Systemaattisen revision eliminointi palkkasummien suhdannekuvasta logistista regressiomallia soveltaen

Godfrey M. Lowndes

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
9.10.2014

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Lowndes Godfrey: Systemaattisen revision eliminointi palkkasummien suhdannekuvasta logistista regressiomallia soveltaen
Tilastotieteen pro gradu -tutkielma, 30 sivua
9.10.2014

TIIVISTELMÄ

Tilastokeskus julkaisee kuukausittain palkkasummien viimeisestä kehityksestä kertovan suhdannekuvan. Viimeistä kehitystä estimoidaan muutoksella edellisvuodesta. Estimointia varten verohallinnolta saadaan kausiveroaineisto.

Aineiston puuttuvien havaintojen käsittelytavan on huomattu aiheuttavan palkkasummien muutoksiin pientä systemaattista harhaa ylöspäin tuoreilla kuukausilla. Nykyisellään puuttuvat havainnot poistetaan, kunnes yritykseltä on jäänyt palkat ilmoittamatta neljältä peräkkäiseltä kuukaudelta, ja siten yrityksen lopetuksesta seuraava palkkasummien lasku näkyy viiveellä palkkasummien kehityksessä.

Mallintamalla logistisella regressiolla todennäköisyyttä, onko puuttuva havainto seurausta yrityksen lopetuksesta tai huonosta taloudellisesta tilanteesta eikä aineistovirheestä, voidaan tilastollisesti päätellä, mitkä puuttuvat havainnot tulee poistaa aineistovirheinä. Mallinnuksessa hyödynnettiin tutkimusaineistoa, jossa oli noin 27 000 havaintoa. Mallin tärkeimmiksi muuttujiksi osoittautuivat puuttuvaa havaintoa edeltävä palkkatieto sekä puuttuvien havaintojen lukumäärä.

Mallin tulokset ja tilastotuotanto-ohjelmien rajoitukset huomioonottaen simuloitiin vuoden 2012 ja 2013 alkuvuoden palkkasummien kehitys uudelleen ottamalla puuttuvat havainnot laskentaan mukaan yrityksiltä, joiden viimeinen ilmoitettu palkkatieto oli alle 100 000 ja virhetodennäköisyys mallin mukaan alle 0,1. Simulointikokeesta nähdään, miten puuttuvien havaintojen käsittelytapaa muuttamalla saadaan harha eliminoitua palkkasummien suhdannekuvasta.

Avainsanat: harha, logistinen regressiomalli, palkkasumma, puuttuva havainto, revisio, suhdannekuva

Sisältö

1 Johdanto

Tilastokeskuksen Yritystilastot-yksikkö julkaisee kuukausittain koko talouden palkkasumman ja eri toimialojen palkkasummien viimeisestä kehityksestä kertovan suhdannekuvan. Koko talouden ja eri toimialojen palkkasummien kehitys estimoidaan muutoksella edellisvuodesta käyttäen kuukausittain verohallinnosta saatavaa kausiveroaineistoa ja Tilastokeskuksen Yritysrekisterin toimialatietoja.

Verohallinnolta saatu palkka-aineisto täydentyy vuoden ajan. Aineiston täydentymisestä johtuen Tilastokeskuksessa on käytäntönä julkaista, miten jo julkaistut vuosimuutokset muuttuvat aineiston täydentyessä. Aineiston täydentymisestä johtuvaa vuosimuutoksessa tapahtuvaa tarkentumista kutsutaan yleisesti nimellä *revisio*. Koko talouden palkkasummien vuosimuutosten on huomattu tarkentuvan aina alaspäin aineiston täydentyessä. *Systemaattiseksi revisioksi* kutsutun ilmiön on esitetty olevan seurausta tavasta tulkita yrityksen toiminta loppuneeksi puuttuvien palkkahavaintojen perusteella.

Tutkimuksen tavoitteena on etsiä keino korjata koko talouden palkkasummien vuosimuutoksissa esiintyvä systemaattinen revisio. Tavoitteena on, että korjaussääntö ottaa huomioon tuotanto-ohjelmien rajoitukset eikä pitkitä ohjelmien ajoaikaa. Ratkaisussa mallinnetaan logistisella regressiomallilla todennäköisyyttä, että puuttuva havainto on seurausta yrityksen lopetuksesta tai huonosta taloudellisesta tilasta eikä puutteellisesta aineistosta. Mallinnuksessa hyödynnetään peräkkäisten puuttuvien havaintojen lukumäärää, puuttuvaa havaintoa edeltävää palkkatietoa, sen toista potenssia, havaintoneljännestä sekä yrityksen toimialan liikevaihdon kehitystä. Mallin tulosten

avulla ja tuotanto-ohjelmien rajoitukset huomioonottaen yrityksen puuttuva havainto luokitellaan joko aineistovirheeksi tai seuraukseksi yrityksen lopetuksesta. Suodattamalla aineistosta oikeassa suhteessa aineistovirheet ja päästäen yrityksen lopetuksesta johtuvat puuttuvat havainnot palkkasummien muutoksien laskentaan saadaan simulointitulosten perusteella systemaattinen revisio eliminoitua palkkasummien suhdannekuvasta.

Tutkinnon rakenne on seuraava. Luvussa 2 kerrotaan tutkimusongelmasta ja mallinnukseen käytetystä aineistosta. Työn luvussa 3 käydään läpi käytetyn logistisen regressiomallin, parametrien estimoinnin sekä mallin diagnostiikan ja tulkinnan yleinen teoria. Luvussa 4 esitellään mallin tulokset sekä niiden ja simulointitulosten perusteella muodostettu uusi käsittelytapa puuttuville palkkahavainnoille systemaattisen revision eliminoimiseksi. Luvussa 5 on yhteenveto työn tavoitteista ja tuloksista.

2 Aineisto ja tutkimusongelma

2.1 Palkkasummien suhdannekuva ja revisio

Tilastokeskus julkaisee palkkasummien viimeisestä kehityksestä kertovan suhdannekuvan kuukausittain. Tiedot julkaistaan noin 40 päivän viiveellä tilastoitavasta ajankohdasta. Julkaisu kertoo koko talouden sekä eri toimialojen kokonaispalkkasummien kehityksestä. Julkaisun palkka-aineistona käytetään verohallinnolta saatavaa kausiveroaineistoa, jota täydennetään tiedonkeruutiedoilla suurimpien yritysten palkkatietojen osalta. Lisäksi julkaisun toimialat määräytyvät Tilastokeskuksen Yritys- ja toimipaikkarekisterin vuoden 2008 toimialaluokituksen mukaisesti. Palkkasummien kehityksestä julkaistaan kuukausittain vuosimuutos, joka saadaan tilastoitavan kuukauden ja vertailuvuoden vastaavan kuukauden kokonaispalkkasummien suhteesta

$$M_t = \frac{\sum_{i=1}^{N} Y_{it} - \sum_{i=1}^{N} Y_{it-12}}{\sum_{i=1}^{N} Y_{it-12}}, \tag{1}$$

missä Y_{it} on yrityksen i palkkasumma tilastoitavalta kuukaudelta t ja Y_{it-12} on yrityksen i vertailuvuoden palkkasumma. Käytettävää palkka-aineistoa voidaan pitää kokonaisaineistona Suomessa toimivien yritysten maksamista palkoista. Palkkasumma lasketaan summaamalla yritysten maksamat bruttomääräiset palkat. Palkkasummiin lasketaan kaikki tuloverojen ja sosiaaliturvamaksujen alaiset palkat sekä erilaisista lisätöistä, bonuksista ja lomarahoista koituvat kulut. Palkkasummiin ei lasketa lähdeveron alaisia palkkoja, yrityksien maksamia optioita, työntekoon liittyviä kuluja eikä työnantajalta perittäviä sosiaaliturvamaksuja [1]. Taulukossa 1 on maaliskuussa 2013 julkaistut vuoden 2013 tammikuun vuosimuutokset päätoimialoittain.

Taulukko 1: Maaliskuussa 2013 julkaistut tammikuun palkkasummien vuosimuutokset päätoimialoittain.

Toimiala	M_t %
Koko talous	1,5
Teollisuus	-3,2
Rakentaminen	2,5
Kauppa	3,2
Palvelut	3,0

Koko talouden palkkasumma kasvoi tammikuussa 2013 1,5 prosenttia edellisvuodesta [2]. Julkaisun jälkeen vuosimuutosten laskentaan käytetty palkka-aineisto täydentyy virheellisten ja puuttuvien havaintojen osalta. Aineiston täydentyessä myös edelliskuukausien muutokset M_t lasketaan uudelleen.

Olkoon k alaindeksi, joka kertoo kuinka monta kuukautta kuukauden t aineistoa on täydennetty. Tarkentuneella aineistolla laskettujen vuosimuutosten M_{tk} ja ensimmäisen julkaisun vastaavan kuukauden vuosimuutoksen M_{t0} välistä erotusta

$$R_{tk} = M_{tk} - M_{t0}, \tag{2}$$

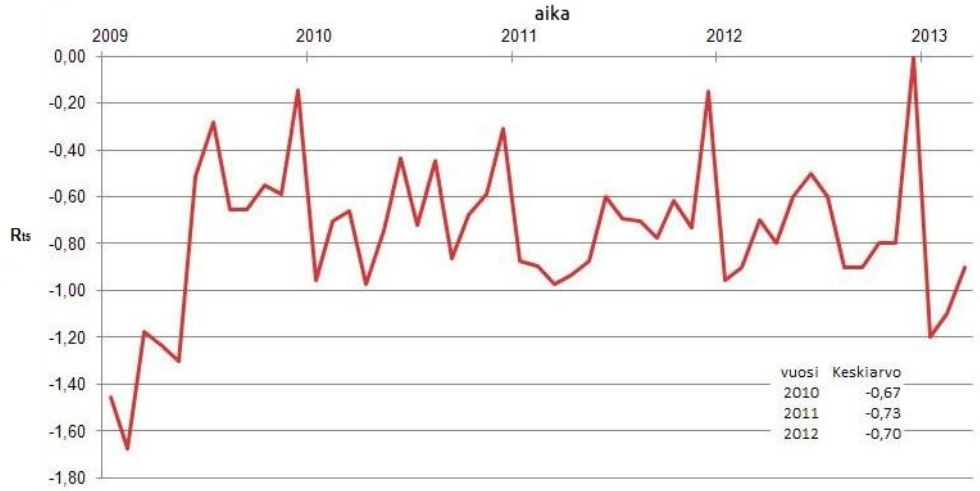
missä k voi saada arvoja $0, 1, 2, ..., 12$, kutsutaan revisioksi. Nyt revisio R_{tk} kuvaa aineiston tarkentumisesta johtuvaa vuosimuutoksessa tapahtuvaa muutosta. Aineisto täydentyy vuoden ajan, mutta aineiston voidaan katsoa olevan kokonaan kertynyt viiden kuukauden jälkeen. Taulukossa 2 on tammikuun 2013 vuosimuutos ensimmäisestä julkaisusta M_{t0}, tammikuun vuosimuutos viiden kuukauden tarkentumisen jälkeen M_{t5} sekä näistä laskettu revisio R_{t5} [3].

Taulukko 2: Vuoden 2013 tammikuun palkkasummien vuosimuutokset päätoimialoittain ensimmäiseltä M_{t0} ja viidenneltä julkaisukerralta M_{t5} sekä näistä laskettu revisio R_{t5}.

Toimiala	M_{t0} %	M_{t5} %	R_{t5} %-yks.
Koko talous	1,5	0,3	-1,2
Teollisuus	-3,2	-3,8	-0,6
Rakentaminen	2,5	-0,3	-2,8
Kauppa	3,2	2,6	-0,6
Palvelut	3,0	1,9	-1,1

Taulukosta 2 nähdään, miten aineiston täydentymisen jälkeen koko talouden palkkasummien tammikuun 2013 vuosimuutos on tarkentunut alaspäin 1,2 prosenttiyksikköä ensimmäisestä julkistuskerrasta. Taulukosta 2 nähdään myös, miten päätoimialojen vuosimuutokset ovat kaikki tarkentuneet alaspäin. Revisioita on laskettu vuodesta 2009 alkaen eikä koko talouden palkkasumman lopullinen revisio R_{t5} ole ollut kertaakaan positiivinen. Kuvassa 1 on koko talouden palkkasummien revisiot R_{t5} vuodesta 2009 alkaen.

Kuvasta 1 nähdään, miten koko talouden palkkasummien vuosimuutokset tarkentuvat systemaattisesti alaspäin. Esimerkiksi vuonna 2012 ensimmäisen ja viidennen kerran välinen vuosimuutoksessa tapahtuva tarkentuminen oli tammikuussa noin -0,2 prosenttiyksikköä ja koko vuonna keskimäärin -0,7 prosenttiyksikköä. Kuvan perusteella voidaan sanoa, että ensimmäisen kierroksen vuosimuutokset näyttävät olevan ylöspäin harhaisia. Tutkielman tavoitteena on korjata tämä pieni harha palkkasummien suhdannekuvassa.



Kuva 1: Koko talouden palkkasumman vuosimuutoksien viidennen julkaisukerran revisiot R_{t5} prosenttiyksikköinä vuodesta 2009 alkaen.

2.2 Systemaattinen revisio

Revisiota käytetään eräänlaisena tilastollisen virheen tunnuslukuna. Tästä syystä vuosimuutosten revision tulisi olla odotusarvoisesti nolla tilanteessa, missä vuosimuutos on harhaton estimaattori palkkasummien kehityksestä. Koska koko talouden palkkasumman vuosimuutoksen revision odotusarvo on negatiivinen, ensimmäisen julkaisun vuosimuutokset ovat ylöspäin harhaisia. Tämän on esitetty olevan seurausta puuttuvien havaintojen käsittelystä palkka-aineistossa.

Palkkasummien vuosimuutoksella halutaan kuvata eri toimialojen ja koko talouden palkkasummien kehitystä edelliseen vuoteen verrattuna. Tästä syystä tulee vuosimuutoksen laskentaan ottaa mukaan vain vertailukelpoiset palkkatiedot aineiston yrityksiltä kuluvalta ja vertailuvuodelta. Kahden vuoden palkkasummat eivät ole vertailukelpoisia, jos yrityksen tapa raportoida palk-

koja on muuttunut vuoden takaisesta, palkkatieto on virheellinen tai yritykseltä on jäänyt palkkatieto raportoimatta.

Lisätään kaavaan (1) indikaattorimuuttuja v_{it}, joka kertoo, ovatko yrityksen i palkkatiedot Y_{it} ja Y_{it-12} vertailukelpoiset. Indikaattori v_{it} saa arvon yksi, kun yrityksen palkkatiedot Y_{it} ja Y_{it-12} ovat vertailukelpoiset ja muuten arvon nolla. Nyt kaavan (1) suhde M_t voidaan esittää muodossa

$$M_t = \frac{\sum_{i=1}^{N} v_{it} Y_{it} - \sum_{i=1}^{N} v_{it} Y_{it-12}}{\sum_{i=1}^{N} v_{it} Y_{it-12}}, \tag{3}$$

jolloin suhde (3) kertoo yritysten maksamien palkkojen muutoksesta ilman raportointikäytänteiden muutoksien, aineistovirheiden tai puuttuvien havaintojen vaikutusta. Aineistossa puuttuva havainto ja nolla ovat sama asia.

Nykyisellään yrityksen i puuttuva havainto Y_{it} poistetaan palkka-aineistosta ($v_{it} = 0$), jos peräkkäisiä puuttuvia havaintoja on alle viisi. Tämä sen takia, ettei aineistovirheiden vaikutus näkyisi palkkasummien suhdannekuvassa. Poistetut havainnot Y_{it} palautetaan laskentaan ($v_{it} = 1$), kun yritykseltä on jäänyt palkat raportoimatta neljältä peräkkäiseltä kuukaudelta ja yrityksen toiminta todetaan oikeasti loppuneeksi. Puuttuvien palkkahavaintojen ja niitä vastaavien vertailuarvojen liiallinen poistaminen palkka-aineistosta ja osan palauttaminen neljän kuukauden viiveellä aiheuttaa vuosimuutoksissa harhaa ylöspäin tuoreilla kuukausilla. Puuttuvien havaintojen käsittely on haastavaa, koska yrityksen i puuttuva havainto tulee olla mukana aineistossa ($v_{it} = 1$), jos se on seurausta yrityksen toiminnan loppumisesta. Toisaalta jos puuttuva havainto on aineistovirhe, tulee se poistaa aineistosta ($v_{it} = 0$).

Tavoitteena on parantaa puuttuvien havaintojen käsittelytapaa luokittelemalla ne aineistovirheiksi tai yrityksen lopetuksiksi jo ensimmäisen puuttuvan havainnon pohjalta. Tätä varten tulee palkka-aineiston tietoja hyväksi käyttäen estimoida todennäköisyys, että puuttuva havainto on aineistovirhe eikä seurausta yrityksen toiminnan loppumisesta tai huonosta taloudellisesta tilanteesta. Todennäköisyyden estimointiin käytetään yleistettyä lineaarista mallia, joka tässä tapauksessa on logistinen regressiomalli. Mallin vaste saa arvon 1 (*aito*), kun puuttuva tieto on seurausta yrityksen lopetuksesta, ja se tulee ottaa mukaan muutoksen laskentaan ($v_{it} = 1$) ja toisaalta arvon 0 (*virhe*), kun puuttuva havainto on aineistovirhe ja se tulee poistaa muutoksen laskennasta ($v_{it} = 0$).

2.3 Tutkimusaineisto ja mallin muuttujat

Tutkimusta varten koottiin kausiveroaineiston historiatiedoista sekä Yritys- ja toimipaikkarekisteristä tutkimusaineisto, jossa on 26 739 havaintoa. Näistä 26 156 sai arvon '*aito*' ja loput 583 arvon '*virhe*'. Valtaosa puuttuvista havainnoista näyttäisi olevan siten aitoja ja ne tulisi olla mukana vuosimuutosten laskennassa.

Aineistoon sovitettavan mallin vaste saadaan poimimalla talteen kunkin kuukauden palkka-aineistosta sen tuoreen kuukauden puuttuvat havainnot ehdolla, että puuttuvaa havaintoa edeltävä havainto oli ei-puuttuva. Vertaamalla kyseisiä havaintoja tietokannan nykytilanteeseen saadaan tieto, onko puuttuva tieto korjaantunut aineiston täydentyessä vai ei. Mallin vaste saa arvon

nolla, '*virhe*', kun puuttuva havainto on aineistovirhe ja muutoin vaste saa arvon yksi, '*aito*'. Vasteen arvot ovat toisistaan riippumattomia, koska aineistossa kultakin yritykseltä on vain yksi vasteen arvo.

Tutkimusaineistossa on 18 selittävää muuttujaa: peräkkäisten puuttuvien havaintojen lukumäärä (*lkm*), puuttuvaa havaintoa edeltävä palkka sadoissatuhansissa euroissa (*palkka*), samaisen palkkatiedon toinen potenssi sadoissamiljoonissa euroissa (*palkka2*), vuosineljännes, jona puuttuva havainto tuli (*kausi*) sekä Yritys -ja toimipaikkarekisteristä poimittu yrityksen vuosiliikevaihto miljoonissa euroissa (*kokoluokka_lv*) ja toimiala (*toimiala*). Näiden lisäksi aineistossa on suhdannetilanteen vaikutusta vasteeseen selittävät muuttujat, joita ovat puuttuvaa havaintoa edeltävän kuuden kuukauden palkkasummien sekä liikevaihtojen kehitys edellisvuodesta yrityksen toimialalla (*pa1*, *pa2*, *pa3*, *pa4*, *pa5* ja *pa6* sekä *lv1*, *lv2*, *lv2*, *lv3*,*lv4*, *lv5* ja *lv6*).

Luvussa 3 määritellään logistinen regressiomalli yhtenä yleisten lineaaristen mallien erikoistapauksena, parametrien estimointiin käytetty suurimman uskottavuuden menetelmä sekä mallin diagnostiikan ja tulkinnan yleinen teoria.

3 Yleinen teoria

3.1 Eksponentiaalinen perhe ja yleistetyt lineaariset mallit

Määritellään aluksi *eksponentiaalinen perhe* satunnaismuuttujan Y avulla. Satunnaismuuttujan Y todennäköisyysjakauma on osa eksponentiaalista perhettä, jos sen tiheysfunktio voidaan esittää muodossa

$$f(y;\theta) = s(y)t(\theta)\exp(a(y)b(\theta)), \tag{4}$$

missä funktiot s, t, a ja b ovat tunnettuja funktioita. Tiheysfunktio (4) voidaan esittää muodossa

$$f(y;\theta) = \exp(a(y)b(\theta) + c(\theta) + d(y)), \tag{5}$$

kun $s(y) = \exp(d(y))$ ja $t(\theta) = \exp(c(\theta))$ [4]. Määritellään seuraavaksi yleistetty lineaarinen malli, kun vastemuuttujan todennäköisyysjauma on osa eksponentiaalista perhettä.

Oletetaan satunnaismuuttujat Y_i riippumattomiksi ja jakautuneiksi parametreilla θ_i, $i = 1, ..., N$. Oletetaan lisäksi, että satunnaismuuttujien jakauma on osa eksponentiaalista perhettä. Nyt satunnaismuuttujien $Y_1, ..., Y_N$ yhteistiheysfunktio

$$f(y_1, ..., y_N; \theta_1, ..., \theta_N) = \prod_{i=1}^{N} \exp(a(y_i)b(\theta_i) + c(\theta_i) + d(y_i)) \quad (6)$$

$$= \exp(b(\theta_i) \sum_{i=1}^{N} a(y_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)). \quad (7)$$

Tiheysfunktion (6) sanotaan olevan kanonisessa muodossa, kun $a(y_i) = y_i$. Funktio $a(y_i)$ on parametrin θ_i tyhjentävä tunnusluku ja funktio $b(\theta_i)$ sen luonnollinen parametri, joka on jokin odotusarvon $E(Y_i) = \mu_i$ muunnos.

Määritellään seuraavaksi parametrit $\beta_1, ..., \beta_p$, missä $p < N$ siten, että niiden lineaarikombinaatio on odotusarvon $E(Y_i) = \mu_i$ monotoninen funktio

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (8)$$

jossa $\mathbf{x}_i$ on havainnon Y_i selittävät muuttujat sisältävä $(p \times 1)$-vektori ja $\boldsymbol{\beta}$ estimoitavat regressiokertoimet sisältävä $(p \times 1)$-vektori. Yhtälön (8) funktiota $g(\mu_i)$ kutsutaan yleisesti *linkkifunktioksi*. Koska luonnollinen parametri $b(\theta_i)$ on odotusarvon $E(y_i)$ funktio kaikilla $i = 1, ..., N$, voidaan lineaarikombinaatio $\mathbf{x}_i^T \boldsymbol{\beta}$ esittää luonnollisen parametrin $b(\theta_i)$ avulla siten, että

$$b(\theta_i) = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (9)$$

kun satunnaismuuttujan Y_i todennäköisyysjakauma on osa eksponentiaalista perhettä [4]. Edellä määriteltyä yleistettyä lineaarista mallia hyödyntäen luvussa 3.2 määritellään logistinen regressiomalli, kun vastemuuttuja on kaksitasoinen eli dikominen.

3.2 Logistinen regressiomalli

Vastemuuttujan ollessa kaksiluokkainen ei voida käyttää tavallista regressiomallia, missä vastemuuttujan odotusarvoa μ mallinnetaan lineaarisella funktiolla $\mathbf{x}^T\boldsymbol{\beta}$. Tavallisen regressiomallin sijaan käytetään logistista regressiomallia. Määritellään logistinen regressiomalli satunnaismuuttujan Y_i avulla siten, että

$$Y_i = \begin{cases} 1, & \text{mittauksen ollessa tosi,} \\ 0, & \text{muuten.} \end{cases}$$

Olkoon lisäksi $P(Y_i = 0) = \pi_i$ ja $P(Y_i = 1) = 1 - \pi_i$ kaikilla $i = 1, ..., N$. Kun $Y_i \sim Bin(n, \pi_i)$, on satunnaismuuttujan Y_i tiheysfunktio

$$f(y_i; \pi_i) = \binom{n}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n - y_i}. \tag{10}$$

Olettamalla satunnaismuuttujat Y_i riippumattomiksi niiden yhteistiheysfunktio

$$f(y_1, ..., y_N; \pi_1, ..., \pi_N) = \prod_{i=1}^{N} \binom{n}{y_i} \pi_i^{y_i} (n - \pi_i)^{1 - y_i}. \tag{11}$$

Näytetään seuraavaksi, että binomijakauma on osa eksponentiaalista perhettä. Esitetään tiheysfunktio (11) eksponentiaaliperheen yleisessä muodossa (7). Käyttämällä potenssiin korotuksen ja logaritmin laskusääntöjä voidaan tiheysfunktio (11) esittää muodossa

$$f(y_1, ..., y_N; \pi_1, ..., \pi_N) = \exp\left[\sum_{i=1}^{N} y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \sum_{i=1}^{N} \log(1 - \pi_i) + \sum_{i=1}^{N} \log \binom{n}{y_i}\right]. \tag{12}$$

Tiheysfunktio on osa eksponentiaalista perhettä, kun yhtälön (7) funktiot nimetään siten, että $a(y_i) = \sum_{i=1}^{N} y_i$, $b(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$, $c(\pi_i) = \log(1-\pi_i)$ ja $d(y_i) = \sum_{i=1}^{N} \log \binom{n}{y_i}$. Lisäksi nähdään, että tiheysfunktio on kanonisessa muodossa, $\log(\frac{\pi_i}{1-\pi_i})$ on luonnollinen parametri ja $\sum_{i=1}^{N} y_i$ sen tyhjentävä tunnusluku. Nyt kaavan (9) lineaarikombinaatio $\mathbf{x}_i^T\boldsymbol{\beta}$ voidaan esittää luonnollisen parametrin $\log(\frac{\pi_i}{1-\pi_i})$ avulla siten, että

$$g(\mu_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}. \tag{13}$$

Yhtälöstä (13) voi ratkaista todennäköisyyden

$$P(Y_i = 0) = \pi_i = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}. \tag{14}$$

3.3 Suurimman uskottavuuden menetelmä

Olkoon satunnaismuuttujien $Y_1, ..., Y_N$ yhteystiheysfunktio

$$f(y_1, ..., y_N; \theta_1, \ldots, \theta_p). \tag{15}$$

Merkitään muuttujia $y_1, ..., y_N$ vektorilla $\mathbf{y}$ ja parametreja $\theta_1, ..., \theta_p$ vektorilla $\boldsymbol{\theta}$. Kun $\boldsymbol{\Omega}$ on kaikki vektorin $\boldsymbol{\theta}$ mahdolliset arvot sisältävä parametriavaruus, parametrivektorin $\boldsymbol{\theta}$ suurimman uskottavuuden estimaatti $\hat{\boldsymbol{\theta}}$ maksimoi uskottavuusfunktion $L(\boldsymbol{\theta}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})$ siten, että

$$L(\hat{\boldsymbol{\theta}}; \mathbf{y}) \geq L(\boldsymbol{\theta}; \mathbf{y}) \tag{16}$$

kaikilla $\boldsymbol{\theta}$ kuuluu parametriavaruuteen $\boldsymbol{\Omega}$. Koska logaritmifunktio on aidosti kasvava, maksimoi suurimman uskottavuuden estimaatti $\hat{\boldsymbol{\theta}}$ logaritmisen uskottavuusfunktion $l(\mathbf{y}; \boldsymbol{\theta}) = \log(L(\mathbf{y}; \boldsymbol{\theta}))$ siten, että

$$l(\hat{\boldsymbol{\theta}}; \mathbf{y}) \geq l(\boldsymbol{\theta}; \mathbf{y}), \tag{17}$$

jossa $\boldsymbol{\theta}$ kuuluu parametriavaruuteen $\boldsymbol{\Omega}$. Suurimman uskottavuuden estimaatti $\hat{\boldsymbol{\theta}}$ saadaan ratkaisemalla uskottavuusyhtälö

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_j} = 0 \tag{18}$$

kaikilla $j = 1, \ldots, p$. Koska yhtälöryhmä (18) on yleensä epälineaarinen, tulee sen ratkaisu etsiä numeerisesti. Työssä käytetty SAS Instituten tarjoama proseduuri LOGISTIC käyttää parametrien estimoimiseen Newton-Raphson-algoritmia [7]. Algoritmille annetaan alkuarvausvektori $\boldsymbol{\theta}^{(0)}$. Seuraava arvo $\boldsymbol{\theta}^{(1)}$ saadaan päivityskaavalla

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \left[\frac{\partial^2 l}{\partial \theta_j \partial \theta_k}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} \left[\frac{\partial l}{\partial \theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} \tag{19}$$

kaikilla $j = 1, ..., p$ ja $k = 1, ..., p$. Newton-Raphson-algoritmin yleinen päivityskaava vektorille $\boldsymbol{\theta}^{(m)}$ on

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} - \left[\frac{\partial^2 l}{\partial \theta_j \partial \theta_k}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m-1)}} \left[\frac{\partial l}{\partial \theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m-1)}}. \tag{20}$$

Yhtälön (20) tulokset maksimoivat logaritmisen uskottavuusfunktion, jos logaritmisen uskottavuusfunktion toisten derivaattojen

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_j \partial \theta_k} \tag{21}$$

muodostama matriisi on negatiivisesti definiitti, kun $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ kaikilla $j = 1, \ldots, p$ ja $k = 1, \ldots, p$. Lisäksi tulee tarkistaa, saavuttaako logaritminen uskottavuusfunktio lokaalin maksimin jossain parametriavaruuden $\boldsymbol{\Omega}$ reunaarvoissa [7]. Tässä tapauksessa suurimman uskottavuuden estimaatti $\hat{\boldsymbol{\theta}}$ on

näistä suurin. Olkoon $g(\boldsymbol{\theta})$ jokin parametrien $\theta_1, \ldots, \theta_p$ aidosti kasvava funktio. Funktion $g(\boldsymbol{\theta})$ suurimman uskottavuuden estimaattori on $g(\hat{\boldsymbol{\theta}})$. Nyt kaavassa (14) esitellyn todennäköisyyden suurimman uskottavuuden estimaatti

$$\hat{\pi}_i = \frac{\exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})}. \tag{22}$$

Luvussa 3.4 määritellään vielä Waldin testi parametrien tilastollisen merkitsevyyden testaamiseksi, parametrien ja niistä laskettujen ristitulosuhteiden luottamusvälit ja kerrotaan, miten parametriestimaattien arvoja tulee tulkita.

3.4 Mallin valinta, diagnostiikka ja tulkinta

Ennen mallin parametrien estimointia tulee valita käytettävä malli. Mallin valintaan käytetään askeltavaa menetelmää (*Stepwise*), joka minimoi logaritmisesta uskottavuusfunktiosta laskettavan *Akaiken informaatiokriteerin*

$$AIC = -2 \log L + 2k, \tag{23}$$

missä k on mallin parametrien lukumäärä. Askeltava mallin valinta aloittaa mallista, jossa on vain vakiotermi, ja lisää malliin selittäviä muuttujia. Jokaisen selittävän muuttujan lisäyksen jälkeen askeltava menetelmä tarkistaa jo selittävät muuttujat ja poistaa mahdollisesti jo lisättyjä muuttujia [7].

Mallin valinnan ja sen parametrien estimoimisen jälkeen tulee selvittää, onko estimoitua parametria vastaavan selittävän muuttujan vaikutus vastemuuttujaan tilastollisesti merkitsevä. Estimaattien $\hat{\beta}$ tilastollisen merkitsevyyden

testaamiseksi suoritetaan estimaateille Waldin testi, jonka nollahypoteesi on, ettei parametria β vastaavan selittävän muuttujan vaikutus vastemuuttujaan ole tilastollisesti merkitsevä. Waldin testin testisuure

$$W = \left(\frac{\hat{\beta}}{SE(\hat{\beta})} \right)^2, \tag{24}$$

joka on asymptoottisesti χ^2-jakautunut yhdellä vapausasteella. Kun $P(\chi^2_{(1)} \geq W) < 0,05$, testin nollahypoteesi hylätään 5 prosentin riskillä [5].

Kun estimaattien tilastollinen merkitsevyys on testattu, voidaan aloittaa mallin antamien tulosten tulkinta. Logistisen regressiomallin parametriestimaattien tulkitsemisen helpottamiseksi tehdään estimaateille $\hat{\beta}$ usein eksponentiaalinen korotus $\exp(\hat{\beta})$, jolla saadaan parametriestimaatin ristitulosuhde. Selittävän muuttujan ollessa välimatka- tai suhdeasteikollinen kertoo ristitulosuhde, kuinka moninkertaiseksi riski vastemuuttujan olla 'virhe' kasvaa selittävän muuttujan yksikön kasvaessa yhdellä. Tutkittavan muuttujan ollessa luokittelu- tai järjestysasteikollinen ristitulosuhde vertaa kutakin selittävän muuttujan luokkaa valittuun selittävän muuttujan verrokkiluokkaan ja kertoo, miten moninkertaiseksi riski vastemuuttujan olla 'virhe' kasvaa.

Olettamalla, että parametriestimaattorit ovat normaalistijakautuneita, saadaan parametreille ja vastaaville ristitulosuhteille luottamusvälit seuraavasti. Parametrin β 95 prosentin luottamusväli on $\hat{\beta} \pm 1,96SE(\hat{\beta})$. Nollan osuessa välille $[\hat{\beta} - 1,96SE(\hat{\beta}), \hat{\beta} + 1,96SE(\hat{\beta})]$ tulkitaan, että estimaatti $\hat{\beta}$ ei ole tilastollisesti merkitsevä. Ristitulosuhteen $\exp(\beta)$ 95 prosentin luottamusväli on $\exp(\hat{\beta} \pm 1,96SE(\hat{\beta}))$. Parametriestimaatti $\hat{\beta}$ ei ole tilastollisesti merkitsevä, jos yksi on luottamusvälillä $[\exp(\hat{\beta} - 1,96SE(\hat{\beta})), \exp(\hat{\beta} + 1,96SE(\hat{\beta}))]$.

Estimaatin $\hat{\pi}_i$ ollessa suurimman uskottavuuden estimaatti todennäköisyydelle $P(Y_i = 0)$ ja $\hat{\mathbf{V}}_\beta$ parametriestimaattoreiden estimoitu kovarianssimatriisi, on $SE(\hat{\pi}_i)$ neliöjuuri neliömuodosta $\mathbf{x}_i^T \hat{\mathbf{V}}_\beta \mathbf{x}_i$ ja estimoitavan todennäköisyyden $P(Y_i = 0) = \pi_i$ 95 prosentin luottamusväli on

$$[\hat{\pi}_i - 1,96SE(\hat{\pi}_i), \hat{\pi}_i + 1,96SE(\hat{\pi}_i)] \tag{25}$$

[7].

Logistisen regressiomallin sopivuutta ja mallin ennustuskykyä voidaan tarkastella luottamusvälien lisäksi Hosmer-Lemeshown testillä, joka perustuu estimoitujen todennäköisyyksien $\hat{\pi}_1, ..., \hat{\pi}_N$ jakamiseen yhtäsuuriin joukkoihin $n_1, ..., n_g$. Joukoille $n_1, ..., n_g$ määrätään raja-arvot k/g, missä $k = 1, ..., g-1$. Esimerkiksi, kun $g = 10$, sisältää joukko n_1 estimoidut todennäköisyydet $\hat{\pi}_i < 0,1$ ja joukko n_{10} estimoidut todennäköisyydet $\hat{\pi}_i \geq 0,9$. Hosmer-Lemeshow testisuure

$$H = \sum_{k=1}^{g} \frac{(o_k - n_k^{'} \tilde{\pi}_k)^2}{n_k^{'} \tilde{\pi}_k (1 - \tilde{\pi}_k)}, \tag{26}$$

missä $n_k^{'}$ on uniikkien kovariaattiyhdistelmien lukumäärä joukossa g ja o_k uniikkeja yhdistelmiä vastaavien vasteen arvojen lukumäärä luokassa g. Testisuureen kaavassa (26) merkintä $\tilde{\pi}_k$ on ryhmään g kuuluvien estimoitujen todennäköisyyksien $\hat{\pi}_i$ keskiarvo. Mallin ollessa oikea ja mallin ennusteiden ollessa hyviä $P(H \geq \chi^2_{g-2}) > 0,05$ [8].

4 Mallin sovitus ja tulosten implementointi tilastotuotantoon

4.1 Mallin sovitus ja tulokset

Olettamalla yritykset i riippumattomiksi saadaan virhetodennäköisyys π_i estimotua sovittamalla aineistoon logistinen regressiomalli

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta} \tag{27}$$

ja ratkaisemalla todennäköisyysestimaatit

$$\hat{\pi}_i = P(yrityksen\ i\ puuttuva\ havainto\ on\ virhe) \tag{28}$$

sekä

$$1-\hat{\pi}_i = P(yrityksen\ i\ puuttuva\ havainto\ on\ aito). \tag{29}$$

Mallin valinta suoritettiin askeltavalla menetelmällä (*Stepwise*), jonka antaman parhaan mallin AIC oli $5691,85$. Valitussa mallissa on seuraavat selittävät muuttujat: yrityksen i puuttuvaa havaintoa edeltävä palkkatieto (*palkka*), sen toinen potenssi (*palkka2*), peräkkäisten puuttuvien havaintojen lukumäärä (*lkm*, arvot: 1, 2 tai 3), havaintoneljännes (*kausi*, arvot: $Q_1,\ Q_2,\ Q_3$ sekä Q_4) ja suhdannevaikutusmuuttuja (*lv3*).

Malli ennustaa puuttuvan havainnon tilaa varsin hyvin ja sopii aineistoon. Estimaateista $\hat{\pi}_i$ lasketun Hosmer-Lemeshown testin p-arvo 0,11 ei anna viitteitä testin nollahypoteesia vastaan. Taulukossa 3 on mallista estimoidut suurimman uskottavuuden estimaatit $\hat{\beta}$, niiden keskihajonnat, Waldin testisuureet, p-arvot ja ristitulosuhteet $\exp(\hat{\beta})$.

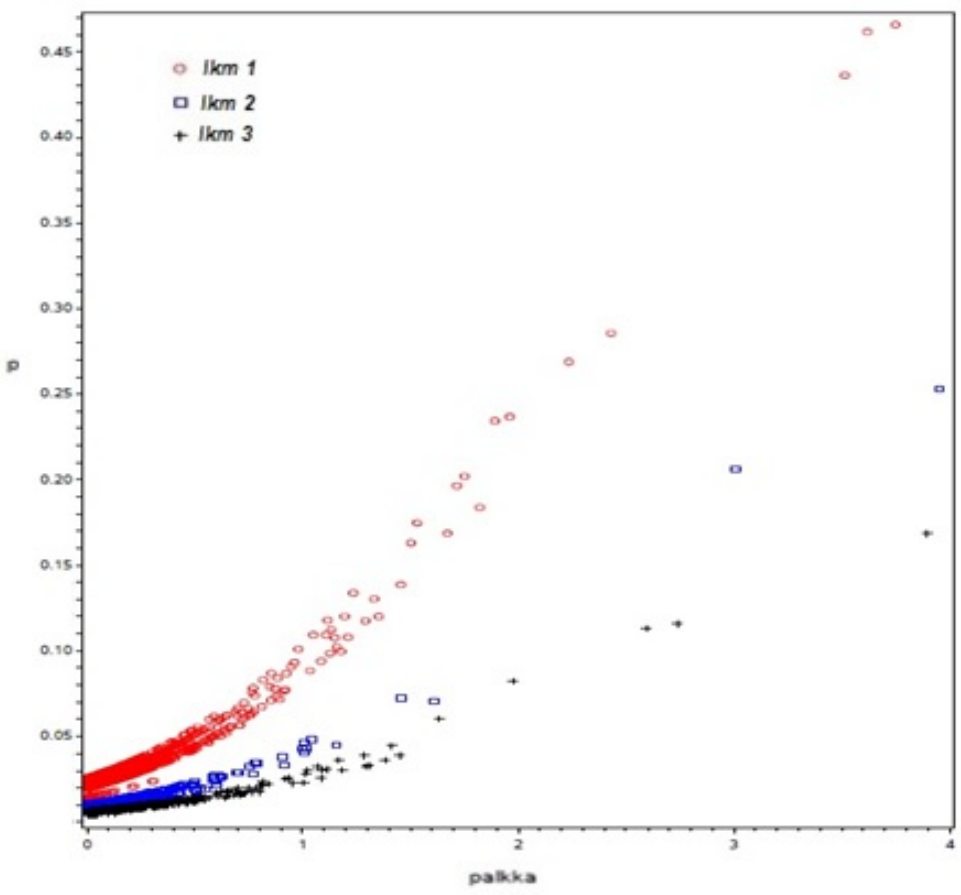
Taulukko 3: Suurimman uskottavuuden estimaatit, keskivirheet, Waldin testisuureiden arvot, p-arvot sekä ristitulosuhteet.

Muuttuja		*df*	$\hat{\beta}$	*s.e*	*W-arvo*	*p-arvo*	$\exp(\hat{\beta})$
Intercept		1	-4.0187	0.0512	6166.6014	<.0001	0.018
lkm	*2*	1	-0.1372	0.0394	12.1099	0.0005	0.872
lkm	*3*	1	-0.5943	0.0453	171.9519	<.0001	0.552
palkka		1	1.6675	0.0980	289.2896	<.0001	5.299
palkka2		1	-0.00189	0.000203	86.7351	<.0001	0.998
kausi	*Q2*	1	-0.0804	0.0366	4.8326	0.0279	0.923
kausi	*Q3*	1	-0.2534	0.0361	49.1940	<.0001	0.776
kausi	*Q4*	1	0.6168	0.0333	342.2289	<.0001	1.853
lv3		1	-0.0181	0.00391	21.4592	<.0001	0.982

Taulukosta 3 nähdään, miten kaikki valitun mallin selittävät muuttujat ovat tilastollisesti merkitseviä ($p < 0,05$). Taulukon 3 ristitulosuhteiden estimaateista nähdään, miten niitä vastaavat muuttujat vaikuttavat todennäköisyyteen, että yrityksen i puuttuva havainto on *virhe*. Puuttuvaa havaintoa edeltävän palkkatiedon vaikutus virhetodennäköisyyteen on estimoidun mallin mukaan suurin. Puuttuvaa havaintoa edeltävän palkkatiedon kasvaessa sadallatuhannella kasvaa virhetodennäköisyys mallin mukaan 5,30-kertaiseksi. Taulukosta nähdään myös, miten ensimmäinen puuttuva havainto on mallin mukaan todennäköisimmin virhe kuin sitä edeltävät. Toisen puuttuvan havainnon todennäköisyys olla virhe on 0,87-kertainen ensimmäiseen nähden ja kolmannen vain 0,55-kertainen. Se, onko puuttuva havainto tullut ensimmäisellä, toisella, kolmannella vai viimeisellä vuosineljänneksellä on tilastollisesti merkitsevä vaikutus virhetodennäköisyyteen. Viimeisellä neljännek-

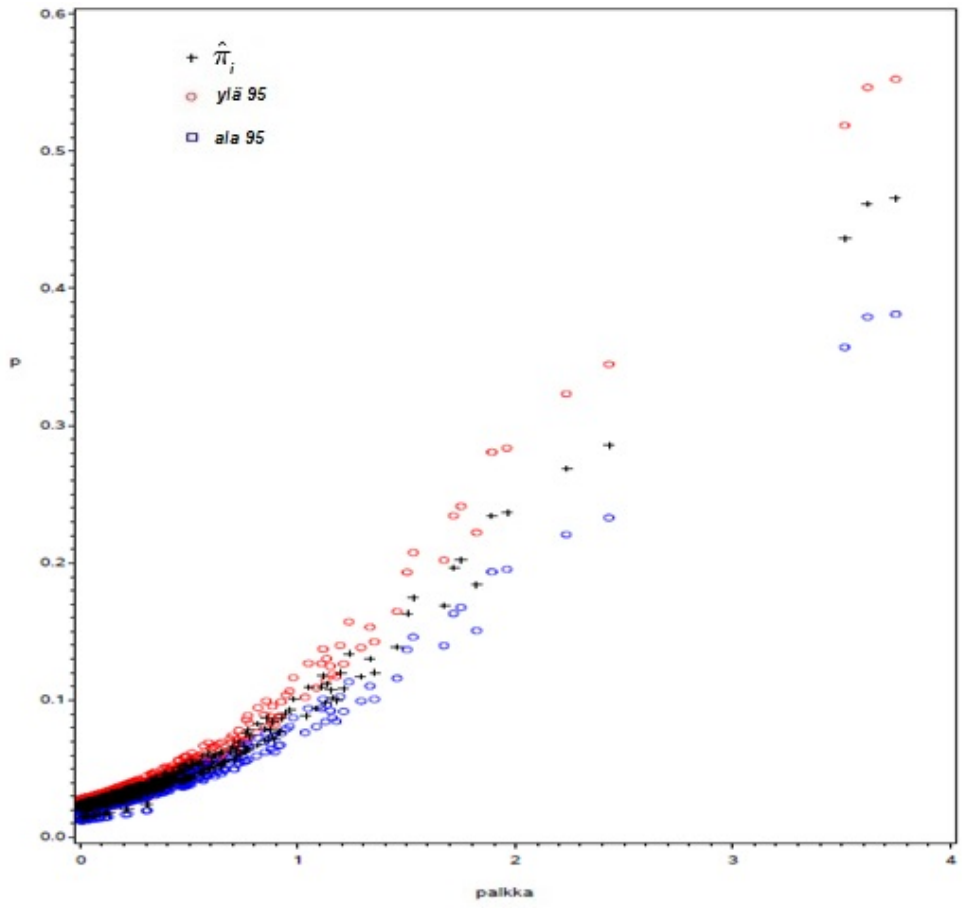
sellä tullut puuttuva havainto on 1,85 kertaa todennäköisemmin virhe kuin ensimmäisellä neljänneksellä tullut puuttuva havainto, kolmannella neljänneksellä tulleella puuttuvalla havainnolla on 0,78-kertainen todennäköisyys olla virhe verrattuna ensimmäiseen neljännekseen ja toisella neljänneksellä 0.92-kertainen. Viimeisen palkan toisen potenssin ja suhdannevaikutusmuuttujan vaikutus on myös tilastollisesti merkitsevä, mutta näiden käytännön vaikutus virhetodennäköisyyteen on hyvin pieni (0,99 ja 0,98-kertainen).

Kuvassa 2 on estimoidut todennäköisyydet $\hat{\pi}_i$, että puuttuva havainto on *virhe*, kun $kausi = Q_1$ ja puuttuvia havaintoja on yksi, kaksi tai kolme (*lkm*) puuttuvaa havaintoa edeltävää palkkatietoa (*palkka*) vasten.



Kuva 2: Mallista estimoidut yrityksien i virhetodennäköisyydet $\hat{\pi}_i$ eroteltuina muuttujan *lkm* suhteen, kun $kausi = Q_1$ ja x-akseli on puuttuvaa havaintoa edeltävä palkka 100 000 euroissa (*palkka*).

Kuvasta 2 nähdään, miten ensimmäinen puuttuva havainto on toista ja kolmatta puuttuvaa havaintoa aina todennäköisimmin virhe. Kuvassa 3 on kuvan 2 estimoidut virhetodennäköisyydet $\hat{\pi}_i$, kun $lkm = 1$ ja $kausi = Q_1$ sekä virhetodennäköisyyksien π_i 95 prosentin luottamusvälit (25).



Kuva 3: Mallista estimoidut yrityksien i virhetodennäköisyydet $\hat{\pi}_i$ sekä 95 prosentin luottamusvälit, kun $lkm = 1$ ja $kausi = Q_1$ sekä puuttuvaa havaintoa edeltävä palkka 100 000 euroissa (*palkka*) x-akselilla.

Kuvasta 3 nähdään, kuinka todennäköisyys, että ensimmäinen puuttuva havainto on virhe, kasvaa viimeisen palkan mukana. Esimerkiksi yrityksellä, jonka puuttuvaa havaintoa edeltävä palkka oli 100 000, ensimmäisen puuttuvan havainnon todennäköisyys olla virhe ensimmäisellä neljänneksellä on suuruusluokaltaan 0,07-0,13.

Kuvista 2 ja 3 nähdään lisäksi, että valtaosa puuttuvista havainnoista tulee

yrityksiltä, joiden viimeinen ilmoitettu palkka on alle 100 000 euroa. Luvussa 4.2 esitetään taulukon 3 ja kuvien 2 ja 3 avulla, miten puuttuvia havaintoja tulisi ottaa palkka-aineistoon mukaan *systemaattisen revision* eliminoimiseksi palkkasummien suhdannekuvasta.

4.2 Tulosten implementointi tilastotuotantoon

Puuttuvat havainnot, jotka ovat seurausta puutteellisesta aineistosta (*virhe*), tulee poistaa palkka-aineistosta tai imputoida, koska ne korjaantuvat ei-puuttuviksi aineiston täydentyessä. *Aidot* puuttuvat havainnot tulee olla mukana palkka-aineistossa, koska ne jäävät nolliksi aineiston täydentyessä ja ovat seurausta yrityksen lopetuksesta tai yrityksen huonosta taloudellisesta tilasta. Koska luvussa 4.1 todettiin ensimmäisen puuttuvan havainnon olevan mallin mukaan todennäköisimmin virhe verrattuna toiseen ja kolmanteen, poistettaessa ensimmäinen puuttuva havainto voidaan poistaa myös peräkkäiset puuttuvat havainnot tämän jälkeen.

Mallilla voidaan laskea kullekin yritykselle i virhetodennäköisyys palkkatiedon jäädessä ilmoittamatta. Tavan, jolla puuttuvia havaintoja otetaan vuosimuutosten laskentaan mukaan, tulee ottaa huomioon tuotanto-ohjelmien määräämät rajoitukset. Yksikkötieto-ohjelma, joka käsittelee puuttuvat havainnot, ajetaan kerran kuussa aineiston päivityksen yhteydessä kaikille yrityksille. Samaista ohjelmaa ajetaan kuitenkin päivittäin kymmeniä kertoja yksittäisille yrityksillä päivittäisen tarkastustyön yhteydessä. Ohjelman ajoaika ei näin ollen saa merkittävästi pitkittyä.

Ohjelmassa on nykyisellään tieto päivitettävän yrityksen palkkahistoriasta

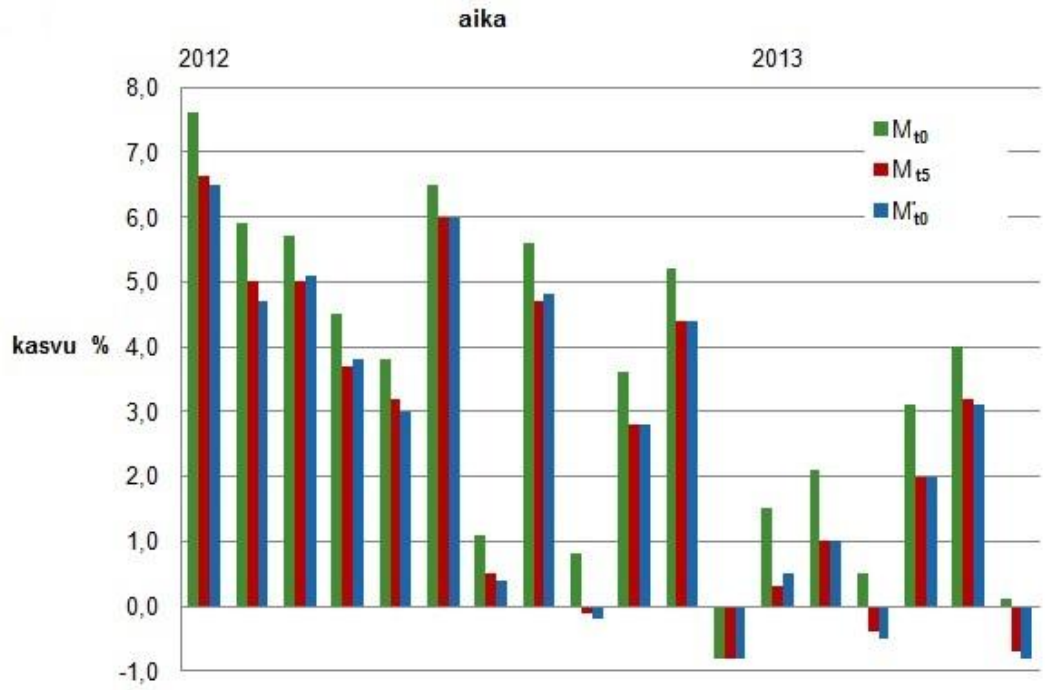
ja kuukausien lukumäärä viimeisestä ei-puuttuvasta palkkatiedosta. Ohjelmasta ei löydy tietoa eri toimialojen palkkojen kehityksestä tai tietoa yrityksen vuosiliikevaihdosta. Edellä mainitut tiedot tulisi hakea tietokannasta tai johtaa yksikkötasolta, joka hidastaa ohjelmaa.

Kuvasta 3 nähdään, miten puuttuvaa havaintoa edeltävän palkkatiedon ollessa alle 100 000, on puuttuvan havainnon virhetodennäköisyys alle 0,1. Tämä tarkoittaa, että näiden puuttuvien havaintojen todennäköisyys olla aitoja on yli 0,9. Tietoa hyväksikäyttäen suoritetaan simulointikoe, missä puuttuvat havainnot otetaan mukaan ilman viivettä yrityksiltä, joiden puuttuvaa havaintoa edeltävä palkkatieto on alle 100 000. Taulukossa 4 on alkuperäinen muutos M_{t0}, tarkentunut muutos M_{t5}, näistä laskettu revisio R_{t5} ja simuloitu muutos $M_{t0}^{'}$ sekä erotus $M_{t0}^{'} - R_{t5}$ ajanjaksolta 01/2012 - 06/2013.

Kuvaamalla taulukon 4 muuttujat M_{t0}, M_{t5} ja $M_{t0}^{'}$ aikaa vasten rinnakkain kuvaan 4 nähdään, miten päästämällä puuttuvat havainnot palkka-aineistoon ilman viivettä yrityksiltä, joilla puuttuvaa havaintoa edeltä palkkatieto on alle 100 000, ja laskemalla saadusta aineistosta muutos $M_{t0}^{'}$ vastaa se nykyisen puuttuvien havaintojen käsittelytavan tarkentunutta muutosta M_{t5}.

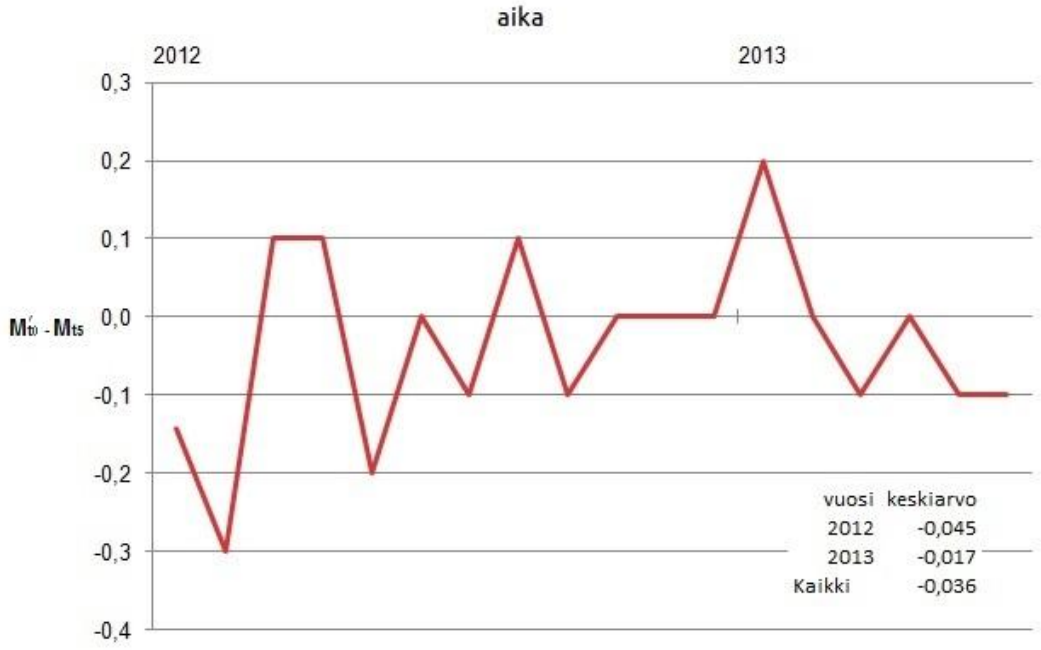
Taulukko 4: Uuden puuttuvien havaintojen käsittelytavan testitulokset verrattuna vanhaan käsittelytapaan.

aika	M_{t0}	M_{t5}	R_{t5}	M'_{t0}	$M'_{t0} - M_{t5}$
1/2012	7,6	6,6	-1,0	6,5	-0,1
2/2012	5,9	5,0	-0,9	4,7	-0,3
3/2012	5,7	5,0	-0,7	5,1	0,1
4/2012	4,5	3,7	-0,8	3,8	0,1
5/2012	3,8	3,2	-0,6	3,0	-0,2
6/2012	6,5	6,0	-0,5	6,0	0,0
7/2012	1,1	0,5	-0,6	0,4	-0,1
8/2012	5,6	4,7	-0,9	4,8	0,1
9/2012	0,8	-0,1	-0,9	-0,2	-0,1
10/201	3,6	2,8	-0,8	2,8	0,0
11/201	5,2	4,4	-0,8	4,4	0,0
12/201	-0,8	-0,8	0,0	-0,8	0,0
1/2013	1,5	0,3	-1,2	0,5	0,2
2/2013	2,1	1,0	-1,1	1,0	0,0
3/2013	0,5	-0,4	-0,9	-0,5	-0,1
4/2013	3,1	2,0	-1,1	2,0	0,0
5/2013	4,0	3,2	-0,8	3,1	-0,1
6/2013	0,1	-0,7	-0,8	-0,8	-0,1



Kuva 4: Muutokset M_{t0}, M_{t5} ja $M_{t0}^{'}$ rinnakkain aikaa vasten.

Ottamalla erotus $M_{t0}^{'} - M_{t5}$ ja piirtämällä se kuvaan 5 aikaa vasten nähdään, miten uuden menetelmän odotettu revisio eli virhe on odotusarvoisesti nolla.



Kuva 5: Erotus $M_{t0}^{'} - M_{t5}$ aikaa vasten.

Kuvasta 5 nähdään, miten erotusten $M_{t0}^{'} - M_{t5}$ keskiarvo on vuonna 2012 -0,045 ja vuoden 2013 ensimmäisellä puoliskolla -0,017. Lisäksi erotusten kokonaiskeskiarvo aikavälillä 01/2012 - 06/2013 on -0,036. Tämä tarkoittaa, että ottamalla palkka-aineistoon mukaan puuttuvat havainnot ilman viivettä yrityksiltä, joilla puuttuvaa havaintoa edeltävä palkka on alle 100 000, voidaan koko talouden palkkasumman muutoken lopullisen revision R_{t5} odottaa olevan yhden desimaalin tarkkuudella odotusarvoisesti nolla.

Simulointitulosten ja käytännön syiden valossa esitetään puuttuvien havaintojen käsittelyyn muutosta, missä viimeisen palkan ollessa alle 100 000 anne-

taan puuttuvan havainnon tulla mukaan palkka-aineistoon ja muutoin ei. Suodattamalla puuttuvat havainnot näin näyttää simulointitulosten perusteella systemaattinen revisio poistuvan palkkasummien suhdannekuvasta.

5 Yhteenveto

Palkkasummien vuosimuutosten on huomattu tarkentuvan aina alaspäin palkka-aineiston täydentyessä. Tämä systemaattiseksi revisioksi kutsuttu ilmiö on esitetty olevan seurausta tavasta käsitellä puuttuvia havaintoja. Nykyisellään yrityksen lopetuksesta seuraavat puuttuvat havainnot tulevat mukaan neljän kuukauden viiveellä, kun kaikki puuttuvat havainnot poistetaan ensin, kunnes yritys on jättänyt vastaamatta neljällä peräkkäisellä kuukaudella.

Puuttuvien havaintojen käsittelytapaan esitetään muutosta, missä puuttuvia havaintoja otetaan mukaan palkka-aineistoon oikeassa suhteessa ilman nykyistä viivettä. Mallintamalla logistisella regressiolla todennäköisyyttä, että yritykseltä raportoimatta jäänyt palkka on aineistovirhe eikä seurausta yrityksen lopetuksesta tai huonosta taloudellisesta tilanteesta, saadaan kunkin yrityksen puuttuvalle havainnolle virhetodennäköisyys.

Mallin tuloksia hyväksikäyttäen ja tuotanto-ohjelmien rajoitukset huomioonottaen muodostettiin palkka-aineiston puuttuville havainnoille uusi käsittelytapa, missä puuttuvat havainnot otetaan palkka-aineistoon ilman viivettä yrityksiltä, joilla puuttuvaa havaintoa edeltä palkkatieto on alle 100 000. Uutta käsittelytapaa testattiin laskemalla uudelleen koko vuoden 2012 ja vuoden 2013 alkupuoliskon palkkasummien vuosimuutokset ja suhdannekuvassa esiintynyt systemaattinen revisio näyttää häviävän.

Kiitokset

Työn aiheesta, aineistosta ja teknisestä tuesta haluan antaa erityiskiitokset Tilastokeskuksen Yritystilastot-yksikön Liiketoiminnan kuukausikuvaajien vastuualueelle. Lisäksi haluan kiittää hyvin sujuneesta etäyhteistyöstä ohjaajaani FT Salme Kärkkäistä sekä hyvistä kommenteista työn toista tarkastajaa professori Antti Penttistä Jyväskylän yliopistosta.

Lähteet

[1] Suomen virallinen tilasto (SVT): Palkkasummakuvaajat, Laatuseloste. Tilastokeskus, Helsinki, tammikuu 2013.

[2] Suomen virallinen tilasto (SVT): Palkkasummakuvaajat, Liitetaulukko 1. Palkkasumman vuosimuutos toimialoittain. Tilastokeskus, Helsinki, tammikuu 2013.

[3] Suomen virallinen tilasto (SVT): Palkkasummakuvaajat, Tietojen tarkentuminen. Tilastokeskus, Helsinki, kesäkuu 2013.

[4] Dobson, A. J. *An Introduction to Generalized Linear Models*. Chapman and Hall, London, 1991.

[5] Moore D. S., McCabe G. P. & Craig B. A., *Introduction to the Practice of Statistics*. W. H. Freeman and Company, New York, 2009.

[6] Searle, S. R., *Matrix Algebra useful for Statistics*. Wiley Interscience, New York, 2006.

[7] SAS Institute Inc., *SAS/STAT User Guide*, Version 8, SAS Institute, New York, 1999.

[8] Hosmer, D. & Lemeshow, S., *Applied Logistic Regression*. Wiley Interscience, New York, 1989.