Introducing the BCHOICE Procedure for Bayesian Discrete Choice Models

Allen McDowell and Amy Shi, SAS Institute Inc.

ABSTRACT

The new, experimental BCHOICE procedure in SAS/STAT® 13.1 enables you to perform Bayesian analysis for discrete choice models. PROC BCHOICE fits multinomial logit, nested logit, mixed multinomial logit, multinomial probit, and mixed multinomial probit models. Brief summaries of the properties of the various models are provided along with a series of examples that highlight the capabilities of PROC BCHOICE.

INTRODUCTION

The new BCHOICE procedure, which is introduced in SAS/STAT 13.1 as an experimental procedure, provides Bayesian analysis for discrete choice models. Discrete choice models are used to analyze the choices that are made by decision makers who face a finite and exhaustive set of mutually exclusive alternatives. Under such conditions, the response variable has a multinomial distribution and the analyst attempts to model the relationship between the decision makers' choices and explanatory variables such as the attributes of the available alternatives and the decision makers. Examples are decisions about labor force participation, occupation, educational level, marital status, family size, residential and work location, travel mode, and brands of commodity purchases (McFadden 1981). The BCHOICE procedure was designed specifically for analyzing choice data; it enables you to fit the following types of models:

- multinomial logit
- nested logit
- multinomial probit
- mixed multinomial logit
- mixed multinomial probit

PROC BCHOICE samples from the posterior distributions and produces summary and diagnostic statistics when you specify the model or the priors or both. The BCHOICE procedure's model-building syntax, which includes the CLASS, MODEL, and RANDOM statements, is similar to linear and generalized linear modeling procedures such as the GENMOD, GLM, GLIMMIX, and MIXED procedures. PROC BCHOICE also provides the following features:

- is multithreaded
- uses the following sampling algorithms:
  - Gamerman algorithm
  - random walk Metropolis
  - latent variables via the data augmentation method
  - conjugate sampling
- provides a variety of Markov chain convergence diagnostics
- works with the postprocessing autocall macros that are designed for Bayesian posterior samples
- enables you to save an output data set to contain the posterior samples of all parameters
- creates a new SAS data set that contains random samples from the posterior predictive distribution of the choice probabilities

CHOICE SETS

To better understand the derivation of discrete choice models and how the BCHOICE procedure works, it is useful to have a clear picture of how observations of discrete choices are represented in a SAS data set. Imagine a population of consumers who make choices among three brands of ice cream, A, B, and C. Suppose you are interested in how the choices made are related to the prices of the alternative brands and the incomes of the individuals making the choices. To record the choice made by a particular individual in this example, you need three rows of data. Each of the three rows represents a binary choice for a specific alternative (brand), so the data for the first individual look like Table 1.

Table 1 Choice Set for Individual

ID	Alternative	Choice	Price	Income
1	A	0	2.59	30,000
1	B	1	2.99	30,000
1	C	0	3.29	30,000

The variable **ID** identifies the individual, and thus the choice sets in this example. One or more variables that identify the choice sets is required by PROC BCHOICE. The variable **Alternative** identifies the alternatives. Technically, this variable is not required, but having such a variable provides added flexibility for model specification. The variable **Choice** represents the dependent variable in this example. It is a binary variable that indicates whether a particular alternative is chosen. PROC BCHOICE requires that the response variable indicate the chosen alternative by the value 1 and the unchosen alternatives by the value 0. In this example, the variable **Price** varies across the alternatives but the variable **Income** does not. Variables that vary across alternatives are called alternative-specific attributes. Sociodemographic variables that are constant across alternatives are called individual-specific attributes.

DISCRETE CHOICE MODELS

Discrete choice models are derived under the assumption that the behavioral process that determines how a decision-maker makes a choice can be adequately represented by some mathematical function, $y = h(x, \epsilon)$, which relates the observed outcome of a decision, y, to a set of factors that collectively determine the outcome. The factors that are labeled x are observable by the researcher, and the factors labeled ϵ are not. An observer would perceive the unobserved factors as random influences. Because ϵ is random, the outcome is not deterministic and cannot be predicted exactly. However, if you know the distribution function of the unobserved factors, $f(\epsilon)$, you can derive the probability of any particular outcome as

$$\begin{aligned} P(y|x) &= \text{Prob}(\epsilon \text{ s.t. } h(x,\epsilon) = y) \\ &= \text{Prob}(I[h(x,\epsilon) = y] = 1) \\ &= \int I[h(x,\epsilon) = y] f(\epsilon) d\epsilon \end{aligned}$$

where $I[h(x,\epsilon) = y]$ is an indicator function that takes the value of 1 when the statement in brackets is true and 0 when the statement is false (Train 2009). You can derive a variety of discrete choice models by choosing different specifications for the behavioral process function $h(x,\epsilon)$ and the density function $f(\epsilon)$.

When the choices being made are of an economic nature, the decision makers are usually assumed to behave as though they are attempting to maximize utility subject to either budget or technological constraints. When the indirect utility function (which is a mathematical representation of the maximum utility achievable given the observed attributes) includes random factors, the behavioral process is called a random utility model (RUM); see McFadden (1981) for a formal derivation.

Suppose the utility that individual i obtains from alternative j is

$$u_{ij} = v_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, N, \text{ and } j = 1, \ldots, J$$

where v_{ij} is a nonstochastic function that relates the observed factors to the utility and ϵ_{ij} is a random component that represents the unobserved factors that determine utility. Dictated in part by economic theory and in part by computational feasibility, v_{ij} has historically been specified as

$$v_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}$$

where $\mathbf{x}_{ij}$ is a p-dimensional design vector of observed attribute levels that relate to alternative j and $\boldsymbol{\beta}$ is a corresponding vector of fixed regression coefficients. This specification is linear in the parameters, $\boldsymbol{\beta}$, and represents what is known in the economics literature as a quasilinear indirect utility function.

Decision makers choose the alternative that gives them the greatest utility. Suppose $\mathbf{y}_i$ is the multinomial response for the ith individual. The value y_{ij} takes 1 if the jth component of $\mathbf{u}_i = (u_{i1}, \ldots, u_{iJ})$ is the largest, and 0 otherwise:

$$\begin{aligned} u_{ij} &= \mathbf{x}_{ij}'\boldsymbol{\beta} + \epsilon_{ij} \\ y_{ij} &= \begin{cases} 1 & \text{if } u_{ij} \geq \max(\mathbf{u}_i) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The probability that the individual i chooses alternative j is

$$\begin{aligned} P(y_{ij} = 1|\boldsymbol{\beta}) &= \text{Pr}\ (u_{ij} > u_{ik}\ \forall\ k \neq j) \\ &= \text{Pr}\ (v_{ij} + \epsilon_{ij} > v_{ik} + \epsilon_{ik}\ \forall\ k \neq j) \\ &= \text{Pr}\ (\epsilon_{ik} - \epsilon_{ij} < v_{ij} - v_{ik}\ \forall\ k \neq j) \\ &= \int_{\epsilon} \text{I}(\epsilon_{ik} - \epsilon_{ij} < v_{ij} - v_{ik}\ \forall\ k \neq j) f(\boldsymbol{\epsilon}_i) d\boldsymbol{\epsilon}_i \end{aligned}$$

where I(·) is the indicator function and $f(\boldsymbol{\epsilon}_i)$ denotes the joint density of the random vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})$. The cumulative distribution function $P(y_{ij} = 1|\boldsymbol{\beta})$ is the probability that each random term $(\epsilon_{ik} - \epsilon_{ij})$ is less than the observed quantity $(v_{ij} - v_{ik})$. It is computed as the multidimensional integral over the density of the unobserved portion of utility, $f(\boldsymbol{\epsilon}_i)$. Different discrete choice models are obtained from different specifications of this density.

An important feature of the choice probabilities that are derived from a random utility model that affects model specification and estimation is that the probability depends only on the differences in utility. This is consistent with the fact that the only relevant feature of a utility function is the ordinal ranking of preferences that it provides. Any monotonic transformation of a utility function preserves the ordinal preference rankings of the original utility function (Varian 1978). This implies that utility has no unique origin or scale and that any parameters of the utility function that are related to location or scale cannot be identified and are not estimable. Therefore, when you fit a discrete choice model, you must normalize any constants that enter the utility function and you must normalize the scale of utility. For example, consider the following indirect utility function:

$$u_{ij} = \alpha_j + \mathbf{x}_{ij}'\boldsymbol{\beta} + \epsilon_{ij}\ \forall j$$

where α_j is the main effect that is related to alternative j, $\mathbf{x}_{ij}$ is the design variable vector that is specific to alternative j, and $\boldsymbol{\beta}$ contains the corresponding coefficients of $\mathbf{x}_{ij}$. The main effect α_j is the average effect for alternative j on utility of all factors that are not included in the model. Because only differences in utility affect the choice probabilities, only differences in the alternative-specific constants affect the choice probabilities. Any two models that have the same difference in constants are equivalent. To account for this fact, PROC

BCHOICE automatically normalizes alternative-specific constants by setting the last main effect to 0 when you specify the alternative-specific main fixed effects in the CLASS statement; it also provides an option that enables you to designate which alternative to set to 0. After normalization, the estimates of the main fixed effects are interpreted as the average effect of excluded factors relative to the main effect that is normalized to 0.

Similarly, individual-specific fixed effects such as age, gender, race, and income are constant across alternatives. You can include these types of fixed effects in your model only by specifying them in a way that creates differences in utilities among alternatives. In PROC BCHOICE, you do this by creating interactions between the individual-specific fixed effects and the alternative-specific main effects, one of which has been normalized to 0.

The scale of utility is also irrelevant for choice models. For example, the utility function $u_{ij} = v_{ij} + \epsilon_{ij}$ yields exactly the same choice probabilities as does the utility function $u_{ij} = \lambda v_{ij} + \lambda \epsilon_{ij}$. To account for this fact, you must normalize the scale of utility. But notice that when you multiply the utility function by λ, the variance of the error term becomes $Var(\lambda)\epsilon_{ij} = \lambda^2 Var(\epsilon_{ij})$. So, if you normalize the variance of the error terms, you normalize the scale of utility. PROC BCHOICE automatically normalizes the scale of utility by normalizing the variance of the error terms, but it uses different normalization methods for different types of models.

MODELING HETEROGENEITY IN PREFERENCES

Suppose you assume that preferences are consistent with the quasilinear indirect utility function $u_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij}$ and that the explanatory variables $\mathbf{x}_{ij}$ contain only alternative-specific factors. When you fit a discrete choice model that is derived from such a behavioral process, you are assuming that the preferences of all the individuals in the population can be represented by a common functional form and that all individuals have exactly the same preference rankings over the observed attributes of the alternatives. All variations in their choices are due to the random, unobserved component of the utility function. If this reflected the true state of nature, it would be convenient because these two features satisfy the sufficient conditions for the existence of aggregate demand functions (McFadden 1981). However, empirical evidence rarely, if ever, supports such a view of the world. On the contrary, there is considerable evidence that individuals' preferences are heterogeneous. Although you cannot escape the use of a common functional form to represent individuals' utilities, discrete choice models do provide two methods for incorporating some degree of heterogeneity in preferences into a model. One method is to include individual-specific attributes as fixed effects in your model. The second method, which is not exclusive of the first, is to include random effects in your model. In the statistics and econometrics literature, discrete choice models that include random effects are often called as mixed models.

MIXED MODELS

In mixed models that include both fixed effects and random effects, the utility of individual i from alternative j is written as

$$u_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\boldsymbol{\gamma}_i + \epsilon_{ijt}$$

where $\mathbf{x}_{ij}$ is the fixed design vector for individual i and alternative j, $\boldsymbol{\beta}$ is the vector of fixed coefficients, $\mathbf{z}_{ij}$ is the random design vector for individual i and alternative j, and $\boldsymbol{\gamma}_i$ is the vector of random coefficients for individual i that correspond to $\mathbf{z}_{ij}$.

It is assumed that each $\boldsymbol{\gamma}_i$ is drawn from a superpopulation and this superpopulation is normal, so the prior distribution for $\boldsymbol{\gamma}_i$ is

$$\pi(\boldsymbol{\gamma}_i) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}})$$

Specifying the mean of $f(\boldsymbol{\gamma})$ as $\mathbf{0}$ means that you are assuming either that the random effects are truly centered around 0 or that they have been centered by the fixed effects.

The covariance matrix $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$ characterizes the extent of unobserved heterogeneity among individuals. Large diagonal elements of $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$ indicate substantial heterogeneity. Off-diagonal elements indicate patterns in the evaluation of attribute levels. $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$ is treated as a model hyperparameter, and a prior distribution is specified as

$$\pi(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}) = \text{inverse Wishart}(\nu_0, \mathbf{V}_0)$$

RANDOM COEFFICIENTS MODELS

PROC BCHOICE also supports a hierarchical Bayesian random-effects-only model that is proposed by Allenby and Rossi (1999) and Rossi, Allenby, and McCulloch (2005). In this model, there are no fixed effects. The indirect utility function is specified as

$$u_{ij} = \mathbf{z}'_{ij}\boldsymbol{\gamma}_i + \epsilon_{ij}$$

The prior distributions for $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$ are

$$\begin{aligned}\pi(\boldsymbol{\gamma}_i) &= \text{N}(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}})\\ \pi(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}) &= \text{inverse Wishart}(\nu_0, \mathbf{V}_0)\end{aligned}$$

In this model, $\bar{\boldsymbol{\gamma}}$ is a mean vector of regression coefficients, which models the central locations of the random coefficients' distributions.

As an alternative, you can also fit a random coefficients model that is proposed by Rossi, McCulloch, and Allenby (1996) and Rossi, Allenby, and McCulloch (2005). In this variation, the prior mean of the random coefficients is modeled as a function of individual-specific variables.

MULTINOMIAL LOGIT MODEL

The multinomial logit (MNL) model is derived by assuming that each ϵ_{ij} is independently and identically distributed (iid) with the Type I extreme-value distribution (also known as the Gumbel distribution) so that the cumulative distribution is

$$F(\epsilon_{ij}) = e^{-e^{-\epsilon_{ij}}}$$

This distribution has a variance equal to $\pi^2/6$. By setting the variance of ϵ_{ij} to $\pi^2/6$, the scale of utility is implicitly normalized. Assuming an indirect utility function that is linear in parameters, the choice probability is

$$P(y_{ij} = 1|\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{\sum_{k=1}^{J}\exp(\mathbf{x}'_{ik}\boldsymbol{\beta})}, \quad i = 1, \ldots, N \text{ and } j = 1, \ldots, J$$

where $0 < P(y_{ij} = 1|\boldsymbol{\beta}) < 1 \; \forall \; i, j$ and where $\sum_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\beta}) = 1 \; \forall \; i$.

The likelihood for the MNL model is formed by the product of the N independent multinomial distributions:

$$L(\mathbf{Y}|\boldsymbol{\beta}) = \prod_{i=1}^{N}\prod_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\beta})^{y_{ij}}$$

The MNL model was originally derived by Luce (1959) from an axiom of *independence from irrelevant alternatives* (IIA), which restricts the relative odds of two alternatives to be independent of the attributes. This implies proportionate substitution between alternatives when there are changes in the attributes. Although the IIA restriction is appropriate for some choice situations, there are situations where it is clearly inappropriate. Much of the development of alternative choice models has been motivated, at least in part, by a desire to find models that do not impose the IIA restriction.

Bayesian analysis for the MNL model requires the specification of a prior over the coefficient parameters $\boldsymbol{\beta}$ and computation of the posterior density. When you have no strong prior beliefs about the location of the parameters, it is recommended that diffuse, but proper, priors be used. The BCHOICE procedure currently supports the use of a normal prior on $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}) = \mathrm{N}(\bar{\boldsymbol{\beta}}, \boldsymbol{\Omega}_{\boldsymbol{\beta}})$$

In PROC BCHOICE, the default specification for the prior is $\pi(\boldsymbol{\beta}) = \mathrm{N}(\mathbf{0}, 10^2\mathbf{I})$, but you can specify an option in the MODEL statement to assign the mean and covariance information for the normal prior.

The posterior density of the parameter $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\mathbf{Y}) \propto L(\mathbf{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})$$

PROC BCHOICE uses Markov chain Monte Carlo (MCMC) simulation methods to sample from the posterior distribution. For MNL models, PROC BCHOICE uses the Metropolis-Hastings approach of Gamerman (1997) by default, but you can request that it use the random walk Metropolis algorithm instead.

EXAMPLE: CONJOINT ANALYSIS OF CHOCOLATE CANDY

In this example from Kuhfeld (2010), each of 10 subjects is presented with eight different chocolate candies and asked to choose one. The eight candies consist of the 2^3 combinations of dark or milk chocolate, soft or chewy center, and nuts or no nuts. Each subject sees all eight alternatives and makes one choice. The following statements read the data:

```
title 'Conjoint Analysis of Chocolate Candies';

data Chocolate;
   input Subj Choice Dark Soft Nuts;
   datalines;
1 0 0 0 0
1 0 0 0 1
1 0 0 1 0
1 0 0 1 1
1 1 1 0 0
1 0 1 0 1
1 0 1 1 0
1 0 1 1 1
2 0 0 0 0

   ... more lines ...

10 0 1 1 0
10 0 1 1 1
;
```

The data set contains 10 subjects and 80 observations. Each line of the data represents one alternative in the choice set for each subject. The response variable, **Choice**, indicates the chosen alternative by the value 1 and the unchosen alternatives by the value 0. The variable **Dark** is 1 for dark chocolate and 0 for milk chocolate, the variable **Soft** is 1 for soft center and 0 for chewy center, and the variable **Nuts** is 1 if the candy contains nuts and 0 if it does not contain nuts.

The following statements fit a multinomial logit model:

```
ods graphics on;
proc bchoice data=Chocolate nmc=10000 thin=2 nthreads=8 seed=124;
   class Dark(ref='0') Soft(ref='0') Nuts(ref='0') Subj;
   model Choice = Dark Soft Nuts / choiceset=(Subj) cprior=normal(var=1000);
   preddist nalter=8 outpred=Predout;
run;
```

The PROC BCHOICE statement invokes the procedure, and the DATA= option specifies the input data set Chocs. The NMC= option specifies the number of posterior simulation iterations. The THIN= option controls the thinning of the Markov chain and requests that one of every two samples be kept. PROC BCHOICE is multithreaded, and the NTHREADS= option specifies the number of threads for analytic computations. The SEED= option specifies a seed for the random number generator, which guarantees the reproducibility of the random stream.

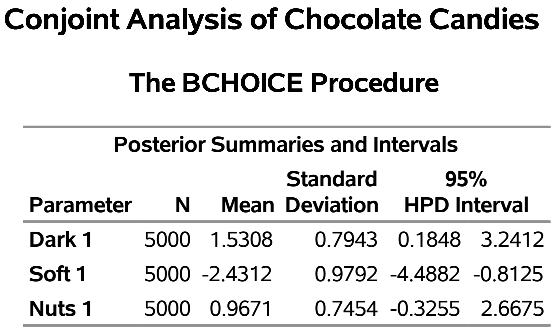
The CLASS statement names the classification variables to be used in the model. The CLASS statement must precede the MODEL statement. The REF= option enables you to specify the reference level of each CLASS variable.

The MODEL statement specifies **Choice** as the response variable and includes **Dark**, **Soft**, and **Nuts** as fixed effects. The CHOICESET= option, which is required, specifies that **Subj** identifies the choice sets. The variables that you specify in the CHOICESET= option must be classification variables that appear in the CLASS statement. The CPRIOR= option specifies the prior distribution for the fixed-effects coefficients as $\mathrm{N}(0, 1000\mathbf{I})$.

The PREDDIST statement creates a new SAS data set that contains random samples from the posterior predictive distribution of the choice probabilities. The NALTER= option specifies the number of alternatives in each choice set. The OUTPRED= option creates an output data set to contain the samples from the posterior predictive distribution of the choice probabilities.

Figure 1 reports posterior summary statistics (posterior means, standard deviations, and highest posterior density (HPD) intervals) for each parameter.

Figure 1 PROC BCHOICE Posterior Summary Statistics



Conjoint Analysis of Chocolate Candies

The BCHOICE Procedure

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
Dark 1	5000	1.5308	0.7943	0.1848	3.2412
Soft 1	5000	-2.4312	0.9792	-4.4882	-0.8125
Nuts 1	5000	0.9671	0.7454	-0.3255	2.6675

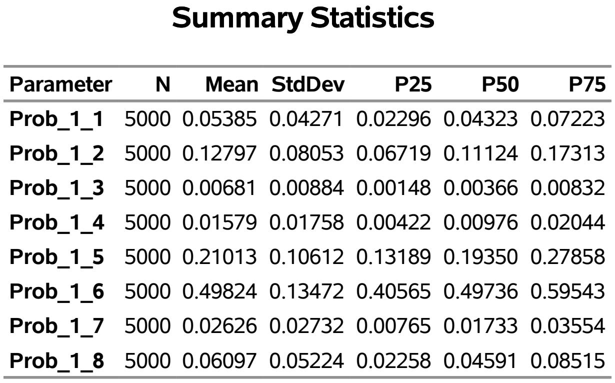
Recall that in this example the variable **Dark** is 1 for dark chocolate and 0 for milk chocolate, the variable **Soft** is 1 for soft center and 0 for chewy center, and the variable **Nuts** is 1 if the candy contains nuts and 0 if it does not contain nuts. So the reported parameter estimates (posterior means) are contrasts with respect to 0. Contrasts are sometimes referred to as “utility part-worths.” The contrast for dark chocolate is 1.5, the contrast for soft center is –2.4, and the contrast for containing nuts is 1.0. A positive contrast implies that an attribute is more favorable; so you conclude that dark chocolate is preferred over milk chocolate, soft centers are less popular than chewy centers, and candies with nuts are more popular than candies without nuts.

You can use SAS autocall macros to analyze the posterior predictive distribution samples. For example, the %POSTSUM macro provides summary statistics.

```
%POSTSUM(data=Predout, var=Prob_1_:);
```

Figure 2 shows the results from using the %POSTSUM macro. In this example, there is only one choice set (which has eight alternatives) for choice probability prediction. This explains the parameter names in the first column of the output, where the first number indexes the choice sets and the second number indexes the alternatives in each choice set. The most preferred chocolate candy is the sixth one, Dark/Chewy/Nuts, which takes about half the market.

Figure 2 PROC BCHOICE Posterior Summary Statistics of Predictive Distribution



Summary Statistics

Parameter	N	Mean	StdDev	P25	P50	P75
Prob_1_1	5000	0.05385	0.04271	0.02296	0.04323	0.07223
Prob_1_2	5000	0.12797	0.08053	0.06719	0.11124	0.17313
Prob_1_3	5000	0.00681	0.00884	0.00148	0.00366	0.00832
Prob_1_4	5000	0.01579	0.01758	0.00422	0.00976	0.02044
Prob_1_5	5000	0.21013	0.10612	0.13189	0.19350	0.27858
Prob_1_6	5000	0.49824	0.13472	0.40565	0.49736	0.59543
Prob_1_7	5000	0.02626	0.02732	0.00765	0.01733	0.03554
Prob_1_8	5000	0.06097	0.05224	0.02258	0.04591	0.08515

NESTED LOGIT MODEL

The nested logit model is derived by making two assumptions: The first is that the set of J alternatives can be partitioned into K nonoverlapping subsets, called nests, such that the IIA assumption holds within each nest and that the IIA assumption does not hold, in general, for alternatives in different nests. The second assumption is that the ϵ_{ij} are jointly distributed as a generalized extreme value (GEV) with a cumulative distribution,

$$F(\epsilon_{ij}) = \exp\left(-\sum_{k=1}^{K} \left(\sum_{j \in S_k} \exp\left(-\frac{\epsilon_{ij}}{\lambda_k} \right) \right)^{\lambda_k} \right)$$

where $S_1, S_2, \ldots, S_K$ are the K nonoverlapping nests. In a nested logit model, the ϵ_{ij} are correlated within nests. If alternatives j and m belong to the same nest, then ϵ_{ij} is correlated with ϵ_{im}. But if any two alternatives are in different nests, the unobserved part of their utility is still independent. The parameter λ_k measures the degree of independence among alternatives in nest k. The higher the value of λ_k, the less correlation there is, but the correlation is actually more complicated than the parameter λ_k. The equation $\lambda_k = 1$ represents no correlation in nest k. If $\lambda_k = 1$ for all nests, the nested logit model reduces to the standard logit model. The λ_k in nest k is often called the log-sum coefficient. The value of λ_k must be positive for the model to be consistent with utility-maximizing behavior. If $\lambda_k \in [0, 1]$ for all k, the model is consistent with utility maximization for all possible values of the explanatory variables. But if $\lambda_k > 1$, the model is consistent only for some range of the covariates but not for all values. A nest that has only one alternative is degenerate, and the λ_k for that nest is not estimable.

The choice probability for alternative $j \in S_k$ has a closed form:

$$P(y_{ij} = 1|\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}/\lambda_k)(\sum_{m \in S_k} \exp(\mathbf{x}_{im}'\boldsymbol{\beta}/\lambda_k))^{\lambda_k - 1}}{\sum_{l=1}^{K}(\sum_{m \in S_l} \exp(\mathbf{x}_{im}'\boldsymbol{\beta}/\lambda_l))^{\lambda_l}}$$

The variance of ϵ_{ij} in nest $j \in S_k$ is $\frac{\pi^2}{6\lambda_k}$. Thus, the model is homoscedastic within nests and potentially heteroscedastic between nests. However, as you can see from the choice probabilities, because the indirect utility function is linear in the parameters $\boldsymbol{\beta}$ and the observable portion of the indirect utility function is scaled by λ_k, the overall scale of utility is normalized.

The likelihood function for the nested logit model is the product of the N multinomial distributions:

$$L(\mathbf{Y}|\boldsymbol{\beta},\boldsymbol{\lambda}) = \prod_{i=1}^{N}\prod_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\beta})^{y_{ij}}$$

For Bayesian estimation, you specify prior distributions for the parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ and $\boldsymbol{\beta}$. Noninformative priors are not ideal for $\boldsymbol{\lambda}$. Flat priors on different versions of the parameter space can yield different posterior distributions. Lahiri and Gao (2002) suggest the following semi-flat priors by using the parameter ϕ for each λ:

$$\pi(\lambda) = \begin{cases} 0 & \text{if } \lambda \le 0 \\ \phi & \text{if } 0 < \lambda < 1 \\ \phi \exp[\frac{\phi}{1-\phi}(1-\lambda)] & \text{if } \lambda \ge 1 \end{cases}$$

PROC BCHOICE uses this prior with a default value of 0.8 for ϕ. You can specify other values for ϕ. PROC BCHOICE currently supports the use of a normal prior for $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}) = \mathrm{N}(\bar{\boldsymbol{\beta}}, \boldsymbol{\Omega}_{\boldsymbol{\beta}})$$

with a default specification of $\pi(\boldsymbol{\beta}) = \mathrm{N}(\mathbf{0}, 10^2\mathbf{I})$. However, there are options that enable you to specify the mean and covariance information for the normal prior.

The posterior density for the parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is $p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda})$.

EXAMPLE: NEST LOGIT MODEL FOR TRAVEL DEMAND

Consider an example of travel demand. People are asked to choose among travel by auto, plane, or public transit (bus or train). The following SAS statements create the data set Travel. The variables **AutoTime**, **PlanTime**, and **TranTime** represent the total travel time that is required to get to a destination by using auto, plane, or public transit, respectively. The variable **Age** represents the age of each individual who is surveyed, and the variable **Chosen** contains each individual's choice of travel mode.

```
data Travel;
input Subject $ Mode $ Choice Age AgeCtr TravTime;
datalines;
1 Auto 0 32 -2 10.0
1 Plane 1 32 -2 4.5
1 Transit 0 32 -2 10.5
2 Auto 1 13 -21 5.5
2 Plane 0 13 -21 4.0
2 Transit 0 13 -21 7.5
3 Auto 0 41 7 4.5

   ... more lines ...

20 Transit 0 35 1 15.5
21 Auto 1 22 -12 1.5
21 Plane 0 22 -12 4.0
21 Transit 0 22 -12 2.0
;
```

In this example, the **AutoTime**, **PlanTime**, and **TranTime** variables apply to the alternatives, whereas **Age** is a characteristic of the individuals. **AgeCtr**, a centered version of **Age**, is created by subtracting the sample's mean age from each individual's age. To study how the choice depends on both the travel time and age of the individuals, you need to incorporate both types of variables.

It seems plausible that auto and public transit might be more similar to each other than either of them is to plane, because the probability of choosing auto and public transit might rise by about the same proportion whenever the option of taking a plane is unavailable. A nested logit model that places auto and public transit in one nest and plane in another nest might seem more reasonable than the standard logit model.

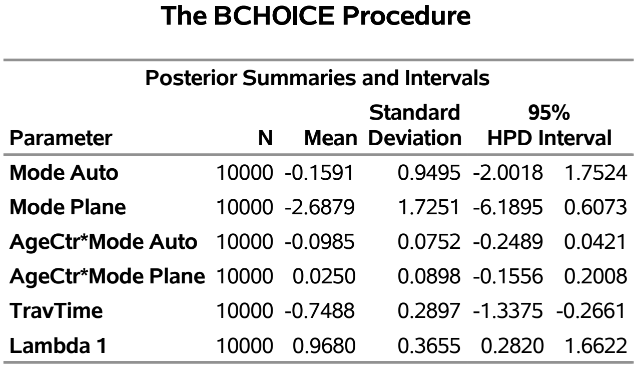
The following SAS statements specify a nested logit model:

```
proc bchoice data=Travel seed=531 nthreads=8 nmc=20000 thin=2;
   class Mode Subject / param=ref order=data;
   model Choice = Mode Mode*AgeCtr TravTime  / choiceset=(Subject)
                                               type=nlogit
                                               nest=(1 2 1);
run;
```

The CLASS statement specifies **Model** and **Subject** as classification variables. The PARAM=REF option requests that reference cell coding be used, and the ORDER=DATA option requests that the sort order for the levels of classification variables be based on the order of their appearance in the data. The MODEL statement specifies **Choice** as the response variable and includes **Mode**, **TravTime**, and the interaction of **Mode** and **AgeCtr** as fixed effects. **AgeCtr** is not estimable by itself because it is the same throughout a choice set for an individual, so you have to create an interaction between **AgeCtr** and **Mode**. The TYPE=NLOGIT option requests that a nested logit model be fit, the NEST= option specifies the nests, and NEST=(1 2 1) specifies that travel alternatives 1 (Auto) and 3 (Transit) are in the first nest and travel alternative 2 (Plane) is in the second nest.

Figure 3 reports posterior summary statistics for each parameter.

Figure 3 PROC BCHOICE Posterior Summary Statistics



The BCHOICE Procedure

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
Mode Auto	10000	-0.1591	0.9495	-2.0018	1.7524
Mode Plane	10000	-2.6879	1.7251	-6.1895	0.6073
AgeCtr*Mode Auto	10000	-0.0985	0.0752	-0.2489	0.0421
AgeCtr*Mode Plane	10000	0.0250	0.0898	-0.1556	0.2008
TravTime	10000	-0.7488	0.2897	-1.3375	-0.2661
Lambda 1	10000	0.9680	0.3655	0.2820	1.6622

The parameter estimate for **Mode Auto** reflects the part-worth of **Auto** for an individual of mean age (34 years), whereas the parameter estimate for **Mode Plane** is the part-worth of **Plane** for an individual of mean age. There are two interaction effects: the first is the effect of a one-unit change in age on the probability of choosing **Auto** over **Transit**, and the second is the effect of a one-unit change in age on the probability of choosing **Plane** over **Transit**.

There are two alternatives in the first nest and one alternative in the second nest. A nest that has only one alternative is said to be degenerate, and its λ is not estimable. That is why there is only one λ estimate in the output. The estimate of λ for nest 1 is 0.97. The proximity of this value to 1 indicates that there might be some, but not much, correlation between alternatives 1 (Auto) and 3 (Transit).

MULTINOMIAL PROBIT MODEL

The multinomial probit (MNP) model is derived by assuming that the unobserved components of utility, $\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ})$, have a multivariate normal (MVN) distribution with a mean vector of $\mathbf{0}$ and a covariance matrix $\boldsymbol{\Sigma}$. By default, PROC BCHOICE estimates a full covariance matrix, which can accommodate any pattern of correlation and heteroscedasticity. However, you can use the COVTYPE=VC option in the MODEL statement to fit a variance components model, which restricts the off-diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$ to equal 0, thus reducing the number of parameters that are estimated.

Assuming an indirect utility function that is linear in parameters, the choice probability is

$$\begin{aligned} P(y_{ij} = 1|\boldsymbol{\beta}) &= \text{Prob}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij} > \mathbf{x}'_{ik}\boldsymbol{\beta} + \epsilon_{ik}) \ \ \forall j \neq k \\ &= \int \text{I}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij} > \mathbf{x}'_{ik}\boldsymbol{\beta} + \epsilon_{ik} \ \ \forall j \neq k)\phi(\boldsymbol{\epsilon}_i)d\boldsymbol{\epsilon}_i \end{aligned}$$

This probability does not have a closed form.

The likelihood for the MNP model is represented symbolically as

$$L(\mathbf{Y}|\boldsymbol{\beta}) = \prod_{i=1}^{N}\prod_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\beta})^{y_{ij}}$$

Probit models require normalization with respect to both location and scale. The solution to the location shift is differencing with respect to the last alternative in each choice set. That is, suppose you take the differences against the last alternative in the choice set and you define the following variables:

$$\begin{aligned} \tilde{\mathbf{x}}_{ij} &= \mathbf{x}_{ij} - \mathbf{x}_{iJ} \\ \tilde{\epsilon}_{ij} &= \epsilon_{ij} - \epsilon_{iJ} \\ w_{ij} &= \tilde{\mathbf{x}}'_{ij}\boldsymbol{\beta} + \tilde{\epsilon}_{ij} \end{aligned}$$

where $\tilde{\boldsymbol{\epsilon}}_i \sim \text{N}(0, \tilde{\boldsymbol{\Sigma}}), \ \ i = 1, \ldots, N$ and $j = 1 \ldots, J - 1$, and $\tilde{\boldsymbol{\Sigma}}$ is the $(J-1) \times (J-1)$ covariance matrix of the vector of error differences.

Then

$$y_{ij} = \begin{cases} 1 & \text{if } \ w_{ij} \geq \max(0, \mathbf{w}_{i,-j}) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{w}_{i,-j} = (w_{i1}, \ldots, w_{i(j-1)}, w_{i(j+1)}, \ldots, w_{i(J-1)})$.

A scale shift problem still remains because the parameters $(c\boldsymbol{\beta}, c^2\tilde{\boldsymbol{\Sigma}})$ for any constant $c > 0$ are equivalent to $(\boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}})$. One solution to the scaling problem is to normalize the parameters with respect to one of the diagonal elements of the covariance matrix, $\tilde{\boldsymbol{\Sigma}}$. PROC BCHOICE normalizes with respect to the first diagonal entry of $\tilde{\boldsymbol{\Sigma}}$ and reports $(\boldsymbol{\beta}/\sqrt{\sigma_{11}}, \tilde{\boldsymbol{\Sigma}}/\sigma_{11})$ at each draw. Therefore, σ_{11} is always equal to 1 in the BCHOICE procedure's output.

A normal prior is used for $\boldsymbol{\beta}$, and an inverse Wishart prior is used for $\tilde{\boldsymbol{\Sigma}}$:

$$\begin{aligned} \pi(\boldsymbol{\beta}) &= \text{N}(\bar{\boldsymbol{\beta}}, \boldsymbol{\Omega}_{\boldsymbol{\beta}}) \\ \pi(\tilde{\boldsymbol{\Sigma}}) &= \text{inverse Wishart}(\nu, \mathbf{V}) \end{aligned}$$

PROC BCHOICE uses an algorithm proposed by McCulloch and Rossi (1994), which is a multivariate version of the probit regression algorithm of Albert and Chib (1993). The algorithm consists of a Gibbs sampler that is based on a Markov chain that draws directly from the exact posteriors of the MNP model. This approach avoids direct evaluation of the likelihood. Sampling is carried out consecutively from the following three groups of conditional posterior distributions:

$$\begin{aligned} &p(w_{ij}|\mathbf{w}_{i,-j}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \mathbf{Y}), \quad i = 1, \ldots, N \ \text{ and } \ j = 1, \ldots, J-1 \\ &p(\boldsymbol{\beta}|\mathbf{W}, \tilde{\boldsymbol{\Sigma}}, \mathbf{Y}) \\ &p(\tilde{\boldsymbol{\Sigma}}|\mathbf{W}, \boldsymbol{\beta}, \mathbf{Y}) \end{aligned}$$

where **W** is obtained by stacking all $\mathbf{w}_i$. All three groups of conditional distributions have closed forms that are easily drawn from. The $p(w_{ij}|\mathbf{w}_{i,-j}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \mathbf{Y})$ are all truncated normal distributions, $p(\boldsymbol{\beta}|\mathbf{W}, \tilde{\boldsymbol{\Sigma}}, \mathbf{Y})$ is a regular multivariate normal distribution, and $p(\tilde{\boldsymbol{\Sigma}}|\mathbf{W}, \boldsymbol{\beta}, \mathbf{Y})$ is an inverse Wishart distribution.

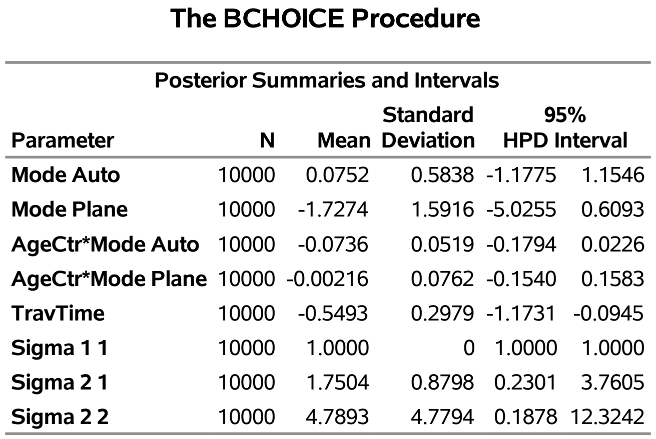
EXAMPLE: MULTINOMIAL PROBIT MODEL FOR TRAVEL DEMAND

This example uses the travel demand data from the previous example. The multinomial probit model completely relaxes the IIA restriction and fits a full covariance matrix for the random component of the utility function. You specify a multinomial probit model by using the TYPE=PROBIT option in the MODEL statement:

```
proc bchoice data=Travel seed=725 nthreads=8 nmc=20000 thin=2;
   class Mode Subject / param=ref order=data;
   model Choice = Mode Mode*AgeCtr TravTime  /
         choiceset=(Subject) type=probit;
run;
```

Figure 4 shows the posterior summary statistics. Although the fixed-effects parameter estimates show numerical differences compared to the nested logit model, the estimates are of similar magnitude and algebraic sign. The MNP model's covariance estimates indicate that there is heteroscedasticity. There is also evidence of correlation between the alternatives Auto and Plane.

Figure 4 PROC BCHOICE Posterior Summary Statistics



The BCHOICE Procedure

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
Mode Auto	10000	0.0752	0.5838	-1.1775	1.1546
Mode Plane	10000	-1.7274	1.5916	-5.0255	0.6093
AgeCtr*Mode Auto	10000	-0.0736	0.0519	-0.1794	0.0226
AgeCtr*Mode Plane	10000	-0.00216	0.0762	-0.1540	0.1583
TravTime	10000	-0.5493	0.2979	-1.1731	-0.0945
Sigma 1 1	10000	1.0000	0	1.0000	1.0000
Sigma 2 1	10000	1.7504	0.8798	0.2301	3.7605
Sigma 2 2	10000	4.7893	4.7794	0.1878	12.3242

MIXED MULTINOMIAL LOGIT MODELS

The mixed multinomial logit (MMNL) model, like the MNL model, assumes that each ϵ_{ij} is independently and identically distributed (iid) with the Type I extreme-value distribution. The indirect utility function is written as

$$u_{ijt} = \mathbf{x}_{ijt}'\boldsymbol{\beta} + \mathbf{z}_{ijt}'\boldsymbol{\gamma}_i + \epsilon_{ijt}$$

where $\boldsymbol{\gamma}_i \sim N(0, \Omega_{\boldsymbol{\gamma}})$.

The probability of person i's observed choices, conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_i$, is

$$P(y_{ij} = 1|\boldsymbol{\beta}, \boldsymbol{\gamma}_i) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}) + \mathbf{z}_{ij}'\boldsymbol{\gamma}}{\sum_{j=1}^{J} \exp(\mathbf{x}_{ij}'\boldsymbol{\beta}) + \mathbf{z}_{ij}'\boldsymbol{\gamma}}$$

The conditional likelihood for the MMNL model is

$$L(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\beta}, \boldsymbol{\gamma}_i)^{y_{ij}}$$

For Bayesian estimation, you specify prior distributions for the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_i$, and $\Omega_{\boldsymbol{\gamma}}$. PROC BCHOICE uses the following priors:

$$\pi(\boldsymbol{\beta}) = \mathrm{N}(\bar{\boldsymbol{\beta}}, \boldsymbol{\Omega}_{\boldsymbol{\beta}})$$

In PROC BCHOICE, the default specification for the prior is $\pi(\boldsymbol{\beta}) = \mathrm{N}(\mathbf{0}, 10^2\mathbf{I})$, but you can specify an option in the MODEL statement to assign the mean and covariance information for the normal prior.

The prior for $\boldsymbol{\gamma}_i$ is

$$\pi(\boldsymbol{\gamma}_i) = \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}})$$

The hyperprior for $\Omega_{\boldsymbol{\gamma}}$ is

$$\pi(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}) = \text{inverse Wishart}(\nu_0, \mathbf{V}_0)$$

The parameter ν_0 specifies the degrees of freedom of the inverse Wishart distribution; the default value in PROC BCHOICE is the dimension of the covariance matrix of the random effects plus 3. The scale parameter of the inverse Wishart distribution, $\mathbf{V}_0$, is specified as $b\mathbf{I}$, where $\mathbf{I}$ is the identity matrix; the default in PROC BCHOICE is the dimension of the covariance matrix of the random effects plus 3. Options in the RANDOM statement enable you to specify both the degrees of freedom and the parameter b.

PROC BCHOICE samples from the following conditional posterior distributions:

$$\begin{aligned} &p(\boldsymbol{\beta}|\boldsymbol{\gamma}_i, \mathbf{Y}) \\ &p(\boldsymbol{\gamma}_i|\boldsymbol{\beta}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}, \mathbf{Y}) \quad i = 1, \ldots, N \\ &p(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_i, \mathbf{Y}) \end{aligned}$$

By default, PROC BCHOICE uses the Gamerman algorithm (Gamerman 1997), but you can use the ALGORITHM= option in the PROC BCHOICE statement to request that PROC BCHOICE use the random walk Metropolis algorithm with a normal proposal distribution instead.

MULTINOMIAL LOGIT RANDOM COEFFICIENTS MODEL

The multinomial logit random coefficients model assumes an indirect utility function of the form

$$u_{ij} = \mathbf{z}_{ij}'\boldsymbol{\gamma}_i + \epsilon_{ij}$$

with $\epsilon_{ij} \sim$ iid extreme value and $\boldsymbol{\gamma}_i \sim \mathbf{N}(\bar{\boldsymbol{\gamma}}, \Omega_{\boldsymbol{\gamma}})$.

The probability of person i's observed choices, conditional on $\boldsymbol{\gamma}_i$, is

$$P(y_{ij} = 1|\boldsymbol{\gamma}_i) = \frac{\exp(\mathbf{z}_{ij}'\boldsymbol{\gamma}_i)}{\sum_{j=1}^{J} \exp(\mathbf{z}_{ij}'\boldsymbol{\gamma}_i)}$$

The conditional likelihood for the random coefficients model is

$$L(\mathbf{Y}|\bar{\boldsymbol{\gamma}}, \Omega_{\boldsymbol{\gamma}}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\gamma}_i)^{y_{ij}}$$

For Bayesian estimation, you specify prior distributions for the parameters $\boldsymbol{\gamma}_i$, $\bar{\boldsymbol{\gamma}}$, and $\Omega_{\boldsymbol{\gamma}}$. PROC BCHOICE assigns the following priors:

$$\begin{aligned}
\pi(\boldsymbol{\gamma}_i) &= \text{N}(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}) \\
\pi(\bar{\boldsymbol{\gamma}}) &= \text{N}(\mathbf{0}, 100\mathbf{I}) \\
\pi(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}) &= \text{inverse Wishart}(\nu_0, \mathbf{V}_0)
\end{aligned}$$

PROC BCHOICE samples from the following conditional posterior distributions:

$$\begin{aligned}
& p(\boldsymbol{\gamma}_i|\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}, \mathbf{Y}) \quad i = 1, \ldots, N \\
& p(\bar{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_i, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}) \\
& p(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_i, \bar{\boldsymbol{\gamma}})
\end{aligned}$$

There is no closed form for $p(\boldsymbol{\gamma}_i|\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}, \mathbf{Y})$. By default, PROC BCHOICE samples by using the Gamerman algorithm, but you can choose to use the random walk Metropolis algorithm instead. Both $p(\bar{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_i, \boldsymbol{\Omega}_{\boldsymbol{\gamma}})$ and $p(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_i, \bar{\boldsymbol{\gamma}})$ have direct sampling distributions: $p(\bar{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_i, \boldsymbol{\Omega}_{\boldsymbol{\gamma}})$ has a normal distribution with a mean of $\sum_{i=1}^{N} \boldsymbol{\gamma}_i/N$ and a covariance of $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}/N$; and $p(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_i, \bar{\boldsymbol{\gamma}})$ is an inverse Wishart$(\nu_0 + N, \mathbf{V}_0 + S)$, where $S = \sum_{i=1}^{N}(\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}})'/N$.

EXAMPLE: RANDOM COEFFICIENTS MODEL FOR MARGARINE SCANNER PANEL DATA

Rossi, Allenby, and McCulloch (2005) studied scanner panel data about purchases of margarine. The data were first analyzed in Allenby and Rossi (1991) and are about purchases of 10 brands of margarine. This example considers a subset of data about six margarine brands: Parkay stick, Blue Bonnet stick, Fleischmann's stick, a house-brand stick, a generic stick, and Shedd's Spread tub. There are 313 households, which made a total of 3,405 purchases.

The data set, which is called Sashelp.Margarin, comes from the Sashelp library and includes the following variables: **HouseID**, **Set**, **Choice**, **Brand**, **LogPrice**, **LogInc**, and **FamSize**. The variable **HouseID** represents the household ID. Each household made at least five purchases, which are defined by **Set**. The variable **Choice** represents the choice made among the six margarine brands for each purchase or choice set. The variable **Brand** has the value PPK for Parkay stick, PBB for Blue Bonnet stick, PFL for Fleischmann's stick, PHse for the house brand stick, PGen for the generic stick, and PSS for Shedd's Spread tub. The variable **LogPrice** is the logarithm of the product price. The variables **LogInc** and variable **FamSize** provide information about household income and family size, respectively.

The following statements fit the random-effects-only logit model by using random walk Metropolis sampling as suggested in Rossi, Allenby, and McCulloch (2005):

```
proc bchoice data=Sashelp.Margarin seed=123 nmc=20000 thin=4 alg=rwm nthreads=8;
   class Brand(ref='PPk') HouseID Set;
   model Choice = / choiceset=(HouseID Set);
   random  Brand LogPrice / subject=HouseID remean=(LogInc FamSize) type=un;
run;
```

The ALG=RWM option in the PROC BCHOICE statement requests the random walk Metropolis sampling algorithm, the NMC=20000 option runs the chain for 20,000 iterations, and the THIN=4 option keeps one of every four samples. The NTHREADS= requests that eight threads be used. The CLASS statement requests that **Brand**, **HouseID**, and **Set** be treated as categorical variables. The MODEL statement specifies

Choice as the response variable but includes no fixed effects. The CHOICESET= option specifies that the combination of **HouseID** and **Set** identifies the choice sets. The RANDOM statement requests that **Brand** and **LogPrice** be included as random effects in the model. The REMEAN=(LOGINC FAMSIZE) option requests estimation of the nonzero mean of the random effects, which is a function of household income and family size. The TYPE=UN option specifies an unstructured covariance matrix for the random effects, thus providing a mechanism for estimating the correlation between the random effects.

Figure 5 displays the posterior summary statistics for the means and covariances of the random coefficients.

Figure 5 Posterior Summary Statistics

The BCHOICE Procedure

Posterior Summaries and Intervals

			Standard	**95%**	
Parameter	**N**	**Mean**	**Deviation**	**HPD Interval**	
REMean Brand PBB	5000	-1.1848	0.6264	-2.3604	0.0984
REMean Brand PFl	5000	-3.2743	1.9054	-6.8877	0.4810
REMean Brand PGen	5000	-5.0670	1.2463	-7.6215	-2.7320
REMean Brand PHse	5000	-3.2251	0.9154	-5.0595	-1.4794
REMean Brand PSS	5000	-0.0333	1.2299	-2.4706	2.3341
REMean LogPrice	5000	-3.3441	0.9011	-5.1967	-1.6316
REMean Brand PBB LogInc	5000	0.0571	0.2060	-0.3370	0.4696
REMean Brand PFl LogInc	5000	0.7307	0.6466	-0.5474	1.9757
REMean Brand PGen LogInc	5000	-0.5484	0.4142	-1.3782	0.2394
REMean Brand PHse LogInc	5000	0.0279	0.3028	-0.5645	0.6040
REMean Brand PSS LogInc	5000	-0.5929	0.4219	-1.3941	0.2429
REMean LogPrice LogInc	5000	-0.3242	0.3106	-0.9570	0.2485
REMean Brand PBB FamSize	5000	-0.0339	0.0966	-0.2241	0.1590
REMean Brand PFl FamSize	5000	-0.7220	0.3148	-1.3473	-0.1178
REMean Brand PGen FamSize	5000	0.5940	0.1862	0.2286	0.9572
REMean Brand PHse FamSize	5000	0.2313	0.1373	-0.0302	0.5093
REMean Brand PSS FamSize	5000	0.0484	0.2019	-0.3592	0.4298
REMean LogPrice FamSize	5000	0.1166	0.1224	-0.1097	0.3707
RECov Brand PBB, Brand PBB	5000	2.1932	0.3785	1.4982	2.9535
RECov Brand PFl, Brand PBB	5000	2.1611	0.9137	0.4284	4.0035
RECov Brand PFl, Brand PFl	5000	12.8291	3.4529	6.5018	19.4713
RECov Brand PGen, Brand PBB	5000	2.0479	0.5618	0.9765	3.1469
RECov Brand PGen, Brand PFl	5000	1.5670	1.8388	-2.0842	5.1846
RECov Brand PGen, Brand PGen	5000	8.5357	1.5046	5.6061	11.4962
RECov Brand PHse, Brand PBB	5000	1.5707	0.4456	0.7194	2.4650
RECov Brand PHse, Brand PFl	5000	2.5443	1.4105	-0.0934	5.4436
RECov Brand PHse, Brand PGen	5000	5.8385	0.9710	4.0326	7.7937
RECov Brand PHse, Brand PHse	5000	5.5638	0.8312	4.0915	7.2899
RECov Brand PSS, Brand PBB	5000	1.2293	0.6139	0.0682	2.4178
RECov Brand PSS, Brand PFl	5000	0.7676	1.7915	-2.7169	4.2290
RECov Brand PSS, Brand PGen	5000	5.1962	1.2905	2.8378	7.8132
RECov Brand PSS, Brand PHse	5000	3.6687	0.8812	1.8585	5.3381
RECov Brand PSS, Brand PSS	5000	8.9329	1.8437	5.8489	12.7281
RECov LogPrice, Brand PBB	5000	-0.2136	0.3379	-0.8814	0.4438
RECov LogPrice, Brand PFl	5000	2.1600	0.8909	0.3472	3.8789
RECov LogPrice, Brand PGen	5000	-1.1061	0.6575	-2.4348	0.1330
RECov LogPrice, Brand PHse	5000	-0.4502	0.5378	-1.4680	0.6476
RECov LogPrice, Brand PSS	5000	0.2339	0.7020	-1.2893	1.5310
RECov LogPrice, LogPrice	5000	2.1049	0.4866	1.1700	3.0405

Table 2 collects the posterior means and standard deviations that are shown in Figure 5. The first column corresponds to the parameters that are specified in the model, namely the brand and the price. The third column shows the average part-worths of each brand (versus the brand Parkay stick) and the price at **LogInc**=0 and **FamSize**=0. The LogInc and FamSize columns list the modifying effects on the preference for each brand and price by household income and family size, respectively. Larger families show more interest in the generic and house brands and tend to stay away from the Fleischmann's brand. For example, consider the part-worth estimates for Fleischmann's. The posterior mean for **REMean Brand PFI FamSize** (the Fleischmann's row and the **Famsize** column) is –0.76 with a standard deviation of 0.32, meaning that an additional unit increase in family size is associated with a reduction of 0.76 in the estimated part-worth for Fleischmann's. In general, the demographics of households are only weakly associated with preference for brand and price. These results are in good agreement with those of Rossi, Allenby, and McCulloch (2005).

Table 2 Posterior Means and Standard Deviations

Parameter		Intercept	LogInc	FamSize
Blue Bonnet	Name	REMean Brand PBB	REMean Brand PBB LogInc	REMean Brand PBB FamSize
	Mean	–1.18	0.06	–0.03
	Std	0.61	0.20	0.10
Fleischmann's	Name	REMean Brand PFI	REMean Brand PFI LogInc	REMean Brand PFI FamSize
	Mean	–3.51	0.83	–0.76
	Std	2.02	0.66	0.32
Generic	Name	REMean Brand PGen	REMean Brand PGen LogInc	REMean Brand PGen FamSize
	Mean	–4.98	–0.55	0.58
	Std	1.17	0.39	0.18
House	Name	REMean Brand PHse	REMean Brand PHse LogInc	REMean Brand PHse FamSize
	Mean	–3.23	0.03	0.23
	Std	0.90	0.30	0.14
Shedd's Spread	Name	REMean Brand PSS	REMean Brand PSS LogInc	REMean Brand PSS FamSize
	Mean	–0.02	–0.59	–0.04
	Std	1.22	0.42	0.20
LogPrice	Name	REMean LogPrice	REMean LogPrice LogInc	REMean LogPrice FamSize
	Mean	–3.32	–0.32	0.11
	Std	0.85	0.30	0.13

You can obtain the utilities of households that have any income levels and sizes. For example, the average part-worth of the Fleischmann's brand for a household that has **LogInc**=3.1 (income) and **FamSize**=3 (family size) is $-3.51 + 0.83 \times 3.1 - 0.76 \times 3 = -3.22$. You can similarly obtain part-worths for all other brands and compare their popularity among average households.

The posterior means and standard deviations of the covariance matrix of the random coefficients are displayed by parameters that are labeled "RECov Brand PBB, Brand PBB," "RECov Brand PFI, Brand PBB," and so on. Some of the diagonal terms are fairly large, indicating that there is quite a bit of heterogeneity among households in margarine brand preference and price sensitivity. The covariance between the generic and house brands, "RECov Brand PHse, Brand PGen," is fairly large, suggesting that household preferences for these two brands are highly correlated.

MIXED MULTINOMIAL PROBIT MODELS

The mixed multinomial probit (MMNP) model is derived by assuming that the unobserved components of utility, $\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ})$, have a multivariate normal (MVN) distribution with a mean vector of $\mathbf{0}$ and a covariance matrix $\boldsymbol{\Sigma}$. The indirect utility function is written as

$$u_{ijt} = \mathbf{x}_{ijt}'\boldsymbol{\beta} + \mathbf{z}_{ijt}'\boldsymbol{\gamma}_i + \epsilon_{ijt}$$

where $\boldsymbol{\gamma}_i \sim N(0, \Omega_{\boldsymbol{\gamma}})$.

The probability of person i's observed choices, conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, is

$$\begin{aligned} P(y_{ij} = 1|\boldsymbol{\beta}, \boldsymbol{\gamma}_i) &= \text{Prob}(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\boldsymbol{\gamma}_i + \epsilon_{ij} > \mathbf{x}_{ik}'\boldsymbol{\beta} + \mathbf{z}_{ik}'\boldsymbol{\gamma}_i + \epsilon_{ik}) \ \ \forall k \neq j \\ &= \int \text{I}(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\boldsymbol{\gamma}_i + \epsilon_{ij} > \mathbf{x}_{ik}'\boldsymbol{\beta} + \mathbf{z}_{ik}'\boldsymbol{\gamma}_i + \epsilon_{ik} \ \ \forall k \neq j)\phi(\boldsymbol{\epsilon}_i)d\boldsymbol{\epsilon}_i \end{aligned}$$

This probability does not have a closed form.

The conditional likelihood for the MNP model is represented symbolically as

$$L(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}_i) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\beta}, \boldsymbol{\gamma}_i)^{y_{ij}}$$

The probit model with random effects has the following parameters: the fixed-coefficients parameters $\boldsymbol{\beta}$, the covariance parameters for the error differences $\tilde{\boldsymbol{\Sigma}}$, the random-coefficients parameters $\boldsymbol{\gamma}_i$, and the covariance parameters for the random coefficients $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$. It has extra parameters $(\boldsymbol{\gamma}_i, \boldsymbol{\Omega}_{\boldsymbol{\gamma}})$ in addition to $(\boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}})$ in a fixed-effects-only model. The MMNP model requires the same normalization as the MNP model with respect to both location and scale, and the method used is the same as described in the section "MULTINOMIAL PROBIT MODEL" on page 10.

For Bayesian estimation, you specify the following prior distributions:

$$\begin{aligned} \pi(\boldsymbol{\beta}) &= \text{N}(\bar{\boldsymbol{\beta}}, \boldsymbol{\Omega}_{\boldsymbol{\beta}}) \\ \pi(\tilde{\boldsymbol{\Sigma}}) &= \text{inverse Wishart}(\nu, \mathbf{V}) \\ \pi(\boldsymbol{\gamma}_i) &= N(0, \Omega_{\boldsymbol{\gamma}}) \\ \pi(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}) &= \text{inverse Wishart}(a, b\mathbf{I}) \end{aligned}$$

PROC BCHOICE samples from the following conditional posterior distributions:

$$\begin{aligned} & p(w_{ij}|\mathbf{w}_{i,-j}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i, \mathbf{Y}) \quad i = 1, \ldots, N \ \text{ and } \ j = 1, \ldots, J-1 \\ & p(\boldsymbol{\beta}|\mathbf{W}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i, \mathbf{Y}) \\ & p(\boldsymbol{\gamma}_i|\mathbf{W}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}, \mathbf{Y}) \quad i = 1, \ldots, N \\ & p(\tilde{\boldsymbol{\Sigma}}|\mathbf{W}, \boldsymbol{\beta}, \mathbf{Y}) \\ & p(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}|\mathbf{W}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i, \mathbf{Y}) \end{aligned}$$

All the groups of conditional distributions have closed forms that are easily drawn from: $p(w_{ij}|\mathbf{w}_{i,-j}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i, \mathbf{Y})$ is a truncated normal distribution, $p(\boldsymbol{\beta}|\mathbf{W}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i, \mathbf{Y})$ and $p(\boldsymbol{\gamma}_i|\mathbf{W}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}, \mathbf{Y})$ are normal distributions, and $p(\tilde{\boldsymbol{\Sigma}}|\mathbf{W}, \boldsymbol{\beta}, \mathbf{Y})$ and $p(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}|\mathbf{W}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i, \mathbf{Y})$ are inverse Wishart distributions. For more information, see McCulloch and Rossi (1994).

MULTINOMIAL PROBIT RANDOM COEFFICIENTS MODEL

The multinomial probit random coefficients model assumes an indirect utility function of the form

$$u_{ij} = \mathbf{z}_{ij}'\boldsymbol{\gamma}_i + \epsilon_{ij}$$

with $\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ}) \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$.

The probability of person i's observed choices, conditional on $\boldsymbol{\gamma}_i$, is

$$\begin{aligned} P(y_{ij} = 1|\boldsymbol{\gamma}_i) &= \text{Prob}(\mathbf{z}_{ij}'\boldsymbol{\gamma}_i + \epsilon_{ij} > \mathbf{z}_{ik}'\boldsymbol{\gamma}_i + \epsilon_{ik}) \;\; \forall k \neq j \\ &= \int \text{I}(\mathbf{z}_{ij}'\boldsymbol{\gamma}_i + \epsilon_{ij} > \mathbf{z}_{ik}'\boldsymbol{\gamma}_i + \epsilon_{ik} \;\; \forall k \neq j)\phi(\boldsymbol{\epsilon}_i)d\boldsymbol{\epsilon}_i \end{aligned}$$

This probability does not have a closed form.

The conditional likelihood for the MNP model is represented symbolically as

$$L(\mathbf{Y}|\boldsymbol{\gamma}) = \prod_{i=1}^{N}\prod_{j=1}^{J} P(y_{ij} = 1|\boldsymbol{\gamma}_i)^{y_{ij}}$$

For Bayesian estimation, you specify the following prior distributions:

$$\begin{aligned} \pi(\tilde{\boldsymbol{\Sigma}}) &= \text{inverse Wishart}(\nu, \mathbf{V}) \\ \pi(\boldsymbol{\gamma}_i) &= N(0, \Omega_{\boldsymbol{\gamma}}) \\ \pi(\bar{\boldsymbol{\gamma}}) &= \text{N}(\mathbf{0}, 100\mathbf{I}) \\ \pi(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}) &= \text{inverse Wishart}(a, b\mathbf{I}) \end{aligned}$$

PROC BCHOICE samples from the following conditional posterior distributions:

$$\begin{aligned} &p(w_{ij}|\mathbf{w}_{i,-j}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i, \mathbf{Y}) \quad i = 1, \ldots, N \;\text{ and }\; j = 1, \ldots, J-1 \\ &p(\boldsymbol{\gamma}_i|\mathbf{W}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}, \mathbf{Y}) \quad i = 1, \ldots, N \\ &p(\bar{\boldsymbol{\gamma}}|\boldsymbol{\gamma}_i, \boldsymbol{\Omega}_{\boldsymbol{\gamma}}) \\ &p(\tilde{\boldsymbol{\Sigma}}|\mathbf{W}, \boldsymbol{\beta}, \mathbf{Y}) \\ &p(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}|\mathbf{W}, \boldsymbol{\beta}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i, \mathbf{Y}) \end{aligned}$$

EXAMPLE: MIXED MULTINOMIAL PROBIT MODEL FOR PEANUT BUTTER SCANNER PANEL DATA

Consider the following fictional scanner panel data set that records purchases and prices of three brands of peanut butter: Nutty, Crunchy, and Gourmet. There are 200 households represented, and five purchases are recorded for each household. The following SAS statements create the data set Pbutter:

```
data pbutter;
  input id task Brand Choice LogPrice ;
datalines;
1 1 Crunchy 0 0.51282
1 1 Gourmet 0 0.61519

   ... more lines ...

200 5 Crunchy 0 0.41211
200 5 Gourmet 0 0.58222
200 5 Nutty 1 0.25464
;
```

The following SAS statements fit a mixed multinomial probit model to the data:

```
proc bchoice data=pbutter seed=9103 nmc=100000 nthreads=8;
  class Brand(ref='Nutty') ID Task;
  model Choice = Brand  LogPrice / choiceset=(ID Task) type=probit;
  random Brand LogPrice / subject=ID type=un;
run;
```

The NMC= option in the PROC BCHOICE statement specifies 100,000 iterations in the MCMC simulation. The sampler for the MMNP model produces samples that exhibit a high degree of autocorrelation, so large nominal sample sizes are needed to produce effective sample sizes that are large enough for reliable inference. The MODEL statement specifies **Choice** as the response variable and includes **Brand** and **LogPrice** as fixed effects. The CHOICESET= option specifies that the combination of **ID** and **Task** identifies the choice sets. The TYPE=PROBIT option requests a probit model. The RANDOM statement requests that **Brand** and **LogPrice** be included as random effects. The SUBJECT= option specifies that **ID** identifies the subjects. The TYPE=UN option specifies an unstructured covariance structure for the random effects.

Figure 6 displays the posterior summary. The first three items report the estimates for the fixed effects. The next three items report the elements of the normalized covariance matrix of the error term differences. There is evidence of both correlation and heteroscedasticity in the error terms. The last six items report the elements of the covariance matrix for the random effects. Likewise, there is evidence of both correlation and heteroscedasticity among the random effects.

Figure 6 Posterior Summary Statistics

The BCHOICE Procedure

Posterior Summaries and Intervals

			Standard	95%	
Parameter	**N**	**Mean**	**Deviation**	**HPD Interval**	
Brand Crunchy	100000	-0.3979	0.2075	-0.8012	0.0147
Brand Gourmet	100000	0.0686	0.2737	-0.4871	0.5627
LogPrice	100000	-3.5740	0.6759	-4.9071	-2.2826
Sigma 1 1	100000	1.0000	0	1.0000	1.0000
Sigma 2 1	100000	-0.0736	0.2616	-0.5736	0.4454
Sigma 2 2	100000	0.7068	0.4420	0.1235	1.5762
RECov Brand Crunchy, Brand Crunchy	100000	4.0201	0.9876	2.2811	5.9889
RECov Brand Gourmet, Brand Crunchy	100000	2.0213	0.9206	0.4726	3.8914
RECov Brand Gourmet, Brand Gourmet	100000	3.1879	1.7519	0.7360	6.6664
RECov LogPrice, Brand Crunchy	100000	0.8388	1.4010	-1.8539	3.7451
RECov LogPrice, Brand Gourmet	100000	0.0823	1.1819	-2.4750	2.2998
RECov LogPrice, LogPrice	100000	1.8061	1.4961	0.1014	4.6748

SUMMARY

PROC BCHOICE is a new, experimental procedure that enables you to perform Bayesian analysis for discrete choice models. The examples in this paper demonstrate how to use PROC BCHOICE to fit multinomial logit, nested logit, mixed multinomial logit, multinomial probit, and mixed multinomial probit models.

REFERENCES

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, 185–205.

Allenby, G. M. and Rossi, P. E. (1999), “Marketing Models of Consumer Heterogeneity,” *Journal of Econometrics*, 89, 57–78.

Gamerman, D. (1997), “Sampling from the Posterior Distribution in Generalized Linear Models,” *Statistics and Computing*, 7, 57–68.

Kuhfeld, W. F. (2010), *Marketing Research Methods in SAS*, Technical report, SAS Institute Inc., `http://support.sas.com/resources/papers/tnote/tnote_marketresearch.html`.

Lahiri, K. and Gao, J. (2002), “Bayesian Analysis of Nested Logit Model by Markov Chain Monte Carlo,” *Journal of Econometrics*, 11, 103–133.

Luce, R. D. (1959), *Individual Choice Behavior: A Theoretical Analysis*, New York: John Wiley & Sons.

McCulloch, R. and Rossi, P. E. (1994), “An Exact Likelihood Analysis of the Multinomial Probit Model,” *Journal of Econometrics*, 64, 207–240.

McFadden, D. (1981), “Econometric Models of Probabilistic Choice,” in C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, MA: MIT Press.

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

Rossi, P. E., McCulloch, R., and Allenby, G. M. (1996), “The Value of Purchase History Data in Target Marketing,” *Marketing Science*, 15, 321–340.

Train, K. E. (2009), *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press.

Varian, H. R. (1978), *Microeconomic Analysis*, New York: W. W. Norton.

ACKNOWLEDGMENTS

The authors are grateful to Fang Chen, Funda Gunes, and Anne Baxter of SAS Institute Inc. for their valuable assistance in the preparation of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Allen McDowell
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-6837
Allen.Mcdowell@sas.com

Amy Shi
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-2936
Amy.Shi@sas.com