$\mathcal{I}^2\mathcal{SDS}$

The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2008-17

September 18, 2008

Probit Versus Logit Models: The Role of the Link Function in the Multivariate Realm

Eugene D. Hahn

Department of Information and Decision Sciences

Salisbury University, USA

Refik Soyer

Department of Decision Sciences

The George Washington University, USA

Probit Versus Logit Models: The Role of the Link Function in the Multivariate Realm

September 18, 2008

Abstract

Current opinion regarding the selection of link function in binary discrete choice models is that the probit and logit links give essentially similar results in terms of fit. This seems to be true for univariate binary choice models; however, for multivariate binary choice models it appears to be an oversimplification. We examine the relationship between link function selection and model fit in two classes of multivariate binary discrete choice models. We find model fit can be improved by the selection of the appropriate link, even in more modestly-sized 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 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 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 for modeling 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). Thus, the selection of the link function tends to be driven more by what we might call contextual factors, such as interpretability, because the models are presumed to be essentially equivalent. For example, one might select the logit link because a log odds ratio is a parameter of interest.

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 choice models. Increasingly, interest has turned to instances where there is more than one binary choice 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 choice data (Stiratelli et al. 1984). Despite these developments, the properties of link functions for binary choice models in the multivariate realm remain largely unexplored. This seems unfortunate, as it turns out that the impact of link function on model fit is affected by the form of the model considered.

In the current paper we address this gap by examining model fit for two families of multivariate binary discrete choice models. In so doing, we take a Bayesian point of view and use fit measures such as Deviance Information Criterion of Spiegelhalter et al. (2002) and marginal likelihoods. In §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 across the response variables using the marginal link functions. Factors that are expected to influence the fit of these models are discussed in §2.2, while measures of model fit are presented in §2.3. Three research propositions are stated in §2.4 and the methods used in the study are described in §3. Our findings are presented in §4 while in §5 we examine link function selection in the context of real-world multivariate binary choice data involving a study of consumer behavior regarding personal finances. We discuss our findings and draw conclusions in §6.

2. Link Function and Model Fit

2.1. Multivariate Binary Choice Models

Many models for multivariate binary choice 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 $|\rho| < 3/\pi^2 \approx .304$. 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 choice 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. Futher, a Q-Q plot (described below) also shows 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

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

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. Figure 1 displays a Q-Q 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 $2/\pi$. Hence, the $t(8)$ distribution provides a quite close approximation to the Logistic $(2/\pi)$ 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. 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 close fit of the $t(8)$, we expect that their formulation would yield essentially equivalent results to the current one. We note here that other symmetric link functions are possible; however, we confine our attention to the probit and logit links since they were the focus of Chambers and Cox and moreover are the most commonly used link functions.

Another frequently used model for multivariate binary choice data is the random effects model. Here, individual-specific terms are introduced to account for heterogeneity at the individual level. In the current context, the random effects model can be written as

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

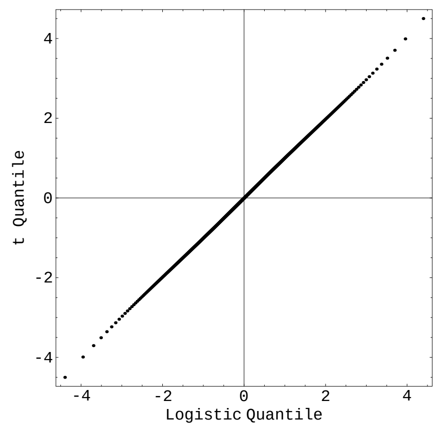

Figure 1: Quantile values of Logistic($2/\pi$) versus $t(8)$ for probabilities from .001 to .999

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 may be more widely used for multivariate binary choice data. Zeger and Karim (1991) provided an early Bayesian development of the model in the context of a Gibbs sampling approach, while Train (2003) provides a contemporary overview of Bayesian and likelihood-based approaches.

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 choice models have not been examined. Here, we examine this issue 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). We also consider 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. Thus, our study involves a $2 \times 2 \times 2 \times 2 \times 2$ factorial examination of model fit in multivariate discrete choice models, as we describe in more detail in §3. 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

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

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 §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,

$$\text{DIC} = \overline{D} + p_D \tag{4}$$

where $\overline{D}$ is the posterior mean of the total deviance and D itself is the sum of the deviance contributions of the individual observations, i.e., $D = \sum_{i=1}^{n} d_i$. Moreover, 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 as a measure of a model's out-of-sample predictive ability. Thus, for comparative purposes we calculate both the DIC values and the log marginal likelihoods for the different models examined here. The log marginal likelihoods can be used to construct Bayes factors. 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 from the theoretical development above.

Research Proposition 1 The presence of extreme independent variable levels in consumer attributes 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 among consumer choices will lead to decreased differences in fit across the two link functions.

In describing this proposition, we may begin by describing the following self-evident statement: all else being equal, any differences in model fit should become more pronounced 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 choice 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 since 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, where consumer heterogeneity is directly modeled, use of the probit link results in model fit that is as good or better than model fit under the logit link.

This 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 choice 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.

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 and 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 in the choice data 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 choice 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 choice models have more modest sample sizes than that 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 among choices: 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

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

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

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

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. Study Results

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. 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 but still well below the threshold. 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.

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.

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

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 difference is 4.0. 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 seem 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.

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 support 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. Application

We also consider the impact of link function selection in the context of real-world consumer choice data. The data comes from the 2001 Survey of Consumer Finances commissioned by the U.S. Federal Reserve and conducted by the National Opinion Research Center (NORC) at the University of Chicago. In the study, 4442 U.S. households drawn from all economic strata were surveyed regarding their personal finances and related personal financial decision-making. Here in order to provide greater comparability with the more moderate sample sizes examined previously we selected a random subset of 500 households. For the analyses reported here, we examined the impact of education on two technologically-oriented behavioral outcomes involving personal finances. In particular, Y_1 was coded 1 if the respondent indicated that he/she or his/her spouse used computer software to help with managing money, and was coded 0 otherwise. Similarly, Y_2 was coded 1 if the respondent indicated that he/she used internet banking to do business with his/her financial institution, and was coded 0 otherwise. The simple Pearson correlation between Y_1 and Y_2 was 0.44. The number of years of respondent education (in terms of grades of school plus years of post-secondary education) was used as the predictor variable. In our $n = 500$ sample, this variable ranged from two years to 17 years, the latter indicating graduate level education beyond the 16 years of primary, secondary, and college education. While in a more detailed analysis we might treat the censoring mechanism occurring at 17 years of education, for the purposes of comparability with the analyses in §4 we do not apply any special treatment to the independent variable.

Table 4 contains the summaries of model coefficients for the consumer survey data. Displayed there are the posterior means for the coefficients as well as the posterior standard deviations in parentheses. The model coefficient results indicate that years of education is positively related to the occurrence of technologically-oriented consumer behavior across models. For example, the results for $\beta_{2,1}$ indicate that years of education are predictive of an increased likelihood for a respondent to utilize computer software for managing finances. Across models, the 95% posterior credible intervals for $\beta_{2,1}$ and $\beta_{2,2}$ exclude the value of zero by a considerable margin.

Table 5 contains the model fit measures for the consumer survey data. For the multivariate link models, we find evidence of differences in fit consistent with the findings of §4. The DICs under the probit and logit links were 951.7 and 948.5 respectively. These results indicate DIC favors the logit model over the probit in the context of the multivariate link formulation as was found previously. Similarly, the log marginal likelihoods for the two respective models were -494.4 and -492.1, indicating a somewhat modest but nonetheless non-trivial support for the logit model in the multivariate link context. Again this is consistent with previous findings. Switching to the random effects context, again we find consistent evidence of considerable differences in fit for the two link functions using DIC. Here, the DICs under the probit and logit links were 833.4 and 842.8 respectively, indicating the fit under the probit model is considerably improved. The respective values of $\overline{D}$ were 617.4 and 629.6 while those of p_D were 216.0 and 213.1. Hence, we can see that the improved

Table 4: Coefficient summary statistics: Consumer survey data

		$\beta_{1,1}$	$\beta_{2,1}$	$\beta_{1,2}$	$\beta_{2,2}$	ρ	τ^{-1}
Multivariate Link	*logit*	-0.85	0.17	-0.89	0.19	0.62	—
		(0.07)	(0.03)	(0.08)	(0.03)	(0.06)	—
	probit	-0.78	0.13	-0.81	0.15	0.64	—
		(0.06)	(0.03)	(0.07)	(0.03)	(0.06)	—
Random Effects	*logit*	-1.44	0.25	-1.56	0.31	—	2.23
		(0.17)	(0.05)	(0.18)	(0.06)	—	(0.61)
	probit	-1.30	0.22	-1.40	0.27	—	1.84
		(0.15)	(0.05)	(0.16)	(0.05)	—	(0.50)

Table 5: Model fit measures: Consumer survey data

		Multivariate Link	*Random Effects*
DIC	*logit*	948.5	842.8
	probit	951.7	833.4
$\overline{D}$	*logit*	943.5	629.6
	probit	946.8	617.4
p_D	*logit*	4.93	213.1
	probit	4.94	216.0
Log marginal likelihood	*logit*	-492.1	-490.0
	probit	-494.4	-490.7

DIC of the probit model is driven by the reduced deviance contribution from $\overline{D}$. The log marginal likelihoods for the two respective models were again essentially equivalent at -490.7 and -490.0, replicating the previous findings.

We have seen in both the study data of §4 and the real-world data of §5 that DIC and Bayes factors lead to differing conclusions about fit in the random effects models. Specifically, the Bayes factors suggest there is little to differentiate the logit and the probit link, while DIC consistently favors the probit link. We examine this issue in more detail. With no loss of generality we consider Y_j for the moment and noting the Bernoulli likelihood,

$$D_{Y_j} = -2 \sum_{i=1}^{n} \big(Y_{i,j} \log p_{i,j} + (1 - Y_{i,j}) \log(1 - p_{i,j}) \big).$$

Since

$$\frac{\partial}{\partial p_{i,j}} = -2 \sum_{i=1}^{n} \left(\frac{Y_{i,j}}{p_{i,j}} - \frac{1 - Y_{i,j}}{1 - p_{i,j}} \right) \tag{5}$$

and

$$\frac{\partial^2}{\partial p_{i,j}} = 2\sum_{i=1}^{n}\left(\frac{Y_{i,j}}{p_{i,j}^2} + \frac{1-Y_{i,j}}{(1-p_{i,j})^2}\right) \tag{6}$$

the deviance gets increasingly large at an increasing rate as p_j diverges from Y_j. Note from our discussion in §2.3 that in the multivariate context we can rewrite (4) as

$$\text{DIC} = \sum_{i=1}^{n}\sum_{j=1}^{J}\overline{d}_{i,j} + p_D. \tag{7}$$

Given the pattern of results in §§4 and 5 we know the sum of the individual deviance contributions in (7) is greater in the context of the logit link than it is in the context of the probit link because the values of p_D are roughly equivalent. This in conjunction with the behavior of the deviance function as described in (5) and (6) implies that values of $p_{i,j}$ that are discrepant tend to be more discrepant under the logit link than under the probit link in random effects models.

We provide here an example of the differences in deviance contributions and values of $p_{i,j}$ across the two link functions. Our illustration involves the respondents with the two largest contributions to the deviance. Respondent 93 was an atypical respondent who with 2 years of formal education nonetheless used computer software to manage money. His/her deviance contribution was 5.9 under the logit model with predicted probabilities of 0.15 and 0.07 for the two outcomes. Under the probit model his/her deviance contribution was 4.9 with predicted probabilities of 0.19 and 0.08 for the two outcomes. While both models were discrepant with regard to $Y_{i,1}$, in this situation the logit model was more extremely discrepant and less correct via $p_{i,1}$ than was the probit by a notable margin. Furthermore, the logit model was more extreme (as well as more correct) with regard to $p_{i,2}$ than was the probit. However, the relative deviance reduction garnered by the logit for $p_{i,2}$ was rather small, as would be expected by both the relative similarity of the probit estimate of $p_{i,2}$ as well as the behavior of the deviance as described in (5) and (6). Similarly, respondent 71 with 4 years of formal education nonetheless used both computer software to manage money and additionally used internet banking. Under the logit model, he/she had a deviance contribution of 4.8. This is because the logit model predicted the probabilities of these behaviors as being 0.52 and 0.34 respectively. The probit model, by contrast, predicted these probabilities as being 0.54 and 0.34, with the respondent having a deviance contribution of 4.4 under the probit. Here, we see the probit was less discrepant (and hence less incorrect) than the logit model with regard to $p_{i,1}$.

We note that for this data set, the probit and the logit models were both approximately equal in their ability to correctly classify a respondent at the binary level using a threshold of 0.5. The logit model correctly classified an observation as a 0 or a 1 approximately 86% of the time for Y_1 and 86% of the time for Y_2. The probit model also correctly classified an observation as a 0 or a 1 approximately 86% of the time for Y_1 and 86% of the time for Y_2. However, as indicated by the deviance quantities and DIC, when the logit predicted probability was discrepant from Y, it tended to be more so than the probit.

6. 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 necessarily carry over to the multivariate realm. In fact, some differences in fit can be found even in small sample sizes. 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 seems to be improved by selecting the probit link as compared to the logit 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, namely 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. We argue that performance considerations should increasingly be considered in link function decisions, particularly since some of the contextual factors influencing these decisions are less relevant given advances in modern computing methods. For example, in an MCMC environment it is easy to specify the odds ratio as a parameter of interest even in a probit model. It is well known that MCMC permits straightforward estimation of functions of parameters, as the relevant function of a parameter can be computed during the MCMC run such that its posterior distribution is available from the output. The odds ratio associated with a probit coefficient is one such function that may easily be computed during a probit run. Hence, there is less reason to strictly prefer the logit link to satisfy this contextual reason alone.

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, at least one such analysis 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

Abe, M. (1995). "A Nonparametric Density Estimation Method for Brand Choice Using Scanner Data." *Marketing Science*, 14, 300–325.

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," pages 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 Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88, 669–679.

Allenby, G. M. and Rossi, P. E. (1991). "Quality Perceptions and Asymmetric Switching between Brands." *Marketing Science*, 10, 185–204.

Ansari, A., Jadedi, K., and Jagpal, S. (2000). "A Hierarchical Bayesian Methodology for Treating Heterogeneity in Structural Equation Models." *Marketing Science*, 19, 328–347.

Ashford, J. R. and Sowden, R. R. (1970). "Multi-variate Probit Analysis." *Biometrics*, 26, 535–546.

Ben-Akiva, M., McFadden, D., et al. (1997). "Modeling Methods for Discrete Choice Analysis." *Marketing Letters*, 8, 273–286.

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

Carlin, B. P., Kass, R. E., Lerch, F. J., and Huguenard, B. R. (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 Cox, D. R. (1967). "Discrimination between Alternative Binary Response Models." *Biometrika*, 54, 573–578.

Chen, M.-H. and Dey, D. K. (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 Greenberg, E. (1998). "Analysis of Multivariate Probit Models." *Biometrika*, 85, 347–361.

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

Edwards, Y. D. and Allenby, G. M. (2003). "Multivariate Analysis of Multiple Response Data." *Journal of Marketing Research*, 40, 321–334.

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

Gelfand, A. E. (1996). "Model Determination using Sampling-Based Methods." In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds., "Markov Chain Monte Carlo in Practice," pages 145–161. Boca Raton, FL: Chapman & Hall.

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

Jadedi, K., Jagpal, H. S., and DeSarbo, W. S. (1997). "Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity." *Marketing Science*, 16, 39–59.

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

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

Lavine, M. and Schervish, M. J. (1999). "Bayes Factors: What They Are and What They Are Not." *The American Statistician*, 53, 119–122.

Lewis, S. M. and Raftery, A. E. (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.

Madigan, D. and Raftery, A. E. (1994). "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window." *Journal of the American Statistical Association*, 89, 1535–1546.

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

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association*, 92, 179–191.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. Chichester, UK: Wiley.

Shively, T. S., Allenby, G. M., and Kohn, R. (2000). "A Nonparametric Approach to Identifying Latent Relationships in Hierarchical Models." *Marketing Science*, 19, 149–162.

Smith, M. D. and Moffatt, P. G. (1999). "Fisher's Information on the Correlation Coefficient in Bivariate Logistic Models." *Australian & New Zealand Journal of Statistics*, 41, 315–330.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). "Bayesian Measures of Model Complexity and Fit (*with discussion*)." *Journal of the Royal Statistical Society: Series B*, 64, 583–639.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). "Random-Effects Models for Serial Observations with Binary Response." *Biometrics*, 40, 961–971.

Swait, J. and Andrews, R. L. (2003). "Enriching Scanner Panel Models with Choice Experiments." *Marketing Science*, 22, 442–460.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Wedel, M., Kamakura, W., et al. (1999). "Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling." *Marketing Letters*, 10, 219–232.

Yang, S., Allenby, G. M., and Fennell, G. (2002). "Modeling Variation in Brand Preference: The Roles of Objective Environment and Motivating Conditions." *Marketing Science*, 21, 14–31.

Zeger, S. L. and Karim, M. R. (1991). "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach." *Journal of the American Statistical Association*, 86, 79–86.