Title
Approximate Generalized Extreme Value Models of Discrete Choice

Author
Small, Kenneth A.

Publication Date
1988-04-01

UCI-ITS-WP-88-5

Approximate Generalized Extreme Value Models of Discrete Choice

UCI-ITS-WP-88-5

Kenneth A. Small

School of Social Sciences and
Institute of Transportation Studies
University of California, Irvine

April 1988

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

ABSTRACT

Estimation of generalized extreme value (GEV) models of discrete choice is hampered by computational complexity and convergence problems. However, the much simpler estimation routine for multinomial logit can be applied in a two-step procedure so as to test the null hypothesis of multinomial logit against any particular GEV model as an alternative hypothesis. The procedure also produces an approximate estimate of the GEV model. Monte Carlo data, generated alternatively by logit and by three different GEV models, provide evidence that both the test statistics and the approximate estimator have small-sample properties superior in important respects to maximum-likelihood estimation of the GEV model.

This work was supported by the National Science Foundation and the Institute of Transportation Studies, University of California, Irvine. I am grateful to Feng Zhang for research assistance. I also thank the following people for comments on earlier drafts: David Brownstone, Lester Johnson, Kevin Lang, Daniel McFadden, Jon Sonstelie, Clifford Winston, two anonymous referees, and seminar participants at California Institute of Technology, University of California at Santa Barbara, and University of Southern California.

APPROXIMATE GENERALIZED EXTREME VALUE MODELS OF DISCRETE CHOICE

I. INTRODUCTION

Recently McFadden (1978, 1981) has specified a broad class of discrete-choice models known as Generalized Extreme Value (GEV), which relax some restrictive properties of multinomial logit. Like logit, GEV models can be derived from random-utility maximization. However, their use has been limited by computational complexity.

This has contributed to a strong interest in specification tests for the logit model. But all such tests have disadvantages. Some require estimating a more general model, either GEV (Hausman and McFadden, 1984) or probit (Horowitz, 1981), and hence are similarly thwarted by computational difficulties. Others involve deleting alternatives from the choice set (McFadden, Train and Tye, 1977; Hausman and McFadden, 1984; Small and Hsiao, 1985); but their power varies widely depending on which alternatives are deleted (Horowitz, 1981, pp. 352-353), a decision for which little guidance is available. A recently proposed Lagrange Multipler test (McFadden, 1987) has thus far been developed only for one particular alternative model--nested logit--and has unknown performance in practice.

This paper considers yet another type of specification test, proposed by McFadden, Train, and Tye (1977) and sometimes called "universal logit." It is based on the fact that a logit estimation routine will accept a "variable" whose value for one alternative depends on traits of other alternatives. Since no such variable can arise in the usual random-utility-based logit model, I call it a "pseudovariable"; it should attain an insignificant coefficient if the logit model is correct.

As yet, no one has provided much guidance for constructing pseudovariables. Furthermore, the test gives little information about the source of error if logit is rejected (Horowitz, 1981, p. 353); indeed, in the absense of a well-defined alternative model, one cannot even distinguish between rejecting the logit model itself and rejecting of the specification of its variables.

This paper solves these problems by presenting a computationally simple pseudovariable test of the logit model against any prespecified alternative model of the GEV class. The test is based on an approximation to the alternative model, whose exact choice probabilities need not even be derived explicitly. The test generalizes a special case proposed by Small (1987), and can easily handle GEV models of considerable complexity. It requires only standard logit software, and provides as a byproduct a crude estimate of the alternative model's parameters. Its asymptotic distribution permits inference using statistics computed by standard logit estimation routines.

Monte Carlo investigation of the test reveals several desirable properties. First, its finite-sample distribution is much closer to its asymptotic distribution than is true for tests based on the maximum-likelihood estimator (MLE) of the alternative model. Second, its power against the prespecified alternative is at least as high as other tests in situations where their nominal sizes are accurate. Third, the test is quite specific, showing much higher power against the model for which it was designed than against other models. Fourth, the approximate estimate of the alternative model that it provides, although inconsistent, is often better behaved in finite samples than the MLE. Hence the test is recommended as an easy way to check for a suspected violation of logit and to explore the properties of the corresponding alternative.

II. THEORY

A GEV model for choice among J discrete alternatives is generated by specifying a function $G(y_1,\ldots,y_J)$ which (a) is non-negative on the hyper-quadrant $y_j > 0$; (b) is homogeneous of degree one in $y \equiv (y_1,\ldots,y_J)$; (c) approaches infinity as any y_j does; and (d) has n-th partial derivatives which are non-negative for odd n and non-positive for even n. The probability of choosing alternative k is defined to be:

$$P_k = \frac{\exp(V_k)G_k}{G} \equiv \frac{\exp(V_k+\log G_k)}{\sum_{j=1}^{J}\exp(V_j+\log G_j)} \tag{1}$$

where G and its first partial derivatives G_j are evaluated at arguments $y_\ell = \exp(V_\ell)$. The quantity V_j is a function of unknown parameter vector β and of data vector z_j describing characteristics of both the sample member and the alternative;[1] for simplicity, I assume a linear relation

1. I could equally well have assumed $V_j = \beta_j' z$, letting the parameters rather than the variables vary across alternatives. This latter parameterization is more familiar to labor economists. As Maddala (1983, p. 42) shows, the two are equivalent.

$$V_j = \beta' z_j \quad . \tag{2}$$

The second version of (1) follows from Euler's theorem, as noted by Ben-Akiva and Lerman (1985, p. 127), since G is homogeneous of degree one.

McFadden (1978) shows that these probabilities can be generated by assuming a random-utility model in which the (indirect) utility achieved with choice of alternative j is

$$U_j = V_j + \varepsilon_j \quad , \tag{3}$$

where $\{\varepsilon_j\}$ are random components whose joint distribution is generalized extreme value. In the special case $G(y) = \sum_j y_j$, (1) and (2) reduce to logit:

$$P_k = \frac{\exp(V_k)}{\sum_{j=1}^{J} \exp(V_j)} \quad . \tag{4}$$

A pseudovariable test is carried out by replacing (4) by

$$P_k = \frac{\exp(V_k+\theta' x_k)}{\sum_j \exp(V_j+\theta' x_j)} \tag{5}$$

which contains pseudovariables x with unknown coefficients θ. For example, if z_j is the cost of alternative j, the alternatives might be divided into two groups B_1 and B_2, and x_j defined as the cost averaged over all the alternatives in whichever group includes j.

The particular pseudovariable test proposed here is based on the resemblance between equations (1) and (5). We need only parameterize G

by $\theta = (\theta^1,\ldots,\theta^I)'$, chosen in such a way that $G(y;\theta) \underset{\theta\to 0}{\to} \sum_j y_j$. When θ is small, representing a small departure from logit, $\log G_\ell$ in equation (1) can be approximated by $\theta' G_{\theta\ell}$, where $G_{\theta\ell}$ is an I-vector whose i-th component is $\partial^2 G/\partial\theta^i \partial y_\ell$ evaluated at $y_j = e^{V_j}$ and $\theta = 0$. This yields equation (5) with pseudovariables[2] defined by

$$x_\ell = G_{\theta\ell}(e^{V_1},\ldots,e^{V_J};\ 0) \qquad \ell = 1,\ldots,J. \tag{6}$$

We may regard equations (5) and (6) as defining an <u>approximate GEV model</u>. Although (5) emphasizes its similarity to the logit model, it cannot be estimated in a single logit step because β appears both directly as the coefficients of z and indirectly in the definitions of x. However, when θ is small, we can estimate β and θ in two steps. First, estimate β using maximum-likelihood on (4). Denote the result $\tilde{\beta}$. Second, using pseudovariables $\tilde{x}$ constructed by substituting $\tilde{\beta}$ for β in (6), estimate both β and θ by maximum-likelihood on (5). Denote these estimates by $\hat{\gamma} = (\hat{\beta}',\hat{\theta}')'$. Note that both estimation steps use only a logit routine. Note also that, since the parameters θ enter (5) linearly, equality restrictions among them are easily imposed by combining variables.

Under the null hypothesis that $\theta = 0$, $\tilde{\beta}$ is consistent for β; hence $\tilde{x}$ is consistent for x; hence the second-stage estimate $\hat{\gamma}$ is

---

2. In the terminology adopted here, there is one pseudovariable x^i for each component of θ; it takes on value x^i_ℓ for alternative ℓ (for a particular member of the sample, whose index is suppressed). Hence $x_\ell \equiv (x^1_\ell,\ldots,{}_\ell x^I)'$.

consistent for $\gamma \equiv (\beta',\theta')'$. The more formal proof given in Small (1987, Appendix B) for ordered GEV applies here as well. Furthermore, it is shown there that, under the null hypothesis, (i) $\sqrt{N}(\hat{\gamma}-\gamma)$ is asymptotically normally distributed with mean zero and variance-covariance matrix $\tilde{\mathcal{I}}^{-1}$, where N is sample size and

$$\tilde{\mathcal{I}} = -\operatorname*{plim}_{N\to\infty} \frac{1}{N} E\, \tilde{L}_{\gamma\gamma'}(\hat{\gamma};\, \tilde{x}) \qquad (7)$$

is the information matrix associated with the second-stage log-likelihood function $\tilde{L}$ formed from (5) holding x at $\tilde{x}$; (ii) either of the variance-covariance estimates commonly employed at the second state, viz. $-\tilde{L}_{\gamma\gamma}^{-1}$ or $(E\tilde{L}_{\gamma}\tilde{L}_{\gamma}')^{-1}$ where E indicates the sample average, is, when multiplied by N, consistent for $\tilde{\mathcal{I}}^{-1}$; and (iii) the usual likelihood-ratio statistic $2[\tilde{L}(\hat{\beta},\hat{\theta};\tilde{x}) - \tilde{L}(\tilde{\beta},0;\tilde{x})]$ is asymptotically chi-square distributed with I degrees of freedom. Hence, one can use the second-stage estimates or their log-likelihood value in the usual ways to form a t-test, Wald test, or likelihood ratio test of the null hypothesis.

These results seem counter-intuitive because there are several examples in econometrics of multi-step tests in which variances are underestimated at the last step. This does not happen here because all parameters are simultaneously reestimated at the second stage. In fact, the test is closely related to the Lagrange Multiplier (LM) test, which is asymptotically efficient. The LM test can be written as a quadratic form in the score L_θ evaluated at $\tilde{\gamma} = (\tilde{\beta}',\, 0')'$. Using (1), this means that each sample member contributes the following quantity to L_θ:

$$l_\theta = \sum_j d_j \frac{\partial \log P_j}{\partial \theta} = \sum_j (d_j - P_j^o)\tilde{x}_j, \qquad (8)$$

where P_j^o is the logit probability (4) and d_j is the choice variable (equal to one if j is chosen, zero otherwise). Hence the LM test, like mine, looks for a tendency to choose alternatives with unusual values of x.

My test is even more closely related to a version of the LM test recently proposed by McFadden (1987, Section 4). His test is based on ordinary least-squares regression of normalized residuals $\tilde{u}$ (from the logit estimation) on the variables z and a set of pseudovariables w. These w satisfy $L_\theta = w'\tilde{u}$, an equation of the same form as (8). McFadden's test is defined against two-level nested logit, and his w are straighforward transformations of my x for that case.[3] Hence it appears that equation (6) could serve as a way to generate regression-based tests against other GEV models.

To summarize, the procedure proposed here provides a rigorous way to derive pseudovariables for testing particular types of suspected departures from logit. At the same time, it provides an estimate of the statistical model (5), which approximates a random-utility model of the GEV class. Although this estimate is not consistent, the Monte Carlo evidence reported later suggests that it is in some ways preferable to the MLE of the alternative model when θ is small.

---

3. Compare McFadden's equation (50) with my equation (A.4).

III. EXAMPLES

Equation (6) is calculated for several GEV models in the Appendix, with the results summarized in Table 1. The models include nested logit (NL) (McFadden, 1978, 1981) with both two and three levels; ordered GEV (OGEV) (Small, 1984); and a new model which I call "nested ordered GEV" (NOGEV) that groups alternatives as in NL but allows OGEV-type ordering within groups. To satisfy the GEV axioms, all these models require the components of θ to lie in the unit interval. These components are denoted by σ_r, ϕ_q, and ϕ_q^r; the corresponding pseudovariables are denoted by N^r, M^q, and M^{rq}.

The intuition behind the approximate estimator is readily grasped by considering the two-level NL model, whose choice probabilities are those of a tree decision structure:

$$P_k = \frac{\exp(V_k/\rho_s)}{\sum_{j\epsilon B_s} \exp(V_j/\rho_s)} \cdot \frac{\exp(\rho_s I_s)}{\sum_{r=1}^{R} \exp(\rho_r I_r)}. \tag{12}$$

In (12), the subsets B_r form a partition of the choice set $\{1,\ldots,J\}$; s indexes the subset that contains alternative k; ρ_r are parameters in the unit interval; and

$$I_r = \log \sum_{j\epsilon B_r} \exp(V_j/\rho_r) \tag{13}$$

is a measure of the desirability of subset B_r. The parameter $\sigma_r = 1-\rho_r$ is related to the correlation among the stochastic terms ϵ_j

within subset B_r; hence when σ_r is large, the conditional choice within B_r is likely to yield whichever alternative in B_r has the largest systematic utility.

This property of NL is reflected in the pseudovariable test. Consider, for example, the constrained model $\sigma_r \equiv \sigma$. There is just a single pseudovariable $N = \sum_r N^r$ which takes values:

$$\begin{aligned} N_j &= -\log \sum_{\ell \epsilon B_s} (P_\ell^\circ / P_j^\circ) \\ &= V_j - I_s^\circ \end{aligned} \tag{14}$$

where P_ℓ° are the logit choice probabilities (4), and

$$I_s^\circ = \log \sum_{\ell \in B_s} \exp(V_\ell) \quad . \tag{15}$$

Thus N_j measures the relative advantage of alternative j within its subset: when N_j is large, indicating that j is grouped in a subset with few and/or unattractive alternatives, j is more likely to be chosen because it dominates its close competitors. The strength of this effect depends on the closeness of these competitors, which is measured by σ. The stronger is this effect, the more important is N in explaining observed choices hence the larger is the estimate σ.

The intuition for other GEV models is similar, and is developed in Small (1987) for ordered GEV.

It is striking how little extra complexity is added by going to a three-level model, such as three-level NL or NOGEV. The approximate estimator simply involves an extra set of pseudovariables which are no more difficult to calculate than those for the two-level models. In contrast, the complexity of maximum-likelihood estimation rises dramatically with additional levels in the tree structure.

IV. SOME MONTE CARLO RESULTS

In order to further investigate properties of the pseudovariable test proposed here, it is useful to measure its performance in samples generated randomly from known statistical models. This section describes experiments with samples generated by three of the models from the previous section: logit, nested logit (NL), and ordered GEV (OGEV). Pseudovariable tests based on approximations to these models and to nested ordered GEV (NOGEV) are constructed, as are t-tests and likelihood-ratio tests based on the maximum-likelihood estimates (MLEs) of the same models. A Hausman-type test of the kind studied by Hausman and McFadden (1984), hereafter referred to as the HM test, is also constructed.

Experimental Design

Three questions guided the experiments' design. First, how do the finite-sample distributions of various test statistics compare with their asymptotic distributions? Second, what is the power of these tests when the true model is the one against which the test is explicitly designed, and when it is not? Third, what are the distributions of the estimators themselves?

Answering these questions required a specification for which different GEV models produce observationally distinguishable results, and for which assuming the wrong model leads to bad coefficient estimates. Given these requirements, I selected as simple a specification as possible. A population of individuals, observationally identical except for their observed choices, faces six alternative travel modes. Alternative j has an observed "full cost" c_j which is listed in Table 2, and its utility V_j is equal to $\log(1/c_j)$ plus an error term. When the error terms are independently extreme-value distributed, the resulting logit choice probabilities P_j^0 are those shown in the last column.

The NL model is a two-level one with the first three and the last three alternatives grouped, and the corresponding parameters constrained equal: $B_1 = \{1,2,3\}$, $B_2 = \{4,5,6\}$, and $\sigma_1 = \sigma_2 \equiv \sigma$. (It could result, for example, from unobserved traits affecting access to a private motorized vehicle.) The OGEV model is that defined in Table 2 with $K=1$, $w_1=w_2=1/2$, and $\sigma_r \equiv \sigma$; this could arise if the error terms for immediately adjacent alternatives are correlated (perhaps due to unobserved preferences regarding physical effort, which decreases with j). The NOGEV model has the same kind of ordering, but within each of the groups B_1 and B_2: i.e., it is defined as in Table 2 with $R=2$, $K=1$, $\sigma_1=\sigma_2=\sigma$, and $\phi_q^r \equiv \phi$.

Table 3 shows just one of the choice probabilities, P_3, under each of these models as σ varies from 0 (logit) to 1 (a limiting case). Also shown is the approximate value of P_3 given by eqns. (5) and (6). The limiting cases as $\sigma \to 1$ are versions of "maximal" models, in which certain conditional probabilities become discontinuous in the utilities -- see McFadden (1978, p. 85) for NL, and Small (1987, pp. 414-415) for OGEV.

The following explanation of how the error distribution affects P_3 may aid intuition. Under NL, $P_3 > P_3^0$ because alternative 3 (bus) is grouped with unattractive alternatives, hence it receives a disproportionate vote from those with a tendency to avoid private motorized vehicles. Under OGEV, $P_3 < P_3^0$ because alternative 3 is adjacent to a very attractive alternative (motorcycle) and hence tends to lose its natural constituency of people with preference for middling levels of physical effort. Under NOGEV, P_3 is higher even than under NL because not only is it grouped with the same unattractive alternatives as in NL, but within that class it is adjacent to an alternative with lower systematic utility than itself (bicycle).

Our hypothetical researcher's specification is:

$$V_j = \beta_1 D_{3j} + \beta_2 \log(1/c_j) \quad , \tag{16}$$

where D_{3j} is an alternative-specific dummy (1 when $j=3$, 0 otherwise). The true parameter vector is $\beta = (\beta_1,\beta_2)' = (0,1)'$. Because P_3 is sensitive to the statistical model, this specification leads to error if the wrong statistical model is assumed.

For each statistical model, 400 samples of size 200 and 400 samples of size 1000 were generated randomly. The smaller samples are "small" in that they produce estimates of β, σ, and ϕ with sizeable standard deviations (typically over 0.3 in the case of σ and ϕ); they also produce occasional convergence problems, indicating that one can have trouble estimating even a correctly specified model on a sample of this size. The samples of 1000 are "large," producing standard deviations that are usually below 0.15. With 400 replications, further replications caused little change in the empirical sampling distributions of the estimators and test statistics. Estimates were computed using the quadratic hill-climbing algorithm of Goldfeld and Quandt (1972, pp. 5-8) as implemented in their computer program GQOPT.

Distributions of Test Statistics Under Null Hypothesis

Table 4 characterizes the empirical sampling distributions of various test statistics when the null hypothesis is true (i.e., using samples generated by the logit model). The t-statistic is $\hat{\sigma}/\hat{SD}(\hat{\sigma})$ or $\hat{\phi}/\hat{SD}(\hat{\phi})$; the test depicted is one-sided. The chi-square statistic is the usual likelihood-ratio statistic except in the last column. That column shows the HM statistic for testing against NL by comparing two logit estimates: one with full choice set, the other with restricted choice set B_2.

The entries in the table are the proportions of replications for which the given test rejects the null at asymptotic significance levels of 0.5, 0.1, and 0.05. Each of these entries has a sampling variance due to the finite number of replications: when the true probability of rejection is π, the proportion of R replications for which rejection occurs has mean $R\pi$ and standard deviation $[\pi(1-\pi)/R]^{1/2}$. With 400 replications, this standard deviation never exceeds .025; it equals .025, .015, and .011 for π = .5, .1, and .05, respectively.

Keeping these in mind, the two-step tests and the HM test appear to have true sizes close to their asymptotic sizes. The tests based on the MLE, however, reject much too often when set for asymptotic significance levels of 0.1 or 0.05. For example, estimating the NL model on data generated by logit produces MLE t-statistics that exceed the 10% critical value about 16% of the time, and that exceed the 5% critical value about 10% of the time (not counting several cases where the MLE did not converge). The NL likelihood-ratio test seems not to be affected by this problem, suggesting that flat spots on the likelihood function are contributing to it. The large samples perform no better than small ones in this respect, though further investigation showed that samples of 10,000 brought the frequencies within the expected range. The OGEV MLE produces even worse results, at least for the small samples: the t-test rejects even more frequently, and the likelihood-ratio test is also unreliable. The NOGEV MLE estimates, not shown in the table, behaved similarly to NL though not quite as badly.

A closer look at these t-statistics, in Table 5, shows just how far their distributions are from normal. The t-statistic obtained from the MLE in the smaller samples has large skewness and kurtosis, combining to cause the heavy right tail just noted. The skewness persists in samples of 1000, but finally dwindles in samples of 10,000, verifying the asymptotic distribution.

One reason for this behavior is lack of precision in estimating $SD(\hat{\sigma})$. In the case of the NL MLE with samples of 200, the actual standard deviation of $\hat{\sigma}$ over the 400 replications was 0.46; yet $\hat{SD}(\hat{\sigma})$ ranged from 0.31 to 3.78, with mean 0.43 and standard deviation

0.45. By contrast, the approximate estimator produced not only a $\hat{\sigma}$ with smaller actual standard deviation (0.32), but a much tighter $\hat{SD}(\hat{\sigma})$, with mean 0.32 and standard deviation only 0.04.

These results cast serious doubt on the practical usefulness of logit specification tests based on MLE estimation of GEV models. The researcher would too often be led to reject the true specification, and to estimate a GEV model that deviates radically from the true model.

<u>Power</u>

Table 6 shows the power of several tests designed to test against NL. The first two panels show the power when the true model is NL, at various values of σ; the last two panels show the power of the same tests when the true model, contrary to the researcher's suspicions, is OGEV. Four tests are shown: two based on the two-step estimator, one on the MLE, and one on the HM statistic. Because the MLE one-sided t-test did not have the correct size, as noted earlier, its power would be misleading and hence is not shown.

This table answers several questions. First, which test has the highest power? Comparing tests against NL when NL in fact is true, we find that, as expected, the one-sided t-test has somewhat higher power than any of the two-sided tests. The three chi-square tests are virtually identical. Hence when the researcher suspects an alternative model for which a one-sided test is appropriate, the good small-sample distribution of the two-step t-statistic gives it a distinct advantage.

Second, how high is the power? It is reasonably high when σ is between 0.1 and 0.3 for large samples, or between 0.3 and 0.5 for small samples. At σ=0.5, the large samples rejected the null 100% of the time no matter what test was used.

Third, are the tests specific against a particular GEV model? The answer is yes: all four tests have much lower power against a true model that differs from the one against which they are designed. Their power against OGEV with σ=0.3 is only about as great as that against NL with σ=0.1. (If this property is not desired, one might design omnibus tests by simultaneously including pseudovariables derived from a variety of possible alternative models; but this possibility is not explored here.)

Although not shown in the table, results with tests designed against OGEV confirmed the first and third conclusions: higher power is attained by one-sided tests and by tests designed for the correct alternative model. However, power was substantially lower for a given value of σ than was the case for NL. Also, the OGEV MLE had numerical and convergence problems on about one-fourth of the small samples generated by NL with σ=0.5; no such problems occured with the two-step estimator.

<u>Distributions of Estimators</u>

Table 7 shows some properties of the empirical distributions of $\hat{\beta}_2$ and $\hat{\sigma}$ estimated by the 2-step approximation procedure and by maximum likelihood, when the researcher correctly guesses the form of the true model. These results are for N=1000.

In all cases, the MLE provides estimates with negligible bias and reasonably low standard deviations. The two-step estimates are very close to the MLE at low values of σ but, as expected from the nature of the approximation, deviate more and more as σ rises to 0.3 and beyond.

The MLE has a decided edge over the two-step estimator in precision, as measured by the actual standard deviations of $\hat{\beta}_2$ and $\hat{\sigma}$. This is especially true at higher values of σ. (Still greater precision is attained by the logit estimator, whose $\hat{\beta}_2$ when the true model is logit was unbiased to three decimal places and had a standard deviation 0.051, skewness 0.30, and kurtosis 3.45.)

The standard-error estimates, which are computed from the Hessian of the log-likelihood function, are all centered very close to the actual standard deviations. However, as noted earlier, the MLE standard-error estimates have a much greater variance (not shown in the table) than those from the two-step procedure.

The non-normality of the t-statistics discussed earlier has its counterpart for the parameter estimates. In all but one instance the MLE estimators depart substantially from normality, with $\hat{\sigma}$ typically having skewness around -0.8 and kurtosis between 3.7 and 4.4; the negatively skewed $\hat{\sigma}$ confirms a finding of Brownstone and Small (forthcoming, Table 2) using Monte Carlo on a much more complex model. The two-step estimators, in contrast, are close to normal.

<u>Root-Mean-Squared Error</u>

Since the real world is rarely precisely logit, we may ask: What degree of departure is needed for better results to be obtained from a more complex econometric model? Table 8 addresses this question in one way, by comparing the root-mean-squared error (RMSE) in various estimators of $\hat{\beta}_1$ or $\hat{\beta}_2$. Of course, this is not the only consideration for choosing a model: others include parsimony, prediction error in various situations, and transferability to other data sets.

With either sample size, the logit estimator is superior on the RMSE criterion at $\sigma = 0$ and $\sigma = 0.1$. The MLE becomes better somewhere between $\sigma = 0.1$ and $\sigma = 0.3$, and it strongly dominates at $\sigma = 0.5$. The two-step estimator is somewhat better than the MLE at very low σ, but this advantage erodes quickly as σ rises and in no case shown does the two-step estimator dominate both of the others.

This suggests that the two-step estimate of β may not be very useful. If the main interest is in estimating β, for example to measure marginal rates of substitution, an attractive procedure would be the following. Use the rather easily constructed pseudovariable test and its approximate estimator of θ to determine if logit is suitable either (a) because it is accepted at conventional significance levels, or (b) because the departure is too small to make a practical difference. If so, use logit. If not, use maximum likelihood to estimate the more general GEV model identified by the pseudovariable test, perhaps on an independent sample; but don't trust the MLE's estimated standard deviations without further verification by bootstrapping or Monte Carlo.

V. CONCLUSION

Approximating arbitrary GEV models by logit models with pseudovariables provides a flexible and powerful way to design logit specification tests. The test statistics are easier to compute and better behaved in finite samples than those based on the corresponding maximum-likelihood estimators. They provide an estimate of the approximate model that, though not consistent, gives strong clues as to the true behavior generating the data.

The Monte Carlo experiments reported here are limited to one basic design and hence cannot describe the full range of interesting conditions. Nevertheless, the evidence suggests that the maximum-likelihood estimators of nested logit and ordered GEV models have noticeably non-normal distributions in practical samples. As a consequence, specification tests based on them may falsely reject the null hypothesis far more often than indicated by the tests' nominal size.

As for the approximate estimator itself, the Monte Carlo evidence is mixed. When the departure from logit is small, the approximate estimator is quite accurate and hence might be suitable for prediction or simulation; but the logit estimator produces utility-parameter estimates with smaller root-mean-squared errors. When the departure from logit is large, the approximate estimator is a reliable indicator of this fact, and could therefore help the researcher decide on an alternative model for more thorough investigation.

Table 1. Some Alternative Models and Corresponding Pseudovariables

Model	Generating G-function and Component(s) of θ	Pseudovariables
NL-2	$\sum_{r=1}^{R} (\sum_{j\in B_r} y_j^{1/\rho_r})^{\rho_r}$	
	$\sigma_r \equiv 1-\rho_r$	$N_j^r = -\delta_{j\in B_r} \cdot \log \sum_{\ell\in B_r} (P_\ell^\circ/P_j^\circ)$
NL-3	$\sum_{r=1}^{R} \sum_{q\in C_r} \otimes \sum_{j\in D_q} y_j^{1/\gamma_q)^{\gamma_q/\rho_r}\ \rho_r}$	
	$\sigma_r \equiv 1-\rho_r$	N_j^r: same as NL-2, with $B_r \equiv \bigcup_{q\in C_r} D_q$
	$\phi_q \equiv 1-(\gamma_q/\rho_r)$	$M_j^q = \delta_{j\in D_q} \cdot \log \sum_{\ell\in D_q} (P_\ell^\circ/P_j^\circ)$
OGEV	$\sum_{r=1}^{J+K} (\sum_{j\in B_r} w_{r-j}\, y_j^{1/\rho_r})^{\rho_r}$	
	$\sigma_r \equiv 1-\rho_r$	$N_j^r = -\delta_{j\in B_r} \cdot w_{r-j} \log \sum_{\ell\in B_r} w_{r-\ell} (P_\ell^\circ/P_j^\circ)$
NOGEV	$\sum_{r=1}^{R} \sum_{q=\underline{Q}_r}^{\bar{Q}_r}{}^{r} (\sum_{j\in D_q^r} w_{q-j}\, y_j^{1/\gamma_q^r)\ \gamma_q^r/\rho_r\ \rho_r}$	
	$\sigma_r \equiv 1-\rho_r$	N_j^r: same as NL-2, with $B_r \equiv \bigcup_{q=\underline{Q}_r}^{\bar{Q}_r}{}^{r} D_q$
	$\phi_q^r \equiv 1-(\gamma_q^r/\rho_r)$	$M_j^{rq} = -\delta_{j\in D_q^r} \cdot w_{q-j} \log \sum_{\ell\in D_q^r} w_{q-\ell} (P_\ell^\circ/P_j^\circ)$

Notes to Table 1:

NL-2: nested logit, 2 levels; $\{B_r\}$ is a partition of $\{1,\ldots,J\}$.

NL-3: nested logit, 3 levels; $\{C_r\}$ is a partition of $\{1,\ldots,J\}$ and $\{D_q, q \in C_r\}$ is a partition of C_r.

OGEV: ordered GEV; $B_r \equiv \{r-K,\ldots,r\} \cap \{1,\ldots,J\}$.

NOGEV: nested OGEV; $\{B_r\}$ is a partition of $\{1,\ldots,J\}$ with $B_r = \{J_{r-1}+1,\ldots,J_r\}$; $\underline{Q}_r = J_{r-1}+1$; $\bar{Q}_r = J_r+K$; $D_q^r = \{q-K,\ldots,q\} \cap B_r$.

$\delta_{j \in A} = 1$ if $j \in A$, 0 otherwise

$$P_i^0 = \exp(V_i) / \sum_{j=1}^{J} \exp(V_j)$$

Restrictions:

$$0 < \rho_r \leq 1$$

$$0 < \gamma_r \leq 1$$

$$\sum_{\ell=0}^{K} w_\ell = 1$$

Table 2. Model for Monte Carlo Experiments

Alternative (j) No.	Name	Full Cost (c_j)	Logit Probability (P_j^o)
1	Walk	2.00	.05
2	Bicycle	2.00	.05
3	Bus	1.00	.10
4	Motorcycle	0.25	.40
5	Carpool	0.40	.25
6	Drive Alone	0.66...	.15

Table 3. Values of P_3 in Monte Carlo Experiment

	σ					
	0	0.1	0.3	0.5	0.8	→1
NL:						
exact	.100	.104	.114	.132	.187	.200
approx.	.100.	.103	.110	.116	.125	.131
OGEV:						
exact	.100	.095	.084	.073	.069	.071
approx.	.100	.096	.087	.079	.069	.062
NOGEV with $\phi=\sigma$:						
exact	.100	.106	.124	.150	.193	.200
approx.	.100	.105	.115	.122	.129	.131

Table 4. Distribution of Test Statistics under Null Hypothesis[a]

	Two-Step Estimator				MLE		Hausmann-McFadden Test[c]
	NL	OGEV	NOGEV σ	NOGEV ϕ	NL[b]	OGEV	
N=200							
Pr[t > 0]	.512	.522	.500	.528	.509	.540	--
Pr[t > 1.282]	.110	.122	.105	.118	.155**	.195**	--
Pr[t > 1.645]	.060	.060	.042	.048	.102**	.125**	--
$Pr[\chi^2 > c_{.50}]$	.475	.475	.502		.471	.485	.470
$Pr[\chi^2 > c_{.10}]$	.112	.098	.100		.109	.262**	.105
$Pr[\chi^2 > c_{.05}]$	.048	.052	.045		.048	.215**	.045
N = 1000							
Pr[t > 0]	.550	.538	.530	.525	.550	.538	--
Pr[t > 1.282]	.120	.112	.090	.102	.162**	.142**	--
Pr[t > 1.645]	.055	.062	.042	.050	.100**	.095**	--
$Pr[\chi^2 > c_{.50}]$	.510	.488	.528		.510	.488	.508
$Pr[\chi^2 > c_{.10}]$	.108	.108	.092		.108	.110	.108
$Pr[\chi^2 > c_{.05}]$	.048	.062	.045		.048	.065	.048

[a]For each test statistic, the three relative frequencies shown are computed from 400 replications (393 in case of NL-MLE), and correspond to 50%, 10%, and 5% significant levels. If the statistic had its asymptotic distribution, these frequencies would have mean (SD) of .500 (.025), .100 (.015), and .050 (.011), respectively.

[b]With sample size N = 200, 7 samples (out of 400) produced a nonconvergent NL MLE estimator, hence are excluded.

[c]Based on comparison of logit estimates: one with full choice set {1,...,6}, the other with restricted choice set {4,5,6}.

[d]Critical values for 50%, 10%, and 5% significance levels are 1.386, 4.605, and 5.991 for the χ^2-statistic based on the NOGEV estimator (2 degrees of freedom); they are .455, 2.706, and 3.841 for all the other χ^2-statistics (1 degree of freedom).

**Differs from asymptotic expectation by more than two standard deviations (see note a).

Table 5. Distribution of t-statistic under Null Hypotheses

	Two-Step Estimator		MLE	
	NL	OGEV	NL	OGEV
N=200				
Mean	0.032	0.075	0.273	0.350
Std. Deviation	0.968	0.973	1.032	1.291
Skewness	-0.070	0.185	1.286	1.768
Kurtosis	2.842	2.917	4.539	7.795
N=1000				
Mean	0.099	0.074	0.222	0.178
Std. Deviation	1.009	1.005	1.034	1.040
Skewness	-0.095	-0.099	0.516	0.493
Kurtosis	2.706	3.096	3.081	3.196
N = 10,000				
Mean	0.055		0.090	
Std. Deviation	0.965		0.967	
Skewness	-0.092		0.106	
Kurtosis	2.842		2.792	

Note: The asymptotic distribution of the t-statistic has mean 0, standard deviation 1, skewness 0, and kurtosis 3.

Table 6. Power of Tests against NL
with Asymptotic Significance Level .10[a]

	σ	NL Two-Step Estimator t-test[b]	NL Two-Step Estimator LR test	NL MLE: LR Test	Hausman-McFadden Test[c]
True Model NL:					
N = 200	.0	.110	.112	.109	.105
	.1	.168	.120	.118	.105
	.3	.488	.328	.328	.318
	.5	.910	.830	.830	.812
N = 1,000	.0	.120	.108	.108	.108
	.1	.322	.228	.228	.225
	.3	.932	.865	.865	.865
	.5	1.000	1.000	1.000	1.000
True Model OGEV:					
N = 200	.0	.110	.112	.109	.105
	.3	.170	.115	.115	.112
N = 1,000	.0	.120	.108	.108	.108
	.3	.308	.220	.221	.215

[a]Each entry is the proportion (of 400 replications) for which the test rejects the null hypothesis. Equivalently, each entry is the empirical value of $Pr[t > 1.282]$ or $Pr[\chi^2 > c_{.10}]$.

[b]One-sided test.

[c]Based on comparison of logit estimates: one with full choice set {1,...,6}, the other with restricted choice set {4,5,6}.

Table 7. Distribution of Estimators When Correct Model is Assumed (N = 1000)

	Estimator of β_2 (true value 1.00)		Estimator of σ (true value as given)	
	2-step (approx.)	MLE	2-step (approx.)	MLE
NL (true σ = 0.0)				
Mean	0.994	1.001	-0.015	-0.003
(Est. S.D.)[a]	(0.080)	(0.078)	(0.142)	(0.137)
Std. Dev.	0.081	0.083	0.145	0.151
Skewness	0.331	0.600	-0.073	-0.776
Kurtosis	3.115	3.553	2.826	3.697
NL (true σ = 0.1)				
Mean	0.992	1.001	0.121	0.097
Std. Dev.	0.086	0.081	0.147	0.132
NL (true σ = 0.3)				
Mean	0.939	1.002	0.414	0.297
(Est. S.D.)[a]	(0.103)	(0.072)	(0.146)	(0.079)
Std. Dev.	0.105	0.077	0.153	0.088
Skewness	0.198	0.519	-0.152	-0.819
Kurtosis	3.041	3.595	2.942	4.354
NL (true σ = 0.5)				
Mean	0.637	0.999	0.912	0.500
Std. Dev.	0.161	0.071	0.169	0.052
OGEV (true σ = .3)				
Mean	1.010	0.995	0.366	0.299
(Est. S.D.)[a]	(0.057)	(0.062)	(0.203)	(0.147)
Std. Dev.	0.053	0.060	0.204	0.153
Skewness	0.358	0.206	-0.073	-0.684
Kurtosis	3.388	3.206	3.066	3.935

[a]Median of the estimated standard deviations of $\hat{\beta}_2$ or of $\hat{\sigma}$. (In all cases shown the mean is within 5% of the median.)

Table 8. Root-Mean-Squared-Error of Utility-Parameter Estimates when True Model is NL

True σ	Estimator of β_1: Logit	Estimator of β_1: NL 2-step (approx.)	Estimator of β_1: NL MLE	Estimator of β_2: Logit	Estimator of β_2: NL 2-step (approx.)	Estimator of β_2: NL MLE
N = 200						
.0	.259	.385	.540	.111	.190	.245
.1	.272	.386	.457	.129	.200	.224
.3	.419	.438	.341	.248	.247	.200
.5	.841	.949	.243	.540	.511	.178
N = 1000						
.0	.116	.164	.172	.051	.081	.083
.1	.141	.167	.157	.072	.086	.081
.3	.347	.236	.127	.209	.122	.077
.5	.804	.785	.091	.510	.397	.071

Note: Each entry is the square root of the average over 400 replications of $(\hat{\beta}_i-\beta_i)^2$, where $\hat{\beta}_i$ is the estimator named in the column label.

APPENDIX A: COMPUTATIONS FOR SELECTED GEV MODELS

Nested Logit (2 levels). See Figure A1.

$$G(y;\sigma) = \sum_{r=1}^{R} \left(\sum_{j \epsilon B_r} y_j^{1/\rho_r} \right)^{\rho_r} \tag{A.1}$$

where $\sigma_r = 1-\rho_r$, $r=1,\ldots,R$. The choice probabilities are given by eqn. (12) in the text. Here the parameter vector θ is denoted by $\sigma = (\sigma_1,\ldots,\sigma_R)'$, and pseudovariables x by N^r, $r = 1,\ldots,R$, so that

$$\theta' x_j = \sum_r \sigma_r N_j^r . \tag{A.2}$$

From eqn. (6) in the text,

$$N_j^r = \begin{cases} V_j - I_r^\circ & j \in B_r \\ 0 & j \notin B_r , \end{cases} \tag{A.3}$$

where $I_r^\circ \equiv \log \sum_{\ell \epsilon B_r} \exp(V_\ell)$ is called the inclusive value of Subset B_r.

Eqn. (A.3) can also be written

$$N_j^r = \begin{cases} -\log \sum_{\ell \epsilon B_r} (P_\ell^\circ / P_j^\circ) & j \in B_r \\ 0 & j \notin B_r \end{cases} \tag{A.4}$$

where P_j° is just the logit choice probability (4).

Nested Logit (3 levels). See Figure A2.

$$G(y;\sigma,\phi) = \sum_{r=1}^{R} \left[\sum_{q \epsilon C_r} \left(\sum_{j \epsilon D_q} y_j^{1/\gamma_q} \right)^{(\gamma_q/\rho_r)} \right]^{\rho_r} \quad \text{(A.5)}$$

where $\sigma_r = 1-\rho_r$ for $r=1,\ldots,R$, and $\phi_q = 1-\gamma_q/\rho_r$ for $q=1,\ldots,Q$. Here $R < Q < J$, $\{D_q,\ r = 1,\ldots,Q\}$ is a partition of $\{1,\ldots,J\}$, and $\{C_r,\ r=1,\ldots,R\}$ is a partition of $\{1,\ldots,Q\}$. (This has, as a special case when $\phi = 0$, the 2-level NL model defined above, provided we adopt the notation that $B_r = \bigcup_{q \in C_r} D_q$.) For this model,

$$\theta' x_j = \sum_{r=1}^{R} \sigma_r N_j^r + \sum_{q=1}^{Q} \phi_q M_j^q \quad \text{(A.6)}$$

where N_j^r is given by (A.3) or (A.4),

$$M_j^q = \begin{cases} V_j - \tilde{I}_q^\circ & j \in D_q \\ 0 & j \notin D_q \end{cases}, \quad \text{(A.7)}$$

and $\tilde{I}^\circ = \log \sum_{\ell \in D_r} \exp(V_\ell)$ is an inclusive value for subset D_q.

Equivalently,

$$M_j^q = \begin{cases} -\log \sum_{\ell \in D_q} (P_e^\circ / P_j^\circ) & j \in D_q \\ 0 & j \notin D_q \end{cases}. \quad \text{(A.8)}$$

Note that restricting all ϕ_q to be equal is accomplished by replacing the pseudovariables M^q with the single pseudovariable M defined by:

$$M_j = \sum_{q=1}^{Q} M_j^q = V_j - \tilde{I}^\circ_{q(j)} \tag{A.9}$$

$$= -\log \sum_{\ell \in D_{q(j)}} (P^\circ_\ell / P^\circ_j) \tag{A.10}$$

where $q(j)$ is the node to which j is attached, i.e., $j \in D_{q(j)}$.

Ordered GEV

$$G(y;\sigma) = \sum_{r=1}^{J+K} \otimes \sum_{j \varepsilon B_r} w_{r-j} y_j^{1/\rho_r})^{\rho_r} \tag{A.11}$$

where $B_r = \{r-K,\ldots,r\} \cap \{1,\ldots,J\}$, $\sigma_r = 1-\rho_r$, $K>0$ is a positive integer, and $\sum_{m=0}^{K} w_m = 1$. Then

$$\theta' x_j = \sum_r \sigma_r N_j^r \tag{A.12}$$

$$N_j^r = \begin{cases} w_{r-j}(V_j - I^\circ_r) & j \in B_r \\ 0 & j \notin B_r \end{cases} \tag{A.13}$$

where $I^\circ_r = \log \sum_{\ell \varepsilon B_r} w_{r-\ell} \exp(V_\ell)$. Equivalently,

$$N_j^r = \begin{cases} -w_{r-j} \log \sum_{\ell \varepsilon B_r} w_{r-\ell}(P^\circ_\ell / P^\circ_j) & j \in B_r \\ 0 & j \notin B_r \end{cases} . \tag{A.14}$$

Nested Ordered GEV. See Figure A3.

This model is identical to the 3-level nested logit except the lowest level (represented by the term inside square brackets in Equation A.5) is ordered instead of nested. As in nested logit, the nesting is described by groups $\{B_r, r=1,\ldots,R\}$; but unlike nested logit, the alternatives in B_r must be labeled consecutively. I denote them by $B_r = \{J_{r-1}+1,\ldots,J_r\}$. The ordering within B_r is described by groups $D_q^r = \{q-K,\ldots,q\} \cap B_r$ for $q = \underline{Q}_r,\ldots,\bar{Q}_r$ where $\underline{Q}_r = J_{q-1}+1$ and $\bar{Q}_r = J_q+K$. (We define $J_0=0$.) That is:

$$G(y;\sigma,\phi) = \sum_{r=1}^{R}\left[\sum_{q=\underline{Q}_r}^{\bar{Q}_r}\left(\sum_{j\epsilon D_q^r} w_{q-j}\, y_j^{1/\gamma_q^r}\right)^{\gamma_q^r/\rho_r}\right]^{\rho_r} \tag{A.15}$$

where $\sigma_r = 1-\rho_r$ and $\phi_q^r = 1 - (\gamma_q^r/\rho_r)$. Then:

$$\theta' x_j = \sum_{r=1}^{R} \sigma_r N_j^r + \sum_{r=1}^{R} \sum_{q=\underline{Q}_r}^{\bar{Q}_r} \phi_q^r\, M_j^{rq} \tag{A.16}$$

where N_j^r is again given by (A.3) or (A.4), and

$$M_j^{rq} = \begin{cases} w_{q-j}(V_j - \tilde{I}_q^{r\circ}), & j \in D_q^r \\ 0 & j \notin D_q^r \end{cases} \tag{A.17}$$

$$= \begin{cases} -w_{q-j}\log \sum_{\ell\epsilon D_q^r} w_{q-\ell}(P_\ell^\circ/P_j^\circ), & j \in D_q^r \\ 0 & j \notin D_q^r . \end{cases} \tag{A.18}$$

Here $\tilde{I}_q^{r\circ} = \log \sum_{\ell\in D_q^r} w_{q-\ell}\exp(V_\ell)$ is an inclusive value for the subset D_q^r.

For the special case of identical ϕ_q^r, the last term of (A.16) becomes a single parameter ϕ multiplied by the pseudovariable M defined by

$$M_j = \sum_{r=1}^{R} \sum_{q=Q_r}^{\bar{Q}_r} M_j^{rq}$$

$$= V_j - \sum_{q=j}^{j+K} w_{q-j} \tilde{I}_q^{r(j)\circ} \qquad (A.19)$$

$$= \sum_{q=j}^{j+K} w_{q-j} \log \sum_{\ell \in D_q^{r(j)}} w_{q-\ell} (P_\ell^\circ / P_j^\circ) \qquad (A.20)$$

where $r(j)$ is defined uniquely by $j \in B_{r(j)}$.

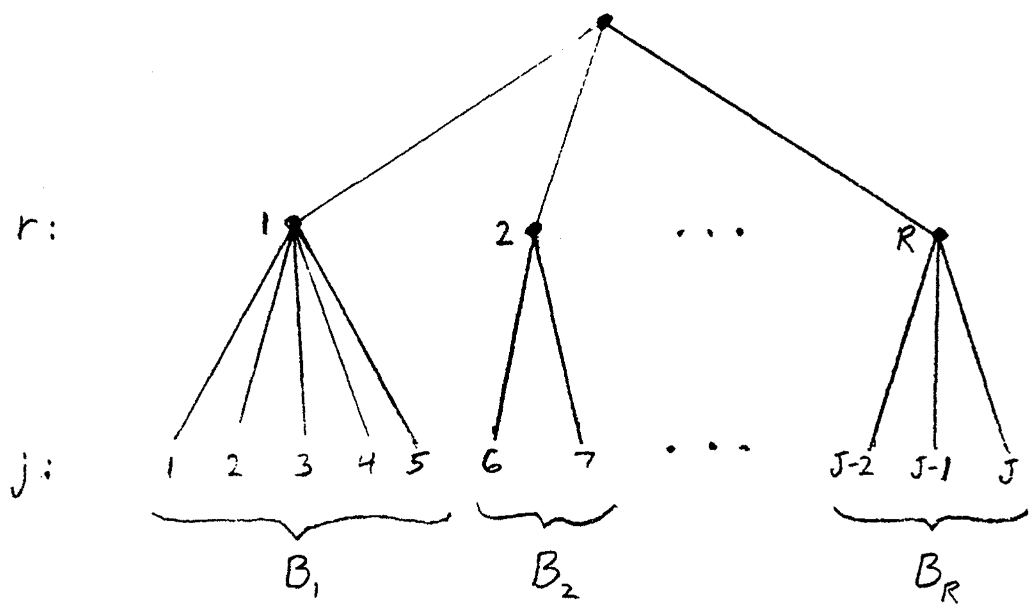

Figure A1. Two-Level Nested Logit

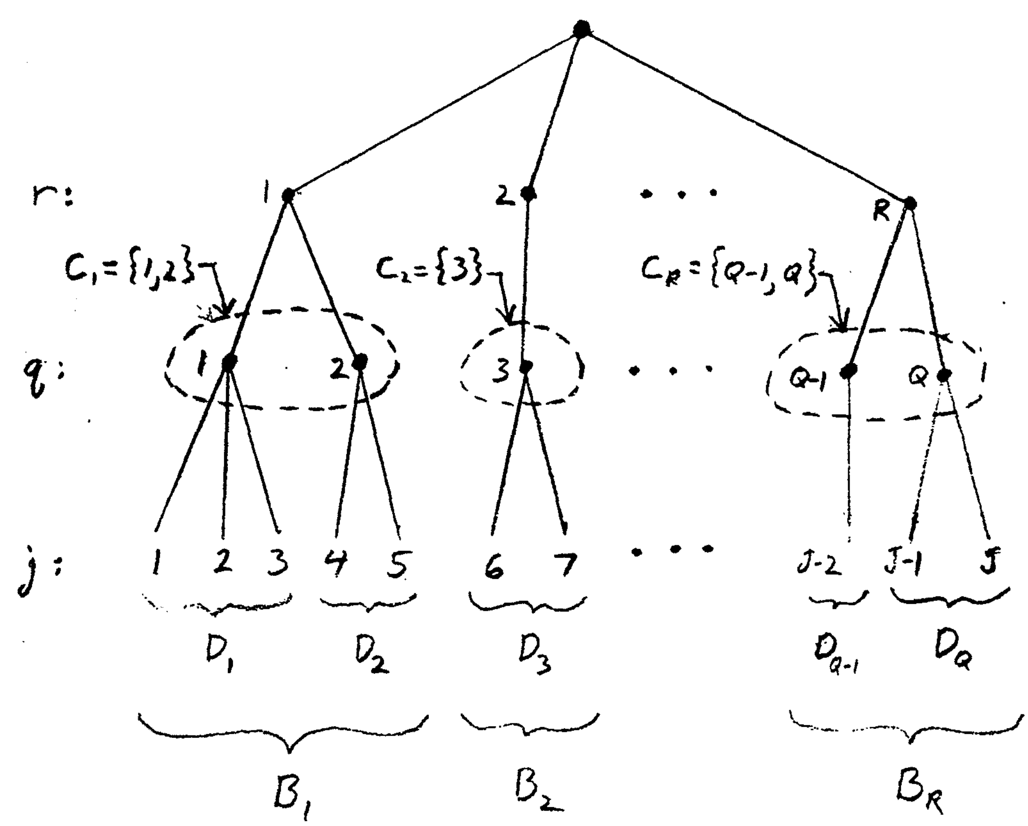

Figure A2. Three-Level Nested Logit

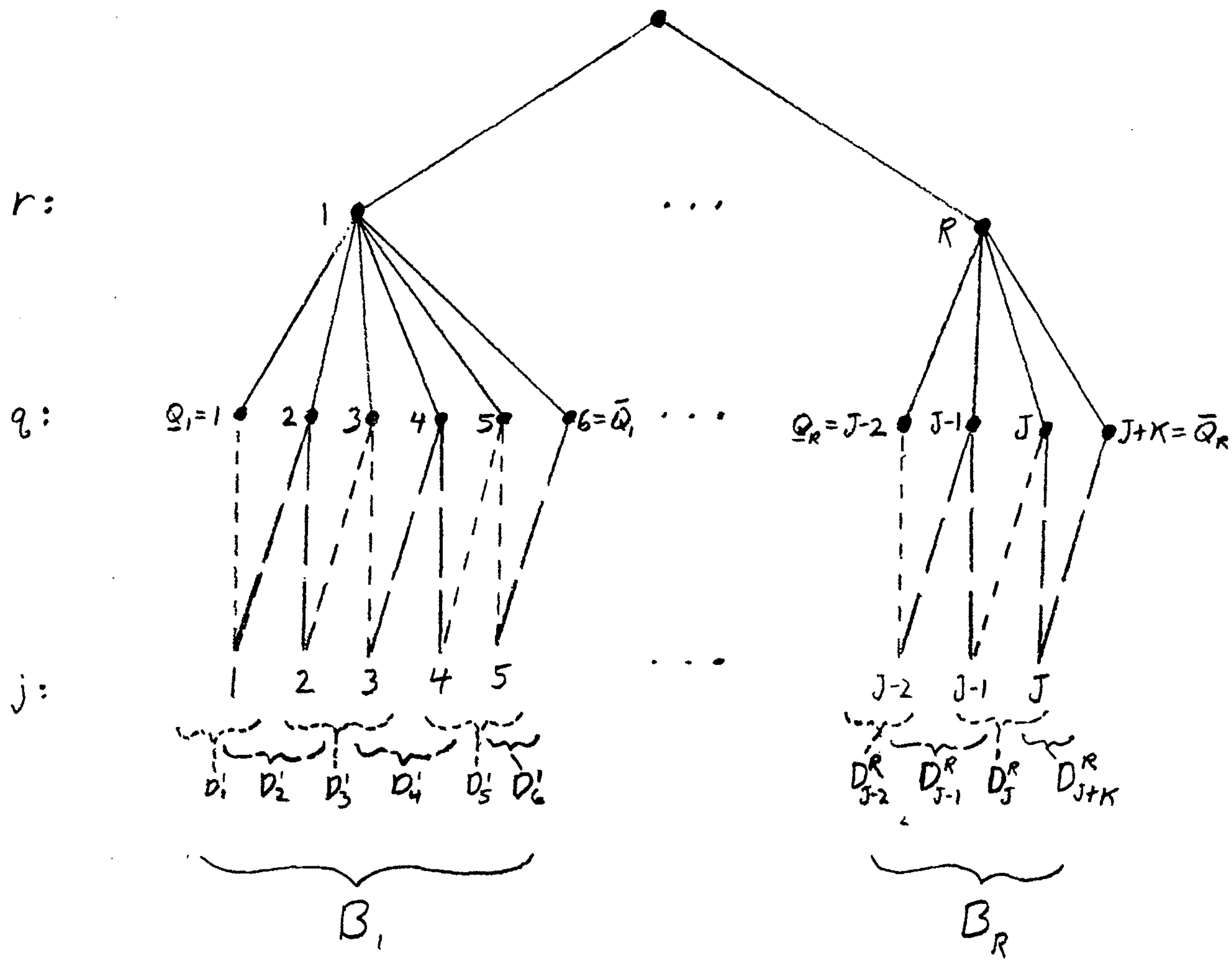

Figure A3. Nested Ordered GEV (K=1)

BIBLIOGRAPHY

Ben-Akiva, Moshe, and Steven R. Lerman, 1985, Discrete Choice Analysis: Theory and Application to Travel Demand (Cambridge, Mass: MIT Press).

Brownstone, David, and Kenneth A. Small, forthcoming, Efficient Estimation of Nested Logit Models, Journal of Business and Economic Statistics.

Goldfeld, Stephen M., and Richard E. Quandt, 1972, Nonlinear Methods in Econometrics (Amsterdam: North-Holland Publishing Company).

Hausman, Jerry, 1978, Specification Tests in Econometrics, Econometrica 46, 1251-1272.

Hausman, Jerry, and Daniel McFadden, 1984, Specification Tests for the Multinomial Logit Model, Econometrica 52, 1219-1240.

Horowitz, Joel, 1981, Identification and Diagnosis of Specification Errors in the Multinomial Logit Model, Transportation Research B, 15B, 345-360.

Maddala, G.S., 1983, Limited-Dependent and Qualitative Variables in Econometrics (Cambridge: Cambridge University Press).

McFadden, Daniel, 1978, Modelling the Choice of Residential Location, in Anders Karlqvist et al. (Eds.), Spatial Interaction Theory and Planning Models (Amsterdam: North-Holland Publishing Company), 75-96.

McFadden, Daniel, 1981, Econometric Models of Probabilistic Choice, in C.F. Manski and D. McFadden (Eds.), Structural Analysis of Discrete Data with Econometric Applications (Cambridge, Mass.: M.I.T. Press).

McFadden, Daniel, 1987, Regression-Based Specification Tests for the Multinomial Logit Model, Journal of Econometrics 34, 63-82.

McFadden, Daniel, Kenneth Train, and William B. Tye, 1977, An Application of Diagnostic Tests for the Independence from Irrelevant Alternatives Property of the Multinomial Logit Model, Transportation Research Record 637, 39-45.

Small, Kenneth A., 1987, A Discrete Choice Model for Ordered Alternatives, Econometrica 55, 409-424.

Small, Kenneth A., and Cheng Hsiao, 1985, Multinomial Logit Specification Tests, Internatioanal Economic Review 26, 619-627.