Annals of Actuarial Science, Vol. 7, part 2, pp. 258–287. © Institute and Faculty of Actuaries 2013
doi:10.1017/S1748499512000346
First published online 2 April 2013

Actuarial Applications of Multivariate Two-Part Regression Models

Edward W. (Jed) Frees*
University of Wisconsin - Madison

Xiaoli Jin
University of Wisconsin - Madison

Xiao (Joyce) Lin
University of Wisconsin - Madison

Abstract

This paper synthesizes and extends the literature on multivariate two-part regression modelling, with an emphasis on actuarial applications. To illustrate the modelling, we use data from the US Medical Expenditure Panel Survey to explore expenditures that come in two parts. In the first part, zero expenditures correspond to no payments for health care services during a year. For the second part, a positive expenditure corresponds to the payment amount, a measure of utilization. Expenditures are multivariate, the five components being (i) office-based, (ii) hospital outpatient, (iii) emergency room, (iv) hospital inpatient, and (v) home health expenditures. Not surprisingly, there is a high degree of association among expenditure types and so we utilize models that account for these associations. These models include multivariate binary regressions for the payment type and generalized linear models with Gaussian copulas for payment amounts.

As anticipated, the strong associations among expenditure types allow us to establish significant model differences on an in-sample basis. Despite these strong associations, we find that commonly used statistical measures perform similarly on a held-out validation sample. In contrast, out-of-sample risk measures used by actuaries reveal differences in the association among expenditure types.

Keywords

Copula; healthcare expenditures

1 Introduction

This paper synthesizes and extends the literature on multivariate two-part regression modelling with an emphasis on actuarial applications.

1.1 What is a Multivariate Two-Part Regression Model?

Multiple linear regression is a widely used statistical methodology; this can be seen from a cursory examination of the scientific literature (through the many applications of regression), popular press (e.g. textbooks) or scientific computing space (e.g. statistical packages). *Regression* is a statistical

*Correspondence to: Edward W. (Jed) Frees. E-mail: jfrees@bus.wisc.edu technique that serves to explain the distribution of an outcome of interest in terms of other variables, often called "explanatory" or "predictor" variables. Although it is possible to consider only a single explanatory variable, applications that involve *multiple* (more than one) variables are more prevalent. Typically, one uses *linear* combinations of explanatory variables as parameters of the distribution, hence giving rise to the phrase "multiple linear regression."

Many actuarial data sets come in "two parts:"

- one part for the frequency, indicating whether or not a claim has occurred or, more generally, the number of claims and
- one part for the severity, indicating the amount of a claim.

In predicting or estimating claims distributions, we often associate the cost of claims with two components: the event of the claim and its amount, if the claim occurs. This is the traditional way of decomposing "two-part" data, where one can think of a zero as arising from a policy without a claim (Frees, 2010, Chapter 16). Because of this decomposition, two-part models are also known as frequency-severity models.

Healthcare data also often feature a large proportion of zeros. Zero values can represent an individual's lack of healthcare utilization, no expenditure, or non-participation in a program. In healthcare, Mullahy (1998) cites some prominent areas of potential applicability:

- outcomes research – amount of health care utilization or expenditures
- demand for health care – amount of health care sought, such as number of physician visits, and
- substance abuse – amount consumed of tobacco, alcohol and illicit drugs.

This paper emphasizes applications where there is more than a single outcome of interest. For example, there could be several components of an automobile claim corresponding to (i) damage to one's own vehicle, (ii) another party's vehicle or other property, or (iii) injury to an insured driver. Of course, simply because an analyst examines several outcomes of interest, the utility of using other variables to explain these outcomes does not disappear – regression techniques are important when examining multivariate outcomes. Although not universal, we use the widely adopted descriptor *multivariate regression* to denote the situation when multiple explanatory variables are used to explain several outcomes of interest.

1.2 Actuarial Applications

To motivate multivariate regression models, we provide a few examples that have appeared recently in the actuarial literature.

Example 1. Health Care Expenditures. Frees *et al.* (2011a) examined medical payments of 9,472 participants from the Medical Expenditure Panel Survey. Over thirty participant characteristics were used to model payment patterns including demography (age, sex, ethnicity), socio-economic (education, marital status, income), health status, employment (status, industry), and availability of health insurance. Payments consisted of two types: inpatient (e.g. hospital) and outpatient expenditures. This paper extends this study and is described in more detail beginning in Section 3.

Example 2. Multi-Peril Homeowners Insurance. Many actuaries interested in pricing homeowners insurance are now decomposing the risk by *peril*, or cause of loss. Homeowners is typically sold as

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

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. However, it 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. Frees *et al.* (2010) examined a dataset containing a sample of over 400,000 policies issued by several major insurance companies in the US. This paper established strong dependencies among perils.

Example 3. Life Insurance Ownership. In Frees & Sun (2010), a bivariate two-part regression model was used to study the demand for life insurance. The study considered 2,150 households from the Survey of Consumer Finances. The paper examined the joint ownership of the amount of term life insurance and the net amount of risk (coverage amount less cash value) for whole life insurance. From the survey, many variables were available to help explain life insurance ownership, including ownership of assets, debt, income, bequests and inheritance, age, education and financial vulnerability. This study revealed that the ownership decision involves substitution, i.e., the two products tend to serve as replacement for one another. Further, for households owning both types of insurance, amounts are positively related. That is, term and whole life insurance are substitutes in the frequency yet complements in the severity.

1.3 Why Multivariate Outcomes?

Analysts and managers gain useful insights by studying insurance risks individually. These insights can be augmented through the study of *joint* behaviour of insurance risks, i.e., a multivariate approach.

1. For some products, insurers must track payments separately by component to meet contractual obligations.

    - In automobile coverage, deductibles and limits depend on the coverage type, e.g. bodily injury, damage to one's own vehicle or to another party.
    - In medical insurance, there are often co-payments in more routine expenditures such as prescription drugs.
    - In personal lines umbrella insurance, there are separate limits for homeowners and auto coverages, as well overall limits for losses from all sources.

    When contract specifications differ by type, it is natural for insurers to decompose a total loss into components; jointly studying several components of loss simultaneously can be accomplished through a multivariate approach.
2. For other products, there may be no contractual reasons to decompose an obligation by components. Nonetheless, insurers often do so to help them better understand the overall risk. Homeowners insurance in Example 2 illustrates this situation. It is intuitively appealing to decompose an overall loss into perils because some variables 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.
3. Multivariate models need not be restricted to only insurance losses.

    - In Example 3 we presented a study of term and whole life insurance ownership. Potential competition among products suggests dependencies in demand. Although demand is not

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

an "insurance loss," actuaries can use the same models to understand demand, an important issue for any business.

- For some additional examples, insurers may also be interested in the type and amount of debt or assets that individuals own.

When considering multivariate outcomes, it is important to understand how each component depends on the others. Intuitively, we may think of each multivariate outcome as a vector of risks. When actuaries consider multiple asset categories, another type of risk, they are well trained to appreciate their inter-dependencies. For example, actuaries almost intuitively consider dependencies among risky stocks, risky bonds and (relatively) risk-free government assets.

The purpose of this paper is to analyze inter-dependent risks that insurers and other financial service firms routinely deal with. Unlike the linear, Gaussian model that is commonly used on the assets side, we also wish to consider modelling losses. Thus, we entertain distributions that have large proportions at zero (corresponding to no loss) and when positive, the distribution tends to have a longer tail than the normal (Gaussian) distribution.

1.4 Modelling Strategy

The multivariate modelling strategy is to build upon well-known models of univariate outcomes to allow for multivariate dependencies. To this end, we review univariate models in Section 2.1 where we compare the two-part models to Tweedie (pure premium) and tobit (censored regression) approaches. Then, Section 2.2 introduces the multivariate two-part model. For frequencies, multivariate binary regressions are considered in Section 2.2.1. For severities, we join the component amount distributions using a copula in Section 2.2.2.

Section 3 will show how to use these models in the context of a detailed case study on healthcare expenditures. After introducing the data, Section 4 describes model estimation results and Section 5 summarizes prediction results. Section 6 provides concluding remarks.

2 Models and Inference

It is helpful to use mathematical notation to introduce technical model ideas compactly. To this end, we use y to denote an outcome of interest, typically an insurance loss. The vector $\mathbf{x}$ represents a set of explanatory variables that can be used to explain the distribution of y. Lower case Greek letters denote model parameters.

2.1 Univariate Models

We begin with a representation of a single outcome, y, which we call a univariate model.

2.1.1 Generalized Linear Model (GLM)

Many insurance and actuarial applications employ the generalized linear model (GLM) to represent the distribution of y. In this model, the outcome distribution is a member of the linear exponential family with location parameter μ that depends on parameters β, explanatory variables $\mathbf{x}$, and a scale parameter ϕ. Important special cases of the linear exponential family allows the analyst to handle:

- linear regression, with approximately normally distributed outcomes,
- logistic regression, with binary outcomes,

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

- Poisson regression, with count outcomes, and
- gamma regression, with heavy tail outcomes.

In actuarial applications, Tweedie GLMs are also prominent. The Tweedie distribution is a Poisson sum of gamma random variables. Thus, it has a mass at zero as well as a continuous component. It is used to model "pure premiums," where the zeros correspond to no claims and the positive part is used for the claim amount.

For a more detailed introduction to GLMs with an actuarial emphasis, see Haberman & Renshaw (1996), Frees (2010), or de Jong & Heller (2008).

2.1.2 Two-Part Model

In a two-part model, one explicitly decomposes the outcome of interest into the occurrence event and a positive part. That is, let $y = \begin{cases} 0 & r = 0 \\ y^* & r = 1 \end{cases}$, where r indicates if an outcome occurs and, conditional on an outcome occurrence ($r = 1$), y^* is the amount.

To model the distribution of the outcome vector (r, y^*), it is convenient to first model the occurrence event and then model the amount. Symbolically, we use $\mathrm{f}(r,\ y^*) = \mathrm{f}_1(r) \times \mathrm{f}_2(y^*|r)$. From this, parameter estimation proceeds in two steps.

- **Part 1.** The random variable r is binary and so its distribution can be estimated using logistic regression.
- **Part 2.** Given $r = 1$ (the occurrence of an outcome), the amount y^* can be represented using a GLM. For example, in personal lines automobile insurance, it is common to use a gamma regression model.

When the parameters from the two parts are functionally independent, we can optimize each part in isolation of one another and thus, treat the likelihood process in "two parts".

Tweedie GLMs. Some two-part data $(y, \mathbf{x})$ can also be analyzed using a Tweedie GLM model. For example, if the outcome is an insurance claim, we interpret $y = r = 0$ to be the event of no claim and so can use a Tweedie distribution. Compared to the two-part model, a strength of the Tweedie approach is that both parts are estimated simultaneously; this means fewer parameters, making the variable selection process simpler. Further, simpler models are preferred for predictive purposes, other things being equal. In contrast, the two-part model employs a set of parameters for the frequency and another set of parameters for the amount. In many applications, variables that affect the frequency may differ from those that affect the amount. For a healthcare example, it is an individual's decision to seek treatment and those characteristics affect the frequency whereas the physician mainly decides the intensity of expenditures, making the individual's characteristics less relevant. The two-part model cleanly captures this joint decision-making process by splitting the likelihood into two parts, one for each decision maker.

Tobits. Another commonly used device for incorporating a mass at zero into an otherwise continuous distribution is through a censored regression model. Here, the dependent variable $y = \max(0,\ y^*)$ is limited, or censored, by zero. For example, we might think of y^* as a continuous measure of a person's unobserved tendency to incur healthcare expenditures; the observed quantity, y, is bounded below by zero. Censored regression models have strengths and limitations that are comparable to the Tweedie

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

GLM model. They typically have fewer parameters than the two-part model at the price of limited flexibility. Moreover, both the Tweedie GLM and the censored regression model require strong distributional assumptions on the amount component. Because the Tweedie distribution is defined as the Poisson sum of gamma random variables, there is an implicit parametric distribution assumption on the positive (amount) component. The censored regression model typically assumes normality, resulting in the so-called "tobit" model. In contrast, the two-part model retains flexibility in the specification of the amount distribution.

2.2 Multivariate Two-Part Model

To define a multivariate two-part model, we use a multivariate outcome of interest **y** where each element of the vector consists of two parts. Thus, we observe

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \text{ as well as } \mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_p \end{pmatrix}$$

and potentially observe

$$\mathbf{y}^* = \begin{pmatrix} y_1^* \\ \vdots \\ y_p^* \end{pmatrix}.$$

We think of **r** as the frequency vector and $\boldsymbol{y}^*$ as the amount, or severity, vector. As with the univariate model, we may decompose the overall likelihood into frequency and severity components. Specifically, we use $\mathrm{f}(\mathbf{r}, \mathbf{y}^*) = \mathrm{f}_1(\mathbf{r}) \times \mathrm{f}_2(\mathbf{y}^*|\mathbf{r})$.

We now discuss each component.

2.2.1 Multivariate Binary Models

Fortunately, there are many good approaches to modelling multivariate binary frequencies in the statistics literature. See, for example, Diggle *et al.* (2002). Unfortunately, none of these approaches appears to be uniformly superior to the others. Thus, this section provides an overview, emphasizing approaches that are most promising for actuarial applications.

Multinomial Logistic Regressions. We first note that there are only a finite number of possibilities associated with the binary vector **r**; specifically, there are 2^p possible events. Thus, a straightforward way of fitting the distribution of **r** is to treat it as a categorical outcome and to use multinomial logistic regression, e.g. Glonek & MuCullagh (1995). This approach allows the analyst to specify a set of explanatory variables for each event. This flexibility means that one can get very good fits. However, analysts prefer models that have explanatory variables and coefficients associated with each outcome variable. These "marginal regression" models are easier to interpret and so are our focus here.

Marginal Binary Regressions. Specifically, suppose we wish to predict the probability that the *i*th individual has the first outcome, i.e., $\pi_{i1} = \Pr(r_{i1} = 1)$. With a marginal logistic regression model, we would employ

$$\pi_{i1} = \frac{\exp(\mathbf{x}_{i1}'\beta_1)}{1 + \exp(\mathbf{x}_{i1}'\beta_1)}, \tag{1}$$

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

resulting in $\text{logit}(\pi_{i1}) = \mathbf{x}'_{i1}\beta_1$. Through this notation, we allow the explanatory variables ($\mathbf{x}$) and regression coefficients (β) to depend on the type of outcome.

Dependence Ratios and Odds Ratios. For association among binary variables, instead of correlations it is customary to use (1) dependence ratios and (2) odds ratios. Intuitively, correlations, such as those due to Pearson and Spearman, are less useful because they rely on higher order moments that do not provide a great deal of information about a binary distribution. We begin with the dependence ratio and drop the "*i*" subscript for the moment. The dependence ratio,

$$\tau_{12} = \frac{\Pr(r_1 = 1, r_2 = 1)}{\Pr(r_1 = 1)\Pr(r_2 = 1)}, \tag{2}$$

is 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.

For the odds ratio approach, first recall that the odds of the event $\{r_1 = 1\}$ is

$$odds(r_1) = \frac{\pi_1}{1 - \pi_1} = \frac{\Pr(r_1 = 1)}{\Pr(r_1 = 0)}.$$

With this, the odds ratio between r_2 and r_1 is

$$\begin{aligned}
\text{OR}(r_2, r_1) &= \frac{odds(r_2|r_1 = 1)}{odds(r_2|r_1 = 0)} \\
&= \frac{\Pr(r_2 = 1|r_1 = 1)/(1 - \Pr(r_2 = 1|r_1 = 1))}{\Pr(r_2 = 1|r_1 = 0)/(1 - \Pr(r_2 = 1|r_1 = 0))} \\
&= \frac{\Pr(r_2 = 1, r_1 = 1)/(\Pr(r_1 = 1) - \Pr(r_2 = 1, r_1 = 1))}{\Pr(r_2 = 1, r_1 = 0)/(\Pr(r_1 = 0) - \Pr(r_2 = 1, r_1 = 0))} \\
&= \frac{\Pr(r_2 = 1, r_1 = 1)\Pr(r_2 = 0, r_1 = 0)}{\Pr(r_2 = 1, r_1 = 0)\Pr(r_2 = 0, r_1 = 1)}.
\end{aligned} \tag{3}$$

As with the dependence ratio, the odds ratio is one under independence. Values greater than one indicate positive dependence and values less than one indicate negative dependence.

Example. Health Care Expenditures. Beginning in Section 3, we describe a dataset of $n = 18{,}908$ subjects who have up to $p = 5$ types of expenditures. Table 1 provides the joint counts as well as total event counts. This table shows that office-based events are the most prevalent; specifically, 10,528 subjects, or 55.7%, had an office-based healthcare expenditure during the year. Equivalently, the odds of an office-based expenditure are 1.256 to one ($= 0.557/(1 - 0.557)$).

Tables 2 and 3 give the corresponding dependence ratios and odds ratios, respectively. To illustrate the calculations, the dependence ratio between office-based and hospital outpatient may be determined as $\frac{1982/18908}{0.557 \times 0.114} = 1.645$. From equation (3), the odds ratio is.

$$\frac{(1982/18908) \times (18908 - 2164 - 10528 + 1982)/18908}{(2164 - 1982)/18908 \times (10528 - 1928)/18908} = 10.447$$

For both association measures, we see that there is positive dependence among all types. The strongest association appears to be between inpatient and home health (dependence ratio is

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 1. Joint Counts Among Types of Events and Total Event Counts

	OB	OP	ER	IP	HH
Office-Based (OB)	–	1,982	1,793	1,212	220
Hospital Outpatient (OP)	1,982	–	511	383	86
Emergency Room (ER)	1,793	511	–	626	74
Inpatient (IP)	1,212	383	626	–	111
Home Health (HH)	220	86	74	111	–
Total Count for the Event	10,528	2,164	2,274	1,339	235
Percent of an Event	55.7	11.4	12.0	7.1	1.2
Odds of an Event	1.256	0.129	0.137	0.076	0.013

Table 2. Dependence Ratios Among Types of Events

	OB	OP	ER	IP	HH
Office-Based (OB)	–	1.645	1.416	1.626	1.681
Hospital Outpatient (OP)	1.645	–	1.963	2.499	3.198
Emergency Room (ER)	1.416	1.963	–	3.887	2.618
Inpatient (IP)	1.626	2.499	3.887	–	6.670
Home Health (HH)	1.681	3.198	2.618	6.670	–

Table 3. Odds Ratios Among Types of Events

	OB	OP	ER	IP	HH
Office-Based (OB)	–	10.447	3.371	8.454	11.902
Hospital Outpatient (OP)	10.447	–	2.627	3.551	4.609
Emergency Room (ER)	3.371	2.627	–	8.482	3.442
Inpatient (IP)	8.454	3.551	8.482	–	12.717
Home Health (HH)	11.902	4.609	3.442	12.717	–

6.670 and odds ratio is 12.717). The weakest association for the dependence ratio is between office-based and emergency room (1.416) and for odds ratio is between hospital outpatient and emergency room (2.627). This example emphasizes that the two statistics measure different aspects of the association. Also note that Tables 2 and 3 do not account for potential explanatory variables, or risk factors; we shall control for these risk factors in subsequent regression modelling.

Both approaches have been used extensively in the statistical literature. The dependence ratio approach was introduced by Ekholm *et al.* (1995) and was used in actuarial applications in Examples 2 and 3. Sun (2011) develops applications in personal lines automobile and homeowners insurance. The odds ratio approach is discussed in Diggle *et al.* (2002), among others, and will be used in the case study in this paper.

Both approaches work well when the marginal probabilities are small. For the $p = 2$ case, one can determine the joint probability of two outcomes through knowledge of the two marginal distributions and the association parameter. For example, in the dependence ratio case, $\Pr(r_{i1} = 1, r_{i2} = 1) = \tau_{12} \times \Pr(r_{i1} = 1)\Pr(r_{i2} = 1)$. As when introducing correlation association

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

parameters in generalized least squares, it is customary to assume that the association parameters are constant over observations. These simplifying assumptions allow one to calculate the likelihood and then use maximum likelihood estimation to determine model parameters. This yields interpretable regression parameters for each marginal logistic distribution and an overall association parameter. As with generalized least squares, one could weaken these assumptions by allowing the association parameters to depend on specific explanatory variables.

For $p > 2$ and small marginal probabilities, strategies are similar. With small marginal probabilities, the probability of the occurrence of three or more events is small, meaning that one can focus on bivariate association measures. This was the case in Example 2 on homeowners insurance where the overall probability of a homeowners' event during the year was only about 6.5%. By decomposing the loss according to peril, the marginal probabilities were very small and bivariate association measures were most important. In this case, there are $\binom{p}{2}$ pairs of association parameters.

In contrast, for the case study beginning in Section 3, marginal probabilities are not small and one needs to be concerned about higher order interactions. We discuss this situation in the context of the case study.

Other Marginal Binary Regression Approaches. There are two other approaches that are used with multivariate binary regression modelling that we do not consider in detail here. One approach is the multivariate probit model. To fit this model using likelihood methods, one needs to compute a multivariate normal distribution function for each subject. This is computationally feasible for problems with $p = 2$ or 3 but for other applications (such as the case study beginning in Section 3), the computational burden becomes cumbersome when p is larger.

Another approach, widely used in the biomedical community, is known as "alternating logistic regressions," (Carey *et al.* 1993). Alternating logistic regressions uses generalized estimating equations (GEE) methods to estimate binary dependencies. This approach is helpful when the main goal of the analysis is to account for the dependence of the outcome vector using a set of explanatory variables. However, for many actuarial applications, the dependence is an important relationship to be modelled and used in a predictive capacity.

2.2.2 Multivariate Severity Models

There are four features that are desirable in a model of the severity vector. We require a multivariate distribution that:

1. can readily incorporate discrete and continuous explanatory variables,
2. has marginal distributions that may be asymmetric with the ability to handle long tails,
3. allow marginal distributions to differ so that each distribution may have a distinctive shape, and
4. can address the unbalanced nature of observed severities (where the imbalance is dictated by the frequency vector).

Historically, the multivariate normal has been the distribution of choice. This distribution allows one to readily handle features (1) and (4). Power transforms, such as a logarithmic transform, allows analysts to handle feature (2) in a limited fashion. To illustrate feature (3) in automobile losses, one might wish to use a long-tail distribution to handle bodily injury claims yet need only a medium- or short-tail distribution for damage to one's own vehicle. This data feature is difficult to address with multivariate normal distributions.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

In contrast, in this paper (as well as Examples 1–3) we use copula regression models. A copula is a multivariate distribution with uniform marginal distributions on the interval (0, 1). As tools to construct multivariate distributions, copulas are being increasingly explored in the statistics, econometrics, finance and insurance literature. Copulas separate the multivariate joint distribution into two parts: one describing the interdependency of the probabilities, the other describing the marginal distributions only. Through their construction, copulas provide an easy way to describe and simulate jointly distributed random variables. This approach readily accommodates the four features that we find desirable in a multivariate severity model.

2.2.3 Other Approaches to Multivariate Two-Part Modelling

To accommodate potential dependencies, Frees *et al.* (2012) introduced an *instrumental variables* approach in the context of multi-peril homeowners insurance. Instrumental variables is an estimation technique that is commonly used in econometrics to handle dependencies that arise among systems of equations. To illustrate, suppose that we are interested in predicting fire claims and believe that there exists an association between fire and theft/vandalism claims. One would like to use the information in theft/vandalism claims to predict fire claims; however, the number and severity of theft/vandalism claims are unknown when making the predictions. We can, however, use *estimates* of theft/vandalism claims as predictors of fire claims. This is the essence of the instrumental variable estimation method where one substitutes proxies for variables that are not available a priori.

Frees *et al.* (2012) showed statistically significant relationships among perils in their homeowners data and that the instrumental variable technique provided desirable out-of-sample forecasts. Although it is not a likelihood-based method, a notable feature of this approach is that it can be implemented in a two-stage fashion, without the requirement of coding specialized routines.

Robinson *et al.* (2006) developed a hierarchical version of the multivariate two-part model. They were concerned with three levels (services within patient within primary care physician) and developed mixed generalized models to handle dependencies. Liu *et al.* (2010) extended this approach, again using latent variables to represent dependencies among observations that are induced by the hierarchical nature of their sampling (e.g. they examine several patients served by the same physician).

3 Case Study: Health Expenditures

National health expenditures in the US exceeded $2.3 trillion in 2008, almost double the $1.2 trillion spent ten years earlier in 1998. During the last ten years, the average annual percent growth of expenditures was about 7%, which was faster than the growth of the gross domestic product (GDP). This resulted in a steady increase of health expenditures' share of GDP from 13.5% in 1998 to 16.2% in 2008. The per capita spending in health care was $7,681 in 2008, a dramatic increase from the $4,295 spent in 1998 and the $2,814 spent in 1990 (e.g. Centers for Medicare and Medicaid Services, National Health Expenditure Data).

Broken down by inpatient stays, office-based medical doctor visits, outpatient department visits, emergency room visits, and home health visits, the expenditures of each component have grown steadily every year from 1998 to 2008 with only a few exceptions. In particular, spending on office-based medical doctor visits, the largest component, was $27 million in 2008, more than doubling

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

the $11 million spending in 1998. Spending on emergency room visits was the fastest-growing component, with an average annual growth of 11.6% over the 10-year period.

We wish to predict health expenditures of an individual, both in total and by type, e.g. inpatient stays, office-based medical doctor visits, outpatient department visits, emergency room visits, and home health visits. One could certainly predict the expenditures for each type of outcome in isolation of the others and then sum the component predictors to get a prediction of total expenditures. However, as we will see, these outcome types turn out to be correlated, both in terms of the possibility of simultaneous occurrence of more than one type of outcome, and the impact of the severity of one outcome on the other. In an economic context, it is important to find out whether these components of health care are complements or substitutes; providers may have the option of substituting one type of service for another while consumers may elect one type of outcome versus another. By taking into account the dependencies among different types of health outcomes, we hope to arrive at better predictors of total health expenditures.

3.1 Medical Expenditure Panel Survey (MEPS)

The Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. The survey includes household, medical provider and insurance components. The household component provides complete data on the cost and use of health care and health insurance coverage. It also provides respondents' health status, demographic and socio-economic characteristics, employment, access to care, and satisfaction with health care.

Expenditures refer to the amount paid for health care services. Specifically, expenditures in MEPS are defined as the sum of payments for care received, including out-of-pocket payments and payments made by private insurance, Medicaid, Medicare and other sources. This definition differs from a "charge," where the former is a more appropriate proxy for medical expenditures during the 1990s due to the increasingly common practice of discounting, as well as not taking into account uncollected liability, bad debt, and charitable care, which are actually not expenditures because there are no payments associated. Appendix Section 7.1 provides a detailed description of the sources of payments.

In our modelling, we examined whether an expenditure occurred and, if so, how much. Special attention needs to be paid to "zero dollar" medical events, such as those arising from non-payment for services and "flat fee groups." In a flat fee groups, a fixed dollar amount is paid for a group of health care services, e.g. orthodontic care. Appendix Section 7.2 describes zero expenditures and flat fee groups in more detail.

For estimation purposes in Sections 3 and 4, we use panels 10 and 11 from the MEPS data from calendar year 2006. In this data set, there were $n = 18,908$ individuals between ages 18 and 65. For out-of-sample validation purposes in Section 5, we use panels 11 and 12 from calendar year 2007.

3.2 MEPS Data

We decomposed expenditures into five categories:

1. Office-based (OB) provider visits occur in places such as a doctor's or group practice office, medical clinic, managed care plan or HMO centre, or community health centre.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 4. Summary Statistics of Expenditures by Event Types

Event Type	Count*	Percent	Mean	Standard Deviation	Median	Maximum
Office-Based (OB)	10,528	55.7	1,653.31	5,336.31	420.02	199,696.65
Hospital Outpatient (OP)	2,164	11.4	2,817.81	7,517.73	909.96	256,741.18
Emergency Room (ER)	2,274	12.0	1,311.67	2,398.80	566.69	33,412.56
Inpatient (IP)	1,339	7.1	16,488.77	36,133.11	7,548.46	693,483.54
Home Health (HH)	235	1.2	14,092.87	36,611.92	3,312.00	394,913.62

*An observation is a person who has this type of medical event during the year.

2. Hospital outpatient (OP) department visits are visits to a unit of a hospital, a facility, or "urgent care centre" owned by or affiliated with a hospital, examples include obesity clinics, cardiology clinic, and internal medicine department.
3. Emergency room (ER) visits are visits to a medical department at a hospital that is open 24 hours a day where no appointment is necessary.
4. Inpatient admissions (IP) include persons who were admitted to a hospital and stayed overnight. Hospital stays with the same date of admission and discharge are excluded from inpatient counts and expenditures, but are included in either outpatient department or emergency room visit counts and expenditures. Payments associated with emergency room visits that immediately preceded an inpatient stay are included in the inpatient expenditures. Prescribed medicines that are linked to hospital admissions are included in inpatient expenditures.
5. Home health (HH) visits are healthcare provided in a patient's home by healthcare professionals. These services may include some combination of professional health care services and life assistance services. Professional home health services could include medical or psychological assessment, wound care, medication teaching, pain management, disease education and management, physical therapy, speech therapy, or occupational therapy. Life assistance services include help with daily tasks such as meal preparation, medication reminders, laundry, light housekeeping, errands, shopping, transportation, and companionship.

Table 4 summarizes the five expenditure types. This table shows that office-based expenditures are the most prevalent (55.7% of respondents had such a visit during the year) whereas home-health expenditures are the least prevalent (1.2%). Among types of expenditures, inpatient expenditures were typically the most expensive, at least in terms of the mean and the median expenditure amount. Each amount distribution appears to be skewed in that the mean exceeds the median and standard deviation is large relative to the mean.

We chose our explanatory variables from a large set of factors that are described in Table 13. Table 13 also provides summary statistics that suggest factor effects on the probability of positive healthcare expenditures for each of the five types of events. For example, females have higher overall utilization than males throughout all of the five types; and Asian people have lower overall utilization. Many other explanatory variables also show great variation among categories.

3.3 Dependencies

As noted above, outcome types turn out to be related, both in terms of the possibility of simultaneous occurrence of outcomes and the impact of the amount of one outcome on the other. To get a sense of the dependencies in the amounts, Table 5 provides correlations among outcome types. This table shows

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 5. Spearman Correlations Among Five Types of Expenditure

	OB	OP	ER	IP	HH
Office-Based (OB)	1.000	0.245	0.126	0.242	0.088
Hospital Outpatient (OP)	0.245	1.000	0.087	0.212	−0.010
Emergency Room (ER)	0.126	0.087	1.000	0.190	0.003
Inpatient (IP)	0.242	0.212	0.190	1.000	0.035
Home Health (HH)	0.088	−0.010	0.003	0.035	1.000

Table 6. Counts by Number and Type of Events

	None	Singles		Doubles		Triples		Quadruples		Quintuples	
	7687	OB	6703	OB&OP	1303	OB&OP&ER	279	OB&OP&ER&IP	163	OB&OP& ER&IP&HH	19
		OP	130	OB&ER	917	OB&OP&IP	154	OB&OP&ER&HH	11		
		ER	396	OB&IP	427	OB&OP&HH	26	OB&OP&IP&HH	27		
		IP	59	OB&HH	63	OB&ER&IP	362	OB&ER&IP&HH	28		
		HH	7	OP&ER	30	OB&ER&HH	14	OP&ER&IP&HH	0		
				OP&IP	10	OB&IP&HH	32				
				OP&HH	2	OP&ER&IP	9				
				ER&IP	44	OP&ER&HH	0				
				ER&HH	1	OP&IP&HH	1				
				IP&HH	3	ER&IP&HH	1				
Subtotal	7687		7,295		2,800		878		229		19

several positive dependencies. For example, the correlation between office-based (OB) and hospital outpatient (OP) expenditures is 24.5%, indicating that when an OB expenditure is large, the OP expenditure tends to be large (and vice-versa). We use a Spearman correlation (a rank-based version of the usual ordinary, or Pearson, correlation) because of the skewness of our expenditure distributions. Of course, these correlations only suggest a positive relationship among amount types; subsequent sections will control for explanatory variable effects and introduce models to assess their statistical significance.

To get a sense of the dependencies in the frequencies, Table 6 provides counts by the number and types of events. These are the same data introduced in our Section 2.2.1 examples; there, we suggested positive association in joint counts using both dependence ratio and odds ratios. One can use the information in Table 6 to see that there are also higher order positive associations in this data set; for example, the probability of the quadruple OB&OP&ER&IP is $\frac{163}{18,908}$=0.86%; this is much larger than would be suggested by a model of independence. We utilize multivariate binary frequency models to account for this feature.

4 In-Sample Modelling Results

4.1 Marginal Regression Models

4.1.1 Frequency

Following the two-part model framework, the frequency part and the severity part of data are fit separately. For the frequency part, we use logistic models to model the binary outcomes of whether a type of medical event happened; for the severity part, we use gamma regression marginal models. For both the frequency and severity parts, there are five outcomes acting as dependent variables. The "marginal" models are estimated assuming independence among different types of events. Table 14 summarizes the fits of the marginal logit regression models.

Demographic factors such as age and gender are significant determinants of medical services utilization. Females are consistently heavier users of all kinds of medical services, especially the office-based events, where they are 1.8 times more likely than men to visit medical doctors in the office. However, the signs of age are not consistent across all types of events. Contrary to expectations, age is negatively associated with emergency room visits and inpatient stays. The ethnicity factors turn out to be significant determinants for office-based, outpatient department, and emergency room visits, with Asians being the least frequent users, which are also indicated from the summary statistics.

The access to care and its proxies (region variables) are useful determinants except for home health events. People who have their usual source of care (USC) are much more likely to use medical services, especially office-based visits. People from the west seem to use less of these medical services.

People with higher education are more likely to have office-based, hospital outpatient department visits, and inpatient stays. Other socioeconomic factors such as marital status, family size, and income level, show some interesting results across different types of medical services utilization. Marital status variables are significant except for inpatient stays. In particular, people who never married use less ordinary medical services, but have more home health events than their ever-married counterparts. Family size has a negative impact on the probability of healthcare utilization except for inpatient stays and home health visits. Income factors show some interesting implications. People in the highest income group are more likely to have office-based medical doctor visits, but less likely to have emergency room or inpatient visits, and the effect of income on utilization of home health is U shaped, indicating a non-linear relationship.

Health status variables are significantly related to all types of medical events with expected signs. People feeling less comfortable with themselves use more health care services, and those who rated their health as "poor" are generally 3–4 times more likely to have any type of medical events than those who rated their health as "excellent." The self-rated mental health factors are less significant determinants, though they show large differentiation in the summary statistics in Table 13, which may be due to the fact that the mental health effects have been incorporated in the physical health effects in the regression models. Another health status variable, any activity limitation, has a large impact on healthcare utilization, especially on home health visits, with a coefficient larger than 2 and highly significant, indicating that people who have some kind of activity limitation are almost 8 times more likely to have home health care.

The employment factors are interesting as well. People who are unemployed use more office-based, inpatient, and home health care. Considering that inpatient and home health visits are generally the most expensive outcomes, these results show possible financial burden on the unemployed. Across industries, there are a lot of variations shown in Table 13. But after controlling the other covariates, only a few occupations remain slightly significant. For example, people working in education and health, or public administration, are more likely to use some types of healthcare, while people in the natural resource industry are the least frequent users.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 7. Association Test Statistics From Logistic Regression Fits

	OB	OP	ER	IP	HH
Office-Based (OB)	–	10.735	9.084	10.290	1.921
Hospital Outpatient (OP)	10.735	–	8.453	9.995	2.752
Emergency Room (ER)	9.084	8.453	–	25.374	1.709
Inpatient (IP)	10.290	9.995	25.374	–	10.049
Home Health (HH)	1.921	2.752	1.709	10.049	–

For the insurance factors, people who have any kind of insurance at the beginning of the year are significantly more likely to use healthcare services. Whether the coverage is involved with any sort of managed care arrangement turns out not as important, and managed care patients even have slightly more office-based and hospital outpatient department visits, contrary to the cost-containment expectation of managed care plans.

Do the explanatory variables account for the dependencies among outcome types noted in Section 3.3? To assess the dependence, we computed the two-sample t-statistics that were described in Frees *et al.* (2010, Section 2.3); the results appear in Table 7. From this table, we see strong association among all event types with the exception of home health (using the usual cut-off rules for statistical significance). For home health, recall from Table 4 that only 1.2% of respondents had home health expenditures during the year. Thus, we attribute the lack of statistical significance to a small sample size effect.

4.1.2 Severity

Table 15 summarizes the marginal distribution fit of the severity part. Compared to the frequency part, there are fewer covariates that are significant. This reinforces the theory that the choice of healthcare utilization and the subsequent expenditure are two different processes; individual factors have larger impacts on the choice of healthcare utilization while the subsequent expenditures may be more likely to be determined by medical providers.

Demographic factors such as age and gender are still quite significant, though their signs are different from those in the frequency part. Age has a positive effect on expenditure, as expected. Females spend less on medical events such as emergency room, inpatient and home health visits, despite the fact that they use the services more frequently. Ethnicity variables turn out not as significant as in the frequency part.

Access to care variables are not as significant either.

Education factors are only significant in the office-based and home health models. Only in the office-based model do they have expected signs, confirming the idea that higher-educated people also spend more. Income levels have mixed effects on expenditures. We cannot simply say that people with higher income spend more, and in fact for some types of events, people under or near the poverty line have quite large expenditures.

The self-rated physical health variables are still important determinants of office-based, emergency room, and inpatient visits, with healthier people spending less, but the results for outpatient department visits and home health visits are not clear. People with any activity limitation tend to spend more on healthcare, with the exception of emergency room visits.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 8. Spearman's Rho after Controlling for Covariates

	OB	OP	ER	IP	HH
Office-Based (OB)	1	0.231	0.142	0.204	0.188
Hospital Outpatient (OP)	0.231	1	0.046	0.120	0.071
Emergency Room (ER)	0.142	0.046	1	0.160	0.101
Inpatient (IP)	0.204	0.120	0.160	1	0.023
Home Health (HH)	0.188	0.071	0.101	0.023	1

Employment factors show that unemployed people spend more on office-based and home health events, while industry classification is not that important.

Insurance factors show significant effects only on some types of healthcare services. People who are insured at the beginning of the year spend more on office-based and inpatient visits, and those who have managed care plans spend even more on hospital outpatient department visits.

Do the explanatory variables account for the dependencies among outcome types noted in Section 3.3? To assess the dependence, we computed the two-sample t-statistics based on residuals from the marginal model fits. The idea is that residuals represent the value of expenditures having "controlled for" the explanatory variables. The results appear in Table 8 where we see the level of correlations are comparable to the correlations in Table 5 (without controlling for covariates). This table suggests that there remains some positive correlation among outcome types.

4.2 Models of Dependence

4.2.1 Frequency

For the healthcare case study, odds ratio approach is used to model the multivariate binary frequencies. This is a likelihood approach, where the likelihood is written in terms of marginal probabilities and odds ratios. See Appendix Section 8 and Liang *et al.* (1992) for details.

For $p = 5$, full parameterization with odds ratio approach needs 26 association parameters: ten bivariate odds ratios, ten triple odds ratios, five quadruple odds ratios, and one quintuple odds ratio. In this case study, we assume independence between home health visits and all other outcome, i.e., the odds ratios associated with home health are assumed to be zero. This assumption is based on two reasons. Intuitively, home health services mainly target the elderly and may include some combination of professional healthcare services and life assistance services. It is quite different from the other four healthcare services in nature and is less likely to be associated with them. As noted earlier, only 1.2% of observations have home health visits, much fewer than the counts of any other events in our sample that leads to difficulties in assessing strong dependence patterns. Under this assumption, the dependence model is parameterized by six bivariate, four triple, and one quadruple odds ratios.

To estimate the dependence model, we use the explanatory variables that were developed under the independence model. There are many explanatory variables for each healthcare, and so we summarize the results only for the dependence parameters.

Likelihood estimates and t-statistics of odds ratios are summarized in Table 9. The empirical estimates without controlling for covariates are also included in the table for comparison.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 9. Odds Ratios Estimates and t-statistics

		Empirical Estimate without Covariates	Likelihood Estimates with Covariates	t-statistic
Bivariate	OB, OP	10.447	5.604	9.789
	OB, ER	3.371	2.906	10.743
	OB, IP	8.454	6.671	8.378
	OP, ER	2.627	1.985	8.060
	OP, IP	3.551	2.532	8.612
	ER, IP	8.482	6.446	13.186
Triple	OB, OP, ER	0.357	0.439	−6.086
	OB, OP, IP	0.218	0.324	−7.442
	OB, ER, IP	0.437	0.395	−7.202
	OP, ER, IP	0.500	0.559	−5.314
Quadruple	OB, OP, ER, IP	2.916	2.076	0.852

Recall that odds ratio is one under independence, values greater than one indicate positive dependence and values less than one indicate negative dependence. All the bivariate odds ratios are significantly greater than one, indicating positive association between any pair of outcomes choosing from office-based, hospital outpatient, emergency room and inpatient stay. From an economic perspective, the positive association suggests that the four healthcare services are complements in frequency. The strongest association appears to be between office-based visits and inpatient stays while the weakest association appears to be between hospital outpatient visits and emergency room visits.

All the triple odds ratios are significantly less than one, indicating negative third-order associations, yet much less significant than the bivariate associations. This means that given the occurrence of an event, the positive association between two other events is weaker than it would be if the first event had not occurred. In other words, given the utilization of one type of healthcare, an individual is less likely to utilize two more types of healthcare simultaneously than he/she would be if he/she had not utilized the first healthcare. This makes sense intuitively. To illustrate, think about two persons who are otherwise identical except their healthcare utilization. Assume Person A had an office-based visit while Person B did not. Compared to B, A is more likely to have one more other type of healthcare, say hospital outpatient visit, to cure his/her condition since the two events are complements. Now suppose both persons had hospital outpatient visits (recall that A had an office-based visit while B did not). Now A may have already received enough care and thus is less likely to seek one more type of service on top of his/her office-based and hospital outpatient visits.

The quadruple odds ratio is not significant, indicating the absence of fourth-order association. Comparison between the likelihood and empirical estimates shows that with only one exception, the likelihood estimates deviate less from one than the empirical estimates do. Hence, it can be concluded that a proportion of, but not all, associations are explained by the covariates.

4.2.2 Severity

Copula regression is a promising method in continuous severity modelling with bivariate associations. For the severity part of our healthcare case study, Gaussian copulas are used to model the dependence between each pair of outcomes. That is, now we focus only on the bivariate associations, which is different from what we did for the frequency model.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 10. Copula Parameters Estimates

	OB	OP	ER	IP	HH
Office-Based (OB)	1	0.234	0.141	0.175	0.157
Hospital Outpatient (OP)	0.234	1	0.023	0.135	0.062
Emergency Room (ER)	0.141	0.023	1	0.167	0.085
Inpatient (IP)	0.175	0.135	0.167	1	−0.019
Home Health (HH)	0.157	0.062	0.085	−0.019	1

Table 11. Copula Parameters *t*-statistics

	OB	OP	ER	IP	HH
Office-Based (OB)	–	11.877	6.223	7.369	2.848
Hospital Outpatient (OP)	11.877	–	0.551	2.602	0.524
Emergency Room (ER)	6.223	0.551	–	5.170	0.843
Inpatient (IP)	7.369	2.602	5.170	–	−0.268
Home Health (HH)	2.848	0.524	0.843	−0.268	–

Unlike the odds ratio model for frequency part, adding one more outcome for the copula model does not increase much computational difficulty. Hence, we now include Home Health in the dependence modelling. We allow the correlation matrix to be unstructured (subject to being symmetric and invertible), resulting in $\binom{5}{2}=10$ association parameters for $p = 5$ to be estimated, one for each pair of healthcare expenditures. The likelihood for severity part under copula model has been documented extensively (see, e.g. the Appendix of Frees *et al.* 2010).

For consistency with the frequency part, we estimate models including regression covariates developed from independence model, but do not report on this portion of the results. Tables 10 and 11 show the likelihood estimates for the 10 association parameters and the corresponding *t*-statistics respectively. Here, we see positive associations between most pairs, the only exception is between Inpatient Stay and Home Health where the parameter is an insignificant small negative value. From an economic perspective, it suggests that the five types of healthcare are complements in severity. Among the 10 parameters, six are statistically significant. They are between Office-Based and any other outcome, Hospital Outpatient and Inpatient Stay, Emergency Room and Inpatient Stay. The strongest association appears to be between Office-Based and Hospital Outpatient. The lack of statistical significance for the other four pairs is possibly due to the fact that there are relatively few individuals having joint expenditures in these pairs.

With only one exception, Home Health is not significantly associated with other expenditures. This is consistent with the frequency dependence modelling where Home Health did not appear to be related to the other types of expenditures.

5 Out-of-Sample Results

We use the MEPS data from calendar year 2006 for estimation, and apply the estimated coefficients to the MEPS data from calendar year 2007 for out-of-sample validation. Section 5.1 compares

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

our predictions to held-out data at an observation level. Section 5.2 summarizes the distribution of our predictions.

5.1 Out-of-Sample Point Predictions

Table 16 summarizes the out-of-sample statistics measured by six criteria for a range of models.

The first set of models are the Section 2.1.1 univariate one part models. Here, we consider only a single (univariate) outcome – the total expenditure of the five types of healthcare events, and do not separate frequency part from severity part. BasicOnePart fits a basic linear regression; LogOnePart assumes log normal distribution to deal with the skewed outcome; SmearOnePart further applies smearing adjustment to the prediction; Tweedie assumes that the outcome has a Tweedie distribution discussed in Section 2.1.2. See, for example, Frees *et al.* (2011a) for a discussion of the smearing adjustment.

The second set of models consists of the Section 2.1.2 univariate two part models. Here, the single outcome is modelled in two steps – the frequency part and the severity part. BasicTPM assumes that the severity part (conditional on positive expense) follows normal distribution; TPMlogNSev takes the log of the severity; TPMSmearSev does a smearing transformation to the log of severity outcome; TPMGammaSev assumes the severity part follows gamma distribution.

The third set consists of models that are multivariate, where we decompose the total expenditure into five types of outcomes, and assume independence among different types of outcomes. We start with one part models, INDBasicOnePart, INDLogOnePart, and INDOnePartTweed, that use full sets of covariates to model each outcome. Next, we consider INDBasicOnePartReduced, INDLogOnePartReduced, and INDOnePartTweedReduced, that use selected covariates for each outcome based on in-sample estimation results.

The last set consists of multivariate two part models that include dependencies, which are the focus of our case study. The first six in this category are the two-part versions of the previous multivariate one part models. We also explored CellTPMlogNSev that fits a multinomial logistic regression to the frequency part, treating the binary vector as a categorical outcome (there are $2^5 = 32$ categories). The last two models incorporate dependencies among different types of outcomes using the modelling techniques and parameter estimates described in Section 4.2. These two models specify a lognormal and a gamma distribution for the severity part, respectively.

We choose six criteria to measure how each model performs in terms of out-of-sample prediction. The first three statistics are standard out-of-sample validation statistics, e.g. Frees (2010); they measure how far away the predicted values deviate from the observed values in the hold-out sample. Thus, the smaller the numbers, the better are the predictions. The mean absolute (percentage) error computes the average of the (percentage) absolute error between the prediction and the observed value; the root mean square error is the square root of the average squared distance between the prediction and the observed values.

The next three statistics measure the correlation between predicted values and observed values in the hold-out sample. The larger the numbers, the better are the predictions. The Pearson correlation is obtained by dividing the covariance of two variables by their standard deviations; the Spearman correlation is defined as the Pearson correlation coefficient between the ranked variables. That is,

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

the original values need to be converted to ranks, and Spearman correlation is less sensitive than Pearson correlation to outliers in the tails of both samples. The Gini coefficient is a newer measure due to Frees *et al.* (2011b, 2013). Essentially, it measures the correlation between the prediction error and the rank of prediction.

Surprisingly, the models that give better estimation results do not predict better, and there is no model that is obviously superior to the others measured by these six criteria. In general, the two part models perform slightly better than the one part models in that they often have higher correlation statistics. Nonetheless, the Tweedie model is also a good choice considering its simplicity.

5.2 Out-of-Sample Risk Measures

Despite the strong in-sample associations documented in Section 4.2, we found no substantial differences among point predictions in Section 5.1 on an out-of-sample basis. Thus, in the section, we examine the entire distribution of our alternative prediction methods. To provide focus, we consider only the Section 4 multivariate two-part models, comparing the model using the independence assumption for both frequency and severity to that of the model incorporating dependence.

We compare the model estimated in Section 4.1 to that in Section 4.2 by simulating the distribution of total expenditures for the group in our held-out validation sample. Specifically, for the marginal regression models, we use the in-sample estimated parameter values and the explanatory variables associated with the held-out sample to compute location and (if relevant) scale parameters for each person in the sample, for both the frequency (logistic) and severity (gamma) distributions. With these parameters, we then simulated both the occurrence and the amount of each expenditure and used these to calculate the simulated value of the total expenditures for each person in the held-out sample. Expenditures from all individuals were summed to get a realization from the entire held-out sample. We repeated this procedure 1,000 times to get our predictive distribution for total group expenditures.

For the model of dependence, the process was similar but more complex. Simulating severity outcomes using a copula is well-known, see, e.g. the description in Frees *et al.* (2009), Appendix A.3. To simulate from the frequency model, one needs to use the odds-ratio dependencies to simulate the joint occurrence of claims and then simulate this multivariate outcome.

Figure 1 and Table 12 summarize the result of our comparison. For the figure, the smooth density is the predictive distribution from the independence model. The rectangular histogram is from the dependence model. The arrow marks the actual held-out 2007 expenditures.

In Table 12, we use standard actuarial risk measures, the value at risk, *VaR*, and conditional tail expectation, *CTE*, although our approach could be easily extended to other risk measures. The *VaR* is simply a quantile or percentile, the $VaR(\alpha)$ gives the $100(1-\alpha)$ percentile of the distribution. The $CTE(\alpha)$ is the expected value conditional on exceeding the $VaR(\alpha)$. See, for example, Frees *et al.* (2009) for another predictive modelling application using these risk measures. In Table 12, *VaRInd* and *CTEInd* are the value at risk and conditional tail expectation for the independence model whereas *VaRDep* and *CTEDep* are the corresponding measures for the dependence model.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

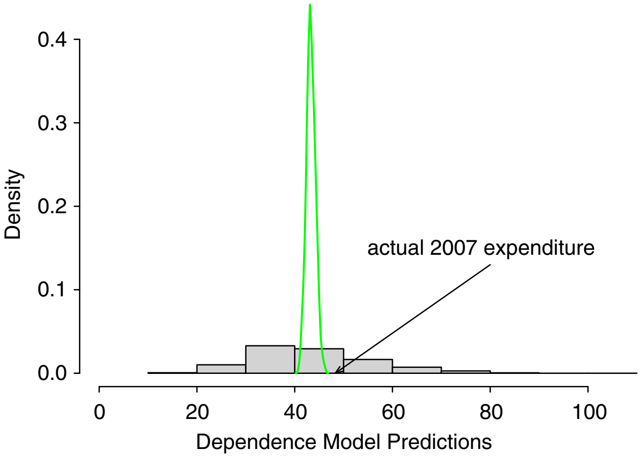

Figure 1. Comparison of Predictive Distributions. The smooth density is the predictive distribution from the independence model. The rectangular histogram is from the dependence model. The arrow marks the actual held-out 2007 expenditures.

Table 12. Out-of-Sample Risk Measures

Risk	Percentile						
Measure	0.50	0.75	0.90	0.95	0.98	0.99	1.00
VaRInd	45.05	45.63	46.18	46.56	46.87	47.32	47.44
VaRDep	41.91	50.51	58.83	65.06	68.86	73.18	79.35
CTEInd	45.76	46.20	46.67	46.97	47.25	47.55	47.70
CTEDep	52.39	58.86	66.37	71.27	75.75	81.61	86.90

As anticipated, both the figure and table show that the predictive distribution for the independence model is much narrower than the model for dependence. In basic probability theory that we teach students, for two random variables X and Y, that $Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)$ (here, "*Var*" is for "variance" and "*Cov*" is for "covariance"). That is, if X and Y are positively associated, then a model that assumes independence will display less uncertainty than a model that accounts for dependence; $Var(X)+Var(Y)<Var(X+Y)$. In the same vein, Table 12 shows that the model incorporating dependence has a wider spread than the model that ignores these features.

Unfortunately, for our analysis, we have only one held-out realization. Total expenditures in 2007 turned out to be $ 48.34 million. This is unlikely to occur in the model of independence although very plausible in the dependence model. Of course, this does not validate the model of dependence but it is consistent with what we learned from our detailed in-sample analysis.

As a final remark, we note that this sampling procedure was done using the estimated parameters as fixed quantities. An alternative procedure to incorporate their sampling variability into this analysis would be to bootstrap the results.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 13. Covariate Description and Percentage of Positive Expenditures by Level of Covariates

				Percentage of Positive Expend				
Category	Variable	Description	Percentage of Data	OB	OP	ER	IP	HH
Demography	FEMALE	1 if female	53.48	63.03	13.58	13.74	9.31	1.53
	MALE	0 if male	46.52	47.23	8.99	10.06	4.52	0.91
Ethnicity	ASIAN	1 if Asian	4.18	46.27	6.32	5.18	4.93	0.25
	BLACK	1 if Black	16.81	49.76	10.16	15.63	8.43	2.08
	WHITE	Reference level is white, multiple races and native American	79.00	57.44	11.99	11.62	6.91	1.12
Region	NORTHEAST	1 if Northeast	14.63	59.81	15.40	12.14	6.94	1.52
	MIDWEST	1 if Midwest	20.26	61.52	15.51	13.60	7.70	1.46
	SOUTH	1 if South	37.35	54.14	10.35	13.00	7.70	1.15
	WEST	Reference level	27.75	51.31	7.87	9.51	5.87	1.07
Access to Care	YES	1 if have USC	70.64	67.75	14.64	13.51	8.33	1.58
	NO	0 if otherwise	29.36	26.63	3.77	8.47	4.09	0.43
Education	COLLEGE	1 if college or higher degrees	29.07	63.23	12.33	9.53	6.60	0.84
	HIGHSCHOOL	1 if high school degree	44.59	55.26	12.22	12.58	6.99	1.22
	LOWERHS	Reference level is lower than high school	26.34	48.05	9.16	13.84	7.77	1.73
Marital Status	MARRIED	1 if married	55.69	59.23	12.34	10.19	6.78	0.73
	WIDIVSEP	1 if widowed or divorced or separated	16.33	63.47	15.58	16.84	9.33	2.69
	NEVMAR	Reference level	27.98	44.07	7.26	12.87	6.37	1.42
Income Compared to Poverty Line	HINCOME	1 if high income	33.49	63.19	13.12	9.70	5.18	0.71
	MINCOME	1 if middle income	29.21	54.39	11.26	10.03	6.10	0.72
	LINCOME	1 if low income	15.35	47.55	9.03	12.65	6.89	1.14
	NPOOR	1 if near poor	5.74	48.80	9.67	15.29	9.39	1.75
	POORNEG	Reference level is poor/negative	16.20	52.61	11.23	18.70	12.14	3.20
Self-rated Physical Health	POOR	1 if poor	4.00	84.66	29.63	31.35	23.68	8.33
	FAIR	1 if fair	11.48	73.98	19.90	21.14	12.34	4.01
	GOOD	1 if good	29.06	57.05	11.83	12.17	6.99	0.80
	VGOOD	1 is very good	30.38	54.33	9.49	9.56	5.41	0.54
	EXCELLENT	Reference level is in excellent health	25.08	42.72	6.60	7.59	4.15	0.21
Self-rated Mental Health	GOOD	0 if good, very good and secellent mental health	92.24	53.92	10.69	10.99	6.41	0.88
	POOR	1 if poor or fair mental health	7.76	76.55	20.45	24.34	15.06	5.52
Any limitation	ANYLIMIT	1 if any functional or activity limitation	22.92	77.89	23.03	21.23	13.92	4.75
		o if otherwise	77.08	49.08	8.00	9.29	5.05	0.20

Table 13 *(Continued)*

				Percentage of Positive Expend				
Category	Variable	Description	Percentage of Data	OB	OP	ER	IP	HH
Unemployment	UNEMPLOYED	1 if unemployed at the beginning of 2006	29.57	60.33	14.47	16.44	11.97	3.27
		0 if otherwise	70.43	53.73	10.17	10.17	5.03	0.39
Industry	NATRESOURCE	1 if natural resources	1.15	46.54	8.76	5.07	1.84	0.00
Classification	MINCONST	1 if mining or construction	6.24	37.46	6.86	10.17	2.80	0.08
	MANUFACT	1 if manufacturing	8.55	54.70	10.09	9.96	3.40	0.37
	SALES	1 if sales	8.83	50.48	9.58	10.54	5.39	0.24
	TRANSINFO	1 if transportation, utilities and information	5.26	54.02	10.06	9.56	4.53	0.50
	FINANCE	1 if finance, insurance, or real estate	4.32	58.14	9.42	10.04	5.39	0.37
	PROFSERV	1 if professional services	7.87	51.28	9.27	9.21	4.57	0.40
	EDUCHEALTH	1 if education, health and social services	14.73	64.92	13.82	11.45	6.97	0.83
	LEISURE	1 if leisure and hospitality	5.98	44.78	6.64	10.35	5.75	0.09
	PUBADMIN	1 if public administration	3.56	65.97	14.12	11.89	6.98	0.30
	MILITARY	1 if active military	0.17	40.63	15.63	3.13	0.00	0.00
	OTHERSERV	1 if other services	3.76	47.68	8.02	7.88	3.52	0.14
		Reference level is inapplicable	29.59	60.29	14.46	16.43	11.96	3.27
Insurance	INSURED	1 if is insured at the beginning of the year 2006	72.02	64.00	14.03	12.85	8.02	1.61
		0 if otherwise	27.98	34.27	4.80	9.91	4.67	0.30
Managed Care	MANAGEDCARE	1 if is enrolled in a managed care plan	52.17	64.26	13.73	12.07	7.32	1.13
		0 if otherwise	47.83	46.32	8.96	11.97	6.82	1.37

Table 14. Marginal Logistic Regressions for Five Types of Events

		Office-Based			Hospital Outpatient			Emergency Room			Inpatient			Home Health		
Category	Variable	Estimate	t-value		Estimate	t value		Estimate	t value		Estimate	t value		Estimate	t value	
	(Intercept)	−2.463	−23.798	***	−5.219	−31.313	***	−2.224	−16.939	***	−4.279	−23.343	***	−8.968	−17.865	***
Demography	AGE	0.007	4.046	***	0.015	6.352	***	−0.019	−9.295	***	−0.007	−2.895	**	0.030	4.415	***
	GENDER	0.584	16.167	***	0.398	7.935	***	0.201	4.188	***	0.563	8.629	***	0.246	1.672	.
Ethnicity	ASIAN	−0.492	−5.853	***	−0.486	−3.163	**	−0.621	−3.764	***						
	BLACK	−0.438	−9.118	***	−0.260	−3.717	***	0.155	2.625	**						
Region	NORTHEAST	0.041	0.735		0.598	7.706	***	0.214	2.748	**	0.117	1.186				
	MIDWEST	0.139	2.733	**	0.585	8.071	***	0.310	4.453	***	0.249	2.825	**			
	SOUTH	0.044	0.996		0.197	2.855	**	0.196	3.142	**	0.197	2.578	**			
Access to Care	USC	1.292	32.451	***	0.822	10.352	***	0.362	6.020	***	0.422	5.189	***			
Education	HIGHSCH	0.104	2.310	*	0.235	3.590	***				0.121	1.628				
	COLLEGE	0.289	5.338	***	0.219	2.979	**				0.302	3.310	***			
Marital Status	MARRIED	0.248	5.093	***	0.209	2.762	**							−1.022	−5.314	***
	WIDIVSEP	0.144	2.337	*	0.128	1.505		0.258	4.175	***				−0.503	−2.675	**
Family Size	FAMSIZE	−0.114	−9.749	***	−0.076	−4.090	***	−0.038	−2.519	*	0.058	3.201	**			
Income	HINCOME	0.131	2.059	*				−0.378	−5.148	***	−0.586	−5.919	***	−0.229	−1.008	
	MINCOME	0.063	1.084					−0.445	−6.404	***	−0.417	−4.664	***	−0.518	−2.464	*
	LINCOME	−0.072	−1.143					−0.241	−3.183	**	−0.338	−3.486	***	−0.372	−1.730	.
	NPOOR	−0.115	−1.363					−0.093	−0.927		−0.080	−0.650		−0.284	−1.063	
Physical Health	POOR	1.308	10.535	***	1.108	9.789	***	1.144	10.122	***	1.184	9.044	***	1.449	3.914	***
	FAIR	1.116	15.785	***	0.863	9.638	***	0.850	9.826	***	0.706	6.438	***	1.294	3.630	***
	GOOD	0.587	12.507	***	0.488	6.421	***	0.425	5.971	***	0.409	4.355	***	0.550	1.522	
	VGOOD	0.421	9.413	***	0.262	3.443	***	0.224	3.123	**	0.228	2.407	*	0.663	1.799	.
Mental Health	MNHPOOR	0.301	3.792	***				0.125	1.608							
Any Limitation	ANYLIMIT	0.777	15.741	***	0.726	12.652	***	0.574	9.918	***	0.631	8.617	***	2.051	9.032	***
Unemployment	UNEMPLOYED	0.139	3.130	**							0.482	6.441	***	1.244	5.481	***
Industry	EDUCHEALTH	0.105	2.021	*							0.160	1.702	.	0.745	2.602	**
	PUBADMIN	0.208	2.251	*							0.336	2.064	*			
	NATRESOURCE							−0.723	−2.300	*	−0.829	−1.621				
Insurance	INSURED	0.672	13.202	***	0.662	7.949	***	0.297	5.027	***	0.532	6.673	***	1.424	5.364	***
Managed Care	MANAGEDCARE	0.157	3.540	***	0.138	2.371	*									
Model fit indices	AIC	20,734.74			11,785.83			12,989.63			8,834.53			1,851.78		
	Log-Likelihood	−10,338.37			−5,871.92			−6,471.82			−4,393.27			−908.89		

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Table 15. Marginal Regressions for Expenditure of Five Types of Events Assuming Gamma Distribution

		Office-Based			Hospital Outpatient			Emergency Room			Inpatient			Home Health		
Category	Variable	Estimate	t-value		Estimate	t value		Estimate	t value		Estimate	t value		Estimate	t value	
	(Intercept)	5.571	33.540	***	6.587	24.176	***	6.694	37.459	***	8.445	39.250	***	8.559	11.913	***
Demography	AGE	0.010	4.628	***	0.012	2.865	**	0.008	2.905	**	0.014	4.105	***	0.009	0.921	
	GENDER	0.272	5.150	***	−0.096	−0.932		−0.112	−1.539		−0.250	−2.716	**	−0.621	−2.853	**
Ethnicity	ASIAN	−0.410	−2.910	**	−0.346	−1.054					−0.369	−1.522				
	BLACK	0.059	0.805		−0.216	−1.524					0.162	1.541				
Region	NORTHEAST	0.011	0.134		0.013	0.081					−0.201	−1.481				
	MIDWEST	0.030	0.403		0.386	2.622	**				−0.103	−0.856				
	SOUTH	−0.145	−2.153	*	0.205	1.430					−0.275	−2.541	*			
Access to Care	USC	0.077	1.009		0.005	0.031		0.116	1.303		0.112	1.008		0.732	2.115	*
Education	HIGHSCH	0.129	1.845	.										−0.530	−2.222	*
	COLLEGE	0.236	2.931	**										−0.538	−1.641	
Marital Status	MARRIED													−0.516	−2.117	*
	WIDIVSEP							−0.206	−2.252	*						
Family Size	FAMSIZE	−0.041	−2.401	*				−0.060	−2.784	**						
Income Compared to Poverty Line	HINCOME	0.205	2.197	*	0.318	1.996	*	0.253	2.389	*	0.278	2.248	*	0.152	0.419	
	MINCOME	0.053	0.602		0.311	1.977	*	0.249	2.435	*	0.151	1.315		−0.940	−2.995	**
	LINCOME	0.026	0.271		0.479	2.550	*	0.036	0.323		−0.112	−0.866		0.301	0.940	
	NPOOR	0.331	2.522	*	−0.105	−0.414		−0.001	−0.004		−0.011	−0.065		0.041	0.103	
Physical Health	POOR	0.956	7.322	***				0.560	3.771	***	0.946	5.319	***	−0.665	−1.124	
	FAIR	0.649	6.850	***				0.249	2.036	*	0.235	1.523		−1.421	−2.462	*
	GOOD	0.315	4.132	***				0.113	1.025		0.118	0.898		−0.983	−1.709	.
	VGOOD	0.121	1.625					0.125	1.106		0.017	0.124		−1.438	−2.432	*
Mental Health	MNHPOOR										−0.232	−1.888	.	−0.430	−1.800	.
Any Limitation	ANYLIMIT	0.644	10.148	***	0.362	3.369	***				0.300	2.977	**	0.930	2.494	*
Unemployment	UNEMPLOYED	0.340	5.409	***										1.104	3.672	***
Insurance	INSURED	0.332	4.629	***							0.391	3.614	***			
Managed Care	MANAGEDCARE				0.321	2.987	**									
Model fit indices	AIC	170,113.40			37,790.91			36,953.51			28,283.05			4,744.50		
	Log-Likelihood	−85,032.69			−18,879.45			−18,461.76			−14,120.53			−2,353.25		
	Dispersion	6.76			5.14			2.73			2.11			2.38		

Table 16. Out-of-Sample Statistics

	Mean Absolute	Mean Absolute	Root Mean	Correlations		
	Error	Percentage Error	Square Error	Pearson	Spearman	Gini
One Part Models						
BasicOnePart	3,874.938	227.540	13,774.179	25.680	44.688	18.296
LogOnePart	2,684.980	6,993.282	14,300.376	22.565	51.515	18.608
SmearOnePart	10,934.435	174.691	35,244.110	22.565	51.515	18.608
Tweedie	3,589.482	150.285	13,734.769	26.646	49.616	18.896
Two Part Models						
BasicTPM	3,665.951	329.365	13,740.511	26.621	45.791	18.498
TPMLogNSev	2,705.836	525.740	14,209.043	27.036	50.521	18.896
TPMSmearSev	3,630.774	160.093	13,718.385	27.036	50.521	18.896
TPMGammaSev	3,579.156	156.046	13,720.311	27.091	50.109	18.893
Multivariate One Part Models						
INDBasicOnePart	3,874.938	227.540	13,774.179	25.680	44.688	18.296
INDLogOnePart	2,719.334	9,387.637	14,433.423	22.584	51.474	18.350
INDOnePartTweed	2,781.582	9,787.627	14,513.104	21.256	47.050	18.376
INDBasicOnePartReduced	3,863.703	211.227	13,773.980	25.694	45.006	18.399
INDLogOnePartReduced	2,719.343	9,415.450	14,433.430	22.583	51.474	18.353
INDOnePartTweedReduced	2,781.683	9,526.059	14,513.085	22.570	46.912	18.509
Multivariate Two Part Models						
INDBasicTPM	3,603.476	265.246	13,731.740	26.811	48.913	18.940
INDTPMLogNSev	2,851.499	258.909	14,075.163	26.351	49.191	18.710
INDTPMGammaSev	3,579.410	146.869	13,736.173	26.677	49.794	18.974
INDBasicTPMReduced	3,597.650	192.595	13,732.308	26.803	48.663	18.834
INDTPMLogNSevReduced	2,848.996	258.562	14,062.045	26.879	49.856	18.743
INDTPMGammaSevReduced	3,574.799	149.014	13,726.846	26.883	50.184	18.956
CellTPMLogNSev	3,053.903	185.672	14,013.866	25.429	50.036	18.773
DepTPMLogNSevReduced	2,822.35	272.848	14082.69	26.967	50.046	18.854
DepTPMGammaSevReduced	3520.812	152.459	13732.83	26.783	50.225	18.996

6 Concluding Remarks

Multivariate two-part regression models can be applied in many different areas of actuarial practice. In this paper, we have cited applications in healthcare, property and casualty (general), and life insurance. Our detailed case study of healthcare expenditures shows that it is important to distinguish between when an event may occur (frequency) and, if it occurs, the amount of expenditure (severity). The explanatory variables that influence an expenditure may differ by whether one is modelling frequency or severity. Moreover, we have found that the association among differ types depends greatly on whether one is modelling frequency or severity.

In this paper, we have used the descriptor “frequency” to describe whether or not an event may occur. We note that there are many important situations in actuarial practice when the analyst has a *count* of claims and would use a frequency distribution such as a Poisson or negative binomial distribution for modelling. This is also referred to as a “frequency” problem. Clearly, an interesting extension of this paper would be to consider multivariate regression problems where the number of claims for each event type is available.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

Acknowledgment

The first author thanks the Assurant Health Insurance Professorship for funding to support this research.

References

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

de Jong, P. & Heller, G.Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press, Cambridge, UK.

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

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

Frees, E.W. (2010). *Regression Modelling with Actuarial and Financial Applications*. Cambridge University Press, New York.

Frees, E.W., Gao, J. & Rosenberg, M. (2011a). The frequency and amount of inpatient and outpatient healthcare expenditures. *North American Actuarial Journal*, **15**, 377–392.

Frees, E.W., Meyers, G. & Cummings, A.D. (2010). Dependent multi-peril ratemaking models. *Astin Bulletin*, **40**(2), 699–726.

Frees, E.W., Meyers, G. & Cummings, A.D. (2011b). Summarizing insurance scores using a Gini index. *Journal of the American Statistical Association*, **106**, 1085–1098.

Frees, E.W., Meyers, G. & Cummings, A.D. (2012). Predictive modelling of multi-peril homeowners insurance. To appear in *Variance*.

Frees, E.W., Meyers, G. & Cummings, A.D. (2013). Insurance ratemaking and a Gini index. To appear in the *Journal of Risk and Insurance*.

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

Frees, E.W. & Sun, Y. (2010). Household life insurance demand – a multivariate two-part model. *North American Actuarial Journal*, **14**(3), 338–354.

Glonek, G.F.V. & MuCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society B*, **57**(3), 533–546.

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

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

Liu, L., Strawderman, R.L., Cowen, M.E. & Shih, Y.T. (2010). A flexible two-part random effects model for correlated medical costs. *Journal of Health Economics*, **29**, 110–123.

Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, **17**(3), 247–281.

Robinson, J.W., Zeger, S.L. & Forrest, C.B. (2006). A hierarchical multivariate two-part model for profiling providers' effects on health care charges. *Journal of the American Statistical Association*, **101**, 911–923.

Sun, Y. (2011). *Micro-Econometric Modelling of Personal Lines Insurance*, unpublished dissertation, University of Wisconsin-Madison.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

7 Appendix – Additional Survey Details

7.1 Sources of Payment

The total expenditure for medical services is the sum of 12 sources of payment: (1) out-of-pocket by the user or family, (2) Medicare, (3) Medicaid, (4) private insurance, (5) Veterans Administration, excluding TRICARE/CHAMPVA, (6) TRICARE/CHAMPVA, (7) other federal sources including Indian health services, Military treatment facilities, and other care by the Federal government, (8) other state and local sources including community and neighbourhood clinics, state and local health departments, and state programs other than Medicaid, (9) workers' compensation, (10) other unclassified sources including sources such as automobile, homeowner's, and liability insurance, and other miscellaneous or unknown sources, (11) other private such as any type of private insurance payments reported for persons not reported to have any private health insurance coverage during the year as defined in MEPS, and (12) other public such as Medicare/Medicaid payments reported for persons who were not reported to be enrolled in the Medicare/Medicaid program at any time during the year.

7.2 Zero Expenditures Flat Fee Groups

There are some medical events reported by respondents where the payments were zero. Zero payment events can occur in MEPS for the following reasons:

1. The event was covered under a flat fee arrangement (flat fee payments are included only on the first event covered by the arrangement).
2. There was no charge for a follow-up stay.
3. The provider was never paid by an individual, insurance plan, or other source for services provided.
4. Charges were included in another bill (e.g. for emergency room visits that have a subsequent inpatient stay).
5. The event was paid for through government or privately-funded research or clinical trials.
6. For office-based or outpatient department files that contain events involving a telephone call rather than a medical provider visits, there are no expenditure associated.
7. All expenditures for home health care provided by informal care providers (family, friends, or volunteers) were assigned -1 "inapplicable" because those types of events were skipped out (never asked) of the questions regarding expenditures.

The approach used to count expenditures for flat fee groups was to place the expenditure on the first visit of the flat fee group (stem event). The remaining visits (leaves) have zero facility payments, while physician's expenditures may be still present. Thus, if the first visit in the flat fee group occurred prior to 2006, all of the events that occurred in 2006 will have zero payments. Conversely, if the first event in the flat fee group occurred at the end of 2006, the total expenditure for the entire flat fee group will be on that event, regardless of the number of events it covered after 2006.

There are no flat fee groups regarding prescribed medicine or home health visits. Outpatient and office-based medical provider visits are the only two event types allowed in a single flat fee group. The stem may have been reported as an outpatient department visit and the leaves may have been reported as office-based medical provider visits.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press

8 Appendix – Odds Ratios

As emphasized in equation (3), odds ratios can be interpreted as conditional odds. This interpretation is particularly convenient when defining higher order associations. We follow standard statistical literature (e.g. Liang *et al.* 1992, Section 2) and define higher order association measures in terms of contrasts of conditional odds ratios. Specifically, define the third order association measure

$$\zeta_{123} = \zeta_{123}(r_1, r_2, r_3) = \ln \mathrm{OR}((r_1, r_2)|r_3 = 1) - \ln \mathrm{OR}((r_1, r_2)|r_3 = 0) \tag{4}$$

and the fourth order association measure

$$\begin{aligned} \zeta_{1234} &= \zeta_{1234}(r_1, r_2, r_3, r_4) \\ &= \ln \mathrm{OR}((r_1, r_2)|r_3 = 1, r_4 = 1) + \ln \mathrm{OR}((r_1, r_2)|r_3 = 0, r_4 = 0) \\ &\quad -\ln \mathrm{OR}((r_1, r_2)|r_3 = 0, r_4 = 1) - \ln \mathrm{OR}((r_1, r_2)|r_3 = 1, r_4 = 0). \end{aligned} \tag{5}$$

Higher order association parameters are possible (e.g. Liang, *et al.* 1992, Section 2) but will not be needed for our application. That is, for our application with $p = 5$, we assume independence between home health visits and all other events, and set the log odds ratios associated with home health to zero.

Bounds

Knowledge of the marginal means π_j and the odds ratios are sufficient to determine joint probabilities and hence evaluate likelihood functions. To illustrate, consider equation (3). Simple algebra shows that

$$\exp(\zeta_{12}) = \frac{\pi_{11}(1 - \pi_1 - \pi_2 + \pi_{11})}{(\pi_1 - \pi_{11})(\pi_2 - \pi_{11})},$$

where $\pi_{11} = \Pr(r_1 = 1, r_2 = 1)$. From this expression, we may determine π_{11} as the solution to a quadratic function involving ζ_{12}, π_1, and π_2:

$$\begin{aligned} (\exp(\varsigma_{12}) - 1)\pi_{11}^2 - [(\exp(\varsigma_{12}) - 1)(\pi_1 + \pi_2) + 1]\pi_{11} \\ + \exp(\varsigma_{12})\pi_1\pi_2 = 0. \end{aligned} \tag{6}$$

With this joint probability and marginal means, other joint probabilities $\Pr(r_1 = j, r_2 = k)$, $\{j, k\}$=$\{0, 1\}$ may be readily determined.

Knowledge of third order association parameters ζ_{123}, together with marginal means π_j, j = 1,2,3 and joint probabilities $\Pr(r_1 = j,\ r_2 = k)$ allows one to determine joint probabilities of the form $\Pr(r_1 = j, r_2 = k, r_3 = l)$, for $\{j, k, l\}$ in $\{0, 1\}$. The triple joint probabilities are the solution of a fourth order polynomial equation. Similarly, quadruple probabilities can be found as the solution of an eighth-order degree polynomial.

There are two difficulties in determining the joint probabilities from knowledge of the marginal means and association parameters. The first is a computational one. Solutions to second, fourth and eighth-order degree polynomials must be found for *each* observation for *each* evaluation of the likelihood function. This means that even for data sets that are moderately sized (about 19,000 observations for our application), computational concerns arise. Second, joint probabilities are bounded by lower-order joint probabilities. For example, it is easy to see that

$$\max(0, \pi_1 + \pi_2 - 1) \leq \pi_{11} \leq \min(\pi_1, \pi_2). \tag{7}$$

Because of this, when we compute the solutions to high-order polynomials for each observation, the roots depend heavily on the specification of the association parameters.

Equation (7) implies that the bounds for bivariate dependence ratio are

$$\frac{\max(0, \pi_1 + \pi_2 - 1)}{\pi_1 \pi_2} \leq \tau_{12} \leq \left(\frac{1}{\pi_1}, \frac{1}{\pi_2} \right), \tag{8}$$

Recall that with dependence ratio approach, it is customary to define a constant dependence ratio over observations. When marginal probabilities are small, the dependence ratio approach works well since equation (8) could be easily satisfied by a constant dependence ratio. For high-frequency events, e.g. office-based visits, the marginal probability is high and equation (8) leads to very strict restrictions on the constant dependence ratio. Care must be taken to set appropriate boundaries for dependence ratio in the modelling algorithm.

With odds ratio approach, it is customary to define a constant odds ratio over observations. While the odds ratio is constant, the dependence ratio varies over observations. This allows one to apply the approach even when the marginal probabilities are relatively large. Taking $\exp(\varsigma_{12}) \rightarrow 0$, it is easy to solve for π_{11} from equation (6):

$$\pi_{11} = \frac{\pi_1 + \pi_2 - 1 + |\pi_1 + \pi_2 - 1|}{2} = \max(0, \pi_1 + \pi_2 - 1).$$

Taking $\exp(\varsigma_{12}) \rightarrow \infty$, it is easy to see that

$$\pi_{11} = \frac{\pi_1 + \pi_2 - |\pi_1 - \pi_2|}{2} = \min(\pi_1, \pi_2).$$

Hence, a mild constraint on the odds ratio, i.e., $\exp(\varsigma_{12}) > 0$ would make equation (7) satisfied even if the marginal probabilities are relatively large. In this sense, given a fixed association parameter over observations, the odds ratio approach is preferred to the dependence ratio approach when the outcomes of interest contain one or more frequent events.

https://doi.org/10.1017/S1748499512000346 Published online by Cambridge University Press