Age-Period-Cohort Models

Jonas Harnau

Oriel College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Economics

Hilary Term 2018

To Sam and Brigitte

In memory of Mario

Acknowledgements

First, I thank my supervisor Bent Nielsen who inspired me to pursue this research, provided timely counsel and supported me in any way possible, great and small. Without his guidance, this journey would have been impossible. To him, I promise never to forget the question "what is the model?" Next, I thank my family and friends from Oxford and around the world for their moral support. Without them, I would never have made it this far. I am particularly grateful to my fellow DPhil students for the time spent together, academic and otherwise, that immeasurably enriched my time at Oxford and greatly benefited my work. I am also thankful for discussions with Søren Johansen and, indirectly through my supervisor, David R. Cox that benefited the chapter on "Over-Dispersed Age-Period-Cohort Models". This chapter and the chapter on "Misspecification Tests for Chain-Ladder Models" also benefited from helpful comments by anonymous referees during the peer review process. Last but not least, I am indebted to the Economic and Social Research Council (ES/J500112/1) and the European Research Council (AdG 694262) for their financial support.

Abstract

While over-dispersed Poisson age-period-cohort and extended chain-ladder models are used in a number of fields, so far no rigorous statistical theory has been available. We consider models for aggregate data organized in a two-way table with age and cohort as indices, but without measures of exposure. In these models, used for example in actuarial science, demography, economics, epidemiology and sociology, the number of parameters grows with the number of observations. Thus, standard asymptotic theory is invalid. In Chapter 2, we propose a repetitive structure that keeps the dimension of the table fixed while increasing the latent exposure. We pair this with the assumptions of infinitely divisible distributions which include a variety of compound Poisson models and Poisson mixture models. We then show that Poisson quasi-likelihood estimation results in asymptotic t parameter distributions, F inference, and t forecast distributions. In Chapter 3, we build on the asymptotic framework from Chapter 2 and develop tests for model specification. The over-dispersed Poisson model assumes that the over-dispersion is common across the data. A further assumption is that effects do not have breaks, for example age effects do not vary over cohorts. A log-normal age-period-cohort model makes similar assumptions. We show that these assumptions can easily be tested and that similar tests can be used in both models. In Chapter 4, we develop a non-nested test that allows one to evaluate whether the over-dispersed Poisson or log-normal model is the better choice for the data. While the over-dispersed Poisson model imposes a fixed variance to mean ratio, the log-normal models assumes the same for the standard deviation to mean ratio. We leverage this insight to propose a test that has high power to distinguish between the two models. Again, the theory is asymptotic but does not build on a large size of the array and instead makes use of information accumulating within the cells.

Contents

1. Introduction

In this thesis, we develop statistical theory for age-period-cohort models for data that are over-dispersed relative to a Poisson model. An important area of application for these results is in actuarial science, particularly in non-life insurance claim reserving. In this literature, age-period-cohort models are more commonly referred to as extended chain-ladder models. The theory put forward in this thesis allows for inference, closed form distribution forecasting, misspecification testing and non-nested testing between rival models.

In this introduction, we give a brief overview of age-period-cohort modeling with a particular focus on applications to reserving in non-life insurance. First, we provide a motivating example that illustrates the claim reserving problem. Next, we more precisely define age-period-cohort data and discuss different types of such data. Then, we take a closer look at the linear age-period-cohort predictor, also known as extended chain-ladder predictor. With this in place, we briefly explain the contributions of this thesis. Finally, we return to the motivating example and illustrate the application of our contributions.

Before we move on, we remark that while age-period-cohort and extended chain-ladder terminologies may seem different, they are in fact interchangeable. The extended chain-ladder terminology is used in actuarial science whereas the age-period-cohort terminology is the standard in most other fields of applications such as demography, economics, epidemiology or sociology. What is referred to as "accident year" or sometimes "underwriting year" in actuarial science is called "cohort" in the remaining strands of literature. Similarly, "development year" translates to "age" and "calendar year" to "period". Depending on the particular application, we use both terminologies somewhat flexibly in a way we hope is not confusing to the reader.

1.1 Motivating example

In this simplified example, which is loosely based on Verrall (1994), we imagine a non-life insurer providing car insurance and focus on the insurer's expenses. The insurer has to pay out claims for car accidents to the insured. However, claims are not necessarily paid out

Dev. Acc.	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 1.1: Example of an insurance run-off triangle. Accident years (Acc.) are in the rows with the most recent year corresponding to accident year ten. Development years (Dev.) in the columns. The cells contain aggregate incremental paid amounts. Data from Taylor & Ashe (1983, Appendix).

in the year of the accident. Instead, there can be substantial delays between the accident year and the payment. The delay is often referred to as the development year. To ensure liquidity and to satisfy the matching principle in accounting which demands attribution of losses to the accident year they originate from (Daykin et al. 1994, §1.2), the insurer needs to forecast the delayed payments. That is, forecasts are needed for claims relating to accidents that occurred in the past but have not yet been paid out.

Forecasts are commonly based on data for the sum of payments over individual accidents within accident and development year, called aggregate incremental payments. In our example, the aggregation would be over a portfolio of car insurance policies. Consequently, there is a single data point for each observed accident and development year combination. This gives rise to a two-way table commonly referred to as a run-off triangle. An example are the data in Table 1.1. The triangular structure may initially seem peculiar but becomes intuitive once we introduce a third time scale: calendar years. The main insight is that together accident and development year determine the calendar year. Counting with an off-set so all three time-scales start at unity,

$$\text{Accident Year} + \text{Development Year} = \text{Calendar Year} + 1.$$

Thus, claims relating to accidents that occurred in year 1 and were paid out in development year 1, the year of the accident, were paid out in calendar year 1. Therefore, cells relating

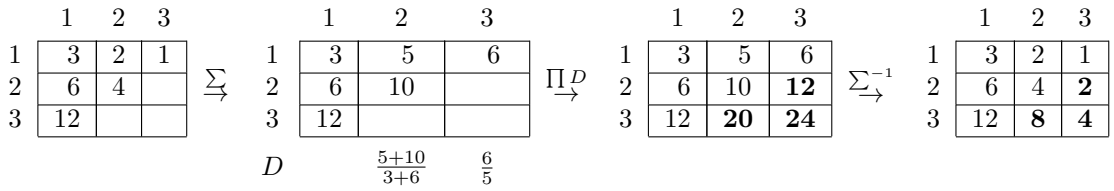


Table 1.2: Illustration of the chain-ladder technique. The first step forms cumulative sums over the rows and computes the development factors D. The second step forms point forecasts for cumulative sums. The third step reverses the cumulative sums. Forecasts are shown in bold.

to the same calendar year are on the same diagonal running from the top right to the bottom left, sometimes referred to as antidiagonal. Calendar years are increasing from the top left to the bottom right. Thus, the final diagonal in Table 1.1 contains payments for calendar year 10, the most recent calendar year.

The task for the actuary is to generate forecasts for cells in the empty lower triangle that contains cells relating to future calendar years. To ensure liquidity, the insurer may be particularly interested in forecasts for the cash-flow. This corresponds to the sum over cells on the diagonals. To satisfy the matching principle in accounting, the insurer requires forecast for payments within a given accident year known as the reserve. This is the sum over cells in a row. Another measure of interest may be the total reserve or total cash-flow, the sum over all cells in the lower triangle.

A widely used method to generate point forecasts is the chain-ladder technique. While this method has an entirely deterministic origin, it is intuitively appealing. The idea is to extrapolate based on observed growth rates of the cumulative sums of payments over development years.

We illustrate the chain-ladder technique in a stylized example in Table 1.2. In the example, we start off with at three-by-three triangle of aggregate incremental payments. In this triangle, we form cumulative sums over the rows which yields the second shown triangle; such entries are often referred to as aggregate cumulative payments. Based on these, we compute the growth factors from one column to the next, also known as development factors. For example, we compute the development factor relating to the growth from the first to the second development year as follows. First, we find all accident years (rows) for

which we have data for the second development year (column). In this case, these are the first two accident years. Then, considering only data for these accident years, we take the ratio of the sum over cells in the second development year, $5 + 10$, relative to sum in the first development year, $3 + 6$. To obtain the forecasts shown in bold in the third table, we extrapolate the aggregate cumulative payments using the development factors D. To do this, we find the last available cumulative aggregate payment in a row and multiply it by the development factor for the next column. For example, the forecast for accident year (row) 3 and development year (column) 2 is computed as $12 \cdot (5 + 10)/(3 + 6) = 20$. We iterate this if needed; for instance the entry in accident and development year 3 is computed by multiplying the already extrapolated entry 20 by the development factor 6/5. In this way, we fill in all cells in the array. Finally, we recover the incremental aggregate payments in the fourth table by reversing the cumulative sums. We do this by taking differences between columns, except for the first column which remains unchanged. We would argue that the forecasts, shown in bold, are quite intuitive for this artificial data.

While the chain-ladder technique has a deterministic origin, a stochastic specification is needed to go beyond point forecasts. For example, the insurer may be interested in the value at risk or the forecast standard deviation. The process is somewhat reversed from the usual statistical procedure. Rather than starting with a statistical model and then finding a reasonable estimator, we start with an estimator and look for a model in which the estimator is reasonable.

One model in which chain-ladder point forecasts are reasonable is an independent Poisson model that allows individual effects for accident and development years. For this model, Kremer (1985) showed that chain-ladder point forecasts are replicated by maximum likelihood estimation. Denoting assumed independent aggregate incremental payments for accident year i and development year j as Y_{ij} and chain-ladder pointforecasts by $\widetilde{Y}_{ij}^{CL}$,

$$Y_{ij} \stackrel{D}{=} \text{Poisson}\{\exp(\mu_{ij})\} \quad \text{where} \quad \mu_{ij} = \alpha_i + \beta_j + \delta \quad \Rightarrow \quad \widetilde{Y}_{ij}^{CL} = \exp(\hat{\mu}_{ij}^{mle})$$

where α is the accident year effect and β the development year effect. We remark that there is an identification issue for the effects which we will discuss below in §1.2. Unfortunately, the Poisson model is not well suited for the insurance data. One issue is that Poisson

random variables are discrete while we would usually think of payments as continuous. Perhaps a bigger challenge is that a Poisson model imposes that variance and mean are identical which often does not fit the data well.

To address the often poor fit of the Poisson model, a common solution is to model the over-dispersion and abandon the Poisson assumption while retaining independence and mean structure. Such over-dispersed Poisson models allow variance and mean to be proportional, rather than equal, across the array. In this model, chain-ladder point forecasts are replicated by Poisson quasi-likelihood estimation:

$$E(Y_{ij}) = \exp(\mu_{ij}) \quad \text{and} \quad \frac{\text{var}(Y_{ij})}{E(Y_{ij})} = \sigma^2 \quad \Rightarrow \quad \widetilde{Y}_{ij}^{CL} = \exp(\hat{\mu}_{ij}^{ql}).$$

See Wedderburn (1974) for quasi-likelihood. An appealing motivation for such an over-dispersed Poisson model in an insurance context is a compound Poisson model. The idea is that the aggregate incremental payments Y_{ij} are a sum over a Poisson number of claims N_{ij} of independent and identically distributed individual claim amounts X_ℓ that are independent of N_{ij}. Then the aggregate incremental payments $Y_{ij} = \sum_{\ell=1}^{N_{ij}} X_\ell$ satisfy the over-dispersed Poisson assumptions if the distribution of the individual payments is identical across accident and development years. However, it is not clear what the distribution of X_ℓ should be. Further, there are other distributions that are consistent with the over-dispersed Poisson model. Thus, the flexibility of the model comes at the cost that it does not pin down a specific distribution for Y_{ij}.

While there is an appealing motivation for the over-dispersed Poisson model and a reasonable motivation for the chain-ladder technique, it is not clear how to form distribution forecasts. A popular approach is bootstrapping as suggested by England & Verrall (1999) and England (2002). For example, Table 1.3 shows the 95% value at risk, the 95th percentile of the bootstrap forecast distribution, for the data in Table 1.1. However, so far there is no theory available showing the validity of the bootstrap in this setting. Further, in practice the results sometimes seem unsatisfactory. An inherent problem that a potential theory has to tackle is that the number of parameters grows with the number of observations. For example, by adding cells relating to new accident or development years, new accident or development effects are needed. This incidental parameter problem means that a standard

Calendar Year	Cash-flow	95% value at risk	Accident Year	Reserve	95% value at risk
11	523	639	2	9	22
12	418	524	3	47	77
13	313	419	4	71	111
14	213	287	5	98	144
15	156	217	6	142	197
16	118	170	7	218	292
17	74	117	8	392	513
18	45	79	9	428	591
19	9	19	10	463	772
			Total	1868	2337

Table 1.3: Bootstrap chain-ladder forecasts for the insurance data in Table 1.1. Results in ten thousands. Bootstrap method by England & Verrall (1999) and England (2002) implemented with help of the R (R Core Team 2016) package `ChainLadder` (Gesmann et al. 2015).

asymptotic theory for quasi-likelihood estimators based on a large number of cells is invalid. Besides that, run-off triangles can often hardly be considered "large" if considered relative to the number of parameters.

Given the difficulty of over-dispersed Poisson models, an apparent solution would be to abandon the chain-ladder altogether, for example in favor of a log-normal model. Kremer (1982) suggests such a model which does not replicate the chain-ladder point forecasts but retains the structure of the linear predictor. In this model, the independent aggregate incremental payments are log-normal so $\log(Y_{ij}) \overset{D}{=} N(\alpha_i+\beta_j+\delta, \omega^2)$. While the distribution of maximum likelihood estimators in this model is well known, distribution forecasting on the original scale is difficult. One problem is that we are usually interested in forecasts for cell sums such as the cash-flow or reserve. Yet, the log-normal model is not closed under convolution such that the sum of log-normally distributed variables is not log-normal.

Overall, there is still a range of unsolved problems when it comes to reserving in non-life insurance some of which we address in this thesis. First, we provide a rigorous statistical foundation for over-dispersed Poisson models that allows for chain-ladder distribution forecasts as well as inference. Second, building on this theory, we propose misspecification tests that allow us to test for violations of crucial assumptions of the over-dispersed Poisson model. We find that related tests can be used to evaluate assumptions of the log-normal model. Third, we develop a test that directly pits log-normal and over-dispersed Poisson

models against each other. In §1.4 below, we explain the contributions in a little more detail. In §1.5, we show how these results can be applied to the data in Table 1.1.

1.2 Data

Insurance run-off triangles are just one case of age-period-cohort data that can come in many different shapes and forms. We discuss their defining feature and introduce generalized trapezoids as a structure that captures the most commonly encountered shapes. We also briefly consider distinctions between aggregate and individual level data and between data with and without covariates. In this thesis, we exclusively consider aggregate data without covariates such as the run-off triangle in Table 1.1.

The defining feature of age-period-cohort data is that information for at least two of the three time-scales age (development year), period (calendar year), and cohort (accident year) is available. Given knowledge of any two time-scales we can then infer the third. For example, the run-off triangle in Table 1.1 comes with information for accident year (cohort) and development year (age) while the calendar year (period) is implicit. We discuss some restrictions for the link between the three time-scales at the end of this section.

Kuang et al. (2008*b*) define an index set in age-cohort space that accommodates rectangular arrays in age-cohort, age-period, and cohort-period space as well as run-off triangles. They refer to this index set as a generalized trapezoid. The advantage of the age-cohort space is that all three time-scales increase in the same direction. For example, in the run-off triangle in Table 1.1 all time-scales increase from the top left. For age (development year) i, cohort (accident year) k and period (calendar year) $j = i + k - 1$, the index set is defined as

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

where I, K and J are the number of ages, cohorts and periods in the sample, respectively, and $L + 1$ is the index of the oldest available period.

We can further distinguish between aggregate and individual level age-period-cohort data. For aggregate data, we have only a single outcome, sometimes called response, for each cell in the array, as is the case for the insurance data in Table 1.1. Individual level

data arises if we observe several outcomes for each cell, for example, if we observed not aggregate but individual payments for the insurance data.

Another distinction is between age-period-cohort data with and without covariates. A special case of a covariate is the exposure. Roughly, a covariate is referred to as exposure if the outcome is intuitively proportional to it. For example, in demography death counts are intuitively proportional to the population size. Consequently, the population size is also called exposure.

Mortality tables are an example of aggregate data that form rectangular arrays in age-period space. These tables contain mortality data for certain age groups collected over periods. Mortality tables are common in demography and life insurance. There, usually both death counts and exposure are available, for example from the Human Mortality Database (2018). Alai & Sherris (2014) discuss age-period-cohort models in this context. Another field that uses mortality data is epidemiology. In this setting, the appropriate exposure is not always available. An example are mesothelioma deaths as modeled without exposure by Martínez Miranda et al. (2015), or with a synthetically constructed measure of exposure by Hodgson et al. (2005) and Tan et al. (2010).

Panel and repeated cross-sectional data are usually rectangular cohort-period arrays of individual level data with covariates. Cross-sectional data are a special case of rectangular cohort-period arrays with a single period. Heckman & Vytlacil (2001) explicitly work with panel data in an age-period-cohort setting to investigate returns to schooling. Ejrnaes & Hochguertel (2013) consider an application to moral hazard in unemployment insurance choice. Deaton & Paxson (1994) investigate saving rates in Taiwan based on repeated cross-sectional data.

Time-series data are a special case of rectangular arrays in age-cohort space with a single cohort. Multi-cohort age-cohort arrays seem harder to come by. This may be because the required sampling seems somewhat odd, the number of sampled age groups would need to increase over periods. We can still come up with settings where such data would be of interest. For example, we may want to study change of societal norms or political leanings within age groups over birth cohorts. We could collect data for political leaning, left or right, for individuals born in a given cohort and consider how this changes as they age.

Then we could repeat this for a number of cohorts holding the age range fixed. Of course, nothing would stop us from artificially creating rectangular age-cohort arrays by mapping, for example, a cohort-period panel into age-cohort space and dropping superfluous cells.

In what follows we abstract from the issue that discrete time recordings break the exact link between age, period and cohort. Taking yearly time information as an example, individuals born in year 1988 turn 30 in 2018 but may be recorded as either age 29 or age 30 depending on whether the sample was taken before or after their birthday. This is illustrated for example by Carstensen (2007) who uses data with information for all three time-scales. However, while information for only two time-scales in discrete time does not allow us to exactly pin down the third, we can pin down a range. For example, individuals born in 1988 must necessarily report either age 29 or 30 in 2018. A direct way to solve the issue is to explicitly model continuous time data if it is available. An example is the "continuous chain-ladder" by Martínez-Miranda, Nielsen, Sperlich & Verrall (2013). Hiabu (2017) considers biases arising from modeling continuous time data in discrete time.

1.3 Linear age-period-cohort predictor

Besides age-period-cohort data, a defining feature of (linear) age-period-cohort or extended chain-ladder models is the predictor

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta.$$

In age-period-cohort terminology, α_i, β_j and γ_k are the age, period and cohort effects for age i, period j and cohort k, respectively. In extended chain-ladder terminology, the ordering is usually changed and α_i, β_j and γ_k represent accident, development and calendar year effects, respectively. This predictor is quite flexible. For example, a linear age effect arises as a special case by placing a parametric assumption on the age effects so $\alpha_i = a_0 + a_1 i$. Heckman & Vytlacil (2001) caution against such parametric assumptions that often go untested. In their application, they find evidence against linear age and time effects.

A special case of the predictor that deserves particular attention is the age-cohort (chain-ladder) predictor which does not feature a period (calendar year) effect. With this

predictor, it is often possible to estimate all relevant parameters for forecasting from the data, making parameter extrapolation unnecessary. Lee et al. (2015), in a continuous time setting, refer to this as in-sample forecasting. For example, in a run-off triangle, all accident and development year effects needed to forecast the empty lower triangle also appear in the upper triangle for which we have data. Of course, if we wanted to forecast cells for accident or development years outside the array or if there are calendar year effects, parameter extrapolation would become a necessity. As an example from epidemiology, Martínez Miranda et al. (2015) use an age-cohort model that does not require parameter extrapolation to forecast mesothelioma mortality.

The age-period-cohort predictor has a well known identification problem. As discussed for example in Kuang et al. (2008*b*), for any $a, b, c, d \in R$, recalling that $j = i + k - 1$,

$$\mu_{ik} = (\alpha_i + a + id) + (\beta_j + b - jd) + (\gamma_k + c + kd) + (\delta - a - b - c - d).$$

Thus, four constraints are necessary. Ad-hoc identification is common. For example, we could require age, period and cohort effects to sum to zero and set one additional parameter to zero. A more involved but still ad-hoc approach is the intrinsic estimator, discussed for example by Yang et al. (2004). Nielsen & Nielsen (2014) point out that ad-hoc identification constraints are not necessarily innocent: parameter estimates become entangled and in plots of, say, age-parameters the linear trend is determined by the ad-hoc constraints, not informed by the data. This is also true for levels so that the sign of first differences cannot be interpreted. However, second differences are informed by the data and invariant to the identification approach used. Kuang et al. (2008*b*) propose a parametrization in terms of freely varying parameters based on the second differences. This parametrization is canonical in a Poisson model. Nielsen (2014) discusses a number of sub-models of the linear age-period-cohort model, such as age-cohort models, and how they relate to the canonical parametrization.

Crucially, ad-hoc identification can introduce arbitrary elements into forecasts when parameter extrapolation is needed. This is investigated in Kuang et al. (2008*a*). In a nutshell, parameter extrapolation has to be done by a method that preserves linear trends; only then does the ad-hoc identification not impact forecasts. Nielsen & Nielsen (2014) also

offer further discussion of the identification issue in this context, including for the intrinsic estimator. Kuang et al. (2011) consider forecasting with parameter extrapolation in an insurance context.

We remark that non-linear age-period-cohort models are available as well. A famous example is the Lee-Carter model (Lee & Carter 1992). This model is a popular choice for mortality modeling and features the non-linear age-period predictor $\mu_{ik} = \alpha_{1,i} + \alpha_{2,i}\beta_j$. Nielsen & Nielsen (2014) discuss the identification problem for this predictor, too. In this thesis, we focus exclusively on linear age-period-cohort and extended chain-ladder models as well as their nested sub-models.

1.4 Contributions of the thesis

We briefly summarize the contributions of this thesis. The contributions are to the theory for over-dispersed Poisson and log-normal age-period-cohort and extended chain-ladder models with a focus on applications to non-life insurance claim reserving. The theory is for aggregate age-period-cohort data without covariates with linear age-period-cohort or extended chain-ladder predictors.

In Chapter 2, we develop an asymptotic framework for over-dispersed Poisson models. The theory is based on infinitely divisible distributions and keeps the dimension of the data array fixed; instead it grows the cell means. In an insurance context this can again be motivated by a compound Poisson story. The aggregate incremental payments are the sum of a, potentially unknown, Poisson number of random individual payments. We consider a scenario in which the number of individual claims is large. Besides many compound Poisson distributions, the class of infinitely divisible distributions also includes Poisson, gamma, and negative binomial distributions. The intuition of the theory is that the array is roughly normally distributed for large means so that results are reminiscent of a classical analysis of variance (ANOVA) setting. We show that Poisson quasi-likelihood estimators are t-distributed and F-tests based on Poisson likelihoods can be used to test for model reduction, such as for the absence of a calendar effect. Finally, forecasts errors are t-distributed, giving rise to closed form distribution forecasts including for cell sums such as

the reserve or cash-flow.

In Chapter 3, we build on the asymptotic framework from Chapter 2 and propose an asymptotic theory for misspecification tests for two crucial assumptions of the over-dispersed Poisson model. First, the variance to mean ratio is common across the array. Second, accident effects cannot vary over development years and vice versa. To test these assumptions, we suggest to split the run-off triangle into sub-samples and then to test whether a reduction from individual models for each sub-sample to a single model for the full array is justifiable. Here, too, the theory is reminiscent of an ANOVA setting. A classical Bartlett test (Bartlett 1937) can be used to assess whether we can justify common variance to mean ratios. This is followed by an independent F-test for the absence of breaks in accident and development effects. Again, the asymptotics needed to arrive at these results keep the dimension of the array fixed, growing instead the cell means. We also show that these misspecification tests can be used in similar fashion in a finite sample log-normal model.

In Chapter 4, we put forward a test that allows us to assess whether an over-dispersed Poisson or a log-normal model is more appropriate for the data. This choice is often not obvious. Yet, the two models are clearly different: the over-dispersed Poisson model imposes a fixed variance to mean ratio while the log-normal model assumes the same for the standard deviation to mean ratio. We leverage this insight to develop a non-nested test that has high power to distinguish between the two models. The idea for the test relates to the encompassing literature. It asks whether the null-model can accurately predict the behavior of statistics from the rival model. Once again, the theory is asymptotic but makes use of information accumulating within the cells, holding the array size fixed.

1.5 Motivating example revisited

To conclude the introduction, we return to the motivating example and illustrate how we could apply the contributions of the thesis.

First, we ask whether the over-dispersed Poisson model is a reasonable choice for the data in Table 1.1 when compared to a log-normal model. We do this for an extended-ladder

model that allows for a calendar effect so $\mu_{ij} = \alpha_i + \beta_j + \gamma_k + \delta$. Based on the test put forward in Chapter 4, we cannot reject the over-dispersed Poisson model with a p-value of 0.92. If we instead considered a log-normal model as the null hypothesis and tested against the over-dispersed Poisson model, we would reject the log-normal model with a p-value of 0.001.

Next, we assess whether the calendar effect is needed in the over-dispersed Poisson model. To do so, we test whether $\gamma_k = 0$ using an F-test as proposed in Chapter 2. We cannot reject the null hypothesis with a p-value of 0.30. Thus, we decide to drop the calendar effect so the linear predictor reduces to $\mu_{ij} = \alpha_i + \beta_j + \delta$. With this model, point forecasts replicate the chain-ladder technique and we sidestep parameter extrapolation.

In the over-dispersed Poisson model without calendar effect, we test for misspecification. As explained in Chapter 3, we first split the array into sub-samples such as depicted in Figure 1.1. Then, we estimate separate models on all sub-samples. Based on these esti-

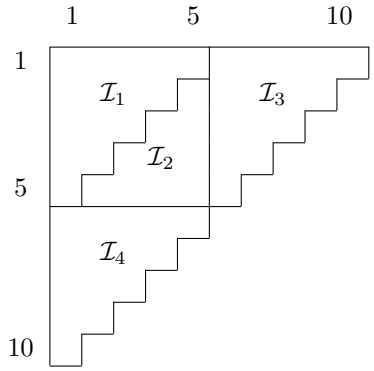


Figure 1.1: Split of a run-off triangle into four sub-samples.

mates, we first look for evidence against the assumption of common variance to mean ratios across the array. As shown in Chapter 3, we use a Bartlett test which cannot convincingly reject that the variance to mean ratio is identical with a p-value of 0.08. Next, we ask whether we can reasonably estimate accident and development year effects from the full run-off triangle rather than the sub-samples. We cannot reject this simplification with a p-value of 0.93 based on an F-test that we show to be independent of the Bartlett test.

Now reassured that the over-dispersed Poisson chain-ladder model without calendar effect is a reasonable choice, we can compute distribution forecasts as shown in Chapter 2.

Calendar Year	Cash-flow	95% value at risk		Accident Year	Reserve	95% value at risk	
		t	bootstrap			t	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 1.4: Chain-ladder distribution forecasts for the insurance data in Table 1.1. Results in ten thousands. t is our distribution forecast, bootstrap refers to the method by England & Verrall (1999) and England (2002). Bootstrap implemented with help of the R (R Core Team 2016) package `ChainLadder` (Gesmann et al. 2015). The t forecast is available in the R package `apc` (Nielsen 2015).

The distribution forecast for $Y_{\mathcal{A}}$, which may be the cash-flow or reserve, is given by

$$\widetilde{Y}_{\mathcal{A}}^{dist} = \widetilde{Y}_{\mathcal{A}}^{CL} + (\text{Process Error} + \text{Estimation Error})^{1/2} t_{df}$$

where t is a t-distributed random variable. The forecast takes into account both the process error, concerning the deviation of $Y_{\mathcal{A}}$ from its mean, and the estimation error, relating to the deviation of the point forecast $\widetilde{Y}_{\mathcal{A}}^{CL}$ from the mean. Both can easily be computed in closed form. Table 1.2 shows the 95% value at risk, the 95th percentile of the forecast distribution, for cash-flow, reserve, and total reserve. Comparing to the bootstrap methodology by England & Verrall (1999) and England (2002), the two methods give a broadly similar indication of the uncertainty for this data.

1.6 Remarks

Chapters 2, 3 and 4 are written in the style of academic papers and meant to be self contained. Thus, each chapter finishes with its own appendix. In contrast, to avoid redundancy, there is a single bibliography at the end of the thesis. Chapter 2 has in similar form, co-authored with Bent Nielsen, been accepted for publication at the *Journal of the American Statistical Association* and this reference is cited in Chapters 3 and 4. In Chapter 2, the terminology is largely aimed at the age-period-cohort literature so that α_i is the age

effect for age i (development year effect), β_j is the period effect for period j (calendar year effect), and γ_k is the cohort affect for cohort k (accident year effect). Chapter 3 has is in similar form been published in the journal *Risks* and this reference is cited in Chapter 4. The terminology is aimed at the actuarial science literature. Thus α_i is the accident year effect for accident year i (cohort effect), β_j is the development year effect for development year j (age effect), and γ_k is the calendar year effect for calendar year k (period effect). Chapter 4 also uses the actuarial science terminology, identical to that in Chapter 3.

2. Over-dispersed Age-Period-Cohort Models

SUMMARY We consider inference and forecasting for aggregate data organized in a two-way table with age and cohort as indices, but without measures of exposure. This is modeled 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 the 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.

2.1 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 (2.2) 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 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 modeling 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 distribution 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 distributions. 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 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) and non-parametric, continuous time variation (Lee et al. 2015). The age-period-cohort specification of the log mean or linear predictor is

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta, \tag{2.1}$$

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 reparameterize 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

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 2.1: 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.

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 parameterizations 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 (2011). We apply the triangle of Taylor & Ashe (1983) as shown in Table 2.1. 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 §2.2 we derive a limit theorem for infinitely divisible distributions. In §2.3 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 §2.4 and results for forecasts that do not require parameter extrapolation in §2.5. We apply the results in a data example in §2.6. The simulation study in §2.7 shows that the asymptotic results give good approximations in finite samples. Finally, we discuss directions for future research in §2.8 where we also briefly consider forecasting with parameter extrapolation.

2.2 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), and 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; all proofs are in the appendix of this chapter.

Theorem 2.1. *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, $skew(Y_\ell) = E[\{Y_\ell - E(Y_\ell)\}/\sqrt{\text{var}(Y_\ell)}]^3 \to 0$, then, in distribution,*

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

Theorem 2.2. *If the conclusion of Theorem 2.1 is met and the ratio of mean to standard deviation increases,* $E(Y_\ell)/\sqrt{\text{var}(Y_\ell)} \to \infty$*, 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.

2.3 Model

2.3.1 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

$$\mathcal{I} = (i, k : 1 \leq i \leq I, 1 \leq k \leq K, L+1 \leq j \leq L+J), \tag{2.2}$$

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 2.1 is a generalized trapezoid with $I = K = J = 10$, $L = 0$ and $n = 55$.

2.3.2 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 (2.2). 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 (2.1) 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 modeling vital data one will sometimes be interested in mortality rates. This would be modeled 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.

2.3.3 Identification

It is well known that the age, period and cohort effects α_i, β_j, γ_k and the level 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 (2.2), Kuang et al. (2008*b*) suggest to represent μ_{ik} as

$$\mu_{ik} = \mu_{11} + (i-1)(\mu_{21} - \mu_{11}) + (k-1)(\mu_{12} - \mu_{11}) + \sum_{t=3}^{i}\sum_{s=3}^{t} \Delta^2 \alpha_s + \sum_{t=3}^{i+k-1}\sum_{s=3}^{t} \Delta^2 \beta_s + \sum_{t=3}^{k}\sum_{s=3}^{t} \Delta^2 \gamma_s. \tag{2.3}$$

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.

2.3.4 Estimation in a Poisson model

A Poisson model satisfies the assumptions in §2.3.2 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

$$i_\xi = -\frac{\partial^2}{\partial\xi\partial\xi^T}\ell_Y(\xi) = \sum_{ik\in\mathcal{I}} \exp(\mu_{ik}) \begin{pmatrix} 1 \\ x_{ik}^{(2)} \end{pmatrix} \begin{pmatrix} 1 \\ x_{ik}^{(2)} \end{pmatrix}^T. \tag{2.4}$$

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.

2.3.5 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

$$\tau = E(Y_{\cdot\cdot}) = \exp(\xi^{(1)}) \sum_{ik\in\mathcal{I}} \exp\{(x_{ik}^{(2)})^T\xi^{(2)}\} \tag{2.5}$$

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 reparameterized log likelihood is

$$\ell_Y(\psi) = \ell_{Y_{\cdot\cdot}}(\tau) + \ell_{T^{(2)}|Y_{\cdot\cdot}}(\xi^{(2)}) \tag{2.6}$$

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 (2.4). To see this, introduce the frequencies

$$\pi_{ik} = \frac{E(Y_{ik})}{\tau} = \frac{\exp\{(x_{ik}^{(2)})^T\xi^{(2)}\}}{\sum_{ik\in\mathcal{I}}\exp\{(x_{ik}^{(2)})^T\xi^{(2)}\}}, \tag{2.7}$$

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

$$j_\psi = -\frac{\partial^2}{\partial\psi\partial\psi^T}\ell_Y(\psi) = \hat{\tau}\begin{pmatrix}\tau^{-2} & 0\\ 0 & \bar{\imath}_{\xi^{(2)}}\end{pmatrix}. \tag{2.8}$$

2.3.6 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.

2.3.7 Sampling scheme in an over-dispersed Poisson model

Consider the over-dispersed Poisson model described in §2.3.2. 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 2.1 and 2.2 apply and we can make asymptotic inference about $\xi^{(2)}$ but neither about τ nor $\xi^{(1)}$; the latter follows from (2.5) 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 (2.8) 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.

2.4 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 $r \le q \le p$ 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 2.1. *In the over-dispersed Poisson model of §2.3.2 and §2.3.7, 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 2.3. *In the over-dispersed Poisson model of §2.3.2 and §2.3.7, 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 2.3 applies to many, but not all, ad-hoc identified parameterizations. 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 2.3 applies.

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

Theorem 2.4. *In the over-dispersed Poisson model of §2.3.2 and §2.3.7, 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 2.4 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.

2.5 Forecasting

2.5.1 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 2.1, 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 §2.3.2 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 §2.8.

2.5.2 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 $\mathcal{I}$.

2.5.3 Distribution forecasting

The aim is to predict the distribution of the difference between the point forecast $\widetilde{Y}_\mathcal{A}$ and the realization $Y_\mathcal{A}$. Defining $\hat{\pi}_\mathcal{A} = \sum_{ik\in\mathcal{A}} \pi_{ik}(\hat{\zeta}^{(2)})$ with π_{ik} as in (2.7) we find three contributions for the forecast error:

$$Y_\mathcal{A} - \widetilde{Y}_\mathcal{A} = Y_\mathcal{A} - E(Y_\mathcal{A}) - \hat{\tau}(\hat{\pi}_\mathcal{A} - \pi_\mathcal{A}) - (\hat{\tau} - \tau)\pi_\mathcal{A}. \tag{2.9}$$

The first contribution is the process error which, extending Theorems 2.1 and 2.2, 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 2.1 and the δ-method,

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

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

where

$$s^2_{\mathcal{A}} = (\sum_{ik\in\mathcal{A}} \pi_{ik}H_{ik})^T(\bar{\imath}_{\zeta^{(2)}})^{-1}(\sum_{ik\in\mathcal{A}} \pi_{ik}H_{ik}). \tag{2.10}$$

The third contribution pertains the estimation uncertainty for τ. By Lemma 2.1,

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

Using Lemma 2.1 again to combine, we arrive at the following theorem.

Theorem 2.5. *In the over-dispersed Poisson model of §2.3.7 and §2.5.1, in distribution,*

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

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.

2.6 Data example

We apply the theory to the insurance run-off triangle shown in Table 2.1. All R (R Core Team 2016) code is given in the supplementary material. We use the R package `apc` (Nielsen 2015).

Table 2.2 shows a deviance analysis based on Theorem 2.4. 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 §2.8, 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 2.4.

Table 2.3 shows 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. In contrast, 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 2.1 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 2.3. Plots of the detrended time-effects (d-f) 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 triangle so these estimates are based on fewer observations.

Table 2.4 shows forecasts for the empty lower triangle from an age-cohort model based on Theorem 2.5. 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

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

is the aggregate over the full lower triangle. We report point forecasts and the 95% quantile of the forecast distribution

$$\widetilde{Y}_{\mathcal{A}} + [\hat{\tau}\{D_1/(n-q)\}\{\hat{\pi}_{\mathcal{A}} + \hat{s}^2_{\mathcal{A}} + (\hat{\pi}_{\mathcal{A}})^2\}]^{1/2} t_{n-q} \tag{2.11}$$

where $\hat{\pi}_{\mathcal{A}}$ and $\hat{s}^2_{\mathcal{A}}$ are (2.7) and (2.10) 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 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 centered 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

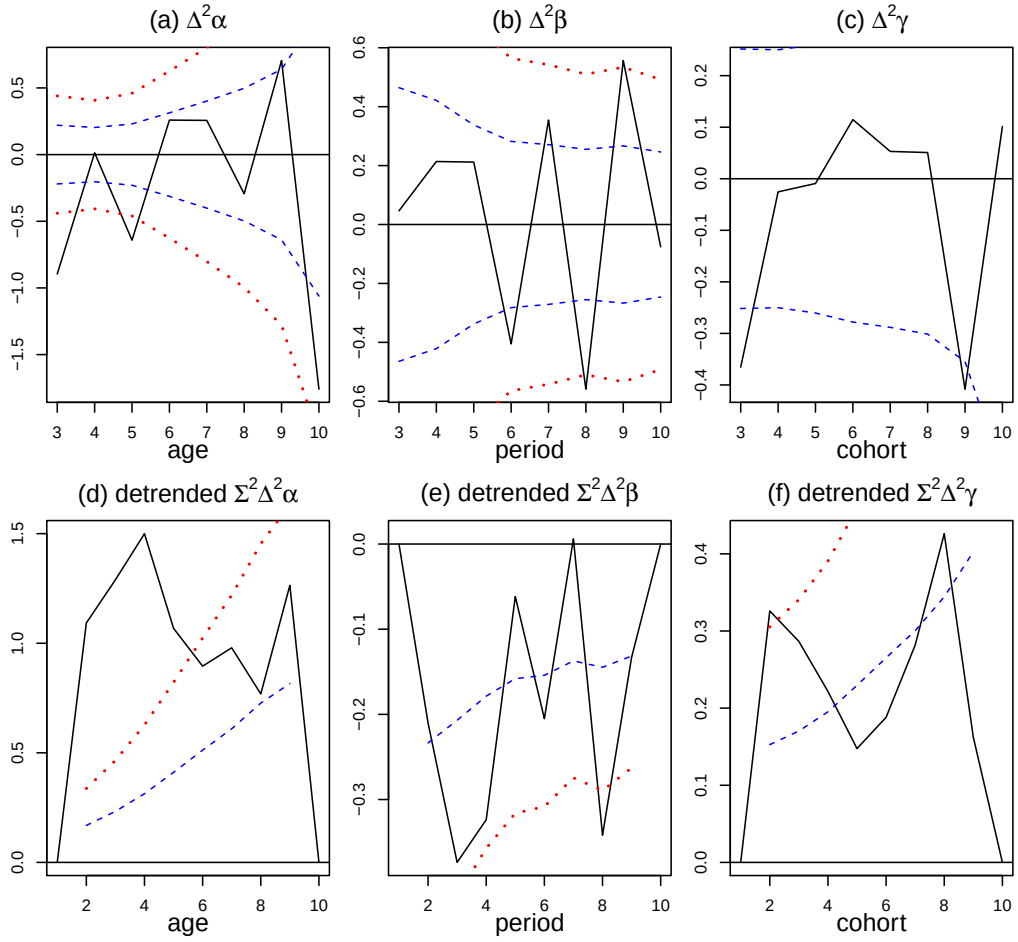


Figure 2.1: Plot of double differences and detrended parameter estimates. Dashed and dotted lines are one and two standard errors se_t around zero, respectively.

the 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.

2.7 Simulation study

2.7.1 Test statistic

We assess the finite sample performance of the asymptotically F distributed specification test F_1 proposed in Theorem 2.4. We simulate under the age-cohort hypothesis H_{ac} so

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 2.4: 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 2.5: Simulation performance of F test. Monte Carlo standard error less than 0.05.

$\Delta^2\beta_j = 0$ 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 2.2 and Table 2.3, respectively, except μ_{11} is chosen as $\log(s) + \hat{\mu}_{11}$ for $s = 0.5, 1, 2$ so $\tau = s\hat{\tau}$. We draw 10^6 repetitions.

Table 2.5 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 2.6: Simulation performance of t and bootstrap forecasts. Results in hundred thousands. Shown are averages across simulations and, in parentheses, root mean square errors.

2.7.2 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 2.5 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 2.6.

We consider the data generating process described in §2.7.1 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 realizations.

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 (2.11) minus the point forecast. For the bootstrap, we proceed as described in §2.6 and draw for every $\mathcal{I}_r$ a bootstrap sample of $B = 999$ realizations 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, respectively.

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.

2.8 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 §2.5 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 (2.9). 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.

Appendix

Proof of Theorem 2.1

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

$$\phi_\ell(t) = \log[E\{\exp(itY_\ell)\}] = i\gamma_\ell t + \int_0^\infty \{\exp(ity) - 1 - ity\}\nu_\ell(dy) \tag{2.12}$$

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

$$\gamma_\ell = E(Y_\ell), \quad \int_0^\infty y^2\nu_\ell(dy) = \mathrm{var}(Y_\ell), \quad \int_0^\infty y^3\nu_\ell(dy) = E[\{Y_\ell - E(Y_\ell)\}^3] \tag{2.13}$$

From Billingsley (1995, page 343) we get

$$\exp(ity) = 1 + ity - \frac{(ty)^2}{2} + r(y,t), \quad |r(y,t)| \leq \min\left\{\frac{|yt|^3}{6}, (yt)^2\right\}. \tag{2.14}$$

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 (2.13). Inserting (2.13) and (2.14) in (2.12),

$$\phi_\ell(t) = itE(Y_\ell) - \frac{t^2}{2}\text{var}(Y_\ell) + \int_0^\infty r(y,t)\nu_\ell(dy). \tag{2.15}$$

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 (2.15) gives

$$\rho_\ell(s) = -\frac{s^2}{2} + \int_0^\infty r\left\{y, \frac{s}{\sqrt{\text{var}(Y_\ell)}}\right\}\nu_\ell(dy). \tag{2.16}$$

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 (2.16) 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 (2.14),

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

where the last inequality follows by the non-negativity of the integrand and (2.13). The minimum is dominated by either of its arguments. It therefore vanishes as $skew(Y_\ell) \to 0$.

Proof of Theorem 2.2

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 2.1

Consider the mixed parametrization of the Poisson likelihood discussed in §2.3.5 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 $(Y_{\cdot\cdot}, T_S^{(2)})$ 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

$$M_\tau^{-1} \frac{\partial \ell_Y}{\partial \psi_S} = \begin{Bmatrix} \hat{\tau} - \tau \\ \hat{\tau}(T_S^{(2)}/\hat{\tau} - \sum_{ik \in \mathcal{I}} s_{ik}^T \pi_{ik}) \end{Bmatrix} = M_\tau^{-1} \frac{\partial \mu^T}{\partial \psi_S} \frac{\partial \ell_Y}{\partial \mu} = M_\tau^{-1} \frac{\partial \mu^T}{\partial \psi_S} \{Y - E(Y)\}. \quad (2.17)$$

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

$$M_\tau^{-1} \frac{\partial \mu^T}{\partial \psi_S} diagonal(\pi) \frac{\partial \mu}{\partial \psi_S^T} M_\tau^{-1} = -\tau^{-1} \left. \frac{\partial^2 \ell_Y}{\partial \psi_S \partial \psi_S^T} \right|_{Y = E(Y)} = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{\imath}_\theta \end{pmatrix} = \tilde{\imath}_{\psi_S}. \quad (2.18)$$

With independent Y_{ik} Theorem 2.1 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 2.2. 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,

$$\hat{\tau}^{-1/2} M_\tau^{-1} \frac{\partial \ell_Y}{\partial \psi_S} = M_\tau^{-1} \frac{\partial \mu^T}{\partial \psi_S} \hat{\tau}^{-1/2} \{Y - E(Y)\} \to N(0, \sigma^2 \tilde{\imath}_{\psi_S}). \quad (2.19)$$

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 2.3

From Lemma 2.1, $\hat{\tau}^{1/2} v^T(\hat{\xi}^{(2)} - \xi^{(2)}) \to N\{0, \sigma^2 v^T(\bar{\imath}_{\xi^{(2)}})^{-1} v\}$, asymptotically independent of $D_{apc} \to \sigma^2 \chi^2_{n-p}$. The studentized estimator is then t_{n-p} distributed.

Proof of Theorem 2.4

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. 1995*b*, page 212). Hence, the F distributed ratios $\{(1 - V_1^2)/df_2\}/\{V_1^2/df_1\}$ and $\{(1 - V_2^2)/df_3\}/\{V_2^2/(df_1 + df_2)\}$ are independent. By Lemma 2.1, 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 2.5

Since $\hat{\tau}/\tau \to 1$ in probability as noted in the proof of Lemma 2.1, $\hat{\tau}^{-1/2}\{Y_{\mathcal{A}} - E(Y_{\mathcal{A}})\}$ 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 2.1. The latter is a function of the future realizations 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 2.1. 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.

3. Misspecification Tests for Chain-Ladder Models

Abstract Despite the widespread use of chain-ladder models, so far no theory was available to test for model specification. The popular over-dispersed Poisson model assumes that the over-dispersion is common across the data. A further assumption is that accident year effects do not vary across development years and vice versa. The log-normal chain-ladder model makes similar assumptions. We show that these assumptions can easily be tested and that similar tests can be used in both models. The tests can be implemented in a spreadsheet. We illustrate the implementation in several empirical applications. While the results for the log-normal model are valid in finite samples, those for the over-dispersed Poisson model are derived for large cell mean asymptotics which hold the number of cells fixed. We show in a simulation study that the finite sample performance is close to the asymptotic performance.

3.1 Introduction

"Can we trust chain-ladder models?" is a central question in non-life insurance claim reserving. It hinges on the model assumptions: if these are violated the answer would be "no". For example, the popular over-dispersed Poisson chain-ladder model assumes a fixed variance to mean ratio across the run-off triangle. If this is false then distribution forecasts are bound to fail. Yet, there is no statistical theory available to test for a violation of this assumption.

We show that testing for a violation of central assumptions is straightforward in two popular chain-ladder models: over-dispersed Poisson and log-normal. While the over-dispersed Poisson model assumes a fixed variance to mean ratio, the log-normal model imposes a common variance of the log data. Further, both models assume a chain-ladder structure. That is, accident year effects do not vary by development year and vice versa.

We show that these assumptions are not only testable, but testable with standard tools that can easily be implemented in a spreadsheet.

The over-dispersed Poisson model arguably owes its special status to the ubiquitous chain-ladder technique. Kremer (1985) showed that this deterministic technique so commonly used in claim reserving is replicated by maximum likelihood estimation in a Poisson model. However, integer support and the implicit assumption that the variance equals the mean cannot be reconciled with insurance claim data. This explains the need for the over-dispersed Poisson model which relaxes both of these assumptions. Unlike the Poisson model, the over-dispersed Poisson model is moment-based and does not come equipped with a distributional framework. Despite this shortcoming, distribution forecasts are needed and bootstrapping (England & Verrall 1999, England 2002) is in widespread use. Yet, so far we do not have a statistical theory for the bootstrap in this setting.

Recently, Harnau & Nielsen (2017) proposed a distributional framework that incorporates the moment assumptions of the over-dispersed Poisson model. This framework allows for a compelling asymptotic theory that does not require a large array but rather large cell means. The practical implication is that for a run-off triangle with a large, potentially unknown, number of payments, we can use a fixed sample size Gaussian distribution theory. They derive parameter distributions, tests for model reduction, such as the absence of calendar effects, and closed form distribution forecasts. Their assumptions accommodate, among others, many compound Poisson distributions. In insurance, these have the interpretation that each cell of aggregate incremental claims is the sum of a Poisson number of claims each with a random individual claim amount. The asymptotic theory then does not assume that we have many such cells, but rather that the mean of the Poisson number of claims is large. We stress that while Harnau & Nielsen (2017) largely use terminology from the age-period-cohort literature, the theory immediately applies to the reserving literature by renaming age, period, and cohort effects to development, calendar, and accident effects.

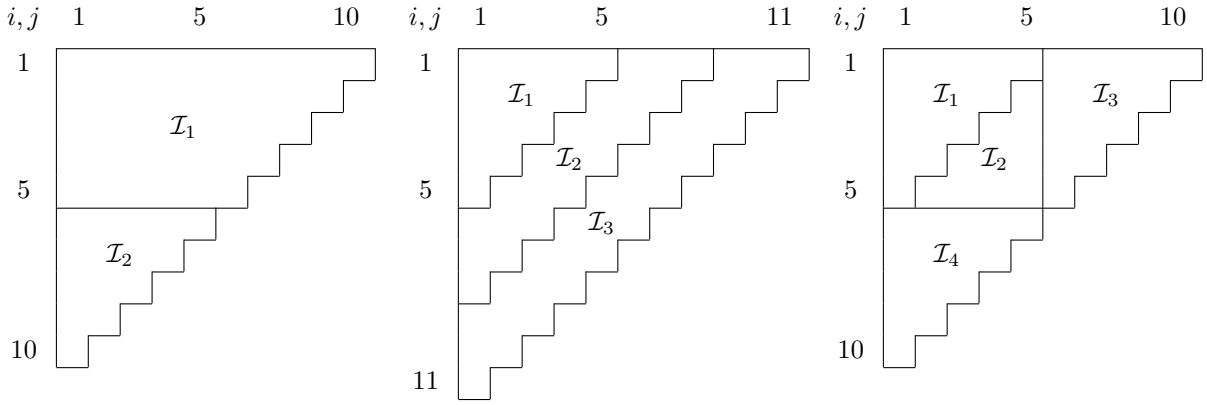
Modeling aggregate incremental claims as log-normal rather than over-dispersed Poisson is also common. Kremer (1982) introduced a log-normal model with multiplicative mean structure mirroring the over-dispersed Poisson chain-ladder model. While this model does not replicate the classic chain-ladder technique, it is easily estimated by least squares.

Recently, Kuang et al. (2015) derived explicit expression for the estimators in the log-normal model. These have interpretation as a geometric, rather than the classic arithmetic chain ladder. Other contributions for the log-normal model are discussed in the excellent overview of stochastic reserving models by England & Verrall (2002).

We are of course not the first to question the validity of the assumptions in these models. Yet, so far the problem was dealt with by specifying more flexible models. For example, Hertig (1983) considers a log-normal model that allows the log data variance to vary by development year. The double-chain-ladder model by Martínez-Miranda et al. (2012) has, conditional on the incurred counts, an approximate over-dispersed Poisson structure where the over-dispersion varies by accident year. The "distribution-free" model by Mack (1993) has separate variance parameters for each development year. We note that while this model also replicates the classical chain-ladder point forecasts, it differs from the over-dispersed Poisson model and so far lacks a distributional framework that would allow for a rigorous statistical theory. Thus, while it is a popular model, we do not consider it further in this paper.

While using more flexible models seems sensible when assumptions are violated, we should not be too quick to dispose of well-known simple models. Particularly for forecasting, such simpler models may be advantageous. A statistical framework for misspecification testing is thus needed. The tests may corroborate the initial modeling choice of the expert, draw attention to an issue, or confirm the suspicion that the model is not well suited for the task. Whichever scenario the expert encounters, the misspecification tests can help to make an informed choice.

The test statistics we propose in this paper are well known in an analysis of variance (ANOVA) context. There, the researcher is usually presented with several samples and wants to test for treatment effects. The data are often assumed to be independent Gaussian. The first step is to test for common variances across samples. This is done with a Bartlett test based on an easily computed likelihood ratio statistic. Then, given common variances, a standard F-test can be used to test for different means between the samples, indicating a treatment effect.



(a) Two sub-samples (b) Three sub-samples (c) Four sub-samples

Figure 3.1: Examples for splits of run-off triangles into two (**a**), three (**b**) and four (**c**) sub-samples. Sub-samples are denoted by $\mathcal{I}_\ell$. Accident years i are in the rows, development years j in the columns.

The difference to the ANOVA application is that we generally have data for only one sample, often a run-off triangle. We thus reverse engineer the ANOVA situation by splitting the data into several artificial sub-samples. This idea has a long history in the econometric literature. For instance, Chow (1960) proposed a test for structural breaks that involved splitting the sample at the known breakpoint. In the (weak) instrumental variable literature, Angrist & Krueger (1995) proposed a split-sample procedure with the objective to break the bias of the instrumental variable estimator towards the ordinary least squares estimator. Figure 3.1 shows examples of how we could split run-off triangles into sub-samples. In §3.2, we give a precise definition for the conditions that both the data set as well as the artificial sub-samples must meet. We note that while we do not provide guidance on how to choose the sub-sample structure in this paper, the choice does not affect the size of the proposed tests under the null hypothesis.

In a log-normal model, taking logs yields Gaussian data such that we can directly apply the Bartlett and F-test from the ANOVA scenario. While the finite sample distribution of the Bartlett test statistic has no closed form, it does not have nuisance parameters and critical values could easily be simulated. However, Bartlett (1937) suggests a χ^2 approximation to the exact distribution that allows us to sidestep simulations. For a special case with just two sub-samples, we can also apply an F-test for the hypothesis for common variances of

the log data ; while Bartlett and F-tests are not identical, simulations indicate that they give similar information. Next, we show that an F-test for common mean parameters is not only straightforward but also independent of the Bartlett test. These results are collected in §3.3.

In the over-dispersed Poisson model, the asymptotic framework by Harnau & Nielsen (2017) catapults us into a finite dimensional Gaussian world. Therefore, the results developed for the log-normal model carry over. We can now asymptotically use a Bartlett test as a test for common over-dispersion across sub-samples. Similarly, an F-test for common mean parameters across sub-samples is asymptotically F-distributed and asymptotically independent of the dispersion parameter tests. We stress again that the asymptotic theory does not require a large triangle but rather large means of the cells in the triangle. As for the log-normal model, we could simulate critical values for the Bartlett test; however a χ^2 approximation can still be justified. We show all this in §3.4.

The Bartlett test is easily implemented and makes an empirical application straightforward. The same is true for an F-test on the means. We illustrate the testing procedure, splitting the data, estimating the sub-models, Bartlett testing for common dispersion parameters, and F-testing for common mean parameters, in §3.5 with several empirical applications.

We clear up remaining questions about the power of the tests and the performance of approximations in a simulation study based on a run-off triangle. First, it would not take much to simulate critical values of the Bartlett test statistic under the null, rather than to use a χ^2 approximation. However, we show in a simulation study that this approximation works so well that simulating critical values seems superfluous. Second, we produce power curves under several alternatives for the test for common variances of the log data in a log-normal model. Third, we find that the asymptotic results for the over-dispersed Poisson model are well approximated in finite samples, at least in our simulations. The simulation study is in §3.6.

Finally, we discuss some open questions for future research such as how to choose the sub-sample structure and whether one can select between the over-dispersed Poisson and the log-normal model. With this, §3.7 concludes the paper.

3.2 Data and Sub-Samples

Our aim is to test model specification by using statistics that are usually employed to test for common parameters across separate samples. However, we are presented with just a single sample, such as a run-off triangle. Thus, we artificially construct separate samples by splitting the data at hand into sub-samples. Many intuitive splits can be accommodated by the theory, for example, all sub-samples in Figure 3.1. Here, we define precisely the permissible structures for data and sub-samples, illustrated on an example of a run-off triangle.

For the theory in this paper, we assume that data are a generalized trapezoid as defined by Kuang et al. (2008*b*). This flexible format allows for different numbers of accident and development years, and can accommodate missing past and future calendar years. Run-off triangles are a special case with as many accident as development years and only future calendar years missing. For accident year i and development year j, we count calendar years k with an offset so $k = i + j - 1$. Generalized trapezoids are characterized by the index set

$$\mathcal{I} = \{(i, j) : I^l \leq i \leq I^u, J^l \leq j \leq J^u, K^l \leq k \leq K^u\},$$

where I^l and I^u, J^l and J^u, and K^l and K^u are the smallest and largest accident, development and calendar year indices available, respectively. We denote the number of cells in $\mathcal{I}$ by n. The run-off triangle in Table 3.1, taken from Taylor & Ashe (1983), are a generalized trapezoid with $I^l = J^l = K^l = 1$, $I^u = J^u = K^u = 10$ and $n = 55$.

We also assume that each sub-sample is a generalized trapezoid. We denote sub-samples by $\mathcal{I}_1, \ldots \mathcal{I}_m$. The sub-samples should be disjoint so $\mathcal{I}_s \cap \mathcal{I}_t = \emptyset$ and their union should be the original sample so $\cup_\ell \mathcal{I}_\ell = \mathcal{I}$. All sub-samples of the examples in Figure 3.1 are generalized trapezoids. For instance, the sub-sample $\mathcal{I}_2$ in Figure 3.1c is specified by $I_2^l = J_2^l = 2$, $I_2^u = J_2^u = 5$, $K_2^l = 6$, $K_2^u = 9$ and $n_2 = 10$.

The purpose of the generalized trapezoid assumption is to ensure parameter identification later on. We note that this assumption is often more restrictive than needed. Examples for arrays that do not fall into the generalized trapezoid category are arrays with missing

i,j	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 3.1: Insurance run-off triangle taken from Taylor & Ashe (1983) as an example for a generalized trapezoid. Entries are aggregate incremental paid amounts for claims of accident year i and development year j. Calendar years $k = i + j - 1$ are on the diagonals increasing from the top left.

cells and disconnected arrays such as the combination of sub-samples $\mathcal{I}_1$ and $\mathcal{I}_3$ in Figure 3.1b. However, for many of these arrays identification may still be given and then the theory developed below will still be valid.

3.3 Log-Normal Model

Given data and sub-samples, we can specify a log-normal model, define estimators, and provide the theory for specification testing. The idea is to start with a model that allows parameters to vary across sub-samples and then to test for reductions to a model with common parameters. The latter, most restrictive, model is commonly used in claim reserving. If we reject a reduction to this model, it is likely misspecified. Estimation is done by least squares. The first hypothesis is that log data variances are common across sub-samples; we can test this with a Bartlett test. The second hypothesis is for common linear predictors and can be assessed with an independent F-test.

3.3.1 Model and Hypotheses

The unrestricted model allows both log data means and variances to vary across sub-samples. For this model, we assume that the aggregate incremental claims $Y_{ij,\ell}$ for accident year i, development year j, and sub-sample ℓ are independent log-normal with

$$M^{LN}: \quad \log(Y_{ij,\ell}) \stackrel{D}{=} N(\mu_{ij,\ell}, \sigma_\ell^2), \quad \mu_{ij,\ell} = \alpha_{i,\ell} + \beta_{j,\ell} + \delta_\ell \quad \forall (i,j) \in \mathcal{I}_\ell,\ \ell \in \{1, \ldots, m\}.$$

While we focus on linear predictors $\mu_{ij,\ell}$ with accident and development year effect, the theory in this paper allows for more general or restrictive linear predictors. For example, we could incorporate calendar year effects as in Zehnwirth (1994) or Kuang et al. (2011).

The first hypotheses restricts log data variances to be common across sub-samples $\mathcal{I}_\ell$. The remaining assumptions are maintained; thus, linear predictors are still allowed to vary across sub-samples. We write the hypothesis as

$$H_{\sigma^2} : \sigma_\ell^2 = \sigma^2 \quad \forall \ell \in \{1, \dots, m\}.$$

The model that arises by imposing this restriction is

$$M_{\sigma^2}^{LN} : \log(Y_{ij,\ell}) \stackrel{D}{=} N(\mu_{ij,\ell}, \sigma^2).$$

The second hypothesis nests the first but also restricts linear predictors to be common across sub-samples. The hypothesis is

$$H_{\mu,\sigma^2} : \sigma_\ell^2 = \sigma^2 \text{ and } \mu_{ij,\ell} = \mu_{ij} = \alpha_i + \beta_j + \delta \quad \forall \ell \in \{1, \dots, m\}.$$

Under this hypothesis, all parameters are common across sub-samples $\mathcal{I}_\ell$. Thus, we can feasibly drop the sub-script ℓ and write the model under this hypothesis as

$$M_{\mu,\sigma^2}^{LN} : \log(Y_{ij}) \stackrel{D}{=} N(\mu_{ij}, \sigma^2).$$

This is the log-normal geometric chain-ladder model.

We can also think about the hypotheses on the original scale. Mean and variance parameters on the log-scale map into median and coefficients of variations on the original scale. Taking the model M_{μ,σ^2}^{LN} under H_{μ,σ^2} as an example,

$$\log(Y_{ij}) \stackrel{D}{=} N(\mu_{ij}, \sigma^2) \implies \text{Median}(Y_{ij}) = \exp(\mu_{ij}), \quad \frac{SD(Y_{ij})}{E(Y_{ij})} = \sqrt{\exp(\sigma^2) - 1}.$$

Thus, the separation between mean and variance on the log-scale translates to separation between median and coefficient of variation on the original scale. Hence, we can alternatively think of H_{σ^2} as the hypothesis of common coefficients of variation and of H_{μ,σ^2} as further imposing common median parameters.

3.3.2 Estimation

We estimate on the log-scale with standard estimators, least squares for log data means and residual sum of squares for log data variances. Since the theory for testing developed below is adapted from a Gaussian framework, estimation on the log-scale is intuitive. Before specifying the estimators, we briefly discuss identification.

The identification problem is that

$$\mu_{ij} = \alpha_i + \beta_j + \delta = (\alpha_i + a) + (\beta_j + b) + (\delta - a - b)$$

for any a, b. Thus, the levels of accident and development effects are not identified. However, the linear predictors μ are identified (Kuang et al. 2008*b*). These are thus invariant to the identification constraints imposed on the individual effects. Therefore, it does not matter whether we impose ad-hoc constraints such as $\alpha_{I_\ell^l} = \beta_{J_\ell^l} = 0$ or non ad-hoc constraints as suggested by Kuang et al. (2008*b*). We choose to discuss estimation based on the latter, which has the advantage that it allows for straightforward counting of degrees of freedom. By way of example, we apply the identification by Kuang et al. (2008*b*) to a run-off triangle with $\mathcal{I} = \{(i, j) : 1 \leq i, j, k \leq I\}$. Defining the first difference operator as Δ, the idea is to re-write

$$\mu_{ij} = \mu_{11} + \sum_{s=2}^{I} 1_{(i \leq s)} \Delta\alpha_s + \sum_{t=s}^{I} 1_{(j \leq s)} \Delta\beta_s.$$

Then, $\mu_{ij} = x'_{ij}\xi$ where the design vector $x_{ij} = (1, 1_{(i \leq 2)}, \ldots, 1_{(i \leq I)}, 1_{(j \leq 2)}, \ldots, 1_{(j \leq I)},)'$ and the identified parameter vector is $\xi = (\mu_{11}, \Delta\alpha_2, \ldots, \Delta\alpha_I, \Delta\beta_2, \ldots, \Delta\beta_I)'$. We denote the number of parameters as $p = \text{length}(\xi)$. The identification method can be extended to generalized trapezoids as well as to linear predictors with calendar year effects.

3.3.2.1 Estimation in Unrestricted Model M^{LN}

For the unrestricted model M^{LN}, we estimate linear predictors as

$$\hat{\mu}_{ij,\ell}^{LN} = x'_{ij,\ell}\hat{\xi}_\ell^{LN} \quad \text{where} \quad \hat{\xi}_\ell^{LN} = \left(\sum_{ij \in \mathcal{I}_\ell} x_{ij,\ell} x'_{ij,\ell} \right)^{-1} \left\{ \sum_{ij \in \mathcal{I}_\ell} x_{ij,\ell} \log(Y_{ij,\ell}) \right\}.$$

With degrees of freedom $df_\ell = n_\ell - p_\ell$, we estimate log data variances by

$$\hat{\sigma}_\ell^{2,LN} = \frac{RSS_\ell}{df_\ell} \quad \text{where} \quad RSS_\ell = \sum_{ij \in \mathcal{I}_\ell} \{\log(Y_{ij,\ell}) - \hat{\mu}_{ij,\ell}^{LN}\}^2. \tag{3.1}$$

3.3.2.2 Estimation with Common Variances in $M^{LN}_{\sigma^2}$

Imposing the restriction of common log data variances H_{σ^2} does not require re-estimation as the estimators from M^{LN} can be re-used. The estimators for the linear predictors $\mu_{ij,\ell}$ are identical to those of M^{LN}. The log data variance in $M^{LN}_{\sigma^2}$ is estimated by

$$\bar{\sigma}^{2,LN} = \sum_{\ell=1}^{m} \frac{df_\ell}{df_\cdot} \hat{\sigma}^{2,LN}_\ell = \frac{RSS_\cdot}{df_\cdot} \tag{3.2}$$

where $df_\cdot = \sum_{\ell=1}^{m} df_\ell$ and $RSS_\cdot = \sum_{\ell=1}^{m} RSS_\ell$.

3.3.2.3 Estimation with Common Variances and Linear Predictors in M^{LN}_{μ,σ^2}

Under the hypothesis H^{LN}_{μ,σ^2}, which imposes common log data mean and variance parameters, both estimators change. We drop the ℓ-subscript indicating the sub-sample since estimation is done over the full sample $\mathcal{I}$. With that, we write the estimators for the linear predictors in M^{LN}_{μ,σ^2} as

$$\hat{\mu}^{LN}_{ij} = x'_{ij}\hat{\xi}^{LN} \quad \text{with} \quad \hat{\xi}^{LN} = \left(\sum_{ij \in \mathcal{I}} x_{ij} x'_{ij} \right)^{-1} \left\{ \sum_{ij \in \mathcal{I}} x_{ij} \log(Y_{ij}) \right\}.$$

We estimate the log data variance σ^2 under this hypothesis, defining $df = n - p$, by

$$\hat{\sigma}^{2,LN} = \frac{RSS}{df} \quad \text{where} \quad RSS = \sum_{ij \in \mathcal{I}} \{\log(Y_{ij}) - \hat{\mu}^{LN}_{ij}\}^2.$$

3.3.2.4 Remarks

Least squares estimation for the identified parameter vector ξ is maximum likelihood estimation in the log-normal model. Kuang et al. (2015) derive a representation of the least squares estimators that is interpretable as a geometric chain-ladder, in contrast to the classic, arithmetic, chain-ladder.

For many regression models, there is little difference between scaling the residual sum of squares by the degrees of freedom or the number of observations; the former yields an unbiased estimator for σ^2, the latter the maximum likelihood estimator. However, the scaling does matter here due to the large parameter to observation ratio. By way of example, the Taylor & Ashe (1983) data has $n = 55$ observations but only $df = 36$

degrees of freedom so that $\hat{\sigma}^{2,LN}$ is some 50% larger then the rival estimator RSS/n. This is amplified for the sub-samples.

3.3.3 Testing for Common Variances

We show how to test for common log data variances, that is for H_{σ^2} in M^{LN} using a Bartlett test. In a special case with two sub-samples, we can use an F-test instead of a Bartlett test.

The Bartlett test (Bartlett 1937) was designed to test for common variances across several Gaussian samples. Thus, it is directly applicable to the log sub-samples. We only give a rough overview of the theory; for a more detailed derivation in contemporary terminology see Jørgensen (1993, pp. 94–96). The test rests on the independent χ^2-distribution of $\hat{\sigma}_\ell^{2,LN}$ in M^{LN}. Rather than deriving a test in the Gaussian model for $\log(Y_{ij,\ell})$, Bartlett (1937) considers a joint χ^2 model for the variance estimators. In this χ^2 model, the log-likelihood ratio statistic for the hypothesis H_{σ^2} is

$$LR^{LN} = df_\cdot \log(\bar{\sigma}^{2,LN}) - \sum_{\ell=1}^{m} df_\ell \log(\hat{\sigma}_\ell^{2,LN}) \tag{3.3}$$

for $\hat{\sigma}^2$ and $\bar{\sigma}^2$ as defined in (3.1) and (3.2), respectively. Define now the Bartlett distribution $\mathrm{Ba}(\cdot)$ such that $LR^{LN} \stackrel{D}{=} \mathrm{Ba}(df_1, \ldots, df_m)$ under the hypothesis. Considering LR^{LN} as a function of the estimators so $LR^{LN} = LR(\hat{\sigma}_1^{2,LN}, \ldots, \hat{\sigma}_m^{2,LN})$, the Bartlett distribution is characterized by

$$P\{\mathrm{Ba}(df_1, \ldots, df_m) \leq y\} = \int_{A(y)} \prod_{\ell=1}^{m} dG_\ell(x_\ell) \tag{3.4}$$

where $G_\ell(\cdot)$ is the $\chi^2_{df_\ell}$ cdf and $A(y) = \{(x_1, \ldots, x_m) : LR(x_1, \ldots, x_m) \leq y\}$. Likelihood theory tells us that $\mathrm{Ba}(df_1, \ldots, df_m)$ and thus LR^{LN} approaches a χ^2_{m-1} as $\min(df_1, \ldots, df_m)$ goes to infinity. However, Bartlett (1937) goes a step further and suggest to divide LR^{LN} by

$$C = 1 + \frac{1}{3(m-1)} \left(\sum_{\ell=1}^{m} \frac{1}{df_\ell} - \frac{1}{df_\cdot} \right).$$

Comparing LR^{LN}/C rather than LR^{LN} to a χ^2_{m-1} substantially improves the quality of the approximation and makes it useful even in rather small samples. That is, under H_{σ^2},

$$B^{LN} = \frac{LR^{LN}}{C} \stackrel{D}{\approx} \chi^2_{m-1}. \tag{3.5}$$

The Bartlett correction factor C improves the order of magnitude of the error term. This idea has been shown to apply generally to likelihood ratio tests; see, for instance, Lawley (1956) and Barndorff-Nielsen & Cox (1984).

While using an asymptotic approximation for the Bartlett test is appealing, we could also simulate critical values of the exact distribution. This is feasible because the exact distribution of LR^{LN}, Ba, is free of nuisance parameters. However, if Ba$/C$ is sufficiently close to χ^2_{m-1}, simulating the critical values may be unnecessary even for rather small degrees of freedom. Looking ahead, we confirm in a simulation study in §3.6.1 that the asymptotic approximation indeed works very well.

As an alternative to the Bartlett test, we can test the equality of dispersion parameters across two sub-samples with an F-test that is not equivalent to a Bartlett test. The F-test follows quickly given independence and distribution of the log data variance estimators $\sigma_\ell^{2,LN}$ in (3.1). Under H_{σ^2},

$$F_{\sigma^2}^{LN} = \frac{\sigma_2^{2,LN}}{\sigma_1^{2,LN}} \overset{D}{=} \mathrm{F}_{df_2,df_1} \tag{3.6}$$

so that we can use a (two-sided) F-test to test the hypothesis; see, for example, Snedecor & Cochran (1967, chp. 4.15). We can write LR^{LN} as a function of $F_{\sigma^2}^{LN}$. With $r = df_2/df_1$,

$$LR^{LN} = LR(F_{\sigma^2}^{LN}) = df_1 \log\left(\frac{1 + rF_{\sigma^2}^{LN}}{1+r}\right) + df_2 \log\left\{\frac{1 + (rF_{\sigma^2}^{LN})^{-1}}{1 + r^{-1}}\right\}. \tag{3.7}$$

This mapping is not monotone. Intuitively, the Bartlett test is one-sided compared to a two-sided F-test. Thus, we would expect LR to be increasing both for small and large $F_{\sigma^2}^{LN}$. We can now find scenarios in which the F-test and the Bartlett test lead to different decisions: for example, with $df_1 = 1$ and $df_2 = 2$ an equal-tailed 5% F-test just about rejects the null for a draw $F_{\sigma^2}^{LN} = 0.025$, while a 5% Bartlett test does not reject with $LR(0.025) = 4.23$ and a (simulated) exact critical value of 4.91. This leaves the question which test should be used; we investigate this in §3.6.2.

Usually, a drawback of both F and Bartlett test is their sensitivity to departures from Gaussianity of the log data $\log(Y_{ij,\ell})$. Box (1953) goes as far as comparing the Bartlett test to a test for Gaussianity and argues in favor of robust tests, prioritizing robustness over other qualities such as power. However, sensitivity to non-Gaussianity is not necessarily

undesirable for an application to insurance claim-reserving since distribution forecasts of the log-normal model would also be invalid if the data is not log-normal. Besides, we find F-test and Bartlett test appealing for their simplicity and because they carry over to over-dispersed Poisson models as we will see later. Thus, we do not consider methods to improve robustness to departures from Gaussianity such as made by Shoemaker (2003) for F-tests.

3.3.4 Testing for Common Linear Predictors

Now that we know how to test for common variances, we turn to testing for common linear predictors. The idea is to test sequentially: first for common variances, then for common linear predictors. We show how to use an F-tests for the latter and prove that this test is independent of Bartlett and F-tests for common variances. Thus, size control is not an issue.

If we take the model with common variances $M_{\sigma^2}^{LN}$ as given, then testing for H_{μ,σ^2} amounts to testing for common linear predictors. Since standard Gaussian theory applies,

$$F_{\mu}^{LN} = \frac{(RSS - RSS_{\cdot})/(df - df_{\cdot})}{RSS_{\cdot}/df_{\cdot}} \stackrel{D}{=} \mathrm{F}_{df-df_{\cdot},df_{\cdot}}$$

under the hypothesis. Thus, we can use a (one-sided) F-test to test for a reduction from $M_{\sigma^2}^{LN}$ to M_{μ,σ^2}^{LN}. Unlike the dispersion Bartlett and F-tests, this F-test is equivalent to the corresponding exact Gaussian likelihood ratio test. However, a χ^2 approximation to the likelihood ratio test may not work well due to rather few degrees of freedom. Thus, we prefer the F-test since it is easier to implement.

A sequential test approach for common variance and common linear predictors is sensible. This is because we can show the tests are independent. We formulate the independence result in a theorem; all proofs are in the appendix of this chapter.

Theorem 3.1. *In model M_{μ,σ^2}^{LN}, the test statistic F_{μ}^{LN} is independent of $F_{\sigma^2}^{LN}$ and LR^{LN}.*

In applications, we would first conduct a, say, 5% Bartlett test for H_{σ^2}. Conditional on non-rejection of the hypothesis, we can conduct an F-test for H_{μ,σ^2} at 5% critical values and be assured that it truly has a 5% size if the hypothesis is correct.

3.4 Over-Dispersed Poisson

The over-dispersed Poisson model is appealing because it naturally links to the classic chain-ladder technique, unlike the log-normal model. Harnau & Nielsen (2017) developed an asymptotically framework in which the over-dispersed Poisson model is asymptotically Gaussian. Using their results, we show that finite sample results from the log-normal model hold asymptotically in the over-dispersed Poisson model. The structure of this section reflects the similarities between the log-normal and over-dispersed Poisson model. After setting up the model, we specify the estimators; these are based on a Poisson quasi-likelihood, thus replicating the chain-ladder. Before we can proceed, the over-dispersed Poisson model needs another ingredient, a sampling scheme for the asymptotic theory that we take from Harnau & Nielsen (2017). Then, we show that we can use test for common over-dispersion with a Bartlett test. Finally, we can use an F-test to test for common mean parameters. We prove that this F-test is independent of the over-dispersion test.

3.4.1 Model and Hypotheses

We set up a model that allows over-dispersion and mean parameters to vary across sub-samples, and specify hypotheses for common over-dispersion, and common mean parameters. This mirrors the process from the log-normal model. The key assumption of the over-dispersed Poisson model involves infinitely divisible distributions: to justify it we provide an example that is appealing for insurance claim-reserving.

We adopt the assumptions for the over-dispersed Poisson model from Harnau & Nielsen (2017). One assumption is distributional and allows for an asymptotic theory, the other imposes the desired over-dispersed Poisson chain-ladder structure. Specifically, we assume that aggregate incremental claims $Y_{ik,\ell}$ are independent across $(i,k) \in \mathcal{I}_\ell$ and $\ell = \{1,\dots,m\}$ with non-degenerate infinitely divisible distribution, at least three moments, and non-negative support. The second assumption imposes a log-linear mean and common over-dispersion within the sub-sample:

$$M^{ODP}: \qquad E(Y_{ij,\ell}) = \exp(\mu_{ij,\ell}), \quad \mu_{ij,\ell} = \alpha_{i,\ell} + \beta_{j,\ell} + \delta_\ell, \quad \frac{\text{var}(Y_{ij,\ell})}{E(Y_{ij,\ell})} = \sigma_\ell^2$$

for all $(i,j) \in \mathcal{I}_\ell$ and $\ell \in \{1,\dots,m\}$.

The first hypotheses imposes common over-dispersion parameters across sub-samples. It matches the hypothesis from the log-normal model:

$$H_{\sigma^2} : \sigma_\ell^2 = \sigma^2 \quad \forall \ell \in \{1,\dots,m\}.$$

The remaining assumptions are maintained. We can write the model under this assumption as

$$M_{\sigma^2}^{ODP} : E(Y_{ij,\ell}) = \exp(\mu_{ij,\ell}), \quad \frac{\text{var}(Y_{ij,\ell})}{E(Y_{ij,\ell})} = \sigma^2.$$

The second hypothesis again nests the first and imposes common linear predictors. The hypothesis is

$$H_{\mu,\sigma^2} : \sigma_\ell^2 = \sigma^2 \text{ and } \mu_{ij,\ell} = \mu_{ij} = \alpha_i + \beta_j + \delta \quad \forall \ell \in \{1,\dots,m\}.$$

Dropping the superfluous ℓ subscript, we write the model under this hypothesis as the familiar

$$M_{\mu,\sigma^2}^{ODP} : E(Y_{ij}) = \exp(\mu_{ij}), \quad \frac{\text{var}(Y_{ij})}{E(Y_{ij})} = \sigma^2.$$

The model under this hypothesis in a run-off triangle replicates the chain-ladder. Thus, M_{μ,σ^2}^{ODP} is the model we would ideally like to use.

We can motivate the assumption of an over-dispersed infinitely divisible distribution for the aggregate incremental claims by a compound Poisson story. We can think of the aggregate incremental claims Y as a random Poisson number of claims N each with an independent random claim amount X so the $Y = \sum_{s=1}^{N} X_s$ are compound Poisson. Compound Poisson distributions are infinitely divisible. The over-dispersion σ^2 simplifies to $E(X^2)/E(X)$. Thus, it is common across the data set if the same is true for the claim amount distribution. If the claim amount distribution varies across sub-samples, so does the over-dispersion.

3.4.2 Estimation

With the model and hypotheses in place, we move on to estimation. The estimators match those in Harnau & Nielsen (2017). Means are estimated by Poisson quasi-likelihood, over-dispersion parameters by Poisson log-likelihood ratios. By estimating means by Poisson

quasi-likelihood, we match the classic arithmetic chain-ladder forecasts in run-off triangles as Kremer (1985) showed. Just as the results for the log-normal model, the theory in this section is invariant to the identification scheme since the statistics are functions of the identified linear predictors. We choose the same identification scheme as in the log-normal model, matching the notation.

3.4.2.1 Estimation in Unrestricted Model M^{ODP}

We estimate linear predictors by Poisson quasi-likelihood

$$\hat{\mu}_{ij,\ell}^{ODP} = x'_{ij,\ell}\hat{\xi}_\ell^{ODP} \quad \text{where} \quad \hat{\xi}_\ell^{ODP} = \underset{\xi_\ell \in R^{p_\ell}}{\arg\max} \sum_{ij \in \mathcal{I}_\ell} \{Y_{ij,\ell}(x'_{ij,\ell}\xi_\ell) - \exp(x'_{ij,\ell}\xi_\ell)\}.$$

The over-dispersion parameter estimators are Poisson quasi log-likelihood ratios; looking ahead, this is justified by their asymptotic χ^2 distribution. Specifically, the estimator for σ_ℓ^2 is the Poisson deviance divided by the degrees of freedom. The deviance is the log-likelihood ratio against a saturated model with as many parameters as observations and perfect fit. Specifically for deviance D_ℓ, the estimator for σ_ℓ^2 is

$$\hat{\sigma}_\ell^{2,ODP} = \frac{D_\ell}{df_\ell} \quad \text{where} \quad D_\ell = 2 \sum_{ij \in \mathcal{I}_\ell} Y_{ij,\ell}\{\log(Y_{ij,\ell}) - \hat{\mu}_{ij,\ell}^{ODP}\}.$$

3.4.2.2 Estimation with Common Variances in $M_{\sigma^2}^{ODP}$

In the model with common variances we can, as in the log-normal model, compute estimators from those for the unrestricted model. Estimators for the linear predictors $\mu_{ij,\ell}$ are unchanged. The estimator for the over-dispersion parameters is the degree of freedom weighted average

$$\bar{\sigma}^{2,ODP} = \sum_{\ell=1}^{m} \frac{df_\ell}{df_\cdot} \hat{\sigma}_\ell^{2,ODP} = \frac{D_\cdot}{df_\cdot}$$

where $D_\cdot = \sum_{\ell=1}^m D_\ell$ and, as before, $df_\cdot = \sum_{\ell=1}^m df_\ell$.

3.4.2.3 Estimation with Common Variances and Linear Predictors in M_{μ,σ^2}^{ODP}

In the model with common linear predictors and over-dispersion parameters, we estimate over the full sample. Dropping the ℓ subscript,

$$\hat{\mu}_{ij}^{ODP} = x'_{ij}\hat{\xi}^{ODP} \quad \text{where} \quad \hat{\xi}^{ODP} = \underset{\xi \in R^p}{\arg\max} \sum_{ij \in \mathcal{I}} \{Y_{ij}(x'_{ij}\xi) - \exp(x'_{ij}\xi)\}$$

and

$$\hat{\sigma}^{2,ODP} = \frac{D}{df} \quad \text{where} \quad D = 2\sum_{ij\in\mathcal{I}} Y_{ij}\{\log(Y_{ij}) - \hat{\mu}_{ij}^{ODP}\}.$$

3.4.3 Sampling Scheme

The asymptotic theory requires a sampling scheme. The challenge is that the number of observations n grows with the number of parameters: new accident or development years would demand their own parameters. Harnau & Nielsen (2017) circumvent this problem. They propose a sampling scheme that requires the means of the cells in the data set $\mathcal{I}$ to grow proportionally. This is reminiscent of multinomial sampling as used, for example, by Martínez Miranda et al. (2015) in a Poisson model. Crucially, the number of observations n, thus the number of parameters, remains fixed. We adopt their sampling scheme and motivate it by a compound Poisson example.

The sampling scheme stipulates that the aggregate mean $E(Y_{\cdot\cdot}) = E(\sum_{ij\in\mathcal{I}} Y_{ij})$ over the array grows in such a way that the skewness $skew(Y_{ij,\ell})$ vanishes while keeping the frequencies $E(Y_{ij,\ell})/E(Y_{\cdot\cdot})$ fixed. The requirement on the skewness is somewhat unconventional and is motivated by a limit theorem proved by Harnau & Nielsen (2017, Theorem 1).

For intuitive appeal, the skewness in the compound Poisson example from §3.4.1 vanishes as the expected number of claims grows. More precisely, considering once again aggregate incremental claims $Y = \sum_{s=1}^{N} X_s$ with N being the random Poisson number of claims and X_s the random claim amounts, the skewness of Y vanishes if the mean of the number of claims N grows for a fixed claim amount distribution X_s.

3.4.4 Asymptotic Testing for Common Over-Dispersion

Having set up the model and sampling scheme, we turn to the asymptotic theory. We show that the asymptotic distribution of the Bartlett test and the two-sample F-test for common over-dispersion match the finite sample distribution of the test for common log data variance in the log-normal model. We can justify a χ^2 approximation to the distribution of the Bartlett test through a sequential asymptotic argument.

To test for common over-dispersion across sub-samples in the over-dispersed Poisson model, we can proceed just as is the log-normal model. This is because the asymptotic distribution of $\hat{\sigma}_\ell^{2,ODP}$ matches the exact distribution of $\hat{\sigma}_\ell^{2,LN}$ in the log-normal model (Harnau & Nielsen 2017, Lemma 1):

$$\hat{\sigma}_\ell^{2,ODP} \xrightarrow{D} \frac{\sigma_\ell^2}{df_\ell}\chi^2_{df_\ell}. \tag{3.8}$$

Therefore, to test H_{σ^2}, we merely replace the estimators from the log-normal model with the over-dispersion estimators and compute

$$LR^{ODP} = df_\cdot \log(\bar{\sigma}^{2,ODP}) - \sum_{\ell=1}^{m} df_\ell \log(\hat{\sigma}_\ell^{2,ODP}). \tag{3.9}$$

Since the theory for the variance tests in the log-normal model hinged on the distribution of the log data variance estimators alone, we can immediately jump to the main result of the paper.

Theorem 3.2. *In the over-dispersed Poisson model with common over-dispersion $M_{\sigma^2}^{ODP}$ of §3.4.1 and 3.4.3, LR^{ODP} converges to the Bartlett distribution* $\mathrm{Ba}(df_1, \ldots, df_\ell)$ *from* (3.4). *Further, the F-statistic $F_{\sigma^2}^{ODP} = \hat{\sigma}_2^{2,ODP}/\hat{\sigma}_1^{2,ODP}$ is asymptotically* F_{df_2,df_1} *distributed.*

In §3.6.3 below, we show that finite sample approximations to the asymptotic results in Theorem 3.2 work well. To make the χ^2 approximation for the Bartlett test work we can use a sequential asymptotic argument. In the log-normal model, the χ^2 approximation followed through large degree of freedom asymptotics. In the over-dispersed Poisson model, we first let the aggregate mean $E(Y_{\cdot\cdot})$ grow such that LR^{ODP}/C is distributed Ba. Then, we can increase the sub-sample dimension and thus the degrees of freedom so Ba becomes χ^2. Then, under H_{σ^2}, we can expect

$$B^{ODP} = \frac{LR^{ODP}}{C} \overset{D}{\approx} \chi^2_{m-1}. \tag{3.10}$$

A simultaneous double asymptotic theory for large $E(Y_{\cdot\cdot})$ and degrees of freedom would have to wrestle with the complication that the number of mean parameters grows with the dimension of the sub-samples. Hence, such a generalization is by no means trivial and the simulations in §3.6 make it seem unnecessary.

3.4.5 Asymptotic Testing for Common Linear Predictors

We show how to F-test for common mean parameters. We also prove asymptotic independence of this F-test and tests for common over-dispersion.

As in the log-normal model, we can use a sequential testing strategy, first testing for H_{σ^2}, then for H_{μ,σ^2}. Harnau & Nielsen (2017, Theorem 4) showed that under H_{μ,σ^2} and thus in M^{ODP}_{μ,σ^2}, an F-statistic has an asymptotic F-distribution:

$$F^{ODP}_{\mu} = \frac{(D - D_.)/(df - df_.)}{D_./df_.} \xrightarrow{D} \mathrm{F}_{df-df_.,df_.}. \tag{3.11}$$

Thus, we can use a (one-sided) F-test to test for a reduction from $M^{ODP}_{\sigma^2}$ to M^{ODP}_{μ,σ^2}. If we compare to the test in the log-normal model, we simply replaced the residual sum of squares RSS with Poisson quasi-deviances D. The difference is that the F-distribution is now asymptotic and not exact.

To justify a sequential testing approach, it is useful to show that the test is independent of the Bartlett and F-test for common dispersion, just as it was for the log-normal model.

Lemma 3.1. *In the over-dispersed Poisson model M^{ODP}_{μ,σ^2} of §3.4.1 and §3.4.3, F^{ODP}_{μ} is asymptotically independent of $F^{ODP}_{\sigma^2}$ and LR^{ODP}.*

Therefore, under H_{μ,σ^2} the distribution of F^{ODP}_{μ} is asymptotically unaffected by conditioning on non-rejection of tests for common over-dispersion. We confirm in simulations below that this result holds approximately in finite samples. Hence, size control is not an issue in sequential testing, just as for the log-normal model.

3.5 Empirical Applications

To illustrate implementation of the theory we take it to the data. A run-off triangle first analyzed by Verrall et al. (2010) is appealing for a log-normal application: Kuang et al. (2015) raised the question of misspecification for this model on this data. As an over-dispersed Poisson example, we chose the data set by Taylor & Ashe (1983) in Table 3.1 which has become a sort of benchmark data set for this model. Verrall (1991), England & Verrall (1999), and Pinheiro et al. (2003) all use this data, to name but a few. Finally,

the data by Barnett & Zehnwirth (2000) seem to require a calendar effect for modeling; we take this opportunity to demonstrate that we can easily test for specification in a model with an extended chain-ladder structure that includes a calendar effect. We use the R (R Core Team 2016) package `apc` (Nielsen 2015) for the empirical applications and simulations below.

3.5.1 Log-Normal Chain-Ladder

Kuang et al. (2015) employ a log-normal chain-ladder model for data in a run-off triangle first analyzed by Verrall et al. (2010). They remark that the largest residuals congregate within the first five accident years, indicating a potential misspecification. Verrall et al. (2010) used the data to illustrate a model that makes use of the number of reported claims that is also available; we do not make use of this information. The data relate to a portfolio of motor policies from the insurer Royal & Sun Alliance. We show this triangle in Table 3.3 in the appendix of this chapter.

We take the remarks about misspecification by Kuang et al. (2015) as an opportunity to apply the specification tests for common log data variance and mean parameters. To do so, we first specify the sub-samples. Then, we set up the unrestricted model and test the hypotheses. Figure 3.2 summarizes the results.

Figure 3.2a shows how we split the data $\mathcal{I}$, a run-off triangle with ten accident and development years. We split into two sub-samples: $\mathcal{I}_1$ contains the first five and $\mathcal{I}_2$ the last five accident years. Choosing this specific structure seems intuitive given Kuang et al. (2015) remarks about the location of large residuals.

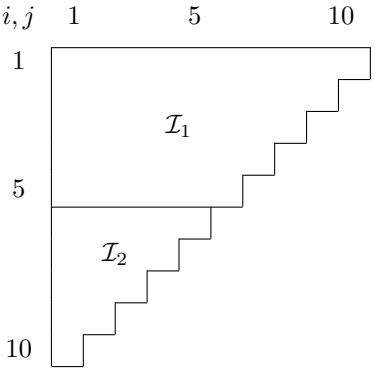
Given the sub-samples, we specify the unrestricted independent log-normal model

$$M^{LN} : \log(Y_{ij,\ell}) \stackrel{D}{=} N(\alpha_{i,\ell} + \beta_{j,\ell} + \delta_\ell, \sigma^2_\ell).$$

We first consider the hypothesis $H_{\sigma^2} : \sigma^2_1 = \sigma^2_2$ for a reduction to

$$M^{LN}_{\sigma^2} : \log(Y_{ij,\ell}) \stackrel{D}{=} N(\alpha_{i,\ell} + \beta_{j,\ell} + \delta_\ell, \sigma^2).$$

Figure 3.2b shows the relevant estimates and test results. Since we have just two sub-samples, we can test the hypothesis either with a Bartlett test or an F-test for common



(a) Sub-sample structure

	$\hat{\sigma}^{2,LN}_\ell$	df_ℓ	$\bar{\sigma}^{2,LN}$	$df_\cdot$	$\hat{\sigma}^{2,LN}$	df
$\mathcal{I}_1$	0.095	26				
$\mathcal{I}_2$	0.027	6				
$\mathcal{I}$			0.082	32	0.075	36

$B^{LN} = 2.79$ $\quad p_{\chi^2_1} = 0.09$

$F^{LN}_{\sigma^2} = 0.28$ $\quad p_{F_{6,26}} = 0.12$

$F^{LN}_{\mu} = 0.24$ $\quad p_{F_{4,32}} = 0.91$

(b) Estimation and test results

Figure 3.2: Log-normal chain-ladder model for Verrall et al. (2010) data. Sub-sample structure shown in (**a**), estimation and test results in (**b**).

variances. The two test give a rather similar indication. The Bartlett statistic B^{LN} has a χ^2 p-value of 0.09 and the F-statistic $F^{LN}_{\sigma^2}$ a two-sided F p-value of 0.12.

If we take the variance test results as an indication not to reject H_{σ^2}, we can take $M^{LN}_{\sigma^2}$ as our primary model and test for H_{μ,σ^2}. That is, we test for a reduction to

$$M^{LN}_{\mu,\sigma^2} : \log(Y_{ij}) \stackrel{D}{=} N(\alpha_i + \beta_j + \delta, \sigma^2).$$

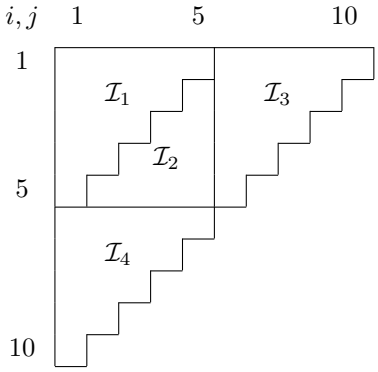
Based on the F-statistic F^{LN}_{μ}, we cannot reject this hypothesis with a p-value of 0.91. Thus, we do not find compelling evidence against a reduction to M^{LN}_{μ,σ^2}.

Alternatively, we could make use of the information that there is not just a discrepancy between the sub-samples when it comes to residuals, but that those in $\mathcal{I}_1$ are larger. With this information, we could alternatively have conducted a one-sided F-test for a one-sided hypothesis $H_{\sigma^2} : \sigma^2_1 \leq \sigma^2_2$. This test yields a p-value of 0.06, a much closer call. Note that we cannot evaluate one-sided hypotheses with a Bartlett test.

3.5.2 Over-Dispersed Poisson Chain-Ladder

The Taylor & Ashe (1983) data in Table 3.1 has served many times as an empirical application for over-dispersed Poisson chain-ladder models. Thus, it seems only appropriate to investigate the model specification. We summarize results in Figure 3.3.

Figure 3.3a shows the chosen sub-sample structure. We split the sample after the fifth accident, development, and calendar year into four sub-samples. Unlike in the case of the



(a) Sub-sample structure

	$\hat{\sigma}^{2,ODP}_\ell$	df_ℓ	$\bar{\sigma}^{2,ODP}$	$df_.$	$\hat{\sigma}^{2,ODP}$	df
$\mathcal{I}_1$	31903	6				
$\mathcal{I}_2$	168293	3				
$\mathcal{I}_3$	104493	6				
$\mathcal{I}_4$	17592	6				
$\mathcal{I}$			68038	21	52862	36

$B^{ODP} = 6.78$ $\quad p_{\chi^2_3} = 0.08$

$F^{ODP}_\mu = 0.46$ $\quad p_{F_{15,21}} = 0.93$

(b) Estimation and test results

Figure 3.3: Over-dispersed Poisson chain-ladder model for Taylor & Ashe (1983) data. Sub-sample structure shown in (**a**), estimation and test results in (**b**).

Verrall et al. (2010) data above, we do not have information indicating a specific sub-sample structure. While arbitrary, we find the chosen structure appealing because all sub-samples are run-off triangles themselves and of relatively similar size. Further, we hope that splits after each of the three time-scales increases our chances to find breaks. We point out that the specific sub-sample structure has no effect on the size of the tests if the hypothesis is true.

Figure 3.3b shows estimates and test results. The unrestricted model is the over-dispersed Poisson model discussed in §3.4.1 so that

$$M^{ODP}_{\sigma^2} : E(Y_{ij,\ell}) = \exp(\mu_{ij,\ell}), \quad \frac{\text{var}(Y_{ij,\ell})}{E(Y_{ij,\ell})} = \sigma^2_\ell.$$

Looking at evidence for varying over-dispersion, we test for H_{σ^2} with a Bartlett test. While we can see quite a bit of variation in the dispersion estimates, ranging from $\hat{\sigma}^{2,ODP}_4 = 17{,}592$ to $\hat{\sigma}^{2,ODP}_2 = 168{,}293$, the test does not convincingly reject the hypothesis with a p-value of 0.08. Even though relative deviations from the degree of freedom weighted average $\bar{\sigma}^{2,ODP} = 68{,}038$ are less stark, it seems to us that making a decision by eyeballing alone would be difficult in this case.

If the Bartlett test results convince us that a reduction to $M^{ODP}_{\sigma^2}$ is sensible, we can test for common linear predictors. Given an F-statistic of $F^{ODP}_\mu = 0.46$, we cannot reject this simplification with a p-value of 0.93.

Overall, the target over-dispersed Poisson model for the Taylor & Ashe (1983) data survives both misspecification tests at a 5% level for this sub-sample structure. Thus, we may be more confident now to model it with an over-dispersed Poisson chain-ladder model.

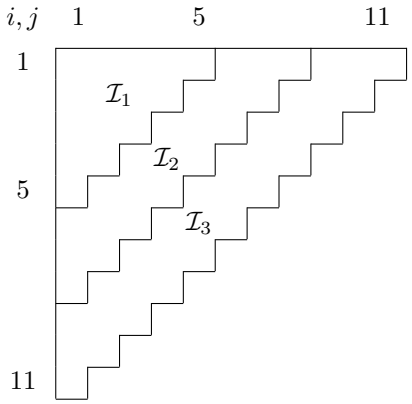
We could also opt to repeat the test for other sub-sample structures, adjusting the size to take into account that tests for different sub-sample structures on the same data are generally not independent. For example, retesting for the split into two sub-samples consider above and shown in Figure 3.1a. For this structure, a Bartlett test statistic of $B^{ODP} = 2.89$ yields a p-value of 0.09 and an F-test statistic of $F_{\mu}^{ODP} = 0.63$ a p-value of 0.64. Further, we can test for a split into three sub-samples after calendar years four and seven, similar to the structure in Figure 3.1b. For this structure, we get $B^{ODP} = 1.27$ with a p-value of 0.53 and $F_{\mu}^{ODP} = 1.84$ with a p-value of 0.11. Controlling the overall size of the thrice repeated sequential tests with a Bonferroni correction, we would reject if any p-value was below $5\%/3 \approx 0.017$. This is not the case so the model survives this battery of tests as well.

3.5.3 Log-Normal (Extended) Chain-Ladder

As a final empirical application, we look at a run-off triangle first considered by Barnett & Zehnwirth (2000). We show this data in Table 3.4 in the appendix of this chapter. These data are known to be modeled best with a predictor with not just accident and development, but also calendar effects. We look at a model with and without calendar effects. Barnett & Zehnwirth (2000) and also Kuang et al. (2011) consider a log-normal model for this data and we follow them in this choice. As before, we split the data, specify the model, and test for the hypotheses. The results are summarized in Figure 3.4.

Figure 3.4a shows the sub-sample structure we choose. Given the apparent need for calendar effects, we aim to maximize power for varying dispersion parameters along the same time dimension and split the run-off triangle, this time with eleven accident and development years, after periods five and eight into three sub-samples.

The top of Figure 3.4b shows estimation and test results for a model without calendar



(a) Sub-sample structure

	$\hat{\sigma}^{2,LN}_\ell$	df_ℓ	$\bar{\sigma}^{2,LN}$	$df_.$	$\hat{\sigma}^{2,LN}$	df
$\mathcal{I}_1$	0.0002	6				
$\mathcal{I}_2$	0.0013	6				
$\mathcal{I}_3$	0.0015	9				
$\mathcal{I}$			0.0011	21	0.0068	45

$B^{LN} = 6.06$ $\quad p_{\chi^2_2} = 0.05$

$F^{LN}_\mu = 11.20$ $\quad p_{F_{20,16}} = 0.00$

	$\hat{\sigma}^{2,LNe}_\ell$	df^e_ℓ	$\bar{\sigma}^{2,LNe}$	$df^e_.$	$\hat{\sigma}^{2,LNe}$	df^e
$\mathcal{I}_1$	0.0003	3				
$\mathcal{I}_2$	0.0013	5				
$\mathcal{I}_3$	0.0016	8				
$\mathcal{I}$			0.0013	16	0.0014	36

$B^{LNe} = 2.06$ $\quad p_{\chi^2_2} = 0.36$

$F^{LNe}_\mu = 1.13$ $\quad p_{F_{20,16}} = 0.41$

(b) Estimation and test results

Figure 3.4: Log-normal chain-ladder (LN) and extended chain-ladder (LNe) model for Barnett & Zehnwirth (2000) data. Sub-sample structure shown in (**a**), estimation and test results in (**b**).

effect. This model is given by

$$M^{LN} : \log(Y_{ij,\ell}) \stackrel{D}{=} N(\alpha_{i,\ell} + \beta_{j,\ell} + \delta_\ell, \sigma^2_\ell).$$

A Bartlett test for the hypothesis H_{σ^2} of common log data variances has a χ^2 p-value of (just under) 0.05. We may consider this as evidence against H_{σ^2}. For comparison with the model with calendar effect considered next, we still compute an F-test for the hypothesis H_{μ,σ^2}. We point out that this test is not strictly a test for common linear-predictors if we are not comfortable to accept $M^{LN}_{\sigma^2}$ as a model. The statistic $F^{LN}_\mu = 11.20$ has a 0.00 p-value so that we reject H_{μ,σ^2}. Thus, M^{LN}_{μ,σ^2} is not well specified.

At the bottom of Figure 3.4b we show results for a model with calendar effects γ for calendar years $k = i + j - 1$. The model is

$$M^{LNe} : \log(Y_{ij,\ell}) \stackrel{D}{=} N(\alpha_{i,\ell} + \beta_{j,\ell} + \gamma_{k,\ell} + \delta_\ell, \sigma^2_\ell).$$

The theory for specification tests is not affected by this change and thus still valid. A Bartlett test for H_{σ^2} in this model yields a χ^2 p-value of 0.36 so we may feel comfortable

to impose common log data variances and take $M_{\sigma^2}^{LNe}$ as given. An F-test for common linear predictors leaves us with a p-value of 0.41. Thus, reducing the model to M_{μ,σ^2}^{LNe} seems sensible. Therefore, we cannot reject the specification of the model with calendar effect.

If we directly compare the two models, we can see that the calendar effect has a substantial impact on the specification tests. While the model with calendar effect seems to be well specified, the model without this effect raises red flags for both a test for common variances and common linear predictors. The test for common linear predictors is much more strongly affected by dropping the calendar effect than the Bartlett test. This indicates that the shift in log data variances is smaller than that in linear predictors.

We look at the shift in linear predictor in two ways. First, we can directly test for for dropping the calendar effects from the well specified M_{μ,σ^2}^{LNe}. A standard F-test for the hypothesis $H_\gamma : \gamma_k = 0 \ \forall k$ yields a p-value of 0.00, consistent with the rejection of the model without calendar effects M_{μ,σ^2}^{LN} above.

Alternatively, we can test for a reduction from $M_{\sigma^2}^{LNe} : \log(Y_{ij,\ell}) \stackrel{D}{=} N(\alpha_{i,\ell} + \beta_{j,\ell} + \gamma_{k,\ell} + \delta_\ell, \sigma^2)$ to $M_{\sigma^2}^{LN}$, corresponding to the hypothesis $H_{\gamma_{k,\ell}} : \gamma_{k,\ell} = 0 \ \forall k, \ell$. This reduction allows for breaks in linear predictors between sub-samples. Interestingly, an F-test cannot reject $H_{\gamma_{k,\ell}}$ (p-value 0.92). As an intuition, we recall that the chain-ladder predictor without calendar effects can accommodate a constant trend in calendar years, but not deviations from that trend. Thus, allowing for separate sets of linear predictors on the sub-samples implicitly allows for three different calendar trends. While still less flexible than the model with an effect for each calendar year, this seems to be good enough. Note, however, that the Bartlett flags the reduction from M^{LN} to $M_{\sigma^2}^{LN}$ (but not from M^{LNe} to $M_{\sigma^2}^{LNe}$).

Overall, the analysis suggests that calendar effects are needed in this data set for two reasons for this sub-sample structure: to capture the structure of the linear predictors themselves, and, to a lesser extent, to achieve homogeneous variance across the log data.

We note that for this data, repeating the tests for different sub-samples structures does affect the results. Indeed, considering sub-samples similar to before, the specification of the log-normal extended chain-ladder model is rejected. Specifically, splitting the data into two sub-samples after the fifth accident year, a Bartlett test yields a p-value of 0.017 and an F-test a p-value of 0.004. Considering four sub-samples with splits after the fifth calendar

year, the fifth development year and the sixth accident year, the p-value of the Bartlett test is 0.03 and that of the F-test 0.05. Again controlling the size of the repeated tests with a Bonferroni correction, we would reject the null hypothesis if we can find a p-value below about 0.017. This is the case for the F-test and a knife-edge decision for the Bartlett test in the two sub-sample scenario. Thus, for this data we may want to consider a different model or at least be somewhat more skeptical of its results.

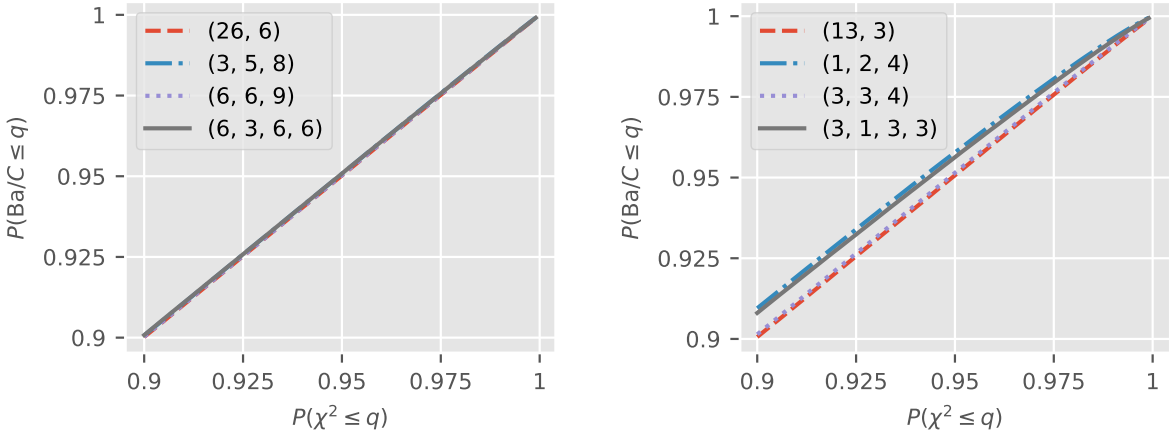
3.6 Simulations

The developed theory begs several questions that we answer in a simulation study. First, we argued that we can sidestep simulating critical values of the Bartlett distribution Ba and instead approximate these by a Bartlett corrected χ^2 critical value. We show that this works very well. Second, we compute power curves of Bartlett and F-test for common log data variances under several alternatives in a log-normal model to get a better understanding for the tests' behavior. Third, we show that an asymptotic approximation in an over-dispersed Poisson model resembles the asymptotic distribution closely, both under the null and the considered alternatives. Finally, we derived above that F-tests for common linear predictors in the over-dispersed Poisson model are asymptotically independent of tests for common over-dispersion. We confirm that the size of the former test seems unaffected by conditioning on the results of the latter, even in finite samples.

3.6.1 Performance of Bartlett test χ^2 Approximation

The theory tells us that the distribution of Ba/C, which is the exact distribution of the Bartlett statistic B^{LN} in the log-normal model, is close to a χ^2 for large degrees of freedom. We show that the approximation works very well for a range of degrees of freedom.

We draw realizations from the adjusted Bartlett distribution $\text{Ba}(df_1, \ldots, df_m)/C$ as follows. For $\ell = 1, \ldots, m$, we draw independent χ^2 distributed V_ℓ with df_ℓ degrees of freedom and compute $s_\ell = V_\ell/df_\ell$ and $\bar{s} = \sum_{\ell=1}^m df_\ell/df_\cdot s_\ell$. Then, $\{df_\cdot \log(\bar{s}) - \sum_{\ell=1}^m df_\ell \log(s_\ell)\}/C$ is Ba/C distributed.



(a) Degrees of freedom from empirical applications

(b) Half the empirical degrees of freedom

Figure 3.5: pp-plots for the adjusted Bartlett distribution Ba/C against χ^2 for varying degrees of freedom. (**a**) and (**b**) show results for degrees of freedom corresponding to the empirical applications and half those degrees of freedom, respectively.

	(13, 3)	(1, 2, 4)	(3, 3, 4)	(3, 1, 3, 3)	(26, 6)	(3, 5, 8)	(6, 6, 9)	(6, 3, 6, 6)
$\alpha = 10\%$	9.94	9.05	9.86	9.20	9.98	9.93	9.97	9.92
$\alpha = 5\%$	4.93	4.22	4.85	4.37	4.98	4.92	4.97	4.92
$\alpha = 1\%$	0.95	0.69	0.92	0.76	0.99	0.95	0.98	0.96

Table 3.2: $P(\text{Ba}/C > c_\alpha)$ where c_α is the χ^2 α critical value. Results are in %. Degrees of freedom shown as $(df_1, \ldots, df_m)$.

Figure 3.5a shows the upper 10% probability spectrum of a pp-plot for the adjusted Bartlett distribution $\text{Ba}(df_1, \ldots, df_m)/C$ against a χ^2_{m-1}. We show plots for the tuples $(26, 6)$, $(3, 5, 8)$, $(6, 6, 9)$, and $(6, 3, 6, 6)$ encountered in the empirical applications above. The plots are based on 10^7 draws for each tuple. The plots seem indistinguishable from the 45-degree line, even though we zoomed in to the upper 10% of the spectrum.

Figure 3.5b is constructed in the same way as Figure 3.5a, except the degrees of freedom are halved and rounded down. Now, we can see some deviations from the 45-degree line. As expected, we can see convergence to the 45-degree line as the degrees of freedom increase.

In Table 3.2, we take a closer look at the approximation at $\alpha = 1\%, 5\%, 10\%$ critical values c_α of a χ^2_{m-1} specifically. The table shows $P(\text{Ba}/C > c_\alpha)$, corresponding to the true size of a Bartlett test in a log-normal model if we use the χ^2 approximation rather than simulated critical values. While we can see some differences for some of the halved

critical values, we would argue that the approximation for the degree of freedom tuples from the empirical applications is so good that using it is reasonable and should not affect the modeling decision.

3.6.2 Rejection Frequencies of Tests for Common Variance in Log-Normal Model

As a supplement to the behavior of the tests for common log data variance under the null hypothesis in the log-normal model given in §3.3.3, we now also take a look at power. We simulate the three sub-sample structures from the empirical applications and consider rejection frequencies of the tests used in the corresponding applications. We find that the Bartlett and F-test for common variance have very similar power, at least in this simulation. Further, we see that the power does not necessarily decrease with the number of sub-samples.

For the sub-sample structures from the empirical applications (see Figure 3.1), we simulate $M^{LN} : \log(Y_{ij,\ell}) \stackrel{D}{=} N(\mu_{ij,\ell}, \sigma^2_\ell)$. Thus, we simulate for $m = 2, 3, 4$ sub-samples. Before specifying the parameter values, we point out that the distribution of tests in this model depends only on ratios σ^2_s/σ^2_t, the degrees of freedom df_ℓ, and the number of sub-samples m. To see this, we first re-write

$$LR^{LN} = df_\cdot \log(\bar{\sigma}^{2,LN}) - \sum_{\ell=1}^{m} df_\ell \log(\hat{\sigma}^{2,LN}_\ell) = \sum_{\ell=1}^{m} df_\ell \log\left\{ \frac{df_\ell}{df_\cdot} \left(\sum_{s=1}^{\ell} \frac{RSS_s}{RSS_\ell} \right) \right\}. \quad (3.12)$$

Now, under M^{LN}, $RSS_\ell \stackrel{D}{=} \sigma^2_\ell \chi^2_{df_\ell}$ independently. Thus, the distribution of LR^{LN} is invariant to common changes in levels of σ^2_ℓ as well as to $\mu_{ij,\ell}$. Therefore, we can normalize the smallest σ^2_ℓ to unity and set $\mu_{ij,\ell} = 0$ without loss of generality. The distribution of the F-statistic $F^{LN}_{\sigma^2}$ shares these properties.

For each sub-sample scenario, we consider a range of values for the log data variance ratios σ^2_s/σ^2_t. For $m > 2$ sub-samples there is more than one ratio such that we cannot effectively visualize all combinations. We thus consider the following special case. For each sub-sample structure, we the compute the spacing of the estimates from the corresponding empirical application. That is, we order the empirical estimates $\hat{\sigma}^2_{(1)} < \cdots < \hat{\sigma}^2_{(m)}$ and

compute the m spacing-coefficients $x_\ell = (\hat{\sigma}^2_\ell - \hat{\sigma}^2_{(1)})/(\hat{\sigma}^2_{(m)} - \hat{\sigma}^2_{(1)})$. We note that $x_{(1)} = 0$ and $x_{(m)} = 1$. The spacings $(x_1, \ldots, x_m)$ in the empirical examples are $(1, 0)$ (Verrall et al. 2010), $(0, 0.76, 1)$ (Barnett & Zehnwirth 2000), and $(0.09, 1, 0.58, 0)$ (Taylor & Ashe 1983). The log data variance for the ℓ-th subset is then

$$\sigma^2_\ell = \sigma^2_{(1)} + x_\ell(\sigma^2_{(m)} - \sigma^2_{(1)}). \tag{3.13}$$

To trace out power curves, we vary the largest ratio $\sigma^2_{(m)}/\sigma^2_{(1)}$ from one, corresponding to $H_{\sigma^2} : \sigma^2_\ell = \sigma^2$, to twenty in 0.5 increments. As noted above, we can set $\sigma^2_{(1)} = 1$ without loss of generality. For each degree of freedom scenario and for each ratio $\sigma^2_{(m)}/\sigma^2_{(1)}$, we draw 10^6 sub-samples.

For each draw, we compute the test statistics used in the corresponding empirical application, $F^{LN}_{\sigma^2}$ as in (3.6) and B^{LN} as in (3.5). We note that for $m = 3$, we compute only the Bartlett test statistic for the model with calendar effect B^{LNe} to make the plot less cluttered. Thus, the degrees of freedom for $\hat{\sigma}^{2,LN}_\ell$ in the three scenarios are $(26, 6)$, $(3, 5, 8)$ and $(6, 3, 6, 6)$. We use χ^2 critical values for the Bartlett tests.

Figure 3.6 shows rejection frequencies at 5% critical values. We can see that all tests have the right size under H_{σ^2}, that is for $\sigma^2_{(m)}/\sigma^2_{(1)} = 1$. The power of two-sided F-test and Bartlett test in the two sub-sample scenario is very similar with a slight advantage for the Bartlett test. Thus, the choice between the two test may mostly depend on taste. Comparing Bartlett tests across scenarios, we see that the power for $m = 4$ sub-samples is larger than that for $m = 3$ sub-samples. Thus, fewer sub-samples do not necessarily imply higher power. Intuition comes from the degree of freedom weighting. For $m = 3$ sub-samples, if we drop the variance with the smallest degree of freedom the larger two variances are relatively homogeneous. Meanwhile, for $m = 4$ sub-samples there is still plenty of variation left among the largest three variances. Thus, since the test attributes more weight to the better estimates with higher degrees of freedom, the scenario with $m = 3$ sub-samples is a rather tough case.

We indicated the $\sigma^2_{(m)}/\sigma^2_{(1)}$ ratios we found in the individual empirical applications by vertical lines. We recall that the spacing of intermediate variances is taken from the empirical applications. Therefore, suppose that the empirical estimates are the truth such

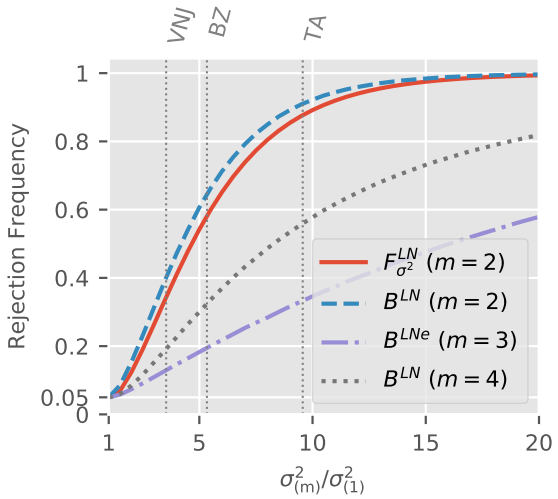


Figure 3.6: Power curves for log-normal dispersion tests based on sub-sample structures from empirical applications. Empirical maximum to minimum ratios indicated by horizontal lines. BZ is short for Barnett & Zehnwirth (2000), VNJ for Verrall et al. (2010), and TA for Taylor & Ashe (1983).

that H_{σ^2} is violated. Then we can read of the power against this scenario directly from the plot. For example, in the application to the Verrall et al. (2010) data, the F-test would have a power of about 35% while the Bartlett test power would be closer to 40%.

3.6.3 Performance of Over-Dispersed Poisson Model Asymptotics

The theoretical results for the over-dispersed Poisson model are asymptotic, rather than exact as in the log-normal model. We show that an asymptotic approximation works well. Tests for common over-dispersion have the right size under the null. The power under the alternative in finite samples is close to the asymptotic power. Further, F-tests for common linear predictors conditional on non-rejection of over-dispersion tests are very close to F distributed in finite samples.

3.6.3.1 Rejection Frequencies of Tests for Common Over-Dispersion

We can use the rejection frequencies from the log-normal simulations as a benchmark for those in the over-dispersed Poisson model. To see this we recall that as the overall mean $E(Y_{..}) \to \infty$, the over-dispersion estimator $\hat{\sigma}^{2,ODP}_{\ell} \xrightarrow{D} \sigma^2_{\ell}\chi^2_{df_\ell}/df_\ell$ in the over-dispersed Poisson model M^{ODP}. This matches the exact distribution of $\hat{\sigma}^{2,LN}_{\ell}$ in M^{LN}. Thus, asymp-

totically, the distribution of LR^{ODP} in M^{ODP} and LR^{LN} in M^{LN} are identical for identical ratios σ_s^2/σ_t^2. The same holds for $F_{\sigma^2}^{ODP}$ and $F_{\sigma^2}^{LN}$.

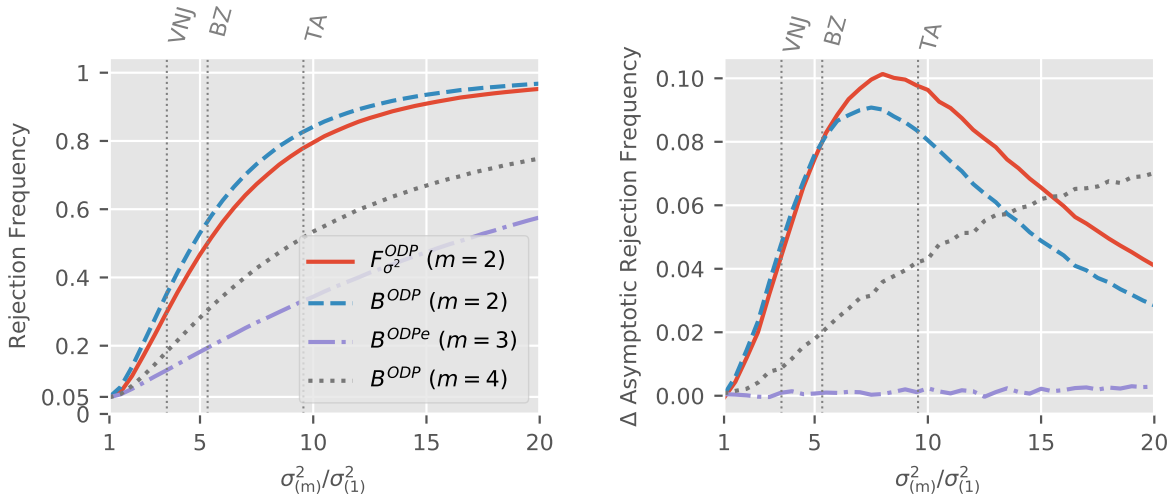
We simulate for the same three sub-sample structures as in the log-normal simulations. For the simulation design, we set-up an unrestricted model M^{ODP} that satisfies the assumptions in §3.4.1. For the distribution of the cells we choose compound Poisson-gamma so $Y_{ij,\ell} = \sum_{s=1}^{C_{ij,\ell}} X_{s,\ell}$ where $C_{ij,\ell} \overset{D}{=} \text{Poisson}\{\exp(\mu_{ij,\ell})\}$ independent of the i.i.d. gamma distributed $X_{s,\ell}$ with scale $\sigma_\ell^2 - 1$ and shape $(\sigma_\ell^2 - 1)^{-1}$. We note that the parametrization for the linear predictors $\mu_{ij,\ell}$ and the level of the over-dispersion σ_ℓ^2 matters in finite samples. This is in contrast to the log-normal model. The reason is that the finite sample distribution of $\hat{\sigma}_\ell^{2,ODP}$ in M^{ODP} is generally not $\sigma_\ell^2 \chi_{df_\ell}^2/df_\ell$. Thus, for each considered scenario, we set the linear predictors $\mu_{ij,\ell}$ to the estimates $\hat{\mu}_{ij,\ell}^{ODP}$ from the data in the corresponding empirical application. Similarly, we set the smallest over-dispersion $\sigma_{(1)}^2 = \hat{\sigma}_{(1)}^{2,ODP}$.

We again vary the ratios $\sigma_{(m)}^2/\sigma_{(1)}^2$ from one to twenty, using the exact same spacing x_ℓ from (3.13) in the log-normal simulations so $\sigma_\ell^2 = \sigma_{(1)}^2 + x_\ell(\sigma_\ell^2 - \sigma_{(1)}^2)$. The only difference is that now $\sigma_{(1)}^2 = \hat{\sigma}_{(1)}^{2,ODP}$. Therefore, asymptotically, the power for a common $\sigma_{(m)}^2/\sigma_{(1)}^2$ is identical in over-dispersed Poisson and log-normal models. We draw 10^6 sub-samples for each over-dispersion ratio $\sigma_{(m)}^2/\sigma_{(1)}^2$ and sub-sample structure.

Figure 3.7a shows the rejection frequencies at 5% critical values for the four test statistics from the empirical applications as in the log-normal model but now computed based on $\hat{\sigma}_\ell^{2,ODP}$.

For $\sigma_{(m)}^2/\sigma_{(1)}^2 = 1$ we are under the null; we can see that the rejection frequencies are very close to 5% so the tests have the correct size. Under the alternative where $\sigma_{(m)}^2/\sigma_{(1)}^2 > 1$, the ordering of the rejection frequencies matches that in the log-normal simulations (Figure 3.6). Generally, the plot is reassuringly reminiscent of its equivalent in the log-normal simulations.

Figure 3.7b shows the gap to the asymptotic rejection frequencies that arises in the finite sample simulations. The simulation set-up implies that this is the difference between rejection frequencies in log-normal and over-dispersed Poisson simulations. Thus, the plots shows the impact of the asymptotic approximation in the over-dispersed model. Since the difference under the alternative is positive throughout, the power in the over-dispersed



(a) Rejection frequencies

(b) Gap to asymptotic rejection frequencies

Figure 3.7: Power gap for log-normal dispersion tests based on sub-sample structures from empirical applications. Empirical maximum to minimum ratios indicated by horizontal lines. Rejection frequencies shown in (**a**), gap to asymptotic rejection frequencies in (**b**).

model is lower than in the log-normal model. We next interpret the plots under the alternative in turn for the three sub-sample scenarios.

For $m = 2$, the power gap of Bartlett and F-test initially increases with $\sigma^2_{(m)}/\sigma^2_{(1)}$, hitting $10pp$ (percentage points)for the F-test at $\sigma^2_{(\mathrm{m})}/\sigma^2_{(1)} \approx 7.5$, before it decreases. The initial increase relates to the asymptotic theory by Harnau & Nielsen (2017) which assumes fixed dispersion parameters. Since we keep $\sigma^2_{(1)}$ constant, the remaining dispersion parameters grow with the ratio. Thus, we would require larger cell-means to achieve the same asymptotic approximation quality. The later decrease reflects the upper bound of one for the power: even as the asymptotic approximation becomes worse, the difference between dispersion parameters becomes so large that it is easily caught. For $m = 4$, the power gap is increasing throughout the considered range for $\sigma^2_{(\mathrm{m})}/\sigma^2_{(1)}$. The intuition for the increase again comes from the asymptotic theory. We do not see a decrease since the power is still quite far from unity, staying below 80% even for the largest maximum to minimum ratio. Meanwhile, for $m = 3$, the power gap is essentially zero so that the finite sample power matches the asymptotic power. The intuition for this follows because the dispersion to mean ratio is small. As a rough indication, dividing the largest considered dispersion

$20 \cdot \hat{\sigma}^{2,ODP}_{(1)}$ by the mean over all cells $n^{-1} \sum_{ij} Y_{ij}$ yields 0.8% for the Barnett & Zehnwirth (2000) simulations compared with 70% and 56% for the Verrall et al. (2010) and Taylor & Ashe (1983) simulations, respectively.

We again indicate the power at the particular alternative generated by taking the estimates in the empirical applications as true values by vertical lines. Figure 3.7b shows that for these alternatives, the power for all asymptotic approximations is within 5*pp* of their asymptotic power.

3.6.3.2 Independence of Test for Common Linear Predictors

We move on to evaluate the quality of a finite sample approximation to the asymptotic independence in Lemma 3.1. Specifically, we consider the finite sample distribution of F^{ODP}_{μ} as in (3.11) given that a tests for common over-dispersion did not reject. Arguably, this is the most interesting case since it matches the natural order of the two specification tests.

We simulate under the null H_{μ,σ^2}, that is for a model with common linear predictors and over-dispersion M^{ODP}_{μ,σ^2}. As before, cells Y_{ij} are compound Poisson-gamma. We consider three scenarios, setting the parameters to the estimates for M^{ODP}_{μ,σ^2} in the three empirical examples. We draws 10^6 triangles per scenario.

For each draw, we compute tests based on the sub-sample structure of the corresponding empirical application. We first conduct a Bartlett test for H_{σ^2} at 5% critical values. If we do not reject H_{σ^2} based on this test, we keep the triangle, otherwise we throw it out. Since we simulate under the null hypothesis, we thus keep about 95% of the draws. Only for the draws we keep do we compute the F-statistic for common linear predictors F^{ODP}_{μ}.

Figure 3.8 shows a pp-plot for the F^{ODP}_{μ} against $\mathrm{F}_{df-df_\cdot,df_\cdot}$ for the triangles that survived Bartlett testing. To be able to tell a difference from the 45-degree line, we limit our attention to the upper 10% of the probability spectrum since. This is also the most interesting range for testing. Even in this spectrum, each plot is very close to the 45-degree line. Therefore, under H_{μ,σ^2}, we can be reassured that an F-test for common linear predictors has the correct size in finite samples even if we apply it only conditionally on non-rejection of a test for common over-dispersion.

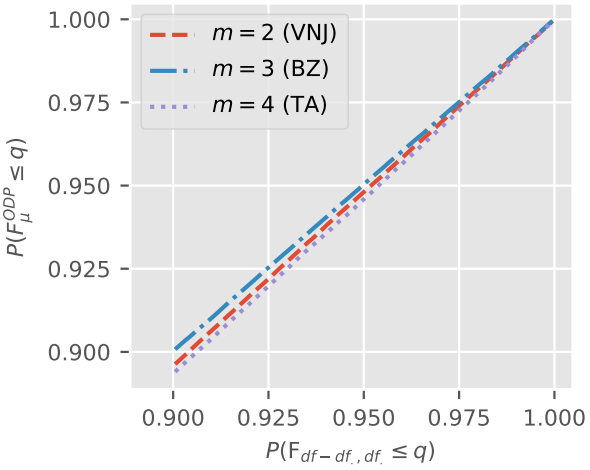


Figure 3.8: Distribution of F_{μ}^{ODP} conditional on non-rejection of a 5% Bartlett test.

3.6.4 Remark

We note that all simulations are for tests that consider the correct sub-sample structure under the alternative. Of course, this does not seem realistic in applications. However, for tests computed on a given sub-sample structure, it appears we would generally be able to choose a true, different, sub-sample structure against which the tests would at best have limited power. For example, say we compute the tests on the two sub-samples with a split after the fifth accident year in Figure 3.1a while really there are three sub-samples with an additional split after the fifth development year. Then, we could choose parameterizations for the three true true sub-samples to balance out the variation between the two incorrectly chosen sub-samples, thus minimizing power. Therefore, it seems to us that such simulation results would be almost entirely driven by our chosen parametrization and provide little insight beyond that. We believe the real answer to this problem must come from a theory that is agnostic to the sub-sample structure as discussed below. However, we stress again that the size of the tests under the null hypothesis is not affected by the chosen sub-sample structure.

3.7 Discussion

Some questions are left open for future research. For example, it is not clear how to best choose the sub-sample structure and the number of sub-samples. Further, the question arises whether we can somehow select between the over-dispersed Poisson and log-normal model. Finally, a misspecification test for independence of the cells would be a useful addition to the modeling toolkit.

So far, we chose the sub-sample structures somewhat arbitrarily if potentially informed by prior knowledge of the data. While the size of the tests under the null is not affected by the sub-sample structure, the power of the tests under the alternative is affected both by the chosen number of sub-samples and their structure. In applications, the expert may consider choosing a range of sub-samples structures and conducting tests for each, adjusting the size based on the number of tests to account for multiple testing as shown in the empirical applications. For future research, it would be useful to derive a theory that is agnostic to the number of sub-samples and their structure while still directly controlling size. It might be fruitful to look for ideas in time-series econometrics which has been concerned with tests for parameter breaks for a long time. In this literature, Chow (1960) had proposed a test for parameter breaks that required knowledge of the breakpoint. By now, there are several test available that are agnostic with respect to the number of breaks, related to the number of sub-samples in our problem, and the position of breaks, akin to the sub-sample structure. Examples include Andrews' test (Andrews 1993), generalizations of Chow tests (Nielsen & Whitby 2015), and indicator saturation (Hendry 1999). However, these tests are designed for data with a single time-scale and results are generally based on long time-series. In contrast, we are confronted with data with three interlinked time-scales and the arrays are often small with a large number of parameters that is growing with the array size. Thus, the known results do not carry over and a it appears that a new theory is needed.

Since we have seen two models in this paper, log-normal and over-dispersed Poisson, a natural question is when we should choose which model. As we have seen, the log-normal model assumes a fixed standard deviation to mean ratio while the over-dispersed Poisson model considers the variance to mean ratio to be fixed. Making use of recent results for

generalized log-normal models by Kuang & Nielsen (2018), a class of models that includes the log-normal but is more general, Harnau (2018*a*) proposes a test to distinguish between (generalized) log-normal and over-dispersed Poisson models based on this discrepancy.

Finally, a misspecification test for the assumption that the cells in the array are independent would be useful. This is an assumption that both the log-normal and the over-dispersed Poisson model impose. In contrast, the "distribution free" model by Mack (1993) relaxes this somewhat, assuming independence only across accident years.

Appendix

Proof of Theorem 3.1

The proof relies on two properties. First, $RSS - RSS_{\cdot}$, the numerator of F_{μ}^{LN}, reduces to a comparison of least squares fitted log-means $\sum_{\ell=1}^{m}\sum_{ij\in\mathcal{I}_\ell}(\hat{\mu}_{ij}^{LN} - \hat{\mu}_{ij,\ell}^{LN})^2$, and is therefore, in the Gaussian framework at hand, independent of the residual sum of squares $RSS_1, \ldots, RSS_m$.

Second, the denominator of F_{μ}^{LN}, the aggregated residual sum of squares $RSS_{\cdot}$ and the relative contributions $\pi_1, \ldots, \pi_m$ for $\pi_\ell = RSS_\ell/RSS_{\cdot}$ are mutually independent. To see this, we first recall that under the hypothesis $RSS_1, \ldots, RSS_m$ are independent $\sigma^2\chi^2$. The proof is unaffected by setting $\sigma^2 = 1$. Thus, let $X_1, \ldots X_m$ be independent χ^2. Recall that the sum of independent χ^2's is χ^2. Let $S_\ell = \sum_{s=1}^{\ell} X_s$ and $V_\ell = S_{\ell-1}/S_\ell$. We note that we can map $V_2, \ldots, V_m$ to the frequencies $X_1/S_m, \ldots, X_m/S_m$. For the special case with $m = 2$, Johnson et al. (1995*b*, p. 212) note the independence of S_2 and V_2. The general case for independence of S_m and $V_2, \ldots, V_m$ can be proved by induction. Here, we partially replicate the (originally Danish) argument from Andersson & Jensen (1987, p. 180). We show the induction step from $m-1$ to m. Suppose $V_2, \ldots, V_{m-1}, S_{m-1}$ and X_m are independent. Then $S_m = S_{m-1} + X_m$ and $V_m = S_{m-1}/S_m$. S_m and V_m match the setting for the special case with $m = 2$ from above and are thus independent. Hence, $V_2, \ldots, V_m$ and S_m are independent, completing the induction step. Independence of $V_2, \ldots, V_m$ and S_m implies independence of $\pi_1, \ldots, \pi_m$ and $RSS_{\cdot}$,

Taken together, $RSS - RSS_{\cdot}$, $RSS_{\cdot}$, and $\pi_1, \ldots, \pi_m$, are mutually independent. Now, we can write the test statistics for the dispersion parameters as functions of the relative contributions π_ℓ:

$$LR^{LN} = LR(\pi_1, \ldots, \pi_\ell) = \sum_{\ell=1}^{m} df_\ell \left[\log\left(\frac{df_\ell}{df_{\cdot}}\right) - \log(\pi_\ell) \right], \quad F_{\sigma^2}^{LN} = \frac{df_1}{df_2}\frac{\pi_2}{\pi_1}.$$

Thus, F_{μ}^{LN} is independent of $F_{\sigma^2}^{LN}$ and LR^{LN}.

Proof of Theorem 3.2

This follows from (3.8), independence of d_ℓ across ℓ due to disjoint sub-samples made up of independent Y_{ij}, the continuous mapping theorem, and the results discussed in §3.3.

Proof of Lemma 3.1

Once we show that $D - D_{\cdot}$, $D_{\cdot}$ and $D_1/D_{\cdot}, \ldots, D_m/D_{\cdot}$ are asymptotically mutually independent, the result follows from the proof of Theorem 3.1 since the asymptotic distribution of the deviances in the over-dispersed Poisson model matches the exact distribution of the residual sum of squares in the log-normal model.

We can set $\sigma^2 = 1$ without loss of generality. Then, to prove mutual independence, we build on the insight of Harnau & Nielsen (2017) that asymptotics for the over-dispersed Poisson model match standard exponential family asymptotics. Thus, $D - D_{\cdot}$ and $D_1, \ldots, D_m$ are asymptotically equivalent to quadratic forms (Johansen 1979, Theorem 7.8) of asymptotically Gaussian projections on orthogonal subspaces (Johansen 1979, Theorem 7.6). Thus, independence and hence Lemma 3.1 follows.

i, j	1	2	3	4	5	6	7	8	9	10
1	451288	339519	333371	144988	93243	45511	25217	20406	31482	1729
2	448627	512882	168467	130674	56044	33397	56071	26522	14346	
3	693574	497737	202272	120753	125046	37154	27608	17864		
4	652043	546406	244474	200896	106802	106753	63688			
5	566082	503970	217838	145181	165519	91313				
6	606606	562543	227374	153551	132743					
7	536976	472525	154205	150564						
8	554833	590880	300964							
9	537238	701111								
10	684944									

Table 3.3: Insurance run-off triangle taken from Verrall et al. (2010, Table 1) as used in the empirical application in §3.5.1 and the simulations in §3.6.

i, j	1	2	3	4	5	6	7	8	9	10	11
1	153638	188412	134534	87456	60348	42404	31238	21252	16622	14440	12200
2	178536	226412	158894	104686	71448	47990	35576	24818	22662	18000	
3	210172	259168	188388	123074	83380	56086	38496	33768	27400		
4	211448	253482	183370	131040	78994	60232	45568	38000			
5	219810	266304	194650	120098	87582	62750	51000				
6	205654	252746	177506	129522	96786	82400					
7	197716	255408	194648	142328	105600						
8	239784	329242	264802	190400							
9	326304	471744	375400								
10	420778	590400									
11	496200										

Table 3.4: Insurance run-off triangle taken from Barnett & Zehnwirth (2000, Table 3.5) as used in the empirical application in §3.5.1 and the simulations in §3.6.

4. Log-Normal or Over-Dispersed Poisson?

Abstract Although both over-dispersed Poisson and log-normal chain-ladder models are popular in claim reserving, it is not obvious when to choose which model. Yet, the two models are obviously different. While the over-dispersed Poisson model imposes the variance to mean ratio to be common across the array, the log-normal models assumes the same for the standard deviation to mean ratio. Leveraging this insight, we propose a test that has power to distinguish between the two models. The theory is asymptotic, but it does not build on a large size of the array and instead makes use of information accumulating within the cells. The test has a non-standard asymptotic distribution, however, saddlepoint approximations are available. We show in a simulation study that these approximations are accurate, and that the test performs well in finite samples and has high power.

4.1 Introduction

Which is the better chain-ladder model for claim reserving: over-dispersed Poisson or log-normal? While the expert may have a go-to model, the answer should be informed by the data. Choosing the wrong model could substantially influence the quality of the reserve forecast. Yet, so far no statistical theory is available that supports the actuary in their decision and allows them to make a solid argument in favor of either model.

We develop a test that can distinguish between over-dispersed Poisson and log-normal data generating processes, both of which have a long history in claim reserving. The test exploits that the former model fixes the variance to mean ratio across the array while the latter assumes a common standard deviation to mean ratio. Consequently, the test statistic is based on estimators for the variation in the respective models. The idea is drawn from the econometric literature on encompassing. Intuitively, the test asks whether the null-model can accurately predict the behavior of the rival model's variation estimator when the null-model is true.

The over-dispersed Poisson model is appealing since it naturally pairs with Poisson quasi-likelihood estimation, replicating the popular chain-ladder technique in run-off trian-

gles (Kremer 1985, pp. 130). Furthermore, this model makes for an appealing story due to its relation to compound Poisson distributions. Such distributions give the aggregate incremental claims an interpretation as the sum over a Poisson number of claims with random individual claim amounts (Beard et al. 1984, §3.2). A popular method to generate distribution forecasts for the over-dispersed Poisson model is bootstrapping (England & Verrall 1999, England 2002). While in widespread use, there is so far no theory proving the validity of the bootstrap in this setting. Also, in some settings the method seems to produce unsatisfactory results.

Recently, Harnau & Nielsen (2017) developed a theory that gives the over-dispersed Poisson model a rigorous statistical footing. They propose an asymptotic framework based on infinitely divisible distributions that keeps the dimension of the data array fixed and instead builds on large cell means. This resolves the incidental parameter problem (Neyman & Scott 1948, Lancaster 2000) that renders a standard asymptotic theory based on a large array invalid and arises since the number of parameters grows with the size of the array. The class of infinitely divisible distributions includes compound Poisson distributions which are appealing in an insurance context as noted above. We can then interpret large cell means as the result of a large latent underlying number of claims. Other infinitely divisible distributions that can be reconciled with the over-dispersed Poisson structure include Poisson, gamma, and negative binomial.

The intuition for the theory by Harnau & Nielsen (2017) is that the array is roughly normally distributed for large cell means so that results remind us of a classical analysis of variance (ANOVA) setting. Harnau & Nielsen (2017) show that Poisson quasi-likelihood estimators are t-distributed and F-tests based on Poisson likelihoods can be used to test for model reduction, such as for the absence of a calendar effect. Finally, chain-ladder forecast errors are t-distributed giving rise to closed form distribution forecasts including for aggregates, such as the reserve or cash-flow. In their simulations, Harnau & Nielsen (2017) find that while the bootstrap (England & Verrall 1999, England 2002) matches the true forecast error distribution better on average, the t-forecast produces fewer outliers and appears more robust.

Building on the asymptotic framework put forward by Harnau & Nielsen (2017), Harnau (2018*b*) proposes misspecification tests for two crucial assumptions of the over-dispersed Poisson model. First, the variance to mean ratio is assumed to be common across the array. Second, accident effects are not allowed to vary over development years and vice versa. To check for a violation of these assumptions, Harnau (2018*b*) suggests to split the run-off triangle into sub-samples and then to test whether a reduction from individual models for each sub-sample to a single model for the full array can be justified. While the idea to split the sample is borrowed from time-series econometrics (Chow 1960), the theory for a reduction to a single model is again reminiscent of an ANOVA setting. A classical Bartlett-test (Bartlett 1937) can be used to assess whether we can justify common variance to mean ratios. This is followed by an independent F-test for the absence of breaks in accident and development effects. Again, the asymptotics needed to arrive at these results keep the dimension of the array fixed, growing instead the cell means. Harnau (2018*b*) also shows that these misspecification tests can be used in similar fashion in a finite sample log-normal model.

The log-normal model introduced by Kremer (1982), who relates it to the ANOVA literature, features a predictor structure that is reminiscent of the classical chain-ladder. Verrall (1994) refers to this as the chain-ladder linear model while Kuang et al. (2015) use the term geometric chain-ladder. The latter authors show that the maximum likelihood estimators in the log-normal model can be interpreted as development factors of geometric averages, compared to an interpretation of arithmetic averages arising for the classical chain-ladder. An advantage of the log-normal model is that an exact Gaussian distribution theory applies to the maximum likelihood estimators. However, since these estimators are computed on the log scale, a bias is introduced on the original scale. Verrall (1991) tackles this issue and derives unbiased estimators for the mean and standard deviation on the original scale. One issue for full distribution forecasts in the log-normal model is that the insurer is usually not interested in forecasts for individual cells, but rather for cell sums such as the reserve or the cash-flow. However, the log-normal distribution is not closed under convolution so that cell sums are not log-normally distributed.

Recently, Kuang & Nielsen (2018) proposed a theory that includes closed form distribution forecasts for cell sums, such as the reserve, in the log-normal model, thus remedying one of its drawbacks. Kuang & Nielsen (2018) combined the insight by Thorin (1977) that the log-normal distribution is infinitely divisible and the asymptotic framework by Harnau & Nielsen (2017). Based on this, they propose a theory for generalized log-normal models, a class that nests the log-normal model but is not limited to it. In particular, the distribution is not assumed to be exactly log-normal, but merely needs to be infinitely divisible with a moment structure close to that of the log-normal model. The asymptotics in this framework again leave the dimension of the array untouched to avoid an incidental parameter problem. In contrast to the theory for large cell means in the over-dispersed Poisson model, results are now for small standard deviation to mean ratios.

For the generalized log-normal model, Kuang & Nielsen (2018) show that least squares estimators computed on the log scale are asymptotically t-distributed and simple F-tests based on the residual sum of squares can be used to test for model reduction. Reassuringly, these results match the exact results in a log-normal model. Beyond that, they also prove that forecasts errors on the original scale are asymptotically t distributed so that distribution forecasting for cell sums is straightforward. Further, they show that the misspecification tests by Harnau (2018*b*) are asymptotically valid for the generalized log-normal model, just as they were in finite samples for the log-normal model.

We remark that besides over-dispersed Poisson and log-normal models, there exist a number of reserving models that we do not consider further in this paper. England & Verrall (2002) give an excellent overview. Perhaps the most popular contender is the "distribution free" model by Mack (1993). This model also replicates the classical chain-ladder but differs from the over-dispersed Poisson model. Mack (1993) derives expression for forecast standard errors. However, so far no full distribution theory exists for this model.

With a range of theoretical results in place for over-dispersed Poisson and (generalized) log-normal models, discussed further in §4.3, a natural question is when we should employ which model. The misspecification tests by Harnau (2018*b*) seem like a natural starting

point. For example, if we can reject the specification of the log-normal but not the over-dispersed Poisson model, the latter seems preferable. However, the misspecification tests may not always have enough power to make this distinction as we show in §4.2.

Since generalized log-normal and over-dispersed Poisson models are not nested, a direct test between them is not trivial. Cox (1961, 1962) introduced a theory for non-nested hypothesis testing with a null model. Vuong (1989) provides theory for non-nested model selection without a null model; in selection, the goal is to choose the better, not necessarily the true, model. However, both procedures are likelihood based so that the results are not applicable here since we did not specify exact distributions and thus do not have likelihoods available.

Given the lack of likelihoods for the models, we look to the econometric encompassing literature for inspiration. The theory for encompassing allows for a more general way of non-nested testing. As Mizon & Richard (1986) put it, "Among other criteria, it seems natural to ask whether a specific model, say M_1, can mimic the DGP [data generating process], in that statistics which are relevant within the context of another model, M_2 say, behave as they should were M_1 the DGP." The encompassing literature originates with Hendry & Richard (1982) and Mizon & Richard (1986); for a less technical introduction see Hendry & Nielsen (2007, §11.5). Ermini & Hendry (2008) applied the encompassing principle in a time-series application. They test whether disposable income is better modeled on the original scale or in logs. Taking the log model as null hypothesis, they evaluate whether the log model can predict the behavior of estimators for mean and variance of the model on the original scale.

Building on the encompassing literature, we find the distribution of the over-dispersed Poisson model estimators under a generalized log-normal data generating process and vice versa. It turns out that both Poisson quasi-likelihood and log data least squares estimators for accident and development effects are asymptotically normal, regardless of the data generating process. Differences arise in the second moments. This manifests in the limiting distributions of the variation estimators. While these are asymptotically χ^2 under the correct model, their distribution is a non-standard quadratic form of normals under the rival model. However, these distributions involve the unknown dispersion parameter which

needs to be estimated. Employing the variation estimator of the correct model for this purpose, we arrive at a test statistic with a non-standard asymptotic distribution: the ratio of dependent quadratic forms. Saddlepoint approximations to such distributions are available (Lieberman 1994, Butler & Paolella 2008). Further, we can show that the power of the tests originates from variation in the means across cells. This is intuitive given that the main difference between the models disappears when all means are identical; then, both standard deviation to mean and variance to mean ratios are constant across the array. These findings are collected in §4.4.

With the theoretical results for encompassing tests between over-dispersed Poisson and generalized log-normal models in place, we show that they perform well in a simulation study. First, we demonstrate that saddlepoint approximations to the limiting distributions of the statistics work very well. Second, we tackle an issue that disappears in the limit: we have the choice between a number of asymptotically identical estimators that generally differ in finite samples. Simulations reveal substantial heterogeneity in finite sample performance, but also show that some choices generally do well. Third, we show that the tests have high power for parameterizations we may realistically encounter in practice. We also find that power grows quickly with the variation in the means. The simulation study is in §4.5.

Having convinced ourselves that the tests do well in simulations, we demonstrate their application in a range of empirical applications in §4.6. First, we revisit the empirical illustration of the problem from the beginning of the paper. We show that the test has no problem to reject one of the two rival models. Second, we consider an example that perhaps somewhat cautions against starting with a model that may be misspecified to begin with. In this application, dropping a clearly needed calendar effect turns the results of the encompassing tests upside down. Third, taking these insights into account, we implement a testing procedure that makes use of a whole range of recent results: deciding between over-dispersed Poisson and generalized log-normal model, evaluating misspecification, and testing for the need of a calendar effect.

We conclude the paper with a discussion of potential avenues for future research in §4.7. These include further misspecification tests, a theory for the bootstrap, and empirical

studies assessing the usefulness of the recent theoretical developments in applications.

4.2 Empirical illustration of the problem

i,j	1	2	3	4	5	6	7	8	9	10
1	451288	339519	333371	144988	93243	45511	25217	20406	31482	1729
2	448627	512882	168467	130674	56044	33397	56071	26522	14346	
3	693574	497737	202272	120753	125046	37154	27608	17864		
4	652043	546406	244474	200896	106802	106753	63688			
5	566082	503970	217838	145181	165519	91313				
6	606606	562543	227374	153551	132743					
7	536976	472525	154205	150564						
8	554833	590880	300964							
9	537238	701111								
10	684944									

Table 4.1: Run-off triangle taken from Verrall et al. (2010) with indication for split into sub-samples corresponding to the first and last five accident years. Accident years i in the rows, development years j in the columns.

We illustrate in an empirical example that the choice between over-dispersed Poisson and (generalized) log-normal model is not always obvious. Table 4.1 shows a run-off triangle taken from Verrall et al. (2010, Table 1) with accident years i in the rows and development years j in the columns. Calendar years $k = i + j - 1$ are on the diagonals.

While Kuang et al. (2015) and Harnau (2018*b*) model the data in Table 4.1 as log-normal, it is not obvious whether a log-normal or an over-dispersed Poisson model is more appropriate. In a log normal model, the aggregate incremental claims Y_{ij} are independent

$$M^{LN}: \quad \log(Y_{ij}) = N(\alpha_i + \beta_j + \delta, \omega^2)$$

where α and β are accident and development effects, respectively. On the original scale, this implies that

$$M^{LN} \implies \quad E(Y_{ij}) = \exp\left(\alpha_i + \beta_j + \delta + \frac{\omega^2}{2}\right), \quad \frac{\text{sd}(Y_{ij})}{E(Y_{ij})} = \sqrt{\exp(\omega^2) - 1}. \tag{4.1}$$

Thus, the standard deviation to mean ratio as well as the log data variance are common across cells. If we instead chose an over-dispersed Poisson model, we would maintain the

independence assumption and specify the first two moments of the claims Y_{ij} as

$$M^{ODP}: \quad E(Y_{ij}) = \exp(\alpha_i + \beta_j + \delta), \quad \frac{\text{var}(Y_{ij})}{E(Y_{ij})} = \sigma^2.$$

Thus, the variance to mean ratio σ^2 is identical for all cells.

To choose between the two models, we could take the misspecification tests by Harnau (2018*b*) as a starting point. To implement the tests, the data is first split into sub-samples. For the Verrall et al. (2010) data, Harnau (2018*b*) considers a split into two sub-samples consisting of cells relating to the first and last five accident years as illustrated in Table 4.1. The idea is then to test for common parameters across the sub-samples. In the log-normal model, we first Bartlett-test for common log data variances ω^2 across sub-samples and, if this is not rejected, F-test for common accident and development effects. Similarly, in the over-dispersed Poisson model we first test for common over-dispersion σ^2 and then again for common accident and development effects.

If one of the models is flagged as misspecified but not the other, the choice becomes obvious. However, in this application we cannot reject either model based on these tests. For the log-normal model, the Bartlett-test for common log data variances yields a p-value of 0.09, the F-test for common effects a p-value of 0.91; the p-values for the equivalent tests in the over-dispersed Poisson model are 0.78 and 0.64, respectively. Therefore, the question remains which model we should choose.

4.3 Overview of the rival models

We first discuss two common elements of the rival models, namely the data structure, and the chain-ladder predictor and its identification. Then, we in turn state assumptions, estimation, and known theoretical results for the over-dispersed Poisson and the generalized log-normal chain-ladder model.

4.3.1 Data

We assume that we have data for a run-off triangle of aggregate incremental claims. We denote the claims for accident year i and development year j by Y_{ij}. Further, we count

calendar years with an offset so the calendar year $k = i + j - 1$. Then we can define the index set for a run-off triangle with I accident, development, and calendar years by

$$\mathcal{I} = \{(i, j) : 1 \leq i, j, k \leq I\}.$$

We define the number of observations in $\mathcal{I}$ as n. We could also allow for data in a generalized trapezoid as defined by Kuang et al. (2008*b*) without changing the results of the paper. Loosely, generalized trapezoids allow for an unbalanced number of accident and development years as well as missing calendar years both in the past and the future.

4.3.2 Identification

We briefly discuss the identification problem of the chain-ladder predictor $\alpha_i + \beta_j + \delta$ that is common to both over-dispersed and generalized log-normal models. Kremer (1985) showed that based on this predictor, Poisson quasi-likelihood estimation replicates the classical chain-ladder point forecasts in a run-off triangle.

The identification problem is that for any a and b,

$$\mu_{ij} = \alpha_i + \beta_j + \delta = (\alpha_i + a) + (\beta_j + b) + (\delta - a - b)$$

where α_i and β_j are accident and development effects, respectively. Thus, no individual effect is identified. Several ad-hoc identification methods are available; for example we could set $\sum_i \alpha_i = \sum_j \beta_j = 0$. Kuang et al. (2008*b*) suggest a parametrization that is canonical in a Poisson model and allows for easy counting of degrees of freedom. The idea is to re-write the linear predictor in terms of a level and deviations from said level as

$$\mu_{ij} = \mu_{11} + \sum_{s=2}^{I} 1_{(i \leq s)} \Delta\alpha_s + \sum_{t=s}^{I} 1_{(j \leq s)} \Delta\beta_s.$$

Thus, we can write

$$\mu_{ij} = x'_{ij}\xi$$

where the design x_{ij} and identified parameter ξ are given by

$$x_{ij} = (1, 1_{(i \leq 2)}, \ldots, 1_{(i \leq I)}, 1_{(j \leq 2)}, \ldots, 1_{(j \leq I)})' \quad \xi = (\mu_{11}, \Delta\alpha_2, \ldots, \Delta\alpha_I, \Delta\beta_2, \ldots, \Delta\beta_I)' \in \mathbb{R}^p.$$

For the asymptotic theory in the over-dispersed Poisson model it turns out to be useful to explicitly decouple the level and its deviations by decomposing as

$$\xi = (\mu_{11}, \xi^{(2)'})' \quad \text{and} \quad x_{ij} = (1, x_{ij}^{(2)'})'.$$

We can then define the aggregate predictor τ and the frequencies π_{ij} as

$$\tau = \sum_{ij \in \mathcal{I}} \exp(\mu_{ij}) \quad \text{and} \quad \pi_{ij} = \frac{\exp(\mu_{ij})}{\tau} = \frac{\exp(x_{ij}^{(2)'} \xi^{(2)})}{\sum_{ij \in \mathcal{I}} \exp(x_{ij}^{(2)'} \xi^{(2)})}. \tag{4.2}$$

Importantly, the frequencies π_{ij} are invariant to the level μ_{11}, the first component of ξ. Therefore, we can vary the aggregate predictor τ by varying μ_{11} without affecting the frequencies π_{ij}. The frequencies π_{ij} are, in turn, functions of $\xi^{(2)}$ alone. Further, we note that given $\xi^{(2)}$, there is a one-one mapping between μ_{11} and τ through $\tau = \exp(\mu_{11}) \sum_{ij \in \mathcal{I}} \exp(x_{ij}^{(2)'} \xi^{(2)})$.

While this choice of identification scheme is useful for derivation of the theory in this paper, any scheme may be used in applications of the results. This is because, as Kuang et al. (2008*b*) point out, the linear predictor μ_{ij} is identified, unlike the individual effects. Since the main results of the paper rely on estimates of the linear predictors μ_{ij} alone, they are unaffected by the choice of a particular identification scheme.

Furthermore, the results in this paper are not limited to the chain-ladder predictor; we could, for example, include a calendar effect. Nielsen (2015) derives the form of the design vector for extended chain-ladder predictors in generalized trapezoids. The identification method is implemented in the R (R Core Team 2016) package `apc` (Nielsen 2015), as well as in the homonymous python package (Harnau 2017).

We note that the identification method can introduce arbitrariness into the forecast for models that require parameter extrapolation, such as the extended chain-ladder model with calendar effects. In the standard chain-ladder model, we can forecast claim reserves without parameter extrapolation; in a continuous setting, Lee et al. (2015) refer to this as in-sample forecasting. In contrast, in the extended chain-ladder model we cannot estimate parameters for future calendar years from the run-off triangle. For this case, Kuang et al. (2008*a*) and Nielsen & Nielsen (2014) explain how forecasts can be influenced by ad-hoc constraints

and lay out conditions for the identification method that make forecasts invariant to these arbitrary and untestable constraints.

4.3.3 Over-dispersed Poisson model

We give the assumptions of the over-dispersed Poisson model and discuss its estimation by Poisson quasi-likelihood. We state the sampling scheme proposed by Harnau & Nielsen (2017) and the asymptotic distribution of the estimators.

4.3.3.1 Asssumptions

The first assumption imposes the over-dispersed Poisson structure on the moments. We can write it as

$$M^{ODP}: \quad E(Y_{ij}) = \exp(\mu_{ij}), \quad \frac{\operatorname{var}(Y_{ij})}{E(Y_{ij})} = \sigma^2.$$

The second assumption is distributional and allows for the asymptotic theory later on. We assume that the independent aggregate claims Y_{ij} have a non-degenerate, non-negative, and infinitely divisible distribution with at least three moments. As noted by Harnau & Nielsen (2017), an appealing example for claim reserving of such a distribution is compound Poisson. The interpretation is that the aggregate incremental claims Y_{ij} can be written as $Y_{ij} = \sum_{\ell=1}^{N_{ij}} X_\ell$ for a Poisson number of claims $N_{ij} \stackrel{D}{=} \text{Poisson}\{\exp(\mu_{ij})\}$ independent of the independent and identically distributed random claim amounts X_ℓ.

4.3.3.2 Estimation

We estimate the over-dispersed Poisson model by Poisson quasi-likelihood. The appeal is that, as noted in §4.3.2, Poisson quasi-likelihood estimation replicates the chain-ladder technique. We explicitly distinguish between the model, subscripted by ODP, and its standard estimators, sub- or super-scripted by ql, to avoid confusion later on when we evaluate the estimators under the rival model.

The fitted values for the linear predictors are given by

$$\hat{\mu}_{ij}^{ql} = x'_{ij}\hat{\xi}_{ql} \quad \text{where} \quad \hat{\xi}_{ql} = \underset{\xi \in \mathbb{R}^p}{\arg\max} \sum_{ij \in \mathcal{I}} \{Y_{ij,\ell}(x'_{ij}\xi) - \exp(x'_{ij}\xi)\}.$$

The fitted value for the aggregate predictor τ is then given by

$$\hat{\tau}_{ql} = \sum_{ij \in \mathcal{I}} \exp(\hat{\mu}_{ij}^{ql}) = \sum_{ij \in \mathcal{I}} Y_{ij},$$

a result implied by the fact that the re-parametrization of the Poisson likelihood in terms of the mixed parameter $(\tau, \xi^{(2)})$ is linearly separable so the parameters are variation independent; see, for example, Martínez-Miranda, Nielsen & Wüthrich (2013), Harnau & Nielsen (2017) or, for a more formal treatment, Barndorff-Nielsen (1978, Theorem 8.4). This implies that the estimator for the aggregate predictor is unbiased for the aggregate mean.

As an estimator for the over-dispersion σ^2, Harnau & Nielsen (2017) use the Poisson deviance D scaled by the degrees of freedom. The deviance is the log likelihood ratio statistic against a model with as many parameters as observations, giving a perfect fit. The estimator is given by

$$\hat{\sigma}^2 = \frac{D}{n-p} \quad \text{where} \quad D = 2 \sum_{ij \in \mathcal{I}} Y_{ij} \{\log(Y_{ij}) - \hat{\mu}_{ij}^{ql}\}. \tag{4.3}$$

4.3.3.3 Sampling Scheme

For the asymptotic theory, we adopt the sampling scheme proposed by Harnau & Nielsen (2017). The idea is to grow the overall mean $\tau = \sum_{ij \in \mathcal{I}} E(Y_{ij})$ while holding the frequencies π_{ij} and thus $\xi^{(2)}$ fixed. We note that this also implies that μ_{11} is $O\{\log(\tau)\}$. In this sampling scheme, information accumulates in the estimated frequencies. In this sense, it is reminiscent of multinomial sampling as used, for example, by Martínez Miranda et al. (2015) in a Poisson model conditional on the data sum. Furthermore, we assume that τ increases in such a way that the skewness vanishes. Harnau & Nielsen (2017) remark that this is implicit for distributions such as Poisson, negative Binomial, and many compound Poisson distributions. Importantly, the sampling scheme holds the number of cells in the run-off triangle fixed. If we instead grew the dimension of the array the number of parameters would also increase, thus making an asymptotic theory difficult.

4.3.3.4 Asymptotic theory

Based on the assumptions in §4.3.3.1 and the sampling scheme §4.3.3.3, Harnau & Nielsen (2017) derived the asymptotic distribution of the estimators.

The theory hinges on Harnau & Nielsen (2017, Theorems 1, 2) which for our purposes can be formulated as

$$\tau^{-1/2}\{Y_{ij} - \exp(\mu_{ij})\} = \tau^{1/2}(Y_{ij}/\tau - \pi_{ij}) \xrightarrow{D} N(0, \sigma^2\pi_{ij}) \quad \text{and} \quad \frac{Y_{ij}}{\tau} \xrightarrow{P} \pi_{ij}. \tag{4.4}$$

An implication of the sampling scheme is that we cannot consistently estimate μ_{11} since the overall mean τ and thus the level μ_{11} grow. However, the remaining parameters $\xi^{(2)}$ are fixed and can be estimated in a consistent way. To ease notation, we define the design matrix X and the diagonal matrix of frequencies Π so

$$X = \{x_{ij} : (i,j) \in \mathcal{I}\}' \quad \text{and} \quad \Pi = \text{diag}\{\pi_{ij} : (i,j) \in \mathcal{I}\}. \tag{4.5}$$

Harnau & Nielsen (2017, Lemma 1) derive the distribution of the estimator for the mean parameters in terms of the mixed parametrization $(\tau, \xi^{(2)'})'$. The advantage is that the two components of the mixed parameter are variation independent so the covariance matrix featured in the asymptotic distribution is block-diagonal. This property turns out to be useful for example in the derivation of distribution forecasts. However, we opt to state the results in terms of the original parameterization by ξ to ease the analogy with the generalized log-normal model below. For our purposes, this does not complicate the theory.

As a corollary to Harnau & Nielsen (2017, Lemma 1), we can then state the distribution of the quasi-likelihood estimator $\hat{\xi}_{ql}$ as follows. All proofs are in the appendix.

Corollary 4.1. *In the over-dispersed Poisson model §4.3.3.1 and §4.3.3.3,*

$$\sqrt{\tau}(\hat{\xi}_{ql} - \xi) = \sqrt{\tau}\begin{Bmatrix} (\hat{\mu}_{11} - \mu_{11}) \\ \hat{\xi}^{(2)}_{ql} - \xi^{(2)} \end{Bmatrix} \xrightarrow{D} N\{0, \sigma^2(X'\Pi X)^{-1}\}.$$

Thus, even though the level $\mu_{11} \to \infty$, the difference between estimator and level $(\hat{\mu}_{11} - \mu_{11})$ vanishes in probability. We note that $\tau X'\Pi X$ corresponds to the Fisher information about ξ in a Poisson model.

Further, Harnau & Nielsen (2017, Lemma 1) find that the asymptotic distribution of the deviance is proportional to a χ^2:

$$D \xrightarrow{D} \sigma^2 \chi^2_{n-p}.$$

Thus, the estimator $\hat{\sigma}^2$ has an asymptotic distribution which is unbiased for σ^2.

4.3.4 Generalized Log-Normal model

Following the same structure as for the over-dispersed Poisson model above, we set up the generalized log-normal model as introduced by Kuang & Nielsen (2018) and discuss its estimation and theoretical results. This model nests the log-normal model. While the log-normal model allows for an exact distribution theory for the estimators, Kuang & Nielsen (2018) provide an asymptotic theory that covers the generalized model. We are going to employ this asymptotic theory for the encompassing tests below.

4.3.4.1 Assumptions

The assumptions for the generalized log-normal model mirror those for the over-dispersed Poisson model closely. The assumption of independent Y_{ij} with non-negative, non-degenerate infinitely divisible distribution and at least three moments is maintained. The difference lies in the moment assumptions which are replaced with

$$M^{GLN} : E(Y_{ij}) = \exp\left(\mu_{ij} + \frac{\omega^2}{2}\right) \quad \text{and} \quad \frac{\mathrm{sd}(Y_{ij})}{E(Y_{ij})} = \sqrt{\omega^2}\{1 + o(1)\}$$

where $o(1)$ is with respect to ω^2. Thus, in the generalized log-normal model the standard deviation to mean ratio, also known as the coefficient of variation, is common across the data for small ω^2. This is in contrast to the variance to mean ratio in the over-dispersed Poisson model. Kuang & Nielsen (2018, Theorem 3.2) point out that the log-normal model $\log(Y_{ij}) \stackrel{D}{=} N(\mu_{ij}, \omega^2)$ satisfies these assumptions. There, the standard deviation to mean ratio is $\sqrt{\exp(\omega^2) - 1}$ as in (4.1).

Based on the infinite divisibility assumption, we can construct a story similar to the compound Poisson story for the over-dispersed Poisson model. By definition, Y is infinitely divisible if for any $m > 0$ there exist independent and identically distributed random

variables $X_1, \ldots, X_m$ so $\sum_{\ell=1}^{m} X_\ell$ has the same distribution as Y. Thus, as pointed out by Kuang & Nielsen (2018), we can again think of m as the unknown number of claims and of X_ℓ as the individual claim amounts.

4.3.4.2 Estimation

We estimate the generalized log-normal model on the log scale by least squares. We define

$$Z = \{\log(Y_{ij}) : (i,j) \in \mathcal{I}\}'.$$

Then, least squares fitted values for the linear predictors μ_{ij} are, with the design X as defined in (4.5), given by

$$\hat{\mu}_{ij}^{ls} = x_{ij}' \hat{\xi}_{ls} \quad \text{where} \quad \hat{\xi}_{ls} = (X'X)^{-1} X'Z.$$

We estimate the variation parameter ω^2 based on the residual sum of squares written as

$$\hat{\omega}_{ls}^2 = \frac{RSS}{n-p} \quad \text{where} \quad RSS = \sum_{ij \in \mathcal{I}} (Z_{ij} - \hat{\mu}_{ij}^{ls})^2 = Z'MZ \quad \text{for} \quad M = I - X(X'X)^{-1}X'. \tag{4.6}$$

The estimator for the aggregate predictor τ as defined in (4.2) is then

$$\hat{\tau}_{ls} = \sum_{ij \in \mathcal{I}} \exp(\hat{\mu}_{ij}^{ls}).$$

Unlike in the over-dispersed Poisson model, this estimator is generally not unbiased. Instead, the sum of linear predictors is unbiased the for the sum of logs since $\sum_{ij \in \mathcal{I}} \hat{\mu}_{ij}^{ls} = \sum_{ij \in \mathcal{I}} Z_{ij}$.

4.3.4.3 Sampling scheme

We adopt the sampling scheme Kuang & Nielsen (2018) put forward for the generalized log-normal model. In this scheme, ω^2 vanishes in such a way that the skewness of Y_{ij} goes to zero while ξ remains fixed. In a log-normal model, this corresponds to letting the log data variance ω^2, thus the standard deviation to mean ratio $\sqrt{\exp(\omega^2) - 1}$, go to zero. Again, the dimension of the array $\mathcal{I}$ remains fixed.

4.3.4.4 Asymptotic theory

The asymptotic theory Kuang & Nielsen (2018) introduced for the generalized log-normal model allows to find parameter uncertainty, testing for nested model reduction and closed form distribution forecasts.

Kuang & Nielsen (2018, Theorem 3.4) find that for small ω^2,

$$(\omega^2)^{-1/2}\{Y_{ij} - \exp(\mu_{ij})\} \xrightarrow{D} N\{0, \exp(2\mu_{ij})\}.$$

Thus, generalized log-normal random variables are asymptotically normal but heteroskedastic on the original scale. Furthermore, Kuang & Nielsen (2018, Theorem 3.3) prove that

$$(\omega^2)^{-1/2}(Z_{ij} - \mu_{ij}) \xrightarrow{D} N(0, 1). \tag{4.7}$$

Therefore, conversion to the log scale yields asymptotic normality as well. The difference is that the variance is now homoskedastic. We recall that μ_{ij} is fixed under the sampling scheme in the generalized log-normal model. Therefore, these results imply that $Y_{ij} \xrightarrow{P} \exp(\mu_{ij})$ and $Z_{ij} \xrightarrow{P} \mu_{ij}$. This also means that the data sum $\sum_{ij\in\mathcal{I}} Y_{ij} \xrightarrow{P} \tau$.

The small ω^2 distribution of the estimators in the generalized log-normal model is given by Kuang & Nielsen (2018, Theorem 3.5) as

$$(\omega^2)^{-1/2}(\hat{\xi}_{ls} - \xi) \xrightarrow{D} N\{0, (X'X)^{-1}\} \quad \text{and} \quad \frac{RSS}{\omega^2} \xrightarrow{D} \chi^2_{n-p}. \tag{4.8}$$

In an exact log-normal model, the results in (4.8) hold for any ω^2.

In contrast to the over-dispersed Poisson model, the full parameter vector ξ, including the level μ_{11}, can now be consistently estimated since it is fixed under the sampling scheme. This comes at the cost that ω^2 and thus the standard deviation to mean ratio moves towards zero.

4.4 Encompassing tests

With the two rival models in place, we aim to test the over-dispersed Poisson against the generalized log-normal model and vice versa. Since the models are generally not nested, we cannot simply test for a reduction from one to the other. Instead, we investigate whether

the null model can correctly predict the behavior of the statistics of the rival model if the null model is true. We first consider identifiable differences between the two models. Then, we in turn look at scenarios where the null is the over-dispersed Poisson model versus where it is the generalized log-normal model.

4.4.1 Identifiable differences

It is interesting to consider what key features let us differentiate between the generalized log-normal and the over-dispersed Poisson model. Looking first at the means, we find that differences between the two models are not identifiable. This is because for any $\xi = (\mu_{11}, \xi^{(2)'})'$ and ω^2 in the generalized log-normal model, we can define $\xi^\dagger = (\mu_{11}+\omega^2/2, \xi^{(2)'})'$ for the over-dispersed Poisson model so

$$E_{GLN}(Y_{ij}; \xi, \omega^2) = \exp\left(x'_{ij}\xi + \frac{\sigma^2}{2}\right) = \exp(x'_{ij}\xi^\dagger) = E_{ODP}(Y_{ij}; \xi^\dagger).$$

Thus, we could not even tell the models apart based on the means if we knew their true values.

In contrast, differences in the second moments are identifiable. In the generalized log-normal model, the standard deviation to mean ratio is constant for small ω^2, while the variance to mean ratio is constant in the over-dispersed Poisson model. Since

$$\frac{\text{var}(Y_{ij})}{E(Y_{ij})} = \left\{\frac{\text{sd}(Y_{ij})}{E(Y_{ij})}\right\}^2 E(Y_{ij}),$$

constancy in one ratio generally implies variation in the other, except when all means are identical. Thus, the standard deviation to mean ratio in an over-dispersed Poisson model varies by cell and so does the variance to mean ratio in a generalized log-normal model. Thus, if nature presented us with the true ratios, we could tell the models apart. As noted, an exception arises when all cells have the same mean, a scenario that seems unlikely in claim reserving. If this were the case, the assumptions of the two models are identical: the over-dispersed Poisson model becomes a generalized log-normal model and vice versa. Thus, non-identifiable differences between the ratios imply that both models are congruent with the data generating process in this dimension. Loosely, the two models become more different as the variation in the means increases. We may thus conjecture that there is a

relationship between the power of tests based on standard deviations and variance to mean ratios and the variation in the means.

4.4.2 Null model: over-dispersed Poisson

We find the asymptotic distribution of the least squares estimators, motivated in the generalized log-normal model, when the data generating process is over-dispersed Poisson. We propose a test statistic based on these estimators and find its limiting distribution under an over-dispersed Poisson data generating process.

The estimators from the log-normal model are computed on the log scale. Thus, we first find the limiting distribution of over-dispersed Poisson Y_{ij} on the log scale.

Lemma 4.1. *In the over-dispersed Poisson model §4.3.3.1 and §4.3.3.3,* $\lim_{\tau\to\infty} P(Y_{ij} = 0) = 0$. *For positive* Y_{ij}, *with* $Z_{ij} = \log(Y_{ij})$,

$$\sqrt{\tau}(Z_{ij} - \mu_{ij}) = \sqrt{\tau}\{\log(Y_{ij}/\tau) - \log(\pi_{ij})\} \xrightarrow{D} N(0, \sigma^2\pi_{ij}^{-1}).$$

We stress again that μ_{ij} is not fixed under the sampling scheme so that the result does not imply that Z_{ij} converges to μ_{ij}, rather it implies that their difference $(Z_{ij} - \mu_{ij})$ vanishes. We can relate this lemma to Harnau & Nielsen (2017, Theorem 2) which states that $Y_{ij}/\exp(\mu_{ij}) \xrightarrow{P} 1$. This implies that $\log\{Y_{ij}/\exp(\mu_{ij})\} = \log(Y_{ij}) - \mu_{ij} \xrightarrow{P} 0$, matching what we find here.

Given the limiting distribution on the log scale, we can find the distribution of the estimators in the same way as we would in a Gaussian model. Since the asymptotic distribution of $\sqrt{\tau}(Z_{ij} - \mu_{ij})$ is now heteroskedastic, unlike in the generalized log-normal model as shown in (4.7), we can anticipate that the results will not match those found in the generalized log-normal model. This is confirmed by the following lemma, using the notation for the design matrix X and the diagonal matrix of frequencies Π introduced in (4.5).

Lemma 4.2. *Define* $\Omega = (X'X)^{-1}X'\Pi^{-1}X(X'X)^{-1}$ *and let* $U = N(0, I)$. *Then, in the over-dispersed Poisson model §4.3.3.1 and §4.3.3.3,*

$$\sqrt{\tau}(\hat{\xi}_{ls} - \xi) = \sqrt{\tau}\begin{Bmatrix}(\hat{\mu}_{11} - \mu_{11}) \\ \hat{\xi}_{ls}^{(2)} - \xi^{(2)}\end{Bmatrix} \xrightarrow{D} N(0, \sigma^2\Omega), \ \frac{\hat{\tau}_{ls}}{\tau} \xrightarrow{P} 1 \text{ and } \tau RSS \xrightarrow{D} \sigma^2 U'\Pi^{-1/2}M\Pi^{-1/2}U.$$

As could be expected given Lemma 4.1, the results in Lemma 4.2 match finite sample results in a heteroskedastic independent Gaussian model. Notably, the residual sum of squares RSS are not asymptotically χ^2. However, the over-dispersion σ^2 enters their distribution only multiplicatively. The frequency matrix Π enters as a nuisance parameter that we can, however, consistently estimate since it is a function of $\xi^{(2)}$ alone. For example, we could use plug-in estimators $\widehat{\Pi}_{ql} = \Pi(\hat{\xi}^{(2)}_{ql})$ or $\widehat{\Pi}_{ls} = \Pi(\hat{\xi}^{(2)}_{ls})$. If we knew σ^2, we could feasibly approximate the limiting distribution of RSS. Besides Monte-Carlo simulation, numerical methods are available; see, for example, Johnson et al. (1995*a*, §18.8). These methods exploit that the distribution of the quadratic form can be written as a weighted sum of χ^2_1. Generally, for a real symmetric matrix A and independent χ^2_1 variables V_{ij},

$$U'AU \overset{D}{=} \sum_{ij \in \mathcal{I}} \lambda_{ij} V_{ij}$$

where λ_{ij} are the eigenvalues of A; this follows directly by the Eigendecomposition of A.

Unfortunately, the over-dispersion σ^2 is generally unknown so that we cannot simply base an encompassing test on the residual sum of squares RSS. Therefore, we require an estimator for σ^2. An obvious choice in the over-dispersed Poisson model is the estimator $\hat{\sigma}^2 = D/(n-p)$. However, computed on the same data, D and RSS are not independent. We could tackle this issue in two ways. First, similar to Harnau (2018*b*), we could split the data $\mathcal{I}$ into disjoint and thus independent sub-samples. Then, we could compute RSS on one sub-sample and D on the other, making the two statistics independent. However, in doing so we would incorporate less information into each estimate and likely lose power. Beyond that, it seems little would gained by this approach since no closed form for the distribution of RSS is available in the first place. The second way to tackle the issue is to find the asymptotic distribution of the ratio RSS/D with each component computed over the full sample. This is the way we are going to go.

Before we proceed, we derive an alternative estimator for the over-dispersion σ^2 that gives us more choice later on for the encompassing test. Lemma 4.1 is suggestive of a weighted least squares approach on the log scale since the form of the heteroskedasticity is known, taking Π as given. For

$$X^* = \Pi^{1/2} X, \quad Z^* = \Pi^{1/2} Z \quad \text{and} \quad M^* = I - X^*(X^{*'}X^*)^{-1}X^{*'}, \tag{4.9}$$

the weighted least squares estimators on the log scale are given by

$$\hat{\xi}^* = (X^{*'}X^*)^{-1}X^{*'}Z^* \quad \text{and} \quad RSS^* = Z^{*'}M^*Z^*.$$

Of course, Π is unknown so these estimators are infeasible. However, we can consistently estimate Π. Thus, we can compute feasible weighted least squares estimators. For a first stage estimation of the weights by least squares we write

$$\widehat{\Pi}_{ls} = \Pi(\hat{\xi}_{ls}^{(2)}), \quad X_{ls}^* = \widehat{\Pi}_{ls}X, \quad Z_{ls}^* = \widehat{\Pi}_{ls}Z \quad \text{and} \quad M_{ls}^* = I - X_{ls}^*(X_{ls}^{*'}X_{ls}^*)^{-1}X_{ls}^{*'},$$

so the (least squares) feasible weighted least squares estimators are

$$\hat{\xi}_{ls}^* = (X_{ls}^{*'}X_{ls}^*)^{-1}X_{ls}^{*'}Z_{ls}^* \quad \text{and} \quad RSS_{ls}^* = Z_{ls}^{*'}M_{ls}^*Z_{ls}^*.$$

Similarly, using instead the quasi-likelihood based plug-in estimator $\widehat{\Pi}_{ql} = \Pi(\hat{\xi}_{ql}^{(2)})$ for the weights we write

$$\widehat{\Pi}_{ql} = \Pi(\hat{\xi}_{ql}^{(2)}), \quad X_{ql}^* = \widehat{\Pi}_{ql}X, \quad Z_{ql}^* = \widehat{\Pi}_{ql}Z \quad \text{and} \quad M_{ql}^* = I - X_{ql}^*(X_{ql}^{*'}X_{ql}^*)^{-1}X_{ql}^{*'},$$

so the (quasi-likelihood) feasible weighted least squares estimators are

$$\hat{\xi}_{ql}^* = (X_{ql}^{*'}X_{ql}^*)^{-1}X_{ql}^{*'}Z_{ql}^* \quad \text{and} \quad RSS_{ql}^* = Z_{ql}^{*'}M_{ql}^*Z_{ql}^*.$$

While we would generally expect them to differ in finite samples, it turns out that the Poisson quasi-likelihood and the (feasible) weighted least squares estimators are asymptotically equivalent. We formulate this in a lemma.

Lemma 4.3. *In the over-dispersed Poisson model §4.3.3.1 and §4.3.3.3,* $\sqrt{\tau}(\hat{\xi}^* - \hat{\xi}_{ql}) \xrightarrow{P} 0$ *and, for the Poisson deviance* D *as in* (4.3), $\tau RSS^* - D \xrightarrow{P} 0$. *These results still hold if* $\hat{\xi}^*$ *is replaced by* $\hat{\xi}_{ls}^*$ *or* $\hat{\xi}_{ql}^*$, RSS^* *is replaced by* RSS_{ls}^* *or* RSS_{ql}^*, *or* τ *is replaced by* $\hat{\tau}_{ql}$ *or* $\hat{\tau}_{ls}$.

We are now armed with four candidate statistics for an encompassing test:

$$R_{ls} = \hat{\tau}_{ls}\frac{RSS}{D}, \quad R_{ql} = \hat{\tau}_{ql}\frac{RSS}{D}, \quad R_{ls}^* = \frac{RSS}{RSS_{ls}^*}, \quad \text{and} \quad R_{ql}^* = \frac{RSS}{RSS_{ql}^*}. \tag{4.10}$$

To find their asymptotic distribution, we exploit that the distribution of each one is asymptotically equivalent to a quadratic form of the same random vector Y. This is reflected in the limiting distribution which we formulate in a theorem.

Theorem 4.1. *In the over-dispersed Poisson model §4.3.3.1 and §4.3.3.3, R_{ls}, R_{ql}, R^*_{ls} and R^*_{ql} are asymptotically equivalent so that the difference of any two vanishes in probability. For $U \overset{D}{=} N(0, I)$, Π as in (4.2), M as in (4.6) and M^* as in (4.9), each statistic is asymptotically distributed as*

$$\mathrm{R}_{ODP} = \frac{U'\Pi^{-1/2}M\Pi^{-1/2}U}{U'M^*U}.$$

Crucially, the asymptotic distribution R_{ODP} is invariant to σ^2. While it is again a function of the unknown but consistently estimable frequencies π_{ij}, for large τ, the plug-in version $\widehat{\mathrm{R}}_{ODP} = \mathrm{R}_{ODP}(\widehat{\Pi})$ has the same distribution as $\mathrm{R}_{ODP}(\Pi)$.

Theorem 4.1 allows us to test whether the over-dispersed Poisson model encompasses the generalized log-normal model. For a given critical value, if we reject that the R-statistic was drawn from $\widehat{\mathrm{R}}_{ODP}$, then we reject that the over-dispersed Poisson model M^{ODP} encompasses the generalized log-normal model. While this indicates that the over-dispersed Poisson model is likely wrong, it could mean that the generalized log-normal model is correct or that some other model is appropriate. Conversely, non-rejection means that we cannot reject that the over-dispersed Poisson model encompasses the generalized log-normal model.

The distribution R_{ODP} does not have a closed form but precise saddlepoint approximations are available as we show below. Furthermore, it is of interest to investigate the impact of the choice among the different test statistics and plug-in estimators for Π appearing in R_{ODP} in finite samples. Above that, we may question the power properties of the test. We discuss these points below in §4.5.

4.4.3 Null model: generalized log-normal

We first derive the small-ω^2 asymptotic distribution of Poisson quasi-likelihood and weighted least squares estimators when the data generating process is generalized log-normal. Then, we find the asymptotic distribution of the R-statistic proposed for an encompassing test above.

First, given asymptotic standard-normality on the log scale as in (4.7), we can easily show asymptotic normality of the weighted least squares estimator. As it turns out, Poisson

quasi-likelihood estimators are also asymptotically equivalent to the weighted least squares estimators when the data generating process is generalized log-normal. We formalize this result in a lemma.

Lemma 4.4. *Define* $\Sigma = (X'\Pi X)^{-1}X'\Pi^2 X(X'\Pi X)^{-1}$ *and let* $U \overset{D}{=} N(0, I)$. *Then, in the generalized log-normal model §4.3.4.1 and §4.3.4.3,*

$$(\omega^2)^{-1/2}(\hat{\xi}^* - \xi) \overset{D}{\to} N(0, \Sigma) \quad \textit{and} \quad (\omega^2)^{-1}RSS^* \overset{D}{\to} U'\Pi^{1/2}M^*\Pi^{1/2}U.$$

Further, $(\omega^2)^{-1/2}(\hat{\xi}^* - \hat{\xi}_{ql}) \overset{P}{\to} 0$ *and* $(\omega^2)^{-1}(RSS^* - D/\tau) \overset{P}{\to} 0$. *These results still hold if* $\hat{\xi}^*$ *is replaced by* $\hat{\xi}^*_{ls}$ *or* $\hat{\xi}^*_{ql}$, RSS^* *is replaced by* RSS^*_{ls} *or* RSS^*_{ql}, *or* τ *is replaced by* $\hat{\tau}_{ql}$ *or* $\hat{\tau}_{ls}$.

With these results in place, we can find the distribution of the R-statistics in the generalized log-normal model.

Theorem 4.2. *In the generalized log-normal model §4.3.4.1 and §4.3.4.3,* R_{ls}, R_{ql}, R^*_{ls} *and* R^*_{ql} *as in* (4.10) *are asymptotically equivalent so that the difference of any two vanishes in probability. For* $U \overset{D}{=} N(0, I)$, Π *as in* (4.2), M *as in* (4.6) *and* M^* *as in* (4.9), *each statistic is asymptotically distributed*

$$\mathrm{R}_{GLN} = \frac{U'MU}{U'\Pi^{1/2}M^*\Pi^{1/2}U}.$$

Thus, the test statistics are asymptotically distributed as the ratio of quadratic forms in both data generating processes. The difference arises in the sandwich-matrices. While the orthogonal projections M and M^* feature in both distributions, the frequency matrix Π acts in different ways on R_{ODP} and R_{GLN}. Intuitively, R_{ODP} is the ratio of "bad" least squares to "good" weighted least squares residuals computed in a heteroskedastic Gaussian model. In contrast, R_{GLN} has the interpretation as the ratio of "good" least squares to "bad" weighted least squares residuals now computed in a homoskedastic model. Thus, we may expect draws from R_{GLN} to likely be smaller than those from R_{ODP}.

4.4.4 Distribution of ratios of quadratic forms

We discuss the support of and numerical saddlepoint approximations to the limiting distributions of the encompassing tests under either data generating process.

The limiting distribution under the null hypothesis in both models is a ratio of dependent quadratic forms in normal random variables. This class of distributions is rather common. Besides standard F distributions, which are a special case, they appear for example in the Durbin-Watson test for serial correlation (Durbin & Watson 1950, 1951). While the distributions generally do not permit closed form computations of the cdf, fast and precise numerical methods are available.

Butler & Paolella (2008) study a setting that includes ours but is more general. They consider $\mathrm{R} = \epsilon' A\epsilon / \epsilon' B\epsilon$ where A and B are symmetric $n \times n$ matrices, B is positive semidefinite, and $\epsilon \overset{D}{=} N(\nu, I)$. In our scenario, both A and B are positive semidefinite, and $\nu = 0$.

Butler & Paolella (2008, Lemma 2) state that R is degenerate if and only if $A = cB$ for some constant c. In our setting, this occurs if $\Pi = n^{-1}I$ so all cells have the same mean. This matches our observation from §4.4.1 that generalized log-normal and over-dispersed Poisson model are indistinguishable if all cells Y_{ij} have the same mean. In that case, both the standard deviation to mean and the variance to mean ratio are constant across cells. This manifests in the collapse of both R_{ODP} and R_{GLN} to a point mass at n.

Further, Butler & Paolella (2008, Lemma 3) derive the support of R for a variety of cases depending on the properties of A and B. Building on their work, we can prove the following result.

Lemma 4.5. *The distributions* R_{GLN} *and* R_{ODP} *have the same support. In non-degenerate cases, the support is* (l, r) *for* $0 < l < r < \infty$.

The cumulative distribution functions and densities of ratios of quadratic forms admit saddlepoint approximations. We adapt the discussion in Butler & Paolella (2008) to our scenario in which $\nu = 0$; a setting which matches Lieberman (1994). We aim to approximate

$$P(\mathrm{R} \leq r) = P\left(\frac{\epsilon' A\epsilon}{\epsilon' B\epsilon} \leq r\right) = P(X_r \leq 0) \quad \text{where} \quad X_r = \epsilon'(A - rB)\epsilon$$

First, we compute the Eigenvalues of $A - rB$ denoted $\lambda_1, \ldots, \lambda_n$. We can write the cumulant generating function $K(s)$, the log of the moment generating function $\varphi(s) = E[\exp\{s(A -$

$rB)\}]$, of X_r and its ℓ-th derivative as

$$K(s) = -\frac{1}{2}\sum_{t=1}^{n}\log(1-2s\lambda_t), \quad K^{(\ell)}(s) = \left(\frac{\partial}{\partial s}\right)^{\ell} K(s) = (2\ell-2)!!\sum_{t=1}^{n}\left(\frac{\lambda_t}{1-2s\lambda_t}\right)^{\ell}$$

where $a!! = a(a-2)(a-4)\cdots$ is the double factorial with the usual definition that $0!! = 1$. The saddlepoint is the root

$$\hat{s}: \quad K^{(1)}(\hat{s}) = \sum_{t=1}^{n}\left(\frac{\lambda_t}{1-2\hat{s}\lambda_t}\right) = 0.$$

Except for the special case when all Eigenvalues λ_t are 0 so $K^{(1)}(s) = 0$, $\hat{s}$ is unique since $K^{(1)}(s)$ is strictly increasing. The former case occurs if and only if $E(X_r) = 0$ which is the case for $r = \text{trace}(A)/\text{trace}(B)$. This case is dealt with separately. For the other cases, we compute

$$\hat{w} = \text{sgn}(\hat{s})\sqrt{-2K(\hat{s})} \quad \text{and} \quad \hat{u} = \hat{s}\sqrt{K^{(2)}(\hat{s})}.$$

Then, denoting by $\Phi(.)$ and $\phi(.)$ the standard normal cdf and density, respectively, the first order approximation to the cdf of R is

$$\widehat{P}(\text{R} \leq r) = \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})(\hat{w}^{-1} - \hat{u}^{-1}), & \text{if } E(X_r) \neq 0 \\ \frac{1}{2} + \frac{K^{(3)}(0)}{6\sqrt{2\pi}K^{(2)}(0)^{3/2}}, & \text{if } E(X_r) = 0 \end{cases}.$$

This saddlepoint approximation is a special case of the more general form in Lugannani & Rice (1980). This is what Lieberman (1994) builds on. Lugannani & Rice (1980) analyzed the error behavior for a sum of independent and identically distributed random variables and showed uniformity of the errors for a large sample. Butler & Paolella (2008) instead consider a fixed sample size and show uniformity of errors in the tail of the distribution. This seems appealing for our scenario since we would expect the rejection region of the test to correspond to the tail of the distribution.

4.4.5 Power

We show that the conjecture of a link between power of the tests and variation in the means raised above in §4.4.1 is correct. To prove this, we consider a sequential asymptotic argument in which first, depending on the data generating process, τ becomes large or ω^2 becomes small and then the means become "more dispersed" in a sense made precise

below. Based on this argument, we can justify a one-sided test where the rejection region corresponds to the upper tail when the null model is generalized log-normal and to the lower tail when it is over-dispersed Poisson.

The sequential asymptotics allows us to exclusively consider the impact of more dispersed means on R_{GLN} and R_{ODP} without worrying about the effect on the distribution of R_{ls}, R_{ql}, R^*_{ls} or R^*_{ql}. However, larger mean dispersion would be linked to changes in $\xi^{(2)}$, a parameter that we keep fixed when deriving the asymptotic distribution of the test statistics in the first stage of the asymptotics. Thus, we would expect the approximation quality achieved in the first stage to be affected by the second stage. The interpretation of the results is thus for a given first stage approximation quality, however large τ or small ω^2 may be needed to achieve this.

We model "more dispersed" means by increasing the variation in the frequencies π_{ij} and specifically by letting some frequencies go to zero. In this way, we do not make a statement about the means in absolute terms but merely say that some cell means become large relative to others.

For our analysis, we exclude cells for which estimation would yield a perfect fit; equivalently we can impose that the frequencies do not exclusively vanish for perfectly fitted cells. For example, in a chain-ladder model for the run-off triangle in Table 4.1, this would correspond to the corner cells $(1, 10)$ and $(10, 1)$ which would be fit perfectly as they have their own parameters $\Delta\beta_{10}$ and $\Delta\alpha_{10}$.

To increase the variation in the frequencies, we decide on $n - q$ cells of the run-off triangle for which we want the frequencies to vanish. We require that the remaining q cells with non-vanishing frequencies make up an array on which we can estimate a model with the same structure for the linear predictor μ_{ij} without obtaining a perfect fit. For example, for a chain-ladder model in which $\mu_{ij} = \alpha_i + \beta_j + \delta$, this would be the case for rectangular arrays with at least two columns and rows or for triangular arrays with at least three rows and columns.

For ease of notation, we sort rows and columns of the frequency matrix Π as defined in (4.5) such that the cells with vanishing frequencies are in the bottom right block of the

matrix. Then, for a $q \times q$ matrix Π_1 and an $n-q \times n-q$ matrix Π_2, we define a new frequency matrix

$$\Pi_{(t)} = s(t) \begin{pmatrix} \Pi_1 & 0 \\ 0 & t\Pi_2 \end{pmatrix} \quad \text{where} \quad s(t) = \{\text{trace}(\Pi_1) + t \cdot \text{trace}(\Pi_2)\}^{-1}$$

so $s(t)$ takes care of the normalization such that the elements of $\Pi_{(t)}$ are still frequencies. Thus, $\Pi_{(1)}$ corresponds to Π whereas $\Pi_{(0)}$ has all frequencies in the the bottom right block equal to zero. We assume that $\Pi_1 \neq q^{-1}I$ so that the limiting case does not correspond to a scenario without variation in the frequencies.

We are now interested in the small-t behavior of the limiting distribution under either data generating process

$$\mathrm{R}^{(t)}_{GLN} = \frac{U'MU}{U'\Pi^{1/2}_{(t)}M^*_{(t)}\Pi^{1/2}_{(t)}U} \quad \text{and} \quad \mathrm{R}^{(t)}_{ODP} = \frac{U'\Pi^{-1/2}_{(t)}M\Pi^{-1/2}_{(t)}U}{U'M^*_{(t)}U}$$

where $M^*_{(t)}$ is the weighted least squares orthogonal projection in (4.9) based on $X^*_{(t)} = \Pi^{1/2}_{(t)}X$. A first intuition based on the behavior of $\Pi_{(t)}$ alone, neglecting the behavior of $M^*_{(t)}$, may tell us that $\mathrm{R}^{(t)}_{GLN}$ should be well behaved while $\mathrm{R}^{(t)}_{ODP}$ blows up for small t. This turns out to be correct.

Theorem 4.3. *As $t \to 0$, for $\mathrm{R}^{(0)}_{GLN}$ as defined in (4.11) in the appendix, $\mathrm{R}^{(t)}_{GLN} \xrightarrow{a.s.} \mathrm{R}^{(0)}_{GLN}$ while $\mathrm{R}^{(t)}_{ODP} \xrightarrow{a.s.} \infty$. Further, for $\alpha \in (0,1)$, let $q^{(t)}_{GLN,\alpha}$ be the α-quantile of $\mathrm{R}^{(t)}_{GLN}$ and similarly for $q^{(t)}_{ODP,\alpha}$. Then $\mathrm{R}^{(t)}_{ODP} > q^{(t)}_{GLN,\alpha}$ and $\mathrm{R}^{(t)}_{GLN} \leq q^{(t)}_{ODP,\alpha}$ almost surely as $t \to 0$.*

Theorem 4.3 justifies one-sided tests and shows that the power of the tests under either data generating process goes to unity in the sequential asymptotic argument. Since the distribution of R_{ODP} and R_{GLN} coincides for equal means, the power of the tests to distinguish between the data generating processes comes entirely from the variation in means. As the mean variation becomes large, R_{ODP} first order stochastic dominates R_{GLN}. Thus we can consider the lower tail of R_{ODP} and the upper tail of R_{GLN} as rejection regions. While still controlling the size of the test under the null, we gain power compared to two-sided tests as the mean variation increases.

4.5 Simulations

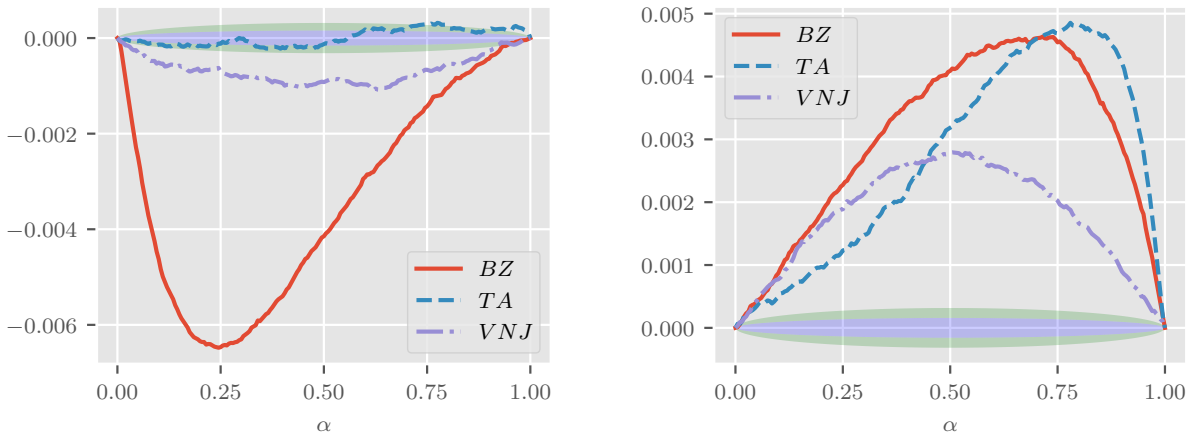
With the theoretical results for encompassing tests between over-dispersed Poisson and generalized log-normal models in place, we show that they perform well in a simulation study. First, we show that saddlepoint approximations to the limiting distributions R_{ODP} and R_{GLN} are very accurate. Second, we tackle an issue that disappears in the limit. Namely, the choice between asymptotically identical estimators that generally differ in finite samples. We show that finite sample performance is indeed affected by this choice. However we find that for some choices finite sample and asymptotic distributions are very close. Third, we show that the tests have high power in finite samples and, considering the behavior of the limiting distributions alone, that power increases quickly with the variation in means. For the simulations and empirical applications below, we use the python packages `quad_form_ratio` (Harnau 2018*c*) and `apc` (Harnau 2017). The package was inspired by the R (R Core Team 2016) package `apc` (Nielsen 2015) with similar functionality.

4.5.1 Quality of saddlepoint approximations

We show that saddlepoint approximations work well compared to large Monte-Carlo simulations.

We consider three parameterizations. First, we let the design X correspond to a that of a chain-ladder model for a ten-by-ten run-off triangle and set the frequency matrix Π to the least squares estimates $\widehat{\Pi}_{ls} = \Pi(\hat{\xi}_{ls}^{(2)})$ of the Verrall et al. (2010) data in Table 4.1 (VNJ). Second, for the same design, we now set the frequency matrix to the least squares plug-in estimates based on a popular data set by Taylor & Ashe (1983) (TA). We provide these data in the appendix. Third, we consider a design X for an extended chain-ladder model in an eleven-by-eleven run-off triangle and set Π to the least squares plug-in estimates of the Barnett & Zehnwirth (2000) data (BZ), also shown in the appendix. We remark that in the computations, we drop the corner cells of the triangles that would be fit perfectly in any case; this helps to avoid numerical issues without affecting the results.

Given a data generating process R chosen from R_{ODP} and R_{GLN}, a design matrix X and a frequency matrix Π, we use a large Monte-Carlo simulation as a benchmark for



(a) $\widehat{P}^{SP}(\mathrm{R}_{GLN} \leq q_\alpha) - \alpha$

(b) $\widehat{P}^{SP}(\mathrm{R}_{ODP} \leq q_\alpha) - \alpha$

Figure 4.1: Approximation error of the first order saddlepoint approximation to R_{GLN}, shown in (a), and R_{ODP}, displayed in (b). Monte-Carlo simulation with 10^7 draws taken as truth. One and two Monte-Carlo standard errors shaded in blue and green, respectively.

the saddlepoint approximation. First, we draw $B = 10^7$ realizations r_b from R. For the Monte-Carlo cdf $\widehat{P}^{MC}(\mathrm{R} \leq q) = B^{-1} \sum_{b=1}^{B} 1(r_b \leq q)$, we then find the quantiles q_α so $\widehat{P}^{MC}(\mathrm{R} \leq q_\alpha) = \alpha$ for $\alpha = 0, 0.005, 0.01, \ldots, 1$. To compute the saddlepoint approximation $\widehat{P}^{SP}(\mathrm{R} \leq q)$ we use the implementation of the procedure described in § 4.4.4 in the package `quad_form_ratio`. Then, for each Monte-Carlo quantile q_α, we compute the difference $\widehat{P}^{SP}(\mathrm{R} \leq q_\alpha) - \alpha$. Taking the Monte-Carlo cdf as the truth, we refer to this as the saddlepoint approximation error.

Figure 4.1a shows the saddlepoint approximation error for the generalized log-normal model $\widehat{P}^{SP}(\mathrm{R}_{GLN} \leq q_\alpha) - \alpha$ plotted against α. One and two (pointwise) Monte-Carlo standard errors $\sqrt{\alpha(1-\alpha)/B}$ are shaded in blue and green, respectively. While the approximation errors for TA are generally not significantly different from zero, the same cannot be said for the other two sets of parameters. For the parameterizations VNJ and BZ, the errors start and end in zero and are negative in between. Despite statistically significant differences, the approximation is very good with a maximum absolute approximation error of just over -0.006. The errors in the tails are much smaller, as we might have expected given the results by Butler & Paolella (2008) discussed in §4.4.4.

Figure 4.1b shows the plot for the approximation error to R_{ODP} produced in the same way as Figure 4.1a. The approximation error is positive and generally significantly different from zero across parameterizations. Yet, the largest error is about 0.005 with smaller errors in the tails.

We would argue that the saddlepoint approximation errors, while statistically significant, are negligible in applications. That is, using a saddlepoint approximation rather than a large Monte-Carlo simulation is unlikely to affect the practitioners modeling decision.

4.5.2 Finite sample approximations under the null

The asymptotic theory above left us without guidance on how to choose between test statistics R and estimators for the nuisance parameter Π that appears in the limiting distributions R. While the choice is irrelevant for large τ or small ω^2, we show that it matters in finite samples and that some combinations perform much better than others when it comes to approximation under the null hypothesis.

In applications, we approximate the distribution of R by $\widehat{\mathrm{R}} = \mathrm{R}(\widehat{\Pi})$. That is, defining the α quantile of $\widehat{\mathrm{R}}$ as $q_\alpha^{\widehat{\mathrm{R}}}$, we hope that $P(R \leq q_\alpha^{\widehat{\mathrm{R}}}) \approx \alpha$ under the null hypothesis. To assess whether this is justified, we simulate the approximation quality across 16 asymptotically identical combinations of R-statistics and ratios of quadratic forms $\widehat{\mathrm{R}}$. We describe the simulation process in three stages. First, we explain how we set up the data generating processes for generalized log-normal and over-dispersed Poisson model. Second, we lay out explicitly the combinations we consider. Third, we explain how we compute the approximation errors. As in §4.5.1, we point out that we drop the corner cells of the triangles in simulations. This aids numerical stability without affecting the results.

For the generalized log-normal model, we simulate independent log-normal variables Y_{ij} so $\log(Y_{ij}) \overset{D}{=} N(x'_{ij}\xi, \omega^2)$. We consider three settings for the true parameters corresponding largely to the estimates from the same three datasets we used in §4.5.1, namely the Verrall et al. (2010) data (VNJ), Taylor & Ashe (1983) data (TA), and Barnett & Zehnwirth (2000) data (BZ). Specifically, we consider pairs (ξ, ω^2) set to the estimated counterparts $(\hat{\xi}_{ls}, \hat{\omega}^2/s)$ for $s = 1, 2$. The estimates $\hat{\omega}^2$ are 0.39 for VNJ, 0.12 for TA and 0.001 for BZ. Theory tells us that the approximation errors should decrease with ω^2, thus as s increases.

For the over-dispersed Poisson model, we use a compound Poisson-gamma data generating process, largely following Harnau & Nielsen (2017) and Harnau (2018*b*). We simulate independent $Y_{ij} = \sum_{\ell=1}^{N_{ij}} X_s$ where $N_{ij} \stackrel{D}{=} \text{Poisson}\{\exp(x'_{ij}\xi)\}$ and X_s are independent Gamma distributed with scale $\sigma^2 - 1$ and shape $(\sigma^2 - 1)^{-1}$. This satisfies the assumptions for the over-dispersed Poisson model in §4.3.3.1 and §4.3.3.3. For the true parameters $(\tau, \xi^{(2)}, \sigma^2)$, we consider three sets of estimates $(s\hat{\tau}_{ql}, \hat{\xi}^{(2)}_{ls}, \hat{\sigma}^2)$ from the same data as for the log-normal data generating process. We use least squares estimates $\hat{\xi}^{(2)}_{ls}$ so that the frequency matrix Π is identical within parameterization between the two data generating processes. The estimates for $\hat{\sigma}^2$ are 10393 for VNJ, 52862 for TA and 124 for BZ. Those for $\hat{\tau}_{ql}$ are $14,633,814$ for VNJ, $34,358,090$ for TA and $10,221,194$ for BZ. Again, we consider $s = 1, 2$ but this time scaling the aggregate predictor. If this increases, so should the approximation quality. We recall that $\xi^{(2)}$ and τ pin down μ_{11} through the one-one mapping $\tau = \exp(\mu_{11}) \sum_{ij \in \mathcal{I}} \exp(x^{(2)'}_{ij} \xi^{(2)})$. Thus, multiplying τ by s corresponds to adding $\log(s)$ to μ_{11}.

For a given data generating process, we independently draw $B = 10^5$ run-off triangles $\triangle_b = \{Y_{ij,b} : (i,j) \in \mathcal{I}\}$ and compute a battery of statistics for each draw. First, we compute the four test statistics R_{ls}, R_{ql}, R^*_{ls} and R^*_{ql} as defined in (4.10). Second, we compute the estimates for the frequency matrices Π based on least squares estimates, quasi-likelihood estimates, and feasible weighted least squares estimates with least squares and with quasi-likelihood first stage. This leads to four different approximations to the limiting distribution which, dropping the subscript for the data generating process, we denote by

$$\widehat{\mathrm{R}}_{ls} = \mathrm{R}\{\Pi(\hat{\xi}^{(2)}_{ls})\}, \quad \widehat{\mathrm{R}}_{ql} = \mathrm{R}\{\Pi(\hat{\xi}^{(2)}_{ql})\}, \quad \widehat{\mathrm{R}}^*_{ls} = \mathrm{R}\{\Pi(\hat{\xi}^{*(2)}_{ls})\}, \quad \text{and} \quad \widehat{\mathrm{R}}^*_{ql} = \mathrm{R}\{\Pi(\hat{\xi}^{*(2)}_{ql})\}.$$

Given a data generating process and a choice of test statistic and limiting distribution approximation $(R, \widehat{\mathrm{R}})$, we approximate $P(R \leq q^{\widehat{\mathrm{R}}}_{\alpha})$ by Monte-Carlo simulation. For each combination $(R, \widehat{\mathrm{R}})$, we have B paired realizations; for example, R_b and the distribution $\widehat{\mathrm{R}}_b$ are based on the triangle $\triangle_b$. Denote the saddlepoint approximation to the cdf of $\widehat{\mathrm{R}}_b$ as $G_b(q) = \widehat{P}^{SP}(\widehat{\mathrm{R}}_b \leq q)$. Neglecting the saddlepoint approximation error, we then compute

$P(R \leq q_\alpha^{\widehat{\mathrm{R}}})$ as $\widehat{P}^{MC}(R \leq q_\alpha^{\widehat{\mathrm{R}}}) = B^{-1}\sum_{b=1}^{B} 1\{G_b(R_b) \leq \alpha\}$, exploiting that $G_b(R_b) \leq \alpha$ whenever $R_b \leq G_b^{-1}(\alpha) = q_{\alpha,b}^{\widehat{\mathrm{R}}}$. We do this for $\alpha \in \mathcal{A} = \{0.005, 0.01, \ldots, 0.995\}$.

To evaluate the performance, we consider three metrics: area under the curve of absolute errors (also roughly the mean absolute error), maximum absolute error, and error at (one-sided) 5% critical values. We compute the area under the curve as $\text{AUC} = \sum_{\ell=1}^{199} |\widehat{P}^{MC}(R \leq q_{\alpha_\ell}^{\widehat{\mathrm{R}}}) - \alpha_\ell|\Delta\alpha_\ell$ where $\alpha_\ell = 0.005 \cdot \ell$ so $\Delta\alpha_\ell = 0.005$; we can also roughly interpret this as the mean absolute error $\text{MAE} = 200/199 \cdot \text{AUC}$ since $\alpha_\ell = 200^{-1}$. The maximum absolute error is $\max_{\alpha\in\mathcal{A}} |\widehat{P}^{MC}(R \leq q_\alpha^{\widehat{\mathrm{R}}}) - \alpha|$. Finally, the error at 5% critical values is $\widehat{P}^{MC}(R > q_{GLN,0.95}^{\widehat{\mathrm{R}}}) - 0.05$ for the generalized log-normal and $\widehat{P}^{MC}(R \leq q_{ODP,0.05}^{\widehat{\mathrm{R}}}) - 0.05$ for the over-dispersed Poisson data generating process.

Figure 4.2 shows bar charts for the area under the curve for all 16 combinations of R and $\widehat{\mathrm{R}}$ stacked across the three parameterizations for $s = 1$. The chart is ordered by the sum of errors across parameterizations and data generating processes within combination, increasing from top to bottom. The maximum absolute error summed over parameterizations is indicated by "+". Since a bar chart for the maximum absolute errors is qualitatively very similar to the plot for the area under the curve we do not discuss it separately and instead provide the figure in the appendix.

Looking first at the sum over parameterizations and data generating processes within combinations, we see large differences in approximation quality both for the area under the curve of absolute errors and the maximum absolute error. The former varies from about $5pp$ for $(R_{ls}^*, \widehat{\mathrm{R}}_{ls}^*)$ to close to $30pp$ for $(R_{ql}, \widehat{\mathrm{R}}_{ls}^*)$, the latter from $8pp$ to $45pp$. It is notable that the four combinations involving R_{ql} are congregated at the bottom of the pack. In contrast, the three best performing combinations all involve R_{ls}^*. These three top-performers have a substantial head start compared to their competition. While their AUC varies from $4.8pp$ and $6.0pp$, there is a jump to $13.9pp$ for fourth place. Similarly the maximum absolute errors of the top three contenders lie between $7.5pp$ and $9.3pp$ while those for fourth place add up to $20.5pp$.

Considering next the contributions of the individual parameterizations to the area under the curve across data generating processes, the influence is by no means balanced. Instead, the average contribution over combinations of the VNJ, TA and BZ parameterizations

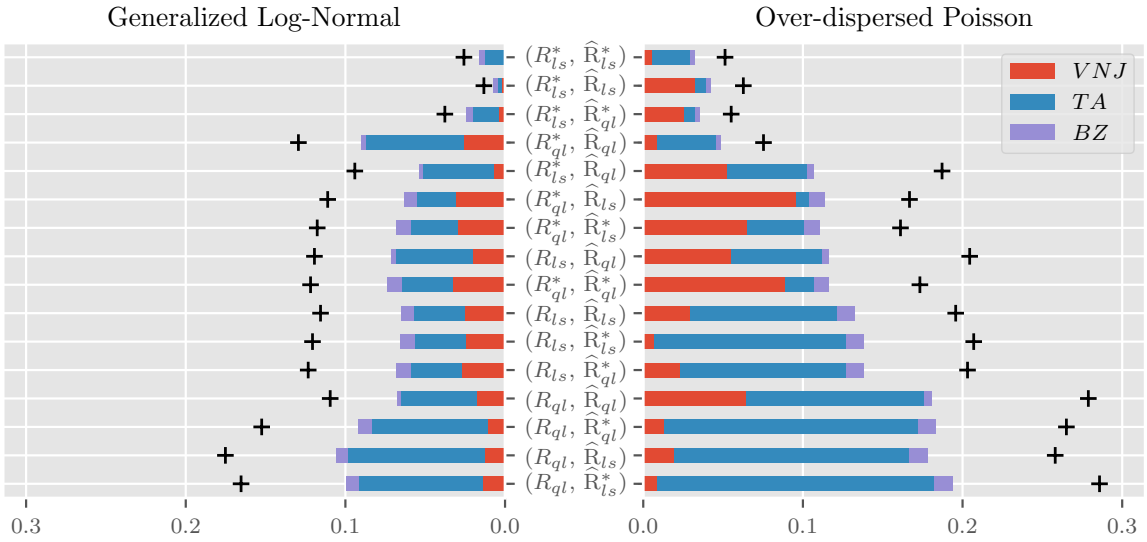


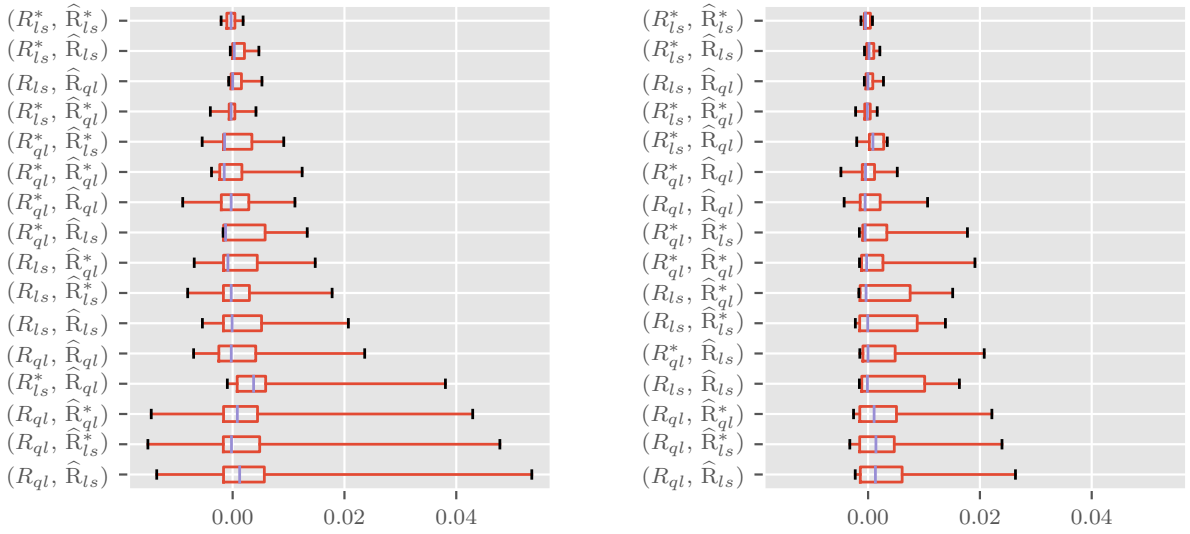
Figure 4.2: Bar chart of area under the curve of absolute approximation errors (also roughly mean absolute error) for the considered combinations of R and $\widehat{\mathrm{R}}$. Ordered by the sum of errors within combination across data generating processes and parameterizations increasing from top to bottom. Sum of maximum absolute errors across parameterizations indicated by "+". VNJ, TA, and BZ is short for parameters set to their estimates from the Verrall et al. (2010) data in Table 4.1, the Taylor & Ashe (1983) data and the Barnett & Zehnwirth (2000) data, respectively. The latter two data sets are provided in the appendix. Based on 10^5 repetitions for each parametrization. $s = 1$.

is about 35%, 57% and 8%, respectively. This ordering is well aligned in magnitude and ordering with that of ω^2 and σ^2/τ, loosely interpretable as a measure for the expected approximation quality. Still, considering the contributions of the parameterizations within combinations, we see substantial heterogeneity. For example, the TA parameterization contributes much less to $(R^*_{ql}, \widehat{\mathrm{R}}^*_{ql})$ than VNJ while the reverse is true for $(R_{ls}, \widehat{\mathrm{R}}_{ls})$.

Finally, we see substantial variation between the two data generating processes. While the range of areas under the curve of absolute errors aggregated over parameterizations for the generalized log-normal is $0.7pp$ to $10pp$, that for the over-dispersed Poisson is $3.2pp$ to $19.4pp$. The best performer for the generalized log-normal is, perhaps unsurprisingly, $(R^*_{ls}, \widehat{\mathrm{R}}_{ls})$. Intuitively, since the data generating process is log-normal, the asymptotic results would be exact for this combination if we plugged the true parameters into the frequency matrices. Just shy of these, we plug in the least squares parameter estimates which are maximum likelihood estimated. It is perhaps more surprising that using R_{ql}

is not generally a good idea for the over-dispersed Poisson data generating process even though the fact that these combinations take the bottom four slots is largely driven by the TA parametrization. Reassuringly, the top three performers across data generating processes also take the top three spots within data generating processes, albeit with a slightly changed ordering.

Figure 4.3a shows box plots for the size error at 5% nominal size computed over the three parameterizations and two data generating processes within combinations $(R, \widehat{\mathrm{R}})$ for $s = 1$. Positive errors indicate an over-sized, negative errors an under-sized test. In the plots, medians are indicated by blue lines inside the boxes. The boxes show the interquartile range. Whiskers represent the full range. The ordering is increasing in the sum of the absolute errors at 5% critical values from top to bottom.



(a) $s = 1$ (b) $s = 2$

Figure 4.3: Box plots of size error at 5% critical values over parameterizations (VNJ, TA, and BZ) and data generating processes (generalized log-normal and over-dispersed Poisson). Results for $s = 1$ shown in (a), for $s = 2$ in (b). Medians indicated by blue lines inside the boxes. The boxes show the interquartile range. Whiskers represent the full range.

Looking at the medians, we can see that these are close to zero, ranging from $-0.15pp$ to $0.37pp$. However, there is substantial variation in the interquartile range, $0.1pp$ for $(R_{ls}, \widehat{\mathrm{R}}^*_{ql})$

to $0.7pp$ for $(R^*_{ql}, \widehat{\mathrm{R}}_{ls})$, and range, $0.4pp$ for $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$ to $6.7pp$ for $(R_{ql}, \widehat{\mathrm{R}}_{ls})$). The best and worst performers from the analysis for the area under the curve and maximum absolute errors are still found in the top and bottom positions. Particularly the performance of $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$ seems close to perfection with a range from $-0.2pp$ to $0.2pp$.

Figure 4.3b is constructed in the same way as Figure 4.3a but for $s = 2$, halving the variance for the generalized log-normal and doubling the aggregate predictor for the over-dispersed Poisson data generating process. Theory tells us that the approximation quality should improve and this is indeed what we see. The medians move towards zero, now taking values between $-0.05pp$ and $0.14pp$, the largest interquartile range is now $1.1pp$ and the largest range $2.9pp$.

Overall, the combination $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$ performs very well across the considered parameterizations and data generating processes. This is not to say that we could not marginally increase performance in certain cases, for example by picking $(R^*_{ls}, \widehat{\mathrm{R}}_{ls})$ when the true data generating process is log-normal. However, even in this case in which we get the data generating process exactly right, not much seems to be gained in approximation quality where it matters most, namely in the tails relevant for testing. Thus it seems reasonable to simply use $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$ regardless of the hypothesized model, at least for size control.

4.5.3 Power

Having convinced ourselves that we can control size across a number of parameterizations, we show that the tests have good power. First, we consider how the power in finite sample approximations compares to power in the limiting distributions. Second, we investigate how power changes as the means become more dispersed based on the impact on the limiting distributions R_{GLN} and R_{ODP} alone as discussed in § 4.4.5.

4.5.3.1 Finite sample approximations under the alternative

We show that combinations of R-statistics and approximate limiting distributions $\widehat{\mathrm{R}}$ that do well for size control under the null hypothesis also do well when it comes to power at 5% critical values. The data generating processes are identical to those in § 4.5.2 and so

are the three considered parameterizations VNJ, BZ and TA. To avoid numerical issues, we again drop the perfectly fitted corner cells of the triangles without affecting the results.

To avoid confusion, we stress that we do not consider the impact of more dispersed means in this section. Thus, if we mention asymptotic results, we refer to large τ when the true data generating process is over-dispersed Poisson and for small ω^2 when it is generalized log-normal, holding the frequency matrix Π fixed.

For a given parametrization, we first find the asymptotic power. When the generalized log-normal model is the null hypothesis, we find the 5% critical values $c_{GLN} : P(\mathrm{R}_{GLN} > c_{GLN}) = 0.05$, using the true parameter values for Π. Then, we compute the power $P(\mathrm{R}_{ODP} > c_{GLN})$. Conversely, when the over-dispersed Poisson is the null model, we find $c_{ODP} : P(\mathrm{R}_{ODP} \leq c_{ODP}) = 0.05$ and compute the power $P(\mathrm{R}_{GLN} \leq c_{ODP})$. Lacking closed form solutions, we again use saddlepoint approximations, iteratively solving the equations for the critical values to a precision of 10^{-4}.

Next, we approximate the finite sample power of the top four combinations for size control in § 4.5.2, $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$, $(R^*_{ls}, \widehat{\mathrm{R}}_{ls})$, $(R_{ls}, \widehat{\mathrm{R}}_{ql})$ and $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ql})$, by the rejection frequencies under the alternative for $s = 1$. For example, say the generalized log-normal model is the null hypothesis and we want to compute the power for the combination $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$. Then, we first draw $B = 10^5$ triangles $\triangle_b$ from the over-dispersed Poisson data generating process. For each draw b, we find 5% critical values $c^{\widehat{\mathrm{R}}}_{GLN,ls,b} : P(\widehat{\mathrm{R}}^*_{GLN,ls,b} > c^{\widehat{\mathrm{R}}}_{GLN,ls,b}) = 0.05$. We compute these based on saddlepoint approximations, solving iteratively up to a precision of 10^{-4}. Then, we approximate the power as $B^{-1} \sum_{b=1}^{B} 1\{R^*_{GLN,ls,b} > c^{\widehat{\mathrm{R}}}_{GLN,ls,b}\}$. For the over-dispersed Poisson null hypothesis we proceed equivalently, using the left tail instead. In this way, we approximate power for all three parameterizations and all four combinations.

Before we proceed, we point out that we should be cautious to interpret power without taking into account the size error in finite samples. A test with larger than nominal size would generally have a power advantage purely due to the size error. One way to control for this is to consider size-adjusted power which levels the playing field by using critical values not at the nominal but at the true size. In our case, this would correspond to critical values from the true distribution of the test statistic R, rather than the approximated distribution $\widehat{\mathrm{R}}$. Therefore, the choice of $\widehat{\mathrm{R}}$ would not play a role anymore. To sidestep this issue, we take

H_0	DGP	Π	$P(\mathrm{R}_{GLN} \leq c_{ODP})$	$P(R \leq c^{\widehat{\mathrm{R}}}_{ODP}) - P(\mathrm{R}_{GLN} \leq c_{ODP})$ $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$	$(R^*_{ls}, \widehat{\mathrm{R}}_{ls})$	$(R_{ls}, \widehat{\mathrm{R}}_{ql})$	$(R^*_{ls}, \widehat{\mathrm{R}}^*_{ql})$
GLN	ODP	VNJ	99.02	0.25	0.14	0.18	0.22
		BZ	94.61	-0.94	-0.96	-0.97	-0.94
		TA	65.39	4.18	3.41	5.28	3.15
			$P(\mathrm{R}_{ODP} > c_{GLN})$	$P(R > c^{\widehat{\mathrm{R}}}_{ODP}) - P(\mathrm{R}_{ODP} > c_{GLN})$ $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$	$(R^*_{ls}, \widehat{\mathrm{R}}_{ls})$	$(R_{ls}, \widehat{\mathrm{R}}_{ql})$	$(R^*_{ls}, \widehat{\mathrm{R}}^*_{ql})$
ODP	GLN	VNJ	99.23	-0.30	-0.25	-0.34	-0.20
		BZ	94.67	-1.14	-1.15	-1.12	-1.12
		TA	64.73	0.81	0.15	4.68	2.49

Table 4.2: Power in % at 5% critical values for large τ (over-dispersed Poisson DGP) and small ω^2 (generalized log-normal DGP) along with the power gap in *pp* for the top four performers from Table 4.3. DGP is short for data generating process. Based on 10^5 repetitions. $s = 1$.

a different approach and compare how close the power of the finite sample approximations matches the asymptotic power.

Table 4.2 shows the asymptotic power and the gap between power in finite sample approximations and asymptotic power. Looking at the asymptotic power first, we can see little variation between data generating processes within parameterizations. The power is highest for the VNJ parameterization with 99%, followed by BZ with 95% and TA with 65%. This ordering aligns with that of the standard deviations of the frequencies π_{ij} under these parameterizations which are given by 0.016, 0.012 and 0.009 for VNJ, BZ and TA, respectively.

When considering the finite sample approximations, we see that their power is relatively close to the asymptotic power. For VNJ absolute deviations range from 0.14*pp* to 0.34*pp*, and for BZ from 0.94*pp* and 1.15*pp*. Compared to that, discrepancies for the TA parameterization are larger. The smallest discrepancy of 0.14*pp* arises for $(R^*_{ls}, \widehat{\mathrm{R}}_{ls})$ when the data generating process is generalized log-normal. As before, this is intuitive since it corresponds to plugging maximum likelihood estimated parameters $\hat{\xi}^{(2)}$ into Π. With 5.28*pp* the largest discrepancy arises for $(R_{ls}, \widehat{\mathrm{R}}_{ql})$ for an over-dispersed Poisson data generating process. Mean absolute errors across parameterizations and data generating processes are rather close, ranging from 1.01*pp* for $(R^*_{ls}, \widehat{\mathrm{R}}_{ls})$ to 2.1*pp* for $(R_{ls}, \widehat{\mathrm{R}}_{ql})$. Our proposed favorite from above $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$ comes in second with 1.27*pp*. We would argue that we can still justify

the use of $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$ regardless of the data generating process.

4.5.3.2 Increasing mean dispersion in R_{ODP} and R_{GLN}

We consider the impact of more dispersed means on power based on the the test statistics' limiting distributions R_{GLN} and R_{ODP}. We show that the power grows quickly as we move from identical means across cells to a scenario where a single frequency hits zero.

For a given diagonal frequency matrix Π with values π_{ij}, we define the linear combination

$$\Pi_{(t)} = t\Pi + (1-t)n^{-1}I_n.$$

Thus, for $t = 1$ we recover Π while for $t = 0$ we are in a setting where all cells have the same frequencies so all means are identical. In the latter scenario, R_{GLN} and R_{ODP} collapse to a point-mass at n, as discussed in §4.4.4. We consider t ranging from just over zero to just under $t^{max} : t^{max} \min_{ij\in\mathcal{I}}(\pi_{ij}) + (1-t^{max})n^{-1} = 0$. The significance of t^{max} is that $\Pi_{(t^{max})}$ corresponds to the matrix where the smallest frequency is exactly zero.

For each t, we approximate one-sided 5% critical values of $\mathrm{R}^{(t)}_{GLN} = \mathrm{R}_{GLN}(\Pi_{(t)})$ and $\mathrm{R}^{(t)}_{ODP} = \mathrm{R}_{ODP}(\Pi_{(t)})$ through

$$c^{(t)}_{GLN} : \widehat{P}^{SP}(\mathrm{R}^{(t)}_{GLN} > c^{(t)}_{GLN}) = 0.05 \quad \text{and} \quad c^{(t)}_{ODP} : \widehat{P}^{SP}(\mathrm{R}^{(t)}_{ODP} \leq c^{(t)}_{ODP}) = 0.05.$$

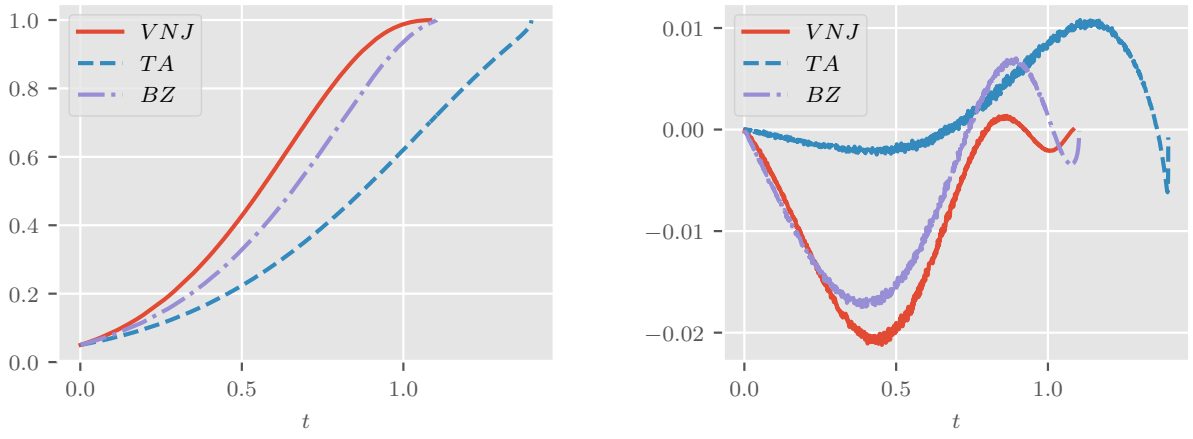
We iteratively solve the equations up to a precision of 10^{-4}. Theorem 4.3 tells us that the critical values should grow for both models, but that $c^{(t)}_{GLN}$ converges as t approaches t^{max} while $c^{(t)}_{ODP}$ goes to infinity.

Then, for given t and critical values, we find the power when the null model is generalized log-normal $P(\mathrm{R}^{(t)}_{ODP} > c^{(t)}_{GLN})$ and when the null model is over-dispersed Poisson $P(\mathrm{R}^{(t)}_{GLN} \leq c^{(t)}_{ODP})$. Again, we use saddlepoint approximation. Based on Theorem 4.3, we should see the power go to unity as t approaches t^{max}.

We consider the same parameterizations VNJ, TA and BZ of frequency matrices Π and design matrices X as above. The values for t^{max} are 1.083 for VNJ, 1.396 for TA and 1.103 for BZ. To avoid numerical issues, we again drop the perfectly fitted corner cells from the triangles. In this case, while the power is not affected, the critical values are scaled down by the ratio of $\hat{\tau}_{ls}$ computed over the smaller array without corner cells to

that computed over the full triangle. Since this is merely proportional, the results are not affected qualitatively.

Figure 4.4a shows the power when the generalized log-normal model is the null hypothesis. For all considered parameterizations, this is close to 5% for t close to zero, increasing monotonically with t, and approaching unity as t approaches t^{max} as expected. For $t = 1$,



(a) $P(\mathrm{R}^{(t)}_{ODP} > c^{(t)}_{GLN})$ (b) $P(\mathrm{R}^{(t)}_{ODP} > c^{(t)}_{GLN}) - P(\mathrm{R}^{(t)}_{GLN} \leq c^{(t)}_{ODP})$

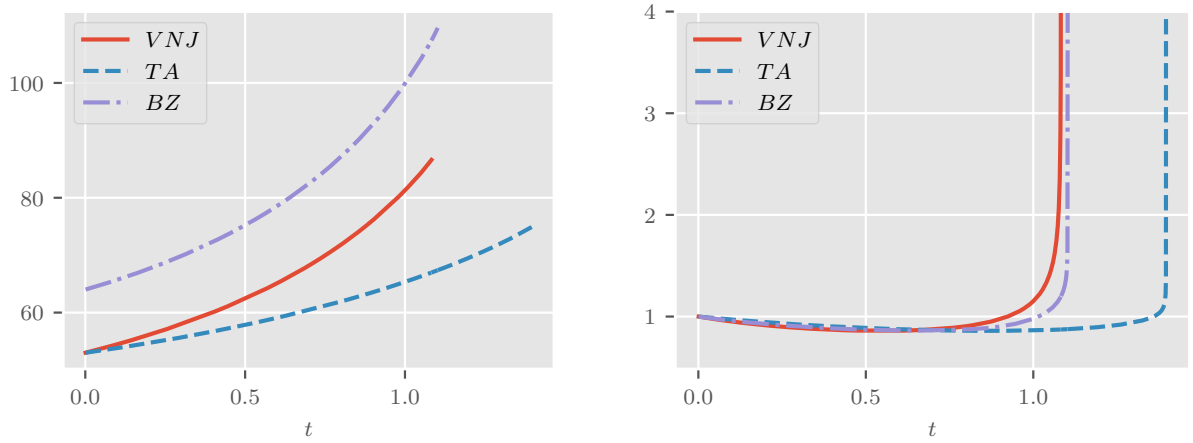
Figure 4.4: Power as t increases from 0 to t^{max}. Values for t^{max} are 1.083 for VNJ, 1.396 for TA and 1.103 for BZ. (a) shows power when the null model is generalized log-normal, (b) shows the difference in power between the two models.

where $\Pi_{(t)}$ corresponds to the least squares estimated frequencies from the data, the power matches what we found in Table 4.2.

Figure 4.4b shows the difference in power between the two models plotted over t. For the three settings we consider, these curves have a similar shape and start and end at zero. Generally, the power is very comparable, with differences between $-2pp$ to $1pp$ again matching our findings from Table 4.2 for $t = 1$.

Figure 4.5a shows the one-sided 5% critical values $c^{(t)}_{GLN}$ plotted over t. As expected, these are increasing for all settings. Figure 4.5b show the ratio of the critical values $c^{(t)}_{ODP}$ to $c^{(t)}_{GLN}$. This starts at unity, initially decreases then increases and finally explodes towards infinity as we approach t^{max}.

Taking the plots together, we get the following interpretation. We recall that the two



(a) $c^{(t)}_{GLN} : P(\mathrm{R}^{(t)}_{GLN} > c^{(t)}_{GLN})$

(b) $c^{(t)}_{ODP}/c^{(t)}_{GLN}$

Figure 4.5: Critical values as t increases. (a) shows critical values of $\mathrm{R}^{(t)}_{GLN}$, (b) the ratio of critical values of $\mathrm{R}^{(t)}_{ODP}$ to $\mathrm{R}^{(t)}_{GLN}$.

distributions are identical for $t = 0$. Further, the rejection regions for the generalized log-normal null is the upper tail while the lower tail is relevant for the over-dispersed Poisson model. However, for small t, the mass of both $\mathrm{R}^{(t)}_{GLN}$ and $\mathrm{R}^{(t)}_{ODP}$ is highly concentrated around n and the distributions are quite similar. This explains why the power is initially close to 5% for either. Further, due to the concentration $c^{(t)}_{GLN}$ and $c^{(t)}_{ODP}$ are initially close. As t increases, both distributions become more spread out and move up the real line, with $\mathrm{R}^{(t)}_{ODP}$ moving faster than $\mathrm{R}^{(t)}_{GLN}$. This is reflected in the increase in power. Initially $c^{(t)}_{GLN}$ increases faster than $c^{(t)}_{ODP}$ so their ratio decreases. Yet, for t large enough, $c^{(t)}_{ODP}$ overtakes $c^{(t)}_{GLN}$, indicating the point at which power reaches 95% for either model. The power differential is necessarily zero at this point. Finally, $c^{(t)}_{ODP}$ explodes while $c^{(t)}_{GLN}$ converges as t approaches t^{max}, so the ratio diverges.

4.6 Empirical applications

We consider a range of empirical examples. First, we revisit the empirical illustration of the problem from the beginning of the paper in §4.2. We show that the proposed test favors the over-dispersed Poisson model over the generalized log-normal model. Second, we consider an example that perhaps somewhat cautions against starting off with a model that may be

	H_0 : generalized log-normal				H_0 : over-dispersed Poisson			
	R_{ls}	R_{ql}	R^*_{ls}	R^*_{ql}	R_{ls}	R_{ql}	R^*_{ls}	R^*_{ql}
$\widehat{\mathrm{R}}_{ls}$	0.43	0.39	0.14	0.27	8.53	9.00	14.59	10.89
$\widehat{\mathrm{R}}_{ql}$	0.32	0.29	0.10	0.19	11.80	12.40	19.35	14.79
$\widehat{\mathrm{R}}^*_{ls}$	0.35	0.32	0.11	0.22	10.42	10.97	17.34	13.14
$\widehat{\mathrm{R}}^*_{ql}$	0.38	0.34	0.13	0.24	9.48	9.99	15.96	12.01

Table 4.3: p-values in % for the Verrall et al. (2010) data.

misspecified to begin with: dropping a clearly needed calendar effect turns the results of the encompassing tests upside down. Third, taking these insights into account, we implement a testing procedure that makes use of a number of recent results: deciding between over-dispersed Poisson and generalized log-normal model, evaluating misspecification, and testing for the need of a calendar effect.

4.6.1 Empirical illustration revisited

We revisit the data in Table 4.1 discussed in §4.2 and show that we can reject that the (generalized) log-normal model encompasses the over-dispersed Poisson model but cannot reject the alternative direction. Thus, the encompassing tests proposed in this paper have higher power to distinguish between the two models than the misspecification tests Harnau (2018*b*) applied to these data. We remark that the encompassing tests were designed explicitly to distinguish between the two models, in contrast to the more general misspecification tests.

Table 4.3 shows p-values for all 16 combinations of R-statistics and $\widehat{\mathrm{R}}$ under both null hypotheses. Computing the four R-statistics in (4.10) yields

$$R_{ls} = 104.87, \quad R_{ql} = 105.61, \quad R^*_{ls} = 113.19 \quad \text{and} \quad R^*_{ql} = 108.39.$$

Thus, while not identical, the test statistics appear quite similar.

First, we consider the generalized log-normal model as the null model so

$$H_0 : \text{generalized log-normal} \quad \text{vs} \quad H_A : \text{over-dispersed Poisson}.$$

This is consistent with the applications in Kuang et al. (2015) and Harnau (2018*b*) who consider these data in a log-normal model. Looking at our preferred combination $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$,

we find a p-value of 0.001, rejecting the model. Reassuringly, we reject the generalized log-normal model for any combination of R and $\widehat{\mathrm{R}}$. The most favorable impression to this null hypothesis is given by $(R_{ls}, \widehat{\mathrm{R}}_{ls})$ with a p-value of 0.004.

If we instead take the over-dispersed Poisson model as the null so

$$H_0 : \text{over-dispersed Poisson} \quad \text{vs} \quad H_A : \text{generalized log-normal},$$

the model cannot be rejected with a p-value of 0.17 for $(R^*_{ls}, \widehat{\mathrm{R}}^*_{ls})$. Again, this decision is quite robust to the choice of estimators with a least favorable p-value of 0.09 obtained based on $(R_{ls}, \widehat{\mathrm{R}}^*_{ql})$. If we accept the null, we can evaluate the power against the generalized log-normal model. For instance, the 5% critical value under the over-dispersed Poisson model is 95.7. The probability of drawing a value smaller than that from the generalized log-normal model is 0.99. Thus, the power at the 5% critical value is close to unity. We can also find the power at the value taken by R^*_{ls}, interpretable as the 17% critical value if we like. This is simply one minus the p-value of the generalized log-normal model, thus equal to $1 - 0.001 = 0.999$.

4.6.2 Sensitivity to invalid model reductions

The Barnett & Zehnwirth (2000) data are known to require a calendar effect for modeling. We show those data in Table 4.5 in the appendix. Barnett & Zehnwirth (2000), Kuang et al. (2015) and Harnau (2018*b*) approached these data set using log-normal models. Here, we find that an encompassing test instead heavily favors an over-dispersed Poisson model. Further, we show that dropping the needed calendar effect substantially affects the test results.

We again first consider a generalized log-normal model, however, we initially allow for a calendar effect. Adding the prefix "extended" to models with calendar effect, we test

$$H_0 : \text{extended generalized log-normal} \quad \text{vs} \quad H_A : \text{extended over-dispersed Poisson}.$$

Our preferred test statistic $R^*_{ls} = 162.2$. Paired with $\widehat{\mathrm{R}}^*_{ls}$, this yields a p-value of 0.0002. Thus, the generalized log-normal model is clearly rejected. For illustrative purposes we

continue anyway and test whether we can drop the calendar effect from the generalized log-normal model. Thus, the hypothesis is

$$H_0 : \text{generalized log-normal} \quad \text{vs} \quad H_A : \text{extended generalized log-normal.}$$

Kuang & Nielsen (2018) show that for small ω^2 we can use a standard F-test for this purpose. If we assumed that the data generating process is not generalized log-normal but log-normal, the F-test would be exact. This test rejects the reduction with a p-value of 0.00. If again we decide to continue anyways, we can now test the generalized log-normal against an over-dispersed Poisson model, both without calendar effect. Thus, the hypothesis is

$$H_0 : \text{generalized log-normal} \quad \text{vs} \quad H_A : \text{over-dispersed Poisson.}$$

Interestingly, the log-normal model does not look so bad anymore now. For this model, $R^*_{ls} = 113.2$ which yields a p-value of 0.10. Of course, this should not encourage us to assume that the generalized log-normal model without calendar effect is actually a good choice. Rather, it draws attention to the fact that tests computed on inappropriately reduced models may yield misleading conclusions. The tests proposed in this paper assume that the null model is well specified and the results are generally only valid if this is correct. In applications, we may relax this statement to "the tests only give useful indications if the null model describes the data well". In this case, we did not only ignore the initial rejection of the generalized log-normal model, but also that calendar effects are clearly needed to model the data well.

We now start over, switching the role of the two models thus starting with an extended over-dispersed Poisson model. The first hypothesis is the mirror image from above:

$$H_0 : \text{extended over-dispersed Poisson} \quad \text{vs} \quad H_A : \text{extended generalized log-normal.}$$

The test statistic is still $R^*_{ls} = 162.2$, but now we cannot reject the null hypothesis with a p-value of 0.56. We may thus feel comfortable to model the data using an over-dispersed Poisson model with a calendar effect. Next, we investigate whether the calendar effect can be dropped, testing

$$H_0 : \text{over-dispersed Poisson} \quad \text{vs} \quad H_A : \text{extended over-dispersed Poisson.}$$

Harnau & Nielsen (2017) showed that for large τ, this can be done with an F-test based on Poisson deviances. This reduction is clearly rejected, again with a p-value of 0.00. We move on anyways, drop the calendar effect, and test

$$H_0 : \text{over-dispersed Poisson} \quad \text{vs} \quad H_A : \text{generalized log-normal.}$$

In this case, the p-value is 0.01 and we reject the null so we get the opposite result.

Comparing the outcomes of the tests, it seems clear that an over-dispersed Poisson model with calendar effect is the most reasonable choice. However, if we had not started at this point, but rather never considered a calendar effect in the first place, we might have come to a very different conclusion. This indicates that the starting point can matter a great deal for the model choice and that it may be a good idea to start with a more general model and test for reductions, even if we were fairly certain that the reduced model is a good choice.

4.6.3 A general to specific testing procedure

The Taylor & Ashe (1983) data have frequently been modeled as over-dispersed Poisson, for example by England & Verrall (1999), England (2002) and Harnau (2018*b*). We provide those data in Table 4.4 in the appendix. Based on the insight from the application to the Barnett & Zehnwirth (2000) data above, we start with a general model with calendar effect and use a whole battery of tests to see if a generalized log-normal or over-dispersed Poisson chain-ladder model can be justified. We find that an over-dispersed Poisson chain-ladder model is reasonable for these data.

We first consider a generalized log-normal model with calendar effect. We test

$$H_0 : \text{extended generalized log-normal} \quad \text{vs} \quad H_A : \text{extended over-dispersed Poisson.}$$

The null hypothesis is clearly rejected with a test statistic of $R^*_{ls} = 81.5$ and a p-value of 0.001. Thus, we do not proceed further with this model.

Instead, we now start with an over-dispersed Poisson model with calendar effect. The hypothesis

$$H_0 : \text{extended over-dispersed Poisson} \quad \text{vs} \quad H_A : \text{extended generalized log-normal}$$

cannot be rejected with a p-value of 0.92. We point out that this indicates that the draw is in the right tail of $\widehat{\mathrm{R}}_{ODP}$. While we would argue this is not the case here, we may worry about values that are too far out in the right tail of $\widehat{\mathrm{R}}_{ODP}$ which would perhaps indicate that we should reject both models.

Next, we apply the misspecification tests by Harnau (2018*b*). We first split the run-off triangle into four sub-samples as indicated in Figure 4.6. We can now test whether the

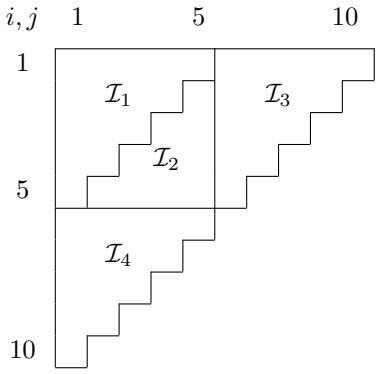


Figure 4.6: Split of a run-off triangle into four sub-samples as in Harnau (2018*b*).

over-dispersion is common across sub-samples:

$$H_0 : \sigma_\ell^2 = \sigma^2.$$

Harnau (2018*b*) showed that we can use a Bartlett test based on the Poisson deviance for this purpose. In the model with calendar effect, this test yields a p-value of just above 0.05, a rather close call. In light of the fact that the ultimate goal of the exercise is forecasting reserves and that forecasting often benefits from simpler models we decide to accept the hypothesis. Next, we consider the hypothesis that there are no breaks in accident, development, and calendar effects between sub-samples:

$$H_0 : \alpha_{i,\ell} + \beta_{j,\ell} + \gamma_{k,\ell} + \delta_\ell = \alpha_i + \beta_j + \gamma_k + \delta.$$

As demonstrated by Harnau (2018*b*), this can be tested with a deviance based F-test that is independent of the Bartlett tests for large τ. This test yields a p-value of 0.07. Based on the same argument as above, we accept the hypothesis.

Now that we are reasonably happy with the over-dispersed Poisson extended chain-ladder model, we test whether the calendar effects can be dropped.

$$H_0 : \text{over-dispersed Poisson} \quad \text{vs} \quad H_A : \text{extended over-dispersed Poisson.}$$

Based on an F-test, this hypothesis cannot be rejected with a p-value of 0.30. Thus, we move on, retesting whether the over-dispersed Poisson model still encompasses the log-normal model.

$$H_0 : \text{over-dispersed Poisson} \quad \text{vs} \quad H_A : \text{generalized log-normal.}$$

Based on a test statistic $R^*_{ls} = 73.5$, this cannot be rejected with a p-value of 0.73. We can now go back and apply the misspecification tests by Harnau (2018*b*) once again, except this time for models without calendar effect. Using the same sub-sample structure, a Bartlett test cannot reject the hypothesis of common over-dispersion $H_0 : \sigma^2_\ell = \sigma^2$ with a p-value of 0.08. Further, an F-test for the hypothesis of the absence of breaks in the mean parameters $H_0 : \alpha_{i,\ell} + \beta_{j,\ell} + \delta_\ell = \alpha_i + \beta_j + \delta$ cannot be rejected with a p-value of 0.93.

In conclusion, an over-dispersed Poisson chain-ladder model for the Taylor & Ashe (1983) data survived a whole battery of specification tests and we may at least be more comfortable with this model choice, having found no strong evidence telling us otherwise. In contrast, the generalized log-normal model was clearly rejected.

4.7 Discussion

While there has been a range of recent advances for both over-dispersed Poisson and (generalized) log-normal models, there are still several areas left for further research. This spans from further misspecification tests and refinements thereof over a potential theory for the bootstrap to empirical studies evaluating the impact of the theoretical procedures in practice.

As pointed out by Harnau (2018*b*), the misspecification tests require a specific choice for the number of sub-samples and their shape. A generalization that is agnostic about these choices would be desirable. Harnau (2018*b*) also remarked that a misspecification test for

independence would be useful. The assumption of independence across cells is common to both over-dispersed Poisson and generalized log-normal models. It seems likely that a test that is valid in one model would translate easily to the other.

The closed form distribution forecasts proposed by Harnau & Nielsen (2017) for the over-dispersed Poisson model and by Kuang & Nielsen (2018) for the generalized log-normal model are both based on t-distributions and thus symmetric. These forecasts seem to perform rather well and, in some settings, appear more robust than the bootstrap by England & Verrall (1999) and England (2002). However, with an appealing asymptotic theory in place for both types of models, it may be worth considering whether a theory for the bootstrap could be developed to allow for potential asymmetry of the forecast distribution that we might expect in finite samples.

Finally, given the range of recent theoretical developments, an empirical study that evaluates the impact of the contributions in applications seems appropriate. Since the main concern in claim reserving is forecasting, such a study would likely require data not just for run-off triangles but also for the realized values in the forecast array, that is the lower triangle. Such data is available, for example, from the Casualty Actuarial Society (2011). For instance, it would be interesting to see how the forecast performance between rival models differs if one was rejected by the theory but not the other.

Appendix

Proof of Corollary 4.1

The proof is similar to that for the distribution of the mixed parameter in Harnau & Nielsen (2017, Lemma 1). We define the Poisson quasi-likelihood estimator $g(Y) = \hat{\xi}_{ql}$ for $Y = (Y_{ij} : i, j \in \mathcal{I})$. We make use of the fact that $g(Y) = (\hat{\mu}_{11}, \hat{\xi}_{ql}^{(2)'})'$ and $g(Y/\tau) = \{\hat{\mu}_{11} - \log(\tau), \hat{\xi}_{ql}^{(2)'}\}'$ for identical $\hat{\mu}_{11}$ and $\hat{\xi}_{ql}^{(2)}$. This follows from the Poisson score equation $\sum_{ij\in\mathcal{I}} Y_{ij}x_{ij} = \exp(\hat{\mu}_{11}) \sum_{ij} x_{ij} \exp(x_{ij}^{(2)'} \hat{\xi}_{ql}^{(2)})$ in which replacing Y_{ij} by Y_{ij}/τ goes hand in hand with replacing $\hat{\mu}_{11}$ by $\hat{\mu}_{11} - \log(\tau)$. Thus, only $\hat{\mu}_{11}$ is affect by scaling the variables. By Johansen (1979, Theorem 7.1), $g(.)$ is Fisher consistent so $g\{E(Y)\} = (\mu_{11}, \xi^{(2)'})'$ and

$g\{E(Y)/\tau\} = (\mu_{11} - \log(\tau), \xi^{(2)'})$. By Johansen (1979, Lemma 7.2), $\partial g/\partial Y|_{Y=\{E(Y)/\tau\}} = (X'\Pi X)^{-1}X'$.

By independence of Y_{ij}, (4.4) generalizes to $\tau^{1/2}\{Y/\tau - E(Y)/\tau\} \xrightarrow{D} N(0, \sigma^2\Pi)$. Applying $g(.)$ to this result using the δ-method and taking into account that $g(Y/\tau) - g\{E(Y)/\tau\} = \hat{\xi}_{ql} - \xi$ yields the desired result. For the δ-method see, for example, Casella & Berger (2002, Theorem 5.5.24); to avoid confusion, we point out that in our notation, the sequence is not over n but rather over τ (for proofs relating to the generalized log-normal model below the sequence is over ω^2).

Proof of Lemma 4.1

First, we show that $\lim_{\tau\to\infty} P(Y_{ij} = 0) = 0$. Recall that $Y_{ij}/\tau \xrightarrow{P} \pi_{ij} > 0$. Thus, $\lim_{\tau\to\infty} P(Y_{ij}/\tau = 0) = 0$. Since $P(Y_{ij} = 0) = P(Y_{ij}/\tau = 0)$ the results follows. Next, we know from (4.4) that $\sqrt{\tau}(Y_{ij}/\tau - \pi_{ij}) \xrightarrow{D} N(0, \sigma^2\pi_{ij})$. For $Y_{ij} > 0$ we can employ the δ-method to apply $\log(.)$ to Y_{ij}/τ. The result follows since $\log(Y_{ij}/\tau) - \log(\pi_{ij}) = \log(Y_{ij}) - \log(\tau) - [\log\{\exp(\mu_{ij})\} - \log(\tau)]$ and $\partial \log(x)/\partial x|_{x=E(Y_{ij}/\tau)} = \pi_{ij}^{-1}$.

Proof of Lemma 4.2

Define $Z = (Z_{ij} : i, j \in \mathcal{I})$ and $\mu = (\mu_{ij} : i, j \in \mathcal{I})$. Taking into account the independence, the multivariate version of Lemma 4.1 is $\sqrt{\tau}(Z - \mu) \xrightarrow{D} N(0, \sigma^2\Pi^{-1})$.

To obtain the distribution of the least-squares estimator, we pre-multiply by $(X'X)^{-1}X'$. This yields $\sqrt{\tau}(\hat{\xi}_{ls} - \xi) \xrightarrow{D} N(0, \sigma^2\Omega)$ with Ω as defined in the lemma. We find the distribution of the residual sum of squares using the continuous mapping theorem. With that, $\tau Z'MZ = \{\sqrt{\tau}(Z - \mu)\}'M\{\sqrt{\tau}(Z - \mu)\} \xrightarrow{D} U'\Pi^{-1/2}M\Pi^{-1/2}U$ for $U \overset{D}{=} N(0, I)$.

Finally, we show that $\hat{\tau}_{ls}/\tau \xrightarrow{P} 1$ where $\hat{\tau}_{ls} = \sum_{ij\in\mathcal{I}} \exp(x'_{ij}\hat{\xi}_{ls})$. Let $f(\xi) = \sum_{ij\in\mathcal{I}} \exp(x'_{ij}\xi)$. For this map, with $\xi = (\mu_{11}, \xi^{(2)'})'$ and defining $\xi^\tau = (\mu_{11} - \log(\tau), \xi^{(2)'})'$, we have $f(\xi^\tau) = f(\xi)/\tau$. Further by the equivalent argument made in the Proof of Corollary 4.1, subtracting $\log(\tau)$ from Z (element-wise) affects only the estimate for the intercept μ_{11}. That is, for the least squares estimator $g(Z) = (\hat{\mu}_{11}^{ls}, \hat{\xi}_{ls}^{(2)'})'$ and with $Z^\tau = \{Z_{ij} - \log(\tau) : i, j \in \mathcal{I}\}$, we have $g(Z^\tau) = (\hat{\mu}_{11}^{ls} - \log(\tau), \hat{\xi}_{ls}^{(2)'})' =: \hat{\xi}_{ls}^\tau$. Thus, $\hat{\xi}_{ls}^\tau - \xi^\tau = \hat{\xi}_{ls} - \xi$ and $\sqrt{\tau}(\hat{\xi}_{ls}^\tau - \xi^\tau) \xrightarrow{D} N(0, \sigma^2\Omega)$.

Now, we apply $f(.)$ by the δ-method to get that $\sqrt{\tau}\{f(\hat{\xi}_{ls}^{\tau}) - f(\xi^{\tau})\} = O_p(1)$. Since $f(\hat{\xi}_{ls}^{\tau}) - f(\xi^{\tau}) = \hat{\tau}_{ls}/\tau - 1$ it follows that $\hat{\tau}_{ls}/\tau = 1 + o_p(1)$.

Proof of Lemma 4.3

Define the vector $Y = (Y_{ij} : i, j \in \mathcal{I})$ and let $\exp(\mu) = \{\exp(\mu_{ij}) : i, j \in \mathcal{I}\}$. Then, using the independence, (4.4) generalizes to $\tau^{-1/2}\{Y - \exp(\mu)\} \xrightarrow{D} N(0, \sigma^2\Pi)$.

Harnau & Nielsen (2017, Lemma 1) derive the asymptotic distribution of the Poisson quasi-likelihood estimator $\hat{\xi}_{ql}$ through the δ-method. Following Johansen (1979, Theorems 7.1, 7.3, Lemma 7.2), they show that the mapping $Y \mapsto \hat{\xi}_{ql}$ estimator is asymptotically equivalent to the linear mapping $Y \mapsto (X'\Pi X)X'\Pi^{1/2}Y$.

Meanwhile, the weighted least squares estimator maps $Z \mapsto (X'\Pi X)^{-1}X'\Pi^{1/2}\Pi^{1/2}Z = \hat{\xi}^*$. Thus, the only non-linear component of the mapping $Y \mapsto \hat{\xi}^*$ is the transformation from Y to Z. However, while this mapping is non-linear in finite sample, for large τ it is equivalent to the linear map from Y to $\Pi^{-1/2}Y$ as seen in the proof of Lemma 4.1. Asymptotically, this conforms to sequentially applying the transformations $\Pi^{-1/2}$ followed by $(X'\Pi X)^{-1}X'\Pi^{1/2}\Pi^{1/2}$ to Y. Taken together, the map reduces to $(X'\Pi X)X'\Pi^{1/2}$. Thus, both the Poisson quasi-likelihood and weighted least squares mapping asymptotically apply the same transformation to $\tau^{-1/2}\{Y - \exp(\mu)\} \xrightarrow{D} N(0, \sigma^2\Pi)$. Thus, $\sqrt{\tau}(\hat{\xi}^* - \hat{\xi}_{ql}) \xrightarrow{P} 0$.

The proof for $\tau RSS^* - D \xrightarrow{P} 0$ follows by the same argument. The main insight is that the asymptotic distribution of the Poisson deviance is asymptotically equivalent to that of the quadratic form $\tau^{-1}\{Y - \exp(\mu)\}'(\Pi - X'(X'\Pi X)^{-1}X')\{Y - \exp(\mu)\}$, as Harnau & Nielsen (2017, Proof of Lemma 1) show building on Johansen (1979, Theorem 7.7, Lemma 7.8). This is again asymptotically identical to the sequential mapping from Y to Z followed by the map from Z to the scaled residual sum of weighted least squares $\tau RSS^* = \{\sqrt{\tau}(Z - \mu)\}'M^*\{\sqrt{\tau}(Z - \mu)\}$.

To show that we can replace the weight matrix Π in the weighted least squares estimator by $\widehat{\Pi}_{ls}$ or $\widehat{\Pi}_{ql}$ we note that both matrices converge in probability to Π and then apply Slutsky's theorem (Casella & Berger 2002, Theorem 5.5.17). Combining this argument with the proof of the equivalence of D and RSS^* in the last paragraph, it also follows that

we can replace the weights in RSS^* without affecting the result. Finally, both $\hat{\tau}_{ls}/\tau \xrightarrow{P} 1$ and $\hat{\tau}_{ql}/\tau \xrightarrow{P} 1$ so we can replace τ as well by Slutsky's theorem.

Proof of Theorem 4.1

Taking into account the results from Lemma 4.3, it follows that $(\tau RSS^*)/D \xrightarrow{P} 1$ and that the result still holds if we replace RSS^* by RSS^*_{ls} or RSS^*_{ql} and τ by $\hat{\tau}_{ls}$ or $\hat{\tau}_{ql}$. Thus, for example, $R_{ls} \xrightarrow{P} R_{ql}$ so their difference vanishes and similarly for any other of the six total combinations.

Both $RSS = Z'MZ$ and $RSS^* = Z'M^*Z$ are quadratic forms in the same random vector Z. It follows from the proofs of Lemma 4.2 and Lemma (4.3) that $\tau RSS \xrightarrow{D} U'\Pi^{-1/2}M\Pi^{-1/2}U$ and $\tau RSS^* \xrightarrow{D} U'M^*U$ for the same $U \overset{D}{=} N(0, I_n)$. The distribution of RSS/RSS^* follows by the continuous mapping theorem. Since $\tau RSS^* - D \xrightarrow{P} 0$ as in Lemma 4.3, $\tau RSS^*/D \xrightarrow{P} 1$ so that $RSS/RSS^* - \tau RSS/RSS^* \xrightarrow{P} 0$ follows.

We can replace τ by $\hat{\tau}_{ql}$ since Harnau & Nielsen (2017, Theorem 2) gives us that $\tau/\hat{\tau}_{ql} \xrightarrow{P} 1$. From Lemma 4.2, $\hat{\tau}_{ls}/\tau \xrightarrow{P} 1$. Further, both $\Pi(\hat{\xi}^{(2)}_{ls})$ and $\Pi(\hat{\xi}^{(2)}_{ql})$ converge to Π in probability. Then, by Slutsky's theorem, we can replace the true parameters with their estimates in $\tau RSS/D$ and RSS/RSS^* without affection the limiting distribution.

Proof of Lemma 4.4

The asymptotic distribution of the weighted least squares estimators follows by the same argument as in Lemma 4.2, except now taking $(\omega^2)^{-1/2}(Z - \mu) \xrightarrow{D} N(0, I_n)$, as shown by Kuang & Nielsen (2018, Theorem 3.3), as a starting point.

The asymptotic equivalence of weighted least squares and Poisson quasi likelihood estimation follows from the same argument as in Lemma 4.3, except now $(\omega^2)^{-1/2}\{Y - \exp(\mu)\} \xrightarrow{D} N[0, \mathrm{diag}\{\exp(\mu)\}]$. The argument for replacing true parameters in the frequency matrix Π and the aggregate means τ by estimates is identical to that in Lemma 4.3 as well.

Proof of Theorem 4.2

This follows by the same argument as the proof for Theorem 4.1 above, except now combining the asymptotic distribution of the least squares estimator in the generalized log-normal model from Kuang & Nielsen (2018, Theorem 3.5) in (4.8) and Lemma 4.4.

Proof of Lemma 4.5

First, we show that R_{GLN} and R_{ODP} share a common support. If we recall that

$$\mathrm{R}_{GLN}(U) = \frac{U'MU}{U'\Pi^{1/2}M^*\Pi^{1/2}U} \quad \text{and} \quad \mathrm{R}_{ODP}(U) = \frac{U'\Pi^{-1/2}M\Pi^{-1/2}U}{U'M^*U},$$

the main insight is that $\mathrm{R}_{GLN}(\Pi^{-1/2}U) = \mathrm{R}_{ODP}(U)$. Formally, both $\mathrm{R}_{GLN} : \mathbb{R}^n \mapsto \mathbb{R}$ and $\mathrm{R}_{ODP} : \mathbb{R}^n \mapsto \mathbb{R}$ are random variables on $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P\}$ where P is the measure associated with $N(0, I_n)$. We now show that $P\{\mathrm{R}_{GLN} \in \mathcal{S}\} = 1$ implies that $P\{\mathrm{R}_{ODP} \in \mathcal{S}\} = 1$; the opposite direction follows. Notationally, $P\{\mathrm{R}_{GLN} \in \mathcal{S}\} = P\{u \in \mathbb{R}^n : \mathrm{R}_{GLN}(u) \in \mathcal{S}\}$ and, for some set $\mathcal{A} \subseteq \mathbb{R}$ we denote by $\mathrm{R}^{-1}_{GLN}(\mathcal{A})$ the pre-image $\{u \in \mathbb{R}^n : \mathrm{R}_{GLN}(u) \in \mathcal{A}\}$. Now, R_{GLN} is measurable since it is continuous almost everywhere, the exception being the measure zero set where the denominator is zero. Thus,

$$P\{\mathrm{R}_{GLN} \in \mathcal{S}\} = P\{U \in \mathrm{R}^{-1}_{GLN}(\mathcal{S})\}.$$

Since the support of $N(0, I_n)$ is $\mathbb{R}^n$, we must have that $\mathrm{R}^{-1}_{GLN}(\mathcal{S}) = \mathbb{R}^n$ and hence $\mathcal{S} = \mathrm{R}_{GLN}(\mathbb{R}^n)$. Since Π is invertible, $\Pi^{-1/2}\mathbb{R}^n = \mathbb{R}^n$ so that $\mathrm{R}_{GLN}(\mathbb{R}^n) = \mathrm{R}_{GLN}(\Pi^{-1/2}\mathbb{R}^n) = \mathrm{R}_{ODP}(\mathbb{R}^n)$. Thus, $\mathrm{R}_{ODP}(\mathbb{R}^n) = \mathcal{S}$ and so $\mathrm{R}^{-1}_{GLN}(\mathcal{S}) = \mathrm{R}^{-1}_{ODP}(\mathcal{S})$. Taken together,

$$P\{\mathrm{R}_{GLN} \in \mathcal{S}\} = P\{U \in \mathrm{R}^{-1}_{GLN}(\mathcal{S})\} = P\{U \in \mathrm{R}^{-1}_{ODP}(\mathcal{S})\} = P\{\mathrm{R}_{ODP} \in \mathcal{S}\} = 1.$$

Since this holds in both directions and for any such $\mathcal{S}$, it holds for the support which is a special case of $\mathcal{S}$.

Now that we showed that R_{GLN} and R_{ODP} have identical support, we show that the support is a bounded compact set (l, r). To do so, we specify the support of R_{GLN}. The key insight is that the real symmetric matrices $A = M$ and $B = \Pi^{1/2}M^*\Pi^{1/2}$ commute so $AB = BA$. Thus, they are simultaneously diagonalizable (Newcomb 1961). Beyond that,

both matrices are of rank $n-p$. Thus, we can find an orthogonal matrix of Eigenvectors P' such that

$$PAP' = \begin{pmatrix} \Lambda_A & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad PBP' = \begin{pmatrix} \Lambda_B & 0 \\ 0 & 0 \end{pmatrix}$$

where Λ_A and Λ_B are diagonal $n-p \times n-p$ matrices of Eigenvalues. Since M is a projection matrix, $\Lambda_A = I$. Then, making use of Butler & Paolella (2008, Lemma 3 Case 2(c)), the upper bound of the support of R_{GLN} is given by the largest element of $\Lambda_B^{-1}\Lambda_A = \Lambda_B^{-1}$. Thus, it is finite. To find the lower bound, we consider the upper bound of $-\mathrm{R}_{GLN}$, thus swapping A for $-A$. By the same argument, the lower bound is the largest element of Λ_B^{-1} times -1, thus the smallest element of Λ_B^{-1}. Since B is positive semi-definite and because Λ_B contains only the non-zero eigenvalues, this must be larger then zero. Denoting the diagonal elements of Λ_B sorted in descending magnitude by $\lambda_B = (\lambda_{B,(1)}, \ldots \lambda_{B,(n-p)})$, we can write the support as $(\mathrm{l}, \mathrm{r}) = (\lambda_{B,(1)}^{-1}, \lambda_{B,(n-p)}^{-1})$.

Proof of Theorem 4.3

First, we find the limit of the weighted least squares orthogonal projection M^* as $t \to 0$. Then we prove the results relating to power. We assume that rows and columns from the design X and the frequency matrix Π that relate to cells with perfect fit, such as the corners in a run-off triangle, have been removed. In this way, there is no need to keep track of the restriction that not only frequencies relating to such cells can go to zero. Further, we assume without loss of generality that X and Π have been sorted such that the $n-q$ cells for which we want $\pi_{ij} \to 0$ are in the last $n-q$ rows of the matrices. Similarly, we sort the columns of X such that parameters relevant for the first q rows of X, those relating to cells with non-vanishing frequencies, are in the first p_1 columns.

To find the limit of M^*, we first note that this orthogonal projection is invariant to scaling of X^* by constants so that we can without loss of generality drop the normalizing constant $s(t)$ from $\Pi_{(t)}$. Thus, we define,

$$\widetilde{X}^*_{(t)} = \begin{pmatrix} \Pi_1^{1/2} & 0 \\ 0 & t\,\Pi_2^{1/2} \end{pmatrix} X = \begin{pmatrix} \Pi_1^{1/2} X_{11} & \Pi_1^{1/2} X_{12} \\ t\,\Pi_2^{1/2} X_{21} & t\,\Pi_2^{1/2} X_{22} \end{pmatrix} = \begin{pmatrix} \widetilde{X}^*_{11} & \widetilde{X}^*_{12} \\ t\,\widetilde{X}^*_{21} & t\,\widetilde{X}^*_{22} \end{pmatrix}$$

where we partition so $\widetilde{X}^*_{11}$ is $q \times p_1$. Since we ruled out scenarios with a perfect fit for the q cells relating to X^*_{11}, we implicitly impose that $q > p_1$. A crucial insight for later is that $X_{12} = 0$ since we sorted X so the relevant parameters for the cells in the first q rows are collected in the first p_1 columns.

We can then write $M^*_{(t)} = I - \widetilde{X}^*_{(t)}(\widetilde{X}^{*'}_{(t)}\widetilde{X}^*_{(t)})^{-1}\widetilde{X}^*_{(t)}$. Now, let $\widetilde{X}^{*+}_{(t)}$ be the Moore-Penrose inverse of $\widetilde{X}^*_{(t)}$. For any $t \neq 0$ we have that $\widetilde{X}^{*+}_{(t)} = (X^{*'}_{(t)}X^*_{(t)})^{-1}X^{*'}_{(t)}$ since $X^*_{(t)}$ has full column rank. Thus, $M^*_{(t)} = I - X^*_{(t)}X^{*+}_{(t)}$. We are interested in

$$\lim_{t\to 0} M^*_{(t)} = I - \lim_{t\to 0} X^*_{(t)}X^{*+}_{(t)}.$$

The limit of $X^*_{(t)}$ is easily seen to exist. The Moore-Penrose inverse, which always exists, takes a simple form for the limit. Recalling that $X_{12} = 0$ so $\widetilde{X}^*_{12} = 0$ as mentioned above,

$$\lim_{t\to 0} \widetilde{X}^*_{(t)} = \begin{pmatrix} \widetilde{X}^*_{11} & 0 \\ 0 & 0 \end{pmatrix} := \widetilde{X}^*_{(0)} \quad \text{and} \quad X^{*+}_{(0)} = \begin{pmatrix} \widetilde{X}^{*+}_{11} & 0 \\ 0 & 0 \end{pmatrix}$$

where, since $\widetilde{X}_{11}$ has full column rank, $\widetilde{X}^{*+}_{11} = (\widetilde{X}^{*'}_{11}\widetilde{X}^*_{11})^{-1}\widetilde{X}^{*'}_{11}$. Thus, the limit is given by

$$M^*_{(0)} := \lim_{t\to 0} M^*_{(t)} = I - \begin{pmatrix} \widetilde{X}^*_{11}(\widetilde{X}^{*'}_{11}\widetilde{X}^*_{11})^{-1}\widetilde{X}^{*'}_{11} & 0 \\ 0 & 0 \end{pmatrix}.$$

For a brief intuition, we consider weighted least squares estimation for $Z = X\xi + \Pi^{-1/2}_{(t)}\epsilon$. We solve this by minimizing $\|\Pi^{1/2}_{(t)}(Z - X\xi)\|_2$. As t becomes small, the last $n - q$ elements of Z and rows of X corresponding to the vanishing frequencies do not play a role anymore. Similarly, the last $p - p_1$ parameters in ξ are only relevant for such cells and thus not identified; they do not contribute to the norm. The norm is entirely determined by the cells with non-vanishing frequencies which we collected in X_{11} and the first p_1 relevant parameters of ξ.

Reassured that the $M^*_{(0)}$ exists, we show that $\mathrm{R}^{(t)}_{GLN}$ converges. The limit sandwich in the denominator exists and is given by

$$\Pi^{1/2}_{(0)}M^*_{(0)}\Pi^{1/2}_{(0)} = \begin{pmatrix} \Pi^{1/2}_1\{I - \widetilde{X}^*_{11}(\widetilde{X}^{*'}_{11}\widetilde{X}^*_{11})^{-1}\widetilde{X}^{*'}_{11}\}\Pi^{1/2}_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

We note that this is not a zero matrix since we imposed that there be no perfect fit for the cells relating to X^*_{11}. We define, for $U \overset{D}{=} N(0, I)$,

$$\mathrm{R}^{(0)}_{GLN} = \frac{U'MU}{U'\Pi^{1/2}_{(0)}M^*_{(0)}\Pi^{1/2}_{(0)}U}. \tag{4.11}$$

i, j	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 4.4: Run-off triangle by Taylor & Ashe (1983, Appendix)

Now, for any $u \in \mathbb{R}^n$ for which $u'\Pi_{(0)}^{1/2} M_{(0)}^* \Pi_{(0)}^{1/2} u > 0$,

$$\lim_{t \to 0} \mathrm{R}_{GLN}^{(t)}(u) = \mathrm{R}_{GLN}^{(0)}(u).$$

Since the set for which $u'\Pi_{(0)}^{1/2} M_{(0)}^* \Pi_{(0)}^{1/2} u = 0$ has measure zero, $\mathrm{R}_{GLN}^{(t)}(u) \overset{a.s.}{\to} \mathrm{R}_{GLN}^{(0)}(u)$ as $t \to 0$.

Next, we consider $\mathrm{R}_{ODP}^{(t)}$. For the denominator, we have $\lim_{t \to 0} U' M_{(t)}^* U = U' M_{(0)}^* U$ which is positive with probability one. Looking at the numerator, we see that the limit of $(\Pi^{(t)})^{-1/2}$ does not exist in $\mathbb{R}$. We look at this in more detail. Partition $U = (U_1, U_2)'$ for U_1 of length q. Similarly, partition M so M_{11} is $q \times q$. Then we can write the numerator as

$$s(t)^{-2}(U_1'\Pi_1^{-1/2} M_{11} \Pi_1^{-1/2} U_1 + t^{-1} U_2 \Pi_2^{-1/2} M_{22} \Pi_2^{-1/2} U_2 + t^{-1/2} U_1 \Pi_1^{-1/2} M_{12} \Pi_2^{-1/2} U_2).$$

The normalization $s(t)$ converges to $\mathrm{trace}(\Pi_1)^{-1}$ as $t \to 0$. The first term of the sum is non-negative and does not vary with t. The second term is non-negative and $O_p(t)$ while the third term is $O_p(\sqrt{t})$ with ambiguous sign. Further, dropping cells with perfect fit from X ensures $M_{22} \neq 0$. Thus, since $U_2 \neq 0$ with probability one, the second term is positive with probability one so overall the numerator $U'\Pi_{(t)}^{-1/2} M \Pi_{(t)}^{-1/2} U \overset{a.s.}{\to} \infty$. Thus, $\mathrm{R}_{ODP}^{(t)} \overset{a.s.}{\to} \infty$.

Finally, $\mathrm{R}_{ODP}^{(t)} > q_{GLN,\alpha}^{(t)}$ almost surely since for $\alpha \in (0, 1)$, the quantile $q_{GLN,\alpha}^{(t)} \overset{a.s}{\to} q_{GLN,\alpha}^{(0)} < \infty$. Conversely, $q_{ODP,\alpha}^{(t)} \overset{a.s}{\to} \infty$ so $\mathrm{R}_{GLN}^{(t)} \leq q_{ODP,\alpha}^{(t)}$ almost surely.

i, j	1	2	3	4	5	6	7	8	9	10	11
1	153638	188412	134534	87456	60348	42404	31238	21252	16622	14440	12200
2	178536	226412	158894	104686	71448	47990	35576	24818	22662	18000	
3	210172	259168	188388	123074	83380	56086	38496	33768	27400		
4	211448	253482	183370	131040	78994	60232	45568	38000			
5	219810	266304	194650	120098	87582	62750	51000				
6	205654	252746	177506	129522	96786	82400					
7	197716	255408	194648	142328	105600						
8	239784	329242	264802	190400							
9	326304	471744	375400								
10	420778	590400									
11	496200										

Table 4.5: Run-off triangle by Barnett & Zehnwirth (2000, Table 3.5)

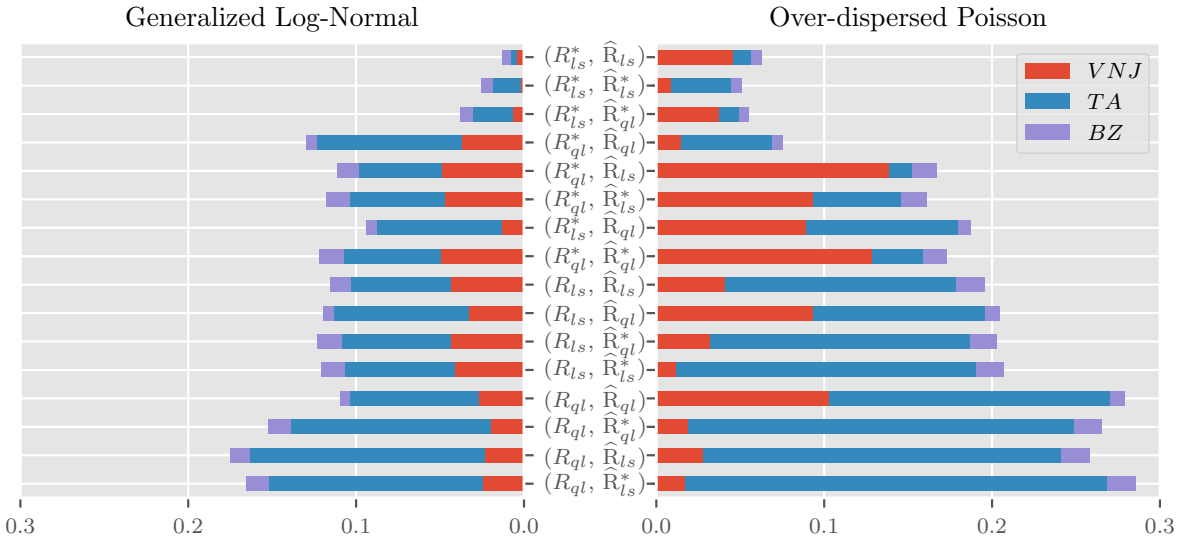


Figure 4.7: Bar chart of maximum absolute errors for the considered combinations of R and $\widehat{\mathrm{R}}$. Ordered by the sum of errors within combination across data generating processes and parameterizations increasing from top to bottom. Sum of maximum absolute errors across parameterizations indicated by "+". VNJ, TA, and BZ is short for parameters set to their estimates from the Verrall et al. (2010) data in Table 4.1, the Taylor & Ashe (1983) data and the Barnett & Zehnwirth (2000) data, respectively. The latter two data sets are provided in the appendix. Based on 10^5 repetitions for each parametrization. $s = 1$.

5. Conclusion

In this thesis, we developed statistical theory for age-period-cohort and extended chain-ladder models. This allowed us to conquer a range of previously unsolved problems including inference, distribution forecasting, misspecification testing, and non-nested model testing.

First, we provided a novel asymptotic framework for over-dispersed Poisson models that opened the door for a number of theoretical contributions. Crucially, the asymptotic framework provided a solution for the incidental parameter problem intrinsic to age-period-cohort models that rendered standard asymptotic theory invalid. We solved this problem by keeping the dimension of the array fixed and instead growing the cell means so that information accumulates within cells rather than across the number of cells.

With this foundation in place, we established a number of results that we hope are useful for the practitioner. We showed that Poisson quasi-likelihood estimators and distribution forecast are asymptotically t-distributed, and F-statistics based on Poisson likelihoods are asymptotically F-distributed. Thus, large sample results match a finite sample Gaussian theory. We showed that the similarity to Gaussian theory carries over to misspecification tests. Specifically, Bartlett-tests can be used to assess a violation in the assumed constancy of variation parameters and F-tests to look for evidence against constancy of mean parameters. Both tests are known from the ANOVA literature for Gaussian models and we pointed out that they are also applicable to log-normal models. This raised the question of when we should choose which model. In our final contribution, we proposed a non-nested test that allows us to determine just that.

Still there remain a number of open problems to solve, including those mentioned at the end of the individual chapters. Already, Kuang & Nielsen (2018) adopted the asymptotic results put forward in Chapters 2 and 3 to develop a theory for generalized log-normal models. This theory allows straightforward distribution forecasting for cell sums in log-normal models, a task that previously proved difficult. We hope that our contributions will provide a useful starting point for further future research.

Bibliography

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* **1238**(3), 208–227.

Andersson, S. A. & Jensen, S. T. (1987), *Forelæsningsnoter i Sandsynlighedsregning*, 3rd edn, Institute of Mathematical Statistics, University of Copenhagen.

Andrews, D. W. K. (1993), 'Tests for Parameter Instability and Structural Change with Unknown Change Point', *Econometrica* **61**(4), 821–856.

Angrist, J. D. & Krueger, A. B. (1995), 'Split-sample instrumental variables estimates off the return to schooling', *Journal of Business and Economic Statistics* **13**(2), 225–235.

Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families*, Chichester, UK: Wiley, New York.

Barndorff-Nielsen, O. E. & Cox, D. R. (1984), 'Bartlett Adjustments to the Likelihood Ratio Statistic and the Distribution of the Maximum Likelihood Estimator', *Journal of the Royal Statistical Society. Series B (Methodological)* **46**(3), 483–495.

Barnett, G. & Zehnwirth, B. (2000), 'Best estimates for reserves', *Proceedings of the Casualty Actuarial Society* **LXXXVII**(167), 245–321.

Bartlett, M. S. (1937), 'Properties of Sufficiency and Statistical Tests', *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **160**(901), 268–282.

Beard, R. E., Pentikäinen, T. & Pesonen, E. (1984), *Risk theory : the stochastic basis of insurance*, Chapman and Hall.

Billingsley, P. (1995), *Probability and Measure*, 3rd ed. edn, Wiley-Interscience, New York; Chichester.

Bliss, C. I. & Fisher, R. A. (1953), 'Fitting the Negative Binomial Distribution to Biological Data', *Biometrics* **9**(2), 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**(3), 1063–1081.

Box, G. E. P. (1953), 'Non-Normality and Tests on Variances', *Biometrika* **40**(3/4), 318–335.

Butler, R. W. & Paolella, M. S. (2008), 'Uniform saddlepoint approximations for ratios of quadratic forms', *Bernoulli* **14**(1), 140–154.

Carstensen, B. (2007), 'Age-period-cohort models for the Lexis diagram', *Statistics in Medicine* **26**(15), 3018–3045.

Casella, G. & Berger, R. L. (2002), *Statistical Inference*, 2nd edn, Duxbury/Thomson Learning, Pacific Grove, Calif.

Casualty Actuarial Society (2011), 'LOSS RESERVING DATA PULLED FROM NAIC SCHEDULE P'.
`http://www.casact.org/research/index.cfm?fa=loss_reserves_data`

Chow, G. C. (1960), 'Tests of Equality Between Sets of Coefficients in Two Linear Regressions', *Econometrica* **28**(3), 591.

Cox, D. R. (1961), 'Tests of separate families of hypothesis', *Proc. 4th Berkeley Symp. on Math. Stat. and Probability* **1**, 105–123.

Cox, D. R. (1962), 'Further Results on Tests of Separate Families of Hypotheses', *Journal of the Royal Statistical Society, Series B* **24**(2), 406–424.

Cox, D. R. (1983), 'Some Remarks on Overdispersion', *Biometrika* **70**(1), 269–274.

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

Daykin, C. D., Pentikäinen, T. & Pesonen, M. (1994), *Practical Risk Theory for Actuaries*, Chapman & Hall, London.

Deaton, A. S. & Paxson, C. H. (1994), Saving , growth , and aging in Taiwan, *in* 'Studies in the Economics of Aging', University of Chicago Press, pp. 331–362.

Durbin, J. & Watson, G. S. (1950), 'Testing for serial correlation in least squares regression. I', *Biometrika* **37**(3-4), 409–428.

Durbin, J. & Watson, G. S. (1951), 'Testing for Serial Correlation in Least Squares Regression. II', *Biometrika* **38**(1/2), 159.

Ejrnaes, M. & Hochguertel, S. (2013), 'Is Business Failure Due to Lack of Effort? Empirical Evidence from a Large Administrative Sample', *The Economic Journal* **123**(571), 791–830.

England, P. D. (2002), 'Addendum to "Analytic and bootstrap estimates of prediction errors in claims reserving"', *Insurance: Mathematics and Economics* **31**(3), 461–466.

England, P. D. & Verrall, R. J. (2002), 'Stochastic Claims Reserving in General Insurance', *British Actuarial Journal* **8**(03), 443–518.

England, P. & Verrall, R. (1999), 'Analytic and bootstrap estimates of prediction errors in claims reserving', *Insurance: Mathematics and Economics* **25**(3), 281–293.

Ermini, L. & Hendry, D. F. (2008), 'Log income vs. linear income: An application of the encompassing principle', *Oxford Bulletin of Economics and Statistics* **70**(s1), 807–827.

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**(1), 87–94.

Fu, W. (2016), 'Constrained Estimators and Consistency of a Regression Model on a Lexis Diagram', *Journal of the American Statistical Association* **111**(513), 180–199.

Gesmann, M., Murphy, D., Zhang, W., Carrato, A., Crupi, G. & Wuthrich, M. (2015), 'ChainLadder: Statistical Methods and Models for Claims Reserving in General Insurance. R package version 0.2.0'.
http://cran.r-project.org/package=ChainLadder

Harnau, J. (2017), 'apc'.
https://pypi.python.org/pypi/apc/

Harnau, J. (2018*a*), 'Log-Normal or Over-Dispersed Poisson?', *Unpublished manuscript; UK: University of Oxford* .

Harnau, J. (2018*b*), 'Misspecification Tests for Log-Normal and Over-Dispersed Poisson Chain-Ladder Models', *Risks* **6**(2), 25.

Harnau, J. (2018*c*), 'quad_form_ratio'.
https://pypi.python.org/pypi/quad-form-ratio/

Harnau, J. & Nielsen, B. (2017), 'Over-dispersed age-period-cohort models', *Journal of the American Statistical Association* .
https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1366908

Heckman, J. & Vytlacil, E. (2001), 'Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return of Schooling', *The Review of Economics and Statistics* **83**(1), 1–12.

Hendry, D. F. (1999), An Econometric Analysis of US Food Expenditure, *in* J. R. Magnus & M. S. Morgan, eds, 'Methodology and Tacit Knowledge: Two Experiments in Econometrics', Wiley, New York, pp. 341–361.

Hendry, D. F. & Nielsen, B. (2007), *Econometric modeling: a likelihood approach*, Princeton University Press, Princeton, N.J.; Oxford.

Hendry, D. F. & Richard, J.-F. (1982), 'On the formulation of empirical models in dynamic econometrics', *Journal of Econometrics* **20**(1), 3–33.

Hertig, J. (1983), 'A statistical approach to IBNR-reserves in marine reinsurance', *ASTIN Bulletin* **15**(2), 171–183.

Hiabu, M. (2017), 'On the relationship between classical chain ladder and granular reserving', *Scandinavian Actuarial Journal* **2017**(8), 708–729.

Hodgson, J. T., McElvenny, D. M., Darnton, A. J., Price, M. J. & Peto, J. (2005), 'The expected burden of mesothelioma mortality in Great Britain from 2002 to 2050', *British journal of cancer* **92**(3), 587–93.

Human Mortality Database (2018), 'Human Mortality Database'.
`http://www.mortality.org/`

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

Johnson, N., Kemp, A. & Kotz, S. (2005), *Univeriate discrete distributions*, Wiley, Hoboken.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995*a*), *Continuous Univariate Distributions Volume 1*, 2nd edn, Wiley, Chichester.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995*b*), *Continuous univariate distributions Volume 2*, 2nd edn, Wiley, Chichester.

Jørgensen, B. (1986), 'Some properties of exponential dispersion models', *Scandinavian Journal of Statistics* **13**(3), 187–197.

Jørgensen, B. (1987), 'Exponential dispersion models', *Journal of the Royal Statistical Society. Series B (Methodological)* **49**(2), 127–162.

Jørgensen, B. (1993), *The theory of linear models*, Chapman & Hall, New York.

Keiding, N. (1990), 'Statistical inference in the Lexis diagram', *Philosophical Transactions: Physical Sciences and Engineering* **332**(1627), 487–509.

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

Kremer, E. (1985), *Einführung in die Versicherungsmathematik*, 7th edn, Vandenhoeck & Ruprecht.

Kuang, D. & Nielsen, B. (2018), 'Generalized Log Normal Chain-Ladder', *Unpublished manuscript; UK: University of Oxford* .

Kuang, D., Nielsen, B. & Nielsen, J. (2015), 'The geometric chain-ladder', *Scandinavian Actuarial Journal* **2015**(3), 278–300.

Kuang, D., Nielsen, B. & Nielsen, J. P. (2008*a*), 'Forecasting with the age-period-cohort model and the extended chain-ladder model', *Biometrika* **95**(4), 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**(4), 979–986.

Kuang, D., Nielsen, B. & Nielsen, J. P. (2009), 'Chain-Ladder as Maximum Likelihood Revisited', *Annals of Actuarial Science* **4**(01), 105–121.

Kuang, D., Nielsen, B. & Perch Nielsen, J. (2011), 'Forecasting in an Extended Chain-Ladder-Type Model', *Journal of Risk and Insurance* **78**(2), 345–359.

Lancaster, T. (2000), 'The incidental parameter problem since 1948', *Journal of Econometrics* **95**(2), 391–413.

Lawley, D. N. (1956), 'A general method for approximating to the distribution of likelihood ratio criteria', *Biometrika* **43**(3/4), 295–303.

Lee, R. D. & Carter, L. R. (1992), 'Modeling and Forecasting U.S. Mortality', *Journal of the American Statistical Association* **87**(419), 659–671.

Lee, Y. K., Mammen, E., Nielsen, J. P. & Park, B. U. (2015), 'Asymptotics for in-sample density forecasting', *Annals of Statistics* **43**(2), 620–651.

Lieberman, O. (1994), 'Saddlepoint approximation for the distribution of a ratio of quadratic forms in normal variables', *Journal of the American Statistical Association* **89**(427), 924–928.

Lugannani, R. & Rice, S. (1980), 'Saddle point approximation for the distribution of the sum of independent random variables', *Advances in Applied Probability* **12**(2), 475.

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

Mack, T. (1993), 'Distribution free calculation of the standard error of chain ladder reserve estimates', *ASTIN Bulletin* **23**(2), 213–225.

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 (Statistics in Society)* **178**(1), 29–55.

Martínez-Miranda, M. D., Nielsen, J. P., Sperlich, S. & Verrall, R. (2013), 'Continuous Chain Ladder: Reformulating and generalizing a classical insurance problem', *Expert Systems with Applications* **40**(14), 5588–5603.

Martínez-Miranda, M. D., Nielsen, J. P. & Verrall, R. (2012), 'Double Chain Ladder', *ASTIN Bulletin* **42**(1), 59–76.

Martínez-Miranda, M. D., Nielsen, J. P. & Wüthrich, M. V. (2013), 'Statistical modelling and forecasting of outstanding liabilities in non-life insurance', **36**(2), 195–218.

Mizon, G. E. & Richard, J.-F. (1986), 'The Encompassing Principle and its Application to Testing Non-Nested Hypotheses', *Econometrica* **54**(3), 657.

Newcomb, R. W. (1961), 'On the simultaneous diagonalization of two semi-definite matrices', *Quarterly of Applied Mathematics* **19**(2), 144–146.

Neyman, J. & Scott, E. L. (1948), 'Consistent estimates based on partially consistent observations', *Econometrica* **16**(1), 1–32.

Nielsen, B. (2014), 'Deviance analysis of age-period-cohort models', *Nuffield Discussion Paper* (W03).
`http://www.nuffield.ox.ac.uk/economics/papers/2014/apc_deviance.pdf`

Nielsen, B. (2015), 'apc: An R Package for Age-Period-Cohort Analysis', *The R Journal* **7**(2), 52–64.
`https://journal.r-project.org/archive/2015-2/nielsen.pdf`

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

Nielsen, B. & Whitby, A. (2015), 'A Joint Chow Test for Structural Instability', *Econometrics* **3**, 156–186.

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**(4), 701–714.

R Core Team (2016), 'R: A Language and Environment for Statistical Computing'.
`http://www.r-project.org`

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

Shoemaker, L. H. (2003), 'Fixing the F Test for Equal Variances', *The American Statistician* **57**(2), 105–114.

Snedecor, G. W. & Cochran, W. G. (1967), *Statistical methods*, 6th edn, The Iowa State University.

Tan, E., Warren, N., Darnton, A. J. & Hodgson, J. T. (2010), 'Projection of mesothelioma mortality in Britain using Bayesian methods.', *British journal of cancer* **103**(3), 430–6.

Taylor, G. C. & Ashe, F. (1983), 'Second moments of estimates of outstanding claims', *Journal of Econometrics* **23**(1), 37–61.

Thorin, O. (1977), 'On the infinite divisibility of the lognormal distribution', *Scandinavian Actuarial Journal* **1977**(3), 121–148.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, Spriger-Verlag, New York.

Verrall, R. J. (1991), 'On the estimation of reserves from loglinear models', *Insurance: Mathematics and Economics* **10**(1), 75–80.

Verrall, R. J. (1994), 'Statistical methods for the chain ladder technique', *Casualty Actuarial Society Forum* **Vol. 1**, 393–446.

Verrall, R., Nielsen, J. P. & Jessen, A. H. (2010), 'Prediction of RBNS and IBNR claims using claim amounts and claim counts', *ASTIN Bulletin* **40**(2), 871–887.

Vuong, Q. H. (1989), 'Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses', *Econometrica* **57**(2), 307.

Wedderburn, R. (1974), 'Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method', *Biometrika* **61**(3), 439–447.

Yang, Y., Fu, W. & Land, K. (2004), 'A Methodological Comparison of Age-Period-Cohort Models: The Intrinsic Estimator and Conventional Genralized Models', *Sociological Methodology* **34**(1), 75–110.

Zehnwirth, B. (1994), 'Probabilistic development factor models with applications to loss reserve variability, prediction intervals and risk based capital.', *Insurance: Mathematics and Economics* **15**(1), 82.