Annals of Actuarial Science, Vol. 11, part 2, pp. 390–411. © Institute and Faculty of Actuaries 2017
doi:10.1017/S1748499517000070
First published online 4 May 2017

Application of bivariate negative binomial regression model in analysing insurance count data

Feng Liu* and David Pitt

Department of Applied Finance and Actuarial Studies, Faculty of Business and Economics, Macquarie University, North Ryde NSW 2109, Australia

Abstract

In this paper we analyse insurance claim frequency data using the bivariate negative binomial regression (BNBR) model. We use general insurance data on claims from simple third-party liability insurance and comprehensive insurance. We find that bivariate regression, with its capacity for modelling correlation between the two observed claim counts, provides both a superior fit and out-of-sample prediction compared with the more common practice of fitting univariate negative binomial regression models separately to each claim type. Noting the complexity of BNBR models and their potential for a large number of parameters, we explore the use of model shrinkage methodology, namely the least absolute shrinkage and selection operator (Lasso) and ridge regression. We find that models estimated using shrinkage methods outperform the ordinary likelihood-based models when being used to make predictions out-of-sample. We find that the Lasso performs better than ridge regression as a method of shrinkage.

Keywords

Bivariate negative binomial regression model; Lasso; Ridge regression

1. Introduction

We explore the use of a bivariate negative binomial regression (BNBR) model in the context of modelling bivariate insurance claim frequency data. Two types of insurance claims, the third-party liability claim and the comprehensive cover claim, made by the same policyholder are assumed to be correlated and to be explained by a set of explanatory variables. By allowing a correlation between the two response variables, the performance of the BNBR is better than if two univariate negative binomial regression (UNBR) models are fitted separately, both in terms of in-sample goodness-of-fit and out-of-sample prediction. We also find that the BNBR also outperforms the bivariate Poisson regression (BPR) model.

In addition, we apply two shrinkage techniques, the least absolute shrinkage and selection operator (Lasso) and ridge regression, to reduce the number of covariates used in the original unshrunken BNBR model. Although an increasing number of explanatory variables will increase in-sample goodness-of-fit, an overfitted model may result which performs less well in out-of-sample prediction. By selecting more relevant risk factors and removing unnecessary explanatory variables, we find that the shrunken models outperform the unshrunken model in out-of-sample prediction.

*Correspondence to: Feng Liu, Department of Applied Finance and Actuarial Studies, Faculty of Business and Economics, Macquarie University, North Ryde NSW 2113, Australia. Tel: 61 2 9850 8455; E-mail: feng.liu4@hdr.mq.edu.au

We use the model specification for BNBR in Famoye (2010*b*), where correlation structure allows for both a negative and a positive relationship between the two claim type frequencies.

The contributions of this paper are threefold. First, we successfully demonstrate the importance of the BNBR model in analysing over-dispersed general insurance claim data, which outperforms the BPR model. Second, the correlation factor is found to be significant, with the implication that BNBR model is more suitable when the two claim counts are correlated. A similar conclusion is not evident in Famoye (2010*b*), where the correlation between the two variables considered is too low for useful dependence modelling, and thus univariate models seem to be adequate. Third, we shrink both BNBR models and UNBR models to reduce the size of coefficients of irrelevant explanatory variables, some of which are eliminated totally from the regression model. The shrinkage results are consistent with James *et al.* (2013: Chapter 6), in that the shrunken models provide much higher out-of-sample prediction accuracy, compared with the original full BNBR models.

The paper is organised as follows: section 2 gives a summary of existing methods to analyse claim counts, including univariate and bivariate generalised linear models (GLMs). Section 3 describes the model used in this study as well as the shrinkage techniques. Section 4 introduces the claims data. Section 5 gives the modelling results and a discussion of findings. Section 6 concludes the paper.

2. Literature Review

Modelling of insurance claim count data has been an active area of research for some decades. The research interest often lies in modelling the relationship between the observed counts and a set of explanatory variables. GLMs are very commonly used for this purpose as a mathematical formulation of the relationship. With a chosen link function, the mean of the distribution can be expressed as a linear function of the explanatory variables.

Under the GLM framework, the response variable is modelled using a member of the exponential dispersion family of distributions. Two common choices for this distribution in the case of insurance count data are the Poisson distribution and the negative binomial distribution (see McCullagh & Nelder, 1989). While the Poisson regression model assumes equality between the underlying mean and variance of the response variable, negative binomial regression relaxes the assumption and accounts for over-dispersion in the data (see Cameron & Trivedi, 2005). Both models have been widely adopted to analyse claim count data in general insurance. For example, Dionne & Vanasse (1989) used both Poisson and negative binomial regression models for automobile insurance risk classification. Haberman & Renshaw (1996) illustrated the use of the over-dispersed Poisson model in analysing life insurance claim counts, after presenting a summary of GLMs in actuarial science. Some other early studies in this area are Samson & Thomas (1987), Hürlimann (1990) and Renshaw (1995).

Various extensions to the basic GLM framework have been proposed in the statistics literature and explored in insurance contexts. For example, generalised additive models (GAMs) are postulated by combining an original GLM with additive models in the linear regression model, where smooth functions with semi-parametric or non-parametric forms are applied to explanatory variables. So with a chosen link function, the mean of the response variable is expressed as a linear function of unknown smooth functions of explanatory variables (see Hastie & Tibshirani, 1990). The GAM framework is adopted in Denuit & Lang (2004) to account for discrete, continuous and spatial risk

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

factors in a Bayesian framework for insurance rate-making purposes. Mixtures of GLMs, such as Poisson mixtures, can be used to accommodate non-homogeneous populations (see Karlis & Xekalaki, 2005). More recently, increasing attention has been given to the application of extended GLMs in accounting for excess zero and over-dispersion in count data, especially for automobile insurance count numbers under no claim discount system. The proposed zero-inflated models are considered as a mixture of zero point mass and a Poisson or negative binomial regression models under the original GLM framework. Yip & Yau (2005) provided a good summary of zero-inflated models with an application in general insurance count data. Heller *et al.* (2007) considered a group of candidate distribution to model claim counts, including Poisson, zero-inflated Poisson and negative binomial. Thorough reviews for count data regression can be found in Denuit *et al.* (2007) and Cameron & Trivedi (1998).

In addition to univariate models, bivariate regression models have been proposed to analyse two response variables that are possibly correlated. These models offer sufficient flexibility by allowing the two response variables to be affected by different predictive factors. Moreover, a bivariate model is more helpful for inference and prediction purposes because it allows us to properly specify the dependency between the two dependent variables (see Shi & Valdez, 2014).

One way to introduce the correlation factor is to use copulas to analyse the correlation structure, by linking univariate marginals to the full multivariate distribution (see Frees & Valdez, 1998). The use of copulas is common in analysing correlation structure related to continuous variables such as claim losses (see Denuit *et al.*, 2006; Frees & Valdez, 2008; Czado *et al.*, 2012). In studying discrete variables such as the number of insurance claims, Cameron *et al.* (2004) used a bivariate copula in modelling the difference between self-reported and true doctor visits, but the application is limited to studying the distribution of the difference between two counts. Shi & Valdez (2014) considered three types of automobile claim counts using a mixture of copulas and the family of elliptical copulas. A review of using couplas to specify correlation structure can be found in a recent study by Chen & Hanson (2017).

Another group of studies analyse the correlation structure through the trivariate reduction method, where the pair of dependent variables are specified using three random variables. For example, by setting $Y_1 = X_1 + X_{12}$ and $Y_2 = X_2 + X_{12}$, where X_1, X_2 and X_{12} are independent Poisson random variables, Y_1 and Y_2 have a bivariate Poisson distribution with a covariance term derived from the use of the common Poisson variable X_{12} (see Kocherlakota & Kocherlakota, 1992; Johnson *et al.*, 1997). The trivariate reduction method has been explored in Jung & Winkelmann (1993), King (1989) and Kocherlakota & Kocherlakota (2001). Karlis & Xekalaki (2005) proposed an extension to allow for a combination of common random variables. Bermúdez & Karlis (2011) postulated a zero-inflated multivariate Poisson model to account for excess of zeros in automobiles insurance claim data. In another context of frequency modelling, a multivariate Poisson-lognormal regression model has been used for prediction of crash counts (Ma *et al.*, 2008; El-Basyouny & Sayed, 2009).

Although the trivariate reduction model can be extended to capture over-dispersion in the data, one drawback is that the correlation can only be positive (see Famoye, 2010*b*; Shi & Valdez, 2014). One way to address the issue is to use an imposed parameter in the bivariate probability function to specify a covariance term to account for correlation. As the value of this correlation parameter can be negative, zero and positive, the limitation of positive correlation is removed. Thus, the model is obviously more flexible with a more straightforward covariance structure.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Lakshminarayana *et al.* (1999) defined a BPR model by including a multiplicative factor to capture the correlation between the two response variables. The probability function for the bivariate distribution is composed of two univariate Poisson probability functions, linked by the multiplicative correlation factor whose value depends on the embedded correlation parameter.

Based on a similar correlation structure, Famoye (2010*b*) applied a BNBR model to analyse the bivariate distribution of two series of count data, while addressing over-dispersion in the sample. The study models marginal means of the two response variables with a set of explanatory covariates in a log-linear relationship. Data from the 1977–1978 Australian health survey is used to illustrate the model and the coefficients are estimated with maximum likelihood technique. The test results show that the BNBR model provides a better fit to the data than the BPR model, and supports the use of BNBR when the variance of the data is very different from the mean. However, the correlation parameter is not significant, thus two UNBR models may be able to provide similar results in his study.

2.1. Shrinkage methods

One drawback of the likelihood-based estimation of the regression models described above in the analysis of count data is that it commonly leads to a large number of variables being used. Although it is very tempting to incorporate as much information as possible to account for the heterogeneity in the population, this strategy is more time consuming in terms of model estimation. Too many explanatory variables in a regression model can also result in overfitting and consequently poor out-of-sample predictions.

The Lasso and ridge regression are two popular methods to shrink models (see Tibshirani, 1996; James *et al.*, 2013). Model shrinkage refers to the process of determining a smaller subset of variables that provide stronger explanatory power. Both techniques constrain the coefficient estimates through a penalty term in the maximum likelihood estimation algorithm, comprised of the coefficient values and a shrinkage parameter ω. The higher the shrinkage parameter, the higher the impact of the shrinkage penalty. As a result, the coefficient values will approach zero as ω increases without bound. The optimal ω is commonly selected using cross-validation.

The two techniques differ in the way coefficient values are incorporated in the shrinkage penalty. The Lasso uses the sum of absolute values of coefficients, and ridge regression uses the sum of squared values. Ridge regression tends to shrink all coefficients towards zero, but will not generally set any of them to exactly zero. The Lasso is an alternative to ridge regression and can force some of the coefficient estimate to exactly zero if ω is sufficiently large. In other words, Lasso performs variable selection (see James *et al.*, 2013: Chapter 6).

The importance of model shrinkage has been recognised in the actuarial literature. First proposed by Tibshirani (1996), the Lasso has been extended to GLMs to handle count data (see Park & Hastie, 2007). Tang *et al.* (2014) applied adaptive Lasso to car insurance data. The risk factor selection improves the model goodness-of-fit both in the Poisson model as well as zero-inflated Poisson model. Wang *et al.* (2015) considered over-dispersed data and added a Lasso penalty to the maximum likelihood function of the negative binomial regression model. Their study concludes that a parsimonious model offers better prediction and interpretation. Both Tang *et al.* (2014) and Wang *et al.* (2015) used univariate regression models and applied the shrinkage technique to only one response variable. Ridge regression is shown to improve mean square error in an early study by

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Hoerl & Kennard (1970). The technique is then applied to many areas of science. Some examples are Shen *et al.* (2013), Douak *et al.* (2013) and Meijer & Goeman (2013).

The two shrinkage methods can be applied to regression models to remove less significant variables. As a consequence, the unnecessary complexity in the model can be reduced and this leads to easier interpretation and potentially improved out-of-sample predicition (see James *et al.*, 2013: Chapter 6). It is these possibilities which we explore in the context of bivariate insurance claim data in this paper.

3. Methodology

3.1. BNBR model

The bivariate Poisson distribution proposed in Lakshminarayana *et al.* (1999) has a probability function as the product of Poisson marginals with a multiplicative factor:

$$P(y_1, y_2) = \prod_{t=1}^{2} \frac{\theta_t^{y_t} e^{-\theta_t}}{y_t!} \times \left[1 + \lambda\left(e^{-y_1} - e^{-d\theta_1}\right)\left(e^{-y_2} - e^{-d\theta_2}\right)\right],\ y_1, y_2 = 0, 1, 2, \ldots \tag{1}$$

where $d = 1 - e^{-1}$. θ_t is the mean of $Y_t (t = 1, 2)$, and Y_1 and Y_2 are both Poisson distributed. The covariance between Y_1 and Y_2 is $\lambda\theta_1\theta_2 d^2 e^{-d(\theta_1+\theta_2)}$ and the correlation is $\rho = \lambda\sqrt{\theta_1\theta_2}\, d^2 e^{-d(\theta_1+\theta_2)}$. Depending on the value of λ, the two response variables Y_1 and Y_2 can be positively or negatively correlated, or independent if λ is equal to zero.

By using a similar approach, Famoye (2010*b*) defined a bivariate negative binomial distribution. Following the same covariance specification as Lakshminarayana *et al.* (1999), a bivariate negative binomial distribution has the following probability function:

$$P(y_1, y_2) = \prod_{t=1}^{2} \binom{y_t + m_t^{-1} - 1}{y_t} \theta^{y_t} (1-\theta)^{m_t^{-1}} \times [1 + \lambda(e^{-y_1} - c_1)(e^{-y_2} - c_2)],\ y_1, y_2 = 0, 1, 2, \ldots \tag{2}$$

Both Y_1 and Y_2 are random variables and follow a negative binomial distribution, with dispersion parameters m_1^{-1} and m_2^{-1}, respectively. The mean of $Y_t (t = 1, 2)$ is $\mu_{t_1} = m_t^{-1}\theta_t/(1-\theta_t)$ and the variance is $\sigma_t^2 = m_t^{-1}\theta_t/(1-\theta_t)^2$. Also, $c_t = E\left(e^{Y_t}\right) = \left[(1-\theta_t)/\left(1-\theta_t e^{-1}\right)\right]^{m_t^{-1}}$.

Let n denote the sample size and Y_{it} $(t = 1, 2;\ i = 1, 2, \ldots, n)$ denote the count response variable, the corresponding vector of k explanatory variables is represented as $x_i = (x_{i0} = 1, x_{i1}, \ldots, x_{ik})$. Assuming a log-linear model and the same set of covariates as possible explanatory variables for both Y_{i1} and Y_{i2}, the means of the two response variables can be modelled as

$$E(Y_{it}|x_i) = \mu_{it} = \exp(x_i\beta_t),\ t = 1, 2 \tag{3}$$

where $\beta_t^T = (\beta_{t0}, \beta_{t1}, \beta_{t2}, \ldots, \beta_{tk})$ and is the vector of the coefficients estimated using the maximum likelihood method. Given that $\theta_{it} = \mu_{it}/\left(m_t^{-1} + \mu_{it}\right)$, equation (2) can be rewritten as:

$$P(y_{i1}, y_{i2}) = \prod_{t=1}^{2} \binom{y_{it} + m_t^{-1} - 1}{y_{it}} \left(\frac{\mu_{it}}{m_t^{-1} + \mu_{it}}\right)^{y_{it}} \left(\frac{m_t^{-1}}{m_t^{-1} + \mu_{it}}\right)^{m_t^{-1}} \times [1 + \lambda(e^{-y_{i1}} - c_1)(e^{-y_{i2}} - c_2)] \tag{4}$$

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Accordingly, the log-likelihood function, which is set to a maximum to estimate the model parameters, for the unshrunken model is:

$$\log L = \sum_{i=1}^{n}\left\{\sum_{t=1}^{2}\left[y_{it}\log\mu_{it} - m_t^{-1}\log m_t - \left(y_{it}+m_t^{-1}\right)\log\left(\mu_{it}+m_t^{-1}\right) - \log(y_{it}!) + \sum_{j=1}^{y_{it}-1}\log\left(m_t^{-1}+j\right)\right] + \log[1+\lambda(e^{-y_{i1}}-c_1)(e^{-y_{i2}}-c_2)]\right\} \quad (5)$$

where $c_t=(1+d\mu_{it}m_t)^{-1/m_t}$ with $d=1-e^{-1}$. Equation (5) can be maximised with respect to β_t, m_t and λ. The asymptotic standard deviations of the estimated parameters are obtained in the usual way from Hessian matrix.

The deviance for the BNBR model, which is a measure of the model's goodness-of-fit, is defined as:

$$D = 2\sum_{i=1}^{n}\left\{\sum_{t=1}^{2}\left[y_{it}\log\left(\frac{y_{it}\left(m_t^{-1}+\hat{\mu}_{it}\right)}{\hat{\mu}_{it}\left(m_t^{-1}+y_{it}\right)}\right) + m_t^{-1}\log\left(\frac{m_t^{-1}+\hat{\mu}_{it}}{m_t^{-1}+y_{it}}\right)\right] \log\left(\frac{1+\lambda\prod_{t=1}^{2}(e^{-y_{it}}-\bar{c}_t)}{1+\lambda\prod_{t=1}^{2}(e^{-y_{it}}-\hat{c}_t)}\right)\right\} \quad (6)$$

where $\bar{c}_t$ and $\hat{c}_t$ are the values of c_t evaluated at $\mu_{it}=y_{it}$ and $\mu_{it}=\hat{\mu}_{it}$, respectively, and $\hat{\mu}_{it}$ the predicted value of μ_{it} found using equation (3) with estimated coefficients that maximise equation (5).

3.2. The Lasso and ridge regression

Given the BNBR model in equation (4), the coefficient vector β_t can be estimated by maximising equation (5). The resulting model will be called the full model in what follows. Here β_t ($t=1, 2$) are vectors each having $k+1$ values. These relate to the model intercept and k explanatory variable coefficients. When k is large, the model may produce poor out-of-sample results because of an overfitting problem. It is therefore useful to shrink the estimated BNBR model using either the Lasso approach or ridge regression, by subtracting a shrinkage penalty from the log-likelihood function.

We define the log-likelihood function of the BNBR model in section 3.1, which is Log L in equation (5). The new functions to be maximised under the two shrinkage approaches, with $2\times k$ coefficients to be analysed are specified as:

$$\begin{aligned} &\text{The Lasso}: && \log L - \omega\sum_{t=1}^{2}\sum_{j=1}^{k}\left|\beta_{tj}\right| \\ &\text{Ridge regression}: && \log L - \omega\sum_{t=1}^{2}\sum_{j=1}^{k}\beta_{tj}^2 \end{aligned} \quad (7)$$

where ω is the shrinkage parameter. Here t takes values 1 and 2, indicating that the shrinkage models consider regression coefficients for both y_1 and y_2. Thus the above equations specify the two shrinkage models in the context of a bivariate model.

Note that we do not shrink the intercept coefficients (β_{t0}), as they simply constitute a measure of the mean value of the response variables when other explanatory variables are set to zero. Similarly, we also exclude the two over-dispersion parameters (m_1, m_2) and the correlation parameter (λ) from

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

shrinkage, as we are focussing on shrinking the estimated association of each explanatory variable with the response. As a result, for each response variable, k regression coefficients are included in the shrinkage penalty.

When ω is equal to zero, both the Lasso and ridge regression will generate the same coefficients as the full model. A larger ω gives greater emphasis to model simplicity compared with in-sample goodness-of-fit. Consequently coefficient values will deviate from the maximum likelihood estimates, resulting in reduced in-sample goodness-of-fit. At the same time, the model is simplified with the potential for improved out-of-sample performance.

It is clear that different ω values will lead to different coefficients in the shrunken model and therefore differing out-of-sample prediction results. In order to perform the two shrinkage techniques as specified in equation (7) using the maximum likelihood method, the optimal value must be chosen for ω based on only the sample data to achieve the possibly best out-of-sample prediction accuracy. In this study we use k-fold cross-validation for this purpose, where commonly k is set to be 5 or 10. In the cross-validation process, the sample data are randomly divided into k groups. One group is chosen as the validation set, while the model is fitted on the remaining $k-1$ groups. The fitted model is applied to the validation set to calculate the out-of-sample deviance, as the validation set is held out in the model fitting process. As there are k groups, the procedure can be repeated k times resulting in k deviances when each of the k groups is held out as the validation set. The average of the k deviance values, each denoted deviance_i ($i = 1, 2, \ldots, k$), is taken as the cross-validation result, or k-fold cross-validation, at a particular ω value:

$$CV_{(\omega)} = \frac{1}{k}\sum_{i=1}^{k} \text{deviance}_i$$

For each of the ω values, we perform the procedure as described previously. Among a grid of ω values, the most appropriate ω is the one that generates the lowest k-fold CV. As the CVs are calculated on the validation set, separated from the data to fit the model, when ω increases the CV is expected to decrease initially and later increase again when the impact from the penalty term is too strong. The ω that gives the minimum CV should be chosen.

We note here that although we develop different log-likelihood functions and shrinkage functions for the bivariate model, the validation process is standard. This is because the validation process only takes into consideration the deviances generate by a model, whether it is univariate or bivariate. Given the specified shrinkage models in equation (5), the validation process mentioned previously is proper for the BNBR model.

The shrinkage parameter, ω, is not assumed to be the same for the two shrinkage methods. A separate cross-validation is performed for each of the methods to locate the best ω value. Once this is achieved, the model is fitted again to the full set of data, disregarding the previously k group classifications. The shrunken models can then be compared with the full model, which is estimated using maximum likelihood without any penalty term.

4. Data

The study is based on data from 14,000 automobile policies from a major insurance company in Spain, randomly selected from a pool of 80,994 policies. A subset of the data is also used in

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Table 1. Explanatory variables in the regression model.

Variables	Description
v1	Equals 1 for women and 0 for men
v2	Equals 1 when driving in urban area, 0 otherwise
v3	Equals 1 when zone is medium risk (Madrid and Catalonia)
v4	Equals 1 when zone is high risk (Norther Spain)
v5	Equals 1 if the driving license is between 4 and 14 years old
v6	Equals 1 if the driving license is 15 or more years old
v7	Equals 1 if the client is in the company for more than 5 years
v8	Equals 1 if the insured is 30 years old or younger
v9	Equals 1 if includes comprehensive coverage (except fire)
v10	Equals 1 if includes comprehensive and collision coverage
v11	Equals 1 if horsepower is $\geq$5,500 cc

Brouhns *et al.* (2003), Bolancé *et al.* (2008), Bolancé *et al.* (2003), Boucher & Denuit (2008), Boucher *et al.* (2007), Bermúdez & Karlis (2011) and Boucher *et al.* (2009). We use 10,000 policies to estimate the model parameters, and the remaining 4,000 policies are used to test the model's out-of-sample prediction accuracy.

We model two types of claims, and their associated claim counts are recorded as Y_1 and Y_2. Y_1 represents the simple third-party liability with basic guarantees, and Y_2 stands for comprehensive cover. The same set of explanatory variables are assumed to affect both Y_1 and Y_2. The explanatory variables are summarised in Table 1. A similar table can also be found in Boucher *et al.* (2009).

We present in Table 2 a summary of the effects of the covariates on claim count based on all 80,994 policies[1]. The covariates are classified into eight groups. In the first column, we present the total number of policies that fall into each subgroup, followed by the percentage of policies with claim counts equal to 0, 1 or 2 (including higher than 2) for Y_1 and Y_2, respectively.

For example, in the case of gender, we see here 12,957 of the policyholders are female. In total, 93% of these female policyholders does not make a third-party liability claim and 91.64% does not make a claim on the comprehensive cover. This is to be compared with the male policyholders, where 93.80% of them does not make a third-party liability claim and 92.59% makes no claim on the comprehensive cover. Ignoring other covariates and factors, female policyholders tend to have a slightly riskier profile compared with male policyholders.

Similar observations can be made for the other groups of covariates. A lower claim count tends to be associated with driving in low-risk zone, a longer driving experience, a longer time with the company, an older age and a smaller car horsepower. The effects of driving area (v2) and insurance cover (v9, v10) seem to be minimal based on this one-way analysis.

The estimated mean and variance of Y_1 and Y_2 are given at the end of Table 2. Y_1 has a lower mean and smaller variance compared with Y_2. Moreover, the variance is much higher than the mean for both claim types. This feature implies that a model capable of handling over-dispersed data,

[1] Similar distribution figures can be generated for the sample chosen in this paper, which are not presented here.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Table 2. Summary statistics of claim frequencies as classified by the explanatory variables.

		Y_1 (Third-party liability claim)			Y_2 (Comprehensive cover claim)		
	Total	Count = 0 (%)	Count = 1 (%)	Count ≥ 2 (%)	Count = 0 (%)	Count = 1 (%)	Count ≥ 2 (%)
Gender							
Female (v1 = 1)	12,957	93.29	5.38	1.33	91.64	6.14	2.22
Male (v1 = 0)	68,037	93.80	4.86	1.34	92.59	5.60	1.81
Area							
Urban (v2 = 1)	54,183	93.81	4.86	1.33	92.21	5.84	1.95
Other (v2 = 0)	26,811	93.53	5.10	1.37	92.89	5.37	1.74
Zone risk level							
Low (v3 = 0, v4 = 0)	45,958	94.03	4.65	1.33	93.78	4.83	1.39
Medium (v3 = 1, v4 = 0)	19,320	93.78	5.01	1.22	88.65	8.14	3.21
High (v3 = 0, v4 = 1)	15,716	92.73	5.73	1.55	93.17	5.17	1.66
Driver license							
Below 4 years (v5 = 0, v6 = 0)	1,894	90.87	7.18	1.95	93.19	5.33	1.48
Between 4 and 14 years (v5 = 1, v6 = 0)	20,854	92.93	5.57	1.51	90.46	7.19	2.35
Above 14 years (v5 = 0, v6 = 1)	58,246	94.09	4.65	1.26	93.12	5.16	1.72
Years with the company							
<5 years (v7 = 0)	11,670	92.60	5.79	1.61	90.26	7.22	2.53
Longer than 5 years (v7 = 1)	69,324	93.90	4.80	1.30	92.80	5.43	1.77
Age							
30 years old or younger (v8 = 1)	7,484	91.98	6.27	1.75	90.62	7.16	2.22
Older than 30 years (v8 = 0)	73,510	93.89	4.81	1.30	92.62	5.54	1.84
Insurance cover							
No extra cover (v9 = 0, v10 = 0)	39,791	93.97	4.75	1.29	98.62	1.17	0.21
Only comprehensive (except fire)							
Cover (v9 = 1, v10 = 0)	12,613	93.61	5.05	0.90	78.36	14.39	7.25
Both comprehensive and collision							
Cover (v9 = 0, v10 = 1)	28,590	93.41	5.17	1.42	90.04	8.13	1.83
Horsepower							
<5,500 cc (v11 = 0)	15,725	94.07	4.67	1.27	96.09	2.93	0.98
≥5500cc (v11 = 1)	65,269	93.63	5.01	1.36	91.56	6.35	2.08
Mean			0.081			0.102	
Variance			0.123			0.168	

such as the negative binomial regression model, is more appropriate compared with Poisson regression model.

The correlation coefficient between Y_1 and Y_2 is 0.187, taking into account all 80,944 observations. The scatter plot is presented in Figure 1, including a trend line. The two variables can only take integer values. The number of observations at each of the dots is relatively indicated by the size of the dot, which is a rough reflection of the exact count summary shown in Table 3.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

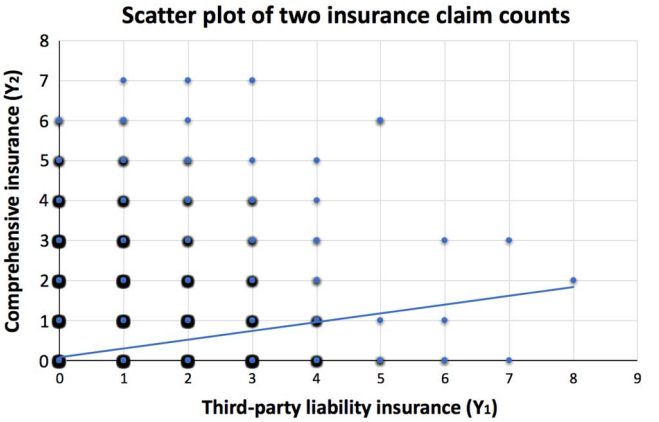

Figure 1. Scatter plot of two insurance claim counts. The size of the dot at each point gives a relative indication of the number of observations. The trend line is also presented.

Table 3. Summary table of two types of insurance counts.

		Y_1								
		0	1	2	3	4	5	6	7	8
Y_2	0	71,087	3,022	574	149	29	4	2	1	0
	1	3,722	686	138	42	15	1	1	0	0
	2	807	184	55	21	3	0	0	0	1
	3	219	71	15	6	2	0	1	1	0
	4	51	26	8	6	1	0	0	0	0
	5	14	10	4	1	1	0	0	0	0
	6	4	3	1	0	0	2	0	0	0
	7	0	1	1	1	0	0	0	0	0

We also analyse the correlation structure in the tail, when at least one of the claim counts is not zero. The correlation coefficient is computed at 0.126, which is lower than if all observations are considered. This is consistent with the statistics in Table 3. If a higher right-tailed correlation is found, modelling tools such as copulas can be used to more accurately model the correlation structure (see Denuit *et al.*, 2006: Chapter 4.4.4). As tail dependency is not presented in this study, the model specified in equation (2) will suffice.

In addition to the variables listed in previous tables, we also consider two-way interaction effects among the variables. Adding interaction terms between independent covariates help relax the assumption that each of those independent variables only has additive effect in the regression model (see Fahrmeir *et al.*, 2013). Interaction effects are frequently analysed in regression models and have been considered in claim counts models (see Yip & Yau, 2005; Shi & Valdez, 2014). We have initially considered 14 potential two-way interactions. These terms cover the interaction effects between different groups of covariates, for example, gender and driving experience, and are summarised in Table 4. We note that after model shrinkage many of the interaction terms were removed from the model.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Table 4. Interaction terms used in the regression model.

With v1	With v2	With v6	With v7	With v8
v1v2	v2v6	v6v7	v7v8	v8v11
v1v6	v2v7	v6v11	v7v11	
v1v7	v2v8			
v1v8	v2v11			
v1v11				

The total number of variables we use in the regression model is 25, excluding the intercept. Although we use the same set of variables for both response variables, we do not expect all explanatory variables to be significant in evaluating the claim counts, nor that the coefficients are the same for Y_1 and Y_2.

5. Results

5.1. BNBR model

We present in Table 5 the results of fitting four models: the BNBR model, UNBR model for Y_1, UNBR model for Y_2, and the BPR model. The four models are classified as full models as opposed to shrunken models, since at this stage we use all available variables including the chosen interaction terms. The BNBR model is specified in equation (4). The two UNBR models are fitted separately for each of the two response variables. The BPR model specification is the same as in Lakshminarayana *et al*. (1999) and is given in equation (1).

The results from the BNBR model are compared with the UNBR models. Coefficients from the BNBR model are consistent with those in UNBR models, both in terms of sign and statistical significance. By introducing a correlation factor λ, which is significant at the 1% level in the BNBR model, it is observed that the deviance of BNBR model is much lower than the sum of the deviances of the two UNBR models. It is true both in sample and out of sample, implying that the BNBR model provides a better in-sample goodness-of-fit, as well as more accurate out-of-sample prediction. It adds value to analyse the two correlated variables in a bivariate model, to properly account for the dependence between the two types of claim counts.

Consistent with expectation, the BNBR model also outperforms the BPR model. Although the BPR model recognises the correlation between the two response variables, the BNBR is more appropriate here when the data are over-dispersed and the variance of the claim counts is much higher than the mean for both types of claims as shown in Table 2. For this reason the BNBR generates both lower in-sample and out-of-sample deviances as expected.

5.2. The Lasso and ridge regression

The first step when applying the two shrinkage techniques is to choose the most optimal shrinkage parameter ω through cross-validation. We choose $k = 10$ and use tenfold cross-validation which is widely used and effective, see for example, Kohavi (1995)[2]. The two intercept coefficients

[2] In addition to tenfold cross-validation, we also conduct fivefold cross-validation and the results are robust to the number of k. Here we present the results for $k = 10$.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Table 5. Modelling results of the BNBR model, two UNBR models and the BPR model, which are all classified as the full models.

Variables	BNBR	UNBR (Y_1)	UNBR (Y_2)	BPR
Y_1 (third-party liability claim)				
Intercept	−1.984 (0.570)***	−1.896 (0.573)***		−1.990 (0.452)***
v1	−0.114 (0.431)	−0.186 (0.434)		−0.123 (0.341)
v2	−0.070 (0.376)	−0.112 (0.377)		−0.091 (0.301)
v3	0.003 (0.108)	0.036 (0.109)		0.013 (0.089)
v4	0.122 (0.115)	0.113 (0.115)		0.115 (0.091)
v5	−0.341 (0.295)	−0.374 (0.300)		−0.316 (0.224)
v6	−0.865 (0.501)*	−0.952 (0.507)*		−0.833 (0.387)**
v7	0.053 (0.437)*	0.056 (0.438)		0.064 (0.349)
v8	−0.655 (0.592)	−0.064 (0.594)		−0.710 (0.476)
v9	−0.012 (0.132)	0 (0.133)		0.074 (0.109)
v10	0.173 (0.099)*	0.179 (0.099)*		0.195 (0.079)
v11	−0.157 (0.433)*	−0.198 (0.434)		−0.155 (0.347)
v1v2	−0.056 (0.253)	−0.017 (0.252)		−0.061 (0.200)
v1v6	0.452 (0.277)	0.481 (0.277)*		0.443 (0.221)**
v1v7	0.011 (0.322)	0.019 (0.322)		0.035 (0.253)
v1v8	−0.100 (0.380)	−0.086 (0.382)		−0.116 (0.308)
v1v11	−0.042 (0.282)	−0.027 (0.282)		−0.052 (0.226)
v2v6	0.023 (0.241)	0.057 (0.242)		0.010 (0.193)
v2v7	−0.255 (0.281)	−0.253 (0.282)		−0.254 (0.224)
v2v8	0.255 (0.366)	0.244 (0.369)		0.266 (0.294)
v2v11	0.335 (0.241)	0.360 (0.242)		0.368 (0.197)
v6v7	0.068 (0.277)	0.059 (0.279)		0.057 (0.216)
v6v11	0.356 (0.295)	0.381 (0.296)		0.345 (0.236)
v7v8	0.666 (0.397)*	0.634 (0.400)		0.688 (0.322)**
v7v11	−0.170 (0.344)	−0.179 (0.346)		−0.197 (0.282)
v8v11	−0.070 (0.430)	−0.044 (0.434)		−0.051 (0.339)
m_1	6.454 (0.649)***	6.440 (0.648)***		
Y_2 (comprehensive cover claim)				
Intercept	−5.104 (0.629)***		−5.041 (0.640)***	−4.732 (0.525)***
v1	0.068 (0.400)		0.039 (0.412)	−0.001 (0.324)
v2	0.489 (0.346)		0.486 (0.355)	0.355 (0.279)
v3	0.136 (0.084)		0.151 (0.087)*	0.184 (0.065)***
v4	−0.257 (0.108)		−0.293 (0.108)***	−0.253 (0.089)***
v5	0.421 (0.314)		0.431 (0.323)	0.328 (0.267)
v6	0.373 (0.489)		0.252 (0.506)	0.143 (0.405)
v7	0.578 (0.477)		0.572 (0.476)	0.403 (0.406)
v8	0.209 (0.605)		0.194 (0.609)	−0.064 (0.510)
v9	2.946 (0.130)***		2.943 (0.131)***	2.942 (0.120)***
v10	1.941 (0.126)***		1.948 (0.127)***	1.955 (0.120)***
v11	0.848 (0.497)		0.843 (0.495)*	0.659 (0.422)
v1v2	−0.351 (0.225)		−0.331 (0.222)	−0.278 (0.181)
v1v6	−0.080 (0.236)		−0.039 (0.239)	−0.129 (0.192)
v1v7	0.274 (0.275)		0.246 (0.275)	0.317 (0.223)
v1v8	0.142 (0.327)		0.132 (0.327)	0.168 (0.261)
v1v11	−0.142 (0.283)		−0.140 (0.283)	−0.183 (0.231)
v2v6	−0.165 (0.203)		−0.143 (0.207)	−0.190 (0.163)
v2v7	−0.213 (0.228)		−0.209 (0.231)	−0.171 (0.177)
v2v8	−0.406 (0.310)		0.506 (0.314)	−0.404 (0.247)
v2v11	−0.023 (0.244)		−0.024 (0.250)	0.119 (0.201)

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Table 5. *Continued*

Variables	BNBR	UNBR (Y_1)	UNBR (Y_2)	BPR
v6v7	0.080 (0.216)		0.098 (0.231)	0.234 (0.172)
v6v11	−0.102 (0.293)		−0.034 (0.297)	−0.058 (0.240)
v7v8	−0.276 (0.305)		−0.199 (0.312)	−0.075 (0.240)
v7v11	−0.884 (0.411)		−0.924 (0.414)**	−0.865 (0.361)***
v8v11	0.177 (0.505)		0.177 (0.513)	0.252 (0.441)
m_2	2.532 (0.254)***		2.504 (0.254)***	
λ	5.663 (0.396)***			5.748 (0.371)***
In-sample log-likelihood	−5,556.90	−2,384.60	−2,945.76	−5,880.94
Out-of-sample log-likelihood	−2,605.69	−1,136.68	−1,519.55	−2,845.35
In-sample deviance	5,215.18	2,384.60	2,866.00	8,764.67
Out-of-sample deviance	2,854.57	1,025.36	1,851.15	4,494.50

Note: The coefficients of each variable are shown, followed by their standard deviation in parentheses.
BNBR, bivariate negative binomial regression; UNBR, univariate negative binomial regression; BPR, bivariate Poisson regression.
*,**,*** represent, respectively, statistical significance at the 10%, 5% and 1% level, calculated based on the *t*-statistics of coefficients of each variable.

(β_{10} and β_{20}), the dispersion parameters (m_1 and m_2) and the correlation parameter (λ) are excluded from the shrinkage process. For each of the two dependent variables, Y_1 and Y_2, 25 coefficients are estimated by maximising the penalised log-likelihood in equation (7).

We select a grid of values for ω ranging from 0 to 50, and perform the procedure as described in section 3.2. The sample data containing 10,000 policyholders is randomly divided into ten groups. One group is held out as the validation group while the model is fitted on the other nine groups at various ω values. This results in a number of different shrunken models and accordingly different deviance values based on equation (6) calculated on the validation set. Repeated ten times for ten different validation sets, we reach a series of $CV_{(a)}$ computed as the average of deviances from the ten validation sets at a, where a denotes different ω values from 0 to 50.

We present in Figure 2 the $CV_{(a)}$ values from the cross-validation process. As expected, $CV_{(a)}$ decreases initially to a minimum before increasing again. When ω is zero, the shrunken models are equivalent to the full model. When ω increases, the deviances calculated using the held-out group first decrease, indicating better out-of-sample prediction results. Both of the curves increase again after reaching a minimum, where the shrinkage penalty is too strong and affects the models' prediction power.

The shrinkage parameter, ω, in the Lasso and ridge regression are chosen using the cross-validation procedure. We get distinct optimal ω values that minimise deviance under the two different methods. As can be seen in Figure 2, the optimal ω chosen for the Lasso was found to be around 13, and the optimal ω for ridge regression was found to be around 4. We refit the BNBR model under two shrinkage approaches at given ω using the penalised log-likelihood as specified in equation (7). The estimated coefficients as well as the chosen ω are all presented in Table 6. The full BNBR model fitted previously in section 5.1 is also included.

Two observations from Table 6 can be made. First, the full model provides the best in-sample goodness-of-fit among the three, indicated by its lowest in-sample deviance. This is as expected as the

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

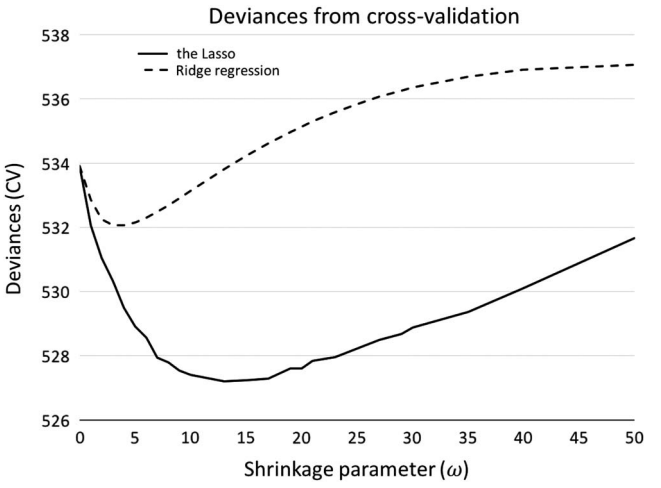

Figure 2. Deviances from cross-validation at different ω values. Each deviance in the graph is calculated as the average of the ten deviances at the same ω generated in the tenfold cross-validation process.

full model is estimated to fit the sample data as closely as possible. Second, both shrunken models outperform the full model in out-of-sample prediction accuracy. The Lasso-shrunken model is the best among the three, with an out-of-sample deviance of 2,586.82, <2,626.77 of the shrunken model obtained using ridge regression.

The shrinkage effect is more obvious in the Lasso-shrunken model. Many of the coefficients are forced to zero, including the insignificant ones identified in the full model. This indicates that those variables are not important in assessing the claim counts, and once removed, the out-of-sample prediction of the model is greatly improved. One possible explanation is that the full model overfits the sample data and thus underperforms shrunken models in making predictions. With fewer explanatory variables, the shrunken model is also much easier to interpret.

The shrinkage effect is not as obvious in the model regularised by ridge regression and none of the coefficients is zero after the shrinkage process. However, many coefficient values are more close to zero than in the full model, while the more significant variables, such the intercept, have a higher absolute coefficient and are still significant. This may explain why the shrunken model also outperforms the full model even when it uses a similar set of variables. Some coefficients of the regression may be reduced as ridge regression can be applied to treat the problem of collinearity between independent variables (see García *et al.*, 2015). In this study, we use categorical variables with values of 0 or 1, which may still lead to some potential for collinearity, for example, between the policyholder's age and driving experience measured in years. As a result, treating the problem of collinearity may further improve the out-of-sample prediction accuracy.

The different results from the Lasso and ridge regression can also be explained with reference to Figure 3, which is similar to that in James *et al.* (2013: Chapter 6, 222). The graph on the left refers to a two-dimensional coefficient scope of the Lasso, and the graph on the right represents the ridge regression. In both graphs, the dot inside the ellipses indicates the maximum likelihood estimator $\hat{\beta}$ without any shrinkage penalty. Assuming the same constraint amount s is used in both

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Table 6. Modelling result for the original full bivariate negative binomial regression model and shrunken models.

Variables	Full model	the Lasso	Ridge regression
ω	0	13	4
Y_1(third-party liability claim)			
Intercept	−1.984 (0.570)***	−2.371 (0.283)***	−2.418 (0.297)***
v1	−0.114 (0.431)	0 (0.009)	−0.008 (0.241)
v2	−0.070 (0.376)	0 (0.009)	−0.018 (0.225)
v3	0.003 (0.108)	0 (0.009)	0.009 (0.103)
v4	0.122 (0.115)	0 (0.009)	0.118 (0.108)
v5	−0.341 (0.295)	−0.012 (0.274)	−0.124 (0.203)
v6	−0.865 (0.501)*	−0.131 (0.268)	−0.224 (0.254)
v7	0.053 (0.437)*	0.082 (0.121)	0.026 (0.230)
v8	−0.655 (0.592)	0 (0.009)	−0.081 (0.258)
v9	−0.012 (0.132)	0 (0.009)	−0.059 (0.122)
v10	0.173 (0.099)*	0.058 (0.091)	0.141 (0.094)
v11	−0.157 (0.433)*	0 (0.009)	−0.023 (0.230)
v1v2	−0.056 (0.253)	0 (0.009)	−0.042 (0.194)
v1v6	0.452 (0.277)	0 (0.009)	0.268 (0.201)
v1v7	0.011 (0.322)	0 (0.009)	0.026 (0.213)
v1v8	−0.100 (0.380)	0 (0.009)	−0.127 (0.246)
v1v11	−0.042 (0.282)	0 (0.009)	−0.053 (0.203)
v2v6	0.023 (0.241)	0 (0.009)	−0.062 (0.175)
v2v7	−0.255 (0.281)	0 (0.009)	−0.141 (0.189)
v2v8	0.255 (0.366)	0 (0.009)	0.071 (0.229)
v2v11	0.335 (0.241)	0 (0.009)	0.241 (0.177)
v6v7	0.068 (0.277)	−0.002 (0.025)	−0.105 (0.186)
v6v11	0.356 (0.295)	0 (0.009)	0.140 (0.189)
v7v8	0.666 (0.397)*	0 (0.009)	0.275 (0.233)
v7v11	−0.170 (0.344)	0 (0.009)	−0.061 (0.200)
v8v11	−0.070 (0.430)	0 (0.009)	−0.159 (0.238)
m_1	6.454 (0.649)***	6.459 (0.643)***	6.501 (0.653)***
Y_2 (comprehensive cover claim)			
Intercept	−5.104 (0.629)***	−4.073 (0.328)***	−3.895 (0.294)***
v1	0.068 (0.400)	0 (0.009)	−0.001 (0.229)
v2	0.489 (0.346)	0.067 (0.081)	0.158 (0.211)
v3	0.136 (0.084)	0.159 (0.085)*	0.184 (0.080)***
v4	−0.257 (0.108)	−0.133 (0.105)	−0.219 (0.101)**
v5	0.421 (0.314)	0.163 (0.084)*	0.270 (0.198)
v6	0.373 (0.489)	0 (0.009)	0.061 (0.245)
v7	0.578 (0.477)	0.010 (0.331)	−0.022 (0.221)
v8	0.209 (0.605)	0 (0.009)	−0.018 (0.249)
v9	2.946 (0.130)***	2.756 (0.123)***	2.509 (0.107)***
v10	1.941 (0.126)***	1.748 (0.120)***	1.545 (0.104)***
v11	0.848 (0.497)	0.375 (0.323)	0.260 (0.224)
v1v2	−0.351 (0.225)	0 (0.009)	−0.215 (0.178)
v1v6	−0.080 (0.236)	0 (0.009)	−0.049 (0.183)
v1v7	0.274 (0.275)	0 (0.009)	0.174 (0.194)
v1v8	0.142 (0.327)	0 (0.009)	0.073 (0.229)
v1v11	−0.142 (0.283)	0 (0.009)	−0.079 (0.195)
v2v6	−0.165 (0.203)	0 (0.009)	−0.065 (0.155)
v2v7	−0.213 (0.228)	0 (0.009)	−0.104 (0.166)
v2v8	−0.406 (0.310)	0 (0.009)	−0.212 (0.210)
v2v11	−0.023 (0.244)	0 (0.009)	0.117 (0.170)

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Table 6. *Continued*

Variables	Full model	the Lasso	Ridge regression
v6v7	0.080 (0.216)	0 (0.009)	0.104 (0.162)
v6v11	−0.102 (0.293)	0 (0.009)	−0.043 (0.184)
v7v8	−0.276 (0.305)	0 (0.009)	−0.174 (0.210)
v7v11	−0.884 (0.411)	−0.361 (0.345)	−0.340 (0.194)*
v8v11	0.177 (0.505)	0 (0.009)	0.186 (0.235)
m_2	2.532 (0.254)***	2.511 (0.253)***	2.545 (0.255)***
λ	5.663 (0.396)***	3.797 (0.429)***	5.774 (0.413)***
In-sample log-likelihood	−5,556.90	−5,587.42	−5,567.22
Out-of-sample log-likelihood	−2,605.69	−2,491.68	−2,493.17
In-sample deviance	5,215.18	5,228.46	5,228.87
Out-of-sample deviance	2,854.57	2,586.82	2,626.77

Note: The coefficients of each variable are shown, followed by their standard deviation in parentheses.
*,**,*** represent, respectively, statistical significance at the 10%, 5% and 1% level, calculated based on the *t*-statistics of coefficients of each variable.

methods, this means $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, which can be represented by the grey area. If s is large enough to reach $\hat{\beta}$, the Lasso and ridge regression estimates will be the same as the maximum likelihood estimates (e.g. when $\omega = 0$).

The ellipses around $\hat{\beta}$ represent regions of constant log-likelihood. The ellipses will expand away from $\hat{\beta}$ and touch the grey constraint area to satisfy the imposed shrinkage penalty. During this process, the Lasso is very likely to end up on one axis while ridge regression will land on the sphere, both shown in the graph as the cross. As a result, in the Lasso selection coefficients are commonly set to zero, while the same cannot be said for ridge regression. This simple graphical example can be extended to the higher dimensional case, when many Lasso estimated coefficients are equal to zero simultaneously.

To support the discussion and to show how the coefficient values react under the two shrinkage techniques, we present the shrunken coefficients at different ω values from the cross-validation procedure, computed as the average value across the ten different models, each fitted when one group is held as the validation set. Note that we only plot the coefficients of explanatory variables, which are directly reduced in the shrinkage process. Figures 4 and 5 show the results from the Lasso and ridge regression, respectively, and present how the 25 coefficients change when the shrinkage parameter increases from 0 to 50 for Y_1 and Y_2 in separate graphes. As expected, all coefficients decrease with an increasing shrinkage parameter. They behave differently for Y_1 and Y_2, with some persistent coefficients significantly different from zero even at large ω values. These are specifically labelled on the figures.

However, it is quite noticeable that when ω is very large (i.e. set to 50), the coefficients in Figure 4 for the Lasso are much closer to zero, compared with those found in ridge regression in Figure 5. In particular, it can be observed that although the coefficients in Figure 5 approach zero initially and a few of them eventually become very close to zero in the end, most coefficients keep a constant distance away from zero which lasts to the end. The findings confirm the discussion made previously, that the two shrinkage techniques affect the coefficients in much distinctive ways.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

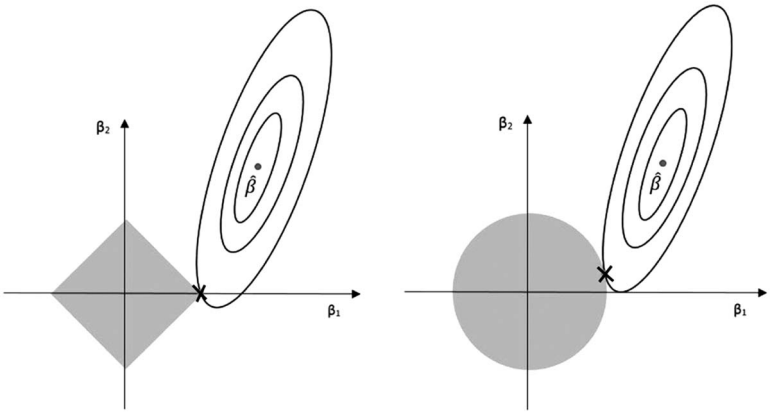

Figure 3. Comparison of the least absolute shrinkage and selection operator (Lasso) (left) and ridge regression (right).

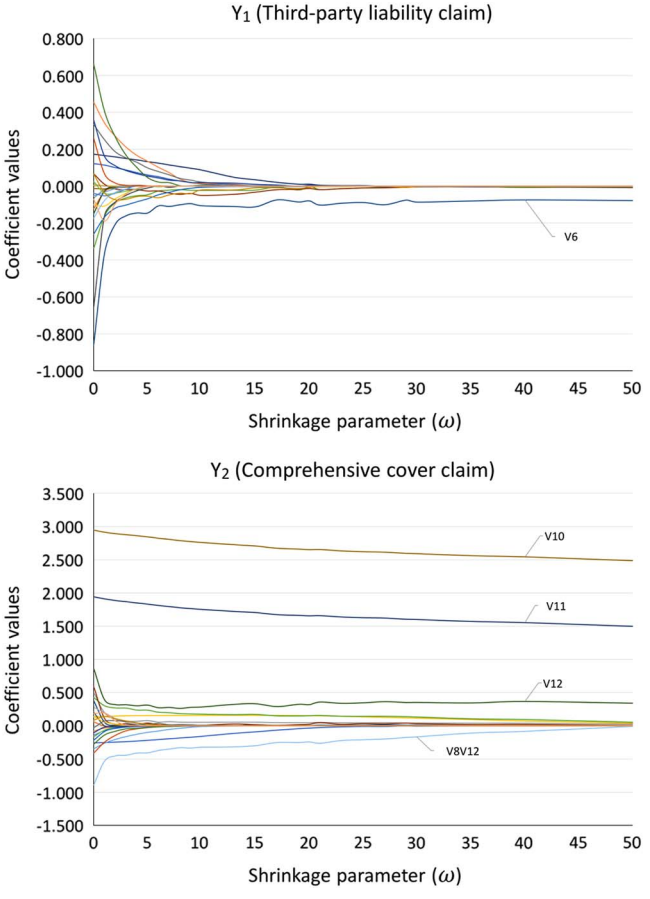

Figure 4. Shrunken coefficients: the least absolute shrinkage and selection operator (Lasso).

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

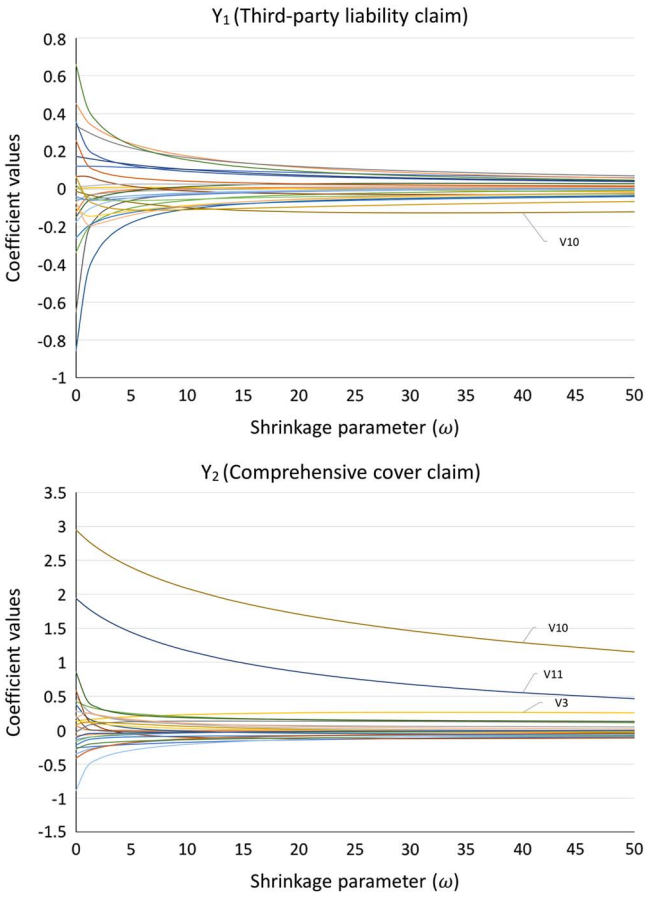

Figure 5. Shrunken coefficients: ridge regression.

The two shrinkage techniques are also applied to UNBR models in a similar way. For each response variable, two shrunken models are generated at given ω values selected by cross-validations. The results are presented in Table 7. Two full UNBR models estimated previously are also presented here.

Similar conclusions can be drawn from the shrunken UNBR models. For Y_1, both of the two shrunken models outperform the full model in out-of-sample prediction, implied by lower deviances. For Y_2, although the full model provides the best in-sample goodness-of-fit, it underperforms the shrunken models out-of-sample, with a lower log-likelihood and higher deviance.

By comparing the results for the Lasso-shrunken BNBR model in Table 6 and the two Lasso-shrunken UNBR models in Table 7, we see that the in-sample deviance of the BNBR model is much lower than the deviances from the two UNBR models combined, implying a better in-sample goodness-of-fit. The out-of-sample deviances are similar for the BNBR model and UNBR models. Obtained using ridge regression, the shrunken BNBR model outperforms the two shrunken UNBR models, providing both lower in-sample and out-of-sample deviances. This is in consistent with the conclusion we draw from the full models. It is beneficial to analyse the two response variables together in a bivariate model and properly account for the correlation structure between them.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Table 7. Modelling results of the original full univariate negative binomial regression (UNBR) model and UNBR models shrunken by the two methods.

	Y_1 (third-party liability claim)			Y_2 (comprehensive cover claim)		
Variables	UNBR	The Lasso	Ridge regression	UNBR	The Lasso	Ridge regression
ω	0	7	29	0	23	2
Intercept	−1.896 (0.573)***	−2.469 (0.206)***	−2.542 (0.163)***	−5.041 (0.64)***	−4.124 (0.349)***	−4.111 (0.373)***
v1	−0.186 (0.434)	0 (0.014)	0.008 (0.110)	0.039 (0.412)	0 (0.007)	−0.014 (0.283)
v2	−0.112 (0.377)	0 (0.012)	0.006 (0.106)	0.486 (0.355)	0.070 (0.085)	0.221 (0.256)
v3	0.036 (0.109)	0 (0.018)	0.021 (0.082)	0.151 (0.087)	0.151 (0.088)*	0.178 (0.084)**
v4	0.113 (0.115)	0.020 (0.110)	0.064 (0.085)	−0.293 (0.108)***	−0.024 (0.106)	−0.269 (0.105)***
v5	−0.374 (0.300)	−0.001 (0.012)	−0.013 (0.107)	0.431 (0.323)	0.134 (0.091)	0.303 (0.240)
v6	−0.952 (0.507)*	−0.101 (0.273)	−0.058 (0.112)	0.252 (0.506)	0 (0.008)	0.015 (0.312)
v7	0.056 (0.438)	−0.054 (0.243)	−0.032 (0.107)	0.572 (0.476)	0 (0.352)	0.060 (0.279)
v8	−0.064 (0.594)	0 (0.013)	0.003 (0.114)	0.194 (0.609)	0 (0.161)	0.008 (0.326)
v9	0 (0.133)	0 (0.013)	−0.011 (0.091)	2.943 (0.131)***	2.656 (0.122)***	2.692 (0.117)***
v10	0.179 (0.099)*	0.120 (0.092)	0.122 (0.077)	1.948 (0.127)***	1.690 (0.119)***	1.719 (0.113)***
v11	−0.198 (0.434)	0 (0.017)	0.032 (0.107)	0.843 (0.495)*	0.416 (0.344)	0.335 (0.283)
v1v2	−0.017 (0.252)	0 (0.013)	−0.011 (0.110)	−0.331 (0.222)	0 (0.007)	−0.249 (0.198)
v1v6	0.481 (0.277)*	0.070 (0.164)	0.097 (0.111)	−0.039 (0.239)	0 (0.007)	−0.020 (0.205)
v1v7	0.019 (0.322)	0 (0.015)	0.026 (0.110)	0.246 (0.275)	0 (0.007)	0.197 (0.224)
v1v8	−0.086 (0.382)	0 (0.013)	−0.040 (0.121)	0.132 (0.327)	0 (0.270)	0.076 (0.266)
v1v11	−0.027 (0.282)	0 (0.013)	−0.010 (0.109)	−0.140 (0.283)	0 (0.008)	−0.096 (0.224)
v2v6	0.057 (0.242)	0 (0.013)	−0.029 (0.100)	−0.143 (0.207)	0 (0.013)	−0.065 (0.175)
v2v7	−0.253 (0.282)	−0.041 (0.166)	−0.064 (0.099)	−0.209 (0.231)	0 (0.023)	−0.125 (0.191)
v2v8	0.244 (0.369)	0 (0.254)	0.022 (0.115)	−0.506 (0.314)	0 (0.007)	−0.316 (0.248)
v2v11	0.360 (0.242)	0.077 (0.161)	0.109 (0.099)	−0.024 (0.250)	0 (0.011)	0.093 (0.196)
v6v7	0.059 (0.279)	−0.032 (0.259)	−0.085 (0.102)	0.098 (0.231)	(0.009)	0.123 (0.188)
v6v11	0.381 (0.296)	0.010 (0.174)	0.052 (0.102)	−0.034 (0.297)	0 (0.007)	−0.004 (0.215)
v7v8	0.634 (0.400)	0 (0.257)	0.071 (0.115)	−0.199 (0.312)	0 (0.007)	−0.134 (0.245)
v7v11	−0.179 (0.346)	0 (0.013)	−0.024 (0.099)	−0.924 (0.414)**	−0.289 (0.365)	−0.458 (0.243)*
v8v11	−0.044 (0.434)	0 (0.012)	−0.028 (0.115)	0.177 (0.513)	0 (0.008)	0.184 (0.299)
m	6.440 (0.648)***	6.47 (0.666)***	6.571 (0.658)***	2.504 (0.254)***	2.547 (0.269)***	2.511 (0.254)***
In-sample log-likelihood	−2,725.21	−2,732.89	−2,731.50	−2,945.76	−2,961.10	−2,949.04
Out-of-sample log-likelihood	−1,136.68	−1,112.983	−1,111.99	−1,519.55	−1,400.59	−1,453.35
In-sample deviance	2,384.60	2,394.80	2,377.77	2,865.60	2,882.55	2,869.90
Out-of-sample deviance	1,025.36	976.22	968.92	1,851.15	1,608.78	1,717.86

Note: The coefficients of each variable are shown, followed by their standard deviation in parentheses.
*,**,*** represent, respectively, statistical significance at the 10%, 5% and 1% level, calculated based on the t-statistics of coefficients of each variable.

6. Conclusion

In this paper we use the BNBR model to analyse general insurance claim data. We show that with a more flexibly specified correlation structure, the BNBR model adequately captures the relationship between the two claim counts and the set of explanatory variables. The correlation, which is totally ignored if two UNBR are fitted separately, proves to be essential in analysing the two types of claim counts from the same policyholder. Note that the correlation coefficient between the two claim count is only 0.187 in this study which is considered as a weak correlation. When a higher correlation coefficient is present, it is likely that a bivariate model with a proper specification of the correlation structure is more suitable than a univariate model.

In addition, we apply two shrinkage techniques to choose core independent variables in modelling claim counts. The results from the Lasso and ridge regression are different, but both shrunken models outperform original full regression models which are likely to suffer from the overfitting problem. The shrunken models provide much better out-of-sample prediction accuracy in both UNBR and BNBR models. This automatic approach to model selection has considerable potential for application in actuarial modelling where very large numbers of variables and data points are often available. Moreover, the shrunken BNBR models also outperforms the two separately fitted shrunken UNBR models, which again emphasises the importance of properly accounting for the correlation structure between response variables.

In addition to BNBR model in this study, some extended Poisson model can also incorporate over-dispersion. For example, the zero-inflated versions of multivariate Poisson models used in Bermúdez & Karlis (2011), where the correlation structure in equation (2) can be implemented instead of the full covariance specification. The bivariate generalised Poisson regression model in Famoye (2010*a*) follows a similar correlation structure as in this study, which also allows for over-dispersion. These potential alternative models can be considered in future research.

Acknowledgment

The authors are grateful to the anonymous reviewers. Their comments improved the quality of the paper.

References

Bermúdez, L. & Karlis, D. (2011). Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, **48**(2), 226–236.

Bolancé, C., Guillén, M. & Pinquet, J. (2003). Time-varying credibility for frequency risk models: estimation and tests for autoregressive specifications on the random effects. *Insurance: Mathematics and Economics*, **33**(2), 273–282.

Bolancé, C., Guillén, M. & Pinquet, J. (2008). On the link between credibility and frequency premium. *Insurance: Mathematics and Economics*, **43**(2), 209–213.

Boucher, J.-P. & Denuit, M. (2008). Credibility premiums for the zero-inflated Poisson model and new hunger for bonus interpretation. *Insurance: Mathematics and Economics*, **42**(2), 727–735.

Boucher, J.-P., Denuit, M. & Guillén, M. (2007). Risk classification for claim counts: a comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*, **11**(4), 110–131.

Boucher, J.-P., Denuit, M. & Guillen, M. (2009). Number of accidents or number of claims? An approach with zero-inflated Poisson models for panel data. *Journal of Risk and Insurance*, **76**(4), 821–846.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press

Brouhns, N., Guillén, M., Denuit, M. & Pinquet, J. (2003). Bonus-Malus scales in segmented tariffs with stochastic migration between segments. *Journal of Risk and Insurance*, **70**(4), 577–599.

Cameron, A.C., Li, T., Trivedi, P.K. & Zimmer, D.M. (2004). Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal*, **7**(2), 566–584.

Cameron, A.C. & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, UK.

Cameron, A.C. & Trivedi, P.K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York, USA.

Chen, Y. & Hanson, T. (2017). Copula regression models for discrete and mixed bivariate responses. *Journal of Statistical Theory and Practice*, **11**, 1–16.

Czado, C., Kastenmeier, R., Brechmann, E.C. & Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, **2012**(4), 278–305.

Denuit, M., Dhaene, J., Goovaerts, M. & Kaas, R. (2006). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. John Wiley & Sons, West Sussex, UK.

Denuit, M. & Lang, S. (2004). Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics*, **35**(3), 627–647.

Denuit, M., Maréchal, X., Pitrebois, S. & Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons, West Sussex, UK.

Denuit, M., Van Keilegom, I., Purcaru, O. *et al.* (2006). Bivariate Archimedean copula models for censored data in non-life insurance. *Journal of Actuarial Practice*, **13**, 5–32.

Dionne, G. & Vanasse, C. (1989). A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. *Astin Bulletin*, **19**(2), 199–212.

Douak, F., Melgani, F. & Benoudjit, N. (2013). Kernel ridge regression with active learning for wind speed prediction. *Applied Energy*, **103**, 328–340.

El-Basyouny, K. & Sayed, T. (2009). Collision prediction models using multivariate Poissonlog-normal regression. *Accident Analysis & Prevention*, **41**(4), 820–828.

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer Science & Business Media, Berlin, Germany.

Famoye, F. (2010*a*). A new bivariate generalized Poisson distribution. *Statistica Neerlandica*, **64**(1), 112–124.

Famoye, F. (2010*b*). On the bivariate negative binomial regression model. *Journal of Applied Statistics*, **37**(6), 969–981.

Frees, E.W. & Valdez, E.A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, **2**(1), 1–25.

Frees, E.W. & Valdez, E.A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, **103**(484), 1457–1469.

García, C., García, J., López Martín, M. & Salmerón, R. (2015). Collinearity: revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics*, **42**(3), 648–661.

Haberman, S. & Renshaw, A.E. (1996). Generalized linear models and actuarial science. *The Statistician*, **45**(4), 407–436.

Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models. Volume 43*. CRC Press, Boca Raton, USA.

Heller, G.Z., Mikis Stasinopoulos, D., Rigby, R.A. & De Jong, P. (2007). Mean and dispersion modelling for policy claims costs. *Scandinavian Actuarial Journal*, **2007**(4), 281–292.

Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Hürlimann, W. (1990). On maximum likelihood estimation for count data models. *Insurance: Mathematics and Economics*, **9**(1), 39–49.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning. Volume 112*. Springer, New York, USA.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1997). *Discrete Multivariate Distributions. Volume 165*. Wiley, New York, USA.

Jung, R.C. & Winkelmann, R. (1993). Two aspects of labor mobility: a bivariate Poisson regression approach. *Empirical Economics*, **18**(3), 543–556.

Karlis, D. & Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review*, **73**(1), 35–58.

King, G. (1989). A seemingly unrelated Poisson regression model. *Sociological Methods & Research*, **17**(3), 235–255.

Kocherlakota, S. & Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Marcel Dekker, New York.

Kocherlakota, S. & Kocherlakota, K. (2001). Regression in the bivariate Poisson distribution. *Communications in Statistics. Theory and Methods*, **30**(5), 815–825.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 1995 International Joint Conference on Artificial Intelligence, pp. 1137–1143.

Lakshminarayana, J., Pandit, S. & Srinivasa Rao, K. (1999). On a bivariate Poisson distribution. *Communications in Statistics-Theory and Methods*, **28**(2), 267–276.

Ma, J., Kockelman, K.M. & Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention*, **40**(3), 964–975.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models. Volume 37*. CRC Press, Boca Raton, USA.

Meijer, R.J. & Goeman, J.J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, **55**(2), 141–155.

Park, M.Y. & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(4), 659–677.

Renshaw, A. (1995). Modelling the claims process in the presence of covariates. *Insurance Mathematics and Economics*, **2**(16), 167.

Samson, D. & Thomas, H. (1987). Linear models as aids in insurance decision making: the estimation of automobile insurance claims. *Journal of Business Research*, **15**(3), 247–256.

Shen, X., Alam, M., Fikse, F. & Rönnegård, L. (2013). A novel generalized ridge regression method for quantitative genetics. *Genetics*, **193**(4), 1255–1268.

Shi, P. & Valdez, E.A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, **55**, 18–29.

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**(6), 1112–1127.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Wang, Z., Ma, S. & Wang, C.-Y. (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biometrical Journal*, **57**(5), 867–884.

Yip, K.C. & Yau, K.K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, **36**(2), 153–163.

https://doi.org/10.1017/S1748499517000070 Published online by Cambridge University Press