Over-dispersed age-period-cohort models

J. Harnau[1] and B. Nielsen[2]
Department of Economics, University of Oxford

31 July 2017

Summary We consider inference and forecasting for aggregate data organised in a two-way table with age and cohort as indices, but without measures of exposure. This is modelled using a Poisson likelihood with an age-period-cohort structure for the mean while allowing for over-dispersion. We propose a repetitive structure that keeps the dimension of table fixed while increasing the latent exposure. For this we use a class of infinitely divisible distributions which include a variety of compound Poisson models and Poisson mixture models. This results in asymptotic F inference and t forecast distributions.

Keywords Chain-ladder model; Forecasting; Generalized linear model; Inference; Infinitely divisible, Two-way table.

1 Introduction

Over-dispersion is often a serious complication in the analysis of two-way tables. We consider the case of a two-way table with two features. First, the indices of the table are two time scales, cohort and age, so that we may be interested in forecasting for combinations of age and cohort that are not observed. Second, there is no information about the exposure. Examples include data with a reporting delay such as AIDS diagnosis (Davison & Hinkley 1997, pages 342–346), asbestos caused mesothelioma deaths (Martínez Miranda et al. 2015), and reserving in non-life insurance (England & Verrall 2002). Closely related examples are mortality data with observed exposure (Alai & Sherris 2014) and reserving with continuous time information (Lee et al. 2015). A basic model is a Poisson model with an age-period-cohort predictor. When faced with over-dispersion there are two strategies: either to change the distribution or to work with a correction factor. The second route is attractive in this case where it is hard to choose an alternative distribution with confidence due to a high parameter to observation ratio. Even so, a model is needed to justify such a correction. We suggest a sampling scheme based on infinitely divisible distributions which include Poisson mixtures such as the negative binomial distribution as well as compound Poisson distributions. This leads to asymptotic inference and distribution forecasts based on standard quasi-Poisson statistics combined with F and t asymptotics. The results apply to data arrays of the generalized trapezoid type, see equation (2) below. These include rectangular arrays in age-cohort, age-period and period-cohort space as well as triangular age-cohort arrays called run-off triangles.

When there is no over-dispersion we can apply a multinomial sampling scheme, since conditioning on the totals in a Poisson table gives a multinomial distribution, see

[1]Oriel College, Oxford, OX1 4EW, U.K.; jonas.harnau@oriel.ox.ac.uk
[2]Nuffield College, Oxford OX1 1NF, U.K.; bent.nielsen@nuffield.ox.ac.uk

Fisher (1922), Agresti (2013, §1.2.5, 9.6.8). Recently, Martínez Miranda et al. (2015) have exploited this idea to solve the inference and forecasting problem for a Poisson age-period-cohort model. The conditioning solution relies on the very particular link between the Poisson and multinomial distributions. This falls away in the over-dispersed case, so we need another solution.

The more principled way to address the over-dispersion problem is to formulate an alternative to the Poisson distribution. A classic solution is a negative binomial model as explored by Bliss & Fisher (1953), Cox (1983), Agresti (2013, §14). This works well in situations with many repetitions and relatively few parameters unlike the present scenario with a high parameter to observation ratio. Another solution is to use an exponential dispersion model. Jørgensen (1987) shows that F-type inference applies under small-dispersion asymptotics. The class of exponential dispersion models is restrictive, however, since no exponential dispersion model with support on the integers exists (Jørgensen 1986).

An alternative way to address the over-dispersion problem is to work with correction factors. There are many warnings against this approach as opposed to modelling of the distribution, for example Venables & Ripley (2002, §7.5). The attraction is, however, that we use the Poisson likelihood as a quasi-likelihood (Wedderburn 1974). Much applied work is carried out this way. Indeed, the quasi-Poisson approach is fundamental to reserving in non-life insurance where it is known as the chain ladder method (England & Verrall 2002). Widely used bootstrap solutions have been developed for the quasi-likelihood by Davison & Hinkley (1997, pages 342–346) and England (2002); see also Pinheiro et al. (2003). These bootstrap solutions are, however, not based in a formal model of the over-dispersion. We therefore formulate a class of infinite divisible distributions where the log mean has an age-period-cohort structure and the variance-to-mean ratio is constant across cell while the skewness vanishes as the sum of the data increases. This setup includes the Poisson distribution as well as classic over-dispersion distributions, such as the negative binomial distributions that is useful for heterogeneous populations and the compound Poisson distribution that is relevant for reserving data. Within this framework we can formally derive F-type inference as well as t-type forecast distributions. A simulation study indicates that the bootstrap solutions perform well within this framework, indicating that the model framework might be amenable to a theory for the bootstrap.

A feature of the proposed class of infinite divisible distributions is that the variance-to-mean ratio is constant across the cells in the table, which is the defining feature of over-dispersed Poisson distribution. In cases with severe over-dispersion an alternative would be to apply a log-normal distribution, in which case the standard deviation-to-mean ratio would be held constant. It would be useful to develop tests to distinguish between these situations for two-way data. In reserving it is common to apply a compound Poisson interpretation for the data (Beard et al. 1984, §3.2) hence the relevance of over-dispersed Poisson distributions. Log-normal age-period-cohort models are, however, also in used to some extent in insurance (Barnett & Zehnwirth 2000, Kuang et al. 2011).

We focus on an age-period-cohort structure for the log-mean of the data, but note that the results also extend to more standard contingency tables. The age-period-

cohort model provides an interesting focal point due to time series interpretation of the parameters and the wide use of this model in demography, economics, epidemiology, sociology and actuarial science. Recent statistical developments of the model include an asymptotic analysis of a class of constrained estimators when the dimension of the array increases (Fu 2016), non-parametric, continuous time variation (Lee et al. 2015) and Bayesian estimation (Smith & Wakefield 2016). The age-period-cohort specification of the log mean or linear predictor is

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta, \tag{1}$$

where i, k are age and cohort indices while $j = i + k - 1$ is the period. The effects α_i, β_j, γ_k and δ are not identified. We reparametrise the log mean in terms of freely varying parameters as suggested by Kuang et al. (2008*b*). A Poisson model becomes a regular exponential family in this way where the freely varying parameters are canonical. One of the parameters measures the level of the data. This is taken to be large in the asymptotic analysis so that the expectations of the data grow proportionally. The other parameters measure contrasts. These are invariant to recursive analysis and are assumed fixed in the limiting experiment. This relates to a mixed parametrization in the sense of Barndorff-Nielsen (1978, Theorem 8.4). Other features of the parametrization are discussed in Nielsen & Nielsen (2014). Different parametrizations would give the same fit, but asymptotic analysis is naturally formulated in terms of the mixed parametrization with freely varying parameters.

We illustrate the results on an insurance run-off triangle. Insurers use these to forecast incurred but not fully reported liabilities. Typically contracts run for a year but liabilities may not be settled for several years. Publicly available triangles are provided by for instance Casualty Actuarial Society (2016). We apply the triangle of Taylor & Ashe (1983) as shown in Table 1. The entries are aggregate paid amounts for cohort (accident year) k in age (development year) i with period (calendar year) j along the diagonals. The two-way table results from the delay between accident and payment. The insurance problem is to forecast the incurred but not yet paid amounts in the empty lower triangle. Row-sums in the lower triangle are payments related to particular accident years and commonly called reserves. Diagonal-sums correspond to payments in specific calendar years and thus represent the future cash-flow. The sum over all cells in the lower triangle is called the total reserve.

In §2 we derive a limit theorem for infinitely divisible distributions. In §3 we set up the model: we describe the data structure, state the assumptions, consider identification, estimation, and the sampling scheme. We derive the distribution of estimators and test statistics in §4 and results for forecasts that do not require parameter extrapolation in §5. We apply the results in a data-example in §6. The simulation study in §7 shows that the asymptotic results give good approximations in finite samples. Finally, we discuss directions for future research in §8 where we also briefly consider forecasting with parameter extrapolation.

k, i	1	2	3	4	5	6	7	8	9	10
1	357848	766940	610542	482940	527326	574398	146342	139950	227229	67948
2	352118	884021	933894	1183289	445745	320996	527804	266172	425046	
3	290507	1001799	926219	1016654	750816	146923	495992	280405		
4	310608	1108250	776189	1562400	272482	352053	206286			
5	443160	693190	991983	769488	504851	470639				
6	396132	937085	847498	805037	705960					
7	440832	847631	1131398	1063269						
8	359480	1061648	1443370							
9	376686	986608								
10	344014									

Table 1: Insurance run-off triangle. The entries are aggregate paid amounts at age (development year) i for claims of cohort (accident year) k. Periods (calendar years) are on the diagonals increasing from the top left.

2 Infinite divisibility

Martínez Miranda et al. (2015) proposed an age-period-cohort model for mesothelioma mortality. In their model, the cells Y_{ik} are independent Poisson distributed. They condition on the data sum $Y_{\cdot\cdot}$ and use a multinomial sampling scheme. This approach does not extend easily to over-dispersed data. Instead, we work with non-negative infinitely divisible distributions.

Recall that a distribution D is infinitely divisible if for any $m \in N$ there are independent identically distributed random variables $X_1, \ldots, X_m$ such that $\sum_{\ell=1}^{m} X_\ell$ has distribution D. For the present applications, non-negativity seems reasonable. Examples of such distributions include Poisson, compound Poisson, negative binomial (Johnson et al. 2005, pages 164, 388, 218), log normal (Thorin 1977), gamma and generalized gamma convolutions (Thorin 1977, Bondesson 2015). Infinitely divisible distributions are linked to Lévy processes (Sato 1999, Theorem 7.10). Non-negative Lévy processes are known as subordinators. The following results lie at the heart of our asymptotic theory.

Theorem 1 *Let $\{Y_\ell\}$ be a sequence of random variables with non-degenerate, non-negative infinitely divisible distributions with at least three moments. If the skewness vanishes, $skew(Y_\ell) = E[\{Y_\ell - E(Y_\ell)\}/\sqrt{\mathrm{var}(Y_\ell)}]^3 \to 0$, then, in distribution,*

$$\frac{Y_\ell - E(Y_\ell)}{\sqrt{\mathrm{var}(Y_\ell)}} \to N(0,1).$$

Theorem 2 *If the conclusion of Theorem 1 is met and the ratio of mean to standard deviation increases, $E(Y_\ell)/\sqrt{\mathrm{var}(Y_\ell)} \to \infty$, then $Y_\ell/E(Y_\ell) \to 1$ in probability.*

For some distributions, such as the Poisson or negative binomial, the skewness vanishes if and only if the mean increases. This is not a necessary condition. For a log normal variable, the skewness vanishes if and only if the variance of the associated normal distribution vanishes; similarly, the gamma requires the shape to increase. Neither

requires the mean to grow. For all these examples, the ratio of mean to standard deviation grows if and only if the skewness vanishes.

A more complicated example is the compound Poisson distribution $\sum_{m=1}^{Z_\ell} X_{m,\ell}$ where Z_ℓ is Poisson distributed and, for each ℓ, the jumps $X_{m,\ell}$ are non-negative independent identically distributed across m with at least three moments, independent of Z_ℓ. A special case arises when the jump distribution does not depend on ℓ. Then, a necessary and sufficient condition for the skewness to vanish and the mean to standard deviation ratio to grow is that $E(Z_\ell)$ becomes large.

3 Model

3.1 Data

Due to the wide range of applicability, age-period-cohort data arrays take different forms. In a mortality setting, Keiding (1990) summarizes the three principle sets of dead related to Lexis diagrams. These are data organized as rectangles in an age-cohort array, a cohort-period array or an age-period array. The latter two form trapezoids in an age-cohort array. Insurance reserving data known as run-off triangles are triangular age-cohort arrays. The three principle sets of dead and insurance run-off triangles are special cases of generalized trapezoid data arrays

$$\mathcal{I} = (i, k : 1 \leq i \leq I, 1 \leq k \leq K, L + 1 \leq j \leq L + J), \tag{2}$$

where I, J and K indicate the numbers of age, period and cohort indices available while $L + 1$ is the lower period index (Kuang et al. 2008*b*). The number of elements of $\mathcal{I}$ is the number of observations, n. Table 1 is a generalized trapezoid with $I = K = J = 10$, $L = 0$ and $n = 55$.

3.2 Assumptions

We define the over-dispersed Poisson model with age-period-cohort structure. Consider observations Y_{ik} for $(i, k) \in \mathcal{I}$ where $\mathcal{I}$ is a generalized trapezoid as in (2). We assume that the Y_{ik} are independent with non-degenerate and non-negative infinitely divisible distribution with at least three moments. Moreover, suppose $E(Y_{ik}) = \exp(\mu_{ik})$ where μ_{ik} satisfies the age-period-cohort structure (1) while variance and mean are proportional so $\text{var}(Y_{ik})/E(Y_{ik}) = \sigma^2 > 0$. A Poisson model satisfies this with $\sigma^2 = 1$.

The model has no explicit assumptions to the unobserved exposure. However, as we consider aggregates, the data need to be on the same scale. That is, we either need population data or a representative sample; this would be violated if some age-cohort groups are over-represented in the sample. When modelling vital data one will sometimes be interested in mortality rates. This would be modelled by conditioning on exposure. The present model does not give information about the rates unless the exposure and the rates have a separable structure; see Martínez Miranda et al. (2015) for further discussion.

3.3 Identification

It is well known that the age, period and cohort effects α_i, β_j, γ_k and the δ are not identified. Kuang et al. (2008*b*) proposed an identified parametrization in terms of three initial points and three sets of double differences. The parameter vector is $\xi = \{\xi^{(1)}, (\xi^{(2)})^T\}^T$ where $\xi^{(1)} = \mu_{\ell m}$ and, with $\nu_a = \mu_{\ell^\dagger m} - \mu_{\ell m}$ and $\nu_c = \mu_{\ell m^\dagger} - \mu_{\ell m}$ for distinct (ℓ, m), $(\ell^\dagger, m)$, $(\ell, m^\dagger)$ in $\mathcal{I}$,

$$\xi^{(2)} = (\nu_a, \nu_c, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_I, \Delta^2\beta_{L+3}, \ldots, \Delta^2\beta_{L+J}, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_K)^T,$$

so ξ has length $p = I + J + K - 3$. Thus, p grows with n. Generally, μ_{ik} is a linear function of ξ of the form

$$\mu_{ik} = x_{ik}^{(1)}\xi^{(1)} + (x_{ik}^{(2)})^T\xi^{(2)},$$

where $x_{ik}^{(1)} = 1$ while the $p-1$ vector $x_{ik}^{(2)}$ depends on the choice of the data array $\mathcal{I}$. Kuang et al. (2008*b*, Corollary 2) show that if $\xi \neq \xi^\dagger$ then $\mu_{ik}(\xi) \neq \mu_{ik}(\xi^\dagger)$. They also note that μ_{ik} is identified. Martínez Miranda et al. (2015) point out that the double differences have log-odds ratio interpretation.

As an example, for rectangular or triangular age-cohort data arrays so $L = 0$ in (2), Kuang et al. (2008*b*) suggest to represent μ_{ik} as

$$\begin{aligned}\mu_{ik} = \mu_{11} + (i-1)(\mu_{21} - \mu_{11}) + (k-1)(\mu_{12} - \mu_{11}) \\ + \sum_{t=3}^{i}\sum_{s=3}^{t}\Delta^2\alpha_s + \sum_{t=3}^{i+k-1}\sum_{s=3}^{t}\Delta^2\beta_s + \sum_{t=3}^{k}\sum_{s=3}^{t}\Delta^2\gamma_s.\end{aligned} \tag{3}$$

Then, the initial points can be taken as a point $\xi^{(1)} = \mu_{11}$ while the slopes in the age and cohort directions are $\nu_a = \mu_{21} - \mu_{11}$ and $\nu_c = \mu_{12} - \mu_{11}$. Taken together these three terms determine a linear plane. The three double sums of double differences represent time effects constrained to zero for their first two values. The key to this representation is that the three time scales age, period and cohort all increase from the coordinate $i = k = j = 1$. For other data arrays the presentation has a more tedious appearance. Two different solutions are proposed in Martínez Miranda et al. (2015) and in Nielsen (2015).

Ad-hoc identification of the time effects α_i, β_j, γ_k, can be done in three ways. We discuss an example for each. First, Nielsen (2015) suggests a representation of μ_{ik} in terms of a linear plane and time effects that are detrended to start and end in zero. This decouples the time effects that can be interpreted individually. There is a bijective map from $\xi^{(2)}$ to the linear slopes and the detrended time effects. Second, one may employ a restriction such as $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = \beta_2 = 0$. Now, the linear slopes are distributed onto the time effects in a particular way and the three time effects cannot be interpreted individually. There is now an injective map from $\xi^{(2)}$ to the three time effects and the exponential family is no longer regular; see Nielsen & Nielsen (2014). The presented asymptotic theory covers these two cases. Third, if the identification restricts the intercept, for example $\delta = 0$, the time effects are functions of both $\xi^{(1)}$ and $\xi^{(2)}$; this is outside the scope of the asymptotic theory. None of these identification schemes is amenable to recursive analysis: for example expanding the data array and

adding observations for a newly observed period changes the constraints and thus the parameters.

3.4 Estimation in a Poisson model

A Poisson model satisfies the assumptions in §3.2 without over-dispersion so $\sigma^2 = 1$. The model is a regular exponential family with canonical parameter ξ, log likelihood

$$\ell_Y(\xi) = Y_{\cdot\cdot}\xi^{(1)} + (T^{(2)})^T\xi^{(2)} - \exp(\xi^{(1)})\sum_{ik\in\mathcal{I}}\exp\{(x_{ik}^{(2)})^T\xi^{(2)}\}$$

and minimal sufficient statistic given by $Y_{\cdot\cdot}$ and $T^{(2)} = \sum_{ik\in\mathcal{I}} Y_{ik}x_{ik}^{(2)}$. The information is

$$i_\xi = -\frac{\partial^2}{\partial\xi\partial\xi^T}\ell_Y(\xi) = \sum_{ik\in\mathcal{I}}\exp(\mu_{ik})\begin{pmatrix}1\\ x_{ik}^{(2)}\end{pmatrix}\begin{pmatrix}1\\ x_{ik}^{(2)}\end{pmatrix}^T. \tag{4}$$

The maximum likelihood estimator is unique if and only if $(Y_{\cdot\cdot}, T^{(2)})$ takes a value in the interior of its convex support (Barndorff-Nielsen 1978, Theorem 9.13).

Kuang et al. (2009) analyze this condition when $\mathcal{I}$ is triangular and the period parameter absent, $\Delta^2\beta_s = 0$. In this special case, the estimators have closed form expressions.

3.5 Mixed parametrization of the Poisson model

Martínez Miranda et al. (2015) consider a Poisson age-period-cohort model. They condition on the data sum and base asymptotics on a multinomial sampling scheme, keeping the array dimension and consequently the number of parameters fixed. The cost is that asymptotic inference on the overall mean is not possible.

The link between Poisson and multinomial model can be made explicit. The Poisson model has mixed parametrization given by $\psi = \{\tau, (\xi^{(2)})^T\}^T$, where

$$\tau = E(Y_{\cdot\cdot}) = \exp(\xi^{(1)})\sum_{ik\in\mathcal{I}}\exp\{(x_{ik}^{(2)})^T\xi^{(2)}\} \tag{5}$$

is the aggregate mean. The mapping from ξ to ψ is homeomorph and the parameters τ and $\xi^{(2)}$ are variation independent (Barndorff-Nielsen 1978, Theorem 8.4). The reparametrized log likelihood is

$$\ell_Y(\psi) = \ell_{Y_{\cdot\cdot}}(\tau) + \ell_{T^{(2)}|Y_{\cdot\cdot}}(\xi^{(2)}) \tag{6}$$

where

$$\ell_{Y_{\cdot\cdot}}(\tau) = Y_{\cdot\cdot}\log(\tau) - \tau, \quad \ell_{T^{(2)}|Y_{\cdot\cdot}}(\xi^{(2)}) = (T^{(2)})^T\xi^{(2)} - Y_{\cdot\cdot}\log[\sum_{ik\in\mathcal{I}}\exp\{(x_{ik}^{(2)})^T\xi^{(2)}\}].$$

Here, $\ell_{Y_{\cdot\cdot}}$ is a Poisson likelihood for τ based on $Y_{\cdot\cdot}$ and $\ell_{T^{(2)}|Y_{\cdot\cdot}}$ a multinomial likelihood for $\xi^{(2)}$ based on $T^{(2)}$ and conditional on $Y_{\cdot\cdot}$. An implication is that Poisson and multinomial

log likelihood ratios coincide for unrestricted τ. The maximum likelihood estimator for τ is $\hat{\tau} = Y_{\cdot\cdot}$. The estimator for $\xi^{(2)}$ can be obtained either from the multinomial likelihood or by dropping the first element from the Poisson regression estimator for ξ. The model by Martínez Miranda et al. (2015) does not allow inference on τ which goes to infinity, but inference on $\xi^{(2)}$ is feasible.

The corresponding observed information is closely linked to the expected information i_ξ in (4). To see this, introduce the frequencies

$$\pi_{ik} = \frac{E(Y_{ik})}{\tau} = \frac{\exp\{(x_{ik}^{(2)})^T \xi^{(2)}\}}{\sum_{ik\in\mathcal{I}} \exp\{(x_{ik}^{(2)})^T \xi^{(2)}\}}, \tag{7}$$

which are functions of $\xi^{(2)}$. The average information about $\xi^{(2)}$ is

$$\bar{\imath}_{\xi^{(2)}} = -\hat{\tau}^{-1} \frac{\partial^2}{\partial \xi^{(2)} \partial (\xi^{(2)})^T} \ell_Y(\psi) = \sum_{ik\in\mathcal{I}} \pi_{ik} H_{ik} H_{ik}^T, \quad H_{ik} = x_{ik}^{(2)} - \sum_{lm\in\mathcal{I}} \pi_{lm} x_{lm}^{(2)},$$

so that the inverse information $(\tau \bar{\imath}_{\xi^{(2)}})^{-1}$ equals the bottom right element of i_ξ^{-1}. The observed information on the mixed parameter can now be written as

$$j_\psi = -\frac{\partial^2}{\partial \psi \partial \psi^T} \ell_Y(\psi) = \hat{\tau} \begin{pmatrix} \tau^{-2} & 0 \\ 0 & \bar{\imath}_{\xi^{(2)}} \end{pmatrix}. \tag{8}$$

3.6 Estimation in an over-dispersed Poisson model

In an over-dispersed model σ^2 is left unrestricted. Then, the scaled log likelihood $\sigma^2 \ell_Y$ is a quasi-likelihood in the sense of Wedderburn (1974) with the Poisson likelihood as objective function. Thus, properties resulting from the functional form of the Poisson likelihood such as the variation independence in the mixed parametrization are still valid. The Poisson estimators for τ and $\xi^{(2)}$ coincide with the quasi-likelihood estimators. The mixed parametrization makes the derivation of the asymptotic theory below easier and more insightful due to the diagonal structure of the information. For applications, however, there is no need to estimate a multinomial model: as we showed above multinomial estimates for $\xi^{(2)}$ are simply the last $p-1$ parameters of the Poisson estimate, the estimate for τ is the data sum, Poisson and multinomial log-likelihood ratios coincide, and the inverse average multinomial information $(\bar{\imath}_{\xi^{(2)}})^{-1}$, playing a role in the results below, does not require extra computation, being the bottom right block of i_ξ^{-1}/τ.

3.7 Sampling scheme in an over-dispersed Poisson model

Consider the over-dispersed Poisson model described in §3.2. Unlike the Poisson model, the over-dispersed Poisson model does not allow for conditioning. Our sampling scheme stipulates that the index set $\mathcal{I}$ and the frequencies π_{ik} are fixed, while τ increases in such a way that $skew(Y_{ik})$ vanishes. Then, Theorems 1 and 2 apply and we can make asymptotic inference about $\xi^{(2)}$ but neither about τ nor $\xi^{(1)}$; the latter follows from (5) since $\xi^{(1)}$ is increasing in τ for fixed $\xi^{(2)}$. An example is a compound Poisson

distributed array $Y_{ik} = \sum_{m=1}^{Z_{ik}} X_{ikm}$ where the means of the Poisson counts Z_{ik} grow proportionally. The advantage of this sampling scheme, compared to one with increasing array dimension, is that the number of parameters is fixed.

We note that in the Poisson model, which is a special case of the over-dispersed model, we have from (8) that the expected information about τ is τ^{-1} while that for $\xi^{(2)}$ is $\tau\bar{\imath}_{\xi^{(2)}}$. Hence, they move in opposite directions as τ increases. Thus, decompose the expected information so

$$E(j_\psi) = \tau\bar{\imath}_\psi = \tau M_\tau \tilde{\imath}_\psi M_\tau, \quad M_\tau = \begin{pmatrix} \tau^{-1} & 0 \\ 0 & I \end{pmatrix}, \quad \tilde{\imath}_\psi = \begin{pmatrix} 1 & 0 \\ 0 & \bar{\imath}_{\xi^{(2)}} \end{pmatrix}.$$

M_τ is a normalization matrix and $\tilde{\imath}_\psi$ the normalized average information that is invariant to τ.

4 Inference

We derive asymptotic distributions for quasi-likelihood estimators and test statistics for hypotheses about $\xi^{(2)}$. For $\xi^{(2)} \in R^{p-1}$, $\zeta^{(2)} \in R^{q-1}$ and $\varphi^{(2)} \in R^{r-1}$ with $r \leq q \leq p$ we consider nested smooth hypotheses (Johansen 1979, page 39)

$$H_{apc} : \mu_{ik} = \xi^{(1)} + (x_{ik}^{(2)})^T \xi^{(2)}, \quad H_1 : \xi^{(2)} = \xi^{(2)}(\zeta^{(2)}), \quad H_2 : \zeta^{(2)} = \zeta^{(2)}(\varphi^{(2)}).$$

H_{apc} is the age-period-cohort model. H_1 restricts to a sub-model such as an age-cohort model $\mu_{ik} = \alpha_i + \gamma_k + \delta$ in which $\zeta^{(2)}$ is $\xi^{(2)}$ with $\Delta^2\beta = 0$. H_2 restricts to another nested sub-model such as an age model $\mu_{ik} = \alpha_i$ so $\varphi^{(2)}$ is $\zeta^{(2)}$ with $\Delta^2\gamma = \nu_c = 0$. An overview of linear sub-models is given in Nielsen (2015).

Since τ is unrestricted for the hypothesis considered, Poisson and multinomial log likelihood ratio statistics and deviances coincide, irrespective of the identification method for the time trends since the deviances are functions of the identified μ_{ik}. Let LR_{st} be the log likelihood ratio statistic for H_s against H_t and D_s be the deviance for H_s, that is the log likelihood ratio against the saturated model where μ_{ik} is completely unrestricted. The asymptotic distribution of the estimators and test statistics in the over-dispersed model is as follows.

Lemma 1 *In the over-dispersed Poisson model of §3.2 and §3.7, in distribution,*

$$\hat{\tau}^{1/2} M_{\hat{\tau}}(\hat{\psi} - \psi) = \begin{Bmatrix} \hat{\tau}^{-1/2}(\hat{\tau} - \tau) \\ \hat{\tau}^{1/2}(\hat{\xi}^{(2)} - \xi^{(2)}) \end{Bmatrix} \to N\{0, \sigma^2(\tilde{\imath}_\psi)^{-1}\}.$$

D_{apc}, $LR_{1,apc}$ and $LR_{2,1}$ are asymptotically independent $\sigma^2\chi^2$ with $n-p$, $p-q$ and $q-r$ degrees of freedom, respectively. $\hat{\tau}^{1/2} M_{\hat{\tau}}(\hat{\psi}-\psi)$ and D_{apc} are asymptotically independent.

We note that no consistent estimator for τ is available under the sampling scheme; in the Poisson special case this is reflected by the vanishing information about τ prior to normalization by $M_{\hat{\tau}}$. With $\sigma^2 = 1$, the distributions match those in a Poisson model as well as, leaving $\hat{\tau}^{-1/2}(\hat{\tau} - \tau)$ aside, a multinomial model conditional on $Y_{\cdot\cdot}$. We can

exploit the asymptotic distribution of $D_{apc}/(n-p)$ with expectation σ^2 to find statistics that are asymptotically invariant to σ^2.

Theorem 3 *In the over-dispersed Poisson model of §3.2 and §3.7, in distribution,*

$$\hat{\tau}^{1/2}\frac{v^T(\hat{\xi}^{(2)}-\xi^{(2)})}{\{D_{apc}/(n-p)\}^{1/2}} \to \{v^T(\bar{\imath}_{\xi^{(2)}})^{-1}v\}^{1/2}t_{n-p}, \quad \textit{for all } v \in R^{p-1}.$$

In particular, the distribution of elements of the estimator is approximately proportional to a t_{n-p} distribution. Theorem 3 applies to many, but not all, ad-hoc identified parametrizations. If the identification does not constrain the intercept δ and the identified time effects α, β, γ are linear injective functions of $\xi^{(2)}$, then Theorem 3 applies.

The next theorem allows independent successive testing of H_1 and H_2.

Theorem 4 *In the over-dispersed Poisson model of §3.2 and §3.7, in distribution,*

$$\mathrm{F}_1 = \frac{LR_{1,apc}/(p-q)}{D_{apc}/(n-p)} \to F_{p-q,n-p}, \quad \mathrm{F}_2 = \frac{LR_{2,1}/(q-r)}{D_1/(n-q)} \to F_{q-r,n-q}.$$

F_1 *and* F_2 *are asymptotically independent.*

The models we consider typically have a high parameter to observation ratio. One could wonder how much the degree of over-dispersion depends on the specific hypothesis. Theorem 4 gives some insight to this. Given a valid restriction $E(\mathrm{F}_1)$ is close to one as the F_{v_1,v_2} distribution has mean $v_2/(v_2-2)$. In particular, $\mathrm{F}_1 = 1$ is equivalent to $D_1/(n-q) = D_{apc}/(n-p)$, noting that $LR_{1,apc} = D_1 - D_{apc}$, so the over-dispersion should not change much by imposing valid restrictions. Imposing invalid restrictions would by the same argument lead to an increase in over-dispersion. In applications, one would first compare D_{apc} with a χ^2_{n-p}, effectively asking if a Poisson model is appropriate. If this is large, for instance if $D_{apc}/(n-p) = 2$, for sufficiently large degrees of freedom, say ten, we would reject a Poisson model and switch to an over-dispersed model. Confidence bands are then about 50% wider compared to a Poisson model. With ten degrees of freedom for $D_{apc}/(n-p) = 1.5$ we would not reject the Poisson model; here the over-dispersed confidence bands would have been some 25% wider.

5 Forecasting

5.1 Assumptions

We consider a forecasting array that is triangular in age-cohort space:

$$\mathcal{J} = (i,k : 1 \le i \le I, 1 \le k \le K, L+J+1 \le j \le I+K-1).$$

In Table 1, the forecasting array $\mathcal{J}$ is the empty lower triangle. Forecasting arrays of this type are not only of interest for run-off triangles, but also arise naturally for data that is rectangular in age-period or period-cohort space. We assume that the over-dispersed Poisson model in §3.2 is satisfied out of sample for $(i,k) \in \mathcal{J}$. We consider an age-cohort

model $\mu_{ik} = \alpha_i + \gamma_k + \delta$ and denote the restricted parameter vector by ζ. With this, any parameters in the forecasting array $\mathcal{J}$ also appear in the data array $\mathcal{I}$ so parameter extrapolation is not necessary. Lee et al. (2015) refer to this as in-sample forecasting. Models with period effect or forecasting arrays that are not triangular in age-cohort space generally require parameter extrapolation; see the discussion in §8.

5.2 Point forecasting

We may be interested in forecasting individual cells as well as sums of cells over any subset $\mathcal{A} \subseteq \mathcal{J}$. Summations over $(i,k) \in \mathcal{A}$ are indicated by the subscript $\mathcal{A}$. Point forecasts for $E(Y_\mathcal{A}) = \tau\pi_\mathcal{A}(\zeta^{(2)})$ are $\widetilde{Y}_\mathcal{A} = \hat{\tau}\pi_\mathcal{A}(\hat{\zeta}^{(2)})$. The point forecasts are not consistent under the sampling scheme but $\widetilde{Y}_\mathcal{A}/E(Y_\mathcal{A}) \to 1$. We note that $\pi_\mathcal{A}$ does not have interpretation as a frequency outside the index set $\mathcal{I}$.

5.3 Distribution forecasting

The aim is to predict the distribution of the difference between the point forecast $\widetilde{Y}_\mathcal{A}$ and the realisation $Y_\mathcal{A}$. Defining $\hat{\pi}_\mathcal{A} = \sum_{ik\in\mathcal{A}} \pi_{ik}(\hat{\zeta}^{(2)})$ with π_{ik} as in (7) we find three contributions for the forecast error:

$$Y_\mathcal{A} - \widetilde{Y}_\mathcal{A} = Y_\mathcal{A} - E(Y_\mathcal{A}) - \hat{\tau}(\hat{\pi}_\mathcal{A} - \pi_\mathcal{A}) - (\hat{\tau} - \tau)\pi_\mathcal{A}. \tag{9}$$

The first contribution is the process error which, extending Theorems 1 and 2, satisfies

$$\hat{\tau}^{-1/2}\{Y_\mathcal{A} - E(Y_\mathcal{A})\} \to N(0, \sigma^2\pi_\mathcal{A}).$$

The second contribution is the estimation error for $\zeta^{(2)}$. By Lemma 1 and the δ-method,

$$\hat{\tau}^{1/2}(\hat{\pi}_\mathcal{A} - \pi_\mathcal{A}) \to N(0, \sigma^2 s_\mathcal{A}^2)$$

where

$$s_\mathcal{A}^2 = (\sum_{ik\in\mathcal{A}} \pi_{ik}H_{ik})^T(\bar{\imath}_{\zeta^{(2)}})^{-1}(\sum_{ik\in\mathcal{A}} \pi_{ik}H_{ik}). \tag{10}$$

The third contribution pertains the estimation uncertainty for τ. By Lemma 1,

$$\hat{\tau}^{-1/2}(\hat{\tau} - \tau)\pi_\mathcal{A} \to N\{0, \sigma^2(\pi_\mathcal{A})^2\}.$$

Using Lemma 1 again to combine, we arrive at the following theorem.

Theorem 5 *In the over-dispersed Poisson model of §3.7 and §5.1, in distribution,*

$$\hat{\tau}^{-1/2}\frac{Y_\mathcal{A} - \widetilde{Y}_\mathcal{A}}{\{D_1/(n-q)\}^{1/2}} \to \{\pi_\mathcal{A} + s_\mathcal{A}^2 + (\pi_\mathcal{A})^2\}^{1/2}t_{n-q}.$$

Martínez Miranda et al. (2015) investigate forecasting in a Poisson model conditional on $Y_{\cdot\cdot}$. Then there is no estimation uncertainty for $\hat{\tau}$ so the third contribution, $(\pi_\mathcal{A})^2$, is switched off.

sub	df_{sub}	D_{sub}	p_{χ^2}	D_{sub}/df_{sub}	$\mathrm{F}_{sub,apc}$	p_F	$\mathrm{F}_{sub,ac}$	p_F	$\mathrm{F}_{sub,ad}$	p_F
apc	28	1395518	0	49840						
ap	36	1780577	0	49460	0.97	0.48				
ac	36	1903014	0	52862	1.27	0.30				
ad	44	2269756	0	51585	1.10	0.40	0.87	0.55		
a	45	2474053	0	54979	1.27	0.28	1.20	0.32	3.96	0.05

Table 2: Deviance analysis of insurance data. D_{sub}/df_{sub} are estimates for σ^2.

6 Data example

We apply the theory to the insurance run-off triangle shown in Table 1. All R (R Core Team 2016) code is given in the supplementary material. We use the R package `apc` (Nielsen 2015).

Table 2 shows a deviance analysis based on Theorem 4. First, we can consider whether a Poisson model with $\sigma^2 = 1$ is appropriate. Under this hypothesis, the deviance of the age-period-cohort (apc) model is χ^2_{28}. This is clearly rejected.

We proceed with the over-dispersed Poisson model. As discussed in §8, for future work it would be of interest to develop specification tests for this model. Given it is correct, the reported F tests show that the model can be reduced to the age-period (ap, $\Delta^2\gamma = 0$), age-cohort (ac, $\Delta^2\beta = 0$), age-drift (ad, $\Delta^2\beta = \Delta^2\gamma = 0$) and age (a, $\Delta^2\beta = \Delta^2\gamma = \nu_c = 0$) model. In this actuarial context, the age-cohort model is our preferred model for forecasting; it is known as the chain ladder model and widely used. The estimates for the over-dispersion parameter D_{sub}/df_{sub} do not vary much among models as expected in light of the discussion after Theorem 4.

Table 3 show the estimated parameters for the age-period-cohort and age-cohort models with $n - p = 28$ and $n - q = 36$ degrees of freedom, respectively. We report standard errors se_N for a Poisson and se_t for an over-dispersed Poisson model. For the age-period-cohort model, se_N are the diagonal elements of $(\hat{\tau}\bar{\imath}_{\hat{\xi}^{(2)}})^{-1}$ evaluated at $\hat{\xi}^{(2)}$ while $se_t = se_N\{D_{apc}/(n-p)\}^{1/2}$ and similarly for the age-cohort model. Studentized estimators are asymptotically standard normal distributed in the Poisson and asymptotically t distributed in the over-dispersed model. 95% critical values for the normal, t_{n-p} and t_{n-q} are 1.96, 2.05 and 2.03, respectively. Estimates for the two models are similar. The Poisson and over-dispersed Poisson models give very different indications of the parameter uncertainty due to the proportionality factors $(D/df)^{1/2}$ that are close to 230.

Figure 1 shows plots of the the age-period-cohort estimates along with point-wise t-standard errors. The plots for the double difference (a-c) show the estimates presented in Table 3. Plots of the detrended time-effects (d-i) follow from a linear transformation of $\xi^{(2)}$. In these, the detrended time effects have interpretation as deviations from a linear plane and can be interpreted separately. Nielsen (2015) offers a more in depth discussion for interpretation of this representation. We notice that standard errors are increasing with age and cohort, and decreasing with period. This is because larger age and cohort indices, and lower period indices are associated with the corners of the data

	apc model			ac model		
	$\hat{\xi}$	se_N	se_t	$\hat{\zeta}$	se_N	se_t
μ_{11}	12.79			12.51		
$\mu_{21} - \mu_{11}$	0.70	0.001	0.22	0.91	0.001	0.12
$\mu_{12} - \mu_{11}$	0.11	0.001	0.25	0.33	0.001	0.13
$\Delta^2\alpha_3$	-0.90	0.001	0.22	-0.87	0.001	0.20
$\Delta^2\alpha_4$	0.01	0.001	0.20	0.02	0.001	0.21
$\Delta^2\alpha_5$	-0.64	0.001	0.23	-0.66	0.001	0.23
$\Delta^2\alpha_6$	0.26	0.001	0.31	0.24	0.001	0.32
$\Delta^2\alpha_7$	0.26	0.002	0.40	0.27	0.002	0.41
$\Delta^2\alpha_8$	-0.29	0.002	0.50	-0.30	0.002	0.51
$\Delta^2\alpha_9$	0.71	0.003	0.64	0.79	0.003	0.66
$\Delta^2\alpha_{10}$	-1.76	0.005	1.06	-1.79	0.005	1.09
$\Delta^2\beta_3$	0.05	0.002	0.46			
$\Delta^2\beta_4$	0.21	0.002	0.42			
$\Delta^2\beta_5$	0.21	0.002	0.34			
$\Delta^2\beta_6$	-0.41	0.001	0.28			
$\Delta^2\beta_7$	0.35	0.001	0.27			
$\Delta^2\beta_8$	-0.56	0.001	0.26			
$\Delta^2\beta_9$	0.56	0.001	0.27			
$\Delta^2\beta_{10}$	-0.08	0.001	0.25			
$\Delta^2\gamma_3$	-0.37	0.001	0.25	-0.34	0.001	0.24
$\Delta^2\gamma_4$	-0.03	0.001	0.25	-0.01	0.001	0.26
$\Delta^2\gamma_5$	-0.01	0.001	0.26	-0.07	0.001	0.27
$\Delta^2\gamma_6$	0.11	0.001	0.28	0.14	0.001	0.28
$\Delta^2\gamma_7$	0.05	0.001	0.29	0.05	0.001	0.29
$\Delta^2\gamma_8$	0.05	0.001	0.30	0.08	0.001	0.31
$\Delta^2\gamma_9$	-0.41	0.002	0.35	-0.37	0.002	0.36
$\Delta^2\gamma_{10}$	0.10	0.003	0.57	0.06	0.003	0.58

Table 3: Estimates for insurance data. The data sum is $\hat{\tau} = 34,358,090$.

triangle so these estimates are based on fewer observations.

Table 4 shows forecasts for the empty lower triangle from an age-cohort model based on Theorem 5. That is, forecasts of future payments for liabilities of incurred but not fully reported claims. The forecasts are aggregated diagonally and row-wise, thus by period and cohort, respectively. Period aggregates indicate the cash-flow period by period wheres cohort aggregates are the necessary reserves for particular accident years. The total reserve is the aggregate over the full lower triangle. We report point forecasts and the 95% quantile of the forecast distribution

$$\widetilde{Y}_{\mathcal{A}} + [\hat{\tau}\{D_1/(n-q)\}\{\hat{\pi}_{\mathcal{A}} + \hat{s}^2_{\mathcal{A}} + (\hat{\pi}_{\mathcal{A}})^2\}]^{1/2} t_{n-q} \tag{11}$$

where $\hat{\pi}_{\mathcal{A}}$ and $\hat{s}^2_{\mathcal{A}}$ are (7) and (10) evaluated at $\hat{\zeta}^{(2)}$, respectively. The quantile has interpretation as the 95% value at risk.

We also report results based on the bootstrap by England (2002) implemented using

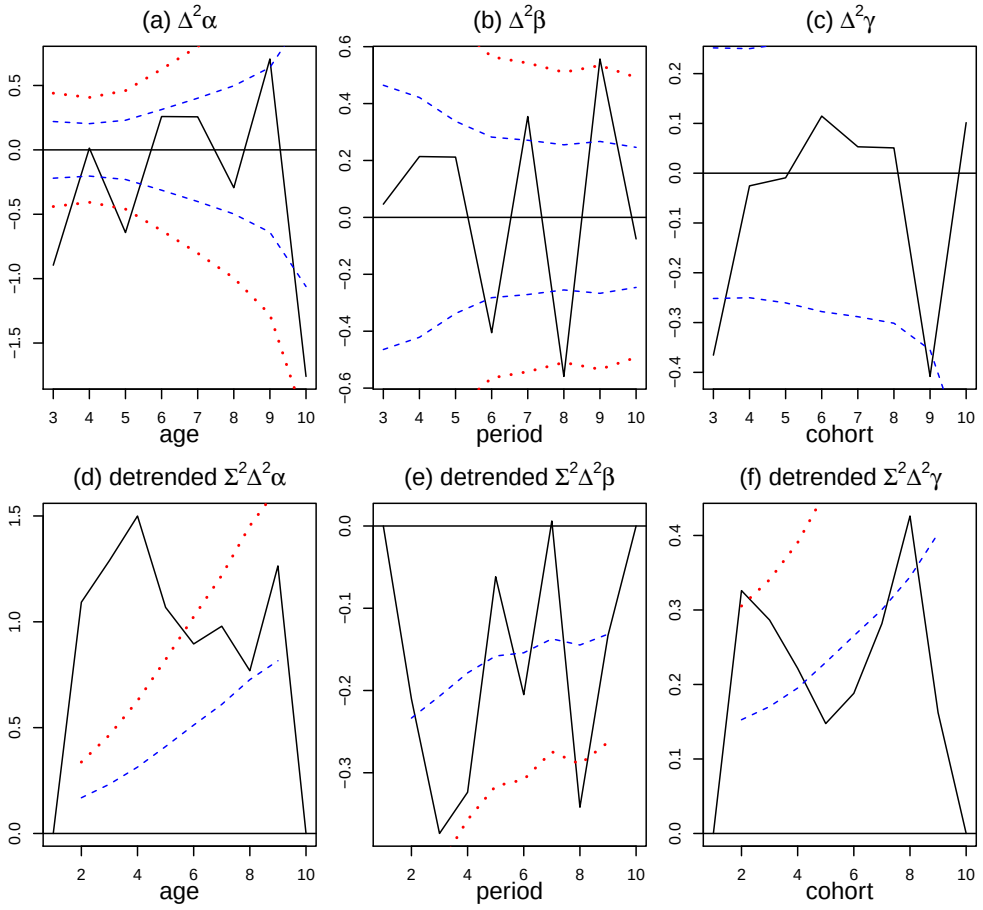

Figure 1: Plot of double differences and detrended parameter estimates. Dashed and dotted lines are one and two standard errors se_t around zero, respectively.

the R package `ChainLadder` (Gesmann et al. 2015). We draw a bootstrap sample of $B = 999$ point forecasts $\widetilde{Y}_{\mathcal{A},b}$ and then add process error variation by drawing $Y^{bs}_{\mathcal{A},b}$ from a gamma distribution centred at $\widetilde{Y}_{\mathcal{A},b}$. The distribution of $-(Y^{bs}_{\mathcal{A},b} - \widetilde{Y}_{\mathcal{A}})$ should then approximate that of $Y_{\mathcal{A}} - \widetilde{Y}_{\mathcal{A}}$ and its 95th quantile added to the point forecast approximates the 95% value at risk. We note that there is no formal theory for the validity of the bootstrap in the present situation. The t forecast is usually larger than the bootstrap, but not always. The two methods are closer for larger values at risk, that is, for earlier periods and later cohorts.

Period	Cash-flow	95% value at risk t	95% value at risk bootstrap	Cohort	Reserve	95% value at risk t	95% value at risk bootstrap
11	523	643	639	2	9	28	22
12	418	532	524	3	47	83	77
13	313	417	419	4	71	115	111
14	213	291	287	5	98	149	144
15	156	222	217	6	142	205	197
16	118	177	170	7	218	301	292
17	74	123	117	8	392	523	513
18	45	86	79	9	428	602	591
19	9	27	19	10	463	786	772
				Total	1868	2330	2337

Table 4: Age-cohort forecasts for insurance data. Results in ten thousands.

	Size under H_{ac}			Power under H_{apc}		
Target	1.00%	5.00%	10.00%	1.00%	5.00%	10.00%
$s = 0.5$	1.24%	5.81%	11.32%	9.78%	26.68%	39.41%
$s = 1$	1.12%	5.41%	10.66%	23.57%	48.92%	63.06%
$s = 2$	1.03%	5.16%	10.31%	58.30%	82.62%	90.58%

Table 5: Simulation performance of F test. Monte Carlo standard error less than 0.05.

7 Simulation study

7.1 Test statistic

We assess the finite sample performance of the asymptotically F distributed specification test F_1 proposed in Theorem 4. We simulate under the age-cohort hypothesis H_{ac} so $\Delta^2\beta_j = 0$ for all j, as well as under H_{apc}, the age-period-cohort hypothesis.

The Y_{ik} are simulated as independent compound Poisson gamma variables so $Y_{ik} = \sum_{\ell=1}^{Z_{ik}} X_l$ where Z_{ik} is Poisson with mean $\exp(\mu_{ik})$ and independent of the independent gamma distributed X_ℓ with scale $\sigma^2 - 1$ and shape $(\sigma^2 - 1)^{-1}$. We choose the data array $\mathcal{I}$ and parameters to match the insurance data, and estimates in Table 2 and Table 3, respectively, except μ_{11} is chosen as $\log(s) + \hat{\mu}_{11}$ for $s = 0{\cdot}5, 1, 2$ so $\tau = s\hat{\tau}$. We draw 10^6 repetitions.

Table 5 shows the simulated rejection frequencies under the age-cohort hypothesis H_{ac} and the age-period-cohort unrestricted model H_{apc}. The size control is good for all values of s. The power is increasing in s. For $s = 1$ we get a 50% power for a 5% test which indicates that one should perhaps be cautious not to reduce the model too far for the insurance data. However, a parsimonious model can be advantageous for forecasting.

		Moments				Quantiles							
s		First		Second		1%		5%		50%		95%		99%	
0.5	true	-2		22		-63		-40		1		30		40	
	boot	-2	*(1)*	22	*(6)*	-61	*(20)*	-40	*(12)*	0	*(1)*	30	*(7)*	40	*(10)*
	t	0	*(2)*	20	*(5)*	-47	*(20)*	-32	*(11)*	0	*(1)*	32	*(8)*	47	*(13)*
1	true	-2		30		-82		-55		1		44		59	
	boot	-2	*(1)*	30	*(6)*	-79	*(20)*	-54	*(12)*	0	*(1)*	44	*(8)*	59	*(11)*
	t	0	*(2)*	28	*(6)*	-66	*(21)*	-46	*(12)*	0	*(1)*	46	*(9)*	66	*(14)*
2	true	-2		42		-110		-74		1		64		87	
	boot	-2	*(1)*	42	*(7)*	-107	*(21)*	-73	*(13)*	0	*(2)*	64	*(10)*	87	*(14)*
	t	0	*(2)*	40	*(7)*	-94	*(22)*	-65	*(14)*	0	*(1)*	65	*(10)*	94	*(16)*

Table 6: Simulation performance of t and bootstrap forecasts. Results in hundred thousands. Shown are averages across simulations and, in parentheses, root mean square errors.

7.2 Forecasting

We first simulate the true distribution of the forecast error $Y_{\mathcal{A}} - \widetilde{Y}_{\mathcal{A}}$. Then, we evaluate the quality of the t forecast in Theorem 5 and the bootstrap method by England (2002). We consider forecast errors for the sum of all entries in the lower triangle so $\mathcal{A} = \mathcal{J}$, known in insurance as the chain ladder reserve. Results are reported in Table 6.

We consider the data generating process described in §7.1 for the age-cohort model and simulate for $s = 0.5, 1, 2$. Due to the age-cohort structure, this defines the distribution in both the data array $\mathcal{I}$ and the forecast array $\mathcal{J}$.

We approximate the first two moments and α quantiles of $Y_{\mathcal{A}} - \widetilde{Y}_{\mathcal{A}}$ by Monte Carlo with 10^6 draws. The moments have interpretation as bias and prediction error of $\widetilde{Y}_{\mathcal{A}}$. The distribution is left skew since the distribution of the point forecasts is more right skew than that of the realisations.

For the forecast approximations we draw $R = 5,000$ data triangles $\mathcal{I}_r$. We report averages across r and, in parentheses, root mean square errors. For the t forecast, for every $\mathcal{I}_r$, we compute approximations to moments and quantiles based on (11) minus the point forecast. For the bootstrap, we proceed as described in §6 and draw for every $\mathcal{I}_r$ a bootstrap sample of $B = 999$ realisations of $-(Y^{bs}_{\mathcal{A},r,b} - \widetilde{Y}_{\mathcal{A},r})$. For each r, moments and quantiles are computed as sample averages across b and αB order statistics.

The bootstrap clearly performs better on average. Root mean square errors of the t forecast are mostly close to those of the bootstrap and sometimes smaller, indicating that the bootstrap produces more outliers than the t.

8 Discussion

The presented sampling scheme provides a framework for developing specification tests for the over-dispersed Poisson model. We are currently working on a test for the assumption of common over-dispersion across the full sample. Such a test might also

be a starting point for model selection between over-dispersed Poisson and log-normal model. Rather than a fixed variance to mean ratio, the log-normal model implies a fixed standard deviation to mean ratio.

In §5 we referred to forecasting scenarios that require parameter extrapolation. If ad-hoc identification is used in this case, care is needed to prevent an impact of the ad-hoc decision on the forecasts (Kuang et al. 2008*a*). Kuang et al. (2011) and Martínez Miranda et al. (2015) discuss forecasting with period-effect extrapolation in a log-normal model and a Poisson model, respectively. Extrapolation would add additional terms to the forecast error decomposition (9). A formal analysis of this would be of interest.

This paper considers a model for responses only. In other scenarios there is additional information available about exposure. It would be interesting to derive a theory for such a setting.

Acknowledgement

The authors thank two anonymous referees for numerous helpful comments. Discussions with D. R. Cox and S. Johansen are gratefully acknowledged. The first author was supported by the Economic and Social Research Council, grant number ES/J500112/1. The second author was supported by the programme for Economic Modelling, Oxford and European Research Council, grant AdG 694262.

Appendix

Proof of Theorem 1

From Sato (1999, pages 37–39 and Theorem 21.5) we have a general form for the the logarithm of the characteristic function of Y_ℓ. Since Y_ℓ is non-negative infinitely divisible with $E(|Y_\ell|) < \infty$, then

$$\phi_\ell(t) = \log[E\{\exp(itY_\ell)\}] = i\gamma_\ell t + \int_0^\infty \{\exp(ity) - 1 - ity\}\nu_\ell(dy) \tag{12}$$

with Lévy measure ν_ℓ. Since $E(|Y_\ell^3|) < \infty$, we can find the first three cumulants by differentiating $\phi_\ell(t)$ (Lukacs 1960, pages 33–34) and get

$$\gamma_\ell = E(Y_\ell), \quad \int_0^\infty y^2\nu_\ell(dy) = \mathrm{var}(Y_\ell), \quad \int_0^\infty y^3\nu_\ell(dy) = E[\{Y_\ell - E(Y_\ell)\}^3] \tag{13}$$

From Billingsley (1995, page 343) we get

$$\exp(ity) = 1 + ity - \frac{(ty)^2}{2} + r(y,t), \quad |r(y,t)| \le \min\left\{\frac{|yt|^3}{6}, (yt)^2\right\}. \tag{14}$$

The remainder $r(y,t)$ is ν_ℓ-integrable for any $t \in R$ since it is dominated by $(yt)^2$ and y^2 is ν_ℓ-integrable due to (13). Inserting (13) and (14) in (12),

$$\phi_\ell(t) = itE(Y_\ell) - \frac{t^2}{2}\text{var}(Y_\ell) + \int_0^\infty r(y,t)\nu_\ell(dy). \tag{15}$$

Let $U_\ell = \{Y_\ell - E(Y_\ell)\}/\sqrt{\text{var}(Y_\ell)}$ with log characteristic function

$$\rho_\ell(s) = \log[E\{\exp(isU_\ell)\}] = \phi_\ell\left(\frac{s}{\sqrt{\text{var}(Y_\ell)}}\right) - \frac{isE(Y_\ell)}{\sqrt{\text{var}(Y_\ell)}}.$$

Inserting the expression (15) gives

$$\rho_\ell(s) = -\frac{s^2}{2} + \int_0^\infty r\left\{y, \frac{s}{\sqrt{\text{var}(Y_\ell)}}\right\}\nu_\ell(dy). \tag{16}$$

The standard normal distribution has log characteristic function $-s^2/2$ (Lukacs 1960, page 26). Thus, the distribution of U_ℓ converges weakly to a standard normal distribution if and only if its characteristic function converges point-wise to the standard normal characteristic function (Lukacs 1960, Theorem 3.6.1). Hence, we want to show that for each $s \in R$ the second term in (16) vanishes as $skew(Y_\ell) \to 0$. Denoting the integrand by r for shortness, we find $|\int_0^\infty r\nu_\ell(dy)| \leq \int_0^\infty |r|\nu_\ell(dy)$. With (14),

$$\int_0^\infty |r|\nu_\ell(dy) \leq \int_0^\infty \min\left\{\frac{|ys|^3}{6\text{var}(Y_\ell)^{3/2}}, \frac{(ys)^2}{\text{var}(Y_\ell)}\right\}\nu_\ell(dy) \leq \min\left\{\frac{|s|^3}{6}skew(Y_\ell), s^2\right\}$$

where the last inequality follows by the non-negativity of the integrand and (13). The minimum is dominated by either of its arguments. It therefore vanishes as $skew(Y_\ell) \to 0$.

Proof of Theorem 2

With $\{Y_\ell - E(Y_\ell)\}/\sqrt{\text{var}(Y_\ell)} \to N(0,1)$ and $E(Y_\ell)/\sqrt{\text{var}(Y_\ell)} \to \infty$ the results follows since

$$\frac{Y_\ell}{E(Y_\ell)} = 1 + \frac{Y_\ell - E(Y_\ell)}{\sqrt{\text{var}(Y_\ell)}}\left\{\frac{E(Y_\ell)}{\sqrt{\text{var}(Y_\ell)}}\right\}^{-1}.$$

Proof of Lemma 1

Consider the mixed parametrization of the Poisson likelihood discussed in §3.5 for a saturated model in which μ_{ik} is unrestricted. This nests the age-period-cohort model and its sub models. The saturated model has mixed parametrization $\psi_S = \{\tau, (\theta)^T\}^T$ where the vector θ contains $\theta_{ik} = \mu_{ik} - \mu_{\ell m}$ for $(i,k) \in \mathcal{I} \setminus (\ell, m)$. Define the design vectors $s_{ik} \in R^{n-1}$ for $(i,k) \in \mathcal{I}$ so $s_{\ell m} = 0$ and s_{ik} for $(i,k) \neq (\ell, m)$ is a unit vector so $\theta_{ik} = s_{ik}^T\theta$. The minimal sufficient statistic for ψ_S in a saturated Poisson model is $(Y_{..}, T_S^{(2)})$ where $T_S^{(2)} = \sum_{ik\in\mathcal{I}} s_{ik}^T Y_{ik}$. We have $\mu_{ik} = \log(\tau) + \log\{\pi_{ik}(\theta)\}$. Recalling that

$\hat{\tau} = Y_{\cdot\cdot}$ and organizing Y_{ik} and μ_{ik} for $(i,k) \in \mathcal{I}$ in vectors, Y and μ, say, we find

$$M_\tau^{-1}\frac{\partial \ell_Y}{\partial \psi_S} = \begin{Bmatrix} \hat{\tau} - \tau \\ \hat{\tau}(T_S^{(2)}/\hat{\tau} - \sum_{ik\in\mathcal{I}} s_{ik}^T \pi_{ik}) \end{Bmatrix} = M_\tau^{-1}\frac{\partial \mu^T}{\partial \psi_S}\frac{\partial \ell_Y}{\partial \mu} = M_\tau^{-1}\frac{\partial \mu^T}{\partial \psi_S}\{Y - E(Y)\}. \tag{17}$$

Organize $\{\pi_{ik} : (i,k) \in \mathcal{I}\}$ as a vector, π, say. With Johansen (1979, Proof of Lemma 7.2) we verify that

$$M_\tau^{-1}\frac{\partial \mu^T}{\partial \psi_S} diagonal(\pi) \frac{\partial \mu}{\partial \psi_S^T} M_\tau^{-1} = -\tau^{-1}\left.\frac{\partial^2 \ell_Y}{\partial \psi_S \partial \psi_S^T}\right|_{Y=E(Y)} = \begin{pmatrix} 1 & 0 \\ 0 & \bar{\imath}_\theta \end{pmatrix} = \tilde{\imath}_{\psi_S}. \tag{18}$$

With independent Y_{ik} Theorem 1 extends to $\tau^{-1/2}\{Y - E(Y)\} \to N\{0, \sigma^2 diagonal(\pi)\}$. This implies that $\hat{\tau}/\tau \to 1$ in probability by Theorem 2. Then, by Slutsky's theorem, $\hat{\tau}^{-1/2}\{Y - E(Y)\}$ and $\tau^{-1/2}\{Y - E(Y)\}$ have the same asymptotic distribution. Thus,

$$\hat{\tau}^{-1/2} M_\tau^{-1}\frac{\partial \ell_Y}{\partial \psi_S} = M_\tau^{-1}\frac{\partial \mu^T}{\partial \psi_S}\hat{\tau}^{-1/2}\{Y - E(Y)\} \to N(0, \sigma^2 \tilde{\imath}_{\psi_S}). \tag{19}$$

In particular, the asymptotic distribution of the two components of the normalized sufficient statistics $\hat{\tau}^{-1/2}(\hat{\tau} - \tau)$ and $\hat{\tau}^{1/2}(T_S^{(2)}/\hat{\tau} - \sum_{ik\in\mathcal{I}} s_{ik}^{(2)} \pi_{ik})$ are asymptotically independent. Quasi likelihood estimators for $\xi^{(2)}$ and its restrictions, as well as deviances and log likelihood ratio statistics are functions of the second component only, thus asymptotically independent of $\hat{\tau}^{-1/2}(\hat{\tau} - \tau)$. We note that $(\hat{\tau}/\sigma^2)^{1/2}(T_S^{(2)}/\hat{\tau} - \sum_{ik\in\mathcal{I}} s_{ik}^{(2)} \pi_{ik})$ has the same asymptotic distribution as in a multinomial model conditional on $Y_{\cdot\cdot}$ and that we are interested in the same data transformations as in that model. Thus, for the asymptotic argument, we can exploit results from exponential family theory. The asymptotic distribution of $\hat{\tau}^{1/2}(\hat{\xi}^{(2)} - \xi^{(2)})$ follows from Johansen (1979, Theorem 7.3). With Johansen (1979, Theorems 7.6, 7.7, 7.8), the asymptotic distributions and independence of D_{apc}, $LR_{1,apc}$ and $LR_{2,1}$ follow, as does asymptotic independence of $LR_{1,apc}$ and $\hat{\tau}^{1/2}(\hat{\xi}^{(2)} - \xi^{(2)})$.

Proof of Theorem 3

From Lemma 1, $\hat{\tau}^{1/2} v^T(\hat{\xi}^{(2)} - \xi^{(2)}) \to N\{0, \sigma^2 v^T(\bar{\imath}_{\xi^{(2)}})^{-1} v\}$, asymptotically independent of $D_{apc} \to \sigma^2 \chi^2_{n-p}$. The studentized estimator is then t_{n-p} distributed.

Proof of Theorem 4

If $W_1^2 = \chi^2_{df_1}$, $W_2^2 = \chi^2_{df_2}$ and $W_3^2 = \chi^2_{df_3}$ are mutually independent then $V_1^2 = W_1^2/(W_1^2 + W_2^2) = beta(df_1/2, df_2/2)$ and $V_2^2 = (W_1^2 + W_2^2)/(W_1^2 + W_2^2 + W_3^2) = beta\{(df_1 + df_2)/2, df_3/2\}$ are independent (Johnson et al. 1993, page 212). Hence, the F distributed $\{(1 - V_1^2)/df_2\}/\{V_1^2/df_1\}$ and $\{(1 - V_2^2)/df_3\}/\{V_2^2/(df_1 + df_2)\}$ are independent. By Lemma 1, D_{apc}, $LR_{1,apc}$ and $LR_{2,1}$ are asymptotically mutually independent $\sigma^2\chi^2$ distributed with $n - p$, $p - q$ and $q - r$ degrees of freedom. By taking ratios as in $\mathrm{F}_{1,apc}$ and $\mathrm{F}_{2,1}$, σ^2 cancels out in the asymptotic distribution. Setting $W_1^2 = D_{apc}$, $W_2^2 = LR_{1,apc}$ and $W_3^2 = LR_{2,1}$, the asymptotic F distribution and independence follows.

Proof of Theorem 5

Since $\hat{\tau}/\tau \to 1$ in probability as noted in the proof of Lemma 1, $\hat{\tau}^{-1/2}\{Y_{\mathcal{A}} - E(Y_{\mathcal{A}})\}$ has the same asymptotic distribution as $\tau^{-1/2}\{Y_{\mathcal{A}} - E(Y_{\mathcal{A}})\} \to N(0, \sigma^2\pi_{\mathcal{A}})$, using Theorem 1. The latter is a function of the future realisations in $\mathcal{J}$ and thus independent of both estimation error components which are functions of the data in $\mathcal{I}$. The distribution of the two estimation error components and D_1 are asymptotically independent by Lemma 1. Since D_1 is a function of the data, it is also asymptotically independent of the process error component. The studentized forecast error is then t_{n-q} distributed.

References

Agresti, A. (2013), *Categorical Data Analysis*, 3rd edn, John Wiley & Sons, Hoboken, NJ.

Alai, D. H. & Sherris, M. (2014), 'Rethinking age-period-cohort mortality trend models', *Scandinavian Actuarial Journal* **2014**, 208–227.

Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, John Wiley & Sons, Chichester.

Barnett, G. & Zehnwirth, B. (2000), 'Best estimates for reserves', *Proceedings of the Casualty Actuarial Society* **87**, 245–321.

Beard, R. E., Pentikäinen, T. & Pesonen, E. (1984), *Risk Theory*, 3rd edn, Chapman and Hall, London.

Billingsley, P. (1995), *Probability and Measure*, 3rd edn, John Wiley & Sons, Chichester.

Bliss, C. I. & Fisher, R. A. (1953), 'Fitting the negative binomial distribution to biological data', *Biometrics* **9**, 176–200.

Bondesson, L. (2015), 'A class of probability distributions that is closed with respect to addition as well as multiplication of independent random variables', *Journal of Theoretical Probability* **28**, 1063–1081.

Casualty Actuarial Society (2016), 'Casualty actuarial society — loss reserves', www.casact.org/research/index.cfm?fa=loss_reserves_data. Accessed: 2016-10-28.

Cox, D. R. (1983), 'Some remarks on overdispersion.', *Biometrika* **70**, 269–274.

Davison, A. C. & Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge.

England, P. D. (2002), 'Addendum to 'analytic and bootstrap estimates of prediction errors in claims reserving'', *Insurance: Mathematics and Economics* **31**, 461–466.

England, P. D. & Verrall, R. J. (2002), 'Stochastic claims reserving in general insurance', *British Actuarial Journal* **8**, 443–518.

Fisher, R. A. (1922), 'On the interpretation of χ^2 from contingency tables, and the calculation of p', *Journal of the Royal Statistical Society* **85**, 87–94.

Fu, W. (2016), 'Constrained estimators and consistency of a regression model on a lexis diagram', *Journal of the American Statistical Association* **111**, 180–199.

Gesmann, M., Murphy, D., Zhang, Y., Carrato, A., Crupi, G., Wuthrich, M. & Concina, F. (2015), 'Chainladder: Statistical methods and models for claims reserving in general insurance'.
URL: *CRAN.R-project.org/package=ChainLadder*

Johansen, S. (1979), *Introduction to the Theory of Regular Exponential Families*, Institute of Mathematical Statistics, University of Copenhagen, Copenhagen.

Johnson, N. L., Kemp, A. W. & Kotz, S. (2005), *Univariate discrete distributions*, 3rd edn, Wiley, Hoboken, NJ.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1993), *Continuous univariate distributions*, Vol. 2, 2nd edn, John Wiley & Sons, New York.

Jørgensen, B. (1986), 'Some properties of exponential dispersion models', **13**, 187–197.

Jørgensen, B. (1987), 'Exponential dispersion models (with discussion)', **49**, 127–162.

Keiding, N. (1990), 'Statistical inference in the lexis diagram', *Philosophical Transactions of the Royal Society of London, Series A* **332**, 487–509.

Kuang, D., Nielsen, B. & Nielsen, J. P. (2008*a*), 'Forecasting with the age-period-cohort model and the extended chain-ladder model', *Biometrika* **95**, 987–991.

Kuang, D., Nielsen, B. & Nielsen, J. P. (2008*b*), 'Identification of the age-period-cohort model and the extended chain-ladder model', *Biometrika* **95**, 979–986.

Kuang, D., Nielsen, B. & Nielsen, J. P. (2009), 'Chain-ladder as maximum likelihood revisited', *Annals of Actuarial Science* **4**, 105–121.

Kuang, D., Nielsen, B. & Nielsen, J. P. (2011), 'Forecasting in an extended chain-ladder-type model', *Journal of Risk and Insurance* **78**, 345–359.

Lee, Y. K., Mamman, E., Nielsen, J. P. & Park, B. U. (2015), 'Asymptotics for in-sample density forecasting', *Annals of Statistics* **43**, 620–651.

Lukacs, E., ed. (1960), *Characteristic functions*, Griffin, London.

Martínez Miranda, M. D., Nielsen, B. & Nielsen, J. P. (2015), 'Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality', *Journal of the Royal Statistical Society, Series A* **178**, 29–55.

Nielsen, B. (2015), 'apc: An r package for age-period-cohort analysis', *The R Journal* **7**, 52–64.

Nielsen, B. & Nielsen, J. P. (2014), 'Identification and forecasting in mortality models.', *The Scientific World Journal* **2014**, 347043.

Pinheiro, P. J. R., Andrade e Silva, J. M. & De Lourdes Centeno, M. (2003), 'Bootstrap methodology in claim reserving', *Journal of Risk and Insurance* **70**, 701–714.

R Core Team (2016), 'R: A language and environment for statistical computing'. **URL:** *www.R-project.org*

Sato, K., ed. (1999), *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.

Smith, T. R. & Wakefield, J. (2016), 'A review and comparison of age-period-cohort models for cancer incidence', *Statistical Science* **31**, 591–610.

Taylor, G. C. & Ashe, F. R. (1983), 'Second moments of estimates of outstanding claims', *Journal of Econometrics* **23**, 37–61.

Thorin, O. (1977), 'On the infinite divisibility of the lognormal distribution', *Scandinavian Actuarial Journal* **1977**, 121–148.

Venables, W. N. & Ripley, B. D., eds (2002), *Modern Applied Statistics with S*, 4th edn, Spring-Verlag, New York.

Wedderburn, R. W. M. (1974), 'Quasi-likelihood functions, generalized linear models, and the gauss-newton method', *Biometrika* **61**, 439–447.