Probit and Logit Models: Differences in the Multivariate Realm

Eugene D. Hahn

Salisbury University, Salisbury, MD, USA.

Refik Soyer

The George Washington University, Washington, DC, USA.

Summary. Current opinion regarding the selection of link function in binary response models is that the probit and logit links give essentially similar results. This seems to be true for univariate binary response models; however, for multivariate binary response models such advice is misleading. We address a gap in the literature by empirically examining the relationship between link function selection and model fit in two classes of multivariate binary response models. We find clear evidence that model fit can be improved by the selection of the appropriate link even in small data sets. In multivariate link function models, the logit link provides better fit in the presence of extreme independent variable levels. Conversely, model fit in random effects models with moderate size data sets is improved generally by selecting the probit link.

Key Words generalized linear models, link function, Bayesian inference, Markov chain Monte Carlo (MCMC), DIC.

1. Introduction

Probit and logit models are among the most widely used members of the family of generalized linear models in the case of binary dependent variables. In probit models, the link function relating the linear predictor $\eta = \mathbf{x}\boldsymbol{\beta}$ to the expected value μ is the inverse normal cumulative distribution function, $\Phi^{-1}(\mu) = \eta$. In the logit model the link function is the logit transform, $\ln(\mu/1-\mu) = \eta$. The conventional wisdom is that in most cases the choice of the link function is largely a matter of taste. For example, Greene (1997, p. 875) concludes his discussion of the issue with the summary "in most applications, it seems not to make much difference." Gill puts it especially plainly; in discussing link functions including the cloglog, he indicates that they "provide identical substantive conclusions" (Gill, 2001, p. 33). Elsewhere, similar advice appears regularly when the topic is discussed (e.g., Maddala, 1983; Davidson and MacKinnon, 1993; Long, 1997; Powers and Xie, 2000; Fahrmeir and Tutz, 2001; Hardin and Hilbe, 2001). Empirical support for the recommendations regarding both the similarities and differences between the probit and logit models can be traced back to results obtained by Chambers and Cox (1967). They found that it was only possible to discriminate between the two models when sample sizes were large and certain extreme patterns were observed in the data. We discuss their work in greater detail below.

Since the time of Chambers and Cox, a great number of developments have occurred in the area of binary response models. Increasingly, interest has turned to instances where there is more than one binary response variable to consider. For example, Ashford and Sowden (1970) proposed a multivariate probit model. More recently, the linear mixed models framework has been extended to binary response data (Stiratelli et al., 1984). Despite these developments, the properties of link

functions for binary response models in the multivariate realm remain largely unexplored. This is unfortunate, as it turns out that the impact of link function on model fit is strongly affected by the form of the model considered.

In the current paper we address this gap in the literature by examining model fit for two families of multivariate binary response models. In Section 2.1, we review the bivariate probit model of Ashford and Sowden (1970) and propose an approximate bivariate logistic model by exploiting the relationship between the logistic distribution and the t distribution with degrees of freedom $\nu = 8$. As an alternative dependence structure a random effects model is presented by introducing a common intercept term to the marginal link functions across the response variables. Factors that are expected to influence the fit of these models are discussed in Section 2.2. The deviance information criterion and Bayes factors are presented in Section 2.3 as the measures of fit that are used in our study. Three research propositions are stated in Section 2.4 and the methods used in the study are described in Section 3. Our findings and discussion of these are presented in Section 4 and 5, respectively,

2. Link function and model fit

2.1. Multivariate binary response models

Many models for multivariate binary response data are possible (e.g., Fahrmeir and Tutz, 2001). Here, we review two of the more widely used frameworks. The first involves specifying a joint multivariate link function for the multiple binary responses. For example, the bivariate probit model described in Ashford and Sowden (1970) can be written as

$$\begin{aligned} P(Y_{i,j} = 1|\mathbf{x}_{i,j}) &= \Phi(\eta_{i,j}), \qquad j = 1, 2 \\ P(Y_{i,1} = 1, Y_{i,2} = 1|\mathbf{x}_{i,j}) &= \Phi_2(\eta_{i,1}, \eta_{i,2}, \rho) \end{aligned} \tag{1}$$

where Φ_2 is the bivariate standard normal cumulative distribution function, and i and j index respondents and dependent variables respectively. This approach could be applied directly to obtain a bivariate logistic model. However, the various extant multivariate logistic distributions have properties such as restrictions on possible values of correlation coefficients and asymmetric non-elliptical distributions (Kotz et al., 2000, ch. 51) that make such a direct approach less practical. For example, the Type II distribution of Gumbel (1961, Eq. 6.3) is among the more attractive of the bivariate logistic distributions as it is not asymmetric. However, as Smith and Moffatt (1999, p. 318) recently pointed out, the correlation is restricted such that $|\rho| < 3/\pi^2 \approx .304$. Clearly this is a considerable limitation. Therefore, an attractive alternative is to capitalize on the logistic distribution's relationship to the t distribution.

Albert and Chib (1993) examined the choice of link function in binary response models from the Bayesian perspective. They discussed the similarities between the logistic distribution and the t distribution with degrees of freedom $\nu = 8$. In particular, their plot of the logistic quantiles against the quantiles of the $t(8)$ distribution shows an approximately linear relationship between the two distributions. Albert and Chib (1993) determined that a $t(8)$ variable is approximately .634 times a standard logistic variable. Examining further, it turns out that we can show there is almost a one-to-one relationship between these two distributions with the appropriate parameterization. The logistic pdf with location parameter c and scale parameter d is

$$P(x) = \frac{\exp[(x-c)/d]}{d\big\{1 + \exp\big[(x-c)/d\big]\big\}^2}\,.$$

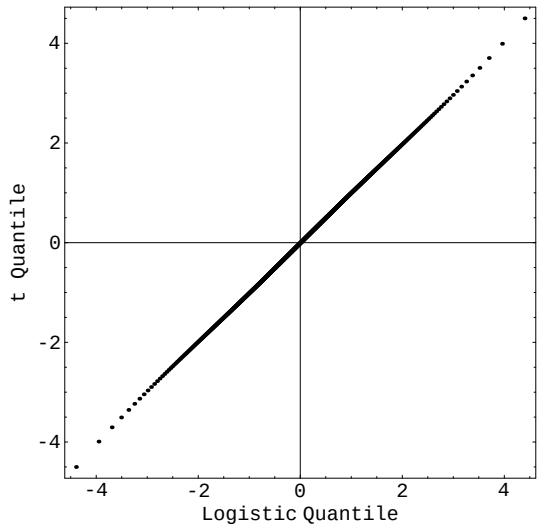

Fig. 1. Quantile values of Logistic$(2/\pi)$ versus $t(8)$ for probabilities from .001 to .999

Note that the $t(8)$ distribution has variance $4/3$ and that the standard logistic distribution with $c = 0$ and $d = 1$ has variance $\pi^2/3$. We may therefore equate the variances of the two distributions by setting the logistic distribution's scale parameter to $2/\pi$. With $c = 0$ the first three moments of the two distributions are then identical, with standardized fourth moments being very close ($\gamma_2 = \mu_4/\mu_2^2 = 4.2$ in the case of the Logistic$(2/\pi)$ and $\gamma_2 = 4.5$ for the t). Thus, we can see the $t(8)$ has approximately the same distribution as the Logistic$(2/\pi)$ but is just marginally more leptokurtic. To obtain a more concrete understanding of the similarities between the two distributions, we may examine the linear relationship between the quantiles of the two distributions for probabilities between .001 to .999, e.g., as in Albert and Chib (1993). Figure 1 displays a plot of this relationship. We find the linear relationship between these quantiles is described by the equation $tq = 5.6616 \times 10^{-17} + .9976 \times lq$, where tq is the $t(8)$ quantile and lq is the quantile for the logistic distribution with scale $2/\pi$. The R^2 between the two sets of quantiles is in excess of .9999. Hence, the $t(8)$ distribution provides a very satisfactory approximation to the logistic distribution. As such, we propose the following approximate multivariate logistic model

$$\begin{aligned} P(Y_{i,j} = 1|\mathbf{x}_{i,j}) &= F_{t(8)}(\eta_{i,j}), \qquad j = 1, 2 \\ P(Y_{i,1} = 1, Y_{i,2} = 1|\mathbf{x}_{i,j}) &= \mathbf{F}_{t(8)}(\eta_{i,1}, \eta_{i,2}, \rho) \end{aligned} \tag{2}$$

where $F_{t(8)}$ is the $t(8)$ cdf and $\mathbf{F}_{t(8)}$ is the bivariate $t(8)$ cdf. Note that Chen and Dey (1998) developed a Bayesian multivariate logistic model using a scaled multivariate t proposal distribution involving somewhat heavier tails ($\nu = 5$). Given the excellent fit of the $t(8)$, we expect that their formulation would yield essentially equivalent results to the ones obtained here.

Another frequently used model for multivariate binary response data is the random effects model. Here, we have

$$P(Y_{i,j}|\mathbf{x}) = g(\eta_j + b_i), \quad i = 1, \ldots, n \quad j = 1, \ldots, J \tag{3}$$

in which the probability of observing the response on variable j in individual i is related both to the linear predictor η_j as well as an individual-specific random intercept, b_i. The intercepts are specified to arise from a common distribution. Thus, in the random effects model, dependence is introduced at the respondent level by the presence of a shared intercept term across the J dependent variables. We can see therefore that the link function $g(\cdot)$ does not need to be given a multivariate characterization. This is especially convenient as multivariate link functions are more computationally expensive to evaluate and sometimes, as in the case of the logistic distribution, are simply unavailable in a sufficiently flexible form. As such, random effects approaches seem to be much more widely used for multivariate binary response data. Zeger and Karim (1991) provided an early Bayesian development of the model in the context of a Gibbs sampling approach.

2.2. Factors influencing fit

As mentioned above, Chambers and Cox (1967) established that under certain conditions it was possible to distinguish the results from probit and logit models. In particular, they were able to distinguish between the link functions when sample sizes were large (e.g., $n \geq 1000$) and where there were what can be termed extreme independent variable levels. An extreme independent variable level involves the confluence of three events. First, an extreme independent variable level occurs at the upper or lower extreme of an independent variable. For example, say the independent variable x were to take on the values 1, 2, and 3.2. The extreme independent variable level would involve the values at $x = 3.2$ (or $x = 1$). Second, a substantial proportion (e.g., 60%) of the total n must be at this level. Third, the probability of success at this level should itself be extreme (e.g., greater than 99%).

While the conditions under which univariate probit and logit models could be distinguished were established by Chambers and Cox, the conditions under which the two link functions can be distinguished in multivariate binary response models have not been examined. Here, we remedy this gap using two major families of models: the multivariate link function models such as (1) and (2), and the random effects models of (3). We consider the bivariate case here. As such, we utilize the bivariate probit model, first considered from a Bayesian perspective by Chib and Greenberg (1998), as well as the new formulation of the multivariate logit model proposed in (2). We also consider the random effects model under the probit link as well as under the Logistic$(2/\pi)$ link. We explore the behavior of these models in the presence of extreme independent variable levels as well as in the absence thereof. We also explore these models' behavior in the context of both moderate and high levels of dependent variable correlation. It may seem that, for a given level of dependent variable correlation, numerous data sets will need to be randomly sampled and analyzed via a Monte Carlo study to ensure the robustness of the findings. However, note that bivariate binary data can be expressed as a contingency table with four cells: a, b, c and d. The measure of association for the contingency table for any given n can be calculated deterministically via Pearson's phi, which is

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{4}$$

It is easy to show that for any fixed values of n and φ, at most one data set can be generated up to a relabeling of the cells. A somewhat similar argument applies to the generation of a predictor with extreme independent variable levels. A predictor with an extreme independent variable level as described by Chambers and Cox (1967) has a fairly well-specified set of properties. Deviating from these properties will likely lead to a predictor that does not have extreme independent variable levels. Thus, we examine here data sets that either do or do not have the property of extreme independent

variable levels as defined by Chambers and Cox. Moreover, we examine data sets with a particular moderate or high level of dependent variable correlation. Details regarding these data sets appear in Section 3.

2.3. Measures of fit

Traditional Bayesian model comparison is performed using Bayes factors (Kass and Raftery, 1995). More recently, Spiegelhalter et al. (2002) introduced the Deviance Information Criterion (DIC) which combines measures of both model fit and model complexity. Specifically,

$$\text{DIC} = \overline{D} - p_D$$

where $\overline{D}$ is the posterior mean of the deviance and p_D is a measure of model complexity which may be termed the effective number of parameters. In fixed effects models, p_D should approximately equal the actual number of model parameters. In random effects models, p_D will typically be less than the actual number of model parameters. Nonetheless, p_D gives an indication of how much these terms are contributing to the model's overall performance. p_D itself is defined as $\overline{D} - D(\overline{\theta})$, where $D(\overline{\theta})$ is the deviance evaluated at the posterior means of the parameters. Models with greater values of p_D are penalized for their greater complexity *ceteris paribus*, as smaller values of DIC are preferred. Thus, DIC is similar in interpretation and in spirit to another information-theoretic model comparison criterion, AIC (Akaike, 1973). Based on this similarity, Spiegelhalter et al. (2002) cite work in Burnham and Anderson (1998) which suggests that models with a DIC which is 3–7 greater than a 'better' model deserve less consideration. We adopt this criterion here for assessing model fit. It is perhaps natural to want to compare the more recently-developed DIC measure with the traditional Bayes factor, although Spiegelhalter et al. (2002) caution against this since the two methods have different purposes. Specifically, the Bayes factor summarizes how well the prior has predicted the obtained data whereas DIC summarizes how well the posterior might predict future data that had been generated by the same process as that which generated the obtained data. Therefore another way of describing two approaches is that the Bayes factor has a prior predictive emphasis while DIC has a posterior predictive emphasis. Nonetheless, there is some preliminary evidence to suggest that the two approaches may provide substantively similar results, at least in some circumstances. Berg et al. (2004) conducted a simulation study to compare the performance of DIC against Bayes factors calculated by the marginal likelihood method of Chib (1995) as well as by the harmonic mean method of Newton and Raftery (1994). Performance comparisons were also made using empirical financial data drawn from the 1993–1998 Standard & Poors 100 market index. In both studies, DIC and Chib's method yielded similar results.

In summary, Bayes factors have long been used in the context of Bayesian inference. Subject to certain caveats (e.g., Lavine and Schervish, 1999), they may be preferred in certain situations. However, here it is of interest to examine model fit and model complexity simultaneously because of the random effects models in (3). Additional insights are available by examining model complexity via p_D under different extreme independent variable levels and different levels of dependent variable correlation, particularly for the random effects models. By contrast, Bayes factors do not provide measures of model complexity. Furthermore, DIC is a measure of a model's out-of-sample predictive ability. Thus, DIC appears to be the more relevant criterion here. Nonetheless, as a point of comparison, we calculate the log marginal likelihoods for the different models here. We use the Laplace-Metropolis method of Lewis and Raftery (1997). The Laplace method is known to perform well with regard to accuracy for marginal likelihood calculations even in the presence of small sample sizes. For example, in probit models with a sample size approximately half of what we consider

here Chib (1995) compared the performance of his method with the Laplace method. He found agreement between the two approaches up to the second decimal place.

2.4. Research propositions

We present here three research propositions derived in part from the theoretical development above. Findings with respect to these propositions should help provide assistance in decisions regarding the selection of the link function in multivariate binary response models.

Research Proposition 1. The presence of extreme independent variable levels will lead to increasingly pronounced differences in fit across the two link functions.

We arrive at this proposition directly from the work of Chambers and Cox (1967). Specifically, to the extent that there are differences in model fit, they will be exacerbated by the occurrence of extreme independent variable levels.

Research Proposition 2. Increasingly positive correlation will lead to decreased differences in fit across the two link functions.

In describing this proposition, we may begin by describing the following trivially obvious statement: all else being equal, any differences in model fit will increase as the sample size increases. For example, part of Chambers and Cox's work involved finding at what sample size one may discern differences in binary response model fit across the two link functions. In Research Proposition 2 the observation is as follows: as the correlation increases, what may be termed the effective sample size decreases. Phrased differently, the amount of new information provided by y_2 relative to y_1 decreases with increasing correlation. In the limiting case when $\rho = 1$, the bivariate model could be replaced by a univariate one as y_2 provides no information that has not already been provided by y_1. Hence, differences in fit will be diminished at higher correlations.

Research Proposition 3. In random effects models, use of the probit link results in model fit that is as good or better than model fit under the logit link.

This possibly surprising proposition does not directly stem from the work of Chambers and Cox but instead can be obtained as follows. Recall that the logistic distribution is leptokurtic relative to the normal distribution and so in fixed effects models having some overdispersion, we might expect logit models to fit somewhat better. While in the current study the principal use for the random effects terms is as a means for introducing dependence between the binary response variables, note that random effects terms also can be used as a device to model overdispersion. Therefore, in a random effects model where the random effects terms are adequately modeling any existing overdispersion, the logit link should likely not fit better than the probit link. Clearly the random effects terms will already be capturing the overdispersion, so the heavy tails of the logit will likely not contribute to further improvements in fit. Instead, we would expect that the more compact normal distribution associated with the probit model to provide a more precise fit. It would of course be possible to construct a random effects model in which the random effects terms did a poor job of modeling overdispersion. For example, one could assign a highly informative prior to the random effects terms such that the prior was very discrepant from the actual patterns in the data. However, in a random effects model that is functioning well (i.e., fitting the overdispersion accurately), the probit model should lead to improved fit.

3. Methods

We previously described how extreme independent variable levels were those in which the ability to discriminate between probit and logit links are maximized. We now describe their operationalization in the current study. Chambers and Cox (1967) investigated the case where there were three levels for a single independent variable. To be more specific about what constitutes an extreme independent variable level, they found that the three levels of x should be 1, 2, and 3.2 respectively. They also found that, depending on the baseline link (probit vs. logit), either 11.7% or 16.7% of the total responses should be placed at Level 1 ($x = 1$). Due to the constraints of needing to have an integer number of successes in the data as well as of working with considerably smaller sample sizes, we approximate the average of these proportions by placing 13.3% of the observations at this level when extreme independent variable levels are desired. They also found that the proportion of successes at Level 1 should be either 21.5% or 17.1%. Here, our slightly crude approximation of these proportions (resulting again from much smaller sample sizes) is that the number of successes at Level 1 will be 16.7%. Note that, if anything, the crudeness of this approximation (and any others we might consider) will make it more difficult for us to demonstrate differences between the link functions since Chambers and Cox described the optimal points at which discriminability was globally maximized. Level 2 should contain either 21.4% or 26.3% of the responses, with the proportion of successes being either 78.5% or 82.9%. Here, we place 20.0% of the observations at this level with 77.8% being successes. Finally, Level 3 should contain either 66.9% or 57.5% of the responses, with the proportion of successes being either 99.64% or 99.87%. We place 66.7% of the observations at this level with 96.7% being successes.

For the case of non-extreme independent variable levels, we create data in such a way that the exact opposite of the three conditions above are obtained. First, we divide the data evenly among the levels so that each level contains $n/3$ observations. Second, less extreme proportions of successes are placed at each level. In particular, the proportion of successes are 60.0%, 80.0%, and 86.7% for Levels 1, 2, and 3 respectively. Then the third condition is also satisfied: given that all of the levels have equal sample sizes and more modest proportions of successes, then the necessary conditions do not exist at the extreme levels of the independent variable since they do not exist at any of its levels. We take the three levels of x to be 1, 2, and 3. In a departure from the recommendations of Chambers and Cox, we consider smaller sample sizes of $n = 90$ and $n = 450$. This is because in many occasions sample sizes used in binary response models have more modest sample sizes than the one considered by Chambers and Cox. Data sets having $n = 450$ were generated by stacking 5 copies of the respective $n = 90$ data set.

We also consider two levels of dependent variable correlation: moderate and high. In the extreme independent variable level conditions, the correlation φ will be set at .544 as a moderate amount of correlation, and .848 for a high amount. In the conditions where independent variable levels are not extreme, φ will be set at .519 as a moderate amount of correlation, and .819 for a high amount. The values of φ vary slightly across the extreme/non-extreme conditions here because of the limitations of having to specify an integer number of cases at each level for a smaller sample size. Nonetheless, the across-condition correlations are quite close to one another; the differences are all less than 0.03. Given these factors of interest, the Monte Carlo study design had a 2 (extreme or non-extreme independent variable level) $\times$ 2 (small or large n) $\times$ 2 (moderate or high dependent variable correlation level) $\times$ 2 (model type: multivariate link versus random effects model) $\times$ 2 (logit or probit link) factorial structure. The first three of these factors involve differences that may be encountered in data whereas the latter two factors involve model choice which is under the

control of the statistician.

We estimate the models in (1) and (2) as well as logit and probit versions of (3). To further facilitate comparability, the logit version of (3) utilizes the Logistic($2/\pi$) distribution as opposed to the standard Logistic(1). This is easily accomplished by using the data augmentation approach of Albert and Chib (1993). The probit version of (3) is also estimated using data augmentation. We complete the specification of the multivariate link function models (1) and (2) as follows

$$\begin{aligned} Y_{i,j} &\sim \text{Bernoulli}(p_{i,j}) \\ p_{i,j} &= g(\eta_{i,j}) \\ \eta_{i,j} &= \beta_{1,j} + \beta_{2,j}\, x_i \\ \beta_{1,j} &\sim N(0, 0.02) \\ \beta_{2,j} &\sim N(0, 0.02) \\ \rho_j &\sim U(-1, 1), \end{aligned}$$

where $g(\cdot)$ is the link function. In the random effects models of (3), we complete the specification as

$$\begin{aligned} Y_{i,j} &\sim \text{Bernoulli}(p_{i,j}) \\ p_{i,j} &= g(\eta_{i,j}) \\ \eta_{i,j} &= \beta_{1,j} + \beta_{2,j}\, x_i + b_i \\ \beta_{1,j} &\sim N(0, 0.02) \\ \beta_{2,j} &\sim N(0, 0.02) \\ b_i &\sim N(0, \tau) \\ \tau &\sim G(0.05, 0.05). \end{aligned}$$

Under the high dependent variable correlation conditions, convergence of the random effects models is improved by adopting mildly informative priors. Accordingly, the β parameters were given normal priors with precisions of 0.02 (i.e., variances of 50) and prior means of zero. These priors are not particularly informative (especially given the modest values of β associated with binary response models) and they gave considerable leeway for the parameters to move toward their posteriors. As mentioned previously, the b_i parameters are assumed to arise from a common distribution. The distribution used here for the b_is is the normal with mean zero and precision τ. The prior for τ was also a mildly informative Gamma prior with prior shape and scale of 0.05. For consistency purposes, the βs in models (1) and (2) were also given prior means and precisions of zero and 0.02. The correlation parameter, ρ, in (1) and (2) was given a flat uniform prior over the interval $[-1, 1]$. Estimation was conducted using MCMC. For all models, 5,000 iterations of burn-in were discarded and 150,000 samples from the posteriors were retained for use.

4. Results

From the preceding discussion, we examine four conditions in the current research. In Condition 1, the data has non-extreme independent variable levels with moderate dependent variable correlation. In Condition 2, the data has non-extreme independent variable levels with high dependent variable correlation. In Condition 3, the data has extreme independent variable levels with moderate dependent variable correlation and the data has extreme independent variable levels with high

Table 1. Model fit measures: small sample size

		Multivariate Link				*Random Effects*			
		NEI Mod	*NEI High*	*EI Mod*	*EI High*	*NEI Mod*	*NEI High*	*EI Mod*	*EI High*
DIC	*logit*	180.8	146.5	103.4	83.5	145.5	60.7	100.9	45.8
	probit	180.6	146.3	104.8	84.0	143.2	59.0	98.0	42.0
$\overline{D}$	*logit*	176.0	142.0	98.7	79.1	100.5	38.8	79.9	28.9
	probit	175.9	141.9	100.0	79.6	98.4	37.8	74.7	26.7
p_D	*logit*	4.79	4.50	4.72	4.40	44.9	21.9	21.0	16.9
	probit	4.76	4.47	4.77	4.42	44.8	21.1	23.2	15.2

NEI indicates non-extreme independent variable levels; EI indicates extreme independent variable levels; Mod (moderate) and High refer to dependent variable correlation levels.

dependent variable correlation in Condition 4. Table 1 contains the results for the models under the four conditions for the small sample size ($n = 90$).

We first examine the results for the multivariate link models. We see that fit as measured by DIC under the non-extreme independent variable level conditions is comparable across links since the differences in DIC across links are well below 3. The values of DIC as well as $\overline{D}$ may provide a very modest indication that the probit models may be fitting trivially better, but the differences are at best slight. As the dependent variable correlation moves from moderate to high, we see the value of p_D drop from around 4.8 to the vicinity of 4.5. This reflects the increasing parameter redundancy under high dependent variable correlation. In the extreme independent variable level conditions, the differences in fit become slightly more pronounced with the logit model fitting marginally better. The heavier tails of the logistic distribution seem to provide a minimally better fit under moderate or high levels of correlation in the presence of extreme independent variable levels. The values of p_D suggest that the probit model is less susceptible to increased parameter redundancy under high correlation and extreme independent variable level in small sample sizes.

For the random effects models, the DIC results are only clearly delineated in the high correlation extreme independent variable level condition. There we see that the DIC difference is 3.8 with the probit model having a DIC of 42.0 versus a DIC of 45.8 for the logit model. Nonetheless, under the other conditions the probit looks to be the more competitive, although the differences are rather small due to the small sample size. We also see that the effective number of parameters as measured by p_D is considerably smaller than the true number of parameters, 95; this is not uncommon in random effects models due to parameter redundancy (see Spiegelhalter et al., 2002). The values of p_D across link function tend to be relatively similar within condition. When the dependent variable correlation becomes high or when extreme independent variable levels are present, the values of p_D fall markedly.

Table 2 displays the results for the models when the sample size is larger ($n = 450$). Consistent with expectations, we find here that differences between the two link functions become increasingly distinct. For example, in the multivariate link models the logit model becomes noticeably more preferred by DIC in the extreme independent variable level conditions. Under moderate dependent variable correlation the difference in DIC in favor of logit is 7.9; under high correlation the

Table 2. Model fit measures: large sample size

		Multivariate Link				*Random Effects*			
		NEI Mod	*NEI High*	*EI Mod*	*EI High*	*NEI Mod*	*NEI High*	*EI Mod*	*EI High*
DIC	*logit*	864.4	690.0	477.5	373.8	720.0	303.2	475.1	205.5
	probit	863.8	689.8	485.4	377.8	710.7	292.8	465.2	195.9
$\overline{D}$	*logit*	859.4	685.2	472.6	369.0	514.6	194.2	406.0	132.9
	probit	858.8	685.1	480.5	373.1	503.1	188.4	383.2	127.4
p_D	*logit*	4.97	4.85	4.97	4.74	205.3	109.0	69.1	72.6
	probit	4.93	4.79	4.93	4.74	207.6	104.4	82.2	68.5

Table 3. Log marginal likelihoods

		Multivariate Link				*Random Effects*			
		NEI Mod	*NEI High*	*EI Mod*	*EI High*	*NEI Mod*	*NEI High*	*EI Mod*	*EI High*
Small sample size	*logit*	-102.0	-85.8	-60.6	-51.0	-98.4	-76.1	-62.3	-46.3
	probit	-102.5	-86.3	-62.8	-52.8	-99.2	-76.7	-62.5	-46.1
Large sample size	*logit*	-448.0	-362.5	-251.8	-201.6	-443.8	-351.4	-262.3	-199.9
	probit	-448.1	-363.0	-257.0	-204.7	-444.6	-352.2	-261.8	-200.4

difference is 4.0. We might expect though that with further increases in sample size the relative superiority of the logit's fit would continue to grow. In the random effects models, the probit link provides a considerably better fit with all of the differences in DIC favoring probit by 9.3 or more. There is a notable amount of consistency in the DIC differences favoring probit: the differences all lie within a relatively narrow band from 9.3 to 10.4 despite the variation in the data across the four conditions. The values of p_D are relatively similar in the non-extreme independent variable level conditions. They become more dissimilar in the moderate correlation extreme independent variable level condition. Here, the heavier tails of the logistic distribution seems to allow the model to be estimated with a smaller amount of effective parameters. By contrast, the more compact normal distribution generates a greater number of distinct effective parameters. This offsets the relatively large reduction in deviance (difference in $\overline{D}$ = 22.8) that the probit provides over the logit. In the final condition, the probit's advantage is reduced by the high correlation and so the models' p_D values are again more similar.

Table 3 contains the log marginal likelihoods for the models under consideration. We first examine the multivariate link models. In the small sample size condition, there is little to distinguish the logit and probit links in the two non-extreme independent variable level conditions. In the extreme independent variable level conditions, the Bayes factors somewhat tend toward the logit link over the probit with values of 8.57/1 in the moderate correlation condition and 6.10/1 in the high correlation condition. In the large sample size condition, this pattern is repeated with the extreme independent variable level condition Bayes factors in support of the logit link being considerably larger (172.7/1 and 23.8/1 for the moderate and large correlation conditions respectively). Thus, we see that DIC and the Bayes factors are in agreement with respect to these fixed effects models:

the logit is preferred in the case of extreme independent variable levels. With the random effects models, however, DIC and Bayes factors provide different pictures. As described earlier, the values of both DIC and also $\overline{D}$ in Tables 1 and 2 are substantially smaller for the probit models, indicating that from a minimum-deviance perspective probit models perform noticeably better. However, the log marginal likelihoods for the random effects models in Table 3 are approximately equal across links, indicating little support for one link function over the other.

5. Discussion

Tables 1 and 2 illustrate that the conventional wisdom about the relative similarity of the logit and probit link functions in binary response models does not carry over to the multivariate realm. In fact, differences in fit can be found even in small sample sizes. With regard to the research propositions, considerable support for Proposition 1 was found in the results for the multivariate link models. In the small sample size condition (Table 1), the differences in DIC for these models in the non-extreme independent variable level conditions were relatively small (0.2 or less for multivariate link and 2.2 or less in the random effects models), while in the extreme independent variable level conditions they increased. In Table 2 Proposition 1 was again supported in the multivariate link models, and it is in this context that the differences were most pronounced. However, Table 2 also shows that support was not quite obtained in the context of large sample random effects models under high correlation. It might be expected that Proposition 1 would be largely supported based on the work of Chambers and Cox (1967). However, to the best of our knowledge it has never been verified in the case of random effects models. The point deserved examination for two reasons. First, models with random effects terms are qualitatively different from the purely fixed effects models that were considered by Chambers and Cox (e.g., in non-MCMC approaches, an integration of the random effects is involved in the formulation of these models). Second, the proposition did not appear to hold up as well in the case of random effects models with larger sample sizes. Evidence supporting Proposition 2 was primarily found in the multivariate link model conditions with extreme independent variable levels. For example, in Table 2 we find the difference in DIC across links declined from 7.9 in the moderate correlation extreme independent variable level condition to 4.0 in the high correlation extreme independent variable level condition. In the non-extreme independent variable level conditions, a similar pattern is visible but the magnitudes are far smaller. When the sample size was small, the evidence declined. By contrast, the results Proposition 2 were equivocal for random effects models. Finally, from a deviance perspective evidence for Proposition 3 was also obtained and in terms of statistical practice these findings arguably have the most important implications. Simply stated, the probit link appears to offer a consistent advantage over the logit link in random effects models from the perspective of minimizing deviance and enhancing model fit. For the small sample size results of Table 1, the evidence was not of sufficient magnitude to constitute a notable difference (except in the case of the high correlation extreme independent variable level condition) but the probit link was consistently favored. For the larger sample size of Table 2, the differences in DIC all exceeded 9, well over the threshold adopted here.

In summary, judicious selection of the link function seems likely to help improve model fit in multivariate binary response models according to a deviance-based perspective. Model fit in random effects models seem to be improved generally by selecting the probit link. By contrast, the logit link seems preferable for multivariate link models when there are extreme independent variable levels. However, we note that when a perspective based on Bayes factors is adopted, the interpretation of the findings becomes somewhat less clear cut. For the fixed effects multivariate link models, the findings were consistent across the DIC and Bayes factor measures. For example, the logit link is

selected by both approaches in the context of multivariate link models with extreme independent variable levels. However, in the random effects models there were little differences to be found between the link functions according to the Bayes factors. Given their prior-predictive nature, this indicates that in the random effects models the prior predicted the data equally well across the two link functions. So, from a prior predictive viewpoint, there is little to differentiate the models. However, if we are interested in both in-sample predictive ability (as measured by the deviance) and out-of-sample predictive ability (as measured by DIC), then in the random effects models the probit is clearly preferable.

It is not uncommon to find disagreements between the Bayes factors and deviance based measures such as DIC. It was noted by Kass and Raftery (1995) that Bayesian Information Criterion (BIC), another deviance based measure, does not approximate Bayes factors well in cases where the number of parameters is large relative to the sample size. Similar findings were reported by Carlin et al. (1992) where authors used random effects logistic models. Furthermore, evaluation of Bayes factors in random effects models under the probit and logit links poses computational challenges and therefore the disagreements may be attributed to the accuracy of these results, although as discussed previously the Laplace method has attractive performance properties. We consider this as a future research topic.

One might speculate as to whether the results presented here would replicate to other situations. There appear to be relatively few instances of published analyses involving link function comparison and the use of DIC in the context of multivariate binary response models. However, some such research has appeared. In particular, Spiegelhalter et al. (2002, §8.3) also happened to provide an example in which results for random effects models under the probit and logit link were contrasted (as were the results under the cloglog). The data set was that of a real world study of the effects of air pollution. Interestingly, the probit link was again preferred, in both the canonical and mean parameterizations (DICs 1411.3 and 1307.3 respectively), over the logit (DICs 1415.1 and 1335.3 respectively).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csàki (Eds.), *Proceedings of the Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akadémiai Kiadó. Reprinted in *Breakthroughs in Statistics*, vol. 1, pp. 610-624, eds. Kotz, S. and Johnson, N. L., 1992, New York: Springer.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Ashford, J. R. and R. R. Sowden (1970). Multi-variate probit analysis. *Biometrics 26*, 535–546.

Berg, A., R. Meyer, and J. Yu (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business & Economic Statistics 22*(1), 107–120.

Burnham, K. P. and D. R. Anderson (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Carlin, B. P., R. E. Kass, F. J. Lerch, and B. R. Huguenard (1992). Predicting working memory failure: A subjective Bayesian approach to model selection. *Journal of the American Statistical Association 87*, 319–327.

Chambers, E. A. and D. R. Cox (1967). Discrimination between alternative binary response models. *Biometrika 54*, 573–578.

Chen, M.-H. and D. K. Dey (1998). Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhyā, Series A 60*, 322–343.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chib, S. and E. Greenberg (1998). Analysis of multivariate probit models. *Biometrika 85*, 347–361.

Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York: Oxford.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer.

Gill, J. (2001). *Generalized Linear Models: A Unified Approach*. Thousand Oaks, CA: Sage.

Greene, W. H. (1997). *Econometric Analysis* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Gumbel, E. J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association 56*, 335–349.

Hardin, J. and J. Hilbe (2001). *Generalized Linear Models and Extensions*. College Station, TX: Stata Press.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–794.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions Volume 1: Models and Applications* (2nd ed.). New York: Wiley.

Lavine, M. and M. J. Schervish (1999). Bayes factors: What they are and what they are not. *The American Statistician 53*, 119–122.

Lewis, S. M. and A. E. Raftery (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association 92*, 648–655.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (*with discussion*). *Journal of the Royal Statistical Society: Series B 56*, 3–48.

Powers, D. A. and Y. Xie (2000). *Statistical Methods for Categorical Data Analysis*. San Diego: Academic Press.

Smith, M. D. and P. G. Moffatt (1999). Fisher's information on the correlation coefficient in bivariate logistic models. *Australian & New Zealand Journal of Statistics 41*, 315–330.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (*with discussion*). *Journal of the Royal Statistical Society: Series B 64*, 583–639.

Stiratelli, R., N. Laird, and J. H. Ware (1984). Random-effects models for serial observations with binary response. *Biometrics 40*, 961–971.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association 86*, 79–86.