risks

MDPI

Article

Misspecification Tests for Log-Normal and Over-Dispersed Poisson Chain-Ladder Models

Jonas Harnau

Department of Economics, University of Oxford & Oriel College, Oxford OX1 4EW, UK; jonas.harnau@oriel.ox.ac.uk

Received: 28 February 2018; Accepted: 21 March 2018; Published: 23 March 2018

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.

Keywords: Bartlett test; *F*-test; over-dispersed Poisson; log-normal

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 2002; England and Verrall 1999) is in widespread use. Yet, so far we do not have a statistical theory for the bootstrap in this setting.

Recently, Harnau and 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 and 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.

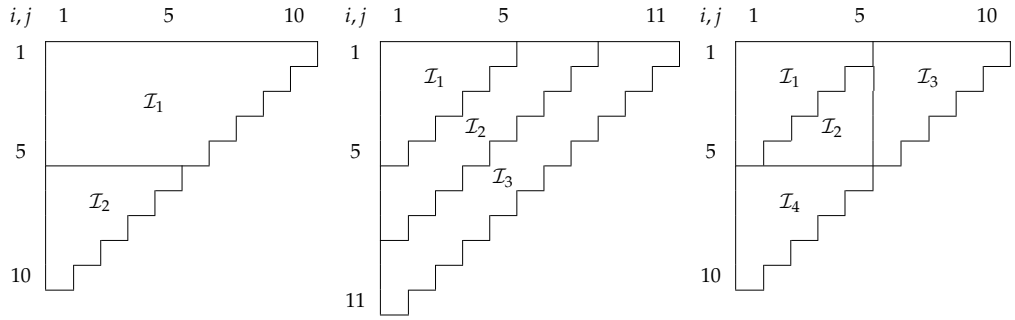
Modeling aggregate incremental claims as log-normal rather than over-dispersed Poisson is also common. Kremer (1982) introduced a log-normal model with multiplicative mean structure mirroring the over-dispersed Poisson chain-ladder model. While this model does not replicate the classic chain-ladder technique,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

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 and 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 1 shows examples of how we could split run-off triangles into sub-samples.



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

Figure 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.

In Section 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 chose 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 Section 3.

In the over-dispersed Poisson model, the asymptotic framework by Harnau and 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 Section 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 Section 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 Section 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 over-dispersed Poisson and log-normal model. With this, Section 7 concludes the paper.

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 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). 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 1, taken from Taylor and Ashe (1983), are a generalized trapezoid with $I^l = J^l = K^l = 1$, $I^u = J^u = K^u = 10$ and $n = 55$.

Table 1. Insurance run-off triangle taken from Taylor and 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.

i, j	1	2	3	4	5	6	7	8	9	10
1	357,848	766,940	610,542	482,940	527,326	574,398	146,342	139,950	227,229	67,948
2	352,118	884,021	933,894	1,183,289	445,745	320,996	527,804	266,172	425,046	-
3	290,507	1,001,799	926,219	1,016,654	750,816	146,923	495,992	280,405	-	-
4	310,608	1,108,250	776,189	1,562,400	272,482	352,053	206,286	-	-	-
5	443,160	693,190	991,983	769,488	504,851	470,639	-	-	-	-
6	396,132	937,085	847,498	805,037	705,960	-	-	-	-	-
7	440,832	847,631	1,131,398	1,063,269	-	-	-	-	-	-
8	359,480	1,061,648	1,443,370	-	-	-	-	-	-	-
9	376,686	986,608	-	-	-	-	-	-	-	-
10	344,014	-	-	-	-	-	-	-	-	-

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 1 are generalized trapezoids. For instance, the sub-sample $\mathcal{I}_2$ in Figure 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 cells and disconnected arrays such as the combination of sub-samples $\mathcal{I}_1$ and $\mathcal{I}_3$ in Figure 1b. However, for many of these arrays identification may still be given and then the theory developed below will still be valid.

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.

3.1. 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

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

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

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

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

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

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^{LN}_{μ,σ^2} 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.

3.2. 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). 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). 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) 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.

3.2.1. 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

$$\hat{\sigma}_\ell^{2,LN} = \frac{RSS_\ell}{df_\ell} \quad \text{where} \quad RSS_\ell = \sum_{ij \in \mathcal{I}_\ell} \{\log(Y_{ij,\ell}) - \hat{\mu}_{ij,\ell}^{LN}\}^2. \tag{1}$$

3.2.2. Estimation with Common Variances in $M_{\sigma^2}^{LN}$

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 $M_{\sigma^2}^{LN}$ is estimated by

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

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

3.2.3. Estimation with Common Variances and Linear Predictors in M_{μ,σ^2}^{LN}

Under the hypothesis H_{μ,σ^2}^{LN}, 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_{μ,σ^2}^{LN} as

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

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

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

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 and 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. 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

$$LR^{LN} = df_\cdot \log(\bar{\sigma}^{2,LN}) - \sum_{\ell=1}^{m} df_\ell \log(\hat{\sigma}_\ell^{2,LN}) \tag{3}$$

for $\hat{\sigma}^2$ and $\bar{\sigma}^2$ as defined in (1) and (2), respectively. Define now the Bartlett distribution $\mathrm{Ba}(\cdot)$ such that $LR^{LN} \overset{D}{=} \mathrm{Ba}(df_1, \ldots, df_m)$ 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

$$P\{\mathrm{Ba}(df_1, \ldots, df_m) \le y\} = \int_{A(y)} \prod_{\ell=1}^{m} dG_\ell(x_\ell) \tag{4}$$

where $G_\ell(\cdot)$ is the $\chi^2_{df_\ell}$ cdf and $A(y) = \{(x_1, \ldots, x_m) : LR(x_1, \ldots, x_m) \le y\}$. 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},

$$B^{LN} = \frac{LR^{LN}}{C} \overset{D}{\approx} \chi^2_{m-1}. \tag{5}$$

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 and 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 Section 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 $\sigma^{2,LN}_\ell$ in (1). Under H_{σ^2},

$$F^{LN}_{\sigma^2} = \frac{\sigma^{2,LN}_2}{\sigma^{2,LN}_1} \overset{D}{=} \mathrm{F}_{df_2,df_1} \tag{6}$$

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

$$LR^{LN} = LR(F^{LN}_{\sigma^2}) = df_1 \log\left(\frac{1 + rF^{LN}_{\sigma^2}}{1+r}\right) + df_2 \log\left\{\frac{1 + (rF^{LN}_{\sigma^2})^{-1}}{1+r^{-1}}\right\}. \tag{7}$$

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 $F^{LN}_{\sigma^2}$. 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 $F^{LN}_{\sigma^2} = 0.025$, 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 Section 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.

3.4. 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 $M^{LN}_{\sigma^2}$ as given, then testing for H_{μ,σ^2} amounts to testing for common linear predictors. Since standard Gaussian theory applies,

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

under the hypothesis. Thus, we can use a (one-sided) F-test to test for a reduction from $M^{LN}_{\sigma^2}$ to M^{LN}_{μ,σ^2}. 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 Appendix A.

Theorem 1. *In model M^{LN}_{μ,σ^2}, the test statistic F^{LN}_{μ} is independent of $F^{LN}_{\sigma^2}$ 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.

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 and 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 and 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.

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 and 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:

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

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:

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

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

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

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

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

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

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

The model under this hypothesis in a run-off triangle replicates the chain-ladder. Thus, M^{ODP}_{μ,σ^2} 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.

4.2. Estimation

With the model and hypotheses in place, we move on to estimation. The estimators match those in Harnau and 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.

4.2.1. Estimation in Unrestricted Model M^{ODP}

We estimate linear predictors by Poisson quasi-likelihood

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

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^2_ℓ 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^2_ℓ is

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

4.2.2. Estimation with Common Variances in $M^{ODP}_{\sigma^2}$

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

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

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

4.2.3. Estimation with Common Variances and Linear Predictors in M^{ODP}_{μ,σ^2}

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

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

and

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

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

For intuitive appeal, the skewness in the compound Poisson example from Section 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.

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 and Nielsen 2017, Lemma 1):

$$\hat{\sigma}_\ell^{2,ODP} \xrightarrow{D} \frac{\sigma_\ell^2}{df_\ell}\chi^2_{df_\ell}. \tag{8}$$

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

$$LR^{ODP} = df_{\cdot}\log(\bar{\sigma}^{2,ODP}) - \sum_{\ell=1}^{m} df_\ell \log(\hat{\sigma}_\ell^{2,ODP}). \tag{9}$$

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 2. *In the over-dispersed Poisson model with common over-dispersion $M_{\sigma^2}^{ODP}$ of Sections 4.1 and 4.3, LR^{ODP} converges to the Bartlett distribution* $\text{Ba}(df_1,\dots,df_\ell)$ *from (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 Section 6.3 below, we show that finite sample approximations to the asymptotic results in Theorem 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 $E(Y_{..})$ 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

$$B^{ODP} = \frac{LR^{ODP}}{C} \overset{D}{\approx} \chi^2_{m-1}. \tag{10}$$

A simultaneous double asymptotic theory for large $E(Y_{..})$ 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 Section 6 make it seem unnecessary.

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 and 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:

$$F^{ODP}_{\mu} = \frac{(D - D_{.})/(df - df_{.})}{D_{.}/df_{.}} \overset{D}{\to} \mathrm{F}_{df-df_{.},df_{.}}. \tag{11}$$

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 1. *In the over-dispersed Poisson model M^{ODP}_{μ,σ^2} of Sections 4.1 and 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.

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 and Ashe (1983) in Table 1 which has become a sort of benchmark data set for this model. Verrall (1991), England and Verrall (1999), and Pinheiro et al. (2003) all use this data, to name but a few. Finally, the data by Barnett and 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.

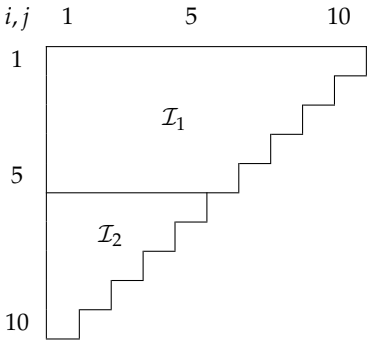
5.1. Log-Normal Chain-Ladder

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 A1.

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 2 summarizes the results.

Figure 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.



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

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 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 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

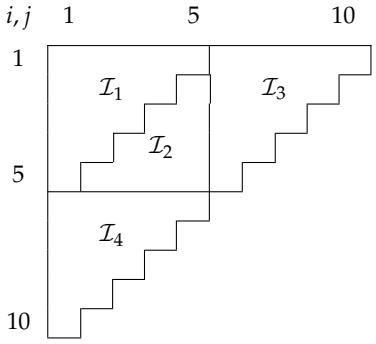
$$M^{LN}_{\mu,\sigma^2} : \log(Y_{ij}) \stackrel{D}{=} N(\alpha_i + \beta_j + \delta, \sigma^2).$$

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 $H_{\sigma^2} : \sigma^2_1 > \sigma^2_2$. 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.

5.2. Over-Dispersed Poisson Chain-Ladder

The Taylor and Ashe (1983) data in Table 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.



(**a**) Sub-sample structure

	$\hat{\sigma}^{2,ODP}_\ell$	df_ℓ	$\bar{\sigma}^{2,ODP}$	$df_\cdot$	$\hat{\sigma}^{2,ODP}$	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. Over-dispersed Poisson chain-ladder model for Taylor and Ashe (1983) data. Sub-sample structure shown in (**a**), estimation and test results in (**b**).

Figure 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 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 3b shows estimates and test results. The unrestricted model is the over-dispersed Poisson model discussed in Section 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.

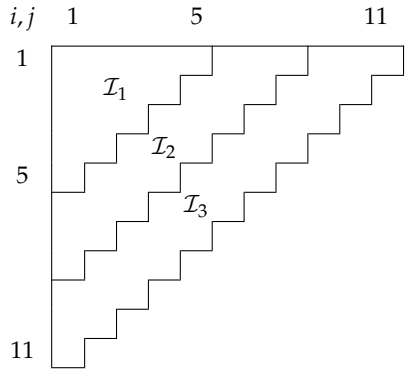
Overall, the target over-dispersed Poisson model for the Taylor and Ashe (1983) data survives both misspecification tests at a 5% level for this sub-sample structure. Thus, we may be more confident now to model it with an over-dispersed Poisson chain-ladder model.

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 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 $F^{ODP}_\mu = 0.63$ 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 1b. For this structure, we get $B^{ODP} = 1.27$ with a p-value of 0.53 and $F^{ODP}_\mu = 1.84$ 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.

5.3. Log-Normal (Extended) Chain-Ladder

As a final empirical application, we look at a run-off triangle first considered by Barnett and Zehnwirth (2000). We show this data in Table A2. 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 and 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 4.



(**a**) Sub-sample structure

	$\hat{\sigma}^{2,LN}_\ell$	df_ℓ	$\bar{\sigma}^{2,LN}$	$df_\cdot$	$\hat{\sigma}^{2,LN}$	df
$\mathcal{I}_1$	0.0002	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	$\bar{\sigma}^{2,LNe}$	$df^e_\cdot$	$\hat{\sigma}^{2,LNe}$	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 4. Log-normal chain-ladder (LN) and extended chain-ladder (LNe) model for Barnett and Zehnwirth (2000) data. Sub-sample structure shown in (**a**), estimation and test results in (**b**).

Figure 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 4b shows estimation and test results for a model without calendar 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 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 $M^{LNe}_{\sigma^2}$ as given. An F-test for common linear predictors leaves us with a p-value of 0.41. Thus, reducing the model to M^{LNe}_{μ,σ^2} 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^{LNe}_{μ,σ^2}. 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^{LN}_{μ,σ^2} above.

Alternatively, we can test for a reduction from $M^{LNe}_{\sigma^2} : \log(Y_{ij,\ell}) \overset{D}{=} N(\alpha_{i,\ell} + \beta_{j,\ell} + \gamma_{k,\ell} + \delta_\ell, \sigma^2)$ to $M^{LN}_{\sigma^2}$, 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 $M^{LN}_{\sigma^2}$ (but not from M^{LNe} to $M^{LNe}_{\sigma^2}$).

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.

6. Simulations

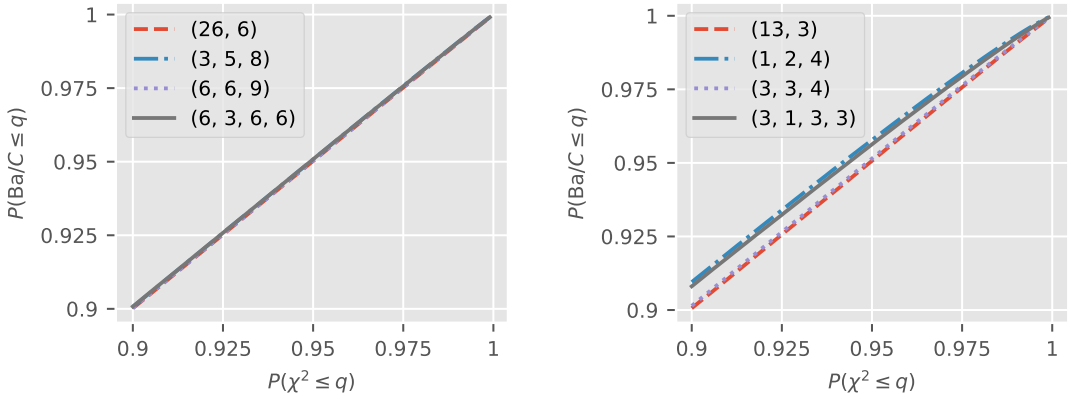
The developed theory begs several questions that we answer in a simulation study. First, we argued that we can sidestep simulating critical values of the Bartlett distribution Ba and instead approximate these by a Bartlett corrected χ^2 critical value. We show that this works very well. Second, we compute 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.

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.

Figure 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.



(**a**) Degrees of freedom from empirical applications (**b**) Half the empirical degrees of freedom

Figure 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.

Figure 5b is constructed in the same way as Figure 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 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.

Table 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 $(df_1, \ldots, df_m)$.

	(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

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.

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 Section 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 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

$$LR^{LN} = df_\cdot \log(\bar{\sigma}^{2,LN}) - \sum_{\ell=1}^{m} df_\ell \log(\hat{\sigma}^{2,LN}_\ell) = \sum_{\ell=1}^{m} df_\ell \log\left\{ \frac{df_\ell}{df_\cdot} \left(\sum_{s=1}^{\ell} \frac{RSS_s}{RSS_\ell} \right) \right\}. \tag{12}$$

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 and Zehnwirth 2000), and $(0.09, 1, 0.58, 0)$ (Taylor and Ashe 1983). The log data variance for the ℓ-th subset is then

$$\sigma^2_\ell = \sigma^2_{(1)} + x_\ell(\sigma^2_{(m)} - \sigma^2_{(1)}). \tag{13}$$

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 (6) and B^{LN} as in (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 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 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%.

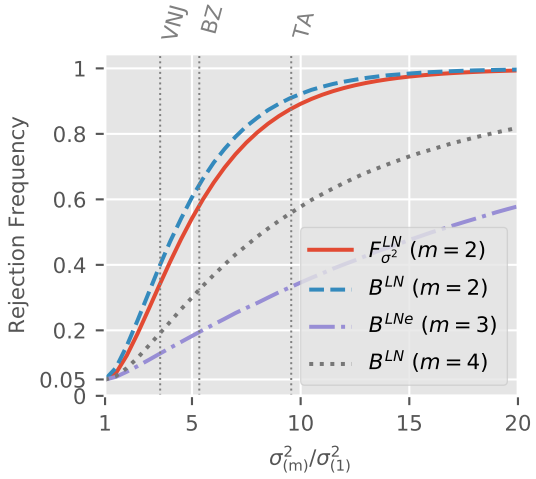


Figure 6. Power curves for log-normal dispersion tests based on sub-sample structures from empirical applications. Empirical maximum to minimum ratios indicated by horizontal lines. BZ is short for Barnett and Zehnwirth (2000), VNJ for Verrall et al. (2010), and TA for Taylor and Ashe (1983).

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.

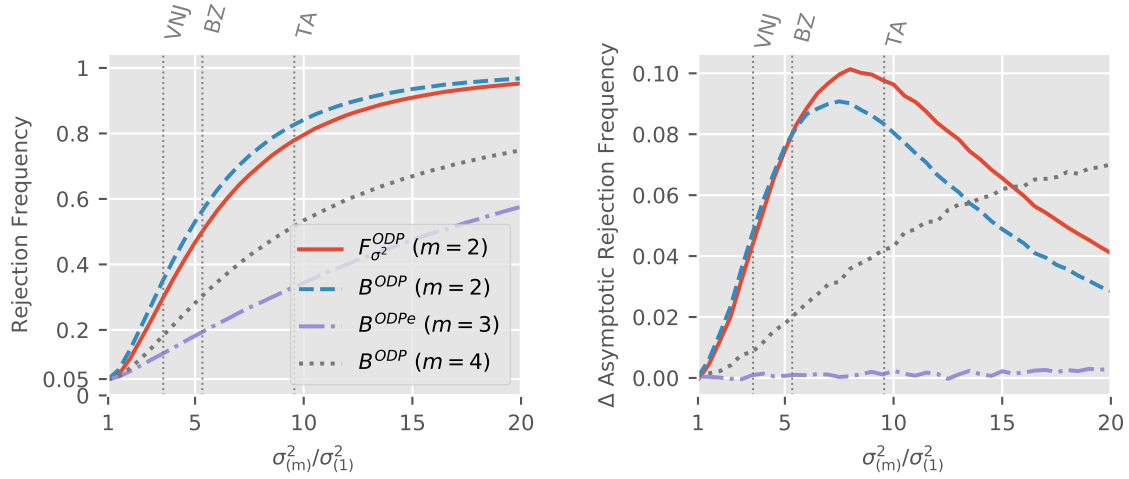
6.3.1. Rejection Frequencies of Tests for Common Over-Dispersion

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, asymptotically, the distribution of LR^{ODP} in M^{ODP} and LR^{LN} in M^{LN} are identical for identical ratios σ^2_s/σ^2_t. The same holds for $F^{ODP}_{\sigma^2}$ and $F^{LN}_{\sigma^2}$.

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 Section 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 $\sigma^2_\ell - 1$ and shape $(\sigma^2_\ell - 1)^{-1}$. We note that the parametrization for the linear predictors $\mu_{ij,\ell}$ and the level of the over-dispersion σ^2_ℓ matters in finite samples. This is in contrast to the log-normal model. The reason is that the finite sample distribution of $\hat{\sigma}^{2,ODP}_{\ell}$ in M^{ODP} is generally not $\sigma^2_\ell\chi^2_{df_\ell}/df_\ell$. Thus, for each considered scenario, we set the linear predictors $\mu_{ij,\ell}$ to the estimates $\hat{\mu}^{ODP}_{ij,\ell}$ from the data in the corresponding empirical application. Similarly, we set the smallest over-dispersion $\sigma^2_{(1)} = \hat{\sigma}^{2,ODP}_{(1)}$.

We again vary the ratios $\sigma^2_{(m)}/\sigma^2_{(1)}$ from one to twenty, using the exact same spacing x_ℓ from (13) in the log-normal simulations so $\sigma^2_\ell = \sigma^2_{(1)} + x_\ell(\sigma^2_\ell - \sigma^2_{(1)})$. The only difference is that now $\sigma^2_{(1)} = \hat{\sigma}^{2,ODP}_{(1)}$. Therefore, asymptotically, the power for a common $\sigma^2_{(m)}/\sigma^2_{(1)}$ is identical in over-dispersed Poisson and log-normal models. We draw 10^6 sub-samples for each over-dispersion ratio $\sigma^2_{(m)}/\sigma^2_{(1)}$ and sub-sample structure.

Figure 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 $\hat{\sigma}^{2,ODP}_{\ell}$.



(**a**) Rejection frequencies (**b**) Gap to asymptotic rejection frequencies

Figure 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**).

For $\sigma^2_{(m)}/\sigma^2_{(1)} = 1$ 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 $\sigma^2_{(m)}/\sigma^2_{(1)} > 1$, the ordering of the rejection frequencies matches that in the log-normal simulations (Figure 6). Generally, the plot is reassuringly reminiscent of its equivalent in the log-normal simulations.

Figure 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 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 $\sigma^2_{(m)}/\sigma^2_{(1)} \approx 7.5$, before it decreases. The initial increase relates to the asymptotic theory by Harnau and 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 $\sigma^2_{(m)}/\sigma^2_{(1)}$. 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 and Zehnwirth (2000) simulations compared with 70% and 56% for the Verrall et al. (2010) and Taylor and 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 7b shows that for these alternatives, the power for all asymptotic approximations is within $5pp$ of their asymptotic power.

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 1. Specifically, we consider the finite sample distribution of F_{μ}^{ODP} as in (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_{μ,σ^2}^{ODP}. As before, cells Y_{ij} are compound Poisson-gamma. We consider three scenarios, setting the parameters to the estimates for M_{μ,σ^2}^{ODP} 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_{μ}^{ODP}.

Figure 8 shows a pp-plot for the F_{μ}^{ODP} against $\mathrm{F}_{df-df.,df.}$ for the triangles that survived Bartlett testing.

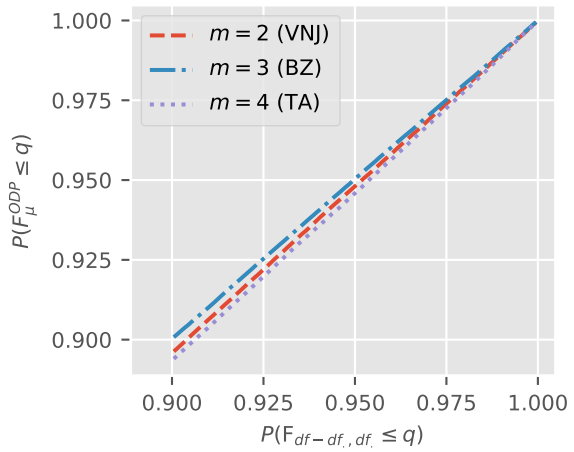


Figure 8. Distribution of F_{μ}^{ODP} conditional on non-rejection of a 5% Bartlett test.

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.

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 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.

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 and 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 and Nielsen (2018), a class of models that includes the log-normal but is more general, Harnau (2018) 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.

Acknowledgments: The author thanks two anonymous referees for their helpful comments. Discussions with Bent Nielsen (Department of Economics, University of Oxford & Nuffield College, UK) are gratefully acknowledged. The author was supported by the Economic and Social Research Council, grant number ES/J500112/1, and the European Research Council, grant AdG 694262.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Appendix A.1. Proof of Theorem 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, 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 and 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}.

Appendix A.2. Proof of Theorem 2

This follows from (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 Section 3.

Appendix A.3. Proof of Lemma 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 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 and 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 1 follows.

Table A1. Insurance run-off triangle taken from Verrall et al. (2010, Table 1) as used in the empirical application in Section 5.1 and the simulations in Section 6.

i, j	1	2	3	4	5	6	7	8	9	10
1	451,288	339,519	333,371	144,988	93,243	45,511	25,217	20,406	31,482	1729
2	448,627	512,882	168,467	130,674	56,044	33,397	56,071	26,522	14,346	-
3	693,574	497,737	202,272	120,753	125,046	37,154	27,608	17,864	-	-
4	652,043	546,406	244,474	200,896	106,802	106,753	63,688	-	-	-
5	566,082	503,970	217,838	145,181	165,519	91,313	-	-	-	-
6	606,606	562,543	227,374	153,551	132,743	-	-	-	-	-
7	536,976	472,525	154,205	150,564	-	-	-	-	-	-
8	554,833	590,880	300,964	-	-	-	-	-	-	-
9	537,238	701,111	-	-	-	-	-	-	-	-
10	684,944	-	-	-	-	-	-	-	-	-

Table A2. Insurance run-off triangle taken from Barnett and Zehnwirth (2000, Table 3.5) as used in the empirical application in Section 5.1 and the simulations in Section 6.

i, j	1	2	3	4	5	6	7	8	9	10	11
1	153,638	188,412	134,534	87,456	60,348	42,404	31,238	21,252	16,622	14,440	12,200
2	178,536	226,412	158,894	104,686	71,448	47,990	35,576	24,818	22,662	18,000	-
3	210,172	259,168	188,388	123,074	83,380	56,086	38,496	33,768	27,400	-	-
4	211,448	253,482	183,370	131,040	78,994	60,232	45,568	38,000	-	-	-
5	219,810	266,304	194,650	120,098	87,582	62,750	51,000	-	-	-	-
6	205,654	252,746	177,506	129,522	96,786	82,400	-	-	-	-	-
7	197,716	255,408	194,648	142,328	105,600	-	-	-	-	-	-
8	239,784	329,242	264,802	190,400	-	-	-	-	-	-	-
9	326,304	471,744	375,400	-	-	-	-	-	-	-	-
10	420,778	590,400	-	-	-	-	-	-	-	-	-
11	496,200	-	-	-	-	-	-	-	-	-	-

References

Andersson, Steen A., and Søren T. Jensen. 1987. *Forelæsningsnoter i Sandsynlighedsregning*, 3rd ed. Copenhagen: Institute of Mathematical Statistics, University of Copenhagen.

Andrews, Donald W. K. 1993. Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica* 61: 821–56.

Angrist, Joshua D., and Alan B. Krueger. 1995. Split-sample instrumental variables estimates off the return to schooling. *Journal of Business and Economic Statistics* 13: 225–35.

Barndorff-Nielsen, Ole E., and David R. Cox. 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: 483–95.

Bartlett, Maurice S. 1937. Properties of Sufficiency and Statistical Tests. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 160: 268–82.

Barnett, Glen, and Ben Zehnwirth. 2000. Best estimates for reserves. *Proceedings of the Casualty Actuarial Society* 87: 245–321.

Box, George E. P. 1953. Non-Normality and Tests on Variances. *Biometrika* 40: 318–35.

Chow, Gregory C. 1960. Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica* 28: 591.

England, Peter D. 2002. Addendum to "Analytic and bootstrap estimates of prediction errors in claims reserving". *Insurance: Mathematics and Economics* 31: 461–66.

England, Peter D., and Richard J. Verrall. 1999. Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics* 25: 281–93.

England, Peter D., and Richard J. Verrall. 2002. Stochastic Claims Reserving in General Insurance. *British Actuarial Journal* 8: 443–518.

Harnau, Jonas. 2018. Log-Normal or Over-Dispersed Poisson? Unpublished manuscript, University of Oxford, UK.

Harnau, Jonas, and Bent Nielsen. 2017. Over-dispersed age-period-cohort models. *Journal of the American Statistical Association*. doi:10.1080/01621459.2017.1366908.

Hendry, David F. 1999. An Econometric Analysis of US Food Expenditure. In *Methodology and Tacit Knowledge: Two Experiments in Econometrics*. Edited by Jan R. Magnus and Mary S. Morgan. New York: Wiley, pp. 341–61.

Hertig, Joakim. 1983. A Statistical Approach To Ibnr-Reserves in Marine Reinsurance. *ASTIN Bulletin* 15: 171–83.

Johansen, Søren. 1979. *Introduction to the Theory of Regular Exponential Families*. Copenhagen: Institute of Mathematical Statistics, University of Copenhagen.

Johnson, Norman L., Samuel Kotz, and Narayanaswamy Balakrishnan. 1995. *Continuous Univariate Distributions Volume 2*, 2nd ed. Chichester: Wiley.

Jørgensen, Bent. 1993. *The Theory of Linear Models*. New York: Chapman & Hall.

Kremer, Erhard. 1985. *Einführung in die Versicherungsmathematik*, 7th ed. Göttingen: Vandenhoeck & Ruprecht, pp. 130–36.

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

Kuang, Di, and Bent Nielsen. 2018. Generalized Log Normal Chain-Ladder. Unpublished manuscript, University of Oxford, UK.

Kuang, Di, Bent Nielsen, and Jens P. Nielsen. 2011. Forecasting in an Extended Chain-Ladder-Type Model. *Journal of Risk and Insurance* 78: 345–59.

Kuang, Di, Bent Nielsen, and Jens P. Nielsen. 2008. Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika* 95: 979–86.

Kuang, Di, Bent Nielsen, and Jens P. Nielsen. 2015. The geometric chain-ladder. *Scandinavian Actuarial Journal* 2015: 278–300.

Lawley, Derrick N. 1956. A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika* 43: 295–303.

Mack, Thomas. 1993. Distribution free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin* 23: 213–25.

Martínez Miranda, Maria D., Bent Nielsen, and Jens P. Nielsen. 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: 29–55.

Martínez Miranda, Maria D., Jens P. Nielsen, and Richard J. Verrall. 2012. Double Chain Ladder. *ASTIN Bulletin* 42: 59–76.

Nielsen, Bent. 2015. apc: An R Package for Age-Period-Cohort Analysis. *The R Journal* 7: 52–64.

Nielsen, Bent, and Andrew Whitby. 2015. A Joint Chow Test for Structural Instability. *Econometrics* 3: 156–86.

Pinheiro, Paulo J. R., João M. Andrade e Silva, and Maria de Lourdes Centeno. 2003. Bootstrap methodology in claim reserving. *Journal of Risk and Insurance* 70: 701–14.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Shoemaker, Lewis H. 2003. Fixing the F Test for Equal Variances. *The American Statistician* 57: 105–14.

Snedecor, George W., and William G. Cochran. 1967. *Statistical Methods*, 6th ed. Ames: The Iowa State University.

Taylor, Greg C., and Frank R. Ashe. 1983. Second moments of estimates of outstanding claims. *Journal of Econometrics* 23: 37–61.

Verrall, Richard J. 1991. On the estimation of reserves from loglinear models. *Insurance: Mathematics and Economics* 10: 75–80.

Verrall, Richard J., Jens P. Nielsen, and Anders H. Jessen. 2010. Prediction of RBNS and IBNR claims using claim amounts and claim counts. *ASTIN Bulletin* 40: 871–87.

Zehnwirth, Ben. 1994. Probabilistic development factor models with applications to loss reserve variability, prediction intervals and risk based capital. *Insurance: Mathematics and Economics* 15: 82.