

# Multivariate probit models for conditional claim-types

GARY YOUNG  
School of Economics  
Faculty of Business  
University of New South Wales  
Sydney, Australia 2052  
e-mail: g.young@unsw.edu.au

EMILIANO A. VALDEZ  
School of Actuarial Studies  
Faculty of Business  
University of New South Wales  
Sydney, Australia 2052  
e-mail: e.valdez@unsw.edu.au

ROBERT KOHN  
School of Economics and Finance  
Faculty of Business  
University of New South Wales  
Sydney, Australia 2052  
e-mail: r.kohn@unsw.edu.au

December 22, 2006

## Abstract

This paper considers statistical modeling of the types of claim in a portfolio of insurance policies. For some classes of insurance contracts, in a particular period, it is possible to have a record of whether or not there is a claim on the policy, the types of claims made on the policy, and the amount of claims arising from each of the type. A typical example is automobile insurance where in the event of a claim, we are able to observe the amounts that arise from say injury to oneself, damage to one's own property, damage to a third party's property, and injury to a third party. Modeling the frequency and the severity components of the claims can be handled using traditional actuarial procedures. However, modeling the claim-type component is less known and in this paper, we recommend analyzing the distribution of these claim types using multivariate probit models, which can be viewed as latent variable threshold models for the analysis of multivariate binary data. A recent article by Valdez and Frees (2005) considered this decomposition to extend the traditional model by including the conditional claim-type component, and proposed the multinomial logit model to empirically estimate this component. However, it is well-known in the literature that this type of model assumes independence across the different outcomes. We investigate the appropriateness of fitting a multivariate probit model to the conditional claim-type component in which the outcomes may in fact be correlated, with possible inclusion of important covariates. Our estimation results show that when the outcomes are correlated, the multinomial logit model produces substantially different predictions relative to the true predictions; and second, through a simulation analysis, we find that even in ideal conditions under which the outcomes are independent, multinomial logit is still a poor approximation to the true underlying outcome probabilities relative to the multivariate probit model. The results of this paper serve to highlight the trade-off between tractability and flexibility when choosing the appropriate model.

*Keywords:* Correlation; Insurance; Multinomial Logit.

## 1 Introduction and previous literature

Consider a portfolio of insurance policies with  $J$  possible claim types. Denote by  $I_j$  to be the indicator variable that a contract from this portfolio had a claim of type  $j$  where  $j = 1, 2, \dots, J$ . Denote the vector

$$M = (I_1, I_2, \dots, I_J)'$$

to be the vector of claim types for this policy. One can think of the zero vector as the vector of claim types that gives rise to the situation when there is no claim. Thus, there are  $2^J$  possible combinations of claim type vectors  $M$ . This random variable is clearly a discrete random variable with possible values describing the different possible combinations of  $I_j$ . In this paper, we restrict our considerations to observations in which a claim was made so that we simply have  $2^J - 1$  claim-types. For our purposes, we call our observations conditional claim-type components.

As a simple illustration, consider a portfolio of automobile insurance policies where we can observe two possible claim types: personal injury and damage to property. Denote the case  $j = 1$  for personal injury and  $j = 2$  for damage to property. Thus, there are 4 possible combinations of  $M$ :  $(0, 0)$  which corresponds to the case where there is no claim,  $(1, 0)$  which corresponds to the case where there is claim due to personal injury only,  $(0, 1)$  which corresponds to the case where there is a claim due to damage to property only, and  $(1, 1)$  which corresponds to the case where there is a claim arising from personal injury as well as property damage. Our model formulation does not preclude us from considering the possible dependency between the occurrence of personal injury and the occurrence of damage to property. One would suspect some positive form of dependency. In the case of an automobile accident, when there is damage to property, the chances are high that there will also be personal injuries associated with it.

As is already well-known in the actuarial literature, the aggregate loss distribution is obtained by its traditional decomposition into the frequency and severity components, where each component is then modeled separately. There has been almost no attempt in the actuarial literature to model the conditional claim-type component of the aggregate loss distribution. Recently, however, Valdez and Frees (2005), in a working paper, proposed a hierarchical model structure for the estimation and prediction of the aggregate loss distribution using a highly detailed, micro-level, automobile insurance dataset. To illustrate this decomposition, consider a risk class  $i$  at calendar year  $t$  for which the potential observable responses for observational unit  $\{it\}$  consist of:

- $K_{it}$  the indicator of a claim within a year;
- $M_{it,j}$  the type of claim, available for each claim,  $j = 1, \dots, N_t$ ;
- $A_{it,jk}$  the loss amount, available for each claim,  $j = 1, \dots, N_t$ , and  
for each type of claim  $k = 1, \dots, m$ .

where  $N_t$  denotes the number of claims within the calendar year  $t$ . The joint density function of the aggregate loss can then be decomposed as:

$$f(K, \mathbf{M}, \mathbf{A}) = f(K) \times f(\mathbf{M}|K) \times f(\mathbf{A}|K, \mathbf{M})$$

joint = frequency  $\times$  conditional claim-type  $\times$  conditional severity,

where  $f(K, \mathbf{M}, \mathbf{A})$  denotes the joint aggregate loss density and is equal to the product of the frequency, conditional claim-type, and conditional severity components. In the notation used,  $f(K)$  denotes the frequency component and is equal to the probability of having a claim (or no claim) in a given calendar year;  $f(\mathbf{M}|K)$  denotes the conditional claim-type component and is equal to the probability of having a claim-type of  $\mathbf{M}$ , given  $K$ ; and  $f(\mathbf{A}|K, \mathbf{M})$  denotes the conditional severity component, and is equal to the probability density of the claim vector  $\mathbf{A}$  given  $K$  and  $\mathbf{M}$ . Here, conditional on having observed a claim, the random variable  $\mathbf{M}$  describes the combination observed. Each combination observed is an  $m$ -tuple of the form  $(i_1, \dots, i_m)$ , where each  $i_k$ , for  $k = 1, \dots, m$ , is equal to one if the  $k$ th type of claim is observed and zero otherwise. Thus, the  $m$ -tuple  $(1, 0, \dots, 0)$  means that there was a claim

with respect to “first type of claim” only, and similarly,  $(1, 1, \dots, 1)$  means that all types of claims were observed.

Suppose that we have multinomial observations such that  $M_i$  is the observation on unit (or individual)  $i$ , with  $M_i$  taking  $m$  possible values. We assume for now that the  $M_i$  are generated by the multinomial logistic model

$$\Pr(M_i = j|x_i) = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta}_j)}{\sum_{k=1}^m \exp(\mathbf{x}'_i \boldsymbol{\beta}_k)} \quad (1)$$

The work of McFadden (1974) shows that we can express (1) in latent variable form as follows. Let

$$U_{ij} = \mathbf{x}'_i \boldsymbol{\beta}_j + \epsilon_{ij} \quad j = 1, \dots, m \quad (2)$$

where the  $\epsilon_{ij}$  are independent and have type I extreme-value distribution. By McFadden (1974),  $M_i = j$  if and only if  $U_{ij} > U_{ik}$  for all  $k \neq j$ . See Amemiya (1985) and Maddala (1991) for a proof. As with any multinomial model, parameter identification is achieved by taking one of  $\beta_j$  as zero.

In the MNL framework under which the latent variables depend on covariate values that vary across individuals but not across alternatives, the assumption of independent and identically distributed  $\epsilon_{ij}$  for  $j = 1, \dots, m$  implies that the ratio of the probability of outcome  $j$  to the probability of some other outcome  $k$  is independent of every other alternative  $l, l = 1, \dots, m, l \neq j, k$ . This property is known as the Independence of Irrelevant Alternatives (IIA). Valdez and Frees (2005) use a single covariate, gross premiums, which varies across individuals but not across different outcomes. Their results suggest that higher levels of premium are associated with higher probabilities of observing each of the remaining six claim-type combinations relative to the base outcome of  $M = 2$ , but only two of these coefficient estimates are precisely estimated at the traditional levels of significance. In the context of the IIA assumption, it is not difficult to envisage scenarios in which the stochastic component associated with observing a particular claim-type combination may in fact be correlated with that of another combination. In these scenarios, the MNL is inappropriate.

Often one’s preferred probabilistic choice model stems not only from considerations which pertain largely to the type of data one has, but also those of tractability and flexibility. In relation to the latter two considerations, McFadden (1981, p. 217) comments that, for a given parameterization of the chosen probabilistic choice model, there should be “sufficient flexibility to capture patterns of substitution between alternatives” and the chosen parameterization must also be computationally tractable. Thus, in the present context of modeling the claim-type combination which manifests as a discrete, and unordered, categorical variable, these two considerations naturally point towards the use of the multinomial probit model, hereafter referred to as MNP, as a natural alternative to the MNL model. The formulation of the MNP model is similar to that of the MNL under the random utility framework, with the exception that the unobserved disturbances of each outcome, for  $m$  mutually exclusive and exhaustive outcomes in the choice set, now have a multivariate Normal distribution with a zero mean vector and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1J} \\ \vdots & \ddots & \vdots \\ \sigma_{J1} & \cdots & \sigma_J^2 \end{bmatrix}, \quad (3)$$

instead of being independently and identically distributed according to the type I extreme value distribution under the MNL model. Here,  $\sigma_{sj}$  is the correlation between  $\epsilon_s$  of outcome  $s$  and  $\epsilon_j$  of outcome  $j$ ; that is, the correlation between the unobservables influencing the utilities of outcomes  $s$  and  $j$ . Formally, when expressed in utility differences, there are at most  $m(m-1)/2$  free parameters in the covariance matrix which can be estimated (Daganzo, 1979). Thus, for a trinomial probit case where  $m = 3$ , for example, there are at most three elements of the covariance matrix which we can estimate; these elements correspond to the lower triangular (or upper triangular) off-diagonal elements in  $\boldsymbol{\Sigma}$ , where the

variances on the diagonal are normalized to one. It is clear from this formulation, then, that the MNP relaxes the stringent IIA property inherent in MNL and is therefore more flexible in its specification of the covariance matrix. However, the viability of the MNP as a framework for modeling multinomial choice has received much attention in the literature. Juxtaposed against this flexibility are issues of tractability in the MNP model. Weeks (1997) brings to light the trade-off between the tractability and flexibility considerations alluded to by McFadden (1981). Here, tractability considerations pertain to the evaluation of high dimensional Normal integrals at each iteration of the optimization procedure, and is particularly restrictive for a large number of alternatives.<sup>1</sup> However, with the development of a computationally practical approach of simulated method of moments estimator, along with convenient reparameterization methods, this issue is now of a much lesser concern. In addition, the key papers of Bunch (1991), Bolduc (1992) and Keane (1992) stress yet another practical problem with estimating the MNP; not all parameters in the MNP may be identified and thus estimated in the MNP even if the data are well-behaved. More specifically, parameter identification is, in the words of Keane (1992, p. 193), “tenuous” or “fragile” in the absence of exclusion restrictions. What this means is that the practical estimation of the MNP requires that certain covariates do not affect the utility levels of certain outcomes, despite there being no requirement for such restrictions under *formal* identification of MNP models. The implication here is that estimation of the MNP model is likely to be problematic if the outcomes are being explained by characteristics which vary for each individual but not across the different outcomes. The characteristics which we propose to use vary for each policyholder, but not across different claim-type combinations. These practical considerations form the motivation for using the multivariate probit model to estimate the claim-type component. As shall be seen, our choice of this model is also more parsimonious in its application to the insurance portfolio considered in the paper.

The multivariate probit model has never been considered in the context of modeling motor insurance claims. In particular, considerations of an additional claim-type component of the aggregate loss distribution have been very limited, with Valdez and Frees (2005) providing a clear example of such with the use of the MNL discrete choice model. A related, but somewhat remote, example which considers the claim-type component is that of Pinquet (1998) who considers two types of claims, claims at fault and those not at fault. Pinquet (1998) models the frequency (claim count) component for each of these two types using a fixed effect Poisson model with covariates. In the general model with  $q$  types of claims, the fixed effect is assumed to be time-invariant for each policyholder  $i$  and each type of claim  $j$  for  $j \in q$  (Pinquet, 1998, p. 208). Thus, for a portfolio of  $p$  insurance policies at time  $t$ , the number of claims,  $N$ , associated with each claim-type  $j$  is a random variable which follows a Poisson distribution with mean parameter

$$\lambda_j^{it} = \exp(\boldsymbol{\theta}'_j \mathbf{x}_j^{it} + u_j^i)$$

specified as a function of “rating” factors observed at the fleet or individual vehicle level, and

$$N_j^{it} \sim \text{Poisson}(\lambda_j^{it}), \quad i = 1, \dots, p \quad j = 1, \dots, q \quad t = 1, \dots, T,$$

where  $\mathbf{x}_j^{it}$  and  $\boldsymbol{\theta}_j$  are vectors of covariates and parameters, respectively, which differ for each claim-type  $j$ , and  $u_j^i$  is the time-invariant fixed effect which accounts for the heterogeneity in the distribution. For  $q$  claim-types then, the associated claims frequency is modeled from a  $q$ -variate Poisson distribution (Pinquet, 1998, p. 208).

The approach taken in this article to model the conditional claim-type component is different from Valdez and Frees (2005) because we do not assume independence of the disturbances arising from each claim-type. It is also different from Pinquet (1998) because we focus on modeling the joint probability of observing a particular claim-type combination, conditional on there being at least one claim. Furthermore, in this paper, we explore the modeling of the multivariate claim types arising from automobile insurance policies. However, there are other possible situations and other possible contexts for which

<sup>1</sup>Some authors consider more than three or four alternatives as “large”; see Maddala (1991, p. 63), for example.

the model formulation proposed in this paper can be applied. Take, for example, the case of modeling operational risks which is receiving a lot of attention as of late. Some classify operational losses into several types including losses arising from process risk, people risk, system risk, business strategy risk and external environment risk. See Saunders, Boudoukh, and Allen (2003). The traditional (actuarial) approach of modeling operational losses is to consider the frequency of loss and the severity of the loss. If the company is able to classify losses according to the various types, it may detect possible dependencies on the types of losses and answer the question: how do losses from one type of risk affect the losses arising from the other types?

For the empirical investigation section of this paper, we consider the claims experience data derived from vehicle insurance portfolios of general insurance companies in Singapore. The primary source of this data is the General Insurance Association of Singapore, an organization consisting of most of the general insurers in Singapore. The observations are from each policyholder over a period of nine years: January 1993 until December 2001. To provide focus, we restrict our considerations to “fleet” policies from a single insurance company. These are policies issued to customers whose insurance covers more than a single vehicle. A typical situation of “fleet” policies is motor insurance coverage provided to a taxicab company, where several taxicabs are insured. The unit of observation in our analysis is therefore a registered vehicle insured under a fleet policy. We further break down these registered vehicles according to their exposure in each calendar year 1993 to 2001. Moreover, records from our databases show that when a claim payment is made, we can also identify the type of claim. For our data, there are three types: (1) claims for injury to a party other than the insured, (2) claims for damages to the insured including injury, property damage, fire and theft, and (3) claims for property damage to a party other than the insured.

This paper is organized as follows. Section 2 introduces the theoretical foundations and motivations of the multivariate probit model in analyzing potentially correlated multivariate outcomes. Here, we also provide a summary of a selection of literature which has used the multivariate probit model. Section 3 discusses the data used in the analysis and identifies its stylized features which motivates our preferred model of choice. We estimate a multivariate probit model for claim-type and discuss the important implications of the results. Section 4 provides the numerical results of our simulation which substantiate the conclusions reached in Section 5.

## 2 The formulation of the multivariate probit model

This section details the specification of the multivariate probit model, hereafter referred to as MVP, that is used to fit the distribution of different claim-types. To reiterate, conditional on there being at least one claim, we observe any  $2^J - 1$  combinations of the  $J$  different claim-types. We begin by first defining the notation consistent with that used in the introduction. Let  $I_j^o$  denote the underlying latent response associated with the  $j$ th type of claim, for  $j = 1, \dots, J$ , and  $I_j$  denote the binary response outcome associated with the same type. Using the indicator function,  $I_j$  is equal to one if there is a claim with respect to the  $j$ th type, and zero otherwise. Therefore, our MVP may be specified as a linear combination of a deterministic and stochastic components as follows:

$$\begin{aligned} I_1^o &= \mathbf{x}'\boldsymbol{\beta}_1 + \epsilon_1, & \text{for } I_1 &= \mathbb{I}_{\{I_1^o > 0\}} \\ I_2^o &= \mathbf{x}'\boldsymbol{\beta}_2 + \epsilon_2, & \text{for } I_2 &= \mathbb{I}_{\{I_2^o > 0\}} \\ &\vdots & &\vdots \\ I_J^o &= \mathbf{x}'\boldsymbol{\beta}_J + \epsilon_J, & \text{for } I_J &= \mathbb{I}_{\{I_J^o > 0\}} \end{aligned} \tag{4}$$

where  $\mathbf{x} = (1, x_1, \dots, x_p)'$  is a vector of  $p$  covariates which do not differ for each claim-type (the deterministic component) and  $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jp})'$  is a corresponding vector of parameters, including an intercept, which we seek to estimate. Note that the observation subscript  $i$  has been suppressed for notational convenience. The stochastic component,  $\epsilon_j$ , may be thought of as consisting of those

unobservable factors which explain the marginal probability of making a type  $j$  claim. Each  $\epsilon_j$  is drawn from a  $J$ -variate Normal distribution with zero conditional mean and variance normalized to unity (for reasons of parameter identifiability), where  $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$ , and the covariance matrix  $\Sigma$  is given by

$$\Sigma = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1J} \\ \rho_{21} & 1 & \cdots & \rho_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{J1} & \rho_{J2} & \cdots & 1 \end{bmatrix}. \quad (5)$$

Of particular interest are the off-diagonal elements in the covariance matrix,  $\rho_{sj}$ , which represents the unobserved correlation between the stochastic component of the  $s$ th and the  $j$ th types of claim. Moreover, because of symmetry in covariances, we necessarily have  $\rho_{sj} = \rho_{js}$ . As we saw previously, this covariance matrix is similar to that of the MNP, except the variances here are normalized to unity. We stress that our motivation for the joint estimation of correlated claim-types is not from the potential gain in efficiency, but of the ability to estimate the joint probabilities of the outcomes.

Note that in this formulation of the MVP model, we can derive marginal probabilities directly. For instance, the marginal probability of observing the  $j$ th type of claim can be expressed as

$$\Pr(I_j = 1) = \Phi(\mathbf{x}'\beta_j), \quad \text{for } j = 1, \dots, J \quad (6)$$

where  $\Phi(\cdot)$  denotes the cumulative distribution function of the standard Normal. Moreover, the joint probability of observing all possible types of claim comes from a  $J$ -variate standard Normal distribution

$$\Pr(I_1 = 1, \dots, I_J = 1) = \Phi_J(\mathbf{x}'\beta_1, \dots, \mathbf{x}'\beta_J; \Sigma), \quad (7)$$

where  $\Sigma$  is the covariance matrix.

Multivariate correlated binary observations arise in numerous contexts. An oft-cited example in the literature on animal studies is that of the ‘‘litter effect’’, where there is a greater tendency of likeness in individual responses within a litter relative to responses across different litters. Another example is that of the responses of different and separable physiological systems of an organism to the exposure of stimuli. It is perceived that the biological response of one physiological system to an injection of stimuli may be correlated with that of another physiological system. These examples point towards a central issue; that in an analysis of correlated quantal response data, one must account for the correlation structure between different levels of response if, *a priori*, there is a perceived possibility that these responses may in fact be correlated. This was the motivating theme behind the development of the multivariate probit model.

The seminal paper by Ashford and Sowden (1970) marks the development of the multivariate probit model. The authors generalized the univariate probit model for binary responses in consideration of a multi-level, vector-valued, response structure to different physiological systems of an organism. The quantal response of each system is manifested as an underlying continuous latent variable which is discretized subject to a threshold specification. Since this seminal paper, other works have applied the MVP in various contexts; see, for example, the works of Gibbons and Wilcox-Gök (1998) and Balia and Jones (2004).

### 3 Empirical investigation

#### 3.1 Data characteristics

The data used in our empirical investigation to illustrate fitting the multivariate probit models has been sourced from the General Insurance Association of Singapore (GIA), an organization of all general insurers in the country, whose primary objectives are, among others, to ‘‘foster public confidence in

and respect for the insurance industry” and to “establish a sound insurance structure and promotion of greater efficiency within the industry”. The motor insurance industry is the single largest class of insurance in the country, comprising approximately a third of the market share of the entire general insurance market. Each motor vehicle must have a valid insurance policy for it to be legally operated, where the minimum required is that of coverage against third party personal injury. Moreover, there are three major types of coverage available, which consist of third party, third party fire and theft, and comprehensive.

To give a general overview of the size of the data, the complete data set consists of over five million records from some forty-nine insurance companies. Each record is an observation corresponding to the claims experience of a registered motor vehicle in a given calendar year. The data spans over a period of ten calendar years, starting from 1 January 1993 until 31 December 2002. The length of a typical motor vehicle insurance policy is one year. Because insurance policies are very infrequently purchased at the start of the year, a peculiar feature of this data set is the treatment of the length of coverage as a fraction of a calendar year; this is known as ‘exposure’, and is the period (as a fraction of a calendar year) during which a policyholder had insurance coverage.

To provide focus, we restrict our consideration to the claims experience of “fleet” policies of one randomly-selected general insurer in our dataset. Fleet policies are those under which several motor vehicles are insured in a single contract. A typical example is that of an insurance policy for a taxicab company. See Valdez and Frees (2005, p. 3). The trivariate binary response vector consists of the following three types of claims:

- (1) claims for personal injury to a party other than the insured (**Injury**);
- (2) claims for damages to the insured, including personal injury, property damage, and fire and theft (**Own**); and
- (3) claims for property damage to a party other than the insured (**Property**).

All possible claim-type combinations that we observe from this dataset are illustrated in Table 1. Furthermore, the probit model is now a trivariate probit model which we hereafter refer to as the TVP model.

**Table 1:** Sample space of the conditional claim-type variable  $M$ .

$M$	1	2	3	4	5	6	7
Binary Triplet	(1, 0, 0)	(0, 1, 0)	(0, 0, 1)	(1, 1, 0)	(1, 0, 1)	(0, 1, 1)	(1, 1, 1)

Although the data consist of observations that included a vector of driver, policy and vehicle-specific characteristics, we find that in our model investigation, only the level of gross premiums, adjusted for the length of exposure, provides the single most important covariate. Table 2 provides summary statistics of premiums by claim-type combinations. From Table 2, we observe that, generally, higher average premiums are associated with multiple claim-types. In particular, the average premium associated with personal injury to a third party ( $M = 1$ ) and that associated with an insureds own injuries and damages ( $M = 2$ ) are higher than the average premium associated with third party property damage ( $M = 3$ ). Those who claim with respect to all three types face the third highest average premium; this is an odd result which may be simply be attributable to random variation in the data. Policyholders who claim with respect to “Injury” and “Own” damages ( $M = 4$ ), face the highest average premium, which is followed by claims with respect to “Own” injuries only ( $M = 2$ ). An important characteristic of this distribution is that, in any binary triplet, a claim with respect to “Own” injuries and “Property” damages entails a higher average premium, relative to the other claim-types.

We can also deduce from Table 2 the sample joint and marginal probabilities. The probability of observing “Property” damage is the highest (82.42%) among all types of claim, with the other two

**Table 2:** Summary statistics for Premium sorted by Claim-Type.

$M$	Binary Triplet	Frequency	%	Mean	Std. Dev.	Minimum	Maximum
1	(1,0,0)	152	6.80%	0.743	0.516	0.015	4.253
2	(0,1,0)	211	9.44%	1.153	1.163	0.012	6.778
3	(0,0,1)	1,673	74.82%	0.719	0.513	0.007	6.688
4	(1,1,0)	30	1.34%	1.192	1.318	0.039	4.570
5	(1,0,1)	119	5.32%	0.856	0.841	0.006	5.026
6	(0,1,1)	33	1.48%	0.953	1.240	0.080	6.457
7	(1,1,1)	18	0.81%	1.132	0.949	0.023	4.045

The binary triplets represent the possible claim-type combinations. Each element in the triplet is a binary variable for a claim-type (Injury, Own, Property), which is equal to one if there was a claim with respect to that type and zero otherwise. Units are in thousands of Singaporean dollars.

**Table 3:** Sample unconditional and conditional probabilities

	“Injury”	“Own”	“Property”
Prob( $\cdot$ )	0.143	0.131	0.824
Prob( $\cdot$   I=1)	1.000	0.150	0.429
Prob( $\cdot$   O=1)	0.164	1.000	0.175
Prob( $\cdot$   P=1)	0.074	0.028	1.000
Prob( $\cdot$   I=1, O=1)	1.000	1.000	0.375
Prob( $\cdot$   I=1, P=1)	1.000	0.131	1.000
Prob( $\cdot$   O=1, P=1)	0.353	1.000	1.000

types with relatively the same marginal probabilities; the marginal probabilities of observing “Injury” is 14.27% while that of observing “Own” damages is 13.06%. The probability of observing all three types at once is at the rate of 0.81%. From the table, we also find that the sample probability of observing only “Injury” claims is 6.80%, only “Own” damages is 9.44% and only “Property” damages is 74.82% (the highest among all three types of claim).

Although these empirical joint and marginal probabilities provide interesting results, the sample conditional probabilities in Table 3 provide an interesting indication of the existence of possible dependence between the types of claim. Consider for example the case of “Property” damages. The unconditional probability of a claim with “Property” damages is 0.82. However, among all accidents with “Injury” claims, the sample probability with “Property” damages is 0.43; among all accidents with “Own” damages, the probability would be only 0.18. This means that the probability of a claim with “Property” damages could be substantially reduced if there is additional information that another type of claim would have occurred.

### 3.2 Estimation and discussion of results

Estimation was carried out in Stata<sup>®</sup>. To estimate the TVP model, we used the command `-mvprobit-` coded by Cappellari and Jenkins (2003); the MNL models were estimated by maximum likelihood using `-mlogit-`. The multivariate probit model was estimated using the method of simulated maximum likelihood (SML) using a smooth recursive simulator, known as the GHK simulator, to evaluate multivariate Normal probabilities. The specific details of this algorithm are omitted here, but see Train (2003, p. 126–37), Greene (2003, p. 932–33) and the references cited therein. Furthermore, the SML estimator is asymptotically consistent as the number of observations and the number of draws tend to infinity. Cappellari and Jenkins (2003) recommend that, so long as the number of draws,  $R$ , is greater than the square root of the sample size,  $\sqrt{N}$ , then parameter estimates obtained through `-mvprobit-` are robust to different initial seed values. We adopt this “rule of thumb” in our estimation.

Under the multivariate probit framework, the variances of the disturbances are normalized to unity. In the context of our insurance data, however, we conjecture that the variation in the probability of observing a particular outcome may in fact vary with different levels of premium. For example, there may be less variation in the probability of observing “extreme outcomes” for lower levels of premium, as well as for higher levels of premium, but a larger variation in the probability of outcomes which fall in-between. In the context of simple univariate non-linear models, such as probit and logit, heteroskedasticity causes parameter estimates to be inconsistent (Davidson and MacKinnon, 1984). A Lagrange multiplier (LM) test was proposed by Davidson and MacKinnon (1984) to test for heteroskedasticity of a known functional form; but the applicability of their testing procedure has not, to our knowledge, been verified in a multivariate context. As a preliminary analysis, we used the `HETTEST` procedure in `SHAZAM` to test for heteroskedasticity using the Davidson and MacKinnon (1984)  $LM_2$  test statistic.<sup>2</sup> We ran three separate univariate probit regression on the fleet policies for each claim type using Premium as the only covariate. In all three tests, the null hypothesis of homoskedasticity was strongly rejected in favour of heteroskedastic disturbances.<sup>3</sup> However, it is important to note that these tests do not point definitively towards the presence of heteroskedasticity, but may, instead, pick up some other form of misspecification (Greene, 2003, p. 681). Furthermore, computation of a robust covariance matrix in this particular setting is unclear and will be inappropriate if the heteroskedasticity is of an unknown form. Finally, after fitting MNL models to the policies, the Small and Hsiao (1985) test was used to test the hypothesis that the IIA property holds. In all of these tests, there were sufficient evidence at the 1% significance level leading to unequivocal rejections of the hypothesis that the IIA property holds.

The unit of observation in our analysis is a registered vehicle insured under a fleet policy. Tables 4 and 6 summarize the results of fitting our TVP and MNL models, respectively. As expected, Premium is an important predictor of claim-type; the coefficient estimates are both statistically and economically significant, as well as having the *a priori* expected sign. A somewhat surprising result is the negative marginal effect of Premium on the marginal probability of claiming with respect to “Property”, which we have defined as a claim with respect to third party property damage. To evaluate the marginal effect of Premium on each of the marginal probabilities, we calculate the linear prediction on each claim-type  $j$ ,  $\mathbf{x}'\hat{\beta}_j$ , and, using  $\partial \mathbb{E}(y_j|\mathbf{x})/\partial x_1 = \phi(\mathbf{x}'\hat{\beta}_j) \times \hat{\beta}_{j1}$ , we averaged out the marginal effects for each observation (Greene, 2003, p. 668). Here,  $\hat{\beta}_{j1}$  is the coefficient estimate on Premium from the claim-type  $j$  equation, for  $j = 1, 2, 3$ , and  $\phi(\cdot)$  is the probability density function of a standard normal distribution with zero mean and unit variance. The average return to Premium for each claim-type is reported in Table 5. These marginal effects suggest that, for every \$1,000 increase in Premium, the marginal probabilities of making an “Injury” and “Own” claim are increased by approximately 2.5% and 6.8%, respectively. For “Property” claims, however, a same \$1,000 increase in Premium is associated with a fall in its marginal probability by approximately 6.2%.

The estimated correlation coefficients,  $\hat{\rho}_{sj}$ , between each of the three claim-types are statistically and economically significant. Surprisingly, the correlations between the disturbances of the “Injury” and “Property” equations, and the “Own” and “Property” equations are negative. This suggests that the unobservable factors which increase the probability of claiming with respect to “Injury”, for example, actually reduce the probability of claiming with respect to “Property”; a similar interpretation also applies to the negative correlation between “Own” and “Property”. The positive correlation between the “Injury” and “Own” equations is intuitive. Here, unobservable factors which increase the probability of claiming with respect to “Injury” also increase the probability of claiming with respect to “Own”. One can think of latent driver characteristics, such as responsiveness or restlessness, as important factors which influence the probability. Furthermore, the likelihood ratio test for independence between the disturbances is strongly rejected, implying correlated binary responses between different claim-types.

<sup>2</sup>The code for the `HETTEST` procedure may be downloaded from <http://shazam.econ.ubc.ca/intro/logit3.htm>.

<sup>3</sup>The  $LM_2$  statistics from the two probit regressions of Injury on Premium, using fleet policies, were small relative to the critical value.

**Table 4:** TVP model fitted to the fleet policies.

	<b>Coefficient Estimate</b>	<b>Standard Error</b>	<i>z</i> -statistic	<i>p</i> -value
<b>Injury</b>				
Premium	0.109	0.0426	2.58	0.010
Intercept	-1.150	0.0485	-23.75	0.000
<b>Own</b>				
Premium	0.330	0.0409	7.98	0.000
Intercept	-1.397	0.0495	-28.08	0.000
<b>Property</b>				
Premium	-0.254	0.0379	-6.72	0.000
Intercept	1.189	0.0438	27.13	0.000
<b>Correlation Coefficient</b>				
$\hat{\rho}_{21}$	0.274	0.0373	7.62	0.000
$\hat{\rho}_{31}$	-0.615	0.0258	-23.89	0.000
$\hat{\rho}_{32}$	-0.844	0.0158	-54.14	0.000
$R = 100$	Log pseudolikelihood = -2265.3486, $n = 2236$			

† Likelihood Ratio Test  $H_0 : \rho_{21} = \rho_{31} = \rho_{32} = 0$ ,  $\chi^2_{(3)} = 1007.25$ ,  $p$ -value = 0.000;  $R$  is the number of pseudo-random draws. Premium is in \$'000s of Singaporean dollars.

**Table 5:** The average marginal effect of Premium on the marginal claim-type probabilities.

<b>Claim-Type</b>	<b>Marginal Effect</b>
Injury	0.025
Own	0.068
Property	-0.062

**Table 6:** MNL model fitted to the fleet policies.

$M$	Binary Triplet	Intercept			Slope		
		Coefficient Estimate	Standard Error	$p$ -value	Coefficient Estimate	Standard Error	$p$ -value
1	(1,0,0)	-2.460	0.137	0.000	0.085	0.145	0.558
2	(0,1,0)	-2.626	0.112	0.000	0.636	0.092	0.000
4	(1,1,0)	-4.605	0.262	0.000	0.661	0.165	0.000
5	(1,0,1)	-2.907	0.149	0.000	0.339	0.137	0.014
6	(0,1,1)	-4.305	0.287	0.000	0.467	0.240	0.052
7	(1,1,1)	-5.071	0.297	0.000	0.622	0.169	0.000
Log-pseudolikelihood = -2060.594, $n = 2236$ , Pseudo $R^2 = 0.0171$							

Note: The response variable is the conditional claim-type,  $M$ ; the single covariate is Premium (in \$'000s of Singaporean dollars); and the omitted category is  $M = 3$ .

The MNL results are summarized under Table 6, where, like the TVP, a single covariate was used to explain the conditional claim-type,  $M$ . Here, the omitted category is a claim with respect to “Property” damage only ( $M = 3$ ). Again, we observe that Premium is an important predictor of claim-type; in all but one combination ( $M = 1$ ), the coefficient on Premium is statistically and economically significant, as well as having the *a priori* expected sign. The coefficient estimate on Premium is the marginal effect of Premium on the log of the ratio of probabilities; therefore, one can exponentiate the index function to produce a probability of a given outcome relative to the omitted category. To illustrate, a 100 Singaporean dollar increase in Premium is associated with an increase in the probability of making an “Injury” only claim ( $M = 1$ ) by approximately 1.1% ( $= (e^{0.109/10} - 1) \times 100\%$ ), relative to the probability of making a “Property” only claim ( $M = 3$ ). In contrast, for two policies which differ by 100 Singaporean dollars, the more expensive policy is 6.42% ( $= (e^{0.622/10} - 1) \times 100\%$ ) more likely to claim with respect to all three types of claim ( $M = 7$ ), relative to the probability of claiming for “Property” damages only. The largest coefficient estimate on Premium is associated with claiming with respect “Injury” and “Own” ( $M = 4$ ). Here, a 100 Singaporean dollar increase in Premium is expected to increase the probability of claiming for “Injury” and “Own” by approximately 6.83% ( $= (e^{0.661/10} - 1) \times 100\%$ ), relative to the omitted category.

### 3.3 Comparing the results of the MNL and the MVP models

The individual and joint statistical significance of the correlation coefficients from the TVP, using the fleet policies, is supporting evidence for correlated latent responses amongst different claim-types. As mentioned before, the MNL model assumes that the latent responses are independent between different claim-type combinations. The implication here is that the ratio of the probabilities of observing two different claim-types is not affected by the presence of other claim-type combinations in the same choice set (the IIA property). We use predicted probabilities of each possible outcome as the basis for comparing the MNL to the TVP models in the presence of cross-correlations in the latent responses. These probabilities are conditional on there being a claim.

The predicted probabilities of each outcome from the TVP model are computed from a simulation exercise, which draws from a trivariate Normal distribution using pseudo-random sequences derived from a standard Uniform density; see Cappellari and Jenkins (2006). To illustrate the complexity of evaluating these probabilities, suppose we wish to compute the predicted probability of observing a claim-type combination of the form (1,0,1); here, there is a claim with respect to “Injury” and “Property”, but none for “Own” damages. Now, if we let  $w_j = q_j \mathbf{x}' \hat{\beta}_j$  and  $q_j = 2I_j - 1$  for  $j = 1, 2, 3$ , then the

integral to be evaluated is

$$\begin{aligned} \Pr(I_1 = 1, I_2 = 0, I_3 = 1 | w_1, w_2, w_3) &= \Phi_3(w_1, w_2, w_3; \rho_{21}, \rho_{31}, \rho_{32}) \\ &= \int_{-\infty}^{\mathbf{x}'\hat{\beta}_1} \int_{-\infty}^{-\mathbf{x}'\hat{\beta}_2} \int_{-\infty}^{\mathbf{x}'\hat{\beta}_3} \phi_3(\epsilon_1, \epsilon_2, \epsilon_3; \rho_{21}, \rho_{31}, \rho_{32}) d\epsilon_3 d\epsilon_2 d\epsilon_1, \end{aligned} \quad (8)$$

where  $\Phi_3(\cdot)$  and  $\phi_3(\cdot)$  denote a trivariate Normal distribution function and probability density function, respectively, and the upper support is the linear prediction or index value corresponding to each claim-type. Note that the sign on each upper support depends on whether the observed binary outcome is one or zero, so that it is positive if the observed outcome is one, and negative if the observed outcome is zero. This specification follows the parameterization of the bivariate case considered in Greene (2003, p. 710). Table 7 summarizes the average predicted probabilities from the TVP and the MNL models for each possible outcome, using the fleet policies. For each observation in the relevant subsample, we computed the predicted probability of observing the given outcome, and then took the average of the probabilities for the same outcome.

Furthermore, we observe that the probability of claiming with respect to ‘‘Property’’ damage only ( $M = 3$ ) is the highest, given that there is a claim. This result is not surprising, since it is by far the most frequently observed outcome in our sample; see Table 2. The two last columns in the table measure the difference between the two sets of predicted probabilities and the proportionate differences, respectively. Here, we define

$$\Delta_M = \hat{p}_{1M} - \hat{p}_{2M} \quad (9)$$

to be the difference between the predicted probabilities of the TVP and MNL models for outcome  $M$ . These differences are small for outcomes  $M = \{2, 7\}$  but relatively larger for outcome  $M = 4$ . For the proportionate differences column, the largest divergences are observed in outcomes  $M = \{1, 7\}$ . Figures 1a through to 1g provide graphical representations of the distribution of the differences in predicted probabilities for each outcome across Premium. For outcome  $M = 1$ , there is apparent fall in the differences between the probabilities for higher amounts of premium. In contrast, for outcomes  $M = \{2, 4, 6, 7\}$ , the differences increase over larger amounts of premium. There is no such apparent pattern in the differences in probabilities for outcomes  $M = \{3, 5\}$ ; for outcome  $M = 3$ , the differences fall for premium amounts up to \$2,500, but increase thereafter. Similarly, for outcome  $M = 5$ , differences fall up to premium amounts close to \$1,400, but increase thereafter. The least probable outcome is a claim with respect to all three types of damages ( $M = 7$ ). A peculiar result is that in spite of its slightly higher observed frequency relative to outcome  $M = 6$ , its predicted probability of occurrence is substantially smaller. Their predicted probabilities are substantially over-estimated, on average, by the MNL model.

**Table 7:** Average predicted probabilities for each outcome, using the *fleet* policies.

$M$	Outcome	TVP ( $\hat{p}_{1M}$ )	MNL ( $\hat{p}_{2M}$ )	$\Delta_M$	$\left(\frac{\hat{p}_{1M} - \hat{p}_{2M}}{\hat{p}_{1M}}\right) \%$
1	(1,0,0)	0.037	0.068	-0.031	-83.78
2	(0,1,0)	0.079	0.094	-0.015	-18.99
3	(0,0,1)	0.734	0.748	-0.014	-1.91
4	(1,1,0)	0.049	0.013	0.036	73.47
5	(1,0,1)	0.071	0.053	0.018	25.35
6	(0,1,1)	0.039	0.015	0.024	61.54
7	(1,1,1)	0.001	0.008	-0.007	-700.00

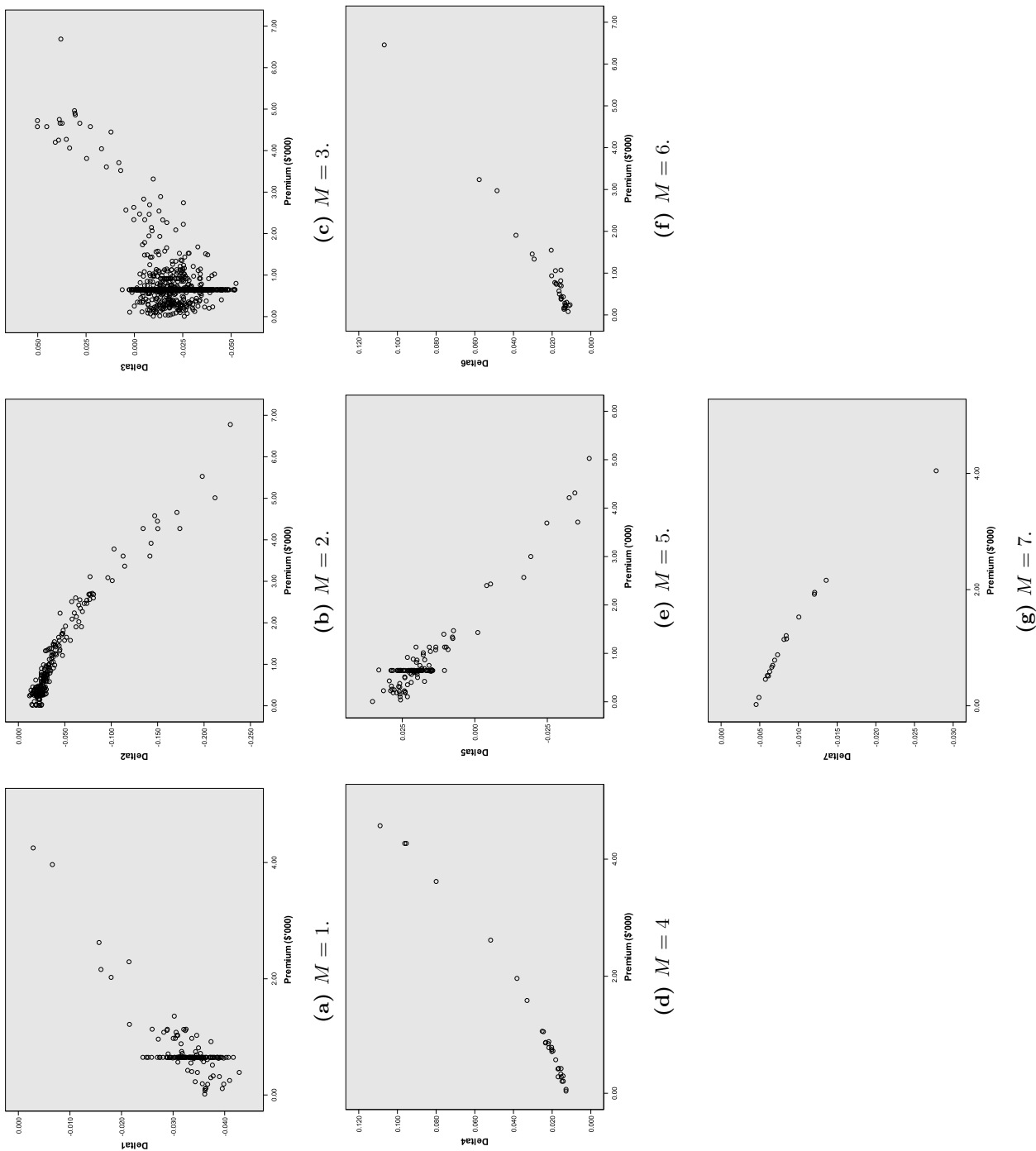


Figure 1: Differences in predicted probabilities for each outcome across Premium.

## 4 Monte Carlo simulation

The estimation results reported in the previous section suggest that fitting the MNL model to correlated binary response data results in considerable divergences in the predicted probabilities relative to those produced by our TVP model. In particular, we saw that these divergences varied systematically over premiums for various claim-type combinations; for outcomes  $M = \{2, 4, 6, 7\}$ , the divergences tend to increase over higher amounts of premium, but not clearly so for the remaining outcomes. Moreover, the proportionate differences between the predicted probabilities are substantially higher for claim-type combinations in which there was more than one type of claim made, with the exception of outcome  $M = 1$ ; see the last column of Table 7. The aim of this section is to now substantiate these findings. Through a controlled experimental design, we carry out a number of Monte Carlo simulations, whereby we fit the MNL and TVP models of the specification found in §3 to multivariate response outcomes drawn from a randomly generated trivariate Normal distribution, using previously estimated coefficients as the true parameters. In addition, we investigate the extent of these divergences under various experimental values of the correlation coefficients in the trivariate Normal covariance matrix, which we denote, respectively, by  $\{\boldsymbol{\Sigma}_F, \boldsymbol{\Sigma}_H, \boldsymbol{\Sigma}_Z\}$ . Using the standard deviation of these divergences as the standard error of the differences in the predictions, our results provide strong evidence in support of the TVP model over the MNL model when the binary responses are potentially correlated.

### 4.1 The experimental design

We extract the vector of premium values from fleet policies under the same insurer and use this as the single covariate, which is consistent with our TVP and MNL models fitted to fleet policies in §3. There are 2,236 premium amounts in this vector. For the purposes of our experiment, we assume that the underlying data generating process (DGP) for latent disturbances of each claim-type follows a trivariate Normal distribution with zero mean vector and covariance matrix  $\boldsymbol{\Sigma}$  as defined in equation (5). Our method of drawing from the multivariate Normal distribution follows that of Cappellari and Jenkins (2006, pp. 10–11). Here, we use coefficient estimates extracted from Table 4 as the *a priori* known true population parameters. For each claim-type  $j$ , the underlying model specification is given by

$$I_j^{\circ} = \mathbf{x}'\boldsymbol{\beta}_j + \epsilon_j, \quad \text{for } I_j = \mathbb{I}_{\{I_j^{\circ} > 0\}}, \quad j = 1, 2, 3 \quad (10)$$

where  $\mathbf{x} = (1, \text{Premium})'$  is a vector of covariates which varies across policyholders, and

$$\boldsymbol{\beta}_1 = (-1.150, 0.109)', \quad \boldsymbol{\beta}_2 = (-1.397, 0.330)', \quad \boldsymbol{\beta}_3 = (1.189, -0.254)' \quad (11)$$

are vectors of estimated parameters which we treat as the true population parameters of the underlying DGP. Finally, to complete the specification, we have  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_F)$ , where the covariance matrix is

$$\boldsymbol{\Sigma}_F = \begin{bmatrix} 1 & & \\ 0.274 & 1 & \\ -0.615 & -0.844 & 1 \end{bmatrix}, \quad (12)$$

and the subscript  $F$  denotes that the specification is of the *full* correlations. Note that the upper triangular elements have been omitted due to symmetry. Generating the trivariate realizations  $I_j$  is then straightforward by (10). In each replication, we have a  $2,236 \times 3$  matrix of random disturbances drawn from a trivariate Normal distribution. We use this sample to estimate the TVP and MNL models. In summary, our methodology can be decomposed into three steps:

- (1) Using the trivariate probit parameter estimates from fleet policies as the “true” parameters, generate a simulated trivariate normal distribution for the underlying latent responses for each claim-type  $j$ , with correlation structure defined by  $\hat{\rho}_{21}, \hat{\rho}_{31}$  and  $\hat{\rho}_{32}$ . The observed response of each binary outcome,  $y_j$ , is the realisation of the underlying latent response for claim-type  $j$ . For each observation, we observe a claim-type combination in the form of a binary triplet and a corresponding premium amount.

- (2) Estimate the TVP and MNL models using the random sample generated in step (1), with 2,236 premium values as the single covariate. Note that observations for which the outcome is  $M = 0$  have been excluded from estimation.
- (3) Repeat steps (1) and (2) 100 times.

Figure 2 shows the distribution of the MNL coefficient estimates for each outcome  $M$  over the 100 replications. In each replication,  $M = 3$  was treated as the base outcome, so that the coefficient estimate on Premium measures the return to Premium on the probability of observing outcomes  $M = \{1, 2, 4, 5, 6, 7\}$  relative to  $M = 3$ . Figures 2a through to 2e show that estimates for the intercept and Premium coefficient converge to approximately to similar values over each replication. For outcome  $M = 7$ , Figure 2f shows a more sporadic pattern in the coefficient estimates over the replications. Similarly, Figure 3 shows the distribution of the TVP estimates for each claim-type  $j$  over the 100 replications. Not surprisingly, these estimates are consistent with the true population parameters as specified under our experimental design.

## 4.2 Predicted probabilities

In each replication  $r$ , we simulated the trivariate Normal probabilities for each observation; these are stored under a new variable `prob‘r’` in the simulated data set. These are the predicted probabilities corresponding to the TVP model. For each outcome  $M$ , we then averaged the predicted probabilities over Premium. This results in seven mean predicted probabilities, with each corresponding to an outcome. The same methodology is applied to the MNL predicted probabilities, and these are stored in new variables `M‘i’-1` through to `M‘i’-7` in the data set. The differences between these two sets of predicted probabilities for each outcome are computed as follows. Define a variable  $\Delta_M^{(r)}$  to be the difference between the mean TVP ( $\hat{p}_{1M}$ ) and MNL ( $\hat{p}_{2M}$ ) predicted probabilities for outcome  $M$ , so that at each replication  $r$ , we have

$$\Delta_M^{(r)} = \hat{p}_{1M}^{(r)} - \hat{p}_{2M}^{(r)}, \quad r = 1, \dots, 100. \quad (13)$$

This same process is then repeated over 100 times so that for each outcome  $M$ , there are 100  $\Delta_M$ 's. Our aim is to test the null hypothesis that for the same outcome the mean predicted produced by the TVP model is not different to that of the MNL model. This is equivalent to testing the null hypothesis that  $\mu_M = 0$ , where  $\mu_M$  is the population mean of the difference in the predictions. The alternative hypothesis is that the means of the differences are nonzero ( $\mu_M \neq 0$ ), so that correlated binary responses should determine the model. Here, we appeal to the central limit theorem (CLT) so that the difference in mean predicted probabilities ( $\Delta$ ), for each outcome, over the 100 replications is approximately standard Normal. The implication here is that even if the individual replications are not Normally distributed, the mean of the replications will be Normally distributed by the CLT. As an aside, Figure 4 shows that the replications for outcomes one through to six are normally distributed, with the exception of  $\Delta_7$ . Quantile-quantile (QQ-) plots are used as a graphic representation of normality, so that if the observations are drawn from a normal distribution, then they should not deviate from the 45° reference line. For  $\Delta_7$ , there are clear deviations from this reference line, but its non-normality does not affect our inferences.

Finally, the difference in sample means<sup>4</sup> is still itself a sample mean, so that the relevant test statistic for outcome  $M$  is

$$\frac{\Delta_M - \mu_M}{\sigma_{\Delta_M}/\sqrt{n}} \sim \mathcal{N}(0, 1), \quad (14)$$

where  $\Delta_M$  is the difference between the mean predicted probabilities of the TVP and MNL models,  $n$  is the number of simulations run, and  $\mu_M$  is the population mean. Equation (14) holds only if the  $\Delta$ 's are

<sup>4</sup>That is, the difference between the mean TVP and MNL predicted probabilities.

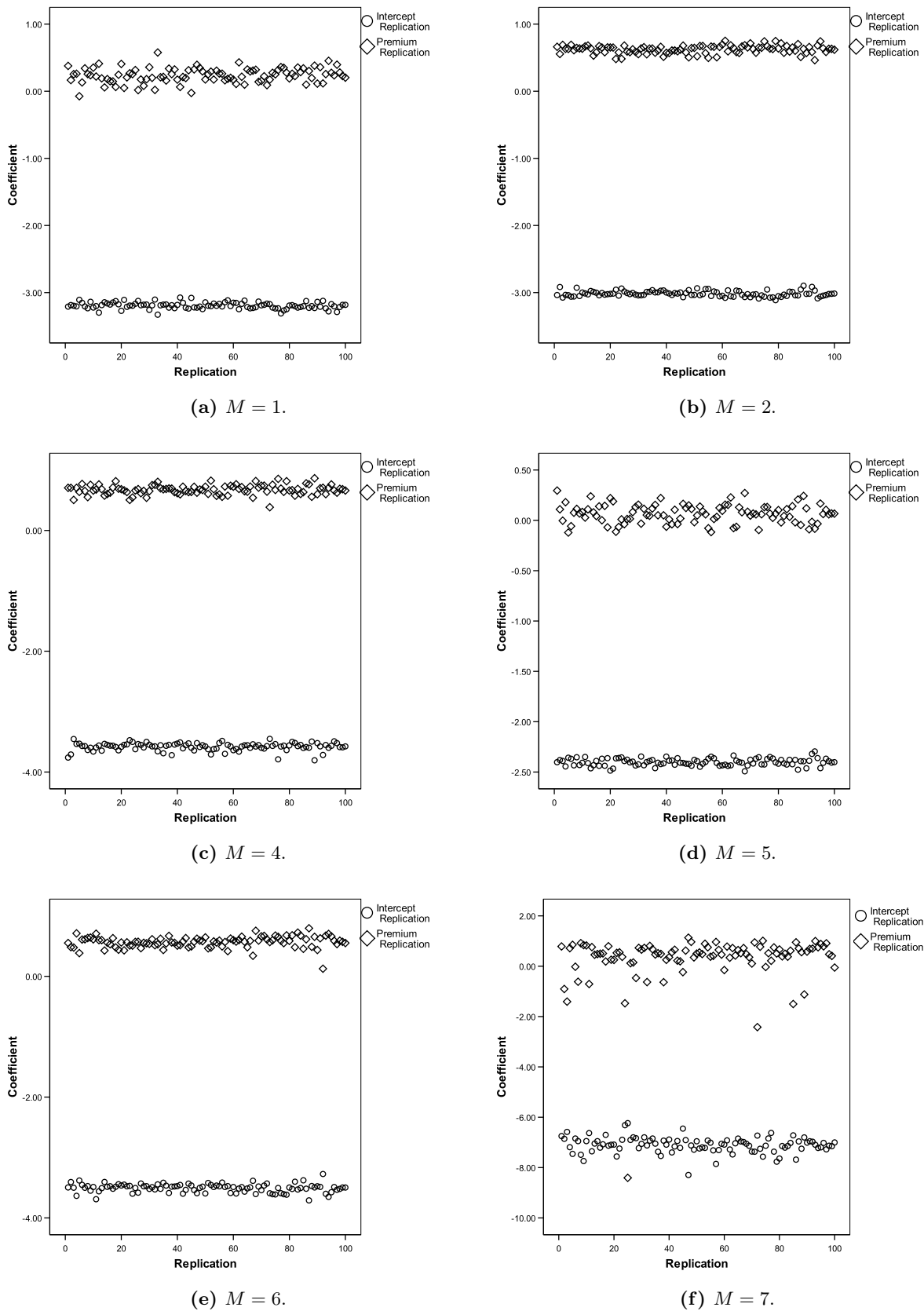


Figure 2: MNL coefficient estimates relative to the base outcome  $M = 3$  over 100 simulations.

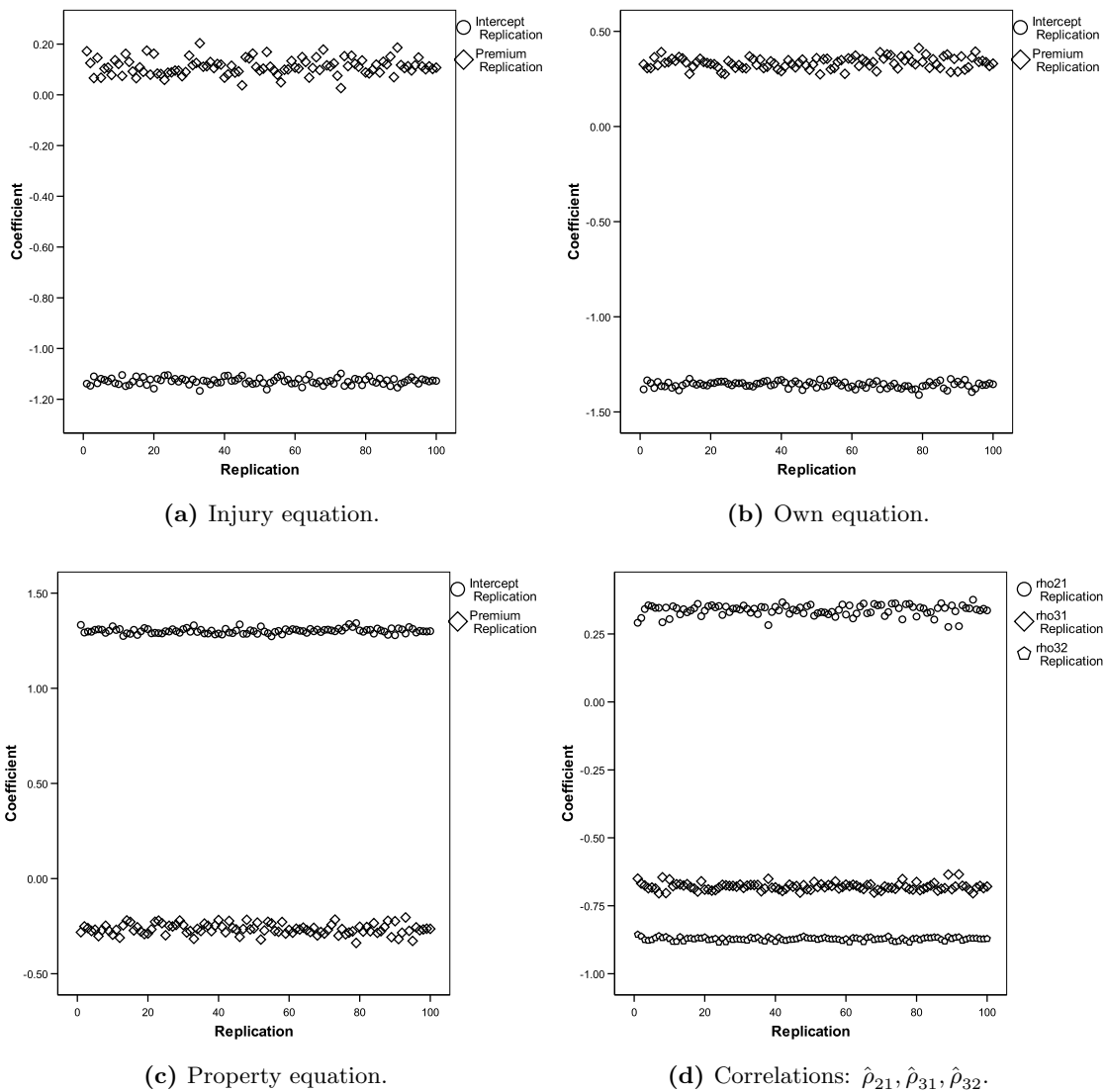


Figure 3: TVP coefficient estimates over 100 simulations.

drawn from independent replications. Recall our experimental design where our sample consists of 2,236 observations in each replication. In each subsequent replication, the premium amounts will be the same, so such that the  $\Delta$ 's can no longer be regarded as drawn from independent samples. To account for this dependence in the samples, we estimate the variance of  $\Delta_M$ ,  $\text{Var}(\Delta_M)$ , whilst allowing for covariances between replications. The derivation of  $\text{Var}(\Delta_M)$  is outlined below.

Assume that for any outcome  $M$ , the variance of  $\Delta_M^{(r)}$ , for each replication  $r$ , and the covariance between  $\Delta_M^{(r)}$  and  $\Delta_M^{(s)}$ , for replications  $r$  and  $s$ , are equal. Then, the variance of  $\Delta_M$  is derived by the following:

$$\begin{aligned} \text{Var}(\Delta_M) &= \frac{1}{n^2} \text{Var} \left[ \sum_{r=1}^n \Delta_M^{(r)} \right] \\ &= \frac{1}{n^2} \left[ n \text{Var} \left( \Delta_M^{(r)} \right) + \binom{n}{2} \text{Cov} \left( \Delta_M^{(r)}, \Delta_M^{(s)} \right) \right] \\ &= \frac{1}{n} \left[ \text{Var} \left( \Delta_M^{(r)} \right) + \left( \frac{n-1}{2} \right) \text{Cov} \left( \Delta_M^{(r)}, \Delta_M^{(s)} \right) \right], \end{aligned} \tag{15}$$

where  $\binom{n}{2} = \frac{n!}{2!(n-2)!}$  denotes the number of combinations of choosing any two different replications from  $n$  simulations. Both the variance and covariance terms in (15) can be estimated from the generated data where

$$\text{Var} \left[ \Delta_M^{(r)} \right] = \mathbb{E} \left[ \left( \Delta_M^{(r)} \right)^2 \right] - \mathbb{E} \left[ \Delta_M^{(r)} \right]^2 \tag{16}$$

is the variance of  $\Delta_M$  for each replication  $r$ , and

$$\text{Cov} \left[ \Delta_M^{(r)}, \Delta_M^{(s)} \right] = \mathbb{E} \left[ \Delta_M^{(r)} \cdot \Delta_M^{(s)} \right] - \mathbb{E} \left[ \Delta_M^{(r)} \right] \cdot \mathbb{E} \left[ \Delta_M^{(s)} \right], \tag{17}$$

is the covariance between replications  $r$  and  $s$ , for any  $r \neq s$ .

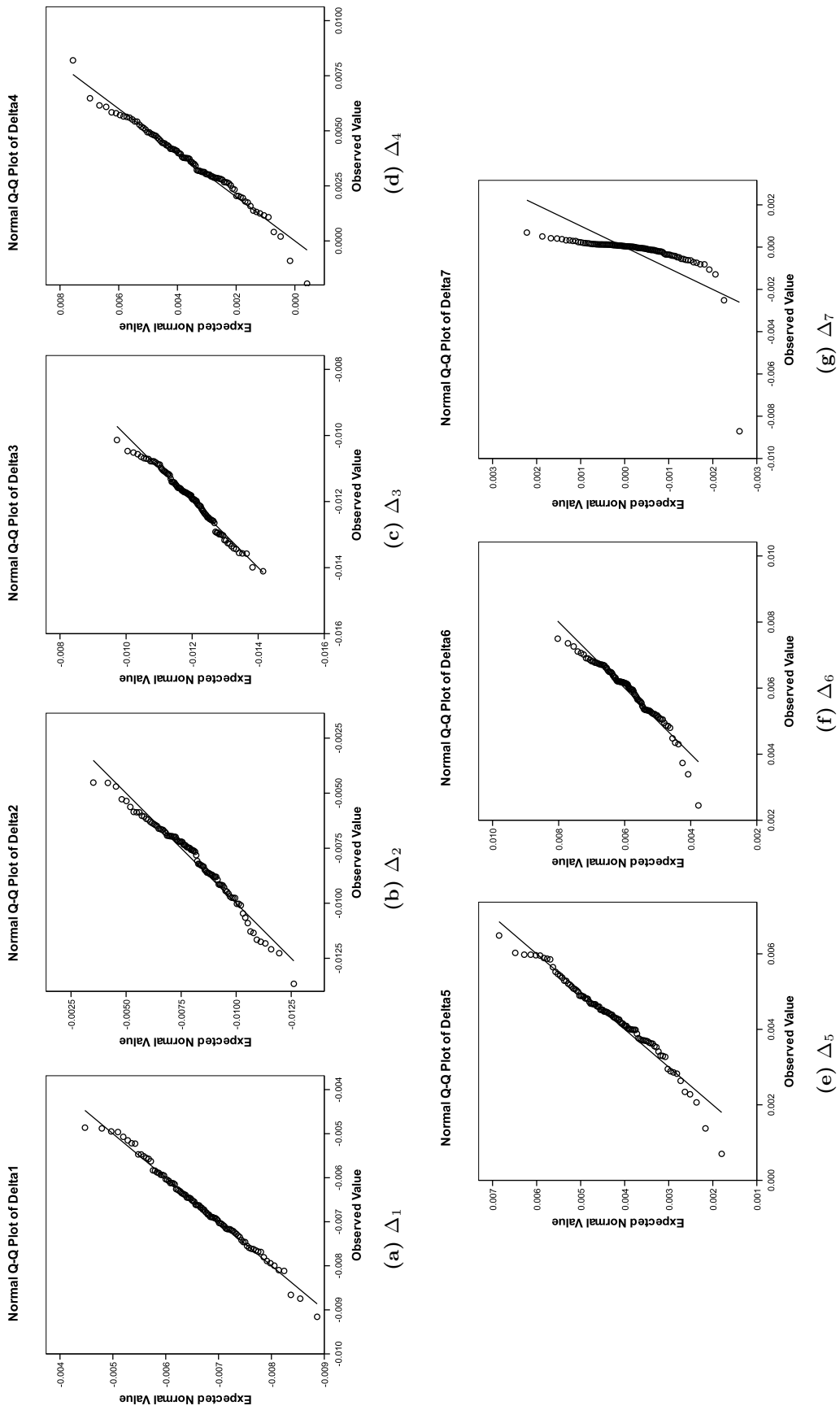


Figure 4: Quantile-Quantile plots for each  $\Delta_M$ .

### 4.3 Results and discussion

The results of our simulation exercise are summarized in Table 8 below. The square root of the estimated variance is the standard error of the difference in mean predictions, and corresponds to taking the square root of the variance derived the previous subsection. These results collectively suggest that implications of correlated disturbances from a multivariate outcome setting should not be neglected; a finding of such correlation should drive one's decision to use a model which allows for flexibility in the specification of the covariance matrix for the disturbances as opposed to another which conveniently assumes independence. Again, this brings to light the aforementioned trade-off between flexibility and tractability; see §1 for a discussion. From our experimental design in which we are able to specify *a priori* correlations in the underlying DGP, the means of the MNL predicted probabilities of six from the seven outcomes are statistically different from those of the TVP model. Specifically, there is strong evidence against the null hypothesis of no difference for outcomes  $M = \{1, 2, 3, 5, 6\}$  at the 1% significance level; for outcomes  $M = \{4, 7\}$ , however, the evidence is less convincing. Moreover, these results further substantiate those in Table 7. For an actuary seeking to model the conditional claim-type, the implication is that the TVP model should be preferred over MNL when the outcomes are correlated.

**Table 8:** Simulation results for the standard deviation of  $\Delta_M$ , with *full correlations* specification.

$\Delta_M$	Mean	Variance	Standard Error	Test Statistic <sup>1</sup>	<i>p</i> -value <sup>2</sup>
$\Delta_1$	-0.00667	$9.52621 \times 10^{-8}$	0.00031	-21.52	0.000
$\Delta_2$	-0.00806	$6.08742 \times 10^{-6}$	0.00247	-3.26	0.002
$\Delta_3$	-0.01194	$5.64243 \times 10^{-7}$	0.00075	-15.92	0.000
$\Delta_4$	0.00357	$7.76953 \times 10^{-6}$	0.00279	1.28	0.200
$\Delta_5$	0.00433	$3.28590 \times 10^{-8}$	0.00018	24.06	0.000
$\Delta_6$	0.00590	$2.00991 \times 10^{-7}$	0.00045	13.11	0.000
$\Delta_7$	-0.00019	$1.07663 \times 10^{-8}$	0.00010	-1.90	0.058

<sup>1</sup> The test statistic is  $\Delta_M / \sqrt{\text{Var}(\Delta_M)} \sim \mathcal{N}(0, 1)$ .

<sup>2</sup> The *p*-value corresponds to the probability value of the relevant test statistic under a two-sided alternative.

### 4.4 Experimental correlations

A benefit of our experimental design is the ability to specify *a priori* the true population parameters of the underlying DGP. In the previous section we saw that under a specification of the covariance matrix in (12) with full correlations derived from our TVP results, differences between the TVP and MNL predicted probabilities were found to be statistically significant. In the discussion which follows, we experiment with two other specifications for the covariance matrix, namely, one where the correlations are halved and the other where they are set to zero (that is, the error disturbances in each claim-type equation are orthogonal). In both of these experiments, the simulation methodology is the same as before.

#### 4.4.1 Half correlations

Suppose that each of the correlations between the claim-type disturbances are halved, so that instead of the covariance matrix in (12), we have

$$\Sigma_H = \begin{bmatrix} 1 & & & \\ 0.137 & 1 & & \\ -0.308 & -0.422 & 1 & \\ & & & \ddots \end{bmatrix}. \quad (18)$$

**Table 9:** Simulation results for the standard deviation of  $\Delta_M$ , with *half correlations* specification.

$\Delta_M$	Mean	Variance	Standard Error	Test Statistic <sup>1</sup>	<i>p</i> -value <sup>2</sup>
$\Delta_1$	-0.00907	$6.92655 \times 10^{-8}$	0.00026	-34.47	0.000
$\Delta_2$	-0.01181	$5.01097 \times 10^{-6}$	0.00224	-5.28	0.000
$\Delta_3$	-0.00905	$3.73666 \times 10^{-7}$	0.00061	-14.80	0.000
$\Delta_4$	0.00860	$1.14378 \times 10^{-5}$	0.00338	2.54	0.010
$\Delta_5$	0.00657	$7.26321 \times 10^{-7}$	0.00085	7.71	0.000
$\Delta_6$	0.00711	$6.05012 \times 10^{-7}$	0.00078	9.14	0.000
$\Delta_7$	-0.00461	$7.13183 \times 10^{-8}$	0.00027	-17.26	0.000

<sup>1</sup> The test statistic is  $\Delta_M/\sqrt{\text{Var}(\Delta_M)} \sim \mathcal{N}(0,1)$ .

<sup>2</sup> The *p*-value corresponds to the probability value of the relevant test statistic under a two-sided alternative.

Here, the direction of the correlations  $\rho_{sj}$  remains unchanged, so that unobservables affecting the marginal probability of an ‘‘Injury’’ claim remain positively correlated with unobservables affecting the marginal probability of an ‘‘Own’’ injuries and damages claim (conversely for ‘‘Injury’’ and ‘‘Property’’ marginal probabilities, and ‘‘Own’’ and ‘‘Property’’ marginal probabilities). The results of the simulation are summarized under Table 9. These results suggest that, even if population correlations are halved, there still exist statistically significant differences between TVP and MNL predicted probabilities of the outcomes. In contrast to the results in Table 8, all  $\Delta$ ’s are statistically different from zero at the 5% significance level.

#### 4.4.2 Zero correlations

The aim of this experiment is to determine whether the MNL model can be used to approximate the outcome probabilities under an artificial setup whereby the disturbances are orthogonal. If the MNL model can be used in place of the TVP model at all, one would expect that the ideal conditions under which MNL may be appropriate would be such that the disturbances are independent and hence uncorrelated. In the previous two setups where we experimented with full and half correlation specifications for the covariance matrix, predicted probabilities of the MNL model were found to be statistically different from those of the TVP model. These results were *a priori* expected, and serve to reinforce our choice of the TVP over MNL when the disturbances are correlated.

Suppose now that the correlation between each claim-type disturbance is zero in the underlying population, so that the estimation of the TVP model is equivalent to the estimation of three independent univariate probit models. Therefore, instead of using our TVP coefficient estimates as the true population parameters in the experimental design, we use those estimated by three separate probit regressions with all the covariance terms set to zero, such that

$$\Sigma_Z = \mathbf{I}_3, \tag{19}$$

where  $\mathbf{I}_3$  is an identity matrix of dimension  $3 \times 3$ . Here,  $\Sigma_Z$  is consonant with either the independent probit<sup>5</sup> model or the MNL, depending on the distributional assumptions placed on the disturbances. The reason this follows is that both models assume that the disturbances are independent and identically distributed; the independent probit model assumes multivariate normality, whereas MNL assumes that the errors are type I extreme value. Where the parameters in the covariance matrix are allowed to vary freely, then we end up with the MNP model discussed in §1.

<sup>5</sup>The term ‘independent probit’ as used in this context refers to the MNP setup under which the covariance matrix is of the specification in (19), and is different from the independent *univariate* probit models mentioned in the preceding paragraph. This usage is consistent with Hausman and Wise (1978, p. 412).

**Table 10:** Simulation results for the standard deviation of  $\Delta_M$ , with zero correlations specification.

$\Delta_M$	Mean	Variance	Standard Error	Test Statistic <sup>1</sup>	$p$ -value <sup>2</sup>
$\Delta_1$	-0.00726	$1.02581 \times 10^{-7}$	0.00032	-22.66	0.000
$\Delta_2$	-0.01519	$8.71023 \times 10^{-5}$	0.00933	-1.63	0.104
$\Delta_3$	-0.00548	$7.47460 \times 10^{-8}$	0.00027	-20.05	0.000
$\Delta_4$	0.01278	$5.54073 \times 10^{-5}$	0.00744	1.72	0.086
$\Delta_5$	0.00403	$1.21972 \times 10^{-8}$	0.00011	36.47	0.000
$\Delta_6$	0.00496	$1.32017 \times 10^{-6}$	0.00115	4.31	0.000
$\Delta_7$	-0.00416	$8.74481 \times 10^{-7}$	0.00094	-4.45	0.000

<sup>1</sup> The test statistic is  $\Delta_M / \sqrt{\text{Var}(\Delta_M)} \sim \mathcal{N}(0, 1)$ .

<sup>2</sup> The  $p$ -value corresponds to the probability value of the relevant test statistic under a two-sided alternative.

The results of this simulation are summarized under Table 10. The results suggest that, even under ideal zero correlation conditions, the MNL fails to “correctly” estimate the predicted probabilities of each outcome; here, six out of seven  $\Delta$ ’s are statistically significant at the 10% level. There are two primary reasons for this finding. First, the TVP and MNL models are non-nested models and are inherently different choice processes. In the TVP model, we have three distinct choices, where each choice is represented as a binary outcome in each of the three equations in the setup of (4). By contrast, MNL models one single choice, over seven possible choices. Moreover, there are three sets of parameters as well as three additional covariance terms to be estimated in the TVP model; under MNL, there are six sets of parameters to be estimated, where the remaining choice is the normalized base outcome. Second, whilst the covariance structure assumed in (19) implies independent outcomes, the results of a multinomial choice model that is fitted to these generated data may not necessarily reflect this independence. This is illustrated under the MNP framework for three choices. Consider again the latent model specification in (4) for three choices. In a multinomial choice model, only differences in utility matter. Thus, normalizing the first choice yields

$$\begin{aligned} I_2^\circ - I_1^\circ &= \mathbf{x}'(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1) + \epsilon_2 - \epsilon_1 \\ I_3^\circ - I_1^\circ &= \mathbf{x}'(\boldsymbol{\beta}_3 - \boldsymbol{\beta}_1) + \epsilon_3 - \epsilon_1, \end{aligned} \quad (20)$$

so that it is one dimension smaller than before. Now, following the parameterization in Hausman and Wise (1978), we define

$$\eta_{21} = \epsilon_2 - \epsilon_1 \quad (21a)$$

$$\eta_{31} = \epsilon_3 - \epsilon_1, \quad (21b)$$

where the joint distribution of  $\eta_{j1}$ , for  $j = 1, 2$ , is bivariate normal, with alternative-specific covariance matrix

$$\boldsymbol{\Omega}_1 = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & \\ \sigma_1^2 - \sigma_{13} - \sigma_{12} + \sigma_{23} & \sigma_1^2 + \sigma_3^2 - 2\sigma_{13} \end{bmatrix}, \quad (22)$$

so that the probability of the first outcome is chosen is given by  $\Pr(I_2^\circ - I_1^\circ > \eta_{21}, I_3^\circ - I_1^\circ > \eta_{31})$ . Here,  $\boldsymbol{\Omega}_1$  is the actual covariance matrix to be estimated under the MNP framework for a three-choice setting. Clearly, then, even if the covariance terms in the initial covariance matrix are set to zero (that is,  $\sigma_{12} = \sigma_{31} = \sigma_{32} = 0$ ), the covariances in the estimated  $\boldsymbol{\Omega}_1$  may still be nonzero. This result extends to the current seven-choice setting of the MNL under consideration.

## 5 Summary and concluding remarks

The extension of the aggregate claims distribution to include the conditional claim-type component provides an actuary with additional information which is beneficial in two respects. First, it improves the accuracy in the prediction of future claims, and hence risks; and second, it facilitates a more equitable pricing of motor insurance contracts, whereby the amount of premium to be charged can be determined commensurately with the most likely claim-type combination, in addition to other factors. Conditional on there being at least one claim, an actuary is able to predict the precise combination of claim-types, given specific policy, driver and vehicle-specific characteristics. Our use of the general multivariate probit model is not limited to motor insurance contracts and can be generalized to other forms of insurance contracts in which multiple different claim-types may be observed. A benefit of our model is that it is flexible, allowing estimation of correlations between different claim-types. The following is a summary of the key findings of this paper.

First, fitting a multivariate probit model to our data is superior to the MNL model when the outcomes are correlated. The flexible specification of the multivariate probit model allows an actuary to estimate the extent of correlation between different claim-types. On the other hand, whilst the MNL model is definitely more computationally tractable than the multivariate probit model, its primary drawback is the IIA assumption.

Second, the results suggest that the amount of premium charged is an important predictor of claim-type. Specifically, for fleet policies, we found that higher premiums were associated with higher marginal probabilities of claiming with respect to “Injury” and “Own” damages, but the converse for “Property” damages. We found statistically and economically significant correlations between all three different types of claims; in particular, “Injury”–“Property” and “Own”–“Property” were negatively correlated, whilst “Injury”–“Own” were positively correlated. The latter suggests that those who claim for third party injury are also more likely to claim for own personal injuries and property damage.

Third, despite the finding of significant correlations between different claim-types, the MNL model produces qualitatively consistent predictions compared to those of the multivariate probit model. In the event of a claim, a policyholder is most likely to claim for third party property damages, followed by claims for own personal injuries and damages.

Finally, the adequacy of MNL as an appropriate model for claim-type was investigated under a controlled experiment where the true underlying DGP was known *a priori*. We substantiated the deviations in the predictions between the two models through an artificial setup in which three different specifications of the multivariate Normal covariance matrix were characteristic of the DGP: full, half and zero correlations. The results suggest that, even in ideal conditions under which the claim-types are uncorrelated, MNL is still a less-favored approximation to the “true” underlying outcome probabilities relative to the multivariate probit model when the MVP model is correct.

We have assumed in this paper that the MVP model specification is functionally correct and well-specified and that our model contains the true underlying DGP in that all covariates which help predict claim-type have been accounted for. We have mentioned before the possible presence of heteroskedasticity, and have tested for heteroskedasticity under a univariate probit setting using the Davidson and MacKinnon (1984)  $LM_2$  test statistic. We suspect that a finding of heteroskedasticity in the univariate case carries forward to the multivariate case, so that the standard errors of our estimates may in fact be incorrect. However, to explicitly model the form of the heteroskedasticity is beyond the scope of this paper.

Moreover, due to limitations in the data which are exacerbated by a large number of missing observations for potentially important covariates, there is a possibility that the coefficient estimates may have been affected by omitted variable bias. Whilst this is unfortunate, the problem is primarily one of the data, over which we have no control. We have implicitly assumed that the data were missing at random. A possible area of future investigation would be to explicitly model this data imbalance.

Finally, we acknowledge that our results pertain specifically to the randomly chosen insurer. We did not pool historical claims data from multiple insurers so as to avoid the problems of heterogeneity induced by the pooling of data which may have come from different underlying DGPs. Nonetheless, the implications of our results are that under the presence of potentially correlated responses, the MVP model is superior to the MNL model and thus should be preferred. Finally, there exists other alternatives to the multivariate probit model which are both flexible as well as computationally tractable, but have not been considered. An example is the class of generalized extreme-value (GEV) models, which assume that the unobservable component of the utility for all alternatives are jointly distributed as a generalized extreme value. More importantly, models within this class are able to capture sources of correlation between outcomes and are therefore not restrained by the IIA property. This is an interesting area for future research.

## References

- AMEMIYA, T. (1985): *Advanced Econometrics*. Oxford, UK: Basil Blackwell.
- ASHFORD, J., AND R. SOWDEN (1970): "Multivariate probit analysis," *Biometrics*, 26(3), 535–46.
- BALIA, S., AND A. M. JONES (2004): "Mortality, lifestyle and socio-economic status," University of York, Working Paper, dated October 2004.
- BOLDUC, D. (1992): "Generalized autoregressive errors in the multinomial probit model," *Transportation Research B*, 26B(2), 155–70.
- BUNCH, D. S. (1991): "Estimability in the multinomial probit model," *Transportation Research B*, 25B(1), 1–12.
- CAPPELLARI, L., AND S. P. JENKINS (2003): "Multivariate probit regression using simulated maximum likelihood," *The Stata Journal*, 3, 278–94.
- (2006): "Calculation of multivariate normal probabilities by simulation, with applications to maximum simulated likelihood estimation," ISER Working Paper 2006–16. Colchester: University of Essex.
- DAGANZO, C. (1979): *Multinomial Probit: The Theory and its Application to Demand Forecasting*. NY: Academic Press.
- DAVIDSON, R., AND J. G. MACKINNON (1984): "Convenient specification tests for logit and probit," *Journal of Econometrics*, 25, 241–62.
- GIBBONS, R. D., AND V. WILCOX-GÖK (1998): "Health service utilization and insurance coverage: A multivariate probit approach," *Journal of the American Statistical Association*, 93(441), 63–72.
- GREENE, W. H. (2003): *Econometric Analysis*. New Jersey: Prentice Hall, 5th edn.
- HAUSMAN, J. A., AND D. A. WISE (1978): "A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences," *Econometrica*, 46(2), 403–26.
- KEANE, M. P. (1992): "A note on identification in the multinomial probit model," *Journal of Business & Economic Statistics*, 10(2), 193–200.
- MADDALA, G. S. (1991): *Limited-Dependent and Qualitative Variables in Econometrics*. New York, NY: Cambridge University Press.
- McFADDEN, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Frontiers of Econometrics*, NY: Academic Press, pp. 105–42.

- (1981): “Econometric Models of Probabilistic Choice,” in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press, 198–272.
- PINQUET, J. (1998): “Designing optimal Bonus-Malus systems from different types of claims,” *ASTIN Bulletin*, 28(2), 205–20.
- SAUNDERS, A., J. BOUDOUKH, AND L. ALLEN (2003): *Understanding Market, Credit, and Operational Risk: The Value-at-Risk Approach*. Oxford: Blackwell Publishing.
- SMALL, K. A., AND C. HSIAO (1985): “Multinomial logit specification tests,” *International Economic Review*, 26, 619–27.
- TRAIN, K. E. (2003): *Discrete Choice Models with Simulation*. Cambridge: Cambridge University Press. Pre-print version, available from <<http://elsa.berkeley.edu/books/choice2.html>>.
- VALDEZ, E. A., AND E. W. FREES (2005): “Longitudinal modeling of Singapore motor insurance,” University of New South Wales and the University of Wisconsin-Madison, Working Paper, dated 28 December 2005, available from <<http://wwwdocs.fce.unsw.edu.au/actuarial/research/papers/2006/Valdez-Frees-2005.pdf>>.
- WEEKS, M. (1997): “The multinomial probit model revisited: A discussion of parameter estimability, identification and specification testing,” *Journal of Economic Surveys*, 11(3), 297–320.