DEPENDENT MULTI-PERIL RATEMAKING MODELS

BY

EDWARD W. (JED) FREES[1], GLENN MEYERS[2] AND A. DAVID CUMMINGS[2]

ABSTRACT

This paper considers insurance claims that are available by cause of loss, or peril. Using this multi-peril information, we investigate multivariate frequency and severity models, emphasizing alternative dependency structures. Although dependency models may be used for many risk management strategies, we focus on ratemaking.

Motivation for this research comes from homeowners insurance and so, for the frequency portion, we consider binary response models. Specifically, we examine several multivariate binary regression models that have appeared in the biomedical literature, focusing on a dependence ratio model. For multivariate severity, we use Gaussian copulas to represent dependencies among gamma regressions.

We calibrate competing models based on a representative sample of over 400,000 records and validate them using a held-out sample of over 350,000 records. We find that methods that allow for cross-dependencies among perils provide important economic value in pricing.

KEYWORDS

Copulas, multivariate binary regression, insurance pricing.

1. INTRODUCTION

This paper explores the use of modern statistical predictive models that can be used for pricing and ratemaking in personal lines insurance. Specifically, we focus on homeowners insurance although, as described below, the range of applications is broader. Homeowners represents a large segment of the personal property and casualty (general) insurance business; for example, in the US, homeowners accounted for 12% of all property and casualty insurance premiums and 25% of personal lines insurance, for a total of over $53 billions of US dollars (*I.I.I. Insurance Fact Book 2007*).

[1] University of Wisconsin and Insurance Services Office.
[2] Insurance Services Office.

Astin Bulletin 40(2), 699-726. doi: 10.2143/AST.40.2.2061134 © 2010 by Astin Bulletin. All rights reserved.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

In the traditional actuarial literature (e.g., Bowers et al. 1997, Chapter 2), ratemaking is based on the *individual risk model*. This model decomposes a short-term risk, such as a homeowners claim, into frequency and amount (known as "severity") components. Specifically, let r_i be a binary variable indicating whether or not the ith policyholder has an insurance claim and y_i describe the amount of the claim. Then, the claim is modeled as

$$(claim\ recorded)_i = r_i \times y_i.$$

This is the basis of the individual risk model.

In homeowners, typically insurers have available many characteristics of the home upon which rates are based. For notation, let $\mathbf{x}_i$ be a complete set of explanatory variables that are available to the analyst. The approach adopted here is often known as a "frequency-severity model", where models are specified for both the frequency and severity components. For example, for the frequency component we might fit a logit regression model with r_i as the dependent variable and $\mathbf{x}_{1i}$ as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_1$. For the severity component, we condition on the occurrence of a claim ($r_i = 1$), and might use a gamma regression model with y_i as the dependent variable and $\mathbf{x}_{2i}$ as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_2$.

In this frequency-severity, also known as a two-part, model, one need not have the same set of explanatory variables influencing the frequency and amount of response. However, it is not uncommon to find overlap in the sets of explanatory variables, where variables are members of both $\mathbf{x}_1$ and $\mathbf{x}_2$. Typically, one assumes that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are not related so that the joint likelihood of the data can be separated into two components and run separately, as described above.

Many actuaries interested in pricing homeowners insurance are now decomposing the set of dependent variables (r_i, y_i) by *peril*, or cause of loss (e.g., Modlin, 2005). Homeowners is typically sold as an all-risk policy, which covers all causes of loss except those specifically excluded. By decomposing losses into homogenous categories of risk, actuaries seek to get a better understanding of the determinants of each component, resulting in a better overall predictor of losses.

Table 1 illustrates this decomposition for a data set that we will describe further in Section 2. This table displays summary statistics for nine perils from a sample of 404,664 records. This table shows that WaterNonWeather is the most frequently occurring peril whereas Liability is the least frequent. (WaterNonWeather is water damage from causes other than weather, e.g., the bursting of a water pipe in a house.) When a claim occurs, Hail is the most severe peril (according to the median severity) whereas the "Other" category is the least severe. In Table 1, we note that neither the frequency nor the number sum to the totals due to jointly occurring perils within a policy. That is, for each policy, we record the claims amount for each of the nine perils.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

TABLE 1

HOMEOWNERS SUMMARY STATISTICS

Peril	Frequency (in percent)	Number of Claims	Median Claim Amount
Fire	0.310	1,254	4,152
Lightning	0.527	2,134	899
Wind	1.226	4,960	1,315
Hail	0.491	1,985	4,484
WaterWeather	0.776	3,142	1,481
WaterNonWeather	1.332	5,391	2,167
Liability	0.187	757	1,000
Other	0.464	1,877	875
Theft-Vandalism	0.812	3,287	1,119
Total	5.889*	23,834*	1,661

In a multi-peril model, one decomposes the risk into one of c types ($c = 9$ in Table 1). To set notation, define $r_{i,j}$ to be a binary variable indicating whether or not the ith policy-year has an insurance claim due to the jth type, $j = 1, \ldots, c$. Similarly, $y_{i,j}$ describes the amount of the claim due to the jth type. To relate the multi- to the single-peril variables, we have the following relationships

$$r_i = 1 - (1 - r_{i,1}) \times \ldots \times (1 - r_{i,c}) = \max(r_{i,1}, \ldots, r_{i,c}) \tag{1}$$

and

$$(\textit{claim recorded})_i = y_i = \sum_{j=1}^{c} r_{i,j} \times y_{i,j}. \tag{2}$$

Current actuarial practice involves modeling each peril in isolation of the others. Thus, for example, from the full set of explanatory variables $\mathbf{x}$, the analyst selects a set of variables $\mathbf{x}_{1,j}$ to predict the frequency and another a set $\mathbf{x}_{2,j}$ to predict the severity for each peril, $j = 1, \ldots, c$. This is intuitively appealing because some predictors do well in predicting certain perils but not others. For example, "dwelling in an urban area" may be an excellent predictor for the theft peril but provides little useful information for the hail peril. To implement this modeling strategy, it is straightforward in principle to use a logistic regression for each frequency and gamma regression for each severity.

Although easy to interpret, this procedure uses the same dataset to calibrate several models. From a modeling point of view, this amounts to assuming that perils are independent of one another and that sets of parameters from each peril are unrelated to one another. Although making sets of parameters unrelated to another (sometimes called functionally independent) is plausible, it

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

seems unlikely that perils are independent. Event classification can be ambiguous (e.g., fires triggered by lightning) and unobserved latent characteristics of policyholders (cautious homeowners who are sensitive to potential losses due to theft-vandalism and liability) may induce dependencies among perils. Our preliminary empirical examinations in Section 2 will also suggest that perils may be related to one another.

To investigate potential dependence relationships, we retain the basic hierarchical approach of the frequency-severity model and treat each component as a *multivariate* response. Specifically, we first analyze the frequency component and then model the severity component conditional on the frequency. The multivariate multi-peril model is:

1. Use a multivariate binary regression model with $\mathbf{r}_i = (r_{i,1}, \ldots, r_{i,c})'$ as the dependent variable.
2. Conditional on the frequency $\mathbf{r}_i$, for the severity we specify a multivariate regression with $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,c})'$ as the dependent variable.

We will apply this modeling strategy to homeowners insurance, where a claim type may be due to fire, liability, and so forth. One could also use this strategy to model homeowners and automobile policies jointly. As another example, in healthcare, expenditures are often broken down by diagnostic related groups.

In the actuarial and insurance literatures, there has been a recent surge of interest in modeling short-term coverages at the individual policyholder level, e.g., "micro-level" data. Examples include works by Angers, Desjardins, Dionne and Guertin (2006), Boucher and Denuit (2006), Frees, Peng and Valdez (2009), Loo, Fung and Zhu (2007) and Mahmoudvan and Hassani (2009). However, these papers deal with automobile insurance insurance, not homeowners. Moreover, because they do not split by cause of loss as we do in this paper, for the frequency model they examine count models such as Poisson regression in lieu of the binary logistic regression models in this paper.

Section 3 will introduce our multivariate severity model. It is based on gamma regressions for the marginal peril distributions with a Gaussian copula to quantify the association among severities. In this context, a strength of the copula framework is that marginals are preserved — this is important when we have relatively few joint severities upon which to base our measures of dependence.

Section 4 will introduce multivariate binary regression models, focusing on the dependence ratio approach introduced by Ekholm et al (1995). Appendix Section B describes alternative models, including log-linear, quadratic exponential and alternating logistic regressions.

As we discuss estimation of the models in Sections 3 and 4, we will be able to provide assessments using in-sample measures. In-sample measures include hypothesis testing statistics such as t-statistics and likelihood ratio tests to judge the statistical significance of parameters and goodness of fit statistics such as *AIC* to judge the overall model fit. Section 5 will introduce our validation of models based on a held-out sample that is unrelated to our estimation

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

sample. As discussed in this section and Appendix Section D, insurance claims out-of-sample model validation can be difficult due to the mixture of zeros and positive outcomes, as well as positive outcomes that are skewed and fat-tailed.

Concluding remarks are provided in Section 6.

2. Data

To calibrate our models, we drew two random samples from a homeowners database maintained by the Insurance Services Office. This database contains over 4.2 million policyholder years. It is based on the policies issued by several major insurance companies in the United States, thought to be representative of most geographic areas in the US. These policies were almost all for one year and so we will use a constant exposure (one) for our models.

Our in-sample, or "training", dataset consists of a representative sample of 404,664 records taken from this database. The original database had an oversampling of claims; we adjusted our sampling procedures so that the in-sample dataset could be treated as a random sample from the population of policies. The summary measures in this section are based on this training sample. In Section 5, we will test our calibrated models on a second held-out, or "validation" subsample that was also randomly selected from this database.

Table 1 in Section 1 summarized the main tendencies of frequency and severity. Prior to introducing formal mathematical models that account for dependence, we first examine the homeowners data to establish, or at least suggest, the presence of dependence among perils.

2.1. Severities

Dependence among continuous variables is more well-known than discrete variables, so we begin with the severity portion. Table 2 presents correlations among perils. Because claims distributions are typically right-skewed, entries in the tables are Spearman correlations. Recall that a Spearman correlation is a regular (Pearson) correlation among the ranks, not the actual claims. In this way, they do not depend on the scale and so, for example, would be unchanged if we calculated the correlation of logarithmic claims.

Table 2 shows many positive correlations but also some large negative ones. For example, the correlation between Lightning and Liability is –1! The explanation for this is that correlations are based on observed pairs of severity perils — there are relatively few observations to base these correlations upon. Table 1 showed that there were only 23,834 claims in our data base. Moreover, it turns out that 96.1% of these were from policies with only a single claim. Thus, the number of policies with two or more claims is small, making the estimation of severity correlations imprecise.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

TABLE 2

SPEARMAN CORRELATIONS AMONG SEVERITY PERILS

	Fire	Light-ning	Wind	Hail	Water Weather	Water Non Weather	Liability	Other	Theft Vand
Fire	1.00	0.29	–0.19	0.32	0.50	0.32	0.80	0.19	0.03
Lightning	0.29	1.00	–0.04	0.45	–0.28	0.16	–1.00	0.26	0.53
Wind	–0.19	–0.04	1.00	–0.15	0.11	–0.07	0.62	–0.15	0.12
Hail	0.32	0.45	–0.15	1.00	0.82	0.21	–0.50	–1.00	–0.09
WaterWeath	0.50	–0.28	0.11	0.82	1.00	0.17	–0.46	0.45	0.09
WaterNWeath	0.32	0.16	–0.07	0.21	0.17	1.00	0.39	0.04	0.45
Liability	0.80	–1.00	0.62	–0.50	–0.46	0.39	1.00	0.41	–0.63
Other	0.19	0.26	–0.15	–1.00	0.45	0.04	0.41	1.00	0.41
TheftVand	0.03	0.53	0.12	–0.09	0.09	0.45	–0.63	0.41	1.00

2.2. Frequencies

Table 3 gives the number of joint claims among perils. For example, we see that there were only 3 records that had a Lightning and a Liability claim within the year. The (rank) correlation between Lightning and Liability noted above, –1, is not meaningful as it is based on only 3 observations.

TABLE 3

JOINT CLAIM COUNTS AMONG PERILS

	Fire	Light-ning	Wind	Hail	Water Weather	Water Non Weather	Liability	Other	Theft Vand
Fire	–	4	23	7	23	27	4	16	20
Lightning	4	–	17	11	12	32	3	18	25
Wind	23	17	–	23	62	92	17	45	55
Hail	7	11	23	–	13	43	3	2	16
WaterWeather	23	12	62	13	–	93	7	18	38
WaterNWeath	27	32	92	43	93	–	16	48	71
Liability	4	3	17	3	7	16	–	13	9
Other	16	18	45	2	18	48	13	–	31
TheftVand	20	25	55	16	38	71	9	31	–
Subtotal	131	147	352	118	266	422	72	191	265
Totals	1254	2134	4960	1985	3142	5391	757	1877	3287

Note: Totals refer to all claims from a peril, not just those occurring jointly with another peril.

To measure association in our binary frequency data, correlations can be misleading summary statistics. Instead, we will examine dependence ratios of the form

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

$$\tau_{12} = \frac{\Pr(r_1 = 1, r_2 = 1)}{\Pr(r_1 = 1)\ \Pr(r_2 = 1)} \tag{3}$$

the ratio of the joint probability to the product of the marginal probabilities. In the case of independence, we would expect the dependence ratio τ_{12} to be 1. Values of $\tau_{12} > 1$ indicate positive dependence and values of $\tau_{12} < 1$ indicate negative dependence.

Table 4 provides empirical dependence ratios for each pair of perils. To understand the calculation of this table, consider the relationship between Fire and WaterWeather. To begin, the marginal empirical probability of a fire claim is

$$\widehat{\Pr}(r_{Fire} = 1) = \frac{1254}{404664} = 0.00310,$$

where from Table 3 the number of fire claims is 1,254. Similarly, the probability of a "water due to weather" claim is

$$\widehat{\Pr}(r_{WaterWeather} = 1) = \frac{3142}{404664} = 0.00776.$$

Further, the joint probability of a policy having claims due to both Fire and WaterWeather is

$$\widehat{\Pr}(r_{Fire} = 1, r_{WaterWeather} = 1) = \frac{23}{404664} = 0.00006.$$

Putting these together, the estimated dependence ratio is

$$\hat{\tau}_{Fire,WaterWeather} = \frac{0.00006}{0.00310 \times 0.00776} = 2.362.$$

TABLE 4

DEPENDENCE RATIOS AMONG PERILS

	Fire	Light-ning	Wind	Hail	Water Weather	Water Non Weather	Liability	Other	Theft Vand
Fire	1.000	1.663	1.496	1.138	2.362	1.616	1.705	2.751	1.963
Lightning	1.663	1.000	1.338	1.051	0.724	1.126	0.751	1.818	1.442
Wind	1.496	1.338	1.000	0.945	1.610	1.392	1.832	1.956	1.365
Hail	1.138	1.051	0.945	1.000	0.843	1.626	0.808	0.217	0.992
WaterWeath	2.362	0.724	1.610	0.843	1.000	2.222	1.191	1.235	1.489
WaterNWeath	1.616	1.126	1.392	1.626	2.222	1.000	1.587	1.920	1.621
Liability	1.705	0.751	1.832	0.808	1.191	1.587	1.000	3.702	1.464
Other	2.751	1.818	1.956	0.217	1.235	1.920	3.702	1.000	2.033
TheftVand	1.963	1.442	1.365	0.992	1.489	1.621	1.464	2.033	1.000

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

Table 4 summarizes these calculations for all 36 pairs of perils. Here we see much dependence among perils, typically positive dependence but some negative dependence as well.

The data (and intuition) suggests dependencies among perils. As we have seen, the severity correlations are based on only a few observations — these associations turn out to be hard to estimate. In contrast, for frequency the dependence ratios are based on 404,664 observations. Although the joint probabilities are small, we can get precise estimates.

2.3. Dependencies among Marginal Frequency Models

The data presented (so far) do not account for explanatory variables that could induce correlations. For these effects, we used many explanatory variables for each peril (from 8 for the "Other" peril to 19 for the "Water Weather" peril) for over 100 predictors. These are a variety of geographic-based plus several standard industry variables that account for:

- weather and elevation,
- vicinity,
- commercial and geographic features,
- experience and trend, and
- rating variables.

The web site http://www.iso.com/Products/ISO-Risk-Analyzer/ISO-Risk-Analyzer-Homeowners.html provides more information on these explanatory variables. Because the focus of this paper is on the association aspects, in this paper we summarize the results only for the dependency parameters.

For assessing frequency dependencies, recall that r denotes the binary variable that indicates a claim ($y = 1$). In our sample, we have r_{ij}, $i = 1, \ldots, n = 404{,}664$, and $j = 1, \ldots, c = 9$. Let q_{ij} be the corresponding probability of a claim. The number of claims that is joint between the jth and kth perils is $\sum_{i=1}^{n} r_{ij} \times r_{ik}$. Assuming independence among perils, this has mean and variance

$$E\left(\sum_{i=1}^{n} r_{ij} \times r_{ik}\right) = \sum_{i=1}^{n} q_{ij} \times q_{ik}$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} r_{ij} \times r_{ik}\right) = \sum_{i=1}^{n} q_{ij}\, q_{ik} - \left(q_{ij}\, q_{ik}\right)^2.$$

To assess dependencies among the claim frequencies, we employ the t-statistic

$$t_{jk} = \frac{\sum_{i=1}^{n} r_{ij} \times r_{ik} - \sum_{i=1}^{n} q_{ij} \times q_{ik}}{\sqrt{\sum_{i=1}^{n} q_{ij}\, q_{ik} - \left(q_{ij}\, q_{ik}\right)^2}}. \tag{4}$$

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

TABLE 5

TEST STATISTICS FROM LOGISTIC REGRESSION FITS

	Fire	Light-ning	Wind	Hail	Water Weather	Water Non Weather	Liability	Other	Theft Vand
Fire	–	1.472	1.662	0.754	3.955	2.732	1.023	4.048	3.085
Lightning	1.472	–	1.530	0.247	–1.166	0.837	–0.485	2.229	1.816
Wind	1.662	1.530	–	–1.240	3.185	3.369	2.436	3.919	2.270
Hail	0.754	0.247	–1.240	–	–0.100	1.697	–0.303	–2.616	–0.235
WaterWeath	3.955	–1.166	3.185	–0.100	–	7.429	0.333	0.478	2.227
WaterNWeath	2.732	0.837	3.369	1.697	7.429	–	1.825	4.004	3.503
Liability	1.023	–0.485	2.436	–0.303	0.333	1.825	–	4.929	1.147
Other	4.048	2.229	3.919	–2.616	0.478	4.004	4.929	–	3.766
TheftVand	3.085	1.816	2.270	–0.235	2.227	3.503	1.147	3.766	–

The t-statistic in equation (4) would be a standard two-sample t-statistic except that we allow the probability of a claim to vary by policy i. To estimate these probabilities, we fit a logistic regression model for each peril j, where the explanatory variables are peril-specific. Each model was fit in isolation of the others, thus implicitly using the null hypothesis of independence among perils.

Table 5 summarizes the test statistics for assessing independence among the frequencies. Not surprisingly, the strongest relationship was between water damage due to weather and water damage from causes other than weather. The largest dependence ratio in Table 4, between fire and the "Other" category, was the second largest t-statistic — this indicates strong dependence even after covariates are introduced. Interestingly, the only significant negative relationship was between hail and the "Other" category.

For the degrees of freedom of the t-statistic, we have followed the usual rule of the number of observations minus the number of parameters. Because our sample size is large ($n = 404{,}664$) relative to the number of parameters, the reference distribution is essentially normal. We acknowledge that the asymptotic distribution may be slightly mis-specified because the probabilities q are only known up to the estimated regression parameters. Thus, for some applications, one may wish to determine the reference distribution via alternative means such as bootstrapping. However, given our large sample size and because we are using the statistic only for diagnostic purposes, we recommend this procedure to uncover dependencies among the frequencies.

3. MULTIVARIATE SEVERITY MODEL

To accommodate dependencies among claim severities, we use a parametric copula. A copula allows us to use different gamma regression models for each type, thus permitting a direct comparison with the independence model.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

One could allow the distributions to vary by peril within the copula framework. We use only gamma regression marginals in part to conform with industry practice. See Frees and Wang (2005) for a longitudinal application that employs copulas to relate dependencies among gamma regression models.

3.1. Marginal Distributions

Suppose that there are c potential claims for the ith policy, $\mathbf{y}_i = (y_{i1}, \ldots, y_{ic})'$. The joint distribution function is denoted by

$$F_i(a_{i1}, \ldots, a_{ic}) = \Pr(y_{i1} \le a_{i1}, \ldots, y_{ic} \le a_{ic}),$$

with marginal distribution functions $\mathrm{F}_{ij}(a_{ij}) = \mathrm{F}_{ij} = \Pr(y_{ij} \le a_{ij})$.

The marginals follow a gamma distribution with parameters that vary by peril and covariates that depend on the policy. Specifically, let $\theta_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta}_{2,j}$ be a systematic component where $\mathbf{x}_{ij}$ is a vector of known explanatory variables and $\boldsymbol{\beta}_{2,j}$ is a vector of unknown parameters. Assuming a logarithmic link function, the distribution F_{ij} is specified to be a gamma distribution, with mean $\mu_{ij} = \exp(\theta_{ij})$ and scale parameter that also varies by peril, $scale_j$. Below, we use $\mathrm{f}(., \theta_{ij}, scale_j)$ to denote this density.

3.2. Modeling the Dependence

The joint distribution function of claim severities can be expressed as a function of the marginal distributions through a copula. Suppressing the $\{i\}$ subscripts, let F_j be the distribution function associated with the jth type, y_j. We may write the joint distribution of claims $\mathbf{y} = (y_1, \ldots, y_c)'$ as

$$\begin{aligned} \mathrm{F}(a_1, \ldots, a_c) &= \Pr(y_1, \ldots, y_c \le a_c) \\ &= \Pr(\mathrm{F}_1(y_1) \le \mathrm{F}_1(a_1), \ldots, \mathrm{F}_c(y_c) \le \mathrm{F}_c(a_c)) \\ &= \mathrm{COP}(\mathrm{F}_1(a_1), \ldots, \mathrm{F}_c(a_c)). \end{aligned}$$

Here, $\mathrm{COP}(\cdot)$ is the copula linking the marginals to the joint distribution. See, for example, Frees and Valdez (1998) for an introduction to copulas. Let f_j be the density function associated with the jth type. The multivariate density is

$$\mathrm{f}(a_1, \ldots, a_c) = \mathrm{cop}(\mathrm{F}_1(a_1), \ldots, \mathrm{F}_c(a_c)) \prod_{j=1}^{c} \mathrm{f}_j(a_j). \tag{5}$$

Here, $\mathrm{cop}(.)$ is the density function corresponding to the copula distribution function $\mathrm{COP}(.)$.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

To illustrate, we will work extensively with the normal (also known as the Gaussian) copula. We may write the normal copula as

$$\text{cop}_N(u_1, \ldots, u_c) = \phi_N\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_c)\right) \prod_{j=1}^{c} \frac{1}{\phi(\Phi^{-1}(u_j))}. \quad (6)$$

Here, Φ and ϕ are the distribution and density functions of the standard normal distribution, respectively. The multivariate normal density is

$$\phi_N(\mathbf{z}) = \frac{1}{(2\pi)^{c/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}\mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z}\right).$$

The matrix $\boldsymbol{\Sigma}$ is a correlation matrix, with ones on the diagonal.

3.3. Estimation Results

The copula allows for a fully parametric specification of the probability model. We exploit this specification by using maximum likelihood estimation. We assume independence among policies and use the copula to model dependencies among perils.

Using equation (5), when there are claims from all perils, the log-likelihood of the ith policy is

$$l_i = \sum_{j=1}^{c} \ln f\left(y_{ij}, \theta_{ij}, scale_j\right) + \ln \text{cop}_N(F_{i1}, \ldots, F_{ic}). \quad (7)$$

Gaussian copulas are preserved under the marginals, so having only a subset of perils does not present a difficulty in evaluating the likelihood expression. A broader set of difficulties in the evaluation of the likelihood is summarized in Appendix A.

To estimate the copula model, we use the explanatory variables that were developed under the independence model. As with the frequency model, there were many explanatory variables for each peril and so we summarize the results only for the dependency parameters.

We examined three models of association by varying the specification of the correlation matrix $\boldsymbol{\Sigma}$. In the most complex specification, we allowed $\boldsymbol{\Sigma}$ to be unstructured (subject to being symmetric and invertible), resulting in $\binom{9}{2} = 36$ association parameters to be estimated, one for each pair of perils. In the least complex specification, we specified all association parameters to be equal, so that $\boldsymbol{\Sigma}$ has a structure known as a "uniform correlation" or "compound symmetry" model. As an intermediate choice, we grouped the perils into five classes. As described below, this grouping resulted in twelve association parameters.

Using a single association parameter, the maximum likelihood estimator turned out to be 0.0746 with a corresponding t-statistic equal to 3.256. This indicates statistically significant positive association among claims.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

TABLE 6

COPULA PARAMETER ESTIMATES

Association	Estimate	*t*-statistic
Between Groups 1 and 2	0.0984	1.838
Between Groups 1 and 3	0.2014	1.245
Between Groups 1 and 4	0.1933	1.376
Between Groups 1 and 5	0.1542	1.896
Between Groups 2 and 3	0.0151	0.134
Between Groups 2 and 4	0.0352	0.558
Between Groups 2 and 5	0.1018	1.879
Between Groups 3 and 4	0.3937	1.353
Between Groups 3 and 5	–0.3436	–1.296
Between Groups 4 and 5	0.2127	1.417
Within Group 1	0.0313	0.146
Within Group 2	0.0356	0.943

Note: The five groups are: (1) Fire, Lightning, (2) Wind, Hail, WaterWeather, WaterNonWeather, (3) Liability, (4) Other and (5) Theft.

Results for the twelve parameter model appear in Table 6. Here, we see positive association between most groups and within the two larger groups — the exception is between groups 3 and 5. However, none of the association parameters are strongly statistically significant, despite estimation using 23,384 claims severities. Further, when we estimated the copula model with 36 parameters, only one of the 36 of the parameter estimates turned out to be strongly statistically significant (with a p-value < 0.01 — this was between "Fire" and "Water Nonweather"); thus, these parameter estimates are not reported here. We conjecture that the reason is that there were relatively few policies having joint claims that would contribute to the copula portion of the likelihood (see Table 3).

We also examined alternative copula (e.g., t-copula) and marginal regression specifications (heavier tail than gamma). Because we had relatively few joint claims, the out-of-sample analysis showed that this line of research was not fruitful for our data.

4. DEPENDENCE RATIO MULTIVARIATE FREQUENCY MODEL

Fortunately, in the statistics literature, there are many good approaches to modeling multivariate binary frequencies. To keep this paper contained, Section 4.3 and Appendix Section B provide brief overviews of these methods with appropriate references for readers who wish to explore this topic further.

This paper features the *dependence ratio approach*, introduced by Ekholm, Smith and McDonald (1995). This is a likelihood approach, where the likelihood

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

is written in terms of means and dependence ratios. See also Ekholm, McDonald and Smith (2000).

Suppressing the $\{i\}$subscripts, recall the mean parameters $\pi_j = \mathrm{E}\, r_j = \Pr(r_j = 1)$ and similarly define higher order moments $\pi_{jk} = \mathrm{E}\, r_j r_k = \Pr(r_j = r_k = 1)$, $\pi_{ijk} = \mathrm{E}\, r_i r_j r_k$, …, $\pi_{12\cdots c} = \mathrm{E}\, r_1 r_2 \cdots r_c$. Now, if the responses are independent, then $\pi_{12} = \pi_1 \pi_2$ and so forth. To assess this, as in equation (3) we may define the *dependence ratio*

$$\tau_{12} = \frac{\pi_{12}}{\pi_1 \pi_2} = \frac{\Pr(r_1 = r_2 = 1)}{\Pr(r_1 = 1)\Pr(r_2 = 1)}.$$

Interpret $\tau_{12} \times 100$ to be the percentage that r_1 and r_2 are both one (claims) under the dependence model compared to the independence model.

Similarly, define higher order dependence ratios as

$$\tau_{jk} = \frac{\pi_{jk}}{\pi_j \pi_k}, \tau_{ijk} = \frac{\pi_{ijk}}{\pi_i \pi_j \pi_k}, \ldots, \tau_{12\cdots c} = \frac{\pi_{12\cdots c}}{\pi_1 \pi_2 \cdots \pi_c}.$$

The approach is to use regression covariates to estimate the means π_j and simpler specifications, typically constants, to estimate the dependence ratios.

4.1. Basic Likelihood

We will use maximum likelihood estimation. Writing the likelihood in terms of the means π is straightforward yet tedious. To see some of the difficulties, consider the case with only $c = 3$ perils. Then, we have

$$\begin{aligned}
\Pr(r_1 = 1, r_2 = 1, r_3 = 1) &= \mathrm{E}\, r_1 r_2 r_3 = \pi_{123} \\
\Pr(r_1 = 1, r_2 = 1, r_3 = 0) &= \mathrm{E}\, r_1 r_2 (1 - r_3) = \pi_{12} - \pi_{123} \\
\Pr(r_1 = 1, r_2 = 0, r_3 = 1) &= \pi_{13} - \pi_{123} \\
\Pr(r_1 = 1, r_2 = 0, r_3 = 0) &= \pi_1 - \pi_{12} - \pi_{13} + \pi_{123} \\
&\vdots \\
\Pr(r_1 = 0, r_2 = 0, r_3 = 0) &= 1 - (\pi_1 + \pi_2 + \pi_3) + (\pi_{12} + \pi_{13} + \pi_{23}) - \pi_{123}.
\end{aligned}$$

So this gives each possible likelihood outcome in terms of marginal means π_j and dependency parameters τ_{jk} and τ_{123}.

For $c = 9$ perils, the pattern is similar. Most policies result in zero claims for all perils:

$$\begin{aligned}
&\Pr(r_1 = 0, r_2 = 0, \ldots, r_c = 0) \\
&= 1 - \sum_{j=1}^{c} \pi_j + \sum_{j<k} \pi_{jk} - \sum_{i<j<k} \pi_{ijk} + \ldots + (-1)^c \pi_{12\cdots c}
\end{aligned} \tag{8}$$

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

For a policy with a claim in the first peril and no other claims, we have

$$\Pr(r_1 = 1,\, r_2 = 0, \ldots, r_c = 0)$$
$$= \pi_1 - \sum_{j=2}^{c} \pi_{1j} + \sum_{1<j<k} \pi_{1jk} - \ldots + (-1)^{c-1} \pi_{12\cdots c}.$$

Other policies with singleton claims can be calculated via symmetry. For a claim in the first two perils and no other claims, we have

$$\Pr(r_1 = 1,\, r_2 = 1,\, r_3 = 0, \ldots, r_c = 0)$$
$$= \pi_{12} - \sum_{j=3}^{c} \pi_{12j} + \sum_{2<j<k} \pi_{12jk} - \ldots + (-1)^{c} \pi_{12}.$$

Other policies with two claims can be calculated via symmetry.

For our data set, policies with three and more claims represent an extremely small fraction of the data. Thus, Section 4.2 presents estimates τ_{jk} for each pair of perils (j,k) but assume that higher order ratios, such as τ_{jkl} and τ_{jklm}, are equal to one. Appendix Section C provides further calculation details.

Calculation of the likelihood estimates uses the independence model to provide initial values. For some data sets, there could be a small issue with constraints on the optimization. For example, because $r_2 \le 1$, we have $\pi_{12} = \mathrm{E}\, r_1 r_2 \le \mathrm{E}\, r_1 = \pi_1$. Thus,

$$\pi_{12} \le \min(\pi_1, \pi_2), \quad \text{and} \quad \tau_{12} \le \min\left(\frac{1}{\pi_1}, \frac{1}{\pi_2}\right).$$

Because of the small size of our π_js, this has not been an issue for the homeowners application.

4.2. Estimation Results

Our focus is on estimation of association between pairs of perils. For consistency with the severity section, we consider three models of association, a single parameter model, a model with 36 parameters, one for each pair of perils, and an intermediate version formed by taking groups of perils. Also for consistency with the severity portion, we estimate models including regression covariates but do not report on this portion of the results.

For the single association parameter, the maximum likelihood estimator turned out to be 1.3325 with a standard error of 0.0386, 8.61 standard errors above 1. This is clearly statistically significant and indicates positive dependence.

Results for a 12 parameter "intermediate" model are summarized in Table 7. Here, we see that all of the parameter estimates indicated positive dependence and most are statistically significant. To link this to the summary statistics

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

TABLE 7

DEPENDENCE RATIO PARAMETER ESTIMATES

Association	Estimate	*t*-statistic
Between Groups 1 and 2	1.2182	2.393
Between Groups 1 and 3	1.1487	0.369
Between Groups 1 and 4	1.8476	2.649
Between Groups 1 and 5	1.4826	2.159
Between Groups 2 and 3	1.4698	2.288
Between Groups 2 and 4	1.4090	3.200
Between Groups 2 and 5	1.2562	2.788
Between Groups 3 and 4	3.4071	2.706
Between Groups 3 and 5	1.4571	0.993
Between Groups 4 and 5	1.6776	2.172
Within Group 1	1.3236	0.774
Within Group 2	1.3956	5.404

Notes: The five groups are: (1) Fire, Lightning, (2) Wind, Hail, WaterWeather, WaterNonWeather, (3) Liability, (4) Other and (5) Theft.
The t-statistic provides the number of standard errors that the estimate differs from 1, the null value under the independence hypothesis.

presented in Table 4, Table 8 gives the parameter estimates in a matrix format. This allows one to compare the maximum likelihood results from our dependence ratio model that includes covariate information to the empirical dependence ratios that are calculated without covariate information.

We also calculated the 36 parameter dependency ratio model where each pair of perils had a unique parameter. These maximum likelihood estimates also turned out to be close to the empirical ones presented in Table 4 and so are not presented here.

TABLE 8

MATRIX OF DEPENDENCE RATIO PARAMETER ESTIMATES

	Fire	Light-ning	Wind	Hail	Water Weather	Water Non Weather	Liability	Other	Theft
Fire	1	1.3236	1.2182	1.2182	1.2182	1.2182	1.1487	1.8476	1.4826
Lightning	1.3236	1	1.2182	1.2182	1.2182	1.2182	1.1487	1.8476	1.4826
Wind	1.2182	1.2182	1	1.3956	1.3956	1.3956	1.4698	1.4090	1.2562
Hail	1.2182	1.2182	1.3236	1	1.3956	1.3956	1.4698	1.4090	1.2562
WaterWeath	1.2182	1.2182	1.3236	1.3956	1	1.3956	1.4698	1.4090	1.2562
WaterNWeath	1.2182	1.2182	1.3236	1.3956	1.3956	1	1.4698	1.4090	1.2562
Liability	1.1487	1.1487	1.4698	1.4698	1.4698	1.4698	1	3.4071	1.4571
Other	1.8476	1.8476	1.4090	1.4090	1.4090	1.4090	3.4071	1	1.6776
Theft	1.4826	1.4826	1.2562	1.2562	1.2562	1.2562	1.4571	1.6776	1

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

4.3. Alternative Multivariate Models

There are many approaches for modeling multivariate variables with none being uniformly superior to the others. This section documents our findings in exploring alternative models for multivariate claims.

One widely used multivariate frequency model that we did not empirically estimate is the multivariate probit. To fit this model using likelihood methods, one needs to compute a multivariate normal distribution function for each policyholder. For our case, this would involve a 9-dimensional multivariate normal distribution function evaluation for each of the 404,664 records. Because of these computational issues, multivariate probits were not further explored.

We did explore a method for multivariate frequency that is widely used in the biomedical community known as "alternating logistic regressions." Alternating logistic regressions, also known by the acronym ALR, uses generalized estimating equations (GEE) methods to estimate binary dependencies. This method is attractive in that it is available in the commercial statistical software SAS. However, for the number of subjects considered here it is coded inefficiently and this canned procedure is not a reasonable alternative. More importantly, as a GEE method, alternating logistic regression does not probabilistically model dependencies but rather allows for them in the moment structure to enhance the efficiency of estimators. For data sets of the size we are considering, efficiency is less of an issue. Moreover, the lack of a probabilistic model makes it more difficult to specify an optimal predictor, a key goal of ratemaking. We do not report further results for ALR, although Appendix Section B3 contains a brief overview with references for readers wishing to learn more about the alternating logistic regression approach.

5. Out-of-Sample Validation

In predictive modeling, one validates a model by examining performance on an independent held-out sample of data (e.g., Hastie, Tibshirani and Friedman, 2001). Standard performance measures include an assessment of bias, root mean square error and related measures. In insurance claims, these standard measures are not the most informative due to the high proportions of zeros (corresponding to no claim) and the skewed, fat tailed distribution of the positive values. Appendix Section D discusses this point and develops a summary measure that we will use in this paper.

We call this measure a "Gini" index, after a standard measure of income inequality. As described in Appendix Section D, this measure summarizes the typical profitability that an insurer will enjoy for a held-out sample when taking on a new scoring method relative to an existing pricing structure. The larger the index, the more effective is the scoring mechanism. We do not have actual prices charged by the insurance companies. Moreover, because these are intercompany data, even if we did it is not clear that these prices would be

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

TABLE 9

COMPARISON OF ALTERNATIVE APPROACHES TO INDEPENDENCE MODEL

Frequency Model	Severity Model	Gini Index
Independence	Independence	–
Independence	Copula with 36 parameters	–1.477
Independence	Copula with 1 parameter	–0.272
Dep Ratio with 1 parameter	Independence	2.478
Dep Ratio with 12 parameters	Independence	1.673
Dep Ratio with 36 parameters	Independence	2.322

comparable because different companies use different pricing strategies. Instead, as our "base" price we use the "independence" model predicted values computed assuming no dependence among perils. Appendix Section D provides additional details.

Table 9 summarizes the out-of-sample performance of the several approaches to dependence modeling considered in this paper. Here, we see that introducing association among severities using the copula framework provides no additional predictive ability. This is not surprising — as described in Section 3, because of the large number of perils (9), we have relatively few policies with joint severities, meaning that it is difficult to assess their association. For other applications, it would certainly make sense to investigate other parametric specifications or choices of copulas. Because of this lack of predictive ability, Table 9 reports only two models using copulas.

Table 9 shows that introducing dependence ratios to model multivariate binary responses provides additional predictive power. All three specifications performed well on an out-of-sample basis. Of these three, the one parameter specification is preferred based on the principle of parsimony and out-of-sample performance.

6. SUMMARY AND CONCLUDING REMARKS

This work provides evidence regarding the effects of dependence on multi-peril multivariate frequency and severity models. Our foundation essentially assumes independence among perils. We have examined several alternative models, using in-sample and out-of-sample measures to validate the model selection.

For multivariate severity, copula modeling with gamma regression marginals provided little benefit, either on an in-sample or out-of-sample basis. In contrast, recent work in Frees and Valdez (2008) shows that copula models of multivariate severity (in auto) can be an effective modeling tool — they considered three types of automobile claims. We conjecture that the large dimension (number of perils is nine) of our severity response vector contributes to our inability to capture severity dependencies.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

For multivariate frequency, we surveyed a number of models that could be used. For our data, the dependence ratio approach was most effective. We used a special case of this framework that features an association structure similar to a correlation matrix. The dependence ratio model displayed statistically significant association parameters and had desirable out-of-sample performance.

Acknowledgement

We thank colleagues at the Innovative Analytics unit of the Insurance Services Office and an anonymous review for comments that help to improve the paper.

References

Angers, J.-F., Desjardins, D., Dionne, G. and Guertin, F. (2006) Vehicle and fleet random effects in a model of insurance rating for fleets of vehicles. *Astin Bulletin* **36(1)**, 25-77.

Boucher, J.-P. and Denuit, M. (2006) Fixed versus random effects in Poisson regression models for claim counts. A case study with motor insurance. *Astin Bulletin* **36(1)**, 285-301.

Boucher, J.-P. and 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.

Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. and Nesbitt, C.J. (1997) *Actuarial Mathematics*. Society of Actuaries, Schaumburg, IL.

Carey, V., Zeger, S.L. and Diggle, P. (1993) Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80(3)**, 517-526.

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002) *Analysis of Longitudinal Data, Second Edition*. Oxford University Press.

Ekholm, A., Smith, P.W.F. and McDonald, J.W. (1995) Marginal regression analysis of a multivariate binary response. *Biometrika* **82(4)**, 847-854.

Ekholm, A., McDonald, J.W. and Smith, P.W.F. (2000) Association models for a multivariate binary response. *Biometrics* **56**, 712-718.

Frees, E.W. and Wang, P. (2005) Credibility using copulas. *North American Actuarial Journal* **9(2)**, 31-48.

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

Frees, E.W., Shi, P. and Valdez, E.A. (2009) Actuarial applications of a hierarchical insurance claims model. *Astin Bulletin* **39(1)**, 165-197.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Liang, K.-Y. and Zeger, S.L. (1992) Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society B* **54(1)**, 3-40.

Lo, C.H., Fung, W.K. and Zhu, Z.Y. (2007) Structural parameter estimation using generalized estimating equations for regression credibility models. *Astin Bulletin* **37(2)**, 323- 343.

Mahmoudvand, R. and Hassani, H. (2009) Generalized bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *Astin Bulletin* **39(1)**, 307 -315.

Modlin, C. (2005) *Homeowners' modeling*. Presentation at the 2005 Casualty Actuarial Society Seminar on Predictive Modeling, available at http://www.casact.org/education/specsem/f2005/handouts/modlin.pdf.

Zhao, L.P. and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642-648.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

APPENDICES

A. Multivariate Severity Likelihoods

The copula allows for a fully parametric specification of the probability model. We exploit this specification by using maximum likelihood estimation. We assume independence among policies and use the copula to model dependencies among perils.

The difficulty is in the practical evaluation. There are between 10 and 20 regression covariates for each peril for over 100 regression parameters. There are an additional 9 scale parameters plus 36 correlation parameters. The large number of parameters can be handled under the independence model. Under independence, we have that $\text{cop}(.) \equiv 1$ and so the copula portion of the log-likelihood is zero. Because the first part of the log-likelihood is additive and with the assumption of no common parameters, the log-likelihood can be decomposed into $c = 9$ separate problems.

To get a better handle on the computational difficulties under dependence, let us examine the second expression on the right-hand side of equation (7) in the normal copula case. Using equation (6), we have

$$\begin{aligned}\ln \text{cop}_N\left(F_{i1}, \ldots, F_{ic}\right) &= \ln \phi_n\left(\Phi^{-1}(F_{i1}), \ldots, \Phi^{-1}(F_{ic})\right) - \sum_{j=1}^{c} \ln \phi\left(\Phi^{-1}(F_{ij})\right) \\ &= -\frac{c}{2}\ln(2\pi) - \frac{1}{2}\ln\det \boldsymbol{\Sigma} - \frac{1}{2}\boldsymbol{v}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{v}_i - \sum_{j=1}^{c} \ln \phi(v_{ij})\end{aligned}$$

where $v_{ij} = \Phi^{-1}(F_{ij})$ and $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{ic})'$.

As described before, it is rare to encounter a policy with claims from all perils. When there is a claim from a single peril, say the jth, the contribution to the log-likelihood is $l_i = \ln f(y_{ij}, \theta_{ij}, scale_j)$. This is because the copula reduces to a uniform distribution over $[0, 1]$ that has logarithmic density $\ln \text{cop}(\cdot) = 0$.

When there is a claim from two perils, say the first and second, the contribution to the log-likelihood is $l_i = \ln f(y_{i1}, \theta_{i1}, scale_1) + \ln f(y_{i2}, \theta_{i2}, scale_2) + \ln \text{cop}_N(F_{i1}, F_{i2})$, where

$$\begin{aligned}\ln \text{cop}_N\left(F_{i1}, F_{i2}\right) &= -\ln(2\pi) - \frac{1}{2}\ln\det\begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix} \\ &\quad - \frac{1}{2}\begin{pmatrix} v_{i1} & v_{i2} \end{pmatrix}\begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}^{-1}\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} - \sum_{j=1}^{2} \ln \phi(v_{ij}).\end{aligned}$$

Note that this likelihood expression involves only coefficients from the first two perils.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

Thus, we investigate the following algorithm:

1. Determine initial estimates of regression and scale coefficients assuming independence. Call these estimates $\widehat{\boldsymbol{\beta}}_j$ and $\widehat{scale_j}$, for $j = 1, \ldots, c$.
2. Assume that the regression and scale parameters are fixed. Maximize the likelihood over correlation parameters. Call these estimates $\widehat{\boldsymbol{\Sigma}}$.
3. Update the parameter estimates for the jth peril, $j = 1, \ldots, c$.
   - Assume that the correlation parameters ($\widehat{\boldsymbol{\Sigma}}$) are fixed.
   - Assume that the regression and scale parameters from other perils ($\widehat{\boldsymbol{\beta}}_k$ and $\widehat{scale_k}$, for $k = 1, \ldots, c$, $k \neq j$) are fixed.
   - Find the regression and scale parameters to maximize the likelihood.
4. Return to Step 2, until convergence.

We provide the following remarks. Step 1 provides estimates under the independence model. Step 2 involves only data from policies with two or more claims. For our data, this is only 3.9% of claims. For Step 3, each maximization step involves only claims from that peril.

B. Multivariate Binary Regression Models

This appendix reviews key features of multivariate binary regression models. We rely on Diggle et al. (2002), as well as Ekholm et al. (1995) and Liang and Zeger (1992).

B.1. Log-Linear Model

The vector of dependent variables is $\mathbf{r} = (r_1, \ldots, r_c)'$; there are $d = 2^c$ possible responses. One way of organizing the d possible outcomes is through the sufficient statistic

$$s(\mathbf{r}) = \left(r_1, \ldots, r_c,\ r_1 r_2, \ldots, r_{c-1} r_c, \ldots, r_1 r_2 \cdots r_c\right)'$$

consisting of all singletons, possible product pairs, and so on up to a product of all c variables. For example, if $c = 3$, then the $d = 8$ outcomes consist of

$$s(\mathbf{r}) = \left(r_1, r_2, r_3,\ r_1 r_2, r_1 r_3, r_2 r_3,\ r_1 r_2 r_3\right)'.$$

With a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_c, \theta_{12}, \ldots, \theta_{c-1,c}, \ldots, \theta_{12\cdots c})'$, we may write the likelihood as

$$\begin{aligned} \Pr(\mathbf{r}) &= c(\boldsymbol{\theta}) \exp\left(s(\mathbf{r})'\boldsymbol{\theta}\right) \\ &= c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{c} \theta_j r_j + \sum_{i<j} \theta_{ij} r_i r_j + \sum_{i<j<k} \theta_{ijk} r_i r_j r_k + \ldots + \theta_{12\cdots c} r_1 r_2 \cdots r_c \right). \end{aligned} \tag{9}$$

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

Here, $c(\boldsymbol{\theta})$ is a scaling term, so that probabilities sum to one. Equation (9) provides the basis for the *log-linear model*, a widely use representation for multivariate binary data.

Although widely used, the log-linear model does not readily handle regression explanatory variables. Instead, the parameters are interpreted in terms of *conditional* odds and odds ratios. To see this, take $c = 3$ and use $r_3 = 0$. Then,

$$\Pr(r_1, r_2, 0) = c(\boldsymbol{\theta}) \exp(\theta_1 r_1 + \theta_2 r_2 + \theta_{12} r_1 r_2).$$

From this, the odds for r_1, conditional on r_2 and $r_3 = 0$, are

$$\frac{\Pr(r_1 = 1 \,|\, r_2, 0)}{\Pr(r_1 = 0 \,|\, r_2, 0)} = \frac{\Pr(r_1 = 1, r_2, 0)}{\Pr(r_1 = 0, r_2, 0)} = \exp(\theta_1 + \theta_{12} r_2). \tag{10}$$

We may interpret θ_1 to be the logarithmic odds for r_1, conditional on all the other responses equal to zero (including r_2). We can interpret θ_{12} to be the association between r_1 and r_2, conditional on the values of the other dependent variables.

To introduce regression explanatory variables $\mathbf{x}$, one can always make each parameter a function of $\mathbf{x}$. One limitation is that the parameter interpretation depends on the other responses. Another limitation is that it is difficult to relate the parameters to the independence model, our baseline.

Special Case – Quadratic Exponential Model

Assuming that coefficients associated with more than product pairs are zero, Zhao and Prentice (1990) introduced the quadratic exponential model

$$\Pr(\mathbf{r}) = c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{c} \theta_j r_j + \sum_{i<j} \theta_{ij} r_i r_j \right). \tag{11}$$

An advantage of this model is that coefficients can be interpreted in terms of "conditional log odds ratios."

As with equation (10), from equation (11), one can check that

$$\ln\left\{ \frac{\Pr(r_j = 1 \,|\, r_k, r_l = 0, l \neq j, k)}{\Pr(r_j = 0 \,|\, r_k, r_l = 0, l \neq j, k)} \right\} = \theta_j + \theta_{jk} r_k.$$

We may interpret θ_j to be the logarithmic odds for r_j, conditional on all the other responses equal to zero (including r_k). We can interpret θ_{jk} to be the association between r_j and r_k, conditional on the values of the other dependent variables.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

To introduce regression explanatory variables $\mathbf{x}$, one can always make each parameter a function of $\mathbf{x}$. One limitation is that the parameter interpretation depends on the number of responses c. This is particularly important in longitudinal data, where the number depends on each subject. This is less important in our set-up where the number of perils is fixed. As mentioned earlier, perhaps the main limitation is that it is difficult to relate the parameters to the independence model, our baseline.

B.2. Bahadur's Representation

For marginal regression models, we use the c means π_j for parameters. There are several ways to specify the remaining $2^c - c - 1$ parameters. This subsection briefly describes Bahadur's representation.

Here, second order moments are given in terms of correlations. Specifically, define the mean of each response to be $\pi_j = \mathrm{E}\, r_j$ and a standardized version as

$$r_j^* = \frac{r_j - \pi_j}{\sqrt{\pi_j(1-\pi_j)}}.$$

Parameters are given as $\rho_{jk} = \mathrm{Corr}(r_j, r_k) = \mathrm{E}\, r_j^* r_k^*$, $\rho_{ijk} = \mathrm{E}\, r_i^* r_j^* r_k^*$ and so on up to $\rho_{12\cdots c} = \mathrm{E}\, r_1^* r_2^* \cdots r_c^*$.

With this notation, Bahadur's representation is

$$\begin{aligned}\Pr(\mathbf{r}) = &\prod_{j=1}^{c} \pi_j^{r_j}(1-\pi_j)^{(1-r_j)} \\ &\times \left(1 + \sum_{i<j} \rho_{ij} r_j^* r_k^* + \sum_{i<j<k} \rho_{ijk} r_i^* r_j^* r_k^* + \cdots + \rho_{12\cdots c} r_1^* r_2^* \cdots r_c^*\right).\end{aligned} \tag{12}$$

The strength of Bahadur's representation is that one can see how the joint probability depends on the correlations and more complex interactions among dependent variables. The limitation is that correlations are constrained by marginal means in the binary model.

B.3. Alternating Logistic Regressions

This section discusses an "alternating logistic regression", or ALR, due to Carey et al. (1993). The ALR uses generalized estimating equations (GEE) to estimate mean and association parameters. The ALR algorithm is described that uses means and odds ratios only, and is silent on the role of the remaining parameters.

To describe the ALR procedure, we begin with the same mean parameters that one would use for a univariate binary regression model. Define $\pi_{ij} = \Pr(r_{ij} = 1)$ to be the mean that is modeled through a systematic component as $\mathrm{logit}(\pi_{ij}) =$

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

$\mathbf{x}_{ij}'\boldsymbol{\beta}_j$. Through this notation, we allow the explanatory variables ($\mathbf{x}$) and regression coefficients ($\boldsymbol{\beta}$) to depend on peril j. For association, instead of correlations we use the odds ratio. Specifically, define the odds ratio between r_{ij} and r_{ik} to be

$$\psi_{ijk} = \frac{\Pr(r_{ij}=1, r_{ik}=1)/\Pr(r_{ij}=0, r_{ik}=1)}{\Pr(r_{ij}=1, r_{ik}=0)/\Pr(r_{ij}=0, r_{ik}=0)} \tag{13}$$

This is one under independence. As with the means, we estimate parameters that will summarize this dependence and use the relation $\ln \psi_{ijk} = \mathbf{z}_{ijk}'\boldsymbol{\alpha}$. Here, $\mathbf{z}_{ijk}$ is a set of explanatory variables that could be used to model the association.

The ALR algorithm is based on two stages, one for the regression coefficients ($\beta's$) and one for the association parameters ($\alpha's$).

Specifically, the first stage is the usual GEE estimation procedure with fixed association parameters. To compute the variance, recall the well-known relation that if means and correlations are known, then one can compute variances and covariances. Similarly, if means and odds ratios are known, then one can compute variances and covariances of binary variables. To see this, for the variances, we have $\text{Var}\, r_{ij} = \pi_{ij} - \pi_{ij}^2$, so the mean parameter determines the variance. For covariances, we have $\text{Cov}(r_{ij}, r_{ik}) = \text{E}\, r_{ij}r_{ik} - \pi_{ij}\pi_{ik} = \Pr(r_{ij}=1, r_{ik}=1) - \pi_{ij}\pi_{ik}$. Thus, only the joint probability $\Pr(r_{ij}=1, r_{ik}=1)$ needs to be determined to compute the covariances. Using the relations, $\pi_{ij} = \Pr(r_{ij}=1, r_{ik}=1) + \Pr(r_{ij}=1, r_{ik}=0)$, similarly for π_{ik} and equation (13), we can solve for $\Pr(r_{ij}=1, r_{ik}=1)$ and hence compute the covariances. To summarize the variances and covariances, define $\mathbf{V}_i$ to be the variance-covariance matrix of $\mathbf{r}_i$.

With this notation, the stage one algorithm is to solve the estimating equation for $\boldsymbol{\beta}$

$$\mathbf{0} = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\pi}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} (\mathbf{r}_i - \boldsymbol{\pi}_i).$$

Stage two of the algorithm is for estimation of the association parameters, fixing the regression coefficients. To this end, we focus on the conditional mean $\zeta_{ijk} = \text{E}(r_{ij} | r_{ik} = y)$, $y = 0,1$ and define the residual $R_{ijk} = r_{ij} - \zeta_{ijk}$. The vector of residuals $\mathbf{R}_i$ has dimension $c(c-1)$. The calculation of the conditional mean uses the expression

$$\text{logit}(\zeta_{ijk}) = y \ln \psi_{ijk} + \ln \frac{\pi_{ij} - \Pr(r_{ij}=1, r_{ik}=1)}{1 - \pi_{ij} - \pi_{ik} + \Pr(r_{ij}=1, r_{ik}=1)} \tag{14}$$

that can easily be derived from straightforward calculations. In the algorithm, the second term on the right-hand side of equation (14) is taken to be an offset term (although it involves $\boldsymbol{\alpha}$ in the joint probability). Following

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

usual GEE practices, only the diagonal term is used in the variance component and so define $\mathbf{S}_i = diag(\zeta_{ijk}(1-\zeta_{ijk}))$. This is a $c(c-1)\times c(c-1)$ diagonal matrix.

With this notation, the stage two algorithm is to solve the estimating equation for $\boldsymbol{\alpha}$

$$\mathbf{0} = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\zeta}_i'}{\partial \boldsymbol{\alpha}} \mathbf{S}_i^{-1} \mathbf{R}_i.$$

One begins the recursion using estimates from the independence model as initial values. Then, one alternates between stage one and stage two until convergence.

C. Dependence Ratio Likelihood With only Bivariate Dependencies

The idea is to use a logit specification to introduce explanatory variables to estimate the marginal means π_j. For our data set, policies with three and more claims represent an extremely small fraction of the data. Thus, in our first parameterization we will estimate τ_{jk} for each pair of perils (j,k) but assume that higher order ratios, such as τ_{jkl} and τ_{jklm}, are equal to one. With this convention, from equation (8) we may write for the probability of zero claims for all perils:

$$\begin{aligned}\Pr(r_1 = 0, \ldots, r_c = 0) &= 1 - \sum_{j=1}^{c} \pi_j + \sum_{j<k} \pi_j \pi_k - \sum_{i<j<k} \pi_i \pi_j \pi_k + \ldots \\ &\quad + (-1)^c \pi_1 \cdots \pi_c + \sum_{j<k} \left(\pi_{jk} - \pi_j \pi_k\right) \\ &= \prod_{j=1}^{c} \left(1 - \pi_j\right) + \sum_{j<k} \left(\tau_{jk} - 1\right) \pi_j \pi_k\end{aligned}$$

Similarly, for the probability of a claim in the first peril and no other claims, we have

$$\begin{aligned}\Pr(r_1 = 1, r_2 = 0, \ldots, r_c = 0) &= \pi_1 - \sum_{j=2}^{c} \pi_1 \pi_j + \sum_{1<j<k} \pi_1 \pi_j \pi_k - \ldots \\ &\quad + (-1)^{c-1} \pi_1 \pi_2 \cdots \pi_c - \sum_{j=2}^{c} \left(\pi_{1j} - \pi_1 \pi_j\right) \\ &= \pi_1 \prod_{j=2}^{c} \left(1 - \pi_j\right) - \pi_1 \sum_{j=2}^{c} \left(\tau_{1j} - 1\right) \pi_j.\end{aligned}$$

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

For a claim in the first two perils and no other claims, we have

$$\begin{aligned}\Pr(r_1 = 1,\, r_2 = 1,\, r_3 = 0, \ldots, r_c = 0) \\ &= \pi_{12} - \sum_{j=3}^{c} \pi_1\pi_2\pi_j + \sum_{2<j<k} \pi_1\pi_2\pi_j\pi_k - \ldots + (-1)^c \pi_1\pi_2\cdots\pi_c \\ &= \pi_1\pi_2 \prod_{j=3}^{c} \left(1-\pi_j\right) - \pi_1\pi_2(\tau_{12}-1).\end{aligned}$$

For a claim in the first three perils and no other claims, we have

$$\Pr(r_1 = 1,\, r_2 = 1,\, r_3 = 1,\, r_4 = 0, \ldots, r_c = 0) = \pi_1\pi_2\pi_3 \prod_{j=4}^{c} \left(1-\pi_j\right)$$

and similarly for a claim in the first four perils.

Maximization Algorithm

The procedure is similar to the copula model. Specifically, we have implemented the following:

1. Determine initial estimates of regression coefficients assuming independence. Call these estimates $\widehat{\boldsymbol{\beta}}_j$, for $j = 1, \ldots, c$.
2. Assume that the regression are fixed. Maximize the likelihood over dependence ratio parameters. Call these estimates $\hat{\tau}_{jk}$.
3. Update the parameter estimates for the jth peril, $j = 1, \ldots, c$.
   - Assume that the dependence ratio parameters ($\hat{\tau}_{jk}$) are fixed.
   - Assume that the regression parameters from other perils ($\widehat{\boldsymbol{\beta}}_k$, for $k = 1, \ldots, c$, $k \neq j$) are fixed.
   - Find the regression parameters to maximize the likelihood.
4. Return to Step 2, until convergence.

We provide the following remarks. Unlike the copula estimation scheme, Step 2 involves data from all the policies. Thus, we have spent considerable amount of time making the computing efficient. For Step 3, each maximization step involves data from all the policies (again, unlike the copula estimation scheme).

D. Out-of-Sample Validation Measures

In insurance claims modeling, standard out-of-sample validation measures are not the most informative due to the high proportions of zeros (corresponding to no claim) and the skewed fat tailed distribution of the positive values.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

TABLE 10

OUT-OF-SAMPLE DISTRIBUTIONS

Variable	Mean	Minimum	Percentiles 1st	5th	25th	50th	75th	95th	99th	Maximum
Score1	294.93	33.05	97.14	126.61	185.07	244.99	333.68	606.04	1,106.18	2,402.49
Claims	332.89	0	0	0	0	0	0	660.000	5,916.36	350,000.00

To underscore this point, Table 10 compares a score computed from an in-sample model to held-out insurance claims. Specifically, as described in Section 2, we use a held-out, or "validation" subsample of 359,454 records, whose claims we wish to predict. The total Claims in Table 10 consist of 93.5% zeros, with positive claims being skewed to the right and fat tailed. A basic score, Score1, is the predicted amount assuming no dependencies among claims frequencies nor severities. Using notation, we write this as

$$\widehat{\text{Score1}}_i = \sum_{j=1}^{c} \widehat{\text{Prob}}_{i,j} \times \widehat{\text{Fit}}_{i,j} = \sum_{j=1}^{c} \frac{\exp(\mathbf{x}_{1,ij}' \mathbf{b}_{1,j})}{1+\exp(\mathbf{x}_{1,ij}' \mathbf{b}_{1,j})} \times \exp(\mathbf{x}_{2,ij}' \mathbf{b}_{2,j}). \quad (15)$$

Here, $\widehat{\text{Prob}}_{i,j}$ is the predicted probability using logistic regression model parameter estimates, $\mathbf{b}_{1,j}$, and frequency covariates $\mathbf{x}_{1,ij}$, for the jth peril. Further, $\widehat{\text{Fit}}_{i,j}$ is the predicted amount based on a logarithmic link using gamma regression model parameter estimates, $\mathbf{b}_{2,j}$, and severity covariates $\mathbf{x}_{2,ij}$, for the jth peril.

One standard measure of out-of-sample performance is the mean absolute error (MAE). Using $y_{(i)}$ to denote the ith order statistic, we can express this as

$$MAE = \frac{1}{359{,}454} \sum_{i=1}^{359{,}454} \left|y_{(i)} - \hat{y}_{(i)}\right| = \frac{1}{359{,}454} \left(\sum_{i=1}^{\xi} \hat{y}_{(i)} + \sum_{i=\xi+1}^{359{,}454} \left|y_{(i)} - \hat{y}_{(i)}\right| \right),$$

where ξ is the index corresponding to the first non-zero percentile (for our dataset, $\xi = \xi_{.935} = 336294$, corresponding to 93.5% of the held-out 359,454 records that do not have claims). From this expression, we can see that lowering predicted values $\hat{y}_{(i)}$ will reduce MAE because this would reduce the first sum dramatically and not increase the second sum substantially. Thus, the predicted value that produces the best MAE is zero! Hence, MAE is not an appropriate criterion for choosing a pricing model.

Similar criticisms can be made of related standard measures, including root mean square error, mean absolute percentage error and so forth. Thus, we now introduce a measure motivated by the economics of insurance.

Assume that Score1 introduced in equation (15) is our "base score" that an insurer is currently using for pricing. Under consideration is an alternative

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

Score2 that is superior in the sense that $\mathrm{E}\, y_i \approx \mathrm{Score2}_i$ for the ith member of a held-out sample. An insurer has available these two scores and may not wish to underwrite or retain policies where the price is much lower than expected claims. To this end, define R_i to be the relative price, $R_i = \mathrm{Score2}_i \,/\, \mathrm{Score1}_i$. To help the insurer identify policies with a score (Score2) that is low relative to price (Score1), sort the policies by R_i. If the insurer retains policies with the k lowest relative prices, then the profit earned is $\mathrm{Profit}_k = \sum_{i=1}^{k} (\mathrm{Score1}_i - y_i)$. This has approximate expectation

$$\mathrm{E\ Profit}_k \approx \sum_{i=1}^{k} \left(\mathrm{Score1}_i - \mathrm{Score2}_i\right) = \sum_{i=1}^{k} \mathrm{Score1}_i (1 - R_i).$$

Thus, an insurer's optimal strategy is to continue retaining policies as long as the relative price is less than one. Of course, the optimality of this strategy depends heavily on the assumption $\mathrm{E}\, y_i \approx \mathrm{Score2}_i$.

Moreover, prices often contain components to meet costs that are not part of the claims payments, such as expenses for underwriting, marketing and so on. Thus, when examining a portfolio of risks, it may be more robust to compare the percentage of premium to the percentage of claims retained. Thus, we also examine

$$\mathrm{PercentScore1}_k - \mathrm{PercentClaims}_k = 100 \times \left(\frac{\sum_{i=1}^{k} \mathrm{Score1}_i}{\sum_{i=1}^{n} \mathrm{Score1}_i} - \frac{\sum_{i=1}^{k} y_i}{\sum_{i=1}^{n} y_i} \right) \quad (16)$$

This is another measure of the profitability of a block of business.

To see how this works for the homeowners data, Table 11 compares the out-of-sample performance between Score1 and another score simply labeled as "Score2" for now. There are 359,454 records in this held-out dataset that was broken into ten subgroups, each of approximate size 35,945, by rank of relative price. For example, Table 11 shows that the first subgroup had an average relative price of 1.014 with average claims of $\$387.77$. The profit for this group is $\mathrm{Profit}_1 = 35945 \times (408.98 - 387.77) = 762{,}393$. From another perspective, this first group generated 13.87% of premiums and was responsible for only 11.65% of claims. Thus, this is clearly a desirable group of policies to retain.

These profit figures depend upon the number of policies retained by the insurer. Although some insurers might wish to retain only those policies with a relative price below one, others will wish to have more aggressive or conservative marketing strategies. To assess the out-of-sample performance of different scoring models, we will use an average of the profit measure defined in equation (16), defined as

$$\mathit{Gini} = \frac{1}{2n} \sum_{k=1}^{n} \left(\mathrm{PercentScore1}_k - \mathrm{PercentClaims}_k\right). \quad (17)$$

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press

TABLE 11

AVERAGE OUT-OF-SAMPLE CLAIMS AND SCORES

Group	1	2	3	4	5	6	7	8	9	10	Total
Total Claims	387.77	326.09	255.82	335.51	292.39	323.76	337.52	277.44	374.51	418.11	332.89
Score1	408.98	274.36	264.06	261.18	262.91	262.85	266.80	271.28	290.12	386.79	294.93
Score2	414.37	279.86	269.94	267.42	269.56	269.85	274.26	279.24	299.13	400.26	302.39
Relative Price	1.014	1.020	1.022	1.024	1.025	1.027	1.028	1.029	1.031	1.034	1.025
Percent Claims	11.65	21.44	29.13	39.21	47.99	57.72	67.86	76.19	87.44	100.00	
Percent Score 1	13.87	23.17	32.12	40.98	49.89	58.80	67.85	77.05	86.89	100.00	
Percent Score 2	13.70	22.96	31.89	40.73	49.64	58.57	67.64	76.87	86.76	100.00	

Notes: Groups are based on deciles ordered by relative prices.
Average claims, scores and relative prices are reported for each group.

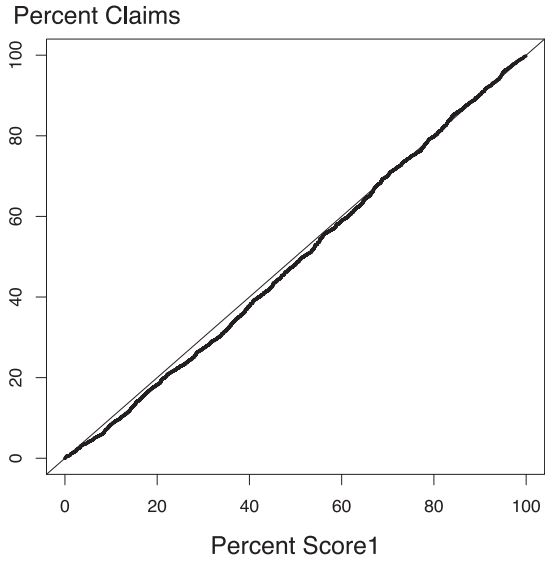


FIGURE 1: Percentage of total claims versus independence model scores. The Gini index is 2.32%

It is an "average" in the sense that we are taking a mean over all decision-making strategies, that is, each strategy retaining the k policies with lowest relative prices.

By comparing percent premium to percent claims in Table 11, we can see how profit evolves as the insurer retains larger portfolios. Figure 1 displays a plot of percentage claims versus percentage premiums. The Gini index can be interpreted as the area between the two lines (hence the reason for dividing by the 2 in equation 17). We call it a "Gini" index because it is reminiscent of the measure of income inequality. Note, however, that in our definition we are ordering by relative prices and so this is not a standard Gini measure.

https://doi.org/10.2143/AST.40.2.2061134 Published online by Cambridge University Press