

Lecture 7:

Estimating the Multinomial Logit Model

by Maximum Likelihood

(Greene, 19.7)

1. Multinomial choice models in general

2. MNL

Notational note: In this lecture, observations (individuals/choosers) are indexed by $n = 1, \dots, N$. Choice options are indexed by $j = 1, \dots, J$. “ i ” and “ j ” are used to indicate two different options between which the chooser is choosing.

1. Multinomial choice models

Very often in economics, we are interested in *which one* of a finite number of non-ordered, mutually exclusive alternatives is selected by a decisionmaker:

- make of car chosen by a consumer (GM, Ford, Toyota, Honda...)
- occupation selected by a worker
- heat home by natural gas, oil, electric
- travel to work by car, bus, subway, walk, bike...

How can we model/measure the effects of (a) characteristics of the chooser (e.g. consumer age and income), and (b) characteristics of the choice (e.g. price, features, tax incentives) on such decisions?

We need a model of multinomial choice, preferably one consistent with utility-maximization by the chooser.

In general, a multinomial choice model imagines N decisionmakers (these will be the observations in the data) choosing among J alternatives (the chosen alternative is the outcome).

Denote the utility obtained by decisionmaker n from choice j by U_{nj} , $j = 1, \dots, J$. Let $U_{nj} = V_{nj} + \varepsilon_{nj}$, where V_{nj} is a function of things the econometrician observes (prices of alternatives, demographics of consumers); and ε_{nj} is an unobserved *choice-specific* error term (for example, consumer n 's taste for fast cars).

Assume that decisionmakers pick the alternative yielding the highest utility to them, i.e. they pick alternative i iff:

$$U_{ni} > U_{nj}, \forall j \neq i. \text{ , or}$$

$$\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i \quad (1)$$

From the point of view of the researcher, who does not observe the ε 's, the probability that individual n chooses alternative i is thus:

$$\begin{aligned} & \Pr(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i) \\ &= \int_{\varepsilon_n} I(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i) f(\varepsilon_n) d\varepsilon_n, \end{aligned} \quad (2)$$

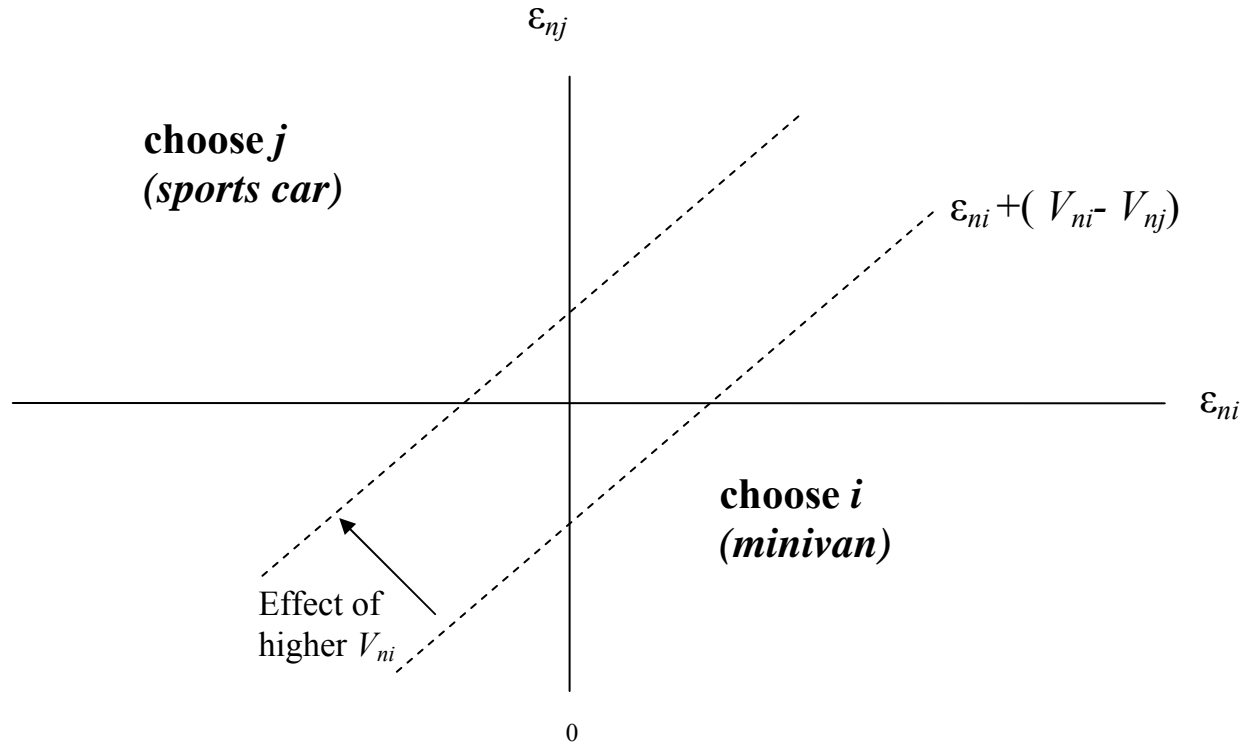
Where I is the indicator function (equals one when the condition is true, zero when false).

Computationally, this is a pretty challenging integral.

To illustrate, in the case of just two choices, i and j , (1) reduces to:

Choose option i iff $\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}$.

Diagrammatically:



For any given observable determinants of preferences (V_{ni} and V_{nj}), the level of ε_{nj} needed for the household to prefer option j rises with the level of ε_{ni} .

When V_{ni} rises (e.g. the consumer has a child, which typically raises preferences for household's preference for a option i —the minivan). The households who change their behavior when this happens are those whose ε_{ni} and ε_{nj} lie between the dotted lines. For households who continue to buy the sports car (option j) both before and after they have a kid, the econometrician infers that they must have a high ε_{nj} (more specifically, an ε_{ni} and ε_{nj} northwest of the higher dotted line), and uses this information to estimate parameters of the model.

As the number of choices expands, the dimensionality of these integrals rises –which can dramatically complicate matters computationally-- but the same types of areas must be integrated over.

Speaking of parameters, typically we assume that $V_{nj} = \beta_j \mathbf{x}_n + \alpha \mathbf{z}_j$, thus:

$$U_{nj} = \beta_j \mathbf{x}_n + \alpha \mathbf{z}_j + \varepsilon_{nj} \quad (3)$$

In the above notation, \mathbf{x}_n varies across choosers (households), but not across choices (e.g. income, presence of children). Note that it has a *different* coefficient vector for each choice, j . (Why? If kids had the same effect on the utility of every type of car, they would not affect anyone's choice. Chooser characteristics only matter if they have *different* effects on the utility of various choices.) [it follows that we can only estimate $J-1$ β vectors—one must be treated as the reference category].

On the other hand, \mathbf{z}_j varies across choices (e.g. the price of the car model) but not (in the above formulation) across households. So we estimate only one α vector.

In the logit context, models where (3) does not include any \mathbf{z}_j 's are often called multinomial logit models, models *with* \mathbf{z}_j 's are called conditional logits. But they are just different versions of the same thing.

2. MNL

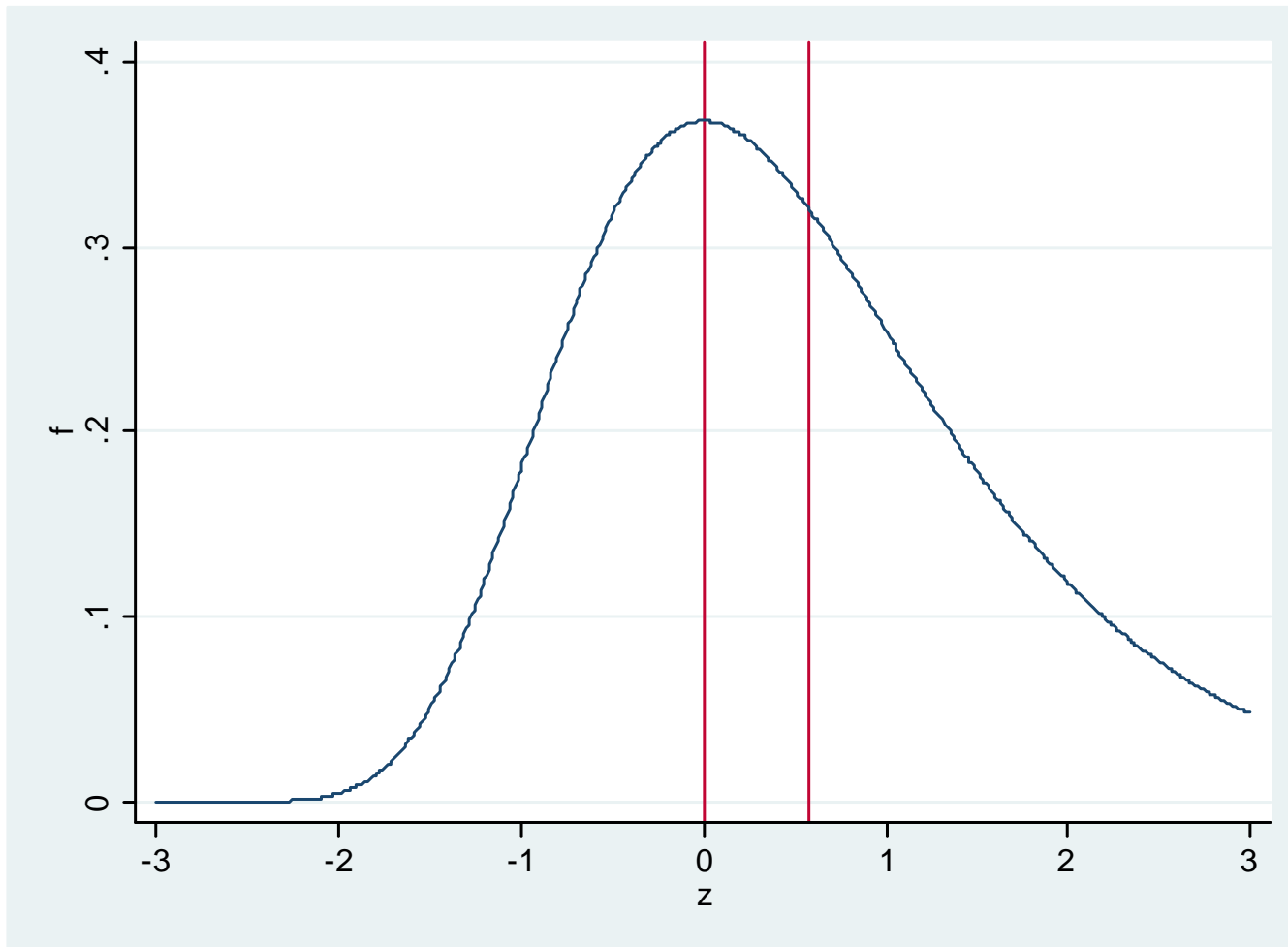
In general, the likelihood function for a multinomial choice model will be a product of n $J-1$ dimensional integrals of the form in (1) above.

Even when all the ε_{nj} 's are independent and identically distributed (both across n 's and across j 's) these are computationally difficult to compute for almost all distributions.

An easy case, however, is when $f(\varepsilon_{nj}) = \exp(-\varepsilon_{nj}) \cdot \exp(-\exp(-\varepsilon_{nj}))$, and therefore:

$$F(\varepsilon_{nj}) = \exp(-\exp(-\varepsilon_{nj}))$$

This is the “standard” *type-1 extreme value* distribution. Here's what it looks like:



The “standard” extreme value density is defined on the entire real line, has a single mode at 0 and a mean equal to Euler’s γ (about .577).

Theorem: If $f(\varepsilon_1)$ and $f(\varepsilon_2)$ are iid (standard) extreme value, then $\varepsilon_1 - \varepsilon_2$ follows the logistic distribution.

Recall that in the general multinomial choice model with only two choices i and j , the decisionmaker chooses i iff:

$\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}$; using $V_{nj} = \beta_j \mathbf{x}_n$ (no choice-specific variables) this becomes:

Choose i iff: $\varepsilon_{nj} - \varepsilon_{ni} < \beta_i \mathbf{x}_n - \beta_j \mathbf{x}_n$

or: $\varepsilon_{nj} - \varepsilon_{ni} < (\beta_i - \beta_j) \mathbf{x}_n \equiv \beta \mathbf{x}_n$, so (if $\varepsilon_n \equiv \varepsilon_{nj} - \varepsilon_{ni}$ is logistic), the likelihood of choosing option i is just $\frac{\exp(\beta \mathbf{x}_i)}{1 + \exp(\beta \mathbf{x}_i)}$.

This is just the logit model for a binary choice that we developed in the previous lecture. Just as in the binary choice model, where the coefficient vector, β , represented the effect of the x 's on the utility *difference* between the two options (in that example the options were "work (W)" and "not work(N)", so β equalled $\beta^W - \beta^N$), here with two choices all we can estimate is a single parameter vector, $\beta = \beta_i - \beta_j$, that gives the effect of the x 's on the relative utility of the two options.

Generalizing the above to the case of $J > 2$ choices, the number of parameter vectors that can be estimated will always be $J-1$. We will handle this from here on by letting choice 1 be the ‘reference’ choice; thus setting $\boldsymbol{\beta}_1 = \mathbf{0}$. We then estimate a parameter vector $\boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \dots, \boldsymbol{\beta}_J$ for each of the remaining choices, and interpret these $\boldsymbol{\beta}_j$ ’s as the effect of \mathbf{x} on the utility of option j *relative to* option 1.

More formally, McFadden has (famously) shown that when

1) Decisionmakers pick the choice yielding the highest utility from among J choices with utility given by $U_{nj} = \boldsymbol{\beta}_j \mathbf{x}_n + \varepsilon_{nj}$

2) All of the ε_{nj} ’s are iid (standard) extreme value

Then the probability that option i is chosen (i.e. that its utility exceeds that of all the other choices) is given by:

$$\Pr(U_{ni} > U_{nj}, \forall j \neq i) = \frac{\exp(\boldsymbol{\beta}_i \mathbf{x}_{ni})}{\sum_{j=1}^J \exp(\boldsymbol{\beta}_j \mathbf{x}_{ni})} \quad (4)$$

(recall that $\boldsymbol{\beta}_1 = \mathbf{0}$, and note how the probabilities of all J choices sum to 1).

Finally, to estimate the MNL model by maximum likelihood, we need an expression for the log likelihood of our entire observed sample, given its x 's and its observed choices from among the J alternatives.

Using (4), we can do this observation by observation, calculating each observation's likelihood based on its \mathbf{x} and the actual choice it made. Then (as always) we take logs, and sum log likelihoods across observations to obtain the log likelihood for the entire sample.

You will do this in this week's problem set.

A final note:

Just as in the probit and logit models, the estimated β_j 's (recall you will estimate $J-1$ coefficient vectors) do **not** tell you the effect of the \mathbf{x} 's on the probability that choice j will occur. (Instead, they give the effect of \mathbf{x} on the “utility” of option j *relative to* option 1 –at least if we adopt a utility-maximizing interpretation of our estimates).

To understand the effects of a change in \mathbf{x} on the predicted distribution of choices (i.e. effect of population aging on the share of the population buying minivans, sedans, sports cars versus SUVs), it is best and quite easy to calculate the predicted distribution of choices at actual and counterfactual values of \mathbf{x} , as we did in the logit and probit cases.