# ON DURATION EFFECTS IN NON-LIFE INSURANCE PRICING

MATHIAS LINDHOLM[*] AND TAARIQ NAZAR[†]

Abstract. The paper discusses duration effects on the consistency of mean parameter and dispersion parameter estimators in exponential dispersion families (EDFs) that are the standard models used for non-life insurance pricing. Focus is on the standard generalised linear model assumptions where both the mean and variance, conditional on duration, are linear functions in terms of duration. We derive simple convergence results that highlight consequences when the linear conditional moment assumptions are not satisfied. These results illustrate that: (i) the resulting mean estimators always have a relevant asymptotic interpretation in terms of the duration adjusted actuarially fair premium – a premium that only agrees with the standard actuarial premium using a duration equal to one, given that the expected value is linear in the duration; (ii) deviance based estimators of the dispersion parameter in an EDF should be avoided in favour of Pearson estimators; (iii) unless the linear moment assumptions are satisfied consistency of dispersion and plug-in variance estimators can not be guaranteed and may result in spurious over-dispersion.

The results provide explicit conditions on the underlying data generating process that will lead to spurious over-dispersion that can be used used for model checking. This is illustrated based on real insurance data, where it is concluded that the linear moment assumptions are violated, which results in non-negligible spurious over-dispersion.

*Keywords:* Consistency, Duration, Exponential dispersion family, Generalised linear model, Over-dispersion, Dispersion estimators

## 1. Introduction

When using generalised linear models (GLMs) for insurance pricing, it is standard to include information about duration as a known exposure see e.g. Ohlsson & Johansson (2010), Wüthrich & Merz (2023). More specifically, the data being observed consists of triplets $(Z, X, W)$, where $Z$ corresponds to the response e.g. claim amount, $X$ is a vector of covariates and $W$ is an exposure weight e.g. policy duration. The standard GLM assumption is

$$\text{(A1)} \qquad \mathbb{E}[Z \mid X, W] = W\mu(X), \quad \text{and} \quad \operatorname{Var}[Z \mid X, W] = W\sigma^2(X).$$

That is, one assumes that both the mean and the variance are linear in the exposure $W$. Furthermore, concerning the variance assumption, it is common to use Tweedie models from the exponential dispersion family (EDF), see e.g. Jørgensen (1987), Ohlsson & Johansson (2010), Wüthrich & Merz (2023), whose variances are expressed according to

$$\text{(A2)} \qquad \operatorname{Var}[Z \mid X, W] = W\varphi\mu(X)^{\xi},$$

for some $\xi \in \mathbb{R}$, and $\varphi > 0$. The parameter $\varphi$ is referred to as a dispersion parameter, and based on (A2) it is clear that this parameter can only be correctly estimated, given that $\mu(X)$ is correctly specified, see e.g. Lindholm et al. (2023). Further, when it comes to parameter estimation the influence of $W$ when using

[*] Department of Mathematics, Stockholm University; lindholm@math.su.se

[†] Department of Mathematics, Stockholm University; taariq.nazar@math.su.se

model assumption (A1) is not obvious. This is commented on in Lindholm et al. (2023).

Moreover, estimation of $\mu$ is typically based on deviance loss functions, where the minimum is attained by the maximum likelihood estimator (MLE), see e.g. Ohlsson & Johansson (2010), Wüthrich & Merz (2023). More specifically, if we consider Bregman deviance losses, that include Gaussian models with fixed dispersion, Poisson models, and Gamma models with fixed dispersion, all under assumption (A1), the deviance function for $Y = Z/W$ can be written as

$$D_{\text{Breg}}(Y, \mu) \propto W d(Y, \mu),$$

where $d(Y, \mu)$ is the so-called unit deviance function, see (5) below, with $m = 1$. Furthermore, as discussed in Lindholm et al. (2023) Section 2.1.1, the MLE for $\mu$ corresponds to the empirical version of the minimiser

$$\pi(X) \in \operatorname{argmin}_f \mathbb{E}[W d(Y, f(X))], \tag{1}$$

where the minimisation is over all $X$-measurable functions $f$ such that

$$\mathbb{E}[W d(Y, f(X))] < \infty.$$

Further, as shown in Lindholm et al. (2023) the population minimiser $\pi(X)$ is given by

$$\pi(X) = \frac{\mathbb{E}[Z \mid X]}{\mathbb{E}[W \mid X]}. \tag{2}$$

Note that (2) holds regardless of whether (A1) is satisfied or not, and does not rely on any specific assumptions regarding the dependence between, $Z, X$, and $W$. In addition, from (2) it is clear that $\pi(X)$ will differ from $\mu(X)$ unless assumption (A1) is satisfied. Thus, given that we assume a parametric form of $f$ to which the true model is a special case, this suggests that the corresponding MLE will be a consistent estimator of $\pi(X)$, which may, or may not, coincide with $\mu(X)$. Another observation is that $\pi(X)$ corresponds to the duration adjusted actuarially fair premium, since $\pi(X)$ satisfies the relation

$$\mathbb{E}[W\pi(X) \mid X] = \mathbb{E}[Z \mid X],$$

regardless of whether assumption (A1) holds or not, and does not rely on any specific assumptions regarding the dependence between, $Z, X$, and $W$. Continuing, from (2) it is clear that $\pi(X)$ will capture randomness in $W$ due to, e.g., new policies being written, existing policies that are not renewed, or policies that are annulled. On the other hand, it is clear that the reasonability in charging a premium corresponding to $\mu(X)$ depends on whether (A1) holds or not. In the numerical illustrations in Section 4 we provide an example where this does not hold, see Section 3.2 in Lindholm et al. (2023) for an additional example.

In the present paper the effects of $W$ on parameter estimation is analysed w.r.t. consistency. Note that as the sample size tends to infinity, there will eventually be a sufficiently large number of observations in any small (and shrinking) neighbourhood of any $X = x$, and we may estimate $\mu(x)$ locally as a parameter. That is, we will assume that we have data on the form $(Z_i, X_i, W_i)_{i=1}^m = (Z_i, x, W_i)_{i=1}^m$. This allows us to avoid introducing bias from a poorly specified mean model $\mu(x)$ and makes it possible to assess the validity of the linearity assumption in terms of $W$ from assumption (A1). Moreover, in practice the trouble with using a too rigid model for $\mu(X)$ can also be reduced by using flexible machine learning models. The $\mu$ function will be estimated based on deviance functions that are in agreement with (A1), but where the true underlying data generating process does not necessarily comply with model assumption (A1).

Particular focus will be on the situation when the data generating process is *assumed* to follow an EDF model with moments given by (A1) typically also assuming (A2). In agreement with the above, it is verified, see Lemma 2.1, that regardless of whether the true data generating process agrees with the EDF model assumptions, the maximum likelihood estimator (MLE) of $\mu(x)$ will always coincide with the duration adjusted actuarially fair premium $\pi(x)$ from (2). This, however, does not justify that $\widehat{\mu}$ is a consistent estimator of the mean of $Y = Z/W$, since this latter statement relies on that (A1) is satisfied. Note that from Lemma 2.1 it is seen that the estimator $\widehat{\mu}$ is a function of scaled sums of $Z$s and $W$s, which means that in order to be able to discuss convergence of $\widehat{\mu}$ under potential model misspecification we need to assume that there exists a limiting distribution for $W$, or that scaled sequences of $W$s converge.

Moreover, when analysing dispersion it is important that the mean estimator is unbiased. If this is not the case it is not meaningful to proceed with an analysis of dispersion, and local bias adjustment techniques should be applied first, see e.g. Denuit et al. (2021), Lindholm et al. (2023), Wüthrich & Ziegel (2023).

Given that the mean estimator is unbiased, we derive results for the *local* dispersion parameter estimators based on $(Z_i, X_i, W_i)_{i=1}^m = (Z_i, x, W_i)_{i=1}^m$ data and their corresponding plug-in variances. Based on the analyses of local estimators of $\varphi(x)$ we obtain conditions on the underlying data generating process that will lead to spurious over-dispersion caused by violating (A1) and (A2), see Lemmas 3.2 and 3.3. These results also allows us to assess whether a single *global* over-dispersion parameter estimator complies with (A1) and (A2). Further, Lemmas 3.2 and 3.3 rely on Pearson estimators, since these estimators turn out to be consistent, given that the underlying data generating process satisfies assumptions (A1) and (A2). This cannot be guaranteed for the corresponding Bregman deviance based over-dispersion estimators, see Lemma 3.1. In fact, Corollary 3.1 shows that even under ideal circumstances the special case of the Poisson deviance will produce an over-dispersion estimator that is not consistent unless also the mean tends to infinity. Due to this we suggest to use Pearson estimators in favour of Bregman deviance based dispersion parameter estimators.

As an alternative to using plug-in estimators of the variance, direct estimation of the variance of $Z$ is discussed in Section 3.1. This provides us with additional results that can be used for model checking purposes.

The remainder of the paper is structured as follows: Section 2 introduces relevant results relating to GLMs, EDFs, and discusses consistency of mean estimators and the relation to the duration adjusted actuarially fair premium in more detail. In Section 3 estimation and consistency of the dispersion parameter estimator is discussed together with consistency of the corresponding plug-in estimator of the variance of $Z$. This is followed by Section 3.1 where direct estimation of the variance of $Z$ using the $L^2$-loss is discussed and related to the previous results. Finally, Section 4 illustrates the theoretical results using real insurance data, and the paper ends with concluding remarks in Section 5.

## 2. Consistency of mean estimators

Assume that $Y \mid X, W$ has the following conditional density (or probability function) in the exponential dispersion family (EDF), see e.g. (Ohlsson & Johansson 2010, Sec. 2.1) and (Wüthrich & Merz 2023, Sec. 2.2),

$$f(y; \theta(x), w/\varphi) = \exp\left\{ \frac{y\theta(x) - \kappa(\theta(x))}{\varphi/w} + a(y, w/\varphi) \right\}, \tag{3}$$

where $\kappa'(\theta(x)) = \mu(x) = \mathbb{E}[Y \mid X = x, W = w]$, and where $a$ is a function normalising the density (probability) function. Thus, if we let $h(\mu(x)) = \theta(x)$ the density (probability) function (3) can be re-written according to

$$f(y; \mu(x), w/\varphi) = \exp\left\{\frac{yh(\mu(x)) - \kappa(h(\mu(x)))}{\varphi/w} + a(y, w/\varphi)\right\}. \tag{4}$$

It is now possible to introduce the weighted Bregman deviance function:

$$D_{\text{Breg}}(Y, \mu) := 2\sum_{i=1}^{m} W_i d(Y_i, \mu), \tag{5}$$

where

$$d(y, \mu) := \phi(y) - \phi(\mu) - \phi'(\mu)(y - \mu),$$

and where $\phi$ is a convex function with sub-gradient $\phi'$, see Savage (1971), Gneiting (2011) and (Wüthrich & Merz 2023, Sec. 2.3.2). By setting

$$\phi(x) := xh(x) - \kappa(h(x)),$$

it follows that

$$D_{\text{Breg}}(Y, \mu) = 2\varphi\sum_{i=1}^{m}(\log(f(Y_i, Y_i, W_i/\varphi)) - \log(f(Y_i, \mu, W_i/\varphi))), \tag{6}$$

where $f$ is from (4), see (Wüthrich & Merz 2023, Eqs. (2.28) - (2.29)). Thus, the $\mu$ obtained by minimising (5) is equivalent to the $\mu$ obtained by maximising (4) with $\mu(x) = \mu$, see also (Wüthrich & Merz 2023, Cor. 4.5).

In order to obtain a GLM for $Y = Z/W$, conditional on $X$ of dimension $k$ and $W$, introduce the so-called link-function $g$ defined such that

$$g(\mu(x)) = \beta_0 + \sum_{j=1}^{k}\beta_j x_j.$$

Examples of GLMs of particular interest in actuarial applications that are special cases of (A1) and (A2) are Gaussian models with constant dispersion parameter ($\xi = 0, \varphi = \sigma^2$), Poisson models with unit dispersion parameter ($\xi = 1$, $\varphi \equiv 1$), and Gamma models with constant dispersion parameter ($\xi = 2$), see e.g. (Wüthrich & Merz 2023, Sec. 5).

Further, by using (3) it follows that (A1) is satisfied, and, in particular, it holds that

$$\mathbb{E}[Y \mid X, W] = \mathbb{E}\left[\frac{Z}{W} \;\middle|\; X, W\right] = \mu(X),$$

together with

$$\text{Var}[Y \mid X, W] = \text{Var}\left[\frac{Z}{W} \;\middle|\; X, W\right] = \frac{1}{W}\sigma^2(X),$$

Moreover, for GLMs it is well known, that subject to regularity conditions, the maximum likelihood estimator (MLE) based on (3) of the vector $\beta$ using i.i.d. observations of triplets $(Z_i, X_i, W_i)_{i=1}^{m}$, denoted $\widehat{\beta}_m$, is consistent, see e.g. (Wüthrich & Merz 2023, Sec. 5.1.6). That is,

$$\widehat{\beta}_m \xrightarrow{p} \beta, \text{ as } m \to \infty.$$

This result, gives us, for a fixed covariate vector $X = x$, that

$$\widehat{\mu}_m(x) \xrightarrow{p} \mu(x), \text{ as } m \to \infty.$$

Concerning premium calculations, recall from the introduction that the duration adjusted premium $\pi(X)$ from (2) is the solution to the population minimisation from (1), to which the MLE is the empirical counterpart. Consequently, by using a GLM, conditional on $X = x$, it follows that $\widehat{\mu}_m(x)$ is a consistent estimator of the actuarially fair premium $\pi(x)$, and in this case $\pi(x) = \mu(x)$. Note, however, that this consistency relies on that we know that the functional form of the true model is a GLM. In practice the true model is unknown, and a misspecified mean model $\mu(X)$, or more generally a model for $\mathbb{E}[Z \mid X]$, will lead to local bias, see e.g. Lindholm et al. (2023) and Wüthrich & Ziegel (2023). This local bias will also contaminate estimation of the dispersion parameter $\varphi$.

Due to this, we will focus on the situation where we start by estimating the mean function for a specific covariate vector $X = x$, without assuming any specific functional form of $\mu(X)$. That is, we estimate $\mu(x)$ as a parameter in order to avoid introducing bias from a misspecified mean model. This is also a reasonable assumption when we consider the situation of letting the sample size tend to infinity. This also covers the situation with a mean function that can only take on a finite number of unique values, see e.g. Lindholm et al. (2023) and Wüthrich & Ziegel (2023), and the univariate local regression considered in Denuit et al. (2021).

We start with the following basic result:

**Lemma 2.1.** *Consider an i.i.d. sample* $(Z_i, X_i, W_i)_{i=1}^m = (Z_i, x, W_i)_{i=1}^m$ *and define* $Y_i := Z_i/W_i$. *The estimator* $\widehat{\mu}_m(x)$ *that minimises the duration weighted Bregman deviance from* (5) *is given by*

$$\widehat{\mu}_m(x) = \frac{\widehat{\widehat{\mathbb{E}}}_m[Z \mid X = x]}{\widehat{\widehat{\mathbb{E}}}_m[W \mid X = x]},$$

*where*

$$\widehat{\widehat{\mathbb{E}}}_m[Z \mid X = x] := \frac{1}{m}\sum_{i=1}^m Z_i, \quad \text{and} \quad \widehat{\widehat{\mathbb{E}}}_m[W \mid X = x] := \frac{1}{m}\sum_{i=1}^m W_i,$$

*for which it holds that*

$$\widehat{\mu}_m(x) \xrightarrow{p} \frac{\mathbb{E}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]}, \quad \text{as } m \to \infty.$$

*Proof.* The first part follows by direct minimisation of (5) w.r.t. $\mu$ by using that $Z = WY$, and the second part follows from Cramér-Slutsky. □

**Remark 2.1.**

(a) *Note that Lemma 2.1 does not assume independence between* $Z$ *and* $W$*, or that the true data belongs to an EDF with expectation and variance being linear in* $W$.

*Moreover, the setup used in Lemma 2.1 does not assume any explicit functional form of* $\mu(x)$*, since it is estimated as a parameter. This, of course, is identical to the situation where we assume that* $\mu$ *is a constant, when letting* $x$ *vary over the entire population, instead of conditioning on* $X = x$*. Related to this, in practice the limiting object of Lemma 2.1 will be evaluated based on local approximations in small neighbourhoods of observed points* $X = x$*. An example of this based on real insurance data is given in Section 4 where local piece-wise constant estimators are used.*

*Further, the convergence of* $\widehat{\mu}_m(x)$ *relies on that* $\widehat{\widehat{\mathbb{E}}}_m[W \mid X = x]$ *converges, or alternatively that there exists a limiting population distribution for* $W$*. This assumption is needed in order to be able to analyse convergence properties of* $\widehat{\mu}_m(x)$ *under possible model misspecification.*

(b) *The predictor* $\widehat{\mu}_m(x)$ *in Lemma 2.1 is always asymptotically actuarially fair in the sense of* $\pi(x)$ *from* (2)*. Further,* $\widehat{\mu}_m(x)$ *is not a consistent estimator of* $\mathbb{E}[Y \mid X = x] = \mathbb{E}[Z/W \mid X = x]$*, unless the true intensity is given by*

*$\mathbb{E}[Z \mid X = x, W] = W\mu(x)$, see also Proposition 2.1 in Lindholm et al. (2023) and the discussion following* (2) *above.*

(c) *As discussed in Gneiting (2011), since $d(y,\mu)$ is a consistent scoring function for predicting the mean, viewing the deviance as a loss function will always bring us closer to the true unknown mean regardless of the underlying data generating process. For a longer text book treatment, see (Wüthrich & Merz 2023, Sec. 4.1.3). For more on the situation with using duration weighted Bregman deviances (i.e. with $W_i \neq 1$), see (Lindholm et al. 2023, Sec. 2.1.1).*

Recall that standard actuarial pricing practice, see e.g. (Ohlsson & Johansson 2010, Sec. 2), suggests that the premium to be charged to policyholders is obtained by annualising the expected cost by setting $W \equiv 1$ based on (A1), i.e.,

$$\pi^*(x) := \mathbb{E}[Z \mid X = x, W = 1] = \mathbb{E}[Y \mid X = x, W = 1] = \mu(x).$$

This means that the corresponding $\widehat{\pi}^*_m(x) = \widehat{\mu}_m(x)$ is only guaranteed to be asymptotically actuarially fair in the sense of (2), which is not necessarily equal to the expectation of $Y$, since this latter property relies on assumption (A1) being true. In particular, unless (A1) holds, then

$$\widehat{\pi}^*_m(x) := \frac{\widehat{\widehat{\mathbb{E}}}_m[Z \mid X = x]}{\widehat{\widehat{\mathbb{E}}}_m[W \mid X = x]} \xrightarrow{p} \pi^*(x) \geq \mathbb{E}[Z \mid X = x], \tag{7}$$

since, in practice it will typically be the case that $\mathbb{E}[W \mid X = x] \leq 1$.

## 3. Dispersion estimators and variation

When discussing variation in observed data, and focus on the predictive setting of Lindholm et al. (2023) where $W$ is not a priori known, it is natural to consider the variance

$$\operatorname{Var}[Z \mid X] = \mathbb{E}[\operatorname{Var}[Z \mid X, W]] + \operatorname{Var}[\mathbb{E}[Z \mid X, W]]. \tag{8}$$

If we consider Tweedie models, assuming (A1) with variance functions given by (A2), we arrive at

$$\operatorname{Var}[Z \mid X] = \varphi\mu(X)^{\xi}\mathbb{E}[W \mid X] + \mu(X)^2 \operatorname{Var}[W \mid X]. \tag{9}$$

Further, even if we restrict our attention to Poisson data, assuming (A1), when we know that $\varphi = 1$ should hold, it is of interest to estimate $\varphi$ for model checking purposes. Alternatively, by estimating $\varphi$, allowing this parameter to deviate from 1, gives us a so-called over-dispersed Poisson (ODP) model, see e.g. (Ohlsson & Johansson 2010, Sec. 3.5.2) or (Wüthrich & Merz 2023, Sec. 5.4). The ODP model, however, lacks a proper distribution function and can be seen as a simple way of keeping a Poisson model for obtaining the mean function, while correcting a variance function that is seen not to agree with observed data.

When turning to dispersion and variance estimators these estimators will be centered based on mean estimators. That is, we will throughout this section assume that

$$\widehat{\mu}_m(x) \xrightarrow{p} \frac{\mathbb{E}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]}, \text{ as } m \to \infty, \tag{10}$$

and if this is not the case, local bias adjustment techniques as those in e.g. Denuit et al. (2021), Lindholm et al. (2023), Wüthrich & Ziegel (2023) should be applied first.

In order to be able to evaluate plug-in estimators of (9), we will use

$$\widehat{\widehat{\operatorname{Var}}}_m[W \mid X = x] := \frac{1}{m}\sum_{i=1}^{m}(W_i - \widehat{\widehat{\mathbb{E}}}_m[W \mid X = x])^2.$$

Further, in analogy with the argumentation in Section 2 this means that we will start by analysing the asymptotic *local* dispersion and variance estimators in small neighbourhoods surrounding $X = x$. That is, we will start by analysing estimators of $\varphi(x)$, and the results for these local estimators then allow us to assess the validity of a single global dispersion estimator.

Moreover, when it comes to estimation of $\varphi$, one alternative is to use the following deviance based estimator:

$$\widehat{\varphi}_m^{\mathrm{Breg}}(x) := \frac{1}{m-1} D_{\mathrm{Breg}}(Y, \widehat{\mu}_m(x)), \tag{11}$$

see e.g. (McCullagh & Nelder 1989, Sec. 2.3), (Ohlsson & Johansson 2010, Sec. 3.1.1) and (Wüthrich & Merz 2023, Sec. 5.3.1). When suppressing the dependence on $X$ we have the following result:

**Lemma 3.1.** *Given an i.i.d. sample $(Z_i, X_i, W_i)_{i=1}^m = (Z_i, x, W_i)_{i=1}^m$, and $\phi$ is continuous, it holds that*

$$\widehat{\varphi}_m^{\mathrm{Breg}}(x) \xrightarrow{p} 2\left(\mathbb{E}\left[W\phi\left(\frac{Z}{W}\right) \;\Big|\; X = x\right] - \mathbb{E}[W \mid X = x]\phi\left(\frac{\mathbb{E}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]}\right)\right),$$

*as $m \to \infty$.*

The proof of Lemma 3.1 is a direct consequence of Cramér-Slutsky, and does not rely on assumed independence between $Z$ and $W$, or that the true data generating process satisfies (A1).

From Lemma 3.1 it is not obvious that $\widehat{\varphi}_m^{\mathrm{Breg}}(x)$ converges in probability to $\varphi(x)$, even if the true data generating process has a likelihood which is consistent with an EDF. We continue with analysing the Poisson deviance under assumption (A1) given by

$$D_{\mathrm{Pois}}(Y, \mu) = 2\sum_{i=1}^{m} W_i(Y_i \log(Y_i) - Y_i \log(\mu) - Y_i + \mu), \tag{12}$$

see e.g. Section 2.3.1 in McCullagh & Nelder (1989), which is a special case of the Bregman deviance. This gives us the following result:

**Corollary 3.1.** *Given an i.i.d. sample $(Z_i, X_i, W_i)_{i=1}^m = (Z_i, x, W_i)_{i=1}^m$ it holds that*

$$\widehat{\varphi}_m^{\mathrm{Pois}}(x) := \frac{1}{m-1} D_{\mathrm{Pois}}(Y, \widehat{\mu}_m(x)) \xrightarrow{p} \varphi^*(x), \quad \text{as } m \to \infty,$$

*where*

$$\begin{aligned}\varphi^*(x) := 2\Bigg(&\mathbb{E}[Z\log(Z) \mid X = x] - \mathbb{E}[Z\log(W) \mid X = x] \\ &- \mathbb{E}[Z \mid X = x]\log\left(\frac{\mathbb{E}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]}\right)\Bigg).\end{aligned}$$

*If data is truly Poisson according to* (A1) *with $W = 1$ then $\varphi^*(x) := \varphi^*(\mu(x))$ satisfies*

$$\varphi^*(\mu(x)) = 1 + O(1/\mu(x)).$$

The first part of the proof follows directly from Lemma 3.1, and the proof of the asymptotic behaviour of $\varphi^*$ in terms of $\mu$ is given in Appendix B.

Corollary 3.1 reveals that the estimate for the over-dispersion parameter is inconsistent even when data is truly Poisson according to (A1) with $W = 1$. In an insurance context $\mu$ usually takes on small values, e.g., in the range 0.05 to 0.15. Clearly, in this case the dispersion parameter estimate will be erroneous. More specifically, from the proof of Corollary 3.1 it follows that a crude upper bound of $\varphi^*(x)$ is given by

$$\varphi^*(\mu(x)) = 2\mu(x)\mathbb{E}[\log((1+Z)/\mu(x)) \mid X = x] \leq 2\mu(x)\log((1+\mu(x))/\mu(x)),$$

which gives us that, e.g., $\varphi^*(\mu(x) = 0.1) \leq 0.48 < 1$. Consequently, since the estimator $\widehat{\varphi}_m^{\mathrm{Pois}}(x)$ is not consistent even under ideal circumstances, the estimator's performance with non-constant $W$ will likely be even more erratic. A numerical example illustrating this for the Poisson deviance is given in Example 5.17 in Wüthrich & Merz (2023). Lemma 3.2 is concerned with the Poisson deviance, and a similar inconsistency for Binomial data is noted in (McCullagh & Nelder 1989, Sec. 4.5.2), and in (Ohlsson & Johansson 2010, Sec. 3.1.1) numerical issues for Gamma distributed data is discussed. Due to the above we will not continue to analyse the effects of $\widehat{\varphi}_m^{\mathrm{Breg}}$ on plug-in estimators of the variance of $Z$, and advice against using deviance based estimates of $\varphi$ for model checking.

An alternative to using an estimator of $\varphi$ based on deviances is to use the corresponding Pearson estimator, see e.g. (Ohlsson & Johansson 2010, Sec. 3.1.1) and (Wüthrich & Merz 2023, Sec. 5.3.1). By assuming that data is Tweedie distributed according to (A1) with the explicit variance expression given by (A2), the (local) Pearson estimator of $\varphi(x)$ is given by

$$\widehat{\varphi}_m^{\mathrm{P}}(x) := \frac{1}{m-1}\sum_{i=1}^m \frac{W_i(Y_i - \widehat{\mu}_m(x))^2}{\widehat{\mu}_m(x)^\xi}, \tag{13}$$

and we have the following result:

**Lemma 3.2.** *Given an i.i.d. sample* $(Z_i, X_i, W_i)_{i=1}^m = (Z_i, x, W_i)_{i=1}^m$ *it holds that*

$$\begin{aligned}\widehat{\varphi}_m^{\mathrm{P}}(x) \xrightarrow{p} \varphi^{*,\mathrm{P}}(x) &:= \left(\frac{\mathbb{E}[W \mid X = x]}{\mathbb{E}[Z \mid X = x]}\right)^\xi \left(\mathbb{E}\left[\frac{Z^2}{W} \;\middle|\; X = x\right] - \frac{\mathbb{E}[Z \mid X = x]^2}{\mathbb{E}[W \mid X = x]}\right)\\ &= \overline{\varphi}(x) - \frac{\mathbb{E}[W \mid X = x]^{\xi-1}}{\mathbb{E}[Z \mid X = x]^\xi}\operatorname{Cov}\left[\frac{Z^2}{W}, W \;\middle|\; X = x\right],\end{aligned}$$

*as* $m \to \infty$*, where*

$$\overline{\varphi}(x) := \frac{\mathbb{E}[W \mid X = x]^{\xi-1}\operatorname{Var}[Z \mid X = x]}{\mathbb{E}[Z \mid X = x]^\xi}. \tag{14}$$

*If the underlying data generating process agrees with moment assumptions* (A1) *and* (A2) *then* $\varphi^{*,\mathrm{P}}(x) = \varphi(x)$.

The first part of the proof of Lemma 3.2 follows from Cramér-Slutsky without making any assumptions about the dependence between $Z$ and $W$ or the distribution of the data generating process. That $\varphi^{*,\mathrm{P}} = \varphi$ follows from direct verification. Also recall from Remark 2.1 that $\varphi(x)$ can be replaced by $\varphi$ if $\mu(x) \equiv \mu$.

From Lemma 3.2 it is seen that unless assumptions (A1) and (A2) hold the dependence between $Z$ and $W$ matters. Further, $\overline{\varphi}(x)$ from (14) can be thought of as a dispersion index, and note that given that (A1) and (A2) are satisfied it holds that

$$\overline{\varphi}(x) > \varphi(x). \tag{15}$$

Thus, if we believe that assumptions (A1) and (A2) hold, then it must hold that $\operatorname{Cov}[Z^2/W, W \mid X = x] > 0$. Consequently, by observing

$$\varphi^{*,\mathrm{P}}(x) \geq \overline{\varphi}(x), \tag{16}$$

implies that assumptions (A1) and (A2) are violated. Example A.1 shows an example where the covariance term in $\varphi^{*,\mathrm{P}}$ from Lemma 3.2 can be both positive and negative. Consequently, observing (16) based on real data indicates that assumptions (A1) and (A2) are violated.

The above analysis is concerned with *local* dispersion parameter estimators, i.e. estimators of $\varphi(x)$. In practice, however, the assumption complying with (A2) tells us that we should use a *single* over-dispersion parameter, whose Pearson estimator is given by

$$\widehat{\varphi}_m^{\mathrm{P}} := \frac{1}{m-1}\sum_{i=1}^{m}\frac{W_i(Y_i - \widehat{\mu}_m(X_i))^2}{\widehat{\mu}_m(X_i)^{\xi}}, \tag{17}$$

see e.g. Equation (3.9) in Ohlsson & Johansson (2010), where we in addition assume that

$$\widehat{\varphi}_m^{\mathrm{P}} \xrightarrow{p} \varphi^{*,\mathrm{P}}, \text{ as } m \to \infty. \tag{18}$$

Still, by observing

$$\varphi^{*,\mathrm{P}} \geq \overline{\varphi}(x), \tag{19}$$

this implies that the global $\varphi^{*,\mathrm{P}}$ violate assumptions (A1) and (A2) locally as well. In Section 4 we will evaluate both (16) and (19) using real insurance data.

Similarly, by observing (16) the plug-in estimator of $\mathrm{Var}[Z \mid X = x, W = 1]$ when assuming (A1) and (A2) satisfies

$$\widehat{\mathrm{Var}}_m[Z \mid X = x, W = 1] \xrightarrow{p} \varphi^{*,\mathrm{P}}(x)\pi(x)^{\xi} \geq \frac{\mathrm{Var}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]}, \text{ as } m \to \infty. \tag{20}$$

Hence, if

$$\frac{\mathrm{Var}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]} \geq \mathrm{Var}[Z \mid X = x, W = 1],$$

then the plug-in variance estimator from (A1) will result in local spurious over-dispersion. A situation when this occurs is described in Example A.1. The same argument applies if we instead use the global dispersion estimator (17) and (19) holds.

This brings us to the final result of this section:

**Lemma 3.3.** *Assume the moment structure given by* (A1) *and* (A2)*, i.e.* $\mathrm{Var}[Z \mid X = x]$ *is given by* (9)*. Let* $\widehat{\mathrm{Var}}_m[Z \mid X = x]$ *denote the corresponding plug-in estimator when using* $\widehat{\mu}_m$ *from Lemma 2.1, and* $\widehat{\varphi}_m^{\mathrm{P}}(x)$ *from Lemma 3.2. It then holds that*

$$\widehat{\mathrm{Var}}_m[Z \mid X = x] \xrightarrow{p} (\sigma^*(x))^2, \quad \text{as } m \to \infty,$$

*where*

$$(\sigma^*(x))^2 := \mathrm{Var}[Z \mid X = x] - \mathrm{Cov}\left[\frac{Z^2}{W}, W \;\middle|\; X = x\right] + \mathrm{Var}[W \mid X = x]\left(\frac{\mathbb{E}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]}\right)^2.$$

*Further, given that assumptions* (A1) *and* (A2) *hold, then the plug-in variance estimator is consistent.*

Thus, unless the moments of the data generating process are in agreement with assumptions (A1) and (A2), the dependence between $Z$ and $W$ matters when using plug-in estimation of the variance of $Z$, and the resulting variance estimator will be biased. Further, it is seen that

$$\mathrm{Cov}\left[\frac{Z^2}{W}, W \;\middle|\; X = x\right] < \mathrm{Var}[W \mid X = x]\left(\frac{\mathbb{E}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]}\right)^2$$

is a necessary condition for the plug-in estimator of the unconditional variance to overestimate the true variance. Overestimation of the variance has been noted in Lindholm et al. (2023) for the `freMTPL2freq` data from `CASdatasets` when analysing Poisson regression models, starting from assumption (A1) and (A2). It can be verified that in fact $\mathrm{Cov}[Z^2/W, W \mid X = x] < 0$ for this data set. For a detailed analyses of another real insurance data set, see Section 4 below.

If we instead consider a single global $\varphi$ estimator, $\widehat{\varphi}^{\mathrm{P}}_m$ from (17), the plug-in variance estimator based on (9) is given by

$$\widehat{\mathrm{Var}}_m[Z \mid X = x] = \widehat{\varphi}^{\mathrm{P}}_m \widehat{\mu}_m(x)^{\xi} \widehat{\widehat{\mathbb{E}}}_m[W \mid X = x] + \widehat{\mu}_m(x)^2 \widehat{\widehat{\mathrm{Var}}}_m[W \mid X = x], \tag{21}$$

which results in

$$\begin{aligned}\widehat{\mathrm{Var}}_m[Z \mid X = x] \xrightarrow{p} \varphi^{*,\mathrm{P}} \frac{\mathbb{E}[Z \mid X = x]^{\xi}}{\mathbb{E}[W \mid X = x]^{\xi-1}} + \mathrm{Var}[W \mid X = x] \left( \frac{\mathbb{E}[Z \mid X = x]}{\mathbb{E}[W \mid X = x]} \right)^2 \\ \geq \varphi^{*,\mathrm{P}} \frac{\mathbb{E}[Z \mid X = x]^{\xi}}{\mathbb{E}[W \mid X = x]^{\xi-1}}.\end{aligned}$$

That is, a sufficient condition for spurious over-dispersion is again given by

$$\varphi^* \geq \frac{\mathbb{E}[W \mid X = x]^{\xi-1}\, \mathrm{Var}[Z \mid X = x]}{\mathbb{E}[Z \mid X = x]^{\xi}} =: \overline{\varphi}(x),$$

where $\overline{\varphi}(x)$ is precisely the dispersion index from (14), which implies that assumptions (A1) and (A2) are violated.

To summarise this far, Lemmas 3.2 and 3.3 provide *local* conditions on the dependence between $Z$ and $W$ in order to observe spurious over-dispersion. Based on these results it is also possible to assess whether a *global* dispersion parameter estimate agrees with assumptions (A1) and (A2) or not. In Section 4 both local and global dispersion parameter estimators will be assessed, and in the next section we discuss direct estimation of $\mathrm{Var}[Z \mid X = x]$.

### 3.1. $L^2$ estimation of mean and variance parameters.

The (unweighted) $L^2$ deviance loss based on $Z$ avoids explicit assumptions on $W$ and is given by

$$\overline{D}_{L^2}(Z, \overline{\mu}) := \sum_{i=1}^{m} (Z_i - \overline{\mu})^2, \tag{22}$$

which applied to $(Z_i, X_i, W_i)_{i=1}^m = (Z_i, x, W_i)_{i=1}^m$ gives us

$$\widehat{\widehat{\mu}}_m(x) = \widehat{\widehat{\mathbb{E}}}[Z \mid X = x], \tag{23}$$

together with the duration adjusted actuarially fair premium from (2):

$$\widehat{\pi}_m(x) := \frac{\widehat{\widehat{\mathbb{E}}}[Z \mid X = x]}{\widehat{\widehat{\mathbb{E}}}[W \mid X = x]} = \frac{\widehat{\widehat{\mu}}_m(x)}{\widehat{\widehat{\mathbb{E}}}[W \mid X = x]} = \widehat{\mu}_m(x),$$

or alternatively

$$\widehat{\widehat{\mu}}_m(x) = \widehat{\widehat{\mathbb{E}}}[W \mid X = x] \widehat{\mu}_m(x). \tag{24}$$

Thus, we can directly assess the variance of $Z$ without making explicit assumptions about $W$ by using the estimator

$$\widehat{\widehat{\mathrm{Var}}}_m[Z \mid X = x] := \frac{1}{m-1} \overline{D}_{L^2}(Z, \widehat{\widehat{\mu}}_m(x)), \tag{25}$$

which is obtained by combining (22) and (23). This is always a consistent estimator of $\mathrm{Var}[Z \mid X = x]$. That is, if we in a first step estimate $\widehat{\mu}_m(x)$ in accordance with

Lemma 2.1, in order to avoid potential issues related to the dependence between $Z$ and $W$, the variance of $Z$ should be assessed using

$$\widehat{\widehat{\mu}}_m(x) = \widehat{\widehat{\mathbb{E}}}[W \mid X = x]\widehat{\mu}_m(x),$$

in accordance with (24) instead of $W_i\widehat{\mu}_m(x)$. This is in agreement with the discussion in Lindholm et al. (2023).

**Remark 3.1.** *By first estimating $\widehat{\mu}_m(x)$ and adjusting with $\widehat{\widehat{\mathbb{E}}}[W \mid X = x]$, the plug-in variance from Lemma 3.3 can be compared with the corresponding variance estimated according to* (25)*, since both estimators are based on the same mean predictor. Further, recall that the $L^2$ loss is a consistent scoring rule for the mean, see Remark 2.1. This implies that differences observed between the two variance estimators are due to violation of assumptions* (A1) *and* (A2)*. That is, the observed differences correspond to spurious over- or under-dispersion.*

## 4. Numerical illustration

In the present note we consider the `freMTPLfreq` data from `CASdatasets`. This data set has $n = 413\ 169$ policies where 96.3% of the policies have 0 claims and 3.5% of the policies have 1 claim, i.e. 0.2% of the policies have more than 1 claim and there is no policy with more than 4 claims. Concerning the policy duration, 99.9% of the policies have a duration no larger than 1, 29.4% of the policies have precisely a duration equal to 1, and the maximum observed duration is 1.99, see Figure 1(A). Further, the data set has seven covariates; four categorical and three numerical. For more information about the data set we refer the reader to Dutang & Charpentier (2020). No additional data cleaning has been done.

To reduce effects from using a poorly specified regression function, i.e., the functional form of $\mu(X)$, we use a Poisson gradient boosting machine (GBM) regression model, see Friedman (2001), in accordance with (A1) and (A2) using $\xi = 1$. The current numerical illustration uses the `gbm`-package in `R`, see Ridgeway (2020), with default parameters except that we set the interaction depth to 2 instead of 1, together with a maximum of 5 000 trees. We here only focus on in-sample behaviour and use 5-fold cross-validation for training the model, see Hastie et al. (2009). The optimal number of trees using seed 201126 for the `freMTPLfreq` data set is 192.

Let $\widehat{\mu}(X)$ denote the mean predictor obtained from the Poisson GBM, henceforth dropping the subscript w.r.t. $n$. In order to analyse the local performance of the GBM model we start by ordering $(\widehat{\mu}(x_i))_{i=1}^n$ from highest to lowest risk to obtain $(\widehat{\mu}(x_{(i)}))_{i=1}^n$, where $x_{(i)}$ corresponds to the covariate resulting in the $i$th largest in-sample mean prediction. That is, by evaluating $\widehat{\mu}$ in $x_{(i)}$ produces an estimate of the $100i/n$%th percentile of the distribution of $\widehat{\mu}(X)$. This allows us to analyse the connection between $\widehat{\mu}(x_{(i)})$ and $\mathbb{E}[W \mid X = x_{(i)}]$. As an empirical locally approximate unbiased estimator of $\mathbb{E}[W \mid X = x_{(i)}]$ we use

$$\widehat{\widehat{\mathbb{E}}}[W \mid X = x_{(i)}] := \frac{1}{2k+1}\sum_{j=i-k}^{i+k} w_{(j)},\ k < i < n-k, \tag{26}$$

where we assume that contracts that are close in $x_{(i)}$ should be reasonably close in terms of the distribution $W \mid X = x_{(i)}$. Figure 1(B) shows $\widehat{\widehat{\mathbb{E}}}[W \mid X = x_{(i)}]$ as a function of $x_{(i)}$, which is equivalent to the $100i/n$%th percentile of the distribution of $\widehat{\mu}(X)$, where the moving average calculation (26) is done using $0.5\%n = 2\ 066$ data points. If $W$ would be independent of $X$, we expect to see uniform noise, but we rather see a pronounced trend where the risk is negatively correlated with policy duration. This is similar to what is observed in Lindholm et al. (2023) for the `freMTPL2freq` data set.

Further, if assumption (A1) holds we expect to see that the expected number of claims obtained by setting $W \equiv 1$, given by

$$\widehat{\mu}_{Z|X,W=1}(x_{(i)}) := \widehat{\mu}(x_{(i)}), \tag{27}$$

should be close to

$$\widehat{\widehat{\mathbb{E}}}[Y \mid X = x_{(i)}] := \frac{1}{2k+1} \sum_{j=i-k}^{i+k} \frac{z_{(j)}}{w_{(j)}},\ k < i < n-k. \tag{28}$$

Recall that from Lemma 2.1 we know that a reasonably well specified $\mu$-model estimated using a sufficient amount of data should be close to

$$\widehat{\widehat{\pi}}(x_{(i)}) := \frac{\widehat{\widehat{\mathbb{E}}}[Z \mid X = x_{(i)}]}{\widehat{\widehat{\mathbb{E}}}[W \mid X = x_{(i)}]}, \tag{29}$$

where

$$\widehat{\widehat{\mathbb{E}}}[Z \mid X = x_{(i)}] := \frac{1}{2k+1} \sum_{j=i-k}^{i+k} z_{(j)},\ k < i < n-k. \tag{30}$$

From Figure 1(C) it is clear that $\widehat{\mu}_{Z|X,W=1}(x_{(i)})$ from (27) closely follows $\widehat{\widehat{\pi}}(x_{(i)})$ from (29), which is in agreement with Lemma 2.1. It is, however, also clear that $\widehat{\mu}_{Z|X,W=1}(x_{(i)})$ is far off from $\widehat{\widehat{\mathbb{E}}}[Y \mid X = x_{(i)}]$ from (28). This again implies that assumption (A1) is violated.

Furthermore, assume that the observed durations are representative for a typical year, neglecting the 0.1% of the data with policy durations in the range $(0, 1.99]$ years. As discussed in relation to Lemma 2.1, see Remark 2.1(b) and (7), assumption (A1) matters: The duration annualised expected number of claims is given by

$$\widehat{\mu}_{Z|X}(x_{(i)}) := \widehat{\widehat{\mathbb{E}}}[W \mid X = x_{(i)}]\widehat{\mu}(x_{(i)}), \tag{31}$$

which can be compared with the expected number of claims obtained by merely setting $W \equiv 1$, i.e. $\widehat{\mu}_{Z|X,W=1}(x_{(i)})$ from (27). From Figure 1(B) we know that the expected duration is smaller than one, which together with (7) suggests that $\widehat{\mu}_{Z|X}(x_{(i)}) \leq \widehat{\mu}_{Z|X,W=1}(x_{(i)})$. This can be observed by comparing Figure 2(A), where the solid black line corresponds to $\widehat{\mu}_{Z|X}(x_{(i)})$, with Figure 1(C), where the dashed black line corresponds to $\widehat{\mu}_{Z|X,W=1}(x_{(i)})$. Consequently, if the distribution of the duration produces Figure 2(B), then Figures 1(C) and 2(A) tells us that we will systematically overestimate the expected number of claims when using $\widehat{\mu}_{Z|X,W=1}(x_{(i)})$ instead of $\widehat{\mu}_{Z|X}(x_{(i)})$. Further, note that $\widehat{\mu}_{Z|X}(x_{(i)})$ is nicely aligned with the empirical counterpart given by (30). If this would not be the case, one should first do local bias corrections as discussed in e.g. Denuit et al. (2021), Lindholm et al. (2023), Wüthrich & Ziegel (2023), since any further analysis of dispersion would not be meaningful.

The final part of the analysis is concerned with over-dispersion and variance estimators, and we start by comparing $\widehat{\varphi}^{\mathrm{P}}(x)$ from (13) with an estimate of $\overline{\varphi}(x)$ from (14). That is, in analogy with (13) define

$$\widehat{\varphi}^{\mathrm{P}}(x_{(i)}) := \sum_{j=i-k}^{i+k} \frac{w_{(j)}(y_{(j)} - \widehat{\mu}(x_{(i)}))^2}{\widehat{\mu}(x_{(i)})}, \tag{32}$$

and as an estimator of (14) use

$$\widehat{\widehat{\varphi}}(x_{(i)}) := \frac{\widehat{\widehat{\sigma}}^2_{Z|X}(x_{(i)})}{\widehat{\mu}_{Z|X}(x_{(i)})}, \tag{33}$$

where

$$\begin{aligned} \widehat{\widehat{\sigma}}^2_{Z|X}(x_{(i)}) :=&\ \widehat{\widehat{\mathbb{E}}}[(Z - \widehat{\mu}_{Z|X}(X))^2 \mid X = x_{(i)}] \\ :=&\ \frac{1}{2k+1} \sum_{j=i-k}^{i+k} (z_{(j)} - \widehat{\mu}_{Z|X}(x_{(i)}))^2,\ k < i < n-k, \end{aligned} \tag{34}$$

which corresponds to (25). This is summarised in Figure 2(A), and it seen that the risk ordered $\widehat{\varphi}^{\mathrm{P}}$s from (32) tend to be higher than the corresponding $\widehat{\widehat{\varphi}}$s from (33); also recall relation (16). Further, when calculating the single global $\widehat{\varphi}^{\mathrm{P}}$ from (17) we get $\widehat{\varphi}^{\mathrm{P}} = 1.72$, which should be compared with the average $\widehat{\widehat{\varphi}} = 1.05$. Thus, this implies that assumptions (A1) and (A2) are violated for this data set, which here will lead to non-negligible spurious over-dispersion, as discussed in Section 3. Furthermore, from Figure 2(B) there are instances where the risk ordered (on the $\widehat{\mu}$-scale) $\widehat{\varphi}^{\mathrm{P}}$s are *lower* than the corresponding $\widehat{\widehat{\varphi}}$s. This, however, is not in conflict with Lemma 3.2, since this is an asymptotic result.

Continuing, the plug-in variance estimator when assuming (A1) and (A2), i.e. using (9), is given by

$$\begin{aligned} \widehat{\sigma}^2_{Z|X}(x_{(i)}) :=&\ \widehat{\mathrm{Var}}[Z \mid X = x_{(i)}] \\ =&\ \widehat{\widehat{\mathbb{E}}}[W \mid X = x_{(i)}]\widehat{\varphi}^{\mathrm{P}}\widehat{\mu}_{Z|X}(x_{(i)}) \\ &+ \widehat{\widehat{\mathrm{Var}}}[W \mid X = x_{(i)}](\widehat{\mu}_{Z|X}(x_{(i)}))^2, \end{aligned} \tag{35}$$

where $\widehat{\widehat{\mathrm{Var}}}[W \mid X = x_{(i)}]$ is calculated in analogy with (26). Recall that the reference observations (i.e. the dots) in Figure 2(A) correspond to $\widehat{\widehat{\mathbb{E}}}[Z \mid X = x_{(i)}]$ from (30) that are weighted sums of responses. That is, each dot in Figure 2(A) is an outcome from the random variable $\overline{Z} \mid X = x_{(i)}$, whose variance is given by

$$\mathrm{Var}[\overline{Z} \mid X = x_{(i)}] = \frac{1}{2k+1} \mathrm{Var}[Z \mid X = x_{(i)}], \tag{36}$$

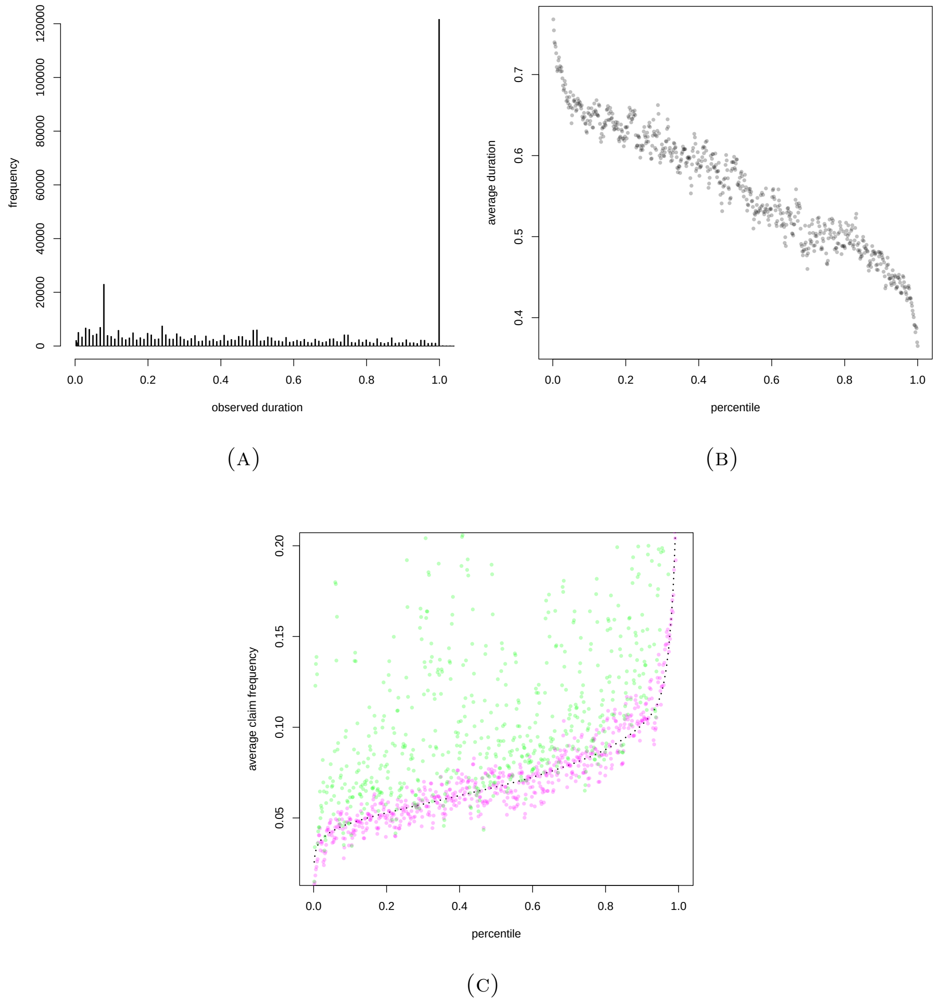
which under assumptions (A1) and (A2) is estimated using (35). Further, as discussed in Remark 3.1, the plug-in variance estimator can be compared with the locally unbiased empirical estimator from (34).

The results for the `freMTPLfreq` data is given in Figure 2(A), where the thin blue lines correspond to ± one standard deviation based on (36) calculated using (35), and the thin red lines are the corresponding standard deviations calculated using (34). From this it is seen that the plug-in variance (35) is affected by spurious over-dispersion indicating that assumptions (A1) and (A2) are violated, see Remark 3.1. Upon closer inspection, the standard deviation based on (35) is on average about 30% higher than the standard deviation based on (34). This is reasonable, since

$$\sqrt{\widehat{\varphi}^{\mathrm{P}}} \approx 1.3,\ \text{and}\quad \widehat{\widehat{\mathbb{E}}}[Z \mid X = x_{(i)}] \approx \widehat{\widehat{\mathrm{Var}}}[Z \mid X = x_{(i)}]$$

when using the corresponding locally unbiased distribution-free estimators. Finally, since we have been using a flexible GBM model, we believe that the observed over-dispersion primarily is due to model misspecification through the violation of assumptions (A1) and (A2) rather than being a consequence of using a too rigid model for describing the influence of $X$ on the response.

Further, the above analyses have been concerned with model evaluation w.r.t. both model assumptions (A1) and (A2), and the corresponding GBM predictor. An alternative is to, e.g., replace the GBM predictor with a piece-wise constant auto-calibrated predictor as in Lindholm et al. (2023). By doing so it is possible to reduce both the model complexity and the variance. For more on this, see e.g. Denuit et al. (2021), Lindholm et al. (2023), Wüthrich & Ziegel (2023).



(A)

(B)

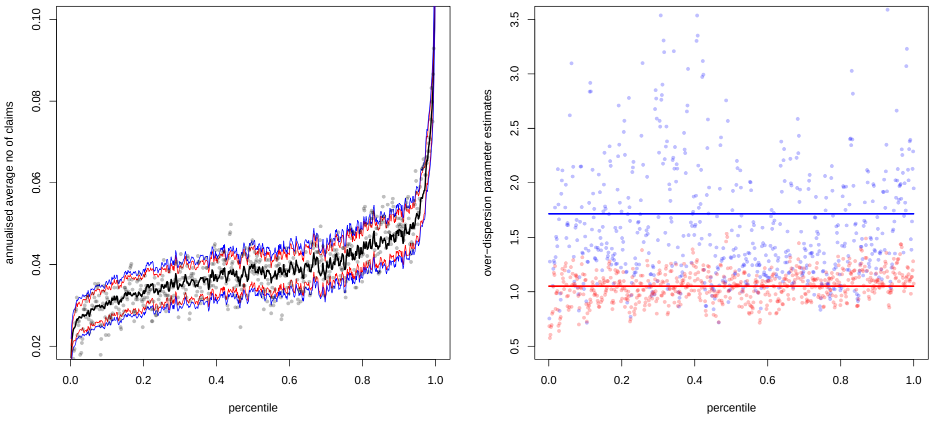
(C)

Figure 1. Analyses of `freMTPLfreq` data, consisting of $n = 413\ 169$ observations, using a Poisson GBM model. All moving averages are calculated using a centered window of size $[0.5\%n] = 2\ 066$ observations. All references to percentiles corresponds to conditioning on $X = x_{(i)}$, where $x_{(i)}$ is the observed covariate value that produces the $100i/n\%$th percentile value in the distribution of $\widehat{\mu}(X)$. The percentile values have been evaluated in $[\sqrt{n}] = 643$ equidistant points. **Panel (A)** shows a histogram of observed policy durations, $W$. Note the cut off at $W = 1$, which excludes 0.1% of the data. **Panel (B)** shows $\widehat{\widehat{\mathbb{E}}}[W \mid X = x_{(i)}]$ calculated according to (26). **Panel (C)** shows average claims frequency; the dashed line shows the GBM predictor $\widehat{\mu}(X)$ using the normalisation $W \equiv 1$, i.e. $\widehat{\mu}_{Z|X,W=1}(x_{(i)})$ from (27); the green dots show $\widehat{\widehat{\mathbb{E}}}[Y \mid X = x_{(i)}]$ from (28), and the pink dots show $\widehat{\widehat{\pi}}(x_{(i)})$ from (29).

As a final note, the above analyses are in agreement with the conclusions from Lindholm et al. (2023), where the `freMTPL2freq` data set is analysed.

## 5. Concluding remarks

For all Bregman deviance functions we will obtain the same predictor for the mean. This predictor will always be a consistent estimator of the duration adjusted



(A) (B)

FIGURE 2. Analyses of `freMTPLfreq` data, consisting of $n = 413\ 169$ observations, using a Poisson GBM model. All moving averages are calculated using a centered window of size $[0.5\%n] = 2\ 066$ observations. All references to percentiles corresponds to conditioning on $X = x_{(i)}$, where $x_{(i)}$ is the observed covariate value that produces the $100i/n\%$th percentile value in the distribution of $\widehat{\mu}(X)$. The percentile values have been evaluated in $[\sqrt{n}] = 643$ equidistant points. **Panel (A)** shows annualised premia; the solid black line shows $\widehat{\mu}_{Z|X}(x_{(i)})$ from (31); the grey dots show $\widehat{\widehat{\mathbb{E}}}[Z \mid X = x_{(i)}]$ from (30), and the thin blue and red lines show the corresponding standard deviations given by (36), centered at $\widehat{\mu}_{Z|X}(x_{(i)})$, calculated according to (35) and (34), respectively. **Panel (B)** shows Pearson estimators of $\varphi$; the blue dots correspond to $\widehat{\varphi}^{\mathrm{P}}(x_{(i)})$s from (32), the solid blue line corresponds to the global estimator $\widehat{\varphi}^{\mathrm{P}}$ from (17), the red dots correspond to $\widehat{\widehat{\varphi}}(x_{(i)})$s from (33), the solid red line corresponds to the mean of the $\widehat{\widehat{\varphi}}(x_{(i)})$s.

actuarially fair premium, regardless of the true underlying data generating process, and regardless of any potential dependence between $Z$ and $W$. It is, however, important to note that a consistent estimator of the annualised mean intensity, $\mathbb{E}[Y \mid X = x] = \mathbb{E}[Z/W \mid X = x]$, is not obtainable, unless the pro rata moment assumption from (A1) is satisfied by the underlying data generating process. These results that concern consistency of the mean estimator only relies on that there exists a limiting population distribution for $W$.

Further, not surprisingly, if the mean estimator is not consistent, this will result in inconsistent estimators of the dispersion parameter in a Tweedie model and the corresponding plug-in variances. In this situation, the dependence between $Z$ and $W$ matters and we derive conditions for when spurious over-dispersion occurs. These conditions are consequences of that the duration $W$ is included incorrectly in the modelling and can be used for model checking purposes.

All of the above has been illustrated for real insurance data using a Poisson GBM claim count model for the `freMTPLfreq` data set. For this data it is concluded that the linear moment assumptions are violated and non-negligible over-dispersion is present.

## Acknowledgments

The authors are grateful to F. Lindskog for comments and discussions. M. Lindholm gratefully acknowledges financial support from the Länsförsäkringar Alliance [Project P9/20 "Machine learning methods in non-life insurance"].

## References

Denuit, M., Charpentier, A. & Trufin, J. (2021), 'Autocalibration and tweedie-dominance for insurance pricing with machine learning', *Insurance: Mathematics and Economics* **101**, 485–497.

Dutang, C. & Charpentier, A. (2020), 'Software package `CASdatasets`'.
**URL:** *http://cas.uqam.ca/pub/web/CASdatasets-manual.pdf*

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *The Annals of Statistics* pp. 1189–1232.

Gneiting, T. (2011), 'Making and evaluating point forecasts', *Journal of the American Statistical Association* **106**(494), 746–762.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer.

Jørgensen, B. (1987), 'Exponential dispersion models', *Journal of the Royal Statistical Society: Series B (Methodological)* **49**(2), 127–145.

Lindholm, M., Lindskog, F. & Palmquist, J. (2023), 'Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells', *Scandinavian Actuarial Journal* pp. 1–28.

McCullagh, P. & Nelder, J. (1989), *Generalized linear models*, Chapman & Hall.

Ohlsson, E. & Johansson, B. (2010), *Non-life insurance pricing with generalized linear models*, Vol. 174, Springer.

Privault, N. (2011), 'Generalized Bell polynomials and the combinatorics of Poisson central moments', *The Electronic Journal of Combinatorics* **18**(1), P54.

Ridgeway, G. (2020), 'Generalized boosted models: A guide to the gbm package'.
**URL:** *https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf*

Savage, L. J. (1971), 'Elicitation of personal probabilities and expectations', *Journal of the American Statistical Association* **66**(336), 783–801.

Wüthrich, M. V. & Merz, M. (2023), *Statistical foundations of actuarial learning and its applications*, Springer Nature.

Wüthrich, M. V. & Ziegel, J. (2023), 'Isotonic recalibration under a low signal-to-noise ratio', *arXiv preprint arXiv:2301.02692* .

## Appendix A. Examples

**Example A.1.** *Assume that* $Z \mid W \sim \mathrm{Pois}(\mu W^{1+\epsilon})$*, which gives us that*

$$\mathrm{Cov}[Z^2/W, W] = \mu\,\mathrm{Cov}[W, W^{\epsilon}] + \mu^2\,\mathrm{Cov}[W, W^{1+2\epsilon}].$$

*Hence, if* $\epsilon = -1/2$ *we get that*

$$\begin{aligned}\mathrm{Cov}[Z^2/W, W] &= \mu\,\mathrm{Cov}\left[W, \frac{1}{\sqrt{W}}\right] \\ &= \mu\left(\mathbb{E}[\sqrt{W}] - \mathbb{E}[W]\mathbb{E}\left[\frac{1}{\sqrt{W}}\right]\right) \\ &\le \mu\mathbb{E}[\sqrt{W}]\left(1 - \mathbb{E}[\sqrt{W}]\mathbb{E}\left[\frac{1}{\sqrt{W}}\right]\right) \le 0,\end{aligned}$$

*due to Jensen's inequality, and if* $\epsilon = 0$

$$\mathrm{Cov}[Z^2/W, W] = \mu^2\,\mathrm{Var}[W] \ge 0,$$

*as it should. Further, if $\epsilon = 1$ it follows that*

$$\mathrm{Cov}[Z^2/W, W] = \mu \,\mathrm{Var}[W] + \mathrm{Cov}[W, W^3] \geq 0,$$

*where the inequality follows by noting that $\mathbb{E}[W] \leq \mathbb{E}[W^4]^{1/4}$ and $\mathbb{E}[W^3] \leq \mathbb{E}[W^4]^{3/4}$, which follows from concavity and Jensen's inequality.*

*The above illustrates that the covariance may become both negative and positive.*

## Appendix B. Proof of Lemma 3.2: the asymptotics of $\varphi^*$ in terms of $\mu$

In order to simplify the exposition (possible) dependence on $X = x$ is suppressed. From the first part of the lemma we have that

$$\varphi^*(\mu) = 2(\mathbb{E}[Z \log(Z)] - \mu \log(\mu)).$$

Taylor expanding the first term inside the expectation around $Z = \mu$ yields

$$\varphi^*(\mu) = 2\left(\frac{1}{2} + \sum_{k=3}^{\infty} \frac{(-1)^k}{k} \frac{\mathbb{E}[(Z-\mu)^k]}{\mu^{k-1}}\right).$$

Thus, if it holds that $\mathbb{E}[(Z-\mu)^k] = O(\mu^{k-2}),\ k \geq 3$, it follows that

$$\varphi^*(\mu) \longrightarrow 1, \text{ as } \mu \longrightarrow \infty.$$

We start by proving the following result which ascertains that the central moments are bounded in the desired way:

**Proposition B.1.** *Let $Z \sim \mathrm{Pois}(\mu), \mu > 0$. It holds that*

$$\mathbb{E}[(Z-\mu)^n] = O(\mu^{n-2}),\ \forall n \geq 3.$$

**N.B.** It is possible to show that certain central moments are of even lower order than $O(\mu^{n-2})$, but this is not needed for the current purposes.

To prove Proposition B.1 we use the following lemmas.

**Lemma B.1.** *For $m = 0, 1, 2$ it holds that*

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} k^m = 0,$$

*together with*

$$\sum_{k=0}^{n-1} (-1)^k \binom{n}{k} k^m = (-1)^{n+1} n^m.$$

*Proof.* The result for $m = 0$ follows directly from the binomial theorem:

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} 1^{n-k} (-1)^k = (1-1)^n = 0.$$

For $m = 1, 2$ we take a different approach. Examine the function

$$g(x) = \sum_{k=1}^{n} (-1)^k \binom{n}{k} x^k = (1-x)^n - 1.$$

The last equality holds from the binomial theorem (subtract 1 since we are summing from $k = 1$). Clearly

$$g'(1) = \sum_{k=1}^{n} (-1)^k \binom{n}{k} k = -n(1-1)^{n-1} = 0,$$

from which it follows that

$$g''(1) = \sum_{k=1}^{n} (-1)^k \binom{n}{k} k(k-1) = n(n-1)(1-1)^{n-2} = 0,$$

and it, hence, holds that

$$\sum_{k=1}^{n} (-1)^k \binom{n}{k} k^2 = \sum_{k=1}^{n} (-1)^k \binom{n}{k} k = 0.$$

This finishes the proof of the first part. The proof of the second part follows by splitting the sum into two parts: the sum from 0 to $n-1$ and the $n$:th term and finally carrying over the $n$:th term to the r.h.s. This concludes the proof of the lemma. □

The second lemma needed to prove the proposition is the following:

**Lemma B.2.** *Let* $Z \sim \text{Pois}(\mu), \mu > 0$*. It then holds that*

$$\mathbb{E}[(Z-\mu)^n] = \sum_{a=0}^{n} \mu^a S_2(n,a), \qquad \forall n \in \mathbb{N},$$

*where*

$$S_2(a,b) = \sum_{k=0}^{b} (-1)^k \binom{a}{k} S(a-k, b-k)$$

*and* $S(m,n)$ *are Stirling numbers of the second kind.*

*Proof.* See Privault (2011). □

*Proof of Proposition B.1.* Given Lemma B.2, what remains to show is that

$$S_2(n,n) = S_2(n,n-1) = 0.$$

To start off, for Stirling numbers of the second kind it holds that

$$S(n,n) = 1 \text{ and } S(n,n-1) = \frac{n(n-1)}{2}, \quad \forall n \in \mathbb{N}.$$

Next, note that

$$S_2(n,n) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} S(n-k, n-k) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0,$$

where the last equality follows from Lemma B.1. Continuing,

$$\begin{aligned}
S_2(n,n-1) &= \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} S(n-k, n-k-1) \\
&= \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} \frac{(n-k)(n-k-1)}{2} \\
&= \frac{1}{2} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (n^2 - 2nk + k^2 - n + k) \\
&= \frac{1}{2} \Bigg( (n^2 - n) \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} - (2n-1) \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} k \\
&\quad + \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} k^2 \Bigg) \\
&= \frac{(-1)^{n+1}}{2} \left( n^2 - n - (2n-1)n + n^2 \right) = 0,
\end{aligned}$$

where the last equality holds from Lemma B.1. This proves the proposition. □