Stockholms universitet

Reassessing the Duration Effects in Non-Life Insurance Pricing

Nils Oscar Potts

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2025:3
Försäkringsmatematik
Februari 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2025:3**
http://www.math.su.se

Reassessing the Duration Effects in Non-Life Insurance Pricing

Nils Oscar Potts*

February 2025

Abstract

Assuming that the observed duration of a non-life insurance policy is linear towards the underlying risk has been shown to induce spurious over-dispersion. This over-dispersion can distort pricing models by introducing variance where there is none and create inaccurate insurance premiums. This paper shows through simulation that said over-dispersion may arise from *detrimental claims* which are claims that cancel the need for coverage and, in effect, terminate the policy prematurely. Further, through parametric assumptions on duration, a method of adjusting for detrimental claims is proposed. This method shows to remove over-dispersion and maintain accurate model estimates in a simulated setting with improved performance in an applied setting.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: oscar.n.potts@gmail.com. Supervisor: Mathias Lindholm.

Contents

Abstract

Assuming that the observed duration of a non-life insurance policy is linear towards the underlying risk has been shown to induce spurious over-dispersion, see [5]. This over-dispersion can distort pricing models by introducing variance where there is none and create inaccurate insurance premiums. This paper shows through simulation that said over-dispersion may arise from *detrimental claims* which are claims that cancel the need for coverage and, in effect, terminate the policy prematurely. Further, through parametric assumptions on duration, a method of adjusting for detrimental claims is proposed. This method shows to remove over-dispersion and maintain accurate model estimates in a simulated setting with improved performance in an applied setting.

Keywords: Duration Effects, Non-Life Insurance, Actuarial Pricing, Actuarial Sciences

1 Introduction

In non-life insurance pricing, one seeks to predict the potential economic compensation of a specific event over an agreed duration of coverage. The purpose of this is to provide security and help individuals manage risk. Usually, these contracts are signed on a year-by-year basis of coverage, which means one would expect a duration of 1 year per contract. In reality, things happen, and the consequences of such things are that the observed duration does not always equal 1 year.

Traditionally we model the pricing of these insurances by assuming that half a year premium corresponds to half a year of claims. As such one should pay half the total annual premium for half a year of coverage. An example would be selling your car or a change of employer which would remove the need for coverage and cancel the policy early [1].

This would be natural reasons for cancellation, and we should expect that the half a year of coverage reflects half a year of claims. But sometimes a policy may be canceled for risk-related reasons. For instance, given a fire insurance with the house burning down after 3 months, this would remove the need for further coverage and cancel the insurance. In this scenario, we should not expect the 3 months of coverage to correspond to 3 months of claims, and expect 4 fires to the same building a year.

This paper seeks to study this type of cancellations, seeing how it affects price modeling in non-life insurances, and what adjustments one should make to correctly account for the risk of *detrimental claims*.

2 Theory

We represent an insurance contract by the triplet (X, Z, W) where Z denotes the claim amounts, X the covariate vector, and W the duration of our contract insurance in years. Contracts tend to be written to provide 1 year of coverage but may be canceled, which leads to $W \in (0, 1]$. Longer coverage periods may occur, but that is not a concern in this thesis.

When the contract is written, an insurance provider provides an actuarial one-year premium $\pi(X)$, which is the expected compensation given X for 1 year of coverage. We call the actuarial premium *fair*, if

$$\mathrm{E}[W\pi(X) \mid X] = \mathrm{E}[Z \mid X]. \tag{1}$$

Fundamentally, this states that the expected amounts of premiums should equal the expected amount of claims. Observe that associated costs such as claim handling or profit margins are not included, which is why it is named as the actuarially fair premium.

As $\pi(X)$ is known given X and since W is unknown when the policy is written we can revise (1) as

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

which defines the actuarially fair premium as a division between 2 expected values given X.

2.1 Probability measure

As the goal is to model insurance premiums to estimate a correct pure premium in (2), dividing two expected values is a bit bothersome. Preferably, one would want a single expected value to estimate, and this can be done by introducing a probability measure, see [6], Let P_W be a duration-weighted probability measure which is defined as

$$\mathrm{E}_{P_W}[A] = \mathrm{E}\left[\frac{W}{\mathrm{E}[W]}A\right], \tag{3}$$

where A is an arbitrary random variable and the random variable W is positive with expected value 1, which it is with $W \in (0, 1]$, see [11].

We now introduce $Y = Z/W$ and rewrite equation (2) as

$$\pi(X) = \frac{\mathrm{E}[WY \mid X]}{\mathrm{E}[W \mid X]} = \mathrm{E}\left[\frac{W}{\mathrm{E}[W \mid X]}Y \mid X\right] \tag{4}$$

[1] There's often regulation and contract specification in how policies are allowed to be canceled.

Which if we compare to equation (3) yields a single expectation formulation of the pure premium

$$\pi(X) = E_{P_W}\left[\frac{Z}{W} \mid X\right]. \tag{5}$$

We will in this thesis use this formulation and define a probability measure P_W for W which is to account for the over-dispersion observed in [5].

2.2 Estimating the Fair Premium

The modeling task in actuarial pricing is to accurately estimate the correct fair premium $\pi(X)$ based on historical data. For historical data $W_i, i = 1, 2, .., n$ are known and in a perfect world, all contracts would have a duration of 1, but as things occur, for one reason or another, the historical data has policies with different periods of coverage. To compensate for this dilemma, one assumes *linearity* or *constant intensity* of the underlying claim generation process, see [11]. This introduces an intensity $\mu(X)$ and variance $\sigma(X)$ such that

$$\begin{aligned} \mathrm{E}[Z \mid X, W] &= W\mu(X), \\ \mathrm{Var}[Z \mid X, W] &= W\sigma(X). \end{aligned} \tag{6}$$

Using this assumption, one uses historical data consisting of triplets $(Z_i, X_i, W_i), i = 1, 2, .., n$ to estimate $\mu(X)$ with $\widehat{\mu}(X)$ and then writes annual premiums as $\pi(X) = 1 \cdot \widehat{\mu}(X)$. Observe here that the linearity assumption allows one to write contracts for arbitrary periods of time.

2.3 Severity and Frequency Models

In the above example, we concern ourselves with the ultimate claim cost Z, which in fact consists of two components: the number of claims and the cost of each separate claim. It is general practice, see [7], to model the number of claims and the claim cost for each claim separately. One thereby rewrites μ as a product of two functions:

$$\mu(X) = \lambda(X)S(X). \tag{7}$$

Where λ is the expected number of claims per year and S claim cost per claim given X. Observe that it is only for the claim intensity λ we need to assume (6) as S is only dependent on claims being observed.

In this thesis, we will concern ourselves with frequency modeling, as this is where the linearity assumption of constant intensity is needed and where the problems noted in [4] are relevant. As such Z will denote the amount of claims moving forward in this thesis.

2.4 Generalized Linear Models

Generalized linear models (GLM), see [7], are a generalization of ordinary linear regression that allows for response variables to have distributions other than a conditional normal distribution. A GLM assumes a parameter vector β, which is used to define a linear predictor $\eta_i = \beta^T X_i = \sum_{j=1}^{p} x_{ij}\beta_j$ and link function g such that

$$E[Z_i \mid X_i] = g^{-1}(\eta_i) = g^{-1}(\beta^T X_i). \tag{8}$$

In addition, we assume that our observations Z_i are independent and distributed according to a distribution of the exponential family, see [11]. In our analysis, with Z_i denoting the number of claims, we will assume that Z_i follows a Poisson distribution, and by using the log link function $g(\eta) = \log(\eta)$, we assume $Z_i \sim Poi(\mu(X_i))$, where $\mu(X_i) = g^{-1}(\beta^T X_i) = \exp(\beta^T X_i)$.

In training a GLM, one seeks to estimate the β through estimators $\widehat{\beta}$, which is done numerically by minimizing a Kullback-Leibler divergence, see [6]. For the Poisson distribution, the Kullback-Leibler results in the unit deviance function.

$$d(y, m) = \sum_{i=1}^{n} 2(y_i \ln(y_i) - \ln(m_i) - y_i + m_i) \tag{9}$$

where m is a vector of fitted values $m_i = \widehat{\mu}(X_i) = \exp(\widehat{\beta}^T X_i), i = 1, 2, .., n$.

In regards to duration and assuming equation (6) we introduce the duration W as an *offset* variable, which implies that we fixate a linear covariate $\beta_W = 1$ towards a covariate $\log(W)$ which defines our estimator as

$$\mu(X_i) = e^{\beta_W log(W_i) + \beta^T X_i} = We^{\beta^T X_i}, \tag{10}$$

which aligns our assumptions in equation (6).

We will in this thesis use GLM:s to analyse how detrimental claims affect model estimation. Specifically, we will simulate data where all the assumptions of the GLM are true, introduce the concept of detrimental claims, and measure how this affects our estimates $\widehat{\beta}$ and estimation of dispersion, which will be explained in further detail in section 2.6.

2.5 Gradient Boosting Machines

A modern alternative to generalized linear models is gradient boosting machines (GBM), see [2]. A gradient boosting machine does, instead of assuming a linear prediction η, fit a predictor f defined as a sum of regression trees as $f(X) = \sum_{i=1}^{M} h_i(X)$ such that

$$E[Z \mid X, W] = Wf(X_i). \tag{11}$$

This generalizes our expression from equation (8) to a more arbitrary curve fitting definition.

To train a GBM, we start with an initial model, often a constant, and iteratively add regression trees, see [2] to the model. Each tree is fit to the residuals of the previous model, aiming to correct the errors made by the previous trees. The process can be summarized as follows:

1. Initialize the model with a constant value: $f_0(x) = \arg\min_c \sum_{i=1}^{n} L(y_i, c)$, where L is the loss function.
2. For (m = 1) to (M) (number of trees):
   (a) Compute the residuals: $r_{im} = Z_i - f(X_i)$.
   (b) Fit a regression tree $h_m(x)$ to r_{im}.
   (c) Update the model: $f_m(x) = f_{m-1}(x) + \epsilon h_m(x)$ where ϵ denotes the learning rate.

Again for our intended purposes, we will assume that Z_i is Poisson distributed according to $Z_i \sim Poi(W_i f(X_i))$, and fit regression tree $h_m(X_i), m = 1, 2, .., M$ to minimize the Kullback-Leibler divergence, see [6] which means using the unit deviance as defined by equation (9). as a loss function L.

In fitting a Poisson GBM, one cares less about specific parametrization; one wants to find the best hyperparameters which provide the best estimator $\widehat{f}(x)$. For our intended purposes this concerns the shrinkage ϵ, tree depth b , and the number of trees M. The number of trees is often selected on a subset of the decisions tree which minimizes the generalization error, see [6]. In this thesis, this will be done by setting aside 20% of the data for validation and selecting the number of trees which minimize the loss on validation data as $M^V = \arg\min_m \sum_{i=1}^{n} d(Z_i, f_m(X_i))$. For the learning rate and tree depth, we will use $\epsilon = 0.1$ and $b = 2$ trees per tree regressor as used by [5].

In regards to offsetting duration for the GBM there is no parametric formulation to (6) as for the GLM. Instead, one supplies the offset by dividing the reference variable with the duration explicitly, creating a variable $Y_i = Z_i/W_i$ which is used as the reference variable.. This is done by the `gbm` package in R using the offset parameter.

We will in this thesis use GBM:s for modeling of real insurance data, this to avoid the tedious work of identifying correct linear parametrization of GLM:s, while still providing sufficiently good models and intensity estimates.

2.6 Dispersion

In assuming a Poisson distribution on the claim arrival process, we assume that the variance σ equals the estimated intensity $\sigma(X) = \mu(X)$. This assumption is not always true. Instead, one can generalize the variance function of equation (6) to assume that the variance is proportional to the intensity, see [7] as

$$\sigma(X) = \rho\mu(X) \tag{12}$$

where $\rho > 0$ defines the *dispersion* parameter. We say that we exhibit under-dispersion if we have $\rho < 1$ and over-dispersion if $\rho > 1$.

One can estimate the dispersion parameter through the Pearson dispersion statistic, see [11], as

$$\widehat{\rho} = \frac{1}{n} \sum_{i=1}^{n} \frac{W_i(Y_i - \widehat{\mu}(X_i))^2}{\widehat{\mu}(X_i)} \tag{13}$$

which will be used in this thesis, as in [5], to measure the dispersion. It is this estimate which indicates over-dispersion which may be observed when assuming (6), see [5].

This thesis will continue the work and confirm observations made in [5] but will also explore approaches to estimating equation (2). Specifically, it is deemed probable that the over-dispersion arises from the so-called *detrimental claims*, which are claims that in effect cancel the policy.

It is shown through simulations that the introduction of *detrimental claims* leads to over-dispersion where the underlying simulated claim arrival process has none. The paper then suggests an analytical approach for adjusting the observed duration to account for the probability that the policy is canceled due to a claim.

3 Background

The work presented in this thesis is an extension of [5] and the observations made in the paper. The observation made in [5] is unique in academia, and few papers note similar observations. There are studies where other measures than duration have been used as a measure of policy exposure, see [11], and discussion of the suitability of duration as an exposure variable. Duration as exposure is often the default, but few papers discuss the consequences of such an assumption.

There is previous work in understanding the duration of the insurance correctly, see [4]. In many cases, this research addresses other aspects than pricing, like churn rates as in [3], reserves in [8] or customer analytics as in [10]. In terms of modeling and methods used in the paper, these are based on standard practice modeling in non-life insurance, see [7, 11], which is explained in chapter 2.

3.1 The inaccuracies of linearity

As a basis for the results in this study, the aim is to confirm that the observation made in [5] is general across multiple datasets. As such, we will use the datasets available in the CAS datasets, see [1], which in addition to `freMTPL2freq`, contains 3 more P&C pricing datasets. In table 1 we highlight the complexity of these datasets where n denotes the number of datapoints, p the number of variables and hyperparameters used and defined in section 2.5. For more details regarding these datasets, see the documentation available at [1].

Dataset	n	p	ϵ	b	M	M^V
`ausPrivateAuto`	67 856	5	0.1	2	300	19
`brveh`	236 514	4	0.1	2	300	221
`freMTPLfreq`	413 169	7	0.1	2	300	157
`swmotor`	47 875	6	0.1	2	300	79

Table 1: Complexity of used datasets with the resulting hyperparameter used to fit corresponding GBM.

In [5] it is shown that by inference a local mean estimate of the duration, based on estimated risk-percentile, creates sensible estimates for the expected duration as in equation (2). More formally, the local mean is defined as

$$\widehat{W}_i^{L\&N} = \mathrm{E}[W \mid X = x_{(i)}] = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} W_{(j)} \tag{14}$$

for $i = k, k+1, .., n-k$ and where $W_{(j)}$ denotes the duration for observation j ordered according to risk percentile of $\widehat{\mu}$. The risk percentiles based on $\widehat{\mu}$ which is estimated using the `GBM` package in R using 5-fold cross-validation to fit an optimal model based on parameters. In the study they use $k = 2066$ which represents 0.5% of the `freMTPL2freq` dataset, which is the same 0.5% window we will use for our other datasets, specified in table 1.

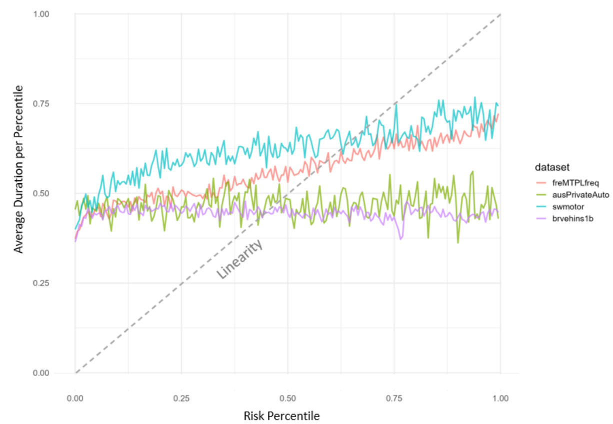


Figure 1: Rolling average duration using windows of 1% of total datapoints. This is across multiple dataset with percentiles ordered by to fitted $\widehat{\mu}(X)$ ascending values using duration as an offset. The results show a non-linear nature of the observed duration as in [5].

Seen in Figure 1 we have the fitted duration curves according to the local mean estimation across multiple datasets. These are the same duration estimates that would be provided by equation (14) if used on new data. If duration is linear, one would expect the highest risk policies to have duration near 1, while the lowest risk policies to have low duration near 0. This is not the case, as observed in 1, where `swmotor` and `freMTPLfreq` show linearity between 0.7 and 0.4; meanwhile, `brvehins1b` and `ausPrivateAuto` show no relation between the risk percentile and duration.

To confirm the observations made by these plots, we will also conduct a student t-test, see [11], to confirm the nonlinear structure of the observed duration towards the underlying risk. Using the estimated pure premium $\widehat{\mu}(X)$ and conducting a Poisson regression towards claims with pure premiums as offset and duration as the dependent variable, we can estimate the linearity of duration towards claims as

$$Z = e^{\beta_W log(W)} \widehat{\mu}(X) \tag{15}$$

where β_W is the parameter we are estimating and $\widehat{\mu}$ is our already estimated intensity assuming (6). One could see this regression to relaxing assumption in equation (6) and the offset requirement on β_W to estimate how duration relates to risk, seeing it as a covariate. If we estimate $\beta_W = 1$ our observed duration is linear towards the risk and risk-independent if $\beta_W = 0$.

As observed in Table 2 we have no dataset with an insignificant difference from 1, meaning that even when assuming linearity for the pure premium, the effect observed after fitting is not linear.

Dataset	$\beta_W \pm 95\%$
`ausPrivateAuto`	0.73 ± 0.045
`brvehins1b`	0.67 ± 0.040
`freMTPLfreq`	0.59 ± 0.020
`swmotor`	0.56 ± 0.085

Table 2: Estimated Poisson GLM regression coefficient on equation (15) with 95% confidence interval.

3.2 Detrimental Claims

In the theory above, we see how the linearity assumption leads to over-dispersion across multiple datasets and that the linear assumption is not statistically significant across those datasets. These are pure observations with little reasoning for why this is the case.

A reasonable argument is that the over-dispersion arises from inflated claims, which occur for policies with claims unproportionate to the duration.

A hypothesis is that there is a probability of *detrimental claims* being claims severe enough that in effect cancel the policy and generate a low observed duration compared to the number of claims. To further this idea and to show how the introduction of detrimental claims confirms the empirical observations made by this thesis, we will show that through the introduction of detrimental claims, one observes over-dispersion where there is none.

Suppose that the underlying claim generation process is Poisson distributed with 5 risk groups, with intensity $\lambda_i = 0.02i = \exp\beta_i, i = 1,2,3,4,5$. We will simulate claim arrivals across these groups across the *natural duration* of a policy $W_i^n, i = 1,2,..,n$ which is conditionally uniformly distributed based on a Bernoulli probability

$$W_i^n \sim \begin{cases} 1, & (1-p_n) \\ \text{Unif}(0,1), & p_n \end{cases} \quad i = 1,2,..,n \tag{16}$$

where we set the probability of cancellation $p_n = 0.2$ to be similar to the observations made in the study of previous datasets. Equation (16) is used to describe the natural dynamics of duration, which for many reasons may be canceled for reasons not related to risk.

We now introduce the concept of detrimental claims, which for a given claim from the above claim arrival process have a probability p_d of being detrimental. This implies that given the underlying natural assumption of linearity, when a claim arrives, there is a probability that it censors the natural duration, yielding a smaller observed duration $W_i < W_i^n$ and also unobserved claims $Z_i \leqslant Z_i^n$ that could have occured after a detrimental claim.

The simulation process is described in algorithm 1 where we will use n =100 000, $p_n = 0.2$ and vary p from 0 to 1 in this thesis. We will then fit a GLM on the data $\{X_i, W_i, Z_i\}, i = 1,2,..,n$ and compare the dispersion estimate ρ according to equation (13) and intensity estimates $\widehat{\lambda}_j = \exp(\widehat{\beta}_j), j = 1,2,..,5$ to the same GLM estimated on the natural data $\{X_i, W_i^n, Z_i^n\}, i = 1,2,..,n$, where our GLM assumptions are true.

Algorithm 1 Simulation Process

Require: n: Number of policies,
  p_d: Probability of a Detrimental claim,
  p_n: Early cancellation probability.
Ensure: Simulation Process
for Policies $i = 1,2,\dots,n$ **do**
  Assign risk group $X_i = j \sim \text{Unif}(\{1,2,..,5\})$.
  Assign risk intensities $\lambda_i = \lambda_j$.
  Simulate natural durations
  $$W_i^n = \begin{cases} 1 & p = 1 - p_n \\ Unif(0,1) & p = p_n \end{cases}$$
  Generate claims $Z_i^n \sim \text{Poisson}(\lambda_i \cdot W_i^n)$.
end for
– Adjust data with detrimental claims
Set $W_i = W^n$, $Z_i = Z_i^n$, $i = 1,2,\dots,n$.
for each policy i with $Z_i^n > 0$ **do**
  Claim arrival $t_{ij} \sim \text{Unif}(0, W_i)$, $j = 1,\dots,Z_i^n$
  for each claim j **do**
    With Probability p_d:
      Set $W_i = t_{ij}$
      Set $Z_i = j$
  end for
end for
return $\{X_i, W_i, Z_i, W_i^n, Z_i^n\}, i = 1,2,..,n$

Seen in Figure 2 we see how the probability of detrimental claims induces over-dispersion. The observation made here is that a relatively low probability of detrimental claims will lead to quite significant over-dispersion. This aligns with the empirical observations, and the results imply that if one were to account for detrimental claims, we would be able to account for the over-dispersion. Do note that the graph implies some linear relationship between the dispersion and probability of detrimental claim, a relationship which became steeper as one increased n.

In terms of how detrimental claims impact intensity estimates, we see in Figure 3 that as p_d increases, there is no significant change to mean intensity estimates, but an increase in the variance of our estimates. This increased variance makes it impossible to distinguish between risk groups for high p_d. This is the case of using the over-dispersion in modeling through a Quasi-Poisson GLM, if one would assume unit dispersion $\rho = 1$ this variance is not observed.

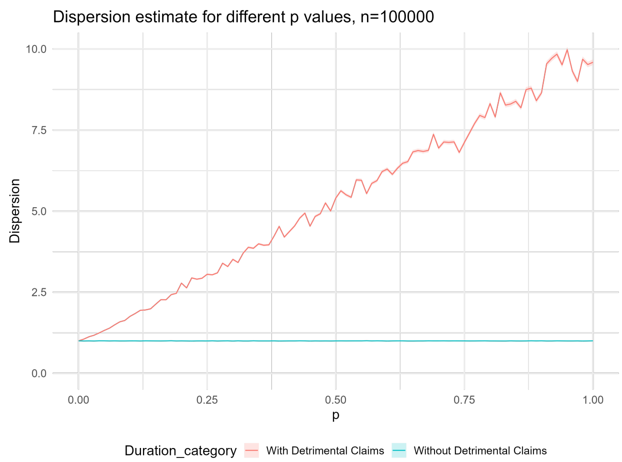


Figure 2: Observed median dispersion estimate $\widehat{\rho}$ using equation (13) with GLM estimate $\widehat{\mu}$ on data with and without detrimental claims and the estimated 95% Chi-squared confidence intervals across 100 experiments per probability $p_d = p$ of a claim being detrimental and sample size n =100 000.

4 Methodology

As observed in the previous chapter, one can use the concept of detrimental claims to explain the over-dispersion observed in [5]. One could thereby imagine that if we adjust our modeling to take the risk of detrimental claims into consideration, we may avoid induced over-dispersion. The problem with this is how does one identify a detrimental claim and what is the expected true duration given that the policy was canceled through a detrimental claim.

In practice, one could identify detrimental claims through seeing if the cancellation is in connection with the claim being reported. However, knowing this, what is the suitable adjustment needed for detrimental claims? Given a detrimental claim, one would want to adjust the duration towards what the duration would have been if the claim was not detrimental. To make the matter even more complicated, given that we do not know if a claim is detrimental, what adjustments should we make to correctly account for detrimental claims?

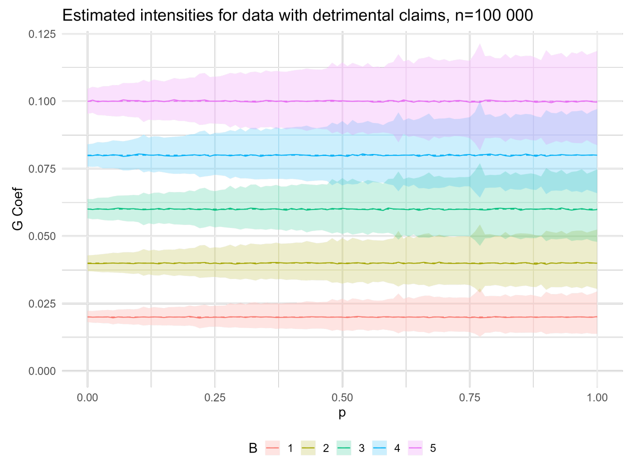


Figure 3: Observed mean intensity estimates of $\widehat{\lambda}_i = \exp \widehat{\beta}_i$ and 95% confidence interval estimates across the 5 risk groups when estimating intensity through a Quasi-Poisson GLM on 100 000 simulated policies without and with detrimental claim probability p a 100 times.

We will now suggest an estimation technique to estimate the unobserved natural duration $W^{adj} = \mathrm{E}[W^n \mid W, Z]$ given a set of observed duration W and potential detrimental claims Z. In this, we will assume linearity and equation (2) of the underlying claim generation process, but now we will assume that the observed duration W may be a censored observation of the natural duration W^n.

To formulate this assumption we reformulate (6) as

$$\begin{aligned} \mathrm{E}[Z \mid X, W] &= W^n \mu(X), \\ \mathrm{Var}[Z \mid X, W] &= W^n \sigma(X). \end{aligned} \tag{17}$$

Observe that the model assumptions are the same as in (6), where we assume linearity towards the insured period, but now W^n is an unknown variable, similar to equation (2).

To adjust to W^n being unknown we will estimate W^n based on the expected natural duration given W and Z as $W^{adj} = \mathrm{E}[W^n \mid W, Z]$. The resulting estimate is similar to using equation (14), see [5] but here we will provide a single parametric form which does not require you to estimate $\hat{\mu}$ and calculate $\widehat{W}^{L\&N}$ retroactively. Do observe that compared to [5] the suggested method can not be used to predict the duration as per equation (6), but to adjust the observed duration in training.

In deriving the parametric estimate of the natural duration W^{adj}, the crucial observation is that we are given true observations of the natural duration when there are no claims. Given a policy with 0 claims, we have pure observations on the natural level of cancellations, and given a policy with $Z = 1$ claims, we are $q_d = 1 - p_d$ certain to observe it as well.

We will begin by deriving the probability of a policy being canceled by a detrimental claim, and then, through parametric assumptions on duration, what the expected duration we would expect given a detrimental claim. Underlying these derivations will be the following definition of a detrimental claim:

Definition 4.1. *Given a natural duration W^n, claims Z^n and observed duration W and claims Z as defined in chapter 3.1, an insurance claim $i \in \{1, 2, .., Z^n\}$ is defined as detrimental if $W < W^n$ and $Z = i$.*

4.1 Probability of Detriment

For a policy with $Z : Z > 1$ claims, we know that the first $Z - 1$ claim was not detrimental, and our probability of a policy being canceled by a detrimental claim is independent of the amount of claims. We can formalize this through the following lemma:

Lemma 4.1. *Given a policy with claims $Z > 0$ and assuming that all claims $i = 1, 2, .., Z$ are independent, we have the probability of the policy being canceled through a detrimental claim, as defined in 4.1, independent of the amount of claims Z and equal p_d.*

Proof. Let d_i be a boolean random variable being 1 if a policy was canceled by claim number i and 0 if not. For a policy with $Z = 1$ we have the probability of the policy being canceled due to a detrimental claim as

$$P(d_1 = 1) = p_d. \tag{18}$$

For a policy with $Z = i$ claims we have the probability of claim i being detrimental given that

$$P(d_i = 1 \mid d_{i-1} = 0) = \frac{P(d_i = 1)}{P(d_{i-1} = 0)} = \frac{q_d^{i-1} p_d}{q_d^{i-1}} = p_d \tag{19}$$

Through induction we can thereby deduce that the probability of a policy being detrimental is independent of the amount of claims Z and equal p_d.

□

Lemma 4.1 states that all policies with $Z > 0$ have a probability p_d of being canceled by a detrimental claim, assuming that our observations are independent. We can also state that for policies with $Z = 0$ or $W = 1$, it is not possible to observe a detrimental claim, meaning the probability is 0 for these policies. As such, we will now derive the probability of detriment, or the probability of a policy being canceled through the following lemma:

Lemma 4.2. *Given an observed duration W, claims Z and by assuming independent claims with a single probability for detriment p_d according to lemma 4.1 we have the probability of detriment, defined as a policy being canceled as a consequence of a detrimental claim, as*

$$p_D = 1 - q_D = \begin{cases} 0, W = 1 \ \mathit{or} \ Z = 0 \\ p_d \left(1 - \int_0^W f_w(w)dw\right), \mathit{else} \end{cases}. \tag{20}$$

where f_w is the pdf of the underlying natural duration W^n.

Proof. For a policy to be canceled by a detrimental claim it is conditional on the policy having not been canceled for natural reasons. Given an observed duration W we have the probability

$$p_c = 1 - q_c = P(W^n < W) = \int_0^W f_w(w)dw \quad (21)$$

that it has been canceled for natural reasons. Combining this with equation (22) we have the probability $p_d q_c$ that our observed duration policy has been canceled due to detrimental claims, and $1 - q_d p_c$ that it has been canceled for natural reasons, yielding equation (20) □

Observe the difference between p_d from lemma 4.1 and p_D from lemma 4.2 where p_d denotes the probability of a single claim being detrimental, while p_D denotes the cancellation of the whole policy due to a detrimental claim. In addition, we will now show through the following corollary that we have a closed form expression for the probability of detriment given p_d and p_n.

Corollary 4.2.1. *Given the assumptions of lemma 4.2 and assuming a conditional uniform Bernoulli distribution according to equation* (16) *we have the probability of detrimental claims as*

$$p_D = 1 - q_D = \begin{cases} 0, W = 1 \ or \ Z = 0 \\ p_d\left(1 - W p_n\right), else \end{cases} . \quad (22)$$

Proof. Given the parametric assumptions on W^n as specified in equation (16) we have the cdf

$$F_W(w) = \int_0^W f_w(w)dw = \begin{cases} W p_n, W < 1 \\ 1, else \end{cases} \quad (23)$$

Inserting this into equation (20) as specified by lemma 4.2, we arrive at equation (22). □

4.2 Expected Natural Duration

We have now derived an expression for the probability of a policy being canceled by a detrimental claim. It now remains that given the probability of detriment as specified by corollary 4.2.1 equation (22) to provide an expression for the expected natural duration W^n given X, W, and Z. This will result in a closed-form expression of the expected natural duration $W^{adj} = \mathrm{E}[W^n|X, W, Z]$, which is derived by the following lemma.

Lemma 4.3. *Given an observed duration W, claims Z, assuming independent claims with a single probability for detriment p_d according to lemma 4.1 and assuming a conditional uniform Bernoulli distribution according to equation* (16)*, we have the expected natural duration*

$$\breve{W}^{adj} = q_D W + p_D \left((1 - p_n) + p_n \frac{(W+1)}{2} \right) \quad (24)$$

Proof. Given p_D from lemma 4.2 we have the expected value of W^n given W, Z as

$$\mathrm{E}[W^n|X, W, Z] = q_D W + p_D \, \mathrm{E}[W^n|W^n > W]. \quad (25)$$

Using the parametric assumptions as per equation (16) we have

$$\begin{aligned} \mathrm{E}[W^n \mid W^n \geqslant W] &= \int_W^1 w f_W(w)dw \\ &= (1 - p_n) \cdot 1 + p_n \left(W + \frac{(1-W)}{2} \right) \\ &= (1 - p_n) + p_n \frac{(W+1)}{2}. \end{aligned} \quad (26)$$

Inserting this equation into equation (25) we arrive at our estimate as defined by equation (24). □

We have now derived a closed form expression of the expected natural duration W^n given W and Z. This is to be used to estimate the true underlying duration of our observed claims. There still are some technicalities; specifically, we have by definition 4.1 that detrimental claims not only censor the observed duration but also the observed amount of claims. As such we need to adjust our estimate in equation (24) to the observed amount of claims and after that estimate our model parameters.

4.3 Unobserved Claims

The censoring of duration does not only censor the remaining duration of our policy, it may also censor the observed amount of claims. One could imagine reformulating (24) as $\mathrm{E}[W^n \mid W] = W + \widetilde{W}$ and estimating outstanding claims according to $\mu(X)$. We will now show that this analogue to equation (6) by the following lemma:

Lemma 4.4. *Given an adjusted duration $W^{adj} = W + \widetilde{W}$, and estimating outstanding claims according to equation* (17) *as*

$$\mu(X) = \frac{\mathrm{E}[Z \mid W] + \widetilde{W}\mu(X)}{W + \widetilde{W}}. \tag{27}$$

is identical to equation (6)*.*

Proof. Rewrite equation (27) as

$$\mu(X)(W + \widetilde{W}) = \mathrm{E}[Z \mid W] + \widetilde{W}\mu(X) \tag{28}$$

We here observe that $\widetilde{W}\mu(X)$ is present on both sides of the equal sign, yielding

$$\mu(X)W = \mathrm{E}[Z \mid X, W] \tag{29}$$

which is analogous to equation (6). $\square$

An alternative approach would be not to expect the full natural duration W^n of the policy, but the *remaining coverage* W^{tail} which would be somewhere between 0 and the expected time for the following unobserved claim. This is illustrated in Figure 4, but no results in estimating the remaining coverage have been found and the topic is left for future research.

4.4 Adjusting for Unobserved Claims

Instead of seeking to estimate unobserved claims or expected tail, we decided to ensure the adjusted duration reflects the same amount of claims as the observed durations. What this means is to ensure that the sum of adjusted duration equals the sum of observed duration, which can be done by the following adjustment:

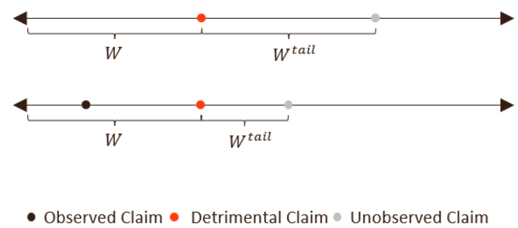


Figure 4: Alternative suggested idea to estimate the tail of remaining coverage W^{tail} to take unobserved claims into consideration.

$$\widehat{W}^{adj} = \widetilde{W}^{adj}\frac{\mathrm{E}[W]}{\mathrm{E}[W^n]} \tag{30}$$

Observe that $\mathrm{E}[\widehat{W}^{adj}] = \mathrm{E}[W]$ and that we have an adjustment that does not increase or decrease the total duration. We will estimate $\mathrm{E}[W^n]$ with $\frac{1}{n}\sum_{i=1}^{n}\widetilde{W}_i^{adj}$ and $\mathrm{E}[W]$ with $\frac{1}{n}\sum_{i=1}^{n} W_i$.

A way to motivate this adjustment is to see equation (30) in terms of a probability measure, see [6], and define a probability measure P_{W^n} as

$$\mathrm{E}_{P_{W^n}}[A \mid W] = \mathrm{E}[\frac{W^n}{\mathrm{E}[W^n]}A \mid W]. \tag{31}$$

Through this definition one can rewrite the equation (17) as

$$\mathrm{E}_{P_{W^n}}[Y \mid W] = \mu(X). \tag{32}$$

Which, if you expand the expression, yields our estimate (30) and assumption (17). This formulation is possible to motivate the adjustment, but for all intended purposes, this adjustment is made to account for unseen claims.

As a conclusion, we now have a closed-form expression for the expected duration of a policy given the observed duration and the amount of claims. Observe that this methodology could be extended to other parametric assumptions on W. Also note that our p_D is generalizable to be 1 if one manually identifies detrimental claims and 0 otherwise. Even so, we still need to estimate the parameters of this model, which will be the final step of the methodology.

4.5 Estimation of Parameters

In the above model, we have two unknown parameters, p_d, the probability of a detrimental claim and p_n, the probability that a policy will be canceled for natural reasons. Again, for $Z = 0$ we only have observations on the duration that naturally occurs, which means that we can estimate the cancellation rate p_n as

$$\widehat{p}_n = \frac{\sum_{i:Z_i=0} 1_{[W_i<1]}}{\sum_{i:Z_i=0} 1}. \tag{33}$$

For p_d the estimation is trickier, for $Z > 0$ our data contains both policies canceled naturally or by detrimental claims. To distinguish these scenarios we will try to estimate p_d by comparing the cancellation rates when $Z = 0$ and $Z > 0$. If the cancellation rate is high for $Z > 0$ compared to $Z = 0$, it means a high probability of detrimental claims, and low probability if the cancellation rates are similar. For $Z > 0$, we have the probability of cancellation p_z as

$$\widehat{p}_z = \frac{\sum_{i:Z_i>0} 1_{[W_i<1]}}{\sum_{i:Z_i>0} 1}. \tag{34}$$

We now show through these estimates how to calculate an estimate p_d based on the discrepancy between p_z and p_n.

Lemma 4.5. *Given a dataset $X_i, W_i, Z_i i = 1, 2, \dots, n$ and cancellation rates p_n for $Z = 0$ and p_z for $Z > 0$, one has the probability of detrimental claims according to*

$$p_d = \frac{p_z - p_n}{1 - 2p_n} \tag{35}$$

Proof. By the definition p_d as per lemma (4.1) and definition 4.1 we have the probability of cancellation for a policy with $Z > 0$ being canceled by either a detrimental claim or naturally as

$$\begin{aligned} p_z &= P(W < 1 \mid Z > 0) \\ &= q_d p_n + p_d q_n \\ &= (1 - p_d)p_n + p_d(1 - p_n) \\ &= p_n + p_d - 2p_d p_n \\ &= p_n + p_d(1 - 2p_n). \end{aligned} \tag{36}$$

Rearranging this equation we arrive at equation (35). □

From lemma 4.5 we can insert our estimates $\widehat{p}_z$ and $\widehat{p}_n$ to estimate $\widehat{p}_d$ as

$$\widehat{p}_d = \frac{\widehat{p}_z - \widehat{p}_n}{1 - 2\widehat{p}_n} \tag{37}$$

Using these estimates, we see in Figure 5 the results from estimating p_d using equation 35 on simulated policies as simulated by algorithm 1. As can be seen, there is some indication of positive bias, but generally, the real value is within the 95% confidence interval.

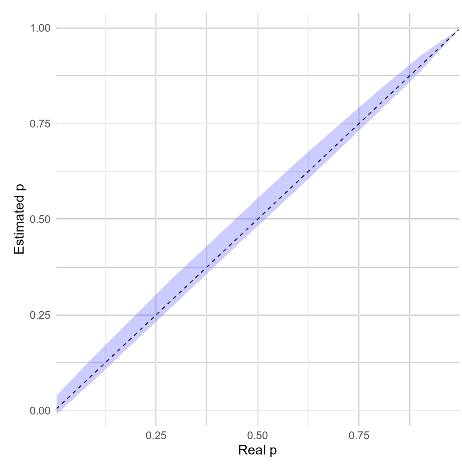


Figure 5: 95% confidence interval of detrimental claim probability estimate $\widehat{p}_d$ using equation (37) on 100 simulations of data per algorithm 1 using different values of p as in chapter 3.2.

4.6 Arbitrary Duration

Duration is sometimes not limited to 1. Sometimes a policy is signed to cover multi-year or arbitrary periods of coverage. One should also mention that many contracts are "renewed", meaning it is the same customer, but a new policy. One could imagine the full coverage period across renewals could be seen as the duration measure.

In these cases, the above terminology would not work as we expect a maximum duration of 1. Instead, we would need to generalize our expression from the maximum period of 1. Let W^{tail} denote the potential remaining period of the policy, in the above case we have $W^{tail} = 1 - W$, much according to the idea presented in chapter 4.3. Using this definition, we can generalize equation (24) to

$$\breve{W}^{adj} = q_D W + p_D \left(\int_{W}^{W+W^{tail}} w f_W(w) dw \right) \quad (38)$$

We here generalize our formula to have an arbitrary remaining coverage W^{tail}. The integral is a generalization of equation (26). Observe that $W^{max} = W^{tail} + W$ does not have to be the maximum expected tail or highest possible duration; it can be below a certain threshold. In such a scenario p_n needs to be redefined as the probability of early cancellation and p_n as the probability of early cancellation due to a detrimental claim.

In this thesis we will strictly study the case of $W^{max} = 1$. As such, the evaluation of said method is left for future research.

4.7 Experiments

To evaluate the suggested duration adjustment, we will begin on the simulated data, where the underlying model parameters are known. Here we will use the same simulation as in chapter 3.2, algorithm 1, and see how the estimates of the dispersion and risk intensities differ. Specifically, we want our adjusted duration to remove the over-dispersion observed in Figure 2 but still accurately estimate the intensities as in Figure 3.

Further, we also want to see how said method would affect dispersion estimates in an applied setting. As such, we will replicate the experiment in 3.1, but now specifically study dispersion and how it is estimated across risk-quantiles.

In estimating the dispersion ρ, we will insert our different duration estimates W, $\widehat{W}^{L\&N}$ as the expected duration according to [5] equation (24) and $\widehat{W}^{adj}$ the adjusted duration as specified by equation (30) into equation (13). Doing so, we will replicate the results in [5] and compare these to the new method suggested by this thesis. The experiments and results can be replicated using [9].

Finally, we will evaluate the predictive performance of the method. Specifically, we want to evaluate performance in an applied situation, where the final duration is unknown. As such, we will perform 10-fold cross validation prediction, where we will use the adjusted duration by equation (30) in training, but in evaluation on the test fold, we will evaluate the realized earned premium compared to realized claims. In doing so, we will evaluate the deviance and absolute error of the resulting earned premium and claims in a setting where the final duration is unknown.

In analyzing real data, we replicated the experiment conducted in [5] and added additional datasets from the CAS dataset, see [1]. Doing so, we fitted a gradient boosting machine using the `gbm` package in R, using a shrinkage factor of $\epsilon = 0.1$, training fraction 80%, maximum depth of 2, and out of 500 maximum trees selected an optimal amount of trees based on the validation loss.

Details regarding the resulting fit is found in section 3.2. In this model, all covariates will be used, and we will have the different duration measures as offset. In the results W^{obs} denotes observed duration, $\widehat{W}^{adj}$ is the detrimental adjusted duration in equation (30) and $\widehat{W}^{L\&N}$ is the local mean estimate in [5] equation (24)

5 Results

As described in section 4.7 we will begin by presenting the results on simulated data, seeing how the suggested method changes dispersion and intensity estimated on known true values. Following this, we will show the results from estimating the dispersion ρ using the adjusted duration and compare it to the results using [5] equation (14). Finally, we will show the results in an applied setting, where our pure premiums $\pi(X)$ are estimated on historical data, but evaluated in an applied setting, where W is unknown.

5.1 Simulated Data

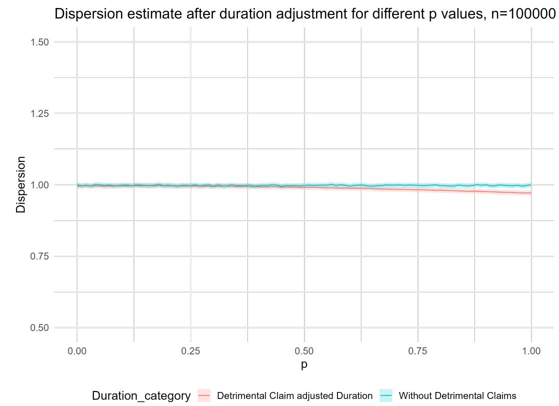
In Figure 6 we observe how adjusting the duration for detrimental claims compares to the underlying estimates in the Poisson claim generating process. These results should be taken in comparison to figures 2 and 3 where the main observation is that we no longer observe over-dispersion and no longer induce variance in our risk estimates. Unfortunately, we do observe small negative bias that arise for high values of p.

5.2 Real Insurance Data

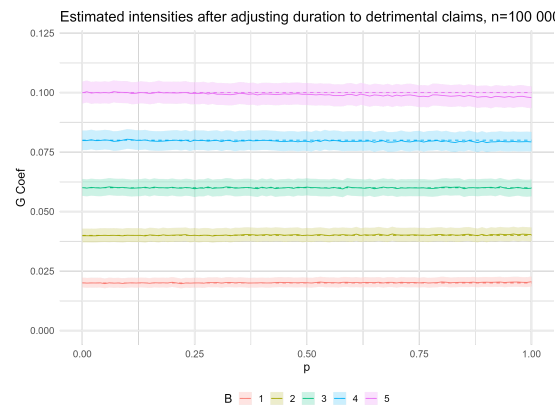
5.2.1 Dispersion Estimates

The resulting dispersion estimates can be seen in table 3. Here we observe that the duration-adjusted dispersion estimate $\widehat{\rho}^{adj}$ is close to 1 and does not indicate over-dispersion. In fact, the results indicate that mostly we cannot dismiss the hypothesis of $\rho = 1$, meaning that there is no over or under-dispersion. An exception is `swmotor` where both using $\widehat{\rho}^{L\&N}$ and $\widehat{\rho}^{adj}$ indicate under-dispersion. Note that we observe unit dispersion and consistency between using $\widehat{\rho}^{L\&N}$ and $\widehat{\rho}^{adj}$.

Continuing by replicating the dispersion plot in [5] we see in Figure 7 the moving average dispersion estimate across the pure premium percentiles. First observe that some local means do not exhibit claims which leads to an estimate of 0. Secondly, observe that the detrimental claim adjusted duration $\widehat{W}^{adj}$ aligns with the local mean estimate $\widehat{W}^{L\&N}$ across the dataset.



(a) Dispersion estimate using the Pearson dispersion estimate with 95% Chi-squared confidence intervals.



(b) Estimated Risk intensities for the 5 different risk groups.

Figure 6: Results comparing Quasi-Poisson GLM model estimates on poisson generated data without detrimental claims compared towards glm model adjusting for simulated detrimental claims by equation (30). Experiment done 100 times simulating n =100 000 policies

5.2.2 Performance

In table 4 and table 5 we see the resulting performance metrics using the above duration adjustment methods on unseen data. This compares the earned premiums in an applied setting, where the duration is unknown, to the realized claims.

Dataset	$\widehat{\rho}$	$\widehat{\rho}^{L\&N}$	$\widehat{\rho}^{adj}$
`ausPrivateAuto`	1.43 ± 0.15	1.07 ± 0.04	0.99 ± 0.04
`freMTPLfreq`	1.68 ± 0.11	1.03 ± 0.02	0.98 ± 0.02
`swmotor`	1.45 ± 0.28	0.85 ± 0.11	0.85 ± 0.11
`brveh`	2.46 ± 1.20	0.87 ± 0.15	0.88 ± 0.16

Table 3: Resulting dispersion estimates $\pm$ the 95% confidence interval by equation (13) where $\widehat{\rho}$ used the observed duration, $\widehat{\rho}^{L\&N}$ expected duration according to [5] equation (24) and $\widehat{\rho}^{adj}$ the adjusted duration as specified by equation (30).

In table 4 we note that the alternative methods provide similar deviance to using the observed duration, except for the scenario of `swmotor`. Also, we see in table 5 that the duration adjustment methods decrease the absolute error in the amount of claims. Specifically, we see that the local mean estimate $\widehat{W}^{L\&N}$ and detrimental claim adjusted duration $\widehat{W}^{adj}$ create similar improvements in the amount of absolute claim error, except for the case of `fretmtpl`.

Dataset	W^{obs}	$\widehat{W}^{adj}$	$\widehat{W}^{L\&N}$
`ausprivate`	25394	25413	25427
`fretmtpl`	103766	104643	104581
`swmotor`	3952	4510	8028
`brveh`	10561	10627	10627

Table 4: Total Deviance from 10 fold cross val predicted pure premiums using the different duration estimates as offset and evaluating using realized earned premium $W\widehat{\mu}(X)$ compared to claims

Dataset	W^{obs}	$\widehat{W}^{adj}$	$\widehat{W}^{L\&N}$
`ausprivate`	8968	8888	8884
`fretmtpl`	31012	31257	30743
`swmotor`	869	782	755
`brveh`	2115	2086	2085

Table 5: Absolute error from 10 fold cross val predicted pure premiums using the different duration estimates as offset and evaluating using realized earned premium $W\widehat{\mu}(X)$ compared to claims Z.

6 Discussion

The results indicate that adjusting for detrimental claims can improve modeling performance. From the simulated data, we observe that over-dispersion is controlled for in Figure 6a and that we no longer induce variance in our risk estimates. We do observe some type of bias towards under-dispersion for high values of p. which seems to arise from the high intensity group 5 in Figure 6b where there likely are censored claims which would arise after cancellation.

A potential reason for this bias is that unobserved claims are not correctly accounted for, and one sees fewer claims for high intensity groups (see risk group 5 in Figure 6b) But one may argue that the scenario of high intensity and a large chance of detrimental claims is rare, and the bias should be taken into relation with the dispersion observed in Figure 2, where the bias is marginal, even for large p. But it is a topic for future research in how one may remove this bias; some suggestions for handling unobserved claims are discussed in chapter 4.3.

In evaluating the model on real data, we observe how the resulting dispersion estimates in table 6a align with the local mean estimate suggested in [5]. In this case, we no longer observe over-dispersion, as we would using the observed duration. As this is real data, we cannot say what the real dispersion would be, but by referring to the results on simulated data, Figure 6 and 2, the over-dispersion observed without adjustments may be induced. An application of these methods could be to test for over-dispersion, as seen this method corrects for over-dispersion caused by detrimental claims.

Lastly, we see from evaluating the methods on unseen data, where the duration is unknown, we decrease the absolute error in the amount of claims, while having similar results observed deviance. This is a welcome addition, indicating that the over-dispersion observed can worsen model performance in an applied setting. Based on the results simulated scenario, see Figure 3, the induced variance of detrimental claim does not significantly affect parameter estimation, even if some bias is indicated.

7 Conclusion

We have in this thesis reassessed the duration effect in non-life insurance. Largely, this is an extension to the work made in [5] which observed that the linear assumption on the observed duration leads to over-dispersion which induces variance in our risk estimates and makes it impossible to distinguish between risk groups. This thesis analyzes the over-dispersion and shows in a simulated experiment that it may arise from *detrimental claims*.

A detrimental claim is a claim severe enough to in effect cancel the policy. It is shown that by introducing said dynamic to a regular Poisson claim arrival process, one exhibits over-dispersion where the underlying claim generation process has none.

The paper suggests a method for adjusting data towards detrimental claims. By estimating the probability of a policy being canceled for natural reasons p_n and the probability of a claim being detrimental p_d, we derive an expression which adjusts the duration for potential censoring by detrimental claims.

Using said method we show that we can remove the over-dispersion caused by detrimental claims, maintain estimates and improve the absolute error on unseen data. The method seems to indicate a small bias for large values of p_d and high intensity, a situation which is speculated to arise from unseen claims which could be observed if the policy was not canceled by a detrimental claim.

This method can be applied directly in the modeling of P&C insurance and is compatible with existing methods. It can also be used as a test to validate if one exhibits over-dispersion or not. Further work could extend on the possibility and modeling of unobserved claims which arise due to detrimental claims and analyze duration which is not limited to 1 as in this thesis. Suitable approaches are suggested in chapter 4.3 and 4.6.

References

[1] C. Dutang, A. Charpentier, and M. C. Dutang. Package 'casdatasets'. *url: https://www. openml. org/search*, 2024.

[2] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[3] A. Groll, C. Wasserfuhr, and L. Zeldin. Churn modeling of life insurance policies via statistical and machine learning methods. *Journal of Insurance Issues*, 47(1):78–117, 2024.

[4] M. Lindholm, F. Lindskog, and J. Palmquist. Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells. *Scandinavian Actuarial Journal*, 2023(10):946–973, 2023.

[5] M. Lindholm and T. Nazar. On duration effects in non-life insurance pricing. *European Actuarial Journal*, pages 1–24, 2024.

[6] F. Lindskog. *Non-life pricing essentials.* Lecture notes Non-life insurance pricing., 2024.

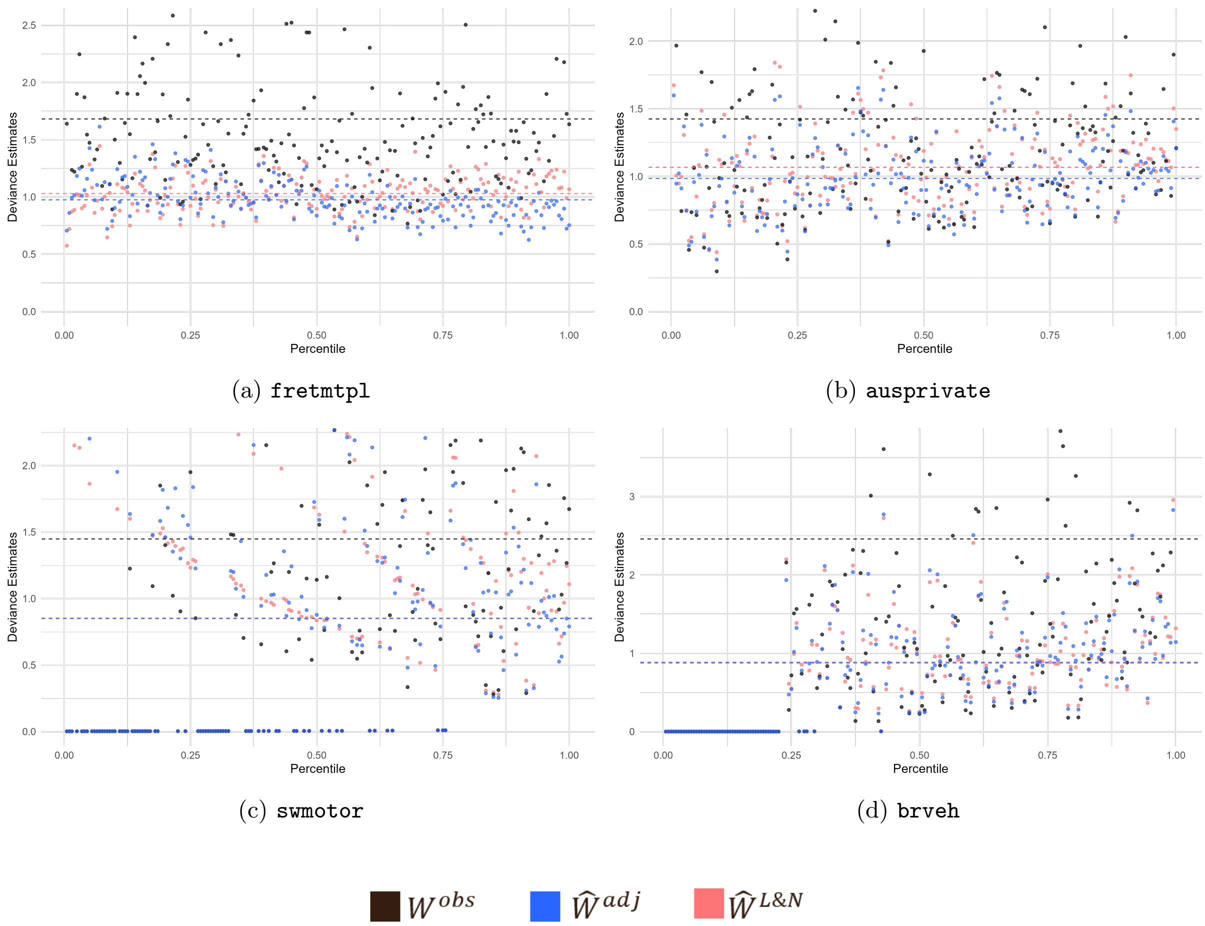
[7] E. Ohlsson and B. Johansson. *Non-life insurance pricing with generalized linear models*, volume 174. Springer, 2010.

[8] E. Ohlsson and J. Lauzeningks. The one-year non-life insurance risk. *Insurance: Mathematics and Economics*, 45(2):203–208, 2009.

[9] O. Potts. Detrimental claim simulation and evaluation. `https://github.com/NisseOscar/Detrimental-Claim-Simulation-Code`, 2024.

[10] Y. Staudt and J. Wagner. What policyholder and contract features determine the evolution of non-life insurance customer relationships? a case study analysis. *International Journal of Bank Marketing*, 36(6):1098–1124, 2018.

[11] M. V. Wüthrich and M. Merz. *Statistical foundations of actuarial learning and its applications.* Springer Nature, 2023.



(a) `fretmtpl` (b) `ausprivate`

(c) `swmotor` (d) `brveh`

Figure 7: Resulting dispersion estimates ordered by predicted pure premium percentile with 0 being lowest risk. $W = W^{obs}$ denote the observed duration, $\widehat{W}^{adj}$ the detrimental claim adjusted duration, and $\widehat{W}^{L\&N}$ the local mean duration, suggested in [5] equation (24).