A linear model with log-linear variance for insurance claims

M.J.P. Nijmeijer[1] and E. Cator[2]

[1]Posthuma Partners
Zuidelijk Halfrond 1
2801 DD Gouda
the Netherlands
e-mail: nijmeijer@posthuma-partners.nl

[2]Institute of Mathematics, Astrophysics and Particle Physics
Radboud University Nijmegen
Nijmegen
the Netherlands

January 29, 2018

Abstract

We study a method to select motor claims with large claim amounts. The amount is judged by its quantile. Two statistical models to estimate the quantile are considered: a classical linear model and a generalization ('lmvar') in which the variance is not held fixed but has a log-linear structure. The latter turns out to improve the fitted model significantly. The choice of model is important when selecting claims with a high quantile. Of all claims with a quantile larger than 80% according to 'lmvar', 20% has a quantile smaller than 80% according to the classical linear model. Assuming 'lmvar' to be the true model, we generate data and show that the classical linear fit misses roughly 20% of all claims with a percentile larger than 80%. Moreover, the linear fit selects claims with a bias towards claims with a large variance.

1 Introduction

Claims handling is a key task of any insurer. Claim volumes can be substantial, especially in the Property and Casualty line of business, and a proper judgment of claim validity and claim amounts is important to ensure a healthy financial performance. Claims handling is often done by the application of a set of business rules, derived from policy terms and internal guidelines, combined with the experience of the claims handler. In this paper, we describe the application of a statistical model to support the claims handling process further.

The model helps the claims handler to judge the height of the claim amount. It compares the actual claim amount against a model-distribution. Given this distribution, it calculates the quantile of the actual claim amount. The quantile (a number between 0 and 1 or, alternatively, between 0% and 100%) is a measure for the height of the claim: a value close to 0 indicates a small claim amount while a value close to 1 means that the amount is large.

We use two models for our model-distribution: a classical linear model [Chatterjee and Hadi, 2006, Seber, 1977] and a model which is linear in the expectation value and log-linear in the variance (denoted as the 'lmvar' model for short) [Verbyla, 1993]. The main difference between these two models is their treatment of the variance. The linear model assumes a constant variance. In other words, the variance of claims within a risk-class, is the same for all risk-classes. The 'lmvar' model allows different risk-classes to have different variances. It allows for a more flexible treatment of the variance in which the effect of the predictors on the variance becomes explicit [Goldstein, 2014]. Careful modeling of the variance is important if one is interested in quantiles because they are sensitive to the variance, not only to the expected value.

The 'lmvar' model has been present in the statistical literature [Aitkin, 1987, Harvey, 1976, Verbyla, 1993] for a while but it appears to be uncommon in the actuarial sciences. We demonstrate its use for a data set of motor claims. Our objective is to set up model-distributions for car repair costs and calculate quantiles for those costs. We show that it is feasible to use the 'lmvar' model and that it leads to better results than a classical linear model.

We describe the 'lmvar' model in the next section. The data set is described in section 3. Section 4 describes the model design for this dataset. Section 5 describes the linear regression and the 'lmvar' regression to the data.

All calculations are carried out in R [R Core Team, 2017], using the package 'lmvar' [Posthuma Partners, 2017] to carry out the fits to the 'lmvar' model.

2 The linear model with log-linear variance

The 'lmvar' model has a response vector $Y \in \mathbb{R}^n$ with a multivariate normal distribution:

$$Y \sim \mathcal{N}_n(\mu, \Sigma), \tag{1}$$

with $\mu \in \mathbb{R}^n$ the expected value of Y and $\Sigma \in \mathbb{R}^{n,n}$ the covariance matrix.

In a classical linear model, Σ takes the form $\Sigma = \sigma^2 I$ with σ^2 the variance (taken to be the same for all observations Y_i) and I the $n \times n$ identity matrix.

The 'lmvar' model, however, takes

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}, \tag{2}$$

i.e., Σ is still a diagonal matrix but each observation Y_i has its own variance σ_i^2.

The average values μ depend linearly on a matrix of explanatory variables (which we will also call 'predictor variables' or 'covariates') $X_\mu \in \mathbb{R}^{n,k_\mu}$ as

$$\mu = X_\mu \beta_\mu \tag{3}$$

with $\beta_\mu \in \mathbb{R}^{k_\mu}$ the parameter vector of betas for μ. Let σ denotes the vector $(\sigma_1, \ldots, \sigma_n)$ and $\log \sigma$ the vector $(\log \sigma_1, \ldots, \log \sigma_n)$, then we have an additional matrix of predictor values $X_\sigma \in \mathbb{R}^{n,k_\sigma}$ such that

$$\log \sigma = X_\sigma \beta_\sigma \tag{4}$$

with $\beta_\sigma \in \mathbb{R}^{k_\sigma}$ the parameter vector of betas for $\log \sigma$.

The model for σ is sometimes defined as $\log \sigma^2 = X_\sigma \beta_\sigma^\star$ [Verbyla, 1993]. This simply amounts to a rescaling of β_σ: $\beta_\sigma^\star$ thus defined is twice the β_σ defined in (4). We stick to convention (4) though.

A particular predictor (such as the brand of the car or the mileage) can appear in X_μ only, in X_σ only, or in both X_μ and X_σ. In the first case it influences the expected value but not the variance, in the second case it influences the variance but not the expected value and in the third case it influences both the expected value and the variance.

3 The data

Our data set consists of $n = 31,168$ car-repair claims [Tzougas et al., 2015, de Jong and Heller, 2008, Heller et al., 2007]. Each claim contains a claim amount and 13 covariates. There are 11 categorical covariates. These are listed in table 1, together with the number of levels they take (in the column 'original'). The 2 continuous covariates are 'number of parts' and 'mileage'.

The 'number of parts' is the number of parts involved in the repair. It ranges from 1 to 147 with a sample mean of 9.5 and a sample standard deviation of 7.8. The mileage ranges from 0 to 999,999 with a sample mean of 75,530 and a sample standard deviation of 58,367.

The variable 'year of construction' has been treated as a categorical variable, and is therefore listed in the table. It ranges from 1962 to 2015 but there are also 800 observations for which it has the value 0 as well as 1 observation for which it has the value 1.

We exclude observations with a number of parts that exceeds 100, which we consider to be an extreme value. There are 7 such observations, so we remain with 31,161 observations.

Levels of categorical covariates that appear 50 times or less amongst the 31,161 observations, have been excluded. As an example, if a particular model appears 50 times or less, all observations with that model do not have 'model' as a predictor. This is to avoid spurious effects due to insufficient statistics.

The removal of the 7 observations and subsequent removal of all levels that appear fewer than 50 times, reduces the number of levels of the categorical covariates. The remaining number of levels is shown in the column 'remaining' in table 1.

After the exclusion of the 7 observations, the claim amount ranges from 7 to 38,012 with a sample mean of 1,177 and a sample standard deviation of 1,107.

The left figure of fig. 1 shows a scatter plot of the claim amounts versus the number of parts N for the 31,161 observations. There are two clear outliers, in the top-left corner of the plot. They have claim amounts 37,843 and 38,012 respectively, and are repairs on the same car although carried out on different dates. The increasing spread of the data with increasing N, indicates that the standard deviation grows with N.

The right figure shows a boxplot of the claim amount versus the cause of damage. The difference in height of the boxes indicates a difference in standard deviation of the amounts between the different causes.

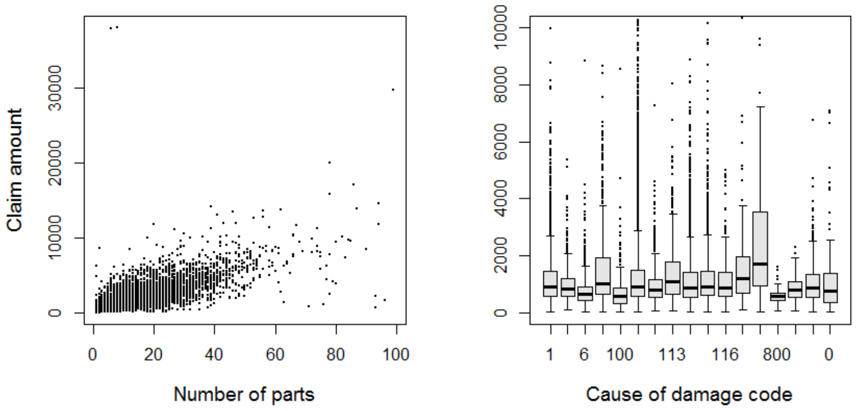

Figure 1: Left figure: scatter plot of the claim amounts versus the number of parts. Right figure: boxplot of the claim amounts versus the cause of damage. The y-scale has been limited to a claim amount of 10,000.

4 Model design

Both the classical linear regression and the 'lmvar' regression take

$$Y = (\log c_1, \dots, \log c_n)$$

as the response vector, where c_i is the claim amount of claim i. The logarithm is the natural logarithm, with base e.

The full model contains the two continuous covariates and all the covariate levels shown in table 1 in the column 'remaining'. The influence of the mileage m is modeled by a polynomial of degree three:

$$m + m^2 + m^3 \tag{5}$$

where each term has its own corresponding β.

The influence of the number of parts N is captured by the following terms:

$$\delta_{N,1} + \delta_{N,2} + \cdots + \delta_{N,10} + f(N) \tag{6}$$

where δ is the Kronecker-delta. The function $f(N)$ is a cubic spline [James et al., 2015] with 5 knots. It is half-sided natural: the spline is linear in the in the region to the right of the last knot. The spline has the terms

$$N + h(N, \zeta_1) + \cdots + h(N, \zeta_5) \tag{7}$$

with

$$h(N, \zeta_i) = \begin{cases} -(N - \zeta_i)^3 & N \leq \zeta_i \\ 0 & N > \zeta_i \end{cases} \tag{8}$$

and ζ_i the i-th knot. The knots are placed at $\zeta_1 = 53.0$, $\zeta_2 = 55.0$, $\zeta_3 = 60.5$, $\zeta_4 = 72.0$ and $\zeta_5 = 83.0$. These are the (1/6)-th, (2/6)-th, (3/6)-th, (4/6)-th and (5/6)-th quantile of the number of parts N that occur in our data-set, restricted to $N > 50$.

Keep in mind that each Kronecker-delta in (6) as well as N and each spline basis-function h in (7) has its own β. Therefore, (6) contains 16 degrees of freedom.

The terms are combined in a design matrix (also called a 'model matrix') which has 31,161 rows and 291 columns (including a column for the intercept term). The matrix is full rank.

5 Classical linear regression versus 'lmvar' regression

In this section we fit the data to a classical linear model (section 5.1) and to a 'lmvar' model (section 5.2).

5.1 Classical linear regression

5.1.1 Full linear model

The full linear model is a classical linear fit of the data set described in section 3, using the full model of section 4 The full linear model contains 31,161 observations and has 292 degrees of freedom.

Various statistics of the fit are shown in table 2 under 'All observations'. Note that $\sqrt{\mathrm{MSE}} = 683$, which is significantly smaller than the sample standard deviation 1,107. The small p-value in table 2 indicates that the Kolmogorov-Smirnov (KS) distance [Bohm and Zech, 2010] is significantly larger than zero. Hence the distribution of the response vector Y differs noticeably from a truly linear model with model matrix X, parameter vector β and variance σ^2 as in the full linear model.

We check the values for bias in a 10-fold cross-validation. This bias can originate from the fact that the statistics are calculated from the same observations which are used in the fit. The cross-validation results are also listed in table 2.

Except for the p-value of the KS-statistic, the results for the cross-validation are essentially the same as calculated from the full fit. The much larger p-values in the cross-validation occur because the number of observations on which the KS test is carried out in the cross-validation, is one-tenth of the total number of observations of the full data-set. A difference between the actual distribution of the response-vector and a truly linear model, is established with less certainty for smaller datasets.

5.1.2 Model reduction

Many of the fitted β's have large p-values, indicating that they can be set to 0 without introducing a large amount of bias in the model. to reduce the number of terms in the model, we look for the subset of terms that results in the lowest value of the Akaike Information Criterion (AIC) [Akaike, 1973]. We try to identify this subset by various means of variable-selection: the LASSO method [Tibshirani, 1996], a forward/backward-step method [James et al., 2015] and by means of simulated annealing [Aarts and Korst, 1988].

The LASSO method generates a sequence of models, corresponding to a sequence of values for the penalty-parameter λ. We select the model with the lowest AIC from this sequence. Different LASSO runs with different random numbers generate slightly different sequences, resulting in differences between the selected models. We find models with degrees of freedom ranging from 151 to 184 and AIC values ranging from 40,643 to 40,645.

The forward/backward step method results in a model with 155 degrees of freedom and AIC = 40,541.

Three subsequent simulated annealing runs result in models with 138 $\sim$ 143 degrees of freedom and AIC values in the range 40,529 $\sim$ 40,533.

5.1.3 Reduced linear model

Of all reduced models, the model with the smallest AIC has 138 degrees of freedom and AIC = 40,529. We call this model A. In comparison with the full model, it has a less than half of the degrees of freedom (138 versus 292). Other values for model A are listed in table 2. Apart from a smaller AIC value, there are no significant differences with the full linear model. Judged from the various error estimates, the reduction of the degrees of freedom did neither reduce the variance of the model, nor introduce extra bias.

The number of levels of each categorical covariate that remains in model A is shown in table 1. All terms from (5) have survived the reduction.

Of (6), all Kronecker-deltas and the terms N and $h(N, \zeta_5)$ remained. The terms $h(N, \zeta_1)$, $h(N, \zeta_2)$, $h(N, \zeta_3)$ and $h(N, \zeta_4)$ dropped out. Figure 2 shows the contribution of the number of parts N to the expected value of the logarithm of the claim amount.

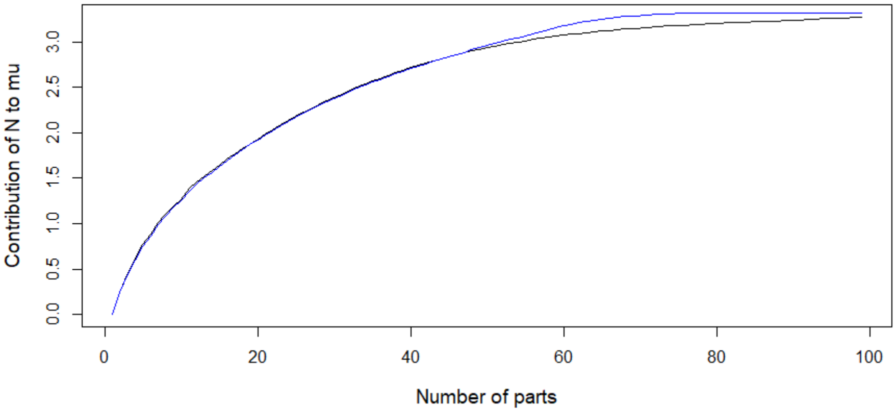

Figure 2: The contribution of the surviving terms of (6) in the reduced linear model (model A, black line) and the reduced 'lmvar' (model B, blue line) to the expected value of the log of the claim amount. The contribution has been set to zero for $N = 1$

Figure 3 shows the QQ-plot of the fit of model A (black points). Ideally, all points should fall on the red line. In the top-right corner, there are a few individual points with higher claim amounts than can be accounted for by the linear model. Of the four points that deviate most from the red line, two are the outliers shown in the left-hand figure of fig 1. They have very high claim amounts (37,843 and 38,012). The remaining 2 points also have relatively high claim amounts (8,536 and 6,269) but appear to have no special features otherwise.

The large set of points in the bottom-left corner falling below the red line, shows that the fit underestimates the probability that the claim amount is small. In other words, the 'true' probability distribution has more probability mass in its left tail than the fitted distribution.

Figure 4 shows the distribution of the quantiles for model A. Ideally, this distribution is uniform. The deviations of the fitted linear model from the true distribution, cause the shape seen in the figure.

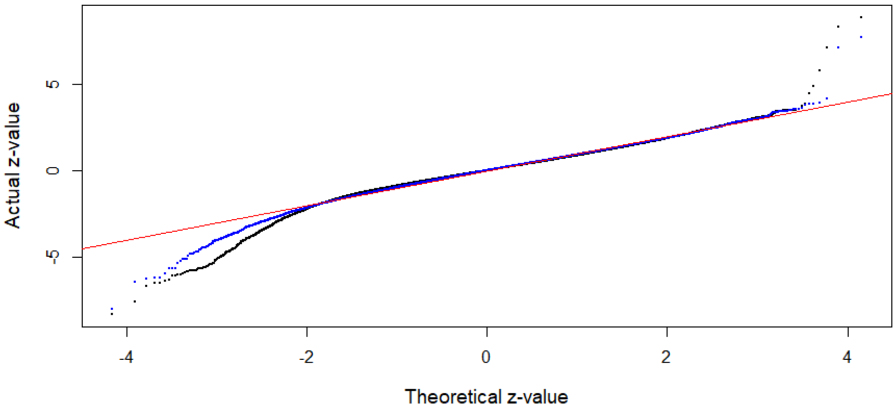

Figure 3: QQ plot for the reduced linear model (model A, black points) and the reduced 'lmvar' model (model B, blue points). The y-axis shows the values $z_{\text{actual}} = (y - \mu)/\sigma$ with μ the expected value and σ the standard deviation for the log of the claim amount, as estimated by the two models. The x-axis shows the z-value that belongs to the sample quantile of z_{actual}, assuming that z_{actual} follows a standard-normal distribution. The red line is the line $y = x$.

5.2 Linear regression model with a log-linear variance

5.2.1 Full linear model with log-linear variance

The full 'lmvar' model is a fit of the data set described in section 3 to an 'lmvar' model. The model matrices X_μ and X_σ are both equal to the model matrix described in section 4. The model has 31,161 observations and $2 \times 291 = 582$ degrees of freedom. Statistics for the model are shown in table 2.

Compared to the full linear model, the AIC has decreased significantly and the KS distance is about half as small. Both are indicative of a better fit. This is confirmed by the log-likelihood ratio between the two models. It is equal to 2535 for 290 additional degrees of freedom, giving a p-value of 0.

Surprisingly, table 2 shows that the MSE has increased sharply compared to the full linear model. The reason for this becomes apparent from fig. 5 which shows that there are a few observations (in particular the two observations in the lower left corner) which have a very large residual in the 'lmvar' model. These points are points with a large number of parts: the two points in the lower left corner have $N = 93$ and $N = 96$. The extreme values for the residuals go together with extreme values for the standard deviations of the claim amounts: the two points have $\sigma = 2.8 \times 10^5$ and $\sigma = 2.6 \times 10^6$ respectively. It can be seen from figs. 2 and 6 (the figures are for the reduced model but the situation is qualitatively the same for the full model) that both μ and σ for the log of

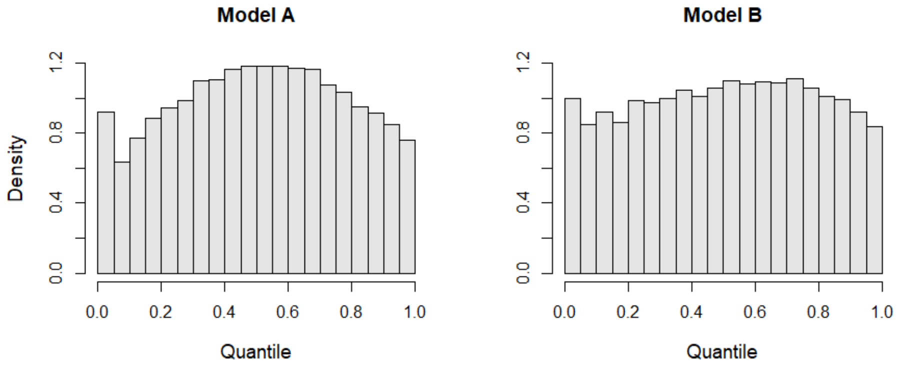

Figure 4: Distribution of the quantiles for the reduced linear model (model A) and the reduced 'lmvar' model (model B)

the claim amount tend to be large for large values of N. This can result in very large values for the expected value μ_{cl} and the standard deviation σ_{cl} of the claim amount itself (with $\mu_{\mathrm{cl}} = \exp(\mu + \sigma^2/2)$ and $\sigma_{\mathrm{cl}} = \mu_{\mathrm{cl}}\sqrt{\exp(\sigma) - 1}$). If we calculate the MSE without these two points, we obtain MSE $= 4.5 \times 10^5$, rather than the value 11.3×10^5 mentioned in table 2.

The cross-validations for the 'lmvar' model must be carried out with some care. It happens occasionally that the log-likelihood is maximized by bringing the standard deviation of one or a few observations to zero, while making the residuals for those observations equal to zero as well. Elements of β_σ will diverge to extreme values if that happens and the algorithm to find the maximum-likelihood estimators becomes unstable. To avoid these instabilities, we impose the condition $\sigma > 0.05$ when fitting the logarithms of the claim amounts to the 'lmvar' model in a cross-validation. The minimum value of σ when fitting the full 'lmvar' model is 0.16, hence our boundary value 0.05 is well below this. We checked that cross-validation results do not depend on the choice 0.05: doubling the minimum standard deviation to 0.1 does not change the results. Results are shown in table 2.

5.2.2 Model reduction

Like we did for the linear model, we weed out model terms that have little predictive power. We look again for the subset of model terms that result in the smallest AIC value. To do this efficiently, we first reduce the number of terms with a 'greedy' backward-step algorithm. The subset that is the result, is the start point for a simulated annealing run. Terms that were removed by the backward-step pre-selection, can be reintroduced by simulated annealing. Terms that survived the pre-selection, can still be removed by simulated annealing.

Three subsequent simulated annealing runs with different random numbers

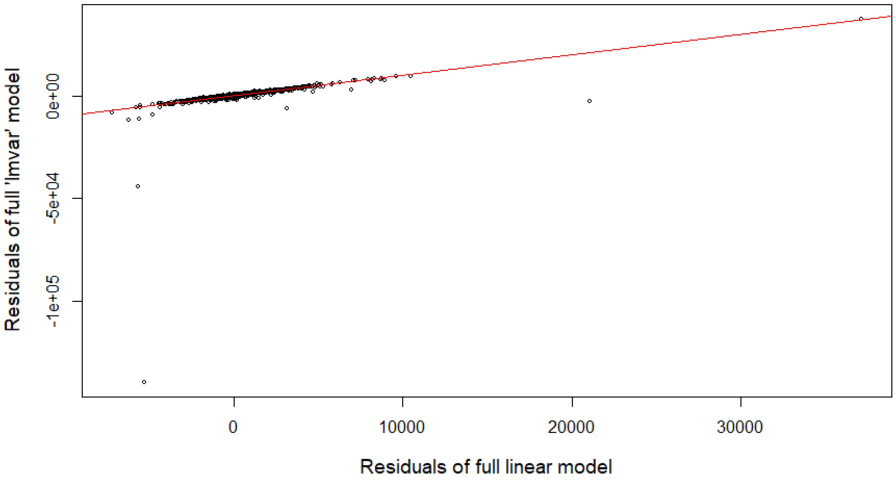

Figure 5: Scatterplot of the residuals of the full linear model versus the residuals of the full ’lmvar’ model. The red line is the line $y = x$. The two points in the lower left corner are mainly responsible for the difference in the mean squared error (MSE) between the two models.

result in similar subsets of model terms. E.g., the three design matrices X_μ have 139, 141 and 149 columns (including an intercept term) of which 121 columns are shared between all three. Likewise, the three matrices X_σ have 128, 128 and 129 columns of which 122 are shared. The degrees of freedom of the three resulting models vary between 267 $\sim$ 277, AIC values range between 35,827 $\sim$ 35,832 and 10-fold cross-validation MAE’s between 374.1 $\sim$ 374.8 with a standard deviation of 8 over the 10 folds.

5.2.3 Reduced linear model with log-linear variance

From the three results of the model reduction, we pick the model with the lowest AIC (35,827). It has 267 degrees of freedom. We call this model B. Its model for μ has 139 degrees of freedom, its model for $\log \sigma$ has 128 degrees of freedom. Other characteristics are shown in table 2. As it does for the full ’lmvar’ model, the tabel shows large values for the MSE and $\sqrt{\text{MSE}}$, compared to the full linear model or model A. The cause is the same as discussed for the full ’lmvar’ model. If we leave out the two observations in the lower-left corner of fig. 5, we obtain values that are comparable with both linear models.

Compared to the full model, we have removed nearly half of the degrees of freedom (267 versus 582). This reduction did not affect the cross-validation error estimate MAE or the results of the Kolmogorov-Smirnov test. The cross-validation estimates for MSE and $\sqrt{\text{MSE}}$ are smaller than for the full 'lmvar' model. However, since these estimates are strongly influenced by two observations only (both for the full 'lmvar' model and for model B), they are not so relevant.

The levels of the categorical covariates that survived the model reduction are shown in table 1, both for the model for μ as for the model for $\log \sigma$. All terms in (5) were kept in the model for μ. Only the terms m^2 and m^3 survived in the model for $\log \sigma$.

Of (6), all Kronecker delta's and the terms $h(N, \zeta_2)$, $h(N, \zeta_3)$ and $h(N, \zeta_5)$ survived in the model for μ. The terms N, $h(N, \zeta_1)$ and $h(N, \zeta_4)$ dropped out. Fig. 2 shows the contribution of the number of parts N to the expected value of the log of the claim amount. In the model for $\log \sigma$, all Kronecker delta's and the terms N, $h(N, \zeta_4)$ and $h(N, \zeta_5)$ survived. Fig. 6 shows the contribution of N to the log the standard deviation of the log of the claim amount. The figure shows a sharp cusp at $N = 10$, raising worries that (6) might not be able to capture the 'true' dependency of log σ on N. However, the p-values for the Kronecker-delta's decrease continuously from $p < 2.2 \times 10^{-16}$ for $\delta_{N,1}$ - $\delta_{N,6}$ to $p = 0.054$ for $\delta_{N,10}$. Hence the latter is already indistinguishable from zero at the 5% confidence level. The fact that $h(N, \zeta_1)$, $h(N, \zeta_2)$ and $h(N, \zeta_3)$ dropped out also indicates that the functional form of (6) is rich enough.

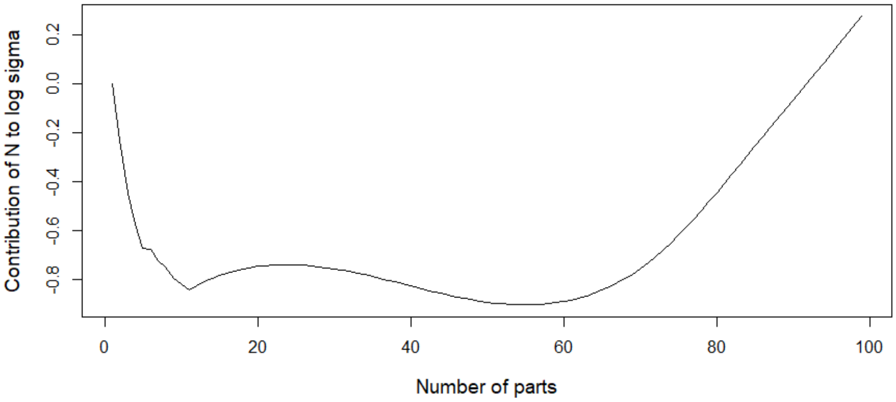

Figure 6: The contribution of the surviving terms in (6) in the reduced 'lmvar' model (model B) to the log of standard deviation of the log of the claim amount. The contribution has been set to zero for $N = 1$.

Fig. 3 shows the QQ-plot for model B. There are fewer outliers in the top-right corner as there are for model A. The two most extreme points are the points with claim amounts 38,012 and 37,843 that also appeared as extreme points for model A. The figure also shows that the point for model B in the bottom-left corner stay closer to the line $y = x$ than for model A.

Figure 4 shows the distribution of quantiles for both model A and model B. It underscores the superiority of the 'lmvar' model which shows a much more uniform distribution of quantiles.

‘

5.3 Comparison of expected values, standard deviations and quantiles

Fig. 5 shows there are a few observations with a very large difference between the expected values according to model A and B. Fig. 7 shows that the differences are small for all observations on the scale of the standard deviation of the claim amounts as estimated by model B. The relative differences run from -0.65 to 0.60. Typically, the difference between expected values is within 20% of the fitted standard deviation.

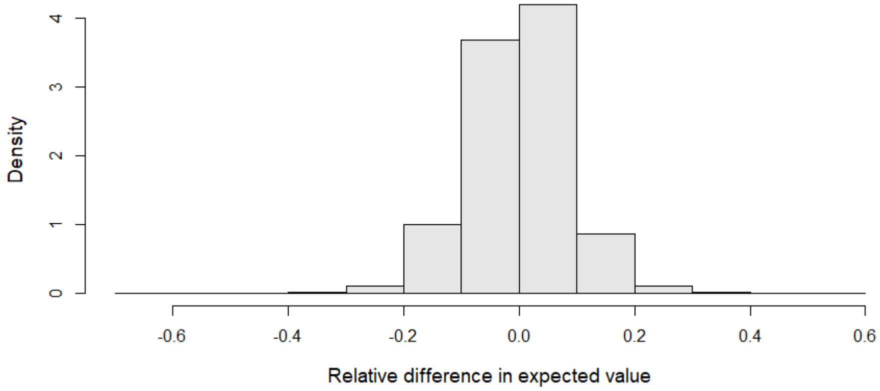

Figure 7: Histogram of the relative difference of the expected values for the claim amounts. The relative difference is defined as $(\mu_{\mathrm{lmvar}} - \mu_{\mathrm{lm}})/\sigma_{\mathrm{lmvar}}$ with μ_{lmvar} the expected value according to model B, μ_{lm} the expected value according to model A and σ_{lmvar} the standard deviation according to model B. Expected values and standard deviations are for the claim amounts themselves, not for the logarithm of the claim amount.

Fig. 8 compares the standard deviations as estimated by model A and model B. The figure at the left shows that model B predicts a large spread in the standard deviation of the logarithm of the claim amount, which can not be captured by model A. The figure at the right shows that model A tends to give larger estimates of the standard deviation of the claim amount for N in the range $30 \sim 80$. For larger values of N, the standard deviation blows up in

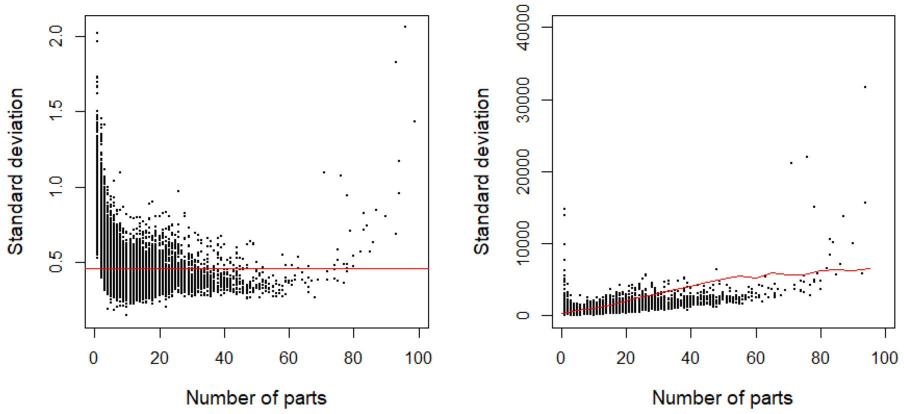

Figure 8: Left figure: scatterplot of the estimated standard deviation of the logarithm of the claim amount, according to the reduced 'lmvar' model (model B). The red line is the estimate according to the reduced linear model (model A). Right figure: scatterplot of the estimated standard deviation of the claim amount, according model B. The red line is the average standard deviation according to model A. The y-scale has been limited to 40,000. This leaves out the points (93, 2.4×10^5), (96, 5.3×10^5) and (99, 0.6×10^5)

.

model B.

Fig. 9 shows the quantile of an observation according to model A versus the quantile according to model B. Globally, both models predict roughly the same quantile, as evidenced by the fact that the band of data-points falls along the line $y = x$. There is a small curvature in the band: observations with a quantile less than 0.5 according to model B, tend to have a quantile larger than 0.5 according to model A. The situation is reversed in the other half of the figure.

An insurer might want to investigate all claims with a claim amount that falls above to 80% quantile. These are the claims above the horizontal dashed line according to model A, or to the right of the vertical dashed line according to model B. Table 3 shows that the would have to investigate 5,417 claims according to model A and 5,857 claims according to model B. These numbers do not differ that much but only 4,702 claims are shared between them.

6 Simulation study

6.1 Identification of quantiles

All of our analyses of section 5 indicate that a model that takes a structure for the variance into account (the 'lmvar' model) is a more suitable model for our data set of motor claims than a model with constant variance (the classical linear model). If we assume the 'lmvar' model to be the true model, how far off are we when we fit with a classical linear model? We study this in an simulation in which we assume model B to be the true model and generate values for the logarithms of the claim amount according to model B. We calculate the quantile (according to model B) of each generated observation and take these to be the 'true' quantiles.

We fit the data set we thus obtain to a classical linear model (with the matrix X_μ of model B as model-matrix) on the one hand, and to an 'lmvar' model (with the matrices X_μ and X_σ of model B as model-matrices) on the other hand. We calculate the quantiles of the generated observations according to the 'lmvar' fit and the classical linear fit. Table 4 shows to which extent the fits are able to identify the 'true' quantiles larger and smaller than 80%.

The table shows that linear model misses 21% (1,313 of the 6,216) of the observations with a quantile larger than 80% whereas this is only 7% (412 of the 6,216) for the 'lmvar' model. The linear model performs slightly worse when looking at the number of observations that is incorrectly classified as having a quantile $\leq$ 80% (573 observations for the linear model versus 464 for the 'lmvar' model). However, the absolute difference is small and table 4 is for one generated response vector only.

6.2 Variance-based bias

Assuming again model B to be the 'true' model, there is a subtle bias in the observations which have a large quantile according to the linear model. Observations with a large 'true' variance are over-represented. This can be understood as follows. If the true variance is small, the linear fit will overestimate it. The linear fit will therefore over-estimate the quantile if the 'true' quantile is smaller than 50%, and under-estimate the quantile if the true quantile is above 50%. Hence if we look at all observations with a small true variance and select the ones with a quantile larger than a minimum quantile, the linear fit will select too many if the minimum quantile is below 50% and too few otherwise. In other words: the linear fit reports an excess of false positives when the minimum quantile is below 50% and an excess of false negatives otherwise. This assumes that the median according to the linear fit is on the average close to the 'true' median, which is indeed the case as we have seen in section 5.3.

The situation is reversed if we look at observations with a large true variance. The linear fit reports an excess of false negatives when the minimum quantile is below 50% and false positives otherwise.

The effects are shown in figure 10. For the left-hand figure, we look at the subset of all observation which true standard deviation falls in the smallest 20%. The figure shows which fraction of this subset is selected, when we select observations with an estimated quantile larger than a minimum quantile. The black curve is the result when the quantiles are estimated by a linear fit, the blue curve when the quantiles are estimated by an 'lmvar' fit.

Both curves are an average over 50 simulations. In each simulation we generate a response vector according to the true model. We then fit a linear model (with the matrix X_μ of model B as model-matrix) and an 'lmvar' model (with the matrices X_μ and X_σ of model B as model-matrices) and calculate the quantile of each observation according to both fits.

For the figure on the right-hand side, we look at the subset of observations with a true standard deviation in the largest 20%. Otherwise, the figure is the same as on the left-hand side.

The two blue curves are approximately the straight lines $y = 100 - x$. This is because the true model is taken to be an 'lmvar' model and the distribution of quantiles according to an 'lmvar' fit is therefore uniform for both subsets of observations. The black curves show the bias of the linear fit discussed above.

7 Conclusions

We want to identify motor claims with a large claim amount. A claim amount is considered large if its quantile is large. We fitted two distributions to calculate the quantile: a classical linear fit and a 'lmvar' fit. The latter is a generalization of the former in which different claims can have different variances.

All statistical measures are in favor of the 'lmvar' fit: the difference in deviance between the two fits has a p-value equal to 0, the AIC and the KS distance are smaller, the QQ-plot shows that the 'lmvar' fit fits the tails of the distribution of the logarithms of the claims better and the distribution of quantiles is more uniform for the 'lmvar' fit.

Error measures such as the MAE and MSE are the same for both types of fits if we disregard two observations with very large residuals and standard deviations in the 'lmvar' fit. A direct comparison of expected values according to the two models shows that the difference is small on the scale of the standard deviation according to the 'lmvar' fit. The primary difference between the models is therefore in the estimate of the variance, with a less important effect on the expected values. This is confirmed by a direct comparison of the variances according to both models. There is a large spread in the variances of the logarithm of the claim amounts according to 'lmvar' that can not be captured by the linear model.

We took the 'lmvar' model to be the 'true' model and carried out a simulation study in which we generated claim amounts according to the 'true' model. We showed that if we try to identify all observations with a quantile $> 80\%$ using a linear fit, we miss about 21% of those observations. About 2% of all observations with a quantile $\leq 80\%$ are incorrectly classified as having a quantile $> 80\%$. These numbers are 7% and 2% for the 'lmvar model.

We showed there is bias in the claims selected by a linear fit, assuming the 'lmvar' model to be the true model. When claims with a large quantile are selected based on a linear fit, claims whose 'true' variance is large are over-represented.

An alternative approach besides a linear or 'lmvar' fit is an estimate of the quantiles by a quantile regression [Koenker and Basset, 1978]. It does not require the choice of a distribution but a separate regression has to be carried out for every quantile. We have not pursued this approach in our current work.

8 Acknowledgments

The appraisal of claim amounts by a statistical method such as described here, has been implemented in the Claims Management Filter of Posthuma Partners (www.posthuma-partners.com). We thank Marc Dijkstra of Posthuma Partners for his support. We thank Bouke Posthuma for a careful reading of the manuscript.

References

Emile Aarts and Jan Korst. *Simulated Annealing and Boltzmann Machines.* Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, 1988. ISBN 978-0-471-92146-2.

Murray Aitkin. Modelling Variance Heterogeneity in Normal Regression Using GLIM. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):332–339, 1987. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2347792.

H. Akaike. Information theory as an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory, Akademiai Kiado*, pages 267–281, 1973.

Gerhard Bohm and Gnter Zech. *Introduction to statistics and data analysis for physicists.* Verlag Deutsches Elektronen-Synchrotron, 2010. ISBN 978-3-935702-41-6. doi: 10.3204/DESY-BOOK/statistics(e-book). URL http://www-library.desy.de/elbook.html.

Samprit Chatterjee and Ali S. Hadi. *Regression Analysis by Example.* Wiley, 2006. ISBN 978-0-470-05545-8.

Piet de Jong and Gillian Z. Heller. *Generalized Linear Models for Insurance Data.* International Series on Actuarial Science. Cambridge University Press, 2008. ISBN 978-0-521-87914-9.

Harvey Goldstein. *Heteroscedasticity and Complex Variation.* John Wiley & Sons, Ltd, 2014. ISBN 9781118445112. doi: 10.1002/9781118445112.stat06249. URL http://dx.doi.org/10.1002/9781118445112.stat06249.

A. C. Harvey. Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica*, 44(3):461–465, 1976. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1913974`.

G. Z. Heller, M. D. Stasinopoulos, R. A. Rigby, and P. de Jong. Mean and dispersion modeling for policy claims costs. *Scandinavian Actuarial Journal*, (4):281–292, 2007.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning.* Springer Texts in Statistics. Springer, 2015. ISBN 978-1-4614-7138-7. doi: 10.1007/978-1-4614-7138-7. URL `http://www-bcf.usc.edu/~gareth/ISL/`.

R. Koenker and G. Basset. Regression quantiles. *Econometrica*, (46):33–50, 1978.

Posthuma Partners. *lmvar: Linear Regression with Non-Constant Variances.* Posthuma Partners, Gouda, Netherlands, 2017. URL `https://CRAN.R-project.org/package=lmvar`.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

G.A.F. Seber. *Linear Regression Analysis.* Wiley series in probability and mathematical statistics. John Wiley & Sons, 1977. ISBN 0-471-01967-4.

Robert Tibshirani. regression shrinkage and selection via the lasso. (58):267–288, 1996.

George Tzougas, Spyridon Vrontos, and Nicholas Frangos. Risk classification for claim counts and losses using regression models for location, scale and shape. *Variance*, 9(1):140–157, 2015. URL `http://www.variancejournal.org/issues/09-01/140.pdf`.

A. P. Verbyla. Modelling Variance Heterogeneity: Residual Maximum Likelihood and Diagnostics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(2):493–508, 1993. ISSN 00359246. URL `http://www.jstor.org/stable/2346209`.

		Number of levels				
				model		
				A	B	
Covariate	Description	original	remaining		μ	$\log \sigma$
Make	The brand of the car	38	29	20	18	12
Model	A combination of make and model of the car	830	161	49	55	55
Point of impact	The point where the car has been impacted	11	10	8	8	6
Direction of impact	The direction in which the impact took place	13	13	9	11	8
Cause of damage	The cause of the damage	32	16	13	11	14
Car segment	A car classification like 'small', 'station', 'hatchback' etc.	14	12	8	5	7
Year of construction	Year of construction of the car	33	17	5	5	4
Fuel type	The type of fuel	5	5	5	5	3
Owner status code	Status of the car owner	3	3	1	1	1
Type of activity 1	Type of activity carried out by repair shop	5	2	1	1	1
Type of activity 2	Type of activity carried out by repair shop	11	3	2	2	1
	Total	995	271	121	122	112

Table 1: categorical covariates present in the data set. The 'number of levels' are the number of distinct values that a covariate takes. The column 'original' shows the number of levels in the full data set. The column 'remaining' shows this number after removing the 7 observations with a number of parts larger than 100 and subsequent removal of levels that appear 50 times or less.

	Full linear model	Model A	Full 'lmvar' model	Model B
	All observations			
AIC	40,749	40,529	36,259	35,827
MAE	374	374	374	373
MSE	4.7×10^5	4.7×10^5	11.3×10^5	6.5×10^5
$\sqrt{\text{MSE}}$	683	684	1065	805
KS distance	0.0433	0.0425	0.0214	0.0212
p-value	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	9.1×10^{-13}	1.4×10^{-12}
	Cross-validation			
MAE	379	377	377	374
sd	13	13	9	8
MSE	5.0×10^5	4.7×10^5	7.9×10^5	6.4×10^5
sd	1.9×10^5	1.9×10^5	6.9×10^5	3.7×10^5
$\sqrt{\text{MSE}}$	694	677	838	775
sd	130	130	319	213
KS distance	0.0452	0.0450	0.0251	0.0258
sd	0.0065	0.0068	0.0058	0.0057
p-value	2.7×10^{-4}	2.4×10^{-4}	0.10	0.079
sd	7.5×10^{-4}	6.1×10^{-4}	0.12	0.093

Table 2: Various statistics for the different models. They are the AIC, the mean absolute error (MAE), the mean squared error (MSE), the square root of the means squared error ($\sqrt{\text{MSE}}$), the Kolmogorov-Smirnov (KS) distance and the p-value of the KS-distance. The MAE, MSE and $\sqrt{\text{MSE}}$ are for the claim amount. The KS-distance and its p-value are for the fit to the logarithm of the claim amount. Results are shown for the full set of observations and for a 10-fold cross-validation. In the latter case, the tabulated values are the sample means over the 10 folds. The corresponding sample standard deviations (sd) are also listed.

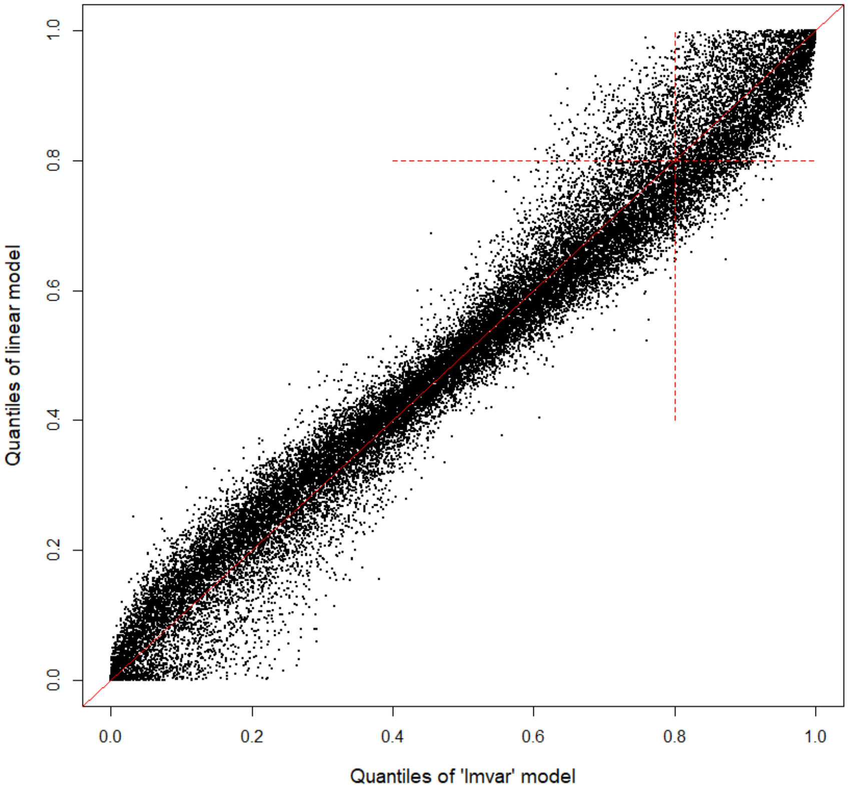

Figure 9: Scatterplot of the quantile of each observation according to the reduced linear model (model A) versus the quantile according to the reduced 'lmvar' model (model B). The solid red line is the line $y = x$. The dashed red lines show the 80% quantiles that appear in table 3.

		According to model B		
		$\leq 80\%$	$> 80\%$	
According to model A	$\leq 80\%$	24,589	1,155	25,744
	$> 80\%$	715	4,702	5,417
		25,304	5,857	31,161

Table 3: Table showing the number of claims with a claim amount with a quantile $\leq 80\%$ and $> 80\%$, according to the reduced linear model (model A) and the reduced 'lmvar' model (model B)

		True quantiles		
		≤ 80%	> 80%	
≤ 80%	’linear’	24,372	1,313	25,685
> 80%		573	4,903	5,476
		24,945	6,216	31,161

	True quantiles		
	≤ 80%	> 80%	
’lmvar’	24,481	412	24,983
	464	5,804	6,268
	24,945	6,216	31,161

Table 4: Table showing the number of claims with a ’true’ quantile ≤ 80% and > 80%, and these numbers according to a linear fit (at the left), and according to an ’lmvar’ fit (at the right).

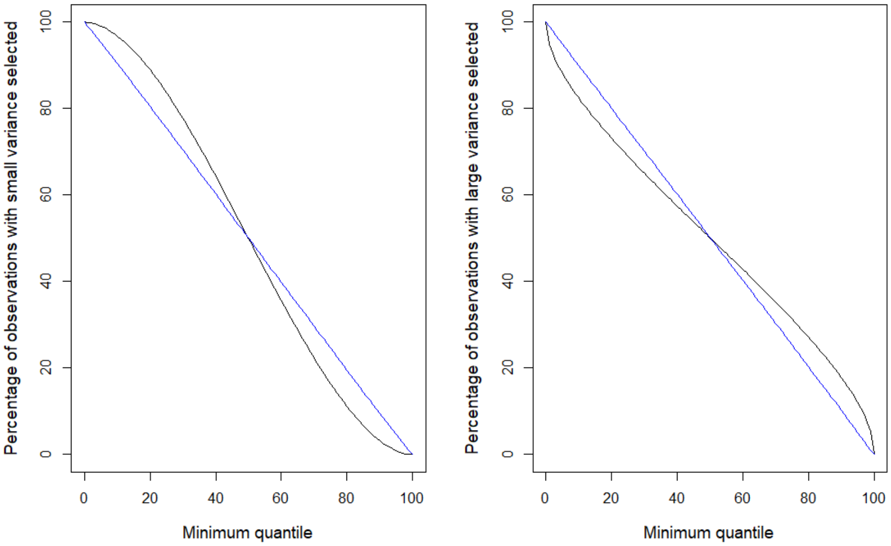

Figure 10: Percentage of a subset observations that is selected when selecting observations with an estimated quantile larger than a minimum value. The black line is for an estimate from a classical linear model, the blue line for an ’lmvar’ model. The figure on the left is the subset of observations with a standard deviation (according to model B) within the smallest 20%. The figure on the right is the subset with a standard deviation within the largest 20%. The curves are an estimate over 50 simulations.