Isotonic regression for variance estimation and its role in mean estimation and model validation

Łukasz Delong* Mario Wüthrich†

January 12, 2024

Abstract

We study isotonic regression which is a non-parametric rank-preserving regression technique. Under the assumption that the variance function of a response is monotone in its mean functional, we investigate a novel application of isotonic regression as an estimator of this variance function. Our proposal of variance estimation with isotonic regression is used in multiple classical regression problems focused on mean estimation and model validation. In a series of numerical examples, we (1) explore the power variance parameter of the variance function within Tweedie's family of distributions, (2) derive a semi-parametric bootstrap under heteroskedasticity, (3) provide a test for auto-calibration, (4) explore a quasi-likelihood approach to benefit from best-asymptotic estimation, (5) deal with several difficulties under lognormal assumptions. In all these problems we verify that the variance estimation with isotonic regression is essential for proper mean estimation and beneficial compared to traditional statistical techniques based on local polynomial smoothers.

Keywords. Isotonic regression, generalized linear model, Tweedie's family, power variance parameter, quasi-likelihood, lognormal model, auto-calibration, T-reliability diagram, bootstrap, best-asymptotic normal.

1 Introduction

We start by introducing our general statistical set-up. We consider a response variable Y with conditional mean and variance, respectively,

$$\mathbb{E}[Y|\boldsymbol{x}] = \mu(\boldsymbol{x}, \boldsymbol{\theta}) \qquad \text{and} \qquad \mathrm{Var}[Y|\boldsymbol{x}] = \phi\, V(\mu(\boldsymbol{x}, \boldsymbol{\theta})), \tag{1.1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is a d-dimensional feature vector (covariate) that characterizes the response Y, and $\boldsymbol{\theta} \in \mathbb{R}^k$ is the k-dimensional parameter vector of the functional $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}, \boldsymbol{\theta})$. We follow the generalized linear model (GLM) framework. We set $k = d$, and we assume GLM structure

$$g(\mu(\boldsymbol{x}, \boldsymbol{\theta})) = \eta(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^\top \boldsymbol{\theta}, \tag{1.2}$$

*University of Warsaw, Faculty of Economic Sciences, l.delong@uw.edu.pl
†ETH Zürich, Department of Mathematics, mario.wuethrich@math.ethz.ch

for a given strictly monotone and smooth link function g. The function η is called linear predictor since it defines a linear regression function in $\boldsymbol{x}$. The function $\mu \mapsto V(\mu)$ in (1.1) is a variance function. It describes the relation between the variance and the expected value of the response Y. $\phi > 0$ is a dispersion parameter. If we deal with claim numbers, we often assume that Y has a Poisson distribution. This assumption implies a linear variance function, that is, $V(\mu) = \mu$. If we deal with claim severities, we may assume that Y has a Gamma distribution or an Inverse Gaussian distribution, which implies $V(\mu) = \mu^2$ or $V(\mu) = \mu^3$. More generally, if we assume that Y has a Tweedie's distribution we have power variance function $V(\mu) = \mu^p$ for some $p \geq 1$ or $p \leq 0$. As far as the variance function is concerned, in this paper we work under the assumption that

$$\mu \mapsto V(\mu) \quad \text{is non-decreasing.} \tag{1.3}$$

This assumption is satisfied in the class of Tweedie's distributions with power variance parameters $p \geq 1$. It is seems to be a reasonable assumption in severity modeling as large claims in size are usually more volatile.

Under the moment assumptions (1.1)-(1.3), the goal is to estimate the regression function or the mean functional from data, i.e., to estimate the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ from independent data points $(Y_i, \boldsymbol{x}_i)_{i=1}^n$, assumed that they all follow the same GLM with the moments as specified above. The classical actuarial example is the estimation of the pure risk premium of motor insurance policies based on the individual features of the policyholders. In this application, an actuary is only interested in the expected value of the loss distribution, as this is sufficient for pricing based on the Law of Large Numbers and the diversification principle.

Our statistical set-up may seem to be rather restrictive as in many statistical, econometrics and actuarial problems the variance function (1.1) may take a more general form. There are more flexible parameterizations of the first two moments of a response variable in regression modeling and there are also fully flexible distributional approaches to regression modeling, such as a double GLM, GAMLSS (Generalized Additive Models for Location, Scale and Skewness), mixtures of experts (MoE), multi-task neural networks for mean and variance estimation, just to name a few approaches. However, the GLM set-up (1.1)-(1.3) is still the most commonly used one in actuarial practice. Actuaries mostly use Tweedie's GLMs with a power variance function with $p \geq 1$, and the power variance function, which is strictly increasing in the mean, has proved to fit many actuarial problems sufficiently well; see, e.g., Denuit et al. (2019) and Wüthrich–Merz (2023). An other argument is that more advanced models may be too complex in applications (e.g., if we have claim sizes from an Inverse Gaussian distribution and we start with a Gamma GLM, it is easier to upgrade the variance function from a quadratic to a cubic behavior, rather than fitting a separate regression to the dispersion coefficient or the second central moment). This motivates our model selection, and we will not add any unnecessary sophistication into our statistical model, but we rather aim at improving and simplifying the estimation procedure of the mean and model validation.

Even though the goal is to estimate the mean functional $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}, \boldsymbol{\theta})$, the variance function $\mu \mapsto V(\mu)$ plays a crucial role in statistical estimation and inference on finite samples:

- Problem 1: Asymptotically efficient estimation of mean functionals by quasi-likelihood methods.

- Problem 2: Back-transformation of maximum likelihood estimates of means and variances on

the log scale to the original scale, and estimation of mean values of lognormally distributed responses.

- Problem 3: Tests, validation plots and bootstrap methods, especially for the purpose of verifying the auto-calibration property of mean estimates.

In the statistical literature one can find many papers that discuss the above problems and the role of the variance function in mean estimation and model validation. Our main contribution is to revise classical regression problems focused on mean estimation and model validation, and to investigate isotonic regression for variance estimation. To the best of our knowledge, our paper presents novel applications of isotonic regression in the field of variance modeling. We believe that the application of isotonic regression in Problems 1–3 is rather straightforward and simpler than traditional statistical methods for variance modeling, such as local polynomial regression. In a series of numerical examples we illustrate the advantages of isotonic regression. The main benefit of isotonic regression is that it does not use hyperparameters, hence, we do not need to spend time on fine–tuning and optimizing hyperparameters, which is particularly beneficial if we have to fit multiple recursive regressions or complex regressions such as neural network models.

Let us start by giving a motivation for considering Problem 1. If the distribution of the response is properly specified, and if the distribution is a member of the exponential dispersion family (EDF), the parameter $\boldsymbol{\theta} \in \mathbb{R}^k$ can be estimated with maximum likelihood estimation (MLE) by minimizing the deviance loss function L of the chosen distribution from the EDF, that is, subject to existence and uniqueness, we solve

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} L(Y_i, \mu(\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{1.4}$$

In practice, it is very unlikely that the response exactly follows the chosen distribution. Hence, in practice, we almost always misspecify the distribution and the likelihood function. Fortunately, we can apply the quasi-likelihood method for GLMs developed, among others, by Wedderburn (1974), McCullagh (1983), Nelder–Pregibon (1987), Firth (1987) and Gourieroux et al. (1984). Let us choose any deviance loss function L derived from a particular distribution from the EDF, e.g., the Poisson, the Gamma or Tweedie's deviance loss. We still use the GLM model equation for the mean value of the response, that is, for $i = 1, \ldots, n$, we assume mean structure

$$\mathbb{E}[Y_i|\boldsymbol{x}_i] = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\theta}), \tag{1.5}$$

but we do not require that the response comes from the distribution used for specifying the deviance loss function L. In fact, we do not even require that the second moment assumption implied by the deviance loss function is correct. We can still estimate $\boldsymbol{\theta}$ by minimizing objective function (1.4). From Theorems 1-3 in Gourieroux et al. (1984) we know that this estimate is strongly consistent; this is related to the fact that deviance loss functions are strictly consistent scoring functions for mean estimation, see Gneiting (2011). The important point in Gourieroux et al. (1984) is that we can gain asymptotic estimation efficiency in the quasi-likelihood method by properly specifying the variance function of the response, that is, if we have the right structure and level in the second

moment

$$\mathrm{Var}[Y_i|\boldsymbol{x}_i] = \phi\, V(\mu(\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{1.6}$$

If we correctly specify the variance function $\mu \mapsto V(\mu)$ and insert it into the deviance loss function (1.4), then the resulting estimator is best-asymptotically normal, i.e., it achieves asymptotically the smallest variance among all estimators found by minimizing deviance loss functions within the EDF; see Theorem 4 in Gourieroux et al. (1984). Moreover, it has asymptotically the smallest variance among all unbiased estimators for which the influence function is linear; see McCullagh (1983) and Firth (1987).

To the best of our knowledge, the estimation of variance function and its impact on mean estimation has not been considered in an actuarial context in the framework of GLMs. At the same time, the problem is well-known in statistics and regression modeling. Carroll (1982) and Müller–Stadtmüller (1987) show that mean estimation of a linear regression function using weighted least squares with non-parametrically estimated variances is asymptotically equivalent to weighted least squares with known variances. Hall–Carroll (1989) and Ruppert et al. (2012) show that a variance function can be estimated non-parametrically with smoothers with an accuracy which is optimal if the mean was known. Chiou–Müller (1999) extend Carroll (1982) and Müller–Stadtmüller (1987). Chiou–Müller (1999) introduce a non-parametric quasi-likelihood method and show that if the variance function is estimated with a local polynomial regression then the asymptotic limiting distribution of the vector of parameters in the mean functional is the same as for the quasi-likelihood estimates obtained under correct specification of the variance function. Hence, Theorem 4 from Gourieroux et al. (1984) applies even if the variance of the response is (properly) estimated. Davidian–Carroll (2012) suggest to iteratively re-fit mean and variance to improve the accuracy of both estimates.

The non-parametric quasi-likelihood estimation from Chiou–Müller (1999) gives a theoretical solution to Problem 1. However, the practical application of this solution might be challenging. It is known that without a carefully chosen kernel, properly specified bandwidth of the kernel and degree of the local polynomial, good estimates with local polynomial regressions are difficult to obtain in the framework of smoothing methods. In particular, one has to carefully fine–tune and optimize the hyperparameters of the smoother, and the results can differ significantly depending on the chosen values of the hyperparameters. Hence, there is the need to have a simple and straightforward estimation of the variance function that does not involve hyperparameters. Our proposal is to use isotonic regression instead of a local polynomial regression. Our first contribution is to demonstrate the impact on mean estimation of a misspecified variance function in an actuarial example with skewed and heavy-tailed claim sizes, and we present a non-parametric quasi-likelihood algorithm for mean estimation where the variance function is estimated non-parametrically with isotonic regression. We also illustrate that the variance estimation becomes less important for increasing sample size, but it is still important even for very large samples if claim sizes are skewed and heavy-tailed. These results are of significant importance to actuaries.

Problem 2 is closely related to Problem 1. In many practical actuarial applications, the loss distribution of the response is (close to) lognormal. If we cannot reject the assumption of lognormality, then it is natural to apply MLE to estimate $\boldsymbol{\theta}$ in (1.2). To apply MLE to a lognormal distribution, we transform the responses to the log scale and we fit a (Gaussian) linear regression model to the

logged responses. The difficulty is that the mean estimates of the lognormal responses on the original scale involve both mean and variance estimates (on the log scale). Consequently, the application of this estimation method is limited since, in addition to the mean estimation on the log scale, we have to accurately estimate the variances of the responses on the log scale. Under heteroskedasticity this may be challenging. To deal with Problem 2 we could apply the non-parametric likelihood method of Chiou–Müller (1999). Alternatively, we propose to use isotonic regression instead of a local polynomial regression for variance estimation. Our second contribution concerns the MLE of mean values of lognormally distributed responses. In particular, we propose a new algorithm for MLE of the mean values under the lognormal assumption and a GLM type parameter specification (1.1)-(1.2), where the conditional variance of the response is estimated non-parametrically with an isotonic regression. We believe that this fills a gap in actuarial statistics by considering the estimation of a lognormal regression model with a GLM type specification of its moments.

Finally, Problem 3 is about validation of estimated mean values. One property that should be validated is the auto-calibration property of a predictor; see Pohle (2020), Gneiting–Resin (2021), Krüger–Ziegel (2021), Denuit et al. (2021), Fissler et al. (2022), Chapters 5.1.5 and 7.4.2 in Wüthrich–Merz (2023) and Wüthrich–Ziegel (2023). The auto-calibration property can be validated with reliability diagrams (in particular, CORP reliability diagrams) which compare predictions with observations (predictions with recalibrated versions of the predictions). Dimitradis et al. (2022) and Dimitradis et al. (2020) advocate the use of isotonic regression to construct CORP reliability diagrams. If the graph of a CORP reliability diagram lies close to its diagonal, the mean estimator is auto-calibrated. From the statistical point of view, we would like to formally evaluate the significance of the observed deviations from the diagonal and we would like to derive a consistency band for a CORP reliability diagram. This problem is new and is still under investigation in the statistical and the actuarial literature. In the framework of GLMs, consistency bands, confidence intervals and critical values of tests can be derived with bootstrap techniques; we follow this approach here. Non-parametric bootstrap usually uses Pearson's residuals. Pearson's residuals require accurate variance estimates which need to be estimated in the quasi-likelihood framework, in addition to the mean values – this is the point where the variance estimation with isotonic regression comes into play. Apart from variance estimation, there is an additional difficulty which we may face when modeling claim sizes. Since the claim sizes are skewed and heavy-tailed, the resulting Pearson's residuals are also skewed and heavy-tailed and, consequently, we should not directly sample from them as the bootstrap samples may violate the positivity constraint of claim sizes. Instead, we propose to sample quantile residuals to construct a consistency band for a CORP reliability diagram. In this paper, we also develop a statistical test for validating the auto-calibration property of a predictor based on a CORP reliability diagram and its consistency band. We show how to find the critical value of the test with bootstrap. Our third contribution is to develop several diagnostic tools for mean estimation which require accurate variance estimates. We demonstrate how to derive consistency bands for CORP reliability diagrams and critical values of miscalibration tests based on strictly consistent scoring functions (deviance losses) with bootstrap techniques.

In Section 2 we describe the synthetic data sets investigated in this paper. We start with assessing the performance of a classical Gamma GLM fitted to our synthetic claim sizes. Next, we improve the initial estimates by applying our estimation methods. We also implement our validation methods. The general workhorse we use is isotonic regression, see Section 3. We illustrate this in

four examples:

- Validation of Tweedie's power variance function (Section 4);
- Quantile residuals and bootstrap techniques used for deriving consistency bands for CORP reliability diagrams and distribution of a miscalibration test (Section 5);
- Quasi-likelihood estimation of mean values with a variance function estimated using an isotonic regression (Section 6);
- MLE of mean values for lognormally distributed responses with variance functions estimated from an isotonic regression and back-transformation of the estimates from the log scale to the original scale (Sections 7 and 8).

Finally, in Section 9 we conclude.

2 Synthetic data sets

In our examples, we use three synthetic data sets that are built on the data set `swmotorcycle` from the R package CASdatasets; see Dutang–Charpentier (2018). In our experiments we consider claims sizes. We apply the same data cleaning and transformations to `swmotorcycle` as described in Chapter 13.2 of Wüthrich–Merz (2023). In the first step, we estimate a regression function for the claim sizes on the original data (after cleaning) with a classical Gamma GLM using the log-link function for g, see (1.2). As features $\boldsymbol{x}$ we include the following variables into the regression function: `OwnerAge`, `OwnerAge`^{2}, `Gender`, `Area`, `RiskClass`, `VehAge`, `VehAge`^{2}, `VehAge`^{3}, `VehAge`^{4}; see Chapter 5.3.7 in Wüthrich–Merz (2023) and Section 4.2 in Delong et al. (2021). In the next step, we create our synthetic data sets.

We discuss how synthetic Data Set 1 was created. We specify the features, the expected values and the variances of the claim sizes for all instances in our synthetic portfolio, together with the distribution function of the claims sizes. Hence, we know the ground truth about the distribution of the claim sizes and their parameters. We sample with replacement $n = 20,000$ instances (feature values $(\boldsymbol{x}_i)_{i=1}^n$) from the original data set. We calculate the expected claim sizes of the given features $\boldsymbol{x}_i$ using the estimated Gamma GLM regression function from the first step, and we assume that these estimates are the true mean values $\big(\mu(\boldsymbol{x}_i)\big)_{i=1}^n$ of the claim sizes Y_i in our synthetic Data Set 1. The empirical density of the true mean values $\big(\mu(\boldsymbol{x}_i)\big)_{i=1}^n$ is presented in Figure 2.1 (top-left). Since we are mainly concerned about variance estimation and its impact on mean estimation, we choose a bit a special variance function which has different regimes for different ranges of expectations:

$$\mu > 0 \;\mapsto\; V(\mu) = \mu^2 \mathbf{1}\{\mu < \mu_1^*\} + \big(c_1 + \mu^2 \log(\mu)\big)\mathbf{1}\{\mu_1^* \le \mu < \mu_2^*\} + \big(c_2 + \mu^3\big)\mathbf{1}\{\mu \ge \mu_2^*\}; \quad (2.1)$$

see Figure 2.1 (top-right). The thresholds $0 < \mu_1^* < \mu_2^* < \infty$ were chosen as the 50% and the 90% deciles of the expected values $\big(\mu(\boldsymbol{x}_i)\big)_{i=1}^n$ calculated from the synthetic Data Set 1. This choice of the variance function is artificial, but the motivation behind this choice is to generate skewed and heavy-tailed claim sizes with a non-trivial variance function which we are going to estimate from the data with our algorithms. With our choice of the variance function, we assume that the observations with expected values below μ_1^* have a light-tailed Gamma variance function, the

observations with expected values in the interval (μ_1^*, μ_2^*) have a heavier tail, between the Gamma and the Inverse Gaussian variance function, and the observations in the far right tail with expected values above μ_2^* have a more heavy-tailed Inverse Gaussian variance function. In actuarial practice, claim sizes usually come from distributions with different tails (claim types), and large claims have more heavy-tailed distributions than attritional claims – our variance function reflects this phenomenon. The constants c_1, c_2 were chosen so that the variance function is continuous at the slicing points μ_1^* and μ_2^*. Finally, we simulate $n = 20,000$ claim sizes $(Y_i)_{i=1}^n$ from lognormal distributions with the two moments $\big(\mu(\boldsymbol{x}_i), \phi V(\mu(\boldsymbol{x}_i))\big)_{i=1}^n$ specified as above. For the simulations, the dispersion parameter was set to $\phi = 0.015$. The coefficient of variation of the claim sizes in the synthetic Data Set 1 is 3.07 compared to 1.48 in the original data set. The original claim sizes from `swmotorcycle` are light-tailed, see Chapter 5.3.7 in Wüthrich–Merz (2023), and our goal is to work with skewed and heavy-tailed claim sizes, since such claim sizes are most challenging to actuaries, violate assumptions of classical GLMs and allow us to present interesting results. In Figure 2.1 (bottom) we present the empirical distribution of the standardized logged claim sizes. The unconditional distribution is indeed very skewed and heavy-tailed due to the assumed variance function. In the sequel, we integrate the dispersion parameter ϕ into the variance function V, and we re-define $V(\mu) := \phi V(\mu)$.

Synthetic Data Set 2 was created in the same way as synthetic Data Set 1, but we sampled $n = 100,000$ instances from the original data set and simulated $n = 100,000$ claim sizes from lognormal distributions. Given that claim frequencies are low in motor insurance, a sample size of $20,000$ reflects a medium-size insurer and a sample size of $100,000$ is a larger insurer (2 mio. insurance policies if the claim frequency is 5%).

Synthetic Data Set 3 was created in a slightly different way. We take $n = 20,000$ instances from the synthetic Data Set 1 and their logged mean values $\eta(\boldsymbol{x}_i) = \log(\mu(\boldsymbol{x}_i))$ calculated with the linear predictor from the Gamma GLM fitted to the original data set. We now assume that these values define the true expected values of the claim sizes on the logarithmic scale (the expected values of $Z_i = \log(Y_i)$). The true variances of the claim sizes on the logarithmic scale are defined based on the variance function (2.1) with the following transformation:

$$\eta \in \mathbb{R} \;\mapsto\; V(\eta) = \log\left(1 + \frac{V\big(e^\eta\big)}{\big(e^\eta\big)^2}\right). \tag{2.2}$$

We simulate $n = 20,000$ claim sizes $(Y_i)_{i=1}^n$ from lognormal distributions with the two moments on the logarithmic scale $\big(\eta(\boldsymbol{x}_i), \phi V(\eta(\boldsymbol{x}_i))\big)_{i=1}^n$ specified as above, with $\phi = 0.015$. The reason for this construction will become clear in Section 7, below.

Unless otherwise stated, we study the claim sizes from the synthetic Data Set 1. We start with the classical approach in actuarial modeling and fit a Gamma GLM with quadratic variance function $V(\mu) = \mu^2$ and log-link function $g(\mu) = \log(\mu)$ to our observations in Data Set 1. The parametric form of the linear regression function for mean estimation of the response is properly specified, i.e., we use the same features as regressors in the Gamma GLM that have also been used in the first regression function to create the synthetic data set. However, based on (2.1), the variance function and the distribution of the responses are misspecified. As a result, the Gamma GLM provides a very poor fit in terms of mean estimates, see Figure 2.2 which compares the estimated means against their true values. In particular, the classical approach with a Gamma

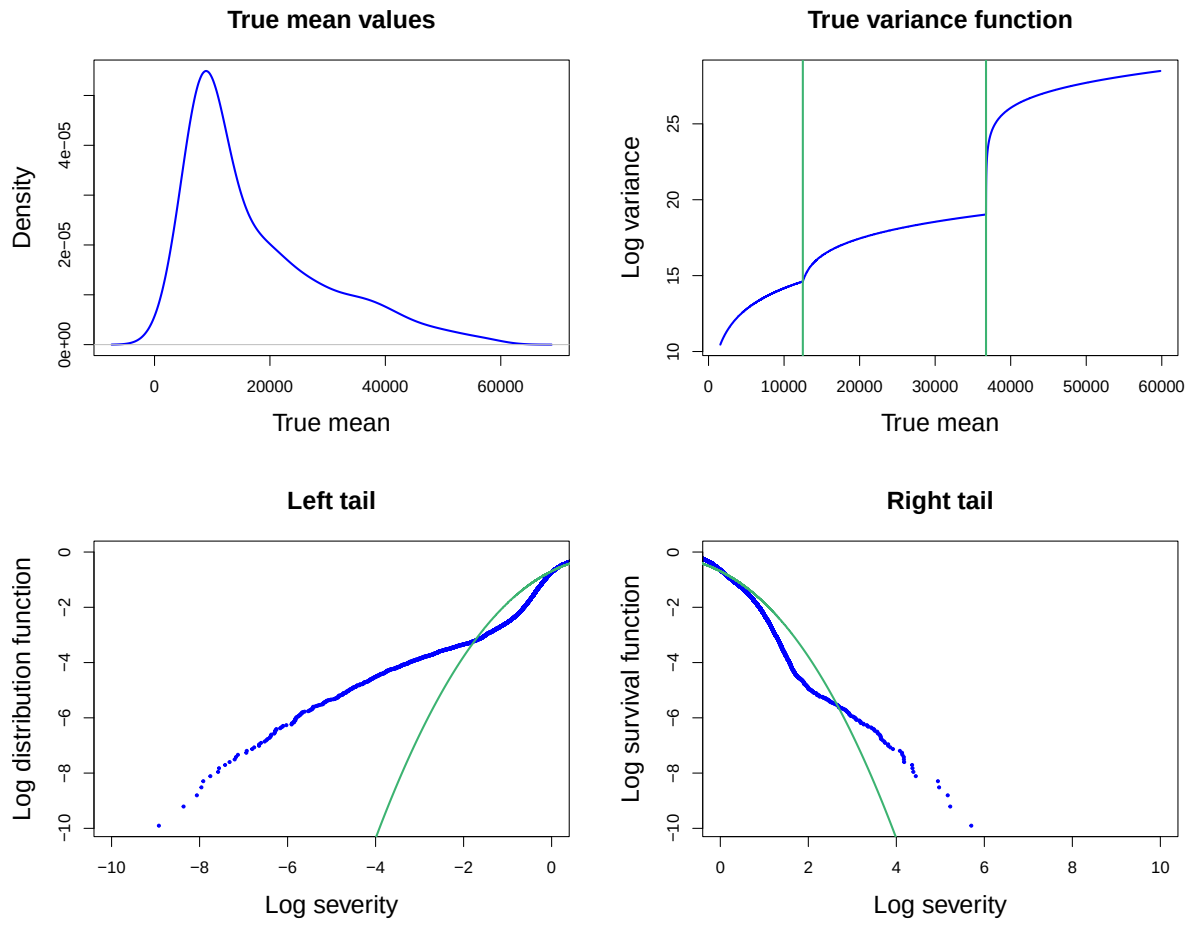

Figure 2.1: Top: Empirical density of the true mean values of the claim sizes, and the choice of the true variance function in the Data Set 1; the green lines show the 50% and 90% deciles of the true mean values used as the splicing points for the variance function. Bottom: Empirical unconditional distribution of the standardized logged claim sizes in the Data Set 1; the green lines show the standard normal distribution.

GLM under- and over-estimates the true means by $\pm 30\%$ for our data set, especially in the tails, see Figure 2.2 (right). This happens because the Gamma model cannot properly explain the small and large claims. Figure 2.2 gives a warning to the modeler that the variance function should not be neglected in the estimation procedure. Instead of a Gamma GLM with quadratic variance function, we could also fit Tweedie's GLMs with different power variance parameters $p \geq 1$. In all cases, the mean estimates turned out to be poor. Our goal is to improve mean estimation by properly modeling the variance function.

In practice, the true means are unknown. In the sequel, we validate the mean estimates with our diagnostic tools. To confirm the conclusions based on the diagnostic tools, we always compare the estimated means also against the true means.

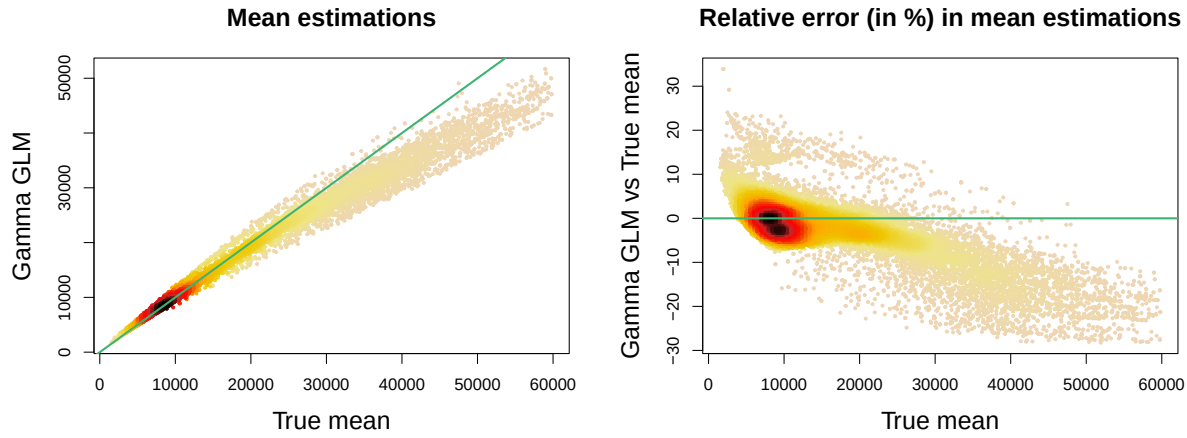

Figure 2.2: Estimated means from the Gamma GLM against true means for all instances; the darker the color, the larger the density of the observations with the particular means.

3 Isotonic regression

The general purpose tool that we are going to use in this paper is isotonic regression. Isotonic regression is a rank based non-parametric regression approach that preserves monotonicity in pre-specified ranks $(\pi(\boldsymbol{x}_i))_{i=1}^n$; we refer to Ayer et al. (1955), Brunk et al. (1957) and Barlow et al. (1972). Assume we have data points $(Y_i, \pi(\boldsymbol{x}_i))_{i=1}^n$ and positive case weights $(w_i)_{i=1}^n$. Would like to fit a non-parametric regression model to the responses $(Y_i)_{i=1}^n$ which respects the ranks $(\pi(\boldsymbol{x}_i))_{i=1}^n$. That is, we consider the following optimization problem

$$\begin{aligned} \widehat{\boldsymbol{\mu}} \;&=\; \underset{\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^\top \in \mathbb{R}^n}{\arg\min} \; \sum_{i=1}^n w_i \left(Y_i - \mu_i\right)^2, \\ &\qquad \text{subject to } \mu_j \le \mu_k \Longleftrightarrow \pi(\boldsymbol{x}_j) \le \pi(\boldsymbol{x}_k) \text{ for all } 1 \le j,k \le n. \end{aligned} \tag{3.1}$$

Let us remark that, generally, $\boldsymbol{x}_i$ is multi-dimensional but $\pi(\boldsymbol{x}_i)$ is one-dimensional, e.g., $\pi(\boldsymbol{x}_i)$ may be the mean estimate of Y_i based on $\boldsymbol{x}_i$. We give some remarks.

Remarks 3.1.

- Optimization problem (3.1) can be solved using the pool adjacent violators (PAV) algorithm of Ayer et al. (1955), Miles (1959) and Kruskal (1964). The resulting solution fulfills the (empirical) auto-calibration property, and it can also be written as the following min-max formula, for $1 \le i \le n$,

$$\widehat{\mu}_i \;=\; \min_{j=i,\ldots,n} \max_{k=1,\ldots,j} \frac{1}{\sum_{l=k}^j w_l} \sum_{l=k}^j w_l Y_l;$$

  see also Wüthrich–Ziegel (2023).

- If there are ties in the ranks $(\pi(\boldsymbol{x}_i))_{i=1}^n$, we adjust the corresponding observations. E.g., if $\pi(\boldsymbol{x}_i) = \pi(\boldsymbol{x}_j)$ for some $i \neq j$, we replace Y_i and Y_j with their weighted average $(w_i Y_i +$

$w_j Y_j)/(w_i + w_j)$ and assign the adjusted case weight $w_i + w_j$ to this new observation. There are other options to deal with ties; we refer to Leeuw et al. (2009).

- The solution to (3.1) only defines the regression values $\widehat{\mu}_i$ in the observations/ranks $\pi(\boldsymbol{x}_i)$. A common way to interpolate between these regression values is to use a step function, see, e.g., Figure 4.1, below. Alternatively, we could connect subsequent estimates $\widehat{\mu}_i$ linearly by straight lines.

- The objective function in (3.1) is given by the (weighted) square loss. However, every Bregman divergence will give the same result because this restricted optimization problem is invariant under different choices of strictly consistent loss functions for mean estimation; see Theorem 1.10 in Barlow et al. (1972). These strictly consistent loss functions are precisely the Bregman divergences; see Savage (1971) and Gneiting (2011). Moreover, every deviance loss function is a Bregman divergence, thus, the solution of (3.1) will not change if we replace the (weighted) square loss by a general (weighted) deviance loss.

- Since the estimates $(\widehat{\mu}_i)_{i=1}^n$ are found under the assumption of monotonicity with respect to $(\pi(\boldsymbol{x}_i))_{i=1}^n$, the choice of the ranks of $(\pi(\boldsymbol{x}_i))_{i=1}^n$ instead of the true values of $(\pi(\boldsymbol{x}_i))_{i=1}^n$ does not lead to any loss of information, i.e., any strictly increasing transformation of $(\pi(\boldsymbol{x}_i))_{i=1}^n$ will not change the result of isotonic regression.

4 Example 1: Validation of Tweedie's power variance function

To start with, we propose a diagnostic tool to assess the variance function of claim size responses under given mean values from the implemented regression model. Often, in actuarial science, one works within the framework of Tweedie's distributions (including the Poisson, the Gamma and the Inverse Gaussian distributions). Tweedie's distributions are special cases of the EDF. Tweedie's distributions have the property that the variance of the corresponding response variable Y_i with feature $\boldsymbol{x}_i$ is given by

$$\mathrm{Var}[Y_i|\boldsymbol{x}_i] = \phi\, V(\mu(\boldsymbol{x}_i)) = \phi\, \mu^p(\boldsymbol{x}_i), \tag{4.1}$$

for a power variance parameter $p \in (-\infty, 0] \cup [1, \infty)$ and a (constant) dispersion parameter $\phi > 0$. Hence, Tweedie's distributions assume a power variance function with power variance parameter p; for Tweedie's distributions we refer to Tweedie (1984) and Jørgensen (1987). Typically, one works with $p \geq 1$. Since we fit a Gamma GLM (which has a quadratic variance function) in the first step of our analysis, we should verify whether $p = 2$ is the right choice for our data set.

Let us fix a certain power variance parameter $p \geq 1$, say $p = 2$ for the fitted Gamma GLM. We can validate the specific choice of p by restructuring (4.1). Namely, we have dispersion coefficients for each instant $i = 1, \ldots, n$

$$\phi_i := \frac{\mathrm{Var}[Y_i|\boldsymbol{x}_i]}{\mu^p(\boldsymbol{x}_i)},$$

and this motivates to consider Pearson's (empirical) residuals

$$\widehat{\phi}_i = \frac{\left(Y_i - \widehat{\mu}(\boldsymbol{x}_i)\right)^2/(1-h_i)}{\widehat{\mu}^p(\boldsymbol{x}_i)}, \tag{4.2}$$

where $(\widehat{\mu}(\boldsymbol{x}_i))_{i=1}^n$ denote the mean values estimated with a Tweedie's GLM with variance function $V(\mu) = \mu^p$, and $(h_i)_{i=1}^n$ denote the hat values from the estimated Tweedie's GLM. The hat values provide a bias corrected version of the empirical variance and they allow us to correct the residuals derived from the fitted model; see Smyth–Verbyla (1999). We can now fit an isotonic regression to the sample $\big(\widehat{\phi}_i, \widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$, where we either assume a monotonically increasing or a monotonically decreasing property of Pearson's residuals $\widehat{\phi}_i$ in $\widehat{\mu}(\boldsymbol{x}_i)$. That is, we consider the estimated means $\widehat{\mu}(\boldsymbol{x}_i)$ as ranks $\pi(\boldsymbol{x}_i)$ in the isotonic regression (3.1), and Pearson's residuals $\widehat{\phi}_i$ play the role of the corresponding responses. Remark that there is no double use of data at this second stage where we fit an isotonic regression since the aim of the hat values is to correct for over-fitting issues from the first stage where we fit the mean values with GLM. If the isotonic regression gives a horizontal straight line in both cases (increasing or decreasing in the ranks), then ϕ in (4.1) is constant in i and the specific choice of the power variance parameter p is correct. Otherwise, the sign of the isotonic regressions will tell us whether the choice of p is too small or too big.

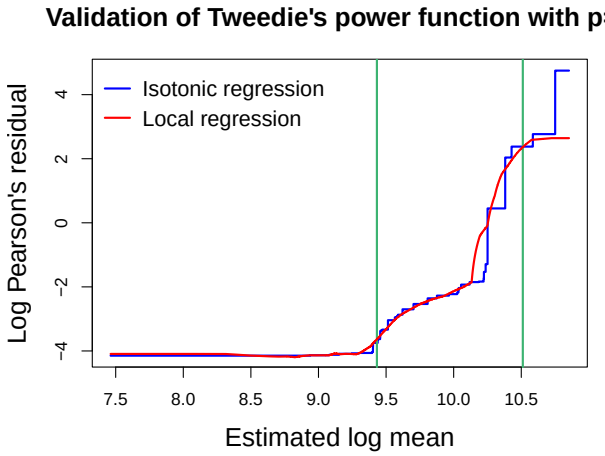

Figure 4.1: Logged Pearson's residuals estimated with an isotonic regression against logged estimated means from the Gamma GLM for all instances. The green lines show the 50% and 90% deciles of the true logged mean values used as the splicing points for the true variance function (2.1).

We apply the above diagnostic tool to the mean estimates from the Gamma GLM, the results are presented in Figure 4.1. We observe that Tweedie's quadratic variance function (with $p = 2$) should be rejected for our data set. From Figure 4.1 we conclude that the quadratic variance function is correct for observations with estimated means below roughly $12,000 \approx e^{9.4}$. Above this mean value, the variance function behaves like a power function with a power parameter of order larger than 2, and the value of the power parameter changes (increases) around the estimated mean of roughly $28,000 \approx e^{10.25}$. These conclusions agree with our assumptions on the true variance function in Figure 2.1. Thus, it is evident that a simple Tweedie's power variance function is not sufficient for our data set, and we should move outside the class of Tweedie's power variance functions and fit a different (and more flexible) variance function. We conclude that performing an isotonic regression on Pearson's residuals (4.2) using the estimated means $\widehat{\mu}(\boldsymbol{x}_i)$ as the ranks equips us in this problem with a simple diagnostic tool that provides us with the correct insights concerning the true variance

behavior under the monotonicity condition.

The analysis of the results as presented in Figure 4.1 is well known in statistical modeling, and Pearson's residuals are usually investigated by fitting a local polynomial regression against the predicted values. We also exploited a local polynomial regression and the conclusions were the same. However, when fitting an isotonic regression we do not have to choose the kernel, the bandwidth of the kernel and the degree of the polynomial. Hence, the curve fitted with isotonic regression is in some sense more objective than curve fitting with local polynomial regression, under the assumption of monotonicity, of course. This argument is in favor of isotonic regression, and it is the main motivation for using isotonic regression instead of other non-parametric smoothers. We point out that, e.g., local polynomial regression could be used to validate the monotonicity assumption for the variance function and an increasing estimate would support the application of an isotonic regression.

5 Example 2: Miscalibration test and consistency bands for reliability diagrams

The importance of the auto-calibration property for insurance pricing has been outlined in the actuarial literature by Denuit et al. (2021), Lindholm et al. (2023), Wüthrich (2023), Wüthrich–Ziegel (2023) and Denuit–Trufin (2023). Based on the literature in statistics on forecast evaluation, see Pohle (2020), Dimitradis et al. (2020), Gneiting–Resin (2021) and Dimitradis et al. (2022), we develop two techniques which allow us to assess if the mean estimates from a fitted model are auto-calibrated. These techniques use quantile residuals which themselves can be used for model diagnostic purposes.

In the sequel we use the notation $(Y, \boldsymbol{X})$ to denote the random selection of a feature $\boldsymbol{X}$ from an insurance portfolio, and Y is the corresponding claim. The observations $(Y_i, \boldsymbol{x}_i)_{i=1}^n$ are then assumed to be an i.i.d. sample of size n with the same distribution as $(Y, \boldsymbol{X})$.

A regression function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$ is auto-calibrated for $(Y, \boldsymbol{X})$ if

$$\mu(\boldsymbol{X}) = \mathbb{E}\left[Y \middle| \mu(\boldsymbol{X})\right], \qquad \mathbb{P}\text{-a.s.} \tag{5.1}$$

In actuarial pricing, auto-calibration means that every price cohort $\mu(\boldsymbol{X})$ is in average self-financing for its claim Y. This implies that there is no systematic cross-financing within the price system $\mu(\cdot)$. Auto-calibration is a quality criterion that every insurance price system $\mu(\cdot)$ should fulfill. We observe that the best-estimate price

$$\mu^*(\boldsymbol{X}) := \mathbb{E}\left[Y \middle| \boldsymbol{X}\right]$$

is auto-calibrated. In fact, this price is called best-estimate because it minimizes every Bregman divergence, see also Remarks 3.1 on strict consistency. In particular, this applies to the gamma deviance loss which we used for the claim sizes. Moreover, the best-estimate price dominates in convex order any other auto-calibrated $\sigma(\boldsymbol{X})$-measurable price, and it also attains the maximal Gini index among all auto-calibrated $\sigma(\boldsymbol{X})$-measurable prices. This gives us model selection criteria that all provide the best-estimate price as the optimal regression function based on information $\sigma(\boldsymbol{X})$; see Wüthrich (2023) and Denuit–Trufin (2023). However, since this best-estimate $\mu^*(\cdot)$ can be any

$\sigma(\boldsymbol{X})$-measurable function, it is typically intractable, and we aim at approximating it within a sufficiently tractable model class of regression functions. The goal is to build a mean estimator $\widehat{\mu}$, which accurately approximates the best-estimate price μ^*, and which fulfills the auto-calibration property (5.1).

Assume that $L(m, Y)$ is a strictly consistent loss function for the mean functional, e.g., the unscaled or scaled gamma deviance loss. Then, we can (empirically) assess the prediction accuracy of a mean estimator $\widehat{\mu}$ by the score

$$S(\boldsymbol{Y}, \widehat{\mu}) := \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \widehat{\mu}(\boldsymbol{x}_i)), \tag{5.2}$$

which is an empirical version of the expected loss

$$\mathbb{E}\left[L(Y, \widehat{\mu}(\boldsymbol{X}))\right]. \tag{5.3}$$

Strict consistency of L for the mean functional implies that this expected loss (5.3) is (uniquely) minimized by the true regression function μ^*.

When we fit a GLM, we restrict the class of regression functions to linear functions in the features $\boldsymbol{x}$ after applying the selected link g, see (1.2). The mean estimator $\widehat{\mu}$ is then found by minimizing the empirical score (5.2) in $\boldsymbol{\theta}$, see (1.4). However, in general, the estimated GLM mean values $\big(\widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$ are not auto-calibrated. As emphasized in Dimitradis et al. (2020), Gneiting–Resin (2021), Dimitradis et al. (2022) and Wüthrich–Ziegel (2023), an isotonic regression can be fitted to $(Y_i, \widehat{\mu}(\boldsymbol{x}_i))_{i=1}^n$, providing an empirically auto-calibrated version $\widehat{\mu}_{\text{rc}}$ of $\widehat{\mu}$. This isotonically recalibrated version $\widehat{\mu}_{\text{rc}}$ then motivates to consider Murphy's score decomposition

$$S(\boldsymbol{Y}, \widehat{\mu}) = \text{UNC}_S(\boldsymbol{Y}, \widehat{\mu}) - \text{DSC}_S(\boldsymbol{Y}, \widehat{\mu}) + \text{MCB}_S(\boldsymbol{Y}, \widehat{\mu}), \tag{5.4}$$

with the uncertainty, discrimination and miscalibration statistics, respectively,

$$\begin{aligned}
\text{UNC}_S(\boldsymbol{Y}, \widehat{\mu}) &= S(\boldsymbol{Y}, \bar{\mu}), \\
\text{DSC}_S(\boldsymbol{Y}, \widehat{\mu}) &= S(\boldsymbol{Y}, \bar{\mu}) - S(\boldsymbol{Y}, \widehat{\mu}_{\text{rc}}) \;\geq\; 0, \\
\text{MCB}_S(\boldsymbol{Y}, \widehat{\mu}) &= S(\boldsymbol{Y}, \widehat{\mu}) - S(\boldsymbol{Y}, \widehat{\mu}_{\text{rc}}) \;\geq\; 0,
\end{aligned}$$

with $\bar{\mu} = \sum_{i=1}^n Y_i/n$ being the empirical mean not considering any features; we refer to Murphy (1973), Pohle (2020), Dimitradis et al. (2020) and Gneiting–Resin (2021). A small value of the miscalibration statistics $\text{MCB}_S(\boldsymbol{Y}, \widehat{\mu})$ supports the hypothesis that the estimated mean values $\widehat{\mu}$ are (empirically) auto-calibrated for $(Y_i, \boldsymbol{x}_i)_{i=1}^n$. This motivates to test the null hypothesis H_0 : $\text{MCB}_S(\boldsymbol{Y}, \widehat{\mu}) = 0$ against the alternative H_1 : $\text{MCB}_S(\boldsymbol{Y}, \widehat{\mu}) > 0$. We formalize this. Let $\widehat{\mu}_i := \widehat{\mu}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, denote the mean estimates to be validated as best estimates, hence auto-calibrated, prices. We set the null hypothesis that under H_0 the observations $(Y_i|\boldsymbol{x}_i)_{i=1}^n$ come from conditional distributions with auto-calibrated best estimates $\widehat{\mu}_i$ against the alternative that they fail to follow the auto-calibrated best estimates $\widehat{\mu}_i$. To perform a statistical test, we need to consider a test statistics that is tractable under the null hypothesis and that can be evaluated on a given signficance level.

A key diagnostic tool in statistics for checking auto-calibration of a regression function is a reliability diagram (also called a lift plot in the actuarial literature). Reliability diagrams plot the observed empirical means against the estimated mean values in pre-defined bins; this was introduced by Hosmer–Lemeshow (1980) in a binary context. It is known that reliability diagrams are sensitive to the choice of the bins; see, e.g., Henzi et al. (2023). Dimitradis et al. (2020) introduce a new approach to reliability diagrams where one compares the estimated mean values $\big(\widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$ against their recalibrated values $\big(\widehat{\mu}_{\mathrm{rc}}(\boldsymbol{x}_i)\big)_{i=1}^n$, where, again, $\widehat{\mu}_{\mathrm{rc}}$ is constructed from $\widehat{\mu}$ with an isotonic regression. Such reliability diagrams are called CORP, since, following Dimitradis et al. (2020):

- C: the reliability diagrams and associated numerical measures of miscalibration are consistent in the classical statistical sense of convergence to population characteristics (consistency);
- O: the reliability diagrams are optimally binned (optimality);
- R: the reliability diagrams do not require any tuning parameters nor implementation decision (reproducibility); and
- P: the reliability diagrams are implemented via the PAV algorithm (pool adjacent violators algorithm).

Later, Gneiting–Resin (2021) introduced a more general notion of conditional T-calibration (auto-calibration with respect to T) and T-reliability diagrams, where T in our case denotes the mean functional; this explains the terminology T-reliability diagram. If $\big(\widehat{\mu}_{\mathrm{rc}}(\boldsymbol{x}_i)\big)_{i=1}^n$ match $\big(\widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$, i.e., the points $\big(\widehat{\mu}(\boldsymbol{x}_i), \widehat{\mu}_{\mathrm{rc}}(\boldsymbol{x}_i)\big)_{i=1}^n$ of a T-reliability diagram lie close to its diagonal, the mean estimator $\widehat{\mu}$ is auto-calibrated for $(Y_i, \boldsymbol{x}_i)_{i=1}^n$. We would like to formally evaluate the significance of the observed deviations from the diagonal. For this we construct a consistency band for a T-reliability diagram. The consistency band is constructed under the null hypothesis H_0 that the observations $(Y_i|\boldsymbol{x}_i)_{i=1}^n$ come from conditional distributions with the auto-calibrated best estimates $\widehat{\mu}_i := \widehat{\mu}(\boldsymbol{x}_i)$ as means.

Gneiting–Resin (2021) propose to explore the problem with bootstrap, see Appendix B in Gneiting–Resin (2021). We develop this proposal of Gneiting–Resin (2021), and it turns out that bootstrapping of residuals from GLMs with non-Gaussian responses poses some challenges. In our case, and in most practical applications in insurance, we have to handle proper support, unknown heteroskedasticity and unknown distribution of the response to correctly implement the bootstrap. We explain this in detail.

Let $\widehat{\mu}$ denote the mean estimator we want to validate. We recommend the following semi-parametric bootstrap approach.

Semi-Parametric Bootstrap Approach

- **Step 1:** To account for heteroskedasticity, we estimate the variances of $Y_i|\boldsymbol{x}_i$ for all instances $1 \le i \le n$ under the estimated means $\big(\widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$ by an isotonic regression assuming that these variances are monotonically increasing in the means. We fit an isotonic regression to $\big(\widehat{v}_i, \widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$, where

$$\widehat{v}_i = \frac{(Y_i - \widehat{\mu}(\boldsymbol{x}_i))^2}{1 - h_i}, \tag{5.5}$$

and $(h_i)_{i=1}^n$ are the hat values from the fitted model. The isotonic regression provides us with variance estimates $\widehat{V}(\boldsymbol{x}_i)$ for all instances $1 \leq i \leq n$ under the estimated means. Hence, we have estimated the first two moments of the observations.

- **Step 2:** We calculate the quantile residuals defined by $\widehat{\varepsilon}_i = F\big(Y_i, \widehat{\mu}(\boldsymbol{x}_i), \widehat{V}(\boldsymbol{x}_i)\big)$ for all instances $1 \leq i \leq n$, where we use a two-parametric cumulative distribution function F with the first two (estimated) moments $\big(\widehat{\mu}(\boldsymbol{x}_i), \widehat{V}(\boldsymbol{x}_i)\big)_{i=1}^n$.

- **Step 3:** Under an i.i.d. assumption for the quantile residuals, we bootstrap $(\varepsilon_i^*)_{i=1}^n$ from $(\widehat{\varepsilon}_i)_{i=1}^n$ to receive the bootstrap observations:

$$Y_i^* = F^{-1}\big(\varepsilon_i^*, \widehat{\mu}(\boldsymbol{x}_i), \widehat{V}(\boldsymbol{x}_i)\big).$$

  If we map $\widehat{\varepsilon}_i \mapsto \Phi^{-1}(\widehat{\varepsilon}_i)$ with the inverse of standard normal cumulative distribution function, we can center and scale the quantile residuals to get better performance of residual re-sampling in small samples.

- **Step 4:** Based on the new sample $\big(Y_i^*, \widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$, we recalibrate the mean estimator $\widehat{\mu}$ with an isotonic regression, plot the T-reliability diagram and calculate the miscalibration statistics. From repreating this semi-parametric bootstrap, we can construct empirical consistency bands for the T-reliability diagram and find critical values for the miscalibration test, under the null hypothesis that $\widehat{\mu}_i := \widehat{\mu}(\boldsymbol{x}_i)$ are the best-estimate prices for $(Y_i|\boldsymbol{x}_i)_{i=1}^n$ (hence, they are auto-calibrated).

Clearly, a good estimate of the variance of the response, $\mathrm{Var}[Y_i|\boldsymbol{x}_i]$, is crucial if we deal with heteroskedastic observations. In practice, the true form of heterskedasticity may be different from the one assumed in the fitted GLM (Poisson, Gamma, Inverse Gaussian, Tweedie's distributions) and we need to estimate this variance as good as possible to obtain a reasonable bootstrap simulation; we point out that the discussion of this step is missing in Gneiting–Resin (2021). In Step 1, we use an isotonic regression for this variance estimation. The next difficulty is that a naive resampling scheme of the responses can be problematic given that the bootstrapped claim sizes Y_i^* must be positive, i.e., should have a restricted support. The classical approach is to use empirical Pearson's residuals calculated based on $\big(\widehat{\mu}(\boldsymbol{x}_i), \widehat{V}(\boldsymbol{x}_i)\big)_{i=1}^n$. However, by construction, this bootstrap approach does not respect support constraints. If we manually correct negative bootstrapped claim sizes by $\max\{Y_i^*, 0\}$, we face a bias. This bias becomes particularly visible for observations with large variances which are (unconditionally) skewed and heavy-tailed, see Figure 2.1, and it makes the Pearson's residuals skewed and heavy-tailed as well. For this reason, we switch to bootstrapping with quantile residuals. Quantile residuals were introduced by Dunn–Smyth (1996) in regression modeling and proposed in Chapter 4.4 of Politis (2015) for a "model-free" bootstrap of regression models; quantile residuals are also called PIT (Probability Integral Transform) residuals, see Gneiting–Resin (2021). The idea of the "model-free" bootstrap from Politis (2015) is that there exists a transformation which maps non-i.i.d. data to i.i.d. data. In Step 2, in order to define quantile residuals, we have to specify a distribution function F, in addition to the mean and variance estimates. This distribution function should be chosen (estimated) based on the observations, so that

the resulting variables $(\widehat{\varepsilon}_i)_{i=1}^n$ are (approximately) i.i.d., but not necessary uniformly distributed. Ideally, if F is the true distribution of the response, then $(\widehat{\varepsilon}_i)_{i=1}^n$ are independent and uniformly distributed. Note that, despite we specify F, we still use a non-parametric bootstrap in Step 3 since we sample from the empirical quantile residuals $(\widehat{\varepsilon}_i)_{i=1}^n$ constructed under the chosen distribution F. Combining Steps 2 and 3, we call this approach a semi-parametric bootstrap. In case the true F is known or we are confident with the estimated F, we can use a parametric bootstrap based on this knowledge (we use a parametric bootstrap in Sections 7 and 8 where we directly use the assumption that the claim sizes come from lognormal distributions). We prefer a semi-parametric bootstrap with quantile residuals since it is slightly more robust to misspecification of the true distribution of the response than a parametric bootstrap, see Lemma 1-2 in the Appendix of Warton et al. (2017). We remark that the quantile residuals will play two roles in the analysis below, namely, they are used for bootstrap, and they are also used to validate the estimates of the mean and the variance.

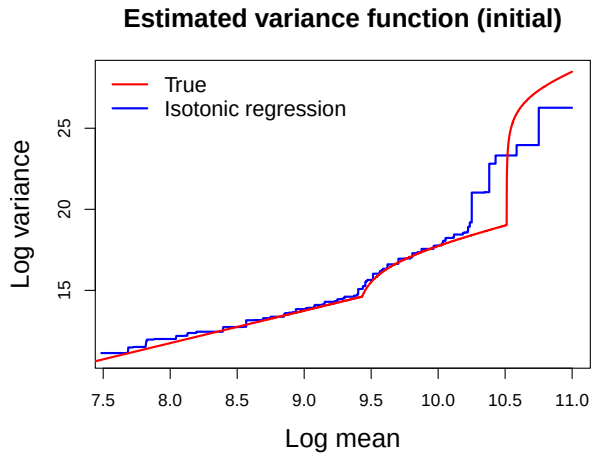

Figure 5.1: Logged variance function estimated with an isotonic regression based on means estimated with the Gamma GLM for a range of logged mean values.

We use the above semi-parametric bootstrap algorithm to investigate the quality of the mean estimates $\big(\widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$ from our first model – the Gamma GLM. The result of Step 1, i.e., the estimation of the variance function based on (5.5) using an isotonic regression is presented in Figure 5.1, and our isotonic regression estimate is compared to the true variance function (2.1). This plot is called 'initial' since we use the initial mean estimates from the Gamma GLM for variance estimation, these estimates will be refined in the next section. At this point, without referring to the true variance function, we can see that our estimate of the variance function is far from a quadratic function, hence a Gamma GLM with a quadratic variance function should not be used for this data set. Figure 5.1 verifies that our initial variance estimate using the isotonic regression already fits the true variance already quite well.

Based on Step 1, we can use the semi-parametric bootstrap steps from the above algorithm to validate auto-calibration of the mean estimator (we point out that for a resonable bootstrap we use the isotonic regression variance estimate from Figure 5.1, and not the quadratic variance of the fitted Gamma GLM). In order to derive the quantile residuals, we choose a Gamma distribution (Step 2) – this choice is the most natural one given we do not want to estimate the full distribution

of the claim sizes. The quantile residuals from the Gamma GLM under the Gamma distribution transformation are presented in Figure 5.2. In this figure we shows the following plots:

- (top-left) The quantile residuals against the estimated logged mean values. In a perfect model, the quantile residuals should be uniformly distributed in any subset of the estimated mean values.
- (top-right) Correlations of the quantile residuals distant at lags $\ell = 1, \ldots, 10$, together with 95% confidence intervals for the null hypothesis that the correlations are zero. In a perfect model, the quantile residuals should be uncorrelated and independent. We randomly shuffle the quantile residuals and we present the results for one sample of shuffled quantile residuals. The results are similar in all random samples.
- (bottom-left) PIT reliability diagram for all quantile residuals, where we compare the empirical cumulative distribution function of the quantile residuals against the uniform distribution (the uniform distribution represents the perfect estimation of the moments and the distribution); see Appendix B in Gneiting–Resin (2021) for a discussion of PIT reliability diagrams.
- (bottom-right) PIT reliability diagrams for quantile residuals grouped into four blocks based on quantiles of the estimated mean values. In a perfect model, the quantile residuals should be identically and uniformly distributed in the four blocks.

From Figure 5.2 we can conclude that the quantile residuals are independent but they are not identically (and uniformly) distributed. Based on Figure 5.2 we can deduce that the Gamma GLM fits poorly to the observations with small estimated means and large estimated means (below the 25th percentile and above the 75th percentile of the estimated means), since for these observations the assumption of common (uniform) distribution of the quantile residuals fails (bottom-right) – a conclusion which agrees with Figure 2.2 based on the ground truth.

We proceed with our (imperfect) quantile residuals and perform Steps 3 and 4 of the above algorithm. We consider 1,000 bootstrap samples. From these bootstrap samples we compute the bootstrap distribution of the miscalibration statistics and the bootstrap versions of the T-reliability diagram, together with the quantiles needed for a 5% significance level of the miscalibration test and a 95% coverage of the consistency band. As strictly consistent scoring function in (5.2)-(5.3) to measure the miscalibration term $\mathrm{MCB}_S(\boldsymbol{Y}, \widehat{\mu})$ of the mean estimator, we use a weighted gamma deviance loss function given by

$$L(m_i, Y_i) = 2w_i\Big(-\log(Y_i/m_i) + (Y_i - m_i)/m_i \Big), \tag{5.6}$$

where we set $w_i = \widehat{V}(\boldsymbol{x}_i)/\big(\widehat{\mu}(\boldsymbol{x}_i)\big)^2$ based on our initial estimates; this is a dispersion estimate in the Gamma GLM using the isotonic regression estimate $\widehat{V}(\boldsymbol{x}_i)$ of Step 1 of the algorithm. These weights are kept fixed in the bootstrap simulations, i.e., we only recalibrate $\big(\widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$ based on the bootstrap sample $\big(Y_i^*, \widehat{\mu}(\boldsymbol{x}_i)\big)_{i=1}^n$. In Section 4 we have concluded that our observations do not have a constant dispersion coefficient, that is why we consider a weighted gamma deviance loss function in (5.6).

We plot the T-reliability diagram, and we first study it without interpreting the bootstrap results, see Figure 5.3 (left). We plot the T-reliability diagram rotated by 45°: the x-axis gives

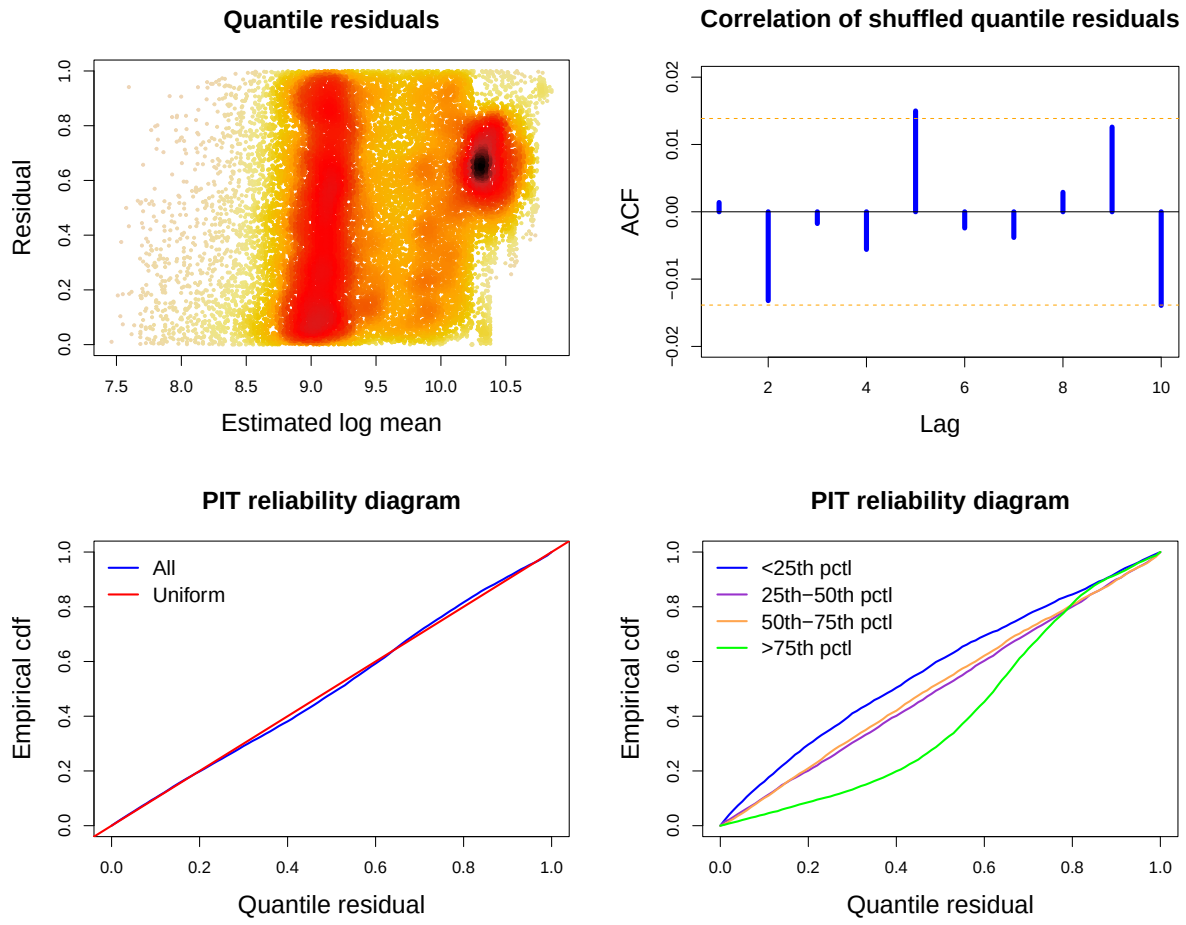

Figure 5.2: Quantile residuals from the Gamma GLM; (top-left) the darker the color, the larger the density of the observations with particular means.

the logged estimated means $\log(\widehat{\mu}(\boldsymbol{x}_i))$, and the y-axis presents the differences between the auto-calibrated mean estimates and their original versions on the logarithmic scale: $\log(\widehat{\mu}_{\text{rc}}(\boldsymbol{x}_i)) - \log(\widehat{\mu}(\boldsymbol{x}_i))$. We expect that the mean estimates from the Gamma GLM are not auto-calibrated, which is verified by this figure (we deviate from the horizontal zero line). The Gamma GLM seems to over-estimate the true means of policyholders with small expected claims, and under-estimate the true means of policyholders with large expected claims. This statement agrees with the conclusions derived above based on the quantile residuals. In practice, such a misestimation leads to a "wrong" premium (failure of auto-calibration and systematic cross-subsidy), which in turn may imply adverse selection and other frictions of the insurance market. We now add the bootstrap results to our analysis in Figure 5.3. The 95% consistency band for the T-reliability diagram is added in orange color on the left hand side of Figure 5.3, and the 5% critical value of the miscalibration test is shown on the right hand side of Figure 5.3. This confirms our guess of miscalibration of the mean estimates from the Gamma GLM. That is, we clearly reject the null hypothesis H_0 of having auto-calibration in the Gamma GLM. We should keep in mind that the quantile residuals used for the bootstrap do not have a perfect structure, hence, these results from the bootstrap should be interpreted with some caution. We could also bootstrap from a subset of the quantile residuals

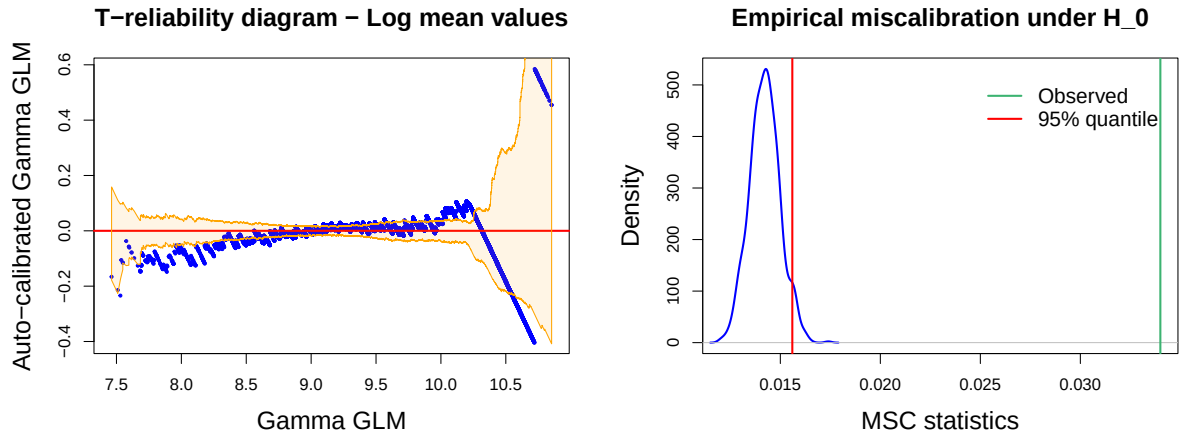

Figure 5.3: (left) T-reliability diagram together with the 95% consistency band for mean estimates from the Gamma GLM, and (right) distribution of the miscalibration statistics, together with the observed value and the critical value at signficance level of 5% – the variance function is estimated with an isotonic regression. T-reliability diagram presents logged mean estimates.

which contains the quantile residuals for the observations with the estimated mean values between the 25th and the 75th percentile of the estimated means, since these quantile residuals are i.i.d. and uniformly distributed, see Figure 5.2. We applied this sampling scheme and we got very similar results.

Summing up, we proposed diagnostic tools for regression modeling of means, in particular, based on the variance function estimated with an isotonic regression. They confirm that the initial Gamma GLM can be improved in many ways, and we see in the sequel that these tools will also be useful to analyze improved versions of this Gamma GLM.

6 Example 3: Quasi-likelihood estimation of mean values with non-parametric variance estimation

Motivated by Gourieroux et al. (1984) and Chiou–Müller (1999), we now iterate the GLM estimation procedure from the previous section to receive better mean estimates. We apply the following algorithm, which applies quasi-likelihood estimation of the mean values together with a non-parametric estimation of the variance function. Compared to the non-parametric quasi-likelihood method of Chiou–Müller (1999), we propose to use an isotonic regression instead of a local polynomial regression as a non-parametric variance estimator. We test the performance of the isotonic regression and study its advantages.

Quasi-likelihood with Non-Parametric Variance Estimation

- **Step 1:** We estimate the expected value of the response with a classical GLM with link function g and variance function V^{GLM} implied by the selected GLM. This provides an initial

parameter estimate $\widehat{\boldsymbol{\theta}}^0$ and the estimated mean values $\big(\widehat{\mu}^0(\boldsymbol{x}_i)\big)_{i=1}^n$, with $\widehat{\mu}^0(\boldsymbol{x}_i) = g^{-1}\big(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}^0\big)$ for all instances $i = 1, \ldots, n$.

- **Step 2:** We estimate the variance of the response using the crude estimator

$$\widehat{v}_i^0 = \frac{\big(Y_i - \widehat{\mu}^0(\boldsymbol{x}_i)\big)^2}{1 - h_i^0},$$

where $(h_i^0)_{i=1}^n$ denote the hat values of the GLM fitted at the initial step.

- **Step 3:** We estimate the variance function of the response based on the observations $\big(\widehat{v}_i^0, \widehat{\mu}^0(\boldsymbol{x}_i)\big)_{i=1}^n$. We use an isotonic regression assuming that the true variance function $V(\mu)$ is monotonically increasing in μ. This step provides us with an estimate of the variance function $\mu \mapsto \widehat{V}^0(\mu)$ and the variances $\widehat{V}^0(\widehat{\mu}^0(\boldsymbol{x}_i))$ for all instances $i = 1, \ldots, n$. This step is identical to Step 1 of the algorithm presented in Section 5.

- **Step 4:** We iterate for $k = 1$ to K:

  (i) Estimate the expected value of the response with a GLM with link function g using the quasi-likelihood method. In the deviance loss function (1.4) we use the estimated variance function $\mu \mapsto \widehat{V}^{k-1}(\mu)$ as the variance function of the EDF family.

  (ii) The regression function for the above quasi-GLM is estimated with the iterative reweighed least squares (IRLS) algorithm. In inner steps $k_\ell = 1, \ldots, M$ of step k, the auxiliary response is given by

$$Y_i^{k_\ell} = g\big(\mu^{k_\ell - 1}(\boldsymbol{x}_i)\big) + g'\big(\mu^{k_\ell - 1}(\boldsymbol{x}_i)\big)\big(Y_i - \mu^{k_\ell - 1}(\boldsymbol{x}_i)\big),$$

  the expected value of the auxiliary response and the regression function are given by

$$g\big(\mu^{k_\ell - 1}(\boldsymbol{x}_i)\big) = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}^{k_\ell},$$

  with the parameter estimate $\widehat{\boldsymbol{\theta}}^{k_\ell}$ to be found in step k_ℓ, and the weights are given by

$$w_i^{k_\ell} = \Big(\widehat{V}^{k-1}(\widehat{\mu}^{k_\ell - 1}(\boldsymbol{x}_i))\Big(g'(\widehat{\mu}^{k_\ell - 1}(\boldsymbol{x}_i))\Big)^2\Big)^{-1}. \tag{6.1}$$

  We initialize with $\widehat{\mu}^{k_0}(\boldsymbol{x}_i) = \widehat{\mu}^{k-1}(\boldsymbol{x}_i)$ for all instances $i = 1, \ldots, n$.

  (iii) We get new estimates of $\widehat{\boldsymbol{\theta}}^k = \widehat{\boldsymbol{\theta}}^{k_M}$ and $\big(\widehat{\mu}^k(\boldsymbol{x}_i)\big)_{i=1}^n$ with $\widehat{\mu}^k(\boldsymbol{x}_i) = g^{-1}\big(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}^k\big)$ for all instances $i = 1, \ldots, n$.

  (iv) We re-estimate the variance of the response using the crude estimator:

$$\widehat{v}_i^k = \frac{\big(Y_i - \widehat{\mu}^k(\boldsymbol{x}_i)\big)^2}{1 - h_i^k}. \tag{6.2}$$

  (v) We re-estimate the variance function of the response based on the observations $\big(\widehat{v}_i^k, \widehat{\mu}^k(\boldsymbol{x}_i)\big)_{i=1}^n$ using an isotonic regression and assuming monotonicity in the mean

estimates. We get new estimates of $\mu \mapsto \widehat{V}^k(\mu)$ and $\widehat{V}^k(\widehat{\mu}^k(\boldsymbol{x}_i))$ for all instances $i = 1, \ldots, n$.

We use $K = 25$ and $M = 10$ – these choices give convergence of the algorithm, but slightly lower values lead to similar estimates. First, we present the final estimate of the variance function of the response, see Figure 6.1. As already pointed out, we estimate the variance function non-parametrically with an isotonic regression as outlined in the above algorithm. In addition, we complement this estimate with an estimate received by a local polynomial regression; see Loader (1999) for local regression. We also compare our final results to the true variance function. We fit a local polynomial regression with the `locfit` function in `R` and we decided to use the polynomial's degree equal to zero, rectangular kernel and smoothing parameter (nearest neighbor fraction) $\alpha = 0.05$; we only consider the degree of the local polynomial equal to zero since local regressions with quadratic or cubic polynomials do not locally satisfy the positivity constraint, hence it was difficult to use them to model the variance function without manual intervention and bounding the estimate by zero. The fit of the local polynomial regression presented in Figure 6.1 is the best we could achieve in this example. We point out that the estimate with the local polynomial regression requires fine–tuning the degree of the polynomial, the type of the kernel and the bandwidth. At the same time, the isotonic regression estimates do not require any selection of hyperparameters, which is a great advantage of this non-parametric method. Remark that in our algorithm we iteratively fit variance functions, hence an estimation method without the need for optimally fine–tuning hyperparameters at each stage of the algorithm is a great advantage. We conclude that the isotonic regression is a very beneficial statistical tool. The isotonic estimate gives very accurate results in our case, slightly better than the local regression. Hence, we propose our non-parametric quasi-likelihood estimation with an isotonic regression as a variance estimator.

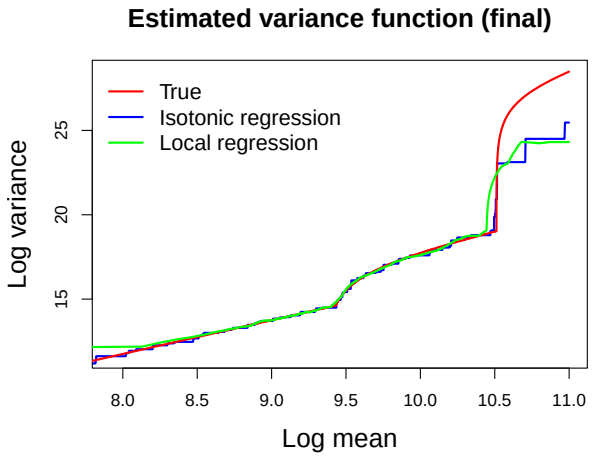

Figure 6.1: Logged variance function estimated with an isotonic regression and a local regression (kernel smoother) based on means estimated with the quasi-GLM for a range of logged mean values.

Next, we validate the quantile residuals from the quasi-GLM, see Figure 6.2. Compared to Figure 5.2, the quantile residuals have a much better structure now due to improved estimations of both the mean and the variance of the responses. However, we still observe a sign of misestimation

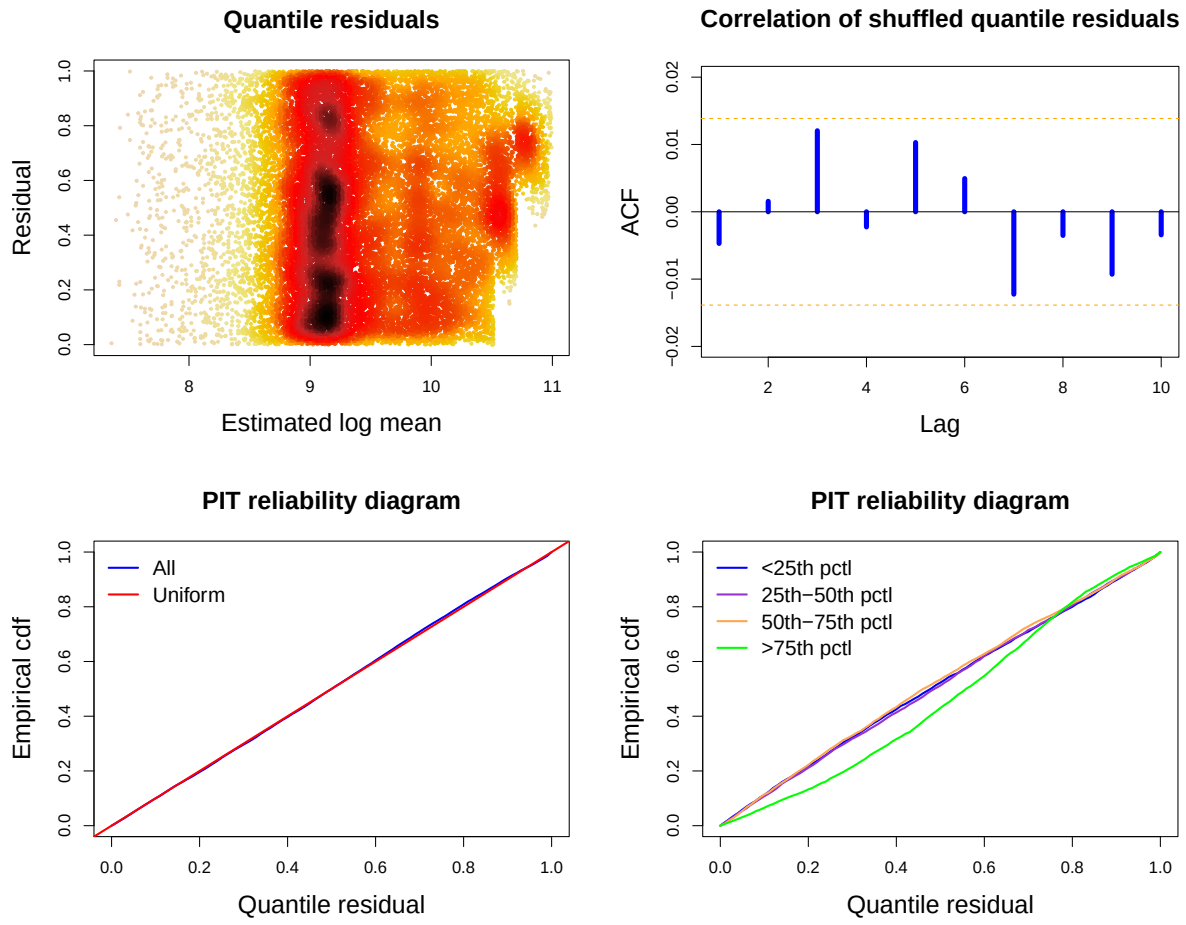

Figure 6.2: Quantile residuals from the quasi-GLM; (top-left) the darker the color, the larger the density of the observations with particular means.

for some observations with the largest estimated mean values (with log means above 10.6, which is the 75% percentile among the estimated mean values; see (bottom-right) of Figure 6.2). We remark that the quantile residuals would not have a better structure if we applied the PIT transformation with lognormal distributions, thus, the problem of non-identical distributions indicated in the quantile residuals is caused by a misestimation of the means and the variances of the responses in the upper tail. Below, we reveal that the misestimation of the means is minor, so this is not the issue here.

In Figure 6.3 (left), we present the T-reliability diagram for the mean estimates from the quasi-GLM, together with the 95% bootstrap consistency band. Figure 6.3 (right) shows the bootstrap distribution of the miscalibration statistics calculated with a weighted gamma deviance function, together with its 5% critical value and the observed miscalibration statistics. We implemented the bootstrap as in the previous section. The results show that the miscalibration observed for the Gamma GLM vanishes by this iteration for most estimated mean values, and the quasi-GLM does not have a bias (the mean estimates from the quasi-GLM are auto-calibrated, i.e., H_0 is not rejected). The only miscalibration of the mean values is potentially observed for the observations with the largest estimated means. From the analysis of the quantile residuals, we may not fully

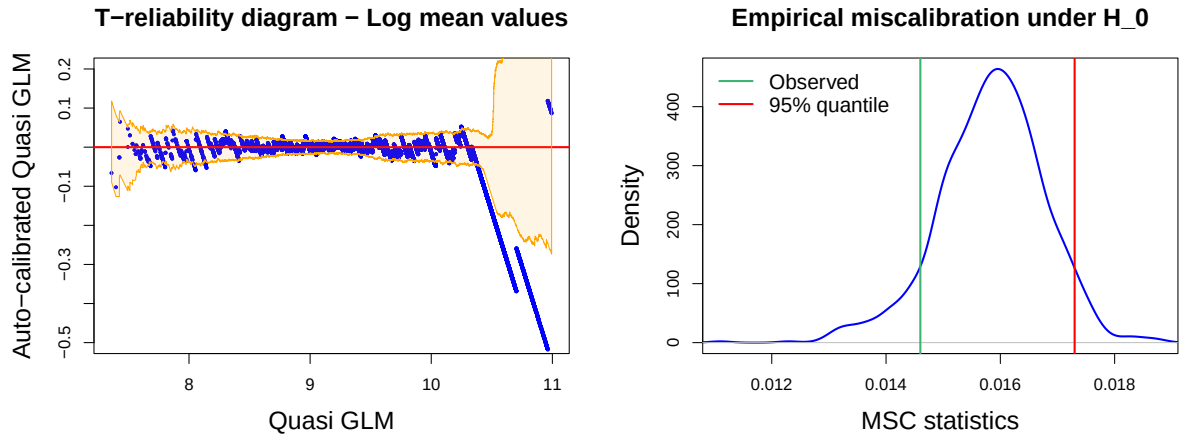

Figure 6.3: (left) T-reliability diagram together with the 95% consistency band for mean estimates from the quasi-GLM, and (right) distribution of the miscalibration statistics, together with the observed value and the critical value at signficance level 5% – the variance function is estimated with an isotonic regression. T-reliability diagram presents logged mean estimates.

trust the bootstrap results for the observations with estimated log means above 10.6, but we are confident about the bootstrap results for the observations with estimated log means below 10.6. To validate the bootstrap results with a different approach, we could also bootstrap from a subset of the quantile residuals which contains the quantile residuals for the observations with estimated log means below 10.6, as discussed in the previous section. The conclusions concerning miscalibration of the mean estimator are the same in this modified approach.

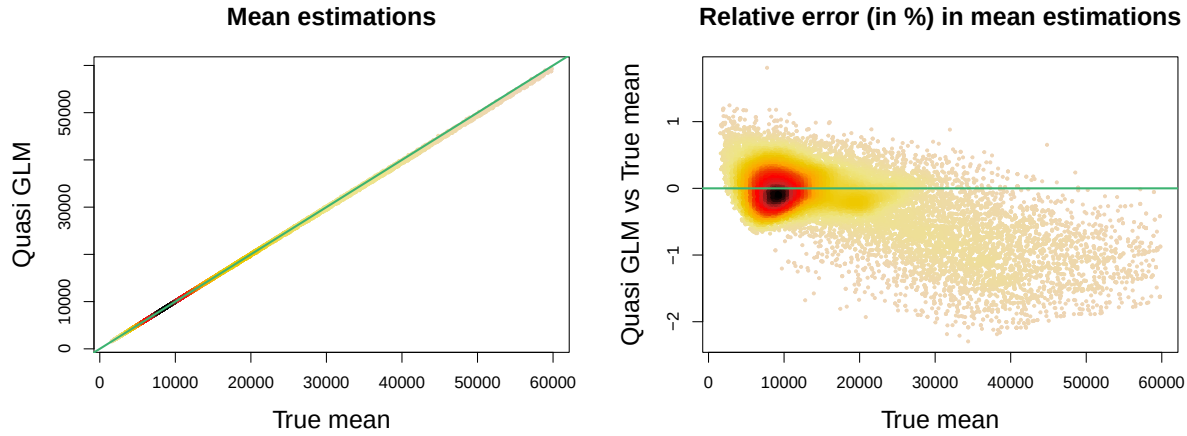

Figure 6.4: Estimated means from the quasi-GLM against true means for all instances; the darker the color, the larger the density of the observations with particular means; remark that the y-scale of the plot on the right hand side differs from the one in Figure 2.2.

Finally, the mean estimates from the quasi-GLM are compared to the true mean values in Figure 6.4. We observe that the mean estimates from the quasi-GLM match much closer the true mean values, compared to the mean estimates from the classical Gamma GLM, see Figure 2.2. This verifies that the estimation iteration with a properly estimated variance leads to a major

improvement in accuracy. In this quasi-GLM approach, we get estimation errors of maximal order $+1\%$ and -2%, respectively, this is substantially smaller than for the Gamma GLM in Figure 2.2. We conclude that on a finite sample our non-parametric quasi-likelihood with an (approximately) correct variance can significantly improve mean estimation.

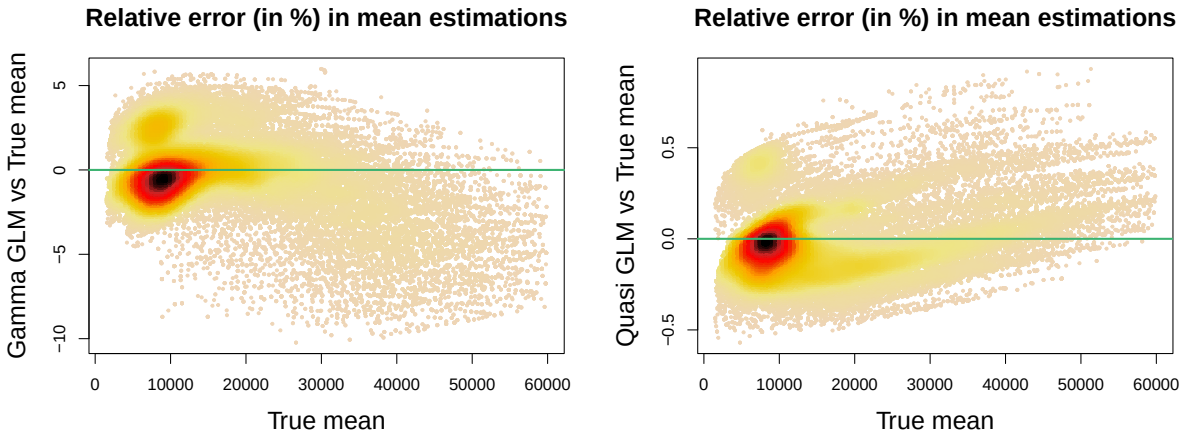

Figure 6.5: Estimated means from the Gamma GLM and the quasi-GLM against true means for all instances in the Data Set 2 with sample size $n = 100,000$; the darker the color, the larger the density of the observations with particular means; note that the scales of the y-axes differ in the two graphs.

We now switch for a moment to the synthetic Data Set 2. The theory of quasi-likelihood methods says that when the sample size increases, we may misspecify the true variance function, but asymptotically the mean estimates still converge to the true means. This is because the deviance loss functions are strictly consistent for mean estimation, see Gneiting (2011) and Gourieroux et al. (1984). In Figure 6.5 we present the estimation errors for Data Set 2 with an increased sample size of $n = 100,000$ compared to Figure 2.2 and 6.4. The classical Gamma GLM leads to estimation errors of order $+5\%$ and -10%, respectively, this is substantially smaller than in Figure 2.2 (right) for sample size $n = 20,000$. If we apply the iterated quasi-likelihood method with the variances estimated with an isotonic regression we get an almost perfect fit with estimation errors of order $+0.5\%$ and -0.5%, respectively, see Figure 6.5 (right). This example clearly shows that even a large insurer with a large data set should still use an optimally estimated variance function to get correct mean estimates.

7 Example 4: MLE of mean with lognormal distributions

In the synthetic Data Set 1, the observations were generated by lognormal distributions. So far we have used a Gamma GLM and a quasi-likelihood approach, under the specification of the first two moments of the distribution, to estimate the mean values of these lognormal observations. Given the information about the true distribution, it is natural to perform MLE under the true distribution to gain full efficiency of the mean estimator. MLE for mean estimation is generally difficult under lognormal distributions since we have to simultaneously, and correctly, estimate the mean and the variance of the response on the log scale (MLE is performed on logged responses and

an exponential transformation is applied to the mean and variance estimates on the log scale to get the mean estimate on the original scale). In contrast to the quasi-likelihood method, if we misspecify the variance function, then the MLE for the mean value is no longer strongly consistent. Hence, variance estimation plays a greater role in MLE of lognormal models than in the quasi-likelihood estimation approach of the mean values of lognormally distributed responses.

In this section we use the same parametrization of the response as in a GLM framework, i.e., we assume that

$$\mathbb{E}[Y_i|\boldsymbol{x}_i] = g^{-1}\big(\boldsymbol{x}_i^\top \boldsymbol{\theta}\big) = g^{-1}\big(\eta(\boldsymbol{x}_i, \boldsymbol{\theta})\big) \quad \text{and} \quad \mathrm{Var}[Y_i|\boldsymbol{x}_i] = \phi V\big(\mathbb{E}[Y_i|\boldsymbol{x}_i]\big), \tag{7.1}$$

for all instances $i = 1, \ldots, n$. In particular, under this parametrization we can directly compare our mean estimation results to the ones of the quasi-likelihood aproach, since the models are estimated under the same assumptions concerning the two moments of the response, which agree with the assumptions used to generate Data Sets 1 and 2. A different parametrization of the lognormal distribution, derived under different assumptions concerning the two moments of the response, is discussed in the next section.

Since we want to apply MLE under an (a priori known) lognormal distribution assumption of the response, we can switch to normally distributed responses. If $Y|\boldsymbol{x}$ has a lognormal distribution with mean $\mathbb{E}[Y|\boldsymbol{x}]$ and variance $\mathrm{Var}[Y|\boldsymbol{x}]$, then $Z = \log(Y)$, conditional on $\boldsymbol{x}$, is normally distributed with mean and variance

$$\mathbb{E}[Z|\boldsymbol{x}] = \log(\mathbb{E}[Y|\boldsymbol{x}]) - \frac{1}{2}\mathrm{Var}[Z|\boldsymbol{x}] \qquad \text{and} \qquad \mathrm{Var}[Z|\boldsymbol{x}] = \log\left(1 + \frac{\mathrm{Var}[Y|\boldsymbol{x}]}{\big(\mathbb{E}[Y|\boldsymbol{x}]\big)^2}\right). \tag{7.2}$$

The variable Z is the response on the log scale, $\mathbb{E}[Z|\boldsymbol{x}]$ is the mean on the log scale and $\mathrm{Var}[Z|\boldsymbol{x}]$ is the variance on the log scale. We introduce an auxiliary response variable on the log scale

$$\widetilde{Z} = Z + \frac{1}{2}\mathrm{Var}[Z|\boldsymbol{x}], \tag{7.3}$$

with the first two moments

$$\begin{aligned}
\widetilde{g}(\mathbb{E}[\widetilde{Z}|\boldsymbol{x}]) &:= g\Big(e^{\mathbb{E}[\widetilde{Z}|\boldsymbol{x}]}\Big) = g\Big(e^{\mathbb{E}[Z|\boldsymbol{x}] + \frac{1}{2}\mathrm{Var}[Z|\boldsymbol{x}]}\Big) = g(\mu(\boldsymbol{x}, \boldsymbol{\theta})) = \boldsymbol{x}^\top \boldsymbol{\theta} = \eta(\boldsymbol{x}, \boldsymbol{\theta}), \\
\mathrm{Var}[\widetilde{Z}|\boldsymbol{x}] &= \log\left(1 + \frac{\mathrm{Var}[Y|\boldsymbol{x}]}{\big(\mathbb{E}[Y|\boldsymbol{x}]\big)^2}\right) \\
&= \log\left(1 + \frac{V(\mu(\boldsymbol{x}, \boldsymbol{\theta}))}{\big(\mu(\boldsymbol{x}, \boldsymbol{\theta})\big)^2}\right) =: \widetilde{V}\big(\eta(\boldsymbol{x}, \boldsymbol{\theta})\big) = \mathrm{Var}[Z|\boldsymbol{x}],
\end{aligned} \tag{7.4}$$

where $\widetilde{V}$ is a variance function of the response on the log scale seen as a function of the linear predictor η. We remark that η is directly linked to the mean value of the response on the original scale by (7.1) and differs from the mean value of the response on the log scale (7.2).

In particular, we choose the logarithmic link $g(\mu) = \log(\mu)$. Models with parametrization (7.1) and logarithmic link are called log-linear models in the statistical literature, see, e.g., Chapter 6 in McCullagh–Nelder (1983). This choice of g is made for our synthetic data sets, and it is further

discussed in the numerical examples, below. In this case, $\widetilde{g}$ is the identity function and the expected value of the auxiliary response $\widetilde{Z}$ directly coincides with the linear predictor we want to estimate for the mean value of Y. That is, we have

$$\mathbb{E}[\widetilde{Z}|\boldsymbol{x}] = \mathbb{E}[Z|\boldsymbol{x}] + \frac{1}{2}\mathrm{Var}[Z|\boldsymbol{x}] = \boldsymbol{x}^\top \boldsymbol{\theta} = \eta(\boldsymbol{x}, \boldsymbol{\theta}). \tag{7.5}$$

The variance function $\widetilde{V}$ becomes here a function of the logarithm of the mean value of the response on the original scale.

There are three difficulties in the estimation procedure on the log scale under the moment assumptions (7.1) on the original scale. First, the auxiliary response $\widetilde{Z}$ is not directly observable, but only Y and Z, respectively. Second, the variance of $\widetilde{Z}$ is not known, and from the previous discussions we know that the information about the variance is beneficial for the estimation of the regression function for $\mathbb{E}[\tilde{Z}|\boldsymbol{x}]$, from which we can next derive the mean estimate of $\mathbb{E}[Y|\boldsymbol{x}]$ by a transformation of $\mathbb{E}[\tilde{Z}|\boldsymbol{x}]$. Third, the definition of $\widetilde{Z}$ involves the unknown variance of $\tilde{Z}$. The simplest approach is to assume that $\mathrm{Var}[Z_i|\boldsymbol{x}_i]$ is constant and independent of $\boldsymbol{x}_i$ for all instances $i = 1, \ldots, n$, hence Z_i is homoskedastic, then $Z_i = \widetilde{Z}_i$ up to a constant and we can fit a linear regression function with parameter $\boldsymbol{\beta}$ to $(Z_i, \boldsymbol{x}_i)_{i=1}^n$ using least squares (MLE on logged responses within a Gaussian linear model). The variance of the error term in the linear homoskedastic regression model can then be estimated in a second step with the classical Pearson's dispersion estimator $\widehat{\sigma}^2$. In this approach, under the logarithmic link, we receive expected value of the response Y on the original scale (after back-transformation)

$$\widehat{\mathbb{E}}[Y_i|\boldsymbol{x}_i] = e^{\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} + \frac{1}{2}\widehat{\sigma}^2}, \tag{7.6}$$

where $\boldsymbol{\beta}$ differs from $\boldsymbol{\theta}$ only in the intercept by the variance of the Gaussian error term. The model estimated with this approach (and the mean estimates derived by (7.6)) is called *Linear Model on log scale (LM on log scale)* in the sequel, since we fit a homoskedastic linear model to logged responses. This approach totally neglects the difficulties mentioned above – in general, it neglects the variance function of Z and the property that $Z = \log(Y)$ is possibly heteroskedastic, which arises if the variance function V of the response Y is not quadratic.

In Figure 7.1 we observe that the LM on log scale leads to poor results. The T-reliability diagram does not look convincing and the miscalibration test rejects the null hypothesis of having an auto-calibrated mean model. Here, we apply a parametric bootstrap since in this section we use the assumption that the response is lognormally distributed. The expected values of the responses for the bootstrap are estimated by (7.6) and the variances of the responses, under the estimated expected values, are estimated non-parametrically with an isotonic regression, as in Section 4. In Figure 7.2 we also compare the mean estimates from the LM on log scale calculated with (7.6) with the true mean values to verify the conclusion that the mean values are not properly estimated, as we do not account for heteroskedasticity.

From (7.1) and (7.4) we can see that the mean estimate of the response on the original scale $\mathbb{E}[Y_i|\boldsymbol{x}_i]$ uses the estimate of the regression function for the response on the log scale $\mathbb{E}[\widetilde{Z}_i|\boldsymbol{x}_i]$, which depends on the variance estimate of the response on the log scale $\mathrm{Var}[Z_i|\boldsymbol{x}_i]$, hence, we should take into account a proper (best possible) estimation of the variance function of the response on the

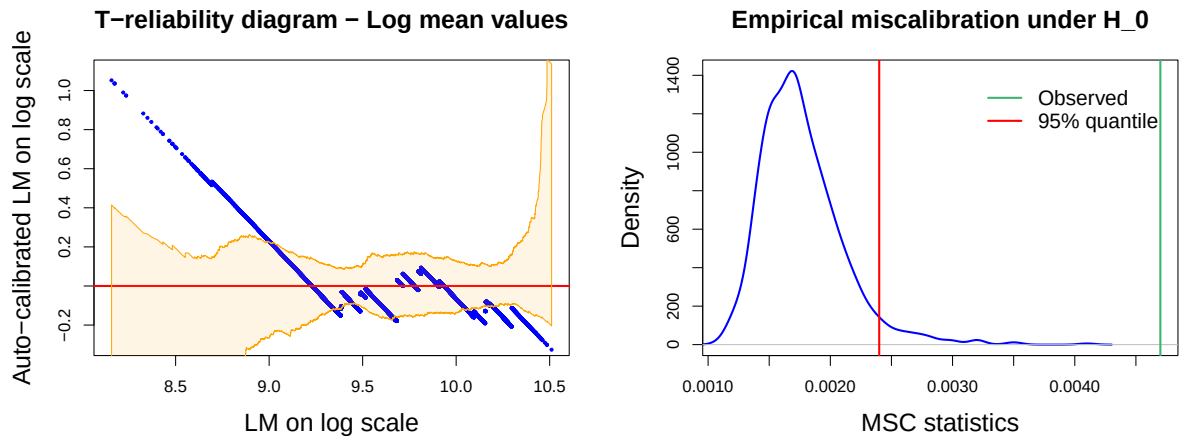

Figure 7.1: (left) T-reliability diagram together with the 95% consistency band for mean estimates from the LM on log scale, and (right) distribution of the miscalibration statistics, together with the observed value and the critical value at signficance level 5% – the variance function is estimated with an isotonic regression. T-reliability diagram presents logged mean estimates.

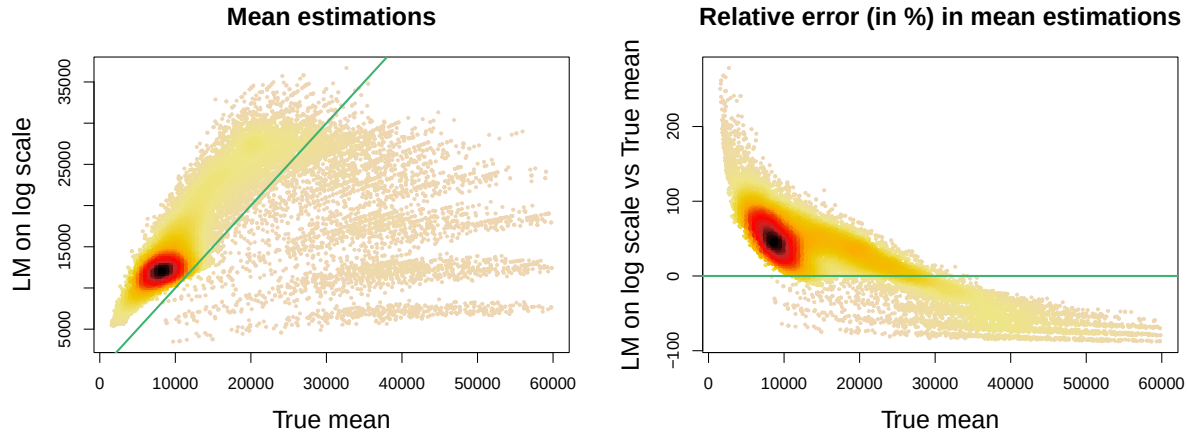

Figure 7.2: Estimated means from the LM on log scale against true means for all instances; the darker the color, the larger the density of the observations with particular means.

log scale to get good predictors on the original scale. In the framework of parametrization (7.1), we propose the following MLE algorithm for the mean values of lognormally distributed responses together with a non-parametric estimation of the variance function. To the best of our knowledge this algorithm is new in the literature.

Lognormal Linear Model with Non-Parametric Variance Estimation

---

- **Step 1:** We estimate the linear predictor (the mean of the response $Y|\boldsymbol{x}$ transformed with g) with a Gaussian GLM with link $\widetilde{g}$ and unit weights based on the observations $(Z_i, \boldsymbol{x}_i)_{i=1}^n$. We get an initial estimate of the parameter $\widehat{\boldsymbol{\theta}}^0$, the linear predictors $\left(\widehat{\eta}^0(\boldsymbol{x}_i)\right)_{i=1}^n$ and the

estimated mean values $\left(\widehat{\mu}^0(\boldsymbol{x}_i)\right)_{i=1}^n$, where $\widehat{\eta}^0(\boldsymbol{x}_i) = \boldsymbol{x}^\top \widehat{\boldsymbol{\theta}}^0$ and $\widehat{\mu}^0(\boldsymbol{x}_i) = g^{-1}\left(\boldsymbol{x}^\top \widehat{\boldsymbol{\theta}}^0\right)$ for all instances $i = 1, \ldots, n$.

- **Step 2:** We estimate the variance of the response on the log scale using the crude estimator

$$\widehat{\widetilde{v}}_i^0 = \frac{\left(Z_i - \log\left(g^{-1}(\widehat{\eta}_i^0)\right)\right)^2}{1 - h_i^0},$$

where $(h_i^0)_{i=1}^n$ denote the hat values of the Gaussian GLM fitted in the initial step.

- **Step 3:** We estimate the variance function $\widetilde{V}$ of the response on the log scale based on the observations $\left(\widehat{\widetilde{v}}_i^0, \widehat{\eta}^0(\boldsymbol{x}_i)\right)_{i=1}^n$. We use an isotonic regression assuming that the true variance function $\widetilde{V}(\eta)$ is monotonically increasing in η. This step gives us an estimate of the variance function $\eta \mapsto \widehat{\widetilde{V}}^0(\eta)$ and the variances $\widehat{\widetilde{V}}^0\left(\widehat{\eta}^0(\boldsymbol{x}_i)\right)$ for all instances $i = 1, \ldots, n$.

- **Step 4:** We iterate for $k = 1$ to K:

  (i) Estimate the linear predictor (the mean of the response $Y|\boldsymbol{x}$ transformed with g) with a Gaussian GLM with link $\widetilde{g}$ and the weights $\left(\widehat{\widetilde{V}}^{k-1}\left(\widehat{\mu}^{k-1}(\boldsymbol{x}_i)\right)\right)^{-1}$ based on the observations $\left(Z_i + \frac{1}{2}\widehat{\widetilde{V}}^{k-1}\left(\widehat{\eta}^{k-1}(\boldsymbol{x}_i)\right), \boldsymbol{x}_i\right)_{i=1}^n$.

  (ii) We get new estimates of $\widehat{\boldsymbol{\theta}}^k$, $\left(\widehat{\eta}^k(\boldsymbol{x}_i)\right)_{i=1}^n$ and $\left(\widehat{\mu}^k(\boldsymbol{x}_i)\right)_{i=1}^n$, where $\widehat{\eta}^k(\boldsymbol{x}_i) = \boldsymbol{x}^\top \widehat{\boldsymbol{\theta}}^k$ and $\widehat{\mu}^k(\boldsymbol{x}_i) = g^{-1}\left(\boldsymbol{x}^\top \widehat{\boldsymbol{\theta}}^k\right)$ for all instances $i = 1, \ldots, n$.

  (iii) We re-estimate the variance of the response on the log scale using the crude estimator

$$\widehat{\widetilde{v}}_i^k = \frac{\left(Z_i + \frac{1}{2}\widehat{\widetilde{V}}^{k-1}\left(\widehat{\eta}^{k-1}(\boldsymbol{x}_i)\right) - \log\left(g^{-1}\left(\widehat{\eta}^k(\boldsymbol{x}_i)\right)\right)\right)^2}{1 - h_i^k}.$$

  (iv) We re-estimate the variance function $\widetilde{V}$ of the response on the log scale based on the observations $\left(\widehat{\widetilde{v}}_i^k, \widehat{\eta}^k(\boldsymbol{x}_i)\right)_{i=1}^n$ using an isotonic regression. We get new estimates of $\eta \mapsto \widehat{\widetilde{V}}^k(\eta)$ and $\widehat{\widetilde{V}}^k\left(\widehat{\eta}^k(\boldsymbol{x}_i)\right)$ for all instances $i = 1, \ldots, n$.

The model fitted with the above algorithm to Data Set 1 is called *Log-Linear Heteroskedastic Model (Log-LHM)*, since we assume a log-linear structure for the mean value of the response (hence it is a log-linear model) and we model the variance of the response (we additionally allow for heteroskedasticity). We use $K = 15$. The results of our estimation are presented in Figures 7.3-7.6. They show that we manage to estimate the mean values correctly with this algorithm. As in the previous section, we estimate the variance function (of the response on the log scale) both with a local polynomial regression and an isotonic regression, see Figure 7.3, and we compare the estimation results to the true variance function. Again, the isotonic regression, applied without any hyperparameter fine-tuning, provides a better fit than the local regression, which requires optimal

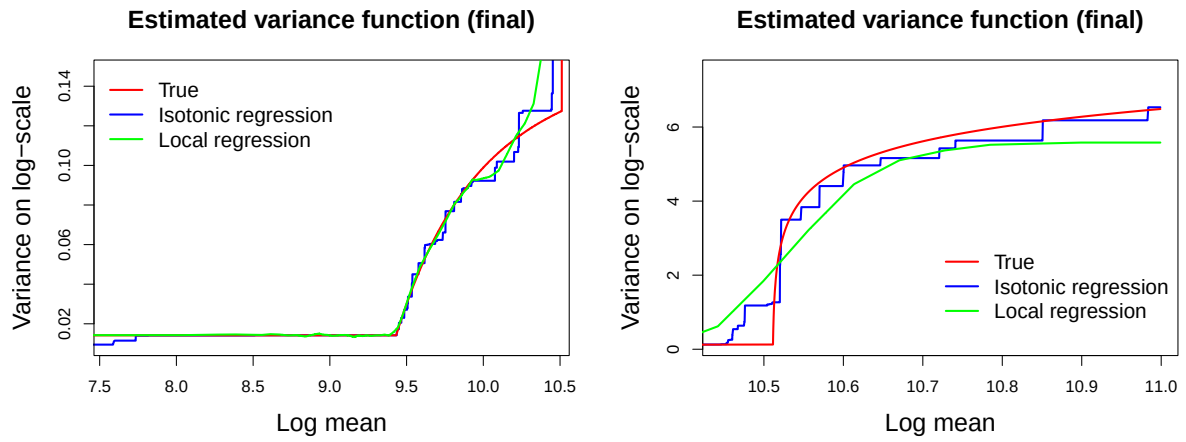

Figure 7.3: Variance function of the response on the log scale estimated with an isotonic regression and a local regression (kernel smoother) based on means estimated with the Log-LHM for different ranges of logged mean values.

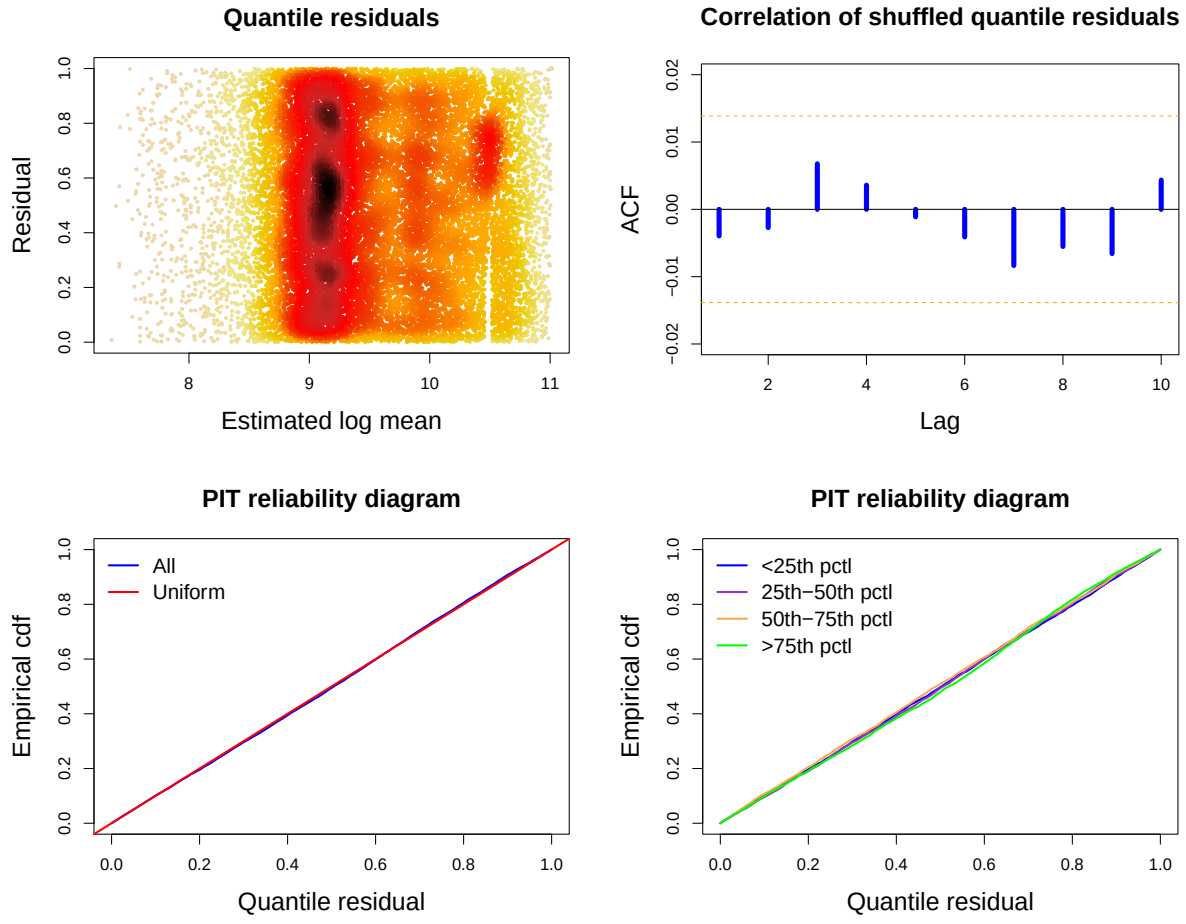

Figure 7.4: Quantile residuals from the Log-LHM; (top-left) the darker the color, the larger the density of the observations with particular means.

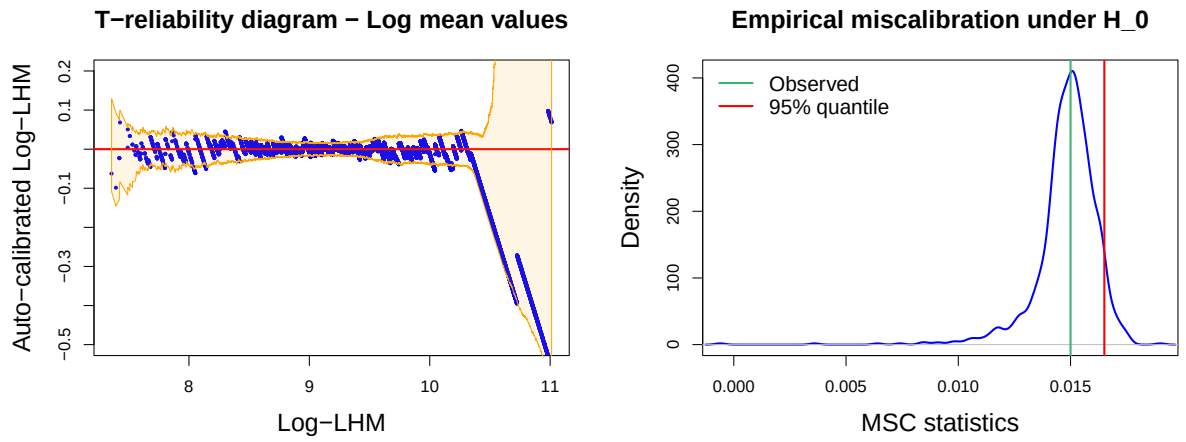

Figure 7.5: T-reliability diagram, together with the 95% consistency band, for mean estimates from the Log-LHM and distribution of the miscalibration statistics, together with the observed value and the critical value at confidence level 95% – the variance function is estimated with an isotonic regression. T-reliability diagram presents logged mean estimates.

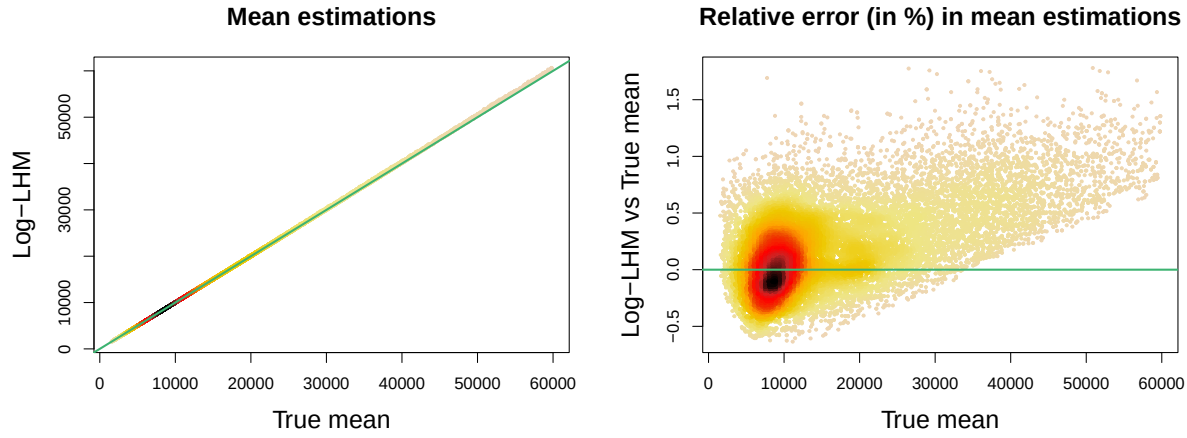

Figure 7.6: Estimated means from the Log-LHM against true means for all instances for Data Set 1; the darker the color, the larger the density of the observations with particular means; remark that the y-scale of the plot on the right hand side differs from the one in Figures 6.4 and 7.2.

selection of hyperparameters (we choose a local regression with degree equal to zero, rectangular kernel and smoothing parameter $\alpha = 0.06$). We point out that by applying the estimation algorithm from this section with non-parametric estimation of the variance, we are now able to fit correctly the variance function also for the observations with the largest mean values, this was not possible in the quasi-likelihood approach of the previous section.

As a part of the model diagnostics, we study the quantile residuals from the Log-LHM approach. The quantile residuals are defined using the lognormal distribution with the first two fitted moments. Figure 7.4 shows that the structure of the quantile residuals looks perfect, which, in particular, confirms that the Log-LHM provides appropriate mean and variance estimates. Using the parametric bootstrap with lognormal distributions with the first two fitted moments, we ob-

tain the T-reliability diagram and its 95% consistency band, as well as the miscalibration test, see Figure 7.5. From this figure we conclude that we cannot reject the null hypothesis of having an auto-calibrated model, hence, everything looks fine, here. Finally, in Figure 7.6 we compare the mean estimates on the original scale from the Log-LHM with the true mean values. This figure shows that we arrive at very accurate estimates of the expected values of the responses. Comparing Figure 7.6 with Figure 6.4, we can also conclude that the approach from this section gives us slightly better estimates of the mean values, closer to the true mean values in terms of relative errors, than the approach from the previous section. This is expected since in this section we perform the MLE of the mean values, in which the whole distribution of the response is correctly specified, whereas in the previous section we apply the quasi-likelihood approach, in which only the first two moments are correctly specified.

We conclude that we can apply our algorithm to fit the mean values with MLE and the variance function with an isotonic regression (under appropriate monotonicity assumption) for lognormal responses in a GLM moment type framework. Moreover, our algorithm does not require any hyperparameter fine-tuning.

8 Example 5: MLE estimation on logged data and non-parametric back-transformation

We continue with MLE for lognormal distributions. In actuarial applications, there is a common practice to logarithmize the observations before the modeling process is started. Taking the logarithmic transformation, heavy-tailed claims become more light-tailed and more symmetric, and one can apply a linear regression with Gaussian errors (under the assumption that the responses come from lognormal distributions), without the drawback of over-fitting to large observations. Once the response is logarithmically transformed, a regression function is specified for the expected value of the response on the log scale. This approach uses a different assumption concerning the first two moments of lognormally distributed responses than the assumption investigated in the previous section.

In this section we use a different parametrization of the lognormal responses compared to Section 7. We now assume that $Y|\boldsymbol{x}$ has a lognormal distribution with mean parameter $\lambda(\boldsymbol{x}) \in \mathbb{R}$ and variance parameter $\sigma^2(\boldsymbol{x}) > 0$ on the log scale, i.e., the response has the first two moments

$$\mathbb{E}[Y|\boldsymbol{x}] = \exp\left\{\lambda(\boldsymbol{x}) + \frac{\sigma^2(\boldsymbol{x})}{2}\right\} \qquad \text{and} \qquad \mathrm{Var}[Y|\boldsymbol{x}] = \mathbb{E}[Y|\boldsymbol{x}]^2 \left(e^{\sigma^2(\boldsymbol{x})} - 1\right). \tag{8.1}$$

This is equivalent to saying that $Z = \log(Y)|\boldsymbol{x}$ has a Gaussian distribution with mean $\lambda(\boldsymbol{x})$ and variance $\sigma^2(\boldsymbol{x})$, that is, the response on the log scale has the first two moments

$$\mathbb{E}[Z|\boldsymbol{x}] = \lambda(\boldsymbol{x}) \qquad \text{and} \qquad \mathrm{Var}[Z|\boldsymbol{x}] = \sigma^2(\boldsymbol{x}) = \mathcal{V}(\lambda(\boldsymbol{x})). \tag{8.2}$$

The difference between (8.2) and (7.1) is that in (8.2) the regression function is specified for the expected value of the logarithmically transformed response ($\lambda(\boldsymbol{x})$ is linked to a linear predictor $\eta(\boldsymbol{x})$), whereas in (7.1) the regression function is specified for the expected value of the response on the original scale ($\mu(\boldsymbol{x})$ is linked to a linear predictor $\eta(\boldsymbol{x})$). By the remark after (7.4), the variance

functions in the two parametrizations are functions of different objects. As far as the mean value of the response on the original scale is concerned, parametrization (8.1) uses the variance on the log scale in the definition of the mean, whereas log-linear models with parametrization (7.1) specify the mean value independently of the variance function. On the log scale, (8.2) leads to a mean function independent of the variance function, whereas (7.4) introduces a dependence of the mean value of the response on its variance. Under both parametrizations, we require a precise estimation of the mean and the variance of the response on the log scale to derive the mean estimate of the response on the original scale. Depending on the application, one of the two parametrizations (7.1) or (8.2) may be more suitable. The form of the parametrization of lognormal distribution to be used in a regression problem should depend on the data and the results of feature engineering on the original and the log scale.

The algorithm presented in Section 7 can easily be adapted to handle the parametrization from this section. We just set $\widetilde{Z} = Z$. In fact, here we just follow the ideas of Carroll (1982), Müller–Stadtmüller (1987) and Davidian–Carroll (2012) on the log scale, and we replace polynomial regression with isotonic regression for variance estimation. Once we get estimates of $\big(\widehat{\lambda}(\boldsymbol{x}_i)\big)_{i=1}^n$ and $\widehat{\mathcal{V}}(\widehat{\lambda}\big(\boldsymbol{x}_i)\big)$ for all instances $i = 1, \ldots, n$ on the log scale, we can predict the responses on the original scale with

$$\widehat{\mathbb{E}}[Y_i|\boldsymbol{x}_i] = \exp\left\{\widehat{\lambda}(\boldsymbol{x}_i) + \frac{\widehat{\mathcal{V}}\big(\widehat{\lambda}(\boldsymbol{x}_i)\big)}{2}\right\}. \tag{8.3}$$

Prediction (8.3) uses a parametric model for the mean parameter $\lambda(\boldsymbol{x})$ and a non-parametric model for the variance parameter $\sigma^2(\boldsymbol{x})$. If this non-parametric estimator comes from an isotonic regression w.r.t. the mean parameter $\lambda(\boldsymbol{x})$, we can also explore a different back-transformation of the estimates from the log scale to the original scale. Namely, under the assumption that $\lambda \mapsto \mathcal{V}(\lambda)$ is monotonically increasing, which allows us to fit an isotonic regression to model the variance function $\mathcal{V}$ on the log scale, $\lambda(\boldsymbol{x})$ and $\exp\{\lambda(\boldsymbol{x}) + \mathcal{V}(\lambda(\boldsymbol{x}))/2\}$ have the same ordering, and we can directly apply an isotonic regression to the sample $\big(Y_i, \widehat{\lambda}(\boldsymbol{x}_i)\big)_{i=1}^n$ providing us with a non-parametric mean estimate for Y_i, given $\boldsymbol{x}_i$.

We investigate the synthetic Data Set 3. For pedagogical reasons, we prefer not to use Data Set 1 since the first moment of the responses in this data set does not fulfill assumption (8.2) with a simple parametric function $\lambda(\boldsymbol{x})$. Hence, we would fail to validate our mean estimation results, achieved with linear regression functions, against the true mean values. For this reason, we create a new synthetic data set with observations which satisfy (8.2). As discussed in Section 2, we set $\lambda(\boldsymbol{x}_i) = \log(\mu(\boldsymbol{x}_i)) = \eta(\boldsymbol{x}_i)$, where $\big(\mu(\boldsymbol{x}_i)\big)_{i=1}^n$ denote the true mean values of the claim sizes from the synthetic Data Set 1. Hence, the assumed true mean function on the log scale $\lambda(\boldsymbol{x})$ is now a linear function of features $\boldsymbol{x}$ (the same function as the logged mean function used in Data Sets 1 and 2). The true variances of the claim sizes on the log scale are defined based on the variance function (2.1) with the following transformation

$$\mathcal{V}(\lambda) = \phi \log\left(1 + \frac{V\big(e^\lambda\big)}{\big(e^\lambda\big)^2}\right).$$

The estimation results are presented in Figures 8.1-8.4. We only need $K = 5$ iterations. To avoid misunderstanding, let us point out that by the mean value (estimate) on the log scale we mean $\lambda(\boldsymbol{x}_i)$

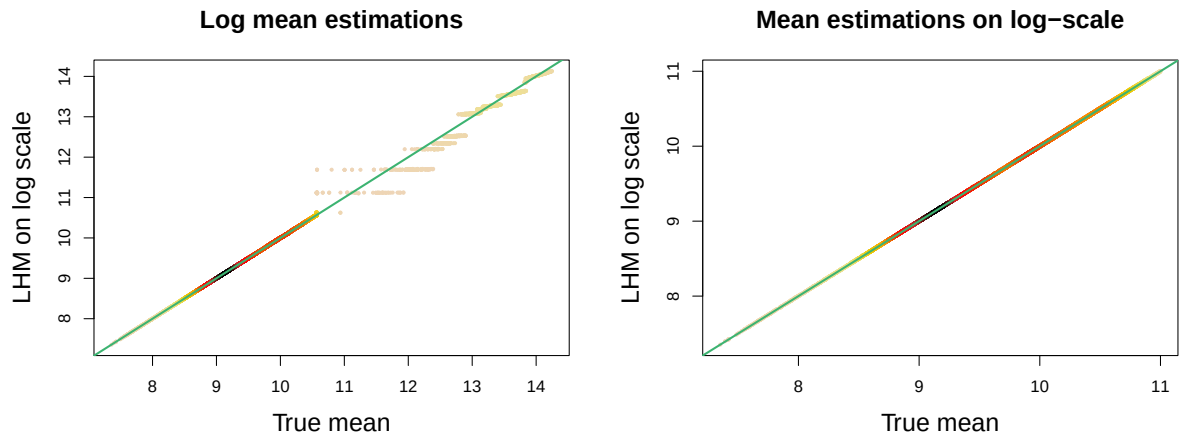

Figure 8.1: (left) Logged estimated means from the LHM on log scale against logged true means, and (right) estimated means on the log scale from the LHM on log scale against true means on the log scale for all instances; the darker the color, the larger the density of the observations with particular means.

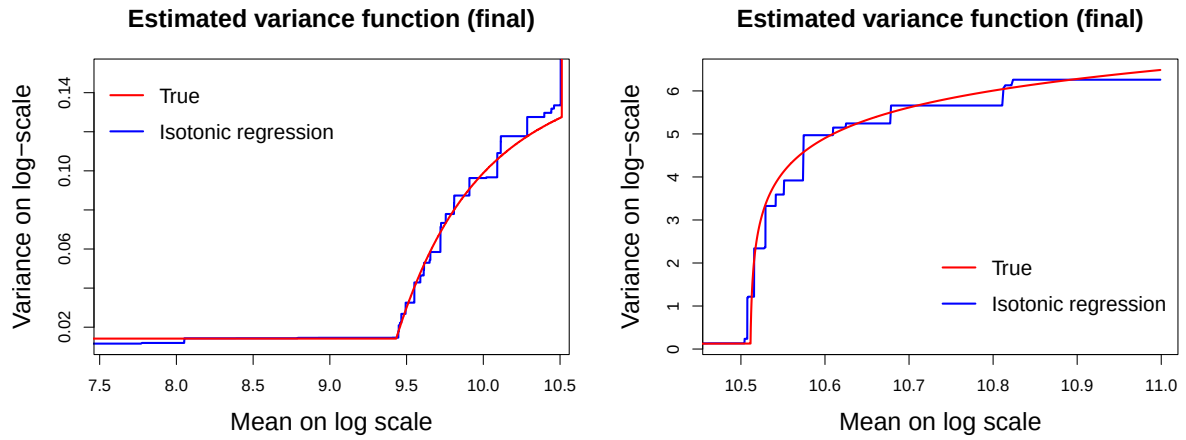

Figure 8.2: Variance function of the response on the log scale estimated with an isotonic regression based on means estimated with the LHM on log scale for different ranges of mean values on the log scale.

and $\widehat{\lambda}(\boldsymbol{x}_i)$, respectively, whereas by the logged mean value (estimate) we mean $\log(\mathbb{E}[Y|\boldsymbol{x}_i])$ and $\log(\widehat{\mathbb{E}}[Y|\boldsymbol{x}_i])$, respectively. Clearly, $\log(\mathbb{E}[Y|\boldsymbol{x}_i]) \neq \lambda(\boldsymbol{x}_i)$, due to Jensen's inequality. We fit a so-called *Linear Heteroskedastic Model on log scale (LHM on log scale)*, since we fit a heteroskedastic linear model to logged responses. If the Gaussian errors on the log scale would have constant variance, then our LHM on log scale would reduce to the LM on log scale, fitted as a naive model in Section 7. The mean and variance estimates on the log scale from the LHM on log scale are transformed to the mean estimates on the original scale by (8.3). We can observe that our approach works very well on the log scale and on the original scale. Of course, the estimates of the mean values of the responses on the original scale have a poorer fit than in the previous sections since under parametrization (8.1) the mean value on the original scale depends on the variance of the log scale,

which is approximated here with a step function from an isotonic regression, and this approximation is crude in the tail. In Figures 8.3 we present the results of applying an isotonic regression directly to $\big(Y_i, \widehat{\lambda}(\boldsymbol{x}_i)\big)_{i=1}^n$. In Figure 8.4 we confirm that our fully non-parametric estimates of the mean responses on the original scale are auto-calibrated (we implemented the parametric bootstrap as in the previous section).

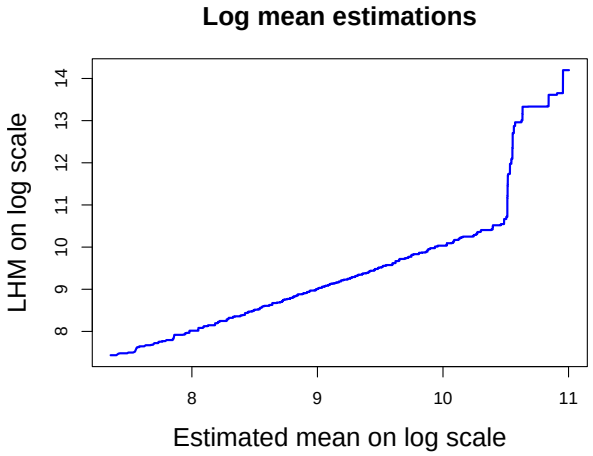

Figure 8.3: Logged non-parametric mean function estimated with an isotonic regression based on mean estimates from the LHM on log scale as a function of the ranks for the mean estimates on the log scale.

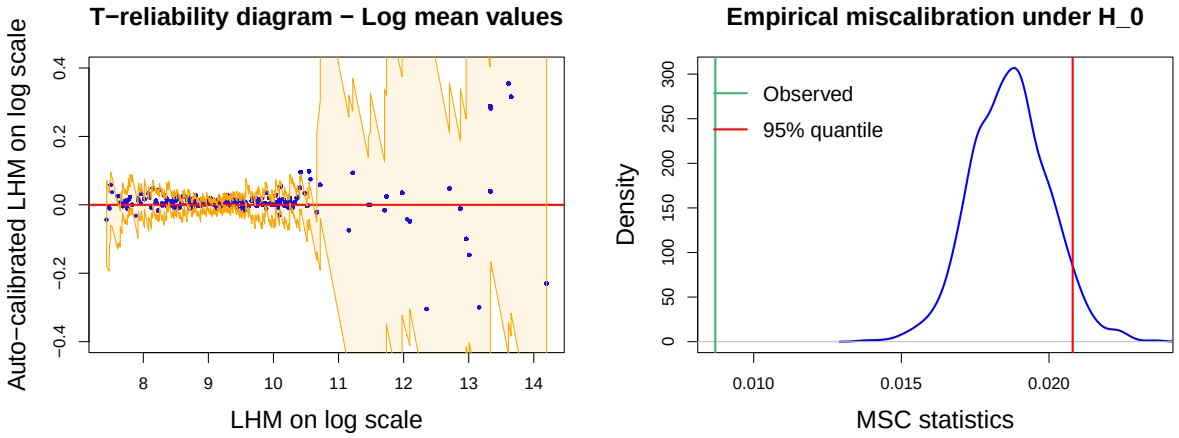

Figure 8.4: (left) T-reliability diagram together with the 95% consistency band for the fully non-parametric mean estimates from the LHM on log scale, and (right) distribution of the miscalibration statistics together with the observed value and the critical value at significance level 5% – the variance function and mean function are estimated with isotonic regressions. T-reliability diagram presents logged mean estimates.

We finally comment on the width of the consistency bands which we construct with bootstrap. Wright (1981) derives the asymptotic distribution of the isotonic regression estimate under some assumptions. From his Theorem 1 we can conclude that the variance of the estimator at a particular

point depends on the total number of observations, the growth rate of the true regression function, the value of the density function and the conditional variance of observations at that point. In our data set the observations with large mean values have the largest variances, hence the consistency band is the widest for these observations.

9 Conclusions

In this paper we have emphasized the significance and benefits of an accurate variance estimation for mean estimation in actuarial regression models. We have proposed an istotonic regression as a tool for non-parametric variance estimation as a monotonic function of the mean. Using five examples, we have demonstrated how to incorporate and validate the estimation of the variance function into the process of mean estimation using a quasi-likelihood method for distributions from the EDF and a MLE for lognormal distributions. We believe that our results are of big relevance in actuarial practice. In the paper, we focus on mean estimation within GLMs but our approach with joint mean-variance estimation also applies to GAMs, regression trees, neural networks or any other more complex regression models for mean estimates. An open problem is to derive asymptotic distributions of mean estimates in our framework with an isotonic regression used for estimating the variance function.

References

Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics* **26**, 641-647.

Barlow, R.E., Bartholomew, D.J., Brenner, J.M., Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*. Wiley.

Brunk, H.D., Ewing, G.M., Utz, W.R. (1957). Minimizing integrals in certain classes of monotone functions. *Pacific Journal of Mathematics* **7**, 833-847.

Carroll, R.J. (1982). Adapting for heteroskedasticity in linear models. *The Annals of Statistics* **10(4)**, 1224-1233.

Chiou, J.-M., Müller, H.-G. (1999). Non-parametric quasi-likelihood. *The Annals of Statistics* **27(1)**, 36-64.

Davidian, M., Carroll, R.J. (2012). Variance function estimation. *Journal of the American Statistical Association* **82**, 1079-1091.

Delong, Ł., Lindholm, M., Wüthrich, M.V. (2021). Making Tweedie's compound Poisson model more accessible. *European Actuarial Journal* **11(1)**, 185-226.

Denuit, M., Trufin, J. (2023). Model selection with Pearson's correlation, concentration and Lorenz curves under autocalibration. *European Actuarial Journal* **13(2)**, 871-878.

Denuit, M., Trufin, J. (2019). *Effective Statistical Learning Methods for Actuaries I*. Springer.

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

Dimitriadis, T., Dümbgen, L., Henzi, A., Puke, M., Ziegel, J. (2023). Honest calibration assessment for binary outcome predictions. *Biometrika* **110(3)**, 663-680.

Dimitriadis, T., Gneiting, T., Jordan, A.I. (2020). Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited. *arXiv*:2008.03033.

Dunn, P.K., Smyth G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5(3)**, 236-244.

Dutang, C., Charpentier, A. (2018). CASdatasets R Package Vignette. Reference Manual. Version 1.0-8, packaged 2018-05-20.

Firth, D. (1987). On efficiency of quasi-likelihood estimation. *Biometrika* **74(2)**, 233-245.

Fissler, T., Lorentzen, C., Mayer, M. (2022). Model comparison and calibration assessment: user guide for consistent scoring functions in machine learning and actuarial practice. *arXiv*:2202.12780.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106(494)**, 746-762.

Gneiting, T., Resin, J. (2021). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams and coefficient of determination. *arXiv*:2108.03210.

Gourieroux, C., Monfort, A., Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica* **52(3)**, 681-700.

Hall, P., Carroll, R.J. (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of Royal Statistical Society* **51(1)**, 3-14.

Henzi, A., Puke, M., Dimitriadis, T., Ziegel, J. (2023). A safe Hosmer–Lemeshow test. *arXiv*:2203.00426v3.

Hosmer, D.W., Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* **9**, 1043-1069.

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

Krüger, F., Ziegel, J.F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economics Statistics* **39(4)**, 972-983.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling. *Psychometrica* **29**, 115-129.

Leeuw, de J., Hornik, K., Mair, P. (2009). Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software* **32(5)**, 1-24.

Lindholm, M., Lindskog, F., Palmquist, J. (2023). Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells. *Scandinavian Actuarial Journal*, in press.

Loader, C. (1999). *Local Regression and Likelihood*. Springer.

McCullagh, P. (1983). Quasi-likelihood functions *Annals of Statistics* **11(1)**, 59-67.

McCullagh, P., Nelder, J.A. (1983). *Generalized Linear Models*. Chapman and Hall.

Miles, R.E. (1959). The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika* **46**, 317-327.

Murphy, A.H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology* **12(4)**, 595-600.

Müller, H.G., Stadtmüller, U. (1987). Estimation of heteroskedasticity in regression analysis. *The Annals of Statistics* **15(2)**, 610-625.

Nelder, J.A., Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74(2)**, 221-232.

Pohle, M.-O. (2020). The Murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. *arXiv*:2005.01835.

Politis, D.N. (2015). *Model-Free Prediction and Regression*. Springer.

Ruppert, D., Wand, M.P., Holst, U., Hosjer, O. (2012). Local polynomial variance function estimation. *Technometrics* **39(3)**, 262-273.

Savage, L.J. (1971). Elicitable of personal probabilities and expectations. *Journal of the American Statistical Association* **66/336**, 783-810.

Smyth, G.K., Verbyla, A.P. (1999). Double generalized linear models: approximate REML and diagnostics. In: *Proceedings of the 14th International Workshop on Statistical Modelling*. Friedl, H., Berghold, A., Kauermann, G. (Eds.). Technical University, Graz, Austria, 66-80.

Tweedie, M.C.K. (1984). An index which distinguishes between some important exponential families. In: *Statistics: Applications and New Directions*. Ghosh, J.K., Roy, J. (Eds.). Proceeding of the Indian Statistical Golden Jubilee International Conference, Indian Statistical Institute, Calcutta, 579-604.

Warton, D.I., Thibaut, L., Wang, Y.A. (2017). The PIT-trap - a general bootstrap procedure for inference about regression models with discrete, multivariate responses. *PLoS ONE* **12(7)**, 1-18.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61(3), 439-447.

Wüthrich, M.V. (2023). Model selection with Gini indices under auto-calibration. *European Actuarial Journal* **13(1)**, 469-477.

Wüthrich, M.V., Merz, M. (2023). *Statistical Foundations of Actuarial Learning and its Applications*. Springer.

Wüthrich, M.V., Ziegel, J. (2023). Isotonic recalibration under a low signal-to-noise ratio. *Scandinavian Actuarial Journal*, in press.

Wright, F.T. (1981). The asymptotic behavior of monotone regression estimates. *The Annals of Statistics* **9(2)**, 443-448.

Yang, F., Barber, R.F. (2019). Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics* **13**, 646-677.