Perusjoukon kokonaissumman estimointi Coxin regressiolla ja logistisella regressiolla: sovellus rakennusten aloitustilastoon

Juha Paju

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
23. tammikuuta 2012

Tiivistelmä

Juha Paju, *Perusjoukon kokonaissumman estimointi Coxin regressiolla ja logistisella regressiolla: sovellus rakennusten aloitustilastoon*
Tilastotieteen pro gradu -tutkielma, Jyväskylä yliopisto, 23. tammikuuta 2012.
43 sivua, 2 liitettä.

Tilastokeskuksen kuukausittain tuottama Rakennus- ja asuntotuotanto -tilasto on keskeinen rakennustoiminnan suhdanteita kuvaava tilastojulkaisu. Suuri osa uusimman tilastoitavan kuukauden tiedoista on tilaston tuotantohetkellä rekisteröitymättä, joten osa tilaston summatiedoista on estimoitava alipeittoisesta rekisteriaineistosta. Tässä tutkielmassa esitellään menetelmä, jonka avulla estimointia voidaan tehostaa rakennushankkeiden taustatiedoilla ja siten parantaa tilaston laatua. Sovelluskohteena on aloitettujen rakennusten kokonaistilavuuden kuukausitieto, mutta esimerkiksi rakennuslupien kuukausittaisen kokonaistilavuuden estimointiin voitaisiin soveltaa samaa menetelmää.

Tilaston tuotantopäivään mennessä rekisteriin tullutta aloitusten kuukausiaineistoa käsitellään otoksena kaikkien tutkittavana kuukautena aloitettujen rakennusten muodostamasta perusjoukosta. Otoksen yksilöille estimoidaan sisältymistodennäköisyydet vanhoista kuukausitiedoista, jotka tunnetaan tuotantohetkellä jo lopullisena. Havaituista tilavuuksista ja estimoiduista sisältymistodennäköisyyksistä lasketaan Horvitz-Thompson-estimaatti aloitusten kokonaistilavuudelle. Sisältymistodennäköisyydet estimoidaan Coxin regressiomallilla ja logistisella regressiomallia. Tuloksia verrataan nykyistä korotusmallia jäljittelevän menetelmän antamiin ennusteisiin.

Malleja testataan jäljittelemällä ensimmäisen ennakollisen aloitusten kokonaistilavuustiedon estimointitilanteita kuukausina 01/2008–06/2010. Logistiseen regressiomalliin perustuva Horvitz-Thompson-estimaattori tuottaa ennusteiden ja toteutuneiden arvojen vertailun perusteella parhaat tulokset. Ennuste on keskimäärin 32 000 kuutiometriä ja 1.2 prosenttia lähempänä toteutuvaa arvoa, kuin vastaava nykyisellä korotusmallilla tuotettu ennuste. Coxin regressiomallilla ei tulosten perusteella sen sijaan saavuteta nykyistä mallia parempia ennusteita.

Avainsanoja: Rakentaminen, Alipeitto, Coxin regressiomalli, Logistinen regressiomalli, Horvitz-Thompson-estimaattori

Sisältö

Luku 1

Johdanto

Rakennus- ja asuntotuotanto -tilasto on Tilastokeskuksen kuukausittain julkaisema katsaus kotimaan rakennusluvanvaraisen rakennustoiminnan suhdanteista. Tilaston tarkoituksena on tarjota informaatiota rakennustoiminnan määrästä, rakennustuotannon volyymista ja niiden kehityksestä Suomessa. Julkaisun rungon muodostavat kuukausittain julkaistavat tiedot myönnettyjen rakennuslupien kokonaistilavuudesta sekä uudisrakentamisen volyymi-indeksin arvot. Neljännesvuosittain julkaisuun tuotetaan myös tiedot aloitetuista ja valmistuneista rakennuksista. Uudisrakentamisen volyymi-indeksi perustuu aloitettuihin rakennuksiin, joten tarkat aloitustiedot ovat tärkeitä myös kuukausitasolla.

Julkaistavat tiedot perustuvat kuntien rakennusvalvontaviranomaisten Väestörekisterikeskukselle toimittamiin tietoihin rakennusluvanvaraisista rakennushankkeista ja rakennusvaiheista. Etenkin aloitettujen rakennusten tiedot kertyvät hitaasti ja uusimman kuukauden tilasto perustuu tuotantohetkellä pahasti alipeittoiseen aineistoon. Tällä hetkellä alipeittoisesta aineistosta laskettuja rakennusten aloitusten kokonaistilavuutta korotetaan yksinkertaisella menetelmällä, joka huomioi rakennuksen käyttötarkoitustyypin ja rakennushankkeen aloituskuukauden (menetelmä kuvattu luvussa 6). Rakennushankkeisiin liittyy kuitenkin myös muita taustatietoja, joiden avulla estimaatin tarkkuutta voidaan mahdollisesti parantaa.

Tässä tutkielmassa esitellään vaihtoehtoinen menetelmä aloitusten kokonaismäärän estimointiin. Tilaston tuotantohetkellä rekisteriin tulleiden aloitustietojen ajatellaan olevan otos kaikkien tilastoitavan kuukauden aloitusten muodostamasta perusjoukosta. Otokseen sisältymisen todennäköisyys on estimoitavissa aikaisempien kuukausien jo tiedossa olevien rekisteröintitapahtumien perusteella. Sisältymistodennäköisyydet vaihtelevat taustatietojen mukaan, mikä johtaa Horvitz-Thompson-estimaattorin käyttöön.

Rekisteröintiviivettä voidaan käsitellä elinaikamuuttujana, joten sisältymis-

todennäköisyydet voidaan estimoida elinaikamallien avulla. Toisaalta rekisteröitymistä kiinnitetyn ajan sisällä voidaan kuvata myös dikotomisella muuttujalla, mikä mahdollistaa logistisen regressiomallin käytön.

Tutkimusongelma on kaksiosainen. Ensiksi halutaan selvittää, onko valittujen taustamuuttujien ja rakennuksen aloituksen rekisteröitymisen välillä havaittavissa riippuvuutta. Tilastollisina menetelminä käytetään Coxin regressiomallia sekä logistista regressiomallia. Toinen, tärkeämpi päämäärä on ennustaa todellista kuukausittaista aloitutusten kokonaistilvuutta molemmilla malleilla käyttämällä selittäjinä merkitseviksi todettuja muuttujia.

Tutkielman luvussa 2 käydään läpi elinaikamallien perusteita ja esitellään Coxin regressiomalli. Luvussa 3 käydään läpi Coxin malliin liittyvää uskottavuuspäättelyä ja esitetään menetelmiä mallin parametrien estimoimiseksi. Luvussa 4 esitellään logistinen regressiomalli ja käydään lyhyesti läpi sen estimointia. Luvussa 5 esitellään kokonaismäärään estimaattorina käytettävä Horvitz-Thompson-estimaattori. Luku 6 on tutkielman soveltava osa, jossa testataan käytännössä esiteltyjen menetelmien soveltuminen rakennusten aloitusten estimointiin. Tuloksien yhteenveto ja johtopäätökset esitellään luvussa 7. Aineiston käsittelyyn, analysointiin ja estimointiin on käytetty sekä R- että SAS-ohjelmistoja (R Development Core Team 2011, SAS Institute Inc. 2009). Liitteessä B on R-kielinen ajojono, jolla ennustamisvaiheen tulokset on tuotettu.

1.1 Merkinnöistä

Tutkielmassa toistuvasti esiintyvät merkinnät määritellään seuraavasti:

$\mathbf{a}$ – lihavoidulla pienellä kirjaimella tarkoitetaan sarakevektoria

$\mathbf{a}^T$ – vektorin $\mathbf{a}$ transpoosi

$\hat{Y}$ – Y:n estimaatti tai estimaattori

$[a, b]$ – suljettu väli alarajana a ja ylärajana b

$[a, b[$ – puoliavoin väli alarajana a ja ylärajana b

Muut merkinnät ja termit määritellään asiayhteyksissään.

Luku 2

Coxin regressiomalli

Elinaika-analyysi tutkii tapahtuman toteutumiseen kuluvaa aikaa. Tapahtumasta käytetään usein nimeä ”kuolema” ja se voikin tarkoittaa kirjaimellisesti esimerkiksi sydänpotilaan kuolemaa. Yleisesti elinaikamallinnuksessa ”kuolema” on geneerinen ilmaisu ja sillä voidaan viitata mihin tahansa muutokseen yksilön tilassa. Tässä tutkielmassa kiinnostava tilanmuutos on aloitetun rakennushankkeen rekisteröityminen.

Elinaikamallit voidaan luokitella kolmeen pääryhmään sen mukaan, millaisia oletuksia mallinnettavasta elinaikamuuttujasta tehdään. Parametrittomat mallit eivät oleta mitään elinaikamuuttujan jakaumasta (paitsi muuttujan positiivisuuden). Parametrisissa malleissa elinajan oletetaan noudattavan jotain tunnettua jakaumaa, esimerkiksi eksponentti- tai Weibullin jakaumaa. Semiparametriset elinaikamallit ovat hybridejä kahdesta edellisestä, eli niihin liittyy sekä parametrinen että parametriton osa.

Tässä luvussa esitellään semiparametrinen Coxin regressiomalli, joka tarjoaa erään mahdollisuuden tehostaa rekisteröintiviiveen jakauman estimointia hankkeen taustatiedoilla. Tätä ennen käydään läpi elinaikamallinnuksen keskeistä teoriaa. Pääasiallisena lähteenä on käytetty kirjaa (Collet 2003).

2.1 Elinaikamallien perusteoriaa

2.1.1 Keskeiset funktiot

Määritellään seuraavaksi elinaika-analyysin tärkeimmät funktiot. Elinaikamuuttujan T oletetaan olevan jatkuva ja positiivinen. Oletetaan, että sen tiheysfunktio on $f(t)$. Kertymäfunktio on tällöin

$$F(t) = P(T < t) = \int_0^t f(u)\,du.$$

Välttöfunktio (myös selviytymisfunktio) määritellään todennäköisyytenä sille, että elinaika saa suuremman tai yhtä suuren arvon kuin t, siis

$$S(t) = P(T \geq t) = 1 - F(t).$$

Vaarafunktio (myös vikataajuus- tai hasardifunktio) kuvaa kuoleman todennäköisyyttä hetkellä t ehdolla, että se ei ole tapahtunut tätä aiemmin. Formaalisti vaarafunktio määritellään raja-arvona

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t}. \tag{2.1}$$

Ehdollisen todennäköisyyden kaavalla[1] saadaan vaarafunktiolle

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t)/P(T \geq t)}{\delta t} \\ &= \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} \times \frac{1}{S(t)}. \end{aligned} \tag{2.2}$$

Oletetaan, että F on derivoituva pisteessä t. Tällöin tulon (2.2) vasen termi on T:n kertymäfunktion derivaatta pisteessä t, eli itse asiassa tiheysfunktio $f(t)$. Vaarafunktiolle pätee siis tällöin tulos

$$h(t) = \frac{f(t)}{S(t)}.$$

Edelleen vaarafunktio voidaan kirjoittaa muodossa

$$h(t) = -\frac{d}{dt} \log S(t).$$

Puolittain integroimalla ja eksponenttiin korottamalla saadaan tulos

$$S(t) = \exp\left(- \int_0^t h(u)\, du \right) = e^{-H(t)},$$

missä funktio

$$H(t) = \int_0^t h(u)\, du. \tag{2.3}$$

on kumulatiivinen vaarafunktio.

[1] $P(A|B) = \frac{P(A \cap B)}{P(B)}$

2.1.2 Sensurointi

Elinaika-analyysissa on hyvin tavallista, että ainakin osa havainnoista on epätäydellisiä. Mikäli yksilön tarkka elinaika ei ole tiedossa, sanotaan havainnon olevan sensuroitu. Sensurointi voi johtua koeasetelmasta tai – kuten rakennusten aloitustietojen tapauksessa – siitä, että tarkkaa elinaikatietoa ei ole saatavilla.

Tyypillisessä koetilanteessa valittujen yksilöiden selviytymistä seurataan ennalta määrättyyn lopettamisajanhetkeen saakka. Havainto sensuroituu, mikäli yksilö on seurannan lopussa vielä elossa. Tilanteesta käytetään nimeä tyypin I sensurointi (Klein & Moeschberger 1997, s. 56–59). Tyypin II sensurointi liittyy puolestaan asetelmaan, jossa koe lopetetaan, kun ennalta sovittu määrä kuolemia on tapahtunut (Klein & Moeschberger 1997, s. 59–61). Sensuroitumista voi molemmissa tapauksissa tapahtua myös koeasetelmasta riippumatta, mikäli kokeeseen valittujen yksilöiden seuranta päättyy ennen kokeen loppumista. Tutkittaessa esimerkiksi tietyn sairauden ilmenemistä, voi koehenkilö kuolla kesken seurannan johonkin muuhun tautiin, jolloin kyseisen henkilön seuranta päättyy.

Elinaikahavainto on oikealta sensuroitu, mikäli tiedossa on ainoastaan aikaisin mahdollinen kuoleman ajankohta. Todellinen elinaika voi siis olla mikä tahansa tiedossa olevaa alarajaa suurempi arvo. Epätavallisempi tapaus on, että elinajalle tunnetaan vain yläraja. Tällöin kyseessä on vasemmalta sensuroitu havainto. Mikäli tapahtuma-aikaa ei tiedetä tarkasti, mutta sen tiedetään olevan jollain aikavälillä, on kyseessä intervalli- eli välisensuroitu havainto. Tässä tutkielmassa tarkastellaan ainoastaan sensuroimattomia ja oikealta sensuroituja elinaikoja. Sensurointi-indikaattori on

$$\delta_i = \begin{cases} 1 & \text{yksilöön } i \text{ liittyy todellinen elinaika} \\ 0 & \text{yksilöön } i \text{ liittyy sensuroitu elinaika.} \end{cases}$$

2.1.3 Välttöfunktion parametriton estimointi

Oletetaan, että on havaittu elinajat $t_1, t_2, \ldots, t_n$. Mikäli havaintoaineisto ei sisällä lainkaan sensuroituja havaintoja, voidaan välttöfunktiota estimoida elossa olevien yksilöiden suhteellisella osuudella hetkellä t. Siis

$$\hat{S}(t) = \frac{\sum_{i=1}^{n} 1_{[t_i \geq t]}}{n},$$

joka on myös t:n empiirinen välttöfunktio.

Aineisto sisältää kuitenkin usein sensuroituja havaintoja. Kaplan-Meier-estimaattori on empiirisen välttöfunktion yleistys aineistolle, joka sisältää oikealta sensuroituja havaintoja. Sen määrittelyä varten jaetaan tarkasteltava

aikaväli sensuroimattomien elinaikahavaintojen mukaan osaväleihin. Käytetään näistä sensuroimattomista elinajoista merkintää $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. Lisäksi merkitään

n_j – niiden yksilöiden lukumäärä, joiden tiedetään olevan elossa ennen ajanhetkeä $t_{(j)}$

d_j – niiden yksilöiden lukumäärä, joiden tiedetään kuolleen välillä $[t_{(j)}, t_{(j+1)}[$.

Kaplan-Meier-välttöfunktioestimaattori määritellään

$$\hat{S}_{KM}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j},$$

kun $t_{(k)} \leq t < t_{(k+1)}$ (Collet 2003, s.19–20). Välillä h sensuroidut havainnot siis huomioidaan elossa olevien määrän n_{h+1} pienenemisellä, mutta kuolleiden määriin d_j ne eivät vaikuta. Mikäli sensuroituja havaintoja ei ole, palautuu $\hat{S}_{KM}(t)$ empiirisen välttöfunktion muotoon.

Käyttäen hyväksi Taylorin sarjan approksimaatiota satunnaismuuttujan varianssille, saadaan Greenwoodin kaavana tunnettu tulos keskivirheelle:

$$\mathbf{s.e.}\left[\hat{S}_{KM}(t)\right] \approx \hat{S}_{KM}(t)\left[\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\right]^{\frac{1}{2}}$$

kun $t_{(k)} \leq t < t_{(k+1)}$ (Collet 2003, s. 23–25). Sensuroimattomalle aineistolle empiirisen välttöfunktion keskivirheen approksimaatio on

$$\mathbf{s.e.}\left[\hat{S}(t)\right] \approx \left[\frac{\hat{S}(t)(1 - \hat{S}(t))}{n_1}\right]^{\frac{1}{2}}.$$

2.2 Coxin suhteellisen vaaran malli

Sir David Coxin vuonna 1972 esittelemä suhteellisen vaaran malli eli Coxin regressiomalli (Cox 1972) on käytetyimpiä elinaika-analyysin työkaluja. Malli koostuu parametrittomasti estimoitavasta perusvaarafunktiosta sekä taustamuuttujien vaikutukset sisältävästä parametrisesta osasta. Se kuuluu siis semiparametristen elinaikamallien perheeseen.

Oletetaan, että kuhunkin elinaikahavaintoon liittyy tieto yhden p-luokkaisen kategorisen taustamuuttujan, faktorin, arvosta. Määritellään dummy-muuttujat $x_{i1}, x_{i2}, \ldots, x_{ip}$ seuraavasti:

$$x_{ik} = \begin{cases} 1 & \text{yksilön } i \text{ faktorin taso on } k \\ 0 & \text{muuten} \end{cases}$$

$k = 1, \ldots, p$. Olkoon vielä $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Suhteellisen vaaran malli on muotoa

$$h(t|\ \mathbf{x}_i) = h_0(t)\psi(\mathbf{x}_i),$$

missä $h_0(t)$ on perustason vaarafunktio (perusvaara) ja ψ tunnettu funktio, joka liittää taustamuuttujan tuoman informaation malliin. Tavallinen valinta funktion $\psi(\mathbf{x}_i)$ on eksponenttifunktio muuttujien x_{ik} lineaarikombinaatiosta, siis

$$h(t|\ \mathbf{x_i}) = h_0(t)e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}} = h_0(t)e^{\boldsymbol{\beta}^T \mathbf{x}_i}. \tag{2.4}$$

Tässä muodossa malli on yliparametrisoitu, ja siksi asetetaan referenssitasoksi $\beta_1 = 0$ (kiinnitettävän parametrin valinta on mielivaltainen). Perusvaara $h_0(t)$ kuvaa siten vaarafunktion arvoa ajanhetkellä t, kun faktori on referenssitasollaan. Regressiokertoimien β_k tulkinta on seuraava: jos yksilöön i liittyy faktorin taso k, on vaarafunktion arvo havainnolle i e^{β_k}-kertainen verrattuna havaintoon, jolla faktorin taso on 1. Tämä nähdään helposti, sillä

$$\frac{h(t|\ x_{ik} = 1)}{h(t|\ x_{j1} = 1)} = \frac{h_0(t)e^{\beta_k x_{ik}}}{h_0(t)e^{\beta_1 x_{j1}}} = \frac{e^{\beta_k}}{e^0} = e^{\beta_k}.$$

Jos faktoreita on enemmän kuin yksi, voidaan kaikki tarvittavat kertoimet, mukaan lukien yhdysvaikutustermit, koota samaan vektoriin $\boldsymbol{\beta}$. Mallin identifioituvuus varmistetaan kiinnittämällä riittävä määrä kertoimia nollaksi. Esimerkiksi kahden kaksiluokkaisen faktorin yhdysvaikutusmalli on muotoa

$$h(t|\ \mathbf{x}_i) = h_0(t)e^{(\beta_2 x_{i2} + \beta_4 x_{i4} + \beta_{22} x_{i8})},$$

kun kertoimet $\beta_1 = 0$ ja β_2 liittyvät ensimmäisen faktorin ja $\beta_3 = 0$ ja β_4 toisen faktorin tasoihin ja kertoimet $\beta_{11} = 0$, $\beta_{12} = 0$, $\beta_{21} = 0$ ja β_{22} ovat interaktiotermejä. Päävaikutusten ensimmäiset tasot ja kolme ensimmäistä interaktiotermiä on siis kiinnitetty nollaksi ja malli on identifioituva.

Selittävä muuttuja voi olla myös jatkuva. Tällöin regressiokertoimen β tulkinta on, että yhden yksikön kasvu selittävän muuttujan arvossa muuttaa vaaran e^{β}-kertaiseksi.

2.3 Mallin diagnostiikkaa

Coxin regressiomallin diagnostiset tarkastelut liittyvät suhteellisen vaaran oletuksen tutkimiseen, residuaalitarkasteluihin sekä yksittäisten havaintojen vaikuttavuuden analysointiin. Suhteellisen vaaran oletusta voi tutkia jo ennen estimointia – tällöin vältytään turhalta työltä, mikäli mallin todetaan olevan lähtökohtaisesti sopimaton aineistoon. Muut tarkastelut tehdään mallin sovittamisen jälkeen.

2.3.1 Suhteellisen vaaran oletus

Integroimalla (2.4) puolittain ja käyttämällä yhtälöön kaavaa (2.3), saadaan

$$\int_0^t h(u|\ \mathbf{x}_i)du = \int_0^t h_0(u)e^{\boldsymbol{\beta}^T\mathbf{x}_i}du$$

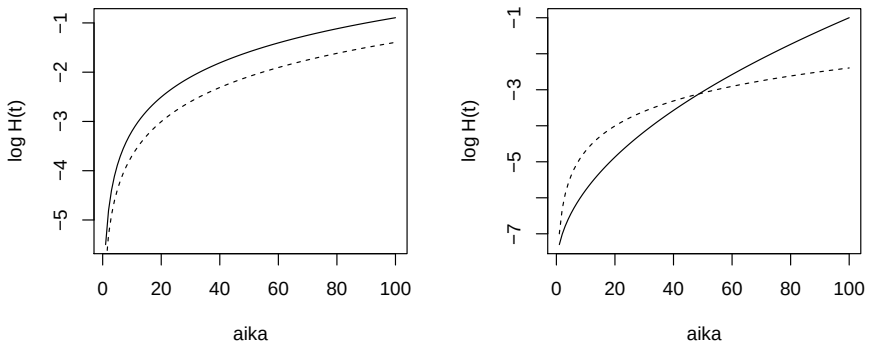
$$\Longleftrightarrow$$

$$H(t|\ \mathbf{x_i}) = \exp(\boldsymbol{\beta}^T\mathbf{x}_i)H_0(t)$$

$$\Longleftrightarrow$$

$$\log H(t|\ \mathbf{x}_i) = \boldsymbol{\beta}^T\mathbf{x}_i + \log H_0(t).$$

Mikäli malli on oikea, ero kahden elinaikahavainnon kumulatiivisen vaarafunktion logaritmissa ei siis riipu ajasta, vaan ainoastaan yksilöiden taustatiedoista. Oletusta voi tutkia jakamalla aineisto osa-aineistoihin tutkittavan selittäjän luokkien mukaisesti. Piirtämällä parametrittomasti estimoidun log-kumulatiivisen vaarafunktion kuvaajat aikaa vastaan samaan kuvioon osa-aineistoittain, pitäisi käyrien olla yhdensuuntaiset, mikäli suhteellisen vaaran oletus on voimassa. Jos käyrien etäisyys vaihtelee rajusti tai käyrät leikkaavat toisiaan, ei oletus ole voimassa, eikä Coxin mallia edes kannata sovittaa aineistoon. Kuvassa 2.1 on esimerkkikäyrät kaksiluokkaisen selittäjän tapauksessa molemmista tilanteista.



Kuva 2.1: Log-kumulatiivisen vaarafunktion kuvaajia. Vasemman kuvion tapauksessa suhteellisen vaaran oletus on voimassa, oikean kuvion tapauksessa ei

2.3.2 Residuaalitarkastelut

Sovitetun mallin hyvyyttä on mahdollista tutkia mallista laskettavien residuaalien (jäännösten) avulla, joita tässä esitellään lyhyesti. Kattavan kuvauksen Coxin mallin residuaaleista ja niiden soveltamisesta mallin sopivuuden tarkasteluun on luettavissa esimerkiksi lähteestä (Collet 2003, s. 111–150).

Cox-Snell residuaalit määritellään

$$r_{Ci} = \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i)\hat{H}_0(t_i),$$

$i = 1, \ldots, n$. Modifioidut Cox-Snell residuaalit ottavat huomioon sensuroidut havainnot ja ne määritellään

$$r'_{Ci} = 1 - \delta_i + r_{Ci},$$

(Collet 2003, s. 112–115). Cox-Snell residuaalit noudattavat likimain eksponenttijakaumaa keskiarvolla 1, mikäli sovitettu malli on oikea. Tätä ominaisuutta voi testata laskemalla residuaaleille Kaplan-Meier-estimaatin: jos malli on oikea, muodostavat termit $-\log(\hat{S}_{KM}(r'_{Ci}))$ tulostettuna residuaaleja r'_{Ci} vasten suoran, jonka kulmakerroin on 1 ja vakiotermi 0 (Collet 2003, s. 122).

Martingaaliresiduaalit ovat

$$r_{Mi} = \delta_i - r_{Ci}$$

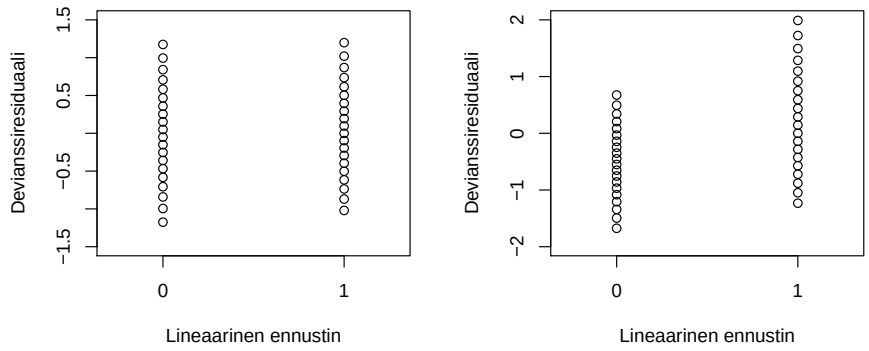
ja devianssiresiduaalit

$$r_{Di} = \text{sign}(r_{Mi}) \left[-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}\right]^{\frac{1}{2}},$$

missä $\text{sign}(r_{Mi})$ on martingaaliresiduaalin r_{Mi} etumerkki.

Coxin mallin osasta $\hat{\boldsymbol{\beta}}^T \mathbf{x}_i$ käytetään nimeä lineaarinen ennustin. Hyödyllinen kuvio saadaan, kun devianssiresiduaalit r_{Di} piirretään ennustimia $\hat{\boldsymbol{\beta}}^T \mathbf{x}_i$ vasten. Jos malli on oikea, ovat devianssiresiduaalit jakautuneet samalla tavoin riippumatta lineaarisen ennustimen arvosta. Kuvio paljastaa myös mahdolliset poikkeavat havainnot. Kuvassa 2.2 on havainnollistettu tilannetta yhden kaksiluokkaisen faktorin sisältävälle mallille. Lineaarisella ennustimella on tällöin kaksi mahdollista arvoa, ja tarkastelu tehdään vertaamalla residuaalien jakautumista pystyakselin suhteen.

Lisäksi voidaan tarkastella Schoenfeld-residuaaleja r_{Pij} tai score-residuaaleja r_{Sij} (Collet 2003, s. 117–119). Ne lasketaan aikaisemmin kuvatuista residuaaleista poiketen vektorin $\hat{\boldsymbol{\beta}}$ jokaiselle parametrille erikseen. Schoenfeld- ja score-residuaalien avulla voidaan tehdä yksityiskohtaisempia diagnostisia tarkasteluja, joihin ei kuitenkaan syvennytä tässä tutkielmassa.



Kuva 2.2: Residuaalien jakaumaa tarkastelevia kuvioita. Vasemmanpuoleinen kuva puoltaa mallin hyvyyttä, oikeanpuoleinen taas viittaa huonosti sopivaan malliin.

2.3.3 Vaikuttavat havainnot

Yksittäisen havainnon vaikuttavuutta mallin parametrien arvoihin voidaan tutkia sovittamalla malli ilman tutkittavaa havaintoa ja havainnon kanssa. Olkoon $\hat{\beta}_j$ parametrin β_j estimaatti koko aineistosta laskettuna ja $\hat{\beta}_{j(i)}$ vastaava estimaatti, kun havainto i on jätetty aineistosta pois. Erotus $\hat{\beta}_j - \hat{\beta}_{j(i)}$ kuvaa havainnon i vaikuttavuutta parametrin β_j estimaatin arvoon. Mikäli aineisto on suuri, on estimaatin $\hat{\beta}_{j(i)}$ laskeminen jokaiselle havainnolle hyvin työlästä. Erotus $\hat{\beta}_j - \hat{\beta}_{j(i)}$ voidaan kuitenkin korvata approksimaatiolla $\Delta_i\hat{\beta}_j$, joka määritellään vektorin

$$\mathbf{r}_{Si}{}^T\mathbf{var}(\hat{\boldsymbol{\beta}})$$

alkiona j. Vektori $\mathbf{r}_{Si}$ sisältää havainnon i score-residuaalit ja $\mathbf{var}(\hat{\boldsymbol{\beta}})$ on vektorin $\hat{\boldsymbol{\beta}}$ kovarianssimatriisi. Vaihtoehtoisesti voidaan laskea myös standardoidut arvot jakamalla $\Delta_i\hat{\beta}_j$:t estimaatin $\hat{\beta}_j$ keskihajonnalla. (Collet 2003, s. 132–133)

Parametrien arvoihin poikkeuksellisen paljon vaikuttavat havainnot paljastuvat yksinkertaisilla kuvioilla, jossa termit $\Delta_i\hat{\beta}_j$ tulostetaan havaintojen indeksejä vasten. Toinen hyödyllinen kuvio saadaan tulostamalla $\Delta_i\hat{\beta}_j$:t elinaikoja tai niiden järjestyslukuja vasten. (Collet 2003, s. 133)

Luku 3

Coxin mallin estimointi

Tässä luvussa käsitellään Coxin regressiomalliin liittyvää uskottavuuspäättelyä. Teoria perustuu pääosin lähteen (Collet 2003) lukuun 3.

3.1 Suurimman uskottavuuden estimointi

Coxin regressiomalli koostuu parametrittomasta perusvaarasta $h_0(t)$ sekä regressiokertoimet sisältävästä parametrisesta osasta $e^{\boldsymbol{\beta}^T \mathbf{x}_i}$. Mallin tärkeä ominaisuus on, että nämä kaksi komponenttia voidaan estimoida erikseen. Ensin estimoidaan vektori $\boldsymbol{\beta}$ ja mikäli kiinnostavaa on ainoastaan regressiokertoimiin liittyvä päättely, perusvaaraa ei tarvitse estimoida lainkaan. Jos myös $h_0(t)$:n estimointi on tarpeellista, onnistuu se vektorin $\hat{\boldsymbol{\beta}}$ avulla. (Collet 2003, s. 63)

Tässä tutkielmassa molemmat komponentit halutaan estimoida. Seuraavista kahdesta kappaleesta ensimmäinen käsittelee regressiokertoimien estimointia ja toisessa syvennytään perusvaaran estimointiin.

3.1.1 Regressiokertoimien estimointi

Olkoon havaintoaineisto $(t_i, \delta_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, kun t_i on havaittu elinaika, δ_i sensurointi-indikaattori ja $\mathbf{x}_i$ taustatiedot sisältävä vektori. Oletetaan ensin, että aineisto ei sisällä lainkaan sidoksia, eli oletetaan, että millä tahansa ajanhetkellä on tapahtunut korkeintaan yksi kuolema. Sensuroinnin oletetaan olevan epäinformatiivista, eli riippumatonta sekä t_i:stä että $\mathbf{x}_i$:stä. Oletetaan vielä, että n elinaikahavainnosta r on sensuroimattomia ja käytetään niistä järjestettyinä merkintöjä $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. Vastaavasti järjestettyihin elinaikoihin liittyvistä selittäjävektoreista käytetään merkintöjä $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(r)}$.

Collet (2003, s. 63–66) osoittaa, että näillä oletuksilla uskottavuusfunktio mallin (2.4) regressiokertoimille on

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}, \tag{3.1}$$

missä $R(t_i)$ on ajanhetkeen t_i liittyvä riskijoukko, eli juuri ennen t_i:tä elossa olevien yksilöiden muodostama joukko. Sensurointi-indikaattorin avulla uskottavuus voidaan esittää muodossa

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)} \right)^{\delta_i}. \tag{3.2}$$

Tavoitteena on löytää sellainen $\boldsymbol{\beta}$:n arvo, joka maksimoi lausekkeen (3.2) ja samalla lausekkeen (3.1). Uskottavuusfunktion logaritmi $l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$ on helpompi maksimoida. Sen lauseke on

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i (\boldsymbol{\beta}^T \mathbf{x}_i - \log \sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)). \tag{3.3}$$

Oletetaan, että $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. Maksimin löytämiseksi tarvitaan osittaisderivaatat jokaisen β_h:n suhteen, $h = 1, \ldots, p$. Merkitään $u_h(\beta_h) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_h}$ ja $\mathbf{u}(\boldsymbol{\beta}) = (u_1, \ldots, u_p)$. Tällöin,

$$u_h(\beta_h) = \sum_{i=1}^{n} \delta_i x_{ih} + \sum_{i=1}^{n} \delta_i \frac{\sum_{l \in R(t_i)} x_{lh} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}.$$

Lausekkeen (3.3) toiset derivaatat etumerkki vaihdettuna muodostavat informaatiomatriisin $\mathbf{I}(\boldsymbol{\beta}) = [I_{hg}(\boldsymbol{\beta})]_{p \times p} = [-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_g}]_{p \times p}$, jonka alkiot ovat

$$I_{hg}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \frac{\sum_{l \in R(t_i)} x_{lh} x_{lg} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}$$
$$- \sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{l \in R(t_i)} x_{lh} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)} \times \frac{\sum_{l \in R(t_i)} x_{lg} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)} \right\}.$$

Uskottavuusfunktion maksimoivalle $\hat{\boldsymbol{\beta}}$:lle pätee

$$\mathbf{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{3.4}$$

Yhtälö (3.4) voidaan ratkaista numeerisesti esimerkiksi Newtonin-Raphsonin menetelmällä (Collet 2003, s. 69) seuraavalla algoritmilla:

1. Valitaan alkuarvaus $\hat{\boldsymbol{\beta}}_0$ ja asetetaan $s = 0$.
2. Lasketaan $\hat{\boldsymbol{\beta}}_{s+1} = \hat{\boldsymbol{\beta}}_s + \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_s)\mathbf{u}(\hat{\boldsymbol{\beta}}_s)$.
3. Mikäli $\hat{\boldsymbol{\beta}}_{s+1}$:n ja $\hat{\boldsymbol{\beta}}_s$:n ero on riittävän pieni, on $\hat{\boldsymbol{\beta}}_{s+1}$ haluttu estimaatti. Muuten asetetaan $s = s + 1$ ja palataan kohtaan 2.

Jos aineisto sisältää sidoksia, ei $\hat{\boldsymbol{\beta}}$ ole löydettävissä lauseketta (3.1) maksimoimalla. Eksakti uskottavuusfunktio on monimutkainen ja sen maksimointi numeerisesti voi olla hyvin hidasta (Collet 2003, s. 67). Sidoksia sisältävän aineiston tapauksessa (3.1) korvataankin tavallisesti uskottavuusfunktion approksimaatiolla, jolle on useita vaihtoehtoja. Käytetyimpiä ovat Breslowin ja Efronin approksimaatiot (Collet 2003, s. 67–68), joiden määrittelyä varten otetaan käyttöön merkinnät:

$\mathbf{s}_j$ – summa ajanhetkellä $t_{(j)}$ kuolleiden yksilöiden selittäjien arvoista selittäjävektorin $\boldsymbol{\beta}$ jokaista parametria kohti

$D(t_{(j)})$ – ajanhetkellä $t_{(j)}$ kuolleiden yksilöiden muodostama joukko.

Breslowin approksimaatio vektorin $\boldsymbol{\beta}$ uskottavuusfunktiolle on

$$L_B(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \mathbf{s}_j)}{\left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)\right]^{d_j}},$$

missä d_j on kuten kappaleessa 2.2. Approksimaatio on yksinkertainen laskea ja se toimii hyvin, kun mitä tahansa yksittäistä ajanhetkeä kohti ei ole montaa sidosta.

Efronin approksimaatio on

$$L_E(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \mathbf{s}_j)}{\prod_{k=1}^{d_j} \left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \mathbf{s}_l) - (k-1) {d_k}^{-1} \sum_{l \in D(t_{(j)})} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)\right]}.$$

Breslowin approksimaatioon verrattuna Efronin approksimaatio on laskennallisesti raskaampi, mutta lähempänä eksaktia uskottavuusfunktiota (Collet 2003, s. 68).

Useimmissa tilasto-ohjelmistoissa (esimerkiksi R ja SAS) käyttäjä voi itse määrätä estimointiohjelman käyttämän laskentamenetelmän. Rakennusten aloitusaineistossa sidoksia on paljon ja siksi aineistoon tullaan soveltamaan Efronin approksimaatiota.

3.1.2 Perusvaaran estimointi

Oletetaan, että $\hat{\boldsymbol{\beta}}$ on estimoitu. Kalbfleisch & Prentice (1973) osoittavat, että suurimman uskottavuuden estimaatti perusvaaralle on

$$\hat{h}_0(t_{(j)}) = 1 - \hat{\xi}_j,$$

missä $\hat{\xi}_j$ toteuttaa yhtälön

$$\sum_{l \in D(t_{(j)})} \frac{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)}{1 - \hat{\xi}_j^{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)}} = \sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l). \tag{3.5}$$

Jos sidoksia ei ole, sisältää yhtälön (3.5) vasen puoli vain yhden summattavan termin ja saadaan helposti ratkaistua

$$\hat{\xi}_j = \left(1 - \frac{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)}\right)^{\exp(-\hat{\boldsymbol{\beta}}^T \mathbf{x}_{(j)})}.$$

Sidoksia sisältävän aineiston tapauksessa yhtälö (3.5) on kompleksisempi, eikä $\hat{\xi}_j$:lle löydy yleistä ratkaisua suljetussa muodossa. Ratkaisu pitää tällöin hakea iteratiivisesti. (Collet 2003, s. 97–99)

Jos oletetaan, että jokaisella välillä $[t_{(j)}, t_{(j+1)}[$ vaara on vakio, on estimaatti perustason välttöfunktiolle

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\xi}_j$$

ja perustason kumulatiiviselle vaarafunktiolle

$$\hat{H}_0(t) = \sum_{j=1}^{k} \log(\hat{\xi}_j),$$

kun $t_{(k)} \leq t < t_{(k+1)}$ (Collet 2003, s. 99).

Yhtälön (3.5) vasen puoli voidaan korvata approksimaatiolla, jolloin $\hat{\xi}_j$:n iteratiivinen ratkaiseminen sidostenkin tapauksessa vältetään. Collet (2003, s. 100–101) mainitsee approksimaation

$$\tilde{\xi}_j = \exp\left(\frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)}\right),$$

joka johtaa perustason välttöfunktion estimaattiin

$$\tilde{S}_0(t) = \prod_{j=1}^{k} \exp\left(\frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)}\right)$$

ja perustason kumulatiivisen vaarafunktion estimaattiin

$$\tilde{H}_0(t) = \sum_{j=1}^{k} \left(\frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)} \right),$$

kun $t_{(k)} \leq t < t_{(k+1)}$. Jälkimmäinen tunnetaan myös nimellä Nelson-Aalen-estimaattori tai Breslow-estimaattori.

Kun regressiokertoimet sekä perustason välttö- ja kumulatiivinen vaarafunktio on estimoitu, saadaan vastaavat ehdolliset funktiot

$$\hat{S}(t|\mathbf{x}_i) = [\hat{S}_0(t)]^{\exp(\hat{\beta}^T \mathbf{x}_i)} \tag{3.6}$$

ja

$$\hat{H}(t|\mathbf{x}_i) = \hat{H}_0(t) \exp(\hat{\beta}^T \mathbf{x}_i).$$

3.2 Uskottavuussuhteen testi

Kahden mallin sanotaan olevan sisäkkäiset, mikäli monimutkaisempi malli sisältää kaikki yksinkertaisemman mallin parametrit. Yleisesti Coxin regressiomallit

M1: $h(t|\ \mathbf{x}_i) = h_0(t) \exp\left(\beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right)$ ja

M2: $h(t|\ \mathbf{x}_i) = h_0(t) \exp\left(\beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \beta_{p+1} x_{(p+1)i} + \cdots + \beta_{p+q} x_{(p+q)i}\right)$

ovat sisäkkäisiä.

Sisäkkäisiä malleja voidaan verrata toisiinsa uskottavuussuhteen testillä. Nollahypoteesin (H_0) mukaan $\beta_{p+1}, \cdots, \beta_{p+q} = 0$. Testiä varten sovitetaan samaan aineistoon sekä **M1** että **M2** ja saadaan kaavan (3.1) mukaiset maksimoidut uskottavuusfunktion arvot $\hat{L}_1$ ja $\hat{L}_2$. Testisuure on

$$-2 \log \left[\frac{\hat{L}_1}{\hat{L}_2} \right]$$

ja se noudattaa asymptoottisesti χ^2-jakaumaa parametrilla q, kun nollahypoteesi on voimassa.(Collet 2003, s. 75–76) Pienet testisuureen arvot puoltavat **M1**:n riittävyyttä suhteessa malliin **M2**. Suuret arvot taas viittaavat siihen, että H_0:n mukaiset kiinnitykset ovat liian rajoittavia.

Luku 4

Logistinen regressiomalli

Oletetaan, että elinaikamuuttuja T on kuten kappaleessa 2.1. Kiinnitetyllä viiveellä t aloituksen rekisteröitymistä voidaan kuvata dikotomisella muuttujalla

$$Y = \begin{cases} 1, & T \leq t \ , \\ 0, & \text{muuten.} \end{cases}$$

Tässä luvussa esitellään logistinen regressiomalli, jolla mallinnetaan dikotomisen vasteen riippuvuutta taustamuuttujista. Tapauksesta $Y = 1$ käytetään geneeristä ilmaisua ”onnistuminen”. Lähteenä on käytetty teosta (Hosmer & Lemeshow 2000).

4.1 Mallin määrittely

Logistisessa regressiossa mallinnetaan onnistumisen todennäköisyyttä

$$\pi(\mathbf{x}_i) = P(Y = 1|\mathbf{x}_i).$$

Koska $\pi(\mathbf{x}_i)$ saa arvoja ainoastaan välillä $[0, 1]$, ei sitä voi mallintaa tavallisen regressiomallin tapaan. Malli onkin

$$\phi\left(\pi(\mathbf{x}_i)\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{\beta}^T \mathbf{x}_i,$$

missä ϕ on käytettävä linkkifunktio. Vektori $\boldsymbol{\beta}$ sisältää Coxin regressiomallista poiketen myös vakiotermin β_0. Tavallisin valinta linkkifunktioksi on logit-linkki, jolloin malli on muotoa

$$\log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} = \boldsymbol{\beta}^T \mathbf{x}_i. \tag{4.1}$$

Ratkaisemalla (4.1) $\pi(\mathbf{x}_i)$:n suhteen saadaan

$$\pi(\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)}$$

Mikäli malli sisältää kategorisia selittäjiä, on identifioituvuus jälleen varmistettava kiinnittämällä riittävä määrä parametreja nollaksi.

Mallin (4.1) osa $\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}$ on vedonlyöntisuhde. Kategoristen selittäjien tapauksessa vedonlyöntisuhde on siis e^{β_k}-kertainen yksilölle, johon liittyy faktorin taso k verrattuna yksilöön, jolla faktori on referenssitasollaan. Mikäli x_k on jatkuva selittäjä, muuttaa yhden yksikön kasvu sen arvossa vedonlyöntisuhdetta e^{β_k}-kertaiseksi.

4.2 Mallin estimointi

Olkoon havaintoaineisto $(y_i, \mathbf{x}_i)$, $i = 1, \cdots, n$, kun y_i on dikotominen vaste ja $\mathbf{x}_i$ taustatiedot sisältävä vektori. Mallin parametrit estimoidaan suurimman uskottavuuden menetelmällä. Y_i noudattaa Bernoullin jakaumaa parametrilla $\pi(\mathbf{x}_i)$, joten uskottavuusfunktio on

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i}.$$

Logaritminen uskottavuusfunktio on

$$\begin{aligned}
l(\boldsymbol{\beta}) &= \log(L(\boldsymbol{\beta})) \\
&= \sum_{i=1}^{n} [y_i \log[\pi(\mathbf{x}_i)] + (1 - y_i) \log[1 - \pi(\mathbf{x}_i)]] \\
&= \sum_{i=1}^{n} [y_i \log[\pi(\mathbf{x}_i)] - y_i \log[1 - \pi(\mathbf{x}_i)] + \log[1 - \pi(\mathbf{x}_i)]] \\
&= \sum_{i=1}^{n} \left[y_i \log \left[\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right] + \log[1 - \pi(\mathbf{x}_i)] \right] \\
&= \sum_{i=1}^{n} \left[y_i \boldsymbol{\beta}^T \mathbf{x}_i + \log \left[1 - \frac{e^{\boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}} \right] \right] \qquad (4.2)
\end{aligned}$$

Derivoimalla (4.2) jokaisen β_j:n suhteen ja asettamalla saadut lausekkeet nollaksi, saadaan $p + 1$ kpl uskottavuusyhtälöitä:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = \sum_{i=1}^{n} \left[y_i - \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \right] = 0 \qquad (4.3)$$

ja

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = \sum_{i=1}^{n} x_{ij}\left[y_i - \frac{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}\right] = 0, j = 1, \ldots, p \quad (4.4)$$

(Hosmer & Lemeshow 2000, s. 33). Yhtälöt (4.3) ja (4.4) ovat epälineaarisia ja ne ratkaistaan tavallisesti numeerisesti, esimerkiksi Newton-Raphson-menetelmällä kappaleen 3.1.1 tapaan. Tarvittavat informaatiomatriisin alkiot ovat

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_l} = \sum_{i=1}^{n} x_{ij}x_{il}\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)), \quad j, l = 0, \ldots, p$$

(Hosmer & Lemeshow 2000, s. 34).

4.3 Mallin diagnostiikkaa

Hosmer & Lemeshow (2000, s. 143–202) esittelee lukuisia testisuureita ja diagnostisia kuvia logistisen regressiomallin hyvyyden tarkasteluun. Tässä tutkielmassa aineiston suuri koko ja jatkuvan muuttujan mukanaolo mallissa aiheuttavat hankaluuksia diagnostisten tarkastelujen tekemiseen. Esimerkiksi yksinkertaisen, koko mallin sopivuutta tarkastelevan residuaalidevianssin χ^2-testin (Hosmer & Lemeshow 1989, s. 146) tekeminen ei ole mielekästä. Logistisen regressiomallin hyvyyttä tullaankin mittaamaan pääasiassa ennusteiden osuvuutta tarkastelemalla.

Hosmerin-Lemeshown testi (Hosmer & Lemeshow 2000, s. 147–149) voidaan kuitenkin tehdä. Siinä yksilöt jaetaan g (usein käytetään $g = 10$) likimain yhtä suureen järjestettyyn ryhmään estimoitujen todennäköisyyksien $\hat{\pi}(\mathbf{x}_i)$ mukaisesti. Ensimmäinen ryhmä sisältää ne ${n_1}' \approx n/g$ kpl yksilöitä, joilla todennäköisyydet $\hat{\pi}(\mathbf{x}_i)$ ovat pienimmät. Toinen ryhmä sisältää jäljelle jääneistä yksilöistä ne ${n_2}' \approx n/g$ kpl yksilöitä, joilla todennäköisyydet $\hat{\pi}(\mathbf{x}_i)$ ovat pienimmät, jne. Nollahypoteesina on, että sovitettu malli on oikea. Testisuure on

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k'\bar{\pi}_k)^2}{n_k'\bar{\pi}_k(1-\bar{\pi}_k)},$$

missä o_k on havaittu onnistumisten määrä ja $\bar{\pi}_k$ on estimoitujen todennäköisyyksien $\hat{\pi}(\mathbf{x}_i)$ keskiarvo ryhmässä k, $k = 1, \ldots, g$. Testisuure $\hat{C}$ noudattaa likimain χ^2-jakaumaa parametrilla $g-2$, kun nollahypoteesi on tosi. Testin tueksi voi piirtää kuvan, jossa termit $\bar{\pi}_k$ esitetään havaittujen onnistumisten suhteellisten osuuksien $\frac{o_k}{n_k'}$ funktiona. Mikäli malli on oikea, tulisi pisteiden sijaita likimain suoralla, jonka kulmakerroin on 1 ja vakiotermi 0.

Uskottavuussuhteen testiä voidaan käyttää myös logistisen regressiomallin valintatyökaluna. Testin konstruointi noudattaa kappaleessa 3.2 esiteltyä proseduuria.

Luku 5

Horvitz-Thompson-estimaattori

Rakennusten aloitusten kokonaistilavuuden estimoinnin perusajatus on, että jokaiselle tutkittavan kuukauden rekisteröityneelle aloitustiedolle estimoidaan sisältymistodennäköisyys otokseen, joka vastaa kaikkia tarkastelupäivään mennessä rekisteriin tulleita havaintoja. Havaitut tilavuudet ja niitä vastaavat sisältymistodennäköisyydet antavat mahdollisuuden estimoida aloitettujen rakennusten kokonaismäärää otantateorian keinoin. Sisältymistodennäköisyydet vaihtelevat yksilöiden välillä, sillä rekisteröitymisviiveen jakauma riippuu rakennushankkeen taustatiedoista. Tämä pitää ottaa huomioon estimaattoria valittaessa.

Tässä luvussa esitellään Horvitz-Thompson-estimaattori (Horvitz & Thompson 1952) (HT-estimaattori) ja johdetaan sille olennaisimpia tuloksia. Lähteenä on käytetty oppikirjaa (Lohr 1999).

Estimoituja sisältymistodennäköisyyksiä ja Horvitz-Thompson-tyyppistä estimaattoria populaation koon laskennassa ovat soveltaneet muun muassa Yip ym. (1999) estimoidessaan erään lintupopulaation kokoa additiivisen vaaran mallin avulla. Heijden ym. (2003) arvioivat laittomien siirtolaisten lukumäärää Hollannissa Poisson-regressiolla tuotettujen sisältymistodennäköisyyksien avulla. Alho (1990) puolestaan käyttää logistista regressiomallia sisältymistodennäköisyyksien estimointiin tutkimuksessaan, jonka empiirisessä osassa arvioidaan eri ammattitauteihin sairastuneiden lukumäärää Suomessa. Rakennusten aloitusten estimointiongelma eroaa edellisistä siinä, että populaation koon sijaan kiinnostava tunnusluku on populaation kokonaissumma tilavuusmuuttujan suhteen. Lisäksi käytössä oleva aineisto mahdollistaa poikkeuksellisella tavalla menetelmien tuottamien ennusteiden vertailun toteutuneisiin summiin.

5.1 Perustuloksia

Oletetaan, että kokoa N olevalle perusjoukolle halutaan laskea summa muuttujan v suhteen. Merkitään

$$\sum_{i=1}^{N} v_i = V.$$

Oletetaan, että perusjoukosta on poimittu riippumattomasti ja palauttamatta n yksilön suuruinen otos. Olkoon Z_i indikaattorimuuttuja yksilön i sisältymiselle otokseen, siis

$$Z_i = \begin{cases} 1, & \text{kun yksilö } i \text{ sisältyy otokseen} \\ 0, & \text{muuten.} \end{cases}$$

Oletetaan, että sisältymistodennäköisyydet ovat

$$\pi_i = P(Z_i = 1), \quad i = 1, \cdots, N$$

ja

$$\pi_{ik} = P(Z_i = 1 \text{ ja } Z_k = 1), \quad i, k = 1, \cdots, N$$

ja aluksi, että π_i ja π_{ik} ovat riippumattomia muuttujasta v. Silloin HT-estimaattori perusjoukon summalle V on

$$\hat{V}_{HT} = \sum_{i=1}^{N} Z_i \frac{v_i}{\pi_i}. \tag{5.1}$$

Lause 5.1. *Edellä mainituin oletuksin pätee*

$$\mathbf{E}(\hat{V}_{HT}) = V,$$

ja

$$\mathbf{var}(\hat{V}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} {v_i}^2 + \sum_{i=1}^{N} \sum_{k=1;k\neq i}^{N} (\pi_{ik} - \pi_i \pi_k) \frac{v_i v_k}{\pi_i \pi_k}. \tag{5.2}$$

Todistus. Z_i noudattaa Bernoullin jakaumaa parametrilla π_i, joten $\mathbf{E}[Z_i] = \pi_i$ ja $\mathbf{var}(Z_i) = \pi_i(1-\pi_i)$. Lisäksi $\mathbf{cov}(Z_i, Z_k) = \pi_{ik} - \pi_i \pi_k$, kun $i \neq k$. Suoraan laskemalla

$$\mathbf{E}(\hat{V}_{HT}) = \mathbf{E}\left[\sum_{i=1}^{N} Z_i \frac{v_i}{\pi_i}\right] = \sum_{i=1}^{N} \left(\mathbf{E}[Z_i] \frac{v_i}{\pi_i}\right) = V$$

ja

$$\begin{aligned}
\mathbf{var}(\hat{V}_{HT}) &= \mathbf{var}\left[\sum_{i=1}^{N} Z_i \frac{v_i}{\pi_i}\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{N}\left(\frac{v_i v_k}{\pi_i \pi_k}\mathbf{cov}(Z_i, Z_k)\right) \\
&= \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}{v_i}^2 + \sum_{i=1}^{N}\sum_{k=1;k\neq i}^{N}(\pi_{ik} - \pi_i\pi_k)\frac{v_i v_k}{\pi_i \pi_k}
\end{aligned}$$

□

Käytännössä varianssi (5.2) tulee estimoida. Lohr (1999, s. 207–209) osoittaa, että harhaton estimaattori on

$$\widehat{\mathbf{var}}(\hat{V}_{HT}) = \sum_{i\in U}\frac{1-\pi_i}{{\pi_i}^2}{v_i}^2 + \sum_{i\in U}\sum_{k\in U;k\neq i}\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}\frac{v_i v_k}{\pi_i \pi_k}, \tag{5.3}$$

missä U on otokseen poimittujen yksilöiden muodostama joukko.

5.2 HT-estimaattori rakennusten aloituksille

Rakennushankkeen aloituksen rekisteröintiviive riippuu hankkeen tilavuudesta, joten myöskään Z_i ei ole riippumaton v_i:stä. Suurten rakennusten aloitustiedot rekisteröityvät keskimäärin pienten hankkeiden tietoja nopeammin. Kaava (5.1) johtaa tällöin keskimäärin liian suuriin estimaatteihin. Harhaton HT-estimaattori saadaan korvaamalla π_i:t niitä vastaaviin tilavuuksiin v_i ehdollistetuilla todennäköisyyksillä. Tästä eteenpäin sisältymistodennäköisyydet määritellään

$$\pi_i = P(Z_i = 1|v_i).$$

Estimaattorin odotusarvo on

$$\begin{aligned}
\mathbf{E}(\hat{V}_{HT}) &= \mathbf{E}\left[\sum_{i=1}^{N} Z_i \frac{v_i}{\pi_i}\right] \\
&= \sum_{i=1}^{N}\left(\mathbf{E}[Z_i v_i]\frac{1}{\pi_i}\right) \\
&= \sum_{i=1}^{N}\left(\mathbf{E}[\mathbf{E}(Z_i v_i|v_i)]\frac{1}{\pi_i}\right) \\
&= \sum_{i=1}^{N}\left(v_i\mathbf{E}[\mathbf{E}(Z_i|v_i)]\frac{1}{\pi_i}\right) \\
&= \sum_{i=1}^{N}\left(v_i P(Z_i=1|v_i)\frac{1}{\pi_i}\right) \\
&= V
\end{aligned}$$

Tässä tutkielmassa varianssi (5.2) yksinkertaistuu olennaisesti. Rakennusten aloitusten tapauksessa Z_i:t ovat nimittäin toisistaan riippumattomia, sillä tieto yksilön i sisältymisestä otokseen ei vaikuta yksilön j sisältymistodennäköisyyteen. Samalla otoskoko n on satunnainen. Riippumattomuudesta seuraa, että varianssi on

$$\begin{aligned}
\mathbf{var}(\hat{V}_{HT}) &= \mathbf{var}\left[\sum_{i=1}^{N} Z_i \frac{v_i}{\pi_i}\right] \\
&= \sum_{i=1}^{N} \mathbf{var}\left[Z_i \frac{v_i}{\pi_i}\right] \\
&= \sum_{i=1}^{N} \frac{1}{{\pi_i}^2}\left[\mathbf{E}({v_i}^2 {Z_i}^2) - (\mathbf{E}(v_i Z_i))^2\right] \\
&= \sum_{i=1}^{N} \frac{1}{{\pi_i}^2}\left[\mathbf{E}(\mathbf{E}[{v_i}^2 Z_i|v_i]) - (\mathbf{E}(\mathbf{E}[v_i Z_i|v_i]))^2\right] \\
&= \sum_{i=1}^{N} \frac{1}{{\pi_i}^2}\left[{v_i}^2 P(Z_i=1|v_i) - {v_i}^2 P(Z_i=1|v_i)^2\right] \\
&= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i}{v_i}^2. \qquad (5.4)
\end{aligned}$$

Varianssin (5.4) estimaattori on

$$\widehat{\mathbf{var}}(\hat{V}_{HT}) = \sum_{i=1}^{N} Z_i {v_i}^2 \frac{1-\pi_i}{{\pi_i}^2}. \qquad (5.5)$$

Se on harhaton, sillä

$$\begin{aligned}
\mathbf{E}\left[\sum_{i=1}^{N} Z_i {v_i}^2 \frac{1-\pi_i}{{\pi_i}^2}\right] &= \sum_{i=1}^{N} \frac{(1-\pi_i)\mathbf{E}[Z_i {v_i}^2]}{{\pi_i}^2} \\
&= \sum_{i=1}^{N} \frac{(1-\pi_i)\mathbf{E}[\mathbf{E}(Z_i {v_i}^2 | v_i)]}{{\pi_i}^2} \\
&= \sum_{i=1}^{N} \frac{(1-\pi_i){v_i}^2 P(Z_i = 1|v_i)}{{\pi_i}^2} \\
&= \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} {v_i}^2 \\
&= \mathbf{var}(\hat{V}_{HT})
\end{aligned}$$

Varianssi (5.4) ja sen estimaattori (5.5) saadaan helposti myös kaavoista (5.2) ja (5.3). Muuttujien Z_i riippumattomuudesta seuraa, että $\pi_{ik} = \pi_i \pi_k$. Kaavoihin jää tällöin ainoastaan ensimmäiset summalausekkeet, jotka vastaavat tässä kappaleessa johdettuja tuloksia.

Esitellyt varianssikaavat eivät ota huomioon sisältymistodennäköisyyksien estimointiin liittyvää tilastollista epävarmuutta ja johtavat harhaisiin estimaatteihin, kun π_i:t eivät ole kiinteitä. Heijden ym. (2003) huomioivat epävarmuuden ja laskevat delta-menetelmällä varianssin approksimaation käyttämälleen HT-estimaattorille. Jo yksinkertaisen Poisson-regression tapauksessa approksimointi johtaa kompleksisiin laskuihin ja tuloksiin. Tässä tutkielmassa sisältymistodennäköisyydet vastaavat Coxin regressiomallilla estimoituja kertymäfunktion arvoja ja logistisella regressiomallilla estimoituja ennusteita. Etenkin Coxin mallin tapauksessa HT-estimaattorin varianssin approksimointi olisi vielä huomattavasti Poisson-regressiota työläämpää (semiparametrisesta mallista johtuen). Tutkielmassa onkin päädytty yksinkertaisuuden vuoksi olettamaan estimoidut sisältymistodennäköisyydet kiinteiksi. Myöskään oletus otoksen riippumattomuudesta ei täysin päde – elinajat ovat luultavasti klusteroituneita kuntien sisällä. Oletuksista aiheutuvan harhan suuruutta arvioidaan ennusteiden osuvuutta ja estimoidun varianssin määräämien luottamusvälien pitävyyttä tarkastelemalla.

Luku 6

Sovellus: aloitettujen rakennusten kokonaistilavuuden estimointi

Tässä luvussa kuvatuista Rakennus- ja asuntotuotanto -tilaston yksityiskohdista voi lukea tarkemmin lähteestä (Suomen virallinen tilasto).

6.1 Rakennus- ja asuntotuotanto -tilaston kuvaus

Rakentaminen on Suomessa luvanvaraista toimintaa. Rakennuslupaa edellyttäviä hankkeita ovat esimerkiksi uuden rakennuksen rakentaminen, rakennuksen laajentaminen sekä rakennuksen rakentamiseen verrattavat korjaus- ja muutostyöt. Lupajärjestelmä takaa muun muassa sen, että rakennus valmistuessaan täyttää asetetut turvallisuus- ja ympäristövaatimukset. Lisäksi järjestelmä mahdollistaa rakennustoiminnan tilastoinnin.

Tilastokeskus tuottaa Väestörekisterikeskuksen ylläpitämän rakennustietorekisterin tiedoista kuukausittain tilaston Suomen rakennustoiminnan suhdanteista. Julkaisu sisältää kuukausittain tiedot myönnetyistä rakennusluvista sekä uudisrakentamisen volyymi-indeksin tuoreimmat luvut. Kuukausitiedot aloitettujen ja valmistuneiden rakennusten määrästä julkaistaan neljännesvuosittain. Uudisrakentamisen volyymi-indeksi perustuu kuitenkin rakennusten aloitustietoihin, joten mahdollisimman tarkkoja tietoja tarvitaan myös kuukausitasolla. Kuhunkin tietoon liittyy varsinaisen tilastoidun arvon lisäksi tieto muutoksesta edellisen vuoden arvoon prosentteina.

Rakennus- ja asuntotuotanto -tilasto tarjoaa tärkeää informaatiota rakentamisen kehityksestä rakentamisen sektorin toimijoille. Tilaston tietoja käytetään rakentamisen suhdannetilanteen seurantaan ja rakentamisen kehityksen

ennakointiin. Rakentamisen määrän ennakointi on tärkeää, jotta esimerkiksi rakennustarvikkeiden menekkiä voidaan ennustaa ja näin suunnitella tulevia tuotantomääriä. Tilaston laadinnan yhteydessä syntyvää tietoa uudisrakentamisen arvosta käytettään myös esimerkiksi kansantalouden tilinpidon talonrakentamisen investointien ja tuotoksen laskennassa.

Rakennus- ja asuntotuotanto -tilasto julkaistaan tilastoitavan kuukauden lopusta noin kahdeksan viikon viiveellä. Tilaston tuotantoajankohta on noin viikko ennen julkaisua. Esimerkiksi kesäkuun 2011 tiedot julkaistiin 30.8.2011. Tiedot täydentyvät kuukausittain lopullisen tilaston tuotantoajankohtaan, tilastoitavaa vuotta seuraavan vuoden puoliväliin saakka. (Suomen virallinen tilasto)

6.1.1 Aineiston synty

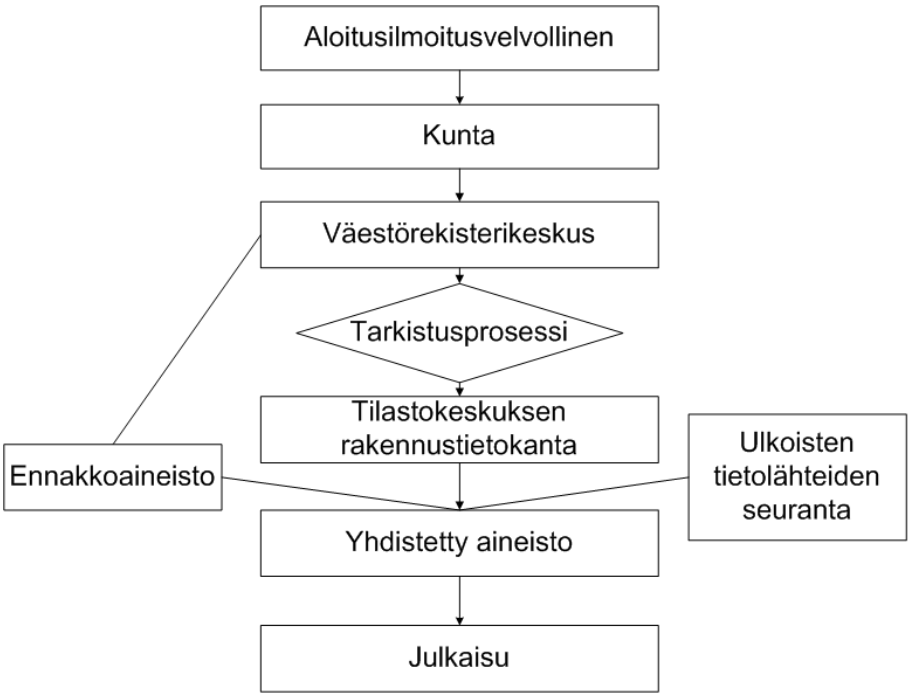
Rakennushanketietojen kulkeutumisprosessi Tilastokeskuksen julkaisuihin on esitetty kuvassa 6.1. Rakennusluvan myöntää kunnan rakennusvalvonnan viranomainen. Aloitustieto perustuu rakennustarkastajan pitämän aloituskatselmukseen, pohjakatselmukseen tai rakentajan ilmoitukseen hankkeen aloituksesta. Omavalvontakohteisssa rakennusvalvonta saa ilmoituksen aloitetusta rakennustyöstä rakennusyritykseltä. Vastaavasti rakentamisen katsotaan päättyvän rakennustarkastajan tekemään käyttöönottotarkastukseen tai loppukatselmukseen, jonka perusteella määräytyy rakennuksen valmistumisajankohta. Käytännöt vaihtelevat hieman kunnittain. Kunta toimittaa tiedon myönnetystä luvasta tai rakennushankkeen aloituksesta Väestörekisterikeskukselle (VRK), missä tiedot käyvät läpi virheentarkastusprosessin ennen kuin ne lisätään varsinaiseen rekisteriin. Tilastokeskus lisää uudet, rekisteriin tulleet havainnot omiin tietokantoihinsa säännöllisesti.

Vuodesta 2001 alkaen Tilastokeskus on täydentänyt rakennustietokantaansa myös VRK:n ennakkoaineistolla, joka sisältää ne havainnot, jotka eivät ole käyneet läpi virheentarkastusta (Väestörekisterikeskus 2001). Vanhoja ennakkotietoja ei ole saatavilla, joten tämän käytännön tuomaa hyötyä ei tarkastella tässä tutkielmassa.

Tietoja suurille rakennushankkeille myönnetyistä luvista sekä suurien hankkeiden aloituksista kerätään myös VRK:sta riippumattomista tietolähteistä. Hankkeen tiedot lisätään tilastoa tuotettaessa aineistoon, mikäli tietoa ei ole saatu VRK:lta tilaston tuotantopäivään mennessä.

6.1.2 Alipeitto-ongelma

Etenkin rakennusten aloitustietojen kertyminen on hidasta. Rekisteröitymisviivettä aiheuttavat kuntien tiedonantokäytännöt sekä aloitusilmoitusvelvol-



Kuva 6.1: Rakennushanketietojen kulkuprosessi

listen viivyttely ilmoitusten teossa. Lisäksi VRK:n tarkistusprosessi viivästyttää tiedonsaantia. Uusimman tilastoitavan kuukauden tieto tulee tuottaa jo noin seitsemän viikkoa kyseisen kuukauden jälkeen, joten tuotantohetkellä tilasto perustuu vielä pahasti alipeittoiseen aineistoon. Alipeitosta johtuen eri ajankohtien tiedot eivät ole suoraan vertailukelpoisia, ennen kuin lopulliset tiedot julkaistaan.

Aloitettujen rakennusten tapauksessa alipeittoa korjataan korotuskertoimilla. Nämä muodostetaan laskemalla käyttötarkoitusryhmittäin kolmen edellisen vuoden keskimääräiset alipeitot rakennusten yhteenlasketussa tilavuudessa tarkasteltavana kuukautena. Käyttötarkoitusryhmiä ovat erilliset asuinpientalot, muut asuinrakennukset, vapaa-ajanrakennukset, maatalousrakennukset ja muut rakennukset. Korottaminen kohdistetaan ainoastaan tilavuudeltaan alle 20 000 kuutiometrin rakennushankkeisiin.

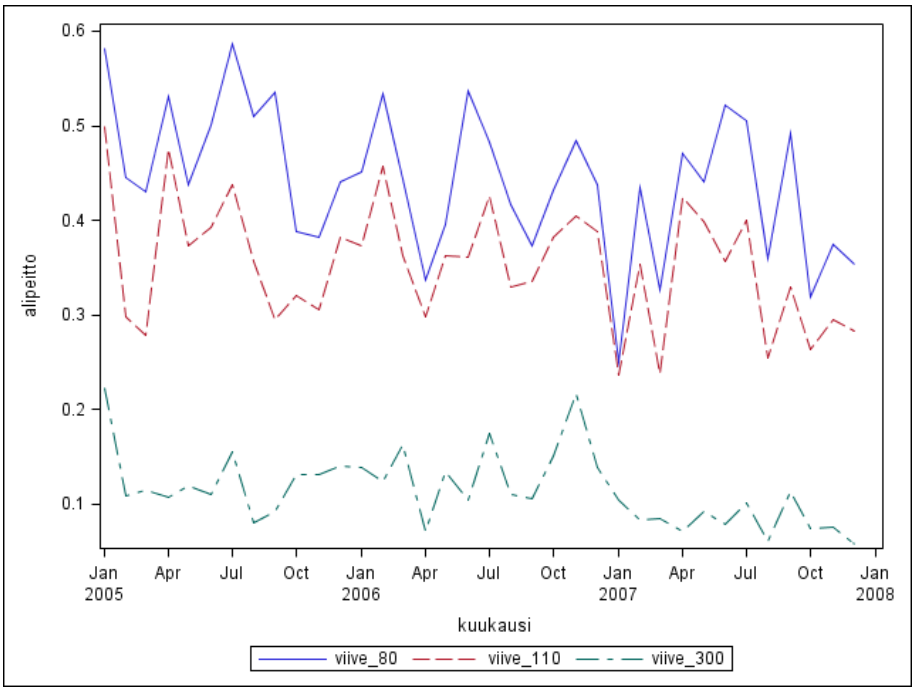
Oletetaan, että käyttötarkoitusluokassa i tutkittavan kuukauden rekisteröityneiden aloitusten kuutiotilavuus on V_i^{hav} ja kolmen edellisen vuoden keskimääräinen alipeitto (tutkittavaa kuukautta vastaavilta kuukausilta) on $\hat{A}_i$, $0 < \hat{A}_i < 1$. Estimaatti kaikkien käyttötarkoitusluokan i aloitettujen rakennusten tilavuudelle on

$$\hat{V}_i = \frac{V_i^{hav}}{1 - \hat{A}_i}.$$

Kaikkien aloitettujen rakennusten tilavuus saadaan aggregoimalla,

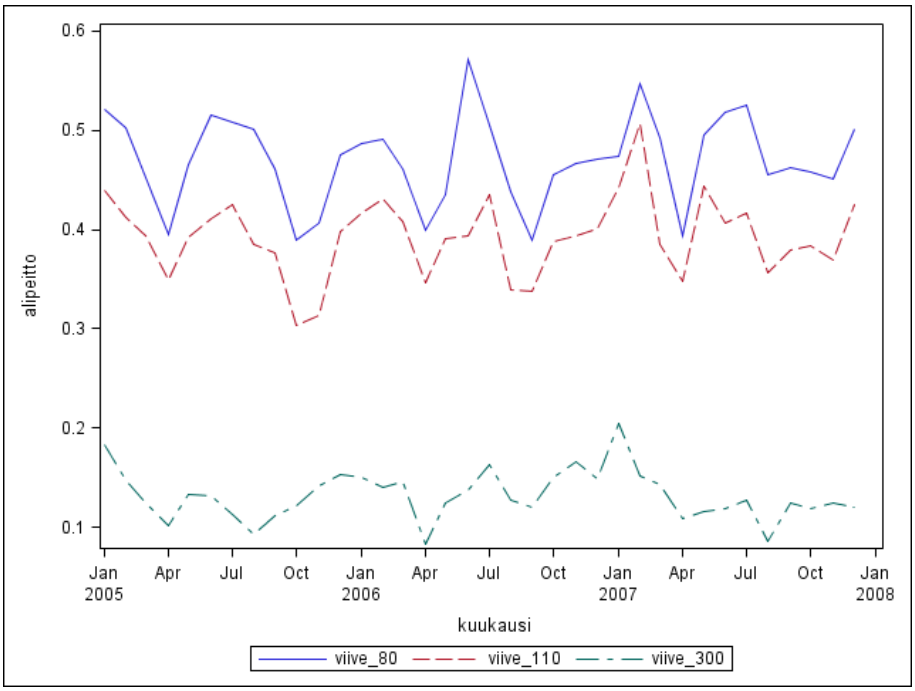
$$\hat{V}_{total} = \sum_{i=1}^{5} \hat{V}_i.$$

Merkittävin tilaston tarkkuutta heikentävä tekijä on alipeiton vaihtelu. Kuvissa 6.2 ja 6.3 on esitetty aikasarjat suhteellisista alipeitoista rakennusten yhteenlasketussa tilavuudessa eri pituisilla viiveillä vuosina 2005–2007. Ensimmäisessä kuvassa on mukana kaikki rakennukset ja toisessa ainoastaan alle 20 000 kuution rakennushankkeet. Kuvat osoittavat, että alipeitto voi vaihdella kymmeniä prosentteja kuukausien välillä. Vaihtelu pienenee, kun suuret hankkeet jätetään tarkastelun ulkopuolelle. Isojen hankkeiden mahdollisimman nopea rekisteröityminen olisikin tilaston laadun kannalta ensiarvoisen tärkeää. Huomattavaa on myös, että vielä 300 vuorokauden viiveellä aineisto on merkittävästi puutteellinen.



Kuva 6.2: Suhteelliset alipeitot aloitettujen rakennusten yhteenlasketussa tilavuudessa vuosina 2005–2007, kaikki rakennukset. Viiveet 80, 110 ja 300 vrk

Muutosprosentit lasketaan vertaamalla uusinta, korottamatonta, tietoa edellisen vuoden vastaavaan ennakkotietoon. Nykyinen käytäntö ei huomioi tilaston tuotantopäivän vaihtelua vertailtavien vuosien välillä, vaan muutosprosentti lasketaan suoraan tilastosta, joka tuotettiin edellisenä vuonna tutkittavalle kuukaudelle. Tämä aiheuttaa harhaa etenkin uusimman tilastoitavan



Kuva 6.3: Suhteelliset alipeitot aloitettujen rakennusten yhteenlasketussa tilavuudessa vuosina 2005–2007, vain alle 20 000 m^3 rakennukset. Viiveet 80, 110 ja 300 vrk

kuukauden tietoihin. Tässä tutkielmassa esitetyt menetelmät hyödyntävät päiväkohtaista rekisteröintitietoa, eikä vaihtelu tilaston tuotantopäivässä aiheuta niiden tuottamiin estimaatteihin harhaa. Suurin epävarmuutta aiheuttava tekijä on kuitenkin myös muutosprosenttien tapauksessa aloitustietojen alipeitto ja sen vaihtelu.

6.2 Aineiston kuvaus

6.2.1 Muuttujat

Aloitustiedon rekisteröintiviive perustuu aloituspäivämäärän ja aloituksen rekisteröintipäivämäärän eroon. Rekisteröitymisajankohta tunnetaan tarkasti, mutta aloituspäivämäärä on tiedossa vain kuukauden tarkkuudella. Esimerkiksi tammikuussa aloitetun ja 15. helmikuuta rekisteröidyn aloitustiedon viive on vähintään 15 ja enintään 45 vuorokautta. Sellaisenaan viive on siis intervallisensuroitu elinaikamuuttuja. Tässä tutkielmassa mallinnettava muuttuja määritellään aloituskuukauden ensimmäisen päivän ja rekisteröintipäivän ajallisena etäisyytenä, jota käsitellään sensuroimattomana elinaikamuuttuja-

na.

Rakennushankkeisiin liittyy lukuisa määrä taustatietoja. Tässä työssä käytetyt muuttujat ovat *kuntalaji*, *käyttötarkoitus* ja *aloituskuukausi*. Muuttuja *kuntalaji* jakaa havainnot viiteen ryhmään sen perusteella, kuinka suuri havaintoon liittyvän kunnan rakennustuotannon vuotuinen arvo on. Muuttuja perustuu Tilastokeskuksessa vuonna 2010 tehtyyn selvitykseen. Muuttuja *käyttötarkoitus* vastaa kappaleessa 6.1.2 kuvattua jakoa. Lisäksi hankkeen tilavuus on jossain muodossa sisällytettävä malliin selittäjäksi, mikäli se todetaan merkitseväksi. Luvun 5 perusteella vain tällöin mallia voi käyttää aloitettujen rakennusten kokonaistilavuuden ennustamiseen. Graafisten tarkastelujen ja pohdinnan jälkeen on päädytty käyttämään tilavuuden logaritmia (*logtilav*). Muuttujien interaktiot jätetään tarkastelujen ulkopuolelle.

6.2.2 Aineiston rajaus

Tilastokeskuksen rakennustietokanta sisältää yli miljoonan rakennushankkeen tiedot. Tässä tutkielmassa rajoitutaan vuonna 2003 tai sitä myöhemmin aloitettuihin hankkeisiin. Vuonna 2000 aloitettu VTJ2000-järjestelmän sisäänajo VRK:ssa ja kunnissa viivästytti tietojen rekisteröitymistä vuosituhannen alussa (Väestörekisterikeskus 2001). Siksi vanhempia tietoja ei sisällytetä aineistoon.

Luvanvaraisia korjaus- ja muutostöitä ei tilastoida Rakennus- ja asuntotuotanto -tilastossa, eikä niitä sisällytetä aineistoon tässäkään. Nykyisen korotuskäytännön mukaisesti myös suuret, yli 20 000 m^3, rakennushankkeet jätetään aineiston ulkopuolelle.

Nykyisin lopullinen tilasto julkaistaan tilastoitavaa vuotta seuraavan vuoden puolivälissä. Tietoja, jotka rekisteröidään tämän jälkeen, ei sisällytetä julkaistaviin tilastoihin. Alkuvuonna aloitetut rakennukset ehtivät kertyä rekisteriin huomattavasti loppuvuoden aloituksia pidempään, joten kuukausitietojen vertailu on mielekästä ainoastaan samannimisten kuukausien välillä. Liukuvan käytännön, jossa jokaisen kuukauden tiedot kertyisivät yhtä pitkän ajan, käyttöönotosta on keskusteltu. Tässä tutkielmassa on päädytty rajaamaan aineisto niihin aloituksiin, joiden rekisteröintiviive on ollut korkeintaan 300 vuorokautta.

Aineistosta poistetaan lisäksi ne aloitukset, joiden kuntakoodi on virheellinen. Niiden osuus koko aineistosta on häviävän pieni.

Poistettu muuttuja	df	χ^2-testisuure	P
käyttötarkoitus	4	246.6	< 0.0001
aloituskuukausi	11	2101.1	< 0.0001
kuntalaji	4	3879.1	< 0.0001
logtilavuus	1	1188.2	< 0.0001

Taulukko 6.1: Uskottavuussuhteen testien tulokset logistiselle regressiomallille

Poistettu muuttuja	df	χ^2-testisuure	P
käyttötarkoitus	4	166.3	< 0.0001
aloituskuukausi	11	2704.1	< 0.0001
kuntalaji	4	3745.6	< 0.0001
logtilavuus	1	495.2	< 0.0001

Taulukko 6.2: Uskottavuussuhteen testien tulokset Coxin regressiomallille

6.3 Tulokset

Tässä kappaleessa testataan esiteltyjen menetelmien soveltuvuus rakennusten aloitusten estimointiin käytännössä. Aineisto jaetaan kahteen osaan: Muuttujien valinta ja mallien diagnostiset tarkastelut tehdään 01/2003–04/2007 aloitettujen rakennusten tietojen pohjalta. Jäljelle jäävää osaa aineistosta käytetään mallien ennustuskyvyn testaamiseen. Kiinnostavaa on uusilla malleilla saavutettavan hyödyn arviointi, joten ennustaminen tehdään myös nykyisellä korotusmallilla.

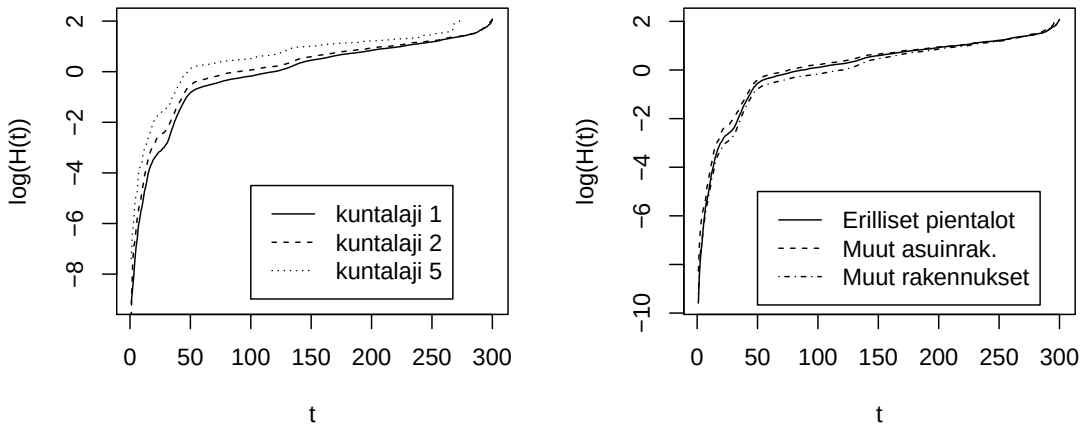
6.3.1 Mallien valinta

Sovitetaan aikavälillä 01/2003–04/2007 aloitettuihin rakennuksiin Coxin regressiomalli sekä logistinen regressiomalli, jossa dikotominen vaste saa arvon 1, kun *viive* on alle 80 vuorokautta ja muuten arvon 0. Valittu 80 vuorokauden raja vastaa likimain viivettä, jolla ensimmäinen ennakollinen tilasto nykyisin tuotetaan.

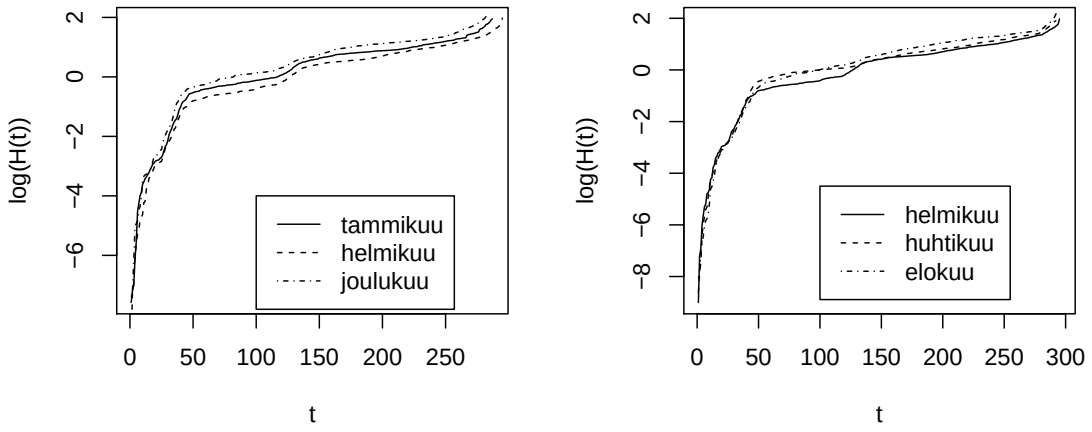
Kun kaikki muuttujat sisältävää mallia verrataan uskottavuussuhteen testillä malliin, josta on vuorotellen jätetty yksi muuttuja pois, saadaan logistiselle regressiomallille taulukon 6.1 ja Coxin regressiomallille taulukon 6.2 mukaiset tulokset. Molempien mallien tapauksessa kaikki neljä selittäjämuuttujaa ovat merkitseviä.

Coxin mallin toimivuuden kannalta ratkaisevaa suhteellisen vaaran oletusta

on tutkittu kuvissa 6.5 ja 6.4. Selvyyden vuoksi kuvissa on vain kolme käyrää kerrallaan. Muuttujien *kuntalaji* ja *käyttötarkoitus* tapauksessa oletus toteutuu kohtuullisesti, mutta muuttuja *aloituskuukausi* tuottaa ongelmia. Tammi-, helmi- ja joulukuun käyrät ovat siististi samansuuntaiset, mutta esimerkiksi helmi-, huhti- ja elokuun käyrät leikkaavat toisiaan.

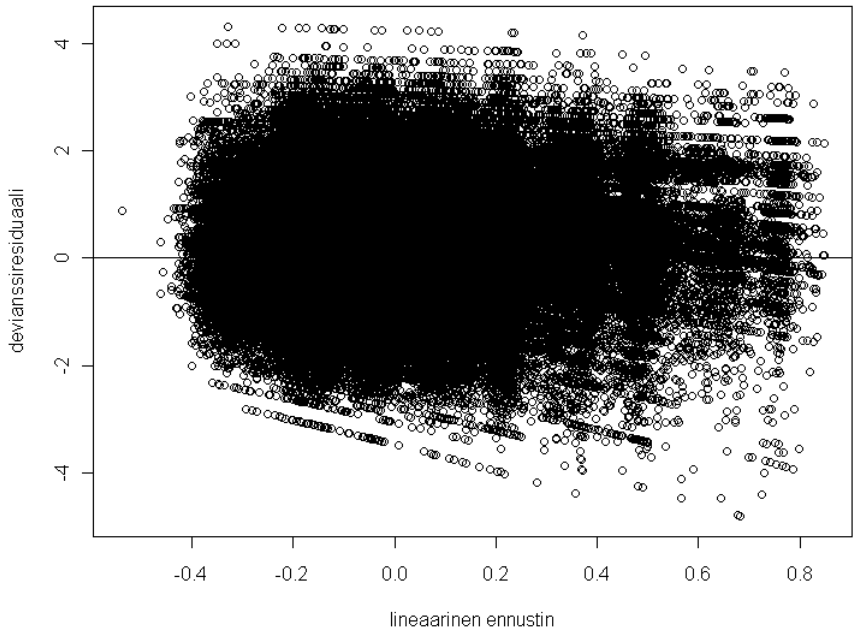


Kuva 6.4: Log-kumulatiivisen vaarafunktion kuvaajat muuttujille *kuntalaji* ja *käyttötarkoitus*



Kuva 6.5: Log-kumulatiivisen vaarafunktion kuvaaja muuttujalle *aloituskuukausi*

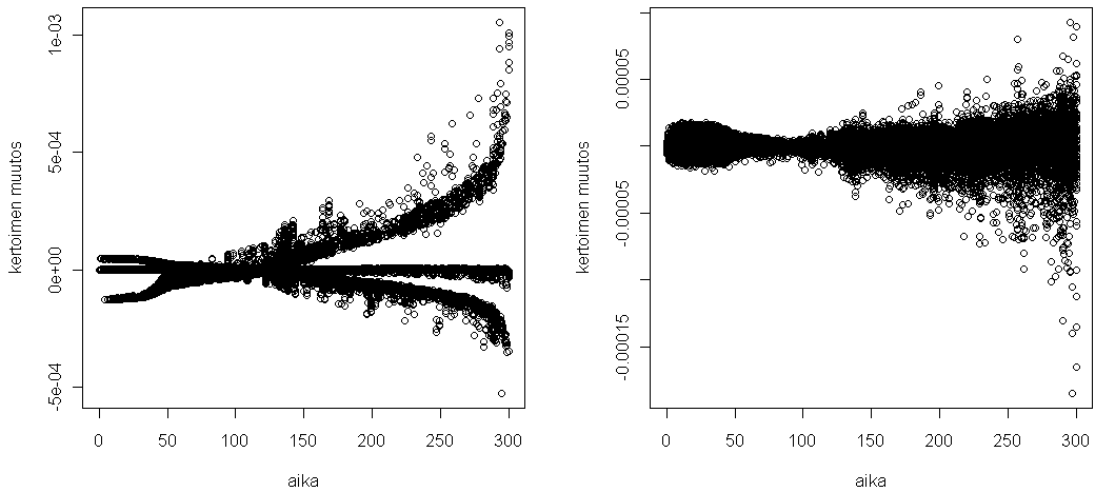
Coxin mallista lasketut devianssiresiduaalit on esitetty lineaarisia ennustimia vasten kuvassa 6.6. Kuvasta nähdään, että pienillä lineaarisen ennustimen arvoilla residuaalit ovat painottuneet positiiviselle puolelle. Tämä voi johtua esimerkiksi siitä, että muuttuja *logtilav* ei välttämättä ole yhteydessä vaarafunktioon mallin olettamalla tavalla.



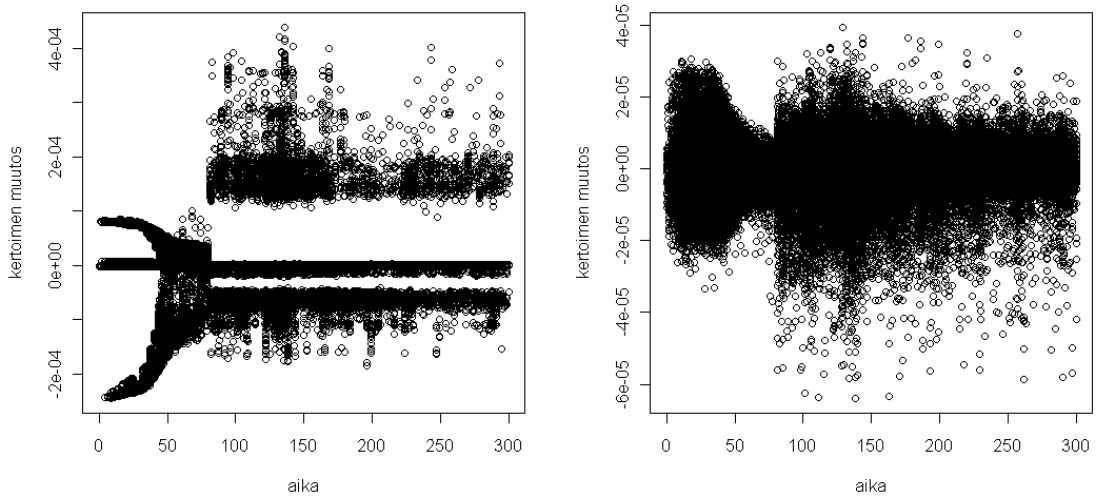
Kuva 6.6: Coxin mallin devianssiresiduaalit vs. lineaariset ennusteet

Yksittäisten havaintojen vaikuttavuutta Coxin mallin parametrien estimaatteihin on tutkittu kuvissa 6.7 ja 6.8. Kuvassa 6.7 termit $\Delta_i\hat{\beta}_j$ on tulostettu elinaikaa vasten parametreille *aloituskuukausi8* ja *logtilav*. Kuvioista nähdään, että mitä pidemmästä elinajasta on kyse, sitä enemmän se vaikuttaa estimaattien arvoihin (sama ilmiö toistuu kaikille parametreille). Tässä tutkielmassa mallia on tarkoitus käyttää rekisteröitymisviiveen kertymäfunktion ennustamiseen tilaston tuotantopäivänä, joka seuraa tilastoitavan kuukauden ensimmäistä päivää noin 80 vuorokauden viiveellä. On mahdollista, että suuret elinaikahavainnot vipuavat parametrien arvoja siten, että malli toimii huonosti tällä kiinnostavalla viiveellä. Siksi tarkastellaan myös mallia, jossa havainnot sensuroidaan sen perusteella, onko havaittu elinaika yli vai alle tuotantopäivän määräämän viiveen. Kuva 6.8 osoittaa, että suurten elinaikojen vaikuttavuus pienenee olennaisesti, kun malli sovitetaan sensuroituun aineistoon. Seuraavassa kappaleessa ennustaminen suoritetaan sekä sensuroimatto-

malla että sensuroidulla Coxin mallilla, jotta sensuroinnin tarpeellisuudesta voidaan tehdä päätelmiä.



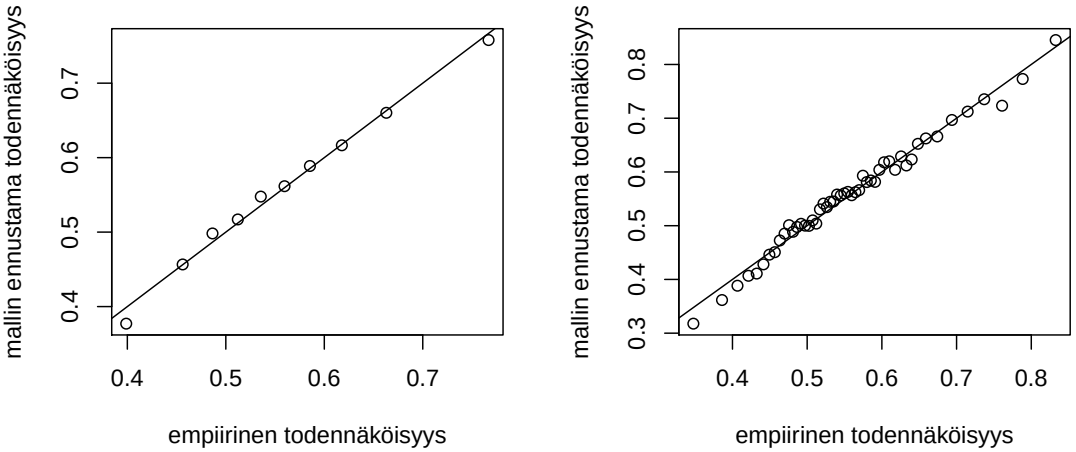
Kuva 6.7: Havaintojen vaikuttavuus parametreihin *aloituskuukausi8* (vasemmalla) ja *logtilav*, sensuroimaton Coxin malli



Kuva 6.8: Havaintojen vaikuttavuus parametreihin *aloituskuukausi8* (vasemmalla) ja *logtilav*, sensuroitu Coxin malli

Logistiselle regressiomallille tehdyn Hosmerin-Lemeshown testin testisuureen

arvo, 79.0, on selvästi merkitsevä ($df = 8$, $P < 0.0001$). Malli ei siis testin perusteella sovitu aineistoon erityisen hyvin. Kuva 6.9 sen sijaan osoittaa, että ennustetuille todennäköisyyksille ryhmittäin lasketut keskiarvot $\bar{\pi}_k$ ovat hyvin lähellä empiirisiä vastineitaan. Selvästi merkitsevä tulos Hosmerin-Lemeshown testissä johtuukin luultavasti aineiston suuresta koosta, eikä mallia ole syytä hylätä.



Kuva 6.9: Ryhmittäin lasketut ennustettujen todennäköisyyksien keskiarvot $\bar{\pi}_k$ tulostettuna vastakkain empiiristen todennäköisyyksien $\frac{o_k}{n'_k}$ kanssa, vasemmalla ryhmien lukumäärä $g = 10$ ja oikealla $g = 50$

Sensuroimattoman Coxin mallin ja logistisen regressiomallin parametrien estimaatit on esitetty liitteessä A taulukoissa A.1 ja A.2. Parametrit osoittavat, että esimerkiksi rakennustoiminnan määrältään suurimpien kuntien (*kuntalaji* = 5) tiedot rekisteröityvät muiden kuntien tietoja nopeammin. Lisäksi huomataan, että loppuvuoden tiedot rekisteröityvät alkuvuoden tietoja nopeammin. Yleisesti parametrien tulkinnat mallien välillä eivät ole ristiriitaisia.

6.3.2 Ennustaminen

Tämän tutkielman kiinnostavin kysymys on, että onko esiteltyjen mallien avulla mahdollista tuottaa nykyistä laadukkaampaa tilastotietoa. Tätä tutkitaan jäljittelemällä todellisia korotustilanteita kuukausitiedoille 01/2008 - 06/2010. Kokeissa oletetaan, että tilasto on tuotettu tilastoitavaa kuukautta

seuraavan toisen kuukauden 20. päivä. Esimerkiksi tammikuun tiedon oletetaan olevan tuotettu 20. maaliskuuta. Ainoastaan uusimman tilastoitavan kuukauden kokonaismäärää ennustetaan. Testausprosessi noudattaa seuraavaa algoritmia:

1. Asetetaan tutkittava kuukausi $kk = 01/2008$ ja tilaston tuotantopäivämäärä $pvm = 20.3.2008$. Lisäksi asetetaan estimointiaineiston rajaavat kuukaudet $kk1 = 01/2005$ ja $kk2 = 04/2007$. Merkitään kuukauden kk ensimmäisen päivän ja päivämäärän pvm eroa symbolilla t_{kk} (vuorokautta).

2. Rajataan korotettava aineisto siten, että se sisältää ne kuukautena kk aloitetut rakennukset, jotka on rekisteröity viimeistään päivämääränä pvm.

3. Rajataan estimointiaineisto aikavälillä *kk1* - *kk2* aloitettuihin rakennuksiin.

4. Sovitetaan estimointiaineistoon Coxin regressiomalli ja logistinen regressiomalli. Jos käytetään sensuroitua Coxin mallia, tehdään sensurointi sen perusteella, onko *viive* yli vai korkeintaan 83 vuorokautta (näin varmistutaan, että sensurointiaika on aina suurempi kuin t_{kk}). Logistisen mallin dikotomisen vasteen arvo on 1, mikäli *viive* on pienempi kuin t_{kk}, muuten 0.

5. Lasketaan molempiin malleihin perustuvat sisältymistodennäköisyyksien estimaatit korotettavan aineiston jokaiselle yksilölle. Coxin regressiomallin tapauksessa yksilön sisältymistodennäköisyys vastaa elinajan kertymäfunktion estimaattia hetkellä t_{kk} ehdolla yksilön taustatiedot, siis $\hat{F}(t_{kk}|\mathbf{x}_k) = 1 - \hat{S}(t_{kk}|\mathbf{x}_k)$, joka saadaan kaavasta (3.6). Logistisen regressiomallin tapauksessa sisältymistodennäköisyys on suoraan $\hat{\pi}(\mathbf{x}_k) = \frac{\exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_k)}{1+\exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_k)}$.

6. Lasketaan HT-estimaatti (5.1) aloitusten kokonaismäärälle molempien mallien tapauksessa. Lasketaan lisäksi nykyisellä käytännöllä tuotettu estimaatti.

7. Lasketaan kk:n todellinen aloitettujen rakennusten kokonaismäärä.

8. Siirretään kuukausia kk, *kk1*ja *kk2* sekä päivämäärää pvm kuukaudella eteenpäin.

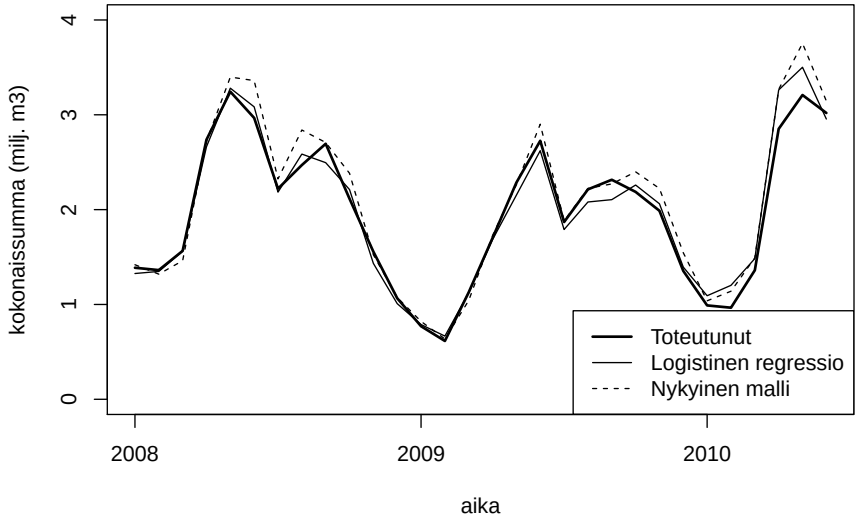
9. Mikäli kk on 07/2010, lopetetaan. Muuten palataan kohtaan 2.

Estimointiaineiston rajaus kohdassa 1. perustuu nykykäytäntöön, jossa korotuskertoimet lasketaan kolmen edellisen vuoden tiedoista.

Logistisella regressiomallilla, sensuroidulla Coxin regressiomallilla (Cox 1) sekä nykyisellä mallilla saadut kuukausikohtaiset tulokset on esitetty liitteessä A taulukossa A.3. Taulukossa A.4 on logistisen regressiomallin ennusteille likimääräiset 95 %:n luottamusvälit, jotka on laskettu varianssiestimaatin (5.5) avulla:

$$\hat{V}_{HT} \pm 2\sqrt{\widehat{\mathbf{var}}(\hat{V}_{HT})}.$$

Tulosten yhteenveto on taulukoissa 6.3 ja 6.4, jotka sisältävät myös sensuroimattomalla Coxin mallilla (Cox 2) tuotettujen ennusteiden tunnuslukuja. Kuvassa 6.10 on esitetty logistisella regressiolla ja nykyisellä korotusmallilla tuotetut ennusteet aikasarjoina sekä toteutuneet kuukausittaiset kokonaistilavuudet.



Kuva 6.10: Logistisella regressiolla ja nykyisellä korotusmallilla tuotetut ennusteet sekä toteutuneet kuukausittaiset kokonaistilavuudet

Tulosten tulkinta riippuu osittain siitä, tarkastellaanko ennustevirheiden keskiarvoja vai mediaaneja. Logistinen regressiomalli tuottaa kuitenkin lähes kaikilla mittareilla parhaat ennusteet. Ennusteen absoluuttinen virhe (ennustetun ja toteutuneen arvon erotuksen itseisarvo) on keskimäärin 103 000 m^2, kun se nykyisellä korotusmallilla on keskimäärin 135 000 m^2. Vastaava suhteellinen virhe on logistisella regressiomallilla keskimäärin 5.5 ja nykyisellä mallilla keskimäärin 6.7 prosenttia. Logistisella regressiomallilla saavutet-

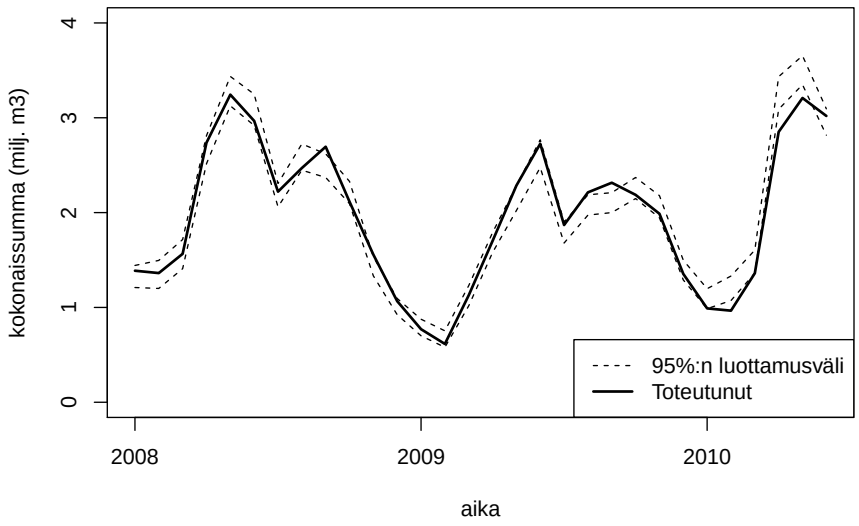
tiin siis keskimäärin 1.2 prosenttiyksikön parannus ennusteen osuvuudessa. Coxin regressiomallin tulokset paranevat oleellisesti, kun pitkien elinaikojen sensurointi otetaan käyttöön. Ennusteet jäävät silti laadultaan nykyisen korotusmallin tuottamien ennusteiden tasolle.

Muutosprosenttiestimaatit on mahdollista laskea liitteen A taulukon A.3 tuloksista kuukausille 01/2009–06/2010. Estimaatit on laskettu kaavalla

$$\frac{\hat{V} - V_{ev}}{V_{ev}},$$

missä $\hat{V}$ on jollakin mallilla laskettu kuukausiestimaatti ja V_{ev} on edellisen vuoden vastaavan kuukauden (tunnettu) aloitusten kokonaistilavuus. Logistinen regressiomalli tuottaa myös parhaat muutosprosenttiestimaatit. Suhteessa nykyiseen korotusmalliin sillä saavutetaan noin 1.1 prosenttiyksikköä pienempi keskimääräinen revisio. Keskiarvoilla mitattuna molemmat Coxin mallit ennustavat muutosprosentteja huonosti, mutta mediaanien perusteella sensuroitu Cox toimii nykymallia paremmin.

Logistisen regressiomallin ennusteille saadut luottamusvälit sisältävät toteutuneen arvon 21 tapauksessa 30:sta. Luottamusvälit ja toteutuneet arvot on esitetty kuvassa 6.11. Vaikuttaisi siltä, että estimoitu varianssi on systemaattisesti liian pieni. Tämä johtuu luultavasti yksinkertaistavista oletuksista, jotka tehtiin varianssin laskennan helpottamiseksi.



Kuva 6.11: Logistisella regressiolla ja tuotetut ennusteiden luottamusvälit sekä toteutuneet kuukausittaiset kokonaistilavuudet

	Logistinen	Cox 1	Cox 2	Nykymalli
Absoluuttinen virhe (m^3)	103 000	139 000	176 000	135 000
Suhteellinen virhe (%)	5.5	7.1	8.9	6.7
Muutospros. revisio (pros.yks.)	8.0	10.9	13.3	9.1

Taulukko 6.3: Mallien ennustuskykyä mittaavien tunnuslukujen keskiarvot

	Logistinen	Cox 1	Cox 2	Nykymalli
Absoluuttinen virhe (m^3)	77 000	85 000	93 000	95 000
Suhteellinen virhe (%)	4.3	4.1	6.5	4.8
Muutospros. revisio (pros.yks.)	3.9	5.5	8.3	5.7

Taulukko 6.4: Mallien ennustuskykyä mittaavien tunnuslukujen mediaanit

Luku 7

Pohdinta

Tutkielman tarkoituksena oli rakennusten aloitustietojen rekisteröitymisen tilastollinen mallintaminen ja aloitusten kokonaistilavuuden ennustaminen alipeittoisesta rekisteriaineistosta. Rekisteröitymisviivettä mallinnettiin Coxin regressiomallilla ja rekisteröitymistä kiinnitetyn aikarajan sisällä kuvaavaa dikotomista muuttujaa logistisella regressiomallilla. Kokonaismäärän estimoinnin ajatuksena oli käsitellä tilaston tuotantopäivään mennessä rekisteriin tullutta kuukausiaineistoa otoksena kaikkien tutkittavana kuukautena aloitettujen rakennusten muodostamasta perusjoukosta. Otoksen yksilölle määrättiin sisältymistodennäköisyydet, jotka estimoitiin edellä mainituilla tilastollisilla malleilla aikaisemmista, jo lopullisena tunnetuista tiedoista. Kokonaismäärän ennuste tuotettiin havaituista tilavuuksista ja estimoiduista sisältymistodennäköisyyksistä Horvitz-Thompson-estimaattorin avulla. Tuloksia verrattiin nykyistä korotusmallia jäljittelevän menetelmän antamiin ennusteisiin, jolloin saatiin tietoa myös nykyisen korotuskäytännön tarkkuudesta.

Kaikki tutkittavat muuttujat osoittautuivat potentiaalisiksi selittäjiksi, kun mallit sovitettiin aikavälillä 01/2003–04/2007 aloitettujen rakennusten tietoihin. Coxin regressiomallin suhteellisen vaaran oletus tuotti ongelmia etenkin muuttujan *aloituskuukausi* tapauksessa. Lisäksi pitkien elinaikojen todettiin vaikuttavan voimakkaasti estimaattien arvoihin. Vaihtoehtoisena mallina päädyttiin siksi tarkastelemaan Coxin regressiomallia, jossa pitkät elinajat sensuroidaan.

Logistinen regressio tuotti parhaat ennustetulokset – ennusteiden suhteellinen virhe oli keskimäärin 1.2 prosenttiyksikköä pienempi kuin nykymallilla tuotettujen ennusteiden virhe. Myös estimoituja muutosprosentteja tarkastelemalla logistinen regressiomalli oli paras. Coxin regressiomallin tulokset paranivat huomattavasti, kun pitkät elinajat sensuroitiin. Lopputulokset jäivät kuitenkin selvästi logistisen mallin tuloksia heikommiksi. Logistista regressio-

mallia voidaan siten pitää suositeltavampana menetelmänä HT-estimaattorin sisältymistodennäköisyyksien estimointiin rakennusten aloitusten tapauksessa.

Aloitusten kokonaismäärän estimaatin tarkkuudesta ei nykyisin tuoteta minkäänlaista tietoa julkaisun yhteydessä. Tässä tutkielmassa esiteltiin HT-estimaattorin varianssin approksimaatio, joka on helposti laskettavissa itse estimaatin laskennan yhteydessä. Varianssiapproksimaation avulla on helppo tuottaa aloitusten kokonaistilavuuden estimaatille esimerkiksi likimääräinen luottamusväli. Tutkielmassa luottamusvälit osoittautuivat hieman liian kapeiksi, mikä johtuu luultavasti varianssin approksimoinnissa tehdyistä yksinkertaistavista oletuksista. Varianssiapproksimaatiolla saatiin kuitenkin kohtuullinen arvio estimaatin tarkkuudesta.

Esiteltyjen menetelmien käyttöönottoa harkittaessa tulee arvioida saavutettavaa hyötyä suhteessa implementoinnin vaatimiin resursseihin. Nykymalli menestyi jo nyt vertailussa melko hyvin ja eri malleilla tuotettujen ennusteiden suhteelliset erot pienenevät, kun suurien hankkeiden kokonaistilavuudet lisätään kaikkiin ennusteisiin korottamattomina. Nykyinen korotusmalli ei myöskään edellytä estimointia tilastollisella ohjelmistolla, vaan korotus voidaan tehdä yksinkertaisilla tietokantaoperaatioilla. Korotusjärjestelmän ylläpito vaatisi lisäksi nykyistä enemmän tilastotieteen asiantuntemusta, mikäli korotus tehtäisiin tilastolliseen malliin perustuen.

On tärkeä muistaa, että muutokset tiedonsaantikäytännöissä vaikuttavat estimaattien laatuun. Jos esimerkiksi VRK onnistuisi nopeuttamaan tiedonsaantia kunnilta, tuottaisivat edellisen kolmen vuoden tietoihin perustuvat korotusmenetelmät systemaattisesti liian suuria estimaatteja kolmen vuoden ajan. Tilaston laatu siis heikkenisi, mikäli estimointikäytäntöä ei muutettaisi, vaikka itse aineisto olisi laadukkaampaa.

Tutkielman tärkein tulos on se, että esitelty Horvitz-Thompson-estimaattoriin pohjautuva korotusmenetelmä soveltuu perusjoukon kokonaissumman estimointiin hitaasti täydentyvästä rekisteriaineistosta. Menetelmää voidaan siten suositella sovellettavan myös muihin vastaavan tyyppisiin ongelmiin. Esimerkiksi rakennuslupiin liittyvä alipeitto-ongelma on täysin analoginen aloitusten ongelman kanssa. Jatkotutkimuksena voisikin selvittää, voidaanko rakennuslupatilaston tai valmistuneiden rakennusten tilaston laatua parantaa esitellyillä menetelmillä. Rakennus- ja asuntotuotanto -tilaston lisäksi Tilastokeskuksessa tuotetaan muitakin tilastoja, joissa menetelmää voitaisiin mahdollisesti käyttää hyväksi.

Lähdeluettelo

Alho, J. (1990): Logistic Regression in Capture-recapture Models. *International Biometric Society*, vol. 46, s. 623–635.

Collet, D. (2003): *Modelling Survival Data in Medical Research*, 2. painos. Chapman & Hall/CRC, Boca Raton.

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

Heijden, P., Bustami, R., Cruyff, M., Engbersen, G. & Houwelingen, H. (2003): Point and Interval Estimation of the Population Size Using the Truncated Poisson Regression Model. *Statistical Modelling*, vol. 3, s. 305–322.

Horvitz, D. & Thompson, D. (1952): A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, vol. 47, s. 663–685.

Hosmer, D. & Lemeshow, S. (2000): *Applied Logistic Regression*, 2. painos. Wiley, New York.

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

Klein, J. & Moeschberger, M. (1997): *Survival Analysis: Statistical Methods for Censored and Truncated Data.* Springer, New York.

Lohr, S. (1999): *Sampling: Design and Analysis.* Brooks/Cole Publishing Company, Pacific Grove.

R Development Core Team (2011): R, A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org

SAS Institute Inc. (2009). SAS OnlineDoc® 9.2. Cary, NC: SAS Institute Inc.

Suomen virallinen tilasto (SVT): *Rakennus- ja asuntotuotanto [verkkojulkaisu]*. ISSN=1796-3257. Helsinki: Tilastokeskus [viitattu: 15.9.2011]. Saantitapa: http://www.stat.fi/til/ras.

Väestörekisterikeskus (2001): *Väestörekisterikeskuksen ja Tilastokeskuksen yhteistyön kehittämistä selvittäneen työryhmän loppuraportti* (julkaisematon). Väestörekisterikeskus, Helsinki.

Yip, P., Zhou, Y., Lin, D. & Fang, X.(1999): Estimation of Population Size Based on Additive Hazards Hodels for Continuous-time Recapture Experiments. *Biometrics*, vol. 55, s. 904–908.

Liite A

Parametrien estimaatteja ja estimointituloksia

Taulukko A.1: Sensuroimattoman Coxin mallin parametrien estimaatit

	β	$\exp(\beta)$	**P**
kuntalaji2	0.01	1.02	0.0093
kuntalaji3	0.26	1.29	< 0.0001
kuntalaji4	-0.03	0.97	0.0016
kuntalaji5	0.53	1.70	< 0.0001
aloituskuukausi2	-0.11	0.90	< 0.0001
aloituskuukausi3	-0.12	0.88	< 0.0001
aloituskuukausi4	0.02	1.02	0.1422
aloituskuukausi5	0.01	1.01	0.6169
aloituskuukausi6	0.03	1.03	0.0436
aloituskuukausi7	0.04	1.04	0.0036
aloituskuukausi8	0.13	1.14	< 0.0001
aloituskuukausi9	0.19	1.21	< 0.0001
aloituskuukausi10	0.28	1.33	< 0.0001
aloituskuukausi11	0.28	1.32	< 0.0001
aloituskuukausi12	0.19	1.21	< 0.0001
käyttötarkoitus2	-0.07	0.94	< 0.0001
käyttötarkoitus3	-0.05	0.95	< 0.0001
käyttötarkoitus4	-0.11	0.89	< 0.0001
käyttötarkoitus5	-0.05	0.95	< 0.0001
logtilav	0.04	1.04	< 0.0001

Taulukko A.2: Logistisen regressiomallin parametrien estimaatit

	β	$\exp(\beta)$	**P**
β_0	-0.75	0.47	< 0.0001
kuntalaji2	0.08	1.08	< 0.0001
kuntalaji3	0.61	1.84	< 0.0001
kuntalaji4	-0.04	0.96	0.0257
kuntalaji5	1.07	2.92	< 0.0001
aloituskuukausi2	-0.32	0.73	< 0.0001
aloituskuukausi3	-0.17	0.84	< 0.0001
aloituskuukausi4	0.32	1.37	< 0.0001
aloituskuukausi5	0.19	1.21	< 0.0001
aloituskuukausi6	0.02	1.02	0.4866
aloituskuukausi7	0.11	1.12	< 0.0001
aloituskuukausi8	0.25	1.29	< 0.0001
aloituskuukausi9	0.43	1.53	< 0.0001
aloituskuukausi10	0.46	1.59	< 0.0001
aloituskuukausi11	0.44	1.55	< 0.0001
aloituskuukausi12	0.38	1.47	< 0.0001
käyttötarkoitus2	-0.13	0.88	< 0.0001
käyttötarkoitus3	-0.10	0.90	< 0.0001
käyttötarkoitus4	-0.32	0.73	< 0.0001
käyttötarkoitus5	-0.12	0.89	< 0.0001
logtilav	0.14	1.15	< 0.0001

Taulukko A.3: Havaitut, ennustetut ja toteutuneet aloitukset (m^3)

Kuukausi	Havaittu	Toteutunut	Sensuroitu Cox	Logistinen	Nykymalli
01/2008	864925	1387104	1334571	1347981	1419288
02/2008	738287	1366383	1354991	1333892	1320060
03/2008	883623	1565828	1563306	1582755	1461331
04/2008	1799209	2737459	2663129	2673568	2720579
05/2008	2065992	3243766	3280977	3271037	3398317
06/2008	1758992	2967419	3084094	3064053	3360919
07/2008	1303392	2221458	2184279	2283637	2323485
08/2008	1673964	2469798	2585852	2638733	2839844
09/2008	1727067	2695605	2495899	2526417	2708525
10/2008	1543586	2115951	2212471	2203730	2384192
11/2008	985176	1556664	1433788	1439199	1524752
12/2008	663369	1066004	1009186	1021128	1074534
01/2009	505645	770468	784165	792532	821610
02/2009	345861	615457	666286	663163	637093
03/2009	620488	1129876	1139419	1162256	1042313
04/2009	1136549	1701970	1684496	1695051	1726650
05/2009	1376643	2285714	2153099	2144959	2263884
06/2009	1570394	2726489	2621695	2603498	2901418
07/2009	1079490	1868987	1800131	1874770	1881334
08/2009	1431547	2215300	2073719	2107997	2223893
09/2009	1467076	2315731	2106975	2113290	2272849
10/2009	1576873	2187955	2260571	2266953	2399869
11/2009	1425567	1987898	2062447	2068136	2223462
12/2009	922279	1356799	1394837	1415786	1546897
01/2010	693600	989851	1088589	1108735	1038457
02/2010	651312	966032	1183826	1182702	1140677
03/2010	858655	1360624	1505567	1542306	1492706
04/2010	2204252	2852925	3278703	3338435	3268638
05/2010	2294507	3207835	3501033	3486359	3752442
06/2010	1823680	3022721	2956955	2944206	3146718

Taulukko A.4: Logistisella regressiolla lasketut luottamusvälit

Kuukausi	Toteutunut	Alaraja	Yläraja
01/2008	1387104	1209944	1442723
02/2008	1366383	1200322	1495841
03/2008	1565828	1408579	1716006
04/2008	2737459	2514652	2810676
05/2008	3243766	3125756	3435619
06/2008	2967419	2919033	3248714
07/2008	2221458	2060901	2307615
08/2008	2469798	2447309	2724060
09/2008	2695605	2370533	2619991
10/2008	2115951	2094449	2328831
11/2008	1556664	1332727	1533915
12/2008	1066004	922973	1094195
01/2009	770468	700922	875683
02/2009	615457	579657	753173
03/2009	1129876	1016877	1229979
04/2009	1701970	1579024	1789432
05/2009	2285714	2021995	2284167
06/2009	2726489	2475468	2767666
07/2009	1868987	1678900	1901343
08/2009	2215300	1973570	2187548
09/2009	2315731	2000121	2210144
10/2009	2187955	2148178	2370496
11/2009	1987898	1944917	2178643
12/2009	1356799	1289602	1498013
01/2010	989851	986605	1197514
02/2010	966032	1072456	1330891
03/2010	1360624	1362359	1604647
04/2010	2852925	3089938	3433696
05/2010	3207835	3348824	3653810
06/2010	3022721	2816733	3094160

Liite B

R-koodit

```
#############################################################################
#                                                                           #
# Tällä ajojonolla testataan coxin regressiomallin ja logisti-              #
# sen regressiomallin soveltuvuutta rakennusten aloitusten kuu-             #
# kausittaisen kokonaistilavuuden estimointiin.                             #
#                                                                           #
# Aineisto ''tiedot'' sisältää seuraavat muuttujat:                         #
#                                                                           #
#    apvm       - aloituskuukausi muodossa vvvvkk                           #
#                                                                           #
#    alorekpvm - aloituksen rekisteröintipäivämäärä muodossa                #
#                vvvvkkpp                                                   #
#                                                                           #
#    viive      - aloituskuukauden ensimmäisen päivän ja rekiste-           #
#                röintipäivämäärän ero vuorokausissa                        #
#                                                                           #
#    logtilav   - rakennushankkeen tilavuuden logaritmi                     #
#                                                                           #
#    akk        - aloituskuukausi (1 - 12)                                  #
#                                                                           #
#    ktark      - rakennuksen käyttötarkoitus (1 - 5)                       #
#                                                                           #
#    kuntalaji - kunnan rakennustoiminnan määrää kuvaava                    #
#                muuttuja (1 - 5)                                           #
#                                                                           #
#############################################################################

## Ladataan elinaikamallien kirjasto
library(survival)

## Luodaan testikehikko
## matriisin 1. sarake kuvaa ennustettavaa kuukautta
## 2. sarake on tilaston tuotantopäivä
## 3. sarake on estimointidatan ensimmäinen kuukausi
## 4. sarake on estimointidatan viimeinen kuukausi
```

```
design <- matrix(nrow = 30, ncol = 4)
design[1,] <- c(200801, 20080320, 200704, 200501)
design[2,] <- c(200802, 20080420, 200705, 200502)
design[3,] <- c(200803, 20080520, 200706, 200503)
design[4,] <- c(200804, 20080620, 200707, 200504)
design[5,] <- c(200805, 20080720, 200708, 200505)
design[6,] <- c(200806, 20080820, 200709, 200506)
design[7,] <- c(200807, 20080920, 200710, 200507)
design[8,] <- c(200808, 20081020, 200711, 200508)
design[9,] <- c(200809, 20081120, 200712, 200509)
design[10,] <- c(200810, 20081220, 200801, 200510)
design[11,] <- c(200811, 20090120, 200802, 200511)
design[12,] <- c(200812, 20090220, 200803, 200512)
design[13,] <- c(200901, 20090320, 200804, 200601)
design[14,] <- c(200902, 20090420, 200805, 200602)
design[15,] <- c(200903, 20090520, 200806, 200603)
design[16,] <- c(200904, 20090620, 200807, 200604)
design[17,] <- c(200905, 20090720, 200808, 200605)
design[18,] <- c(200906, 20090820, 200809, 200606)
design[19,] <- c(200907, 20090920, 200810, 200607)
design[20,] <- c(200908, 20091020, 200811, 200608)
design[21,] <- c(200909, 20091120, 200812, 200609)
design[22,] <- c(200910, 20091220, 200901, 200610)
design[23,] <- c(200911, 20100120, 200902, 200611)
design[24,] <- c(200912, 20100220, 200903, 200612)
design[25,] <- c(201001, 20100320, 200904, 200701)
design[26,] <- c(201002, 20100420, 200905, 200702)
design[27,] <- c(201003, 20100520, 200906, 200703)
design[28,] <- c(201004, 20100620, 200907, 200704)
design[29,] <- c(201005, 20100720, 200908, 200705)
design[30,] <- c(201006, 20100820, 200909, 200706)

## Viiveet, joilla kunkin kuukauden ennustaminen tehdään
## (2008 oli karkausvuosi)
viiveet <- c(80,80,81,81,81,81,82,81,81,81,81,81,79,79,
81,81,81,81,82,81,81,81,81,81,79,79,81,81,81,81)

## Alustetaan vektorit, joihin tulokset kerätään
logistinen <- NULL
cox <- NULL
sd <- NULL
vertailumalli <- NULL
havaittu <- NULL
toteutunut <- NULL

## Suoritetaan testi
for(i in 1:30){

## Haetaan tuotantohetkellä tiedossa oleva ennustettavan kuukauden
## data
pvm1 <- design[i,1] # tutkittava kuukausi
```

```
pvm2 <- design[i,2] # tarkasteluajankohta
pvm3 <- design[i,3] # estimointidatan yläpvm
pvm4 <- design[i,4] # estimointidata alapvm

## Haetaan viive, jolla ennustettavan kuukaden tieto tuotetaan
viive_ajo <- viiveet[i]

## Ennustettavan kuukauden kaikki pvm2:een mennessä rekisteriin
## tulleet havainnot
kk_data <- tiedot[tiedot$apvm == pvm1 & tiedot$alorekpvm <= pvm2,]

## estimointiin käytettävä data
est_data <- tiedot[tiedot$apvm <= pvm3 & tiedot$apvm >= pvm4,]

### Coxin regressio
est_data_cox <- est_data

## tehdään sensurointi
est_data_cox$viive[est_data_cox$viive > 83] <- 83
est_data_cox$sens <- 1*(est_data_cox$viive <= 83)

##sovitetaan malli
fit1 <- coxph(Surv(viive, sens) ~ factor(akk) + factor(kuntalaji) +
  factor(ktark) + logtilav, data = est_data_cox)

## haetaan korotuskertoimet mallin ennusteista
s1 <- survfit(fit1,newdata = data.frame(
  logtilav = kk_data$logtilav,
  akk = kk_data$akk,
  kuntalaji = kk_data$kuntalaji,
  ktark =  kk_data$ktark))

s1 <- cbind(basehaz(fit1)["time"], s1$surv)
k <- s1[,1] - viive_ajo
apu <- min(k[k >= 0])
korotus_cox <- 1 - t(s1[s1[,1] - viive_ajo == apu,-1])

### Logistinen regressio

## Luodaan dikotominen vaste logistista regressiota varten
est_data$dico <- 1*(est_data$viive <= viive_ajo)
## Sovitetaan logistinen regressiomalli

fit2 <- glm(dico ~ factor(kuntalaji) + factor(ktark) + factor(akk) +
  logtilav, family = binomial('logit'), data = est_data)

## Haetaan korotettavan kuukauden havainnoille sisältymis-
## todennäköisyydet logistisen regressiomallin estimoiduista
## parametreista
korotus_log <- predict(fit2,newdata = data.frame(
  logtilav = kk_data$logtilav,
```

```
  kuntalaji = kk_data$kuntalaji,
  akk = kk_data$akk,
  ktark =  kk_data$ktark),type="response")

### 3 vuoden keskimääräiseen alipeittoon perustuva estimaatti

## Edellisen vuoden tilastoitava kuukausi ja tilaston tuotantopäivä
v1 <- pvm1 - 100
u1 <- pvm2 - 10000
## Tilastoitava kuukausi ja tilaston tuotantopäivä 2 vuotta sitten
v2 <- pvm1 - 200
u2 <- pvm2 - 20000
## Tilastoitava kuukausi ja tilaston tuotantopäivä 3 vuotta sitten
v3 <- pvm1 - 300
u3 <- pvm2 - 30000

kk1_ennakkosumma <- aggregate(tiedot[tiedot$apvm == v1 &
  tiedot$alorekpvm <= u1,]$tilav,
  by=list(factor(tiedot[tiedot$apvm == v1 &
  tiedot$alorekpvm <= u1,]$ktark)), FUN = sum)[,2]

kk2_ennakkosumma <- aggregate(tiedot[tiedot$apvm == v2 &
  tiedot$alorekpvm <= u2,]$tilav,
  by=list(factor(tiedot[tiedot$apvm == v2 &
  tiedot$alorekpvm <= u2,]$ktark)), FUN = sum)[,2]

kk3_ennakkosumma <- aggregate(tiedot[tiedot$apvm == v3 &
  tiedot$alorekpvm <= u3,]$tilav,
  by=list(factor(tiedot[tiedot$apvm == v3 &
  tiedot$alorekpvm <= u3,]$ktark)) , FUN = sum)[,2]

kk1_summa <- aggregate(tiedot[tiedot$apvm == v1,]$tilav
  , by=list(factor(tiedot[tiedot$apvm == v1,]$ktark))
  , FUN = sum)[,2]
kk2_summa <- aggregate(tiedot[tiedot$apvm == v2,]$tilav
  , by=list(factor(tiedot[tiedot$apvm == v2,]$ktark))
  , FUN = sum)[,2]
kk3_summa <- aggregate(tiedot[tiedot$apvm == v3,]$tilav
  , by=list(factor(tiedot[tiedot$apvm == v3,]$ktark))
  , FUN = sum)[,2]

# 3 vuoden keskiarvo alipeitolle käyttötarkoitusluokittain
apeitto_ka <- apply(cbind((kk1_summa - kk1_ennakkosumma)/kk1_summa,
  (kk2_summa - kk2_ennakkosumma)/kk2_summa,
  (kk3_summa - kk3_ennakkosumma)/kk3_summa), 1, FUN = mean)
  k_kerroin1 <- 1/(1 - apeitto_ka)

### Tulokset

## havaitut aloitukset
havaittu <- c(havaittu,sum(kk_data$tilav))
```

```
## Coxin regressiolla estimoidut aloitukset
cox <- c(cox, sum(kk_data$tilav/(korotus_cox)))

## Logistisella regressiolla estimoidut aloitukset
logistinen <- c(logistinen, sum(kk_data$tilav/(korotus_log)))
sd <-  c(sd, sqrt(sum((1 - korotus_log)*kk_data$tilav^2/korotus_log^2 )))

## Nykyisellä korotusmallilla estimoidut aloitukset
vertailumalli <- c(vertailumalli, sum(aggregate(kk_data$tilav,
by = list(factor(kk_data$ktark)),
FUN = sum)[,2]*k_kerroin1))

# Toteutunut aloitusten kokonaistilavuus
kk_data2 <- tiedot[tiedot$apvm == pvm1,]
toteutunut <- c(toteutunut, sum(kk_data2$tilav))
}


### Tulosten vertailua
mean(abs(logistinen - toteutunut))
mean(abs(cox - toteutunut))
mean(abs(vertailumalli - toteutunut))

median(abs(logistinen - toteutunut))
median(abs(cox - toteutunut))
median(abs(vertailumalli - toteutunut))

median(abs((logistinen - toteutunut)/ toteutunut))
median(abs((cox- toteutunut)/ toteutunut))
median(abs((vertailumalli - toteutunut)/ toteutunut))

mean(abs((logistinen - toteutunut)/ toteutunut))
mean(abs((cox - toteutunut)/ toteutunut))
mean(abs((vertailumalli - toteutunut)/ toteutunut))
```