Stockholms universitet

Botten av isberget – OBNR Estimering: En tillämpad litteraturöversikt

Patrik Emanuelsson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2009:8
Matematisk statistik
September 2009

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2009:8**,
http://www.math.su.se/matstat

Botten av isberget - OBNR Estimering: En tillämpad litteraturöversikt

Patrik Emanuelsson*

September 2009

Sammanfattning

Historiska data av inträffade händelser karaktäriseras av en förfluten tid mellan företeelse och observation. Eftersom det endast är möjligt att observera händelser fram till och med den senaste mätningen kommer de rapporterade händelserna efter denna tidpunkt inte vara observerbara, således sker en underskattning av det faktiska antalet. Händelser som har inträffat men ej rapporterats förekommer inom en rad och i vitt skilda områden. Litteraturen är mest fokuserad på reservering av oreglerade skulder hos försäkringsbolag och i samband med utvärdering av smittospridning, i synnerhet AIDS. Denna litteraturöversikt beskriver ett antal modeller som kan användas för reglering av inträffade men ej rapporterade händelser. Dessa modeller är alla oberoende av tillämpningsområde och kan därför vara väldigt användbara. En utvärdering av modellerna visar att resultaten blir väldigt snarlika. Detta är en följd av att poisson och multinomialfördelade stokastiska variabler ger samma maximum likelihood skattning av fördröjningsfördelningen mellan inträffande och rapportering. Den huvudsaklig skillnad visar sig ligga i kategoriseringen av data. Det vill säga om data är grupperat efter inträffande och fördröjning eller om hela datamaterialet i form av exakta datum används, samt om endast en delmängd av de observerade händelserna används för att erhålla skattningarna. Avslutningsvis illustreras uppräkning av inträffade men ej rapporterade händelser med två variationer av den poisson log-linjära modellen tillämpade på Statistiska Centralbyråns lagfartsstatistik. Arbetet har utförts på uppdrag av Statistiska Centralbyrån eftersom de kan dra nytta av en minskad fördröjning i statistikrapporteringen.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: patrik.emanuelsson@gmail.com. Handledare: /Åke Svensson.

Abstract

Time to events data is characterized by an elapsed time between the occurrence and observation of an event. Following, it is only possible to observe events up to and including the most recent date of measurement. Events reported subsequent to this date cannot be observed, which leads to an underestimation of the actual numbers. Events that have occurred but not yet been reported arise in numerous and diverse areas. The literature is mostly focused around claims reserving by insurance companies and in connection with the evaluation of infectious diseases, in particular AIDS.

This report deals with a number of models that can be used for adjustment of events that have occurred but have not yet been reported. These models are all independent of the area of application and can therefore be very useful. Evaluations of the models show very similar results. This is a consequence of the fact that poisson and multinomial distributed random variables eventuate in the same maximum likelihood estimates of the delay distribution between occurrence and reporting. The main difference appears to be in the categorization of data. That is, if the data is grouped by occurrence and delay or exact dates and if only a subset of the observed events or all the data are used to obtain estimates. At the end of the report a grossing-up of events that have occurred but not been reported are illustrated with two variations of the poisson log-linear model applied to Statistic Sweden's land registration statistics. The work has been commissioned by Statistics Sweden since they can benefit from a reduction in delays in reporting statistics.

Förord

Denna uppsats utgör ett examensarbete om 15 högskolepoäng och leder till en kandidatexamen i matematisk statistik vid Matematiska institutionen, Stockholms Universitet. Examensarbetet har utförts på uppdrag av utvecklingsavdelningen på Statistiska Centralbyrån.

Jag vill rikta ett stort tack till båda mina handledare professor Åke Svensson vid Matematiska institutionen på Stockholms Universitet och Dan Hedlin från Statistiska Centralbyrån för all vägledning och hjälp utefter arbetets gång.

Patrik Emanuelsson

Stockholm, 2009-09-12

Innehåll

1 Inledning **5**

1.1 Bakgrund 5

1.2 Syfte 8

1.3 Metod 8

2 Modeller **9**

2.1 Icke parametriska modeller 9

2.1.1 Lawlessmodell 9

2.1.2 Trunkeringsmodell 13

2.1.3 Link ratio och Chain ladder metoden 17

2.2 Parametriska modeller 20

2.2.1 Tvåsidig variansanalysmodell, modelltyp I 22

2.2.2 Poisson log-linjär modell 25

3 Tillämpning **30**

3.1 Fastighetsprisstatistik 30

3.1.1 Beskrivning av data 31

3.1.2 Modellering 32

4 Sammanfattning **36**

5 Diskussion **37**

A Appendix **39**

A.1 Härledning 39

A.2 Figurer 40

A.3 Ordlista 43

Referenser **44**

1 Inledning

1.1 Bakgrund

I företag och myndigheter är det viktigt att ha tillgång till aktuell och tillförlitlig information som beslutsunderlag, för att kunna fullfölja åtaganden mot intressenter och effektivisera allokering av befintliga resurser. Vid insamling av data kan det föreligga ett moment av fördröjning, mellan att en händelse inträffar och tills dess att denna rapporteras till berörd part. Följden blir att vid en given tidpunkt har händelser inträffat men ännu ej rapporterats d.v.s. data är till höger trunkerade, beskrivet nedan (Figur 1). För att kunna agera utifrån befintlig information finns ett intresse av att skatta omfattningen av inträffade händelser som ej har rapporterats samt att försöka prediktera framtida händelser. Denna typ av information kan generellt benämnas OBNR, ”Occurred But Not Reported”.

$T_i =$ Tidpunkt för primär händelse $t_i,\ i = 1, \ldots, n$.
$X_i =$ Förfluten tid till sekundär händelse, fördöjning eller rapportförsening $x_i,\ i = 1, \ldots, n$.

OBNR-data erhålls i form av par av tidpunkter $(T_i, T_i + X_i),\ i = 1, \ldots, n$, uppåtbegränsade av det sista observationstillfället $\tau \geq T_i + X_i$. Och grupperas efter lämplig kategorisering av inträffande och fördröjning i antal dagar, veckor, månader eller år, beroende på fördröjningarnas fördelning och stickprovets detaljrikedom. Grupperingen är nödvändigtvis inte densamma för både T och X.

Vid studier av relationen mellan två kategorivariabler kan dess förhållande till varandra sammanfattas i form av en kontingenstabell. Med celler representerande utfall av en given kombination kategorier mellan variablerna. Ibland kan det förekomma tomma celler, antingen till följd av stickprovets variabilitet och att sannolikheten för en händelse klassificerad av denna cell är relativt liten (sampling zero), eller att cellen av undersökningens natur inte kan observeras (structural zero). Det förstnämnda fenomenet kan undvikas genom att öka stickprovets storlek, medan cellen för den senare förblir tom (Bishop et al. (1975)). Om vi dessutom observerar försöksobjekten över ett tidsintervall kan även objekten utsättas för censurering eller trunkering. I överlevnadsanalys innebär censurering att en livstid endast är känd inom ett givet tidsintervall och exakt tidpunkt för en händelse exempelvis dödsfall inte är observerbar till följd av censurering. Trunkering inträffar när försöksobjekt har utsatts för en händelse utanför studiens ram, händelsens existens är därför ej observerbar. Distinktionen mellan trunkerad och censurerad data är att man för censurerad observerar samtliga försöksobjekt oberoende om händelsen inträffar eller ej medan för trunkerad observerar endast objekten för vilka händelsen har inträffat (Klein och Moeschberger (1997)).

Det finns en rad områden som är utsatta för denna typ av undervärdering av antal inträffade händelser till följd av begränsning av observationstid till dagens datum och högertrunkering. Olika tillämpningsområden kan också tänkas vara associerade med områdesspecifika problem så som inflation, ekonomisk aktivitet, företagens storlek, till händelsens associerade belopp eller liknande.

Vad är en händelse?
En händelse är ett inträffande som i sig är belägen i tiden.

Företagsintroduktioner

Vid undersökning av antalet verksamma företag på en marknad påträffas två huvudsakliga problem. Först, nybildade företag har inte hunnit inkluderas i nyttjat register, en födelsefördröjning. Det andra problemet uppstår då företag som ej längre existerar fortfarande finns kvar i registret, en dödsfördröjning. Detta medför att den verkliga mängden verksamma företag under en given period inte är känd. Om ett företag inte inkluderats kommer det kända antalet vara mindre än det verkliga d.v.s en underskattning. Och likaledes det omvända då företag till följd av dödsfördröjning inte har exkluderats, en överskattning (Hedlin et al. (2006)).

Behållning av reserver, IBNR

Om en individ har försäkring kommer en händelse under täckning leda till en ”försäkrad förlust” och följdaktligen ett anspråk på försäkringsgivare. För anspråket föreligger nödvändigtvis en fördröjning mellan tidpunkt för händelsens inträffande och tills dess att ärendet har avslutas hos försäkringsbolaget. Eftersom oreglerade anspråk finns måste försäkringsgivare behålla reserver i syfte att uppfylla dessa förpliktelser vilket även är lagstadgat. Vid en given tidpunkt finns två typer av fordringar på försäkringsgivaren, händelser som inträffat och har rapporterats samt händelser som har inträffat men ännu ej rapporterats, de senare går under beteckningen IBNR (Incurred But Not Reported) (Fac (1997)).

AIDS

För smittsamma sjukdommar är det nödvändigt att veta den verkliga omfattningen och kunna göra en precis framställning av antal inträffade fall för att kunna planera och vidta åtgärder på kort och lång sikt. Efter diagnos av läkare eller laboratorium existerar en tidsfördröjning tills dess att den ansvariga centrala myndigheten tar del av diagnos. Om denna fördröjning ignoreras kommer man att strukturellt underskatta det verkliga antalet diagnostiserade sjukdomsfall (Sellero et al. (1996)). Om smitta dessutom kan

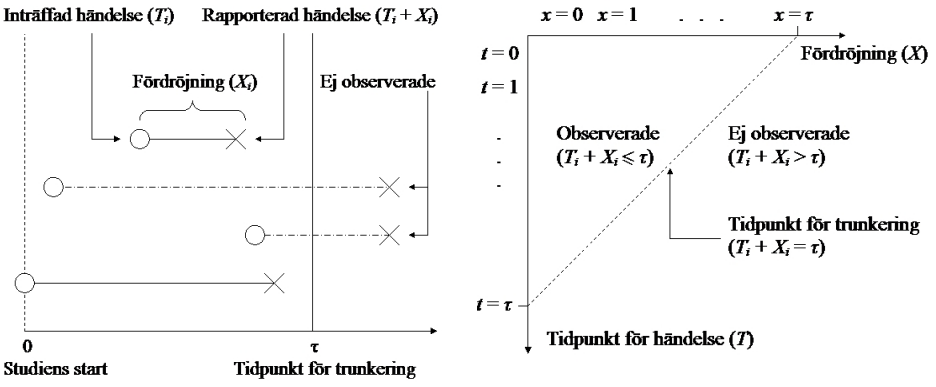
härledas till specifik tidpunkt för infektion kan utvärdering av antal smittade men ej diagnostiserade fall göras.

Garantianspråk

Tillverkare av produkter som omfattas av en garanti lagrar data över produktens historik av anspråk och relaterade problem. Denna information används sedan för att skatta hur många anspråk som uppkommit per tillverkad enhet, för att bedöma framtida anspråk, när dessa uppstår och kostnad för produkt serier. Med dessa data kan man även utvärdera prestanda och om man eventuellt behöver förbättra produkterna i något avseende (Kalbfleisch et al. (1991b)).

Produktion av naturgas

I USA förväntar man sig att konsumtionen av naturgas stadigt kommer att öka snabbare än någon annan energiresurs under de närmaste 30 åren. På grund av detta finns ett ökat intresse av att få tillförlitlig information avseende hur stort det befintliga utbudet är. Den centrala myndigheten EIA (Energy Information Agency) som tillhör energi departementet samlar data från delstaterna genom en månatlig undersökning där de angivna produktionstalen oftast är en underskattning av de verkliga. Generellt är inte den faktiska produktionen känd förrän med ett års eftersläpning. Den initiala prognosen för månadens produktion uppdateras fram tills dess att inga vidare förändringar kan påvisas och man antar att denna nivå är den sanna produktionen (Linkletter och Sitter (2007)).



Figur 1: Till höger trunkerad data till följd av rapportfördröjning.

1.2 Syfte

- Litteraturöversikt av modeller behandlande punktskattning av strukturellt tomma celler för händelser som har inträffat men ännu ej rapporterats.
- Jämförelse av likheter och skillnader mellan modeller.
- Illustration av skattningar på ett faktiskt datamaterial.

När man avser att göra punktskattningar är det inte rimligt att dessa ska prediktera händelser utanför den senast observerade fördröjningen ($x > \tau$). Tillgänglig data behåller inte någon information avseende dessa och kommer därav att vara mycket osäkra. Vidare är den huvudsakliga målsättningen inte nödvändigtvis riktad mot skattning av cellspecifikt värde (T_i, X_j), utan snarare för det totala antalet vid varje tidpunkt för inträffande.

1.3 Metod

Arbetet har utförts med utgångspunkt i artikel av Hedlin et al. (2006), genom ställa upp en referensram av möjliga modeller för att sedan söka centrala artiklar i referenslistor och via databasen Web Of Science.

2 Modeller

2.1 Icke parametriska modeller

Icke-parametriska eller fördelningsfria modeller innebär att man inte apriori antar en specifik fördelning för studerad population. Detta kan vara fördelaktigt då man inte är beroende av lika många underliggande antaganden som nödvändigtvis inte är i linje med observerad data.

2.1.1 Lawlessmodell

Data observerade för en period av längd τ ($0 \leq x + t \leq \tau$), där antal inträffade händelser vid en given tidpunkt $t \geq 0$ och rapporterade med en fördröjning $x \geq 0$ kan illustreras med en till höger avhuggen kontingenstabell definierad av $n_{tx} \in \Delta,\ \Delta = \{n_{tx} |\ t = 0, \ldots, \tau\ x = 0, \ldots, \tau - t\}$ (Tabell 1). Problemet består i att försöka skatta antalet icke observerade händelserna till höger om diagonalen d.v.s. för alla tidpunkter τ_1 som uppfyller villkoret $\tau_1 = t + x > \tau \geq x$.

			$x=0$	$x=1$	$\ldots$	$x=\tau-1$	$x=\tau$
$x \leq \tau$	$\Leftarrow$	$t=0$	n_{00}	n_{01}	$\ldots$	$n_{0(\tau-1)}$	$n_{0\tau}$
$x \leq \tau-1$	$\Leftarrow$	$t=1$	n_{10}	n_{11}		$n_{1(\tau-1)}$	-
	$\vdots$		$\vdots$			$\ddots$	$\vdots$
$x \leq 1$	$\Leftarrow$	$t=\tau-1$	$n_{(\tau-1)0}$	$n_{(\tau-1)1}$	-		-
$x \leq 0$	$\Leftarrow$	$t=\tau$	$n_{\tau 0}$	-	-	$\ldots$	-

Tabell 1: Tillgänglig data för n_{tx} där $t, x \geq 0$ och $t + x \leq \tau$

Beteckningar

n_{tx} = Antal inträffade händelser vid tidpunkt t, rapporterade med fördröjning x i period $t + x$.

N_{tx} = Sammanlagt antal inträffade händelser vid tidpunkt t, rapporterade med fördröjning upp till och med x, $\sum_{u=0}^{x} n_{tu} = N_{tx},\ x = 0, \ldots, \tau - t$

$f_t(x)$ = Sannolikhet för att en händelse inträffar vid tidpunkt t och rapporteras med fördröjning x i period $t + x$.

$F_t(x)$ = Kumulativ sannolikhet att en händelse inträffar vid tidpunkt t och har rapporterats vid fördröjning x.

$g_t(x)$ = Betingade sannolikheten att en händelse rapporteras med fördröjning x givet rapportering upp till och med fördröjning x.

Antaganden

1. Givet tidpunkt för inträffande $t = 0, \ldots, \tau$ är antal händelser i varje cell n_{tx} oberoende för alla $x = 0, \ldots, \tau - t$.

2. Sannolikheterna $g_t(x)$ är stationära för de $m + 1$ senaste perioderna och in i framtiden, dvs $g_t(x) = g(x) \quad \tau - m \leq t + x \leq \tau$

3. Rapportförseningarna för händelser inträffande i olika tidsperioder t är oberoende.

Punktskattningarna eller närmare bestämt prediktionerna av det totala antalet händelser som rapporterats vid kalendertidpunkten τ_1 görs i linje med Lawless (1994). En uppräkning av observerat antal sker genom att de relateras till kvoten $W_t = F_t(\tau - t)/F_t(\tau_1 - t)$ av kumulativa sannolikheter. Detta förfaringsätt kan liknas vid problemet att beräkna storleken av ett isberg då endast toppen är synlig och givet att proportionen av hela berget toppen utgör är känd. För vår del innebär detta att metoden är applicerbar om kvoten W_t, proportionen av sannolikheten för observerade händelser i jämförelse med sannolikheten för totalt antal inträffade händelser fram till och med $x = \tau$ går att skatta. Tidpunkten τ_1 kan lämpligen sättas så att för varje $t \leq \tau,\ x = \tau,\ t + \tau = \tau_1$ men dock inte större då tillgänglig data inte behåller någon information om dessa fördröjningar och skattningarna blir därav ej tillförlitliga.

$$\hat{N}_{t(\tau_1 - t)} = N_{t(\tau - t)} \frac{\hat{F}_t(\tau_1 - t)}{\hat{F}_t(\tau - t)} = \left\{ \hat{W}_t = \frac{\hat{F}_t(\tau - t)}{\hat{F}_t(\tau_1 - t)} \right\} = \frac{N_{t(\tau - t)}}{\hat{W}_t} \tag{1}$$

För att det nu ska vara möjligt att göra punktskattningar behöver vi veta något om hur rapportförseningarna fördelar sig. Eftersom $f_t(x)$ inte är observerbara för $x > \tau - t$ definierar vi de betingade sannolikheterna $g_t(x)$ för rapportering av händelser i period $t + x$, normerade med den kumulativa sannolikheten för samma period.

$$g_t(x) = \frac{f_t(x)}{F_t(x)} \quad x = 0, 1, \ldots, \tau \tag{2}$$

$$1 - g_t(x) = \frac{F_t(x - 1)}{F_t(x)} \quad x \neq 0 \tag{3}$$

$$\frac{F_t(x)}{F_t(\tau)} = \prod_{r=x+1}^{\kappa} [1 - g_t(r)] \quad 0 \leq x < \tau \tag{4}$$

Till följd av egenskapen hos kategorivariabler använder vi att den simultana fördelningen för antalen $\mathbf{n}_t = \{n_{tx} : \ (x = 0, \ldots, \tau - t)\}$ givet det sammanlagda antalet $N_{t(\tau - t)}$ observerade händelser är multinomialfördelad med sannolikheter $f_t(x)/F_t(\tau - t)$. Lawless (1994) noterar att fördelningen för de $m + 1$

sist observerade antalen n_{tx} givet $N_{t(\tau-t)}$ kan nu skrivas som (5) nedan (se Appendix A.1).

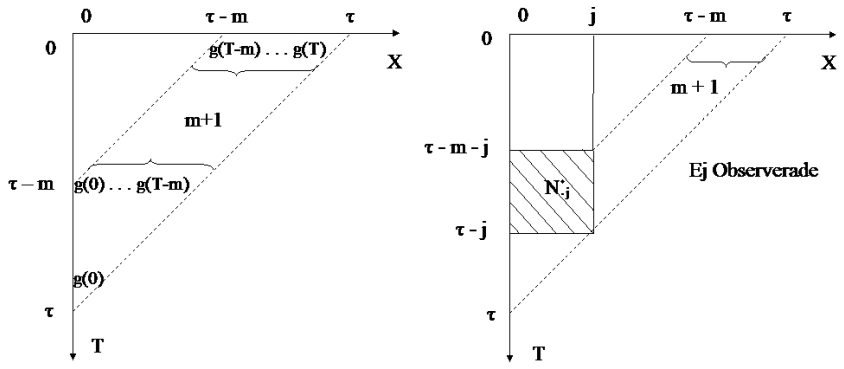
$$\tau - m \leq t + x \leq \tau \quad \Leftrightarrow \quad \tau - m - t \leq x \leq \tau - t$$

$$L_t = \prod_{x=max\{0,\tau-m-t\}}^{\tau-t} \binom{N_{tx}}{n_{tx}} g_t(x)^{n_{tx}} \left[1 - g_t(x)\right]^{N_{tx}-n_{tx}} \tag{5}$$

För att vi sedan ska kunna använda fördelningen av rapportförseningar måste ett antagande beträffande relationen mellan sannolikheterna $g_t(x)$ och tidpunkterna av inträffad händelse införas. Det är troligt att $g_t(x)$ förändras med tiden, exempelvis till följd av strukturella förändringar i rapporterings systemet, politiska faktorer eller ökad ekonomisk aktivitet. För att fördelnings antagandet inte ska vara alltför starkt använder vi antagande 2, sannolikheterna är stationära för $t + x > \tau - m$, det vill säga för de senaste $m + 1$ perioderna och inte för hela observerade intervallet $(t + x) \in [0,\ \tau]$,

$$g_t(x) = g(x) \quad t + x \geq \tau - m. \tag{6}$$

Detta tillåter sannolikheterna $g_t(x)$ att variera medan beroendet av observationer långt i det förflutna minimeras.



Figur 2: Antagande om stationäritet för de senaste $m + 1$ perioderna.

Via antagande 3 får vi likelihood $L = \prod_{t=0}^{\tau} L_t$ för hela stickprovet,

$$L \propto \prod_{x=1}^{\tau} g(x)^{n^*_{\cdot x}} \left[1 - g(x)\right]^{N^*_{\cdot x} - n^*_{\cdot x}} \tag{7}$$

där

$$n^*_{\cdot x} = \sum_{t=max\{0,\, \tau-m-x\}}^{\tau-x} n_{tx} \tag{8}$$

och

$$N^*_{\cdot x} = \sum_{t=max\{0,\tau-m-x\}}^{\tau-x} N_{tx}. \tag{9}$$

Med rapportfördröjning $X = j$ utgörs $N^*_{\cdot j}$ av den skuggade arean i Figur 2 ovan, där ”*” betonar införandet av stationäritets antagande för ett begränsat intervall av observerad period. Utan antagande 2 utgörs $N_{\cdot x}$ av hela rektangeln av observerade händelser $\{n_{tx} |\ 0 \leq t \leq \tau - j,\ 0 \leq x \leq j\}$. Maximering av likelihood (7) ger oss skattningarna,

$$\hat{g}(x) = \frac{n_{\cdot x}}{N_{\cdot x}} \quad x = 0, \ldots, \tau \tag{10}$$

och

$$\hat{W}_t = \prod_{x=\tau-t+1}^{\tau_1-t} \left[1 - \hat{g}(x)\right] \tag{11}$$

samt

$$\hat{N}_{t\tau} = \frac{N_{t(\tau-t)}}{\hat{W}_t}. \tag{12}$$

Den totala antalet av inträffade händelser beräknar vi sedan genom att summera $\hat{N}_{t\tau}$ i ekvation (12) över alla observerade tidsperioder $t = 0, \ldots, \tau$.

2.1.2 Trunkeringsmodell

När man i efterhand kan avgöra vilken tidpunkt som en individ har infekterats av ett virus är det av centralt intresse att skatta fördröjningen till utbrott och diagnos, för att sedan kunna avgöra den verkliga förekomsten av antal fullt insjuknade fall. Tidpunkten för infektion är exempelvis känd då den kan härledas till överföring av virus via blodtransfusion. I studier med medicinsk anknytning relateras problemet till överlevnadsanalys och skattning av riskfunktionen i omvänd tid, detta görs med fördel om hela registret av infektion och rapportering är tillgängligt, då all information kan tillvaratas (Sellero et al. (1996)).

Beteckningar

$d_i =$ Antal händelser vid tidpunkt x_i.
$Y_i =$ Antal individer som riskerar att utsättas för en händelse vid tidpunkt x_i.
$S(x) =$ Överlevelsefunktion, sannolikhet att en individ överlever tidpunkt x.
$h(x) =$ Riskfunktion eller riskfrekvens, sannolikhet att en individ utsätts för en händelse nästa ögonblick givet att han har överlevt till tidpunkt x.

Antaganden

1. Observerade händelser (t_i, x_i) är oberoende realisationer av (T_i, X_i).
2. Stationäritet av sannolikheter för hela observerade perioden.

$$S(x) := P(X > x) \quad \text{och} \quad h(x) := \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x | X \geq x)}{\Delta x} \tag{13}$$

Om tidpunkterna x_i av rapporterade händelser är diskreta med sannolikhetsfunktion $p(x_i)$, $i = 1, \ldots, n$ där $x_1 < x_2 < \ldots < x_n$ kan överlevnadsfunktionen $S(\cdot)$ och riskfunktionen $h(\cdot)$ skrivas som,

$$S(x) = \sum_{x_i > x} P(X = x_i) \quad \text{och} \quad h(x_i) = P(X = x_i | X \geq x_i) = \frac{p(x_i)}{S(x_{i-1})}. \tag{14}$$

$$S(x_i) = \Big\{ S(x_0) = 1 \Big\} = \frac{S(x_1)}{S(x_0)} \frac{S(x_2)}{S(x_1)} \cdots \frac{S(x_i)}{S(x_{i-1})} = \prod_{x_j \leq x_i} \Big[1 - h(x_j) \Big] \tag{15}$$

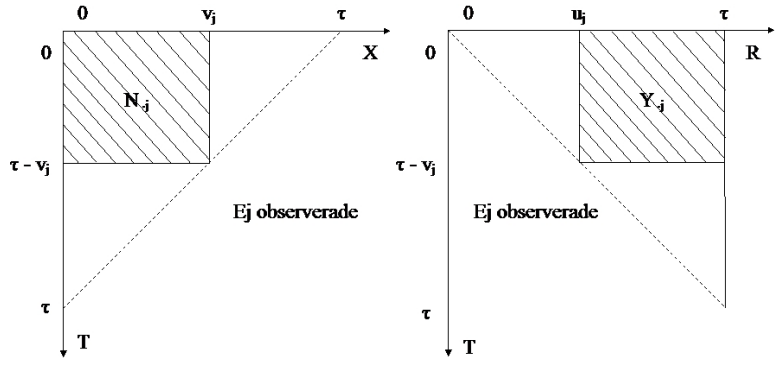
För att sedan skatta $\hat{S}(\cdot)$ använder vi kvoten d_i / Y_i som den betingade sannolikheten $\hat{h}(\cdot)$, vilket är ML skattningen av $h(x_i)$ (Kalbfleisch och Lawless (1991a)). Implementerar vi denna i (15) resulterar skattningen (16) nedan,

vilken går under namnet ”product-limit estimator” eller ”Kaplan-Meier estimator” (Klein och Moeschberger (1997)).

$$\hat{S}(x) = \begin{cases} \prod_{x_i \leq x} \left[1 - \frac{d_i}{Y_i}\right], & \text{om} \quad x_1 \leq x \\ 1, & \text{om} \quad x_1 > x. \end{cases} \tag{16}$$

Observera att om $d_i = Y_i$ för något $x_i \leq x$ kommer följdaktligen $\hat{S}(x_j) = 0$ för alla $j \geq i$.

OBNR-data är nödvändigtvis högertrunkerade med observationstid $(0, \tau)$ och skattning av $S(\cdot)$ görs lämpligen med metod för vänstertrunkerade data. Genom att först transformera fördröjningarnas tidsaxel med $R_i = \tau - X_i$ överförs trunkeringen till vänster i bemärkelsen att vi endast kan observera $R_i \geq T_i$ (figur. 3). Om vi har n oberoende realiserade par av



(a) Höger trunkerad, $T_i + X_i \leq \tau$. (b) Vänster trunkerad, $T_i \leq R_i$.

Figur 3: Transformering av tidsaxel för fördröjning, $R_i = \tau - X_i$

(T_i, X_i), $i = 1, \ldots, n$ där $v_1 < v_2 < \ldots < v_m$ är distinkta värden på X_i och $u_i = \tau - v_i$ så att $u_1 > u_2 > \ldots > u_m$ är distinkta värden på R_i då händelsen observeras eller blir rapporterad.

$$d_j = \sum_{i=1}^{n} \mathrm{I}(R_i = u_j) = \sum_{i=1}^{n} \mathrm{I}(X_i = v_j) = n_{\cdot j} \tag{17}$$

d_i är alltså antalet observerade händelser i rapportförsenings kolumn $x = v_i$, vilket med tidigare beteckning är likaställt med $n_{\cdot x}$.

$$Y_j = \sum_{i=1}^{n} \mathrm{I}(T_i \leq u_j \leq R_i) = \sum_{i=1}^{n} \mathrm{I}(X_i \leq v_j \leq \tau - T_i) = N_{\cdot j} \tag{18}$$

Antalet Y_i som riskerar att utsättas är i ursprunglig skala, de observerade antalet händelser innan fördröjning v_j och med tidpunkt för inträffande före

$\tau - v_j$ (figur 3).

Eftersom vi endast kan observera fördröjnings fördelningen $F(\cdot)$ för ett begränsat tids intervall $(0, \tau)$, kan endast aspekten (19) av $F(\cdot)$ identifieras utan att införa ytterligare overifierbara antaganden (Lagakos et al. (1988), Kalbfleisch och Lawless (1991a)).

$$G(x) := \frac{F(x)}{F(\tau)} \quad \text{och} \quad g(x) := \frac{dG(x)}{dx} = \frac{f(x)}{F(\tau)} \quad x \leq \tau \tag{19}$$

Eftersom vi har omdefinierat antalet individer Y_j som riskerar att utsättas med hänsyn till vänster trunkering, får vi att Kaplan-Meier skattningen av överlevnasfunktionen nu tolkas som,

$$P(R > r|\ R \geq 0) = P(X < x|\ X \leq \tau) = G(x)$$

(Klein och Moeschberger (1997)). Och åter transformerad i ursprunglig ursprunglig skala,

$$\hat{G}(x) = \begin{cases} \prod_{v_j \geq x} \left[1 - \frac{d_j}{Y_j}\right], & \text{om} \quad 0 \leq x \leq v_m \\ 1, & \text{om} \quad v_m < x \leq \tau. \end{cases} \tag{20}$$

(Lagakos et al. (1988)) där d_i/Y_i är ML skattningen av $g(x_i)$ (Kalbfleisch och Lawless (1991a)). Riskfunktionen i omvänd tid $g(x)$ ovan är sannolikheten (2) presenterad under Lawlessmodell och dess kumulativa motsvarighet $G(x)$ i (4). Dock med skillnaden att man i grundmodellen endast antar stationäritet för sannolikheterna $g_t(x)$ i det begränsade intervallet av $m + 1$ senaste tidsperioderna. För att slutligen skatta hela populationen N föreslår Sellero et al. (1996) att om man har n realiserade par av $(\tau - T_i, X_i)$ kan man använda ekvation (21) nedan,

$$\hat{N}_n = \sum_{i=1}^{n} \frac{1}{\hat{G}(\tau - t_i)}. \tag{21}$$

Intuitivt betyder ekvation (21) att varje rapporterat fall ska räknas upp med en faktor av $1/\hat{G}(\tau - t_i)$. Detta inses om vi här inför beteckningarna $A_i,\ i = 0, 1, \ldots, \tau$ där $A_0 = 0$ för de icke observerbara antalet inträffade händelser i varje rad och noterar att de observerade antalen är binomial fördelade.

$$N_{t(\tau - t)} + A_t = N^t \tag{22}$$

$$N_{t(\tau - t)} \sim \text{Bin}(N^t, p_t) \quad t = 0, 1, \ldots, \tau \tag{23}$$

$$p_t = G(\tau - t) \quad \text{och} \quad \hat{p}_t = \frac{N_{t(\tau - t)}}{N^t}$$

n_{00}	n_{01}	$\dots$	$n_{0(\tau-1)}$	$n_{0\tau}$
n_{10}	n_{11}	$\dots$	$n_{1(\tau-1)}$	A_1

$\vdots$

$n_{(\tau-1)0}$	$n_{(\tau-1)1}$	$A_{\tau-1}$
$n_{\tau 0}$	A_τ	

$$\hat{N} = \sum_{t=0}^{\tau} \hat{N}^t = \sum_{t=0}^{\tau} \frac{N_{t(\tau-t)}}{\hat{G}(\tau - t)} \tag{24}$$

Ekvation (21) reduceras till (24) om det för varje t_i observerats $N_{t_i(\tau-t_i)}$ händelser. Om man i Lawlessmodell antar stationäritet av sannolikheter för hela den observerade perioden och med beräkningar baserade på trunkerad kontingenstabell resulterar att skattningarna $\hat{N}^t$ är identiska med $\hat{N}_{t\tau}$ i (12) och därav även $\hat{N}$. Vidare noterar även Sellero et al. (1996) att populationsskattningen är konsistent.

Skillnader i skattning av populationen uppstår därför vid antagande om stationäritet och om man använder fullständig information avseende datum för händelsens inträffande och rapportering. Fördelen med att använda (21) är att man undgår homogenitets antagande av ankomstintensitet under antagande om poisson fördelning (Kaminsky (1986)), diskretisering av data och relaterade grupperings problem beskrivet i kommande avsnitt.

2.1.3 Link ratio och Chain ladder metoden

I försäkringsbranschen är en av de äldsta, mest frekvent använda och vida omdiskuterade metoderna den så kallade Chain ladder metoden som är ett specialfall av Link ratio metoden, för beräkning IBNR reserver. Ursprunget till metodens popularitet är på grund av dess enkla och intuitiva struktur. Och bygger på utvecklingsmönstret av det kumulativa antalet händelser mellan rapport fördröjningar, vilket kan användas i olika variationer. Chain ladder metoden används med syfte att skatta det monetära värdet av utestående fordringar på försäkringsbolag men kan likaväl användas för antalet händelser som vi utvecklar nedan.

Beteckningar

$D_{tx} =$ Länkkvot för tidpunkt av inträffande t och med fördröjning x.
$f_x =$ Utvecklingsfaktor för fördröjning x, $f_x > 0$ $x = 1, \ldots, \tau$.
$v_{tx} =$ Vikt av länk-kvot.

Antaganden

1. Stationäritet i utveklingsmönster för tidpunkt av inträffande t.
2. Det finns utvecklingsfaktorer $f_1, \ldots, f_\tau > 0$ s.a.

$$E\left[N_{tx} \mid N_{t0}, \ldots, N_{t(x-1)}\right] = N_{t(x-1)} f_x \quad t = 0, \ldots, \tau \quad x = 1, \ldots, \tau. \tag{25}$$

3. Variablerna N_{tx} för olika perioder t av händelsens inträffande är oberoende.

Som tidigare använder vi N_{tx}, $0 \leq t + x \leq \tau$, summan av antalet händelser för "utvecklingsperiod" x, givet tidpunkt för inträffande t och betraktar denna som en stokastisk variabel. Antagande 2 och 3 tillsammans med observerad data $\Delta = \{N_{tx} \mid 0 \leq t + x \leq \tau\}$ ger att väntevärdet för den sista möjliga fördröjningen $x = \tau$ i varje period av inträffande ges av,

$$E\left[N_{t\tau} \mid \Delta\right] = N_{t(\tau-t)} f_{\tau-t+1} \cdot \ldots \cdot f_\tau \quad t = 1, \ldots, \tau. \tag{26}$$

(Mack (1993)). Med avsikt att utvinna information om hur det sammanlagda antalet händelser utvecklas med tiden definierar vi nu länkkvoterna,

$$D_{tx} := \frac{N_{tx}}{N_{t(x-1)}} \quad t = 0, \ldots, \tau \quad x = 1, \ldots, \tau - t. \tag{27}$$

Utifrån dessa kan därefter skattningar av utvecklingsfaktorer erhållas, därmed även den icke observerade delen av kontingenstabellen via (26).

$$
\begin{array}{ccccc|c}
N_{00} & N_{01} & \dots & N_{0(\tau-1)} & N_{0\tau} & \\
N_{10} & N_{11} & \dots & N_{1(\tau-1)} & \hat{N}_{1\tau} & \\
\vdots & & \iddots & & \vdots & \\
N_{\tau 0} & & & & \hat{N}_{\tau\tau} &
\end{array}
\qquad
\begin{array}{cccc|c}
D_{01} & D_{02} & \dots & D_{0(\tau-1)} & D_{0\tau} \\
D_{11} & D_{12} & \dots & D_{1(\tau-1)} & \\
\vdots & & \iddots & & \\
D_{\tau 1} & & & &
\end{array}
$$

Om vi för varje kolumn i tabellen av länkkvoter inte kan observera några större avvikelser kan ett *genomsnitt* för given fördröjning vara en god skattning av den generella utvecklingen,

$$\hat{f_x} = \frac{1}{S_{\tau-x}} \sum_{t=0}^{\tau-x} v_{tx} D_{tx} \quad v_{tx} = 1 \quad \text{för alla} \quad t, x \tag{28}$$

där

$$S_{\tau-x} = \sum_{t=0}^{\tau-x} v_{tx} \qquad x = 1, \dots, \tau. \tag{29}$$

Eller då ett större förtroende för senare observationer föreligger ansätter vi ökande vikter v_{tx} för senare perioder av inträffande t. Och kan i likhet med Lawlessmodell göra en geometrisk modifiering, genom att ta bort en symmetrisk triangel i det övre vänstra hörnet av kontingenstabell av kumulativa antal och sedan beräkna de genomsnittliga eller viktat genomsnittliga länkkvoterna. För data bestående av monetärt värde av fordringar istället för antal, skulle vi nu även kunna tillåta inflation (Renshaw och Verall (1998)). Med generella vikter uttrycks f_x för de $m + 1$ senaste perioderna,

$$\hat{f_x} = \frac{1}{T_{\tau-x}} \sum_{t=max\{0,\tau-x-m\}}^{\tau-x} v_{tx} D_{tx} \tag{30}$$

där

$$T_{\tau-x} = \sum_{t=max\{0,\tau-x-m\}}^{\tau-x} v_{tx} \qquad x = 1, \dots, \tau. \tag{31}$$

Vidare med antagande 1, om stationäritet för utvecklingsmönstret och mer konservativ förhållning, skattas *det värsta möjliga utfallet* genom att välja det största värdet på D_{tx} i varje kolumn,

$$\hat{f_x} = max\{D_{0,x}, \dots, D_{(\tau-x)x}\} \qquad x = 1, \dots, \tau. \tag{32}$$

Om stationäritet är ett alldeles för starkt antagande, till följd av att data uppvisar en *trend* för ökande t kan Link ratio metoden även utvidgas ytterligare. Genom att skatta en regressionslinje på formen

$$f_x = \alpha + \beta t \quad t = 0, \dots, \tau - x \quad x = 1, \dots, \tau - 2 \tag{33}$$

för trenden i länkkvoterna, givet fördröjning med minsta kvadrat metoden. Eftersom det endast finns en och två observationer för $x > \tau - 2$ skattas utvecklingsfaktorerna för dessa fördröjningar lämpligen med ett genomsnitt och det enskilt observerade värdet (Fac (1997)).

Det mest populära specialfallet är dock *Chain ladder* som är Link ratio variation (28) med antalet händelser för varje inträffande period som vikt, $v_{tx} = N_{t(x-1)}$.

$$\hat{f_x} = \frac{1}{S_{\tau-x}} \sum_{t=0}^{\tau-x} N_{t(x-1)} D_{tx} = \left[\sum_{t=0}^{\tau-x} N_{t(x-1)} \right]^{-1} \sum_{t=0}^{\tau-x} N_{tx} \quad x = 1, \ldots, \tau \tag{34}$$

För Chain ladder metoden visar Mack (1993) att under antaganden 2 och 3 så är skattningarna av utvecklingsfaktorerna $\hat{f_x}$ ovan väntevärdesriktiga och även okorrelerade. Den slutliga populationsskattningar av rader erhålls därefter genom (25),

$$\hat{N}_{t\tau} = N_{t(\tau-t)} \prod_{x=\tau-t+1}^{\tau} \hat{f_x} \quad t = 0, \ldots, \tau. \tag{35}$$

2.2 Parametriska modeller

Essensen av parametriska modeller är att man minskar dimensionen av problemet, från att antalet händelser kan tillhöra en stor klass av fördelningar till att definieras av ett begränsat antal parametrar. Det kan vara informativt att alternativt använda parametriska modeller, exempelvis när det totala antalet händelser är relativt liten (Kalbfleisch et al. (1991b)), eftersom få antal observationer ger lite information om underliggande fördelning.

Generaliserade linjära modeller

GLM är en klass av modeller tillhörande den naturliga exponentialfamiljen, bestående av tre huvudsakliga komponenter. En slumpmässig- och en systematisk komponent samt en länkfunktion. Den slumpmässiga komponenten, desamma som responsvariabel Y_i av oberoende observationer antar en fördelning tillhörande den ovan nämnda och är till skillnad från den allmäna linjära modellen mer generell i bemärkelsen att den kan anta en hel klass av fördelningar, både diskreta och kontinuerliga. Generellt kan fördelningarna tillhörande exponentialfamiljen skrivas som,

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp\{y_i Q(\theta_i)\} \quad \text{eller}$$

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi)\right\}. \tag{36}$$

Den systematiska komponenten relaterar en vektor $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_N)^{\mathrm{T}}$ till en känd uppsättning av förklarande variabler x_{ij} och okända parametrar β_j genom en linjär modell vilken benämns linjär prediktor.

$$\eta_i = \sum_j \beta_j x_{ij} \qquad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{37}$$

Länkfunktionen $g(\cdot)$ kopplar den slumpmässiga komponenten via väntevärdet till den systematiska

$$E[Y_i] = \mu_i \quad g(\mu_i) = \eta_i \quad i = 1,\ldots,N \tag{38}$$

och med andra moment specificerad av

$$Var(Y_i) = a(\phi)b''(\theta). \tag{39}$$

Variansfunktionen betecknas även som $V(\mu_i)$ och kan för bl.a. Poisson, Normal och Gamma fördelningen skrivas

$$V(\mu_i) = \mu_i^{\zeta}, \quad \zeta \geq 0. \tag{40}$$

Funktionen $a(\phi)$ är oftast på formen ϕ/ω_i där ϕ (>0) är en avvikelse parameter, ω_i apriori vikter som varierar mellan observationer samt $b''(\theta)$ varians

funktionen av den kanoniska parametern θ relaterad till medelvärdet genom $\mu = b'(\theta)$ (McCullagh och Nelder (1989)). I själva verket är det inte nödvändigt att explicit referera till en specifik fördelning utan det räcker att specificera ϕ, ω_i och ζ . För exempelvis poisson erhålls samma skattningar om parametrarna specificeras av,

$$\phi = 1 \quad \zeta = 1 \quad \omega_i = 1 \quad \text{för alla} \quad i \tag{41}$$

d.v.s

$$\theta = \log(\mu) \quad b(\theta) = \exp\{\theta\} \quad V(\mu) = \mu. \tag{42}$$

Fördelarna med modellering av GLM är att teorin är enhetlig och omfattar de viktigaste fördelningarna samt restriktionen av Y_i till exponentialfamiljen då samma algoritm för parameterskattning är tillämpbar för hela familjen oberoende av länkfunktion (Agresti (2002)).

2.2.1 Tvåsidig variansanalysmodell, modelltyp I

I synvinkel av den allmäna linjära modellen är y_{ij} utfall av de okorrelerade stokastiska variablerna $Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, s \quad j = 1, \ldots, s - i + 1$. Där väntevärdena μ_{ij} kan uttrycks på linjär form som,

$$E[Y_{ij}] = \mu + \alpha_i + \beta_j \quad \text{och} \quad \sum_i \alpha_i = \sum_j \beta_j = 0. \tag{43}$$

μ är det gemensamma väntevärdet samt α_i och β_j representerar tillskott av rad respektive kolumneffekter. Eftersom vi endast har en observation per cell är det inte lönsamt att inkludera en samspelseffekt (Sundberg (1997)).

Beteckningar

$N_{tx} =$ Stokastisk variabel betecknande total mängd av anspråk, $t, x = 0, \ldots, \tau$.

$E_t :=$ $E[N_{t\tau}]$, förväntad mängd anspråk för skadeår t. Representation av rad effekt.

$S_j :=$ $E[n_{tx}]/E[N_{t\tau}]$, kvoten av förväntad ökad mängd anspråk i utvecklingsår x och förväntad slutgiltig mängd anspråk, $\sum_{x=0}^{\tau} S_x = 1$. Representation av kolumn effekt.

$R_{tx} =$ Stokastisk variabel med väntevärde $E[R_{tx}] = 1$, $R_{tx} \in \Delta = \{R_{tx} |\ t = 0, \ldots, \tau \quad x = 0, \ldots, \tau - t\}$.

Den multiplikativa representationen av modellen ges av,

$$n_{tx} = E_t S_x R_{tx} \qquad Z_{tx} = \mu + \alpha_t + \beta_x + \varepsilon_{tx} \quad \forall\ t,\ x \leq \tau - t \tag{44}$$

där

$$Z_{tx} = \ln(n_{tx}) \quad \varepsilon_{tx} = \ln(R_{tx}) \quad \text{och} \quad \sum_{t=0}^{\tau} \alpha_t = \sum_{x=0}^{\tau} \beta_x = 0$$

Data för fordringar till följd av en inträffad händelse beskrivs vid beräkning av reserver som heterogen då utbetalningar nödvändigtvis sker över tiden och tiden i sig är en orsak till heterogeniteten. För att stabilisera variansen använder man log-transformen då dess standardavvikelse är proportionell mot medelvärdet (Zehnwirth (1997)).

Antaganden

1. $n_{tx} > 0 \quad \forall\ t,\ x \leq \tau - t$
2. ε_{tx} okorrelerade med $\mathrm{E}[\varepsilon_{tx}] = 0$ och $\mathrm{Var}(\varepsilon_{tx}) < \infty$ för $t,\ x$ så att $0 \leq t + x \leq \tau$

3. $Z_{tx} = \log(n_{tx}) \sim N(m_{tx}, \sigma^2) \quad \Rightarrow \quad n_{tx} \sim \log N\left(E[n_{tx}], Var(n_{tx})\right)$

$$E[X_{tx}] = \exp\{m_{tx} + \sigma^2/2\} \quad Var(X_{tx}) = E[X_{tx}]^2(\exp\{\sigma^2\} - 1)$$

Den multiplikativa modellen given av ekvation (44), n_{tx} är fördelade enligt antagande 3 (Renshaw och Verall (1998)). Men eftersom vi har olika antal observationer för varje faktornivå, en obalanserad modell och kan således inte använda de vanliga ML skattningarna. Kremer (1982) ger tre rekursiva formler för att skatta μ, α_i och β_j utan antagande 3, vilka visas vara de bästa linjära väntevärdesriktiga skattningarna via Gauss-Markovs theorem, dock en aningen trassliga att genomföra. Men det är även möjligt att skatta dessa via EM-algoritmen beskrivet av Zehnwirth (1997) enligt följande.

Steg 0: Fyll ut kontingenstabellen med förväntade värden. Börja exempelvis med att fylla de tomma cellerna på varje rad med det senast observerade värde d.v.s. $z_{tx} = z_{t(\tau-t)}$, för Z_{Δ^c} $\Delta^c = \{t = 0, \ldots, \tau \quad x = \tau - t + 1, \ldots, \tau\}$

Steg 1: Beräkna ML skattningarna med,
$\hat{\mu} = \bar{z}_{..} \quad \hat{\alpha}_t = \bar{z}_{i.} - \bar{z}_{..} \quad \hat{\beta}_x = \bar{z}_{.j} - \bar{z}_{..}$

Steg 2: Använd ML skattningarna ovan för att beräkna nya förväntade värden för de tomma cellerna Z_{Δ^c}. Återgå därefter till steg 1 tills dess att skattningarna förändras mindre än till föreskriven toleransnivå.

Antalet skattade parametrar i modellen är $2(n-1)+1$ och antal observationer $n(n+1)/2$, alltså antalet frihetsgrader för residualkvadratsumman $(n(n-3)/2)+1$ och skattningen av variansen σ^2,

$$\hat{\sigma}^2 = \frac{1}{\left(\frac{n(n-3)}{2}+1\right)} \sum_{t=0}^{\tau} \sum_{x=1}^{\tau-t} (z_{tx} - \bar{z}_{t.} - \bar{z}_{.x} + \bar{z}_{..})^2 \tag{45}$$

De tomma cellerna X_{Δ^c} skattas nu med väntevärdet för X_{tx},

$$\hat{X}_{tx} = \exp\left\{\hat{\mu} + \hat{\alpha}_t + \hat{\beta}_x - \frac{\sigma^2}{2}\right\} \quad \forall\, t, x \in \Delta^c \tag{46}$$

Poisson

Poisson är en gränsfördelning till binomial fördelningen då sannolikheten för att en händelse inträffar i n st försök är liten, beroendet mellan dem är svagt och antalet försök n är stort. Fördelningen har endast en parameter, medelvärdet μ vilket måste vara positivt och bestämmer fördelningen fullständigt. Fördelningen beskriver händelser som inträffar slumpmässigt och oberoende över tiden, sådana händelser genereras av en räkneprocess kallad poissonprocessen. En räkneprocess $\{N(t) \geq 0,\ t \geq 0\}$ definierar Ross (2007) som en poissonprocess med intensitet $\lambda > 0$ om den uppfyller

(i) $N(0) = 0$

(ii) Stationära inkrement, antalet händelser i ett intervall är endast beroende av intervallets längd.

(iii) Oberoende inkrement, antalet händelser i icke överlappande intervall är oberoende.

(vi) $P\left(N(h) = 1\right) = \lambda h + o(h)$

(v) $P\left(N(h) \geq 2\right) = o(h)$.

Baserat på kriterierna ovan kan det visas att antalet händelser i ett intervall av längd t är Poisson fördelat med intensitet λt.

Om $X_1, \ldots, X_k$ är oberoende poissonfördelade stokastiska variabler med parametrar $\lambda_1, \ldots, \lambda_k$ så är vektorn $\mathbf{X} = (X_1, \ldots, X_k)$ givet det totala antalet observationer $\sum_{i=1}^{k} X_i = n$ multinomialfördelad (Bishop et al. (1975)).

$$n \sim Po(\textstyle\sum_{j=1}^{k} \lambda_i) \quad \text{och} \quad \mathbf{X}|\textstyle\sum_{j=1}^{k} X_i = n \sim Multi(n, \boldsymbol{\pi})$$

$$X_i|\textstyle\sum_{j=1}^{k} X_i = n \sim Bin(n, \pi_i) \quad \text{där} \quad \pi_i = \dfrac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}$$

Den simultana fördelningen för $\mathbf{X}$ kan delas upp som produkten $P(\mathbf{X}) = P(\mathbf{X}|n)P(n)$ samt att likelihood funktionen kan faktoriseras som en produkt av två oberoende funktioner $L_M(\boldsymbol{\pi})L_{Po}(\boldsymbol{\lambda})$ då $\boldsymbol{\pi}$ inte bär någon information om n och den omvända relationen gäller.

2.2.2 Poisson log-linjär modell

Eftersom medelvärdet för poissonfördelade stokastiska variabler är strängt positivt är en additiv modell av medelvärdet inte helt tillfredsställande. Länkfunktionen $\eta = \sum_i \beta_i x_i$ kan anta negativa värden för vissa parameter och kovariat kombinationer givet identitetslänken $\mu = \eta$ följaktligen även medelvärdet μ. Dock kan vi försäkra oss att μ förblir positiv, ansättning av modellen med multiplikativa effekter för $\mu = \exp\{\eta\}$ innebärande en log-länkfunktion $\eta = \log(\mu)$ och att den linjära prediktorn följer den additiva linjära modellen snarare än medelvärdet (McCullagh och Nelder (1989)).

Beteckningar

$\lambda_t =$ Medelvärde för antal inträffade fall vid tidpunkt $t = 0, \ldots, \tau$
$p_x =$ Sannolikhet för rapportering med fördröjning $x = 0, \ldots, \tau$

Antagande

1. Antal fall i varje cell är oberoende poissonfördelade med medelvärde μ_{ij}

$n_{tx} \sim Po(\mu_{tx})$	$\boldsymbol{\eta} = g(\boldsymbol{\mu})$	$\boldsymbol{\eta} = \sum_1^p \mathbf{x}_j \beta_j$
Observationerna är poissonfördelade	log-länkfunktion $g(\cdot) = \log(\cdot)$	linjär prediktor av kovariater $\mathbf{x}_1, \ldots, \mathbf{x}_p$

Genom att ansätta den multiplikativa modellen för medelvärdet korrigeras antalet inträffade händelser vid tidpunkterna $t = 0, \ldots, \tau$ med sannolikheten för varje fördröjning

$$\mu_{tx} = \lambda_t p_x \quad t,\ x = 0, \ldots, \tau. \tag{47}$$

Som tidigare noterat är andramoment och variansfunktion lika då $a(\phi) = 1$,

$$V(\mu_{tx}) = \mu_{tx}. \tag{48}$$

Fördröjningensfördelningen är här av huvudsakligt intresse, antas implicit vara stationär för $x \in [0, \tau]$ samt att alla händelser är inträffade vid maximal fördröjning $x = \tau$. Med en separat parameter för varje fördröjning, inget antagande om dess fördelningen och då intervallens längd i både t och x led går mot noll, kan skattningarna av p_x anses som analoga till de erhållna av ”product limit estimate” under trunkering (20) (Sellero et al. (1996)). Fortsatt är $N_{t(\tau-t)}$ oberoende poissonfördelade och i likhet med Lawlessmodell sätter vi medelvärdet till θ_t,

$$\theta_t = \sum_{x=0}^{\tau-t} \mu_{tx} = \lambda_t F_t(\tau - t) \quad t = 0, \ldots, \tau \tag{49}$$

där

$$\mu_{tx} = \lambda_t f_t(x). \tag{50}$$

Till följd av att likelihoodfunktionen för den icke betingade fördelningen av $\mathbf{n}_t = (n_{t0}, \ldots, n_{\tau-t})$ kan faktoriseras i $L_M(\boldsymbol{\pi})$ (5) och $L_{Po}(\theta_t)$ av totalen $N_{t(\tau-t)}$ ges de resulterande ML skattningarna under stationäritets antagande för de senaste $m+1$ perioderna och $t = 0, \ldots, \tau$ av $\hat{\theta}_t = N_{t(\tau-t)}$ samt $\hat{g}(x)$ lika med (10) (Lawless (1994)).

För en poissonfördelad responsvariabel är den kanoniska länkfunktionen $\theta(\mu_{tx})$ log-länken (42), om denna ansätts ges de multiplikativa effekterna på medelvärdet av additiva i den linjära prediktorn och $\theta(\mu_{tx}) \equiv g(\mu_{tx})$,

$$\eta_{tx} = \alpha'_t + \beta'_x. \tag{51}$$

Med $\alpha'_t = \log(\lambda_t)$ $\beta'_x = \log(p_x)$ eller i likhet med variansanalysmodellen,

$$\eta_{tx} = \mu + \alpha_t + \beta_x \quad \alpha_1 = \beta_1 = 0 \quad t, x = 0, \ldots, \tau. \tag{52}$$

Utöver inträdes- och fördröjnings effekter kan även diagonaleffekt motsvarande kalendertidpunkt för observation γ_k, $k \equiv t + x$ (modulo 12) samt säsong för inträffade införas. När sedan ML skattningarna för parametrarna i den linjära prediktorn beräknats, uttrycks det totala antalet händelser av,

$$\hat{N}_{t\tau} = N_{t(\tau-t)} + \sum_{x=\tau-t+1}^{\tau} \hat{\mu}_{tx} \quad t = 0, \ldots, \tau. \tag{53}$$

Dessa skattningar av $\hat{N}_{t\tau}$ visar Renshaw och Verall (1998) är ekvivalenta med skattningar erhållna från Chain-ladder (34), (35) och Lawlessmodell (12) med antagande om stationäritet för hela den observerade perioden. Ett stickprov från poisson, multinomial eller då de utgör en produkt av multinomialfördelade stokastiska variabler resulterar samma ML skattningar (Bishop et al. (1975)). Detta är en viktig pusselbit som knyter ihop båda modellsektionerna. Chain-ladder är endast ett enkelt sätt att hitta ML skattningarna för den betingade likelihooden L_M och följaktligen även Poisson modellen, därav är Chain-ladder metoden lämplig att appliceras på antal IBNR händelser men dock inte för dess monetära värde (Renshaw och Verall (1998)).

I fallet det underliggande antagandet om poissonfördelning är ogiltigt, exempelvis då variationen i data överstiger den given av medelvärdet kan fortfarande fördelningen (36) användas. Genom att utöver parametrarna i den linjära prediktorn även skatta $a(\phi) = \phi$ vilken korrigerar för extra variation, $V(n_{tx}) = V(\mu_{tx})\phi$ (Agresti (2002)).

Garantier

Den log-linjära modellen används även med fördel då problemet utvidgas ytterligare en dimension för att anpassas till antalet garantianspråk som rapporterats till producent. Varje anspråk besitter information avseende tidpunkt för försäljning s, ålder t och rapportfördröjning x. Antalet n_{stx} antas fortfarande var poissonfördelade med multiplikativa effekter $\mu_{stx} = N_s \lambda_t p_x$.

Beteckningar

$N_s =$ Antal produkter som säljs dag $s: \ 0 \leq s \leq v$
$\lambda_t =$ Medelvärde för antalet anspråk t dagar efter försäljning, $0 \leq t \leq \tau$.
$p_x =$ Sannolikhet att ett anspråk rapporteras i databas x dagar efter att det har inträffat.

Antaganden

1. Antalet anspråk t dagar efter försäljning $\sim Po(\lambda_t)$, där λ_t är oberoende faktorer avseende när produkten tillverkades och såldes.
2. $n_{stx} \sim Po(\mu_{stx})$ där $\mu_{stx} = N_s \lambda_t p_x \quad 0 \leq s + t + x \leq \tau$.
3. p_x är oberoende av när anspråket inträffade.

$$\Lambda_t = \sum_{u=0}^{t} \lambda_u \quad \text{och} \quad P_x = \sum_{u=0}^{x} p_u \tag{54}$$

Fokus skiftar till att prediktera det genomsnittliga antalet anspråk $m(t)$ för produkter av ålder t,

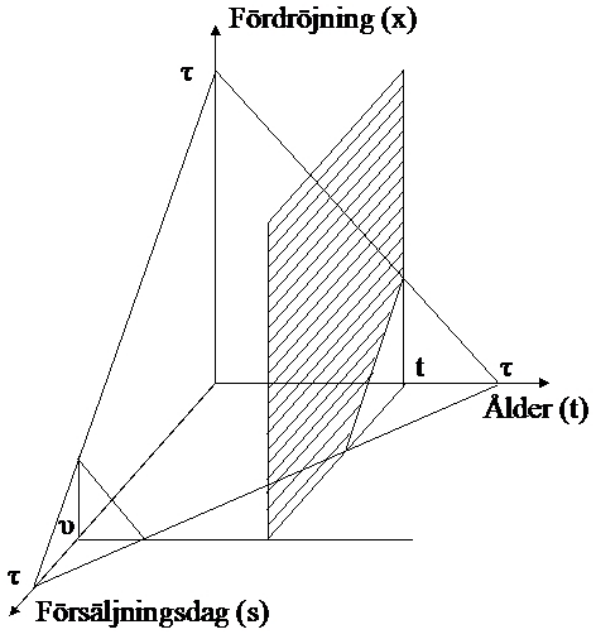
$$m(t) = \frac{\sum_{s=0}^{\tau} \sum_{x=0}^{\infty} n_{stx}}{\sum_{s=0}^{\tau} N_s} \quad t = 0, 1, \ldots, \tau \tag{55}$$

där N_s är antalet sålda enheter tidpunkt s. Och det genomsnittliga antalet anspråk på produkter fram till och med ålder t,

$$M(t) = \sum_{u=0}^{t} m(u). \tag{56}$$

Som tidigare konstrueras och maximeras likelihoodfunktionen, med de resulterande ML skattning av antalet anspråk t dagar efter försäljning,

$$\hat{\lambda}_t = \frac{n_{\cdot t \cdot}}{R_{\tau - t}} \quad t = 0, \ldots, \tau. \tag{57}$$



Figur 4: Anspråk med fördröjning v, där det streckade området representerar icke observerbara händelser $\{s, t, x : \ 0 \leq s \leq \tau, \ \tau - v < t + x \leq \tau\}$.

.

Där $n_{\cdot t \cdot}$ är antalet anspråk på produkter som är t år gamla och $R_{\tau-t}$ är det reglerade antalet produkter av ålder t som det finns risk att ett anspråk tillkommer innan kalender tidpunkt τ.

$$n_{\cdot t \cdot} = \sum_{s+l \leq \tau - t} \sum n_{stx} \quad \text{och} \quad R_{\tau-t} = \sum_{s=0}^{\tau-t} N_s P_{\tau-t-s} \tag{58}$$

Eftersom $n_{\cdot t \cdot}$ är antalet som rapporterats innan kalender tidpunkt τ och $R_{\tau-t}$ reglerar för antalet som riskerar anspråk, skattar kvoten (57) $m(t)$,

$$\hat{m}(t) = \hat{\lambda}_t \quad \text{och} \quad \hat{M}(t) = \sum_{u=0}^{t} \hat{\lambda}_t. \tag{59}$$

Observera att om $P_{\tau-t-s} = 1$ för alla $s \in [0, \ \tau - t]$ d.v.s. alla anspråk rapporterats för produkter av ålder t innan kalender tidpunkt τ blir $\hat{\lambda}$ observerat antal anspråk genom totalt antal sålda produkter fram till kalendertidpunkt $\tau - t$ För att skatta antalet anspråk som har rapporterats och ej $\hat{n}_{\cdot t \cdot}$ kan vi använda det ”nominella antalet” som riskerar garanti anspråk, $R^*_{\tau-t} = \sum_{s=0}^{\tau-t} N_s$ då,

$$\hat{\lambda}_t = \frac{\hat{n}_{\cdot t \cdot}}{R^*_{\tau-t}} = \frac{n_{\cdot t \cdot}}{R_{\tau-t}} \quad \Rightarrow \quad \hat{n}_{\cdot t \cdot} = \frac{R^*_{\tau-t}}{R_{\tau-t}} n_{\cdot t \cdot} \tag{60}$$

För att sedan skatta p_x föreslår Kalbfleisch et al. (1991b) bland annat maximumlikelihood av den trunkerade fördelningen av fördröjningar, på samma sätt som i Lawlessmodell.

3 Tillämpning

3.1 Fastighetsprisstatistik

Prisstatistik används som underlag för att bl.a. utvärdera prisutveckling, vid bestämning av taxeringsvärdenivå och ge information om omsättning av fastigheter. Statistiken tas fram av SCB i två serier, pris- och lagfartsstatistik. Den förstnämnda bestående av marknadsmässiga köp innefattande fastighetsprisindex (FASTPI) – uppdelad för egnahem och fritidshus samt efter region, köpeskillingskoefficient och kvadratmeterpris. För att bedöma värdeförändringen av en fastighet över tiden är det lämpligt att använda FASTPI, dock för den aktuella värdenivån är köpeskillingskoefficienten en mer passande statistika. Den andra är av antalet beviljade lagfarter indelade efter fastighetstyp samt dess underliggande klassificering.

Vid köp av fastighet skrivs ett köpekontrakt mellan köpare och säljare innehållande uppgifter om fastighet, köpeskilling, tillträde etc. Endast överrenskommelse angiven av det skriftliga köpekontraktet är giltig dock kan detta brytas med vissa förbehåll. Förvärv av fastigheten är slutligt genomförd vid tidpunkt T_i då köpebrevet undertecknats av båda parter, vilket fungerar som kvitto för betalning av fastighet. Därefter måste köparen själv via mäklare eller banken göra en ansökan om lagfart d.v.s. officiell registrering av förvärv av fast egendom samt inteckning av fastighet som tryggar bankens panträtt. Lagfart ska enligt lag sökas senast tre månader efter att officiell handling avseende köp upprättades (köpebrev). Ansökan skickas till ett av inskrivningsmyndighetens sju kontor som därefter handlägger ärendet. Beredning av ansökan på respektive inskrivningsenhet tar olika lång tid innan beslut, beroende på arbetsbelastning och ärendets svårighetsgrad.

Figur 5: Händelseförlopp från förvärv till rapportering till SCB.

Uppgifter om lagfart förmedlar inskrivningsmyndigheten (del av lantmäteriet) till skattemyndigheten som i sin tur lämnar uppgifter avseende taxering (årligen). Uppgifterna om förvärv och taxering kombineras i ett register i Fastighetsdatasystemet hos Lantmäteriet som löpande rapporteras till SCB

och bildar lagfarts- och fastighetsprisregistren.

Statistiken som sammanställs av SCB publiceras för varje kvartal och med två månaders eftersläpning för att så stor del som möjligt av lagfarna köp ska hinna registreras. De publicerade siffrorna för varje kvartal revideras därefter kvartalsvis för att även innefatta sent registrerade förvärv, detta fortgår fram till och med maj påföljande år i samband med publicering av årsstatistiken vilket räknas som fastställda siffror. Detta innebär en strukturell snedvridning och underskattning av de senare perioderna då siffror, kvartal ett har ungefär nio månader längre tid på sig att rapporteras i jämförelse med årets sista kvartal.

För kvartal- och årsstatistiken kan det vara av intresse att skatta omsättningen på fastigheter i beståndet till följd av den underrepresentation av den verkliga, beskriven ovan och kan användas för tidiga konjunkturindikationer.

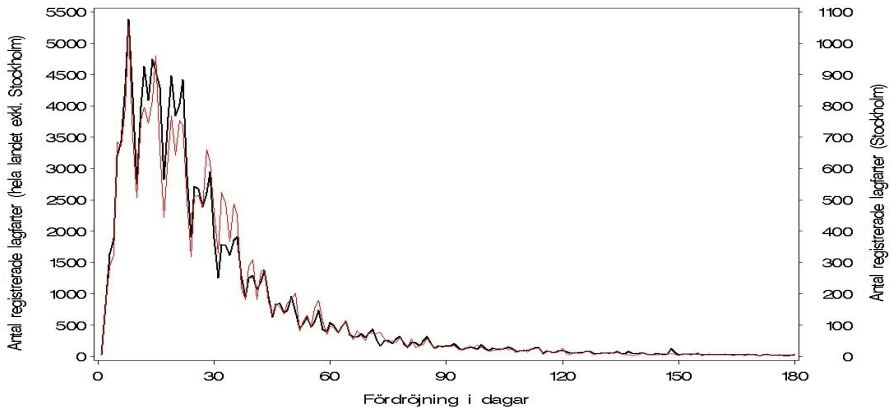
Denna del avser att illustrera hur vi med den log-linjära modellen kan göra en tidig skattning av årsstatistikens slutgiltiga siffror.

3.1.1 Beskrivning av data

Data som har använts för analys av lagfartsstatistik har erhållits från Statistiska Centralbyråns fastighetsprisregister. Databasen består av en huvudtabell med information om fastighetsförvärv registrerade vid Lantmäteriverket (LMV), innefattande 46 variabler bland andra fångeskod (se Appendix A.3 Ordlista), förvärvsdatum, registreringsdatum, länskod och typkod vid fastighetstaxering. Huvudtabellen är i sin tur kopplad via ett unikt överlåtelsenummer till sex mindre tabeller med uppgifter om köpare och säljare samt specifika variabler efter typkodsklassificering.

Avgränsning

Den fullständiga tabellen innehåller 703 164 observationer av förvärv mellan 1996-01-02 och 2009-08-07 varav den längsta fördröjningen mellan förvärv och registrering uppgår till 3 578 dagar. För att mängden data ska bli mer hanterbar avgränsas analysen till att endast omfatta vanliga köp (normal- och specialfall) av småhusenheter avsedda som helårsbostad, gjorda tidigast den 1 januari 2007 och belägna i Stockholms län. Antalet observationer reduceras till 27 302 och längsta registrerade fördröjning 747 dagar. Denna avgränsning är även motiverad då fördelningen av beviljade lagfarter i hela landet i jämförelse med Stockholms län är näst intill identisk med skillnad i skalfaktor (figur 6).



Figur 6: Fördelning av registrerade lagfarter efter rapportfördröjning i dagar, hela landet exklusive Stockholm län (tjock linje) och för Stockholms län (tunn linje), mellan 20070101 och 20090812.

3.1.2 Modellering

Den mest naturliga början är att undersöka våra möjligheter att prediktera antal fastighetsförvärv under ett år efter passerat årsskifte, med tillgänglig data. Genom att dela in antalet förvärv efter månad för inträffande och med en fördröjning grupperad i månader á 30, 4 dagar erhålls en tabell med 12 månader för inträffande och fördröjning. Eftersom januari är den första observationsmånaden motsvarar den $t = 0$ och den sista december $t = 11$, enligt tidigare notation. Vidare bör det noteras att SCB endast får leverans av data en gång i månaden infallande den 12e, vilket innebär att antalet inträffande förvärv i december och som har ett registreringsdatum efter den 12e först kommer att vara känd den 12e januari eller februari påföljande år. För att få ekvidistanta intervall används endast data för upp till och med beviljande per den 31 december d.v.s. en trunkeringstid motsvarande 365 dagar. År 2008 väljs att undersökas då publicering av årsstatistik redan är passerad samt de sanna värdena anses kända.

	0	1	...	8	9	10	11
jan	647	84	...	1	2	0	0
feb	783	133		3	0	0	-
⋮	⋮					⋰	⋮
nov	484	65	-				-
dec	191	-	-			...	-

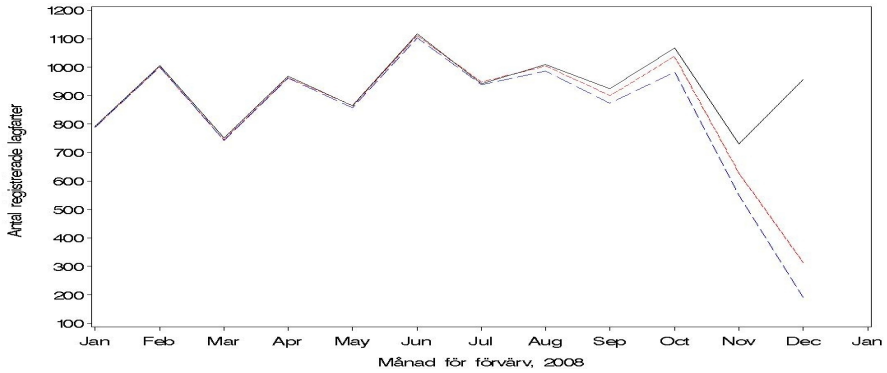
Tabell 2: Trunkerad kontigenstabell med förvärv under 2008 och trunkering med avseende på 365 dagar.

Tabell 2 visar antal registrerade fastighetsförvärv 2008, trunkeringen är med avseende på observation fram till och med den 31 december, $\tau = 365$.

Den multiplikativa modellen för medelvärdet och log-länkfunktionen ger den linjära prediktorn,

$$\eta_{tx} = \alpha_t + \beta_x \quad t = 0, \ldots, 11 \quad x = 0, \ldots, 11 - t. \tag{61}$$

Parametrarna för förvärvsmånad α_t respektive fördröjning β_x skattas genom maximumlikelihood för Poisson fördelad respons $\phi = 1$, med proceduren GENMOD i SAS® 9.1. Till följd av att vi endast kan observera inträffade händelser upp till och med 9 månaders fördröjning skattas endast parametrar för dessa 10 intervall av fördröjning.



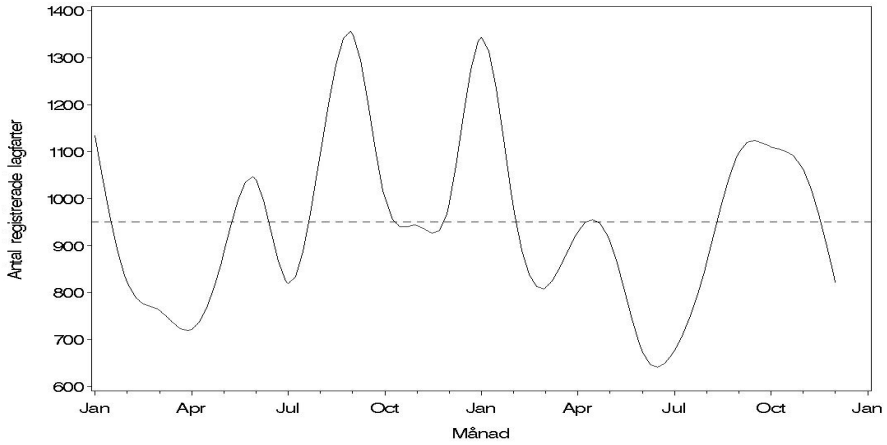
Figur 7: Antal beviljade lagfarter i Stockholms län per månad för registrering 2008. Faktiskt utfall (heldragen linje), skattade värden (61) (kortstreckad linje) och observerade värden (långstreckad linje), trunkering med avseende på 365 dagar.

Genom att undersöka antalet beviljade lagfarter efter månad för registrering (Figur 8, 12) kan vi observera att januari och september har ett högt antal medan juli tenderar att ha ett lågt antal beviljade lagfarter. Samtidigt tenderar speciellt antalet förvärv i juni samt november/december ha långa fördröjningar (Figur 11). Rimligtvis är detta effekter av semester perioder som infaller under jul och sommar, där den senare är känd sedan tidigare. För att anpassa modellen till dessa effekter införs parametrar och indikator variabler $I_{(\cdot)}$ för den första ($0 \equiv t + x$ modulo 12), sjunde ($6 \equiv t + x$, modulo 12) och nionde ($8 \equiv t + x$ modulo 12) diagonalen motsvarande förvärv i januari, juli och september. Den linjära prediktorn ges nu av,

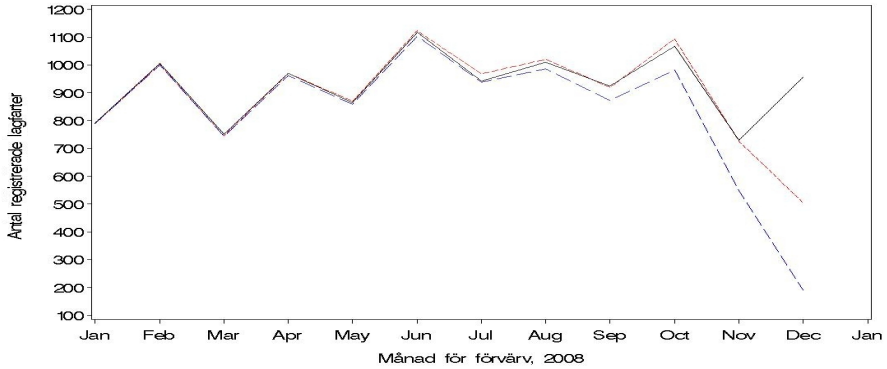
$$\eta_{tx} = \alpha_t + \beta_x + \gamma I_{jan} + \delta I_{jul} + \epsilon I_{sept} \quad t = 0, \ldots, 11 \quad x = 0, \ldots, 11 - t. \tag{62}$$

Modellen presterar bättre i flertalet intervall (figur 9), speciellt avseende november vilken uppvärderas med 98 registreringar, dock förblir december

kraftigt undervärderad. Detta är en effekt av att endast halva cellen för inträde i december är känd, vilket även gäller för övriga utefter samma diagonal. Om hänsyn tas till denna effekt och tabellen istället trunkeras med avseende på hela celler, observeras celler med samtliga fördröjningar och resulterar en bättre anpassning till den sista månaden däremot med en något större diskrepans för övriga månader (se Appendix A.2 Figur 13).



Figur 8: Antal registrerade lagfarter i Stockholms län per månad för registrering, mellan jan 2007 och dec 2008.



Figur 9: Antal registrerade lagfarter i Stockholms län per månad för registrering, 2008. Faktiskt utfall (heldragen linje), skattade värden (62) (kortstreckad linje) och observerade värden (långstreckad linje), trunkering med avseende på 365 dagar.

Månad	Utfall[1]	Obs.[2]	Obs.[3]	(61)[2]	(62)[2]	(62)[3]
jan	791	788	789	788	788	789
feb	1005	1000	1001	1002	1003	1014
mar	752	743	744	744	745	751
apr	969	962	963	964	970	978
maj	864	858	859	864	869	873
jun	1117	1103	1106	1111	1123	1132
jul	942	938	938	948	968	978
aug	1009	986	987	1004	1020	1026
sep	924	873	895	900	919	949
okt	1067	982	1017	1038	1094	1153
nov	730	549	624	628	726	849
dec	956	191	376	313	504	1044
$\sum$	11126	9973	10299	10304	10729	11536
Δ		-1153	-827	-822	-397	410

Tabell 3: Antal registrerade lagfarter per månad 2008.

Först bör vi notera att parameter skattningarna i den linjära prediktorn, erhållna från proceduren GENMOD, inte är definierade för den icke observerbara delen av kontingenstabell 2, utan endast för värden till vänster om tidpunkt för trunkering. Om vi dessutom låter avvikelseparametern variera fritt får vi en kraftig överspridning i samtliga modeller. Parametern ϕ skattad med kvadratroten ur deviance dividerat med antal frihetsgrader resulterar i $\hat{\phi} = 5,66$ för modell (62) under cell trunkering och i samma storleksordning för övriga modeller samt trunkering med $\tau = 365$ dagar. Resultatet är följaktligen inte i linje med det underliggande antagandet om poissonfördelad respons. Parameterskattningarna påverkas dock inte av poissonantagandet och överspridningen då endast länk- och variansfunktion används för att erhålla ML skattningarna.

[1] Utfall per den 12 augusti 2009.
[2] Trunkering med avseende på 365 dagar.
[3] Trunkering med avseende på hela celler.

4 Sammanfattning

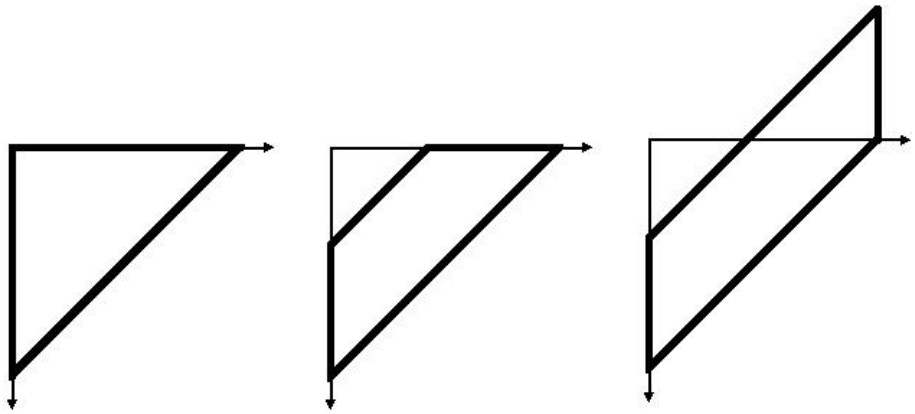
Modeller		
(1)	(2)	Jämförelse
Lawlessmodell	Trunkeringsmodellen	(1) diskretiserar data och antar endast stationäritet i fördröjningsfördelningen för ett begränsat intervall av observerad period, annars resulterar (1) och (2) i lika skattningar.
Lawlessmodell	Poisson log-linjär modell	Den icke betingade fördelningen av $\mathbf{n}_t$ kan faktoriseras i $L_M(\boldsymbol{\pi})$ och $L_{Po}(\theta_t)$. Därför resulterar både (1) och (2) i samma ML skattningar av fördröjningsfördelningen $\hat{g}(x)$.
Chain ladder	Tvåsidig variansanalysmodell	Enligt Kremer (1982) resulterar (1) och (2) i identiska skattningar dock är de inte exakt ekvivalenta (Renshaw och Verall (1998)).
Chain ladder	Poisson log-linjär modell	(2) med den linjära prediktorn given av ekvation (52) resulterar i ekvivalenta skattningar av $\hat{N}_{t\tau}$ med (1). Modell (1) är egentligen endast ett enkelt sätt att hitta ML skattningarna av den betingade likelihooden L_M.
Trunkeringsmodell	Poisson log-linjär modell	Med en separat parameter för varje fördröjning, inget antagande om dess fördelning och då intervallens längd går mot noll i både t och x led kan skattningarna av p_x i (2) anses analoga till $\hat{G}(x)$ (ekvation (20)) erhållna av (1).

5 Diskussion

I samtliga modeller antas implicit att man kan extrapolera skattningar av fördröjningsfördelning, utvecklingsfaktorer eller parametrar på den icke observerbara delen av kontingenstabellen Δ^c.

Modellerna som har avhandlats resulterar i ekvivalenta eller likartade skattningar med skillnad i vilken information som används för att dessa ska uppnås. De observerade händelserna grupperas antingen i celler, representerade av intervall för inträde och rapportering eller så tas hänsyn till hela datamängden i form av exakta datum. Det förstnämnda förfarandet underlåter beaktande av delvis observerade celler eftersom de klassificeras som strukturella nollor. Följaktligen går möjligtvis viktig information förlorad, detta exemplifieras tydligt i tillämpningen på lagfartsstatistik (Figur 9, 13).

Därefter införs antagande om stationäritet av fördröjningarnas fördelning genom att välja en delmängd av observerad triangel som modellen ska anpassas till, antingen efter samtliga observationer/celler alternativt endast för ett viss intervall av $m + 1$ perioder, lämpligen de senast observerade. Eftersom fördröjningarnas fördelning spelar en central roll är det värt att ifrågasätta om dessa är den mest lämpliga formerna att modellera fördröjningen på. Om data existerar för inträffande i tidigare perioder $T < 0$, kan ett parallellogram vara mer passande då varje parameter för samtliga fördröjningar skattas med lika många observationer, fortfarande endast erhållna från de senaste $m + 1$ observerade perioderna (Fac (1997)). Många tillämpningsområden visar rimligtvis på icke stationäritet, händelser påverkas av yttre omständigheter som inflation, ekonomisk aktivitet, trender eller liknade. Därför måste en avvägning göras mellan längden på stationäritets antagandet och förlusten av information.



Figur 10: Möjliga former.

Vidare finns det flera möjliga modeller för skattning av OBNR händelser som inte har inkluderats i denna rapport. Bland annat kan regressions-

modeller för riskfunktionen i omvänd tid ställas upp, där $g(x)$ modelleras med logit eller komplementerande log-log transformation (Kalbfleisch och Lawless (1991a)). Detta kan vara lämpligt om stationäritetsantagandet inte är uppfyllt, och leder oss till att beakta en utvidgning av (50) där fördröjningsfördelningen får fluktuera slumpmässigt genom att anta en sannolikhetsfördelning för $f_t(x)$. Mer specifikt antar Lawless (1994) att vektorn $\mathbf{f}_t = (f_t(0), \ldots, f_t(\tau))$ följer dirichletfördelningen.

Huvudtemat i den andra delen, parametriska modeller bygger på teori om generaliserade linjära modeller. Dessa karaktäriseras av antagandet att de underliggande observationerna är oberoende, detta innebär att tidsserie modeller som bygger på autokorrelation exkluderas (McCullagh och Nelder (1989)). Ett alternativ till ovanstående kan därav tillhöra denna klass av modeller.

Eftersom vårt syftet specifikt varit att erhålla punktskattningar har andra viktiga hänsynstagande åsidosatts, så som undersökning och jämförelse av skattningarnas variabilitet, antaganden om responsvariabelns fördelning och hypotestest för fördröjningsfördelningens stationäritet. Detta återstår för senare studier likaså utvidgning och speciell modellanpassning för tillämpning på lagfartsstatistiken.

Detta examsensarbete likaså fortsatta studier kommer att vara användbara för SCB eftersom snabbare rapportering är högt värderad och modellerna kan användas på ett flertal områden.

A Appendix

A.1 Härledning

Ekvationer (3) och (4)

$$1 - g_t(x) = 1 - \frac{f_t(x)}{F_t(x)} = \frac{F_t(x) - f_t(x)}{F_t(x)} = \frac{F_t(x-1)}{F_t(x)}$$

$$\frac{F_t(x)}{F_t(\tau)} = \frac{F_t(x)}{F_t(x+1)} \frac{F_t(x+1)}{F_t(x+2)} \cdots \frac{F_t(\tau-1)}{F_t(\tau)} = \prod_{r=x+1}^{\tau} (1 - g_t(r))$$

Likelihood (5)

Likelihood funktionen för vektorn $\mathbf{X} = (X_1, \ldots, X_T)$ av kategorier givet totala antalet observationer $\sum_{i=1}^{T} X_i = N$ och vektorn $\mathbf{p} = (p_1, \ldots, p_T)$ av sannolikheter för att en händelse inträffar i respektive kategori.

$$\mathbf{X} \mid N \sim \text{Multi}\Big(N, \mathbf{p}\Big)$$

$$L = \binom{N}{x_1, \ldots, x_T} \prod_{i=1}^{T} p_i^{x_i} = \frac{N!}{x_T! \prod_{i=1}^{T-1} x_i!} \, p_T^{x_T} \prod_{i=1}^{T-1} p_i^{x_i} = \left\{ n = \sum_{i=1}^{T-1} x_i = N - x_T \right\} =$$

$$= \frac{N!}{x_T!(N - x_T)!} \, p_T^{x_T} \, (1 - p_T)^{N - x_T} \left(\left(\frac{1}{1 - p_T} \right)^{\sum_{i=1}^{T-1} x_i} \frac{n!}{\prod_{i=1}^{T-1} x_i!} \prod_{i=1}^{T-1} p_i^{x_i} \right) =$$

$$= \left\{ Z_T \sim \text{Bin}\Big(N, p_T\Big) \right\} = f(Z_T = x_T) \, \frac{n!}{\prod_{i=1}^{T-1} x_i!} \prod_{i=1}^{T-1} \left(\frac{p_i}{\sum_{j=1}^{T-1} p_j} \right)^{x_i} = \ldots =$$

$$= \prod_{i=1}^{T} f(Z_i = x_i) \quad \text{där} \quad Z_i \sim \text{Bin}\Big(N - \textstyle\sum_{j=1}^{i} x_j, \; p_i / \sum_{j=1}^{i} p_j\Big)$$

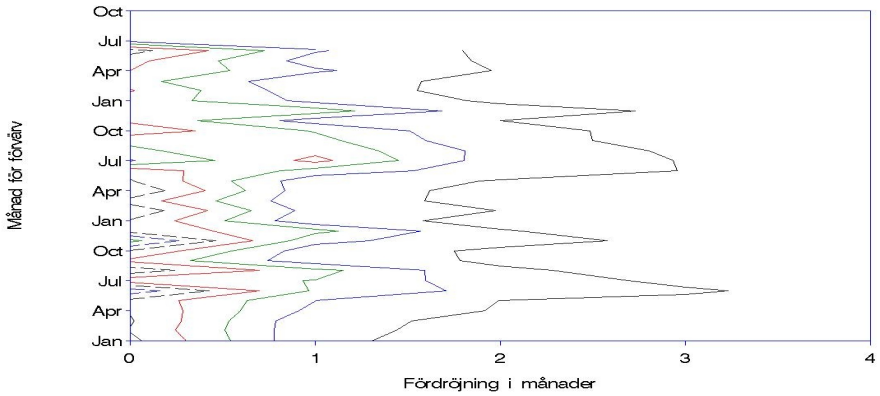
Detta är samma likelihood som presenteras i Lawless (1994) men dock med de normerade sannolikheterna $g_t(x)$ istället för $p_i / \sum_{j=1}^{i} p_j$.

Ekvation (34)

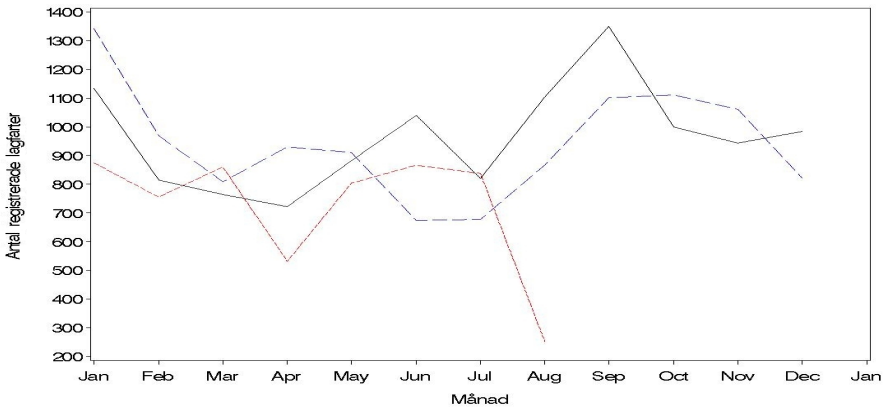
Sista steget i ekvation (34) får vi genom att inse,

$$\sum_{t=0}^{\tau-x} N_{tx} = \sum_{t=0}^{\tau-t} \left\{ N_{tx} \frac{N_{tx}}{N_{t(x-1)}} \frac{N_{t(x-1)}}{N_{tx}} \right\} = \sum_{t=0}^{\tau-x} N_{t(x-1)} D_{tx}.$$

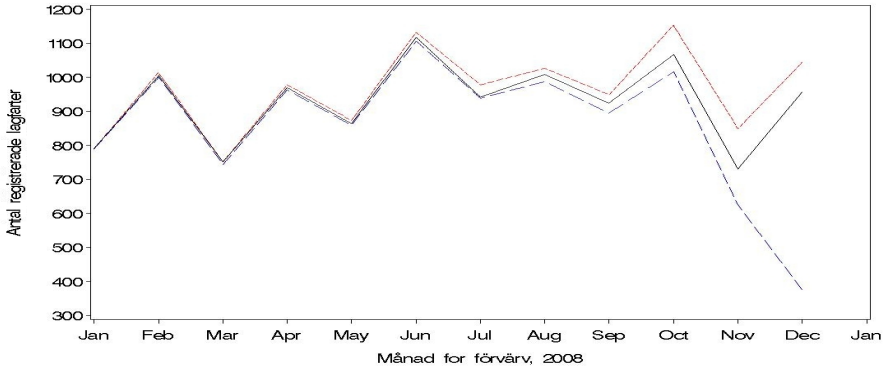
A.2 Figurer



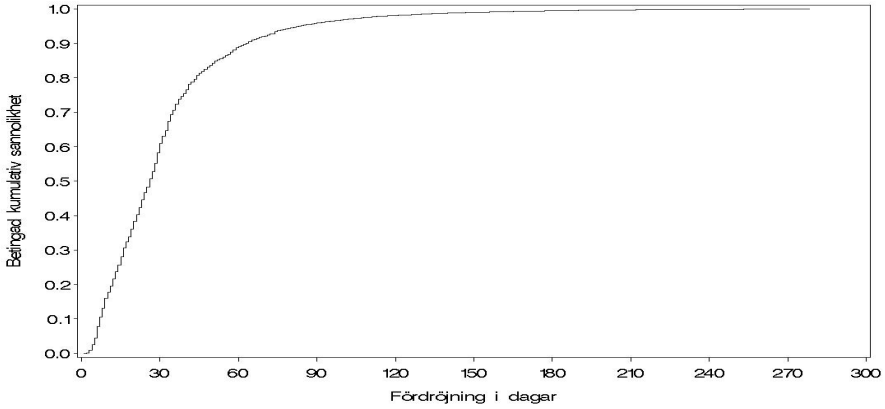
Figur 11: Konturplot av månad för förvärv mot fördröjning, mellan 20070101 och 20090812. De yttre ”ringarna” markerar lägre frekvenser av registreringar och de inre högre frekvenser.



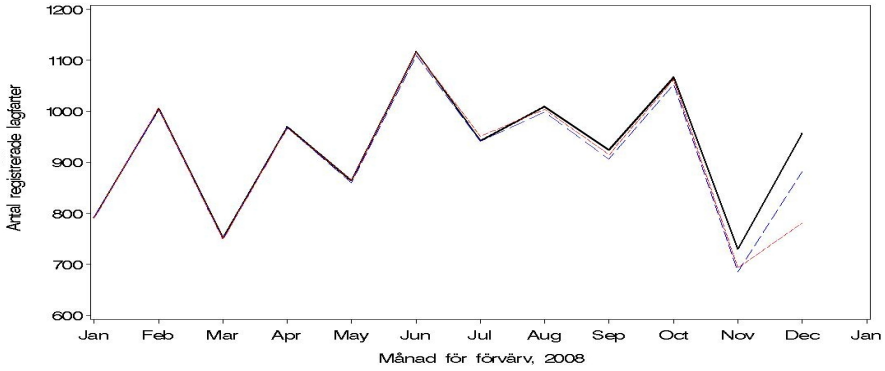
Figur 12: Jämförelse mellan antal registrerade lagfarter per månad 2007 (heldragen linje), 2008 (långstreckad linje) och 2009 (kortstreckad linje).



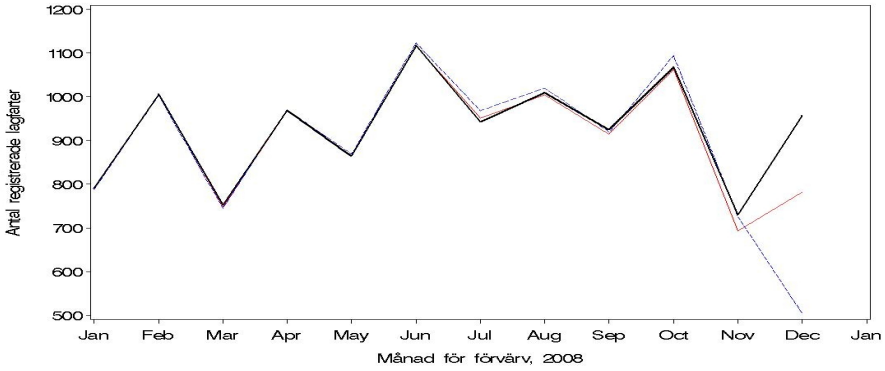
Figur 13: Antal registrerade lagfarter 2008, trunkering med avseendee på celler. Faktiskt utfall (heldragen linje), skattade värden med (62) (kortstreckad linje) och observerade värden (långstreckad linje).



Figur 14: Product Limit Estimate av den betingade sannolikhetsfunktionen $P(X \leq x | X \leq \tau)$, beräknad med proceduren PHREG i SAS® 9.1.



Figur 15: Faktiskt utfall (fet linje), skattade värden med (62) och 13 månaders trunkering (kortstreckad linje) samt observerade värden och 14 månaders trunkering (långstreckad linje).



Figur 16: Faktiskt utfall (fet linje), skattade värden med (62) och 12 månaders trunkering (kortstreckad linje) samt skattade värden med (62) och 13 månaders trunkering (heldragen linje).

A.3 Ordlista

Egnahem,
: Hus för permanentboende, innefattande en- och tvåfamiljevillor samt rad- och kedjehus.

Fastighet,
: Markområde som enligt jordbalken utgör fast egendom.

Fångeskod,
: Kod som anger hur en fastighet har förvärvats, exempelvis vanligt köp, släktköp, byte, gåva, arv m.fl.

Köpebrev,
: Bevis på att villkor för förvärv av fastighet är uppfyllda samt kvitto för betalning av fastighet.

Lagfart,
: Inskrivning i fastighetsregistret av lagfaren ägare och bevis på äganderätt av fastighet.

Småhus,
: Byggnad avsedd för boende, indelad i egnahem och fritidshus.

Referenser

Agresti, A. *Categorcial data analysis*. John Wiley & Sons, andra utgåvan, 2002.

Bishop, Y. M. M., Feinberg, S. E, och Holland, P. *Discrete Multivariate Analysis: Theory And Practice*. Springer, första utgåvan, 1975. Ett samarbete med R.J. Light och F. Mosteller.

Claims reserving manual. Faculty and institute of actuaries, `http://www.actuaries.org.uk/general_insurance/documents/crm`, September 1997. volume 1.

Hedlin, D., Fenton, T., McDonald, J. W., Pont, M., och Wang, S. Estimating the undercoverage of a sampling frame due to reporting delays. *Journal of official statistics*, 22(1):53–70, 2006.

Kalbfleisch, J. D. och Lawless, J. F. Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 1:19–32, 1991a.

Kalbfleisch, J. D., Lawless, J. F., och Robinson, J. A. Methods for the analysis and prediction of warranty claims. *Technometrics*, 33(3):273–285, 1991b.

Kaminsky, K. S. Prediction of IBNR claims count by modelling the distribution of reporting lags. *Insurance: Mathematics and Economics*, 6:151–159, 1986.

Klein, J. P. och Moeschberger, M. L. *Survival analysis: Techniques for censored and truncated data*. Statistics for biology and health. Springer-Verlag, 1997.

Kremer, E. IBNR-claims and the two-way model of ANOVA. *Scandinavian Actuarial Journal*, 1:47–55, 1982.

Lagakos, S. W., Barraj, L. M., och De Gruttola, V. Nonparametric analysis of truncated survival data with application to AIDS. *Biometrika*, 75(3): 515–523, 1988.

Lawless, J. F. Adjustment for reporting delays and the prediction of occurred but not reported events. *The Canadian journal of statistics*, 22(1):15–31, 1994.

Linkletter, C. D. och Sitter, Randy R. Predicting natural gas production in Texas: An application of nonparametric reporting lag distribution estimation. *Journal of official statistics*, 23(2):239–251, 2007.

Mack, T. Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23(2):213–225, 1993.

McCullagh, P. och Nelder, J. A. *Generalized Linear Models*. Chapman & Hall/CRC, andra utgåvan, 1989.

Renshaw, A. E. och Verall, R. J. A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, 4:903–923, 1998.

Ross, S. M. *Introduction to probability models*. Academic Press, nionde utgåvan, 2007.

Sellero, C. S., Fernández, E. V., Manteiga, W. G., Otero, X. L., Hervada, X., Fernández, E., och Taboada, X. A. Reporting delay: A review with a simulation study and application to Spanish AIDS data. *Statistics in medicine*, 15:305–321, 1996.

Sundberg, R. Kompendium i tillämpad matematisk statistik, December 1997.

Zehnwirth, B. *Claims Reserving manual*, vol. 2, kapitel D1: The chain ladder technique - A stochastic model. Faculty and institute of actuaries, `http://www.actuaries.org.uk/general_insurance/documents/crm`, September 1997.

Muntliga referenser

Martin Verhage, *Produktansvarig: Fastighetspris- och lagfartsstatistik*, Statistiska Centralbyrån.

Internet

`http://www.lantmateriet.se`
`http://www.scb.se`