Annals of Actuarial Science (2024), **18**, pp. 579–604
doi:10.1017/S1748499523000271

ORIGINAL RESEARCH PAPER

Nonparametric intercept regularization for insurance claim frequency regression models

Gee Y. Lee[1]* and Himchan Jeong[2]*

[1]Department of Statistics and Probability, Department of Mathematics, Michigan State University, East Lansing, MI, USA; and [2]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada
Corresponding author: Gee Y. Lee; Email: leegee@msu.edu

(Received 06 September 2022; revised 09 October 2023; accepted 11 October 2023; first published online 5 January 2024)

Abstract

In a subgroup analysis for an actuarial problem, the goal is for the investigator to classify the policyholders into unique groups, where the claims experience within each group are made as homogenous as possible. In this paper, we illustrate how the alternating direction method of multipliers (ADMM) approach for subgroup analysis can be modified so that it can be more easily incorporated into an insurance claims analysis. We present an approach to penalize adjacent coefficients only and show how the algorithm can be implemented for fast estimation of the parameters. We present three different cases of the model, depending on the level of dependence among the different coverage groups within the data. In addition, we provide an interpretation of the credibility problem using both random effects and fixed effects, where the fixed effects approach corresponds to the ADMM approach to subgroup analysis, while the random effects approach represents the classic Bayesian approach. In an empirical study, we demonstrate how these approaches can be applied to real data using the Wisconsin Local Government Property Insurance Fund data. Our results show that the presented approach to subgroup analysis could provide a classification of the policyholders that improves the prediction accuracy of the claim frequencies in case other classifying variables are unavailable in the data.

Keywords: Subgroup analysis; credibility; random effects models; dependence modeling; regularization methods; fused lasso; actuarial modeling; insurance claims analysis

1. Introduction

Subgroup analysis in actuarial science is related to the risk classification problem in insurance ratemaking. In property insurance ratemaking applications, risk classification allows for insurance rates to be charged differently according to the policyholder's geographical region, property type, or policy feature. Risk classification is an important technique for the solvency of the insurance company because it prevents adverse selection against the company. In a competitive insurance market, companies with better risk classification are likely to attract more so-called low-risk policyholders, leaving those companies using inferior risk classification methods with high-risk policyholders. The mechanism by which this can influence the insurance company's solvency is explained in detail by authors such as Akerlof (1970), Rothschild and Stiglitz (1976), or Wilson (1977).

In the actuarial literature, authors such as Guo (2003), and Yeo et al. (2001) have suggested using machine learning approaches to segment the subjects in a population into homogenous groups with similar characteristics, while heterogeneity between the groups is maximized. In the statistics literature, subgroup analysis has been known as a technique called cluster analysis.

*Both authors have contributed equally to this work.

For a recent example, Ma and Huang (2017) proposed a subgroup linear model, where the subgroup structure is defined by group-specific intercepts. They applied the alternating direction method of multipliers (ADMM) algorithm to the subgroup analysis in the statistics literature for the first time, according to our understanding. The approach in Ma and Huang (2017) did not have insurance applications in particular in mind, and they used a pairwise penalty on the subject-specific intercepts to regularize the distance among the subjects.

The adoption of the technique described in Ma and Huang (2017) into the actuarial literature was first done by Chen et al. (2019), who implemented the approach for a zero-inflated Poisson model and applied it to an insurance risk classification context. Their approach also used a pairwise penalty in the same way Ma and Huang (2017) did. The authors of Chen et al. (2019) have also suggested using deviance residuals as a preliminary tool to assess the applicability of subgroup analysis methods to a given dataset. The paper was an interesting contribution to the literature, and in our opinion they presented an alternative approach to viewing the risk classification problem. Although the approach was interesting, we believe there are some practical difficulties applying the method to larger datasets because of the computational load caused by the pairwise penalties. Reducing this computational load is part of the contribution we are making in our paper.

Meanwhile, the fused lasso method has been around for a longer period of time. Authors such as Tibshirani et al. (2005) and Tibshirani and Taylor (2011) have contributed to the development of fused lasso approaches to estimation, where the coefficients are assumed to be ordered from the lowest to the largest one. A recent example of a study applying the ADMM algorithm to the fused lasso problem in an actuarial context can be found in Deviendt et al. (2021). In their work, the authors propose an algorithm for solving the regularized regression problem for a multitype lasso penalized problem. The multitype penalty includes the fused lasso and generalized fused lasso as one of the penalties applicable to the objective function.

The approach in Deviendt et al. (2021) is interesting and is closely related to our approach, although the intercepts were not a target of regularization in their work unlike ours. Additional contributions in our work are that in addition to existing work, we attempt to relate the risk classification problem to the credibility problem and make a comparison with existing Bayesian approaches to the problem. In the actuarial literature, authors such as Jeong and Valdez (2020) or Jeong (2020) have explored the credibility problem using Bayesian approaches. As we understand, credibility theory is a way to consider unobserved heterogeneity within the observations from the same object (could be an individual or a group) so that it is strongly related with the ratemaking problem, where a large part of the interest is in attaining better risk classification.

Hierarchical random effects models are also related to our work, as we compare the ADMM approach to the random effects approach to classification. In actuarial science, hierarchical random effects models have been explored by authors such as Norberg (1986), Boucher and Denuit (2006), and Antonio and Beirlant (2007). Random effects models in actuarial science have been studied extensively in relation to the credibility problem. Another interesting related paper would be Frees and Shi (2017), where the authors use a Tweedie model with multiplicative random effects to incorporate collateral information into a credibility ratemaking model. We are unaware of existing work comparing the ADMM approach to classification with the random effects approach.

The approach presented in this paper illustrates how a modified ADMM approach can be used to estimate the parameters for a subgroup analysis problem quickly. We present three different cases of the subgroup analysis problem, where the first case assumes there are no interdependence among the different coverage groups of a multiline insurance company, the second case assumes perfect interdependence, and the third case assumes flexible interdependence. We also propose an approach to speed up the ADMM algorithm by keeping track of the group members for the regularized intercepts at each step of the iteration. To summarize our contributions, the first is a computational one, where we contribute to improving the speed and estimation of the ADMM algorithm by presenting a way to figure out the initial values, penalize adjacent coefficients only,

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

and keep track of the groupings of the coefficients. The second is an analytical one, where we are able to combine the ADMM approach with dependence models with dependencies induced by common effects. The third is an empirical one, where we demonstrate the application of our approach using the Wisconsin Local Government Property Insurance Fund (LGPIF) dataset.

The rest of the paper proceeds in the following order: Section 2 explains the proposed methodology in detail. Section 3 presents a simulation study to determine when the proposed method may perform well, and provides some discussion. Section 4 starts by illustrating the actuarial problem of interest, and the dataset related to the problem. Section 5 presents the empirical results from the application of our approach to the actuarial problem stated in Section 4. Section 6 concludes the paper with closing remarks.

2. Methodology

Let us consider a usual data structure for multi-peril frequency. For an insurance policy of the ith policyholder, we may observe the multi-peril claim frequencies over time t for the jth coverage type as follows:

$$\mathcal{D}_{j,T} = \left\{ \left(n_{jit}, \mathbf{x}_{it} \right) \;\middle|\; i = 1, \ldots, I,\ t = 1, \ldots, T \right\}, \tag{1}$$

where $\mathbf{x}_{it}$ is a p-dimensional vector that captures the observable characteristics of the policy and n_{jit} is defined as the observed number of accident(s) from claim type $j \in \{1, 2, \ldots, J\}$, for the ith policyholder in year t, respectively.

We assume that the number of claims is affected by both observable covariates via associated regression coefficients as well as the (common) unobserved heterogeneity factor θ_{ji} for each coverage. Imagine that $\mathcal{P}(\nu)$ is a Poisson distribution for now. One way to incorporate heterogeneity into this model would be to multiply policyholder and claim-type specific effects:

$$N_{jit} | \mathbf{x}_{it}, \theta_{ji}, \theta_i \sim \mathcal{P}\left(\nu_{jit} \theta_{ji} \theta_i \right), \tag{2}$$

where $\nu_{jit} = \exp(\mathbf{x}_{it}' \boldsymbol{\alpha}_j)$, and the common effect θ_i allows the lines to be dependent among one another. If $\theta_i = 1$ with probability 1, then we have independent lines. In comparison to the hierarchical random effects literature, the θ's in our model would correspond to the multiplicative random effects factor used in papers such as the one published by Frees and Shi (2017). Note that one needs to impose additional constraints on θ_i and θ_{ji} in (2) to assure model identifiability, which are specified in each of the scenarios discussed in this section later.

Our goal is to develop a method that enables the analyst to easily consider individual unobserved heterogeneity θ_{ji}, and the subject-specific effect θ_i. We are also interested in exploring the possibility of capturing the dependence among the claims of the coverages from the same policyholder. Here, we treat θ_{ji} as either random or nonrandom. In the former case, we assume a parametric distribution of θ_{ji} so that Bayesian credibility is used to compute the posterior expectation of θ_{ji}. In the latter case, we perform a nonparametric regression of θ_{ji} with a fused LASSO-type penalty applied to the log-likelihood function using an advanced version of ADMM algorithm from that of Chen et al. (2019). We refer to this approach as the nonparametric approach, because we do not assume a parametric distribution on θ_{ji}. The original ADMM algorithm by Chen et al. (2019) is summarized in Appendix A. The different cases of our model, depending on how θ_{ji} and θ_i are treated are explained in more detail in Appendix D.

We tried implementing our routines using all three penalties explained in Chen et al. (2019), including the L1 penalty, the minimax concave penalty (MCP) penalty, and the smoothly clipped absolute deviation (SCAD) penalty. The definitions of all three penalties are replicated in the Appendix, for the interested reader. Intuitively, the MCP and SCAD are concave penalty functions capable of shrinking pairwise differences to zero, and κ controls the concavity of the penalty function. As $\kappa \to \infty$, both penalties approach the L1 penalty as shown in Figure 1.

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

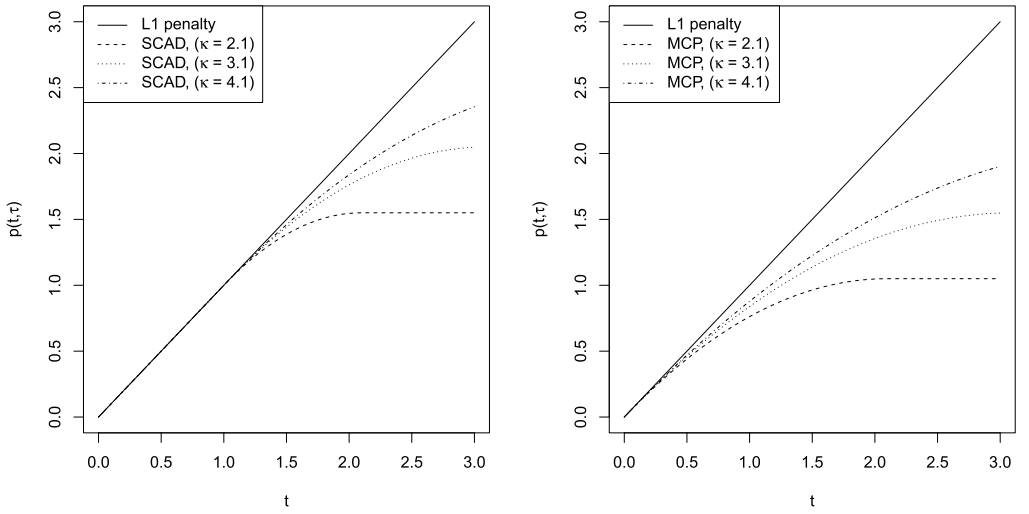

Figure 1. The SCAD and MCP penalty functions with different κ values.

Because the SCAD and MCP penalties are concave, they avoid over-shrinking large differences. The large difference pairs are the pairs that should not be combined into groups, and for these pairs it is optimal to apply smaller penalties. For this reason, the SCAD and MCP are known to reduce biases in the estimated coefficients. As for the choice between SCAD versus MCP, we preferred SCAD because it has a higher curvature near zero and hence is more likely to shrink small coefficients to zero. Furthermore, SCAD gave the cleanest solution paths according to our empirical analyses, so in the rest of this paper and in our implementation, we will assume the penalty term is specifically the SCAD penalty.

To elaborate on this, a solution path is the curve one obtains by adjusting the tuning parameter from a small value to a large value. We prefer that this curve is smooth (clean), without kinks, because only then will the objective function for cross-validation have a smooth shape. Thus, smooth solution paths imply the solutions change continuously as the tuning parameter is altered, which is a desirable property for cross-validation purposes. Yet, even with the SCAD penalty, we found it difficult to obtain nice solution paths in general, where a nice solution path is one where the solution curves are smooth. The problem with the approach taken by Chen et al. (2019) seemed to be that once the policies start forming groups, the group with the largest number of policies start attracting all the other policies into its group. This is because of all the pairwise penalty terms resulting in a large pull from all the policies forming the group. The pairwise penalty approach has other difficulties in the estimation, namely the fact that the estimation algorithm runs very slow when there are a large number of policies. If there are n policies, then there would be $\binom{n}{2}$ pairwise penalty terms. Instead, we propose penalizing the distance between the intercepts of those policies that have similar intercepts. Specifically, we propose using the following penalized likelihood to estimate the fixed effects $\boldsymbol{\alpha}_j$ and coverage specfic unobserved heterogeneity factors $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jI})$:[1]

$$\tilde{\ell}_p(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j | \mathcal{D}_{j,T}) = \tilde{\ell}(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j | \mathcal{D}_{j,T}) - \sum_{i=2}^{I} p\left(|\delta_{ji}^*|; \tau_j\right), \tag{3}$$

[1]Note that we do not estimate nonrandom θ_{ji} and nonrandom θ_i simulataneously in our applications. Therefore, we illustrate the proposed ADMM approach for estimation of only nonrandom θ_{ji} in this section. One can apply essentially the same algorithm with modest modification to estimate nonrandom θ_i, as elaborated in Case 2b of Appendix D.

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

where $\tilde{\ell}(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j|\mathcal{D}_{j,T})$ is the Poisson likelihood given by (2) with $\theta_i = 1$,

$$p(|\delta|;\tau) = \frac{\tau}{2.7}\int_0^{|\delta|} \min\{2.7, (3.7 - x/\tau)_+\}dx,\ \delta_{ji}^* = \log\theta_{ji}^* - \log\theta_{j,i-1}^*,$$

and θ_{ji}^* are the coefficients θ_{ji} ordered by their initial values $\theta_{ji}^{(0)}$:

$$\min\left\{\theta_{ji}^{(0)}\right\} = \theta_{j1}^{*(0)} \le \theta_{j2}^{*(0)} \le \cdots \le \theta_{jI}^{*(0)} = \max\left\{\theta_{ji}^{(0)}\right\} \tag{4}$$

for each j. Note that we may define $\eta_{ji}^* = \log\theta_{ji}^*$ for notational convenience in the algorithm. For a low-dimensional problem with a small number of policyholders, a generalized linear model (GLM) may be used to figure out the initial values, which can be used to obtain the ordering of the coefficients. For a high-dimensional problem, this approach becomes problematic, and our proposal is to use a ridge-regression to obtain the ordering, where the tuning parameter is initially set to a high value and eventually adjusted to a value close to zero. The reason for using ridge regression instead of lasso here, is because we would like to preserve all the coefficients while converging to the initial values of the coefficients, for small values of the tuning parameter. Specifically, let

$$(\boldsymbol{\alpha}_j^{(0)}, \boldsymbol{\theta}_j^{(0)}) = \underset{\boldsymbol{\alpha}_j,\boldsymbol{\theta}_j}{\arg\min}\left[\sum_{i=1}^{I}\sum_{t=1}^{T}\left(n_{jit}(\boldsymbol{x}_{it}'\boldsymbol{\alpha}_j + \log\theta_{ji}) - \nu_{jit}\theta_{ji}\right) - \lambda\sum_{i=1}^{I}\|\theta_{ji}\|^2\right] \tag{5}$$

for a very small (close to zero) tuning parameter λ. The expression in Equation (5) is optimized initially with a large λ value (so that the coefficients form one big group), and the estimates for the θ_{ji}'s from the previous step are used as initial values for the subsequent estimates for a smaller value of λ. The process is repeated until λ is a very small number close to zero (small enough so that it is numerically identical to zero). Because the ridge penalty does not introduce kinks into the objective function, this optimization can be performed using standard maximum likelihood estimation. Once the ordering is obtained, Equation (3) can be optimized for increasing values of τ_j, where the optimal tuning parameter is selected by the one that maximizes the correlation of the predicted claim frequencies with the true frequencies (because that particular tuning parameter would be the one that performs the best in terms of prediction). In practice, this scheme would be implemented so that for each j, we optimize

$$\tilde{\ell}_p^*(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j^*|\mathcal{D}_{j,T}) = \tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j^*|\mathcal{D}_{j,T}) - \sum_{i=2}^{I} p\left(|\delta_{ji}^*|;\tau_j\right). \tag{6}$$

In this case, the augmented Lagrangian is given as follows:

$$L(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j^*, \boldsymbol{\delta}_j, \boldsymbol{\lambda}_j|\mathcal{D}_{j,T}) = -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j^*|\mathcal{D}_{j,T}) + \sum_{i=2}^{I} p\left(|\delta_{ji}^*|;\tau_j\right) + W_j, \tag{7}$$

where $W_j = \sum_{i=2}^{I}\lambda_{j,i-1}(\log\theta_{j,i}^* - \log\theta_{j,i-1}^* - \delta_{j,i}^*) + \frac{\rho}{2}\sum_{i=2}^{I}(\log\theta_{j,i}^* - \log\theta_{j,i-1}^* - \delta_{j,i}^*)^2$.

The problem with the naive Algorithm shown in Appendix A is that the solution path is yet not ideal, because the coefficients within the same group do not merge completely. This problem can be fixed using an encoding of which policies belong to which group at each step s. This approach is illustrated in Algorithm 1.

where

$$L(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*, \boldsymbol{\delta}_j^{(s)}, \boldsymbol{\lambda}_j^{(s)}) = -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*) + \sum_{i=2}^{I} p(|\delta_{ji}^{(s)}|;\tau_j) + W_j^{(s)} \tag{8}$$

$$= -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*) + \frac{\rho}{2}\sum_{i=2}^{I}\left[\eta_{j,i}^* - \eta_{j,i-1}^* - \delta_{j,i-1}^{(s)} + \frac{\lambda_{j,i-1}^{(s)}}{\rho}\right]^2 + C \tag{9}$$

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Algorithm 1: A modified ADMM for fast and stable parameter estimation

1. Set $\boldsymbol{G}_j^{(0)} = \boldsymbol{I}_{I\times I}$, an identity matrix of dimension $I \times I$.
2. Obtain $\boldsymbol{\alpha}_j^{(0)}$ and $\boldsymbol{\theta}_j^{*(0)}$ from Equations (4) and (5), and set $\boldsymbol{\xi}_j^{*(0)} = \boldsymbol{G}_j^{(0)}\boldsymbol{\eta}_j^{*(0)}$
3. Set $\lambda_{j,i}^{(0)} = 0$ and $\delta_{j,i}^{*(0)} = \eta_{j,i}^{*(0)} - \eta_{j,i-1}^{*(0)}$ for $i = 1, \cdots, I-1$.
4. Repeat the following for $\tau_h = \tau_{min}, \cdots, \tau_{max}$, where τ_{min} is close to zero:
   (a) Repeat the following for $s = 0, 1, 2, \cdots$ until convergence:
      i. Let $\left(\boldsymbol{\alpha}_j^{(s+1)}, \boldsymbol{\psi}_j^{*(s+1)}\right) = \underset{\boldsymbol{\alpha}_j, \boldsymbol{\xi}_j^{*(s)}}{\arg\min}\, L\left(\boldsymbol{\alpha}_j, \boldsymbol{G}_j^{(s)\prime}\boldsymbol{\xi}_j^{*(s)}, \boldsymbol{\delta}_j^{(s)}, \boldsymbol{\lambda}_j^{(s)}\right)$ using Equation (8).
      ii. Set $\boldsymbol{\eta}_j^{*(s+1)} = \boldsymbol{G}_j^{(s)\prime}\boldsymbol{\psi}_j^{(s+1)}$.
      iii. Obtain $\delta_{j,i}^{(s+1)}$, $\boldsymbol{G}_j^{(s+1)}$, and $\boldsymbol{\xi}_j^{*(s+1)}$ using Equations (10), (13), and (15), respectively.
      iv. Obtain $\lambda_{j,i}^{(s+1)}$ using Equation (12).
   (b) Use the current estimates of the parameters as initial values for the $\tau_h + 1^{th}$ step.

with a constant C,

$$\delta_{j,i}^{(s+1)} = \begin{cases} ST(u_{j,i}, \tau_h/\rho) & \text{if } |u_{j,i}| \le \tau_h(1+\rho^{-1}) \\ \frac{ST(u_{j,i}, \kappa\tau_h/((\kappa-1)\rho))}{1-((\kappa-1)\rho)^{-1}} & \text{if } \tau_j(1+\rho^{-1}) \le |u_{j,i}| \le \kappa\tau_h \\ u_{j,i} & |u_{j,i}| > \kappa\tau_h, \end{cases} \tag{10}$$

$$\begin{aligned} u_{j,i} &= \eta_{j,i}^{(s+1)} - \eta_{j,i-1}^{(s+1)} + \rho^{-1}\lambda_{j,i}^{(s)}, \\ ST(u,\tau) &= \text{sgn}(u)(|u| - \tau)_+, \end{aligned} \tag{11}$$

$$\lambda_{j,i}^{(s+1)} = \lambda_{j,i}^{(s)} + \rho\left(\eta_{j,i}^{*(s+1)} - \eta_{j,i-1}^{*(s+1)} - \delta_{j,i}^{(s+1)}\right), \tag{12}$$

and

$$\boldsymbol{G}_j^{(s+1)} = (\boldsymbol{g}_{j1}^{(s+1)}, \boldsymbol{g}_{j2}^{(s+1)}, \cdots, \boldsymbol{g}_{j,g_{s+1}}^{(s+1)})', \tag{13}$$

$$\boldsymbol{g}_{jr}^{(s+1)} = \sum_{w\in A(r)} \boldsymbol{g}_{jw}^{(s)}, \quad A(r) = \left\{ w : w \text{ is in the new } r\text{th group} \right\}. \tag{14}$$

Here g_s is the number of groups (number of rows in matrix $\boldsymbol{G}_j^{(s)}$ at step s) and

$$\boldsymbol{\xi}_j^{(s+1)} = \boldsymbol{G}_j^{(s+1)}\boldsymbol{\eta}_j^{*(s+1)}. \tag{15}$$

With Algorithm 1, the gradient and Hessian need to be modified, but it is a simple modification as explained in Appendix C section.

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

The proposed method can be used to capture the unobserved heterogeneity of the policyholders in diverse scenarios on the dependence structure of the insurance claims from multiple lines of business. In that regard, we consider the following cases to elaborate how the proposed method can be applied in the ratemaking practices:

- No interdependence among the lines,
- Perfect interdependence among the lines,
- Flexible interdependence.

In Appendix D, the details of these three cases are explained in full detail. In the following section, we perform a simulation study to compare the different dependence cases in conjunction with the ADMM estimation approach.

3. Simulation study

With the three cases illustrated in Section 2 and Appendix D, we are interested in which approach performs better in different situations of data availablity. We perform a simulation study to seek some answers. Here we assume that $J = 3$ (trivariate coverage case), $I = 1000$, and $T = 5$.
Recall that $\nu_{jit} = \exp(\boldsymbol{x}_{it}'\boldsymbol{\alpha}^{(j)})$ while we set $\boldsymbol{x}_{it} = (x_{1it}, x_{2i}, x_{3i})'$. Here x_{1it} follows a student's t-distribution with degree of freedom of 30, corresponds to a continuous covariate such as age or vehicle value. x_{2i} is a Bernoulli random variable with probability 0.5 that might have varying impacts depending on the coverage type for the same insured. For example, an outdated pick-up truck may have higher risk in bodily injury and property damage liabilities but low exposure on the collision due to relatively less vehicle value, which is vice versa for a compact luxurious electric vehicle. Lastly, x_{3i} follows a normal distribution with mean -0.5 and variance 1, which corresponds to the territorial or individual risk score that affects all coverages simultaneously in the same way, which could be available or not as a covariate. We also fixed $\boldsymbol{\alpha}_1 = (-1, +0.3, 0.1)$, $\boldsymbol{\alpha}_2 = (-1, -0.3, 0.1)$, and $\boldsymbol{\alpha}_3 = (-1, +0.1, 0.1)$.

Note that depending on the data collection scheme or company policies, it is possible that some of the covariates that affect the claim frequency may not be available. In this regard, we assume the following scenarios to fit ratemaking models:

- Scenario 1: All covariates are available,
- Scenario 2: x_{1it} and x_{2i} are available,
- Scenario 3: x_{1it} and x_{3i} are available,
- Scenario 4: Only x_{1it} is available.

Under each scenario, we used the data points $\{\boldsymbol{x}_{it}, N_{jit} \mid t = 1, 2, 3, 4\}$ as a training set, which were fitted with the following models:

- GLM without individual effects: Simple Poisson GLM and no consideration in the unobserved factors, which is equivalent to $\theta_{ji} = 1$ and $\theta_i = 1$ for all i and j.
- Individual effects model: A model where θ_{ji} are nonrandom and estimated by the proposed ADMM approach while $\theta_i = 1$ for all i and j. The details are explained in Case 1b of Appendix D.
- Common effects model: A model where θ_i are nonrandom and estimated by the proposed ADMM approach while $\theta_{ji} = 1$ for all i and j. The details are explained in Case 2b of Appendix D.
- Boosted credibility model: A model where θ_{ji} are nonrandom and estimated by the proposed ADMM approach first, and boosted by the credibility estimate of random θ_i for all i and j. The details are explained in Case 3 of Appendix D.
- True model: A model that used actual ν_{jit}, which is not attainable in practice but only provided as an ideal benchmark.

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Table 1. Out-of-sample validation results with simulated Coverage 1

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
GLM without individual effects	2.6851	2.7409	2.6492	2.7052
Individual effects model	2.8210	2.8423	2.8276	2.8336
Common effects model	3.6703	2.6462	2.5875	2.6252
Boosted credibility model	2.8269	2.8448	2.8377	2.8052
True model	2.6616	2.6741	2.6438	2.6564

Table 2. Out-of-sample validation results with simulated Coverage 2

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
GLM without individual effects	2.1750	2.2427	2.2312	2.2979
Individual effects model	2.1927	2.1893	2.1599	2.1854
Common effects model	2.0942	2.1126	2.1562	2.1863
Boosted credibility model	2.1940	2.1889	2.1745	2.1483
True model	2.1826	2.1970	2.1935	2.2081

Table 3. Out-of-sample validation results with simulated Coverage 3

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
GLM without individual effects	2.4593	2.5299	2.4423	2.5139
Individual effects model	2.4580	2.4500	2.4398	2.4384
Common effects model	2.3996	2.4367	2.3813	2.4251
Boosted credibility model	2.4605	2.4573	2.4505	2.4366
True model	2.4313	2.4467	2.4254	2.4410

After the models are fitted for each dataset, $\{\boldsymbol{x}_{it}, N_{jit} \mid t = 5\}$ was used to assess the model performance via out-of-sample validation with root-mean squared error (RMSE), which is defined as follows:

$$\text{RMSE: } \sqrt{\frac{1}{I}\sum_{i=1}^{I}(N_{ij5} - \hat{N}_{ij5})^2}.$$

Note that we prefer a model with lower RMSE. Tables 1, 2, and 3 summarize the validation measures for all data availability scenarios and models. It is observed that the individual effects model, common effects model, and boosted credibility models tend to perform better than the naive model (GLM without individual effects) as the covariates become unavailable in Scenarios 2, 3, and 4. Note that such improved performance is obtained at the expense of increased computation costs as shown in the Tables E.1, E.2, and E.3 of Appendix E.

4. Data

It is often the case for actuaries to be in a situation, where limited explanatory variables are available besides the exposure variable for modeling the insurance claim frequencies. In this case, the question arises: how would the insurer avoid overcharging those policyholders with little or no claims, while charging a higher, fair, premium for those policies with a recurrently large number of claims. By doing so, the insurer may be interested in avoiding adverse selection, and preventing

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Table 4. Summary of coverage amounts by year in millions of dollars

Year	BC	PN	PO	CN	CO	IM
2006	32.335	0.176	0.532	0.094	0.270	0.772
2007	35.082	0.164	0.588	0.099	0.288	0.819
2008	37.118	0.154	0.613	0.097	0.287	0.815
2009	40.239	0.148	0.674	0.102	0.334	0.903
2010	41.272	0.154	0.684	0.093	0.354	0.958

Table 5. Summary of claim frequencies by year

Year	BC	PN	PO	CN	CO	IM
2006	0.734	0.155	0.092	0.101	0.098	0.040
2007	0.923	0.146	0.089	0.121	0.128	0.057
2008	0.745	0.144	0.084	0.126	0.147	0.055
2009	0.923	0.139	0.129	0.138	0.104	0.069
2010	1.098	0.216	0.126	0.154	0.125	0.062

low-risk policyholders from lapsing. Standard credibility approaches are available for this situation, allowing the actuary to vary the insurance rate depending on past experience, in addition to the exposure variable in hand. Subgroup analysis methods present an alternative approach to solve the same problem, where one may consider the subject-specific intercepts for the frequency regression model as a target of regularization. Which approach performs better in terms of the out-of-sample prediction accuracy would be the question of interest.

We first summarize the dataset used for the empirical analysis. We assume that the insurance company carries multiple coverage groups of business. For the Wisconsin LGPIF dataset that we use, there are six different coverage groups; building and content (BC), contractor's equipment (IM), comprehensive new (PN), comprehensive old (PO), collision new (CN), and collision old (CO) coverage. Table 4 summarizes the coverage amount per policyholder per year for the six different groups over time. The reader may verify that the building and contents coverage group has the highest coverage amounts. Table 5 shows the average claim frequencies for each coverage group per policy per year. Due to its comprehensive characteristics of the coverage, it is observed that the number of BC claims per year is substantially higher than the number of claims from the other coverages. For detailed description and marginal analysis results of the LGPIF dataset, see Frees et al. (2016).

One may expect that the claim frequencies and coverage amounts are highly correlated, and this is verified in Figure 2. The reader may verify that higher claim frequencies are related to higher coverage amounts in general. This justifies using the log coverage amount as the exposure variable. Note that in Figure 2, it is possible for a policyholder to have positive coverage with zero claim. This is a unique feature of insurance claims datasets. Besides the coverage amount, we assume that no other explanatory variables are available for the actuary. This imaginary setup allows us to focus on the risk classification problem, where limited explanatory variables are available. This kind of situation may arise in practice quite naturally in, for example, a guaranteed issue insurance product, where the insurance company is not allowed to underwrite or collect any details regarding the policyholder. In the real dataset we have in our hands, there are in fact other variables available, such as the geographical location, the county code, and the entity type of the policy. Yet, we avoid using them in our analysis to mimic a situation where rating variables are limited. In Section 5, we do provide a comparison of our method with and without the use of

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Table 6. Number of observations by the entity type variable

Entity type	BC	PN	PO	CN	CO	IM
City	795	219	224	191	190	784
County	328	313	313	245	252	328
Misc.	612	54	113	48	98	320
School	1,599	575	779	574	771	1,195
Town	981	182	316	182	313	775
Village	1,346	295	182	289	389	1,220
Total	5,660	1,638	2,138	1,529	2,013	4,622

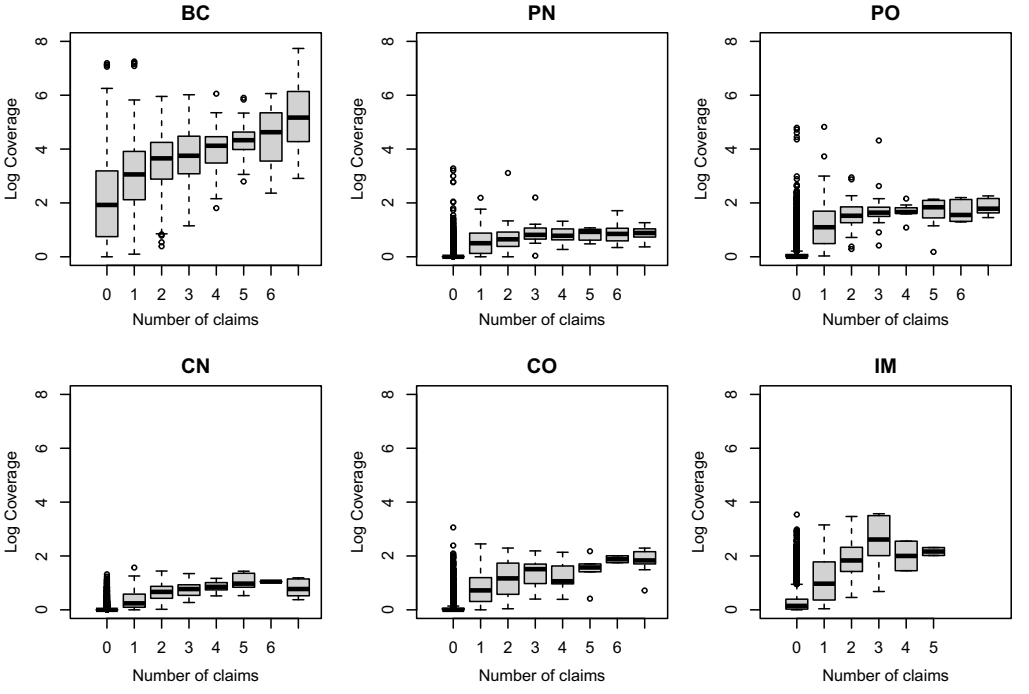

Figure 2. Relationship between coverage amounts and claim frequencies.

additional entity type variables in the regression. For completeness, we include a summary table of the sample sizes by the entity type variable in Table 6.

One of the main goals of our analysis is to figure out a way to prevent policyholders from lapsing. Naturally, it is of interest to see if lapse behaviors are observed in the dataset. Table 7 shows that over time, the number of policyholders with each coverage is in general decreasing. At the same time, the LGPIF has been experiencing a drop in the surplus ratio. The surplus ratio is the ratio of the collected premium to the surplus, and it is a measure of the insurer's financial health. A surplus ratio below the company's target ratio would indicate bad financial health. The fact that this ratio has dropped in recent years makes the property fund manager wonder if there is an alternative approach to the ratemaking problem. Hence, the presence of lapse may provide motivation for the type of analysis we are about the present in this paper.

Figure 3 shows the Cox–Snell residuals for the claim frequencies observed for each policyholder. The Cox–Snell residuals are obtained by performing a simple Poisson frequency regression of the claim frequencies on the exposure and obtaining the fitted distributions first. For example,

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Table 7. Number of policies with positive coverage each year

Year	BC	PN	PO	CN	CO	IM
2006	1,159	359	453	334	424	935
2007	1,143	342	430	318	403	932
2008	1,130	321	422	298	397	922
2009	1,114	307	422	288	400	917
2010	1,098	309	411	291	389	906

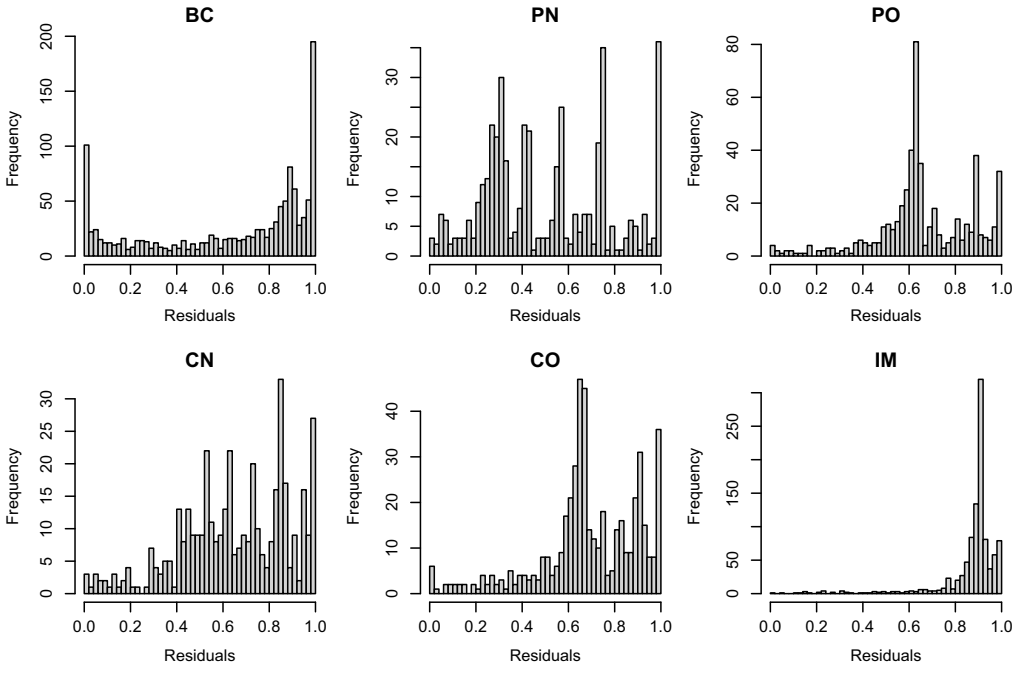

Figure 3. Histogram plots of the Cox–Snell residuals for a basic GLM.

if x_{it} are the exposures (say the log coverage amount) and n_{it} are the observed frequencies for policyholders $i = 1, 2, \cdots, I$ and times $t = 1, 2, \cdots, T$, then one would perform a Poisson regression with the specification

$$N_{jit}|x_{it} \sim \mathcal{P}(\nu_{jit})$$

where $\nu_{jit} = \exp(x_{it}\alpha_j)$, to obtain the parameter estimates $\hat{\alpha}_j$ for each coverage type $j = 1, 2, \cdots, J$ and Poisson distribution functions $\mathcal{P}$ with means ν_{jit}. The observed frequencies are then plugged into the fitted cumulative distribution functions and plotted as histograms. In other words, the Cox–Snell residuals are

$$\hat{F}_{N_{jit}}(n_{jit})$$

where $\hat{F}_{N_{jit}}$ are the estimated cumulative distribution functions of the observed frequencies n_{jit} of the random variables N_{jit}, having coefficients $\hat{\alpha}_j$. The values are plotted as a histogram in each of the panels of Figure 3. This is done as a preliminary analysis to see if there is evidence of the need of subgroup analysis. It is an approach comparable to analyzing the deviance residuals as done in Chen et al. (2019). Figure 3 shows that the residuals are not so uniform and instead have

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

clusters between 0 and 1. This provides motivation to perform subgroup analysis to figure out if the policyholders can be classified into homogenous groups for the ratemaking purposes.

5. Empirical results

To empirically illustrate which approach has the most advantage for the LGPIF dataset, we compare five different approaches to performing a Poisson regression. The five different cases are summarized in the list below. In each case, we may either include or not include the entity type categories as explanatory variables, resulting in ten different special cases. The entity type categories include: city, county, school, town, village, and miscellaneous (six different entity type categories).

- The basic GLM approach: This approach is a plain generalized linear modeling approach, without individual or common effects. It will be used as a benchmark model, which we hope to out-perform using the approaches described in this paper. This would be the case when $\theta_i = \theta_{ji} = 1$ in Equation (2), so that we simply have

$$N_{jit}|\mathbf{x}_{it} \sim \mathcal{P}\left(\nu_{jit}\right), \tag{16}$$

  Note that $\mathbf{x}_{it}$ would include the log coverage amount, and the entity type categories, which are the same across different coverage types. Yet, a different GLM is used for each coverage type, and hence the subscript j. If the entity type categories are not used, then $\mathbf{x}_{it}$ would include a single covariate, which is the log coverage amount.
- The individual effects model: This is one of the ADMM approaches, where each of the six lines is modeled separately with fixed policy-number effects. The details are explained in Case 1b of Appendix D.
- The common effects model: This is another ADMM approach, where we have perfect interdependence among the different coverage groups. The details are explained in Case 2b of Appendix D.
- The offsets model: This is a two-step approach, where the predictions from the individual effects model are used as offsets in the regression model for the common effects model, hoping that the second step common effects model will capture any residual common effects among the different coverage groups.
- The boosted credibility model: This is a hybrid approach, where we have flexible interdependence. The individual effects model is boosted with credibility factors. The details are explained in Case 3 of Appendix D.

We now compare the basic GLM approach, the individual effects model, the common effects model, the offsets model, and the boosted credibility model. Tables 8 and 9 compare the five different approaches. The hold out sample comparisons were performed after omitting new policies for which experience does not exist. The optimal tuning parameters for the models using the ADMM approach are determined using a fourfold cross-validation, where in each fold a particular year of data is used as the validation sample and the rest are used to fit the model. Computation time was a critical issue in completing this project, and we found that a fourfold cross-validation (as opposed to a ten-fold) was enough to deliver the results we wanted in a reasonable amount of time. The model with the lowest validation sample Akaike Information Criteria (AIC) is selected as the optimal model.

In Tables 8 and 9, the GLM without individual effects is a simple Poisson regression model. The individual effects model predictions are obtained by concatenating the predictions from the six different coverage groups ($j = 1, \cdots 6$) and comparing it with the hold out sample frequencies from each coverage group. The same is true for the common effects model, the offsets model, and the boosted credibility model.

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Table 8. Spearman correlations with the hold out sample frequencies

	With entity type	Without entity type
GLM without individual effects	0.509	0.453
Individual effects model	0.509	0.474
Common effects model	0.508	0.471
Individual and common effects using offsets	0.509	0.474
Boosted credibility model	0.510	0.487

Table 9. Number of unique fixed-effect coefficients

	With entity type	Without entity type
GLM without individual effects	42	12
Individual effects model	42	28
Common effects model	37	7
Individual and common effects using offsets	79	35
Boosted credibility model	42	28

Table 10. Number of groups in the intercepts of the individual effects model

	BC	PN	PO	CN	CO	IM
With entity type	1	1	1	1	1	1
Without entity type	4	5	5	4	2	2

From Tables 8 and 9, we can see that the GLM and the individual effects model are giving the same results. This is because with the entity type fixed effects present, the individual effects model converges to the trivial model without any additional fixed effects. The resulting overall intercept is exactly same as the intercept for the GLM model, and so are the other coefficients as well. We can see that even the boosted credibility model does not show a particular advantage in this case.

Table 10 illustrates that with the entity type covariates present, the ADMM approach proposes the trivial model without any subject-specific intercepts as the optimal models. This can be seen in the first row, where the number of intercepts in the resulting model is 1 for all of the six coverage groups. In the second row, we can see that if the entity type covariates are unavailable, then the resulting optimal model has several groups in the intercepts. This is consistent with our expectations from the preliminary analysis using the Cox–Snell residuals in Section 4.

The last column in the tables illustrate when and how our approach can be useful. The individual effects model without entity type would be used in an imaginary situation, where the analyst only has a single exposure variable and no other predictors. In this case, the GLM model is outperformed by the individual effects model, which uses 16 additional coefficients. The common effects model seems to have not much added benefits. The boosted credibility model seems to be the best one with 48.69% Spearman correlation with the hold out sample claim frequencies.

The solution paths for the individual effects models for the six lines are shown in Figures 4 and 5. Figure 4 shows the case when there are no entity type fixed effects, while Figure 5 shows the solution path of the $\eta^*_{j,i}$ values for the model with entity type fixed effects.

To understand the figure intuitively, imagine changing the tuning parameter τ from a small value to a larger value. As the tuning parameter increases, more and more penalty is applied to the likelihood function, forcing more coefficients to merge with one another. As the coefficients merge more and more, the model contains less and less unique coefficients. At some point, there

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

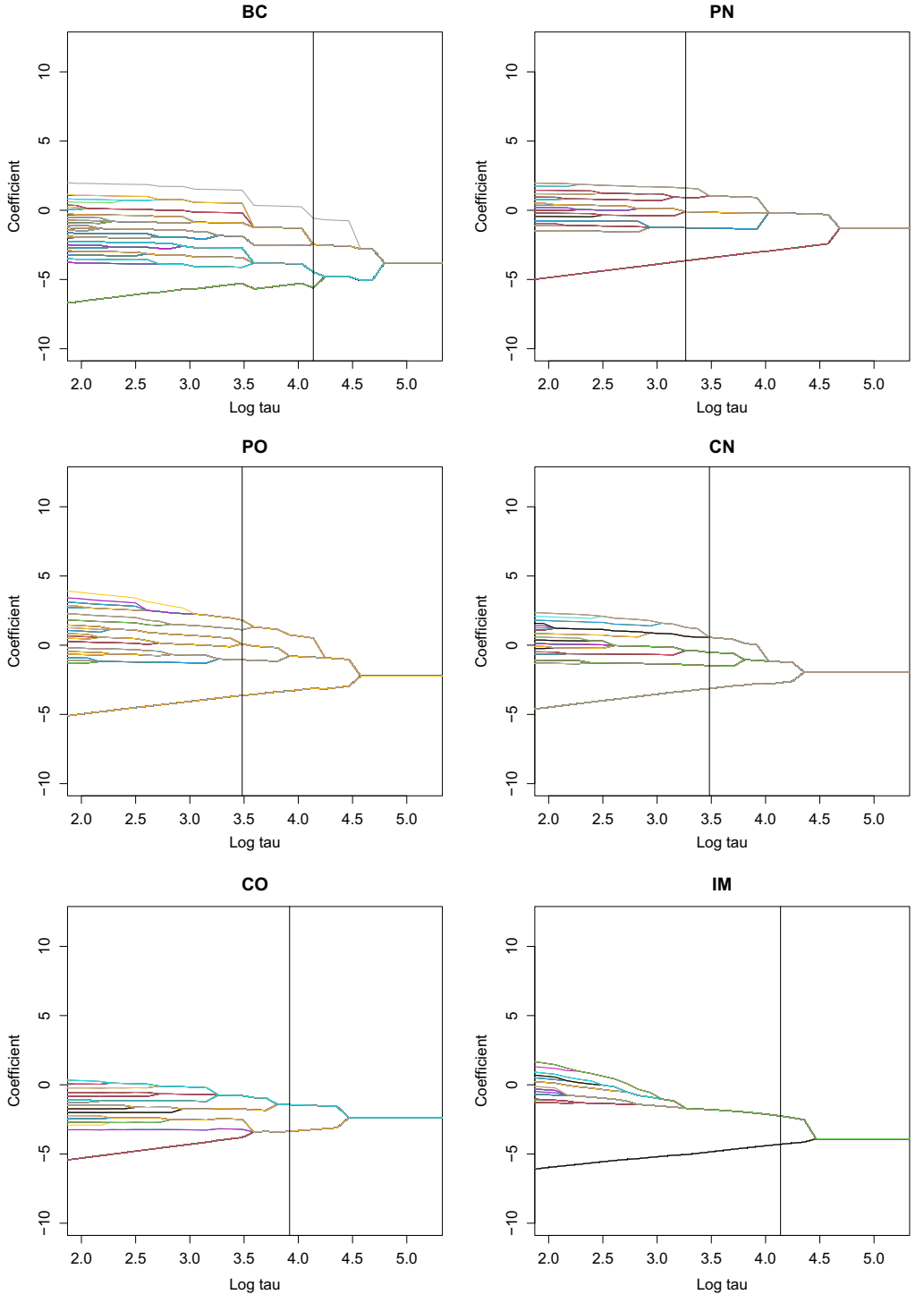

Figure 4. Solution paths for the six coverage groups without entity type predictors.

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

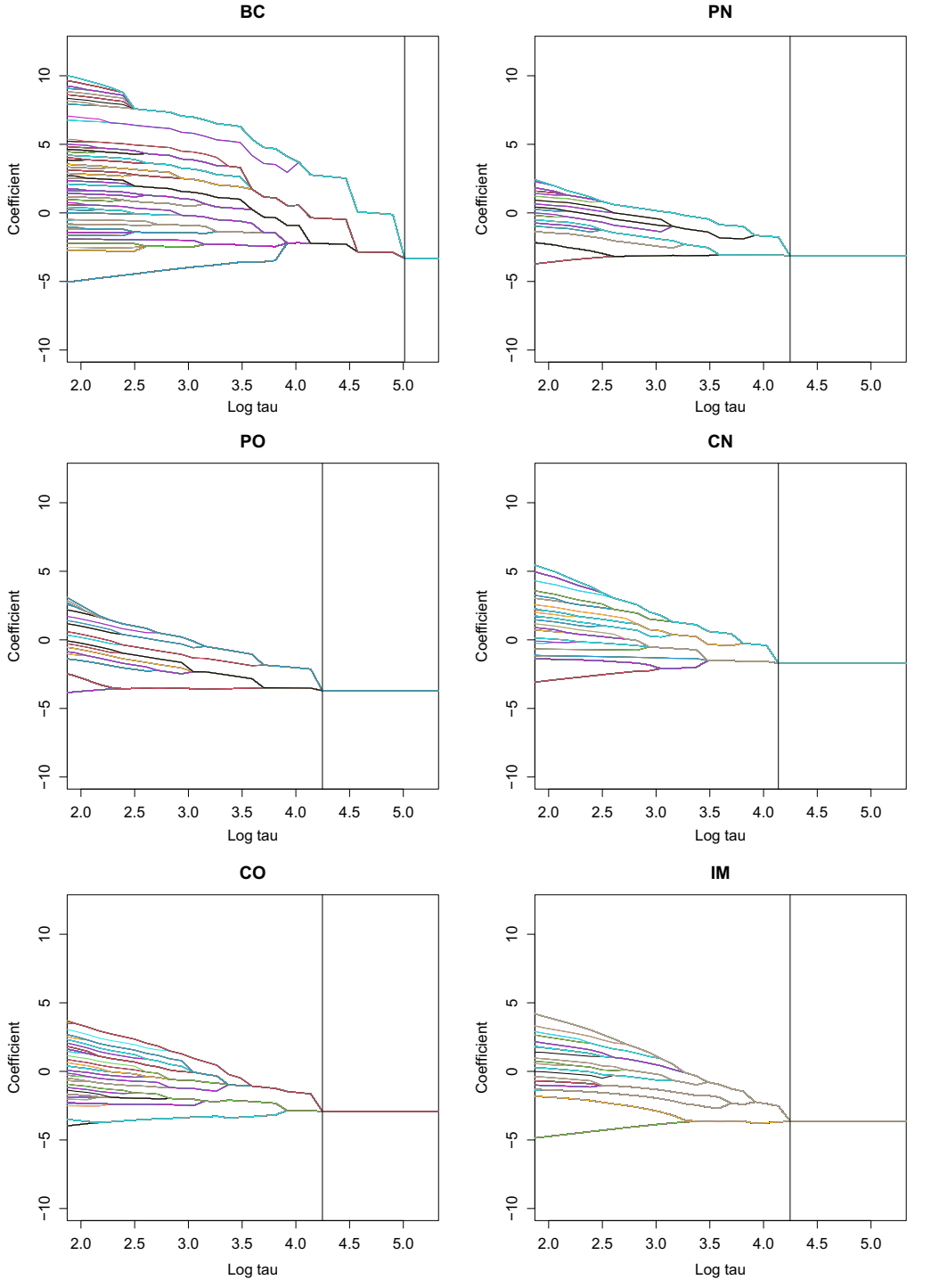

Figure 5. Solution paths for the six coverage groups with entity type predictors.

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

will be an optimal model suggested by cross-validation, and that optimal model is indicated by a vertical line in each panel of the figures. Note that each line would represent one coefficient in the beginning, but that beginning point is now shown in the figures, so as to magnify the most interesting part of the entire figure.

In the figures, we can see that the optimal model chosen by cross-validation contains only one unique coefficient for the intercepts, indicating that the optimal model is the GLM (the optimal tuning parameter is indicated by the solid vertical line). In Figure 4, we can see that the selected optimal model contains more than one unique parameters for the $\eta^*_{j,i}$ values for each coverage group.

6. Conclusion

In this paper, we presented an approach for enhancing the prediction from a GLM model by allowing the model to have subject-specific intercepts. We have presented a modified ADMM algorithm that uses penalties for adjacent coefficients only, as opposed to the pairwise penalty approach in the original ADMM for subgroup analysis. Our approach runs faster with better looking solution paths, and we have been successful at applying the approach to a real insurance dataset with more than thousands of unique policyholders. The optimal number of unique intercepts is determined by cross-validation, and in the empirical studies, we have demonstrated that the approach can be further enhanced using a flexible interdependence approach among the coverage groups.

The benefits and costs of the ADMM approach are as follows: The first benefit of the ADMM approach for the unobserved heterogeneity is that it can be flexibly applied to an actuarial application regardless of the underlying distribution of $N|\theta$. For example, the response $N|\theta$ does not need to follow Poisson but any frequency distribution such as zero-one inflated negative binomial (Shi & Lee, 2022). Another benefit is that one can also use the approach to capture the unobserved heterogeneity of random variables, whose dependence structure is described by a copula. For example, one way this can happen is if one assumes a multivariate Gaussian copula with many dependence parameters (due to the fact that there are many response variables). Each dependence parameter may be a target for subgroup analysis. The disadvantage of the approach seems to be the computation time. If $N|\theta$ indeed follows a Poisson distribution, then it could be more sensible to assume that θ follows a gamma distribution that leads to much simpler estimation and prediction with closed form formulas. Nevertheless, the ADMM approach is conceptually simple, and it provides a simple alternative to the Bayesian approach to credibility in case there is a reasonable number of unique policies.

Our contribution in this paper is the following. First, we have presented a modified version of the ADMM algorithm for subgroup analysis, allowing the penalties to be applied to adjacent coefficients only, and allowing the number of unique coefficients in the model to decrease as the iterations increase, allowing the algorithm to run even faster. Second, we have demonstrated how the approach can be extended to a hybrid approach, so it allows for the flexible incorporation of dependencies among different coverage groups for a multiline insurance company. Third, we presented a case study using the LGPIF dataset, where we compared the performance of the ADMM approach with competing approaches with a full insurance claims dataset.

Possible future work may include the application of subgroup analysis to dependence models involving copulas, where the dependence parameters for an elliptical copula are regularized using pair-wise penalties similar to the one explored in this paper. Another avenue of future work may be the exploration of subgroup analysis for other distributions, including but not limited to the negative binomial, or zero-inflated versions of the frequency models often used in the actuarial literature. For example, in Section 2, Equation (2), Imagine that $\mathcal{P}(\nu)$ is a Poisson distribution with mean ν, although our approach is general it is possible to work with other discrete distributions including the binomial or negative binomial distributions. It may be a little trickier with the

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Tweedie distribution, as the density function requires a big summation, and exploring this may be interesting future work.

Competing interests. The authors have no competing interests to declare.

Data availability statement. The data and code that support the findings of this study are available from the corresponding author, Gee Y. Lee, upon request.

References

Akerlof, G. (1970). The market for lemons: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, **84**(3), 488–500.

Antonio, K. & **Beirlant, J.** (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, **40**(1), 58–76.

Boucher, J.-P. & **Denuit, M.** (2006). Fixed versus random effects in Poisson regression models for claim counts: A case study with motor insurance. *ASTIN Bulletin: The Journal of the IAA*, **36**(1), 285–301.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & **Eckstein, J.** (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, **3**(1), 1–122.

Chen, K., Huang, R., Chan, N. H. & **Yau, C. Y.** (2019). Subgroup analysis of zero-inflated Poisson regression model with applications to insurance data. *Insurance: Mathematics and Economics*, **86**, 8–18.

Deviendt, S., Antonio, K., Reynkens, T. & **Verbelen, R.** (2021). Sparse regression with multi-type regularized feature modeling. *Insurance: Mathematics and Economics*, **96**, 248–261.

Fan, J. & **Li, R.** (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456), 1348–1360.

Frees, E. W., Lee, G. & **Yang, L.** (2016). Multivariate frequency-severity regression models in insurance. *Risks*, **4**(1), 4.

Frees, E. W. & **Shi, P.** (2017). Credibility prediction using collateral information. *Variance*, **11**(1), 45–59.

Guo, L. (2003). *Applying data mining techniques in property/casualty insurance*. Forum of the Casualty Actuarial Society.

Jeong, H. (2020). Testing for random effects in compound risk models via Bregman divergence. *ASTIN Bulletin: The Journal of the IAA*, **50**(3), 777–798.

Jeong, H. & **Dey, D.** (2023). Multi-peril frequency credibility premium via shared random effects. *Variance*, Forthcoming. http://dx.doi.org/10.2139/ssrn.3825435

Jeong, H. & **Valdez, E. A.** (2020). Predictive compound risk models with dependence. *Insurance: Mathematics and Economics*, **94**, 182–195.

Ma, S. & **Huang, J.** (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, **112**(517), 410–423.

Norberg, R. (1986). Hierarchical credibility: Analysis of a random effect linear model with nested classification. *Scandinavian Actuarial Journal*, **1986**(3-4), 204–222.

Rothschild, M. & **Stiglitz, J.** (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics*, **90**(4), 629–649.

Shi, P. & **Lee, G. Y.** (2022). Copula regression for compound distributions with endogenous covariates with applications in insurance deductible pricing. *Journal of the American Statistical Association*, **117**(539), 1094–1109.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & **Knight, K.** (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, **67**(1), 91–108.

Tibshirani, R. & **Taylor, J.** (2011). The solution path of the generalized lasso. *The Annals of Statistics*, **39**(3), 1335–1371.

Tzougas, G. (2020). EM estimation for the Poisson-inverse gamma regression model with varying dispersion: An application to insurance ratemaking. *Risks*, **8**(3), 97.

Tzougas, G. & **Makariou, D.** (2022). The multivariate Poisson-generalized inverse gaussian claim count regression model with varying dispersion and shape parameters. *Risk Management and Insurance Review*, **25**(4), 401–417.

Wilson, C. (1977). A model of insurance markets with incomplete information. *Journal of Economic Theory*, **97**(2), 167–207.

Yeo, A., Smith, K., Willis, R. & **Brooks, M.** (2001). Clustering techniques for risk classification and prediction of claim costs in the automobile insurance industry. *Intelligent Systems in Accounting, Finance, and Management*, **10**(1), 39–50.

Appendix A: The original ADMM algorithm

For completeness, we briefly introduce the algorithm used in Chen et al. (2019), which considered a simpler data structure compared to the one studied in this paper. They neither considerd possible dependence among the lines of business nor serial dependence among the claims from

the same policyholder so that $J = T = 1$ in Equation (1) and $\theta_i = 1$ in Equation (2) of the main paper. The dataset that we studied is more complicated, because there are multiple coverage types $j = 1, \cdots, J$ and times $t = 1, \cdots, T$. Chen et al. (2019) target to maximize the Poisson likelihood with a fused LASSO-type penalty, which is given as follows:

$$\ell_p(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j | \mathcal{D}_{j,T}) = \tilde{\ell}(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j | \mathcal{D}_{j,T}) - \sum_{1 \le i < l \le I} p\left(|\delta_{jil}|; \tau_j\right) \tag{17}$$

subject to $\log \theta_{ji} - \log \theta_{jl} = \delta_{jil}$ for all $1 \le i < l \le I$, where

$$\tilde{\ell}(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j | \mathcal{D}_{j,T}) = \boldsymbol{N}_j' \boldsymbol{X} \boldsymbol{\alpha}_j + \boldsymbol{N}_j' \boldsymbol{U}_j \log(\boldsymbol{\theta}_j) - \boldsymbol{v}_j' \boldsymbol{U}_j \boldsymbol{\theta}_j - C_j \tag{18}$$

and

$$\boldsymbol{N}_j = (\boldsymbol{n}_{j1}', \boldsymbol{n}_{j2}', \cdots, \boldsymbol{n}_{jI}')' \quad \text{and} \quad \boldsymbol{n}_{ji} = (n_{ji1}, n_{ji2}, \cdots, n_{jiT})' \tag{19}$$

$$\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2', \cdots \boldsymbol{X}_I')' \quad \text{and} \quad \boldsymbol{X}_i = (\boldsymbol{x}_{i1}', \boldsymbol{x}_{i2}', \cdots \boldsymbol{x}_{iT}')' \tag{20}$$

$$\boldsymbol{U}_j = (\boldsymbol{U}_{j1}', \boldsymbol{U}_{j2}', \cdots \boldsymbol{U}_{jI}')' \quad \text{and} \quad \boldsymbol{U}_{ji} = (\boldsymbol{e}_{ji}', \boldsymbol{e}_{ji}', \cdots, \boldsymbol{e}_{ji}')_{T \times I}' \tag{21}$$

$$\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \cdots, \theta_{jI})' \tag{22}$$

$$\boldsymbol{v}_j = \exp\left(\boldsymbol{X} \boldsymbol{\alpha}_j\right) \tag{23}$$

$$\text{and } C_j = \sum_{i=1}^{I} \sum_{t=1}^{T} \log\left(n_{jit}!\right). \tag{24}$$

The elements of $\boldsymbol{e}_{ji} = (e_{ji1}, e_{ji2}, \cdots, e_{jiI})$ are basically vectors whose elements are all zeros except for one entry:

$$e_{jil} = \begin{cases} 1 & \text{if } i = l \\ 0 & \text{if } i \ne l \end{cases}$$

and the $\exp(\cdot)$ and $\log(\cdot)$ are element-wise functions in case the input is a vector. $p(\cdot\,; \tau)$ can be any concave penalty function and τ is the corresponding tuning parameter. Examples of penalty functions used are

(a) The LASSO penalty:

$$p(t; \tau) = \tau |t|, \tag{25}$$

(b) The MC (minimax concave) penalty:

$$p(t; \tau) = \tau \int_0^t (1 - x/(\kappa \tau))_+ dx, \quad \kappa > 1, \tag{26}$$

(c) The SCAD penalty:

$$p(t; \tau) = \tau \int_0^t \min\{1, (\kappa - x/\tau)_+ / (\kappa - 1)\} dx, \quad \kappa > 2. \tag{27}$$

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

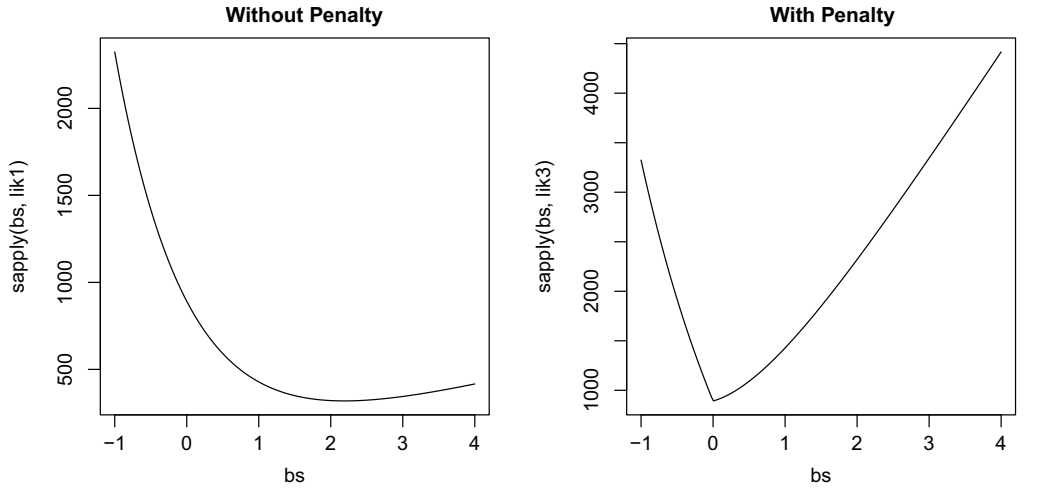

Figure A.1. An illustration of the potential difficulties of direct optimization.

One can observe that (17) is not smooth for certain values of the tuning parameter, which hinders direct maximization of (17) via the use of popular optimization routines for convex optimization. For example, with the LASSO penalty, one may observe that the negative log-likelihood function has a kink as in the right panel of Figure A.1, for certain values of the tuning parameter:

In that regard, Chen et al. (2019) used a version of ADMM algorithm that uses the following augmented Lagrangian as the objective function:

$$L(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j, \boldsymbol{\delta}_j, \boldsymbol{\lambda}_j | \mathcal{D}_{j,T}) = -\tilde{\ell}(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j | \mathcal{D}_{j,T}) + \sum_{1 \le i < l \le I} p\left(|\delta_{jil}|; \tau_j\right) + \overline{W}_j,$$

where $\overline{W}_j = \sum_{1 \le i < l \le I} \lambda_{jil}(\log \theta_{ji} - \log \theta_{jl} - \delta_{jil}) + \frac{\rho}{2} \sum_{1 \le i < l \le I} (\log \theta_{ji} - \log \theta_{jl} - \delta_{jil})^2$. Here λ_{jil} is Lagrange multiplier that corresponds to the equality constraints and ρ is a hyperparameter that controls smoothness of the augmented Lagrangian with an additional quadratic term. According to Boyd et al. (2011), addition of the quadratic term makes the augmented Lagrangian differentiable under mild conditions and facilitates better convergence of the algorithm.

This results in the naive version of the ADMM algorithm are shown below.
where

$$L(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*, \boldsymbol{\delta}_j^{(s)}, \boldsymbol{\lambda}_j^{(s)}) = -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*) + \sum_{i=2}^{I} p(|\delta_{ji}^{(s)}; \tau_j) + W_j^{(s)} \tag{28}$$

$$= -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*) + \frac{\rho}{2} \sum_{i=2}^{I} \left[\eta_{j,i}^* - \eta_{j,i-1}^* - \delta_{j,i-1}^{(s)} + \frac{\lambda_{j,i-1}^{(s)}}{\rho} \right]^2 + C, \tag{29}$$

where C is a constant, and

$$\delta_{j,i}^{(s+1)} = \begin{cases} ST(u_{j,i}, \tau_h/\rho) & \text{if } |u_{j,i}| \le \tau_h(1+\rho^{-1}) \\ \frac{ST(u_{j,i}, \kappa\tau_h/((\kappa-1)\rho))}{1-((\kappa-1)\rho)^{-1}} & \text{if } \tau_j(1+\rho^{-1}) \le |u_{j,i}| \le \kappa\tau_h \\ u_{j,i} & |u_{j,i}| > \kappa\tau_h, \end{cases} \tag{30}$$

$$\begin{aligned} u_{j,i} &= \eta_{j,i}^{(s+1)} - \eta_{j,i-1}^{(s+1)} + \rho^{-1}\lambda_{j,i}^{(s)}, \\ ST(u, \tau) &= \text{sgn}(u)(|u| - \tau)_+, \end{aligned} \tag{31}$$

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Algorithm 0: A naive ADMM algorithm for parameter estimation

1. Obtain the initial values $\boldsymbol{\alpha}_j^{(0)}$ and $\boldsymbol{\eta}_j^{*(0)}$ from Equations (4) and (5) of the main paper.
2. Set $\lambda_{j,i}^{(0)} = 0$ and $\delta_{j,i}^{*(0)} = \eta_{j,i}^{*(0)} - \eta_{j,i-1}^{*(0)}$ for $i = 1, \cdots, I-1$.
3. Repeat the following for $\tau_h = \tau_{min}, \cdots, \tau_{max}$, where τ_{min} is close to zero:
   (a) Repeat the following for $s = 0, 1, 2, \cdots$ until convergence:
      i. Let $\left(\boldsymbol{\alpha}_j^{(s+1)}, \boldsymbol{\eta}_j^{*(s+1)}\right) = \underset{\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j^*}{\arg\min}\, L\left(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*, \boldsymbol{\delta}_j^{(s)}, \boldsymbol{\lambda}_j^{(s)}\right)$ using Equation (28).
      ii. Obtain $\delta_{ij}^{(s+1)}$ using Equation (30).
      iii Obtain $\lambda_{ji}^{(s+1)}$ using Equation (32).
   (b). Use the current estimates of the parameters as initial values for the $\tau_j + 1^{th}$ step.

and

$$\lambda_{j,i}^{(s+1)} = \lambda_{j,i}^{(s)} + \rho\left(\eta_{j,i}^{*(s+1)} - \eta_{j,i-1}^{*(s+1)} - \delta_{j,i}^{(s+1)}\right). \tag{32}$$

To see the equality between Equations (28) and (29), note that expanding the right-hand side of (29), we see that

$$\begin{aligned}
L(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*, \boldsymbol{\delta}_j^{(s)}, \boldsymbol{\lambda}_j^{(s)}) &= -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*) + \frac{\rho}{2}\sum_{i=2}^{I}\left[\eta_{j,i}^* - \eta_{j,i-1}^* - \delta_{j,i}^{(s)} + \frac{\lambda_{j,i-1}^{(s)}}{\rho}\right]^2 + C \\
&= -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*) + \frac{\rho}{2}\sum_{i=2}^{I}(\eta_{j,i}^* - \eta_{j,i-1}^* - \delta_{j,i}^{(s)})^2 + \frac{\rho}{2}\sum_{i-2}^{I}\left(\frac{\lambda_{j,i-1}^{(s)}}{\rho}\right)^2 \\
&\quad + 2 \times \frac{\rho}{2}\sum_{i=2}^{I}\frac{\lambda_{j,i-1}^{(s)}}{\rho}(\eta_{j,i}^* - \eta_{j,i-1}^* - \delta_{j,i}^{(s)}) + C \\
&= -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*) + \frac{\rho}{2}\sum_{i=2}^{I}(\eta_{j,i}^* - \eta_{j,i-1}^* - \delta_{j,i}^{(s)})^2 + \frac{\rho}{2}\sum_{i-2}^{I}\left(\frac{\lambda_{j,i-1}^{(s)}}{\rho}\right)^2 \\
&\quad + \sum_{i=2}^{I}\lambda_{j,i-1}^{(s)}(\eta_{j,i}^* - \eta_{j,i-1}^* - \delta_{j,i}^{(s)}) + C \\
&= -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\eta}_j^*) + \sum_{i=2}^{I} p(|\delta_{ji}^{(s)}|; \tau_j) + W_j^{(s)}
\end{aligned}$$

where, the last equality holds if

$$C = \sum_{i=2}^{I} p(|\delta_{ji}^{(s)}|; \tau_j) - \frac{\rho}{2}\sum_{i-2}^{I}\left(\frac{\lambda_{j,i-1}^{(s)}}{\rho}\right)^2$$

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Note that terms with superscript (s) can all be considered constant. Here κ controls the shape of SCAD penalty, which was set to 3.7 in the original work of Fan and Li (2001), but also could be determined by cross-validation. τ_h corresponds to the magnitude of the penalty term so that as τ_h increases $\delta_{j,i}^{(s)}$ is more likely to shrink to zero, which implies the discrepancy $\eta_{j,i}^{*(s+1)}$ and $\eta_{j,i-1}^{*(s+1)}$ also shrinks to zero. The choice of ρ determines the smoothness of the objective function, and it is left to the modeler. In our case, it has been set to 5 after trial and error. We acknowledge that we do not have a statistical approach to selecting the optimal ρ (neither did the orignal authors of the ADMM paper) and leave this as future work. The real problem with Algorithm 0 is that the solution path is yet not ideal, because the coefficients within the same group do not merge completely, which gives rise to need for a more efficient algorithm as proposed in the main paper.

Appendix B: Derivation of the gradient and Hessian

For fast computation of the ADMM algorithm, an analytic expression for the gradient and Hessian of the likelihood function is required. Here, we replicate Equation (18) and the ADMM objective function.

$$L(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j^*, \boldsymbol{\delta}_j, \boldsymbol{\lambda}_j | \mathcal{D}_{j,T}) = -\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j^* | \mathcal{D}_{j,T}) + \sum_{i=2}^{I} p\left(|\delta_{ji}|; \tau_j\right) + W_j,$$

where

$$\tilde{\ell}^*(\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j^* | \mathcal{D}_{j,T}) = \boldsymbol{N}_j' \boldsymbol{X} \boldsymbol{\alpha}_j + \boldsymbol{N}_j' \boldsymbol{U}_j \log\left(\boldsymbol{\theta}_j^*\right) - \boldsymbol{v}_j' \boldsymbol{U}_j \boldsymbol{\theta}_j^*.$$

To write the gradient down, one can first observe that W_j can be written as a function of $\eta_{jk}^* := \log \theta_{jk}^*$ as follows:

$$W_j(\eta_{jk}^*) = \begin{cases} \lambda_{j,k-1}^*(\eta_{jk}^* - \eta_{j,k-1}^* - \delta_{j,k-1}^*) + \frac{\rho}{2}(\eta_{jk}^* - \eta_{j,k-1}^* - \delta_{j,k-1}^*)^2 + C_{jk}^* & \text{if } k = I \\ \lambda_{j,k-1}^*(\eta_{jk}^* - \eta_{j,k-1}^* - \delta_{j,k-1}^*) + \lambda_{jk}^*(\eta_{j,k+1}^* - \eta_{jk}^* - \delta_{jk}^*) & \\ \quad + \frac{\rho}{2}(\eta_{jk}^* - \eta_{j,k-1}^* - \delta_{j,k-1}^*)^2 + \frac{\rho}{2}(\eta_{j,k+1}^* - \eta_{jk}^* - \delta_{jk}^*)^2 + C_{jk}^* & \text{if } 1 < k < I \\ \lambda_k(\eta_{j,k+1}^* - \eta_{,k}^* - \delta_{jk}^*) + \frac{\rho}{2}(\eta_{j,k+1}^* - \eta_{jk}^* - \delta_{jk}^*)^2 + C_{jk}^* & \text{if } k = 1 \end{cases} \tag{33}$$

where $\partial C_{jk}^* / \partial \eta_{jk}^* = 0$ and it leads to

$$\frac{\partial W_j}{\partial \eta_{jk}^*} = \begin{cases} \rho\left(\eta_{jk}^* - \eta_{j,k-1}^* - \delta_{j,k-1}^* + \frac{\lambda_{j,k-1}^*}{\rho}\right) & \text{if } k = I \\ \rho\left[\left(\eta_{jk}^* - \eta_{j,k-1}^* - \delta_{j,k-1}^* + \frac{\lambda_{j,k-1}^*}{\rho}\right) - \left(\eta_{j,k+1}^* - \eta_{jk}^* - \delta_{jk}^* + \frac{\lambda_{jk}^*}{\rho}\right)\right] & \text{if } 1 < k < I \\ -\rho\left(\eta_{j,k+1}^* - \eta_{jk}^* - \delta_{jk}^* + \frac{\lambda_{jk}^*}{\rho}\right) & \text{if } k = 1 \end{cases} \tag{34}$$

and

$$\frac{\partial^2 W_j}{\partial \eta_{jk}^* \partial \eta_{jk'}^*} = \begin{cases} \rho\left[\mathbb{1}_{\{k'=k\}} - \mathbb{1}_{\{k'=k-1\}}\right] & \text{if } k = I \\ \rho\left[2 \cdot \mathbb{1}_{\{k'=k\}} - \mathbb{1}_{\{k'=k-1\}} - \mathbb{1}_{\{k'=k+1\}}\right] & \text{if } 1 < k < I \\ \rho\left[\mathbb{1}_{\{k'=k\}} - \mathbb{1}_{\{k'=k+1\}}\right] & \text{if } k = 1 \end{cases} \tag{35}$$

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

It is also easy to observe that

$$\frac{\partial \tilde{\ell}^*}{\partial \boldsymbol{\alpha}_j} = \sum_{i=1}^{I}\sum_{t=1}^{T}(n_{jit} - \theta_{ji}^* \nu_{jit})\boldsymbol{x}_{it} = \boldsymbol{N}_j'\boldsymbol{X} - \boldsymbol{X}'\text{diag}\left(\boldsymbol{\nu}_j'\right)\boldsymbol{U}_j\boldsymbol{\theta}_j^*, \tag{36}$$

$$\frac{\partial \tilde{\ell}^*}{\partial \eta_{jk}^*} = \sum_{t=1}^{T}(n_{jkt} - \nu_{jkt}\theta_{jk}^*), \ \frac{\partial \tilde{\ell}}{\partial \boldsymbol{\eta}_j^*} = N_j'\boldsymbol{U}_j - \boldsymbol{\nu}_j'\boldsymbol{U}_j\text{diag}(\boldsymbol{\theta}_j^*), \tag{37}$$

$$\frac{\partial^2 \tilde{\ell}^*}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j'} = -\sum_{i=1}^{I}\sum_{t=1}^{T}\theta_{ji}^*\nu_{jit}\boldsymbol{x}_{it}\boldsymbol{x}_{it}' = -\boldsymbol{X}'\text{diag}(\boldsymbol{\nu}_j')\cdot\text{diag}(\boldsymbol{U}_j\boldsymbol{\theta}_j^*)\boldsymbol{X}, \tag{38}$$

$$\frac{\partial^2 \tilde{\ell}^*}{\partial \boldsymbol{\alpha}_j \partial \eta_{jk}^*} = -\sum_{t=1}^{T}\nu_{jkt}\boldsymbol{x}_{kt}\theta_{jk}^*, \ \frac{\partial^2 \tilde{\ell}^*}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\eta}_j^{*'}} = -\boldsymbol{X}'\text{diag}(\boldsymbol{\nu}_j')\boldsymbol{U}_j\text{diag}(\boldsymbol{\theta}_j^*), \tag{39}$$

$$\frac{\partial^2 \tilde{\ell}^*}{\partial \boldsymbol{\eta}_j^* \partial \boldsymbol{\eta}_j^{*'}} = \text{diag}\left(-\boldsymbol{\nu}_j'\boldsymbol{U}_j\text{diag}(\boldsymbol{\theta}_j^*)\right), \tag{40}$$

Using these expressions for the optimization routine helps improve the convergence of the algorithm and reduces computation time significantly.

Appendix C: Modified gradient and Hessian

With the modified algorithm (Algorithm 1 in the main paper), we simply use the fact that we have

$$\boldsymbol{\xi}_j^* = \boldsymbol{G}_j\boldsymbol{\eta}_j^*, \tag{41}$$

so that using chain rule, we may write the gradient and Hessian in terms of $\boldsymbol{\xi}_j^*$:

$$\frac{\partial^2 \tilde{\ell}^*}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\xi}_j^{*'}} = \left(\frac{\partial^2 \tilde{\ell}^*}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\eta}_j^{*'}}\right)\boldsymbol{G}_j', \tag{42}$$

$$\frac{\partial^2 \tilde{\ell}^*}{\partial \boldsymbol{\xi}_j^* \partial \boldsymbol{\xi}_j^{*'}} = \boldsymbol{G}_j\left(\frac{\partial^2 \tilde{\ell}^*}{\partial \boldsymbol{\eta}_j^* \partial \boldsymbol{\eta}_j^{*'}}\right)\boldsymbol{G}_j' \tag{43}$$

Note that the rows of $\boldsymbol{G}_j$ are indicators of the elements within a specific group of coefficients, and the length of $\boldsymbol{\xi}_j^*$ is equal to the number of unique coefficients left inside the model.

Appendix D: Dependence cases

In the main text of the paper, three different dependence cases are considered. The details of these three cases are explained in full detail here.

Case 1. No interdependence among the lines

In the first case, we do not assume any type of interdependence among $(N_{1it}, \ldots, N_{Jit})$ for each of $i = 1, \ldots, I$ and $t = 1, \ldots, T$ and estimate $\boldsymbol{\alpha}_j$ and θ_{ji} only using the observations of claims and

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

policy characteristics from jth line of business. In other words, we use the following longitudinal dataset for estimation of $\boldsymbol{\alpha}_j$ and θ_{ji} and no inter-line dependence is considered:

$$\mathcal{D}_{j,T} = \left\{ \left(n_{jit}, \mathbf{x}_{it}\right) \;\middle|\; i = 1, \ldots, I,\ t = 1, \ldots, T \right\}. \tag{44}$$

In this case, we do not consider the effect of θ_i. For this reason, it may be thought of as if $\theta_i = 1$ is fixed. There are two subcases to the no-interdependence case:

Case 1a. When θ_{ji} is random

We assume θ_{ji} follows a gamma distribution with unit mean. More specifically, $\theta_{ji} \sim \mathcal{G}(r_j, 1/r_j)$ so that $\mathbb{E}[\theta_{ji}] = 1$ and $Var[\theta_{ji}] = 1/r_j$. Due to the conjugacy of Poisson and gamma distributions (see Jeong & Valdez, 2020), it turns out that

$$\mathbb{E}[\theta_{ji}|\mathcal{D}_{j,T}] = \frac{r_j + \sum_{t=1}^{T} n_{jit}}{r_j + \sum_{t=1}^{T} \nu_{jit}}, \tag{45}$$

which captures the posterior expectation of the unobserved heterogeneity of the policyholder i on coverage j. Note that r_j can be determined in two ways. First, r_j can be interpreted as degree of dispersion in θ_{ji}, which is unobserved heterogeneity in the claim frequency. In this regard, one can assign a value for r_j so that the range (for example, 95% highest posterior density interval) of θ_{ji} can be in certain level as discussed in Page 6 of Jeong (2020).

Second, one can use a method of moments estimator to retrieve the value of r_j. If we assume that $\theta_i = 1$ and $\theta_{ij} \sim \mathcal{G}(r_j, 1/r_j)$ where θ_{ij} are independent of each other for all $i = 1, \ldots, I$ and fixed j, then $\mathbb{E}\left[\sum_{t=1}^{T} N_{jit}\right] = \sum_{t=1}^{T} \nu_{jit}$ and $Var\left[\sum_{t=1}^{T} N_{jit}\right] = \sum_{t=1}^{T} \nu_{jit} + \frac{(\sum_{t=1}^{T} \nu_{jit})^2}{r_j}$ so that $\mathbb{E}[z_{ji}] = 1/r_j$ and z_{ji} $(i = 1, \ldots, I)$ are independent observations where

$$z_{ji} = \frac{(\sum_{t=1}^{T} N_{jit} - \sum_{t=1}^{T} \nu_{jit})^2 - \sum_{t=1}^{T} \nu_{jit}}{(\sum_{t=1}^{T} \nu_{jit})^2},$$

which lead to the following estimate of r_j:

$$\hat{r}_j = \frac{I}{\sum_{i=1}^{I} z_{ji}}.$$

Once r_j is specified in either way, the predicted claim frequency can be written as follows:

$$\begin{aligned} \hat{N}_{ji,T+1} &= \mathbb{E}[N_{ji,T+1}|\mathcal{D}_{j,T}] = \hat{\nu}_{ji,T+1}\mathbb{E}[\theta_{ji}|\mathcal{D}_{j,T}] \cdot \mathbb{E}[\theta_i|\mathcal{D}_{j,T}] \\ &= \exp(\boldsymbol{x}'_{i,T+1}\hat{\boldsymbol{\alpha}}_j) \cdot \frac{r_j + \sum_{t=1}^{T} n_{jit}}{r_j + \sum_{t=1}^{T} \hat{\nu}_{jit}}, \end{aligned} \tag{46}$$

where we have assumed $\theta_i = 1$ with probability 1, which is basically an assumption that the claims experience from each line is independent from one another.

Note that one can use different distributions instead of gamma distribution to describe the behavior of random θ_{ji} as long as $\mathbb{E}[\theta_{ji}] = 1$ for an identifiability issue. For more details on the use of distributions for random θ_{ji} other than gamma when $N_{jit}|\theta_{ji}$ follows a Poisson distribution, please see Tzougas (2020) and Tzougas and Makariou (2022).

Case 1b. When θ_{ji} is nonrandom

In this case, there is no assumed structure in θ_{ji}, so we may consider finding the point estimates of $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jI})$ via the proposed ADMM algorithm. For the identifiability issue, θ_{j1} was set 0 for all j.

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

After the estimates of $\boldsymbol{\alpha}_j$ and θ_{ji} are obtained, the predicted claim frequency will be written as follows:

$$\hat{N}_{ji,T+1} = \exp(\boldsymbol{x}'_{i,T+1}\hat{\boldsymbol{\alpha}}_j) \cdot \hat{\theta}_{ji}. \tag{47}$$

Case 2. Perfect interdependence among the lines

In this case, we assume there is only a common unobserved heterogeneity factor for each policyholder so that $\theta_{1i} = \cdots = \theta_{Ji} = 1$. In this case, we have

$$N_{jit}|\boldsymbol{x}_{it}, \theta_i \sim \mathcal{P}\left(\nu_{jit}\theta_i\right), \tag{48}$$

for all i, t, and j and we estimate $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_J, \boldsymbol{\theta}_i)$ simultaneously using the multiline dataset.

Case 2a. When θ_i is random

We assume that $\theta_i \sim \mathcal{G}(r, 1/r)$ so that $\mathbb{E}[\theta_i] = 1$ and $\mathrm{Var}[\theta_i] = 1/r$. Due to the conjugacy of Poisson and gamma distributions (see Jeong & Dey, 2023), it turns out that

$$\mathbb{E}[\theta_i|\mathcal{D}_T] = \frac{r + \sum_{j=1}^{J}\sum_{t=1}^{T} n_{jit}}{r + \sum_{j=1}^{J}\sum_{t=1}^{T} \nu_{jit}}, \tag{49}$$

where $\mathcal{D}_T = \left\{ \left(n_{1it}, \ldots, n_{jit}, \ldots, n_{Jit}, \mathbf{x}_{it}\right) \,\middle|\, i = 1, \ldots, I,\ t = 1, \ldots, T \right\}$. It captures the pooled posterior expectation of the common unobserved heterogeneity of the policyholder i. Therefore, the predicted claim frequency will be written as follows:

$$\begin{aligned} \hat{N}_{ji,T+1} &= \mathbb{E}[N_{ji,T+1}|\mathcal{D}_T] = \hat{\nu}_{ji,T+1}\mathbb{E}[\theta_i|\mathcal{D}_T] \\ &= \exp(\boldsymbol{x}'_{i,T+1}\hat{\boldsymbol{\alpha}}_j) \cdot \frac{r + \sum_{j=1}^{J}\sum_{t=1}^{T} n_{jit}}{r + \sum_{j=1}^{J}\sum_{t=1}^{T} \nu_{jit}}. \end{aligned} \tag{50}$$

Case 2b. When θ_i is nonrandom

In this case, there is no assumed distributional structure in θ_i so we find the point estimates of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)$ by maximizing the following penalized likelihood:

$$\begin{aligned} \tilde{\ell}_p(\boldsymbol{\alpha}, \boldsymbol{\theta}|\mathcal{D}_T) = &\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{t=1}^{T} \left(n_{jit}(\boldsymbol{x}'_{it}\boldsymbol{\alpha}_j + \log(\theta_i^*)) - \nu_{jit}\theta_i^* - \log(n_{jit})\right) \\ &- \sum_{i=2}^{I} p\left(|\log\theta_i^* - \log\theta_{i-1}^*|; \tau\right). \end{aligned}$$

Since the only difference between the above penalized likelihood and (3) in the main paper is that θ_{ji} are replaced with θ_i, one can use the proposed ADMM algorithm with modest modification to estimate $\boldsymbol{\alpha}, \boldsymbol{\theta}$. Note that (48) is already of full-rank as long as $J \geq 2$ or $J \geq 2$ for each i, which is easiliy satisfied in our simulation study and empirical analysis so that they is no identification issue either. In that case, the predicted claim frequency will be written as follows:

$$\hat{N}_{ji,T+1} = \exp(\boldsymbol{x}'_{i,T+1}\hat{\boldsymbol{\alpha}}_j) \cdot \hat{\theta}_i.$$

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Case 3. Flexible interdependence

In the third case, we have

$$N_{jit}|\boldsymbol{x}_{it}, \boldsymbol{\theta}, \theta_i \sim \mathcal{P}\left(v_{jit}\theta_{ji}\theta_i\right), \tag{51}$$

where θ_i accounts for the common unobserved heterogeneity of policyholder i shared by all coverages. This is an experimental two-step approach, where we boost the risk classification model using common effects θ_i. The main idea of boosting the risk classification model stepwise is to add common unobserved heterogeneity upon observed covariates and coverage-specific unobserved heterogeneity. We assume θ_{ji} is nonrandom but θ_i is random. The step-wise approach is the following:

1. We find the nonparametric estimates of θ_{ji} for $j = 1, \ldots, J$ and $i = 1, \ldots, I$, as in Case 1 (when θ_{ji} are nonrandom), while θ_i is not considered in this step. In other words, $\theta_i = 1$ with probability 1.
2. After that, we assume that θ_i follows $\mathcal{G}(r, 1/r)$, which leads to the following posterior distribution and Bayesian premium:

$$\pi(\theta_i|\mathcal{D}_{j,T}, \boldsymbol{\theta}) \sim \mathcal{G}\left(\sum_{t=1}^{T}\sum_{j=1}^{J} N_{jit}\hat{\theta}_{ji} + r,\ \left[\sum_{t=1}^{T}\sum_{j=1}^{J} v_{jit}\hat{\theta}_{ji} + r\right]^{-1}\right), \tag{52}$$

$$\mathbb{E}[N_{ji,T+1}|\mathcal{D}_{j,T}] = v_{ji,T+1}\hat{\theta}_{ji}\mathbb{E}[\theta_i|\mathcal{D}_{j,T}] = v_{ji,T+1}\hat{\theta}_{ji}\frac{\sum_{t=1}^{T}\sum_{j=1}^{J} N_{jit}\hat{\theta}_{ji} + r}{\sum_{t=1}^{T}\sum_{j=1}^{J} v_{jit}\hat{\theta}_{ji} + r}, \tag{53}$$

where $\hat{\theta}_{ji}$ are estimated in the first step. In this case, we assume there are *unobserved heterogeneity* factors of the coverages for each policyholder that are intertwined in a way.

Appendix E: Tables

Table E.1. Computation time for simulated Coverage 1 (in sec)

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
GLM without individual effects	0.01	0.01	0.01	0.01
Individual effects model	68.97	58.08	58.59	83.38
Common effects model	49.70	46.40	73.70	71.41
Boosted credibility model	68.99	58.10	58.61	83.40

Table E.2. Computation time for simulated Coverage 2 (in sec)

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
GLM without individual effects	0.01	0.01	0.01	0.01
Individual effects model	71.57	58.69	83.64	81.62
Common effects model	49.70	46.40	73.70	71.41
Boosted credibility model	71.59	58.71	83.66	81.64

https://doi.org/10.1017/S1748499523000271 Published online by Cambridge University Press

Table E.3. Computation time for simulated Coverage 3 (in sec)

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
GLM without individual effects	0.01	0.01	0.01	0.01
Individual effects model	67.85	59.94	93.02	81.31
Common effects model	49.70	46.40	73.70	71.41
Boosted credibility model	67.87	59.96	93.04	81.33

Cite this article: Lee GY and Jeong H (2024). Nonparametric intercept regularization for insurance claim frequency regression models, *Annals of Actuarial Science*, **18**, 579–604. https://doi.org/10.1017/S1748499523000271