Market-based Model Selection with an Application in Commercial Auto Ratemaking

Brian Hartman[a,*], Chris Dixon[a], Alisa Walch[b]

[a] *Department of Statistics, Brigham Young University, Provo, UT, USA*
[b] *Department of Mathematics, University of Texas, Austin, TX, USA*

Abstract

It is often difficult to know which models to use when setting rates for auto insurance. We develop a market-based model selection procedure which incorporates the goals of the business. Additionally, it is easier to interpret the results and better understand the models and data. As an application, we compare many different models (variants of GLMs, GLMMs, regularized regression, random forest, and spike and slab models) on a robust dataset of 70,000 commercial auto policies. We first compare the models using standard metrics (MAPE, MSPE, and runtime), leading to mixed results and interpretation. The market-based model comparison method shows that the random forest model far outperforms the other models in terms of both loss ratio and market share, likely compensating for the increased computational cost.

1. Introduction

In this paper, we propose a market-based model selection method which focuses on the goals of the prediction exercise, to optimally price auto policies. To show the potential benefits of the method, we apply it in a commercial auto insurance ratemaking exercise. We fit many different predictive models to a robust dataset of 70,000 commercial auto policies and compare the model performance.

The ratemaking literature is vast. When modeling pure premium in a property/casualty setting, there are two main streams of research. One option is to model the pure premium directly. Observed pure premium has many zeros for all the policies with no claims, but the positive values are relatively continuous. The Tweedie distribution (Tweedie, 1984) is commonly chosen in this setting because (depending on a power parameter between 1 and 2) it has a point mass at zero and continuous support over the positive reals. Jørgensen and Paes De Souza (1994) first applied the Tweedie distribution as a GLM to insurance claim data. The other option is to model the frequency (count of claims per policy) and severity (cost per claim) separately and then combine them to get the total cost. Some more recent work (e.g. Shi et al., 2016) also incorporates the type of claim (liability or physical damage, for example). We model the frequency, severity, and the pure premium, but will focus on pure premium for our market-based method (it can be applied to frequency and severity, but is most natural when comparing pure premium models).

After choosing how to model the total losses, we need to compare the different available models. When examining many potential ratemaking models, it can be difficult to choose the best model. Some models are more accurate by one measure, while other models are more accurate under a different measure. Even if a model is consistently more accurate, what if it is much more computationally intense, or if its optimization more unstable? In this paper we propose a market-based model comparison technique which can give the user a better understanding of the potential impact of different ratemaking models on the company results. We compare eight different models of pure premium both using standard accuracy and complexity measures

[*] Corresponding author

Email addresses: hartman@stat.byu.edu (Brian Hartman), chrisbr.dixon@gmail.com (Chris Dixon), ahavens@math.utexas.edu (Alisa Walch)

and using our market-based model comparison metric. The market-based metric provides a much clearer picture of the differences between the models.

Many papers use likelihood-based methods like AIC or BIC to choose the best models (e.g. Bermúdez and Karlis, 2011; Shi, 2016; Tang et al., 2014; Yip and Yau, 2005). Others use prediction error metrics like mean squared or median absolute prediction error or the GINI index (e.g. Frees et al., 2014; Guelman, 2012). Still others use both (e.g. Klein et al., 2014). Our market-based model selection method provides additional information to help answer the questions of which model is best and how much better.

In the next section, we outline market-based model selection and the other comparable metrics. In section 3, we describe our data. In section 4, we outline the predictive models we will compare. In section 5, we detail the results of our analysis. In section 6, we conclude and discuss limitations and next steps.

2. Market-based Model Selection Metric

The main contribution of this paper is a market-based model selection metric. Imagine that each of the candidate models is an insurance provider. Each of the pure premium predictions becomes the price the provider offers for each policy. The company with the lowest price earns the policy. We will then compare market share and loss ratios for the providers to see which model best differentiates the good risks from the poor ones. For example, if we are comparing five different models and their (cross-validated) predictions for a single observation (Table 1).

Model	Predicted Pure Premium	Actual Losses
Model 1	29	24
Model 2	27	24
Model 3	29	24
Model 4	31	24
Model 5	35	24

Table 1: Model selection example

Because model 2 has the lowest prediction, it will get to write the policy. It receives 27 in premium and 24 in losses. All of the rest of the models are unaffected. This does not depend on the actual earned premium for this policy in our data. This process is repeated for all of the observations. Then, to compare the models we compute the loss ratio (total losses/total earned premium) and the market share (policies written by the model/total policies). This is also compared for each of the 100 iterations to understand the uncertainty.

In addition to the market-based model comparison, we compare the models using median absolute prediction error (MAPE, the median of the absolute differences between the predicted and actual pure premium), mean squared prediction error (MSPE, the mean of the squared differences), and runtime. All of the computing was performed on an Intel Xeon E5-2689 v4 (3.10GHz) with 192 GB of RAM and 40 available threads. To get a better understanding of the prediction metrics, we ran the cross-validation 100 times. Intervals of the middle 95 iterations are also displayed.

3. Commerical Ratemaking Data

Our dataset consists of bodily injury (BI) and property damage (PD) claims from 77,229 commercial auto policyholders in California in 1999. We will use the following covariates to model frequency, severity, and pure premium.

- Program (standard, substandard, assigned risk, direct excess, and motorcycle)
- Good driver discount (true/false)
- Limits (basic, greater than basic)

- Coverage (bodily injury or property damage)
- Five policy level covariates (named Var1 through Var5, including some continuous and some categorical)

Some of the policy-level variables are included not because we think it is likely they will accurately predict losses, but rather because we hope that it will be removed from the model (or its impact will be minimized) by some of the methods we try below. In this paper we are mainly interested in different model families how to measure which perform best. Discussion of the chosen covariates (and their impact on losses) is deliberately avoided to maintain that focus.

4. Predictive Models

As mentioned in the previous section, we will model frequency, severity, and pure premium. We chose the models below after a review of the most common ratemaking models in the literature. We hope to focus on the model comparison process and also make our models easily replicable by practitioners who would like to apply our methods in their work. To that end, we will only use models which have a published R package. The methods we use in this paper are detailed in the remainder of this subsection.

Basic GLMs. Generalized linear models are a natural starting point. They extend the linear model when Gaussian additive errors don't make sense. Using the `glm` function in base R, we fit basic GLMs as a benchmark for all of the loss metrics. For frequency, we use a Poisson GLM with exposure offset. For severity, we use a Gamma GLM. Finally, for pure premium, we use a Tweedie distribution with exposure offset. For more information about GLMs, please see Nelder and Wedderburn (1972).

Bayesian GLMs. While the `glm` function chooses the parameters with a likelihood-based optimization, Bayesian GLMs use a prior distribution and then sample from the posterior distributions of the parameters using MCMC. These models are ideal for incorporating expert information not included in the data and for properly accounting for all types of uncertainty. In this application, we want a relatively fair comparison between our Bayesian and Frequentist models, so we will use basic non-informative priors on all the parameters. Using the `arm` package, we will fit basic Bayesian GLMs for all of the loss metrics. Similar to the basic GLMs, we use Poisson (with exposure offset) for frequency, gamma for severity, and Tweedie for pure premium (with exposure offset). For more details, please see Dey et al. (2000).

GLMMs. Using the `lme4` package, we fit generalized linear mixed models for all of the loss metrics. The main reason to model a covariate as a random effect instead of a fixed effect is to take advantage of partial pooling (Gelman and Hill, 2006). When you have a categorical variable where some levels have only a few entries, a random effect will shrink that level's coefficient estimate towards the estimates for the other levels. As is the case in the GLM models, we use Poisson for frequency, gamma for severity, and Tweedie for pure premium. We tried first to fit all the categorical variables (program, good driver discount, limits, coverage, and city type) as fixed effects and all the continuous variables as random effects, but the likelihood function became so flat that the model was unable to converge. We then simplified the model to still include all of the fixed effects, but only program as a random effect. This is important to note when reading the results section as the model we use will fit slightly faster and may perform more poorly than the other models because of the limited covariates. For more information on GLMMs, please see McCulloch and Searle (2004).

Regularization (LASSO, Elastic Net, and Ridge). Regularized methods aim to automatically find the balance between overly complex and overly simple models. In all cases, LASSO (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), and ridge (Hoerl and Kennard, 1970), the models are penalized for having large regression coefficients. This makes the models as simple as they can be but no simpler. If there is a large amount of signal for a large regression coefficient, it will overwhelm the penalty term and be included in the model. Elastic net and LASSO can also perform variable selection by penalizing the coefficients of the insignificant variables all the way to zero.

Using the `glmnet` package, we fit Poisson LASSO, elastic net, and ridge regression models on the frequency data with an exposure offset. To fit the severity and pure premium models (gamma and Tweedie with an exposure offset), we use the `HDTweedie` package. When testing and scrutinizing the `HDTweedie` package, we found some concerning results. Setting the shrinkage parameter (λ) to 0 removes the regularization from the model making it a standard GLM. We set the power parameter to 1 (implying a Poisson distribution) and compare the `HDTweedie` to `glm` and find that all the parameter estimates are equal. When we adjust the power parameter, the estimates no longer match, though they should. We decided to still include these results in the paper, but they should be taken with a grain of salt.

Spike and Slab. As an alternative to regularized regression to determine which variables are important, we fit a Bayesian Spike and Slab model. A standard regression model depends on some function of $\sum_k X_{ik}\beta_k$ for observation i. The spike and slab model (Kuo and Mallick, 1998) adds a binary variable to the kernel in the previous sentence to get to $\sum_k \gamma_k X_{ik}\beta_k$. When $\gamma_k = 1$ then X_k is included in the model, but X_k is left out when $\gamma_k = 0$. This model provides a posterior probability that β_k is significantly different from 0 (the probability that $\gamma_k = 1$). In contrast to the other regularization methods above, the spike and slab model does not shrink the variables which stay in the model much, but shrinks the others all the way to zero.

To fit the spike and slab model, we use the package `SpikeSlabGAM`. That package implements the Poisson model, but not the gamma or Tweedie models. The `boral` package states that it implements Tweedie spike and slab models, but we were unable to get reasonable results. Because of this, we decided to use the spike and slab model only in the frequency predictions.

Random Forests. Random forest models are an ensemble of many relatively simple models on subsets of the data. By combining the estimates from many simple models, we can get better results than those from a single complicated model. Each of the individual models are regression trees (Breiman, 2017). Each of the individual trees are built on a random sample of the observations and a random sample of the possible covariates. We use the `randomForest` package to fit random forest models. For the frequency model we include the exposure as a covariate. That is similar to including it as an offset in a standard GLM, though we are using more external information than in the GLM where the coefficient is effectively constrained to be one. The severity values are already per claim and do not need to be adjusted for exposure. In the pure premium model, we first divide the total losses by the exposure and then model that (per exposure) loss amount. When fitting a random forest model, several hyperparameters will need to be set. The main hyperparameters are:

- Number of trees to grow: This is the number of simple models (trees) to fit. Because each model is fit to a subset of the explanatory variables and a subset of the observations, a sufficient number of trees need to be built. This is only really limited by your computational power. As we will show later in the paper, random forest models can be computationally intense.
- Number of explanatory variables randomly sampled for each split in the tree: How many variables should we use in each simple model. The default is one-third of the total number of variables.
- Sample size of each draw: Number of samples drawn for each tree.
- Minimum size of the terminal nodes: Making this value larger will cause smaller trees to be grown. The default value is 5.

The rest of the settings we simply used the defaults. The four in the list we set through cross-validation as described in the next section. For more details on random forests, please see Breiman (2001).

4.1. Model Fitting

In summary, we fit the following models.

Frequency	Severity	Pure Premium
Poisson GLM	Gamma GLM	Tweedie GLM
Bayesian Poisson GLM	Bayesian Gamma GLM	Bayesian Tweedie GLM
Poisson GLMM	Gamma GLMM	Tweedie GLMM
Poisson LASSO	Gamma LASSO	Tweedie LASSO
Poisson Ridge	Gamma Ridge	Tweedie Ridge
Poisson Elastic Net	Gamma Elastic Net	Tweedie Elastic Net
Poisson Spike and Slab (SaS)	-	-
Random Forest	Random Forest	Random Forest

We fit all of the above models using ten-fold cross validation (we divide the dataset into ten parts and hold-out one at a time while fitting the model on the other parts). For the models without hyperparameters to specifically tune (GLMs, GLMMs, and spike and slab), we simply fit the models to 9/10 of the data and then used that model to predict the held out tenth. For each set of predictions, we iterate over each tenth. That way, all of the predictions are on out-of-sample data.

For the other models (LASSO, ridge, elastic net, and random forest) we used nested cross validation where the bottom level determined the hyperparameter settings. Specifically, we used the 9/10 of the data to first tune the hyperparameters (λ in the regularization methods and various hyperparameters for the random forest model). We did that through 10-fold cross-validation on the 9/10 of the original dataset. That way, we choose the hyperparameters that work best when holding out some data, and then use those hyperparameters to fit the model on the entire 9/10. Finally, we use that fitted model to predict the values for the held out tenth, data which the model has not seen yet. This method, while computationally expensive (because of the nested cross-validations), allows us to use the entire dataset more fully and still maintain proper holdout samples.

5. Results

5.1. Frequency

The frequency models produce the following results:

	Runtime (in mins)	MAPE Confidence Interval	MSPE Confidence Interval
Poisson GLM	0.26	(0.3179, 0.3194)	(116.58, 117.28)
Bayesian Poisson GLM	0.64	(0.3179, 0.3196)	(116.52, 117.26)
Poisson GLMM	4.11	(0.3349, 0.3362)	(136.26, 136.66)
Poisson LASSO	1.13	(0.3222, 0.3235)	(116.58, 117.32)
Poisson Ridge	0.55	(0.4906, 0.4919)	(126.75, 127.52)
Poisson Elastic Net	1.16	(0.3221, 0.3236)	(116.55, 117.29)
Poisson SaS	9.90	(0.7776, 0.7804)	(383.11, 385.13)
Random Forest	1403.53	0.3655	88.05

The Poisson GLM was the most computationally efficient model, followed closely by the ridge and Bayesian models. The LASSO and elastic net were next, with the GLMM and SaS models a number of times slower. The random forest model was orders of magnitude slower than the other models because of the many individually optimized trees.

When predicting, we were only able to run the random forest model once because of the long runtime. The simplest models (GLM and Bayes GLM) performed the best in terms of MAPE. The random forest model did really well in terms of MSPE. After the random forest model, the GLM, Bayes GLM, LASSO and elastic net all performed similarly. The GLMM did not do very well, but also did not use all of the same covariates as the other models. MSPE adds more emphasis to single predictions with large errors, in this case exceptionally large predicted (or actual) frequencies. It is interesting that the ridge regression performed worse than the LASSO and elastic net in this situation. The random forest model may better predict large values than the GLM. That is why the GLM has a consistently smaller MAPE, but a larger MSPE. In choosing the best model for frequency on this data, it is difficult to differentiate between the GLM (whether Bayesian or not) and the random forest model.

5.2. Severity

The severity models produce the following metrics:

	Runtime (in mins)	MAPE Confidence Interval	MSPE Confidence Interval (millions)
Gamma GLM	0.10	(1213.0, 1218.6)	(130.12, 130.95)
Bayesian Gamma GLM	0.39	(1212.1, 1218.4)	(130.23, 130.94)
Gamma GLMM	3.46	(1216.6, 1221.4)	(130.57, 131.17)
Gamma LASSO	0.16	(3710.7, 3716.2)	(148.75, 148.81)
Gamma Ridge	0.16	(3710.7, 3716.2)	(148.75, 148.81)
Gamma Elastic Net	0.17	(3710.8, 3716.3)	(148.75, 148.81)
Random Forest	253.14	1169.4	137.88

As in the frequency data, the simple models were very effective. Under MAPE, the random forest model was optimal, but the GLM, Bayes GLM, and GLMM were next. The results are similar with MSPE, only with the simpler models outperforming the random forest model.

The GLMM was slower than all the other models, except for the random forest model which again was orders of magnitude slower than the other models.

Note that these results are similar to the frequency results, except that the random forest model outperformed in terms of MAPE here and was worse in MSPE. Again it is difficult to choose a best model between the GLMs and the random forest models.

5.3. Pure Premium

The pure premium models produce the following results.

	Runtime (in mins)	MAPE Confidence Interval	MSPE Confidence Interval
Tweedie GLM	0.87	(5.751, 5.793)	(18100.47, 18102.30)
Bayesian Tweedie GLM	2.02	(5.756, 5.789)	(18100.36, 18102.17)
Tweedie GLMM	8.31	(5.862, 5.902)	(18100.26, 18101.49)
Tweedie LASSO	0.42	(9.279, 9.318)	(18130.01, 18130.75)
Tweedie Ridge	0.45	(9.279, 9.317)	(18130.01, 18130.75)
Tweedie Elastic Net	0.45	(9.279, 9.318)	(18130.01, 18130.75)
Random Forest	1022.73	3.347	19224.24

Similar to the severity models, the random forest outperformed the rest of the models under MAPE, but was much worse in terms of MSPE. This is likely due to the fact that it severely over-predicted (or under-predicted) some of the policies. This is another potential benefit of the market-based model selection metric. In actual practice, under-predicting premium is much more detrimental than over-predicting. Also, continuing the theme, the random forest model took much longer to fit than any of the other models.

5.4. Market-based Metric

Looking only at these metrics continues to offer a difficult decision, which model is the best? The GLMs are the most computationally efficient and perform the best under some of the metrics, but the random forest model is best under other metrics. Is the increased computational complexity of the random forest model worthwhile? Why is the performance so different between frequency and severity? Does the dramatic improvement in terms of MAPE cancel out the poor performance in terms of MSPE? The market model comparison answers this very question by simulating that if we were to price with a simpler model, but our competitor was to price with a random forest, what would happen to our market share or our loss ratio? This can help us more concretely decide which model to choose and the value of a more complicated or computationally intense method.

In our application, the simulation produced the following results:

	Loss Ratio	Market Share
Tweedie GLM	(1.034, 1.075)	(0.058, 0.065)
Bayesian Tweedie GLM	(1.040, 1.080)	(0.051, 0.057)
Tweedie GLMM	(1.057, 1.074)	(0.085, 0.088)
Tweedie LASSO	(1.367, 1.438)	(0.048, 0.054)
Tweedie Ridge	(1.282, 1.334)	(0.097, 0.113)
Tweedie Elastic Net	(1.321, 1.367)	(0.147, 0.167)
Random Forest	(0.947, 0.953)	(0.484, 0.486)

The random forest model is the only one with a loss ratio less than 1.0 (implying more premiums than losses) and obtains almost half of the market share. The simpler models were next best and the regularization methods performed the worst. This shows that even though the random forest was the least accurate in terms of pure premium MSPE, being most accurate in terms of MAPE lead it to being dominant in the market simulation. One possible explanation is that MSPE excessively penalizes large misses. If the random forest model severely over-predicted some of the policies, that would greatly impact MSPE but the impact on MAPE would be muted. The model wouldn't win those policies, so the over-prediction would not adversely impact the market simulation. These results show that the increased computational time to fit a random forest is definitely worth the advantage gained in the market.

One other nice feature is the ability to look at the results and quickly interpret possible reasons. About half of the market went to the random forest model and the other half went to the remaining models. What if the other models simply split the rest of the policies? How would the random forest perform if compared with just one of the other six models? What would happen if we allowed the models to contain interaction or polynomial terms? What about an ensemble model of many of the models currently explored? All of these questions and more can be answered by market-based model selection.

Seeing the loss ratio less than 1.0, a potential issue emerges. Because there is no accounting for profit or expenses, is a loss ratio less than 1.0 actually desirable? If a model were perfect, its predicted pure premium would always equal the actual losses. Therefore any of the policies that it earns would have a loss ratio of exactly 1.0. Any loss ratio less than 1.0 would be due to over-prediction, less desirable than an accurate prediction.

This turns out to be another scenario where the market-based model selection method aligns statistical techniques with business realities. We can see that by looking at two separate scenarios.

1. Assume a perfect model exists. It would only write the policies where all other models over-predicted the losses. All of the other models would only write policies where they under-predict losses making their loss ratios greater than 1.0 and showing that the perfect model is the best.
2. In reality there are no perfect models. The best model is the one which makes the least costly mistakes. Most common model selection procedures are symmetric, implying that over- and under-predictions are equally penalized. In the market-based method (and in actual practice) an over-prediction simply means that you likely won't write the policy (because the premium is so high), but an under-prediction exposes you to severe risk because you are likely to write the policy and have more losses than premiums if you do.

6. Conclusion and Future Research

When looking at standard metrics, it can be difficult to determine which model is best. This is especially true when the different metrics provide conflicting information. The market-based model comparison method can help an analyst to choose a model more closely based on the desired outcomes and results. It can provide a better story for stakeholders to justify the capital expenditures to roll out a new model, or at least make that decision more data-based. In our commercial auto dataset, it is difficult to decide between GLMs and a random forest when looking at standard model comparison metrics. The market-based model comparison shows that the random forest model is dramatically superior to the other models in metrics that actually affect insurance profits (loss ratio and market share).

There are a few shortcomings and cautions with the results in this paper. The most important issue has to do with the amount of time necessary to fit a quality model in each of the model families. While the random forest model can do much of its own tuning, variable selection, and even a little bit of feature engineering, the other model families require a good bit of testing and adjustment to the variables in the model to get a well-fitting model. Some (or even all) of the advantage of the random forest models may evaporate if certain interaction terms or other covariates were added to the other models. Also, while the random forest model won convincingly in our application, this was just one application. Care should be taken when using these results in a similar situation. All of the methods in this paper are limited by the implementation in their respective R packages. A different implementation of some of the models could perform better. This is also not an exhaustive list of possible models. Two-part models (where frequency and severity are each modeled and then combined) are another great example of a possible model to consider. One other possibility is the increased flexibility of the random forest models, especially when it comes to the relationships between the covariates, helped the random forest to outperform the others. If that is true, some of the benefits of the random forest can be muted by allowing more complicated (interactions, polynomials, etc.) model structures in the GLMs. Additionally, legal and regulatory requirements can limit which ratemaking models are even available to an insurer. Finally, losses in other lines of business can behave very differently than losses in commercial auto. All of these factors should be taken into account when comparing your own set of models.

These models also didn't incorporate the number of exposures as a weight in the regression models, which would properly account for the fact that an observation with 4 exposures is actually four observations with identical characteristics. Essentially all of the models compared have a weight variable that can be set. The notable exception is the random forest model. To incorporate exposures as a weight in that model, the observations in the dataset would be expanded to multiple rows with identical characteristics. For example, if an observation includes 6 exposures, it could be expanded to 6 rows each with a single exposure. This could also work with decimal exposures (dividing a 2.3 exposure observation into 5 observations each with 0.5 exposures or 23 observations each with 0.1 exposures). The more precisely the observations are divided, the larger the computational impact. Because this paper focused on comparative and predictive (not inferential) results, the results are unchanged by this simplification.

Additionally, it would be better if this metric were used over multiple years, allowing the different models to adjust their base rates and drop out of the market completely depending on their performance. Unfortunately, we only have access to a single year of data in this project.

It would be beneficial to continue to test the hypothesis that the random forest outperforms the other models. This can be done with other datasets, other lines of business, more comparison models (e.g. two-part models, ensembles of the current models, more complicated design matrices), and different comparison metrics (e.g. GINI, information criteria).

7. Acknowledgments

The authors are grateful for the financial support of the Casualty Actuarial Society, to the anonymous reviewers for their thoughtful comments which greatly improved the manuscript, and to Jim King for planting the seeds of this project.

References

Bermúdez, L., Karlis, D., 2011. Bayesian multivariate poisson models for insurance ratemaking. Insurance: Mathematics and Economics 48, 226–236.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Breiman, L., 2017. Classification and regression trees. Routledge.

Dey, D.K., Ghosh, S.K., Mallick, B.K., 2000. Generalized linear models: A Bayesian perspective. CRC Press.

Frees, E.W., Meyers, G., Cummings, A.D., 2014. Insurance ratemaking and a gini index. Journal of Risk and Insurance 81, 335–366.

Gelman, A., Hill, J., 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge university press.

Guelman, L., 2012. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Systems with Applications 39, 3659–3667.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Jørgensen, B., Paes De Souza, M.C., 1994. Fitting tweedie's compound poisson model to insurance claims data. Scandinavian Actuarial Journal 1994, 69–93.

Klein, N., Denuit, M., Lang, S., Kneib, T., 2014. Nonlife ratemaking and risk management with bayesian generalized additive models for location, scale, and shape. Insurance: Mathematics and Economics 55, 225–249.

Kuo, L., Mallick, B., 1998. Variable selection for regression models. Sankhyā: The Indian Journal of Statistics, Series B , 65–81.

McCulloch, C.E., Searle, S.R., 2004. Generalized, linear, and mixed models. John Wiley & Sons.

Nelder, J.A., Wedderburn, R.W., 1972. Generalized linear models. Journal of the Royal Statistical Society: Series A (General) 135, 370–384.

Shi, P., 2016. Insurance ratemaking using a copula-based multivariate tweedie model. Scandinavian Actuarial Journal 2016, 198–215.

Shi, P., Feng, X., Boucher, J.P., et al., 2016. Multilevel modeling of insurance claims using copulas. The Annals of Applied Statistics 10, 834–863.

Tang, Y., Xiang, L., Zhu, Z., 2014. Risk factor selection in rate making: Em adaptive lasso for zero-inflated poisson regression models. Risk Analysis 34, 1112–1127.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) , 267–288.

Tweedie, M., 1984. An index which distinguishes between some important exponential families, in: Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference, p. 604.

Yip, K.C., Yau, K.K., 2005. On modeling claim frequency data in general insurance with extra zeros. Insurance: Mathematics and Economics 36, 153–163.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.