Updated 30 August, 2006

The Mixed Logit Estimator: As Implemented in the DISCRETE program.

The Mixed Logit estimator is a varying parameters MNL model introduced by Train. In addition to the natural appeal of varying parameters (that individuals don't all have exact same preference structure), the Mixed Logit's use of varying parameters effectively addresses the IIA problems of the standard logit. And does so without the need of specifying a possibly arbitrary nesting structure (as required by the Nested Logit model).

Contents:

---

1. Synopsis

The MIXED model of the DISCRETE program contains a feature rich implementation of the Mixed Logit estimator. In addition to a standard Mixed Logit, you can specify a number of variants. Among other variants (they grow over time) are:

1. repeated choice occasions for each individual
2. uncertain Y-variables (imprecise measures of dependent variables)
3. uncertain X-variables (imprecise measures of independent variables)
4. grouping of alternatives (a simple nesting)

Note that these variants can be combined; for example, a "repeated choice occasion, uncertain X-variables model with grouped alternatives" is a supported variant!

DISCRETE is based on the GRBL2 package. GRBL2, which uses GAUSS 5.0, is a set of procedures that implement both maximum likelihood estimation, and that provide a user interface. In particular, GRBL2 uses "command files" to specify input

data sets, independent and dependent variables, model options, etc etc etc. These command files can be fairly complex; for example, they support IF/GOTO logic and user interaction at run-time. In this documentation, we will not discuss GRBL2 in detail – see GRBL2_BATCH.TXT for the details on using GRBL2.

Note that this documentation is meant as a supplement to the complete descriptions found in DISCRETE.TXT. Hence, in the remainder of this document we refer the reader to DISCRETE.TXT for further details, and for descriptions of various advanced options.

---

2. The Basic Model

- We suppose n=1..N *individuals*.
  - Each individual has her own, unique, vector of coefficient values: β_n
- Each individual faces one, or several, different *choice occasions*.
- Within a choice occasion, the individual can choose from among k=1..K different *alternatives.*
  - Each alternative is characterized by a vector of characteristics: X_{nk}
  - These characteristics can vary across choice occasions.
- The individual can choose just one of these alternatives one time: $y_k=1$; $y_j=0$ j□k,
  - or she can make *many-choices*, choosing more then one alternative more than one time (i.e.; choosing alternative 1 three times, alternate 6 twelve times, and alternative 9 once): $y_k=V$, v=0…

Following Train, the basic mixed logit model is:

A0) $$L_{nc}(\beta_n) = \exp(X'_{nc}\beta_n) \Big/ \sum_j \exp(X'_{nj}\beta_n)$$

where:

n = the nth individual
c = the chosen alternative (of 1..J alternatives)
L_{nk} = the conditional probability that individual n chooses alternative k
β_n = the individual specific value of β

The unconditional probability is the integral of the conditional probability over all possible values of β_n. DISCRETE uses simulation techniques to solve for the unconditional probability. Basically, an expanded vector of coefficients,θ, is estimated; where θ describes the distribution of β_n. For example, θ may consist of

two sets of parameters: the coefficient-means and the coefficient-standard-deviations. In this case, non-varying parameters would have coefficient-standard-deviations set equal to 0.

The mixed logit is solved using a simulation method that works by drawing, for each observation, r=1..R_n different *replications* of β_n ($\beta_n^{r|\theta}$). These draws are based upon a current estimate of θ. For each of these $\beta_n^{r|\theta}$ values, L_{nk} is computed. The average value of L_{nk}, across these R "replications", is used as the likelihood for this individual (given θ). Using these likelihoods, maximum likelihood techniques (MLE) are used to find the best estimate of θ.

Thus, the simulated, unconditional probability for observation *n* (given a candidate value of β) is:

A1) $$L^*_{nc} = \sum_{r=1}^{R} L_{nc}(\beta_n^{r|\theta}) \Big/ R$$

As noted above,

- θ is the same for all observations.
- β_n varies across the n=1..N observations in the sample;
- the distribution of β_n is described by the elements of θ

Operationally, for each replication of each individual, a $\beta_n^{r|\theta}$ vector is generated.

- For example, the k'th element of $\beta_n^{r|\theta}$ can be modeled as:

  $\beta_{nl} = b_l + \eta_{nrl}$,

  where b_l is a constant (non-varying) component, and η_{nrl} is a stochastic (*n'th* observation and *r'th* replication specific) component, of the *l'th* coefficient

Typically, β_n is typically assumed to follow a multi-variate normal distribution, with θ containing the mean (b), and the covariance matrix of η (Σ_η). Therefore, given an estimate of θ , for observation n and replication r,

$$\beta_n^{r|\theta} = B_\mu + \Omega_\eta \upsilon_{nr}$$

where…

B_μ is a vector of *non-varying components*

$\Omega_\eta \upsilon_{nr}$ is used to generate values of η_{nr} – the *stochastic components*

υ is a generated vector of random deviates (typically with a standard normal distribution).

Ω_η is the "square root" of Σ_η

In the simple case of no-correlation of varying parameters, Ω_η is a diagonal matrix, with the standard deviations of each η on the diagonal.

Notes:

- In practice, only a subset of the β_n are allowed to vary – some coefficients are assumed to be the same for all observations.
- A different set of R_n draws are used for each of the n observations. Hence, the values of υ_{nr} (hence $\beta_n^{r|\theta}$) will vary across both r and n.
- The size of R_n is usually the same value for all observations (say, 500 draws for each observation)

This model can be extended by recognizing that each individual may have more than one *choice occasion.* For example, there may be several years of data for an individual, with one choice made in each year. Over these several *choice occasions*, the choices may change and the explanatory variables (the Xs) may also change, but the β_n are assumed to **not** change.

b) $L_{ntc}(\beta_n) = \exp(X'_{ntc}\beta_n) \Big/ \sum_j \exp(X'_{ntj}\beta_n)$

where:

t = the t'th (of T) choice occasions

L_{ntc} = the conditional probability that individual i, during the t'th choice occasion, chooses alternative c

When solving the likelihood for an observation, the product of the probabilities of each choice occasion is used. That is,

$L_{nT}(\beta_n) = \prod_{t=1}^{T} L_{ntc_t}(\beta_n)$

Where c_t is the alternative chosen on the t'th choice occasion.

Hence, the unconditional likelihood for observation n is:

B1) $L^*_{nT} = \sum_{r=1}^{R} L_{nT}(\beta_n^{r|\theta}) \Big/ R$

Note that (A1) is simply (B1) with T=1.

---

3. Estimating the basic model with DISCRETE

DISCRETE expects that each observation consists of multiple rows in your GAUSS dataset. Each row corresponds to an alternative. If there are more than one choice occasion, than each row corresponds to an alternative belonging to a choice occasion.

DISCRETE can convert "one row per observation" GAUSS datasets to the "one row per alternative" format. See the description of the CONVERT model in DISCRETE.TXT.

Each row of the dataset should contain the following variables:

1. X (independent) variables
2. A response variable, that indicates whether the alternative was chosen
3. An observation identifier
4. If you have multiple choice occasions, a choice occasion identifier.

Appendix 1 has a short example of a dataset.

If you have "balanced" observations, you don't need to include the observation (and choice occasion) identifiers. See DISCRETE.TXT for the details.

The "command file", that you will feed to DISCRETE, needs to specify variables and options. Lets use an example:

```
@ this is a sample command file for DISCRETE @
reset2 ;
file d:\stuff\myproject ;
output d:\stuff\myproject.out overwrite  ;
title My mixed logit model ;

model mixed ;
  REPS 50 ;
  ID id ;
  RESPONSE visit1 ;
  X X_001 x_002 x_003 x_004 x_005 ;
  NORMAL x_001 x_002  x_005;

RUN;
```

Notes:

- semi-colons (;) end each command.
- Text between @ …. @ are comments (and are ignored)
- Commands are case-insensitive. Thus, `FILE and file` do the same thing.

Lets go over each of the commands

`MODEL` Estimate the mixed logit model. DISCRETE also can be used for other models, such as bivariate probit or nested-MNL.

You can specify one option after `MODEL`: `FULL=1.`
The `FULL=1` option effects what model is estimated:

- If a FULL=1 option is **not** specified, a diagonal variance/covariancematrix (of the varying parameters) is estimated -- each "varying" oefficient is effected only by a single random component.
- If you specify FULL=1, a full covariance matrix be estimated.

Example: MODEL MIXED FULL=1 ;

FILE	The GAUSS data file to use. If no path is specified, the "GAUSS working directory" is assumed
OUTPUT	The output file. If overwrite is specified, then a pre-existing file (with this name) will be erased before new output is written.
TITLE	An optional title to be displayed above the results (written to the output file)
MODEL	Specify which MODEL to use. Note that DISCRETE supports a variety of models, including two stage nested logit.
REPS	How many draws to use (the R in A1 and B1).
ID	A variable name, in the input file, that identifies the observation. The rows for each observation must be grouped together. Thus, if you have 100 rows with an ID of 1, then 100 with an ID of 2, then 100 with an ID of 1 – DISCRETE assumes you have three separate observations.
RESPONSE	The response variable. Typically, this is a 0/1 dummy – all rows will have a 0 value, except the row that describes the chosen alternative.
X	The explanatory variables. These values of these variable describe the alternatives – the values are expected to vary across alternatives (for an observation).
NORMAL, And TRIANGLE UNIFORM LOGNORMAL	The X variables that are allowed to vary. Variables not listed after NORMAL are assumed to **not** vary across observations.

Note that the NORMAL "varying coeffients" are assumed to have coefficients that are normally distributed. DISCRETE will estimate the parameters of this distribution; in particular, the mean values and standard deviation; or the covariance matrix (of the several varying coefficients). The "mean values" are equivalent to the coefficient estimates from a standard MNL, the "standard deviations" measure how much these coefficients vary across the population.

In addition to NORMAL, you can also use

- UNIFORM to specify –SD to +SD uniformally distributed coefficients (where SD is the estimated standard deviation)
- TRIANGLE to specify –SD to +SD triangularly distributed coefficients (where SD is the estimated standard deviation)
- LOGNORMAL to specify log normally distributed coefficients.
  Thus if B_U is the "estimated mean" and B_SD is the estimated standard deviation of a lognormally distributed coefficient, the actual coefficient will have:
  median = exp(b_u)
  mean = exp(b_u + ((b_sd^2)/2))
  sd = [exp(b_u + ((b_sd^2)/2))] * sqrt[exp(b_sd^2)-1]

`RUN` Estimate the model just specified.

If you have more than one choice occasion per observation, you would add an

```
OCC varname ;
```

For example:

```
@ this is a sample command file for DISCRETE @
file d:\stuff\myproject ;
output d:\stuff\myproject.out overwrite  ;
title My mixed logit model ;

model mixed ;
  DRAWS 50 ;
  ID ID1 ;
  RESPONSE visit1 ;
  OCC year ;
  X X1 x2 x3 x4 ;
  NORMAL x1 x3 ;
 RUN;
```

Thus, each observation may have several *years* of observations. For each observation DISCRETE will compute a separate probability using the rows (the alternatives) belonging to each *year*.

Hint:	*to estimate a standard MNL, use* `DRAWS 0` All the other MIXED options can be used (such as OCC). Note that DISCRETE also supports a **MNL** model, which can also be used to estimate standard MNL models. **MNL** is faster, but it does not support all the bells and whistles (such as uncertain dependent variable correction) offered by **MIXED**.

3a. Many alternatives chosen & multiple choice occasions

DISCRETE can also estimate *many choice* models; models where each alternative may be chosen several different times.

For example, you may have data on the number of times individuals visited each of several parks during a single year. Say: an individual may have visited park A ten times, park B three times, park C zero times.

Furthermore, the explanatory variables for these parks doesn't change across the year.

You could estimate this using a choice occasion model, but it is much simpler to use a non 0/1 value for the response variable. To do this, the value of the Y (response0 variable should simply be the count of times the alternative was chosen. You do not need to specify any options – DISCRETE will detect that a *many choice* model is being estimated.

In fact, you can combine *many choice* with "choice occasions" – during each choice occasion, many alternatives can be chosen. Just remember that explanatory variables may change across choice occasions, but not across the many choices made within a choice occasion.

Formally, for the many-choices model (b) becomes:

$$\text{b2)}\quad L_{ntK|y}(\beta_n) = \prod_{k=1..K}\left(\exp(X'_{ntk}\beta_n) \Big/ \sum_j \exp(X'_{ntj}\beta_n)\right)^{y_k}$$

Note that if y_k=1 for only one value of k (and zero for all other values), (b2) is equivalent to (b).

4. The Uncertainty Models

Following Hellerstein, the mixed logit model can be extended to account for uncertain measures of either the dependent variable or the independent variable. Dependent variable uncertainty means you aren't sure of exactly which alternative was chosen. Independent variable uncertainty means you do not have exact measures of the explanatory variables.

4a. Dependent variable uncertainty

Instead of knowing exactly which site was chosen, you may have a posterior-probability of choice. This posterior probability measures your guess as to what alternative was chosen, given that an alternative was chosen. A measure that is not based on the explanatory variables you include in the model, one that uses ancillary information. In a sense, the uncertain dependent variable controls for mismeasuring which choice was made.

There are two uncertain dependent variables probability models:

The "quasi-bayesian" (additive) model is:

$$\text{Ci)} \qquad L_{ni}(\beta_n) = \sum_{k=1..K} \exp(X'_{nk}\beta_n)\pi_{nk} \Big/ \sum_j \exp(X'_{nj}\beta_n)$$

The "implicit replications" (multiplicative) model is:

$$\text{Cii)} \quad L_{ni}(\beta_n) = \prod_{k=1..K} \exp(X'_{nk}\beta_n)^{\pi_{nk}} \Big/ \sum_j \exp(X'_{nj}\beta_n)$$

where π_{nk} is the "reporting accuracy" – it is the posterior probability that observation n (on this choice occasion) actually chose alternative k. If $\pi_{nk} = 0$ for k not k', and $\pi_{nk} = 1$ for k=k', this collapses to the standard MNL.

Note that π_{nk} is often computed using some external data, say D_n that contains information used to infer the posterior probability of each of k=1..K alternatives being chosen.

The WEIGHT variable should contain measures of the posterior probability (the probability based on what the observation reports, rather then on attributes of the alternatives) that a row was chosen. A value of 0 means "0 probability". Larger values imply larger probabilities. This posterior probability is formed by using the

normalized value of the weights; the normalization being the sum (across all rows belonging to a choice occasion) of the WEIGHT variable.

Example:

```
@ this is a sample command file for DISCRETE @
file d:\stuff\myproject ;
output d:\stuff\myproject.out overwrite  ;
title My mixed logit model ;

model mixed ;
  reps 50 ;
  ID ID1 ;
  RESPONSE visit1 ;
  OCC year ;
  UNCERTAIN ytype=additive type=no ;
  WEIGHT ACCURACY ;
  X X1 x2 x3 x4 ;
  NORMAL x1 x3 ;

RUN;
```

Using weights and a many-choices model

Use of weights changes how DISCRETE constructs likelihoods.

In the non-weighted model, each row can uniquely contribute to the value of the likelihood. This means that the value of the response variable, for each row, can be used as a separate multiplier – leading to the many choices variant. However, with weighting, this separation does not make sense – the weighting means that the probability of a single choice is a weighted average of each alternative's probability.

Hence, when non 0/1 response variable is used, DISCRETE will sum Y, and use this sum as a multiplier for the weighted probability.

Formally, for an individual n, facing a single choice occasion comprising K different alternatives, who can choose each of the alternatives several times (many choices):

Weighted: $$\left(\sum_{k=1..K} \pi_{nk} \exp(X'_{nk}\beta_n) \Big/ \sum_{k=1...K} \exp(X'_{nj}\beta_n) \right)^{\sum_{j=1...K} y_{nj}}$$

Where:

y_j = times alternative j is chosen

π_k = "posterior probablity" of choosing alternative k

β_n = Coefficient vector for individual n

This can be contrasted to (b2) above, the many-choices non-weighted model probability.

4.a.i) Details: the *quasi-bayesian* (addtive) uncertain dependent variables model

The uncertain dependent variables model can be thought of as predicting the probability of some observed "choice indicator" (D), given knowledge of prior and posterior probability. Thus, for each choice k from a set of K choice ….

- the prior probability is the standard MNL probability – it measures the probability, given observable attributes (X_k), of selecting this choice.
- The posterior probability assigns the probability of observing the choice-indicator (D) given that a particular choice was made. It is based on ancillary, non-estimable information particular to each choice (R_k).

Thus, call D_i the "choice indicator for an observation" (assume just one choice occasion, and a given value of β). Then, the probability of observing D_i is:

$$P(D_i) = \sum_{k=1..K} \left[\pi(D_i \mid R_k) * \left(\exp(X'_k \beta) \Big/ \sum_{j=1..K} \exp(X'_j \beta) \right) \right]$$

where the first term, of observing D given that site k was chosen (and given R_k info on site k).

π(D) has a bayesian feel: it considers all possible choices (k=1..K), the prior probability of choosing each choice (the MNL term), and the probability of observing the indicator variable (D), given a choice (R_k).

However, it is not a true probability: the product of $P(D_i)$ across all D (across all possible values of the "choice indicator") does not necessarily sum to 1.0. A more complete model would normalize $P(D_i)$ by

$$P(D)^* = \sum_{i \in I} \left[\pi(D_i \mid R_k) \right]$$

where $i \in I$ indicates the set of all possible values of the "choice indicator" variable.

While simulations methods might be able to proxy for this (possibly infinite) set, the information requirements are still this is onerous – a measure of R_k, for use in to $\pi(D \mid R_k)$ generated for all the "simulated" i would be needed.

Hence, MIXED does not attempt this, more formally correct, normalization.

4.a.ii) Details: the *implicit replications* uncertain dependent variables model

Instead of the bayesian framework, consider a replication framework. The notion is that the probabilities predict the distribution of results over C replications (say, C identical individuals)

This uses $\pi_k(D,R_k)$ -- the fraction of times, over the C replications, that k was chosen. This fraction is conditioned on information D and R_k.

The definition of π_k implies that over C choices "replications", site k is chosen $c_k=\pi_k(|D,R_k)*C$ times. Thus, over C replications the joint probability is

$$P_{N*} = \prod_k \left(\exp(X'_k\beta) \Big/ \sum_{j=1..K} \exp(X'_j\beta) \right)^{c_k} = \prod_k \left(\exp(X'_k\beta)^{c_k} \Big/ \sum_{j=1..K} \exp(X'_j\beta) \right)$$

where $(\Sigma_{k=1..K}\ c_k) = C$

Since in reality only one choice (and not C choices) is made, some form of mean is required. A geometric mean is appropriate:

$$P_N = \left(\prod_k \left(\exp(X'_k\beta)^{c_k} \Big/ \sum_{j=1..K} \exp(X'_j\beta) \right) \right)^{1/C} =$$

$$\prod_k \exp(X'_k\beta)^{c_k/C} \Big/ \left\{ \prod_k \left(\sum_j \exp(X'_k\beta) \right)^{c_k} \right\}^{1/C} =$$

$$\prod_k \exp(X'_k\beta)^{\pi_k} \Big/ \sum_j \exp(X'_k\beta)$$

Note that the quasi-bayesian model uses a linear mean (P *(D))*, while the implicit-replications model uses a multiplicative mean (*P_N.)*.

Note that for both the "quasi-bayesian" and "implicit replications" models, when B_k = 1.0 for one value of k and are otherwise zero, both models are identical to the standard MNL.

4b . Independent variable uncertainty

In addition to dependent variables uncertainty, DISCRETE can correct for independent-variables uncertainty.

Independent variable uncertainty is an errors in variable correction, where you provide some form of data describing the scope of this uncertainty.

DISCRETE support three methods for controlling for independent variable uncertainty.

1. *Static* X-Replication (replicated observations included in dataset)
2. *Dynamic* x-replication (replicated observations generated on the fly)
3. *Xβ-Replications*, uses variance information to generate $X\exists$ values on the fly

Static replication requires that you replicate, within your dataset, the rows describing each alternative. Thus, instead of having one row per alternative per observation (or per choice occasion within an observation), you would have many "x-replicated" rows. Each of these "x-replicated" rows would have alternate measures of the explanatory variables – that is, each of these rows would contain a plausible value of the explanatory variables for an alternative. Presumably, these plausible values reflect a probability distribution (perhaps a non-parametric probability distribution) describing the distribution of the true values of the explanatory variables.

For each observation, DISCRETE will form a likelihood by taking the average value across all the x-replicated rows. Thus, DISCRETE computes an average across two dimension of data – the x-replicated rows and the draws used to simulate the mixed logit probability density function.

> *Example:* if you specify 10 alternatives to chose from, and 100 x-replications, there would be 1000 rows for this observation.

Dynamic replication is similar to static replication – they both use alternative measures of X to compute an average likelihood (given a beta). However, with dynamic replication, these x-replicated observations are generated on the fly. This generation is based on covariance-of-X information that you provide in a second dataset. Note that this covariance may be the same across many (or even all) of the alternatives included in the dataset.

Xβ-Replication is a simplification of dynamic replication – it uses covariance-of-X information that you provide in a second dataset. This covariance information is used to estimate a distribution of $X\beta$ for each alternative. DISCRETE will generate *Xβ-replications* -- values of $X\beta$ drawn from this distribution -- and use these values to compute an average likelihood.

The UNCERTAIN command is used to specify both of these methods:

1. Static X-replicated observations : `uncertain type=X var=xrep ;`

2. Dynamicc X-replicated observations : `uncertain type=XREP VCI=vci_file REP=nreps CLASS=class_id ;`
3. Xβ-replicated observations: `uncertain type=XB in=vcfile rep=nn id=vcid ;`

Note that the TYPE option is used to select which method to use:

TYPE=X for X-Replicated observations, and
TYPE=XB for Xβ-Replications.

Static X-replicated observations

Example:

```
@ this is a sample command file for DISCRETE @
file d:\stuff\myproject ;
output d:\stuff\myproject.out overwrite  ;
title My mixed logit model ;

model mixed ;
  DRAWS 50 ;
  ID ID1 ;
  RESPONSE visit1 ;
  OCC year ;
  UNCERTAIN ytype=no type=X var=MYXREP ;
  X X1 x2 x3 x4 ;
  NORMAL x1 x3 ;

RUN;
```

The VAR=MYXREP is used to identify a "x-replication identifier" variable in the dataset. This should be a numeric variable with values that are unique to an observation (you can use the same values across different observations).

Thus, for each observation DISCRETE will:

1. Extract all rows with the same observation ID
2. Extract, from the rows chosen in step 1, all rows with the same X-replication ID
3. If there are more than one choice occasion, extract (from the rows chosen in step 2) all rows with the same choice-occasion ID.
4. These rows form the set of alternatives from which a choice is made.

For each observation (or for each choice occasion within an observation) DISCRETE will find the value of A1 (or B1) across all the X-replications. That is, for each unique value of the x-replication variable, A1 (or B1) is computed using rows that satisfy step 3. The average of these values (across all unique values of the x-replication variable) is used as the probability for this observation.

Note that the Mixed logit draws (of different values of beta) are also done. Thus, if there are 50 draws specified, and 100 x-replications per observation, the probability for an observation is the average across 5000 (50*100) different values of A1 (or B1).

Dynamic X-replicated observations

Example:

```
@ this is a sample command file for DISCRETE @
file d:\stuff\myproject ;
output d:\stuff\myproject.out overwrite  ;
title My mixed logit model ;

model mixed ;
  DRAWS 50 ;
  ID ID1 ;
  RESPONSE visit1 ;
  OCC year ;
  UNCERTAIN ytype=no type=XREP VCI=d:\stuff\myxvc rep=122
            class=aclass ;
  X X1 x2 x3 x4 ;
  NORMAL x1 x3 ;

RUN;
```

Dynamic X-replication requires three options:

1. `rep=nn :` nn is the number of replications, of each alternative
2. `vci=afile: afile` is the name of a GAUSS dataset that contains *variance-covariance-information* of the explanatory variables (VCI file)
3. `class=aclass: aclass` is a variable that identifies the "class" to which this row belongs. Each class is associated with a variance covariance matrix. The `aclass` variable must exist in your gauss dataset (and have a value defined for each row of the dataset), and must exist in the variances file specified by the `vci=afile` statememt.

There are several other options that can also be set (FRAC, NROOT, and ROUNDS) – if you do not set them, default values are used. See DISCRETE.TXT for the details.
.

Note that the Mixed logit draws (of different values of beta) are also done. Thus, if there are 50 draws specified, and 100 x-replications per observation, the probability for an observation is the average across 5000 (50*100) different values of A1 (or B1).

Xβ-replicated observations

Example:

```
@ this is a sample command file for DISCRETE @
file d:\stuff\myproject ;
output d:\stuff\myproject.out overwrite  ;
title My mixed logit model ;

model mixed ;
  DRAWS 50 ;
  ID ID1 ;
  RESPONSE visit1 ;
  OCC year ;
  UNCERTAIN  ytype=no type=XB VCI=d:\stuff\myxvc rep=122
            class=aclass ;
   X X1 x2 x3 x4 ;
  NORMAL x1 x3 ;

RUN;
```

Xβ-replication is similar to dynamic X-replication, in that a number of "replications" of each row are generated on the fly, and the probability of an observation is the average value computed using these replications. However, rather then generating random values of X, random values of Xβ are generated.

Xβ-replication requires three options:

4. `rep=nn :` nn is the number of replications, of each alternative
5. `vci=afile: afile` is the name of a GAUSS dataset that contains *variance-covariance-information* of the explanatory variables (VCI file)
6. `class=aclass: aclass` is a variable that identifies the "class" to which this row belongs. Each class is associated with a variance covariance matrix. The `aclass` variable must exist in your gauss dataset (and have a value defined for each row of the dataset), and must exist in the variances file specified by the `vci=afile` statememt.

The variance-covariance-information file

The *variance-covariance-information* file is a gauss dataset that contains variance/covariance information on the x variables of the alternatives.

This file should contain variable names that are identical to the variable names in your main gauss dataset. In addition, two other variables must be present:

CLASS_ID	The class ID. This identifies the "class", and will be matched against a variable with the same name in the main gauss dataset.

	The CLASS_ID is specified using the `CLASS=vvv` option in the `UNCERTAIN command.` *Important note:* different alternatives in the main gauss dataset can share the same CLASS_ID! In fact, every single alternative (every single row) in the main gauss dataset can use the same value for CLASS_ID.
NAME *must have a name of* `_NAME_`	The variance/covariance matrix for an alternative (in the main gauss dataset) is specified by: 1. Get the class-id for this alternative 2. Find all rows in the *VCI* file that have the same value for their class-id 3. For the i,j cell of the variance covariance matrix (the covariance of **XNAME_i** and **Xname_j** variables) for this "class" of rows: • find the row with a `_NAME_` value that matches a **Xname_i,** • read the value from the column with a name that matches **Xname_j** • If either are not found, use a value of 0.

Example of a portion of a VCI file

ID	**CLASSI**	**_NAME_**	**X1**	**X12**	**X5**	**X4**
1	1	X1	20	1	5	-3
1	1	X12	1	51	-2	6
1	1	X5	5	-2	22	12
1	1	X4	-3	6	12	8
1	2	X1	23	2	0	-13
1	2	X12	2	31	0	12
1	2	X4	-13	12	0	2

Notes:

- Symmetry should be maintained: the value of row XA and column XB should match that of row XB column XA.
- DISCRETE will enforce symmetry, but not necessarily in a way you intend.
- The X5 row is missing for CLASSI=2 -- values of 0 are used (as reflected in the X5 column for CLASSI=2)
- If a model uses the X variables: `X1 Z1 X12 X4,`
  then the variance/covariance matrix used for CLASSI=1 will be

```
20   0   1  -3
 0   0   0   0
 1   0  51   6
-3   0   6   8
```

Note that since Z1 is not specified in either columns are rows, 0s are used.

- Thus, if a set of *alternatives* have different "observed" values of X, but the same variance/covariance (say, the same noisiness features), one should use the same CLASSI value for each row -- even if these rows are associated with different observations.
- You can use MAKEDATA to create a VCI file from a raw dataset (say, a panel a dataset containing individual observations)

4b.1. Details: the Xβ-replicated observations model (for uncertain independent variables)

The likelihood for the Xβ-replicated observations model is (suppressing observation subscripts):

$$\text{Ciii) } P_{xb} = \sum_{s=1..R_{xb}} \left(\exp\left(X_k \beta + \delta_{sk} \sqrt{\beta' \Omega_x \beta}\right) \Big/ \sum_{j=1..K} \exp(X_j' \beta + \delta_{sk} \sqrt{\beta' \Omega_x \beta}) \right) \Big/ R_{xb}$$

Where
δ_{sk} is a draw from a standard normal
Ω_x is the variance matrix of the X variables.
R_{xb} is a number of draws.

For the mixed logit, β_n (the n'th observation specific value of β). Since this is unknown, R draws are used to simulate β_n. Hence, the mixed logit version of Ciii is estimated using simulations, with a likelihood:

$$L_{xb}^{*} = \sum_{r=1..R} P_{xb}(\beta_r) \Big/ R$$

Where β_r is the r'th simulated value of β, and R is the number of "mixed logit" replications. Thus, estimation of a likelihood (for a single observation and a single guess at the coefficient vector) requires R_{xb} * R different computations of the interior of Ciii.

To estimate uncertain dependent and independent variables, replace the numerator in Ciii with the corresponding numerator (from Ci or Cii). Thus, for the quasi-bayesian (linear) model, the numerator becomes:

$$\sum_{k=1..K} \exp\left(X_k \beta + \delta_{sk} \sqrt{\beta' \Omega_x \beta}\right) \pi_k$$

4b.2. Details: the dynamic X-replicated observations model (for uncertain independent variables)

The dynamic X-replication model, for correcting for uncertain X, uses a multi-round approach.

At the beginning of the d'th round, a mixed-logit model is used to estimate β_d.

> A no-X-uncertainy model is estimated, with almost all of the other DISCRETE model options available. For example, you can specify choice occasions and Y uncertainty. Two-stage models can also be estimated, though the first stage variables are assumed to be measured with no uncertainty.

After estimating β_d, DISCRETE uses β_d compute the likelihood for each observation. Then, DISCRETE will identify which observations seem *funny*, hence may be suffering from an uncertain measure of X.

For each of these *funny* observations, DISCRETE will dynamically generate a number of X-replications. These are cenetered around the "measured" X values, with a deviation that is a function of the variance information for this observation/alternative (as recorded in the VCI file).

> Within an observation, DISCRETE will compute a likelihood for each of these dynamically generated X-replication. The X-replication that generates the "best" likelihood is then used as a "better measure" of the value of X (for this observation).

Once these "better measures of X" have been determined for each of the *funny* observations, a new round commences, that uses these "better measures" as the independent variables. Note that for the non- *funny* observations, the X values are used without modification.

Also note that an observation may be one of the *funny* observations for none, one, several, or all of the rounds. If an observation is *funny* for more one (not-necessarily sequential) rounds, DISCRETE will generate x-replications around the "better measure" (determined in a prior round). Thus, over the course of several rounds DISCRETE will modify its prior guess of the "better measure of X".

Generation of x-replications uses several factors:
1) The variance of the X data in a row --- this uses the CLASSID variable
2) The round. In later rounds, the variance is shrunk. Thus, in later rounds the neighborhood of searching (for "better values of X") shrinks.

In practice, choosing what is the "better measure of X", from a set of X-replications, is based on two factors: the value of the likelihood (at this X-replcation, given a beta), and the distance between the X-replication and the original X. Note that this is done

on a 1-observation-at-a-time basis -- each observation is treated as an independent entity when determining its "better X" value.

The use of distance is meant to discourage over-fitting of X. You can use NROOT variable to adjust how important distance is. If NROOT=0, then distance is used fully -- X-replications that are far from the measured X are less likely to be used. Large NROOT values (say, 100), cause distance to be de-emphasized; so what matters is the likelihood.

> Setting `NROOT=large_value (say, NROOT=100)` will lead to better likelihoods (for the entire dataset), but greatly increases the chance of overfitting -- it is more likely that X values will be chosen that just happen to work with an arbitrary (hence incorrect) beta value.

5. The aggregated alternatives models

In many cases, the actual alternative chosen may not be known, but you do know which of several, separate, aggregated alternative "groups" was chosen. That is, you know which "group" was chosen, but not which of the alternatives in the group.

For example, one may not know what site a trip was taken to, but one may know the region that the visited site is in.

DISCRETE supports two kinds of aggregated-alternatives models: full information and limited information.

Full information:

- Information on the attributes of all the alternatives is available.
- You know which group each alternative belongs to.
- You don't know which alternative was chosen, you do know which group was chosen.

Limited information

- Information on the attributes of alternatives is **not** available.
- Aggregate information on the attributes of groups is avaialable. In particular, information on the mean and variance-covariance of attributes (across alternatives belonging to a group), and the size of the group.
- You don't know which alternative was chosen, you do know which group was chosen.
- Ths model is based on the Ben-Akiva/Lerman Logit Model With Aggregate Alternatives (Discrete Choice Analysis: Theory and Applications to Travel Demand, MIT Press 1985):

In a sense, the aggregation group is a simple nest, as might be used in a nested-MNL. However, in the nested-MNL you know the elements of a subset, you know if a

subset was chosen, and you know which alternative within a subset was chosen (the knowledge of the chosen subset being an obvious result of knowing with alternative was chosen).

5.a) The *full information* aggregated alternatives model

The *full information* likelihood:

E1) $L_{ntK|y}(\beta_n) =$

$$\prod_{k \notin G} \left(\exp(X'_{ntk}\beta_n) \Big/ \sum_j \exp(X'_{ntj}\beta_n) \right)^{y_k} \left(\sum_{g \in G} \exp(X'_{ntg}\beta_n) \Big/ \sum_j \exp(X'_{ntj}\beta_n) \right)^{\sum_{g \in G} y_g}$$

Where:

$Y_G = \Sigma_{g \in G}\, y_g$ = total number of times that aggregation-group G was chosen

$k \notin G$ = choice k is not in the aggregation group – y_k is known

$k \in G$ = choice k is in the aggregation group -- y_k is **not** known

Note that for the weighted model (uncertain dependent variables models), the first term (the product) is replaced with a weighted sum – the second term (the aggregation group term) is the same. That is, the weighted model ignores weight information for choices within the aggregation group

Actually, DISCRETE allows you to have multiple aggregation groups. The second terms is replicated for each aggregation group, using $g \in G_m$ (choice g is in aggregation group m), and the first terms products only using alternatives that are not in any of the G_m groups.

Example –

```
@ this is a sample command file for DISCRETE @
file d:\stuff\myproject ;
output d:\stuff\myproject.out overwrite  ;
title My mixed logit model, with full info aggregates ;

model mixed ;
  REPS 50 ;
  ID ID1 ;
  RESPONSE visit1 ;
  AGG TYPE=FULL VAR=GRP_VAR ;
  X X1 x2 x3 x4 ;
  NORMAL x1 x3 ;
 RUN;
```

The AGG command specifies the type of aggregated sites model (TYPE=FULL), and specifies a variable that identifies which group a row (alternative) belongs to.

Example:,

- there are 4 alternative,
- alternatives 3 and 4 are in a "group",

then

- you can distinguish between:
  - Alt 1 was chosen
  - Alt 2 was chosen
  - Alt 3 or 4 was chosen (but not which of 3 or 4)

To repeat, you must have X information on all the alternatives. What is unclear is which alternative (within a group) gets the "y" values.

The VAR option identifies which group an alternative is part of.

- If an alternative is **not** n a group, the value of this variable should be 0.
- Otherwise, all alternatives with the same value of the variable will be placed in the same group.

For the set of alternatives comprising a group, the sum of Y is used as the dependent variable.

- It does **not** matter how this sum is achieved.
- Thus, using the above example with a group containing alternatives 3 and 4, the following rows are analytically equivalent:

Y_for_Choice 3	**Y_for_choice_4**
2	1
3	0
1	2
0	3

Notes:

- There can be more than one group in a choice occasion
- The number of groups can vary across observations -- some observations can have 0 groups, some many.
- Or the number of groups can vary across choice occasions within a single observation

Example of a dataset with a GROUP variable

ID	**GRPVARI**	**X1**	**X5**	**X4**
1	0	20	5	-3
1	0	1	-2	6
1	1	5	22	12
1	1	-3	12	8
1	1	23	0	-13
1	2	2	0	12
1	2	5	2	52

2	0	162	5	1
2	0	123	2	2
2	0	612	2	2
2	1	6	5	1
2	1	-13	0	2

5.b) The limited *information* aggregated alternatives model

The *limited information* likelihood (abstracting from varying parameters logit):

E2) $L(\beta_n) = \exp(V_k + \eta' \ln(M_k) + \eta' \ln(B_k)) \Big/ \sum_j \exp(V_j + \eta' \ln(M_j) + \eta' \ln(B_j))$

Where:

$V_k = X\beta$

$M_k = \sum_m \exp(Z_m \exp(\beta_{zm}))$; where $\beta_{z1} = 0$

$B_k = \sum_{\kappa \in k} \exp(\eta(V_\kappa - V_k)) \Big/ M_k$

$\eta`$ is the ratio η^*/η, where η is the "aggregated alternatives" scale parameter, and η^* is the "elemental" scale parameter.

$\eta`$ should have a value between 0 and 1. It is a measure of the correlation of alternatives within a group: corr= $(1 - \eta`^2)$. Thus, if $\eta` = 1$, alternatives within each group are not correlated.

B_k requires information on all the κ alternatives within the k'th aggregate (the k'th group). Since, by definition, this information is not available, DISCRETE uses a Ben-Akiva/Lerman approximation. This approximation is based on a normality assumption: that within an aggregate the alternatives have X (independent) variables that are drawn from a normal distribution.

The approximation is:

E2a) $\mathbf{B} = \eta` \ln(B_k) \approx \sigma^2_k / 2$

where

$\sigma^2_k = \beta' \Omega_k \beta$

and

Ω_k is the variance-covariance matrix of X variables in the k'th group.

However, experimentation suggest that E2a may be a poor approximation if Ω_k is relatively large. Using artificial data, we found a linear approximation that uses σ^2_k to better measure $\eta` \ln(B_k)$.

This linear approximation is:

E2b) $\mathbf{B} = \exp(-1.6 + 1.06 * \ln(\sigma^2_k) - 0.028 * \ln(\sigma^2_k)^2)$

Thus, E2 becomes

E3) $$\exp(V_k + \eta' \ln(M_k) + B_k) \Big/ \sum_j \exp(V_j + \eta' \ln(M_j) + B_j))$$

Note that M_k is a function of Z variables that measures the "size" of the aggregate (i.e.; number of alternatives in the group). Note that $\exp(\beta_z)$ is used to guarantee that M_k is greater than 0, and that the first element of β_z is always set to 0.

Also note that $\mathbf{B_k}$ can be from E2a or E2b.

Example – using the "linear approximation" (E2b):

```
@ this is a sample command file for DISCRETE @
file d:\stuff\myproject ;
output d:\stuff\myproject.out overwrite  ;
title My mixed logit model, with limited info aggregates ;

model mixed ;
  REPS 50 ;
  ID ID1 ;
  RESPONSE visit1 ;
  AGG TYPE=LIM VCI=d:\stuff\myxvc class=aclass ;
  Z SIZE1 ;
  X X1 x2 x3 x4 ;
  NORMAL x1 x3 ;
 RUN;
```

To use the standard approximation (E2a), replace the AGG line with:

```
AGG TYPE=LIM VCI=d:\stuff\myxvc class=aclass APPROX=NO ;
```

Notes:

- You can change the values of the parameters used in E2b – see DISCRETE.TXT for the details.
- If you do not specify Z variable(s), the M (size) terms will **not** be included, and η` will not be estimated.
- If you specify just one Z variable, its β_z coefficient is not estimated (the first coefficient of β_z is always set to 0, and note the $\exp(\beta_z)$ is used).
- If you do not specify a VCI=filename, the **B** (heterogeneity) terms will **not** be included.

6. The 2-stage model

To account for participation/decision 2-stage models, MIXED offers a probit/MNL estimator. This estimator allows for the random factor in the first stage (that is estimated with a PROBIT model) to be correlated with the varying parameters. Thus, a large positive shock in the first stage (ceteris paribus leading to a higher probability of participating) can be associated with an increased (or decreased) importance of one or several of the varying parameters.

One use of this model is to account for non-response bias. In particular, the set of choices available to individuals may vary across individuals, with each choice described by explanatory variables. In such circumstances, there is no natural way to assume a correlation between the "decision to participate random factor" and the random factors of a particular alternative – since the "alternatives" may be completely different things for each observation.

The varying parameters MNL offers a mechanism that lets this participation decision effect alternative choice in a sensible fashion. In fact, if there are particular choices that might be correlated with participation decisions, one can define dummies (with varying parameters) for each of these special choices (or for subsets of the special choices).

The basic model is

$$(E)\ L_{ntk}(\beta) = \Phi(Z\beta_z)^{Y1=1}\left(\exp(X'_{nc}\beta_n)\Big/\sum_{j=1..J}\exp(X'_{nj}\beta_n)\right) + (1-\Phi(Z\beta_z))^{Y1=0}$$

Where
- n = individual n
- $\beta = \beta_n$ and β_z
  - β_n = 2nd stage parameters for individual n, some of which may by varying
  - β_z = Coefficients on first stage variables (non-varying)
- Z = First stage independent variables
- Y1 = First stage (0/1) dependent variable
- X_{nj} = 2nd stage independent variables, individual n alternative j
- c = the chosen alternative
- Φ = CDF of standard normal
- ϕ = PDF of standard normal

The first stage decision is:
Y1=1 if $Z\beta_z + \xi > 0$,
otherwise, Y1=0
where ξ is a standard normal random variable.

We assume that ξ_n and some of the η_{nk} (the varying components of β_n) are correlated.

Hence,

- For observations for which one observes a choice, the expected value of the correlated elements of η will not be strictly zero (since the expectation of ξ_n is non-zero for these observations) .

This means,

- Joint estimation of (E) requires integrating over a bivariate, truncated distribution.

Given the complexities of such an integration, MIXED currently uses a two-step strategy.

First, a PROBIT estimator is used to compute β_z. Note that an expanded set of observations is required, including observations for which choices are not observed (the "non-respondents"). Furthermore, "respondents" contribute just one observation to this PROBIT – regardless of how many "choice occasion" recorded for this observation, or the number of alternatives per choice.

Second, using the estimate of β_z, two observation specific statistics are computed.

These use the following functions:

1) $\lambda(a_n) = \phi(a_n) / \Phi(a_n)$ -- this is a "mills ratio".
2) $\delta(a_n) = \lambda(\lambda + a_n)$

Adapting from Greene (4th edition, chapter 20.4), λ and δ are used to estimate the formulae for the distribution of truncated normal variates (Y and Y1):

$$E[Y \mid Y1 > -\mathbf{a_n}\ ] = \rho\sigma\lambda$$

$$Var[Y \mid Y1 > -\mathbf{a_n}] = \sigma^2 (1 - \rho^2 \delta)$$

Where

σ is the variance of Y.
λ and δ are computed using $\mathbf{a_n}$ (not using - a_n)

Note that the functional form of λ and δ, as specified in equations 1 and 2 above, account for Y1 being greater than -a_n.

Consider the k'th varying parameter. If Y1 is ξ and Y is η_k, then

$$\beta_{nk}^{r|\theta} = B_{\mu k} + B_{\sigma k}\, \upsilon_{nkr}$$

becomes

(E1) $\beta_n^{r|\theta} = B_{\mu k} + (B_{\sigma k} B_{\rho k} \lambda_n) + B_{\sigma k} (1- B_{\rho k}^2 \delta_n)^{1/2} \upsilon_{nkr}$

Where

$B_{\sigma k}$ = the stdev of η_k -- *the k'th element on the diagonal of* Ω_n

$B_{\rho k}$ = the correlation coefficient between η_k and ξ

a_n $=Z\beta_z$ (a_n is the same for all alternatives for observation n)

The mixed logit replications, as described above, proceeds using (E1) to generate candidate values of replications $\beta_n^{r|\theta}$

Notes:

- This version of MIXED only supports 2-stage models with non correlated η (with diagonal Ω_n). Furthermore, log-normally distributed varying parameters are not supported (they can be included, but they are assumed to be uncorrelated with ξ).
- As with the regular mixed logit, each of the *K* varying parameters has its own value of B_σ ($B_{\sigma k}$). In addition, the *K'* parameters (a subset, possibly full, of the *K* parameters) that covary with the first-stage error have a value of B_ρ ($B_{\rho k}$). Alternatively, the same value of B_ρ can be used for all the *K'* parameters ($B_{\rho k}=B_\rho$; k=1..*K'*).

- λ_n and δ_n are observation specific (1x1) constants – they do not vary across replications (or across MLE iterations). They are computed using the first stage coefficients and first stage independent variables.

- The relevant gradients for observation replication/observation n|r, using E1, are:
  - $$dX\beta_n^{r|\Theta} \Big/ d\beta_{\sigma k} = X_n \left[\beta_{\rho k} \lambda_n + \nu_{nr} \sqrt{1 - \delta_{nr} \beta_{\rho k}^2} \right]$$
  - $$dX\beta_n^{r|\Theta} \Big/ d\beta_{\rho k} = X_n \beta_{\sigma k} \left[\lambda_n - \nu_{nr} \frac{\delta_n \beta_{\rho k}}{\sqrt{1 - \delta_n \beta_{\rho k}^2}} \right]$$

Other options

There are a number of other options that you can set in DISCRETE. The following briefly describes some of these options – please see DISCRETE.TXT for the details.

SEQ	The type of random values to generate. By default, MIXED uses a "scrambled halton sequence" to generate the random values used when simulating integration. If desired you can specify other sequences.
AUX and BAUX	Fixed-coefficient variables and betas
WEIGHT	A weight variable
SEED	Random number seed
HEADER	A short text message to display while the model is running
NORMALIZE	Normalize the independent variables

Appendix 1: Example of a dataset:

ID1	ALT	YEAR	VISIT1	X1	X2	X3	X4
1	1	80	1	2	51	62	22
1	1	81	0	0	12	63	4
1	1	82	0	0	52	77	6
1	2	80	2	0	56	11	1
1	2	81	1	1	1	4	6
1	2	82	0	2	2	426	12
1	3	80	0	5	6	16	6
1	3	81	1	10	12	33	3
1	3	82	0	0	61	123	61
1	1	80	0	0	12	6	3
2	1	81	0	1	1	6	6
Etc…							

Notes:

- Rows belonging to the same observation must be grouped together
- Rows belonging to the same choice occasion (for a given observation) do **not** have to be grouped together
- Each observation can have a different number of alternatives.
- Similarly, each choice occasion can have a different number of alternatives. For example, year 80 of observation 1 has four alternatives (years 81 and 82 have three alternatives).
- Each observation can have a different number of choice occasions
- Note that for observation 1 in year 80, alternative 1 was chosen once, while alternative 2 was chosen twice.

Appendix: Gradients and Likelihoods

Basic MNL

For individual n (suppressing the n subscript) the probability of choosing alternative c (from j=1…J alternatives)

$$\mathbf{P_c} = \exp(\mu_c) \Big/ \sum_j \exp(\mu_j)$$

where:

$u_j = X_j \beta$

Then,

$$d\,\mathbf{P_c}/d\beta = \left(\sum_j \left(D_{j=c} - P_j\right) \quad \partial_j \right) P_c$$

where

$D_{m=c} = 1$ if m=c, otherwise $D_{m=c} = 0$

$\partial_j = du_j / d\beta = X_j$ (a vector defined for the j'th alternative)

Note that the above is expressed in terms of likelihood – not log-likelihood. For the basic MNL, use of a log-likelihood is more efficient. However, in the models that follow, the log-likelihood is overly cumbersome.

Mixed Logit

We start with individual specific beta values:

$$\beta_n = b + \eta_n,$$

Estimation uses r=1..R replications, with a simulated β_n vector of

$$\beta_n^{r|\theta} = b + \Omega_0 \upsilon_{nr}$$

where Ω_0 is the "square root" of the covariance matrix of η, and υ_{nr} is a n'th observation specific vector of standard-normal random deviates.

This yields: $u_{nj|r} = X_{nj}\, \beta_n^{r|\theta}$

And a probability of choosing alternative c, for each replication (suppressing n):

$$\mathbf{P_{c|r}} = \exp(\mu_{c|r}) \Big/ \sum_j \exp(\mu_{j|r})$$

Across all replications, the likelihood for an observation is:

$$\mathbf{P_{c|R}} = \prod_{r=1..R} P_{c|r} \Big/ R$$

For the rth replication, and *l*'th parameter, ∂j has two components.

$\partial_{jbl} = du_j / db_l = X_{jl}$ *(∂_{jbl} does not depend on replication)*

and

$\partial_{j\Omega l|r} = du_j / d\Omega_l = \upsilon_{rl} X_{jl}$

Note that for non-varying parameters, $\partial_{j\Omega l} = 0$

Also note that if Ω is not diagonal, it gets more complicated (we ignore that for now).

Across all replications, the gradient of the log-likelihood is (suppressing the *l* subscripts for the L different parameters, and using vector notation for ∂):

$$d\ \mathbf{ln(P_{c|R}}\ )/db = \frac{\sum_{r=1..R}\left[\left(\sum_j \left(D_{j=c} - P_{j|r}\right)\ \ \partial_{jb}\right)P_{c|r}\right] \Big/ R}{P_{c|R}}$$

$$d\ \mathbf{ln(P_{c|R}}\ )/d\Omega = \frac{\sum_{r=1..R}\left[\left(\sum_j \left(D_{j=c} - P_{j|r}\right)\ \ \partial_{j\Omega|r}\right)P_{c|r}\right] \Big/ R}{P_{c|R}}$$

Uncertain dependent variables – quasi-bayesian (additive)

The probability of a choice occasion *(suppresing the n subscript on X and* β*):*

$$\mathbf{P_u} = \left(\sum_{k=1..K} \pi_k \exp(X_k' \beta\ ) \Big/ \sum_{j=1...K} \exp(X_j' \beta\ ) \right) = \sum_{k=1..K} \pi_k P_k$$

With a gradient of:

$$d\ \mathbf{P_u}/d\beta\ = \sum_j \Pi_j P_j \left(\partial_j - \sum_k \left[\partial_k P_k\right]\right)$$

If estimating a mixed model with uncertain Y: $\partial_{kbl|}$ and $\partial_{k\Omega l|r}$ should replace ∂_j, and a summation (over r=1..R) is added to the equation. For the log-likelihood this final summation is divided by the averaged (over the beta replications) likelihood.

Uncertain dependent variables – implicit replications (multiplicative)

The probability of a choice occasion:

$$\mathbf{P_u}=\left(\prod_{j=1..k}\exp(X_j\beta)^{\pi_j}\Big/\sum_{j=1...K}\exp(X_j\beta)\right)$$

For the gradient, we need the derivative of the numerator:

$$\frac{d}{d\beta}\prod_{j=1..K}\left[\exp(X_j\beta)^{\pi_j}\right]=\sum_{k=1..K}\left\{\left(\pi_k\exp(X_k\beta)^{\pi_k-1}X_k\exp(X_k\beta)\right)\left(\prod_{j=1,..,k-1,k+1,..,K}\exp(X_j\beta)^{\pi_j}\right)\right\}=$$

$$=\sum_{k=1..K}\left\{\left(\pi_k X_k\right)\left(\prod_{j=1..K}\exp(X_j\beta)^{\pi_j}\right)\right\}$$

Hence,

$$\frac{d\mathrm{Pu}}{d\beta}=\frac{\left\langle\sum_{k=1..K}\left\{\left(\pi_k\partial_k\right)\left(\prod_{j=1..K}\exp(X_j\beta)^{\pi_j}\right)\right\}\sum_{j=1...K}\left\{\exp(X_j\beta)\right\}\right\rangle-\left\langle\prod_{j=1..K}\exp(X_j\beta)^{\pi_j}\sum_{j=1..K}\partial_j\exp(X_j\beta)\right\rangle}{\left(\sum_{j=1...K}\exp(X_j\beta)\right)^2}$$

$$=\left(\sum_{k=1..K}\left(\pi_k\partial_k\right)\mathrm{P_u}\right)-\left(\mathrm{P_u}\sum_{j=1..K}\partial_j P_k\right)$$

$$=\left(\sum_{k=1..K}\left(\pi_k-P_k\right)\partial_k\right)\mathrm{P_u}$$

where, as above, $\partial_k = du_k / d\beta = X_k$

Xβ-replicated observations model uncertain independent variables model, with quasi-bayesian (linear) uncertain dependent variables

The probability of a choice occasion:

$$P_{xb}= \sum_{s=1..R_{xb}} \left(\sum_{k=1..K} \exp\left(X_k \beta + \delta_{sk} \sqrt{\beta' \Omega_x \beta}\right) \pi_k \Big/ \sum_{j=1..K} \exp(X_j' \beta + \delta_{sj} \sqrt{\beta' \Omega_j \beta}) \right) \Big/ R_{xb}$$

With a gradient

$$dP_{xb}/d\beta_= \sum_{s=1..R_{xb}} \left\{ \sum_{k=1...K} \left(\sum_{\kappa=1..K} D_{\kappa=k} - P_{s\kappa} \right) \mu'_{r\kappa} \right) \pi_k P_{sk} \Big\} \Big/ R_{xb}$$

where…

P_{sk} = probability of choice k in replication s: $\exp(\mu_{sk}) \Big/ \sum_{j=1...K} \exp(\mu_{sj})$

$$\mu_{sk} = X_k \beta + \delta_{sk} \sqrt{\beta' \Omega_x \beta}$$

$$\mu'_{sk} = d\mu_{sk} / d\beta = X_k + \delta_{sk} \frac{\Omega_x \beta}{\sqrt{\beta' \Omega_x \beta}}$$

$D_{k=6}$ equals 1 if k=6, other wise 0.
δ_{sk} is a draw from a standard normal
Ω_x is the variance matrix of the X variables.
R_{xb} is a number of uncertain independent variables draws.
π_k is the uncertain dependent variables posterior probability for choice k.

For the log likelihood, just divide $dP_{xb}/d\beta$ by the probability of the observation (Ciii, or its mixed logit approximation).

Multiple Choice Occasions

This section NEEDS WORK…
An observation may contain information on more then one choice occasion. This may consist of having many choices, having multiple choice occasions, or some combination of the two.

The former (multiple There may be more than one choicehere is more than one choice, m=1…M of the alternatives are chosen Y_m times, simply replicate these M terms. Thus, for an observation,