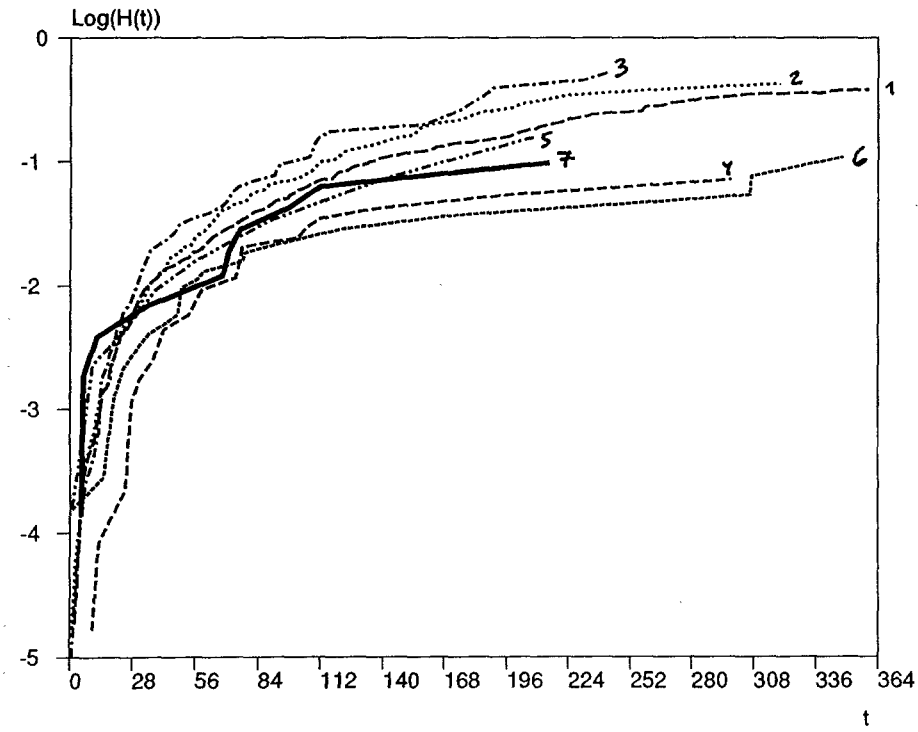
Taulukko 3.10. Testejä nollahypoteesille $S_1(t) = \ldots = S_7(t)$		
Testi	**testisuureen arvo**	**p-arvo**
Log-Rank	28.2507	0.0001
Wilcoxon	19.1830	0.0039
-2 Log(LR)	45.7643	0.0001

Liite 4. Estimointikokeiluja

Taulukko 4.1. Malli 1.: regressiokertoimien estimaatit				
muuttuja	**estimaatti**	**keskivirhe**	**p-arvo**	**riskisuhde**
ikä	-0.0026	0.0138	0.8506	0.997
nainen	0.3790	0.0952	0.0001	1.461
ruotsink	0.2691	0.1905	0.1577	1.309
ulkom	-0.4208	0.4174	0.3133	0.657
jäsen	0.1077	0.0878	0.2197	1.114
työraj	-0.8997	0.3066	0.0033	0.407
työkok	-0.0031	0.0606	0.9598	0.997
hakal	0.1519	0.1345	0.2589	1.164
työaikat	0.0920	0.1177	0.4342	1.096
lukio	0.4113	0.1250	0.0010	1.509
keski	0.5557	0.1373	0.0001	1.743
alkorkea	0.7754	0.2590	0.0028	2.171
alkandi	0.6909	0.4246	0.1036	1.996
akateem	0.2634	0.3677	0.4738	1.301
opetus	-0.1319	0.4419	0.7653	0.876
kauppa	-0.4122	0.1550	0.0078	0.662
tekniikk	-0.4009	0.1497	0.0074	0.670
maatal	-0.5300	0.3276	0.1057	0.589
hoito	0.5492	0.1975	0.0054	1.732
liikenne	0.1360	0.4030	0.7357	1.146
opisk	0.0501	0.1117	0.6538	1.051
asevelv	0.2017	0.1656	0.2234	1.223
kotona	-0.5893	0.2226	0.0081	0.555
muutoim	-0.4880	0.1989	0.0141	0.614
ongelmat	-0.2177	0.3420	0.5245	0.804
työllkou	-0.3844	0.3103	0.2155	0.681
aluekys	-0.8040	0.3197	0.0119	0.448
vuoalku	0.0368	0.1168	0.7527	1.037

Taulukko 4.2. Malli 1: Testejä nollahypoteesille $\beta_1 \ldots = \beta_{28} = 0$				
testi	$-2\log L(0)$	$-2\log L(\hat{\beta})$	**testisuureen arvo**	**p-arvo**
$-2\log(LR)$	7874.260	7732.132	142.129	0.0001
Score			147.849	0.0001
Wald			141.543	0.0001

Kuva 4.1. Log kumulatiiviset hasardit työnhakua edeltävän toiminnan mukaan



1: työssä
2: opiskelemassa
3: armeijassa
4: kotona
5: ongelmatausta
6: muu toiminta
7: työllisyyskoulutuksessa

Taulukko 4.3. Malli 3.: regressiokertoimien estimaatit				
muuttuja	**estimaatti**	**keskivirhe**	**p-arvo**	**riskisuhde**
nainen	0.4145	0.0985	0.0001	1.514
mjäsen	0.2154	0.1242	0.0829	1.240
mtyöraj	-1.6523	0.7111	0.0202	0.192
ntyöraj	-0.7724	0.3835	0.0440	0.462
lukkeski	0.5579	0.1061	0.0001	1.747
mkorkea	0.8497	0.2140	0.0001	2.339
r2akateem	1.9027	0.7313	0.0093	6.704
kauptekn	-0.3917	0.1019	0.0001	0.676
hoito	0.6806	0.1832	0.0002	1.975
kotona	-0.5434	0.2173	0.0124	0.581
muutoim	-0.5912	0.2085	0.0046	0.554
tyollkou	-0.6214	0.3577	0.0824	0.537
aluekys	-0.7460	0.2324	0.0013	0.474
tammi	0.4049	0.1228	0.0010	1.499

Taulukko 4.4. Testejä nollahypoteesille $\beta_1 = \ldots = \beta_{14} = 0$				
Testi	$-2\log(0)$	$-2\log(\hat{\beta})$	**testisuureen arvo**	**p-arvo**
$-2\log(LR)$	7131.507	6981.901	149.606	0.0001
Score			159.726	0.0001
Wald			147.131	0.0001

VATT-TUTKIMUKSIA -SARJASSA AIEMMIN ILMESTYNEET JULKAISUT

PUBLISHED VATT-RESEARCH REPORTS

1. Osmo Kuusi: Uusi biotekniikka, mahdollisuuksien ja uhkien teknologia. Helsinki: Tammi 1991.

2. Seija Parviainen: The Effects of European Integration on the Finnish Labour Market. Helsinki 1991.

3. Esko Mustonen: Julkiset palvelut: Tehokkuus ja tulonjako. Helsinki 1991.

4. Juha Rantala: Työpaikan avoinnaolon keston mittaaminen. Helsinki 1991.

5. Tuomo Mäki: Työvoiman riittävyys ja kohdentuminen 1990-luvulla. Helsinki 1991.

6. Martti Hetemäki: On Open Economy Tax Policy. Helsinki 1991.

7. Tanja Kirjavainen: Koulutuksen oppilaskohtaisten käyttömenojen eroista. Helsinki 1991.

8. Pentti Puoskari: Talouspolitiikan funktiot ja instituutiot. Helsinki 1992.

9. Pekka Parkkinen: Koulutusmenojen kehityspiirteitä vuoteen 2030. Helsinki 1992.

10. Seppo Laakso: Kotitalouksien sijoittuminen, asuinkiinteistöjen hinnat ja alueelliset julkiset investoinnit kaupunkialueella. Helsinki 1992.

11. Tanja Kirjavainen - Heikki A. Loikkanen: Ollin oppivuosi 13 000 - 56 000 markkaa. Helsinki 1992.

12. Teuvo Junka: Suurten teollisuusyritysten toimintasopeutus 1980-luvulla. Helsinki 1993.

13. Hannu Törmä - Thomas Rutherford: Integrating Finnish Agriculture into EC´s Common Agricultural Policy. Helsinki 1993.

14. Mika Kuismanen: Progressiivisen tuloverotuksen vaikutus miesten työn tarjontaan. Helsinki 1993.

15. Estonia and Finland - A Retrospective Socioeconomic Comparison. Helsinki 1993.

16. Tanja Kirjavainen - Heikki A. Loikkanen: Lukioiden tehokkuuseroista. DEA-menetelmän sovellus lukioiden tehokkuuserojen arvioimiseksi. Helsinki 1993.

17. Mikko Räsänen: Pankkien talletusvakuuden arvo ja riskikäyttäytyminen vuosina 1982 - 1992: optiohinnoittelumallin sovellus. Helsinki 1994.

18. Pasi Holm: Essays on International Trade and Tax Policy in Vertically Related Markets. Helsinki 1994.

19. Pekka Mäkelä: Markkinat ja ympäristö - Euroopan unionin ympäristöpolitiikan tarkastelua. Helsinki 1994.

20. Hannu Vartiainen: Rahoitusmarkkinat ja talouden tasapaino informaation taloustieteen näkökulmasta. Helsinki 1994.

21. Janne Känkänen: Elinkeinotuen vaikutukset hyvinvointiin: efektiiviset tukiasteet elinkeinotuen mittaamisessa. Helsinki 1994.

VALTION TALOUDELLINEN TUTKIMUSKESKUS

Hämeentie 3

PL 269

00531 HELSINKI

Seppo Leppänen

Ylijohtaja