# Discrete Choice Models

## Multinomial Models

Mohammad Arshad Rahman
(webpage: https://www.arshadrahman.com)

Course: Econometrics
Chennai Mathematical Institute

# Multinomial Models

Many choice situations are over alternative that cannot be logically ranked. Such choices give rise to multinomial models. Examples: mode of transportation (car, bus, train, $\cdots$), job categories, recreational destinations, etc.

Two such models are the Multinomial Logit (MNL) and Multinomial Probit (MNP) models.

The MNL model is easy to implement, but often biased as it imposes independence across choice options. The MNP corrects this flaw, but is somewhat more difficult to estimate.

# Introduction

The MNL that we study here is what Greene, in the book *Econometric Analysis*, calls "Conditional Logit" model.

Greene's "Multinomial Logit" is a variation of that baseline model with option-specific parameters.

However, in most micro-economic applications the "Conditional Logit" setting makes more sense. Besides, our interpretation of the "MNL" is consistent with the broader statistical literature.

# Random Utility Framework (RUM)

Assume that an individual faces $j = 1, 2, \cdots, J$ options for a given good or service. He will derive utility from each option–this is the indirect utility function, since we are assuming that the individual has already optimized his consumption over all other goods.

As researchers, we relate this utility to a set of observable option features and an error term, which captures everything unobserved that links utility to a given option. Thus:

$$U_{ij}^{*} = x_{ij}'\beta + \varepsilon_{ij}, \quad i = 1, \cdots, n;\ j = 1, \cdots, J. \tag{1}$$

# Random Utility Framework (RUM)

In general, $x_{ij}$ will only include attributes corresponding to the $j$-th option. Here, we index $x$ by $j$ AND $i$, because of the possibilities:

(a) In many applications not every individual may face the exact same mix of attributes for a given option. For example, in contingent experiments[1], different individuals will be given different choice menus *by design* to enhance the properties of the resulting estimator.

(b) It is possible to introduce observed respondent characteristics into the model via *interactions* with the choice attributes.

[1] *Contingent experiments* are surveys that ask people how much they are willing to pay for or give up a product or service. They are used to estimate the value of a good or service.

# Random Utility Framework (RUM)

(c) One may allow $\beta$ vectors to vary over choices (yielding Green's version of the MNL model).

For simplicity, we will abstract from possibilities (b) and (c).

In addition, we will assume a linear utility function for simplicity. This is not a requirement, and many other functional forms have been explored in the literature.

# Random Utility Framework (RUM)

For each individual, we observe his chosen option out of the $J$ possibilities, Let's label the observed choice $y_i$. We can link these observed choices to the underlying latent framework via:

$$y_i = k \qquad \text{if} \quad \max\{U_{ij}\}_{j=1}^{J} = U_{ik} \tag{2}$$

Alternatively, we can write this as:

$$y_i = k \qquad \text{if} \quad U_{ik}^{*} > U_{ij}^{*}, \ \ \forall j \neq k, \qquad \text{or,} \tag{3}$$

# Random Utility Framework (RUM)

$$\begin{pmatrix} U_{ik}^* > U_{i1}^* \\ U_{ik}^* > U_{i2}^* \\ \vdots \\ U_{ik}^* > U_{i,k-1}^* \\ U_{ik}^* > U_{i,k+1}^* \\ \vdots \\ U_{ik}^* > U_{iJ}^* \end{pmatrix} \leftrightarrow \begin{pmatrix} \varepsilon_{ik} - \varepsilon_{i1} > (x_{i1} - x_{ik})'\beta \\ \varepsilon_{ik} - \varepsilon_{i2} > (x_{i2} - x_{ik})'\beta \\ \vdots \\ \varepsilon_{ik} - \varepsilon_{i,k-1} > (x_{i,k-1} - x_{ik})'\beta \\ \varepsilon_{ik} - \varepsilon_{i,k+1} > (x_{i,k+1} - x_{ik})'\beta \\ \vdots \\ \varepsilon_{ik} - \varepsilon_{iJ} > (x_{iJ} - x_{ik})'\beta \end{pmatrix}$$

# Multinomial Logit

The exact estimation framework now depends on the stipulated error distribution in equation (1). The MNL results if we assume a *Type 1 Extreme Value (EV Type 1)* or *Gumbel* distribution for the error terms, i.e.,

$$f(\varepsilon_{ij}) = \frac{1}{b} \exp\left[\frac{(\varepsilon_{ij} - a)}{b}\right] * \exp\left[-\exp\left[\frac{(\varepsilon_{ij} - a)}{b}\right]\right]$$
$$F(\varepsilon_{ij}) = \exp\left[-\exp\left[\frac{(\varepsilon_{ij} - a)}{b}\right]\right]. \tag{4}$$

The location "$a$" and scale "$b$" are customarily set to 0 and 1 in the standard MNL.

# Multinomial Logit

McFadden (1974) has shown that *iff* the error terms follow a Gumbel distribution, the choice probabilities can be expressed as,

$$\Pr(y_i = j) = P_{ij} = \frac{\exp(x'_{ij}\beta)}{\sum_{j=1}^{J} \exp(x'_{ij}\beta)}. \tag{5}$$

This expression is simple and leads to well-behaved MLE estimation. However, it also implies that any person-specific explanatory variables would drop out of the model. If we want to include them, we will have to do so via interaction with alternative-specific constants (ASCs) or any of the alternative-specific attributes[2].

[2] Inclusion of alternative-specific indicator terms only makes sense if every respondent receives the same set of options (e.g. choice of the same set of ski areas, vehicles, holiday destinations etc.).

# Multinomial Logit

The likelihood for the MNL model can be written as,

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} \left[ \sum_{j=1}^{J} \Pr(y_i = j) I(y_i = j) \right] \\
&= \prod_{i=1}^{n} \left[ P_{i1}^{\max[0,1-abs(2(y_i-1))]} * P_{i2}^{\max[0,1-abs(2(y_i-2))]} \ldots \right. \qquad (6) \\
&\qquad \left. P_{iJ}^{\max[0,1-abs(2(y_i-J))]} \right].
\end{aligned}
$$

# The IIA Assumption

A shortcoming of the MNL is its dependence on a strong assumption, called "Independence from Irrelevant Alternatives (IIA)". By construct, any ratio of choice probabilities is independent of all other available options, i.e.

$$\frac{\Pr(y_i = j)}{\Pr(y_i = k)} = \frac{\exp(x_{ij}'\beta)/\sum_{j=1}^{J}\exp(x_{ij}'\beta)}{\exp(x_{ik}'\beta)/\sum_{j=1}^{J}\exp(x_{ij}'\beta)} = \frac{\exp(x_{ij}'\beta)}{\exp(x_{ik}'\beta)} \tag{7}$$

Or, expressed as **log-odds ratios**:

$$\log\left[\frac{\Pr(y_i = j)}{\Pr(y_i = k)}\right] = (x_{ij}' - x_{ik}')\beta. \tag{8}$$

# The IIA Assumption

The IIA assumption is rarely consistent with consumer preferences and behavior. The classic example to illustrate the estimation bias that can arise when using the MNL model and the IIA assumption does not hold is the "`car-blue bus-red bus`" scenario:

Assume that originally a person has two transportation options, `car (c)` and `blue bus (bb)`. Assume both provide equal utility, so $\Pr(y_i = c) = \Pr(y_i = bb) = \frac{1}{2}$, leading to an odd-ratio of 1.

Now a third choice is introduced, say a `red bus (rb)`, which in all other aspects is identical to the blue bus. Thus the individual is indifferent between `bb` and `rb`, leading to $\Pr(y_i = bb)/\Pr(y_i = rb) = 1$.

# The IIA Assumption

However, the IIA assumption forces the MNL to hold the original odds-ratio of $\Pr(y_i = c)/\Pr(y_i = bb)$ at 1 as well. The only probabilities that lead to both $c/bb$ and $bb/rb$ having odds-ratios of 1 are: $\Pr(y_i = c) = \Pr(y_i = bb) = \Pr(y_i = rb) = \frac{1}{3}$.

This implies that the individual's probability of taking the car has decreased to $\frac{1}{3}$ with the introduction of the new alternative, and the probability of taking the bus (any bus) has increased to $\frac{2}{3}$. This is clearly counter-intuitive.

One would expect the probability of `car` remaining constant at $\frac{1}{2}$ and the probability of `bus` to be split between blue and red, $\frac{1}{4}$ each. However, the MNL would predict $\frac{1}{3}$ for all three–a clearly biased result.

# Test for IIA Assumption

Hausman and McFadden (1984) propose a simple test for the IIA assumption. Conceptually, it works as follows:

(a) Estimate the original MNL model with all available alternatives. Call this "full" estimator and its estimated variance, $\hat{\beta}_f$ and $\hat{V}_f$, resp.

(b) Then eliminate a subset of options and re-estimate the MNL model. Call this "restricted" estimator and its estimated variance, $\hat{\beta}_r$ and $\hat{V}_r$, respectively.

(c) If the IIA assumption holds, we would not expect the two estimators to be significantly different. This sets a natural stage for a standard Hausman test, computed as,

$$H = (\hat{\beta}_r - \hat{\beta}_f)'(\hat{V}_r - \hat{V}_f)^{-1}(\hat{\beta}_r - \hat{\beta}_f) \sim \chi^2_k. \tag{9}$$

where $k$ is the dimension of $\beta$.

# Model Fit and Comparison

The standard asymptotic tests (LR, Wald, and LM) are available to compare nested models. A generally applicable, useful measure of fit is McFadden's (1973) pseudo-$R^2$, given as,

$$R^2 = 1 - \frac{LL_f}{LL_0}, \tag{10}$$

where $LL_f$ is the value of the log-likelihood at convergence for the full model, and $LL_0$ is the analogous statistic for an intercept only model that estimates the probability for each alternative to be the sample average.

# Post-estimation Constructs

The *marginal effect* of a given attribute $k$ is associated with option $j$ on the choice probability for option $j$ can be derived as (assuming a linear utility function)

$$\frac{\partial \Pr(y_i = j)}{\partial x_{kj}} = \frac{\partial P_{ij}}{\partial x_{kj}} = \beta_k P_{ij}(1 - P_{ij}). \tag{11}$$

Similarly, the effect of this change on some other option, say $l$, becomes

$$\frac{\partial \Pr(y_i = l)}{\partial x_{kj}} = \frac{\partial P_{il}}{\partial x_{kj}} = -\beta_k P_{ij} P_{il}. \tag{12}$$

Thus, the sign of $\beta_k$ can be unambiguously interpreted as the direction of its `own-effect` and the negative of the direction of its `cross-effect`. The standard error of ME can be computed via the Krinsky-Robb method or equivalent asymptotic techniques.

# Hit-rate

Hit-rate is defined as the number of cases, out of the sample, for which the actually observed choice receives the highest probability. Formally, for a sample of size $n$, the hit-rate is

$$HR = \frac{1}{n} \sum_{i=1}^{n} I\left( \left( \max_j \{\hat{p}_{ij}\}_{j=1}^{J} \right) = y_i \right), \tag{13}$$

where $\hat{p}_{ij}$ is the predicted probability that individual $i$ selects outcome $j$, and $I(\cdot)$ is the indicator function as defined earlier. The hit-rate is generally considered an "informal but useful" statistic.

# New Alternative

In some case, the MNL has also been used to forecast choice probabilities for a new alternative, characterized by attribute vector $x_p$ and added to the bundle of existing options (see Train's BART example in Ch. 3). The point estimate for the corresponding predicted probability can be computed as,

$$P_p = \frac{\exp(x_p'\hat{\beta})}{\left[\sum_{j=1}^{J} \exp(x_{ij}'\hat{\beta}) + \exp(x_p'\hat{\beta})\right]}, \tag{14}$$

and standard errors can be obtained via the Krinsky-Robb method or equivalent asymptotic techniques.

# Multinomial Probit

The Multinomial Probit (MNP) is an alternative to the MNL that relaxes the IIA assumption by introducing correlations across error terms in the utility function.

At the individual level the latent model can be written as,

$$U_i = X_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \Sigma). \tag{15}$$

In the absence of any researcher-imposed structural restrictions on $\Sigma$, the model needs to normalized for *level* and *scale*.

(a) *Level* normalization if not done, adding a constant to all utilities won't change the observed outcome.

(b) *Scale* normalization if not done, multiplying all utilities by the same scalar won't change the observed outcome.

# Level Normalization

Level normalization can be accomplished by declaring the utility of for one of the alternatives (say the first) as "baseline", and differencing all other utilities with respect to that baseline.

Naturally, this only makes sense if the exact same baseline alternative appears in all choice menus for all individuals. If this is not the case, the researcher needs to make arbitrary structural restrictions on $\Sigma$ such as declaring it an identity matrix (in which case the IIA assumption kicks again into full gear).

# Level Normalization

In choice experiments, the baseline is often some "status quao" alternative, such as not developing some piece of land, or "staying home" instead of choosing a recreation site. Let's assume the first option ($j = 1$) is the baseline. We can then express the model in terms of utility differences from the baseline, i.e.,

$$\begin{aligned} U_{ij}^* &= (x_{ij} - x_{i1})'\beta + \varepsilon_{ij}^*, \qquad j = 2, \ldots, J, \quad i = 1, \cdots, n \\ U_{ij}^* &= x_{ij}^{*'}\beta + \varepsilon_{ij}^*, \end{aligned} \tag{16}$$

where in this case the $*$ symbol indicates a differenced construct.

# Level Normalization

The system of $J-1$ random utility differences for person $i$ can be written as

$$U_i^* = X_i^* \beta + \varepsilon_i^*. \tag{17}$$

Train (2003) shows a convenient "trick" to quickly compute the variance matrix for the differenced errors.

Assume you start out with $J$ alternatives (including the baseline). After differencing, you are left with $J-1$ utility differences. Thus, declare a $(J-1) \times (J-1)$ identity matrix and "squeeze in" an extra column of "-1"s in the position of the original baseline alternative.

# Level Normalization

For example, if $J = 4$ and the first alternative is the baseline, the resulting *Differencing Matrix* becomes:

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \tag{18}$$

and the differenced errors have the following distribution:

$$\varepsilon_i^* = D\varepsilon_i \sim N(0, \Sigma^*), \qquad \Sigma^* = D\Sigma D'. \tag{19}$$

# Scale Normalization

For normalization with respect to *scale*, one of the variance elements in $\Sigma^*$ needs to be fixed. The standard approach is to set $\sigma_{11} = 1$, as we did for the Probit model. Thus, in theory, $\Sigma^*$ can include $((J-1)J/2)-1$ free elements.

Train (2003) shows in detail how the elements in $\Sigma^*$ relate to the elements in the original $\Sigma$. He points out that while $\Sigma^*$ still allows for substitution patterns (and thus overcomes the restrictive IIA requirement for logit-type models), no intuition can be gained from inspection of its estimated elements with respect to the original variances and covariances. This subtle but important point is often missed in applied research.

# Remark

Keep in mind that for applications with changing choice options across respondents, it is not meaningful to exante specify an unrestricted $\Sigma$ matrix, since the definition of "option 1, 2, 3, etc" is not consistent over respondents.

In such a case, it may be preferable to set $\Sigma = I$ as in the seminal paper by Hausman and Wise (1978). This automatically normalizes the model for level and scale, and we do not need to exante work with utility differences. However, in absence of any other model refinements, a MNP with $\Sigma = I$ fares no better than the MNL with respect to the IIA dilemma.

In summary the MNP will estimate $\Sigma^*$. This provides no intuition upon the original $\Sigma$, but at least relaxes the IIA assumption.

# Multinomial Probit

For each respondent we observe $y_i$, i.e., a scalar denoting the actual choice (or–equivalently–a $J \times 1$ vector with zeros and a "1" for the chosen option).

We can relate the observed choice index to latent utility differences as follows:

$$\begin{aligned} y_i = 1 \quad &\text{if} \quad \max\{U_{ij}^*\}_{j=2}^J \leq 0 \\ y_i = k \quad &\text{if} \quad \max\left\{0, \{U_{ij}^*\}_{j=2}^J\right\} = U_{ik}^* \end{aligned} \tag{20}$$

# Multinomial Probit

An individual's contribution to the likelihood function can thus be expressed as a multivariate normal cdf, with choice-specific truncation bounds. For example, a choice of "1" implies:

$$\Pr(y_i = 1) = \Pr\begin{pmatrix} U_{i2}^* \leq 0 \\ U_{i3}^* \leq 0 \\ \cdots \\ U_{iJ}^* \leq 0 \end{pmatrix} = \Pr\begin{pmatrix} \varepsilon_{i2}^* \leq -x_{i2}^{*\prime}\beta \\ \varepsilon_{i3}^* \leq -x_{i3}^{*\prime}\beta \\ \cdots \\ \varepsilon_{iJ}^* \leq -x_{iJ}^{*\prime}\beta \end{pmatrix} = \Phi(0, \Sigma^*; R_1) \tag{21}$$

where $\Phi(\cdot)$ denotes the cdf of the truncated multivariate normal density with mean 0 and variance matrix $\Sigma^*$, and truncation region $R_1$ implicitly defined by the condition $(y_i = 1)$.

# Multinomial Probit

For some other choice, say "$k$", we can write the joint probability vector as,

$$\begin{aligned}
&\Pr(y_i = k) \\
&= \Pr\begin{pmatrix} U^*_{ik} - U^*_{i2} > 0 \\ U^*_{ik} - U^*_{i3} > 0 \\ \cdots \\ U^*_{ik} - U^*_{iJ} > 0 \end{pmatrix} = \Pr\begin{pmatrix} \varepsilon^*_{ik} - \varepsilon^*_{i2} \leq -(x^*_{ik} - x^*_{i2})'\beta \\ \varepsilon^*_{ik} - \varepsilon^*_{i3} \leq -(x^*_{ik} - x^*_{i3})'\beta \\ \cdots \\ \varepsilon^*_{ik} - \varepsilon^*_{iJ} \leq -(x^*_{ik} - x^*_{iJ})'\beta \end{pmatrix} \qquad (22) \\
&= \Phi(0, D_k \Sigma^* D_k'; R_k)
\end{aligned}$$

Note: the truncated cdf is defined based on `differences` of differenced errors. This step alters the variance-covariance matrix from $\Sigma^*$ to $D_k \Sigma^* D_k'$, where $D_k$ is a $(J-1)\text{by}(J-1)$ differencing matrix. It is crucial to account for this second differencing step.

# Multinomial Probit

The differencing matrix can be quickly constructed for all possible choice cases as follows: Start with a negative identity matrix of dimension $(J-1)$. Replace the $k$-th column (corresponding to an observed choice of "$k$") with a vector of 1's.

For example, for $J = 4$ and $k = 2$, we would get:

$$D_2 = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix}. \tag{23}$$

# Multinomial Probit

Generally, the likelihood contribution for the $i$th individual can then be concisely expressed as,

$$l_i(\beta, \Sigma^*) = \sum_{j=1}^{J} \Big( \Phi(0, D_j \Sigma^* D_j') * I(y_i = j) \Big) \tag{24}$$

where, for $y_i = 1$, $D_j$ is simply an identity matrix. This leads to the following sample likelihood function:

$$L(\beta, \Sigma^*) = \prod_{i=1}^{n} \left[ \sum_{j=1}^{J} \Big( \Phi(0, D_j \Sigma^* D_j') * I(y_i = j) \Big) \right] \tag{25}$$

# Multinomial Probit

The evaluation of the multivariate truncated normal terms is a major challenge in classical estimation. Usually, this is accomplished via simulation methods such as the GHK algorithm (after Geweke, Hajivasilou, Keane). For a good discussion, see Train (2003).

Luckily Matlab's inbuilt function `mvncdf` allows for the added specification of truncation bounds. Based on my (limited) experience with this function it is quite accuarte, but can be very slow, depending on the spacing of the truncation bounds relative to the mean and to each other.

# Multinomial Probit

Another frequent problem is that the joint probability terms can become too small to be computationally distinguishable from zero, leading to a ln(0) problem in the likelihood.

Some creative programming can circumvent these issues, but in any case some serious tuning with starting values, step size, and perturbations of numerical derivatives is generally needed to get the MLE algorithm to converge.

# Multinomial Probit

In summary, I would suggest to start with the MNL, and only consider the MNP if the IIA test results indicates a violation of the IIA assumption for a given application. However, even in that case there easier to implement and (therefore) more popular models, such as the "Mixed Logit" (see Greene's Book).

All predictive constructs are again based on joint probabilities associated with a truncated multivariate normal density. They are rarely applied and we will thus abstract from a further exploration of this topic.

# Thank you!