

## GEV Defined

Consider a function  $Y(\cdot)$  with the following properties:

1.  $Y(\nu_1, \dots, \nu_J)$  is a non-negative function with non-negative arguments;
2.  $Y(\nu_1, \dots, \nu_J)$  is linearly homogenous in its  $J$  arguments;
3.  $\lim_{\nu_i \rightarrow \infty} Y(\nu_1, \dots, \nu_J) = +\infty$  for all  $i = 1, \dots, J$ ;
4. The partial derivative  $\partial^k Y / \partial \nu_{i_1} \partial \nu_{i_2} \dots \partial \nu_{i_k}$  is non-negative for  $k$  odd and non-positive for  $k$  even, provided that  $i_1, \dots, i_k$  are distinct.

The random variables  $\epsilon_1, \dots, \epsilon_J$  are said to have a **generalized extreme value** (GEV) distribution if their c.d.f. is expressible in the form:

$$F_\epsilon(E_1, \dots, E_J) = \exp(-Y[\exp(-E_1), \dots, \exp(-E_J)])$$

and the joint p.d.f. is found in the usual way by partial differentiating with respect to each of the  $\epsilon_i$ . (p.d.f, awkward and not informative.)

It is called the GEV because its marginal distributions are all of displaced univariate extreme value form.

An important special case of the GEV corresponds to linearity in the function  $Y$ . Can show that a linear  $Y$  produces the joint distribution for  $J$  independent extreme value variables. Thus,  $Y$  must be nonlinear to introduce non-independent distributions.

McFadden shows that the choice probabilities are

$$\begin{aligned} P(j|w_n; \theta) &= \frac{\partial \ln Y[\exp(V_{n1}^*), \dots, \exp(V_{nJ}^*)]}{\partial V_{nj}^*} \\ &= \frac{\exp(V_{nj}^*) Y_j[\exp(V_{n1}^*), \dots, \exp(V_{nJ}^*)]}{Y[\exp(V_{n1}^*), \dots, \exp(V_{nJ}^*)]} \end{aligned}$$

where  $Y_j$  is the partial derivative of  $Y$  with respect to its  $j^{\text{th}}$  argument.

The MNL is a member of the GEV with  $Y(\nu_1, \dots, \nu_J) = \sum \nu_i$ .

The choice probabilities above imply elasticities of the form

$$\frac{\partial \ln P(j|w_n)}{\partial \ln x^i} = \left[ \frac{Y_{ji}}{Y_j} \exp(V_{ni}^*) - P(i|w_n) \right] \frac{\partial V_i^*(x^i, z_n)}{\partial \ln x^i}$$

where  $Y_j$  and  $Y_{ji}$  are the first and second derivatives of  $Y(\cdot)$  evaluated at  $\exp(V_{n1}^*), \dots, \exp(V_{nJ}^*)$ . The elasticities are independent of  $j$  only if the cross-derivatives  $Y_{jk}$  are proportional to  $Y_j$ .

## Nested MNL

We need to be more specific about the choice set  $B$  and its representation. McFadden assumed that alternatives are defined by their characteristics. For example, extend the choice set in the transportation example from (car, bus) to (car, bus, car pool). In this case, the alternative car pool shares certain features with each car and bus and so one partition into subsets (as is necessary for the Nested MNL) would be to include the option car pool to appear in each subset, (car, car pool) and (bus, car pool). Thus, the nested logit framework does not presume that subsets are mutually exhaustive. Indeed, the formal requirement only is that the partition be collectively exhaustive: for a choice set with  $J$  choices, a partition with  $K$  subsets,  $(B_k)_{k=1}^K$ , then  $\bigcup B_k = B$ .

The function  $Y$  that generates the nested logit is:

$$Y(\nu_1, \dots, \nu_J) = \sum_{k=1}^K a(Z_k) \left[ \sum_{i \in B_k} \nu_i^{\rho(z_c)} \right]^{\frac{1}{\rho(z_c)}}$$

where  $a(z_c) \geq 0$  and  $\rho(z_c) = 1/1 - \sigma(z_c)$  with  $0 \leq \sigma(z_c) < 1$ . Where  $B_k$  is a subset of alternatives from the choice set  $B = (1, \dots, J)$ .  $z_c$  corresponds to the vector of attributes of all the choices in subset  $B_k$ ,  $z_k = \{z_i\}_{i \in B_k}$ . The choice probabilities can be written as

$$P(i|B) = \sum_{k=1}^K P(i|B_k)Q(B_k|B)$$

where

$$P(i|B_k) = \frac{\exp(\rho(z_k)V_i^*)}{\sum_{i \in B_k} \exp(\rho(z_k)V_i^*)} \quad \text{for } i \in B_k$$

$$Q(B_k|B) = \frac{a(z_k) \exp(h_k/\rho(z_k))}{\sum_{k=1}^K a(z_k) \exp(h_k/\rho(z_k))}$$

with  $h_k = \ln \sum_{i \in B_k} \exp(V_i^*)$  where I have dropped the explicit conditioning on person or case index  $n$ .

The system can be interpreted as a choice system in which decision makers invoke a subset of alternatives  $B_k$  from (the complete choice set)  $B$  and then select an alternative from  $B_k$ . So the  $P(i|B_k)$  represents the selection at the lower level of the system (“selection of the twig”). While  $Q(B_k|B)$  is the choice of the subset (“selection of the branch”). Notice that  $P(i|B_k)$  has a **MNL** form. Associated with each subset  $B_k$  is an **inclusive value**,  $h_k$ . Choice probabilities for the invoked set  $k$  are multinomial logit functions of the inclusive values. The function  $\sigma(z_k)$  is a measure of the similarity of alternatives with  $B_k$ . When alternatives in  $B_k$  are very similar  $\sigma(z_c)$  is near one, the conditional choice probability  $P(i|B_k)$  selects with high probability the alternative with the highest value of  $B_k$  of  $V_i^*$ . Then  $h_k$  is approximately  $\max V_i^*$  and the choice of an invoked set using  $Q(B_k|B)$ , the set  $B_k$  is assessed approximately as if it contained a single alternative with a scale value of

$$\max_{i \in B_k} V_i^*.$$

I re-learned the hard way last Thursday there are many different parameterizations of the Nested Logit. Some as above in terms of  $\rho$  while others are in terms of  $\sigma$ , with  $\rho = 1/(1 - \sigma)$ , and just to be confusing still others define  $\rho = 1 - \sigma$ .

I have shown the most general representation of the Nested Logit with the  $a(z_k)$  functions. More common, the “standard” sets  $a(z_k) \equiv 1$ . Cameron and Trevedi report parameterization sensitive to the the parameterization of  $a$ .

More complicated models possible with nesting of higher levels. Inclusive values arise at each level.

Useful to give a concrete example to show the basic representation. But the key, as mentioned above is that MNL characterize choice probabilities at the bottom level (among the “twigs”) while dependence in preferences are across the subsets.

**Give example of parameterization in yellow–page handwritten notes or from Train.**

## MVN

Alternatively, can work off first difference form and recognize that that dimension of the integral defining the choice probabilities is  $J - 1$  which can sometimes help.

**Go to Hand written notes.**

## Computation of MNP

Let me follow the notation of Geweke and Keane (chapter 56, Vol 5 Handbook of Econometrics) and denote the discrete choice model as

$$y_{ij}^* = Z_{ij}\gamma + X_i\beta_j^* + \epsilon_{ij}^*, \quad j = 1, \dots, J, \quad i = 1, \dots, N.$$

Notice that to retain individual-specific characteristics in the model it is necessary to include choice specific coefficients  $\beta_j^*$ . The idea is that  $X_i$  are taste shifters (e.g., if  $X_i$  is age, then  $\beta_j^*$  captures differences in preferences by age).  $y_{ij}^*$  are latent random variables.

Define alternative  $J$  as the base alternative and define

$$\begin{aligned} y_{ij} &= y_{ij}^* - y_{iJ}^* = (Z_{ij} - Z_{iJ})\gamma + X_i(\beta_j^* - \beta_J^*) + (\epsilon_{ij}^* - \epsilon_{iJ}^*) \\ &= (Z_{ij} - Z_{iJ})\gamma + X_i\beta_j + \epsilon_{ij} \quad j = 1, \dots, J - 1 \\ y_{iJ} &= 0 \end{aligned}$$

where  $\epsilon_i \equiv (\epsilon_{i1}, \dots, \epsilon_{iJ-1}) \sim N(0, \Sigma)$  and  $\Sigma$  is a  $(J-1) \times (J-1)$  covariance matrix obtained from  $\Sigma^*$ . A further scale normalization is usually imposed by setting  $\Sigma_{11} = 1$ .

It is convenient to write  $y_{ij} = \bar{y}_{ij} + \epsilon_{ij}$  for  $j = 1, \dots, J$  and adopt the convention that  $\bar{y}_{iJ} = \epsilon_{iJ} = 0$ . Then the agent chooses option  $j$  if and only if  $y_{ik} - y_{ij} \leq 0 \forall k$ , which generates the  $J-1$  dimension partition of the  $\epsilon$  space  $\epsilon_{ik} - \epsilon_{ij} \leq \bar{y}_{ij} - \bar{y}_{ik} \quad \forall k$ . Define  $\tilde{\epsilon}_{ik}^j = \epsilon_{ik} - \epsilon_{ij}$  for  $k = 1, \dots, J$  and further define  $\tilde{\epsilon}_i^j = (\tilde{\epsilon}_{i1}^j, \dots, \tilde{\epsilon}_{ij-1}^j, \tilde{\epsilon}_{ij+1}^j, \dots, \tilde{\epsilon}_{iJ}^j) \sim N(0, \tilde{\Sigma}^j)$ .

Then the probability that agent  $i$  chooses option  $j$  can be written as the  $J-1$  dimension integral:

$$\begin{aligned} p(j|Z_i, X_i, \gamma, \beta, \Sigma) &= \int_{-\infty}^{\bar{y}_{ij} - \bar{y}_{i1}} \cdots \int_{-\infty}^{\bar{y}_{ij} - \bar{y}_{iJ}} p\left(\tilde{\epsilon}_1^j, \dots, \tilde{\epsilon}_J^j | \tilde{\Sigma}^j\right) d\tilde{\epsilon}_J^j \cdots d\tilde{\epsilon}_1^j \\ &= P\left(\bar{y}_{ij} - \bar{y}_{i1}, \dots, \bar{y}_{ij} - \bar{y}_{iJ} | \tilde{\Sigma}^j\right). \end{aligned}$$

Letting  $d = (d_1, \dots, d_J)$  the MNP likelihood function is:

$$p(d|\gamma, \beta, \Sigma) = \prod_{i=1}^N \prod_{j=1}^J p\left(\bar{y}_{ij} - \bar{y}_{i1}, \dots, \bar{y}_{ij} - \bar{y}_{iJ} | \tilde{\Sigma}^j\right)^{1[d_j=j]}.$$

The computationally demanding task of MNP estimation is the evaluation of the  $J-1$  dimension integrals that comprise the likelihood function.

Geweke and Keane state that deterministic methods (e.g., quadrature) are useful only for very low dimensions,  $J = 3, 4$ . **Because** the likelihood has to be evaluated at a large number of *trial* values of the  $K \times 1$  parameter vector  $\theta = (\gamma, \beta, \Sigma)$ . At each trial value  $\theta_T$  the  $J-1$  dimensional integrals must be evaluated for all  $N$  agents in the population. Moreover, if using a **derivative** based search algorithm to maximize the likelihood function, it is necessary to evaluate the likelihood at several different values of the parameter vector. These include the (1) initial value of  $\theta_T$ , and (2)  $K$  incremented parameter values  $\theta_T + h\Delta_k$ , where  $h$  is the increment size, and  $\Delta_k = 1$  for  $k = 1, \dots, K$  and zero otherwise (used to evaluate the derivatives), and (3) the new trial parameter vector  $\theta'_T$  of which a line search algorithm will always try at least two. Thus, each iteration involves **at least**  $K + 3$  evaluations of the likelihood.

Geweke and Keane note that in a small problem with  $N = 500$  and  $J = 4$ , with twelve parameters, and 5 free elements of  $\Sigma$ , optimization may require 50 iterations. Then approximately  $50 \cdot (12 + 3) \cdot 500 = 375000$  three dimensional integrals must be evaluated. At 100 evaluations per second (which seems fast even today), would require more than an hour of computation.

It is this curse of dimensionality that stimulated researchers to consider simulation based procedures. **Simulated Maximum Likelihood** (SML) applies fast simulation methods to approximate  $p(j|Z_i, X_i, \gamma, \beta, \Sigma)$  and to insert these approximations into the likelihood function.

Early attempts (e.g., Manski and Lerman in *The White Monster*) used crude frequency simulators

$$\hat{p}(j|Z_i, X_i, \gamma, \beta, \Sigma) = M^{-1} \sum_{m=1}^M 1 \left[ \tilde{\epsilon}_1^{j(m)} \leq \bar{y}_{ij} - \bar{y}_{i1}, \dots, \tilde{\epsilon}_J^{j(m)} \leq \bar{y}_{ij} - \bar{y}_{iJ} \right],$$

where  $\left\{ \tilde{\epsilon}_k^{j(m)} \right\}_{k=1}^J$  for  $m = 1, \dots, M$  are iid draws from the joint  $\tilde{\epsilon}_i^j \sim N(0, \tilde{\Sigma}^j)$ .

The frequency simulators are most intuitive and yet consensus is that frequency simulators did not perform satisfactory.