

ORDERED PROBIT JA JALKAPALLO-OTTELUN MALLINTAMINEN

Random Variable
31. heinäkuuta 2005

<i>SISÄLTÖ</i>	1
----------------	---

Sisältö

1 Johdantoa	2
2 Ordered probit	3
2.1 Mallin määrittely	3
3 Jalkapallo-ottelun analyysi ordered probit-mallilla	5
3.1 Jalkapallovedonlyönnistä	5
3.2 Kirjallisuuskatsaus	7
3.3 Mallin ja aineiston esittely	10
3.3.1 Aineisto	10
3.4 Selittäjien identifiointi	10
3.4.1 Voimalukujärjestelmä	11
3.4.2 Voittosuhde-lähestymistapa	11
3.4.3 Muita selittäjiä	14
4 Alkuperäinen mallinnus	16
5 Jatkoanalyysijä aineistoa jaottelemalla	20
5.1 Mallien estimointi	23
5.2 Estimointitulosten arviointia	24
5.3 Mallien vertailu muilla tavoin	28
5.4 Regressiokertoimien tulkintaa	30
6 Johtopäätökset	31
Lähteet	33
Liitteet	

1 Johdantoa

Tämä on lyhennelmä kevään ja kesän 2005 aikana tekemästäni tilastotieteiden syventävien opintojen erikoiskurssin harjoitustyöstä, jonka aiheena on kvalitatiivinen ordered probit-malli ja sen soveltaminen jalkapallo-ottelun lopputuloksen analysoinnissa. Lisäksi työssä tarkastellaan urheiluviedonlyönin - (tai pelisijoittamistoiminnan) periaatteita sekä tehdään kirjallisuuskat-saus eri tieteellisiin artikkeleihin.

Alkuperäisestä harjoitustyöstä olen poistanut useita eri osia, ennenkaikkea teoreettisia lukuja ym. Koska tämä on tällainen kasaan parsittu dokumentti, niin joissakin kohdin saattaa olla esim. viittauksia kohtiin joita ei löydy vaan ne ovat vain alkuperäisessä työssä. Muutenkin esitys saattaa paikoin olla hie-man "töksähtelevä", mikä johtuu juuri tästä leikkaa-liimaa-kuviosta.

Viimeisten kymmenien vuosien aikana ekonometrisessa kirjallisuudessa on yhä enemmän tutkittu analysoitu mallityyppejä, joissa selitettävä muuttuja ei ole perinteiseen tapaan jatkuva vaan diskreetti. Näiden ns. kvalitatiivisten mallien estimointi ei tyypillisesti onnistu perinteisellä lineaarisella pienimmän neliösumman estimoinnilla vaan mallien estimointi on suoritettava epälineaarisilla menetelmillä, mikä ei viime vuosien aikana tapahtuneen laskentakapasiteetin kasvun tuloksena ole kovinkaan ongelmallista. Tässä harjoitustyössä tarkasteltava mallityyppi, ordered-probit-malli, on esitelty ensimmäisen kerran vuonna 1957 Atchisonin ja Simleyn toimesta. Tämän jälkeen useissa tutkimuksissa on esitelty mm. mallin estimointia ja sovituksen hyvyyden mittaamista, ja tutkimus on yhä kesken.

Miksi sitten juuri tämä malli? Kevättalvella 2005 eteen osunut John Goddardin ja Ioannis Asimakopulosin artikkeli vuodelta 2004 pohti ordered probit-mallin käyttämistä jalkapallo-ottelun lopputuloksen analysoinnissa. Oma kiinnostukseni tilastotieteisiin on alunperin lähtenyt omalta osaltaan joskus 1990-luvun loppupuolella juuri tämän saman tutkimusongelman kohdalla ja tällöin ratkaisuna esiin nousi Poisson-jakauma ja sen soveltaminen. Poisson-jakauman käyttäminen on ollut perinteinen lähtökohta jalkapal-lopelien analysointiin jo vuosien ajan, mutta aivan viime vuosien aikana on siis kokeiltu myös muita malleja varsin hyvin tuloksin. Intuitiivisesti, lopputuloksen mallintaminen suoraan ilman jakaumaapproksimaation käyttämistä kuulostaa houkuttelevalta mahdollisuudelta.

2 Ordered probit

2.1 Mallin määrittely

Edellisen luvun binäärinen vastemuuttuja voidaan yleistää myös koskemaan useampaa vastemuuttujan arvoa. Ordered probit-mallin keskeinen oletus on myös, että kyseiset vastemuuttujan arvot voidaan järjestää (järkevästi) järjestykseen. Ensimmäisenä tämän mallin esittelivät Aitchison ja Silvey (1957) jotka käyttivät siis jaksossa 2.1. esitellyn epälineaarisen funktion F paikalla normaalijakauman kertymäfunktioita sovellettuna aineistoon, jossa vastemuuttujan arvot voidaan järjestää.

Oletetaan nyt, että käytössä on $t = 1, 2, \dots, n$ kokoinen aineisto ja olkoot y_t edellä mainittujen oletusten mukainen diskreetti vastemuuttuja. Edelleen, olkoot y_t^* edellisessä luvussa esitelty latentti muuttuja, jota ei siis varsinaisesti havaita, mutta sen saama arvo ratkaisee havaittavan y_t :n saaman arvon. Tämä latentti muuttuja saa arvoja väliltä $-\infty < y_t < \infty$, ja sen saama arvo määräytyy lineaarisesti selittävien muuttujien muodostaman selittäjävektorin $\mathbf{X}_t$ perusteella, jonka oletetaan sisältävän oleellista informaatiota tarkasteltavasta ilmiöstä. Toisin sanoen, ordered probit-malli perustuu oletukselle, jonka mukaan y_t^* ja sitä myöden y_t , riippuu lineaarisesti $\mathbf{X}_t$:stä (vastaava oletus kuin binäärisessäkin tilanteessa)

$$y_t^* = \mathbf{X}_t \boldsymbol{\beta} + u_t, \quad u_t \sim NID(0, 1) \quad t = 1, 2, \dots, n \quad (1)$$

Latentti muuttuja y_t^* on siis ei-havaittava ja sen yhteys havaittavaan muuttujaan y_t voidaan formalisoida seuraavasti

$$\begin{aligned} y_t &= 1 & \text{jos} & & -\infty < y_t^* \leq \gamma_1 \\ y_t &= 2 & \text{jos} & & \gamma_1 < y_t^* \leq \gamma_2 \\ y_t &= 3 & \text{jos} & & \gamma_2 < y_t^* \leq \gamma_3 \\ &\vdots & & & \\ y_t &= m & \text{jos} & & \gamma_{m-1} < y_t^* < \infty \end{aligned} \quad (2)$$

Parametrit $\boldsymbol{\gamma} = [\gamma_1 : \gamma_2 : \dots : \gamma_{m-1}]$ ovat tunnettuja kynnysarvoja (leikkauspisteitä) joiden mukaan y_t^* :n määrää y_t :n arvon.

Yllä esitetyn yleisen mallin usein käytetty erikoistapaus on kahden kynnyksen eli kolmen eri vastemuuttujan arvon malli, joka tulee otettua käyttöön myöhemmin tässä harjoitustyössä

$$\begin{aligned} y_t &= 1 & \text{jos} & & y_t^* < \gamma_1 \\ y_t &= 2 & \text{jos} & & \gamma_1 \leq y_t^* < \gamma_2 \\ y_t &= 3 & \text{jos} & & \gamma_2 \leq y_t^* \end{aligned} \quad (3)$$

Tässä mallissa todennäköisyys sille, että $y_t = 0$ on

$$\Pr(y_t = 0) = \Pr(y_t^* < \gamma_1) = \Pr(\mathbf{X}_t\boldsymbol{\beta} + u_t < \gamma_1) = \Pr(u_t < \gamma_1 - \mathbf{X}_t\boldsymbol{\beta}) = \Phi(\gamma_1 - \mathbf{X}_t\boldsymbol{\beta})$$

ja todennäköisyys $y_t = 1$

$$\begin{aligned} \Pr(y_t = 1) &= \Pr(\gamma_1 \leq y_t^* < \gamma_2) = \Pr(\gamma_1 \leq \mathbf{X}_t\boldsymbol{\beta} + u_t < \gamma_2) = \Pr(u_t < \gamma_2 - \mathbf{X}_t\boldsymbol{\beta}) \\ &= \Pr(u_t < \gamma_2 - \mathbf{X}_t\boldsymbol{\beta}) - \Pr(u_t < \gamma_1 - \mathbf{X}_t\boldsymbol{\beta}) \\ &= \Phi(\gamma_2 - \mathbf{X}_t\boldsymbol{\beta}) - \Phi(\gamma_1 - \mathbf{X}_t\boldsymbol{\beta}) \end{aligned}$$

sekä todennäköisyys $y_t = 2$

$$\Pr(y_t = 2) = \Pr(y_t^* \geq \gamma_2) = \Pr(\mathbf{X}_t\boldsymbol{\beta} + u_t \geq \gamma_2) = \Pr(u_t \geq \gamma_2 - \mathbf{X}_t\boldsymbol{\beta}) = \Phi(\mathbf{X}_t\boldsymbol{\beta} - \gamma_2)$$

3 Jalkapallo-ottelun analyysi ordered probit-mallilla

3.1 Jalkapallovedonlyönnistä

Jalkapallo on monien mittausten mukaan maailman suosituin urheilulaji. Se on levittäytynyt ympäri maailmaa, mutta vankin sija sillä on edelleen Euroopassa, mutta toki myös Etelä-Amerikassa sekä viimeisten vuosien aikana myös aasialaiset ovat kiinnostuneet "Pelian Kuninkaasta". Jalkapalloon kuuluu kiinteänä osana varsinaisten fanien ohella myös miljardibisnekseksi kasvanut vedonlyönti. Internetin tuomat mahdollisuudet ovat yhä vain kihdyttäneet jalkapallovedonlyönnin suosiota erilaisine vetomuotoineen.

Urheiluvedonlyönnillä tai pelisijoittamisella on kohtuuttoman huono maine. Vanhoilliset piirit pitävät tätä yhä melkoisena syntinä ja toiset käyttävät sanaa uhkapeli, mikä ei todellakaan pidä paikkaansa silloin kun pelipanokset pysyvät kohtuullisina omaan varallisuuteen nähden. Itse asiassa, panostamalla kuuluisan amerikkalaisen matematikon J.L.Kellyn mukaan nimetyn kaavan tai sen modifikaation perusteella, pelaaja ei voi periaatteessa "mennä poikki" eli menettää kaikkia rahojaan. Kelly (1956) johti optimaalisen panososuuden pelikassan kasvunopeuden maksimointitehtävästä

$$B = \frac{pk - 1}{k - 1}, \quad (4)$$

missä p on peli todennäköisyysarvio, k on pelin kerroin ja näin saadaan pelikassan kasvun kannalta optimaalinen panososuus pelikassasta. Tästä Kellyn kaavasta käytetään usein modifikaatiota, jossa optimaalinen panososuus B jaetaan sopivalla jakajalla, tyypillisesti neljä tai viisi. Tätä jakajaa voidaan motivoida sillä, että usein Kellyn antamat optimaaliset osuudet ovat varsin suuria ja koska todennäköisyysarviot ovat vain "hyviä arvioita", niin on perusteltua käyttää hieman pienempiä panoskokoja ja varmistaa näin pelikassan kasvu.

Tämän kaavan perusteella on selvää, miksi menestykselliseen sijoittamiseen vaaditaan mahdollisimman tarkkoja ja hyviä todennäköisyysarvoja, joita tässä harjoitustyössä estimoidaan ordered probit-malleilla. Kellyn kaava siis antaa optimaalisen panos osuuden pelikassasta esim. 2 prosenttia. Tämän vuoksi periaatteessa pelikassa ei koskaan voi mennä nolnaan vaan tappioiden jälkeen vain rahamääräiset pelipanokset putoavat. Kellyn kaavaa on käytetty myös mm. erilaisten rahoitusinstrumenttien tuottojen ja optimaalisten sijoitussummien arviointiin.

Perinteisesti eri otteluihin on voinut lyödä vetoa otteluiden voittajista sekä myös otteluiden lopputuloksista. Tällöin vedonvälittäjät, bookkerit, tarjoavat kiinteitä kertoimia eri otteluiden voittajista ja tasapeleistä ja myös

lopputuloksista. Tämän vetomuodon ovat haastaneet aivan viime vuosina ns. tasoitusvedot joissa tyypillisesti ennalta heikompi joukkue saa maalin tai puoli maalia tasoitusta. Alla oleva esimerkki on valioliigan ottelusta keväältä 2004. Kyseisessä pelissä siis Charltonin voittaessa kerroin on 2,06 ja mikäli Leicester voittaa tai peli päättyy tasapeliin, niin kerroin on 1,94

$$Charlton(-0.5) \quad vs \quad Leicester(+0.5) \quad 2.06 \quad 1.94$$

Nämä edellä mainitut vetomuodot ovat perustuneet siihen, että eri vedonlyöntiyritykset tarjoavat kertoimia siten, että ne raapaisevat panosten jakautuessa tasaisesti eri tulosvaihtoehtojen kesken viidestä jopa kahteenkymmenen prosenttiin "hyvää" itselleen joka tapauksessa. Tässä yhteydessä puhutaan ns. *palautusprosentista*. Johtuen ed. mainitusta panosten tasaisesta jakautumisesta, palautusprosentit ovat "teoreettisia" eivätkä siis eksakteja kuten esim. ruletissa. Vedonvälittäjän teoreettinen palautusprosentti voidaan laskea seuraavasti: lasketaan halutun urheiluottelun kaikille eri lopputulosvaihtoehdoille kertoimien käänteisluvut ja ynnätään nämä. Palautusprosentti on saadun summan käänteisluku.

Nyt kuitenkin kilpaikutilanne näiden vedonvälittäjäyritysten kesken on muuttunut ratkaisevasti. Vedonlyöjät eivät enää tyydy vedonvälittäjien tarjoamiin "mataliin kertoimiin" vaan tarjoavat toisilleen paljon parempia kertoimia pelipörssissä. Pelipörssien ideana on yksinkertaisesti toimia pelaajien kohtauspaikkana jossa pelaajat tarjoavat omia kertoimiaan eri kohteista ja kysynnän ja tarjonnan kohdatessa tietyn kertoimen kohdalla toinen pelaaja hyväksyy toisen tarjouksen. Esimerkiksi pelaaja A on valmis tarjoamaan vedon Chelseaan voitosta Arsenalia vastaan kertoimella 2,00 panoksella 10 euroa ja mikäli jokin toinen pelaaja katsoo tämän kertoimen edulliseksi hän tarttuu tarjoukseen ja pelaa kertoimella 2,00 panoksen ollessa 10 euroa tai vähemmän.

Pelipörssien, kuten maailman suurimman pörssin Betfairin, tuotto tulee siitä, että ne ottavat 1-5 prosenttia voitetusta rahamäärästä itselleen eli edellisen esimerkin tapauksessa voittava osapuoli joutuu maksamaan tällaisen varsin marginaalisen osuuden voittosummastaan pelinjärjestäjälle.

Suomen kuuluisin pelisijoittaja Jorma Vuoksenmaa on kirjoittanut eräässä nettikolumnissaa seuraavasti pelipörssistä: "*Juuri nyt elämme muutosten aikaa. Uudet pelaajapörssit ovat hurjassa tahdissa valtaamassa markkinasijaa perinteisiltä vedonvälittäjiltä. Syy on yksinkertainen: pelaajapörssi tarjoaa ylivoimaiset ehdot kaikille. Vain alle 1 prosentti urheiluvedonlyöjistä pystyi aikaisemmin voittamaan rahaa perinteisiltä vedonvälittäjiltä pitkässä sarjassa. Nyt jo niukastikin keskivertovedonlyöjää paremmilla tiedoilla pelaava alkaa saavuttaa positiivisia tuottolukuja. Pelipörssien tileillä rahat ovat turvassa.*"

Pelipörssin vaara mennä konkurssiin on olematon. Pörssihän ei kanna riskiä tappioista, koska vedon osapuolina on aina kaksi vedonlyöntiharrastajaa jostakin päin maailmaa. Lisäksi asiakkaiden pelitilien varat ovat erillään yhtiöiden varoista. Perinteiset vedonvälittäjät saattavat sen sijaan joutua pian taloudellisiin vaikeuksiin. Kohta kukaan ei tyydy vedonvälittäjien tarjoamiin ker-toimiin vaan alkaa ottaa lähes 10 prosenttia suurempia voittoja pelipörssistä".

Lisää pelisijoittamisen periaatteista mm. Marttinen (2001)

3.2 Kirjallisuuskatsaus

Ensimmäiset artikkelit jalkapallopelien tilastollisesta mallintamisesta ovat keskittyneet lähinnä tehtyjen maalien jakaumien tutkimiseen Poisson-jakauman avulla. Moroney (1956) ehdotti, että huolimatta Poisson- jakauman melko tyydyttävästä kyvystä mallintaa maaleja, negatiivinen binomijakauma saattaisi tarjota paremman vaihtoehdon. Myös Reep ym. (1971) totesivat että negat.binomijakauma olisi varsin hyvä approksimaatio maaleilla, mutta varsinainen heidän tutkimustuloksensa oli se, että "sattuma hallitsee peliä" eli heidän mukaansa ei ollut mahdollista löytää sen aikaisista malleista sellaista jonka avulla olisi voitu päästä järkeviin tuloksiin. Näinä vuosina esitettiin vieläkin tiukempia tuloksia, joiden mukaan vain "sattuma ratkaisee tuloksen".

Intuitiivisesti on helppo yhtyä Hillin (1974) ajatuksiin. Hänen mukaansa jokainen joka on katsonut jalkapalloa on voinut todeta, että ottelu ratkeaa taidon ja sattuman yhteisvaikutuksen tuloksena. Hill osoitti, että asiantuntijoiden ennen kautta muodostamat käsitykset joukkueiden paremmuuseroista sekä joukkueiden lopullisen sijoituksen välillä on merkitsevä korrelaatio. Erityisesti viime vuosina eurooppalaisessa jalkapallossa on havaittu, että pitkällä aikajänteellä huippujoukkueet voittavat lähestulkoot kaikki merkittävät kilpailut.

Tilastollisesti, jalkapallo-ottelu voidaan nähdä satunnaistapahtumana, jossa ottelun lopputuloksena voi olla kolme vaihtoehtoa: kotivoitto, tasapeli ja vierasvoitto. Jokaisella näistä vaihtoehdoista on tietty todennäköisyys ja ne summautuvat todennäköisyyden tulkinnan mukaisesti ykköseksi. Ottelun lopputulosta mallinnettaessa koti- ja vierasjoukkueiden maalien avulla, oletetaan että estimoidut maalimäärät joukkueittain kuvaata niiden kykyä saavuttaa ottelusta tietty tulos (voitto, tasapeli, tappio). Tämän tyypistä lähestymistapaa nimitetään Maher- Poisson- lähestymistavaksi Maherin artikkelin (1982) mukaisesti.

Siinä hän oletti että joukkueiden A (=kotijoukkue) ja

B (=vierasjoukkue) tekemät maalit ottelussa noudattavat toisistaan riippumatta Poisson-jakaumia maaliodotusarvoina λ_a ja λ_b . Tämä lähestymistapa on tuottanut muitakin artikkeleita, mm. Dixon ja Coles (1997) lähtivät liikkeellä tästä Maherin lähtökohdasta. He tutkivat vuosien 1992-1995 välisenä aikana pelattujen Englannin neljällä ylimmällä sarjatasolla pelattujen otteluiden tuloksia. Aineistosta saatujen tulosten reunajakaumat koti- ja vierasjoukkueen maalimäärille noudattivat erittäin tarkasti Poisson-jakaumaa. Kuitenkin, he havaitsivat lievää korrelaatiota koti- ja vierasmaalien ,erityisesti alhaisten maalimäärien $0 - 0, 1 - 0, 0 - 1, 1 - 1$, välillä. Siksi he esittivät Poisson-jakaumaan korjaustermiä

$$P(X_{i,j} = x, Y_{i,j}) = \tau_{\lambda_a, \lambda_b}(x, y) \frac{\lambda_a^x e^{-\lambda_a}}{x!} \frac{\lambda_b^y e^{-\lambda_b}}{y!}$$

missä korjaustermi $\tau_{\lambda_a, \lambda_b}(x, y)$

$$\tau_{\lambda_a, \lambda_b}(x, y) = \begin{cases} 1 - \lambda_a \lambda_b \rho & \text{jos } x=y=0 \\ 1 + \lambda_a \rho & \text{jos } x=0, y=1 \\ 1 + \lambda_b \rho & \text{jos } x=1, y=0 \\ 1 - \rho & \text{jos } x=y=1 \\ 1 & \text{muulloin} \end{cases}$$

Tässä siis ρ on korrelaatiokerroin koti-sekä vierasmaalien välillä.

Maaliodotusarvot λ_a ja λ_b muodostettiin eri joukkueiden hyökkäys- ja puolustusparametrien α ja β mukaisesti. Lisäksi huomioon otetaan kotietu γ

$$\begin{aligned} \lambda_a &= \alpha_a \beta_b \gamma \\ \lambda_b &= \alpha_b \beta_a \end{aligned}$$

Dixonin ja Colesin periaatteiden mukaisesti Raitanen (1999) tutki Englannin valioligan tuloksia vuosien 1994-1998 väliseltä ajalta ja havaitsi, että aineistossa maalilukujen varianssi on suurempi kuin keskiarvo, mikä viittaisi siihen, että Poisson-jakauma ei pätsisi. Siksi hän esittää yleistetyn Poisson-jakauman käyttöönottoa, minkä hän toteaa sopivan aineistoon selkäesti paremmin χ^2 - yhteensopivuustestin perusteella. Edelleen, Raitanen ehdottaa Dixonin ja Coles- korjaustermiä korvaamista yleisemmällä korjaustermillä, joka ottaa huomioon myös muiden teoreettisten tulosten poikkeamisen toteutuneista maaliluvuista kuin vain vähämaalisten otteluiden kohdalla.

Rue ja Salvesen (2000) toteavat että vaikka perinteinen Poisson-oletus on järkeenkäypä, se ei välttämättä pidä paikkaansa silloin, kun toinen joukkueista tekee runsaasti maaleja ja siirtyy käytännössä tavoittamattomaan johtoon.

3 JALKAPALLO-OTTELUN ANALYYSI ORDERED PROBIT-MALLILLA9

Tällainen tilanne lannistaa tappiolla olevan joukkueen, ja johtaa maalintekointensiteetin muutokseen, joka on ristiriidassa oletuksen kanssa että maalintekointensiteetti ei ole riippuvainen ottelussa aiemmin tehdyistä maaleista. Rue ja Salvesen muokkaavat Dixonin ja Colesin mallia niin että joukkueen viidennen maalin jälkeen tehtyjä maaleja ei oteta huomioon joukkueiden hyökkäys- ja puolustusparametreja määriteltäessä. Otteluiden lopputulosten ennustamisessa on käytetty myös erilaisia stokastisia Markov-malleja, joissa lasketaan edellä mainittujen maalintekointensiteettien perusteella erilaisia tapahtuman kulkuja.

Kuten tästä kirjallisuuskatsauksesta voidaan todeta, kirjallisuudessa on pääpiirteittäin keskitytty maalimäärien mallintamiseen ja tämän jälkeen käytettävään Poisson-approksimaatioon. Kuitenkin aivan viime vuosina jotkut ekonometrikot ovat ehdottaneet otteluiden mallintamisessa käytettäväksi suoraan diskreettejä tulosmalleja. Ensimmäiset viitteet Probit- mallien käytöstä jalkapallo-otteluiden analysoinnista löytyvät Koningin (2000) ja Goddardin ja Asimakopoulousin (2004) artikkeleista. Jälkimmäinen artikkelissa viitataan myös mm. Forrestin ja Simmonsien artikkeleihin (2000), joissa he käyttivät muita diskreettejä vastemuuttujamalleja tutkiessaan englantilaisten lehtien kerroin-arvioiden osumisia.

Probit-mallinnuksella voidaan katsoa olevan useampia etuja verrattuna maaliennusteisiin perustuvaan Poisson-mallinnukseen. Luultavasti suurin hyöty koituu siitä, että näillä malleilla vältetään jo tässäkin kirjallisuuskatsauksessa esiin tuleesta ongelmasta koskien koti- ja vierasjoukkueiden maalimäärien riippumattomuudesta. Korrelaatio näiden välille voidaan osoittaa varsin vähäiseksi, mutta kuten jo esim. Dixonin ja Colesin käyttämä korjaustermi osoittaa, joidenkin tulosten kohdalla Poisson-jakauman antamat approksimaatiot ovat joko liian suuria tai liian pieniä verrattuna toteutuneisiin tuloksiin. Lisäksi tämän Poisson- lähestymistavan heikkoutena voidaan pitää sitäkin, että maaliarvot pitäisi saada estimoitua varsin tarkasti. Kuten Rue ja Salvesenkin totesivat, poikkeukselliset tulokset saattavat muuttaa ratkaisevasti maaliennusteita. Esimerkiksi käyttämällä yksinkertaisesti keskiarvoa tehdyistä maaleista suoraan ennusteena, niin varsinkin alkukauden aikana, kun pelejä on pelattuna vain muutamia, yksi suurimallinen ottelu aiheuttaa ennusteiden "räjähtämisen".

3.3 Mallin ja aineiston esittely

Kirjallisuudessa on esitetty ainakin kahta erilaista Probit-mallityyppiä mallintamaan jalkapallo-otteluita. Vaihtoehtoinen tapa tässä harjoitustyössä tarkasteltavalle mallille löytyy Koningin artikkelista (2000).

Tässä työssä käytetään kolmen vaihtoehdon probit-mallia, joka on vastaavanlainen kuin Dobsonin ja Goddardin (2001) sekä Dobsonin että Asimakopouloson (2004) tarkastelema malli

$$\begin{array}{llll}
 \text{vierasvoitto} & y_t = 1 & \text{jos} & y_t^* < \gamma_1 \\
 \text{tasapeli} & y_t = 2 & \text{jos} & \gamma_1 \leq y_t^* < \gamma_2 \\
 \text{kotivoitto} & y_t = 3 & \text{jos} & \gamma_2 \leq y_t^*
 \end{array} \quad (5)$$

Selittävien muuttujien esittelyä seuraavassa

3.3.1 Aineisto

Aineistona tässä käytetään Englannin jalkapallon pääsarjan Premier Leaguen eli Valioliigan ottelutuloksia kausilta 2000-2004. Saatuja tuloksia sovelletaan kauden 2004-2005 tuloksiin. Lisäksi eräissä tarkasteluissa käytetään lisäksi kausien 1997-2002 väliseltä ajalta joukkuekohtaisia lopputuloksia, joiden avulla osittain muodostetaan mallin selittäviä muuttujia. Näiden kausien aikana Valioliigassa on pelannut 20 joukkuetta, joka merkitsee sitä, että otteluita tulee kautta kohden 380.

Aineisto on koottu internetistä <http://www.soccerway.com> sekä <http://stats.football365.com> - palstojen avulla jotka tarjoavat tilastoja lähes tulkoot kaikkiin maailman merkittäviin jalkapallosarjoihin. Selittäjiä on muodostettu mahdollisimman objektiivisesti sarjatilanteiden sekä "oman kokemuksen ja pelituntemuksen mukaisesti siis sillä kuuluisalla "mutu-tuntumalla". Objektiivisuudella tässä tarkoitetaan sitä, että omien muistiinpanojen lisäksi on paikoin ollut tarpeellista palata menneisyyteen koodaamaan tarvittavia tietoja ja ne on pyritty tekemään sillä ajatuksella että ikäänkuin tulevia tapahtumia ei tiedettäisi. Tarkemmin selittäjiä pohditaan seuraavassa.

3.4 Selittäjien identifiointi

Muodostetaan seuraavaksi potentiaalisia selittäjiä. Seuraavien pohdintojen keskeinen ajatus on vain muodostaa mahdollisimman hyvät mittarit mittamaan joukkueiden välisiä tasoeroja.

3.4.1 Voimalukujärjestelmä

Alkuperäinen idea itse ylläpitämälleni voimalukujärjestelmälle on peräisin Jorma Vuoksenmaan kirjasta (1999). Kertoimenlaskijan keskeinen kysymys jonka edessä hän on eri otteluiden osalta on: **kumpi joukkueista on parempi ja kuinka paljon parempi** ? Kuten Vuoksenmaa toteaa:

"Viime otteluiden tulosten tarkastelu on liian heppoinen menetelmä antaakseen tyhjentävän vastauksen. Yhtä vaarallista on perustaa arvio puhtaasti sarjataulukkoon. Toinen joukkue on saattanut edellisissä otteluissa kohdata paljon kovempia vastustajia kuin toinen. Lisäksi vain muutama epäonninen maali on voinut aiheuttaa suuria muutoksia sarjasijoituksessa."

Pelkän edellisen kauden ottelutulosten tai voittosuhteen perusteella ei välttämättä saada oikeata kuvaa joukkueen pelivoimasta. Joukkueet ailahtelivat jo pelkkien satunnaistekijöiden vuoksi vuodesta toiseen sarjataulukossa. Normaali satunnaisheilahtelu voi hyvin olla jopa ± 12 pistettä (kun voitosta 3pt ja tp:sta 1pt) (Vuoksenmaa,1999).

Voimaluvut ovat osoittautuneet toimivaksi ratkaisuksi näihin kysymyksiin. Tämä menetelmä voidaan tiivistää siten, että ne ovat ikäänkuin "jokamiehen"näppituntuma joukkueiden kyvykkyyksistä matemaattisessa muodossa. Voimalukujärjestelmän voi pallopeleissä perustaa joukkueiden kauden kokonaispisteiden odotusarvoihin.

Oma systeemi perustuu edelliseen ideaan. Ennen kauden alkua muodostetaan eri tekijöiden perusteella joukkuekohtaiset voimaluvut. Voimalukujärjestelmäni pohjana on "oma sarjataulukko" jossa voitosta saa yhden pisteen ja tasapelistä puoli pistettä. Näin ollen aikaisemman kauden "sarjataulukko" sekä käsitykset joukkuekohtaisista muutoksista esim. pelaajapakauppojen jälkeen tulevat huomioiduiksi. Tarkempaan esittelyyn ei ehkä tässä yhteydessä ole syytä, mutta pääideana on siis tarkestella joukkueiden välisiä paremmuuseroja ja muodostaa jokin matemaattinen skaala jonka puitteissa näitä eroja pyritään kuvaamaan. Syitä miksi jokin joukkue saa juuri tietyn arvon on vaikea formalisoida paperille, vaan ne perustuvat pitkälti kokemukseen joukkueiden ja erilaisten pelien seuraamisesta.

Liitteessä 1 on kausien 1997-2005 ajalta ennen pelikauden alkua tehdyt voimaluvut.

3.4.2 Voittosuhte-lähestymistapa

Vastaavasti kuin voimalukujen muodostamisessa tarkasteltavan joukkueen voittosuhte voidaan määritellä siten, että voitosta saa yhden pisteen ja tasapelistä puolikkaa pisteen ja jaetaan tämä summa pelatuilla otteluilla. Tätä lähestymistapaa käyttävät myös Goddard ja Asimakopoulos (2004).

Tätä suhdelukua voidaan tilastoida useamman kauden ajalta joukkuekohtaisesti.

Tähän lähestymistapaan liittyy kuitenkin useampia ongelmia. Ensinnäkin eri sarjoita putoaa joukkueita alempaan ja vastaavasti alemmasta divisioonasta nousee uusia joukkueita ylämpään sarjaan. Tällöin näiden joukkueiden voittosuhteet eivät ole suoraan vertailukelpoisia uudessa/ylemmässä sarjassa. Toisaalta kuten edellä todettiin, joukkueiden kyvykkyys muuttuu aina pelikausien välillä pelaajahankintojen yhteydessä. Näin ollen edellisellä kaudella saavutettu voittosuhte ei välttämättä ole järkevä ja hyvä lähtökohta suoraan seuraavallekin kaudelle.

Monissa artikkeleissa voimasuhteet tai maaliodotusarvot on laskettu aikapainotettuina keskiarvoina muutaman edellisen kauden tuloksista, esim. Goddard (2004). Tällöin sarjakausien välillä tapahtuvia muutoksia joukkueiden välisissä voimasuhteissa ei oteta huomioon. Käytetäänkin tässä työssä erilaista lähestymistapaa. Lasketaan jokaista ottelua kohden molemmille joukkueille oma voittosuhte, joka perustuu kauden alun ennakkoarvioihin sekä sitten toteutuneeseen kehitykseen. Toteutunutta kehitystä käsitellään aikapainotettuna keskiarvona, jossa Dixonin ja Colesin 1997 ideaa mukaisesti aikaisempia tuloksia painotetaan funktiomuodon $\exp(-\zeta t)$ mukaisesti. Dixon ja Coles estimoivat parametrin ζ arvoksi 0,0065, mutta he laskivatkin nämä vastaavat keskiarvot viimeisen 24 kuukauden ajalta. Järkevältä tuntuva arvo tässä työssä on noin 0,01 ja se onkin kiinnitetty siihen. Sinänsä lopputuloksen kannalta ei ole olennaista onko tämän parametrin arvo tämä vai jokin sen välittömästä läheisyydestä 0,005-0,02, koska 38 ottelun sarjassa ja edelleen 19 koti- ja vieraspeliin yhtäkään ottelua ei kannata "dumpata" liian pienellä painoarvolla ennusteita laskettaessa.

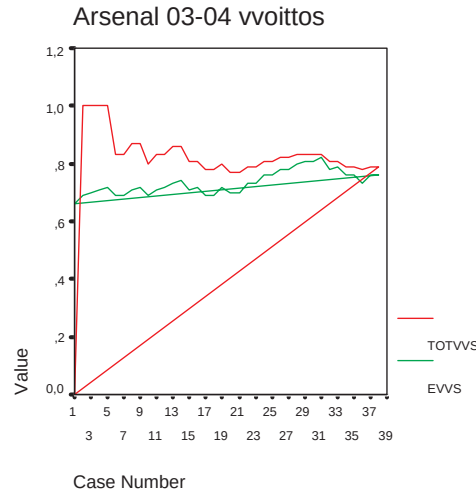
Sanalla "kausianalyysiestimointiin" tarkoitetaan tässä regressioanalyysiä, jossa selittäjinä ovat edellisen kauden voittosuhte sekä ennustevoittosuhte tulevaksi kaudeksi ja selitettävänä muuttujana tarkasteltavan kauden toteutunut voittosuhte. Näitä estimointituloksia on Litteessä 2 niin että joukkueet on jaettu niihin jotka olivat valioliigassa jo edellisenä kautena sekä sarjanousijoihin. Merkitään tätä ennustetta symbolilla Υ

Tarkoitus on siis ennen kauden alkua, tässä esimerkissä 2003-2004, muodostaa mahdollisimman hyvä ennuste tulevaksi kaudeksi. Selitettävänä ovat siis kauden 03-04 arvot ja selittäjinä edellisen kauden sekä kausianalyysin ennakot.

3 JALKAPALLO-OTTELUN ANALYYSI ORDERED PROBIT-MALLILLA13

joukkue	voimaluku	kokvoittos	kvoittos	vvoittos	rank
Arsenal 02-03	27,5	0,72	0,84	0,61	5
kausianalyysi e	28,5	0,75	0,84	0,66	5
Arsenal 03-04	32	0,84	0,89	0,79	5

Taulukko 1 Arsenalin voittosuhteita, missä "kausianalyysi e" on ennuste kaudeksi 03-04



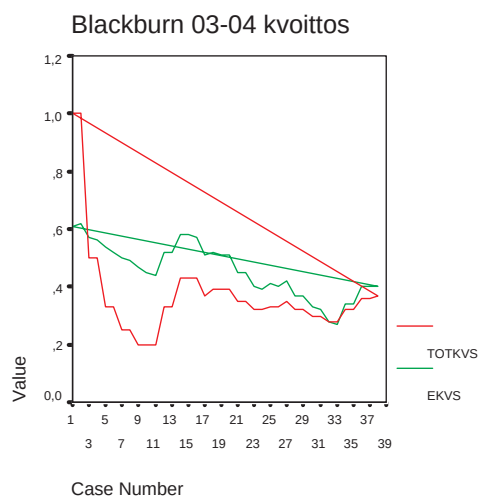
Kuva 1: Arsenalin vierasvoittosuhteen kehitys (totvvs) sekä ennuste (evvs)

Kuten kuvioista voidaan huomata, erityisesti alkukauden aikana voittosuhteiden kohdalla menee runsaasti pelikierroksia jotta todellinen voittosuhte "normalisoituu" järkeväksi. Edelleen, toteutuvaa kehityshän pitäisi voida ennustaa myös ensimmäisen kierroksen osalta. Näiden näkökohtien perusteella on järkevää käyttää eräänlaista ennustetta pelkän painotetun toteutunutta kehitystä kuvaavan keskiarvon sijasta. Merkitään tätä ennustetta notaatiosyistä Ψ

$$\Psi = (1 - \frac{krs.lkm}{38})\Upsilon + (\frac{krs.lkm}{38}) * toteutunutvoittosuhte \quad (6)$$

Tässä siis $krs.lkm$ on pelattujen pelikierrosten lukumäärä ja toteutunut koti- tai vieraspelien voittosuhte riippuen tarkasteltavasta joukkueesta.

Näin muodostuu jokaiseen tiettyyn otteluun Ψ_{koti} ja Ψ_{vieras} jotka pyrkivät kuvaamaan mahdollisimman hyvin juuri sen hetkistä joukkueiden pelivoimaa. Näiden lukujen on tarkoitus olla mallin "pääselittäjät".



Kuva 2: Blackburnin kotivoittosuhteen kehitys (totkvs) sekä ennuste (ekvs)

3.4.3 Muita selittäjiä

Edellä johdettut voimasuhde-ennusteet ovat hyvin keskeisessä asemassa joukkueiden keskinäisiä paremmuuseroja määriteltäessä. Kuitenkin muitakin tekijöitä on hyvä huomioida. Seuraavien selittäjien käyttäminen voidaan perustella aikaisempien, erityisesti Goddardin, muiden artikkeleiden, Vuoksenmaan kirjan sekä omien havaintojen perusteella. Seuraavat muuttujat määrätään niin koti- kuin vierasjoukkueillekin.

RANK=Edellisten voimasuhteiden lisäksi rankataan joukkueet asteikolla 1-5. Tällä onkin keskeinen merkitys lopullisessa mallissa.

TÄRKEYS= asteikolla 1-5 on muodostettu joukkueiden rank-arvojen perusteella approksimaatiot pelin tärkeydestä (5=erittäin tärkeä). Tyypillisesti suuria arvoja havaitaan siis silloin kun lähes tasavahvat joukkueet kohtaavat, jolloin voidaan olettaa ne taistelevat myös samoista sarjasijoituksista kärki- tai häntäpäässä.

RIVALRY= tiettyjen joukkueiden välillä voidaan havaita viha-rakkaussuhteita jotka pohjautuvat yleensä kauas historiaan. Yleensä kyseisten joukkueiden kotipaikat sijaitsevat lähellä toisiaan tai joukkueet esim. ovat kuumimpia mestariuosikkeja jonka vuoksi pelissä on paljon enemmän kuin pelkät kolme sarjapistettä.

PAINE= asteikolla (-1,0,1,2) arvioidaan tarkasteltavan joukkueen kokemaa painetta. Jos joukkue on "pakkovoitto-tilanteessa", niin paine-muuttuja

3 JALKAPALLO-OTTELUN ANALYYSI ORDERED PROBIT-MALLILLA15

saa arvon 2. Lähtökohtaisesti voidaan olettaa, että noin puoleen väliin asti sarjakautta kaikkien joukkuiden painetila on 0. Arvo -1 voidaan antaa kun pelillä ei ole suurta merkitystä, mikä voi johtua mm. sarjasijoituksesta keskeillä taulukkoa tai siitä, että lähitulevaisuudessa on tai oli hyvin tärkeä muun turnauksen peli.

KAMP-muuttujat: MEST =mestaruus KÄRKI =kärkipää PUTOM =putoamis.

HÖYKJÄLK= runsasmaalisen tappion (3 tai useamman maalin) jälkeinen peli. Monesti joukkueet skarppaavat pahan tappion jälkeen

TÄRKJÄLK= erityisen tärkeän pelin jälkeinen ottelu. Onko keskittymisen kohdallaan?

EURO-muuttujat =mikäli joukkue on mukana eurocupeissa (mestareiden liiga, UEFA-cup), niin indikaattori saa arvon 1 kun tarkasteltava peli on joko ennen, jälkeen tai europelien välissä.

CUP= arvo 1 kun joukkue jos joukkue on vielä tammikuun alun cupkierrosten jälkeen mukana F.A.-Cupissa.

4 Alkuperäinen mallinnus

Tarkastellaan aluksi korrelaatioita TULOS ja mahdollisten selittäjien kohdalla. Voidaan havainta, että joukkueiden paremmuutta mittaavat muuttujat RANK (koti- ja vieras) sekä estimoitu VOITTOS korreloivat voimakkaammin TULOS muuttujan kanssa.

Jo tässä vaiheessa on selvää, että useiden muuttujien kohdalla törmätään tilanteeseen jossa ne kertovat enemmänkin joukkueiden tasoeroista kuin tietyn ottelun erityispiirteistä. Esimerkiksi eri KAMP-muuttujat osoittavat jo tämän.

muuttuja	korrelaatio	merkitsevyys	muuttuja	korrelaatio	merkitsevyys
KRANK	0.21	0.00	VRANK	-0.27	0.00
kekvoittos	0.21	0.00	vevvoittos	-0.24	0.00
ktärkeys	0.02	0.48	vtärkeys	-0.06	0.03
krivalry	-0.05	0.08	vrivalry	0.01	0.72
kpaine	0.05	0.08	vpaine	-0.14	0.00
kkampmest	0.10	0.00	vkampmest	-0.18	0.00
kkampkärki	0.05	0.04	vkampkärki	-0.08	0.00
kkampputom	0.12	0.00	vkampputom	0.06	0.03
khöykjäl	0.00	0.98	vhöykjäl	-0.01	0.71
ktärkjäl	0.02	0.56	vtärkjäl	-0.04	0.16
kvahvputjäl	-0.01	0.57	vvahvputjäl	0.02	0.55
kheikputjäl	-0.05	0.04	vheikputjäl	-0.02	0.45
keuroennen	0.09	0.00	veuroennen	-0.04	0.15
keurojälk	0.07	0.01	veurojälk	-0.07	0.00
keurovälis	0.06	0.03	veurovälis	-0.12	0.00
kcup	0.05	0.07	vcup	-0.07	0.01

Taulukko 2 Korrelaatiot TULOS- muuttujan kanssa. Merkitsevyys viittaa testiin jossa nollahypoteesina korreloimattomuus.

Lasketaan seuraavaksi muita perustunnuslukuja. Ottelun lopputulosten suhteelliset osuudet

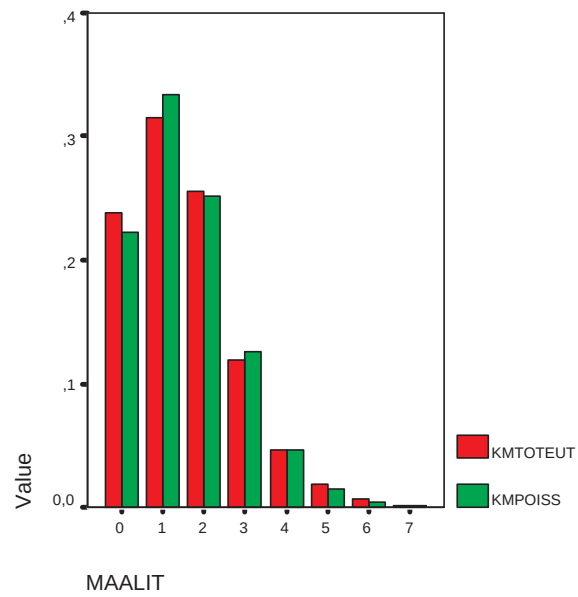
tulos

tulos	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	417	27.43	417	27.43
2	400	26.32	817	53.75
3	703	46.25	1520	100.00

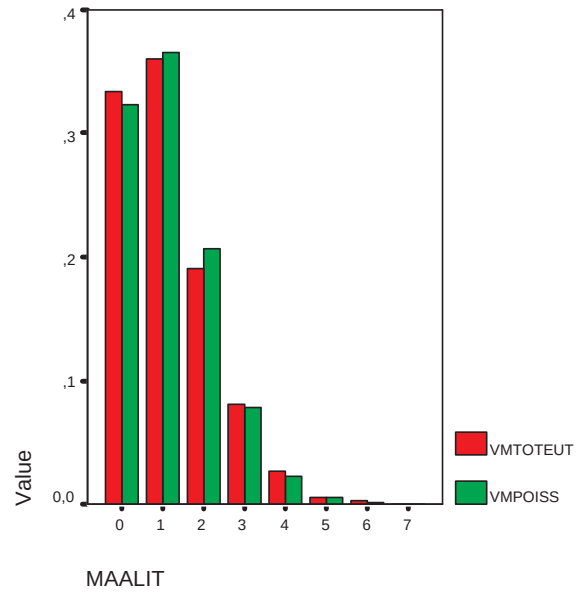
Eri tulosten suhteelliset osuudet noudattavat hyvin pitkän aikavälin vastaavia raportoituja osuuksia. Kotivoittoa siis noin 46.3, tasapelejä 26.3 ja vierasvoittoa 27.4 prosenttia tarkastelluista otteluista.

Alkuperäisen mallinnuksen yhteydessä päädyttiin malliin, jossa kaikki ei ollut aivan kohdallaan. Seuraavassa pala tästä analysoinnista.

Kuten jakson 4.2. kirjallisuuskatsauksesta voidaan todeta, perinteisesti juuri jalkapallopelien todennäköisyysarvioita on tehty Poisson-jakaumaan nojaten. Vertaillaan nyt näitä tekniikoita kausien 2000-2004 aineiston perusteella. Ensinmmäiseksi on estimoitava Poisson-jakauman maali-odotusarvoparametrit λ_a ja λ_b . Kuvatkoott joukkueiden hyökkäys- ja puolustusparametreja joukkueiden tehdyt ja päästetyt maalit. Muodostetaan näin regressiomallit niin koti- ja vierasjoukkueiden odotettavissa oleville maaleille, jossa selittäjinä ovat kotijoukkueen tilanteessa kotijoukkueen aikaisemmissa kotiotteluissa tehdyt maalit sekä vierasjoukkueen vierasotteluiden päästetyt maalit.



Kuva 3: Kotimaalit sekä Poisson-ennusteet



Kuva 4: Vierasmaalit sekä Poisson-ennusteet

Näin saadaan parametreille λ_a ja λ_b estimaatit. Käytetään nyt myös Poissonin-jakauman yhteydessä Dixonin ja Colesin (1997) esittelemää korjaustermiä. Näin saadaan lopulta todennäköisyysarviot eri tulosvaihtoehdoille kuten ordered probit-malleissakin. Kuten kuvista 3 ja 4 voidaan havaita, voidaan hyvin olettaa että koti- ja vierasmaalit generoituvat Poisson-jakauman kautta. Otoksena on käytetty kausien 2000-2004 aineistoa jossa havaintoja on 1520 kappaletta.

Nyt voidaankin pohtia näiden eri lähestymistapojen ominaisuuksia. Kumpi menetelmä antaa parempia tuloksia? Kysymykseen on vaikeaa antaa yksiselitteistä vastausta, ja sitä paitsi tämän "paremmuuden" mittaaminen on tässä yhteydessä vaikeaa. Yksi mahdollinen ja varsin hyvä mittari voisi olla laskea yhteen toteutuneiden tulosten mukaan näiden menetelmien antamat ennusteet kyseiselle lopputulokselle. Menettelemällä näin, havaitaan että ordered probit-ennusteiden todennäköisyyssumma on 633 kun vastaavasti Poisson-menetelmässä tämä on 584. Tällä mittarilla ordered probit olisi merkittävästi parempi menetelmä.

Kuitenkin tarkastelemalla lähemmin menetelmien antamia todennäköisyysarvioita, voidaan tehdä muutamia silmään pistäviä havaintoja. Esinnäkin, ordered probit-mallin ottaa huomattavasti enemmän kantaa kuin varsin konservatiivinen poisson-malli. Esimerkkinä tästä kevään 2004 ottelu

Tottenham-Chelsea	1	X	2	
ordered-probit	0,10	0,21	0,69	
Poisson	0,30	0,26	0,44	

Taulukko 3 Esimerkkiottelu op-mallin ongelmista (op=ordered probit)

Tässä siis silmään pistävää on op-mallin näkemys Chelseaan ylivoimasta. Kokemuksen perusteella voi sanoa, että tämän tyyppisessä otteluissa, tämän tasoisten joukkueiden kesken, jälkimmäiset tn-arviot ovat "järkevämmät". **Op-mallille näyttää siis olevan tyypillistä antaa rank = 5 omaaville huippujoukkueille liiankin suuria tn-arvioita.** Nyrkkisääntönä voidaan pitää, että Englannin valioliigan tasoissa Euroopan huippusarjoissa vierasjoukkueen voitontodennäköisyys ei voi normaalitilanteissa käytännössä ylittää 75 prosenttia.

5 Jatkoanalyysejä aineistoa jaotteleamalla

Aluksi tätä op-mallia pyrittiin siis estimoimaan suoraan koko aineistolle, mutta myöhemmässä vaiheessa havaittiin lopputuloksissa varsin "järjettömiä" ja soveltamisen kannalta hankalia tn-arvioita.

Tämän vuoksi lähdettiin hakemaan toisenlaista ratkaisua. Aivan ilmeisesti aineistoa on tarvetta jaotella jotta eri tyyppiset ottelut sijoittuvat suurinpiirtein samalla tavalla käyttäytyvien pelien kanssa samoihin ryhmiin. **Aineisto on siis varsin heterogeeninen**, ja siksi sitä on syytä jaotella sopivasti ja muodostaa kyseisille ryhmille omat mallit.

Tämän tyyppistä jaottelua voidaan tietysti kritisoida eri perusteilla. Kauden alla valitaan eri seuroille rank-arvot joukkueiden pelillisen kyvykkeen mukaisesti. Ovatko nämä arviot sitten hyviä eli kuvaavatko ne hyvin joukkueiden paremmuuseroja? Tarkastelemalla eri joukkueiden eri kausien rank-arvoja sekä kauden lopun sarjatilanteita voidaan todeta että tämä hyvin karkea jako on perusteltu: ennakkolta vahvat seurat ovat olleet kärjessä ja heikot pääsosan sarjan hännillä. Yksittäisiä epäonnistuneitakin arvioita on toki tapahtunut. Rank-arvoja on myös korjailtu joidenkin seurojen kohdalla muutamien otteluiden pelaamisen jälkeen, kun joukkueen pelillinen taso on ollut yllättävän hyvä tai huono. Esimerkkinä tästä Everton kaudella 2004-2005 jolloin rank nostettiin kahdesta kolmeen.

Tehdään siis uudet luokittelevat muuttujat koti- ja vierasjoukkueiden rank-arvojen mukaan seuraavasti

rankjako		
krank	vrank	luokka
5	1	25
5	2	24
4	1	23
5	3	22
4	2	21
3	1	20
4	3	19
3	2	18
2	1	17
5	4	16
1	1	15
2	2	14
3	3	13
4	4	12
5	5	11
4	5	10
1	2	9
2	3	8
3	4	7
1	3	6
2	4	5
1	4	4
3	5	3
2	5	2
1	5	1

Tämän jaon jälkeen tarkasteltiin otteluiden lopputulosten frekvenssejä eri luokissa. Samassa yhteydessä muodostettiin viitteelliset teoreettiset frekvenssit joita kohden toteutuneiden frekvenssien voisi olettaa konvergoivan havaintomäärien kasvaessa. Nämä luvut perustuvat niin nyt havaittuihin tuloksiin kuin myös ennakoarvioihin eri luokkien tn-jakaumista.

rankjako				
luokka	koti tn.	tasap tn.	vieras tn.	ryhmä
15	0,42	0,29	0,29	6
9	0,34	0,3	0,34	6
6	0,30	0,29	0,41	6
4	0,18	0,27	0,55	3
1	0,15	0,19	0,67	3
17	0,45	0,28	0,27	1
14	0,45	0,29	0,26	1
8	0,45	0,28	0,27	1
5	0,29	0,28	0,43	4
2	0,16	0,23	0,61	3
20	0,52	0,28	0,20	8
18	0,48	0,30	0,22	1
13	0,45	0,31	0,25	1
7	0,37	0,29	0,34	4
3	0,17	0,25	0,58	3
23	0,68	0,20	0,12	7
21	0,62	0,23	0,15	7
19	0,54	0,26	0,20	7
12	0,46	0,28	0,26	4
10	0,30	0,31	0,39	4
25	0,78	0,15	0,07	2
24	0,73	0,17	0,10	2
22	0,65	0,22	0,13	5
16	0,51	0,28	0,21	5
11	0,39	0,31	0,30	4

Nämä "teoreettiset" frekvenssit ovat yksi tärkeä viitekehys jonka perusteella tulevia malleja voidaan vertailla keskenään. Ne toimivat myös yhtenä tärkeänä tekijänä prosessissa, jossa määriteltiin kahdeksan eri ryhmää eri tyyppisille otteluille. Tyypillisesti, tietyssä ryhmässä edellä kuvatut tn-osuudet ovat aikalailla yhdenmukaisia keskenään. Toisaalta, ryhmiä muodostettaessa on pyritty käyttämään ennakkotietoja ja käsityksiä. Ryhmä-jaon voi siis sanoa perustuvan pitkälti henkilökohtaiseen käsitykseen ja mielipiteeseen.

Tämän tyyppinen jaottelu on kuitenkin minusta välttämätön tulevia malleja muodostettaessa.

5.1 Mallien estimointi

Mallit muodostetaan siis tulos-muuttujalle (*koti; tp; vieras*) käyttämällä varsinaisessa harjoitustyössä mainittuja selittäjiä eri ryhmissä (1–8). Aluksi selittäjinä käytettiin vain koti- sekä vierasvoittosuhteita ja myöhemmin selittäjiksi lisättiin muita indikaattorimuuttujia.

Miksi sitten on välttämätöntä puuttua indikaattoriselittäjien regressiokertoimiin? Seuraavassa erään osamaallin estimoitu malli ilman mitään rajoitteita

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.6745	0.3822	-1.4235	0.0746	3.11	0.0776
Intercept2	1	0.8158	0.0543	0.7095	0.9222	226.01	<.0001
kekvoittos	1	1.3134	0.5846	0.1677	2.4592	5.05	0.0247
vevvoittos	1	-0.3581	0.5946	-1.5234	0.8072	0.36	0.5470
kpaine	1	0.3045	0.1093	0.0903	0.5188	7.76	0.0053
vpaine	1	-0.4060	0.1049	-0.6115	-0.2004	14.99	0.0001
krivalry	1	-0.2834	0.1688	-0.6143	0.0474	2.82	0.0931
kkampputom	1	-0.0625	0.1790	-0.4133	0.2882	0.12	0.7268
vkampputom	1	0.1579	0.1648	-0.1651	0.4809	0.92	0.3379

Oletetaan nyt, että ollaan aivan pelikauden lopun viimeisissä otteluissa joissa kotijoukkue taistelee rajusti putoamista vastaan ja vastaavasti vierasjoukkue on turvallisesti sarjan puolenvälin tienoilla ja sille ottelulla ei ole suurempaa merkitystä. Tämän kaltaisessa tilanteessa KPAINE=2 ja VPAINE=-1 ja kotijoukkueen tilanteen seurauksena KKAMPPUTOM=1. Näin ollen todennäköisyysmassat eri tulosten (1, X, 2) kohdalla muuttuvat huolimatta KKAMPPUTOM negatiivisesta arvosta voimakkaasti "plus-suuntaan" eli kotijoukkueen voiton todennäköisyys suorastaan räjähtää näiden regressiokertoimien tilanteessa. Paine-muuttujien yhteenlaskettu vaikutus normaalijakauman kertymäfunktion z-arvoon on nimittäin $0.3045 * 2 - 0.4060 * -1 - 0.0625 * 1 = 0.953$ mikä merkitsee riippuen KEKVOITTOS sekä VEVVOITTOS- arvoista jopa lähemmäs 90 prosentin voitontodennäköisyyttä kotijoukkueelle, joka on kaikesta huolimatta aivan valtavasti liian paljon "oikeaan" todennäköisyyteen nähden.

Korjauksia on siis tehtävä. Siksi näiden indikaattoriselittäjien osalta vaikutusosuuksia mallin antamiin arvoihin rajoitetaan voimakkaasti. Tuon

0.953 sijasta, jota voidaan pitää jonkinlaisena ylärajana suurimmalle mahdolliselle muutokselle, muutosmahdollisuus rajoitettiin tilanteesta riippuen lähelle 0.20:tä. Tämä menettely on mielestäni täysin perusteltu tässä yhteydessä kun tarkoituksena on muodostaa mahdollisimman hyviä ja myös tulkinnallisesti järkeviä **ennusteita**.

Tilastotieteellisen teorian kannalta tilanne on hankalempi. Käytännössähän yllä esitetty malli estimoitujen parametrien kautta tuottaa kyseisen aineiston suurimman uskottavuuden estimaattorin ja muuttamalla estimoituja regressiokertoimia tästä optimaalisesta estimaattorista luovutaan ja käyttöön otetaan ei-optimaalinen estimaattori.

5.2 Estimointitulosten arviointia

Mallinnetaan TULOS-muuttujaa nyt eri ryhmien (1 – 8) osalta. Esimerkiksi ryhmissä 1 ja 2 estimoitu ordered probit-malli on

muuttuja	malli 1	malli 2
intercept1	-0,334	0,751
intercept2	0,794	0,566
kekvoittos	0,866	0,003
vevvoittos	-0,679	-0,324

Kaikissa ryhmissä suoritettujen estimointien jälkeen voittos-muuttujat pidetään ennallaan, mutta muiden selittäjien regressiokertoimien suhteen tarkastellaan tilannetta edellä kuvatun ongelman näkökulmasta rajoittamalla niiden vaikutusmahdollisuuksia.

Näiden jälkeen näiden kahden ryhmän mallit ovat muotoa

muuttuja	malli 1	malli 2
intercept1	-0,334	0,751
intercept2	0,794	0,566
kekvoittos	0,866	0,003
vevvoittos	-0,679	-0,324
kpaine	0,075	0,019
vpaine	-0,100	-0,144
kkampputom	-0,015	0,000
kkampmest	0,000	0,081
vkampputom	0,039	0,056
vkampmest	0,000	0,000
kkampkärki	0,000	0,000
vkampkärki	0,000	0,000
krivalry	-0,070	0,000

Otetaan esimerkiksi kauden 2003-2004 aineistosta otteluita

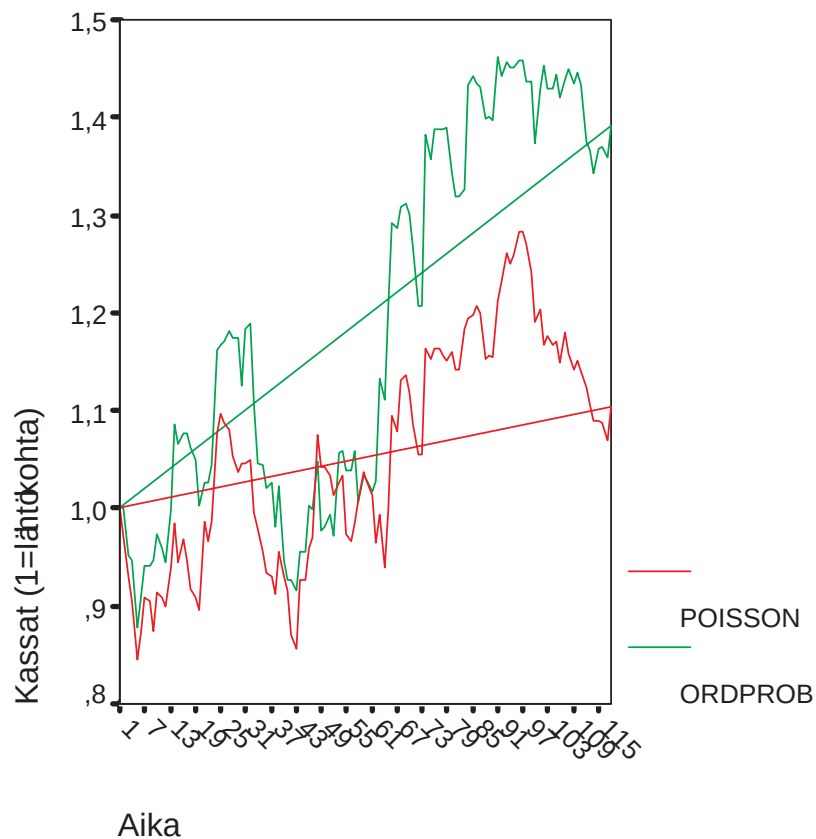
					PO			OP			
kj	vj	km	vm		kv	tp	vv		kv	tp	vv
Liverpool	Everton	0	0		0,52	0,25	0,23		0,47	0,26	0,27
Man City	Chelsea	0	1		0,23	0,26	0,50		0,13	0,21	0,66
Man Utd	Southampton	3	2		0,58	0,24	0,18		0,70	0,20	0,10
Middlesbrough	Blackburn	0	1		0,39	0,29	0,33		0,44	0,30	0,26
Newcastle	Leicester	3	1		0,51	0,25	0,24		0,77	0,15	0,08
Portsmouth	Wolverhampton	0	0		0,59	0,23	0,18		0,54	0,26	0,20
Southampton	Fulham	0	0		0,48	0,27	0,25		0,49	0,29	0,22
Tottenham	Portsmouth	4	3		0,60	0,24	0,16		0,45	0,29	0,26
Wolverhampton	Arsenal	1	3		0,23	0,27	0,50		0,13	0,21	0,66
Arsenal	Charlton	2	1		0,55	0,26	0,19		0,75	0,15	0,09
Aston Villa	Blackburn	0	2		0,41	0,28	0,32		0,49	0,29	0,22
Birmingham	Leicester	0	1		0,42	0,26	0,32		0,48	0,28	0,24
Blackburn	Southampton	1	1		0,43	0,27	0,30		0,44	0,30	0,26
Charlton	Middlesbrough	1	0		0,39	0,26	0,35		0,45	0,30	0,25
Chelsea	Wolverhampton	5	2		0,63	0,24	0,13		0,75	0,14	0,11

Molempien menetelmien hyvyttä voidaan vertailla tarkastelemalla niiden antamien pelisuositusten tuottamia pelikassojen kehittymisiä. Tätä harjoitustyötä varten on käytettävissä ollut keväältä 2004 noin 120 ottelun maailman suurimman vedonlyöntipörssin Betfairin kerroinaineisto. Tämän aineiston perusteella on muodostettu palautusprosentilla 95 otteluille arviot

0.5 maalin tasoitusvedoille niin koti-kuin vierasjoukkueille eli toisin sanoen 4 kerroinsarjaa. Nämä kertoimet ovat siis vain approksimaatioita, valitettavasti todellisia tasoitusvetokertoimia ei ollut käytettävissä. Kuitenkin, on todennäköistä että kerroinasetelma ei suosi kumpaakaan menetelmää.

Näistä neljästä kerroinvaihtoehtoista on valittu jokaisen ottelun kohdalla se pelivaihtoehto, jonka odotusarvo on yli yhden ja kerroin mahdollisimman lähellä kahta. Jälkimmäistä ehtoa on käytetty lähinnä siksi, että käytännön tilanteessa pelivalinta on yleensä samanlainen. Sellaisia otteluita ei olla huomioitu joissa pelivalinnan kerroin on alle 1.5. Panoskoko on muodostettu aiemmin esitellyn Kellyn kaavan mukaisesti jakamalla optimaalinen panoskoko neljälle.

Vaikka edellisistä ehdoista luovuttaisiin, kassojen kehityksissä ei tapahtuisi ratkaisevia muutoksia. Oletuksilla on haluttu vain päästä lähemmäs aitoa päätöksentekotilannetta.



Kuva 5: Pelikassojen kehitys 0,5 maalin tasoitusvetojen kohdalla keväällä 2004

Ordered probit-mallilla muodostettujen ennusteiden tuottama pelikassan kasvu olisi siis ollut huomattavasti parempaa kuin Poisson-mallin vastaava kehitys. Huomattavaa on, että molemmilla tavoilla oltaisiin päädytty voitolliseen peliin. Tuottoluvut ovat jotakuinkin järkeviä, joten ei ole syytä uskoa, että kysymys tai ongelma mallinnuksen ja kerroinaineiston keskinäisestä yhteydestä olisi ratkaiseva. Nämä n.115 ottelua ovat kuitenkin vain noin 7.5 prosenttia koko maallinnukseen käytetystä aineistosta.

Tämä otos on kuitenkin aivan liian pieni osoittaakseen sen kumpi menetelmistä on parempi ja ovatko ne voitollisia pitkällä aikavälillä.

Tulos-muuttujan suhteen voidaan tarkkailla "keskimääräisten" otteluiden tn-arvioiden hyvyttä jakamalla ottelut ryhmä-jaon mukaisesti kahdeksaan eri kategoriaan.

5.3 Mallien vertailu muilla tavoin

Mallien keskenäistä vertailua voidaan suorittaa myös muilla tavoilla. Näitä tekniikoita ei tosin ainakaan allekirjoittaneen tiedossa ole yhtään liikaa, mutta seuraavassa on kuitenkin tarkasteltu mallien antamia sovitteita ja verrattu niitä nyt käytössä olleeseen varsin suureen aineistoon (n=1520).

Toteutuneet frekvenssit

ryhmä	undervv	undertp	underkv	overvv	overtp	overkv
1	0,114	0,223	0,223	0,126	0,070	0,243
2	0,056	0,124	0,292	0,056	0,022	0,449
3	0,260	0,194	0,071	0,321	0,066	0,087
4	0,098	0,264	0,145	0,228	0,078	0,187
5	0,035	0,159	0,221	0,097	0,035	0,451
6	0,139	0,213	0,189	0,172	0,074	0,213
7	0,068	0,130	0,205	0,082	0,027	0,486
8	0,077	0,154	0,215	0,108	0,077	0,369

Poisson-frekvenssit

ryhmä	undervv	undertp	underkv	overvv	overtp	overkv
1	0,128	0,218	0,207	0,141	0,060	0,247
2	0,086	0,197	0,256	0,089	0,049	0,322
3	0,178	0,206	0,135	0,247	0,071	0,164
4	0,144	0,209	0,171	0,185	0,070	0,220
5	0,097	0,199	0,234	0,108	0,057	0,306
6	0,150	0,225	0,188	0,165	0,060	0,212
7	0,096	0,202	0,241	0,102	0,054	0,304
8	0,117	0,217	0,227	0,120	0,055	0,265

Ordered probit-frekvenssit

ryhmä	undervv	undertp	underkv	overvv	overtp	overkv
1	0,099	0,227	0,208	0,140	0,068	0,259
2	0,067	0,119	0,243	0,045	0,027	0,499
3	0,271	0,167	0,074	0,341	0,060	0,087
4	0,122	0,258	0,158	0,204	0,086	0,172
5	0,066	0,156	0,257	0,063	0,040	0,419
6	0,127	0,220	0,183	0,181	0,070	0,219
7	0,087	0,142	0,244	0,075	0,036	0,416
8	0,081	0,180	0,248	0,099	0,051	0,341

Näissä taulukoissa under ja over merkitsevät tilannetta jossa ottelussa tehdään joko alle 2.5 maalia (=under) tai yli 2.5 maalia (=over). Edelleen, vv,tp,kv merkitsevät vierasvoittoa, tasapeliä ja kotivoittoa, joten siis ts. edellä mainitut vaihtoehdot on tässä yhteydessä hajoitettu under ja over-tilanteisiin.

Muodostamalla testisuureet kuvaamaan näiden kahden mallin yhteensopivuutta teoreettisten eli toteutuneiden frekvenssien suhteen. χ^2 -yhteensopivuustestin testisuureiden arvot $X_{orderedprobit} = 19.4$ ja $X_{poisson} = 109.7$. Vapausasteiden lukumäärä on näissä testeissä $(8 - 1) * (6 - 1) = 35$, jolloin testisuureen kriittinen 5 prosentin riskitason arvo on noin 50, joten mikäli tämän 1520-ottelun otoksen toteutuneet frekvenssit eri luokissa ovat "todelliset", niin Ordered probit-malli voidaan hyväksyä malliksi, mutta Poisson-mallia ei. Poisson menettelytapa ei siis tämän testin perusteella sovellu kuvaamaan ilmiötä.

Tulos ei ole yllättävä, sillä op-mallin estimointihan perustuu toteutuneisiin tuloksiin. Poisson-mallissa muutokset maaliodotusarvoissa eivät välttämättä eskaloitu eri tulosten sovite-frekvensseihin riittävällä tavalla. Toisaalta juuri näissä tarkasteluissa havaitaan, että koti- ja vierasjoukkueiden maalimäärät eivät ole kaikissa ryhmissä riippumattomia ja tämä riippumattomuus luultavasti vaihtelee voimakkaasti eri ryhmien kohdalla. Monissa eri ryhmissä on jo nyt niin paljon havaintoja, että mahdolliset korrelaatiot maalimäärien kesken tietäntyyppisissä otteluissa eivät tule näkyviin. Kuitenkin, oletettavaa on että korrelaatio ei ole sittenkään kovin voimakasta, joten myös Poisson-mallinnuksella voidaan johtaa suhteellisen luotettavia arvioita.

Tarkasteltaessa lähemmin edellisten taulokkojen lukuja ja laskemalla yhteen eri under ja over-ryhmät, voidaan havaita että Poisson-mekanismi antaisi jokaisessa ryhmässä yli 50 prosentin todennäköisyyden under-vaihtoehdolle. Tämä ei ole järkevää, sillä jo nyt toteutuneet frekvenssit osoittavat, esimerkiksi ryhmässä 5 under-tn noin 42 prosenttia, että tietyn tyyppisissä otteluissa over-tn ylittää paikoin reilusti under-vaihtoehdon todennäköisyyden. Hyvin suurissa aineistoissa näiden kahden todennäköisyydet ovat olleet hyvin lähellä toisiaan eli 50:tä prosenttia.

Luonnollisesti on täysin mahdollista, että omat maaliodotusarvot eivät ole aivan kohdallaan, mutta on myös ihan selvää, että käytettävä jakauma-approksimaatio on sellainen joka painottaa tätä under-vaihtoehtoa.

Huomattavin yksittäinen tekijä joka aiheuttaa Poisson-mallinnuksen ja toteutuneiden tulosten välille eroja on rank=5 omaavien joukkueiden eli huippujoukkueiden voitonmahdollisuuksiin suhtautuminen. Huippujoukkueet menestyvät otteluissa selvästi paremmin kuin PO-malli antaisi ymmärtää.

5.4 Regressiokertoimien tulkintaa

Seuraavassa hieman eri regressiokertoimien tulkintoja mallin antamiin ennusteisiin.

Kun kotijoukkueena on mestaruudesta pelaava joukkue, hyvin ratkaisevassa osassa on vierasjoukkueen painetila: jos se pelaa putoamista vastaan niin tällön huippujoukkueelle voidaan povata merkittävästi vaikeampaa iltaa kuin sellaista porukkaa vastaan jolla ei pelissä ole suurempaa panosta. Tilanne on vastaavanlainen kun huippujoukkue on vieraana heikkoa kotijoukkuetta vastaan. Tällöin tosin regressiokertoimien valossa ratkaisevempaa on vierasjoukkueen motivaatio eli onko vielä mahdollisuuksia mestaruuteen ym.

Vastaavat tulkinnot kuin yllä pätevät myös kun joukkueista parempi osapuoli taistelee sarjan kärkipään sijoista, eli Valioliigassa UEFA-cup paikoista tms. Tällöin tosin heikomman joukkueen motivaatiotasolla on, kuten olettaa sopiikin, vielä suurempi merkitys.

Ainoastaan silloin kun kotijoukkue on todella heikko (rank=1), niin kotipeleissä pakkovoiton ottaminen laskee aavistuksen verran voitontodennäköisyyttä. Heikot joukkueet siis kestävät paineita heikommin kuin vahvat joukkueet.

Kaikissa ryhmissä paikalliskamppailu tms. tilanteet merkitsevät joukkueiden paremmuuserojen kaventumista. KRIVALRY siis todellakin on merkittävä selittäjä.

Eurocup-muuttujien regressiokertoimien tulkinta on hankalaa, sillä ne eivät ilmeisesti pysty kuvaamaan haluttua vaikutusta huippujoukkueiden suoritustasoon. Kaikesta huolimatta, tämän harjoitustyön yhtenä merkittävänä havaintona voidaan pitää sitä, että eurocup-ottelut eivät häiritse huippuseurojen suoritustasoa kovinkaan olennaisesti jos yleensä ollenkaan. Tämä havainto koskee tosin vain (rank=5) seuroja tyyliin Manchester U, Chelsea ja Arsenal. Aavistuksen verran heikompien seurojen tilanteesta ei voida tehdä luetettavia päätelmiä.

6 Johtopäätökset

Tässä harjoitustyössä on käsitelty jalkapallo-ottelun mallintamista aivan uuden tyyppisellä menetelmällä perinteisen Poisson-jakaumasta lähtevään oletukseen nähden. Ordered probit-mallinnus on varsinkin jatkomallinnuksen perusteella erittäin kilpailukykyinen vaihtoehto.

Lähestymistavat ovat aikalailla erilaiset. Molemmissa perustana on regressiomalli, mutta ratkaiseva ero on siinä, että OP mallintaa ikäänkuin suoraan haluttua tulos-muuttujaa (tai sen johdannaisia) kun vastaavasti PO-mallissa on ensin muodostettava mahdollisimman hyvät maaliodotusarvot ja sen jälkeen käytetään jakauma-approksimaatioita, mikä sinänsä vaikuttaa varsin hyvältä. Jakauman käytössä on kuitenkin ongelmana koti- ja vierasmaalien riippumattomuusoletus. Ilmeistä on, että tämän oletuksen paikkaansapitävyys vaihtelee erityyppisten pelien kohdalla.

Oman erityispiirteen tämän empiirisen ongelman ratkaisemiseen tuo aineiston heterogeenisuus. Tämä on erittäin tärkeä havaita, sillä estimoimalla malli suoraan kokonaisaineistolle tulokset eivät olleet kovinkaan hyviä, kuten harjoitustyön jaksossa 5 voitiin todeta. Jatkomallinnuksessa on pyritty muodostamaan ryhmiä, joihin kuuluvat ottelut käyttäytyisivät mahdollisimman yhtenäisesti. Nämä ryhmät muodostettiin lähinnä ilmiön ennakkotietojen perusteella, mutta tässä yhteydessä voitaisiin ilmeisesti soveltaa myös tilastotieteen monimuuttujamenetelmistä tuttuja ryhmittelyanalyysimenetelmiä. Rank-arvoihin perustuvaa jaotelua voidaan perustella mm. Hillin (1974) ajatusten perusteella.

Millä perusteella sitten eri ryhmissä estimoituja malleja voidaan vertailla Poisson-menetelmän antamiin sovitteisiin. Ryhmittelyjako mahdollistaa erityyppisten yhteensopivuustestien tekemisen olettamalla toteutuneet frekvenssit ikäänkuin "oikeiksi". Näiden lähinnä suuntaa-antavien testien perusteella OP-malli antoi selkeästi parempia tuloksia

Tilastollisesti ajateltuna OP-malli antaa suuremman keskimääräisen toteutuneen todennäköisyyden. Tämä tulos on perusteltu niin kevään 2004 kuin koko kauden 2004-2005 aineistojen perusteella. Tämä tulos on seurausta lähinnä siitä, että OP-malli antaa tyypillisesti suurempia todennäköisyyksiä huippujoukkueille kun vastassa on ennakoita selkeästi heikompia joukkueita niin koti- kuin vierasotteluiden osalta.

Pelisijoittamismielessä tietysti menetelmien paremmuus voidaan ratkaista niiden tuottamien pelikassojen kehittymisten mukaan. Kauden 2004 reilun 100 ottelun otoksen perusteella OP-malli antoi selvästi paremman

pelikassan kasvun kuin PO-malli. Tästä ei kuitenkaan voida tehdä liian pitkälle meneviä johtopäätöksiä. Ensinnäkin, tuo ajankohta sisältyy aineistoon jonka perusteella nämä mallit on muodostettu ja toisaalta otoskoko on aivan liian pieni luotettavien päätelmien tekemiseen. Ilmeistä kuitenkin on, että ordered probit-malli tarvittavine korjauksineen pystyy haastaamaan tai jopa todennäköisesti lyömään Poisson-lähestymistavan pitemmälläkin aikavälillä. Poisson-menetelmän hyvyys sekä heikkous perustuu siihen kuinka hyviä maaliarvot ovat ja siksi muiden harrastajien/ammattilaisten mahdollisesti kehittämät paremmat mallit saattavat muuttaa tilannetta.

Lohduttava tieto niille jotka pitävät tiukasti kiinni Poisson-jakauman soveltamisesta on toki se että tämänkin tutkimus vahvistaa tämän jakauma-approksimaation käyttämisen ja tuloksetkin, kuten tuo pelikassan kasvu osoittaa, olivat varsin hyviä. Poisson-jakauman käyttäminen ja ohjelmoiminen esim. Exceliin on huomattavan helppo tie lähtökohtaprosentteihin, kun taas edes kaikissa tilasto-ohjelmissa ei edes ole OP-mallinnukseen tarvittavia laskentaproseduureja.

Lähteet

Aitchison J. –Silvey S. (1957) : The Generalization of the Probit Analysis to the Case of Multiple Responses. *Biometrika* **44**:131-140.

Davidson R. –MacKinnon J. G. (1993): Estimation and Inference in Econometrics. Oxford University Press.

Daykin A. R. –Moffatt P. G. (2002): Analyzing Ordered Responses: A review of the Ordered Probit Model. *Understanding statistics* **I(3)** :157-166

Dixon M. J. –Coles S.C. (1997) : Modelling association football scores and inefficiencies in the football market. *Applied Statistics* **46**:265-280

Dobson S. –Goddard J. (2001) : The economics of football. Cambridge. Cambridge University Press.

Forrest D. –Simmons R. (2000) : Forecasting sport: the behaviour and performance of football tipster. *International Journal of Forecasting* **16**:317-331

Greene W. H. (2000) : Econometric analysis (4th edition). Upper saddle river. Prentice Hall, New Jersey.

Goddard J. –Asimakopoulos I. (2004): Modelling football match results and the efficiency of fixed-odds betting.

Gould W. (2001) : It is possible to include a constant term in ordered probit model within Stata?. What is the relationship between ordered probit and probit? <http://www.stata.com/support/faqs/stat/constant.html>

Hill I. D. (1974): Association football and statistical inference. *Applied Statistics* **23**:311-323

Koning R. H. (2000) : Balance in competition in Dutch soccer. *The Statistician* **49**:419-431

Maddala G. S. (1983) : Limited-dependent and qualitative variables in econometrics. Cambridge University Press.

Maher M. J. (1982) : Modelling association football scores. *Statistica Neerlandica* **36**:109-118

McFadden D. (1974) : The Measurement of Urban Travel Demand. *Journal of Public Economics* **3**:303-328.

McKelvey W. –Zavoina R. (1975) : A Statistical Model for the Analysis of Ordinal Level Dependent Variables. *Journal of Mathematical Sociology*,. **4**:103-120

Mellin I. (1996) : Johdatus tilastotieteeseen, tilastotieteen jatkokurssi. Helsingin yliopisto.

Moroney M. J. (1956) : Facts from figures, 3 rd. edn. Penguin ,London

Pratt J. W. (1981) : Concavity of the Log-Likelihood. *Journal of the American Statistical Association* **76**:137-159

Raitanen J. (1999) : Jalkapallo-ottelun lopputuloksen tilastollinen mallintaminen. Pro Gradu -tutkielma, Tampereen Yliopisto. (artikkeli Tilastoseuran vuosikertakirjasta vuodelta 2000)

Reep C. –Pollard R. –Benjamin B. (1971)) : Skill and chance in ball games. *Journal of the Royal Statistical Society Series A* **131**:581-585

Rue H. –Salvesen O. (2000) : Prediction and retrospective analysis of soccer matches in a league. *The Statistician* **49**:399-418

Stock J. H. –Watson M. W. (2003) : Introduction to Econometrics. Addison Wesley. Pearson Education Inc.

Vehkalahti K. (2003): Data-analyysi II -kurssimoniste. Helsingin yliopisto.

Vuoksenmaa J. –Kuronen A. –Nåls J. (1999) : Urheiluvetonlyönti, voittajan opas. Gummerus Kirjapaino Oy. Jyväskylä.

**Liite 1. Esimerkki voimaluvuista sekä voittosuhte-
ennusteista ennen kauden alkua kauden 2003-
2004 osalta**

kausi	joukkue	rank	evoimaluku	ekvoittos	evvoittos
03-04	Arsenal	5	28,5	0,84	0,66
03-04	Aston Villa	3	17	0,54	0,36
03-04	Birmingham	3	17,5	0,55	0,37
03-04	Blackburn	3	19	0,63	0,37
03-04	Bolton	2	16,5	0,56	0,31
03-04	Charlton	2	17	0,51	0,38
03-04	Chelsea	5	26	0,81	0,56
03-04	Everton	3	19	0,62	0,38
03-04	Fulham	2	16,5	0,54	0,33
03-04	Leeds	3	18	0,54	0,41
03-04	Leicester	1	12	0,44	0,19
03-04	Liverpool	4	23	0,68	0,53
03-04	Manchester C	3	19	0,58	0,42
03-04	Manchester U	5	28,5	0,83	0,68
03-04	Middlesbrough	3	18,5	0,64	0,33
03-04	Newcastle	4	24	0,78	0,48
03-04	Portsmouth	1	11	0,41	0,17
03-04	Southampton	3	18	0,62	0,33
03-04	Tottenham	3	18	0,57	0,38
03-04	Wolverhampton	1	13	0,48	0,21

Liite 2.Voittosuhteiden estimointitulokset kausien 1997-2004 jälkeen

Linear Regression

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: totkvoittos

Root MSE	0.09535	R-Square	0.5107
Dependent Mean	0.61689	Adj R-Sq	0.5022
Coeff Var	15.45699		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.01317	0.05851	0.23	0.8223
ekvoittos	F6	1	1.34838	0.15650	8.62	<.0001
edkvoittos	F35	1	-0.39872	0.13149	-3.03	0.0030

Dependent Variable: totvvoittos

Root MSE	0.08529	R-Square	0.6695
Dependent Mean	0.41681	Adj R-Sq	0.6638
Coeff Var	20.46233		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.02594	0.02816	0.92	0.3588
evvoittos	F7	1	1.34165	0.13183	10.18	<.0001
edvvoittos	F36	1	-0.39277	0.12042	-3.26	0.0015

Sarjanousijoiden osalta vastaavat estimointitulokset

Dependent Variable: totkvoittos

Root MSE	0.13499	R-Square	0.1884
Dependent Mean	0.49200	Adj R-Sq	0.0929
Coeff Var	27.43795		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.02403	0.28621	0.08	0.9341
ekvoittos	F6	1	0.59031	0.39947	1.48	0.1578
edkvoittos	F35	1	0.24605	0.39930	0.62	0.5459

Dependent Variable: totvvoittos

Root MSE	0.09254	R-Square	0.2845
Dependent Mean	0.31400	Adj R-Sq	0.2003
Coeff Var	29.47257		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.12787	0.20143	0.63	0.5340
evvoittos	F7	1	0.67510	0.27856	2.42	0.0268
edvvoittos	F36	1	0.04642	0.34054	0.14	0.8932

Liite 3.SAS-koodia ordered-probit-mallille

```
PROC IMPORT OUT= SASUSER.AINEISTO
            DATAFILE= "R:\Probit\htaineisto.xls"
            DBMS=EXCEL REPLACE;
            SHEET="aineisto 00-04$";
            GETNAMES=YES;
            MIXED=NO;
            SCANTEXT=YES;
            USEDATE=YES;
            SCANTIME=YES;
run;

proc means data=sasuser.aineisto;
run;

PROC CORR
DATA=sasuser.aineisto;
run;

proc sort data=sasuser.aineisto;
  by descending tulos;
run;

proc probit data=sasuser.aineisto order=data;
  class tulos;
  model tulos = kekvoittos vevvoittos e / itprint d=normal;
  title 'Probití-model for English Premier League ';
  output out=probitdat prob=yhatprobit;
run;
```